Clustering and Prediction

Diabetes Prediction from Pima_Indians_Diabetes Dataset

Angella Ananta Batubara - 71476

Overview

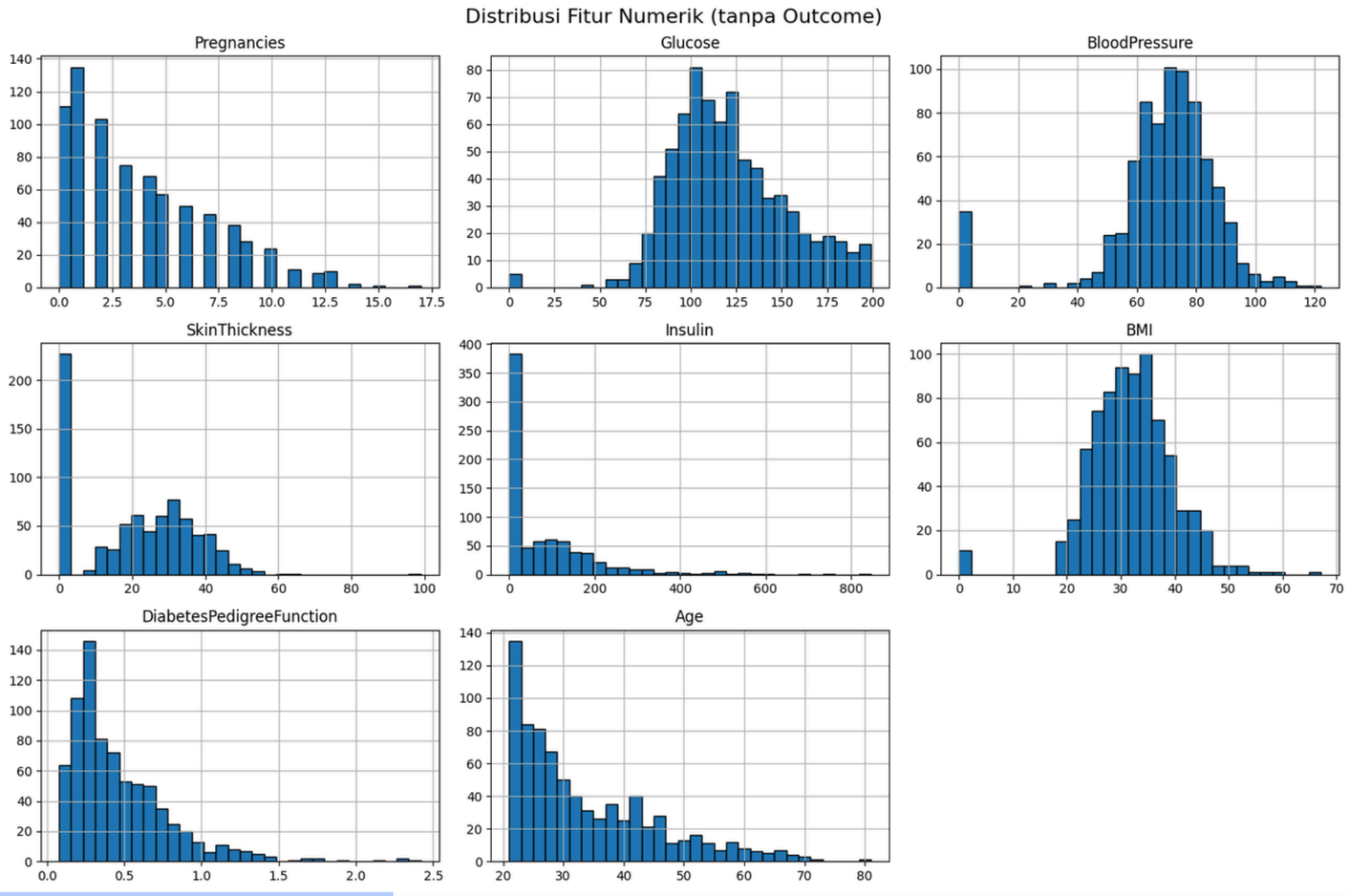
Tujuan

Mengetahui sebaik apa model K-Means dalam memprediksi seseorang diabetes atau tidak

Memberikan gambaran bagaimana K-Means memprediksi data

Memberikan saran dari hasil analisis yang ada

About Data

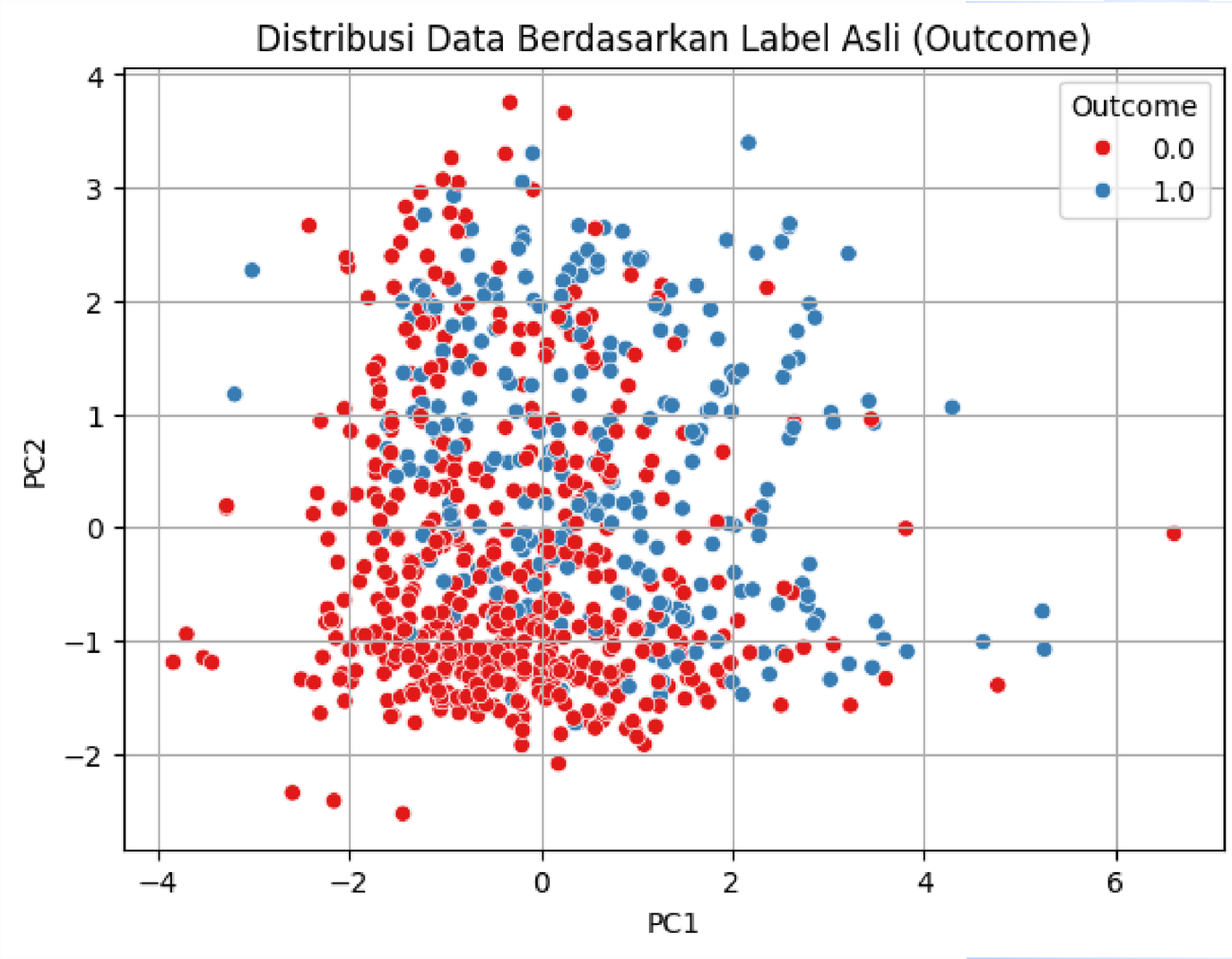


Fitur	Penjelasan
Pregnancies	Jumlah kehamilan yang pernah dialami (angka bulat ≥ 0)
Glucose	Kadar glukosa plasma saat tes toleransi glukosa 2 jam (mg/dL)
BloodPressure	Tekanan darah diastolik (mm Hg)
SkinThickness	Ketebalan lipatan kulit tricep (mm) – indikator lemak tubuh
Insulin	Kadar insulin serum 2 jam setelah makan (mu U/ml)
BMI	Body Mass Index (kg/m²) = berat / tinggi²
DiabetesPedigreeFunction	Indeks risiko berdasarkan riwayat keluarga (nilai antara 0 dan 2+)
Age	Usia pasien (tahun)
Outcome	Label diagnosis:
	0 = Tidak diabetes
	1 = Diabetes

Hubungan Antar Variabel Menurut T-Test

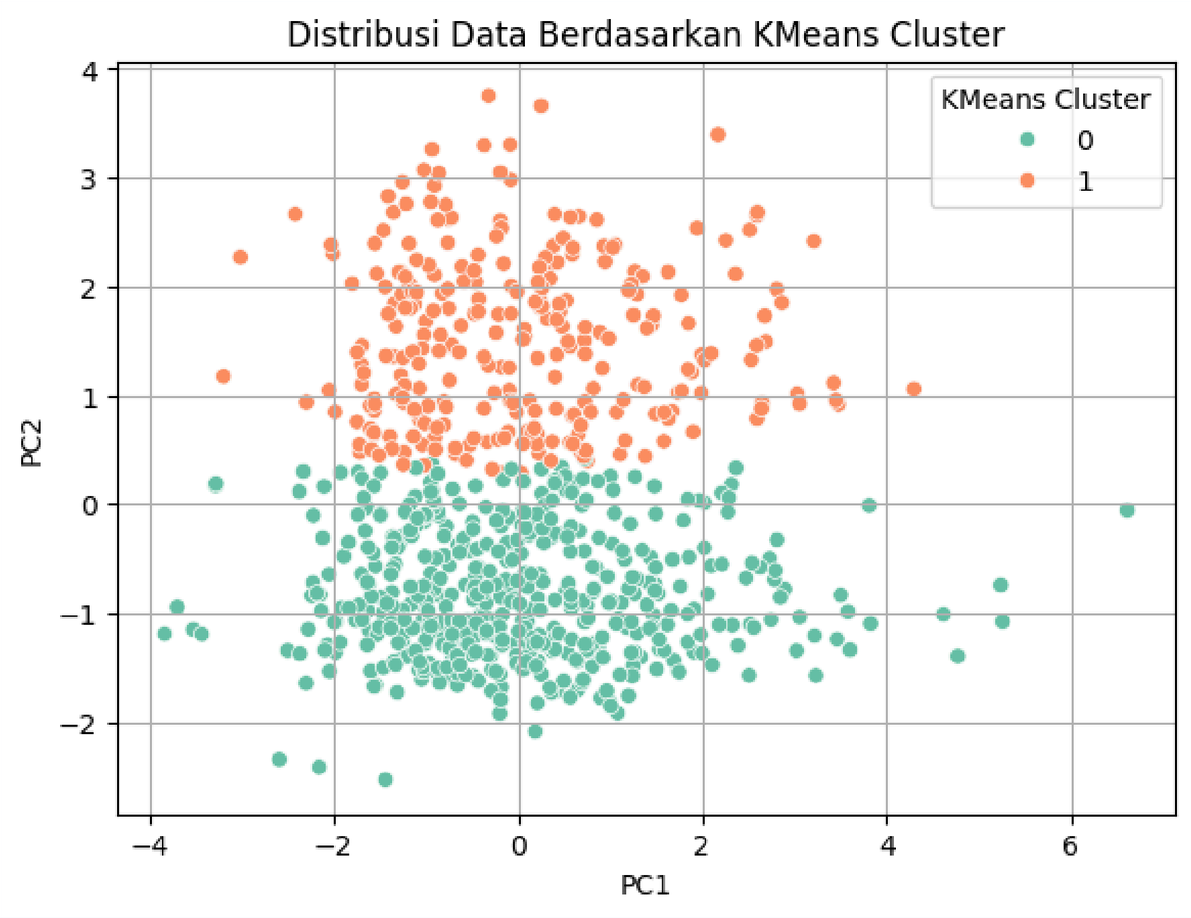
No.	Fitur	T-Statistic	P-Value	Signifikan (<0.05)
1	Glucose	-13.7515	0.000000000000000(	✅ Ya
2	BMI	-8.6193	0.000000000000000(	✅ Ya
3	Age	-6.9207	0.000000000012	✅ Ya
4	Pregnancies	-5.907	0.00000000682	✅ Ya
5	DiabetesPedigreeFunction	-4.5768	0.0000061	✅ Ya
6	Insulin	-3.3009	0.00105	✅ Ya
7	SkinThickness	-1.9706	0.0494	✅ Ya
8	BloodPressure	-1.7131	0.0874	❌ Tidak

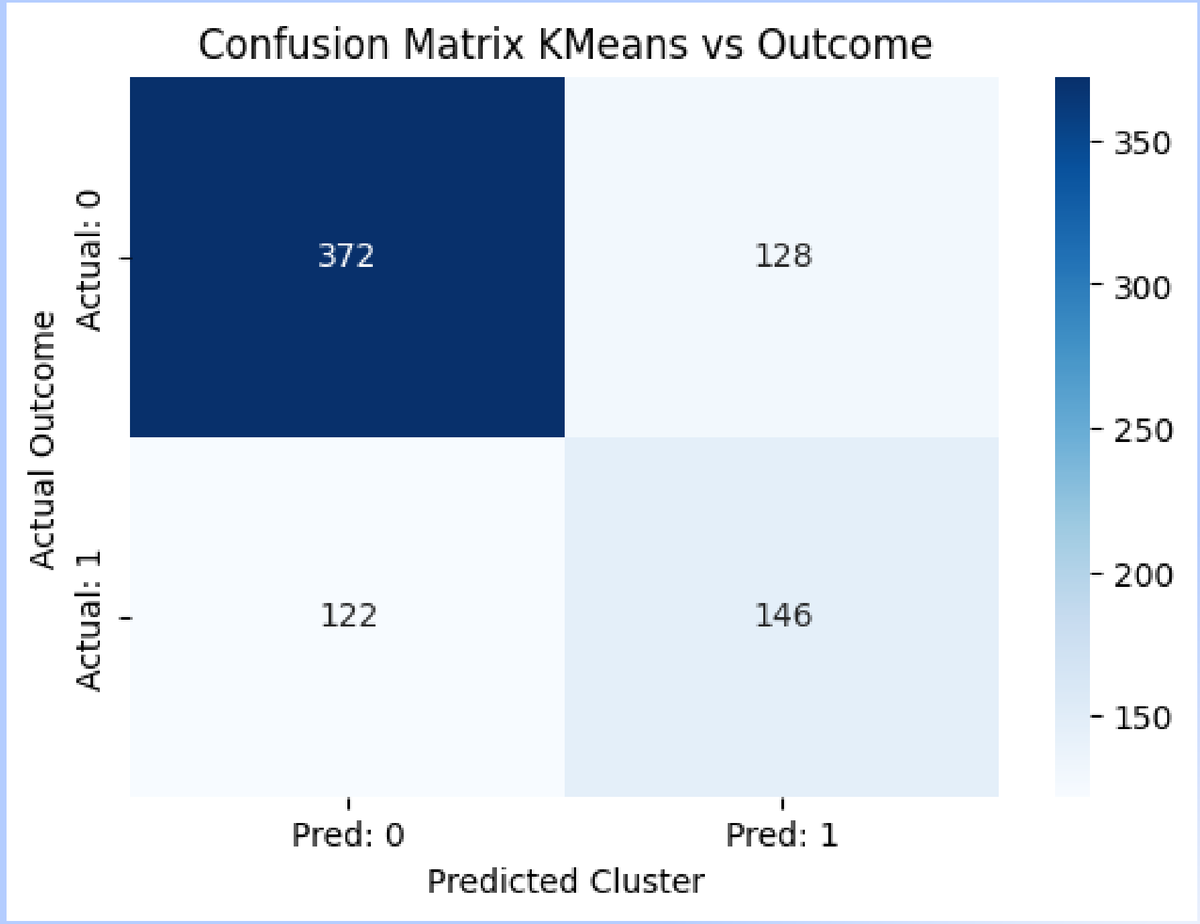
Sebaran Data Asli



0 = Tidak Diabetes
1 = Diabetes

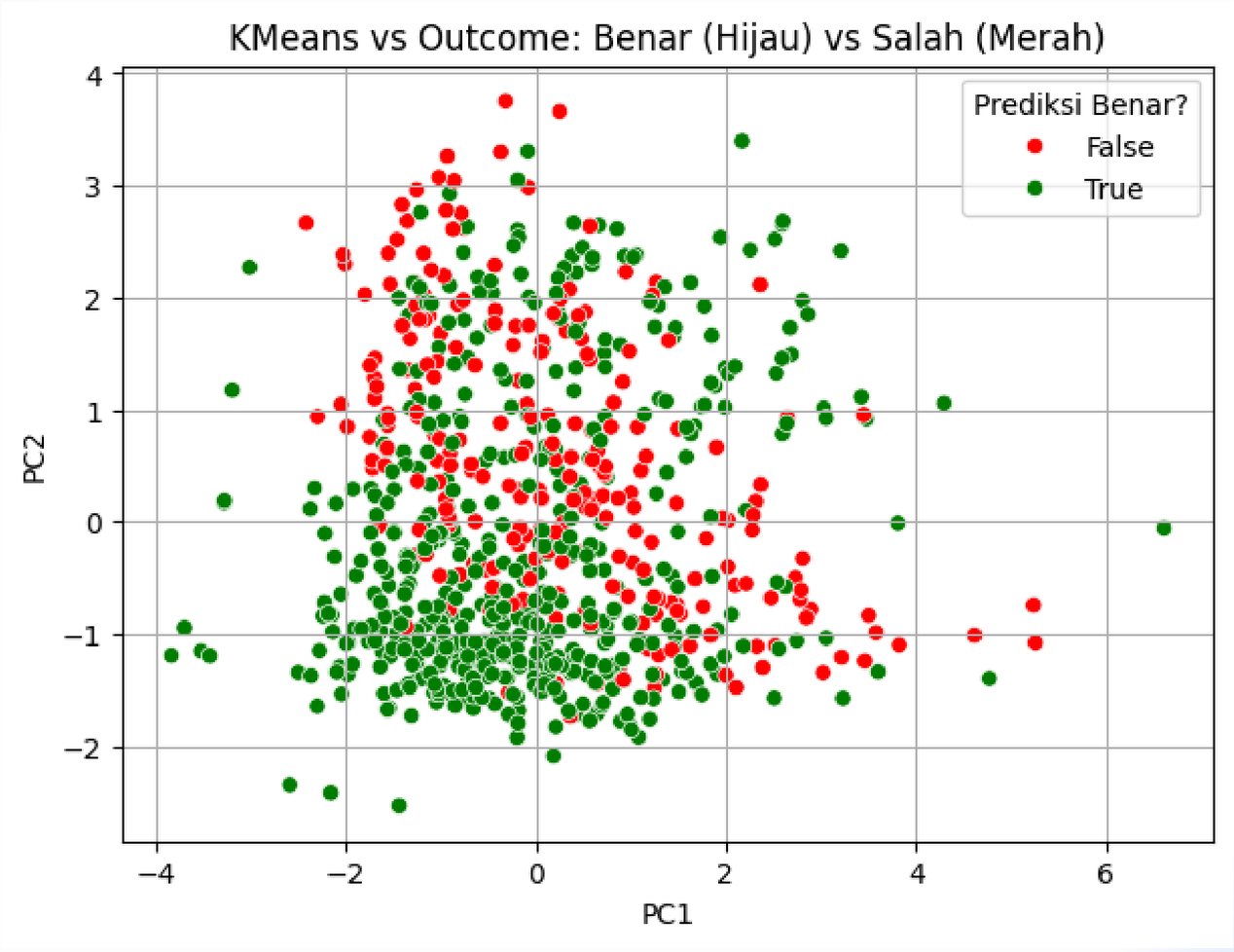
Sebaran Data Hasil Prediksi K-Means



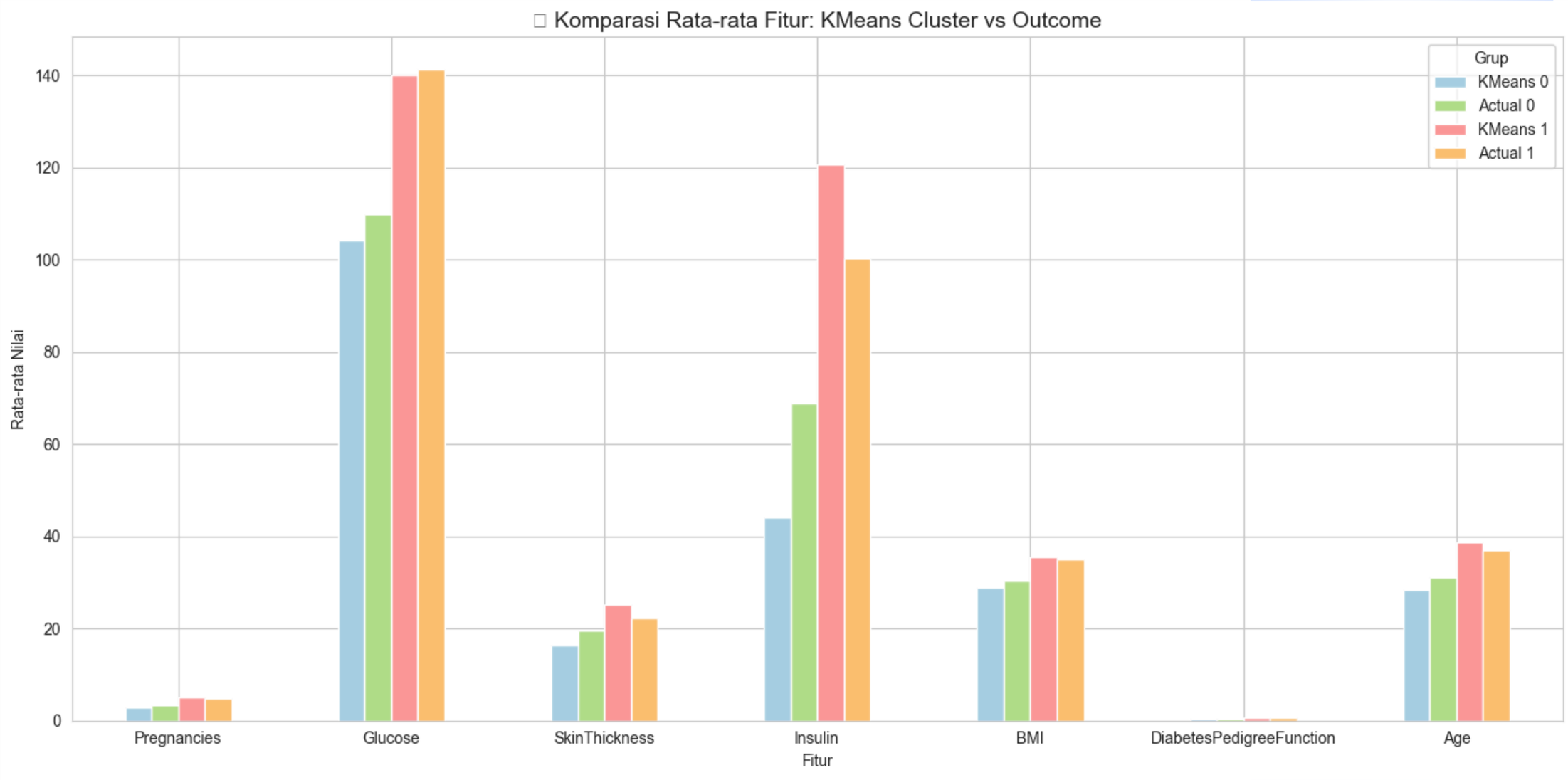


Akurasi : 0.6745

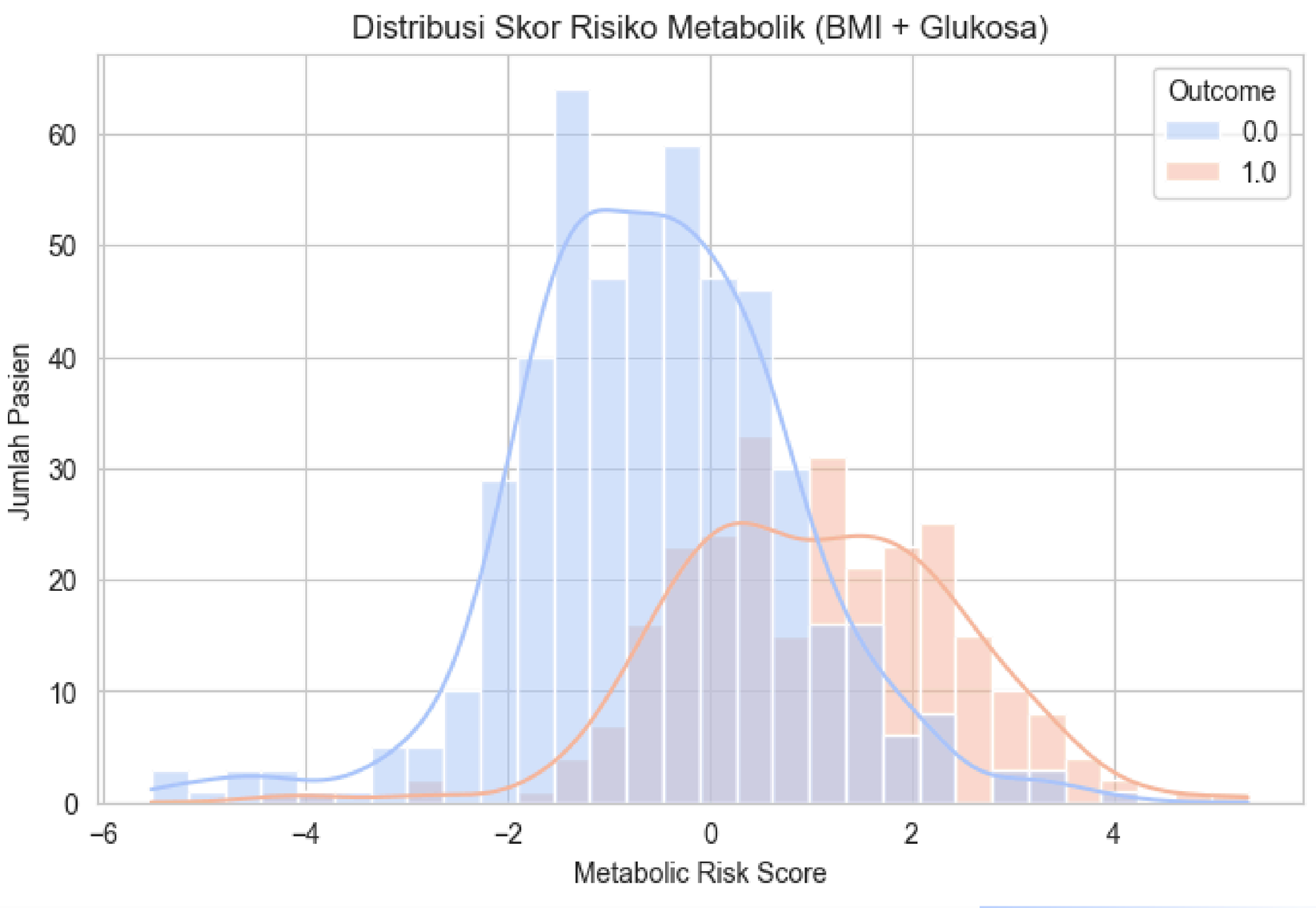
Sebaran Prediksi Benar dan Salah



Perbandingan Hasil Akhir

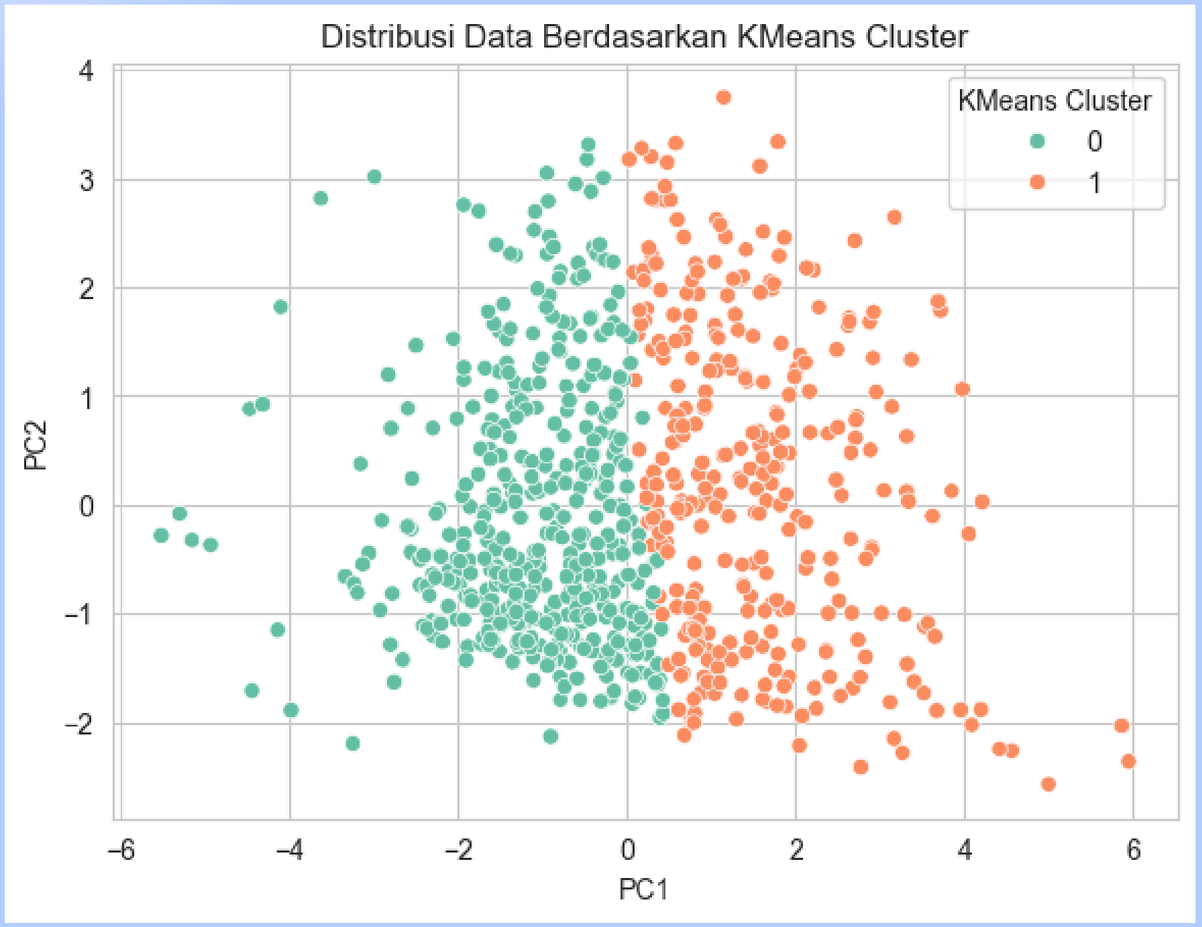


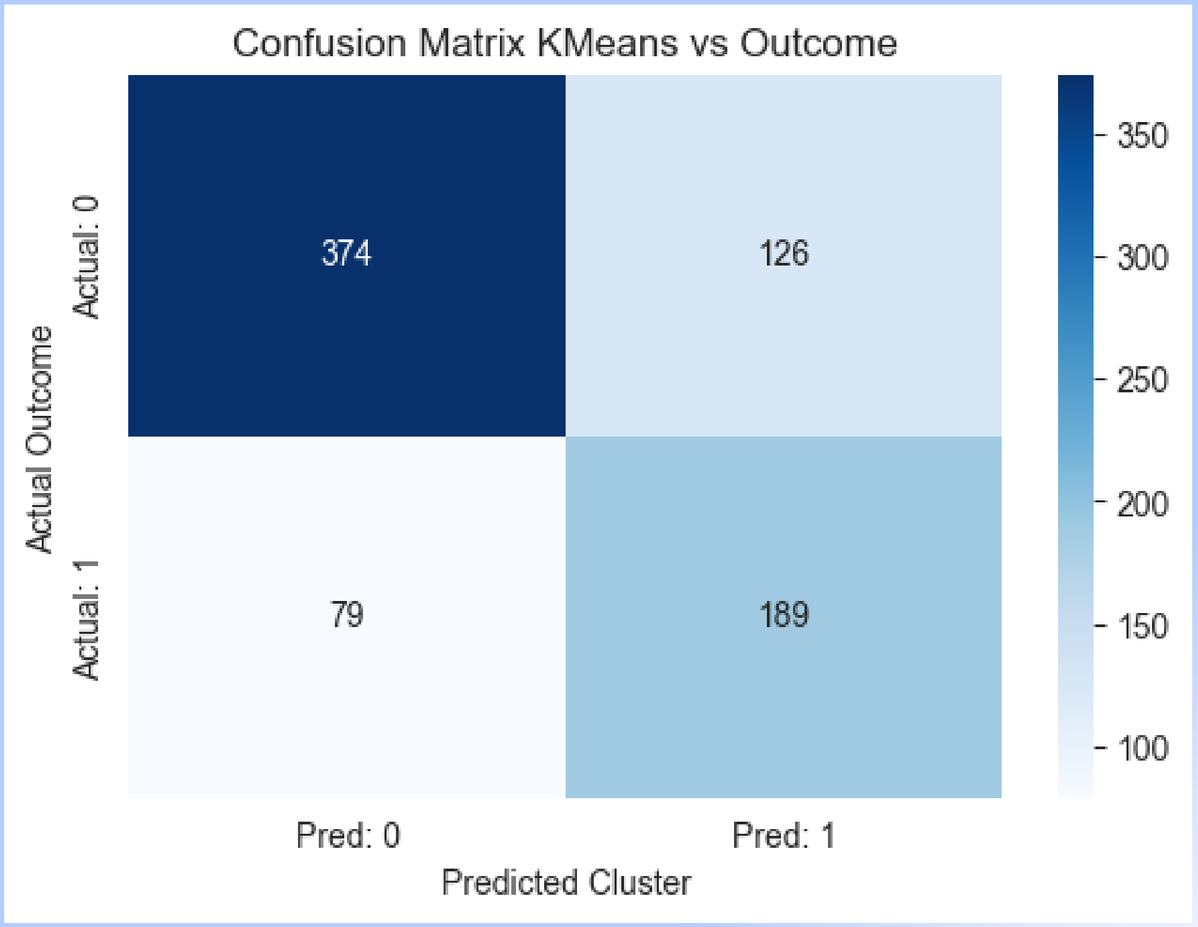
Pembuatan Variabel Baru



Skor Risiko Metabolik merupakan indikator gabungan yang dapat digunakan untuk menilai sejauh mana seseorang memiliki faktor risiko terhadap penyakit metabolik, salah satunya adalah **diabetes tipe 2**

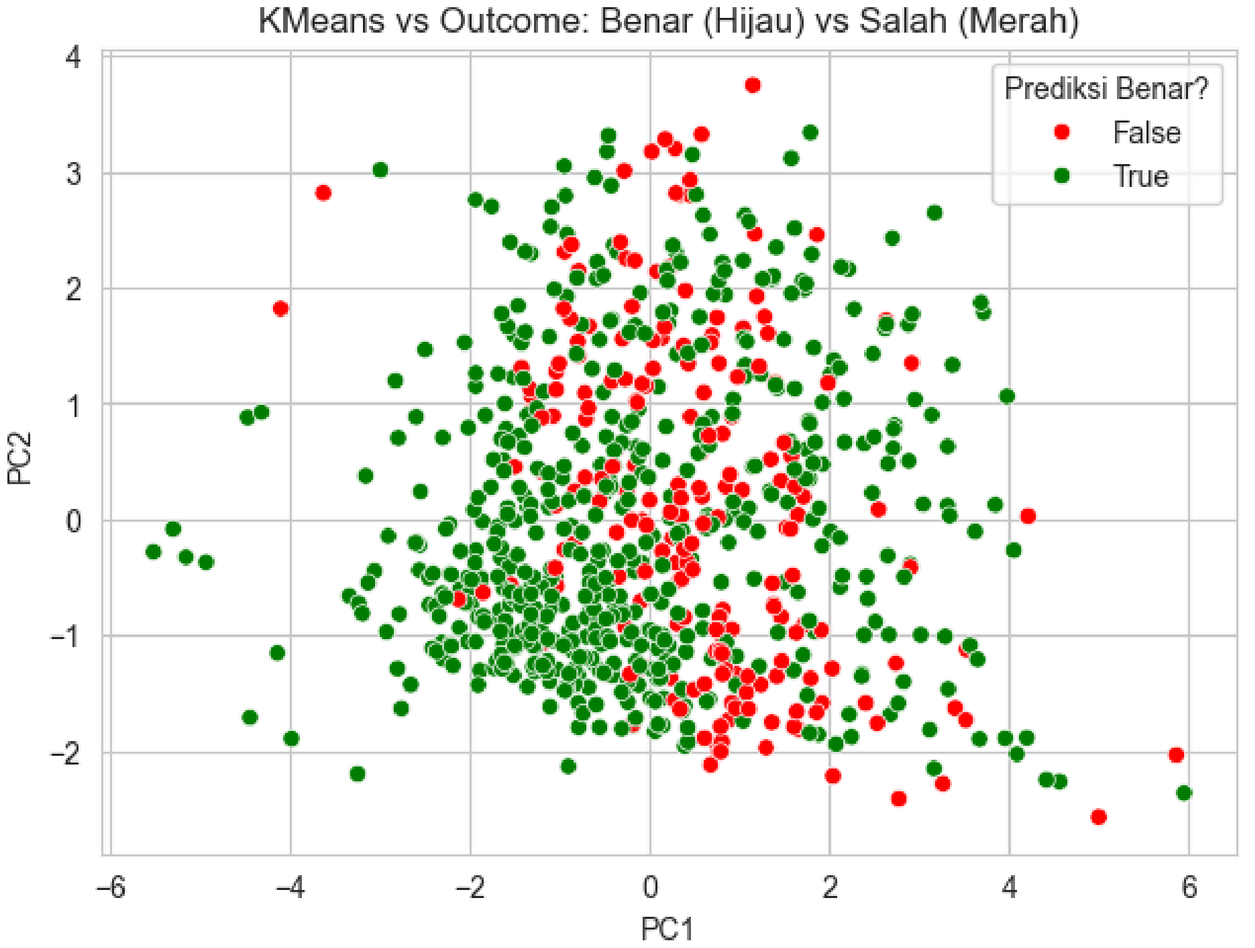
Sebaran Data Hasil Prediksi K-Means



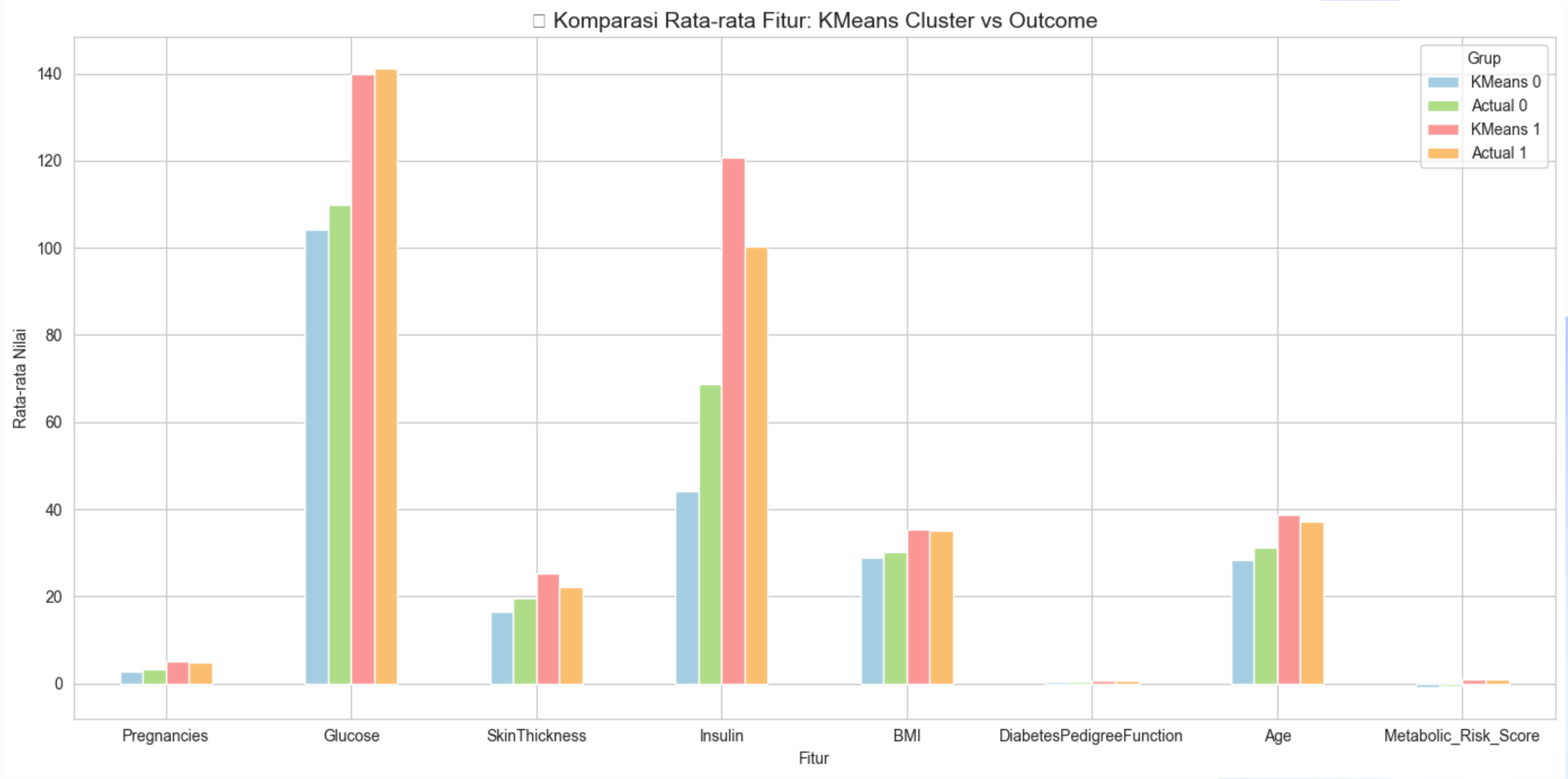


Akurasi : 0.7331

Sebaran Prediksi Benar dan Salah



Perbandingan Hasil Akhir



Kesimpulan

Data yang ada sulit untuk diprediksi atau dikelompokkan dengan baik karena sebaran yang terlalu bervariasi.

K-Means memberikan prediksi clustering yang cukup baik (akurasi 67%). Akurasi meningkat menjadi 73% ketika ada variabel baru yang ditambahkan.

Variabel yang ada di dalam data sudah dapat terprediksi dengan cukup baik.

Saran

Mencoba mencari tahu lebih jauh mana saja variabel yang bisa digabung untuk meningkatkan kebaikan model.

Menganalisis variabel insulin lebih lanjut mengingat gap prediksi dan aktual cukup besar.

Melihat dan mengevaluasi kembali kebaikan prosedur pengambilan data insulin.

Memperbanyak variabel yang dapat diambil melalui sampling secara langsung untuk meningkatkan hasil analisis.

Thank You