ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

PENERAPAN ALGORITMA DBSCAN DALAM MENGIDENTIFIKASI RISIKO STROKE

Hani Istiqomah✉, Khoirun Nisa, Arif Setia Sandi A
Universitas Harapan Bangsa, Banyumas, Indonesia
Email: istiqomahhani724@gmail.com

DOI: https://doi.org/10.46880/jmika.Vol9No2.pp340-349

ABSTRACT

Stroke is a chronic illness that frequently results in death and permanent disability. Using a publicly accessible dataset from Kaggle that contains 5,110 patient records, this study investigates the application of the DBSCAN algorithm to group people according to their risk of stroke. Age, gender, hypertension, history of heart disease, body mass index (BMI), blood sugar level, and smoking status are among the clinical and demographic characteristics included in the dataset. A number of preprocessing techniques were used, including Z-score normalization, categorical data encoding, median imputation for missing BMI values, and dimensionality reduction with PCA for visualization. Using the k-distance plot in conjunction with the Silhouette Score evaluation, DBSCAN's parameters were tuned, yielding ε = 2.5, min_samples = 3, and a score of 0.2158. Although additional feature selection and parameter modification are still required to increase clustering quality, the results demonstrate DBSCAN's promise in identifying groups with comparable stroke risk profiles.

Keywords: Clustering, DBSCAN Algorithm, PCA, Risk Factors, Stroke.

ABSTRAK

Stroke adalah salah satu penyakit jangka panjang yang sering menyebabkan kecacatan permanen atau kematian. Fokus penelitian ini adalah penggunaan algoritma DBSCAN untuk mengelompokkan tingkat risiko stroke dengan menggunakan dataset publik Kaggle 5.110 pasien. Informasi demografis dan klinis seperti usia, jenis kelamin, hipertensi, riwayat penyakit jantung, indeks massa tubuh (BMI), glukosa, dan kebiasaan merokok dimasukkan ke dalam dataset. Praproses termasuk pengkodean data kategorikal, normalisasi dengan Z-score, imputasi median pada variabel BMI, dan reduksi dimensi menggunakan PCA untuk memudahkan visualisasi. Analisis grafik k-distance dan pengujian skor Silhouette digunakan untuk menentukan parameter DBSCAN. Hasilnya adalah nilai ε = 2,5 dan min_samples = 3, dengan skor 0,2158. Hasil menunjukkan bahwa algoritma DBSCAN dapat membantu proses identifikasi kelompok risiko stroke. Namun, kualitas pengelompokan masih dapat ditingkatkan dengan mengubah parameter dan memilih fitur yang lebih baik.

Kata Kunci: Clustering, Algoritma DBSCAN, PCA, Faktor Risiko, Stroke.

PENDAHULUAN

Stroke termasuk penyakit tidak menular yang masih menjadi penyebab utama kematian serta disabilitas, baik di dunia maupun di Indonesia. Data dari *World Health Organization* (WHO) menunjukkan jutaan kasus baru terjadi setiap tahun, dengan banyak penyintas yang harus menghadapi dampak jangka panjang, baik fisik maupun psikologis. Seiring dengan perubahan pola hidup, kebiasaan makan yang kurang sehat, dan bertambahnya usia harapan hidup masyarakat, tren kasus stroke di Indonesia terus meningkat (Ovyawan Herlistiono & Violina, 2024).

Penelitian sebelumnya (Venketasubramanian et al., 2022) menemukan bahwa prevalensi stroke di Indonesia sangat tinggi, dengan 0,0017% di daerah pedesaan dan 0,022% di daerah perkotaan. Menurut Riset Kesehatan Dasar (RISKESDAS), prevalensi meningkat dari 7 per 1.000 penduduk pada 2013 menjadi 10,9 per 1.000 penduduk pada 2018. Kondisi ini menempatkan Indonesia sebagai salah satu negara dengan tingkat kematian akibat stroke tertinggi di Asia Tenggara. Secara global, Amerika Serikat melaporkan sekitar 795.000 kasus stroke baru setiap tahunnya, dengan lebih dari 134.000 di antaranya berakhir pada kematian. Fakta tersebut menggarisbawahi bahwa stroke masih menjadi tantangan kesehatan serius yang perlu diantisipasi melalui deteksi dini serta strategi pencegahan.

Upaya pencegahan dan penanganan stroke sangat bergantung pada kemampuan untuk mendeteksi risikonya sejak dini. Namun, kompleksitas data medis yang melibatkan banyak variabel membuat proses analisis tidak sederhana (Hu et al., 2024). Sebagian besar penelitian terdahulu berfokus pada pendekatan

supervised learning untuk prediksi, sementara penerapan *unsupervised clustering* masih terbatas. Algoritma DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) dipandang relevan karena mampu membentuk klaster dengan bentuk yang bervariasi, tahan terhadap *noise*, serta tidak memerlukan penentuan jumlah klaster sejak awal.

Perkembangan teknologi komputer telah membuka jalan bagi pemanfaatan *data mining* sebagai alternatif dalam menganalisis data medis yang semakin kompleks. Salah satu metode yang cukup banyak digunakan untuk mengelompokkan data tanpa label adalah algoritma DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) (Kim et al., 2022). Algoritma ini dinilai relevan karena mampu menyesuaikan diri dengan karakteristik data kesehatan yang umumnya beragam, termasuk pada kasus stroke.

Data terkait stroke biasanya tidak seimbang, sering kali dipenuhi oleh *noise* pada variabel klinis maupun demografis, serta memperlihatkan pola hubungan antaratribut yang tidak linear. Kondisi semacam ini cenderung menyulitkan metode klasterisasi konvensional yang membutuhkan asumsi bentuk tertentu atau jumlah klaster sejak awal. DBSCAN hadir dengan beberapa kelebihan, di antaranya dapat membentuk klaster dengan bentuk bervariasi, cukup tangguh terhadap *noise*, serta tidak bergantung pada penentuan jumlah klaster. Dengan karakteristik tersebut, DBSCAN dianggap lebih sesuai untuk mengolah data kesehatan yang kompleks dan heterogen, termasuk data risiko stroke (Ma et al., 2023).

Penelitian ini berupaya menerapkan DBSCAN untuk mengelompokkan risiko stroke berdasarkan data terbuka. Dengan memanfaatkan atribut kesehatan seperti tekanan darah, kadar glukosa, BMI, dan riwayat penyakit, hasil pengelompokan diharapkan dapat mendukung analisis lebih lanjut dan menjadi acuan dalam pengambilan keputusan medis, khususnya pada tahap deteksi dini serta pencegahan stroke (Ding et al., 2022).

TINJAUAN PUSTAKA

Berbagai penelitian sebelumnya telah menyoroti penggunaan algoritma DBSCAN di bidang kesehatan dan menghasilkan temuan yang cukup menjanjikan. Misalnya, studi yang dilakukan oleh Puspitasari et al. (2023) menunjukkan bahwa DBSCAN dapat mengelompokkan data pasien secara efektif tanpa memerlukan jumlah klaster yang ditentukan sejak awal, serta mampu mengenali keberadaan *outlier*. Penelitian tersebut menghasilkan dua klaster dengan nilai Davies-Bouldin Index sebesar 1,422. Nilai ini mengindikasikan bahwa kualitas klasterisasi cukup baik sehingga dapat mendukung pengelolaan sumber daya rumah sakit secara lebih efisien. Namun demikian, fokus penelitian tersebut masih terbatas pada konteks manajemen rumah sakit dan belum diarahkan secara spesifik pada analisis risiko stroke.

Potensi DBSCAN dalam bidang diagnosis juga ditunjukkan oleh penelitian Santoso & Syafrianto (2019). Mereka mengombinasikan DBSCAN dengan pendekatan *Case-Based Reasoning* (CBR) untuk mendeteksi hipertensi. Dalam sistem tersebut, DBSCAN berperan sebagai metode *indexing* yang mempercepat pencarian kasus serupa, dan hasilnya terbukti meningkatkan akurasi hingga 100% menggunakan Minkowski distance. Walaupun penelitian ini tidak menyoroti stroke, temuan tersebut memperlihatkan bahwa DBSCAN dapat dimanfaatkan dalam pengolahan data medis berbasis faktor risiko, sehingga relevan bagi penelitian sejenis.

Dari perspektif spasial, Hermanto & Sunandar (2020) menerapkan DBSCAN untuk memetakan distribusi penyakit di salah satu Puskesmas dan berhasil meningkatkan akurasi pemetaan. Peningkatan ini bermanfaat dalam merancang program penyuluhan maupun intervensi medis. Studi lain oleh Sihite et al. (2024) mengembangkan Sistem Informasi Geografis berbasis DBSCAN untuk memetakan kasus gizi buruk di Kota Medan. Hasilnya menunjukkan nilai *Silhouette Index* sebesar 0,5414 dan *Dunn Index* sebesar 0,5124. Indikator tersebut membantu Dinas Kesehatan dalam melakukan pemantauan dan pengambilan keputusan. Walaupun topik kedua penelitian ini berbeda dari stroke, penerapan DBSCAN pada data spasial menegaskan fleksibilitas algoritma ini dalam berbagai konteks kesehatan.

Selain DBSCAN, penelitian lain juga banyak menggunakan algoritma berbasis *supervised learning*. Sari et al. (2024) melaporkan bahwa Random Forest mampu mencapai akurasi 98,58% dalam prediksi stroke, sedangkan Neural Network dan SVM masing-masing mencatat akurasi 95,72% dan 94,11%. Penelitian Wijaya et al. (2024) bahkan menunjukkan bahwa *Voting Classifier* dapat menghasilkan akurasi 98,5% dengan nilai AUC 0,99. Tidak hanya itu, pendekatan *hybrid ensemble* juga terbukti efektif, dengan capaian akurasi 97,2% serta F1-score 97,15% (Islam et al., 2025).

Meskipun berbagai studi tersebut membuktikan efektivitas metode *supervised*, pendekatan *unsupervised clustering* khususnya DBSCAN masih jarang dieksplorasi dalam analisis risiko stroke menggunakan dataset publik. Kebanyakan penelitian

yang ada masih terbatas pada tahap prototipe dan belum disertai validasi empiris yang kuat. Oleh karena itu, penelitian ini berupaya mengisi celah tersebut dengan menerapkan DBSCAN secara khusus untuk menemukan pola risiko stroke pada data terbuka, sehingga dapat memberikan kontribusi baru dalam pengembangan analisis kesehatan berbasis data.

Kebaruan penelitian ini adalah penggunaan algoritma DBSCAN untuk mengidentifikasi pola risiko stroke dengan data publik yang heterogen. Studi ini menggunakan metode *clustering* tanpa supervisi untuk mempelajari struktur data pasien tanpa label. Ini berbeda dengan penelitian sebelumnya yang biasanya berfokus pada pendekatan pembelajaran yang diawasi untuk prediksi. Oleh karena itu, penelitian ini memberikan kontribusi ilmiah tentang cara menggunakan metode berbasis kepadatan untuk mengidentifikasi pola risiko stroke yang tersembunyi. Ini dapat menjadi dasar untuk pengembangan sistem deteksi dini berbasis data di bidang kesehatan.

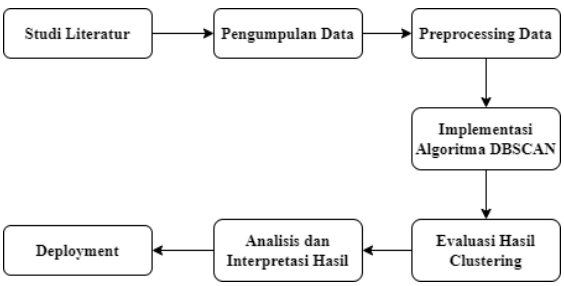
METODE PENELITIAN

Dataset

Penelitian ini menggunakan Stroke *Prediction Dataset* yang tersedia secara terbuka di platform Kaggle. Dataset tersebut berisi 5.110 catatan individu yang mencakup informasi demografis, riwayat medis, serta sejumlah faktor yang diperkirakan berhubungan dengan risiko stroke. Beberapa variabel penting di dalamnya antara lain usia, jenis kelamin, riwayat hipertensi, penyakit jantung, indeks massa tubuh (BMI), kadar glukosa, serta kebiasaan merokok.

Tahap Penelitian

Alur penelitian ditunjukkan pada Gambar 1, yang memperlihatkan rangkaian proses dalam penerapan algoritma DBSCAN untuk menganalisis faktor risiko Stroke.



Gambar 1. Alur Tahapan Penelitian

Secara garis besar, setiap tahap dapat dijelaskan sebagai berikut:

a. Studi Literatur
Kajian pustaka dilakukan untuk memahami penerapan DBSCAN dalam analisis data kesehatan, khususnya dalam pengelompokan pasien berdasarkan tingkat risiko stroke. Pada tahap ini juga dibahas metode evaluasi, salah satunya *Silhouette Score*, yang digunakan untuk menilai kualitas hasil klasterisasi.

b. Pengumpulan Data
Dataset diperoleh dari platform Kaggle dengan total 5.110 entri. Data tersebut mencakup atribut demografis dan klinis yang relevan, seperti usia, jenis kelamin, riwayat hipertensi, kondisi jantung, status pernikahan, jenis pekerjaan, tempat tinggal, kadar glukosa, indeks massa tubuh, serta kebiasaan merokok.

c. Preprocessing Data
Tahap prapemrosesan data dilakukan secara bertahap agar dataset siap digunakan dalam analisis. Adapun langkah-langkah yang ditempuh adalah sebagai berikut:
1. Penanganan data kosong (*missing values)*: Variabel BMI memiliki sebagian nilai yang hilang. Untuk mengatasinya, digunakan metode *median imputation* karena lebih tahan terhadap keberadaan *outlier* dibandingkan dengan mean imputation.
2. Normalisasi data numerik: Variabel numerik seperti usia, kadar glukosa rata-rata, dan BMI kemudian dinormalisasi menggunakan *StandardScaler* dengan pendekatan Z-score. Proses ini bertujuan agar semua fitur berada dalam skala yang sebanding sehingga tidak ada satu variabel pun yang mendominasi analisis.
3. Pengkodean variabel kategorikal: Atribut kategorikal diproses dengan dua cara: variabel dengan jumlah kategori sedikit (gender, status pernikahan, dan tipe tempat tinggal) dikodekan dengan *label encoding*, sementara variabel dengan kategori lebih banyak (jenis pekerjaan dan kebiasaan merokok) diproses menggunakan *one-hot encoding*.
4. Reduksi dimensi: Metode PCA digunakan untuk menyederhanakan sepuluh fitur awal menjadi dua komponen utama untuk mempermudah visualisasi hasil klasterisasi. Ini dilakukan tanpa menghilangkan informasi penting yang terkandung dalam data.

d. Implementasi Algoritma DBSCAN
Pada tahap prapemrosesan, metode median imputation digunakan untuk menghitung nilai kosong pada variabel BMI karena lebih tahan terhadap outlier. Selain itu, ada dua cara untuk

memproses data kategorikal. Label digunakan untuk variabel biner seperti jenis kelamin, status pernikahan, dan tipe tempat tinggal, sedangkan satu-hot encoding digunakan untuk variabel nominal yang memiliki banyak kategori, seperti jenis pekerjaan dan kebiasaan merokok.

Setelah data diproses, algoritma DBSCAN digunakan untuk membentuk kelompok pasien berdasarkan karakteristik yang sebanding. Grafik k-distance digunakan untuk menentukan parameter, dengan titik elbow menunjukkan nilai epsilon ideal pada 2,5. Namun demikian, jumlah minimum tetangga (min_samples) ditetapkan menjadi tiga, berdasarkan hasil evaluasi *Silhouette Score*, di mana nilai tertinggi adalah 0.2158.

e. Evaluasi Hasil *Clustering*
Selanjutnya, skor *Silhouette* digunakan untuk menilai kualitas hasil klasterisasi. Ini digunakan untuk mengevaluasi sejauh mana kelompok yang terbentuk dapat berbeda, terutama dalam hal pasien yang memiliki risiko stroke yang lebih tinggi dibandingkan dengan kelompok lain.

f. Analisis dan Interpretasi Hasil
Hasil klasterisasi menunjukkan pola hubungan antara sejumlah variabel kesehatan, termasuk usia, hipertensi, kebiasaan merokok, dan kadar gula darah. Pola ini kemudian dianalisis lebih lanjut untuk memberikan gambaran tentang risiko stroke dan untuk memberikan rekomendasi pencegahan. Hasil ini juga dibandingkan dengan hasil penelitian sebelumnya untuk validasi ilmiah.

g. Deployment
Pada langkah terakhir, model DBSCAN ini digunakan untuk membuat aplikasi web sederhana. Dengan menggunakan aplikasi ini, tenaga medis dapat memasukkan data klinis pasien dan mendapatkan hasil pengelompokan dengan cepat. Antarmuka dirancang agar mudah digunakan dan membantu pengambilan keputusan medis dengan cepat.

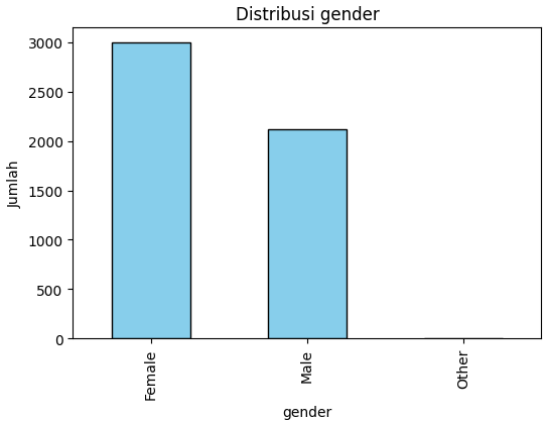
HASIL DAN PEMBAHASAN

Pengumpulan Data

Dataset yang digunakan dalam penelitian ini berasal dari Kaggle dan berisi 5.110 entri individu dengan informasi demografis serta riwayat kesehatan. Variabel-variabel tersebut kemudian dianalisis untuk melihat karakteristik masing-masing.

1. Id
Kolom ini hanya berfungsi sebagai nomor identifikasi unik. Karena tidak memiliki kaitan langsung dengan risiko stroke, variabel ini tidak dilibatkan dalam analisis lebih lanjut.

2. Gender (Jenis Kelamin)
Proporsi responden terdiri dari 59% perempuan dan 41% laki-laki. Faktor jenis kelamin dapat memengaruhi risiko stroke, antara lain melalui perbedaan hormonal maupun pola hidup.



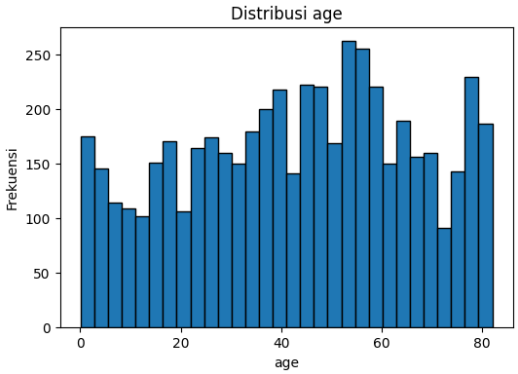
Gambar 2. Distribusi Gender

3. Age (Usia)
Usia menjadi salah satu variabel penting dalam penelitian ini. Secara umum, risiko stroke meningkat seiring bertambahnya umur akibat perubahan kondisi pembuluh darah.
Analisis deskriptif menunjukkan bahwa rata-rata usia responden adalah 43,23 tahun, dengan median 45 tahun. Mereka juga memiliki kadar glukosa darah rata-rata 106,15 dan nilai BMI rata-rata 28,89, menunjukkan bahwa sebagian besar responden berada pada kelompok usia dewasa, dengan kecenderungan memiliki kadar glukosa dan BMI yang lebih tinggi.

Tabel 1. Tabel Statistik Deskriptif

Variabel	Mean	Median	SD	Min	Max
Age	43.23	45.00	22.61	0.08	82.00
Avg_glucose-level	106.15	91.88	45.28	55.12	271.74
BMI	28.89	28.10	7.85	10.30	97.60

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)



Gambar 3. Histogram Distribusi Usia

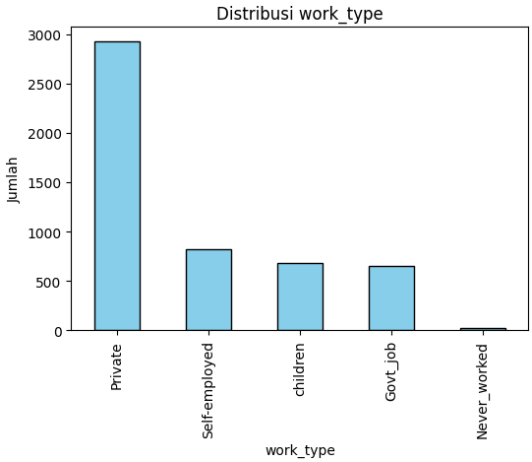
Jumlah responden tersebar dari anak-anak hingga orang dewasa, dengan kelompok usia 40 hingga 60 tahun memiliki jumlah tertinggi. Studi sebelumnya menunjukkan bahwa risiko stroke meningkat secara signifikan pada orang paruh baya dan lanjut usia.

4. Hypertension (Hipertensi)
   Sebanyak 9,74% responden tercatat memiliki riwayat hipertensi. Kondisi ini menjadi salah satu faktor risiko utama stroke karena tekanan darah tinggi dapat merusak pembuluh darah di otak.

5. Heart Disease (Penyakit Jantung)
   Sekitar 5,4% responden yang menjawab memiliki riwayat penyakit jantung. Gangguan jantung dapat mempengaruhi aliran darah ke otak, yang meningkatkan risiko stroke.

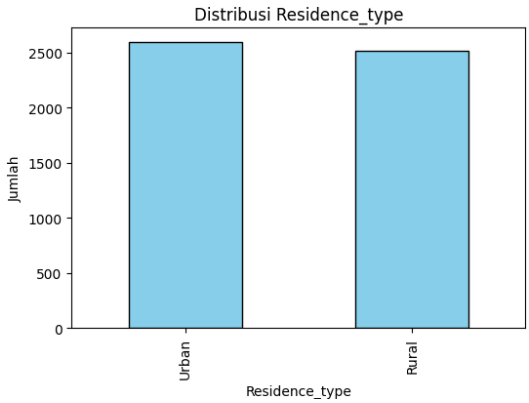
6. Ever Married (Status Pernikahan)
   Dari keseluruhan data, 66% responden diketahui pernah menikah. Status pernikahan dianggap memiliki keterkaitan tidak langsung dengan kesehatan, misalnya melalui tingkat stres atau pola gaya hidup.

7. Work Type (Jenis Pekerjaan)
   Mayoritas responden bekerja di sektor swasta (57,24%), kemudian diikuti oleh wiraswasta, pekerjaan di instansi pemerintah, anak-anak, serta kelompok yang tidak bekerja. Distribusi ini dapat memberikan gambaran mengenai variasi gaya hidup dan tingkat stres yang berhubungan dengan risiko stroke.



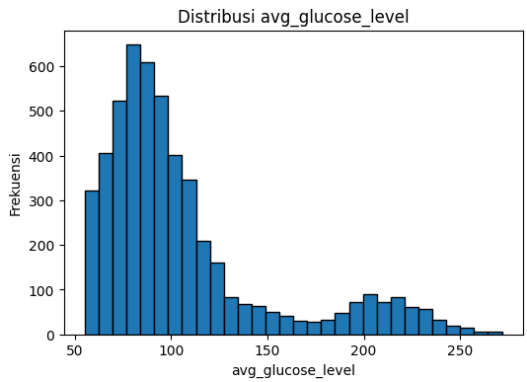
Gambar 4. Distribusi Jenis Pekerjaan

8. Residence Type
   Gambar 5 menunjukkan bahwa komposisi tempat tinggal hampir seimbang, yaitu 51% di perkotaan dan 49% di pedesaan. Lingkungan tempat tinggal memengaruhi akses layanan kesehatan dan gaya hidup.



Gambar 5. Distribusi Residence Type

9. Avg Glucose Level
   Variabel ini menunjukkan kadar gula darah rata-rata. Glukosa tinggi berkaitan erat dengan diabetes, yang juga merupakan faktor risiko stroke.



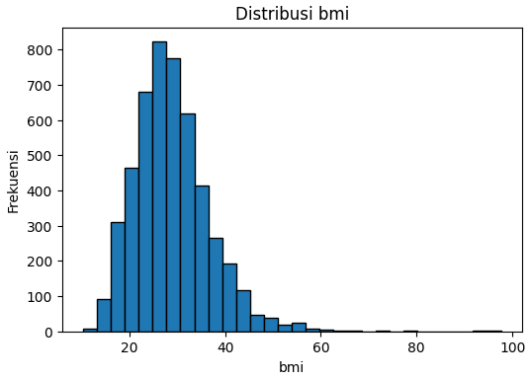
Gambar 6. Distribusi Rata-Rata Kadar Glukosa

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

Sebagian besar responden memiliki kadar glukosa antara 70 dan 120 poin, yang masih dapat dianggap normal, tetapi ada beberapa kelompok yang mencapai nilai lebih dari 200 poin, yang cukup menonjol. Pola ini menunjukkan bahwa banyak responden memiliki kadar glukosa tinggi, yang mungkin terkait dengan risiko diabetes dan peningkatan risiko stroke.

10. BMI (*Body Mass Index*)

BMI digunakan untuk menilai status berat badan seseorang. Individu dengan nilai BMI ≥30 termasuk kategori obesitas, yang diketahui dapat memperbesar risiko stroke melalui pengaruhnya terhadap tekanan darah dan metabolisme tubuh.

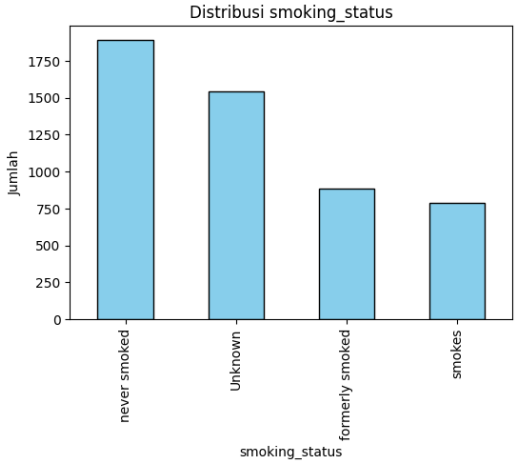


Gambar 7. Distribusi BMI

Mayoritas responden berada pada kisaran BMI 20–35 dengan konsentrasi tertinggi di kelompok *overweight*. Hanya sebagian kecil responden yang memiliki nilai ekstrem di atas 40, sehingga pola distribusi ini menunjukkan dominasi kategori berat badan berlebih ketimbang obesitas morbid.

11. Smoking Status (Status Merokok)

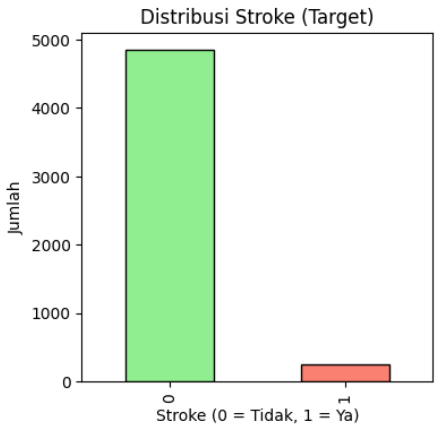
Gambar 8 menunjukkan hasil distribusi menunjukkan mayoritas responden tidak pernah merokok. Kategori berikutnya adalah responden dengan status merokok yang tidak diketahui, kemudian yang pernah merokok, dan terakhir yang masih aktif merokok. Kebiasaan merokok berhubungan erat dengan peningkatan tekanan darah serta kerusakan pembuluh darah, sehingga menjadi salah satu faktor penting dalam analisis risiko stroke.



Gambar 8. Distribusi Smoking Status

12. Stroke

Dari total 5.110 data, sebanyak 249 individu (sekitar 4,87%) tercatat pernah mengalami stroke.



Gambar 9. Distribusi Stroke

Distribusi ini menggambarkan bahwa kasus stroke memang bukan mayoritas dalam dataset, namun tetap signifikan sebagai kelompok yang perlu diperhatikan dalam analisis risiko.

Preprocesing Data

Tahap prapemrosesan dilakukan untuk memastikan data siap dianalisis. Beberapa langkah yang ditempuh dapat dijelaskan sebagai berikut:

1. Penghapusan kolom yang tidak relevan

Kolom ID diabaikan karena hanya berfungsi sebagai nomor identitas unik dan tidak memiliki nilai analitis. Kolom *stroke* juga tidak dilibatkan dalam data latih karena penelitian ini menggunakan pendekatan *unsupervised learning* yang tidak memerlukan label.

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

2. Penanganan nilai kosong (missing values)
   Dari seluruh variabel, hanya BMI yang memiliki nilai kosong sebesar 3,93%. Karena proporsinya masih di bawah 5%, digunakan metode *median imputation* untuk mengganti data yang hilang. Teknik ini dipilih karena lebih stabil terhadap keberadaan *outlier* dibandingkan *mean imputation*.

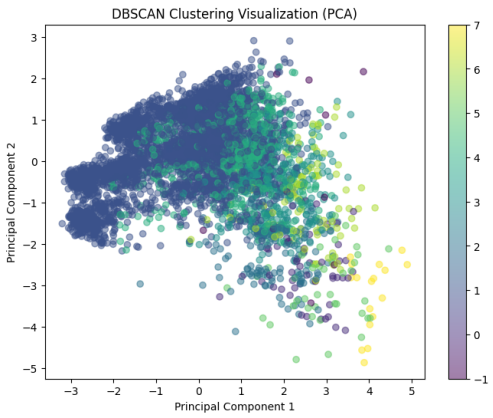
Tabel 2. Persentase Missing Values

Kolom	Sebelum (%)	Sesudah (%)
Gender	0.000	0.000
Age	0.000	0.000
Hypertension	0.000	0.000
heart_disease	0.000	0.000
ever_married	0.000	0.000
work_type	0.000	0.000
Residence_type	0.000	0.000
avg_glucose_level	0.000	0.000
bmi	3.933	0.000
smoking_status	0.000	0.000

3. Pengkodean variabel kategorikal
   Variabel kategorikal dengan jumlah kategori sedikit, seperti gender, status pernikahan, dan tipe tempat tinggal, diproses menggunakan *label encoding*. Sementara itu, variabel dengan kategori lebih banyak, misalnya jenis pekerjaan dan kebiasaan merokok, dikodekan dengan *one-hot encoding* agar tidak menimbulkan urutan semu di antara kategori.

4. Normalisasi variabel numerik
   Variabel numerik seperti usia, BMI, dan kadar glukosa dinormalisasi menggunakan *StandardScaler* berbasis Z-score. Proses ini memastikan setiap variabel berada pada skala yang sebanding, sehingga tidak ada fitur yang mendominasi perhitungan jarak dalam proses klasterisasi.

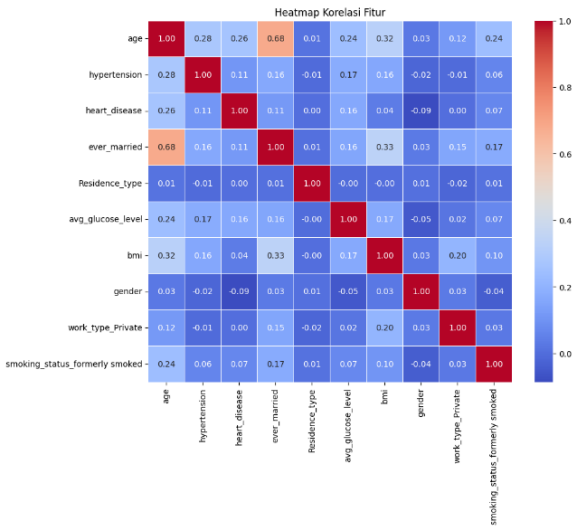
5. Fitur Ekstraksi
   Gambar 10 menunjukkan hasil visualisasi DBSCAN setelah PCA. Warna titik mewakili klaster berbeda, sedangkan titik dengan label -1 menunjukkan *outlier*. PCA tidak mengubah struktur klaster, tetapi membantu memperjelas pola hubungan antar data secara visual.



Gambar 10. Visualisasi PCA

Karena dataset awalnya terdiri dari sepuluh fitur, proses klasterisasi sulit divisualisasikan. Untuk itu, digunakan Principal Component Analysis (PCA) sehingga fitur dapat direduksi menjadi dua komponen utama. Dengan cara ini, pola klaster dapat divisualisasikan lebih jelas tanpa banyak kehilangan informasi penting.

6. Seleksi Fitur
   Analisis korelasi kemudian dilakukan untuk menentukan fitur yang paling relevan dalam pembentukan klaster.



Gambar 11. Heatmap Korelasi Fitur

Hasilnya menunjukkan bahwa:

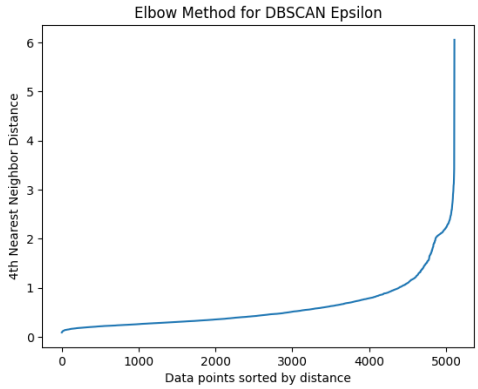
a. Usia memiliki korelasi cukup kuat dengan status pernikahan (0,68).
b. Variabel *smoking_status_formerly_smoked* menunjukkan korelasi tinggi (0,78) terhadap hasil klasterisasi.
c. Hipertensi juga berhubungan cukup signifikan (0,52).

ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

Fitur-fitur tersebut diprioritaskan karena kontribusinya dianggap paling berpengaruh dalam membentuk struktur klaster.

Implementasi Algoritma DBSCAN

Penentuan parameter dilakukan menggunakan grafik k-distance, di mana titik *elbow* muncul pada nilai epsilon sekitar 2,5. Parameter *min_samples* kemudian diuji menggunakan *Silhouette Score*, dan nilai tiga dipilih karena memberikan skor tertinggi, yakni 0,2158.



Gambar 12. Grafik *Elbow* Method

Evaluasi Hasil *Clustering*

Kualitas klasterisasi dievaluasi dengan *Silhouette Score*. Nilai 0,2158 menunjukkan pemisahan antar klaster belum optimal, namun sudah cukup untuk memberikan gambaran awal mengenai struktur data. Evaluasi ini masih terbatas pada satu metrik dan belum melibatkan ukuran lain seperti Davies–Bouldin Index atau Calinski–Harabasz Index, maupun validasi eksternal berbasis label stroke. Oleh sebab itu, hasil klasterisasi ini lebih tepat dipandang sebagai prototipe untuk analisis lanjutan.

Analisis Dan Interpretasi Hasil

Hasil penerapan DBSCAN menunjukkan adanya sejumlah klaster yang merepresentasikan kelompok dengan tingkat risiko stroke yang berbeda-beda. Pola ini membantu membedakan individu dengan risiko tinggi misalnya pasien usia lanjut dengan hipertensi dan kadar glukosa tinggi dari kelompok dengan kondisi kesehatan yang relatif baik.

Selain itu, DBSCAN juga berhasil mendeteksi *outlier*, yaitu individu dengan karakteristik yang berbeda jauh dari mayoritas. Keberadaan *outlier* penting untuk dicermati karena sering kali menggambarkan kasus khusus yang memerlukan perhatian lebih lanjut. Secara umum, temuan ini menunjukkan bahwa DBSCAN dapat digunakan sebagai pendekatan awal dalam membangun sistem pendukung keputusan medis berbasis *unsupervised learning*, terutama untuk deteksi dini dan upaya pencegahan Stroke.

Tabel 3. Tabel Perbandingan Karakteristik *Cluster*

Cluster	**Usia (rata-rata)**	**Hipertensi (%)**	**Penyakit Jantung (%)**	***BMI* (rata-rata)**	***Glukosa* (rata-rata)**	***Ever Married* (%)**	**Interpretasi Risiko**
-1 (Noise)	53	100	100	53	150	80	Individu dengan kondisi ekstrem, berisiko tinggi
0	67	0	100	29	135	90	Risiko penyakit jantung tinggi tanpa hipertensi
1	37	0	0	28	101	20	Kondisi kesehatan baik, risiko rendah
2	67	0	100	29	130	85	Serupa *Cluster* 0, glukosa sedikit lebih rendah
3	60	100	0	32	120	50	Risiko hipertensi tinggi, potensi gangguan metabolik
4	52	0	0	27	105	85	Paruh baya sehat, risiko rendah
5	70	100	100	31	144	80	Kondisi kesehatan buruk, risiko tinggi
6	64	100	0	32.5	132	75	Hipertensi dominan, risiko sedang-tinggi
7	71	100	100	33	181	90	Risiko kesehatan tertinggi (hipertensi + penyakit jantung + glukosa tinggi)

Beberapa kelompok menunjukkan pola risiko kesehatan yang berbeda setelah DBSCAN digunakan. Cluster 1 terdiri dari orang muda yang relatif sehat, tanpa hipertensi atau penyakit jantung, dan nilai BMI dan kadar glukosa rendah. Cluster 0 dan 2 terdiri dari orang lanjut usia yang memiliki riwayat penyakit jantung tetapi tidak memiliki hipertensi.

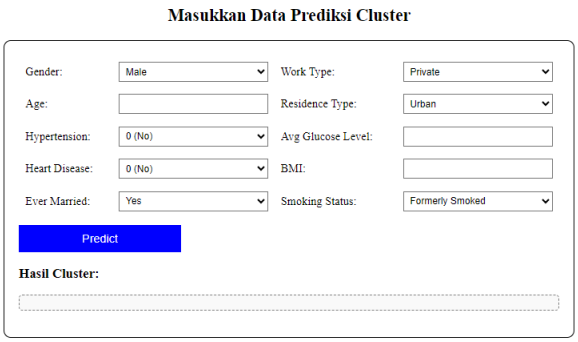
ISSN: 2598-8565 (media cetak)
ISSN: 2620-4339 (media online)

Sementara itu, Cluster 3 dan 6 menunjukkan kebanyakan kasus hipertensi tanpa penyakit jantung, yang membuat mereka dikategorikan sebagai kelompok dengan risiko sedang hingga tinggi. Cluster 5 dan 7, di sisi lain, menunjukkan individu usia lanjut dengan kombinasi hipertensi, penyakit jantung, dan kadar glukosa tinggi, dengan Cluster 7 menunjukkan tingkat risiko tertinggi. Selain itu, ada kelompok -1 atau suara yang menunjukkan kasus ekstrim. Misalnya, orang-orang dengan BMI yang sangat tinggi dan hipertensi serta penyakit jantung termasuk dalam kategori risiko tinggi.

Deployment

Model DBSCAN yang terbukti efektif dalam mengenali pola digunakan dalam aplikasi web berbasis *Flask*. Aplikasi ini memungkinkan pengguna menginput data dan langsung memperoleh hasil prediksi *cluster*.

Tampilan antarmuka (Gambar 13) dibuat sederhana dan mudah digunakan. Saat tombol "Predict" ditekan, model menganalisis data dan menampilkan *cluster* yang sesuai.



Gambar 13. Web Prediksi *Cluster* Risiko Stroke

Aplikasi ini bersifat *prototipe (proof of concept)* dan dapat digunakan untuk membantu peneliti dalam mengidentifikasi kelompok dengan karakteristik serupa, sehingga mendukung analisis data dan pengambilan keputusan secara eksperimental, bukan merupakan sistem siap pakai untuk implementasi operasional.

KESIMPULAN

Penelitian ini menggunakan algoritma DBSCAN pada dataset prediksi stroke publik untuk menilai kemungkinan pengelompokan individu berdasarkan faktor kesehatan. Hasil eksperimen menunjukkan bahwa parameter terbaik berada pada ε = 2,5 dan min_samples = 3, dengan skor *Silhouette* sebesar 0,2158. Skor ini menandakan bahwa pemisahan klaster masih kurang optimal sehingga struktur pengelompokan belum sepenuhnya representatif.

Walaupun demikian, DBSCAN terbukti mampu mendeteksi keberadaan *outlier* yang merepresentasikan individu dengan kondisi klinis tidak lazim. Hal ini meningkatkan kemungkinan bahwa algoritma berbasis kepadatan dapat digunakan sebagai langkah awal dalam mengeksplorasi data risiko stroke. Namun, kualitas pemisahan klaster yang buruk menunjukkan bahwa penyempurnaan lebih lanjut diperlukan agar hasil analisis benar-benar bermanfaat dalam konteks klinis dan dapat membantu pengambilan keputusan medis.

Kualitas pemisahan klaster masih dapat ditingkatkan, seperti yang ditunjukkan oleh nilai *Silhouette Score* yang relatif rendah (0,2158). Agar hasil klasterisasi lebih representatif secara klinis, penelitian selanjutnya disarankan untuk melakukan optimasi parameter ε dan min_samples secara sistematis dengan menggunakan pencarian grid atau pengoptimalan hyperparameter otomatis. Selain itu, disarankan untuk menerapkan seleksi fitur berbasis korelasi atau informasi saling berbagi.

Secara keseluruhan, penelitian ini menunjukkan DBSCAN sebagai metode eksploratif untuk mengelompokkan faktor klinis dan demografis yang berpotensi menyebabkan stroke. Penelitian ini memberikan perspektif baru tentang analisis data medis, khususnya dalam menemukan kelompok risiko tanpa bergantung pada label diagnosis. Hasil klasterisasi masih perlu dioptimalkan, tetapi penelitian ini memberikan landasan awal untuk membangun sistem yang mendukung keputusan medis berbasis data untuk mencegah stroke.

SARAN

1. Untuk meningkatkan kualitas hasil klasterisasi, disarankan menggunakan dataset yang lebih besar, lebih seimbang, atau lebih spesifik pada faktor kesehatan tertentu.
2. Perlu dilakukan perbandingan dengan algoritma klasterisasi lain, seperti *K-Means*, *Agglomerative Clustering*, atau *Gaussian Mixture Model* (GMM), guna menemukan metode paling sesuai dalam mengklasifikasikan risiko stroke.
3. Pola distribusi data dapat diperjelas dan interpretasi hasil analisis dapat diperkuat dengan menggunakan teknik reduksi dimensi alternatif seperti t-SNE atau UMAP.

DAFTAR PUSTAKA

Ding, L., Mane, R., Wu, Z., Jiang, Y., Meng, X., Jing, J., Ou, W., Wang, X., Liu, Y., Lin, J., Zhao, X., Li, H., Wang, Y., & Li, Z. (2022). Data-driven clustering approach to identify novel phenotypes using multiple biomarkers in acute ischaemic stroke: A retrospective, multicentre cohort study. *EClinicalMedicine*.

Hermanto, T. I. S. M. A. (2020). Analisis Data Sebaran Penyakit Menggunakan Algoritma Density Based Spatial Clustering Of Application WIith Noise. *Jurnal Sains Komputer Dan Teknologi Informasi*, *3*, 104–110. https://doi.org/https://doi.org/10.33084/jsakti.v3i1.1775

Hu, Y., Yan, H., Liu, M., Gao, J., Xie, L., Zhang, C., Wei, L., Ding, Y., & Jiang, H. (2024). Detecting cardiovascular diseases using unsupervised machine learning clustering based on electronic medical records. *BMC Medical Research Methodology*, *24*(1). https://doi.org/10.1186/s12874-024-02422-z

Islam, Y., Chowdhury, Md. J. U., & Das, S. C. (2025). *Advancing Tabular Stroke Modelling Through a Novel Hybrid Architecture and Feature-Selection Synergy*. http://arxiv.org/abs/2505.15844

Kim, J. T., Kim, N. R., Choi, S. H., Oh, S., Park, M. S., Lee, S. H., Kim, B. C., Choi, J., & Kim, M. S. (2022). Neural network-based clustering model of ischemic stroke patients with a maximally distinct distribution of 1-year vascular outcomes. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-13636-w

Ma, B., Yang, C., Li, A., Chi, Y., & Chen, L. (2023). A Faster DBSCAN Algorithm Based on Self-Adaptive Determination of Parameters. *Procedia Computer Science*, *221*, 113–120. https://doi.org/10.1016/j.procs.2023.07.017

Ovyawan Herlistiono, I., & Violina, S. (2024). Model Prediksi Risiko Stroke Menggunakan Machine Learning Stroke Risk Prediction Model Using Machine Learning. *Journal of Information Technology and Computer Science (INTECOMS)*, *7*(4).

Puspitasari, D. A., Cahyana, Y., Arum, S., & Lestari, P. (2023). Penerapan Algoritma Density Based Spastial Clustering Algorithm With Noise Untuk Pengelompokkan Penyakit Pasien. *Scientific Student Journal for Information, Technology and Science*, *IV*(1).

Santoso, H. (2019). Case Base Reasoning Untuk Mendiagnosis Penyakit Hipertensi Menggunakan Metode Indexing Density Based Spatial Clustering Application With Noise (DBSCAN). *ETHOS (Jurnal Penelitian Dan Pengabdian)*, *7*(1), 88–100. https://doi.org/10.29313/ethos.v7i1.4206

Sari, W. J., Melyani, N. A., Arrazak, F., Anahar, M. A. Bin, Addini, E., Al-Sawaff, Z. H., & Manickam, S. (2024). Performance Comparison of Random Forest, Support Vector Machine and Neural Network in Health Classification of Stroke Patients. *Public Research Journal of Engineering, Data Technology and Computer Science*, *2*(1), 34–43. https://doi.org/10.57152/predatecs.v2i1.1119

Sihite, E. K., Rangkuti, Y. M., & Karo-Karo, I. (2024). Pembangunan Webgis Untuk Penderita Gizi Buruk Di Kota Medan Berdasarkan Hasil Clustering Algoritma DBSCAN. *SAINTIKOM (Jurnal Sains Manajemen Informatika Dan Komputer)*, *23*, 77–86.

Venketasubramanian, N., Yudiarto, F. L. ;, & Tugasworo, D. (2022). Stroke Burden and Stroke Services in Indonesia. *Cerebrovascular Diseases Extra*, *12*(1), 53–57. https://doi.org/10.1159/000524161

Wijaya, R., Saeed, F., Samimi, P., Albarrak, A. M., & Qasem, S. N. (2024). An Ensemble Machine Learning and Data Mining Approach to Enhance Stroke Prediction. *Bioengineering*, *11*(7). https://doi.org/10.3390/bioengineering11070672