21/04/2024

Heart Stroke Prediction

Presented by Alfiyyah Ajeng Nurardita - SI602003

Studi Independen IBM SkillsBuild For AI & Cybersecurity by Hacktiv8

Daftar Isi

Pendahuluan

Latar Belakang & Problem Statement

Stroke adalah penyebab kematian kedua global (WHO), bertanggung jawab atas 11% kematian dunia. Ditandai dengan defisit neurologis yang berlangsung ≥24 jam akibat sumbatan atau pecahnya pembuluh darah otak.

Dataset ini digunakan untuk memprediksi risiko stroke berdasarkan parameter seperti gender, usia, status merokok, dan tempat tinggal.

Objektif

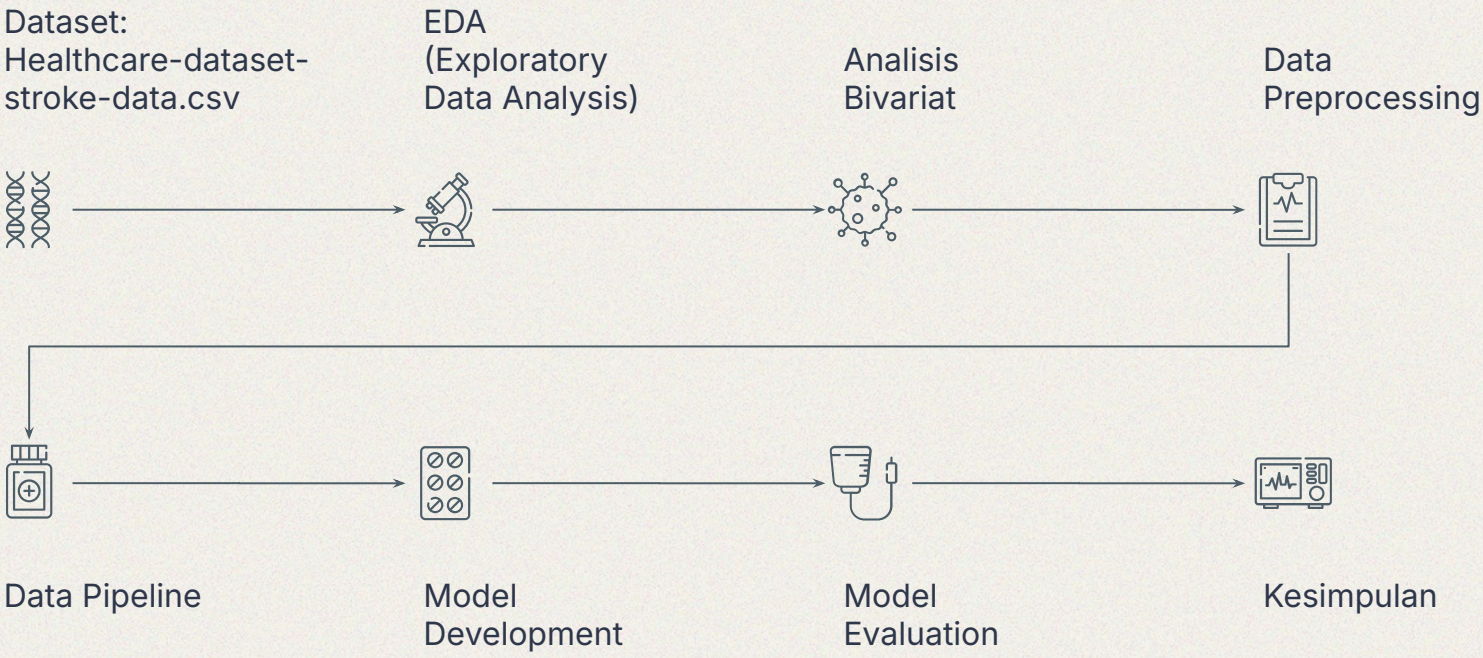
Memprediksi kemungkinan seseorang terkena stroke berdasarkan parameter:

- gender
- umur
- penyakit hipertensi
- penyakit jantung
- status menikah
- jenis pekerjaan
- jenis tempat tinggal
- kadar gula darah
- bmi
- status merokok

Pertanyaan Utama

1. Apakah hipertensi dapat mempengaruhi kemungkinan stroke pada seseorang?
2. Apakah umur dapat mempengaruhi stroke pada seseorang?
3. Apakah menikah dapat mempengaruhi kemungkinan stroke?
4. Apakah merokok dapat mempengaruhi potensi kejadian stroke pada seseorang?

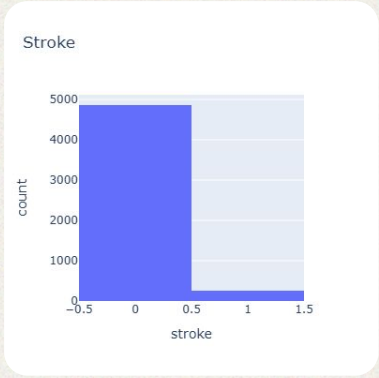
Metodologi


Dataset: Healthcare-dataset-stroke-data.csv
EDA (Exploratory Data Analysis)
Analisis Bivariat
Data Preprocessing
Data Pipeline
Model Development
Model Evaluation
Kesimpulan

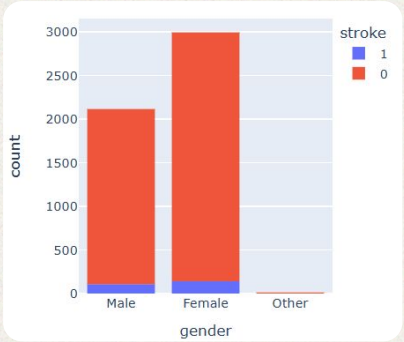
Informasi Fitur

Fitur	Deskripsi
id	unique identifier
gender	'Male','Female', or 'Other'
age	age of the patient
hypertension	0 if the patient doesn't have hypertension, 1 if the patient has hypertension
heart_disease	0 if the patient doesn't have any heart diseases, 1 if the patient has a heart disease
ever_married	"No" or "Yes"
work_type	"children", "Govt_jov", "Never_worked", "Private" or "Self-employed"
Residence_type	"Rural" or "Urban"
avg_glucose_level	average glucose level in blood 10. bmi: body mass index
bmi	body mass index
smoking_status	"formerly smoked", "never smoked", "smokes" or "Unknown"*
stroke	1 if the patient had a stroke or 0 if not

Exploratory Data Analysis (EDA)



95% dari instance adalah **bukan penderita stroke**, sebanyak 4861 pasien.
5% dari instance adalah **penderita stroke**, sebanyak 249 pasien.
Terjadi **Imbalanced Data.**
Penanganan yang dilakukan yaitu menggunakan **SMOTENC**



Terdapat gender **'Other'** yang kita asumsikan bahwa orang tersebut tidak ingin memberitahukan identitas gendernya karena permasalahan tertentu, dan orang tersebut tidak memiliki penyakit stroke. Sehingga bisa dianggap sebagai **MCAR (Missing Completely at Random)** dan bisa kita **drop**

Terdapat **missing values** pada fitur **BMI** sebesar **3,93%** dari keseluruhan baris atau sebanyak **201 baris**.

```
1 df.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
```

Numerical Features

```
df[numerical_features].describe()
```

	age	avg_glucose_level	bmi
count	5110.000000	5110.000000	4909.000000
mean	43.226614	106.147677	28.893237
std	22.612647	45.283560	7.854067
min	0.080000	55.120000	10.300000
25%	25.000000	77.245000	23.500000
50%	45.000000	91.885000	28.100000
75%	61.000000	114.090000	33.100000
max	82.000000	271.740000	97.600000

- Terdapat **usia minimum** yaitu **0.08**, baris ini akan **dianalisis kembali** untuk melihat kewajarannya
- **rata-rata umur** dari keseluruhan populasi yaitu **43 tahun**
- sebanyak **50%** populasi memiliki umur di atas **45 tahun**

Analisis Bivariat

Apakah hipertensi dapat mempengaruhi kemungkinan stroke pada seseorang ?

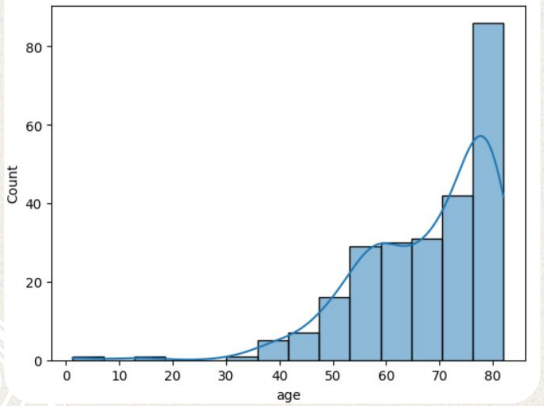
```
prob_stroke_hypertension = round(df[df['hypertension'] ==1]['stroke'].mean()*100,2)
print(prob_stroke_hypertension)
prob_stroke_no_hypertension = round(df[df['hypertension'] ==0]['stroke'].mean()*100,2)
print(prob_stroke_no_hypertension)
```

```
13.25
3.97
```

Seseorang dengan **hipertensi** memiliki kemungkinan **13.25% terkena stroke**, sedangkan seseorang yang **tidak hipertensi** mempunyai kemungkinan **3.97% terkena stroke**. Berdasarkan presentase ini maka **seseorang dengan hipertensi berpotensi terkena stroke lebih besar**.

Apakah umur dapat mempengaruhi stroke pada seseorang ?

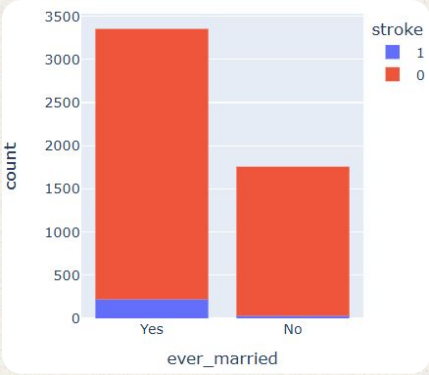
```
sns.histplot(df.query('stroke==1'), x='age',kde=True);
```



Berdasarkan histogram di samping, kejadian stroke **meningkat pada usia 40** dan terus meningkat seiring bertambahnya umur. Sehingga dapat disimpulkan **semakin tua usia seseorang maka berpotensi terkena stroke lebih besar**.

Dalam beberapa artikel, stroke umumnya terjadi pada seseorang yang sudah memasuki usia lanjut yaitu di atas 55 tahun. Namun, risiko stroke juga dapat terjadi pada orang dengan usia muda

Apakah menikah dapat mempengaruhi kemungkinan stroke ?



Seseorang yang sudah menikah memiliki kemungkinan stroke lebih besar yaitu 6.56% dibandingkan orang yang **belum menikah** yaitu **1.65%**.

Berdasarkan penelitian tahun 2016 yang diterbitkan oleh Journal of American Heart Association stabilitas perkawinan berdampak pada risiko stroke suatu pasangan bahkan pada anak-anak mereka yang akan dewasa di kemudian hari.

Depresi yang membuat perubahan kepribadian penderita stroke seperti hilangnya empati, selera humor dan perasaan cemburu mempengaruhi kualitas dan kepuasan pernikahan pasangannya.

Apakah merokok dapat mempengaruhi potensi kejadian stroke pada seseorang ?

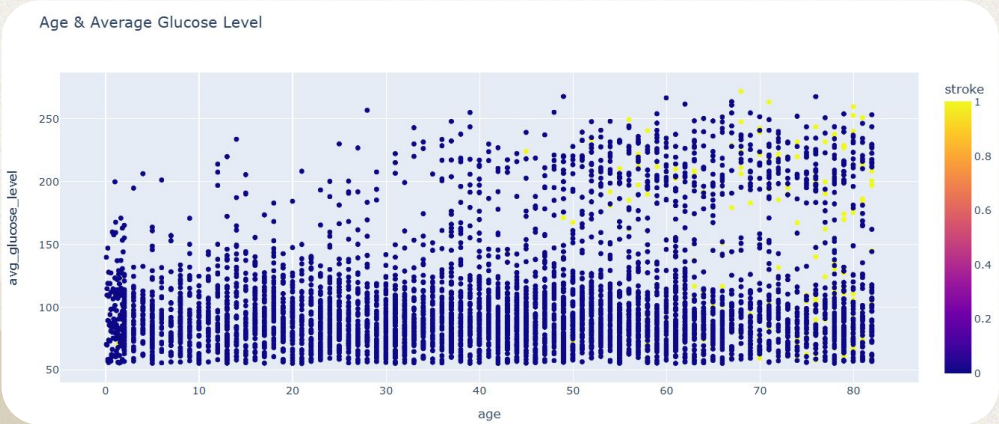
Persentase paling tinggi berisiko terkena stroke adalah penderita yang **dahulunya merokok** yaitu sebesar **7.91%** disusul dengan penderita yang sedang merokok sebesar 5.32%. Maka dapat disimpulkan **seseorang yang pernah merokok berisiko tinggi terkena stroke.**

Merokok dapat menyebabkan masalah dengan aliran darah ke otak dan berkontribusi terhadap pembentukan gumpalan di pembuluh darah. Kedua hal tersebut meningkatkan risiko terjadinya stroke. Berdasarkan *Journal of the American Heart Association* jika seseorang memutuskan berhenti merokok dapat mengurangi risiko stroke berulang hingga 29% dibandingkan dengan orang yang tetap merokok.

Correlation Matrix

	age	avg_glucose_level	bmi	stroke
age	1.000000	0.238323	0.333314	0.245239
avg_glucose_level	0.238323	1.000000	0.175672	0.131991
bmi	0.333314	0.175672	1.000000	0.042341
stroke	0.245239	0.131991	0.042341	1.000000

Korelasi paling tinggi terhadap stroke adalah age atau **usia**, yaitu sebesar 0.24 disusul avg_glucose_level atau **kadar glukosa** dalam darah sebesar 0.13 dan **bmi** sebesar 0.04.



Visualisasi disamping ini menunjukkan hubungan usia dan stroke, Titik kuning atau indikasi stroke lebih banyak muncul pada kelompok **usia 50 - 80 tahun**. Indikasi stroke juga sering muncul pada **kadar glukosa di atas 150.**

Data Pipeline

```
#Pipeline Fitur Numerik
numeric_transformers = Pipeline([
    ('imputer', SimpleImputer(strategy='median')),
    ('scaler', MinMaxScaler())
])
```

```
#Pipeline Fitur Kategorikal
cat_enc_transfromers = Pipeline([
    ('imputer', SimpleImputer(strategy='most_frequent'))
])
```

```
#Pipeline Fitur Kategorikal belum Encode
cat_features = ['gender','ever_married']
cat_transformers = Pipeline([
    ('imputer', SimpleImputer(strategy='most_frequent')),
    ('bin_encoder', OrdinalEncoder())
])
```

```
cat_ohe_features = ['work_type','Residence_type','smoking_status']
cat_ohe_tranformers = Pipeline([
    ('imputer', SimpleImputer(strategy='most_frequent')),
    ('onehot_encoder', OneHotEncoder())
])
```

```
#Menggabungkan keempat pipeline
preprocessor = ColumnTransformer([
    ('numeric', numeric_transformers, numerical_features),
    ('categorical_enc',cat_enc_transfromers, cat_enc_feature),
    ('categorical_bin',cat_transformers, cat_features),
    ('categorical_ohe',cat_ohe_tranformers, cat_ohe_features)
], remainder='passthrough', verbose=True)
```

- age → **SimpleImputer(median), MinMaxScaler**
- avg_glucose_level → **SimpleImputer(median), MinMaxScaler**
- bmi → **SimpleImputer(median), MinMaxScaler**
- gender → **SimpleImputer(modus), OrdinalEncoder**
- ever_married →**SimpleImputer(modus), OrdinalEncoder**
- work_type → **SimpleImputer(modus), OneHotEncoder**
- Residence_type → **SimpleImputer(modus), OneHotEncoder**
- smoking_status → **SimpleImputer(modus), OneHotEncoder**
- hypertension → **SimpleImputer(modus)**
- heart_disease → **SimpleImputer(modus)**

Model Development

Splitting Data

- 80/20

```
X_train,X_test,y_train,y_test = train_test_split(X,y,test_size=0.2,random_state=23,stratify=y)
```

Algoritma

Decision Tree

```
dt_randomcv = RandomizedSearchCV(
    estimator=DecisionTreeClassifier(random_state=42),
    param_distributions=random_search_params,
    n_iter=20,
    cv=5,
    n_jobs=-1,
    scoring='recall',
    random_state=42
)

dt_randomcv.fit(X_train_resampled, y_train_resampled)
```

Random Search

```
rf_randomcv = RandomizedSearchCV(
    estimator=RandomForestClassifier(random_state=42),
    param_distributions=random_search_params,
    n_iter=20,
    cv=5,
    n_jobs=-1,
    scoring='recall',
    random_state=42
)

rf_randomcv.fit(X_train_resampled, y_train_resampled)
```

Hyperparameter Tuning

- **Random Search**

```
random_search_params = {
    'max_depth': range(3, 13),
    'min_samples_split': [2,3,4],
    'min_samples_leaf': [2,3,4]
}
```

Evaluasi

Decision Tree dengan Pipeline

```
print(classification_report(y_train_resampled,y_train_pred))
```

	precision	recall	f1-score	support
0	0.96	0.87	0.91	3888
1	0.88	0.97	0.92	3888
accuracy			0.92	7776
macro avg	0.92	0.92	0.92	7776
weighted avg	0.92	0.92	0.92	7776

Decision Tree tanpa Pipeline

```
print(classification_report(y_test,y_test_pred))
```

	precision	recall	f1-score	support
0	0.97	0.83	0.89	972
1	0.13	0.48	0.20	50
accuracy			0.81	1022
macro avg	0.55	0.65	0.55	1022
weighted avg	0.93	0.81	0.86	1022

Random Forest dengan Pipeline

```
print(classification_report(y_train_resampled,y_train_pred))
```

	precision	recall	f1-score	support
0	0.92	0.78	0.84	3888
1	0.81	0.93	0.86	3888
accuracy			0.85	7776
macro avg	0.86	0.85	0.85	7776
weighted avg	0.86	0.85	0.85	7776

Random Forest tanpa Pipeline

```
print(classification_report(y_test,y_test_pred))
```

	precision	recall	f1-score	support
0	0.98	0.79	0.88	972
1	0.16	0.76	0.26	50
accuracy			0.79	1022
macro avg	0.57	0.77	0.57	1022
weighted avg	0.94	0.79	0.85	1022

Kesimpulan

Berdasarkan *Confusion Matrix*, performa terbaik ditunjukkan oleh model **Decision Tree dengan Pipeline**, dengan precision, recall, F1 dan Accuracy **92%**. Pada nilai recall model terlihat cukup seimbang dengan perbedaan kelas **recall 10%** dan **precision 8%**.

Implementasi **pemrosesan fitur** yang tepat dan teknik **SMOTENC** untuk menangani data tidak seimbang terbukti **meningkatkan akurasi model** serta **kemampuannya dalam mendeteksi kasus stroke** secara lebih andal.

Dalam konteks dunia nyata, interpretasi Confusion Matrix dapat berarti:

- 97 dari 100 orang yang sakit mendapat peringatan dini, sehingga peluang sembuh lebih besar berkat penanganan pada stadium awal.
- Tim medis dapat fokus pada kelompok berisiko tinggi, mengoptimalkan waktu dan biaya karena mayoritas (88%) dari yang diperiksa ulang memang membutuhkan penanganan.
- Model unggul dalam menemukan kasus & tetap akurat dan Ideal untuk dunia nyata, menyeimbangkan tujuan menemukan hampir semua kasus (recall tinggi) tanpa membuang sumber daya untuk alarm palsu yang berlebihan.

Terima Kasih

Lets Connect


https://www.linkedin.com/in/ajeng-nurardita/


https://github.com/alfiyyahajeng