

Analisis KNN dan *Decision Tree* untuk Prediksi Curah Hujan di Tangerang Selatan

Rizky Agung Setiawan^{*1}, Bodi Santoso²

^{1,2,3}Universitas Pamulang; Kota Tangerang Selatan, Banten, (021) 7412566 / 74709855

^{1,2,3}Jurusan Teknik Informatika, Fakultas Ilmu Komputer, Universitas Pamulang

e-mail: *1agungrizky733@gmail.com, 2dosen00233@unpam.ac.id,

Abstrak

Fenomena Variabilitas curah hujan yang tinggi di wilayah perkotaan seperti Kota Tangerang Selatan menuntut adanya metode prediksi cuaca yang akurat dan andal. Penelitian ini bertujuan untuk membandingkan kinerja algoritma K-Nearest Neighbor (KNN) dan *Decision Tree* C4.5 dalam mengklasifikasikan kategori cuaca berdasarkan data curah hujan harian. Dataset yang digunakan merupakan data meteorologis dari BMKG periode 2022–2024 yang telah melalui tahap *preprocessing*, pelabelan kategori cuaca, dan pemilihan atribut. Pengujian model dilakukan menggunakan skenario tanpa dan dengan *cross-validation* (5-fold dan 10-fold). Evaluasi kinerja model dilakukan menggunakan metrik akurasi, *precision*, *recall*, *F1-score*, dan *kappa*. Hasil penelitian menunjukkan bahwa *Decision Tree* C4.5 menghasilkan performa dan akurasi yang lebih tinggi serta lebih stabil dibandingkan KNN pada seluruh skenario pengujian. Kontribusi utama penelitian ini adalah penyajian analisis komparatif kinerja KNN dan *Decision Tree* C4.5 menggunakan pendekatan *cross-validation* pada data curah hujan Kota Tangerang Selatan, sehingga memberikan gambaran metode klasifikasi yang lebih efektif untuk prediksi cuaca berbasis data historis.

Kata kunci—Prediksi, Hujan, Data, Mining, KNN, C4.5

I. PENDAHULUAN

Perubahan iklim global dan peningkatan variabilitas cuaca dalam beberapa tahun terakhir telah memicu terjadinya fenomena hujan ekstrem di berbagai wilayah Indonesia, termasuk kawasan perkotaan. Bencana banjir dipicu oleh adanya jumlah curah hujan ekstrem yang menyebabkan kerugian terutama di daerah dengan kepadatan penduduk tinggi dan sistem drainase terbatas (Ananda & Gaol, 2024). Kondisi ini menjadi tantangan serius bagi wilayah perkotaan seperti Kota Tangerang Selatan, mengingat bahwa Kota Tangerang Selatan adalah salah satu wilayah penyangga Daerah Khusus Jakarta menjadikan kota ini mendapat perhatian jika terjadi bencana banjir sehingga dalam beberapa tahun terakhir sering mengalami hujan deras yang berdampak pada aktivitas masyarakat dan infrastruktur (Pertiwi & Paski, 2022). Ketidakpastian pola curah hujan menyebabkan tantangan besar bagi perencanaan wilayah dan mitigasi bencana banjir (Supriatna & Uliyatunisa, 2025).

Kota Tangerang Selatan memiliki luas wilayah sekitar 164,85 km² atau sebesar 1.63% dari luas wilayah provinsi Banten dengan terbagi menjadi 7 kecamatan dan 54

kelurahan. Sebagai daerah perkotaan yang berkembang pesat, Tangerang Selatan memiliki tingkat kepadatan penduduk yang tinggi dan aktivitas sosial ekonomi yang dinamis. Menurut Badan Pusat Statistik Kota Tangerang Selatan (Badan Pusat Statistik, 2025) memiliki jumlah penduduk sebesar 1.429.529 jiwa pada tahun 2024 dengan kepadatan penduduk sebesar 8.672 jiwa per km², artinya bahwa di Kota Tangerang Selatan setiap 1 km² rata-rata dihuni oleh 8.672 orang penduduk. Kondisi ini menjadikan wilayah tersebut strategis dalam berbagai sektor, termasuk perencanaan tata ruang dan pengelolaan lingkungan, sehingga analisis terhadap fenomena cuaca seperti curah hujan menjadi sangat penting untuk mendukung perencanaan pembangunan dan mitigasi risiko bencana.

Dalam analisis prediksi, diperlukan informasi yang akurat berupa data historis yang kemudian dianalisis menggunakan pendekatan ilmiah. Prediksi berbasis data historis tersebut dapat memberikan dasar yang kuat dalam membantu upaya pengelolaan risiko (Susilawati Sugiana & Musty, 2023). Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) memiliki peran penting dalam menyediakan data cuaca dan iklim, termasuk data curah hujan harian yang digunakan dalam berbagai penelitian serta perencanaan kebijakan publik. Data dari BMKG menjadi sumber utama yang kredibel dalam menggambarkan kondisi meteorologis di

Indonesia, termasuk Tangerang Selatan. Selain itu, BMKG juga menetapkan klasifikasi intensitas curah hujan yang menjadi dasar dalam proses pelabelan data untuk keperluan analisis dan prediksi. Berdasarkan klasifikasi dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG), curah hujan dapat dibedakan menjadi beberapa kategori, antara lain cerah (0 mm/hari), hujan ringan (0–20 mm/hari), hujan sedang (21–50 mm/hari), dan hujan lebat (>50 mm/hari), (Limahelu et al., 2020). Oleh karena itu, dibutuhkan pendekatan ilmiah yang mampu mengklasifikasi cuaca berdasarkan intensitas curah hujan secara akurat untuk mengantisipasi kondisi-kondisi tersebut.

Seiring dengan berkembangnya teknologi informasi, data mining menjadi salah satu pendekatan modern yang dapat dimanfaatkan untuk menganalisis data curah hujan dalam jumlah besar. Data mining adalah teknik yang menggunakan data dalam jumlah besar untuk menemukan informasi berharga yang sebelumnya tidak terlihat dan dapat digunakan dalam pengambilan keputusan penting (Zidane et al., 2024). Dalam konteks cuaca, teknik *data mining* memungkinkan analisis yang lebih cepat, efisien, dan sistematis dibandingkan metode konvensional. Di antara teknik yang sering digunakan dalam klasifikasi dan prediksi cuaca adalah *K-Nearest Neighbor* dan *Decision Tree C4.5*. Kedua teknik ini memiliki pendekatan yang berbeda dalam mengenali pola data, sehingga menarik untuk dibandingkan secara langsung, khususnya dari segi keakuratan dan efisiensi ketika diterapkan pada data curah hujan harian.

Dalam penelitian ini, metode *K-Nearest Neighbor* dan *Decision Tree C4.5* dipilih karena keduanya banyak digunakan dalam permasalahan klasifikasi pada bidang meteorologi. Pada pendekatan ini, proses prediksi dilakukan dalam bentuk kategori cuaca yang dijadikan representasi tingkat intensitas curah hujan. Hal ini sejalan dengan pernyataan Fatmawati (2016) bahwa klasifikasi merupakan bagian dari prediksi, dimana nilai yang diprediksi berupa label. Selain itu, Miller et al. (2024) juga menegaskan bahwa algoritma klasifikasi mencakup semua metode machine learning yang menggunakan data berlabel untuk melatih sebuah algoritma dalam memprediksi kelas yang benar, di mana kelas merujuk pada seluruh data yang memiliki label tertentu. Dengan demikian, penggunaan klasifikasi kategori cuaca seperti cerah, hujan ringan, hujan sedang, dan hujan lebat, tetap berada dalam ruang lingkup prediksi, meskipun output yang dihasilkan tidak berupa nilai numerik curah hujan secara langsung. Pendekatan ini memungkinkan perbandingan performa algoritma *K-Nearest Neighbor* dan *Decision Tree C4.5* dalam mengenali pola curah hujan historis secara sistematis. Metode *K-Nearest Neighbor* melakukan proses klasifikasi berdasarkan kedekatan jarak antara data uji dan data pelatihan, sehingga mampu mengenali pola hujan yang berulang dari data historis tanpa perlu asumsi distribusi data tertentu (Agung et al., 2023). Pendekatan ini sesuai untuk data curah hujan yang bersifat numerik dan musiman. Sementara itu, *Decision Tree* memadukan antara eksplorasi data dan pemodelan, sehingga sangat bagus sebagai langkah awal dalam proses pemodelan bahkan ketika dijadikan sebagai

model akhir dari beberapa teknik (Fatimah & Khazizah, 2022). Karakteristik ini memungkinkan metode *C4.5* untuk mengelompokkan data curah hujan berdasarkan atribut-atribut meteorologis secara sistematis dan mudah ditelusuri kembali.

Penelitian terdahulu yang dilakukan oleh Andi Sandri Agung, dkk (2023) menerapkan algoritma *K-Nearest Neighbor* dengan beberapa nilai k pada data curah hujan harian. Hasilnya menunjukkan bahwa akurasi terbaik diperoleh pada $k = 3$ dengan nilai 82,21%, sedangkan penggunaan $k = 5$ menghasilkan akurasi pelatihan sebesar 84,54% dan akurasi pengujian 79,31%. Temuan ini mengindikasikan bahwa algoritma *KNN* cukup efektif digunakan dalam memprediksi kategori cuaca, terutama ketika nilai k dipilih secara tepat untuk menyesuaikan karakteristik data yang digunakan.

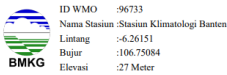
Penelitian lain oleh Al Arif, Firdaus, dkk (2022) melakukan perbandingan beberapa algoritma klasifikasi pada data cuaca, yaitu *Decision Tree C4.5*, *K-Nearest Neighbor*, dan *Naïve Bayes*. Hasil pengujian menunjukkan bahwa algoritma *C4.5* memiliki performa terbaik dengan akurasi 88,03%, diikuti oleh *KNN* dengan akurasi 85,61%, dan *Naïve Bayes* sebesar 81,15%. Penelitian tersebut menegaskan bahwa model berbasis pohon keputusan mampu menangkap pola data meteorologis secara lebih efektif, terutama ketika variabel input memiliki karakteristik non-linear dan saling berinteraksi.

Beberapa penelitian sebelumnya menunjukkan bahwa algoritma *K-Nearest Neighbor* dan *Decision Tree C4.5* mampu menghasilkan akurasi yang baik dalam klasifikasi cuaca berbasis data historis. Namun, sebagian besar penelitian tersebut belum secara khusus membandingkan kinerja kedua algoritma pada wilayah dengan karakteristik curah hujan yang memiliki variabilitas tinggi serta belum menguji kestabilan model menggunakan skenario *cross-validation* yang berbeda.

Oleh karena itu, penelitian ini difokuskan untuk mengisi celah tersebut dengan melakukan analisis perbandingan kinerja algoritma *K-Nearest Neighbor* dan *Decision Tree C4.5* dalam mengklasifikasikan kategori cuaca berdasarkan data curah hujan harian di Kota Tangerang Selatan periode 2022–2024.

II. METODE PELAKSANAAN

Penelitian ini menggunakan data sekunder yang diperoleh dari Badan Meteorologi, Klimatologi, dan Geofisika (BMKG) melalui layanan BMKG Online pada periode 2022–2024 untuk wilayah Kota Tangerang Selatan.



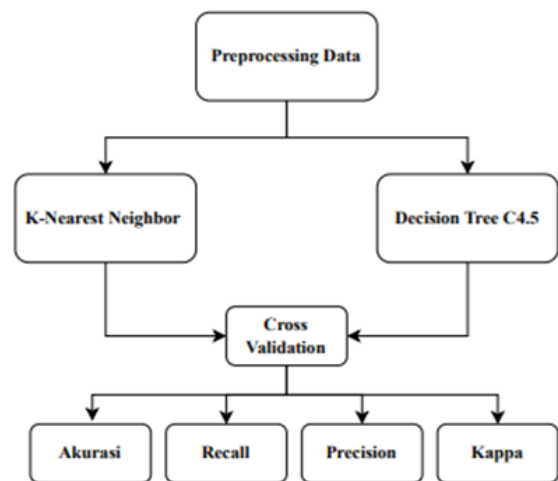
Tanggal	Tn	Tx	Tavg	RH_avg	RR	ss	ff_x	ddd_x	ff_avg	ddd_car
01-01-2024	24.7	32.1	30.2	78.0	21.0	6.6	2.0	310	1.0	C
02-01-2024	25.6	33.2	30.5	77.0	-	0.7	3.0	170	1.0	C
03-01-2024	25.2	30.0	27.9	86.0	10.7	1.0	2.0	290	1.0	C
04-01-2024	25.2	30.9	27.0	88.0	4.6	0.1	3.0	290	1.0	C
05-01-2024	23.7	35.5	29.1	78.0	63.8	0.0	5.0	230	1.0	C
06-01-2024	24.4	32.8	27.7	84.0	13.2	4.8	4.0	40	1.0	C
07-01-2024	23.4	31.6	27.3	86.0	75.2	4.4	3.0	270	1.0	SW
08-01-2024	24.0	32.0	29.0	78.0	7.0	4.2	3.0	170	1.0	C
09-01-2024	24.8	34.8	30.3	73.0	8888.0	4.4	5.0	250	2.0	W
10-01-2024	25.0	33.1	28.9	81.0	0.3	5.4	5.0	250	2.0	SW
11-01-2024	25.2	32.9	28.6	81.0	8888.0	4.9	3.0	270	1.0	W
12-01-2024	24.6	32.6	28.8	79.0	3.6	4.2	4.0	250	2.0	W
13-01-2024	24.6	32.8	28.6	81.0	-	3.0	3.0	10	1.0	C

Data mentah yang ditampilkan berisi beberapa kolom utama seperti tanggal, temperatur rata-rata (Tavg), kelembapan rata-rata (RH_avg), lama penyinaran matahari (ss), kecepatan angin rata-rata (ff_avg), dan curah hujan (RR). Pada tampilan data mentah tersebut juga dicantumkan keterangan dari BMKG, yaitu:

KETERANGAN:

- 8888: Data tidak terukur
- 9999: Tidak Ada Data (tidak dilakukan pengukuran)
- Tn: Temperatur minimum (°C)
- Tx: Temperatur maksimum(°C)
- Tavg: Temperatur rata-rata(°C)
- RH_avg: Kelembapan rata-rata(%)
- RR: Curah hujan(mm)
- ss: Lamanya penyinaran matahari(jam)
- ff_x: Kecepatan angin maksimum(m/s)
- ddd_x: Arah angin saat kecepatan maksimum(°)
- ff_avg: Kecepatan angin rata-rata(m/s)
- ddd_car: Arah angin terbanyak(°)

Variabel yang terdapat dalam dataset meliputi temperatur rata-rata harian (Tavg), kelembapan rata-rata (RH_avg), kecepatan angin rata-rata (ff_avg), lama penyinaran matahari (ss), serta curah hujan harian (RR). Nilai RR menjadi acuan utama dalam proses pelabelan kategori cuaca. Pelabelan dilakukan mengacu pada standar BMKG, yaitu cerah (0 mm), hujan ringan (0–20 mm), hujan sedang (21–50 mm), dan hujan lebat (>50 mm). Dengan demikian, seluruh atribut numerik seperti Tavg, RH_avg, ff_avg, dan ss berfungsi sebagai variabel prediktor, sedangkan kategori cuaca hasil pelabelan RR menjadi variabel target yang akan diprediksi oleh model klasifikasi. Sebelum memasuki tahap pemodelan, dilakukan proses preprocessing agar struktur data siap digunakan. Berikut adalah alur metode penelitian ini:



Tahapan ini dimulai dengan mengubah seluruh nilai 8888, 9999, dan “-” menjadi NaN agar dapat diperlakukan sebagai missing value. Setelah itu, proses penanganan missing value dilakukan menggunakan dua pendekatan, yaitu mean imputation dan interpolasi, sehingga celah data pada variabel meteorologis dapat terisi dengan nilai representatif. Tahap berikutnya adalah deteksi outlier menggunakan metode Interquartile Range (IQR), di mana nilai ekstrem pada variabel Tavg, RH_avg, ff_avg, dan ss ditangani agar tidak memengaruhi pola pembelajaran model. Khusus variabel RR (curah hujan), outlier tidak dihapus karena nilai ekstrem justru merepresentasikan kejadian hujan intensitas tinggi yang bersifat valid secara meteorologis. Setelah itu, dilakukan normalisasi menggunakan z-score agar seluruh variabel numerik berada dalam skala seragam sehingga algoritma yang sensitif terhadap jarak dapat bekerja optimal.

Tahap pemodelan dilakukan menggunakan dua algoritma klasifikasi, yaitu K-Nearest Neighbor (KNN) dan Decision Tree C4.5. RapidMiner digunakan sebagai platform utama karena menyediakan alur proses visual melalui rangkaian operator terstruktur yang memudahkan pemantauan setiap langkah seperti partitioning, training, testing, dan evaluasi. KNN bekerja berdasarkan kedekatan jarak antara data uji dan data pelatihan, sementara Decision Tree C4.5 membangun pohon keputusan menggunakan gain ratio untuk menentukan atribut pemisah terbaik pada setiap node. Kedua algoritma ini dipilih karena banyak digunakan dalam penelitian meteorologi dan memiliki karakteristik berbeda sehingga menarik untuk dibandingkan.

Evaluasi model dilakukan melalui dua skema pengujian, yaitu pengujian tanpa cross-validation dengan pembagian data 80:20 serta pengujian cross-validation 5-fold dan 10-fold untuk memastikan performa model stabil pada berbagai struktur pembagian data. Metrik evaluasi yang digunakan meliputi akurasi, precision, recall, F1-score, dan kappa. Akurasi digunakan untuk mengukur keseluruhan ketepatan prediksi, precision menilai ketepatan prediksi setiap kelas, recall mengukur kemampuan model mengenali data sebenarnya, F1-score menjadi indikator gabungan precision

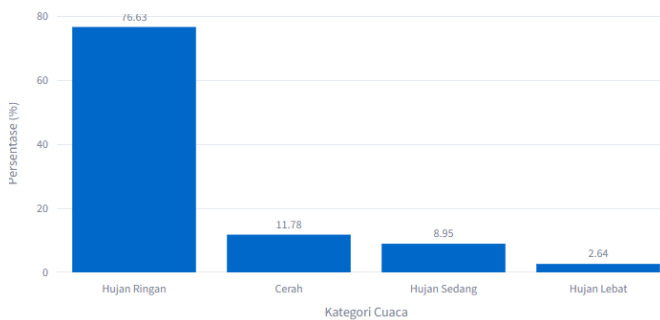
dan recall, sedangkan kappa digunakan untuk mengukur reliabilitas prediksi dengan mempertimbangkan peluang prediksi benar secara acak.

Seluruh tahapan mulai dari pengumpulan data, pembersihan data, penanganan missing value, normalisasi, pemodelan, hingga evaluasi dilakukan untuk memperoleh gambaran yang akurat mengenai performa KNN dan Decision Tree C4.5 dalam mengklasifikasikan kategori cuaca berdasarkan data curah hujan harian di Kota Tangerang Selatan serta menentukan model yang paling stabil dan efektif dalam membaca pola historis curah hujan.

III. HASIL DAN PEMBAHASAN

Dataset yang digunakan dalam penelitian ini berjumlah 1.061 baris data harian yang diperoleh dari BMKG periode 2022–2024 dengan 7 jumlah kolom. Setelah dilakukan preprocessing yang meliputi penghapusan nilai tidak valid seperti 8888, 9999, dan karakter “-”, seluruh nilai tersebut dikonversi menjadi NaN agar dapat ditangani pada tahap berikutnya. Penanganan missing value dilakukan dengan dua teknik, yaitu median imputasi untuk atribut yang memiliki banyak nilai kosong dan interpolasi linier untuk atribut deret waktu seperti suhu rata-rata, kelembapan, dan lamanya penyinaran matahari. Selain itu, deteksi outlier dilakukan menggunakan metode Interquartile Range (IQR) kecuali untuk atribut curah hujan karena nilai ekstrem pada curah hujan tidak dapat dianggap sebagai anomali, melainkan merupakan bagian dari fenomena cuaca itu sendiri. Setelah seluruh proses pembersihan selesai, dilakukan normalisasi agar rentang nilai antar atribut lebih beragam sebelum masuk ke tahap pemodelan.

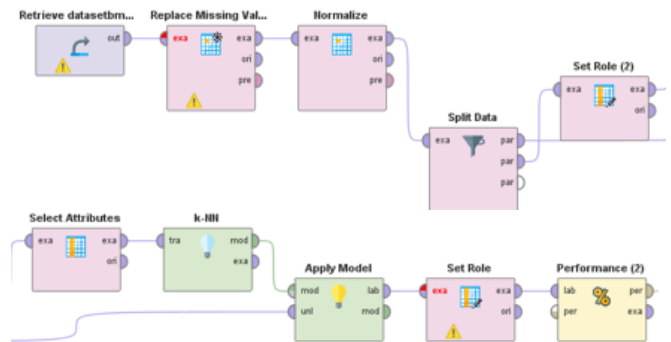
Persentase Data Berdasarkan Kategori Cuaca



Sebaran kategori cuaca menunjukkan bahwa dataset ini memiliki ketidakseimbangan kelas. Kategori Hujan Ringan mendominasi sekitar 76,6% dari total data, diikuti oleh Hujan Sedang sebesar 8,9%, Cerah sebesar 11%, dan Hujan Lebat sebesar 2,6%. Ketidakseimbangan ini berpotensi memengaruhi performa algoritma tertentu, terutama metode yang sensitif terhadap distribusi data seperti K-Nearest Neighbor.

Setelah dataset siap, proses pemodelan dilakukan menggunakan perangkat lunak RapidMiner. Pada tahap awal, dilakukan pengujian menggunakan pembagian data

80% sebagai data latih dan 20% sebagai data uji untuk memperoleh gambaran awal kinerja masing-masing model.

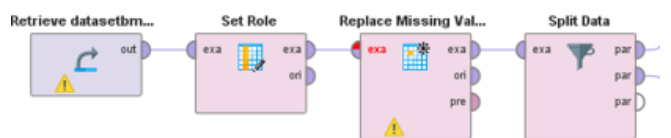


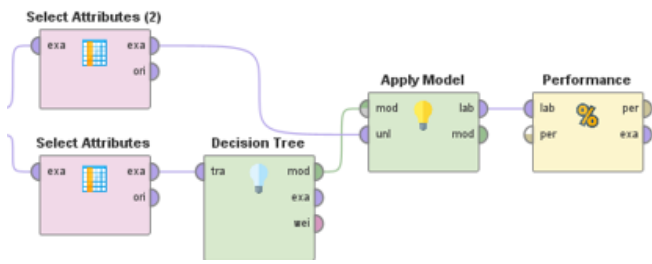
Pemodelan K-Nearest Neighbor dimulai dari proses pengambilan dataset, penanganan missing value, normalisasi, seleksi atribut, penentuan label, hingga proses pelatihan model dengan parameter $k = 3$ menggunakan jarak Euclidean. Alur proses ini menghasilkan model KNN yang kemudian diterapkan pada data uji untuk menghasilkan metrik evaluasi awal dengan hasil sebagai berikut ini:

	true Hujan Ringan	true Cerah	true Hujan Sedang	true Hujan Lebat	class precision
pred Hujan Ringan	141	15	24	0	78.33%
pred Cerah	8	0	0	0	0.00%
pred Hujan Sedang	14	0	2	2	11.11%
pred Hujan Lebat	4	0	1	1	16.67%
class recall	84.43%	0.00%	7.41%	33.33%	

Berdasarkan hasil pengujian, algoritma K-Nearest Neighbor memperoleh nilai akurasi sebesar 67,92%, dengan classification error sebesar 32,08%. Selain itu, diperoleh pula nilai weighted mean recall sebesar 31,29% dan weighted mean precision sebesar 26,53% sebagai indikator tambahan kinerja model pada seluruh kelas. Pada metrik Cohen's kappa, model menghasilkan nilai sebesar -0,011. Seluruh hasil evaluasi tersebut menjadi gambaran performa awal model K-Nearest Neighbor dalam melakukan klasifikasi kategori cuaca berdasarkan data intensitas curah hujan harian pada penelitian ini, yang kemudian akan dibandingkan lebih lanjut dengan performa model Decision Tree C4.5.

Pemodelan Decision Tree C4.5 mengikuti alur yang hampir serupa, namun menggunakan pendekatan rule-based dengan kriteria pemilihan atribut Gain Ratio, sesuai karakteristik algoritma C4.5.



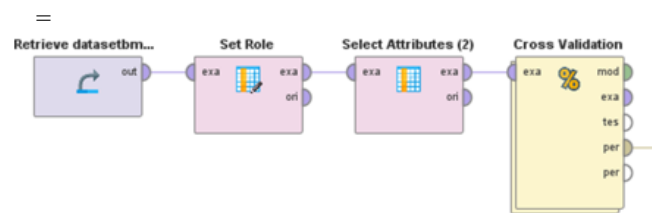


Parameter yang digunakan meliputi nilai confidence = 0.25, minimum leaf size, dan pemilihan atribut terbaik pada setiap node. Model kemudian diuji menggunakan data uji hasil pembagian 80:20 untuk melihat performa awal. Didapatkan hasil sebagai berikut:

	True Hujan Ringan	True Cerah	True Hujan Sedang	True Hujan Lebat	class precision
pred Hujan Ringan	148	16	17	7	78.72%
pred Cerah	6	12	1	1	60.00%
pred Hujan Sedang	2	0	1	0	33.33%
pred Hujan Lebat	1	0	0	0	0.00%
class recall	94.27%	42.86%	5.26%	0.00%	

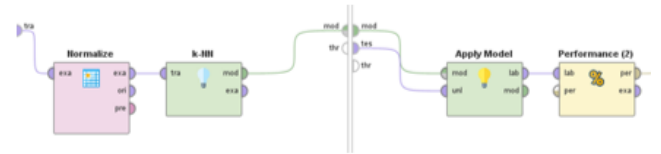
Berdasarkan hasil pengujian, algoritma Decision Tree C4.5 memperoleh nilai akurasi sebesar 75,94%, dengan classification error sebesar 24,06%. Selain itu, diperoleh nilai weighted mean recall sebesar 35,60% dan weighted mean precision sebesar 43,01% dalam mengukur ketepatan dan kemampuan model dalam mengenali setiap kelas. Pada metrik Cohen's kappa, model menghasilkan nilai sebesar 0,270 sebagai indikator tambahan performa klasifikasi. Hasil ini akan menjadi bahan perbandingan dengan model K-Nearest Neighbor untuk menentukan algoritma dengan kinerja terbaik pada penelitian ini.

Setelah mendapatkan hasil pengujian awal menggunakan pembagian data 80:20, tahapan berikutnya adalah melakukan validasi model menggunakan teknik k-Fold Cross-Validation untuk memperoleh hasil evaluasi yang lebih stabil dan tidak bergantung pada satu skenario pembagian data saja, K-fold yang dipilih adalah 10-Fold dan 5-Fold. Proses ini dilakukan secara langsung di RapidMiner dengan membangun workflow khusus cross-validation yang terdiri dari dua bagian utama, yaitu operator Cross Validation dan sub-process Training dan Testing di dalamnya. Alur lengkap proses cross-validation dapat dilihat pada diagram berikut:



Gambar diatas menunjukkan alur proses permodelan dengan cross validation dari kedua model. Terjadi

perbedaan didalam bagian sub-process Training dan Testing yaitu sebagai berikut:



Alur diatas menunjukkan alur didalam operator cross validation. Pada algoritma K-Nearest Neighbor diterapkan cross validation dengan dua skenario pengujian, yaitu menggunakan 10-fold cross validation dan 5-fold cross validation. Kedua skenario tersebut bertujuan untuk membandingkan konsisten kinerja model berdasarkan jumlah pembagian data yang berbeda. alur proses utama penerapan cross validation dalam RapidMiner di tingkat luar workflow. Proses dimulai dari input dataset, lalu set role untuk menjadi cuaca sebagai cuaca sebagai label, setelah itu select attributes digunakan untuk mengecualikan atribut cuaca, curah hujan, dan tanggal, dan cross validation. Selanjutnya alur didalam cross validation, pada tahap training, data terlebih dahulu diproses melalui operator Normalize untuk memastikan seluruh fitur berada pada skala yang seragam. Setelah itu, data dilatih menggunakan algoritma K-Nearest Neighbor untuk membangun model klasifikasi. Pada tahap testing, model yang telah terbentuk diaplikasikan menggunakan operator Apply Model terhadap data uji dari setiap fold. Hasil prediksi kemudian dianalisis menggunakan operator Performance untuk memperoleh metrik evaluasi model.



Alur diatas memaparkan alur operator cross validation Decision Tree C4.5. Pada algoritma Decision Tree C4.5 dilakukan proses skenario yang serupa dengan algoritma sebelumnya, yaitu 10-fold dan 5-fold cross validation. Pendekatan ini digunakan agar performa model dapat dinilai lebih menyeluruh dengan pembagian data pelatihan serta pengujian yang dilakukan secara berulang berdasarkan jumlah fold yang berbeda. rangkaian proses utama sebelum data masuk ke tahap pelatihan dan pengujian berulang di dalam cross validation. Tahapan dimulai dari pemanggilan dataset, kemudian penetapan atribut cuaca sebagai label melalui set role. Selanjutnya, select attributes digunakan untuk menyaring atribut yang dipakai dalam proses pemodelan sehingga hanya atribut prediktor yang diteruskan ke operator Cross Validation. ditunjukkan alur yang berjalan di dalam operator cross validation. Seluruh data yang masuk ke tahap training langsung diproses oleh algoritma Decision Tree C4.5, karena metode ini tidak membutuhkan

normalisasi fitur. Model yang dihasilkan kemudian diuji pada bagian data yang berperan sebagai testing set pada setiap iterasi fold melalui operator Apply Model. Setelah proses prediksi selesai, operator Performance digunakan untuk menghasilkan metrik evaluasi, sehingga performa klasifikasi dapat dianalisis secara lebih objektif pada tiap pengujian.

Setelah implementasi permodelan diatas maka hasil penelitian ini menunjukkan hasil kinerja dan perbandingan dari kedua metode. Proses awal pengujian mode dilakukan dengan pembagian data secara tunggal (singl train-test split) untuk memperoleh gambaran awal kinerja masing-masing algoritma. Tahap ini bertujuan untuk meliht seberapa baik model dapat melakukan klasifikasi cuacaberdasarkan data historis curah hujan tanpa validasi berulang.

Model	Akurasi	Recall	Precision	Kappa
C4.5	75.94%	35.60%	43.01%	0.270
KNN	67.92%	31.29%	26.53%	-0.011

Berikut adalah hasil yang didapatkan pada algoritma K-Nearest Neighbor dengan 10-Fold dan 5-Fold:

F	K-Nearest Neighbor			
	Akurasi	Recall	Precision	Kappa
5-F	56.03%	29.77%	22.11%	0.076
	± 26.42%	± 6.97%	± 11.14%	± 0.118
10-F	60.41%	32.77%	25.29%	0.127
	± 20.20%	± 5.87%	± 9.76%	± 0.120

Berikut adalah hasil yang didapatkan pada algoritma Decision Tree dengan 10-Fold dan 5-Fold:

F	Decision Tree C4.5			
	Akurasi	Recall	Precision	Kappa
5-F	78.70%	35.13%	49.57%	0.235
	± 1.47%	± 4.06%	± 13.67%	± 0.058
10-F	78.70%	35.50%	43.92%	0.247
	± 1.72%	± 5.21%	± 12.39%	± 0.082

Dengan hasil yang didapatkan maka dilakukan analisis perbandingan hasil algoritma dengan metrik akurasi, weight

mean recall, weighted mean precision, kappa. Serta menganalisis stabilitas model.

1. Akurasi

Hasil evaluasi menunjukkan bahwa algoritma Decision Tree C4.5 secara konsisten unggul dalam hal akurasi dibandingkan K-Nearest Neighbor. baik pada pengujian tanpa cross validation maupun pada skenario cross validation. Perbedaan performa ini terlihat jelas dari kemampuan kedua model dalam menggeneralisasi pola data. Pada pengujian tanpa cross validation, K-Nearest Neighbor memperoleh akurasi 67.92%, sedangkan Decision Tree C4.5 mencapai 75.94%. Sejak tahap awal, terlihat bahwa Decision Tree C4.5 mampu menangkap karakteristik data dengan lebih baik. Hal ini mengindikasikan bahwa struktur aturan keputusan yang dibangun oleh Decision Tree C4.5 lebih selaras dengan pola curah hujan yang terdapat pada dataset.

Ketika dilakukan pengujian menggunakan 10-fold cross validation, akurasi K-Nearest Neighbor menurun menjadi 60.41% ± 20.20%. Penyimpangan akurasi yang sangat besar ini menunjukkan bahwa performa K-Nearest Neighbor sangat dipengaruhi variasi komposisi data pelatihan pada setiap fold. Sebaliknya, Decision Tree C4.5 tetap menunjukkan performa tinggi dan stabil dengan akurasi 78.70% ± 1.72%, yang menandakan bahwa model ini memiliki kemampuan generalisasi kuat pada data yang tidak dilibatkan saat pelatihan.

Pada 5-fold cross validation, performa K-Nearest Neighbor kembali menurun dengan akurasi 56.03% ± 26.42%, dan variasinya semakin besar dibandingkan 10-fold. Ini mengindikasikan bahwa semakin sedikit data latih yang tersedia, semakin sulit bagi K-Nearest Neighbor untuk membentuk batas keputusan yang baik. Sementara itu, Decision Tree C4.5 tetap konsisten dengan akurasi 78.70% ± 1.47%, menunjukkan bahwa jumlah fold tidak memberikan dampak signifikan terhadap kualitas model.

Berdasarkan seluruh hasil tersebut, terdapat beberapa alasan teknis yang menjelaskan dominasi Decision Tree C4.5 dibandingkan K-Nearest Neighbor. Pertama, dataset yang digunakan memiliki ketidakseimbangan kelas yang cukup tinggi, dengan kelas Hujan Ringan mendominasi sekitar 76.6% dari total data. Pada kondisi tersebut, K-Nearest Neighbor cenderung bias terhadap kelas mayoritas karena proses klasifikasi berbasis kedekatan titik tetangga. Sementara Decision Tree C4.5 mampu menghasilkan aturan pemisahan yang lebih adil antar kelas. Kedua, K-Nearest Neighbor sangat sensitif terhadap variasi skala fitur dan keberadaan noise pada data, sehingga performanya tidak stabil saat distribusi data berubah pada setiap fold cross validation. Sebaliknya, Decision Tree C4.5 lebih mampu mengatasi variasi nilai dan menghasilkan pemisahan fitur yang optimal. Ketiga, pola pada data curah hujan bersifat non-linear, sehingga model berbasis pohon keputusan lebih mampu menggambarkan struktur batas keputusan yang kompleks dibandingkan K-Nearest Neighbor yang hanya mengandalkan jarak.

2. Weighted Mean Recall

Hasil menunjukkan bahwa algoritma K-Nearest Neighbor memiliki nilai recall sebesar 31.29% tanpa cross validation. Ketika dilakukan skenario cross validation dengan 10-fold didapatkan hasil recall $32.77\% \pm 5.87\%$ yang menandakan adanya peningkatan, namun saat skenario 5-fold performa justru menurun menjadi $29.77\% \pm 6.97\%$.

Sedangkan algoritma Decision Tree C4.5 menunjukkan performa lebih baik dan konsisten dibandingkan dengan K-Nearest Neighbor, dengan nilai recall sebesar 35.60% pada pengujian tanpa cross validation. Kemudian dilakukan skenario cross validation dengan jumlah 10-fold didapatkan nilai recall sebesar $35.60\% \pm 5.21\%$, serta $35.13\% \pm 4.06\%$ pada skenario 5-fold. Perbedaan angka yang tidak terlalu jauh menegaskan bahwa Decision Tree C4.5 mampu mempertahankan stabilitas model secara stabil meskipun jumlah pembagian data berubah.

Kedua algoritma masih menghasilkan nilai recall yang cukup rendah, hal ini menunjukkan bahwa model sering gagal mengenali kelas hujan yang sebenarnya terjadi, khususnya pada kelas dengan jumlah data lebih sedikit seperti kategori hujan sedang atau hujan lebat. Penyebab utama kondisi ini adalah ketidakseimbangan kelas dalam dataset, di mana kelas hujan ringan jauh lebih dominan sehingga model cenderung selalu mengarah pada prediksi kelas tersebut.

Kelemahan ini paling terlihat pada K-Nearest Neighbor yang sangat bergantung pada data tetangga di sekitarnya. Ketika mayoritas data mengelompok pada kelas tertentu, model cenderung memberikan keputusan yang sama meskipun sebenarnya kondisi cuaca dapat berbeda. Deviasi nilai recall yang lebih besar pada K-Nearest Neighbor juga menunjukkan bahwa performa model sangat dipengaruhi oleh perubahan distribusi data pada setiap proses pelatihan.

Di sisi lain, Decision Tree C4.5 masih memiliki keterbatasan dalam memahami pola cuaca yang lebih kompleks, sehingga tidak mampu meningkatkan recall secara signifikan meskipun pembagian data dilakukan berulang. Hal ini memberi gambaran bahwa untuk meningkatkan kemampuan deteksi kejadian cuaca yang lebih beragam, diperlukan pengembangan lanjutan seperti penyeimbangan data, pemilihan fitur yang lebih relevan, atau penambahan teknik pra-pemrosesan yang lebih komprehensif sebelum pelatihan model dilakukan.

3. Weighted Mean Precision

Dalam evaluasi weighted mean precision, terlihat dengan jelas bahwa kedua algoritma memiliki kemampuan berbeda dalam mengendalikan kesalahan prediksi. Pada pengujian awal tanpa cross validation, algoritma K-Nearest Neighbor hanya mampu mencapai precision sebesar 26,53%. Nilai ini menunjukkan bahwa sebagian besar prediksi yang diberikan model tidak tepat sesuai dengan kelas cuaca sebenarnya. Ketika dilakukan cross validation 10-fold, precision K-Nearest Neighbor tidak menunjukkan adanya perbaikan, justru turun menjadi $25,29\% \pm 9,76\%$, dan bahkan semakin menurun pada 5-fold yaitu $22,11\% \pm 11,14\%$. Fluktuasi precision yang cukup besar pada setiap skenario ini

menegaskan bahwa K-Nearest Neighbor sangat bergantung pada struktur tetangga terdekat yang dipengaruhi oleh perubahan komposisi data pada masing-masing fold, sehingga menghasilkan performa yang tidak stabil.

Berbeda dengan K-Nearest Neighbor, algoritma Decision Tree C4.5 menunjukkan performa yang jauh lebih baik dalam menghasilkan prediksi yang benar terhadap kelas yang dipilih model. Precision tanpa cross validation mencapai 43,01%, dan ketika divalidasi 10-fold meningkat menjadi $43,92\% \pm 12,39\%$. Bahkan pada 5-fold, precision mencapai nilai tertinggi yaitu $49,57\% \pm 13,67\%$. Kondisi ini mengindikasikan bahwa Decision Tree C4.5 mampu melakukan pemilahan fitur yang lebih efektif melalui pembentukan aturan keputusan yang spesifik terhadap pola data, sehingga kesalahan prediksi dapat diminimalkan.

Walaupun demikian, besarnya deviasi pada hasil precision Decision Tree C4.5 menunjukkan bahwa model masih cukup terpengaruh oleh variasi distribusi data, terutama karena adanya ketimpangan jumlah kelas yang sangat signifikan. Kelas hujan ringan yang mendominasi sekitar $\pm 76.6\%$ data membuat model cenderung lebih yakin saat melakukan prediksi pada kelas tersebut dibandingkan kelas dengan jumlah sangat sedikit seperti hujan sedang dan hujan lebat.

Secara keseluruhan, hasil weighted mean precision memperlihatkan bahwa Decision Tree C4.5 secara konsisten memberikan prediksi yang lebih tepat dibandingkan K-Nearest Neighbor, sekaligus menjaga kemampuan model dalam memilih kelas yang benar meskipun kondisi distribusi data tidak ideal. Dengan demikian, Decision Tree C4.5 dapat dinilai lebih handal sebagai model prediksi dalam konteks penelitian ini, khususnya dalam mengurangi prediksi salah sasaran pada data cuaca.

4. Kappa

Dalam pengujian yang mempertimbangkan pengaruh prediksi benar yang terjadi secara kebetulan (by chance), nilai Kappa memberikan gambaran yang lebih adil mengenai kemampuan model dalam mengklasifikasikan data yang tidak seimbang jumlah kelasnya. Pada algoritma K-Nearest Neighbor, nilai Kappa tanpa cross validation menunjukkan kinerja yang sangat rendah, yaitu -0,011, yang berarti model hampir tidak lebih baik dari sekadar menebak secara acak. Hasil ini menunjukkan bahwa K-Nearest Neighbor gagal melakukan pembelajaran pola yang konsisten pada dataset ini.

Ketika diterapkan cross validation, kinerja K-Nearest Neighbor sedikit meningkat, dengan nilai Kappa $0,127 \pm 0,120$ pada 10-fold, dan $0,076 \pm 0,118$ pada 5-fold. Namun, peningkatan tersebut tetap belum cukup untuk menunjukkan bahwa model memiliki kemampuan klasifikasi yang kuat dan dapat dipercaya. Fluktuasi nilai yang cukup besar ini juga mengindikasikan bahwa performa K-Nearest Neighbor tidak stabil karena sangat sensitif terhadap komposisi data pelatihan di tiap fold validasi.

Berbanding terbalik, hasil dari algoritma Decision Tree C4.5 menunjukkan performa yang jauh lebih baik dalam menangani ketidakseimbangan kelas. Nilai Kappa tanpa

cross validation mencapai 0,270, menandakan bahwa model mampu memberikan prediksi yang lebih baik daripada sekadar tebakan acak dengan tingkat konsistensi yang cukup kuat. Pada 10-fold cross validation, nilai Kappa memang sedikit menurun menjadi $0,247 \pm 0.082$, namun penurunan ini masih dalam batas wajar dan tetap menunjukkan konsistensi performa model. Pada 5-fold validasi, nilai Kappa $0,235 \pm 0.058$ memperkuat bahwa Decision Tree C4.5 tetap stabil meskipun jumlah data pelatihan di tiap fold berbeda.

Dari hasil ini, dapat disimpulkan bahwa Decision Tree C4.5 secara signifikan lebih unggul dibandingkan K-Nearest Neighbor dalam hal keandalan prediksi, terutama ketika mempertimbangkan faktor peluang prediksi benar yang terjadi secara acak. Dengan kata lain, Decision Tree C4.5 bukan hanya memberikan hasil yang lebih baik, tetapi juga lebih konsisten dan dapat diandalkan dalam menangani data cuaca yang memiliki distribusi kelas tidak merata.

5. Stabilitas Model

Stabilitas model dapat diamati dari konsistensi performanya ketika dilakukan pengujian berulang melalui metode cross validation. Model yang stabil ditandai dengan perubahan nilai evaluasi yang tidak terlalu besar ketika dataset dilatih dan diuji dalam struktur fold data yang berbeda.

Berdasarkan hasil pengujian, algoritma K-Nearest Neighbor memperlihatkan ketidakstabilan yang cukup tinggi. Hal ini tampak dari rentang deviasi yang besar pada nilai akurasi, recall, precision, dan Kappa pada cross validation baik 10-fold maupun 5-fold. Variasi performa yang jauh berbeda antar fold menunjukkan bahwa K-Nearest Neighbor sangat dipengaruhi oleh pemilihan data pelatihan. Dengan kata lain, ketika komposisi data sedikit berubah, kinerja K-Nearest Neighbor pun ikut menurun secara drastis. Kondisi ini terjadi karena model K-Nearest Neighbor bekerja berdasarkan kedekatan jarak antar data, sehingga dominasi kelas hujan ringan menyebabkan model sulit menemukan pola representatif yang merata pada seluruh kelas.

Di sisi lain, algoritma Decision Tree C4.5 menunjukkan tingkat stabilitas yang jauh lebih baik. Hal ini dibuktikan dengan nilai deviasi yang kecil pada setiap metrik evaluasi di tiap pengujian cross validation. Model tetap mampu menghasilkan performa yang konsisten meskipun data dibagi ke dalam fold yang berbeda. Struktur pohon keputusan yang berfokus pada pemilihan fitur pemisah terbaik di setiap node membantu model menyesuaikan diri dengan variasi komposisi data tanpa terjadi penurunan performa yang signifikan.

Dengan demikian, dapat dinyatakan bahwa dalam konteks prediksi curah hujan pada dataset yang digunakan, Decision Tree C4.5 merupakan model yang lebih stabil dan lebih dapat diandalkan dibandingkan K-Nearest Neighbor. Model ini tidak hanya memberikan performa yang lebih baik, tetapi juga mempertahankan kualitas prediksinya secara konsisten pada berbagai skenario pembagian data.

6. Perhitungan Manual

F1-Score merupakan metrik evaluasi yang menggabungkan precision dan recall dalam satu nilai terukur yang seimbang. Berikut hasil perhitungan manualnya.

a. Tanpa Cross Validation

Hasil F1-score berdasarkan nilai precision diperoleh sebesar 26.53% dan recall sebesar 31.29% pada algoritma K-Nearest Neighbor, diperoleh:

$$F1_{KNN} = 2 \times \frac{0.2653 \times 0.3129}{0.2653 + 0.3129} = 0.2871 = 28.71\%$$

Dari perhitungan diatas, F1-score dari algoritma K-Nearest Neighbor diperoleh sebesar 28.71%. Pada algoritma Decision Tree C4.5 dengan nilai precision 43.01% dan recall 35.60% didapatkan :

$$F1_{C4.5} = 2 \times \frac{0.4301 \times 0.3560}{0.4301 + 0.3560} = 0.3896 = 38.96\%$$

Dari perhitungan diatas, F1-score yang diperoleh dari algoritma Decision Tree C4.5 sebesar 38.96%.

b. Cross Validation

Karena RapidMiner tidak langsung menghasilkan nilai F1-Score, maka nilai ini diturunkan dari metrik Precision dan Recall pada masing-masing skenario cross validation. Berikut adalah hasil perhitungan manual F1-score pada Algoritma K-Nearest Neighbor:

Dengan Cross Validation 10-Fold:

Recall = 32.77%

Precision = 25.29%

$$F1_{KNN} = 2 \times \frac{0.2529 \times 0.3277}{0.2529 + 0.3277} = 0.2858 = 28.58\%$$

Dengan Cross Validation 5-Fold:

Recall = 29.77%

Precision = 22.11%

$$F1_{KNN} = 2 \times \frac{0.2211 \times 0.2977}{0.2211 + 0.2977} = 0.2546 = 25.46\%$$

Dari perhitungan diatas pada algoritma K-Nearest Neighbor diperoleh nilai F1-score sebesar 28.58% dengan skenario 10-fold dan nilai F1-score sebesar 25.46% dengan skenario 5-fold.

Sedangkan pada algoritma Decision Tree C4.5 diperoleh nilai sebagai berikut ini:

Dengan Cross Validation 10-Fold:

Recall = 35.50%
Precision = 43.92%

$$F1_{C4.5} = 2 \times \frac{0.4392 \times 0.3550}{0.4392 + 0.3550} = 0.3929 = 39,29\%$$

Dengan Cross Validation 5-Fold:
Recall = 35.13%
Precision = 49.57%

$$F1_{C4.5} = 2 \times \frac{0.4957 \times 0.3513}{0.4957 + 0.3513} = 0.4117 = 41,17\%$$

Dari hasil perhitungan manual diatas, didapatkan nilai F1-score sebesar 39.29% pada 10-fold dan 41.17% pada 5-fold. Berdasarkan hasil evaluasi dengan perhitungan manual menggunakan rumus akurasi, model K-Nearest Neighbor dan Decision Tree C4.5 menunjukkan performa yang berbeda dalam melakukan prediksi curah hujan. Pada metode Decision Tree C4.5, akurasi yang diperoleh sebesar 61,90% dari 22 data uji dengan 13 prediksi benar dan 9 prediksi salah. Sementara itu, metode K-Nearest Neighbor menghasilkan akurasi sebesar 57,14% dengan 12 prediksi benar dan 10 prediksi salah dari jumlah data yang sama. Perbedaan akurasi tersebut mengindikasikan bahwa Decision Tree C4.5 memiliki kemampuan klasifikasi yang lebih baik dalam memahami pola cuaca yang terdapat pada data, karena metode ini mampu memetakan hubungan antar atribut secara lebih eksplisit melalui struktur pohon keputusan. Ketika diuji menggunakan teknik validasi silang 10-fold dan 5-fold, pola performa yang sama juga terlihat. Pada validasi silang 10-fold, Decision Tree C4.5 mencatat akurasi sebesar 60,42% yang lebih tinggi dibandingkan K-Nearest Neighbor sebesar 56,03%, sedangkan pada validasi silang 5-fold Decision Tree C4.5 kembali unggul dengan akurasi 59,68% dibandingkan model K-Nearest Neighbor sebesar 54,88%. Konsistensi hasil tersebut menunjukkan bahwa Decision Tree C4.5 lebih stabil dan adaptif terhadap variasi data dalam proses pelatihan dan pengujian, sehingga lebih layak digunakan sebagai model prediksi dalam penelitian ini. Perbedaan performa ini juga memperlihatkan bahwa K-Nearest Neighbor lebih rentan terhadap distribusi data dan jarak antar fitur, terutama apabila proses normalisasi tidak optimal. Dengan demikian, berdasarkan keseluruhan pengujian, dapat disimpulkan bahwa metode Decision Tree C4.5 memberikan hasil performa prediksi yang lebih baik dan lebih andal dalam memprediksi curah hujan dengan cuaca sebagai label kategori di Kota Tangerang Selatan dibandingkan dengan K-Nearest Neighbor.

Secara keseluruhan, hasil ini membuktikan bahwa Decision Tree C4.5 merupakan metode yang lebih sesuai

digunakan dalam prediksi kategori cuaca berdasarkan curah hujan harian di wilayah Tangerang Selatan, karena memiliki kinerja lebih stabil dan kemampuan generalisasi yang lebih baik dibandingkan KNN.

IV. SIMPULAN

Berdasarkan hasil penelitian mengenai analisis perbandingan kinerja algoritma K-Nearest Neighbor dan Decision Tree C4.5 dalam prediksi curah hujan dengan melakukan klasifikasi kategori cuaca sebagai label target di Kota Tangerang Selatan, dapat ditarik kesimpulan sebagai jawaban dari rumusan masalah dan tujuan penelitian, yaitu:

1. Dari hasil penelitian analisis perbandingan teknik data mining metode K-Nearest Neighbor dan Decision Tree C4.5 menunjukkan bahwa Decision Tree C4.5 memiliki performa yang lebih baik dibandingkan K-Nearest Neighbor dalam mengklasifikasikan kategori cuaca berdasarkan data curah hujan harian di Kota Tangerang Selatan yang memiliki variabilitas tinggi. Tanpa penerapan cross-validation, K-Nearest Neighbor memperoleh akurasi sebesar 67,92%, weighted mean recall 31,29%, weighted mean precision 26,53%, dan kappa -0,011. Sementara itu, Decision Tree C4.5 tanpa cross-validation berhasil mencapai akurasi 75,94%, weighted mean recall 35,60%, weighted mean precision 43,01%, dan kappa 0,270. Analisis ini menegaskan bahwa algoritma Decision Tree C4.5 lebih efektif dalam mengenali pola data historis curah hujan di wilayah tersebut, khususnya ketika data menunjukkan variasi yang signifikan.
2. Berdasarkan penerapan skenario cross validation 10-fold dan 5-fold pada metode K-Nearest Neighbor dan Decision Tree C4.5, menunjukkan bahwa metode klasifikasi Decision Tree C4.5 memiliki stabilitas dan akurasi tertinggi dibandingkan K-Nearest Neighbor. Untuk Decision Tree C4.5, akurasi cross-validation 10-fold sebesar 78,70% ± 1,72% dan 5-fold sebesar 78,70% ± 1,47%, sedangkan K-Nearest Neighbor hanya mencapai 60,41% ± 20,20% pada 10-fold dan 56,03% ± 26,42% pada 5-fold. Nilai weighted mean recall dan precision Decision Tree C4.5 juga lebih tinggi dan lebih konsisten dibandingkan K-Nearest Neighbor. Temuan ini menegaskan bahwa Decision Tree C4.5 merupakan metode yang lebih andal dalam mengenali pola curah hujan harian, sehingga dapat dijadikan pilihan utama dalam prediksi cuaca berbasis klasifikasi kategori cuaca di Kota Tangerang Selatan. Namun demikian, penelitian ini masih memiliki keterbatasan, yaitu adanya ketidakseimbangan distribusi kelas pada data kategori cuaca serta penggunaan dua algoritma klasifikasi sebagai pembanding. Oleh karena itu, penelitian selanjutnya disarankan untuk menerapkan teknik penyeimbangan data dan mengeksplorasi

algoritma atau metode klasifikasi lain guna meningkatkan akurasi dan generalisasi model.

DAFTAR PUSTAKA

- Agung, A. S., Fauzi, A. A., Nur Risal, A. A., & Adiba, F. (2023). Implementasi Teknik Data Mining terhadap Klasifikasi Data Prediksi Curah Hujan BMKG Di Sulawesi Selatan. *Jurnal Tekno Insentif*, 17(1), 22–23. <https://doi.org/10.36787/jti.v17i1.955>
- Al Arif, A., Firdaus, M., Rahmadden, & Maruhawa, Y. (2022). Perbandingan Metode Data Mining untuk Prediksi Curah Hujan dengan Algoritma C4.5, Naïve Bayes, dan KNN. *SENTIMAS: Seminar Nasional Penelitian Dan Pengabdian Masyarakat*, 187–197.
- Ananda, N., & Gaol, G. . M. L. (2024). Estimasi Curah Hujan Bulanan Menggunakan Multi Layer Perceptron Di Kota Tangerang Selatan. *Buletin Meteorologi, Klimatologi, Dan Geofisika*, 5(4), 17–26.
- Badan Pusat Statistik. (2025). Statistik Daerah Kota Tangerang Selatan 2025. *BPS Kota Tangerang Selatan*, 1–22.
- Fatikah, I. H., & Khazizah, N. (2022). Analisis cuaca di Kota Jakarta bulan Januari tahun 2018 menggunakan Algoritma Decision Tree. *Jurnal Poros Teknik*, 14(1), 33–37. <https://www.kaggle.com/datasets/msf1203/pr>
- Fatmawati. (2016). Perbandingan Algoritma Klasifikasi Data Mining Model C4.5 Dan Naive Bayes Untuk Prediksi Penyakit Diabetes. *Jurnal Techno Nusa Mandiri*, XIII(1), 50.
- Limahelu, E. L., Thenu, Y. M., Umar, N., Putri, A. P. S., P.A.B, M. S., Antariksa, A. J., & Hakim, L. (2020). Buletin Meteorologi Analisis Cuaca Bulan Juni 2020. *Buletin Meteorologi Stasiun Meteorologi Umbu Mehang Kunda Sumba Timur*, 1–20.
- Miller, C., Portlock, T., Nyaga, D. M., & O’Sullivan, J. M. (2024). A review of model evaluation metrics for machine learning in genetics and genomics. *Frontiers in Bioinformatics*, 4, 1–13. <https://doi.org/10.3389/fbinf.2024.1457619>
- Pertiwi, D. A. S., & Paski, J. A. I. (2022). Analisis Dinamika Atmosfer Kejadian Hujan Ekstrem (Studi Kasus Banjir Di Tangerang Analysis Of The Atmospheric Dynamics Of Extreme Rainfall Event (Case Study Of Flood In Tangerang Selatan On 7 Th November 2021). *Buletin Meteorologi, Klimatologi, Dan Geofisika*, 2(2), 1–10.
- Supriatna, D., & Uliyatunisa. (2025). Komparasi Model LSTM dan CNN-LSTM untuk Peramalan Curah Hujan di Kota Tangerang Selatan. *Buletin Of Information Technology*, 6(3), 294–301. <https://doi.org/10.47065/bit.v5i2.2235>
- Susilawati Sugiana, N. S., & Musty, B. (2023). Analisis Data Sistem Informasi Monitoring Marketing; Tools Pengambilan Keputusan Strategic. *Jutisi: Jurnal Ilmiah Teknik Informatika Dan Sistem Informasi*, 12(2), 696. <https://doi.org/10.35889/jutisi.v12i2.1240>
- Zidane, M. Y., Nurina Sari, B., Maulana, I., Primaya, A., & Garno, G. (2024). Penerapan Data Mining Dalam Klasifikasi Data Transaksi Produk Koperasi Di Smk Pgri 2 Karawang. *JATI (Jurnal Mahasiswa Teknik Informatika)*, 9(1), 263–269. <https://doi.org/10.36040/jati.v9i1.12196>