Machine Learning Team Project

Cerebral Stroke Prediction

Team11

2017170827 이병주
2018160339 차수지
2019250007 조여정

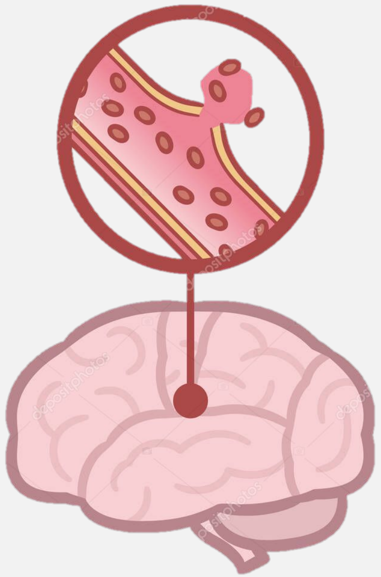
Contents 목차

1. Problem
2. Data Preprocessing
3. Model
4. Result & Conclusion

1. Problem 문제

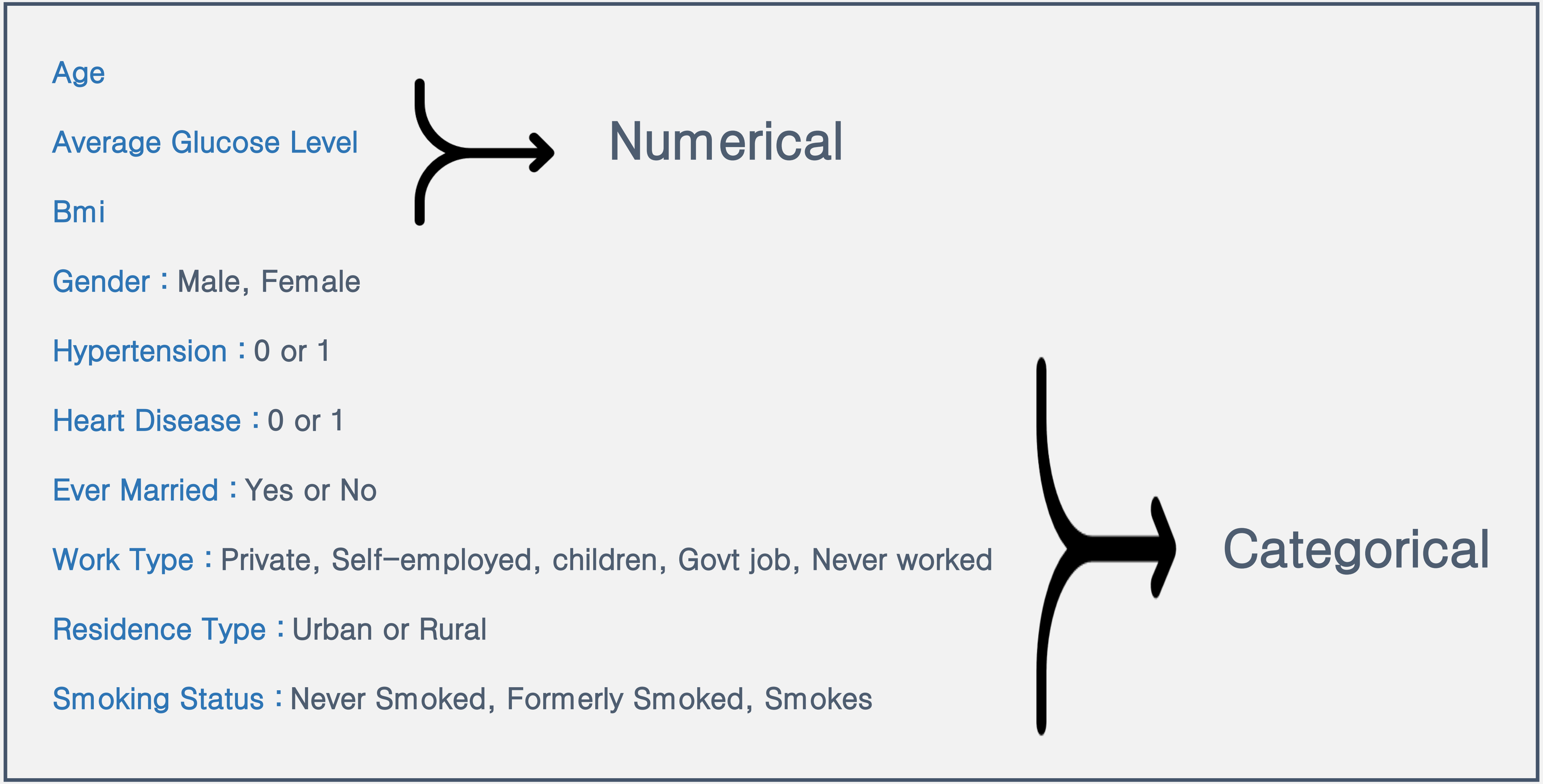
Cerebral Stroke Prediction-Imbalanced Dataset from Kaggle

43400 people



Predict the cerebral stroke with 10 attributes

1. Problem 문제



Age

Average Glucose Level

Bmi

Numerical

Gender : Male, Female

Hypertension : 0 or 1

Heart Disease : 0 or 1

Ever Married : Yes or No

Work Type : Private, Self-employed, children, Govt job, Never worked

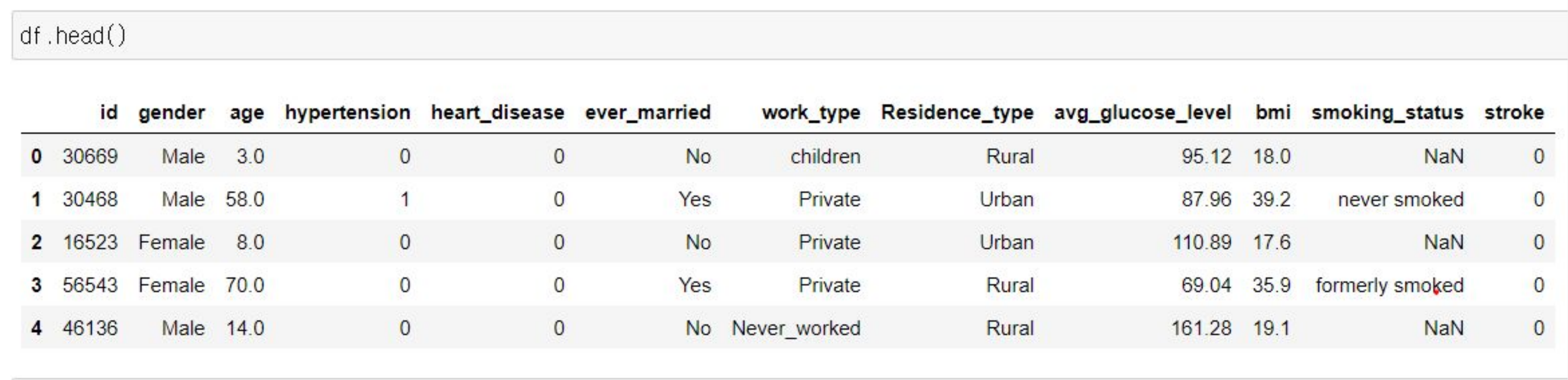
Residence Type : Urban or Rural

Smoking Status : Never Smoked, Formerly Smoked, Smokes

Categorical

2. Data Preprocessing 데이터 전처리

Data Overview

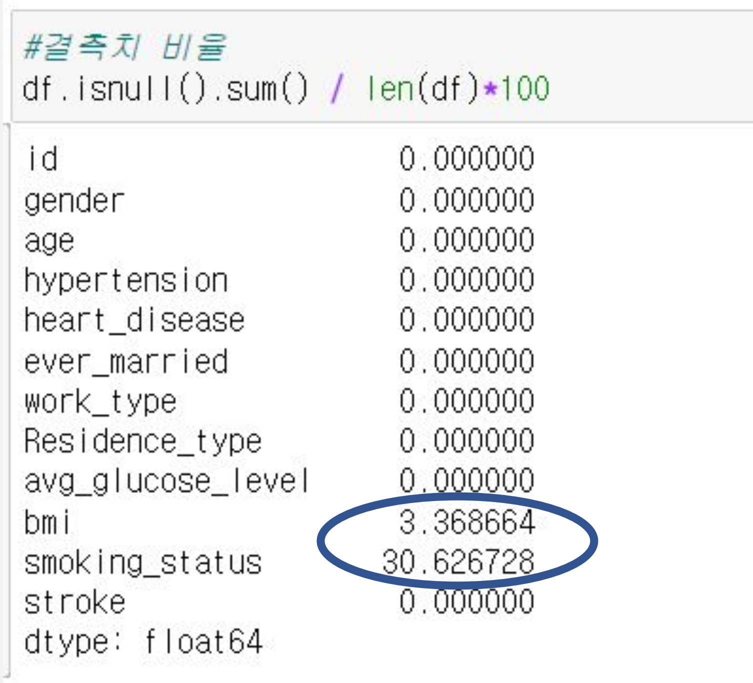


```
df.head()
```

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0	30669	Male	3.0	0	0	No	children	Rural	95.12	18.0	NaN	0
1	30468	Male	58.0	1	0	Yes	Private	Urban	87.96	39.2	never smoked	0
2	16523	Female	8.0	0	0	No	Private	Urban	110.89	17.6	NaN	0
3	56543	Female	70.0	0	0	Yes	Private	Rural	69.04	35.9	formerly smoked	0
4	46136	Male	14.0	0	0	No	Never_worked	Rural	161.28	19.1	NaN	0

2. Data Preprocessing 데이터 전처리

1) Handling Missing Value



```
#결측치 비율
df.isnull().sum() / len(df)*100
```

```
id                   0.000000
gender               0.000000
age                  0.000000
hypertension         0.000000
heart_disease        0.000000
ever_married         0.000000
work_type            0.000000
Residence_type       0.000000
avg_glucose_level    0.000000
bmi                  3.368664
smoking_status      30.626728
stroke               0.000000
dtype: float64
```

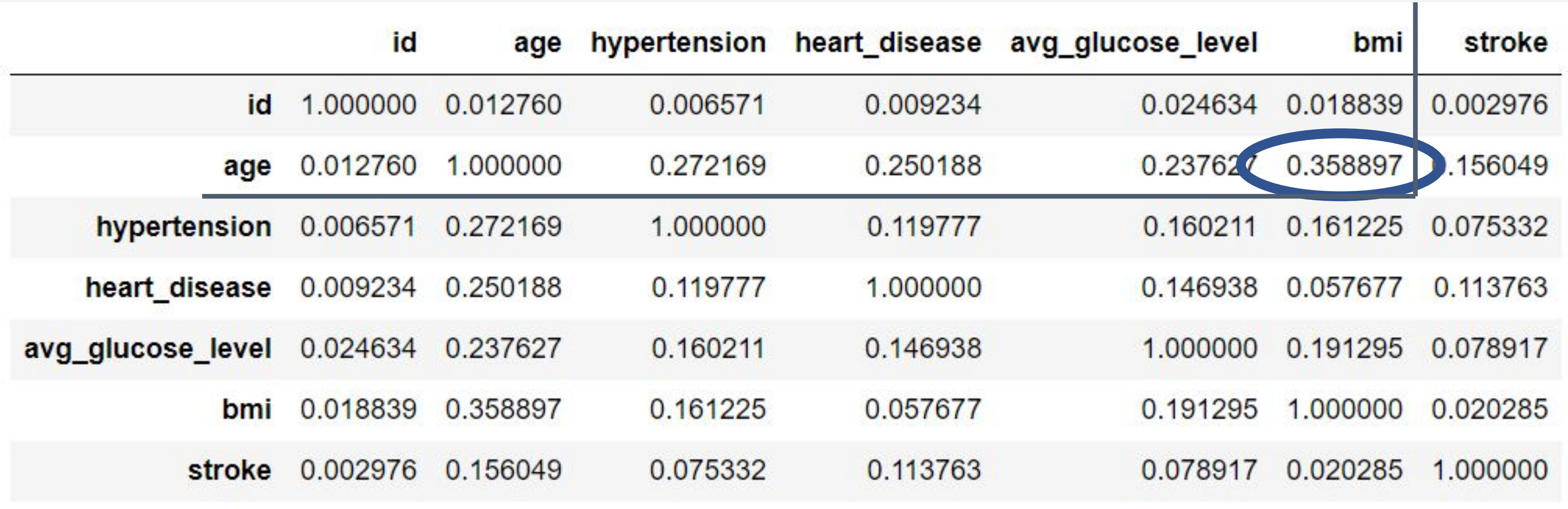
Bmi : 3.3% missing

Smoking Status : 30.6% missing

2-1. Handling Missing Value 데이터 전처리

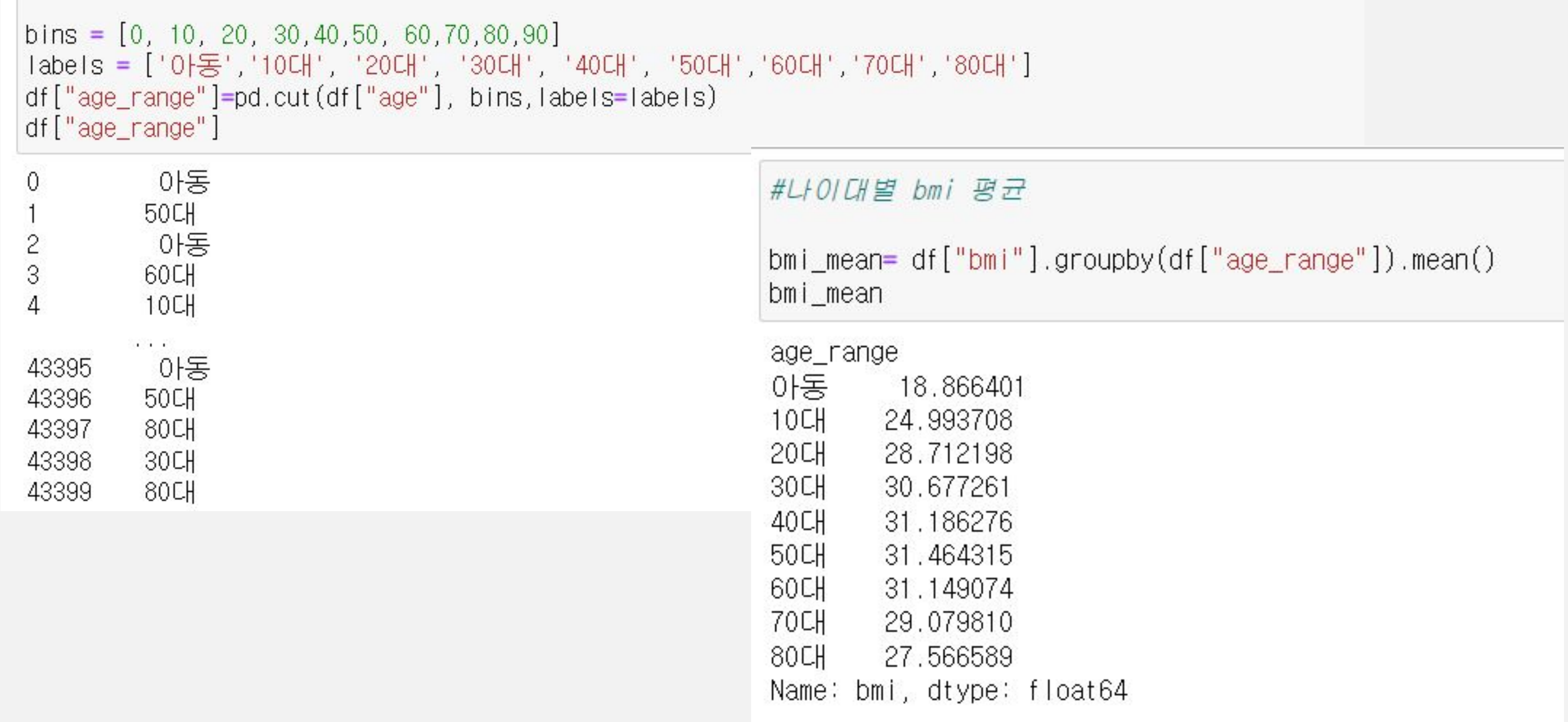
(1) Missing value : bmi

Correlation bmi



	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
id	1.000000	0.012760	0.006571	0.009234	0.024634	0.018839	0.002976
age	0.012760	1.000000	0.272169	0.250188	0.237627	0.358897	0.156049
hypertension	0.006571	0.272169	1.000000	0.119777	0.160211	0.161225	0.075332
heart_disease	0.009234	0.250188	0.119777	1.000000	0.146938	0.057677	0.113763
avg_glucose_level	0.024634	0.237627	0.160211	0.146938	1.000000	0.191295	0.078917
bmi	0.018839	0.358897	0.161225	0.057677	0.191295	1.000000	0.020285
stroke	0.002976	0.156049	0.075332	0.113763	0.078917	0.020285	1.000000

2-1. Handling Missing Value 데이터 전처리



```
bins = [0, 10, 20, 30,40,50, 60,70,80,90]
labels = ['아동','10대', '20대', '30대', '40대', '50대','60대','70대','80대']
df["age_range"]=pd.cut(df["age"], bins,labels=labels)
df["age_range"]
```

```
0         아동
1        50대
2         아동
3        60대
4        10대
          ...
43395     아동
43396    50대
43397    80대
43398    30대
43399    80대
```

```
#나이대별 bmi 평균

bmi_mean= df["bmi"].groupby(df["age_range"]).mean()
bmi_mean
```

```
age_range
아동      18.866401
10대     24.993708
20대     28.712198
30대     30.677261
40대     31.186276
50대     31.464315
60대     31.149074
70대     29.079810
80대     27.566589
Name: bmi, dtype: float64
```

Average bmi grouped by age group

2-1. Handling Missing Value 데이터 전처리

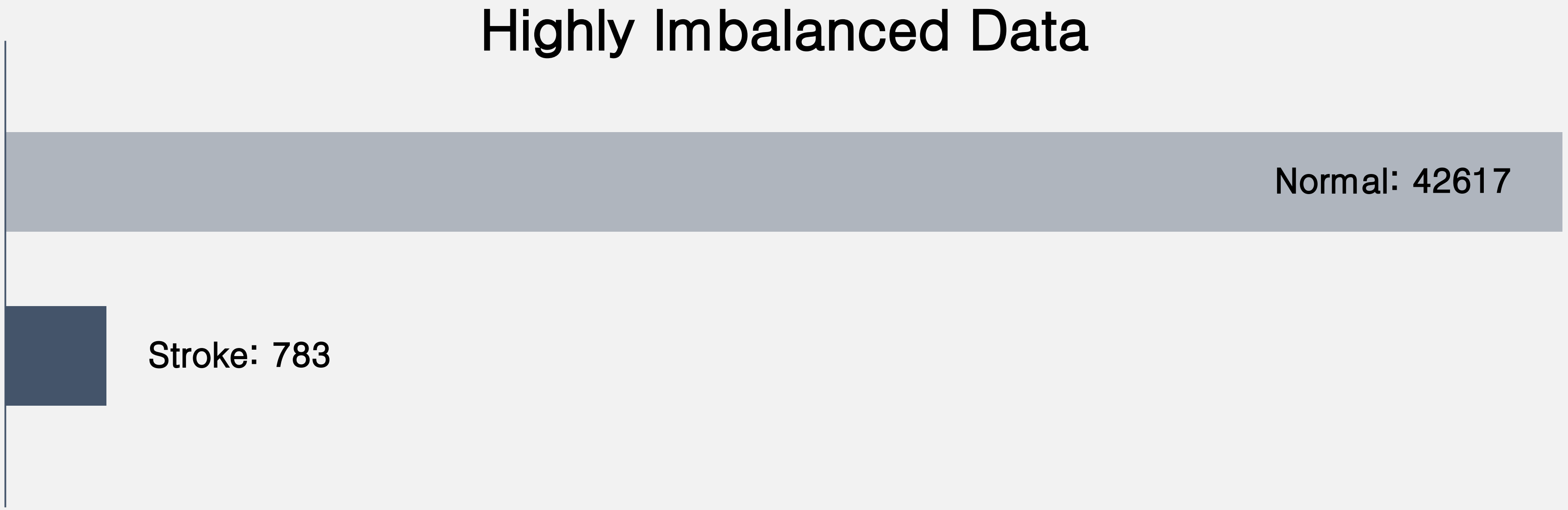
(2) Missing value : Smoking Status

Correlation bmi

2. Data Preprocessing 데이터 전처리

2) Data Analysis

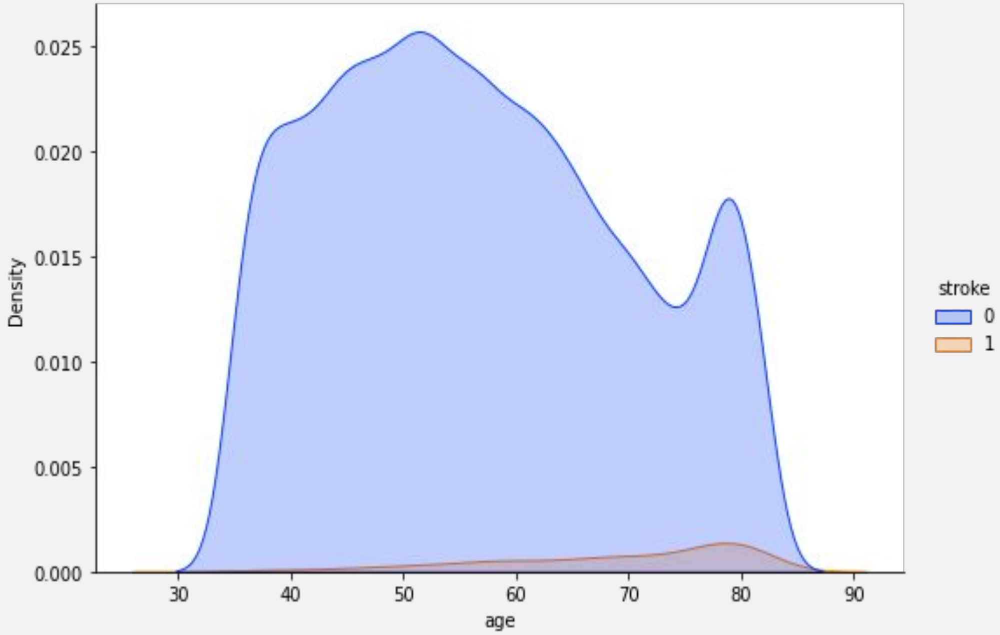
Stroke Occurrence

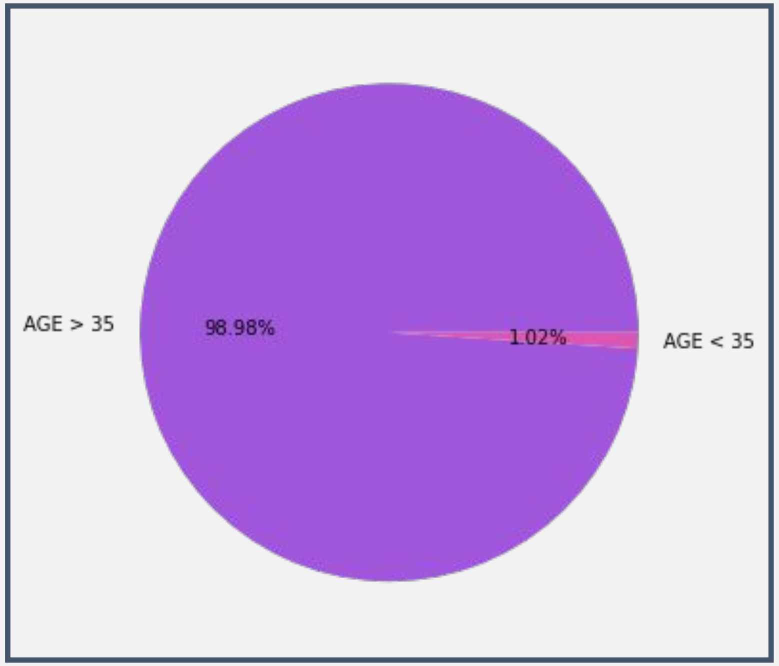


2. Data Preprocessing 데이터 전처리

(1) Discard Data according to Age

Stroke Occurrence according to Age

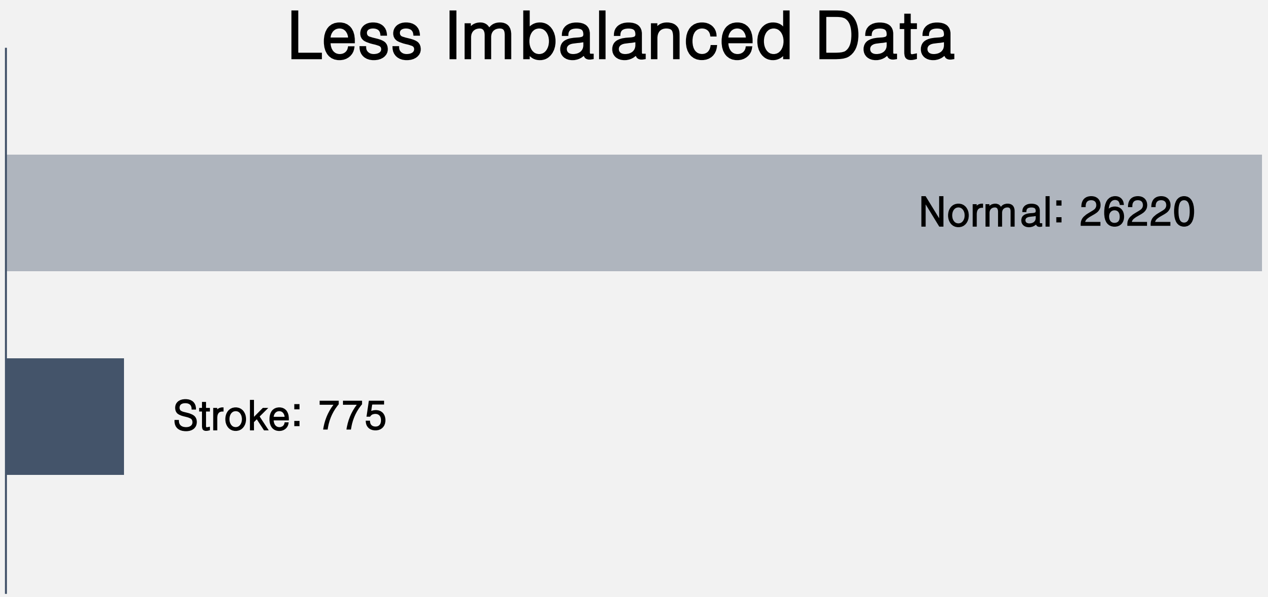




2. Data Preprocessing 데이터 전처리

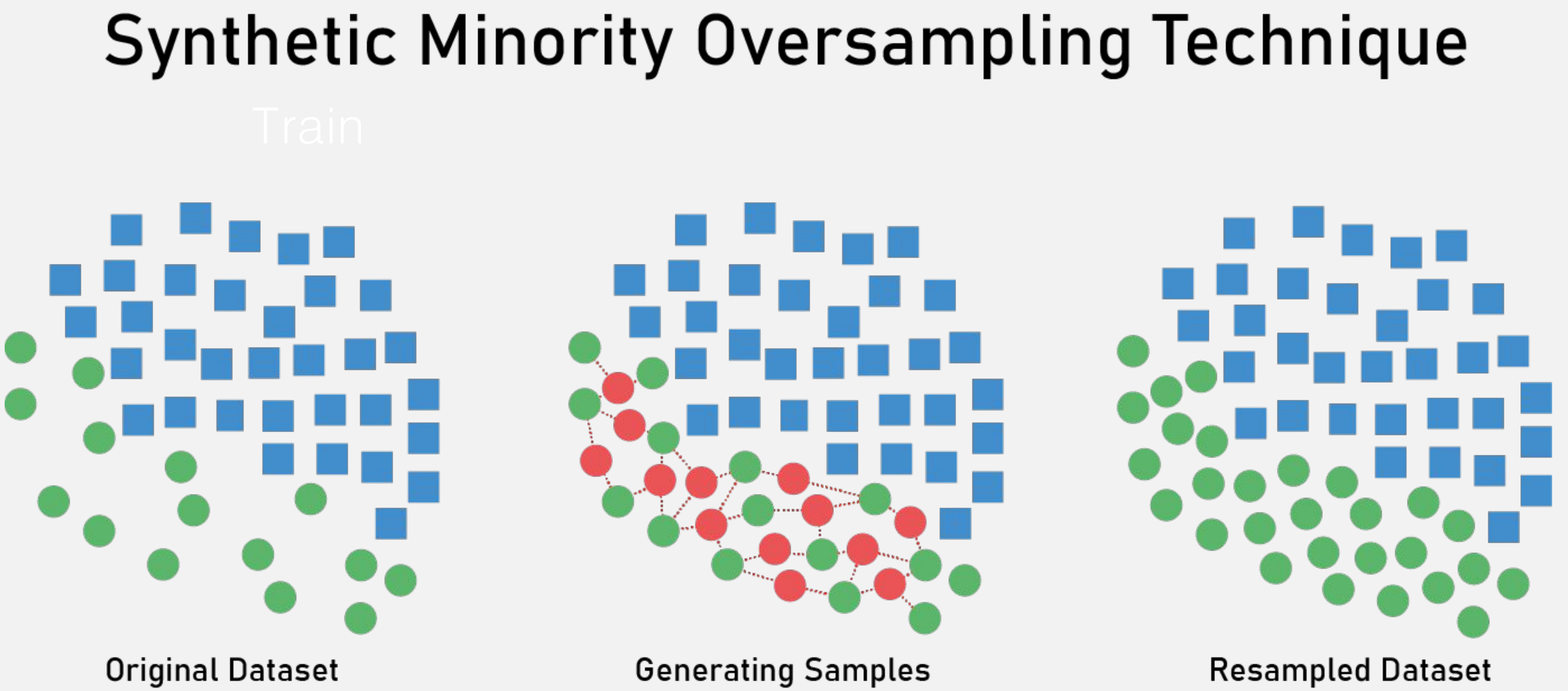
(1) Discard Data according to Age

Stroke Occurrence



2. Data Preprocessing 데이터 전처리

(2) Oversampling: SMOTE

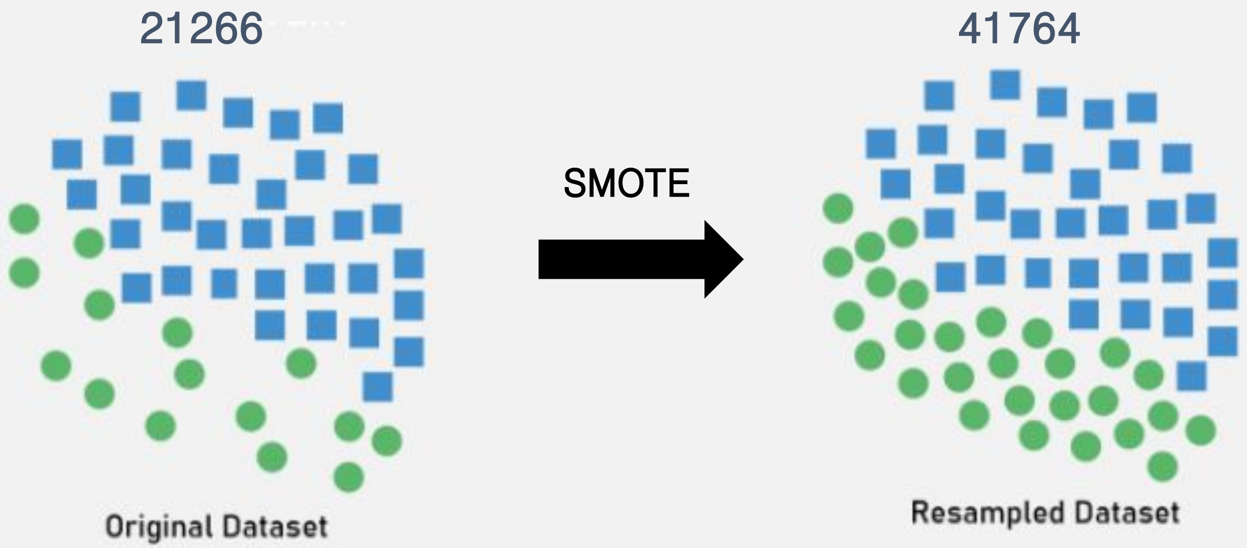


출처:
https://john-analyst.medium.com/smote%EB%A1%9C-%EB%8D%B0%EC%9D%B4%ED%84%B0-%EB%B6%88%EA%B7%A0%ED%98%95-%ED%95%B4%EA%B2%B0%ED%95%98%EA%B8%B0-5ab674ef0b32

2. Data Preprocessing 데이터 전처리

(2) Oversampling: SMOTE

Less Imbalanced Data



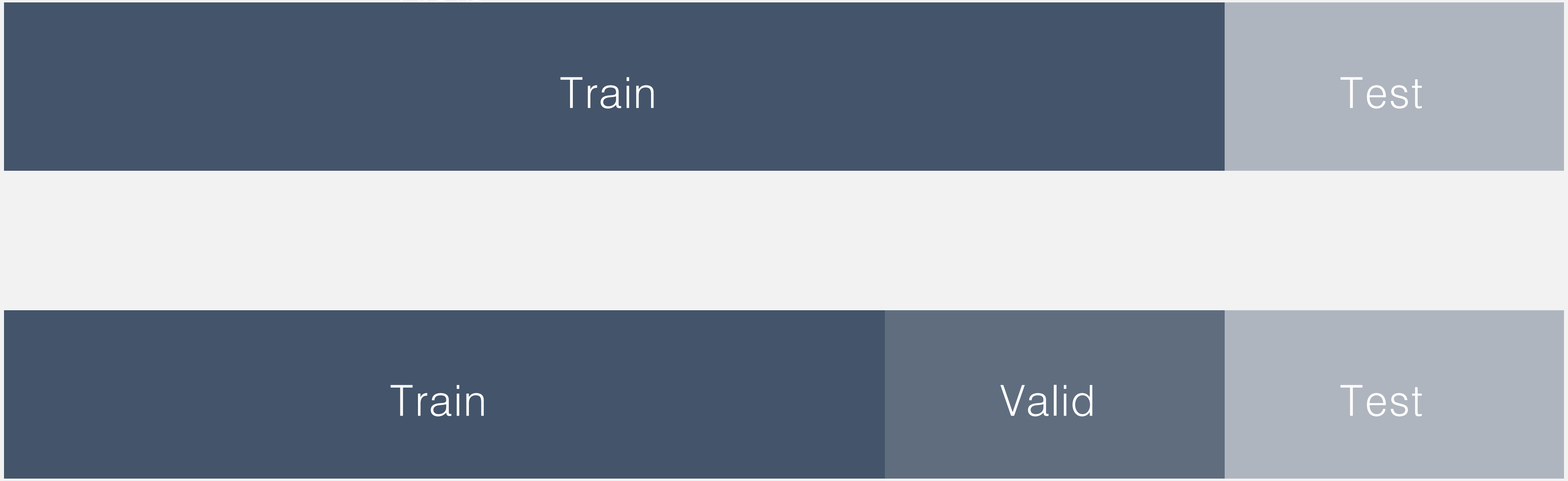
21266

SMOTE

41764

2. Data Preprocessing 데이터 전처리

(3) Splitting Data Set



Train

Test

Train

Valid

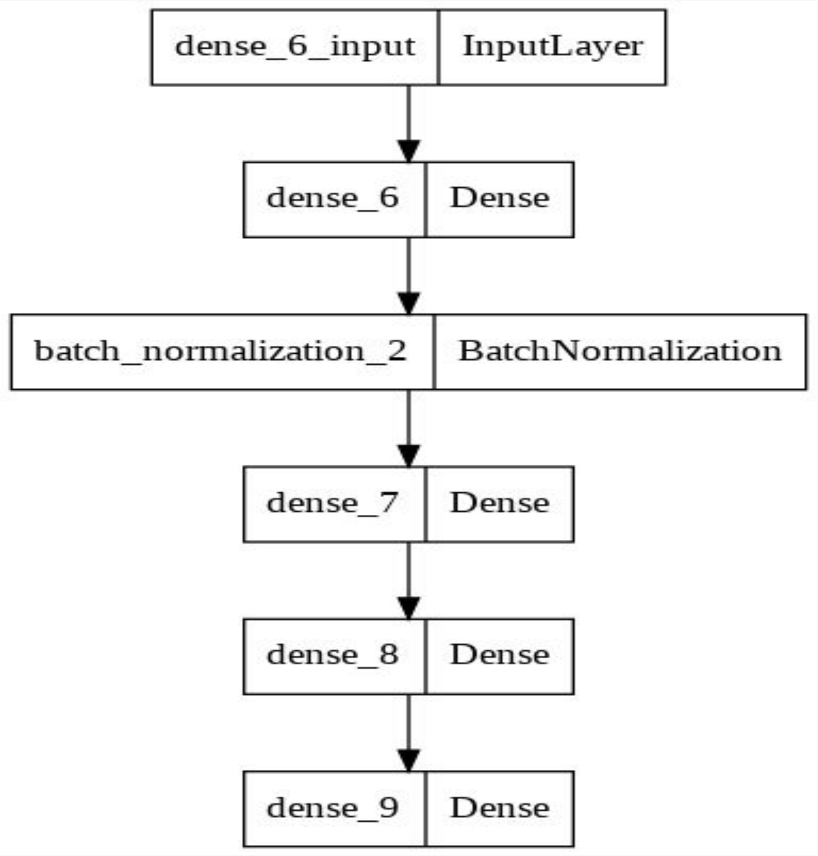
Test

2. Data Preprocessing 데이터 전처리

3) Pipeline

3. Model Selection 모델 선정

1) MLP



MLP result

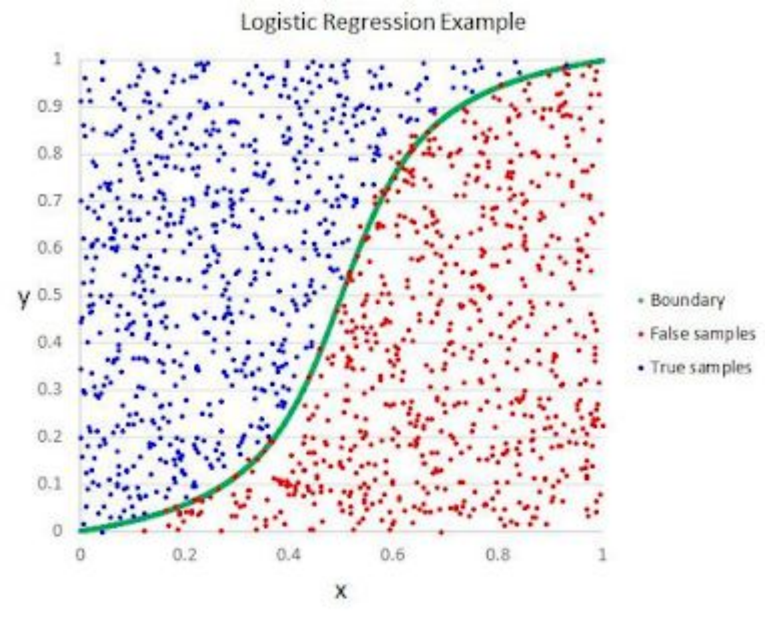
Train accuracy : 95.58%

Valid accuracy : 89.64%

ROC auc score: 0.79

3. Model Selection 모델 선정

2) Logistic Regression



출처:
https://sonsnotation.blogspot.com/2020/11/2-logistic-regression.html

Logistic Regression result

Train accuracy : 78.44%

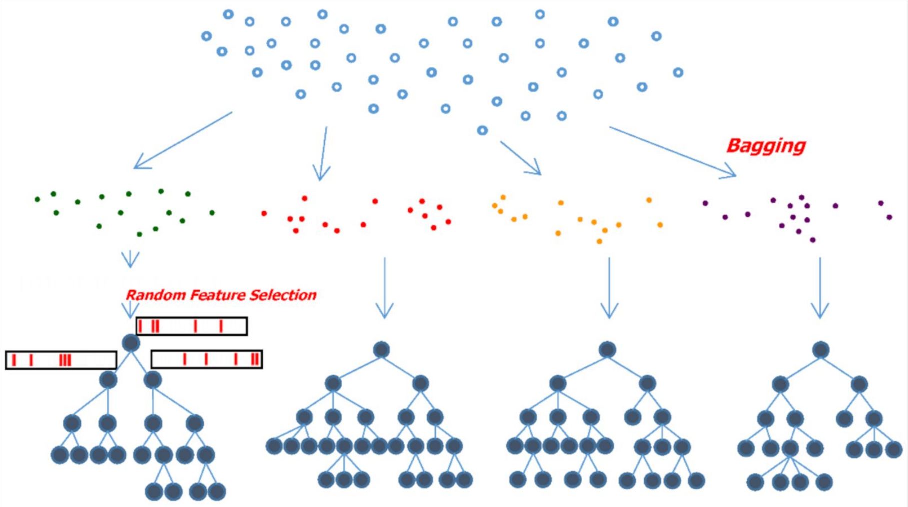
Valid accuracy : 73.89%

ROC auc score: 0.77

f1 score: 0.8

3. Model Selection 모델 선정

3) Random Forest Classifier



출처: https://www.mdpi.com/2227-7102/11/3/92

Random Forest Classifier result

Train accuracy : 100.0%

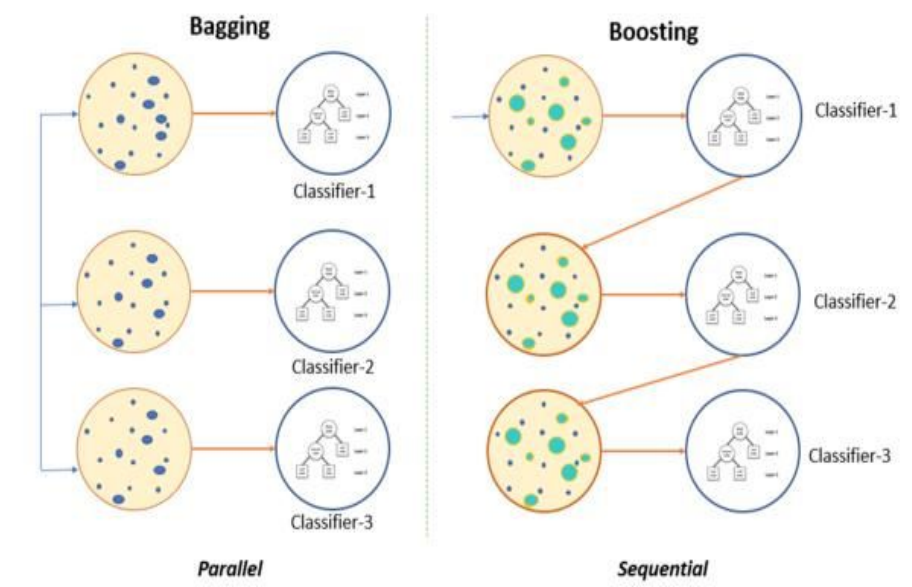
Valid accuracy : 92.58%

ROC auc score: 0.57

f1 score: 1.0

3. Model Selection 모델 선정

4) XGB Classifier



출처: http://egloos.zum.com/incredible/v/7478695

XGB Classifier result

Train accuracy : 86.84%

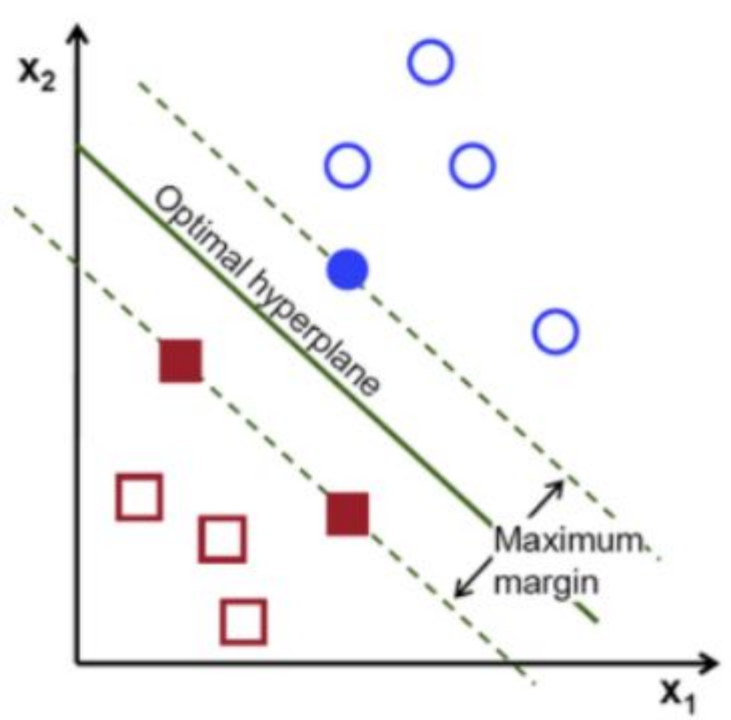
Valid accuracy : 74.71%

ROC auc score: 0.75

f1 score: 0.86

3. Model Selection 모델 선정

5) Linear SVC + class weight



출처: https://wooono.tistory.com/111

Linear SVC result

Train accuracy : 78.64%

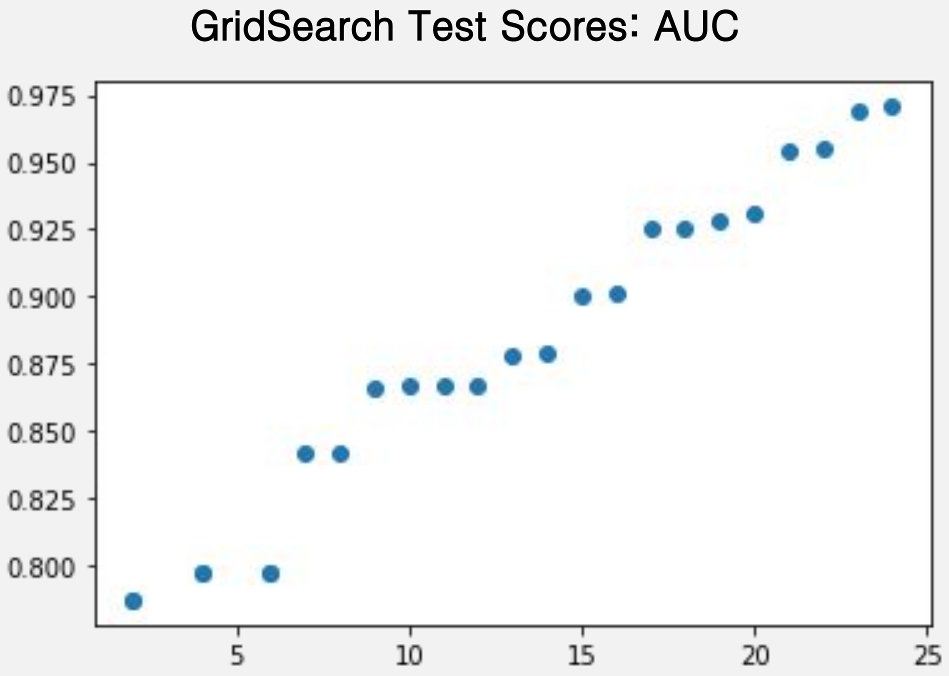
Valid accuracy : 73.28%

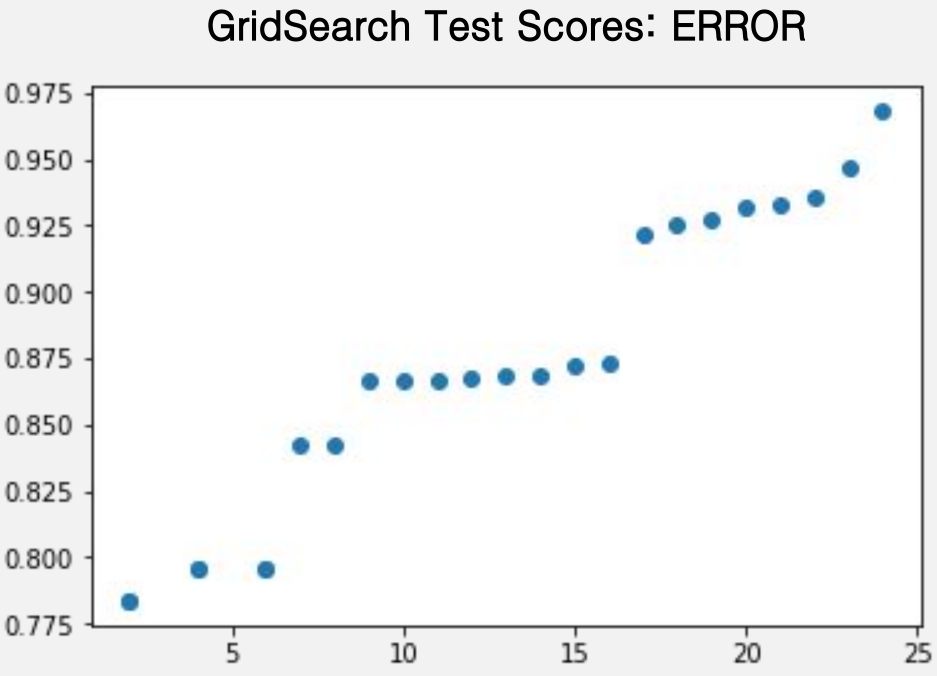
ROC auc score: 0.77

f1 score: 0.80

4. Model Tuning 모델

1) Hyperparameter Tuning


GridSearch Test Scores: AUC


GridSearch Test Scores: ERROR

5. Model Test 모델 평가

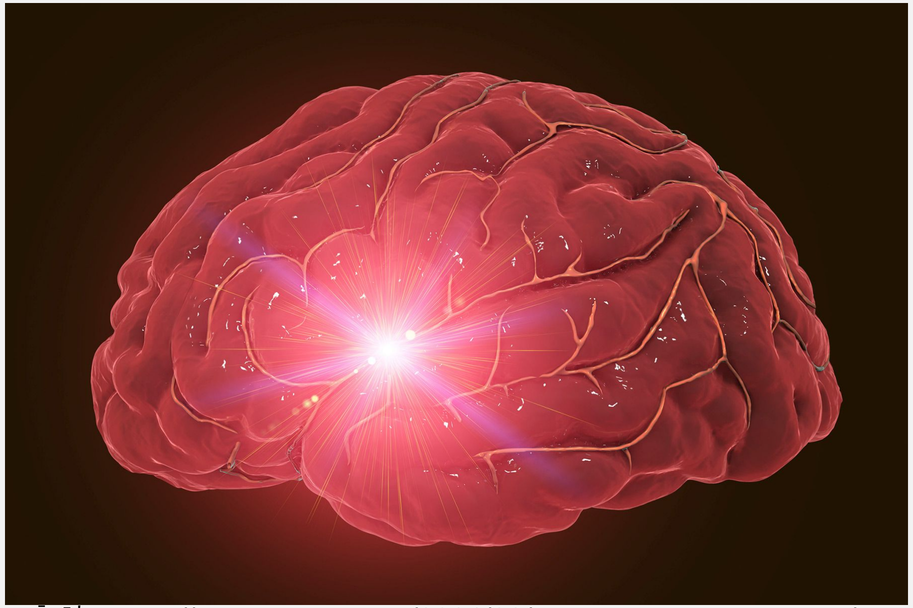
Test results

Test accuracy : 79.78%

Test stroke precision : 0.73

ROC auc score: 0.80

f1 score: 0.82



출처: https://today.uconn.edu/2021/02/stopping-stroke-damage/