

PENGELOMPOKAN PROVINSI MEROKOK DI INDONESIA BERDASARKAN IPM MENGGUNAKAN K-MEANS DAN KNN

Oleh:

Baihaqi Asa'ari Lubis¹

Indry Widiyani²

Elkin Rilvani³

Universitas Pelita Bangsa

Alamat: JL. Inspeksi Kalimalang No.9, Cibatu, Cikarang Sel., Kabupaten Bekasi, Jawa Barat (17530).

Korespondensi Penulis: baihaqi.lubis@mhs.pelitabangsa.ac.id,
indrywidiyani5@gmail.com, elkin.rilvani@pelitabangsa.ac.id.

Abstract. *One of the challenges in national health development is the prevalence of smoking in Indonesia, including among the productive age group. The aim of this study is to group and classify provinces in Indonesia based on the percentage of the population aged 15 and above who smoke, as well as the Human Development Index (HDI) values in 2024. This study uses K-Means Clustering to group provinces into risk clusters, and K-Nearest Neighbor (KNN) is used to classify provinces into risk clusters based on HDI values relative to the K-Means results. Data is provided by the Central Bureau of Statistics (BPS) and the Regional Data Management Information System (SIMREG-Bappenas). The clustering results show that there are three main clusters that differ in terms of smoking rates and HDI. The KNN model is used to predict clusters based on HDI values and produces a high classification accuracy, indicating that HDI is related to the pattern of smoking behavior distribution at the provincial level. This research contributes to mapping regions with high-risk smoking behavior and can serve as a foundation for data-driven health policy formulation at the regional level.*

Keywords: *Data Mining, K-Means Clustering, K-Nearest Neighbor, IPM, HDI.*

PENGELOMPOKAN PROVINSI MEROKOK DI INDONESIA BERDASARKAN IPM MENGGUNAKAN K-MEANS DAN KNN

Abstrak. Salah satu kendala dalam pembangunan kesehatan nasional adalah prevalensi perokok di Indonesia, termasuk di kelompok usia produktif. Tujuan dari penelitian ini adalah untuk mengelompokkan dan mengklasifikasikan provinsi-provinsi di Indonesia berdasarkan persentase penduduk usia 15 tahun ke atas yang merokok serta nilai Indeks Pembangunan Manusia (IPM) pada tahun 2024. Studi ini menggunakan *K-Means Clustering* untuk mengelompokkan provinsi ke dalam klaster risiko, dan *K-Nearest Neighbor* (KNN) digunakan untuk mengelompokkan provinsi ke dalam klaster risiko berdasarkan nilai IPM terhadap klaster hasil K-Means. Badan Pusat Statistik (BPS) dan Sistem Informasi Manajemen Data Regional (SIMREG-Bappenas) menyediakan data. Hasil pengelompokan menunjukkan bahwa terdapat tiga klaster utama yang berbeda dalam hal tingkat merokok dan IPM. Model KNN digunakan untuk memprediksi klaster berdasarkan nilai IPM, dan menghasilkan tingkat akurasi klasifikasi yang tinggi, yang menunjukkan bahwa IPM memiliki keterkaitan terhadap pola persebaran kebiasaan merokok di tingkat provinsi. Penelitian ini memberikan kontribusi dalam pemetaan wilayah dengan risiko perilaku merokok tinggi, serta dapat menjadi dasar dalam penyusunan kebijakan kesehatan berbasis data di tingkat regional.

Kata Kunci: Data Mining, *K-Means Clustering*, *K-Nearest Neighbor*, IPM, HDI.

LATAR BELAKANG

Merokok menjadi masalah kesehatan utama di setiap negara maju maupun negara berkembang. Aktivitas Perokok tidak hanya dilakukan oleh orang dewasa tetapi juga bisa kita lihat anak usia belasan tahun bahkan wanita melakukan aktivitas merokok (Suharmanto, W. S. Utami, N. Pratiwi, and F. Muhammad, 2023). Banyak penyakit yang ditimbulkan dari mengonsumsi rokok, seperti penyakit kanker paru-paru, bronchitis kronik, jantung, gangguan janin dalam kandungan dan masih banyak lagi (M. H. Setiyani and S. Kristiyanto, 2024). Menurut data Badan Pusat Statistik (BPS) tahun 2024, persentase penduduk berusia 15 tahun ke atas yang merokok masih tinggi di banyak provinsi. Merokok pada usia produktif tidak hanya berdampak langsung pada kesehatan individu, tetapi juga berdampak terhadap produktivitas dan pembiayaan sektor kesehatan dalam jangka panjang.

Di sisi lain, untuk melihat sejauh mana keberhasilan pembangunan manusia United Development Programme (UNDP) telah mengeluarkan suatu indikator yaitu

Human Development Index atau Indeks Pembangunan Manusia (IPM). IPM atau Human Development Index (HDI) adalah suatu indeks yang digunakan untuk mengukur perkembangan kondisi sosial ekonomi di suatu wilayah berdasarkan kombinasi perhitungan komponen pendidikan, kesehatan, serta pendapatan per kapita yang disesuaikan. Peningkatan IPM dapat diartikan sebagai semakin baiknya ketiga komponen tersebut secara umum, dan berlaku sebaliknya untuk penurunan IPM (E. Harlan, 2023). IPM merupakan tolak ukur dalam pencapaian pembangunan manusia yang lebih berkualitas (L. S. Hasibuan, 2020). Oleh karena itu, terdapat potensi untuk mengkaji apakah nilai IPM suatu provinsi berkaitan erat dengan tingginya tingkat perokok.

Data mining adalah proses menemukan hubungan yang tidak terduga dalam data. Ini dilakukan dalam scatterplots yang terdiri dari data besar (N. Purwati, H. Kurniawan, and S. Karnila, 2025). Melalui pendekatan data mining, khususnya metode K-Means Clustering, provinsi-provinsi di Indonesia dapat dikelompokkan berdasarkan pola kombinasi antara tingkat merokok dan IPM. Selanjutnya, dengan menggunakan metode K-Nearest Neighbor (KNN), dilakukan klasifikasi untuk memprediksi kluster provinsi berdasarkan nilai IPM-nya. Kombinasi dua metode ini diharapkan dapat membantu dalam memahami pola persebaran risiko dan menjadi dasar bagi perumusan kebijakan kesehatan masyarakat berbasis data.

KAJIAN TEORITIS

Data mining merupakan proses eksplorasi data dalam jumlah besar untuk menemukan pola atau hubungan tersembunyi yang bermakna. Salah satu metode yang umum digunakan dalam data mining adalah K-Means Clustering, yang bekerja dengan mengelompokkan data ke dalam kluster berdasarkan kemiripan atribut. Metode ini telah banyak diterapkan dalam berbagai bidang, termasuk analisis kebiasaan merokok berdasarkan usia dan wilayah.

Penelitian sebelumnya oleh Suharmanto et al. (2023) menerapkan algoritma K-Means untuk mengelompokkan perokok berusia di atas 15 tahun berdasarkan data statistik kesehatan. Penelitian lain oleh Setiyani dan Kristiyanto (2024) menyoroti korelasi antara kebiasaan merokok dan angka kemiskinan di Pulau Jawa. Selain itu, Harlan (2023) mengidentifikasi bahwa HDI memiliki keterkaitan dengan konsumsi tembakau dan faktor sosial lainnya.

PENGELOMPOKAN PROVINSI MEROKOK DI INDONESIA BERDASARKAN IPM MENGGUNAKAN K-MEANS DAN KNN

Dalam konteks klasifikasi, metode K-Nearest Neighbor (KNN) digunakan untuk memprediksi kelompok atau klaster berdasarkan kemiripan data input terhadap data yang sudah terklasifikasi sebelumnya. Rukmana et al. (2022) menunjukkan bahwa algoritma KNN efektif dalam klasifikasi penyakit menggunakan pendekatan normalisasi dan seleksi fitur.

Dengan merujuk pada teori dan penelitian sebelumnya, studi ini mengintegrasikan metode K-Means dan KNN dalam konteks analisis kebiasaan merokok berdasarkan indikator pembangunan manusia, yang hingga kini masih terbatas dalam literatur nasional. Hal ini menjadikan penelitian ini memiliki urgensi dan nilai kebaruan dalam pemetaan risiko kebiasaan merokok di tingkat provinsi.

METODE PENELITIAN

Analisis Masalah

Penelitian ini menekankan bahwa prevalensi perokok yang tinggi di Indonesia, terutama di kelompok usia 15 tahun ke atas, memiliki dampak langsung maupun tidak langsung terhadap kemajuan negara, terutama dalam bidang kesehatan. Sikap yang muncul sebagai cerminan pribadinya yang masih labil dipengaruhi oleh berbagai kesulitan dalam penyesuaian diri. Seringkali, keinginan untuk diakui sebagai orang dewasa diikuti dengan meniru kebiasaan orang dewasa tanpa pemikiran yang matang (F. Almaidah et al, 2020). Indeks Pembangunan Manusia (IPM) adalah indikator yang dapat dikaitkan dengan perilaku merokok. Oleh karena itu, analisis data diperlukan untuk menentukan apakah IPM suatu provinsi berkorelasi dengan tingkat perokok di daerah tersebut.

Sumber Data

Data yang digunakan dalam penelitian ini diperoleh dari dua sumber utama, yaitu:

1. Badan Pusat Statistik (BPS): Menyediakan data persentase penduduk berumur 15 tahun ke atas yang merokok tembakau selama sebulan terakhir menurut provinsi pada tahun 2024.

Tabel 1. Sumber Data BPS

Provinsi	Persentase Merokok	IPM 2024
ACEH	28,61	74.03

SUMATERA UTARA	26,69	74.02
SUMATERA BARAT	31,45	74.49
RIAU	28,02	74.79
JAMBI	29,76	73.43
...	...	...
PAPUA SELATAN	32,28	67.90
PAPUA TENGAH	19,48	59.75
PAPUA PEGUNUNGAN	19,54	53.42

2. Sistem Informasi Manajemen Regional (SIMREG – Bappenas): Menyediakan data IPM (Indeks Pembangunan Manusia) provinsi di Indonesia tahun 2024.

Tabel 2. Sumber Data Bappenas

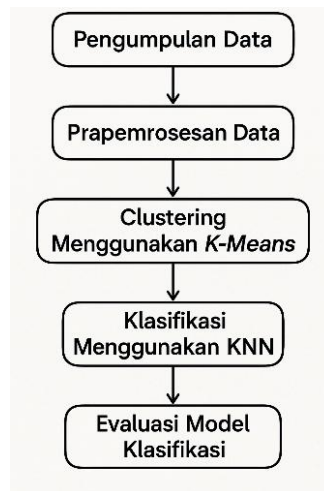
Provinsi	IPM 2024
ACEH	74.03
SUMATERA UTARA	74.02
SUMATERA BARAT	74.49
RIAU	74.79
JAMBI	73.43
...	...
PAPUA	73.00
PAPUA SELATAN	67.90
PAPUA TENGAH	59.75
PAPUA PEGUNUNGAN	53.42

Arsitektur Penelitian

Arsitektur penelitian ini disusun dalam beberapa tahapan utama yang diimplementasikan menggunakan perangkat lunak RapidMiner Studio. Perangkat lunak ini bersifat open source dan dibuat dengan menggunakan program Java yang dilisensikan dengan lisensi publik GNU. RapidMiner dapat digunakan di sistem operasi apa pun, dan bersifat open source (B. G. Sudarsono, M. I. Leo, A. Santoso, and F. Hendrawan, 2021).

PENGELOMPOKAN PROVINSI MEROKOK DI INDONESIA BERDASARKAN IPM MENGGUNAKAN K-MEANS DAN KNN

Proses kerja RapidMiner berbasis *workflow* visual sehingga tiap tahap dapat diatur melalui operator dan proses modular yang saling terhubung. Tahapan tersebut meliputi:



Gambar 1. Workflow Data Mining

Metode K-Means Clustering

K-Means adalah salah satu algoritma clustering yang paling populer dan sederhana. Algoritma ini bekerja dengan cara mengelompokkan data ke dalam jumlah kluster yang ditentukan sebelumnya (M. Rafi Nahjan, Nono Heryana, and Apriade Voutama, 2023). Algoritma K-Means Cluster dapat diterapkan dalam permasalahan Analisis data Perokok dengan rentang usia 15 tahun keatas, dapat mengidentifikasi peluang baru dari perokok di suatu provinsi.

Metode K-Nearest Neighbor (KNN)

Perhitungan algoritma KNN menggunakan jarak antara seluruh data training dan data testing, dan kemudian data terdekat atau data yang paling banyak memiliki kemiripan akan diambil untuk klasifikasi (S. Zulaikhah Hariyanti Rukmana, A. Aziz, and W. Harianto, 2022).

Rapidminer

Berdasarkan www.rapidminer.com, perangkat lunak rapidminer bertindak sebagai perancang proses visual untuk menganalisis ilmu data dan pembelajaran mesin pada tim dimulai dari analis sampai pakar. Perangkat lunak ini mudah digunakan dan bias menghimpun data dari semua sumber seperti database, media sosial, cloud, dokumen atau

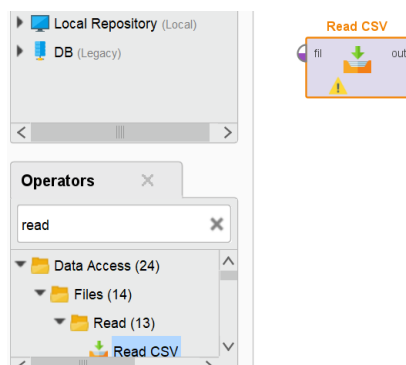
aplikasi bisnis. Selain itu, memungkinkan eksplorasi statistik dan visualisasi data. Berbagai model pembelajaran mesin dan model validasi tersedia (D. Zamri, 2022).

Implementasi

Implementasi penelitian dilakukan dengan menggunakan perangkat lunak RapidMiner Studio, yang mendukung proses data mining berbasis antarmuka visual (drag-and-drop). Alur proses yang dibangun mencerminkan tahapan mulai dari input data, praproses, pemodelan, hingga evaluasi. Berikut adalah langkah-langkah implementasi secara rinci:

1. Input Data

Dataset yang telah dikombinasikan antara persentase perokok dan IPM dimuat menggunakan operator Read CSV. Data telah melalui tahap prapemrosesan awal di luar RapidMiner, seperti penggabungan berdasarkan provinsi dan penggantian format angka desimal.



Gambar 2. Read CSV

2. Normalisasi Data

Dilakukan normalisasi menggunakan *operator Normalize* untuk menghindari bias skala antar atribut (misalnya IPM dan persentase merokok). Metode normalisasi yang digunakan adalah Z-Transformation, karena lebih sesuai untuk model berbasis jarak seperti K-Means dan KNN. Filter diterapkan hanya pada atribut numerik, sehingga atribut non-numerik tidak terpengaruh.

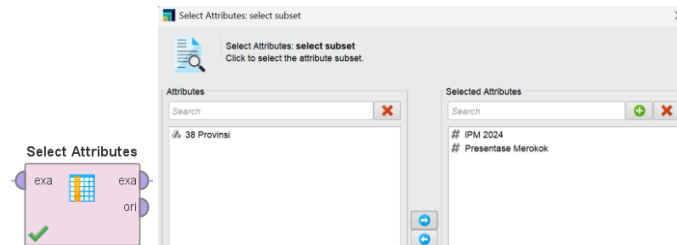


Gambar 3. Metode Normalisasi

PENGELOMPOKAN PROVINSI MEROKOK DI INDONESIA BERDASARKAN IPM MENGGUNAKAN K-MEANS DAN KNN

3. Pemilihan Atribut

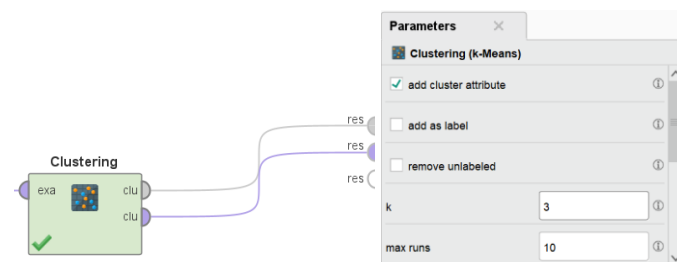
Operator Select Attributes digunakan untuk memilih hanya atribut-atribut relevan, yaitu IPM dan persentase merokok, sebelum dilakukan *clustering*. Hal ini bertujuan untuk menyederhanakan model dan menghindari noise dari data yang tidak relevan.



Gambar 4. Pemilihan Atribut

4. Clustering dengan K-Means

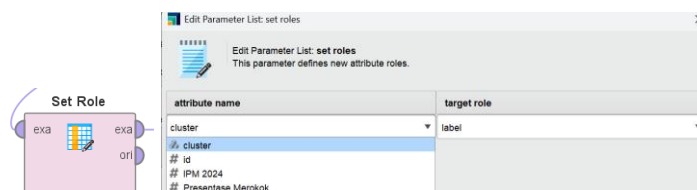
Operator Clustering (K-Means) digunakan untuk membagi provinsi ke dalam 3 kelompok berdasarkan IPM dan persentase merokok. Parameter jumlah kluster ditentukan secara manual sebanyak 3 kluster, yang diasumsikan sebagai tingkat risiko rendah, sedang, dan tinggi.



Gambar 5. Clustering dnegan K-Means

5. Penyetelan Peran Atribut (*Set Role*)

Operator Set Role digunakan untuk menjadikan hasil kluster (cluster ID) sebagai label klasifikasi. Atribut IPM tetap digunakan sebagai fitur utama dalam proses klasifikasi selanjutnya.

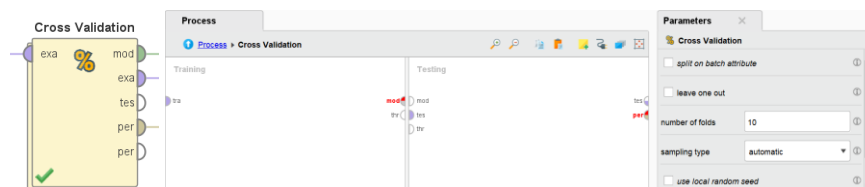


Gambar 6. Set Role untuk Penentuan Label Klasifikasi

6. Validasi Model dengan *Cross Validation*

Dengan membagi data menjadi dua kelompok, data proses pembelajaran dan data validasi/evaluasi, *Cross-Validation* (CV) adalah teknik statistik yang digunakan

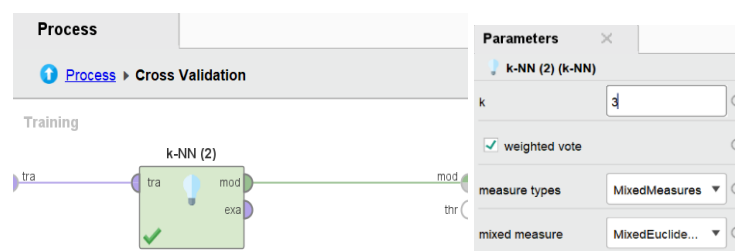
untuk mengevaluasi kinerja model atau algoritma [7]. Untuk menghindari *overfitting* dan mendapatkan estimasi performa yang lebih adil, digunakan operator *Cross Validation* dengan jumlah 10 folds. Proses dibagi menjadi dua bagian: Training: model KNN dibangun menggunakan nilai IPM. Testing: model diaplikasikan pada data validasi, kemudian hasil prediksi dibandingkan dengan label klaster.



Gambar 7. Cross Validation

7. Klasifikasi dengan *K-Nearest Neighbor*

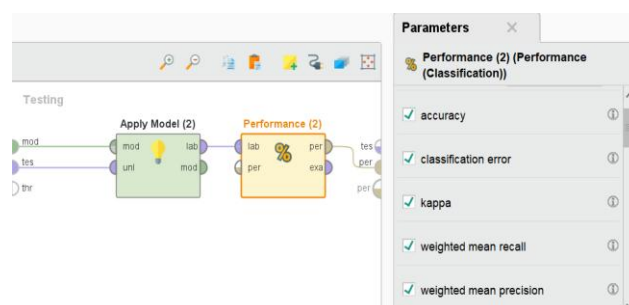
Operator KNN digunakan dalam bagian *training Cross-Validation*. Parameter K diset ke 3, dan Euclidean distance digunakan sebagai metrik jarak. Input utama adalah nilai IPM, sedangkan label target adalah hasil klaster dari K-Means.



Gambar 8. Klasifikasi dengan KNN

8. Aplikasi Model & Evaluasi

Operator Apply Model diterapkan pada data testing untuk menghasilkan prediksi klaster. *Operator Performance (Classification)* digunakan untuk mengukur akurasi dan performa model klasifikasi. Evaluasi dilakukan menggunakan metrik *accuracy*, *confusion matrix*, serta tambahan metrik seperti *precision* dan *recall*.



Gambar 9. Aplikasi Model & Evaluasi

PENGELOMPOKAN PROVINSI MEROKOK DI INDONESIA BERDASARKAN IPM MENGGUNAKAN K-MEANS DAN KNN

HASIL DAN PEMBAHASAN

Hasil Clustering Menggunakan K-Means

Proses *Clustering* dilakukan menggunakan dua atribut, yaitu persentase merokok dan IPM. Berdasarkan pengujian dengan jumlah klaster = 3, diperoleh hasil segmentasi sebagai berikut:

1. Klaster 0 (Risiko Rendah): Provinsi dengan persentase merokok rendah dan IPM tinggi.

Klaster ini menunjukkan bahwa daerah ini memiliki tingkat prevalensi merokok yang rendah, yang berarti bahwa hanya sebagian kecil dari populasinya yang aktif merokok. Provinsi ini dapat memberikan layanan dasar yang baik kepada masyarakatnya, seperti yang ditunjukkan oleh IPM yang tinggi.

Tabel 3. Hasil Cluster 0

Provinsi	Persentase Merokok	IPM 2024	Cluster
ACEH	28.61	74.03	0
SUMATERA UTARA	26.69	74.02	0
SUMATERA BARAT	31.45	74.49	0
RIAU	28.02	74.79	0
JAMBI	29.76	73.43	0
SUMATERA SELATAN	31.01	72.30	0
BENGKULU	32.96	73.39	0
LAMPUNG	33.84	71.81	0
KEP. BANGKA	27.97	73.33	0
BELITUNG	27.97	73.33	0
JAWA BARAT	32.98	74.43	0
JAWA TENGAH	29.13	73.88	0
JAWA TIMUR	28.72	74.09	0
BANTEN	31.31	74.48	0
NUSA TENGGARA BARAT	32.40	70.93	0
NUSA TENGGARA TIMUR	27.51	67.39	0
KALIMANTAN BARAT	28.06	70.13	0
KALIMANTAN TENGAH	28.02	72.73	0
KALIMANTAN UTARA	25.93	73.02	0
SULAWESI UTARA	26.93	75.03	0
SULAWESI TENGAH	28.98	71.56	0
GORONTALO	31.61	71.23	0
SULAWESI BARAT	26.99	68.20	0

MALUKU	29.43	71.57	0
MALUKU UTARA	29.72	71.03	0
PAPUA BARAT	25.90	67.02	0
PAPUA BARAT DAYA	25.00	68.63	0
PAPUA SELATAN	32.28	67.90	0

2. Klaster 1 (Risiko Sedang): Provinsi dengan persentase merokok sedang dan IPM sedang.

Klaster ini mencakup provinsi-provinsi yang berada di tengah-tengah, baik dari segi persentase perokok maupun tingkat IPM (Indeks Pembangunan Manusia). Provinsi dalam klaster ini tidak terlalu buruk, namun juga belum optimal dalam aspek pembangunan dan kesehatan.

Tabel 4. Hasil Cluster 1

Provinsi	Persentase Merokok	IPM 2024	Cluster
KEP. RIAU	25.46	77.97	1
DKI JAKARTA	22.56	83.08	1
DI YOGYAKARTA	25.18	81.55	1
BALI	19.22	77.76	1
KALIMANTAN SELATAN	23.07	73.03	1
KALIMANTAN TIMUR	23.99	78.83	1
SULAWESI SELATAN	24.66	74.05	1
SULAWESI TENGGARA	24.80	73.48	1
PAPUA	22.11	73.00	1

3. Klaster 2 (Risiko Tinggi): Provinsi dengan persentase merokok tinggi dan IPM rendah.

Dalam klaster ini, provinsi-provinsi ini menghadapi dua masalah utama: Rendahnya Indeks Pembangunan Manusia (IPM) dan tingkat perokok yang tinggi. Kondisi ini menunjukkan bahwa provinsi-provinsi tersebut memiliki risiko kesehatan dan pembangunan manusia yang tinggi.

Tabel 5. Hasil Cluster 1

Provinsi	Persentase Merokok	IPM 2024	Cluster
PAPUA TENGAH	19.48	59.75	2
PAPUA PEGUNUNGAN	19.54	53.42	2

PENGELOMPOKAN PROVINSI MEROKOK DI INDONESIA BERDASARKAN IPM MENGGUNAKAN K-MEANS DAN KNN

Distribusi provinsi di setiap klaster menunjukkan pola geografis tertentu; sebagian besar provinsi di Indonesia timur termasuk dalam klaster risiko tinggi. Hasil ini menunjukkan bahwa prevalensi perokok lebih tinggi di daerah dengan IPM yang lebih rendah.

Hasil Klasifikasi Menggunakan *K-Nearest Neighbor* (KNN)

Label klaster digunakan sebagai target (label) dalam proses klasifikasi setelah proses clustering selesai. Nilai IPM model KNN dilatih untuk memprediksi provinsi baru dengan menggunakan atribut IPM saja. Dengan menggunakan nilai $k = 3$, model KNN menunjukkan akurasi sebesar 94,17% untuk pengujian pada data yang sama (validasi hold-out). Hasil matrix confusion juga menunjukkan bahwa prediksi KNN sebagian besar sesuai dengan hasil klaster K-Means.

accuracy: 94.17% +/- 12.45% (micro average: 94.74%)

	true cluster_0	true cluster_1	true cluster_2	class precision
pred_cluster_0	27	0	2	93.10%
pred_cluster_1	0	9	0	100.00%
pred_cluster_2	0	0	0	0.00%
class recall	100.00%	100.00%	0.00%	

Gambar 10. Confusion Matrix

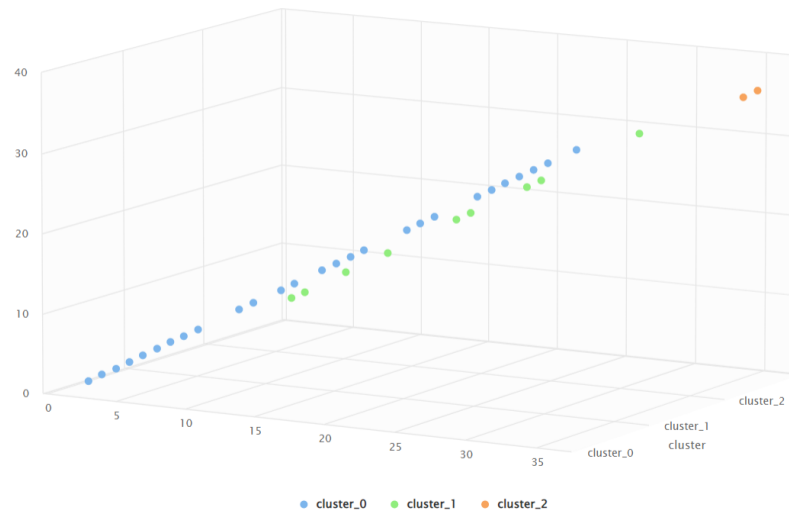
Interpretasi Hasil

Hasil klasifikasi yang tinggi mengindikasikan bahwa nilai IPM memiliki korelasi yang cukup kuat terhadap pengelompokan tingkat risiko merokok. Hal ini memperkuat hipotesis bahwa pembangunan manusia yang baik dalam hal pendidikan, kesehatan, dan daya beli dapat berkontribusi terhadap perilaku hidup sehat, termasuk penghindaran terhadap kebiasaan merokok.

Visualisasi

Visualisasi hasil clustering dan klasifikasi ditampilkan dalam bentuk scatter plot dan tabel perbandingan antar provinsi. Grafik menunjukkan pemisahan yang cukup jelas antar klaster, dengan IPM sebagai garis pembeda utama. Provinsi seperti DKI Jakarta dan Yogyakarta masuk dalam klaster rendah, sedangkan provinsi seperti Papua dan NTT masuk dalam klaster tinggi.

Dengan demikian, hasil penelitian ini mampu memberikan gambaran umum yang jelas terhadap segmentasi wilayah berdasarkan kebiasaan merokok dan indeks pembangunan, serta potensi pemanfaatan IPM sebagai indikator klasifikasi.



Gambar 11. Visualisasi Hasil dengan Scatter 3D

KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa terdapat hubungan yang kuat antara Indeks Pembangunan Manusia (HDI) dan tingkat prevalensi perokok di Indonesia. Pengelompokan provinsi menggunakan metode *K-Means Clustering* berhasil membentuk tiga klaster utama, yaitu klaster risiko rendah, sedang, dan tinggi, berdasarkan kombinasi nilai HDI dan persentase perokok usia 15 tahun ke atas. Selanjutnya, klasifikasi menggunakan metode *K-Nearest Neighbor* (KNN) dengan variabel tunggal HDI menunjukkan akurasi yang tinggi dalam memprediksi klaster risiko, yaitu sebesar 94,17%. Hal ini menunjukkan bahwa HDI memiliki keterkaitan signifikan terhadap pola kebiasaan merokok di tingkat provinsi dan dapat digunakan sebagai indikator prediktif dalam analisis kebijakan kesehatan masyarakat. Penelitian ini memberikan kontribusi terhadap pemetaan wilayah dengan risiko perilaku merokok tinggi berbasis *data-driven*, yang dapat dimanfaatkan oleh pembuat kebijakan dalam merancang program intervensi yang lebih tepat sasaran.

Namun, hasil ini perlu disikapi dengan hati-hati karena generalisasi terhadap seluruh populasi tidak sepenuhnya dapat dilakukan. Keterbatasan penelitian ini terletak pada jumlah variabel yang digunakan, yaitu analisis hanya didasarkan pada dua atribut

PENGELOMPOKAN PROVINSI MEROKOK DI INDONESIA BERDASARKAN IPM MENGGUNAKAN K-MEANS DAN KNN

utama, yaitu HDI dan persentase perokok. Oleh karena itu, argumen "tidak ada bukti bukan berarti tidak ada" harus diperhatikan, dan penelitian selanjutnya perlu memasukkan variabel lain yang relevan, seperti tingkat kemiskinan, pengangguran, akses layanan kesehatan, serta aspek budaya dan sosial yang dapat memengaruhi perilaku merokok.

Selain itu, implementasi sistem berbasis *dashboard* interaktif dari hasil analisis ini dapat dipertimbangkan untuk mendukung pengambilan keputusan secara *real-time* di tingkat pemerintah daerah. Uji banding menggunakan algoritma klasifikasi lainnya seperti *Random Forest*, *Naïve Bayes*, atau *Support Vector Machine* juga dapat menjadi alternatif untuk mengukur efektivitas dan konsistensi hasil klasifikasi. Dengan pendekatan yang lebih komprehensif, hasil penelitian di masa mendatang diharapkan dapat memberikan gambaran yang lebih mendalam mengenai faktor-faktor yang memengaruhi perilaku merokok di Indonesia.

UCAPAN TERIMA KASIH

Penulis berterima kasih kepada Badan Pusat Statistik (BPS), Sistem Informasi Manajemen Data Regional (SIMREG—Bappenas), dan portal data terbuka nasional data.go.id atas data yang digunakan dalam penelitian ini. Data yang akurat dan dapat diakses publik sangat penting untuk mendukung penelitian berbasis data. Selain itu, penulis mengucapkan terima kasih kepada para dosen pembimbing, institusi akademik, dan semua orang yang telah membantu, dan berkontribusi pada proses penyusunan jurnal ini.

DAFTAR REFERENSI

- B. G. Sudarsono, M. I. Leo, A. Santoso, and F. Hendrawan.(2021). "ANALISIS DATA MINING DATA NETFLIX MENGGUNAKAN APLIKASI RAPID MINER," JBASE - J. Bus. Audit Inf. Syst., vol. 4, no. 1, doi: 10.30813/jbase.v4i1.2729.
- D. Zamri.(2022). "Comparison of Data Mining Methods for Predition of Floods with Naïve Bayes and KNN Algorithm".
- E. Harlan.(2023). "Bagaimana Konsumsi Tembakau, IPM, Ketimpangan, Pendapatan dan Pengangguran Mempengaruhi Kemiskinan di Lampung?," Tirtayasa Ekon., vol. 18, no. 2, p. 22, doi: 10.35448/jte.v18i2.22007.

- F. Almaidah et al.(2020). “SURVEI FAKTOR PENYEBAB PEROKOK REMAJA MEMPERTAHANKAN PERILAKU MEROKOK,” J. Farm. Komunitas, vol. 8, no. 1, p. 20, doi: 10.20473/jfk.v8i1.21931.
- L. S. Hasibuan.(2020). “ANALISIS DETERMINAN INDEKS PEMBANGUNAN MANUSIA (IPM) DI INDONESIA,” vol. 5, no. 2.
- M. H. Setiyani and S. Kristiyanto.(2024). “ROKOK, KEBIASAAN MEROKOK DAN ANGKA KEMISKINAN DI PULAU JAWA,” OIKOS J. Kaji. Pendidik. Ekon. Dan Ilmu Ekon., vol. 8, no. 1, doi: 10.23969/oikos.v8i1.12534.
- M. Rafi Nahjan, Nono Heryana, and Apriade Voutama.(2023). “IMPLEMENTASI RAPIDMINER DENGAN METODE CLUSTERING K-MEANS UNTUK ANALISA PENJUALAN PADA TOKO OJ CELL,” JATI J. Mhs. Tek. Inform., vol. 7, no. 1, pp. 101–104, doi: 10.36040/jati.v7i1.6094.
- N. Purwati, H. Kurniawan, and S. Karnila.(2021). Data mining, vol. 1. Zahira Media Publisher. Accessed: Jul. 20, 2025. [Online]. Available: <https://books.google.com/books?hl=id&lr=&id=Q3NHEAAAQBAJ&oi=fnd&pg=PR5&dq=apa+itu+data+mining&ots=cN4aTPpHM2&sig=j8DGIt1mIDLSEcosD7DkyBeQjKg>.
- S. Zulaikhah Hariyanti Rukmana, A. Aziz, and W. Harianto.(2022) “OPTIMASI ALGORITMA K-NEAREST NEIGHBOR (KNN) DENGAN NORMALISASI DAN SELEKSI FITUR UNTUK KLASIFIKASI PENYAKIT LIVER,” JATI J. Mhs. Tek. Inform., vol. 6, no. 2, pp. 439–445, doi: 10.36040/jati.v6i2.4722.
- Suharmanto, W. S. Utami, N. Pratiwi, and F. Muhammad.(2023). “Penerapan Data Mining Menggunakan Algoritma K-Means Untuk Clustering Perokok Usia Lebih dari 15 Tahun,” Bull. Inf. Technol. BIT, vol. 4, no. 4, pp. 501–507, doi: 10.47065/bit.v4i4.1067.