



Machine Learning

DIABETES CLASSIFICATION

DECISION TREE
PRINCIPLE COMPONENT ANALYSIS

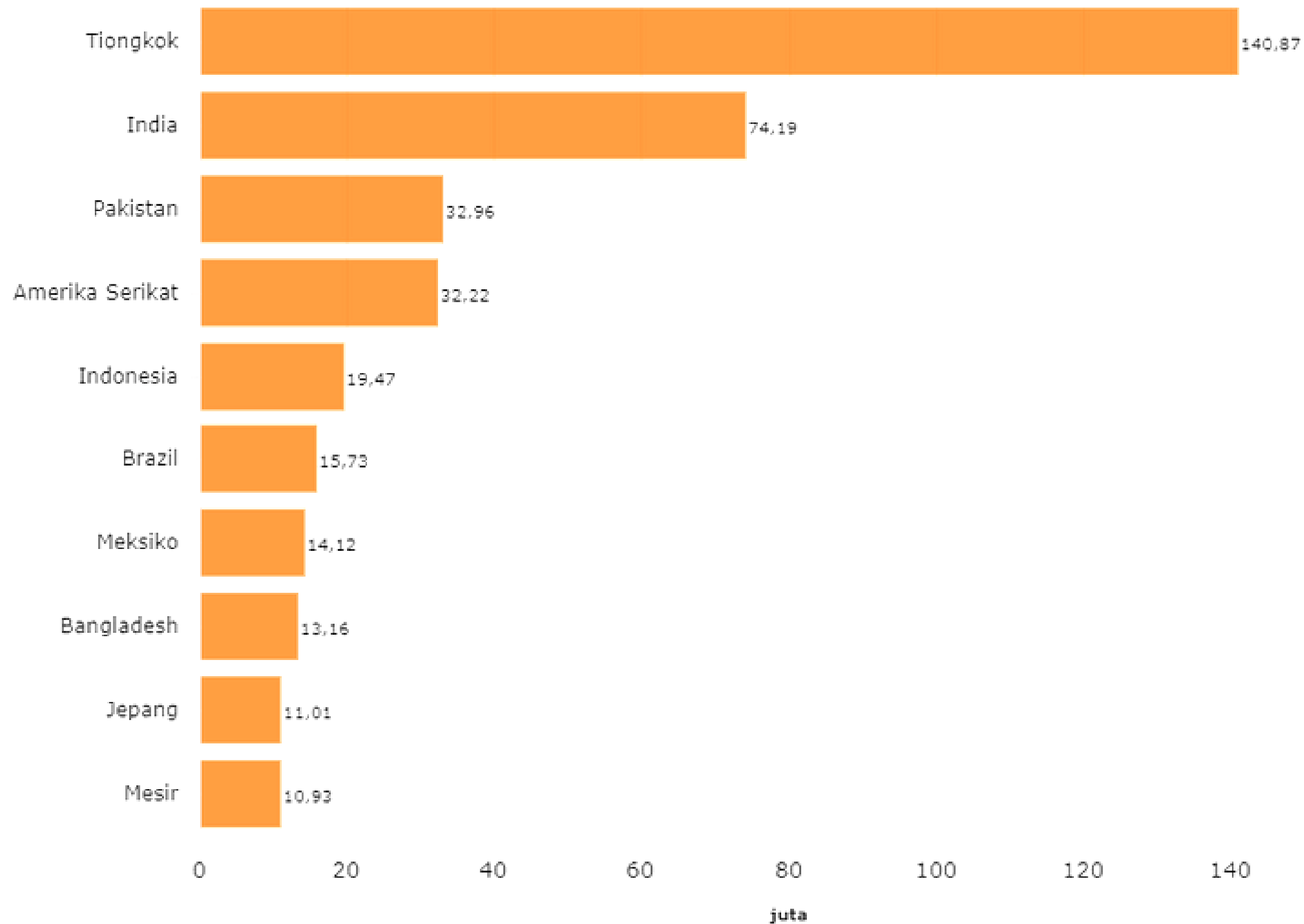
data from UCI Machine Learning

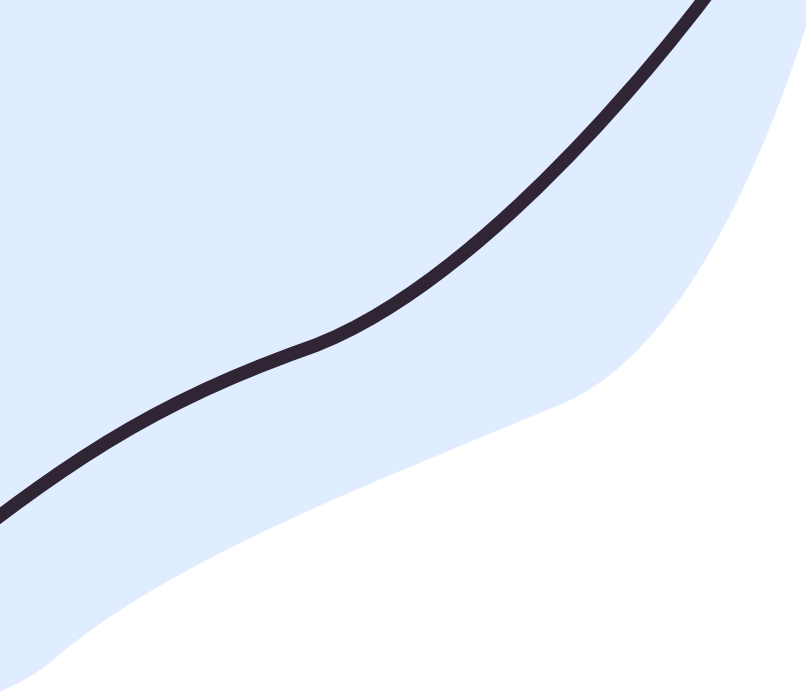
DATA UNDERSTANDING



**INDONESIA
MASUK 5 BESAR**

Jumlah Pengidap Diabetes Berdasarkan Negara 2021



- 
- Diabetes melitus adalah suatu penyakit atau gangguan metabolisme kronis yang ditandai dengan tingginya kadar gula darah disertai adanya gangguan metabolisme karbohidrat.
 - Di Indonesia, penyakit diabetes mellitus tipe 2 merupakan diabetes yang paling dominan ditemukan.

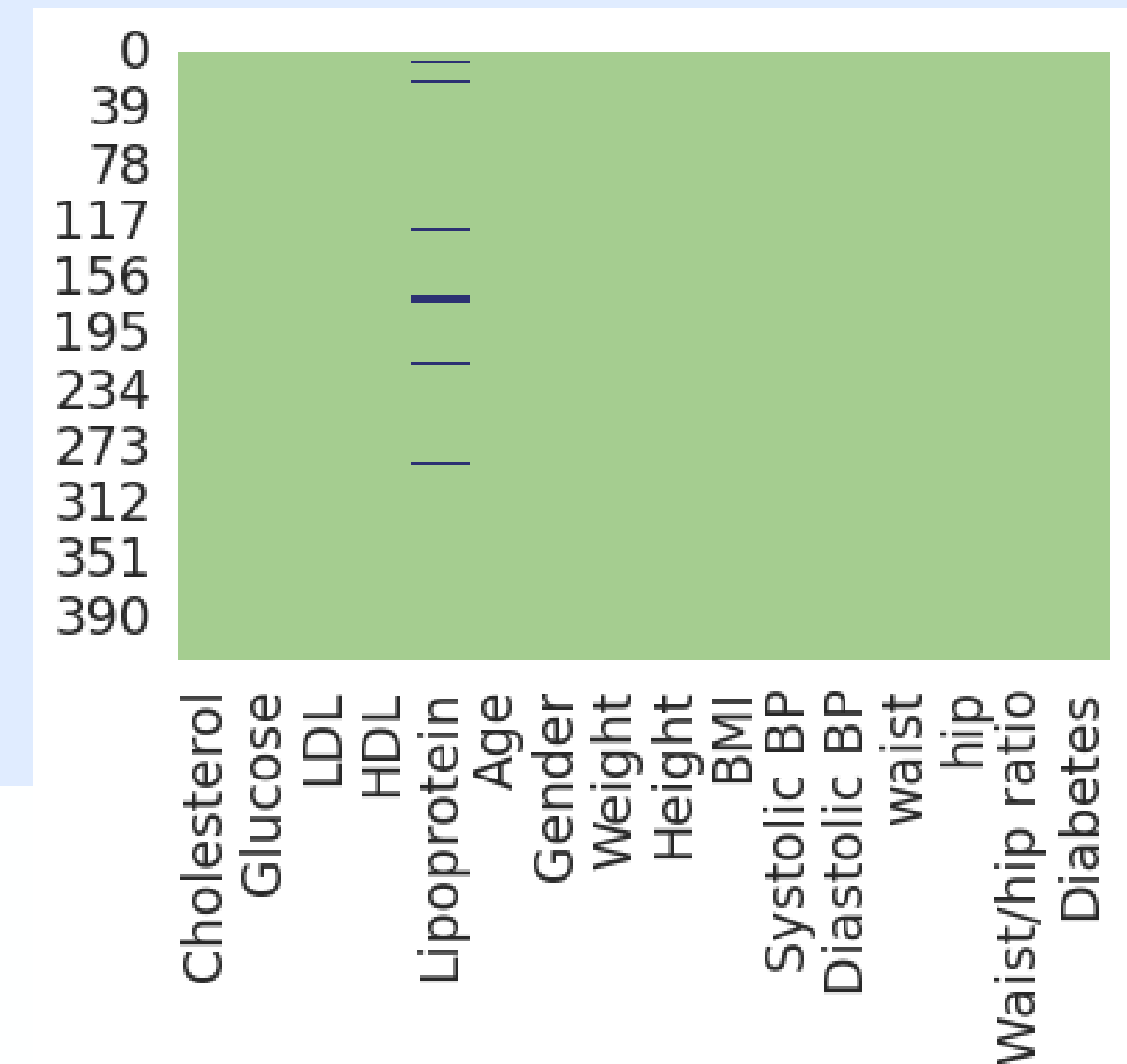
VARIABEL DATA = 15 VARIABEL PREDIKTOR & 1 VARIABEL TARGET

- | | |
|----------------------------------|---------------------------|
| • Kadar kolestrol | • Jenis kelamin |
| • Konsentrasi tingkat gula darah | • Tinggi |
| • Konsentrasi kolesterol baik | • Berat |
| • Rasio kolesterol | • Indeks massa tubuh |
| • Lipoprotein | • Tekanan darah sistolik |
| • Usia pasien | • Tekanan darah diastolik |
| • Ukuran lingkar pinggang | • Ukuran lingkar pinggul |
| • Rasio pinggang pinggul | • Diabetes (YA/TIDAK) |

Check Nilai Hilang

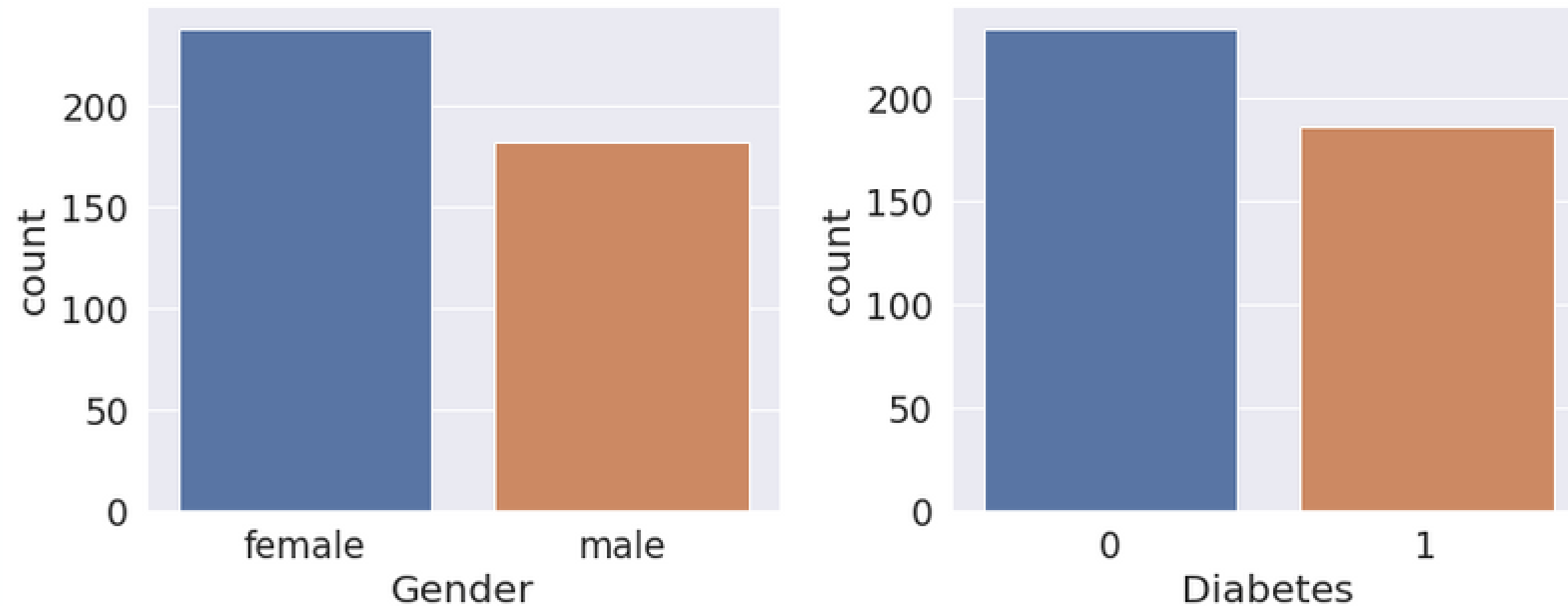
HANYA VARIABEL LIPPOPROTEIN = 3% NILAI HILANG

Analisis Deskriptif



	Cholesterol	Glucose	LDL	HDL	Lipoprotein	Age	Weight	Height	BMI	Systolic BP	Diastolic BP	waist	hip	Waist/hip ratio	Diabetes
count	419.000000	419.000000	419.000000	419.000000	403.000000	419.000000	419.000000	419.000000	419.000000	419.000000	419.000000	419.000000	419.000000	419.000000	419.000000
mean	208.984200	174.341289	102.317422	69.591885	1.445562	37.326969	93.225680	174.558473	29.961575	144.233890	91.226730	41.689737	44.494033	0.933948	0.443914
std	42.839426	43.178117	18.951511	17.078570	0.370320	13.911735	33.250447	22.573713	5.882590	34.209637	24.051851	9.750189	6.809917	0.143045	0.497438
min	98.000000	60.000000	49.000000	0.000000	0.471698	21.000000	0.000000	105.000000	0.000000	90.000000	60.000000	27.000000	30.000000	0.681818	0.000000
25%	179.500000	140.500000	93.000000	66.000000	1.201266	25.000000	66.415000	161.000000	26.150000	123.500000	78.000000	35.000000	40.000000	0.850532	0.000000
50%	202.000000	168.000000	103.000000	72.000000	1.408451	33.000000	85.140000	172.000000	30.700000	138.000000	88.000000	39.000000	43.000000	0.900000	0.000000
75%	230.000000	200.000000	116.000000	78.000000	1.653846	49.000000	115.670000	188.000000	34.150000	150.000000	96.000000	45.000000	48.000000	0.988455	1.000000
max	331.000000	300.000000	133.000000	122.000000	3.833333	76.000000	197.280000	246.000000	38.700000	254.000000	195.000000	70.000000	68.000000	1.750000	1.000000

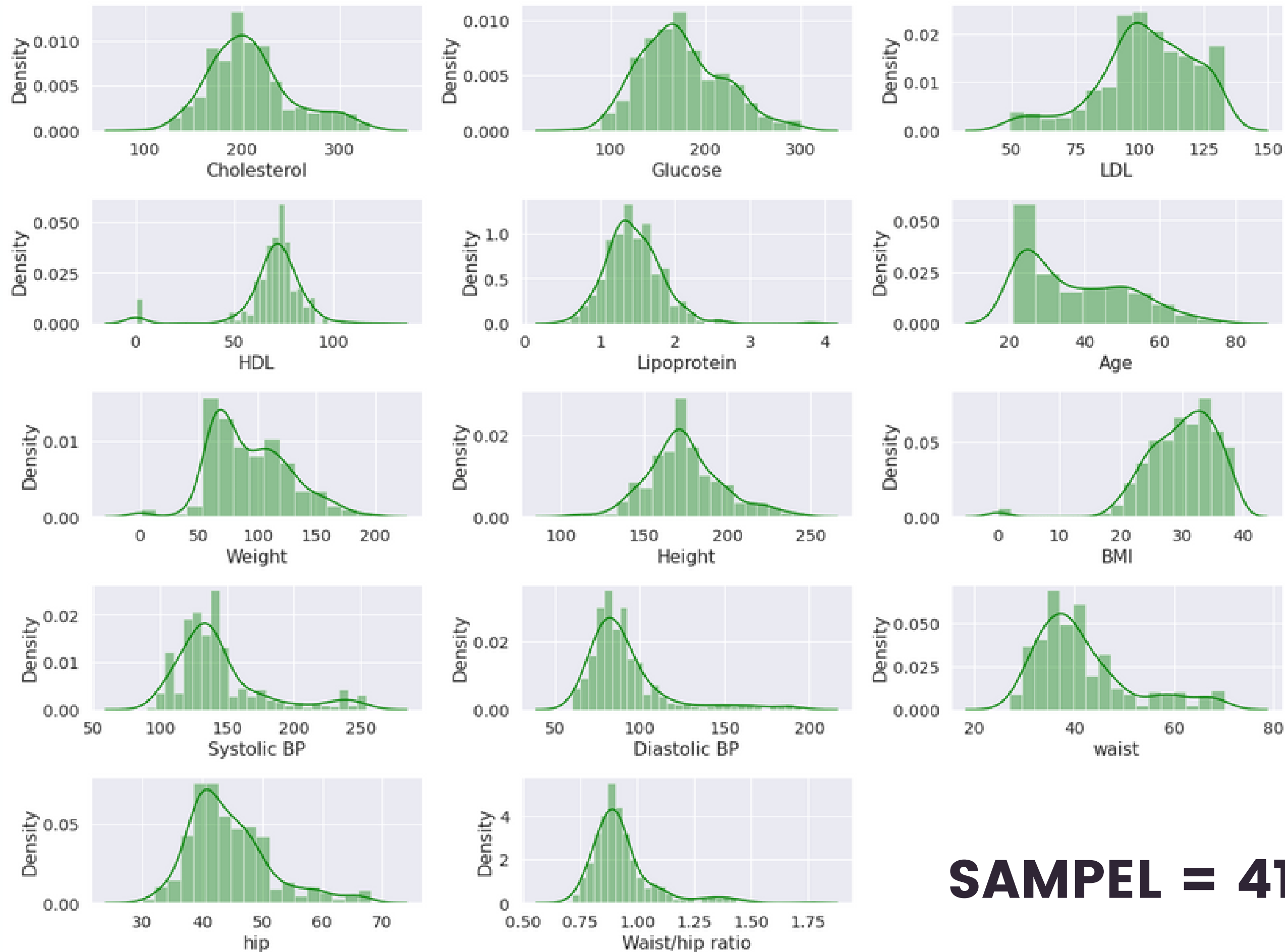
EDA



- Tidak terdapat class imbalance pada variabel prediktor (Gender) dan variabel target (Diabetes)
- Semua variabel terdapat outlier



Variabel	Persentase Outlier
Cholesterol	6.68
Glucose	4.06
LDL	5.49
HDL	5.01
Lipoprotein	3.58
Age	3.58
Weight	4.77
Height	6.44
BMI	1.67
Systolic BP	7.88
Diastolic BP	6.44
waist	6.92
hip	5.73
Waist/hip ratio	5.97



INSIGHT ;

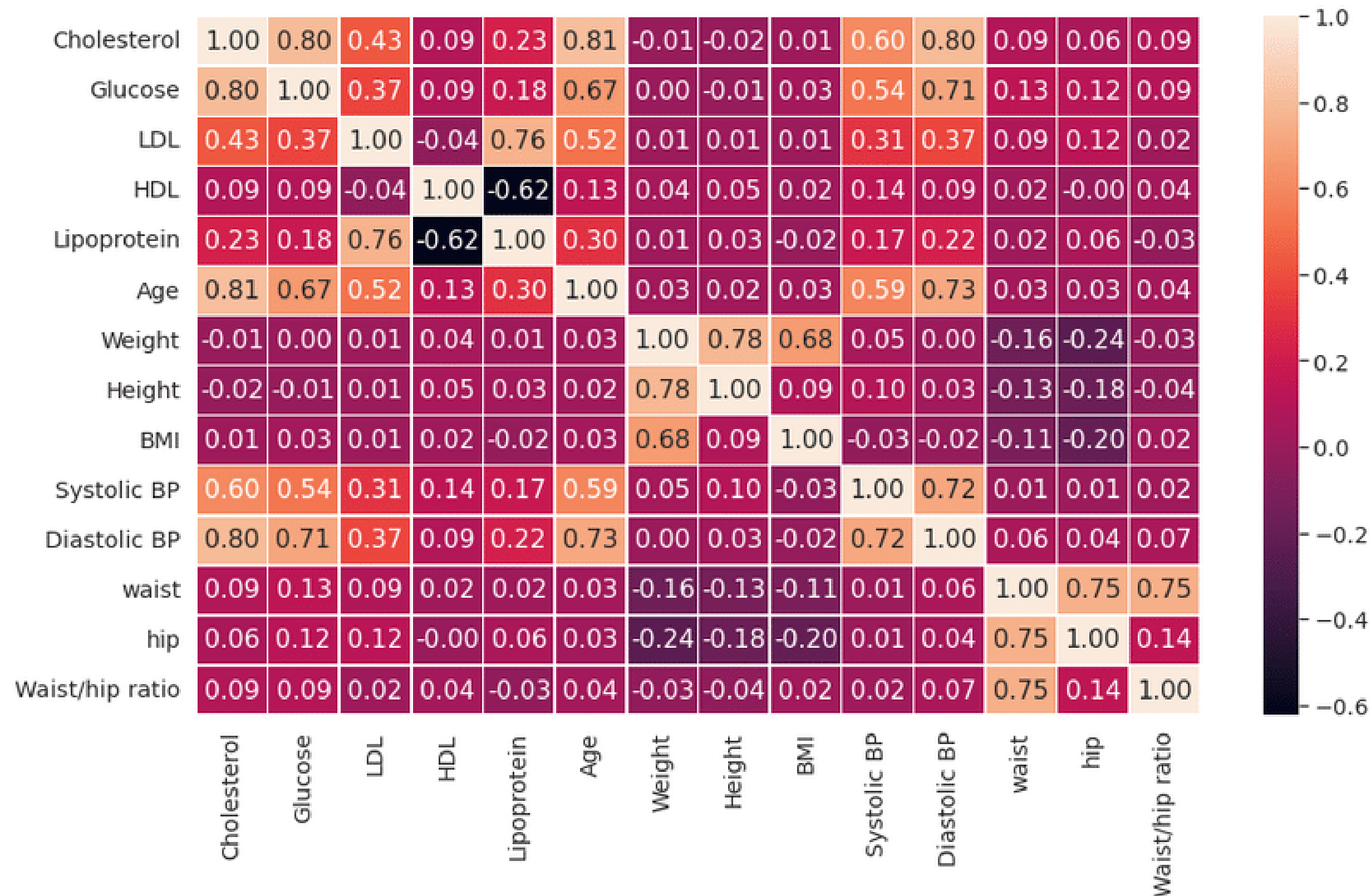
**RATA-RATA
VARIABEL
PREDIKTOR
BERSIFAT SKEW**

ATAU

**MENDEKATI
DISTRIBUSI NORMAL**

**PLOT OUTLIER ADA
PADA FILE KODING**

SAMPEL = 419 DATA



Insight ;

corr (0 – 0,5) lemah
 corr (0,5 – 0,7) sedang
 corr (0,7 – 0,9) kuat
 corr (0,9 – 1) sangat kuat

PCA = SELEKSI VARIABEL – MULTIKOLINEARITAS – CORR \geq 0,7
SEMUA VARIABEL KECUALI HDL TERKENA MULTIOLINEARITAS

Data Preprocessing

- Mengisi nilai hilang dengan nilai median variabel
- Melakukan dummy pada variabel kategori
- Melakukan standarisasi data
- Membagi data latih – 70% & data uji – 30%

Data Modelling

- 1. OUTLIER TIDAK DITANGANI KARENA DECISION TREE TAHAN TERHADAP OUTLIER (JURNAL)**
- 2. OUTLIER DITANGANI DENGAN MEREDUKSI DATA MENGGUNAKAN PCA (JURNAL)**
- 3. ATRIBUT-ATRIBUT KOMPONEN UTAMA – KLASIFIKASI DECISION TREE**
- 4. MEMILIH MODEL DENGAN AKURASI TERTINGGI UNTUK TUNING PARAMETER**
- 5. MEMBANDINGKAN MODEL SEBELUM DAN SESUDAH DILAKUKAN TUNING**

Model tanpa PCA

AKURASI = 0.6349

PRESISI = 0.5833

SENSITIVITAS = 0.7241

Training set accuracy: 1.0

Test set accuracy: 0.6349206349206349

Model dengan PCA

DIPILIH 7 ATRIBUT KOMPONEN UTAMA = 90% VARIANSI DATA

AKURASI = 0.7143

PRESISI = 0.7639

SENSITIVITAS = 0.7432

Training set accuracy: 1.0

Test set accuracy: 0.7142857142857143

Terlihat C4.5 - PCA dapat meningkatkan akurasi tetapi masih terjadi overfitting yang relatif besar dalam model

Tuning Parameter

Model tuning PCA

AKURASI = 0.8016

PRESISI = 0.9722

SENSITIVITAS = 0.7527

Training set accuracy: 0.7337883959044369

Test set accuracy: 0.8015873015873016

PCA-DT-sebelum tuning: 0.7143

jumlah simpul : 107

jumlah cabang : 10

PCA-DT-sesudah tuning: 0.8016

jumlah simpul : 9

jumlah cabang : 4

Kesimpulan

- Melakukan tuning parameter berhasil mengurangi risiko overfitting model
- Tuning parameter mampu mengoptimalkan pembentukan pohon yang lebih akurat