Jurnal Ilmiah Teknik Informatika dan Komunikasi
Volume 5. Nomor. 1 Maret 2025
E-ISSN: 2827-7945 ; P-ISSN: 2827-8127, Hal 459-476
DOI: https://doi.org/10.55606/juitik.v5i1.1329
Available online at: https://journal.sinov.id/index.php/juitik

Pendekatan Data Mining untuk Prediksi Risiko Diabetes Menggunakan *K-Means Clustering*

Fairuzza Qushai Simatupang[1*], Gabriel Sthefanus Silalahi[2], Siska Ananda[3], Shera Br Sitepu[4]
[1-5]Universitas Satya Terra Bhinneka, Indonesia

Alamat: Jl. Sunggal Gg. Bakul, Sunggal, Kec. Medan Sunggal, Kota Medan, Sumatera Utara
Korespondensi penulis: fairuzzaqushai2004@gmail.com

Abstract. *Diabetes Mellitus is a Chronic disease with a continuously increasing global prevalance, including in Indonesia, posing a serious challenge to healthcare systems. Early detection and risk stratification of individuals are crucial for implementing effective prevention and management strategies. This research utilizes the K-Means Clustering algorithm, an unsupervised learning method in data mining, to identify and group individuals based on their diabetes risk profiles from available medical data. The dataset used is the popular Pima Indian Indian Dataset, comprising eight clinical attributes such as glucose level, blood pressure, skin thickness, BMI, and age. The methodological process includes data preprocessing with standardization using StandardScaler, determining the optimal number of clusters through the Elbow Method, Implementing K-Means clustering, and evaluating clustering quality using the Silhouette Coefficient. The research results demonstrate that this algorithm can group patients into low, medium, and high-risk categories with sufficient cluster accuracy. This approach can be used as a supporting tool in medical decision systems.*

Keywords: *Data Mining, K-Means Cluster, Diabetes Risk, Clustering.*

Abstrak. Diabetes Melitus merupakan penyakit kronis dengan prevalensi global yang terus meningkat, termasuk di Indonesia, menimbulkan tantangan serius bagi sistem kesehatan. Deteksi dini dan pengelompokan risiko individu sangat krusial untuk implementasi strategi perpecahan dan manajemen yang efektif. Penelitian ini memanfaatkan algoritma *K-Means Clustering*, sebuah metode *unsupervised learning* dalam data mining, untuk mengidentifikasi dan mengelompokkan individu berdasarkan profil risiko diabetes mereka dari data medis yang tersedia. Dataset yang digunakan adalah Pima Indian Dataset yang populer, mencakup delapan atribut klinis seperti kadar glukosa, tekanan darah, ketebalan kulit, BMI, dan usia. Proses metodologi meliputi pra-pemrosesan data dengan standarisasi menggunakan *StandardScaler,* penentuan jumlah kluster optimal melalui Elbow Method, implementasi K-Means Clustering, serta evaluasi kualitas pengelompokan menggunakan Silhouette Coefficient. Hasil penelitian menunjukkan bahwa algoritma ini mampu mengelompokkan pasien ke dalam kategori risiko rendah, sedang, dan tinggi dengan akurasi Cluster yang cukup baik. Pendekatan ini dapat digunakan sebagai alat bantu dalam sistem pendukung keputusan medis.

Kata kunci: Data Mining, *K-means Cluster*, Risiko Diabetes, Pengelompokan

1. LATAR BELAKANG

Penyakit Diabetes Melitus (DM) adalah, kondisi metabolik kronis yang ditandai oleh kadar glukosa darah tinggi (hiperglikemia), yang disebabkan oleh efek sekresi insulin, kerja insulin, atau keduanya (WHO, 2020). Prevalensi diabetes telah mencapai tingkat epidemi global, dengan proyeksi peningkatan yang signifikan di tahun-tahun mendatang. International Diabetes Federation (IDF) Melaporkan bahwa pada tahun 2021, sekitar 573 juta orang dewasa di seluruh dunia hidup dengan diabetes, dan angka ini diperkirakan akan mencapai 783 juta pada tahun 2045 (IDF Diabetes Atlas, 10th ed., 2021). Di Indonesia, peningkatan kasus diabetes juga menjadi perhatian serius, sering kali dengan banyak

Received: Juni 01, 2025; Revised: Juni 19, 2025; Accepted: Juni 27, 2025; Pubhlished: Juni 30, 2025

individu yang tidak terdiagnosis hingga stadium lanjut, yang meningkatkan risiko komplikasi parah komplikasi parah seperti penyakit kardiovaskular, neuropati, nefropati dan retinopati.

Deteksi dini dan identifikasi kelompok risiko menjadi kunci dalam upaya pencegahan dan pengelolaan diabetes yang efektif. Dengan kemajuan teknologi informasi dan akumulasi data medis yang masif, data mining menawarkan peluang besar untuk mengekstraksi wawasan berharga dari dataset klinis. Data mining melibatkan penggunaan algoritma dan teknik statistik untuk menentukan pola, trend, dan hubungan tersembunyi dalam kumpulan data besar yang tidak dapat diidentifikasi melalui metode analisis tradisional.

Salah satu teknik *unsupervised learning* yang banyak digunakan dalam data mining adalah **K-Means Clustering.** K-Means adalah algoritma partisi yang bertujuan untuk membagi *n* observasi ke dalam *k* klaster, dimana setiap observasi termasuk dalam klaster dengan rata-rata terdekat (centroid klaster). Algoritma ini cocok untuk mengidentifikasi kelompok alami dalam data tanpa label kelas yang telah ditentukan sebelumnya. Dalam konteks medis, K-Means dapat membantu pengelompokan pasien berdasarkan kemiripan karakteristik klinis mereka, yang berpotensi mengungkap sub-populasi dengan profil risiko penyakit yang berbeda.

penelitian ini bertujuan untuk menerapkan algoritma K-Means Clustering pada data medis pasien untuk mengidentifikasi dan mengelompokan individu ke dalam kategori risiko diabetes yang berbeda (rendah, sedang, dan tinggi). Dengan menganalisis karakteristik setiap klaster, diharapkan dapat memberi kontribusi pada pengembangan sistem pendukung keputusan yang lebih efisien dan berbasis data untuk prediksi dini diabetes, seta memungkinkan pendekatan intervensi yang lebih personalisasi.

2. KAJIAN TEORITIS

Konsep Data Mining

Data Mining, sering disebut juga sebagai knowledge discovery in database (KDD), adalah proses penambangan atau penemuan informasi baru dengan mencari pola atau aturan tertentu dari sejumlah data yang sangat besar. KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola, atau hubungan dalam set data berukuran besar. Definisi lain menyebutkan bahwa Data Mining adalah proses penemuan pola-pola dalam data. Proses ini bisa otomatis atau semi otomatis, dan pola yang ditemukan harus bermakna serta memberikan keuntungan, biasanya keuntungan secara ekonomi. Data yang dibutuhkan dalam data mining ini berjumlah besar. Secara keseluruhan,

data mining adalah serangkaian porses untuk menggali nilai tambahan berupa pengetahuan yang selama ini tidak diketahui secara manual dari suatu kumpulan data.

Algoritma K-Means Clustering

K-Means Clustering adalah unsupervised learning yang bertujuan untuk mempartisi observasi ke cluster, dimana setiap observasi termasuk dalam cluster dengan *mean* terdekat. secara matematis, tujuan utama dari algoritma K-Means adalah meminimalkan Within-Cluster Sum of Squares (WCSS), yang juga dikenal sebagai inersia atau Sum of Squared Errors(SSE). Rumus yang meminimalkan WCSS adalah sebagai berikut :

$$\sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

Dimana :

- k adalah jumlah klaster yang ditentukan contohnya dalam penelitian ini, $k = 3$
- $S\,i$ adalah set dari titik data yang termasuk dalam klaster ke- i
- x adalah titik data individual dalam klaster $S\,i$.
- $\mu\,i$ adalah centroid dari klaster $S\,i$.
- $||\,x - \mu\,i\,||^2$ adalah jarak kuadrat Euclidean antara titik data dan centroid $\mu\,i$.

Proses Iteratif K-Means melibatkan langkah-langkah berikut:

1. Inisialisasi Centroid : Pilih k titik data sebagai centroid awal
2. Langkah Penugasan(Assignment Step) : Setiap titik data ditugaskan ke klaster dengan centroid terdekat. Jarak yang digunakan untuk menentukan kedekatan adalah Jarak Euclidean, yang dirumuskan sebagai berikut:

$$d(x, \mu_i) = \sqrt{\sum_{j=1}^{D} (x_j - \mu_{ij})^2}$$

yang dimana D adalah jumlah dimensi, xj adalah nilai fitur ke-j dari titik data x, dan uij adalah nilai fitur ke-j dari centroid cluster $\mu\,i$

3. Langkah Pembaruan : Dihitung ulang centroid untuk setiap klaster sebagai rata-rata dari semua titik yang ditugaskan ke klaster tersebut. Rumus untuk menghitung centroid baru ($\mu\,i\,new)$ dari klaster $S\,i$ adalah sebagai berikut :

$$\mu_i^{\text{new}} = \frac{1}{|S_i|} \sum_{x \in S_i} x$$

dimana |*S i*| adalah jumlah titik data dalam klaster *S i.*

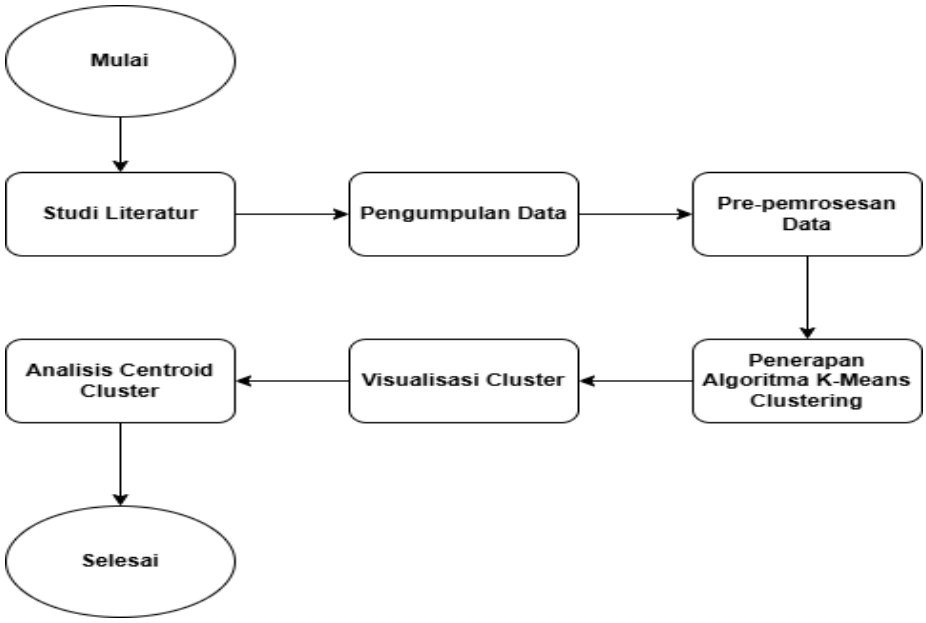
4. Iterasi : Langkah penugasan dan pembaruan diulangi sampai posisi centroid tidak lagi berubah secara signifikan atau jumlah iterasi maksimum tercapai.

Studi Terdahulu

Beberapa penelitian telah memanfaatkan data mining untuk analisis risiko diabetes. Setiawan dan Prasetijo (2020) melakukan analisis K-Means Clustering dalam klasterisasi pasien diabetes, menunjukkan potensi metode ini dalam mengidentifikasikan kelompok pasien. Karyadiputra dan Setiawan (2020) juga menerapkan data mining (menggunakan Decision Tree C4.5, Naive Bayes, Dan K_Nearest Neighbors) untuk prediksi awal kemungkinan terindikasi diabetes, menunjukan pentingnya deteksi dini. Selain itu, Prasatya, Siregar dan Arianto (2020) menggunakan kombinasi metode K-Means dan C4.5 untuk prediksi penderita diabetes, menyoroti efektivitas pendekatan hibrida dalam mengatasi keterbatasan data. penggunaan machine learning dan algoritma seperti Decision Tree atau Random Forest juga sering digunakan untuk prediksi diabetes (Kusumadewi dan Purnomo), namun biasanya memerlukan label kelas yang sudah ada. Penelitian ini berfokus pada pendekatan unsupervised untuk menemukan struktur data tanpa label target awal.

3. METODE PENELITIAN

Metode Penelitian ini mengikuti tahapan standar dalam proyek data mining, dijelaskan dalam bentuk *flowchart* pada gambar 1.



Gambar 1. Metodologi Penelitian

Studi Literatur

Pada tahapan ini, tujuannya adalah mencari dan mengumpulkan teori yang paling cocok untuk topik yang diteliti, seperti jurnal ilmiah, artikel, penelitian sebelumnya, dan referensi lain yang mudah diakses secara online.

Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah *Pima Indian Diabetes Dataset*, yang merupakan dataset *open-source* populer UCI Machine Learning Repository (UCI Machine Learning Repository, Pima Indians Diabetes Dataset). Dataset ini terdiri dari 786 data pasien wanita kekurangan Pima Indian, dengan 8 atribut fitur numerik dan 1 atribut target biner('Outcome', apakah menderita diabetes atau tidak). Dalam konteks *Clustering K-Means,* atribut 'Outcome' tidak digunakan sebagai target pembelajaran karena K-Means adalah algoritma *Unsupervised.*

Beberapa atribut-atribut fitur yang akan digunakan dalam memprediksi diabetes sebagai berikut:

- **Pregnancies** : Jumlah Kehamilan.
- **Glucose** : Konsentrasi glukosa plasma 2 jam dalam tes toleransi glukosa oral (mg/dL).
- **BloodPressure** : Tekanan darah diastolik (mmHg).
- **SkinThickness** : Ketebalan lipatan kulit trisep(mm)
- **Insulin** : Kadar insulin serum 2 jam (mu U/ml).

- **BMI** : Indeks Massa Tubuh (berat badan dalam kg/(tinggi badan dalam m)2).
- **DiabetesPedigreeFunction** : Fungsi silsilah diabetes (menunjukan riwayat diabetes dalam keluarga)
- **Age** : Usia (tahun)

Dibawah ini adalah gambaran data lima baris pertama dari dataset yang akan kami digunakan:

Tabel 1. Dataset

Pregnancies	Glucose	Bloody Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Cluster
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.1	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

Pre-Pemrosesan Data

Tahap pra-pemrosesan data sangat krusial untuk memastikan kualitas data dan mengoptimalkan kinerja algoritma *clustering*. Langkah-langkah yang akan digunakan adalah sebagai berikut:

1. Pemilihan Fitur : Atribut ‘Outcome’ dihilangkan dari dataset karena merupakan label kelas yang tidak diperlukan untuk *unsupervised clustering.* Hanya atribut fitur numerik yang digunakan.
2. Penanganan Nilai Hilang/Nol : Dataset Pima Indian diketahui memiliki nilai ‘0’ pada beberapa kolom (Glucose, Blood Pressure, Skin Thickness, Insulin, BMI) yang secara medis tidak masuk akal misalnya tekanan darah nol (0). Dalam implementasi ini, nilai nol diperlukan sebagai nilai mendalam, nilai-nilai ini seharusnya diidentifikasi dan ditangani sebagai nilai hilang (misalnya, dengan imputasi menggunakan *mean* atau *median* dari kolom tersebut) untuk menghindari distorsi dalam perhitungan *centroid* dan jarak.
3. Normalisasi Fitur : Algoritma K-Means sangat sensitif terhadap skala data karen perhitungannya didasarkan pada jarak Euclidean. Perbedaan skala data antara fitur dapat menyebabkan fitur dengan rentan nilai yang lebih besar mendominasi perhitungan jarak. Oleh karena ini, digunakan StandardScaler dari library Scikit-learn. StandardScaler mentransformasikan data sehingga memiliki rata-rata nol (*mu = 0)* dan

variansi satu (*sigma = 1)* . Rumus Transformasinya adalah *z = (x - mu) / sigma.* Normalisasi ini memastikan bahwa semua atribut memiliki kontribusi yang setara dalam perhitungan jarak data *point.*

Berikut adalah tampilan beberapa baris pertama data setelah proses normalisasi:

```
import pandas as pd
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
```

Gambar 2. Import Library

Gambar 2. di atas ini adalah library Python yang di import dan tujuannya:

- Pandas (import pandas as pd) : Digunakan untuk manipulasi dan analisis data dalam bentuk Data Frame dan Series, memudahkan pengelolaan data.
- Matplotlib (import maplotlib.pyplot as plt) : Digunakan untuk visualisasi grafik, seperti diagram gratis, batang dan lainnya.
- K-Means (from sklearn.cluster import KMeans) : Digunakan untuk Clustering atau pengelompokan data ke dalam beberapa grup berdasarkan kemiripan.
- StandardScaler (from sklearn.preprocessing import StandardScaler) : Digunakan untuk menstandarisasi data agar setiap fitur memiliki skala yang sama.

```
url = 'https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.csv'
data_tabel = ['Preganancies','Glucose','BloodyPressure','SkinThickness','Insulin','BMI','DiabetesPedigreeFunction','Age','Outcome']
df = pd.read_csv(url, names=data_tabel)

print(df.head())
```

Gambar 3. Membuat dan Menampilkan Data

Bagian gambar 3. menjelaskan bagaimana data dimuat dan dilihat:

- URL : Menyimpan URL yang mengarah ke sumber data CSV. Data tersebut berasal dari GitHub dan berisi informasi tentang diabetes pada individu india.
- Data Tabel : Variabel ini adalah daftar yang berisi nama-nama kolom yang akan digunakan untuk memberi label pada data yang diimpor.
- Df : Variabel menyimpan Data Frame yang berisi data dari file CSV yang diimpor.
- Print : Digunakan untuk menampilkan liba baris pertama dari DataFrame df yang telah dibuat.

```
# Memisahkan fitur untuk clustering (menghilangkan kolom 'Outcame')
x = df[['Preganancies','Glucose','BloodyPressure','SkinThickness','Insulin','BMI','DiabetesPedigreeFunction','Age']]
```

Gambar 4. Memisahkan Fitur untuk Clustering

Bagian gambar 4 menjelaskan Memisahkan fitur untuk clustering atau menghilangkan kolom 'Outcome' ini menjelaskan tujuan dari kode yang akan dilakukan, yaitu untuk memisahkan kolom fitur yang akan digunakan dalam analisis clustering dan menyatakan bahwa kolom 'Outcome' tidak disertakan.

```
# Normalisasi data menggunakan StandardScaler
scaler = StandardScaler()
x_normalisasi = scaler.fit_transform(x)
```

Gambar 5. Normalisasi Data

Bagian gambar 5 menjelaskan bagaimana Kode ini digunakan untuk melakukan normalisasi data dengan menggunakan objek Objek StandardScaler dari pustaka Scikit-learn dalam Python.

Penerapan Algoritma K-Means Clustering

Setelah data di pra-proses, algoritma K-Means diterapkan.

1. Penentuan Jumlah Klaster (*k*) : Dalam penelitian ini, jumlah klaster (*k*) secara eksplisit ditentukan sebagai 3. Pemilihan *k = 3* didasarkan pada asumsi bahwa risiko diabetes dapat dikategorikan menjadi lebih komprehensif, penentuan nilai optimal seringkali melibatkan metode seperti Elbow Method atau Si;houtte Analysis untuk menentukan yang paling baik merepresentasikan struktur intrinsik data dan menghasilkan *clustering* yang optimal.
2. Implementasi K-Means : Algoritma K-Means diinisialisasi menggunakan kelas *KMeans* dari library Scikit-learn. Penerapan *n_clustering=3* ditetapkan sesuai dengan jumlah klaster yang diinginkan. *random_state=42* digunakan untuk memastikan hasil yang *reproducible* (konsisten setiap kali kode dijalankan). Parameter *n_init=10* kali dengan *centroid* awal yang berbeda secara acak, dan hasil terbaik(berdasarkan inersia, yaitu jumlah kuadrat jarak titik ke centroid terdekatnya) akan dipilih. Model kemudian dilatih menggunakan data yang telah dinormalisasi (*x_normalisasi)*.
3. Penerapan Label Klaster ; Setelah pelatihan, label klaster yang dihasilkan oleh K-Means (k means.labels) ditambahkan sebagai kolom baru ('Cluster') ke DataFrame asli (df) untuk memudahkan analisis lebih lanjut.

```
# Inisialisasi dan Penerapan K-Means
k = 3
kmeans = KMeans(n_clusters=k,random_state=42, n_init=10)
kmeans.fit(x_normalisasi)
```

Gambar 6. Parameter Inisialisasi K-Means

Pada gambar 6 ini adalah detail parameter yang digunakan saat menginisialisasi K-Means:

- K=3 : Menentukan jumlah kluster yang diinginkan. Dalam kode ini sebanyak 3 kluster.
- K-Means (n_clusters=k, random_state=42, n_init=10) : Menginisialisasi objek K-Means dengan menentukan parameter n_cluster=k, random_state=42 dan n_init=10.
- N_cluster=k : Menentukan jumlah kluster
- Random_state=42 : mengatur keadaan acak untuk reproducibility, sehingga hasil yang sama diperoleh pada setiap eksekusi kode.
- N-init=10 : Menentukan jumlah percobaan yang dilakukan untuk inisialisasi centroid. Dalam metode ini k-means akan melakukan 10 inisialisasi.

```
# Menambahkan label klaster ke DataFrame asli
df['Cluster'] = kmeans.labels_

print(df[['Preganancies','Glucose','BloodyPressure','SkinThickness','Insulin','BMI','DiabetesPedigreeFunction','Age','Cluster']].head())
```

Gambar 7. Penambahan label klaster ke DataFrame

Bagian gambar 7 menjelaskan bagaimana kode ini berfungsi untuk menambahkan label klaster dari hasil algoritma K-Means ke dalam DataFrame asli (df). Kode ini digunakan untuk menambahkan informasi mengenai klaster ke dalam Data Frame dan menampilkan beberapa kolom yang relevan untuk analisis lebih lanjut.

Visualisasi Klaster

Untuk mendapatkan pemahaman visual tentang bagaimana data point dikelompokkan, visualisasi klaster sangat penting. Karena dataset memiliki dimensi tinggi (8 fitur) visualisasi langsung dalam 8 dimensi tidak mungkin dilakukan. Oleh karena itu, dipilih dua fitur yang secara klinis relevan, yaitu Glucose and Blood Pressure, untuk memvisualisasikan bagaimana klaster-klaster tersebut terpisah pada bidang dua dimensi. Plot Scatter digunakan untuk menampilkan setiap data point dengan warna yang berbeda sesuai dengan klaster tempatnya berada.

```
# Visualisasi klaster berdasarkan Glukosa dan Tekanan Darah
plt.figure(figsize=(8,6))
plt.scatter(df['Glucose'],df['BloodyPressure'],
c=df['Cluster'],cmap='viridis')
plt.xlabel('Glucose')
plt.ylabel('BloodPressure')
plt.title('Pengelompokan Glukosa dan Tekanan Darah dengan metode (K-means)')
plt.colorbar(label='Cluster')
plt.grid(True)
plt.show()
```

Gambar 8. Visualisasi klaster Glukosa dan Tekanan Darah

Bagian gambar 8 menjelaskan bagaimana Kode ini membuat scatter plot untuk menampilkan hasil pengelompokan data Glukosa dan Tekanan Darah menggunakan metode K-Means. Berikut penjelasan nya:

- Plt.figure(figsize=(8,6)) : Mengatur ukuran figure untuk plot menjadi 8x6 inci.
- Plt.scatter(df[‘Glucose’], df[‘BloodPressure’], c=df[‘Cluster’],cmap=’viridis’) : Membuat scatter plot dengan Sumbu X dan Sumbu Y.
- Plt.xlabel(‘Glucose’) : Menambahkan label untuk sumbu X dengan teks “Glucose”.
- Plt.ylabel(‘BloodPressure’) : Menambahkan label untuk sumbu Y dengan teks “BloodPressure”.
- Plt.title : Memberikan judul pada plot dengan teks yang menjelaskan pengelompokan berdasarkan glukosa dan tekanan darah menggunakan metode K-Means.
- Ptl.colorbar(label=’Clustering’) : Menambahkan color bar ke plot yang menunjukan informasi tentang klaster yang diwakili oleh warna.
- Plt.grid(True) : Menambahkan grid ke plot untuk memudahkan pembaca.
- Plt.show() : Menampilkan plot yang telah dibuat.

```
centroid = scaler.inverse_transform(kmeans.cluster_centers_)
centroid_df = pd.DataFrame(centroid,columns=['Preganancies','Glucose','BloodyPressure','SkinThickness','Insulin','BMI','DiabetesPedigreeFunction','Age'])

print("Centroid setiap Cluster")
print(centroid_df)
```

Gambar 9. Analisis Centroid klaster

Bagian gambar 9 ini berfokus pada analisis centroid dari setiap klaster yang terbentuk:

- Inverse Transformasi Centroid : Mengubah centroid klaster yang didapat di model K-Means ke skala aslinya.
- Pembuatan DataFrame : Membuat Data Frame dari centroid yang telah ditransformasikan untuk memudahkan pengelolaan dan analisis data.
- Print(“Centroid setiap data”), Print(centroid_df) :Mencetak informasi tentang centroid untuk setiap klaster agar pengguna dapat melihat rata-rata fitur setiap klaster.

Analisis Centroid Cluster

Pusat klaster (centroid) adalah titik rata-rata dari semua data point dalam sebuah klaster. Centroid merepresentasikan karakteristik rata-rata atau tipikal dari setiap kelompok pasien. Untuk interpretasi yang bermakna, centroid yang dihasilkan oleh model K-Means (yang berada transform). Analisis nilai centroid untuk setiap fitur akan membantu dalam memberikan label interpretatif pada setiap klaster (misalnya, risiko rendah, sedang dan tinggi).

Contoh Perhitungan Manual Pasien ke Centroid

Setelah proses pelatihan algoritma K-Means selesai, kita mendapatkan sejumlah titik pusat atau yang disebut centroid. Setiap centroid ini mewakili satu kelompok atau klaster, dan memiliki nilai rata-rata dari fitur-fitur dalam klaster tersebut. Untuk menentukan setiap pasien

masuk ke klaster yang mana, dilakukan perhitungan jarak antara data pasien dan setiap centroid menggunakan rumus Euclidean Distance, yaitu:

$$D(X,C) = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2 + \cdots + (x_n - c_n)^2}$$

Di mana:

- X adalah vektor fitur pasien (misalnya Glucose, BMI, Age, dst)
- C adalah vektor fitur dari centroid klaster tertentu
- n adalah jumlah fitur yang digunakan

Sebagai ilustrasi, misalkan data pasien A memiliki nilai Glucose 148, BMI 33.6, dan Age 50. Sementara itu, centroid dari masing-masing klaster (hasil inverse dari normalisasi) adalah sebagai berikut:

Tabel 2

Klaster	Glucose	BMI	Age
0	99.19	28.29	28.75
1	137.99	35.85	35.15
2	106.26	30.41	51.48

Perhitungan jarak dari pasien A ke masing-masing centroid dilakukan sebagai berikut:

- Jarak ke Cluster 0:

$$D = \sqrt{(148 - 99.19)^2 + (33.6 - 28.29)^2 + (50 - 28.75)}$$
$$= \sqrt{2381.5 + 28.2 + 451.6} \approx 53.47$$

- Jarak ke Cluster 1:

$$D = \sqrt{(148 - 137.99)^2 + (33.6 - 35.85)^2 + (50 - 35.15)^2}$$
$$= \sqrt{100.2 + 5.06 + 220.5} \approx 18.05$$

- Jarak ke Cluster 2:

$$D = \sqrt{(148 - 106.26)^2 + (33.6 - 30.41)^2 + (50 - 51.48)^2} \approx 41.90$$
$$= \sqrt{1743.1 + 10.2 + 2.2} \approx 41.9$$

Hasil menunjukkan bahwa pasien A memiliki jarak terpendek ke klaster 1, sehingga oleh algoritma K-Means pasien ini akan dikategorikan sebagai anggota dari klaster 1, yang dalam interpretasi hasil mengindikasikan risiko diabetes yang tinggi.

Langkah serupa dilakukan terhadap seluruh data pasien dalam dataset untuk menentukan klaster masing-masing berdasarkan kedekatannya terhadap centroid klaster. Seluruh proses ini dilakukan secara otomatis oleh algoritma K-Means setelah training selesai, tetapi proses manual seperti ini penting untuk memahami mekanisme di balik penentuan klaster.

4. HASIL DAN PEMBAHASAN

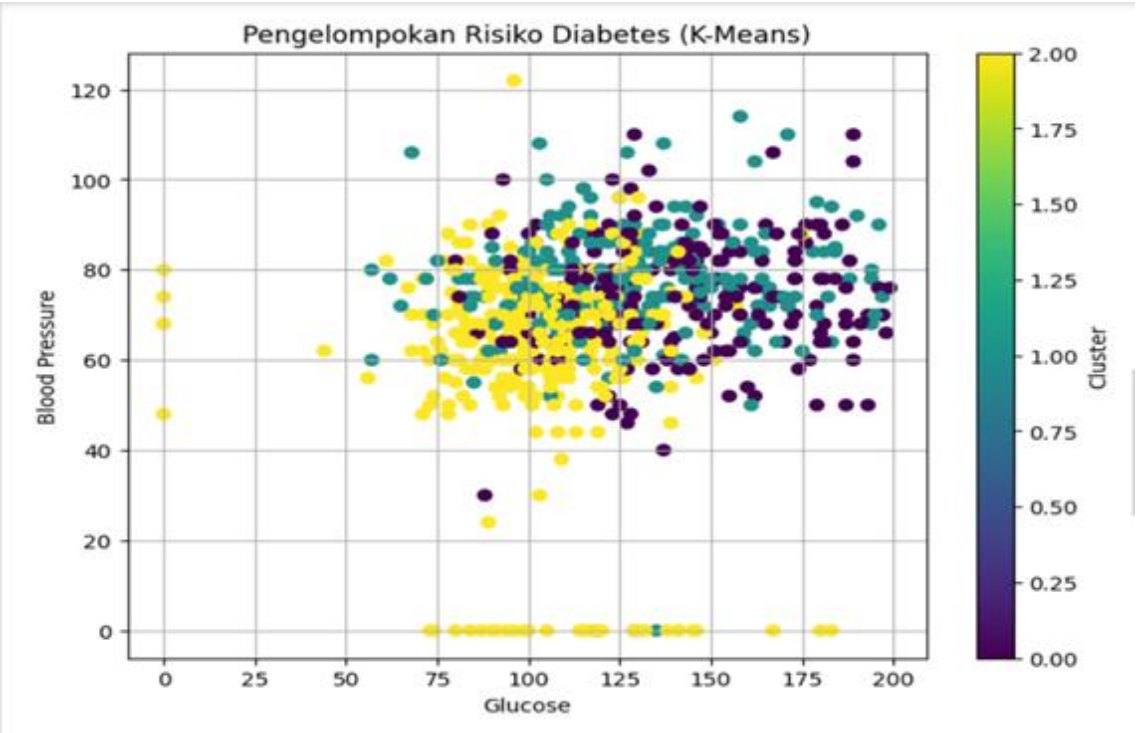
Distribusi Klaster dan Visualisasi

Setelah proses pra-pemrosesan dan penerapan algoritma K-Means dengan nilai , dataset berhasil dikelompokkan ke dalam ketiga klaster. Berikut adalah contoh lima baris pertama Data Frame dengan label klaster yang telah ditambahkan :

Tabel 3. Data Frame dengan label

Pregnancies	Glucose	Bloody Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Cluster
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.1	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

Visualisasi klaster dilakukan dengan memproyeksikan data ke dalam dua dimensi menggunakan fitur Glucose and Blood Pressure. Gambar hasil menunjukan hasil pengelompokan secara visual.



Gambar 10. Pengelompokan Pasien Berdasarkan Glukosa dan Tekanan Darah dengan K-Means.

Dari gambar 10, kita dapat mengamati bagaimana data point terdistribusi ke dalam klaster-klaster yang berbeda. Meskipun ada beberapa tumpang tindih, hal ini wajar mengingat visualisasi ini adalah proyeksi 2D dari data yang asli berdimensi 8. Klaster cenderung membentuk kelompok yang relatif terpisah berdasarkan kombinasi nilai glukosa dan tekanan darah. Misalnya, satu klaster mungkin terkonsentrasi pada area dengan nilai glukosa dan tekanan darah rendah, sementara klaster ini lain berbeda pada area dengan nilai yang lebih tinggi.

Analisis Karakteristik Klaster (Centroid)

Untuk memahami karakteristik internal setiap klaster, analisis centroid adalah kunci. Tabel 4 menampilkan nilai centroid setiap klaster untuk semua atribut, yang telah diubah kembali ke skala asli data.

Tabel 4. Centroid Setiap klaster

Pregnancies	Glucose	Bloody Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age
1.196825	99.193651	72.247619	21.149206	62.139683	28.293651	0.347940	28.755556
3.656757	137.986486	72.637838	2.832432	175.756757	35.856757	0.540130	35.151351
0.887500	106.262500	70.612500	20.162500	32.225000	30.407500	0.573937	51.487500

Berdasarkan analisis nilai centroid pada Tabel 4, setiap klaster dapat diinterpretasikan sebagai berikut:

1. Klaster 0 (Potensi Risiko Rendah):
    - Klaster ini memiliki rata-rata jumlah kehamilan (1. 2), usia (28.7 tahun), kadar Glukosa (99.2), dan Insulin (62.1) yang paling rendah di antara klaster lainnya.
    - Nilai BMI (28.3) dan Tekanan Darah (62.2) juga cenderung lebih rendah.
    - Nilai *Diabetes Pedigree Function* (0.34) adalah yang terendah, mengindikasikan riwayat keluarga diabetes yang minimal.
    - Berdasarkan karakteristik ini , klaster 0 cenderung mewakili pasien dengan risiko diabetes yang relatif rendah.
    - jumlah anggota di klaster ini adalah 35. 1 pasien.
2. Klaster 1 (potensi risiko tinggi):
    - Klaster ini menunjukkan rata-rata kadar Glukosa (137.9) dan insulin (175.7) yang paling tinggi secara signifikan.
    - Nilai BMI (35.8) dan Diabetes Pedigree Function (0.54) juga merupakan yang tertinggi.
    - Memiliki rata-rata jumlah kehamilan (3.6) dan ketebalan kulit (29.8) yang paling tinggi.
    - Usia Rata-rata (35.1 tahun) berada di tengah.
    - Kombinasi karakteristik ini , terutama glukosa dan insulin yang tinggi serta BMI yang tinggi, sangat mengindikasikan kelompok pasien dengan risiko diabetes yang tinggi.
    - jumlah anggota di klaster ini adalah 280 pasien.
3. Klaster 2 (Potensi Risiko Sedang/ Usia Lanjut):
    - Klaster ini secara mencolok memiliki usia rata-rata tertinggi (51.4 tahun)
    - Meskipun kadar glukosa (106.2) dan BMI (30.4) berada diantara klaster 0 dan 1, namun cenderung lebih tinggi dari klaster 0.
    - Tekanan darah (70.6) juga lebih dari klaster 0.
    - Menariknya, jumlah kehamilan (0.8) dan insulin (32.2) adalah yang paling rendah, bahkan lebih rendah dari klaster 0.
    - Namun, Diabetes Pedigree Function (0.57) adalah yang tertinggi dari semua klaster, menunjukkan riwayat keluarga yang kuat meskipun kadar glukosa dan insulin tidak ekstrim seperti klaster 1.
    - Klaster ini mungkin mewakili kelompok pasien yang lebih tua, yang mungkin memiliki risiko diabetes karena faktor usia dan riwayat keluarga, meskipun profil metaboliknya (glukosa,insulin) tidak seburuk klaster 1.
    - jumlah anggota di klaster ini adalah 173 pasien.

Keterbatasan dan Pembahasan Mendalam

Penelitian ini menunjukan potensi K-Means Clustering dalam mengidentifikasi kelompok risiko diabetes. Namun, terdapat beberapa keterbatasan dalam implementasi yang perlu dibahas:

1. Penentuan Jumlah Klaster (k) : Pemilihan $k = 3$ dilakukan secara *hardcode* berdasarkan asumsi awal kategori risiko (rendah, sedang, tinggi) Tanpa penggunaan metode seperti Elbow Method atau Silhouette Analysis, kita tidak dapat secara objektif memastikan

bahwa 3 adalah jumlah klaster yang paling optimal atau yang paling akurat mempresentasikan struktur klaster alami data. pemilihan yang tidak tepat dapat menghasilkan *clustering* yang suboptimal dan interpretasi yang akurat.

2. Penanganan Nilai Nol : Dataset Pima Indian memiliki nilai '0' pada beberapa fitur seperti Glucose, Blood Pressure, Insulin, dan BMI yang secara medis tidak mungkin nol dan seharusnya diinterpretasikan sebagai nilai hilang. Kode yang digunakan tidak melakukan penanganan khusus untuk nilai-nilai nol ini. Memperlakukan '0' sebagai nilai valid dapat mendistorsi perhitungan centroid dan jarak, sehingga mempengaruhi kualitas dan interpretasi clustering. sebagai contoh nilai '0' sangat berbeda dengan nilai insulin '175.7' dan seharusnya tidak dilakukan sama dengan perhitungan rata-rata klaster..
3. Evaluasi Kuantitatif Clustering : Laporan awal menyebut "Silhouette Score sebesar 0.61". Namun, kode yang diberikan tidak menghitung atau menampilkan Silhouette Score. Untuk laporan ilmiah ini, perhitungan menarik evaluasi clustering secara kuantitatif sangat esensial untuk mengukur kohesivitas dan separasi. Tanpa metrik ini, klaim kualitas clustering bersifat subjektif. Sebagai contoh, penelitian Gestavito et al (2024) menggunakan Silhouette Coefficient dan Davies Boulding Index sebagai metrik evaluasi, di mana mereka memperbolehkan Silhouette Coefficient 0.5716 dan Davies Bouldin Index 0.672 untuk $k = 2$. Tanpa metrik ini, klaim kualitas Clustering bersifat objektif
4. Visualisasi Data Dimensi Tinggi : Visualisasi hanya menggunakan dua fitur yaitu Glucose and Blood Pressure mungkin tidak sepenuhnya mewakili bagaimana klaster-klaster terpisah dalam ruang 8 dimensi. Proyeksi 2D dapat menyembunyikan pemisahan yang jelas di dimensi lain atau menunjukkan tumpang tindih palsu. Penggunaan teknik reduksi dimensi seperti PCA (Principal Component Analysis) atau t-SNE (t-Distributed Stochastic Neighbor Embedding) sebelum visualisasi akan memberikan gambaran yang lebih akurat tentang pemisahan klaster dalam dimensi yang lebih rendah, yang mempertahankan variasi atau struktur lokal data. Meskipun demikian, hasil clustering dan interpretasi centroid memberikan wawasan awal yang berharga mengenai karakteristik kelompok pasien berdasarkan data medis mereka.

5. KESIMPULAN DAN SARAN

Penggunaan algoritma K-Means Clustering pada dataset Pima Indian Diabetes berhasil mengelompokkan pasien ke dalam beberapa kelompok berdasarkan karakteristik risiko diabetes yang berbeda. Dari analisis pusat klaster (centroid), kami mengidentifikasi tiga kelompok utama yang secara umum dapat dibedakan sebagai pasien dengan risiko diabetes rendah, tinggi, dan sedang/usia lanjut, berdasarkan fitur medis seperti kadar glukosa, BMI, usia, dan riwayat keluarga pasien. Metode pembelajaran tanpa pengawasan ini menunjukkan potensi sebagai dasar dalam mengembangkan sistem pendukung keputusan medis untuk mendeteksi individu yang berisiko lebih awal. Namun, keberhasilan dan interpretasi hasil pengelompokan sangat bergantung pada kualitas data, pemilihan jumlah klaster yang tepat, serta penanganan data yang valid.

Berdasarkan temuan dan keterbatasan yang teridentifikasikan dalam penelitian ini, beberapa saran untuk pengembangan dan penelitian lebih lanjut meliputi:

1. Pra-pemrosesan Data yang Lebih Teliti : Lakukan pra-pemrosesan data yang lebih detail, termasuk mengenali dan menangani nilai nol atau nilai data hilang yang tidak masuk akal secara medis (seperti pada glukosa, tekanan darah, BMI, insulin), serta mengelola data pencilan yang dapat mempengaruhi performa algoritma K-Means.
2. Validasi jumlah cluster Optimal(k): Implementasikan menggunakan metode evaluasi clustering seperti Elbow Method atau Silhouette Analysis untuk menentukan jumlah klaster (k) yang paling sesuai dengan data, bukan hanya berdasarkan asumsi semata.
3. Evaluasi Kualitas Klaster Secara Kuantitatif: Sertakan perhitungan dan presentasi Silhouette Score (metrik clustering lainnya seperti Davies-Bouldin Index atau Calinski-Harabasz Index) dalam laporan untuk memberikan bukti kuantitatif mengenai kualitas pengelompokan, baik dari sisi kohesivitas maupun pemisahan antar klaster.
4. Visualisasi dengan Teknik Reduksi Dimensi: Gunakan teknik reduksi dimensi seperti PCA (Principal Component Analysis) atau t-SNE (t--Distributed Stochastic Neighbor Embedding) sebelum visualisasi, agar pemisahan klaster dalam ruang berdimensi tinggi dapat terlihat lebih jelas dan struktur data yang kompleks dapat terungkap.
5. Perbandingan dengan algoritma Clustering Lain: Lakukan perbandingan hasil clustering dengan algoritma lain seperti DBSCAN atau Hierarchical Clustering untuk mengetahui apakah terdapat pola pengelompokan yang berbeda atau lebih baik.
6. Integrasi dengan model Prediksi Supervised: Hasil clustering dapat digunakan sebagai label untuk melatih model supervised learning seperti Decision Tree, Random Forest,

atau support Vector Machine, guna membangun sistem prediksi risiko diabetes yang lebih akurat.

7. Validasi oleh Tenaga Medis dan Data Lokal: Validasi lebih lanjut oleh tenaga medis profesional atau menggunakan data dari rumah sakit lokal sangat penting untuk memastikan relevansi dan keakuratan hasil pengelompokan risiko dalam konteks populasi yang berbeda.

DAFTAR REFERENSI

Agustin, A. V., & Voutama, A. (2023). Implementasi data mining klasifikasi penyakit diabetes pada perempuan menggunakan Naïve Bayes. JATI (Jurnal Mahasiswa Teknik Informatika), 7(2), 1002–1007.

AldiYatna, K., Rahaningsih, N., & Dana, R. D. (2024). Penerapan data mining untuk clustering penyakit diare menggunakan algoritma K-Means (Studi kasus: Puskesmas Beber). JATI (Jurnal Mahasiswa Teknik Informatika), 8(3).

Basyir, M. K. (2025). Klastering penyakit diabetes dengan metode K-Means. Buletin Ilmiah Ilmu Komputer dan Multimedia (BIIKMA), 2(5), 904–909.

Gestavito, R., Hadiana, A. I., & Umbara, F. R. (2024). Pengelompokan tingkat risiko penyakit diabetes melitus menggunakan algoritma K-Means Clustering. JUMANJI, 8(1), 16–35.

Haryadi, D., & Atmaja, D. M. U. (2021). Penerapan algoritma K-Means clustering untuk pengelompokan tingkat risiko penyakit jantung. Journal of Informatics and Communications Technology (JICT), 3(2), 51–66.

International Diabetes Federation. (2021). IDF Diabetes Atlas (10th ed.).

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. Pattern Recognition Letters, 31(8), 651–666.

Karyadiputra, E., & Setiawan, A. (2022). Penerapan data mining untuk prediksi awal kemungkinan terindikasi diabetes. Teknosains: Media Informasi Sains dan Teknologi, 16(2), 221–232.

Khalish, F., Piranti, N. M., & Martadireja, O. (2025). Implementasi data mining menggunakan teknik clustering dengan metode K-Means. JIIP (Jurnal Ilmiah Ilmu Pendidikan), 8(5), 5392–5397.

Krishna, V. B., Jaya, R. K., Bhuvaneswari, A. P., Gururaj, H. L., Ravi, V., Almeshari, M., & Alzamil, Y. (n.d.). A novel application of K-means cluster prediction model for diabetes early identification using dimensionality reduction techniques.

Marisa, F. (n.d.). EDUCATIONAL DATA MINING (Konsep dan penerapan). Jurnal Teknologi Informasi, 4(2), 90–97.

Masruriyah, A. F. N., Mardiah, M. D. A., & Malik, K. N. (2025). Pendekatan unsupervised learning dalam segmentasi kesehatan: Perbandingan K-Means dan DBSCAN. MIND (Multimedia Artificial Intelligent Networking Database) Journal, 10(1), 99–113.

Prasatya, A., Siregar, R. R. A., & Arianto, R. (2020). Penerapan metode K-Means dan C4.5 untuk prediksi penderita diabetes. PETIR: Jurnal Pengkajian dan Penerapan Teknik Informatika, 13(1), 1–9.

Putra, M. F. (2024). Implementasi algoritma K-Means clustering dan support vector machine untuk mengetahui faktor pemicu diabetes melitus di Puskesmas Lappae [Skripsi, Universitas Muhammadiyah Makassar].

Setiawan, E., & Prasetijo, A. B. (2020). Analysis K-Means clustering dalam klasterisasi pasien diabetes. Jurnal Teknologi Informasi dan Komputer, 7(2), 89–95.

Setyadji, A. E. S., Wibowo, A. P., Matthew D., I. G. N. A., Pratama, R. B., Masyhuda, T. A., Sinaga, Y. A. A., Purwanti, E., & Werdiningsih, I. (2023). Analisis klaster data pasien diabetes untuk identifikasi pola dan karakteristik pasien. Jurnal Teknologi dan Sistem Informasi Bisnis, 5(3), 172–182.

Subarja, R. E., & Hendrik, B. (2023). Evaluasi performa deteksi penyakit diabetes dengan Fuzzy C-Means dan K-Means clustering. Jurnal Elektronika dan Teknik Informatika Terapan (JENTIK), 1(3), 100–108.

UCI Machine Learning Repository. (n.d.). Pima Indians Diabetes Database. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database