

2024

PRESENTATION

Adult Cencus Income Prediction Model

Hoang Trung - Data Team

Tổng quan về dữ liệu



Tổng quan về Dataset

Dữ liệu được lấy từ Kaggle theo link sau: [Link](#)

Name	Type	Compressed size
adult.csv	Microsoft Excel Comma S...	460 KB
adult.test.csv	Microsoft Excel Comma S...	231 KB

File adult là dữ liệu train, adult.test là dữ liệu Test (**Không có tiêu đề**)

Tổng cộng có 15 cột dữ liệu và 45.232 dòng dữ liệu. Không có dữ liệu null.

Dữ liệu trên có các vấn đề sau:

- Các dữ liệu chữ trong Data chứa rất nhiều dấu cách thừa.
- Trong các cột dữ liệu tuy không có dữ liệu null nhưng chứa nhiều giá trị “?”

```
<class 'pandas.core.frame.DataFrame'>
Index: 45232 entries, 0 to 16280
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype  
---  --  
 0   Age              45232 non-null   int64  
 1   Workclass        45232 non-null   object  
 2   Final Weight    45232 non-null   int64  
 3   Education        45232 non-null   object  
 4   EducationNum    45232 non-null   int64  
 5   Marital Status  45232 non-null   object  
 6   Occupation       45232 non-null   object  
 7   Relationship     45232 non-null   object  
 8   Race             45232 non-null   object  
 9   Gender           45232 non-null   object  
 10  Capital Gain   45232 non-null   int64  
 11  capital loss   45232 non-null   int64  
 12  Hours per Week 45232 non-null   int64  
 13  Native Country  45232 non-null   object  
 14  Income           45232 non-null   object  
dtypes: int64(6), object(9)
memory usage: 5.5+ MB
```



Mục tiêu: Tạo lập mô hình ước tính thu nhập của 1 người có vượt mức 50K USD/năm hay không?

Xử lý dữ liệu

Ta sẽ tiến hành theo các bước sau:

- Loại bỏ các dấu cách thừa trong từng cột dạng chữ.
- Thay thế toàn bộ các giá trị “?” trong bảng thành giá trị null.
- Thay thế và loại bỏ các giá trị null trước khi tiến hành phân tích dữ liệu.
- Tiến hành bỏ dấu “.” với cột Income.

Lọc ra các giá trị null cột Occupation và không bị null cột Workclass
Với các giá trị này ta sẽ thay thế bằng “No-occupation”

```
df[df['Occupation'].isnull() & ~df['Workclass'].isnull()]
```

Executed at 2024.04.17 22:40:41 in 149ms

	Age	Workclass	Final Weight	Education	EducationNum	Marital Status	Occupation	Relationship	Race	Gender	
5361	18	Never-worked	206359	10th		6	Never-married	NaN	Own-child	White	Male
10845	23	Never-worked	188535	7th-8th		4	Divorced	NaN	Not-in-family	White	Male
14772	17	Never-worked	237272	10th		6	Never-married	NaN	Own-child	White	Male
20337	18	Never-worked	157131	11th		7	Never-married	NaN	Own-child	White	Female
23232	20	Never-worked	462294	Some-college		10	Never-married	NaN	Own-child	Black	Male
32304	30	Never-worked	176673	HS-grad		9	Married-civ-spouse	NaN	Wife	Black	Female
32314	18	Never-worked	153663	Some-college		10	Never-married	NaN	Own-child	White	Male
8785	17	Never-worked	131593	11th		7	Never-married	NaN	Own-child	Black	Female
11607	20	Never-worked	273905	HS-grad		9	Married-spouse-absent	NaN	Other-relative	White	Male
13898	18	Never-worked	162908	11th		7	Never-married	NaN	Own-child	White	Male

Occupation	5.75
Workclass	5.73
Native Country	1.75
Age	0.00
Final Weight	0.00
Education	0.00
EducationNum	0.00
Marital Status	0.00
Relationship	0.00
Race	0.00
Occupation	2809
Workclass	2799
Native Country	857
Age	0
Final Weight	0
Education	0
EducationNum	0
Marital Status	0
Relationship	0
Race	0

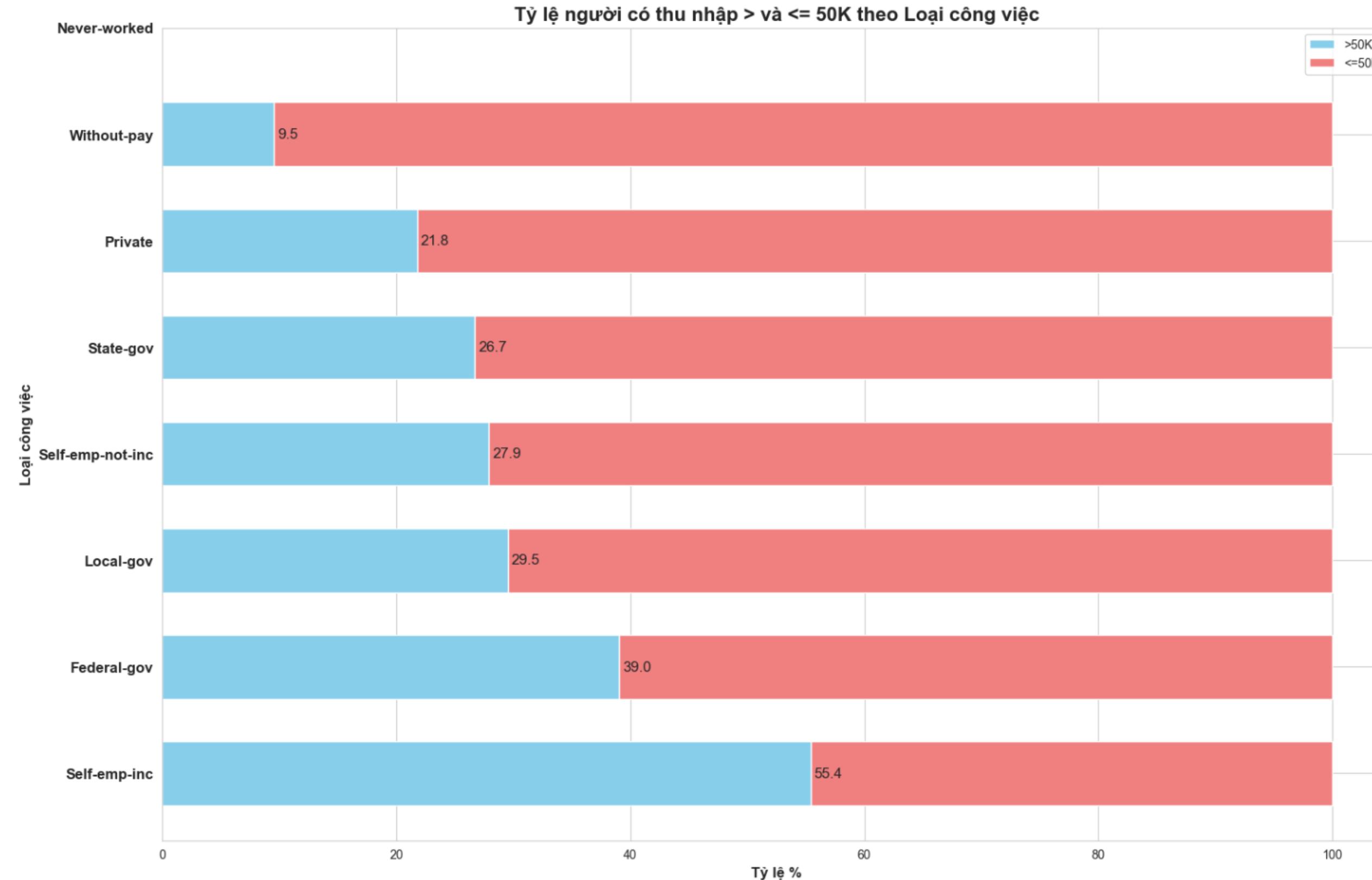
Danh sách các feature trong dataset

- **Age**: Tuổi của người tham gia khảo sát
- **Workclass**: Loại công việc của người tham gia khảo sát
- **Final Weight**: Trọng số trên các tệp CPS được kiểm soát theo ước tính độc lập về dân số phi thể chế dân sự của Hoa Kỳ
- **Education**: Trình độ học vấn cao nhất đã hoàn thành
- **EducationNum**: Số năm học đã hoàn thành
- **Marital Status**: Tình trạng hôn nhân
- **Occupation**: Công việc của người tham gia khảo sát
- **Relationship**: Tình trạng mối quan hệ
- **Race**: Chủng tộc của người tham gia khảo sát
- **Gender**: Giới tính của người tham gia khảo sát
- **Capital Gain**: Số tiền vốn, lợi nhuận tài chính gia tăng
- **capital loss**: Số tiền vốn, lợi nhuận tài chính sụt giảm
- **Hours per Week**: Số giờ làm việc hàng tuần
- **Native Country**: Quê hương của người được khảo sát
- **Income**: Thu nhập của người được khảo sát (Gồm 2 mức là $\leq 50K$ và $> 50K$)

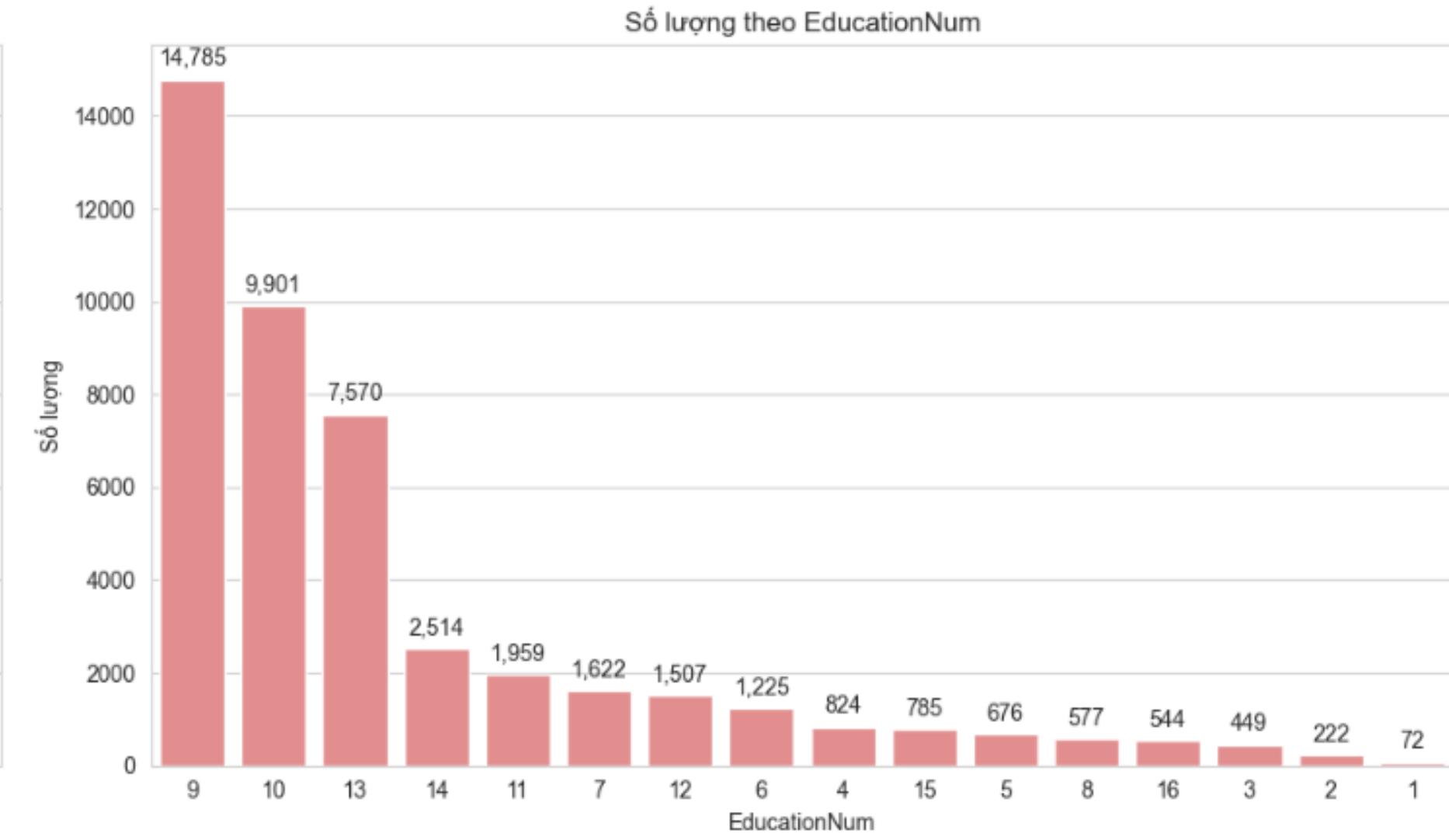
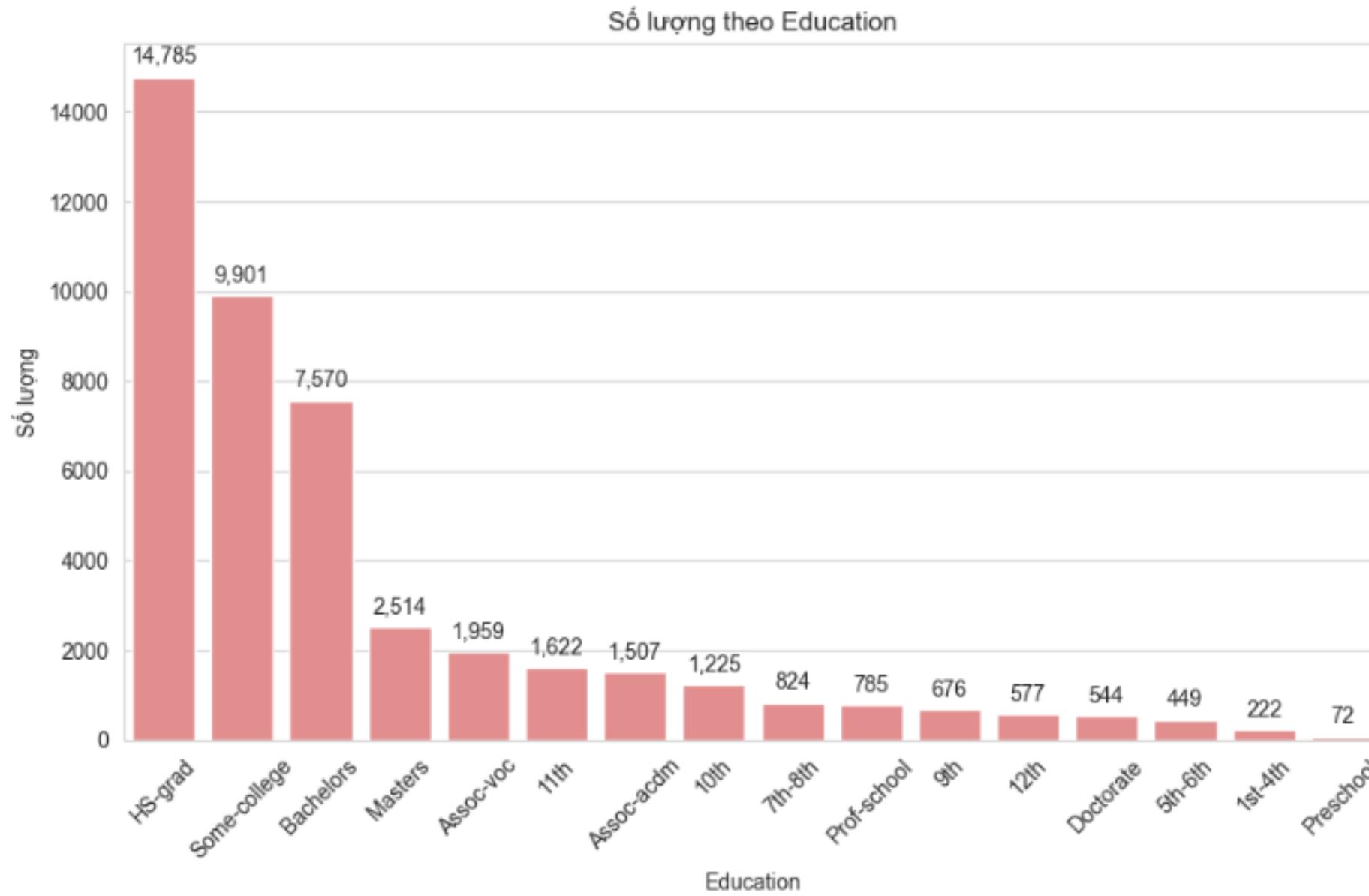
EDA dữ liệu

Workclass

Những cá nhân làm chủ doanh nghiệp (**Self-emp-inc**) và làm việc cho liên bang (**Federal-gov**) thường có mức thu nhập **tốt hơn so với các loại công việc khác**



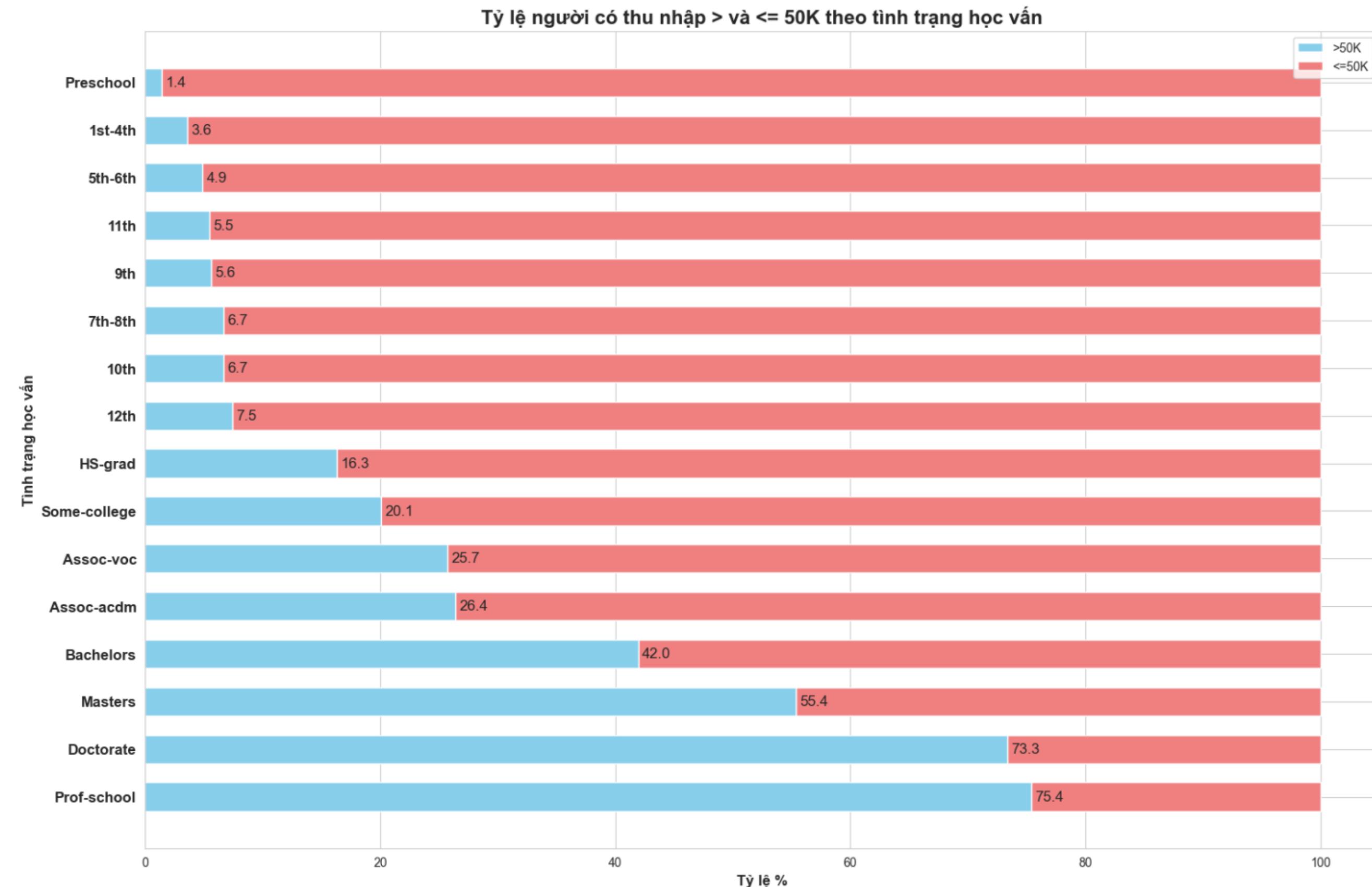
Education - EducationNum



2 cột dữ liệu Education và EducationNum về cơ bản là 1 nên khi xây dựng model ta sẽ loại bỏ cột Education

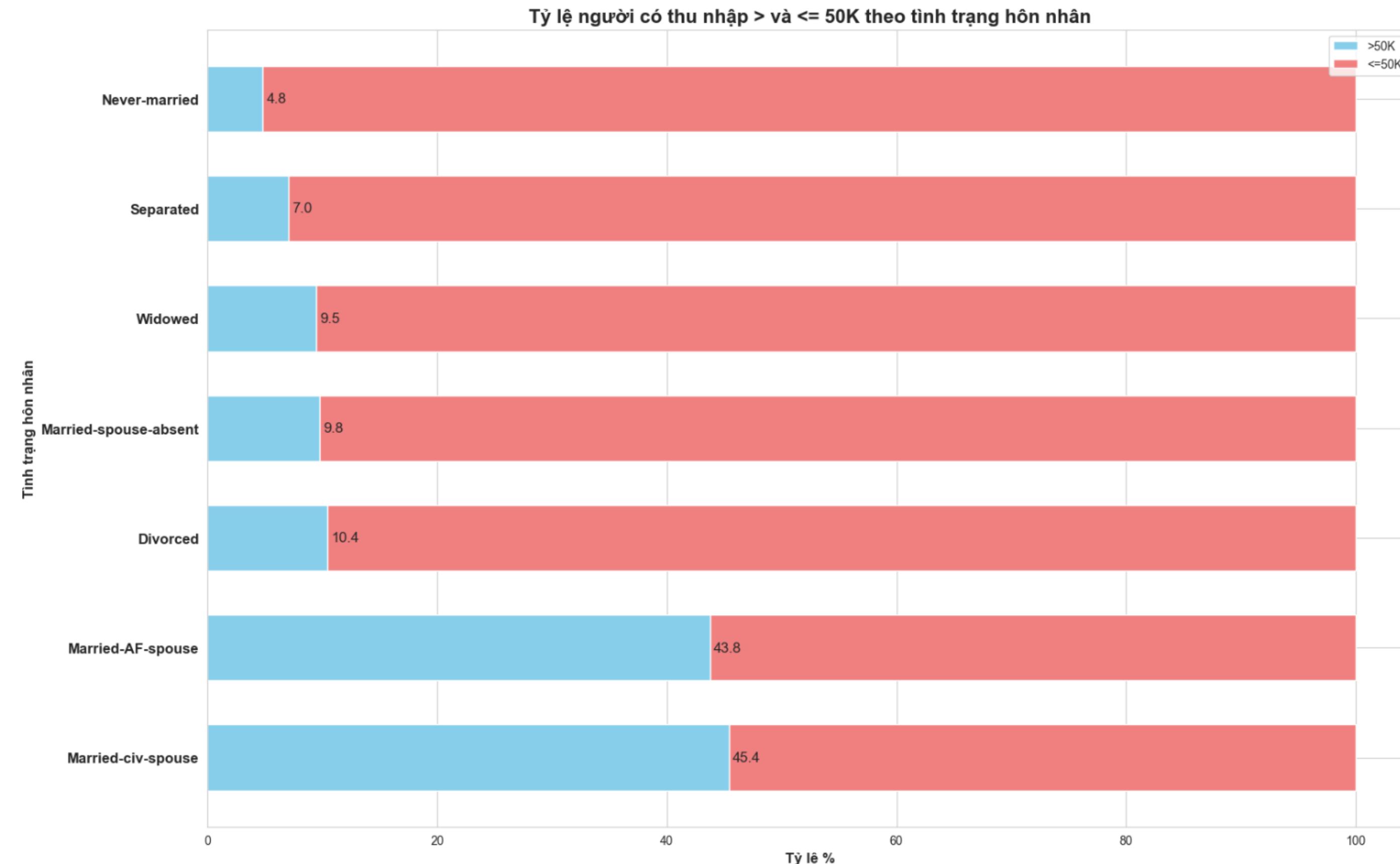
Education - EducationNum

Bạn học càng **cao** thì **tỷ lệ có mức thu nhập >=50K/năm** cũng **cao hơn**



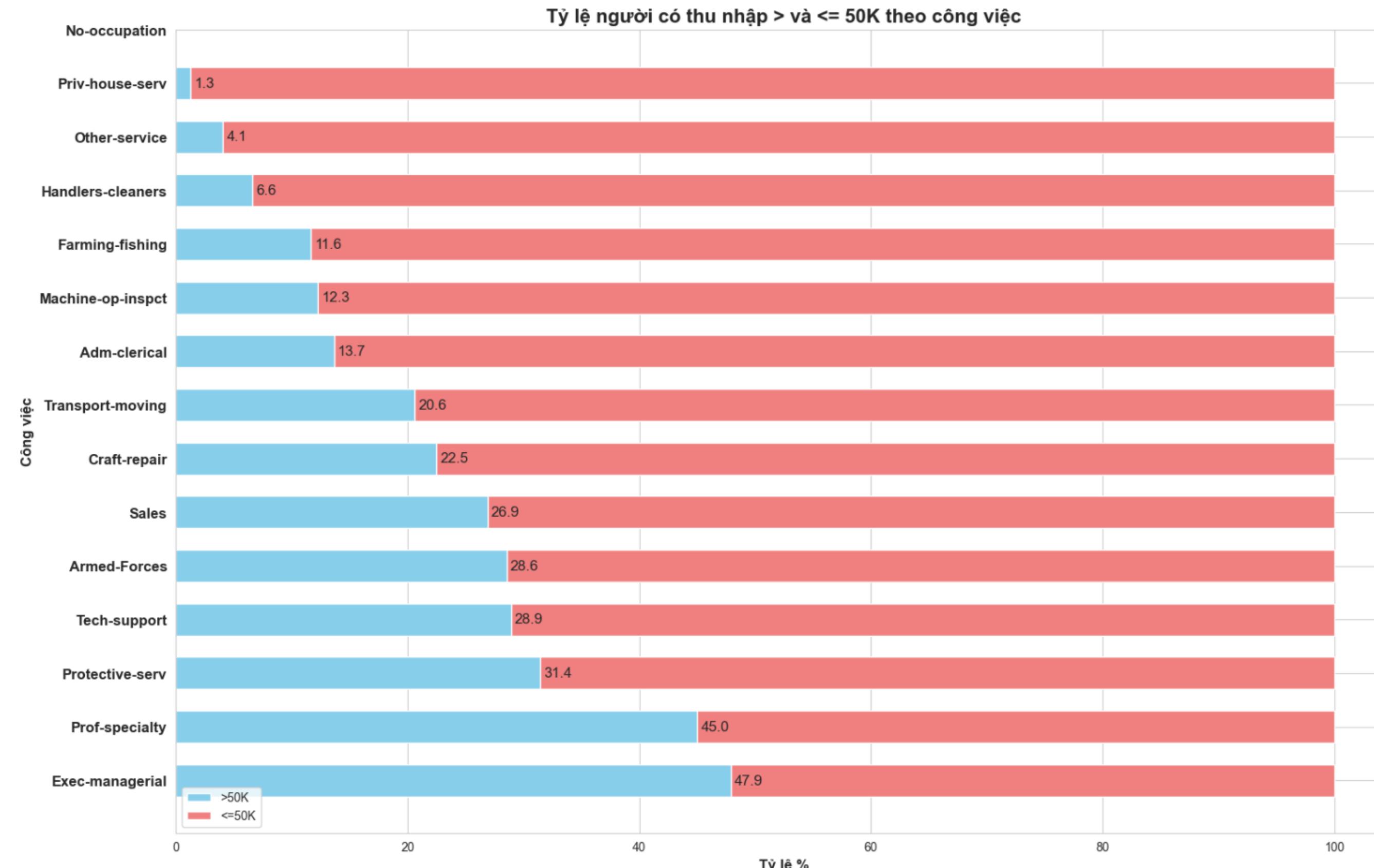
Marital Status

Nhóm có thu nhập $\geq 50K$ thường rơi vào nhóm các cá nhân **đã có gia đình**.



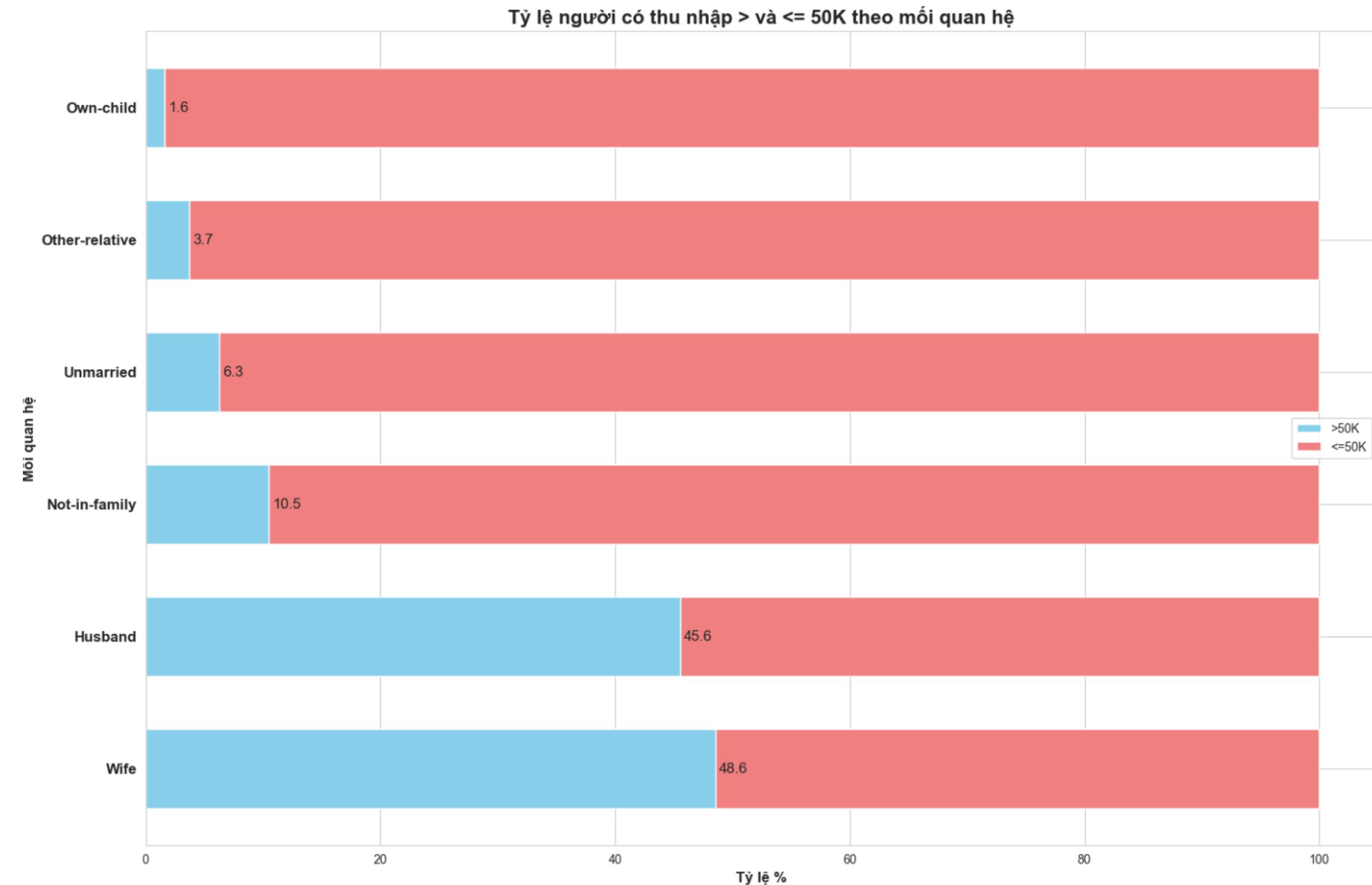
Occupation

Các cá nhân làm **quản lý (Exec-managerial)** và **chuyên gia (Prof-specialty)** thường có mức thu nhập **cao hơn**



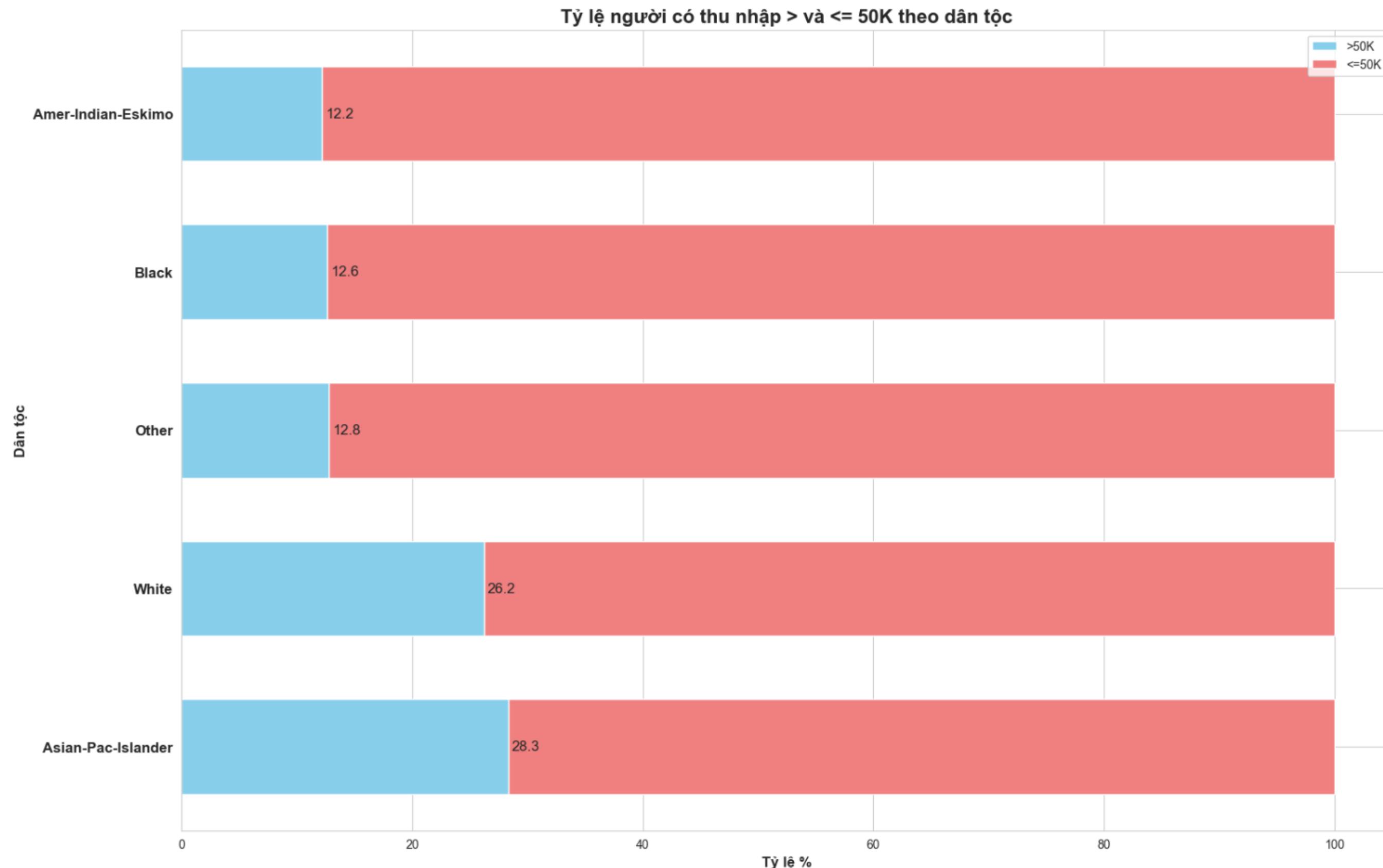
Relationship

Các cá nhân **có gia đình (Husband, Wife)** thường có thu nhập **cao hơn các nhóm khác**.



Race

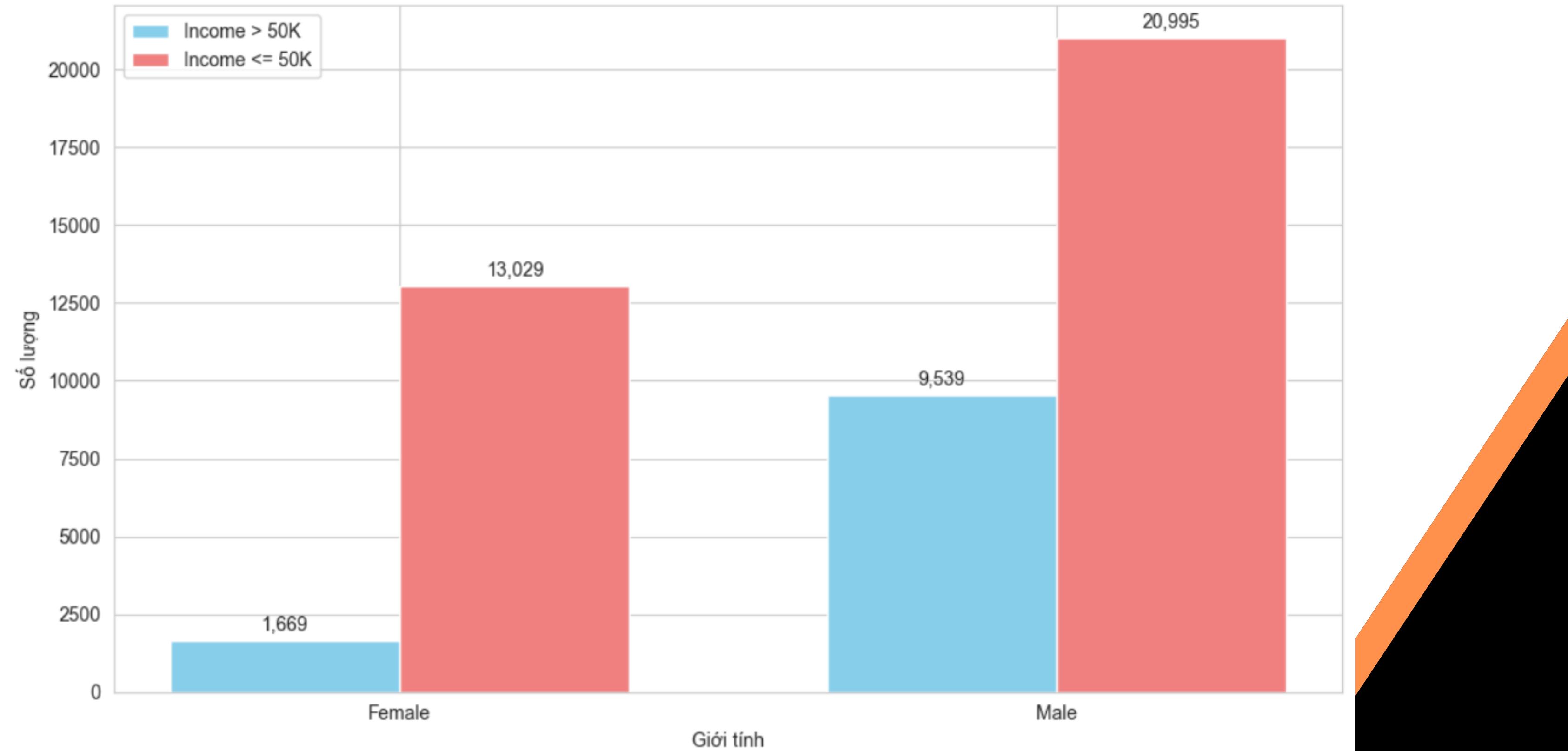
Người da trắng (White) và **người châu Á** là nhóm các dân tộc có thu nhập **cao hơn** so với các dân tộc khác.



Gender

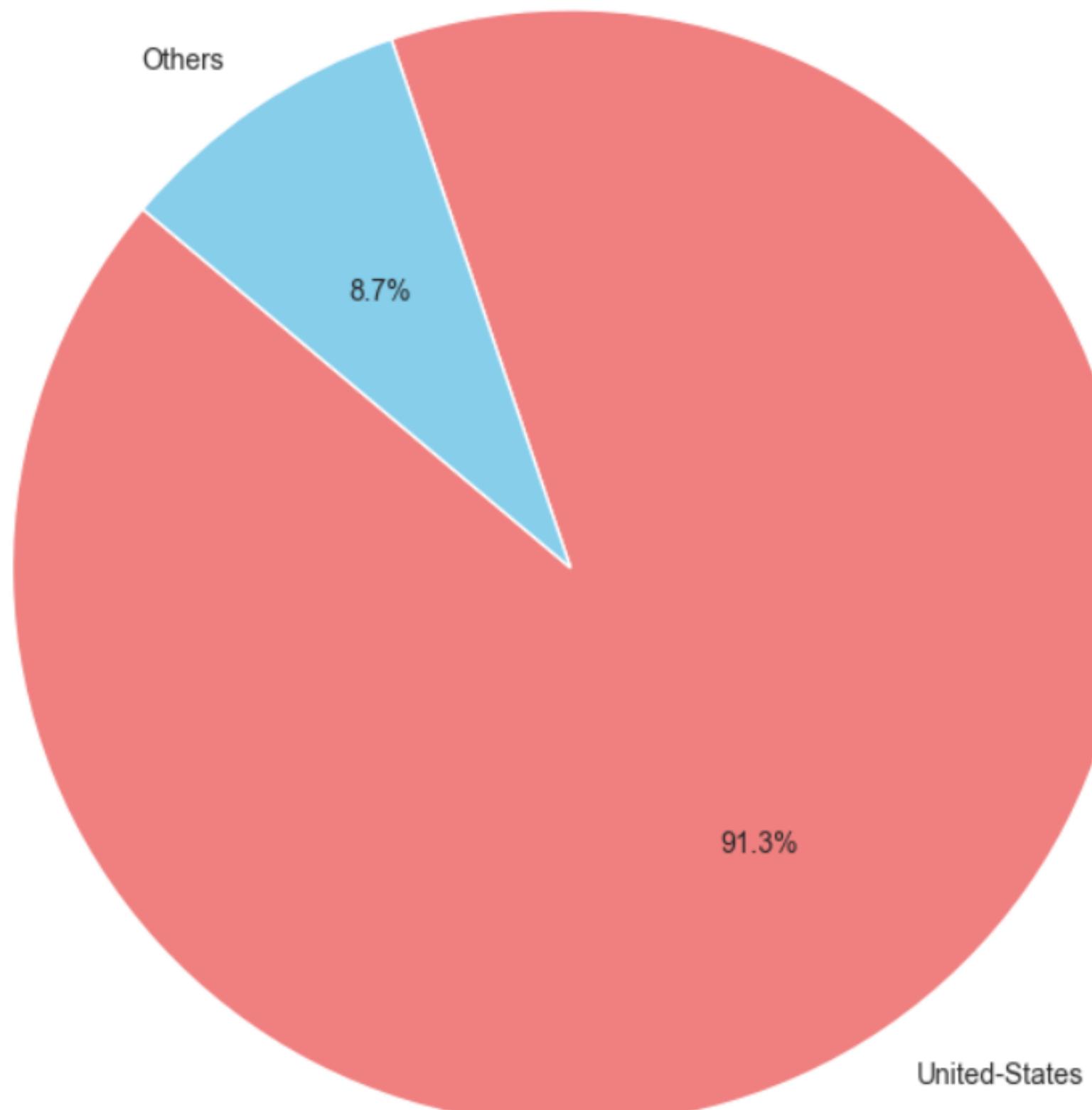
Thu nhập của **đàn ông** có xu hướng cao hơn so với **phụ nữ**

Sự phân phối thu nhập theo giới tính



Native Country

Phân phối theo quốc gia

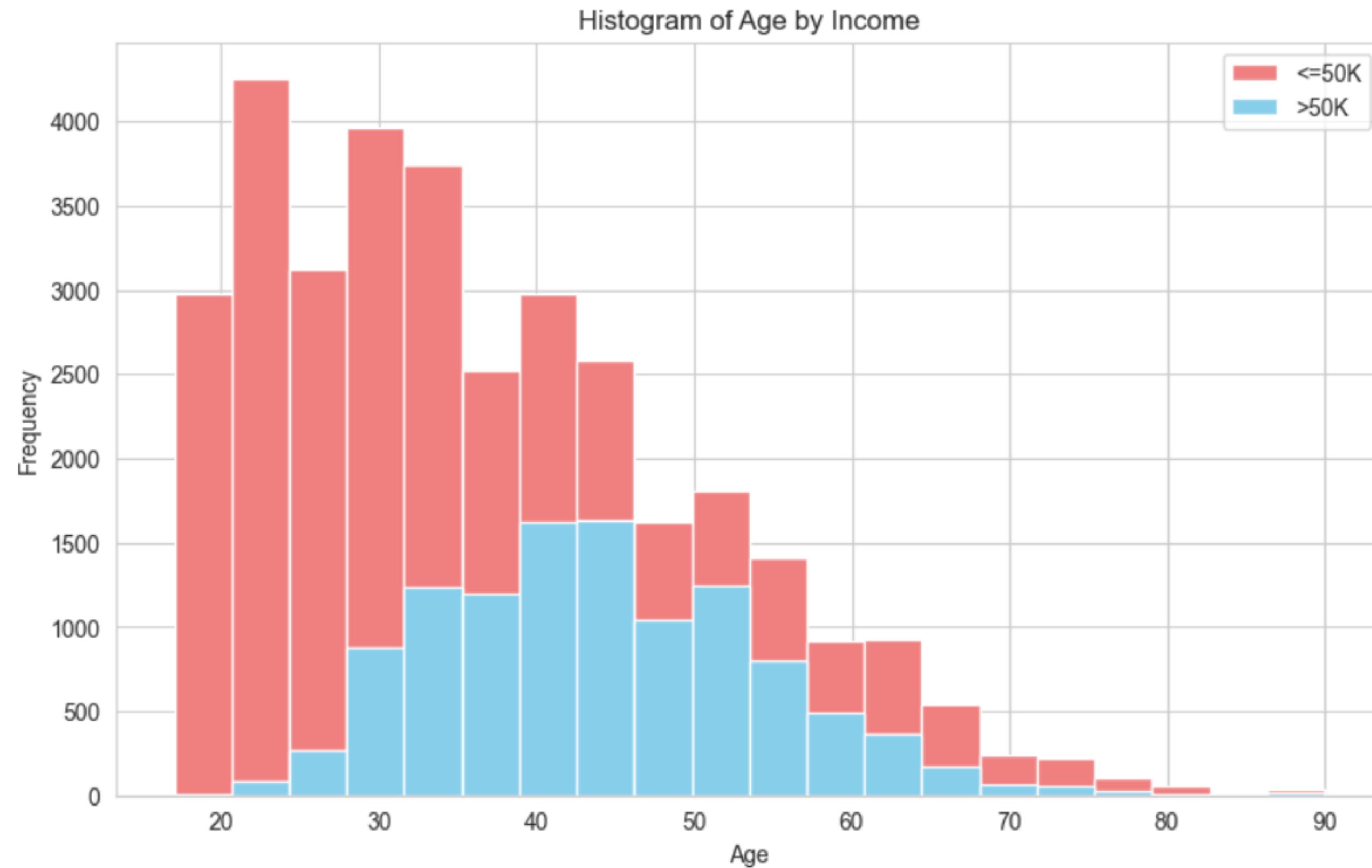


Dữ liệu trên được thu thập ở nước Mỹ nên nước Mỹ (**United-States**) chiếm đa số so với các nước khác

Native Country	total_count	% Count
United-States	41302	91.31
Mexico	903	2.00
Philippines	283	0.63
Germany	193	0.43
Puerto-Rico	175	0.39
Canada	163	0.36
El-Salvador	147	0.32
India	147	0.32
Cuba	133	0.29
England	119	0.26

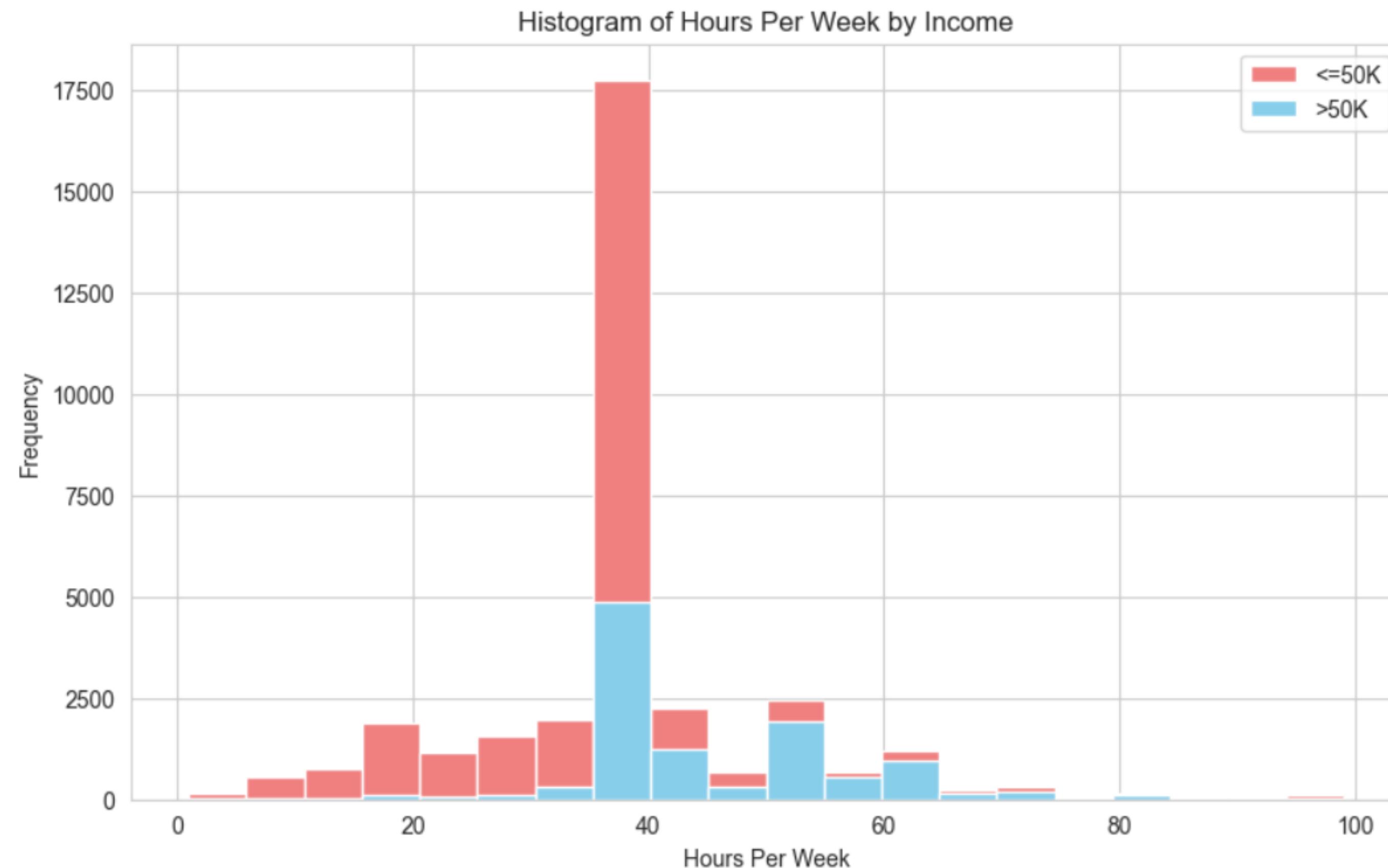
Age

Đa phần các lao động có độ tuổi lớn **từ 30 đến 53 tuổi** (thâm niên lao động nhiều năm) thì thường sẽ có **mức lương cao**



Hours Per Week

Những người lao động **làm việc từ 40 giờ trở lên** thường có thu nhập **cao hơn** các nhóm còn lại.



Xây dựng mô hình dự đoán thu nhập

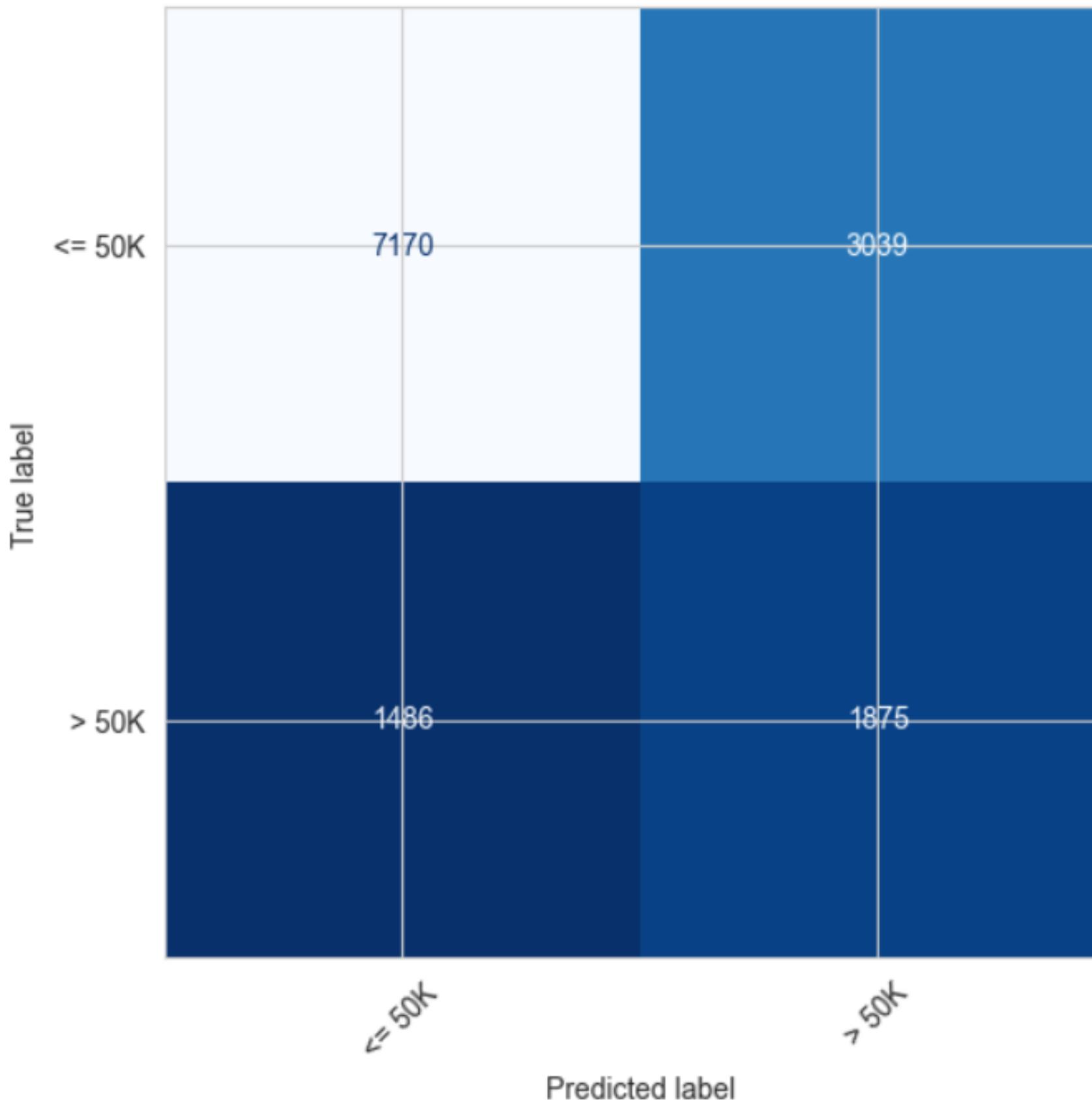


Các bước xây dựng Model

- Drop các Feature có thể ảnh hưởng đến kết quả của model (**Education, Native Country**)
- Encode dữ liệu (**get_dummies** và **map lại dữ liệu cột Income**)
- Chia tập X, y và tập train, test
- Xây dựng model dự đoán bằng các mô hình **Logistic Regression, Decision Tree Classifier, Random Forest Classifier, XGBoost Classifier, Gradient Boosting Classifier, SVC**

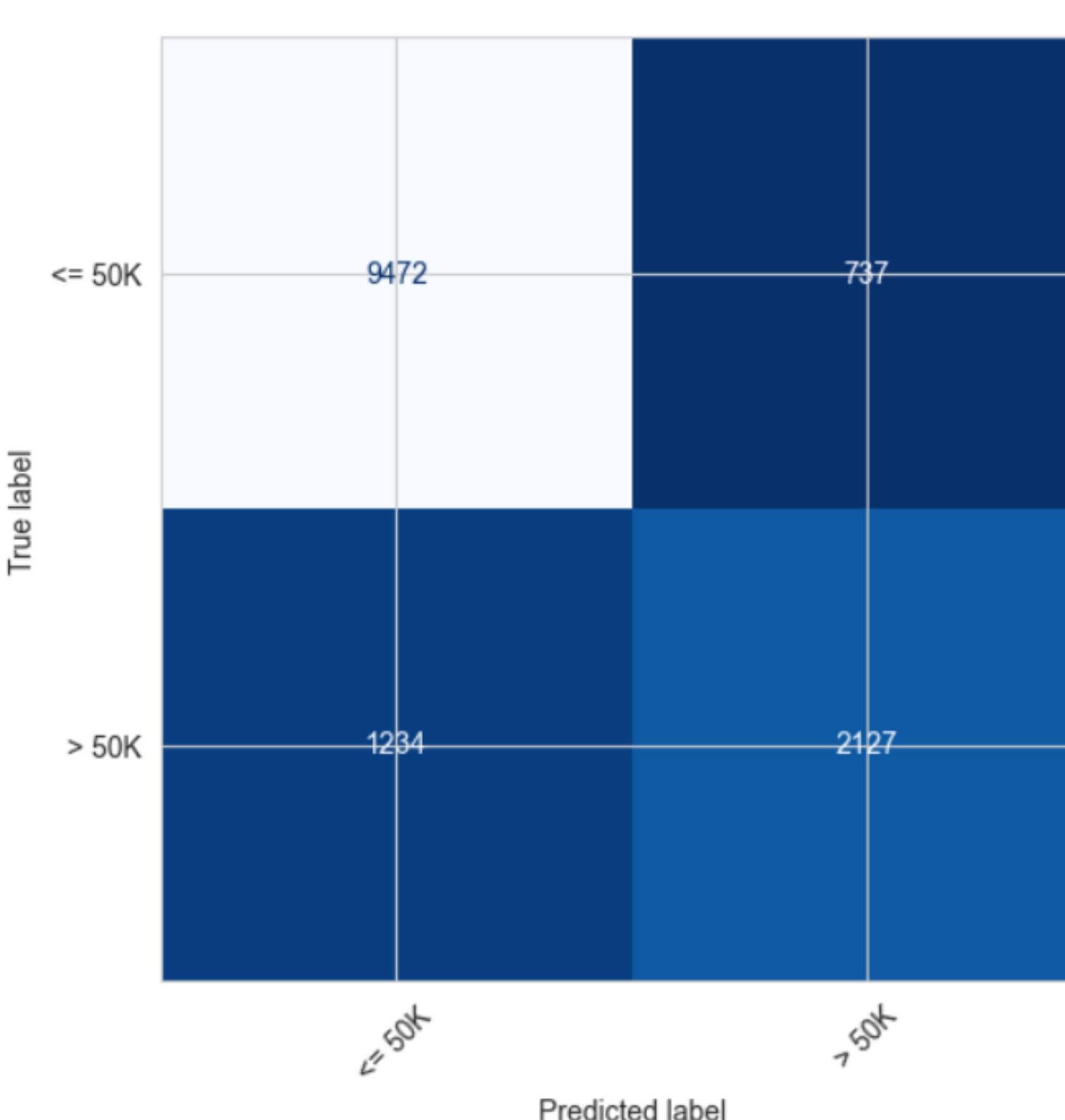


Mô hình Logistic Regression



	precision	recall	f1-score	support
0	0.83	0.70	0.76	10209
1	0.38	0.56	0.45	3361
accuracy			0.67	13570
macro avg	0.60	0.63	0.61	13570
weighted avg	0.72	0.67	0.68	13570

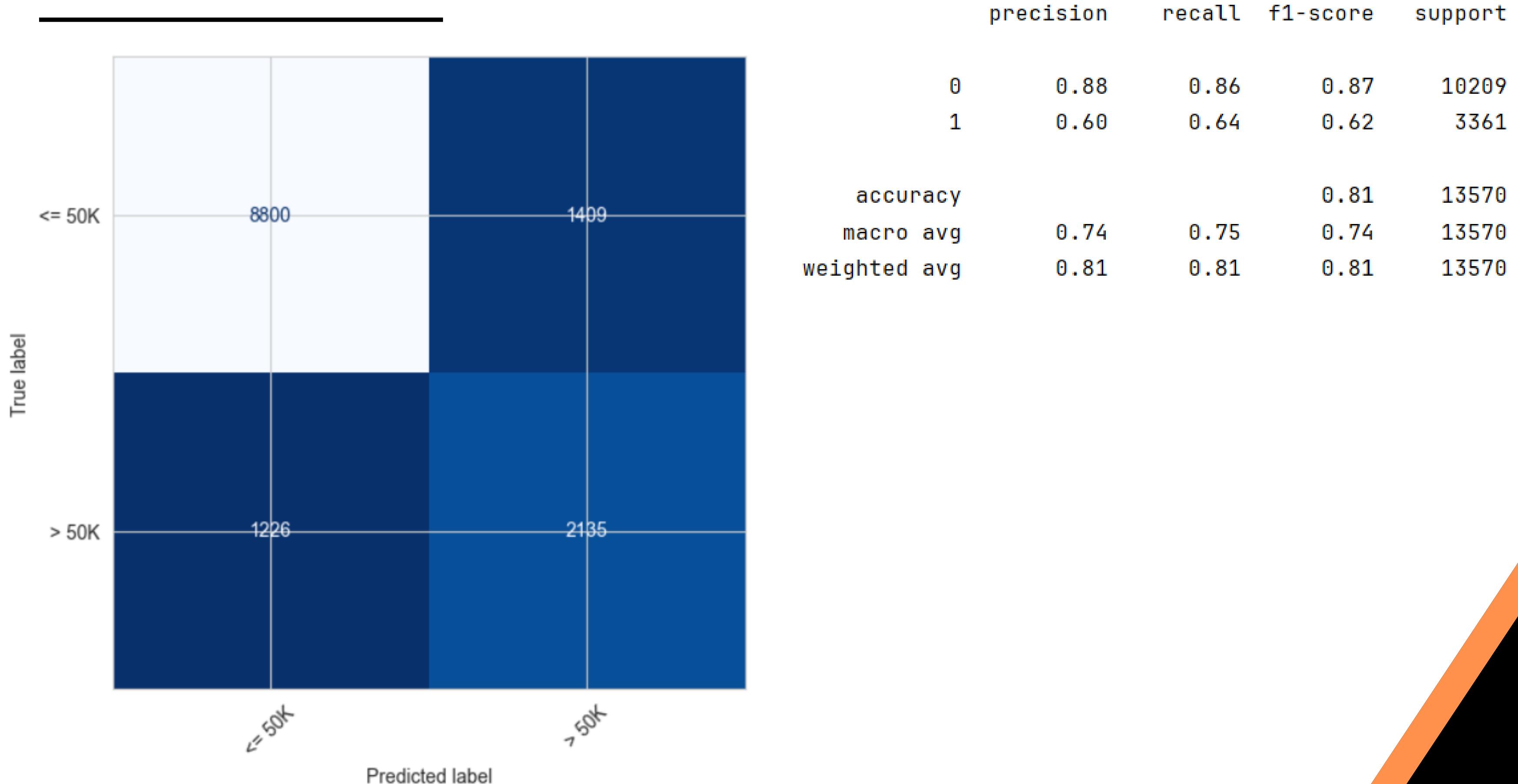
Mô hình Logistic Regression + GridSearchCV



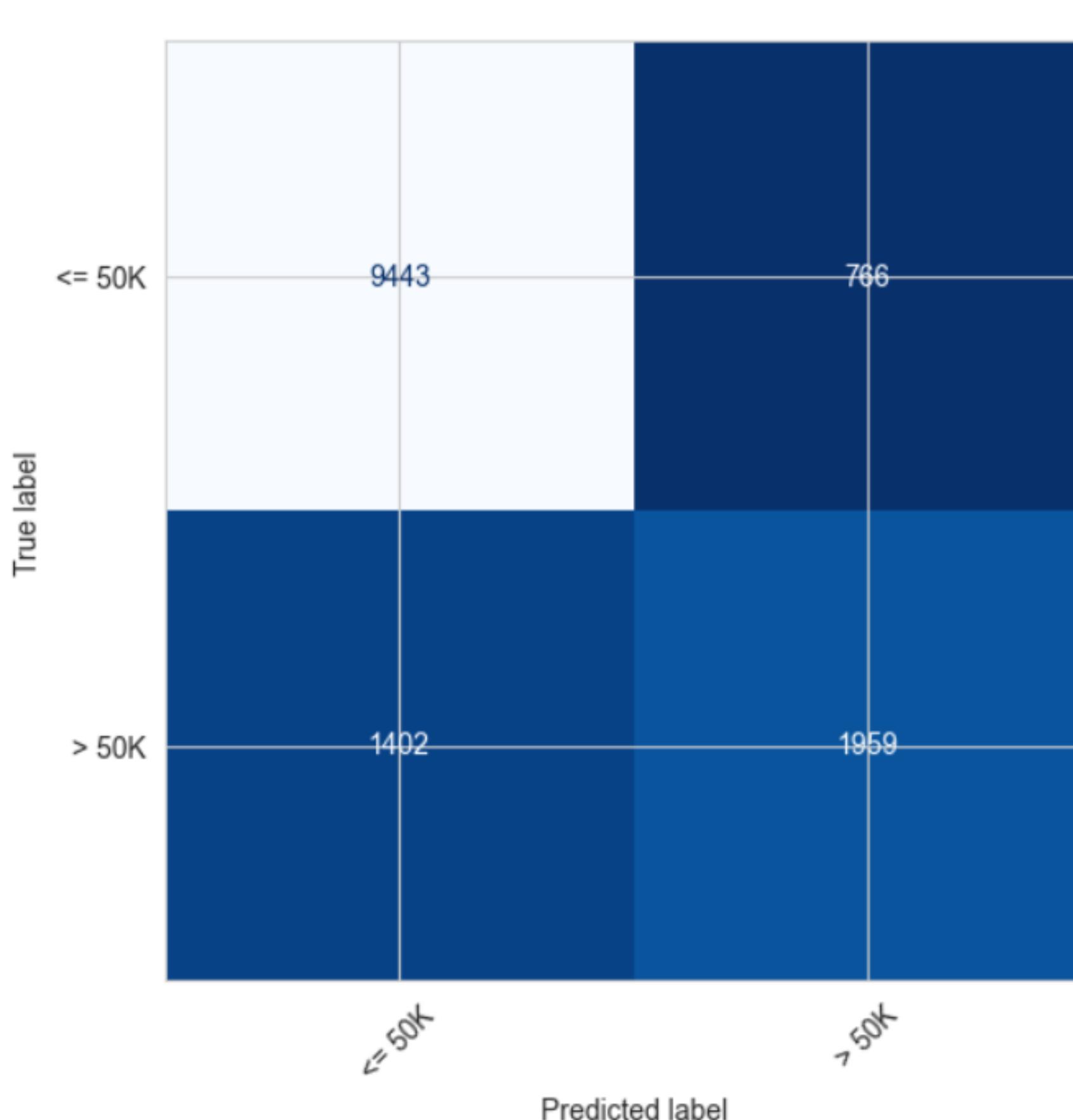
	precision	recall	f1-score	support
0	0.88	0.93	0.91	10209
1	0.74	0.63	0.68	3361
accuracy			0.85	13570
macro avg	0.81	0.78	0.79	13570
weighted avg	0.85	0.85	0.85	13570

```
param_grid = {
    'C': [78.47599703514607], #78.47599703514607
    'max_iter': [5000],
    'penalty' : ['l2'],
    'solver' : ['newton-cg'],
    'multi_class': ['multinomial']
}
```

Mô hình Decision Tree Classifier



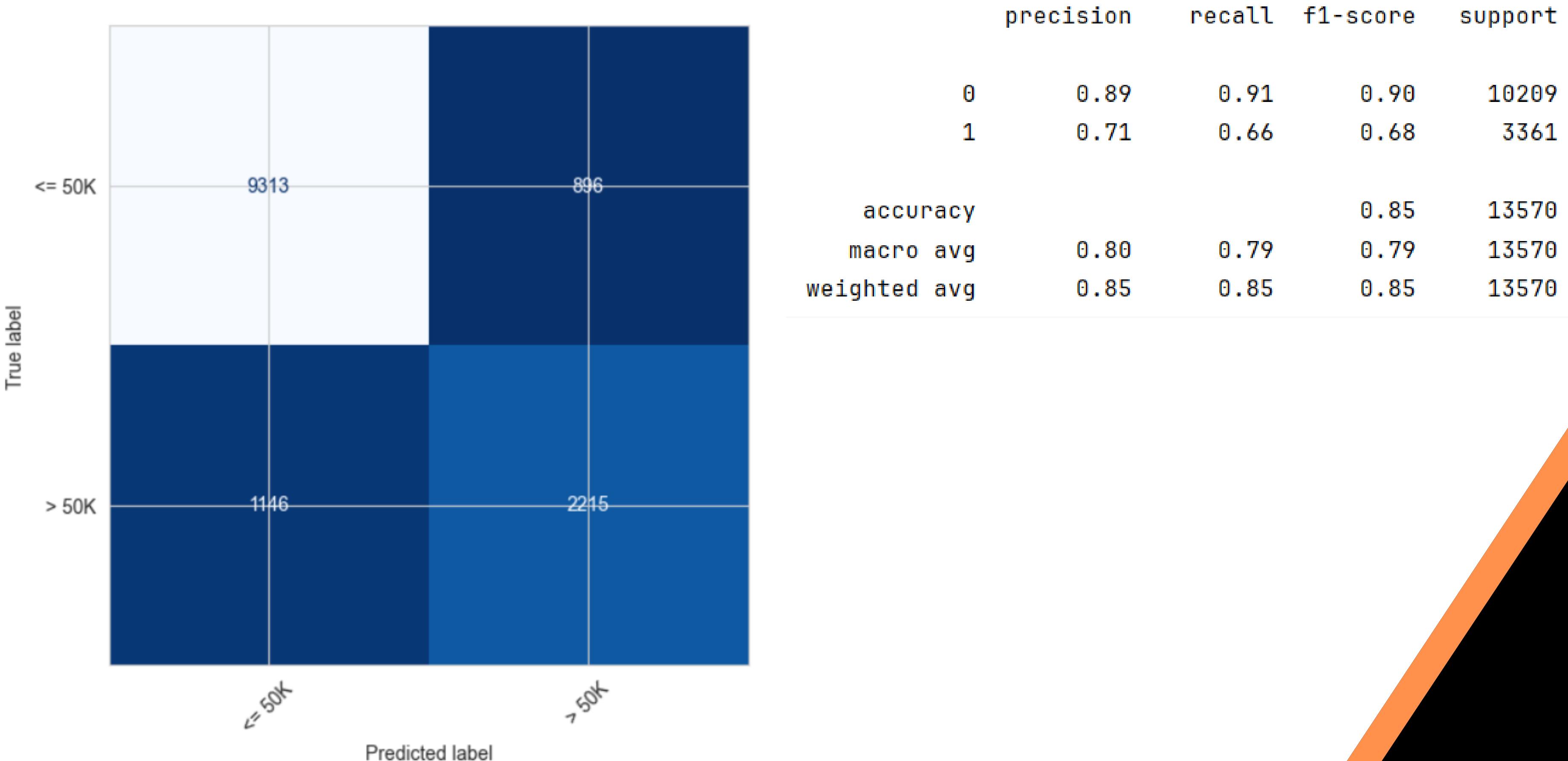
Mô hình Decision Tree Classifier + Tìm max_depth



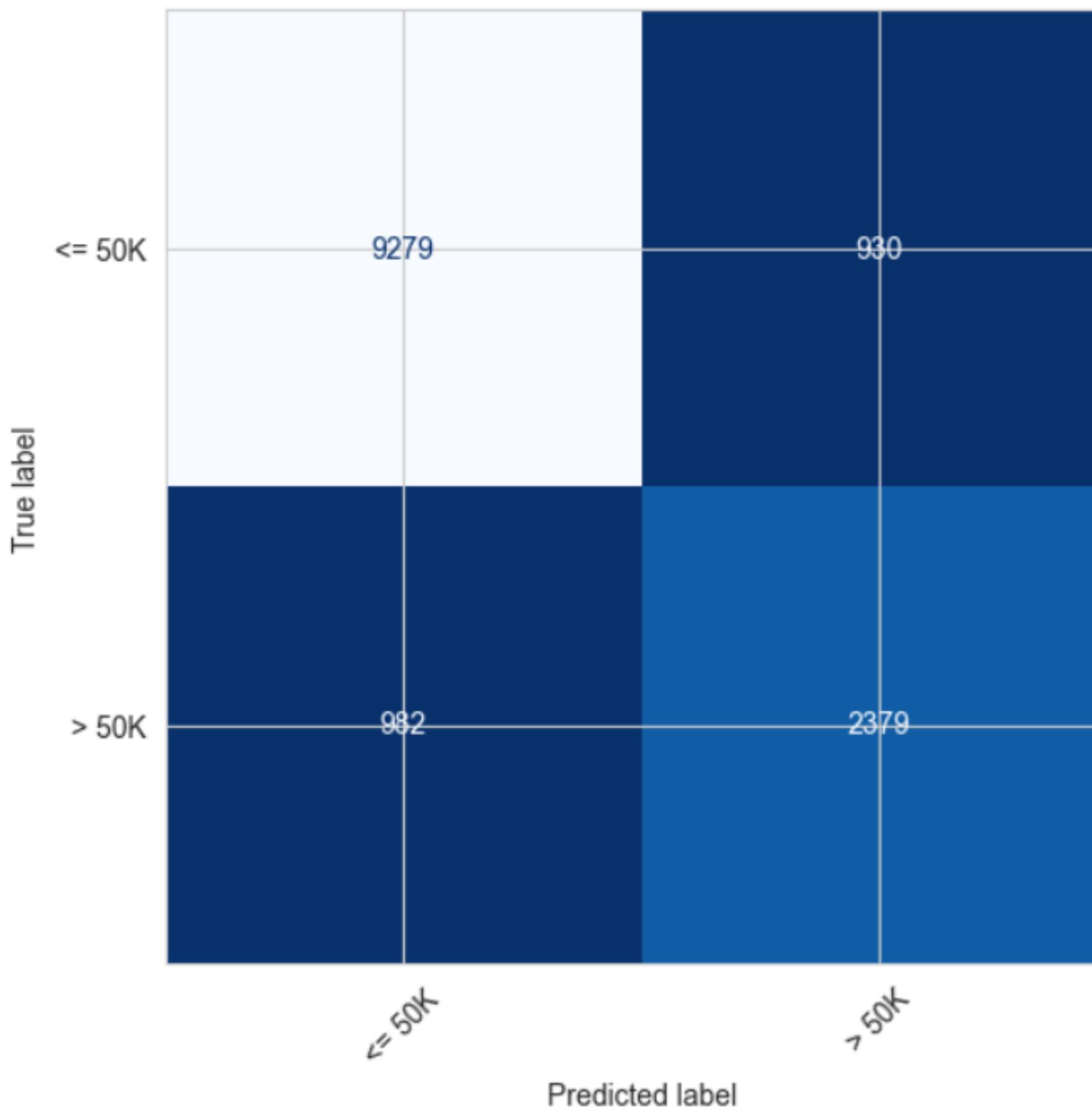
	precision	recall	f1-score	support
0	0.89	0.88	0.89	10209
1	0.65	0.68	0.67	3361
accuracy			0.83	13570
macro avg	0.77	0.78	0.78	13570
weighted avg	0.83	0.83	0.83	13570

```
param_grid = {
    'criterion': ['gini'],
    'max_depth': range(1, 30, 2),
    'min_samples_split': [20]
}
```

Mô hình Random Forest Classifier



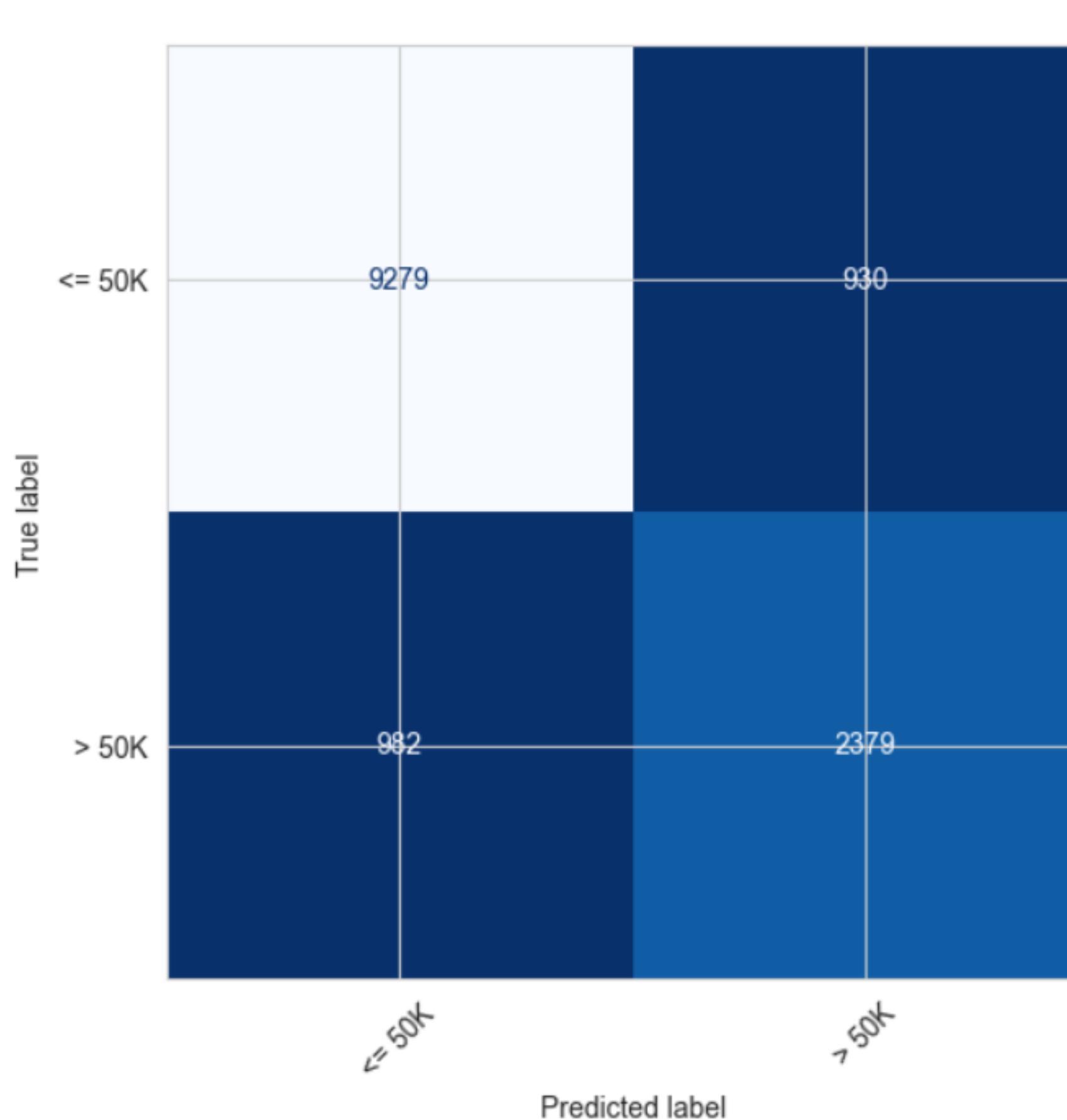
Mô hình Random Forest Classifier + RandomizedSearchCV



	precision	recall	f1-score	support
0	0.90	0.91	0.91	10209
1	0.72	0.71	0.71	3361
accuracy			0.86	13570
macro avg	0.81	0.81	0.81	13570
weighted avg	0.86	0.86	0.86	13570


```
param_grid = {  
    'max_depth': [23],  
    'n_estimators': [280],  
    'max_features': ["sqrt"],  
    'min_samples_split': [6],  
    'min_samples_leaf': [1]  
}
```

Mô hình Random Forest Classifier + GridSearchCV



	precision	recall	f1-score	support
0	0.90	0.91	0.91	10209
1	0.72	0.71	0.71	3361
accuracy			0.86	13570
macro avg	0.81	0.81	0.81	13570
weighted avg	0.86	0.86	0.86	13570

`param_grid = {`

`'max_depth': [23],`

`'n_estimators': [280],`

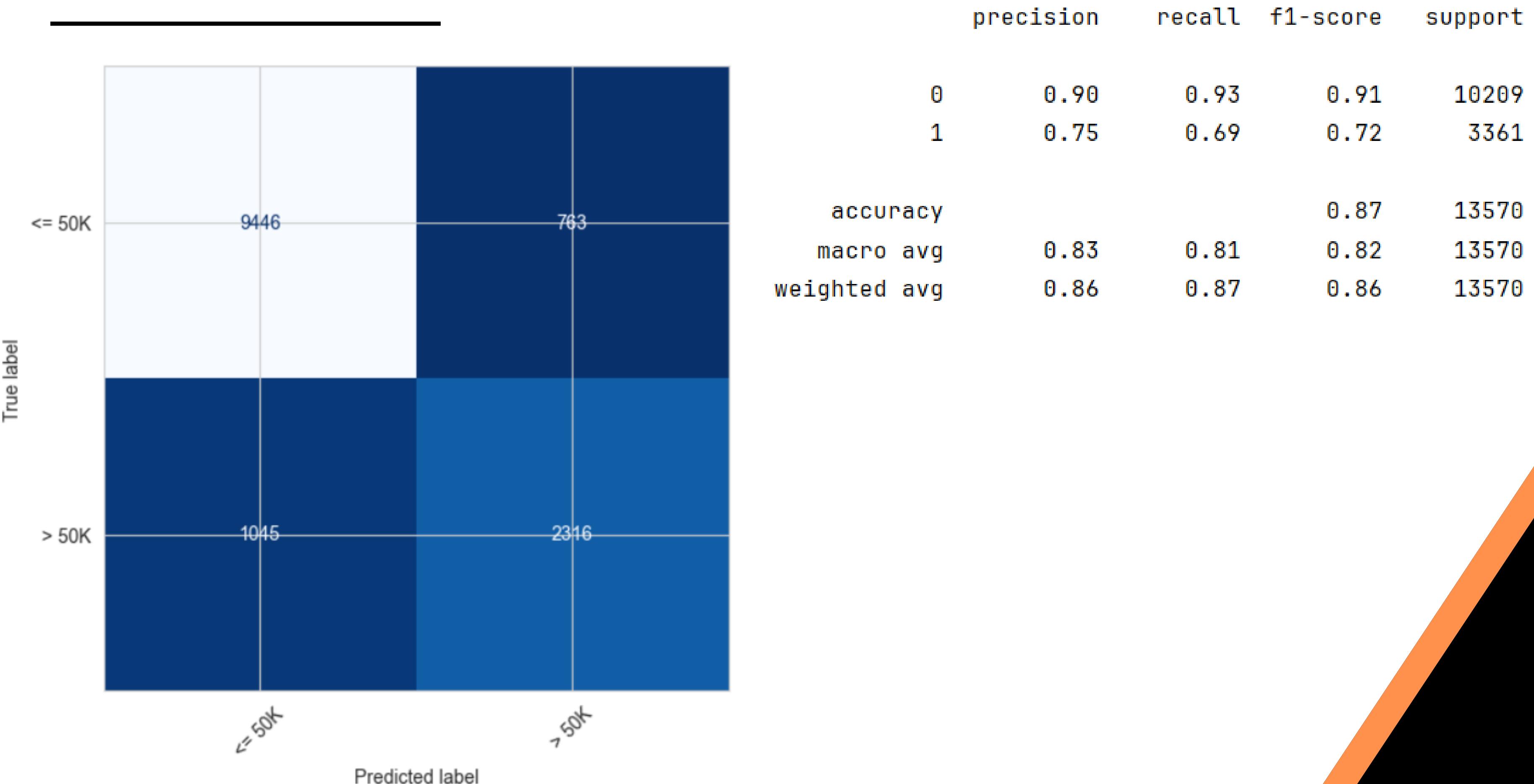
`'max_features': ["sqrt"],`

`'min_samples_split': [6],`

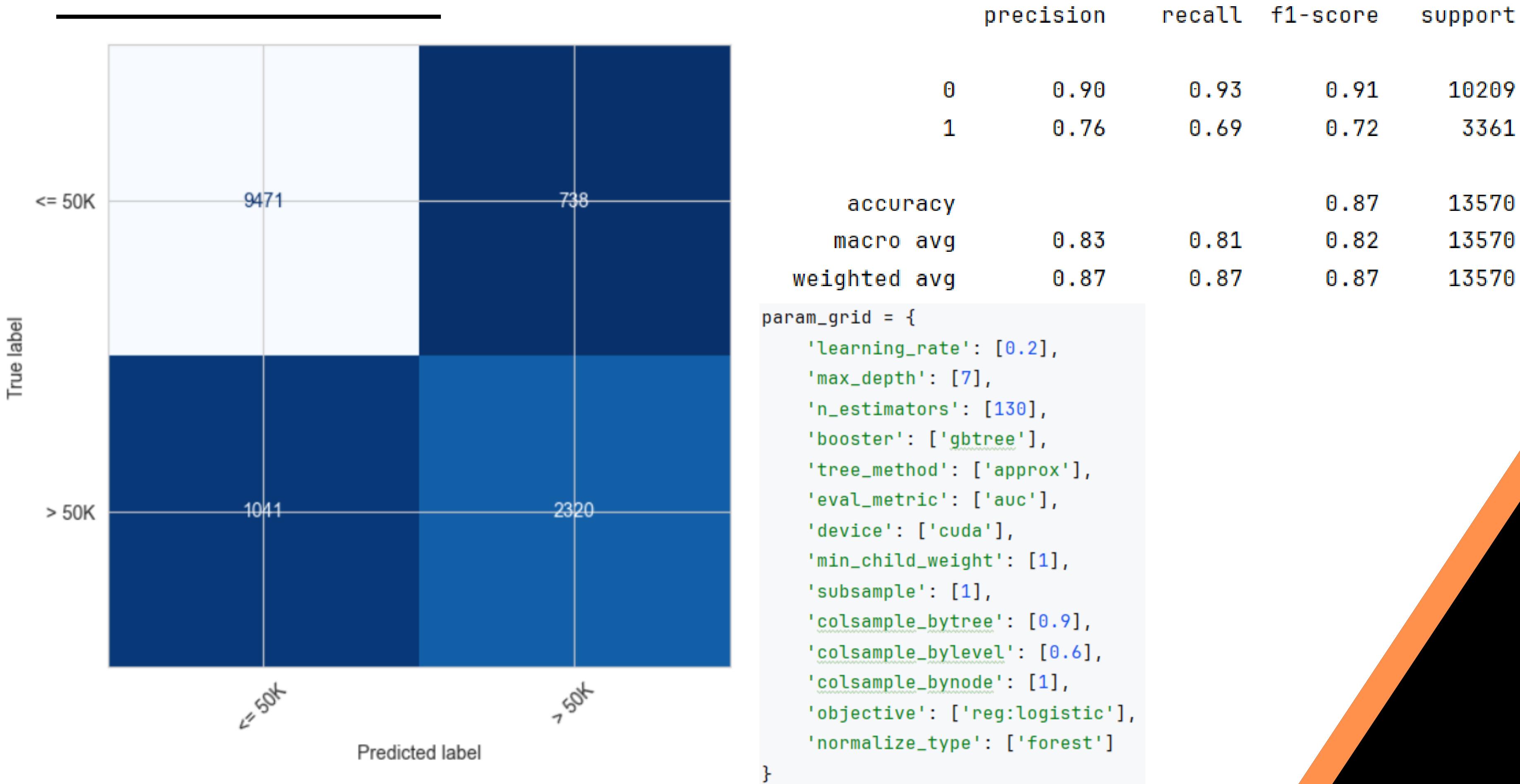
`'min_samples_leaf': [1]`

`}`

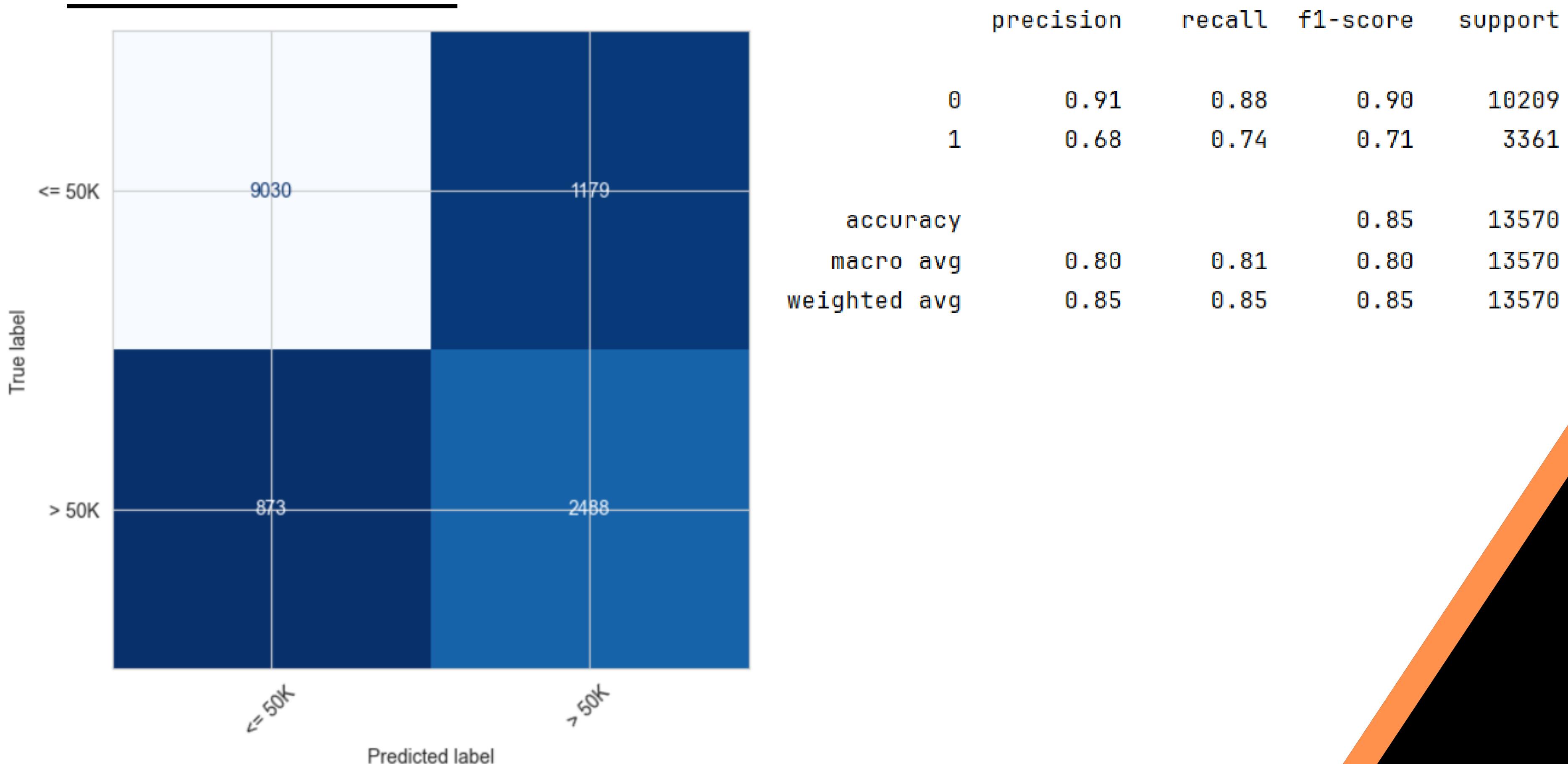
Mô hình XGBoost Classifier



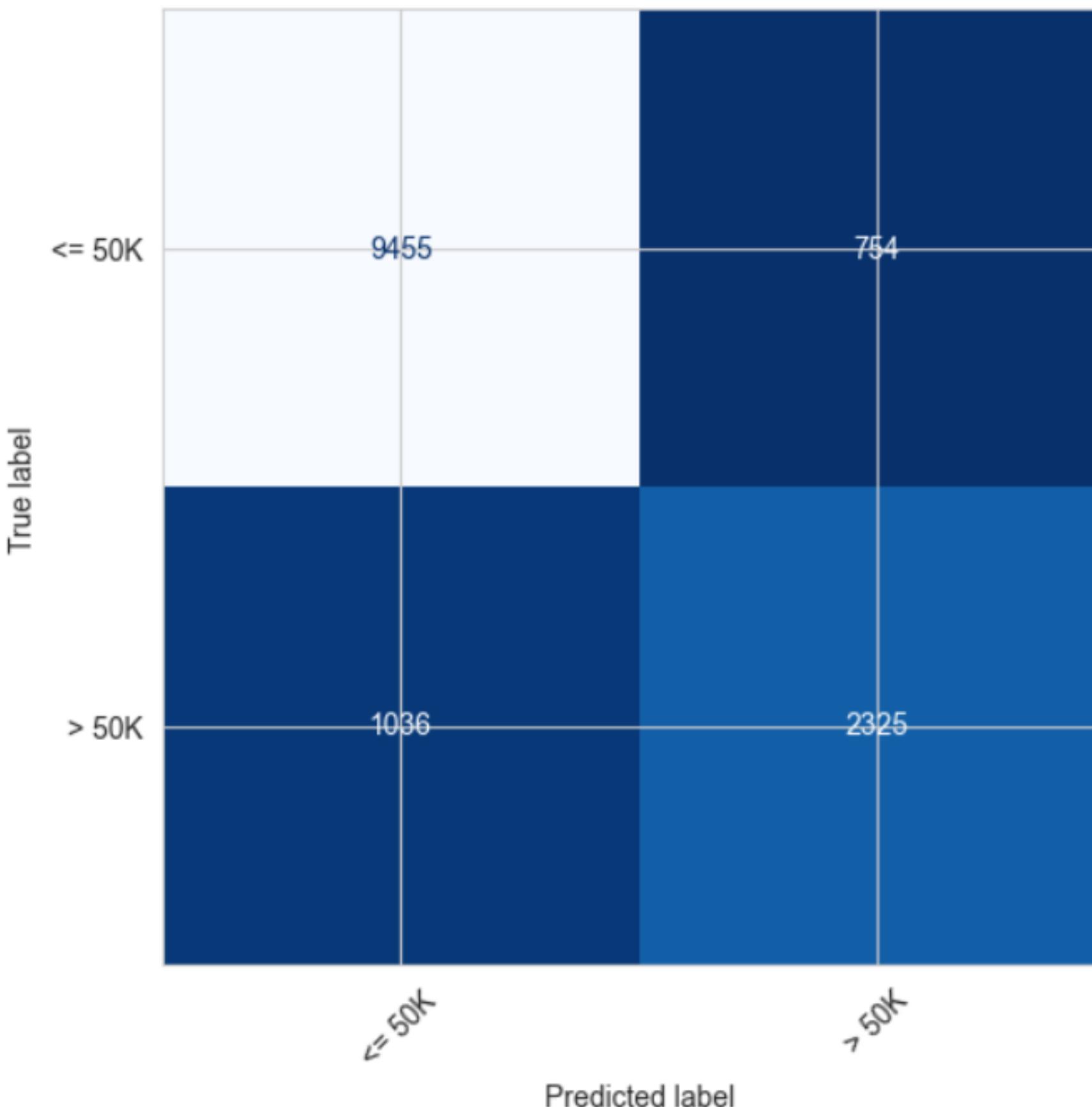
Mô hình XGBoost Classifier + GridSearchCV



Mô hình Gradient Boosting Classifier



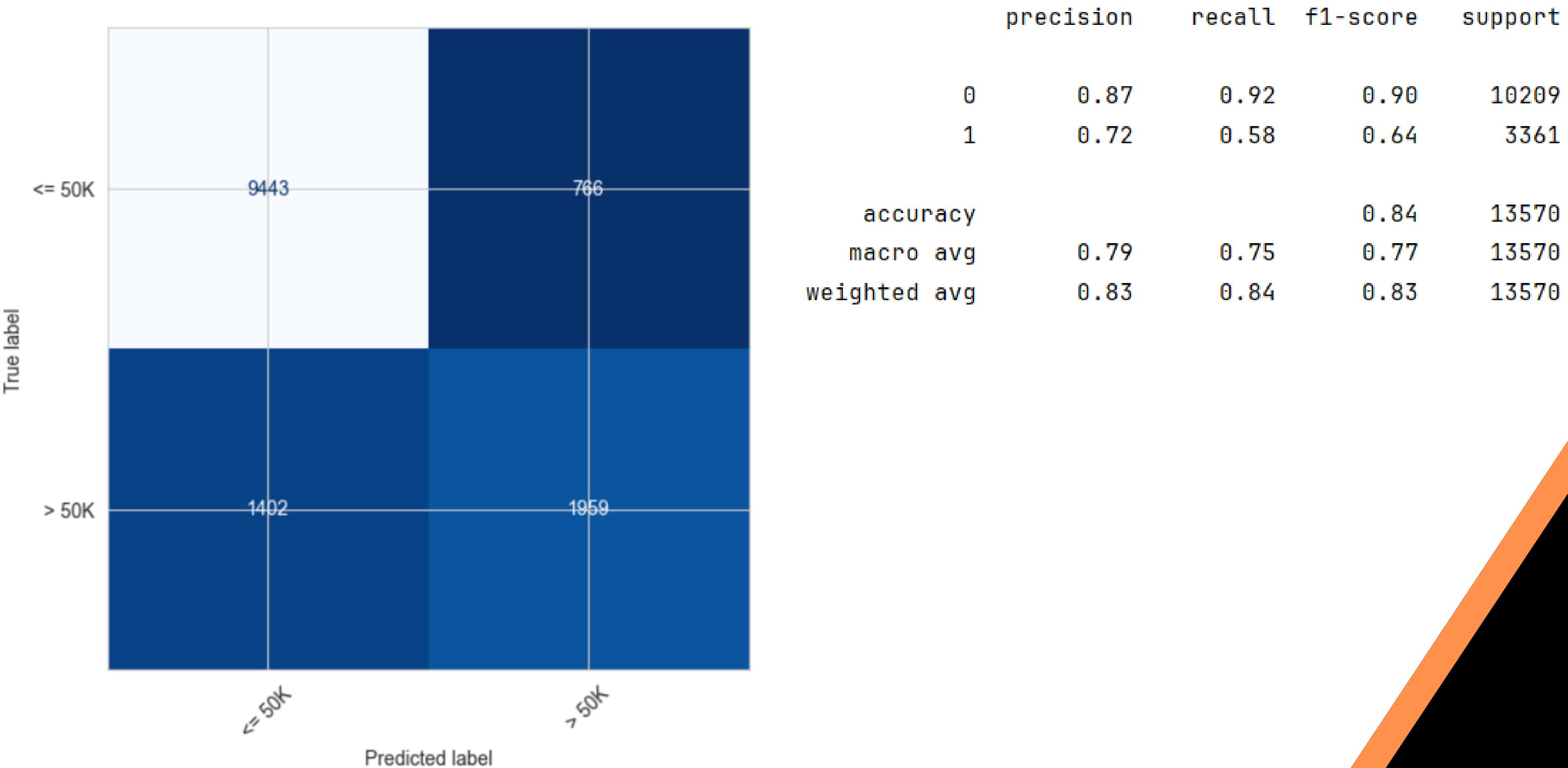
Mô hình Gradient Boosting Classifier + GridSearchCV



	precision	recall	f1-score	support
0	0.90	0.93	0.91	10209
1	0.76	0.69	0.72	3361
accuracy				0.87
macro avg	0.83	0.81	0.82	13570
weighted avg	0.87	0.87	0.87	13570


```
param_grid = {
    'n_estimators': [200],
    'learning_rate': [0.15],
    'max_depth': [5],
    'min_samples_split': [6],
    'min_samples_leaf': [1]}
```

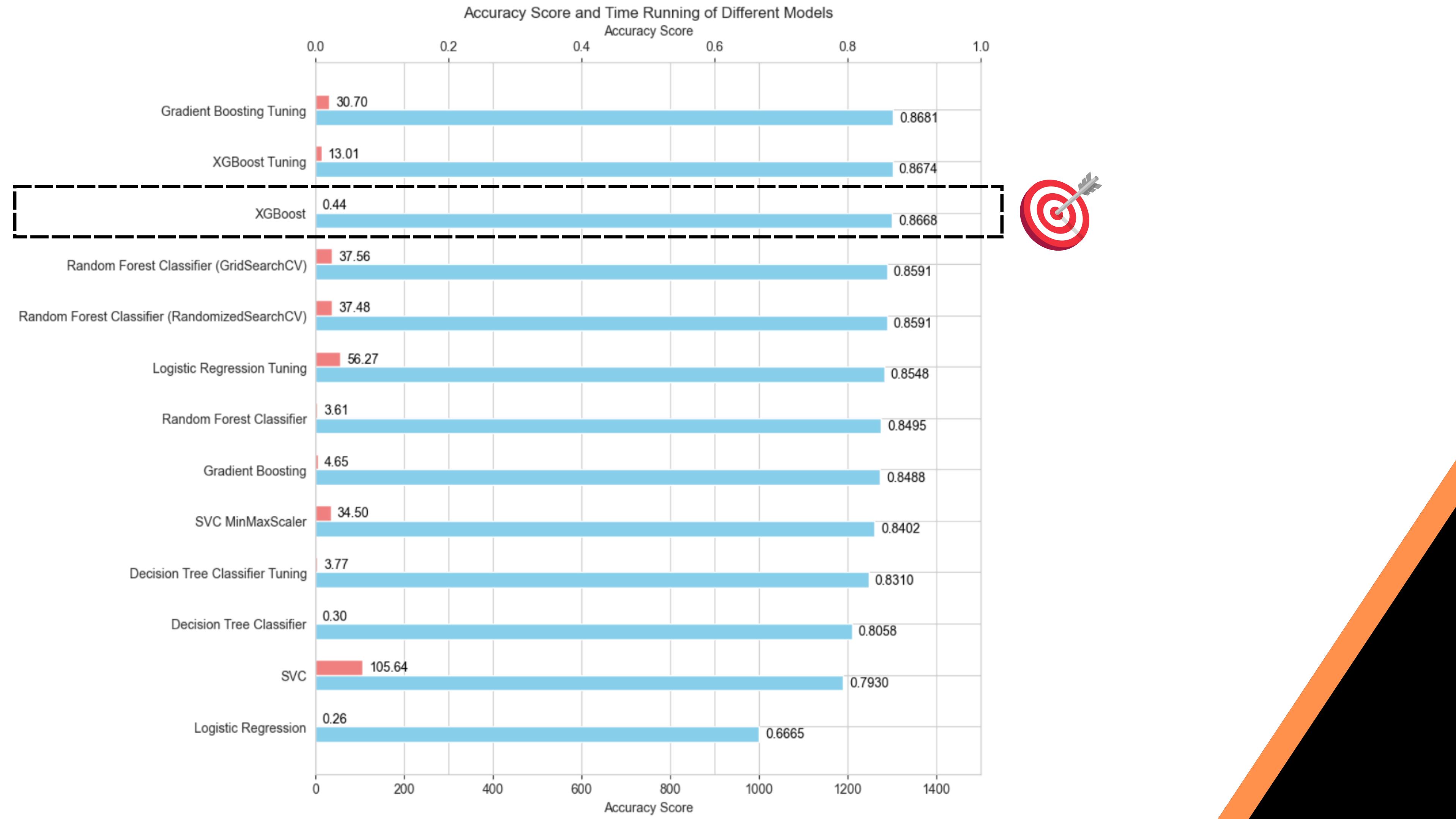
Mô hình SVC + MinMaxScaler



Tổng kết

Model	Accuracy Score	Time Running
0 Gradient Boosting Tuning	0.868091	29.30
1 XGBoost Tuning	0.867354	13.04
2 XGBoost	0.866765	0.41
3 Random Forest Classifier (RandomizedSearchCV)	0.859101	37.31
4 Random Forest Classifier (GridSearchCV)	0.859101	37.29
5 Logistic Regression Tuning	0.854753	52.90
6 Random Forest Classifier	0.849521	3.66
7 Gradient Boosting	0.848784	4.62
8 SVC MinMaxScaler	0.840236	35.65
9 Decision Tree Classifier Tuning	0.830951	4.02
10 Decision Tree Classifier	0.805822	0.30
11 SVC	0.792999	105.33
12 Logistic Regression	0.666544	0.28

Accuracy Score and Time Running of Different Models



2024

PRESENTATION

**Thank you
for attention**

Hoang Trung - Data Team