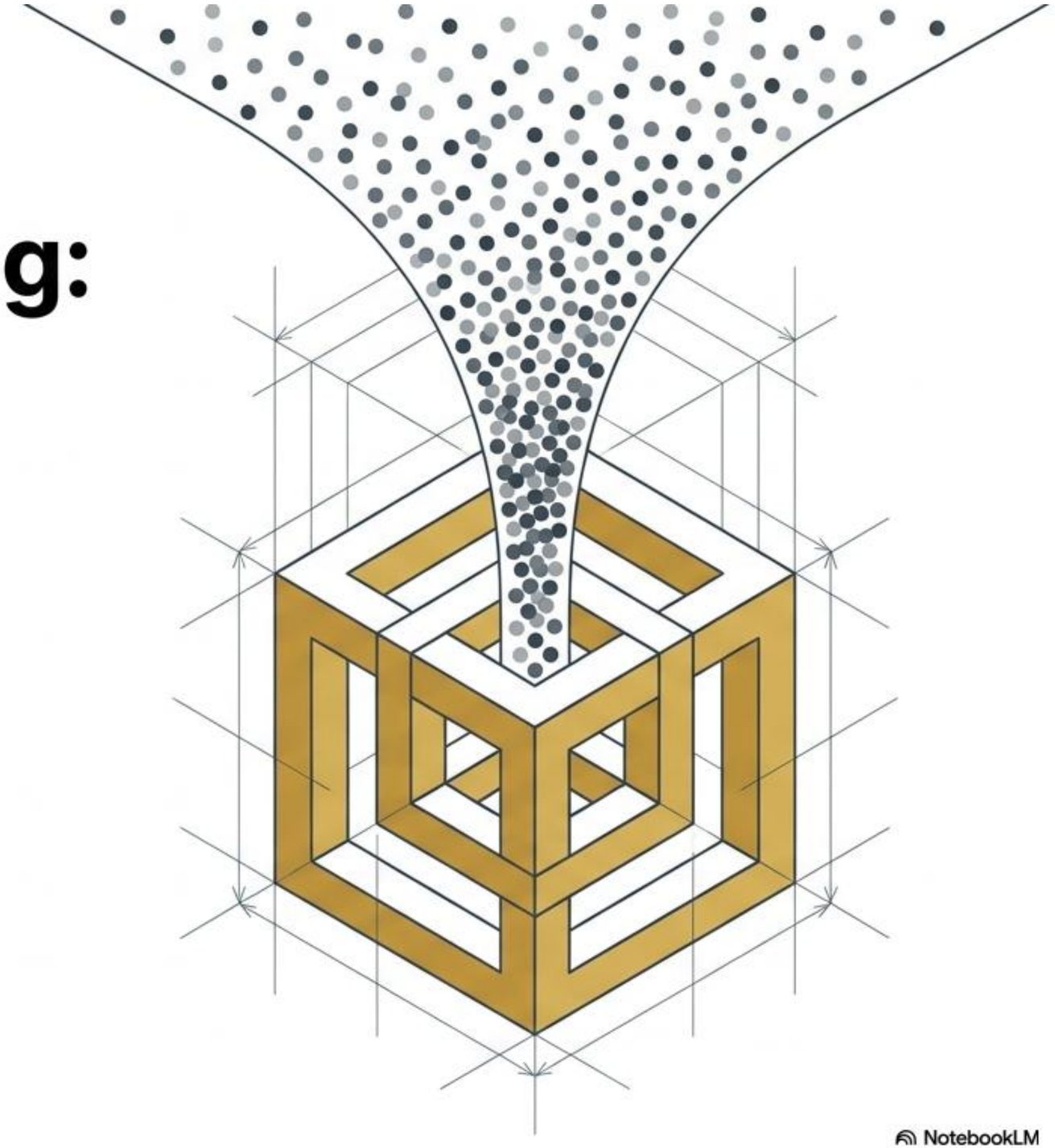
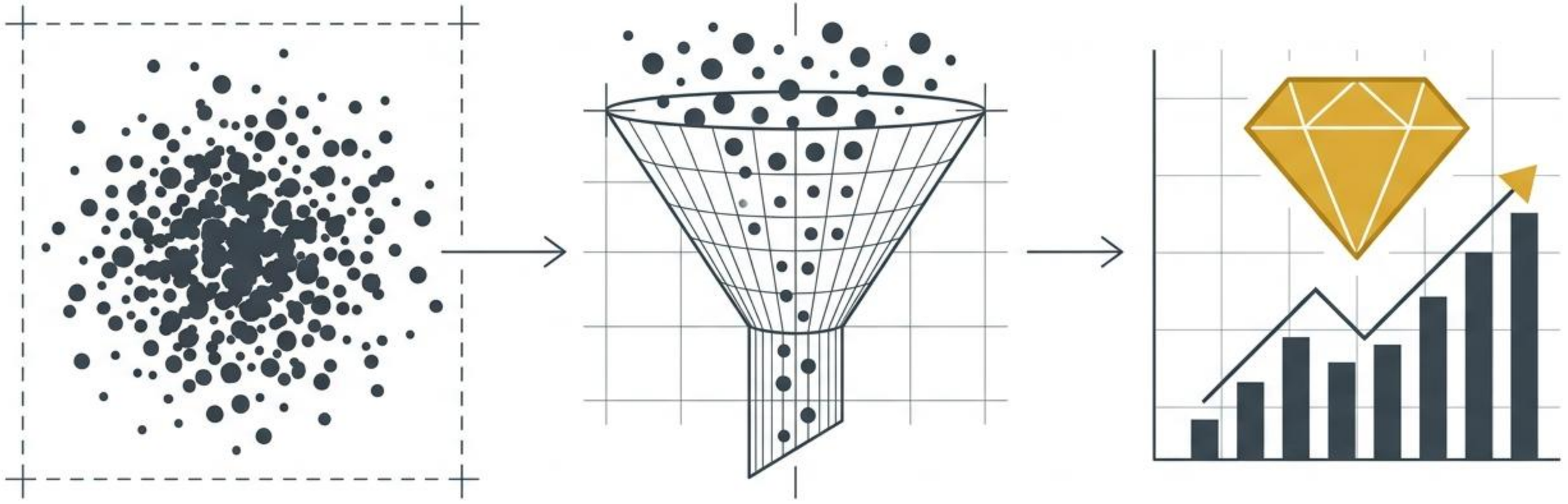
Konsep Data Mining: Mengubah Data Besar Menjadi Keputusan Bisnis

Panduan Komprehensif: Materi, Studi Kasus, & Visualisasi untuk Eksekutif.



Apa itu Data Mining?



Data Besar

Bukan sekadar proses administratif mengumpulkan data.

Menggali Pola

Penemuan pola tersembunyi secara otomatis.

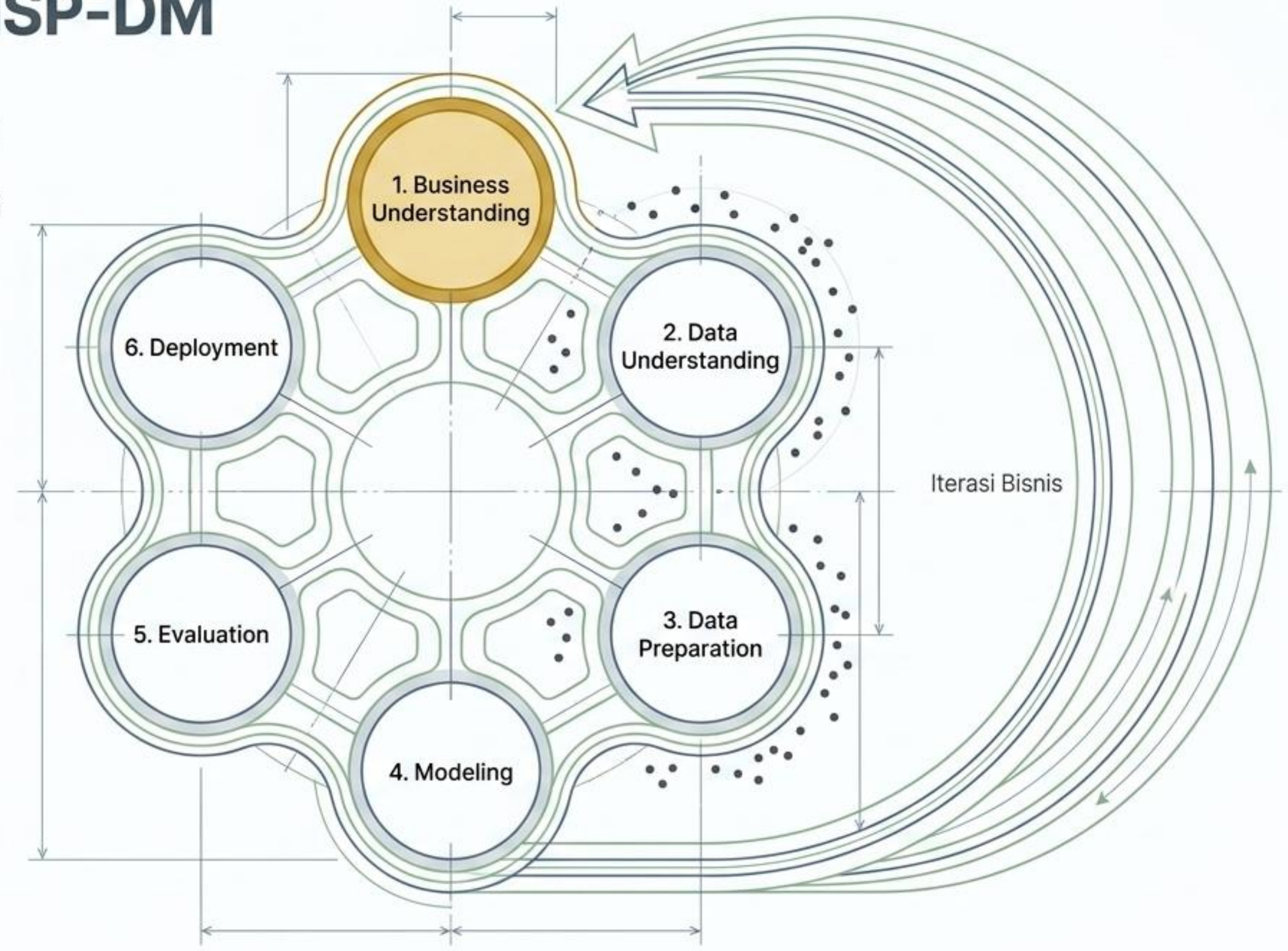
Insight Bisnis

Menghasilkan keputusan strategis yang presisi.

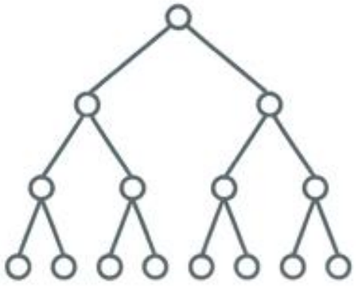
CRISP DM
CROSS-INDUSTRY STANDARD PROCESS - DATA MINING

Alur Proses CRISP-DM

Metodologi standar industri global untuk proyek data. Proses ini bersifat iteratif—selalu dimulai dari pemahaman target bisnis dan berputar kembali untuk penyempurnaan terus-menerus.



Toolkit Teknik Data Mining



Classification

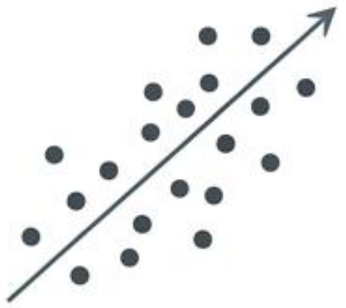
Memprediksi kelas atau kategori spesifik dari data baru berdasarkan data historis.

Clustering

Segmentasi data ke dalam kelompok yang memiliki karakteristik serupa tanpa label awal.

Association

Menemukan aturan dan pola kombinasi barang yang sering muncul bersamaan (pola belanja).



Regression

Memprediksi nilai numerik atau tren kuantitatif di masa depan.

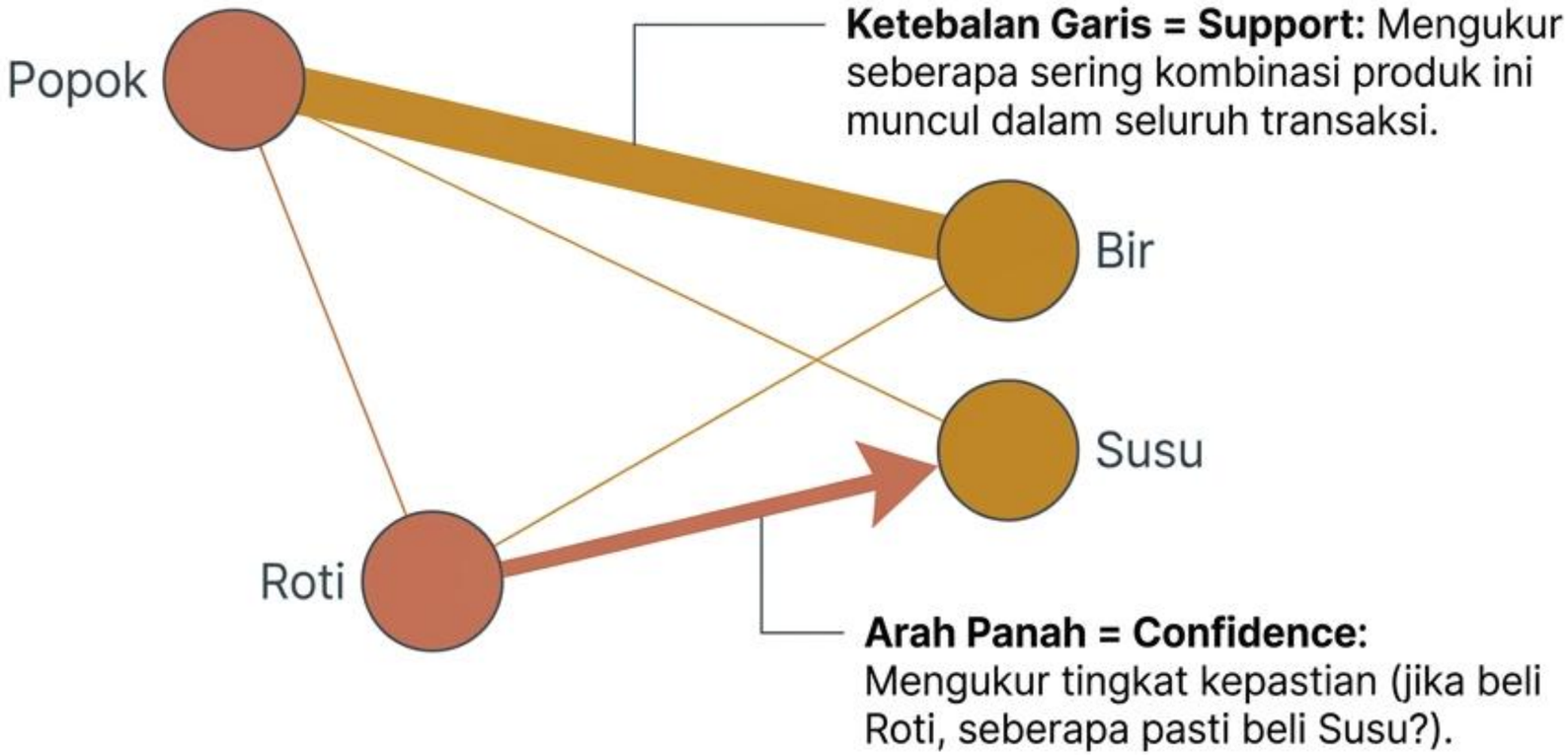
Studi Kasus 1: Analisis Market Basket

Tujuan Bisnis

Mencari pola item yang sering dibeli bersama dari ribuan baris data transaksi pelanggan harian.

Visualisasi Pola Asosiasi

Algoritma tidak menebak; ia menggunakan metrik Support dan Confidence untuk memastikan korelasi yang ditemukan memiliki bobot statistik yang valid.



Insight Bisnis: Eksekusi Analisis Market Basket

1. Optimasi Tata Letak

Menempatkan produk yang memiliki korelasi tinggi secara berdekatan (Cross-merchandising) untuk memicu pembelian impulsif.

2. Bundling Promo

Menciptakan paket promosi berbasis data nyata, bukan sekadar insting, untuk meningkatkan ukuran keranjang belanja (basket size).

3. Rekomendasi Otomatis

Menggerakkan mesin rekomendasi otomatis di platform e-commerce berdasarkan pola riwayat belanja.

Studi Kasus 2: Mencegah Customer Churn

Tujuan Bisnis

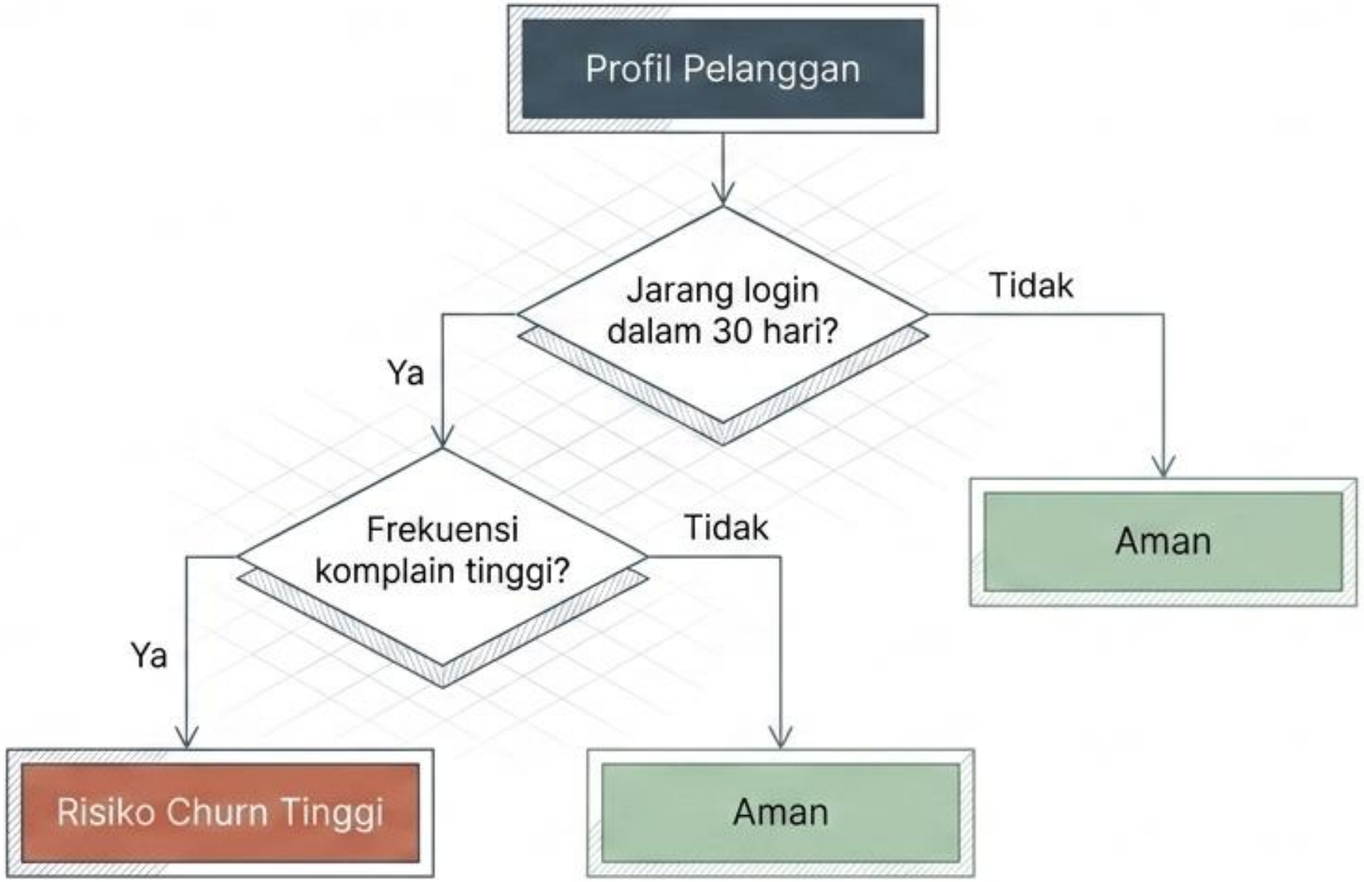
Memprediksi secara akurat pelanggan mana yang berpotensi berhenti berlangganan atau pindah ke kompetitor.

Pendekatan Teknis

Menggunakan teknik Classification (Algoritma Decision Tree atau Logistic Regression) untuk memilah pelanggan ke dalam kategori Aman atau Berisiko.

Logika Prediksi Churn

Pohon Keputusan (Decision Tree) memetakan profil dan perilaku historis untuk **mengidentifikasi Red Flags** secara otomatis sebelum pelanggan pergi.



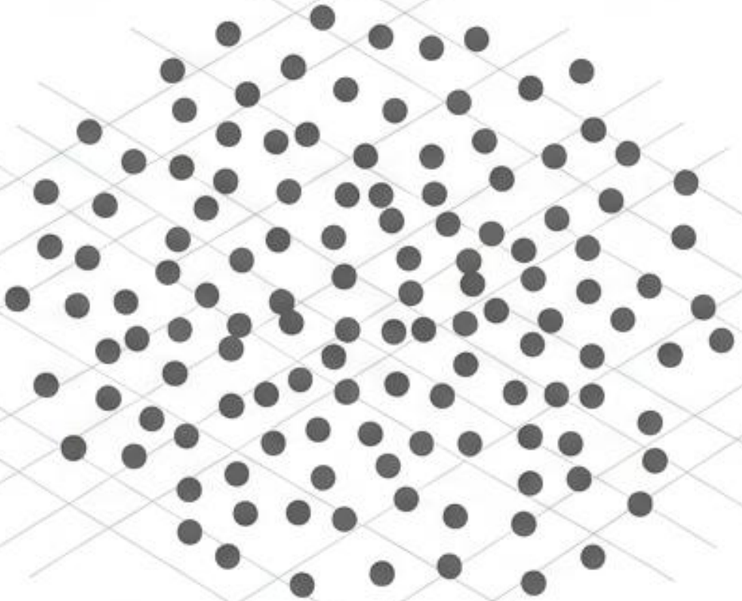
Strategi Retensi Berbasis Prediksi

Tingkat Risiko	Fase	Aksi Bisnis
Risiko Tinggi	Identifikasi Dini	Mengetahui pelanggan berisiko tinggi lebih awal memberi waktu bagi tim untuk bertindak secepatnya.
Risiko Sedang	Promo Penyelamatan (Win-Back)	Memberikan diskon atau insentif yang sangat ditargetkan hanya kepada mereka yang hampir churn, menghemat anggaran promo masal.
Risiko Pasif	Peningkatan Layanan (SLA)	Menaikkan prioritas penanganan keluhan (Customer Service) untuk akun-akun yang terdeteksi tidak puas.

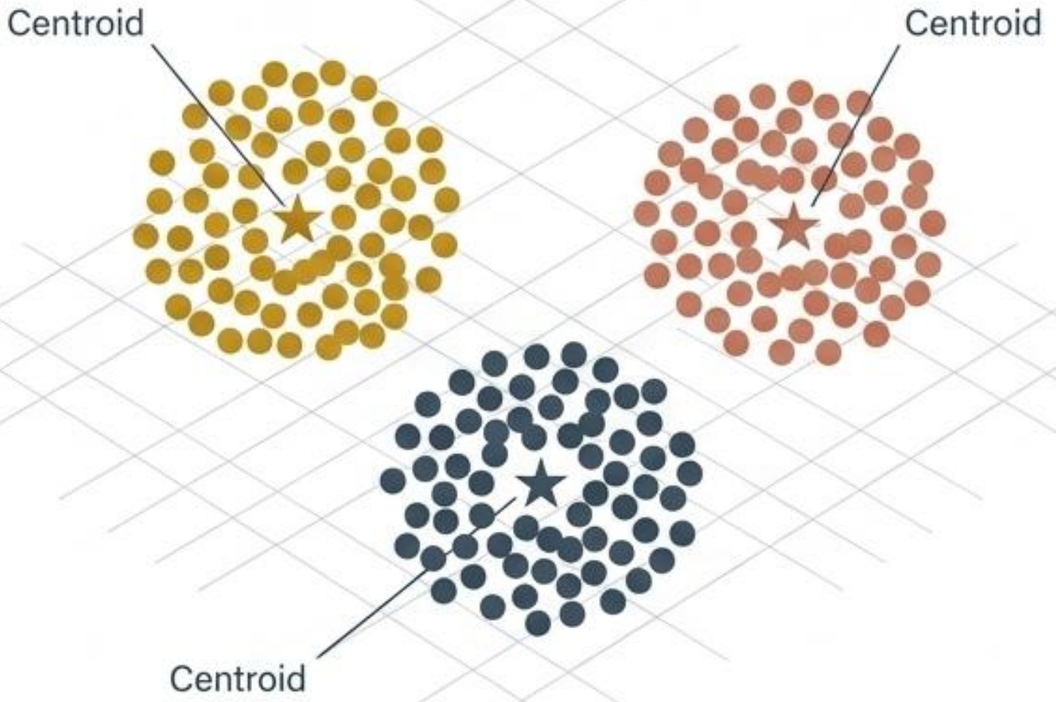
Studi Kasus 3: **Segmentasi Pelanggan** (Clustering)

Tujuan: Mengelompokkan pelanggan berdasarkan perilaku alami mereka.
Pendekatan: Menggunakan algoritma K-Means tanpa label asumsi.

Sebelum K-Means



Sesudah K-Means



Profil Hasil Segmentasi & Aksi Bisnis

Premium

Karakteristik

Nilai transaksi tinggi, loyalitas tinggi.

Aksi Bisnis

Program VIP, layanan eksklusif (Concierge), prioritas akses produk baru.

Reguler

Karakteristik

Konsisten, nilai transaksi rata-rata.

Aksi Bisnis

Program loyalitas poin, penawaran up-selling bertahap.

Pasif

Karakteristik

Aktivitas rendah, dormant.

Aksi Bisnis

Kampanye reaktivasi, diskon pemicu awal.

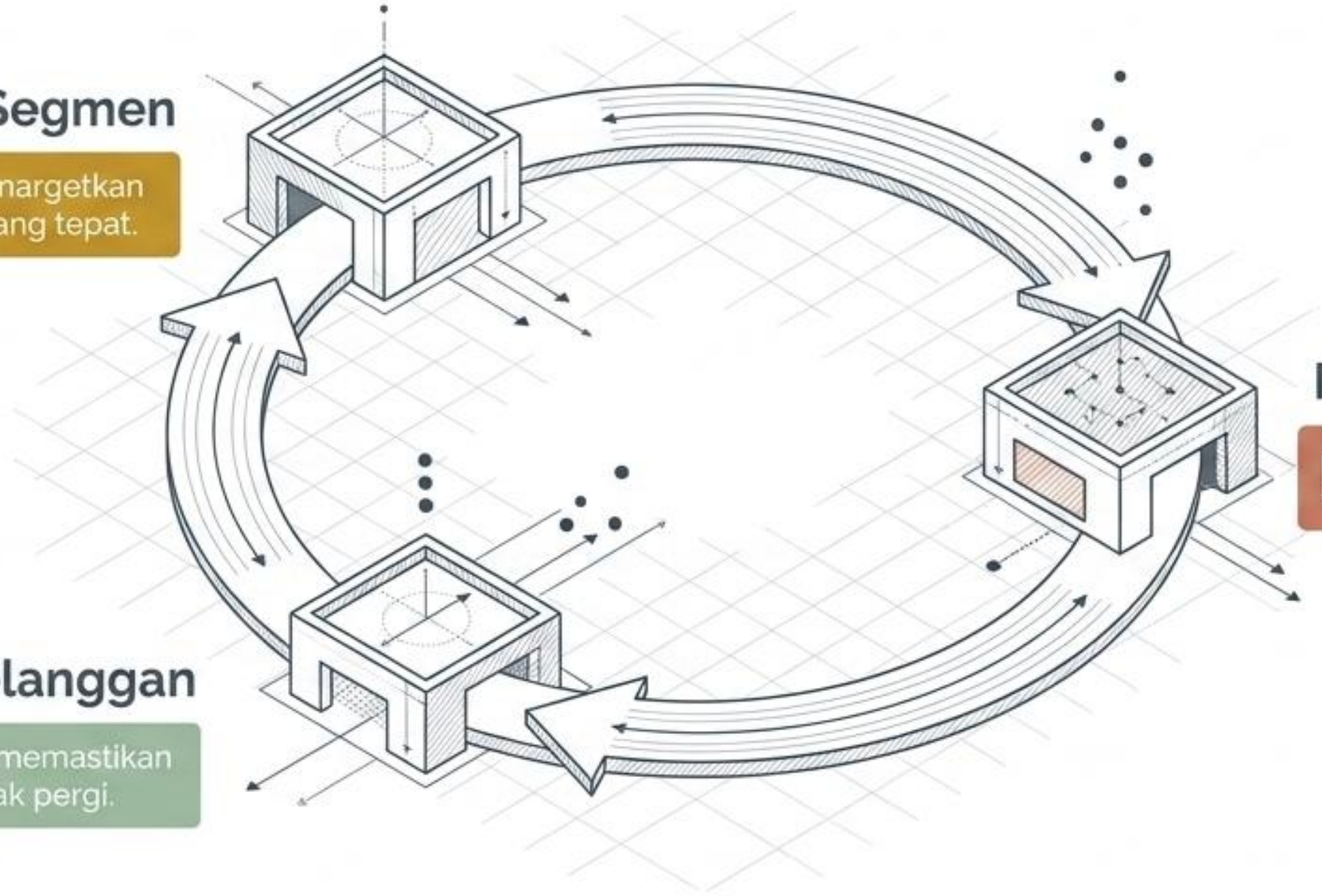
Personalisasi berskala besar: Komunikasi pemasaran disesuaikan secara otomatis dengan karakteristik unik setiap segmen.

Cetak Biru Ekosistem Bisnis Berbasis Data

Sinergi Data Mining. Ketiga teknik ini tidak berdiri sendiri. Bersama-sama, mereka membentuk mesin pertumbuhan yang berkelanjutan.

Akuisisi & Segmen

K-Means menargetkan pelanggan yang tepat.



Maksimalkan Nilai

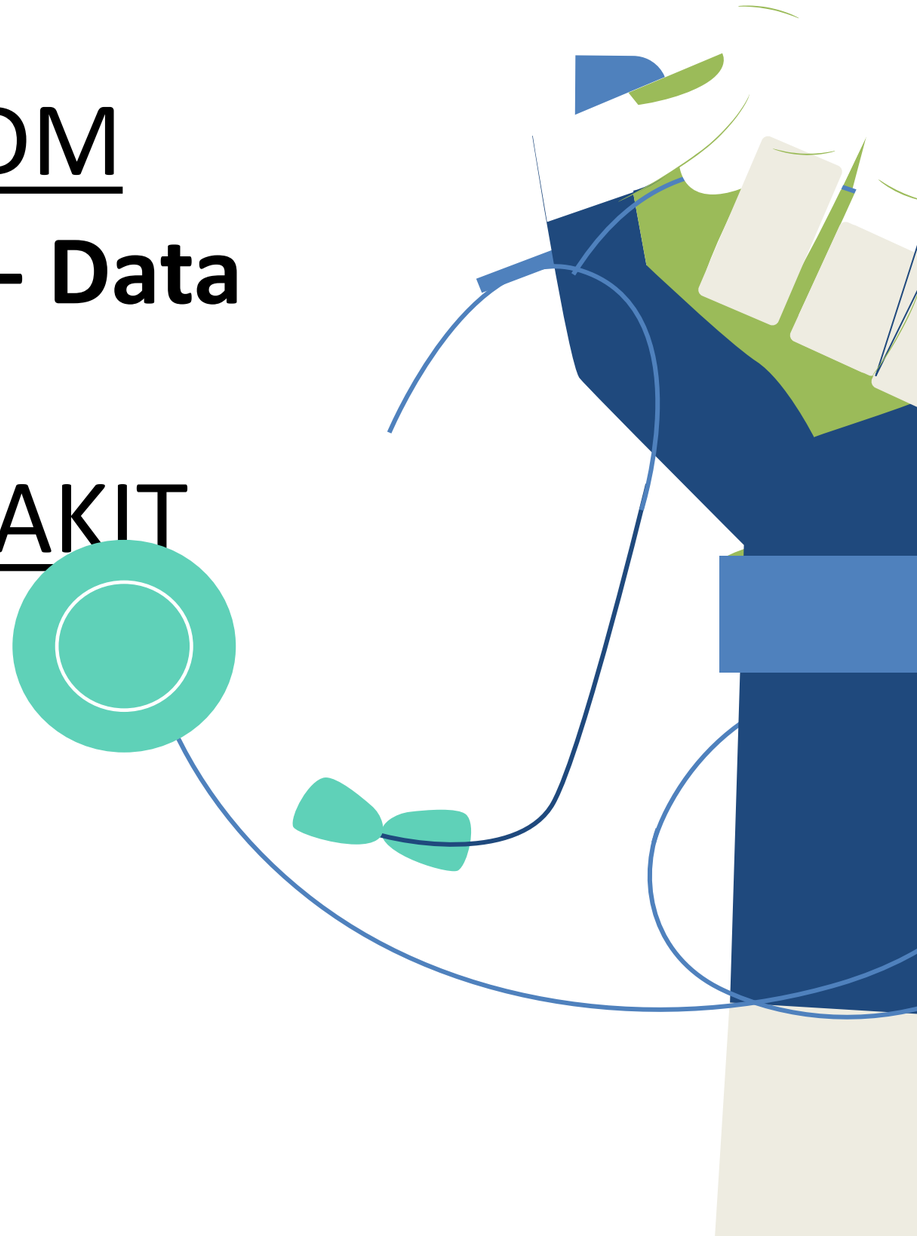
Association mengoptimalkan ukuran keranjang belanja.

Retensi Pelanggan

Classification memastikan pelanggan tidak pergi.

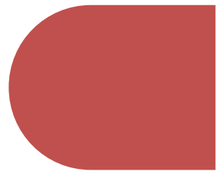
Contoh

PENERAPAN METODE CRISP-DM (**Cross-Industry Standard Process - Data Mining**) DALAM SISTEM PREDIKSI PENYAKIT



BUSINESS UNDERSTANDING

- Menurut *World Health Organization* (WHO), stroke merupakan penyakit dengan tingkat kematian tertinggi ke-2 secara global.
- Oleh karena itu diperlukan upaya preventif untuk mengurangi angka kejadian stroke, salah satunya dengan pemodelan sistem diagnosa dini menggunakan *decision tree classifier* dan *random forest classifier.*
- Tujuan : untuk mengetahui faktor penyebab yang berpengaruh dalam diagnosa stroke.



DATA UNDERSTANDING

Dataset diperoleh dari Kaggle dengan banyak data sebesar 5110 data dan terdiri dari 12 variable berikut :

DATA PREPARATION

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	No	Yes	Yes	Private	Urban	228.69	36.6	formerly smoked	Yes
1	51676	Female	61.0	No	No	Yes	Self-employed	Rural	202.21	NaN	never smoked	Yes
2	31112	Male	80.0	No	Yes	Yes	Private	Rural	105.92	32.5	never smoked	Yes
3	60182	Female	49.0	No	No	Yes	Private	Urban	171.23	34.4	smokes	Yes
4	1665	Female	79.0	Yes	No	Yes	Self-employed	Rural	174.12	24.0	never smoked	Yes

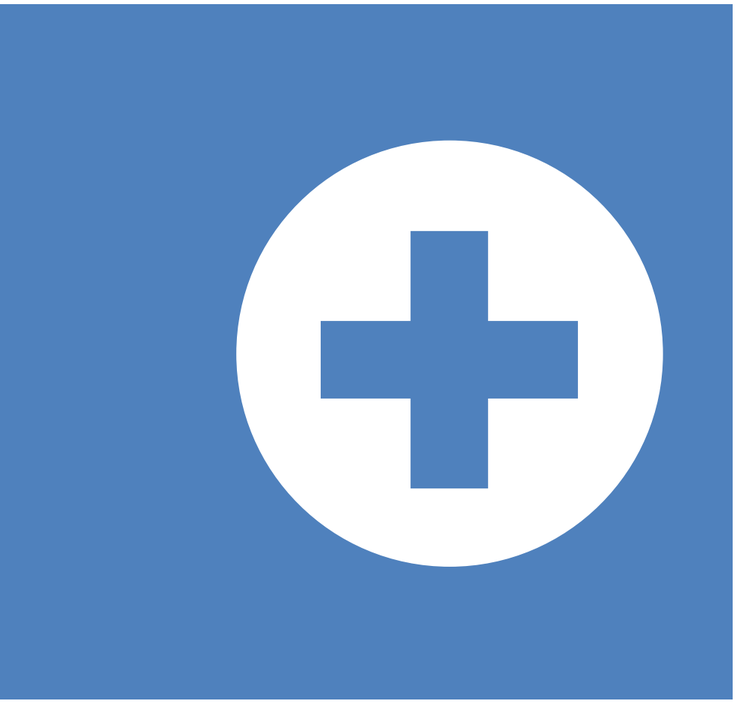
```
[ ] df_data.shape
```

```
(5110, 12)
```

2020

DATA PREPARATION

TABEL



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   id                 5110 non-null   int64  
 1   gender             5110 non-null   object 
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   object 
 4   heart_disease      5110 non-null   object 
 5   ever_married       5110 non-null   object 
 6   work_type          5110 non-null   object 
 7   Residence_type     5110 non-null   object 
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object 
 11  stroke             5110 non-null   object 
dtypes: float64(3), int64(1), object(8)
memory usage: 479.2+ KB
```

```
id                   0
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                201
smoking_status       0
stroke               0
dtype: int64
```

DATA PREPARATION

HANDLING MISSING VALUE

Persentase Data yang Menghilang

```
Persentase missing value setiap variabel:

bmi    3.93
dtype: float64
```

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
1	51676	Female	61.0	No	No	Yes	Self-employed	Rural	202.21	NaN	never smoked	Yes
8	27419	Female	59.0	No	No	Yes	Private	Rural	76.15	NaN	Unknown	Yes
13	8213	Male	78.0	No	Yes	Yes	Private	Urban	219.84	NaN	Unknown	Yes
19	25226	Male	57.0	No	Yes	No	Govt_job	Urban	217.08	NaN	Unknown	Yes
27	61843	Male	58.0	No	No	Yes	Private	Rural	189.84	NaN	Unknown	Yes
...	...	...	...	...	...	...	...	...	...	...	...	...
5039	42007	Male	41.0	No	No	No	Private	Rural	70.15	NaN	formerly smoked	No
5048	28788	Male	40.0	No	No	Yes	Private	Urban	191.15	NaN	smokes	No
5093	32235	Female	45.0	Yes	No	Yes	Govt_job	Rural	95.02	NaN	smokes	No
5099	7293	Male	40.0	No	No	Yes	Private	Rural	83.94	NaN	smokes	No
5105	18234	Female	80.0	Yes	No	Yes	Private	Urban	83.75	NaN	never smoked	No

201 rows × 12 columns

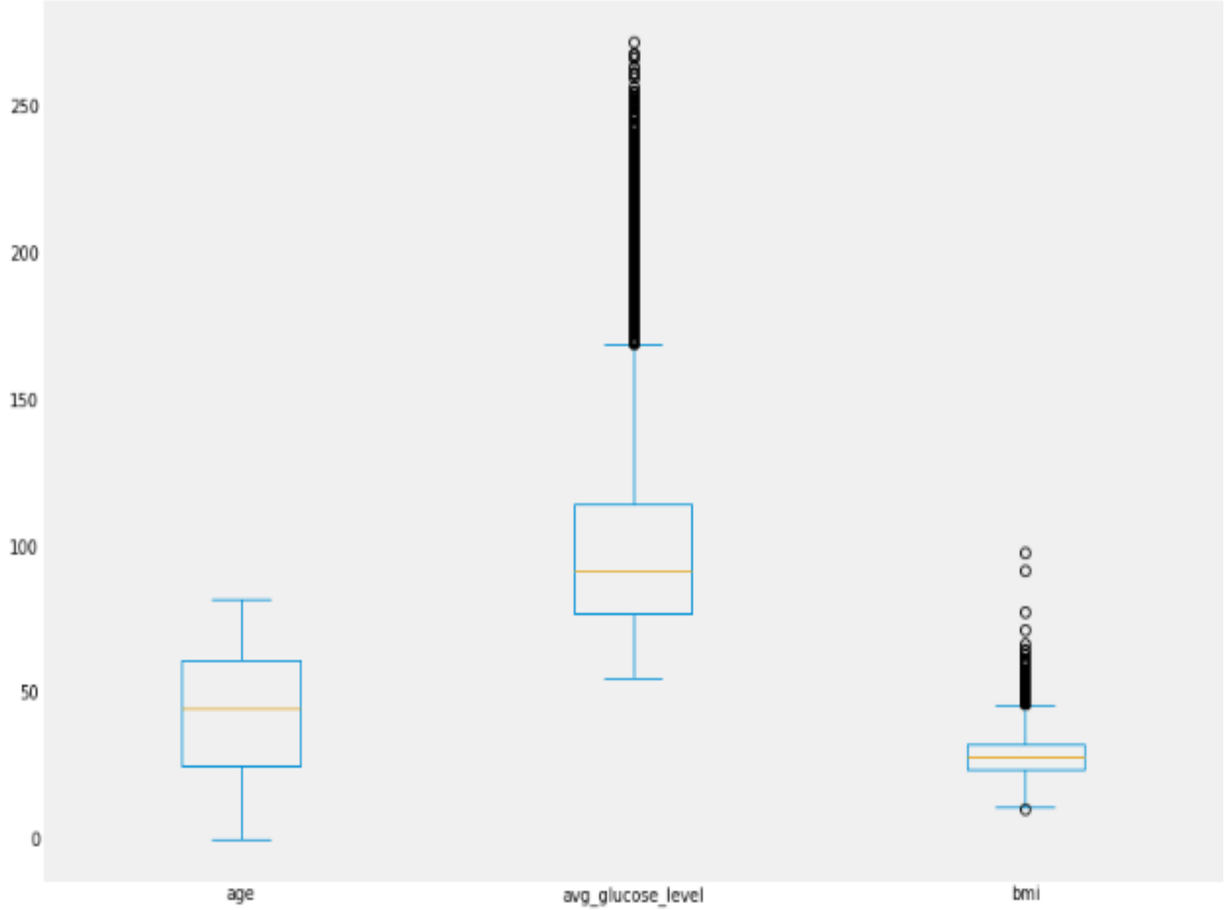
DATA PREPARATION

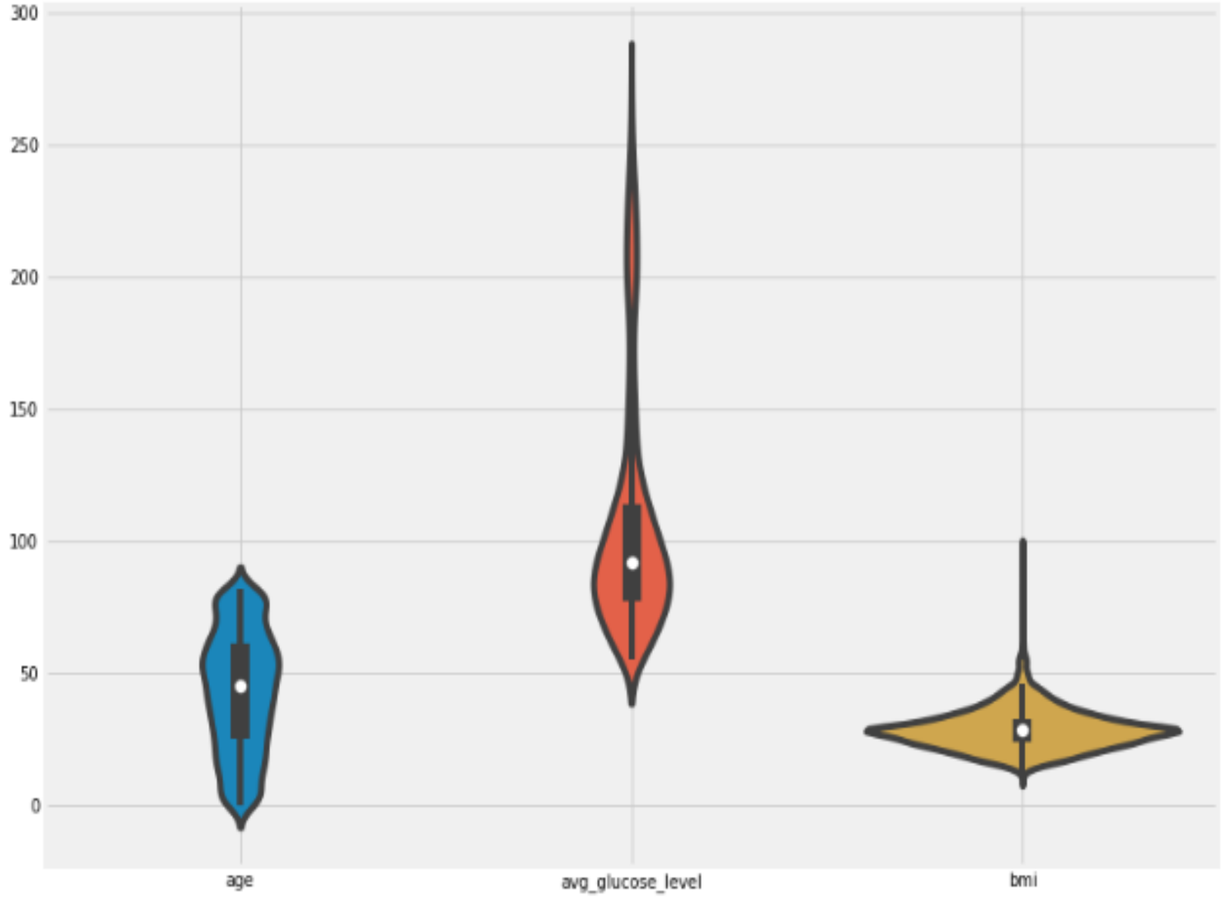
IMPUTE NaN dengan CENTRAL TENDENCY

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 5110 non-null   string
 1   gender             5110 non-null   object
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   object
 4   heart_disease      5110 non-null   object
 5   ever_married       5110 non-null   object
 6   work_type          5110 non-null   object
 7   Residence_type     5110 non-null   object
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                5110 non-null   float64
 10  smoking_status     5110 non-null   object
 11  stroke             5110 non-null   object
dtypes: float64(3), object(8), string(1)
memory usage: 479.2+ KB
```

DATA PREPARATION

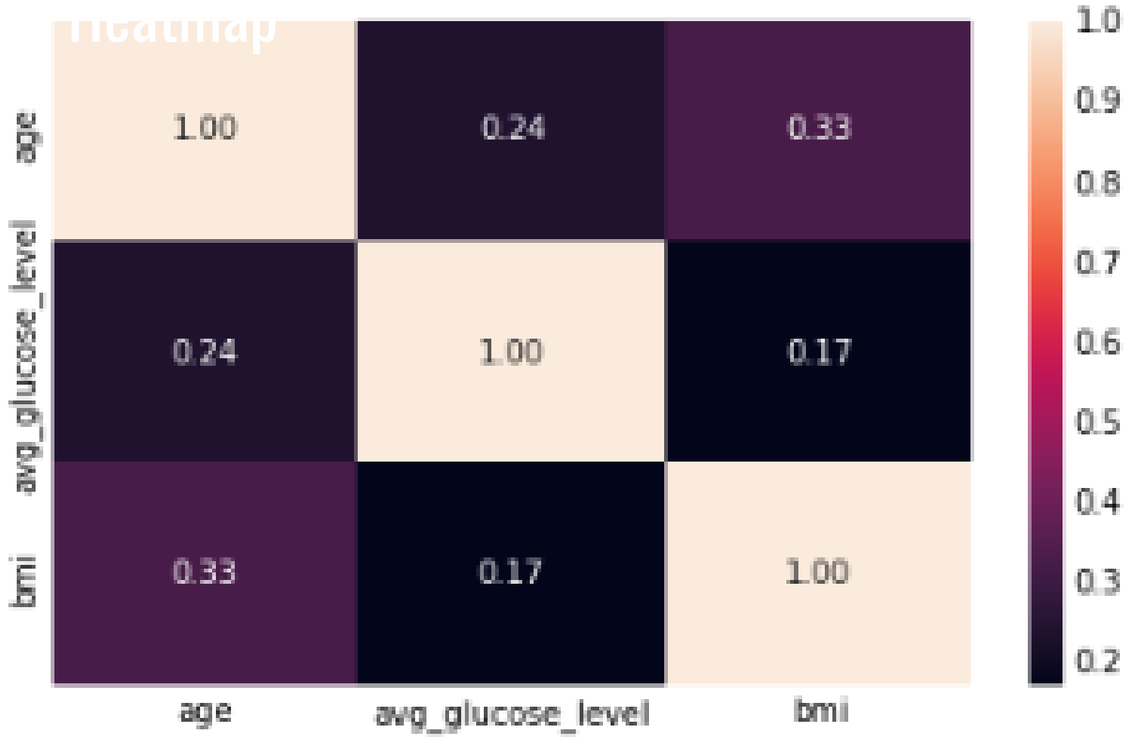
BOXPLOT





DATA PREPARATION

CORRELATION



	age	avg_glucose_level	bmi
age	1.000000	0.238171	0.325942
avg_glucose_level	0.238171	1.000000	0.168751
bmi	0.325942	0.168751	1.000000

DATA PREPARATION

EDA on CATEGORICAL FEATURE

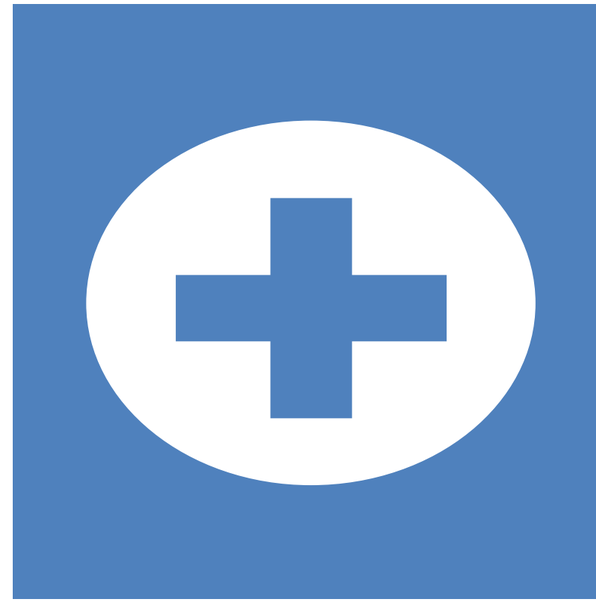
	gender	hypertension	heart_disease	ever_married	work_type	Residence_type	smoking_status	stroke
0	Male	No	Yes	Yes	Private	Urban	formerly smoked	Yes
1	Female	No	No	Yes	Self-employed	Rural	never smoked	Yes
2	Male	No	Yes	Yes	Private	Rural	never smoked	Yes
3	Female	No	No	Yes	Private	Urban	smokes	Yes
4	Female	Yes	No	Yes	Self-employed	Rural	never smoked	Yes
5	Male	No	No	Yes	Private	Urban	formerly smoked	Yes
6	Male	Yes	Yes	Yes	Private	Rural	never smoked	Yes
7	Female	No	No	No	Private	Urban	never smoked	Yes
8	Female	No	No	Yes	Private	Rural	Unknown	Yes
9	Female	No	No	Yes	Private	Urban	Unknown	Yes
10	Female	Yes	No	Yes	Private	Rural	never smoked	Yes
11	Female	No	Yes	Yes	Govt_job	Rural	smokes	Yes
12	Female	No	No	Yes	Private	Urban	smokes	Yes
13	Male	No	Yes	Yes	Private	Urban	Unknown	Yes
14	Female	No	Yes	Yes	Private	Urban	never smoked	Yes
15	Female	Yes	No	Yes	Self-employed	Rural	never smoked	Yes
16	Male	No	Yes	Yes	Private	Urban	smokes	Yes
17	Male	Yes	No	Yes	Private	Urban	smokes	Yes
18	Female	No	No	No	Private	Urban	never smoked	Yes
19	Male	No	Yes	No	Govt_job	Urban	Unknown	Yes

	gender	hypertension	heart_disease	ever_married	work_type	Residence_type	smoking_status	stroke
5090	Male	No	No	No	Govt_job	Rural	smokes	No
5091	Male	Yes	No	Yes	Private	Rural	never smoked	No
5092	Male	No	No	Yes	Govt_job	Urban	never smoked	No
5093	Female	Yes	No	Yes	Govt_job	Rural	smokes	No
5094	Male	No	No	No	children	Urban	Unknown	No
5095	Male	No	No	No	children	Rural	Unknown	No
5096	Male	No	No	Yes	Govt_job	Rural	never smoked	No
5097	Male	No	No	Yes	Self-employed	Urban	Unknown	No
5098	Male	No	No	No	children	Urban	Unknown	No
5099	Male	No	No	Yes	Private	Rural	smokes	No
5100	Male	Yes	No	Yes	Self-employed	Rural	never smoked	No
5101	Female	No	No	Yes	Private	Urban	Unknown	No
5102	Female	No	No	Yes	Private	Rural	never smoked	No
5103	Female	No	No	No	Private	Urban	Unknown	No
5104	Female	No	No	No	children	Rural	Unknown	No
5105	Female	Yes	No	Yes	Private	Urban	never smoked	No
5106	Female	No	No	Yes	Self-employed	Urban	never smoked	No
5107	Female	No	No	Yes	Self-employed	Rural	never smoked	No
5108	Male	No	No	Yes	Private	Rural	formerly smoked	No
5109	Female	No	No	Yes	Govt_job	Urban	Unknown	No

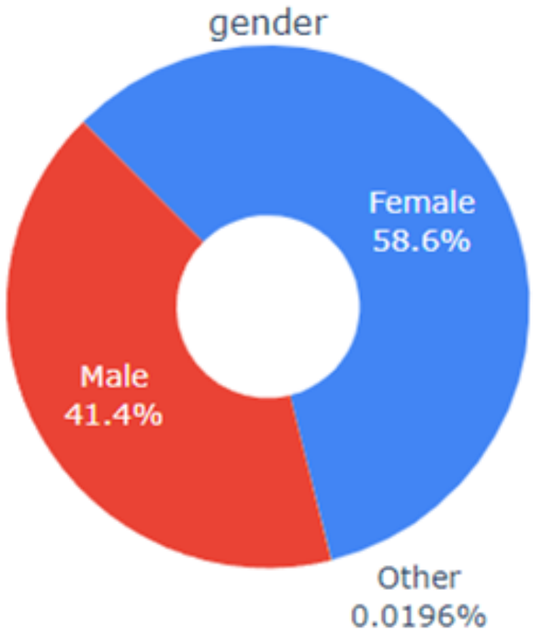
DATA PREPARATION

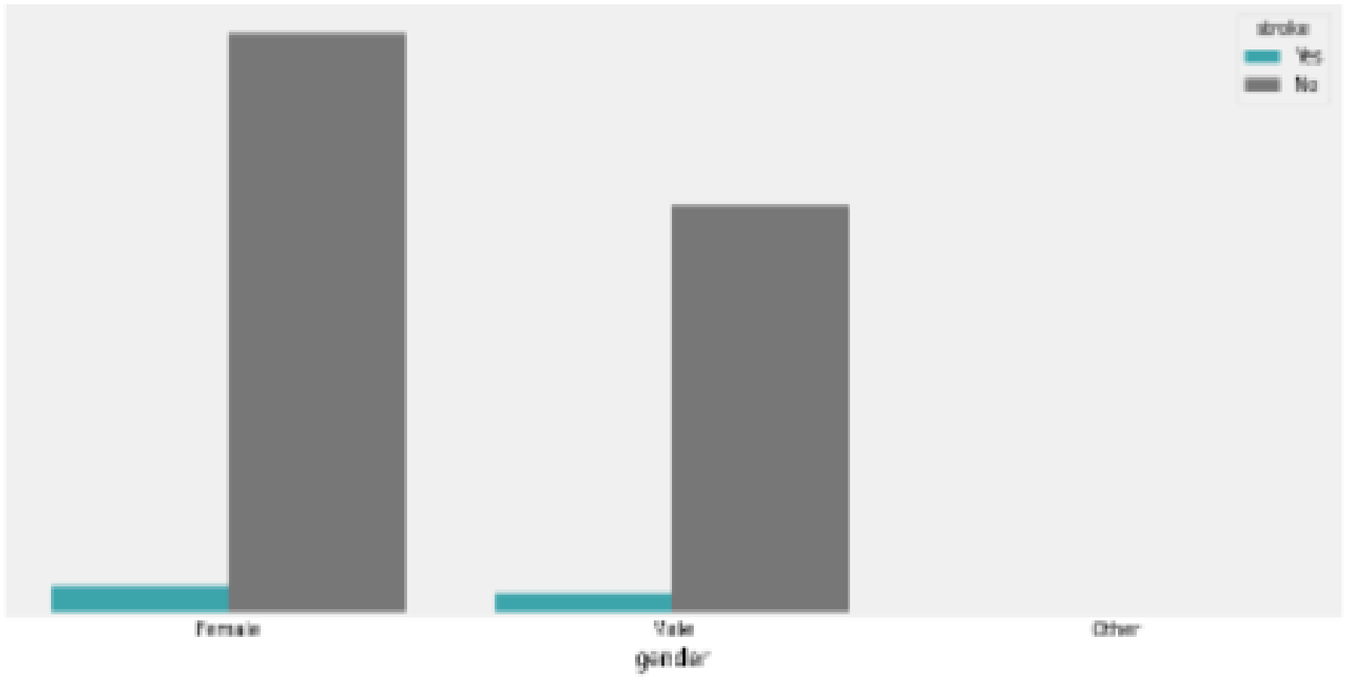
```
gender             3
hypertension       2
heart_disease      2
ever_married       2
work_type          5
Residence_type     2
smoking_status     4
stroke             2
dtype: int64
```

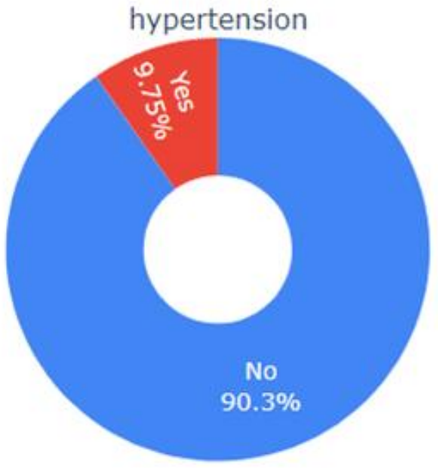
```
Female    2994
Male      2115
Other        1
Name: gender, dtype: int64
------
No     4612
Yes     498
Name: hypertension, dtype: int64
------
No     4834
Yes     276
Name: heart_disease, dtype: int64
------
Yes    3353
No     1757
Name: ever_married, dtype: int64
------
Private          2925
Self-employed     819
children          687
Govt_job          657
Never_worked       22
Name: work_type, dtype: int64
------
Urban    2596
Rural    2514
Name: Residence_type, dtype: int64
------
never smoked       1892
Unknown            1544
formerly smoked     885
smokes              789
Name: smoking_status, dtype: int64
------
No     4861
Yes     249
Name: stroke, dtype: int64
------
```

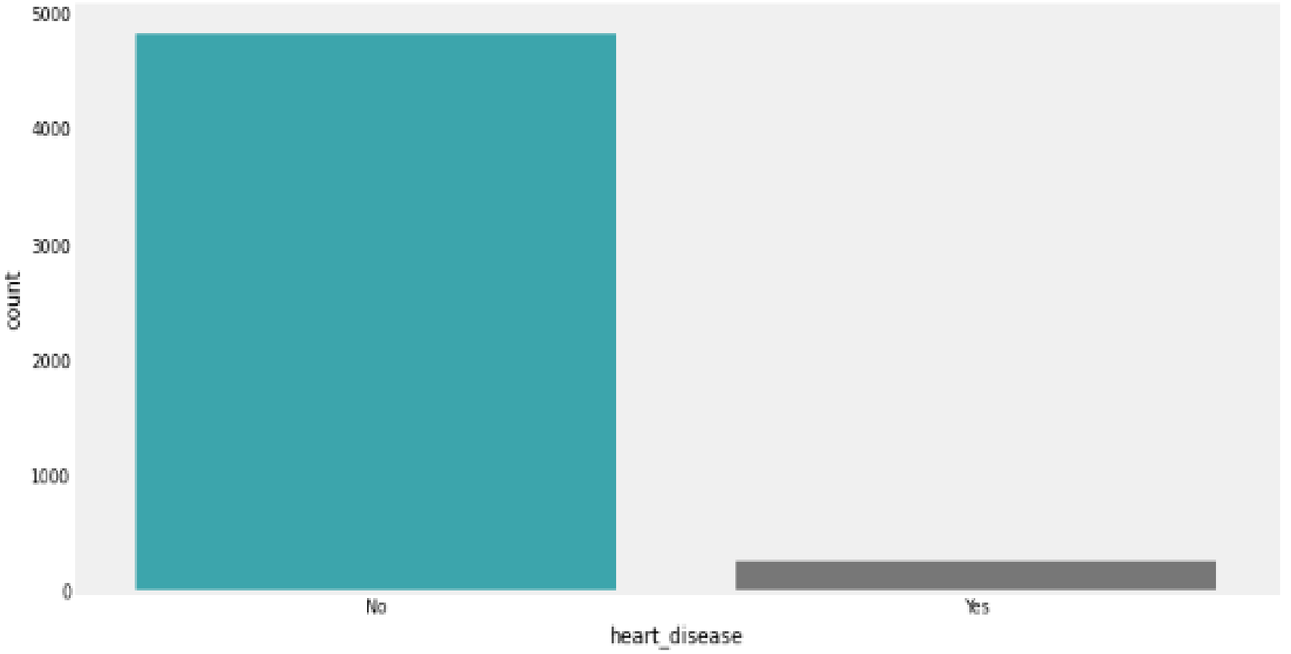


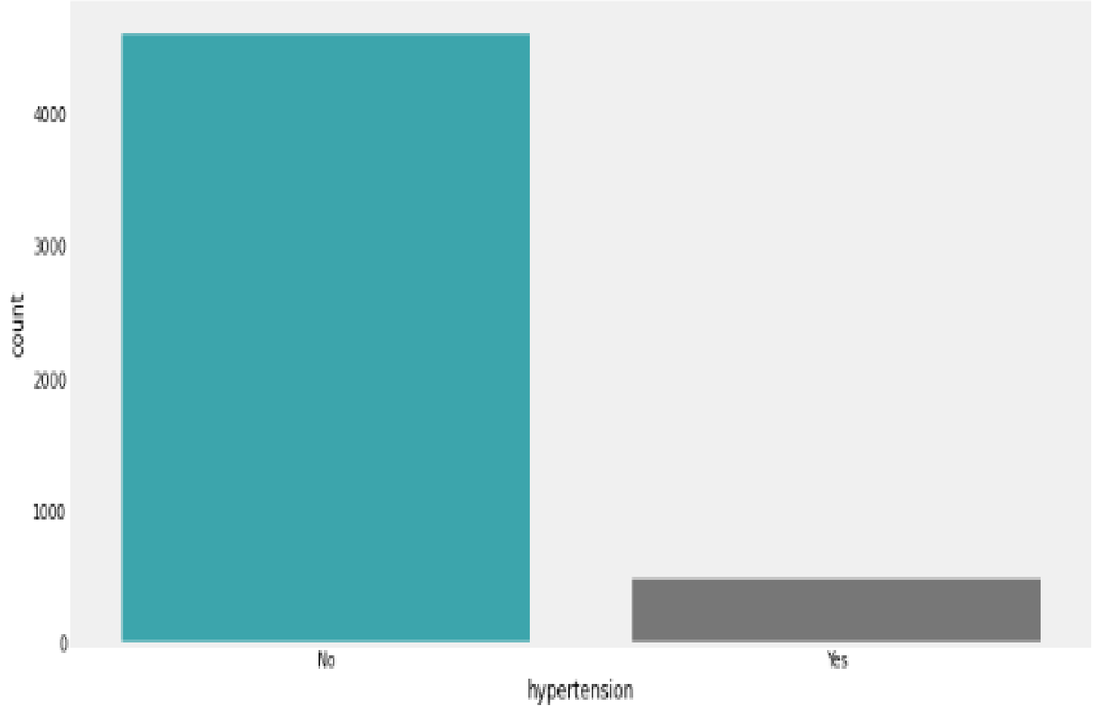
DATA PREPARATION/VISUALISASI

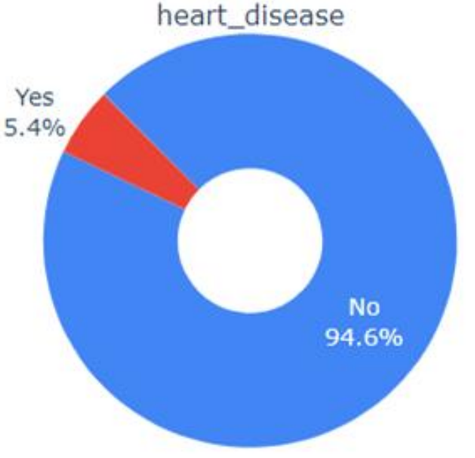


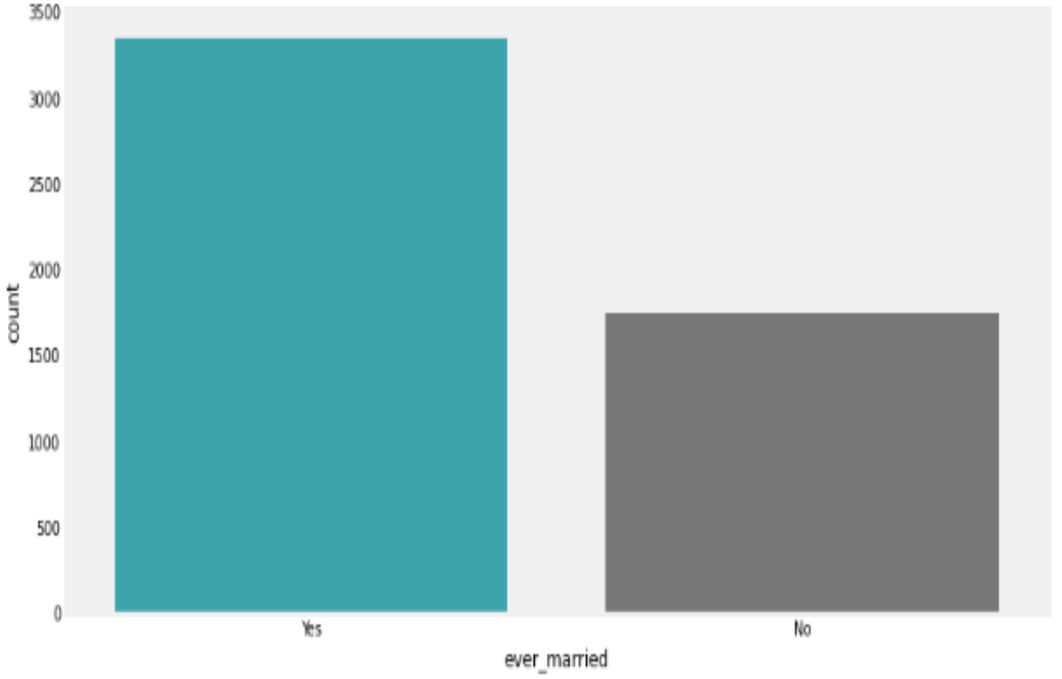


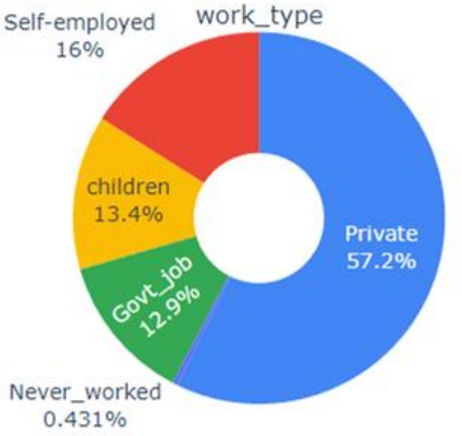

hypertension
Yes
9.75%
No
90.3%

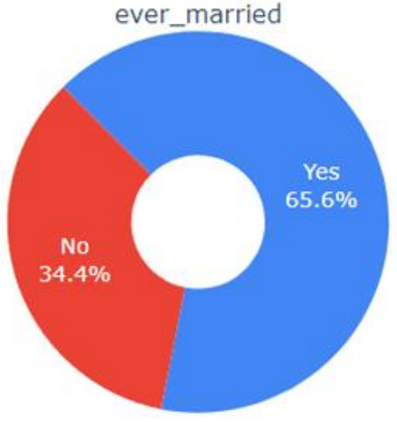

5000
4000
3000
2000
1000
0
count
No
Yes
heart_disease

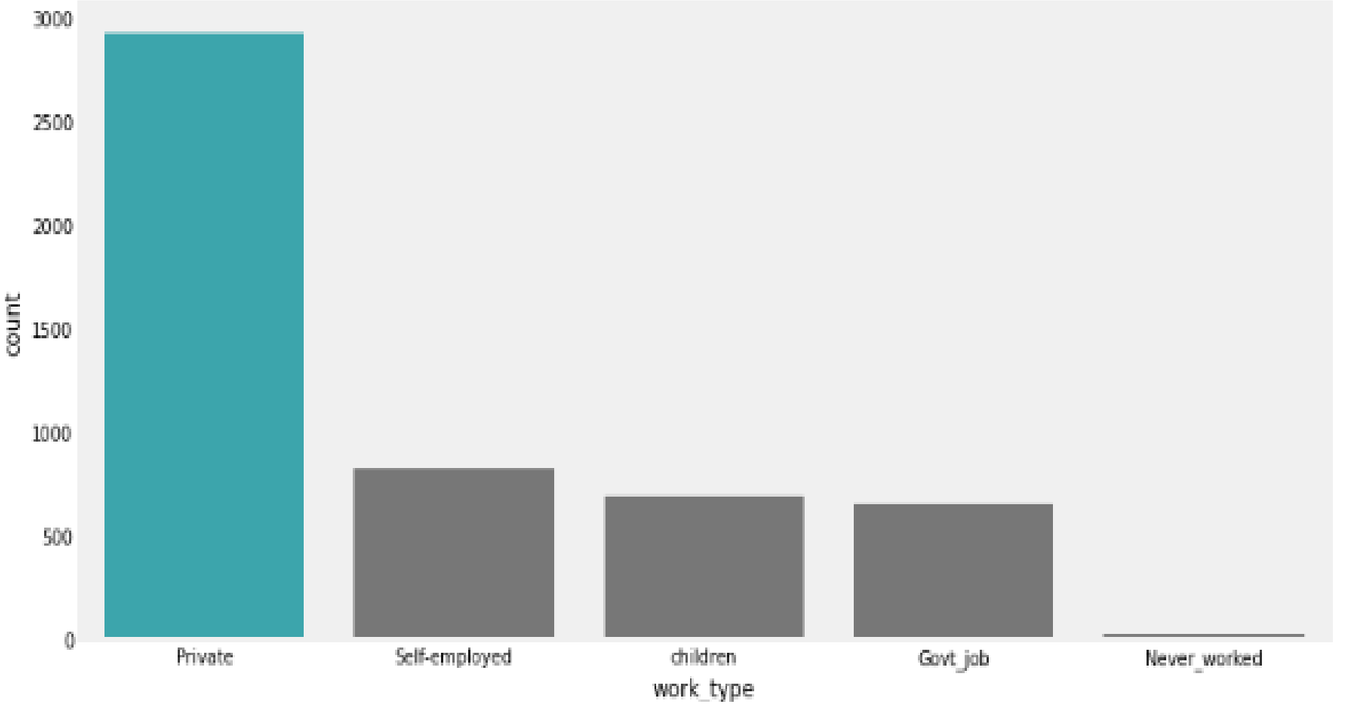

4000
3000
2000
1000
0
count
No
Yes
hypertension

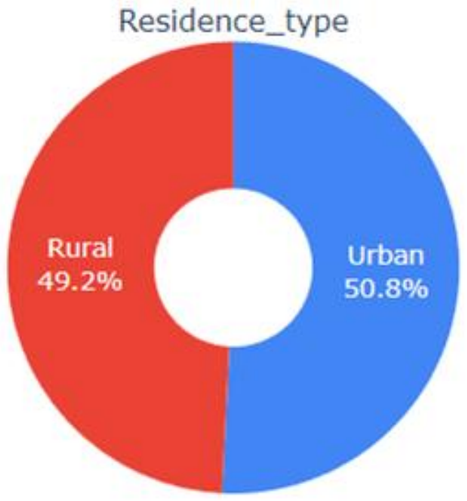

heart_disease
Yes
5.4%
No
94.6%

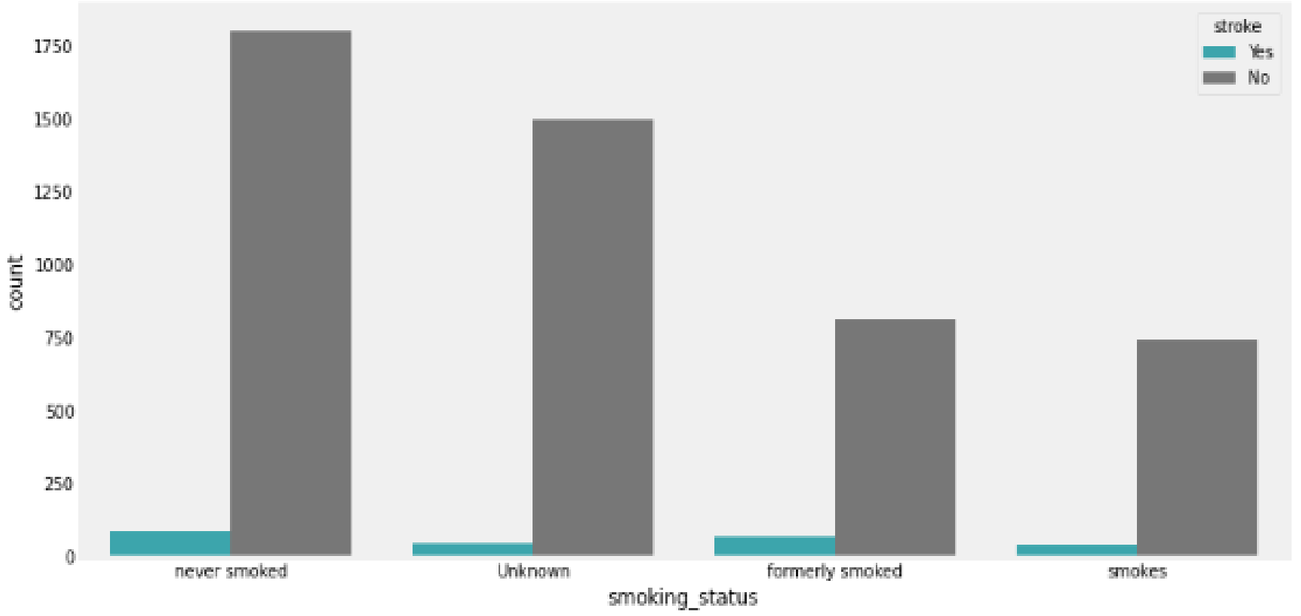

3500
3000
2500
2000
1500
1000
500
0
count
Yes
No
ever_married

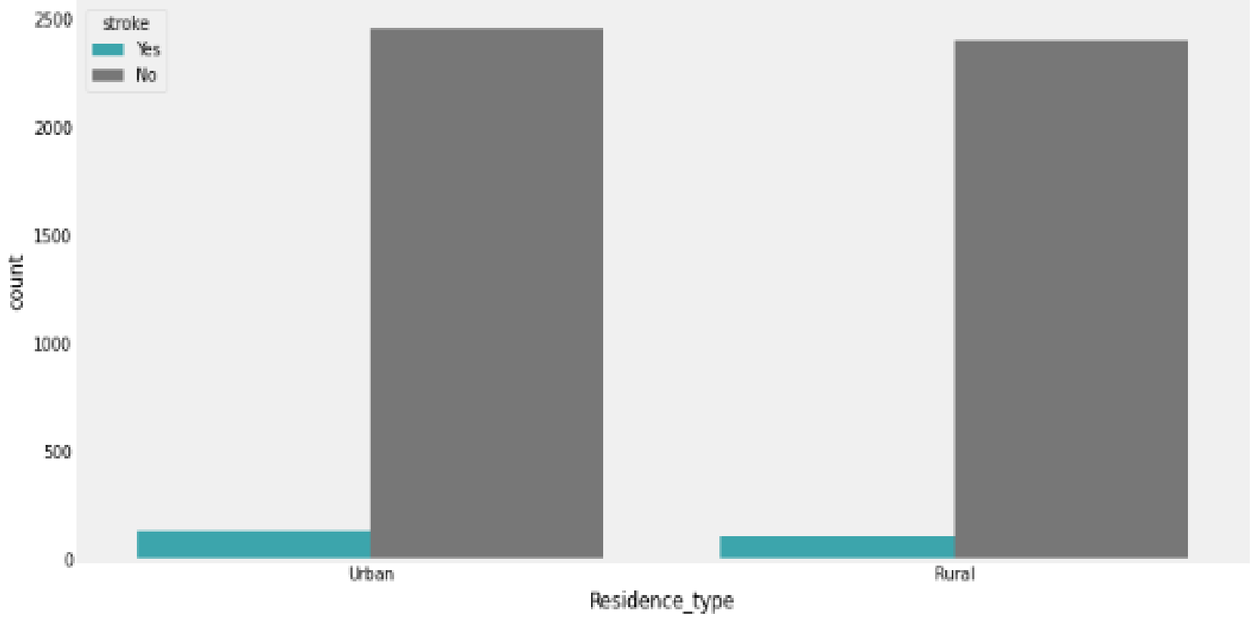

work_type
Self-employed
16%
children
13.4%
Private
57.2%
Govt_job
12.9%
Never_worked
0.431%

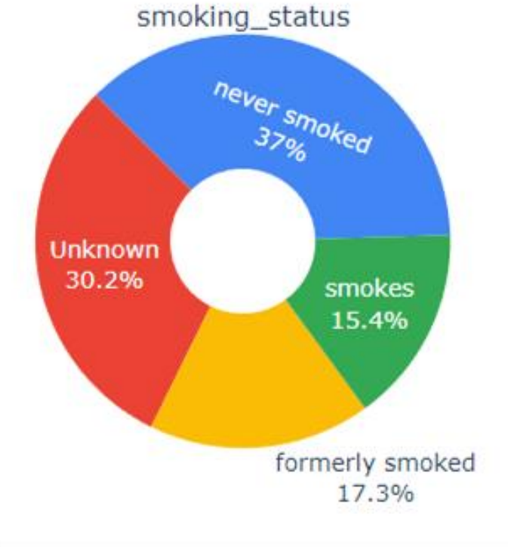

ever_married
Yes
65.6%
No
34.4%

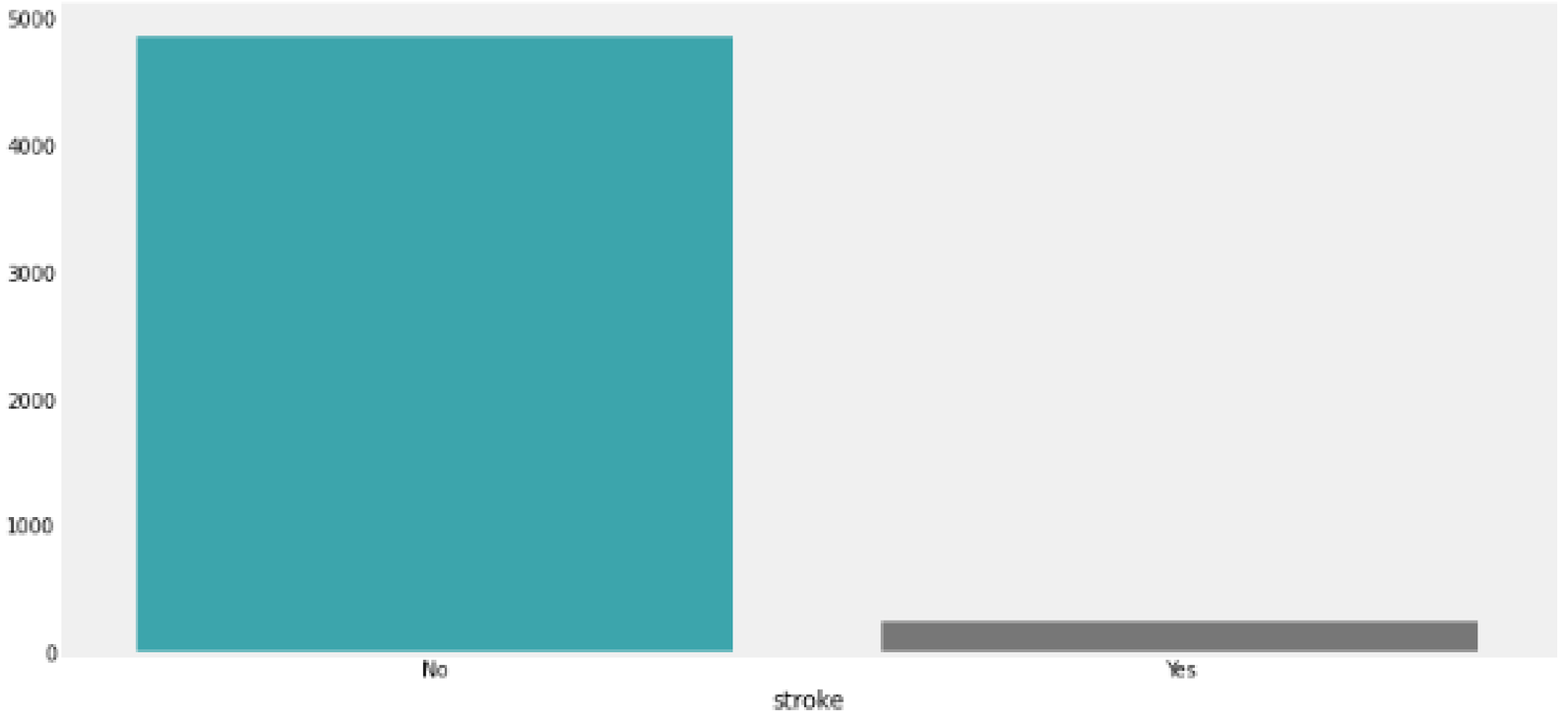

3000
2500
2000
1500
1000
500
0
count
Private
Self-employed
children
Govt_job
Never_worked
work_type

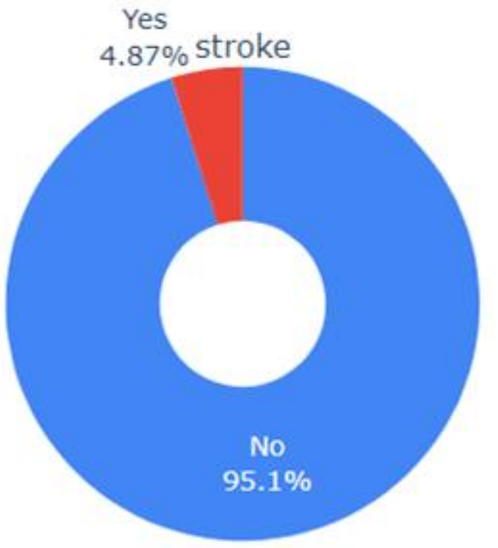

Residence_type
Rural
49.2%
Urban
50.8%


stroke
Yes
No
1750
1500
1250
1000
750
500
250
0
count
never smoked
Unknown
formerly smoked
smokes
smoking_status


stroke
Yes
No
2500
2000
1500
1000
500
0
count
Urban
Rural
Residence_type


smoking_status
never smoked
37%
Unknown
30.2%
smokes
15.4%
formerly smoked
17.3%


5000
4000
3000
2000
1000
0
No
Yes
stroke


Yes
4.87%
stroke
No
95.1%

DATA PREPARATION

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	No	Yes	Yes	Private	Urban	228.69	36.600000	formerly smoked	Yes
1	51676	Female	61.0	No	No	Yes	Self-employed	Rural	202.21	28.893237	never smoked	Yes
2	31112	Male	80.0	No	Yes	Yes	Private	Rural	105.92	32.500000	never smoked	Yes
3	60182	Female	49.0	No	No	Yes	Private	Urban	171.23	34.400000	smokes	Yes
4	1665	Female	79.0	Yes	No	Yes	Self-employed	Rural	174.12	24.000000	never smoked	Yes
...	...	...	...	...	...	...	...	...	...	...	...	...
5105	18234	Female	80.0	Yes	No	Yes	Private	Urban	83.75	28.893237	never smoked	No
5106	44873	Female	81.0	No	No	Yes	Self-employed	Urban	125.20	40.000000	never smoked	No
5107	19723	Female	35.0	No	No	Yes	Self-employed	Rural	82.99	30.600000	never smoked	No
5108	37544	Male	51.0	No	No	Yes	Private	Rural	166.29	25.600000	formerly smoked	No
5109	44679	Female	44.0	No	No	Yes	Govt_job	Urban	85.28	26.200000	Unknown	No

5110 rows × 12 columns

DATA PREPARATION

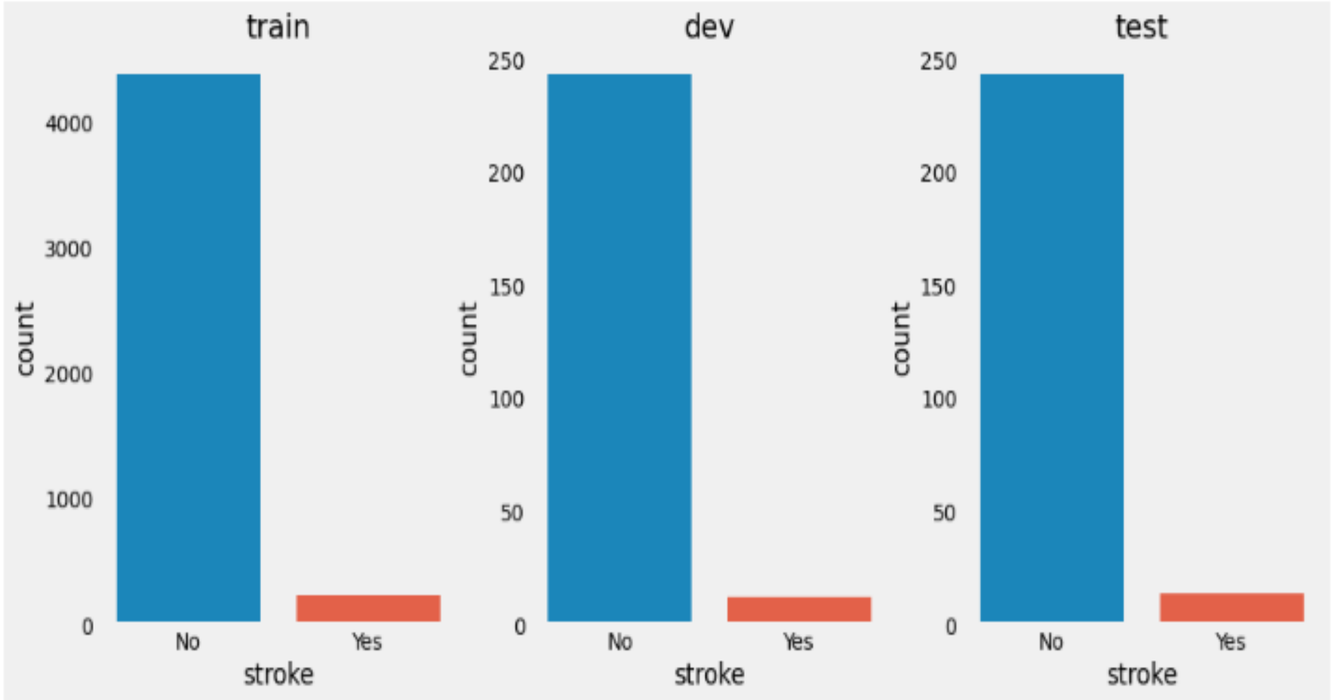
	work_type_Govt_job	work_type_Never_worked	work_type_Private	work_type_Self-employed	work_type_children	Residence_type_Rural	Residence_type_Urban
0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
1	0.0	0.0	1.0	0.0	0.0	0.0	1.0
2	0.0	0.0	1.0	0.0	0.0	0.0	1.0
3	0.0	0.0	1.0	0.0	0.0	1.0	0.0
4	0.0	0.0	1.0	0.0	0.0	1.0	0.0
...	...	...	...	...	...	...	...
4594	0.0	0.0	1.0	0.0	0.0	0.0	1.0
4595	0.0	0.0	1.0	0.0	0.0	0.0	1.0
4596	0.0	0.0	1.0	0.0	0.0	0.0	1.0
4597	0.0	0.0	1.0	0.0	0.0	0.0	1.0
4598	0.0	0.0	0.0	0.0	1.0	0.0	1.0

4599 rows × 7 columns

DATA PREPARATION

	gender	hypertension	heart_disease	ever_married	smoking_status
0	1.0	0.0	0.0	1.0	0.0
1	1.0	0.0	0.0	1.0	3.0
2	0.0	0.0	0.0	0.0	3.0
3	0.0	0.0	0.0	1.0	1.0
4	1.0	0.0	1.0	1.0	3.0
...	...	...	...	...	...
4594	0.0	0.0	0.0	1.0	0.0
4595	1.0	0.0	0.0	1.0	1.0
4596	0.0	0.0	0.0	1.0	1.0
4597	0.0	0.0	0.0	0.0	3.0
4598	0.0	0.0	0.0	0.0	0.0

4599 rows × 5 columns

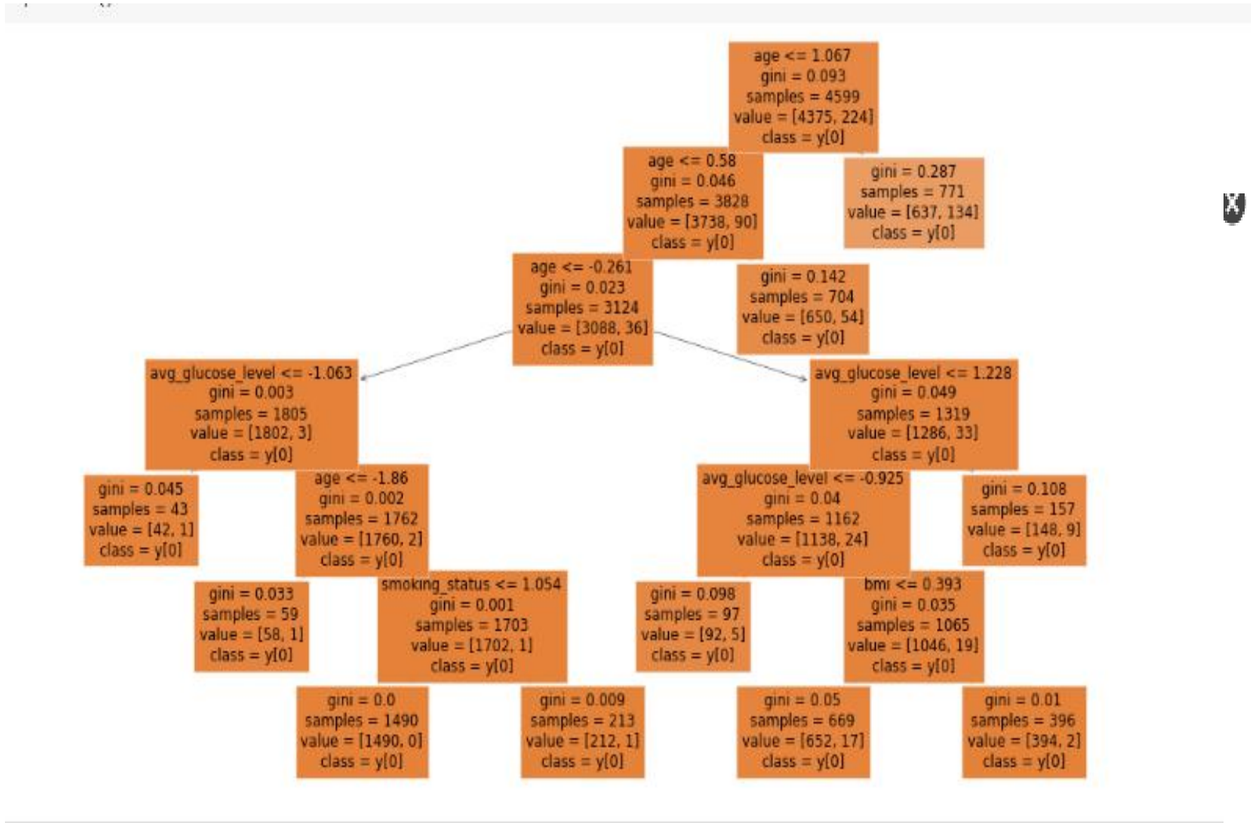


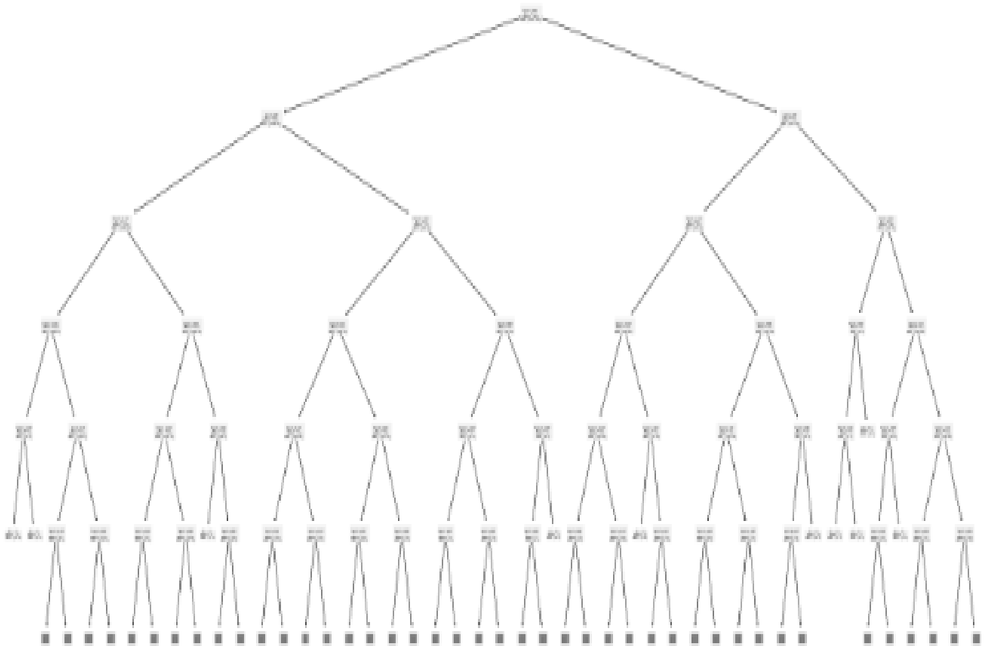
DATA PREPARATION

	age	avg_glucose_level	bmi	work_type_Govt_job	work_type_Never_worked	work_type_Private	work_type_Self-employed	work_type_children	Residence_type_Rural	Residence_type_Urban	gender	hypertension	heart_disease	ever_married	smoking_status
0	65.0	105.61	27.9	0.0	0.0	0.0	1.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	0.0
1	62.0	206.98	36.8	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	1.0	3.0
2	26.0	82.61	28.5	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	3.0
3	61.0	71.40	29.2	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0
4	80.0	82.41	26.3	0.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	1.0	3.0

	age	avg_glucose_level	bmi	work_type_Govt_job	work_type_Never_worked	work_type_Private	work_type_Self-employed	work_type_children	Residence_type_Rural	Residence_type_Urban	gender	hypertension	heart_disease	ever_married	smoking_status
0	0.956497	-0.009357	-0.131530	-0.385116	-0.06933	-1.161389	2.300300	-0.39142	-0.977424	0.977424	1.195608	-0.328927	-0.242284	0.723431	-1.280419
1	0.823707	2.231844	1.020192	-0.385116	-0.06933	0.861038	-0.434726	-0.39142	-0.977424	0.977424	1.195608	-0.328927	-0.242284	0.723431	1.520586
2	-0.769773	-0.517887	-0.053886	-0.385116	-0.06933	0.861038	-0.434726	-0.39142	-0.977424	0.977424	-0.834895	-0.328927	-0.242284	-1.382302	1.520586
3	0.779444	-0.765710	0.036699	-0.385116	-0.06933	0.861038	-0.434726	-0.39142	1.023098	-1.023098	-0.834895	-0.328927	-0.242284	0.723431	-0.346751
4	1.820447	-0.522288	-0.338581	-0.385116	-0.06933	0.861038	-0.434726	-0.39142	1.023098	-1.023098	1.195608	-0.328927	4.127383	0.723431	1.520586

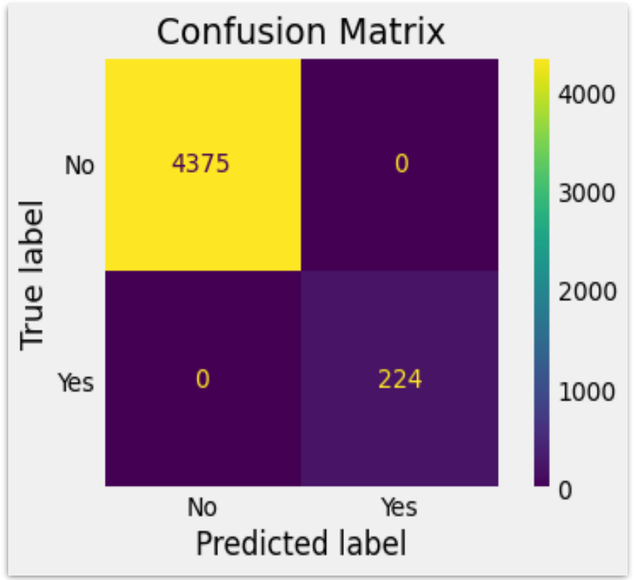
MODELLING


age <= 1.067
gini = 0.093
samples = 4599
value = [4375, 224]
class = y[0]
age <= 0.58
gini = 0.046
samples = 3828
value = [3738, 90]
class = y[0]
gini = 0.287
samples = 771
value = [637, 134]
class = y[0]
age <= -0.261
gini = 0.023
samples = 3124
value = [3088, 36]
class = y[0]
gini = 0.142
samples = 704
value = [650, 54]
class = y[0]
avg_glucose_level <= -1.063
gini = 0.003
samples = 1805
value = [1802, 3]
class = y[0]
avg_glucose_level <= 1.228
gini = 0.049
samples = 1319
value = [1286, 33]
class = y[0]
gini = 0.045
samples = 43
value = [42, 1]
class = y[0]
age <= -1.86
gini = 0.002
samples = 1762
value = [1760, 2]
class = y[0]
avg_glucose_level <= -0.925
gini = 0.04
samples = 1162
value = [1138, 24]
class = y[0]
gini = 0.108
samples = 157
value = [148, 9]
class = y[0]
gini = 0.033
samples = 59
value = [58, 1]
class = y[0]
smoking_status <= 1.054
gini = 0.001
samples = 1703
value = [1702, 1]
class = y[0]
gini = 0.098
samples = 97
value = [92, 5]
class = y[0]
bmi <= 0.393
gini = 0.035
samples = 1065
value = [1046, 19]
class = y[0]
gini = 0.0
samples = 1490
value = [1490, 0]
class = y[0]
gini = 0.009
samples = 213
value = [212, 1]
class = y[0]
gini = 0.05
samples = 669
value = [652, 17]
class = y[0]
gini = 0.01
samples = 396
value = [394, 2]
class = y[0]



TESTING & EVALUATION

01. Data Training - Default Model



Confusion Matrix

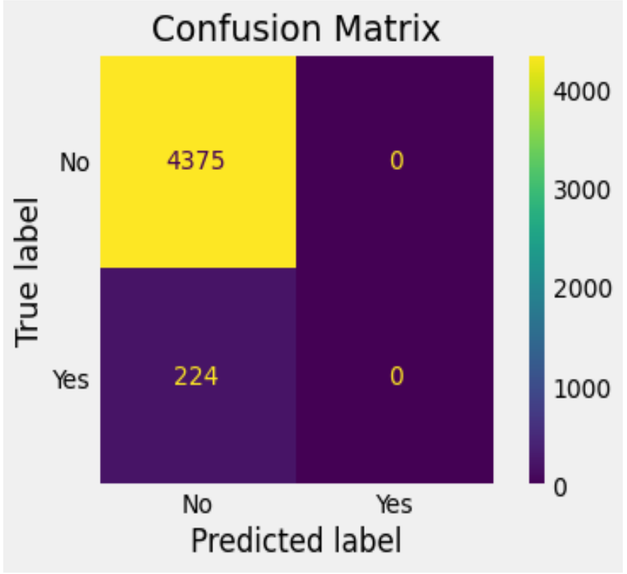
Default model performance on training set..

	precision	recall	f1-score	support
No	1.00	1.00	1.00	4375
Yes	1.00	1.00	1.00	224
accuracy			1.00	4599
macro avg	1.00	1.00	1.00	4599
weighted avg	1.00	1.00	1.00	4599

Classification Report

TESTING & EVALUATION

01. Data Training - Simpler Model



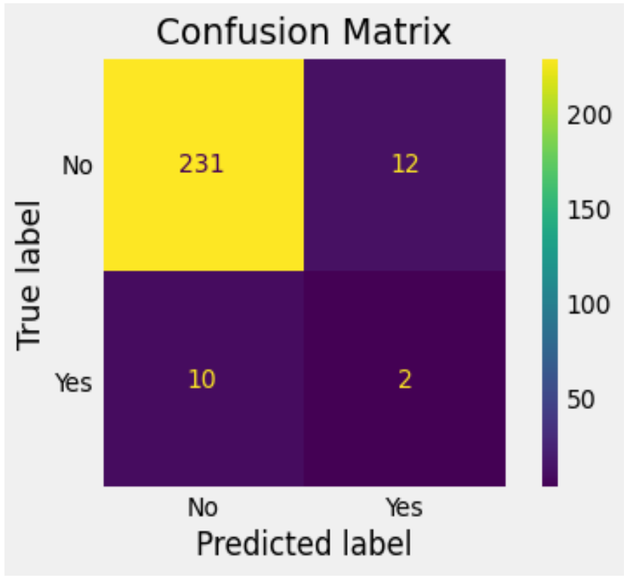
Confusion Matrix

	precision	recall	f1-score	support
No	0.95	1.00	0.98	4375
Yes	0.00	0.00	0.00	224
accuracy			0.95	4599
macro avg	0.48	0.50	0.49	4599
weighted avg	0.90	0.95	0.93	4599

Classification Report

TESTING & EVALUATION

02. Data Development - Default Model



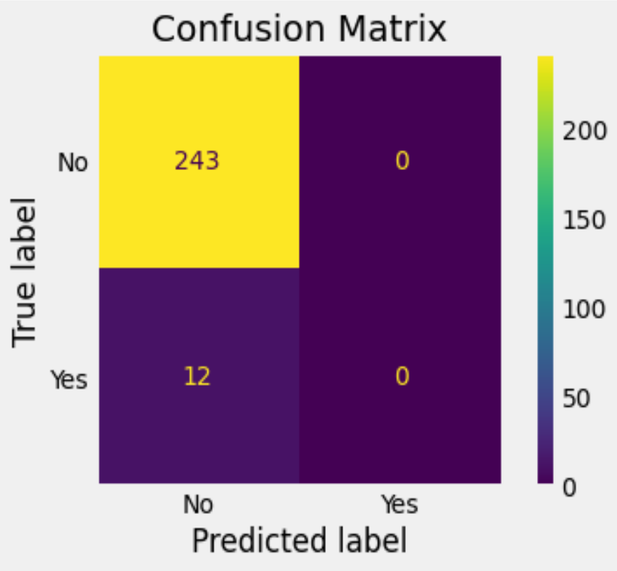
Confusion Matrix

	precision	recall	f1-score	support
No	0.96	0.95	0.95	243
Yes	0.14	0.17	0.15	12
accuracy			0.91	255
macro avg	0.55	0.56	0.55	255
weighted avg	0.92	0.91	0.92	255

Classification Report

TESTING & EVALUATION

02. Data Development - Simpler Model



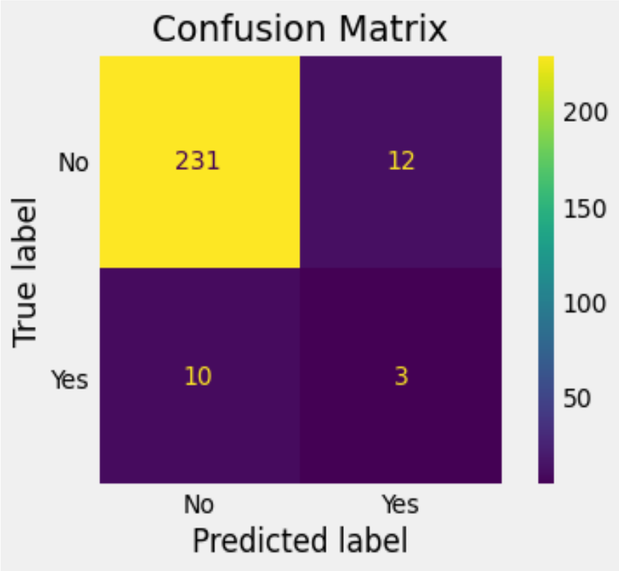
Confusion Matrix

	precision	recall	f1-score	support
No	0.95	1.00	0.98	243
Yes	0.00	0.00	0.00	12
accuracy			0.95	255
macro avg	0.48	0.50	0.49	255
weighted avg	0.91	0.95	0.93	255

Classification Report

TESTING & EVALUATION

02. Data Testing - Default Model



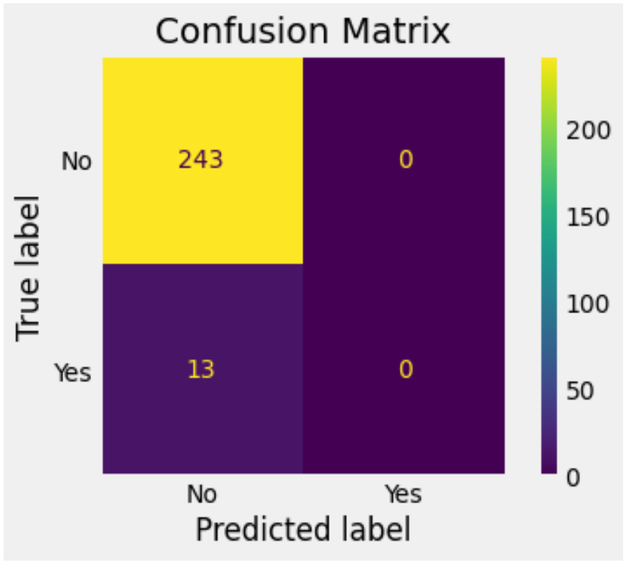
Confusion Matrix

	precision	recall	f1-score	support
No	0.96	0.95	0.95	243
Yes	0.20	0.23	0.21	13
accuracy			0.91	256
macro avg	0.58	0.59	0.58	256
weighted avg	0.92	0.91	0.92	256

Classification Report

TESTING & EVALUATION

02. Data Testing - Simpler Model



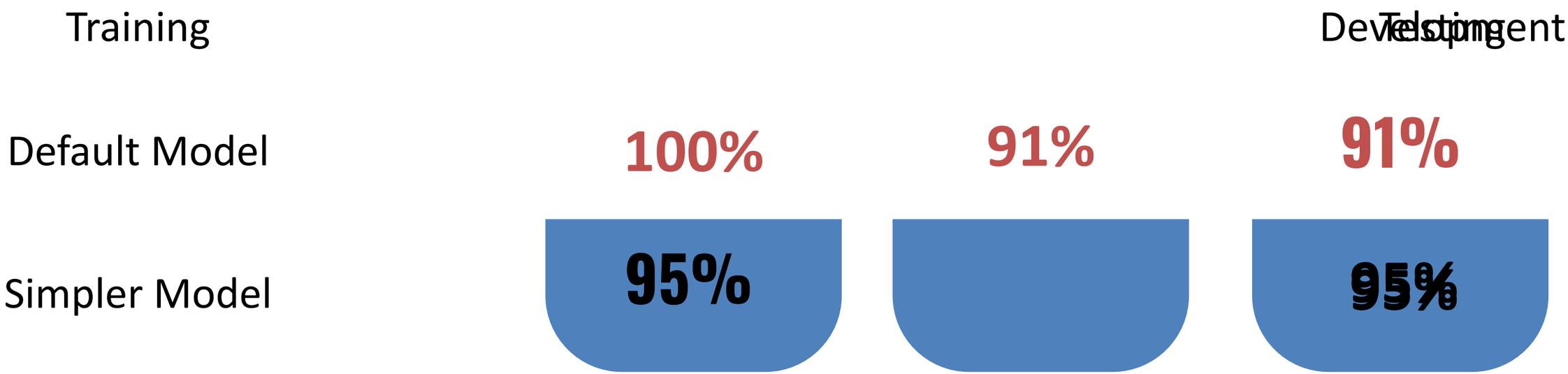
Confusion Matrix

	precision	recall	f1-score	support
No	0.95	1.00	0.97	243
Yes	0.00	0.00	0.00	13
accuracy			0.95	256
macro avg	0.47	0.50	0.49	256
weighted avg	0.90	0.95	0.92	256

Classification Report

TESTING & EVALUATION

02. Kesimpulan - Akurasi Model



TESTING & EVALUATION

03. Prediction

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	prediction
0	9046	Male	67.0	No	Yes	Yes	Private	Urban	228.69	36.600000	formerly smoked	Yes
1	51676	Female	61.0	No	No	Yes	Self-employed	Rural	202.21	28.893237	never smoked	Yes
2	31112	Male	80.0	No	Yes	Yes	Private	Rural	105.92	32.500000	never smoked	Yes
3	60182	Female	49.0	No	No	Yes	Private	Urban	171.23	34.400000	smokes	Yes
4	1665	Female	79.0	Yes	No	Yes	Self-employed	Rural	174.12	24.000000	never smoked	Yes

- Program pencegahan penyakit stroke dapat dilakukan dengan pengecekan kesehatan secara berkala (rata-rata glukosa, BMI, tekanan darah, serta kemungkinan memiliki riwayat penyakit lain).
- Mengurangi merokok juga dapat menjadi salah satu cara untuk meningkatkan kesehatan dan mengurangi resiko penyakit stroke.
- menerapkan program pola makan yang sehat ,hindari stress berlebih dan melakukan olahraga yang teratur dapat menngobati hypertensi dan mengurangi risiko terkena penyakit stroke
- menggunakan masker di lingkungan perkotaan dengan polusi udara yang tinggi dapat mengurangi risiko terkena penyakit radiovaskular dan penyakit stroke

LAMPIRAN

- **Link Dataset :**
  https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset