PENARAPAN ALGORITMA K-MEANS CLUSTERING DALAM MENGANALISIS RESIKO PENYAKIT STROKE

Bani Nurhakim, Intan septiani, Khaerul Anam, Denni Pratama
Teknik Informatika, STMIK IKMI CIREBON
Jl. Perjuangan No.10, B, Karyamulya, Kec.Kesambi, Kota.Cirebon, Jawa Barat 45135
baninurhakim@gmail.com

ABSTRAK

Di kawasan Asia Tenggara, stroke menempati peringkat ketiga sebagai penyebab kecacatan dan menjadi penyakit dengan risiko kematian tertinggi kedua. penyakit Stroke terjadi ketika pembuluh darah mengalir ke otak terhambat atau terjadi pecah, mengakibatkan beberapa sel jaringan otak tidak dapat menerima oksigen yang dibutuhkan dari aliran darah. Permasalahan mencakup mengidentifikasi faktor risiko penyakit stroke dengan akurat sehigga menjadi suatu tantangan dan Menentukan jumlah cluster (K) yang paling sesuai untuk mencerminkan kelompok resiko penyakit stroke dengan optimal. Dengan menerapkan pendekatan data mining K-means clustering, pasien dapat dikelompokkan ke dalam kategori resiko yang beragam. Beberapa kelompok pasien menunjukkan dominasi faktor risiko tertentu, sehingga memberikan pemahaman lebih rinci mengenai variasi resiko stroke dalam populasi. Temuan dari analisis K-means clustering telah diperkuat melalui validasi menggunakan data klinis tambahan, menegaskan kehandalan dan relevansi hasil. Proses validasi melibatkan uji model pada dataset yang berbeda untuk memastikan keberlakuan temuan secara umum. Penelitian ini melibatkan langkah-langkah teliti dalam penentuan jumlah cluster (K) untuk menjamin hasil clustering yang optimal. Pendekatan ini membantu mengurangi risiko overfitting atau underfitting, sehingga meningkatkan akurasi analisis. Visualisasi hasil clustering memberikan gambaran yang jelas tentang sebaran faktor risiko di antara kelompok pasien, dengan grafik dan diagram yang memudahkan interpretasi serta komunikasi temuan kepada pihak-pihak yang berkepentingan.

Kata kunci : *algoritma k-means, penyakit stroke, analisis*

1. PENDAHULUAN

Masalah penyakit stroke menjadi sorotan utama dalam ranah kesehatan global, mengingat dampak seriusnya terhadap kesehatan dan tingkat kematian yang signifikan di berbagai negara. Dalam menghadapi tantangan ini, deteksi dini dan pemahaman mendalam mengenai faktor-faktor yang mempengaruhi risiko stroke menjadi krusial.

Pentingnya teknologi dan metode analisis data mendorong penelitian untuk menjelajahi cara-cara baru dalam memahami risiko penyakit stroke. Tujuan penelitian ini adalah untuk menggali potensi penerapan algoritma K-means clustering dalam menganalisis risiko penyakit stroke. Dengan memanfaatkan teknik clustering, penelitian ini diharapkan dapat mengungkap pola-pola yang mungkin sulit terdeteksi secara manual, serta membantu mengelompokkan individu ke dalam kategori risiko yang berbeda.

Pendekatan ini diharapkan memberikan kontribusi signifikan dalam pengembangan metode analisis risiko penyakit stroke yang lebih efisien dan akurat. Dengan memahami faktor-faktor yang berkontribusi pada risiko stroke melalui analisis data menggunakan algoritma K-means clustering, diharapkan upaya pencegahan dan intervensi dapat dilakukan dengan lebih tepat sasaran, sehingga dapat mengurangi dampak buruk penyakit ini pada tingkat individu maupun populasi secara keseluruhan.[1].

2. TINJAUAN PUSTAKA

2.1. Analisis

Analisis merupakan proses terstruktur yang melibatkan pemeriksaan, pemecahan, atau penyelidikan terhadap informasi, situasi, atau materi tertentu dengan tujuan memahami atau mengevaluasi komponen-komponennya. Tujuan dari analisis adalah memperoleh pemahaman yang lebih mendalam, mengidentifikasi pola atau hubungan, serta mengorganisir informasi menjadi elemen-elemen yang lebih kecil sehingga dapat diartikan atau digunakan dalam pengambilan keputusan. Analisis mencakup kegiatan seperti pengumpulan, pengolahan, dan interpretasi data untuk mencapai pemahaman yang lebih holistic.

2.2. *Algoritma K-means*

Metode Algoritma K-Means adalah teknik analisis data yang mengelompokkan sekumpulan data ke dalam kelompok atau cluster. Tujuan utamanya adalah untuk membagi data menjadi beberapa kelompok. menunjukkan kesamaan atau keterkaitan tertentu.

Sebuah algoritma yang mengatasi kelemahan dari algoritma K-Means, di maa kelemahan K-Means terletak pada kendala dalam perhitungan yang relatif lambat dan memerlukan pengguna untuk menentukan nilai kluster K, adalah algoritma pembelajaran ini atau memprediksi nilai variabel target. Sebaliknya, algoritma pengelompokan mencoba membagi seluruh kumpulan data menjadi subkelompok Dengan

memanfaatkan algoritma ini, data yang sudah dikumpulkan dapat dikelompokkan menjadi beberapa cluster. Implementasi proses Clustering K-Means dilakukan menggunakan perangkat RapidMiner.[3]

2.3. Penyakit Stroke

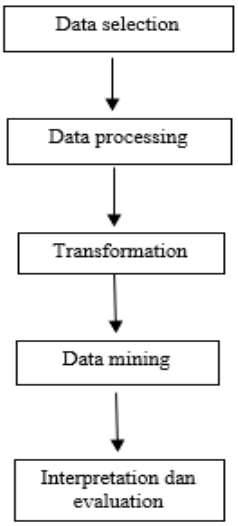
Penyakit Stroke merupakan suatu kesehatan yang terjadi ketika aliran darah menuju otak terganggu atau terhenti, mengakibatkan kerusakan pada sel-sel otak karena kekurangan suplai oksigen dan nutrisi. Gangguan ini dapat disebabkan oleh penyumbatan pembuluh darah otak oleh bekuan darah (*infark serebral*) atau perdarahan dari pecahnya pembuluh darah (perdarahan serebral). Terdapat dua jenis utama stroke Yang pertama adalah Stroke iskemik terjadi ketika penyumbatan di pembuluh darah memotong aliran darah ke otak. Penyumbatan dapat berupa bekuan darah yang terbentuk di dalam pembuluh darah otak atau berasal dari bagian tubuh lain (emboli). Yang kedua adalah *Stroke Hemoragik* yaitu Terjadi saat pembuluh darah pecah, mengakibatkan darah tumpah ke dalam jaringan otak. Ini mungkin disebabkan oleh tekanan darah tinggi, pecahnya aneurisma (pelebaran tidak normal pembuluh darah), atau kondisi medis lainnya.

2.4. Clustering

Clustering merupakan suatu proses di mana objek-objek yang mirip dikelompokkan bersama dalam subset yang berbeda. Lebih spesifiknya, data set dipartisi ke dalam subset sehingga setiap subset memiliki makna yang signifikan. Sebuah klaster terbentuk dari kumpulan objek yang serupa satu sama lain, membedakannya dari objek-objek yang terdapat dalam klaster lainnya [4]. Dalam konteks clustering, objek-objek yang termasuk dalam satu kelompok seharusnya memiliki kemiripan yang tinggi satu sama lain, sedangkan perbedaan antara objek-objek di kelompok yang berbeda seharusnya signifikan. Teknik clustering dalam data mining adalah proses pengelompokan atau klasifikasi objek berdasarkan informasi yang diperoleh dari data. Tujuan utamanya adalah untuk Memahami hubungan antar objek sesuai prinsip serta meminimalkan antar cluster satu dan cluster lainnya .[5]

3. METODE PENELITIAN

Diagram di bawah ini menunjukkan proses tahap penelitian menggunakan KDD (Knowledge Discovery Database). Proses KDD adalah pendekatan yang menggunakan teknik penambangan data untuk mengekstrak pengetahuan yang berpotensi relevan sesuai dengan batasan dan batasan ukuran tertentu.



Gambar 1. Alur Penelitian

Pada tabel dibawah ini merupakan pembahasan dari masing-masing proses dari tahapan KDD (*knowledge discovey data base*), berikut penjelasannya seperti dibawah ini:

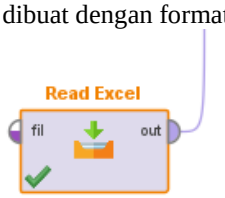
Tabel 1. Tahapan KDD (knowledge discovery data base)

Tahapan	Aktivitas	Deskripsi Aktivitas
Data selection	Mengumpulkan data medis atau Lembaga Kesehatan yang mencakup parameter- parameter yang berkaitan dengan resiko penyakit stroke	Pada tahapan ini Pengumpulkan data dari berbagai sumber yang relevan seperti data medis, Riwayat pasien, faktor gaya hidup, dan faktor resiko stroke lainnya
Data processing	Melakukan pembersihan data dan pemprosesan data	Pada fase ini, penulis melakukan pembersihan data, mengidentifikasi outlier dalam data set untuk mempertimbangkan apakah oulier tersebut harus dihapus atau dilakukan trnasformasi nilai.
transformation	Melakukan transformasi logaritmik pada variabel tertentu yang memiliki distribusi yang tidak normal	pada tahapan ini penulis **melakukan transformasi data** untuk distribusi variabel tertentu yang tidak normal maka akan dilakukan log transform pada variabel,
data mining	Menentukan jumlah optimal cluster	Pada tahapan ini mengimplementasikan algoritma k-means untuk **mengelompokkan** data pasien resiko penyakit stroke menjadi clustercluster yang sesuai, dan menentukan jumlah optimal cluster.
interpretation / evaluation	Mengidentifikasi ciri-ciri khusus dari setiap kelompok dan memberikan karakteristik prediksi resiko penyakit stroke	Pada tahapan ini **mengidentifikasi** ciri ciri khusus dari setiap kelompok dan memberikan label yang mencerminkan karakteristik prediksi resiko penyakit stroke melibatkan analisis pola unik pada setiap cluster hasil clustering.

4. HASIL DAN PEMBAHASAN

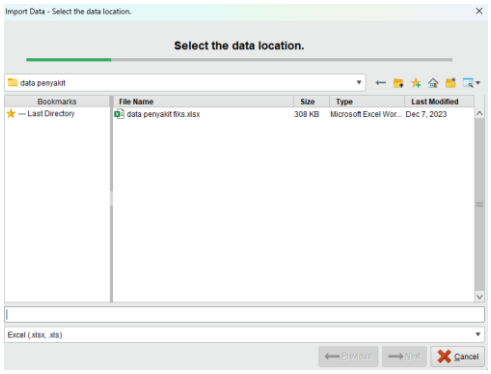
4.1. Data Selection

Tahap awal yang perlu dilakukan dalam metode KDD (Knowledge Discovery in Databases) adalah memilih data [6]. Pada langkah awal, pencarian dilakukan untuk menemukan file lokasi yang sebelumnya telah dibuat dengan format xlsx.



Gambar 2. File excel

Setelah memasukan operator read excel, Langkah berikutnya ialah mengimport data akan diolah, berikut gambar dibawah ini adalah tampilan import file data set excel dengan nama data penyakit fiks.xlsx yang terdapat pada folder data penyakit.



Gambar 3. Tampilan Data Yang Akan Diimport

Setelah semua data selesai diimport, langkah selanjutnya adalah memasukkan data tersebut ke dalam halaman kerja rapid miner, dengan cara klik dan tahan pada data, lalu menarik data tersebut ke dalam halaman kerja RapidMiner.

4.2. Data processing

Pre-processing merupakan langkah yang bertujuan untuk membersihkan data dari segala gangguan, duplikasi, ketidaksesuaian, dan kesalahan tipografi selama proses Data Mining. Pada tahap ini, kegiatan mencakup pemeriksaan nilai yang hilang, [7]

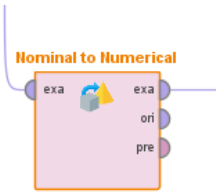
Gambar 4. Hasil Processing Data

Pada tahap ini, tujuannya melakukan pembersihan atau mengatasi nilai yang hilang agar data yang digunakan tidak mengalami kesalahan atau masalah saat diproses. Gambar di bawah ini menunjukkan visualisasi sebelum dilakukannya pembersihan data atau sebelum penanganan nilai yang hilang. Sebagai contoh.

Dari hasil data setelah diprocessing yang terdapat pada gambar diatas dapat disimpulkan bahwa tidak terdapat missing values, atau tidak terdapat data yang kosong dan bisa melanjutkan untuk tahapan selanjutnya.

4.3. Transformation

Transformasi data digunakan untuk mengubah bentuk data agar sesuai dengan kebutuhan selama proses penambangan data [8] Pada tahapan ini penulis menggunakan tahapan transformation bertujuan untuk mengubah attribute non numerik menjadi attribute numerik, Langkah selanjutnya menambahkan operator "nominal to numerical" dengan mencarinya pada kolom search for operator setelah mencari pada kolom pencarian, dengan mengklik dan Tarik ke lembar kerja rapid miner.

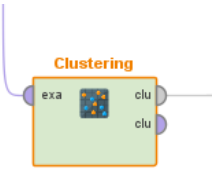


Gambar 5. Tampilan Nominal To Numerical

Pada gambar diatas ini berfungsi mengubah data nominal ke data numerical pada penelitian ini data yang digunakan oleh penulis ialah data nominal dalam clusterisasi data yang harus digunakan ialah harus data numerical oleh karena itu penulis menggunakan operator nominal to numerical untuk mengubah dari data nomial ke data numerical.

4.4. Data Mining

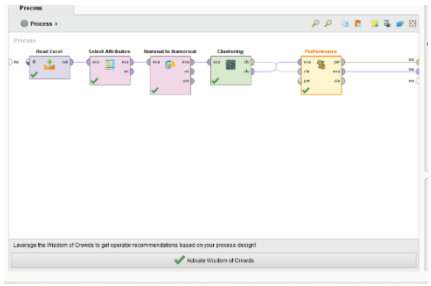
Data mining dapat digunakan untuk melakukan tugas-tugas seperti klasifikasi, klasterisasi, dan penemuan pola asosiasi, sehingga memungkinkan untuk melakukan prediksi atau peramalan [9]. Pada tahapan Penentuan jumlah kluster adalah tahapan penting yang perlu dijalankan. Hasil dari pengelompokan tersebut akan menjadi dasar untuk melaksanakan suatu tindakan atau kegiatan tertentu. prediksi resiko yang terkena penyakit stroke, pada berikut ini.



Gambar 6. Operator K-Means

Clustering merupakan operator Cluster K-Means, operator ini adalah operator utama dari pemodelan, pada proses ini agar dapat mengetahui hasil dari pengklasteran tersebut.

Pada tahap menghubungkan semua operator setelah semua terhubung lalu Klik tombol Jalankan, setelah klik tombol Jalankan Hasil akurasi data ditampilkan secara otomatis.



Gambar 7. Hasil Data dan operator Yang sudah Terhubung

Pada gambar diatas ialah menghubungkan semua operator bertujuan untuk mengetahui output kearah result ini bertujuan untuk melihat semua hasil data akurasi dengan menggunakan model algoritma K-means clustering yang sudah diproses.

4.5. Interpretation/ evaluation

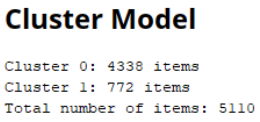
Pada fase cluster model menampilkan nilai titik tengah dari setiap cluster dalam fase ini. Ditunjukkan di bawah ini adalah hasil pengelompokan menggunakan algoritma K-Means.

Tabel 2. hasil percobaan cluster

Cluster	DBI
K2	**0.082**
K3	0.128
K4	0.142
K5	0.157
K6	0.173
K7	0.182
K8	0.177
K9	0.170
K10	0.162
K11	0.163
K12	0.166
K13	0.170
K14	0.168
K15	0.168
K16	0.162
K17	0.167
K18	0.165
K19	0.162
K20	0.168

Dapat disimpulkan dari hasil percobaan diatas dari percobaan *k 2 – k 20.* Hasil *k* yang baik adalah hasil *k* yang mendekati angka *0* / nilai cluster yang paling kecil. Dengan melihat semua hasil *Davies bouldin* dapat dilihat dimana cluster dengan nilai yang baik yang mendekati angka 0. *K* yang mendekati angka 0/ dengan nilai cluster yang paling terendah. Dari hasil percobaan dari *k2 – k20* adalah *K2= 0.082*, dimana *davies bouldin* apabila mendekati 0 nilai berarti hasil algoritma nya semakin baik, maka disimpulkan jumlah cluster terbaik disini ada 2 cluster yang terbaik yang mendekati 0 / dengan nilai cluster yang paling terendah.

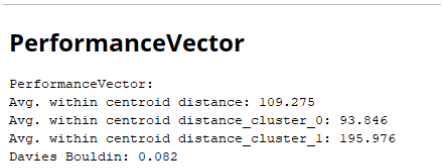
Pada tahap cluster model pada tahapan ini menampilkan nilai titik pusat pada setiap cluster. Nilai tersebut akan menjadi acuan perhitungan pada setiap data set dengan mengukur nilai pada masing-masing titik pusat.



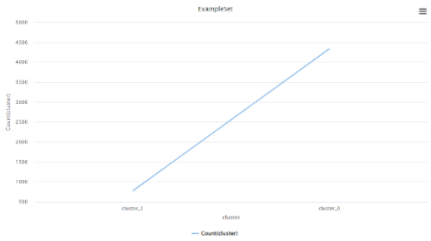
Gambar 8. Tampilan Cluster Model

Pada gambar diatas merupakan hasil klasterisasi dengan algoritma K-Means yang berarti bahwa resiko penyakit stroke dilihat dari penyebabnya tersebut akan dikelompokkan menjadi 2 klaster yaitu cluster 0, cluster 1 untuk cluster 0 berjumlah 4338 items sedangkan cluster 1 ada 772 item.

Dibawah ini merupakan hasil performance vector dari hasil percobaan untuk mencari nilai cluster terbaik.



Gambar 9. Hasil Performance Vector



Gambar 9. Grafik hasil pengelompokkan cluster

Dihasilkan nilai davies bouldin nya adalah 0.082. dengan hasil avg within centroiddistance 109.275. Tahap pengujian ini membantu menentukan jumlah cluster yang optimal melalui hasil komputasi dan pemodelan cluster. Evaluasi dilakukan dengan menggunakan metode Davies-Bouldin Index (DBI), yang merupakan salah satu cara untuk menilai kualitas cluster [10] Pada gambar dibawah ini merupakan hasil

plot cluster dalam pengelompokkan resiko penyakit stroke.

Pada hasil pengelompokkan diatas cluster yang paling tinggi berada pada cluster 0 sedangkan cluster terendah berada pada cluster 1, dimana jarak nilai cluster 0 dihasilkan 4338 sedangkan hasil cluster 1 dihasilkan 772.

5. KESIMPULAN DAN SARAN

Dari permasalahan penelitian ini dapat disimpulkan bahwa pelaksanaan analisis dengan menggunakan algoritma K-means clustering dapat digunakan untuk mengidentifikasi pola risiko penyebab stroke. Hasil analisis menggunakan algoritma K-means clustering dibuat dua kelompok berupa Cluster 0 dengan hasil 4338 elemen dan Cluster 1 dengan hasil 772 elemen. Hasil pengujian menggunakan aplikasi Rapid Miner menghasilkan nilai K optimal 2 dengan nilai DBI (Davies Bouldin Index) dengan hasil cluster 0.082.

Saran untuk penelitian di masa depan adalah untuk mengatasi kekurangan penelitian ini dengan memperluas jumlah data yang digunakan untuk mendapatkan hasil yang lebih akurat. Selain itu, penelitian di masa depan dapat mempertimbangkan untuk menggunakan algoritma penambangan data tambahan atau bahkan perbandingan yang berbeda dari kedua algoritma. Hal ini memungkinkan hasil pengelompokan data menjadi lebih beragam dan detail

DAFTAR PUSTAKA

[1] Y. Tedju, "Pengelompokan Penyakit Hipertensi Di Kota Kupang Menggunakan Metode K-Means," *HOAQ (High Educ. Organ. Arch. Qual. J. Teknol. Inf.*, vol. 13, no. 2, pp. 98–108, 2022, doi: 10.52972/hoaq.vol13no2.p98-108.

[2] F. Rahmadayanti, I. Anggraini, and T. Susanti, "Pengklasterisasian Data Penyakit Hipertensi 94–902, 2023, doi: 10.47065/josh.v4i3.3363.

dengan Menggunakan Metode K-Means," *J. Inf. Syst. Res.*, vol. 4, no. 2, pp. 737–741, 2023, doi: 10.47065/josh.v4i2.2905.

[3] F. S. Napitupulu, I. S. Damanik, I. S. Saragih, and A. Wanto, "Algoritma K-Means untuk Pengelompokkan Dokumen Akta Kelahiran pada Tiap Kecamatan di Kabupaten Simalungun," *Build. Informatics, Technol. Sci.*, vol. 2, no. 1, pp. 55–63, 2020, doi: 10.47065/bits.v2i1.323.

[4] M. R. Sulistio, N. Suarna, and O. Nurdiawan, "Analisa Penerapan Metode Clustering X-Means Dalam Pengelompokan Penjualan Barang," *J. Teknol. Ilmu Komput.*, vol. 1, no. 2, pp. 37–42, 2023, doi: 10.56854/jtik.v1i2.49.

[5] B. Kristanto, A. T. Zy, and M. Fatchan, "Analisis Penentuan Karyawan Tetap Dengan Algoritma K-Means Dan Davies Bouldin Index," *Bull. Inf. Technol.*, vol. 4, no. 1, pp. 112–120, 2023.

[6] R. K. Purnomo, H. A. Saufi, and R. Hs, "Experimental Student Experiences," *Exp. Student Exp.*, vol. 1, no. 1, pp. 1–7, 2023.

[7] M. Walid and F. Halimiyah, "Klasifikasi Kemandirian Siswa SMA/MA Double Track Menggunakan Metode Naive Bayes," *J. ICT Inf. Commun. Technol.*, vol. 22, pp. 190–197, 2022.

[8] O. Nurdiawan, A. Irma Purnamasari, and I. Ali, "Analisa Penjualan Mobil Dengan Menggunakan Algoritma K-Means Di PT. Mulya Putra Kencana," *J. Data Sci. dan Inform.*, vol. 1, no. 2, pp. 32–35, 2021.

[9] O. N. Washilaturrizqi, "Implementasi Algoritma C4 . 5 Untuk Menentukan Penerima," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, pp. 373–377, 2023.

[10] A. Maulana, A. Nazir, R. M. Candra, S. Sanjaya, and F. Syafria, "Clustering Vaksinasi Penyakit Mulut dan Kuku Menggunakan Algoritma K-Means," *J. Inf. Syst. Res.*, vol. 4, no. 3, pp. 8