**TRƯỜNG ĐẠI HỌC THỦY LỢI**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÀI TẬP LỚN**

**KHAI PHÁ DỮ LIỆU**

**ĐỀ TÀI:**

**KHAI PHÁ DỮ LIỆU BẠO LỰC HỌC ĐƯỜNG**

|  |  |
| --- | --- |
| **Giảng viên hướng dẫn:**  **Sinh viên:**  **Lớp:** | TS. Trần Mạnh Tuấn  Nguyễn Thị Phương Anh - 2251161942  Nguyễn Minh Hiếu - 2251162001  64TTNT2 |
|  |  |

Hà Nội - 2024

# LỜI NÓI ĐẦU

Bạo lực học đường từ lâu đã trở thành một vấn đề nhức nhối, gây ảnh hưởng nghiêm trọng đến sự phát triển của thế hệ tương lai trên toàn thế giới. Qua những số liệu thực tế, chúng em nhận thấy sự cần thiết trong việc ứng dụng thực tiễn và sự hữu ích đề tài này sẽ mang lại cho quá trình học tập của mình. Như vậy, chúng em đã quyết định triển khai Đề tài *“Khai phá dữ liệu về bạo lực học đường”* để tìm ra các yếu tố chính gây ảnh hưởng đến việc học sinh trở thành nạn nhân của bạo lực học đường.

Mục tiêu chính của chúng em là áp dụng các kỹ thuật khai phá dữ liệu để phân tích mối quan hệ giữa các yếu tố như giới tính, độ tuổi, và loại bạo lực, từ đó xác định những yếu tố quan trọng nhất dẫn tới bạo lực học đường.

Chúng em tin rằng đề tài sẽ giúp các trường học và cơ quan giáo dục có cái nhìn rõ ràng hơn về tình trạng bạo lực học đường hiện nay, qua đó thực hiện những biện pháp can thiệp và phòng ngừa bạo lực học đường một cách hiệu quả.

Xin chân thành cảm ơn thầy Trần Mạnh Tuấn đã tận tình hướng dẫn và hỗ trợ chúng em hoàn thành Đề tài “*Khai phá dữ liệu về bạo lực học đường*”.

Chúng em xin chân thành cảm ơn!

MỤC LỤC

[**LỜI NÓI ĐẦU 2**](#_gjdgxs)

[**PHẦN I. MÔ TẢ BÀI TOÁN 5**](#_uza4uwhaaoru)

[1. Đặt vấn đề 5](#_uait3eatgnfc)

[2. Quy trình thực hiện 6](#_8a8c0au9xuwf)

[3. Phân tích dữ liệu thô 6](#_kf1zdd4qj4v6)

[3.1. Tổng quan 6](#_ed1ckx7ihxi1)

[3.2. Ý nghĩa các thuộc tính 7](#_6qt8i8na4swo)

[**PHẦN II. TIỀN XỬ LÝ DỮ LIỆU 11**](#_rijdja1rq6ny)

[1. Tiền xử lý dữ liệu 11](#_q24xi7tavb9l)

[1.1. Làm sạch dữ liệu 11](#_i4ubhu95flnx)

[1.1.1. Loại bỏ thuộc tính trùng lặp 11](#_p75y60604od4)

[1.1.2. Xử lý nhiễu 15](#_nk76kr3a7zh4)

[1.1.3. Xử lý dữ liệu thiếu 34](#_kjfdt1iuunxh)

[1.1.4. Xử lý dữ liệu không nhất quán 35](#_vlon4ppmls9r)

[1.2. Tích hợp dữ liệu 35](#_6nhon32fxxqt)

[2. Phân tích dữ liệu sau khi tiền xử lý lần đầu 36](#_wlkq90hdp8eu)

[2.1. Loại bỏ thuộc tính dư thừa 37](#_4olxvb1a4jes)

[2.2. Biến đổi dữ liệu 40](#_z9c1x5yc03)

[2.3. Phân tích dữ liệu sau khi tiền xử lý 45](#_thy8oetsrphs)

[2.4. Đánh giá dữ liệu 46](#_5ubm0z2d4q8f)

[**PHẦN III. PHÂN LỚP DỮ LIỆU 49**](#_6o2dcod3fs69)

[1. Phân lớp dữ liệu 49](#_ywhavmaf9v1b)

[2. Phương pháp chia theo tỉ lệ 49](#_1dxeaty2d78m)

[3. Thuật toán ID3 50](#_6a3icizarojc)

[3.1. Lý thuyết 50](#_t1otajrgsc60)

[3.2. Quy trình thực hiện 51](#_g5f8b4l55vvg)

[3.3. Giải thích các thống kê và độ đo theo lớp 52](#_ciclniqatixl)

[3.3.1. Các thống kê 52](#_eevi3klnw3a)

[3.3.2. Các độ đo theo lớp 52](#_4429ohi3qhch)

[3.4. Kết quả, nhận xét 53](#_pegzn9peptaf)

[3.4.1. Use training test 53](#_mnp74jhcqrh5)

[3.4.1.1. Kết quả thu được 53](#_ke5s1lvv9vps)

[3.4.1.2. Nhận xét 54](#_c4d9932c2z16)

[3.4.2. Cross-validation 55](#_3dv6fta1zgxl)

[3.4.2.1. Kết quả thu được 55](#_nfp4452q3hyc)

[3.4.2.2. Nhận xét 55](#_kz0kc0gn4kb0)

[3.4.3. Supplied test set 56](#_35j72bkkkvrp)

[3.4.3.1. Kết quả thu được 56](#_n5i86u6vaw2i)

[3.4.3.2. Nhận xét 56](#_9zuwb8mboiit)

[3.4.4. Cây quyết định 57](#_b31q50nviuob)

[3.5. Nhận xét 58](#_glx5c136hqs)

[3.6. Đánh giá 58](#_tr1tuaj8jbrk)

[**PHẦN IV. TRIỂN KHAI THUẬT TOÁN 60**](#_2e6k9me5hmzo)

[1. Triển khai thuật toán (Python) 60](#_fr7of8nz6u6c)

[2. Sử dụng mô hình để dự đoán kết quả 62](#_iordq78vujc1)

[**KẾT LUẬN 64**](#_vy8kkcm27spc)

[**TÀI LIỆU THAM KHẢO 65**](#_9ze2ecdnooi3)

# PHẦN I. MÔ TẢ BÀI TOÁN

## Đặt vấn đề

Theo báo cáo của Tổ chức Giáo dục, Khoa học và Văn hóa Liên hợp quốc (UNESCO) vào năm 2021, khoảng ⅓ số lượng học sinh từ 13 đến 15 tuổi trên toàn thế giới đã từng trải qua ít nhất một hình thức bạo lực học đường. Con số này tương đương với hơn 150 triệu học sinh trên toàn cầu cho thấy quy mô thật sự đáng báo động của vấn đề này.

Tại Việt Nam, theo một khảo sát được thực hiện vào năm 2022, hơn 40% học sinh đã từng chứng kiến hoặc trải nghiệm bạo lực học đường, với các hình thức như đánh đập, châm chọc, và bắt nạt qua mạng xã hội. Các con số này cho thấy rằng không chỉ học sinh mà cả giáo viên và phụ huynh đều cần phải quan tâm nhiều hơn đến tình trạng này. Nhiều nghiên cứu đã chỉ ra rằng những học sinh từng trải qua bạo lực học đường có khả năng gặp phải các vấn đề về tâm lý như lo âu, trầm cảm, và khó khăn trong việc thiết lập các mối quan hệ xã hội.

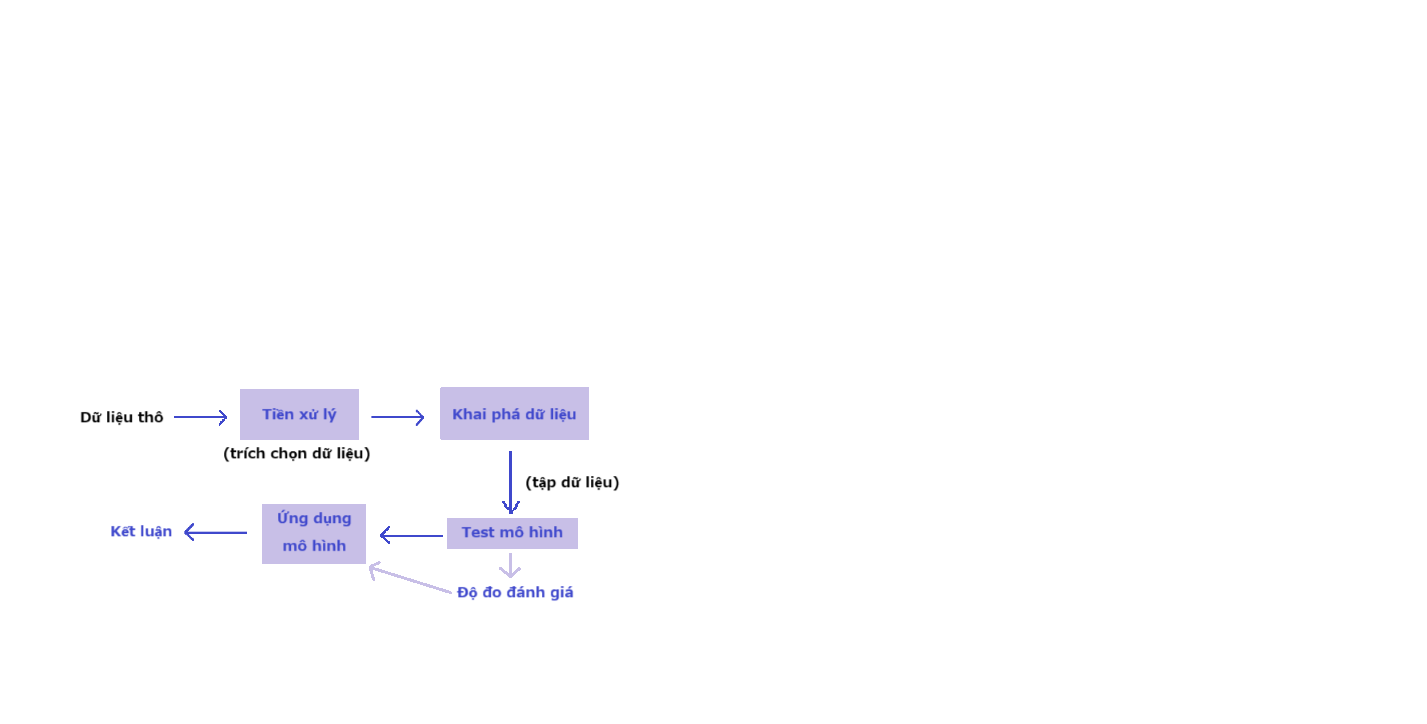
Bên cạnh đó, bạo lực học đường còn ảnh hưởng đến thành tích học tập. Một nghiên cứu tại Hoa Kỳ cho thấy rằng những học sinh bị bắt nạt có nguy cơ bỏ học cao gấp đôi so với các bạn cùng trang lứa. Điều này không chỉ làm giảm chất lượng giáo dục mà còn dẫn đến việc tăng cường sự chênh lệch trong cơ hội phát triển cá nhân và nghề nghiệp sau này.

Trước tình trạng báo động này, chúng em triển khai đề tài không chỉ để tìm hiểu nguyên nhân mà còn làm rõ các yếu tố ảnh hưởng cũng như mức độ nghiêm trọng của bạo lực học đường.

**Tổng quan bài toán**

Dataset bao gồm các thuộc tính mô tả tình trạng bạo lực học đường. Mục tiêu của dự án là áp dụng các kỹ thuật khai phá dữ liệu để phân tích mối quan hệ giữa các yếu tố như giới tính, độ tuổi, và loại bạo lực, từ đó xác định những yếu tố quan trọng nhất dẫn tới bạo lực học đường.

## Quy trình thực hiện



* Quy trình thực hiện khai phá bao gồm 6 bước:

Bước 1: Thu thập tập dữ liệu đầu vào

Bước 2: Tiền xử lý, làm sạch tập dữ liệu

Bước 3: Chọn tác vụ khai phá dữ liệu: Phương pháp phân lớp

Bước 4: Khai phá dữ liệu: tìm kiếm tri thức

Bước 5: Đánh giá mẫu tìm được

Bước 6: Biểu diễn tri thức

* Nguồn dữ liệu thô: [Introducing a New Dataset of Datasets: Where, When, and How Much Data Exists on School Violence | Center For Global Development (cgdev.org)](https://www.cgdev.org/blog/introducing-new-dataset-datasets-where-when-and-how-much-data-exists-school-violence)

## Phân tích dữ liệu thô

### Tổng quan

Tập dữ liệu này được xây dựng nhằm hệ thống hóa các khảo sát quốc tế đại diện cho các quốc gia có thu nhập thấp và trung bình về bạo lực học đường từ năm 2013 đến năm 2023. Bao gồm 14 khảo sát trên phạm vi quốc tế, trong đó có 10 khảo sát có câu hỏi đo lường mức độ bạo lực mà trẻ em báo cáo. Quá trình thu thập dữ liệu bao gồm phân tích các bảng câu hỏi, ghi nhận thông tin về phạm vi quốc gia, năm thực hiện, đối tượng khảo sát, loại bạo lực (vật lý, tâm lý và tình dục) và tính khả dụng của dữ liệu.



* Dữ liệu bao gồm 1066 bản ghi và 44 thuộc tính.

### Ý nghĩa các thuộc tính

|  |  |  |
| --- | --- | --- |
| **STT** | **Thuộc tính** | **Ý nghĩa** |
| 1 | country | Tên quốc gia hoặc khu vực |
| 2 | countrycode | Mã quốc gia gồm ba chữ cái (ISO 3166-1 alpha-3) |
| 3 | region | Tên vùng |
| 4 | regioncode | Mã vùng gồm 3 chữ cái (theo World Bank) |
| 5 | incomegroup | Nhóm thu nhập |
| 6 | incomecode | Mã nhóm thu nhập gồm 3 chữ cái (theo World Bank) |
| 7 | last\_survey\_year | Tên và năm của (các) cuộc khảo sát được cập nhật trên phạm vi quốc gia |
| 8 | pending\_survey | Tên của (các) khảo sát đang chờ xử lý không có sẵn dữ liệu công khai |
| 9 | survey\_number | Tổng số cuộc khảo sát quốc tế với các câu hỏi về bạo lực trong khoảng từ năm 2013 đến năm 2023 |
| 10 | round\_number | Tổng số vòng (round) khảo sát quốc tế có câu hỏi về bạo lực học đường từ năm 2013 đến năm 2023. |
| Chú thích: Vòng đề cập đến những năm mà một cuộc khảo sát được thực hiện. Ví dụ, nếu khảo sát PISA được thực hiện trong 3 năm khác nhau từ năm 2013 đến năm 2023, tổng số vòng PISA sẽ là 3. Chỉ số này cung cấp tổng số vòng xem xét TẤT CẢ các cuộc điều tra quốc tế (chỉ số không phải là khảo sát cụ thể) |
| 11 | dataset\_number | Tổng số bộ dữ liệu (dataset) của các cuộc điều tra quốc tế có câu hỏi về bạo lực học đường từ năm 2013 đến năm 2023 |
| Chú thích: Bộ dữ liệu đề cập đến số lượng khảo sát đã được quản lý theo thời gian và bởi người trả lời khảo sát. Ví dụ: nếu một cuộc khảo sát PISA được thực hiện trong 3 năm khác nhau và có hai người trả lời khảo sát khác nhau (tức là trẻ em và giáo viên), tổng số bộ dữ liệuPISA sẽ là 6. Chỉ số này cung cấp tổng số bộ dữ liệu xem xét TẤT CẢ các cuộc điều tra quốc tế (chỉ số này không phải là khảo sát cụ thể). |
| 12 | surveyname | Tên khảo sát |
| 13 | year | Năm thu thập dữ liệu của vòng khảo sát |
| 14 | status | Giá trị thể hiện tình trạng hoàn thành của vòng khảo sát |
| 15 | school\_hh\_based\_survey | Giá trị thể hiện vòng khảo sát có phải là một cuộc khảo sát tại trường học hay không. |
| 16 | representativeness | Giá trị thể hiện cuộc khảo sát có đại diện cấp quốc gia hay không |
| 17 | male | Giá trị thể hiện người thực hiện cuộc khảo sát là nam |
| 18 | female | Giá trị thể hiện người thực hiện cuộc khảo sát là nữ |
| 19 | respondant\_children | Giá trị thể hiện vòng khảo sát này dành cho trẻ em, thanh thiếu niên hoặc thanh niên, với câu hỏi về trải nghiệm bạo lực học đường ở trường học họ học tập |
| 20 | respondant\_teacher | Giá trị thể hiện vòng khảo sát dành cho giáo viên, với câu hỏi về mức độ phổ biến của bạo lực học đường ở trường học họ giảng dạy |
| 21 | respondant\_principal | Giá trị thể hiện vòng khảo sát dành cho hiệu trưởng, với câu hỏi về mức độ phổ biến của bạo lực học đường ở trường học họ quản lý |
| 22 | age\_6\_9 | Giá trị thể hiện cuộc khảo sát được thực hiện trên đối tượng từ 6 đến 9 tuổi |
| 23 | age\_10\_12 | Giá trị thể hiện cuộc khảo sát được thực hiện trên đối tượng từ 10 đến 12 tuổi |
| 24 | age\_13\_17 | Giá trị thể hiện cuộc khảo sát được thực hiện trên đối tượng từ 13 đến 17 tuổi |
| 25 | age\_18\_22 | Giá trị thể hiện cuộc khảo sát được thực hiện trên đối tượng từ 18 đến 22 tuổi |
| 26 | mentions\_teacher | Giá trị thể hiện vòng khảo sát đề cập đến bạo lực do giáo viên gây ra |
| 27 | mentions\_peer | Giá trị thể hiện bạo lực do bạn bè đồng trang lứa gây ra |
| 28 | mentions\_physical | Giá trị thể hiện bạo lực học đường về mặt thể chất |
| 29 | mentions\_emotional | Giá trị thể hiện bạo lực học đường về mặt tinh thần |
| 30 | mentions\_sexual | Giá trị thể hiện bạo lực học đường về mặt tình dục |
| 31 | mentions\_teacher\_physical | Giá trị thể hiện bạo lực về mặt thể chất do giáo viên gây ra |
| 32 | mentions\_teacher\_emotional | Giá trị thể hiện bạo lực về mặt tinh thần do giáo viên gây ra |
| 33 | mentions\_teacher\_sexual | Giá trị thể hiện bạo lực về mặt tình dục do giáo viên gây ra |
| 34 | mentions\_peer\_physical | Giá trị thể hiện bạo lực về mặt thể chất do bạn bè đồng trang lứa gây ra |
| 35 | mentions\_peer\_emotional | Giá trị thể hiện bạo lực về mặt tinh thần do bạn bè đồng trang lứa gây ra |
| 36 | mentions\_peer\_sexual | Giá trị thể hiện bạo lực về mặt tình dục do bạn bè đồng trang lứa gây ra |
| 37 | prev\_sexual\_victim | Giá trị ước tính độ phố biến của việc bạo lực tình dục có liên quan tới trường học được báo cáo bởi các nạn nhân |
| 38 | prev\_sexual\_staff | Giá trị ước tính độ phố biến của việc bạo lực tình dục có liên quan tới trường học được báo cáo bởi cán bộ nhân viên nhà trường |
| 39 | prev\_physical\_victim | Giá trị ước tính độ phố biến của việc bạo lực thể chất có liên quan tới trường học được báo cáo bởi các nạn nhân |
| 40 | prev\_physical\_staff | Giá trị ước tính độ phố biến của việc bạo lực thể chất có liên quan tới trường học được báo cáo bởi cán bộ nhân viên nhà trường |
| 41 | prev\_fight\_victim | Giá trị ước tính độ phố biến của tỷ lệ đánh nhau, gây gổ có liên quan tới trường học được báo cáo bởi các nạn nhân |
| 42 | prev\_fight\_staff | Giá trị ước tính độ phố biến của tỷ lệ đánh nhau, gây gổ có liên quan tới trường học được báo cáo bởi cán bộ nhân viên nhà trường |
| 43 | prev\_emo\_victim | Giá trị ước tính độ phố biến của việc bạo lực tinh thần có liên quan tới trường học được báo cáo bởi các nạn nhân |
| 44 | prev\_emo\_staff | Giá trị ước tính độ phố biến của việc bạo lực tinh thần có liên quan tới trường học được báo cáo bởi cán bộ nhân viên nhà trường |

# 

# PHẦN II. TIỀN XỬ LÝ DỮ LIỆU

## Tiền xử lý dữ liệu

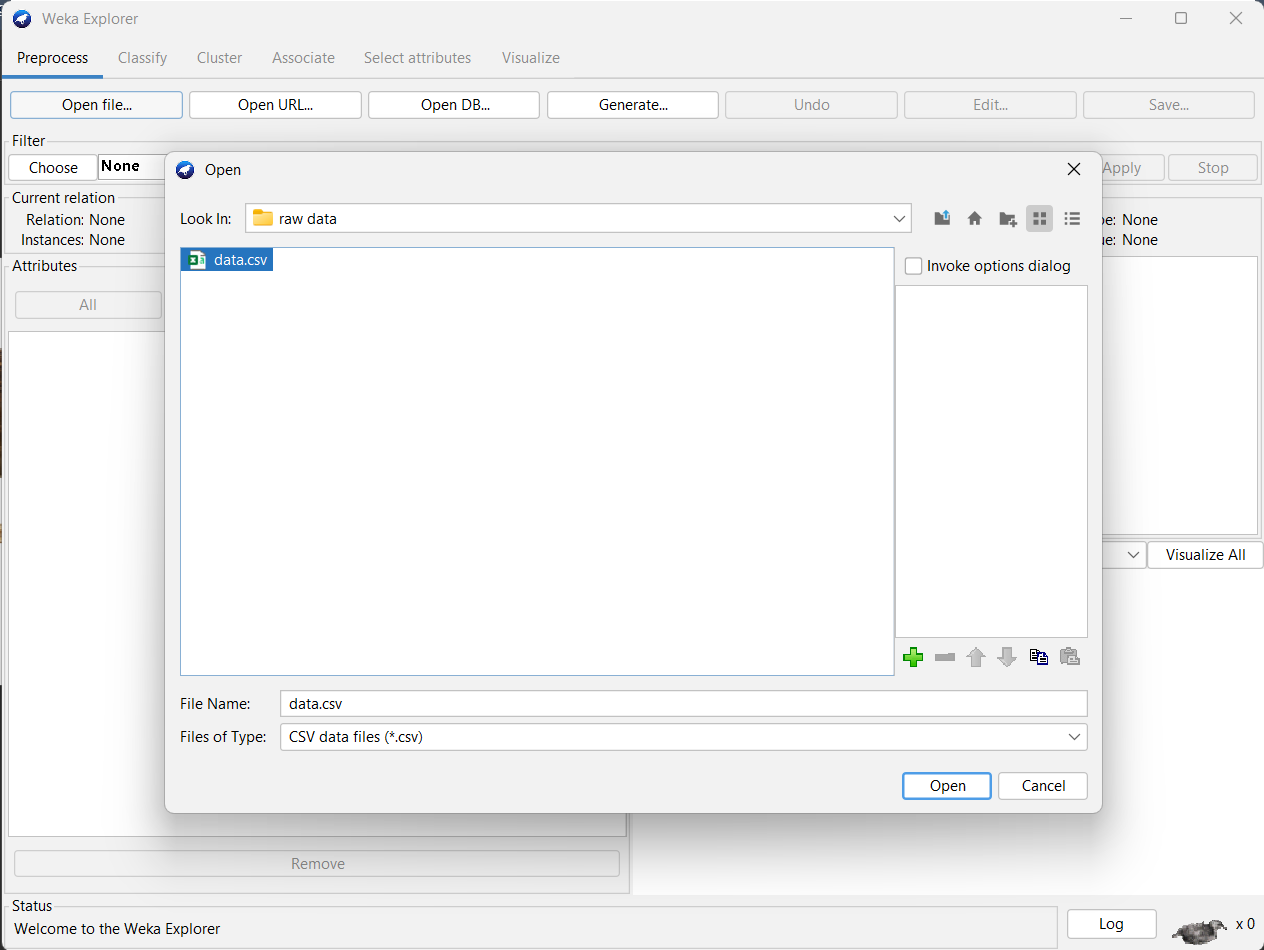
Là quá trình xử lý dữ liệu thô/gốc nhằm cải thiện chất lượng dữ liệu và chất lượng của kết quả khai phá.

### Làm sạch dữ liệu

Là quá trình xử lý dữ liệu bị thiếu, nhận diện phần tử biên và giảm thiểu nhiễu, xử lý dữ liệu không nhất quán

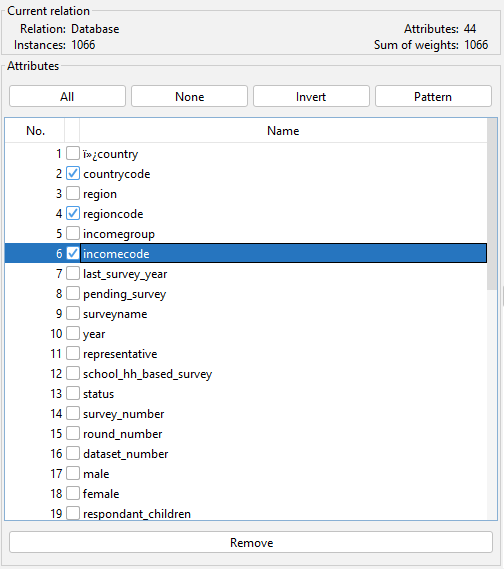
#### Loại bỏ thuộc tính trùng lặp

* Đọc dữ liệu vào Weka:



* Xét những thuộc tính dư thừa có thể suy diễn từ thuộc tính khác:
* *country* (tên quốc gia hoặc khu vực) và *countrycode* (mã quốc gia hoặc khu vực) có cùng ý nghĩa, dựa vào *countrycode* ta có thể xác định *country* và ngược lại⇒ loại bỏ *countrycode.*
* *region* (tên vùng)và *regioncode* (mã vùng) có cùng ý nghĩa, dựa vào *regioncode* ta có thể xác định *region* và ngược lại⇒ loại bỏ *regioncode.*
* *incomegroup* (nhóm thu nhập) và *incomecode* (mã nhóm thu nhập) có cùng ý nghĩa, dựa vào *incomecode* ta có thể xác định *incomegroup* và ngược lại⇒ loại bỏ *incomecode.*

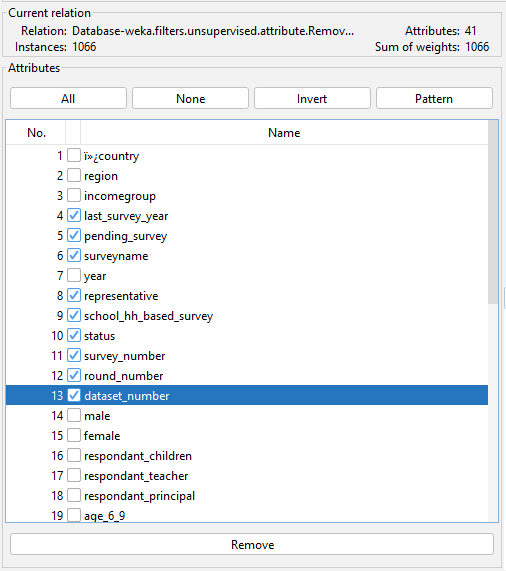
**⇒ Thực hiện**: Bấm chọn các ô *countrycode, regioncode, incomecode*, sau đó nhấn **Remove** để loại bỏ các thuộc tính này:



* Xét những thuộc tính dư thừa, không cần thiết:
* *last\_survey\_year*
* *pending\_survey*
* *survey\_number*
* *round\_number*
* *dataset\_number*
* *surveyname*
* *status*
* *school\_hh\_based\_survey*
* *representativeness*

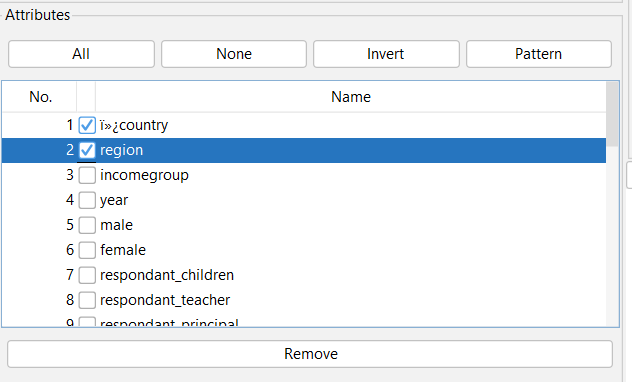
Những thuộc tính trên chỉ dùng để thể hiện độ tin cậy và hiện hành của dữ liệu và không ảnh hưởng tới việc phân tích dữ liệu. Ta lược bỏ những thuộc tính này nhằm thu gọn dữ liệu.

**⇒ Thực hiện**: lựa chọn các ô *last\_survey\_year*, *pending\_survey*, *survey\_number, round\_number, dataset\_number, surveyname, status, school\_hh\_based\_survey, representativeness* và bấm **Remove** để loại bỏ các thuộc tính trên.

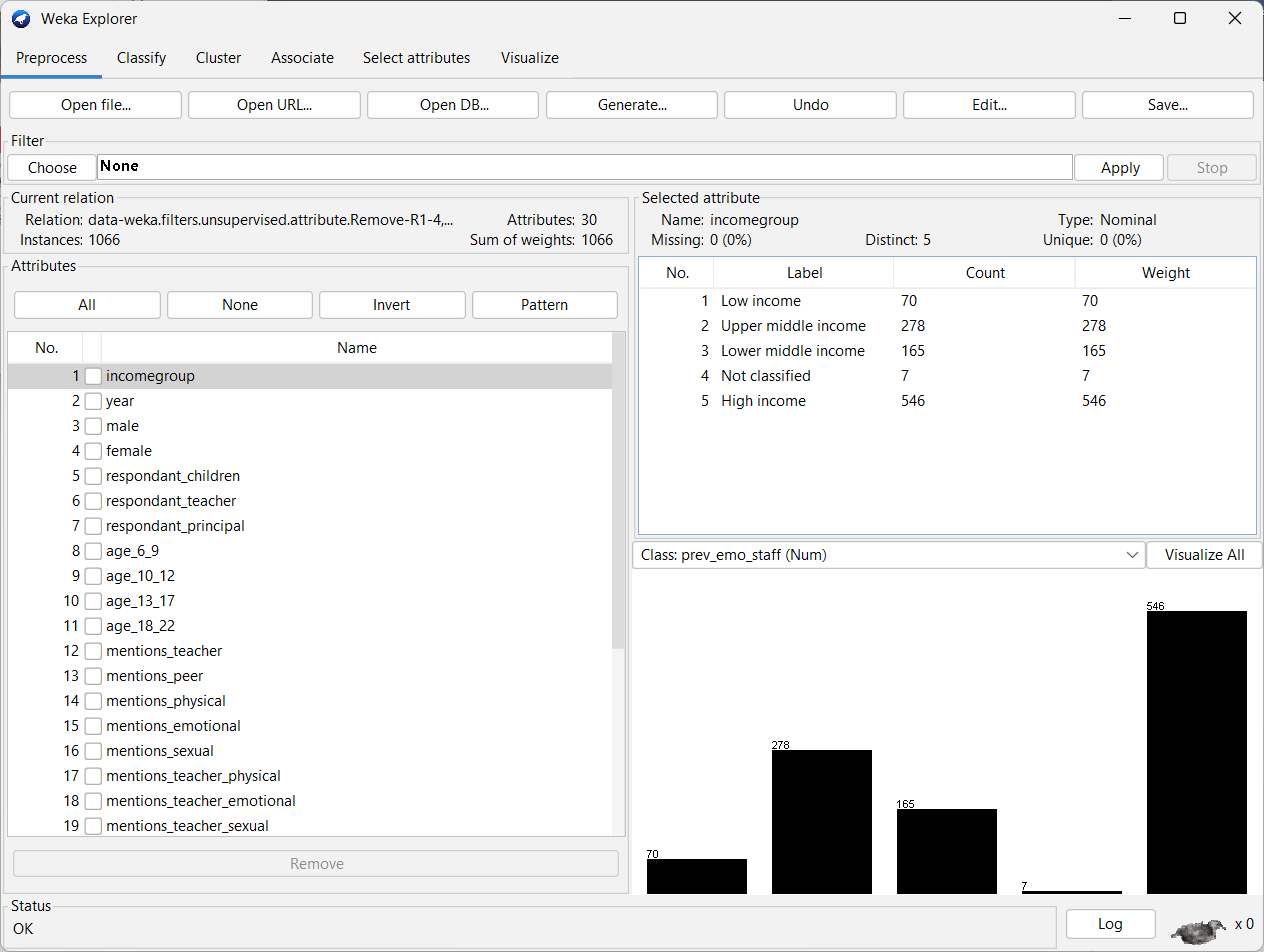


* Xét những thuộc tính dư thừa, không cần thiết đối với mục tiêu phân tích (do mục tiêu của bài toán không hướng đến từng khu vực riêng lẻ):
* *country*
* *region*

**⇒ Thực hiện**: lựa chọn các ô *country, region* và bấm **Remove** để loại bỏ các thuộc tính trên.



* Dữ liệu thu được sau khi loại bỏ các thuộc tính dư thừa:

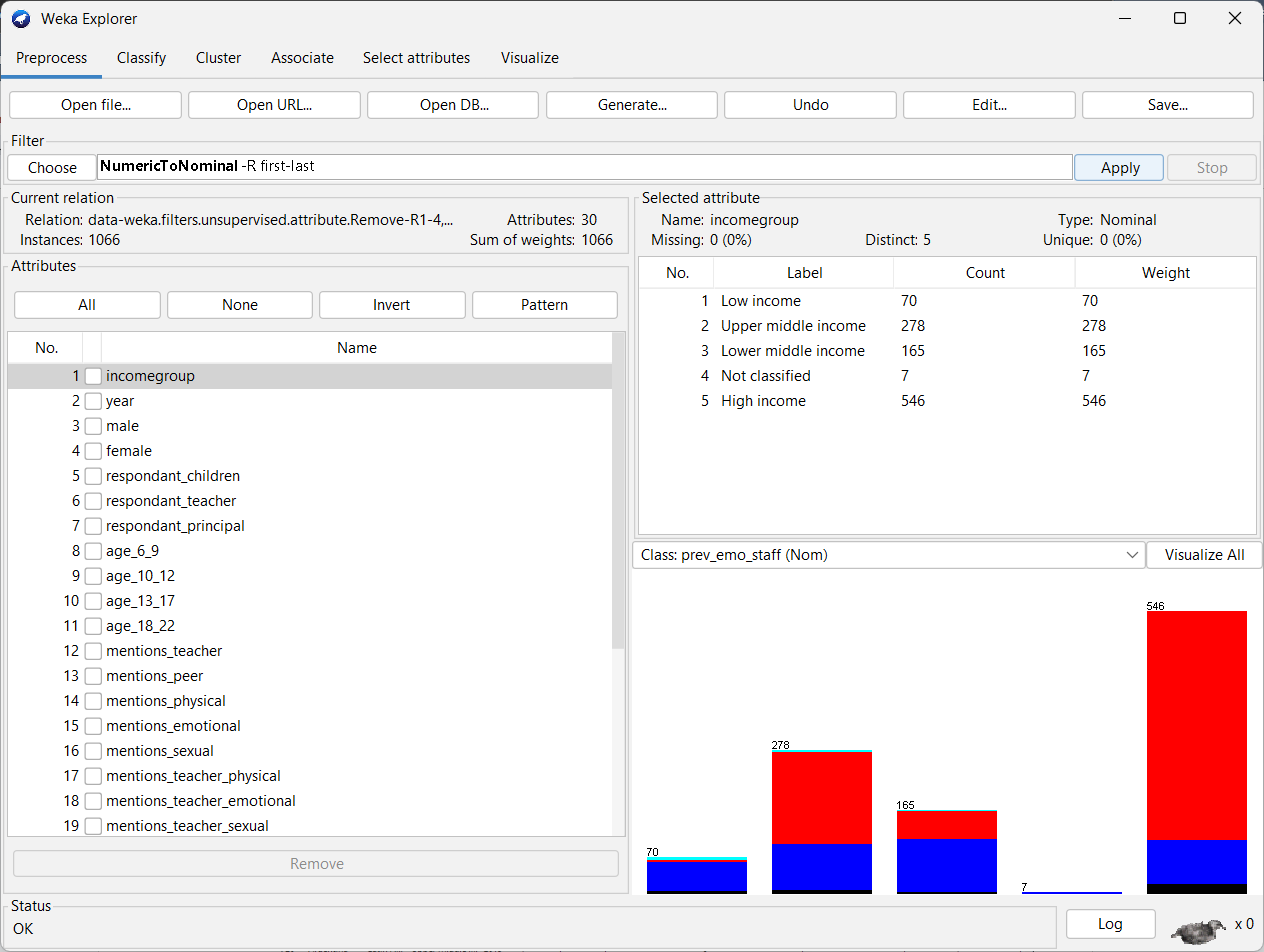


#### Xử lý nhiễu

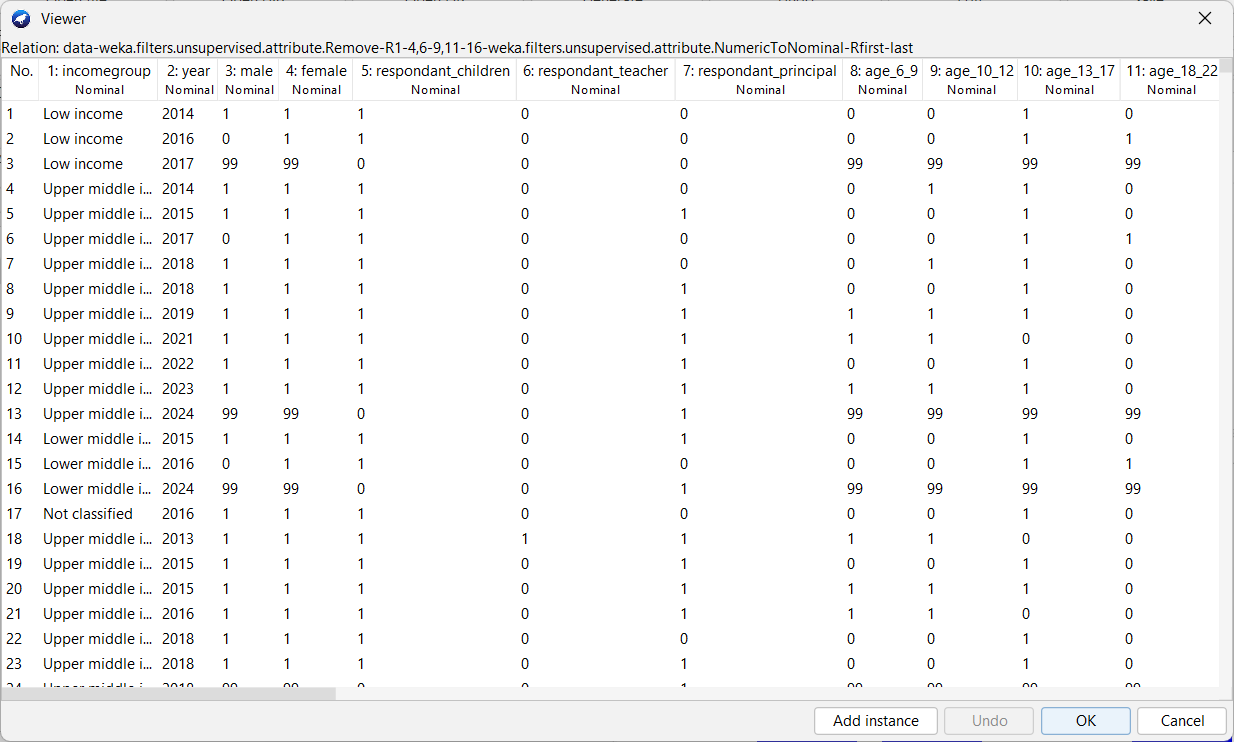
Để thuận tiện cho việc xử lý và phân tích sau đó, ta thực hiện rời rạc hoá dữ liệu bằng cách chuyển kiểu thuộc tính về Nominal

**Thực hiện:**

Filter → Unsupervised → Attribute → NumericToNominal → Apply



Dữ liệu sau khi chuyển về dạng Nominal:

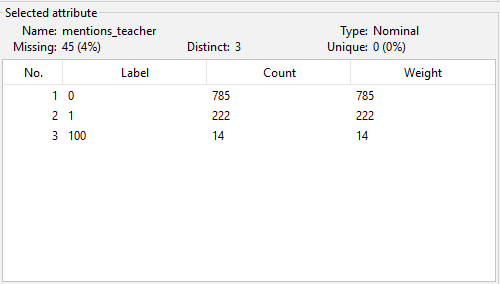


Các giá trị nhiễu gồm:

|  |  |
| --- | --- |
| **Thuộc tính** | **Đánh giá** |
| year | Những giá trị “9999” là những giá trị không xác định được năm mà cuộc khảo sát được thực hiện. |
| age\_6\_9 | Những giá trị “99” là những dòng không được áp dụng. |
| age\_10\_12 | Những giá trị “99” là những dòng không được áp dụng. |
| age\_13\_17 | Những giá trị “99” là những dòng không được áp dụng. |
| age\_18\_22 | Những giá trị “99” là những dòng không được áp dụng. |
| mentions\_teacher | Những giá trị “100” là những dòng cần được xác nhận (do đây là các cuộc khảo sát đang chờ xử lý mà không có bảng câu hỏi có sẵn trong các vòng khảo sát mới) |
| mentions\_peer | Những giá trị “100” là những dòng cần được xác nhận (do đây là các cuộc khảo sát đang chờ xử lý mà không có bảng câu hỏi có sẵn trong các vòng khảo sát mới) |
| mentions\_physical | Những giá trị “100” là những dòng cần được xác nhận (do đây là các cuộc khảo sát đang chờ xử lý mà không có bảng câu hỏi có sẵn trong các vòng khảo sát mới) |
| mentions\_emotional | Những giá trị “100” là những dòng cần được xác nhận (do đây là các cuộc khảo sát đang chờ xử lý mà không có bảng câu hỏi có sẵn trong các vòng khảo sát mới) |
| mentions\_sexual | Những giá trị “100” là những dòng cần được xác nhận (do đây là các cuộc khảo sát đang chờ xử lý mà không có bảng câu hỏi có sẵn trong các vòng khảo sát mới) |
| mentions\_teacher\_physical | Những giá trị “100” là những dòng cần được xác nhận (do đây là các cuộc khảo sát đang chờ xử lý mà không có bảng câu hỏi có sẵn trong các vòng khảo sát mới) |
| mentions\_teacher\_emotional | Những giá trị “100” là những dòng cần được xác nhận (do đây là các cuộc khảo sát đang chờ xử lý mà không có bảng câu hỏi có sẵn trong các vòng khảo sát mới) |
| mentions\_teacher\_sexual | Những giá trị “100” là những dòng cần được xác nhận (do đây là các cuộc khảo sát đang chờ xử lý mà không có bảng câu hỏi có sẵn trong các vòng khảo sát mới) |
| mentions\_peer\_physical | Những giá trị “100” là những dòng cần được xác nhận (do đây là các cuộc khảo sát đang chờ xử lý mà không có bảng câu hỏi có sẵn trong các vòng khảo sát mới) |
| mentions\_peer\_emotional | Những giá trị “100” là những dòng cần được xác nhận (do đây là các cuộc khảo sát đang chờ xử lý mà không có bảng câu hỏi có sẵn trong các vòng khảo sát mới) |
| mentions\_peer\_sexual | Những giá trị “100” là những dòng cần được xác nhận (do đây là các cuộc khảo sát đang chờ xử lý mà không có bảng câu hỏi có sẵn trong các vòng khảo sát mới) |
| prev\_sexual\_victim | Những giá trị “100” là những dòng cần được xác nhận (do đây là các cuộc khảo sát đang chờ xử lý mà không có bảng câu hỏi có sẵn trong các vòng khảo sát mới) |
| prev\_sexual\_staff | Những giá trị “100” là những dòng cần được xác nhận (do đây là các cuộc khảo sát đang chờ xử lý mà không có bảng câu hỏi có sẵn trong các vòng khảo sát mới) |
| prev\_physical\_victim | Những giá trị “100” là những dòng cần được xác nhận (do đây là các cuộc khảo sát đang chờ xử lý mà không có bảng câu hỏi có sẵn trong các vòng khảo sát mới) |
| prev\_physical\_staff | Những giá trị “100” là những dòng cần được xác nhận (do đây là các cuộc khảo sát đang chờ xử lý mà không có bảng câu hỏi có sẵn trong các vòng khảo sát mới) |
| prev\_fight\_victim | Những giá trị “100” là những dòng cần được xác nhận (do đây là các cuộc khảo sát đang chờ xử lý mà không có bảng câu hỏi có sẵn trong các vòng khảo sát mới) |
| prev\_fight\_staff | Những giá trị “100” là những dòng cần được xác nhận (do đây là các cuộc khảo sát đang chờ xử lý mà không có bảng câu hỏi có sẵn trong các vòng khảo sát mới) |
| prev\_emo\_victim | Những giá trị “100” là những dòng cần được xác nhận (do đây là các cuộc khảo sát đang chờ xử lý mà không có bảng câu hỏi có sẵn trong các vòng khảo sát mới) |
| prev\_emo\_staff | Những giá trị “100” là những dòng cần được xác nhận (do đây là các cuộc khảo sát đang chờ xử lý mà không có bảng câu hỏi có sẵn trong các vòng khảo sát mới) |

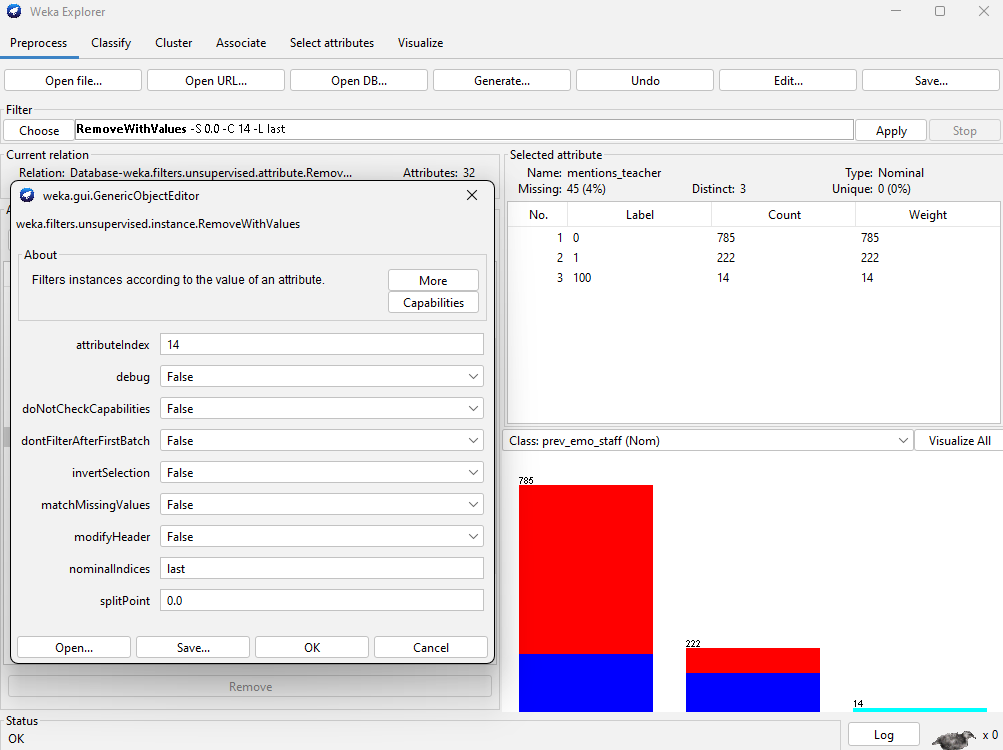
Các bước xử lý:

* Tại thuộc tính mentions\_teacher: giá trị nhiễu là “100”, gồm có 14 bản ghi

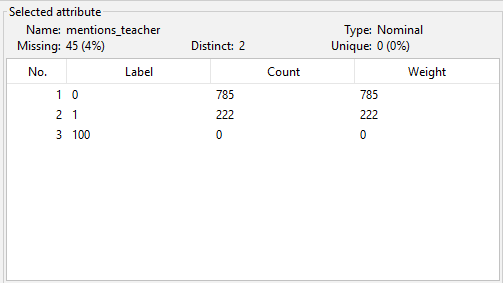


**Thực hiện:** loại bỏ những bản ghi có giá trị nhiễu

Filter → Unsupervised → Instance → RemoveWithValues

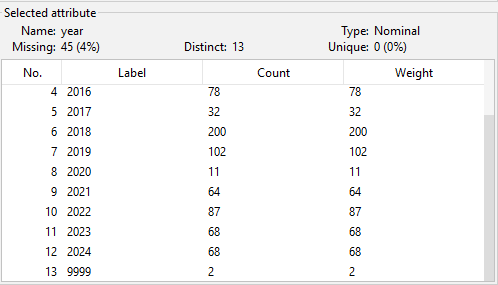


Dữ liệu tại mentions\_teacher sau khi xoá:

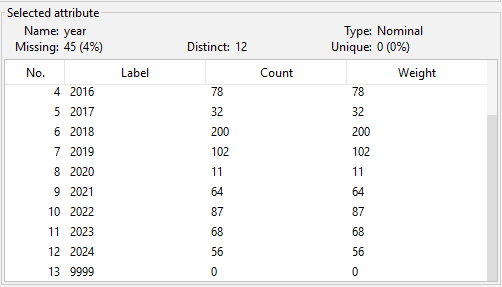


Xử lý tương tự với các thuộc tính còn lại.

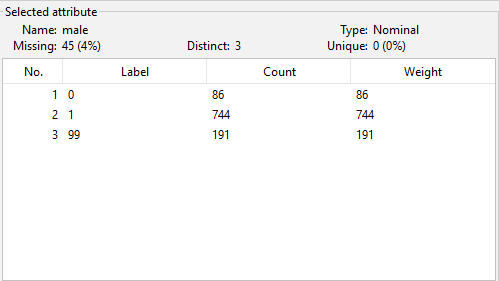
* Dữ liệu tại thuộc tính *year* ban đầu: giá trị “9999” có 2 bản ghi



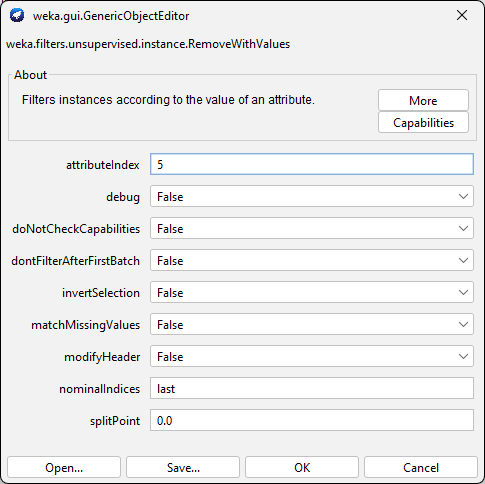
Dữ liệu tại thuộc tính *year* sau khi xử lý:



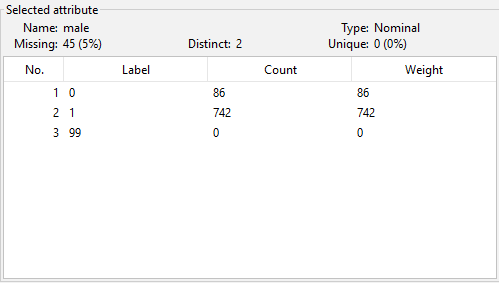
* Dữ liệu tại thuộc tính *male* ban đầu: giá trị “99” gồm 191 bản ghi



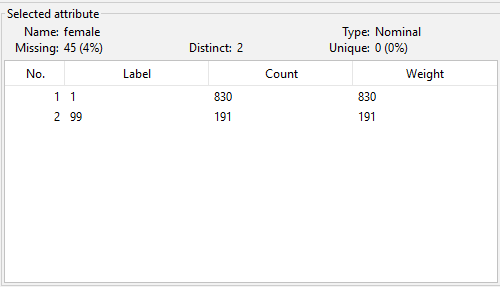
* Áp dụng RemoveWithValues:



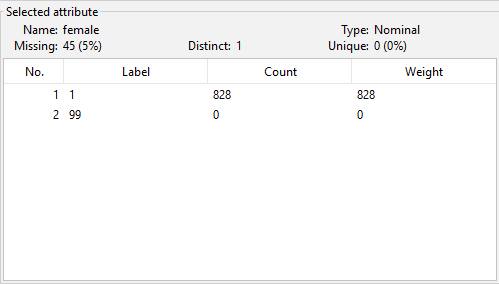
Dữ liệu tại thuộc tính *male* sau khi xử lý:



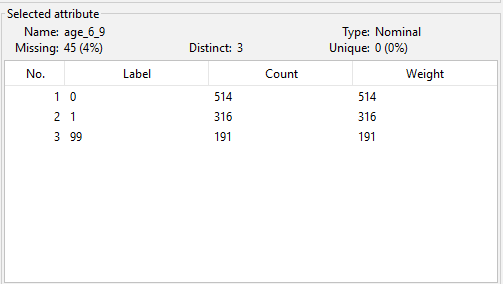
* Dữ liệu tại thuộc tính *female* ban đầu: giá trị “99” gồm 191 bản ghi



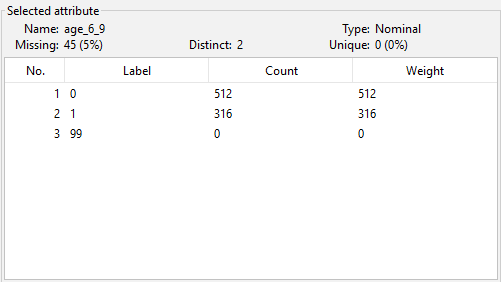
Dữ liệu tại thuộc tính *female* sau khi xử lý:



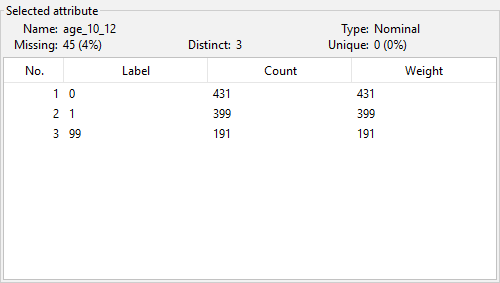
* Dữ liệu tại thuộc tính *age\_6\_9* ban đầu:



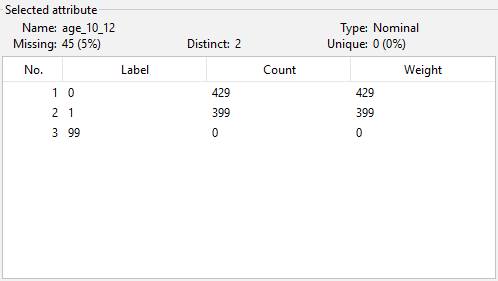
Dữ liệu tại thuộc tính *age\_6\_9* sau khi xử lý:



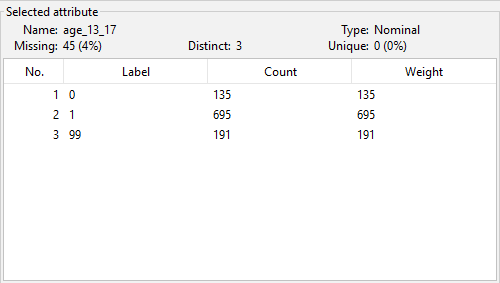
* Dữ liệu tại thuộc tính *age\_10\_12* ban đầu:



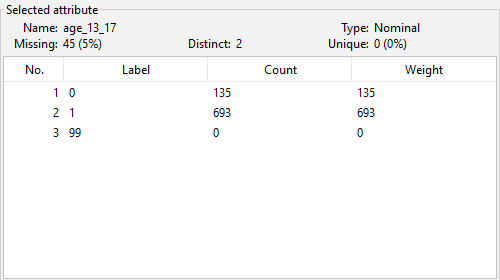
Dữ liệu tại thuộc tính *age\_10\_12* sau khi xử lý:



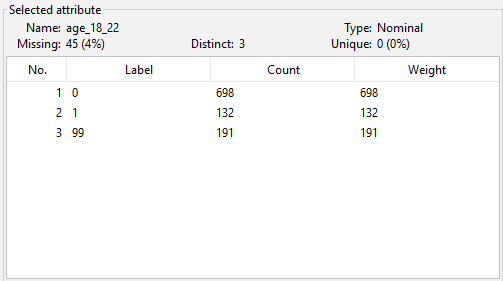
* Dữ liệu tại thuộc tính *age\_13\_17* ban đầu:



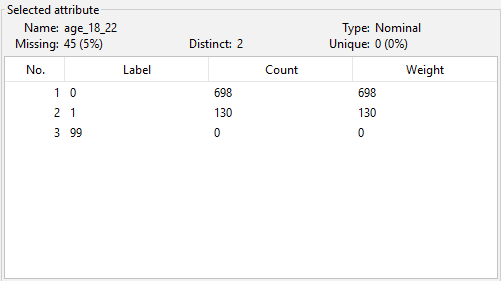
Dữ liệu tại thuộc tính *age\_13\_17* sau khi xử lý:



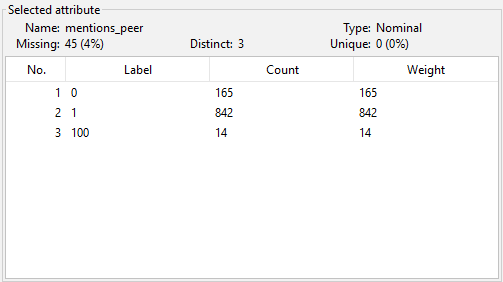
* Dữ liệu tại thuộc tính *age\_18\_22* ban đầu:



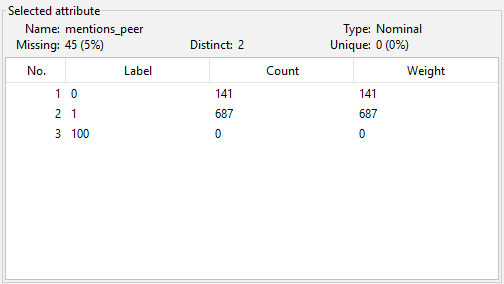
Dữ liệu tại thuộc tính *age\_18\_22* sau khi xử lý:



* Dữ liệu tại thuộc tính *mentions\_peer* ban đầu:



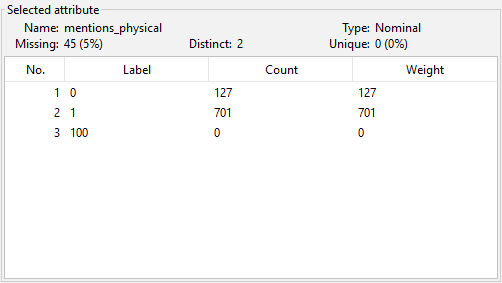
Dữ liệu tại thuộc tính *mentions\_peer* sau khi xử lý:



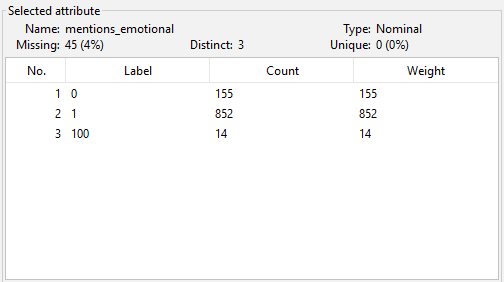
* Dữ liệu tại thuộc tính *mentions\_physical* ban đầu:



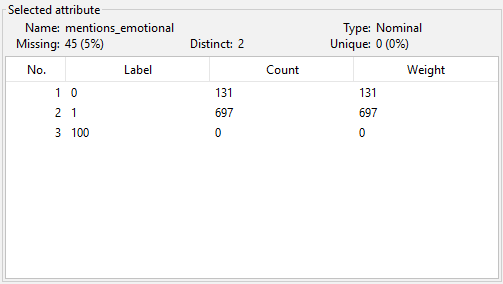
Dữ liệu tại thuộc tính *mentions\_physical* sau khi xử lý:



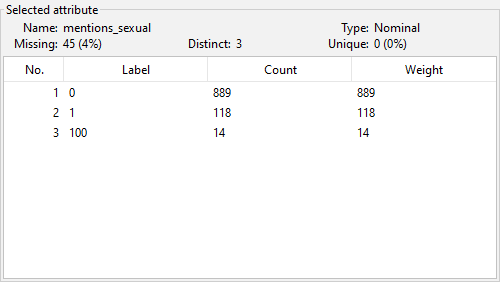
* Dữ liệu tại thuộc tính *mentions\_emotional* ban đầu:



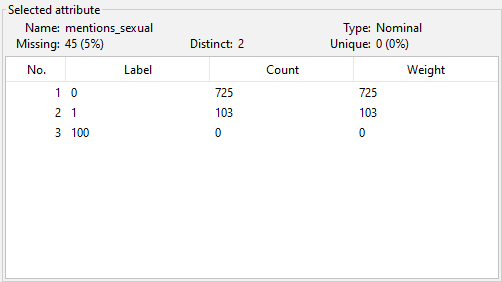
Dữ liệu tại thuộc tính *mentions\_emotional* sau khi xử lý:



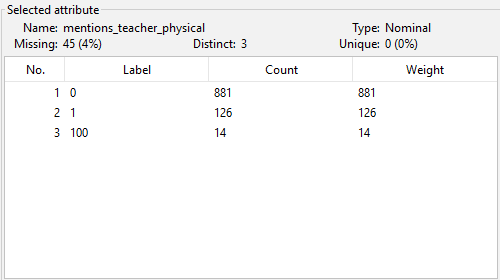
* Dữ liệu tại thuộc tính *mentions\_sexual* ban đầu:



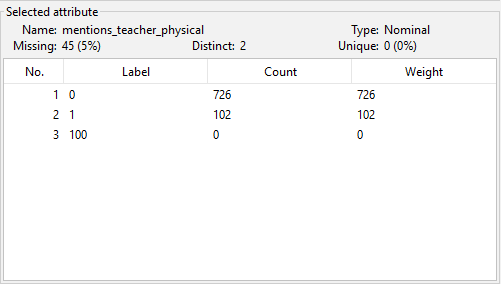
Dữ liệu tại thuộc tính *mentions\_sexual* sau khi xử lý:



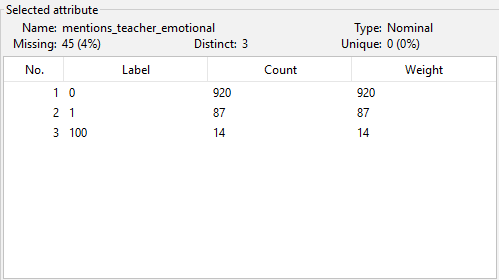
* Dữ liệu tại thuộc tính *mentions\_teacher\_physical* ban đầu:



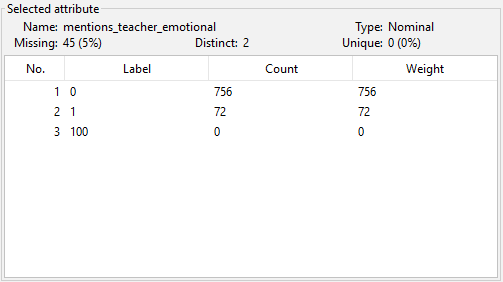
Dữ liệu tại thuộc tính *mentions\_teacher\_physical* sau khi xử lý:



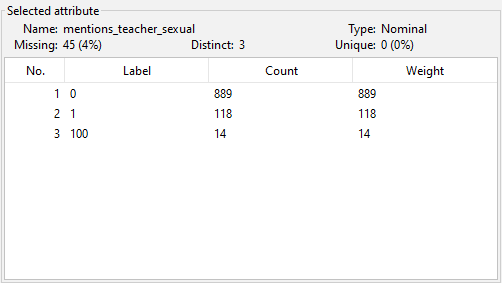
* Dữ liệu tại thuộc tính *mentions\_teacher\_emotional* ban đầu:



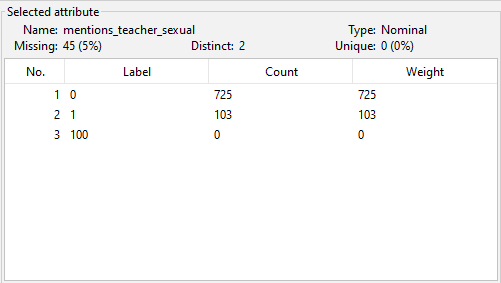
Dữ liệu tại thuộc tính *mentions\_teacher\_emotional* sau khi xử lý:



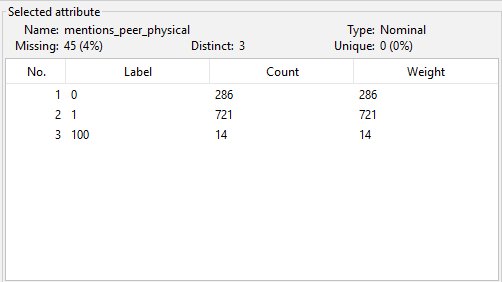
* Dữ liệu tại thuộc tính *mentions\_teacher\_sexual* ban đầu:



Dữ liệu tại thuộc tính *mentions\_teacher\_sexual* sau khi xử lý:



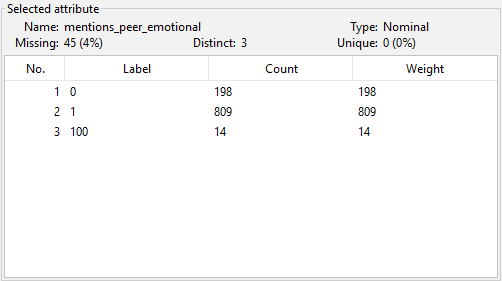
* Dữ liệu tại thuộc tính *mentions\_peer\_physical* ban đầu:



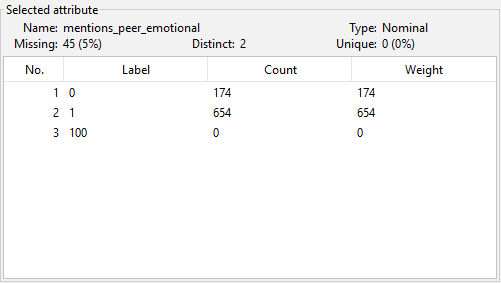
Dữ liệu tại thuộc tính *mentions\_peer\_physical* sau khi xử lý:



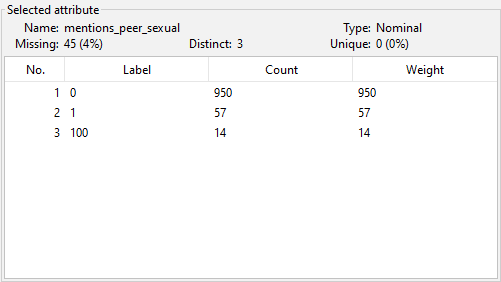
* Dữ liệu tại thuộc tính *mentions\_peer\_emotional* ban đầu:



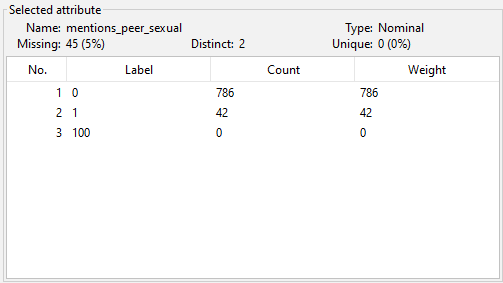
Dữ liệu tại thuộc tính *mentions\_peer\_emotional* sau khi xử lý:



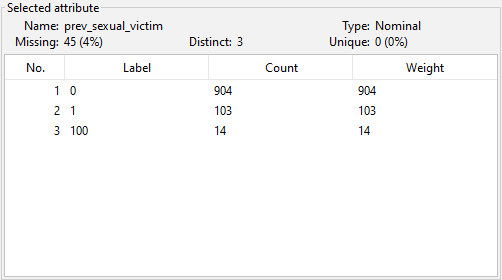
* Dữ liệu tại thuộc tính *mentions\_peer\_sexual* ban đầu:



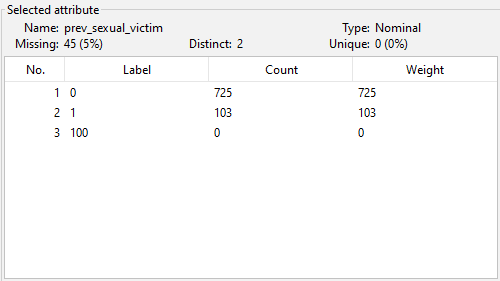
Dữ liệu tại thuộc tính *mentions\_peer\_sexual* sau khi xử lý:



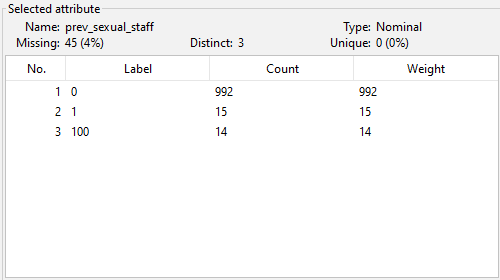
* Dữ liệu tại thuộc tính *prev\_sexual\_victim* ban đầu:



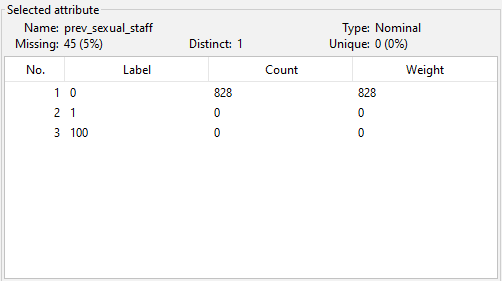
Dữ liệu tại thuộc tính *prev\_sexual\_victim* sau khi xử lý:



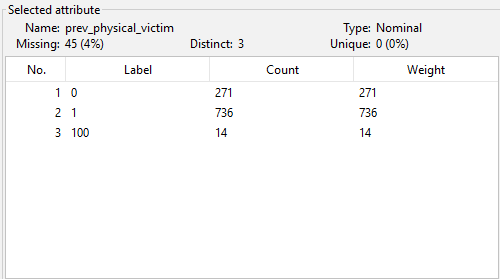
* Dữ liệu tại thuộc tính *prev\_sexual\_staff* ban đầu:



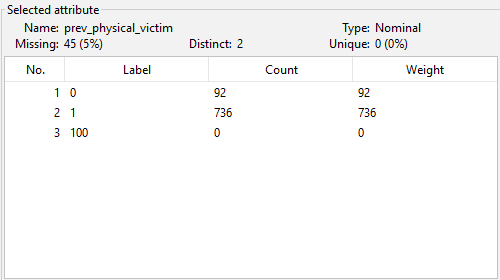
Dữ liệu tại thuộc tính *prev\_sexual\_staff* sau khi xử lý:



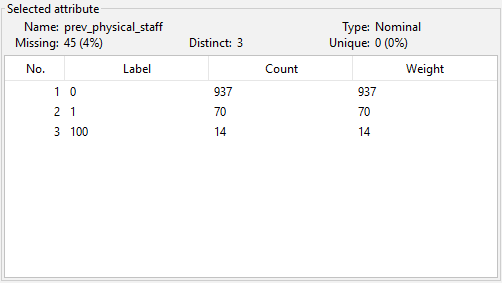
* Dữ liệu tại thuộc tính *prev\_physical\_victim* ban đầu:



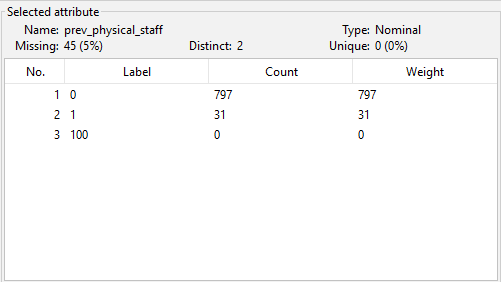
Dữ liệu tại thuộc tính *prev\_physical\_victim* sau khi xử lý:



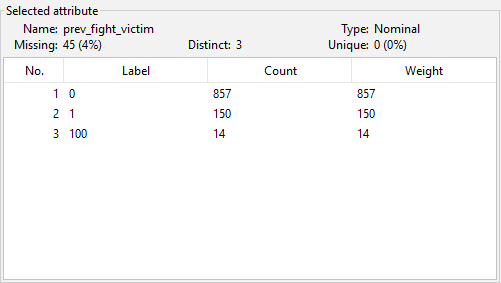
* Dữ liệu tại thuộc tính *prev\_physical\_staff* ban đầu:



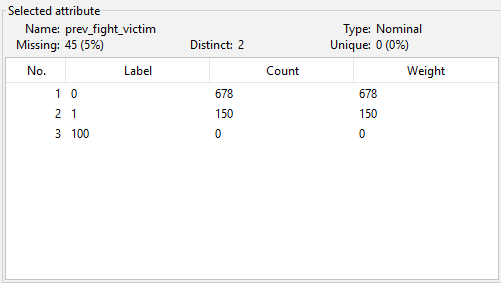
Dữ liệu tại thuộc tính *prev\_physical\_staff* sau khi xử lý:



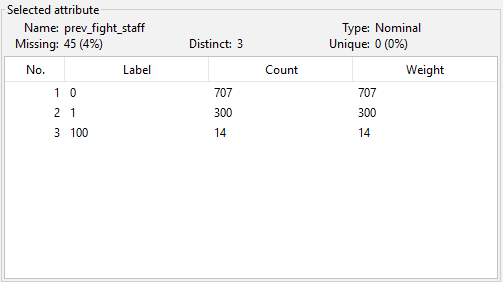
* Dữ liệu tại thuộc tính *prev\_fight\_victim* ban đầu:



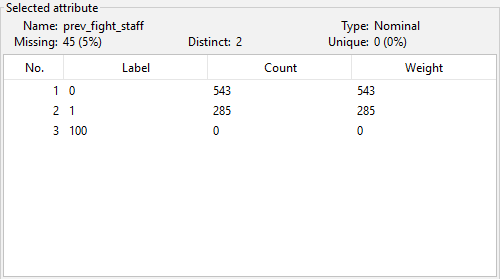
Dữ liệu tại thuộc tính *prev\_fight\_victim* sau khi xử lý:



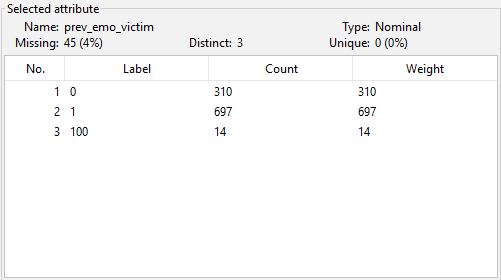
* Dữ liệu tại thuộc tính *prev\_fight\_staff* ban đầu:



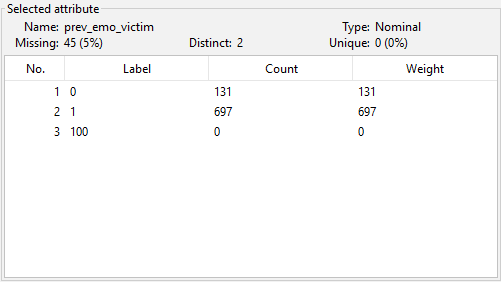
Dữ liệu tại thuộc tính *prev\_fight\_staff* sau khi xử lý:



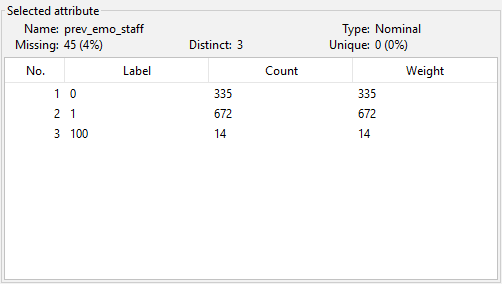
* Dữ liệu tại thuộc tính *prev\_emo\_victim* ban đầu:



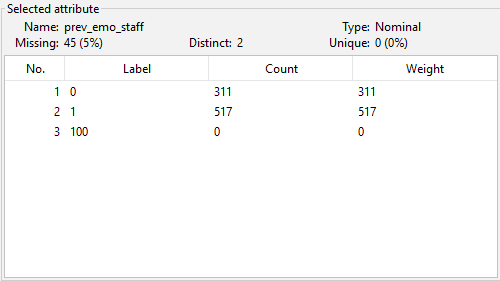
Dữ liệu tại thuộc tính *prev\_emo\_victim* sau khi xử lý:



* Dữ liệu tại thuộc tính *prev\_emo\_staff* ban đầu:

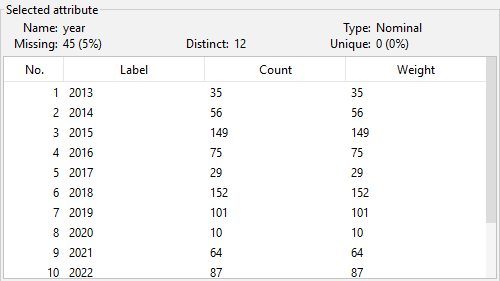


Dữ liệu tại thuộc tính *prev\_emo\_staff* sau khi xử lý:



#### Xử lý dữ liệu thiếu

Có 45 bản ghi chỉ có dữ liệu ở các cột *country, countrycode, region, regioncode, incomegroup, incomecode* mà không có dữ liệu trên các cột còn lại. Điều này là do các cuộc khảo sát ở các khu vực này chưa được thực hiện hoặc chưa có dữ liệu được cập nhật.

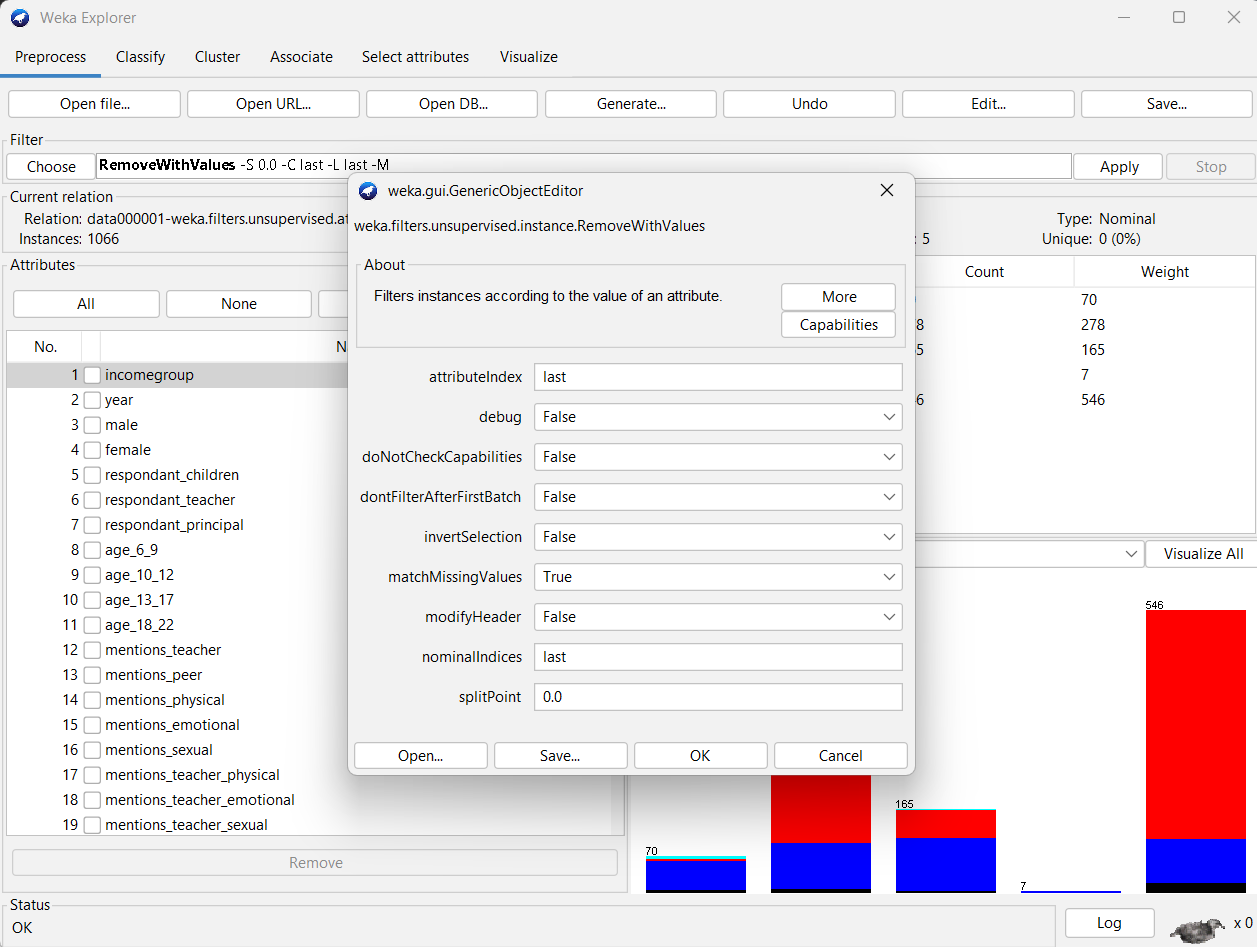


*Missing tại thuộc tính year*

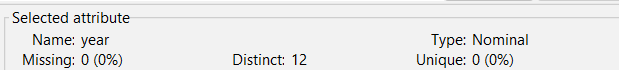
Do dữ liệu bị thiếu này chỉ chiếm một lượng không đáng kể (4%), không ảnh hưởng kết quả phân tích, ta loại bỏ những dữ liệu này.

**Thực hiện:**

Filter → Unsupervised → Instance → RemoveWithValues



Kết quả:



#### Xử lý dữ liệu không nhất quán

Tập dữ liệu đã nhất quán.

### Tích hợp dữ liệu

Tích hợp dữ liệu là quá trình trộn dữ liệu từ các nguồn khác nhau vào một kho dữ liệu sẵn sàng cho quá trình khai phá dữ liệu.

Tích hợp dữ liệu sẽ hỗ trợ việc giảm và tránh dư thừa và không nhất quán về dữ liệu, qua đó cải thiện tính chính xác và tốc độ quá trình khai phá dữ liệu.

⇒ Do dữ liệu này lấy từ một nguồn nên quy trình này bỏ qua.

## Phân tích dữ liệu sau khi tiền xử lý lần đầu

Phân tích dữ liệu nhằm hiểu rõ hơn về dữ liệu và mối quan hệ giữa các thuộc tính do đó ta cần phân tích, nhận biết thêm về sự liên kết giữa chúng.

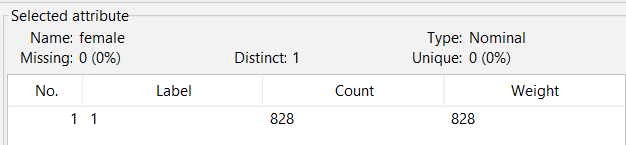
Dữ liệu sau khi tiền xử lý lần đầu:



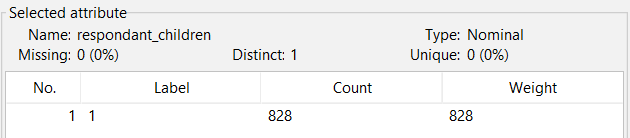
Dựa vào thống kê này, ta nhận thấy có một số thuộc tính chỉ còn lại 1 giá trị sau khi thực hiện tiền xử lý. Những thuộc tính này có thể được coi là không cung cấp thông tin hữu ích cho việc phân tích và mô hình hóa do không có sự biến thiên và không thể giúp phân biệt hay đóng góp vào việc dự đoán kết quả.

Những thuộc tính đó là:

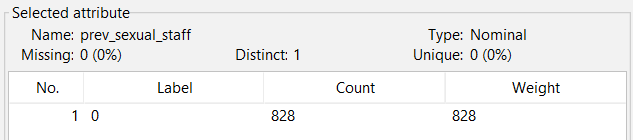
* *female*:



* *respondant\_children*:



* *prev\_sexual\_staff*:



### Loại bỏ thuộc tính dư thừa

Sử dụng Weka loại bỏ những thuộc tính trên:

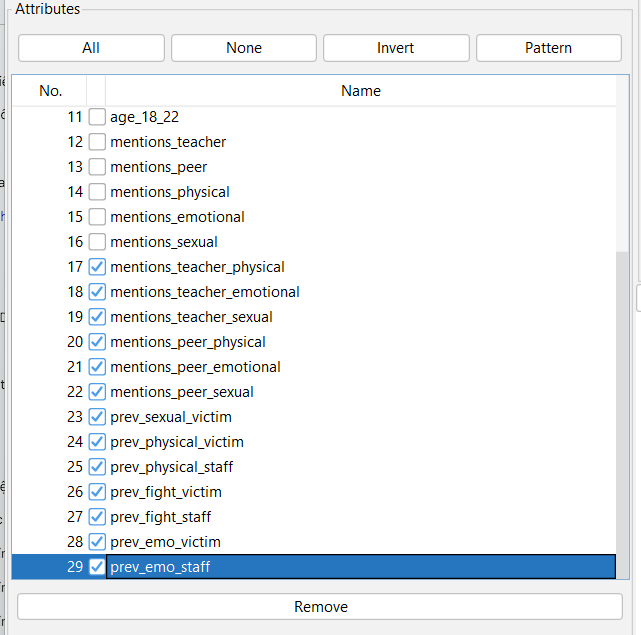




Ta nhận thấy rằng một số thuộc tính có thể được suy diễn từ thuộc tính khác:

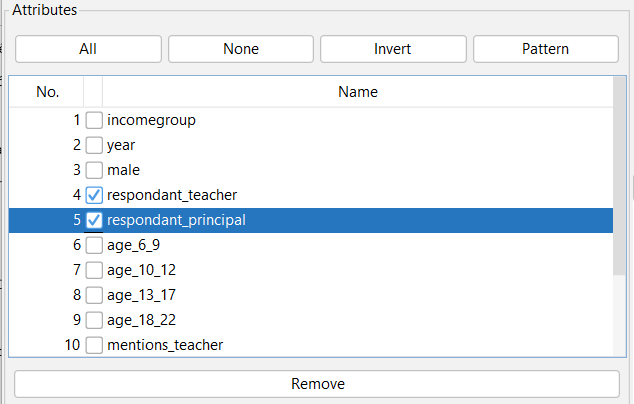
* Những thuộc tính có thể được suy diễn từ *mentions\_teacher* (do ý nghĩa của các thuộc tính này đều đề cập tới đối tượng thực hiện hành vi bạo lực là giáo viên):
* *mentions\_teacher\_physical*
* *mentions\_teacher\_emotional*
* *mentions\_teacher\_sexual*
* Những thuộc tính có thể được suy diễn từ *mentions\_peer* (do ý nghĩa của các thuộc tính này đều đề cập tới đối tượng thực hiện hành vi bạo lực là bạn bè):
* *mentions\_peer\_physical*
* *mentions\_peer\_emotional*
* *mentions\_peer\_sexual*
* Những thuộc tính có thể được suy diễn từ *mentions\_sexual* (do ý nghĩa của các thuộc tính này đều đề cập đến bạo lực liên quan tới tình dục):
* *prev\_sexual\_victim*
* *prev\_sexual\_staff*
* Những thuộc tính có thể được suy diễn từ *mentions\_physical* (do ý nghĩa của các thuộc tính này đều đề cập đến bạo lực liên quan tới thể xác):
* *prev\_physical\_victim*
* *prev\_physical\_staff*
* *prev\_fight\_victim*
* *prev\_fight\_staff*
  + Những thuộc tính có thể được suy diễn từ *mentions\_emotional* (do ý nghĩa của các thuộc tính này đều đề cập đến bạo lực liên quan tới tinh thần):
* *prev\_emo\_victim*
* *prev\_emo\_staff*

Thực hiện loại bỏ các thuộc tính trên bằng Weka: Lựa chọn các thuộc tính tương ứng và nhấn Remove



Loại bỏ những thuộc tính không ảnh hưởng tới việc phân tích (những thuộc tính này mang ý nghĩa thể hiện vai trò của người thực hiện khảo sát):

* *respodant\_teacher*
* *respodant\_principal*

**

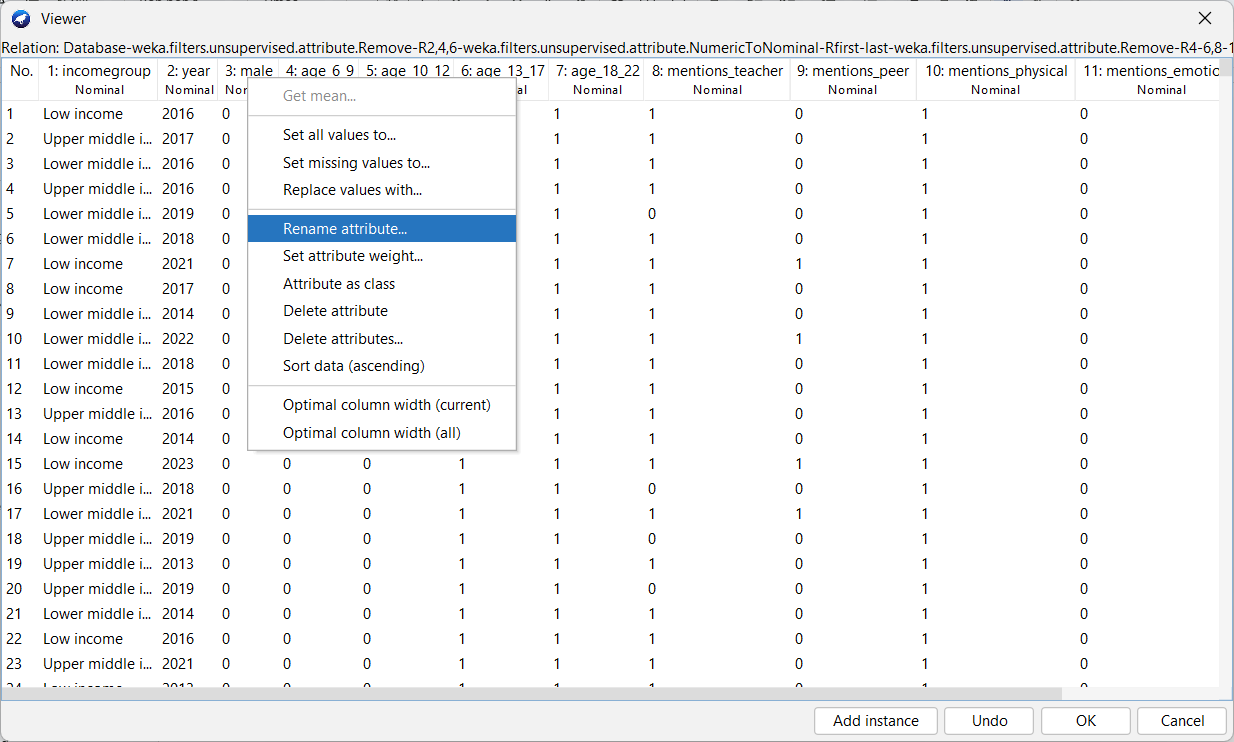
### Biến đổi dữ liệu

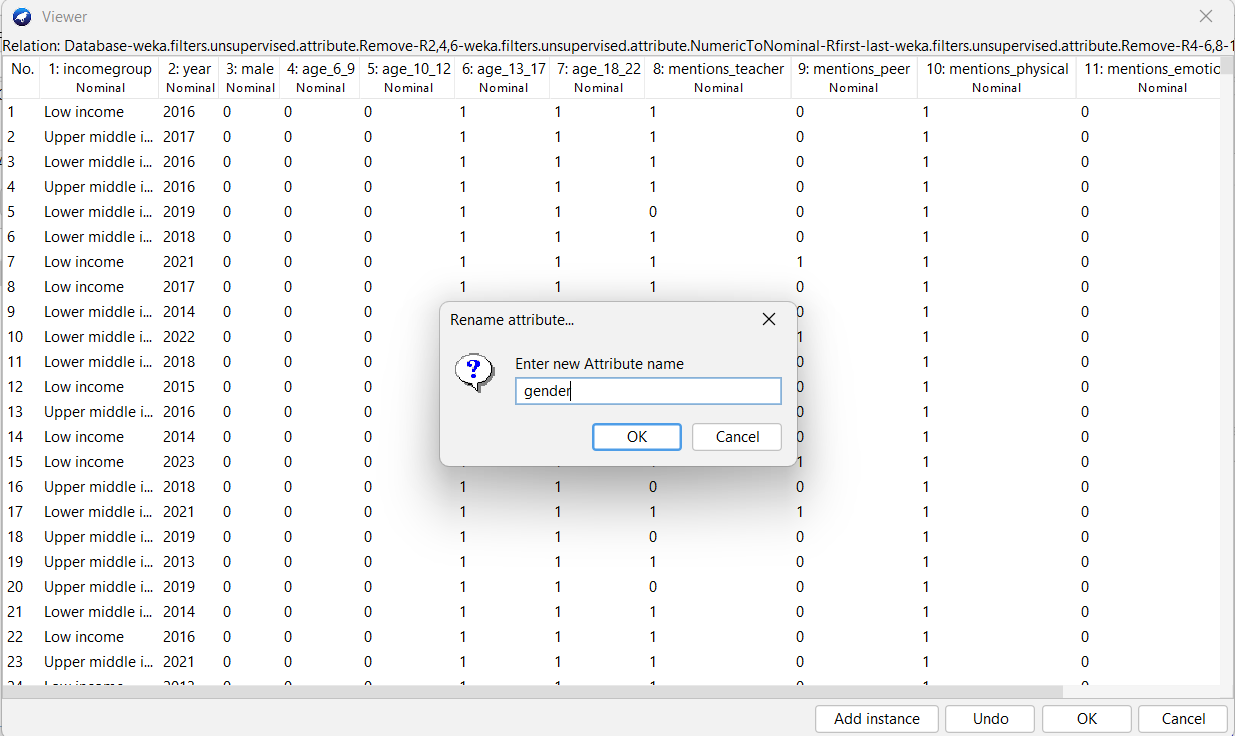
Là quá trình biến đổi hay kết hợp dữ liệu vào những dạng thích hợp cho quá trình khai phá dữ liệu.

* Có thể thấy thuộc tính *male* có hai giá trị là 0 (=No) và 1 (=Yes). Vì thuộc tính *female* (đã được loại bỏ ở bước trước đó) có giá trị 1 (=Yes) ở tất cả bản ghi, ta có thể hiểu giá trị 0 (=No) ở *male* có nghĩa là *female.* Để tổng quát hơn, ta thực hiện đổi tên thuộc tính *male* thành *gender,* có giá trị 1 đại diện cho “male” và giá trị 0 đại diện cho “female”.

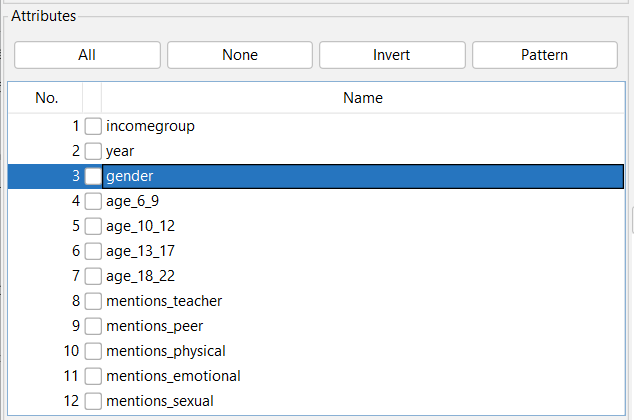
Thực hiện:

* Mở giao diện Edit
* Nhấp chuột phải vào cột *male* cần đổi tên





Kết quả:



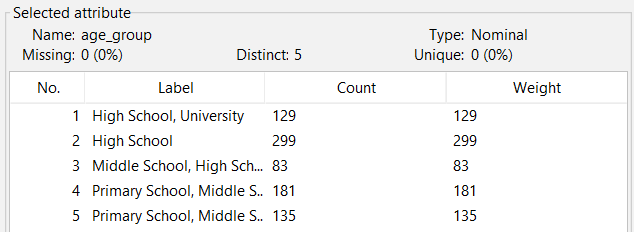
* Nhận thấy rằng các cột *age\_6\_9, age\_10\_12, age\_13\_17, age\_18\_22* có cùng ý nghĩa thể hiện tuổi của đối tượng được khảo sát, ta gộp các thuộc tính này thành một thuộc tính thể hiện độ tuổi: *age\_group*. Với các giá trị:

|  |  |
| --- | --- |
| **Giá trị** | **Ý nghĩa** |
| Primary School | Độ tuổi từ 6 đến 9 tuổi |
| Middle School | Độ tuổi từ 10 đến 12 tuổi |
| High School | Độ tuổi từ 13 đến 17 tuổi |
| University | Độ tuổi từ 18 đến 22 tuổi |

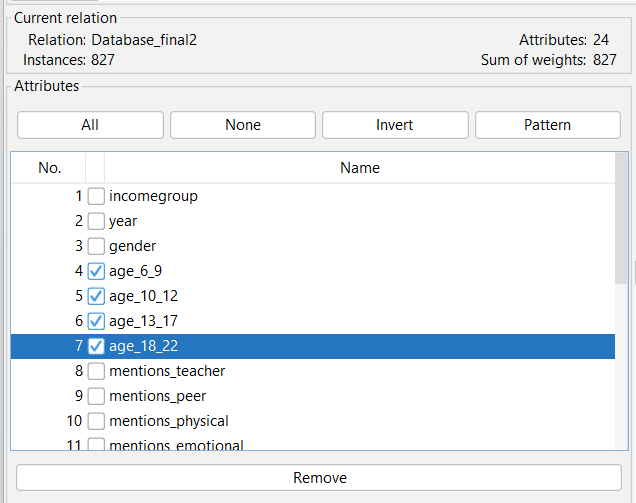
Sử dụng Python:



Kết quả:

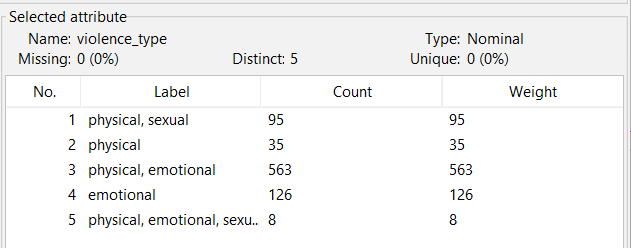


Đồng thời thực hiện xoá những cột *age\_6\_9, age\_10\_12, age\_13\_17, age\_18\_22*: Chọn các thuộc tính này và Remove



* Nhận thấy rằng các cột *mentions\_physical, mentions\_emotional, mentions\_sexual,* có cùng ý nghĩa thể hiện về mức độ và loại bạo lực, ta gộp các thuộc tính này thành cột *violence\_type* với cách thức tương tự như *age\_group.*

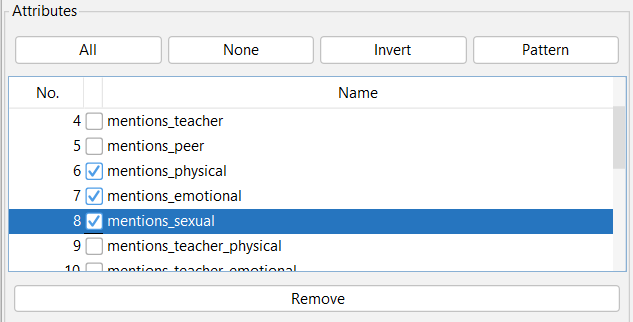
Kết quả:



Ý nghĩa:

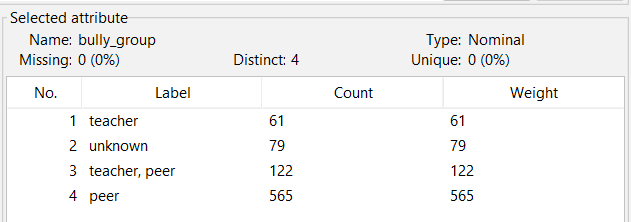
|  |  |
| --- | --- |
| **Giá trị** | **Ý nghĩa** |
| physical | Bị ảnh hưởng bởi bạo lực thể xác |
| emotional | Bị ảnh hưởng bởi bạo lực tinh thần |
| sexual | Bị ảnh hưởng bởi bạo lực tình dục |

Xoá bỏ những thuộc tính cũ:



* Tương tự, cột *mentions\_teacher* và *mentions\_peer* có cùng ý nghĩa về đối tượng gây ra bạo lực, ta gộp giá trị của các thuộc tính này thành *bully\_group.*

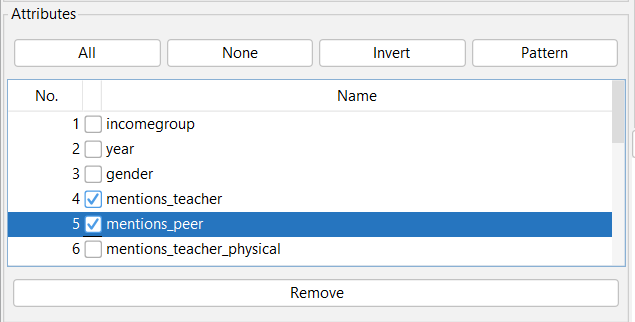
Kết quả:



Ý nghĩa:

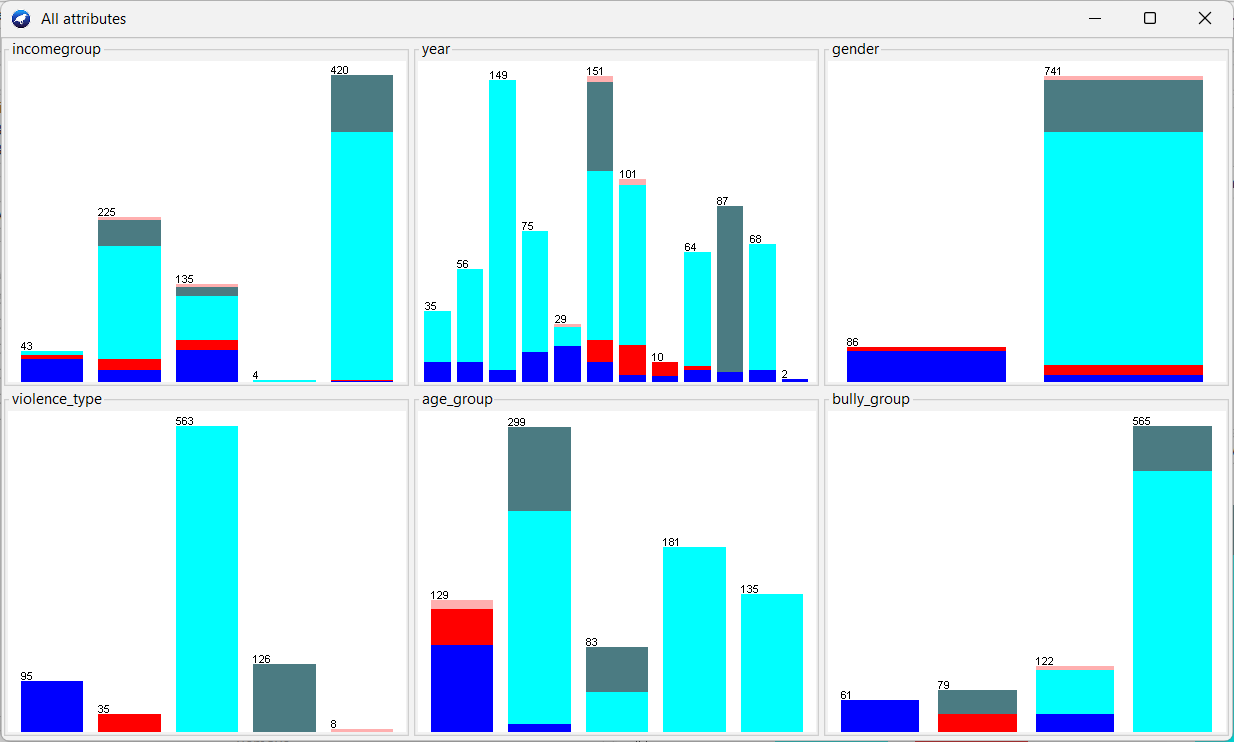
|  |  |
| --- | --- |
| **Giá trị** | **Ý nghĩa** |
| teacher | Bị bạo lực bởi giáo viên |
| peer | Bị bạo lực bởi bạn bè cùng trang lứa |
| unknow | Không rõ đối tượng |

Xoá những thuộc tính cũ:



## Phân tích dữ liệu sau khi tiền xử lý

Thống kê chi tiết dữ liệu:

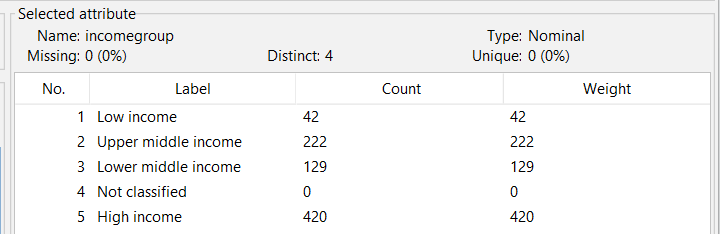


Ta nhận thấy giá trị “Not classified” (không được xếp loại) ở thuộc tính *incomegroup* chỉ có một số lượng bản ghi rất nhỏ (4 bản ghi), bên cạnh đó còn có “2024” của *year* (2 bản ghi) và “physical, emotional, sexual” của *violence\_type* (8 bản ghi) trên tổng số 827 bản ghi. Những giá trị này có thể ảnh hưởng tới độ chính xác của mô hình. Ta thực hiện loại bỏ những giá trị này với filter RemoveWithValue của Weka:

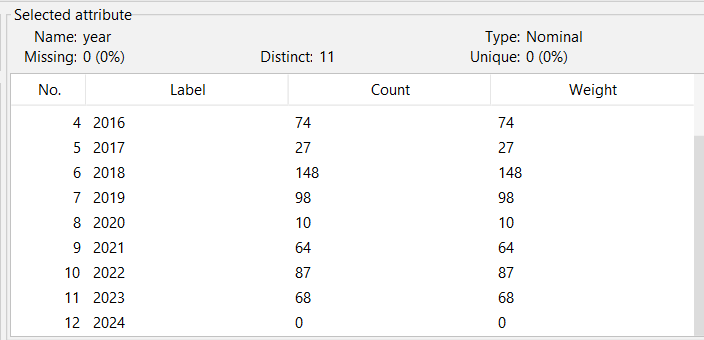
Filter → Unsupervised → Instance → RemoveWithValues

Kết quả sau khi thực hiện:

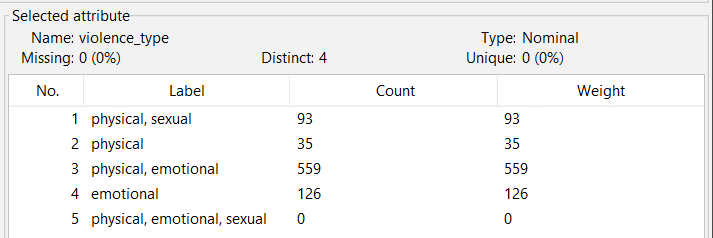
* Với thuộc tính *incomegroup*:



* Với thuộc tính *year:*



* Với thuộc tính *violence\_type*:

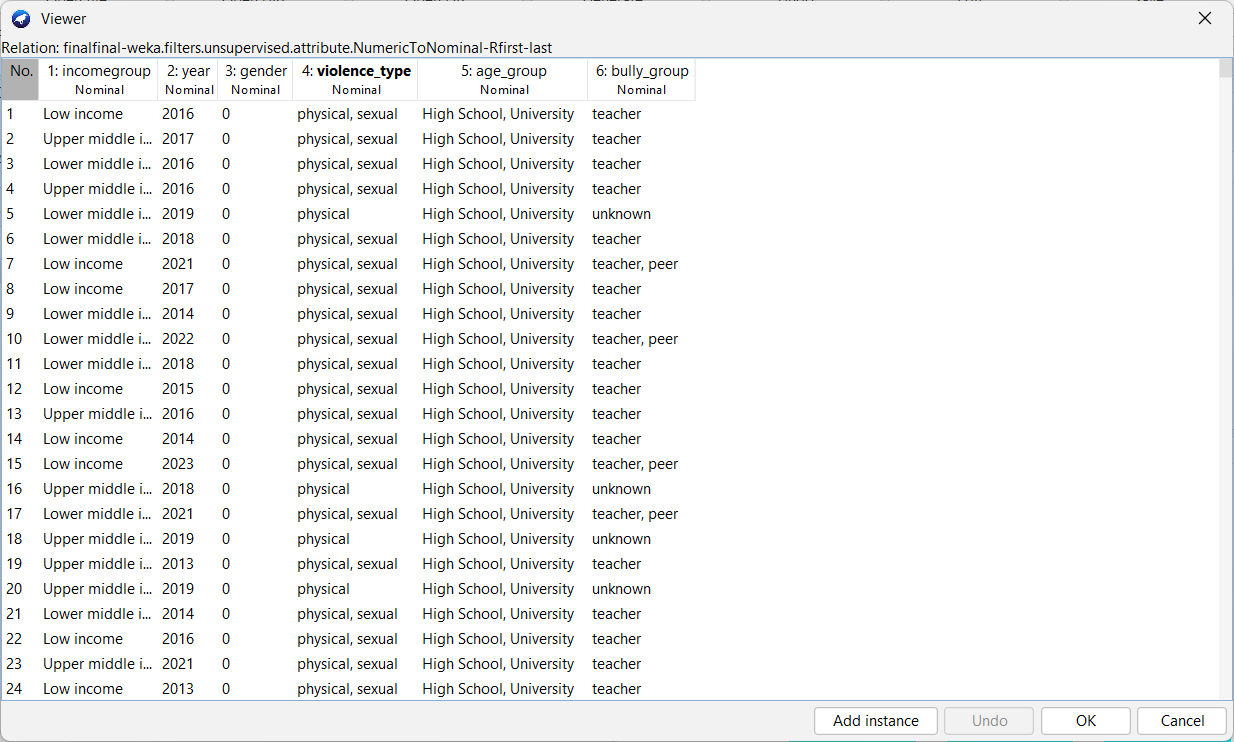


## Đánh giá dữ liệu

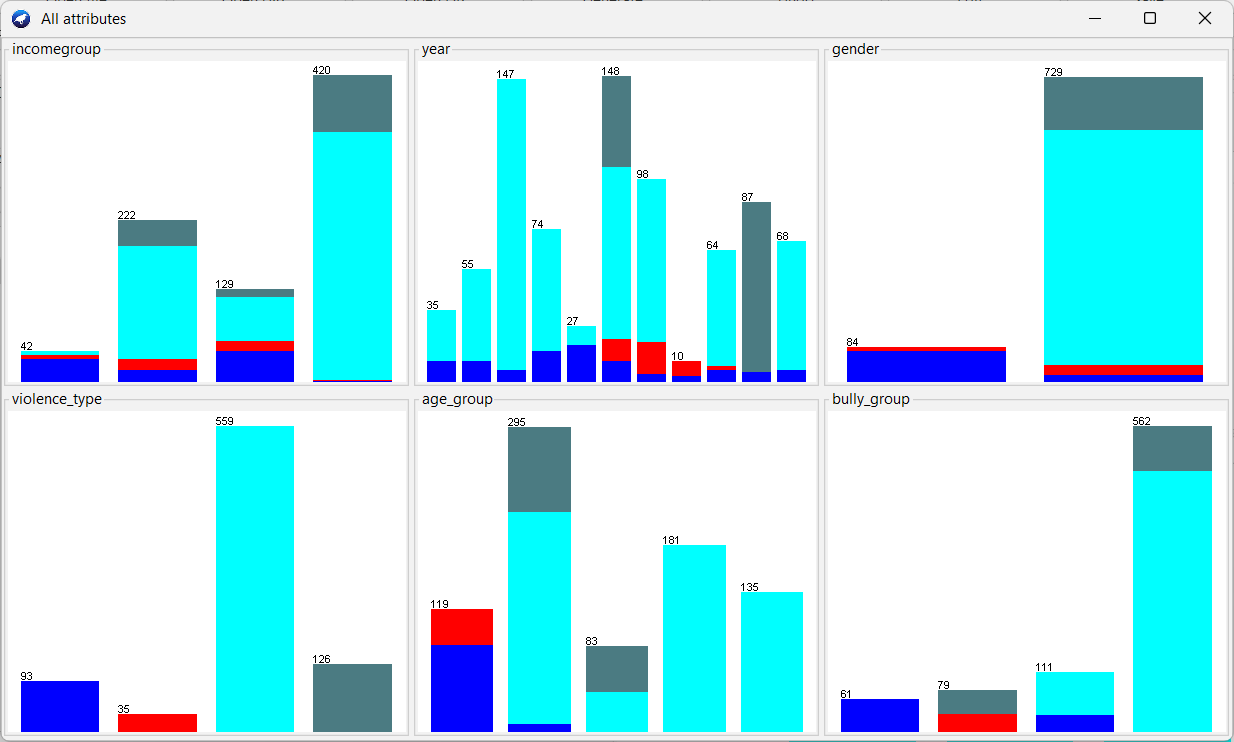
Dữ liệu đưa vào khai phá gồm **6 thuộc tính** và **813 mẫu**

|  |  |  |
| --- | --- | --- |
| **STT** | **Thuộc tính** | **Ý nghĩa** |
| 1 | incomegroup | Nhóm thu nhập của phạm vi đất nước hoặc khu vực thực hiện khảo sát |
| 2 | year | Năm thực hiện khảo sát |
| 3 | gender | Giới tính thực hiện khảo sát |
| 4 | age\_group | Nhóm tuổi thực hiện khảo sát |
| 5 | bully\_group | Đối tượng (hoặc nhóm đối tượng) thực hiện hành vi bạo lực |
| 6 | violence\_type | Loại bạo lực |

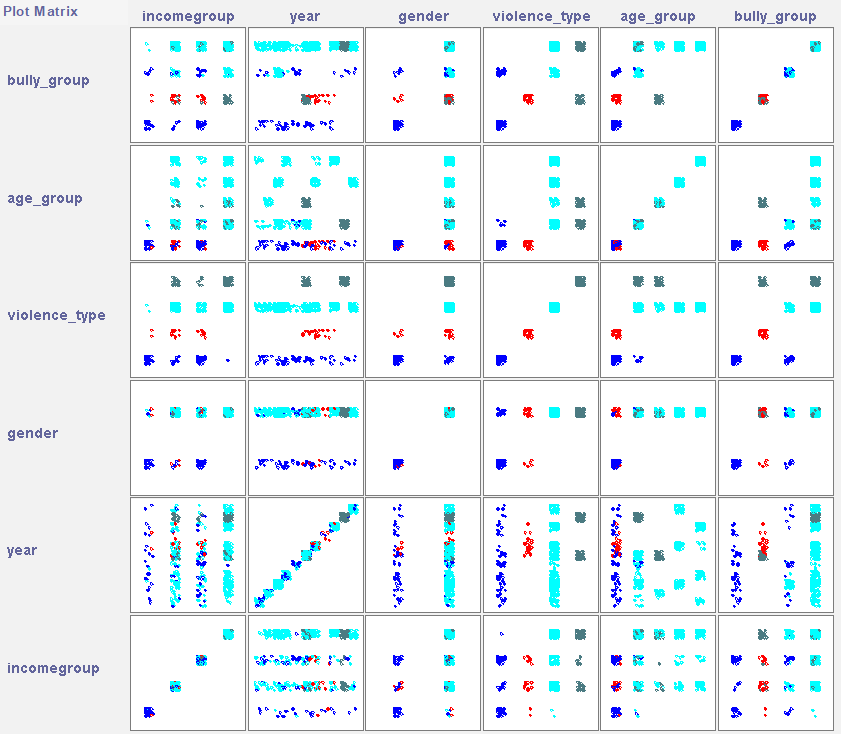
Dữ liệu sau khi tiền xử lý:



Thống kê chi tiết các thuộc tính:

****

Tương quan dữ liệu giữa các thuộc tính:

****

Nhận xét:

* Có sự phân tách giữa *gender* (giới tính) với *violence\_type* (loại bạo lực) và *bully\_group* (nhóm thực hiện bạo lực).
* *violence\_type* có sự phân tách với *bully\_group*, cho thấy mối liên hệ giữa loại bạo lực và các nhóm thực hiện hành vi bạo lực.
* *age\_group* có mối quan hệ rõ ràng với *violence\_type* và *bully\_group*, cho thấy nhóm tuổi là một yếu tố quan trọng.
* *incomegroup* không có sự phân cụm rõ ràng, nhưng vẫn có một số ảnh hưởng đến các yếu tố khác.
* *year* có biến động trong *violence\_type* và *bully\_group*, thể hiện sự thay đổi theo thời gian.

⇒ Các yếu tố như *gender, age\_group* và *violence\_type* có mối tương quan chặt chẽ, trong khi *incomegroup* và *year* có ảnh hưởng yếu hơn.

# PHẦN III. PHÂN LỚP DỮ LIỆU

## Phân lớp dữ liệu

Phân lớp dữ liệu là kĩ thuật dựa trên tập huấn luyện, những giá trị hay nhãn dán của lớp trong một thuộc tính phân lớp và sử dụng nó trong việc phân lớp dữ liệu mới.

Phân lớp là một hình thức học được giám sát: dữ liệu mới được phân lớp dựa trên tập huấn luyện.

Quá trình phân lớp dữ liệu gồm hai bước:

* Bước 1: Xây dựng mô hình
* Bước 2: Sử dụng mô hình

## Phương pháp chia theo tỉ lệ

Phân chia dữ liệu huấn luyện và kiểm thử trong phân lớp giúp đánh giá hiệu suất của mô hình trên tập dữ liệu mới. Dữ liệu huấn luyện (data training) được sử dụng để huấn luyện mô hình và dữ liệu kiểm thử (data test) được sử dụng để đánh giá hiệu suất của mô hình trên tập dữ liệu mới.

Phân chia dữ liệu huấn luyện và kiểm thử cũng giúp tránh tình trạng overfitting (giảm thiểu hạn chế của ID3).

Trong phương pháp này, ta thực hiện chia dữ liệu ngẫu nhiên và phải đảm bảo ràng buộc là tỉ lệ dữ liệu của các lớp trong cả tập dữ liệu huấn luyện và kiểm thử giống nhau.

**Thực hiện:**

* Sử dụng Weka, ta trộn dữ liệu ngẫu nhiên:

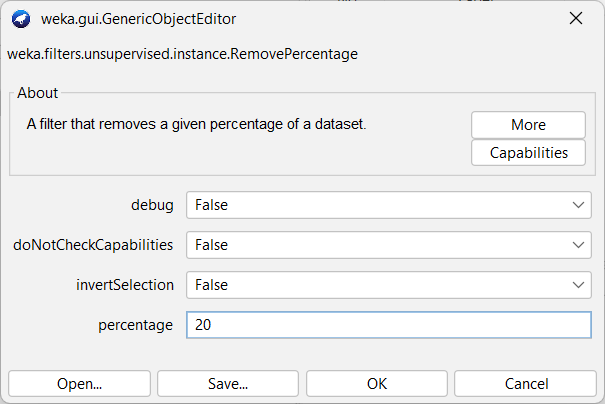
Filter → Unsupervised → Instance → Randomize

* Sử dụng RemovePercentage để chia tập dữ liệu huấn luyện và kiểm thử:

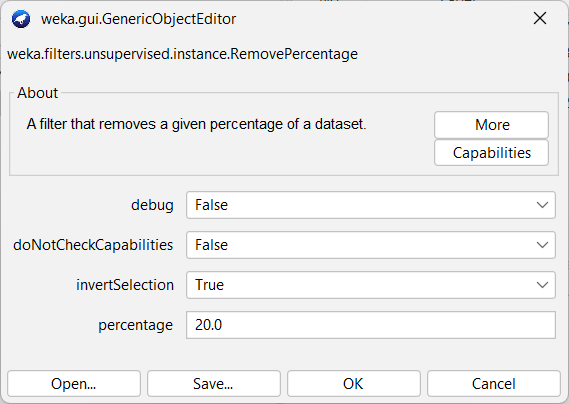
Filter → Unsupervised → Instance → RemovePercentage

Ở đây ta chia tập kiểm thử là 20% và dữ liệu tập huấn luyện là 80%.

* Dữ liệu tập huấn luyện:

****

* Dữ liệu tập kiểm thử:



Kết quả: Sau khi chọn tỉ lệ, ta xáo trộn dữ liệu thu được tập dữ liệu huấn luyện và tập dữ liệu kiểm thử với tỉ lệ 80/20.

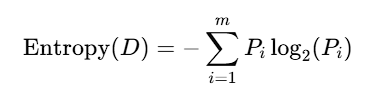
## Thuật toán ID3

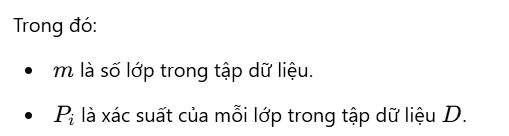
### Lý thuyết

ID3 (Iterative Dichotomiser 3) là một thuật toán học có giám sát (supervised learning) dùng để tạo ra cây quyết định. Nó hoạt động bằng cách chọn các thuộc tính tốt nhất để phân chia dữ liệu và xây dựng cây theo các giá trị của thuộc tính đó cho đến khi đạt được lá đại diện cho lớp của dữ liệu.

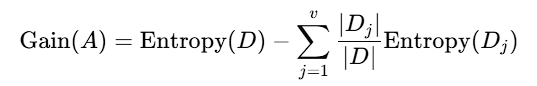
Thuật toán ID3 sử dụng Entropy và độ lợi thông tin (Information Gain) để đánh giá thuộc tính nào tốt nhất để phân chia tập dữ liệu.

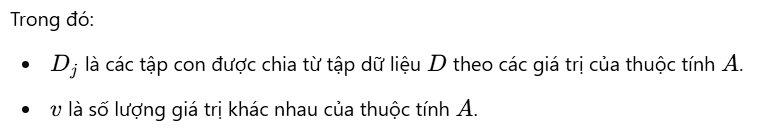
Hàm Entropy dùng để đo tính thuần nhất (không đồng đều) của một tập mẫu dữ liệu. Công thức Entropy cho một tập dữ liệu D là:





Gain là đại lượng dùng để đo độ ưu tiên của một thuộc tính được lựa chọn cho việc phân lớp. Đại lượng này được tính thông qua hai giá trị Information và Entropy. Tại mỗi nút thuộc tính dùng để kiểm tra được chọn dựa vào lượng Information gain lớn nhất.





Thuộc tính có độ lợi thông tin lớn nhất sẽ được chọn để phân chia tập dữ liệu. Điều này giúp giảm độ bất định của dữ liệu và tiến gần hơn đến việc phân loại chính xác.

### Quy trình thực hiện

Bước 1: Xác định thuộc tính chính để phân loại

Bước 2: Tạo một nút trên cây quyết định

Bước 3: Chia tập dữ liệu thành các tập con

Bước 4: Lặp lại các bước trên cho các tập con

Bước 5: Điều kiện dừng

* Tất cả các đối tượng trong tập con thuộc cùng một lớp.
* Không còn thuộc tính nào để chọn (tức là các thuộc tính đã được sử dụng hết).
* Tập dữ liệu quá nhỏ để tiếp tục phân chia.

Bước 6: Gắn nhãn cây với lớp được phổ biến nhất của các đối tượng trong tập dữ liệu đã xét

### Giải thích các thống kê và độ đo theo lớp

#### Các thống kê

* ***Correctly Classified Instances:*** Số lượng các mẫu trong tập dữ liệu được phân loại chính xác bởi mô hình
* ***Incorrectly Classified Instances:*** Số lượng các mẫu trong tập dữ liệu bị phân loại sai bởi mô hình
* ***Kappa statistic*** (Thống kê Kappa): Đây là một chỉ số đo lường sự đồng ý giữa nhãn lớp được dự đoán và thực tế, cho biết rằng mô hình này dự đoán rất hiệu quả so với một mô hình dự đoán ngẫu nhiên hay không
* ***Mean absolute error*** (Sai số trung bình tuyệt đối): Sai số trung bình giữa xác suất dự đoán của mô hình và xác suất thực tế tương ứng với mỗi mẫu
* ***Root mean squared error*** (Căn bậc hai của sai số trung bình bình phương): là căn bậc hai của sai số trung bình bình phương giữa xác suất dự đoán của mô hình và xác suất thực tế tương ứng với mỗi mẫu
* ***Relative absolute error*** (Sai số tuyệt đối tương đối): Đây là sai số trung bình giữa xác suất dự đoán của mô hình và xác suất thực tế, được chuẩn hóa bởi xác suất thực tế trung bình.
* ***Root relative squared error*** (Căn bậc hai của sai số bình phương tương đối): Đây là căn bậc hai của sai số trung bình bình phương giữa xác suất dự đoán của mô hình và xác suất thực tế, được chuẩn hóa bởi xác suất thực tế trung bình
* ***Total Number of Instances***: Tổng số lượng mẫu trong tập dữ liệu

#### Các độ đo theo lớp

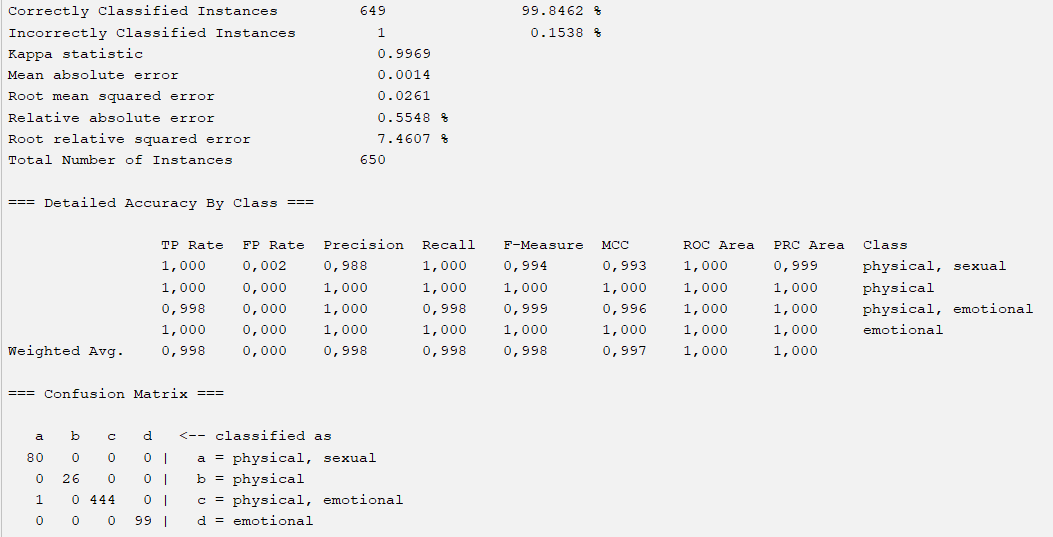
* ***TP Rate:*** Tỷ lệ số lượng các mẫu được phân loại chính xác vào nhãn Positive (dương tính) so với tổng số mẫu Positive trong tập dữ liệu
* ***FP Rate***: Tỷ lệ số lượng các mẫu bị phân loại sai vào nhãn Positive so với tổng số mẫu Negative trong tập dữ liệu.
* ***Precision***: Tỷ lệ số lượng các mẫu được phân loại chính xác vào nhãn Positive so với tổng số các mẫu được phân loại vào nhãn Positive.
* ***Recall***: Tỷ lệ số lượng các mẫu được phân loại chính xác vào nhãn Positive so với tổng số mẫu Positive trong tập dữ liệu.
* ***F-Measure***: Kết hợp giữa Precision và Recall để đánh giá hiệu quả phân loại. F-Measure càng lớn thì phân loại càng chính xác.
* ***MCC***: Độ đo tính tương đồng của hai chuỗi số. MCC = 1 tương đương với việc phân loại hoàn hảo và MCC = -1 tương đương với việc phân loại hoàn toàn ngược lại.
* ***ROC Area:*** Đường cong ROC được sử dụng để đánh giá hiệu quả của thuật toán phân loại trong bài toán dự đoán nhị phân. ROC Area là diện tích dưới đường cong ROC.
* ***PRC Area:*** Đường cong Precision-Recall được sử dụng để đánh giá hiệu quả của thuật toán phân loại trong bài toán dự đoán nhị phân. PRC Area là diện tích dưới đường cong Precision-Recall

### Kết quả, nhận xét

#### Use training test

Sử dụng chính tập dữ liệu huấn luyện để kiểm nghiệm

##### Kết quả thu được



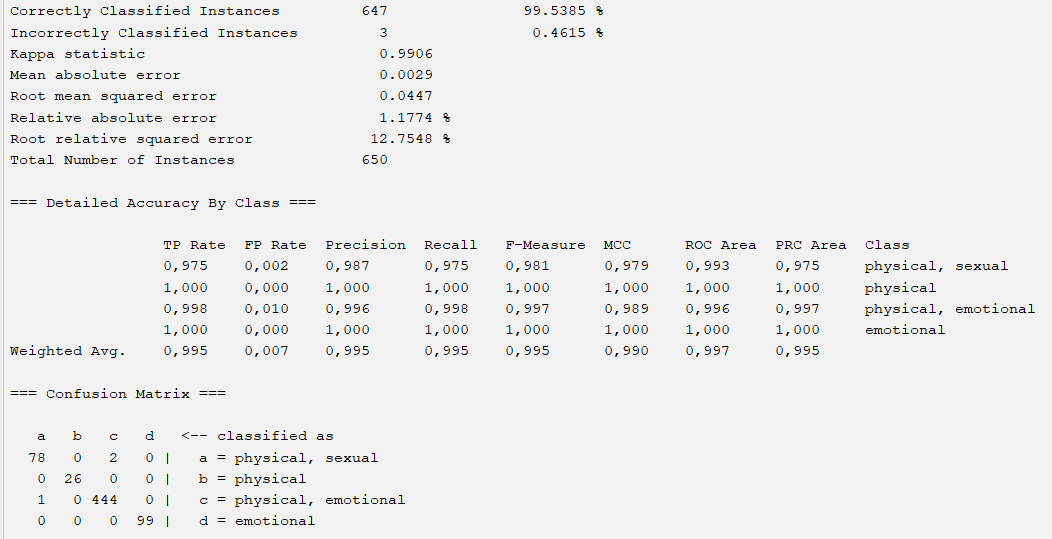
##### Nhận xét

* Thống kê:
* Correctly Classified Instances: Độ chính xác rất cao, 99.85%
* Incorrectly Classified Instances: Chỉ có 1 mẫu bị phân loại sai, chiếm 0.1538%
* Kappa statistic (Thống kê Kappa): Đạt 0.9969, cho thấy sự phù hợp cao giữa dự đoán của mô hình và kết quả thực tế.
* Mean absolute error (Sai số trung bình tuyệt đối): 0.0014, cho thấy sự khác biệt trung bình giữa dự đoán và giá trị thực là rất nhỏ
* Root mean squared error (Căn bậc hai của sai số trung bình bình phương): 0.0261, một sai số bình phương trung bình căn bậc hai rất thấp, cho thấy mức độ chính xác rất cao.
* Relative absolute error (Sai số tuyệt đối tương đối): 0.5548%, tỷ lệ sai số tương đối so với sai số dự đoán trong tập dữ liệu là cực kỳ nhỏ.
* Root relative squared error (Căn bậc hai của sai số bình phương tương đối): 7.4607%, con số này cũng rất thấp, chỉ ra rằng mô hình dự đoán chính xác cao.
* Chi tiết theo lớp:
* TP Rate: Tất cả các lớp đều có tỷ lệ TP rất cao (0.998 đến 1.000), điều này chỉ ra rằng mô hình nhận dạng rất tốt các mẫu thuộc từng lớp cụ thể.
* FP Rate: Rất thấp (từ 0.000 đến 0.002), tức là mô hình hiếm khi phân loại nhầm giữa các lớp.
* Precision: Tất cả các lớp có độ chính xác từ 0.988 đến 1.000, điều này cho thấy mô hình rất ít khi dự đoán sai khi đưa ra kết quả dương tính.
* Recall: Đều ở mức cao (từ 0.998 đến 1.000), tức là mô hình gần như không bỏ sót các trường hợp cần phát hiện.
* F-Measure: Giá trị F-Measure cao (từ 0.994 đến 1.000) cho thấy sự cân bằng tốt giữa precision và recall cho từng lớp.
* MCC: Từ 0.993 đến 1.000, chỉ ra rằng mô hình có hiệu quả cao trong việc phân loại.
* Ma trận nhầm lẫn: Hầu hết các mẫu đều được phân loại chính xác theo từng lớp. Duy nhất có một lỗi phân loại giữa lớp "physical" (bạo lực vật lý) và "physical, emotional" (bạo lực vật lý, cảm xúc), điều này có thể là do hai lớp này có đặc điểm tương đồng.

#### Cross-validation

Chia dữ liệu thành nhiều phần (Folds) để thực hiện nhiều lần đánh giá kết quả.

##### Kết quả thu được



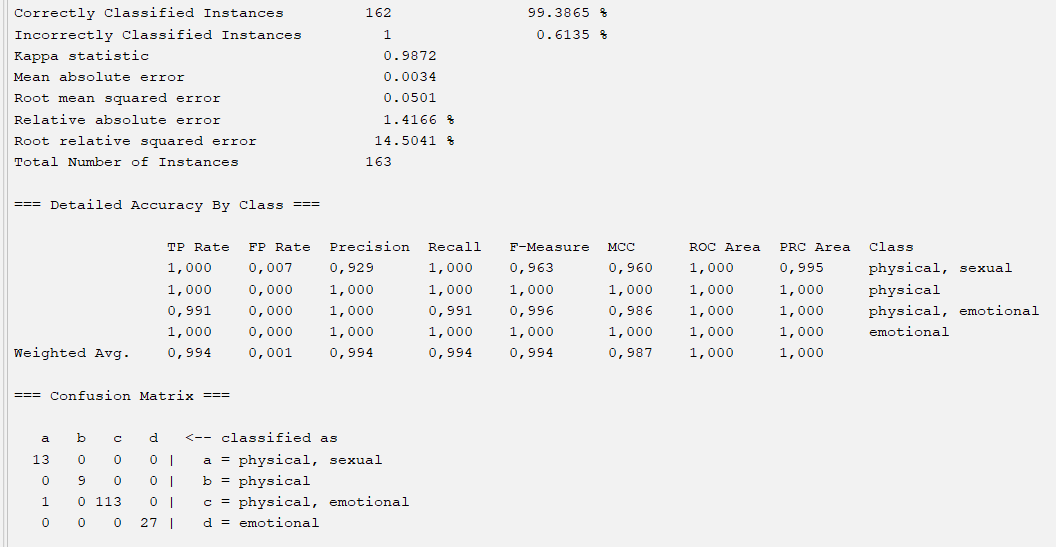
##### Nhận xét

* Theo thống kê:
* Correctly Classified Instances: mô hình cho ra độ chính xác cao: 99,54%
* Incorrectly Classified Instances: Chỉ có 3 mẫu bị phân loại sai
* Kappa statistic (Thống kê Kappa): Chỉ số Kappa rất cao là 0,9906 cho thấy sự hiệu quả với mô hình dự đoán ngẫu nhiên
* Mean absolute error (Sai số trung bình tuyệt đối): 0.0029, cho thấy sự khác biệt trung bình giữa dự đoán và giá trị thực là rất nhỏ
* Root mean squared error (Căn bậc hai của sai số trung bình bình phương) là 0.0447, thể hiện độ lệch bình phương giữa dự đoán và thực tế cũng khá thấp.
* Relative absolute error (Sai số tuyệt đối tương đối): 1.1774%, thể hiện sai số thấp giữa xác suất dự đoán của mô hình và thực tế
* Root relative squared error (Căn bậc hai của sai số bình phương tương đối): 12.7584%, mô hình hoạt động ổn định
* Độ đo theo lớp:
* Lớp "physical, sexual": Tỷ lệ dự đoán đúng (TP Rate) là 0.975, Precision là 0.987, Recall là 0.975 và F-Measure là 0.981, cho thấy mô hình chạy rất tốt với lớp này.
* Lớp "physical": Tỷ lệ TP, Precision, Recall và F-Measure đều là 1.000, cho thấy mô hình không có lỗi phân loại nào trong lớp này.
* Lớp "physical, emotional": Tỷ lệ TP là 0.998, Precision là 0.996, Recall là 0.998, và F-Measure là 0.997, thể hiện độ chính xác gần như 100%.
* Lớp "emotional": Mọi chỉ số đều đạt 1.000, cho thấy mô hình phân loại tuyệt đối chính xác cho lớp này
* Ma trận nhầm lẫn (Confusion Matrix): Ma trận cho thấy một số ít nhầm lẫn giữa các lớp "physical, sexual" và "physical, emotional", với 2 mẫu bị phân loại sai từ lớp "physical, sexual" sang lớp "physical, emotional". Các lớp còn lại được phân loại hoàn toàn chính xác.

#### Supplied test set

Sử dụng tập dữ liệu kiểm thử

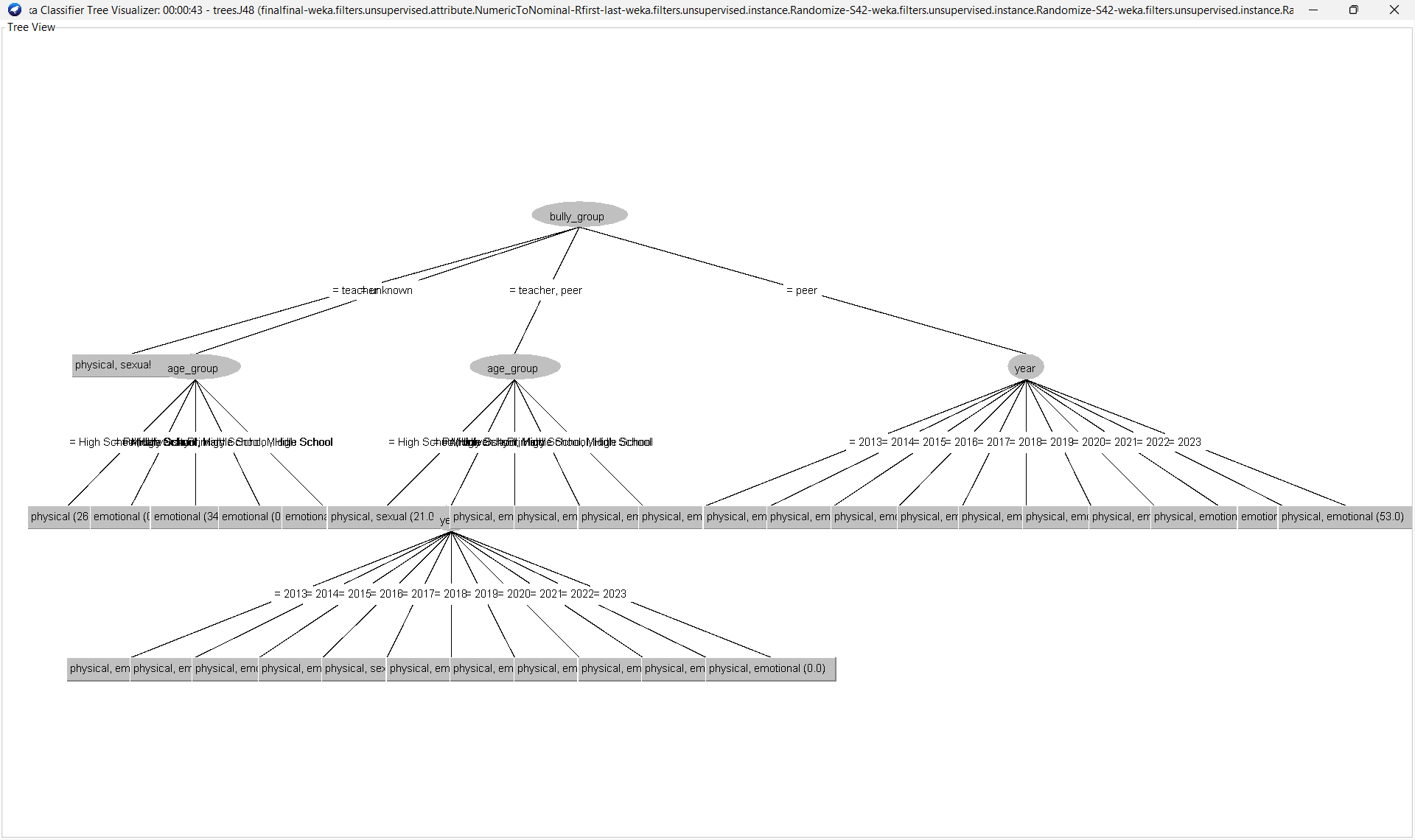
##### Kết quả thu được



##### Nhận xét

* Thống kê:
* Correctly Classified Instances: mô hình cho ra độ chính xác cao 99,34%.
* Incorrectly Classified Instances: Chỉ có 3 mẫu bị phân loại sai
* Kappa statistic (Thống kê Kappa): Chỉ số Kappa rất cao là 0.9872, cho thấy sự hiệu quả với mô hình dự đoán ngẫu nhiên
* Mean absolute error (Sai số trung bình tuyệt đối): 0.0034, cho thấy sự khác biệt trung bình giữa dự đoán và giá trị thực là rất nhỏ
* Root mean squared error (Căn bậc hai của sai số trung bình bình phương): 0.0501, thể hiện độ lệch bình phương giữa dự đoán và thực tế cũng khá thấp.
* Relative absolute error (Sai số tuyệt đối tương đối): 1.4166%, thể hiện sai số thấp giữa xác suất dự đoán của mô hình và thực tế
* Root relative squared error (Căn bậc hai của sai số bình phương tương đối): 14.5041%
* Các độ đo theo lớp:
* Lớp "physical, sexual": Precision là 0.929, Recall 1.0, F-measure 0.963, với FP Rate rất thấp là 0.007. Điều này cho thấy mô hình có thể phân loại đúng nhưng có một số nhầm lẫn nhỏ với các lớp khác.
* Lớp "physical" và "emotional": Precision, Recall, F-Measure đều đạt 1.0, cho thấy mô hình này phân loại cực kỳ chính xác cho hai lớp này.
* Lớp "physical, emotional": Precision và F-measure vẫn cao, nhưng Recall giảm nhẹ xuống 0.991.
* Ma trận nhầm lẫn: Có duy nhất một trường hợp bị phân loại nhầm. Điều này thể hiện rõ trong ma trận nhầm lẫn ở lớp “physical, emotional”, khi 1 mẫu được phân loại nhầm vào lớp “physical, sexual”. Không có bất kỳ lỗi phân loại nào giữa các lớp “physical” và “emotional”.

#### Cây quyết định



### Nhận xét

Mô hình phân loại bắt đầu với nút gốc là *bully\_group,* cho thấy độ ảnh hưởng đối tượng gây ra bạo lực là quan trọng nhất, tiếp đến là *age\_group* và *year* theo từng nhánh. Thuộc tính *incomegroup* (nhóm thu nhập) và *gender* (giới tính) không xuất hiện trong cây.

⇒ Điều này cho thấy các yếu tố độ tuổi và người gây ra bạo lực có vai trò quan trọng trong việc dự đoán loại bạo lực.

Bên cạnh đó, *year* (năm khảo sát) cũng ảnh hưởng tới các nút lá, điều này có ý nghĩa về sự thay đổi xu hướng bạo lực theo thời gian.

### Đánh giá

Đối với bài toán phân tích mối quan hệ giữa các yếu tố như giới tính, độ tuổi và loại bạo lực để xác định những yếu tố quan trọng nhất dẫn tới bạo lực học đường, việc sử dụng thuật toán ID3 là một lựa chọn hợp lý. Bởi vì dữ liệu trong bài toán này có cấu trúc rõ ràng và có mối liên kết logic giữa các thuộc tính như giới tính, độ tuổi và các hình thức bạo lực. ID3 giúp xây dựng một cây quyết định có thể dễ hiểu và trực quan, dễ dàng giải thích quá trình phân loại và đưa ra các yếu tố quan trọng ảnh hưởng đến hành vi bạo lực.

Tuy nhiên, tương tự như các thuật toán khác, ID3 cũng có những hạn chế. Một trong những điểm yếu chính của thuật toán này là khi làm việc với tập dữ liệu lớn và phức tạp, nó dễ gặp hiện tượng quá khớp (overfitting). Ngoài ra, với những dữ liệu có nhiều thuộc tính và sự tương tác phức tạp giữa các yếu tố, ID3 có thể bỏ sót một số quy tắc quan trọng.

Tóm lại, việc sử dụng ID3 cho bài toán phân tích bạo lực học đường là một lựa chọn phù hợp. Tuy nhiên, cần lưu ý xử lý hiện tượng quá khớp (overfitting) và cân nhắc sử dụng các biến thể của thuật toán để đạt được kết quả chính xác và khả năng tổng quát hóa tốt nhất.

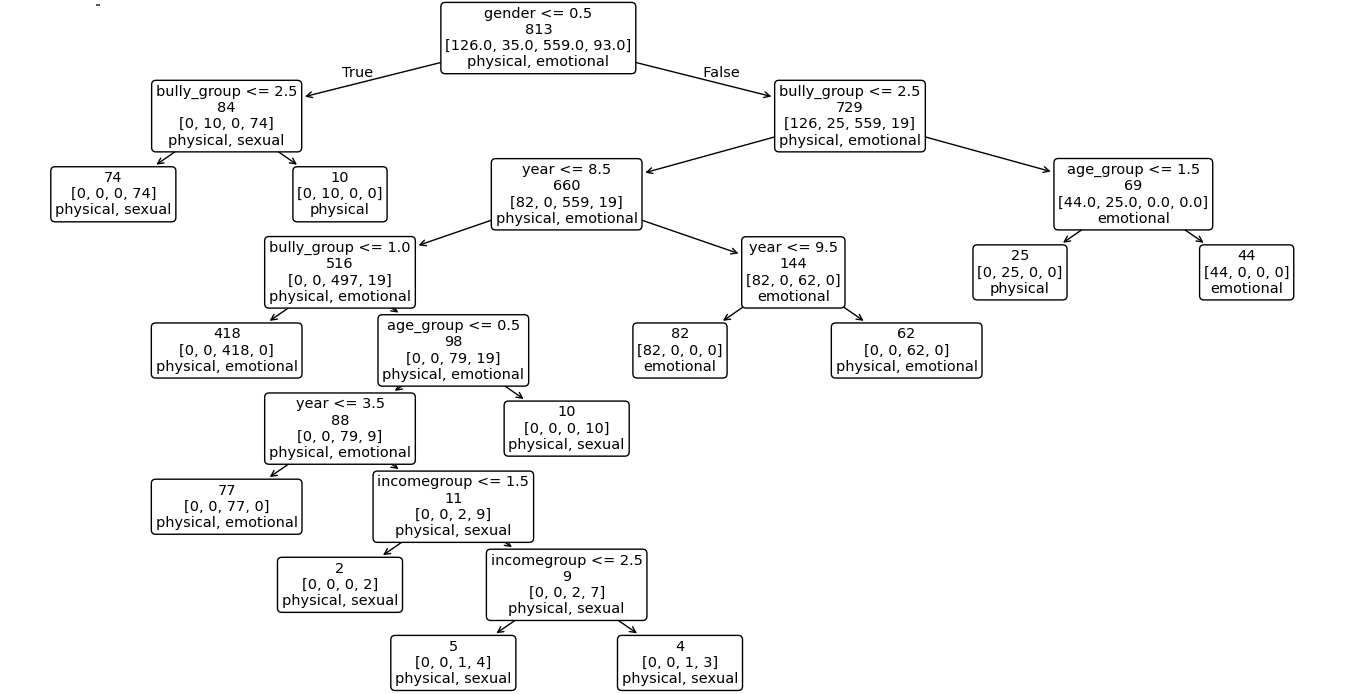
# 

# PHẦN IV. TRIỂN KHAI THUẬT TOÁN

## Triển khai thuật toán (Python)



* Kết quả cây quyết định:



* Khi sử dụng LabelEncoder để mã hóa các giá trị chuỗi thành số, chúng sẽ được chuyển đổi như sau:
* *income\_group*:

'Low income' → 0

'Lower middle income' → 1

'Upper middle income' → 2

'High income' → 3

* *year*:

2013 → 0

2014 → 1

2015 → 2

2016 → 3

2017 → 4

2018 → 5

2019 → 6

2020 → 7

2021 → 8

2022 → 9

2023 → 10

* *gender*:

0 → 0 (female)

1 → 1 (male)

* *violence\_type*:

'physical, sexual' → 0

'physical' → 1

'physical, emotional' → 2

'emotional' → 3

* *age\_group*:

'Primary School, Middle School' → 0

'Primary School, Middle School, High School' → 1

'Middle School, High School' → 2

'High School' → 3

'High School, University' → 4

* *bully\_group*:

'teacher' → 0

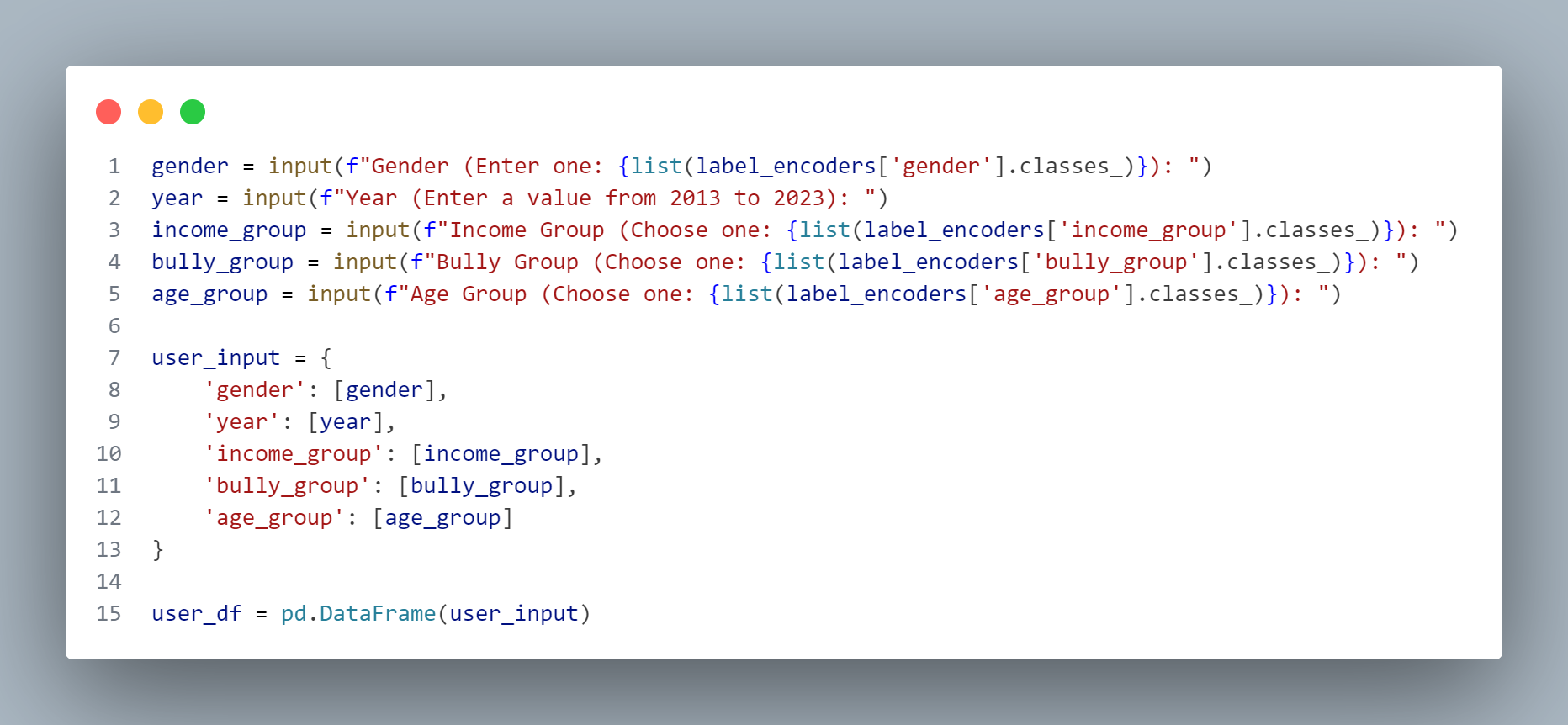
'unknown' → 1

'teacher, peer' → 2

'peer' → 3

## Sử dụng mô hình để dự đoán kết quả

* Nhập Gender
* Nhập Year
* Nhập Income Group
* Nhập Bully Group
* Nhập Age Group

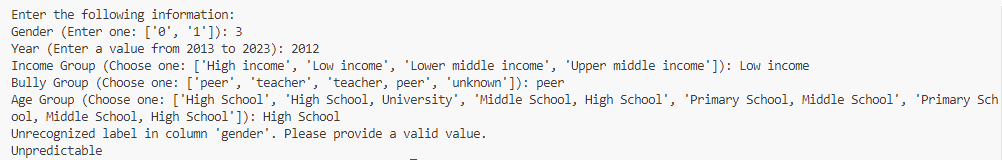


* Sau khi nhập dữ liệu đầu vào tương ứng ở trên , mô hình sẽ sử dụng hàm dự đoán trong code (xem chi tiết ở file đính kèm) để cho ra kết quả:

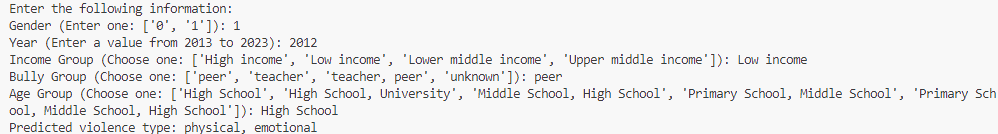
|  |  |
| --- | --- |
| **Nhãn** | **Output** |
| physical, sexual | Đây là nhãn chỉ ra rằng hành vi bạo lực bao gồm cả **bạo lực thể chất** và **bạo lực tình dục** |
| physical | Nhãn này chỉ ra hành vi **bạo lực thể chất** |
| physical, emotional | Nhãn này mô tả hành vi bao gồm cả **bạo lực thể chất** và **bạo lực tinh thần** |
| emotional | Nhãn này chỉ ra hành vi bạo lực là **bạo lực tinh thần** |

* Trường hợp dữ liệu không nằm trong phạm vi dự đoán:

Kết quả sẽ trả về: Unpredictable



* Thực hành chạy code dự đoán:



# KẾT LUẬN

Sau khi hoàn thành đề tài *"Khai phá dữ liệu về bạo lực học đường",* nhóm chúng em đã đạt được một số kết quả như sau:

* Tổng quan: Nhóm đã tìm hiểu các khái niệm như tiền xử lý, phương pháp phân lớp và đặc biệt là thuật toán ID3 để xây dựng cây quyết định, qua đó tìm hiểu được mối liên hệ giữa các yếu tố dẫn tới bạo lực học đường.
* Nhóm đã thu thập và xử lý dữ liệu liên quan đến bạo lực học đường bằng các công cụ Weka và Python. Quá trình tiền xử lý giúp làm sạch và chuẩn hóa dữ liệu để đảm bảo tính chính xác cho mô hình.
* Nhóm đã xây dựng thành công mô hình phân lớp bằng thuật toán ID3 trên phần mềm Weka và Python.

Mặc dù nhóm đã cố gắng hoàn thiện đề tài, nhưng vẫn còn những thiếu sót và hạn chế trong quá trình thực hiện. Chúng em mong muốn nhận được sự góp ý và chỉ dẫn từ các thầy cô để cải thiện kỹ năng, kiến thức và có thể ứng dụng tốt hơn các phương pháp khai phá dữ liệu trong các dự án thực tế trong tương lai.

# 

# TÀI LIỆU THAM KHẢO

1. Dữ liệu thô: [Introducing a New Dataset of Datasets: Where, When, and How Much Data Exists on School Violence | Center For Global Development (cgdev.org)](https://www.cgdev.org/blog/introducing-new-dataset-datasets-where-when-and-how-much-data-exists-school-violence)
2. Trần Mạnh Tuấn, Slide bải giảng Khai phá dữ liệu
3. Trần Mạnh Tuấn, Hoàng Thị Minh Châu, Trần Thanh Đại, Vũ Mỹ Hạnh, Vũ Anh Tuấn, 2022, Giáo trình khai phá dữ liệu, Trường Đại học Kinh tế - Kỹ thuật Công nghiệp.