Dự đoán nguy cơ đột quỵ

Present by: Duong Hieu - Le

Guided by: Assoc. Prof. Dr. Van Hau - Nguyen

Table of contents

01

Introduce the problem

Bệnh đột quỵ là gì?

Theo Tổ chức Y tế Thế giới (WHO), đột quỵ là nguyên nhân gây tử vong đứng thứ hai trên toàn cầu. Đây là một bệnh có tỷ lệ mắc bệnh, tàn tật và tử vong cao, trong đó các nạn nhân có thể đột ngột trải qua tình trạng liệt, suy giảm khả năng nói hoặc mất thị lực do sự gián đoạn dòng máu (thiếu máu cục bộ) gây ra bởi huyết khối và thuyên tắc.

Hình 1: Minh họa bệnh đột quỵ [1]

Lý do chọn đề tài

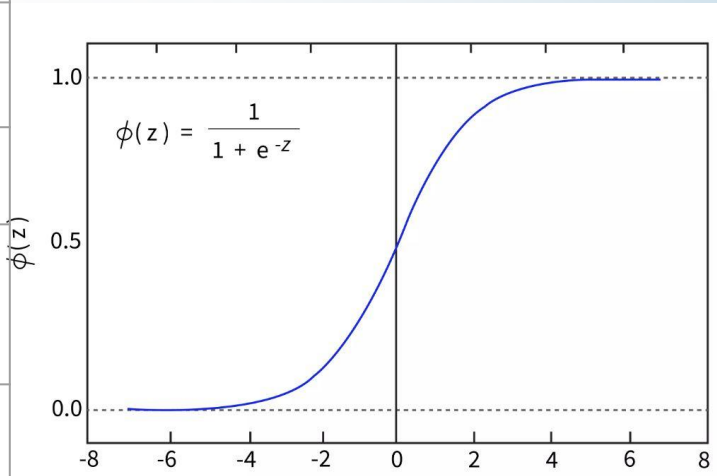
Hình 2: Thống kê số ca đột quỵ hàng năm tại Việt Nam [2]

02

Module

Logistic Regression

Tham số	sklearn mặc định	Từ bài báo	Ghi chú
penalty	'l2'	'l2'	Giảm trọng số của các đặc trưng mà không loại bỏ chúng
solver	'lbfgs'	'liblinear'	Tối ưu hơn cho dữ liệu nhỏ & imbalance
C	1.0	1.0	Là quy chuẩn hóa mạnh hơn (phạt lớn hơn), buộc mô hình phải đơn giản hơn (giảm overfitting).
class_weight	None	'balanced'	Tự động tính trọng số tỷ lệ nghịch với tần suất lớp.
max_iter	100	200	Số lần lặp tối đa để thuật toán tối ưu hóa hội tụ
random_state	None	42	Để tái lập kết quả

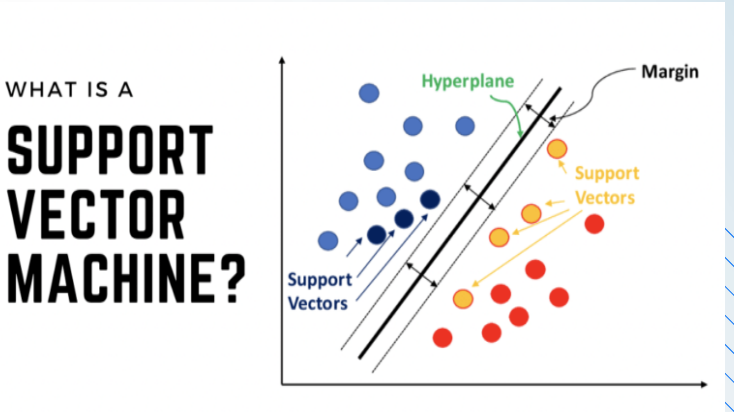


Hình 3: Thuật toán Hồi quy Logistic [3]

https://www.cambridge.org/core/journals/political-analysis/article/logistic-regression-in-rare-events-data/1E09F0F36F89DF12A823130FDF0DA462

Support Vector Machine

Tham số	sklearn mặc định	Từ bài báo	Ghi chú
kernel	'rbf'	'rbf'	Xử lý các quan hệ phi tuyến tính
C	1.0	10	Tăng độ phạt để tách lớp tốt hơn
gamma	'scale'	0.01	Gamma cố định nhỏ giúp ổn định trên imbalance
class_weight	None	'balanced'	Xử lý mất cân bằng lớp
random_state	None	42	Để tái lập kết quả

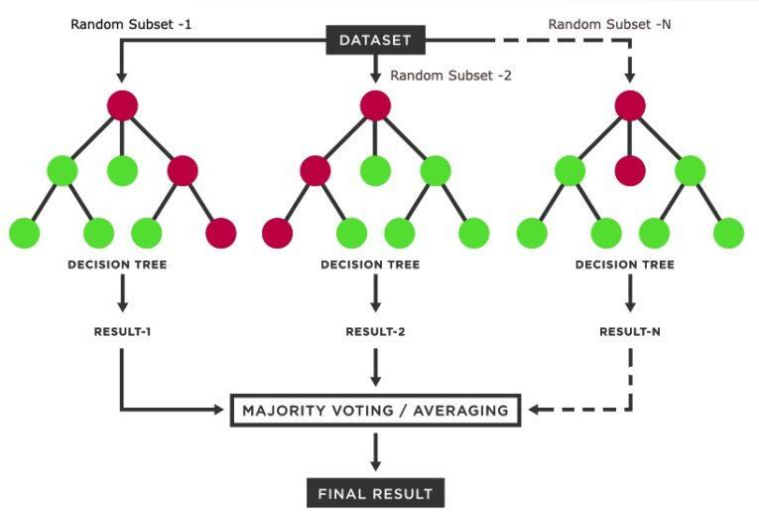


Hình 4: Thuật toán Support Vector Machine [4]

https://www.sciencedirect.com/science/article/abs/pii/S0957417419305573?via%3Dihub

Random Forest

Tham số	sklearn mặc định	Từ bài báo	Ghi chú
n_estimators	100	300	Nhiều cây → ổn định hơn
max_depth	None	20	Tránh overfitting
min_samples_split	2	2	Số lượng mẫu tối thiểu cần thiết để một nút bên trong được phép chia.
min_samples_leaf	1	4	Giúp giảm noise lớp thiểu số
max_features	'sqrt'	'sqrt'	Số lượng đặc trưng tối đa được xem xét khi tìm kiếm sự phân chia tốt nhất tại một nút.
class_weight	None	'balanced_subsample'	Xử lý imbalance
random_state	None	42	Để tái lập



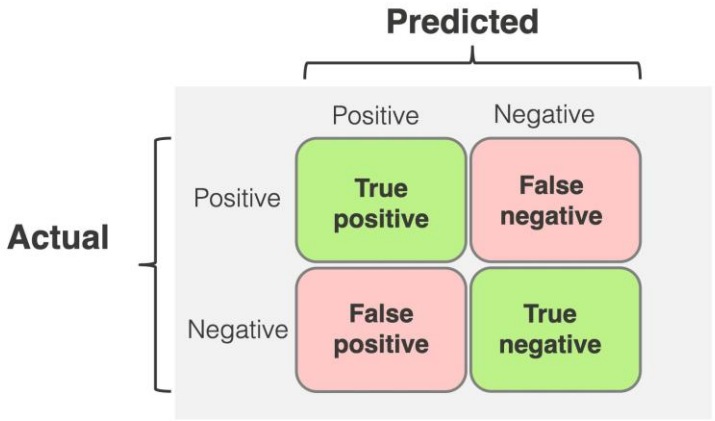
Hình 5: Thuật toán Random Forest [5]

https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf

Phương pháp đánh giá

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$



Hình 6: Ma trận nhầm lẫn [6]

03

Processing and prediction

Data

	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	No	children	Rural	95.12	18	NaN	0
1	30468	Male	58	1	0	Yes	Private	Urban	87.96	39.2	never smoked	0
2	16523	Female	8	0	0	No	Private	Urban	110.89	17.6	NaN	0
3	56543	Female	70	0	0	Yes	Private	Rural	69.04	35.9	formerly smoked	0
4	46136	Male	14	0	0	No	Never_worked	Rural	161.28	19.1	NaN	0
...	...	...	...	...	...	...	...	...	...	...	...	...
43399	36271	Female	82	0	0	Yes	Private	Urban	79.48	20.6	never smoked	0

Bảng 1: Dữ liệu

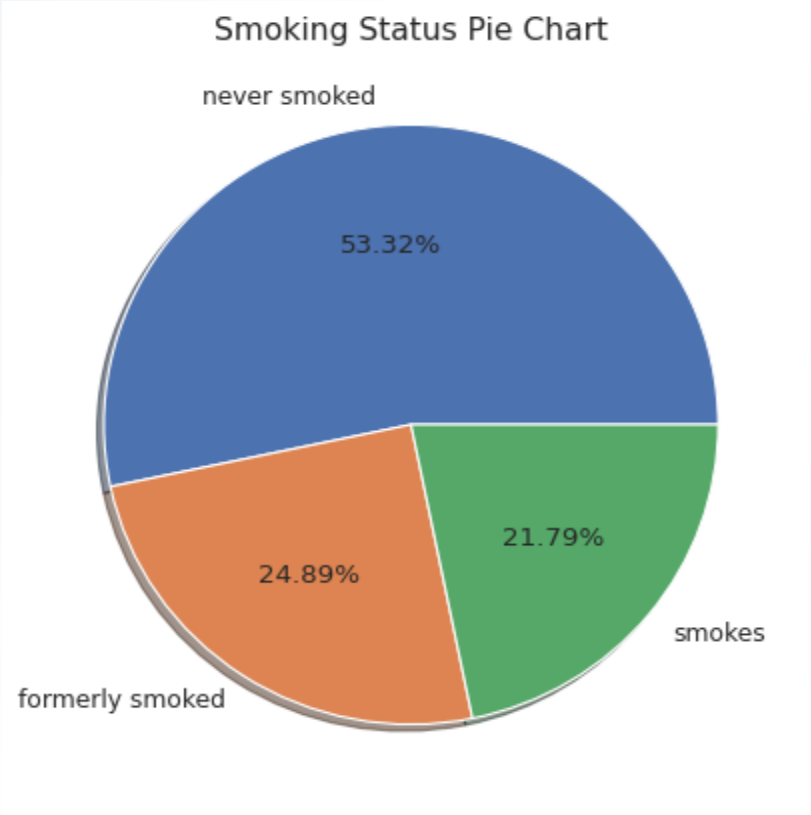
(43400,12)

Dữ liệu của cột

Cột	Kiểu dữ liệu	Mục đích
id	int64	Mã định danh của từng bệnh nhân.
gender	object	Giới tính của bệnh nhân.
age	float64	Tuổi của bệnh nhân.
hypertension	int64	Cho biết liệu bệnh nhân có bị cao huyết áp hay không.
heart_disease	int64	Cho biết liệu bệnh nhân có mắc bệnh tim mạch hay không.
ever_married	object	Cho biết liệu bệnh nhân đã từng kết hôn hay chưa.
work_type	object	Loại công việc của bệnh nhân.
Residence_type	object	Loại nơi cư trú của bệnh nhân (thành thị hoặc nông thôn).
avg_glucose_level	float64	Mức đường huyết trung bình của bệnh nhân.
bmi	float64	Chỉ số khối cơ thể (Body Mass Index) của bệnh nhân.
smoking_status	object	Tình trạng hút thuốc của bệnh nhân.

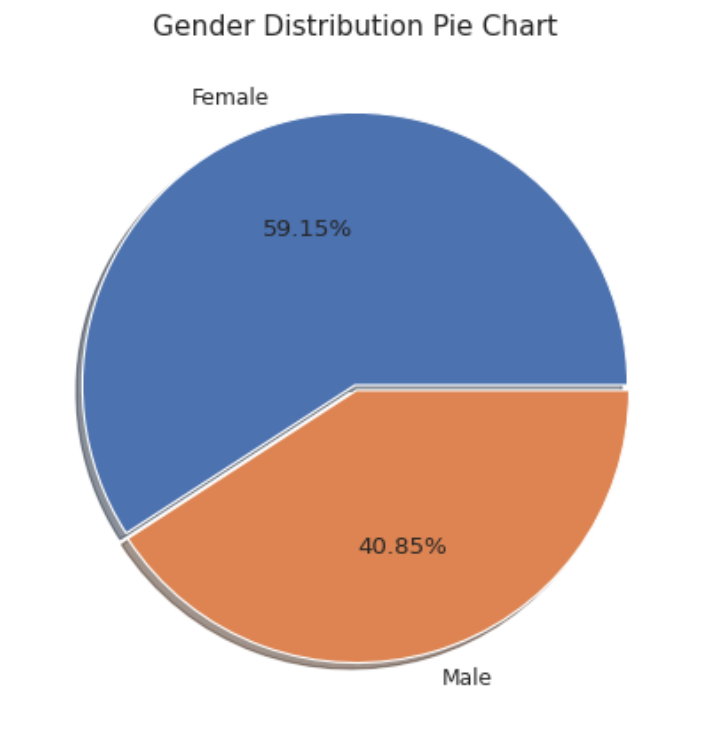
Bảng 2: Kiểu dữ liệu của các Feature

Tương quan dữ liệu

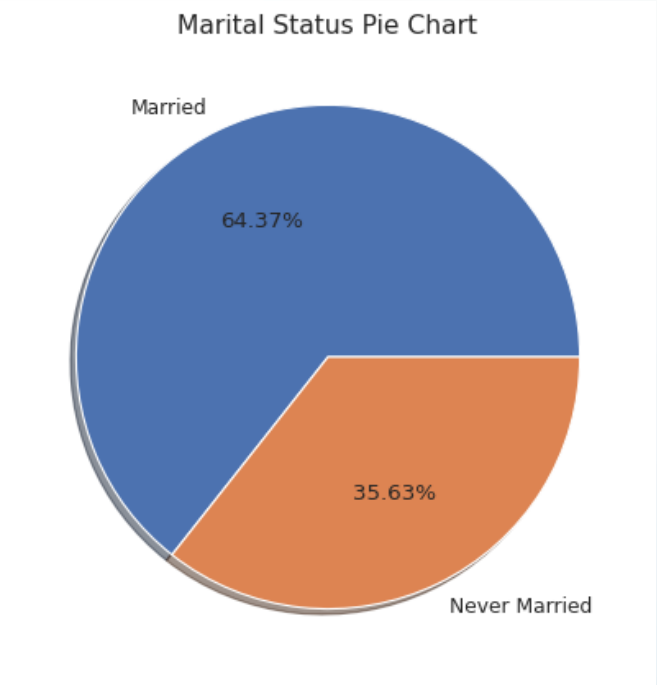


Hình 7: Feature Smoking_status

Tương quan dữ liệu

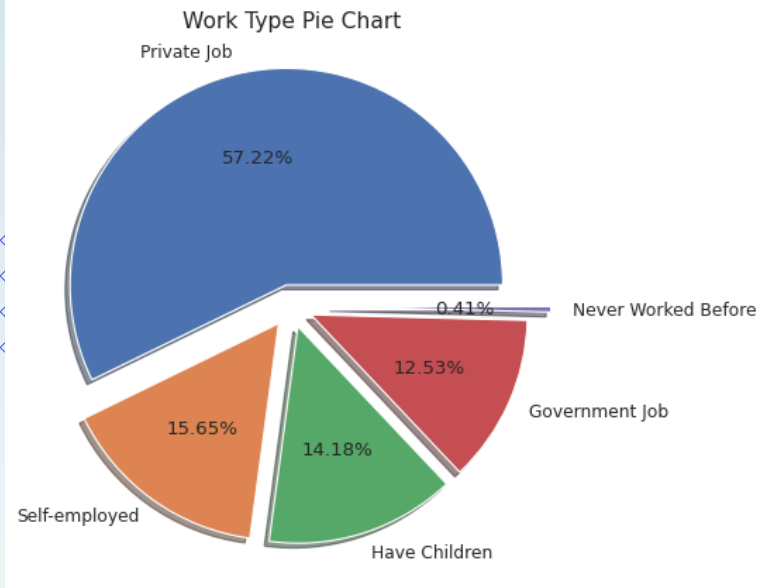


Hình 8: Feature Gender

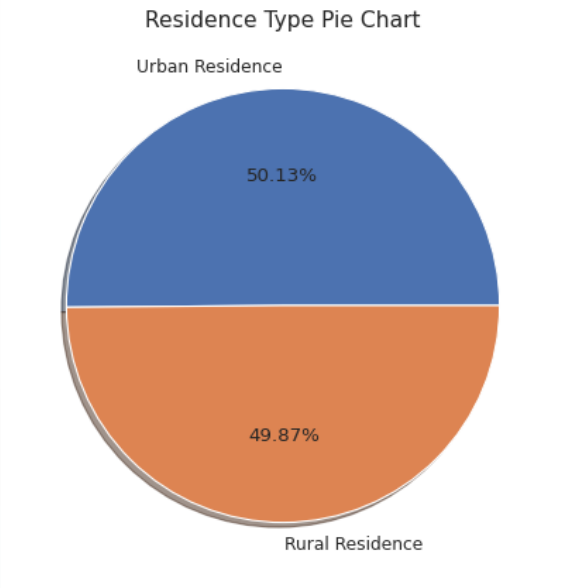


Hình 9: Feature Ever_married

Tương quan dữ liệu

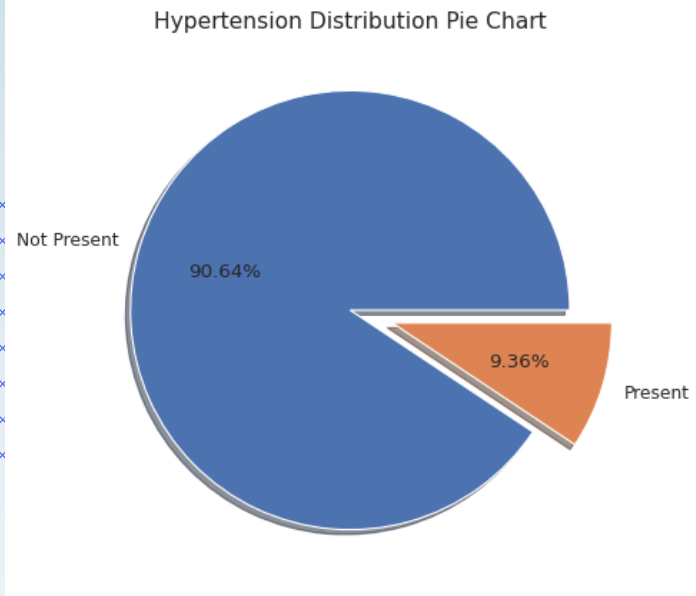


Hình 10: Feature Work_type

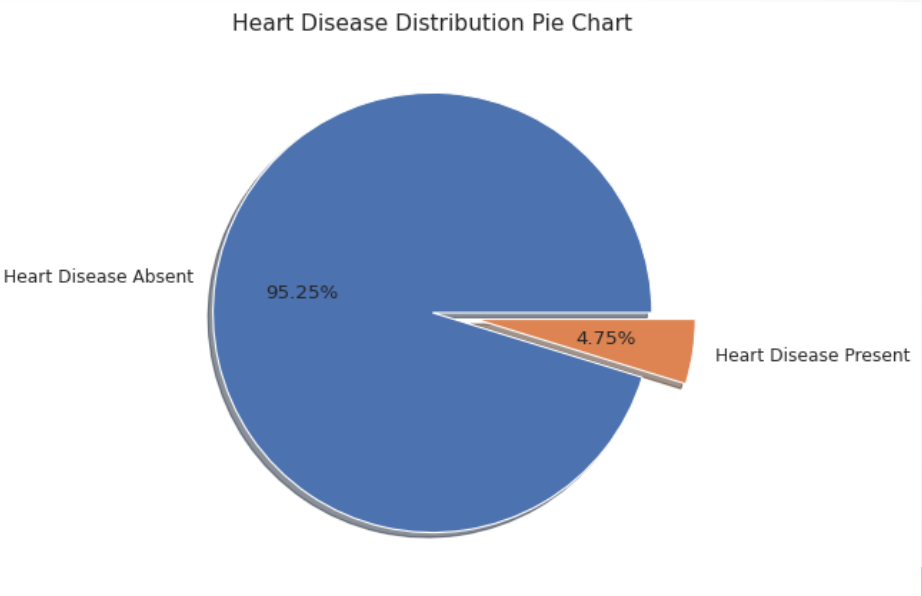


Hình 11: Feature Residence_type

Tương quan dữ liệu



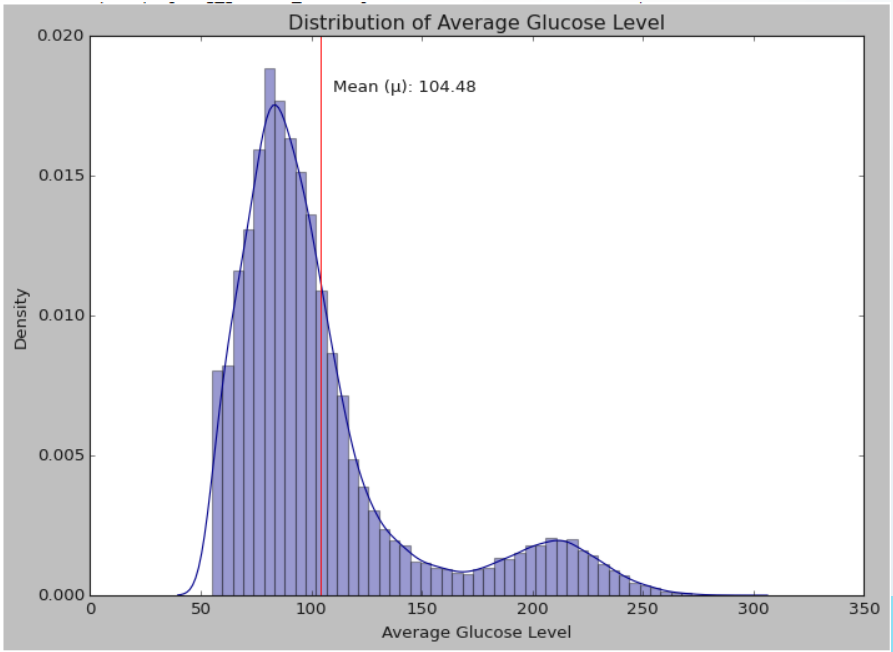
Hình 12: Feature Hypertension



Hình 13: Feature Heart Disease

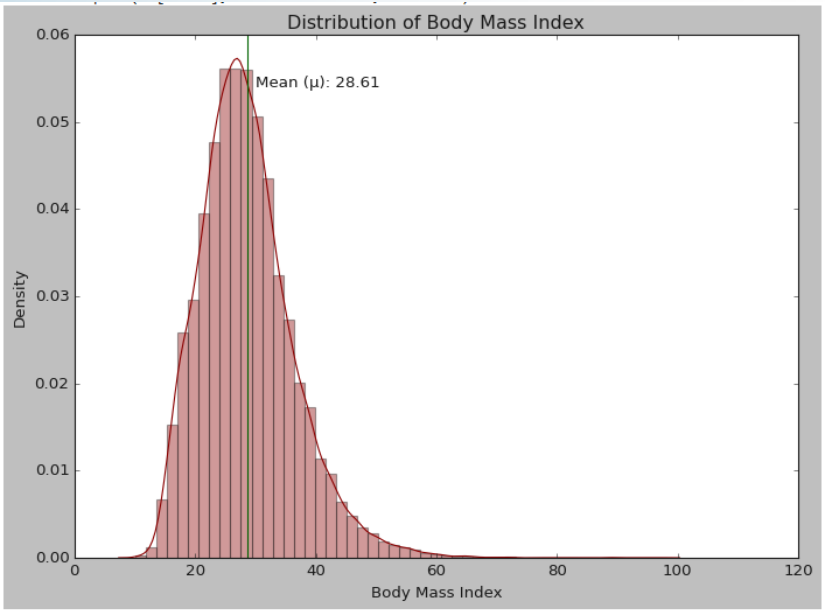
Tương quan dữ liệu

Hình 14: Feature Ages

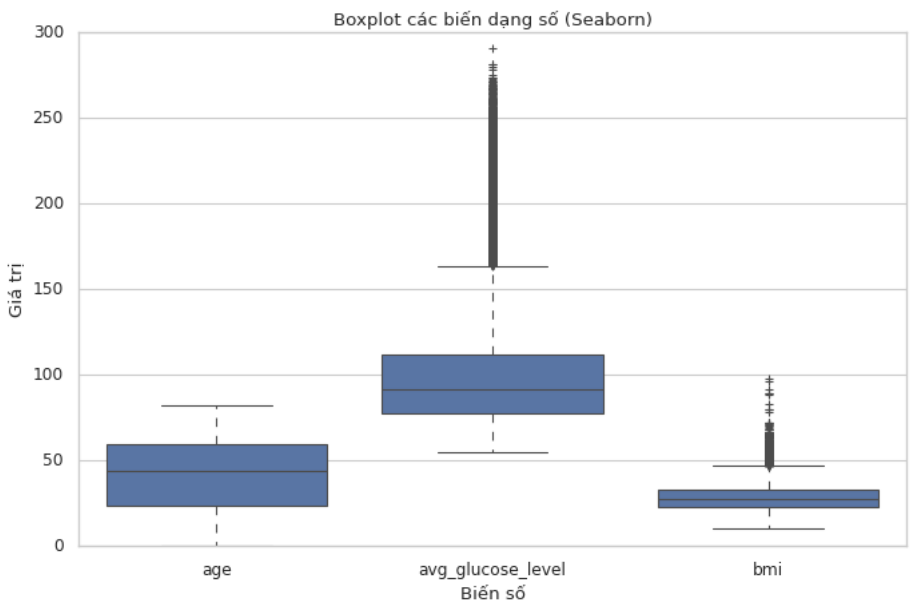


Hình 15: Feature Average Glucose

Tương quan dữ liệu

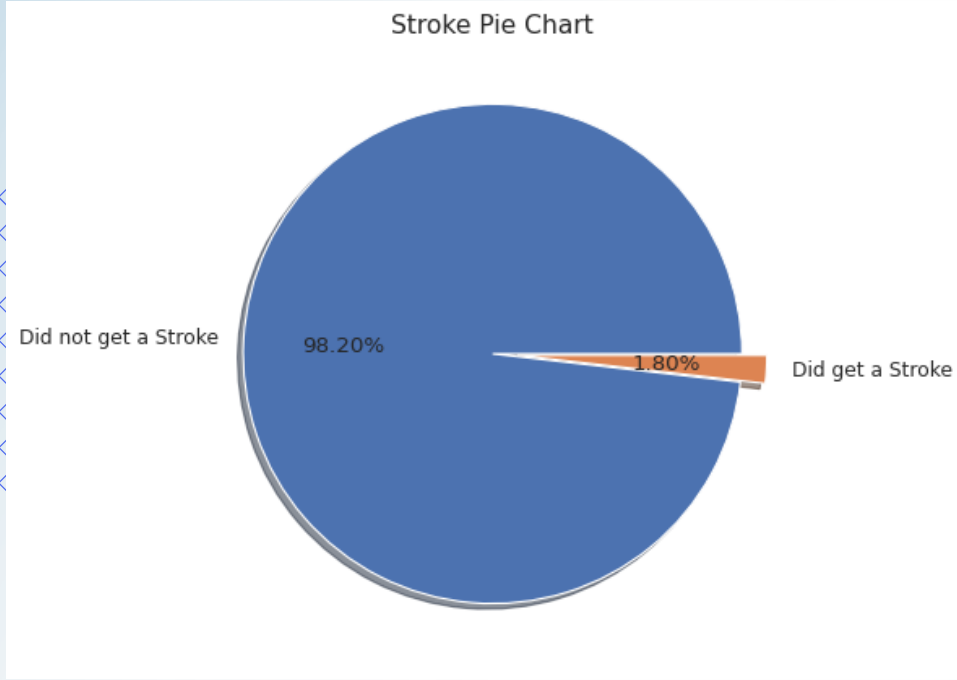


Hình 16: Feature BMI

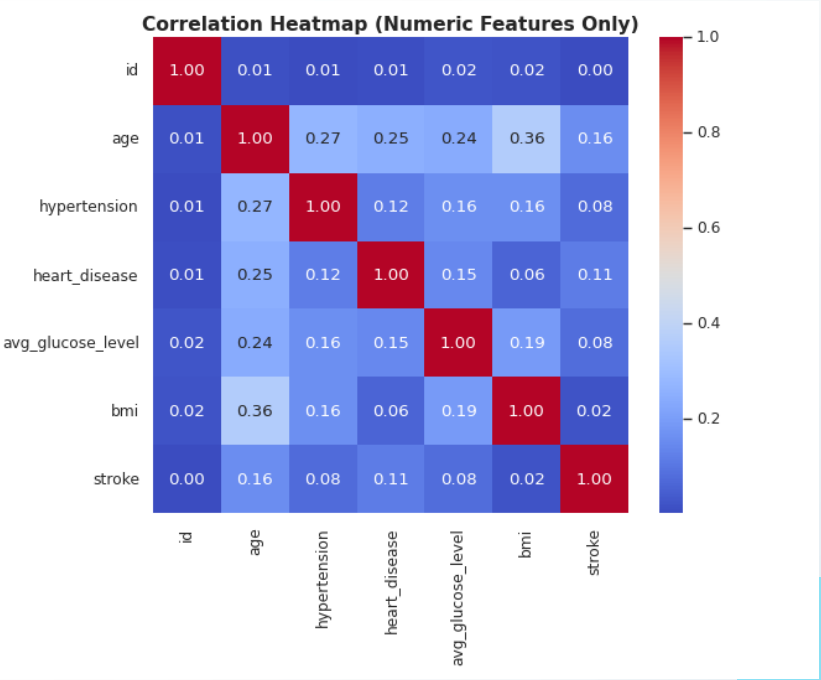


Hình 17: Boxplot

Tương quan dữ liệu

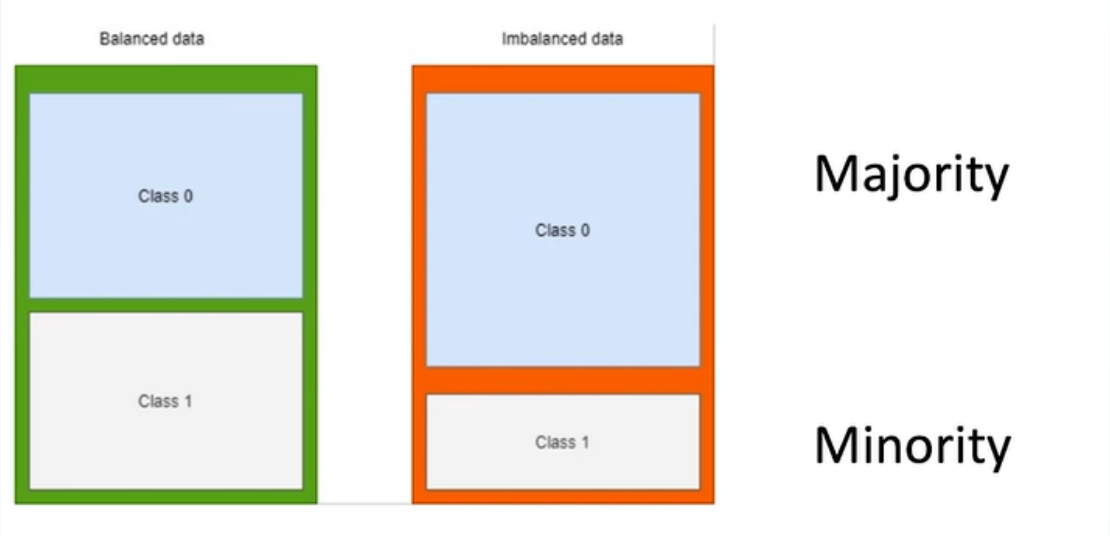


Hình 18: Lable



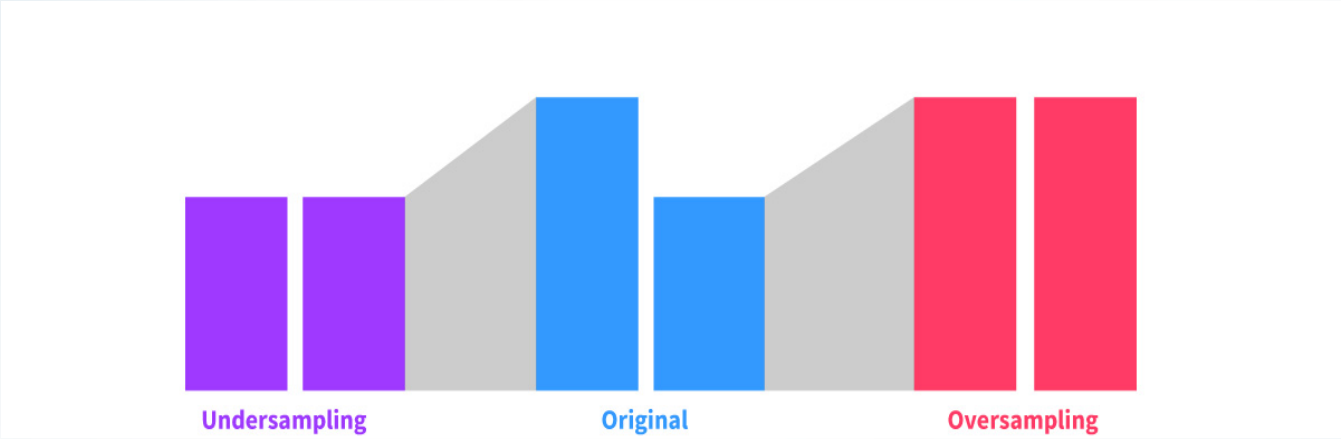
Hình 19: Heat map

Vấn đề Imblance



Hình 20: Các lớp trong dữ liệu Imblance [7]

Cách xử lý



Hình 21: Các cách xử lý dữ liệu Imblance [8]

Cách xử lý

NearMiss-1:

Các mẫu thuộc lớp đa số có khoảng cách trung bình nhỏ nhất tới ba mẫu gần nhất của lớp thiểu số.

NearMiss-2:

Các mẫu thuộc lớp đa số có khoảng cách trung bình nhỏ nhất tới ba mẫu xa nhất của lớp thiểu số.

NearMiss-3

Các mẫu thuộc lớp đa số có khoảng cách nhỏ nhất tới mỗi mẫu của lớp thiểu số.

Hình 22: Near miss

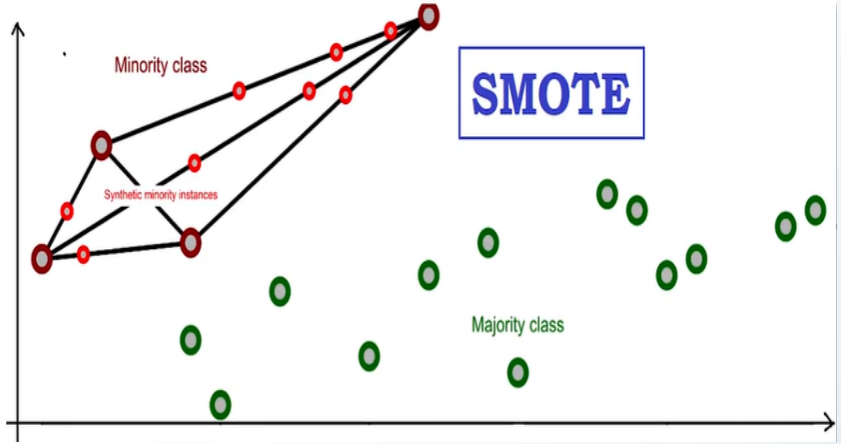
Cách xử lý

Ưu điểm (Advantages)

- Giảm thiểu vấn đề overfitting do random oversampling gây ra, vì các mẫu tổng hợp (synthetic samples) được tạo ra thay vì chỉ sao chép lại các mẫu có sẵn.
- Không làm mất thông tin hữu ích.

Nhược điểm (Disadvantages)

- Khi tạo ra các mẫu tổng hợp, SMOTE không xét đến các mẫu lân cận thuộc lớp khác. Điều này có thể làm tăng sự chồng lấn giữa các lớp và có thể gây thêm nhiễu cho dữ liệu.
- SMOTE không thực sự hiệu quả với dữ liệu có số chiều cao (high-dimensional data).

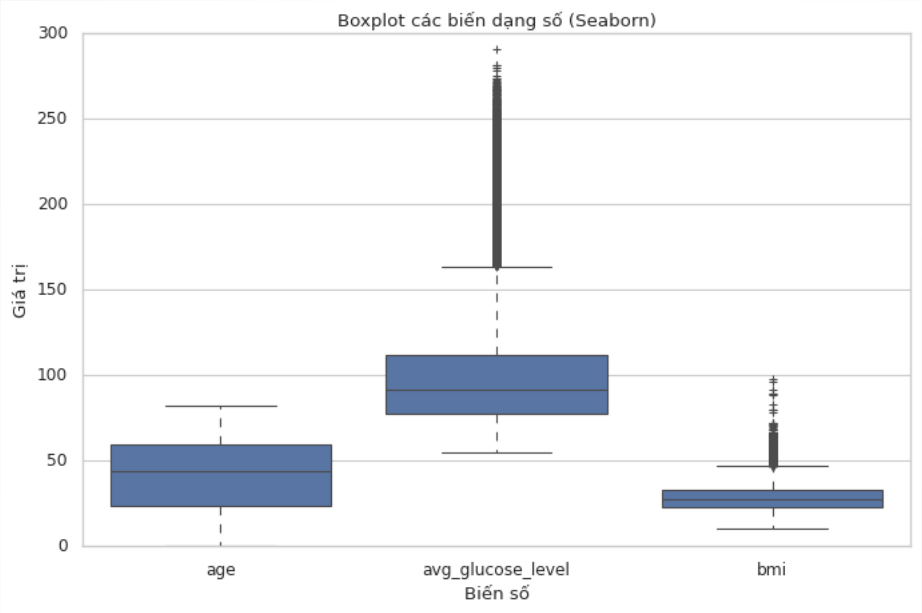


Hình 23: SMOTE

Pre-Proccesing

Bỏ cột "ID"

Bỏ Out Layer



Hình 24: Biểu đồ Boxplot của các biến Numberic

Pre-Proccesing

Bảng 3: Dữ liệu sau khi loại bỏ OutLayer

gender	0
age	0
hypertension	0
heart_disease	0
ever_married	0
work_type	0
Residence_type	0
avg_glucose_level	0
bmi	140
smoking_status	12043
stroke	0

Pre-Proccesing

Xử lý Feature "BMI"

Bỏ các giá trị thiếu có nhãn là 0

Điền các giá trị còn lại theo phân phối xác suất

Bảng 4: Sau khi xử lý

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
36240	Female	1.88	0	0	No	children	Urban	101.41	16.9	NaN	0
13214	Male	13.00	0	0	No	children	Rural	85.40	26.3	NaN	0
1648	Female	23.00	0	0	No	Private	Rural	92.26	17.1	NaN	0

Pre-Proccesing

Chia "Train" và "Test"

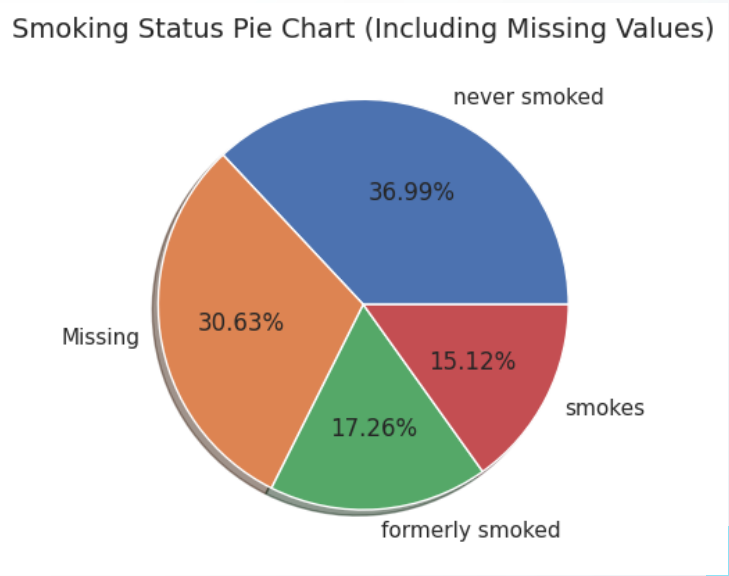
Bảng 5: Chia Dataset

Data	Kích thước
Train	75%
Test	25%

Pre-Proccesing

Xử lý Feature "smoking_status"

- Loại bỏ hoàn toàn thuộc tính “smoking_status”
- Thay thế giá trị “Unknown” bằng giá trị ngẫu nhiên dựa theo phân phối thực tế
- Thay thế giá trị “Unknown” bằng giá trị ngẫu nhiên nhưng KHÔNG giữ phân phối
- Điền giá trị phổ biến nhất (Mode Imputation)



Hình 25: Tương quan dữ liệu ở feature “smoking_status”

Pre-Proccesing

Xử lý Feature "Gender"

Bỏ các sample có giá trị “Other” trong cột “gender”

Bảng 6: Lable feature “gender”

gender	Dạng số
Male	**1**
Female	**0**

Xử lý Feature "Ever_married:"

Bảng 7: Label encoder “ever_married”

Ever_maried	Dạng số
No	**0**
Yes	**1**

Pre-Proccesing

Xử lý Feature "Work type"

Do công việc children cũng là never_worked nên có thể chuyển thành Never_Worked

Bảng 8: One hot "work_type"

gender	Dạng số			
work_type_Never_worked	1	0	0	0
work_type_Private	0	1	0	0
work_type_Self-employed	0	0	1	0
Govt_job	0	0	0	1

Xử lý Feature "Residence_type"

Bảng 9: One hot "Residence_type"

Residence_type	Dạng số
Rural	0
Urban	1

Pre-Proccesing

Xử lý Feature "Smoking_status"

Bảng 10: One hot “work_type”

Smoking_status	Dạng số		
never smoked	1	0	0
smokes	0	1	0
formerly smoked	0	0	1

Các Feature còn lại (“avg_glucose_level”, “bmi”, “age”):

Dùng MinmaxScaler

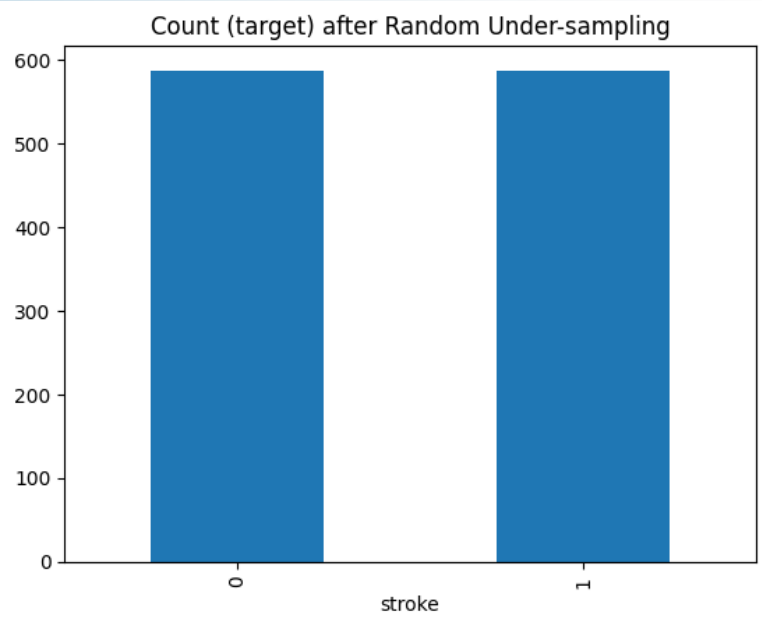
$$x' = \frac{(x - x\_min)}{(x\_max - x\_min)}$$

Pre-Proccesing

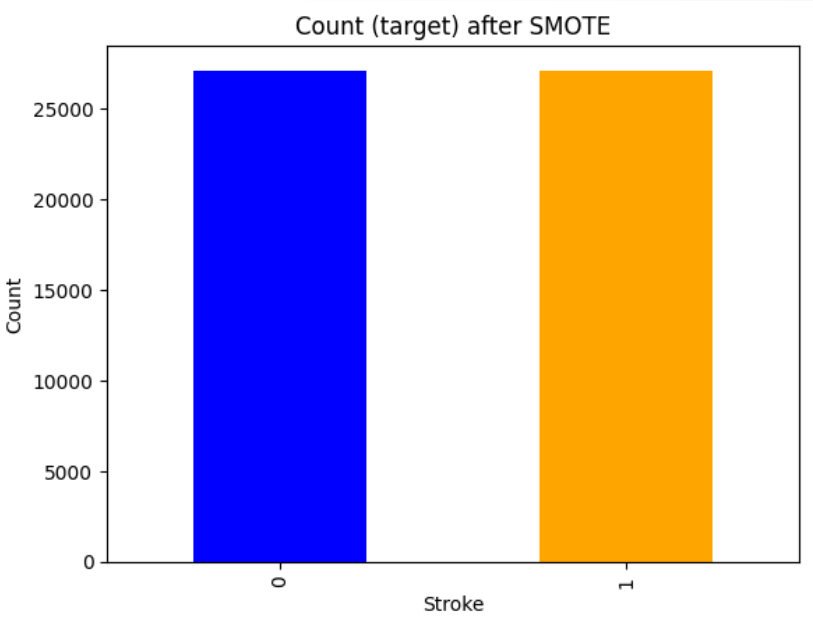
Bảng 11: One hot “work_type”

gender	age	hypertension	heart_disease	ever_married	Residence_type	avg_glucose_level	bmi	stroke	Never_worked	Private	Self-employed	never smoked	smokes
0	0.021973	0	0	0	1	0.214128	0.146237	0	1	0	0	1	0
0	0.279785	0	0	0	0	0.171911	0.150538	0	0	1	0	0	1
0	0.426270	0	0	1	0	0.384654	0.615054	0	0	1	0	1	0

Handel Sampling

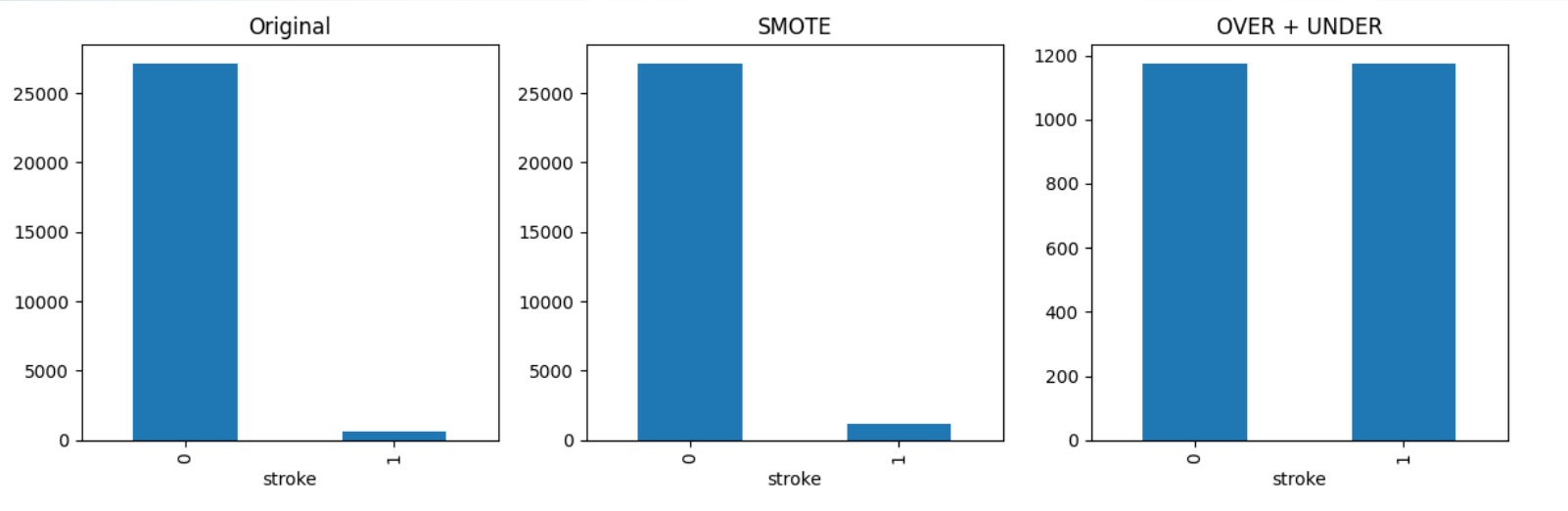


Hình 26: One hot “work_type”



Hình 27: One hot “work_type”

Handel Sampling



Hình 28: One hot “work_type”

04

Result

Kết quả - Riêng

Phương pháp 1:

Bảng 12: Kết quả áp dụng Phương pháp 1

Mô hình	Sampleing method	Accuracy	Precision		Recall		F1	
			0	1	0	1	0	1
Logistic Regression	Under	0.7782	0.9954	0.0895	0.7761	0.8599	0.8722	0.1622
Logistic Regression	Over	0.7790	0.9948	0.0882	0.7774	0.8408	0.8728	0.1596
SVM	Under	0.7631	0.9957	0.0853	0.7603	0.8726	0.8622	0.1553
SVM	Over	0.7695	0.9947	0.0848	0.7677	0.8408	0.8666	0.1540
Random Forest	Under	0.7881	0.9957	0.0939	0.7861	0.8662	0.8785	0.1695
Random Forest	Over	0.9062	0.9891	0.1525	0.9139	0.6051	0.9500	0.2436

Kết quả - Riêng

Phương pháp 2:

Bảng 13: Kết quả áp dụng Phương pháp 2

Mô hình	Sampleing method	Accuracy	Precision		Recall		F1	
			0	1	0	1	0	1
Logistic Regression	Under	0.7869	0.9965	0.0809	0.7850	0.8724	0.8782	0.1481
Logistic Regression	Over	0.7825	0.9966	0.0798	0.7805	0.8776	0.8754	0.1463
SVM	Under	0.7682	0.9973	0.0770	0.7653	0.9031	0.8660	0.1419
SVM	Over	0.7754	0.9964	0.0770	0.7733	0.8724	0.8708	0.1416
Random Forest	Under	0.7953	0.9971	0.0856	0.7932	0.8929	0.8835	0.1562
Random Forest	Over	0.8885	0.9912	0.1146	0.8940	0.6327	0.9401	0.1941

Kết quả - Riêng

Phương pháp 3:

Bảng 14: Kết quả áp dụng Phương pháp 3

Mô hình	Sampleing method	Accuracy	Precision		Recall		F1	
			0	1	0	1	0	1
Logistic Regression	Under	0.7862	0.9968	0.0815	0.7841	0.8827	0.8778	0.1491
Logistic Regression	Over	0.7907	0.9965	0.0823	0.7889	0.8724	0.8806	0.1503
SVM	Under	0.7648	0.9972	0.0760	0.7618	0.9031	0.8638	0.1401
SVM	Over	0.7847	0.9963	0.0797	0.7829	0.8673	0.8768	0.1460
Random Forest	Under	0.7856	0.9970	0.0820	0.7832	0.8929	0.8773	0.1502
Random Forest	Over	0.9064	0.9902	0.1272	0.9135	0.5816	0.9503	0.2088

Kết quả - Riêng

Phương pháp 4:

Bảng 15: Kết quả áp dụng Phương pháp 4

Mô hình	Sampleing method	Accuracy	Precision		Recall		F1	
			0	1	0	1	0	1
Logistic Regression	Under	0.7884	0.9968	0.0822	0.7863	0.8827	0.8791	0.1504
Logistic Regression	Over	0.7911	0.9962	0.0816	0.7896	0.8622	0.8809	0.1491
SVM	Under	0.7685	0.9971	0.0767	0.7657	0.8980	0.8662	0.1414
SVM	Over	0.7878	0.9962	0.0804	0.7862	0.8622	0.8789	0.1471
Random Forest	Under	0.7958	0.9965	0.0842	0.7941	0.8724	0.8839	0.1535
Random Forest	Over	0.9105	0.9904	0.1346	0.9175	0.5918	0.9526	0.2193

Kết quả - Kết hợp

Bảng 16: Kết quả áp dụng Phương pháp 1 - 2

Mô hình	Phương pháp	Accuracy	Precision		Recall		F1	
			0	1	0	1	0	1
Logistic Regression	1	0.7870	0.9933	0.0900	0.7867	0.8000	0.8780	0.1617
Logistic Regression	2	0.8086	0.9954	0.0855	0.8083	0.8265	0.8921	0.1549
SVM	1	0.7761	0.9937	0.0870	0.7752	0.8125	0.8709	0.1571
SVM	2	0.7967	0.9952	0.0804	0.7962	0.8214	0.8846	0.1464
Random Forest	1	0.8164	0.9932	0.1019	0.8172	0.7875	0.8966	0.1805
Random Forest	2	0.8282	0.9950	0.0926	0.8287	0.8061	0.9043	0.1661

Kết quả - Kết hợp

Bảng 17: Kết quả áp dụng Phương pháp 3 - 4

Mô hình	Phương pháp	Accuracy	Precision		Recall		F1	
			0	1	0	1	0	1
Logistic Regression	3	0.8078	0.9956	0.0860	0.8071	0.8367	0.8915	0.1560
Logistic Regression	4	0.8088	0.9956	0.0864	0.8081	0.8367	0.8921	0.1566
SVM	3	0.7948	0.9950	0.0792	0.7943	0.8163	0.8834	0.1445
SVM	4	0.7950	0.9954	0.0806	0.7942	0.8316	0.8835	0.1469
Random Forest	3	0.8250	0.9952	0.0920	0.8252	0.8163	0.9023	0.1653
Random Forest	4	0.8238	0.9949	0.0904	0.8242	0.8061	0.9015	0.1626

Kết quả

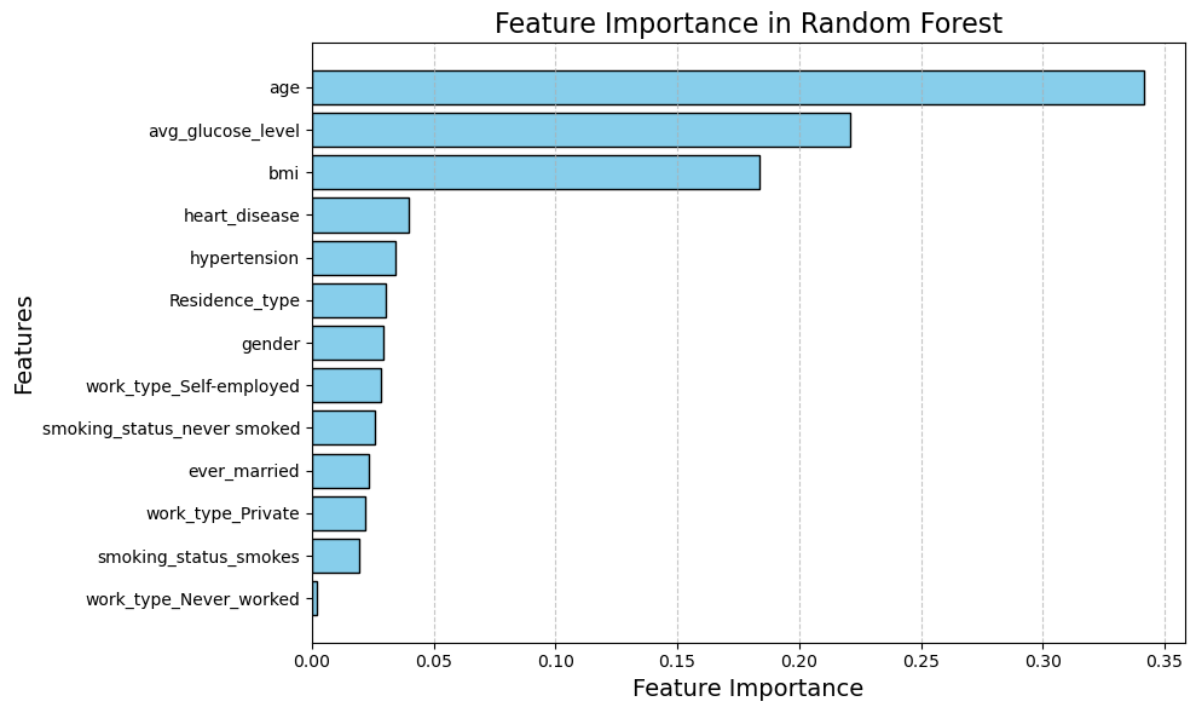
Phương pháp 2 tốt nhất:

Bảng 18: Kết quả tốt nhất

Mô hình	Sampleing method	Accuracy	Precision		Recall		F1	
			0	1	0	1	0	1
Logistic Regression	Under	0.7869	0.9965	0.0809	0.7850	0.8724	0.8782	0.1481
Logistic Regression	Over	0.7825	0.9966	0.0798	0.7805	0.8776	0.8754	0.1463
SVM	Under	0.7682	0.9973	0.0770	0.7653	0.9031	0.8660	0.1419
SVM	Over	0.7754	0.9964	0.0770	0.7733	0.8724	0.8708	0.1416
Random Forest	Under	0.7953	0.9971	0.0856	0.7932	0.8929	0.8835	0.1562
Random Forest	Over	0.8885	0.9912	0.1146	0.8940	0.6327	0.9401	0.1941

Kết quả

Feature Importance



Hình 29: Feature Importance

Tài liệu tham khảo

[1] Kiến thức Y khoa | Đột Quỵ Não Và Dấu Hiệu Nhận Biết Sớm Đột Quỵ Não

[2] Nguyên Nhân Khiến Tỷ Lệ Đột Quỵ Ngày Càng Trẻ Hóa?

[3] Logistic Regression là gì? Ví dụ bài toán Logistic Regression in Python

[4] What is a Support Vector Machine? - Datatron

[5] Random Forest - Regression and Classification - Explained using Sklearn - Python | I N F O A R Y A N

[6] Tìm hiểu về Confusion matrix trong Machine Learning? - Viblo

[7] FAKE AND REAL

[8] What Is Undersampling? | Master's in Data Science

Thanks

Do you have any questions?
hieuleduonghy@gmail.com