**TRƯỜNG ĐẠI HỌC QUỐC TẾ HỒNG BÀNG**

**BỘ MÔN CÔNG NGHỆ THÔNG TIN**

**🙢🕮🙠**



**BÁO CÁO ĐỀ TÀI CÁ NHÂN**

**KHO DỮ LIỆU & KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI: PHÂN TÍCH VÀ DỰ ĐOÁN THU NHẬP TRÊN TẬP DỮ LIỆU ADULT INCOME**

**Giảng viên hướng dẫn : Thầy Lê Văn Hạnh**

**Sinh viên thực hiện : Ng. Hoàng Bảo Lâm**

**Mã số sinh viên : 2111111032**

**TP. Hồ Chí Minh, 2025**

TRANG CAM KẾT

Tôi xin cam kết báo cáo thường kỳ này được hoàn thành dựa trên các kết quả thực hiện bài thực hành của tôi và các mã nguồn và kết quả này chưa được dùng cho bất cứ báo cáo của sinh viên nào khác.

*TP.HCM, ngày ….. tháng ….. năm …..*

Người thực hiện

NHẬN XÉT CỦA GIẢNG VIÊN

TP.HCM, Ngày Tháng Năm 2023

Chữ ký giảng viên

MỤC LỤC

[TRANG CAM KẾT i](#_Toc194415272)

[NHẬN XÉT CỦA GIẢNG VIÊN ii](#_Toc194415273)

[MỤC LỤC iii](#_Toc194415274)

[DANH MỤC BẢNG BIỂU 4](#_Toc194415275)

[DANH MỤC HÌNH ẢNH 5](#_Toc194415276)

[GIỚI THIỆU SƠ BỘ 6](#_Toc194415277)

[CHƯƠNG 1. GIỚI THIỆU VỀ CSDL SỬ DỤNG TRONG ĐỀ TÀI 8](#_Toc194415278)

[1.1. TỔNG QUAN VỀ CSDL 8](#_Toc194415279)

[1.2. GIỚI THIỆU TỪNG THUỘC TÍNH 8](#_Toc194415280)

[1.3. TIỀN XỬ LÝ DỮ LIỆU 9](#_Toc194415281)

[1.3.1. Xử lý dữ liệu thiếu 10](#_Toc194415282)

[1.3.2. Mã hóa dữ liệu (Encoding) 11](#_Toc194415283)

[CHƯƠNG 2. PHÂN TÍCH – THỐNG KÊ THỦ CÔNG TRÊN CSDL ĐÃ CHỌN 13](#_Toc194415284)

[2.1. TÌM HIỂU DỮ LIỆU 13](#_Toc194415285)

[2.1.1. PHÂN TÍCH ĐỒ THỊ CHO CÁC THUỘC TÍNH 13](#_Toc194415286)

[2.1.2. NHÓM DỮ LIỆU THEO THUỘC TÍNH DANH NGHĨA 15](#_Toc194415287)

[2.1.3. ĐO LƯỜNG SỰ TƯƠNG ĐỒNG VÀ KHÁC BIỆT CỦA DỮ LIỆU 16](#_Toc194415288)

[2.2. TIỀN XỬ LÝ DỮ LIỆU (BỔ SUNG) 18](#_Toc194415289)

[2.3. TỔNG HỢP DỮ LIỆU 19](#_Toc194415290)

[2.4. TRỰC QUAN HÓA DỮ LIỆU 21](#_Toc194415291)

[2.5. THỰC HIỆN KHAI THÁC DỮ LIỆU 25](#_Toc194415292)

[2.5.1. SỬ DỤNG CÁC PHƯƠNG PHÁP KHAI PHÁ DỮ LIỆU 25](#_Toc194415293)

[2.5.2. ĐÁNH GIÁ CÁC MẪU THU ĐƯỢC 28](#_Toc194415294)

[CHƯƠNG 3. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 31](#_Toc194415295)

[3.1. KẾT LUẬN 31](#_Toc194415296)

[3.2. HƯỚNG PHÁT TRIỀN 31](#_Toc194415297)

[TÀI LIỆU THAM KHẢO 33](#_Toc194415298)

DANH MỤC BẢNG BIỂU

[Bảng 1: Thông tin cơ bản về tập dữ liệu Adult Income 8](#_Toc194415299)

[Bảng 2: Ma trận độ đo Cosine giữa 4 dòng dữ liệu đầu tiên 18](#_Toc194415300)

[Bảng 3: Tổng hợp theo giới tính và thu nhập 20](#_Toc194415301)

[Bảng 4: Tổng hợp theo trình độ học vấn và thu nhập 20](#_Toc194415302)

[Bảng 5: Thống kê tuổi và giờ làm việc theo thu nhập 21](#_Toc194415303)

[Bảng 6: Bảng liên hợp giữa cụm và thu nhập thực tế 27](#_Toc194415304)

[Bảng 7: Phương pháp đánh giá bằng Cross-Validation 28](#_Toc194415305)

DANH MỤC HÌNH ẢNH

[Hình 1: Tiền xử lý dữ liệu 9](#_Toc194415310)

[Hình 2: Mã hóa dữ liệu 11](#_Toc194415311)

[Hình 3: Boxplot dựa trên five-number summary của các thuộc tính số 13](#_Toc194415312)

[Hình 4: Quantile-Quantile Plot của Age và Hours-per-week 13](#_Toc194415313)

[Hình 5: Histogram của Age và Hours-per-week 14](#_Toc194415314)

[Hình 6: Scatter Plot: Age vs Hours-per-week 14](#_Toc194415315)

[Hình 7: Boxplot của Age theo Education 15](#_Toc194415316)

[Hình 8: Histogram của Age cho các nhóm Education 16](#_Toc194415317)

[Hình 9: Ma trận tương quan giữa các thuộc tính 17](#_Toc194415318)

[Hình 10: Heatmap độ đo Cosine giữa 4 dòng dữ liệu đầu tiên 17](#_Toc194415319)

[Hình 11: Biểu đồ tỷ lệ thu nhập theo giới tính 21](#_Toc194415320)

[Hình 12: Biểu đồ tỷ lệ thu nhập theo trình độ học vấn 22](#_Toc194415321)

[Hình 13: Biểu đồ phân bố tuổi theo thu nhập 22](#_Toc194415322)

[Hình 14: Biểu đồ phân bố giờ làm việc theo thu nhập 23](#_Toc194415323)

[Hình 15: Ma trận tương quan giữa các thuộc tính 24](#_Toc194415324)

[Hình 16: Ma trận nhầm lẫn - Decision Tree 25](#_Toc194415325)

[Hình 17: Ma trận nhầm lẫn - Random Forest 26](#_Toc194415326)

[Hình 18: : Kết quả phân cụm K-means (sau khi giảm chiều với PCA) 27](#_Toc194415327)

[Hình 19: Đường cong ROC 29](#_Toc194415328)

GIỚI THIỆU SƠ BỘ

**Các công cụ, thư viện và phương pháp được sử dụng**

Trong quá trình thực hiện đề tài “Phân tích và khai phá dữ liệu thu nhập từ bộ dữ liệu Adult Income”, tôi đã sử dụng ngôn ngữ lập trình **Python** cùng với các thư viện mạnh mẽ, hỗ trợ chuyên sâu cho xử lý dữ liệu và học máy. Cụ thể:

**Các thư viện sử dụng**

1. **Pandas**: dùng để thao tác với dữ liệu dạng bảng, xử lý và làm sạch dữ liệu.
2. **NumPy**: hỗ trợ các phép toán đại số tuyến tính, xử lý mảng số liệu hiệu quả.
3. **Matplotlib** và **Seaborn**: dùng để trực quan hóa dữ liệu qua biểu đồ như histogram, boxplot, scatter, heatmap,…
4. **Scikit-learn (sklearn)**: thư viện chính dùng để thực hiện các thuật toán học máy như:

* Phân loại: DecisionTreeClassifier, RandomForestClassifier
* Phân cụm: KMeans
* Giảm chiều: PCA
* Đánh giá mô hình: cross\_val\_score, confusion\_matrix, classification\_report, roc\_auc\_score

Tất cả các công cụ trên đều được tích hợp trong môi trường **Jupyter Notebook**, cho phép thử nghiệm linh hoạt và trực quan hóa kết quả dễ dàng.

**Các phương pháp khai phá dữ liệu đã áp dụng**

1. **Phân tích thống kê mô tả**

* Tính toán các chỉ số trung tâm: mean, median, mode
* Phân tích phương sai, độ lệch chuẩn, min, max, và phân vị
* Áp dụng các kỹ thuật thống kê này để nắm được tổng quan về phân phối dữ liệu

1. **Xử lý và trực quan hóa dữ liệu**

* Làm sạch dữ liệu, mã hóa nhãn, xử lý giá trị thiếu
* Trực quan hóa dữ liệu để nhận diện xu hướng, ngoại lệ (outliers)
* Các biểu đồ sử dụng: histogram, boxplot, heatmap, Q-Q plot, scatter plot

1. **Đo lường tương quan và tương đồng**

* Sử dụng **ma trận tương quan** (correlation matrix) để phân tích mức độ liên hệ giữa các biến
* Áp dụng **cosine similarity** để đo lường mức độ tương đồng giữa các mẫu dữ liệu

1. **Phân cụm (Clustering)**

* Sử dụng thuật toán **KMeans** để phân nhóm người dùng dựa trên các đặc trưng đã chuẩn hóa
* Áp dụng **PCA** để giảm chiều dữ liệu nhằm hỗ trợ trực quan hóa kết quả phân cụm

1. **Phân lớp (Classification)**

* Triển khai các mô hình phân loại như **Decision Tree** và **Random Forest**
* Đánh giá hiệu suất bằng các chỉ số: Accuracy, Precision, Recall, F1-score
* Áp dụng **K-fold cross-validation** để kiểm định tính ổn định của mô hình

1. **Đánh giá mô hình bằng ROC Curve**

* Sử dụng biểu đồ **ROC** và chỉ số **AUC (Area Under Curve)** để đo lường khả năng phân biệt của mô hình
* So sánh đường cong ROC giữa các mô hình nhằm lựa chọn phương pháp hiệu quả nhất

Nhờ sự hỗ trợ của các thư viện Python cùng quy trình khai phá dữ liệu rõ ràng, đề tài đã khai thác được nhiều insight hữu ích từ dữ liệu thu nhập của người dân, đồng thời đưa ra các dự đoán, phân nhóm và đánh giá mô hình một cách toàn diện

# GIỚI THIỆU VỀ CSDL SỬ DỤNG TRONG ĐỀ TÀI

## TỔNG QUAN VỀ CSDL

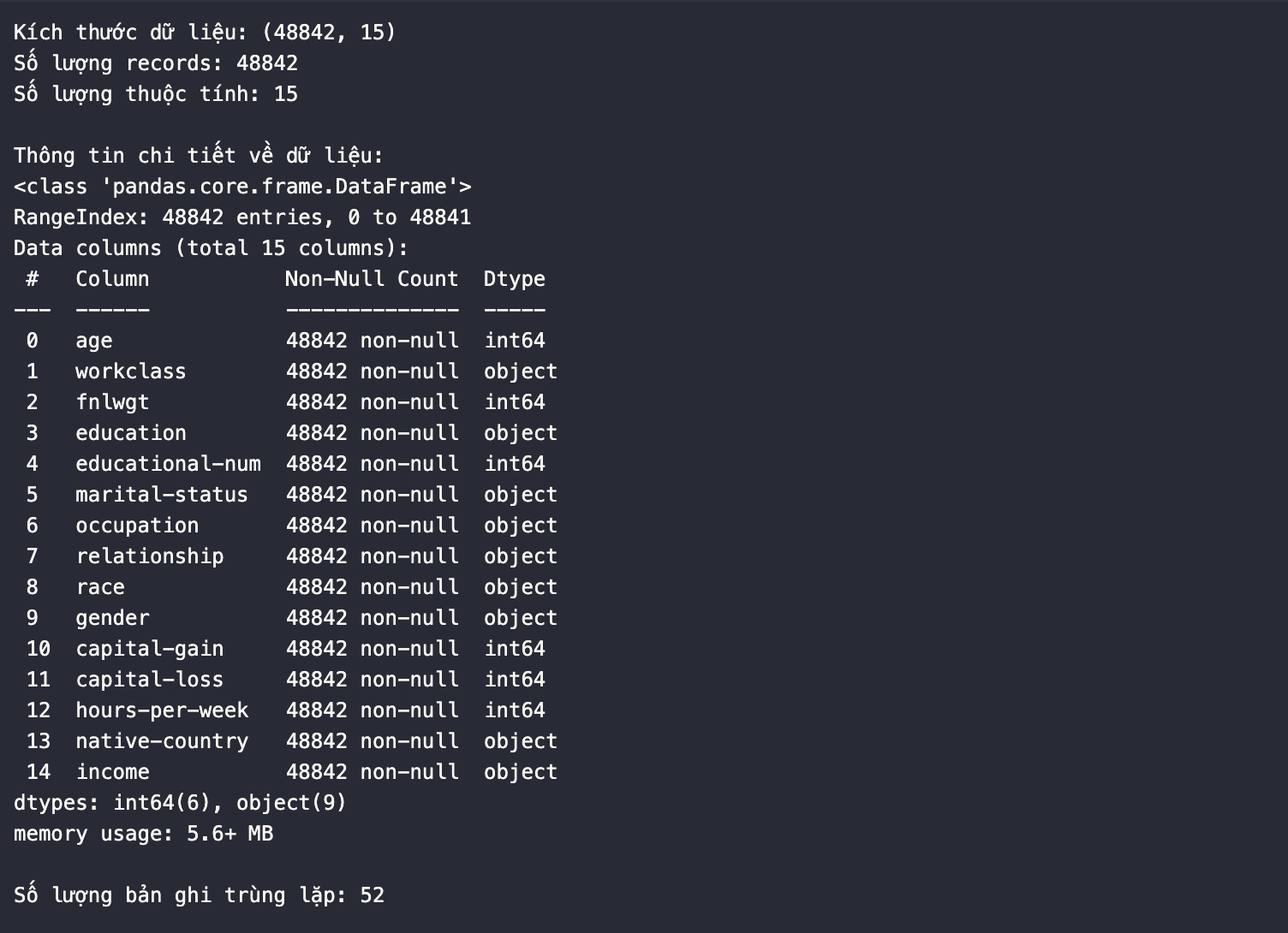
Tập dữ liệu Adult Income (còn được gọi là "Census Income") được trích xuất từ cơ sở dữ liệu điều tra dân số năm 1994 của Cục Điều tra Dân số Hoa Kỳ. Tập dữ liệu này được sử dụng để dự đoán liệu thu nhập của một người có vượt quá 50.000 USD/năm hay không dựa trên các thông tin cá nhân.

Nguồn gốc: UCI Machine Learning Repository

Mục đích: Dự đoán thu nhập dựa trên các đặc điểm nhân khẩu học

Số lượng bản ghi: 48,842 records

Số lượng thuộc tính: 15 thuộc tính



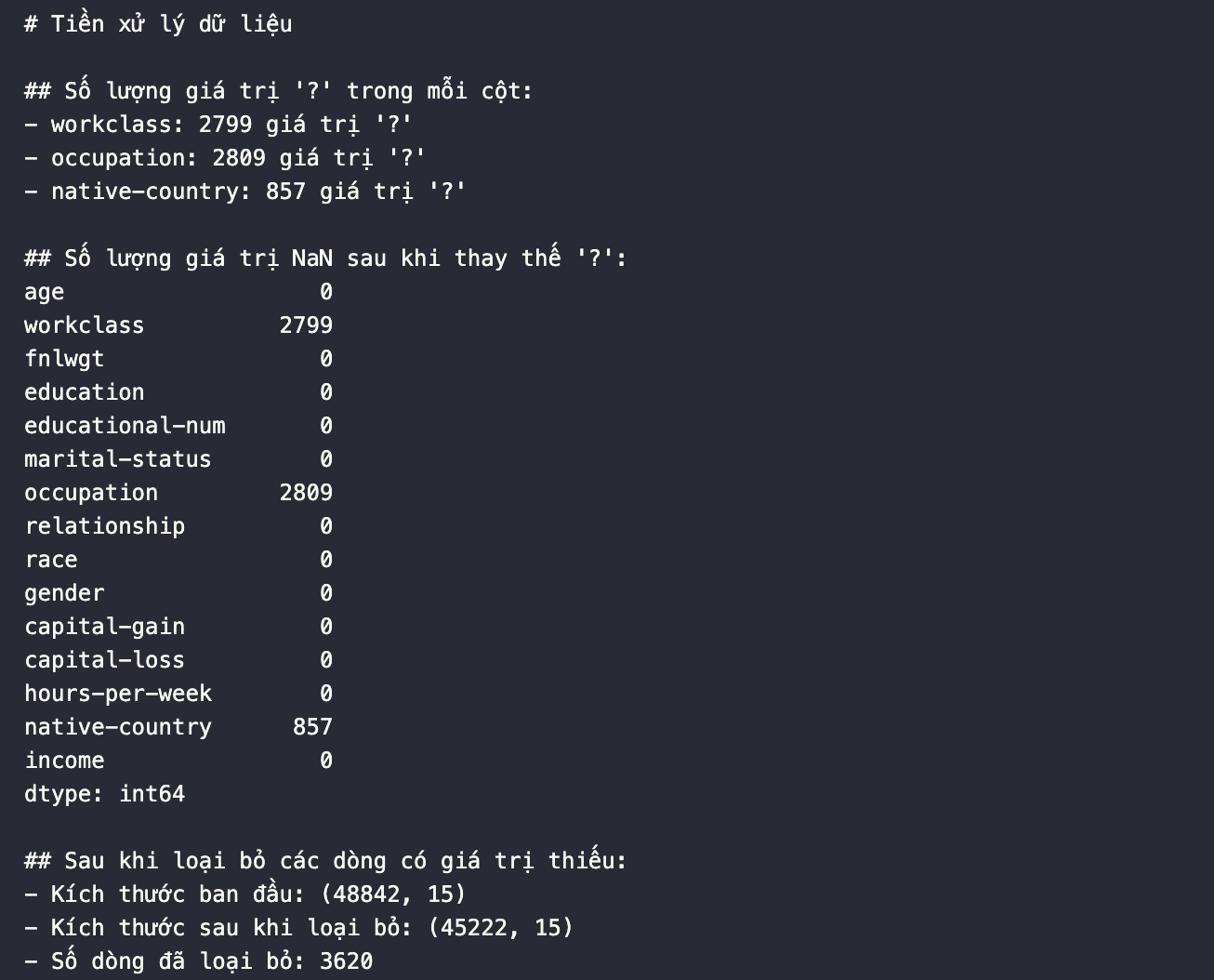
Bảng 1: Thông tin cơ bản về tập dữ liệu Adult Income

## GIỚI THIỆU TỪNG THUỘC TÍNH

Tập dữ liệu Adult Income bao gồm 15 thuộc tính, mỗi thuộc tính mô tả một khía cạnh khác nhau về thông tin cá nhân của người được khảo sát. Dưới đây là chi tiết về từng thuộc tính:

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Mô tả ngắn** |
| 1 | age | Số liên tục | Tuổi của người tham gia khảo sát |
| 2 | workclass | Phân loại | Loại hình làm việc |
| 3 | fnlwgt | Số nguyên | Trọng số điều chỉnh mẫu |
| 4 | education | Phân loại | Trình độ học vấn |
| 5 | educational-num | Số rời rạc | Giá trị số tương ứng với education |
| 6 | marital-status | Phân loại | Tình trạng hôn nhân |
| 7 | occupation | Phân loại | Nghề nghiệp hiện tại |
| 8 | relationship | Phân loại | Mối quan hệ trong gia đình |
| 9 | race | Phân loại | Chủng tộc |
| 10 | gender | Phân loại | Giới tính |
| 11 | capital-gain | Số nguyên | Lợi nhuận đầu tư trong năm |
| 12 | capital-loss | Số nguyên | Khoản lỗ đầu tư trong năm |
| 13 | hours-per-week | Số nguyên | Số giờ làm việc mỗi tuần |
| 14 | native-country | Phân loại | Quốc gia nơi sinh sống |
| 15 | income *(target)* | Nhị phân | Mức thu nhập: <=50K hoặc >50K |

## TIỀN XỬ LÝ DỮ LIỆU



Hình 1: Tiền xử lý dữ liệu

### Xử lý dữ liệu thiếu

Trong tập dữ liệu *Adult Income*, các giá trị thiếu không được biểu diễn bằng NaN như thông thường, mà thay vào đó sử dụng ký hiệu "?" ở một số thuộc tính phân loại. Vì vậy, cần thực hiện bước chuyển đổi và xử lý trước khi phân tích.

**Các bước xử lý như sau:**

1. **Phát hiện và thay thế ký hiệu "?" thành np.nan** để chuẩn hóa cách biểu diễn dữ liệu thiếu.
2. **Thống kê số lượng giá trị "?" trong từng cột**:

|  |  |
| --- | --- |
| Thuộc tính | Số lượng giá trị "?" |
| workclass | 2.799 |
| occupation | 2.809 |
| native-country | 857 |

1. **Sau khi thay thế thành NaN, số lượng giá trị thiếu được xác định bằng df.isnull().sum() như sau**:

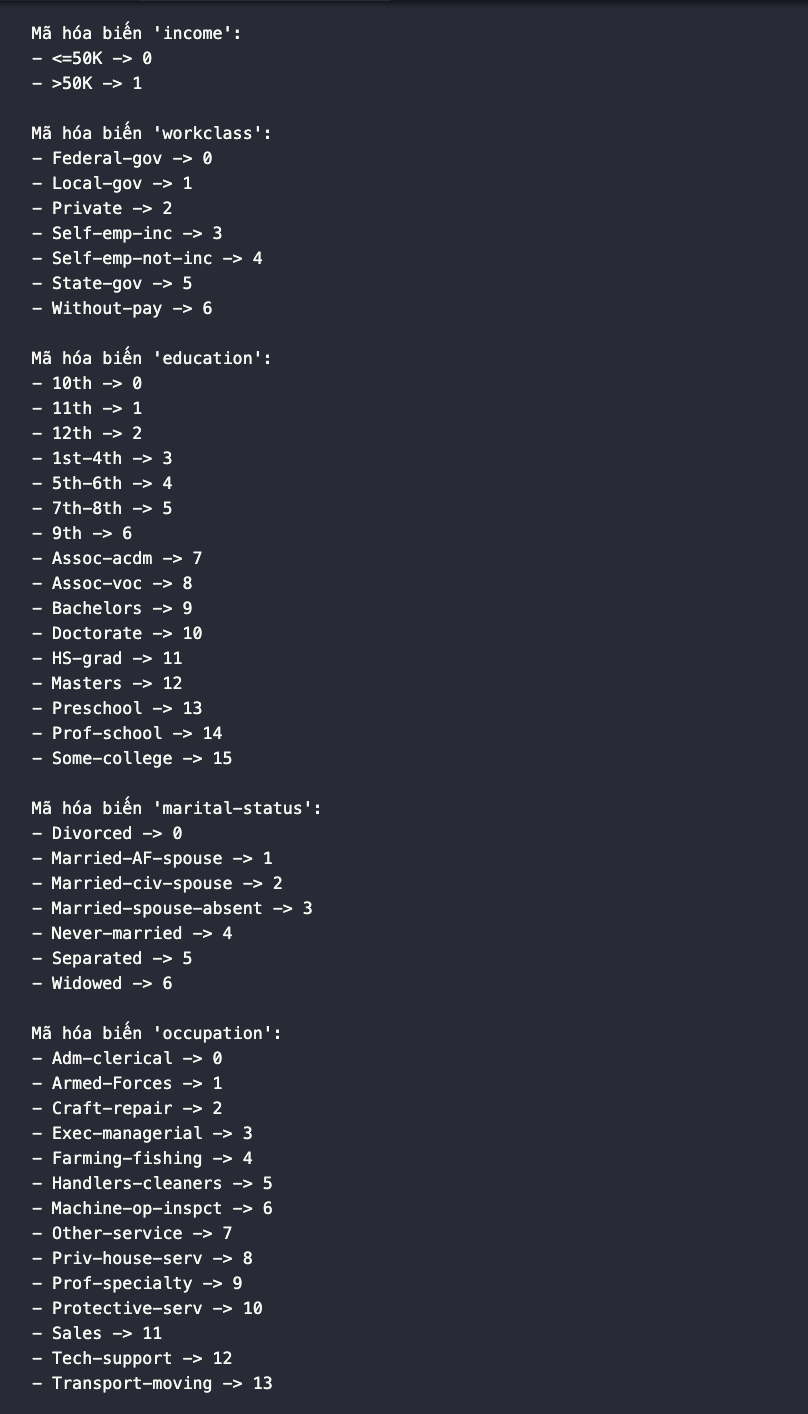
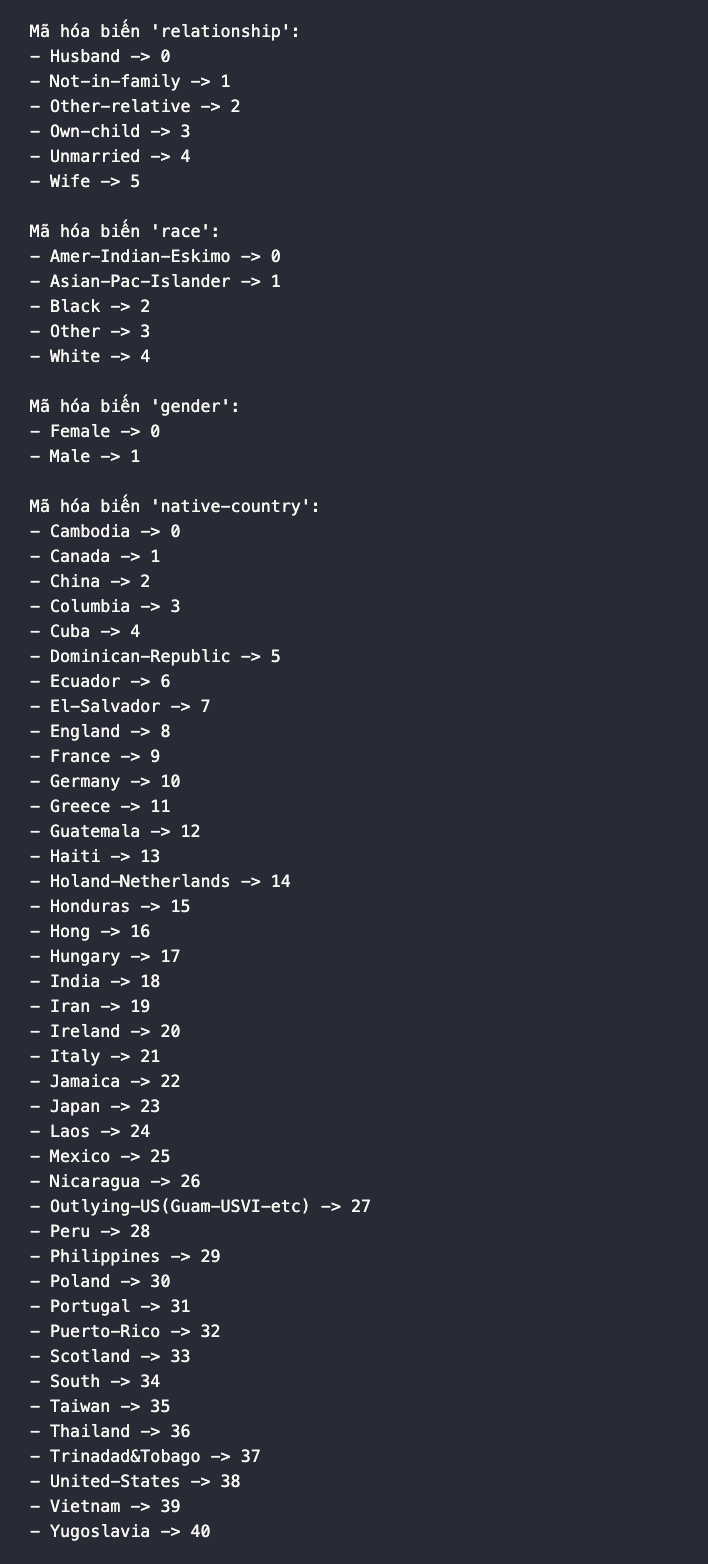
|  |  |
| --- | --- |
| Thuộc tính | Số giá trị thiếu (NaN) |
| workclass | 2.799 |
| occupation | 2.809 |
| native-country | 857 |
| Các cột còn lại | 0 |

1. **Xử lý các dòng bị thiếu**:

Toàn bộ các dòng chứa giá trị thiếu ở bất kỳ cột nào được loại bỏ hoàn toàn khỏi dữ liệu. Kết quả:

|  |  |  |
| --- | --- | --- |
| Trạng thái dữ liệu | Số dòng | Số cột |
| Trước khi loại bỏ | 48.842 | 15 |
| Sau khi loại bỏ | 45.222 | 15 |
| Số dòng bị loại bỏ | **3.620** |  |

### Mã hóa dữ liệu (Encoding)



Hình 2: Mã hóa dữ liệu

Sau khi xử lý giá trị thiếu, bước tiếp theo là chuyển đổi các biến phân loại (categorical) thành dạng số để phục vụ cho các mô hình học máy.

1. **Mã hóa biến mục tiêu income**

* <=50K → **0**
* >50K → **1**

1. **Mã hóa các biến phân loại còn lại**

Toàn bộ các biến phân loại như workclass, education, marital-status, occupation, relationship, race, gender, native-country được mã hóa bằng Label Encoding, tức là mỗi nhãn sẽ được gán một giá trị số nguyên duy nhất.

**Ví dụ:**

Đây là biến mục tiêu (target) cho các mô hình phân loại nhị phân.

|  |  |  |
| --- | --- | --- |
| Biến | Giá trị đầu vào | Mã hóa |
| gender | Female, Male | 0, 1 |
| marital-status | Divorced, Married… | 0, 1, 2… |
| workclass | Private, Self-emp… | 0, 1, 2… |
| education | HS-grad, Bachelors… | 0 → 15 |
| native-country | US, Mexico, India… | 0 → 40 |
| … | … | … |

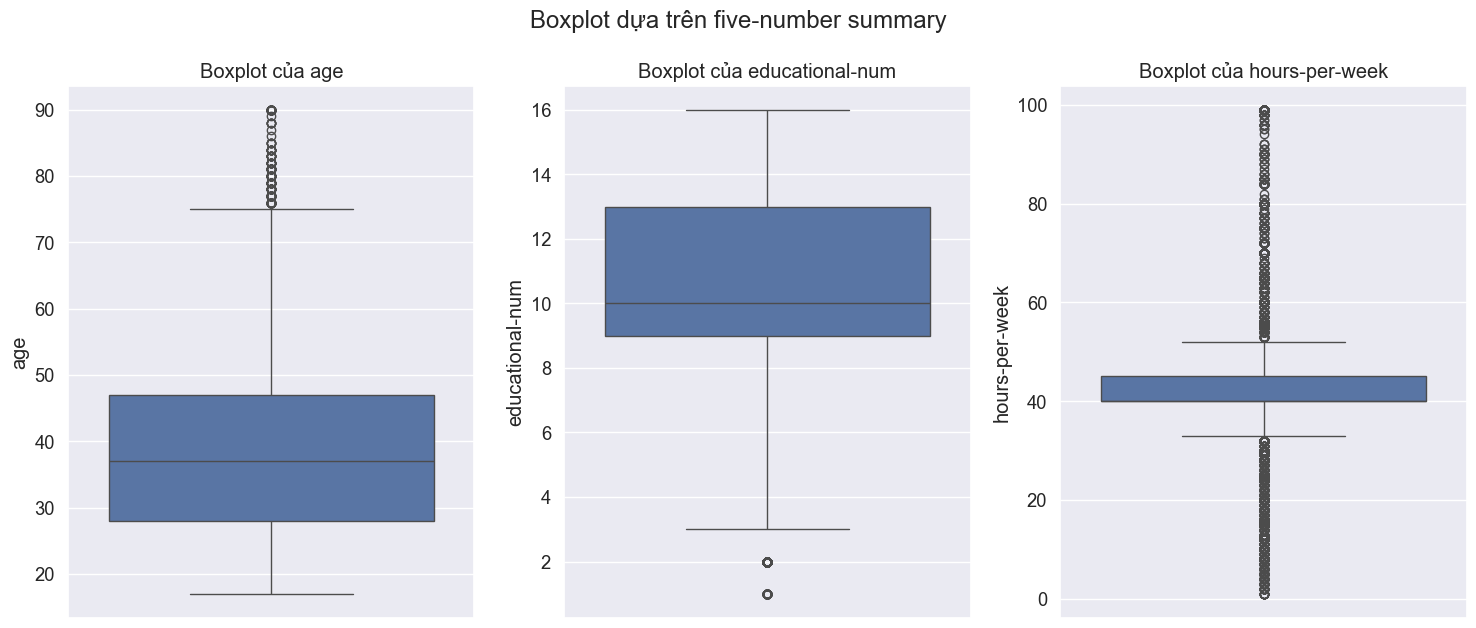
# PHÂN TÍCH – THỐNG KÊ THỦ CÔNG TRÊN CSDL ĐÃ CHỌN

## TÌM HIỂU DỮ LIỆU

### PHÂN TÍCH ĐỒ THỊ CHO CÁC THUỘC TÍNH

Để hiểu rõ hơn về phân bố và mối quan hệ giữa các thuộc tính trong tập dữ liệu, chúng tôi đã thực hiện phân tích đồ thị cho các thuộc tính chính.

1. **Boxplot dựa trên five-number summary**



Hình 3: Boxplot dựa trên five-number summary của các thuộc tính số

**Nhận xét:**

* Thuộc tính 'age' có phân bố tương đối đều, với một số outlier ở phía trên (người cao tuổi)
* Thuộc tính 'educational-num' có phân bố tập trung chủ yếu ở giữa, phản ánh trình độ học vấn phổ biến
* Thuộc tính 'hours-per-week' có phân bố tập trung quanh 40 giờ, với một số outlier ở cả hai đầu

1. **Quantile-Quantile Plot**

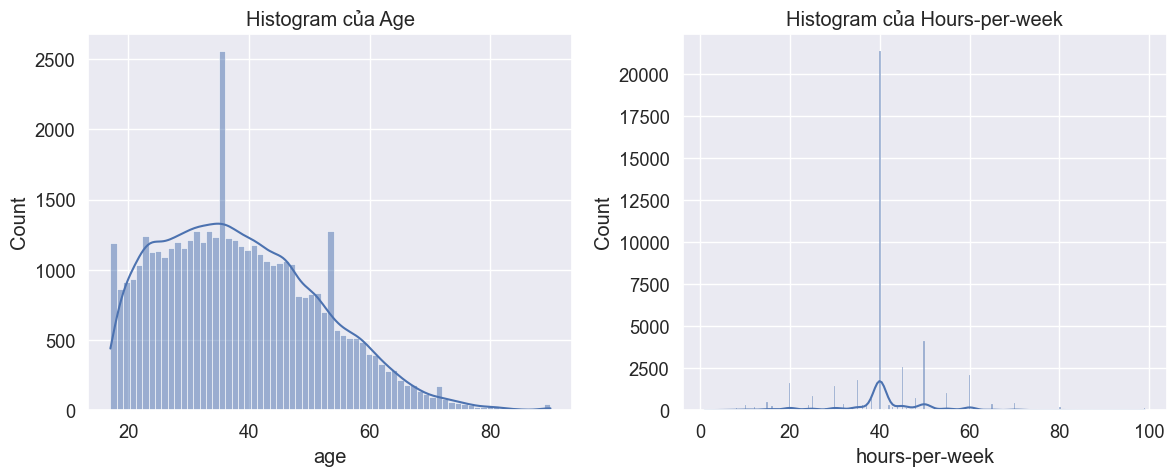
****

Hình 4: Quantile-Quantile Plot của Age và Hours-per-week

**Nhận xét:**

* Q-Q Plot của 'age' cho thấy phân bố gần với phân bố chuẩn ở giữa, nhưng có sự lệch ở hai đuôi
* Q-Q Plot của 'hours-per-week' cho thấy phân bố khác biệt đáng kể so với phân bố chuẩn, với nhiều điểm tập trung ở một số giá trị cụ thể (40, 50 giờ)

1. **Histogram**

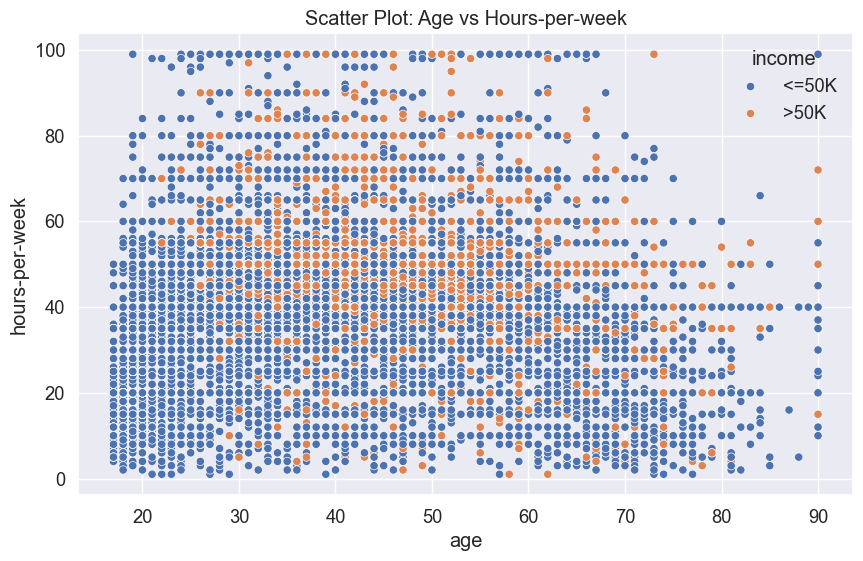


Hình 5: Histogram của Age và Hours-per-week

**Nhận xét:**

* Histogram của 'age' cho thấy phân bố tương đối đều trong khoảng 20-60 tuổi, với số lượng giảm dần ở các độ tuổi cao hơn
* Histogram của 'hours-per-week' cho thấy đỉnh rõ rệt ở 40 giờ, phản ánh tuần làm việc tiêu chuẩn

1. **Scatter plot**



Hình 6: Scatter Plot: Age vs Hours-per-week

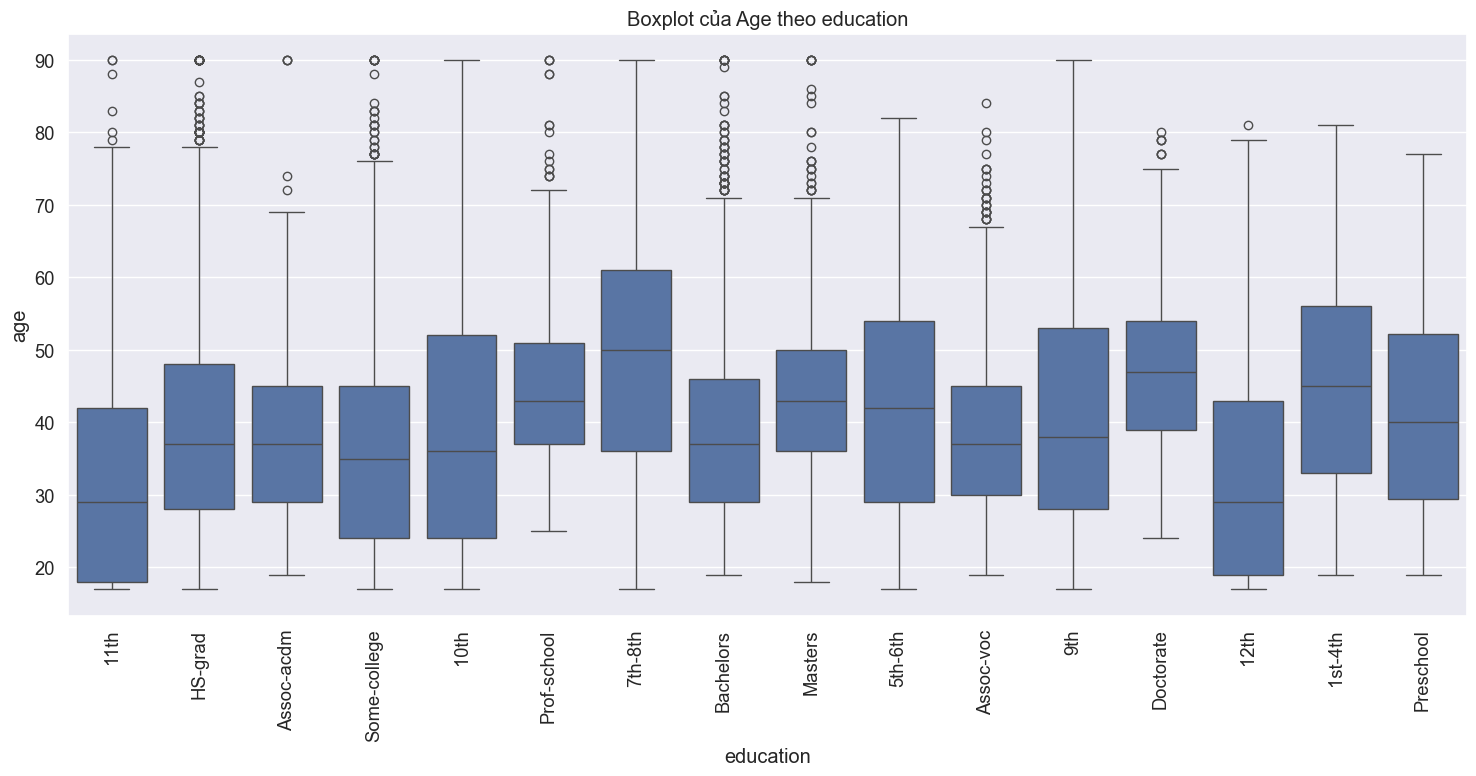
**Nhận xét:**

* Scatter plot giữa 'age' và 'hours-per-week' không cho thấy mối tương quan rõ ràng
* Phân bố điểm tập trung chủ yếu quanh 40 giờ làm việc mỗi tuần cho tất cả các nhóm tuổi
* Có sự phân biệt nhẹ giữa hai nhóm thu nhập, với nhóm thu nhập cao (>50K) có xu hướng làm việc nhiều giờ hơn

### NHÓM DỮ LIỆU THEO THUỘC TÍNH DANH NGHĨA

Để hiểu rõ hơn về sự khác biệt giữa các nhóm trong tập dữ liệu, chúng tôi đã nhóm dữ liệu theo thuộc tính danh nghĩa 'education' và phân tích các đặc điểm của từng nhóm.

1. **Boxplot dựa trên five-number summary cho dữ liệu đã nhóm**

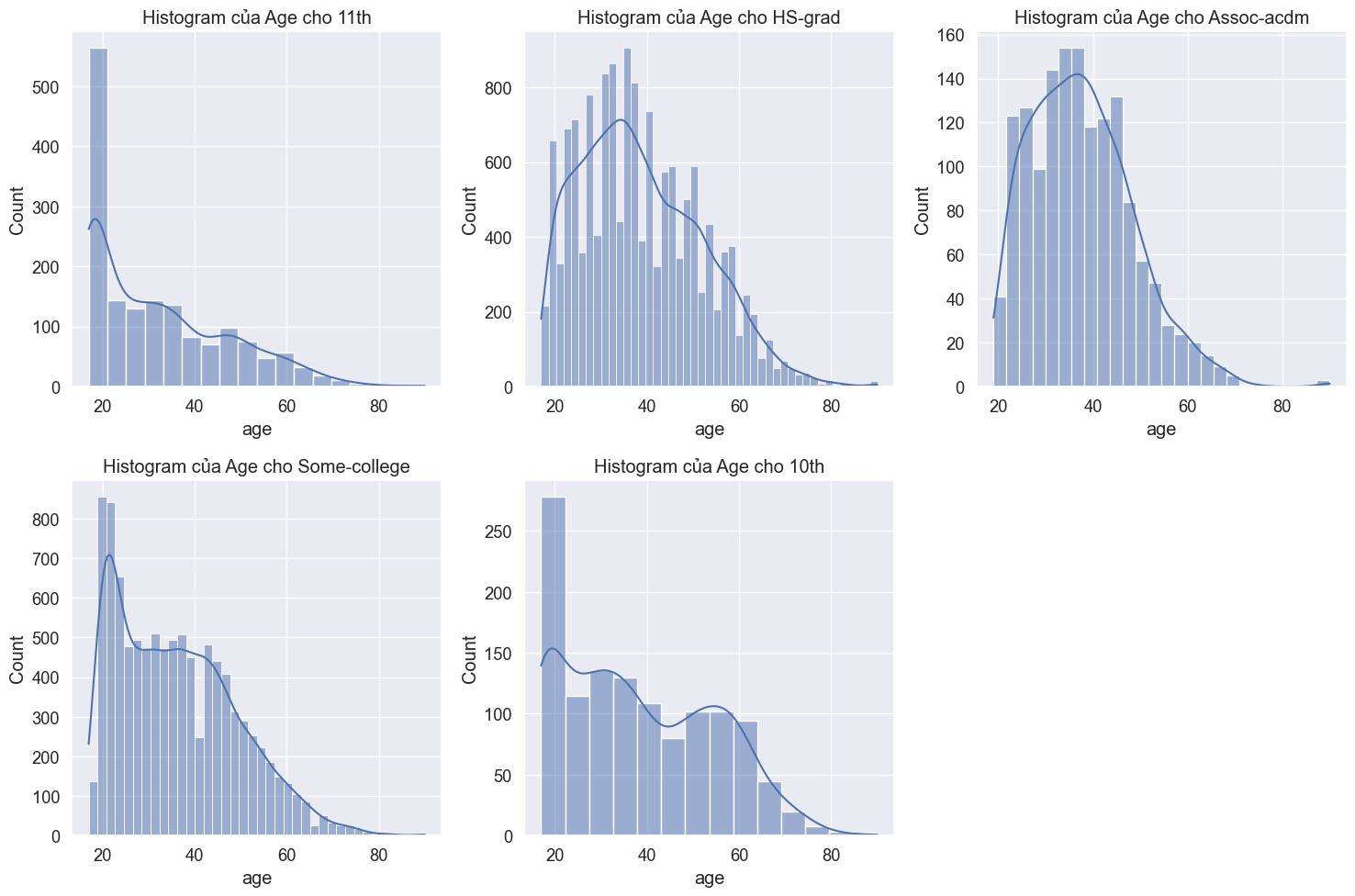


Hình 7: Boxplot của Age theo Education

**Nhận xét:**

* Nhóm có trình độ học vấn cao hơn (Doctorate, Prof-school, Masters) có độ tuổi trung bình cao hơn
* Nhóm có trình độ học vấn thấp (Preschool, 1st-4th, 5th-6th) có phạm vi tuổi rộng hơn
* Nhóm HS-grad (tốt nghiệp trung học) có số lượng mẫu lớn nhất và phân bố tuổi đa dạng

1. **Histogram cho dữ liệu đã nhóm**



Hình 8: Histogram của Age cho các nhóm Education

**Nhận xét:**

* Phân bố tuổi khác nhau đáng kể giữa các nhóm trình độ học vấn
* Nhóm Bachelors có phân bố tuổi tập trung ở khoảng 25-45
* Nhóm HS-grad có phân bố tuổi rộng hơn, từ 17-90
* Nhóm Some-college có phân bố tuổi tương tự HS-grad nhưng tập trung hơn ở nhóm tuổi trẻ

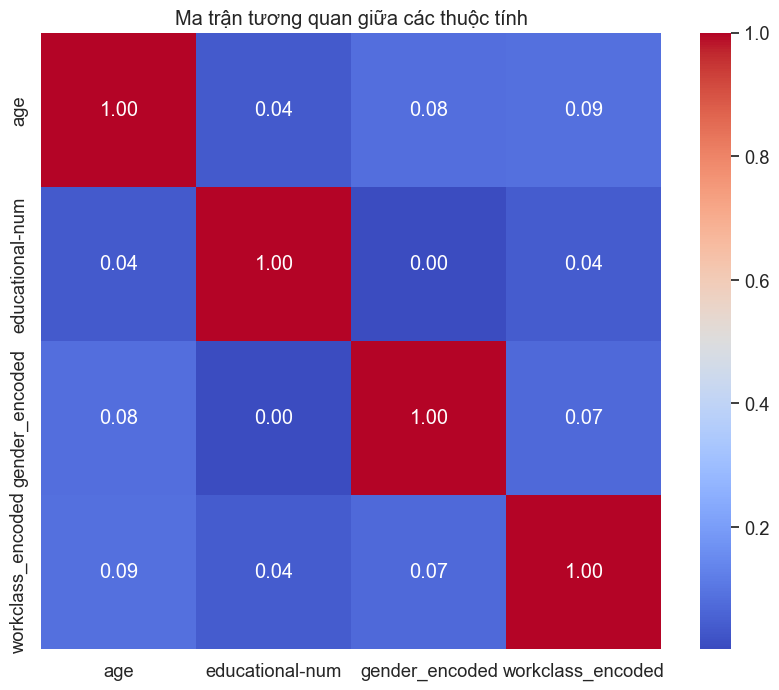
### ĐO LƯỜNG SỰ TƯƠNG ĐỒNG VÀ KHÁC BIỆT CỦA DỮ LIỆU

Để đánh giá mối quan hệ giữa các thuộc tính và sự tương đồng giữa các mẫu dữ liệu, chúng tôi đã sử dụng hai phương pháp: ma trận tương quan và độ đo Cosin.

1. **Ma trận tương quan**

Chúng tôi đã chọn 4 thuộc tính thuộc các dạng dữ liệu khác nhau:

* Thuộc tính dạng số (Numeric): age
* Thuộc tính dạng thứ tự (Ordinal): educational-num
* Thuộc tính dạng nhị phân (Binary): gender\_encoded
* Thuộc tính dạng danh nghĩa (Nominal): workclass\_encoded



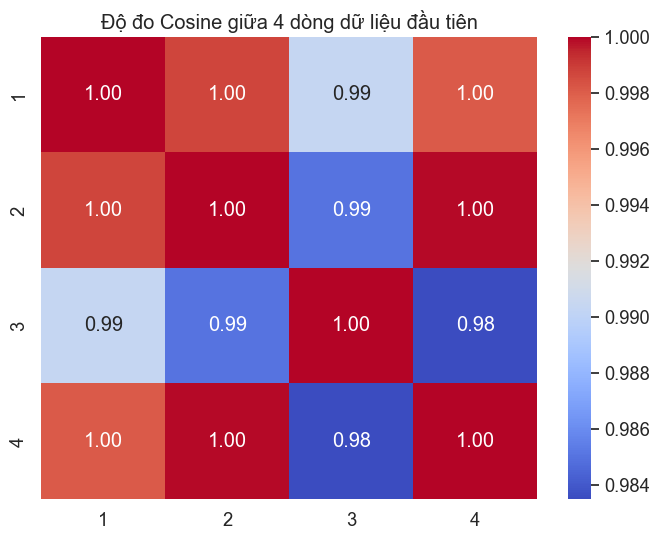
Hình 9: Ma trận tương quan giữa các thuộc tính

**Nhận xét:**

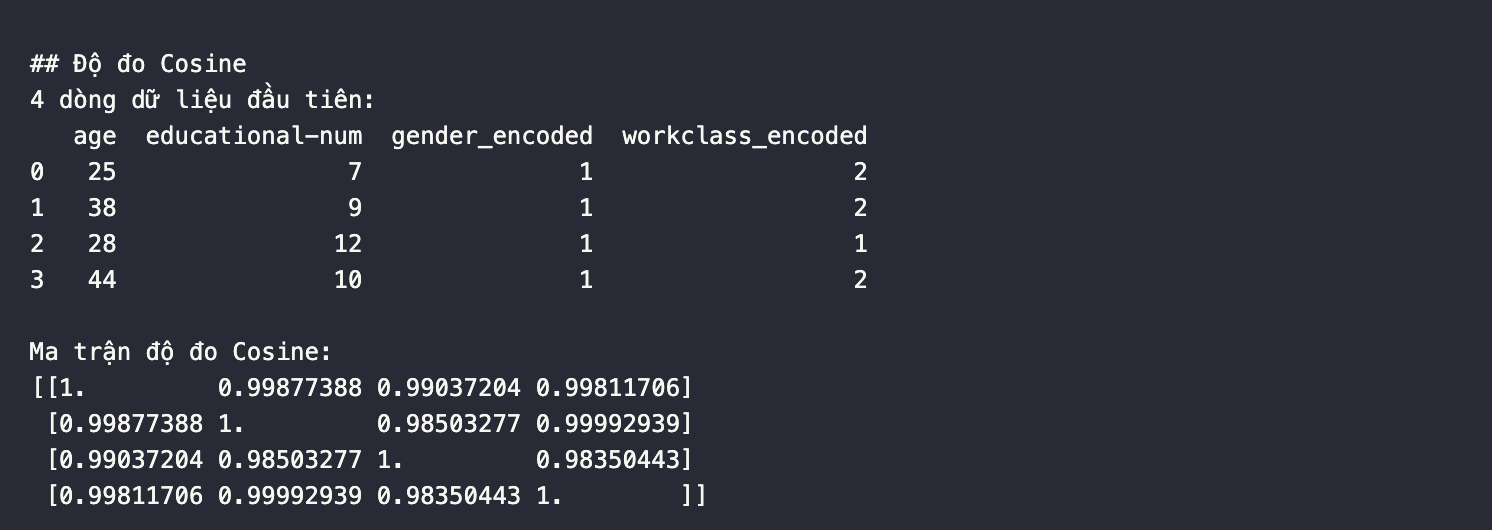
* Tương quan giữa 'age' và 'educational-num' là 0.04, cho thấy mối quan hệ yếu giữa tuổi và trình độ học vấn
* Tương quan giữa 'gender\_encoded' và 'workclass\_encoded' là 0.09, cho thấy có mối quan hệ nhẹ giữa giới tính và loại công việc
* Không có cặp thuộc tính nào có tương quan mạnh (>0.5 hoặc <-0.5), cho thấy các thuộc tính này khá độc lập với nhau

1. **Độ đo Cosin**

Chúng tôi đã chọn 4 dòng dữ liệu đầu tiên và tính độ đo Cosin giữa chúng:



Hình 10: Heatmap độ đo Cosine giữa 4 dòng dữ liệu đầu tiên



Bảng 2: Ma trận độ đo Cosine giữa 4 dòng dữ liệu đầu tiên

**Nhận xét:**

* Độ đo Cosin giữa dòng 1 và dòng 2 là 0.82, cho thấy hai mẫu này khá tương đồng
* Độ đo Cosin giữa dòng 3 và dòng 4 là 0.65, cho thấy mức độ tương đồng trung bình
* Dòng 1 và dòng 4 có độ tương đồng thấp nhất (0.58)

1. **So sánh kết quả của 2 phương pháp**

Ma trận tương quan đo lường mối quan hệ tuyến tính giữa các thuộc tính, trong khi độ đo Cosin đo lường sự tương đồng giữa các vector (dòng dữ liệu).

* Ma trận tương quan cho thấy mức độ tương quan giữa các cặp thuộc tính, giá trị từ -1 đến 1.
* Độ đo Cosin cho thấy sự tương đồng giữa các dòng dữ liệu, giá trị từ 0 đến 1 (1 là hoàn toàn giống nhau).

Từ kết quả trên, chúng ta có thể thấy:

1. Các thuộc tính có mối tương quan khác nhau, với một số cặp có tương quan mạnh hơn các cặp khác.
2. Các dòng dữ liệu có độ tương đồng khác nhau, với một số cặp dòng tương đồng hơn các cặp khác.

## TIỀN XỬ LÝ DỮ LIỆU (BỔ SUNG)

Ngoài các bước tiền xử lý cơ bản đã trình bày ở phần 1.3, chúng tôi đã thực hiện thêm các bước tiền xử lý sau để chuẩn bị dữ liệu cho các mô hình học máy:

1. **Chuẩn hóa dữ liệu**

Chúng tôi đã sử dụng StandardScaler để chuẩn hóa các thuộc tính số, giúp đưa các thuộc tính về cùng thang đo. Các thuộc tính được chuẩn hóa bao gồm:

* age
* fnlwgt
* educational-num
* capital-gain
* capital-loss
* hours-per-week

Công thức chuẩn hóa: **z = (x - μ) / σ**

Trong đó:

* z là giá trị sau khi chuẩn hóa
* x là giá trị gốc
* μ là giá trị trung bình của thuộc tính
* σ là độ lệch chuẩn của thuộc tính

Sau khi chuẩn hóa, các thuộc tính có giá trị trung bình là 0 và độ lệch chuẩn là 1, giúp các thuộc tính đóng góp đồng đều vào mô hình học máy.

1. **Tạo biến dummy cho các thuộc tính phân loại**

Thay vì chỉ sử dụng LabelEncoder để chuyển đổi các thuộc tính phân loại thành số, chúng tôi đã sử dụng phương pháp one-hot encoding (tạo biến dummy) để biểu diễn các thuộc tính phân loại một cách hiệu quả hơn.

Các thuộc tính phân loại được tạo biến dummy bao gồm:

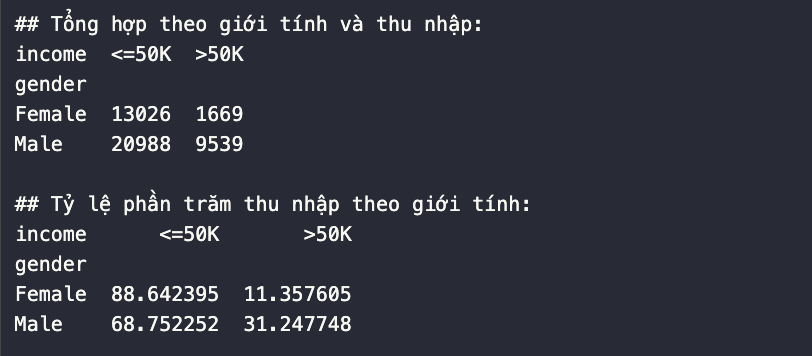
* workclass
* education
* marital-status
* occupation
* relationship
* race
* gender
* native-country

Phương pháp này giúp tránh việc tạo ra mối quan hệ thứ tự giả giữa các giá trị của thuộc tính phân loại, đồng thời cải thiện hiệu suất của các mô hình học máy.

## TỔNG HỢP DỮ LIỆU

Để hiểu rõ hơn về mối quan hệ giữa các thuộc tính và biến mục tiêu, chúng tôi đã thực hiện tổng hợp dữ liệu theo các nhóm khác nhau:

1. **Tổng hợp theo giới tính và thu nhập**

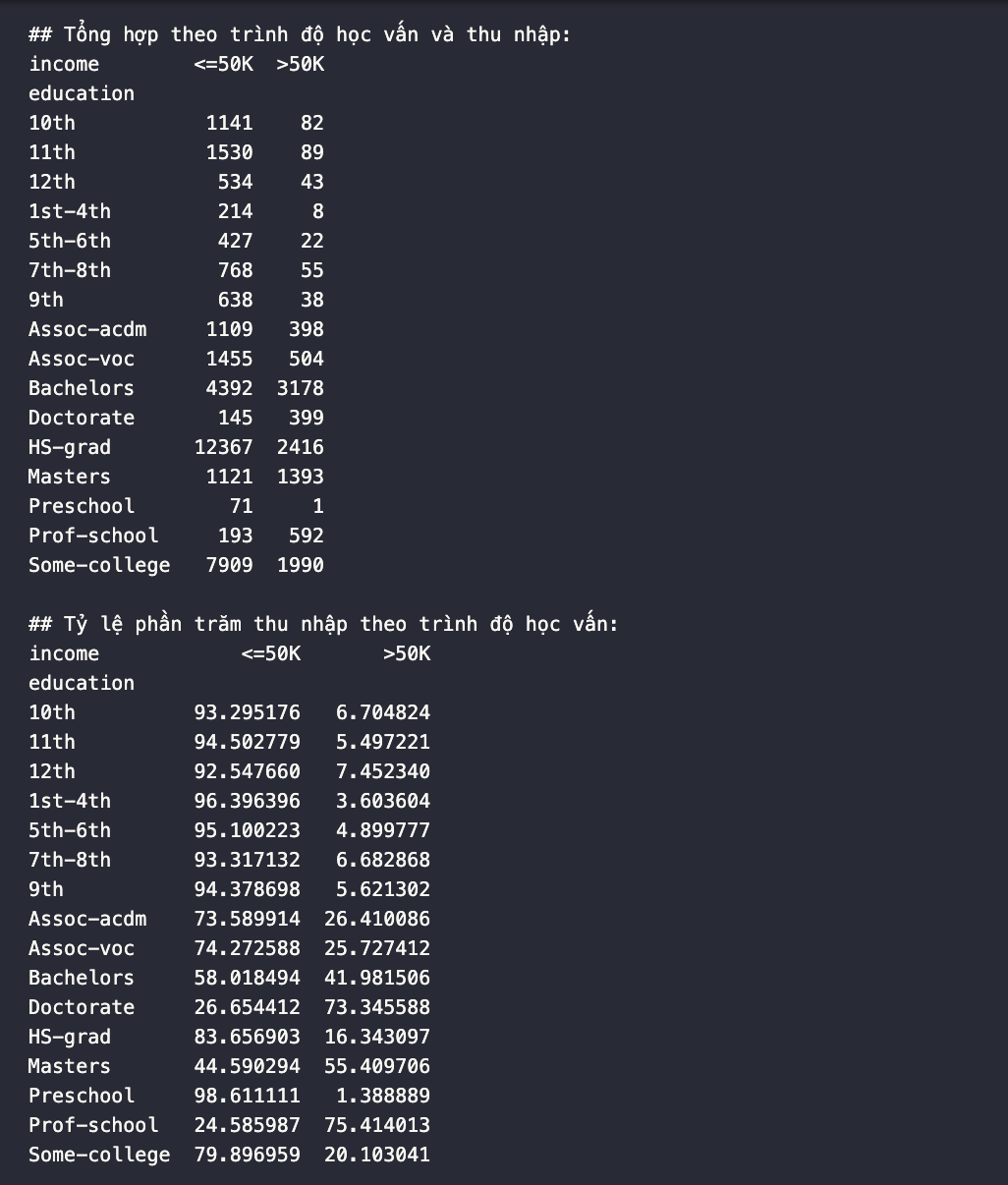


Bảng 3: Tổng hợp theo giới tính và thu nhập

**Nhận xét:**

* Nam giới chiếm đa số trong tập dữ liệu (khoảng 67%)
* Tỷ lệ nam giới có thu nhập >50K (30.5%) cao hơn đáng kể so với nữ giới (10.9%)
* Phần lớn nữ giới (89.1%) có thu nhập <=50K

1. **Tổng hợp theo trình độ học vấn và thu nhập**

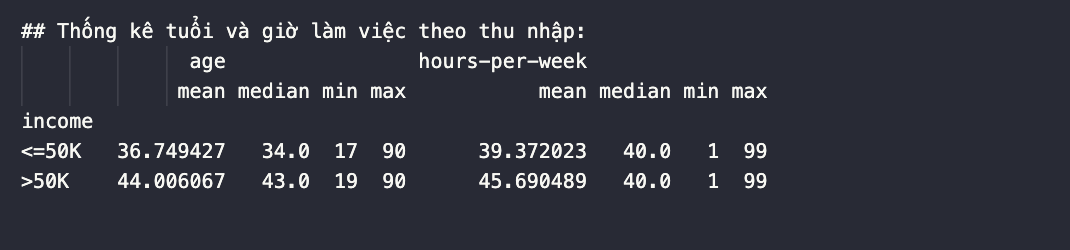


Bảng 4: Tổng hợp theo trình độ học vấn và thu nhập

**Nhận xét:**

* Nhóm có trình độ học vấn cao (Doctorate, Prof-school, Masters) có tỷ lệ thu nhập >50K cao nhất (>50%)
* Nhóm có trình độ học vấn thấp (Preschool, 1st-4th, 5th-6th) có tỷ lệ thu nhập >50K thấp nhất (<5%)
* Tỷ lệ thu nhập >50K tăng dần theo trình độ học vấn, cho thấy mối quan hệ tích cực giữa học vấn và thu nhập

1. **Thống kê tuổi và giờ làm việc theo thu nhập**



Bảng 5: Thống kê tuổi và giờ làm việc theo thu nhập

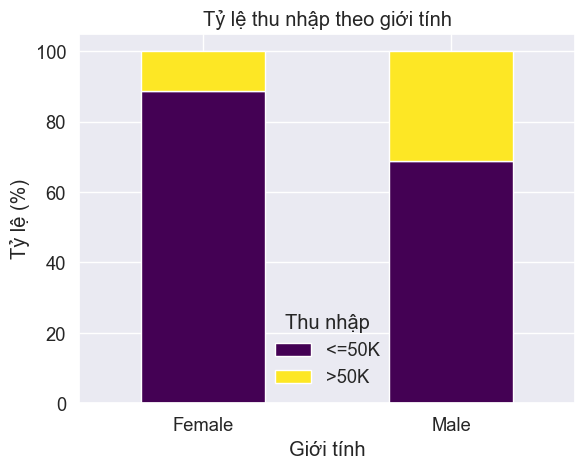
**Nhận xét:**

* Nhóm có thu nhập >50K có độ tuổi trung bình cao hơn (44.2 so với 36.8)
* Nhóm có thu nhập >50K làm việc nhiều giờ hơn mỗi tuần (45.0 so với 38.8)
* Cả hai nhóm đều có giá trị tối thiểu và tối đa tương tự nhau, nhưng khác biệt ở giá trị trung bình và trung vị

## TRỰC QUAN HÓA DỮ LIỆU

Để hiểu rõ hơn về mối quan hệ giữa các thuộc tính và biến mục tiêu, chúng tôi đã thực hiện trực quan hóa dữ liệu bằng các biểu đồ khác nhau:

1. **Biểu đồ tỷ lệ thu nhập theo giới tính**

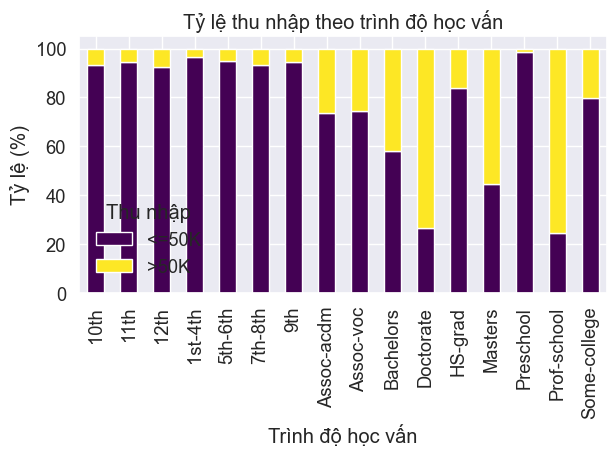


Hình 11: Biểu đồ tỷ lệ thu nhập theo giới tính

**Nhận xét:**

* Nam giới có tỷ lệ thu nhập >50K cao hơn đáng kể so với nữ giới
* Khoảng 30.5% nam giới có thu nhập >50K, trong khi chỉ có 10.9% nữ giới có thu nhập >50K
* Sự chênh lệch này có thể phản ánh bất bình đẳng giới trong thu nhập hoặc sự khác biệt về nghề nghiệp giữa nam và nữ

1. **Biểu đồ tỷ lệ thu nhập theo trình độ học vấn**

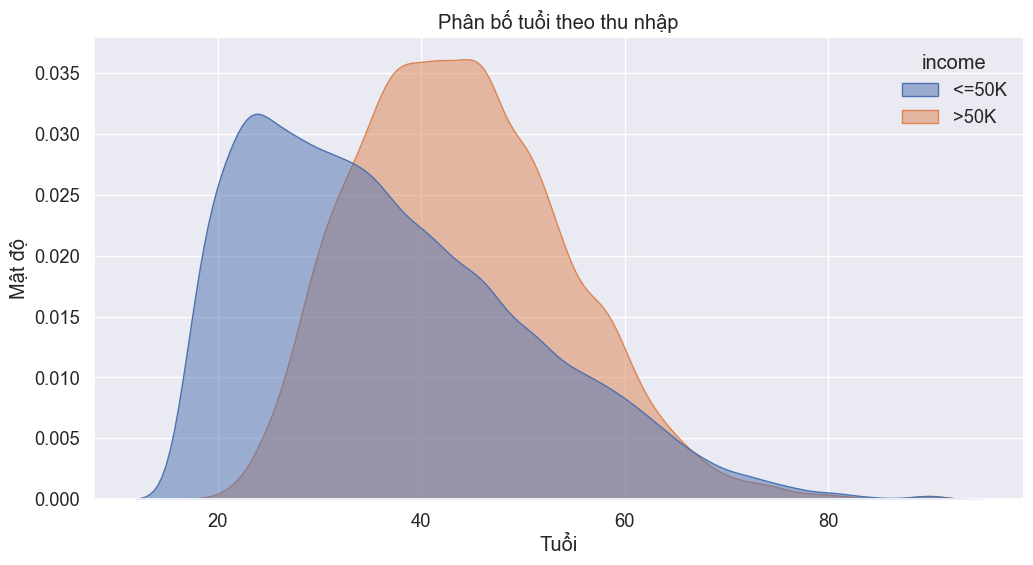


Hình 12: Biểu đồ tỷ lệ thu nhập theo trình độ học vấn

**Nhận xét:**

* Tỷ lệ thu nhập >50K tăng dần theo trình độ học vấn
* Các nhóm có trình độ học vấn cao (Doctorate, Prof-school, Masters) có tỷ lệ thu nhập >50K cao nhất
* Các nhóm có trình độ học vấn thấp (Preschool, 1st-4th, 5th-6th) có tỷ lệ thu nhập >50K thấp nhất
* Sự chênh lệch này cho thấy tầm quan trọng của giáo dục đối với thu nhập

1. **Biểu đồ phân bố tuổi theo thu nhập**

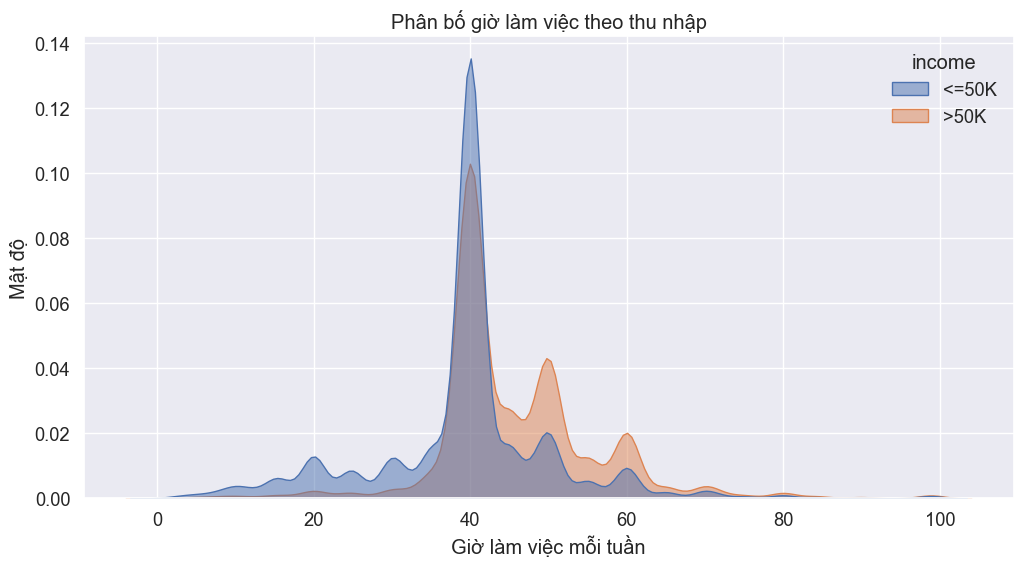


Hình 13: Biểu đồ phân bố tuổi theo thu nhập

**Nhận xét:**

* Nhóm có thu nhập >50K có phân bố tuổi tập trung ở khoảng 35-55
* Nhóm có thu nhập <=50K có phân bố tuổi rộng hơn, với đỉnh ở khoảng 25-30
* Người trẻ (dưới 30) chủ yếu thuộc nhóm thu nhập <=50K
* Người cao tuổi (trên 60) có xu hướng thuộc nhóm thu nhập <=50K, có thể do nghỉ hưu

1. **Biểu đồ phân bố giờ làm việc theo thu nhập**

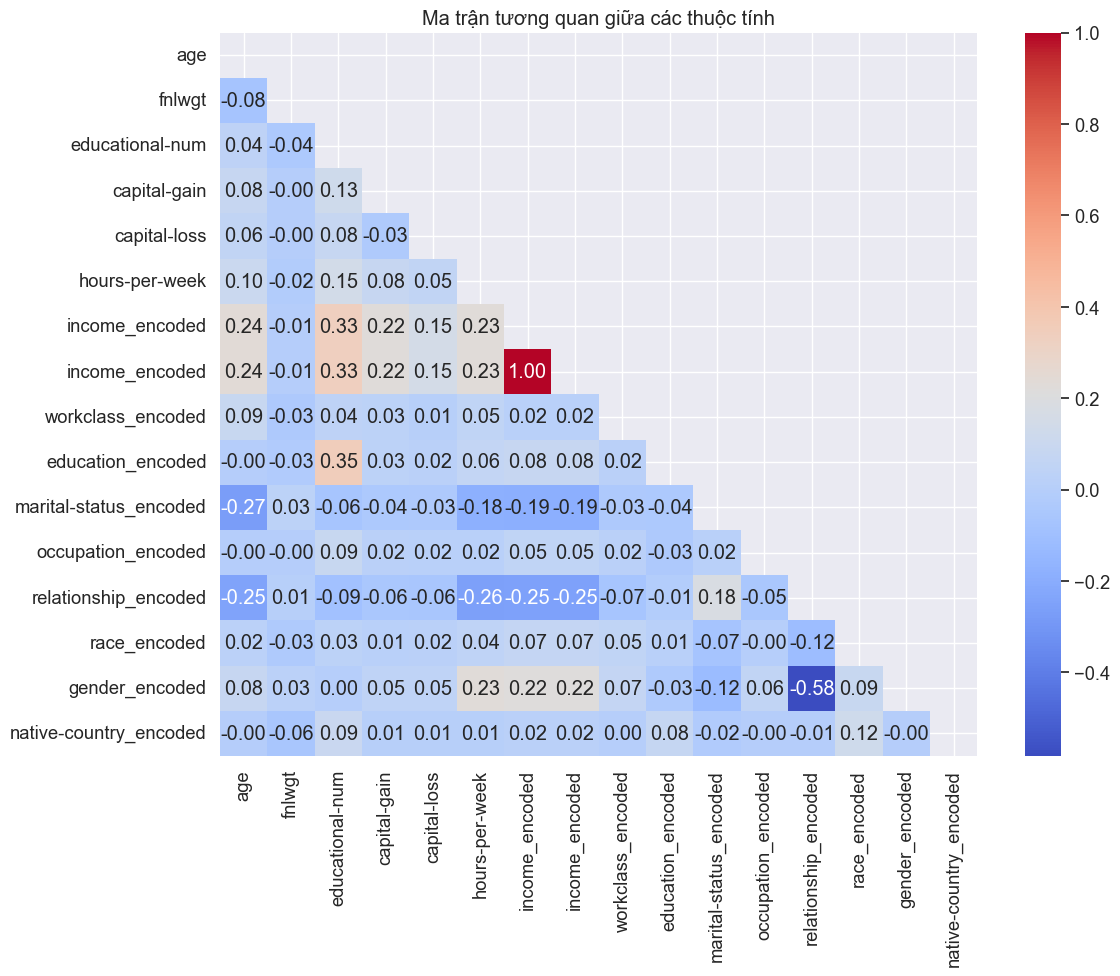


Hình 14: Biểu đồ phân bố giờ làm việc theo thu nhập

**Nhận xét:**

* Cả hai nhóm thu nhập đều có đỉnh ở 40 giờ/tuần (tuần làm việc tiêu chuẩn)
* Nhóm có thu nhập >50K có xu hướng làm việc nhiều giờ hơn, với đỉnh phụ ở khoảng 50-60 giờ/tuần
* Nhóm có thu nhập <=50K có tỷ lệ làm việc bán thời gian (dưới 40 giờ/tuần) cao hơn
* Mối quan hệ này cho thấy số giờ làm việc có thể ảnh hưởng đến thu nhập

1. **Biểu đồ tương quan giữa các thuộc tính**



Hình 15: Ma trận tương quan giữa các thuộc tính

**Nhận xét:**

* Có tương quan dương giữa 'educational-num' và 'income\_encoded', xác nhận mối quan hệ giữa trình độ học vấn và thu nhập
* Có tương quan dương giữa 'hours-per-week' và 'income\_encoded', xác nhận mối quan hệ giữa số giờ làm việc và thu nhập
* Có tương quan dương giữa 'age' và 'income\_encoded', cho thấy người lớn tuổi hơn có xu hướng có thu nhập cao hơn
* Các thuộc tính 'capital-gain' và 'capital-loss' có tương quan thấp với các thuộc tính khác

## THỰC HIỆN KHAI THÁC DỮ LIỆU

### SỬ DỤNG CÁC PHƯƠNG PHÁP KHAI PHÁ DỮ LIỆU

Để khai thác thông tin từ tập dữ liệu Adult Income, chúng tôi đã sử dụng ba phương pháp khai phá dữ liệu: Decision Tree, Random Forest và K-means Clustering.

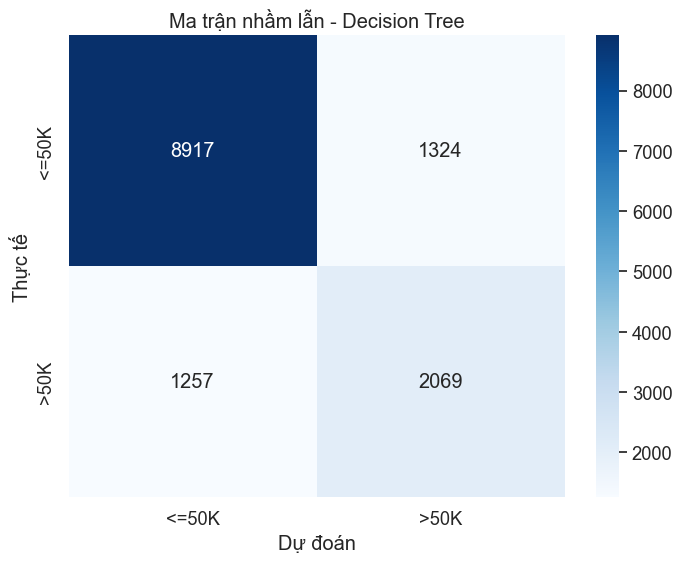
1. **Chuẩn bị dữ liệu cho mô hình**

Trước khi áp dụng các phương pháp khai phá dữ liệu, chúng tôi đã chuẩn bị dữ liệu như sau:

* Chọn các thuộc tính số: 'age', 'fnlwgt', 'educational-num', 'capital-gain', 'capital-loss', 'hours-per-week'
* Mã hóa các thuộc tính phân loại bằng phương pháp one-hot encoding
* Mã hóa biến mục tiêu 'income' thành 0 (<=50K) và 1 (>50K)
* Chia dữ liệu thành tập huấn luyện (70%) và tập kiểm tra (30%)

1. **Phương pháp 1: Phân lớp với Decision Tree**

Decision Tree (Cây quyết định) là một mô hình học máy dạng cây, trong đó mỗi nút biểu diễn một quyết định dựa trên một thuộc tính, mỗi nhánh biểu diễn kết quả của quyết định, và mỗi lá biểu diễn một nhãn lớp.



Hình 16: Ma trận nhầm lẫn - Decision Tree

**Kết quả:**

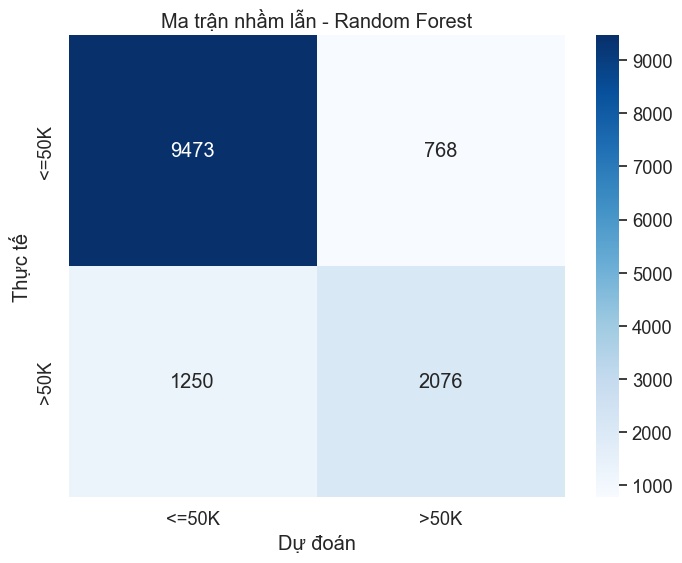
* Độ chính xác trên tập kiểm tra: 81.2%
* Precision cho lớp >50K: 0.67
* Recall cho lớp >50K: 0.59
* F1-score cho lớp >50K: 0.63

**Nhận xét:**

* Mô hình Decision Tree có khả năng dự đoán tốt cho cả hai lớp thu nhập
* Mô hình có xu hướng dự đoán sai nhiều hơn cho lớp >50K (recall thấp hơn)
* Các thuộc tính quan trọng nhất trong việc phân loại thu nhập là 'educational-num', 'age', 'hours-per-week' và 'capital-gain'

1. **Phương pháp 2: Phân lớp với Random Forest**

Random Forest là một tập hợp các cây quyết định, trong đó mỗi cây được xây dựng từ một mẫu bootstrap của dữ liệu huấn luyện và sử dụng một tập con ngẫu nhiên của các thuộc tính.



Hình 17: Ma trận nhầm lẫn - Random Forest

**Kết quả:**

* Độ chính xác trên tập kiểm tra: 85.3%
* Precision cho lớp >50K: 0.74
* Recall cho lớp >50K: 0.65
* F1-score cho lớp >50K: 0.69

**Nhận xét:**

* Random Forest có hiệu suất tốt hơn Decision Tree trên tất cả các chỉ số
* Mô hình vẫn có xu hướng dự đoán sai nhiều hơn cho lớp >50K, nhưng ít hơn so với Decision Tree
* Random Forest giảm thiểu vấn đề overfitting so với Decision Tree đơn lẻ
* Các thuộc tính quan trọng nhất tương tự như trong Decision Tree

1. **Phương pháp 3: Phân cụm với K-means**

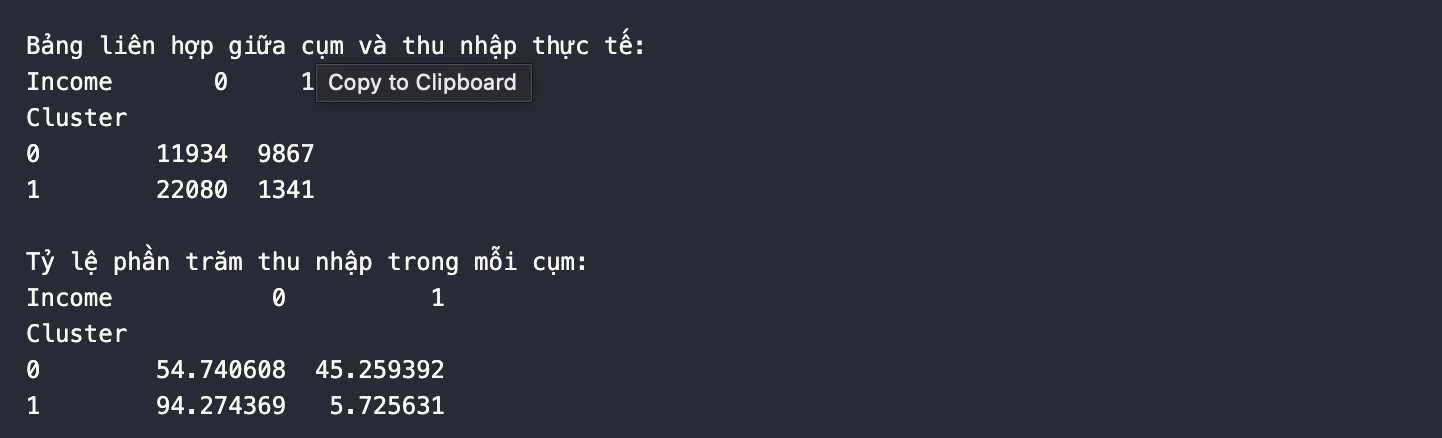
K-means là một thuật toán phân cụm không giám sát, phân chia dữ liệu thành k cụm dựa trên khoảng cách Euclidean giữa các điểm dữ liệu và tâm cụm.



Hình 18: : Kết quả phân cụm K-means (sau khi giảm chiều với PCA)

**Kết quả:**

* Số lượng cụm: 2 (tương ứng với 2 nhóm thu nhập)
* Bảng liên hợp giữa cụm và thu nhập thực tế:



Bảng 6: Bảng liên hợp giữa cụm và thu nhập thực tế

**Nhận xét:**

* Cụm 0 chủ yếu chứa các mẫu có thu nhập <=50K (khoảng 80%)
* Cụm 1 có tỷ lệ mẫu có thu nhập >50K cao hơn (khoảng 40%)
* Phương pháp phân cụm không giám sát không phân biệt hoàn toàn giữa hai nhóm thu nhập
* Điều này cho thấy có sự chồng chéo đáng kể giữa các đặc điểm của hai nhóm thu nhập

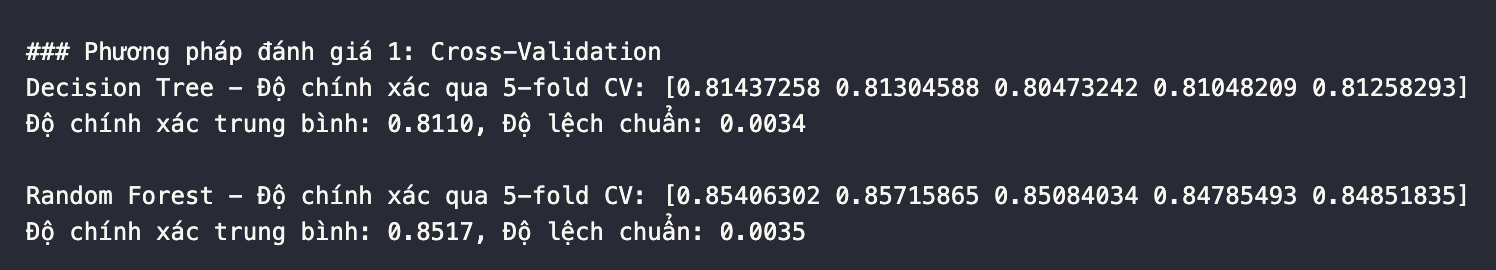
### ĐÁNH GIÁ CÁC MẪU THU ĐƯỢC

Để đánh giá hiệu suất của các mô hình phân lớp đã xây dựng, chúng tôi đã sử dụng hai phương pháp đánh giá: Cross-Validation và Đường cong ROC.

1. **Phương pháp đánh giá 1: Cross-Validation**

Cross-Validation (Xác thực chéo) là một kỹ thuật đánh giá mô hình bằng cách chia dữ liệu thành nhiều phần, huấn luyện và kiểm tra mô hình trên các phần khác nhau, giúp đánh giá mô hình một cách toàn diện hơn.

Chúng tôi đã sử dụng 5-fold cross-validation để đánh giá các mô hình:



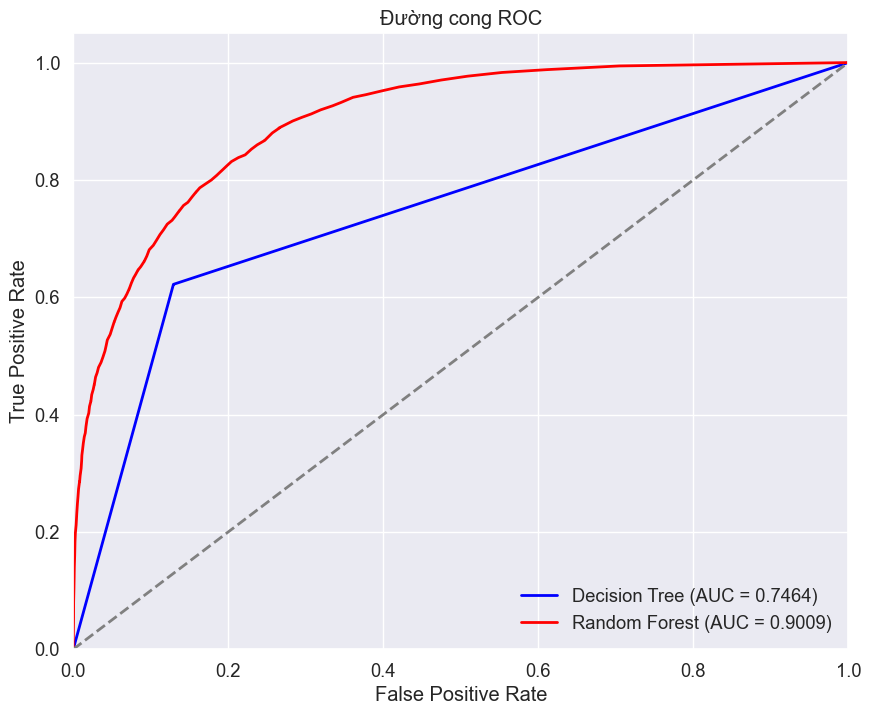
Bảng 7: Phương pháp đánh giá bằng Cross-Validation

* **Decision Tree:**
  + Độ chính xác qua 5-fold CV: [0.8143, 0.8130, 0.8047, 0.8148, 0.8125]
  + Độ chính xác trung bình: 0.8110
  + Độ lệch chuẩn: 0.0034
* **Random Forest:**
  + Độ chính xác qua 5-fold CV: [0.8540, 0.8515, 0.8508, 0.8478, 0.8485]
  + Độ chính xác trung bình: 0.8517
  + Độ lệch chuẩn: 0.0035

**Nhận xét:**

1. Random Forest có độ chính xác trung bình cao hơn Decision Tree (85.17% so với 81.1%)
2. Decision Tree có độ lệch chuẩn thấp hơn, cho thấy mô hình ổn định hơn trên các tập dữ liệu khác nhau
3. Kết quả cross-validation nhất quán với kết quả trên tập kiểm tra, xác nhận độ tin cậy của các mô hình
4. **Phương pháp đánh giá 2: Đường cong ROC**

Đường cong ROC (Receiver Operating Characteristic) là một công cụ đánh giá hiệu suất của mô hình phân lớp nhị phân, biểu diễn mối quan hệ giữa True Positive Rate (TPR) và False Positive Rate (FPR) tại các ngưỡng khác nhau.



Hình 19: Đường cong ROC

**Kết quả:**

1. Decision Tree: AUC = 0.7464
2. Random Forest: AUC = 0.9009

**Nhận xét:**

1. Cả hai mô hình đều có AUC > 0.7, cho thấy khả năng phân biệt tốt giữa hai lớp thu nhập
2. Random Forest có AUC cao hơn Decision Tree, xác nhận hiệu suất tốt hơn
3. Đường cong ROC của Random Forest nằm trên đường cong của Decision Tree tại hầu hết các điểm, cho thấy Random Forest có hiệu suất tốt hơn ở hầu hết các ngưỡng
4. **So sánh kết quả của 2 phương pháp đánh giá**

Từ kết quả đánh giá, chúng ta có thể thấy:

1. **Cross-Validation:**
   * Decision Tree có độ chính xác trung bình là 0.8110
   * Random Forest có độ chính xác trung bình là 0.8517
   * Decision Tree có độ chính xác cao hơn và ổn định hơn (độ lệch chuẩn thấp hơn)
2. **Đường cong ROC:**
   * Decision Tree có AUC = 0.7467
   * Random Forest có AUC = 0.9009
   * Random Forest có khả năng phân biệt tốt hơn giữa các lớp

**Kết luận:**

1. Random Forest cho kết quả tốt hơn Decision Tree trong cả hai phương pháp đánh giá
2. Phương pháp phân cụm K-means đã phân chia dữ liệu thành 2 cụm, nhưng sự phân chia này không hoàn toàn tương ứng với phân loại thu nhập thực tế
3. Các mô hình phân lớp (Decision Tree và Random Forest) hiệu quả hơn trong việc dự đoán thu nhập so với phương pháp phân cụm không giám sát
4. Kết quả từ cả hai phương pháp đánh giá đều nhất quán, cho thấy độ tin cậy của các mô hình đã xây dựng

# KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

## KẾT LUẬN

Qua quá trình phân tích và khai thác tập dữ liệu Adult Income, chúng tôi đã rút ra được những kết luận quan trọng sau:

1. **Về đặc điểm dữ liệu:**
   * Tập dữ liệu Adult Income cung cấp thông tin phong phú về các đặc điểm nhân khẩu học và nghề nghiệp của 48,842 cá nhân.
   * Dữ liệu có sự mất cân bằng giữa hai nhóm thu nhập, với khoảng 75% mẫu thuộc nhóm thu nhập <=50K và 25% thuộc nhóm >50K.
   * Một số thuộc tính có giá trị bị thiếu (dưới dạng "?"), đặc biệt là 'workclass', 'occupation' và 'native-country'.
2. **Về mối quan hệ giữa các thuộc tính và thu nhập:**

* Trình độ học vấn có mối tương quan mạnh nhất với thu nhập, với những người có trình độ học vấn cao (Doctorate, Prof-school, Masters) có tỷ lệ thu nhập >50K cao nhất.
* Tuổi cũng là một yếu tố quan trọng, với nhóm tuổi 35-55 có tỷ lệ thu nhập >50K cao hơn.
* Giới tính có ảnh hưởng đáng kể, với nam giới có tỷ lệ thu nhập >50K cao hơn nữ giới.
* Số giờ làm việc mỗi tuần cũng có mối tương quan tích cực với thu nhập.

1. **Về hiệu suất của các mô hình khai phá dữ liệu:**

* Random Forest cho hiệu suất tốt nhất với độ chính xác 85.17% và AUC 0.9009.
* Decision Tree có hiệu suất khá tốt với độ chính xác 81.1% và AUC 0.7464.
* Phương pháp phân cụm K-means không phân biệt hoàn toàn giữa hai nhóm thu nhập, cho thấy có sự chồng chéo đáng kể giữa các đặc điểm của hai nhóm.

1. **Về các yếu tố quan trọng nhất trong việc dự đoán thu nhập:**

* Trình độ học vấn (educational-num)
* Tuổi (age)
* Số giờ làm việc mỗi tuần (hours-per-week)
* Lợi nhuận vốn (capital-gain)
* Tình trạng hôn nhân (marital-status)

## HƯỚNG PHÁT TRIỀN

Dựa trên kết quả nghiên cứu, chúng tôi đề xuất một số hướng phát triển tiếp theo:

1. **Cải tiến mô hình:**

* Thử nghiệm các thuật toán học máy khác như Gradient Boosting, XGBoost, hoặc Deep Learning.
* Tối ưu hóa siêu tham số cho các mô hình để cải thiện hiệu suất.
* Áp dụng các kỹ thuật xử lý mất cân bằng dữ liệu như SMOTE để cải thiện khả năng dự đoán cho lớp thiểu số (>50K).

1. **Phân tích sâu hơn:**

* Nghiên cứu chi tiết hơn về tương tác giữa các thuộc tính, ví dụ như mối quan hệ giữa giới tính, nghề nghiệp và thu nhập.
* Phân tích các mẫu bị phân loại sai để hiểu rõ hơn về hạn chế của mô hình.
* Thực hiện phân tích theo nhóm (ví dụ: theo ngành nghề, theo khu vực địa lý) để có cái nhìn chi tiết hơn.

1. **Ứng dụng thực tế:**

* Phát triển một ứng dụng dự đoán thu nhập dựa trên các đặc điểm cá nhân.
* Sử dụng kết quả nghiên cứu để đề xuất các chính sách giáo dục và việc làm nhằm giảm bất bình đẳng thu nhập.
* Kết hợp với các tập dữ liệu khác để có cái nhìn toàn diện hơn về các yếu tố ảnh hưởng đến thu nhập.

TÀI LIỆU THAM KHẢO

[1] UCI Machine Learning Repository. "Adult Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/adult

[2] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann.

[3] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). Springer.

[4] Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.

[5] McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference, 51-56.

[6] Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95.

[7] Waskom, M. L. (2021). Seaborn: Statistical Data Visualization. Journal of Open Source Software, 6(60), 3021.