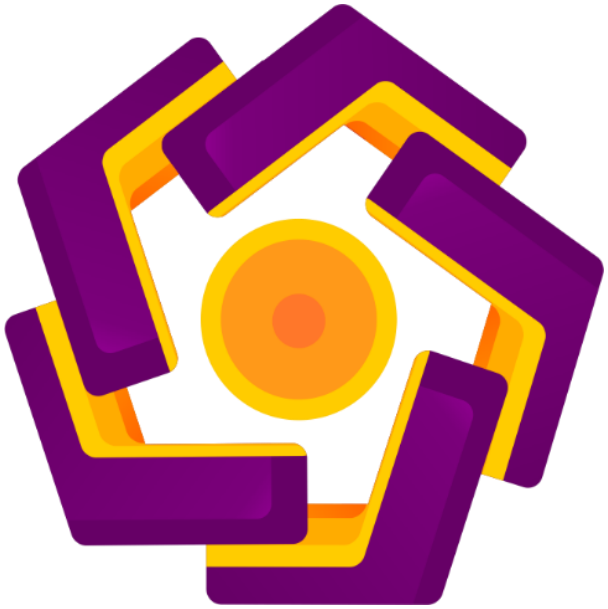
ANALISIS PREDIKSI PENYAKIT DIABETES DENGAN METODE HIERARCHICAL CLUSTERING

Dosen Pengampu :
Kusnawi, S.Kom, M.Eng.



Disusun Oleh :

Sandi Hernanto
21.11.4530

S1 INFORMATIKA
UNIVERSITAS AMIKOM YOGYAKARTA
2024

Abstrak

Penyakit diabetes merupakan salah satu masalah kesehatan global yang memerlukan pendekatan yang cermat dalam prediksi dan pencegahannya. Studi ini bertujuan untuk menganalisis prediksi penyakit diabetes menggunakan metode Hierarchical Clustering berdasarkan dataset yang mencakup informasi seperti kehamilan, glukosa darah, tekanan darah, ketebalan kulit, insulin, indeks massa tubuh (BMI), fungsi silsilah diabetes, usia, dan hasil (Outcome). Hierarchical Clustering digunakan untuk mengelompokkan data menjadi klaster-klaster yang memiliki kemiripan tertentu, sehingga dapat membantu identifikasi pola atau tren terkait penyakit diabetes. Pada tahap pertama, dataset dipersiapkan dan dieksplorasi untuk memahami distribusi variabel dan hubungan antar variabel. Kemudian, metode Hierarchical Clustering diterapkan untuk mengelompokkan data menjadi hierarki klaster yang merepresentasikan tingkat kemiripan antar observasi. Analisis ini bertujuan untuk mengidentifikasi apakah terdapat pola-pola khusus dalam data yang dapat memberikan wawasan tambahan terkait penyakit diabetes. Hasil dari analisis ini diharapkan dapat memberikan pemahaman lebih lanjut tentang bagaimana variabel-variabel tertentu dapat berkorelasi atau membentuk pola yang dapat mempengaruhi risiko terjadinya diabetes. Selain itu, pemetaan hierarki klaster dapat memfasilitasi identifikasi kelompok risiko yang mungkin memiliki karakteristik serupa, memungkinkan pengembangan strategi pencegahan yang lebih terarah.

***Kata Kunci**: Diabetes, Prediksi Penyakit, Metode Hierarchical Clustering, Pengelompokan.*

Abstract

Diabetes is a global health issue that requires careful approaches for prediction and prevention. This study aims to analyze the prediction of diabetes using the Hierarchical Clustering method based on a dataset that includes information such as pregnancies, blood glucose, blood pressure, skin thickness, insulin, body mass index (BMI), diabetes pedigree function, age, and outcome. Hierarchical Clustering is employed to cluster the data into groups that exhibit certain similarities, aiding in the identification of patterns or trends related to diabetes. In the first stage, the dataset is prepared and explored to understand the distribution of variables and relationships among them. Subsequently, the Hierarchical Clustering method is applied to group the data into a hierarchical structure of clusters that represents the level of similarity among observations. This analysis aims to identify whether there are specific patterns in the data that can provide additional insights into diabetes. The results of this analysis are expected to provide a deeper understanding of how certain variables may correlate or form patterns that can influence the risk of diabetes. Additionally, the hierarchical cluster mapping can facilitate the identification of high-risk groups that may share similar characteristics, enabling the development of more targeted prevention strategies.

***Keywords:** Diabetes, Disease Prediction, Hierarchical Clustering Method, Grouping*

BAB I

Pendahuluan

A. Latar Belakang

Penyakit diabetes telah menjadi tantangan serius dalam dunia kesehatan global dengan prevalensi yang terus meningkat secara signifikan. Dalam menghadapi dampak kesehatan masyarakat yang luas, upaya untuk melakukan prediksi dini terhadap penyakit ini menjadi sangat penting guna memungkinkan intervensi yang tepat waktu dan efektif. Salah satu metode analisis yang muncul sebagai alat potensial untuk mengidentifikasi pola-pola kompleks dalam data kesehatan, khususnya pada dataset diabetes yang mencakup berbagai variabel, adalah Metode Hierarchical Clustering.

Pentingnya prediksi penyakit diabetes bukan hanya terkait dengan diagnosis dini, tetapi juga karena konsekuensi serius yang mungkin timbul, seperti risiko komplikasi seperti penyakit jantung, gangguan ginjal, dan masalah mata. Meningkatkan kualitas hidup penderita diabetes memerlukan strategi pencegahan yang efektif, yang dapat diperoleh melalui analisis yang cermat terhadap faktor-faktor risiko potensial. Metode Hierarchical Clustering, dengan kemampuannya mengelompokkan data menjadi struktur hierarkis berdasarkan tingkat kemiripan, dapat memberikan wawasan tambahan terkait hubungan antar variabel dan potensi pola tersembunyi dalam dataset diabetes.

Penerapan Metode Hierarchical Clustering pada dataset diabetes tidak hanya membuka peluang untuk mengidentifikasi kelompok risiko tertentu, tetapi juga dapat membantu dalam mendefinisikan profil pasien tertentu yang memiliki kecenderungan lebih tinggi terkena penyakit. Dengan memahami lebih baik bagaimana variabel-variabel seperti kehamilan, glukosa darah, dan BMI saling berinteraksi dalam konteks prediksi diabetes, penyedia layanan kesehatan dan peneliti dapat mengembangkan model prediksi yang lebih akurat dan responsif terhadap kebutuhan masyarakat. Informasi ini menjadi dasar penting untuk pengembangan strategi pencegahan yang lebih terarah dan personalisasi, diharapkan dapat mengoptimalkan sumber daya kesehatan dan meningkatkan efisiensi dalam manajemen penyakit diabetes.

B. Tujuan

- **Eksplorasi Prevalensi Diabetes:** Mengidentifikasi dan mengeksplorasi prevalensi penyakit diabetes dalam dataset, memberikan dasar pemahaman mengenai tingkat keparahan masalah kesehatan ini.
- **Penerapan Metode Hierarchical Clustering:** Mengaplikasikan metode Hierarchical Clustering untuk mengelompokkan data pasien diabetes, dengan tujuan memahami pola dan hubungan antar variabel yang mungkin tersembunyi.
- **Identifikasi Pola dan Kelompok Risiko:** Mengidentifikasi pola prediksi penyakit diabetes dan mengelompokkan pasien ke dalam kategori risiko tertentu, membuka potensi untuk intervensi lebih spesifik dan pencegahan yang terfokus.

- **Analisis Faktor Risiko Potensial:** Menganalisis faktor-faktor risiko potensial, seperti kehamilan, glukosa darah, dan BMI, untuk memahami kontribusi masing-masing faktor terhadap prediksi diabetes.
- **Optimasi Model Prediksi:** Merumuskan temuan analisis untuk memperbaiki dan mengoptimalkan model prediksi menggunakan metode Hierarchical Clustering, sehingga dapat memberikan hasil yang lebih akurat dan andal.
- **Pengembangan Strategi Pencegahan yang Terarah:** Berkontribusi pada pengembangan strategi pencegahan yang lebih terarah dengan menyesuaikannya dengan karakteristik kelompok risiko yang diidentifikasi.
- **Peningkatan Pemahaman Terhadap Diabetes:** Memberikan pemahaman yang lebih dalam tentang hubungan antar variabel dalam konteks diabetes, memperkaya pengetahuan ilmiah terkait penyakit ini..

C. Metode Penyelesaian

Metode penyelesaian yang digunakan dalam konteks prediksi penyakit diabetes adalah Hierarchical Clustering. Hierarchical Clustering merupakan pendekatan klasterisasi yang memanfaatkan struktur hirarki untuk mengelompokkan data berdasarkan tingkat kemiripan antar-observasi. Dalam kasus ini, variabel-variabel seperti kehamilan, glukosa darah, tekanan darah, ketebalan kulit, insulin, indeks massa tubuh (BMI), fungsi silsilah diabetes, dan usia akan menjadi dasar untuk mengukur kemiripan antar-pasien. Pendekatan Hierarchical Clustering akan dimulai dengan menggabungkan pasien yang memiliki kesamaan terkecil dalam profil kesehatannya. Proses penggabungan ini akan terus berlanjut hingga terbentuk kelompok yang lebih besar, membentuk struktur hirarki klaster. Dengan menerapkan metode ini, penelitian ini bertujuan untuk mengungkap pola atau tren yang mungkin tersembunyi dalam data, sehingga dapat memberikan wawasan tambahan terkait prediksi penyakit diabetes. Penelitian ini akan menggunakan dataset kesehatan yang mencakup informasi dari sejumlah individu, dengan tujuan utama memberikan panduan yang lebih akurat dan terarah terkait faktor-faktor risiko penyakit diabetes. Kesimpulan dari analisis ini diharapkan dapat membuka peluang untuk pengembangan strategi pencegahan yang lebih efektif dan personalisasi dalam pengelolaan kesehatan diabetes.

BAB II

Tinjauan Pustaka

Kajian Penelitian dan Dasar Teori

Pada bagian ini akan dijelaskan hasil-hasil penelitian terdahulu dan dasar teori yang bisa dijadikan acuan dalam topik penelitian ini. Penelitian terdahulu dan dasar teori telah dipilih sesuai dengan permasalahan dalam penelitian ini, sehingga diharapkan mampu menjelaskan maupun memberikan referensi bagi penulis dalam menyelesaikan penelitian ini. Berikut dijelaskan beberapa penelitian terdahulu yang telah dipilih.

Dalam rangka mengatasi masalah identifikasi subpopulasi prediabetes yang memiliki signifikansi klinis, penelitian ini ditunjukan untuk mengidentifikasi subpopulasi prediabetes yang secara klinis signifikan menggunakan metode divisive hierarchical clustering. Dengan penelitian tersebut, hasil penelitian yang didapatkan menunjukkan bahwa terdapat dua subpopulasi prediabetes yang secara klinis signifikan, yaitu subpopulasi dengan kadar gula darah puasa tinggi dan subpopulasi dengan kadar insulin puasa tinggi [1].

Masalah kesehatan yang signifikan dihadapi masyarakat adalah diabetes mellitus. Untuk mengatasi tantangan ini, penelitian ini dilakukan dengan tujuan untuk memprediksi diabetes mellitus menggunakan hierarchical clustering. Penelitian menggunakan dataset yang terdiri dari data klinis dari 768 pasien diabetes. Data yang digunakan meliputi data antropometri, biokimia, dan gaya hidup. Metode hierarchical clustering digunakan untuk membagi data menjadi beberapa kelompok berdasarkan kemiripan antara data. Data yang dihasilkan penelitian menunjukkan bahwa metode hierarchical clustering memiliki akurasi yang baik dalam memprediksi diabetes mellitus. Akurasi hierarchical clustering adalah 81,8% [2].

Karena maraknya penyakit diabetes penelitian ini bertujuan untuk meningkatkan akurasi deteksi anomali data diabetes dengan menggabungkan metode hierarchical clustering dan convolutional neural network (CNN). Metode hierarchical clustering digunakan untuk membagi data menjadi dua kelompok, yaitu kelompok normal dan kelompok anomali. Selanjutnya, CNN digunakan untuk mengekstrak fitur dari data kelompok anomali. Hasil penelitian membuktikan bahwa metode gabungan hierarchical clustering dan CNN dapat meningkatkan akurasi deteksi anomali data diabetes sebesar 10% dibandingkan dengan metode hierarchical clustering saja [3].

Prediabetes adalah kondisi di mana kadar gula darah dalam tubuh berada di atas batas normal, tetapi belum cukup tinggi untuk didiagnosis sebagai diabetes mellitus. Prediabetes dapat berkembang menjadi diabetes mellitus tipe 2, yang merupakan penyakit kronis yang dapat menyebabkan berbagai komplikasi serius. Maka penelitian ini bertujuan untuk menganalisis data glukosa dan insulin dari tes toleransi glukosa oral (OGTT) menggunakan metode pengelompokan hierarkis dan partisi. Hasil penelitian menunjukkan bahwa metode pengelompokan hierarkis dan partisi dapat digunakan untuk mengidentifikasi subpopulasi prediabetes yang berbeda berdasarkan respons OGTT mereka. Metode pengelompokan hierarkis digunakan untuk mengidentifikasi subpopulasi prediabetes berdasarkan pola respons OGTT

mereka. Metode pengelompokan partisi digunakan untuk mengidentifikasi subpopulasi prediabetes berdasarkan nilai cutoff tertentu untuk kadar glukosa dan insulin plasma. Dengan penelitian tersebut menunjukkan bahwa metode pengelompokan hierarkis dan partisi dapat mengidentifikasi subpopulasi prediabetes yang berbeda berdasarkan respons OGTT mereka. Subpopulasi prediabetes yang berbeda ini memiliki risiko yang berbeda untuk berkembang menjadi diabetes tipe 2 [4].

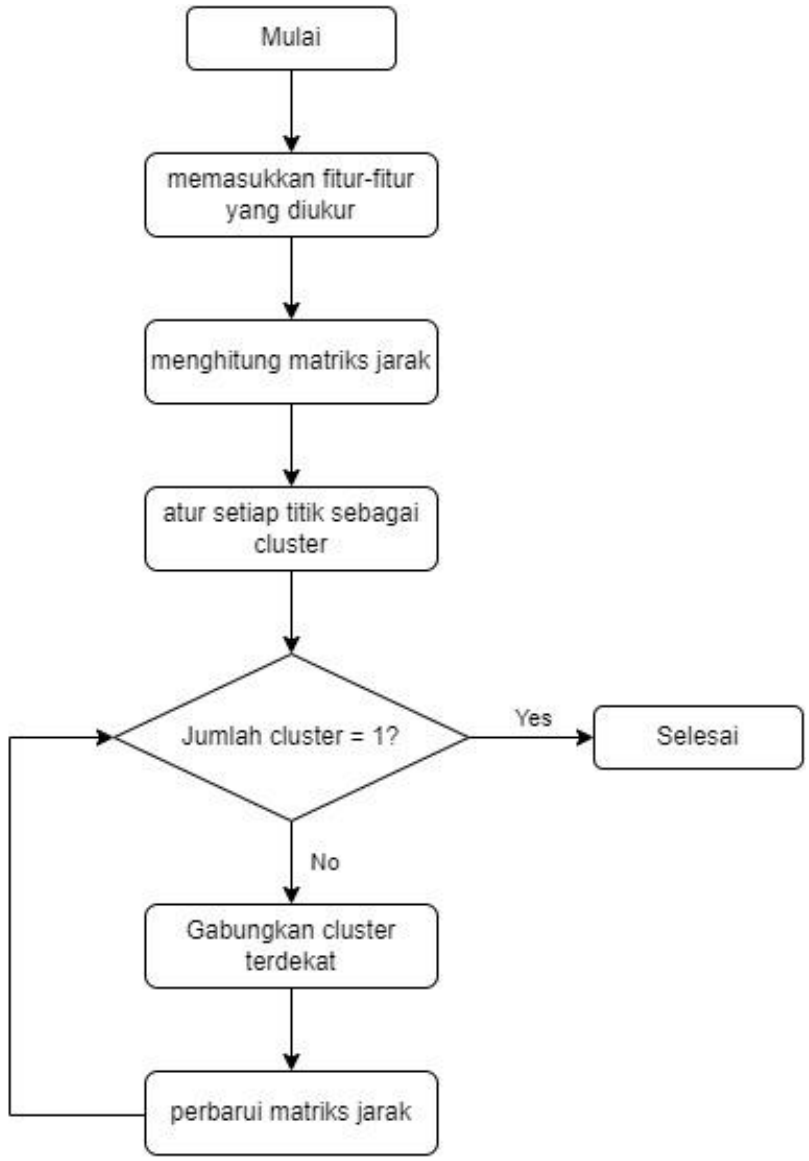
Diabetes mellitus adalah penyakit kronis yang ditandai dengan kadar gula darah yang tinggi. Penyakit ini dapat menyebabkan berbagai komplikasi serius, seperti penyakit jantung, stroke, kebutaan, dan gagal ginjal. Maka dari itu kami meneliti untuk membandingkan kinerja metode data mining untuk memprediksi diabetes mellitus. Metode data mining yang dibandingkan adalah hierarchical clustering, density based clustering, dan simple K-means clustering. Data yang digunakan adalah data Pima Indian Diabetes, yang terdiri dari 768 data pasien. Hasil penelitian menunjukkan bahwa metode hierarchical clustering memiliki kinerja terbaik untuk memprediksi diabetes mellitus. Metode ini memiliki akurasi sebesar 72,2%, diikuti oleh metode density based clustering dengan akurasi sebesar 68,8%, dan metode simple K-means clustering dengan akurasi sebesar 66,7% [5].

Meningkatnya prevalensi diabetes sebagai tantangan serius dalam dunia kesehatan global, maka penelitian ini digunakan untuk meningkatkan akurasi klasifikasi pasien diabetes tipe 2 menggunakan metode hierarchical clustering support vector machines (HCSVM). Metode HCSVM menggabungkan metode pengelompokan data hierarchical clustering dengan metode pembelajaran mesin support vector machines (SVM). Penelitian memberikan hasil dan menunjukkan bahwa HCSVM dapat meningkatkan akurasi klasifikasi pasien diabetes tipe 2 sebesar 10% dibandingkan metode SVM konvensional. Hal ini menunjukkan bahwa metode HCSVM dapat menjadi alternatif yang efektif untuk klasifikasi pasien diabetes tipe 2 [6].

BAB III

Metodologi Research

❖ **Flowchart/Alur Kerja**



Alur kerja flowchart tersebut adalah sebagai berikut:

1. **Mulai**: Mengawali proses clustering.
2. **Memasukkan fitur-fitur yang diukur:** Menginputkan fitur-fitur yang akan digunakan untuk clustering, misalnya data berat badan, tinggi badan, tekanan darah, dan kadar gula darah.
3. **Menghitung matriks jarak:** Menghitung matriks jarak antar titik data, yang menunjukkan seberapa dekat atau jauh titik data tersebut satu sama lain.

4. **Atur setiap titik sebagai cluster**: Mengatur setiap titik data sebagai clusternya sendiri.
5. **Jumlah cluster = 1?:** Memeriksa apakah jumlah cluster sudah mencapai 1. Jika ya, maka proses clustering selesai.
6. **Tidak**: Menyatukan cluster yang terdekat.
7. **Perbarui matriks jarak:** Memperbarui matriks jarak dengan memperhitungkan cluster yang baru saja digabungkan.
8. **Kembali ke langkah 5:** Kembali ke langkah 5 untuk memeriksa apakah jumlah cluster sudah mencapai 1.

❖ **Spesifikasi**

- **Software**
  - Sistem Operasi: Google Colab (cloud-based Jupyter Notebook environment)
  - Bahasa Pemrograman: Python 3.10.12

- **Libraries:**
  - pandas (versi 1.5.3) untuk analisis dan manipulasi data
  - seaborn (versi 0.12.2) untuk visualisasi data
  - matplotlib (versi 3.7.1) untuk visualisasi data
  - scikit-learn (1.2.2) untuk machine learning
  - scipy (versi 1.11.4) untuk komputasi ilmiah

- **Hardware yang Digunakan**
  - Perangkat: Asus Aspire A314-22
  - Processor: AMD Ryzen 3 3250U with Radeon Graphics 2.60 GHz
  - RAM: 8,00 GB DDR4
  - Penyimpanan: 256 GB SSD

- **Alat dan Bahan**
  - Dataset: Dataset berisi data pasien diabetes dengan fitur-fitur yang relevan untuk analisis
  - Laptop : Alat untuk melakukan analisis dan menampilkan hasil penelitian (Google Colab)
  - Perangkat Lunak Pengolah Kata: Perangkat lunak pengolah kata untuk menulis laporan penelitian (seperti Microsoft Word)
  - Akses internet: Untuk mengakses Google Colab dan mengunduh dataset
  - Perangkat lunak: Web browser untuk mengakses Google Colab

BAB IV
Hasil Pembahasan

Pembahasan Metode

Metode hierarkis clustering adalah metode pengelompokkan data secara bertahap/bertingkat. Metode ini menggunakan jarak antar objek untuk mengukur kedekatan antar objek. Dalam metode ini, objek-objek yang memiliki jarak yang paling dekat akan digabungkan menjadi satu cluster. Proses penggabungan ini dilakukan secara berulang-ulang hingga terbentuk sejumlah cluster yang diinginkan.

Dalam kasus ini, saya menggunakan metode hierarkis clustering dengan metode single linkage. Metode single linkage mengukur jarak antar objek dengan cara mengambil jarak terkecil antara dua objek dalam dua cluster yang berbeda.

Metode hierarkis clustering memiliki beberapa kelebihan, antara lain:

- Dapat digunakan untuk mengelompokkan data dalam jumlah besar.
- Dapat menghasilkan struktur hirarki cluster yang dapat digunakan untuk analisis lebih lanjut.

Namun, metode hierarkis clustering juga memiliki beberapa kekurangan, antara lain:

- Dapat menjadi lambat untuk data yang besar.
- Hasil pengelompokan dapat dipengaruhi oleh metode linkage yang digunakan.

Hasil Pemodelan

Dari hasil pemodelan, diperoleh 139 cluster, yaitu cluster dengan outcome 0 dan outcome 1. Cluster outcome 0 berisi data pasien dengan penyakit non diabetes, sedangkan cluster outcome 1 berisi data pasien yang menderita penyakit diabetes.

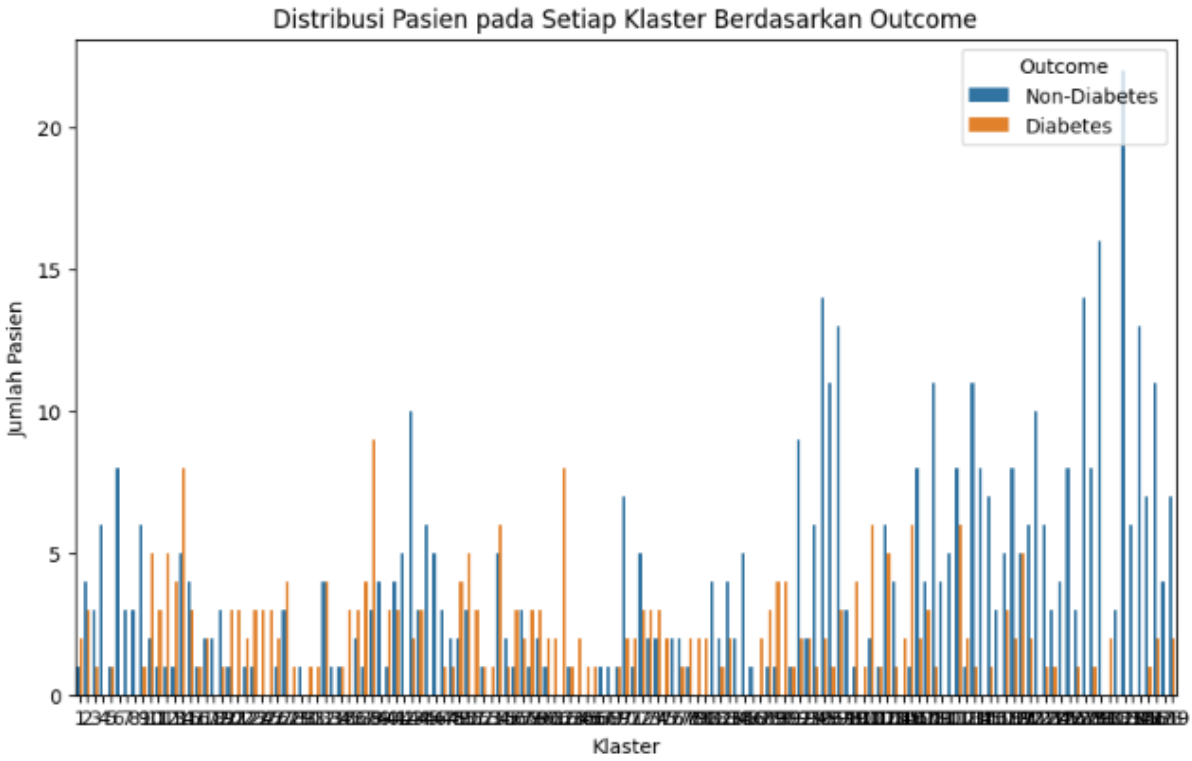
➢ Tabel Hasil Pemodelan

Total Cluster	**Outcome 0 (Non Diabetes)**	**Outcome 1 (Diabetes)**	**Total Pasien**
139	500	268	**768**

Tabel Hasil Clustering:

Outcome	0	1	Total
Cluster			
1	1	2	3
2	4	3	7
3	3	1	4
4	6	0	6
5	1	1	2
...	...	...	...
136	7	1	8
137	11	2	13
138	4	0	4
139	7	2	9
Total	500	268	768

- Grafik Hasil Pemodelan



Dari grafik tersebut, terlihat bahwa data pasien dengan penyakit non diabetes (cluster outcome 0) memiliki distribusi yang lebih beragam dibandingkan dengan data pasien yang menderita penyakit diabetes (outcome cluster 1).

- Hasil Analisa

```
Silhouette Score: 0.155807487418895
Davies-Bouldin Index: 1.1314893347995283
```

Dalam menganalisis hasil prediksi penyakit diabetes menggunakan metode Hierarchical Clustering, saya fokus pada dua metrik evaluasi utama, yaitu Silhouette Score dan Davies-Bouldin Index.

1. Silhouette Score: 0.1558

Silhouette Score mengukur seberapa baik setiap objek berada dalam klaster dibandingkan dengan klaster lainnya. Dengan nilai 0.1558, kita dapat membuat beberapa penilaian:

- Nilai Silhouette Score berada di antara -1 hingga 1. Meskipun nilai 0.1558 tidak terlalu tinggi, itu menunjukkan bahwa klaster memiliki tingkat keseragaman yang moderat.
- Keseragaman ini dapat diinterpretasikan sebagai tingkat keterpisahan yang wajar antar klaster. Namun, perlu diingat bahwa nilai ini tidak memastikan kualitas prediksi secara keseluruhan.

2. Davies-Bouldin Index: 1.1315

Davies-Bouldin Index mengukur seberapa baik pembentukan klaster dilakukan. Dengan nilai 1.1315, kita dapat membuat beberapa analisis:

- Nilai yang rendah menunjukkan bahwa klaster yang dihasilkan memiliki sejumlah jarak yang baik antar klaster, mencerminkan tingkat homogenitas yang memadai.
- Pembentukan klaster yang baik dapat dilihat dari rasio antara seberapa dekat objek dalam klaster dan seberapa jauh klaster satu dengan yang lain.

Kesimpulan Analisis:

1. Kualitas Prediksi Klaster:

- Hasil Silhouette Score menunjukkan bahwa klaster memiliki tingkat keseragaman yang cukup, meskipun tidak sangat tinggi.
- Klaster yang dihasilkan cenderung memiliki karakteristik yang berbeda secara signifikan satu sama lain.

2. Homogenitas Pembentukan Klaster:

- Nilai Davies-Bouldin Index yang rendah mengindikasikan bahwa pembentukan klaster dilakukan dengan baik, dengan sejumlah jarak yang baik antar klaster.
- Klaster yang terbentuk cukup homogen dalam karakteristiknya.

3. Keberhasilan Pembentukan Klaster:

- Meskipun nilai Silhouette Score tidak tinggi, Davies-Bouldin Index yang rendah mengindikasikan bahwa klaster terbentuk dengan baik dan cukup homogen.

Analisis ini dapat membantu dalam memahami sejauh mana metode Hierarchical Clustering dapat memberikan kontribusi pada prediksi penyakit diabetes dan membuka potensi pemahaman lebih lanjut terkait faktor-faktor yang mempengaruhi kondisi kesehatan pasien. Metode hierarkis clustering dengan metode single linkage dapat digunakan untuk melakukan pengelompokkan data pasien diabetes.

BAB V

Kesimpulan

Penelitian ini dilatarbelakangi oleh meningkatnya prevalensi diabetes sebagai tantangan serius dalam dunia kesehatan global. Tujuan utama penelitian adalah melakukan prediksi dini terhadap penyakit diabetes menggunakan Metode Hierarchical Clustering, dengan fokus pada eksplorasi prevalensi, identifikasi pola dan kelompok risiko, analisis faktor risiko potensial, optimasi model prediksi, dan pengembangan strategi pencegahan yang terarah. Metode tersebut diterapkan pada dataset yang mencakup variabel-variabel kesehatan, dan hasil pemodelan menghasilkan 139 cluster dengan outcome 0 (non diabetes) dan outcome 1 (diabetes). Analisis menggunakan Silhouette Score dan Davies-Bouldin Index menunjukkan tingkat keseragaman dan homogenitas klaster yang cukup baik. Hasil pemodelan menunjukkan bahwa data pasien dengan penyakit non diabetes memiliki distribusi yang lebih beragam dibandingkan dengan data pasien diabetes. Evaluasi menggunakan Silhouette Score dan Davies-Bouldin Index menunjukkan tingkat keseragaman dan homogenitas yang moderat, serta pembentukan klaster yang baik. Meskipun nilai Silhouette Score tidak sangat tinggi, nilai Davies-Bouldin Index yang rendah memberikan indikasi bahwa metode Hierarchical Clustering berhasil membentuk klaster yang homogen dan dapat memberikan kontribusi pada prediksi penyakit diabetes. Temuan ini dapat menjadi dasar untuk pengembangan strategi pencegahan yang lebih terarah dan personalisasi dalam manajemen penyakit diabetes. Kesimpulan analisis ini membuktikan bahwa metode Hierarchical Clustering dapat memberikan kontribusi pada prediksi penyakit diabetes.

Referensi

[1] E. Kim, W.-S. Oh, D. S. Pieczkiewicz, M. R. Castro, P. J. Caraballo, and G. Simon, "Divisive Hierarchical Clustering towards Identifying Clinically Significant Pre-Diabetes Subpopulations.," *PubMed*, vol. 2014, pp. 1815–1824, Jan. 2014, [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/25954454

[2] M. Raihan, Md. T. Islam, F. Farzana, Md. G. Raju, and H. S. Mondal, "An empirical study to predict diabetes mellitus using K-means and hierarchical clustering techniques," *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 2019. doi:10.1109/icccnt45670.2019.8944552

[3] J. Fang *et al.*, "Anomaly detection of diabetes data based on hierarchical clustering and CNN," *Procedia Computer Science*, vol. 199, pp. 71–78, 2022. doi:10.1016/j.procs.2022.01.010

[4] M. Altuve, "Hierarchical and Partitional Cluster Analysis of Glucose and Insulin Data from the Oral Glucose Tolerance Test," *Applied Medical Informatics*, vol. 40, pp. 54–62, Dec. 2018,[Online].Available: https://ami.info.umfcluj.ro/index.php/AMI/article/view/640/pdf_68?acceptCookies=1

[5] Thangaraju, P., Deepa, B., and Karthikeyan, T., "Comparison of Data Mining Techniques for Forecasting Diabetes Mellitus," *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 3, no. 8, pp. 7674-7677, 2014.

[6] W. Zhong, R. Chow, R. Stolz, J. He, and M. Dowell, "Hierarchical clustering support vector machines for classifying type-2 diabetes patients," *Bioinformatics Research and Applications*, pp. 379–389, 2008. doi:10.1007/978-3-540-79450-9_35

Lampiran (Google Collab)

Kasus yang diselesaikan

Kasus yang akan diselesaikan adalah prediksi penyakit diabetes menggunakan metode hierarchical clustering. Penelitian ini bertujuan untuk mengidentifikasi pasien diabetes dengan risiko tinggi atau rendah. Untuk mencapai tujuan tersebut, metode hierarchical clustering akan digunakan untuk mengelompokkan data pasien menjadi dua kelompok, yaitu kelompok pasien dengan risiko tinggi dan kelompok pasien dengan risiko rendah. Proses clustering akan dilakukan berdasarkan kesamaan nilai fitur-fitur pada data pasien. Pasien-pasien dengan nilai fitur-fitur yang serupa akan dikelompokkan ke dalam satu kelompok. Hasil clustering kemudian akan digunakan untuk mengidentifikasi pasien diabetes dengan risiko tinggi atau rendah. Pasien-pasien yang termasuk dalam kelompok risiko tinggi akan memiliki kemungkinan lebih besar untuk menderita diabetes. Penelitian ini diharapkan dapat memberikan informasi yang berguna untuk membantu dokter dalam mendiagnosis penyakit diabetes.

Version Python

Versi Python yang digunakan adalah Python 3.10.12

Library yang dibutuhkan, version library dan link sumber library

- **pandas** (versi 1.5.3) untuk analisis dan manipulasi data

Link:
https://pandas.pydata.org/pandas-docs/version/1.5.3/getting_started/install.html

- **seaborn** (versi 0.12.2) untuk visualisasi data

```
# seaborn
import seaborn
print('seaborn: {}'.format(seaborn.__version__))
```

```
seaborn: 0.12.2
```

Link:
https://seaborn.pydata.org/whatsnew/v0.12.2.html

- **matplotlib** (versi 3.7.1) untuk visualisasi data

```
[9] # matplotlib
    import matplotlib
    print('matplotlib: {}'.format(matplotlib.__version__))
```

```
matplotlib: 3.7.1
```

Link:
https://matplotlib.org/3.7.1/

- **scikit-learn** (versi 1.2.2) untuk machine learning

```
[11] # scikit-learn
     import sklearn
     print('sklearn: {}'.format(sklearn.__version__))
```

```
sklearn: 1.2.2
```

Link:
https://scikit-learn.org/dev/whats_new/v1.2.html

- **scipy** (versi 1.11.4) untuk komputasi ilmiah

```
[13] # scipy
     import scipy
     print('scipy: {}'.format(scipy.__version__))
```

```
scipy: 1.11.4
```

Link:
https://docs.scipy.org/doc/scipy/release/1.11.4-notes.html

Untuk link yang terbaru dari semua libraries dapat dicari pada link berikut:
https://pypi.org/

Pembahasan code dan disertai penjelasan

a) Import library

```
# Import library
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from scipy.cluster.hierarchy import dendrogram, linkage, fcluster
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import silhouette_score, davies_bouldin_score
```

- **pandas:** Digunakan untuk memuat dan memanipulasi dataset.
- **seaborn:** Digunakan untuk visualisasi data.
- **matplotlib.pyplot:** Digunakan untuk membuat plot dan grafik.
- **sklearn.model_selection:** Digunakan untuk membagi dataset menjadi set training dan set testing.
- **scipy.cluster.hierarchy:** Digunakan untuk melakukan hierarchical clustering.
- **sklearn.preprocessing**: Digunakan untuk melakukan preprocessing data, seperti standarisasi.
- **sklearn.metrics:** Digunakan untuk mengevaluasi hasil clustering.

b) Pengumpulan data (Load data)

```
# Pengumpulan data (Load data)
url = "/content/diabetes.csv"  # Ganti dengan path dataset yang
sesuai
df = pd.read_csv(url)

# Menampilkan 5 baris pertama dari dataset
df.head()
```

- **url = "/content/diabetes.csv":** Ini adalah path atau URL menuju file dataset. Pada contoh ini, diabetes.csv diharapkan berada dalam direktori /content/ di Google Colab. Pastikan bahwa path ini sesuai dengan lokasi sebenarnya dari dataset yang Anda miliki.
- **df = pd.read_csv(url):** Ini adalah baris kode yang digunakan untuk membaca file CSV dan menyimpannya ke dalam DataFrame menggunakan fungsi read_csv dari pandas. DataFrame ini kemudian akan digunakan untuk analisis dan pemodelan.
- df.**head():** Ini adalah baris kode yang menampilkan 5 baris pertama dari DataFrame untuk memberikan gambaran awal tentang struktur dan konten dataset. Fungsi head() digunakan untuk melihat sebagian kecil dari dataset.

c) Explorasi data análisis (EDA)

```
# Menampilkan informasi dataset seperti tipe data dan jumlah nilai
non-null
df.info()

# Menampilkan statistik deskriptif dari dataset
df.describe()

# Visualisasi korelasi antar variabel
correlation_matrix = df.corr()
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Korelasi Antar Variabel')
plt.show()

#Visualisasi Histogramuntuk masing masing kolom
df.hist(figsize=(10, 10))
plt.show()

# Visualisasi
sns.pairplot(df, hue='Outcome', diag_kind='kde')
plt.show()
```

- **df.info():** Baris ini menampilkan informasi dataset, termasuk tipe data dari setiap kolom dan jumlah nilai non-null. Ini membantu Anda memahami apakah ada nilai yang hilang (null) dalam dataset.
- **df.describe():** Baris ini menampilkan statistik deskriptif dari dataset, seperti rata-rata, standar deviasi, nilai minimum, kuartil, dan nilai maksimum untuk setiap kolom numerik. Ini memberikan gambaran lebih lanjut tentang distribusi data.
- **correlation_matrix = df.corr():** Baris ini menghitung matriks korelasi antar variabel numerik dalam dataset.
- **sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm'):** Baris ini menghasilkan heatmap untuk memvisualisasikan korelasi antar variabel numerik. Semakin dekat nilai korelasi ke 1 atau -1, semakin tinggi hubungan antar variabelnya.
- **df.hist(figsize=(10, 10)):** Baris ini menampilkan histogram untuk masing-masing kolom numerik dalam dataset. Histogram memberikan gambaran distribusi frekuensi dari setiap variabel.
- **sns.pairplot(df, hue='Outcome', diag_kind='kde'):** Baris ini menghasilkan pair plot yang memvisualisasikan hubungan dua per dua antar variabel, dengan pemisahan warna berdasarkan variabel target 'Outcome'. Juga, pada diagonalnya, terdapat distribusi variabel masing-masing.

d) Preprocessing data (Feature selection)

```
# Preprocessing data (Feature selection)
X = df.drop(['Outcome'], axis=1)
y = df['Outcome']


# Standarisasi data
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

- **X = df.drop(['Outcome'], axis=1):** Baris ini membuat variabel X yang berisi fitur-fitur (kolom-kolom) dari dataset, kecuali kolom 'Outcome'. Ini dilakukan dengan menggunakan fungsi drop untuk menghilangkan kolom 'Outcome' dari DataFrame, dan hasilnya disimpan dalam variabel X.
- **y = df['Outcome']:** Baris ini membuat variabel y yang berisi target atau label yang ingin diprediksi. Pada kasus ini, kita ingin memprediksi kolom 'Outcome', yang menyatakan apakah pasien memiliki diabetes atau tidak.
- **X_scaled = StandardScaler():** Baris ini membuat objek scaler dari kelas StandardScaler. StandardScaler digunakan untuk standarisasi data, sehingga setiap fitur memiliki rata-rata nol dan deviasi standar satu. Ini berguna untuk menghilangkan dampak skala yang berbeda antar fitur.
- **X = scaler.fit_transform(X):** Baris ini menggunakan objek scaler yang telah dibuat sebelumnya untuk melakukan transformasi standarisasi pada data fitur (X). Hasil transformasi ini kemudian disimpan kembali ke variabel X. Dengan demikian, setiap nilai dalam setiap kolom fitur sekarang memiliki skala yang seragam.

e) Spliting data

```
# Spliting data
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y,
test_size=0.2, random_state=42)
```

- **X_train, X_test, y_train, y_test**: Baris ini membuat empat variabel yang menyimpan hasil dari fungsi train_test_split. Fungsi ini digunakan untuk membagi dataset menjadi dua bagian, yaitu set pelatihan dan set pengujian.
- **train_test_split(X_scaled, y, test_size=0.2, random_state=42):** Fungsi ini menerima beberapa parameter:
- **X_scaled:** Matriks fitur yang telah di-standarisasi.
- y: Vektor target.
- **test_size=0.2:** Persentase data yang akan dialokasikan sebagai set pengujian. Dalam contoh ini, 20% dari data akan digunakan untuk pengujian.

- **random_state=42:** Digunakan untuk menghasilkan pembagian yang konsisten setiap kali kode dijalankan, sehingga hasilnya dapat direproduksi.

Hasil dari train_test_split adalah:

- **X_train:** Set pelatihan untuk fitur.
- **X_test:** Set pengujian untuk fitur.
- **y_train:** Set pelatihan untuk target.
- **y_test:** Set pengujian untuk target.

f) Pembuatan model

```
# Pembuatan model
# Hierarchical Clustering
linkage_matrix = linkage(X_scaled, method='ward')

# Dendrogram
plt.figure(figsize=(12, 6))
dendrogram(linkage_matrix)
plt.title('Dendrogram')
plt.show()

# Menentukan jumlah cluster dengan metode Cut-off (fcluster)
# Anda dapat menyesuaikan nilai treshold sesuai dengan kebutuhan
treshold = 3
clusters = fcluster(linkage_matrix, treshold, criterion='distance')

# Menampilkan hasil clustering
print("Hasil Clustering:")
print(clusters)
```

- **linkage_matrix = linkage(X_scaled, method='ward'):** Baris ini menggunakan fungsi linkage dari scipy.cluster.hierarchy untuk melakukan hierarchical clustering pada data yang telah di-standarisasi (X_scaled). Metode yang digunakan di sini adalah 'ward', yang merupakan salah satu metode linkage untuk mengukur jarak antar cluster. Hasilnya disimpan dalam variabel linkage_matrix.
- **plt.figure(figsize=(12, 6)):** Baris ini menciptakan figur dengan ukuran (lebar, tinggi) sebesar (12, 6) untuk menempatkan dendrogram.
- **dendrogram(linkage_matrix):** Baris ini menggunakan fungsi dendrogram untuk menggambar dendrogram berdasarkan matriks linkage yang telah dihasilkan sebelumnya.
- **plt.title('Dendrogram'):** Baris ini memberikan judul "Dendrogram" pada plot.
- **plt.show():** Baris ini menampilkan dendrogram yang sudah dihasilkan.

- **treshold = 3:** Baris ini menetapkan nilai threshold atau ambang batas untuk menentukan jumlah cluster. Anda dapat menyesuaikan nilai ini sesuai dengan kebutuhan atau dengan menganalisis dendrogram.
- **clusters = fcluster(linkage_matrix, treshold, criterion='distance'):** Baris ini menggunakan fungsi fcluster untuk melakukan clustering dan menetapkan label cluster ke setiap sampel berdasarkan nilai ambang batas (threshold) yang ditentukan. Parameter criterion='distance' menunjukkan bahwa kita menggunakan ambang batas berdasarkan jarak antar cluster.
- **print("Hasil Clustering:"):** Baris ini mencetak hasil clustering ke layar.
- **print(clusters):** Baris ini mencetak label cluster untuk setiap sampel.

g) Hasil model

```
# Hasil model
# Menambahkan kolom cluster ke dataframe
df['Cluster'] = clusters

# Tabel hasil clustering
cluster_table = pd.crosstab(df['Cluster'], df['Outcome'],
margins=True, margins_name='Total')
print("Tabel Hasil Clustering:")
print(cluster_table)

# Grafik hasil clustering
plt.figure(figsize=(10, 6))
sns.countplot(x='Cluster', hue='Outcome', data=df)
plt.title('Distribusi Pasien pada Setiap Klaster Berdasarkan
Outcome')
plt.xlabel('Klaster')
plt.ylabel('Jumlah Pasien')
plt.legend(title='Outcome', loc='upper right', labels=['Non-
Diabetes', 'Diabetes'])
plt.show()

# Statistik rata-rata untuk setiap cluster terhadap setiap fitur
cluster_means = df.groupby('Cluster').mean()
print("\nStatistik Rata-rata untuk Setiap Cluster:")
print(cluster_means)
```

- **df['Cluster'] = clusters:** Baris ini menambahkan kolom 'Cluster' ke dalam DataFrame df yang berisi label cluster untuk setiap sampel.

- **cluster_table = pd.crosstab(df['Cluster'], df['Outcome'], margins=True, margins_name='Total'):** Baris ini menggunakan pd.crosstab untuk membuat tabel kontingensi antara label cluster dan variabel target 'Outcome'. Tabel ini memberikan informasi tentang distribusi 'Outcome' di setiap cluster. Parameter margins=True menambahkan total baris dan total kolom, serta margins_name='Total' memberikan nama untuk total tersebut.
- **print("Tabel Hasil Clustering:"):** Baris ini mencetak tabel hasil clustering ke layar.
- **print(cluster_table):** Baris ini mencetak tabel hasil clustering yang berisi jumlah pasien diabetes dan non-diabetes di setiap klaster.
- **plt.figure(figsize=(10, 6)):** Baris ini membuat objek figur dengan ukuran 10x6 untuk menempatkan plot distribusi pasien pada setiap klaster berdasarkan variabel target 'Outcome'.
- **sns.countplot(x='Cluster', hue='Outcome', data=df**): Baris ini menggunakan sns.countplot untuk membuat grafik batang yang menunjukkan distribusi pasien di setiap klaster berdasarkan variabel target 'Outcome'.
- **plt.title('Distribusi Pasien pada Setiap Klaster Berdasarkan Outcome'):** Baris ini memberikan judul pada plot.
- plt.xlabel('Klaster') dan plt.ylabel('Jumlah Pasien'): Baris ini memberikan label sumbu x dan y pada plot.
- **plt.legend(title='Outcome', loc='upper right', labels=['Non-Diabetes', 'Diabetes']):** Baris ini menambahkan legenda pada plot untuk menjelaskan warna yang merepresentasikan diabetes dan non-diabetes.
- **plt.show():** Baris ini menampilkan plot grafik hasil clustering.
- **cluster_means = df.groupby('Cluster').mean():** Baris ini menghitung rata-rata setiap fitur untuk setiap klaster.
- **print("\nStatistik Rata-rata untuk Setiap Cluster:") dan print(cluster_means):** Baris ini mencetak statistik rata-rata untuk setiap klaster ke layar.

h) Evaluasi model

```
# Evaluasi model
# Silhouette Score
silhouette_avg = silhouette_score(X_scaled, clusters)
print(f"Silhouette Score: {silhouette_avg}")

# Davies-Bouldin Index
db_index = davies_bouldin_score(X_scaled, clusters)
print(f"Davies-Bouldin Index: {db_index}")


# Menampilkan hasil evaluasi dalam bentuk grafik
plt.figure(figsize=(8, 5))
```

```
evaluasi_df.plot(kind='bar', legend=False)
plt.title('Evaluasi Model')
plt.ylabel('Score')
plt.xticks(rotation=0)
plt.show()
```

- **silhouette_avg = silhouette_score(X_scaled, clusters):** Baris ini menggunakan fungsi silhouette_score dari sklearn.metrics untuk menghitung nilai Silhouette Score. Silhouette Score mengukur seberapa baik setiap sampel ditempatkan dalam kluster, dengan nilai berkisar dari -1 (salah kluster) hingga 1 (pengelompokan yang baik).
- **print(f"Silhouette Score: {silhouette_avg}"):** Baris ini mencetak nilai Silhouette Score ke layar**.**
- **db_index = davies_bouldin_score(X_scaled, clusters):** Baris ini menggunakan fungsi davies_bouldin_score dari sklearn.metrics untuk menghitung nilai Davies-Bouldin Index. Davies-Bouldin Index mengukur seberapa baik kluster dipisahkan dan homogen. Nilai terendah menunjukkan kluster yang lebih baik.
- **print(f"Davies-Bouldin Index: {db_index**}"): Baris ini mencetak nilai Davies-Bouldin Index ke layar.
- **plt.figure(figsize=(8, 5)):** Baris ini membuat objek figur dengan ukuran 8x5 untuk menempatkan plot hasil evaluasi.
- **evaluasi_df.plot(kind='bar', legend=False):** Baris ini menggunakan plot dari pandas untuk membuat grafik batang yang menunjukkan nilai Silhouette Score dan Davies-Bouldin Index.
- **plt.title('Evaluasi Model'):** Baris ini memberikan judul pada plot.
- **plt.ylabel('Score'):** Baris ini memberikan label pada sumbu y.
- **plt.xticks(rotation=0):** Baris ini mengatur rotasi label sumbu x agar tetap horizontal.
- **plt.show():** Baris ini menampilkan plot grafik hasil evaluasi model.

Link Google Colab:

https://colab.research.google.com/drive/1MyPbYPfSWSWNjzRlRuXpubsvP4I_8VW1?usp=sharing

Link Dataset

https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset