

PENERAPAN ALGORITMA C4.5 DALAM KLASIFIKASI DOMINASI JENIS KELAMIN PADA DATA PENDUDUK KOTA DEPOK BERDASARKAN KELOMPOK USIA DI TAHUN 2019-2023

Tintania Dwi Yuliani ¹, Devan Suryavijaya ², Rifqy Zanuar Ferdiansyah ³, Nur Muhammad Hibatullah ⁴, Tasya Anggraini ⁵, Rezi Arfian Cahyadi ⁶, Baginda Oloan Lubis ⁷

^{1,2,3,4,5,6} Sistem Informasi, Bina Sarana Informatika

⁷ Rekayasa Perangkat Lunak, Bina Sarana Informatika

Jl. Kramat Raya No. 98 Jakarta, Indonesia

¹titaniadwiulianii@gmail.com

ABSTRAK

Pertumbuhan penduduk di Kota Depok pada periode 2019-2023 menunjukkan peningkatan yang signifikan, menghadirkan berbagai tantangan seperti ketidakseimbangan daya dukung lingkungan serta meningkatnya kebutuhan infrastruktur yang kompleks. Penelitian ini bertujuan untuk mengelompokkan dominasi jenis kelamin berdasarkan kelompok usia dengan memanfaatkan algoritma C4.5. Algoritma ini dipilih karena kemampuannya dalam membangun pohon keputusan dari data kategori maupun kontinu. Data penduduk yang dianalisis diperoleh dari Badan Pusat Statistik (BPS), mencakup kelompok usia, jenis kelamin, dan tahun (2019-2023), dengan bantuan perangkat lunak Orange. Hasil analisis menunjukkan bahwa kelompok anak-anak (0-9 tahun) didominasi oleh laki-laki pada semua tahun, sedangkan kelompok dewasa muda (25-29 tahun) lebih banyak didominasi oleh perempuan. Untuk kelompok remaja (15-19 tahun), distribusi jenis kelamin relatif seimbang. Pohon keputusan yang dihasilkan memiliki tingkat akurasi yang tinggi, dengan kelompok usia dan tahun sebagai atribut utama. Penelitian ini menunjukkan bahwa algoritma C4.5 mampu mengklasifikasikan dominasi jenis kelamin di Kota Depok berdasarkan kelompok usia dan tahun. Kelompok usia tertentu, seperti anak-anak dan remaja, memiliki tingkat entropi rendah yang mencerminkan dominasi jenis kelamin yang lebih jelas, sedangkan kelompok dewasa dan lansia menunjukkan distribusi yang lebih merata. Temuan ini memberikan wawasan strategis untuk mendukung perencanaan kebijakan berbasis data, khususnya dalam pengelolaan populasi dan penyediaan fasilitas yang adil serta merata di Kota Depok.

Kata kunci : Algoritma C4.5, klasifikasi, dominasi jenis kelamin, kelompok usia

1. PENDAHULUAN

Penduduk merupakan salah satu komponen penting dalam pembangunan suatu wilayah. Komposisi penduduk, termasuk distribusi berdasarkan jenis kelamin dan kelompok usia, memiliki dampak signifikan terhadap perencanaan dan pelaksanaan program pembangunan. Kota Depok, sebagai salah satu kota penyangga ibu kota dengan pertumbuhan penduduk yang pesat, memerlukan analisis yang cermat terhadap data demografis untuk mendukung pengambilan keputusan strategis.

Dominasi jenis kelamin dalam suatu populasi dapat memberikan gambaran tentang struktur sosial dan ekonomi masyarakat. Misalnya, dominasi penduduk laki-laki di kelompok usia produktif dapat menunjukkan potensi besar dalam sektor tenaga kerja, sedangkan dominasi perempuan pada kelompok usia tertentu dapat berimplikasi pada kebutuhan layanan sosial seperti kesehatan ibu dan anak.

Analisis data penduduk berdasarkan jenis kelamin dan kelompok usia pada periode 2019-2023 sangat relevan untuk memahami dinamika perubahan populasi di Kota Depok. Dengan memanfaatkan data ini, dapat dilakukan klasifikasi dominasi jenis kelamin pada setiap kelompok usia. Hasil dari klasifikasi ini dapat menjadi dasar untuk merumuskan kebijakan pembangunan yang lebih inklusif dan sesuai dengan kebutuhan masyarakat.

Algoritma C4.5 merupakan program yang bekerja dengan memproses sekumpulan data berlabel

dan menghasilkan pohon keputusan sebagai outputnya. Pohon keputusan yang dihasilkan kemudian diuji menggunakan data uji berlabel yang belum pernah dilihat sebelumnya untuk mengevaluasi kemampuan generalisasinya.[1]

Algoritma C4.5 memiliki kegunaan dalam mengklasifikasikan data, baik yang berbasis angka maupun kategori. Setelah proses klasifikasi selesai, algoritma ini menghasilkan sejumlah aturan pola yang dapat dimanfaatkan untuk melakukan prediksi, menggunakan nilai-nilai yang diambil dari data historis hingga data terbaru. Algoritma C4.5 merupakan pengembangan dari algoritma ID3 dan umumnya digunakan untuk membangun struktur pohon keputusan (decision tree).[2]

Data Mining menjadi salah satu solusi untuk mengatasi permasalahan ini dengan mengekstraksi informasi penting dari data dalam jumlah besar.[3]

Data mining dapat didefinisikan sebagai proses seleksi, eksplorasi dan pemodelan dari sejumlah besar data untuk dapat menemukan suatu pola ataupun kecenderungan yang biasanya tidak disadari keberadaannya [3].

Melalui Data Mining, kita dapat melakukan klasifikasi, prediksi, estimasi, serta menemukan informasi berharga dari kumpulan data yang sangat besar [4].

Salah satu algoritma yang sering digunakan dalam Data Mining adalah algoritma C4.5, yang berfungsi membangun pohon keputusan (decision

tree) berdasarkan informasi yang terdapat dalam data tersebut.[5]

Penerapan Data Mining memberikan berbagai keuntungan, seperti kemampuan untuk melakukan pengklasifikasian, prediksi, dan pengambilan keputusan berdasarkan pola-pola yang ditemukan dalam data. [6]

Menurut Han dan Kamber, klasifikasi merupakan proses untuk membangun model atau fungsi yang mampu menggambarkan serta membedakan kelas-kelas data. Tujuan utamanya adalah agar model tersebut dapat digunakan untuk memprediksi kelas data yang belum diketahui.[7]

Oleh karena itu, algoritma C4.5 sangat ideal untuk mengklasifikasikan dominasi jenis kelamin berdasarkan kelompok usia di Kota Depok. Hasil klasifikasi ini dapat dimanfaatkan untuk merumuskan kebijakan kependudukan yang lebih efektif. Selain berguna dalam proses klasifikasi, pohon keputusan juga dapat digunakan untuk memprediksi tren berdasarkan pola-pola yang terdapat dalam data.[8]

Untuk mengukur tingkat ketidakpastian atau kerancuan dalam data digunakan Entropy. Dalam konteks pembobotan kriteria, entropy dimanfaatkan untuk menilai sejauh mana suatu kriteria mampu memisahkan atau mengklasifikasikan data dengan baik.[9]

Information Gain adalah metode yang digunakan untuk menentukan atribut terbaik dalam proses pembuatan pohon keputusan. [10]

Selain itu, penelitian ini memanfaatkan aplikasi *Orange* untuk mendukung proses klasifikasi dan evaluasi model.[11] Perangkat lunak *Orange* dirancang untuk penambangan data dan pembelajaran mesin. *Orange* mempermudah visualisasi data, seleksi subset, dan evaluasi model, sehingga mendukung pembuatan model prediktif dan mempermudah pengolahan data. Perangkat lunak ini memungkinkan proses klasifikasi data dengan cara yang lebih interaktif dan intuitif.[12]

Penelitian ini bertujuan untuk menganalisis dominasi jenis kelamin di berbagai kelompok usia dan memberikan wawasan yang berguna untuk perencanaan kebijakan di Kota Depok.

2. TINJAUAN PUSTAKA

2.1. Algoritma C4.5 dalam Data Mining

Algoritma C4.5 adalah salah satu metode yang populer dalam data mining, digunakan untuk membangun pohon keputusan berdasarkan atribut yang ada dalam data. Algoritma ini sering dipilih dalam penelitian klasifikasi karena kemampuannya dalam mengelola data berlabel dan menghasilkan output berupa aturan dalam bentuk pohon keputusan.[13]

Penelitian sebelumnya menunjukkan bahwa algoritma ini efektif dalam memprediksi hasil klasifikasi yang berbasis atribut di berbagai bidang, seperti kesehatan, bisnis, dan pendidikan. [1][2][14]

2.2. Klasifikasi dengan Algoritma C4.5

Klasifikasi merupakan salah satu fungsi utama dalam data mining yang bertujuan untuk mengelompokkan data ke dalam kelas-kelas tertentu.[15]

Algoritma C4.5 beroperasi dengan membagi data berdasarkan information gain tertinggi, yang memudahkan dalam memilih atribut terbaik untuk setiap pembagian dalam pohon keputusan. [5]

Metode ini telah banyak diterapkan pada berbagai jenis data untuk meningkatkan akurasi klasifikasi dan mendukung pengambilan keputusan yang berbasis data. [7]

2.3. Pemanfaatan Entropy dan Information Gain

Entropy dan Information Gain adalah konsep kunci dalam algoritma C4.5 untuk menentukan atribut terbaik dalam membagi dataset.[14]

Entropy mengukur tingkat ketidakpastian dalam dataset, sementara Information Gain berfungsi untuk mengurangi ketidakpastian tersebut dengan memilih atribut yang memberikan pemisahan terbaik. [8]

Gabungan kedua konsep ini memastikan bahwa pohon keputusan yang dihasilkan menjadi lebih efisien dan akurat dalam proses klasifikasi.[9]

2.4. Perangkat Lunak dalam Proses Data Mining

Berbagai perangkat lunak digunakan untuk mendukung analisis data mining, salah satunya adalah *Orange*, yang dikenal karena fitur visualisasi datanya dan kemudahan dalam pemodelan klasifikasi. *Orange* menawarkan antarmuka yang intuitif untuk melakukan klasifikasi menggunakan algoritma C4.5 dan mendukung evaluasi model dengan metrik seperti akurasi, presisi, dan recall. [10]

Dalam penelitian ini, *Orange* digunakan untuk menguji model prediksi dominasi jenis kelamin berdasarkan data penduduk.[11]

2.5. Penggunaan Algoritma C4.5 dalam Studi Kependudukan

Algoritma C4.5 telah digunakan dalam berbagai penelitian kependudukan untuk menganalisis dominasi jenis kelamin, distribusi usia, dan pola demografis lainnya. Studi ini menunjukkan bahwa penerapan pohon keputusan dalam analisis kependudukan dapat memberikan wawasan mendalam tentang tren populasi dan mendukung pengambilan keputusan dalam kebijakan publik.[12]

2.6. Dominasi Jenis Kelamin Berdasarkan Kelompok Usia

Penggunaan algoritma C4.5 pada data kependudukan membantu menganalisis dominasi jenis kelamin di berbagai kelompok usia, sehingga dapat memberikan panduan dalam kebijakan publik [2][3].

2.7. Menghitung Entropy untuk Atribut Utama

Entropy adalah ukuran ketidakpastian atau kekacauan dalam data. Dalam konteks pengambilan keputusan, terutama dalam pembelajaran mesin, seperti pohon keputusan, entropy digunakan untuk mengevaluasi tingkat impurity (ketidakmurnian) suatu dataset berdasarkan distribusi label target.

Secara matematis, entropy $E(S)$ untuk dataset S dapat dihitung menggunakan rumus berikut:

$$E(S) = - \sum_{i=1}^n p_i \cdot \log_2(p_i) \quad (1)$$

Penjelasan:

- S : himpunan kasus
- N : jumlah partisi S
- P_i : proporsi dari S_i terhadap S
- (-) : Digunakan karena logaritma probabilitas selalu negatif

2.8. Menghitung Information Gain (G)

Gain (atau Information Gain) adalah ukuran yang digunakan untuk menentukan efektivitas suatu atribut dalam memisahkan dataset menjadi subset yang lebih homogen (lebih murni). Dengan kata lain, Gain mengukur pengurangan Entropy setelah data dipisahkan berdasarkan suatu atribut tertentu.

$$GAIN(S, A) = E(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} \cdot E(S_v) \quad (2)$$

Penjelasan:

- S : himpunan data awal
- A : atribut yang di gunakan untuk membagi data
- Values : nilai-nilai unik dari atribut A
- $|S_v|$: subset data S yang memiliki nilai v pada atribut A
- $\frac{|S_v|}{|S|}$: proporsi data dalam subset S_v terhadap total S
- $E(S_v)$: entropy dari subset S_v

2.9. Confusion Matrix

Confusion Matrix adalah alat yang digunakan untuk mengevaluasi kinerja model klasifikasi dalam machine learning. Confusion matrix memberikan gambaran yang terperinci tentang bagaimana model klasifikasi bekerja, dengan menunjukkan jumlah prediksi yang benar dan salah yang dibuat untuk setiap kelas.

Matriks ini berbentuk tabel $n \times n$, di mana n adalah jumlah kelas dalam dataset. Untuk kasus dua kelas (binary classification), confusion matrix biasanya berbentuk 2×2 , dengan elemen-elemen berikut: Tabel 1. Merupakan Elemen pada Confusion Matrix (Binary Classification)

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Tabel 1. Elemen pada Confusion Matrix (Binary Classification)

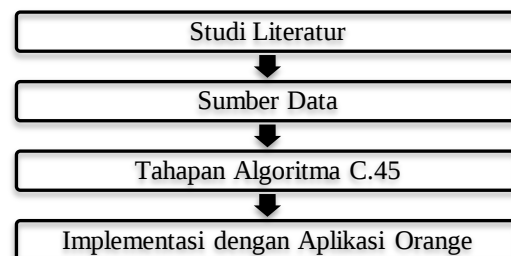
Dari table 1 dapat dijelaskan:

- a. True Positive (TP): Jumlah data positif yang benar-benar diprediksi sebagai positif.
- b. False Positive (FP): Jumlah data negatif yang salah diprediksi sebagai positif (dikenal juga sebagai *type I error*).
- c. True Negative (TN): Jumlah data negatif yang benar-benar diprediksi sebagai negatif.
- d. False Negative (FN): Jumlah data positif yang salah diprediksi sebagai negatif (dikenal juga sebagai *type II error*).

Untuk multi-class classification, confusion matrix diperluas menjadi matriks $n \times n$, di mana n adalah jumlah kelas. Setiap sel (i, j) merepresentasikan jumlah data yang kelas sebenarnya adalah i , tetapi diprediksi sebagai j .

3. METODE PENELITIAN

Berikut merupakan tahapan penelitian pada penelitian ini:



Gambar 1. Tahapan Penelitian

3.1. Studi Literatur

Tinjauan literatur bertujuan untuk mencari solusi terbaik yang telah ada dalam penelitian sebelumnya. Dengan memahami algoritma klasifikasi, Anda tidak hanya dapat mengidentifikasi algoritma mana yang paling relevan untuk masalah Anda tetapi juga memberikan inovasi untuk menyempurnakan metode yang sudah ada. Proses ini merupakan landasan bagi penelitian berkualitas tinggi dan kontribusi yang signifikan terhadap ilmu pengetahuan.

3.2. Sumber Data

Pengumpulan data dilakukan dengan menggunakan dataset yang diperoleh dari Badan Pusat Statistik (BPS). Data uji diambil berdasarkan prediksi jumlah penduduk di Kota Depok, yang dikelompokkan berdasarkan dominasi jenis kelamin dan kelompok usia.

3.3. Tahapan Algoritma C4.5

Algoritma C4.5 merupakan algoritma yang digunakan untuk membangun pohon keputusan (decision tree) berdasarkan data yang tersedia. Pohon keputusan ini berperan dalam membantu proses klasifikasi atau prediksi. Berikut adalah penjelasan mengenai tahapan dalam algoritma C4.5:

- Entropy: Entropy digunakan untuk mengukur tingkat ketidakpastian atau impuritas dalam sebuah dataset. Semakin besar nilai entropy, semakin tinggi ketidakpastian data tersebut.
- Information Gain: Information Gain mengukur pengurangan entropy setelah dataset dibagi berdasarkan nilai atribut tertentu. Atribut dengan nilai Information Gain tertinggi dipilih untuk memisahkan data dalam pohon keputusan.

3.4. Aplikasi Orange

Orange adalah perangkat lunak yang kuat dan mudah digunakan untuk analisis data dan

pembelajaran mesin. Dengan fitur drag-and-drop dan berbagai alat visualisasi interaktif, Orange memungkinkan pengguna dari berbagai latar belakang untuk memanfaatkan data secara maksimal. Cocok digunakan baik untuk pendidikan, penelitian, maupun aplikasi bisnis, Orange adalah pilihan yang tepat untuk eksplorasi data secara intuitif dan cepat.

4. HASIL DAN PEMBAHASAN

4.1. Hasil Sumber Data Jumlah Penduduk Menurut Kelompok Umur dan Jenis Kelamin di Kota Depok (Ribu)

Tabel 1. Sumber data

Kategori	Kelompok Umur (Tahun)	Jumlah Penduduk Menurut Kelompok Umur dan Jenis Kelamin di Kota Depok (Jiwa)							
		Laki-laki	Perempuan	Laki-laki	Perempuan	Laki-laki	Perempuan	Laki-laki	Perempuan
		2.019	2.019	2.020	2.020	2.022	2.022	2.023	2.023
Anak-anak	0-4	111.748	105.472	113.507	107.144	87.995	83.443	86.501	82.541
Anak-anak	5-9	103.348	99.409	105.500	101.885	83.649	80.328	83.968	80.408
Anak-anak	10-14	89.973	85.809	92.691	88.250	89.367	83.330	86.585	82.318
Remaja	15-19	93.394	97.379	95.481	99.384	86.030	81.314	89.555	83.688
Remaja	20-24	106.985	108.033	109.340	110.287	82.705	79.463	83.100	79.829
Dewasa	25-29	106.760	110.477	109.699	113.702	84.801	84.902	82.433	81.621
Dewasa	30-34	111.690	109.919	114.708	112.959	88.376	90.258	86.289	88.166
Dewasa	35-39	106.307	105.833	108.353	107.716	89.354	88.939	87.893	89.302
Dewasa	40-44	100.289	96.398	103.382	99.970	89.339	86.128	87.510	86.064
Dewasa	45-49	83.491	81.369	86.874	85.003	80.316	77.681	83.018	80.196
Dewasa	50-54	66.856	65.634	70.482	69.263	67.825	66.225	70.565	68.670
Dewasa	55-59	51.027	50.554	54.152	53.806	53.235	52.641	56.627	56.362
Dewasa	60-64	34.398	33.463	36.684	35.971	39.425	37.574	40.940	42.458
Lansia	65-69	23.343	20.620	25.291	22.454	25.912	27.664	27.527	29.736
Lansia	70-74	12.033	12.089	13.093	12.866	14.307	16.042	16.633	18.752
Lansia	75+	9.245	13.481	9.881	14.408	10.982	13.799	11.397	14.751
Jumlah		1.210.887	1.195.939	1.249.118	1.235.068	1.071.173	1.052.176	1.080.541	1.064.859

Dari tabel data yang tersedia menunjukkan distribusi dominasi gender berdasarkan kategori usia, kelompok umur, dan tahun, yang memberikan wawasan tentang pola dominasi laki-laki dan perempuan dalam berbagai demografi. Pada kategori anak-anak, dominasi laki-laki sangat jelas di semua kelompok umur dari tahun 2019 hingga 2023. Sementara itu, pada kategori remaja, terdapat pergeseran dominasi gender, di mana perempuan lebih dominan pada kelompok usia 15-19 dan 20-24 pada tahun awal, tetapi laki-laki mulai mendominasi pada tahun-tahun berikutnya. Pada kategori dewasa, dominasi laki-laki terlihat konsisten dalam rentang usia 30-44 tahun, sementara kelompok usia lainnya menunjukkan variasi dominasi antara laki-laki dan perempuan. Untuk kategori lansia, perempuan mendominasi kelompok usia 70 tahun ke atas, sedangkan laki-laki lebih dominan di kelompok usia 65-69 pada tahun awal, dengan pergeseran dominasi ke perempuan di tahun-tahun selanjutnya. Analisis ini menggambarkan pola yang dinamis dalam dominasi gender berdasarkan usia dan waktu.

4.2. Hasil Data di Olah

Tabel 1. Data set setelah diolah

Kategori	Kelompok Usia	Tahun	Dominasi
Anak-anak	0-4	2019	Laki-laki
Anak-anak	0-4	2020	Laki-laki
Anak-anak	0-4	2022	Laki-laki
Anak-anak	0-4	2023	Laki-laki
Anak-anak	05-09	2019	Laki-laki
Anak-anak	05-09	2020	Laki-laki
Anak-anak	05-09	2022	Laki-laki
Anak-anak	05-09	2023	Laki-laki
Anak-anak	10-14	2019	Laki-laki
Anak-anak	10-14	2020	Laki-laki
Anak-anak	10-14	2022	Laki-laki
Anak-anak	10-14	2023	Laki-laki
Remaja	15-19	2019	Perempuan
Remaja	15-19	2020	Perempuan
Remaja	15-19	2022	Laki-Laki
Remaja	15-19	2023	Laki-laki
Remaja	20-24	2019	Perempuan
Remaja	20-24	2020	Perempuan
Remaja	20-24	2022	Laki-laki
Remaja	20-24	2023	Laki-laki
Dewasa	25-29	2019	Perempuan
Dewasa	25-29	2020	Perempuan
Dewasa	25-29	2022	Perempuan
Dewasa	25-29	2023	Laki-laki

Kategori	Kelompok Usia	Tahun	Dominasi
Dewasa	30-34	2019	Laki-laki
Dewasa	30-34	2020	Laki-laki
Dewasa	30-34	2022	Perempuan
Dewasa	30-34	2023	Perempuan
Dewasa	35-39	2019	Laki-laki
Dewasa	35-39	2020	Laki-laki
Dewasa	35-39	2022	Laki-laki
Dewasa	35-39	2023	Perempuan
Dewasa	40-44	2019	Laki-laki
Dewasa	40-44	2020	Laki-laki
Dewasa	40-44	2022	Laki-laki
Dewasa	40-44	2023	Laki-laki
Dewasa	45-49	2019	Laki-laki
Dewasa	45-49	2020	Laki-laki
Dewasa	45-49	2022	Laki-laki
Dewasa	45-49	2023	Laki-laki
Dewasa	50-54	2019	Laki-laki
Dewasa	50-54	2020	Laki-laki
Dewasa	50-54	2022	Laki-laki
Dewasa	50-54	2023	Laki-laki
Dewasa	55-59	2019	Laki-laki
Dewasa	55-59	2020	Laki-laki
Dewasa	55-59	2022	Laki-laki
Dewasa	55-59	2023	Laki-laki
Dewasa	60-64	2019	Laki-laki
Dewasa	60-64	2020	Laki-laki
Dewasa	60-64	2022	Laki-laki
Dewasa	60-64	2023	Perempuan
Lansia	65-69	2019	Laki-laki
Lansia	65-69	2020	Laki-laki
Lansia	65-69	2022	Perempuan
Lansia	65-69	2023	Perempuan
Lansia	70-74	2019	Perempuan
Lansia	70-74	2020	Laki-laki
Lansia	70-74	2022	Perempuan
Lansia	70-74	2023	Perempuan
Lansia	75+	2019	Perempuan
Lansia	75+	2020	Perempuan
Lansia	75+	2022	Perempuan
Lansia	75+	2023	Perempuan

Dari tabel 2 dapat dilihat Data yang telah diolah ini memberikan gambaran mengenai distribusi penduduk Kota Depok berdasarkan kategori usia, kelompok umur, jenis kelamin, dan tahun dari 2019 hingga 2023. Analisis menunjukkan bahwa jumlah penduduk anak-anak (usia 0-14 tahun) mengalami penurunan konsisten selama periode tersebut, dengan dominasi laki-laki yang lebih tinggi dibandingkan perempuan di setiap tahun. Pada kategori remaja (usia 15-24 tahun), terdapat pergeseran jumlah penduduk, terutama dengan dominasi perempuan pada awal periode, yang kemudian berubah menjadi dominasi laki-laki pada tahun-tahun berikutnya. Untuk kategori dewasa (usia 25-64 tahun), variasi jumlah penduduk terlihat cukup signifikan, terutama pada kelompok usia produktif (25-39 tahun), dengan penurunan yang paling mencolok terjadi pada tahun 2022 sebelum stabil kembali pada 2023. Sementara itu, kategori lansia (usia 65 tahun ke atas) menunjukkan peningkatan jumlah penduduk di semua kelompok umur, mencerminkan tren populasi yang menua di wilayah tersebut. Informasi ini memberikan wawasan penting untuk perencanaan kebijakan publik yang lebih tepat sasaran, terutama dalam bidang kesehatan, pendidikan, dan pelayanan sosial, sesuai dengan kebutuhan tiap kelompok usia.

4.3. Hasil Data Set di Proses

Pada tahap ini, dataset yang telah dikumpulkan akan diproses untuk memastikan data siap digunakan dalam analisis atau pembelajaran mesin. Proses ini melibatkan beberapa langkah penting untuk membersihkan, memanipulasi, dan menyiapkan data agar dapat menghasilkan hasil yang akurat dan dapat dipercaya.

Tabel 3. Hasil Perhitungan Entropy dan Gain

Node		Dominasi Laki-laki	Dominasi Perempuan	Jumlah Keduanya	Entropy	Gain
Total		44	20	64	0,89604	
Kel Umur Tahun						0,41025
	2019	11	5	16	0,00000	
	2020	12	4	16	0,00000	
	2022	9	7	16	0,98870	
	2023	10	6	16	0,95443	
Kategori Umur						0,32535
	Anak-anak	12	0	12	0,00000	
	Remaja	4	4	8	0,00000	
	Dewasa	23	9	32	0,85715	
	Lansia	3	9	12	0,81128	
Kelompok Umur						0,73213
	0 - 04	4	0	4	0,00000	
	05 - 09	4	0	4	0,00000	
	10 - 14	4	0	4	0,00000	
	15 - 19	2	2	4	0,00000	
	20 - 24	2	2	4	1,00000	

Node		Dominasi Laki-laki	Dominasi Perempuan	Jumlah Keduanya	Entropy	Gaint
	25 - 29	1	3	4	0,00000	
	30 - 34	2	2	4	0,00000	
	35 - 39	3	1	4	0,81128	
	40 - 44	4	0	4	0,00000	
	45 - 49	4	0	4	0,00000	
	50 - 54	4	0	4	0,00000	
	55 - 59	4	0	4	0,00000	
	60 - 64	3	1	4	0,00000	
	65 - 69	2	2	4	0,00000	
	70 - 74	1	3	4	0,81128	
	75+	0	4	4	0,00000	

Data yang disajikan di atas merupakan hasil olahan yang telah dihitung menggunakan metode entropi dan gain. Pada tabel ini, terlihat distribusi dominasi antara laki-laki dan perempuan dalam beberapa kategori, seperti tahun, kategori umur, dan kelompok umur. Setiap kategori memiliki nilai entropi yang menunjukkan tingkat ketidakpastian atau keragaman data dalam kelompok tersebut, sementara nilai gain mengukur seberapa besar kontribusi setiap kategori dalam mempengaruhi pemisahan data. Data ini memberikan gambaran tentang bagaimana variabel-variabel tersebut berperan dalam menentukan distribusi dominasi berdasarkan jenis kelamin, serta tingkat pengaruhnya dalam analisis data secara keseluruhan.

4.4. Pembahasan Data Set yang Telah Di Proses

a) Anak anak

Hasil entropy nya adalah 0

Karena jika ada data yang hasil nya 0 maka hasil keseluruhan entropy adalah 0

b) Remaja

Hasil entropy adalah 0

Karena jika ada data yang hasil nya 0 maka hasil keseluruhan entropy adalah 0

c) Dewasa

Jumlah laki-laki (dewasa) = 23

Jumlah perempuan (dewasa) = 9

Total dewasa = 32

Dengan demikian penjelasannya :

$$P1 \text{ untuk laki-laki} = \frac{23}{32}$$

$$P2 \text{ untuk perempuan} = \frac{9}{32}$$

Dengan rumus yang sama kita gunakan menghitung di excel maupun manual :

Entropy Dewasa

$$= - \left(\frac{23}{32} \times \log_2 \left(\frac{23}{32} \right) + \frac{9}{32} \times \log_2 \left(\frac{9}{32} \right) \right)$$

Maka hasilnya adalah 0.85715

d) Lansia

Jumlah Laki-laki = 3

Jumlah Perempuan = 9

Total = 12

Dengan demikian penjelasannya :

$$P1 \text{ untuk laki-laki} = \frac{3}{12}$$

$$P2 \text{ untuk perempuan} = \frac{9}{12}$$

Entropy Lansia

$$= - \left(\frac{3}{12} \times \log_2 \left(\frac{3}{12} \right) + \frac{9}{12} \times \log_2 \left(\frac{9}{12} \right) \right)$$

Maka hasilnya adalah 0.81128

e) Contoh menghitung gain total kelompok dewasa

Untuk menghitung gain hal utama yang harus kita hitung adalah total keseluruhan data yang ada di dalam tabel

Dengan entropy awal 0.8960

Dari data :

Anak-anak (12 data, 0 entropi)

Remaja (8 data, 0 entropi)

Dewasa (32 data, 0.85715 entropi)

Lansia (12 data, 0.81128 entropi)

Maka kita akan menghitung total keseluruhan yang ada pada tabel

$$\text{Anak-anak} = \frac{12}{64}$$

$$\text{Remaja} = \frac{8}{64}$$

$$\text{Dewasa} = \frac{32}{64}$$

$$\text{Lansia} = \frac{12}{64}$$

Dengan rumus yang sama kita gunakan menghitung di excel maupun manual

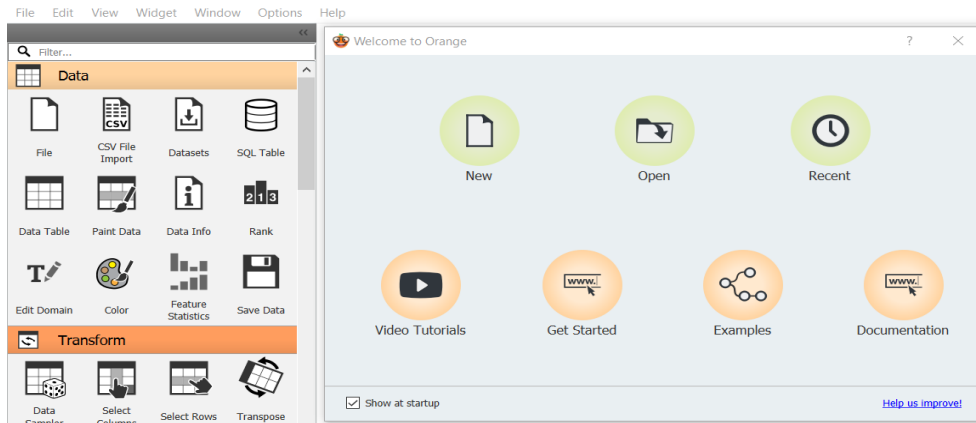
Gain Kategori Umur

$$= 0.8960 - \left(\frac{12}{64} \times 0 + \frac{8}{64} \times 0.85715 + \frac{12}{64} \times 0.81128 \right)$$

Maka hasilnya adalah 0.31535

4.5. Hasil Pembuktian Menggunakan Aplikasi Tools Orange

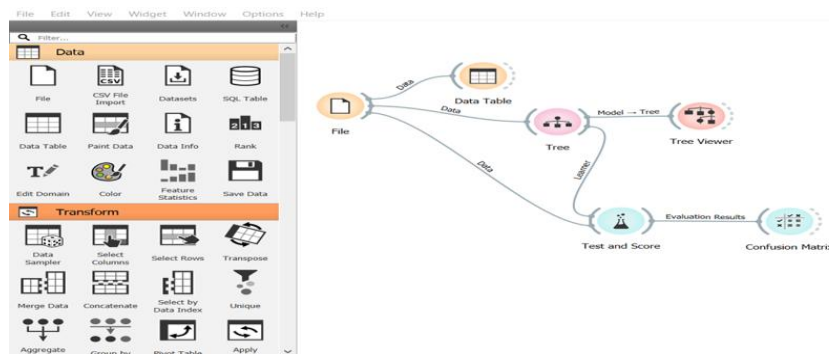
Pada gambar dibawah adalah tampilan halaman utama Orange ketika dibuka.



Gambar 2. Halaman utama aplikasi orange

Pada tahap penelitian ini, yang menggunakan metode klasifikasi Decision Tree, diperlukan penggunaan widget tree. Selain itu, beberapa widget lain juga diperlukan sesuai kebutuhan penelitian. Widget tambahan yang digunakan antara lain widget

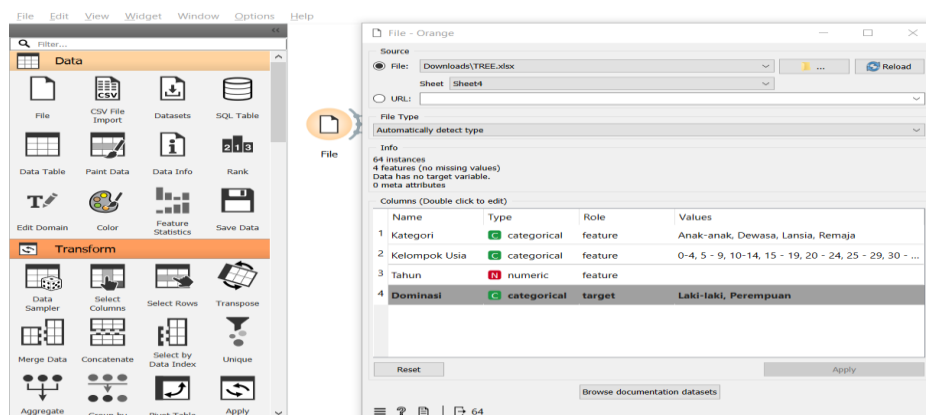
file, treeviewer, data table, test and score, dan confusion matrix. Widget file dihubungkan dengan widget data table, yang berfungsi untuk menampilkan informasi terkait file yang telah diimpor.



Gambar 3. Tools orange

Data pada decision tree memiliki konsep dasar berupa penyajian data dalam bentuk tabel yang terdiri dari atribut dan record. Dalam hal ini, nama produk berfungsi sebagai sampel, kategori sebagai atribut, dan keterangan sebagai atribut target. Selain kategori, atribut lain yang digunakan dalam penelitian ini

meliputi kelompok usia, tahun, dan dominasi. Semua data tersebut disimpan dalam format Excel yang kemudian diimpor ke aplikasi Orange. Setelah data dimasukkan ke dalam aplikasi Orange, langkah selanjutnya adalah memilih kolom yang akan dijadikan target.



Gambar 4. Tampilan tools file

Gambar 4 merupakan tampilan dari tools data table, yang berfungsi untuk melihat dan memeriksa

data dalam format tabel, menyaring dan memfilter data, serta mengedit data secara manual.

Data Table - Orange

Info
64 instances (no missing data)
3 features
Target with 2 values
No meta attributes.

Variables
☒ Show variable labels (if present)
☒ Visualize numeric values
☒ Color by instance classes

Selection
☒ Select full rows

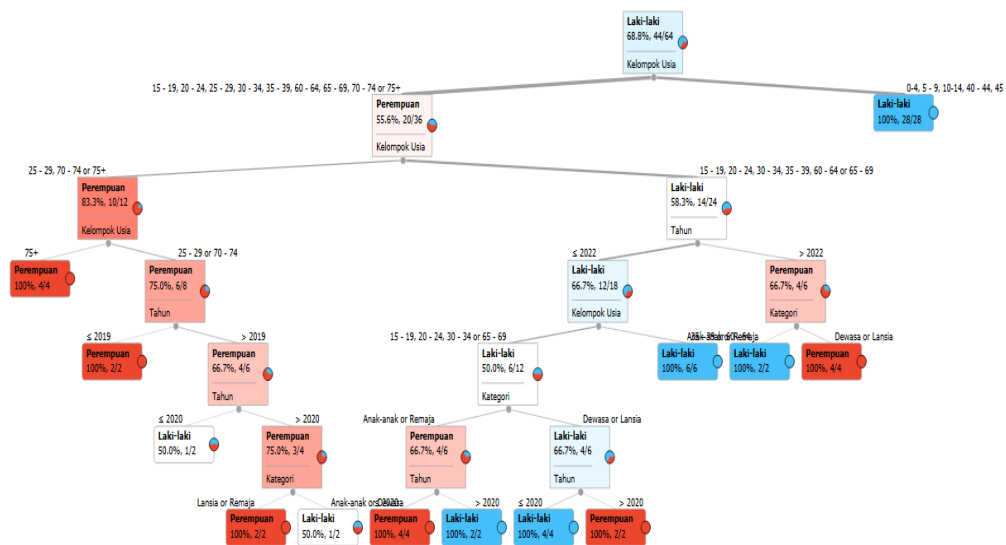
Restore Original Order
Send Automatically

	Dominasi	Kategori	Kelompok Usia	Tahun
1	Laki-laki	Anak-anak	0-4	2019
2	Laki-laki	Anak-anak	0-4	2020
3	Laki-laki	Anak-anak	0-4	2022
4	Laki-laki	Anak-anak	0-4	2023
5	Laki-laki	Anak-anak	5 - 9	2019
6	Laki-laki	Anak-anak	5 - 9	2020
7	Laki-laki	Anak-anak	5 - 9	2022
8	Laki-laki	Anak-anak	5 - 9	2023
9	Laki-laki	Anak-anak	10-14	2019
10	Laki-laki	Anak-anak	10-14	2020
11	Laki-laki	Anak-anak	10-14	2022
12	Laki-laki	Anak-anak	10-14	2023
13	Perempuan	Remaja	15 - 19	2019
14	Perempuan	Remaja	15 - 19	2020
15	Laki-laki	Remaja	15 - 19	2022
16	Laki-laki	Remaja	15 - 19	2023
17	Perempuan	Remaja	20 - 24	2019
18	Perempuan	Remaja	20 - 24	2020
19	Laki-laki	Remaja	20 - 24	2022
20	Laki-laki	Remaja	20 - 24	2023
21	Perempuan	Dewasa	25 - 29	2019
22	Perempuan	Dewasa	25 - 29	2020
23	Perempuan	Dewasa	25 - 29	2022
24	Laki-laki	Dewasa	25 - 29	2023
25	Laki-laki	Dewasa	30 - 34	2019
26	Laki-laki	Dewasa	30 - 34	2020
27	Perempuan	Dewasa	30 - 34	2022
28	Perempuan	Dewasa	30 - 34	2023
29	Laki-laki	Dewasa	35 - 39	2019
30	Laki-laki	Dewasa	35 - 39	2020
31	Laki-laki	Dewasa	35 - 39	2022
32	Perempuan	Dewasa	35 - 39	2023
33	Laki-laki	Dewasa	40 - 44	2019
34	Laki-laki	Dewasa	40 - 44	2020
35	Laki-laki	Dewasa	40 - 44	2022
36	Laki-laki	Dewasa	40 - 44	2023
37	Laki-laki	Dewasa	45 - 49	2019

Gambar 5. Tampilan tools data table

Berdasarkan pengolahan data menggunakan aplikasi Orange, berikut adalah hasil dari fitur tree viewer. Untuk membuat pohon keputusan secara singkat, langkah pertama adalah memilih atribut yang akan digunakan sebagai pembagi data. Setelah itu,

algoritma akan mengkalkulasi information gain atau entropy untuk menentukan atribut mana yang paling optimal untuk membagi data. Proses ini diulang pada setiap cabang pohon hingga mencapai kondisi di mana data dapat diklasifikasikan dengan jelas pada masing-masing daun pohon keputusan.



Gambar 6. Hasil pohon keputusan

Gambar 6 memperlihatkan diagram yang menggambarkan pengaruh atribut-atribut tertentu terhadap hasil klasifikasi. Sebagai contoh, pada kelompok usia 0-4, seluruh data terdiri dari "Laki-laki" (100%, 28/28), sedangkan pada kelompok usia 75+, sebagian besar data adalah "Perempuan." Diagram ini dapat dimanfaatkan untuk menganalisis pola atau melakukan prediksi berdasarkan data yang telah ada sebelumnya.

Test and Score di Orange adalah widget yang digunakan untuk menilai performa model

pembelajaran mesin dengan menghitung metrik seperti akurasi, presisi, recall, F1-score, dan AUC. Widget ini memungkinkan penerapan validasi silang (cross-validation) untuk evaluasi yang lebih akurat, serta mendukung perbandingan antara beberapa model untuk membantu pengguna memilih model terbaik. Dengan Test and Score, pengguna juga dapat melihat hasil uji statistik untuk memastikan apakah perbedaan antar model signifikan.

Tabel 4. Hasil test and score

Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0,785	0,766	0,755	0,756	0,766	0,423

Confusion matrix adalah tabel yang digunakan untuk mengevaluasi kinerja model klasifikasi dengan membandingkan hasil prediksi model terhadap data sebenarnya. Tabel ini menampilkan jumlah true positives (TP), true negatives (TN), false positives (FP), dan false negatives (FN). Dari data ini, berbagai metrik seperti akurasi, presisi, recall, dan F1-score dapat dihitung, memberikan gambaran menyeluruh tentang kinerja model klasifikasi.

Tabel 5. Hasil confusion matrix

		Predicted	
Actual		Laki-laki	Perempuan
	Laki-laki	39	5
	Perempuan	10	10
	Σ	49	15

5. KESIMPULAN DAN SARAN

Hasil penelitian ini menunjukkan bahwa algoritma C4.5 berhasil mengklasifikasikan dominasi jenis kelamin di Kota Depok berdasarkan kelompok umur dan tahun. Kelompok usia tertentu, seperti anak-anak dan remaja, memiliki entropi rendah, yang mencerminkan dominasi jenis kelamin yang lebih jelas, sementara kategori dewasa dan lansia menunjukkan distribusi yang lebih seimbang. Penerapan algoritma C4.5 dalam data mining terbukti efektif untuk analisis demografi yang mendalam, serta dapat memberikan kontribusi dalam merancang kebijakan kependudukan yang lebih tepat. Penelitian di masa mendatang disarankan untuk menambahkan atribut lain, seperti tingkat pendidikan atau jenis pekerjaan, guna meningkatkan akurasi klasifikasi. Selain itu, diperlukan penelitian lebih lanjut untuk menganalisis tren dominasi jenis kelamin di daerah lain serta hubungannya dengan faktor sosial lainnya.

DAFTAR PUSTAKA

- [1] L. Bachtiar and M. Mahradianur, “Analisis Data Mining Menggunakan Metode Algoritma C4.5 Menentukan Penerima Bantuan Langsung Tunai,” *J. Inform.*, vol. 10, no. 1, pp. 28–36, 2023, doi: 10.31294/inf.v10i1.15115.
- [2] M. A. C, “Analisis Kepuasan Pelanggan PT Wook Global Technology,” pp. 160–168, 2023.
- [3] D. Jatmika and G. Taufik, “Penerapan Algoritma C4.5 Untuk Klasifikasi Keberhasilan Pengiriman Barang,” *Invotek Polbeng*, vol. 8, no. 1, pp. 12–26, 2021.
- [4] B. O. Lubis, D. Oscar, F. W. Fibriany, B. Santoso, Jefa, and A. Rusman, “Classification of tomato leaf disease and combination extraction features using K-NN algorithm,” *AIP Conf. Proc.*, vol. 2714, no. November 2021, 2023, doi: 10.1063/1.5000000.
- [5] A. P. Sari, G. Rahmadini, H. Carlina, M. I. Ramadan, and Z. E. Pradani, “Analisis Masalah Kependudukan Di Indonesia.,” *J. Econ. Educ.*, vol. 2, no. 1, pp. 29–37., 2023.
- [6] D. M. Musa *et al.*, “Penerapan Data Mining Untuk Klasifikasi Data Penjualan Pakan Ternak Terlaris Dengan Algoritma C4.5,” *J. Teknol. Inform. dan Komput.*, vol. 10, no. 1, pp. 168–182, 2024, doi: 10.37012/jtik.v10i1.1985.
- [7] C. I. Ifaru *et al.*, “Penentuan Sparepart Genset Paling Sering Digunakan Pada Operator Indosat Ooredoo Hutchison Dengan Algoritma C4.5,” *J. Teknol. Sist. Inf.*, vol. 5, no. 1, pp. 300–309, 2024, doi: 10.35957/jtsi.v5i1.7481.
- [8] A. P. Sari, G. Rahmadini, H. Carlina, M. I. Ramadan, and Z. E. Pradani, “Analisis Masalah Kependudukan Di Indonesia.,” *J. Econ. Educ.*, vol. 2, no. 1, pp. 29–37., 2023.