ANALISIS KLASTERISASI UNTUK PREDIKSI JUMLAH KASUS DBD BERDASARKAN JENIS KELAMIN DAN KABUPATEN/KOTA DI JAWA BARAT

Ali Ikbal [1], Ade Irma Purnamasari [2], Irfan Ali [3]
[1,2]Teknik Informatika, STMIK IKMI Cirebon
[3] Rekayasa Perangkat Lunak, STMIK IKMI Cirebon
Jl. Perjuangan No.10B, Karyamulya, Kec. Kesambi, Kota Cirebon, Jawa Barat 45135
aliikbal260401@gmail.com

ABSTRAK

Penyakit Demam Berdarah Dengue (DBD) adalah penyakit menular yang sering fatal, terutama di daerah tropis dan subtropis, menjadi isu kesehatan global dengan penyebaran luas dan dampak serius. Lingkungan memiliki peran penting dalam kondisi ini. Penelitian ini menggunakan metode *k-means*, *naive bayes*, dan linear regresi, masing-masing memiliki fungsinya sendiri. K-means berhasil mengelompokkan kasus DBD menjadi 3 cluster (tinggi, sedang, rendah) dengan indeks kinerja *Davies Bouldin* rata-rata 0.71. *Naive bayes* digunakan untuk memprediksi hasil cluster 2023 dengan akurasi model 88.27%. Linear regresi untuk menentukan jumlah kasus DBD tahun 2023, dengan kasus tertinggi Kota Bandung 1590 laki laki, 1585 perempuan dan Kota Bekasi 1576 laki-laki dan 1572 perempuan, Kota Depok masuk kategori sedang 948 laki-laki dan 944 perempuan, 24 kota lainnya masuk dalam kategori rendah. Penyebaran DBD cenderung lebih tinggi pada kaum laki-laki, mencapai puncak tertinggi pada tahun 2022. Analisis data kasus DBD dapat memberikan informasi dan kontribusi penting untuk pembangunan strategi pencegahan dan penanggulangan yang lebih efektif oleh masyarakat, pemerintah, dan dinas kesehatan.

Kata kunci : *Data Mining, K-means, Naïve Bayes, Linear Regression*

1. PENDAHULUAN

Demam Berdarah *Dengue* (DBD) adalah penyakit menular yang disebabkan oleh virus *dengue* yang ditularkan melalui gigitan nyamuk *Aedes aegypti* biasanya terjadi didaerah tropis dan subtropis[1]. Demam Berdarah *Dengeu* menjadi masalah kesehatan global karena penyebarannya yang luas dan dampaknya yang serius. Demam berdarah *dengue* sangat erat sekali dengan lingkungan, meningkatnya mobilitas dan kepadatan masyarakat serta kurangya implementasi 3M (Menguras, Mengubur, Menutup) menjadikan faktor utama penyebaran [2]. Salah satu indikator penting dalam mengukur tingkat keparahan penyakit DBD adalah jumlah kematian yang disebabkan oleh penyakit ini [3].

Penyebaran DBD mudah dan dapat menyerang semua kelompok umur, berpotensi menyebabkan kematian. Pemerintah dan masyarakat harus fokus pada pencegahan penyakit ini. Kebersihan lingkungan yang buruk menjadi faktor utama penyebaran DBD [4]. WHO mencatat bahwa Indonesia memiliki angka kasus DBD paling tinggi di Asia Tenggara, Pada tahun 2022, jumlah kasus DBD mencapai 143.266, dengan 1.237 kematian tercatat. Angka ini menunjukkan peningkatan dibandingkan dengan tahun 2021, yang mencatatkan 73.518 kasus dan 705 kematian[5]. Jumlah kasus DBD di Provinsi Jawa Barat pada tahun 2022 meningkat signifikan menjadi 36.608, dibandingkan dengan 23.959 kasus pada tahun 2021[6]. Penelitian ini membantu menentukan prioritas alokasi sumber daya kesehatan, memungkinkan efisiensi dalam penanganan klaster peningkatan kasus DBD. Kontribusinya terhadap penelitian kesehatan masyarakat adalah memberikan wawasan baru tentang faktor penyebab penyebaran penyakit. Teknik data mining dengan algoritma clustering dan forecasting digunakan untuk mencari pola dan tren, memprediksi jumlah kasus DBD 1 tahun ke depan. Hasil analisis dapat menjadi dasar penelitian lebih lanjut.

Clustering adalah teknik dalam data mining yang digunakan untuk mengelompokkan data ke dalam kelompok-kelompok yang memiliki kesamaan[7]. Dengan menggunakan metode data mining, khususnya klasterisasi, hasil dari penelitian akan menghasilkan tentang jumlah kasus tertinggi, kasus sedang dan kasus terendah. Prediksi dapat digunakan untuk menentukan kejadian dimasa yang akan datang belajar dengan menggunakan data history sebelumnya [8].

Penelitian analisis pengelompokan kasus DBD di Jawa Barat menghasilkan prediksi tingkat kasus tertinggi, sedang, rendah tahun 2023. Hasilnya memungkinkan identifikasi pola dan trend berdasarkan jenis kelamin dan kabupaten/kota, memberikan pemahaman mendalam untuk upaya kesehatan masyarakat, serta mendukung pemerintah, terutama Dinas Kesehatan, dalam perencanaan kebijakan dan strategi efektif penanggulangan kasus DBD.

2. TINJAUAN PUSTAKA

2.1. Data Mining

Data mining merupakan inti dari proses identifikasi pola, tren data, dan pengetahuan informasi yang tersembunyi dalam dataset. Dalam proses data mining, terdapat berbagai teknik seperti klasifikasi, regresi, pengelompokan, asosiasi, dan sebagainya [9].

2.2. K-Means

K-means digunakan untuk melakukan clustering data, setiap data point kemudian *diassign* ke cluster yang memiliki *centroid* terdekat berdasarkan suatu metrik jarak, yang biasanya adalah jarak *Euclidean*. Dengan kata lain, setiap data point akan diatributkan ke cluster yang memiliki *centroid* paling dekat [10].

1. Menetapkan jumlah cluster K.
2. Menginisialisasi nilai pusat K sebagai titik centroid secara acak.
3. Mengalokasikan ke cluster terdekat dengan menggunakan rumus jarak Euclidean.

$$v = \frac{\sum_{i=1}^{n} x_1}{\mathrm{n}} : i = \ 1,2,3 \ldots n$$

v = Nilai rata-rata dari himpunan data x
i = Indeks dari elemen data, dimulai dari i=1 hingga n.
n = jumlah total elemen dalam himpunan data x.

4. Melakukan iterasi pada titik centroid dengan menghitung nilai rata-rata dari seluruh data.
5. Mengulangi proses penetapan cluster hingga titik centroid tidak mengalami perubahan[11].

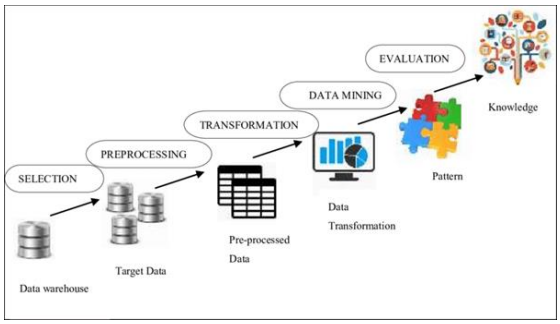
2.3. Naïve Bayes

Naïve bayes merupakan metode klasifikasi dalam bidang ilmu data dan statistik yang mengandalkan prinsip probabilitas. Meskipun dinamakan "naif" karena mengasumsikan independensi antara fitur-fitur data, teknik ini umumnya digunakan untuk meramalkan kategori entitas dengan mempertimbangkan probabilitas keterkaitan fitur-fitur tersebut[12].

2.4. Linear Regression

Regresi linear digunakan untuk menilai keterkaitan antara dua variabel, yang satu sebagai dependen dan yang lain sebagai independen. Tujuannya adalah menciptakan suatu model dari hubungan tersebut melalui persamaan garis lurus, yang dapat digunakan untuk melakukan prediksi. [13].

3. METODE PENELITIAN



Gambar 1. Metode knowledge discovery in database (Sumber: KDD)

Dalam penelitian ini, teknik analisis data yang akan diterapkan adalah *metode Knowledge Discovery in Database* (KDD), yang melibatkan serangkaian tahap, termasuk pemilihan data, pra-pemrosesan data, transformasi data, data mining, evaluasi pola, dan pengembangan pengetahuan. [11].

3.1. Data Selection

Sebelum memasuki tahap pra-pemrosesan data, dilakukan seleksi data terlebih dahulu untuk menentukan data yang akan digunakan dan yang paling relevan dengan fokus penelitian yang akan dilakukan.

3.2. Data Preprocessing

Pada tahap ini, dataset akan mengalami proses eliminasi, di mana data yang memiliki nilai null, duplikat, atau tidak terkait dengan tujuan penelitian akan dihapus. Langkah ini dilakukan untuk membersihkan dataset dan memastikan bahwa hanya data yang relevan dan lengkap yang akan digunakan dalam analisis.

3.3. Data Transformation

Dalam tahap transformasi data, data yang telah dibersihkan akan diproses untuk penggabungan atau penggabungan (merge), perubahan format tipe atribut, atau ekstraksi fitur yang lebih representatif. Berbagai teknik, seperti pemilihan atribut, normalisasi, pengurangan dimensi, dan transformasi lainnya, dapat diterapkan pada tahap ini.

3.4. Data Mining

Mencari pola atau tren sesuai dengan tujuan penelitian dengan teknik clustering k-means dan prediksi menggunakan *naïve bayes* dan linear regression. Clustering k-means akan mengelompokan kabupaten/kota masuk kedalam kategori status tinggi, sedang, rendah. Lalu metode naive bayes akan berperan untuk memprediksi kategori status pada tahun 2023, sedangkan metode linear regression akan meramal jumlah kasus DBD pada tahun 2023.

3.5. Evaluasi

Tahap evaluasi digunakan untuk menentukan tingkat akurasi dan validasi data mining. Tahap ini akan dilakukan secara terus menerus untuk menemukan *performance* dan validasi data paling tepat, tentunya sesuai dengan tujuan yang sudah ditetapkan.

3.6. Knowledge

Setelah data dievaluasi maka dari hasil evaluasi tadi akan menghasilkan sebuah informasi ataupun pengetahuan. Hasil digunakan untuk membuat keputusan, mengembangkan strategi, atau meningkatkan kinerja.

4. HASIL DAN PEMBAHASAN

4.1. Data Selection

Data yang digunakan ini menggunakan data tahun 2014-2022 dengan jumlah 486 record dan 7

attribut diantaranya ada id, kode_kabupaten_kota, nama_kabupaten_kota, jenis_kelamin, jumlah_kasus, satuan, dan tahun. Seperti pada tabel 1.

4.2. Data Preprocessing

Dalam tahap ini, dataset akan menjalani proses eliminasi, di mana data yang memiliki nilai *null*, duplikat, atau tidak terkait dengan tujuan penelitian akan dihapus. Tindakan ini bertujuan untuk membersihkan dataset, memastikan bahwa hanya data yang relevan dan lengkap yang akan digunakan dalam analisis. Karena dalam dataset peneliti ini tidak ada data yang *null* ataupun duplikat maka lanjut pada tahap berikutnya.

4.3. Data Transformation

4.3.1. Data Transformation Clustering K-means

Attribut yang digunakan untuk clustering yaitu nama_kabupaten_kota, jenis_kelamin, jumlah_kasus, dan tahun. Attribut jenis kelamin akan di rubah menjadi numerical

4.3.2. Data Transformation Prediction Naïve Bayes

Attribut yang digunakan untuk melakukan prediksi menggunakan *naive bayes* yaitu id, nama_kabupaten_kota, jenis_kelamin, jumlah_kasus, tahun dan status.

Table 1 Data selection

id	kode_kabupaten_kota	nama_kabupaten_kota	jenis_kelamin	jumlah_kasus	tahun
1	3201	KABUPATEN BOGOR	LAKI-LAKI	915	2014
2	3201	KABUPATEN BOGOR	PEREMPUAN	919	2014
3	3202	KABUPATEN SUKABUMI	LAKI-LAKI	409	2014
4	3202	KABUPATEN SUKABUMI	PEREMPUAN	295	2014
5	3203	KABUPATEN CIANJUR	LAKI-LAKI	200	2014
…	…	…	…	…	…
482	3277	KOTA CIMAHI	PEREMPUAN	370	2022
483	3278	KOTA TASIKMALAYA	LAKI-LAKI	912	2022
484	3278	KOTA TASIKMALAYA	PEREMPUAN	943	2022
485	3279	KOTA BANJAR	LAKI-LAKI	56	2022
486	3279	KOTA BANJAR	PEREMPUAN	57	2022

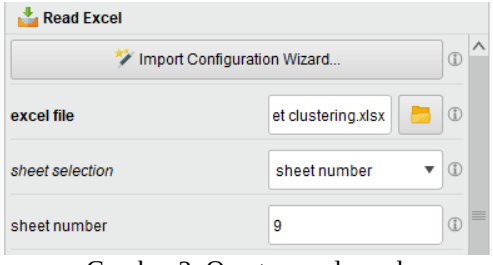
4.3.3. Data Transformation Forcasting Linear Regression

Attribut yang digunakan untuk melakukan *forcasting* menggunakan *linear regression* yaitu id, kode_kabupaten_kota, jenis_kelamin, jumlah_kasus, tahun ,status dan kode_status. Jenis kelamin akan di konversi menjadi numerical.

4.4. Data Mining

4.4.1. Data Mining Clustering K-means

Tahap clustering menggunakan operator K-means melibatkan serangkaian langkah menggunakan operator Read Excel, Nominal to Numerical, Multiply, Clustering K-Means, dan Cluster Distance Performance. Operator Read Excel digunakan untuk membaca dataset dalam format file Excel, dengan beberapa parameter seperti path atau URL file Excel (excel file), pemilihan lembar kerja (sheet selection), dan penentuan lembar kerja berdasarkan nomor atau nama sheet (sheet number).

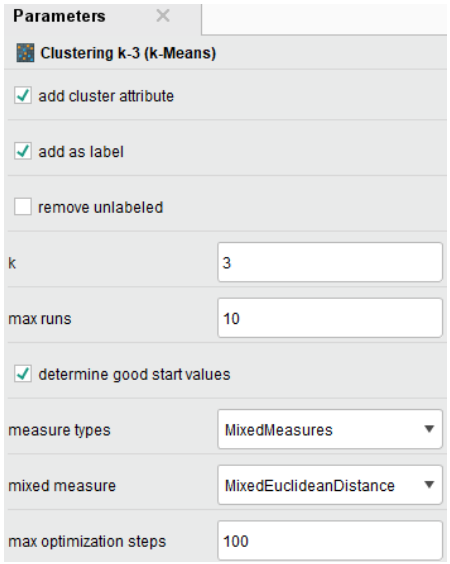


Gambar 2. Oprator read excel

Operator Nominal to Numerical berfungsi untuk mengubah atribut dengan tipe data nominal menjadi bentuk numerik, untuk mengonversi atribut jenis kelamin dalam suatu dataset. Sementara itu, operator Multiply memungkinkan eksekusi lebih dari satu operator dalam satu proses running.

Gambar 3. Oprator nominal to numerical

Operator clustering digunakan untuk grup data berdasarkan karakteristik yang serupa. Di dalam operator ini, ada parameter "*add cluster attribute*" untuk menciptakan atribut klaster baru, dan jika "*add as label*" diaktifkan, atribut klaster dapat dijadikan sebagai label. Parameter K menentukan jumlah klaster, dan "max runs" melakukan iterasi untuk menetapkan klaster. Pada bagian "*measure types*" "*mixed measure*" dipilih karena data bersifat numerik dan nominal. "*Max optimization steps*" melakukan iterasi untuk menyesuaikan pusat *centroid* sesuai dengan nilai yang ditentukan.



Gambar 4. Oprator clustering k-means

a. Hasil cluster 2014

Table 2. Hasil cluster tahun 2014

Tahun	Nilai Davies Bouldin Index				
	K-2	K-3	K-4	K-5	K-6
2014	0.081	0.095	0.068	0.078	0.084

Hasil clustering dari Tabel 2 menunjukkan bahwa nilai DBI paling mendekati 0 pada K-4, DBD tahun 2014 dibagi menjadi 4 cluster. Cluster 2, yang merupakan Kota Bandung, memiliki kasus tertinggi dengan 1698 laki-laki dan 1434 perempuan. Cluster 3 memiliki 2 entitas, yaitu 1 Kabupaten Bogor; cluster 1 memiliki 24 entitas, mewakili 12 Kabupaten/kota; dan cluster 0 memiliki 26 entitas, mewakili 13 Kabupaten/kota.

b. Hasil cluster 2015

Table 3. Hasil cluster tahun 2015

Tahun	Nilai Davies Bouldin Index				
	K-2	K-3	K-4	K-5	K-6
2015	0.021	0.074	0.077	0.067	0.071

Hasil clustering pada Tabel 3 menunjukkan bahwa nilai DBI paling mendekati 0 pada K-2, menandakan bahwa kasus DBD tahun 2015 dibagi menjadi 2 cluster. Cluster 1, yang mencakup Kota Bandung, memiliki kasus tertinggi dengan 1822 laki-laki dan 1818 perempuan. Cluster 0 memiliki 52 entitas, yang mewakili 26 kabupaten/kota.

c. Hasil cluster 2016

Table 4. Hasil cluster tahun 2016

Tahun	Nilai Davies Bouldin Index				
	K-2	K-3	K-4	K-5	K-6
2016	0.051	0.075	0.078	0.067	0.073

Hasil dari clustering pada Tabel 3 menunjukkan bahwa nilai DBI mendekati 0 pada K-2, menunjukkan bahwa kasus DBD tahun 2016 terbagi menjadi 2 cluster. Cluster 1, terutama di Kota Bandung, jumlah kasus tertinggi dengan 2054 laki-laki dan 1827 perempuan, diikuti oleh Kabupaten Bogor, Kabupaten Bandung, Kota Bekasi, dan Kota Depok. Sementara itu, Cluster 0 memiliki 44 entitas, yang mewakili 22 Kabupaten/kota.

d. Hasil cluster 2017

Table 5. Hasil cluster tahun 2017

Tahun	Nilai Davies Bouldin Index				
	K-2	K-3	K-4	K-5	K-6
2017	0.106	0.067	0.100	0.114	0.102

Hasil clustering pada tabel 5 menunjukkan bahwa nilai DBI mendekati 0 pada K-3, mengindikasikan bahwa kasus DBD tahun 2017 dibagi menjadi 3 cluster. Cluster 1, terutama di Kota Bandung, menunjukkan kasus tertinggi dengan 980 laki-laki dan 806 perempuan. Sementara itu, Cluster 2 memiliki 8 entitas, dan Cluster 0 memiliki 44 entitas.

e. Hasil cluster 2018

Table 6. Hasil cluster tahun 2018

Tahun	Nilai Davies Bouldin Index				
	K-2	K-3	K-4	K-5	K-6
2018	0.063	0.076	0.051	0.085	0.096

Hasil dari clustering pada tabel 6 menunjukkan bahwa nilai DBI mendekati 0 pada K-4, mengelompokkan kasus DBD tahun 2018 menjadi 4 cluster. Cluster 1, yang terdapat di Kota Bandung, menunjukkan kasus tertinggi dengan 1453 laki-laki dan 1373 perempuan. Cluster 3 melibatkan 2 entitas (Kabupaten Bandung), Cluster 2 melibatkan 10 entitas, dan Cluster 0 melibatkan 40 entitas.

f. Hasil cluster 2019

Table 7. Hasil cluster tahun 2019

Tahun	Nilai Davies Bouldin Index				
	K-2	K-3	K-4	K-5	K-6
2019	0.078	0.052	0.088	0.071	0.077

Hasil clustering pada Tabel 7 menunjukkan bahwa nilai DBI mendekati 0 pada K-3, mengelompokkan kasus DBD tahun 2019 menjadi 3 cluster. Cluster 1, yang terdapat di Kota Bandung, menunjukkan kasus tertinggi dengan 2322 laki-laki dan 2102 perempuan. Cluster 2 melibatkan 8 entitas, sedangkan Cluster 0 melibatkan 44 entitas.

g. Hasil cluster 2020

Table 8. Hasil cluster tahun 2020

Tahun	Nilai Davies Bouldin Index				
	K-2	K-3	K-4	K-5	K-6
2020	0.093	0.064	0.065	0.069	0.088

Hasil clustering pada Tabel 8 menunjukkan bahwa nilai DBI mendekati 0 pada K-3, mengelompokkan kasus DBD tahun 2020 menjadi 3 cluster. Cluster 1, yang terdapat di Kota Bandung, menunjukkan kasus tertinggi dengan 2322 laki-laki dan 2102 perempuan. Cluster 2 melibatkan 17 entitas, sedangkan Cluster 0 melibatkan 35 entitas.

h. Hasil cluster 2021

Table 9. Hasil cluster tahun 2021

Tahun	Nilai Davies Bouldin Index				
	K-2	K-3	K-4	K-5	K-6
2021	0.075	0.064	0.070	0.087	0.073

Hasil clustering pada Tabel 9 menunjukkan bahwa nilai DBI mendekati 0 pada K-3, mengelompokkan kasus DBD tahun 2021 menjadi 3 cluster. Cluster 2, terutama di Kota Bandung, menunjukkan kasus tertinggi dengan 1951 laki-laki dan 1792 perempuan, diikuti oleh Kota Depok dengan 1707 laki-laki dan 1448 perempuan. Cluster 1 melibatkan 8 entitas, sementara Cluster 0 melibatkan 42 entitas.

i. Hasil cluster 2022

Table 10. Hasil cluster tahun 2022

Tahun	Nilai Davies Bouldin Index				
	K-2	K-3	K-4	K-5	K-6
2022	0.051	0.070	0.079	0.081	0.074

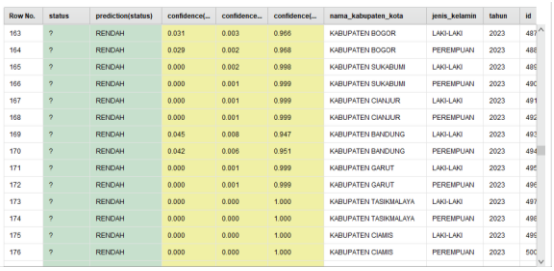
Hasil dari clustering pada Tabel 10 K-2 untuk kasus DBD tahun 2022 menunjukkan bahwa Cluster 2 di Kota Bandung memiliki jumlah kasus tertinggi: 2646 laki-laki dan 2559 perempuan. Di Kabupaten Bandung, pada Cluster 2, jumlah kasus perempuan mendominasi dengan 2177 dibandingkan dengan laki-laki sebanyak 2014. Sementara itu, Cluster 0 melibatkan 50 entitas..

4.4.2. Data Mining Prediction Naïve Bayes

Dalam tahap prediksi *Naïve Bayes*, dilakukan prediksi hasil dari clustering untuk *attribut* status pada tahun 2023. Operator yang digunakan meliputi *read excel*, *set role*, *Naïve Bayes*, *apply model*, *dan performance*. Data terbagi menjadi data latih (378 record, 2014-2020) dan data uji (162 record, 2021-2023). Setelah diimport, label diberikan menggunakan operator set role, dengan atribut status diisi sebagai "status" dan target role diisi sebagai "label". Atribut status akan menjadi label dari dataset.

Gambar 5. Set role naïve bayes

Dalam tahap inti, perhatian difokuskan pada pemrosesan dengan menggunakan operator Naive Bayes. Model Naive Bayes dilatih menggunakan data historis untuk mengidentifikasi pola dan hubungan antara atribut status dengan variabel lainnya. Melalui proses pelatihan ini, model dapat memahami struktur data dan membuat estimasi dengan memanfaatkan asumsi independensi fitur yang dimiliki oleh Naive Bayes.



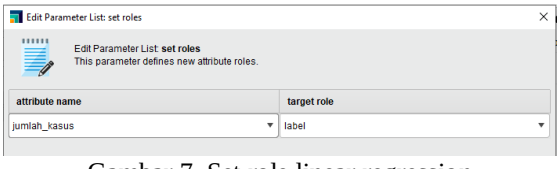
Row No.	status	prediction(status)	confidence(...	confidence...	confidence(...	nama_kabupaten_kota	jenis_kelamin	tahun	id
163	?	RENDAH	0.031	0.003	0.966	KABUPATEN BOGOR	LAKI-LAKI	2023	487
164	?	RENDAH	0.029	0.002	0.968	KABUPATEN BOGOR	PEREMPUAN	2023	488
165	?	RENDAH	0.000	0.002	0.998	KABUPATEN SUKABUMI	LAKI-LAKI	2023	489
166	?	RENDAH	0.000	0.001	0.999	KABUPATEN SUKABUMI	PEREMPUAN	2023	490
167	?	RENDAH	0.000	0.001	0.999	KABUPATEN CIANJUR	LAKI-LAKI	2023	491
168	?	RENDAH	0.000	0.001	0.999	KABUPATEN CIANJUR	PEREMPUAN	2023	492
169	?	RENDAH	0.045	0.008	0.947	KABUPATEN BANDUNG	LAKI-LAKI	2023	493
170	?	RENDAH	0.042	0.006	0.951	KABUPATEN BANDUNG	PEREMPUAN	2023	494
171	?	RENDAH	0.000	0.001	0.999	KABUPATEN GARUT	LAKI-LAKI	2023	495
172	?	RENDAH	0.000	0.001	0.999	KABUPATEN GARUT	PEREMPUAN	2023	496
173	?	RENDAH	0.000	0.000	1.000	KABUPATEN TASIKMALAYA	LAKI-LAKI	2023	497
174	?	RENDAH	0.000	0.000	1.000	KABUPATEN TASIKMALAYA	PEREMPUAN	2023	498
175	?	RENDAH	0.000	0.000	1.000	KABUPATEN CIAMIS	LAKI-LAKI	2023	499
176	?	RENDAH	0.000	0.000	1.000	KABUPATEN CIAMIS	PEREMPUAN	2023	500

Gambar 6. Hasil prediksi naive bayes

Hasil penerapan model Naïve Bayes pada tahun 2023, seperti dalam Gambar 6, menunjukkan bahwa kota-kota seperti Kota Bandung dan Kota Bekasi termasuk dalam kategori kasus tertinggi. Kota Depok, selain dari ketiga kota tersebut, tergolong dalam kategori kasus sedang, sementara kota-kota lainnya masuk ke dalam kategori kasus rendah.

4.4.3. Data Mining Forcasting Linear Regression

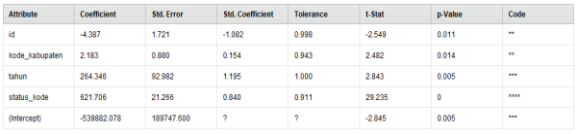
Tahap forecasting menggunakan metode regresi linear bertujuan meramalkan jumlah kasus DBD tahun 2023 dengan memanfaatkan prediksi dari algoritma naive bayes. Proses ini melibatkan operator read excel, set role, model regresi linear, generate attribut, dan performance (regresi). Dalam melaksanakan forecasting, dua set data yang berbeda diperlukan, yakni data training (379 rekam) dan data testing (162 rekam). Setelah mengimpor data melalui operator read excel, langkah selanjutnya adalah penyesuaian data dengan memberikan label menggunakan operator set role. Dalam proses forecasting, label diperlukan untuk atribut jumlah kasus, yang akan menjadi variabel yang diprediksi dalam analisis. Operator set role memiliki peran penting dalam menentukan label untuk atribut tersebut, membantu mempersiapkan data untuk tahap peramalan. Dengan memberikan label, sistem dapat memahami dan mengidentifikasi variabel target yang akan diestimasi dalam konteks peramalan.



Gambar 7. Set role linear regression

Dalam tahap Regresi Linear, analisis dilakukan untuk memahami dan memodelkan hubungan linier antara variabel independen dan variabel dependen. Proses ini dimulai dengan mengidentifikasi atribut atau variabel input yang akan digunakan untuk memprediksi nilai variabel target, dengan fokus

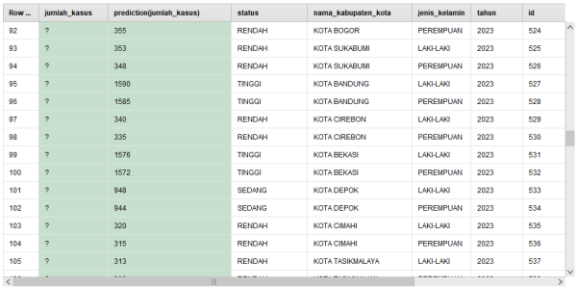
khusus pada atribut yang akan dijadikan variabel dependen, yaitu jumlah kasus. Selanjutnya, model Regresi Linear diterapkan menggunakan operator Apply Model. Model ini menggunakan pendekatan matematis untuk menemukan hubungan linier terbaik antara variabel independen dan variabel dependen, sebagaimana terlihat pada Gambar 8 yang menunjukkan bahwa variabel yang paling berpengaruh adalah status_kode, tahun, kode_kabupaten, dan id. Setelah model diaplikasikan, langkah Generate Attributes digunakan untuk menghasilkan atribut prediksi yang berisi nilai yang diprediksi berdasarkan model yang telah dilatih.



Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
id	-4.387	1.721	-1.082	0.998	-2.549	0.011	**
kode_kabupaten	2.183	0.880	0.154	0.943	2.482	0.014	**
tahun	264.346	92.982	1.195	1.000	2.843	0.005	***
status_kode	621.706	21.266	0.840	0.911	29.235	0	****
(Intercept)	-539882.078	189747.600	?	?	-2.845	0.005	***

Gambar 8. Model linear regression

Dengan menggunakan regresi linear untuk peramalan pada tahun 2023, hasilnya menunjukkan bahwa Kota Bandung memiliki jumlah kasus tertinggi, yakni 1590 laki-laki dan 1585 perempuan, diikuti oleh Kota Bekasi dengan 1576 laki-laki dan 1572 perempuan. Kota Depok tergolong dalam kategori kasus sedang dengan jumlah 948 laki-laki dan 944 perempuan, sedangkan kota-kota lainnya masuk ke dalam kategori kasus rendah.



Row ...	jumlah_kasus	prediction(jumlah_kasus)	status	nama_kabupaten_kota	jenis_kelamin	tahun	id
92	?	355	RENDAH	KOTA BOGOR	PEREMPUAN	2023	524
93	?	353	RENDAH	KOTA SUKABUMI	LAKI-LAKI	2023	525
94	?	348	RENDAH	KOTA SUKABUMI	PEREMPUAN	2023	526
95	?	1590	TINGGI	KOTA BANDUNG	LAKI-LAKI	2023	527
96	?	1585	TINGGI	KOTA BANDUNG	PEREMPUAN	2023	528
97	?	340	RENDAH	KOTA CIREBON	LAKI-LAKI	2023	529
98	?	335	RENDAH	KOTA CIREBON	PEREMPUAN	2023	530
99	?	1576	TINGGI	KOTA BEKASI	LAKI-LAKI	2023	531
100	?	1572	TINGGI	KOTA BEKASI	PEREMPUAN	2023	532
101	?	948	SEDANG	KOTA DEPOK	LAKI-LAKI	2023	533
102	?	944	SEDANG	KOTA DEPOK	PEREMPUAN	2023	534
103	?	320	RENDAH	KOTA CIMAHI	LAKI-LAKI	2023	535
104	?	315	RENDAH	KOTA CIMAHI	PEREMPUAN	2023	536
105	?	313	RENDAH	KOTA TASIKMALAYA	LAKI-LAKI	2023	537

Gambar 9 Hasil forcasting linear regression

4.5. Evaluasi

4.5.1. Evaluasi Transformation Clustering K-means

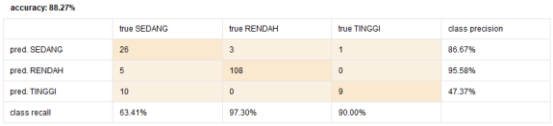
Mengevaluasi hasil clustering dengan menggunakan operator *distance performance* dan parameter *davies bouldin index*. Hasil clustering dikatakan baik karena nilai DBI mendekati 0, seperti yang terlihat pada tabel 11 dengan penggunaan K-3 terbanyak. Kelompok hasil pengelompokan dibagi menjadi tinggi, sedang, dan rendah.

Table 11 Hasil evaluasi nilai davies bouldin index

Tahun	Nilai Davies Bouldin Index				
	K-2	K-3	K-4	K-5	K-6
2014	0.081	0.095	0.068	0.078	0.084
2015	0.021	0.074	0.077	0.067	0.071
2016	0.051	0.075	0.078	0.067	0.073
2017	0.106	0.067	0.100	0.114	0.102
2018	0.063	0.076	0.051	0.085	0.096
2019	0.078	0.052	0.088	0.071	0.077
2020	0.093	0.064	0.065	0.069	0.088
2021	0.075	0.064	0.070	0.087	0.073
2022	0.051	0.070	0.079	0.081	0.074

4.5.2. Evaluasi Transformation Prediction Naïve Bayes

Model *Naive Bayes* dievaluasi menggunakan *Operator Performance* (Classification) untuk mengukur akurasi. Pada kategori status sedang, terdapat 26 prediksi benar, namun terdapat 5 kesalahan masuk kategori rendah dan 10 kesalahan masuk kategori tinggi, menghasilkan kelas *recall* sebesar 63.41%. Kategori status rendah memiliki 108 prediksi benar, 3 prediksi salah masuk kategori sedang sehingga kelas recall mencapai 97.30%. Pada kategori status tinggi, terdapat 9 prediksi benar dan 1 kesalahan masuk kategori sedang, menghasilkan kelas recall 90.00%. Akurasi keseluruhan model mencapai 88.27%, mencerminkan kemampuan model dalam mengklasifikasikan data secara benar.



accuracy: 88.27%

	true SEDANG	true RENDAH	true TINGGI	class precision
pred. SEDANG	26	3	1	86.67%
pred. RENDAH	5	108	0	95.58%
pred. TINGGI	10	0	9	47.37%
class recall	63.41%	97.30%	90.00%	

Gambar 10. Performance naive bayes

4.5.3. Evaluasi Transformation Forcasting Linear Regression

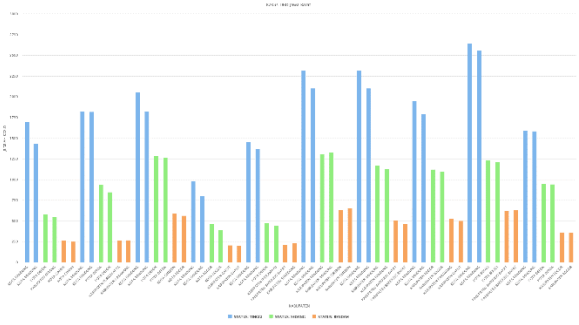
Dari beberapa percobaan evaluasi yang dilakukan oleh peneliti dengan menggunakan data training sebanyak 379 record atau mulai dari tahun 2015 hingga 2021, hasil menunjukkan bahwa Root Mean Square Error (RMSE) sekitar 284.829. Hal ini menggambarkan bahwa rata-rata kesalahan prediksi model adalah sekitar 284.829 unit (dalam satuan yang sesuai dengan variabel dependen). Dalam konteks ini, nilai RMSE yang relatif rendah dapat diartikan sebagai indikasi bahwa model memberikan prediksi yang baik secara keseluruhan. Relative Error sekitar 47.28%, dengan deviasi standar sekitar +/-73.14%, mengindikasikan bahwa, secara rata-rata, model memiliki kesalahan sekitar 47.28%. Semakin kecil nilai Relative Error, semakin baik performa model. Dengan nilai ini, dapat disimpulkan bahwa model memberikan prediksi yang cukup mendekati nilai sebenarnya.

Table 12. Performance linear regression

Performance Linear Regression				
No	Parameter	2014	2015	2016
1	RMSE	288.731 +/- 0.000	284.829 +/- 0.000	297.956 +/- 0.000
2	relative error	51.13% +/- 89.23%	47.28% +/- 73.14%	42.91% +/- 52.43%

4.6. Knowledge

Tiga kelompok cluster (Tinggi, Sedang, Rendah) diidentifikasi sebagai yang terbaik, dengan K-3 sebagai nilai terbanyak, menunjukkan kualitas cluster yang lebih baik. Kasus DBD tertinggi terjadi pada 2022 di Kota Bandung, mencapai 2646 kasus laki-laki dan 2559 perempuan. Puncak kasus sedang terjadi pada 2019 di Kabupaten Bandung, sementara kasus terendah terjadi pada 2017 di Kota Bogor, didominasi oleh kaum laki-laki. Seperti pada gambar 10 warna biru menunjukan kasus tetinggi, warna Hijau kasus sedang, warna orange kasus rendah. Model *naive bayes* memberikan akurasi 88.27%, menandakan kemampuannya dalam memprediksi variabel status. Variabel "status_kode" dan "tahun" berpengaruh signifikan dalam model linear regression, dengan RMSE sekitar 284.829, deviasi standar 0.000, dan *Relative Err*or sekitar 47.28% (+/- 73.14%), menunjukkan tingkat kesalahan yang relatif rendah dalam memprediksi jumlah kasus DBD.



Gambar 11. Hasil viasualisasi cluster tahun 2014-2023

5. KESIMPULAN DAN SARAN

Dengan menggunakan algoritma clustering k-means, berhasil mengelompokkan daerah berdasarkan tingkat kasus DBD menjadi tinggi, sedang, dan rendah, dengan evaluasi menggunakan *Davies Bouldin Index*. Hasil menunjukkan bahwa nilai terbaik untuk jumlah cluster (K) adalah K-3, dengan rata-rata *Davies Bouldin Index* sebesar 0.71. Selanjutnya, menggunakan metode linear regression untuk forecasting, Kota Bandung menduduki peringkat pertama dengan 1590 kasus pada laki-laki dan 1585 kasus pada perempuan, diikuti oleh Kota Bekasi dan Kota Depok. Hasil forecasting juga mengungkap pola penyebaran DBD yang didominasi oleh kasus pada laki-laki, terutama di kota-kota besar, menunjukkan ketidaksetaraan gender dalam dampak epidemi DBD. Hal ini menyoroti perlunya strategi pencegahan yang lebih fokus pada populasi laki-laki di kota-kota besar untuk mengatasi penyebaran penyakit dengan lebih efektif.

DAFTAR PUSTAKA

[1] D. Amelia, T. N. Padilah, and A. Jamaludin, "Optimasi Algoritma K-Means Menggunakan Metode Elbow dalam Pengelompokan Penyakit Demam Berdarah Dengue (DBD) di Jawa Barat," *Jurnal Ilmiah Wahana Pendidikan*, vol. 8, no. 11, pp. 207–215, 2022, doi: 10.5281/zenodo.6831380.

[2] T. P. Lestari, S. Sholikhah, and N. H. Qowi, "Factors Influencing the Incidence of Dengue Haemorrhagic Fever," *Jurnal Ners*, vol. eISSN, no. 3, p. 2019, 2019, doi: 10.20473/jn.v14i3(si).17153.

[3] F. Fadli and B. Butar Butar, "Penerapan Decision Tree Menggunakan Algoritma C4.5 Untuk Deteksi Demam Berdarah Pada RS. IMC Bintaro," *IJSE-Indonesian Journal on Software Engineering*, vol. 5, no. 1, pp. 75–86, 2019.

[4] T. M. M. Tyas and A. I. Purnamasari, "Penerapan Algoritma K-means dalam Mengelompokkan Demam Berdarah Dengue Berdasarkan Kabupaten," *Blend Sains Jurnal Teknik*, vol. 1, no. 4, pp. 277–283, Mar. 2023, doi: 10.56211/blendsains.v1i4.231.

[5] S. E. M. A. Ph. D. S. J. K. R. Kunta Wibawa Dasa Nugraha and S. H. , M. A. P. K. P. D. dan T. I. Tiomaida Seviana H.H., "PROFIL KESEHATAN INDONESIA 2022," 2023. Accessed: Dec. 23, 2023. [Online]. Available: http://www.kemenkes/go.id

[6] P. H. dr Raden Vini Adiani Dewi Kepala Dinas Kesehatan Provinsi Jawa Tengah Pengarah Firman Adam *et al.*, "PROFIL KESEHATAN PROVINSI JAWA BARAT 2022," 2023. Accessed: Dec. 23, 2023. [Online]. Available: https://diskes.jabarprov.go.id

[7] R. Gustriansyah, N. Suhandi, and F. Antony, "Clustering optimization in RFM analysis based on k-means," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 18, no. 1, pp. 470–477, 2019, doi: 10.11591/ijeecs.v18.i1.pp470-477.

[8] S. Saru, "ANALYSIS AND PREDICTION OF DIABETES USING MACHINE LEARNING," 2019. [Online]. Available: https://ssrn.com/abstract=3368308

[9] S. Mohamed and A. Ezzati, "A data mining process using classification techniques for employability prediction," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 14, no. 2, pp. 1025–1029, May 2019, doi: 10.11591/ijeecs.v14.i2.pp1025-1029.

[10] M. A. Ahmed, H. Baharin, and P. N. E. Nohuddin, "k-means variations analysis for translation of English Tafseer Al-Quran text," *International Journal of Electrical and Computer Engineering*, vol. 13, no. 3, pp. 3255–3265, Jun. 2023, doi: 10.11591/ijece.v13i3.pp3255-3265.

[11] G. Gustientiedina, M. H. Adiya, and Y. Desnelita, "Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 5, no. 1, pp. 17–24, Apr. 2019, doi: 10.25077/teknosi.v5i1.2019.17-24.

[12] M. B. Ressan and R. F. Hassan, "Naïve-Bayes family for sentiment analysis during COVID-19 pandemic and classification tweets," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 1, pp. 375–383, Oct. 2022, doi: 10.11591/ijeecs.v28.i1.pp375-383.

[13] V. R. Prasetyo, H. Lazuardi, A. A. Mulyono, and C. Lauw, "Penerapan Aplikasi RapidMiner Untuk Prediksi Nilai Tukar Rupiah Terhadap US Dollar Dengan Metode Linear Regression," *Jurnal Nasional Teknologi dan Sistem Informasi*, vol. 7, no. 1, pp. 8–17, 2021, doi: 10.25077/teknosi.v7i1.2021.8-17.