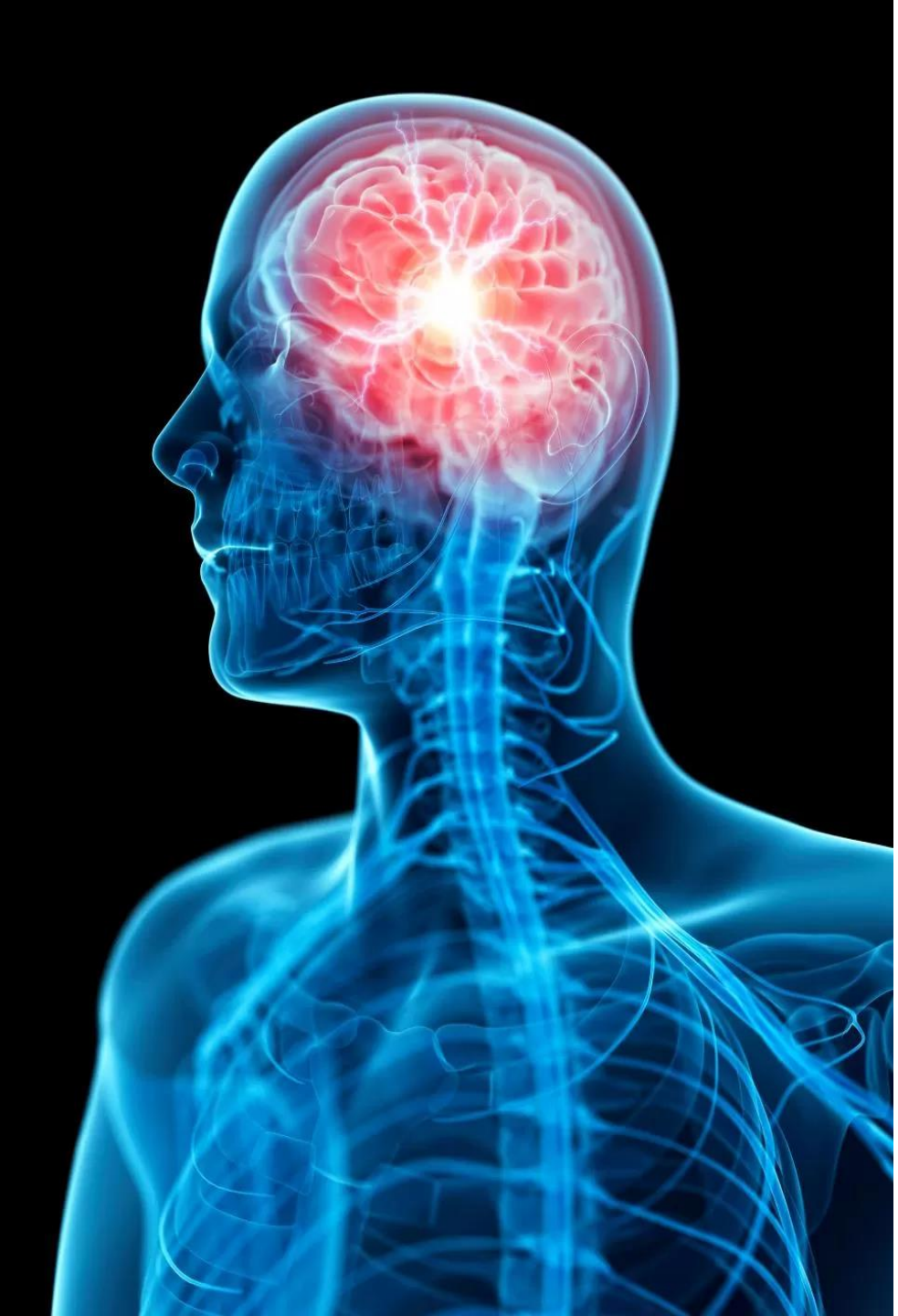


Prediksi Pasien Berpotensi Terkena Stroke

Final Project :

Adi Kurnia Wardhana

Daftar Isi :

Tujuan :

Final project ini bertujuan untuk mengembangkan model yang memperkirakan kemungkinan seseorang mengalami stroke berdasarkan beberapa faktor risiko penyebab stroke.

Gambaran Masalah :

Stroke merupakan salah satu penyakit serius yang dapat menyebabkan cacat dan kematian. Identifikasi faktor-faktor resiko yang dapat memprediksi kemungkinan seseorang terkena stroke dapat membantu dalam pencegahan kondisi ini, dengan harapan dapat menghasilkan model yang akurat untuk membantu dalam penanganan dan pencegahan stroke.

Deskripsi Data

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

Dataset yang digunakan terdiri dari 5110 baris dan 12 kolom, dimana setiap kolom berisi informasi sebagai berikut :

- id : identifikasi unik untuk setiap entri dalam dataset.
- gender : variabel ini mencatat jenis kelamin pasien.
- age : variabel ini menunjukkan usia pasien dalam tahun.
- hypertension : variabel yang menunjukkan apakah pasien memiliki riwayat hipertensi (tekanan darah tinggi) atau tidak.
- heart_disease : variabel yang menunjukkan apakah pasien memiliki riwayat penyakit jantung atau tidak.
- ever_married : variabel ini menunjukkan apakah pasien pernah menikah atau tidak.
- work_type : veriabel yang menunjukkan jenis pekerjaan.
- Residence_type : variabel ini menunjukkan tipe tempat tinggal pasien.
- avg_glucose_level : variable ini menunjukkan rata-rata tingkat glukosa dalam darah pasien.
- bmi : variable ini menunjukkan Indeks Massa Tubuh (BMI) pasien, yang mengukur proporsi berat badan terhadap tinggi badan.
- smoking_status : variable ini menunjukkan status merokok pasien.
- stroke : variabel target yang menunjukkan apakah pasien pernah mengalami stroke

Prapemrosesan Data

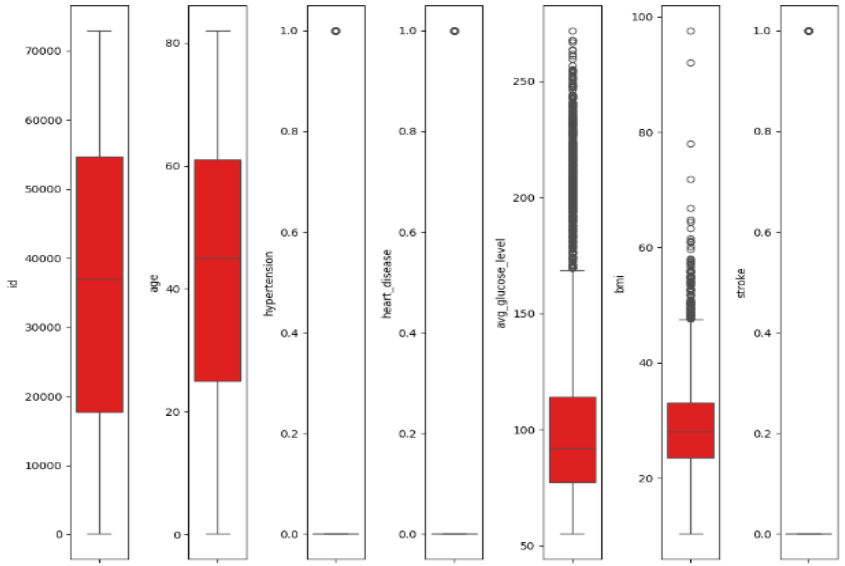
Info Data

```
data.info()
```

```
<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   int64
 4   heart_disease      5110 non-null   int64
 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   int64
dtypes: float64(3), int64(4), object(5)
memory usage: 479.2+ KB
```

- Dataset terdiri dari 5110 baris dan 12 Kolom.
- Teridiri dari 7 data numericals dan 5 data categoricals.
  - numericals = id, age, hypertension, heart_disease, avg_glucose_level, bmi, stroke.
  - categoricals = gender, ever_married, work_type, Residence_type, smoking_status.
- Ada missing value pada kolom bmi.
- Tidak ada duplikat data.
- Target feature : stroke

Cek Outliers



- Kolom id dan age tidak terdeteksi outliers.
- Kolom hypertension, heart_disease dan stroke terdeteksi outliers.
- Kolom avg_glucose_level dan bmi terdeksi banyak outliers

Cek Missing Value

```
data.isna().sum()
```

```
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
```

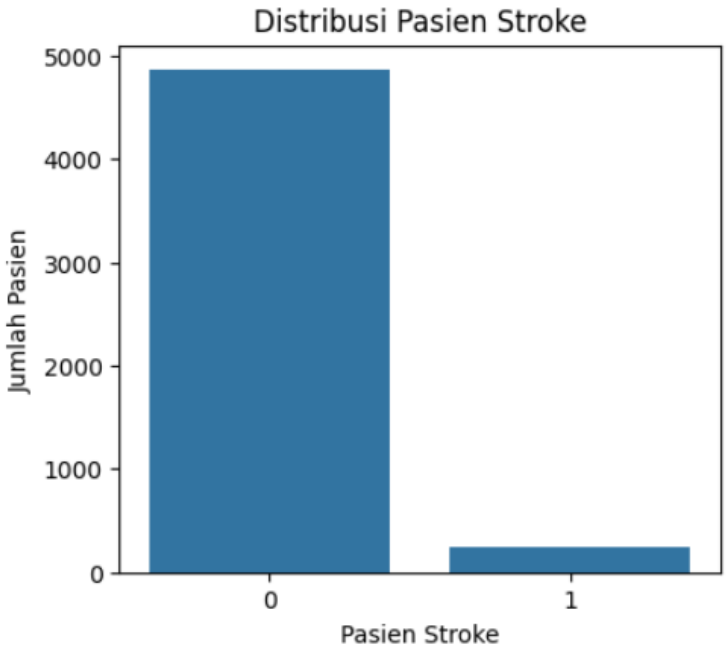
- Terdapat missing value pada kolom ‘bmi’ sebanyak 201 atau 3.93%.

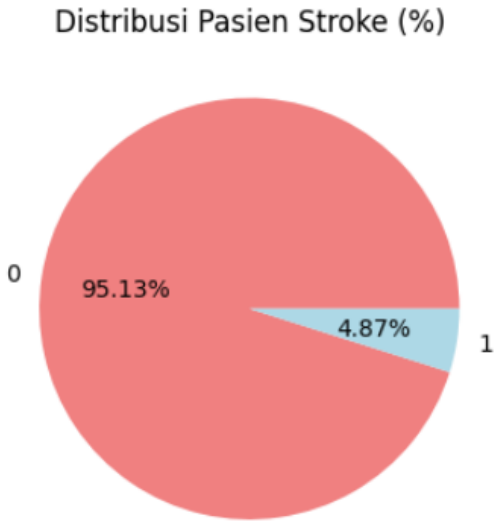
Handling Outliers dan Missing Value

- Setelah dilakukan pengecekan lebih lanjut didapat ada sebanyak 627 outliers pada kolom ‘avg_glucose_level’ dan 110 outliers pada kolom ‘bmi’. Kedua data tersebut tidak dilakukan drop data karena merupakan data yang dapat berpengaruh terhadap kejadian stroke.
- Kolom ‘hypertension’, ‘heart_disease’ dan ‘stroke’ hanya memiliki 2 kondisi yaitu 0 (minimum) dan 1 (maksimum). Kondisi 0 memiliki jumlah jauh lebih banyak dibanding dengan kondisi 1, sehingga 1 terdeteksi sebagai outliers. Ketiga data tersebut tidak dilakukan drop data karena terbilang wajar.
- Untuk missing value pada kolom ‘bmi’ dilakukan imputasi dengan median dan tidak dilakukan drop data.

Exploratory Data Analysis (EDA)

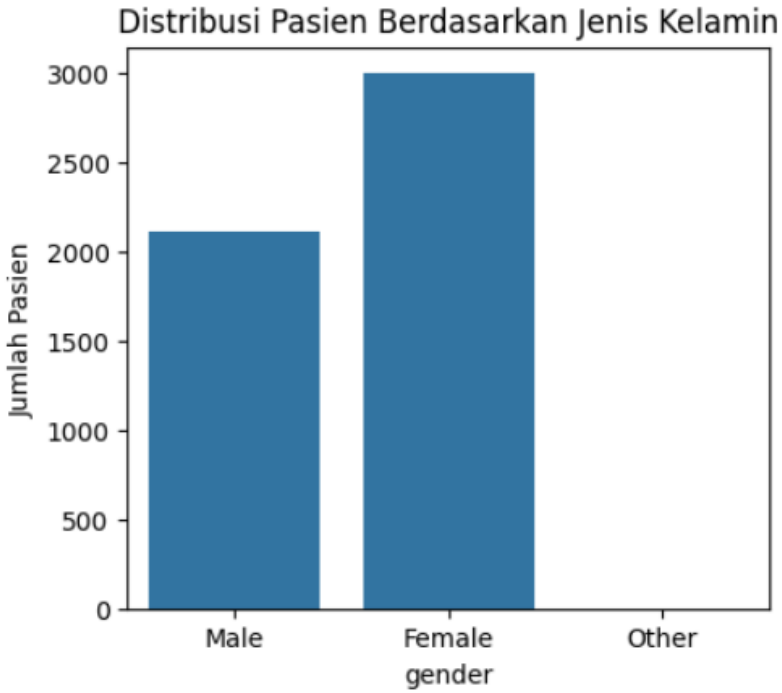
- **Distribusi Pasien Stroke**

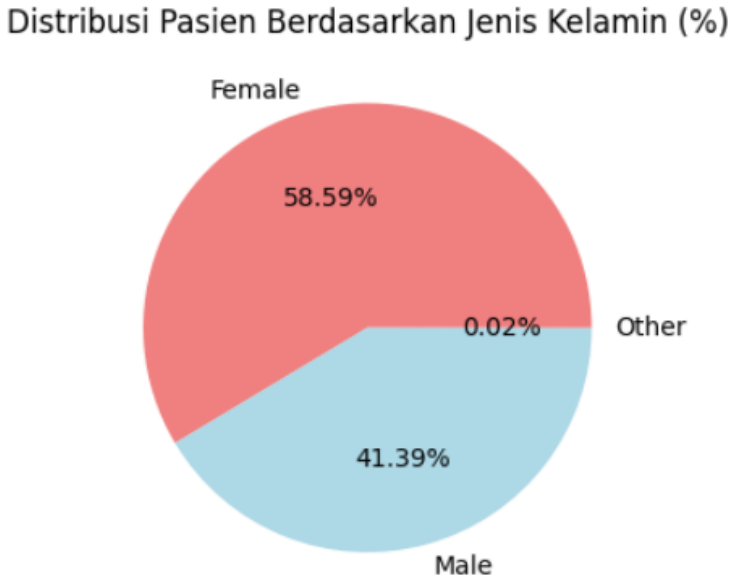




Dari total data sebanyak 5110 pasien terdapat 4861 pasien yang tidak stroke dan ada 249 pasien stroke.

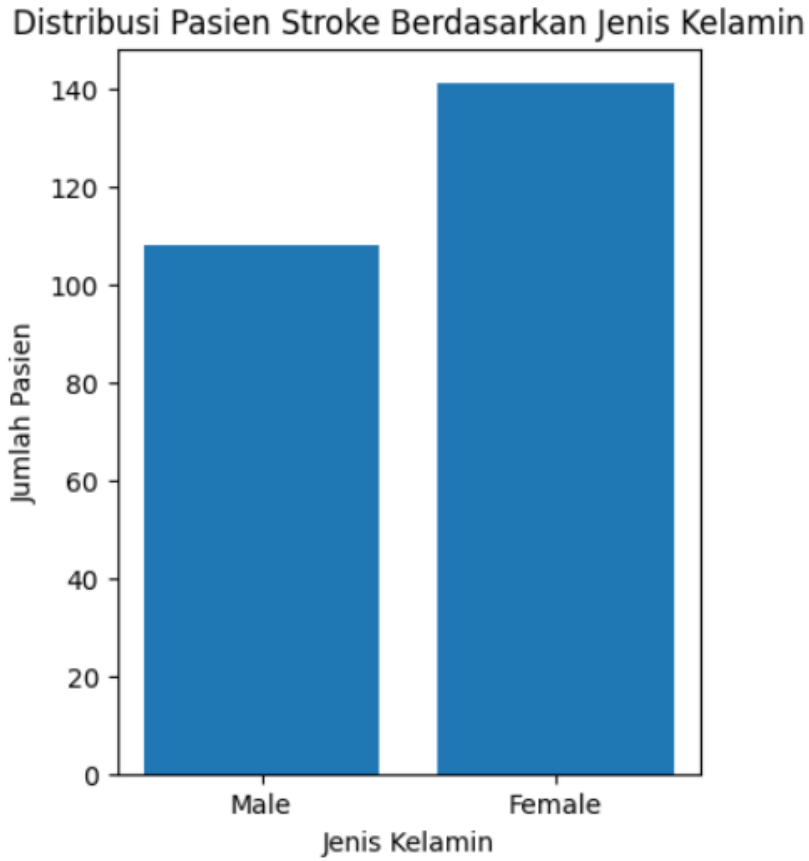
❑ Distribusi Pasien Berdasarkan Jenis Kelamin





- Jumlah pria 41.39% dan wanita 58.59% lebih banyak 17.2% dibanding pria.

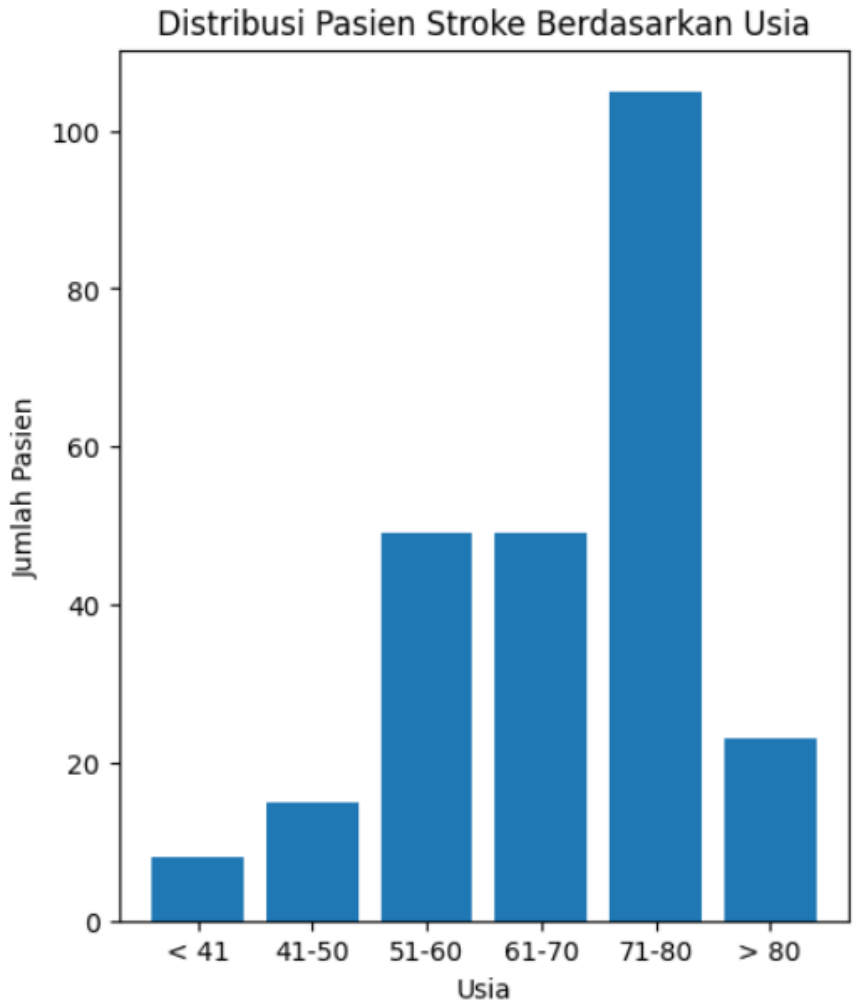
❑ Distribusi Pasien Stroke Berdasarkan Jenis Kelamin




```
Jumlah pasien pria terkena stroke   : 108 pasien
Jumlah pasien wanita terkena stroke : 141 pasien
```

- Dari data terlihat bahwa wanita lebih banyak terkena stroke dibanding pria.

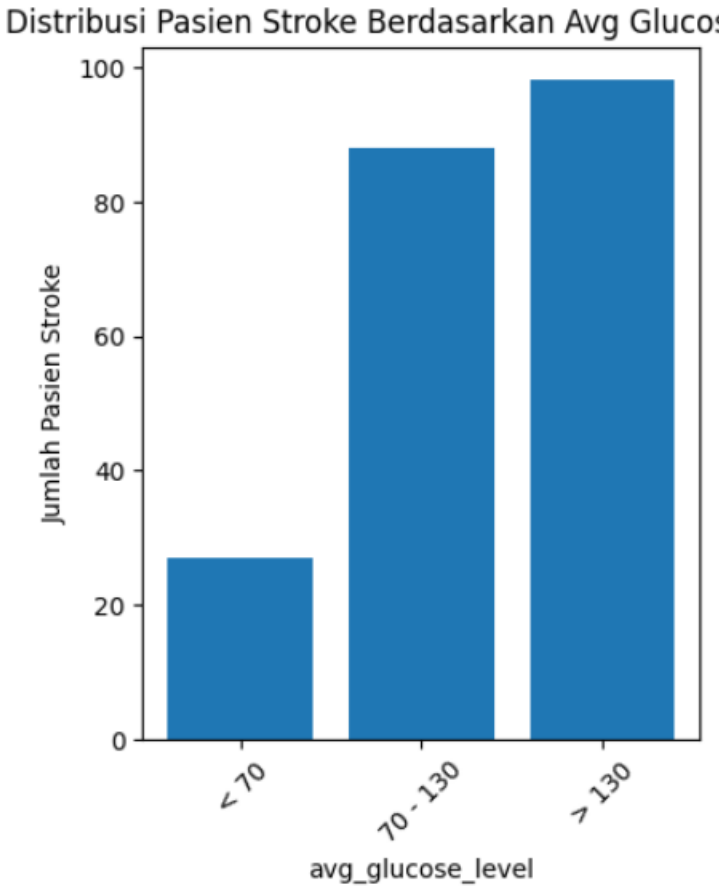
❑ Distribusi Pasien Stroke Berdasarkan Usia




```
Usia < 41      : 8 pasien
Usia 41 s/d 50 : 15 pasien
Usia 51 s/d 60 : 49 pasien
Usia 61 s/d 70 : 49 pasien
Usia 71 s/d 80 : 105 pasien
Usia > 80      : 23 pasien
```

- Dari data terlihat bahwa penderita stroke sudah mulai terlihat banyak pada pasien diusia 51 s/d 70 tahun dan bertambah banyak di usia 71 s/d 80 tahun

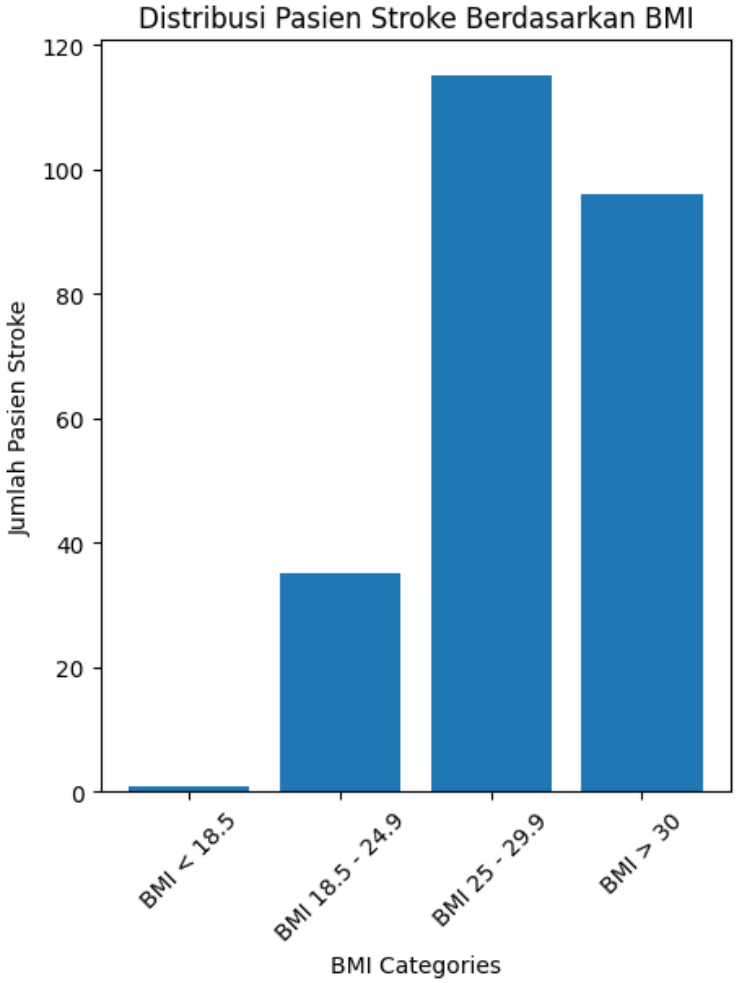
❑ Distribusi Pasien Stroke Berdasarkan Level Gula Darah




```
Avg glucose level < 70 mg/dL                : 27 pasien
Avg glucose level antara 70 dan 130 mg/dL : 88 pasien
Avg glucose level > 130 mg/dL               : 98 pasien
```

- ❑ Stroke banyak terkena pada pasien yang memiliki gula darah normal (70 -130 mg/dL).
- ❑ Stroke bertambah banyak terkena pada pasien dengan level gula darah >130 mg/dL.

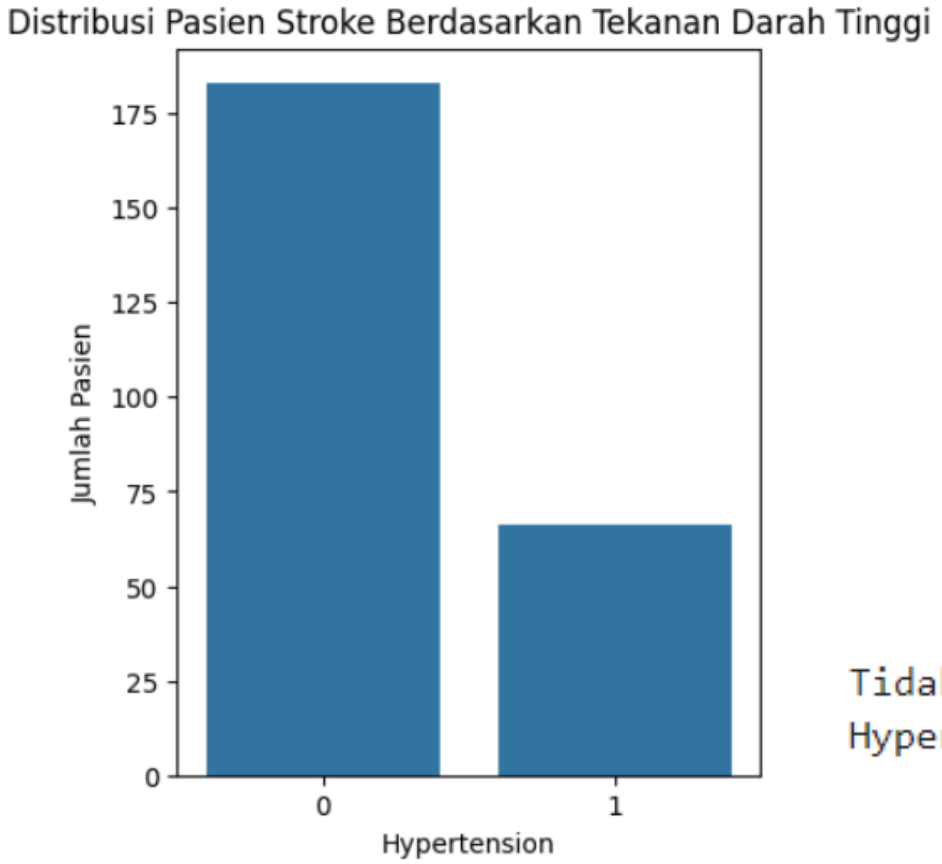
❑ Distribusi Pasien Stroke Berdasarkan Body Mass Index (BMI)




```
Berat badan kurang (BMI < 18.5)    : 1 pasien
Berat badan normal (BMI 18.5-24.9) : 35 pasien
Berat badan lebih (BMI 25-29.9)    : 115 pasien
Pasien stroke obesitas (BMI > 30)  : 96 pasien
```

- Stroke banyak terkena pada pasien yang memiliki berat badan lebih.
- Stroke juga banyak terkena pada pasien yang memiliki berat badan obesitas.

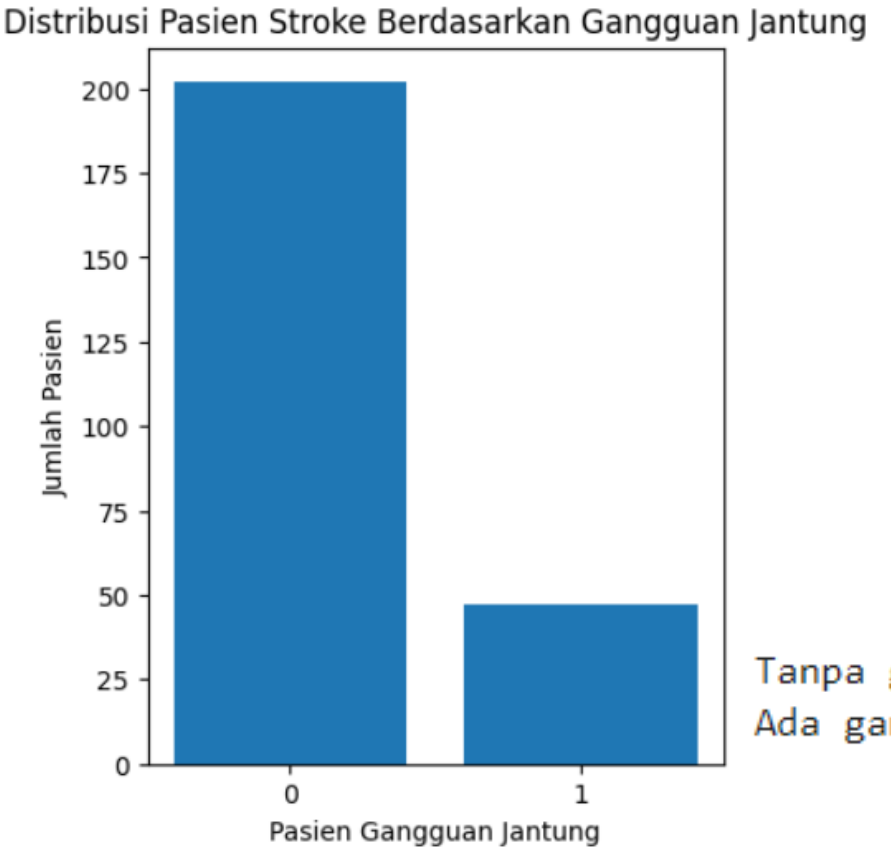
❑ Distribusi Pasien Stroke Berdasarkan Tekanan Darah




```
Tidak hypertensi  : 183 pasien
Hypertensi        : 66 pasien
```

- Dari data terlihat bahwa stroke banyak terkena pada pasien yang tidak memiliki hypertensi.

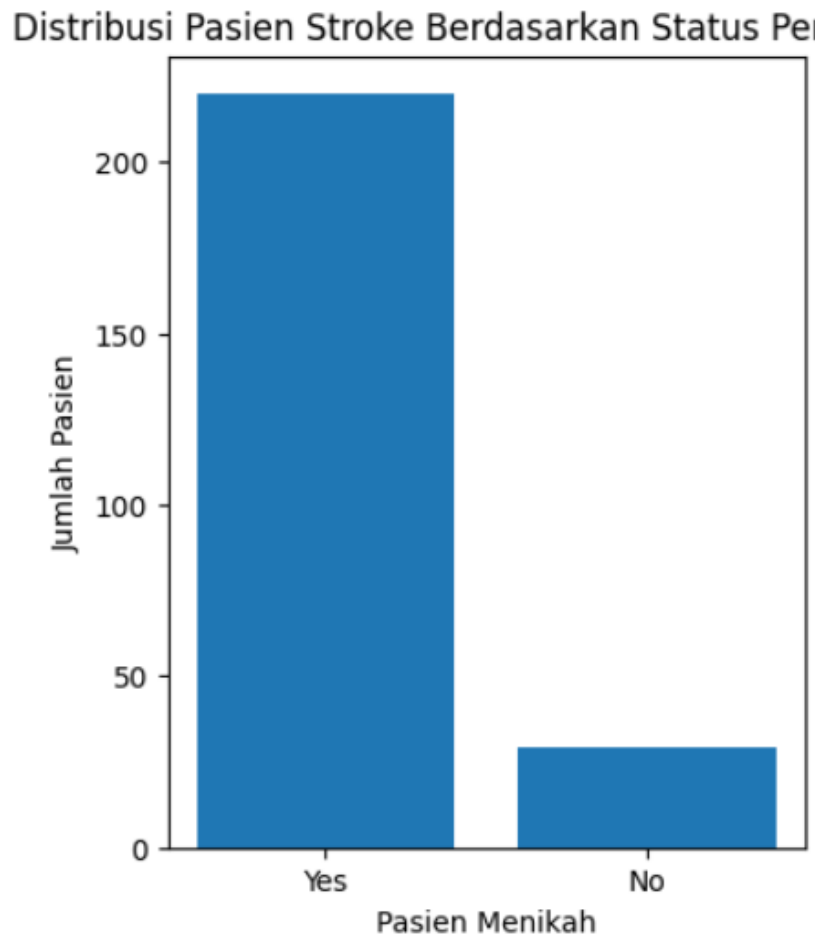
❑ Distribusi Pasien Stroke Berdasarkan Gangguan Jantung




```
Tanpa gangguan jantung  : 202 pasien
Ada gangguan jantung : 47 pasien
```

- Dari data terlihat bahwa stroke banyak terkena pada pasien yang tidak memiliki gangguan jantung.

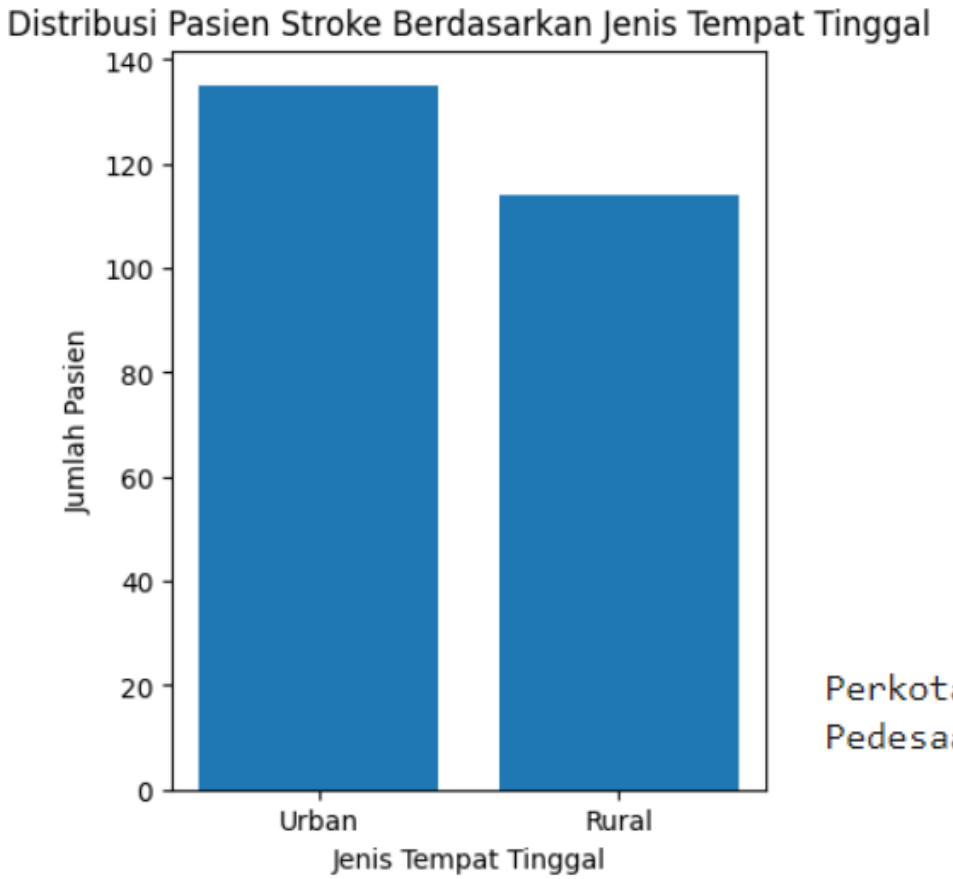
❑ Distribusi Pasien Stroke Berdasarkan Status Pernikahan




```
Menikah       : 220 pasien
Tidak menikah: 29 pasien
```

- Dari data terlihat bahwa stroke banyak terkena pada pasien yang sudah menikah.

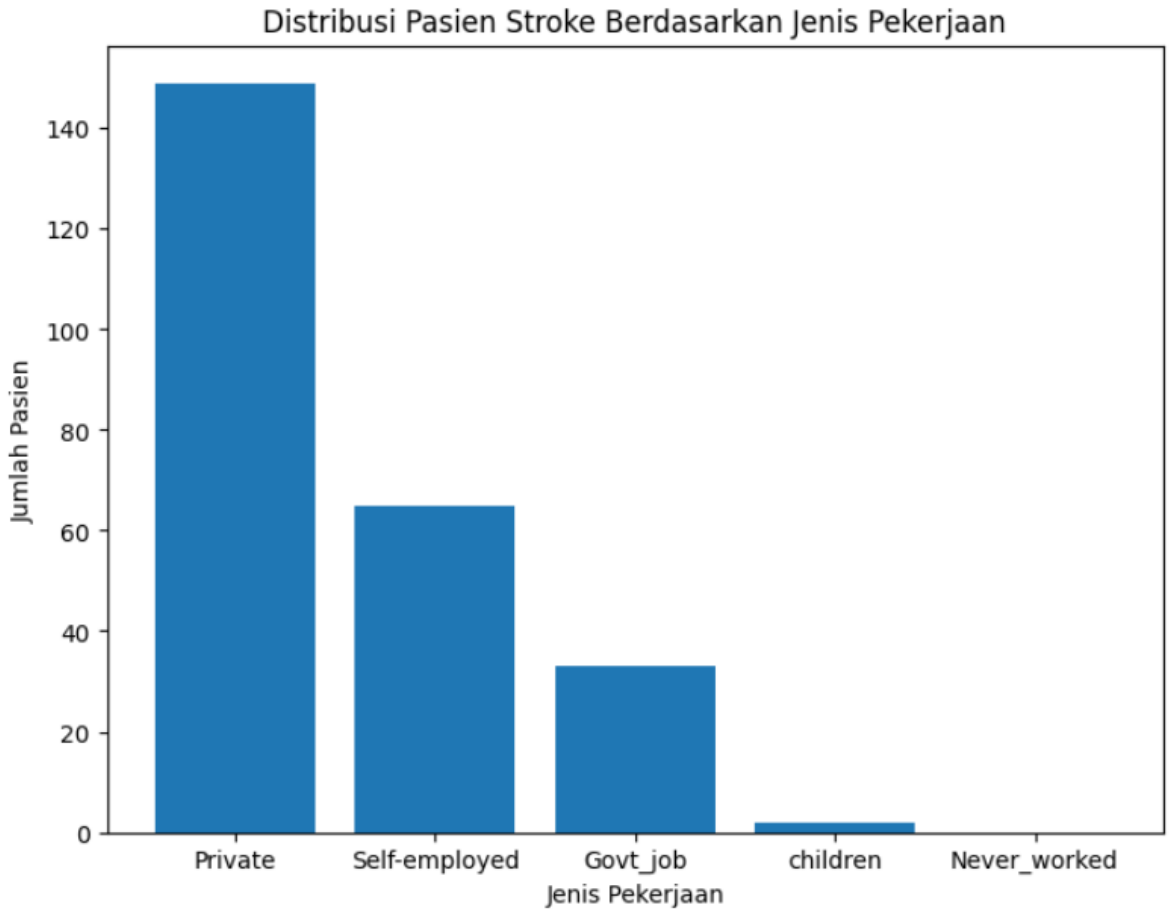
❑ Distribusi Pasien Stroke Berdasarkan Jenis Tempat Tinggal




```
Perkotaan (Urban): 135 pasien
Pedesaan (Rural) : 114 pasien
```

- Dari data terlihat bahwa stroke banyak terkena pada pasien yang tinggal di perkotaan dibanding di pedesaan.

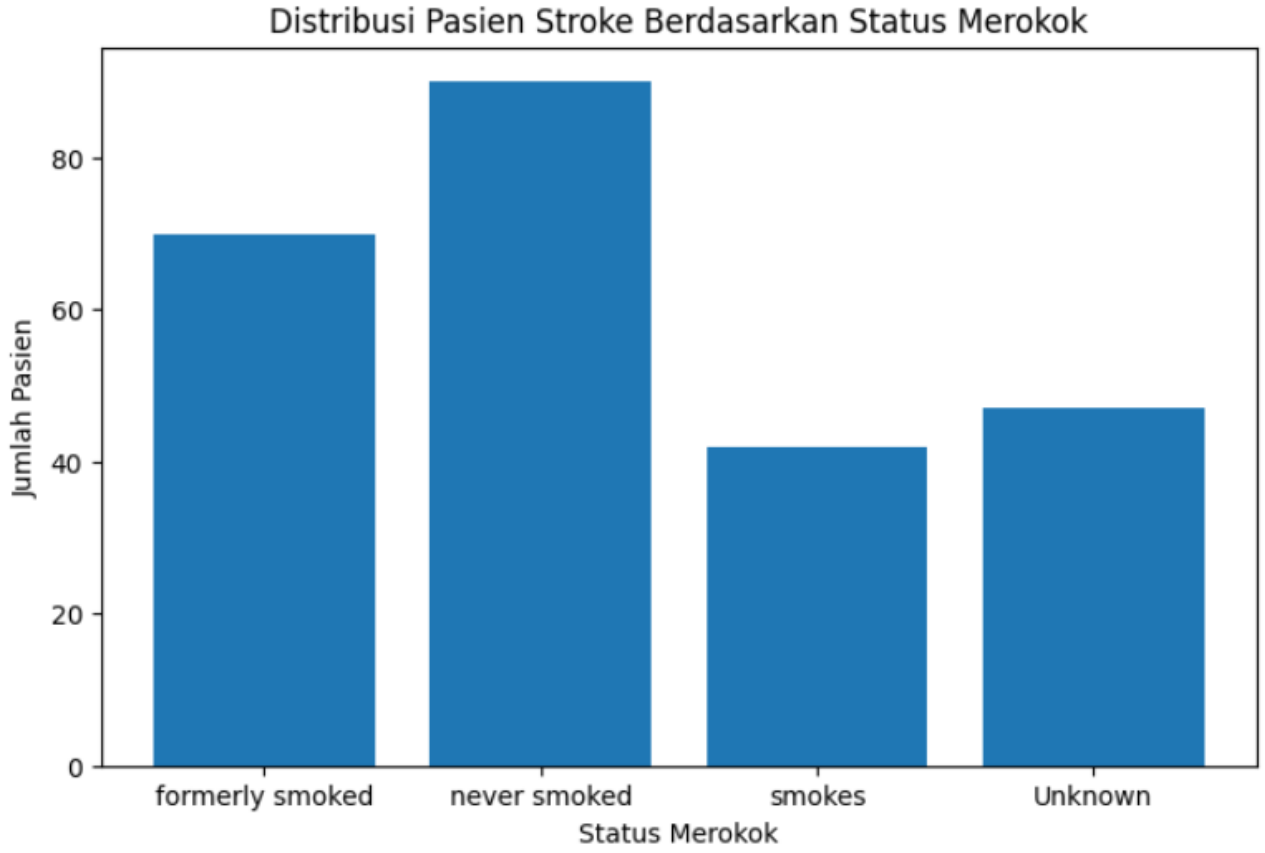
❑ Distribusi Pasien Stroke Berdasarkan Jenis Pekerjaan




```
Pekerja swasta        : 149 pasien
Usaha sendiri         : 65 pasien
Pekerja pemerintah    : 33 pasien
Anak-anak             : 2 pasien
Belum pernah bekerja  : 0 pasien
```

- Dari data terlihat bahwa stroke banyak terkena pada pasien yang bekerja sebagai karyawan swasta.

❑ Distribusi Pasien Stroke Berdasarkan Status Merokok




```
Pernah merokok: 70 pasien
Tidak merokok : 90 pasien
Merokok: 42 pasien
Unknown : 47 pasien
```

- Dari data terlihat bahwa stroke banyak terkena pada pasien yang tidak pernah merokok walaupun untuk pasien yang pernah merokok pun cukup banyak yang terkena stroke.

Pemodelan Data

- K-NN Classification
- Random Forest Classification
- Logistic Regression
- Desicion Tree

Hypertuning parameter,
scoring = recall

Model Dasar

Model	Class	Accuracy	Precision	Recall	f1-Score	ROC-AUC
K-NN Classification	1	0.93	0.18	0.03	0.05	0.65
Random Forest	1	0.94	0.00	0.00	0.00	0.81
Logistic Regression	1	0.94	0.00	0.00	0.00	0.84
Decision Tree	1	0.92	0.24	0.16	0.19	0.58

Dari ke 4 pemodelan yang sudah dilakukan, model tidak dapat membaca data dengan baik.

Hal ini bisa terjadi karena dataset yang digunakan sangat imbalance (extreme imbalance) dimana kelas "0" merupakan data mayoritas dan kelas "1" merupakan data minoritas. Sehingga model akan sangat baik membaca data pada kelas “0” sedangkan kita ingin agar model sensitive terhadap kelas “1” untuk penderita stroke.

Model Undersampling

Model	Class	Accuracy	Precision	Recall	f1-Score	ROC-AUC
K-NN Classification	1	0.72	0.69	0.78	0.73	0.77
Random Forest	1	0.79	0.75	0.86	0.80	0.86
Logistic Regression	1	0.74	0.72	0.78	0.75	0.86
Decision Tree	1	0.71	0.72	0.67	0.69	0.75

Model Oversampling

Model	Class	Accuracy	Precision	Recall	f1-Score	ROC-AUC
K-NN Classification	1	0.91	0.85	0.98	0.91	0.94
Random Forest	1	0.82	0.76	0.94	0.84	0.92
Logistic Regression	1	0.84	0.82	0.86	0.84	0.92
Decision Tree	1	0.90	0.88	0.94	0.91	0.94

Model Terbaik

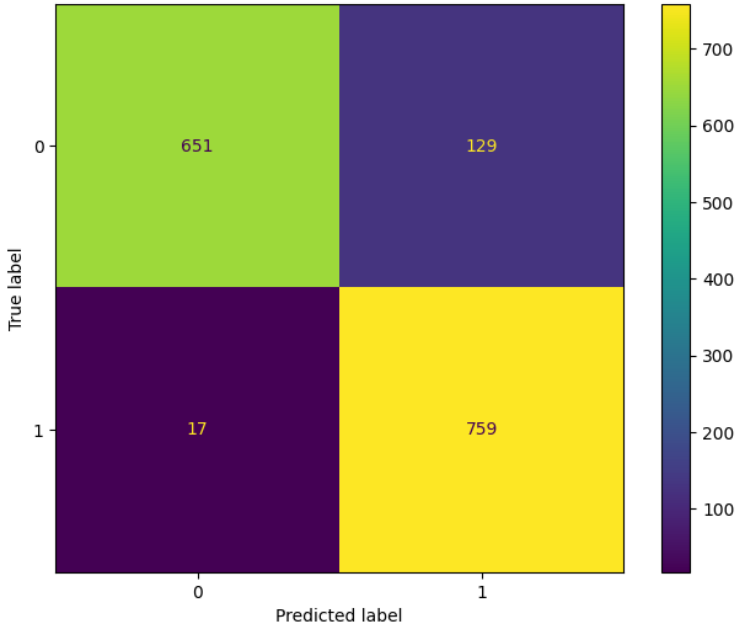
Dari 12 kali percobaan yang sudah dilakukan terlihat bahwa model oversampling memiliki nilai yang paling tinggi diantara yang lainnya.

Untuk itu model yang terbaiknya adalah K-NN Classification.

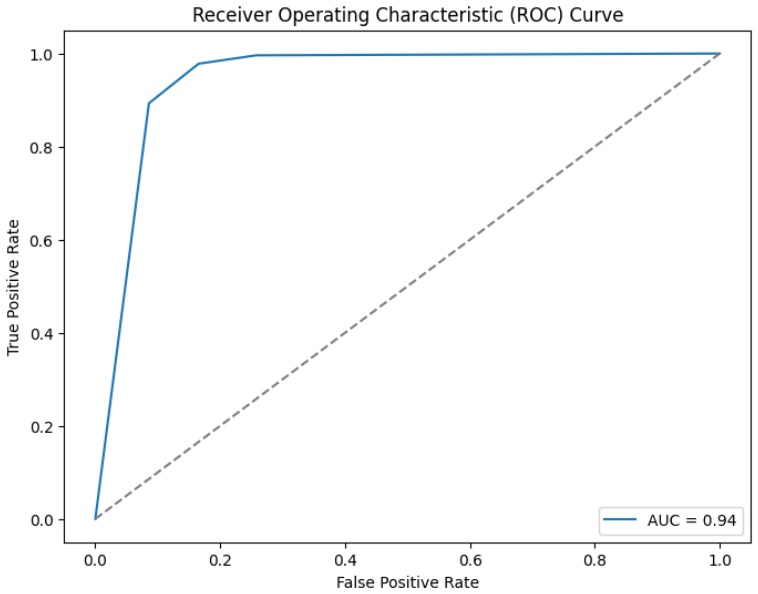
- Memiliki nilai accuracy = 91%
- Memiliki nilai precision 85% yang artinya dari 100 pasien yang diprediksi stroke, 85 pasien benar stroke.
- Memiliki nilai recall 98% yang artinya dari 100 pasien yang diprediksi stroke, hanya 2 pasien yang terprediksi tidak stroke tetapi sebenarnya stroke.

K-NN Classification

Confusion Matrix



ROC-AUC



Dari hasil confusion matrix dapat dijelaskan bahwa :

651 ▶ ada 651 pasien yang terprediksi tidak stroke dan benar tidak stroke.

129 ▶ ada 129 pesien yang terprediksi stroke tetapi sebenarnya tidak stroke.

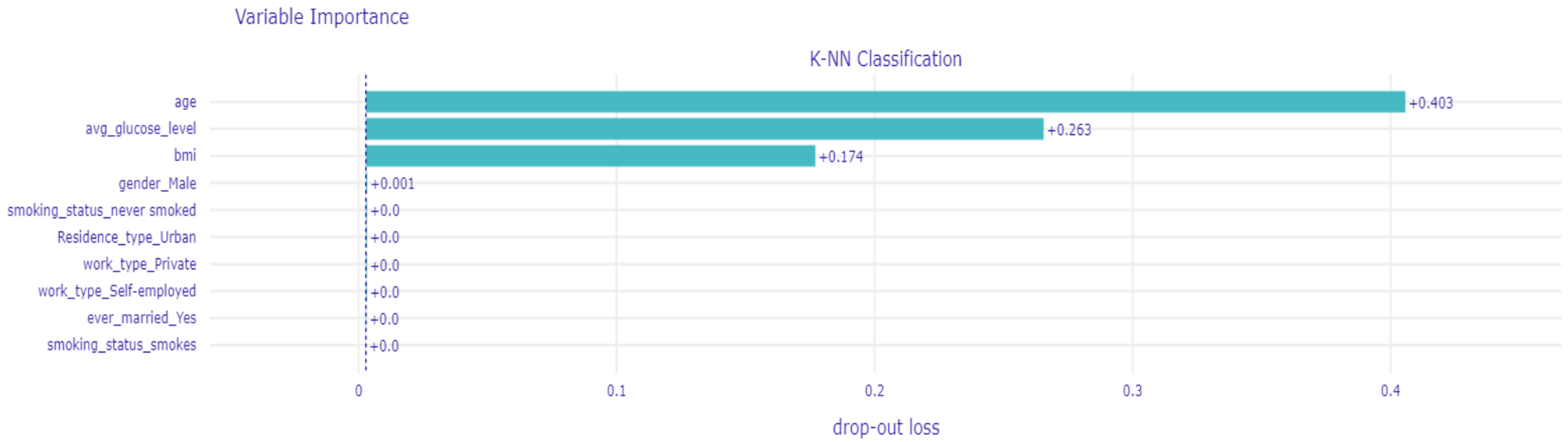
17 ▶ ada 17 pasien yang terprediksi tidak stroke tetapi sebenarnya stroke.

759 ▶ ada 759 pasien yang terprediksi stroke dan benar stroke .

Pada gambar, AUC adalah 0,94, yang menunjukkan kinerja model yang sangat baik. Kurva ROC berada di atas garis putus-putus. Ini berarti model dapat membedakan antara kelas positif dan negatif dengan baik.

K-NN Classification

Feature Omportance



Dari hasi visualisasi feature importance diatas, terlihat bahwa "usia" merupakan faktor utama yang berpengaruh terhadap kejadian stroke. Selain “usia”, faktor lain yang harus diperhatikan adalah “level gula darah” dan “berat badan” yang juga dapat berpengaruh terhadap kejadian stroke.

Insights

- Dari data menunjukkan bahwa resiko stroke pada wanita lebih tinggi dari pria.
- Data menunjukkan bahwa jumlah pasien yang terkena stroke cenderung meningkat seiring bertambahnya usia, dan pentingnya pencegahan pada usia lanjut.
- Jumlah pasien yang terkena stroke cenderung lebih tinggi pada kelompok dengan rata-rata glukosa darah yang lebih tinggi.
- Jumlah pasien stroke yang tergolong memiliki BMI diatas normal (lebih tinggi) merupakan faktor resiko stroke. Resiko stroke menurun dengan berat badan normal.

Rekomendasi

- Lakukan pemeriksaan kesehatan secara rutin dan konsultasi dengan dokter mengenai faktor risiko dan langkah-langkah pencegahan stroke yang sesuai.
- Untuk yang sudah menginjak usia 50 tahun sebaiknya melakukan pola hidup sehat seperti diet seimbang, rutin berolahraga dan menghindari stress yang berlebihan .
- Memantau kadar gula darah secara teratur, terutama bagi mereka yang memiliki riwayat diabetes, mengikuti diet sehat yang rendah gula dan karbohidrat, serta membatasi asumsi makanan olahan dan makan tinggi gula.
- Melakukan penurunan berat badan yang sehat tentunya dengan arahan dokter atau ahli gizi.

Kesimpulan

Menggunakan model K-NN Classification memiliki performa yang paling bagus dibanding dengan model lainnya, memiliki akurasi 91%, recall 98% yang artinya model sangat sensitif mengindektifikasi pasien penderita stroke.

Terimakasih