



ỨNG DỤNG HỒI QUY TRONG PHÂN TÍCH DỮ LIỆU: NGHIÊN CỨU TỪ 4 BỘ MẪU

By Nhóm 1



Thành viên

Quỳnh Anh

Viết báo cáo
Code
Làm slide

Nguyễn Vũ

Viết báo cáo
Code
Làm slide

Gia Vĩ

Viết báo cáo
Code
Làm slide

Trường Phát

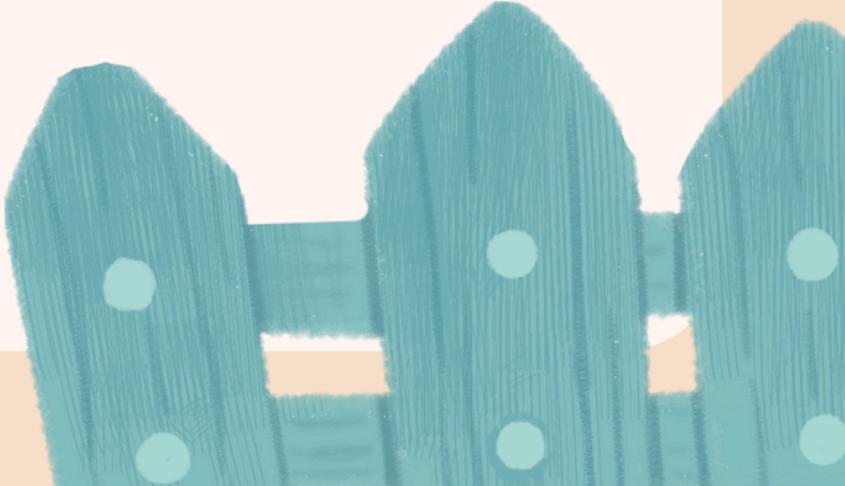
Viết báo cáo
Code
Làm slide



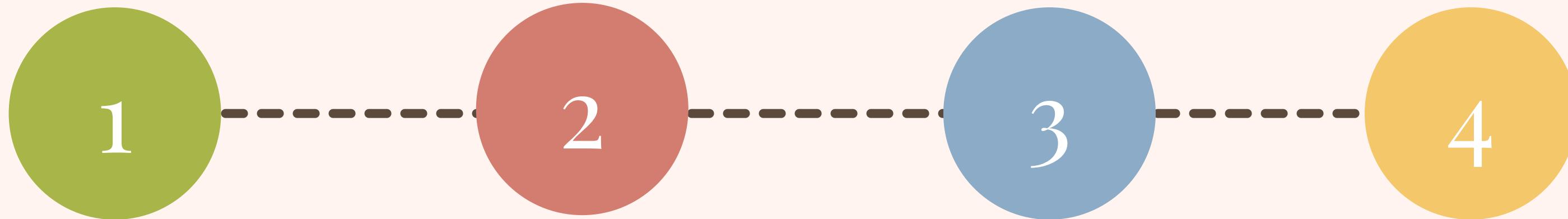


Giới thiệu

Đề tài "Ứng dụng hồi quy trong phân tích dữ liệu: Nghiên cứu từ 4 bộ mẫu" tập trung vào việc áp dụng các phương pháp hồi quy để phân tích dữ liệu từ nhiều lĩnh vực khác nhau. Mục tiêu là khám phá mối quan hệ giữa các biến, xây dựng mô hình dự đoán và đưa ra những kết luận hữu ích từ dữ liệu.



Nội dung



BỘ DỮ LIỆU 1

Stroke Prediction Dataset
Hồi quy đa biến sử dụng
Logistic regression.

BỘ DỮ LIỆU 2

MAGIC Gamma Telescope
Hồi quy đa biến sử dụng
Logistic regression.

BỘ DỮ LIỆU 3

Airline Passenger Satisfaction
Hồi quy thành phần chính sử
dụng Logistic regression.

BỘ DỮ LIỆU 4

Air Quality Index
Hồi quy phân tích thành
phân chính sử dụng Logistic
regression.

BỘ DỮ LIỆU 1

Stroke Prediction Dataset

Biến	Mô tả	Kiểu dữ liệu	Giá trị/Loại dữ liệu
Id	Mã định danh duy nhất của mỗi bệnh nhân.	Số nguyên (Integer)	Duy nhất, không lặp lại.
Gender	Giới tính của bệnh nhân.	Chuỗi (Categorical)	"Nam", "Nữ".
Age	Tuổi của bệnh nhân.	Số thực (Float)	Giá trị từ 0 trở lên.
hypertension	Tình trạng tăng huyết áp.	Số nhị phân (Binary)	0: Không bị tăng huyết áp, 1: Bị tăng huyết áp.
heart_disease	Tình trạng bệnh lý tim mạch.	Số nhị phân (Binary)	0: Không có bệnh lý tim mạch, 1: Có bệnh lý tim mạch.
ever_married	Tình trạng hôn nhân của bệnh nhân.	Chuỗi (Categorical)	"Không", "Có".
work_type	Loại công việc của bệnh nhân.	Chuỗi (Categorical)	"Trẻ em", "Công chức", "Chưa từng làm việc", "Tư nhân", "Tự làm chủ".

Nguồn gốc: Kaggle
Bộ dữ liệu gồm 5110 mẫu và 12 đặc trưng.

Residence_type	Loại nơi ở của bệnh nhân.	Chuỗi (Categorical)	"Nông thôn", "Thành thị".
avg_glucose_level	Mức đường huyết trung bình trong máu.	Số thực (Float)	Giá trị ≥ 0 (mmol/L).
bmi	Chỉ số khối cơ thể (Body Mass Index).	Số thực (Float)	Giá trị từ 0 trở lên.
smoking_status	Tình trạng hút thuốc của bệnh nhân.	Chuỗi (Categorical)	"Đã từng hút thuốc", "Chưa từng hút thuốc", "Đang hút thuốc", "Không rõ".
stroke	Tình trạng bị đột quỵ.	Số nhị phân (Binary)	0: Không bị đột quỵ, 1: Đã bị đột quỵ.

<https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>

Stroke Prediction Dataset

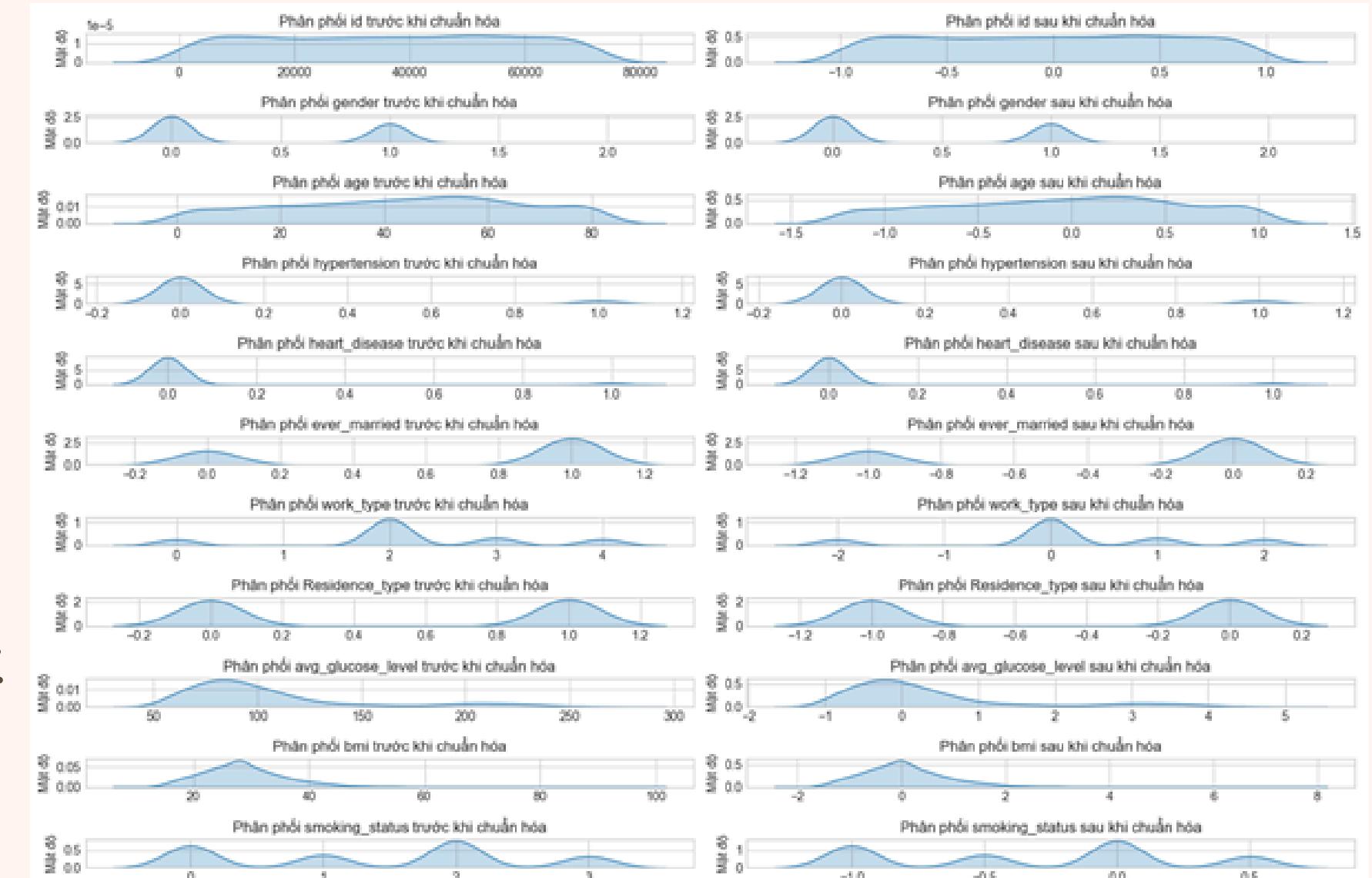


Phân bố ban đầu của các biến và trung bình tổng của từng biến theo biến phụ thuộc stroke.

Stroke Prediction Dataset

Tiền xử lý dữ liệu

- Giá trị thiếu: biến BMI - 0.05% tổng thể
=> Điền bằng median
- Mã hoá các biến object đưa về dạng số: LabelEncoder
- Chuẩn hoá dữ liệu đưa về cùng một thang đo.



Stroke Prediction Dataset

	feature	VIF
0	x1	1.480447
1	x2	2.135948
2	x3	1.209098
3	x4	1.171113
4	x5	2.412230
5	x6	1.311718
6	x7	1.497101
7	x8	1.197649
8	x9	1.210303
9	x10	1.455448

Kiểm tra hiện tượng đa cộng tuyến: Sử dụng chỉ số VIF để kiểm tra kết quả cho thấy không có hiện tượng này xảy ra trong bộ dữ liệu. Tất cả chỉ số đều dưới 5.

```
def calculate_vif(df):
    # trích xuất các giá trị của dataframe
    x = df.values
    # tạo một dataframe rỗng để lưu kq vif
    vif_data = pd.DataFrame()
    vif_data["feature"] = df.columns
    vif_data["VIF"] = [variance_inflation_factor(x, i) for i in range(x.shape[1])]
    return vif_data
```

Stroke Prediction Dataset

Tiến hành đưa bộ dữ liệu vào mô hình Logistic Regression ta thu được bảng kết quả sau:

Mô hình mô tả được 20.14% phuơng sai so với biến phụ thuộc.

Một số thuộc tính không có ý nghĩa thống kê P_value > 0.05 => loại bỏ.

Optimization terminated successfully. Current function value: 0.155535 Iterations 9							
Logit Regression Results							
Dep. Variable:	stroke	No. Observations:	5110 <th>Model:</th> <td>Logit</td> <th>Df Residuals:</th> <td>5099</td>	Model:	Logit	Df Residuals:	5099
Method:	MLE	Df Model:	10	Date:	Sun, 05 Jan 2025	Pseudo R-squ.:	0.2014
Time:	14:13:21	Log-Likelihood:	-794.79	converged:	True	LL-Null:	-995.19
Covariance Type:	nonrobust	LLR p-value:	6.358e-30				
	coef	std err	z	P> z	[0.025	0.975]	
const	-4.0279	0.172	-23.371	0.000	-4.366	-3.690	
x1	0.0510	0.140	0.364	0.716	-0.224	0.326	
x2	2.5227	0.191	13.178	0.000	2.147	2.898	
x3	0.3896	0.164	2.374	0.018	0.068	0.711	
x4	0.3203	0.190	1.688	0.091	-0.052	0.692	
x5	-0.1889	0.219	-0.862	0.389	-0.619	0.241	
x6	-0.0531	0.072	-0.735	0.462	-0.195	0.089	
x7	0.0990	0.138	0.719	0.472	-0.171	0.369	
x8	0.1523	0.044	3.466	0.001	0.066	0.238	
x9	-0.0059	0.101	-0.058	0.954	-0.203	0.191	
x10	0.0009	0.144	0.006	0.995	-0.281	0.283	

Stroke Prediction Dataset

Sau khi loại bỏ các thuộc tính không có ý nghĩa ta thu được kết quả sau.

Dữ liệu quan sát 5110, Pseudo R-squared = 0.1989, giải thích 19,89% biến động khả năng đột quy.

Các biến X₂, X₃, X₈ đều P-value < 0.05, có ý nghĩa thống kê.

Optimization terminated successfully. Current function value: 0.156910 Iterations 9							
Logit Regression Results							
Dep. Variable:	stroke	No. Observations:	5110	Model:	Logit	Df Residuals:	5106
Method:	MLE	Df Model:	3	Date:	Tue, 07 Jan 2025	Pseudo R-squ.:	0.1989
Time:	18:44:08	Log-Likelihood:	-797.21	converged:	True	LL-Null:	-995.19
Covariance Type:	nonrobust	LLR p-value:	1.672e-85				
	coef	std err	z	P> z	[0.025	0.975]	
const	-4.0025	0.132	-30.319	0.000	-4.261	-3.744	
x2	2.5411	0.182	13.944	0.000	2.184	2.898	
x3	0.3845	0.162	2.368	0.018	0.066	0.703	
x8	0.1604	0.042	3.779	0.000	0.077	0.244	

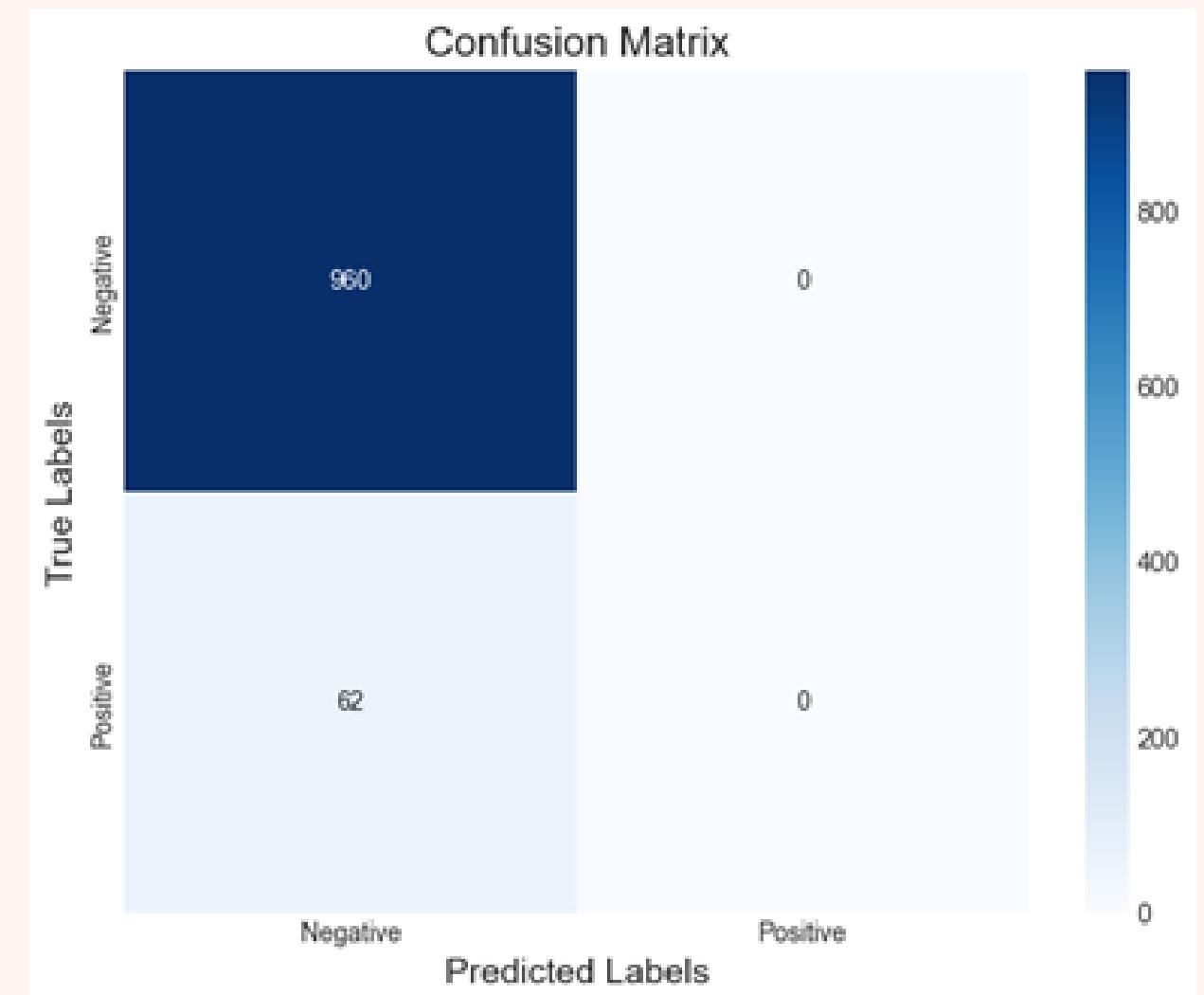
Stroke Prediction Dataset

Accuracy: 93.93%, nhưng mô hình chỉ dự đoán tốt lớp 0 (không đột quy), hoàn toàn thất bại với lớp 1 (đột quy).

Precision, recall, F1-score của lớp 1: Bằng 0.

Mất cân bằng dữ liệu: Lớp 0 chiếm 95.1%, lớp 1 chỉ 4.9%, gây thiên lệch mạnh về lớp 0.

```
He so chan Intercept: [-3.99222363]
He so hoi quy voi tung dac trung Coefficients: [[2.42542491 0.35243037 0.15033073]]
Accuracy: 0.9393346379647749
Classification Report:
precision    recall    f1-score   support
          0       0.94      1.00      0.97      960
          1       0.00      0.00      0.00       62
accuracy                           0.94      1022
macro avg       0.47      0.50      0.48      1022
weighted avg    0.88      0.94      0.91      1022
```



Stroke Prediction Dataset

KẾT LUẬN:

Sự mất cân bằng dữ liệu nghiêm trọng trong biến stroke, với lớp 0 chiếm ưu thế, là nguyên nhân khiến mô hình dự đoán chủ yếu lớp 0 và không nhận diện được lớp 1. Mặc dù độ chính xác đạt 94%, nhưng mô hình không hiệu quả trong việc dự đoán đột quỵ. Độ chính xác không phải là chỉ số duy nhất đánh giá mô hình, đặc biệt với dữ liệu mất cân bằng.

ĐỀ XUẤT:

Cần xử lý mất cân bằng dữ liệu bằng oversampling lớp 1 (SMOTE) hoặc undersampling lớp 0. Sử dụng thuật toán hỗ trợ dữ liệu mất cân bằng và đánh giá mô hình bằng F1-score, recall lớp 1 và ROC AUC để có cái nhìn toàn diện.

BỘ DỮ LIỆU 2

MAGIC Gamma Telescope

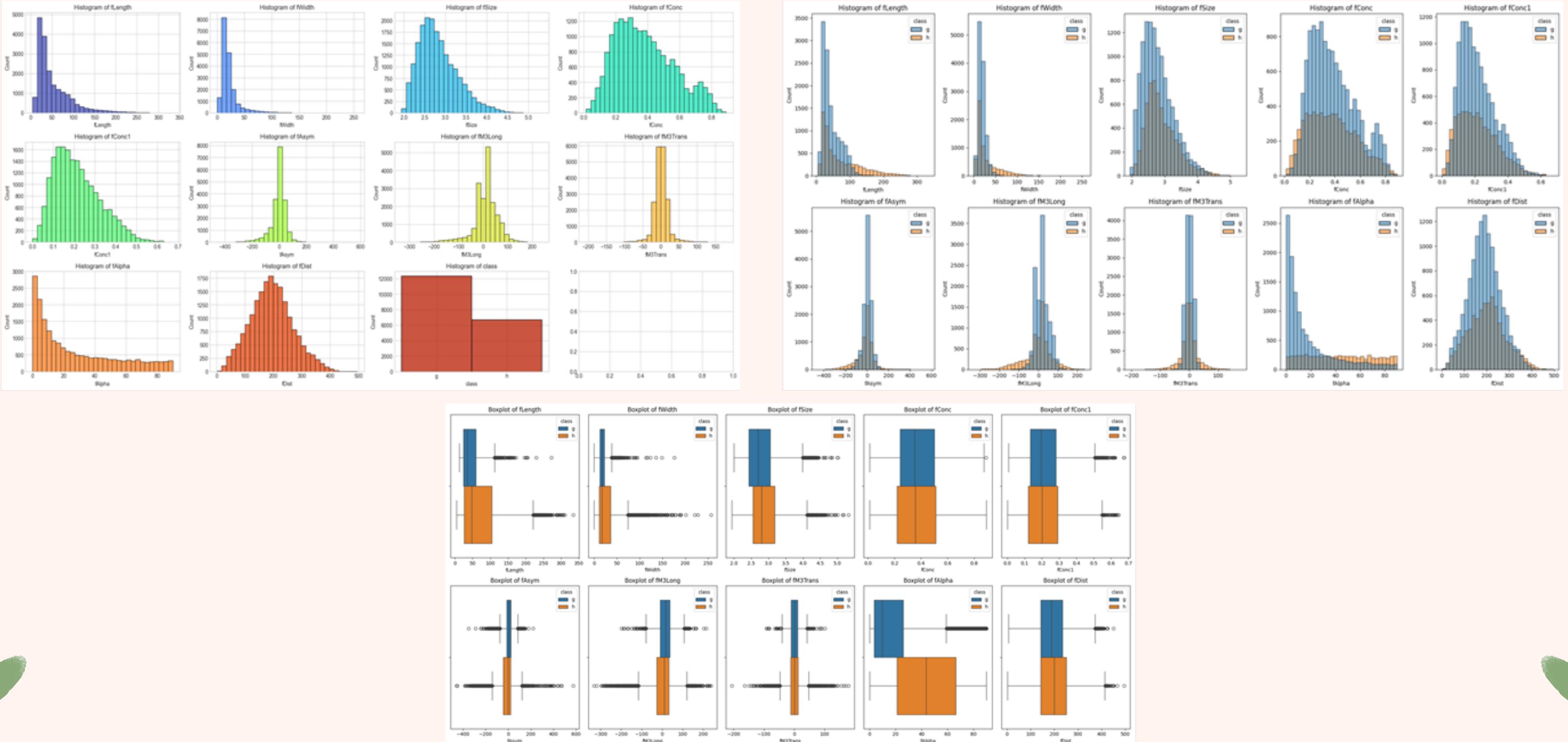
Biến	Mô tả	Kiểu dữ liệu	Giá trị/Loại dữ liệu
fLength	Trục lớn của hình ellipse (mm), đo độ dài của đối tượng theo chiều dài nhất.	Số thực (Float)	Giá trị dương (mm).
fWidth	Trục nhỏ của hình ellipse (mm), đo chiều rộng của đối tượng theo chiều ngắn nhất.	Số thực (Float)	Giá trị dương (mm).
fSize	Logarithm cơ số 10 của tổng giá trị ánh sáng trên toàn bộ các pixel (số photon).	Số thực (Float)	Giá trị dương, logarit của tổng giá trị ánh sáng.
fConc	Tỷ lệ của tổng hai pixel sáng nhất trên tổng giá trị ánh sáng (fSize).	Số thực (Float)	Giá trị từ 0 đến 1, tỷ lệ ánh sáng tập trung tại khu vực sáng nhất.
fConc1	Tỷ lệ của pixel sáng nhất trên tổng giá trị ánh sáng (fSize).	Số thực (Float)	Giá trị từ 0 đến 1, chi tiết hóa mức độ tập trung tại pixel sáng nhất.
fAsym	Khoảng cách từ pixel sáng nhất đến tâm hình ellipse, chiều lên trục lớn (mm).	Số thực (Float)	Giá trị dương, đo sự bất đối xứng trong phân bố ánh sáng.

Nguồn gốc: UCI Machine Learning Repository
 Bộ dữ liệu gồm 19.020 mẫu và 11 đặc trưng.

fM3Long	Căn bậc ba của moment bậc ba dọc theo trục lớn (mm).	Số thực (Float)	Giá trị dương (mm), đo sự lệch ánh sáng theo chiều dài.
fM3Trans	Căn bậc ba của moment bậc ba dọc theo trục nhỏ (mm).	Số thực (Float)	Giá trị dương (mm), đo sự lệch ánh sáng theo hướng ngang.
fAlpha	Góc giữa trục lớn của hình ellipse và vector hướng đến gốc tọa độ (độ).	Số thực (Float)	Giá trị từ 0 đến 360 độ, đo hướng của đối tượng trong không gian.
fDist	Khoảng cách từ gốc tọa độ đến tâm của hình ellipse (mm).	Số thực (Float)	Giá trị dương (mm), đo vị trí tổng thể của đối tượng.
class	Nhãn phân loại của đối tượng: "g" (gamma tín hiệu), "h" (hadron nhiễu nền).	Chuỗi (Categorical)	"g" hoặc "h"

<https://archive.ics.uci.edu/dataset/159/magic+gamma+telescope>

MAGIC Gamma Telescope

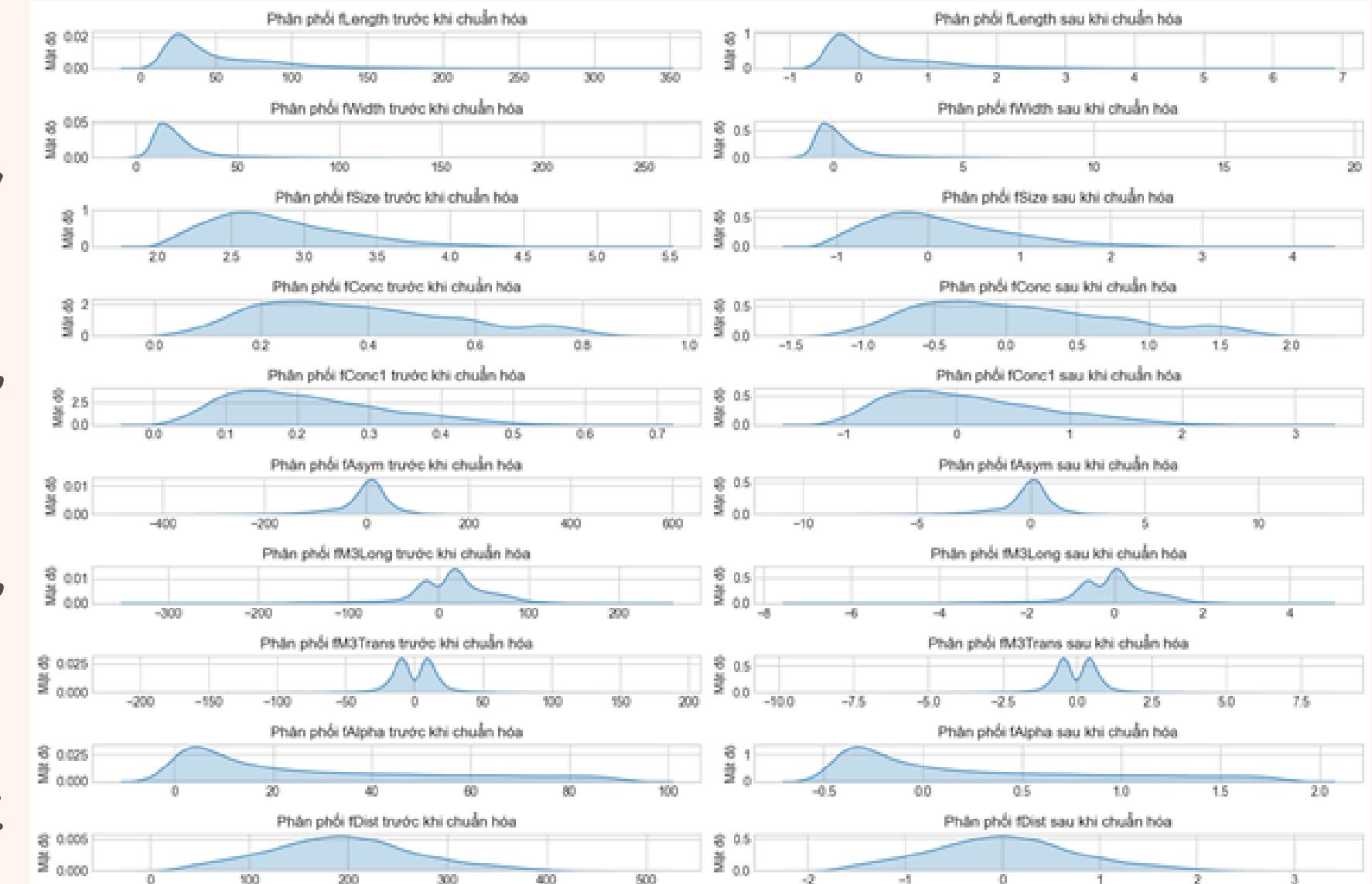


Phân bố ban đầu của các biến và trung bình tổng của từng biến theo biến phụ thuộc class.

MAGIC Gamma Telescope

Tiền xử lý dữ liệu

- Giá trị outliers: giữ lại tất cả vì có ý nghĩa vật lý quan trọng và có thể đại diện cho các sự kiện năng lượng cao từ thiên văn.
- chuyển đổi giá trị của biến 'class' từ chuỗi sang số.
- Chuẩn hoá dữ liệu đưa về cùng một thang đo.



MAGIC Gamma Telescope

	feature	VIF
0	x1	3.361700
1	x2	3.489899
2	x3	5.082901
3	x4	28.056350
4	x5	22.969670
5	x6	1.284958
6	x7	1.247024
7	x8	1.002697
8	x9	1.341576
9	x10	1.356828

Kiểm tra hiện tượng đa cộng tuyến: Sử dụng chỉ số VIF để kiểm tra, kết quả cho thấy x4, x5 đang có chỉ số VIF > 10. Có hiện tượng đa cộng tuyến trong bộ dữ liệu. => loại bỏ đa cộng tuyến.

```
def calculate_vif(df):
    # trích xuất các giá trị của dataframe
    x = df.values
    # tạo một dataframe rỗng để lưu kq vif
    vif_data = pd.DataFrame()
    vif_data["feature"] = df.columns
    vif_data["VIF"] = [variance_inflation_factor(x, i) for i in range(x.shape[1])]
    return vif_data
```

MAGIC Gamma Telescope

Tiến hành loại bỏ đa cộng tuyến : x_4 có giá trị VIF cao nhất loại bỏ đầu tiên. Sau khi loại bỏ ta xem lại giá trị VIF cho các biến khác. lúc này các biến khác đều có giá trị VIF dưới 5.

=>Tuy nhiên để chính xác hơn cần phải xem xét thêm các đặc trưng có ý nghĩa thống kê hay không. ($P > |z| < 0.05$.

feature	VIF
0 x1	3.354834
1 x2	3.488598
2 x3	4.430091
3 x5	2.605865
4 x6	1.284953
5 x7	1.246189
6 x8	1.002685
7 x9	1.326874
8 x10	1.355750

MAGIC Gamma Telescope

Optimization terminated successfully.						
Current function value: 0.455592						
Iterations 7						
Logit Regression Results						
=====						
Dep. Variable:		class	No. Observations:		18905	
Model:		Logit	Df Residuals:		18895	
Method:		MLE	Df Model:		9	
Date:	Sat, 04 Jan 2025		Pseudo R-squ.:		0.2948	
Time:		13:54:53	Log-Likelihood:		-8613.0	
converged:		True	LL-Null:		-12213.	
Covariance Type:		nonrobust	LLR p-value:		0.000	
=====						
	coef	std err	z	P> z	[0.025	0.975]

const	-1.8101	0.028	-64.831	0.000	-1.865	-1.755
x1	1.3525	0.048	28.068	0.000	1.258	1.447
x2	0.0699	0.032	2.209	0.027	0.008	0.132
x3	0.3966	0.056	7.079	0.000	0.287	0.506
x5	0.8274	0.046	17.980	0.000	0.737	0.918
x6	0.0035	0.019	0.179	0.858	-0.034	0.041
x7	-0.3492	0.026	-13.440	0.000	-0.400	-0.298
x8	-0.0125	0.025	-0.500	0.617	-0.062	0.037
x9	1.8153	0.034	52.945	0.000	1.748	1.883
x10	0.0534	0.030	1.805	0.071	-0.005	0.111
=====						

Đưa bộ dữ liệu vào mô hình hồi quy và ta thu được bảng kết quả sau.

Sau khi xem xét chúng tôi phát hiện ra x6, x8, x10 đại diện cho fAsym, fM3Trans, fDist có $(P > |z|) > 0.05$ không có ý nghĩa thống kê nên loại bỏ.

MAGIC Gamma Telescope

Sau khi loại bỏ các thuộc tính không có ý nghĩa ta thu được kết quả sau.

Pseudo R-squared = 0.2946 mô hình giải thích được khoảng 29.46% sự biến thiên của biến mục tiêu.

Các biến $X_1, X_2, X_3, X_5, X_7, X_9$ đều P-value < 0.05, có ý nghĩa thống kê.

```
Optimization terminated successfully.  
    Current function value: 0.455685  
    Iterations 7
```

Logit Regression Results

```

Dep. Variable: class No. Observations: 18905
Model: Logit Df Residuals: 18898
Method: MLE Df Model: 6
Date: Sat, 04 Jan 2025 Pseudo R-squ.: 0.2946
Time: 13:54:53 Log-Likelihood: -8614.7
converged: True LL-Null: -12213.
Covariance Type: nonrobust LLR p-value: 0.000
=====

```

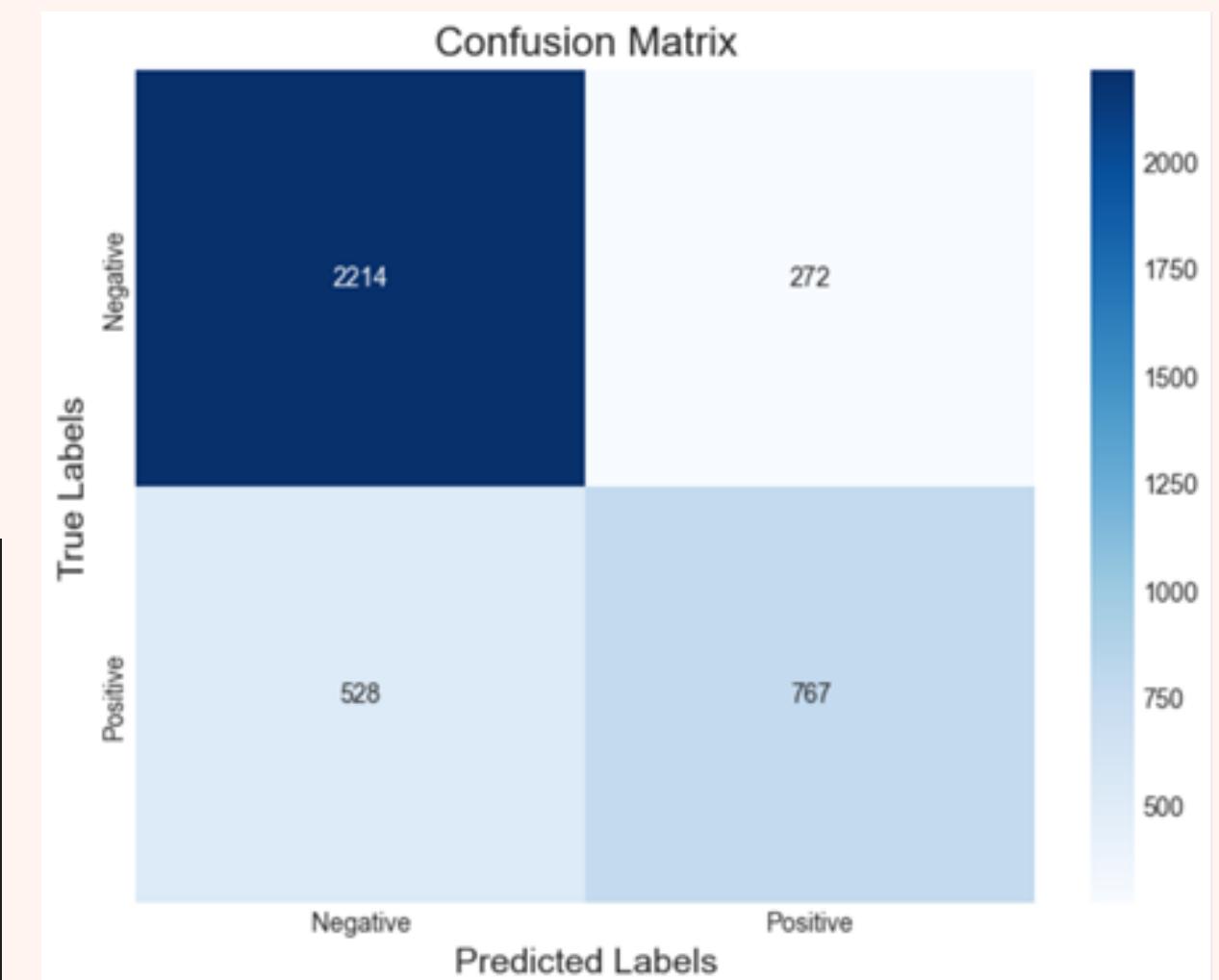
	coef	std err	z	P> z	[0.025	0.975]
const	-1.8147	0.028	-65.362	0.000	-1.869	-1.760
x1	1.3740	0.045	30.345	0.000	1.285	1.463
x2	0.0692	0.032	2.184	0.029	0.007	0.131
x3	0.4140	0.055	7.505	0.000	0.306	0.522
x5	0.8409	0.045	18.565	0.000	0.752	0.930
x7	-0.3509	0.025	-13.872	0.000	-0.400	-0.301
x9	1.8025	0.033	53.830	0.000	1.737	1.868

MAGIC Gamma Telescope

Mô hình đạt độ chính xác 78.84% và AUC-ROC 0.842, chứng tỏ khả năng phân biệt tốt giữa hai lớp. Lớp 0 có hiệu suất cao hơn với precision 81%, recall 89%, và F1-score 85%, trong khi lớp 1 đạt precision 74%, recall 59%, và F1-score 66%, cho thấy hạn chế trong nhận diện lớp 1. AUC trên 0.8 khẳng định hiệu suất phân loại khá tốt.

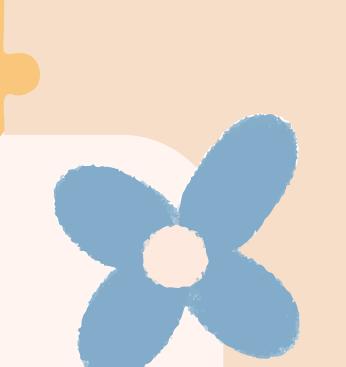
```
He so chan Intercept: [-1.79953135]
He so hoi quy ung voi tung dac trung Coefficients: [[ 1.38892417  0.05838024  0.38214339  0.82745869 -0.35185517  1.79998189]]
ROC AUC: 0.8421793784979545
False Positive Rate: [0.          0.          0.          ... 0.99879324 0.99879324 1.          ]
True Positive Rate: [0.00000000e+00 7.72200772e-04 1.12741313e-01 ... 9.99227799e-01
1.00000000e+00 1.00000000e+00]
Thresholds: [ inf 0.99996841 0.98388772 ... 0.0470001 0.046946  0.04252942]
Accuracy: 0.7884157638256546
Classification Report:
precision    recall   f1-score   support
0            0.81      0.89      0.85     2486
1            0.74      0.59      0.66     1295

accuracy           0.79      3781
macro avg       0.77      0.74      0.75     3781
weighted avg    0.78      0.79      0.78     3781
```

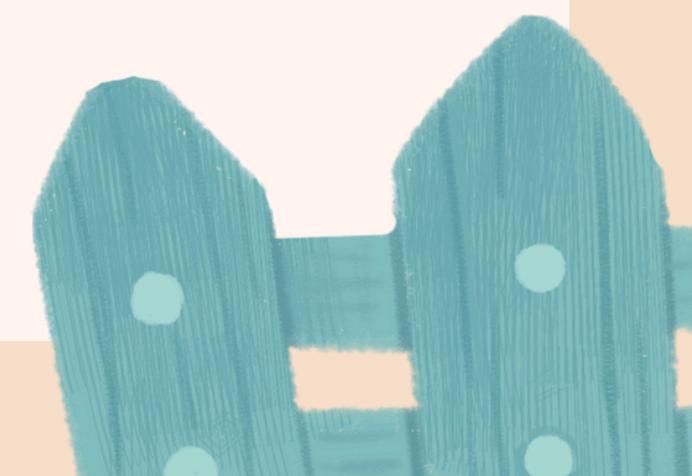




MAGIC Gamma Telescope



Mô hình hồi quy logistic đạt Pseudo R-squared 0.2946, AUC 0.84, và độ chính xác 78.84%, cho thấy hiệu suất khá tốt và khả năng giải thích ý nghĩa các biến. Tuy nhiên, recall thấp ở lớp 1 (59%) và hiệu suất không cân bằng giữa hai lớp cho thấy mô hình bỏ sót nhiều trường hợp thuộc lớp 1. Để cải thiện, có thể áp dụng các kỹ thuật xử lý mất cân bằng dữ liệu như SMOTE, hoặc thử nghiệm các mô hình phi tuyến như Random Forest, Gradient Boosting hay SVM để tăng độ chính xác, đặc biệt với lớp 1.



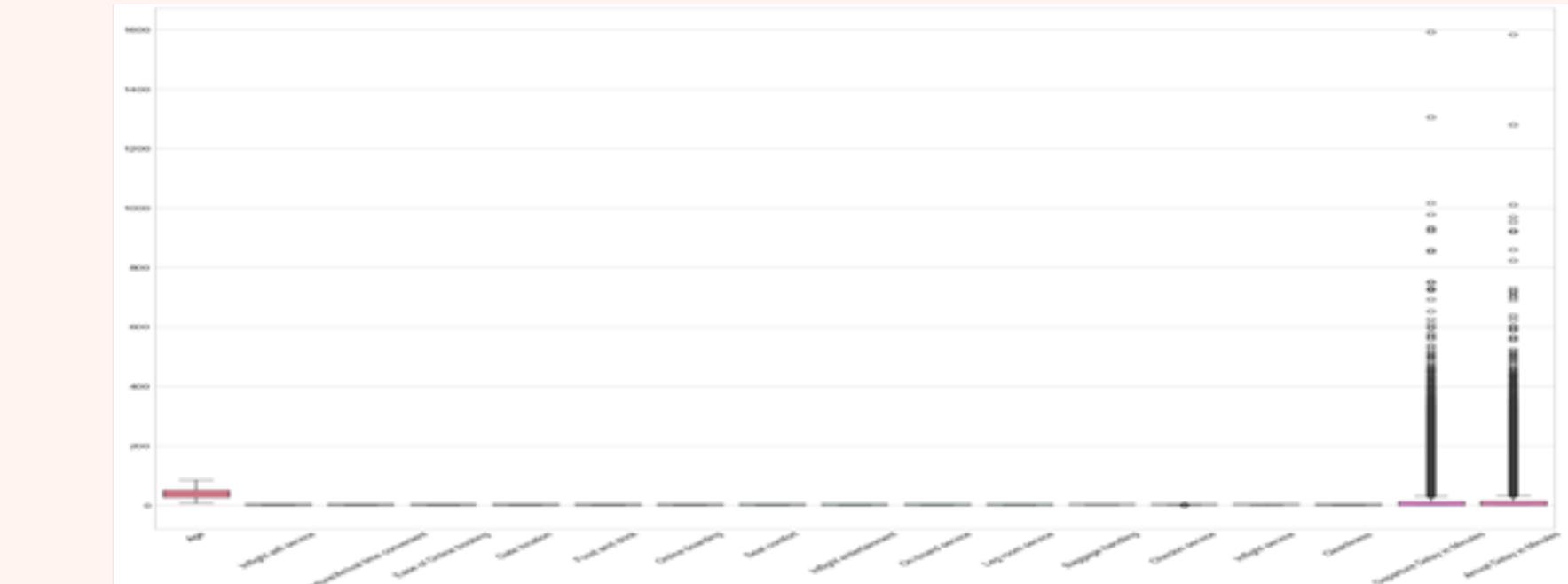
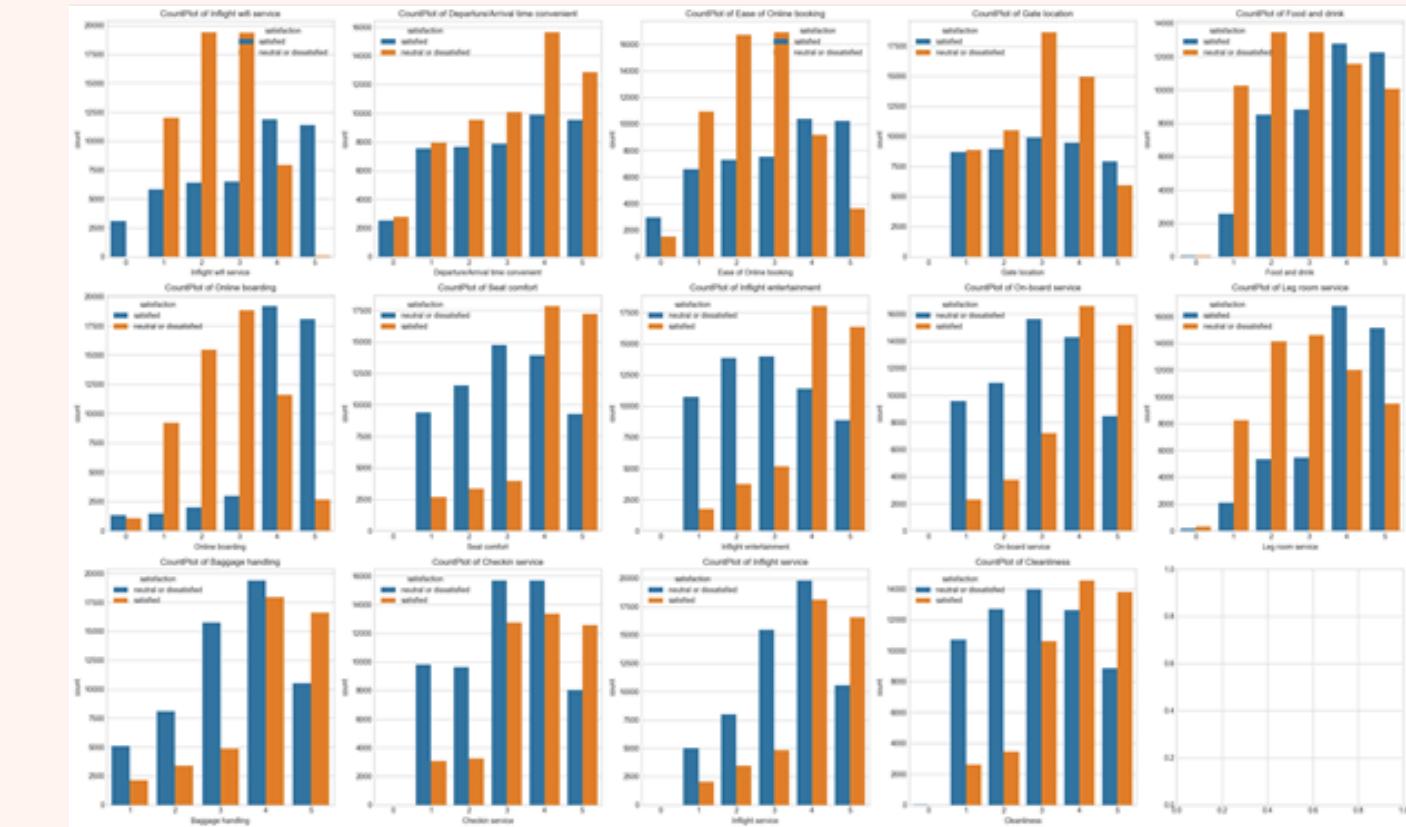
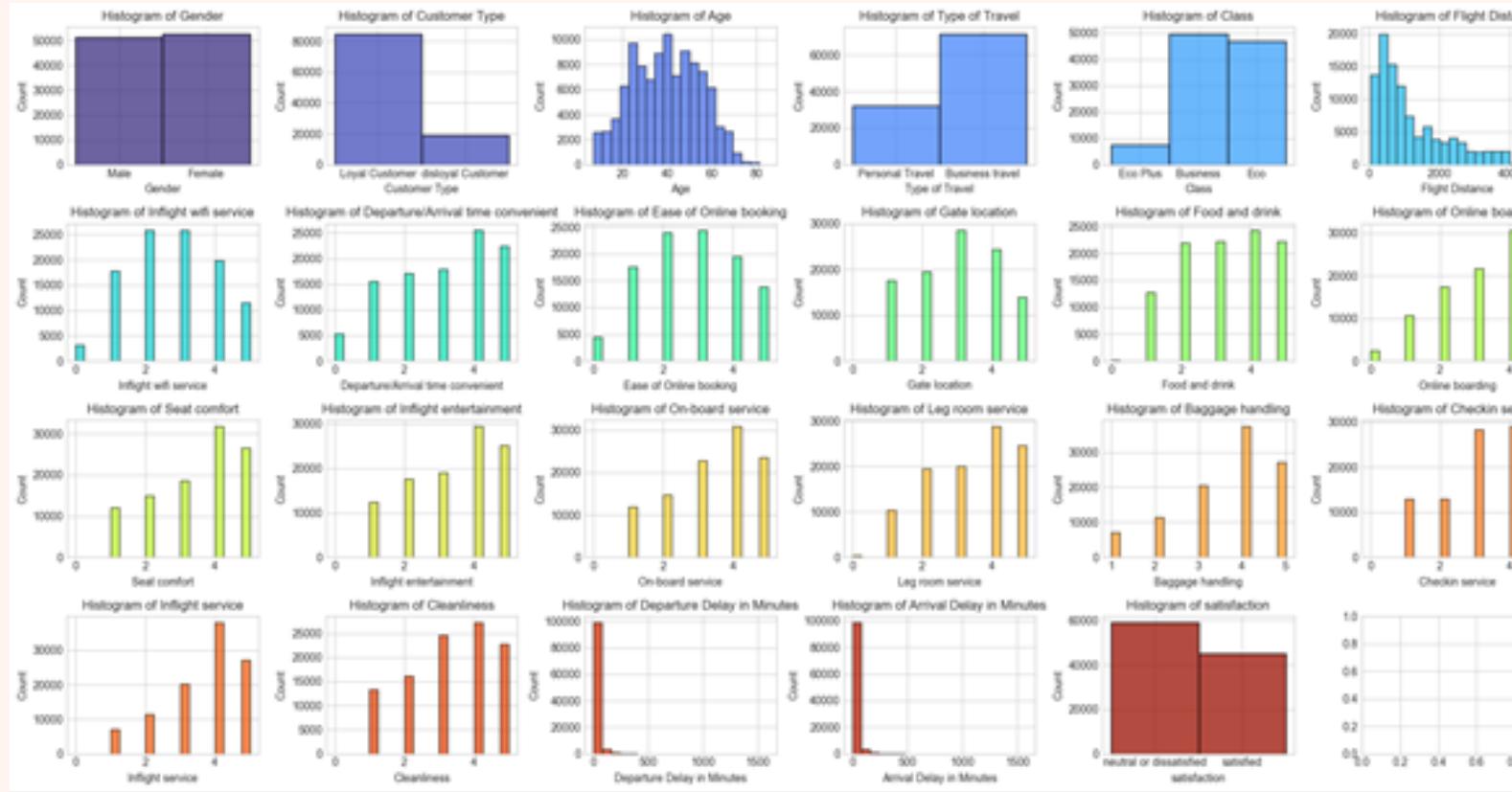
BỘ DỮ LIỆU 3

Airline Passenger Satisfaction

Tên đặc trưng	Tên tiếng Việt	Mô tả
id	Mã định danh	Mã định danh duy nhất cho mỗi hành khách.
Gender	Giới tính	Giới tính của hành khách (Nam hoặc Nữ).
Customer Type	Loại khách hàng	Phân loại khách hàng: Khách hàng trung thành hoặc không trung thành.
Age	Tuổi	Tuổi của hành khách.
Type of Travel	Loại chuyến đi	Mục đích chuyến đi: Cá nhân hay Công tác.
Class	Hạng ghế	Hạng ghế: Hạng Thương gia, Hạng Phổ thông đặc biệt, Hạng Phổ thông.
Flight Distance	Khoảng cách chuyến bay	Khoảng cách của chuyến bay (dặm).
Inflight wifi service	Dịch vụ wifi trên máy bay	Mức độ hài lòng với dịch vụ wifi trên máy bay (1-5).
Departure/Arrival time convenient	Thời gian khởi hành/đến thuận tiện	Mức độ hài lòng với sự thuận tiện của thời gian khởi hành/đến (1-5).
Ease of Online booking	Dễ dàng đặt vé trực tuyến	Mức độ hài lòng với việc đặt vé trực tuyến (1-5).
Gate location	Vị trí cổng lên máy bay	Mức độ hài lòng với vị trí cổng lên máy bay (1-5).
Food and drink	Thức ăn và đồ uống	Mức độ hài lòng với chất lượng thức ăn và đồ uống (1-5).
Online boarding	Lên máy bay trực tuyến	Mức độ hài lòng với quy trình lên máy bay trực tuyến (1-5).
Seat comfort	Sự thoải mái của ghế ngồi	Mức độ hài lòng với sự thoải mái của ghế ngồi (1-5).
Inflight entertainment	Giải trí trên máy bay	Mức độ hài lòng với các dịch vụ giải trí trên máy bay (1-5).
On-board service	Dịch vụ trên máy bay	Mức độ hài lòng với dịch vụ trên máy bay (1-5).
Leg room service	Dịch vụ chỗ để chân	Mức độ hài lòng với không gian để chân (1-5).
Baggage handling	Xử lý hành lý	Mức độ hài lòng với việc xử lý hành lý (1-5).
Check-in service	Dịch vụ làm thủ tục	Mức độ hài lòng với dịch vụ làm thủ tục (1-5).
Inflight service	Dịch vụ trong chuyến bay	Mức độ hài lòng với dịch vụ trong chuyến bay (1-5).
Cleanliness	Sự sạch sẽ	Mức độ hài lòng với sự sạch sẽ của máy bay (1-5).
Departure Delay in Minutes	Thời gian trễ khởi hành (phút)	Thời gian trễ khi khởi hành tính bằng phút.
Arrival Delay in Minutes	Thời gian trễ đến (phút)	Thời gian trễ khi đến tính bằng phút.
satisfaction	Sự hài lòng	Mức độ hài lòng tổng thể của hành khách (Hài lòng hoặc Không hài lòng).

Nguồn gốc: Kaggle
Bộ dữ liệu gồm 103904 mẫu và 25 đặc trưng.
<https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction>

Airline Passenger Satisfaction



Phân bố ban đầu của các biến và trung bình tổng của từng biến theo biến phụ thuộc satisfaction.

Airline Passenger Satisfaction

Tiền xử lý dữ liệu

- Giá trị thiếu: biến Arrival Delay in Minutes - 0.02 % tổng thể
=> Điền bằng mean
- Mã hoá các biến object đưa về dạng số: OrdinalEncoder
- Chuẩn hoá dữ liệu đưa về cùng một thang đo.



Airline Passenger Satisfaction

Vì bộ dữ liệu có quá nhiều đặc trưng và có mối tương quan cao nên chúng tôi quyết định sử dụng PCA phân tích thành phần chính để biến đổi dữ liệu về không gian có số chiều nhỏ hơn mà vẫn giữ được nhiều thông tin nhất có thể của bộ dữ liệu. ta thu được số thành phần chính gồm 8 thành phần

```
# Ti le phuong sai  
expalained_var = pca.explained_variance_ratio_ # khoan  
print(f'Ty le phuong sai giai thich {expalained_var}')
```

```
# chon tren phuong sai tich luy  
cumsum_explained_var = np.cumsum(pca.explained_variance_ratio_)  
print(f'Phuong sai tich luy: {cumsum_explained_var}')
```

```
# Lua chon thanh phan chinh ty le > 0.9  
n_components = np.argmax(cumsum_explained_var >=0.9) + 1  
print(f'So luong thanh phan chinh duoc chon : {n_components}')
```

So luong thanh phan chinh duoc chon : 8

Airline Passenger Satisfaction

Dữ liệu 93513 quan sát, Pseudo R-squared = 0.4261, giải thích 42.61% bộ dữ liệu.

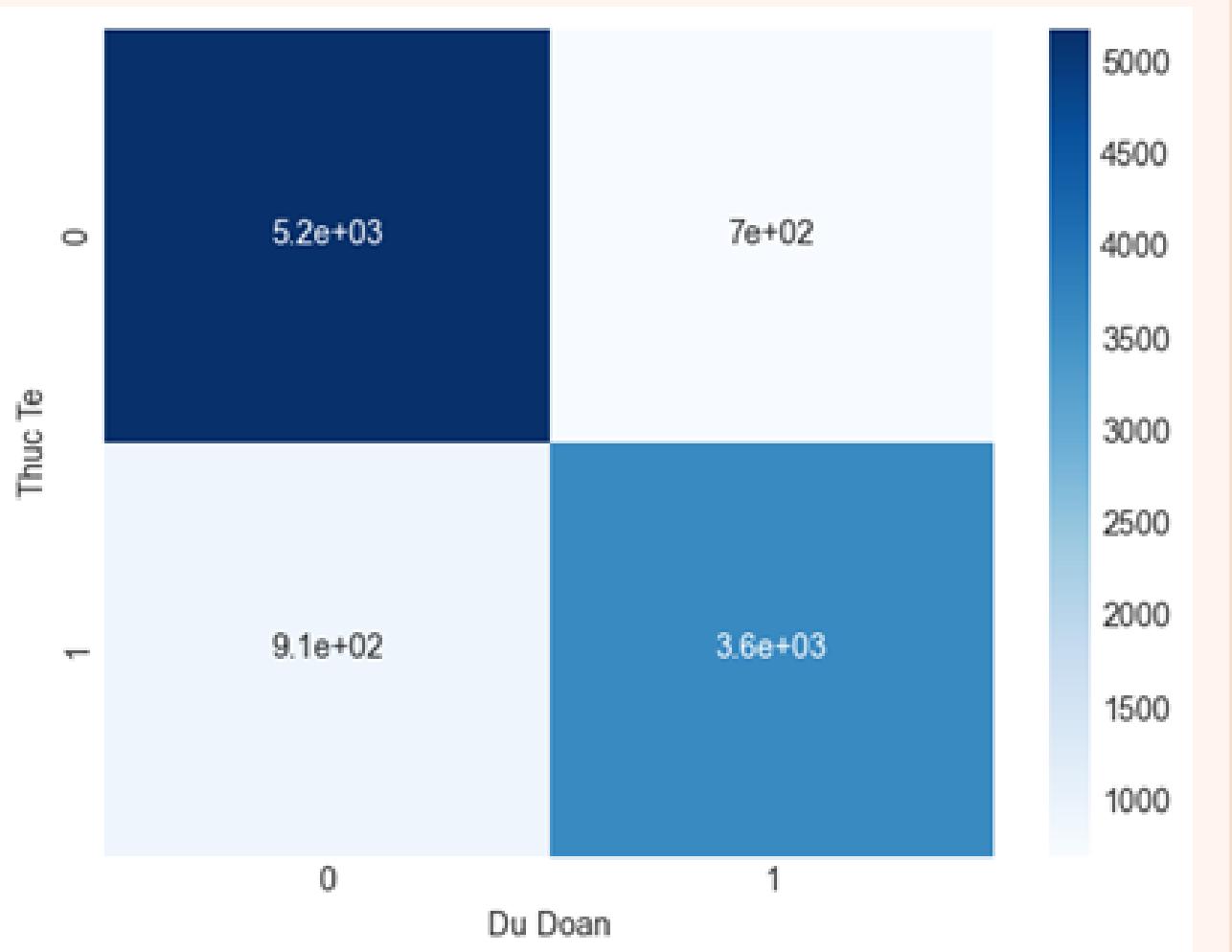
Hầu hết các biến độc lập có ý nghĩa thống kê ($p\text{-value} < 0.05$), khẳng định mối quan hệ đáng kể với sự hài lòng. Giá trị LLR p-value = 0.000 cho thấy mô hình tổng thể có ý nghĩa thống kê cao.

Optimization terminated successfully. Current function value: 0.392697 Iterations 7							
Logit Regression Results							
Dep. Variable:	satisfaction	No. Observations:	103904	Model:	Logit	Df Residuals:	103895
Method:	MLE	Df Model:	8	Date:	Tue, 07 Jan 2025	Pseudo R-squ.:	0.4261
Time:	19:01:34	Log-Likelihood:	-40803.	converged:	True	LL-Null:	-71094.
Covariance Type:	nonrobust	LLR p-value:	0.000				
	coef	std err	z	P> z	[0.025	0.975]	
const	-0.5092	0.009	-54.962	0.000	-0.527	-0.491	
x1	-0.0476	0.002	-22.204	0.000	-0.052	-0.043	
x2	1.1998	0.008	147.051	0.000	1.184	1.216	
x3	-0.7664	0.009	-89.137	0.000	-0.783	-0.750	
x4	-0.6643	0.009	-75.029	0.000	-0.682	-0.647	
x5	0.4239	0.010	42.806	0.000	0.404	0.443	
x6	1.0140	0.011	92.170	0.000	0.992	1.036	
x7	-0.7818	0.013	-58.346	0.000	-0.808	-0.756	
x8	0.2009	0.013	15.082	0.000	0.175	0.227	

Airline Passenger Satisfaction

Mô hình dự đoán đạt độ chính xác 84.58% với F1-score 87% (lớp 0) và 82% (lớp 1), phản ánh hiệu suất tốt và ổn định trên cả hai lớp. Dữ liệu cân bằng giúp giảm thiểu thiên lệch, nhưng mô hình hoạt động tốt hơn với lớp "không hài lòng," dự đoán đúng phần lớn nhưng vẫn bỏ sót 901 mẫu "hài lòng" (False Negative). Các hệ số hồi quy thể hiện rõ tác động của từng đặc trưng đến sự hài lòng.

```
He so chan Intercept: [-0.51174533]
He So Hoi Quy Ung Voi Tung Dac Trung Coefficients: [[-0.04856878  1.19925777 -0.77827722 -0.66238748  0.39843121  1.02715335
 -0.77932459  0.1966884 ]]
-----
Accuracy: 0.8458281284888846
-----
Classification Report:
precision    recall   f1-score  support
0.0          0.85     0.88      0.87     5868
1.0          0.84     0.80      0.82     4523
accuracy
macro avg
weighted avg
```



Airline Passenger Satisfaction

Mô hình đạt độ chính xác 84.58% với F1-score đồng đều, cho thấy hiệu suất tốt, đặc biệt ở lớp "không hài lòng." Tuy nhiên, lớp "hài lòng" vẫn gặp vấn đề với 720 mẫu bị dự đoán sai (False Negative), làm giảm độ nhạy. Các đặc trưng như thời gian trễ và sự thoải mái ngồi có tác động tiêu cực mạnh đến sự hài lòng, trong khi một số yếu tố tích cực như x2 cải thiện đáng kể khả năng hài lòng. Để nâng cao hiệu quả, cần tập trung giảm False Negative, tối ưu ngưỡng dự đoán, và cải thiện các yếu tố dịch vụ như giảm trễ chuyến và nâng cao tiện nghi bay.

BỘ DỮ LIỆU 4

Air Quality Index

Bộ dữ liệu này chúng tôi tiến hành thu thập trên trang web [Visual Crossing](#).

Dữ liệu được thu thập thông qua phương pháp sử dụng API (Application Programming Interface).

Sau khi hoàn thành quá trình thu thập dữ liệu, chúng tôi đã thu được một bộ dữ liệu bao gồm 407 quan trắc và 18 đặc trưng.

Air Quality Index

Tên Biến	Mô tả	Đơn Vị	Ghi Chú
Temp	Nhiệt độ hiện tại	Độ C	
Feelslike	Nhiệt độ cảm nhận được	Độ C	
Humidity	Độ ẩm không khí	%	
Dew	Điểm sương - nhiệt độ tại đó hơi nước ngưng tụ thành sương	Độ C	
Precip	Lượng mưa	mm	
Precipprob	Xác suất xảy ra mưa	%	
Snow	Lượng tuyết rơi	mm	
Snowdepth	Độ sâu lớp tuyết	mm	
Preciptype	Loại hình giáng thủy	-	Mưa, tuyết, mưa đá, v.v.
Windgust	Cơn gió mạnh nhất	km/h	
Windspeed	Tốc độ gió trung bình	km/h	

Mô tả dữ liệu

Winddir	Hướng gió	Độ (0° - 360°)	0° là hướng Bắc
Pressure	Áp suất khí quyển	hPa	
Visibility	Tầm nhìn xa	km	
Cloudcover	Mức độ che phủ của mây	%	
Uvindex	Chỉ số tia cực tím (UV)	-	Thang điểm từ 0 đến 11+
Severerisk	Mức độ rủi ro thời tiết nghiêm trọng	-	Thang điểm
Conditions	Mô tả điều kiện thời tiết	-	Như mây, nắng, mưa, v.v.

Air Quality Index

```
def classify_air_quality_adjusted(row):

    if (row['humidity'] < 80 and row['uvindex'] < 5 and
        row['conditions'] in ['Clear', 'Partially cloudy', 'Overcast']):
        return 'Tốt'
    else:
        return 'Kém'

data['air_quality_adjusted'] = data.apply(classify_air_quality_adjusted, axis=1)

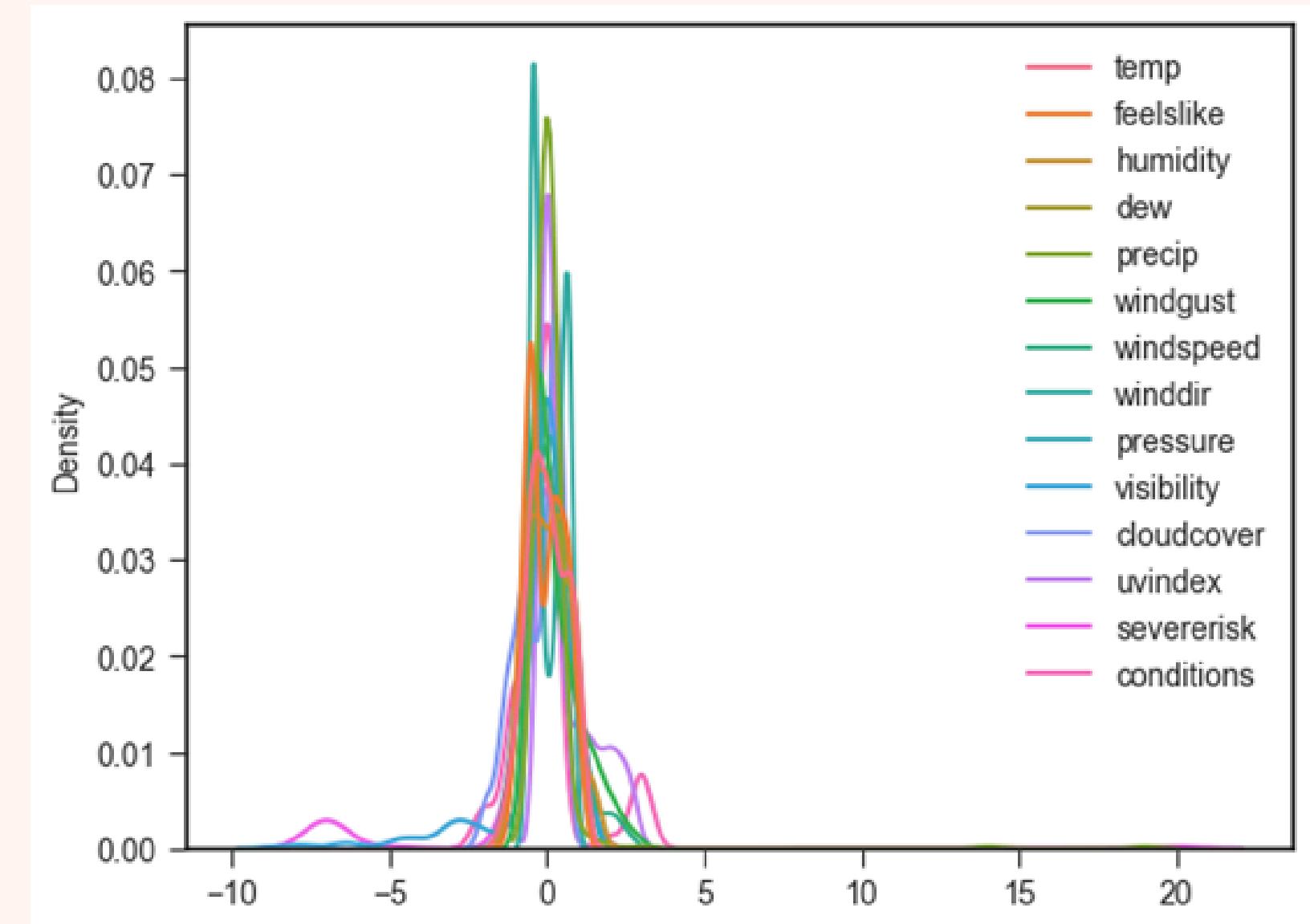
data.isna().sum()
data = data.drop(columns=["Unnamed: 0", "preciptype", "snowdepth", "snow", "precipprob"])
data['precip'].value_counts()
```

Chúng tôi xây dựng mô hình dự đoán chất lượng không khí với hai trạng thái: "Tốt" và "Kém". Để tăng hiệu quả, các thuộc tính ít liên quan như snow, snowdepth, preciptype, và precipprob đã được loại bỏ. Một biến mục tiêu mới, air_quality_adjusted, được tạo dựa trên độ ẩm (humidity), chỉ số UV (uvindex), và điều kiện thời tiết (conditions). Quy tắc phân loại mở rộng tiêu chí "Tốt": độ ẩm dưới 80%, UV dưới 5, và điều kiện thời tiết như "Clear", "Partially cloudy", hoặc "Overcast". Các trường hợp khác được phân loại là "Kém", đảm bảo phản ánh chính xác hơn chất lượng không khí.

Air Quality Index

Tiền xử lý dữ liệu

- Giá trị outliers: các giá trị này chủ yếu nằm gần ngưỡng của khoảng từ phân vị (IQR), cho thấy khả năng cao chúng không phải là ngoại lai thực sự.
- Mã hoá các biến object đưa về dạng số: OrdinalEncoder
- Chuẩn hoá dữ liệu đưa về cùng một thang đo.



Air Quality Index

Tiến hành sử dụng PCA để giảm chiều dữ liệu để loại bỏ các thông tin dư thừa hoặc không quan trọng mà vẫn giữ được phần lớn thông tin cần thiết.

Kết quả cho thấy số lượng thành phần chính được chọn là 7

```
# Ti le phuong sai  
expalained_var = pca.explained_variance_ratio_ # khoan tin cay  
print(f'Ty le phuong sai giai thich {expalained_var}')
```

```
# chon tren phuong sai tich luy  
cumsum_explained_var = np.cumsum(pca.explained_variance_ratio_)  
print(f'Phuong sai tich luy: {cumsum_explained_var}')
```

```
# Lua chon thanh phan chinh ty le > 0.9  
n_components = np.argmax(cumsum_explained_var >=0.9) + 1  
print(f'So luong thanh phan chinh duoc chon : {n_components}')
```

So luong thanh phan chinh duoc chon : 7

Air Quality Index

Dữ liệu 407 quan sát, Pseudo R-squared = 0.3780, giải thích 37,8% bộ dữ liệu.

Hầu hết các biến độc lập có ý nghĩa thống kê ($p\text{-value} < 0.05$), khẳng định mối quan hệ đáng kể với sự hài lòng.

Giá trị LLR p-value = 3.541e-42 cho thấy mô hình tổng thể có ý nghĩa thống kê cao.

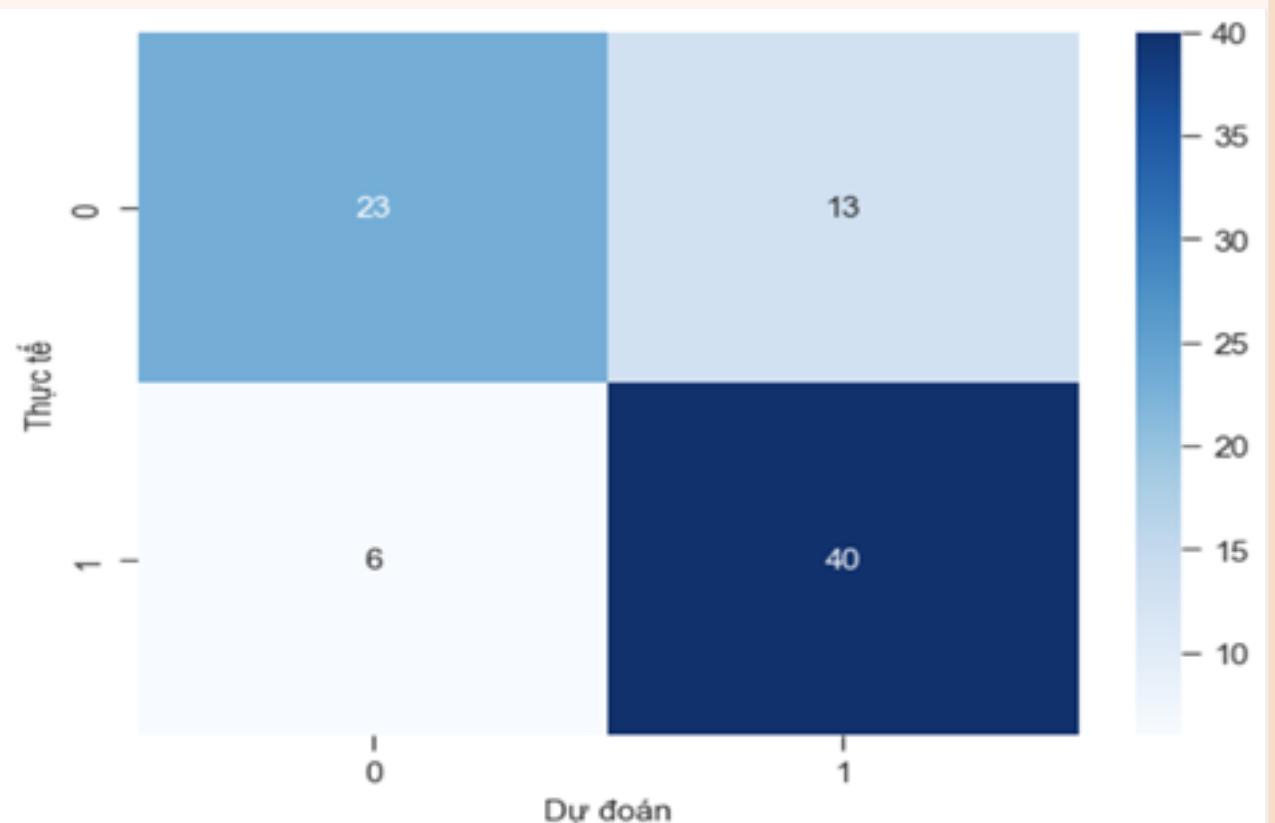
Optimization terminated successfully. Current function value: 0.428295 Iterations 9							
Logit Regression Results							
Dep. Variable:	air_quality_adjusted	No. Observations:	407	Model:	Logit	Df Residuals:	399
Method:		MLE		Df Model:			7
Date:	Mon, 06 Jan 2025	Pseudo R-squ.:	0.3780	Time:	20:43:08	Log-Likelihood:	-174.32
converged:		True	LL-Null:				-280.24
Covariance Type:		nonrobust	LLR p-value:				3.541e-42
	coef	std err	z	P> z	[0.025	0.975]	
const	-1.4780	0.314	-4.709	0.000	-2.093	-0.863	
x1	-0.1505	0.068	-2.200	0.028	-0.285	-0.016	
x2	-0.2774	0.135	-2.058	0.040	-0.542	-0.013	
x3	-8.7297	1.312	-6.651	0.000	-11.302	-6.157	
x4	-8.3904	1.431	-5.863	0.000	-11.195	-5.585	
x5	6.9061	1.180	5.853	0.000	4.594	9.219	
x6	2.4004	0.323	7.428	0.000	1.767	3.034	
x7	-0.4793	0.229	-2.096	0.036	-0.927	-0.031	

Air Quality Index

Mô hình dự đoán chất lượng không khí đạt độ chính xác 77%. Precision là 79% cho lớp "kém" và 75% cho lớp "tốt", recall lần lượt là 64% và 87%. F1-score đạt 71% cho lớp "kém" và 81% cho lớp "tốt", với giá trị trung bình 76%. Dữ liệu phân bố cân bằng, đảm bảo tính chính xác và khách quan trong kết quả phân loại.

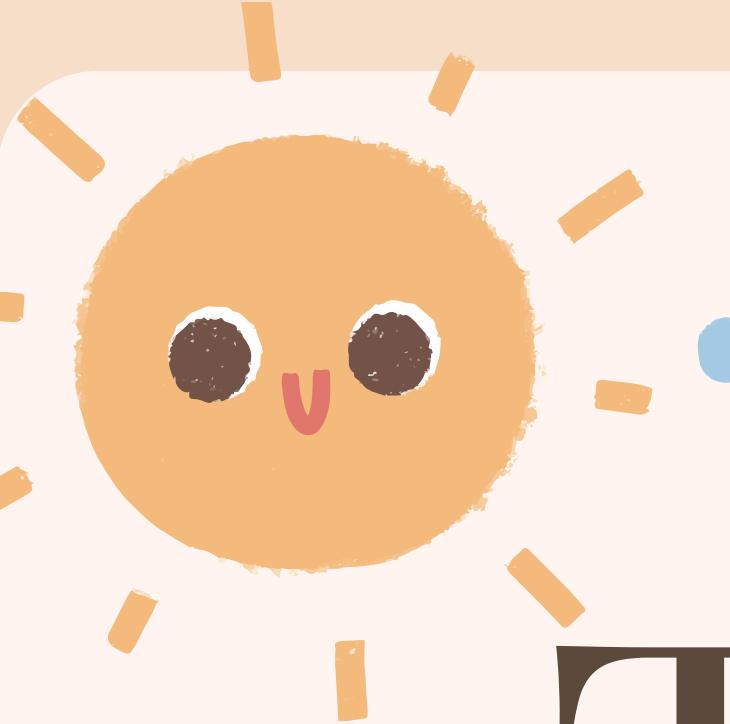
```
He so chan Intercept: [-0.10788385]
He so hoi quy ung voi tung dac trung Cofficients: [[-0.05966694  0.0594603 -2.53772548 -1.41386207  1.05959538  1.22623186
 0.06626928]]
Accuracy:  0.7682926829268293
Classification Report:
precision    recall   f1-score  support
0.0          0.79     0.64      0.71      36
1.0          0.75     0.87      0.81      46

accuracy
macro avg
weighted avg
```



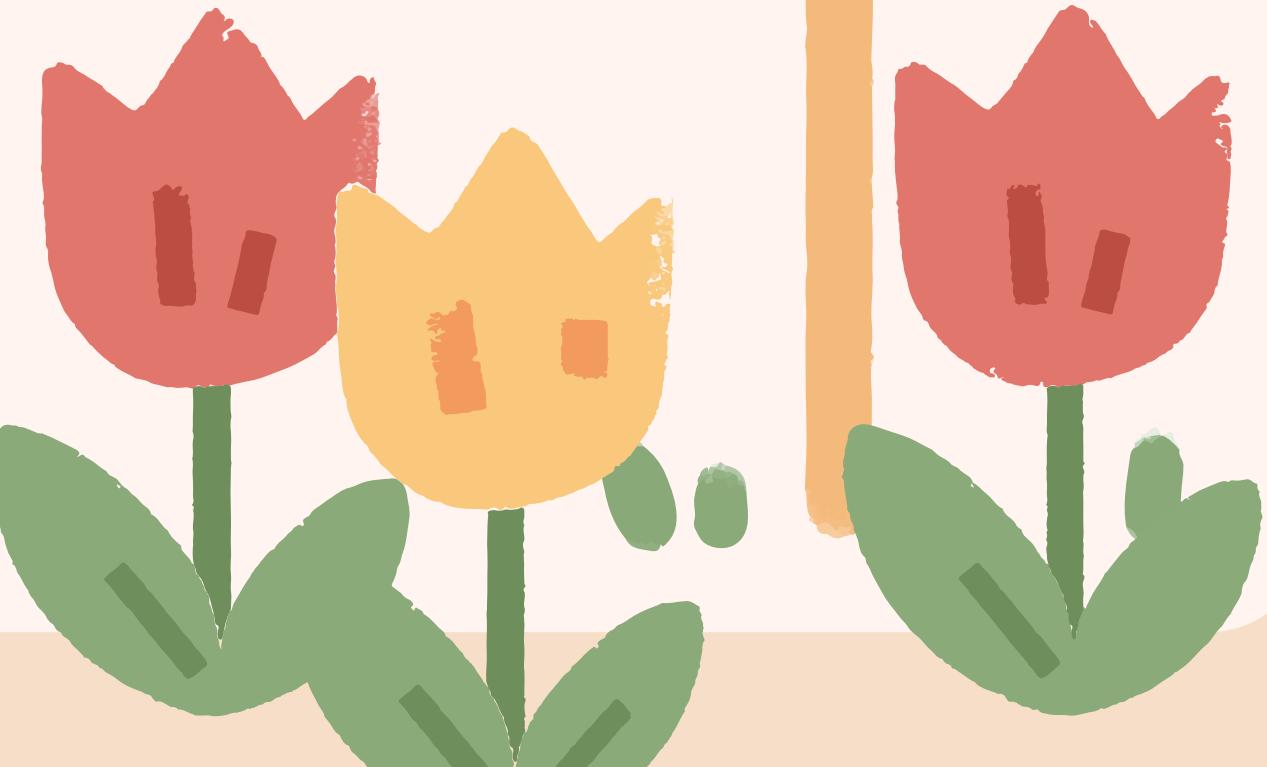
Air Quality Index

Mô hình đạt độ chính xác 77%, với precision và recall của lớp "tốt" cao hơn lớp "kém". Tuy nhiên, recall của lớp "kém" chỉ đạt 64%, cho thấy mô hình bỏ sót nhiều trường hợp. Mô hình có thể sử dụng, nhưng cần cải thiện khả năng phân loại lớp "kém" và cân bằng hiệu suất giữa hai lớp. Cần xem xét tối ưu ngưỡng dự đoán hoặc áp dụng thuật toán phân loại bổ sung để cải thiện kết quả.



Thank you!

www.reallygreatsite.com



Resource Page



Our values

Leadership

Innovation

Integrity

Teamwork

Diversity

Quality

