DỰ BÁO PM2.5 VÀ CẢNH BÁO AQI THEO TRẠM

MiniProject

Sinh viên thực hiện: Lưu Thanh Tùng

Lớp: KHMT - 1701

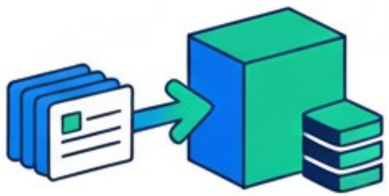
,

ĐƯỜNG DẪN GITHUB

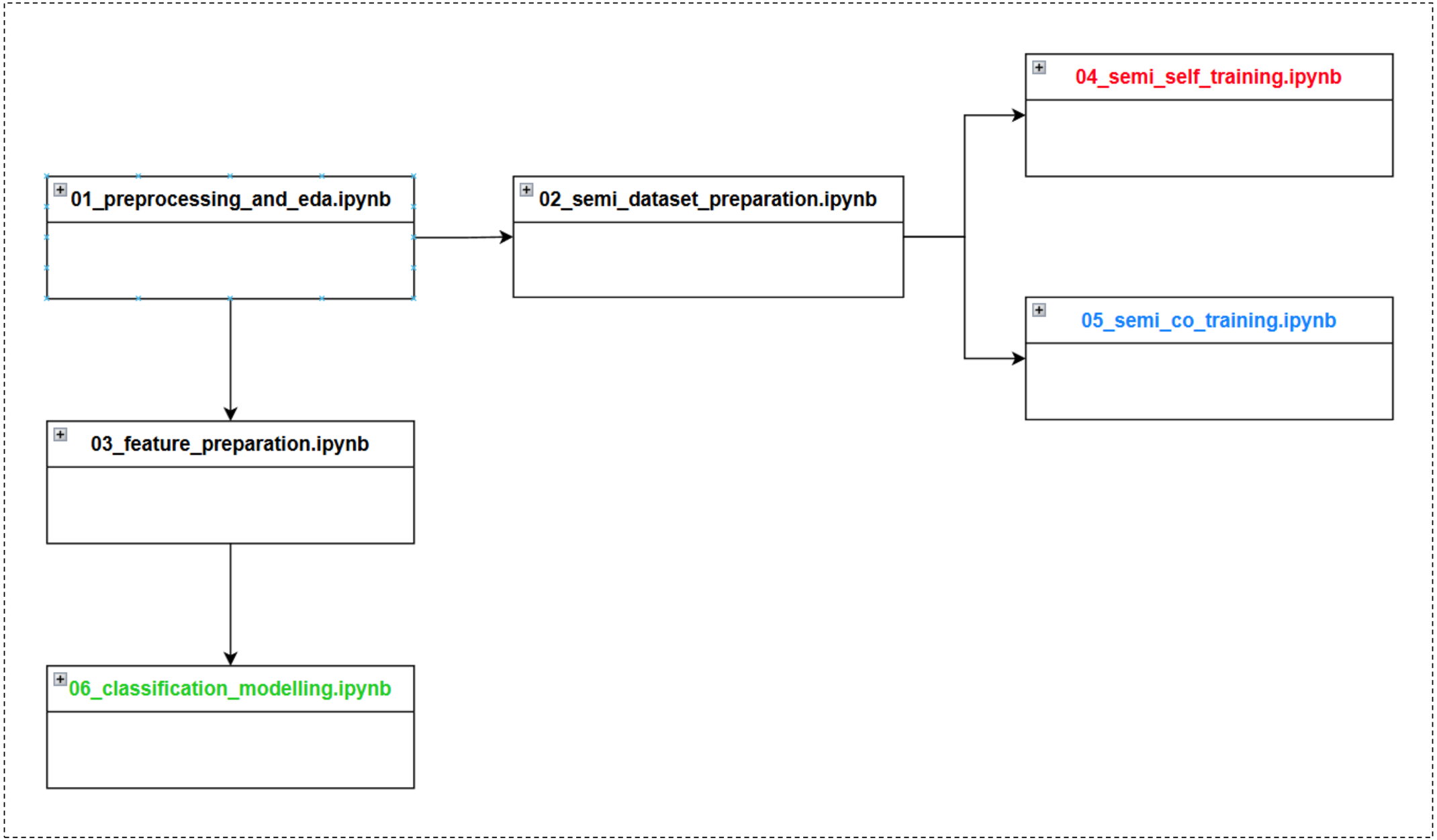
https://github.com/ThanhTung-KHMT-1701/AirGuard

Why Semi-Supervised Learning for AQI Forecasting?

- Urbanization & climate change raise health risks
- Labeled AQI data (PM2.5) is scarce and costly
- Unlabeled time-series data is abundant and underused



QUY TRÌNH XỬ LÝ DỮ LIỆU



KẾT QUẢ THỰC NGHIỆM

❖ File: 01_preprocessing_and_eda.ipynb

05_semi_co_training.ipynb | 01_cleaned_data_sample.csv

data > processed > 01_cleaned_data_sample.csv > data

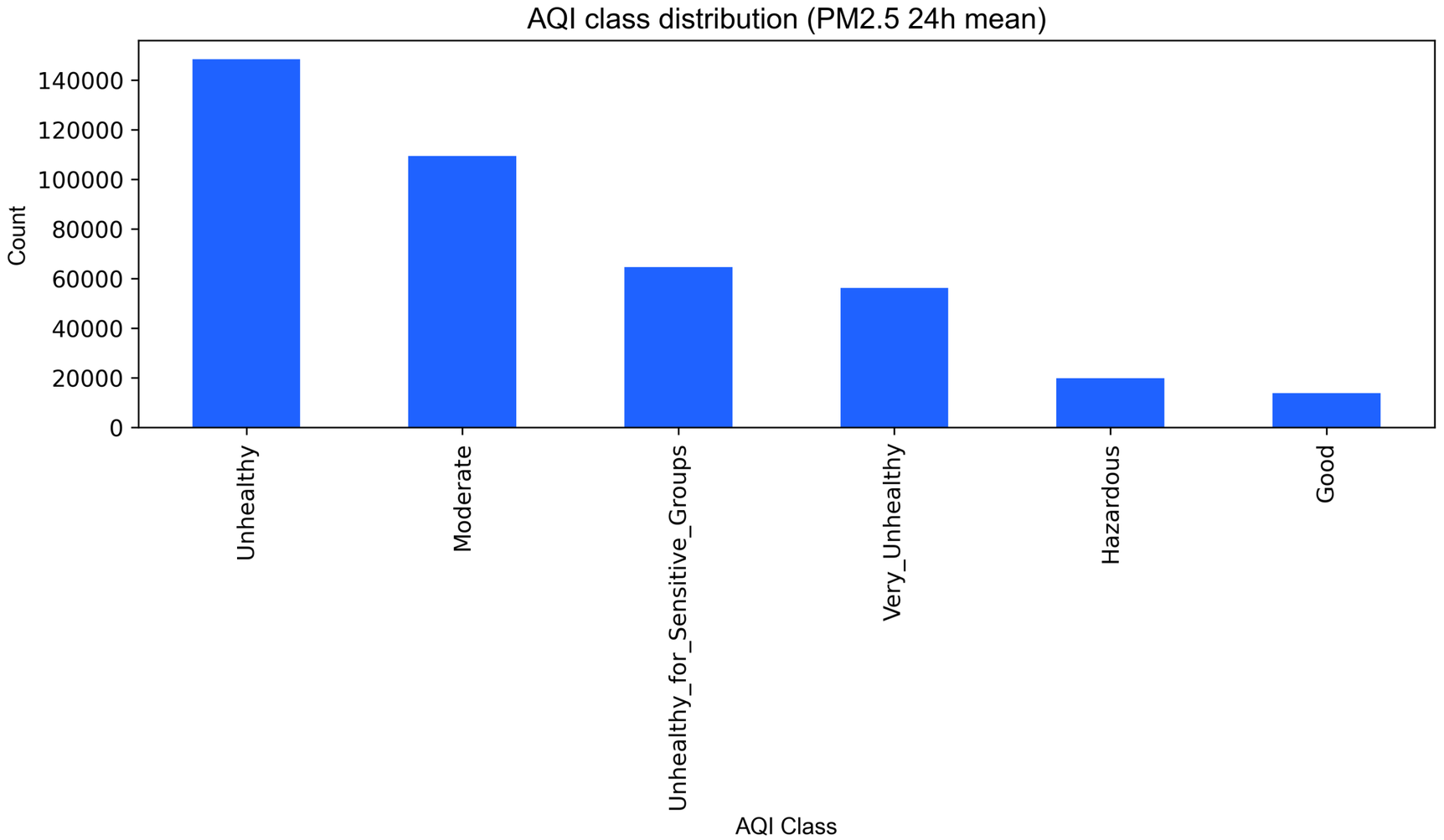
```
1  datetime,station,PM2.5,pm25_24h,aqi_class
29 2013-03-02 03:00:00,Aotizhongxin,3.0,8.25,Good
30 2013-03-02 04:00:00,Aotizhongxin,3.0,8.25,Good
31 2013-03-02 05:00:00,Aotizhongxin,9.0,8.416666666666666,Good
32 2013-03-02 06:00:00,Aotizhongxin,4.0,8.458333333333334,Good
33 2013-03-02 07:00:00,Aotizhongxin,3.0,8.458333333333334,Good
34 2013-03-02 08:00:00,Aotizhongxin,3.0,8.458333333333334,Good
35 2013-03-02 09:00:00,Aotizhongxin,10.0,8.75,Good
36 2013-03-02 10:00:00,Aotizhongxin,11.0,9.083333333333334,Moderate
37 2013-03-02 11:00:00,Aotizhongxin,18.0,9.708333333333334,Moderate
38 2013-03-02 12:00:00,Aotizhongxin,26.0,10.666666666666666,Moderate
39 2013-03-02 13:00:00,Aotizhongxin,25.0,11.583333333333334,Moderate
40 2013-03-02 14:00:00,Aotizhongxin,26.0,12.416666666666666,Moderate
41 2013-03-02 15:00:00,Aotizhongxin,37.0,13.625,Moderate
```

data\processed\01_cleaned_data_sample.csv

01_preprocessing_and_eda.ipynb	
Đầu vào	USE_UCIMLREPO = False RAW_ZIP_PATH = "PRSA2017*.zip" LAG_HOURS = [1, 3, 24]
	Dữ liệu ban đầu có 18 cột
Đầu ra	File Parquet: 01_cleaned.parquet
	55 cột, bao gồm các cột: + Cột nhãn phân loại "**aqi_class**" + Các cột về thời gian "datetime", "**hour_cos**", "**hour_sin**", "dow", "month", "is_weekend" + Các cột **LAG_FEATURES**

KẾT QUẢ THỰC NGHIỆM

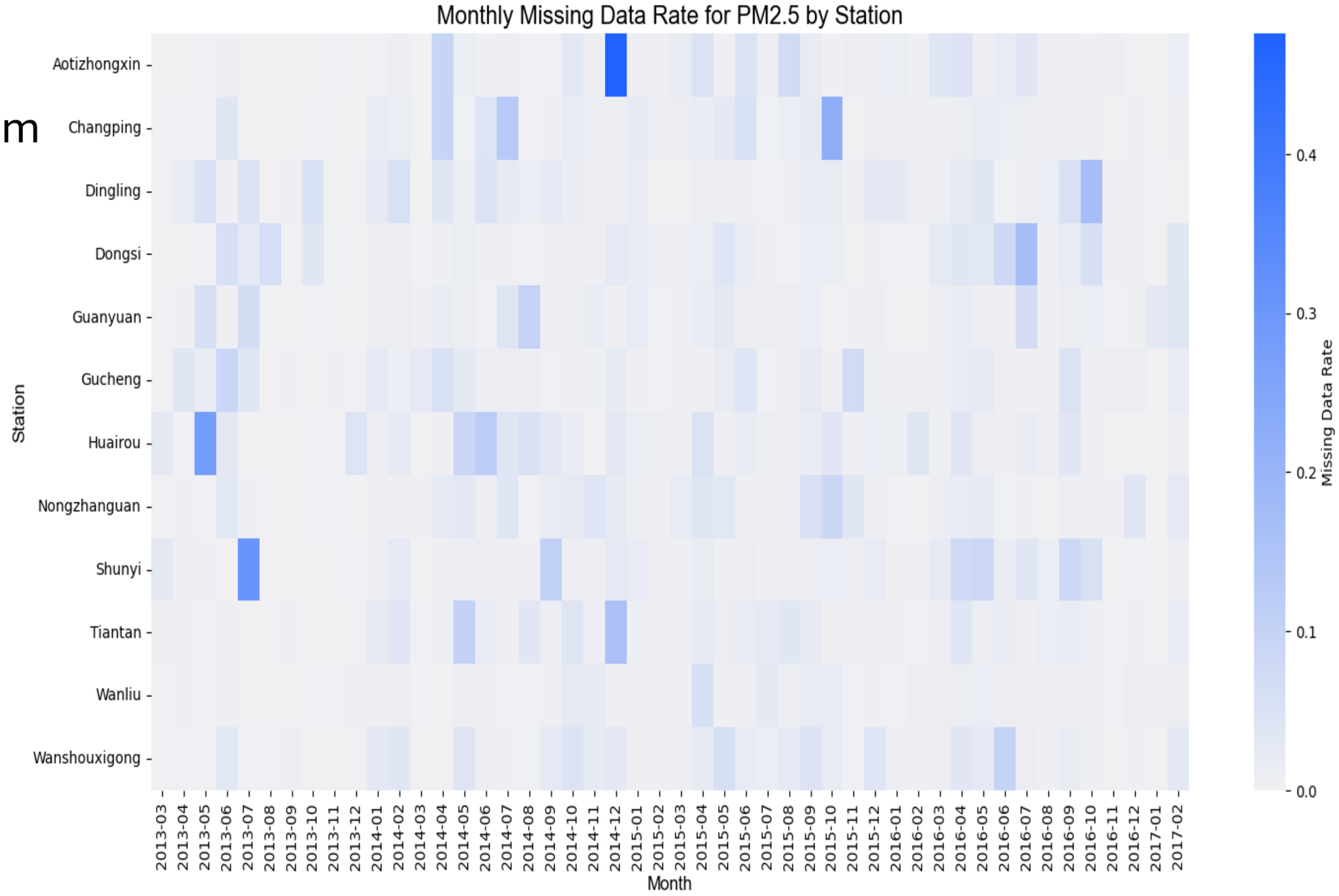
- File: 01_preprocessing_and_eda.ipynb
- Dữ liệu bị mất cân bằng



KẾT QUẢ THỰC NGHIỆM

- File: 01_preprocessing_and_eda.ipynb
- Dữ liệu bị mất cân bằng
- Dữ liệu bị thiếu nhiều nhất ở trạm

  Aotizhongxin



KẾT QUẢ THỰC NGHIỆM

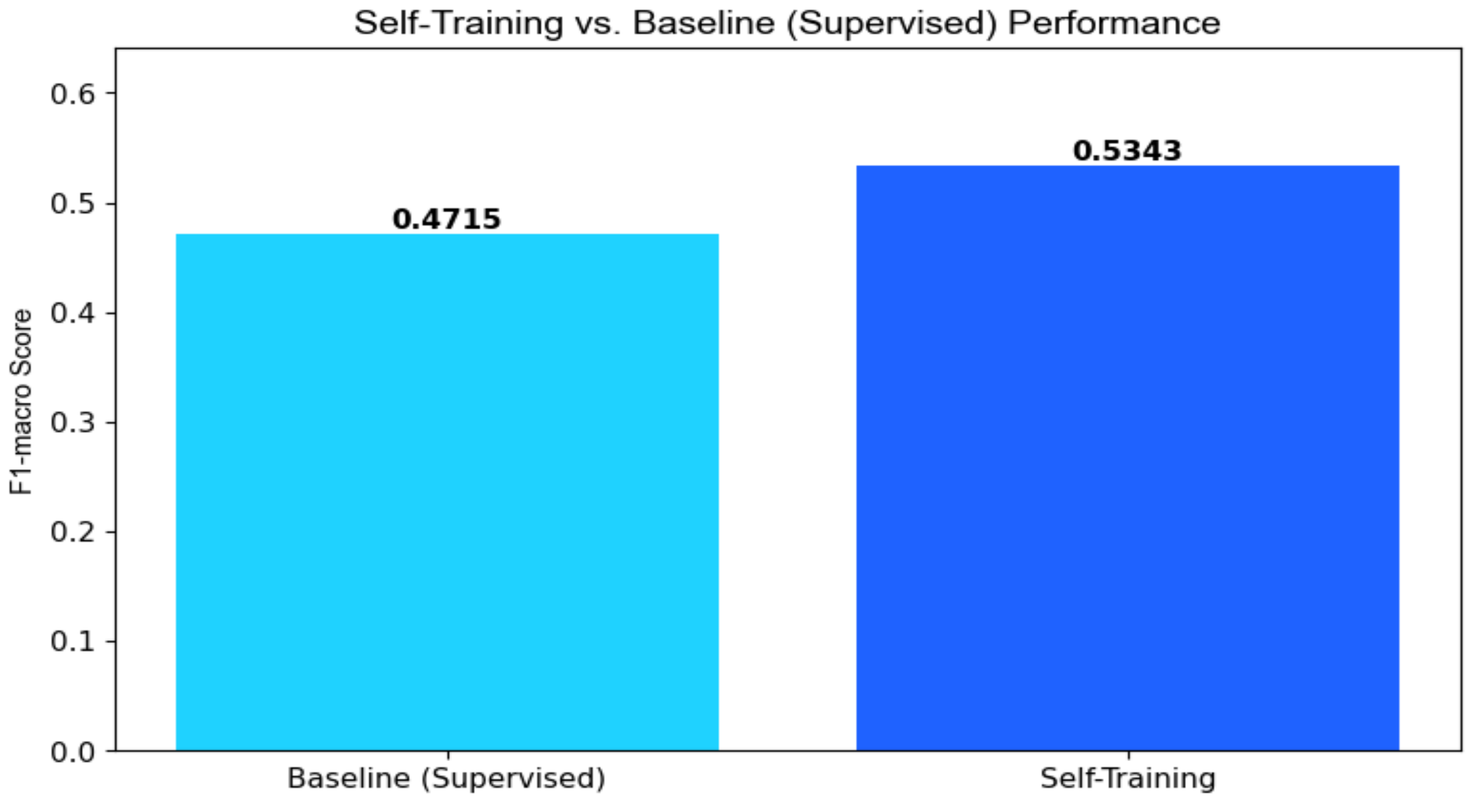
- File: 02_semi_dataset_preparation.ipynb
- Dữ liệu có nhãn được “che đi” 95%

02_semi_dataset_preparation.ipynb	
Đầu vào	File Parquet: 01_cleaned.parquet CUTOFF="2017-01-01" LABEL_MISSING_FRACTION=0.95 RANDOM_STATE=42
Đầu ra	File Parquet: 02_dataset_for_semi.parquet
	56 cột, bao gồm cột **"is_labeled"**

AW	AX	AY	AZ	BA	BB	BC	BD
CO_lag24	O3_lag24	TEMP_lag24	PRES_lag24	DEWP_lag24	RAIN_lag24	WSPM_lag24	is_labeled
500	22	3.8	1018.9	-11.4	0	2.3	FALSE
600	17	6.5	1019.7	-10.7	0	2.3	FALSE
700	21	9	1020.7	-12.2	0	3.1	FALSE
500	52	10.6	1020.8	-12.1	0	1.3	FALSE
600	57	11.9	1020.4	-12.4	0	2.1	FALSE
400	68	13.1	1020	-13	0	3	FALSE
400	77	14.2	1018.9	-13.9	0	2.7	FALSE
300	82	15.3	1017.8	-13	0	2	FALSE
400	84	15.2	1017	-13.1	0	0.7	FALSE
400	84	15.3	1016.3	-13	0	2.4	FALSE
500	70	14.5	1015.7	-13.7	0	2.6	FALSE
500	78	12.7	1015.9	-11.8	0	2.6	FALSE
600	75	11.6	1016.5	-11.3	0	3.4	FALSE

KẾT QUẢ THỰC NGHIỆM

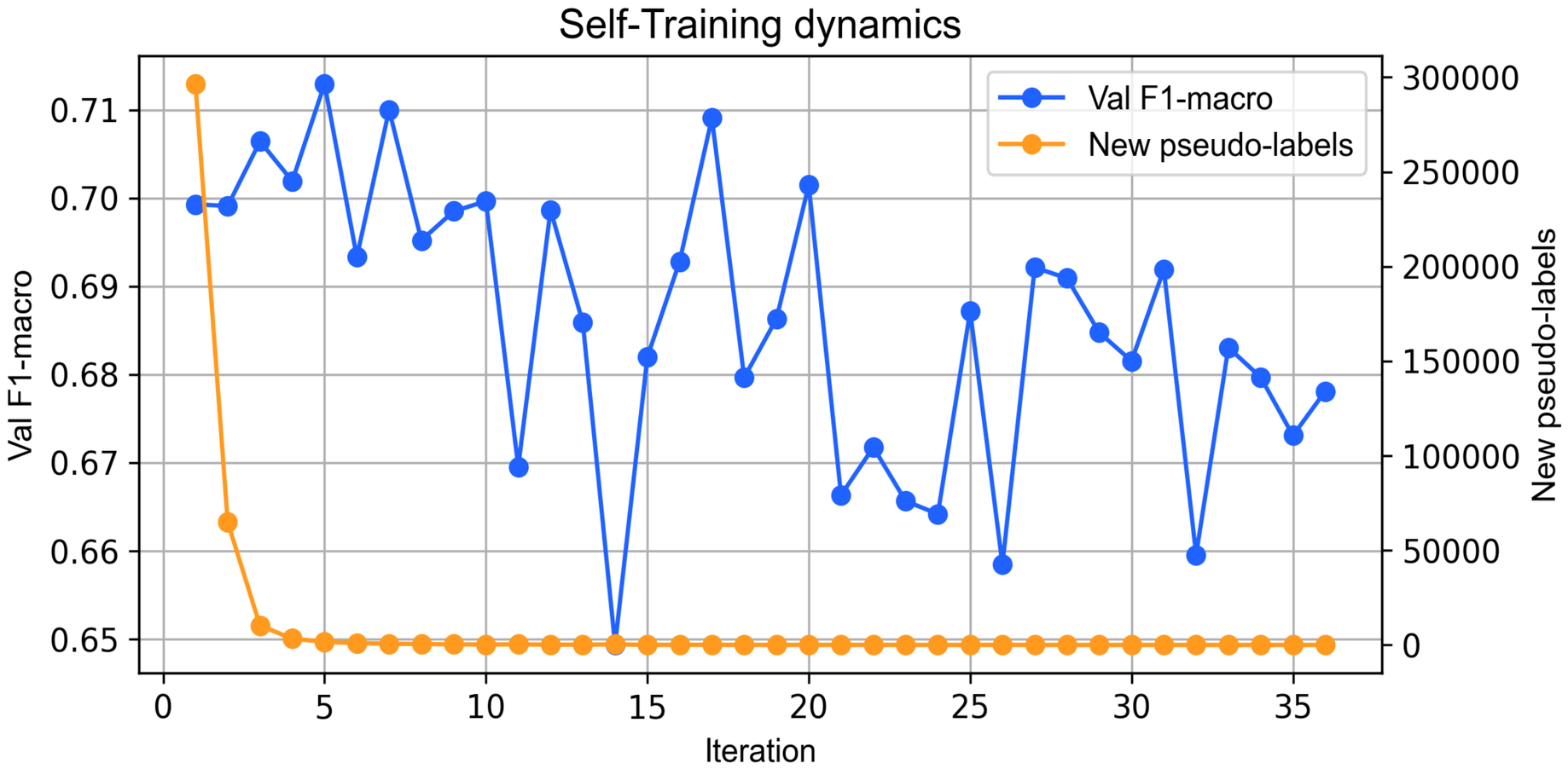
- File: 04_semi_self_training.ipynb
- Áp dụng thuật toán SelfTraining với thông số mặc định ban đầu, thực nghiệm cho thấy chỉ số F1-macro tốt hơn thuật toán Supervised



04_semi_self_training.ipynb	
Đầu vào	File Parquet: 02_dataset_for_semi.parquet CUTOFF = "2017-01-01" **TAU = 0.90** MAX_ITER=10 MIN_NEW_PER_ITER=20 VAL_FRAC=0.20 RANDOM_STATE=42 ALERT_FROM_CLASS="**Unhealthy**"
Đầu ra	File JSON: 04_metrics_self_training.json File CSV: 04_predictions_self_training_sample.csv File CSV: 04_alerts_self_training_sample.csv
	File "04_predictions_self_training_sample.csv" có các cột sau đây: "datetime", "station", "**y_true**", "**y_pred**"
	File "04_alerts_self_training_sample.csv" có các cột sau đây: "datetime", "station", "y_true", "y_pred", "**severity_rank**", "**is_alert**"
	File "04_metrics_self_training.json" có các trường thông tin sau đây: + method + data_cfg: {target_col, **cutoff**, **random_state**, leakage_cols: [ ] } + st_cfg: {**tau**, **max_iter**, **min_new_per_iter**, **val_frac**} + history: [{iter, val_accuracy, val_f1_marco, unlabeled_pool, new_pseudo, tau}]

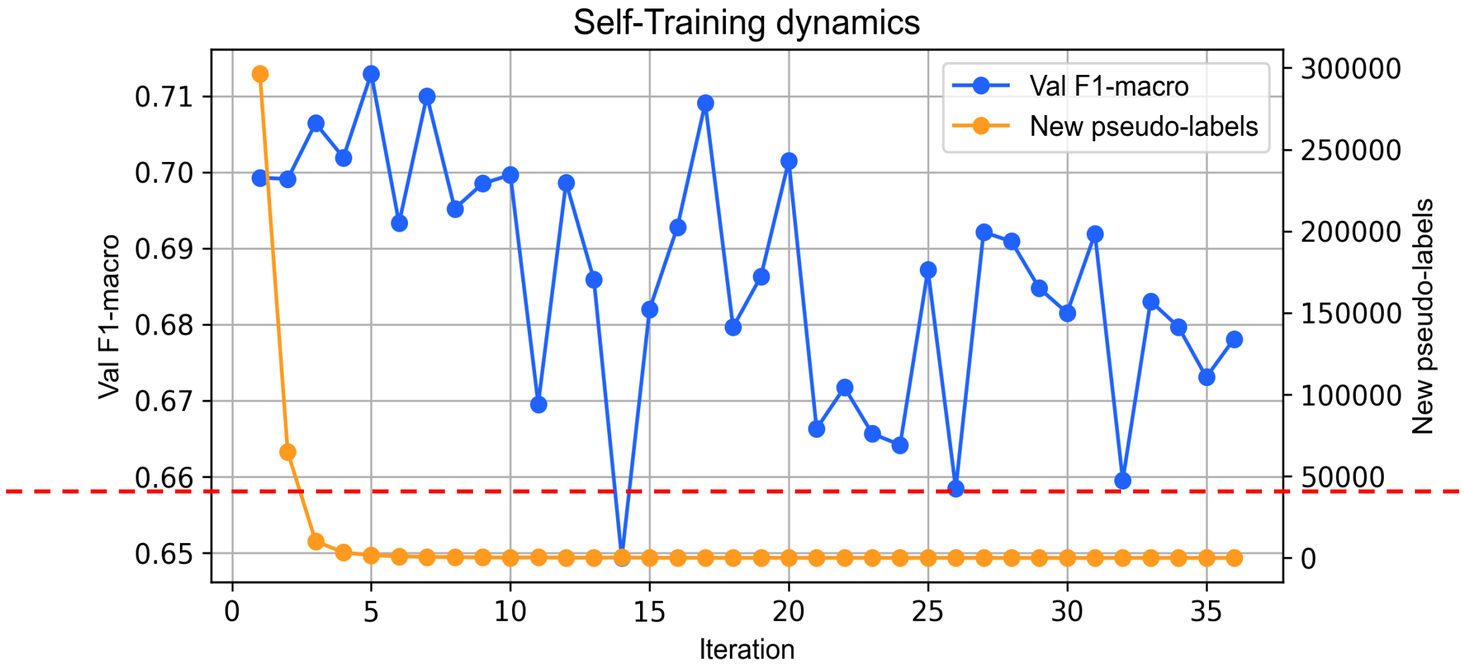
KẾT QUẢ THỰC NGHIỆM

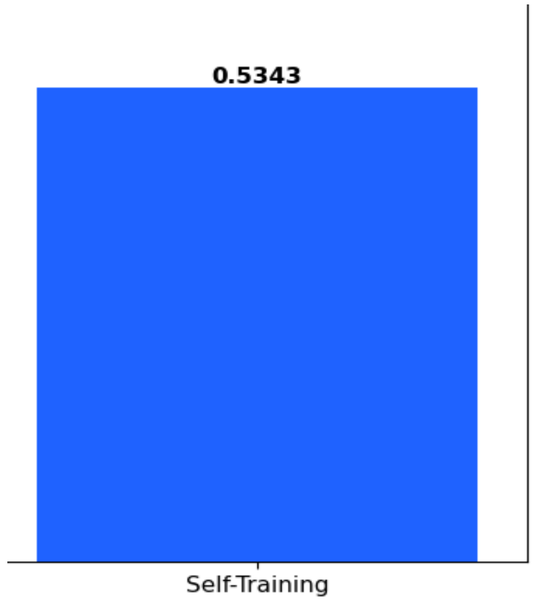
- File: 04_semi_self_training.ipynb
- Áp dụng thuật toán SelfTraining cho chỉ số F1-macro tốt hơn khi áp dụng thuật toán Supervised
- Chỉ số F1-marco biến thiên không phụ thuộc vào số lần lặp trong quá trình huấn luyện



KẾT QUẢ THỰC NGHIỆM

- File: 04_semi_self_training.ipynb
- Áp dụng thuật toán SelfTraining cho chỉ số F1-macro tốt hơn khi áp dụng thuật toán Supervised
- Chỉ số F1-marco biến thiên không phụ thuộc vào số lần lặp trong quá trình huấn luyện
- Quá trình kiểm tra cho thấy chỉ số F1-marco thấp hơn nhiều so với quá trình huấn luyện.

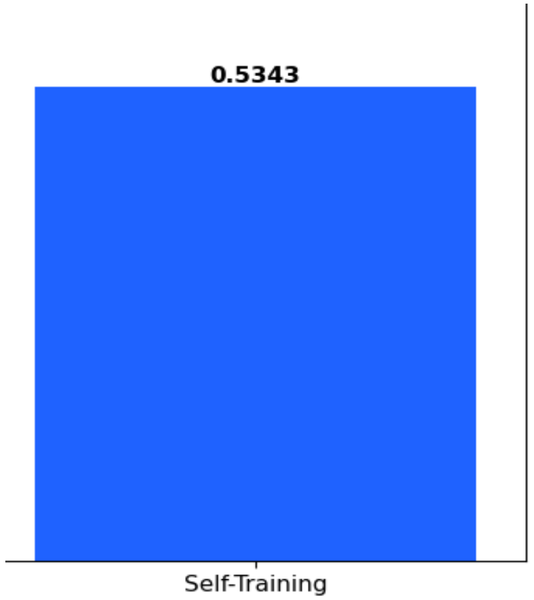




KẾT QUẢ THỰC NGHIỆM

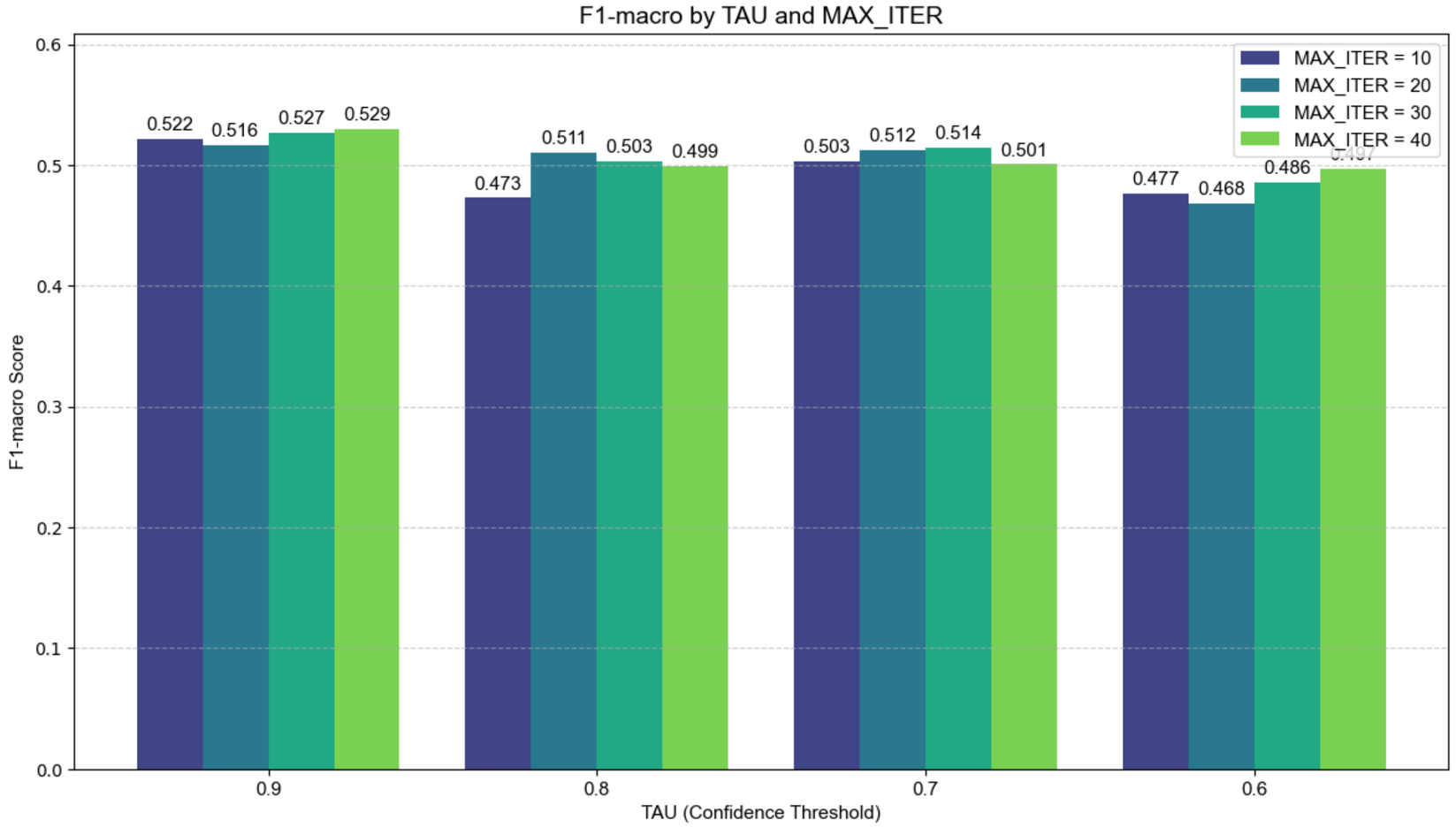
- File: 04_semi_self_training.ipynb
- Áp dụng thuật toán SelfTraining cho chỉ số F1-macro tốt hơn khi áp dụng thuật toán Supervised
- Chỉ số F1-marco biến thiên không phụ thuộc vào số lần lặp trong quá trình huấn luyện
- Mô hình bị "overfitting"

Chúng ta có thể thử nghiệm với nhiều trường hợp **TAU** khác nhau



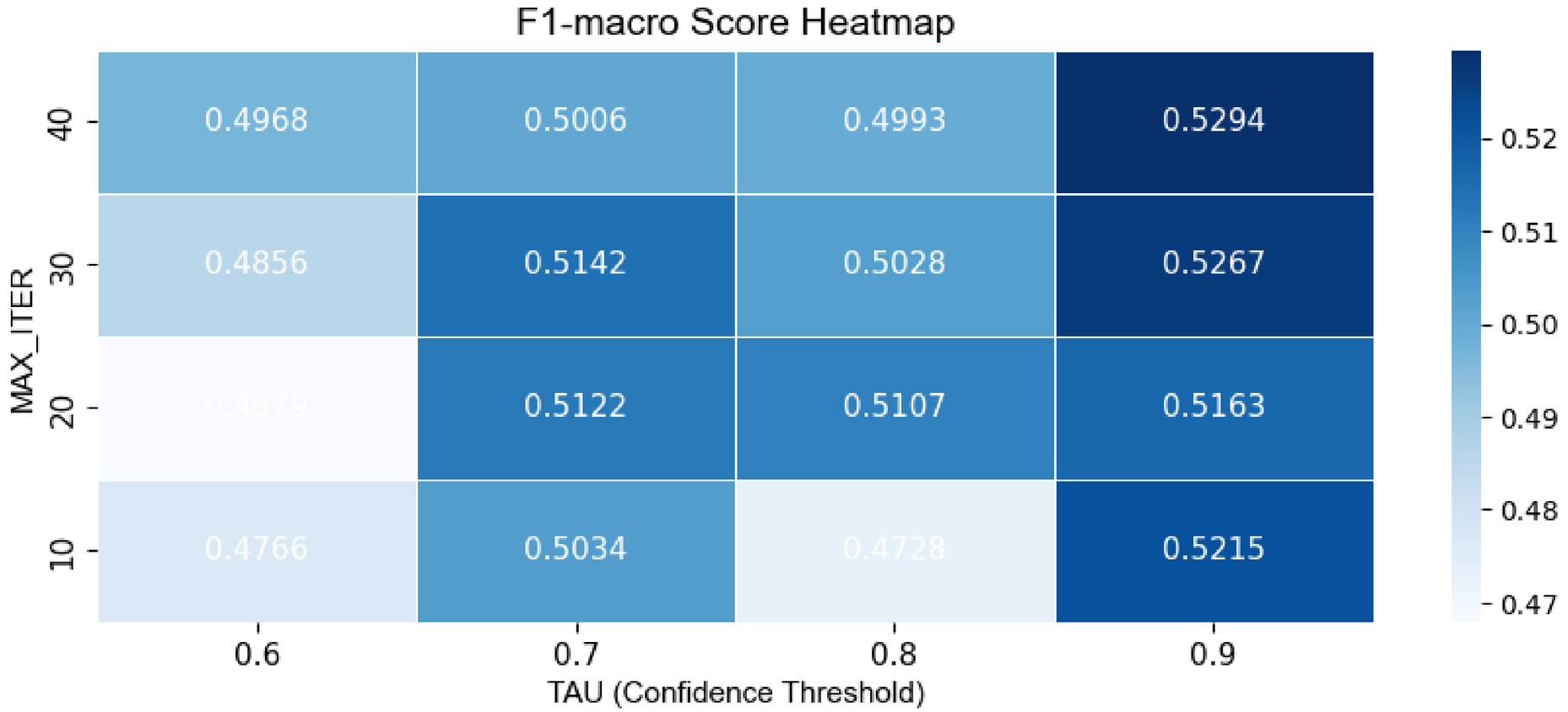
KẾT QUẢ THỰC NGHIỆM

- File: 04_semi_self_training.ipynb
- Thử nghiệm với nhiều trường hợp TAU (mặc định) và MAX_ITER, file: 10_Question01.ipynb
- Với cùng một giá trị TAU, quan hệ giữa thông số **MAX_ITER** với chỉ số **F1-macro** không tuyến tính



KẾT QUẢ THỰC NGHIỆM

- File: 04_semi_self_training.ipynb
- Thử nghiệm với nhiều trường hợp TAU (mặc định) và MAX_ITER, file: 10_Question01.ipynb
- Với cùng một giá trị TAU, quan hệ giữa thông số **MAX_ITER** với chỉ số **F1-macro** không tuyến tính
- Với cùng số lần lặp, quan hệ giữa thông số **TAU** với chỉ số **F1-marco** không tuyến tính



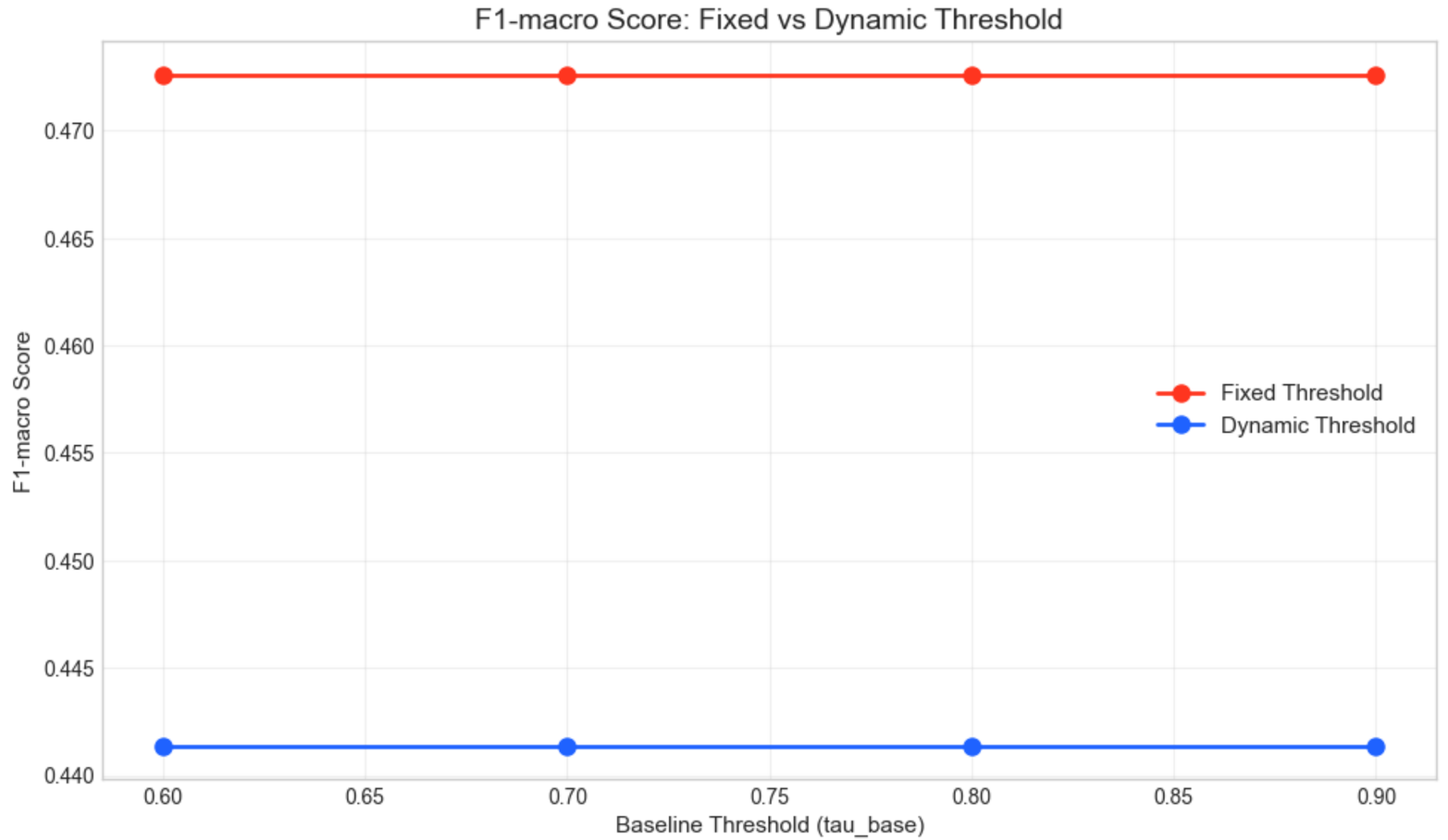
KẾT QUẢ THỰC NGHIỆM

- File: 04_semi_self_training.ipynb
- Thử nghiệm với nhiều trường hợp TAU (mặc định) và MAX_ITER, file: 10_Question01.ipynb
- Với cùng một giá trị TAU, quan hệ giữa thông số **MAX_ITER** với chỉ số **F1-macro** không tuyến tính
- Với cùng số lần lặp, quan hệ giữa thông số **TAU** với chỉ số **F1-marco** không tuyến tính

Cấu hình mặc định ban đầu cho kết quả đầu ra có chỉ số F1-marco cao nhất

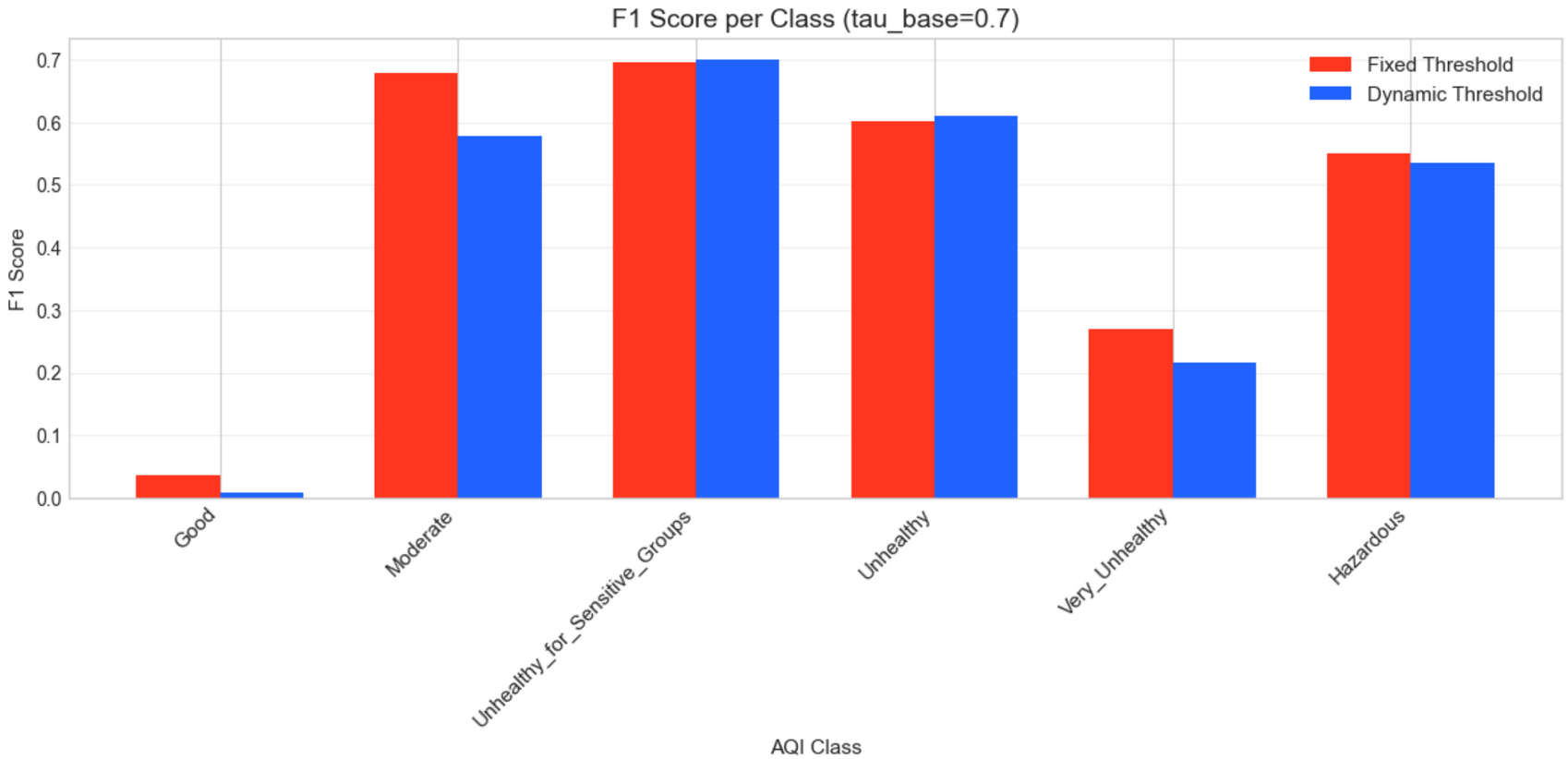
KẾT QUẢ THỰC NGHIỆM

- File: 04_semi_self_training.ipynb
- Thử nghiệm với nhiều trường hợp TAU sử dụng Dynamic Threshold, file: 13_Question04.ipynb
- Kết quả thực tế cho thấy chỉ số F1-macro **bị giảm**



KẾT QUẢ THỰC NGHIỆM

- File: 04_semi_self_training.ipynb
- Thử nghiệm với nhiều trường hợp TAU sử dụng Dynamic Threshold, file: 13_Question04.ipynb
- Kết quả thực tế cho thấy chỉ số F1-macro **bị giảm**
- Giá trị **F1 score** trên các lớp đều **bị giảm**



KẾT QUẢ THỰC NGHIỆM

❖ File: 05_semi_co_training.ipynb

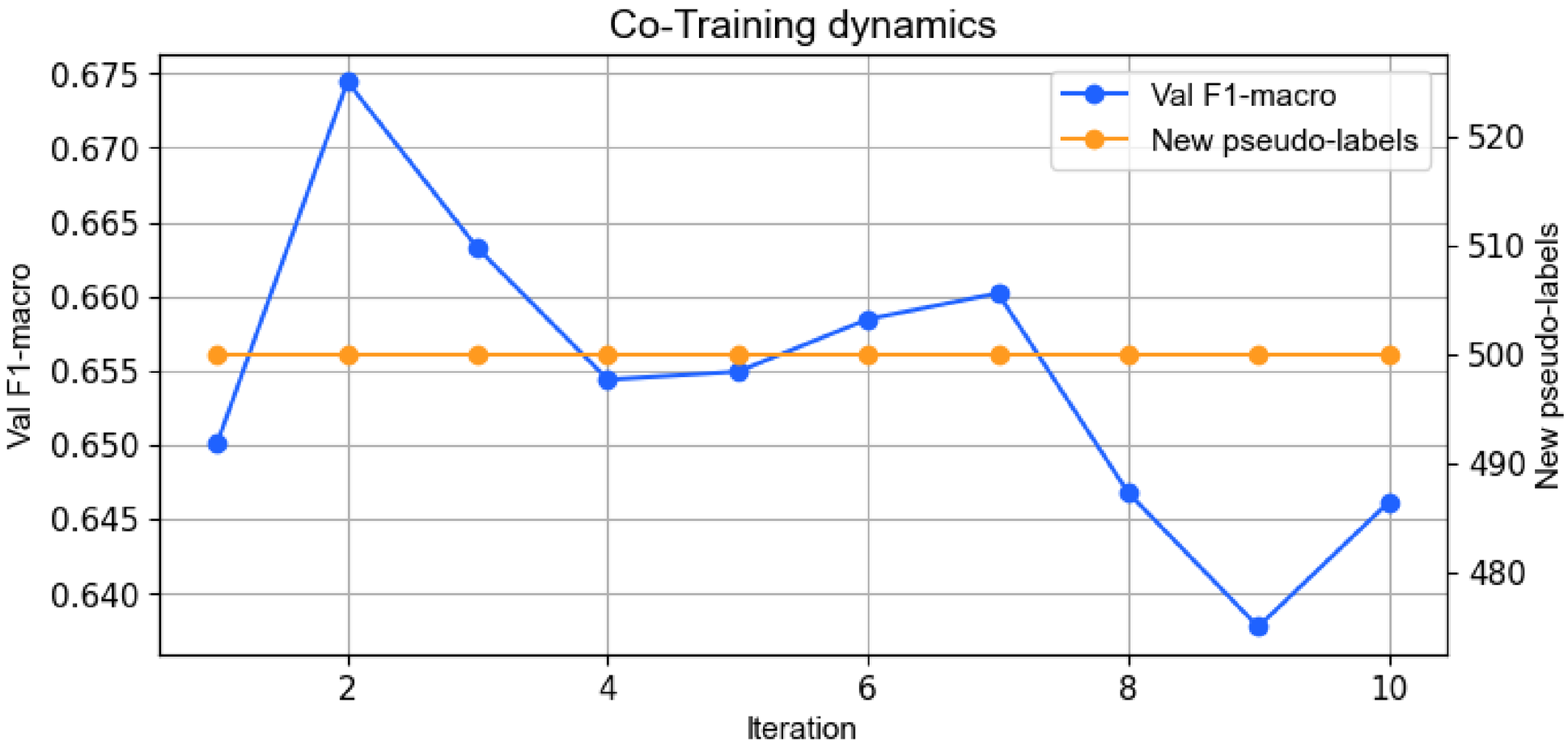
05_semi_co_training.ipynb	
Đầu vào	File Parquet: 02_dataset_for_semi.parquet CUTOFF="2017-01-01" **TAU = 0.90** MAX_ITER=10 MIN_NEW_PER_ITER=20 **MAX_NEW_PER_ITER=500** VAL_FRAC=0.20 RANDOM_STATE=42 ALERT_FROM_CLASS="**Unhealthy**" **VIEW1_COLS=None** **VIEW2_COLS=None**

Đầu ra	File JSON: 05_metrics_co_training.json File CSV: 05_predictions_co_training_sample.csv File CSV: 05_alerts_co_training_sample.csv

	File CSV: 05_predictions_co_training_sample.csv có các cột sau đây: "datetime", "station", "**y_true**", "**y_pred**"
	
	File CSV: 05_alerts_co_training_sample.csv có các cột sau đây: "datetime", "station", "y_true", "y_pred", "**severity_rank**", "**is_alert**"
	
	File JSON: 05_metrics_co_training.json có các trường thông tin sau đây:
	+ method
	+ data_cfg: {target_col, **cutoff**, **random_state**, leakage_cols: [ ] }
	+ st_cfg: {**tau**, **max_iter**, **min_new_per_iter**, **val_frac**}
	+ history: [{iter, val_accuracy, val_f1_marco, unlabeled_pool, new_pseudo, tau}]
	+ test_metrics: {
	cutoff, n_train, n_test, accuracy, f1_marco,
	report: {
	Good: {precision, recall, f1-score, support},
	Hazardous: {precision, recall, f1-score, support}
	},
	confusion_matrix: [[38, 980, 0, 14, 0, 0], [...]],
	labels: ["Good", "Moderate", ...]
	feature_cols: [ ],
	categorical_cols: [ ],
	numeric_cols: [ ]
	}
	+ model_info: {
	view1_cols: [ ],
	view2_cols: [ ],
	view1_numeric_cols: [ ],
	view1_categorical_cols: [ ],
	view2_numeric_cols: [ ],
	view2_categorical_cols: [ ]
	}

KẾT QUẢ THỰC NGHIỆM

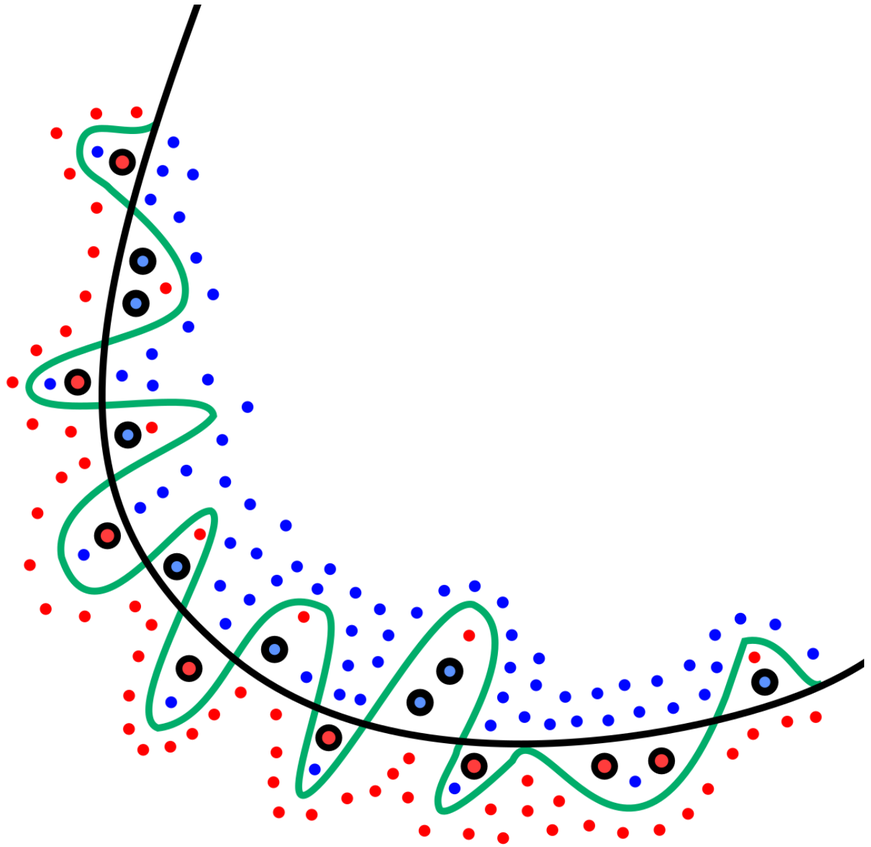
- File: 05_semi_co_training.ipynb
- Kết quả thực nghiệm cho thấy chỉ số F1-marco đạt xấp xỉ 0.645 khi kết thúc quá trình huấn luyện



KẾT QUẢ THỰC NGHIỆM

- File: 05_semi_co_training.ipynb
- Kết quả thực nghiệm cho thấy chỉ số F1-marco đạt xấp xỉ 0.645 khi kết thúc quá trình huấn luyện
- Tuy nhiên, trên tập kiểm tra thì F1-marco chỉ đạt **0.4044**

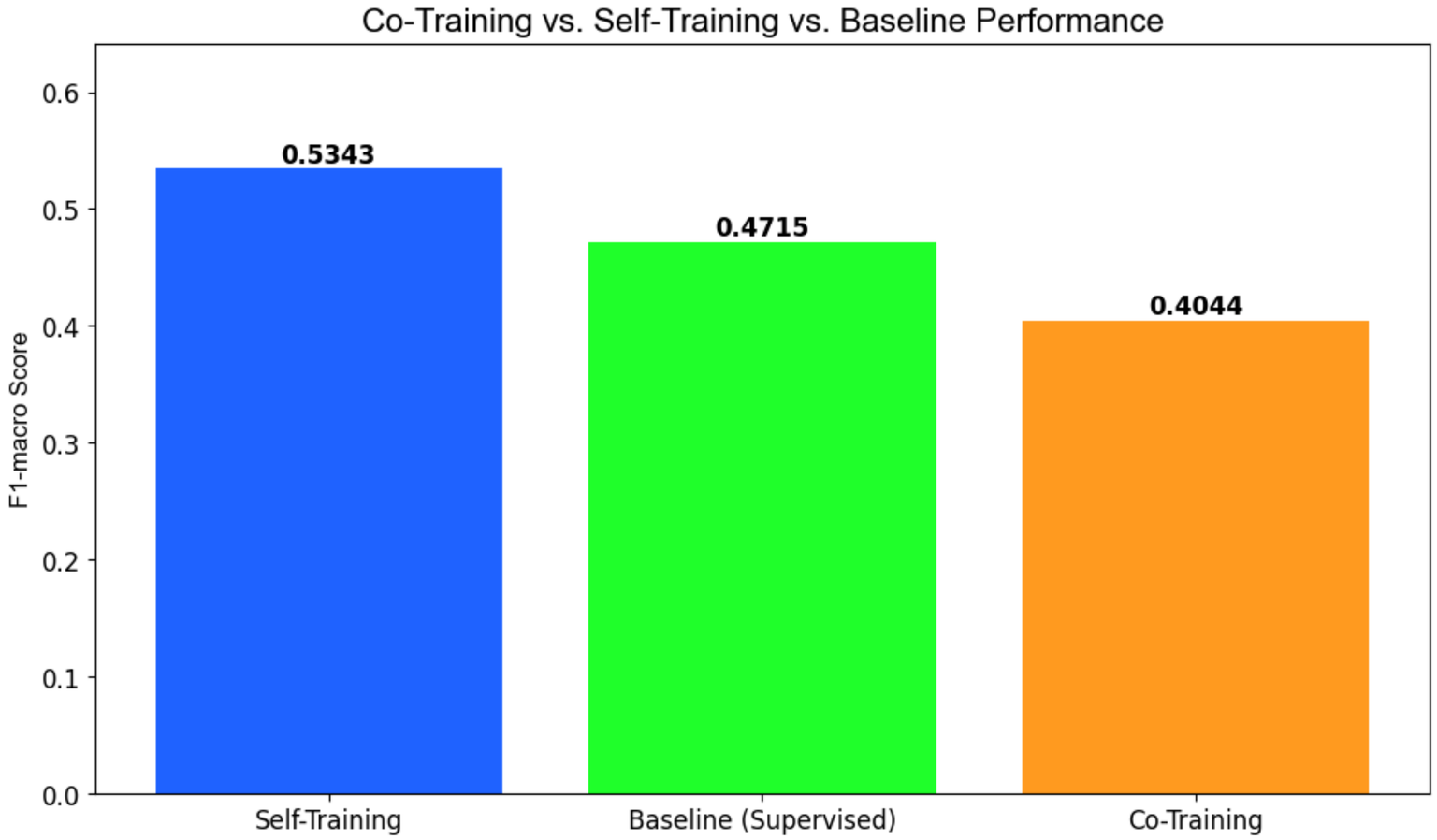
Mô hình bị "overfitting"



KẾT QUẢ THỰC NGHIỆM

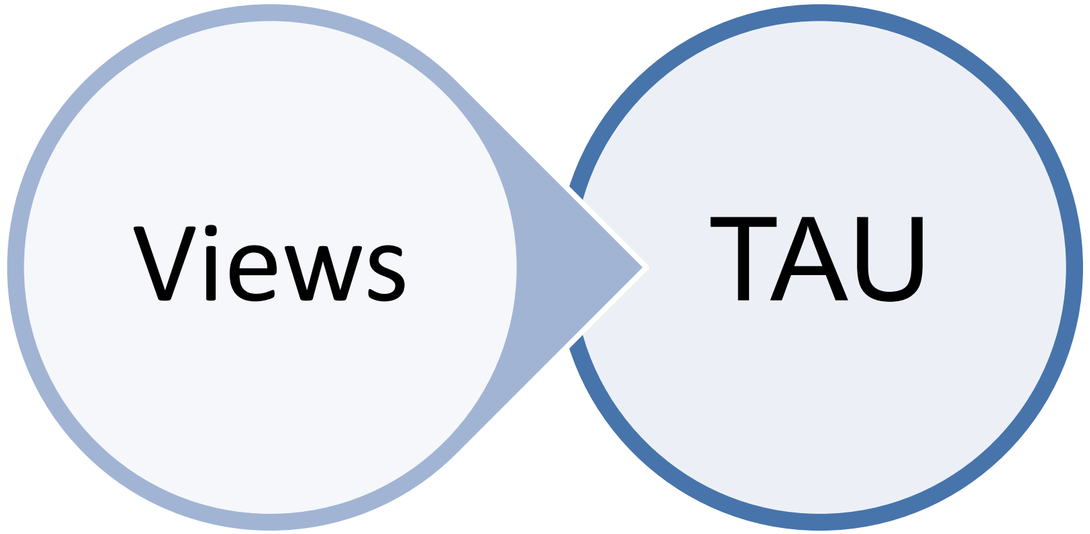
- File: 05_semi_co_training.ipynb
- Để theo thông số mặc định thì mô hình bị "overfitting"
- Để theo thông số mặc định thì kết quả của thuật toán Co-Training thấp nhất trong 3 thuật toán

Chúng ta sẽ tác động đến các chỉ số như thế nào để kết quả của thuật toán Co-Training có thể tốt hơn?



KẾT QUẢ THỰC NGHIỆM

- File: 05_semi_co_training.ipynb
- Để theo thông số mặc định thì mô hình bị "overfitting"
- Để theo thông số mặc định thì kết quả của thuật toán Co-Training thấp nhất trong 3 thuật toán



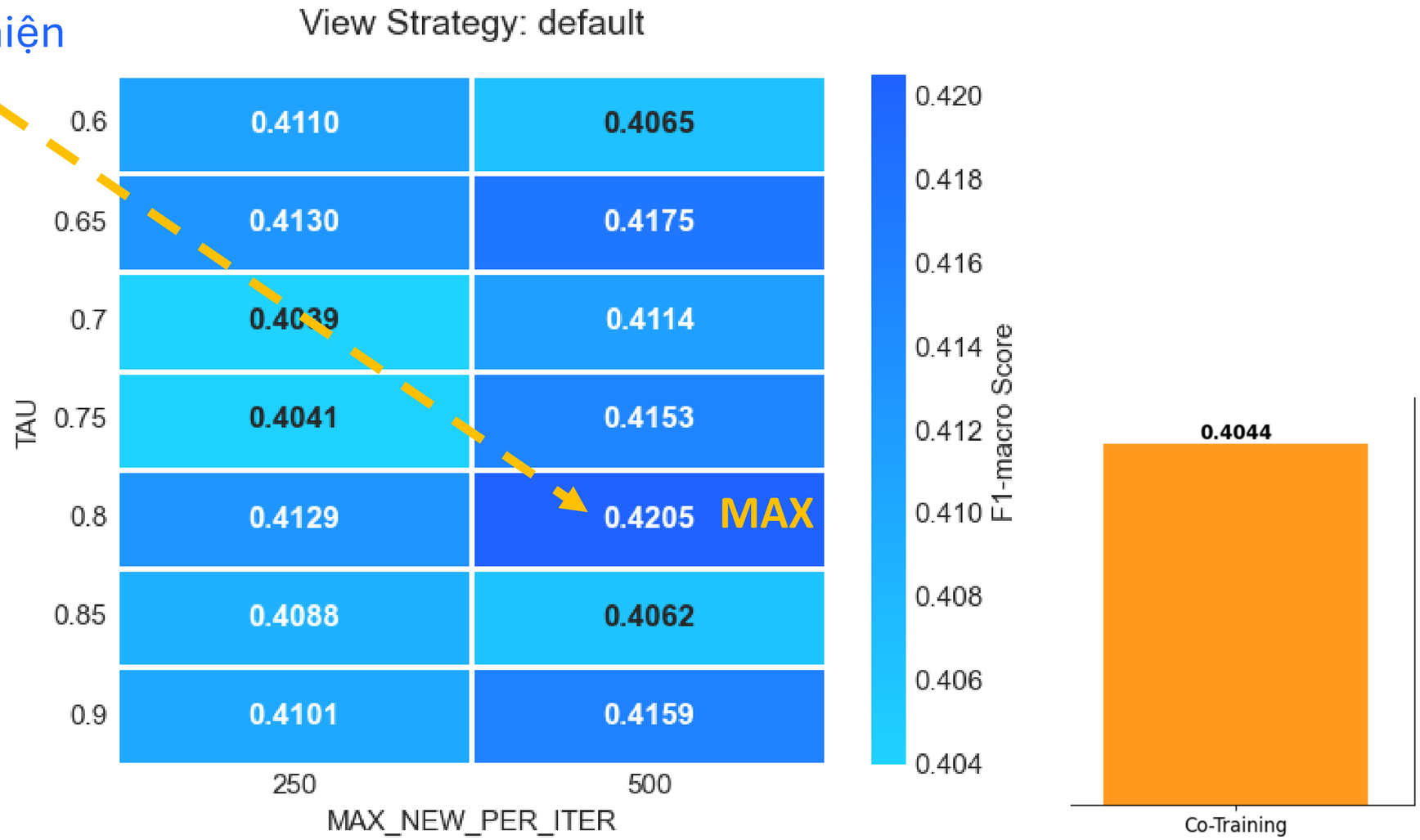
KẾT QUẢ THỰC NGHIỆM

- File: 05_semi_co_training.ipynb
- Thử nghiệm kết hợp TAU (*) MAX_NEW_PER_ITER_LIST (*) Views, file: 11_Question02.ipynb

```
68 TAU_LIST = [0.9, 0.85, 0.8, 0.75, 0.7, 0.65, 0.6]
69 MAX_NEW_PER_ITER_LIST = [250, 500]
70
71 VIEW_STRATEGIES = {
72     "default": {
73         "VIEW1_COLS": None,
74         "VIEW2_COLS": None,
75     },
76     "manual_weather_split": {
77         "VIEW1_COLS": manual_view1_cols,
78         "VIEW2_COLS": manual_view2_cols,
79     },
80     "pca_based_split": {
81         "VIEW1_COLS": pca_view1_cols,
82         "VIEW2_COLS": pca_view2_cols,
83     }
84 }
```

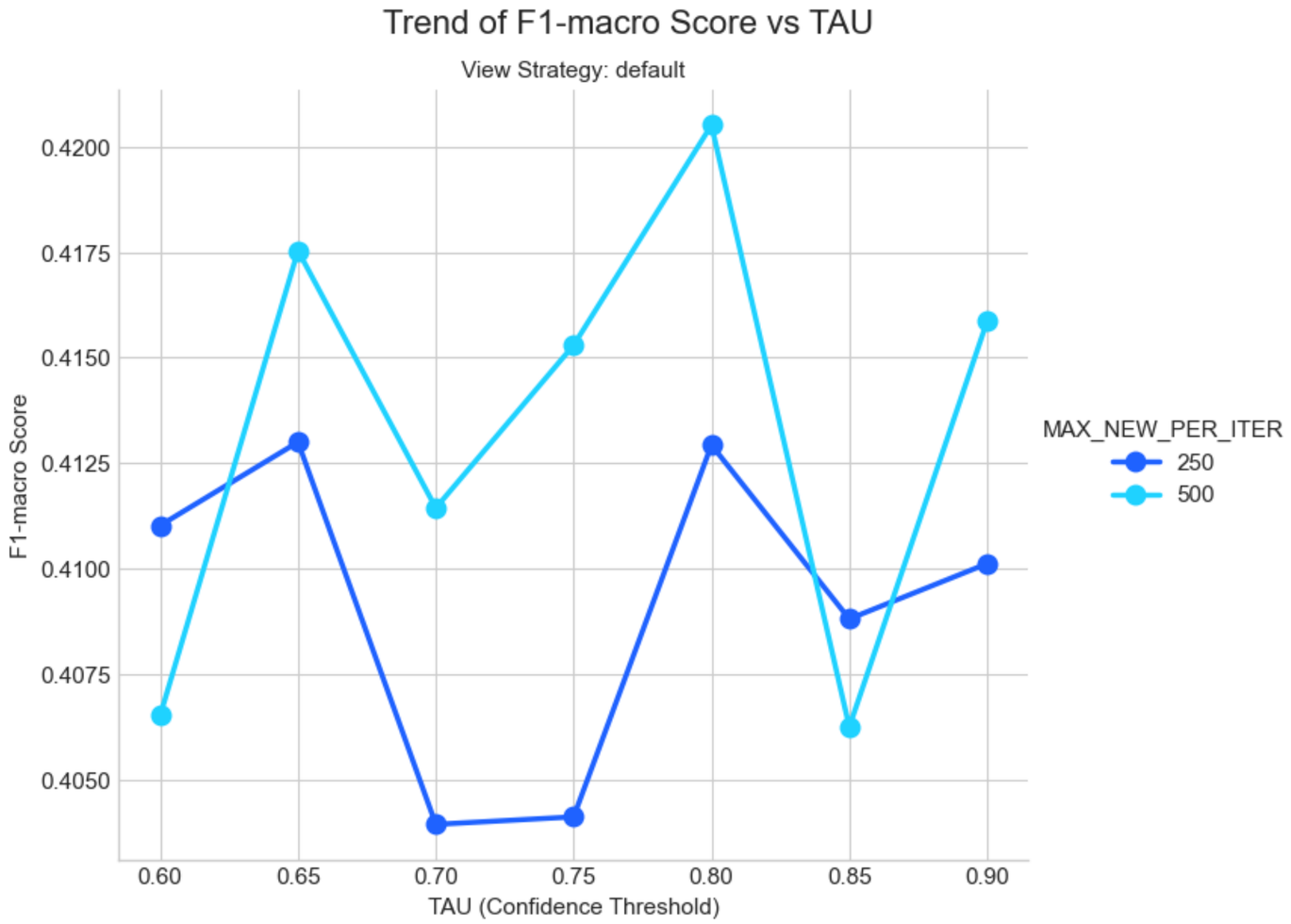
KẾT QUẢ THỰC NGHIỆM

- File: 05_semi_co_training.ipynb
- Thử nghiệm kết hợp TAU (*) MAX_NEW_PER_ITER_LIST (*) Views, file: 11_Question02.ipynb
- Kết quả đầu ra có sự cải thiện



KẾT QUẢ THỰC NGHIỆM

- File: 05_semi_co_training.ipynb
- Thử nghiệm kết hợp TAU (*) MAX_NEW_PER_ITER (*) Views, file: 11_Question02.ipynb
- Kết quả đầu ra có sự cải thiện
- Khi tăng chỉ số MAX_NEW_PER_ITEM thì 71.42% (5/7) trường hợp cho kết quả chỉ số F1-macro lớn hơn



KẾT QUẢ THỰC NGHIỆM

- File: 05_semi_co_training.ipynb
- Thử nghiệm trường hợp chỉ thay đổi mỗi thông số TAU, file: 11_Question02.ipynb
- Giá trị TAU bằng 0.75 hoặc 0.85 mang lại kết quả chỉ số F1-score tốt nhất

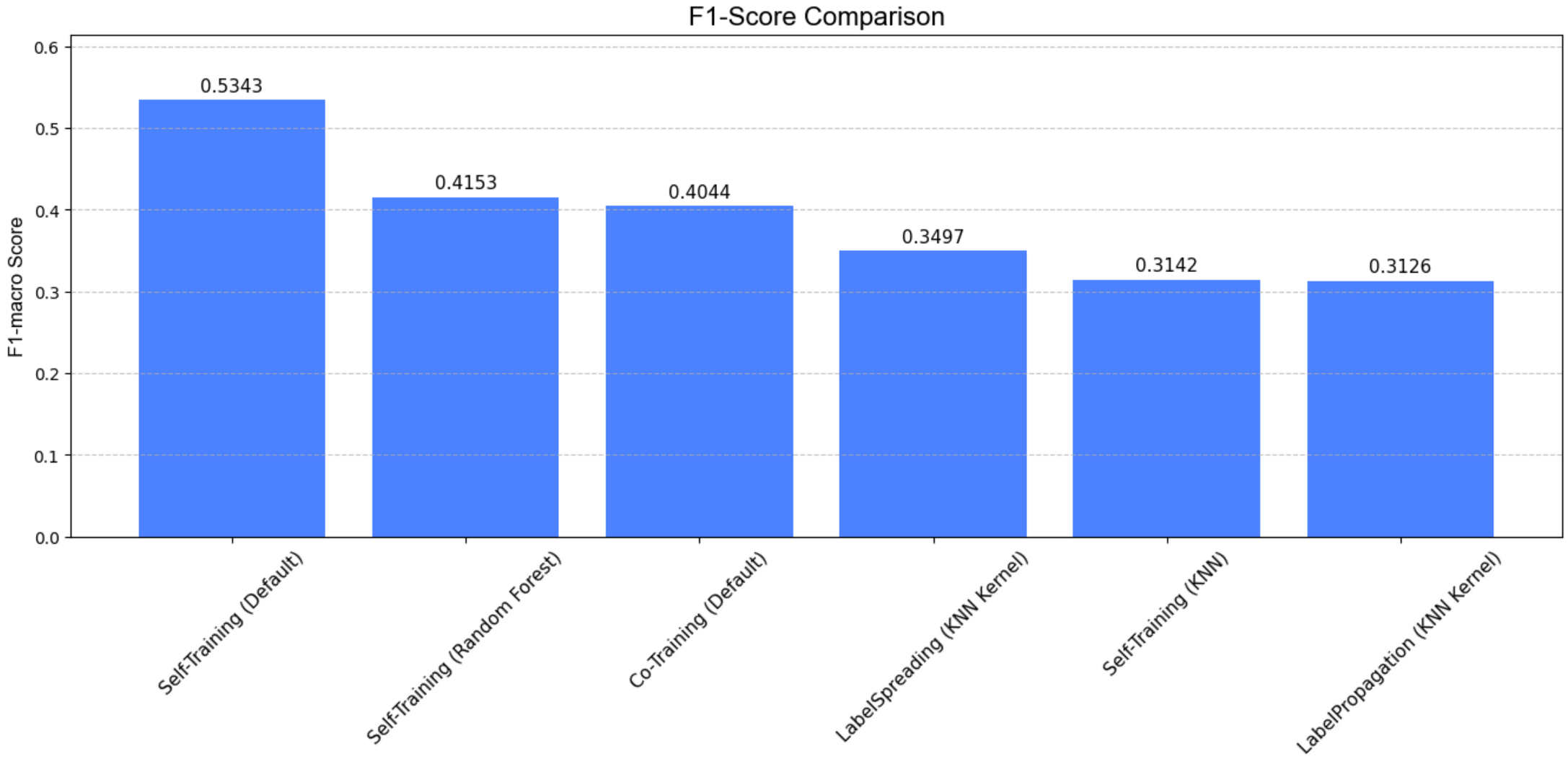

```
# Danh sách các giá trị TAU cần thử nghiệm
TAU_LIST = [0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95]

# Tham số cố định cho Co-Training
MAX_ITER = 10
MAX_NEW_PER_ITER = 500
MIN_NEW_PER_ITER = 20
VAL_FRAC = 0.20
```

KẾT QUẢ THỰC NGHIỆM

❖ File: 16_Question06.ipynb

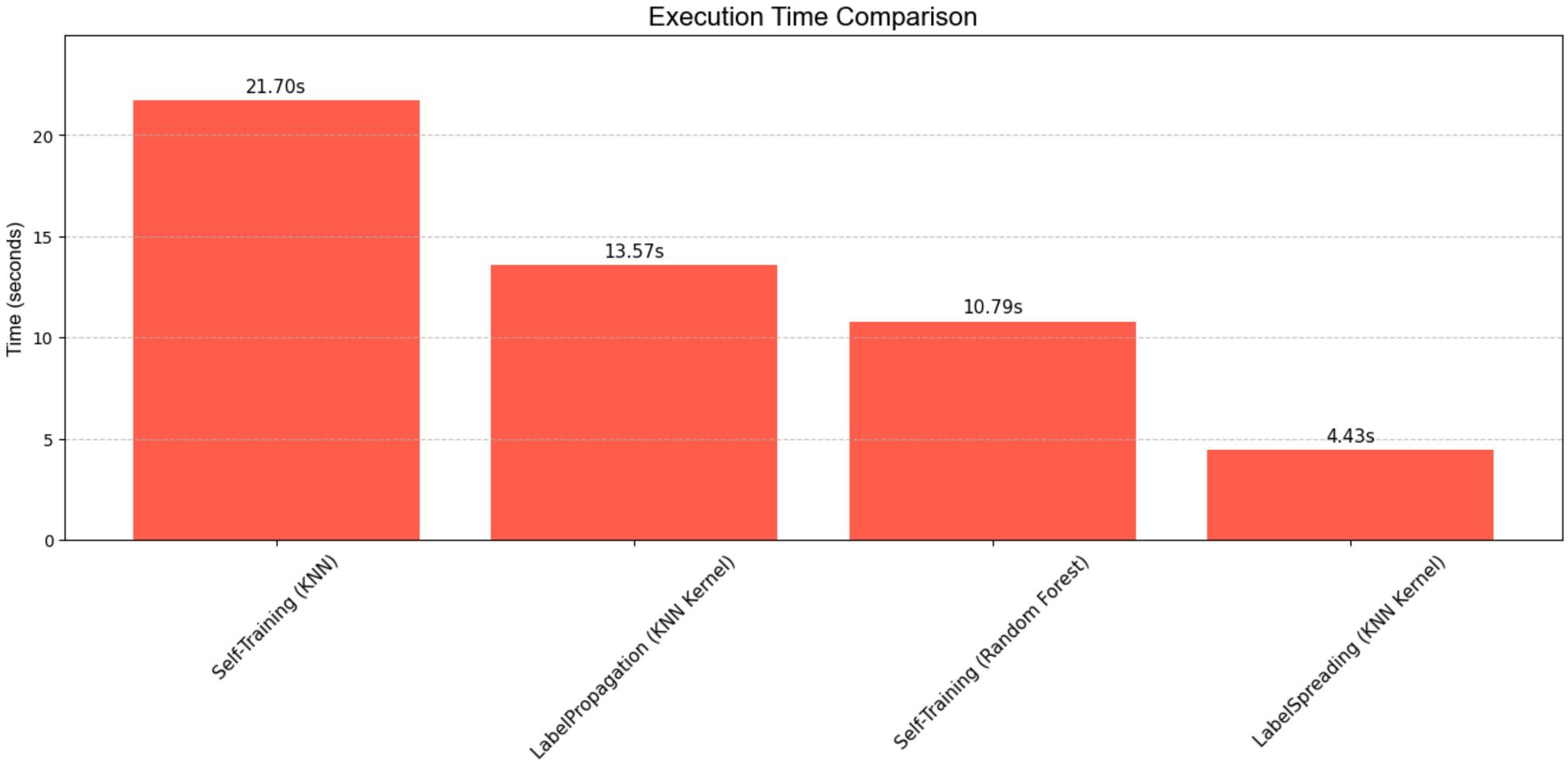
❖ Thuật toán Self-Training (với thông số mặc định) vẫn cho kết quả chỉ số F1-marco tốt nhất



KẾT QUẢ THỰC NGHIỆM

❖ File: 16_Question06.ipynb

❖ Thuật toán Self-Training (với thông số mặc định) vẫn cho kết quả chỉ số F1-marco tốt nhất



ĐẠI NAM
UNIVERSITY

Thank You