p-ISSN 2085-7829
e-ISSN 2798-3455

Penerapan CRISP-DM Model dengan Algoritma Decision Tree, Naive Bayes, SVM, dan KNN untuk Klasifikasi Curah Hujan di Kabupaten Malang

Implementation of the CRISP-DM Model with Decision Tree, Naive Bayes, SVM, and KNN Algorithms for Rainfall Classification in Malang Regency

Arizki Dwi Cahyo[1], Arie Wahyu Wijayanto[2]
[1,2] Program Studi D-IV Statistika, Politeknik Statistika STIS, Indonesia

[a)] *Corresponding author: 212212517@stis.ac.id*

ABSTRACT

Rainfall refers to the amount of rain that falls over a specific period and is measured on a flat surface that does not absorb or channel the water. Rainfall is classified into six categories: cloudy, light rain, moderate rain, heavy rain, very heavy rain, and extreme rain. The objective of this study is to compare several classification methods in data mining, namely Decision Tree, Naïve Bayes, Support Vector Machine (SVM), and K-Nearest Neighbors (KNN) to determine the method with the highest accuracy for classifying rainfall in Malang Regency. To achieve this objective, the procedure used refers to CRISP-DM, which is a methodology for planning data mining projects. The data used in this study were obtained from NASA and consist of daily records collected from June 1, 2019, to June 1, 2025. The total number of observations in this study is 2,193. Data mining classification was employed to identify patterns and similarities in rainfall characteristics in the area. Based on the analysis results, the KNN method demonstrated the best performance with an accuracy of 87.3% and is therefore recommended as the most effective method for rainfall classification in Malang Regency.

Keywords: *crisp-dm, rainfall, decision tree, naïve bayes, svm, knn*

1. Pendahuluan

Kondisi iklim bumi yang berubah secara global telah memberi dampak pada beberapa hal. Salah satu hal yang terdampak akibat perubahan iklim bumi ialah curah hujan (Susilowati & Kusumastuti, 2023). Curah hujan adalah banyaknya air hujan yang jatuh dalam periode tertentu dan diukur melalui tempat yang datar, tidak meresap, dan tidak mengalir (BMKG, 2020). Ukuran 1 mm pada curah hujan menandakan dalam luasan satu persegi, terdapat air hujan setinggi satu milimeter. Artinya, semakin besar angka curah hujan (dalam mm), semakin banyak volume air yang turun ke permukaan dalam periode waktu tertentu.

Menurut BMKG, curah hujan memiliki 6 kategori, yaitu berawan, hujan ringan, hujan sedang, hujan lebat, hujan sangat lebat, dan hujan ekstrem. Kondisi berawan terjadi ketika suatu wilayah tidak terjadi hujan. Sedangkan, hujan ringan terjadi apabila suatu wilayah memiliki curah hujan hingga 20 mm/hari. Untuk kategori lain hujan sedang berkisar antara 20–50 mm/hari, hujan lebat berkisar antara 50-100 mm/hari, hujan sangat lebat berkisar antara 100-150 mm/hari, dan hujan ekstrem terjadi apabila curah hujan mencapai lebih dari 150 mm/hari. Tentunya, hujan yang turun tidak hanya terjadi pada musim hujan, melainkan juga di musim kemarau. Hal ini yang membuat suatu wilayah mengalami perubahan curah hujan.

Penyebab terjadinya perubahan curah hujan disebabkan oleh beberapa faktor. Secara umum, faktor-faktor tersebut dapat berasal dari pola angin, kondisi fisiografis, dan perubahan iklim (Damiri et al., 2024). Pola angin memainkan peran penting dalam menentukan distribusi curah hujan musiman di berbagai wilayah, terutama di daerah tropis, seperti Indonesia. Kondisi fisiografis juga memengaruhi intensitas curah hujan melalui proses orografis. Selain itu, perubahan iklim akibat meningkatnya konsentrasi gas rumah kaca juga menyebabkan ketidakstabilan atmosfer dan intensitas curah hujan. Fenomena yang terjadi akibat ketidakstabilan atmosfer ialah La Nina (BMKG, 2025). Kemunculan La Nina dapat menyebabkan variabilitas dan peningkatan curah hujan (Harahap et al., 2023).

Tentunya, curah hujan dengan intensitas tinggi/lebat lebih berbahaya dibanding curah hujan yang lebih rendah. Salah satu akibat dari curah hujan yang tinggi ialah terjadinya longsor dan erosi (Sutedjo & Kartasapoetra, 2002). Hal itu dapat terjadi karena terdapat pengikisan terhadap tanah yang dilaluinya. Selain itu, curah hujan yang tinggi juga dapat menyebabkan terjadinya banjir (Yatimah et al., 2024). Peristiwa banjir ditandai dengan meningkatnya volume air hingga meluap ke suatu daerah. Bukti-bukti tersebut menandakan bahwa curah hujan yang tinggi sangat berbahaya untuk suatu wilayah. Pasalnya, peristiwa-peristiwa yang datang akan membawa dampak buruk lain bagi wilayah yang bersangkutan. Berbagai dampak buruk lain yang dimaksud ialah rusaknya infrastruktur, kehilangan material, dan terancamnya kesehatan makhluk hidup (Nabila et al., 2024).

Salah satu wilayah di Indonesia yang memiliki curah hujan ekstrem ialah Kabupaten Malang. Sepanjang tahun 2024, Kabupaten Malang menjadi kabupaten/kota tertinggi kedua yang memiliki curah hujan ekstrem,

p-ISSN 2085-7829
e-ISSN 2798-3455

yaitu sebesar 385 mm/hari (DataIndonesia, 2025). Kejadian tersebut menyebabkan banjir setinggi 2 meter dan mengakibatkan gangguan aktivitas serta kerusakan infrastruktur.

Data curah hujan yang terus diperbarui dan muncul secara berkala memiliki relevansi tinggi dengan data mining. Hal itu disebabkan pola dan tren dari data tersebut dapat dianalisis untuk menghasilkan pengetahuan yang bermanfaat, seperti prediksi cuaca kedepannya. Dalam menganalisis data mining, perlu dilakukan prosedur yang mengacu pada CRISP-DM. CRISP-DM adalah suatu metodologi yang memberikan kerangka kerja umum dalam menyusun dan merancang proyek data mining (Brasjö & Lindovsky, 2019). Dalam CRISP-DM, suatu data mining tidak langsung dibentuk menjadi sebuah model, melainkan harus melalui serangkaian persiapan agar model yang terbentuk dapat menghasilkan akurasi yang baik. Persiapan tersebut seperti *business understanding*, *data understanding*, dan *data preparation.* Salah satu hal yang perlu diperhatikan dalam data preparation adalah *imbalance data*, yaitu kondisi ketika jumlah data antarkategori tidak seimbang. Solusi yang harus dilakukan jika hal itu terjadi adalah dengan melakukan *SMOTE*, yaitu alternatif pada *handling imbalance data* yang memungkinkan distribusi antarkelompok menjadi seimbang, sehingga hasil analisis data tidak bias pada kelas mayoritas.

Salah satu teknik data mining ialah klasifikasi. Klasifikasi bertujuan untuk melihat pola dan kesamaan karakteristik dari suatu *dataset* (Plotnikova et al., 2020). Untuk melakukan klasifikasi tersebut, Nantinya, model yang terbentuk digunakan untuk memprediksi nilai suatu kelas yang belum diketahui. Dengan begitu, klasifikasi pada data mining menjadi alat yang sangat penting dalam mengoptimalkan pemanfaatan data curah hujan guna mendukung pengambilan keputusan, terutama untuk mitigasi bencana.

Penelitian terdahulu tentang curah hujan menggunakan algoritma klasifikasi dilakukan oleh Ramadhan et al. (2024) yang mendapatkan bahwa algoritma KNN memiliki akurasi sebesar 86% di Indonesia. Aris Gunadi & Kusuma Dewi (2018) menggunakan algoritma Naïve Bayes terkait curah hujan di Provinsi Bali dan mendapatkan akurasi tertinggi sebesar 86,4%. Sedangkan, Hasanah et al. (2021) menemukan bahwa algoritma *Decision Tree* mampu menghasilkan akurasi sebesar 89,4% dalam mengklasifikasikan curah hujan di Indonesia. Penelitian lain dilakukan oleh Indra Pratama et al. (2022) yang menemukan bahwa akurasi tertinggi untuk klasifikasi curah hujan di Indonesia dengan algoritma SVM ialah 79%. Terakhir ialah penelitian dari Aris Gunadi et al. (2022) yang meneliti tentang klasifikasi curah hujan di Provinsi Bali dengan algoritma *learning vector quantization*. Hasilnya, algoritma tersebut dapat menghasilkan akurasi sebesar 49,6%.

Berdasarkan uraian sebelumnya, dapat diketahui bahwa penelitian yang menganalisis curah hujan dengan membandingkan berbagai algoritma data mining masih relatif terbatas. Selain itu, kajian curah hujan di Kabupaten Malang yang memanfaatkan pendekatan data mining juga masih jarang dilakukan. Oleh karena itu, penelitian ini bertujuan untuk membandingkan beberapa algoritma klasifikasi dalam data mining, yaitu Decision Tree, Naïve Bayes, Support Vector Machine (SVM), dan K-Nearest Neighbors (KNN), dengan menilai tingkat akurasi masing-masing algoritma guna menentukan model terbaik dalam mengklasifikasikan curah hujan di Kabupaten Malang.

2. Metodologi Penelitian

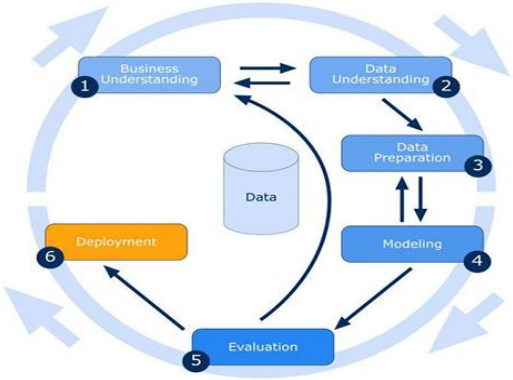
Pada penelitian ini, sumber data yang digunakan berasal dari NASA melalui laman resmi https://power.larc.nasa.gov di Kabupaten Malang. Penelitian ini menggunakan 5 atribut dan 1 label, dengan rinciannya disajikan pada Tabel 1.

Tabel 1. Variabel Penelitian

Notasi	Variabel Penelitian	Jenis Variabel	Satuan
CURAH_H	Curah Hujan	Label	mm
KEC_ANGIN	Kecepatan Angin	Atribut	m/s
KELEM_UDARA	Kelembapan Udara	Atribut	g/kg
KEC_ARAH_ANGIN	Kecepatan Arah Angin	Atribut	Derajat
SUHU	Suhu	Atribut	*Celcius*
TGL	Tanggal	Atribut	-

Data yang digunakan berupa data harian dari tanggal 1 Juni 2019 sampai 1 Juni 2025 dengan total 2.193 records. Selain itu, metodologi data mining yang digunakan ialah klasifikasi dan berpedoman pada CRISP-DM. CRISP-DM ialah metodologi yang digunakan untuk membentuk dan merencanakan proyek *data mining* (Brasjö & Lindovsky, 2019). CRISP-DM memiliki enam tahapan, yaitu *business understanding*, *data understanding*, *data preparation*, *modelling*, *evaluation*, dan *deployment*. Tahapan CRISP-DM memiliki keterkaitan satu sama lain. Hal itu digambarkan dengan arah panah yang tertera pada Gambar 1.

p-ISSN 2085-7829
e-ISSN 2798-3455



Gambar 1. Tahapan CRISP-DM
Sumber: Hasanah, et al., 2021

2.1. *Business Understanding*

Pada tahap ini, dilakukan sejumlah langkah penting seperti memahami kebutuhan dan tujuan dari perspektif bisnis, kemudian menerjemahkan pemahaman tersebut ke dalam bentuk perumusan masalah yang sesuai dengan konteks data mining. Setelah itu, disusun rencana dan strategi yang tepat untuk mencapai tujuan yang telah ditetapkan dalam proses data mining.

2.2. *Data Understanding*

Tahap ini terdiri atas mengumpulkan, melakukan visualisasi, dan mengevaluasi kualitas data. Tentunya, data yang divisualisasikan harus menggambarkan kondisi data. Selain itu, sumber data yang dikumpulkan juga harus jelas agar menghasilkan hasil yang berkualitas.

2.3. *Data Preparation*

Tahapan ini menjadi tahap terakhir sebelum dilakukan permodelan. Berbagai cara dilakukan agar permodelan dapat menghasilkan kualitas yang bagus, diantaranya *cleaning data*, pemeriksaan konstruksi data, dan transformasi. *Cleaning data* dilakukan dengan penanganan *missing value*, *outlier*, dan lain-lain. Konstruksi data dilakukan melalui *frature engineering*. Dan tranformasi dilakukan melalui normalisasi dan standarisasi. Selain itu, tahapan *data preparation* harus memastikan bahwa jumlah amatan setiap kategori dalam kondisi seimbang. Jika antarkategrori tidak seimbang, dilakukan *handling imbalance data* melalui *SMOTE*, yaitu alternatif yang memungkinkan distribusi antarkelompok menjadi seimbang.

2.4. *Modelling*

Tahap ini melibatkan *machine learning*, *software*, dan algoritma dalam *data mining*. Pada penelitian ini, algoritma yang digunakan adalah *Decision Tree*, Naïve Bayes, SVM, dan KNN.

2.4.1 *Decision Tree*

Decision Tree merupakan algoritma yang mengubah bentuk data berupa tabel menjadi sebuah pohon (Tree). Struktur *Decision Tree* dibuat berdasarkan proses rekursi yang memanfaatkan nilai *information gain* yang paling tinggi. Atribut yang memiliki *information gain* tertinggi dipilih sebagai *root* atau titik yang memulai partisi data. Proses ini dilakukan berulang dengan membagi *subset* data pada setiap node internal hingga tidak ada atribut yang tersisa untuk klasifikasi. Pada akhirnya, algoritma ini akan menghasilkan *rule* tertentu dan disederhanakan (Basuki & Syarif, 2003).

2.4.2 Naïve Bayes

Naïve Bayes menjadi algoritma dalam data mining yang terkenal (Wu et al., 2008). Naïve Bayes merupakan teknik klasifikasi yang dapat memprediksi peluang untuk menjadi anggota kelas. Algoritma ini menggunalan probabilitas sederhana sesuai dengan teorema Bayes dengan perkiraan independensi yang kuat (Pebdika et al., 2023). Naïve Bayes memperkirakan bahwa nilai suatu atribut tidak dipengaruhi oleh nilai lainnya.

2.4.3 SVM

SVM merupakan algoritma yang memanfaatkan pemetaan non linier untuk mengganti data *training* menjadi dimensi tinggi (Indra Pratama et al., 2022). SVM digunakan untuk melakukan pengoptimalan lebar celah antar kelas (Zhao & Cen, 2014). Setiap objek diplot sebagai titik dalam ruang berdimensi n. Setiap fungsi menjadi biaya koordinat yang dipilih dan dilakukan klasikasi dengan mendesain *hyperplane*. *Hyperplane* berperan membagi bidang menjadi beberapa kelas agar masing-masing kelas berbeda.

2.4.4 KNN

KNN merupakan algoritma dalam *data mining* yang melakukan klasifikasi berdasarkan jarak suatu objek dengan objek lain. KNN menggunakan sampel baru untuk diklasifikasikan berdasarkan kemiripan

p-ISSN 2085-7829
e-ISSN 2798-3455

dengan sampel pada data latih (Kataria & Singh, 2013). Algoritma ini digunakan untuk *dataset* yang memiliki beberapa kelas. Dalam KNN, terdapat teknik *lazy learning* karena sebagian besar perhitungan dilakukan pada data latih untuk melihat objek terdekat pada kelompok k data uji (Ullah et al., 2019).

2.5. Evaluation

Tahap ini digunakan untuk melihat performa dari algoritma yang dipakai. Terdapat parameter berupa *confusion matrix* yang digunakan untuk evaluasi komparasi algoritma. Evaluasi yang digunakan adalah *accuracy* yang diformulasikan sebagai berikut:

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \quad (1)$$

dengan
TP = *True Positive*
FP = *False Positive*
FN = *False Negative*
TN = *True Negative*.

2.6. *Deployment*

Tahapan ini dilakukan pembuatan laporan berupa jurnal menggunakan model yang dihasilkan. Disajikan seluruh proses, mulai dari perumusan masalah, pengolahan dan analisis data, hingga interpretasi hasil. Laporan disusun secara sistematis agar dapat dipahami oleh pembaca, mencakup penjelasan algoritma yang digunakan, alasan pemilihan algoritma, serta evaluasi performa model.

3. Hasil dan Pembahasan

3.1 *Business Understanding*

Pada penelitian ini dilakukan penerapan data mining dengan data curah hujan. Penerapan tersebut dilakukan untuk melihat pola intensitas curah hujan di Kabupaten Malang. *Software* yang digunakan pada penelitian ini adalah RStudio, dilengkapi dengan beberapa *packages* untuk mendukung pengoperasian penelitian.

3.2 *Data Understanding*

Data yang digunakan terdiri atas lima atribut dan satu label. *Dataset* yang digunakan pada penelitian ini dapat dilihat pada tabel 2.

Tabel 2. *Dataset* Penelitian

Tgl	SUHU	KEC_ANGIN	KELEM_ UDARA	KEC_ARAH_ ANGIN	CURAH_H
1 Juni 2019	22,64	2,77	15,52	137,2	0,00
2 Juni 2019	22,84	3,34	15,64	140,1	0,00
3 Juni 2019	23,34	2,39	16,39	137,5	0,00
4 Juni 2019	23,74	1,81	16,48	125,1	0,00
5 Juni 2019	23,95	1,83	16,72	130,1	0,00
28 Mei 2025	25,03	0,65	18,25	157,1	4,38
29 Mei 2025	25,34	0,80	18,39	156,5	3,65
30 Mei 2025	25,53	0,69	18,22	146,3	6,45
31 Mei 2025	25,34	0,68	18,92	115,3	11,33
1 Juni 2025	25,30	0,64	18,59	115,7	5,90

3.3 *Data Preparation*

Pada tahap ini dilakukan klasifikasi curah hujan menjadi 6 kategori (berawan, ringan, sedang, lebat, sangat lebat, dan ekstrem). Hasilnya, kategori curah hujan lebat, sangat lebat, dan ekstrem tidak mendominasi pada *dataset* ini. Oleh karena itu, tiga kategori tersebut dilakukan penggabungan menjadi kategori Lebat. Penggabungan tersebut didasarkan pada kemiripan karakteristik yang ada pada kategori tersebut, yaitu menggambarkan hujan yang turun memiliki curah yang besar dan berisiko menimbulkan berbagai permasalahan, seperti banjir. Jadi, *dataset* hanya memiliki 4 kategori, yaitu curah hujan berawan, ringan, sedang, dan lebat. Setelah itu, dilakukan *cleaning data* dengan menghilangkan atribut yang tidak diperlukan, yaitu tanggal dan curah hujan yang bersifat numerik. Struktur data yang digunakan berupa numerik untuk atribut dan faktor untuk label.

Dalam *dataset* tersebut, terdapat imbalance data atau perbedaan jumlah data yang terlalu jauh

p-ISSN 2085-7829
e-ISSN 2798-3455

antarkelompok label. *Handling imbalance* data dilakukan menggunakan SMOTE. Cara ini dapat meningkatkan performa klasifikasi dengan menyeimbangkan distribusi antarkelompok sehingga model tidak bias terhadap kelas mayoritas. Hasilnya, *dataset* menjadi seimbang/balance dan berjumlah 7.492 *records*. Tahap selanjutnya ialah melakukan transformasi data berupa normalisasi. Tahap ini perlu dilakukan agar rentang nilai setiap atribut tidak berbeda jauh. Normalisasi yang digunakan ialah min-max scaler. Hasil normalisasi menggunakan min-max scaler dapat dilihat pada tabel 3.

Tabel 3. Hasil Normalisasi

SUHU	KEC_ANGIN	KELEM_UDARA	KEC_ARAH_ANGIN
0,2356	0,6072	0,4885	0,3813
0,2585	0,5036	0,5022	0,3894
0,3157	0,5156	0,5878	0,3821
0,3615	0,3759	0,5981	0,3474
0,3855	0,3807	0,6255	0,3614

3.4 *Modelling*

Setelah melakukan data preparation, tahap selanjutnya ialah melakukan permodelan. Permodelan harus dilakukan dengan data yang *clean* agar menghasilkan kualitas yang baik. Permodelan dilakukan dengan menggunakan algoritma *Decision Tree*, Naïve Bayes, SVM, dan KNN. Selain itu, permodelan dilakukan dengan membagi data *training* dan *testing* masing-masing sebesar 80% dan 20%.

3.5 *Evaluation*

Hasil yang didapatkan pada proses *modelling* akan dievaluasi melalui *confusion matrix*. Selain itu, semua alogitma yang terbentuk akan dibandingkan dengan melihat nilai *accuracy*.

3.5.1 *Decision Tree*

Tabel 4. *Confusion Matrix* dengan *Decision Tree*

Kelas Prediksi	Kelas Aktual			
	Berawan	Ringan	Sedang	Lebat
Berawan	362	81	0	0
Ringan	10	82	16	0
Sedang	2	168	314	178
Lebat	0	43	44	196

Berdasarkan Tabel 4, model *Decision Tree* berhasil memprediksi dengan benar 362 curah hujan berawan, 82 curah hujan ringan, 314 curah hujan sedang, dan 196 curah hujan lebat. Namun, masih terjadi kesalahan klasifikasi, seperti curah hujan ringan yang diprediksi sebagai berawan sebanyak 81 kasus, curah hujan lebat sebagai sedang sebanyak 178 kasus, serta beberapa kesalahan lain antar kelas. Kesalahan terbesar terjadi pada prediksi curah hujan lebat yang diklasifikasikan sebagai sedang, menunjukkan bahwa model masih kesulitan membedakan antara kategori sedang dan lebat.

3.5.2 Naïve Bayes

Tabel 5. *Confusion Matrix* dengan *Naïve Bayes*

Kelas Prediksi	Kelas Aktual			
	Berawan	Ringan	Sedang	Lebat
Berawan	363	86	0	0
Ringan	11	112	32	5
Sedang	0	96	221	143
Lebat	0	80	121	226

Berdasarkan Tabel 5, model Naïve Bayes berhasil memprediksi dengan benar 363 curah hujan berawan, 112 curah hujan ringan, 221 curah hujan sedang, dan 226 curah hujan lebat. Meski begitu, masih terjadi kesalahan klasifikasi seperti 86 curah hujan ringan diprediksi sebagai berawan, 121 curah hujan sedang diprediksi sebagai lebat, serta 143 curah hujan lebat diprediksi sebagai sedang. Sama seperti algoritma Decision Tree, kesalahan terbesar terjadi antara kelas sedang dan lebat, menunjukkan bahwa model mengalami kesulitan dalam membedakan kedua kategori tersebut.

p-ISSN 2085-7829
e-ISSN 2798-3455

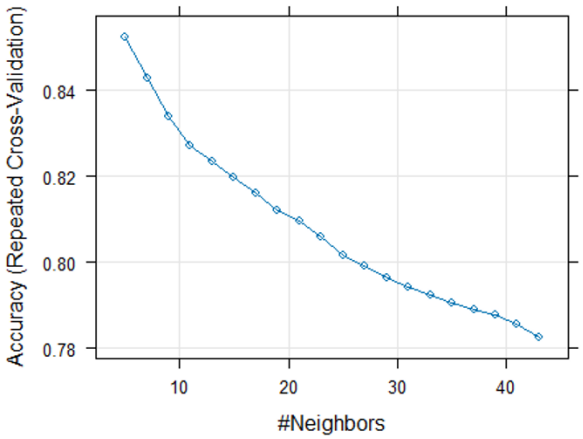
3.5.3 SVM

Tabel 6. *Confusion Matrix* dengan SVM

Kelas Prediksi	Kelas Aktual			
	Berawan	**Ringan**	**Sedang**	**Lebat**
Berawan	365	54	0	0
Ringan	9	196	41	0
Sedang	0	86	265	25
Lebat	0	38	68	349

Berdasarkan Tabel 6, model SVM mampu memprediksi dengan benar 365 curah hujan berawan, 196 curah hujan ringan, 265 curah hujan sedang, dan 349 curah hujan lebat. Meskipun akurasi prediksi tergolong tinggi, masih terdapat beberapa kesalahan klasifikasi, seperti 86 curah hujan ringan diprediksi sebagai sedang, 68 curah hujan sedang sebagai lebat, dan 41 curah hujan sedang sebagai ringan. Kesalahan paling menonjol terjadi antara kelas ringan dan sedang.

3.5.4 KNN



Gambar 2. Akurasi dari berbagai nilai *k*

Pada penelitian ini, didapatkan nilai *k* atau tetangga terdekat sebanyak 5. Artinya, pemilihan kategori pada suatu amatan didasarkan pada lima tetangga terdekat dari amatan tersebut. Hal itu dapat dilihat pada gambar 2 yang memiliki akurasi tertinggi.

Tabel 7. *Confusion Matrix* dengan KNN

Kelas Prediksi	Kelas Aktual			
	Berawan	**Ringan**	**Sedang**	**Lebat**
Berawan	374	58	0	0
Ringan	0	216	21	0
Sedang	0	83	343	1
Lebat	0	17	10	373

Berdasarkan Tabel 7, model KNN dengan nilai *k* sebesar 5 mampu memprediksi dengan benar 374 curah hujan berawan, 216 curah hujan ringan, 343 curah hujan sedang, dan 373 curah hujan lebat. Model menunjukkan performa sangat baik dengan jumlah kesalahan klasifikasi yang sangat kecil, seperti 58 curah hujan ringan diprediksi sebagai berawan, 21 curah hujan sedang sebagai ringan, dan 10 curah hujan lebat sebagai sedang.

Tabel 8. Perbandingan Evaluasi Model

Alogitma	***Accuracy***
Decision Tree	63,77%
Naïve Bayes	61,63%
SVM	78,54%
KNN	87,3%

Berdasarkan Tabel 8, terlihat bahwa masing-masing algoritma klasifikasi memiliki tingkat akurasi yang

p-ISSN 2085-7829
e-ISSN 2798-3455

berbeda-beda. Algoritms KNN menunjukkan performa terbaik dengan akurasi sebesar 87,3%, disusul oleh SVM yang mencapai akurasi sebesar 78,54%. Sementara itu, *Decision Tree* dan Naïve Bayes menunjukkan hasil akurasi yang lebih rendah, yaitu masing-masing sebesar 63,77% dan 61,63%. Dari hasil ini dapat disimpulkan bahwa KNN merupakan algoritma yang paling optimal dalam melakukan klasifikasi pada data yang digunakan, sehingga dapat dipertimbangkan untuk diterapkan pada tahap implementasi guna memperoleh hasil yang lebih akurat dan sesuai dalam melihat pola curah hujan di Kabupaten Malang.

Akurasi yang dihasilkan oleh algoritma KNN tersebut juga konsisten dengan temuan Ramadhan et al. (2024) yang melaporkan bahwa metode KNN mampu mencapai tingkat akurasi sebesar 86,09% dalam mengklasifikasikan curah hujan. Kemiripan pada penelitian sebelumnya menunjukkan bahwa KNN memiliki kemampuan yang baik dalam mengenali pola dan kedekatan antar data, terutama pada data curah hujan yang cenderung memiliki karakteristik nonlinier dan variabilitas tinggi. Hal ini disebabkan oleh prinsip kerja KNN yang mengklasifikasikan data berdasarkan kedekatan jarak terhadap data latih, sehingga mampu menangkap kemiripan pola secara lebih fleksibel dibandingkan metode lain. Dengan demikian, konsistensi performa KNN pada berbagai penelitian memperkuat validitas bahwa algoritma ini merupakan salah satu metode yang efektif dan andal untuk digunakan dalam klasifikasi curah hujan, khususnya pada wilayah yang memiliki karakteristik mirip dengan Kabupaten Malang.

3.6 *Deployment*

Tahapan ini dilakukan pembuatan laporan berupa jurnal menggunakan model yang dihasilkan. Setelah tahap evaluasi selesai, langkah berikutnya adalah mengimplementasikan seluruh model yang telah dikembangkan. Pada tahap ini juga dimungkinkan untuk melakukan penyesuaian terhadap model agar hasilnya lebih selaras dengan tujuan awal yang telah ditetapkan dalam proses CRISP-DM.

4. Kesimpulan

Berdasarkan hasil dan pembahasan yang telah dilakukan, dapat disimpulkan bahwa algoritma *K-Nearest Neighbors* (KNN) menghasilkan tingkat akurasi tertinggi dibandingkan dengan algoritma lainnya yang digunakan dalam penelitian ini, yaitu sebesar 87,3%. Selain itu, model klasifikasi curah hujan di Kabupaten Malang yang dibangun menggunakan algoritma KNN menunjukkan performa yang sangat baik. Oleh karena itu, KNN direkomendasikan sebagai algoritma yang paling tepat untuk mengklasifikasikan curah hujan di Kabupaten Malang.

Sebagai saran bagi Pemerintah Daerah Kabupaten Malang, dalam mengidentifikasi pola curah hujan diperlukan kajian lanjutan serta kerja sama dengan para ahli di bidang data mining. Hal ini penting agar setiap kebijakan yang diambil dapat didasarkan pada analisis data yang kuat dan menghasilkan keputusan yang lebih tepat sasaran, sehingga Kabupaten Malang menjadi lebih siap dalam menghadapi risiko yang timbul akibat tingginya curah hujan.

5. Daftar Pustaka

Aris Gunadi, I. G., & Kusuma Dewi, A. A. (2018). Klasifikasi Curah Hujan di Provinsi Bali Berdasarkan Metode Naïve Bayesian. *Wahana Matematika Dan Sains: Jurnal Matematika, Sains, Dan Pembelajarannya*, *12*(1), 14–25. https://doi.org/10.23887/wms.v12i1.13843

Aris Gunadi, I. G., Oka Gunawan, I. M. A., Hary Candana, P. E. W., Widyantari Arnawa, I. A., & Edo Kharisma Putra, K. A. (2022). Klasifikasi Curah Hujan Harian Menggunakan Learning Vector Quantization (Studi Kasus: Stasiun Pengamatan Ngurah Rai). *Jurnal Ilmu Komputer Indonesia (JIK)*, *7*(2), 1–7. https://doi.org/10.23887/jik.v7i2.4060

Basuki, A., & Syarif, I. (2003). *Decision Tree*. Politeknik Electronika Negeri.

BMKG. (2020). *Buletin Meteorologi Edisi Juli 2020*. Badan Meteorologi, Klimatologi, dan Geofisika.

BMKG. (2025). *Catatan Iklim dan Kualitas Udara Indonesia 2024*. Badan Meteorologi, Klimatologi, dan Geofisika.

Brasjö, C., & Lindovsky, M. (2019). *Machine Learning Project Management: A Study of Project Requirements and Processes in Early Adoption*.

Damiri, B. A., Ramadhan, W. N., & Supriyanti, K. R. (2024). Pengelompokan Faktor-Faktor yang Mempengaruhi Curah Hujan di Provinsi Sumatera Utara Menggunakan Metode Fuzzy C-Means. *Jurnal Statistika Dan Komputasi*, *3*(1), 1–10. https://doi.org/10.32665/statkom.v3i1.2623

DataIndonesia. (2025). Data wilayah dengan curah hujan ekstrem harian tertinggi di Indonesia pada 2024. Bersumber dari BMKG. Diakses pada 29 Maret 2026, dari https://dataindonesia.id/varia/detail/data-wilayah-dengan-curah-hujan-ekstrem-harian-tertinggi-di-indonesia-pada-2024

Harahap, W. N., Yuniasih, B., & Gunawan, S. (2023). Dampak La Nina 2021-2022 terhadap Peningkatan Curah Hujan. *AGROISTA : Jurnal Agroteknologi*, *7*(1), 26–32. https://doi.org/10.55180/agi.v7i1.364

p-ISSN 2085-7829
e-ISSN 2798-3455

Hasanah, M. A., Soim, S., & Handayani, A. S. (2021). Implementasi CRISP-DM Model Menggunakan Metode Decision Tree dengan Algoritma CART untuk Prediksi Curah Hujan Berpotensi Banjir. *Journal of Applied Informatics and Computing (JAIC)*, *5*(2), 103–108. https://doi.org/10.30871/jaic.v5i2.3200

Indra Pratama, A. R., Latipah, S. A., & Sari, B. N. (2022). Optimasi Klasifikasi Curah Hujan Menggunakan Support Vector Machine (SVM) dan Recursive Feature Elimination (RFE). *Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika*, *7*(2), 314–324. https://doi.org/10.29100/jipi.v7i2.2675

Kataria, A., & Singh, M. D. (2013). A Review of Data Classification Using K-Nearest Neighbour Algorithm. *International Journal of Emerging Technology and Advanced Engineering*, *3*(6), 354–360.

Nabila, M. P., Tanjung, M. R., Abelia, & Usiono. (2024). Waspada! Curah Hujan yang Cukup Tinggi: Sumatera Utara. *Jurnal Media Akademik (JMA)*, *2*(12), 1–11.

Pebdika, A., Herdiana, R., & Solihudin, D. (2023). Klasifikasi Menggunakan Metode Naïve Bayes untuk Menentukan Calon Penerima PIP. *Jurnal Mahasiswa Teknik Informatika*, *7*(1), 452–458.

Plotnikova, V., Dumas, M., & Milani, F. (2020). Adaptations of data mining methodologies: A systematic literature review. *PeerJ Computer Science*, *6*, 1–43. https://doi.org/10.7717/PEERJ-CS.267

Ramadhan, M. A., Anggraeny, F. T., & Putra, C. A. (2024). Klasifikasi Curah Hujan Harian Menggunakan Metode K-Nearest Neighbor. *Jurnal Mahasiswa Teknik Informatika*, *8*(3), 3863–3869. https://doi.org/10.36040/jati.v8i3.9817

Susilowati, & Kusumastuti, D. I. (2023). Analisa Karakteristik Curah Hujan dan Kurva Intensitas Durasi Frekuensi (IDF) di Provinsi Lampung. *Jurnal Rekayasa Teknik Sipil Universitas Lampung*, *14*(1), 47–56.

Sutedjo, M. M., & Kartasapoetra, A. G. (2002). *Pengantar Ilmu Tanah*. Bina Aksara.

Ullah, R., Khan, A. H., & Emaduddin, S. M. (2019). ck-NN: A Clustered k-Nearest Neighbours Approach for Large-Scale Classification. *Advances in Distributed Computing and Artificial Intelligence Journal*, *8*(3), 67–77. https://doi.org/10.14201/ADCAIJ2019836777

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z. H., Steinbach, M., Hand, D. J., & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37. https://doi.org/10.1007/s10115-007-0114-2

Yatimah, N., Kumalawati, R., & Muhtar, G. A. (2024). Analisis Kerentanan Bencana Banjir Berdasarkan Data Curah Hujan Kota Samarinda. *Jurnal Multidisiplin Raflesia*, *3*(1), 28–32.

Zhao, Y., & Cen, Y. (2014). *Data Mining Applications with R*. USA : Academic Press.