DigitalSkola
Uncover The World of Digital Skills with Us

Presented by **Tim Hore** DS30 | 2023

Stroke Prediction

Overview

- Background
- Dataset Characteristic
- Data Preparation
- EDA
- Data Preprocessing
- Feature Engineering
- Model Explanatory
- Modelling
- Overfitting Testing
- Save Model to Pickle
- Deployment
- Conclusion

Presented by **Tim Hore**

DS30 | 2023

Background

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. This research is to predict whether a patient is likely to get stroke based on the input parameters / dataset.

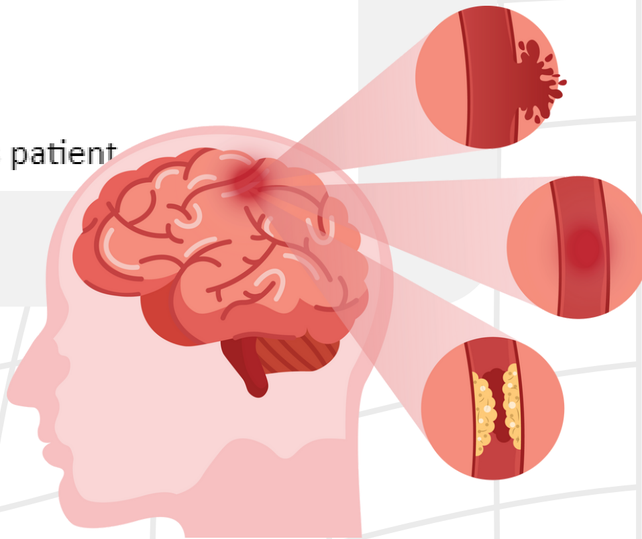
Presented by **Tim Hore**

DS30 | 2023

Dataset Characteristic

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

*Note: "Unknown" in smoking_status means that the information is unavailable for this patient



Data Preparation

1. Load Dataset
2. Check Duplicate Values
3. Check Missing Values

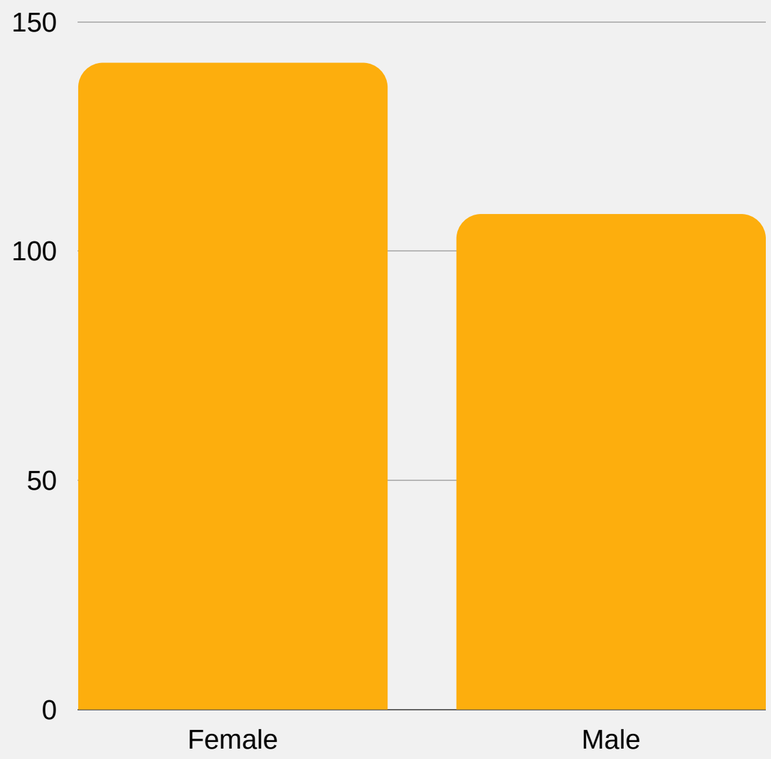
Presented by **Tim Hore**

DS30 | 2023

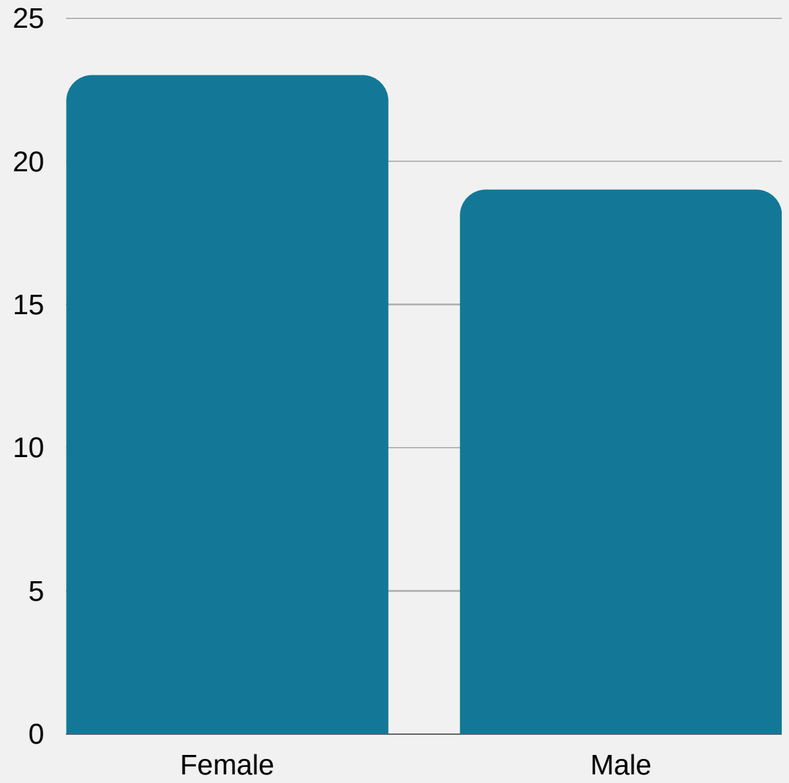
Data Preparation Output

index	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
5	56669	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked	1
6	53882	Male	74.0	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
7	10434	Female	69.0	0	0	No	Private	Urban	94.39	22.8	never smoked	1
8	27419	Female	59.0	0	0	Yes	Private	Rural	76.15	NaN	Unknown	1
9	60491	Female	78.0	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
10	12109	Female	81.0	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
11	12095	Female	61.0	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
12	12175	Female	54.0	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
13	8213	Male	78.0	0	1	Yes	Private	Urban	219.84	NaN	Unknown	1
14	5317	Female	79.0	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1
15	58202	Female	50.0	1	0	Yes	Self-employed	Rural	167.41	30.9	never smoked	1
16	56112	Male	64.0	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
17	34120	Male	75.0	1	0	Yes	Private	Urban	221.29	25.8	smokes	1
18	27458	Female	60.0	0	0	No	Private	Urban	89.22	37.8	never smoked	1
19	25226	Male	57.0	0	1	No	Govt_job	Urban	217.08	NaN	Unknown	1
20	70630	Female	71.0	0	0	Yes	Govt_job	Rural	193.94	22.4	smokes	1
21	13861	Female	52.0	1	0	Yes	Self-employed	Urban	233.29	48.9	never smoked	1
22	68794	Female	79.0	0	0	Yes	Self-employed	Urban	228.7	26.6	never smoked	1
23	64778	Male	82.0	0	1	Yes	Private	Rural	208.3	32.5	Unknown	1
24	4219	Male	71.0	0	0	Yes	Private	Urban	102.87	27.2	formerly smoked	1

EDA

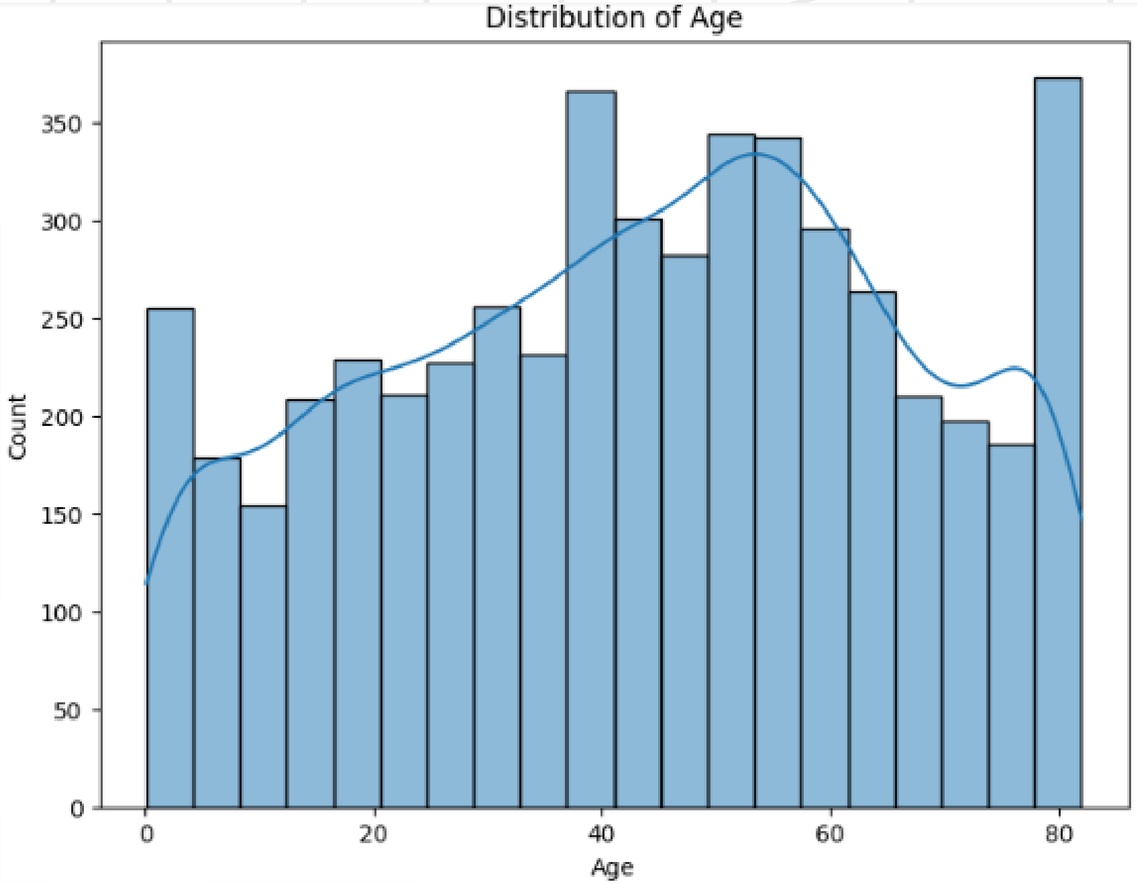


Stroke Count by Gender

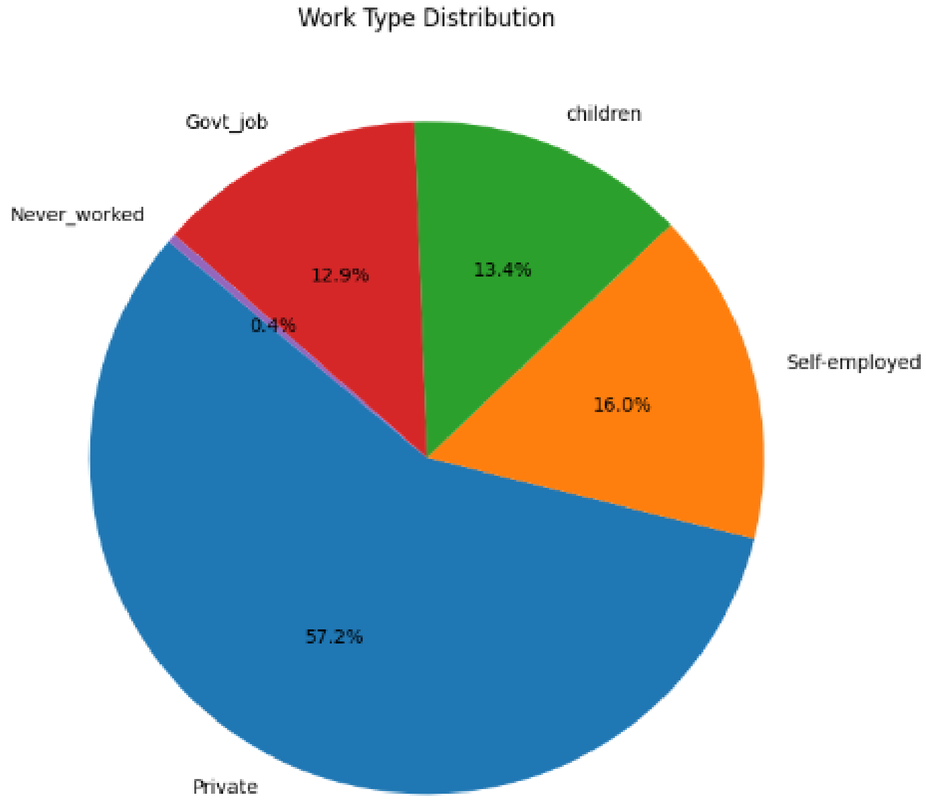


Stroke counts for Smokers by genders

EDA



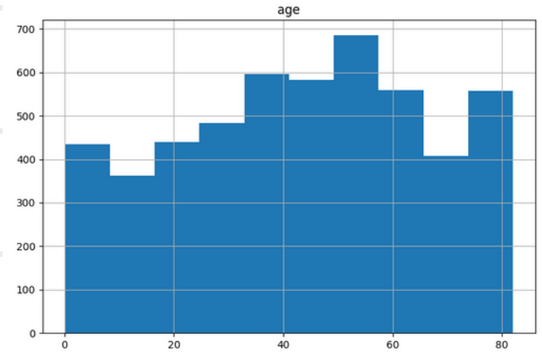
Distribution of Age

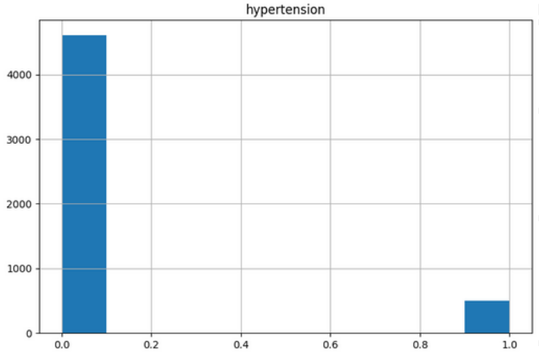


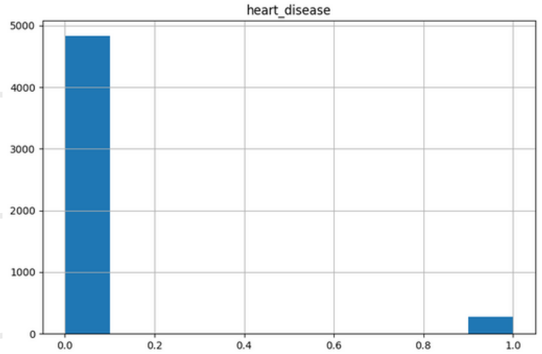
Work Type Distribution

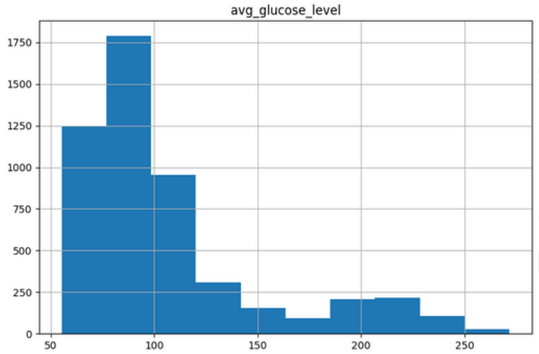
EDA

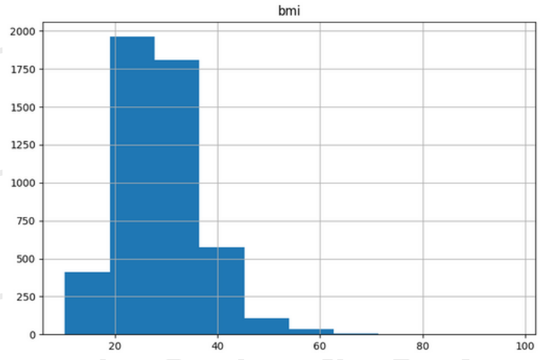
Data Distribution

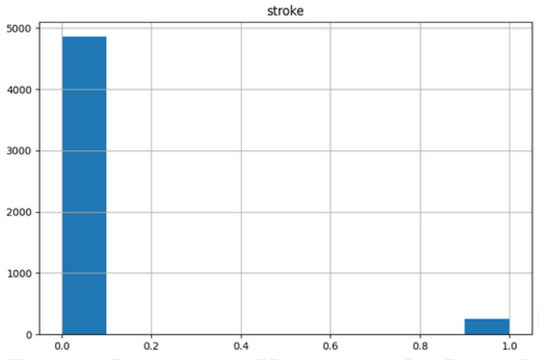


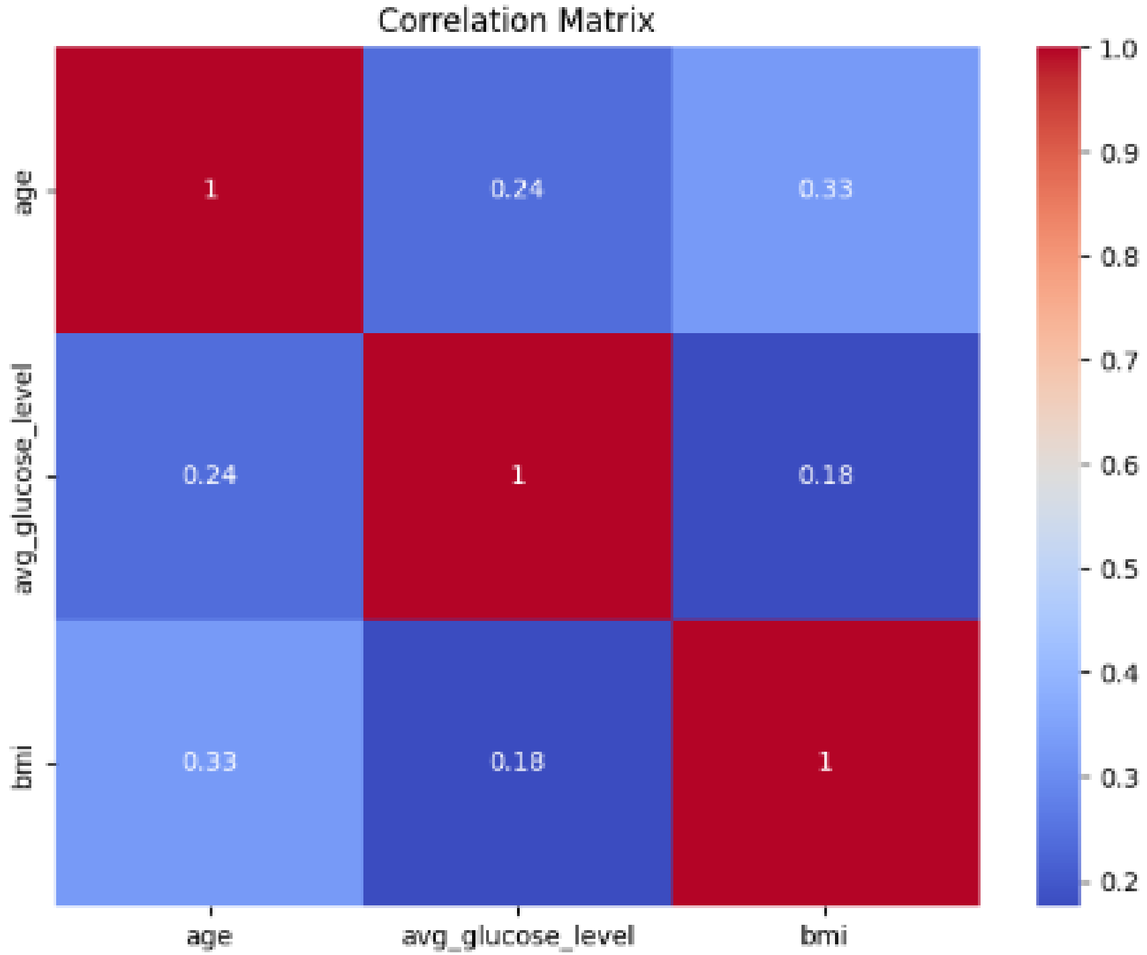












correlation Matrix

Data Preprocessing

01

Fix Missing Value

```
[67] # We can fill in NaN values with a median according to the target
     stroke_0_median = df[df["stroke"] == 0]["bmi"].median()
     stroke_1_median = df[df["stroke"] == 1]["bmi"].median()

     df.loc[(df["stroke"] == 0) & (df["bmi"].isnull()), "bmi"] = stroke_0_median
     df.loc[(df["stroke"] == 1) & (df["bmi"].isnull()), "bmi"] = stroke_1_median
```

```
[68] df.columns.isnull().sum()

     0
```

02

Fix Outlier

Penanggulangan Outlier sukses

```
df.describe(
    percentiles=[0.05, 0.25, 0.50, 0.75, 0.90, 0.95, 0.99]).T
```

1 to 6 of 6 entries Filter

index	count	mean	std	min	5%	25%	50%	75%	90%	95%	99%	max
age	5109.0	43.22998629868859	22.613575307650944	0.08	5.0	25.0	45.0	61.0	75.0	79.0	82.0	82.0
hypertension	5109.0	0.09747504403992954	0.29663257162781625	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0
heart_disease	5109.0	0.0540223135642983	0.22608385143795806	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0
avg_glucose_level	5109.0	106.1403992536114	45.2850036657364	55.12	60.712	77.24	91.88	114.09	192.20199999999997	216.3039999999999	240.7084	271.74
bmi	5109.0	28.858625954198473	7.608339507508262	10.3	17.7	23.8	28.0	32.8	38.7	42.65999999999995	52.892	66.93
stroke	5109.0	0.0487375220199647	0.21533985314518492	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0

Show 25 per page

Feature Engineering

01 Encoding using one-hot-encode

1 to 5 of 5 entries Filter

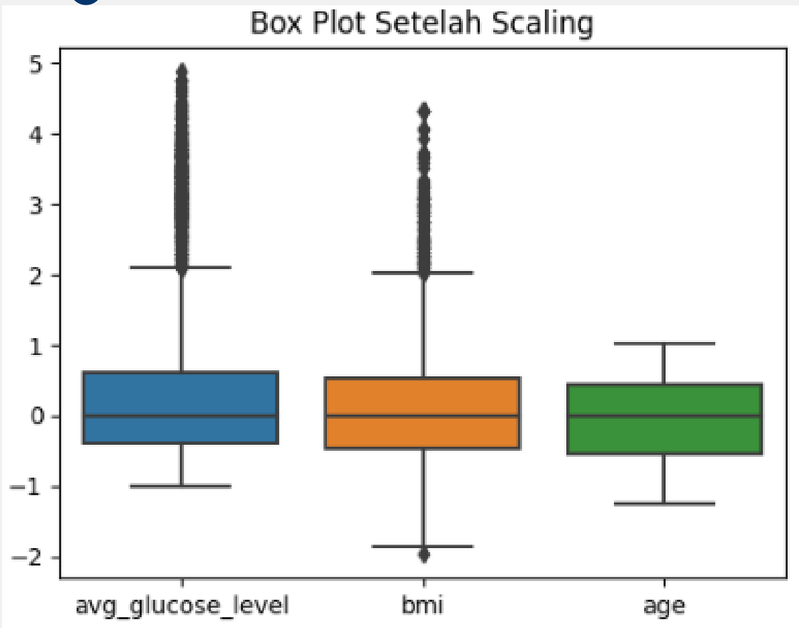
nce_type_Rural	Residence_type_Urban	ever_married_No	ever_married_Yes	gender_Female	gender_Male	smoking_status_Unknown	smoking_status_formerly smoke
0	1	0	1	0	1	0	
1	0	0	1	1	0	0	
1	0	0	1	0	1	0	
0	1	0	1	1	0	0	
1	0	0	1	1	0	0	

Show 25 per page

Like what you see? Visit the data table notebook to learn more about interactive tables.

Warning: Total number of columns (26) exceeds max_columns (20) limiting to first (20) columns.

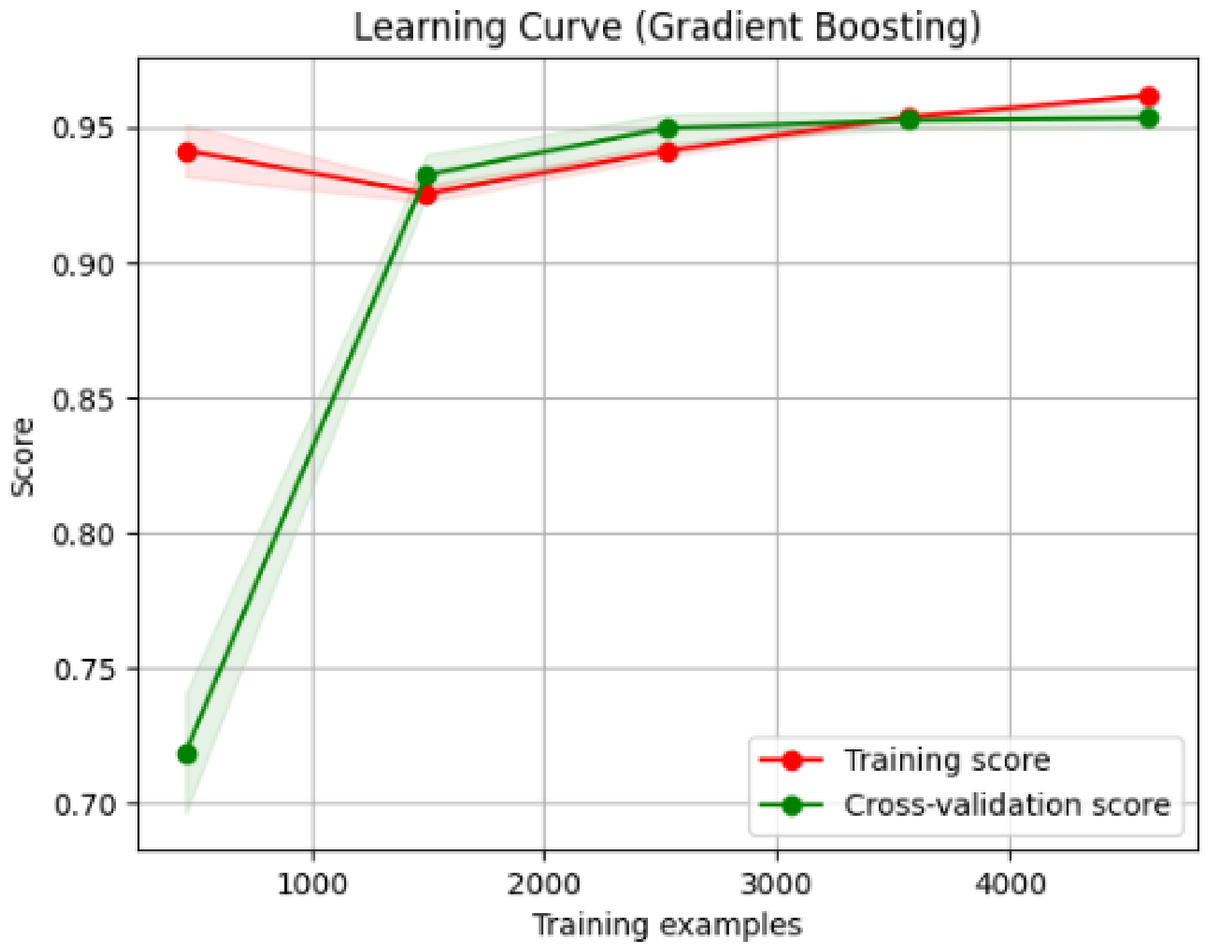
02 Scalling



Model Explanatory

Comparing several models to find which model provides the best accuracy. Based on the model comparison, **we chose Gradient Boosting**, because it provides the best accuracy, A small Std value indicates stable model performance.

	Model_Type	Mean_Accuracy	Std_Accuracy
0	LR	0.951459	0.010886
1	KNN	0.949698	0.012061
2	CART	0.912118	0.009748
3	RF	0.949697	0.011710
4	SVR	0.951263	0.010944
5	XGBM	0.946761	0.007714
6	GB	0.954003	0.009636
7	LightGBM	0.950871	0.010839



Overfitting Test

overfitting of the model is not very significant. Still good to use

Deployment

Code on Streamlit

Membangun visualisi untuk memprediksi stroke menggunakan streamlit

01

Save Model to Pickle

Model yang telah dibangun disimpan kedalam format pickle

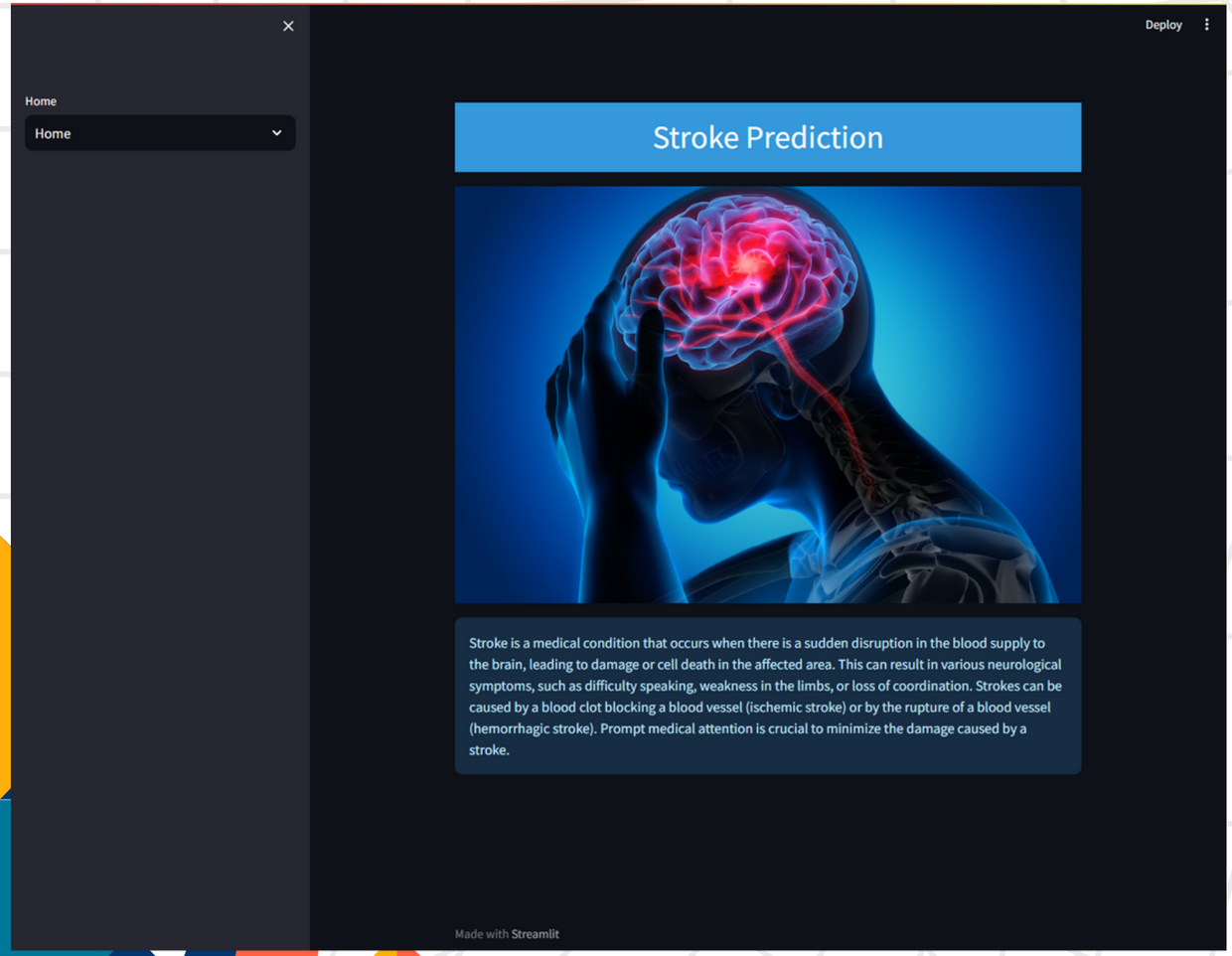
02

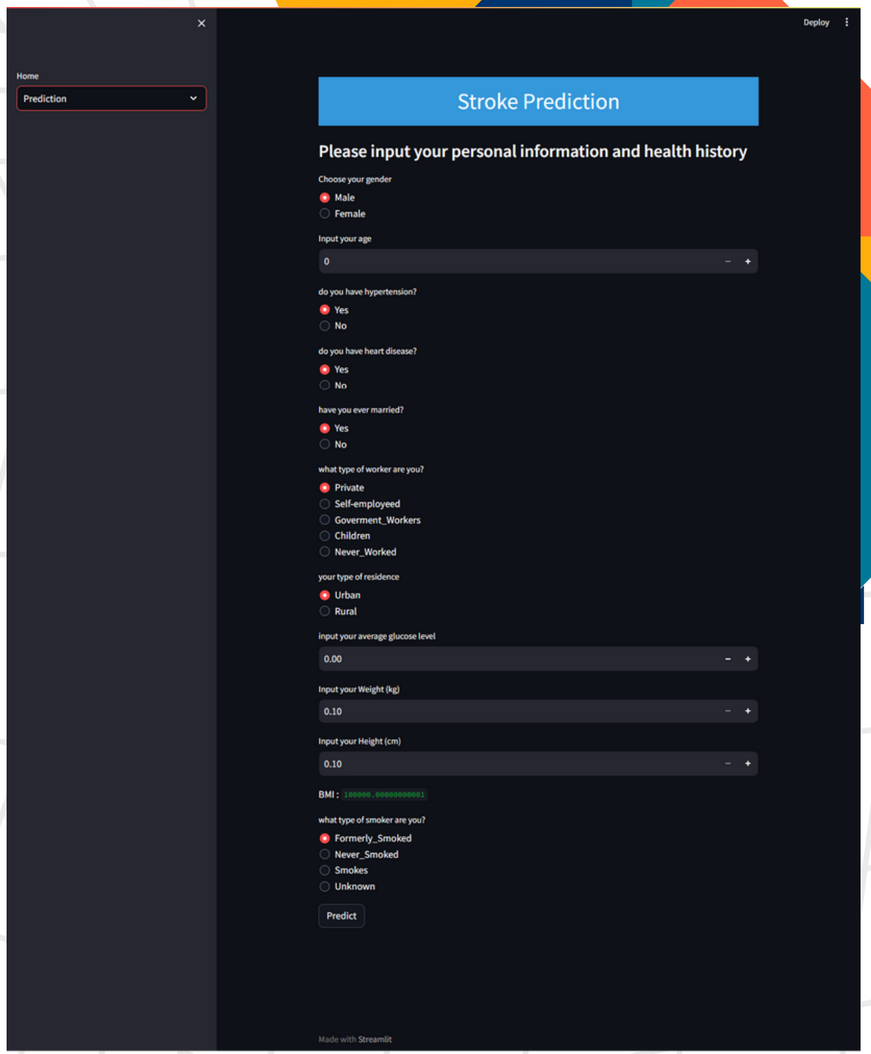
03

Running Streamlit

Apabila sudah selesai deployment, bisa running streamlit untuk memprediksi stroke

Deployment





Presented by **Tim Hore**

DS30 | 2023

Conclusion

- **age distribution** : Rata-rata usia responden adalah sekitar 43 tahun, dengan rentang usia antara 0.08 hingga 82 tahun. Dengan median usia sekitar 45 tahun, distribusi usia cenderung simetris.
- **work type distribution** : Dalam sampel atau populasi yang direpresentasikan oleh pie chart, mayoritas individu (57%) bekerja di sektor swasta (private), yang menjadikannya tipe pekerjaan yang paling umum, dan Fakta bahwa sekitar 13.4% individu merupakan anak-anak menunjukkan bahwa sampel atau populasi juga mencakup anak-anak yang belum masuk ke dalam angkatan kerja. Meskipun hanya sekitar 0.4%, ada individu yang tidak pernah bekerja dalam populasi ini. Ini bisa mencakup individu yang masih belajar atau yang belum memulai karier mereka.
- **the gender that suffers the most strokes** : jumlah kasus stroke lebih tinggi pada individu dengan gender "Female" (Perempuan) dengan jumlah 141(56.63%) kasus dibandingkan dengan "Male" (Laki-laki) yang memiliki 108(43.37%) kasus.
- **stroke counts for smokers by gender** : Data menunjukkan bahwa jumlah kasus stroke pada individu yang merokok adalah Laki-laki (Male) memiliki 23 kasus stroke dan Perempuan (Female) memiliki 19 kasus stroke. Meskipun jumlah kasus stroke pada individu yang merokok adalah lebih tinggi pada laki-laki (23 kasus) dibandingkan perempuan (19 kasus), perbedaan ini tidak signifikan. Data ini menekankan pentingnya kesadaran akan risiko stroke yang dapat terkait dengan merokok. Terlepas dari perbedaan jumlah kasus, merokok dapat meningkatkan risiko stroke pada laki-laki dan perempuan.

Conclusion

Correlation matrix :

- korelasi age terhadap rata-rata glukosa darah : Korelasi antara usia dan rata-rata glukosa darah adalah sekitar 0.24. Ini menunjukkan bahwa ada korelasi positif yang lemah antara usia dan rata-rata glukosa darah. Artinya, dengan bertambahnya usia, rata-rata glukosa darah cenderung sedikit meningkat, meskipun korelasinya lemah.
- korelasi age terhadap bmi : Korelasi antara usia dan BMI adalah sekitar 0.33. Ini menunjukkan bahwa ada korelasi positif yang sedang antara usia dan BMI. Dengan kata lain, dengan bertambahnya usia, BMI cenderung meningkat dalam tingkat korelasi yang lebih kuat daripada korelasi dengan glukosa darah.
- korelasi antara bmi terhadap rata-rata glukosa darah : Korelasi antara rata-rata glukosa darah dan BMI adalah sekitar 0.18. Ini menunjukkan adanya korelasi positif yang lemah antara rata-rata glukosa darah dan BMI. Artinya, individu dengan BMI yang lebih tinggi cenderung memiliki rata-rata glukosa darah yang sedikit lebih tinggi, meskipun korelasinya lemah.

Conclusion

- **Data Preprocessing** : Terdapat data missing value dan outlier di kolom BMI kita melakukan imputasi pada kolom BMI menggunakan Median, dan juga melakukan penekanan nilai outliers pada kolom BMI.
- **Feature Engineering** : Kita melakukan encoded pada data kategorik gender', 'ever_married', 'work_type', 'Residence_type', smoking_status dan scalling pada data numerik avg_glucose_level, bmi, age
- **Modeling** : kita mencoba membuat model Classifier dengan beberapa algoritma dengan metrik accuracy_score. menggunakan validasi silang Kfold Best Model yang kita pilih **GRADIENT BOOSTING**
- **Test Overfit** : Setelah membuat model kita mencoba melakukan test overfit pada model dengan hasil overfitt tidak signifikan dan masih aman digunakan. Training Accuracy: 0.9650110105211647 dan Validation Accuracy: 0.9412915851272016

Thank You!

Presented by **Tim Hore**

DS30 | 2023