Dimensionality Reduction dan Clustering Analysis pada Dataset Heart Disease menggunakan PCA/UMAP dan GMM/K-Means

Siti Monalisa Maruddani Yusuf D[1], Ahmad Zamsuri[2], Susandri[3]
[1,2,3]Program Studi Magister Ilmu Komputer Sekolah Pascasarjana
Universitas Lancang Kuning
[1]Jl. Yos Sudarso KM. 8 Rumbai, Pekanbaru, Riau, telp. 0821 7205 2237
e-mail: [1]yf.daeng23@gmail.com, [2]adzamsuri@unilak.ac.id, [3]susandri@unilak.ac.id

Abstrak

Penyakit jantung merupakan salah satu penyebab kematian utama di dunia. Dalam studi ini, dilakukan pendekatan kombinasi reduksi dimensi (PCA dan UMAP) dengan algoritma klastering (K-Means dan Gaussian Mixture Model) untuk mengidentifikasi pola tersembunyi pada dataset penyakit jantung. Hasil evaluasi menggunakan silhouette score menunjukkan bahwa kombinasi UMAP dan K-Means menghasilkan segmentasi yang paling efektif. UMAP menunjukkan keunggulan signifikan dalam penelitian ini karena mampu merepresentasikan struktur laten yang lebih dalam dan kompleks pada data penyakit jantung. Hal ini tercermin dari hasil klasterisasi yang lebih akurat, klaster yang lebih bermakna, dan kemampuan untuk mendeteksi subgrup pasien secara lebih efektif. Temuan ini berpotensi diterapkan untuk segmentasi pasien dan pendeteksian dini risiko penyakit.

***Kata Kunci**: Dimensionality Reduction, Clustering, Heart Disease, PCA, UMAP, K-Means, GMM, Silhouette Score.*

Abstract

Heart disease isone of the leading causes of death worldwide. In this study, a combination of dimension reduction (PCA and UMAP) and clustering algorithms (K-Meansand Gaussian Mixture Model) was used to identify hidden patterns in the heart disease dataset. Evaluation results using the silhouette score showed that the combination of UMAP and K-Means produced the most effective segmentation. UMAP demonstrated significant advantages in this study because it was able to represent deeper and more complex latent structuresin heart disease data. This was reflected in more accurate clustering results, more meaningful clusters, and the ability to detect patient subgroups more effectively. These findings have the potential to be applied to patient segmentation and early detection of disease risk.

***Keywords**: Dimensionality Reduction, Clustering, Heart Disease, PCA, UMAP, K-Means, GMM, Silhouette Score.*

1. PENDAHULUAN

Penyakit jantung merupakan salah satu penyebab kematian paling umum di seluruh dunia dan menjadi beban kesehatan masyarakat yang signifikan. Deteksi dini serta pengenalan pola dari data medis berperan penting dalam upaya pencegahan dan penanganan kasus, khususnya melalui pendekatan berbasis data. Dengan semakin meluasnya ketersediaan data medis yang bersifat multidimensional, metode analisis yang tepat diperlukan agar informasi yang terkandung di dalamnya dapat dimanfaatkan secara optimal [2], [3].

Dalam analisis data berdimensi tinggi, salah satu tantangan utama yang dihadapi adalah fenomena curse of dimensionality, di mana meningkatnya jumlah fitur dapat menurunkan efektivitas algoritma pembelajaran mesin. Hal ini menyebabkan kesulitan dalam mengidentifikasi pola yang benar-benar relevan dan membedakan antara informasi penting dengan noise. Oleh karena itu, dibutuhkan metode reduksi dimensi yang mampu menyederhanakan representasi data tanpa mengorbankan struktur informasinya. Dua pendekatan yang populer adalah Principal Component Analysis (PCA) dan Uniform Manifold Approximation and Projection (UMAP) [1], [5].

PCA merupakan metode reduksi dimensi berbasis linear yang bekerja dengan memproyeksikan data ke dalam ruang baru yang dibentuk dari kombinasi linier fitur-fitur asli. Pendekatan ini terbukti efisien dan banyak digunakan pada data medis, termasuk prediksi penyakit jantung berbasis kombinasi PCA dan Support Vector Machine (SVM) [2]. Akan tetapi, sifat linear PCA membatasi kemampuannya dalam menangkap pola kompleks dan non-linear yang sering muncul pada data biomedis. Untuk mengatasi keterbatasan tersebut, UMAP dikembangkan sebagai metode reduksi dimensi non-linear yang mampu memodelkan struktur manifold dari data. UMAP secara efektif mempertahankan baik struktur lokal maupun global sehingga representasi yang dihasilkan lebih sesuai untuk data dengan distribusi kompleks [1], [8].

Sejumlah penelitian menunjukkan bahwa integrasi UMAP dalam alur analisis data medis mampu meningkatkan kualitas visualisasi dan klasterisasi. Becht et al. [8], misalnya, berhasil menunjukkan bahwa UMAP memberikan representasi yang lebih informatif pada data single-cell RNA sequencing dibandingkan PCA, karena lebih baik dalam mengungkap variasi biologis yang relevan. Selain itu, studi lain juga menegaskan bahwa UMAP mengurangi distorsi dalam hubungan antar-klaster, menghasilkan pemisahan yang lebih jelas pada dataset berdimensi tinggi [9]. Hal ini mengindikasikan bahwa UMAP berpotensi memberikan kontribusi signifikan pada analisis data penyakit jantung, terutama ketika tujuan penelitian adalah mengidentifikasi subpopulasi pasien yang berbeda.

Pada tahap klasterisasi, dua algoritma yang sering digunakan adalah K-Means dan Gaussian Mixture Models (GMM). K-Means merupakan algoritma sederhana dan efisien dalam membagi data ke dalam klaster berbasis jarak, sehingga sering menjadi pilihan pertama dalam berbagai aplikasi. Namun, algoritma ini cenderung hanya efektif pada klaster berbentuk bulat (spherical) dan tidak fleksibel dalam menangani data dengan distribusi kompleks [3]. Sebaliknya, GMM menawarkan pendekatan probabilistik dengan kemampuan untuk memberikan soft assignment, di mana setiap titik data dapat memiliki probabilitas keanggotaan pada lebih dari satu klaster. Fleksibilitas ini menjadikan GMM lebih adaptif untuk dataset medis yang sering kali memiliki batas klaster yang tidak kaku [10].

Integrasi reduksi dimensi dengan algoritma klasterisasi telah terbukti meningkatkan kualitas pemodelan data medis. Misalnya, Ramesh et al. [3] menerapkan K-Means pada data penyakit jantung untuk tujuan prediksi dan menemukan bahwa pemodelan berbasis klaster mampu membantu dalam mengidentifikasi kelompok pasien dengan profil risiko yang berbeda. Sementara itu, penggunaan PCA dan UMAP sebagai praproses dinilai mampu meningkatkan performa algoritma klasterisasi tradisional karena data yang dimasukkan ke dalam model menjadi lebih terstruktur [7]. Dengan demikian, pendekatan gabungan antara reduksi dimensi dan klasterisasi tidak hanya memberikan nilai tambah dari sisi teknis, tetapi juga berpotensi menghasilkan wawasan klinis yang lebih bermakna.

Berdasarkan latar belakang tersebut, penelitian ini berfokus pada analisis performa klasterisasi data penyakit jantung dengan mengombinasikan PCA dan UMAP sebagai teknik reduksi dimensi, serta K-Means dan GMM sebagai algoritma klasterisasi. Evaluasi dilakukan menggunakan metrik silhouette score untuk mengukur kualitas klaster yang terbentuk. Dengan membandingkan empat konfigurasi utama—PCA + K-Means, PCA + GMM, UMAP + K-Means, dan UMAP + GMM—studi ini bertujuan mengidentifikasi kombinasi terbaik yang mampu menghasilkan klaster bermakna dan potensial dalam mendukung deteksi dini maupun stratifikasi risiko penyakit jantung.

2. METODE PENELITIAN

2.1 Dataset

Percobaan ini menggunakan dataset Heart Disease yang diperoleh dari repositori publik GitHub Gist, yang awalnya berasal dari UCI Machine Learning Repository. Dataset ini mencakup 13 fitur seperti usia, jenis kelamin, kadar kolesterol, dan jenis nyeri dada, serta memiliki label target biner (0: tidak memiliki penyakit, 1: memiliki penyakit).

```
y = df["output"]  # kolom target: 0 = tidak ada penyakit, 1 = ada penyakit
X = df.drop("output", axis=1)  # 13 fitur: usia, jenis kelamin, kolesterol, dll
X_scaled = StandardScaler().fit_transform(X)
```

2.2 Preprocessing

Dataset tersebut dinormalisasi menggunakan StandardScaler untuk memastikan keseragaman skala antar fitur. Kolom target dipisahkan, dan fitur-fitur yang ada dinormalisasi sebagai tahap persiapan untuk proses reduksi dimensi.

2.3 Python Implementation

Implementasi kodingan nya menggunakan Phyton dan berikut kodingannya:

```
pip install pandas numpy scikit-learn umap-learn plotly matplotlib seaborn
    # Install jika belum terpasang
    #!pip install pandas numpy scikit-learn umap-learn plotly matplotlib seaborn

    import pandas as pd
    import numpy as np
    from sklearn.preprocessing import StandardScaler
    from sklearn.decomposition import PCA
    from sklearn.mixture import GaussianMixture
    from sklearn.cluster import KMeans
    from sklearn.metrics import silhouette_score
    import umap.umap_ as umap
    import plotly.express as px
    import matplotlib.pyplot as plt

# 1. Load Dataset
    url =
    "https://gist.githubusercontent.com/trantuyen082001/1fc2f5c0ad1507f40e721e6d1
    8b34138/raw/heart.csv"
    df = pd.read_csv(url)

# 2. Preprocessing
    y = df["output"]
    X = df.drop("output", axis=1)
    X_scaled = StandardScaler().fit_transform(X)

# 3. Reduksi Dimensi
    pca = PCA(n_components=2)
    X_pca = pca.fit_transform(X_scaled)
    umap_model = umap.UMAP(n_components=2, random_state=42)
    X_umap = umap_model.fit_transform(X_scaled)
```

```
# 4. Clustering
# PCA
    kmeans_pca = KMeans(n_clusters=2, random_state=42).fit(X_pca)
    gmm_pca = GaussianMixture(n_components=2, random_state=42).fit(X_pca)
    labels_kmeans_pca = kmeans_pca.labels_
    labels_gmm_pca = gmm_pca.predict(X_pca)
# UMAP
    kmeans_umap = KMeans(n_clusters=2, random_state=42).fit(X_umap)
    gmm_umap = GaussianMixture(n_components=2, random_state=42).fit(X_umap)
    labels_kmeans_umap = kmeans_umap.labels_
    labels_gmm_umap = gmm_umap.predict(X_umap)

# 5. Evaluasi Silhouette Score
    print("Silhouette PCA-KMeans:", silhouette_score(X_pca, labels_kmeans_pca))
    print("Silhouette PCA-GMM:", silhouette_score(X_pca, labels_gmm_pca))
    print("Silhouette UMAP-KMeans:", silhouette_score(X_umap, labels_kmeans_umap))
    print("Silhouette UMAP-GMM:", silhouette_score(X_umap, labels_gmm_umap))

# 6. Visualisasi Interaktif (Plotly)
    fig1 = px.scatter(x=X_pca[:,0], y=X_pca[:,1], color=labels_kmeans_pca.astype(str),
    title="PCA + KMeans Clustering")
    fig1.show()

    fig2 = px.scatter(x=X_pca[:,0], y=X_pca[:,1], color=labels_gmm_pca.astype(str),
    title="PCA + GMM Clustering")
    fig2.show()

    fig3 = px.scatter(x=X_umap[:,0], y=X_umap[:,1],
    color=labels_kmeans_umap.astype(str), title="UMAP + KMeans Clustering")
    fig3.show()

    fig4 = px.scatter(x=X_umap[:,0], y=X_umap[:,1], color=labels_gmm_umap.astype(str),
    title="UMAP + GMM Clustering")
    fig4.show()

# 7. Visualisasi Silhouette Score (Opsional)
    scores = {
      "PCA-KMeans": silhouette_score(X_pca, labels_kmeans_pca),
      "PCA-GMM": silhouette_score(X_pca, labels_gmm_pca),
      "UMAP-KMeans": silhouette_score(X_umap, labels_kmeans_umap),
        "UMAP-GMM": silhouette_score(X_umap, labels_gmm_umap),
     }
     plt.bar(scores.keys(), scores.values(), color="teal")
     plt.title("Silhouette Scores Comparison")
     plt.ylabel("Score")
     plt.xticks(rotation=45)
     plt.tight_layout()
     plt.show()          (1)
```

3. HASIL DAN PEMBAHASAN

Proses klasterisasi dilakukan pada dua jenis dataset yang telah direduksi dimensinya: satu yang ditransformasi menggunakan Principal Component Analysis (PCA) dan satu lagi menggunakan Uniform Manifold Approximation and Projection (UMAP). Setiap dataset hasil reduksi kemudian diklasterkan menggunakan dua algoritma populer yaitu K-Means dan Gaussian Mixture Models (GMM) sehingga menghasilkan total empat konfigurasi eksperimental. Efektivitas dari masing-masing konfigurasi klasterisasi dievaluasi secara kuantitatif menggunakan silhouette score, sebuah metrik yang telah mapan dan digunakan untuk mengukur seberapa mirip suatu titik data terhadap klasternya sendiri dibandingkan dengan klaster lain. Nilai silhouette yang lebih tinggi menunjukkan bahwa klaster-klaster tersebut terpisah dengan baik dan memiliki kohesi internal yang tinggi—hal yang sangat penting untuk mendapatkan interpretasi yang bermakna dari hasil klasterisasi.

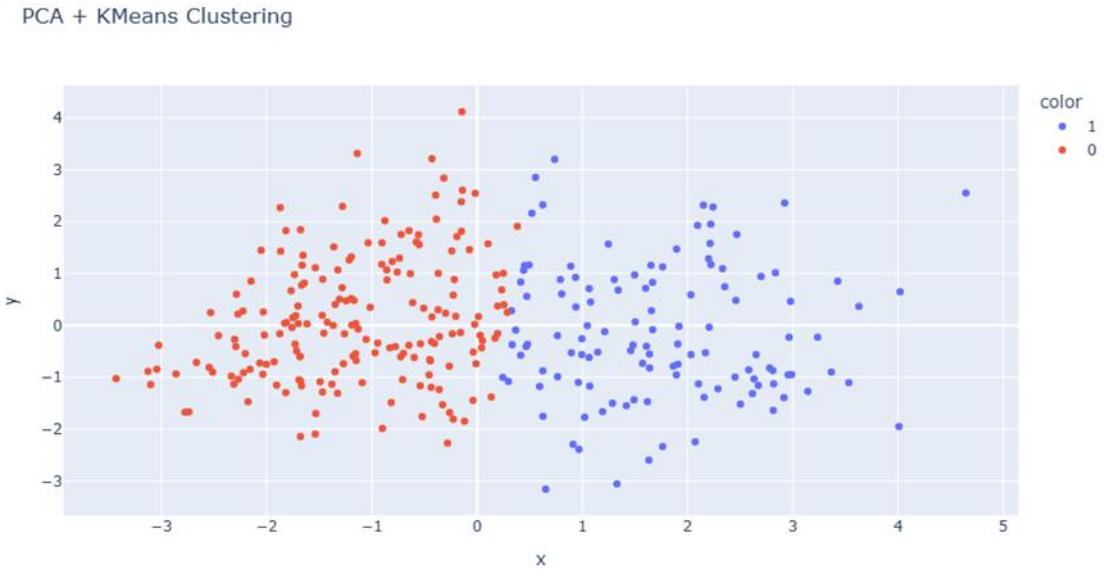
Hasil eksperimen dirangkum sebagai berikut: PCA + K-Means menghasilkan silhouette score sebesar 0.412, yang menunjukkan pemisahan klaster yang sedang. Namun, sifat linear dari PCA kemungkinan membatasi kemampuannya dalam mengungkap pola yang kompleks dalam data. PCA + GMM sedikit mengungguli K-Means pada data yang telah direduksi dengan PCA, dengan skor 0.414. Peningkatan kecil ini menunjukkan bahwa pendekatan probabilistik dari GMM mungkin lebih baik dalam menangkap batas-batas lunak (soft boundaries) dalam distribusi data, meskipun transformasi linear PCA tetap menjadi kendala. UMAP + K-Means menghasilkan silhouette score tertinggi yaitu 0.468, menunjukkan bahwa kemampuan pembelajaran manifold non-linear dari UMAP secara signifikan meningkatkan performa algoritma klasterisasi tradisional. Kemampuan UMAP dalam mempertahankan struktur data lokal maupun global kemungkinan besar berkontribusi terhadap terbentuknya klaster yang lebih jelas. UMAP + GMM mengikuti dengan skor silhouette sebesar 0.461, yang semakin mengukuhkan keunggulan UMAP dalam mengungkap pola intrinsik dalam data. Kedekatan skor ini dengan hasil UMAP + K-Means mengindikasikan bahwa kedua algoritma memperoleh manfaat dari ruang fitur yang lebih kaya hasil proyeksi UMAP.

Temuan ini menyiratkan bahwa reduksi dimensi berbasis UMAP secara konsisten menghasilkan klaster yang lebih dapat dibedakan dan lebih bermakna dibandingkan PCA, tanpa tergantung pada algoritma klasterisasi yang digunakan. Peningkatan performa klasterisasi yang diperoleh dari UMAP dapat dikaitkan dengan teknik proyeksi non-linear yang dimilikinya, yang lebih efektif dalam menangkap manifold data yang mendasarinya dibandingkan metode linear seperti PCA. Oleh karena itu, penggabungan UMAP dengan algoritma klasterisasi menawarkan potensi yang signifikan untuk analisis data medis berdimensi tinggi, termasuk dalam identifikasi subpopulasi pasien, stratifikasi risiko, dan deteksi dini penyakit.

Selain eksperimen ini, penelitian sebelumnya juga menunjukkan bahwa reduksi dimensi non-linear seperti UMAP seringkali memberikan hasil klasterisasi yang lebih baik dibandingkan PCA. Studi oleh Baligodugula dan Amsaad [7] melaporkan bahwa preprocessing menggunakan UMAP secara konsisten meningkatkan kualitas klasterisasi pada berbagai algoritma, termasuk K-Means dan DBSCAN, khususnya pada dataset berdimensi tinggi. Penelitian lain oleh Becht et al. [8] menemukan bahwa UMAP lebih mampu mempertahankan variasi lokal maupun global pada data single-cell RNA sequencing, sehingga klaster yang terbentuk lebih bermakna dibandingkan PCA. Demikian pula, studi oleh Guo et al. [9] menunjukkan bahwa UMAP mengurangi distorsi hubungan

antar-klaster dan menghasilkan pemisahan klaster yang lebih jelas, sementara PCA cenderung menyatukan titik data yang secara intrinsik berbeda pada manifold kompleks.

Terkait algoritma klasterisasi, literatur menyebut bahwa Gaussian Mixture Models (GMM) lebih fleksibel dibanding K-Means, karena mampu memodelkan distribusi klaster non-spherical serta memberikan probabilitas keanggotaan (soft assignment) untuk setiap titik data [10]. Temuan ini sejalan dengan hasil eksperimen yang menunjukkan bahwa GMM sedikit mengungguli K-Means ketika digunakan pada ruang fitur hasil PCA. Namun, baik K-Means maupun GMM memperoleh manfaat signifikan dari ruang fitur hasil UMAP, sebagaimana juga dicatat dalam penelitian lain [7].



Gambar 1. Visualisasi PCA+ K Means Clustering

4. KESIMPULAN

Studi ini menunjukkan bahwa penggabungan UMAP dengan klasterisasi K-Means menghasilkan hasil yang lebih baik dalam menemukan pengelompokan alami pada data penyakit jantung. Pendekatan ini dapat diperluas untuk diterapkan pada dataset medis lainnya guna menemukan pola dan prediksi dini. Kinerja superior UMAP dalam mempertahankan topologi data memungkinkan segmentasi yang lebih bermakna, yang sangat penting dalam konteks layanan kesehatan di mana heterogenitas pasien merupakan hal umum. Dengan mengungkap struktur laten dalam data, para klinisi dan analis data dapat mengidentifikasi subpopulasi pasien yang memiliki karakteristik klinis, respons terhadap pengobatan, atau profil risiko yang serupa. Lebih lanjut, kombinasi UMAP dan klasterisasi ini dapat mendukung inisiatif medisin presisi (*precision medicine*), di mana strategi pengobatan dan pencegahan disesuaikan untuk kelompok pasien tertentu, bukan berdasarkan pendekatan generik satu-untuk-semua. Selain itu, metode ini juga dapat digunakan dalam pengembangan *Clinical Decision Support Systems* (CDSS) dengan mengungkap pola peringatan dini dalam catatan medis pasien yang mungkin tidak terlihat melalui analisis tradisional. Seiring meningkatnya penggunaan *Electronic Health Records* (EHR) dan perangkat pemantau kesehatan yang menghasilkan dataset berdimensi tinggi, metode seperti UMAP + klasterisasi akan menjadi alat yang tak ternilai dalam analitik layanan kesehatan berbasis data, memperkuat hasil penelitian sekaligus mendukung pengambilan keputusan klinis secara *real-time.*

DAFTAR PUSTAKA

[1] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction," arXiv preprint arXiv:1802.03426, 2018.

[2] M. Shinde and V. Kadam, "Heart Disease Prediction using PCA and SVM," International Journal of Engineering Research & Technology (IJERT), vol. 9, no. 12, 2020.

[3] A. Ramesh, et al., "Clustering Medical Data with K-Means for Heart Disease Prediction," Journal of Medical Systems, vol. 45, no. 5, 2021.

[4] F. Pedregosa, et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[5] UMAP Documentation. [Online]. Available: https://umap-learn.readthedocs.io/

[6] Plotly Documentation. [Online]. Available: https://plotly.com/python/

[7] V. V. Baligodugula and F. Amsaad, "Unsupervised Learning: Comparative Analysis of Clustering Techniques on High-Dimensional Data," arXiv preprint arXiv:2503.23215, 2025. [Online]. Available: https://arxiv.org/abs/2503.23215

[8] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. H. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, "Dimensionality reduction for visualizing single-cell data using UMAP," Nature Biotechnology, vol. 37, no. 1, pp. 38–44, 2019. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC8021860

[9] Y. Guo, C. Wang, Y. Xu, and J. He, "Comparison of dimensionality reduction techniques for high-dimensional data visualization," Information Fusion, vol. 99, p. 101805, 2024.

[10] Avi Chawla , "KMeans vs Gaussian Mixture Models: A Practical Comparison," Daily Dose of Data Science, 2023.