BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil

Berdasarkan metodologi yang telah dirancang pada kasus prediksi risiko terkena penyakit *stroke* dengan metode *Random Forest*, berikut hasil dari tahapan metodologi tersebut.

4.1.1 *Data Understanding*

Penelitian menggunakan *dataset* penyakit *stroke* yang diperoleh dari repositori *Kaggle*. Dataset yang digunakan berjumlah 4981 data dalam format *comma separated values (csv).*

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
2	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
3	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
4	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1

Gambar 4. 1 Sample Data

Tabel 4. 1 Meta Data

Atribute	*Description*	*Value*	*Type Data*
gender	Jenis Kelamin Pasien	*Male* dan *Female*	*Categorical*
age	Usia pasien (dalam tahun)	0.08 - 9.0	*Numerical*
hypertension	Apakah Pasien Pernah Menderita Hipertensi (0 = *No*, 1 = *Yes*)	0 atau 1	*Categorical*

heart_disease	Apakah Pasien Memiliki Riwayat Penyakit Jantung (0 = *No*, 1 = *Yes*)	0 atau 1	*Categorical*
ever_married	Apakah Pasien Sudah Menikah (*Yes* = Ya, *No* = Tidak)	*Yes* atau *No*	*Categorical*
work_type	Jenis Pekerjaan : *Children* (anak-anak), *Private* (Swasta), *Self-employed* (Wiraswasta), *Govt_job* (Pekerja pemerintah / PNS) *Never_worked* (Tidak pernah bekerja)	*Children, Private, Self-employed, Govt_job, Never_worked*	*Categorical*
Residence_type	Jenis Tempat Tinggal : *Rural* (Pedesaan), *Urban* (perkotaan)	*Rural, Urban*	*Categorical*
avg_glucose_level	Rata-rata kadar glukosa dalam darah	100.01 - 99.97	*Numerical*
bmi	Indeks Massa Tubuh (dalam Kg)	14.0 - 48.9	*Numerical*
smoking_status	tatus Pasien Merokok : *Smokes* (merokok), *Never smoked* (tidak pernah merokok), *formerly smoked* (pernah merokok), Unknown (Tidak diketahui)	*Smokes, Never smoked, formerly smoked, Unknown*	*Categorical*
stroke	Apakah pasien mengalami *stroke* (1) atau tidak (0)	0 atau 1	*Categorical*

a. *Exploratory Data Analysis (EDA)*

1) Menangani *Missing Values*

Pada tahap ini dilakukan pemeriksaan pada *dataset* yang betujuan mengetahui nilai yang rumpang atau *missing values* untuk menghindari *overfitting* pada model, data yang rumpang sangat sensitif terhadap performa model. Oleh karena, itu harus dilkukan pemeriksaan *missing value* dan setelah dilakukan pemeriksaan

missing value diketahui bahwa dataset tidak memiliki *missing values* sehingga data dapat diolah lebih lanjut.

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

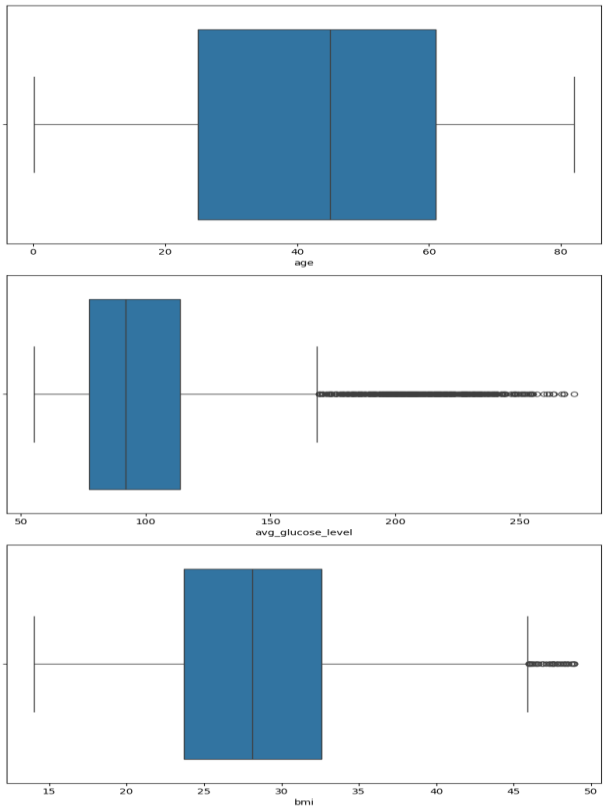
Gambar 4. 2 Output Missing Value

2) Memeriksa dan Menangani *Outlier*

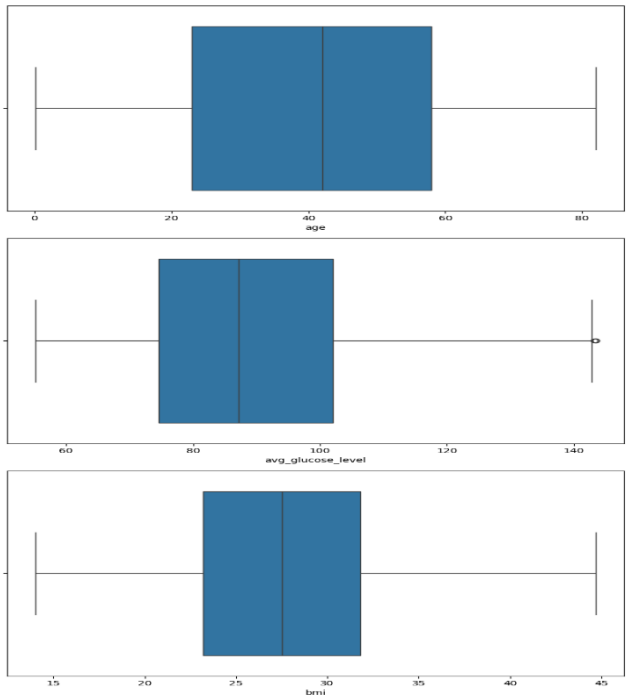
Mengidentifikasi nilai pada dataset apakah *logic* dan rasional sesuai dengan fakta. Data yang tidak sesuai dengan faktanya biasanya disebut sebagai data *outlier* atau data yang sangat menyimpang pada sebaran didalam dataset.

Data *outlier* dapat mempengaruhi performa model karena dapat mengakibatkan *overfitting* dan *underfitting* karena data tersebut bukan termasuk ke dalam data sebarannya, maka perlu dilakukan pemeriksaan *outlier*.

Pada penelitian ini pemeriksaan *outlier* menggunakan IQR (*Inter Quartile Range*) dimana data dikatakan sebagai *outlier* karena data tersebut melebihi nilai maksimum dari pada Q3 atau dibawah nilai minimum. Setelah dilakukan pemeriksaan menggunakan IQR didapatkan nilai yang memiliki *outlier* adalah variabel bmi dan avg_glucose_level.



Gambar 4. 3 Diagram Boxplot dengan outlier



Gambar 4. 4 Output Avg_glucose_level Setelah Menghilangkan Nilai Outlier

Hasil pemeriksaan *outlier* pada variabel ***avg_glucose_level*** menunjukkan terdapat **602** *outlier* yang berada di atas ambang batas maksimal, menandakan bahwa distribusi data sangat dipengaruhi oleh nilai-nilai ekstrem. Karena jumlah *outlier* pada *`avg_glucose_level`* cukup signifikan, diperlukan tindakan droping menggunakan metode IQR *(Interquartile Range*) untuk menjaga keakuratan analisis data dan hasil model. Setelah proses *droping* dilakukan, jumlah *outlier* yang melebihi ambang batas maksimal berkurang menjadi hanya **5**. Hal ini menunjukkan bahwa distribusi data telah menjadi lebih normal dan kurang dipengaruhi oleh nilai-nilai ekstrem, sehingga data lebih siap untuk dianalisis dan dapat memberikan hasil yang lebih baik dalam model.

Hasil dari pemeriksaan *outlier* menunjukkan bahwa pada atribut **BMI**, terdapat **42 *outlier*** yang berada di atas ambang batas maksimal. Hal ini menandakan adanya beberapa nilai yang signifikan lebih tinggi daripada kebanyakan data lainnya, yang dapat menyebabkan analisis dan model menjadi bias. Oleh karena itu, diperlukan proses *droping* untuk menghilangkan nilai yang melewati ambang batas IQR agar mengurangi dampak nilai ekstrem. Setelah melakukan *droping*, jumlah *outlier* yang berada di atas ambang batas maksimal menurun dari **42 menjadi 0**. Ini menunjukkan peningkatan signifikan dalam distribusi data yang kini lebih merata dan *representatif*.

3) Memeriksa Distribusi dan Pemusatan Data

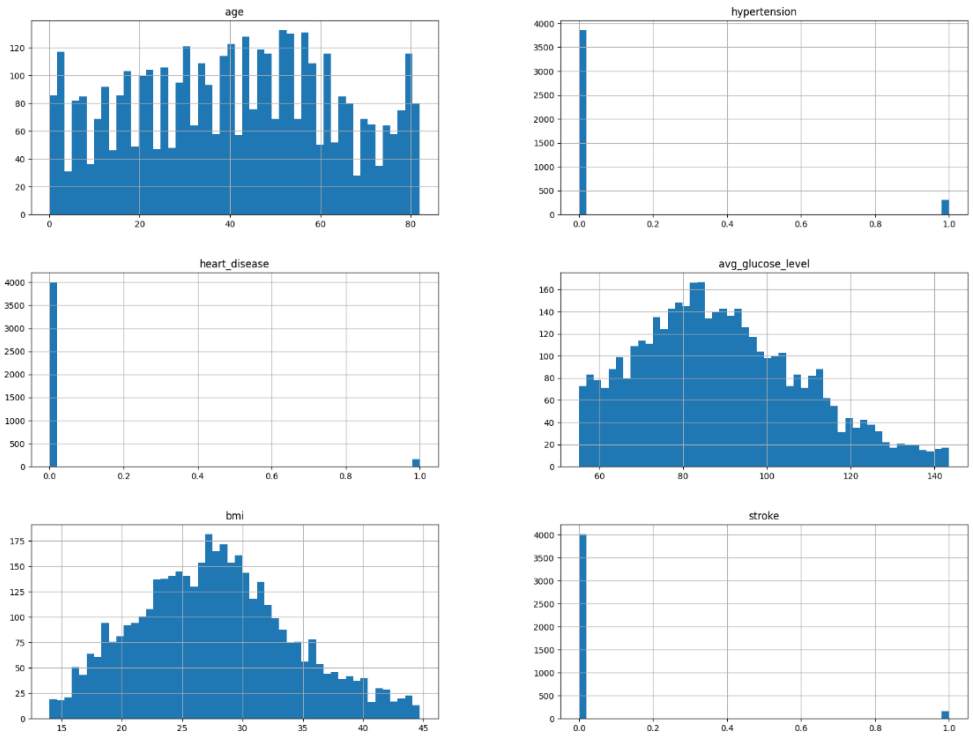
Data yang telah bersih dari *outlier* selanjutnya diperiksa terkait distribusi datanya, apakah distribusi datanya sudah normal, metrik atau ukuran dari distribusi data atau sebaran data mengacu pada nilai *mean, median,* dan *standar deviation (std)*.

Adapun jika melihat pada data distribusinya diketahui bahwa variabel bmi sudah hampir memenuhi distribusi normal karena nilai mean 27,72 dan nilai rata-ratanya 27,50. sedangkan pada variabel

avg_glucose_level belum memenuhi distribusi normal karena nilai *Mean* 89,18 > Median 87,13 sehingga *avg_glucose_level* tergolong *Right-skewed* (*positive* skew).

	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	4174.000000	4174.00000	4174.000000	4174.000000	4174.000000	4174.000000
mean	41.023067	0.07379	0.039530	89.182166	27.729828	0.037374
std	22.512539	0.26146	0.194877	19.449182	6.401379	0.189700
min	0.080000	0.00000	0.000000	55.120000	14.000000	0.000000
25%	23.000000	0.00000	0.000000	74.630000	23.200000	0.000000
50%	42.000000	0.00000	0.000000	87.135000	27.500000	0.000000
75%	58.000000	0.00000	0.000000	101.997500	31.800000	0.000000
max	82.000000	1.00000	1.000000	143.470000	44.700000	1.000000

Gambar 4. 5 Output Distribusi Data

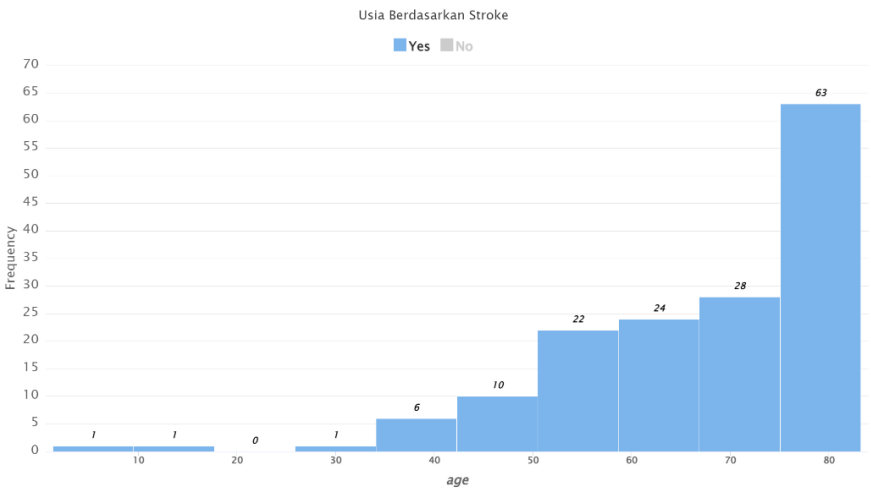


Gambar 4. 6 Histogram Distribusi Data

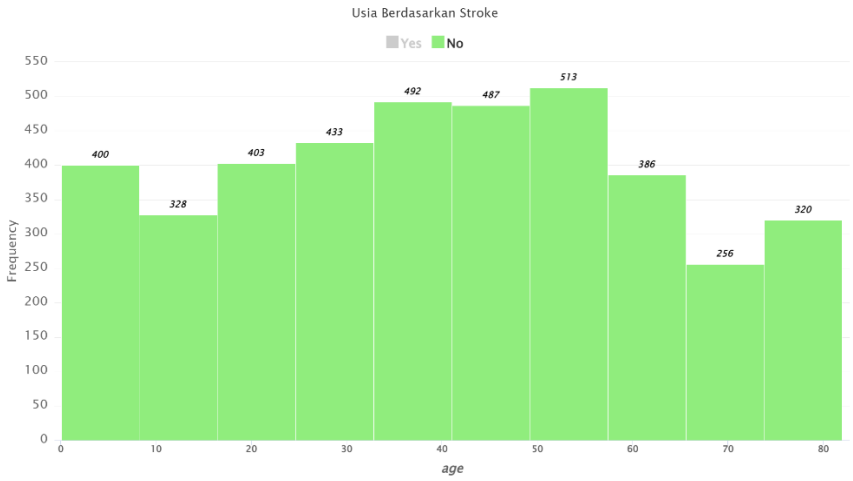
4) Visualisasi Fitur antar Target

a) *Age* Berdasarkan *Stroke*

Penyakit *stroke* paling umum terjadi pada individu berusia 35-83 tahun. Dalam rentang usia tersebut, jumlah kasus *stroke* paling tinggi ditemukan pada individu usia 75-83 tahun dengan 63 kasus, diikuti oleh individu usia 66-75 tahun sebanyak 28 kasus, usia 58-66 tahun sebanyak 24 kasus, usia 50-58 tahun sebanyak 22 kasus dan usia 34-42 sebanyak 6 kasus. Sementara itu, sangat sedikit kasus *stroke* terjadi pada individu usia 0-33 tahun, dengan hanya 3 kasus terdeteksi. Sehingga disimpulkan bahwa individu yang tergolong lansia yaitu 50 tahun keatas rentan terkena *stroke*.



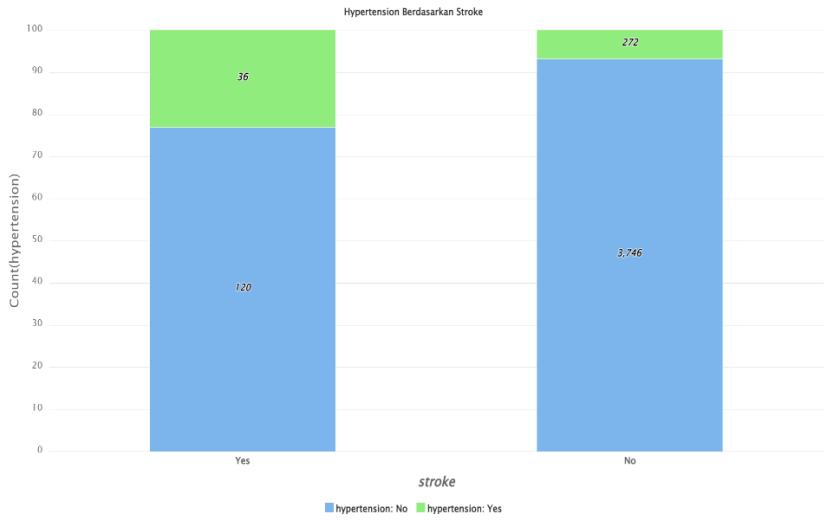
Gambar 4. 7 Visualisasi *Histogram Age* Berdasarkan *Stroke (yes)*



Gambar 4. 8 Visualisasi *Histogram Age* Berdasarkan *Stroke (No)*

b) *Hypertension* berdasarkan *stroke*

Dari data yang ada, terdapat 36 individu penderita *stroke* yang juga mengalami hipertensi, sementara 120 penderita *stroke* tidak mengalami hipertensi. Sebaliknya, di antara individu yang tidak mengalami *stroke*, 272 memiliki hipertensi, sedangkan 3.746 tidak mengalami hipertensi. Ini menunjukkan bahwa hipertensi lebih umum pada penderita *stroke* dibandingkan dengan yang tidak mengalami *stroke*.

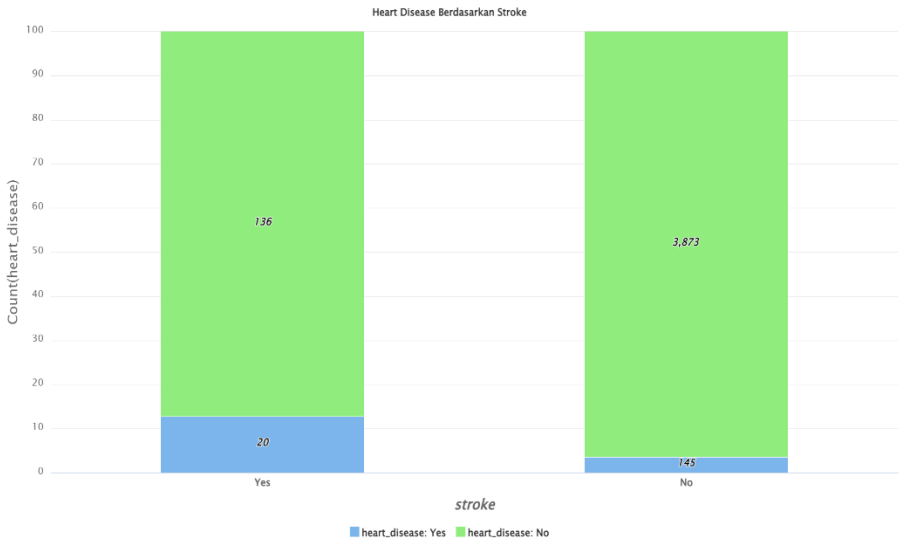


Gambar 4. 9 Visualisasi *Bar Chart Hypertension* Berdasarkan *Stroke*

c) *Heart Disease* Berdasarkan *Stroke*

Individu yang memiliki penyakit *stroke (yes)* dan juga memiliki penyakit jantung atau *heart_disease* sebanyak 20 kasus, sedangkan yang tidak memiliki riwayat penyakit jantung sebanyak 136 kasus. Individu yang tidak *stroke (no)* tetapi memiliki riwayat penyakit jantung sebanyak 145 kasus, sedangkan yang tidak memiliki riwayat penyakit jantung sebanyak 3.873 kasus.

Individu dengan *stroke* cenderung memiliki lebih sedikit kasus penyakit jantung dibandingkan dengan individu tanpa *stroke*, tetapi perbandingan ini menunjukkan bahwa mayoritas dari kedua kelompok (baik *stroke* maupun *non-stroke*) tidak memiliki riwayat penyakit jantung.



Gambar 4. 10 Visualisasi Bar Chart Heart Disease Berdasarkan *Stroke*

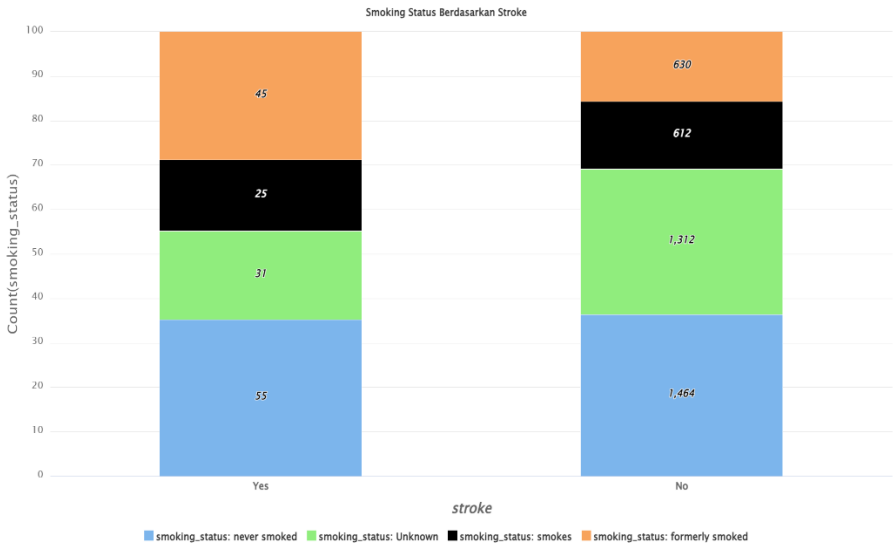
d) *Smoking Status* Berdasarkan *Stroke*

Individu penderita *stroke (yes)* tercatat memiliki berbagai status merokok dengan rincian sebagai berikut yaitu 45 kasus merupakan mantan perokok *(formerly smoked)*, 25 kasus masih aktif merokok *(smokes),* 31 kasus tidak diketahui status

merokoknya *(unknown),* dan 55 kasus tidak pernah merokok *(never smoked).*

Sementara itu, individu yang tidak menderita *stroke (no)* juga memiliki status merokok yang bervariasi, yaitu 630 kasus mantan perokok *(formerly smoked),* 612 kasus masih aktif merokok *(smokes),* 1.312 kasus tidak diketahui status merokoknya *(unknown),* dan 1.464 kasus tidak pernah merokok *(never smoked).*

Jumlah individu yang tidak pernah merokok lebih tinggi dibandingkan dengan kategori lainnya, baik pada kelompok *stroke* maupun *non-stroke*, yang dapat mengindikasikan bahwa merokok bukan satu-satunya faktor risiko yang dominan bagi kejadian *stroke*.



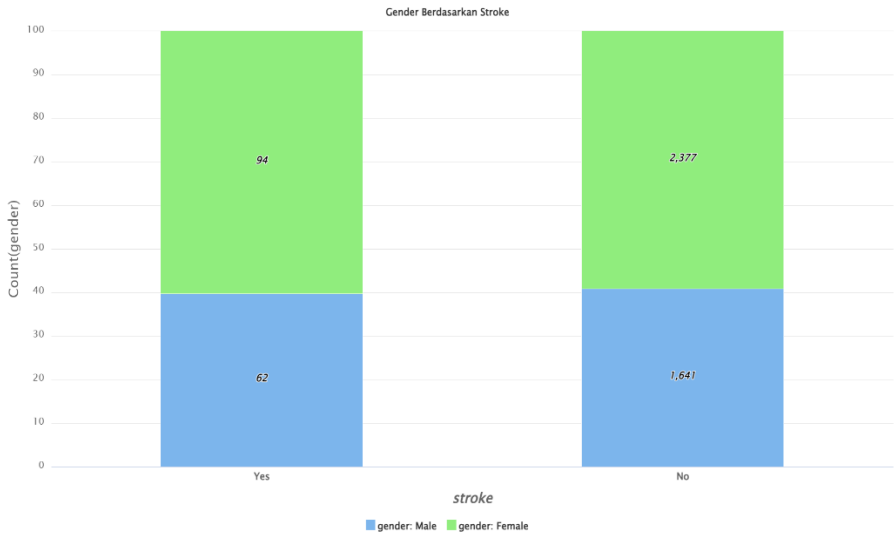
Gambar 4. 11 Visualisasi Bar Chart Smoking Status Berdasarkan *Stroke*

e) *Gender* Berdasarkan *Stroke*

Pada data yang diperoleh, terdapat 62 individu berjenis kelamin laki-laki (*male*) yang menderita *stroke* dan 94 individu berjenis kelamin perempuan (*female*) yang juga menderita *stroke*.

Sementara itu, jumlah individu yang tidak menderita *stroke* terdiri dari 1.639 laki-laki dan 2.375 perempuan.

Dari data tersebut, dapat disimpulkan bahwa jumlah penderita *stroke* lebih banyak pada perempuan dibandingkan laki-laki. Selain itu, jumlah individu yang tidak menderita *stroke* juga lebih banyak pada perempuan dibandingkan laki-laki.



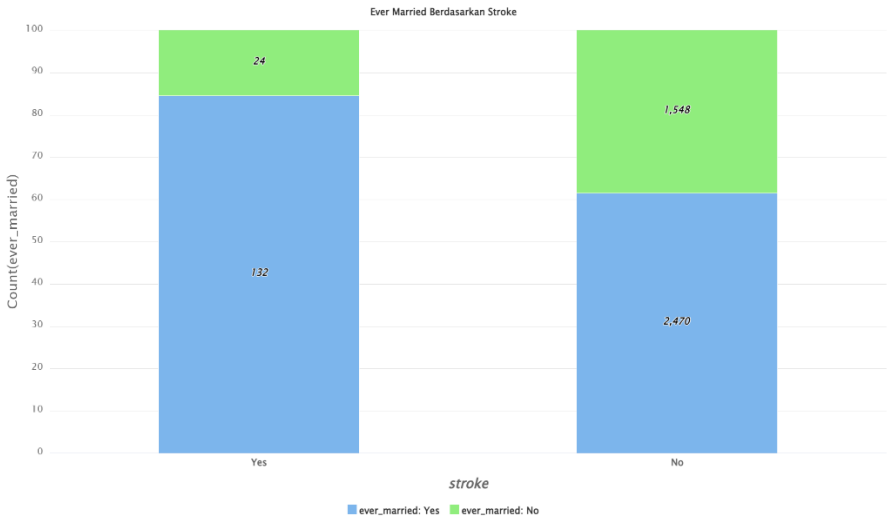
Gambar 4. 12 Visualisasi *Bar Chart Gender* Berdasarkan *Stroke*

f) *Ever Married* berdasarkan *stroke*

Dari data tersebut, tampak bahwa sebagian besar penderita *stroke* (132 dari 156 individu, atau sekitar 84.6%) sudah menikah. Hal ini sejalan dengan tren pada populasi yang tidak menderita *stroke*, di mana lebih banyak individu yang sudah menikah (2,468 dari 4,014 individu, atau sekitar 61.5%).

Namun, persentase individu menikah yang menderita *stroke* (84.6%) jauh lebih tinggi dibandingkan dengan persentase individu menikah yang tidak menderita *stroke* (61.5%). Ini menunjukkan bahwa faktor lain, seperti usia atau kebiasaan hidup yang berbeda, mempengaruhi prevalensi *stroke* di kalangan individu yang sudah menikah.

Sebaliknya, hanya 15.4% dari penderita *stroke* yang belum menikah, dibandingkan dengan 38.5% dari individu yang tidak menderita stroke. Ini menunjukkan bahwa status pernikahan bisa menjadi variabel yang penting untuk dipertimbangkan dalam penelitian lebih lanjut, meskipun tidak cukup untuk menyimpulkan hubungan kausal antara pernikahan dan *stroke*.

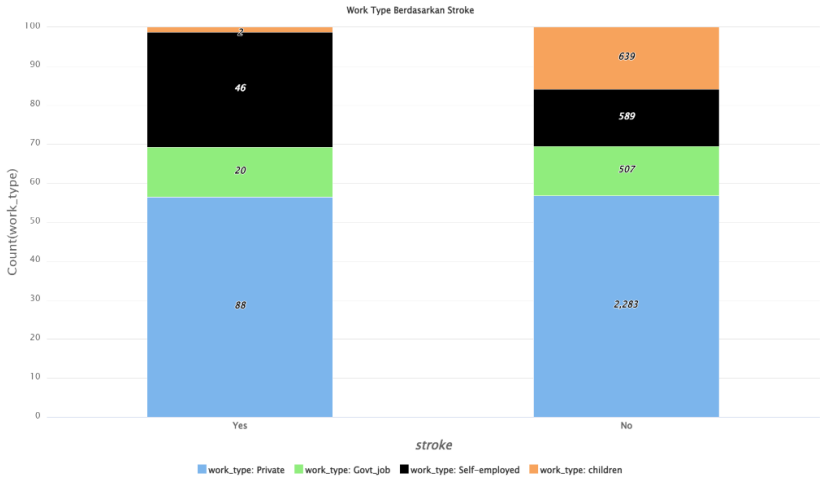


Gambar 4. 13 Visualisasi *Ever Married* berdasarkan *Stroke*

g) *Work Type* berdasarkan *stroke*

Mayoritas penderita *stroke* berasal dari sektor swasta (*private*) dengan jumlah 88 orang, penderita *stroke* dari kalangan PNS (*govt_job*) berjumlah 20 orang, diikuti oleh wiraswasta (*self-employed*) sebanyak 46 orang sedangkan anak-anak (*children*) penderita *stroke* hanya berjumlah 2 orang.

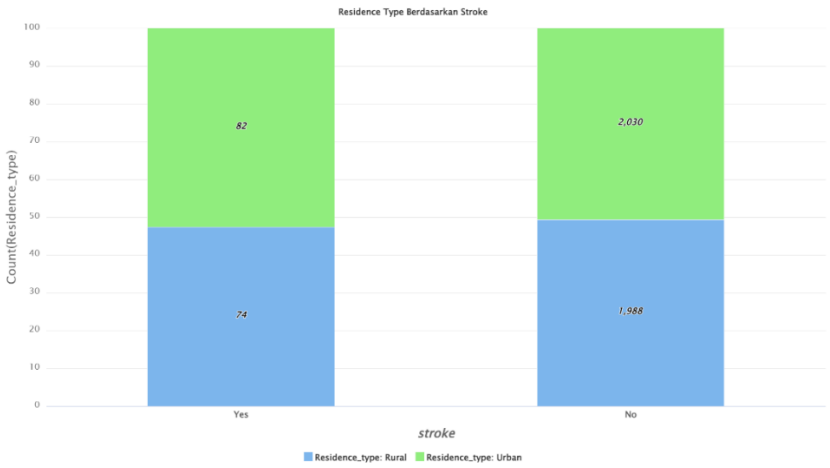
Data ini menunjukkan bahwa pekerjaan dapat menjadi faktor penting dalam risiko *stroke*. Sektor swasta dan wiraswasta menunjukkan jumlah penderita *stroke* yang lebih tinggi, mungkin disebabkan oleh tekanan kerja yang lebih tinggi, gaya hidup yang kurang sehat, atau faktor risiko kesehatan lainnya. Perbedaan ini dapat menjadi fokus untuk intervensi kesehatan kerja yang lebih baik dan program pencegahan *stroke* di berbagai sektor pekerjaan.



Gambar 4. 14 Visualisasi *Work Type* Berdasarkan *Stroke*

h) *Residence Type* berdasarkan *stroke*

Jenis tempat tinggal memiliki pengaruh yang signifikan terhadap jumlah penderita *stroke*. Berdasarkan data, terdapat 74 individu penderita *stroke* yang tinggal di pedesaan (*rural*). Sementara itu, di perkotaan (*urban*) terdapat 82 individu yang menderita *stroke*. Hal ini menunjukkan bahwa meskipun jumlah penderita *stroke* di perkotaan sedikit lebih banyak, faktor tempat tinggal, baik di pedesaan maupun perkotaan, tidak secara signifikan mempengaruhi prevalensi *stroke*.



Gambar 4. 15 Visualisasi *Residence Type Berdasarkan Stroke*

5) Univariate Analysis

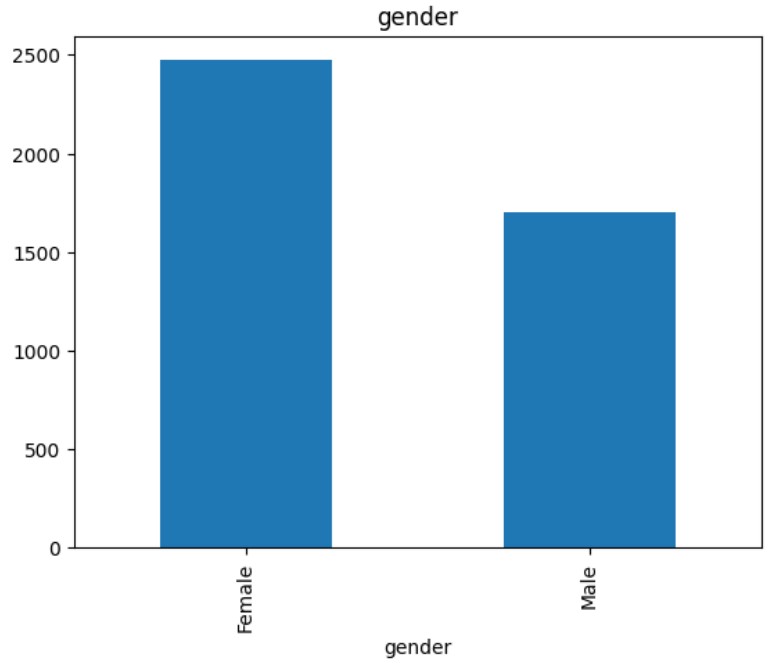
Data yang telah bersih akan dilakukan analisis data untuk memahami dan menggambarkan karakteristik satu variabel tunggal dalam dataset.

a) Categorical Features

Menghitung frekuensi dan proporsi dari setiap kategori, membantu dalam memahami distribusi kategori dalam dataset.

Distribusi gender dari sampel yang dianalisis. Dalam data ini, terdapat dua kategori, yaitu *Female* dan *Male*. Total jumlah sampel untuk kategori *Female* adalah 2.471, yang mencakup 59,2% dari keseluruhan data. Sementara, kategori *Male* memiliki total 1.703 sampel, yang setara dengan 40,8% dari data.

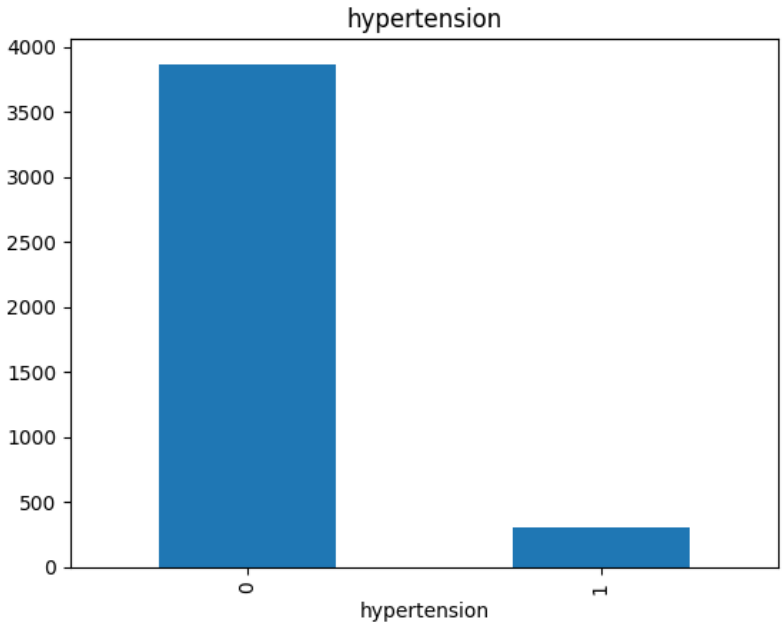
Berdasarkan distribusi ini, terlihat bahwa kategori *Female* lebih dominan dibandingkan dengan *Male*, dengan selisih 18% lebih banyak dalam jumlah sampel. Hal ini mungkin mencerminkan populasi yang lebih besar dari perempuan dalam sampel yang dianalisis.



Gambar 4. 16 Bar Chart Univariate Categorical-Gender

Dataset ini berisi sampel, yang terdiri dari dua kategori pada variabel *hypertension* yaitu individu yang tidak memiliki hipertensi (0) dan individu yang memiliki hipertensi (1). kategori (0) terdapat 3.866 sampel yang merupakan bagian terbesar dari dataset, mencakup 92,6% dari total sampel.

Hal ini menunjukkan bahwa mayoritas individu dalam dataset ini tidak memiliki hipertensi. Kategori (1) sebanyak 308 sampel yang menyumbang 7,4% dari total sampel. Ini menunjukkan bahwa hanya sebagian kecil dari individu dalam dataset yang memiliki hipertensi.



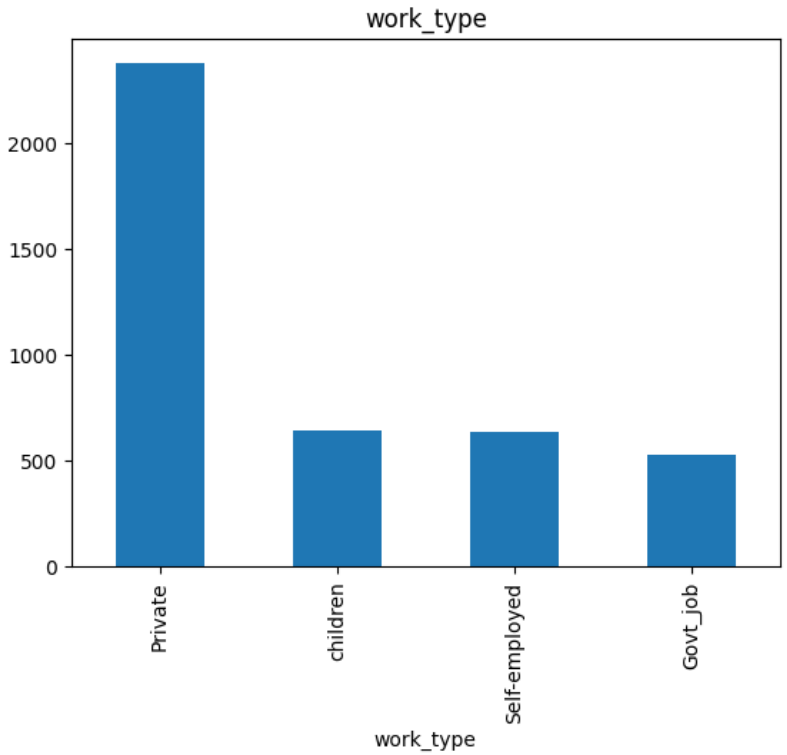
Gambar 4. 17 Bar Chart Univariate Categorical- Hypertension

Pada distribusi sampel berdasarkan jenis pekerjaan (*work_type*) didominasi *private* dengan jumlah sampel sebanyak 2.371, yang menyumbang 56.8% dari total sampel. Hal ini menunjukkan bahwa lebih dari setengah populasi dalam dataset bekerja di sektor swasta.

Kelompok kedua terbesar adalah *Children* dengan 641 sampel, yang mencakup 15.4% dari total. Ini menunjukkan bahwa sebagian besar dataset terdiri dari individu yang belum memasuki usia kerja dan masih berada dalam kelompok usia anak-anak.

Selanjutnya adalah kelompok *Self-employed*, dengan 635 sampel yang mewakili 15.2% dari total. Ini menunjukkan bahwa ada sejumlah besar individu dalam dataset yang bekerja secara mandiri atau memiliki usaha sendiri.

Kelompok terkecil adalah *Govt_job* dengan 527 sampel, yang mencakup 12.6% dari total sampel. Ini menunjukkan bahwa proporsi individu yang bekerja di sektor pemerintah adalah yang paling kecil dibandingkan dengan kelompok pekerjaan lainnya. Hal ini memberikan gambaran tentang profil pekerjaan dari populasi yang dianalisis.



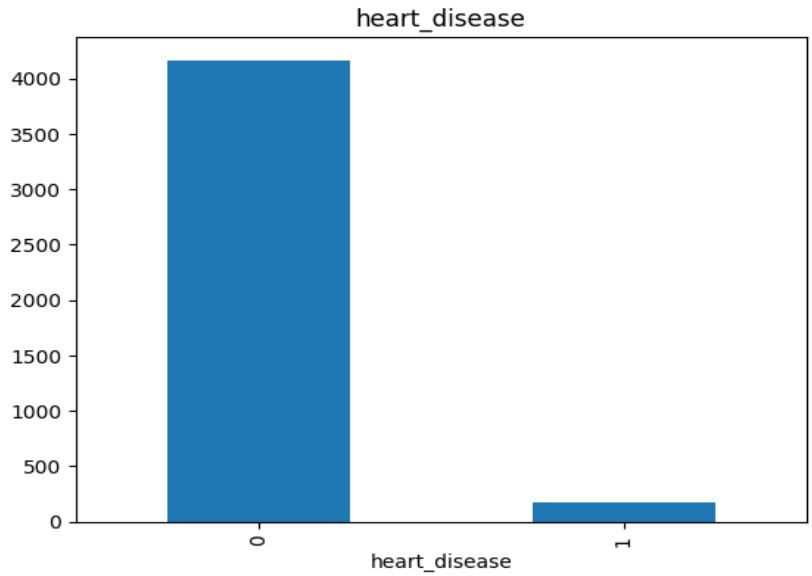
Gambar 4. 18 Bar Chart Univariate Categorical -Work_Type

Data ini terdapat dua kelompok yaitu individu tanpa penyakit jantung (0) dan individu dengan penyakit jantung (1).

Dari keseluruhan sampel, sebanyak 4009 individu atau 96.0% dari total populasi tidak memiliki penyakit jantung (0). Ini

menunjukkan bahwa mayoritas besar dari sampel tidak terdiagnosis dengan kondisi tersebut.

Sebaliknya, terdapat 165 individu atau 4.0% dari sampel yang menderita penyakit jantung (1). Proporsi yang kecil ini menunjukkan bahwa kasus penyakit jantung relatif jarang dalam dataset yang dianalisis. Distribusi ini menunjukkan ketidakseimbangan dalam data, dengan jumlah kasus tanpa penyakit jantung jauh lebih tinggi (0) dibandingkan dengan kasus penyakit jantung (1).



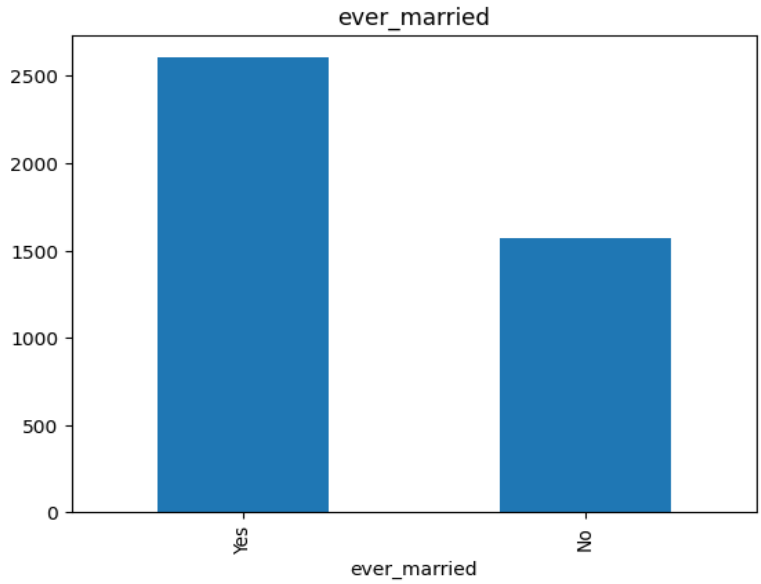
Gambar 4. 19 Bar Chart Univariate Categorical- Heart_Diseas

Data ini menunjukkan distribusi jumlah sampel berdasarkan status pernikahan (*ever married*). Terdapat dua kategori yaitu individu yang pernah menikah (yes) dan individu yang belum pernah menikah (no).

Kategori (yes) menunjukkan bahwa ada 2602 individu (62.3%) dari total sampel yang pernah menikah. Ini menunjukkan bahwa mayoritas dari populasi sampel ini telah memiliki pengalaman pernikahan.

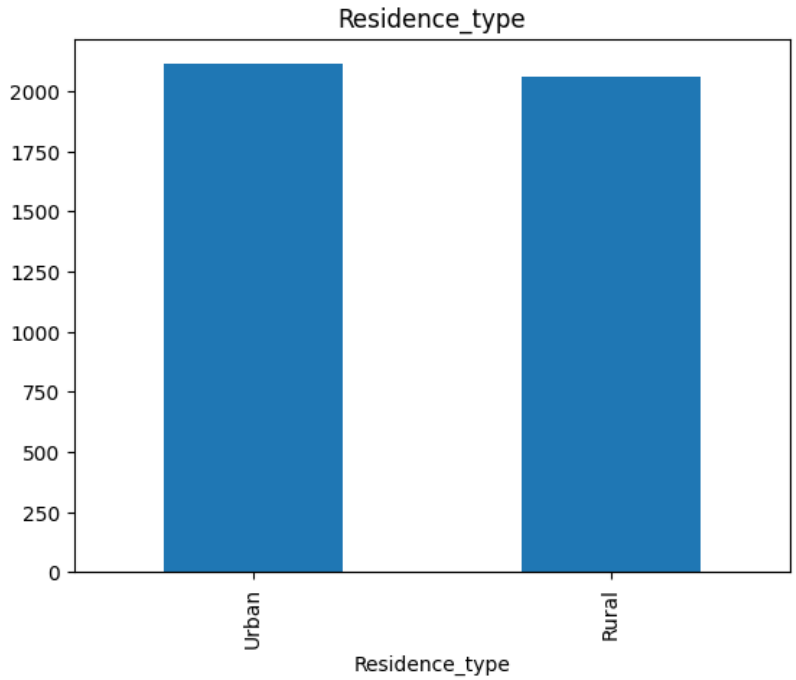
Kategori (no) mencakup 1572 individu (37.7%) dari total sampel yang belum pernah menikah. Persentase ini menunjukkan

bahwa lebih dari sepertiga populasi sampel ini belum memiliki pengalaman pernikahan.



Gambar 4. 20 Bar Chart Univariate Categorical -Ever_Married

Jumlah sampel berdasarkan tipe tempat tinggal *Residence_type*. Terdapat dua kategori utama yaitu perkotaan (*Urban*) dan pedesaan (*Rural*). Kategori (*Urban*) memiliki 2.112 sampel menyumbang 50,6% dari keseluruhan data, sementara kategori (*Rural*) memiliki 2.062 sampel menyumbang 49,4%. Hal ini mengindikasikan bahwa distribusi data antara kedua kategori tempat tinggal tersebut hampir seimbang, dengan sedikit lebih pada kategori *urban*.



Gambar 4. 21 Bar Chart Univariate Categorical -Residence_Type.

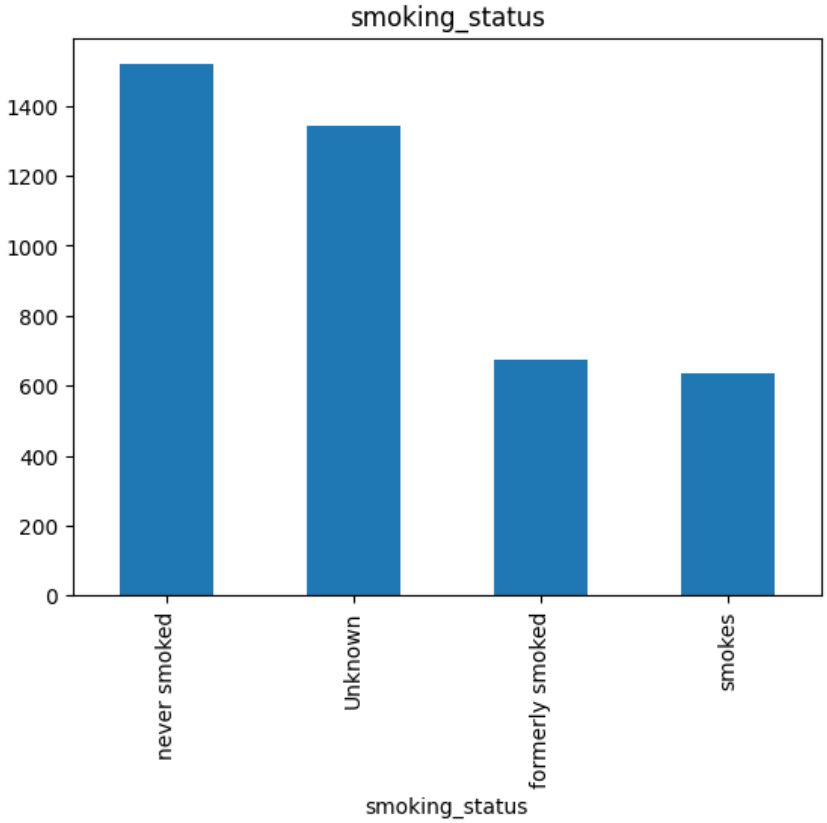
Pada distribusi sample berdasarkan status merokok, *Never Smoked* dengan jumlah sampel 1.519 kategori ini mewakili individu yang tidak pernah merokok. Dengan persentase 36.4%, kelompok ini merupakan yang terbesar dalam *dataset*, menunjukkan bahwa mayoritas dari sampel terdiri dari orang-orang yang tidak memiliki riwayat merokok.

Unknown dengan jumlah sampel 1.343 ini adalah kategori bagi individu yang status merokoknya tidak diketahui atau tidak tercatat. Dengan persentase 32.2%, kategori ini adalah yang kedua terbesar, dan menunjukkan adanya data yang tidak lengkap atau hilang mengenai kebiasaan merokok pada sejumlah besar sampel.

Formerly Smoked dengan jumlah sampel 675 kategori ini terdiri dari individu yang dulunya merokok tetapi telah berhenti. Mencakup 16.2% dari total sampel, menunjukkan individu yang mungkin memiliki riwayat kesehatan yang berbeda dibandingkan dengan non-perokok atau perokok aktif.

Smokes dengan *lower case* 637, kategori ini mencakup individu yang saat ini merokok. Dengan persentase 15.3%, kategori ini adalah kelompok terkecil dalam dataset, namun tetap signifikan untuk analisis terkait dampak kesehatan dari kebiasaan merokok.

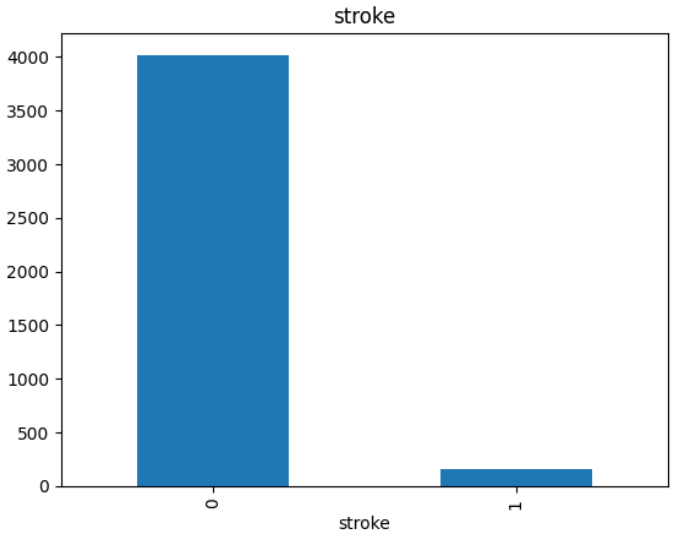
Dari data ini, dapat dilihat bahwa lebih dari sepertiga populasi tidak pernah merokok, sementara hampir sepertiga lainnya memiliki status merokok yang tidak diketahui. Sebagian kecil dari populasi adalah perokok aktif atau mantan perokok, yang mungkin memberikan wawasan penting tentang faktor risiko kesehatan dalam analisis lebih lanjut.



Gambar 4. 22 Bar Chart Univariate Categorical- Smoking Status

Dalam dataset yang dianalisis, terdapat dua kategori untuk atribut `*stroke*`: 0 dan 1 dari total sampel, 4018 (96.3%) tidak mengalami *stroke*, sedangkan 156 (3.7%) mengalami *stroke*. Ini menunjukkan

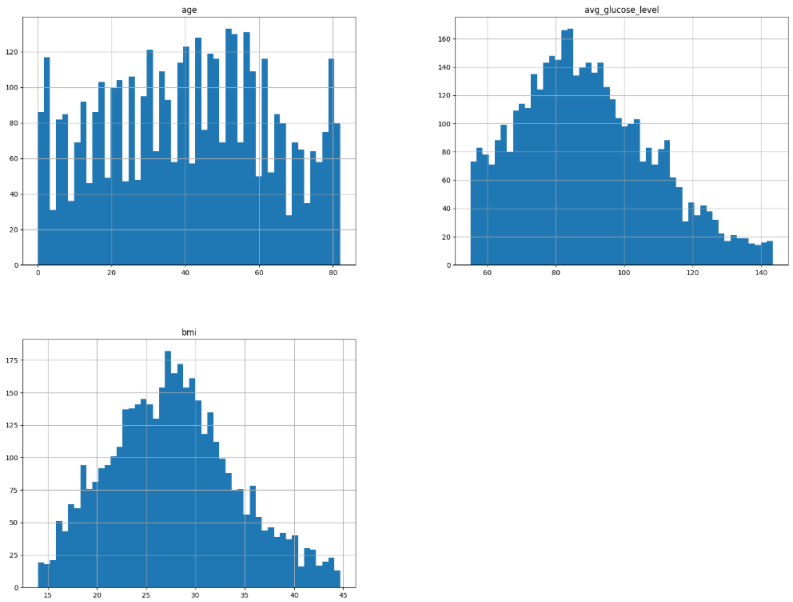
bahwa mayoritas data (96.3%) tidak mengalami *stroke*, sementara proporsi yang mengalami *stroke* relatif kecil (3.7%).



Gambar 4. 23 Univariate Categorical -Stroke

b) Numerical Features

Melihat data dari histogramnya dimana numerical features dari univariate analysis untuk data numerik dimana grafik histogram pada fitur bmi grafik nya sudah terdistribusi cukup baik sedangkan untuk fitur *avg_glucose_level* belum terdistribusi dengan baik.



Gambar 4. 24 Histogram Univariate Numerical Features

6) *Multivariate Analysis*

Analisis *multivariate* ini bertujuan untuk mengetahui korelasi antar fitur terhadap target.

a) *Categorical Features*

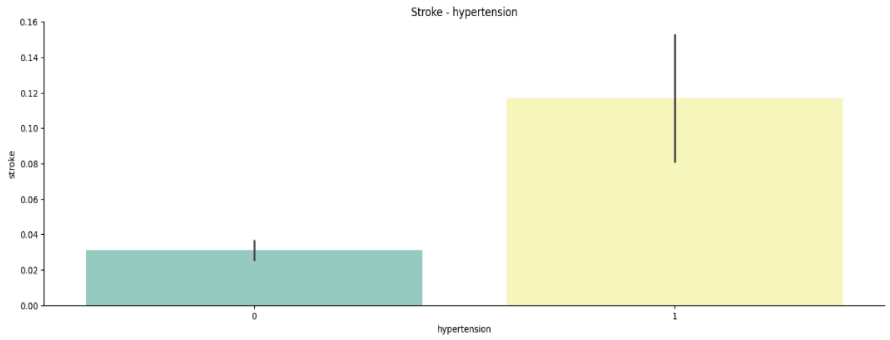
Dapat dilihat bahwa fitur gender terhadap *stroke*, fitur *gender* menunjukkan bahwa baik laki-laki (*male*) maupun perempuan (*female*) memiliki pengaruh yang sama terhadap risiko *stroke*. Hal ini berarti tidak ada perbedaan signifikan dalam pengaruh kemungkinan terjadinya *stroke* antara kedua kelompok tersebut.

Output :

Gambar 4. 25 Diagram Catplot Stroke – Gender

Fitur *hypertension* terhadap *stroke*, fitur hipertensi menunjukkan pengaruh signifikan terhadap risiko *stroke*. Penderita hipertensi yang memiliki status "ya" (1) memiliki risiko *stroke* yang jauh lebih tinggi dibandingkan dengan yang tidak mengalami hipertensi (0). Hal ini menunjukkan bahwa hipertensi merupakan faktor risiko utama untuk terjadinya *stroke*.

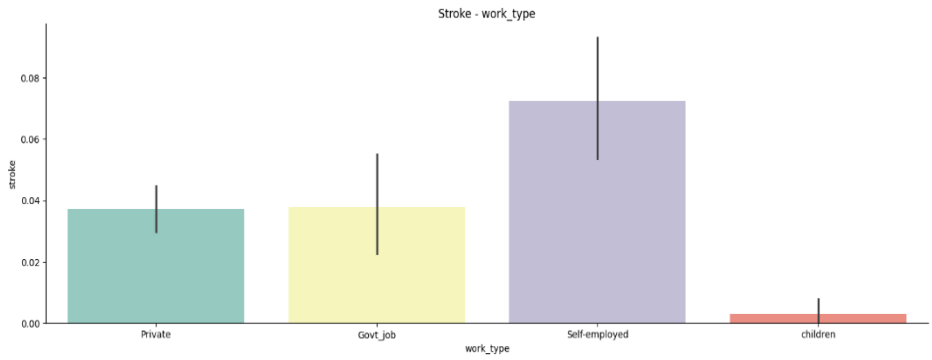
Output :



Gambar 4. 26 Diagram Catplot Stroke – Hypertension

Fitur *work_type* terhadap *stroke*, kategori yang paling berpengaruh terhadap kejadian *stroke* adalah ***self-employed***. Hal ini menunjukkan bahwa individu yang bekerja sebagai **wiraswasta** memiliki risiko *stroke* yang lebih tinggi dibandingkan dengan kategori pekerjaan lainnya. Selanjutnya, ***govt_job*** dan ***private*** juga menunjukkan pengaruh signifikan terhadap risiko *stroke*, namun keduanya memiliki rentang pengaruh yang lebih kecil, yaitu sekitar 0,04 dari pengaruh kategori *self-employed*.

Output :

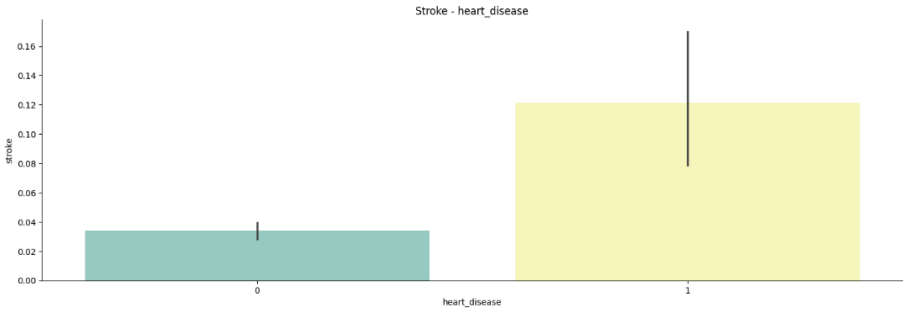


Gambar 4. 27 Diagram Catplot Stroke - Work_Type

Fitur *heart_disease* terhadap *stroke*, individu yang memiliki penyakit jantung (1) menunjukan pengaruh yang signifikan terhadap kejadian *stroke* dibandingkan dengan individu yang tidak memiliki

penyakit jantung (0). Hal ini menunjukkan bahwa penderita penyakit jantung lebih berisiko mengalami *stroke*.

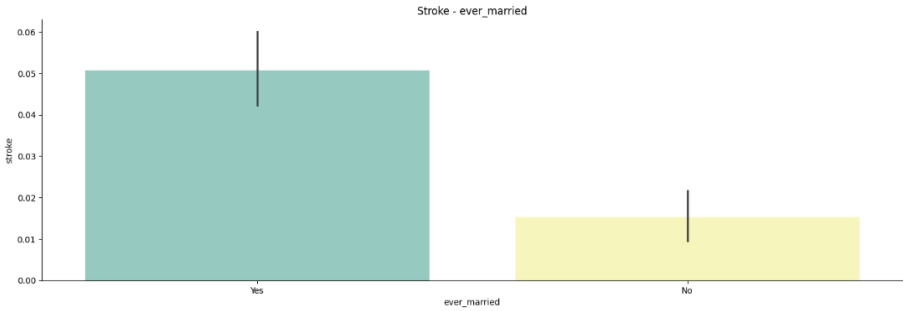
Output :



Gambar 4. 28 Diagram Catplot Stroke - Heart_Disease

Fitur ever_marrid terhadap *stroke*, individu yang memiliki status "pernah menikah" (*yes*) memiliki risiko *stroke* yang lebih tinggi dibandingkan dengan mereka yang "tidak pernah menikah" (*no*). Status pernikahan terbukti berpengaruh signifikan terhadap kemungkinan terjadinya *stroke*.

Output :

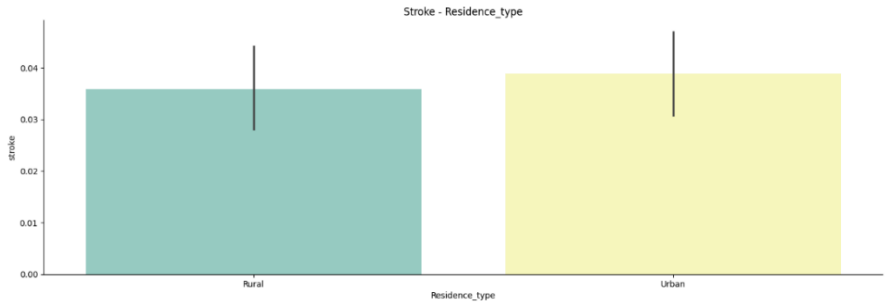


Gambar 4. 29 Diagram Catplot Stroke - Ever_Marrid

Fitur *residence_type* terhadap *stroke*, penderita yang tinggal di pedesaan (*rural*) maupun di perkotaan (*urban*) menunjukkan pengaruh signifikan terhadap risiko *stroke*. Meskipun keduanya

memiliki dampak yang besar, penderita yang tinggal di perkotaan (*urban*) memiliki risiko *stroke* yang lebih tinggi dibandingkan dengan mereka yang tinggal di pedesaan.

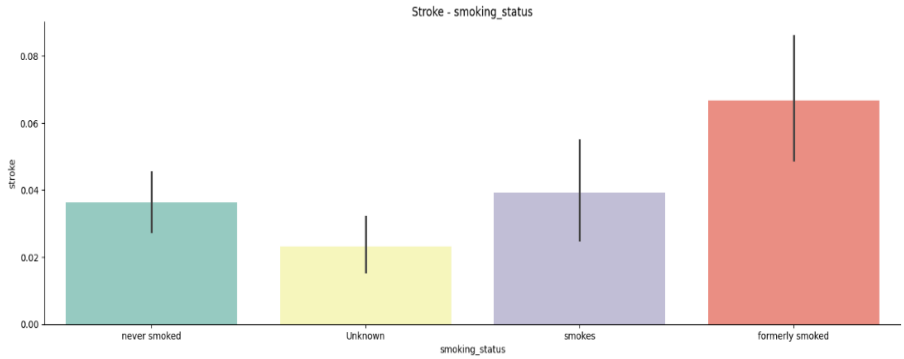
Output :



Gambar 4. Diagram Catplot Stroke - Residence_Type

Fitur *smoking_status* terhadap *stroke*, penderita *stroke* dipengaruhi oleh status merokok, dengan individu yang memiliki status *"formerly smoked"* (pernah merokok) menunjukkan dampak paling signifikan terhadap risiko *stroke*. Status *"never smoked"* (tidak pernah merokok) dan *"smokes"* (merokok) juga berkontribusi pada risiko *stroke*, namun dengan pengaruh yang lebih kecil dibandingkan dengan status *"formerly smoked"*. Status *"unknown"* (tidak diketahui) memiliki pengaruh yang lebih rendah dibandingkan dengan ketiga status merokok tersebut.

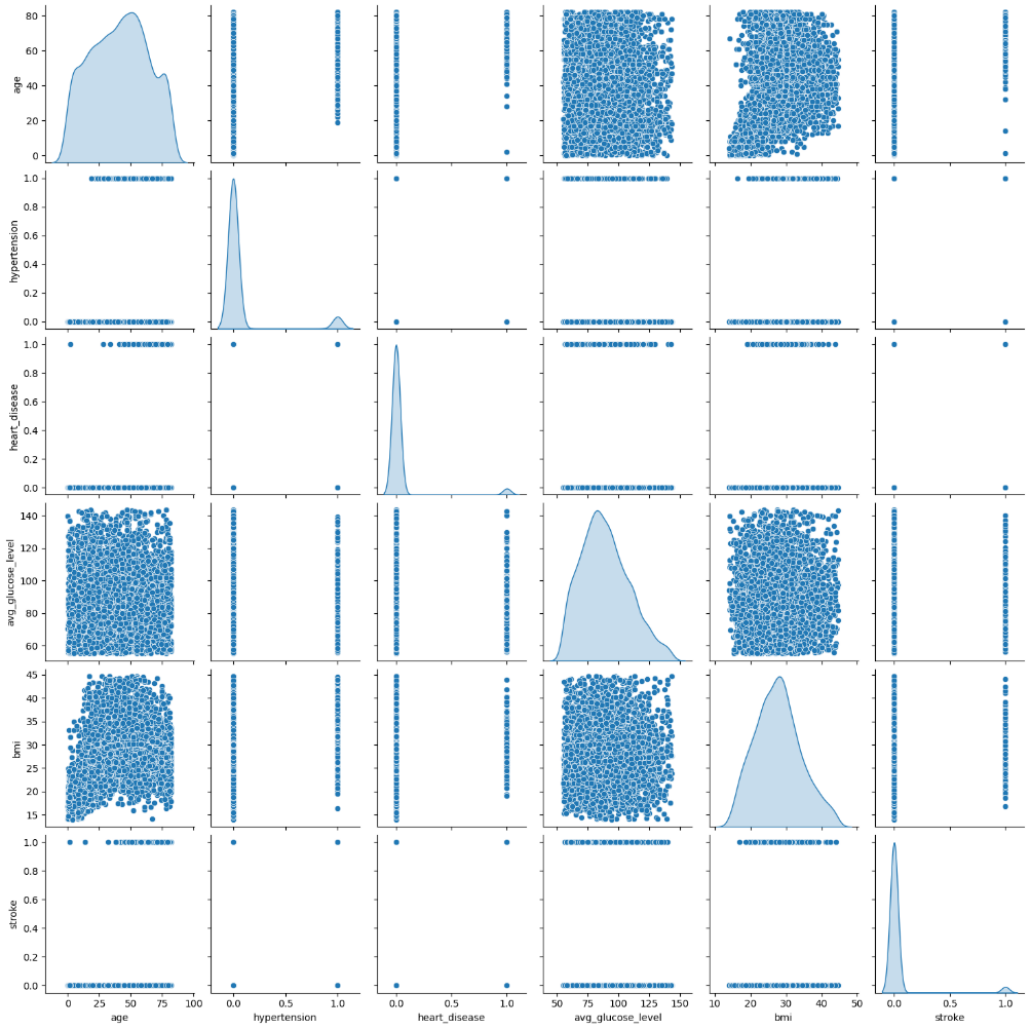
Output :



Gambar 4. 30 Diagram Catplot Stroke – Smoking_Status

b) *Numerical Features*

Korelasi antara fitur numerik terhadap fitur numerik dan antara fitur numerik dengan target menunjukkan adanya hubungan, meskipun tidak terlalu kuat. Untuk meningkatkan efisiensi model dan mengurangi kompleksitas data, akan dilakukan reduksi dimensi untuk menyederhanakan variabel yang terlibat.



Gambar 4. 31 Diagram Korelasi Numerical Features

7) *Matrix* Korelasi

Untuk dapat melihat niai korelasi antar fitur melalui diagram *heatmap*. Korelasi antara *age* dan bmi nilai korelasinya 0.39 merupakan korelasi positif lemah. Ketika usia (*age*) meningkat, nilai BMI cenderung sedikit

meningkat. Namun, hubungan ini tidak terlalu kuat, sehingga keduanya masih dapat dianggap cukup independen.

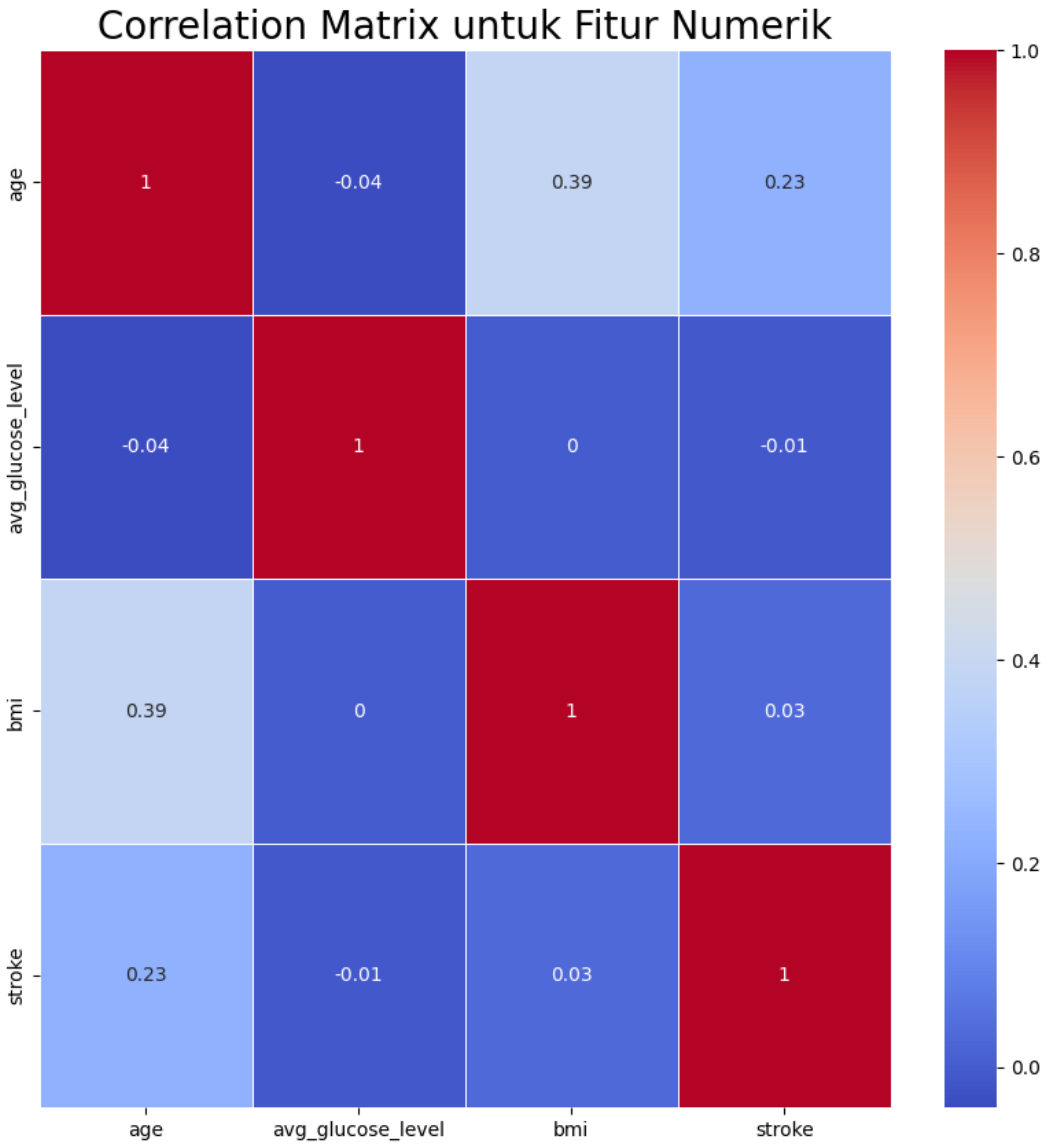
Korelasi antara *age* dan *avg_glucose_level* nilai korelasi -0.04 merupakan korelasi negatif yang sangat lemah. Tidak ada hubungan signifikan antara usia dan tingkat glukosa rata-rata**. Korelasi antara *age* dan *stroke*** nilai korelasi 0.23, merupakan korelasi positif lemah. Ini menunjukkan bahwa peningkatan usia mungkin sedikit berhubungan dengan peningkatan kejadian *stroke*, tetapi hubungan ini lemah.

Korelasi antara *avg_glucose_level* dan bmi nilai korelasi 0, tidak ada korelasi. Ini berarti perubahan tingkat glukosa rata-rata tidak berhubungan sama sekali dengan perubahan BMI. **Korelasi antara *avg_glucose_level* dan *stroke*** nilai korelasi -0.01, korelasi negatif yang sangat lemah. Tidak ada hubungan signifikan antara tingkat glukosa rata-rata dan kejadian stroke. **Korelasi antara bmi dan *stroke*** nilai korelasi 0.03, korelasi positif yang sangat lemah. Tidak ada hubungan signifikan antara BMI dan kejadian stroke.

Korelasi Positif Lemah *age* dengan *bmi* dan *age* dengan *stroke* menunjukkan korelasi positif lemah. Artinya, ada sedikit kecenderungan bahwa seiring bertambahnya usia, BMI dan risiko *stroke* juga meningkat, tetapi hubungan ini tidak kuat.

Korelasi Lemah atau Tidak Ada: *avg_glucose_level* dengan fitur lainnya (*age, bmi,* dan *stroke*) menunjukkan korelasi yang sangat lemah atau tidak ada. Ini menunjukkan bahwa perubahan tingkat glukosa rata-rata tidak banyak berhubungan dengan fitur lain dalam data ini.

Pada diagram *heatmap* dapat dilihat bahwa fitur *age* dan *bmi* memiliki korelasi namun tidak kuat sehingga dapat dilakukan *dimensional reduction.*



Gambar 4. 32 Matrix Korelasi

4.1.2 Data *Preparation*

Pada tahap data *preparation*, perlu mengubah nilai numerik dalam kolom *hypertension, heart_disease,* dan *stroke* pada *dataframe* bernama data menjadi representasi teks. Pada *dataframe* tersebut, nilai 0 diganti dengan 'No', dan nilai 1 diganti dengan *'Yes'*. Digunakan untuk meningkatkan interpretabilitas data, saat melakukan analisis data atau visualisasi. Dengan mengganti nilai numerik dengan teks, bisa lebih mudah memahami dan menjelaskan status kesehatan individu terkait hipertensi, penyakit jantung, dan *stroke* berdasarkan data tersebut.

a. Define Function

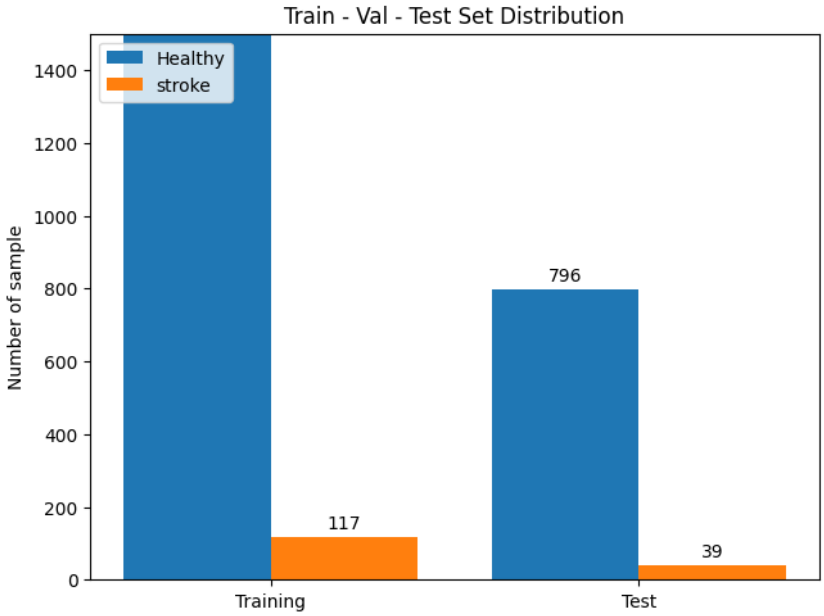
Dalam *Machine Learning* data dapat digunakan oleh model dengan efektif jika data berada pada rentang 0-1, oleh karena itu perlu mengubah nilai fitur numerik agar semua fitur dalam skala yang sama dan dapat membuat algoritma *Machine Learning* lebih mudah dalam memahami hubungan antar fiturnya. Fungsi *scaling* dan *encoding* ini membantu mempersiapkan data sehingga model pembelajaran mesin dapat bekerja dengan lebih efektif.

Normalisasi (*scaling*) membantu menjaga konsistensi antar fitur numerik, sementara encoding mengubah fitur kategori menjadi bentuk yang dapat diproses oleh algoritma pembelajaran mesin. Dengan menyimpan objek *scaler* dan *encoder*, dapat memastikan bahwa transformasi yang sama diterapkan pada data baru di masa depan, menjaga integritas dan konsistensi model pembelajaran. Langkah-langkah ini penting untuk memaksimalkan akurasi dan efisiensi model dalam mengolah data yang diberikan.

b. Spliting Data

Tahap berikutnya adalah membagi dataset menjadi dua bagian, yaitu data latih (*training set*) dan data uji (*test set*) rasio perbandingannya 80:20, proses ini penting untuk mengevaluasi performa model secara objektif. Data latih digunakan untuk membangun dan mengoptimalkan model, sementara data uji digunakan untuk mengevaluasi seberapa baik model dapat menggeneralisasi data baru yang tidak terlihat selama pelatihan.

Dengan pendekatan ini, dapat meminimalkan risiko overfitting, di mana model berperforma baik pada data latih tetapi kurang efektif pada data baru. Pembagian data yang tepat sangat penting untuk memastikan bahwa hasil evaluasi model adalah representasi akurat dari performa sebenarnya.



Gambar 4. 33 Test Set Distribution

c. *Encoding Fitur Kategori*

Dalam proses ini, *LabelEncoder* digunakan untuk mengubah label kategori menjadi bentuk numerik agar dapat digunakan oleh model pembelajaran mesin. Dengan menyimpan *encoder* menggunakan *joblib*, dapat memastikan bahwa label dari data pengujian dan data baru dapat di-transformasi secara konsisten dengan cara yang sama seperti data pelatihan. Langkah ini penting untuk menjaga integritas dan akurasi model saat menangani data baru atau saat mengulangi eksperimen.

d. *Normalisasi Data*

Diketahui bahwa data numerik memiliki skala yang berbeda-beda, agar data tersebut nilainya dapat dikomputasi dengan mudah dan lebih cepat pada model maka data atau fiturnya harus dibubah menjadi skala yang sama dengan mengubah menjadi rentang skala 0-1.

Variabel *avg_glucose_level* kemungkinan besar merujuk pada rata-rata kadar glukosa dalam tubuh, sedangkan pc_1 merupakan nilai yang dihasilkan dari analisis komponen utama (PCA) yang menggambarkan

dimensi baru yang merangkum variasi dalam data. Nilai *avg_glucose_level* bervariasi mulai dari 0.617 hingga 0.282, sedangkan nilai pc_1 berkisar antara 0.594 hingga 0.193. Tabel ini menunjukkan bahwa ada variasi tingkat glukosa rata-rata yang terasosiasi dengan dimensi pertama dari analisis komponen utama.

Output :

	avg_glucose_level	pc_1
554	0.617110	0.594160
2661	0.713881	0.135229
2413	0.373711	0.620877
1954	0.418584	0.620737
1945	0.282833	0.193192

Gambar 4. 34 Output Normalisasi Data

Setelah itu dilakukan pengecekan terhadap *statistic* data numerik pada data *frame* dengan menghitung statistik deskriptifnya. Dari analisis statistik deskriptif bahwa kedua variabel memiliki rentang nilai dari 0 hingga 1. Distribusi data *avg_glucose_level* lebih terpusat di bawah rata-rata dibandingkan dengan pc_1, yang memiliki nilai median yang lebih mendekati rata-rata.

Hal ini menunjukkan bahwa data pada kedua variabel telah dinormalisasi. Selain itu, variasi nilai lebih tinggi pada pc_1 dibandingkan avg_glucose_level, berdasarkan standar deviasi masing-masing. Analisis ini memberikan wawasan awal untuk mengeksplorasi lebih lanjut hubungan antara variabel ini dan bagaimana mempengaruhi model.

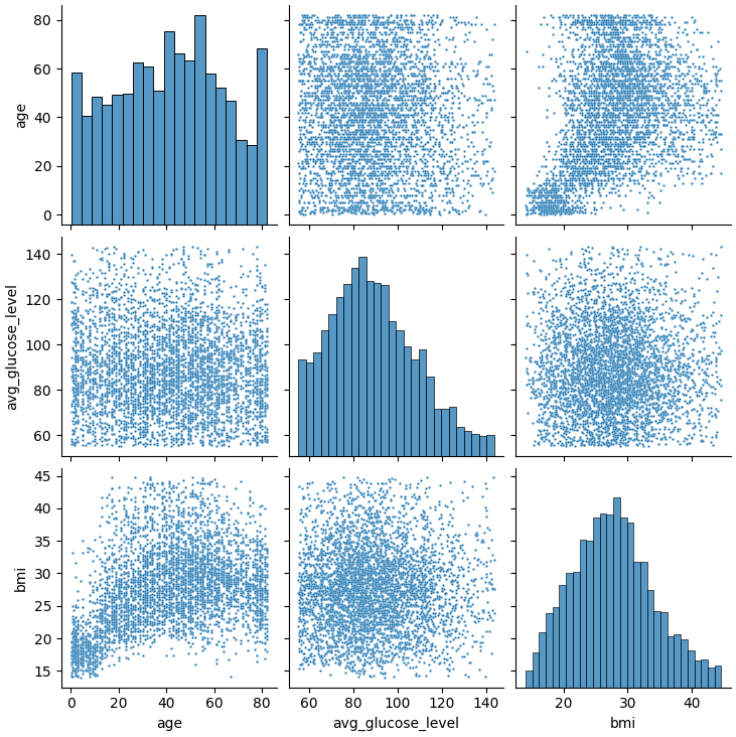
	avg_glucose_level	pc_1
count	3339.0000	3339.0000
mean	0.3862	0.4942
std	0.2197	0.2709
min	0.0000	0.0000
25%	0.2233	0.2824
50%	0.3619	0.4803
75%	0.5311	0.7114
max	1.0000	1.0000

Gambar 4. 35 Output Descriptif Normalisasi Data

e. Dimensional Reduction

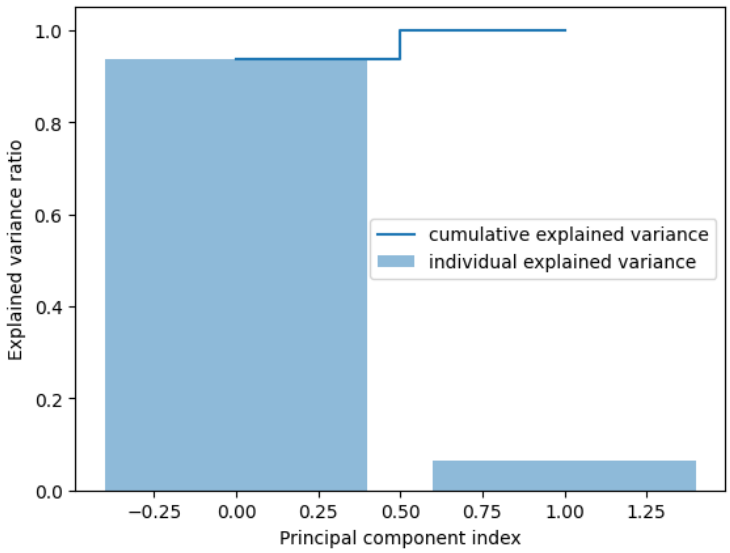
Dimensional Reduction untuk menyederhanakan data dengan menghilangkan fitur yang tidak perlu atau tidak relevan. Ini sangat berguna ketika memiliki data dengan banyak fitur (berdimensi tinggi), karena dapat membuat proses analisis lebih efisien, mengurangi risiko *overfitting* (model terlalu cocok dengan data latihan), dan memudahkan visualisasi.

Teknik ini juga mempercepat waktu komputasi dan meningkatkan kinerja model dengan fokus pada fitur yang paling penting. Dengan *dimensional reduction*, bisa lebih mudah memahami data, menemukan pola tersembunyi, dan meningkatkan hasil model pembelajaran mesin. Untuk meningkatkan efisiensi model dan mengurangi kompleksitas data, maka dilakukan reduksi dimensi untuk menyederhanakan variabel menggunakan PCA.



Gambar 4. 36 Dimensional Reduction

Hasil PCA diketahui bahwa *feature* 'bmi' dan 'age' dilakukan *dimensionality reductioon*, jumlah komponen yang paling ideal ialah berjumlah 1 komponen *principal*. Meskipun nilai korelasi antar dua *feature* tersebut *relative* lemah.



Gambar 4. 37 PCA

Penggunaan PCA membantu dalam menyederhanakan dataset dengan fitur yang berkorelasi tinggi, menghasilkan dataset yang lebih ramping dan informatif. Dengan menambahkan komponen utama ke *dataset* yaitu pc_1 dan menghapus fitur asli yang berkorelasi, analisis dapat dilakukan lebih efektif dan efisien.

Gambar menunjukkan grafik ***Explained Variance Ratio*** dari hasil analisis ***Principal Component Analysis*** **(PCA)**. Grafik ini digunakan untuk menggambarkan seberapa banyak variabilitas dalam data yang bisa dijelaskan oleh masing-masing komponen utama (*principal component*).

Pada grafik ini, batang biru muda mewakili ***individual explained variance***, yang menunjukkan proporsi varians yang dijelaskan oleh masing-masing komponen utama secara terpisah. Komponen pertama memiliki *explained variance* yang sangat tinggi, hampir 100%, yang menunjukkan bahwa sebagian besar informasi atau variabilitas dalam data dapat diwakili oleh komponen ini.

Sedangkan komponen kedua hanya menjelaskan sebagian kecil dari varians. Garis biru di atasnya menggambarkan **cumulative explained variance**, yang menunjukkan total varians yang dijelaskan ketika komponen-komponen tersebut digabungkan. Dengan garis bitu, bisa melihat bahwa dengan hanya satu komponen utama, hampir semua variabilitas data sudah tercakup. Hal ini menunjukkan bahwa PCA berhasil mereduksi dimensi data secara signifikan tanpa kehilangan banyak informasi, menjadikannya sangat efektif untuk digunakan dalam analisis data yang lebih lanjut atau untuk keperluan visualisasi.

4.1.3 *Modeling*

Penelitian ini menggunakan metode *Random Forest*. Adapun pada metode *Random Forest* melakukan *Hyperparamater Tunning* dengan proses sebagai berikut.

a) Training

Proses pelatihan model melibatkan penyesuaian parameter untuk meminimalkan *error* dengan menggunakan data pelatihan untuk menemukan pola atau hubungan antara fitur dan target. Setelah model dilatih, metrik evaluasi digunakan untuk menilai performa model pada data pelatihan dan data pengujian. *Dataframe* yang disiapkan berfungsi untuk menyimpan dan membandingkan metrik evaluasi dari model *RandomForest*.

b) Hyperparamater Tuning

Menggunakan *Hyperparameter Tuning* untuk mengoptimalkan performa model dengan mencari kombinasi *Hyperparameter* yang paling efektif. Parameter yang di-tune meliputi **n_estimators** yaitu jumlah pohon keputusan dalam *Random Forest*, nilai yang diuji adalah 50, 100, dan 200. **max_features** jumlah fitur yang akan dipertimbangkan saat membagi node, nilai yang diuji adalah *'auto'* (semua fitur), 'sqrt' (akar kuadrat dari jumlah fitur), dan 'log2' (logaritma basis 2 dari jumlah fitur).

max_depth kedalaman maksimum pohon keputusan, nilai yang diuji adalah none (kedalaman tidak terbatas) serta 10, 20, dan 30. ***Criterion*** kriteria untuk mengukur kualitas split, nilai yang diuji adalah *'gini'* (*indeks Gini*) dan *'entropy'* (entropi). ***Bootstrap*** menentukan apakah bootstrap samples akan digunakan saat membangun pohon, nilai yang diuji adalah *true* (menggunakan *sampling bootstrap*) dan *false* (tidak menggunakan *sampling bootstrap*).

1) Hyperparameter Search

Menjalankan proses pelatihan model dengan menggunakan *GridSearchCV*. Setelah *GridSearchCV* selesai menjalankan proses pencarian, atribut *best_params_* menyimpan kombinasi *hyperparameter* yang memberikan performa terbaik berdasarkan

hasil *cross-validation.* Hasil *tuning* menunjukkan bahwa kombinasi parameter terbaik adalah:

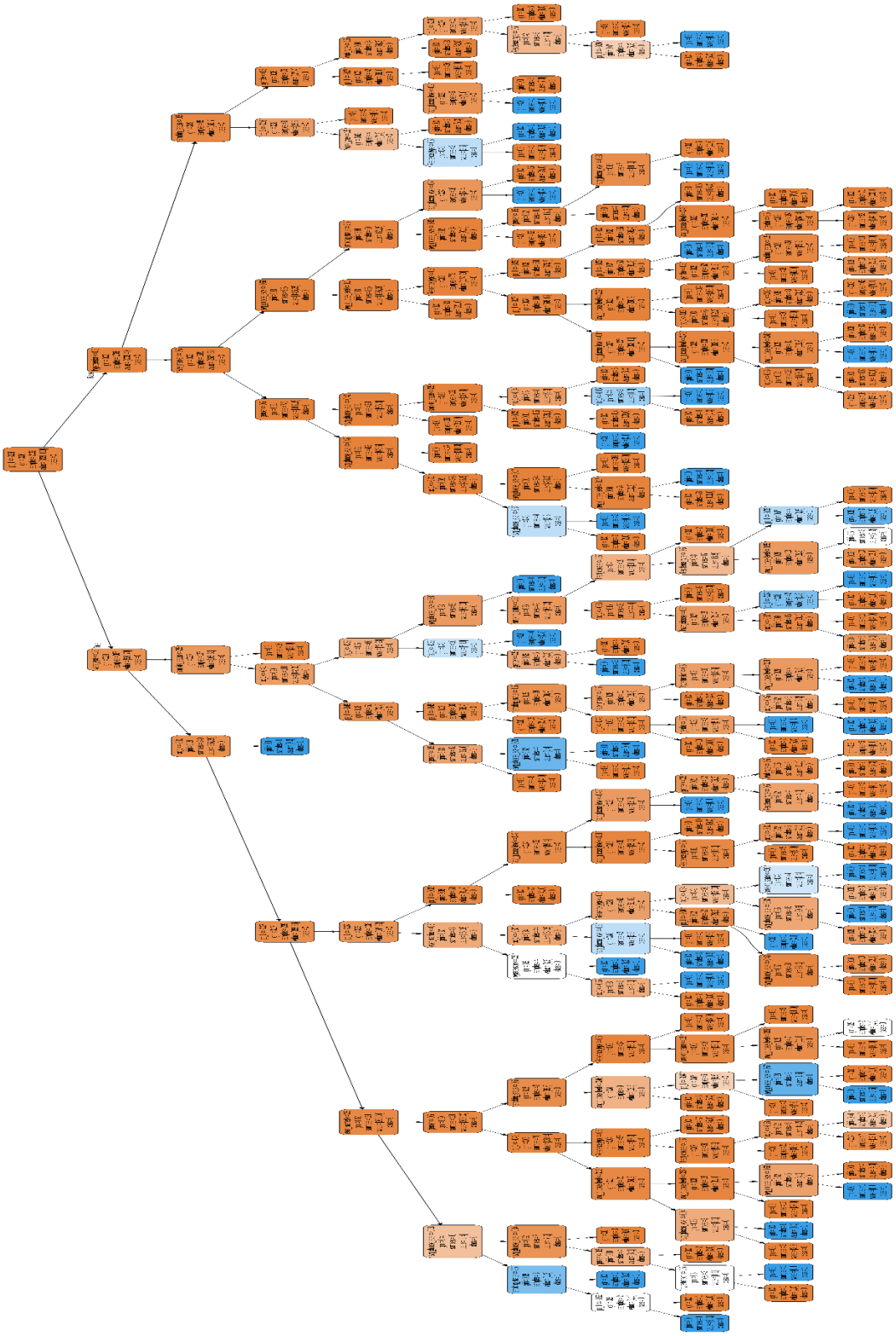
Tabel 4. 2 Hasil *Tuning* Paramater Terbaik

Name	Value
bootstrap	*True*
criterion	gini
max_depth	10
max_features	sqrt
n_estimators	100

Parameter ini dianggap sebagai konfigurasi yang optimal untuk meningkatkan performa model *Random Forest*.

2) Re-Training

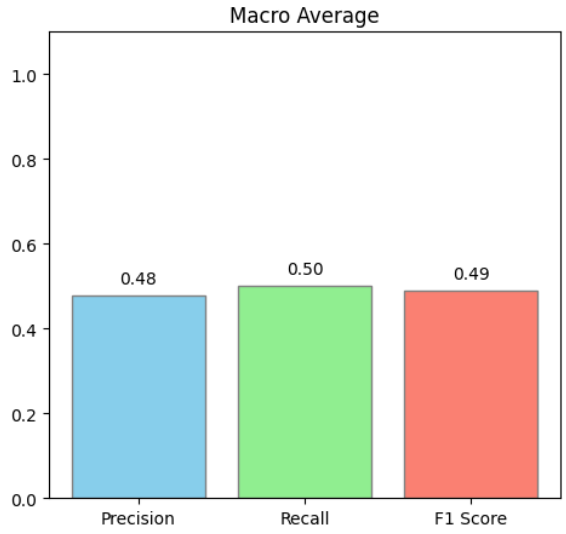
Setelah *GridSearchCV* menemukan kombinasi *hyperparameter* terbaik, *best_estimator* menyimpan model yang dilatih dengan parameter terbaik tersebut. Ini adalah model yang telah dioptimalkan melalui proses tuning. Selanjutnya dilatih kembali *model Random Forest Classifier* yang sudah dioptimalkan pada data pelatihan (*X_train* dan *y_train).* Ini berarti model akhir yang akan digunakan untuk prediksi adalah model dengan konfigurasi *hyperparameter* terbaik yang ditemukan oleh *GridSearchCV*. Setelah itu menggunakan model yang telah dilatih untuk melakukan prediksi pada data pengujian (*X_test*). y_pred merupakan *array* yang berisi prediksi model untuk setiap sampel dalam data pengujian.



Gambar 4. 38 Random Forest

4.1.4 *Evaluation*

Rata-rata metrik kinerja dari sebuah model klasifikasi: *Macro Average*, *macro average* menampilkan nilai rata-rata dari *precision, recall,* dan F1-*score*, tanpa mempertimbangkan proporsi masing-masing kelas. Dalam hal ini*, precision, recall*, dan F1-*score* berada di sekitar nilai 0.48 hingga 0.50, menunjukkan performa yang cukup rendah, terutama jika dataset tidak seimbang. Ini menunjukkan bahwa model mungkin bekerja lebih baik pada kelas yang lebih dominan, sementara kelas minoritas mungkin kurang terwakili dengan baik.



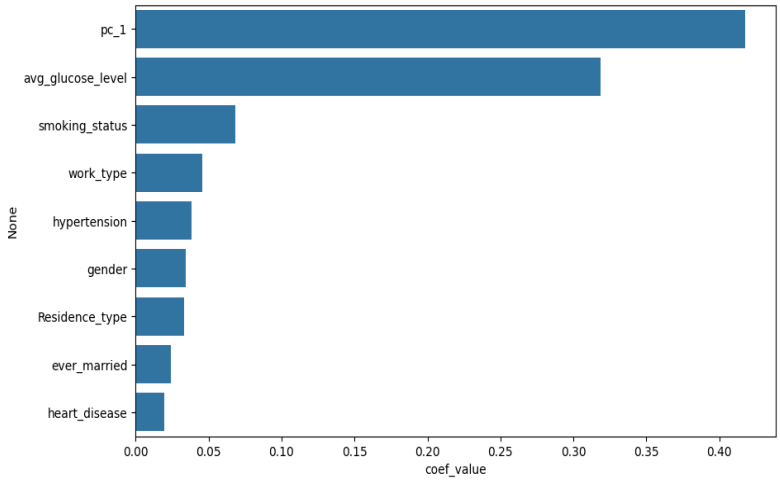
Gambar 4. 39 Macro Average

a) Menampilkan *Best Feature*

Tabel yang diberikan menunjukkan nilai koefisien dari berbagai fitur dalam model, yang mungkin merupakan model regresi linier atau model lain yang menggunakan koefisien untuk menentukan pengaruh fitur terhadap hasil prediksi.

Kesimpulan: pc_1 dan *avg_glucose_level* memiliki koefisien yang paling tinggi, menunjukkan bahwa mereka memiliki pengaruh terbesar terhadap prediksi model. Fitur-fitur lainnya memiliki koefisien yang lebih kecil, menunjukkan pengaruh yang lebih rendah terhadap hasil prediksi. Ini memberikan

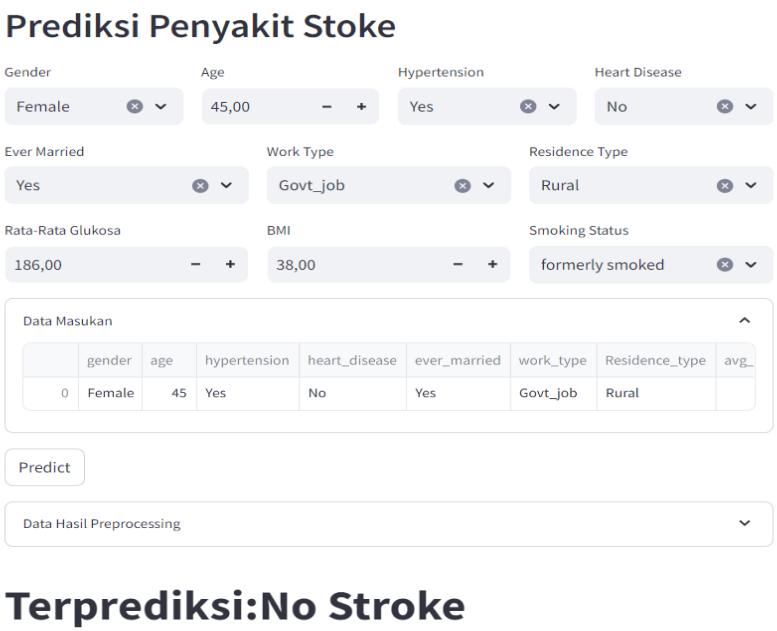
wawasan tentang fitur-fitur mana yang paling berkontribusi dalam model dan bisa menjadi dasar untuk interpretasi atau penyesuaian model lebih lanjut.



Gambar 4. 40 Best Feature

4.1.5 Testing /Deployment

Berdasarkan hasil performa maka akan diujicoba langsung melalui *platform streamlit* untuk memprediksi waktu kelangsungan hidup pasien berdasarkan data masukan.



Gambar 4. 41 Deployment

Gambar 4.41 merupakan uji coba sistem hasil dari *deployment* model, pada uji coba ini memasukan inputan sesuai yang telah disediakan, setelah memasukkan semua indikator lalu klik tombol *predict* sehingga tampil klasifikasi yang terdapat pada gambar 4.41 yang menunjukan bahwa pasien tidak terindikasi penyakit *stroke* berdasarkan indikator yang dimasukan.

4.2 Pembahasan

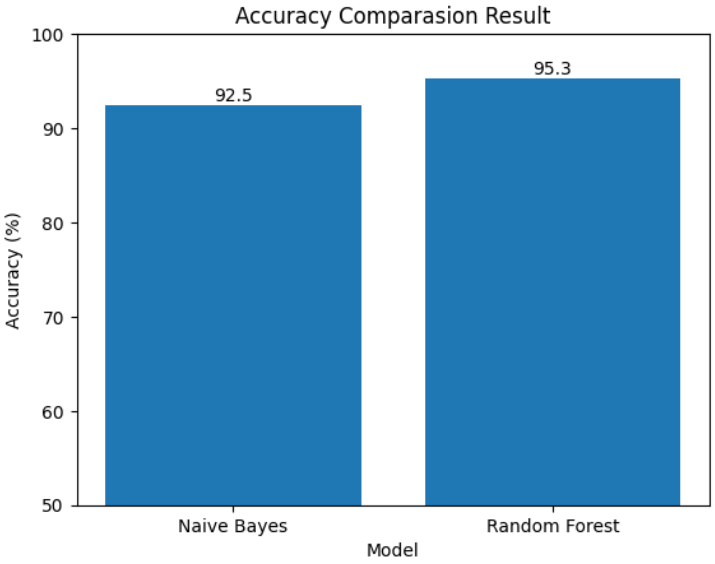
Prediksi risiko penyakit *stroke* dengan metode *Random Forest* memberikan performa yang cukup baik dalam menentukan apakah seseorang terkena penyakit *stroke* atau tidak. Hal ini didasarkan pada nilai metrik *precision, recall,* dan F1-*score*. Yang mencapai performa dimana masing-masing metrik memiliki nilai 47,66 % *precision*, 50% *recall*, dan 48,80% F1-*score*.

Beberapa metrik tersebut dicomparasi dengan performa model oleh Agus Fajar Riany dan Gusmelia Testiana pada penelitiannya yang menggunakan penerapan data mining pada algoritma *naïve bayes* [17] dengan kasus serupa yaitu penyakit *stroke*.

a. Accuracy

Metrik *accuracy* menentukan performa model dalam memprediksi sampel yang terprediksi secara benar dari total sampel yang diujicoba ke dalam model. Artinya semakin banyak prediksi yang benar pada sampel yang diujicoba maka nilai *accuracy* model semakin tinggi.

Nilai *Accuracy* yang didapatkan oleh *Random Forest* mencapai 95,3 % pada test set. Total keseluruhan *test set* yang digunakan adalah sejumlah 835 sampel yang berarti dari 835 sample yang termasuk didalamnya sampel *stroke* dan sehat, 796 sampel terprediksi / terklasifikasi oleh model dengan benar dan 39 diantaranya terprediksi salah. Bila dibandingkan dengan penelitian [17] performa accuracy dapat dilihat pada gambar 4.42

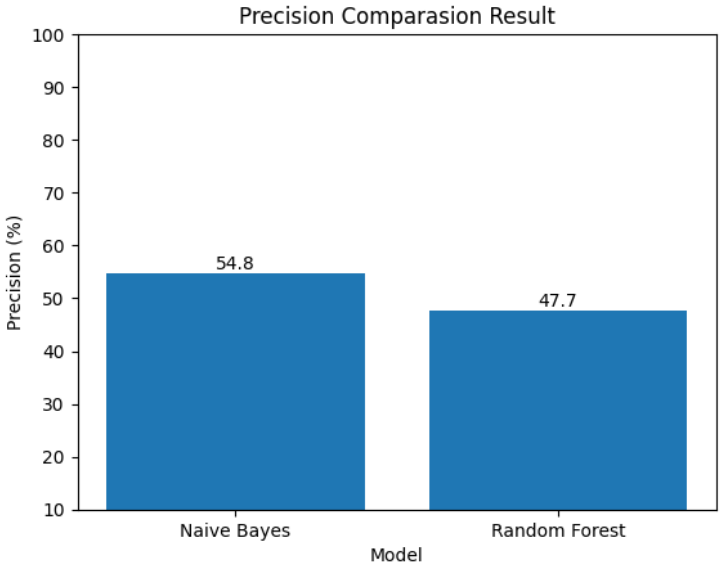


Gambar 4. 42 Bar Chart Comparation Accuracy

b. *Precision*

Metric Precision menentukan performa model dalam memprediksi sampel yang terprediksi sebagai *stroke* dengan benar dari keseluruhan sampel terprediksi *stroke*. Artinya semakin banyak sampel terprediksi *stroke* dengan benar dari keseluruhan sample yang terprediksi *stroke* maka *Precision* semakin tinggi.

Nilai *Precision* yang didapatkan oleh *Random Forest* mencapai 47.7% pada test set. Total keseluruhan *test set* yang digunakan adalah sejumlah 39 sampel. Artinya 19 sampel terprediksi oleh model dengan benar dan 20 diantaranya terprediksi salah. Bila dibandingkan dengan penelitian [17] performa *Precision* dapat dilihat pada gambar 4.43

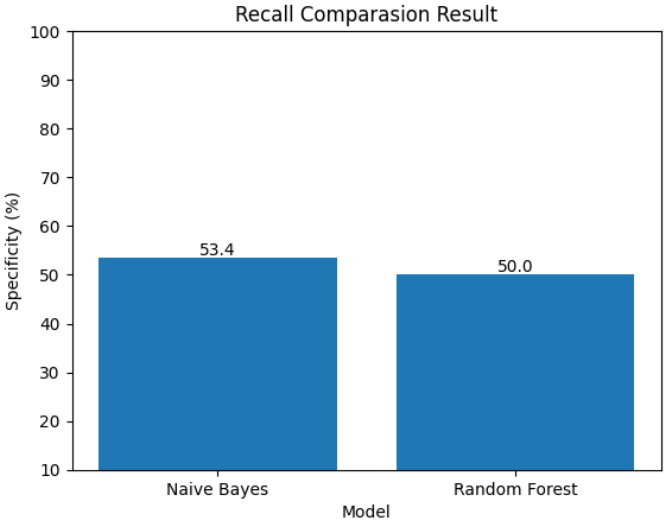


Gambar 4. 43 Bar Chart Comparation Precission

c. *Recall*

Metric *recall* menentukan performa model dalam memprediksi sampel yang tidak terprediksi sebagai *stroke* dengan benar dari keseluruhan sampel terprediksi *stroke*. Artinya semakin banyak sampel terprediksi *stroke* dengan benar dari keseluruhan sample yang terprediksi *stroke* maka *recall* semakin tinggi.

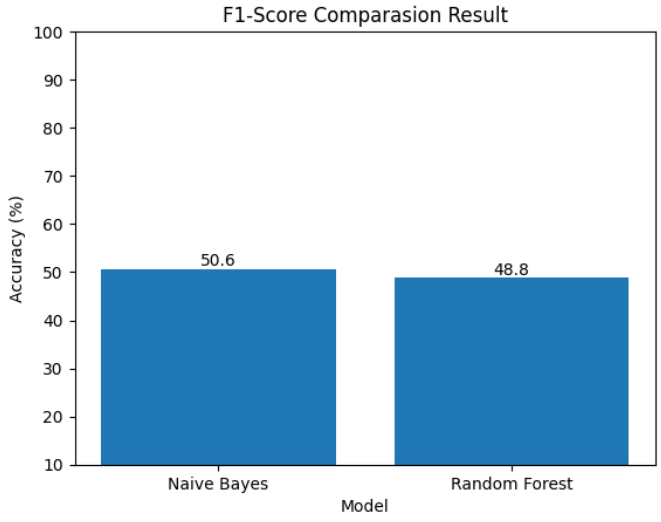
Nilai *recall* yang didapatkan oleh *random forest* mencapai 50.0% pada test set. Total keseluruhan test set yang digunakan adalah sejumlah 796 sampel. Artinya 398 sampel terprediksi oleh model dengan benar dan 398 diantaranya terprediksi salah. Bila dibandingkan dengan penelitian [17] performa *Recall* dapat dilihat pada gambar 4.44.



Gambar 4. 44 Bar Chart Comparation Recall

d. *f1-score*

Metrik *F1-Score* menentukan performa model dalam keseimbangan terprediksi secara benar positif dan benar negatif *stroke*. Artinya semakin banyak sample yang terprediksi benar positif *stroke* dan benar negatif maka nilai *F1-Score* semakin tinggi (mencapai keseimbangan). *F1-Score* yang didapatkan oleh *Random Forest* adalah sebesar 48.8% .



Gambar 4. 45 Bar Chart Comparation F1-score