

# Analisis Perbandingan Algoritma Klasifikasi Data Mining untuk Penentuan Lokasi Perumahan

Andi Ernawati\*, Muhammad Iqbal

Program studi Magister Teknologi Informasi, Universitas Pembangunan Pancabudi, Medan  
Jl. Gatot Subroto, Simpang Tj., Kec. Medan Sunggal, Kota Medan, Sumatera Utara, Indonesia

Email: <sup>1,\*</sup>aernawati296@gmail.com, <sup>2</sup>muhammadiqbal@dosenpancabudi.ac.id

Email Penulis Korespondensi: aernawati296@gmail.com

Submitted: 25/12/2024; Accepted: 14/01/2025; Published: 16/01/2025

**Abstrak**—Penelitian ini bertujuan untuk menganalisis penerapan algoritma C5.0 dan K-Nearest Neighbor (K-NN) dalam proses klasifikasi untuk penentuan lokasi perumahan. Proses klasifikasi ini melibatkan beberapa faktor seperti harga tanah, aksesibilitas, fasilitas publik, tingkat kejahatan, infrastruktur, ketersediaan lahan, dan preferensi konsumen. Dalam penelitian ini, dilakukan pengujian terhadap kedua algoritma tersebut untuk membandingkan kinerja dalam menghasilkan prediksi yang akurat. Hasil pengujian menunjukkan bahwa algoritma C5.0 memberikan kinerja yang lebih baik dibandingkan dengan K-NN. Algoritma C5.0 mencapai tingkat akurasi sebesar 100%, sedangkan K-NN hanya mencapai akurasi sebesar 66,67%. Hal ini menunjukkan bahwa C5.0 lebih efektif dalam memodelkan data dan menghasilkan klasifikasi yang lebih tepat. Dengan demikian, dapat disimpulkan bahwa penggunaan algoritma data mining, khususnya C5.0, sangat membantu dalam proses klasifikasi untuk penentuan lokasi perumahan, memberikan hasil yang lebih optimal dibandingkan dengan K-NN.

**Kata Kunci:** C5.0; K-Nearest Neighbor; Klasifikasi; Data Mining; Lokasi Perumahan; Akurasi

**Abstract**—This study aims to analyze the application of C5.0 and K-Nearest Neighbor (K-NN) algorithms in the classification process for determining the optimal location for housing. The classification process involves several factors such as land price, accessibility, public facilities, crime rate, infrastructure, land availability, and consumer preferences. The research conducted tests on both algorithms to compare their performance in generating accurate predictions. The results show that the C5.0 algorithm outperforms K-NN, achieving an accuracy rate of 100%, compared to K-NN, which achieved an accuracy of 66.67%. This demonstrates that C5.0 is more effective in modeling data and producing more precise classifications. Therefore, it can be concluded that the use of data mining algorithms, particularly C5.0, greatly assists in the classification process for determining housing locations, providing more optimal results compared to K-NN.

**Keywords:** C5.0; K-Nearest Neighbor; Classification; Data Mining; Housing Location; Accuracy

## 1. PENDAHULUAN

Menurut Undang-Undang Nomor 1 Tahun 2011 tentang Perumahan dan Pemukiman, perumahan diartikan sebagai sekumpulan rumah yang difungsikan sebagai tempat tinggal atau kawasan hunian. Lingkungan perumahan ini harus didukung oleh sarana dan prasarana yang memadai untuk mendukung kenyamanan, keamanan, dan kebutuhan penghuni. Lokasi perumahan merupakan salah satu aspek penting dalam perencanaan pembangunan kota. Pemilihan lokasi yang tepat tidak hanya memengaruhi nilai properti, tetapi juga berdampak pada kualitas hidup masyarakat, efisiensi transportasi, dan akses terhadap fasilitas publik seperti sekolah, rumah sakit, dan tempat kerja (Smith et al., 2020). Dengan semakin kompleksnya kebutuhan masyarakat modern, proses pemilihan lokasi perumahan kini semakin menuntut pendekatan berbasis data yang lebih akurat dan komprehensif.

Data mining adalah salah satu metode yang telah berkembang pesat dalam membantu pengambilan keputusan berbasis data. Teknik ini memungkinkan analisis pola dan hubungan tersembunyi dalam data besar, sehingga memberikan wawasan yang lebih mendalam untuk mendukung keputusan strategis [1]. Dalam konteks pemilihan lokasi perumahan, algoritma data mining, khususnya algoritma klasifikasi, dapat digunakan untuk memetakan dan mengevaluasi berbagai faktor seperti aksesibilitas, infrastruktur, tingkat kejahatan, harga tanah, hingga preferensi konsumen [2]–[7]

Namun, permasalahan utama yang masih dihadapi adalah kurangnya studi komparatif yang mendalam terkait kinerja berbagai algoritma klasifikasi dalam konteks penentuan lokasi perumahan. Sebagai contoh, algoritma seperti Decision Tree, Random Forest, dan Support Vector Machine memiliki keunggulan masing-masing dalam hal akurasi, efisiensi, dan interpretasi, tetapi belum ada panduan pasti tentang algoritma mana yang paling sesuai untuk digunakan pada data dengan karakteristik tertentu. Hal ini menjadi tantangan sekaligus peluang untuk penelitian lebih lanjut.

Penelitian ini bertujuan untuk membandingkan kinerja beberapa algoritma klasifikasi dalam menentukan lokasi perumahan yang optimal. Algoritma yang akan dibandingkan meliputi Decision Tree, dan K-Nearest Neighbor [4]. Penilaian akan dilakukan menggunakan metrik evaluasi seperti akurasi, precision, recall, dan F1-score, dengan metode validasi yang ketat seperti k-fold cross-validation. Hasil penelitian diharapkan tidak hanya memberikan wawasan tentang kinerja algoritma, tetapi juga menjadi panduan praktis bagi pengembang properti, pembuat kebijakan, dan akademisi yang tertarik pada implementasi data mining di sektor perumahan.

Melalui penelitian ini, diharapkan dapat ditemukan algoritma yang paling efektif dan efisien dalam menganalisis data lokasi perumahan, sehingga mampu memberikan kontribusi nyata terhadap pengambilan keputusan berbasis data di era digital ini.

Algoritma C5.0 adalah salah satu alat untuk menyelesaikan masalah dalam memilih lokasi perumahan. Salah satu permasalahan yang muncul dalam lokasi perumahan adalah ketidakpahaman terhadap karakteristik spesifik dari lokasi

terbaik yang akan digunakan dalam pembangunan. Untuk mengatasi masalah ini, algoritma C5.0 dan K-Nearest Neighbor diusulkan untuk melakukan proses perbandingan dalam menentukan kebutuhan.

Dalam konteks pemilihan lokasi usaha, algoritma ini dapat memberikan informasi yang akurat dan dapat digunakan sebagai bahan pertimbangan, sehingga tidak terjadi kesalahan dalam pemilihan lokasi dan dapat memberikan informasi yang relevan kepada calon pembeli. Penggunaan algoritma K-Nearest Neighbor dalam analisis lokasi usaha pada penelitian ini diharapkan dapat menghasilkan rekomendasi yang tepat. Dengan demikian, pemilihan lokasi usaha dapat dilakukan secara cepat, akurat, dan efisien.

Dalam penelitian sebelumnya yang dilakukan oleh Vinna Rahmayanti Setyaning Nastiti dan tim, berjudul “Penerapan Algoritma C5.0 Pada Analisis Faktor-Faktor Pengaruh Kelulusan Tepat Waktu Mahasiswa Teknik Informatika Universitas Muhammadiyah Malang,” disimpulkan bahwa Algoritma C5.0 mampu melakukan seleksi fitur secara efektif. Algoritma ini berhasil mengidentifikasi pengaruh faktor-faktor terhadap kelulusan tepat waktu mahasiswa Teknik Informatika UMM dengan memilih 8 fitur dari total 15 fitur yang tersedia, serta mencapai tingkat akurasi sebesar 91,9%. [8].

Dalam penelitian yang dilakukan oleh Tedi Permana dan tim, berjudul “Perbandingan Hasil Prediksi Kredit Macet Pada Koperasi Menggunakan Algoritma KNN dan C5.0,” disimpulkan bahwa algoritma C5.0 memiliki kinerja yang lebih baik dibandingkan dengan algoritma KNN dalam memprediksi kredit macet di koperasi. Berdasarkan analisis terhadap 30 data transaksi, algoritma C5.0 mencapai tingkat akurasi sebesar 86,67%, sedangkan algoritma KNN hanya mencapai akurasi sebesar 83,33%. Oleh karena itu, algoritma C5.0 dinilai lebih sesuai untuk dijadikan acuan dalam proses prediksi kredit macet [9].

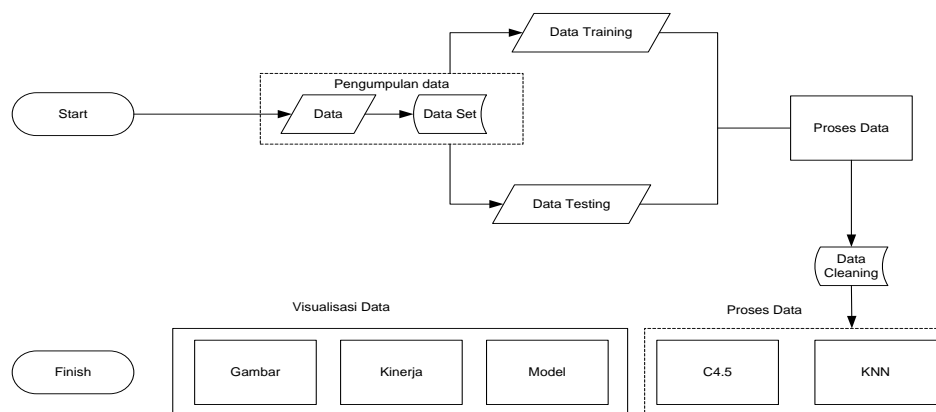
Dalam penelitian yang dilakukan oleh Fatiyah Nur Umma dan tim berjudul “Klasifikasi Status Kemiskinan Rumah Tangga Dengan Algoritma C5.0 di Kabupaten Pematang,” disimpulkan bahwa algoritma C5.0 mampu mengklasifikasikan status kemiskinan rumah tangga dengan tingkat akurasi rata-rata sebesar 91,16%. Hal ini menunjukkan bahwa algoritma C5.0 memiliki kinerja yang sangat baik dalam mengolah data untuk menentukan status kemiskinan rumah tangga di wilayah tersebut. [10].

Dalam penelitian yang dilakukan oleh Inna Alvi Nikmatun dan Indra Waspada berjudul “Implementasi Data Mining Untuk Klasifikasi Masa Studi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor,” disimpulkan bahwa algoritma K-Nearest Neighbor menunjukkan kinerja terbaik pada skenario percobaan yang menggunakan atribut mata kuliah pilihan. Pada skenario tersebut, tingkat akurasi tertinggi yang dicapai adalah sebesar 75,95%. Hasil ini menunjukkan bahwa mata kuliah pilihan memiliki pengaruh signifikan terhadap masa studi mahasiswa [11].

## 2. METODOLOGI PENELITIAN

### 2.1 Tahapan Penelitian

Tahapan penelitian adalah urutan langkah-langkah yang akan ditempuh dalam suatu penelitian. Setiap tahapan ini menggambarkan model penyelesaian yang digunakan dalam penelitian. Berikut ini adalah tahapan-tahapan yang dapat dijumpai dalam penelitian.:



**Gambar 1.** Tahapan Penelitian

Keterangan:

1. Start  
Proses dimulai dengan langkah awal pengolahan data untuk keperluan analisis atau pemodelan.
2. Pengumpulan Data
  - a. Data yang diperlukan dikumpulkan dari berbagai sumber.
  - b. Data yang terkumpul disimpan dalam bentuk dataset, yang siap untuk diproses lebih lanjut.
3. Data Training dan Data Testing  
Dataset yang terkumpul dibagi menjadi dua bagian:
  - a. Data Training: Digunakan untuk melatih model.



- b. Data Testing: Digunakan untuk menguji performa model.
- 4. Proses Data
  - Sebelum model dibuat, data perlu melalui tahap pemrosesan:
  - Data Cleaning: Langkah ini melibatkan pembersihan data dari nilai yang hilang, duplikasi, atau ketidakkonsistenan agar data lebih siap untuk analisis.
- 5. Pemodelan dengan Algoritma
  - Data yang sudah bersih kemudian diproses menggunakan algoritma tertentu:
  - C4.5: Algoritma pohon keputusan yang digunakan untuk membangun model klasifikasi.
  - KNN (K-Nearest Neighbor): Algoritma berbasis tetangga terdekat yang digunakan untuk klasifikasi atau regresi.
- 6. Visualisasi Data
  - Setelah model selesai dibuat, hasilnya divisualisasikan untuk mempermudah interpretasi:
  - Gambar: Menampilkan hasil dalam bentuk visual, seperti grafik atau diagram.
  - Kinerja: Mengukur performa model berdasarkan metrik evaluasi seperti akurasi, presisi, atau recall.
  - Model: Visualisasi struktur model, seperti pohon keputusan atau aturan klasifikasi.
- 7. Finish
  - Proses selesai setelah hasil analisis atau pemodelan data telah divisualisasikan dan siap untuk disimpulkan atau digunakan lebih lanjut.

Proses ini menggambarkan alur kerja yang terstruktur mulai dari pengumpulan data hingga evaluasi model. Dengan adanya visualisasi, hasil akhir menjadi lebih mudah dipahami dan dapat digunakan untuk pengambilan keputusan.

## 2.2 Data Mining

Data Mining bertujuan untuk mengungkapkan pengetahuan dari kumpulan data, sehingga menghasilkan struktur yang dapat dipahami oleh manusia. Data Mining, atau penambangan data, adalah teknik yang efisien dan cepat dalam mengidentifikasi pengetahuan, pola, dan/atau hubungan antar data secara otomatis. Proses ini merupakan bagian dari analisis dalam penemuan pengetahuan dalam basis data, yang dikenal dengan istilah Knowledge Discovery in Databases (KDD) [4], [5], [12], [13]. Pengetahuan dapat berupa pola data atau hubungan antar data yang valid. Data Mining merupakan gabungan dari berbagai disiplin ilmu komputer, yang didefinisikan sebagai proses untuk menemukan pola-pola baru dari kumpulan data yang sangat besar. Proses ini melibatkan metode-metode yang merupakan kombinasi dari kecerdasan buatan (artificial intelligence), pembelajaran mesin (machine learning), statistika, dan sistem basis data. Data Mining adalah serangkaian langkah untuk menggali informasi bernilai tambah yang sebelumnya hanya diketahui secara manual dari suatu basis data [14]–[16].

## 2.3 Algoritma C5.0

Algoritma C5.0 adalah penyempurnaan dari algoritma ID3 dan C4.5. Dalam proses pengembangan pohon keputusan, informasi yang paling penting akan digunakan sebagai akar untuk node berikutnya. Algoritma ini dimulai dengan menggunakan seluruh data sebagai akar pohon keputusan, sementara atribut yang dipilih berfungsi sebagai pembagi untuk sampel-sampel yang ada. Nilai entropy digunakan sebagai parameter untuk menentukan tingkat heterogenitas (keberagaman) dalam suatu himpunan data. Semakin banyak data yang ada, semakin tinggi pula nilai entropy-nya. Secara matematis, entropy dapat dibandingkan dengan hipotesis berikut [17]–[19]:

$$\text{Entropy (S)} = \sum_i^c - p_i \log_2 p_i \tag{1}$$

Dimana:

S = Himpunan data

c = jumlah nilai / kelas

p<sub>i</sub> = Rasio antara sampel di kelas i dengan jumlah sampel di himpunan data

Gain didefinisikan sebagai ukuran efektivitas suatu atribut dalam mengklasifikasi data. Secara matematis gain dari suatu atribut A, dapat dilihat pada persamaan dibawah [17]–[19].

$$\text{Gain (S,A)} = \text{Entropy (S)} - \sum v \in \text{Values A} \frac{|S_v|}{|S|} * \text{Entropy (S}_v) \tag{2}$$

Dimana:

A = atribut

V = menyatakan suatu nilai yang mungkin untuk atribut A

Values(A) = himpunan nilai-nilai yang mungkin untuk atribut A

|S<sub>v</sub>| = jumlah sampel untuk nilai v

|S| = jumlah seluruh sampel data

Entropy(S<sub>v</sub>) = entropy untuk sampel-sampel yang memiliki nilai v

## 2.4 Algoritma K-Nearest Neighbor

K-Nearest Neighbor (K-NN) adalah salah satu teknik klasifikasi data yang efektif, yang bekerja dengan cara mencari kasus yang paling mirip dengan kasus baru dengan mengukur kedekatan antara keduanya. Proses ini dilakukan dengan menghitung jarak atau kedekatan antara data baru dan data lama, berdasarkan pencocokan atribut atau bobot yang relevan.

K-NN kemudian mengklasifikasikan data baru berdasarkan mayoritas kelas dari tetangga terdekat yang telah dihitung jaraknya. Teknik ini sangat bergantung pada pemilihan nilai K yang optimal untuk menentukan jumlah tetangga terdekat yang digunakan dalam klasifikasi[4]. K-Nearest Neighbor (K-NN) termasuk dalam kelompok instance-based learning. K-NN adalah algoritma pembelajaran berbasis instance, di mana data pelatihan (training set) disimpan, dan klasifikasi untuk data baru yang belum terklasifikasi dilakukan dengan membandingkan data baru tersebut dengan data pelatihan yang paling mirip. Untuk menghitung jarak antara dua titik, yaitu titik pada data pelatihan (x) dan titik pada data pengujian (y), digunakan rumus Euclidean. Rumus tersebut digunakan untuk mengukur kedekatan atau jarak antara titik-titik dalam ruang fitur, yang kemudian digunakan untuk menentukan klasifikasi berdasarkan tetangga terdekat[20]–[22].

$$d(x,y) = \sqrt{\sum_{k=1}^n (X_{\text{training}} - Y_{\text{testing}})^2} \tag{3}$$

Proses perhitungan dimulai dengan menghitung selisih antara setiap pasangan fitur dari X training dan Y testing pada masing-masing dimensi (dari 1 hingga n), kemudian dikuadratkan selisih tersebut. Hasil kuadrat ini dijumlahkan untuk seluruh dimensi, dan akhirnya akar kuadrat dari total jumlah tersebut diambil untuk mendapatkan jarak akhir. Rumus ini sering digunakan dalam algoritma pembelajaran mesin, seperti k-Nearest Neighbors (k-NN), untuk mengukur kemiripan atau kedekatan antara dua data dalam ruang multi-dimensi. Semakin kecil nilai d(x, y), semakin mirip kedua data tersebut. Langkah-langkah dalam menyelesaikan algoritma K-Nearest Neighbor adalah sebagai berikut[5]:

1. Menentukan parameter K.
2. Menghitung kuadrat jarak euclid (queri instance) masing-masing objek terhadap data sampel yang ada.
3. Mengurutkan objek-objek tersebut ke dalam kelompok yang mempunyai jarak eucliden terkecil.
4. Mengumpulkan kategori Y (Klasifikasi Nearest Neighbor).
5. Dengan menggunakan kategori nearest neighbor yang paling mayoritas maka dapat diprediksi nilai queri instance yang telah dihitung.

### 3. HASIL DAN PEMBAHASAN

Analisis merupakan suatu proses yang bertujuan untuk menguraikan sebuah permasalahan menjadi elemen-elemen terkecil, sehingga memudahkan dalam penyelesaian masalah yang dihadapi. Tahap analisis memerlukan pendekatan yang hati-hati dan sistematis untuk menghindari potensi kesalahan yang dapat muncul pada tahap selanjutnya. Proses analisis ini sangat krusial, karena pada tahap ini, masalah dalam sistem yang sedang berjalan didefinisikan dengan jelas, sekaligus melakukan evaluasi terhadap cara kerja sistem yang sedang berlangsung.

Salah satu permasalahan yang terjadi dalam proses pemilihan Lokasi Perumahan adalah masih dilakukan oleh staf khusus, yang berpotensi menimbulkan kesalahan manusia (human error) dalam menentukan lokasi perumahan yang tepat. Oleh karena itu, proses pemilihan ini seharusnya dipermudah dengan menggunakan pendekatan yang lebih sistematis, seperti penerapan algoritma untuk mengidentifikasi pola yang relevan dalam memilih lokasi Perumahan.

Dalam memilih lokasi perumahan yang bermutu dan berkualitas, diperlukan kebijakan yang tepat agar pemilihan lokasi dapat dilakukan dengan lebih efisien. Kendala yang sering dihadapi dalam pemilihan lokasi perumahan adalah kurangnya pemahaman tentang spesifikasi lokasi terbaik yang sesuai untuk dibangun. Untuk mengatasi masalah ini, dilakukan perbandingan antara algoritma C5.0 dan K-Nearest Neighbor (K-NN) guna menganalisis tingkat keakuratan dalam memilih lokasi perumahan, berdasarkan data yang tersedia. Data yang digunakan dalam penelitian ini akan disajikan pada Tabel 1 sebagai berikut:

**Tabel 1.** Sampel Data Penelitian

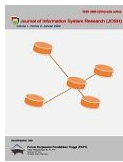
Lokasi	Harga_Tanah (Rp/m <sup>2</sup> )	Aksesibilitas (km)	Fasilitas_Publik (skor)	Tingkat_Kejahatan (skor)	Infrastruktur (skor)	Ketersediaan_Lahan (%)	Preferensi_Konsumen (%)	Kelas_Lokasi
Lokasi_A	1.500.000	2.5	85	20	90	70	80	Dipilih
Lokasi_B	1.200.000	3.0	75	35	80	60	70	Tidak Dipilih
Lokasi_C	2.000.000	1.5	90	15	95	80	85	Dipilih
Lokasi_D	800.000	4.0	65	50	70	50	60	Tidak Dipilih
Lokasi_E	1.000.000	3.5	70	40	75	55	65	Tidak Dipilih
Lokasi_F	2.500.000	1.0	95	10	98	85	90	Dipilih

#### 3.1 Penerapan Algoritma C5.0

Tahapan pertama dalam penelitian ini adalah proses klasifikasi menggunakan algoritma C5.0. Sebelum proses klasifikasi dilakukan, langkah awal yang harus dilakukan adalah preprocessing terhadap data. Tujuan dari preprocessing ini adalah untuk menyetarakan dan menyesuaikan data agar siap digunakan dalam penerapan algoritma. Proses preprocessing yang dilakukan dapat dijelaskan pada Tabel 2 sebagai berikut.:

**Tabel 2.** Preprocessing Harga tanah

No	Data Awal	Hasil Preprocessing
1	500.000 – 800.000	Kurang



No	Data Awal	Hasil Preprocessing
2	1.000.000 – 1.400.000	Cukup
3	1.500.000 – 2.000.000	Baik
4	>2.000.000	Baik Sekali

Pada Tabel 2, dapat dilihat hasil dari proses preprocessing pada data harga Tanah. Selanjutnya, proses preprocessing juga diterapkan pada data lainnya, yang dapat dilihat pada Tabel 3 berikut.:

**Tabel 3.** Preprocessing Aksesibilitas (km)

No	Data Awal	Hasil Preprocessing
1	2.5	Cukup
2	3.0	Cukup
3	1.5	Baik
4	4.0	Kurang

Pada Tabel 3 dapat terlihat hasil dari proses preprocessing data Aksesibilitas, kemudian dilakukan proses preprocessing data terhadap data lainnya seperti tabel berikut:

**Tabel 4.** Preprocessing Fasilitas\_Publik

No	Data Awal	Hasil Preprocessing
1	85	Sangat_Baik
2	75	Baik
3	90	Sangat_Baik
4	65	Cukup

Pada Tabel 4 dapat terlihat hasil dari proses preprocessing data Fasilitas\_Publik, kemudian dilakukan proses preprocessing data terhadap data lainnya seperti tabel berikut:

**Tabel 5.** Tingkat\_Kejahatan

No	Data Awal	Hasil Preprocessing
1	20	Sangat_Baik
2	35	Baik
3	15	Sangat_Baik
4	50	Kurang

Pada Tabel 5 dapat terlihat hasil dari proses preprocessing data Tingkat\_Kejahatan, kemudian dilakukan proses preprocessing data terhadap data lainnya seperti tabel berikut:

**Tabel 6.** Infrastruktur

No	Data Awal	Hasil Preprocessing
1	90	Sangat_Baik
2	80	Baik
3	95	Sangat_Baik
4	70	Cukup

Pada Tabel 6 dapat terlihat hasil dari proses preprocessing data Infrastruktur, kemudian dilakukan proses preprocessing data terhadap data lainnya seperti tabel berikut:

**Tabel 7.** Ketersediaan\_Lahan

No	Data Awal	Hasil Preprocessing
1	70	Baik
2	60	Baik
3	80	Sangat_Baik
4	50	Kurang

Pada Tabel 7 dapat terlihat hasil dari proses preprocessing data Ketersediaan\_Lahan, kemudian dilakukan proses preprocessing data terhadap data lainnya seperti tabel berikut:

**Tabel 8.** Preferensi\_Konsumen

No	Data Awal	Hasil Preprocessing
1	80	Baik
2	70	Baik
3	85	Sangat_Baik

No	Data Awal	Hasil Preprocessing
4	60	Cukup

Setelah diketahui terhadap proses preprocessing data, maka selanjutnya dilakukan proses perubahan ataupun penyesuaian data. Adapun sampe data hasil preprocessing data dapat dilihat pada Tabel 9, sebagai berikut:

**Tabel 9.** Data Hasil Preprocessing

Lokasi	Kelas Lokasi	Harga Tanah (Rp/m <sup>2</sup> )	Aksesibilitas (km)	Fasilitas Publik (skor)	Tingkat Kejahatan (skor)	Infrastruktur (skor)	Ketersediaan Lahan (%)	Preferensi Konsumen (%)
Lokasi_A	Dipilih	Cukup	Cukup	Sangat_Baik	Sangat_Baik	Sangat_Baik	Baik	Baik
Lokasi_B	Tidak	Cukup	Cukup	Baik	Baik	Baik	Baik	Baik
Lokasi_C	Dipilih	Baik	Baik	Sangat_Baik	Sangat_Baik	Sangat_Baik	Baik	Sangat_Baik
Lokasi_D	Tidak	Kurang	Kurang	Cukup	Cukup	Baik	Cukup	Baik
Lokasi_E	Dipilih	Kurang	Kurang	Baik	Baik	Baik	Cukup	Baik
Lokasi_F	Dipilih	Baik	Sangat_Baik	Sangat_Baik	Sangat_Baik	Sangat_Baik	Sangat_Baik	Sangat_Baik

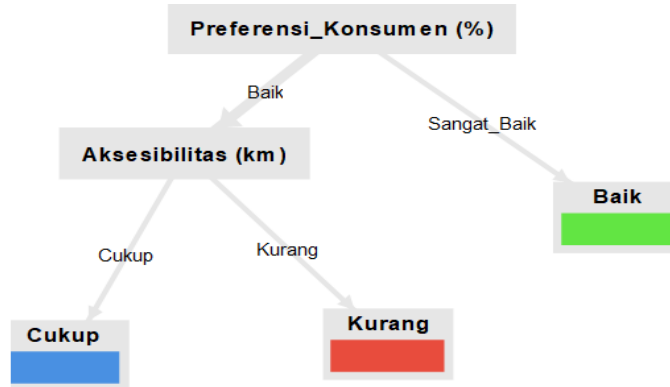
Setelah didapatkan hasil data preprocessing pada tabel 9, maka langkah selanjutnya dapat dilakukan terhadap proses penerapan algoritma C5.0 terhadap penyelesaian proses klasifikasi. Dimana proses pertama sekali dilakukan dengan mencari nilai akar atau node 1. Adapun hasil perhitungan nilai entropy dan gain pada node 1 dapat dilihat pada Tabel 10 sebagai berikut:

**Tabel 10.** Hasil Perhitungan Node I

Node	Atribut	Kelas Lokasi	Jumlah Kasus / Data Set	Dipilih	Tidak Dipilih	Entropy	Gain
1	Harga_Tanah (Rp/m <sup>2</sup> )	Kurang	2	0	3	0	0,666666667
		Cukup	2	1	1	1	
		Baik	2	2	0	0	
		Sangat Baik	0	0	0	0	
		Baik	0	0	0	0	
	Aksesibilitas (km)	Kurang	2	0	2	0	0,333333333
		Cukup	2	1	1	1	
		Baik	1	1	0	0	
		Sangat Baik	1	1	0	0	
		Baik	0	0	0	0	
	Fasilitas_Publik (skor)	Kurang	0	0	0	0	0
		Cukup	1	0	1	0	
		Baik	2	0	2	0	
		Sangat Baik	3	3	0	0	
		Baik	0	0	0	0	
	Tingkat_Kejahatan (skor)	Kurang	0	0	0	0	0
		Cukup	1	0	1	0	
		Baik	2	0	2	0	
		Sangat Baik	3	3	0	0	
		Baik	0	0	0	0	
Infrastruktur (skor)	Kurang	0	0	0	0	0	
	Cukup	0	0	0	0		
	Baik	3	0	3	0		
	Sangat Baik	3	3	0	0		
	Baik	0	0	0	0		
Ketersediaan_Lahan (%)	Kurang	0	0	0	0	0,9182958	
	Cukup	2	0	2	0		
	Baik	3	2	1	3		
	Sangat Baik	1	1	0	0		
	Baik	0	0	0	0		

Nod e	Atribut	Kelas_Lokasi	Jumlah Kasus / Data Set	Dipilih	Tidak Dipilih	Entropy	Gain
	Preferensi_Konsumen (%)						
		Kurang	0	0	0		
		Cukup	0	0	0		
		Baik	4	1	3	0,8112781	-
		Sangat Baik	2	2	0	2	0,540852083

Dari proses yang dilakukan pada tabel 10, maka dapat dilihat bahwasannya atribut Harga\_Tanah (Rp/m<sup>2</sup>) terpilih sebagai akar pada porses. Hal tersebut dikarenakan atribut Harga\_Tanah (Rp/m<sup>2</sup>) merupakan atribut dengan nilai Gain tertinggi. Dari proses yang dilakukan masih terdapat percabangan. Maka proses dilanjutkan dengan menggunakan cara yang sama, atapun hasil akhir pohon keputusan dapat dilihat pada Gambar 2 sebagai berikut:



**Gambar 2.** Hasil Pohon Keputusan

### 3.2 Penerapan Algoritma K-NN

Penerapan Algoritma K-Nearest Neighbor (K-NN) dimulai dengan langkah pertama, yaitu mempersiapkan data yang akan digunakan untuk proses klasifikasi. K-NN adalah algoritma berbasis instance, yang berarti algoritma ini tidak membangun model secara eksplisit, melainkan menggunakan data pelatihan yang ada untuk memprediksi kelas dari data baru berdasarkan kedekatannya dengan data pelatihan yang sudah ada. Pada penerapannya, data pelatihan yang telah disiapkan akan dibandingkan dengan data uji untuk menghitung kedekatannya. Biasanya, jarak yang digunakan untuk menghitung kedekatan ini adalah jarak Euclidean, meskipun ada berbagai jenis pengukuran jarak lainnya yang bisa digunakan sesuai dengan karakteristik data. Dalam proses ini, nilai K (jumlah tetangga terdekat) perlu ditentukan terlebih dahulu. Nilai K yang optimal akan memberikan hasil prediksi yang paling akurat.

Setelah kedekatan antara data uji dan data pelatihan dihitung, algoritma K-NN akan mengklasifikasikan data uji berdasarkan mayoritas kelas dari K tetangga terdekat yang telah dihitung. Proses ini diulang untuk setiap data uji yang ingin diklasifikasikan. Penerapan algoritma K-NN ini sangat bergantung pada pemilihan fitur yang relevan dan penentuan nilai K yang tepat, karena kedua faktor ini dapat mempengaruhi akurasi hasil klasifikasi. Tabel 12 adalah contoh perhitungan pada pembahasan algoritma Nearset Neighbor dengan menggunakan data sampel.

**Tabel 12.** Data Training

Lokasi	Harga_Tanah (Rp/m <sup>2</sup> )	Aksesibilitas (km)	Fasilitas_Publik (skor)	Tingkat_Kejahatan (skor)	Infrastruktur (skor)	Ketersediaan_Lahan (%)	Preferensi_Konsumen (%)	Kelas_Lokasi
Lokasi_A	1.500.000	2.5	85	20	90	70	80	Dipilih
Lokasi_B	1.200.000	3.0	75	35	80	60	70	Tidak Dipilih
Lokasi_C	2.000.000	1.5	90	15	95	80	85	Dipilih
Lokasi_D	800.000	4.0	65	50	70	50	60	Tidak Dipilih
Lokasi_E	1.000.000	3.5	70	40	75	55	65	Tidak Dipilih
Lokasi_F	2.500.000	1.0	95	10	98	85	90	Dipilih

Dari Tabel 12 di atas kita akan mencari data jarak confiden dari semua data diatas kita mendapatkan data yang terdapat pada Gambar 3 Sebagai berikut:

Row No.	Kelas_Lokasi	prediction(Kelas_Lokasi)	confidence(Baik)	confidence(Cukup)	confidence(Kurang)	confidence(Sangat_Baik)	Lokasi	Harga_Tana...	Aksesibilitas ...	Fasilitas_Pu...	Ti
1	Baik	Baik	0.438	0.400	0.162	0	Lokasi_A	150000	2.500	85	20
2	Cukup	Cukup	0.338	0.471	0.191	0	Lokasi_B	120000	3	75	30
3	Baik	Baik	0.465	0.339	0	0.206	Lokasi_C	200000	1.500	90	10
4	Kurang	Cukup	0.210	0.440	0.250	0	Lokasi_D	80000	4	65	50
5	Cukup	Cukup	0.303	0.474	0.224	0	Lokasi_E	100000	3.500	70	40
6	Sangat_Baik	Baik	0.413	0.337	0	0.250	Lokasi_F	250000	1	95	10

**Gambar 3.** Pencarian nilai Confiden

### 3.3 Pembahasan

Setelah dilakukan pengujian atau penerapan terhadap algoritma C5.0 dan K-Nearest Neighbor (K-NN), langkah selanjutnya adalah melakukan perhitungan kinerja atau akurasi terhadap hasil yang diperoleh. Proses ini bertujuan untuk mengevaluasi sejauh mana kedua algoritma tersebut dapat menghasilkan prediksi yang tepat. Hasil dari perhitungan kinerja atau akurasi ini dapat dilihat dalam Gambar 4

accuracy: 100.00%

	true Baik	true Sangat_Baik	class precision
pred. Baik	4	0	100.00%
pred. Sangat_Baik	0	2	100.00%
class recall	100.00%	100.00%	

**Gambar 4.** Hasil Kinerja Algoritma C5.0

Pada Gambar 4 di atas, dapat dilihat hasil kinerja yang diperoleh. Berdasarkan gambar tersebut, kinerja algoritma C5.0 menunjukkan hasil akurasi sebesar **100%**. Selanjutnya, dilakukan proses pengujian terhadap algoritma K-Nearest Neighbor, yang hasilnya dapat dilihat pada Gambar 5, berikut:

accuracy: 66.67%

	true Baik	true Cukup	true Kurang	true Sangat_Baik	class precision
pred. Baik	2	0	0	1	66.67%
pred. Cukup	0	2	1	0	66.67%
pred. Kurang	0	0	0	0	0.00%
pred. Sangat_Baik	0	0	0	0	0.00%
class recall	100.00%	100.00%	0.00%	0.00%	

**Gambar 5.** Hasil Kinerja Algoritma K-Nearest Neighbor

Pada Gambar 5 di atas, terlihat bahwa hasil kinerja algoritma K-Nearest Neighbor mencapai 66,67%. Ini menunjukkan bahwa algoritma C5.0 memiliki kinerja yang lebih baik dalam melakukan proses klasifikasi pada bibit unggul.

## 4. KESIMPULAN

Proses akhir dari penelitian ini adalah penarikan kesimpulan. Berdasarkan hasil penelitian, dapat disimpulkan bahwa penggunaan algoritma data mining dapat mempermudah proses klasifikasi dalam penentuan bibit unggul. Pengujian yang dilakukan menggunakan algoritma C5.0 dan K-Nearest Neighbor telah berhasil dilaksanakan. Hasil pengujian menunjukkan bahwa algoritma C5.0 memiliki kinerja yang lebih baik, dengan tingkat akurasi sebesar **100%**, dibandingkan dengan algoritma K-Nearest Neighbor yang mencapai **66,67%**.

## REFERENCES

- [1] Mesran et al., Data Mining for Decision Support System, 1st ed. Jawa Tengah: Pena Persada Kerta Utama, 2024.
- [2] Y. R. S. Winanjaya, "Implementasi Data Mining Dalam Mengelompokkan Jumlah Penduduk Miskin Berdasarkan Provinsi Menggunakan Algoritma K-Means," KESATRIA J. Penerapan Sist. Inf. (Komputer Manajemen), vol. 2, no. 2, pp. 125-132, 2021.
- [3] I. Nasution, A. P. Windarto, and M. Fauzan, "Penerapan Algoritma K-Means Dalam Pengelompokan Data Penduduk Miskin Menurut Provinsi," vol. 2, no. 2, pp. 76-83, 2020.



- [4] B. Bangun and A. K. Karim, “Pengembalian Data Yang Hilang Pada Dataset Dengan Menggunakan Algoritma K-Nearest Neighbor Imputation Data Mining,” *J. Media Inform. Budidarma*, vol. 8, no. 3, p. 1706, 2024, doi: 10.30865/mib.v8i3.8014.
- [5] A. Karim, S. Esabella, K. Kusmanto, M. Hidayatullah, and S. Suryadi, “Penerapan Data Mining Untuk Pengelompokan Terhadap Kualitas Kinerja Karyawan Dengan Menggunakan Algoritma K-Medoids Clustering,” *J. Media Inform. Budidarma*, vol. 8, no. 2, p. 1001, 2024, doi: 10.30865/mib.v8i2.7445.
- [6] M. S. Pangestu and M. A. Fitriani, “Perbandingan Perhitungan Jarak Euclidean Distance, Manhattan Distance, dan Cosine Similarity dalam Pengelompokan Data Bibit Padi Menggunakan Algoritma K-Means,” *Sainteks*, vol. 19, no. 2, p. 141, 2022, doi: 10.30595/sainteks.v19i2.14495.
- [7] A. D. Adhi Putra, “Analisis Sentimen pada Ulasan pengguna Aplikasi Bibit Dan Bareksa dengan Algoritma KNN,” *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 8, no. 2, pp. 636–646, 2021, doi: 10.35957/jatisi.v8i2.962.
- [8] A. T. Yulindari, Z. Sari, and V. R. S. Nastiti, “Pemetaan Mata Kuliah Yang Berpengaruh Pada Kelulusan Tidak Tepat Waktu Mahasiswa Informatika UMM Menggunakan SOM,” *J. Repos.*, vol. 3, no. 1, pp. 111–120, 2024, doi: 10.22219/repositor.v3i1.31016.
- [9] T. Permana, A. M. Siregar, A. F. N. Masruriyah, and A. R. Juwita, “Perbandingan Hasil Prediksi Kredit Macet pada Koperasi Menggunakan Algoritma KNN dan C5.0,” *Conf. Innov. Appl. Sci. Technol.*, vol. 3, no. 1, pp. 737–746, 2020.
- [10] 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.
- [11] E. Novianto, A. Hermawan, and D. Avianto, “Klasifikasi Algoritma K-Nearest Neighbor, Naive Bayes, Decision Tree Untuk Prediksi Status Kelulusan Mahasiswa S1,” *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 8, no. 2, pp. 146–154, 2023, doi: 10.36341/rabit.v8i2.3434.
- [12] D. P. Indini, Mesran, and Dito Putro Utomo, “Penerapan Data Mining Dalam Pengelompokan Data Reseller di Telkomsel Authorized Partner (TAP) Deli Tua Dengan Algoritma K-Means,” *J. Ilm. Media Sisfo*, vol. 17, no. 2, pp. 189–202, 2023, doi: 10.33998/mediasisfo.2023.17.2.1391.
- [13] M. Mesran, M. Syahrizal, S. Sarwandi, S. Aripin, D. P. Utomo, and A. Karim, “A comparison of the performance of data mining classification algorithms on medical datasets with the application of data normalization,” *AIP Conf. Proc.*, vol. 3048, no. 1, 2024, doi: 10.1063/5.0207994.
- [14] U. R. Amanda and D. P. Utomo, “Penerapan Data Mining Algoritma Hash Based Pada Data Pemesanan Buah Impor Cv. Green Uni Fruit,” *KOMIK (Konferensi Nas. Teknol. Inf. dan Komputer)*, vol. 5, no. 1, 2021.
- [15] B. S. Pranata and D. P. Utomo, “Penerapan Data Mining Algoritma FP-Growth Untuk Persediaan Sparepart Pada Bengkel Motor (Study Kasus Bengkel Sinar Service),” *Bull. Inf. Technol.*, vol. 1, no. 2, pp. 83–91, 2020.
- [16] I. Arfyanti, M. Fahmi, and P. Adytia, “Penerapan Algoritma Decision Tree Untuk Penentuan Pola Penerima Beasiswa KIP Kuliah,” *Build. Informatics, Technol. Sci.*, vol. 4, no. 3, pp. 1196–1201, 2022, doi: 10.47065/bits.v4i3.2275.
- [17] N. H. Harani and F. S. Damayanti, “Implementasi Algoritma C5.0 Untuk Menentukan Pelanggan Potensial Di Kantor Pos Cimahi,” *J. SITECH Sist. Inf. dan Teknol.*, vol. 4, no. 1, pp. 69–76, 2021, doi: 10.24176/sitech.v4i1.6281.
- [18] D. Fitrianih, W. Gunawan, and A. Puspita Sari, “Studi Komparasi Algoritma Klasifikasi C5.0, SVM dan Naive Bayes dengan Studi Kasus Prediksi Banjir Comparative Study of Classification Algorithm between C5.0, SVM and Naive Bayes with Case Study of Flood Prediction,” *Februari*, vol. 21, no. 1, pp. 1–11, 2022.
- [19] R. N. Amalda, N. Millah, and I. Fitria, “Implementasi Algoritma C5.0 Dalam Menganalisa Kelayakan Penerima Keringanan Ukt Mahasiswa Itk,” *Teorema Teor. dan Ris. Mat.*, vol. 7, no. 1, p. 101, 2022, doi: 10.25157/teorema.v7i1.6692.
- [20] A. Putri et al., “Komparasi Algoritma K-NN, Naive Bayes dan SVM untuk Prediksi Kelulusan Mahasiswa Tingkat Akhir,” *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 3, no. 1, pp. 20–26, 2023, doi: 10.57152/malcom.v3i1.610.
- [21] J. Supriyanto, D. Alita, and A. R. Isnain, “Penerapan Algoritma K-Nearest Neighbor (K-NN) Untuk Analisis Sentimen Publik Terhadap Pembelajaran Daring,” *J. Inform. dan Rekayasa Perangkat Lunak*, vol. 4, no. 1, pp. 74–80, 2023, doi: 10.33365/jatika.v4i1.2468.
- [22] R. Sari, “Analisis Sentimen Pada Review Objek Wisata Dunia Fantasi Menggunakan Algoritma K-Nearest Neighbor (K-NN),” *EVOLUSI J. Sains dan Manaj.*, vol. 8, no. 1, pp. 10–17, 2020, doi: 10.31294/evolusi.v8i1.7371.