1. **Giới thiệu**
2. **Giới thiệu đề tài**

Thu nhập cá nhân là thuật ngữ để đề cập đến tất cả khoản thu nhập của một cá nhân nào đó kiếm được trong một khoảng thời gian nhất định cụ thể là hàng năm hoặc hàng tháng từ các nguồn như tiền lương, đầu tư và các khoản khác. Qua thu nhập hàng năm của một người có thể đánh giá xem người đó có làm việc chăm chỉ và có hiệu quả hay không, tuy nhiên điều này còn rất nhiều yếu tố khác nhau có thể ảnh hưởng đến nó. trong cuộc tổng điều tra dân số năm 1994 của Barry Becker ông đã thu thập tìm hiểu và ghi lại trong cơ sở dữ liệu có tên là Adult có đến 14 yếu tố khác nhau là nhân khẩu học và các đặc điểm khác để mô tả một người có thể ảnh hưởng đến thu nhập hàng năm của một người. Nhưng thực tế có thể có nhiều hơn không có nhiều các cuộc điều tra về thu nhập cá nhân của một người vì vậy việc dự đoán thu nhập cá nhân của một người nào đó qua các yếu tố trên làm cho vấn đề điều tra về thu nhập cá nhân trở nên dễ hơn mà cụ thể trong cuộc tổng điều tra dân số năm 1994 của Barry Becker ông đã sắp xếp xem một người có thu nhập trên 50000 hay dưới 50000 trong một năm hay không thông qua các thuộc tính mà ông đã thu thập được.

1. **Mô tả dữ liệu**

* Dữ liệu được sử dụng là nhân khẩu học và các đặc điểm khác để mô tả một người, việc trích xuất tập dữ liệu này được thực hiện bởi Ronny Kohavi và Barry Becker (Data Mining and Visualization, Silicon Graphics) từ cơ sở dữ liệu của cục Điều tra dân số năm 1994. Một tập hợp các bản ghi rõ ràng hợp lý được trích xuất bằng cách sử dụng các điều kiện sau: ((AAGE> 16) && (AGI> 100) && (AFNLWGT> 1) && (HRSWK> 0)) với nhiệm vụ dự đoán là xác định xem một người có kiếm được trên 50K/ năm hay không.
* Tập dữ liệu với 32561 hàng và 15 biến bao gồm 14 biến đầu vào và 1 biến mục tiêu:

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Kiểu dữ liệu** | **Giá trị** | **Mô tả** |
| age | integer | Min: 17  Max: 90 | Age – Tuổi |
| workclass | factor | Federal-gov, Local-gov, Never-worked, Private, Self-emp-inc, Self-emp-not-inc, State-gov, Without-pay | Class of work  - Cấp công việc |
| fnlwgt | integer | Min: 12285 Max: 1490400 | Final weight of how much of the population it represents |
| education | Factor | 10th, 11th, 12th, 1st-4th, 5th-6th, 7th-8th, 9th, Assoc-acdm, Assoc-voc, Bachelors, Doctorate, HS-grad, Masters, Preschool, Prof-school, Some-college | Education level  - Trình độ học vấn |
| education\_num | Integer | Min: 1 Max: 16 | Numeric education level  - Số trình độ học vấn |
| marital\_status | Factor | Divorced, Married-AF-spouse, Married-civ-spouse, Married-spouse-absent, Never-married, Separated, Widowed | Marital status  of the person – Tình trạng hôn nhân |
| occupation | Factor | Adm-clerical, Armed-Forces, Craft-repair, Exec-managerial, Farming-fishing, Handlers-cleaners, Machine-op-inspct, Other-service, Priv-house-serv, Prof-specialty, Protective-serv, Sales, Tech-support, Transport-moving | Occupation  of the person – Nghề nghiệp |
| relationship | Factor | Husband, Not-in-family, Other-relative, Own-child, Unmarried, Wife | Type of relationship  - Loại mối quan hệ |
| race | Factor | Amer-Indian-Eskimo, Asian-Pac-Islander, Black, Other, White | Race of the person  - Chủng tộc |
| sex | Factor | Female, Male | Sex of the person – Giới tính |
| capital\_gain | Integer | Min: 0 Max: 99999 | Capital gains obtained -  Lợi nhuận vốn thu được |
| capital\_loss | Integer | Min: 0 Max: 4356 | Capital loss – Lỗ vốn |
| hours\_per\_week | Integer | Min: 1 Max: 99 | Average number of hour working per week - Số giờ làm việc trung bình mỗi tuần |
| native\_country | Factor | Cambodia, Canada, China, Columbia, Cuba, Dominican-Republic, Ecuador, El-Salvador, England, France, Germany, Greece, Guatemala, Haiti, Holand-Netherlands, Honduras, Hong, Hungary, India, Iran, Ireland, Italy, Jamaica, Japan, Laos, Mexico, Nicaragua, Outlying-US(Guam-USVI-etc), Peru, Philippines, Poland, Portugal, Puerto-Rico, Scotland, South, Taiwan, Thailand, Trinadad&Tobago, United-States, Vietnam, Yugoslavia | Country of origin  - Quốc gia |
| imcome | Factor | <=50K, >50K | Income level – Mức thu nhập |

* References: <https://archive.ics.uci.edu/ml/datasets/adult>

1. **Tiêu chí đánh giá**
   1. **Accuracy**: được dùng để chỉ mức độ khác biệt giữa các giá trị đo so với giá trị thực hoặc giá trị chuẩn dùng để tham chiếu.

Acc = ((TP + TN)/số lượng mẫu) = ((TP + TN)/(TP + TN + FP + FN))

* 1. **F1-score**: là trung bình điều hòa (harmonic mean) của precision và recall, có tầm quan trọng tương tự như với FNs và FPs.

**F1 Score = ((2 x Precision x Recall)/(Precision + Recall))**

* 1. **Precision:** là tỉ lệ số điểm True Positive trong số những điểm được phân loại là Positive (TP + FP).

**Precision = TP/((TP+FP))**

* 1. **Recall**: là tỉ lệ số điểm True Positive trong số những điểm thực sự là Positive (TP+FN).

**Recall = TP/((TP+FN))**

* 1. **Confusion matrix**
* **Confusion Matrix** (ma trận nhầm lẫn hay ma trận lỗi) là một bố cục bảng cụ thể cho phép hình dung hiệu suất của một thuật toán.
* **Ma trận nhầm lẫn** là một trong những kỹ thuật đo lường hiệu suất phổ biến nhất và được sử dụng rộng rãi cho các mô hình phân loại. Ma trận nhầm lẫn chỉ ra được cụ thể mỗi loại được phân loại như thế nào, lớp nào được phân loại đúng nhiều nhất, và dữ liệu thuộc lớp nào thường bị phân loại nhầm vào lớp khác. Bên cạnh đó, ma trận nhầm lẫn còn thể hiện có bao nhiêu điểm dữ liệu thực sự thuộc vào một class, và được dự đoán là rơi vào một class.

**Trong đó:**

* True Positive: TP – Số lượng điểm của lớp positive được phân loại đúng là positive.
* False Positive: FP – Số lượng điểm của lớp negative bị phân loại nhầm thành positive.
* True Negative: TN – Số lượng điểm của lớp negative được phân loại đúng là negative.
* False Negative: FN – Số lượng điểm của lớp positive bị phân loại nhầm thành negative.

1. **Mô hình máy học**
   1. **Naïve Bayes** **Classification (NBC)**

* Naive Bayes Classification (NBC) là một thuật toán dựa trên định lý Bayes về lý thuyết xác suất để đưa ra các phán đoán cũng như phân loại dữ liệu dựa trên các dữ liệu được quan sát và thống kê.
* Naive Bayes Classification là một trong những thuật toán được ứng dụng rất nhiều trong các lĩnh vực Machine learning dùng để đưa các dự đoán chính xác nhất dự trên một tập dữ liệu đã được thu thập, vì nó khá dễ hiểu và độ chính xác cao. Nó thuộc vào nhóm Supervised Machine Learning Algorithms (thuật toán học có hướng dẫn), tức là máy học từ các ví dụ từ các mẫu dữ liệu đã có.
  1. **Logistic Regression**

**-** Logistic Regression là 1 thuật toán phân loại được dùng để gán các đối tượng cho 1 tập hợp giá trị rời rạc (như 0, 1, 2, ...). Logistic Regression lại được sử dụng nhiều trong các bài toán Classification.

**-** Đầu ra dự đoán của logistic regression thường được viết chung dưới dạng:

**f(x) = θ(wTx)**

**Trong đó:** θ được gọi là logistic function

* 1. **Gradient Boosting Model (GBM)**

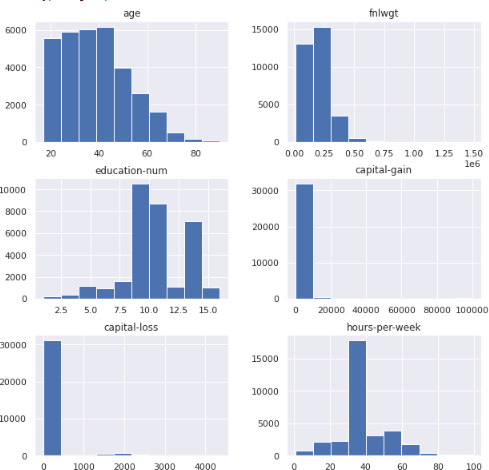
Thuật toán là xây dựng một decision tree và tìm cách để cải thiện độ chính xác tại mỗi nhánh của decision tree bằng các thuật toán gradient boosting. Quá trình này được lặp lại nhiều lần vào sau mỗi một vòng lặp thì metric của model (có thể là loss function hoặc accuracy) sẽ được cải thiện

* 1. **XGBoost**

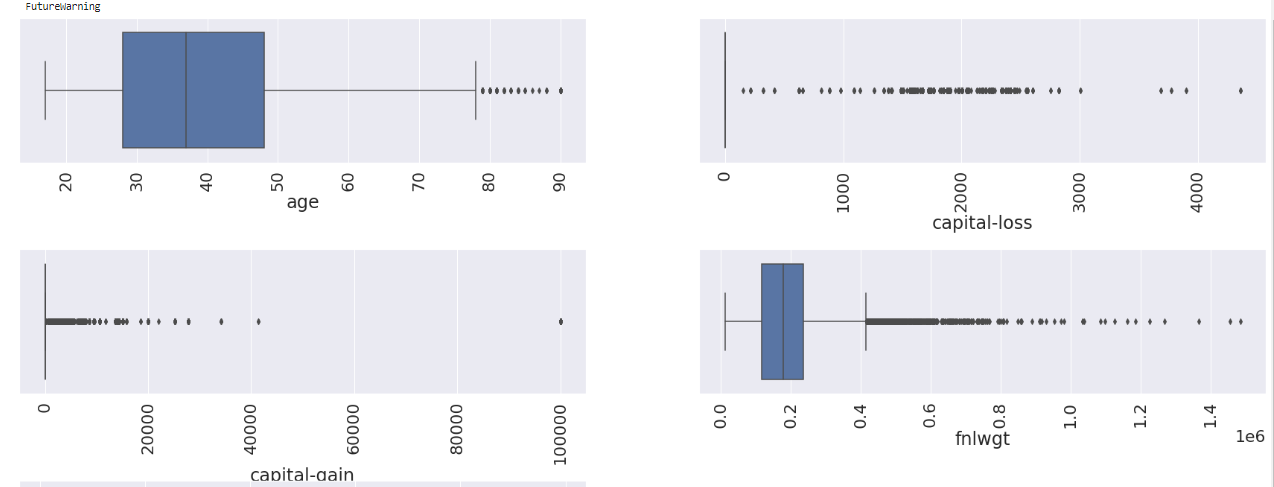
XGBoost là một trong những thuật toán học máy phổ biến và được sử dụng rộng rãi nhất hiện nay vì đơn giản là nó rất mạnh mẽ. Nó tương tự như Gradient Boost nhưng có một vài tính năng bổ sung làm cho nó mạnh hơn nhiều bao gồm:

* Sự thu nhỏ theo tỷ lệ của các nút lá (cắt tỉa) – được sử dụng để cải thiện tính tổng quát của mô hình
* Newton Boosting – cung cấp một tuyến đường trực tiếp đến cực tiểu thay vì giảm độ dốc, làm cho nó nhanh hơn nhiều
* Một tham số ngẫu nhiên bổ sung – giảm mối tương quan giữa các cây, cuối cùng cải thiện sức mạnh của nhóm
  1. **Histogram Based Boosting Model (HGB)**
* Histogram Based Boosting Model là một kỹ thuật để đào tạo cây quyết định nhanh hơn được sử dụng trong nhóm gradient boosting.
* Cách sử dụng triển khai thử nghiệm tăng cường độ dốc dựa trên biểu đồ trong thư viện scikit-learning.
* Cách sử dụng các nhóm tăng cường độ dốc dựa trên biểu đồ với các thư viện bên thứ ba XGBoost và LightGBM.
  1. **LightGBM**
* LightGBM là một loại thuật toán thúc đẩy khác được chứng minh là nhanh hơn và đôi khi chính xác hơn XGBoost.
* Điều làm cho LightGBM trở nên khác biệt là nó sử dụng một kỹ thuật độc đáo được gọi là Lấy mẫu một phía dựa trên Gradient (GOSS – Gradient-based One-Side Sampling) để lọc ra các cá thể dữ liệu nhằm tìm ra giá trị phân tách.

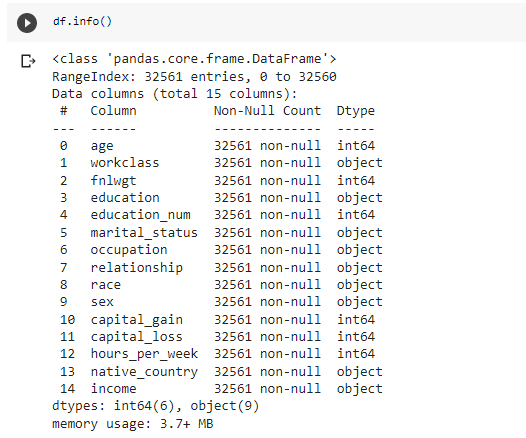
1. **Phân tích**
2. **EDA**
   1. Trực quan các biến định lượng
      1. Histogram



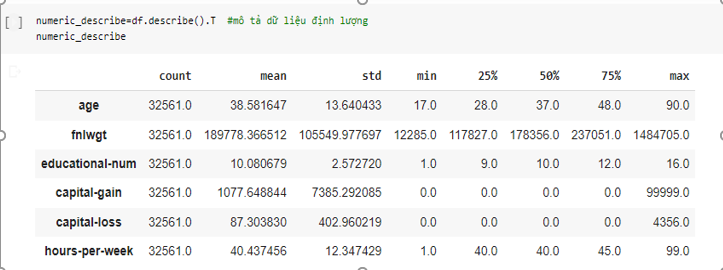
* + 1. Boxplot



* 1. **Mô tả dữ liệu**



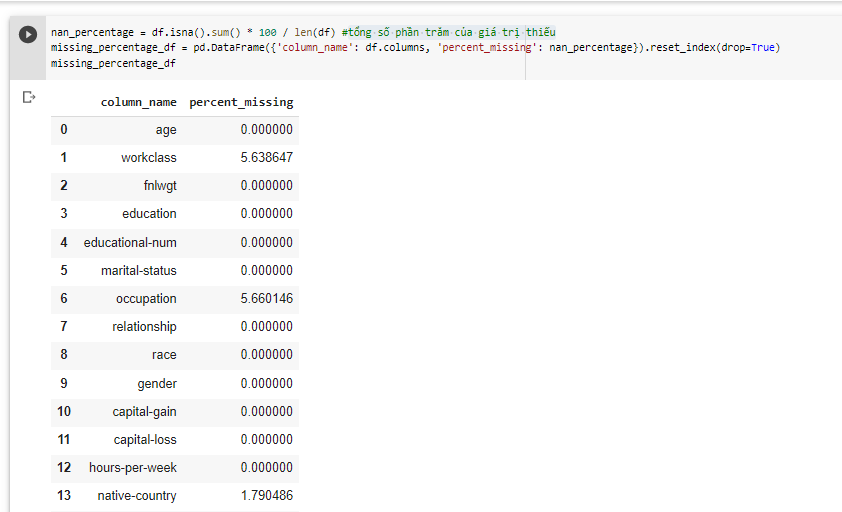
* 1. **Mô tả dữ liệu định lượng**



* 1. **Mô tả dữ liệu định tính**



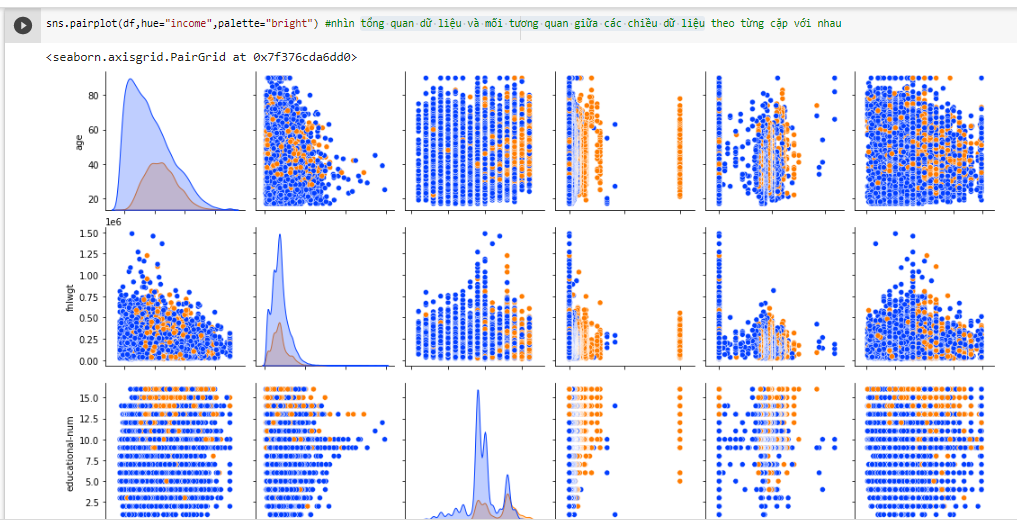
* 1. **Giá trị thiếu (%)**



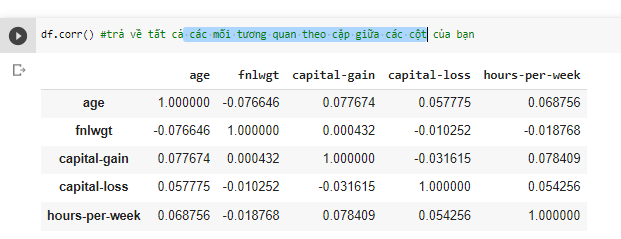
* 1. **Ma trận của giá trị thiếu**



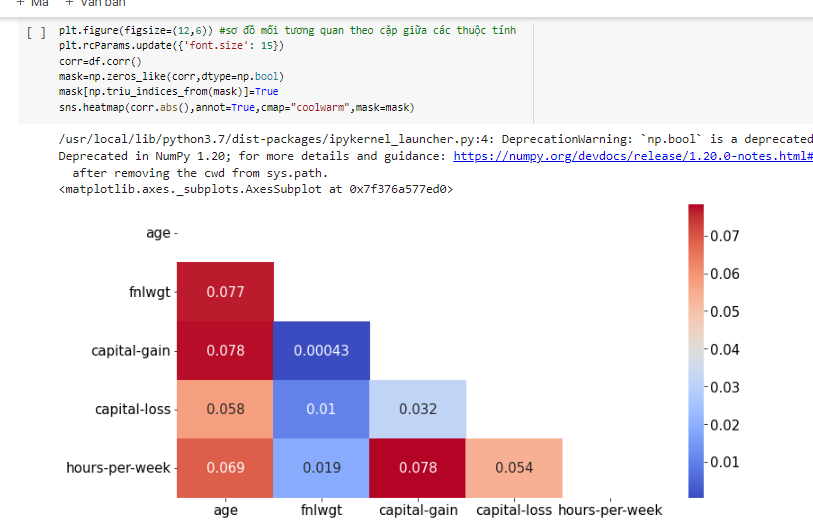
* 1. **Tổng quan dữ liệu**



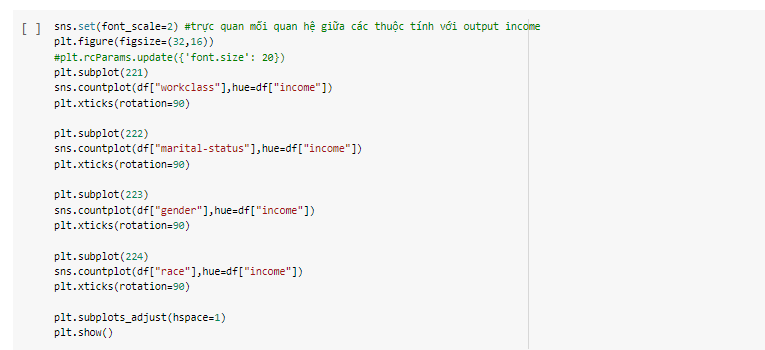
* 1. **Mối tương quan theo cặp giữa các thuộc tính định lượng**

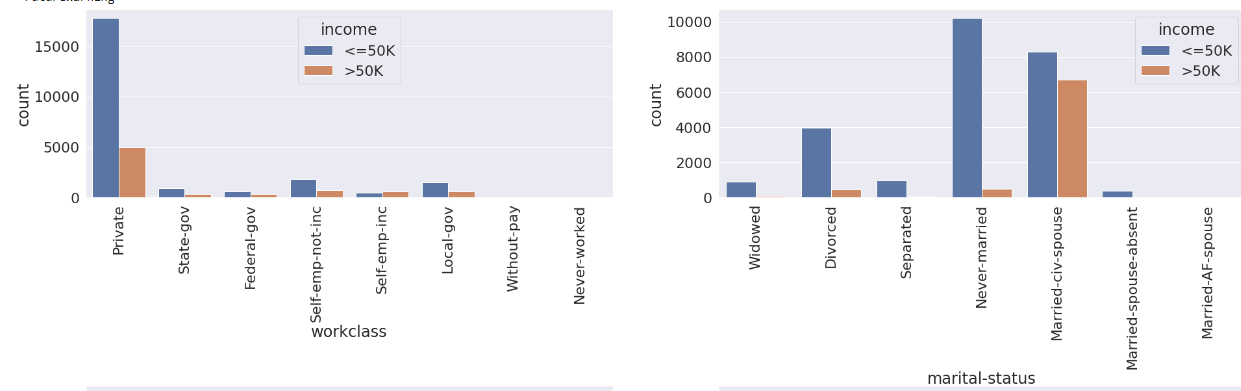


**Sơ đồ:**



* 1. **Mối quan hệ giữa các biến định tính với biến mục tiêu**







1. **Kết quả phân tích**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model\Độ đo** | **Acc** | **F1-Score** | **Precision** | **Recall** |
| **Gaussian Naïve Bayes** | 0.8083 | <=50K: 0.86  > 50K: 0.67 | <=50K: 0.93  >50K: 0.57 | <=50K: 0.81  >50K: 0.80 |
| **Logistic Regression** | 0.8524 | 0.6563 | 0.5964 | 0.7296 |
| **GBM** | 0.8628 | 0.6808 | 0.7659 | 0.6128 |
| **XGBoost** | 0.8615 | 0.6734 | 0.7707 | 0.5979 |
| **HGB** | 0.8659 | 0.7010 | 0.7496 | 0.6583 |
| **LightGBM** | 0.8683 | 0.7063 | 0.7551 | 0.6634 |