

# UJIAN MEMPERTAHANKAN SKRIPSI

Peneliti : Michael Julius Sitanggang  
Pembimbing : Dr. Elmanani Simamora, S.Si., M.Si  
Narasumber : Dr. Arnita, M.Si  
Narasumber : Sudianto Manullang, S.Si, M.Sc  
Narasumber : Lasker P. Sinaga, M.Si



JURUSAN MATEMATIKA  
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM  
UNIVERSITAS NEGERI MEDAN

2022

”PENERAPAN ANALISIS KOMPONEN UTAMA UNTUK MENINGKATKAN  
AKURASI KLASIFIKASI PADA ALGORITMA DECISION TREE C4.5  
DALAM MENDIAGNOSA PENYAKIT DIABETES MELITUS”

Michael Julius Sitanggang (4183230026)

Di bawah bimbingan

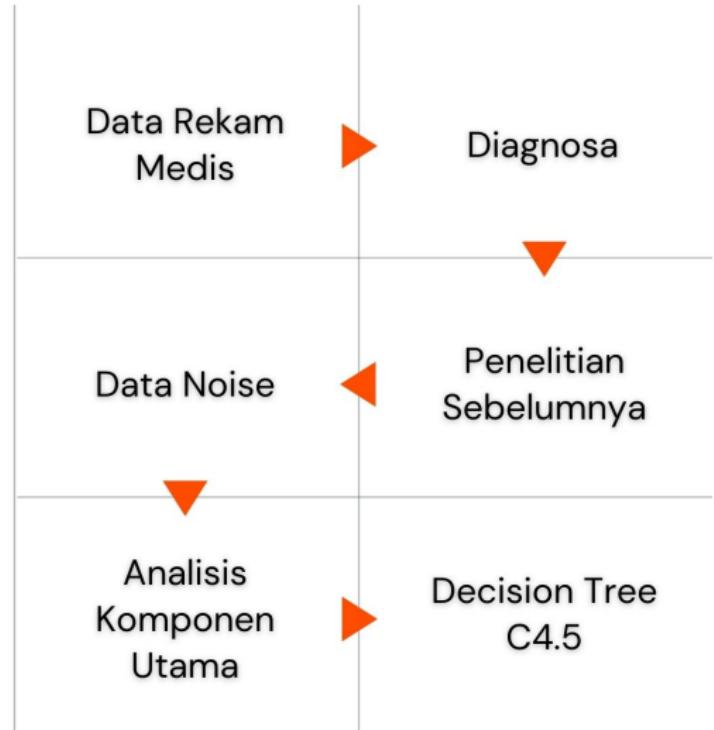
Dr. Elmanani Simamora, S.Si., M.Si



UNIVERSITAS NEGERI MEDAN  
2022

# Pendahuluan

## Latar Belakang



# Tujuan Penelitian



Mengimplementasikan metode Analisis Komponen Utama untuk mengatasi masalah data noise yang terjadi pada model klasifikasi algoritma C4.5.

Mendapatkan perbandingan akurasi hasil klasifikasi data Diabetes Melitus menggunakan algoritma C4.5 sebelum dan sesudah direduksi.

# Hasil dan Pembahasan



- **Analisis Data Eksploratif**

**VARIABEL DATA = 15 VARIABEL PREDIKTOR & 1 VARIABEL TARGET**

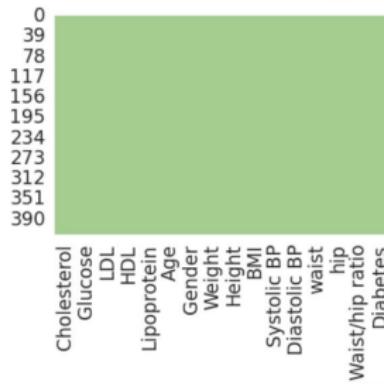
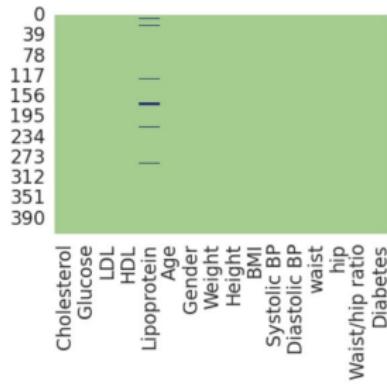
- Kadar kolesterol
- Konsentrasi tingkat gula darah
- Konsentrasi kolesterol baik
- Rasio kolesterol
- Lipoprotein
- Usia pasien
- Ukuran lingkar pinggang
- Rasio pinggang pinggul
- Jenis kelamin
- Tinggi
- Berat
- Indeks massa tubuh
- Tekanan darah sistolik
- Tekanan darah diastolik
- Ukuran lingkar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

**15 VARIABEL PREDIKTOR = 14 VARIABEL NUMERIK & 1 VARIABEL KATEGORIK**

# Check Nilai Hilang

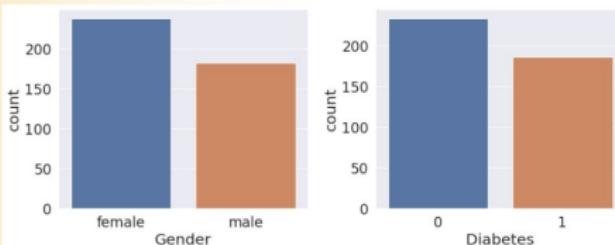
HANYA VARIABEL LIPOPROTEIN = 3% NILAI HILANG

ISI DENGAN MEDIAN (LEBIH ROBUST)



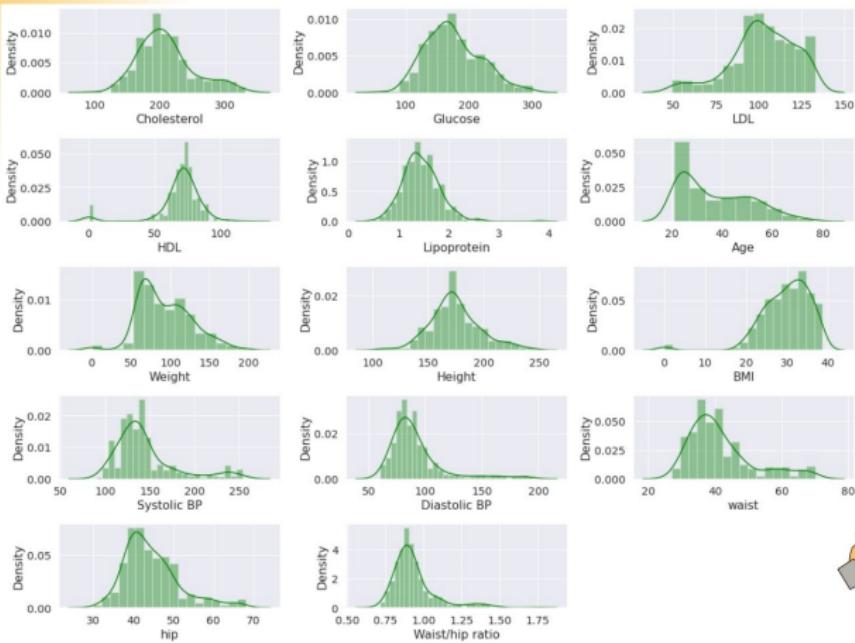
Check Data Duplikat > TIDAK ADA

Nama Variabel	Mean	Median	Min	Max
Kadar kolesterol	208.96	202	98	331
Konsentrasi tingkat gula darah	174.34	168	60	300
Konsentrasi kolesterol baik	102.31	103	49	133
Rasio kolesterol	69.59	72	0	122
Lipoprotein	1.44	1.40	0.47	3.83
Usia pasien	37.32	33	21	76
Tinggi	174.55	172	105	246
Berat	93.22	85.14	0	197.28
Indeks massa tubuh	29.96	30.70	0	38.7
Tekanan darah sistolik	144.23	136	90	254
Tekanan darah diastolik	91.22	86	60	195
Ukuran lingkar pinggang	41.68	39	27	70
Ukuran lingkar pinggul	44.49	43	30	68
Rasio pinggang pinggul	0.93	0.90	0.68	1.75



## Analisis Deskriptif

- batas minimum dan batas maksimum valid disesuaikan / didasarkan artikel kesehatan
- berdasarkan nilai mean dan median didapatkan dugaan dimana rata-rata variabel prediktor mendekati distribusi normal
- tidak terdapat data imbalance pada variabel kategori prediktor maupun variabel target



**INSIGHT ;**

**RATA-RATA  
VARIABEL  
PREDIKTOR  
BERSIFAT SKEW &**

**TERDETEKSI OUTLIER**



# Uji Outlier dan Multikolinearitas

	Cholesterol	Glucose	LDL	HDL	Lipoprotein	Age	Weight	Height	BMI	Systolic BP	Diastolic BP	Waist	Hip	Waist/hip ratio
Cholesterol	1.00	0.80	0.43	0.09	0.23	0.81	-0.01	-0.02	0.01	0.60	0.80	0.09	0.06	0.09
Glucose	0.80	1.00	0.37	0.09	0.18	0.67	0.00	-0.01	0.03	0.54	0.71	0.13	0.12	0.09
LDL	0.43	0.37	1.00	-0.04	0.76	0.52	0.01	0.01	0.01	0.31	0.37	0.09	0.12	0.02
HDL	0.09	0.09	-0.04	1.00	-0.62	0.13	0.04	0.05	0.02	0.14	0.09	0.02	-0.00	0.04
Lipoprotein	0.23	0.18	0.76	-0.62	1.00	0.30	0.01	0.03	-0.02	0.17	0.22	0.02	0.06	-0.03
Age	0.81	0.67	0.52	0.13	0.30	1.00	0.03	0.02	0.03	0.59	0.73	0.03	0.03	0.04
Weight	-0.01	0.00	0.01	0.04	0.01	0.03	1.00	0.78	0.68	0.05	0.00	-0.16	-0.24	-0.03
Height	-0.02	-0.01	0.01	0.05	0.03	0.02	0.78	1.00	0.09	0.10	0.03	-0.13	-0.18	-0.04
BMI	0.01	0.03	0.01	0.02	-0.02	0.03	0.68	0.09	1.00	-0.03	-0.02	-0.11	-0.20	0.02
Systolic BP	0.60	0.54	0.31	0.14	0.17	0.59	0.05	0.10	-0.03	1.00	0.72	0.01	0.01	0.02
Diastolic BP	0.80	0.71	0.37	0.09	0.22	0.73	0.00	0.03	-0.02	0.72	1.00	0.06	0.04	0.07
waist	0.09	0.13	0.09	0.02	0.02	0.03	-0.16	-0.13	-0.11	0.01	0.06	1.00	0.75	0.75
hip	0.06	0.12	0.12	-0.00	0.06	0.03	-0.24	-0.18	-0.20	0.01	0.04	0.75	1.00	0.14
Waist/hip ratio	0.09	0.09	0.02	0.04	-0.03	0.04	-0.03	-0.04	0.02	0.02	0.07	0.75	0.14	1.00



klasifikasi koefisien korelasi; (Cohen 1998) - practical statistics for students

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) %>% (koef korelasi  $\geq 0,7$ )

**semua variabel terkena multiokolinearitas kecuali variabel HDL**

## • Data Preprocessing

- Melakukan dummy pada variabel gender (nominal)
- Melakukan standarisasi data
- Membagi data latih - 70% & data uji - 30% (untuk model tanpa pca)

**standarisasi termasuk tahapan pada PCA**

**DATA SETELAH UJI ASUMSI ;**

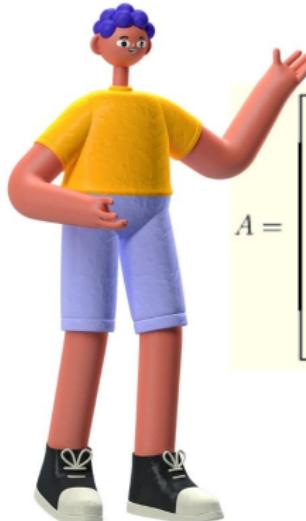
Cholesterol	Glucose	LDL	Lipoprotein	Age	Weight	Height	BMI	Systolic BP	Diastolic BP	waist	hip	Waist/hip ratio
127	100	50	0.793651	21	146.75	196	38.2	90	70	35	38	0.921053
130	98	54	0.870968	23	58.28	150	25.9	98	71	29	35	0.828571
132	98	81	1.408451	22	64.58	150	28.7	100	70	33	38	0.868421
270	231	121	1.635135	54	105.97	174	35.0	100	94	39	41	0.951220
178	150	115	1.642857	27	50.90	141	25.6	100	60	40	45	0.888889

**DATA SETELAH PREPROCESSING ;**

Cholesterol	Glucose	LDL	Lipoprotein	Age	Weight	Height	BMI	Systolic BP	Diastolic BP	waist	hip	Waist/hip ratio	female	male	Diabetes
-1.915576	-1.723794	-2.763894	-1.792986	-1.175014	1.611657	0.950980	1.402150	-1.587235	-0.883595	-0.686934	-0.954754	-0.090256	1	0	0
-1.845464	-1.770169	-2.552577	-1.579874	-1.031079	-1.052240	-1.089224	-0.691265	-1.353103	-0.841969	-1.303042	-1.395815	-0.737541	1	0	0
-1.798722	-1.770169	-1.126185	-0.098386	-1.103047	-0.862543	-1.089224	-0.214716	-1.294570	-0.883595	-0.892303	-0.954754	-0.458630	1	0	0
1.426461	1.313778	0.986988	0.526434	1.199920	0.383741	-0.024770	0.857522	-1.294570	0.115442	-0.276195	-0.513694	0.120884	0	1	1
-0.723661	-0.564415	0.670012	0.547719	-0.743208	-1.274457	-1.488394	-0.742324	-1.294570	-1.299861	-0.173510	0.074387	-0.315373	0	1	0

## • Analisis Komponen Utama

DIBANGUN MENGGUNAKAN PEMROGRAMAN PYTHON



$A =$

$$\begin{bmatrix} 1.00 & 0.80 & 0.43 & 0.22 & 0.82 & -0.01 & -0.02 & 0.01 & 0.60 & 0.80 & 0.09 & 0.06 & 0.09 & -0.00 & 0.00 \\ 0.80 & 1.00 & 0.37 & 0.17 & 0.67 & 0.00 & -0.01 & 0.03 & 0.54 & 0.71 & 0.13 & 0.12 & 0.09 & 0.01 & -0.01 \\ 0.43 & 0.37 & 1.00 & 0.75 & 0.52 & 0.01 & 0.01 & 0.01 & 0.31 & 0.37 & 0.09 & 0.12 & 0.02 & 0.05 & -0.05 \\ 0.22 & 0.17 & 0.75 & 1.00 & 0.30 & 0.02 & 0.03 & -0.02 & 0.17 & 0.22 & 0.02 & 0.06 & -0.03 & 0.07 & -0.07 \\ 0.82 & 0.67 & 0.52 & 0.30 & 1.00 & 0.03 & 0.02 & 0.03 & 0.59 & 0.73 & 0.03 & 0.03 & 0.04 & 0.01 & -0.01 \\ -0.01 & 0.00 & 0.01 & 0.02 & 0.03 & 1.00 & 0.78 & 0.68 & 0.05 & 0.00 & -0.16 & -0.24 & -0.03 & 0.01 & -0.01 \\ -0.02 & -0.01 & 0.01 & 0.03 & 0.02 & 0.78 & 1.00 & 0.09 & 0.10 & 0.03 & -0.13 & -0.18 & -0.04 & 0.00 & -0.00 \\ 0.01 & 0.0 & 0.01 & -0.02 & 0.03 & 0.68 & 0.09 & 1.00 & -0.03 & -0.02 & -0.11 & -0.20 & 0.02 & 0.00 & -0.00 \\ 0.60 & 0.54 & 0.31 & 0.17 & 0.59 & 0.05 & 0.10 & -0.03 & 1.00 & 0.72 & 0.01 & 0.01 & 0.02 & -0.02 & 0.02 \\ 0.80 & 0.71 & 0.37 & 0.22 & 0.73 & 0.00 & 0.03 & -0.02 & 0.72 & 1.00 & 0.06 & 0.04 & 0.07 & 0.01 & -0.01 \\ 0.09 & 0.13 & 0.09 & 0.02 & 0.03 & -0.16 & -0.13 & -0.11 & 0.01 & 0.06 & 1.00 & 0.76 & 0.75 & 0.01 & -0.01 \\ 0.06 & 0.12 & 0.12 & 0.06 & 0.03 & -0.24 & -0.18 & -0.20 & 0.01 & 0.04 & 0.76 & 1.00 & 0.14 & 0.10 & -0.10 \\ 0.09 & 0.09 & 0.02 & -0.03 & 0.04 & -0.03 & 0.04 & 0.02 & 0.02 & 0.07 & 0.75 & 0.14 & 1.00 & 0.09 & 0.09 \\ -0.00 & 0.01 & 0.05 & 0.07 & 0.01 & 0.01 & 0.00 & 0.00 & -0.02 & 0.01 & 0.01 & 0.10 & -0.09 & 0.25 & -0.25 \\ 0.00 & -0.01 & -0.05 & -0.07 & -0.01 & -0.01 & -0.00 & -0.00 & 0.02 & -0.01 & -0.01 & -0.10 & 0.09 & -0.25 & 0.25 \end{bmatrix}$$

SELANJUTNYA HITUNG NILAI EIGEN DAN VEKTOR EIGEN DARI Matriks Kovarians ; MEREPRESENTASIKAN SEBARAN DATA DARI SUATU DATASET

n	Nilai Eigen ( $\lambda_n$ )	Vektor Eigen ( $x_n$ )
1	4.255	-0.437, -0.02 , -0.033, 0.174, -0.052, 0.043, 0.288, -0.026, -0.077, -0.148, 0.81 , -0.065, 0.01 , 0.002, -0.
2	2.488	-0.401, 0.002, 0.011, 0.191, -0.035, 0.14 , 0.391, -0.045, 0.642, 0.325, -0.298, -0.138, -0.007, 0.01 , 0.
3	1.761	-0.31 , -0.003, -0.032, -0.564, -0.106, -0.097, 0.021, -0.16, -0.121, 0.459, 0.055, 0.557, 0.004, -0.001, 0.
4	1.424	-0.213, -0.021, -0.056, -0.685, -0.073, -0.175, -0.089, 0.083, 0.271, -0.406, -0.015, -0.438, -0.005, -0.003, 0.
5	0.96	-0.424, -0.061, -0.051, 0.054, -0.054, 0.012, 0.195, -0.084, -0.686, 0.02 , -0.41 , -0.349, 0.003, 0.004, -0.
6	0.858	-0.002, -0.485, 0.471, -0.058, 0.075, 0.114, 0.001, -0.047, -0.003, -0.033, 0.01 , 0.013, -0.07 , 0.716, 0.
7	0.497	-0.008, -0.382, 0.352, -0.067, 0.619, -0.188, 0.163, -0.019, -0.007, 0.011, 0.009, 0.006, 0.042, -0.528, -0.
8	0.406	0.002, -0.333, 0.346, 0 , -0.624, 0.38 , -0.177, -0.036, 0.01 , -0.035, 0.004, -0.007, 0.04 , -0.447, -0.
9	0.315	-0.361, -0.077, -0.043, 0.191, 0.177, -0.072, -0.787, 0.032, 0.062, 0.327, 0.105, -0.217, 0.009, 0.005, -0.
10	0.23	-0.425, -0.04 , -0.045, 0.197, 0.064, 0.006, -0.153, 0.182, 0.075, -0.58 , -0.265, 0.552, -0.006, -0.011, 0.
11	0.136	-0.082, 0.474, 0.482, -0.022, 0.038, 0.019, -0.036, -0.01, -0.023, -0.027, -0.012, -0.016, 0.726, 0.065, -0.
12	0.182	-0.071, 0.434, 0.227, -0.146, 0.294, 0.561, -0.094, -0.273, -0.032, -0.092, 0.011, -0.022, -0.485, -0.058, 0.
13	0.003	-0.061, 0.288, 0.489, 0.12 , -0.247, -0.538, 0.03 , 0.269, -0.023, 0.07 , -0.005, -0.006, -0.479, -0.031, 0.
14	-0.01	-0.01 , 0.007, -0.015, -0.108, 0.084, 0.26 , 0.055, 0.621, -0.073, 0.135, 0.027, -0.007, 0.004, 0.003, 0.707
15	-0.0	0.01 , -0.007, 0.015, 0.108, -0.084, -0.26 , -0.055, -0.621, 0.073, -0.135, -0.027, 0.007, -0.004, -0.003, 0.707

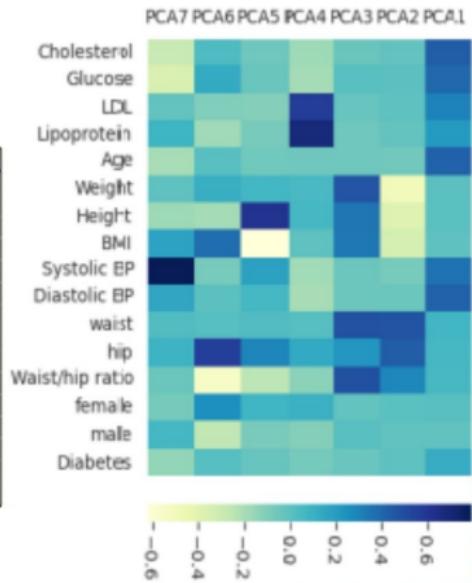
## SELANJUTNYA MENGHITUNG PROPORSI VARIAN DARI 15 VARIABEL UNTUK SETIAP KOMPONEN UTAMA :

Komponen Utama	Nilai Eigen ( $\lambda_n$ )	(%) Variansi	(%) Kumulatif
1	4.255	31.5%	31.5%
2	2.488	18.4%	49.9%
3	1.761	13.0%	62.9%
4	1.424	10.5%	73.4%
5	0.96	7.1%	80.5%
6	0.858	6.3%	86.8%
7	0.497	3.7%	90.5%
8	0.406	3.0%	93.5%
9	0.315	2.3%	95.8%
10	0.23	1.7%	97.5%
11	0.136	1.0%	98.5%
12	0.182	1.3%	99.8%
13	0.003	0.1%	99.9%
14	-0.01	0.0%	99.9%
15	-0.0	0.0%	99.9%

Persentase varians total dianggap cukup mewakili varians jika data 75% atau lebih (Arista 2015) maka dipilih PC1 – PC7 (90.5%).

SELANJUTNYA UNTUK MELIHAT KONTRIBUSI VARIABEL PREDIKTOR DI DALAM TUJUH ATRIBUT KOMPONEN UTAMA MAKA DILAKUKAN PERHITUNGAN BOBOT ATRIBUT MENGGUNAKAN VEKTOR EIGEN ;

Nama Variabel	<i>PCA<sub>1</sub></i>	<i>PCA<sub>2</sub></i>	<i>PCA<sub>3</sub></i>	<i>PCA<sub>4</sub></i>	<i>PCA<sub>5</sub></i>	<i>PCA<sub>6</sub></i>	<i>PCA<sub>7</sub></i>
Kadar kolesterol	0.436	-0.019	-0.033	-0.17	-0.051	0.042	-0.282
Konsentrasi tingkat gula darah	0.398	0.002	0.012	-0.186	-0.033	0.139	-0.356
Konsentrasi kolesterol baik	0.305	-0.003	-0.026	0.567	-0.107	-0.097	-0.014
Lipoprotein	0.208	-0.021	-0.05	0.687	-0.075	-0.175	0.083
Usia pasien	0.422	-0.061	-0.05	-0.05	-0.053	0.011	-0.198
Tinggi	0.001	-0.484	0.471	0.054	0.074	0.115	-0.003
Berat	0.007	-0.382	0.353	0.066	0.618	-0.185	-0.168
Indeks massa tubuh	-0.003	-0.333	0.345	-0.005	-0.625	0.378	0.18
Tekanan darah sistolik	0.357	-0.077	-0.041	-0.183	0.182	-0.075	0.787
Tekanan darah diastolik	0.421	-0.039	-0.044	-0.19	0.068	0.004	0.161
Ukuran lingkar pinggang	0.08	0.474	0.482	0.019	0.037	0.019	0.034
Ukuran lingkar pinggul	0.069	0.434	0.228	0.145	0.292	0.562	0.096
Rasio pinggang pinggul	0.06	0.288	0.488	-0.124	-0.245	-0.539	-0.037
Wanita	0.01	0.007	-0.015	0.107	0.082	0.261	-0.07
Pria	-0.01	-0.007	0.015	-0.107	-0.082	-0.261	0.07



SETELAH ITU DATA KOMPONEN UTAMA DIBAGI MENJADI 70% DATA LATIH DAN 30% DATA UJI.

- Pembentukan Model Decision Tree C4.5**

DIBANGUN MENGGUNAKAN PEMROGRAMAN PYTHON

setelah menghitung entropi dan gain lalu dipilih gain tertinggi sebagai root dilakukan pada kedua model yaitu ;

**C4.5 dengan Data Asli dan C4.5 dengan Data Komponen Utama ;**

```
Cholesterol <= 0.316
gini = 0.495
samples = 293
value = [161, 132]
class = No
```

&

```
PC 1 <= 0.676
gini = 0.495
samples = 293
value = [161, 132]
class = No
```

**-> ROOT**

selanjutnya didapatkan perbandingan pola treeplot dari kedua model, yaitu ;

Jenis Data	Cabang	Simpul
Data Asli	10	89
Data PCA	9	105

- Evaluasi Model

DECISION TREE C4.5 DENGAN DATA ASLI



AKURASI = 64.29%

PRESISI = 59.72%

SENSITIVITAS = 72.88%

DECISION TREE C4.5 DENGAN DATA PCA



AKURASI = 70.84%

PRESISI = 73.61%

SENSITIVITAS = 73.61%

- **Pembahasan**

**PERBANDINGAN AKURASI MODEL :**

Jenis Data	Akurasi	Presisi	Sensitivitas
Data Asli	64.29%	59.72%	72.88%
Data PCA	70.84%	73.61%	73.61%
<b>Selisih</b>	<b>6.55%</b>	<b>13.89%</b>	<b>0.73%</b>

**ANALISIS POLA YANG DITEMUKAN :**

Pohon keputusan yang terbentuk dengan menggunakan **data komponen utama memiliki bentuk lebih sederhana dengan bentuk 9 cabang tetapi simpul yang lebih banyak yaitu 105 simpul** dan menghasilkan akurasi yang lebih tinggi jika dibandingkan dengan menggunakan data asli dengan selisih sebesar 6.55%.

Pohon keputusan yang terbentuk dengan menggunakan **data komponen utama lebih membentuk banyak rule atau aturan baru sehingga membentuk pohon lebih kompleks dan terbukti meningkatkan nilai akurasi, presisi dan sensitivitas secara signifikan.**



## Kesimpulan Penelitian



1. Penerapan analisis komponen utama (PCA) pada data diagnosa penyakit diabetes melitus digunakan sebagai seleksi fitur. PCA mereduksi dimensi dataset yang terdiri dari sejumlah besar variabel yang mungkin saling berkorelasi, dengan mempertahankan sebagian besar variansi dari keseluruhan variabel-variabel aslinya. Dalam hal ini fitur atau variabel dapat juga disebut sebagai atribut. Data yang semula memiliki 15 atribut dan 1 kelas menjadi 7 atribut dan 1 kelas setelah diterapkan PCA.
2. Dari data komponen utama yang terbentuk, didapatkan informasi bahwa data komponen utama mewakili varians sebesar 90.5% dari data asli diabetes melitus. Bobot yang lebih terinci mempengaruhi pembentukan data komponen utama antara lain konsentrasi kolesterol baik, lipoprotein, usia pasien, tekanan darah sistolik, tekanan darah diastolik, ukuran lingkar pinggang dan ukuran lingkar pinggul.
3. Model Decision Tree C4.5 yang digunakan dalam klasifikasi data komponen utama menghasilkan banyak rule atau aturan baru sehingga membentuk pohon lebih kompleks dibandingkan klasifikasi data aslinya. Dalam hal ini data komponen utama dapat mengurangi ditemukan banyak cabang yang terdapat noise atau outlier dalam proses klasifikasi algoritma C4.5 sehingga membuat kinerjanya menjadi lebih optimal.

### lanjutan kesimpulan

4. Hasil akurasi yang didapatkan ketika diterapkan PCA pada algoritma C4.5 adalah sebesar 70.84%. Sementara C4.5 tanpa PCA menghasilkan akurasi sebesar 64.29%. Sehingga dapat diketahui adanya peningkatan sebesar 6.55% berdasarkan perbandingan akurasi yang dihasilkan. Hal ini menunjukkan analisis komponen utama (PCA) berperan penting dalam mengoptimalkan kinerja model Decision Tree C4.5 sehingga dapat menghasilkan akurasi yang lebih baik.

### Saran



1. Menambahkan variabel pada data rekam medis diagnosa diabetes melitus seperti hal-hal yang menjadi pendukung adanya penyakit diabetes melitus (dapat dilihat berdasarkan kadar kolesterol, hipertensi, kondisi fisik, riwayat keturunan dan lainnya).
2. Jumlah data ditambah, sehingga dapat diperoleh hasil akurasi fungsi algoritma yang lebih baik.
3. Menggunakan seleksi fitur lain untuk melakukan reduksi dimensi data sehingga akurasi model Decision Tree C4.5 yang dihasilkan dapat lebih baik.

# Terima Kasih

UNIVERSITAS NEGERI MEDAN – MATEMATIKA

# Sekian dan Terima Kasih

#datascientist #machinelearningengineer

