
Klasifikasi Kejadian Banjir di DKI Jakarta Berdasarkan Data Historis

BPBD Menggunakan Algoritma C4.5

Berbasis Decision tree

Irena Dewi¹, M. Farid Alfikri², Rame Santoso³, Indah Purwandani⁴,

¹19200679@bsi.ac.id, ²19210722@bsi.ac.id, ³rame.rms@bsi.ac.id, ⁴indah@bsi.ac.id

¹Program Studi Sistem Informasi, ²Fakultas Teknik dan Informatika, ³Universitas Bina Sarana Informatika

Abstrak

Banjir merupakan permasalahan utama di DKI Jakarta yang berdampak besar pada aspek ekonomi, sosial, dan infrastruktur. Penelitian ini bertujuan mengklasifikasikan tingkat kejadian banjir berdasarkan data historis BPBD DKI Jakarta periode 2020–2023 menggunakan algoritma C4.5 berbasis *Decision Tree*. Analisis dilakukan melalui dua pendekatan, yaitu perhitungan manual (*entropy* dan *information gain*) serta implementasi menggunakan RapidMiner. Tahapan penelitian mencakup pengumpulan dan pra-pemrosesan data, penerapan algoritma, serta evaluasi model menggunakan *confusion matrix*. Hasil menunjukkan bahwa algoritma C4.5 berhasil membentuk pohon keputusan yang konsisten antara perhitungan manual dan implementasi, serta mampu mengklasifikasikan wilayah ke dalam dua kategori, yaitu aman dan rawan, dengan akurasi yang baik. Model ini diharapkan menjadi dasar pengembangan sistem pendukung keputusan dalam mitigasi dan penanganan banjir di DKI Jakarta.

Kata kunci: Algoritma C4.5, Decision Tree, Klasifikasi Banjir, Rapidminer, DKI Jakarta

Abstract

Flooding is a major problem in DKI Jakarta, significantly affecting the economic, social, and infrastructure sectors. This study aims to classify the level of flood occurrences based on historical data from BPBD DKI Jakarta for the 2020–2023 period using the C4.5 algorithm based on the Decision Tree method. The analysis was carried out through two approaches: manual calculation (entropy and information gain) and implementation using RapidMiner. The research stages included data collection, preprocessing, algorithm application, and model evaluation using a confusion matrix. The results show that the C4.5 algorithm successfully generated a decision tree consistent between manual calculation and implementation, and it was able to classify areas into two categories, namely safe and vulnerable, with good accuracy. This model is expected to serve as a foundation for developing decision support systems for flood mitigation and management in DKI Jakarta.

Keywords: C4.5 Algorithm, Decision Tree, Flood Classification, RapidMiner, DKI Jakarta

1. Pendahuluan

1.1 Latar Belakang

Banjir merupakan salah satu bencana yang paling sering terjadi di Indonesia, khususnya di wilayah perkotaan padat penduduk seperti DKI Jakarta. Kondisi geografis Jakarta yang berupa dataran rendah dan dilintasi oleh 13 sungai besar menyebabkan wilayah ini rentan terhadap genangan, terutama saat curah hujan tinggi. Minimnya area resapan air serta pesatnya pembangunan yang tidak diimbangi dengan sistem drainase memadai memperparah risiko banjir di kawasan ini [1]. Menurut Bappeda DKI Jakarta, kemiringan lahan yang hanya berkisar 0–3% serta alih fungsi lahan yang masif menyebabkan peningkatan limpasan air permukaan [2].

Bencana banjir menimbulkan dampak besar terhadap sektor ekonomi, sosial, dan infrastruktur kota [3]. Oleh karena itu, diperlukan strategi mitigasi yang berbasis data untuk mengenali pola kejadian banjir secara lebih akurat. Analisis data historis banjir dapat memberikan wawasan penting dalam upaya pengambilan keputusan yang cepat dan tepat. Dalam konteks ini, teknik data mining menjadi pendekatan yang efektif untuk mengolah data besar dan mengidentifikasi pola tersembunyi pada kejadian banjir.

1.2 Penelitian Terdahulu

Berbagai penelitian sebelumnya telah menerapkan algoritma C4.5 dalam proses klasifikasi dan prediksi kejadian banjir di beberapa wilayah di Indonesia. Risnawati et al. [4] mengklasifikasikan curah hujan yang berpotensi menyebabkan banjir di wilayah Jawa Tengah dan Jawa Timur menggunakan data dari BMKG dan NASA, dengan tingkat akurasi sebesar 83,33%.

Selanjutnya, Nasrullah et al. [5] membandingkan kinerja algoritma C4.5 dan *K-Nearest Neighbor (KNN)* untuk klasifikasi curah hujan berdasarkan data iklim di Indonesia. Hasil penelitian menunjukkan bahwa meskipun algoritma KNN memiliki akurasi sedikit lebih tinggi (83,37%), algoritma C4.5 dinilai lebih mudah diinterpretasikan karena menghasilkan pohon keputusan yang transparan.

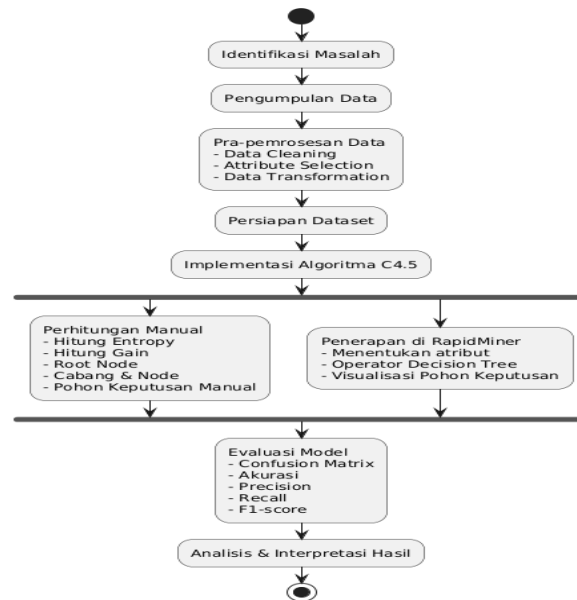
Sementara itu, Abdullah [6] menerapkan metode *Decision Tree* C4.5 untuk memprediksi kejadian banjir di Kota Pontianak dan memperoleh tingkat akurasi tertinggi sebesar 98%, yang menunjukkan potensi besar algoritma ini dalam pemodelan prediktif bencana berbasis data historis.

1.3 Novelty dan State of The Art

Berdasarkan kajian terhadap penelitian-penelitian terdahulu, terlihat bahwa algoritma C4.5 telah banyak digunakan untuk memprediksi curah hujan dan potensi banjir. Namun, sebagian besar studi tersebut menggunakan data cuaca atau data simulatif dari wilayah selain DKI Jakarta, seperti Jawa Timur dan Pontianak, sehingga belum mencerminkan kondisi geografis dan hidrologis Jakarta secara spesifik. Penelitian ini hadir untuk mengisi kesenjangan tersebut dengan menggunakan data aktual dari Badan Penanggulangan Bencana Daerah (BPBD) DKI Jakarta yang mencatat kejadian banjir nyata pada periode 2020–2023.

Pendekatan ini menawarkan nilai tambah berupa pemanfaatan data empiris dan kontekstual yang menggambarkan kondisi riil di lapangan. Selain itu, implementasi algoritma C4.5 dilakukan melalui perangkat lunak RapidMiner yang memungkinkan visualisasi hasil dalam bentuk pohon keputusan serta evaluasi model menggunakan *confusion matrix*. Dengan demikian, penelitian ini tidak hanya memberikan kontribusi teoritis dalam pengembangan metode klasifikasi banjir berbasis *decision tree*, tetapi juga manfaat praktis dalam mendukung sistem pendukung keputusan untuk mitigasi dan penanganan bencana banjir di DKI Jakarta.

2. Metode



Gambar 1. Tahapan Penelitian

2.1 Tahapan Penelitian

Penelitian ini menggunakan pendekatan klasifikasi berbasis *Decision Tree* dengan algoritma C4.5 untuk mengidentifikasi tingkat kerawanan banjir di wilayah DKI Jakarta. Secara umum, penelitian dilaksanakan melalui beberapa tahapan utama yang saling berkaitan, dimulai dari pengumpulan data hingga tahap evaluasi model. Data yang digunakan merupakan data historis kejadian banjir yang diperoleh dari BPBD DKI Jakarta dalam rentang waktu tahun 2020 hingga 2023.

Tahap awal penelitian adalah pengumpulan data, yang mencakup pencarian, penggabungan, dan verifikasi data agar diperoleh informasi yang valid dan representatif. Selanjutnya dilakukan pra-pemrosesan data yang meliputi pembersihan data dari duplikasi dan nilai kosong, seleksi atribut yang relevan, transformasi format data ke bentuk yang dapat diproses oleh algoritma, serta pemberian label kelas berdasarkan tingkat kejadian banjir.

Tahap berikutnya adalah penerapan algoritma C4.5, yang dilakukan melalui dua pendekatan. Pertama, perhitungan manual menggunakan rumus *entropy dan information gain* untuk membentuk struktur pohon keputusan secara teoritis. Kedua, implementasi algoritma menggunakan perangkat lunak RapidMiner guna memperoleh model klasifikasi secara otomatis.

Hasil model dari kedua pendekatan tersebut kemudian dibandingkan untuk memastikan konsistensi struktur dan pola klasifikasi. Selanjutnya, dilakukan evaluasi model menggunakan *confusion matrix* untuk menilai tingkat akurasi, *presisi, recall, dan F1-score*. Seluruh tahapan penelitian ini dirancang secara sistematis agar menghasilkan model klasifikasi yang akurat, dapat dipahami, dan berpotensi digunakan sebagai dasar dalam sistem pendukung keputusan mitigasi banjir di DKI Jakarta.

2.2 Identifikasi Masalah

Identifikasi masalah merupakan fondasi utama dalam penelitian karena menetapkan arah dan tujuan ilmiah [7]. Tahap ini bertujuan menentukan fokus penelitian melalui kajian literatur dan observasi lapangan. Berdasarkan data BPBD DKI Jakarta, banjir merupakan permasalahan tahunan yang belum diolah secara sistematis dalam bentuk klasifikasi wilayah. Selama ini data banjir masih bersifat deskriptif,

sehingga diperlukan model klasifikasi berbasis algoritma C4.5 untuk mengenali wilayah rawan banjir dengan lebih terukur dan mudah dipahami.

2.3 Pengumpulan Data

Data yang digunakan merupakan data sekunder dari BPBD DKI Jakarta yang mencakup periode 2020–2023, dengan variabel seperti nama wilayah kelurahan, tahun kejadian, frekuensi banjir, dan tinggi muka air. Data diperoleh melalui studi dokumentasi dan diunduh dalam format *spreadsheet* dari situs resmi BPBD. Proses validasi dilakukan untuk memastikan data bebas dari nilai ganda (*duplicate*) dan missing value.

2.4 Pra-pemrosesan Data

Pra-pemrosesan data dilakukan agar data siap digunakan dalam proses klasifikasi. Tahapan ini meliputi:

- a. *Data cleaning*, yaitu menghapus data ganda dan memperbaiki nilai kosong.
- b. *Attribute selection*, yaitu memilih atribut yang paling relevan terhadap kejadian banjir.
- c. *Data transformation*, yaitu mengonversi data numerik menjadi kategorikal agar sesuai dengan format algoritma C4.5

Hasil dari tahap ini berupa dataset bersih dengan variabel target berupa status banjir, dikategorikan menjadi Aman dan Rawan.

2.5 Implementasi Algoritma C4.5

Algoritma C4.5 merupakan Salah satu algoritma yang digunakan untuk membentuk pohon keputusan[8]. Algoritma C4.5 digunakan untuk membangun model *Decision Tree* berdasarkan atribut-atribut yang paling memengaruhi kejadian banjir. Proses perhitungannya melibatkan dua konsep utama, yaitu *Entropy* dan *Information Gain*, sebagaimana ditunjukkan pada rumus berikut:

$$Entropy(S) = \sum_{i=1}^n -p_i * \log_2 p_i$$

Keterangan:

- S : Himpunan Kasus
- A : Fitur
- N : Jumlah Partisi S
- Pi : Proporsi Dari Si Terhadap S

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy(S_i)$$

Keterangan:

- A : Atribut
- N : Jumlah Partisi Atribut A
- |Si| : Jumlah Kasus Pada Partisi Ke-i
- |S| : Jumlah Kasus Dalam S

Atribut dengan nilai *gain* tertinggi akan menjadi akar pohon keputusan (*root node*). Proses ini berlanjut hingga seluruh data terklasifikasi sempurna pada simpul daun (*leaf node*).

Dalam penelitian ini, algoritma C4.5 diterapkan untuk membangun model klasifikasi tingkat kerawanan banjir di wilayah DKI Jakarta berdasarkan data historis dari BPBD periode 2020–2023. Proses ini bertujuan untuk menghasilkan pohon keputusan yang mampu mengelompokkan wilayah ke dalam kategori Aman dan Rawan dengan tingkat akurasi yang baik. Tahapan implementasi dilakukan melalui dua pendekatan, yaitu perhitungan manual dan implementasi otomatis menggunakan RapidMiner Studio. Pada tahap manual, dilakukan serangkaian langkah utama sebagai berikut.

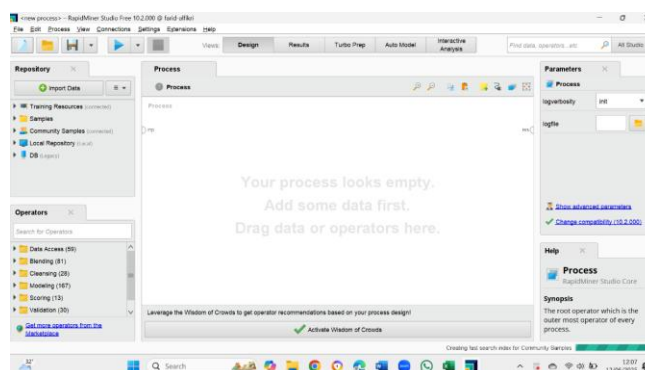
Pertama, dilakukan persiapan dataset, meliputi pengumpulan dan pembersihan data banjir dari BPBD agar sesuai dengan kebutuhan analisis. Proses ini mencakup pengkodean atribut dan penentuan label kelas berdasarkan tingkat kerawanan wilayah.

Kedua, dilakukan perhitungan manual dan pembentukan cabang pohon keputusan. Tahapan ini dimulai dengan perhitungan *entropy* dataset untuk mengukur tingkat ketidakpastian distribusi kategori kerawanan. Nilai *entropy* tersebut menjadi dasar dalam menghitung *information gain* untuk setiap atribut, seperti Tahun 2020–2023 atau Frekuensi kejadian banjir. Atribut dengan nilai *gain* tertinggi ditetapkan sebagai *root node*, karena atribut tersebut paling berpengaruh dalam membedakan kategori data. Selanjutnya, dataset dibagi berdasarkan nilai atribut pada *root node* untuk membentuk cabang dan *node* berikutnya. Proses ini dilakukan secara berulang hingga seluruh data terklasifikasi pada *leaf node*.

Hasil perhitungan manual tersebut menghasilkan pohon keputusan yang menunjukkan bagaimana atribut-atribut tertentu berperan dalam menentukan kategori kerawanan banjir. Model ini kemudian diuji dan diverifikasi menggunakan RapidMiner Studio dengan operator *Decision Tree* (C4.5) untuk membentuk model klasifikasi secara otomatis.

2.6 Implementasi Menggunakan RapidMiner

Setelah proses manual selesai, algoritma C4.5 diimplementasikan menggunakan RapidMiner Studio, karena perangkat lunak ini menyediakan antarmuka visual yang mempermudah proses analisis tanpa harus menulis kode pemrograman. Dataset hasil pra-pemrosesan diimpor ke RapidMiner, kemudian operator *Decision Tree* (C4.5) dijalankan untuk membangun model klasifikasi. Hasil berupa struktur pohon keputusan yang divisualisasikan dan dibandingkan dengan hasil perhitungan manual guna memastikan konsistensi model.



Gambar 2. Tampilan Awal RapidMiner

2.7 Evaluasi Model

Evaluasi model merupakan tahap penting untuk menilai kinerja algoritma C4.5 dalam melakukan klasifikasi data sesuai kategori yang telah ditentukan. Tujuan utama evaluasi adalah untuk mengukur sejauh mana model mampu memberikan hasil prediksi yang akurat, seimbang, dan konsisten antara data latih dan data uji [9]

Dalam penelitian ini, evaluasi dilakukan menggunakan Confusion Matrix, yaitu matriks yang membandingkan hasil prediksi model dengan data aktual. Matriks ini menghasilkan empat komponen utama: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). Nilai-nilai tersebut digunakan untuk menghitung beberapa metrik evaluasi, yaitu akurasi, presisi, *recall*, dan *F1-score*.

2.8 Analisis dan interpretasi Hasil

Setelah proses pengujian selesai dilakukan, langkah selanjutnya adalah memaparkan hasil yang diperoleh, baik melalui penyajian tabel, data numerik, maupun dalam bentuk uraian naratif[10]. Model yang dibangun menggunakan algoritma C4.5 menunjukkan kinerja yang sangat andal dengan tingkat akurasi mencapai hampir 95%. Hasil ini membuktikan bahwa algoritma C4.5 cukup efektif dalam melakukan prediksi tingkat kerawanan banjir berdasarkan data historis yang digunakan.

Struktur pohon keputusan yang dihasilkan terlihat sederhana namun kuat. Atribut tahun 2022 muncul sebagai faktor utama dalam proses klasifikasi, menandakan bahwa data yang lebih mutakhir memiliki pengaruh dominan dibandingkan data tahun-tahun sebelumnya. Hal ini menunjukkan bahwa dinamika kejadian banjir cenderung mengikuti pola terbaru dan relevan dengan kondisi terkini. Meski demikian, model masih memiliki kelemahan berupa adanya 8 kasus *False Negative* (FN), yaitu wilayah yang sebenarnya rawan tetapi diklasifikasikan sebagai aman. Kesalahan jenis ini tergolong berisiko karena dapat menyebabkan wilayah yang rentan banjir tidak mendapatkan perhatian atau upaya mitigasi yang memadai. Dalam konteks kebijakan penanggulangan bencana, hal ini perlu menjadi perhatian khusus agar tidak terjadi under-estimation terhadap potensi risiko banjir.

Sebaliknya, kesalahan *False Positive* (FP) yaitu wilayah aman yang terprediksi rawan—hanya menyebabkan kelebihan alokasi sumber daya, yang secara praktis tidak terlalu berbahaya dibandingkan kesalahan FN. Oleh karena itu, perbaikan model ke depan sebaiknya difokuskan pada pengurangan nilai *False Negative*, agar hasil klasifikasi semakin akurat dan aman untuk dijadikan dasar pengambilan keputusan. Selain itu, nilai *F1-Score* yang seimbang antara kelas “rawan” dan “aman” menunjukkan bahwa model tidak memiliki bias terhadap salah satu kelas. Dengan demikian, dapat disimpulkan bahwa algoritma C4.5 layak digunakan sebagai model klasifikasi kerawanan banjir, meskipun masih perlu dilakukan penyempurnaan untuk meminimalkan kesalahan prediksi, khususnya pada kategori wilayah rawan banjir.

3. Hasil dan Pembahasan

Penelitian ini menghasilkan model klasifikasi tingkat kejadian banjir di wilayah DKI Jakarta menggunakan algoritma C4.5 berbasis *Decision Tree*. Data yang digunakan merupakan data historis kejadian banjir yang diperoleh dari BPBD DKI Jakarta untuk periode 2020–2023. Tujuan utama penelitian ini adalah untuk mengelompokkan wilayah berdasarkan tingkat kerawanan banjir agar dapat mendukung upaya mitigasi dan perencanaan kebijakan yang lebih tepat sasaran.

3.1 Analisis Data dan Pembentukan Label Kelas

Data awal yang diperoleh terdiri dari catatan kejadian banjir di 267 wilayah administrasi DKI Jakarta. Setiap entri berisi informasi mengenai frekuensi banjir dari tahun 2020 hingga 2023. Data kemudian melalui tahap pra-pemrosesan, meliputi pembersihan dari data ganda dan kosong, penyesuaian format atribut, serta transformasi nilai numerik menjadi bentuk kategorikal agar sesuai dengan kebutuhan algoritma C4.5.

Label kelas dibentuk berdasarkan frekuensi kejadian banjir di tiap wilayah. Wilayah yang memiliki frekuensi kejadian 0–1 kali dikategorikan sebagai Aman, sedangkan wilayah dengan kejadian ≥ 2 kali dikategorikan sebagai Rawan. Proses ini menghasilkan dua kelas target yang digunakan dalam tahap pelatihan model.

WILAYAH	LOKASI	BANJIR_2020	BANJIR_2021	BANJIR_2022	BANJIR_2023	FREKUENSI
KOTA JAKARTA PUSAT	Cempaka Putih Barat	3	0	1	0	3
KOTA JAKARTA PUSAT	Cempaka Putih Timur	4	0	0	0	4
KOTA JAKARTA PUSAT	Rawasari	1	0	0	0	1
KOTA JAKARTA PUSAT	Gambir	0	0	0	0	0
KOTA JAKARTA PUSAT	Kebon Kelapa	1	0	0	0	1
KOTA JAKARTA PUSAT	Petjojo Utara	0	0	0	0	0
KOTA JAKARTA PUSAT	Galur	0	0	0	0	0
KOTA JAKARTA PUSAT	Johar Baru	6	0	1	0	7
KOTA JAKARTA PUSAT	Kampung Rawa	1	0	0	0	1
KOTA JAKARTA PUSAT	Tanah Tinggi	1	0	0	0	1
KOTA JAKARTA PUSAT	Cempaka Baru	2	0	0	0	2
KOTA JAKARTA PUSAT	Gunung Sahari Selatan	6	0	0	0	6
KOTA JAKARTA PUSAT	Harapan Mulya	0	0	0	0	0
KOTA JAKARTA PUSAT	Kebon Kosong	10	0	0	0	10
KOTA JAKARTA PUSAT	Kemayoran	1	0	0	0	1
KOTA JAKARTA PUSAT	Serdang	4	0	0	0	4
KOTA JAKARTA PUSAT	Sumur Batu	1	0	0	0	1
KOTA JAKARTA PUSAT	Utan Panjang	0	0	0	0	0
KOTA JAKARTA PUSAT	Gondangdia	0	0	0	0	0
KOTA JAKARTA PUSAT	Kebon Sirih	3	0	0	0	3
KOTA JAKARTA PUSAT	Wenteng	6	0	0	0	6
KOTA JAKARTA PUSAT	Gunung Sahari Utara	2	0	0	0	2
KOTA JAKARTA PUSAT	Karang Anyar	1	0	0	0	1
KOTA JAKARTA PUSAT	Kartini	0	0	0	0	0

Gambar 3. Cuplikan Hasil *Cleaning data* dan *Attribute selection*

	A	B	C	D	E	F
1	WILAYAH	Banjir_2020	Banjir_2021	Banjir_2022	Banjir_2023	STATUS BANJIR
2	KJP	tinggi	tidak banjir	rendah	tidak banjir	rawan
3	KJP	tinggi	tidak banjir	tidak banjir	tidak banjir	aman
4	KJP	rendah	tidak banjir	tidak banjir	tidak banjir	aman
5	KJP	tidak banjir	tidak banjir	tidak banjir	tidak banjir	aman
6	KJP	rendah	tidak banjir	tidak banjir	tidak banjir	aman
7	KJP	tidak banjir	tidak banjir	tidak banjir	tidak banjir	aman
8	KJP	tidak banjir	tidak banjir	tidak banjir	tidak banjir	aman
9	KJP	tinggi	tidak banjir	rendah	tidak banjir	rawan
10	KJP	rendah	tidak banjir	tidak banjir	tidak banjir	aman
11	KJP	rendah	tidak banjir	tidak banjir	tidak banjir	aman
12	KJP	rendah	tidak banjir	tidak banjir	tidak banjir	aman
13	KJP	tinggi	tidak banjir	tidak banjir	tidak banjir	aman
14	KJP	tidak banjir	tidak banjir	tidak banjir	tidak banjir	aman
15	KJP	tinggi	tidak banjir	tidak banjir	tidak banjir	aman
16	KJP	rendah	tidak banjir	tidak banjir	tidak banjir	aman
17	KJP	tinggi	tidak banjir	tidak banjir	tidak banjir	aman
18	KJP	rendah	tidak banjir	tidak banjir	tidak banjir	aman
19	KJP	tidak banjir	tidak banjir	tidak banjir	tidak banjir	aman

Gambar 4. Hasil Transformasi Data

3.2 Pembentukan Pohon Keputusan Secara Manual

Tahap selanjutnya adalah membangun pohon keputusan secara manual menggunakan konsep *Entropy* dan *Information Gain* untuk menentukan atribut terbaik yang menjadi akar pohon (*root node*).

Perhitungan *entropy* dilakukan menggunakan Persamaan (1):

$$Entropy(S) = \sum_{i=1}^n -p_i * \log_2 p_i$$

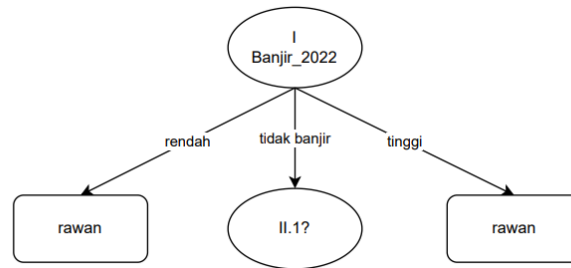
Sedangkan perhitungan *Information Gain* ditunjukkan pada Persamaan (2):

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^n \frac{|S_i|}{|S|} * Entropy(S_i)$$

Untuk menentukan *node* akar pada pohon keputusan, dilakukan perhitungan *entropy* dan *information gain* terhadap seluruh atribut yang digunakan dalam penelitian. Perhitungan ini bertujuan untuk mengetahui atribut mana yang memiliki nilai *gain* tertinggi, karena atribut dengan nilai *gain* tertinggi dipilih sebagai *node* akar. Pada tahap transformasi data, dilakukan penyederhanaan penulisan nama wilayah ,misalnya Kota Jakarta Pusat menjadi KJP, Kota Jakarta Barat menjadi KJB, Kota Jakarta Timur menjadi KJT, Kota Jakarta Selatan menjadi KJS, Kota Jakarta Utara menjadi KJU, dan Pulau Seribu menjadi PS. Adapun ringkasan hasil perhitungan *entropy* dan *gain* untuk setiap atribut dapat dilihat pada table berikut:

Tabel 1. Perhitungan *Node* 1

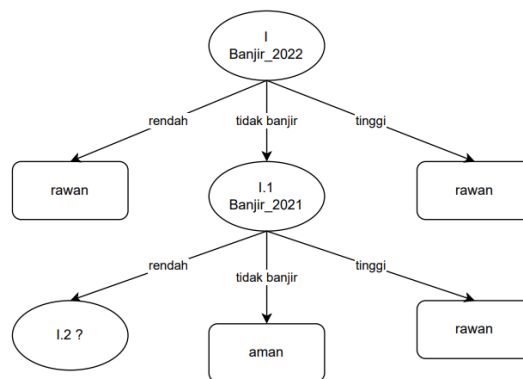
Node	Jumlah (S)	Aman (Si)	Rawan (Si)	Entropy	Gain
Total	245	104	141	0,983484995	
Wilayah					0,135279424
KJP	33	27	6	0,684038436	
KJU	31	17	14	0,99323382	
KJB	52	23	29	0,990374836	
KJS	59	9	50	0,616166193	
KJT	64	24	40	0,954434003	
PS	6	4	2	0,918295834	
Banjir_2020					0,186301452
tidak banjir	30	28	2	0,353359335	
rendah	61	37	24	0,966985296	
tinggi	154	39	115	0,816383668	
Banjir_2021					0,463294208
tidak banjir	127	98	29	0,775117651	
rendah	66	6	60	0,439496987	
tinggi	52	0	52	0	
Banjir_2022					0,464706111
tidak banjir	135	101	34	0,814200068	
rendah	60	3	57	0,286396957	
tinggi	50	0	50	0	
Banjir_2023					0,367333338
tidak banjir	161	104	57	0,937622086	
rendah	66	0	66	0	
tinggi	18	0	18	0	



Gambar 5. Tampilan node 1

Tabel 2. Perhitungan Node 1.1

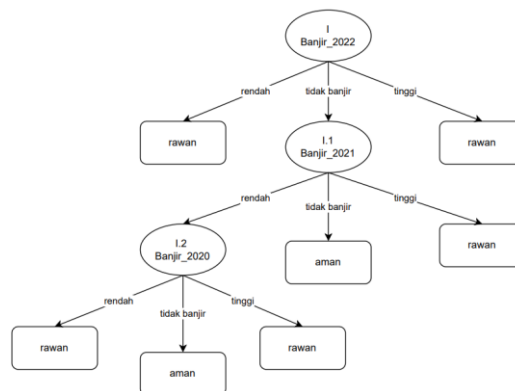
Node	jumlah (S)	aman (Si)	rawan (Si)	Entropy	Gain
Total	135	101	34	0.814200068	
Banjir_2020					0.157326620
tidak banjir	27	27	0	0	
rendah	40	35	5	0.543564443	
tinggi	68	39	29	0.984343203	
Banjir_2021					0.407163978
tidak banjir	103	96	7	0.358268639	
rendah	25	5	20	0.721928095	
tinggi	7	0	7	0	
Banjir_2023					0.160832840
tidak banjir	125	101	24	0.705636606	
rendah	10	0	10	0	
tinggi	0	0	0	0	



Gambar 6. Tampilan node 1.1

Tabel 3. Perhitungan *Node* 1.2

Node	jumlah (S)	aman (Si)	rawan (Si)	Entropy	Gain
Total	25	5	20	0.721928095	
Banjir_2020					0.35958209
tidak banjir	3	3	0	0	
rendah	4	0	4	0	
tinggi	18	2	16	0.50325833	
Banjir_2023					0.02698272
tidak banjir	23	5	18	0.75537541	
rendah	2	0	2	0	
tinggi	0	0	0	0	



Gambar 7. Tampilan *Node* 1.2

Algoritma C4.5 yang diterapkan pada data historis banjir DKI Jakarta menghasilkan struktur pohon keputusan dengan atribut paling dominan sebagai simpul akar, yaitu kejadian banjir tahun 2022. Hal ini logis, karena data terbaru cenderung lebih merepresentasikan kondisi aktual, sehingga menjadi faktor penentu utama dalam klasifikasi wilayah.

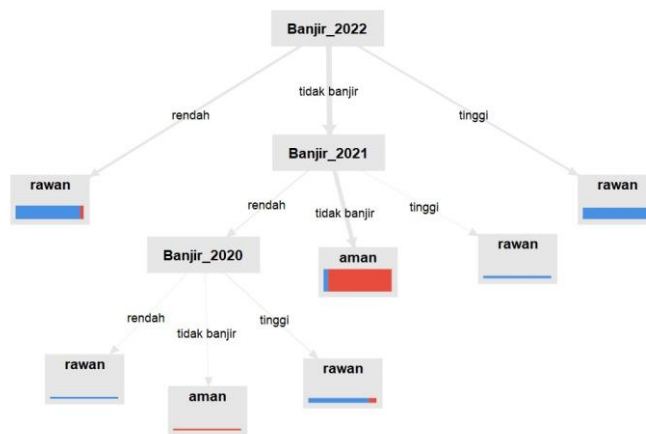
Struktur aturan yang terbentuk dapat dijabarkan sebagai berikut:

1. Jika Banjir 2022 = rendah → kelas: rawan.
2. Jika Banjir 2022 = tinggi → kelas: rawan.
3. Jika Banjir 2022 = tidak banjir → lanjut ke Banjir 2021.
4. Jika Banjir 2021 = tidak banjir → kelas: aman.
5. Jika Banjir 2021 = tinggi → kelas: rawan.
6. Jika Banjir 2021 = rendah → lanjut ke Banjir 2020:
7. Jika Banjir 2020 = rendah → rawan.
8. Jika Banjir 2020 = tinggi → rawan.
9. Jika Banjir 2020 = tidak banjir → aman.

3.3 Implementasi Algoritma Menggunakan RapidMiner

implementasi algoritma dilakukan menggunakan RapidMiner Studio, dengan tujuan untuk memvalidasi hasil manual sekaligus menghasilkan visualisasi model pohon keputusan secara otomatis. Dataset hasil pra-pemrosesan diimpor ke dalam RapidMiner, kemudian ditentukan atribut target “Status Banjir” dengan dua label kelas: Aman dan Rawan.

Hasil visualisasi pohon keputusan dapat dilihat pada Gambar berikut



Gambar 1. Pohon Keputusan Hasil Implementasi Algoritma C4.5

3.4 Evaluasi Model

Evaluasi ini bertujuan untuk mengukur sejauh mana model yang telah dilatih mampu melakukan klasifikasi dengan benar terhadap data baru atau data uji. Dengan kata lain, evaluasi dilakukan untuk mengetahui tingkat kinerja (*performance*) dari model yang dihasilkan.

Berikut gambar hasil evaluasi dengan menggunakan operator *Performance (Classification)*

	true rawan	true aman	class precision
pred. rawan	133	5	96.38%
pred. aman	8	99	92.52%
class recall	94.33%	95.19%	

accuracy: 94.72% +/- 5.44% (micro average: 94.69%)

Gambar Hasil Evaluasi Model

Berikut adalah perhitungan manual dari angka-angka Tersebut:

1. Accuracy (Akurasi)

Mengukur seberapa banyak prediksi yang benar dibanding total data.

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} = \frac{133 + 99}{245} = \frac{232}{245} = 0,9472 = 94,72\%$$

Dengan demikian, secara keseluruhan model mampu menghasilkan tingkat akurasi prediksi sebesar 94,72%.

2. Precision (Ketepatan)

Mengukur seberapa tepat model saat memprediksi suatu kelas.

- Untuk kelas rawan:

$$Precision\ rawan = \frac{TP}{TP + FT} = \frac{133}{133 + 5} = \frac{133}{138} = 0,9638 = 96.38\%$$

Artinya, dari semua prediksi rawan, 96,38% memang benar rawan.

- Untuk kelas aman:

$$Precision\ aman = \frac{TN}{TN + FN} = \frac{99}{99 + 8} = \frac{99}{107} = 0,9252 = 92.52\%$$

Artinya, dari semua prediksi aman, 92,52% memang benar aman.

3. Recall (sensitivitas / keberhasilan mendeteksi)

Mengukur seberapa baik model mengenali suatu kelas yang sebenarnya.

- Untuk kelas rawan:

$$Recall\ rawan = \frac{TP}{TP + FN} = \frac{133}{133 + 8} = \frac{133}{141} = 0,9433 = 94.33\%$$

Artinya, dari semua wilayah yang benar-benar rawan, model berhasil mendeteksi 94,33% dengan benar.

- Untuk kelas aman:

$$Recall\ aman = \frac{TN}{TN + FN} = \frac{99}{99 + 5} = \frac{99}{104} = 0,9519 = 95,19\%$$

Artinya, dari semua wilayah yang benar-benar aman, model berhasil mendeteksi 95,19% dengan benar.

4. F1-Score (harmonic mean precision & recall)

Digunakan untuk menyeimbangkan precision dan recall.

- Untuk kelas rawan:

$$F1_{rawan} = 2 \times \frac{Precision \times Recall}{Precision + Recall} = 2 \times \frac{0,9638 \times 0,9433}{0,9638 + 0,9433} = 2 \times \frac{0,9089}{1,9071} = 0,9528 = 95,28\%$$

- Untuk kelas aman:

$$F1_{aman} = 2 \times \frac{0,9252 \times 0,9519}{0,9252 + 0,9519} = 2 \times \frac{0,8809}{1,8771} = 0,9388 = 93.88\%$$

Artinya, kinerja model seimbang antara ketepatan dan keberhasilan deteksi di kedua kelas.

Hasil evaluasi model diperoleh melalui *confusion matrix*, akurasi, presisi, *recall*, serta perhitungan *F1-score*. Berikut interpretasinya secara rinci:

1. Akurasi Model

Nilai akurasi yang diperoleh adalah 94,72%, yang menunjukkan bahwa hampir semua prediksi model sesuai dengan kondisi sebenarnya di lapangan. Angka ini tergolong sangat tinggi untuk data bencana, yang biasanya kompleks dan penuh variabel tak terduga.

2. Confusion Matrix

Dari *confusion matrix* diketahui:

a. *True Positive* (TP): sebagian besar wilayah yang rawan berhasil dikenali sebagai rawan.

b. *True Negative* (TN): mayoritas wilayah yang aman berhasil diprediksi aman.

c. *False Positive* (FP): terdapat 5 wilayah aman yang diprediksi rawan. Konsekuensinya adalah kemungkinan alokasi sumber daya ke wilayah yang sebenarnya tidak terlalu membutuhkan. Namun dampak ini relatif kecil karena sifatnya hanya “*over-alert*”.

d. *False Negative* (FN): terdapat 8 wilayah rawan yang salah diprediksi aman. Inilah kelemahan paling serius, karena wilayah yang sebenarnya butuh kewaspadaan justru dianggap aman, sehingga berpotensi menurunkan kesiapsiagaan pemerintah maupun masyarakat.

3. Presisi dan *Recall*

Kelas rawan:

Presisi = 96,38% → hampir semua prediksi rawan benar-benar rawan.

Recall = 94,33% → sebagian besar wilayah rawan berhasil terdeteksi, meskipun masih ada 8 wilayah yang terlewat.

Kelas aman:

Presisi = 92,52% → sebagian besar prediksi aman memang benar aman.

Recall = 96,19% → model sangat baik dalam mendeteksi wilayah aman.

Interpretasinya, model lebih hati-hati terhadap prediksi rawan, sehingga tingkat kesalahannya kecil. Namun, *recall* untuk kelas rawan sedikit lebih rendah dibanding *recall* aman, karena masih ada kasus rawan yang tidak terdeteksi.

4. *F1-Score*

Karena RapidMiner tidak menampilkan *F1-score*, maka perhitungan dilakukan manual, dengan hasil :

a. *F1-score* (rawan): ~95,34%

b. *F1-score* (aman): ~94,32%

c. Rata-rata makro: ~94,83%

d. Rata-rata tertimbang: ~94,91%

F1-score memberikan gambaran keseimbangan antara presisi dan *recall*. Hasil di atas menunjukkan bahwa kinerja model sama baiknya pada kedua kelas, dengan perbedaan yang tidak signifikan.

3.5 Interpretasi Hasil

Model yang dibangun menggunakan algoritma C4.5 menunjukkan kinerja yang sangat andal dengan tingkat akurasi mencapai hampir 95%. Hasil ini membuktikan bahwa algoritma C4.5 cukup efektif dalam melakukan prediksi tingkat kerawanan banjir berdasarkan data historis yang digunakan.

Struktur pohon keputusan yang dihasilkan terlihat sederhana namun kuat. Atribut tahun 2022 muncul sebagai faktor utama dalam proses klasifikasi, menandakan bahwa data yang lebih mutakhir memiliki pengaruh dominan dibandingkan data tahun-tahun sebelumnya. Hal ini menunjukkan bahwa dinamika kejadian banjir cenderung mengikuti pola terbaru dan relevan dengan kondisi terkini. Meski demikian, model masih memiliki kelemahan berupa adanya 8 kasus *False Negative* (FN), yaitu wilayah yang sebenarnya rawan tetapi diklasifikasikan sebagai aman. Kesalahan jenis ini tergolong berisiko karena dapat menyebabkan wilayah yang rentan banjir tidak mendapatkan perhatian atau upaya mitigasi

yang memadai. Dalam konteks kebijakan penanggulangan bencana, hal ini perlu menjadi perhatian khusus agar tidak terjadi under-estimation terhadap potensi risiko banjir.

Sebaliknya, kesalahan *False Positive* (FP)—yaitu wilayah aman yang terprediksi rawan—hanya menyebabkan kelebihan alokasi sumber daya, yang secara praktis tidak terlalu berbahaya dibandingkan kesalahan FN. Oleh karena itu, perbaikan model ke depan sebaiknya difokuskan pada pengurangan nilai *False Negative*, agar hasil klasifikasi semakin akurat dan aman untuk dijadikan dasar pengambilan keputusan. Selain itu, nilai *FI-Score* yang seimbang antara kelas “rawan” dan “aman” menunjukkan bahwa model tidak memiliki bias terhadap salah satu kelas. Dengan demikian, dapat disimpulkan bahwa algoritma C4.5 layak digunakan sebagai model klasifikasi kerawanan banjir, meskipun masih perlu dilakukan penyempurnaan untuk meminimalkan kesalahan prediksi, khususnya pada kategori wilayah rawan banjir.

4. Kesimpulan dan Saran

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa penerapan algoritma C4.5 berbasis *Decision Tree* dalam klasifikasi tingkat kerawanan banjir di wilayah DKI Jakarta menunjukkan kinerja yang sangat baik dengan tingkat akurasi mencapai hampir 95%. Hasil ini menegaskan bahwa algoritma C4.5 efektif dalam mengenali pola kejadian banjir berdasarkan data historis dari BPBD. Struktur pohon keputusan yang dihasilkan juga menunjukkan pola yang sederhana namun kuat, dengan atribut tahun 2022 sebagai faktor dominan dalam proses klasifikasi. Hal ini menunjukkan bahwa data terkini memiliki pengaruh yang lebih besar terhadap hasil klasifikasi dibandingkan data pada tahun-tahun sebelumnya.

Meskipun demikian, model masih memiliki kelemahan berupa munculnya kasus *False Negative* pada delapan wilayah, di mana area yang sebenarnya rawan banjir terklasifikasi sebagai wilayah aman. Kondisi ini penting diperhatikan karena dapat menyebabkan kekeliruan dalam penentuan prioritas mitigasi. Sebaliknya, kesalahan *False Positive* (wilayah aman terklasifikasi sebagai rawan) cenderung tidak menimbulkan dampak besar, meskipun dapat menyebabkan alokasi sumber daya yang berlebih. Selain itu, nilai *FI-score* yang seimbang menunjukkan bahwa model mampu mengklasifikasikan kedua kategori, yaitu aman dan rawan, dengan kinerja yang proporsional tanpa bias terhadap salah satu kelas.

Adapun saran untuk penelitian selanjutnya adalah melakukan peningkatan pada proses pra-pemrosesan data, khususnya dalam penanganan data tidak seimbang (*imbalanced data*) agar dapat menekan tingkat kesalahan *False Negative*. Selain itu, disarankan untuk mengintegrasikan algoritma C4.5 dengan metode lain seperti *Random Forest* atau *Gradient Boosting* guna memperoleh hasil klasifikasi yang lebih akurat dan stabil. Penambahan variabel baru seperti curah hujan, kepadatan penduduk, dan kondisi drainase juga dapat meningkatkan kemampuan model dalam memprediksi tingkat kerawanan banjir secara lebih komprehensif.

Daftar Pustaka

- [1] A. M. Alawiyah, “Journal of Geospatial Information Science and Engineering,” vol. 4, no. 2, pp. 95–101, 2021, doi: 10.22146/jgise.
- [2] S. Azizah, *Sustainable Livelihood Strategy , Keberlanjutan Usaha Peternakan sapi Pedaging Pasca Bencana Banjir*, no. December. 2024.
- [3] B. B. Karnisah Iin, Astor. Yackob, *Sistem informasi geografis (sig) pengendalian banjir*, no. 043. 2019.
- [4] I. Risnawati *et al.*, “Klasifikasi Data Mining Untuk Mengestimasi Potensi Curah Hujan Berdampak Banjir Daerah Menggunakan Algoritma C4.5,” *J. Insa. J. Inf. Syst. Manag. Innov.*, vol. 3, no. 2, pp. 78–84, 2023, doi: 10.31294/jinsan.v3i2.3050.
- [5] M. F. Nasrullah, R. R. Saedudin, and F. Hamami, “Perbandingan Akurasi Algoritma C4.5 Dan K-Nearest Neighbors Untuk Klasifikasi Curah Hujan Berdasarkan Iklim Indonesia,” *JUPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 9, no. 2, pp. 628–638, 2024, doi: 10.29100/jupi.v9i2.4655.
- [6] A. Abdullah and I. Koma, “Prediksi Banjir Di Kota Pontianak Menggunakan Metode Decision Tree C4 . 5,” *Justek J. Sains Dan Teknol.*, vol. 8, no. 1, pp. 40–50, 2025.
- [7] R. Nurdianyani Sari and S. S. Azharina, “Penerapan Algoritma C4.5 Untuk Memprediksi Bencana

-
- Gunung Meletus Di Indonesia,” *Jts*, vol. 3, no. 2, pp. 1–9, 2024.
- [8] I. Iddrus and D. W. Sari, “Penerapan Data Mining Menggunakan Algoritma Decision Tree C4.5 Untuk Memprediksi Mahasiswa Drop Out Di Universitas Wiraraja,” *J. Adv. Res. Inform.*, vol. 1, no. 02, pp. 1–7, 2023, doi: 10.24929/jars.v1i02.2684.
- [9] F. N. Umma, B. Warsito, and D. A. I. Maruddani, “Klasifikasi Status Kemiskinan Rumah Tangga Dengan Algoritma C5.0 Di Kabupaten Pematang,” *J. Gaussian*, vol. 10, no. 2, pp. 221–229, 2021, doi: 10.14710/j.gauss.v10i2.29934.
- [10] I. A. Siregar, “ALACRITY : Journal Of Education,” vol. 1, no. 2, pp. 39–48, 2021.