

# K-Means

Michaell Abelard Hendra 71487



Dataset

Tujuan

Metodologi

2 Cluster

3 Cluster

4 Cluster

Conclusion

# Dataset

Dataset ini berisi data medis dari perempuan keturunan Pima Indian berusia di atas 21 tahun, dan digunakan untuk memprediksi kemungkinan seseorang mengidap diabetes berdasarkan hasil pemeriksaan klinis.

Fitur	Penjelasan Singkat
Pregnancies	Jumlah kehamilan
Glucose	Kadar gula darah (glukosa)
BloodPressure	Tekanan darah diastolik (mm Hg)
SkinThickness	Ketebalan lipatan kulit (mm)
Insulin	Kadar insulin serum (mu U/ml)
BMI	Indeks Massa Tubuh ( $\text{kg}/\text{m}^2$ )
DiabetesPedigreeFunction	Skor riwayat keluarga terkait diabetes
Age	Usia (dalam tahun)
Outcome	Label: 0 = Tidak diabetes, 1 = Diabetes

# Tujuan

**vision**

- Mengidentifikasi pola dan struktur dalam data pasien berdasarkan fitur-fitur medis seperti kadar glukosa, insulin, tekanan darah, BMI, dan usia.
- Mengelompokkan pasien berdasarkan kemiripan karakteristik menggunakan metode unsupervised learning (K-Means Clustering) untuk melihat apakah data secara alami membentuk kelompok tertentu yang relevan secara klinis.
- Membangun model klasifikasi menggunakan metode supervised learning (K-Nearest Neighbors) untuk memprediksi status diabetes pasien berdasarkan data historis yang telah berlabel.
- Menganalisis hasil dari pendekatan unsupervised dan supervised dalam mengenali pasien dengan risiko diabetes, serta mengevaluasi performa masing-masing metode.
- Mengeksplorasi kemungkinan adanya sub-kelompok pasien yang tidak secara langsung sesuai dengan label “diabetes” atau “tidak”, seperti prediabetes atau sindrom metabolik.

# Metodologi

Penentuan  
Cluster

Visualisasi  
Scatter Plot

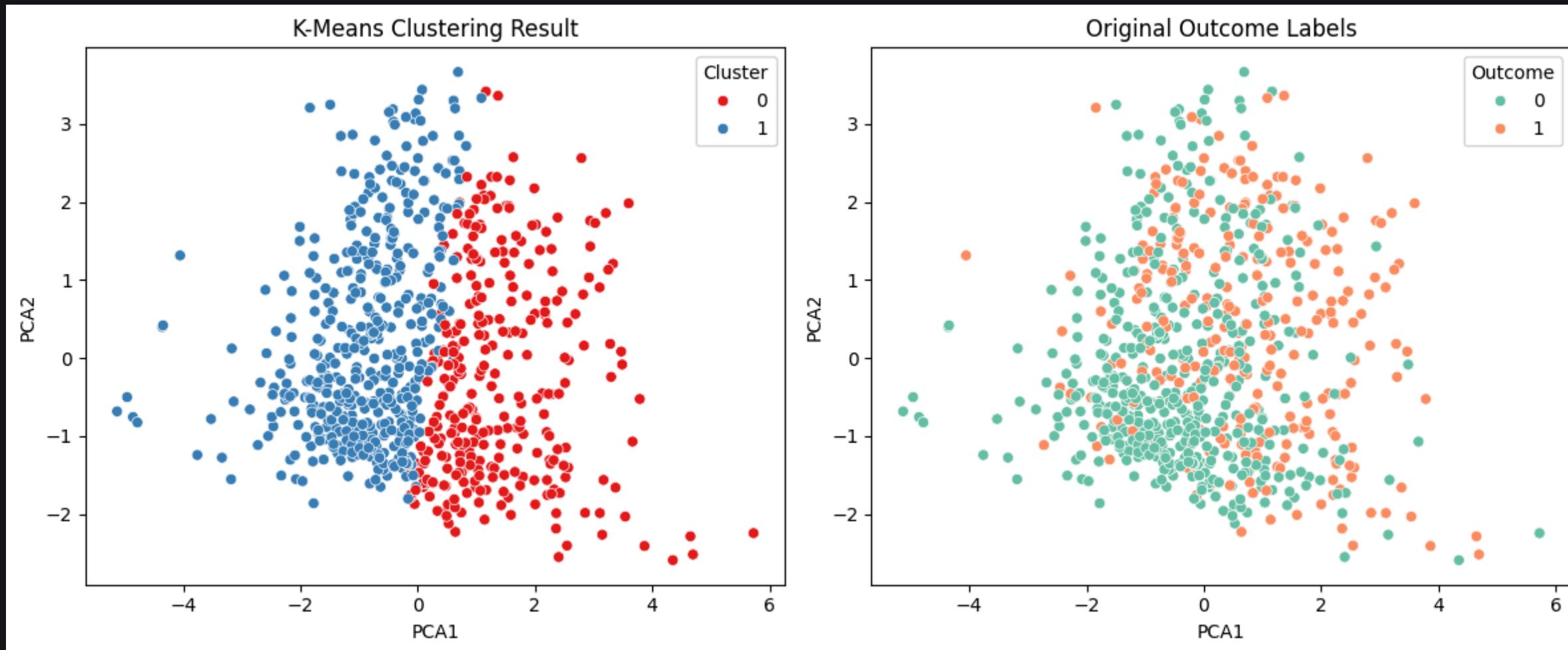
Visualisasi  
Pie Chart

Evaluasi  
Model

Analisis  
Hasil  
Model

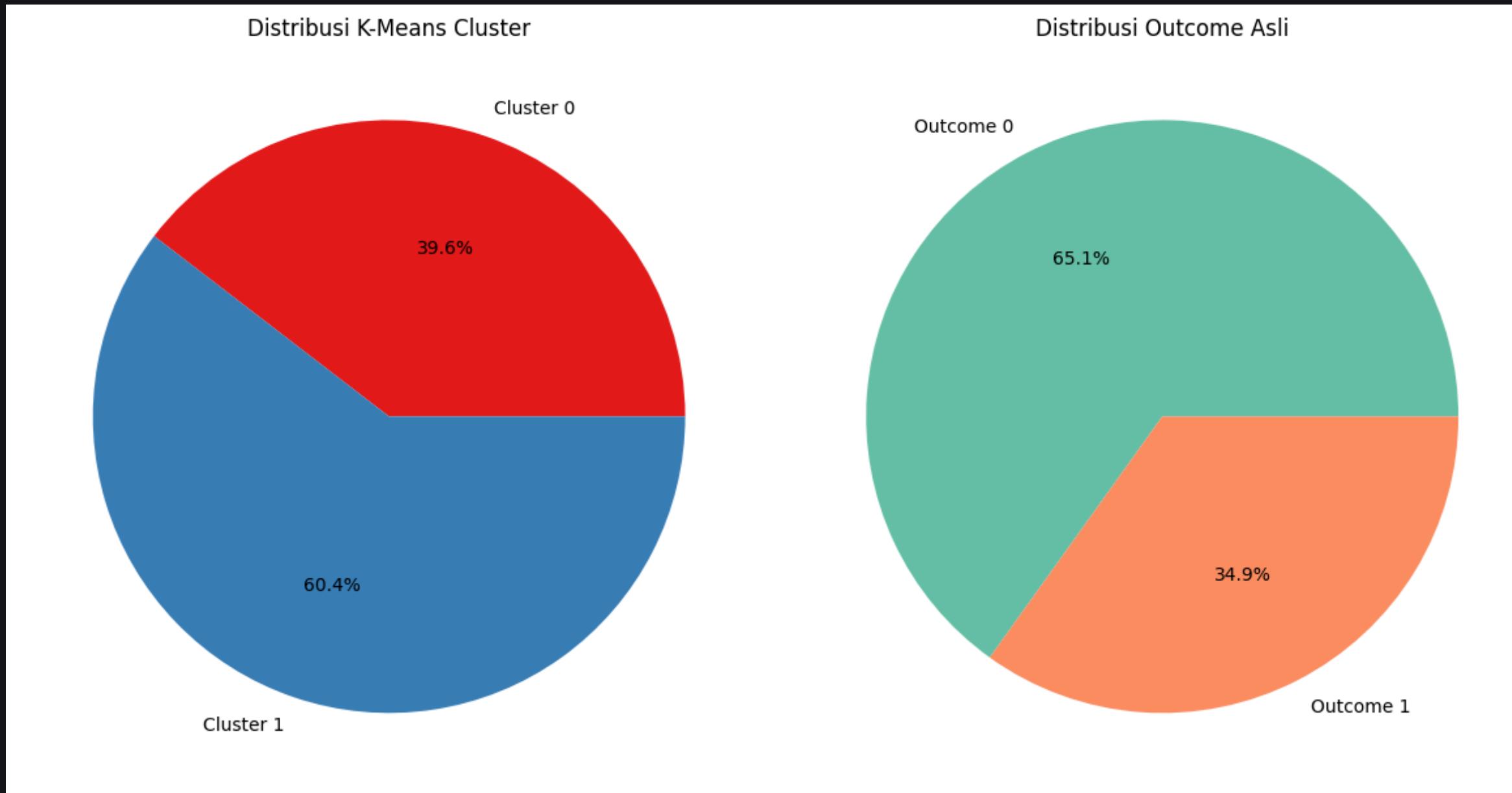


# 2 Cluster



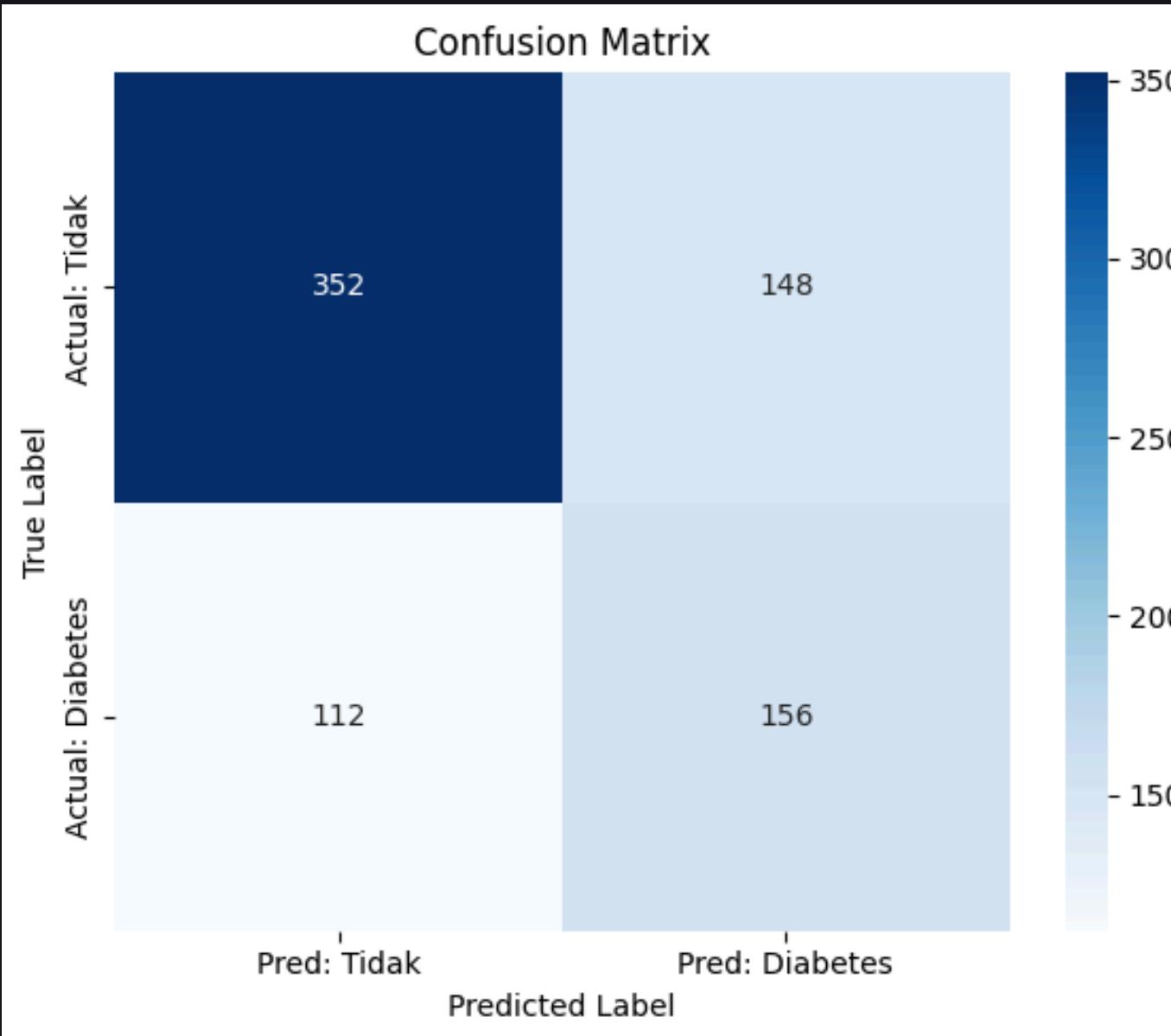
Data hanya terbagi secara vertikal

# 2 Cluster



- Secara proporsi, terbagi secara signifikan
  - Cluster 0 = Outcome 1
  - Cluster 1 = Outcome 0

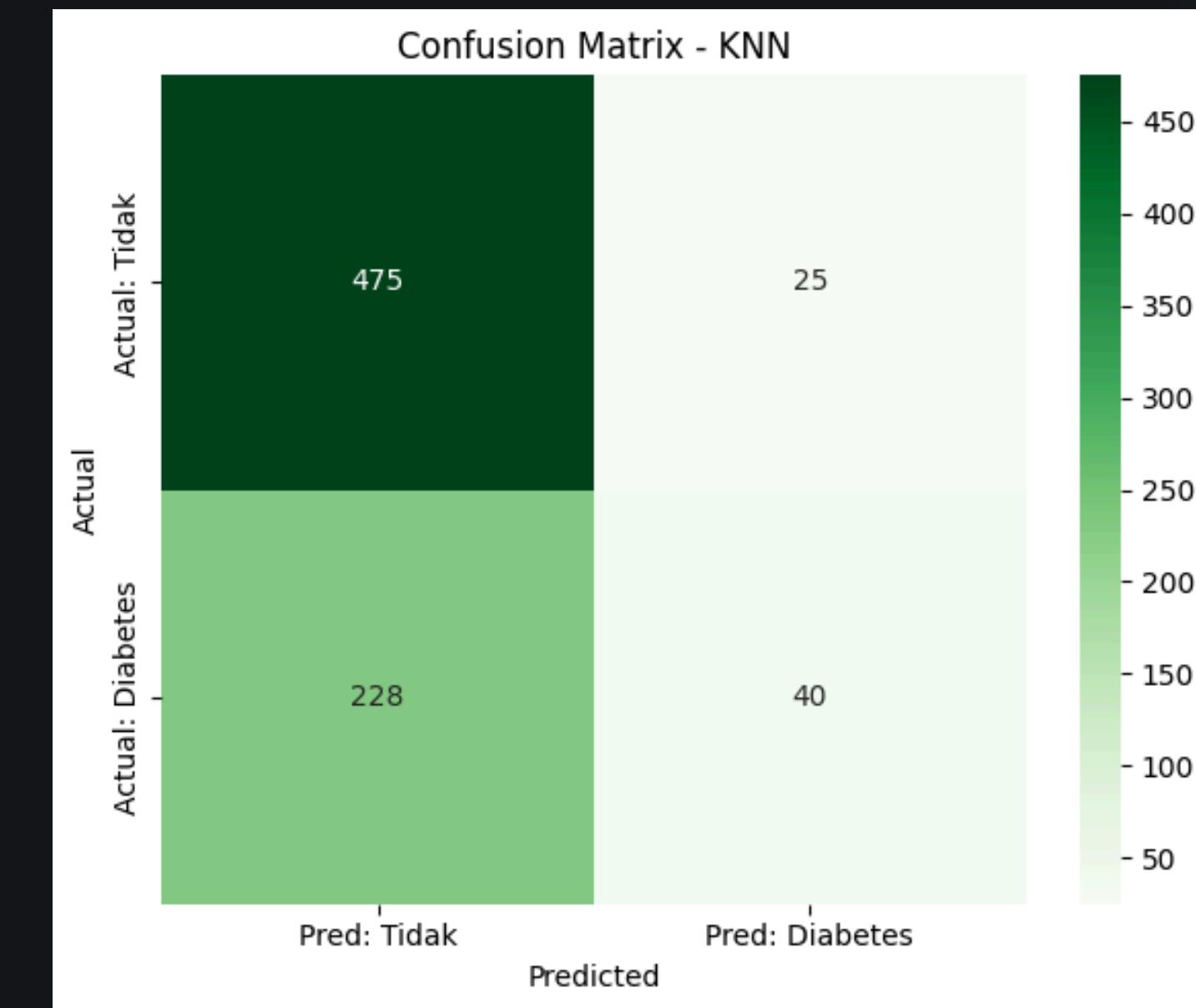
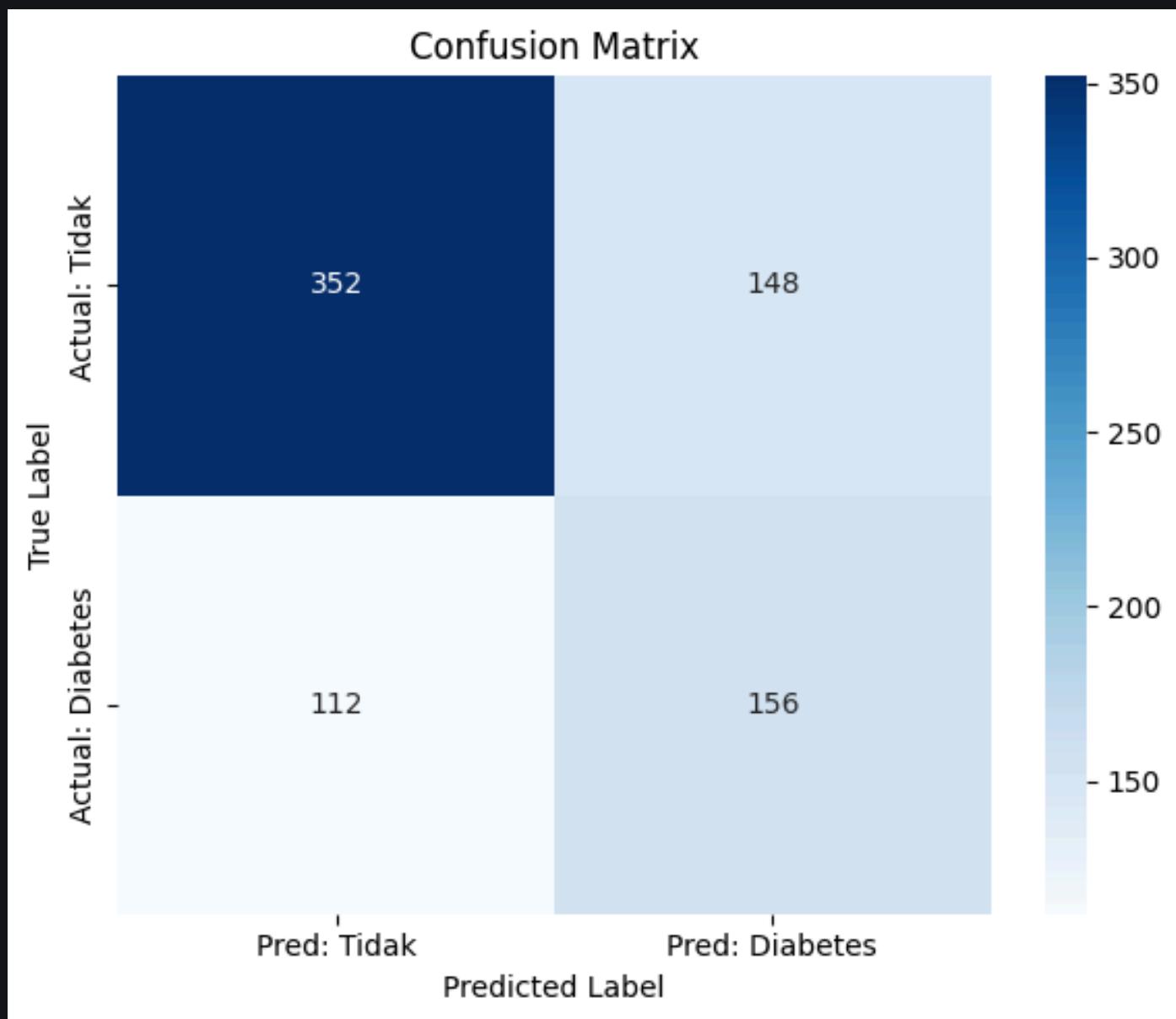
# 2 Cluster



Classification Report:

	precision	recall	f1-score	support
Tidak Diabetes	0.76	0.70	0.73	500
Diabetes	0.51	0.58	0.55	268
accuracy			0.66	768
macro avg	0.64	0.64	0.64	768
weighted avg	0.67	0.66	0.67	768

# 2 Cluster



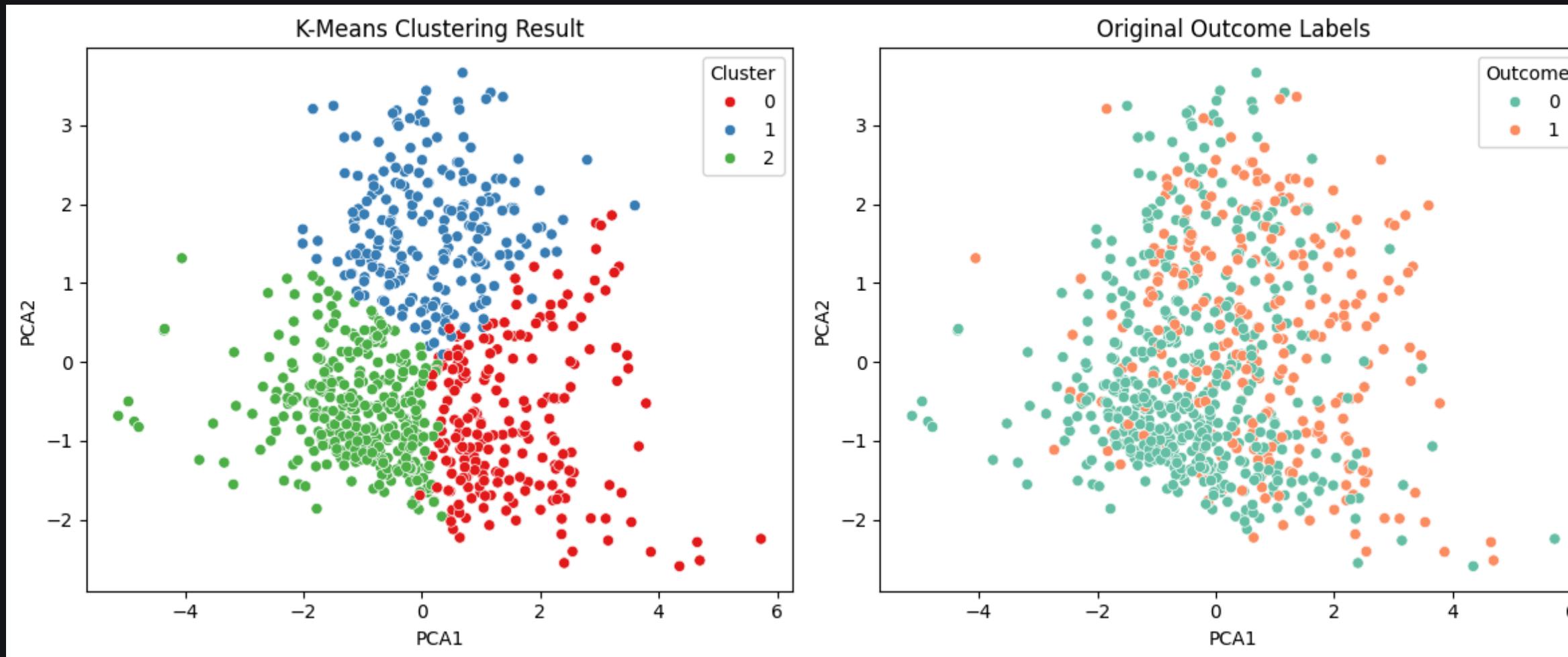
Classification Report:

	precision	recall	f1-score	support
Tidak Diabetes	0.76	0.70	0.73	500
Diabetes	0.51	0.58	0.55	268
accuracy			0.66	768
macro avg	0.64	0.64	0.64	768
weighted avg	0.67	0.66	0.67	768

KNN Classification Report:

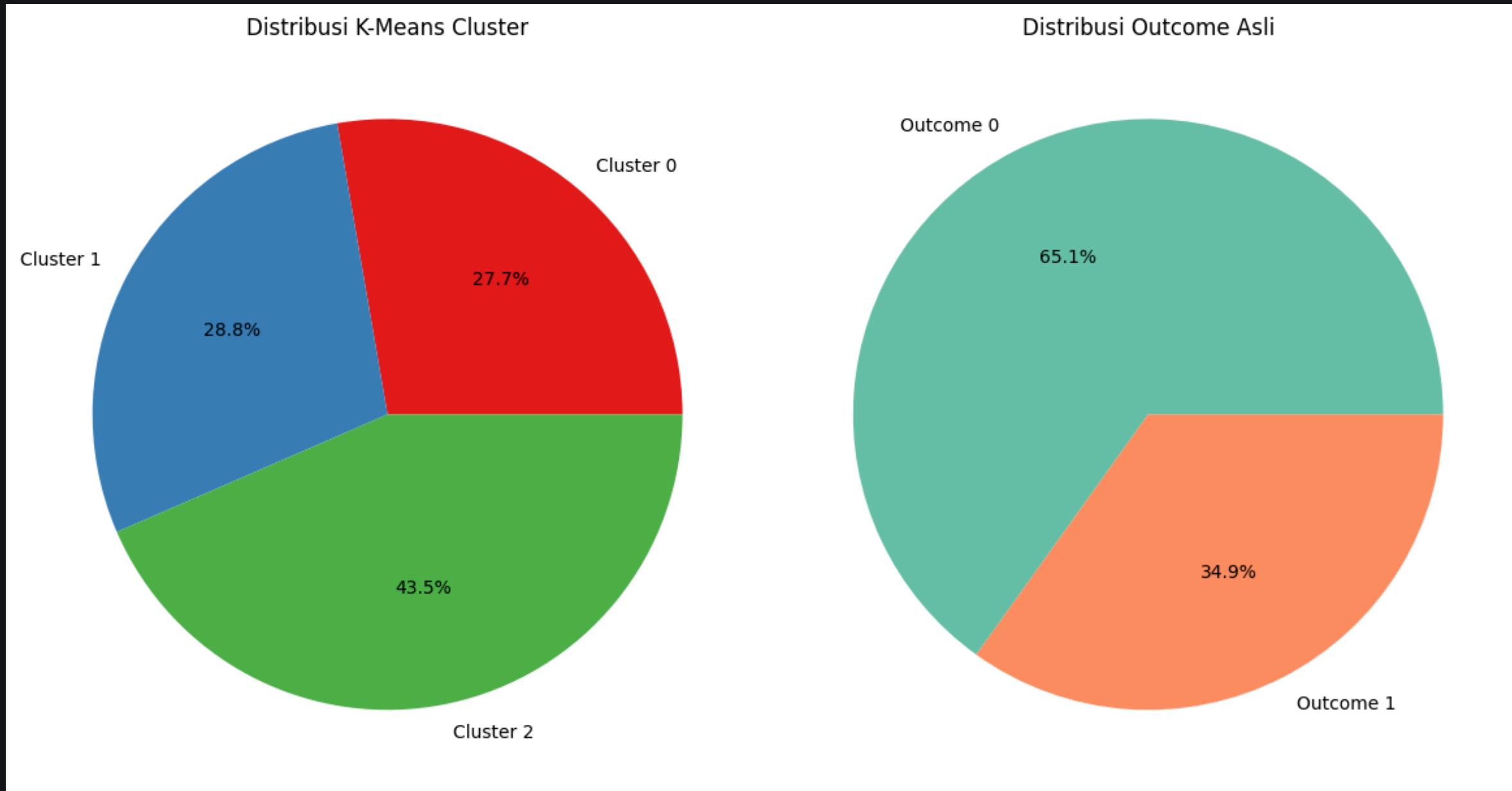
	precision	recall	f1-score	support
Tidak Diabetes	0.68	0.95	0.79	500
Diabetes	0.62	0.15	0.24	268
accuracy			0.67	768
macro avg	0.65	0.55	0.51	768
weighted avg	0.65	0.67	0.60	768

# 3 Cluster



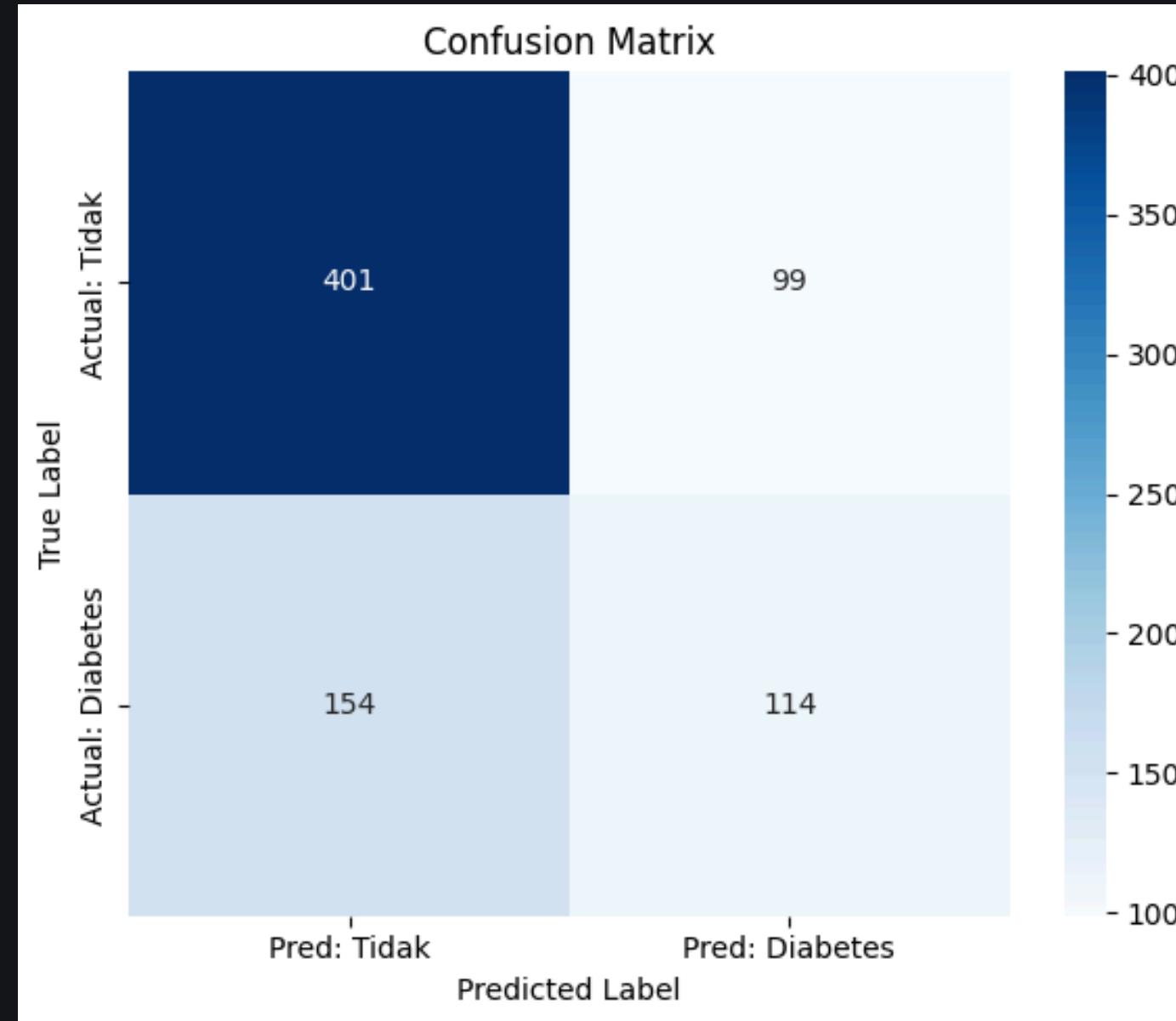
Data terbagi secara proporsional

# 3 Cluster



- Cluster 0 = Outcome 1
- Cluster 1 = Outcome 0
- Cluster 2 = Outcome 0

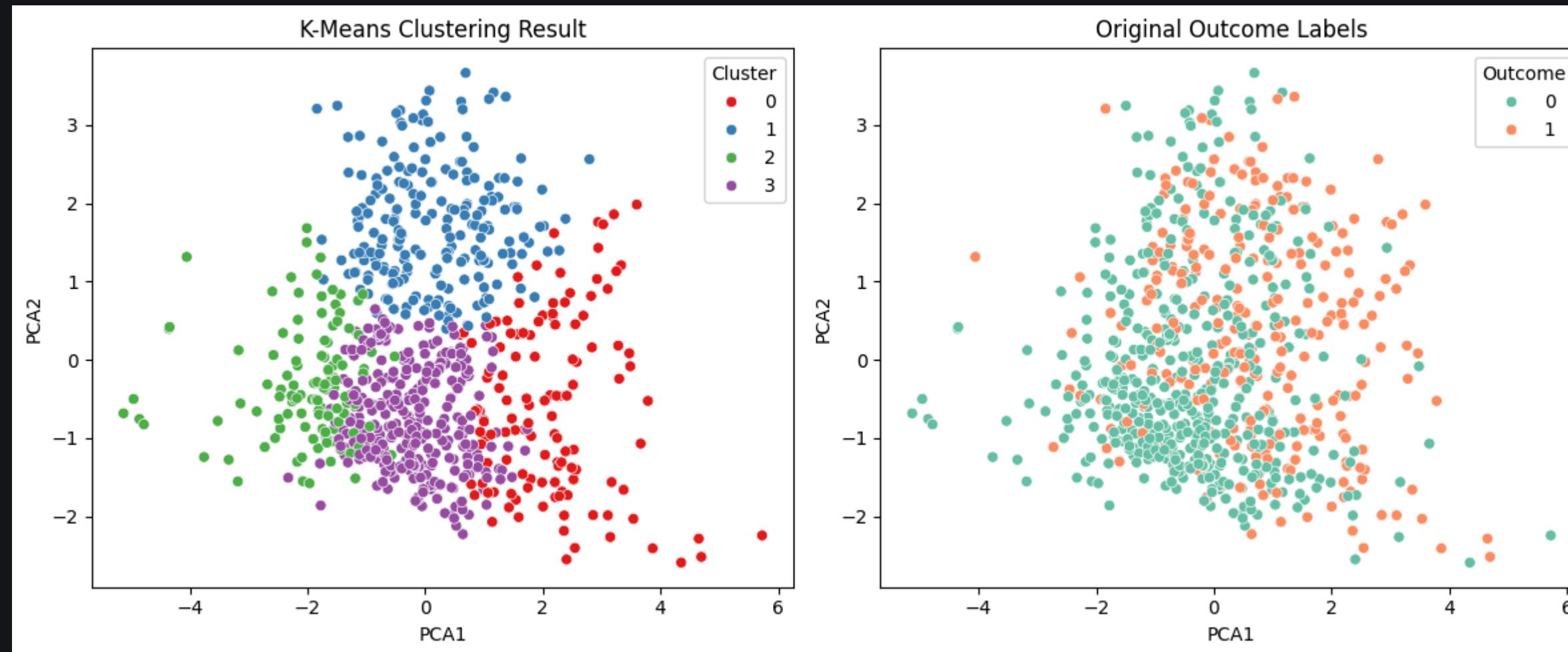
# 3 Cluster



Classification Report:

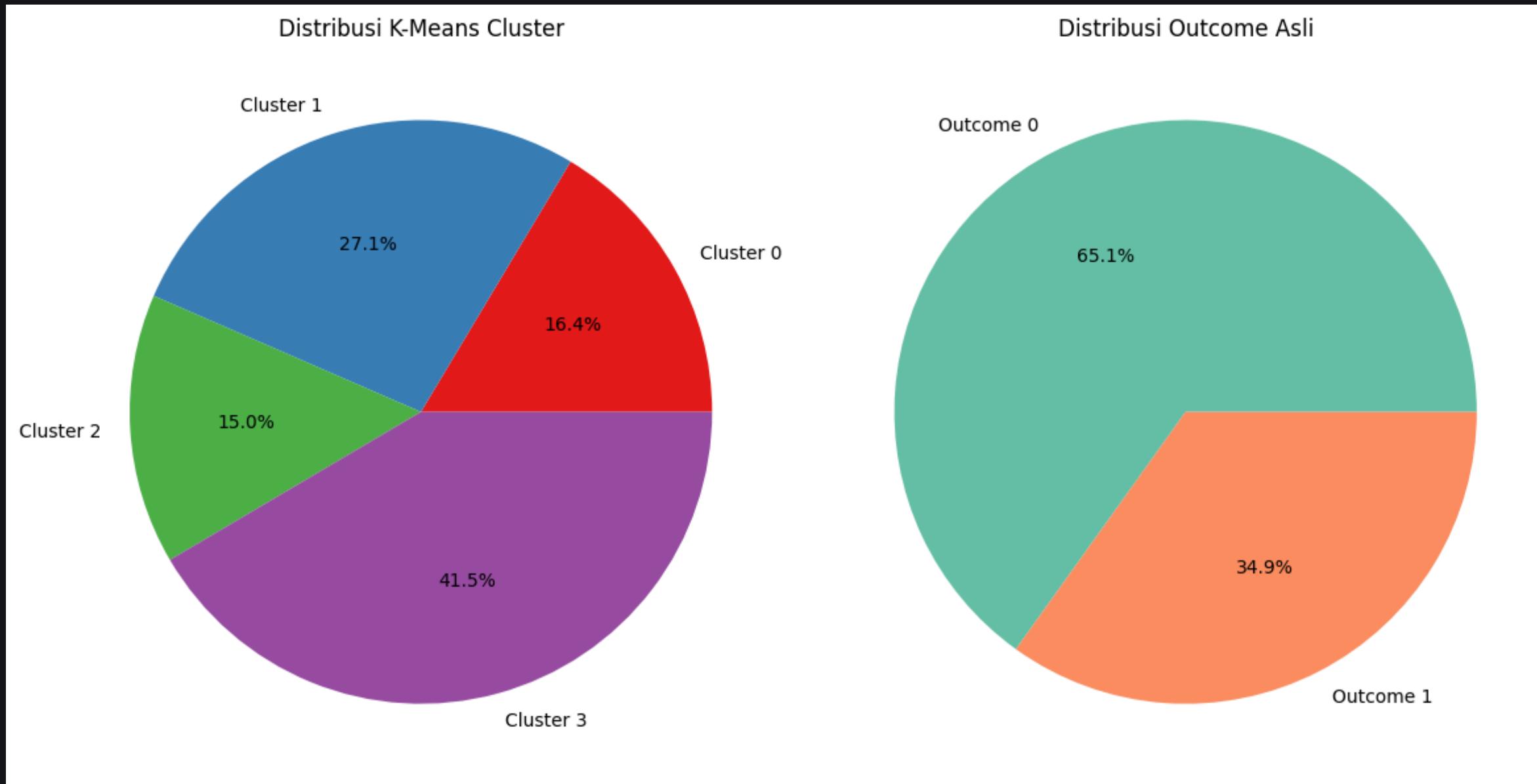
	precision	recall	f1-score	support
Tidak	0.72	0.80	0.76	500
Diabetes	0.54	0.43	0.47	268
accuracy			0.67	768
macro avg	0.63	0.61	0.62	768
weighted avg	0.66	0.67	0.66	768

# 4 Cluster

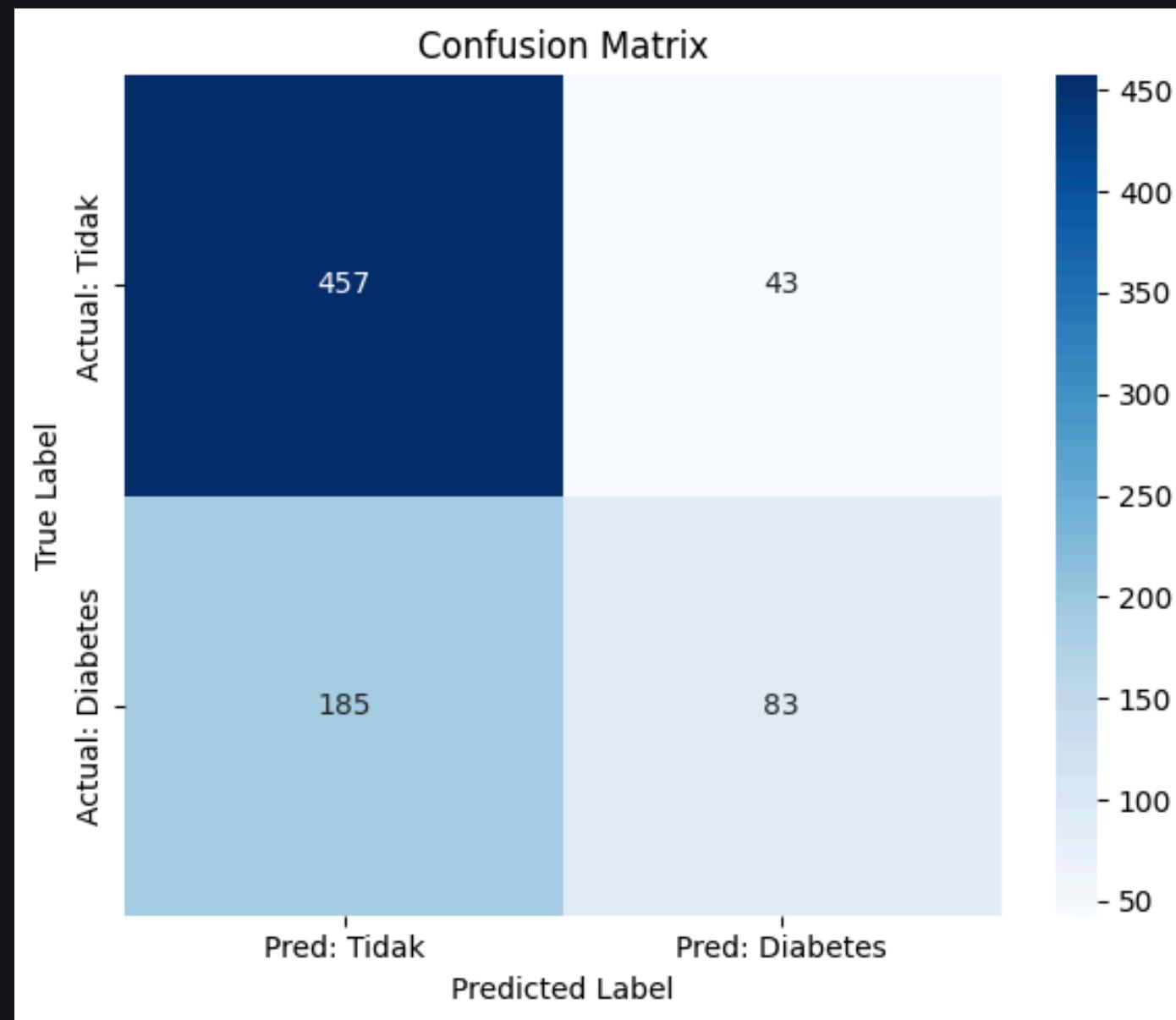


Data mulai terbagi  
dengan persebaran lebih  
baik

# 4 Cluster



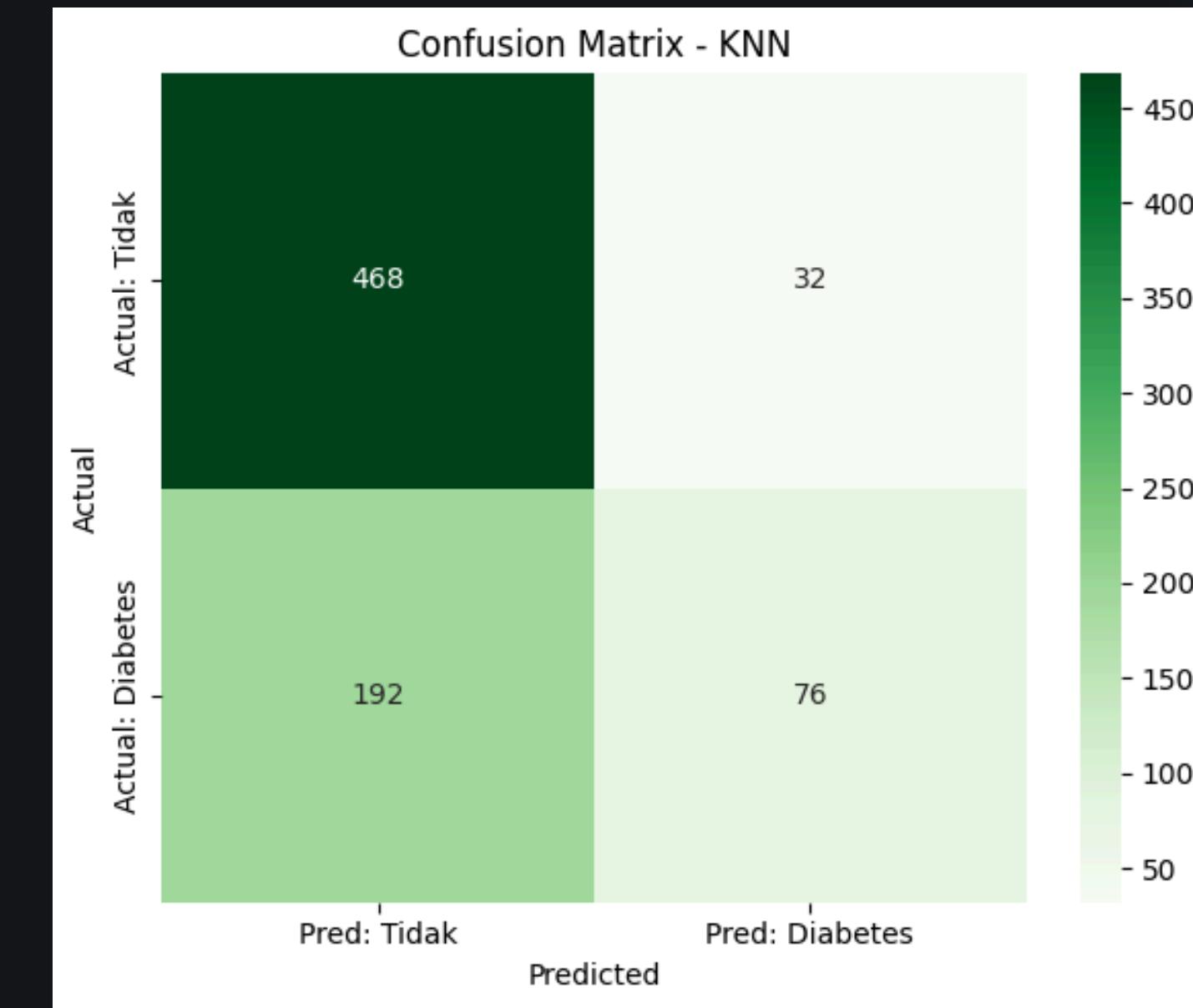
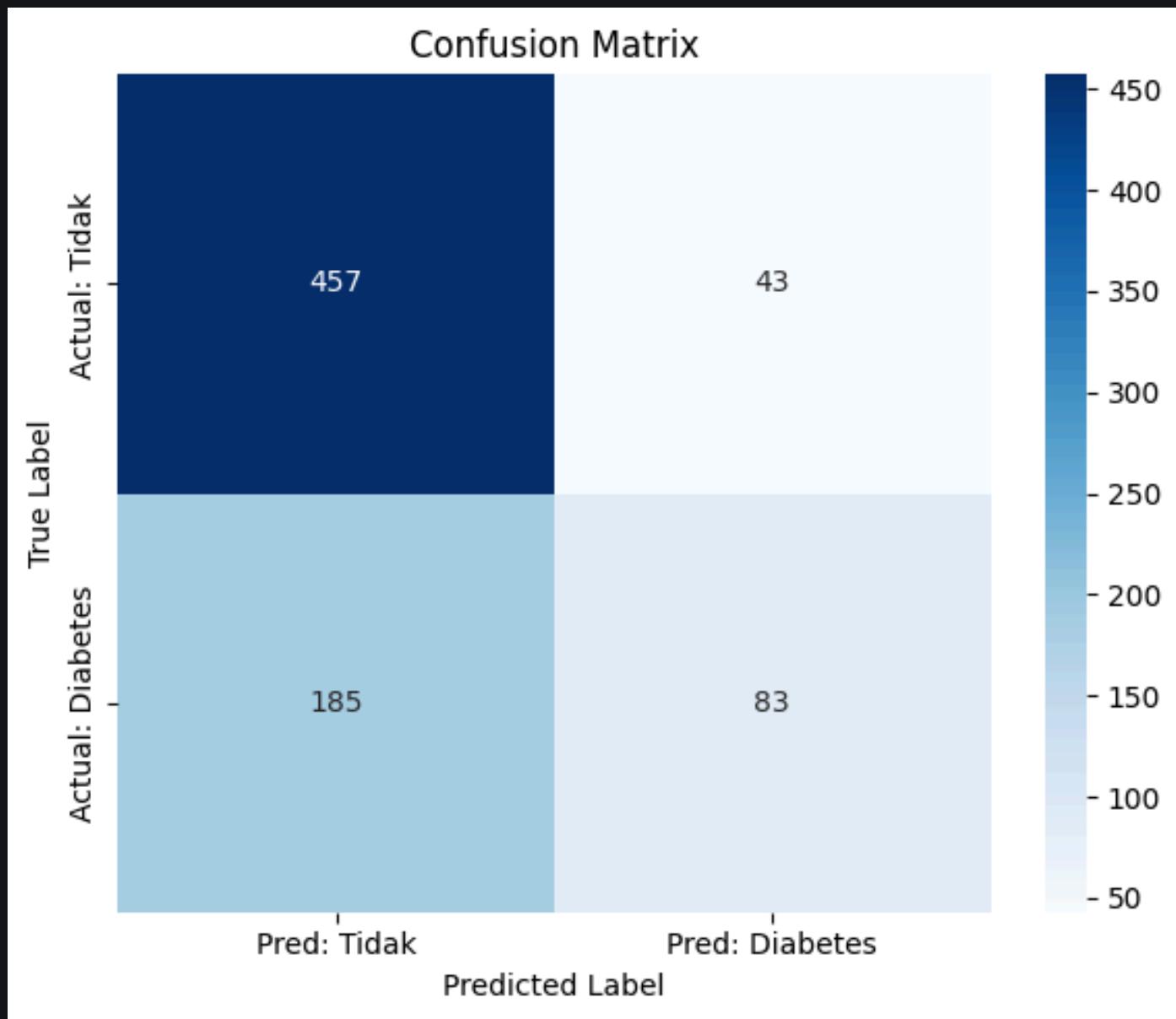
# 4 Cluster



Classification Report:

	precision	recall	f1-score	support
Tidak	0.71	0.91	0.80	500
	0.66	0.31	0.42	268
accuracy			0.70	768
macro avg	0.69	0.61	0.61	768
weighted avg	0.69	0.70	0.67	768

# 4 Cluster



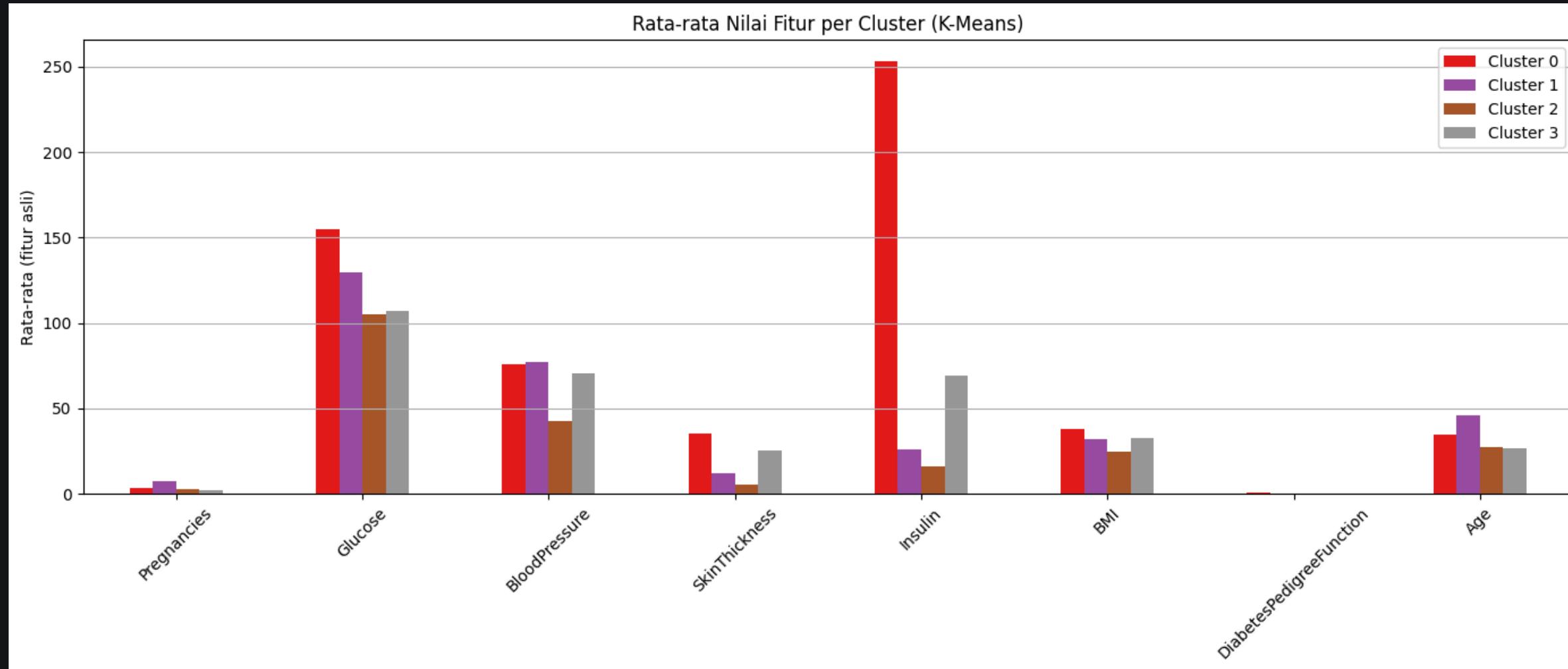
Classification Report:

	precision	recall	f1-score	support
Tidak Diabetes	0.71	0.91	0.80	500
Diabetes	0.66	0.31	0.42	268
accuracy			0.70	768
macro avg	0.69	0.61	0.61	768
weighted avg	0.69	0.70	0.67	768

KNN Classification Report:

	precision	recall	f1-score	support
Tidak Diabetes	0.71	0.94	0.81	500
Diabetes	0.70	0.28	0.40	268
accuracy			0.71	768
macro avg	0.71	0.61	0.61	768
weighted avg	0.71	0.71	0.67	768

# 4 Cluster



Classification Report:				
	precision	recall	f1-score	support
Tidak Diabetes	0.71	0.91	0.80	500
Diabetes	0.66	0.31	0.42	268
accuracy			0.70	768
macro avg	0.69	0.61	0.61	768
weighted avg	0.69	0.70	0.67	768

- Precision Tinggi (0.66) untuk "Diabetes" → Artinya: Saat K-Means bilang seseorang "diabetes", 66% benar-benar diabetes → cukup bagus.
- Tapi Recall Rendah (0.31) → Artinya: Dari semua pasien yang benar-benar diabetes, hanya 31% yang berhasil ditemukan oleh K-Means.
- Banyak pasien diabetes tidak terdeteksi oleh model.

# 4 Cluster Analysis

## Fitur Diabetes Tidak Terpisah Secara Geo

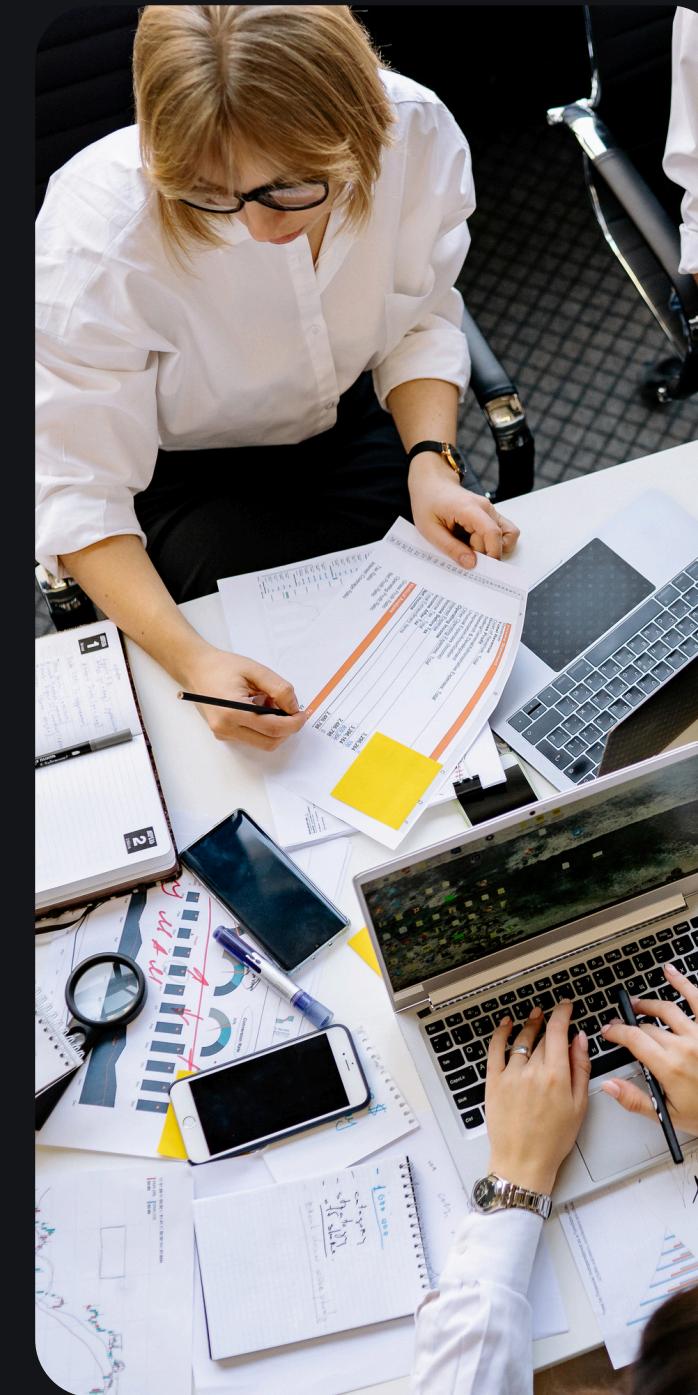
- K-means hanya bisa bekerja kalau pasien diabetes dan tidak diabetes berkelompok secara jelas di ruang fitur
- Pasien Diabetes tersebar di berbagai cluster karena variasi fitur yang tinggi
- Model sulit membentuk batas jelas antara dua golongan

## Ada Subgroup dalam Diabetes

- Insulin rendah
  - BMI rendah
  - Glucose sedang
- Model mengenali tipe diabetes yang khas, tapi gagal tangkap untuk “tidak biasa”

## Cluster 0 mungkin tidak hanya 1 penyakit

- Banyak orang diabetes kabur ke cluster lain
- Bisa terdapat campuran dari beberapa penyakit



# conclusion

- Model hanya bisa melakukan deteksi diabetes terhadap fitur yang sangat signifikan
- Terdapat fitur-fitur dominan yang menjadi penentu model memberikan label diabetes
- Model dapat memprediksi dengan baik orang yang tidak diabetes, tetapi buruk dalam memprediksi orang yang diabetes
- K-Means terbaik dengan 4 Cluster

Fitur	Penjelasan Singkat	Bisa Terkait Dengan Penyakit Lain
Pregnancies	Jumlah kehamilan	Sindrom metabolik, komplikasi kehamilan
Glucose	Gula darah	<b>Diabetes</b> , hipoglikemia, insulinoma
BloodPressure	Tekanan darah	Hipertensi, penyakit jantung, gagal ginjal
SkinThickness	Ketebalan kulit lemak	Gizi buruk, obesitas
Insulin	Kadar insulin darah	<b>Diabetes</b> , resistensi insulin, PCOS
BMI	Indeks massa tubuh	Obesitas, malnutrisi, jantung, liver
DiabetesPedigreeFunction	Skor genetik keluarga	Diabetes, tapi tidak pasti
Age	Usia	Faktor risiko umum, semua penyakit

thank you