ANALISIS PENYAKIT JANTUNG DENGAN PCA UNTUK IDENTIFIKASI KESEHATAN KARDIOVASKULAR

Sebastian Wibowo [1] Dillon Majesson [2] Junardi Chailesia [3]

[1) 2) 3)] Teknik Informatika, FTI, Universitas Tarumanagara
Jl. Letjen S Parman No. 1, Jakarta 11440 Indonesia
email : sebastian.535230187@stu.untar.ac.id [1] dillon.535230209@stu.untar.ac.id [2]
junardi.535230225@stu.untar.ac.id [3]

ABSTRAK

Tujuan dari penelitian ini membahas analisis data klinis untuk mengidentifikasi risiko penyakit jantung menggunakan metode unsupervised learning. Metode yang digunakan adalah Principal Component Analysis (PCA) untuk reduksi dimensi data numerik, dan K-Medoids Clustering untuk mengelompokkan pasien berdasarkan kemiripan atribut klinis. Dataset heart.csv yang digunakan mencakup variabel numerik dan kategorikal, seperti tekanan darah, kolesterol, jenis kelamin, dan hasil EKG. Outlier dideteksi menggunakan Mahalanobis Distance, kemudian dilakukan clustering pada variabel kontinu dengan PCA dan jarak Euclidean, serta pada variabel diskrit dengan jarak Hamming dan visualisasi t-SNE. Hasil clustering menunjukkan adanya kelompok pasien dengan karakteristik klinis serupa yang mencerminkan tingkat risiko penyakit jantung yang berbeda. Kombinasi PCA dan K-Medoids terbukti efektif untuk mengungkap pola risiko secara visual dan sistematis, sehingga berpotensi mendukung pengambilan keputusan klinis

Key words

Penyakit Jantung, PCA, K-Medoids, t-SNE, Data Mining

1. Pendahuluan

Penyakit jantung merupakan salah satu masalah kesehatan yang paling mematikan dan kompleks di dunia. Berdasarkan laporan dari World Health Organization (WHO), penyakit kardiovaskular menyebabkan sekitar 17,9 juta kematian setiap tahun, yang mencakup 31% dari seluruh penyebab kematian global [1]. Penyakit jantung koroner menjadi salah satu jenis penyakit kardiovaskular yang paling sering terjadi dan memiliki risiko tinggi di Indonesia, dengan angka prevalensi yang terus meningkat secara signifikan dalam beberapa tahun terakhir [2]. Tingginya angka kematian ini menekankan pentingnya deteksi dini dan pencegahan penyakit jantung melalui pendekatan berbasis data dan teknologi.

Dalam praktik medis, diagnosis penyakit jantung biasanya bergantung pada evaluasi data klinis seperti usia, jenis kelamin, tekanan darah, kadar kolesterol, detak jantung, dan hasil tes elektrokardiogram (electrocardiogram atau ECG). Namun, banyaknya atribut dan kompleksitas hubungan antar variabel membuat proses diagnosis menjadi sulit dan rawan kesalahan jika dilakukan secara manual. Oleh karena itu, pendekatan berbasis data mining dan machine learning semakin populer untuk mengekstraksi pola dan informasi penting dari data medis yang besar dan kompleks [3].

Pendekatan supervised learning seperti K-Nearest Neighbors (K-NN), Decision Tree, dan Logistic Regression telah banyak digunakan untuk memprediksi risiko penyakit jantung berdasarkan data yang telah dilabeli [4]. Namun, pendekatan ini memiliki keterbatasan karena memerlukan data berlabel, yang tidak selalu tersedia atau akurat. Sebaliknya, unsupervised learning tidak memerlukan data berlabel dan berguna dalam eksplorasi data untuk menemukan pola tersembunyi. Dua metode utama dalam unsupervised learning yang relevan dalam konteks ini adalah Principal Component Analysis (PCA) dan K-Medoid Clustering.

Principal Component Analysis adalah metode reduksi dimensi yang digunakan untuk menyederhanakan data multivariat dengan memproyeksikannya ke dimensi yang lebih rendah tanpa kehilangan informasi penting [5]. PCA dapat membantu mengurangi kompleksitas data dengan menghilangkan redundansi dan noise, sehingga membuat proses analisis menjadi lebih efisien dan akurat [6]. Setelah data direduksi dengan PCA, metode K-Medoid Clustering dapat digunakan untuk mengelompokkan pasien ke dalam beberapa grup berdasarkan kemiripan atribut klinis mereka. Dengan metode ini, kita dapat mengidentifikasi kelompok pasien dengan tingkat risiko kardiovaskular yang serupa [7].

Penelitian ini bertujuan untuk mengidentifikasi pola risiko penyakit jantung dengan menerapkan metode PCA untuk reduksi dimensi dan K-Medoid untuk pengelompokan. Dengan pendekatan ini, diharapkan dapat diperoleh wawasan mengenai pengelompokan pasien berdasarkan kondisi klinis mereka, yang nantinya dapat menjadi dasar dalam merancang strategi pencegahan dan intervensi medis yang lebih tepat sasaran.

2. Landasan Teori

2.1. Principal Component Analysis

Principal Component Analysis (PCA) adalah teknik statistik yang digunakan untuk mengurangi jumlah variabel dalam suatu dataset besar dengan mempertahankan sebanyak mungkin informasi yang terkandung di dalamnya [8]. PCA bekerja dengan mengubah data asli menjadi himpunan baru dari variabel yang tidak saling berkorelasi, yang disebut dengan principal components. Setiap principal component adalah kombinasi linier dari variabel asli dan diurutkan berdasarkan jumlah varians data yang dijelaskan olehnya.

Tujuan utama dari PCA adalah untuk menyederhanakan struktur data tanpa kehilangan informasi yang signifikan [9]. Dalam konteks penyakit jantung, banyaknya atribut klinis yang saling berkorelasi dapat menyulitkan analisis. Dengan menggunakan PCA, atribut-atribut tersebut dapat direduksi menjadi beberapa komponen utama yang tetap merepresentasikan data asli dengan baik. Komponen utama ini kemudian dapat digunakan sebagai input untuk metode analisis lanjutan seperti clustering.

Langkah-langkah utama dalam penerapan PCA antara lain:

1. Menstandardisasi data agar setiap atribut memiliki skala yang sama.
2. Menghitung matriks kovarians dari data.
3. Menghitung eigenvalue dan eigenvector dari matriks kovarians.
4. Memilih sejumlah komponen utama berdasarkan besar eigenvalue.
5. Mengubah data asli menjadi ruang baru menggunakan komponen utama tersebut.

Dengan menggunakan PCA, analisis dapat dilakukan lebih cepat dan efisien karena pengurangan dimensi juga mengurangi kompleksitas komputasi serta menghilangkan noise yang tidak relevan [10].

2.2. Dataset heart.csv

Dataset yang digunakan dalam penelitian ini adalah Heart Disease Dataset yang tersedia di platform Kaggle. Dataset ini terdiri dari 1025 data pasien dan memiliki 14 atribut klinis yang berkaitan dengan kesehatan jantung.

Beberapa atribut utama dalam dataset ini antara lain:

1. Age: usia pasien
2. Sex: jenis kelamin pasien
3. ChestPainType: tipe nyeri dada
4. RestingBP: tekanan darah saat istirahat
5. Cholesterol: kadar kolesterol dalam darah
6. FastingBS: gula darah puasa
7. RestingECG: hasil elektrokardiogram saat istirahat
8. MaxHR: detak jantung maksimum
9. ExerciseAngina: indikasi angina akibat latihan fisik
10. Oldpeak: depresi ST yang disebabkan oleh olahraga
11. ST_Slope: kemiringan segmen ST
12. HeartDisease: status diagnosis penyakit jantung (label, tetapi tidak digunakan karena pendekatan unsupervised.

Dataset ini dipilih karena memiliki atribut yang cukup lengkap dan telah digunakan secara luas dalam berbagai penelitian terkait prediksi dan analisis penyakit jantung. Sebelum diterapkan ke dalam algoritma PCA dan K-Medoid, data terlebih dahulu dilakukan pra-pemrosesan seperti penghapusan nilai kosong, konversi variabel kategori ke numerik, dan normalisasi.

2.3. Mahalanobis Distance

Mahalanobis Distance adalah suatu ukuran jarak multivariat yang mengukur seberapa jauh suatu titik dari distribusi sekelompok data, dengan mempertimbangkan korelasi antar variabel dalam dataset [11]. Berbeda dengan jarak Euclidean biasa, Mahalanobis Distance mempertimbangkan struktur kovarians data, sehingga lebih cocok digunakan pada data yang memiliki korelasi antar variabel atau skala yang berbeda-beda.

Secara matematis, Mahalanobis Distance D_M antara suatu vektor observasi x dan mean dari distribusi μ dengan matriks kovarians Σ dapat dihitung dengan rumus:

$$D_M = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)} \quad (1)$$

Dalam konteks data mining dan analisis data medis, Mahalanobis Distance sering digunakan untuk deteksi outlier, yaitu observasi yang berada jauh dari pusat distribusi data [12]. Misalnya, dalam studi ini, Mahalanobis Distance digunakan untuk mengidentifikasi pasien dengan atribut klinis yang tidak lazim atau ekstrem, yang dapat merusak hasil pengelompokan jika tidak dihilangkan terlebih dahulu.

Metode ini unggul karena memperhitungkan varian dan kovarians antar atribut, sehingga mampu mendeteksi outlier bahkan jika data memiliki orientasi atau penyebaran yang kompleks.

2.4. K-Medoid Clustering

K-Medoid Clustering adalah salah satu algoritma unsupervised learning yang digunakan untuk membagi data menjadi beberapa kelompok atau klaster berdasarkan kemiripan (similarity) antar data [13].

Berbeda dengan algoritma K-Means yang menggunakan mean sebagai pusat klaster, K-Medoid menggunakan medoid, yaitu titik data yang paling representatif atau paling "sentral" dalam suatu klaster [14].

Proses kerja K-Medoid meliputi:

1. Memilih secara acak sejumlah k buah data sebagai medoid awal.
2. Mengelompokkan data lain berdasarkan jarak terdekat dengan medoid.
3. Menghitung total biaya (dalam bentuk total jarak) dari klaster yang terbentuk.
4. Menukar medoid dengan anggota klaster lain dan mengevaluasi apakah biaya menurun.
5. Mengulangi proses hingga tidak ada lagi penurunan biaya.

Dalam penelitian ini, algoritma K-Medoid diterapkan untuk dua jenis variabel:

1. Variabel kontinu (dengan jarak Euclidean).
2. Variabel diskrit (dengan jarak Hamming) untuk menggambarkan kesamaan kategori.

2.5. Euclidean Distance

Euclidean Distance adalah salah satu metrik jarak paling sederhana dan umum digunakan untuk mengukur jarak lurus antara dua titik dalam ruang Euclidean [15]. Diberikan dua titik data x=(x1,x2,...,xn) dan y=(y1,y2,...,yn), maka jarak Euclidean didefinisikan sebagai:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \qquad (2)$$

Dalam konteks clustering data kontinu, jarak Euclidean digunakan sebagai dasar pengukuran kemiripan antar observasi. Namun, penggunaan Euclidean Distance memerlukan data numerik dan skala yang seragam. Oleh karena itu, dilakukan standardisasi (z-score normalization) terhadap data agar semua atribut memiliki rata-rata nol dan deviasi standar satu [16] . Hal ini menghindari dominasi atribut tertentu dalam proses perhitungan jarak.

2.6. Hamming Distance

Hamming Distance adalah metrik jarak yang digunakan untuk membandingkan dua vektor diskrit berdimensi sama dan menghitung jumlah posisi berbeda antar keduanya. Metrik ini sangat sesuai untuk data kategorikal atau biner, di mana Euclidean Distance tidak lagi relevan. Jika dua vektor x dan y memiliki dimensi n, maka jarak Hamming didefinisikan sebagai:

$$d_H(x,y) = \sum_{i=1}^{n} \delta(x_i, \mathrm{y_i}) \qquad (3)$$

$$\delta(x_i, y_i) = \{0, jika\ x_i = y_i\} \qquad (4)$$

$$\delta(x_i, y_i) = \{1, jika\ x_i \neq y_i\} \qquad (5)$$

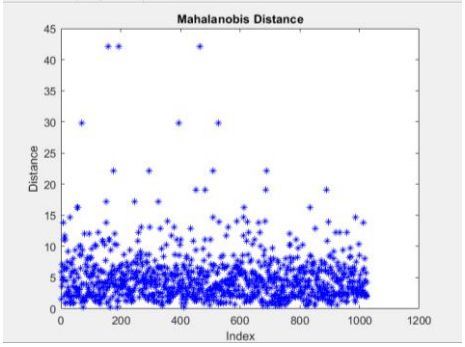
Dalam proyek ini, Hamming Distance digunakan untuk melakukan K-Medoid Clustering terhadap data diskrit seperti jenis kelamin, tipe nyeri dada, atau hasil tes EKG. Karena data tersebut tidak dapat diinterpretasikan secara numerik dalam konteks jarak biasa, Hamming Distance menjadi metrik yang tepat.

Visualisasi hasil clustering data diskrit kemudian dilakukan menggunakan t-SNE (t-distributed Stochastic Neighbor Embedding), yang mampu mereduksi dimensi data dan menampilkan pola clustering secara visual dalam 2D.

3. Hasil Percobaan

3.1. Deteksi Outlier dengan Mahalanobis Distance

Pada bagian plot ini memperlihatkan distribusi jarak Mahalanobis dari seluruh observasi dalam dataset berdasarkan atribut klinis kontinu seperti usia (age), tekanan darah (trestbps), kolesterol (chol), detak jantung maksimum (thalach), dan depresi segmen ST (oldpeak). Mahalanobis Distance mempertimbangkan kovariansi antar fitur sehingga mampu mengenali outlier multivariat, yaitu data yang secara statistik menyimpang dari distribusi umum pasien lainnya.



Gambar 1. Plot Mahalanobis Distance

Dari plot terlihat bahwa sebagian besar data memiliki nilai jarak di bawah 15, namun terdapat beberapa titik yang jauh di atas ambang batas, bahkan mendekati atau melebihi jarak 40. Titik-titik ini ditandai sebagai outlier dan dihapus untuk menjaga integritas proses clustering.

Total terdeteksi 42 outlier, yang kemudian dihapus dari dataset. Keberadaan outlier seperti ini penting untuk diidentifikasi, karena dapat menunjukkan pasien dengan kondisi klinis ekstrem atau pencatatan data yang tidak valid.

3.2. Evaluasi dan Visualisasi Clustering pada Variabel Kontinu

3.2.1 Evaluasi Silhouette Score (Variabel Kontinu)

Clustering dilakukan terhadap variabel kontinu menggunakan algoritma K-Medoids dengan jarak

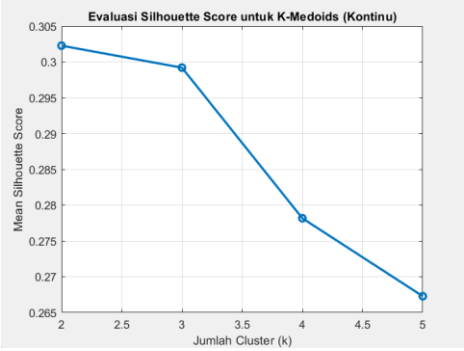
Euclidean. Untuk menilai kualitas hasil clustering, digunakan metrik Mean Silhouette Score, yaitu ukuran yang mengindikasikan seberapa dekat suatu titik dengan anggota klasternya dibandingkan dengan klaster lainnya.

Evaluasi dilakukan terhadap jumlah klaster k dari 2 hingga 5, yang ditunjukkan pada Tabel 1 berikut :

Jumlah Cluster (k)	Mean Silhouette Score
2	0.3023
3	0.2992
4	0.2782
5	0.2673

Tabel 1 Silhouette Score Variabel Kontinu

Nilai Silhouette Score yang mendekati 1 menunjukkan bahwa data dikelompokkan dengan sangat baik, sedangkan nilai mendekati 0 menunjukkan ketidakyakinan terhadap keanggotaan klaster.



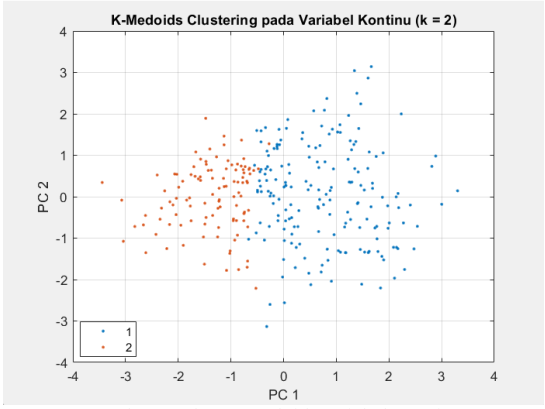
Gambar 2. Grafik Silhouette Score Variabel Kontinu

Grafik pada Gambar 2 tersebut menunjukkan grafik visualisasi nilai silhouette untuk tiap nilai k. Terlihat bahwa nilai tertinggi dicapai saat $k = 2$, yaitu sebesar 0.3023. Meskipun nilai pada $k = 3$ sedikit lebih rendah (0.2992), skor tersebut terus mengalami penurunan pada $k = 4$ dan $k = 5$. Penurunan ini menandakan bahwa pemaksaan jumlah klaster lebih dari dua justru mengurangi kualitas pengelompokan karena terjadi overlapping antar kelompok.

Dengan demikian, struktur dua klaster dapat dikatakan paling optimal dalam memisahkan pasien berdasarkan atribut numerik, seperti tekanan darah, usia, kolesterol, dan detak jantung maksimum.

3.2.2 Visualisasi PCA Clustering (Variabel Kontinu)

Untuk memvisualisasikan hasil clustering variabel kontinu, digunakan metode Principal Component Analysis (PCA) untuk mereduksi dimensi ke dua komponen utama (PC1 dan PC2). Gambar 3 menunjukkan penyebaran data dalam dua dimensi utama, di mana terdapat dua klaster yang cukup terpisah secara visual.



Gambar 3. Plot K-Medoid Variabel Kontinu

Hal ini menunjukkan bahwa atribut-atribut seperti usia, tekanan darah, dan kadar kolesterol memiliki kontribusi signifikan dalam membentuk pola clustering. Cluster biru mewakili kelompok pasien dengan status klinis stabil, sementara cluster oranye menunjukkan kelompok dengan abnormalitas yang lebih mencolok, yang berkorelasi dengan risiko kardiovaskular yang lebih tinggi.

3.3. Evaluasi dan Visualisasi Clustering pada Variabel Diskrit

3.3.1 Evaluasi Silhouette Score (Variabel Diskrit)

Clustering dilakukan pada variabel diskrit menggunakan algoritma K-Medoids dengan metrik Hamming Distance, yang ideal digunakan untuk data kategorikal seperti jenis kelamin, tipe nyeri dada, hasil EKG, dan indikasi angina akibat olahraga. Tujuan dari proses ini adalah untuk mengelompokkan pasien berdasarkan kesamaan atribut diskrit tanpa perlu label diagnosis eksplisit.

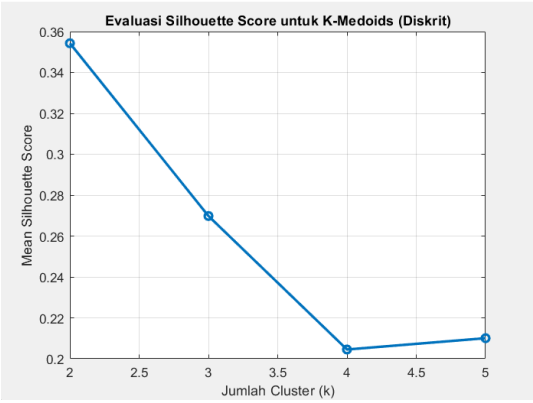
Untuk mengevaluasi performa clustering, digunakan metrik Mean Silhouette Score, dengan variasi jumlah klaster k dari 2 hingga 5. Hasil evaluasi ini dapat dilihat pada Tabel 2 berikut:

Jumlah Cluster (k)	Mean Silhouette Score
2	0.3543
3	0.2583
4	0.203
5	0.2058

Tabel 2 Silhouette Score Variabel Diskrit

Dari tabel tersebut, dapat dilihat bahwa nilai Silhouette Score tertinggi diperoleh saat $k = 2$, yaitu sebesar 0.3543, menunjukkan pemisahan klaster yang paling optimal. Nilai tersebut menurun cukup drastis saat $k = 3$ hingga $k = 5$, yang menandakan bahwa jumlah klaster yang lebih besar tidak memberikan peningkatan pemisahan antar data, bahkan cenderung menurunkan kualitas klasterisasi.

Untuk memperkuat pemahaman terhadap pola ini, digunakan juga representasi visual dalam bentuk grafik Silhouette Score, yang ditampilkan pada Gambar 4. Grafik ini memudahkan pengamatan terhadap tren penurunan nilai Silhouette Score seiring bertambahnya jumlah klaster.



Gambar 4.Grafik Silhouette Score Variabel Diskrit

Gambar 4 menunjukkan dengan jelas bahwa puncak kurva terjadi saat $k = 2$, dan setelahnya kurva menurun tajam. Secara visual, grafik ini memperkuat kesimpulan bahwa dua klaster sudah cukup representatif untuk membedakan pasien berdasarkan karakteristik kategorikal yang ada.

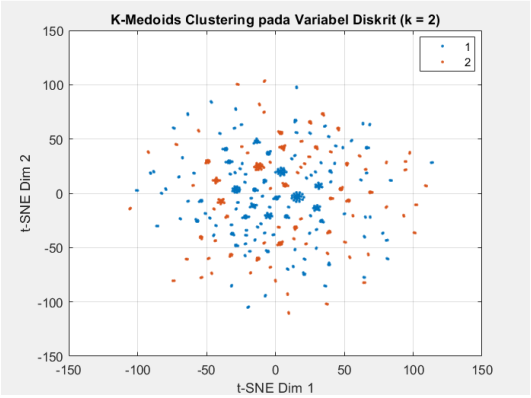
Dari sisi klinis, hasil pengelompokan ini mengindikasikan bahwa:

1. Cluster 1 berisi pasien dengan kondisi umum yang stabil, seperti nyeri dada atipikal atau non-angina, hasil EKG normal, dan tidak mengalami angina saat latihan.
2. Cluster 2 menggambarkan pasien dengan gejala klinis yang lebih berat, seperti angina tipikal, abnormalitas EKG, serta angina saat olahraga.

Dengan demikian, hasil clustering ini dapat dimanfaatkan untuk deteksi dini dan pengelompokan risiko pada pasien jantung berdasarkan informasi kategorikal.

3.3.2 Visualisasi t-SNE Clustering (Variabel Diskrit)

Visualisasi hasil clustering diskrit dilakukan menggunakan t-SNE (t-distributed Stochastic Neighbor Embedding). Gambar 5 menunjukkan penyebaran dua klaster berdasarkan kemiripan kategori pasien. Meski titik-titik tampak tersebar, pola pengelompokan tetap dapat diidentifikasi.



Gambar 5. Plot K-Medoid Variabel Diskrit

T-SNE efektif dalam memproyeksikan data berdimensi tinggi ke dalam ruang dua dimensi, memungkinkan pengamatan visual terhadap kecenderungan klaster berdasarkan karakteristik diskrit pasien. Secara klinis, cluster pertama terdiri dari pasien dengan kondisi normal (seperti nyeri dada tidak khas, EKG normal, tidak ada angina saat olahraga), sedangkan cluster kedua mencerminkan pasien dengan gejala yang lebih serius dan berisiko tinggi.

3.4 Analisis Gabungan dan Interpretasi Klinis

Berdasarkan evaluasi Silhouette Score dan visualisasi untuk variabel kontinu dan diskrit, ditemukan bahwa jumlah klaster optimal adalah k = 2. Hal ini berlaku untuk kedua jenis data, yang menunjukkan bahwa data pasien dapat dikategorikan menjadi dua kelompok besar berdasarkan profil klinis mereka yang dapat dilihat pada Tabel 3 berikut :

Tabel 3 Perbandingan Silhouette Score

Jenis Data	Metode Jarak	Silhouette Score Tertinggi (k=2)
Kontinu	Euclidean	0.3023
Diskrit	Hamming	0.3543

Pengelompokan ini secara tidak langsung memisahkan pasien berdasarkan risiko kardiovaskular. Cluster pertama mencerminkan pasien dengan kondisi klinis yang relatif stabil dan parameter normal. Sementara itu, cluster kedua menunjukkan pasien dengan sejumlah indikator klinis yang tidak normal, seperti tekanan darah tinggi, EKG abnormal, dan nyeri dada khas.

Dengan demikian, kombinasi metode PCA dan K-Medoids pada data kontinu, serta Hamming dan K-Medoids pada data diskrit terbukti efektif dalam mengidentifikasi pola risiko penyakit jantung, bahkan tanpa data label diagnosis. Hasil ini dapat digunakan dalam praktik medis untuk mendukung deteksi dini, skrining massal, dan perencanaan intervensi yang lebih tepat sasaran.

4. Kesimpulan

Penerapan Principal Component Analysis (PCA) secara signifikan meningkatkan efisiensi dan interpretabilitas analisis data klinis dengan mereduksi dimensi atribut numerik tanpa kehilangan informasi varians yang substansial. Transformasi PCA memungkinkan algoritma K-Medoids melakukan pengelompokan yang lebih stabil, karena bekerja dengan medoid sebagai representasi pusat klaster. Evaluasi kualitas clustering menggunakan Mean Silhouette Score menunjukkan nilai tertinggi saat $k = 2$, yaitu sebesar 0.3023 untuk data kontinu (menggunakan jarak Euclidean) dan 0.3543 untuk data diskrit (menggunakan jarak Hamming). Sebelum proses klasterisasi, deteksi outlier dilakukan menggunakan Mahalanobis Distance, dengan total 42 outlier dihapus untuk menjaga konsistensi struktur klaster.

Visualisasi hasil klasterisasi menggunakan PCA (untuk variabel kontinu) dan t-SNE (untuk variabel diskrit) menunjukkan pemisahan dua klaster yang jelas. Secara klinis, klaster pertama merepresentasikan pasien dengan parameter stabil seperti hasil EKG normal dan tidak mengalami angina saat aktivitas fisik. Sementara itu, klaster kedua mengelompokkan pasien dengan karakteristik abnormal seperti angina tipikal dan depresi segmen ST. Pendekatan unsupervised learning ini, yang menggabungkan PCA, K-Medoids, serta metrik jarak yang sesuai, menunjukkan potensi dalam melakukan segmentasi risiko penyakit kardiovaskular tanpa memerlukan label diagnosis eksplisit.

REFERENSI

[1] A. O. Salau, T. A. Assegie, E. D. Markus, J. N. Eneh, and T. I. Ozue, "Prediction of the risk of developing heart disease using logistic regression," International Journal of Electrical & Computer Engineering (2088-8708), vol. 14, no. 2, 2024.

[2] D. Triyono, R. Liani, A. Wahyu Utami, S. Tristiyanti, and A. Supriatna, “PENYAKIT JANTUNG KORONER DI INDONESIA: PERAN FAKTOR RISIKO DAN UPAYA PENCEGAHAN”, HUMANIS: Jurnal Ilmu-Ilmu Sosial dan Humaniora, vol. 17, no. 1, pp. 86-94, Jan. 2025.

[3] P. Sathyaprakash, P. Alagarsundaram, M. V. Devarajan, A. Alkhayyat, P. Poovendran, D. R. Rani, and V. Savitha, "Medical practitioner-centric heterogeneous network powered efficient E-Healthcare risk prediction on Health Big Data," International Journal of Cooperative Information Systems, vol. 34, no. 02, p. 2450012, 2025.

[4] P. Sree, K. Joshi, and S. Jadhav, "A comprehensive analysis on risk prediction of heart disease using machine learning models," Int. J. Res. Inf. Technol. Comput. Commun., vol. 11, no. 11s, p. 8295, 2023.

[5] T. M. Usman, Y. K. Saheed, I. Djitog, and A. Nsang, "Diabetic retinopathy detection using principal component analysis multi-label feature extraction and classification," International Journal of Cognitive Computing in Engineering, vol. 4, pp. 78–88, 2023.

[6] D. Hediyati dan I. M. Suartana, “Penerapan Principal Component Analysis untuk Reduksi Dimensi pada Data Pertanian,” Jurnal Matematika, vol. 8, no. 1, pp. 45–52, 2020.

[7] R. A. et al., "Perbandingan Algoritma K-Means dan K-Medoid dalam Pengelompokan Data Pasien Berdasarkan Rekam Medis di Puskesmas M. Thaha Bengkulu Selatan," Journal of Science and Social Research, vol. 6, no. 3, pp. 580–586, Oct. 2023.

[8] D. R. P. Sari, "Metode Principal Component Analysis (PCA) sebagai Penanganan Asumsi Multikolinearitas," *PARAMETER: Jurnal Matematika, Statistika dan Terapannya*, vol. 2, no. 02, pp. 115–124, 2023.

[9] A. Riyadi and F. Fauziah, "Algoritma Principal Component Analysis untuk Meningkatkan Performa Fuzzy C-Means pada Klasterisasi Dataset Berdimensi Tinggi," *Jurnal Ilmiah Teknologi dan Rekayasa*, vol. 29, no. 2, pp. 99–115, 2024.

[10] I. N. Y. Setyawan and L. P. A. S. Tjahyanti, "Optimasi Sistem Pengenalan Wajah Dengan Teknik Pengolahan Citra Untuk Meningkatkan Akurasi dan Efisiensi," *KOMTEKS*, vol. 3, no. 1, 2024.

[11] R. A. Yusda, R. Risnawati, S. Santoso, P. Z. M. Siregar, and W. P. Nurani, "Analisa Model Clustering Untuk Pemetaan Kualitas Lulusan Mahasiswa Berdasarkan Dataset Tracer Study," *TAMIKA: Jurnal Tugas Akhir Manajemen Informatika & Komputerisasi Akuntansi*, vol. 4, no. 2 (SEMNASTIK), pp. 18–23, 2024.

[12] J. E. Simarmata, M. G. B. Bifel, and F. Mone, "Identifikasi Pengaruh Fasilitas Belajar Terhadap Hasil Belajar Matematika Peserta Didik Menggunakan Pendekatan Structural Equation Modeling," *Jurnal Jendela Matematika*, vol. 3, no. 01, pp. 27–35, 2025.

[13] R. Meiyanti, M. M. Munauwar, R. Fitria, and H. A. Kautsar, "Implementasi Algoritma K-Medoid pada Clustering Sayuran Unggulan di Kabupaten Aceh Utara," *TEKNIKA*, vol. 19, no. 1, pp. 327–337, 2025.

[14] N. Almajid, P. D. Atika, and K. F. Ramdhania, "Regional Mapping Based on Tourism Destinations in West Java: K-Medoid Clustering Analysis," International Journal on Information and Communication Technology (IJoICT), vol. 10, no. 2, pp. 287–296, 2024.

[15] B. He, R. Guo, M. Li, Y. Jing, Z. Zhao, W. Zhu, C. Zhang, C. Zhang, and D. Ma, "The fractal or scaling perspective on progressively generated intra-urban clusters from street junctions," International Journal of Digital Earth, vol. 16, no. 1, pp. 1944–1961, 2023, doi: 10.1080/17538947.2023.2218118.

[16] K. Cabello-Solorzano, I. Ortigosa de Araujo, M. Peña, L. Correia, and A. J. Tallón-Ballesteros, "The impact of data normalization on the accuracy of machine learning algorithms: a comparative analysis," in International Conference on Soft Computing Models in Industrial and Environmental Applications, 2023, pp. 344–353.

Sebastian Wibowo, Mahasiswa S1 Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara, Jakarta.

Dillon Majesson, Mahasiswa S1 Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara, Jakarta.

Junardi Chailesia, Mahasiswa S1 Program Studi Teknik Informatika, Fakultas Teknologi Informasi, Universitas Tarumanagara, Jakarta.