

****

**TRƯỜNG ĐẠI HỌC SƯ PHẠM KĨ THUẬT TP.HỒ CHÍ MINH**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CUỐI KỲ**

**MÔN HỌC: TƯƠNG TÁC DỮ LIỆU TRỰC QUAN**

**ĐỀ TÀI: TRỰC QUAN HOÁ THU NHẬP CỦA NGƯỜI DÂN**

**Giảng viên hướng dẫn: ThS. Lê Quang Thái**

**Mã môn học: 222IDVI333677**

**Người thực hiện: Lê Phước Yên MSSV: 20133119**

**Nguyễn Duy Thái MSSV: 20133020**

**Phan Thanh Tín MSSV: 20133095**

**Võ Hữu Nghĩa MSSV: 20133074**

**Tp. Hồ Chí Minh, tháng 5 năm 2023**

**MỤC LỤC**

[**1.** **Tóm tắt (Abstract):** 3](#_Toc134890329)

[**2.** **Giới thiệu (Introduction) :** 3](#_Toc134890330)

[**3.** **Dữ liệu (Data):** 4](#_Toc134890331)

[**4.** **Trực quan hóa dữ liệu (Data Visulization):** 5](#_Toc134890332)

[**5.** **Mô hình hóa dữ liệu (Model):** 13](#_Toc134890333)

[**6.** **Nhận xét và kết luận (Comment, Conclusions):** 17](#_Toc134890334)

[**6.1.** **Sử dụng Logistic Regression:** 17](#_Toc134890335)

[**6.2.** **Tổng kết (Summary):** 17](#_Toc134890336)

[**7.** **Đóng góp (Contributions):** 18](#_Toc134890337)

[**8.** **Tham khảo:** 18](#_Toc134890338)

**Report Guides**

1. **Tóm tắt (Abstract):**

Mỹ là nước phát triển rất mạnh, những người dân Mỹ luôn có mức thu nhập rất cao, đời sống của họ cũng thế. Ở bộ dữ liệu điều tra về thu nhập cá nhân của người dân ở đây nhằm mục đích dò khảo xem muốn được thu nhập cao như họ cụ thể là thu nhập hơn 50000USD cần những yếu tố gì? Có phải là ai cũng có cơ hội để mà có thu nhập cao như vậy không? Nhờ đó cũng có thể đưa ra những kế hoạch cho bản thân để cũng có thể thu nhập cao như vậy.

Các phương pháp nghiên cứu được sử dụng:

Thu thập và xử lý số liệu: ở đây thì phần dữ liệu còn khá thô có khi có cả những dữ liệu không xác định được nên khá là rắc rối trong việc sử nó để tiến hành nghiên cứu. Vì vậy cần chuẩn hóa, tính toán, trình bày,… Để đưa về dạng mình cần thiết và dễ dàng sử dụng.

Điều tra chọn mẫu: sử dụng 1 phần nhỏ của tổng thể để có thể suy luận các hiện tượng và tất nhiên nó vẫn đảm bảo trên độ tin cậy đặt ra.

Nguyên cứu mối liện hệ giữa các hiện tượng: kiểm tra các mối quan hệ.

Dự đoán: phương pháp này nhờ vào các phương pháp kể ở trên để đưa ra dự đoán.

Kết quả nghiên cứu: Sau khi sử dụng các phương pháp trên những thuộc tính như: age, occupation, sex, hours\_per\_week, gây ảnh hưởng đến việc mình có kiếm được 50000USD hay không. Từ đó có thể đưa ra kế hoạch cho bản thân. Tuy nhiên vẫn là còn thiếu và chúng em sẽ tiếp tục phát triễn dự án để đưa ra kết quả chính xác nhất.

1. **Giới thiệu (Introduction) :**

Các câu hỏi nghiên cứu của nhóm :

* Giới tính có ảnh hưởng đến việc thu nhập trên 50000USD hay không. Câu hỏi này muốn thể hiện rằng liệu có sự phân biệt giữa nam và nữ trong việc thu nhập hay không hay nó chỉ thể hiện sự điều đặn ai cũng như ai và cũng có khả năng dễ dàng có thu nhập trên 50000USD
* Người trẻ tuổi kiếm được 50000USD dễ hơn người trung niên. Câu hỏi này muốn kiểm tra xem liệu độ tuổi có ảnh hưởng đến việc thu nhập hay không.
* Người có số giờ làm việc cao dễ kiếm được 50000USD. Câu hỏi này quan tâm đến việc có phải chắc chắn rằng thời gian làm việc của bạn nhiều thì thu nhập của bạn càng cao đúng hay không.
* Dự đoán thu nhập của họ thông qua các đặc điểm đã phân tích. Ở đây muốn bạn mờ tưởng và đưa ra quyết định của mình sau khi nhìn nhận các kết quả thông qua các phương pháp nghiên cứu trên trong việc thu nhập của 1 người qua các đặc điểm, thuộc tính: age, sex, hours\_per\_week,…
* Có thể thấy các yếu tố được liệt kê ở trên cũng có thể xem xét trên thực tế ví dụ như thuộc tính “age” thì dù ở độ tuổi nào cũng có thể tạo ra thu nhập cho cá nhân mình cụ thể ở đây là người có mức tuổi trên 16 ta dùng kiểm định trên khoảng tin cậy có thể đưa ra khoảng tuổi nào là có khả năng sẽ có được thu nhập trên 50000USD. Tương tự thuộc tính hours\_per\_week cũng sử dụng phương pháp trên.

1. **Dữ liệu (Data):**

Nguồn thu thập dữ liệu: <https://www.kaggle.com/datasets/uciml/adult-census-income>. Dữ liệu này được trích xuất từ ​​cơ sở dữ liệu của Cục điều tra dân số năm 1994 bởi Ronny Kohavi và Barry Becker (Khai thác và Trực quan Dữ liệu, Đồ họa Silicon).

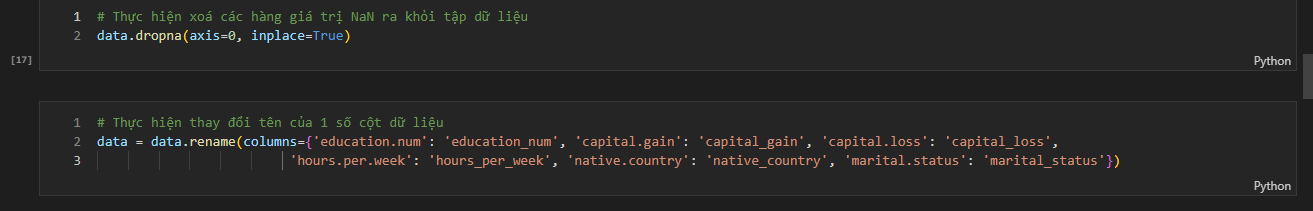
15 thuộc tính bao gồm:

* age: độ tuổi của từng cá nhân.
* workclass: lớp làm việc.
* fnlwgt: final weight – trọng lượng cuối cùng.
* education: giáo dục.
* education.num: số năm học của một cá nhân.
* marital.status: tình trạng hôn nhân (đã li hôn, góa vợ, đã kết hôn...).
* occupation: nghề nghiệp.
* relationship: mối quan hệ trong gia đình (chồng, vợ, không có gia đình, con riêng...).
* race: chủng tộc (màu da: da trắng, da đen, châu Á...).
* sex: giới tính.
* capital.gain: vốn lãi của mức thu nhập.
* capital.loss: lỗ vốn.
* hours.per.week: số giờ làm việc mỗi tuần.
* native.country: quốc tịch.
* income: dự đoán liệu người đó có kiếm được nhiều hơn 50.000 USD hay không.

Thực hiện in vài dòng trong tập dữ liệu:

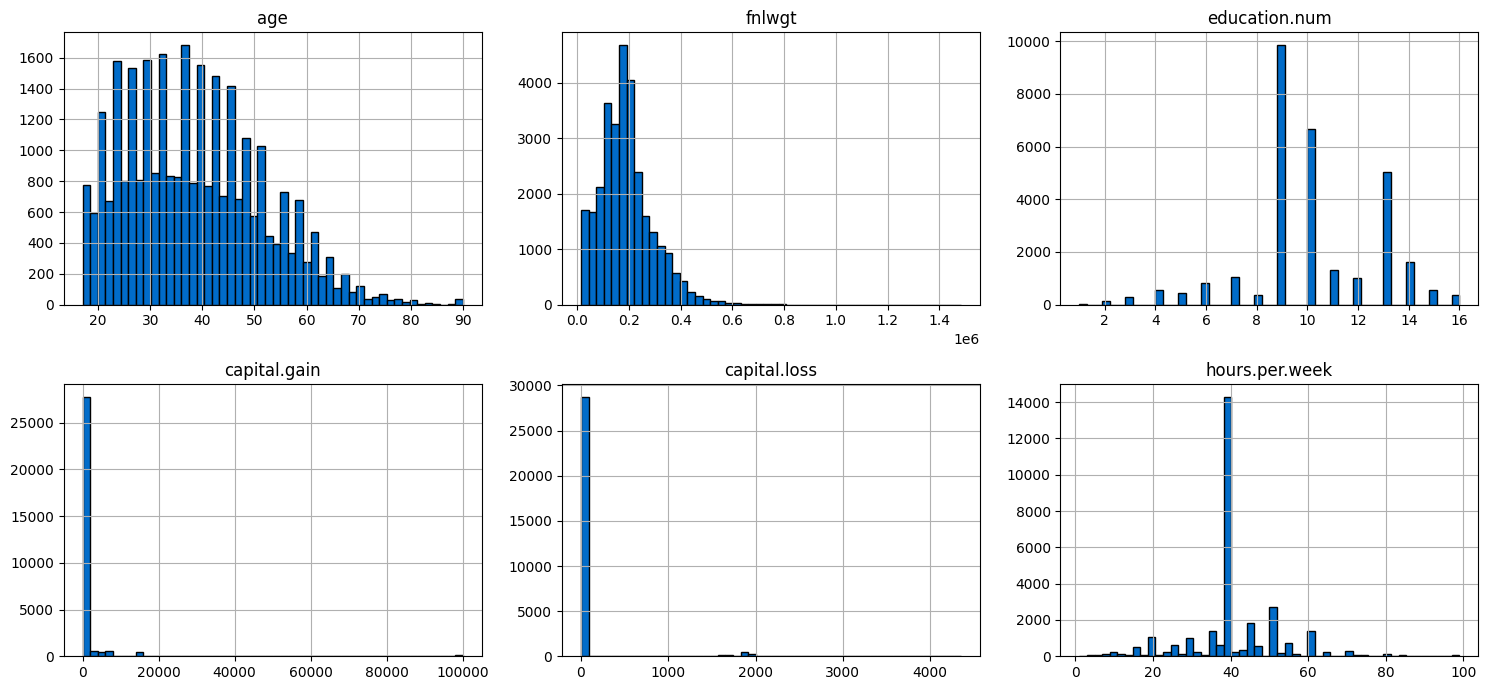


Tiền sử lý dữ liệu:

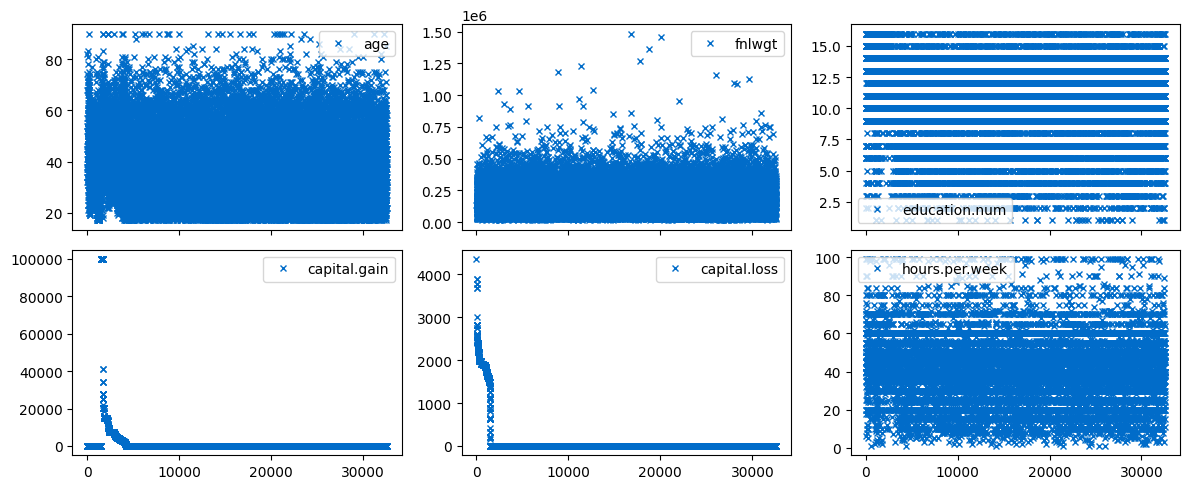


1. **Trực quan hóa dữ liệu (Data Visulization):**

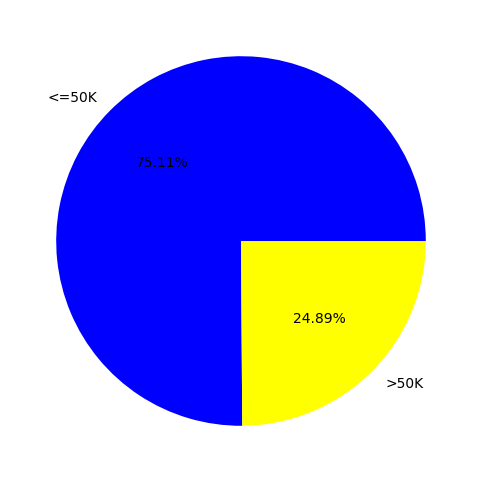
* Biểu đồ Histogram cho các biến số:

****

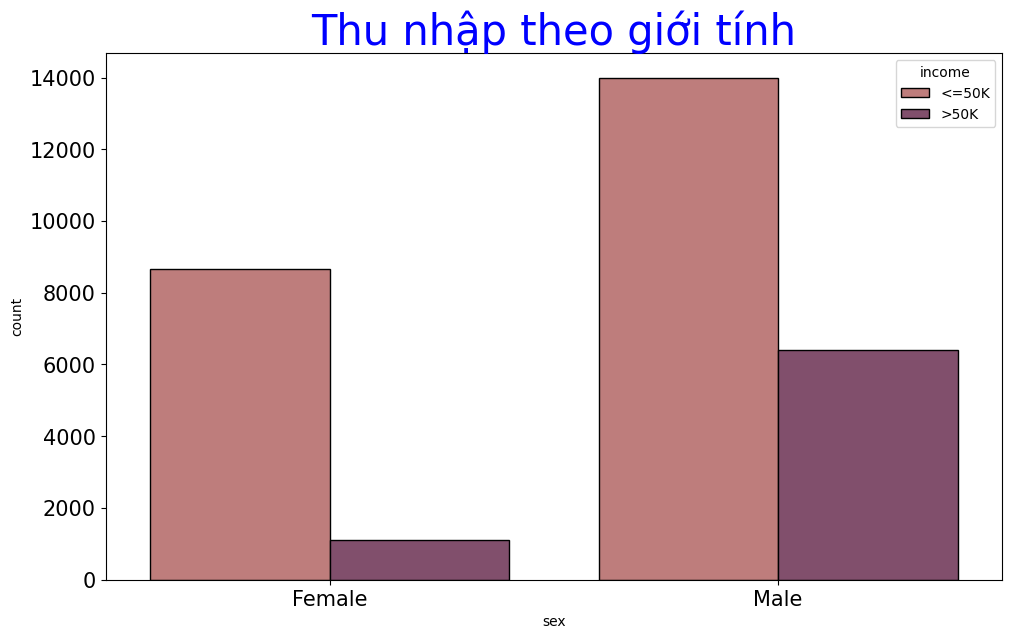
* Biểu đồ thể hiện mật độ phân bó dữ liệu của các biến số:

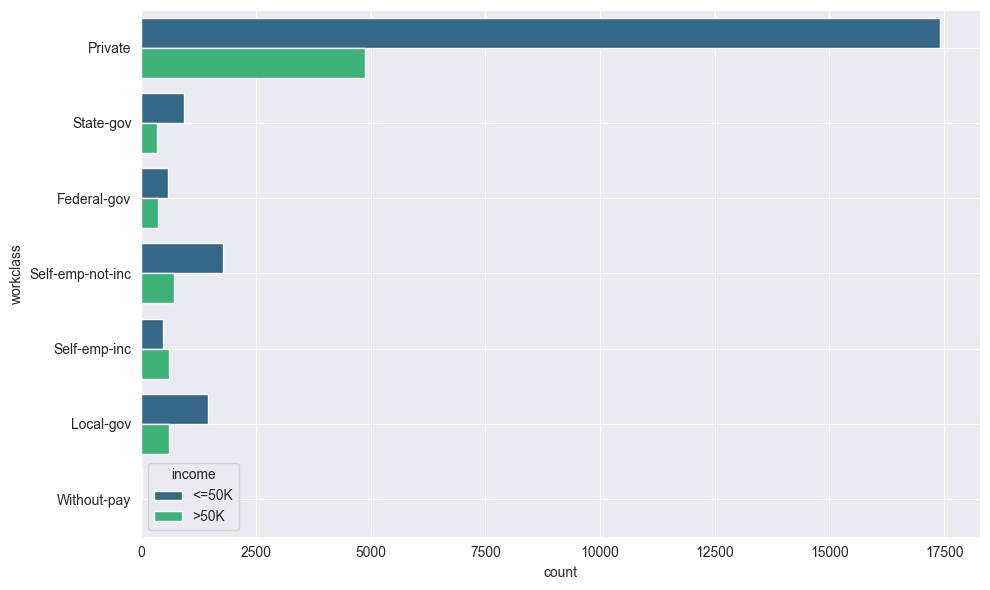
****

* Biểu đồ biểu thị phần trăm phân bố dữ liệu của biến income

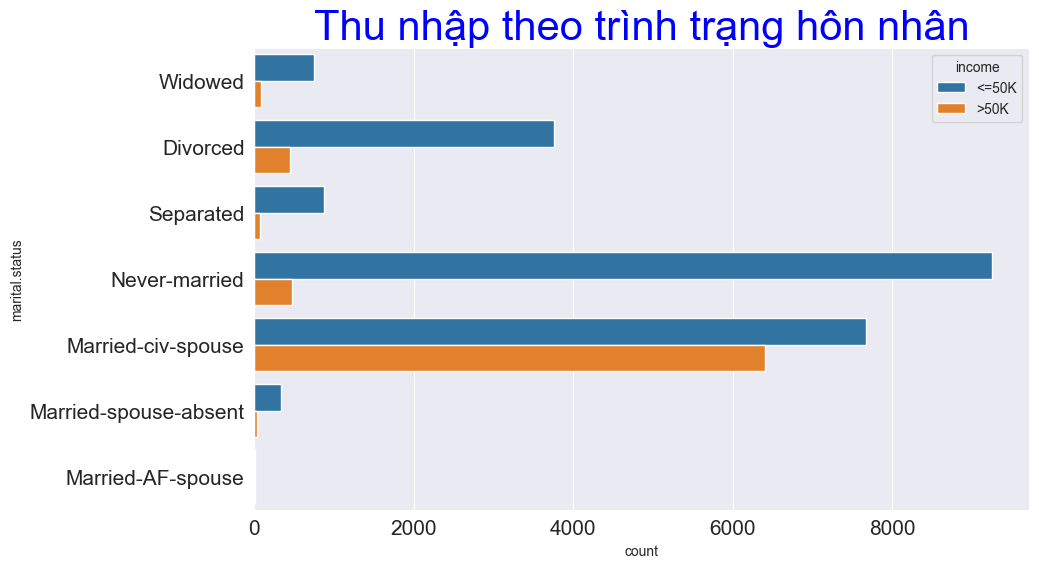
****

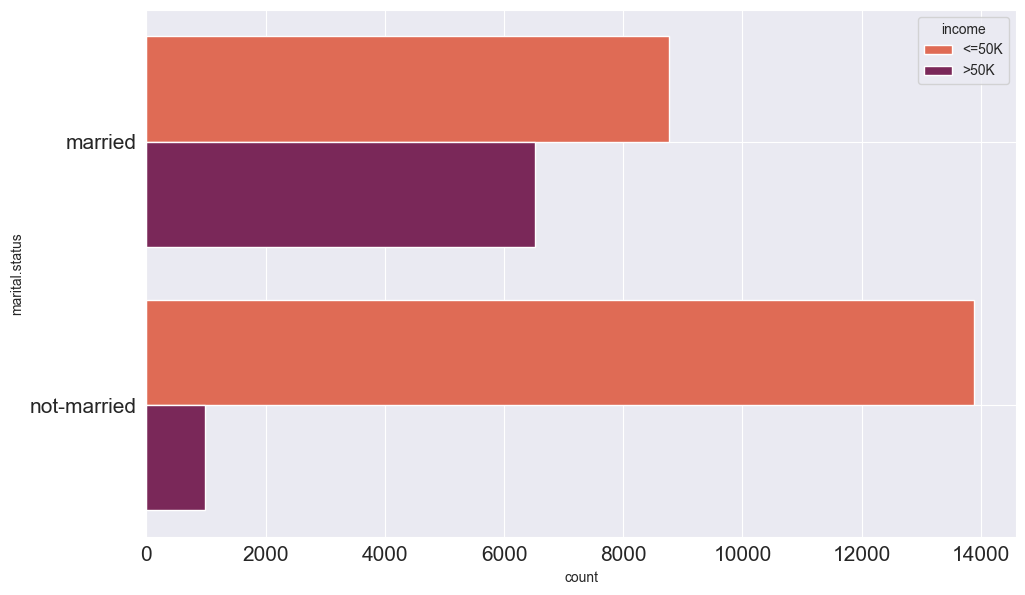
* Biểu đồ cột so sánh mức thu nhập theo giới tính

****

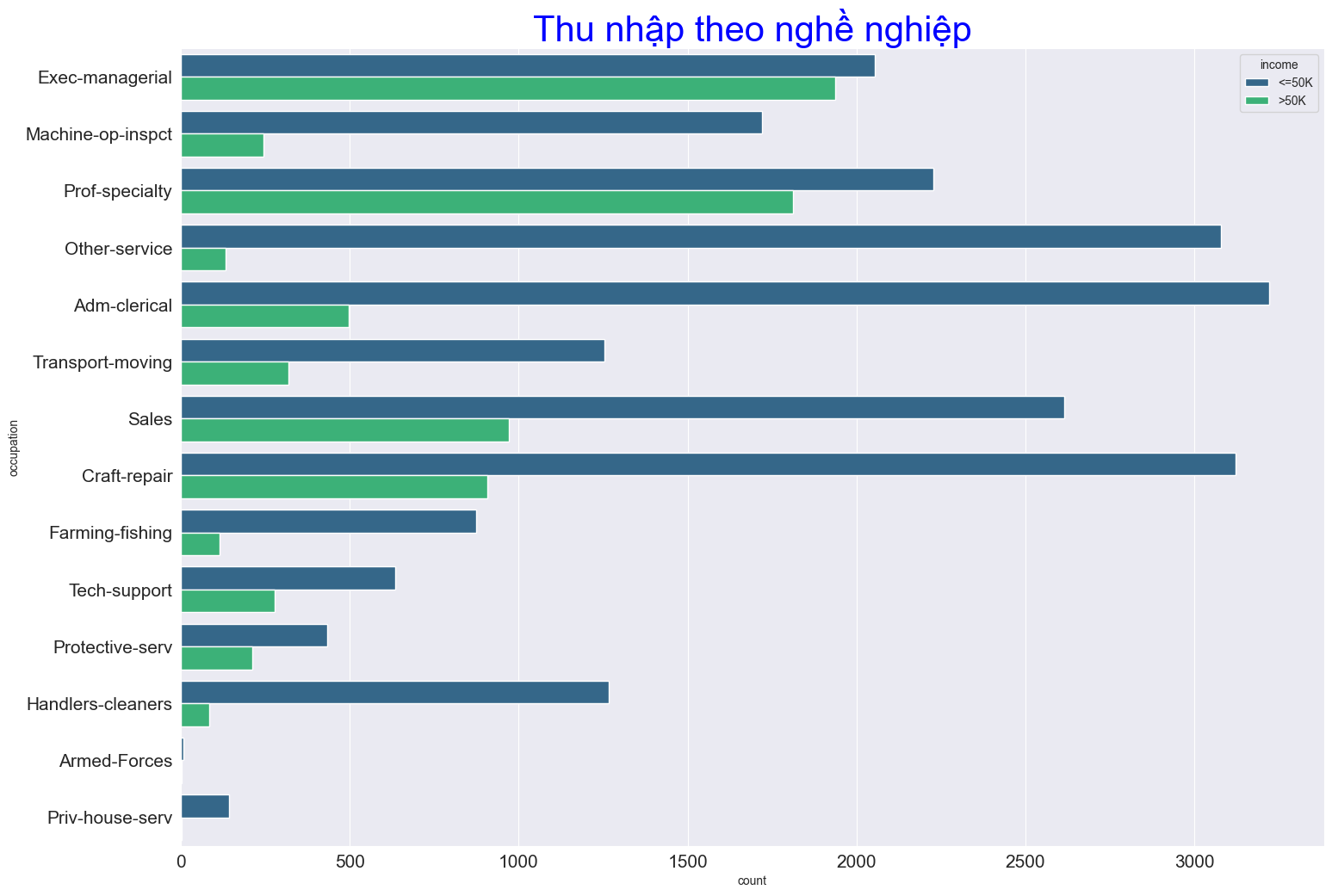
****

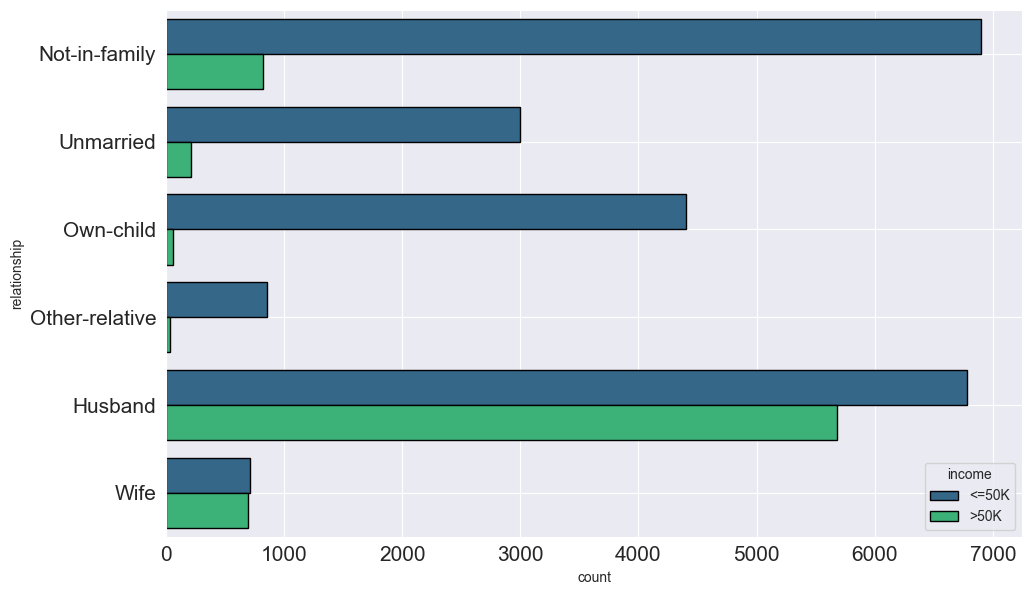
* Biểu đồ cột so sánh mức thu nhập theo tình trạng hôn nhân

****

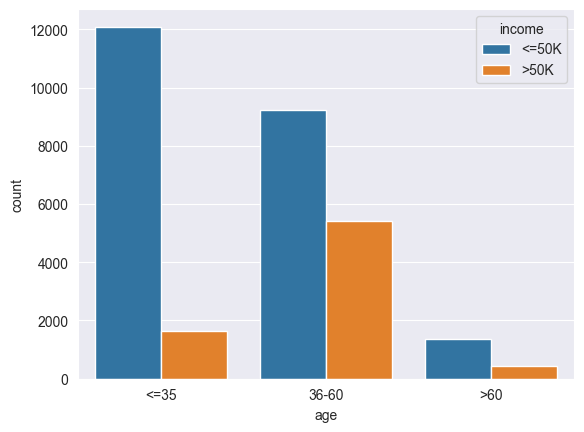
****

* Biểu đồ cột so sánh mức thu nhập theo nghề nghiệp

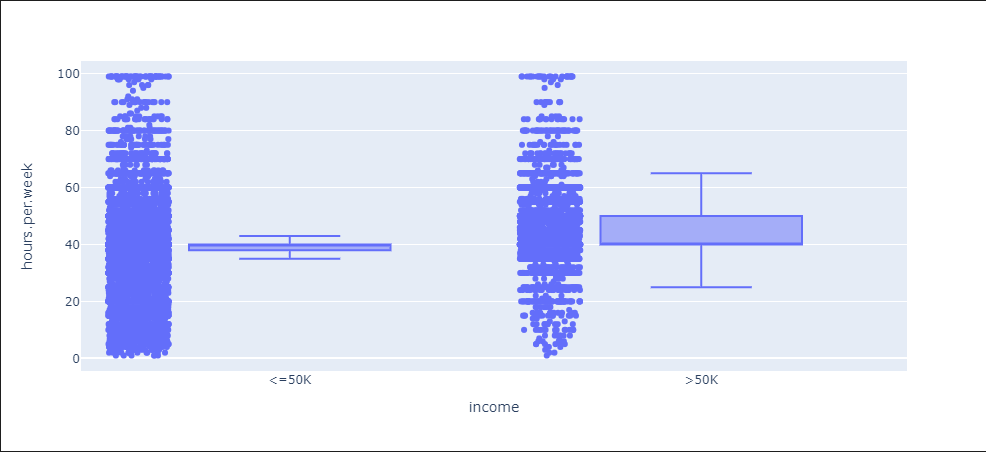
****

****

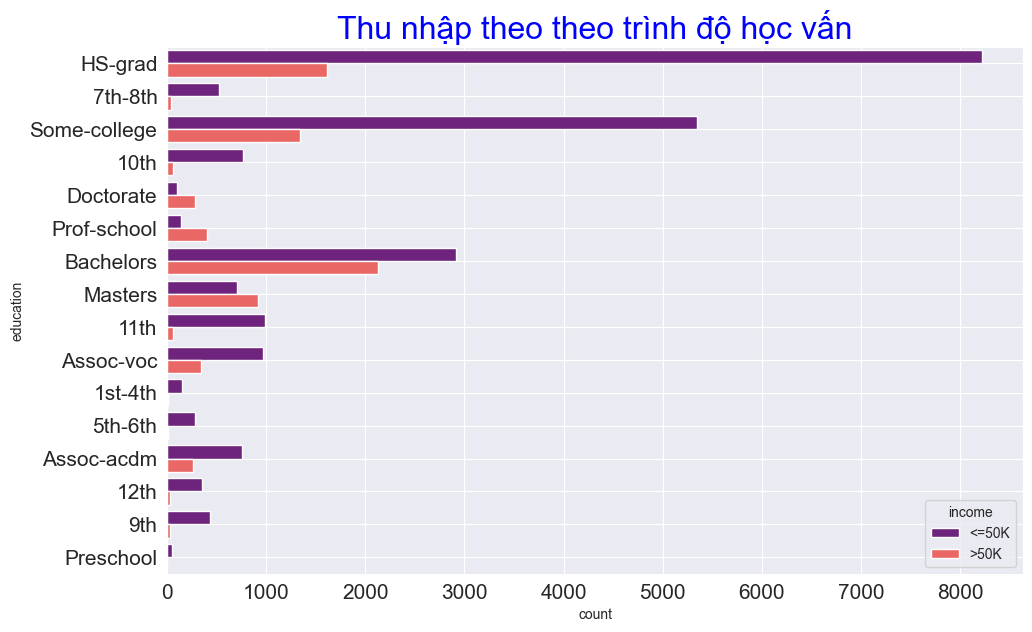
* Biểu đồ cột so sánh mức thu nhập ở các độ tuổi

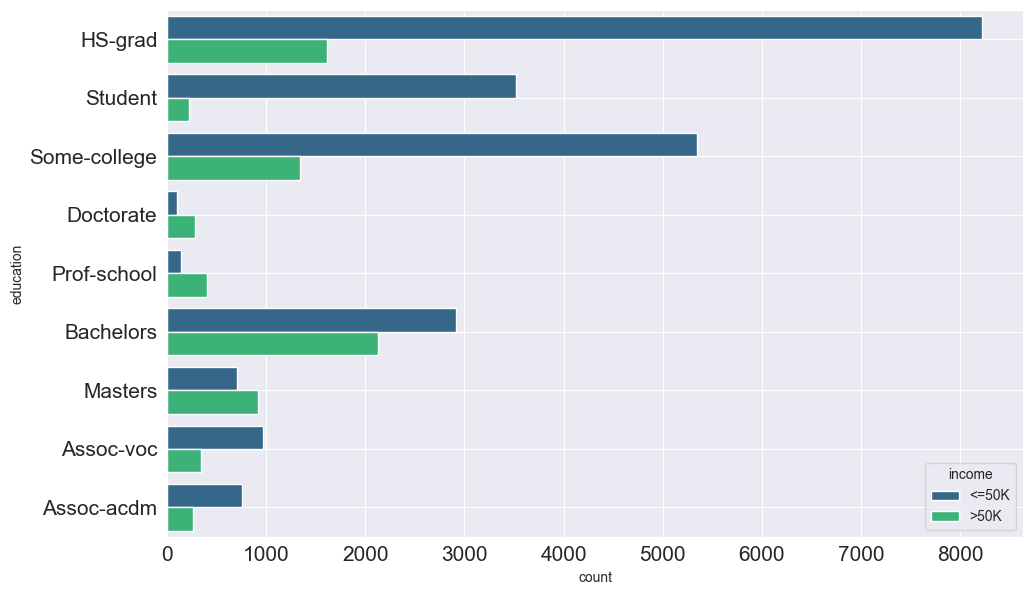
****

* Biểu đồ thể hiện sự phân bố dữ liệu của thuộc tính hours.per.week theo thu nhập

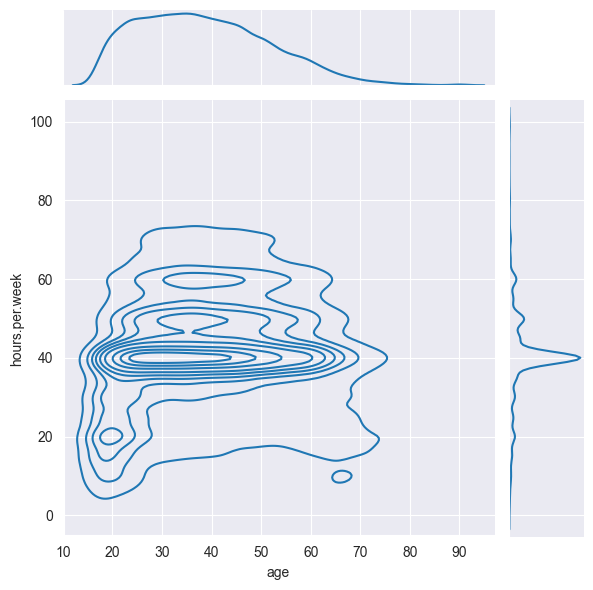
****

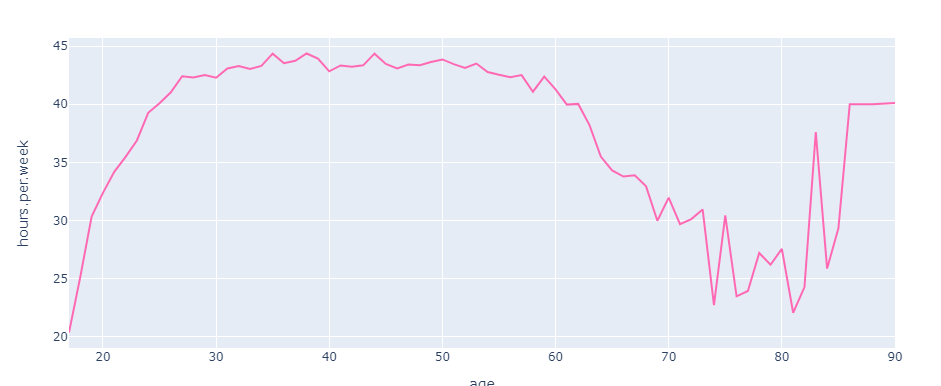
* Biểu đồ cột so sánh mức thu nhập theo trình độ học vấn

****

****

* Biểu đồ thể hiện sự tương quan giữa thuộc tính Age và hours.per.week



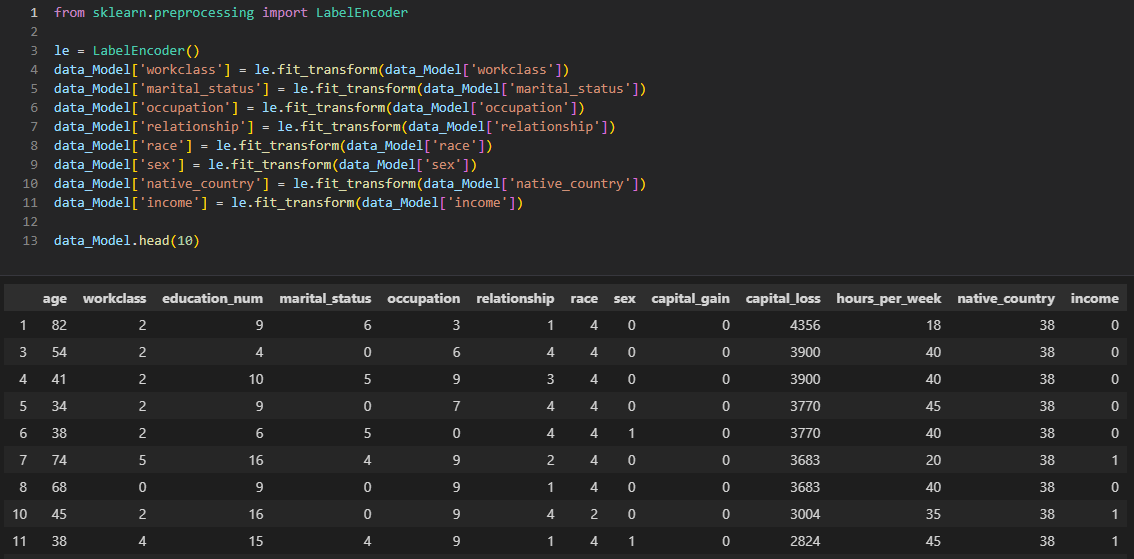


1. **Mô hình hóa dữ liệu (Model):**

* Sử dụng Logistic Regression:

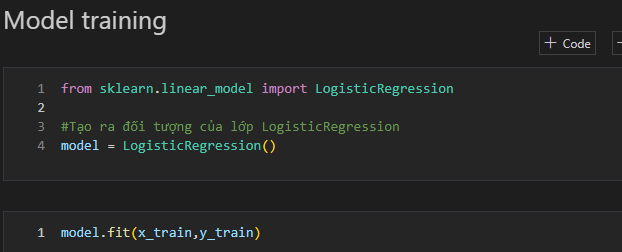
Hồi quy Logistic là một kỹ thuật phân tích dữ liệu sử dụng toán học để tìm ra mối quan hệ giữa hai yếu tố dữ liệu. Sau đó, kỹ thuật này sử dụng mối quan hệ đã tìm được để dự đoán giá trị của những yếu tố đó dựa trên yếu tố còn lại. Dự đoán thường cho ra một số kết quả hữu hạn, như có hoặc không.

Các bước thực hiện :

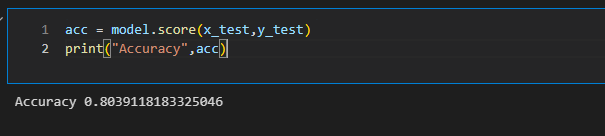
* Chuẩn bị dữ liệu: biến đổi các giá trị quy về thành các số.
* Chuẩn bị mô hình: chia tập dữ liệu thành các tập train và test.



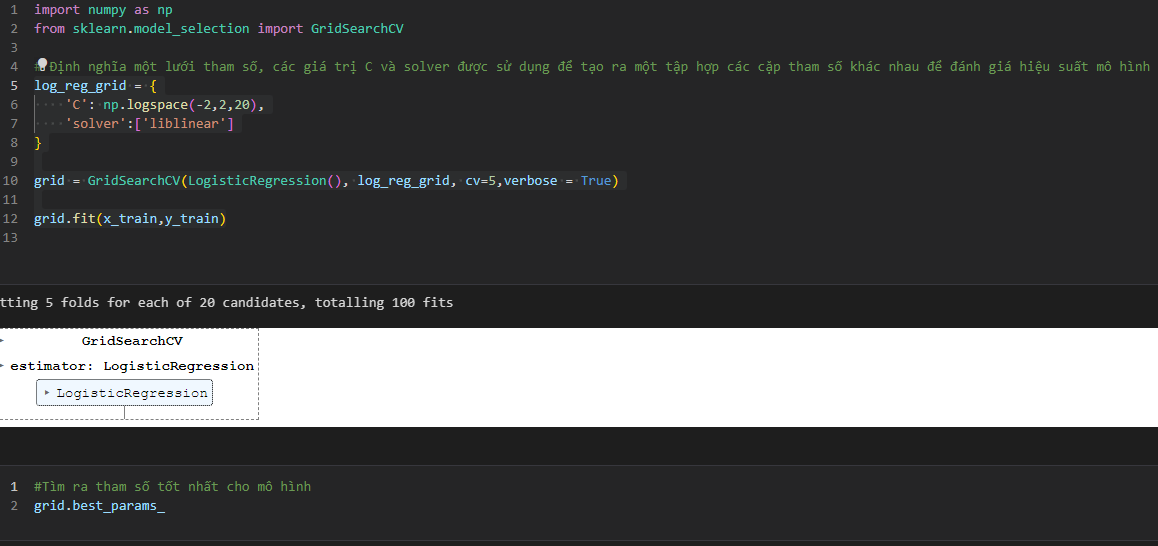
* Bắt đầu huấn luyện mô hình:



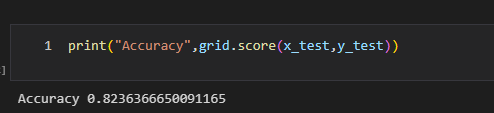
* Kiểm tra lại độ chính xác của mô hình:

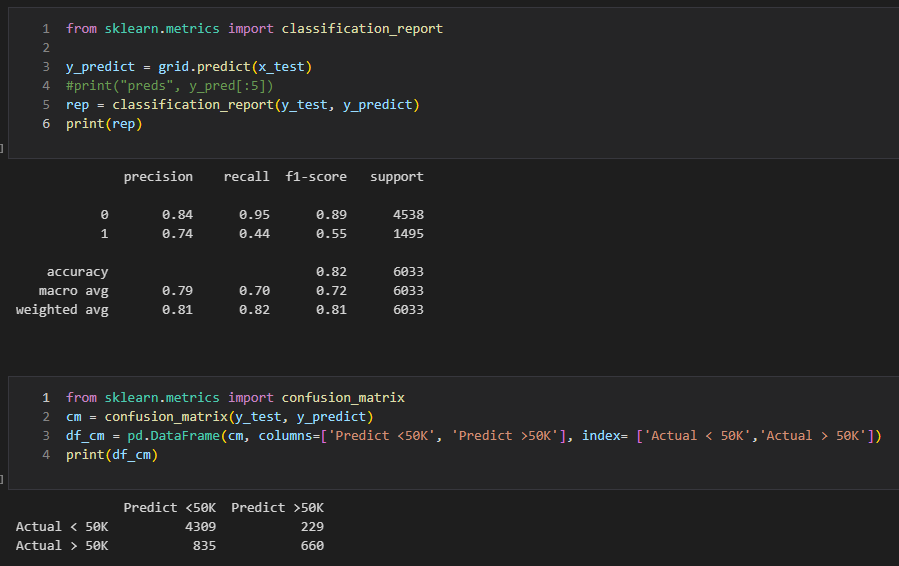


* Lựa chọn ra tham số tốt nhất cho mô hình

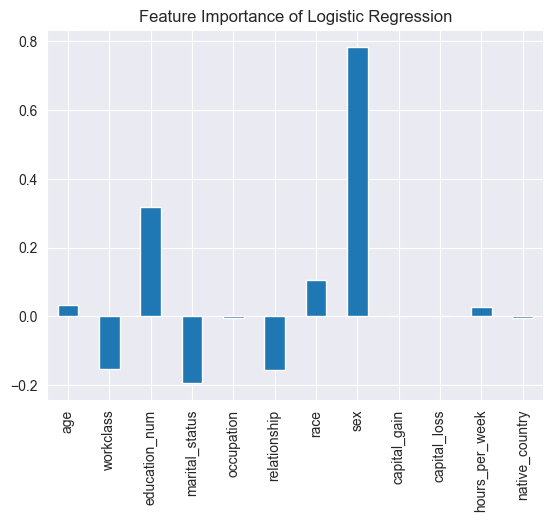


* Thử nghiệm lại với tập test:

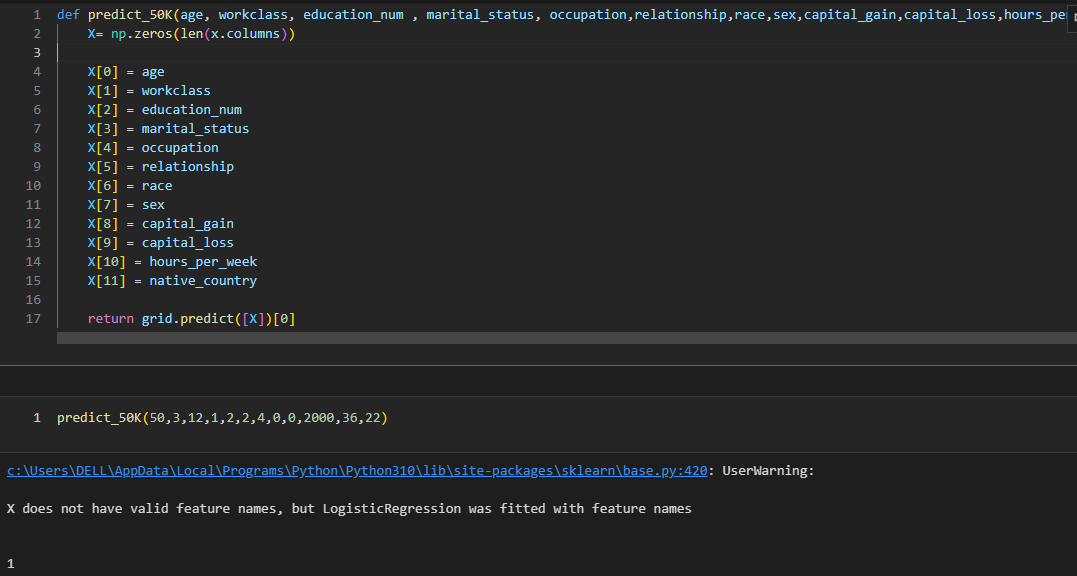




* Đưa ra cái nhìn về các thuộc tính mang tính quyết định đến việc có thu nhập được trên 50K USD hay không:



* Chạy thử mô hình của nhóm vừa huấn luyện được:



1. **Nhận xét và kết luận (Comment, Conclusions):**
   1. **Sử dụng Logistic Regression:**

Tỉ lệ dự đoán đúng (precision) người có thu nhập dưới 50K/năm : 84%, recall : 95%.

Tỉ lệ dự đoán đúng (precision) người có thu nhập trên 50k/năm : 74%, recall: 44% , tỉ lệ bỏ sót cao những người có thu nhập trên 50K/năm .

Mô hình phân loại rất tốt những người có thu nhập dưới 50K/năm, nhưng mô hình chưa tốt trong việc phân loại người đó có thu nhập trên mức 50K/năm.

* 1. **Tổng kết (Summary):**

Qua bài phân tích từ những nhận xét, biểu đồ từ đó ta có thể trả lời được các giả thuyết của đầu bài. Nhờ đó cũng giúp chúng có thể dự đoán được 1 người là có thể có thu nhập lớn hơn 50000USD hay không thông qua các đặc điểm mà ta phân tích được.

Từ những số liệu trên có thể thấy độ tuổi lao mà để kiếm được số tiền 50000USD nó nằm khoảng trung bình 44 tuổi và có độ lệch chuẩn khoảng 10 tuổi tức nghĩa từ độ tuổi 33 - 53 tuổi.

Nhìn vào biểu đồ có thể thấy được ở nam(Male) thì thường sẽ kiếm được số tiền trên 50000USD hơn là nữ.

Nhìn qua thì ta thấy được những ngành nghề mà giúp có thể kiếm được trên 50000USD như: Exec - managerial(Người điều hành, quản lý), Prof-specialyti(Giáo sư đặc biệt), Sales(Người bán hàng), Craft-repair(Sửa chửa). Tuy nhiên nó cũng chưa giúp ít gì cho ta kết luận được việc người đó có kiếm 50000USD hay không bởi vì tỉ lệ mà những công việc này kiếm được lượng tiền nhỏ hơn 50000USD cũng rất là cao. Nhưng nhờ vậy ta cũng có thể loại trừ được những ngành nghề khác.

Qua biểu đồ về thuộc tính hours\_per\_week cho ta biết được khoảng thời gian làm việc trung bình mà ta có thể kiếm được trên 50000USD là khoảng 45.7h/tuần .

Do là còn không có đủ thời gian nên nếu được phát triển thêm nhóm muốn khai thác sâu thêm nhiều biến để đưa ra kết quả chính xác nhất.

1. **Đóng góp (Contributions):**

|  |  |  |
| --- | --- | --- |
| Tên thành viên | Nhiệm vụ | Mức độ hoàn thành |
| Lê Phước Yên | + Sử dụng Python thực hiện trực quan hoá các thuộc tính.  + Tổng hợp. | 100% |
| Nguyễn Duy Thái | + Xây dựng mô hình dự đoán.  + Kết luận. | 100% |
| Phan Thanh Tín | + Lên ý tưởng.  + Sử dụng Python thực hiện trực quan hoá các thuộc tính. | 100% |
| Võ Hữu Nghĩa | + Sử dụng PowerBI để trực quan hoá các thuộc tính.  + Kết luận. | 100% |

1. **Tham khảo:**

Nhóm đã tham khảo tài liệu của mấy anh chị năm trước kèm trên Youtube

<https://www.analyticsvidhya.com/blog/2021/09/guide-for-building-an-end-to-end-logistic-regression-model/>

<https://www.youtube.com/watch?v=g6sIXyF496E>