INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17, No. 1, Mei 2025
eISSN : 2580-3042
pISSN : 1979-0694

IMPLEMENTASI ALGORITMA K-NEAREST NEIGHBOR DALAM MEMPREDIKSI JUMLAH KASUS HIV BERDASARKAN KELOMPOK UMUR DI JAWA BARAT

Gilang Saputra[1], Anita Diana[2*]
[1,2]Sistem Informasi, Fakultas Teknologi Informasi, Universitas Budi Luhur, Jakarta
[1,2]Jl. Ciledug Raya, Petukangan Utara, Jakarta Selatan, Jakarta 12260
e-mail : [1]gilang.26saputra@gmail.com, [2*]anita.diana@budiluhur.ac.id

ABSTRAK

HIV (Human Immunodeficiency Virus) adalah virus yang menyerang sistem imun tubuh manusia, sehingga membuat tubuh menjadi lebih mudah terserang berbagai jenis infeksi. Penelitian ini bertujuan untuk memprediksi jumlah kasus HIV berdasarkan kelompok umur di Provinsi Jawa Barat menggunakan algoritma K-Nearest Neighbor (KNN). Metode CRISP-DM diterapkan untuk mengelola dan menganalisis data melalui tahapan pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan implementasi. Pengujian dilakukan dengan membagi data menjadi tiga rasio (80:20, 70:30, 60:40) untuk melatih dan menguji model. Hasil evaluasi menunjukkan bahwa rasio 60:40 menghasilkan nilai yang lebih tinggi dengan membandingkan nilai k=3 dengan k=5, sehingga menjadikan split data tersebut menjadi nilai dengan performa terbaik dengan nilai akurasi sebesar 31%, presisi 33%, recall 33%, dan F1-score 31%. Penggunaan algoritma K-Nearest Neighbors dapat membantu dalam mengidentifikasi kelompok risiko tinggi untuk pola penyebaran HIV dan dapat membantu lembaga terkait dalam merancang strategi pencegahan yang lebih efektif. Upaya pencegahan dan pengobatan yang lebih baik sangat diperlukan untuk mengatasi masalah ini.

Kata kunci *: HIV, K-Nearest Neighbor, Prediksi, Jawa Barat*

ABSTRACT

HIV (Human Immunodeficiency Virus) is a virus that attacks the human immune system, making the body more susceptible to various types of infections. This study aims to predict the number of HIV cases based on age groups in West Java using the K-Nearest Neighbor (KNN) algorithm. The CRISP-DM methodology is applied to manage and analyze data through the stages of business understanding, data understanding, data preparation, modeling, evaluation, and implementation. Testing is conducted by splitting the data into three ratios (80:20, 70:30, 60:40) to train and test the model. The evaluation results show that the 60:40 ratio produces higher values when comparing k=3 with k=5, making this data split the best-performing one with an accuracy of 31%, precision of 33%, recall of 33%, and an F1-score of 31%. The use of the K-Nearest Neighbors algorithm can help identify high-risk groups for HIV transmission patterns and assist relevant institutions in designing more effective prevention strategies. Better prevention and treatment efforts are essential to addressing this issue.

Kata kunci *: HIV, K-Nearest Neighbor, Prediction, West Java*

1. PENDAHULUAN

Provinsi Jawa Barat merupakan salah satu provinsi dengan jumlah penduduk terpadat di Indonesia. Jumlah penduduk Jawa Barat pada tahun 2023 telah mencapai 49,86 juta jiwa. Jumlah ini menjadikan Jawa Barat sebagai provinsi dengan jumlah penduduk terbesar di Indonesia. Pada tahun 2023 penduduk Jawa Barat bertambah sebanyak 454 ribu jiwa (Badan Pusat Statistik Provinsi Jawa Barat, 2024). Karena menjadi provinsi dengan populasi

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17, No. 1, Mei 2025
eISSN : 2580-3042
pISSN : 1979-0694

terbesar di Indonesia, masalah kesehatan masyarakat semakin menjadi perhatian, terutama penyebaran HIV. Menurut laporan Dinas Kesehatan Jawa Barat, pada tahun 2021, terdapat lebih dari 16.000 kasus HIV yang tercatat di provinsi ini, dengan sebagian besar kasus terjadi di kota-kota besar seperti Bandung dan Bekasi. Angka ini menunjukkan tren peningkatan yang signifikan dibandingkan tahun-tahun sebelumnya, yang mencerminkan perluasan penyebaran virus dalam populasi yang lebih luas (Badan Pusat Statistik Provinsi Jawa Barat, 2024).

HIV (Human Immunodeficiency Virus) adalah virus yang menyebabkan sistem kekebalan tubuh menjadi lemah dengan menginfeksi dan merusak sel darah putih. Penyakit ini sangat serius dan perlu diobati segera karena perkembangannya sangat cepat. Di Jawa Barat, tingkat kasus HIV masih sangat tinggi. HIV juga dapat mengganggu sistem kekebalan tubuh, menyebabkan reaksi alergi yang parah di tubuh, terutama pada saluran pernapasan. Hal ini berkontribusi pada melemahnya sistem kekebalan tubuh individu HIV-positif (Samudra et al., 2022).

HIV menjadi salah satu masalah kesehatan global yang telah menelan banyak korban jiwa. Kasus HIV diJawa Barat dari tahun 2019 hingga 2023 menunjukkan tren yang signifikan. Menurut data yang dirilis Kementrian Kesehatan RI per tanggal 27 Agustus 2019, penderita HIV di Jawa Barat berjumlah cukup banyak, yaitu 36.853 orang (Nurani et al., 2022).

Pada tahun 2020, menurut data berdasarkan Dinas Kesehatan Jawa Barat yang dipublikasikan melalui opendata.jabarprov.go.id penderita HIV di Jawa Barat berjumlah 4.938. Pada tahun 2021, tercatat jumlah kasus 4.758 orang positif HIV. Kasus HIV berdasarkan kelompok umur dengan kasus yang terbanyak berada di umur 25- 49 tahun sebesar 59,35 %. Sementara, jumlah kumulatif HIV di Jawa Barat sampai Oktober 2022 sebanyak 57.914 (Herawati et al., 2023).

Sementara itu, pada tahun 2023, selama periode Januari-September ditemukan 7.383 kasus positif dengan kumulatif pemeriksaan HIV hingga 700.938 orang (Ruhyani, 2023).

Nilai rata-rata jumlah kasus tiap tahun adalah 6.397,2 dalam 5 Tahun Terakhir. Data tersebut dipublikasikan Dinas Kesehatan Provinsi Jawa Barat melalui opendata.jabarprov.go.id. Daerah dengan kepadatan penduduk yang tinggi cenderung memiliki tingkat interaksi sosial yang lebih tinggi, yang dapat meningkatkan risiko penularan HIV. Kontak langsung yang lebih sering di antara individu dapat mempercepat penyebaran virus. Dalam situasi ini, diperlukan strategi yang lebih efektif untuk mengelola dan mengurangi jumlah kasus HIV, sehingga memungkinkan konseling yang lebih cepat dan akurat. Mengingat jumlah kasus HIV yang diprediksi, diharapkan hal ini dapat membantu lembaga pemerintah dan organisasi terkait dalam mengembangkan strategi pencegahan dan tanggapan yang lebih efektif, khususnya di wilayah Jawa Barat (Taniwan et al., 2024).

Prediksi jumlah kasus HIV berdasarkan kelompok umur di provinsi Jawa Barat sangatlah penting sebagai langkah preventif dan pengambilan kebijakan yang lebih tepat. Analisis prediktif memungkinkan para pembuat kebijakan dan instansi kesehatan untuk merencanakan langkah-langkah strategis dalam penanggulangan HIV pada tahun 2019 sampai 2023. Pada tahun 2022, penderita terbanyak pada rentang usia 25 - 49 tahun yakni hampir 70% (2.614). Usia 20 hingga 24 tahun sebanyak 18,4 % (690), usia di atas 50 tahun sekitar 6 % (229), remaja usia 15-19 tahun di urutan berikutnya 3,4 % (126), dan sisanya anak/balita. Jika melihat dari data penderita mayoritas, usia 24 - 49 tahun, mereka terbanyak berada di wilayah kota besar. Di antaranya Kota Bandung (276), Kabupaten Bogor (270), Kota Bekasi (250), Kabupaten Indramayu (188), dan Bekasi (157) (Teguh, 2022).

Dalam publikasinya (Siahaan et al., 2020) mengungkapkan pada klasifikasi penyakit yang dapat ditransfer jenis kelamin pria, K-Nearest Neighbor dapat digunakan untuk mengakhiri dengan menghitung klasifikasi penyakit yang dapat ditransfer jenis kelamin pria antara data, mencari nilai validitas, dan penyesuaian berat badan. Keakuratan sistem diukur berdasarkan tes yang dilakukan pada tes yang ditentukan. Pada penelitian sebelumnya (Sari et al., 2023) menyatakan bahwa dari seluruh informasi data ODHA yang tersedia, diperoleh empat klaster penyebaran HIV/AIDS dengan rincian sebagai berikut: Klaster 1 dikategorikan sangat tinggi dengan 1 data (3%), Klaster 2 kategori sedang terdiri dari 2 data (7%), Klaster 3 termasuk kategori tinggi dengan 5 data (17%), dan klaster 0 tergolong rendah dengan 22 data (73%). Dalam penelitian yang dilakukan oleh Munawar dan Purnamasari, menjelaskan bahwa hasil klasterisasi tingkat penyebaran HIV berdasarkan kelompok usia menggunakan algoritma K-Means Clustering menghasikan dua klaster. Klaster_0 dikategorikan sebagai tingkat penyebaran rendah, sementara klaster_1 termasuk dalam kategori tinggi. Kelompok dengan penyebaran tinggi ditemukan pada usia 0-4 tahun di Kabupaten Subang dan Kota Cirebon, usia 20-24 tahun di Kabupaten Karawang dan Kota Cimahi, serta usia 25-49 tahun di sejumlah wilayah seperti Kabupaten Bogor, Cianjur, Bandung,

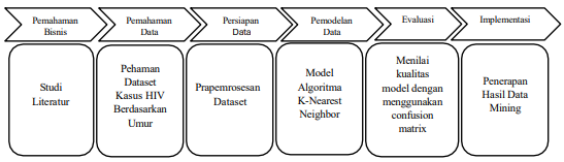
INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17, No. 1, Mei 2025
eISSN : 2580-3042
pISSN : 1979-0694

Garut, Cirebon, Indramayu, Subang, Purwakarta, Karawang Bekasi, dan juga di Kota Bogor, Bandung, Cirebon, Bekasi, Depok, serta Cimahi (Kodratul Munawar & Irma Purnamasari, 2023). Pada publikasinya (Noor et al., 2021) menyatakan bahwa hasil penelitian menunjukkan terbentuknya tiga klaster dengan karakteristik masing-masing yang diharapkan dapat menjadi perhatian pihak terkait dalam upaya menurunkan jumlah ODHA di Kabupaten Banjar. Pada penelitian sebelumnya oleh (Rosida & Wijaya, 2023) dijelaskan bahwa proses pengujian data dilakukan menggunakan tools RapidMiner untuk klasterisasi HIV/AIDS di wilayah Jawa Barat dengan menerapkan algoritma K-Means Clustering. Setelah dilakukan perhitungan nilai DBI pada variasi jumlah klaster dari k=2 hingga k=20, ditemukan bahwa nilai DBI terendah, yaitu 0,414, diperoleh pada k=2. Karena nilai ini paling mendekati 0 dibandingkan dengan nilai k lainnya, maka dapat disimpulkan bahwa klaster terbaik dalam penelitian adalah k=2.

Tujuan penelitian ini adalah untuk menganalisis penyebaran kasus HIV di Jawa Barat agar dapat mengidentifikasi kelompok umur yang berisiko tinggi dan merancang strategi pencegahan yang lebih efektif dengan menggunakan algoritma K-Nearest Neighbors (KNN). Karena metode ini sederhana dan mudah dipahami. KNN bekerja dengan membandingkan data baru dengan data yang sudah ada untuk menemukan pola-pola yang mirip.

2. METODE PENELITIAN

CRISP-DM (Proses Standar Lintas Industri untuk Pemrosesan Data) adalah standar pemrosesan penambangan data yang secara jelas dan akurat mentransfer data terkini ke setiap komponen yang terstruktur dan ditentukan. Selain penggunaan model dalam proses data mining, pemilihan algoritma memiliki dampak yang signifikan terhadap kinerja metode data mining (Hasanah et al., 2021).



Gambar 1. Kerangka Penelitian

Pada gambar 1. dapat dilihat tahapan yang digunakan dalam penelitian dengan menggunakan CRISP-DM. Berikut penjelasannya.

a) Pemahaman Bisnis
Pada tahapan pertama yang dilakukan adalah memahami topik dan tujuan yang ingin dicapai dengan melihat studi literatur penelitian menggunakan klasifikasi KNN yang pernah dibuat sebelumnya. Penelitian ini bertujuan mengukur tingkat akurasi model algoritma klasifikasi KNN dalam memprediksi kasus HIV di Provinsi Jawa Barat.

b) Pemahaman Data
Tahapan kedua merupakan tahapan penting dalam memahami kebutuhan data yang akan diolah. Data yang akan diolah adalah data dinas kesehatan tentang HIV tahun 2019 sampai 2023 dengan memuat 1617 baris dan 10 kolom yang didapat dari website yaitu opendata.jabarprov.go.id.

c) Persiapan Data
Tahapan ketiga adalah tahapan pengolahan data untuk menghasilkan data yang siap untuk pemodelan data. Data akan diolah dengan cara membersihkan data yang tidak relevan dan mengatasi data yang hilang.

d) Pemodelan
Tahapan keempat yang dilakukan adalah proses atau metode pembuatan model matematika atau statistik yang dapat digunakan untuk memahami, menganalisis, atau memprediksi perilaku atau pola data. Pemodelan dilakukan dengan mengidentifikasi teknik data mining yang digunakan, mengidentifikasi tools data mining, menentukan algoritma data mining yang digunakan, dan menetapkan parameter model dengan nilai optimal. Proses pengambilan sampel data penelitian ini digunakan dari website Google Colab. Pada tahap ini, model yang digunakan adalah algoritma model klasifikasi K-Nearest Neighbor yang digunakan untuk mengklasifikasikan kasus HIV di Jawa Barat.

e) Evaluasi
Tahapan kelima yang dilakukan adalah proses pengukuran dan mengevaluasi kualitas data untuk mengetahui keandalan, keakuratan, dan kelengkapan data. Evaluasi dilakukan dengan model standar menggunakan data pelatihan dan data pengujian menggunakan confusion matrix. Produk ini memberikan accuracy, precission, recall dan f1-score untuk semua kondisi.

f) Impelementasi

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17, No. 1, Mei 2025
eISSN : 2580-3042
pISSN : 1979-0694

Tahap ini melakukan pembuatan laporan yang didefinisikan sebagai aplikasi data mining dari hasil pemodelan. Hasil pemodelan tersebut memberikan masukan kepada pemerintah Jawa Barat dan dinas sosial Jawa Barat.

2.1 Pengumpulan Data

Data yang digunakan dalam penelitian ini adalah data sekunder. Data tersebut diambil dari situs resmi yang dikelola oleh Dinas Kesehatan Jawa Barat yaitu opendata.jabarprov.go.id. Data yang digunakan adalah data jumlah kasus HIV di Jawa Barat pada tahun 2019 sampai 2023 dengan memuat 1617 baris dan 10 kolom. Berikut Dataset yang Jumlah Kasus HIV Berdasarkan Kelompok Umur di Jawa Barat pada gambar 2.

id	nama_provinsi	nama_kabupaten_kota	kelompok_umur	jenis_kelamin	jumlah_kasus	satuan	tahun
1220	JAWA BARAT	KOTA BANDUNG	25-49	LAKI-LAKI	633	ORANG	2022
1543	JAWA BARAT	KOTA BANDUNG	25-49	LAKI-LAKI	567	ORANG	2023
1244	JAWA BARAT	KOTA BEKASI	25-49	LAKI-LAKI	468	ORANG	2022
1567	JAWA BARAT	KOTA BEKASI	25-49	LAKI-LAKI	450	ORANG	2023
1303	JAWA BARAT	KABUPATEN BOGOR	25-49	LAKI-LAKI	407	ORANG	2023
1483	JAWA BARAT	KABUPATEN BEKASI	25-49	LAKI-LAKI	392	ORANG	2023
980	JAWA BARAT	KABUPATEN BOGOR	25-49	LAKI-LAKI	371	ORANG	2022
1160	JAWA BARAT	KABUPATEN BEKASI	25-49	LAKI-LAKI	293	ORANG	2022
1471	JAWA BARAT	KABUPATEN KARAWANG	25-49	LAKI-LAKI	257	ORANG	2023
225	JAWA BARAT	KOTA BOGOR	25-49	LAKI-LAKI	254	ORANG	2019

Gambar 2. Tabel Dataset

2.2 Data Pra-processing

Pada tahap ini dilakukan proses persiapan dataset penelitian yang bersumber dari "Jumlah Kasus HIV Berdasarkan Kelompok Umur di Jawa Barat. Tahap persiapan data melalui beberapa tahapan seperti menghilangkan data noise, menghapus missing value, dan melakukan seleksi data. Atribut yang digunakan yaitu kode provinsi, nama provinsi, kode kabupaten kota, nama kabupaten kota, jumlah kasus, kelompok umur, satuan, dan tahun.

2.3 Pemilihan Dataset

Pemilihan dataset dilakukan berdasarkan masalah dan tujuan yang telah ditentukan untuk menghasilkan model terbaik. Variabel yang digunakan antara lain.

Nama variabel	Penjelasan
Nama_Kabupaten_Kota	Menyatakan lingkup data berasal dari setiap kabupaten dan kota di Provinsi Jawa Barat sesuai penamaan BPS merujuk pada aturan Peraturan Badan Pusat Statistik Nomor 3 Tahun 2019 dengan tipe data teks.
Jumlah_Kasus	Menyatakan jumlah kasus hiv dengan tipe data numerik.
Kelompok_Umur	Menyatakan kategori kelompok umur dengan tipe data teks.
Tahun	Menyatakan tahun produksi data dengan tipe data numerik.

Gambar 3. Pemilihan Variabel

2.4 Pembersihan Dataset

Pada tahap ini menggunakan Microsoft Excel untuk melakukan pembesihan data atau variabel dengan menggunakan fitur delete. Variabel yang dihapus adalah id, kode provinsi, nama provinsi, kode kabupaten kota, dan satuan. Berikut alasannya :

a. Id : Tidak memiliki nilai prediktif, bukan sebagai variabel yang memengaruhi prediksi jumlah kasus HIV.
b. Kode Provinsi : Memuat informasi yang sama sehingga tidak memberikan variasi atau pengaruh untuk analisis.
c. Nama provinsi : Karena semua data berasal dari provinsi yang sama, informasi ini tidak berkontribusi pada variasi data.
d. Kode Kabupaten Kota: Karena analisis biasanya lebih mudah dipahami dengan nama kabupaten/kota, kode ini tidak relevan untuk proses lebih lanjut.
e. Jenis Kelamin : Dihapus karena penelitian hanya mempertimbangkan distribusi kasus berdasarkan kelompok umur tanpa membedakan jenis kelamin.
f. Satuan : Seluruh dataset menggunakan satuan yang sama yaitu"orang", sehingga variabel ini tidak menambah nilai dalam analisis atau pemodelan.

2.5 Algoritma K-Nearest Neighbor

K-Nearest Neighbor (KNN) adalah algoritma klasifikasi dalam data mining yang memanfaatkan data terdekat untuk memprediksi data baru maupun data yang tidak diketahui (data uji). Algoritma ini bekerja dengan cara mencari sejumlah tetangga terdekat dari data uji, lalu menentukan kelasnya berdasarkan mayoritas kelas dari data latih tersebut.

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17, No. 1, Mei 2025
eISSN : 2580-3042
pISSN : 1979-0694

Algoritma KNN merupakan metode yang digunakan untuk mengklasifikasikan data berdasarkan jarak terpendek terhadap objek data. Pemilihan nilai K yang optimal sangat bergantung pada karakteristik data yang dimiliki. Kelebihan algoritma KNN mudah dalam mengolah data latih dan data uji dalam jumlah besar dengan cara yang sederhana (Cholil et al., 2021).

2.6 Euclidean Distance

Euclidean distance merupakan metode perhitungan untuk mengukur jarak dua titik dalam euclidean space yang berkaitan dengan hubungan antara sudut dan jarak. Dalam matematika, euclidean distance digunakan untuk mengukur dua titik dalam satu dimensi dengan hasil yang menyerupai perhitungan menggunakan pythagoras (Miftahuddin et al., 2020). Rumus Euclidean Distance dapat dilihat pada formula (1) berikut.

$$d = \sqrt{(x-x)^2+(y-y)^2} \text{.....(1)}$$

Keterangan :
d = Jarak
x = Data Latih
y = Data Uji

2.7 Penentuan Data Latih dan Data Uji

Pemisahan data yang dilakukan pada penelitian ini menggunakan perbandingan 80% data latih : 20% data uji, 70% data latih : 30% data uji, dan 60% data latih : 40% data uji.

1. Rasio 80:20: Rasio ini sering digunakan karena menyediakan cukup banyak data untuk melatih model agar belajar pola dengan baik, sambil tetap menyisihkan data uji yang cukup untuk mengevaluasi performa model.
2. Rasio 70:30: Rasio ini menyediakan lebih banyak data untuk evaluasi dibandingkan 80:20. Cocok jika dataset besar atau jika ingin hasil evaluasi yang lebih komprehensif pada data yang tidak dilihat model sebelumnya.

Rasio 60:40: Rasio ini digunakan jika dataset relatif kecil atau jika evaluasi lebih mendalam terhadap performa model diinginkan, meskipun ini mengurangi jumlah data latih. Dari variasi perbandingan tersebut digunakan untuk menentukan stabilitas dan keakuratan model dalam situasi yang berbeda. Membandingkan model dengan ketiga perbandingan ini memungkinkan penulis untuk melihat konsistensi hasil terbaik.

3. HASIL DAN PEMBAHASAN

3.1 Evaluasi Perbandingan Model

Evaluasi model menggunakan metode split data (0.2;0.3;0.4) dan membandingkan nilai k=3 dengan k=5 dengan menggunakan metrik accuracy, precision, recall, dan F1-score. Pengujian ini menggunakan bahasa pemrograman Python dengan bantuan software Google Colab. Berikut hasil perbandingan prediksi menggunakan algoritma KNN:

1. Split Data 0.2 (80:20)
   Pada gambar 4 terlihat bahwa Hasil evaluasi model dengan menggunakan metrik dan split data 0.2 dengan data latih berjumlah 1.294 sampel dan data uji berjumlah 324 sampel.

Metrik	K=3	K=5
Accuracy	0.25	0.28
Precision	0.29	0.29
Recall	0.27	0.29
F1-Score	0.25	0.28

Gambar 4. Tabel Split Data 0.2

2. Split Data 0.3 (70:30)
   Pada gambar 5 terlihat bahwa Hasil evaluasi model dengan menggunakan metrik dan split data 0.3 dengan data latih berjumlah 1.132 sampel dan data uji berjumlah 486 sampel.

Metrik	K=3	K=5
Accuracy	0.26	0.29
Precision	0.30	0.31
Recall	0.28	0.31
F1-Score	0.27	0.30

Gambar 5. Tabel Split Data 0.3

3. Split Data 0.4 (80:40)
   Pada gambar 6 terlihat bahwa Hasil evaluasi model dengan menggunakan metrik dan split data 0.4 dengan data latih berjumlah .294 sampel dan data uji berjumlah 324 sampel.

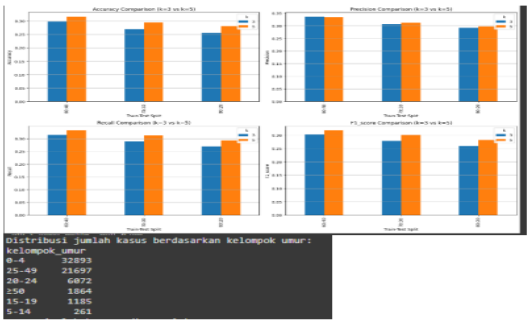
Metrik	K=3	K=5
Accuracy	0.29	0.31
Precision	0.33	0.33
Recall	0.31	0.33
F1-Score	0.30	0.31

Gambar 6. Tabel Split Data 0.4

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17, No. 1, Mei 2025
eISSN : 2580-3042
pISSN : 1979-0694

3.2 Penyajian Model Terbaik

Berdasarkan hasil pengujian performa model, dengan menggunakan metode split data dan membandingkan k=3 dengan k=5. Nilai k pada algoritma K-Nearest Neighbors (KNN) adalah parameter penting yang menentukan jumlah tetangga terdekat (neighbors) yang akan digunakan untuk menentukan kelas atau nilai suatu data baru. Pemilihan nilai k sangat memengaruhi kinerja algoritma KNN. Pemilihan k=3 dan k=5 pada algoritma KNN dilakukan untuk membandingkan performa model dengan ukuran tetangga yang berbeda. Keduanya adalah nilai ganjil untuk menghindari seri (tie) dalam klasifikasi. Dengan membandingkan k=3 dan k=5. Dapat dilihat bahwa pada gambar 8., bahwa split data 60:40 dengan nilai k=5 menghasilkan performa terbaik karena mempunyai nilai metrik yang lebih tinggi. Kelompok umur 0-4 tahun memiliki jumlah kasus tertinggi dengan 32,893 kasus, jauh melebihi kelompok lainnya. Hal ini terlihat pada gambar 7 yang menunjukkan bahwa anak-anak usia 0-4 tahun merupakan kelompok yang paling rentan atau paling banyak terlapor dalam dataset ini.



Gambar 7. Hasil Grafik Algoritma KNN

3.3 Perbandingan Data Aktual dan Data Prediksi

Tahun	Data Aktual	Data Prediksi
2019	4.537	4.464
2020	4,398	4.534
2021	4,531	6.604
2022	8.810	7.120
2023	9.710	6.670
Jumlah	**31.986**	**29.392**
Rata-rata	**6.3972**	**5.8784**

Gambar 8. Perbandingan Data Aktual dan Data Pribadi

Pada gambar 8, terlihat perbandingan antara data aktual yang diambil dari dataset penelitian dan data prediksi dicari menggunakan metode split data dan algoritma KNN untuk melakukan prediksi. Split data dilakukan dengan cara semua data tahun kecuali tahun target digunakan sebagai data pelatihan dan data tahun target digunakan sebagai data pengujian. Misalnya, jika ingin memprediksi tahun 2020 berarti data tahun 2019, 2021, 2022, dan 2023 dijadikan sebagai data pelatihan dan tahun 2020 dijadikan sebagai data pengujian.

3.4 Perbandingan Sample pada Tahun 2022 dan 2023

Menentukan prediksi sample berdasarkan mayoritas sample pada tahun 2022 dan 2023 dari jarak terdekat.

1. Untuk Data Uji: ‘jumlah_kasus = 567’, ‘tahun = 2023’:
    a. Jarak terdekat: ‘Umur 25-49’, ‘Umur 20- 24’,‘Umur 0-4'
    b. Prediksi: ‘Umur 25-49’ (mayoritas)
2. Untuk Data Uji: ‘jumlah_kasus = 468’, ‘tahun = 2022’:
    a. Jarak terdekat: ‘Umur 25-49’, ‘Umur 20- 24’,‘Umur 0-4’
    b. Prediksi: ‘Umur 25-49’ (mayoritas)

3.5 Evaluasi Matriks

Aktual/Prediksi	Umur 25-49	Selain Umur 25-49
Umur 25-49	2 (TP)	0 (FN)
Selain Umur 25-49	0 (FP)	0 (TN)

Gambar 9. Evaluasi Matriks

Pada gambar 9 dapat terlihat bahwa :

1. True Positives (TP) : 2 kasus di mana prediksi sesuai dengan kategori aktual (25-49).
2. False Negatives (FN): 0 kasus di mana kategori aktual adalah 25-49 tetapi diprediksi sebagai "Lainnya".
3. False Positives (FP): 0 kasus di mana kategori aktual adalah "Lainnya" tetapi diprediksi sebagai 25-49.
4. True Negatives (TN): 0 kasus di mana kategori aktual dan prediksi adalah "Lainnya".

Hasilnya menunjukan bahwa model memprediksi kedua data uji dengan benar.

Menghitung evaluasi metrik :

INFORMATIKA
Jurnal Informatika, Manajemen dan Komputer, Vol. 17, No. 1, Mei 2025
eISSN : 2580-3042
pISSN : 1979-0694

1. Accuracy
$$= \frac{TP + TN}{TP + TN + FP + FN} = \frac{2+0}{2+0+0+0} = 1.0$$
2. Precision
$$= \frac{TP}{TP + FP} = \frac{2}{2+0} = 1.0$$
3. Recall
$$= \frac{TP}{TP + FN} = \frac{2}{2+0} = 1.0$$
4. F1-Score
$$= 2 \times \frac{Precision \times Recall}{Precision + Recall} = 2 \times \frac{1.0 \times 1.0}{1.0 + 1.0} = 1.0$$

Hasil Akhir penelitian, dimana Model menunjukkan performa yang baik pada data uji ini dengan akurasi, presisi, recall, dan F1-score semuanya bernilai 1.0. Hal ini mengindikasikan bahwa model berhasil memprediksi semua data uji dengan benar, semua metrik menunjukkan nilai maksimal.

3.6 Pengujian

Split Data	Metrix	K=3	K=5
0.2	*Accuracy*	0.25	0.28
	Precision	0.29	0.29
	Recall	0.27	0.29
	F1-Score	0.25	0.28
0.3	*Accuracy*	0.26	0.29
	Precision	0.30	0.31
	Recall	0.28	0.31
	F1-Score	0.27	0.30
0.4	*Accuracy*	0.29	0.31
	Precision	0.33	0.33
	Recall	0.31	0.33
	F1- Score	0.30	0.31

Gambar 10. Hasil Pengujian

Pada gambar 10 dapat dilihat bahwa, tahapan pengujian pada klasifikasi dilakukan dengan membandingkan nilai accuracy, precission, recall dan F1-Score menggunakan nilai k=3 dan k=5 pada algoritma KNN dengan split data 0.2, 0.3, 0.4. Dilakukan dengan menggunakan bahasa pemrograman Python dan tools Google Colab. Dapat dilihat bahwa pada tabel 4.10, split data 0.4 menggunakan nilai k=5 mempunyai nilai yang lebih tinggi dibandingkan split data yang lain dengan nilai accuracy 0.31, nilai precision 0.33, nilai recall 0.33, dan nilai f1-score 0.31. Sehingga dapat disimpulkan, split data 0.4 mempunyai nilai dengan performa terbaik untuk memprediksi jumlah kasus HIV di Jawa Barat pada tahun 2019 sampai 2023.

4. KESIMPULAN

Pada penelitian ini menggunakan algoritma *K-Nearest Neighbor* untuk memprediksi jumlah kasus HIV di Jawa Barat pada tahun 2019 sampai 2023 dengan metode split data dan menggunakan *tools Google Colab*. Berdasarkan dari hasil pada penelitian ini dapat dilihat bahwa, algoritma KNN dengan metode split data 0.2 menghasilkan nilai yang lebih tinggi sehingga menjadikan split data tersebut menjadi nilai dengan performa terbaik dengan nilai *accuracy* 0.31, nilai *precision* 0.33, nilai *recall* 0.33, dan nilai *f1-score* 0.31. Selain itu, analisis distribusi jumlah kasus menunjukkan bahwa kelompok umur 0-4 tahun memiliki jumlah kasus tertinggi dengan 32,893 kasus. Hal ini mengindikasikan bahwa anak-anak usia 0-4 tahun merupakan kelompok paling rentan, sementara kelompok usia produktif juga menunjukkan tingkat kerentanan yang signifikan. Penggunaan algoritma *K-Nearest Neighbors* dapat membantu mengidentifikasi pola penyebaran dan memungkinkan kebijakan kesehatan yang lebih tepat sasaran. Upaya pencegahan dan pengobatan yang lebih baik sangat diperlukan untuk mengatasi masalah ini.

5. REFERENSI

Badan Pusat Statistik Provinsi Jawa Barat. (2024). *Statistik Daerah Provinsi Jawa Barat 2024*. *15*, 1–88. https://jabar.bps.go.id/id/publication/2024/09/26/00c711fd0b1ccdb85d746607/statistik-daerah-provinsi-jawa-barat-2024.html

Cholil, S. R., Handayani, T., Prathivi, R., & Ardianita, T. (2021). Implementasi Algoritma Klasifikasi K-Nearest Neighbor (KNN) Untuk Klasifikasi Seleksi Penerima Beasiswa. *IJCIT (Indonesian Journal on Computer and Information Technology)*, *6*(2), 118–127. https://doi.org/10.31294/ijcit.v6i2.10438

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

Herawati, I., Iswarawanti, D. N., Febriani, E., & Badriah, D. L. (2023). *Faktor - faktor yang berhubungan dengan kepatuhan minum obat antiretroviral (arv) pada odha di rsud 45 kuningan 2023*. 149–164.

Kodratul Munawar, K., & Irma Purnamasari, A. (2023). IMPLEMENTASI ALGORITMA K-

MEANS CLUSTERING PADA KLASTERISASI KASUS HIV DI JAWA BARAT. In *Jurnal Mahasiswa Teknik Informatika* (Vol. 7, Issue 2).

Miftahuddin, Y., Umaroh, S., & Karim, F. R. (2020). Perbandingan Metode Perhitungan Jarak Euclidean, Haversine, Dan Manhattan Dalam Penentuan Posisi Karyawan. *Jurnal Tekno Insentif*, *14*(2), 69–77. https://doi.org/10.36787/jti.v14i2.270

Noor, H., Dharmawati, A., & Qur'ana, T. W. (2021). Penerapan Algoritma K-Means Clustering Analysis Pada Kasus Penderita Hiv/Aids (Studi Kasus Kabupaten Banjar). *Technologia: Jurnal Ilmiah*, *12*(2), 72. https://doi.org/10.31602/tji.v12i2.4573

Nurani, I. A., HIdayat, R., & Nurfitri. (2022). *Tingkat Stress Mempengaruhi Kepatuhan Minum Obat Orang dengan HIV/AIDS di Rumah Singgah Peka Bogor Intan Asri Nurani*. *13*(April), 534–537.

Rosida, W., & Wijaya, Y. A. (2023). Klasterisasi Penyakit HIV/AIDS di Jawa Barat Menggunakan Algoritma K-Means Clustering. *Blend Sains Jurnal Teknik*, *1*(4), 306–315. https://doi.org/10.56211/blendsains.v1i4.235

Ruhyani, A. (2023). *Kasus HIV/AIDS Kota Bandung Tertinggi di Jawa Barat!* Detik.Com. https://www.detik.com/jabar/berita/d-7038958/kasus-hiv-aids-kota-bandung-tertinggi-di-jawa-barat

Samudra, A. W. P., Susanto, R. A., Putra, A. R., Kurniadi, F. I., & Juarto, B. (2022). *Klasifikasi HIV AIDS dengan Aplikasi Rapid Miner*. *1*, 15–19.

Sari, T. P., Hananto, A. L., Novalia, E., Tukino, T., & Hilabi, S. S. (2023). Implementasi Algoritma K-Means dalam Analisis Klasterisasi Penyebaran Penyakit Hiv/Aids. *Infotek : Jurnal Informatika Dan Teknologi*, *6*(1), 104–114. https://doi.org/10.29408/jit.v6i1.7423

Siahaan, Y. M., Cholissodin, I., & Adikara, P. P. (2020). Penerapan Metode Modified K-Nearest Neighbor pada Klasifikasi Penyakit Menular Seksual Pria. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *4*(11), 4194–4199.

Taniwan, P., Bilbina, A. F., Rafael, C., Ganap, S., & Faidah, D. Y. (2024). *PEMODELAN FAKTOR-FAKTOR YANG MEMENGARUHI KEJADIAN HIV / AIDS DI PROVINSI JAWA BARAT*. *5*(3), 3298–3308.

Teguh, R. (2022). *Dinkes Jabar Rutin Tes HIV pada Kelompok Rentan*. JABARPROVGOID. https://jabarprov.go.id/berita/dinkes-jabar-rutin-tes-hiv-pada-kelompok-rentan-6801