TRƯỜNG ĐẠI HỌC THỦY LỢI

**KHOA CÔNG NGHỆ THÔNG TIN**

A blue and white logo

Description automatically generated

**BÀI TẬP LỚN**

HỌC PHẦN: KHAI PHÁ DỮ LIỆU

**ĐỀ TÀI:**

**KHAI PHÁ DỮ LIỆU VỀ CÁC YẾU TỐ ẢNH HƯỞNG ĐẾN HIỆU SUẤT HỌC TẬP CỦA HỌC SINH BẰNG PHƯƠNG PHÁP PHÂN LỚP**

**Giảng viên hướng dẫn: Trần Mạnh Tuấn**

Nhóm sinh viên thực hiện:

1. Nguyễn Tiến Hưng, 2151062793

2. Nguyễn Thị Diệu Minh, 2151061174

3. Ngô Thị Út Thương, 2151062880

4. Triệu Minh Thảo, 2151062871

**Hà Nội, năm 2024**

MỤC LỤC

[**LỜI CẢM ƠN** 3](#_Toc181607427)

[**I.** **Mô tả bài toán.** 4](#_Toc181607428)

[**1.** **Lý do chọn đề tài** 4](#_Toc181607429)

[**2.** **Quy trình khai phá tri thức** 4](#_Toc181607430)

[**3.** **Mô tả dữ liệu ban đầu** 4](#_Toc181607431)

[**II.** **Quy trình khai phá dữ liệu** 6](#_Toc181607432)

[**1.** **Tiền xử lý dữ liệu** 6](#_Toc181607433)

[**1.1.** **Làm sạch dữ liệu** 6](#_Toc181607434)

[**1.2.** **Tích hợp dữ liệu** 9](#_Toc181607435)

[**1.3.** **Biến đổi dữ liệu** 9](#_Toc181607436)

[**2.** **Phân tích dữ liệu** 14](#_Toc181607437)

[**3.** **Phân tách dữ liệu Train, Test.** 17](#_Toc181607438)

[**III.** **Phương pháp phân lớp và thuật toán phân lớp** 18](#_Toc181607439)

[**1.** **Bài toán phân lớp** 18](#_Toc181607440)

[**2.** **Thuật toán ID3** 18](#_Toc181607441)

[**2.1. Lý thuyết** 18](#_Toc181607442)

[**2.2. Quy trình thực hiện** 19](#_Toc181607443)

[**2.3. Thực hiện phân lớp với weka** 19](#_Toc181607444)

[3. **Thuật toán Naïve Bayes** 20](#_Toc181607445)

[3.1. **Giới thiệu thuật toán** 20](#_Toc181607446)

[**3.2.** **Thuật toán** 20](#_Toc181607447)

[**IV.** **Cài đặt thuật toán** 22](#_Toc181607448)

[**1.** **Sử dụng thư viện Scikit-learn để thực hiện huấn luyện mô hình cây quyết định ID3** 22](#_Toc181607449)

[**2.** **Xây dựng thuật toán ID3 không sử dụng thư viện** 24](#_Toc181607450)

[**3.** **Thuật toán Naïve Bayes sử dụng thư viện sklearn** 26](#_Toc181607451)

[**4.** **Xây dựng thuật toán Naïve Bayes không sử dụng thư viện** 28](#_Toc181607452)

[**KẾT LUẬN** 30](#_Toc181607453)

[**TÀI LIỆU THAM KHẢO** 31](#_Toc181607454)

# **LỜI CẢM ƠN**

Ngày nay, việc ứng dụng công nghệ thông tin đã trở nên phổ biến trong hầu hết mọi cơ quan, doanh nghiệp, trường học đặc biệt là việc áp dụng các giải pháp tin học trong công tác quản lý. Trong ít năm trở lại đây, với tốc độ phát triển như vũ bão, CNTT đang dần làm cho cuộc sống của con người trở nên thú vị và đơn giản hơn. Để bắt kịp với nhịp độ phát triển của xã hội, những kiến thức học được trên giảng đường là vô cùng quan trọng đối với mỗi sinh viên chúng em. Vì vậy chúng em chọn đề tài ***“Khai phá dữ liệu về các yếu tố ảnh hưởng đến hiệu suất học tập của học sinh bằng phương pháp phân lớp”*** để làm báo cáo kết thúc môn học của mình. Chúng em chân thành xin gửi lời cảm ơn đặc biệt đến thầy giáo Trần Mạnh Tuấn – người đã tận tình giảng dạy môn Khai phá dữ liệu cho chúng em trong từng buổi học. thầy đã giúp chúng em trang bị kiến thức môn học và hơn cả là động lực tiếp tục trên con đường chinh phục công nghệ. Bên cạnh những kết quả mà chúng em đạt được thì sẽ không khó tránh khỏi những thiếu sót trong quá trình làm đề tài vì thời gian không cho phép và chưa có kinh nghiệm thực tế. Chính vì vậy chúng em rất mong được sự cảm thông, chỉ bảo góp ý của thầy cô. Những lời nhận xét, góp ý của thầy cô chính là một bài học, kiến thức cho chúng em trên con đường sau này. Chúng em xin chân thành cảm ơn

**Bảng phân chia công việc**

|  |  |
| --- | --- |
| **Công việc** | **Người phụ trách** |
| Xử lý dữ liệu | Ngô Thị Út Thương, Nguyễn Thị Diệu Minh |
| Thuật toán Naive Bayes | Ngô Thị Út Thương,Triệu Minh Thảo |
| Thuật toán ID3 | Nguyễn Tiến Hưng, Nguyễn Thị Diệu Minh |

1. **Mô tả bài toán.**
2. **Lý do chọn đề tài**

Trong thời đại phát triển mạnh mẽ của CNTT như hiện nay, việc sử dụng và khai thác dữ liệu càng ngày càng được chú trọng. Đối với lĩnh vực giáo dục thì việc dự đoán kết quả học tập của học sinh từ đó đưa ra các biện pháp tăng tính cá nhân hóa trong việc giảng dạy đối với từng học sinh. Đảm bảo các em học sinh có thể phát triển một cách tốt nhất. Vậy nên, nhóm chúng em quyết định chọn đề tài ***“Khai phá dữ liệu về các yếu tố ảnh hưởng đến hiệu suất học tập của học sinh bằng phương pháp phân lớp”*** để có thể dự đoán được kết quả của các em học sinh trong tương lai, từ đó hỗ trợ các giáo viên có thể đưa ra những sự điều chỉnh giáo dục phù hợp.

1. **Quy trình khai phá tri thức**

A diagram of a flowchart

Description automatically generated

1. **Mô tả dữ liệu ban đầu**

* Nguồn dữ liệu: [Student Performance Factors](https://www.kaggle.com/datasets/lainguyn123/student-performance-factors?fbclid=IwY2xjawGAPa5leHRuA2FlbQIxMAABHRgjr52K0UOX2njXJkavVyr2oTw4chpgL53TU13HBR-mjNbxVHQKxHm_ag_aem_Cz6Yu8OLv5bpJmj04DPLOw)
* Dữ liệu gồm 6680 vector cung cấp thông tin tổng quan về những yếu tố ảnh hưởng đến kết quả thi của học sinh. Bộ dữ liệu bao gồm thông tin về thói quen học tập, sự tham gia của phụ huynh và các khía cạnh khác ảnh hưởng đến thnàh công trong học tập của học sinh:

A screenshot of a computer

Description automatically generated

* Mô tả cột:

|  |  |
| --- | --- |
| Hours\_Studied | Số giờ dành cho việc học mỗi tuần |
| Attendance | Tỷ lệ phần trăm tham gia lớp học |
| Parental\_Involvement | Mức độ tham gia của phụ huynh vào việc học tập của học sinh (Low: Thấp, Medium: Trung bình, High: Cao) |
| Acess\_to\_Resources | Khả năng tiếp cần tài nguyên giáo dục |
| Extracurricular\_Activities | Tham gia các hoạt động ngoại khóa (Yes: Có, No:Không) |
| Sleep\_Hours | Số giờ ngủ trung bình mỗi đêm |
| Previous\_Scores | Điểm của bài kiểm tra trước đó |
| Motivation\_Level | Mức độ động lực của học sinh (Low: thấp, Medium: trung bình, High: cao) |
| Internet\_Access | Khả năng truy cập Internet (Yes: có, No: không) |
| Tutoring\_Sessions | Số buổi học thêm mỗi tháng |
| Family\_Income | Mức thu nhập của gia đình (Low: thấp, Medium: trung bình, High: cao) |
| Teacher\_Quality | Chất lượng giáo viên (Low: thấp, Medium: trung bình, High: cao) |
| School\_Type | Loại trường đang theo học (Public: trường công, Private: trường tư) |
| Peer\_Inflluence | Ảnh hưởng của bạn bè đến kết quả học tập (Positive: tích cực, Neutral: trung bình, Negative: tiêu cực) |
| Physical\_Activity | Số giờ hoạt động thể chất trung bình mỗi tuần |
| Learning\_Disabilities | Có bị mắc những hội chứng rối loạn học tập không (Yes: có, No: không) |
| Parental\_Education\_Level | Trình độ học vấn cao nhất của cha mẹ (High School: Trung học, College: Đại học, Postgraduate: Sau đại học) |
| Distance\_from\_Home | Khoảng cách từ nhà đến trường (Near: gần, Moderate: trung bình, Far: xa) |
| Gender | Giới tính của học sinh (Male: nam, Female: nữ) |
| Exam\_Score | Điểm bài thi cuối kì |

1. **Quy trình khai phá dữ liệu**
2. **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 KPDL

* 1. **Làm sạch dữ liệu**
     1. **Xử lý dữ liệu thiếu**
* Việc xử lý dữ liệu thiếu là một phần quan trọng trong quá trình tiền xử lý dữ liệu. Có rất nhiều nguyên nhân có thể gây ra sự thiếu dữ liệu như: lỗi ghi dữ liệu, vấn đề kỹ thuật trong quá trình thu thập dữ liệu, …
* Đối với tập dữ liệu trên, ở các cột “**Teacher\_Quality**”, “**Parental\_Education\_Level**”, “**Distance\_from\_Home**” có các dữ liệu bị thiếu.

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

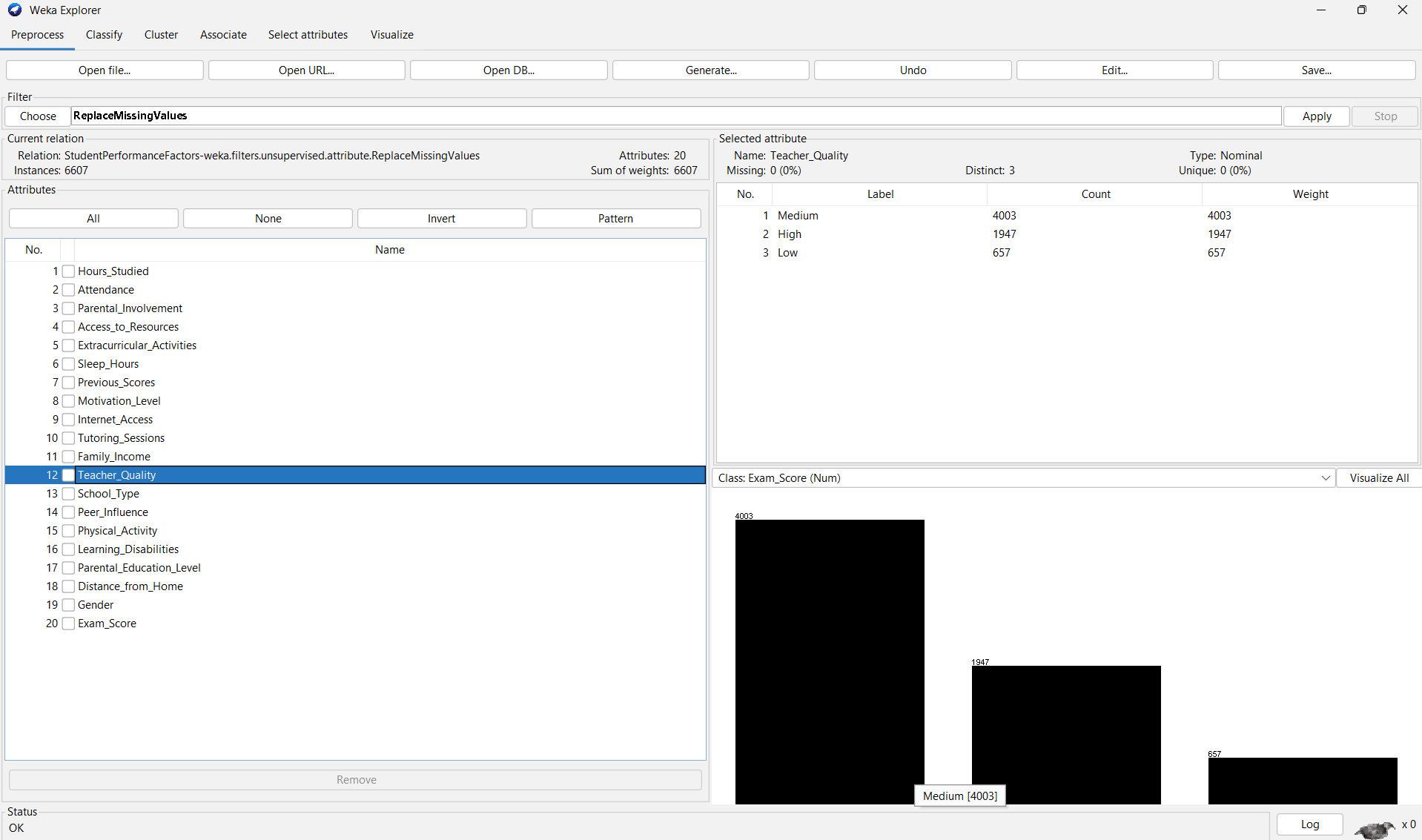
**A screenshot of a computer

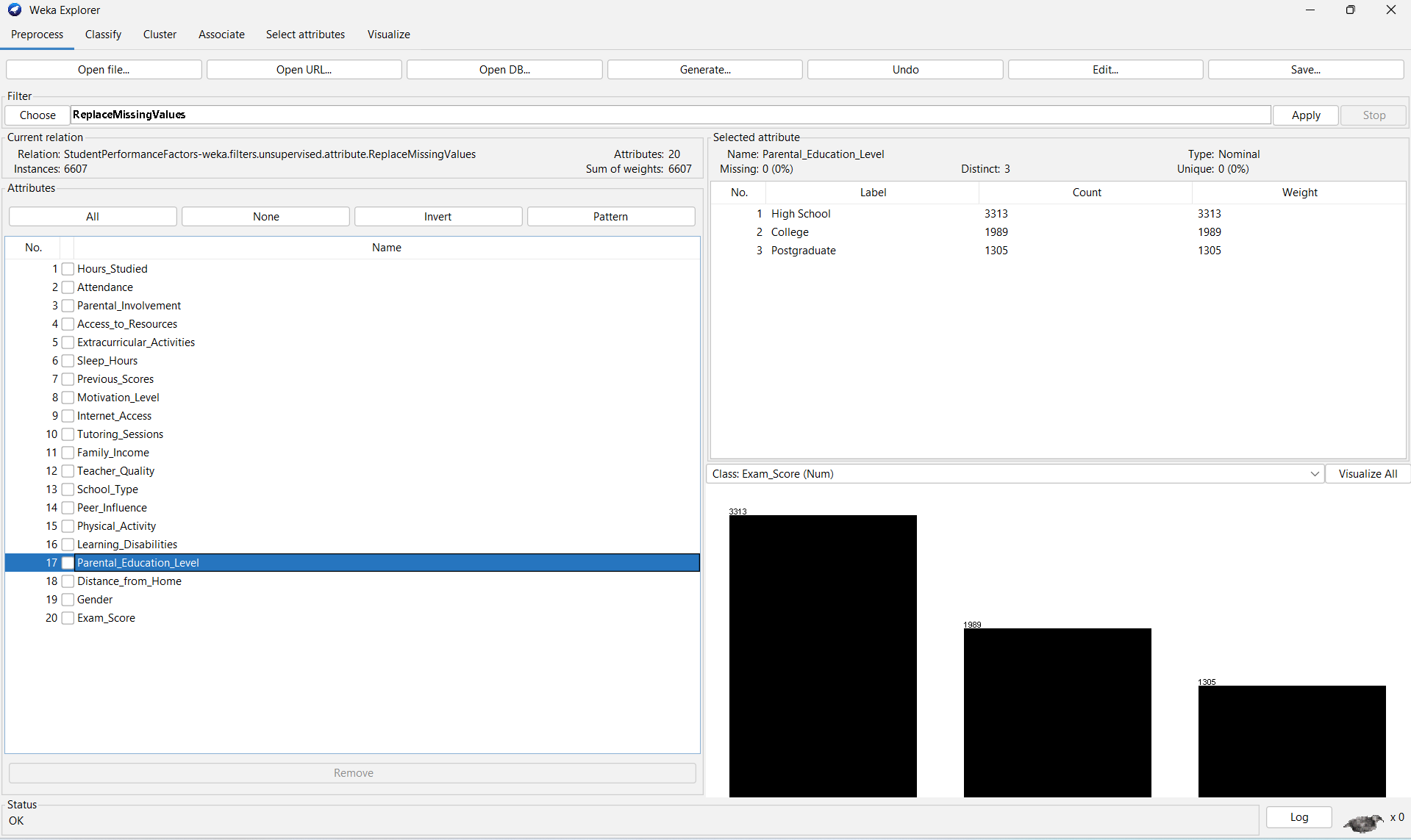
Description automatically generated**

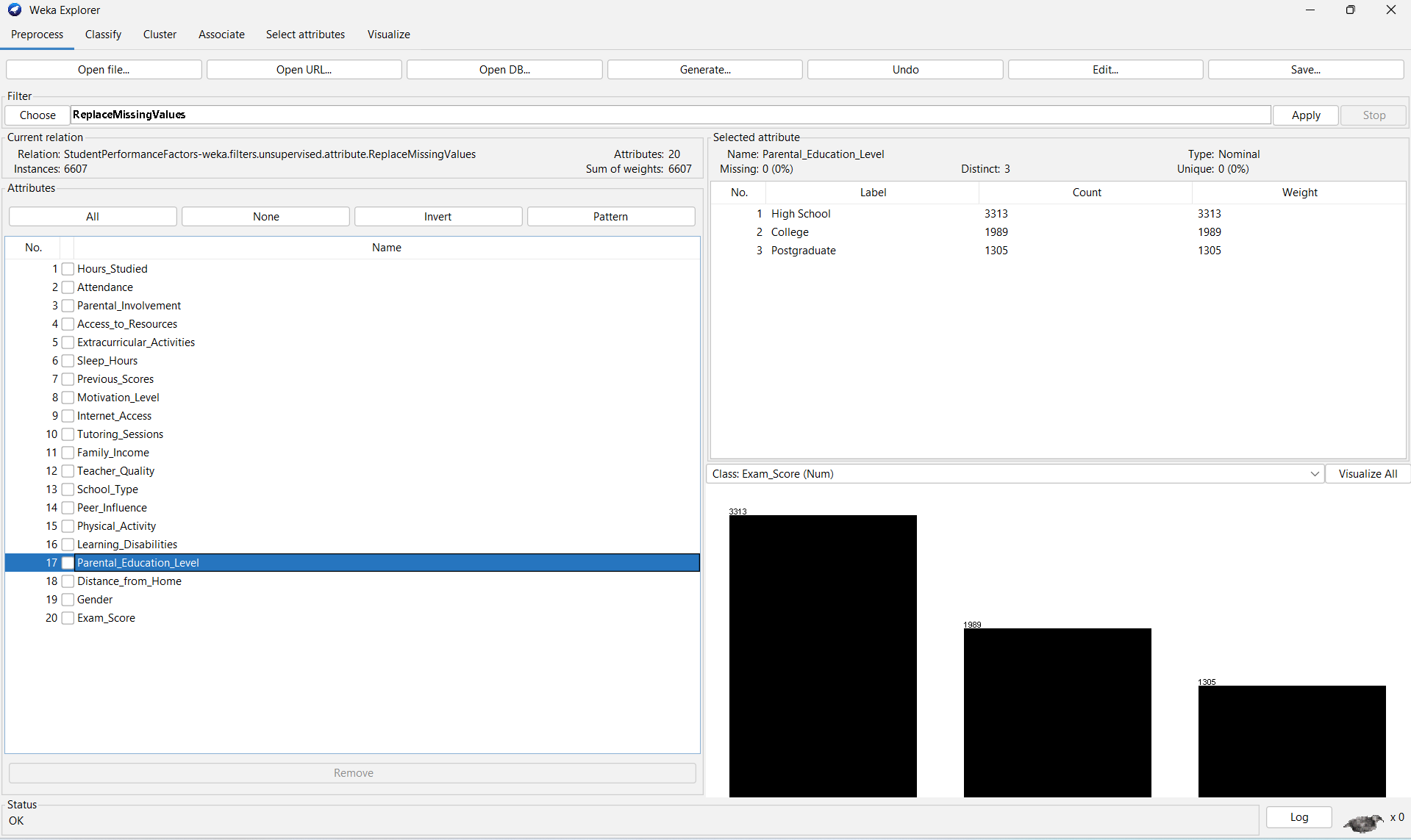
Ta tiến hành xử lý các dữ liệu bị thiếu bằng cách lấy giá trị trung bình của các cột dữ liệu thiếu đó để gán cho ô dữ liệu bị thiếu bằng cách:

***Choose* →  *Weka* → *Filter* → *Unsupervised* →  *ReplaceMissingValues***

* Dữ liệu 3 cột sau khi loại bỏ các dữ liệu bị thiếu:

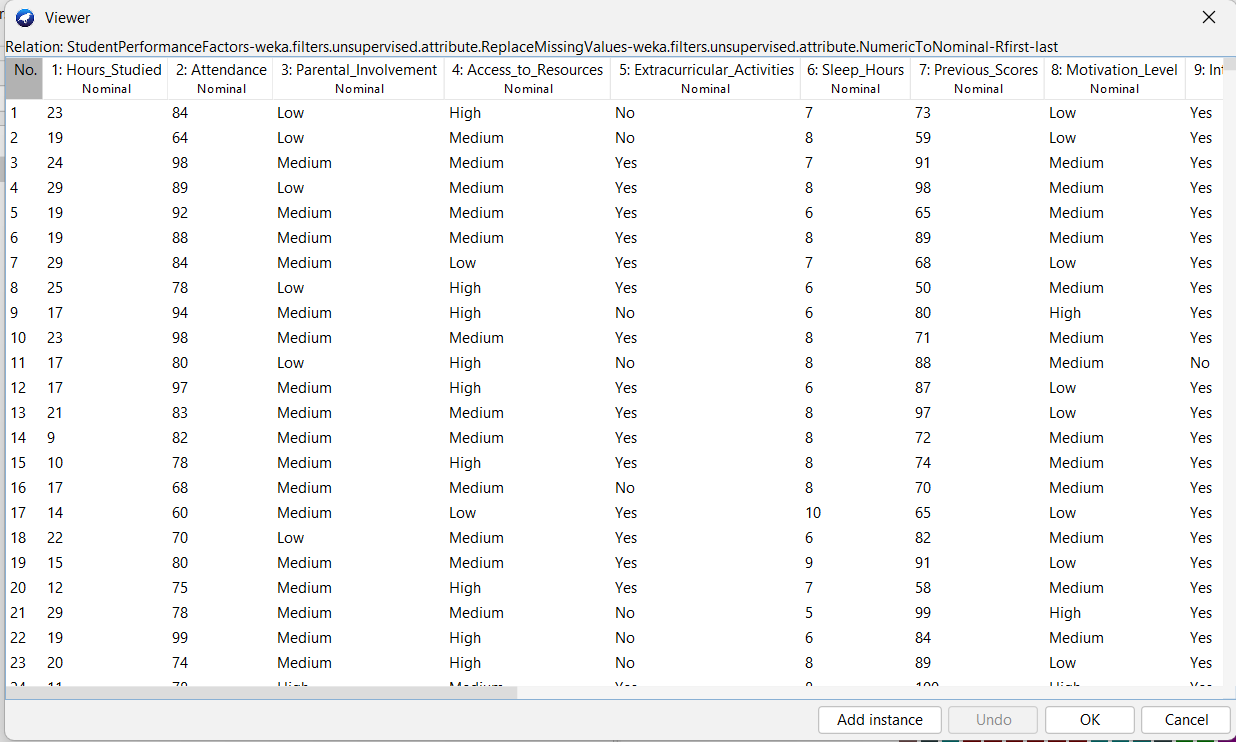






* + 1. **Loại bỏ dữ liệu nhiễu**
* Xử lý dữ liệu nhiễu, không hợp lý có thể ảnh hưởng đến kết quả phân tích:
* Để thuận tiện cho việc xử lý dữ liệu bị nhiễu và phù hợp với mục tiêu khai phá, ta chuyển kiểu dữ liệu của các thuộc tính về dạng nominal bằng cách:

**+ Filter -> Unsupervised -> Attribute -> NumericToNominal -> Apply**



* + 1. **Xử lý dữ liệu không nhất quán**
  1. **Tích hợp dữ liệu**
* Dữ liệu được lấy từ cùng một nguồn vậy nên không xảy ra trường hợp dữ liệu không thuần nhất về mặt ngữ nghĩa. Do đó không cần phải thực hiện bước tích hợp dữ liệu
  1. **Biến đổi dữ liệu**
* Dựa vào mục tiêu ban đầu là thực hiện bài toán phân lớp với tập dữ liệu nên ta sẽ cần chuyển giá trị cột Exam\_Score thành các nhãn điểm A,B,C,D,F. Cụ thể:
  + Điểm < 40 gán nhãn 0 (tương ứng với F)
  + Điềm < 55 gán nhãn 1 (tương ứng với D)
  + Điểm < 70 gán nhãn 2 (tương ứng với C)
  + Điểm < 85 gán nhãn 3 (tương ứng với B)
  + Điềm >85 gán nhãn 4 (tương ứng với A)
* Thực hiện gán nhãn trên Weka cách sử dụng Filter weka.filters.unsupervised.attribute.MathExpression với expression là câu lệnh ifelse:

A screenshot of a computer

Description automatically generated

* Chuyển dữ liệu cột Exam\_Score từ kiểu **Numeric** thành kiểu **Nominal bằng filter: weka.filters.unsupervised.attribute.NumericToNominal**

**A screenshot of a computer

Description automatically generated**

* Dữ liệu cột Exam\_Score sau khi tiến hành chuyển đổi thành các nhãn và rời rạc:

**A screenshot of a computer screen

Description automatically generated**

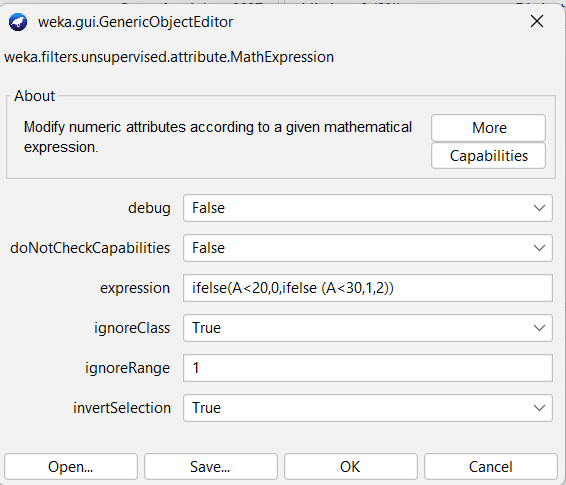
* Thực hiện tương tự với thuộc tính Previos\_Score ta thu được dữ liệu cột Previos\_Score sau khi nhóm

A screenshot of a graph

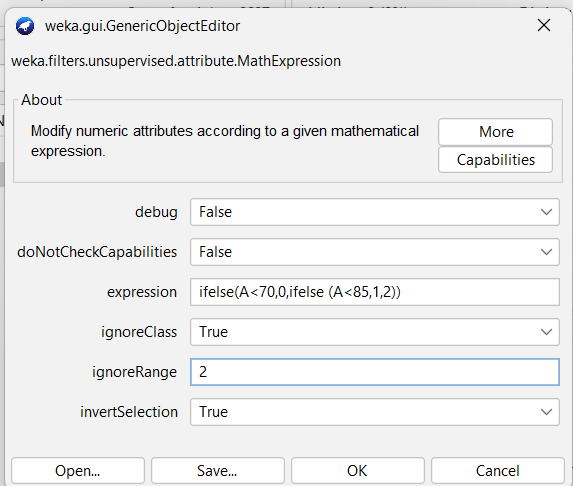
Description automatically generated

* Đối với thuộc tính Hour\_Studies ta cũng sẽ chia các giá trị vào 3 nhóm:
  + Low (0 – 20): thay thế là 0
  + Medium (20 – 30): thay thế là 1
  + High ( > 30): thay thế là 2
* A screenshot of a graph

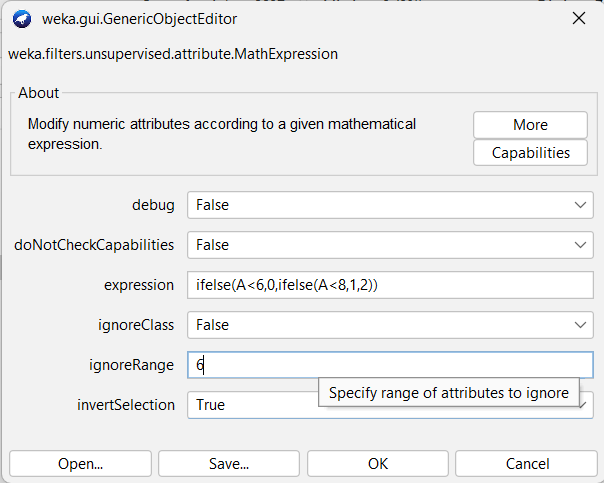
  Description automatically generatedThuộc tính Hours\_Studied sau khi nhóm:



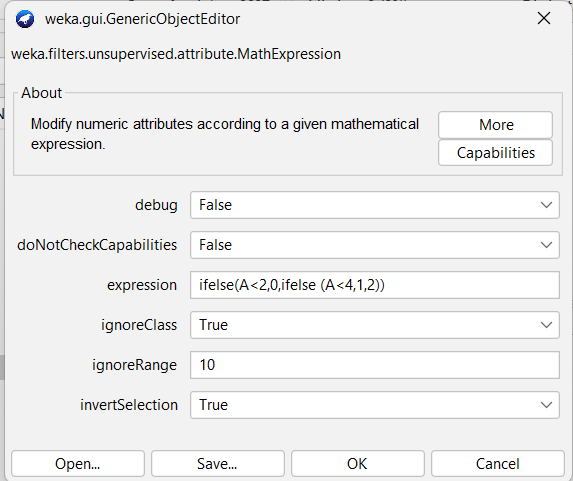
* A screenshot of a graph

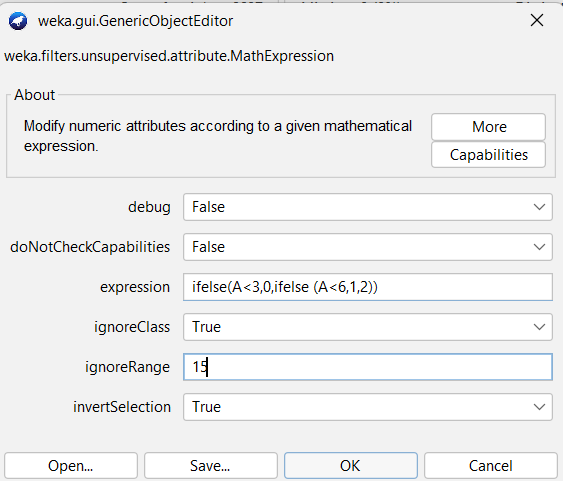
  Description automatically generatedChia dữ liệu thuộc tính Attendance vào 3 nhóm: 0 - Low ( 0 - 70), 1 – Medium (70 – 85), 2 – High (85 - 100)
* Chia dữ liệu thuộc tính Sleep\_Hours vào 3 nhóm: 0 - Low ( 0 - 6), 1 – Medium (6 – 8), 2 – High ( >8 )

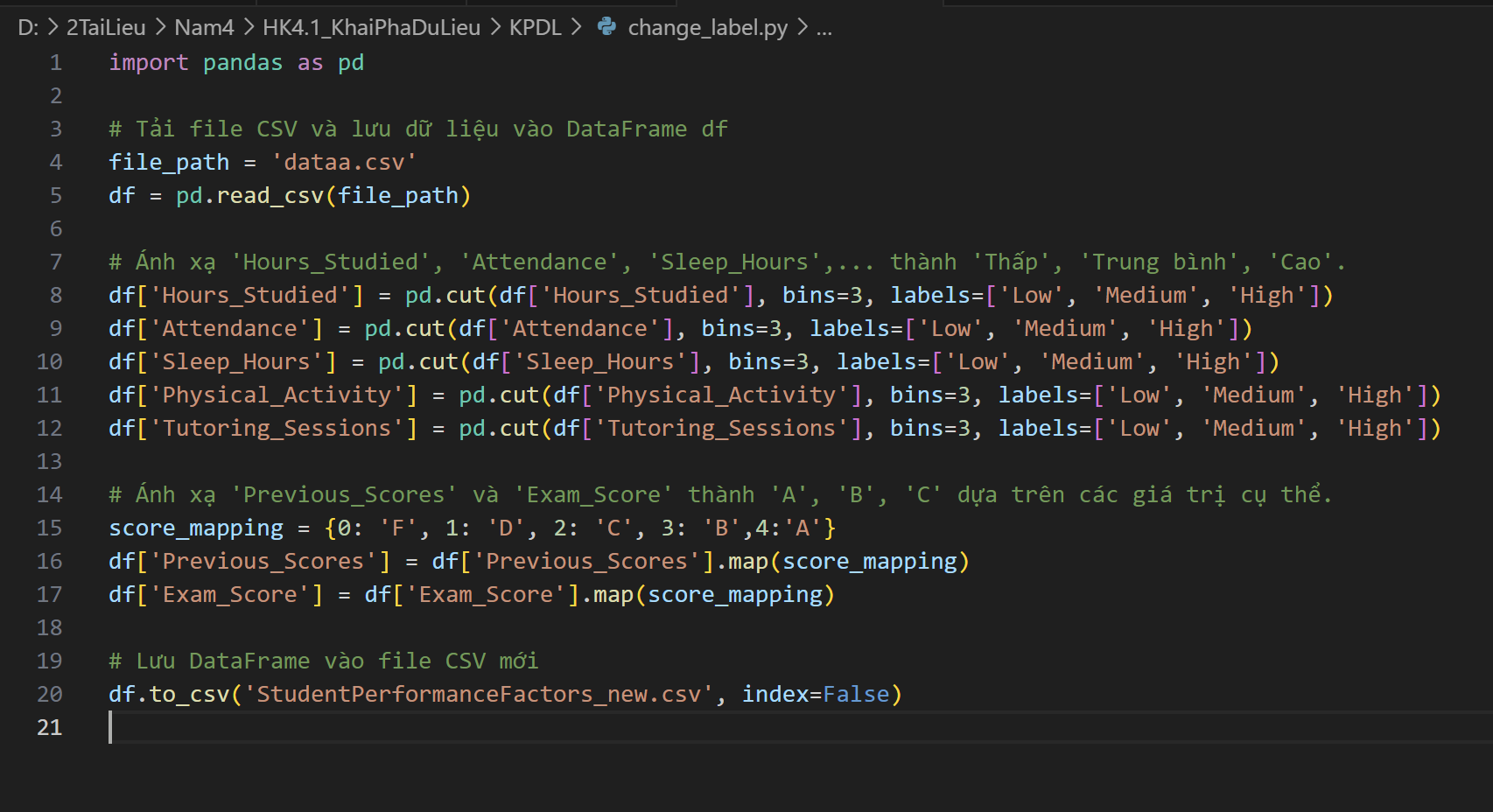
**A screenshot of a graph

Description automatically generated**

* A screenshot of a graph

  Description automatically generatedChia dữ liệu thuộc tính Tutoring\_Sessions vào 3 nhóm: 0 - Low ( 0 - 2), 1 – Medium (2 – 4), 2 – High ( > 4 )
* **A screenshot of a graph

  Description automatically generated**Chia dữ liệu thuộc tính Attendance vào 3 nhóm: 0 - Low ( 0 - 3), 1 – Medium (3 – 6), 2 – High ( > 6 )
* Biến đổi dữ liệu thành các nhãn: Sử dụng python để tiến hành đổi các giá trị của các cột thành các nhãn chữ



* Rời rạc hóa dữ liệu: Chuyển các thuộc tính có kiểu Numeric thành Nominal

1. **Phân tích dữ liệu**

* 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 được sử dụng cho việc khai phá bao gồm **6608** mẫu với **20** thuộc tính

|  |  |
| --- | --- |
| Hours\_Studied | Số giờ dành cho việc học mỗi tuần |
| Attendance | Tỷ lệ phần trăm tham gia lớp học |
| Parental\_Involvement | Mức độ tham gia của phụ huynh vào việc học tập của học sinh (Low: Thấp, Medium: Trung bình, High: Cao) |
| Acess\_to\_Resources | Khả năng tiếp cần tài nguyên giáo dục |
| Extracurricular\_Activities | Tham gia các hoạt động ngoại khóa (Yes: Có, No:Không) |
| Sleep\_Hours | Số giờ ngủ trung bình mỗi đêm |
| Previous\_Scores | Điểm của bài kiểm tra trước đó |
| Motivation\_Level | Mức độ động lực của học sinh (Low: thấp, Medium: trung bình, High: cao) |
| Internet\_Access | Khả năng truy cập Internet (Yes: có, No: không) |
| Tutoring\_Sessions | Số buổi học thêm mỗi tháng |
| Family\_Income | Mức thu nhập của gia đình (Low: thấp, Medium: trung bình, High: cao) |
| Teacher\_Quality | Chất lượng giáo viên (Low: thấp, Medium: trung bình, High: cao) |
| School\_Type | Loại trường đang theo học (Public: trường công, Private: trường tư) |
| Peer\_Inflluence | Ảnh hưởng của bạn bè đến kết quả học tập (Positive: tích cực, Neutral: trung bình, Negative: tiêu cực) |
| Physical\_Activity | Số giờ hoạt động thể chất trung bình mỗi tuần |
| Learning\_Disabilities | Có bị mắc những hội chứng rối loạn học tập không (Yes: có, No: không) |
| Parental\_Education\_Level | Trình độ học vấn cao nhất của cha mẹ (High School: Trung học, College: Đại học, Postgraduate: Sau đại học) |
| Distance\_from\_Home | Khoảng cách từ nhà đến trường (Near: gần, Moderate: trung bình, Far: xa) |
| Gender | Giới tính của học sinh (Male: nam, Female: nữ) |
| Exam\_Score | Điểm bài thi cuối kì |

* Thống kê chi tiết các thuộc tính tương quan với thuộc tính phân lớp (Exam\_Score)

**A screenshot of a computer screen

Description automatically generated**

* **Có thể thấy các thuộc tính Attendance, Hours\_Studied, Access\_to\_Resources có mức độ tương quan cao nhất khi mức điểm cao được phân bố chủ yếu vào các cột giá trị cao của các thuộc tính này. Còn đối với các thuộc tính khác có thể thấy giá trị điểm đạt mức tốt phân bố tương đối đồng đều trên các cột giá trị**
* Dữ liệu sau khi tiền xử lý:

**A screen shot of a computer

Description automatically generated**

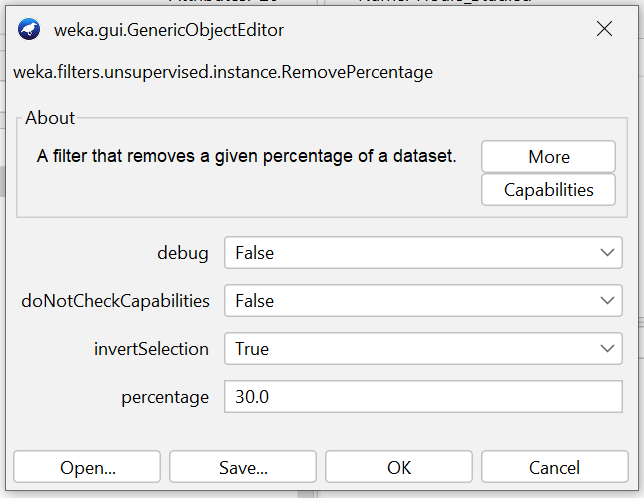
1. **Phân tách dữ liệu Train, Test.**

* Để có thể tối ưu cho việc huấn luyện, ta sẽ chia dữ liệu thành 2 phần là Data Train và Data Test với tỉ lệ là 70/30 với 70 dành cho Data Train và 30 cho Data Test
* Với tập dữ liệu Train: ta sử dụng weka: Filter -> unsupervised -> instance -> RemovePercentage

A screenshot of a computer

Description automatically generated

* + Sau khi thiết lập xong, chọn Ok và apply Filter. Khi này ta đã có được 1 file dữ liệu train với 70% dữ liệu ban đầu
  + Lưu lại file dữ liệu
* Với tập dữ liệu Test: ta sẽ load lại tập dữ liệu ban đầu (hoặc Undo với tab weka). Sau đó chọn Filter -> unsupervised -> instance -> RemovePercentage

****

* + Với lần này invertSelection được thiết lập giá trị **True**
  + Sau đó tiến hành apply Filter, lưu lại file dữ liệu Test

1. **Phương pháp phân lớp và thuật toán phân lớp**
2. **Bài toán phân lớp**

* Phân lớp dữ liệu là kỹ thuật dựa trên tập huấn luyện (Data train), 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 có giám sát

1. **Thuật toán ID3**

### **2.1. Lý thuyết**

**-** Là một thuật toán học các mẫu để tạo cây quyết định. Tạo cây quyết định bằng việc tìm kiếm cơ bản từ trên xuống trên tập huấn luyện

**-** Thuật toán ID3 bắt đầu với một tập ữ liệu ban đầu và một tập các thuộc tính có thể được sử dụng để phân loại. Sử dụng độ lợi thông tin để tìm ra thuộc tính tốt nhất chia các mẫu thành các nhóm. Thuật toán tiếp tục phân tách các nhóm con đến khi tất cả các mẫu trong nhóm đều thuộc cùng một lớp hoặc không còn thuộc tính nào có thể được sử dụng để phân loại

* Khi cây quyết định được xây dựng, việc phân loại của mỗi một mẫu mới được thực hiện bằng cách đi xuống các nhánh cây theo các giá trị của các thuộc tính cho đến khi đại tới một lá cây. Lá này sẽ chỉ ra lớp mà mẫu đó thuộc về
* Hạn chế: Bị overfitting nếu cây quá phức tạp và khả năng bị ảnh hưởng bởi các thuộc tính có giá trị lớn
* Độ lợi thông tin
  + Thông tin mong đợi để phân lớp một mẫu trong D theo nhãn lớp:

A mathematical equation with numbers and symbols

Description automatically generated

* + Thông tin cần thiết để phân chia D theo thuộc tính A:

A mathematical equation with numbers and symbols

Description automatically generated

* + Độ lợi thông tin của sự phân chia dựa trên thuộc tính A:

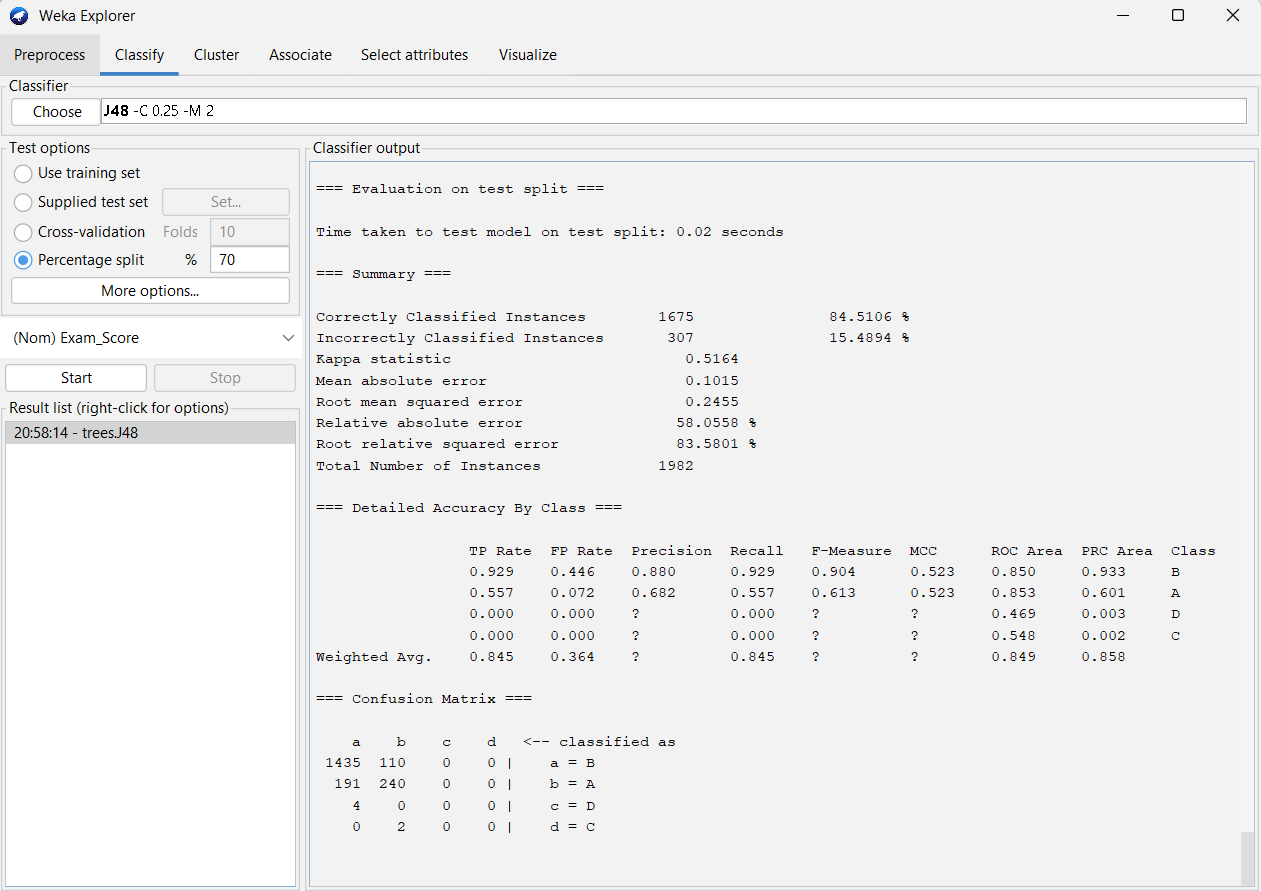
A black and blue text

Description automatically generated

### **2.2. Quy trình thực hiện**

Bước 1: Xác định thuộc tính chính để phân loại đối tượng.  
Bước 2: Tạo một nút trên cây quyết định dựa trên thuộc tính đã chọn  
Bước 3: Chia tập dữ liệu thành các tập con dựa trên giá trị của thuộc tính đã chọn ở bước 1  
Bước 4: Lặp lại các bước trên cho tập dữ liệu con được chọn trong bước 3  
Bước 5: Dừng lại khi một trong các điều kiện sau được thỏa mãn:  
• Tất cả các đối tượng trong tập dữ liệu đang xét đều thuộc cùng một lớp.  
• Không còn thuộc tính nào để chọn.  
• Số lượng đối tượng trong tập dữ liệu đang xét quá nhỏ hoặc không đủ để chia thành  
các tập con.  
Bước 6: Gắn nhãn cây với lớp được phổ biến nhất của các đối tượng trong tập dữ liệu đã xét

### **2.3. Thực hiện phân lớp với weka**



**Kết quả chạy cho thấy:**

Các trường hợp được phân loại chính xác 1674: chiếm 84.5106%

Các trường hợp được phân loại không chính xác là 307: chiếm 15.4894%

Thống kê Kappa: 0.5164

Sai số tuyệt đối trung bình: 0.1015

Lỗi bình phương gốc trung bình: 0.2455

Sai số tuyệt đối tương đối: 58.0558%

Lỗi bình phương tương đối gốc: 83.5801%

Tổng số mẫu: 1982

Confusion Matrix:

* Có 1630 điểm B -> thuật toán ID3 dự đoán đúng 1435 điểm B, dự đoán sai 191 điểm A, dự đoán sai 4 điểm D
* Có 352 điểm A -> thuật toán ID3 dự đoán đúng 240 điểm A, dự đoán sai 110 điểm B, dự doán sai 2 điểm C

1. **Thuật toán Naïve Bayes**
   1. **Giới thiệu thuật toán**

* **Khái niệm:**
  + Naive Bayes là một loại mô hình phân lớp dữ liệu thuộc về phân nhóm các phương pháp học máy có giám sát. Nó dựa trên Định lý Bayes và một giả định "naive" (ngây thơ) rằng tất cả các tính năng (features) đều độc lập với nhau, nghĩa là sự xuất hiện của một đặc trưng cụ thể trong một lớp không liên quan đến sự xuất hiện của các đặc trưng khác.
* **Mô hình và phương trình:** 
  + Mô hình Naive Bayes đưa ra dự đoán dựa trên xác suất hậu nghiệm của mỗi lớp, được tính như sau:

Trong đó:

là nhãn lớp được xét

là các thuộc tính của mẫu dữ liệu

là xác suất cho trước (prior probability) của lớp

là xác suất của thuộc tính nếu biết lớp là (xác suất

có điều kiện)

* 1. **Thuật toán**
* Thuật toán Naive Bayes bao gồm hai phần: học và dự đoán. Trong bước học, thuật toán sẽ tính được xác suất trước của mỗi lớp và xác suất có điều kiện của mỗi thuộc tính cho mỗi lớp. Bước dự đoán sử dụng mô hình đã học để đưa ra dự đoán cho các mẫu dữ liệu mới.
* **Quy trình thực hiện của thuật toán:**
  + **Chuẩn bị dữ liệu:** Nhãn cho tập huấn luyện và đặc trưng của dữ liệu phải được xác định trước
  + **Học mô hình:**
    - Tính cho mỗi nhãn
    - Tính cho mỗi thuộc tính và nhãn
  + **Dự đoán:**
    - Với một mẫu dữ liệu mới, tính cho mỗi lớp dùng công thức trên
    - Chọn giá trị của sao cho là lớn nhất
* **Nhận xét:** Thuật toán này có thể được triển khai sử dụng nhiều ngôn ngữ lập trình và công cụ khác nhau, thường được áp dụng trong các bài toán phân loại, lọc thư rác, dự đoán văn bản và nhiều ứng dụng khác do khả năng xử lý dữ liệu đa chiều và hiệu quả về mặt tính toán.

1. **Cài đặt thuật toán**
2. **Sử dụng thư viện Scikit-learn để thực hiện huấn luyện mô hình cây quyết định ID3**

**A screen shot of a computer program

Description automatically generated**

* Sau khi thực hiện thuật toán ta sẽ thu được cây quyết định

A diagram of a computer network

Description automatically generated with medium confidence

1. **Xây dựng thuật toán ID3 không sử dụng thư viện**

* Class DecisionTreeID3

**A screenshot of a computer program

Description automatically generated**

**A screen shot of a computer program

Description automatically generated**

* Hàm main:

**A screen shot of a computer program

Description automatically generated**

* Thực hiện thuật toán thu được các điểm đánh giá mô hình thuật toán:

A number on a black background

Description automatically generated

1. **Thuật toán Naïve Bayes sử dụng thư viện sklearn**

**A screen shot of a computer program

Description automatically generated**

* Thực thi code ta thu được kết quả:

**A screenshot of a computer

Description automatically generated**

1. **Xây dựng thuật toán Naïve Bayes không sử dụng thư viện**

* Xây dựng lớp NaiveBayes

**A screen shot of a computer program

Description automatically generated**

* Xây dựng chương trình huấn huyện mô hình

**A screen shot of a computer program

Description automatically generated**

* Kết quả chạy mô hình thu được:

A screenshot of a computer

Description automatically generated

# **KẾT LUẬN**

Khai phá dữ liệu với phương pháp phân lớp là một trong những phương pháp phổ biến, mang tính ứng dụng, hiệu quả thực tiễn cao trong các lĩnh vực như thương mại, y tế, giáo dục,... và sẽ tiếp tục phát triển trong tương lai.

Sau khi hoàn thành đề tài “***Khai phá dữ liệu về các yếu tố ảnh hưởng đến hiệu suất học tập của học sinh bằng phương pháp phân lớp*”,** nhóm em đã đạt được kết quả sau:  
- Tìm hiểu tổng quan về khám phá tri thức, khai phá dữ liệu, bài toán phân lớp cùng thuật toán ID3 để xây dựng mô hình phân lớp hỗ trợ dự đoán điểm số của học sinh dựa trên các yếu tố có khả năng ảnh hướng  
- Thu thập dữ liệu đánh giá, tiền xử lý dữ liệu với các công cụ phần mềm như Excel, Weka. Xây dựng được mô hình phân lớp trên Weka.  
- So sánh kết quả mô hình phân lớp với cả phương pháp kiểm thử khác nhau để thu được mô hình tốt nhất.  
Trong quá trình hoàn thành đề tài, dù nhóm em đã cố gắng tìm hiểu và thực hiện bài toán nhưng vẫn không thể tránh những thiếu sót, khiếm khuyết. Do vậy, rất mong nhận được sự đóng góp ý kiến của thầy cô để chúng em rút kinh nghiệm, cải thiện kỹ năng và tích lũy kiến thức trong môn học.

# **TÀI LIỆU THAM KHẢO**

[**1]** TS.Đặng Thị Thu Hiền, Nguyễn Tu Trung, Bài giảng Khai phá dữ liệu  
[2] <https://www.kaggle.com/datasets/lainguyn123/student-performance-factors>  
[3] <https://machinelearningcoban.com/2017/08/08/nbc/>

[4] <https://www.geeksforgeeks.org/multinomial-naive-bayes/>

[5] <https://machinelearningcoban.com/2018/01/14/id3/>