



DIABETES RISK MODELING

Untuk mengidentifikasi individu yang memiliki kemungkinan tinggi terkena diabetes

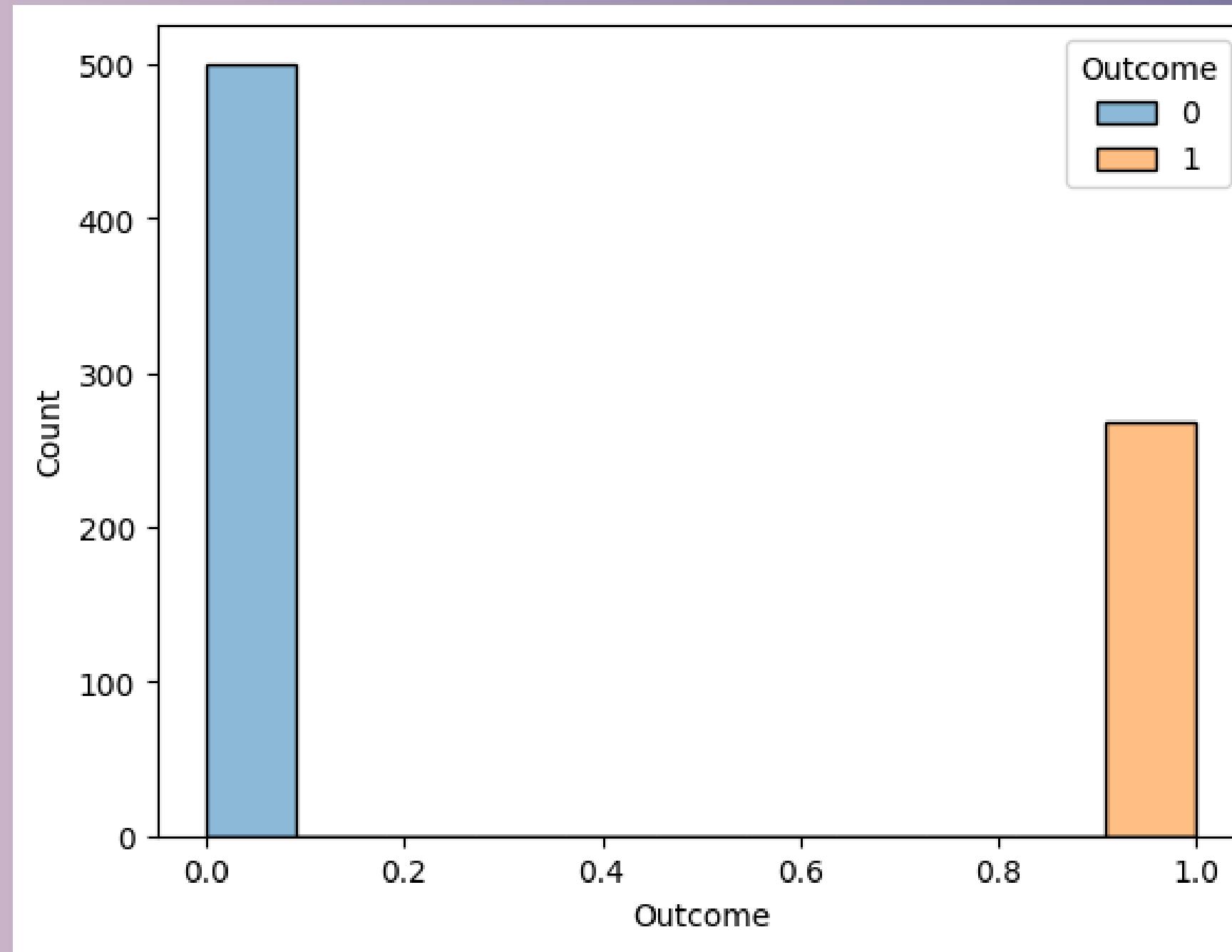
DETAIL DATA

NO	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0	0.232	54	1

TUJUAN

- Menemukan karakteristik individu yang berisiko terkena diabetes
- Mengetahui hubungan antar variabel penting (glukosa, BMI, usia, dll.)
- Memprediksi siapa yang kemungkinan terkena diabetes menggunakan model machine learning

EDA (UNIVARIATE ANALYSIS)



positif diabetes (Outcome = 1)

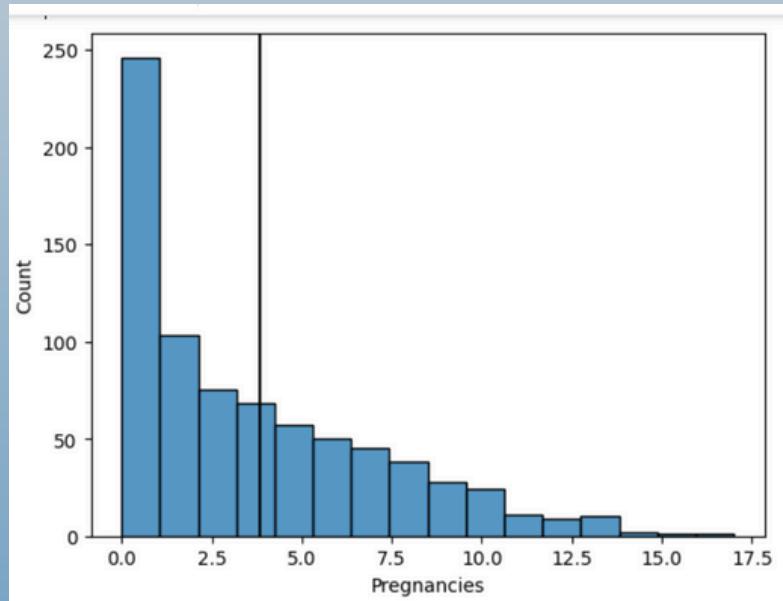
Negatif (Outcome = 0)

Hasil:

- Sekitar **65% tidak diabetes**
- Sekitar **35% diabetes**

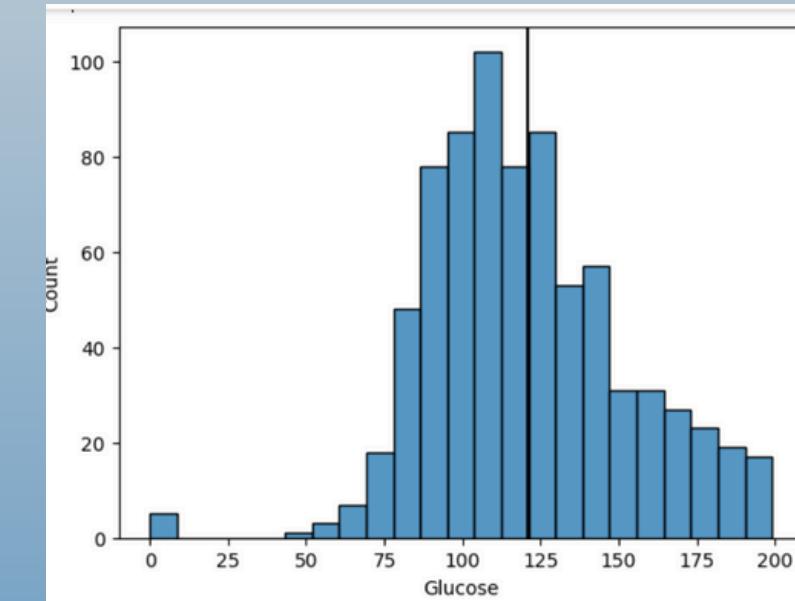
EDA (UNIVARIATE ANALYSIS)

PREGNANCIES



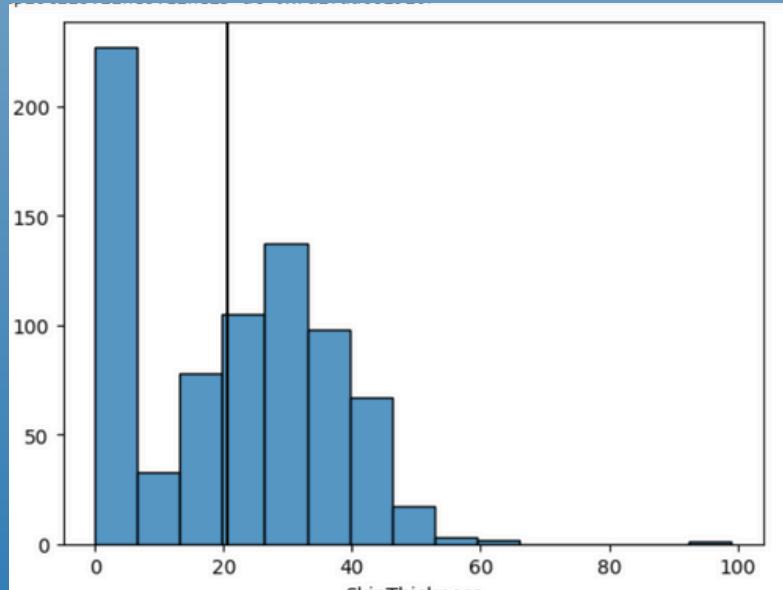
Nilai **0–6** kehamilan **paling sering** muncul.

GLUCOSE



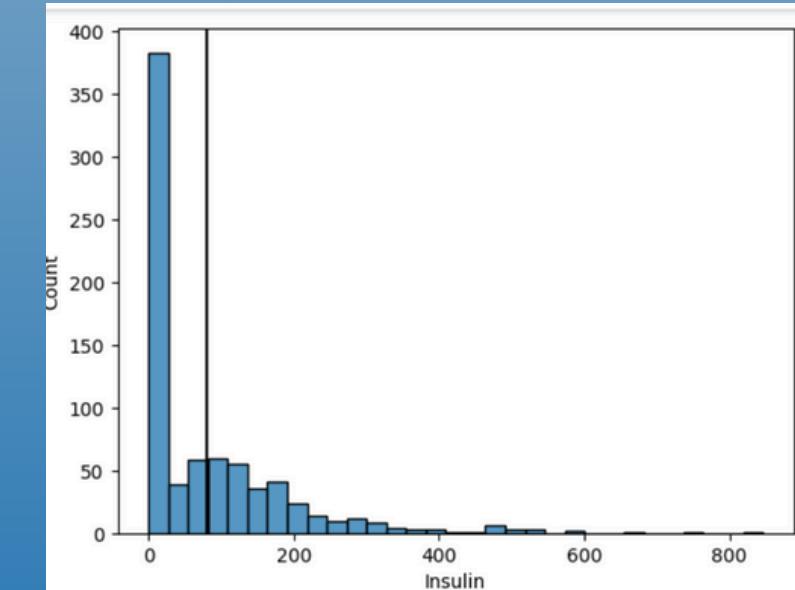
Banyak pasien memiliki glukosa **>100**

SKINTHICKNESS



- Banyak data bernilai 0, menandakan **missing value** atau tidak diukur
- Data tersebar antara **0–50 mm**

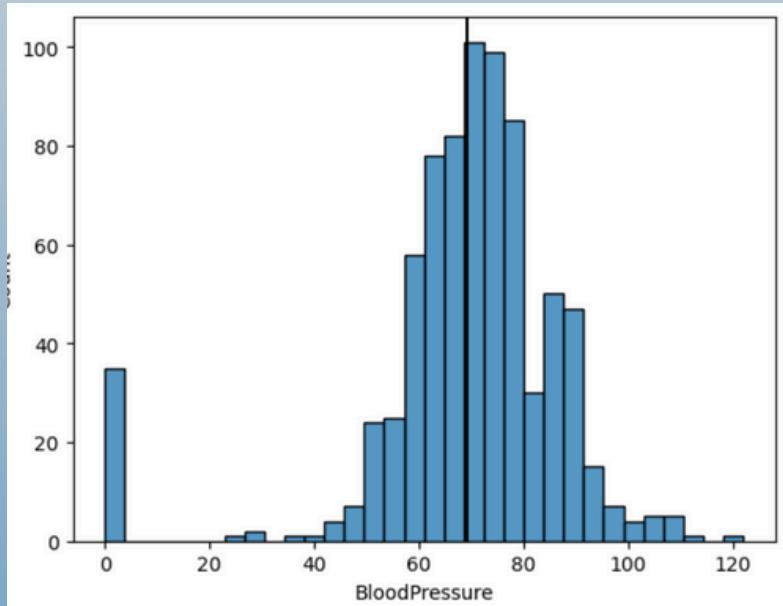
INSULIN



Banyak pasien dengan **nilai 0** insulin, artinya **kemungkinan tidak diuji**

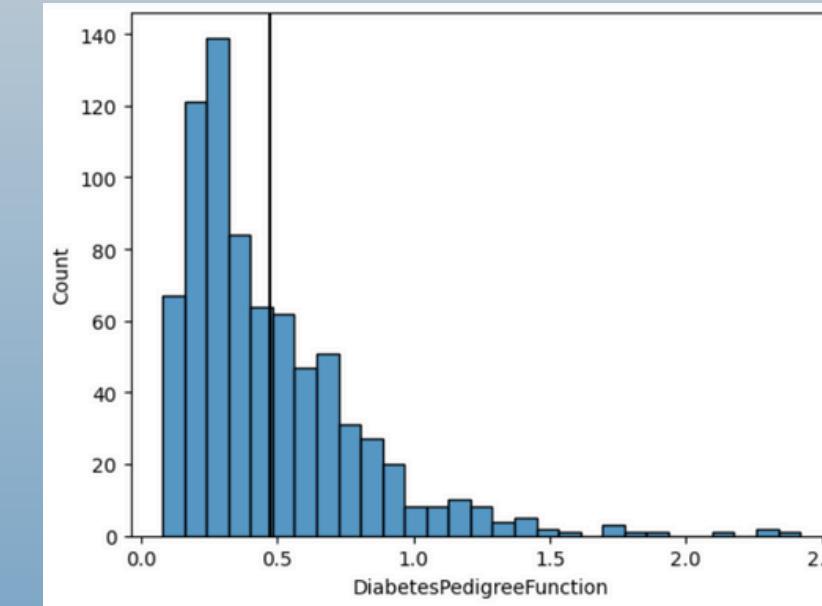
EDA (UNIVARIATE ANALYSIS)

BLOODPRESSURE



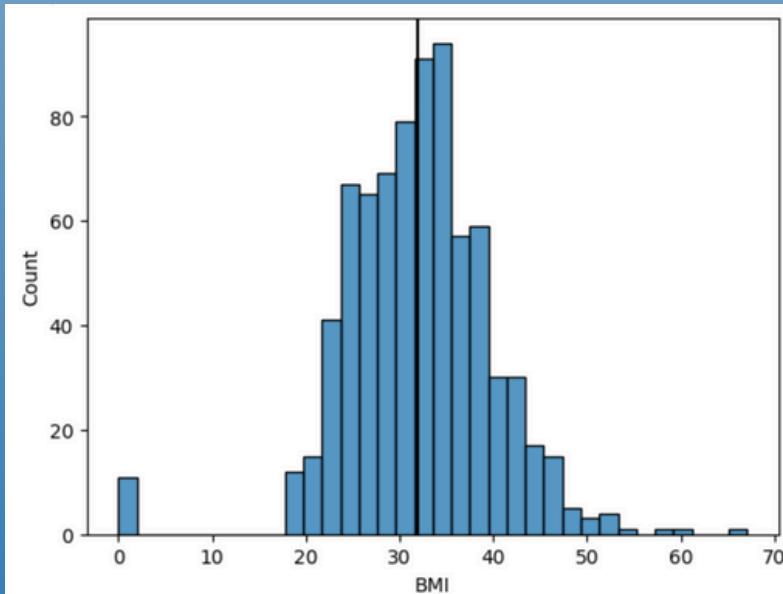
Tekanan darah di kisaran **60–80** paling umum

DIABETES PEDIGREE FUNCTION



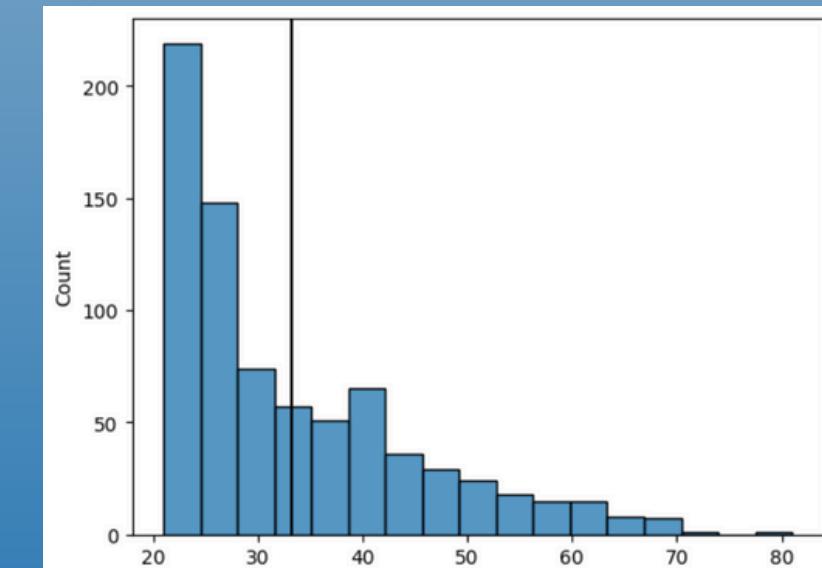
Semakin **tinggi nilainya** → semakin besar **faktor genetik** dalam risiko diabetes

BMI



Sebagian besar pasien memiliki **BMI** antara **30–40**

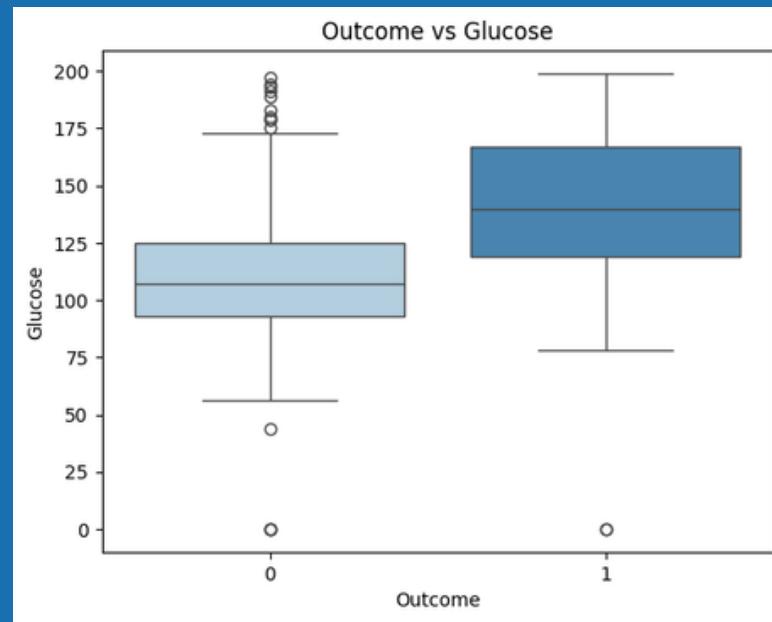
AGE



- Usia **majoritas** pasien antara **20–50 tahun**
- Beberapa pasien usia **di atas 60** → risiko makin tinggi

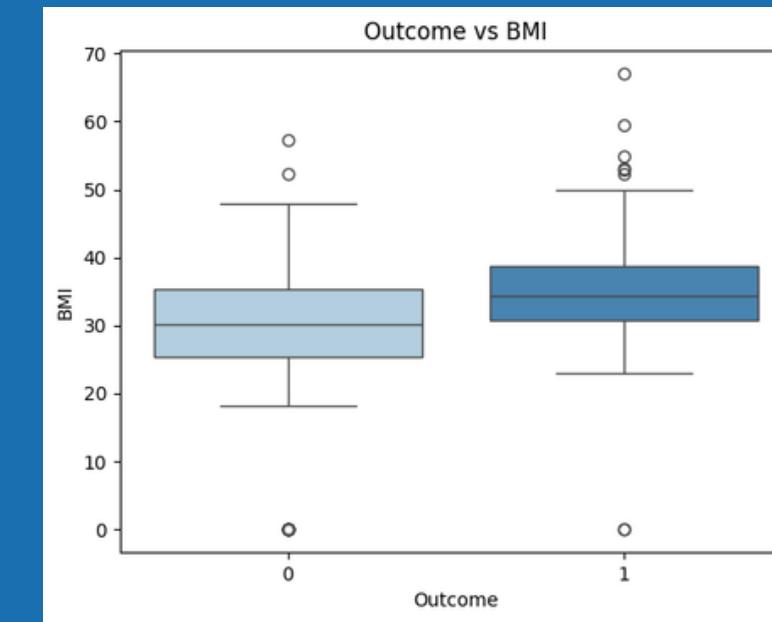
EDA (BIVARIATE DATA ANALYSIS)

OUTCOME VS GLUCOSE



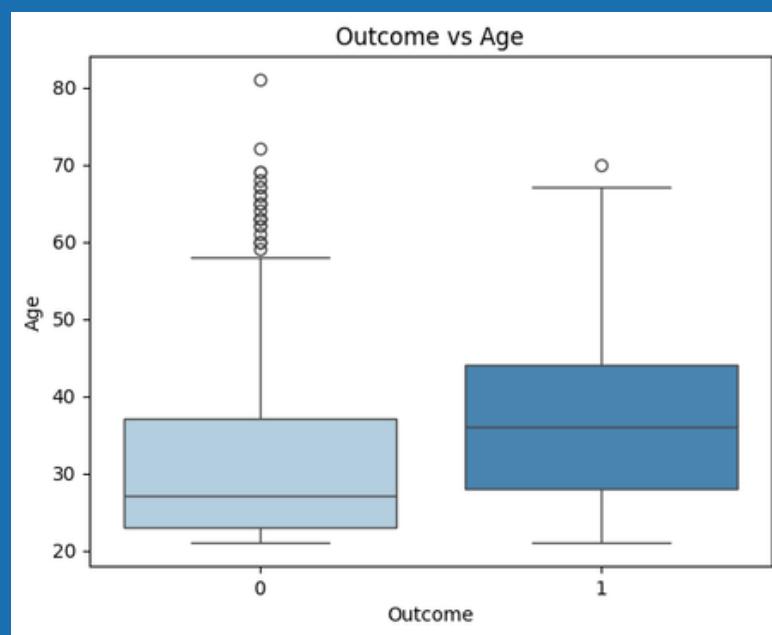
Median glukosa pasien **diabetes > 130**, sedangkan **non-diabetes < 110**.

OUTCOME VS BMI



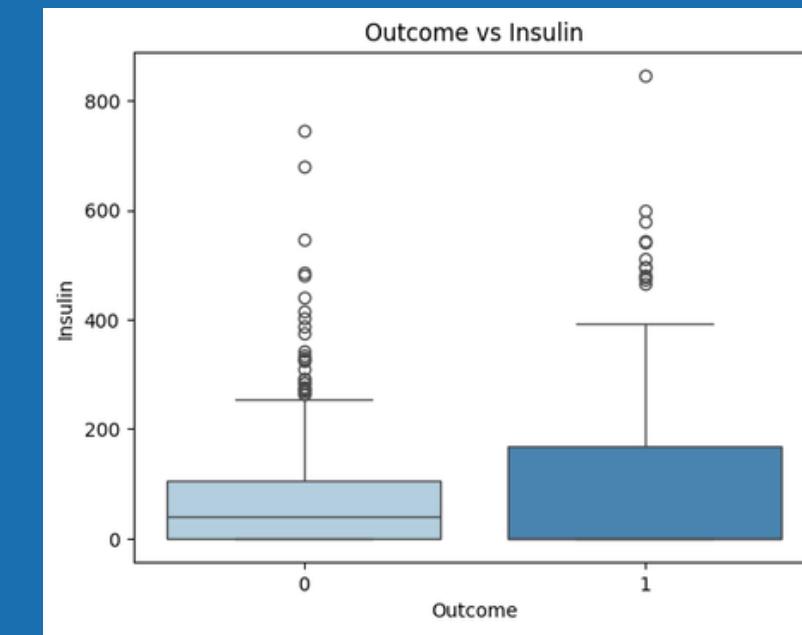
Median BMI pasien **diabetes > 35**, sedangkan **non-diabetes ~30**.

OUTCOME VS AGE



Median usia pasien diabetes **> 35 tahun**.

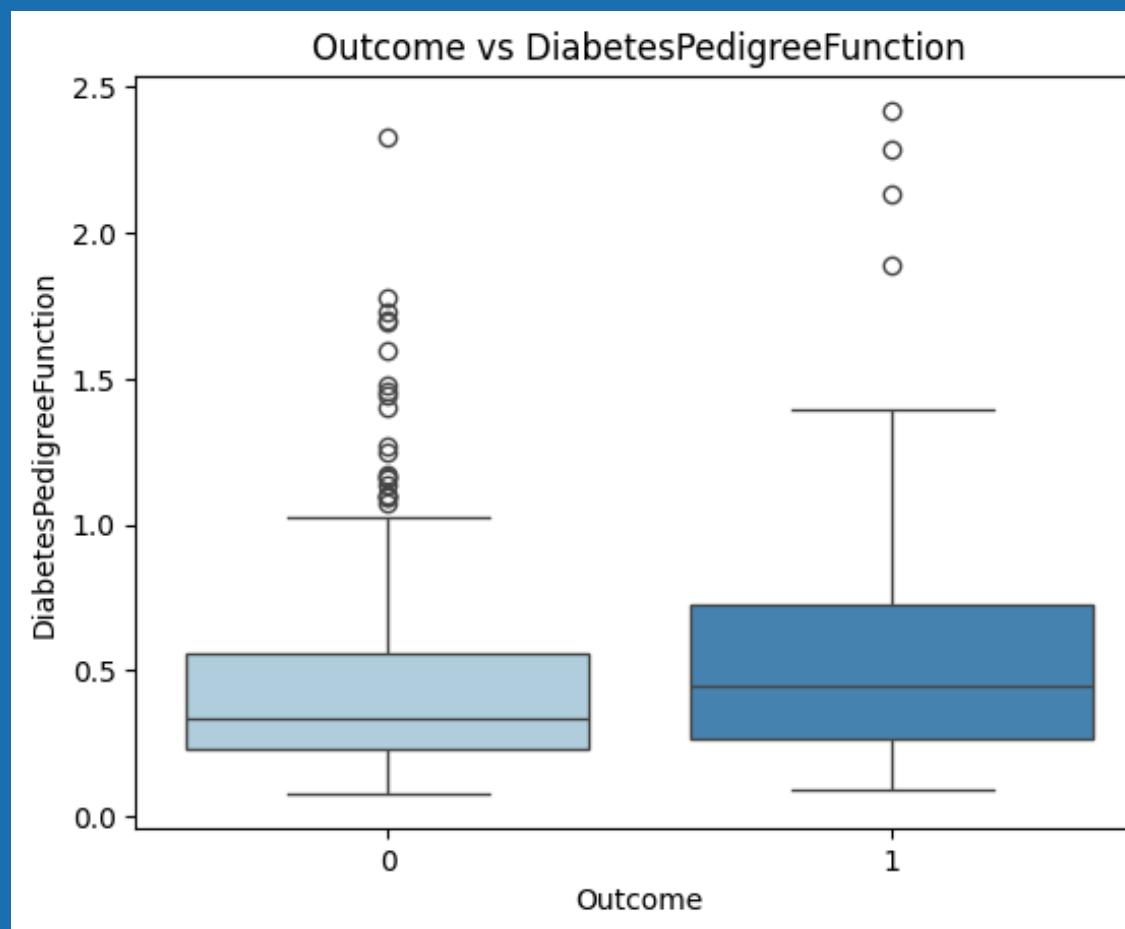
OUTCOME VS INSULIN



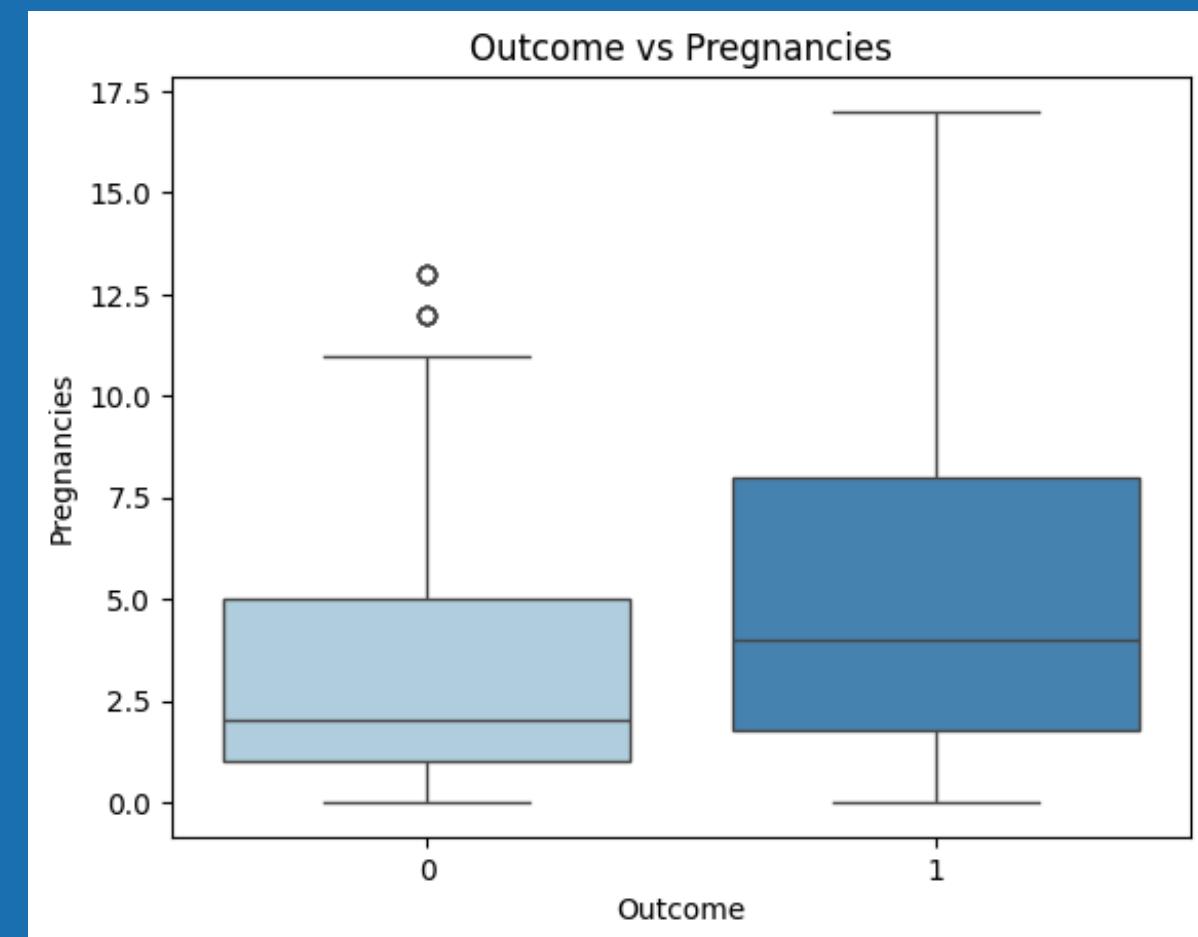
- Sebaran insulin sangat **bervariasi**
- Tidak ada pola yang terlalu jelas karena banyak **data insulin = 0**

EDA (BIVARIATE DATA ANALYSIS)

OUTCOME VS DIABETESPEDIGREEFUNCTION



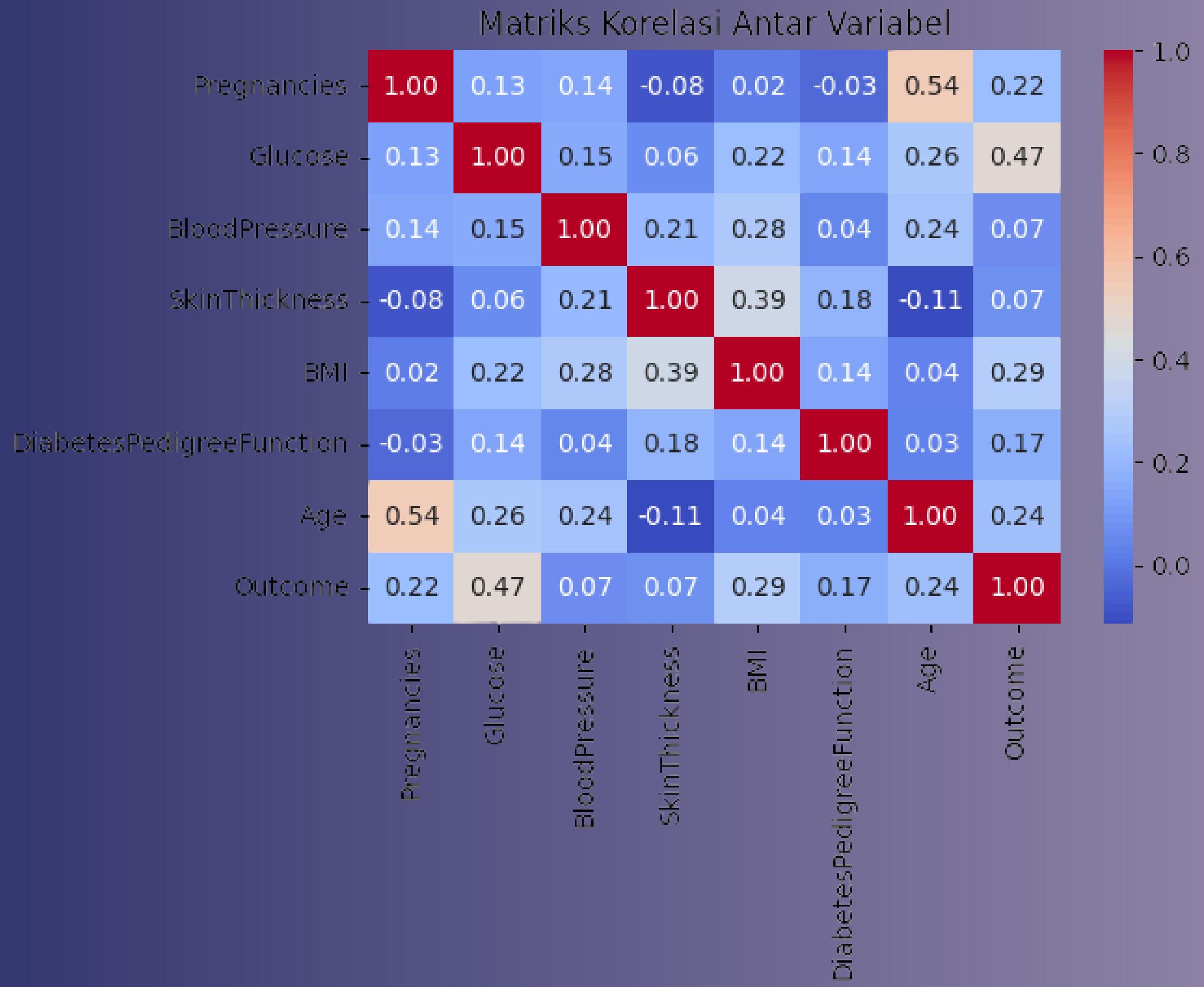
OUTCOME VS PRGNANCIES



DPF pasien diabetes **sedikit lebih tinggi**

Pasien diabetes memiliki **jumlah kehamilan rata-rata lebih tinggi**

CORELLATION HEATMAP



Keterangan:

X1: Pregnancies

X2: Glucose

X3: Blood Pressure

X4: Skin Thickness

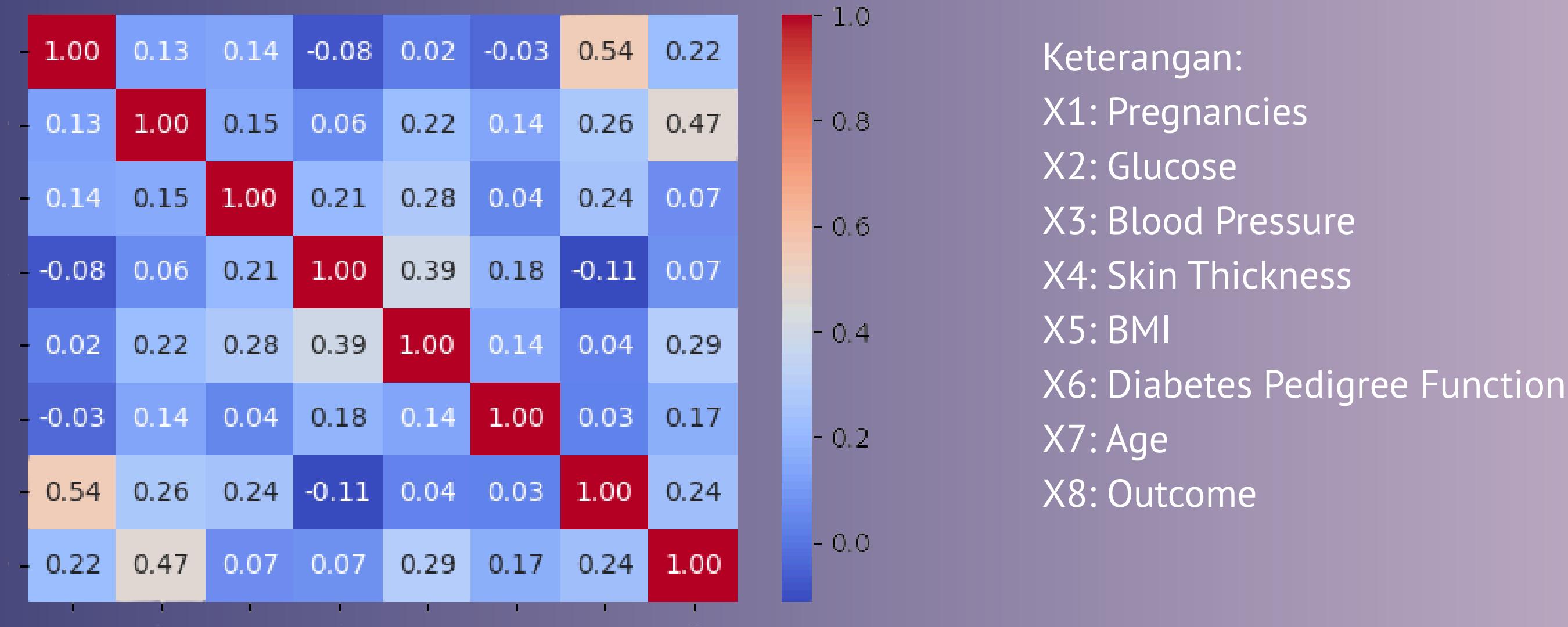
X5: BMI

X6: Diabetes Pedigree Function

X7: Age

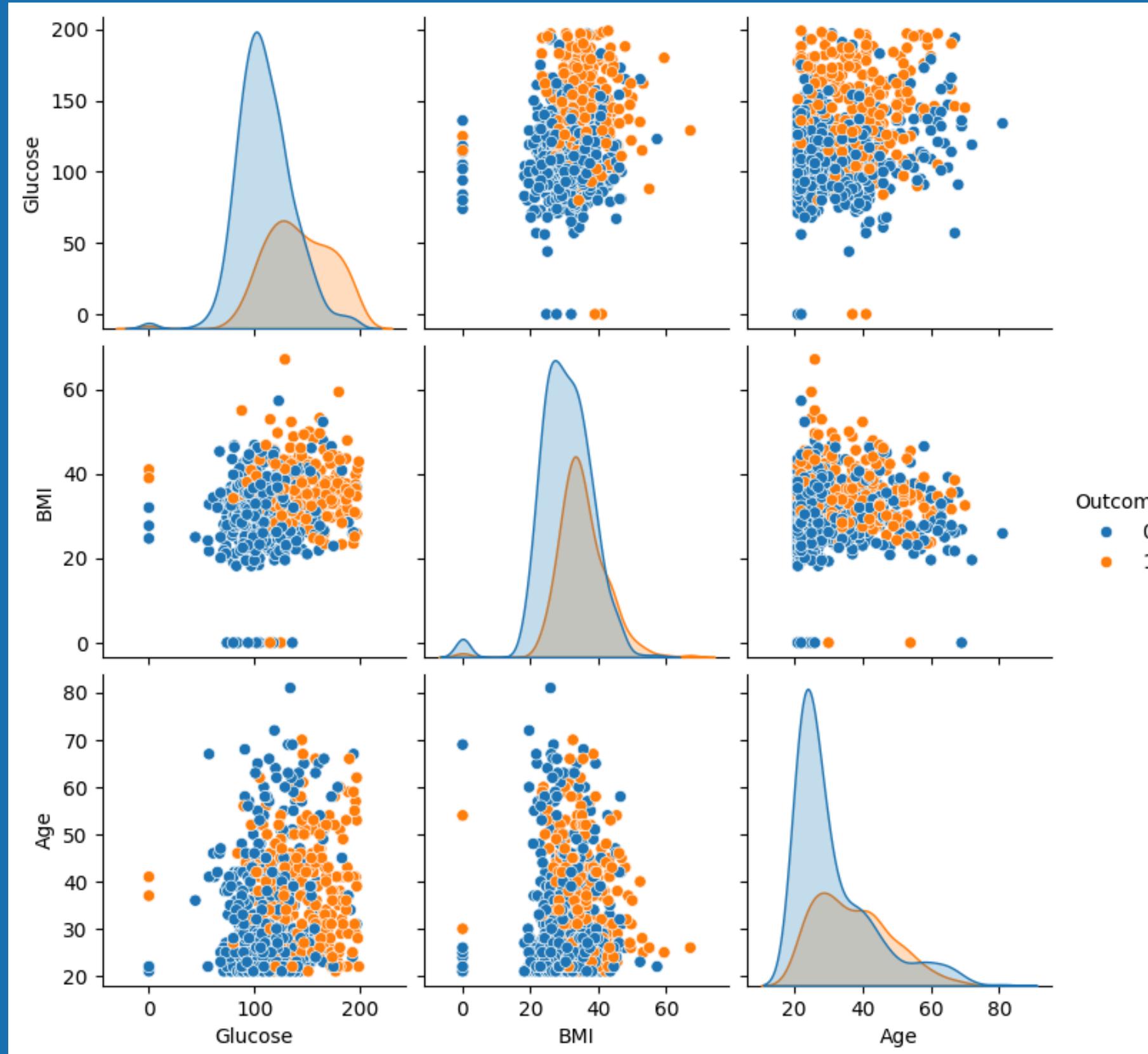
X8: Outcome

CORELLATION HEATMAP



- Korelasi positif tinggi antara **glucose** dengan **outcome** ($R = 0.47$)
- Korelasi positif sedang antara **BMI** dengan **outcome** ($R = 0.29$)
- Korelasi positif antara **Age** dengan **outcome** ($R = 0.24$)
- Korelasi positif antara **Pregnancies** dengan **outcome** ($R = 0.22$)
- Korelasi positif antara **Diabetes Pedigree Function** dengan **outcome** ($R = 0.17$)
- Korelasi positif lemah antara **Skinthickness**, **Bloodpressure** dengan **outcome** (<0.15)

EDA (MULTIVARIATE DATA ANALYSIS)



Merah: Outcome = 1 (**diabetes**)

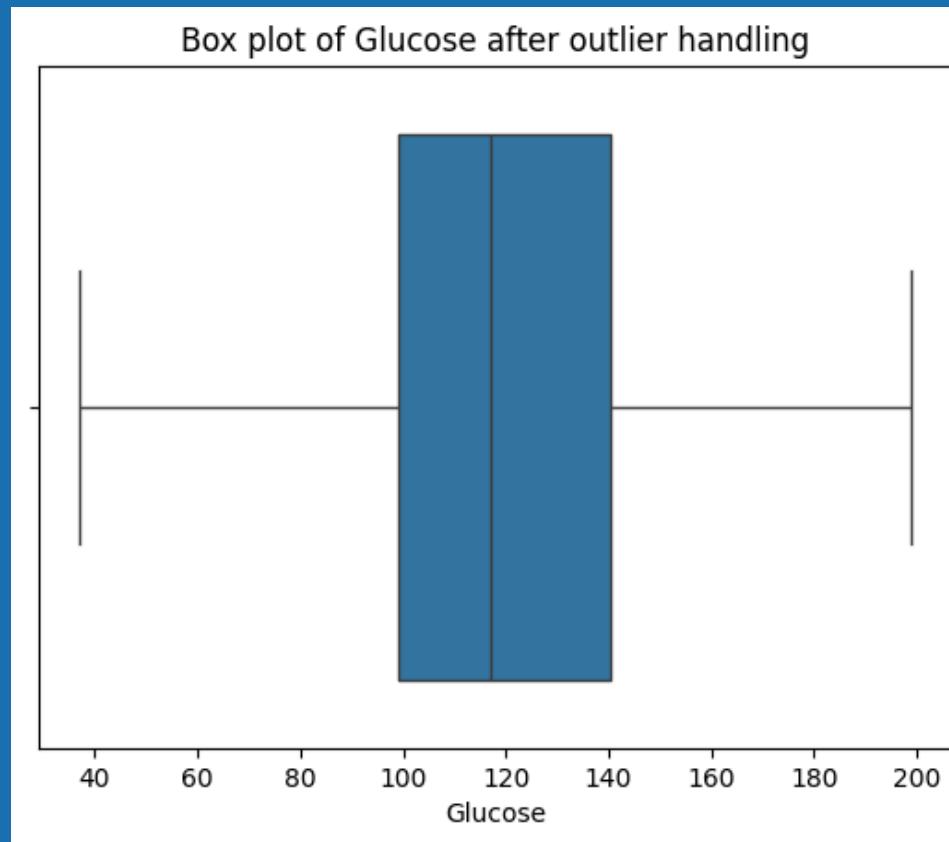
Biru: Outcome = 0 (**tidak diabetes**)

Kelompok pasien diabetes (merah) berkumpul di area dengan **Glucose tinggi dan BMI tinggi**. Pola distribusi pasien diabetes terlihat jelas saat:

- Glucose vs BMI
- Glucose vs Age
- Glucose vs Insulin

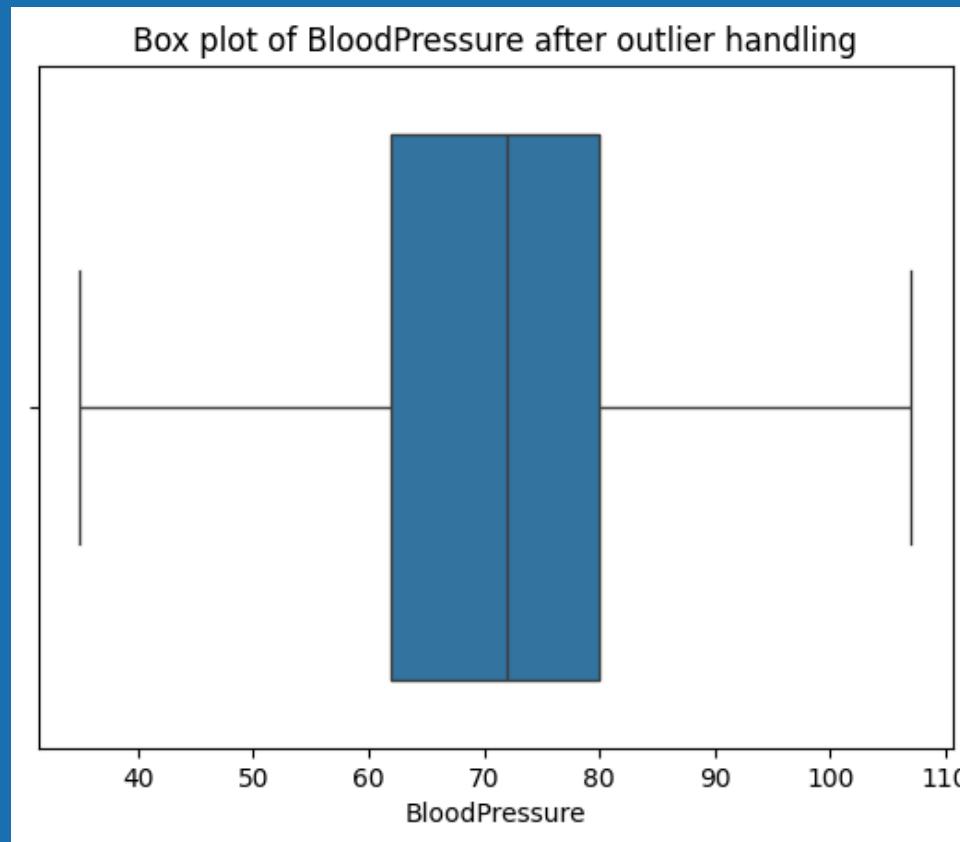
Dimana ketika **Glucose tinggi + BMI tinggi sangat identik dengan risiko diabetes**

OUTLIER HANDLING (NUMERICAL DATA)



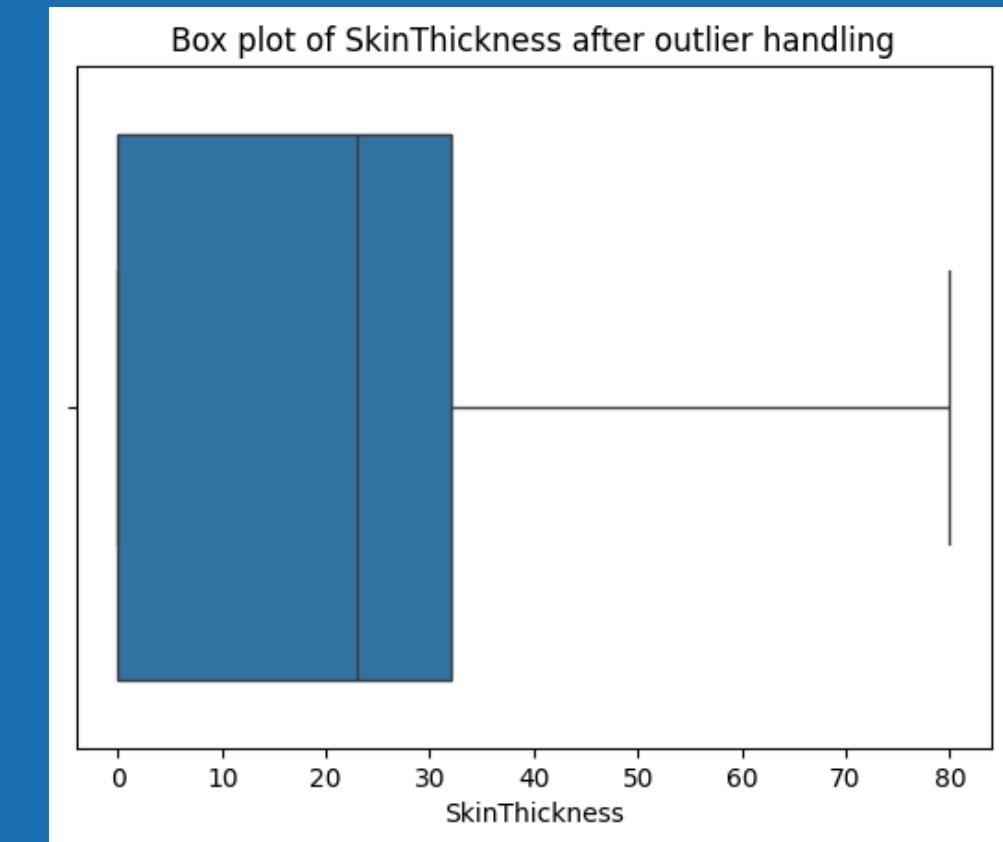
Glucose

- Sebelum: Terlihat **banyak nilai 0** dan beberapa *outlier* di atas 180.
- Sesudah: **Nilai 0** sudah **hilang** (diganti median), **distribusi** jadi lebih **konsisten** dan padat.



Blood Pressure

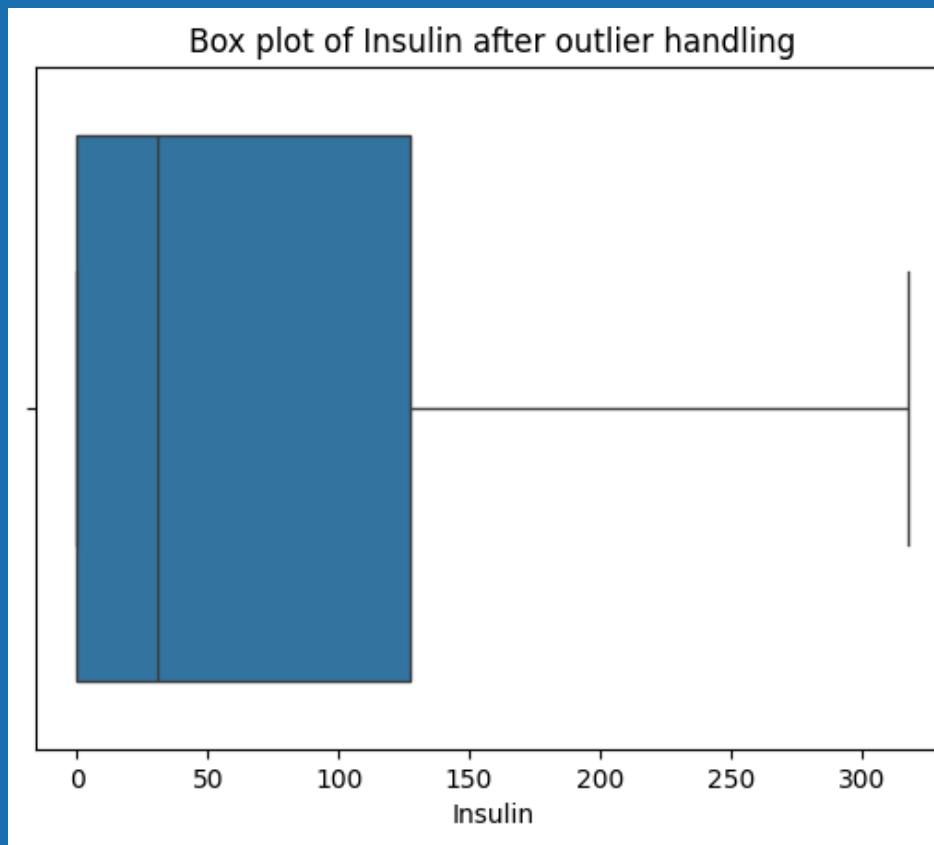
- Sebelum: Ada **titik 0** dan outlier **di atas 100**.
- Sesudah: **Titik 0 hilang** → **distribusi** lebih **stabil**, meski outlier kecil tetap ada.



Skin Thickness

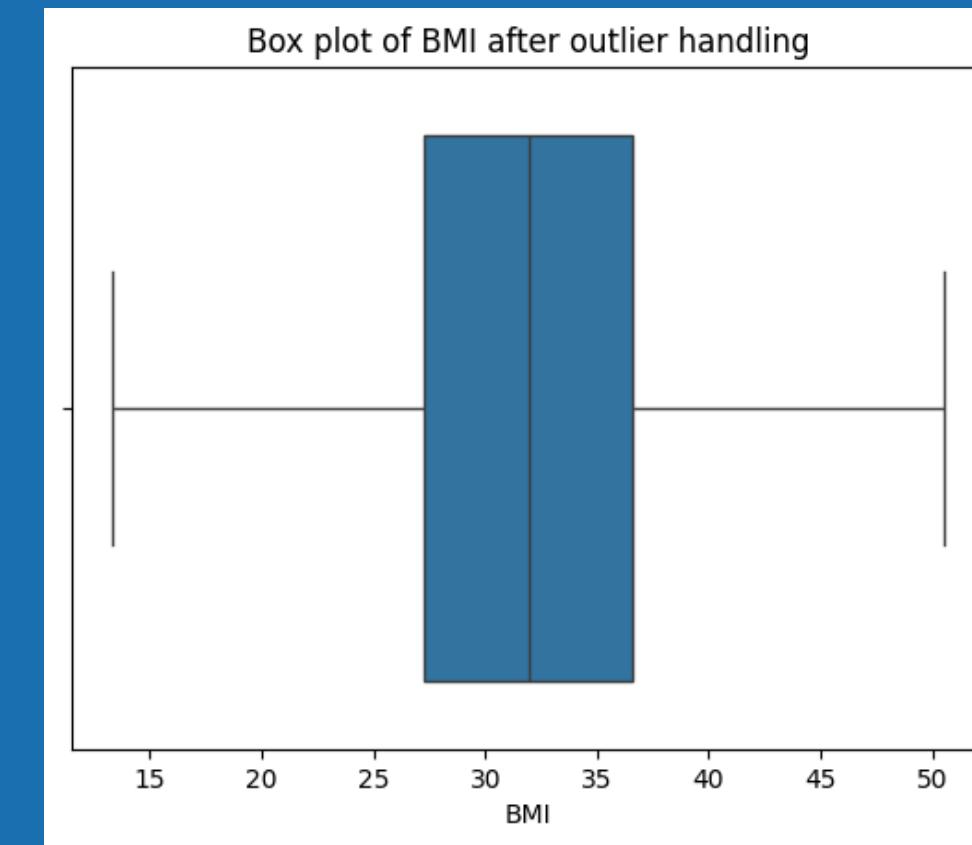
- Sebelum: Banyak data = 0 → **boxplot** terlihat “**kosong**” dan padat di bawah.
- Sesudah: **Median** nilai **muncul**, **distribusi** menjadi lebih **normal**.

OUTLIER HANDLING (NUMERICAL DATA)



Insulin

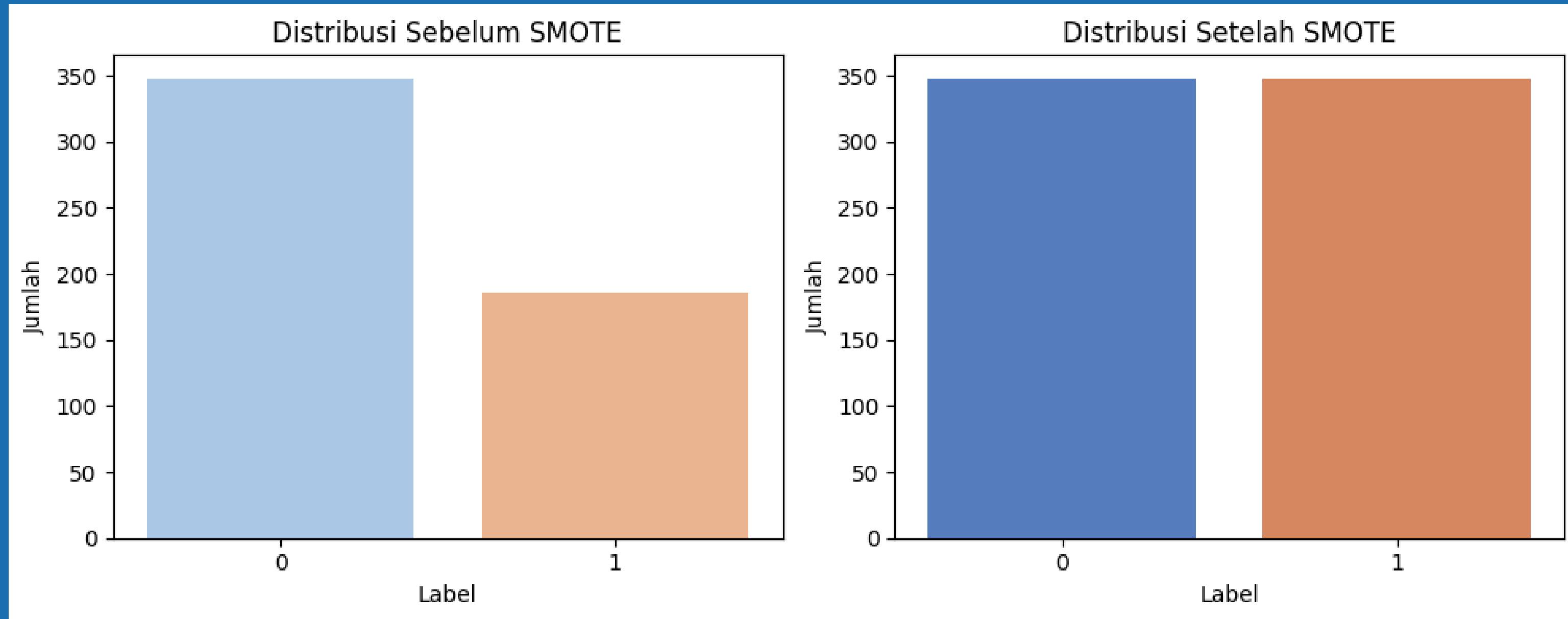
- Sebelum: **Sangat banyak data = 0** → **distribusi tidak informatif.**
- Sesudah: **Nilai tengah lebih tinggi**, meski masih banyak pencilan tinggi (wajar karena insulin memang bervariasi).



BMI

- Sebelum: Ada **BMI = 0** → sangat **tidak logis**.
- Sesudah: Distribusi **BMI normal** kembali (30–40), nilai nol terhapus.

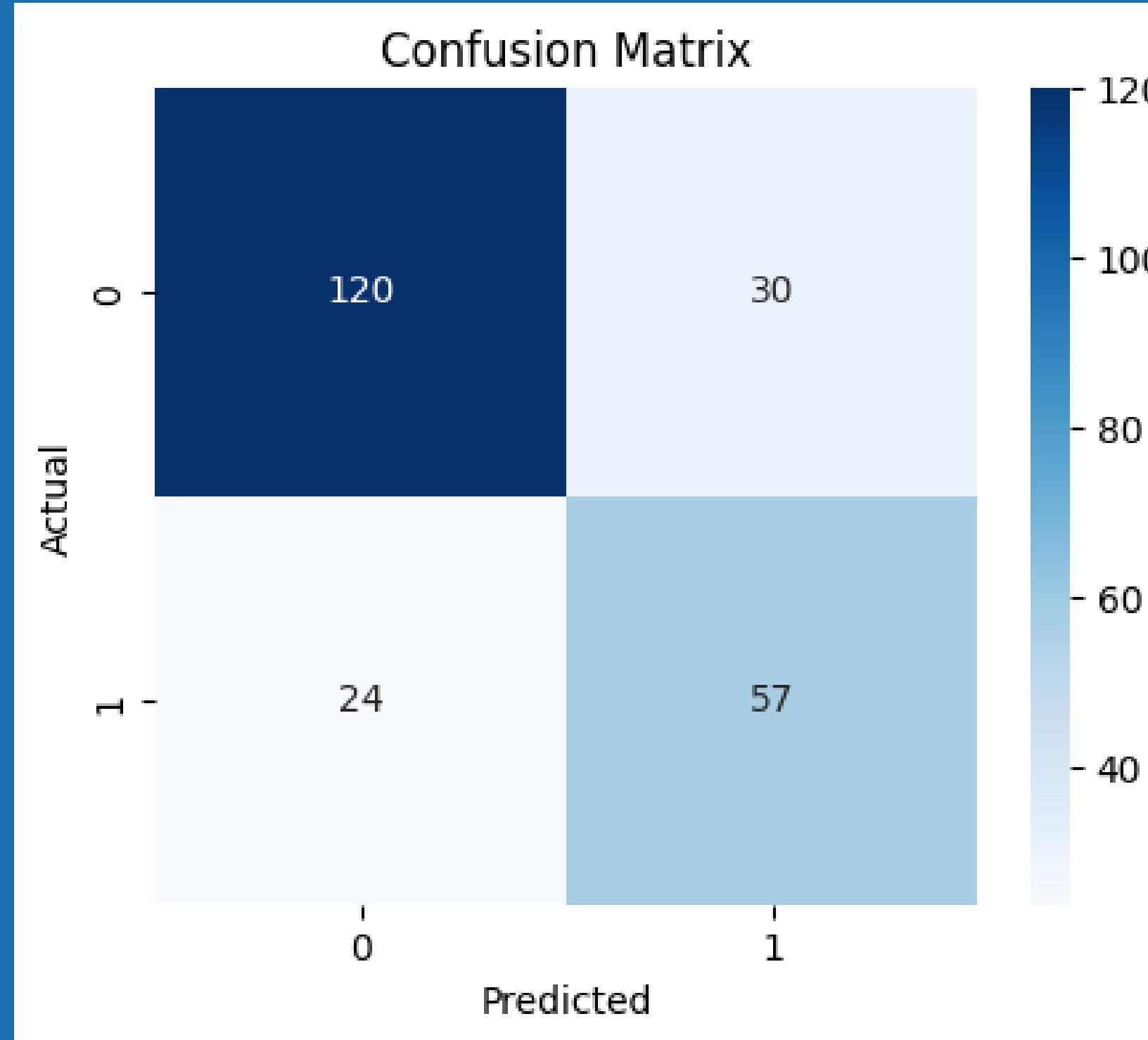
DATA PROCESSING IMBALANCE DATA HANDLING (SMOTE)



Data tidak seimbang (imbalanced) → kasus non-diabetes (0) jauh lebih banyak daripada kasus diabetes (1).

Kedua label (0 dan 1) sekarang hampir sama jumlahnya
Label 1 (diabetes) telah ditambahkan (bukan dari data baru, tapi hasil sintetis)

CONFUSION MATRIX



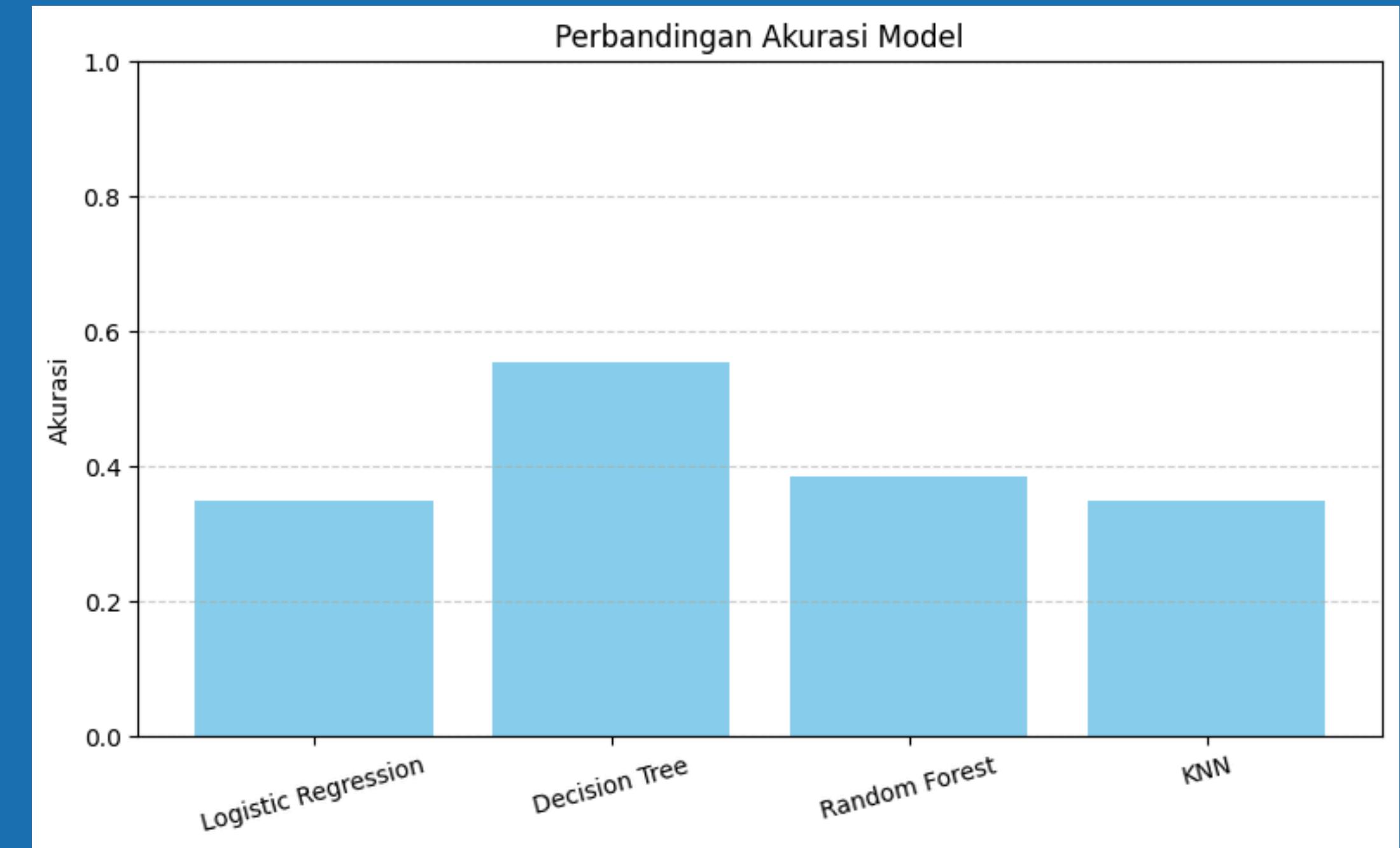
- TP (True Positive): 56 → **Prediksi benar: diabetes**
- TN (True Negative): 120 → **Prediksi benar: tidak diabetes**
- FP (False Positive): 30 → **Salah: diprediksi diabetes, padahal tidak**
- FN (False Negative): 25 → **Salah: diprediksi tidak diabetes, padahal iya**

PERBANDINGAN AKURASI MODEL

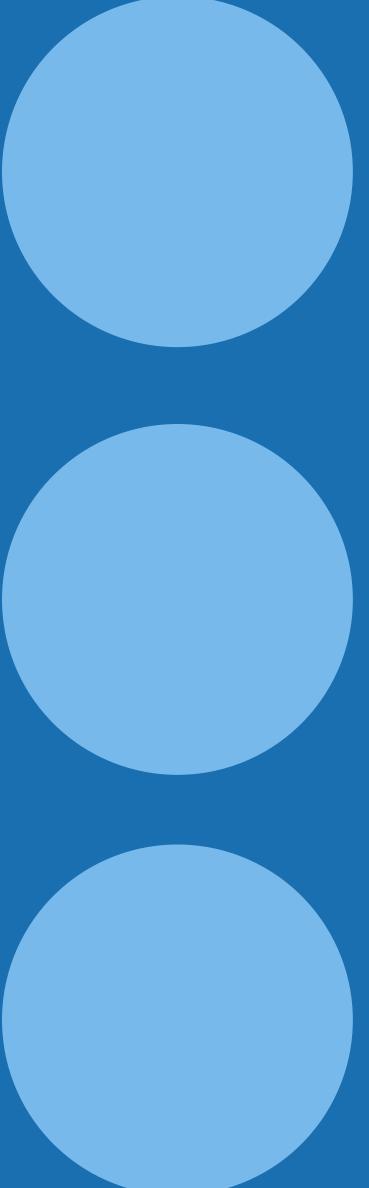
Decision Tree

Accuracy: 0.5541

	precision	recall	f1-score	support
0	0.63	0.77	0.69	150
1	0.26	0.15	0.19	81
accuracy			0.55	231
macro avg	0.44	0.46	0.44	231
weighted avg	0.50	0.55	0.52	231



KESIMPULAN



EDA menunjukkan bahwa beberapa fitur seperti Glucose, BMI, Age, dan DiabetesPedigreeFunction adalah kandidat kuat sebagai prediktor diabetes.

Data missing (nilai 0 pada insulin/skin thickness) perlu perhatian dalam modeling.

Dari perbandingan akurasi model, Random Forest memiliki akurasi paling tinggi.

TERIMA KASIH

