TRƯỜNG ĐẠI HỌC ĐIỆN LỰC

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CHUYÊN ĐỀ HỌC PHẦN**

**KHAI PHÁ DỮ LIỆU**

**SỬ DỤNG THUẬT TOÁN XGBOOST ĐỂ DỰ ĐOÁN KHẢ NĂNG TRẢ NỢ CỦA KHÁCH HÀNG**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Sinh viên thực hiện** | **: LẠI THỊ TÂM**  **VŨ KHÁNH HUYỀN**  **PHẠM LAN ANH** | | **Giảng viên hướng dẫn** | **: VŨ VĂN ĐỊNH** | | | **Ngành** | **: CÔNG NGHỆ THÔNG TIN** | | | **Chuyên ngành** | **: CÔNG NGHỆ PHẦN MỀM** | | | **Lớp** | **: D16CNPM4** | | | **Khóa** | **: 2021-2026** | | |  |
|  |  |

**Hà Nội, tháng 5 năm 2024**

**PHIẾU CHẤM ĐIỂM**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Họ và Tên** | **Nội dung thực hiện** | **Điểm** | **Chữ ký** |
| 1 | Lại Thị Tâm | Báo cáo + tìm hiểu cây quyết định |  |  |
| 2 | Vũ Khánh Huyền | Báo cáo + tìm hiểu cây quyết định |  |  |
| 3 | Phạm Lan Anh | Báo cáo + tìm hiểu cây quyết định |  |  |

|  |  |  |
| --- | --- | --- |
| **Họ và tên** | **Chữ ký** | **Ghi chú** |
| Giảng viên chấm 1: |  |  |
| Giảng viên chấm 2: |  |  |

**MỤC LỤC**

LỜI MỞ ĐẦU

[CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU 2](#_Toc163201178)

[1.1 Khái niệm chung: 2](#_Toc163201179)

[1.1.1 Khái niệm cơ bản: 2](#_Toc163201180)

[1.1.2 Các bước trong quá trình khai phá 2](#_Toc163201181)

[1.1.3 Ứng dụng của khai phá dữ liệu 4](#_Toc163201182)

[1.2 Tiền xử lý dữ liệu 4](#_Toc163201183)

[1.2.1 Dữ liệu 4](#_Toc163201184)

[1.2.2 Làm sạch dữ liệu (data cleaning) 4](#_Toc163201185)

[1.2.3 Tích hợp dữ liệu (data integration) 5](#_Toc163201186)

[1.2.4 Biến đổi dữ liệu (data transformation) 6](#_Toc163201187)

[1.2.5. Thu giảm dữ liệu (data reduction) 6](#_Toc163201188)

[1.3 Phương pháp khai phá 7](#_Toc163201189)

[1.3.1 Phân loại 7](#_Toc163201190)

[1.3.2. Hồi quy 7](#_Toc163201191)

[1.3.3 Luật kết hợp 8](#_Toc163201192)

[1.3.4 Phân cụm 8](#_Toc163201193)

[CHƯƠNG 2: GIẢI QUYẾT BÀI TOÁN CÂY QUYẾT ĐỊNH 9](#_Toc163201194)

[2.1. Tổng quan về thuật toán cây quyết định 9](#_Toc163201195)

[2.1.1. Giới thiệu thuật toán 9](#_Toc163201196)

[2.1.2. Cấu trúc của thuật toán 11](#_Toc163201197)

[2.2. Các thuật toán liên quan đến cây quyết định 11](#_Toc163201198)

[2.2.1. ID3 (Iterative Dichotomiser 3) Algorithm 11](#_Toc163201199)

[2.2.2. ID3 (Examples, Target\_Attribute, Attributes) 11](#_Toc163201200)

[2.2.3. C4.5 Algorithm 13](#_Toc163201201)

[2.3. Ứng dụng của thuật toán 16](#_Toc163201202)

[2.4. Ưu và nhược điểm của thuật toán 16](#_Toc163201203)

[2.4.1. Ưu điểm: 16](#_Toc163201204)

[2.4.2. Nhược điểm: 17](#_Toc163201205)

[CHƯƠNG 3: ỨNG DỤNG THUẬT TOÁN 19](#_Toc163201206)

[3.1 Thuật toán phân lớp cây quyết định dự đoán nơi sản xuất của các ô tô bị hư hỏng do tai nạn 19](#_Toc163201207)

[3.1.1. Phát biểu bài toán: 19](#_Toc163201208)

[3.1.2 Yêu cầu bài toán: 19](#_Toc163201209)

[3.1.3. Xây dựng bộ dữ liệu: 19](#_Toc163201210)

[KẾT LUẬN 26](#_Toc163201211)

[TÀI LIỆU THAM KHẢO 27](#_Toc163201212)

**DANH MỤC HÌNH ẢNH**

[Hình 1. 1 Quá trình khai phá tri thức 3](#_Toc163201270)

[Hình 1. 2 Minh họa về cây quyết định 10](#_Toc163201271)

**LỜI MỞ ĐẦU**

Trong nhiều năm qua, cùng với sự phát triển của công nghệ thông tin và ứng dụng của công nghệ thông tin trong nhiều lĩnh vực của đời sống xã hội, thì lượng dữ liệu được các cơ quan thu thập và lưu trữ ngày một nhiều lên. Người ta lưu trữ những dữ liệu này vì cho rằng nó ẩn chứa những giá trị nhất định nào đó. Tuy nhiên theo thống kê thì chỉ có một lượng nhỏ của những dữ liệu này (khoảng từ 5% đến 10%) là luôn được phân tích, số còn lại họ không biết sẽ phải làm gì và có thể làm gì với những dữ liệu này, nhưng họ vẫn tiếp tục thu thập và lưu trữ vì hy vọng những dữ liệu này sẽ cung cấp cho họ những thông tin quý giá một cách nhanh chóng để đưa ra những quyết định kịp thời vào một lúc nào đó.

Chính vì vậy, các phương pháp quản trị và khai thác cơ sở dữ liệu truyền thống ngày càng không đáp ứng được thực tế đã làm phát triển một khuynh hướng kỹ thuật mới đó là Kỹ thuật phát hiện tri thức và khai phá dữ liệu (KDD - Knowledge Discovery and Data Mining). Kỹ thuật phát hiện tri thức và khai phá dữ liệu đã và đang được nghiên cứu, ứng dụng trong nhiều lĩnh vực khác nhau trên thế giới, tại Việt Nam kỹ thuật này còn tương đối mới mẻ tuy nhiên cũng đang được nghiên cứu và bắt đầu đưa vào một số ứng dụng thực tế.

Vì vậy, hiện nay ở nước ta vấn đề phát hiện tri thức và khai phá dữ liệu đang thu hút được sự quan tâm của nhiều người và nhiều công ty phát triển ứng dụng công nghệ thông tin.

# CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

## 1.1 Khái niệm chung:

### 1.1.1 Khái niệm cơ bản:

"Data Mining" là "quá trình khai phá", trích xuất, khai thác và sử dụng những dữ liệu có giá trị tiềm ẩn từ bên trong lượng lớn dữ liệu được lưu trữ trong các cơ sở dữ liệu (CSDL), kho dữ liệu, trung tâm dữ liệu lớn hơn là Big Data dựa trên kĩ thuật như mạng nơ ron, lí thuyết tập thô, tập mờ, biểu diễn tri thức. Data Mining là một công đoạn trong hoạt động “làm sạch” dữ liệu giúp cho dữ liệu được truyền dẫn một cách nhanh nhất.

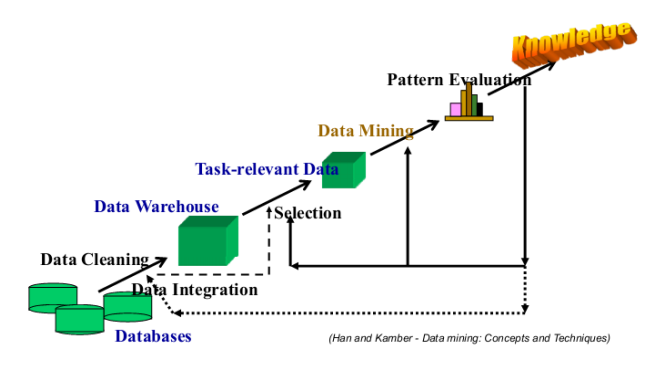
 Quá trình khai phá dữ liệu là một quá trình phức tạp bao gồm kho dữ liệu chuyên sâu cũng như các công nghệ tính toán. Hơn nữa, Data Mining không chỉ giới hạn trong việc trích xuất dữ liệu mà còn được sử dụng để chuyển đổi, làm sạch, tích hợp dữ liệu và phân tích mẫu.Có nhiều tham số quan trọng khác nhau trong Data Mining, chẳng hạn như quy tắc kết hợp, phân loại, phân cụm và dự báo. Một số tính năng chính của Data Mining:

* Dự đoán các mẫu dựa trên xu hướng trong dữ liệu.
* Tính toán dự đoán kết quả
* Tạo thông tin phản hồi để phân tích
* Tập trung vào cơ sở dữ liệu lớn hơn.
* Phân cụm dữ liệu trực quan

### 1.1.2 *Các Bước trong Data mining:*

Các bước quan trọng khi Data Mining bao gồm:

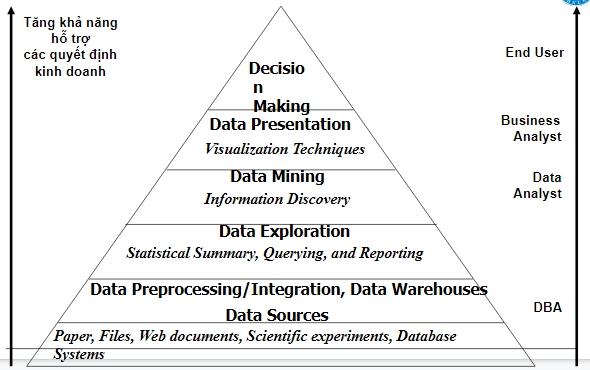
* Bước 1: Làm sạch dữ liệu – Trong bước này, dữ liệu được làm sạch sao cho không có dữ liệu tạp hay bất thường trong dữ liệu.
* Bước 2: Tích hợp dữ liệu – Trong quá trình tích hợp dữ liệu, nhiều nguồn dữ liệu sẽ kết hợp lại thành một.
* Bước 3: Lựa chọn dữ liệu – Trong bước này, dữ liệu được trích xuất từ cơ sở dữ liệu.
* Bước 4: Chuyển đổi dữ liệu – Trong bước này, dữ liệu sẽ được chuyển đổi để thực hiện phân tích tóm tắt cũng như các hoạt động tổng hợp.
* Bước 5: Khai phá dữ liệu – Trong bước này, chúng tôi trích xuất dữ liệu hữu ích từ nhóm dữ liệu hiện có.
* Bước 6: Đánh giá mẫu – Chúng tôi phân tích một số mẫu có trong dữ liệu.
* Bước 7: Trình bày thông tin – Trong bước cuối cùng, thông tin sẽ được thể hiện dưới dạng cây, bảng, biểu đồ và ma trận.
* Đây là một góc nhìn từ hệ thống cơ sở dữ liệu điển hình.
* Khai phá dữ liệu đóng một vai trò thiết yếu trong quá trình khai phá trí thức.



Hình 1. 1 Mô hình mô tả quá trình khai phá dữ liệu

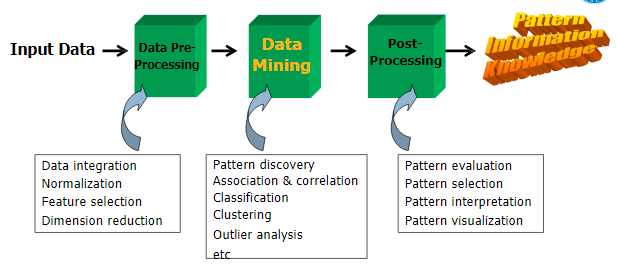
*Quá trình khai phá trí thức là một chuỗi lặp:*

* Data cleaning (làm sạch dữ liệu)
* Data integration (tích hợp dữ liệu)
* Data selection (chọn lựa dữ liệu)
* Data transformation (biến đổi dữ liệu)
* Data mining (khai phá dữ liệu)
* Pattern evaluation (đánh giá mẫu)
* Knowledge presentation (biểu diễn tri thức)
* Quá trình khai phá dữ liệu trong kinh doanh thông minh



*Hình 1.2:Mô hình mô tả quá trình khai phá trong kinh doanh*

- Quá trình khai phá dữ liệu từ góc nhìn máy học thống kê



*Hình 1.3:Mô hình mô tả quá trinh khai phá từ máy học thống kê*

## 1.2. Kiến trúc của một hệ thống khai phá dữ liệu

Khai phá dữ liệu là quá trình rút trích thông tin bổ ích từ những kho dữ liệu lớn, là quá trình chính trong khai phá trí thức từ cơ sở dữ liệu.

Kiến trúc của một hệ thống khai phá dữ liệu có các thành phần như:

### C:\Users\Pham Quynh\OneDrive\Hình ảnh\Cuộn phim\KPDL\Capture.PNG

*Hình 1.4: Khám phá trí thức trong cơ sở dữ liệu điển hình*

* Cơ sở dữ liệu, kho dữ liệu hoặc lưu trữ thông tin khác: Đây là một hay các tập cơ sở dữ liệu, các kho dữ liệu, các trang tính hay các dạng khác của thông tin được lưu trữ. Các kỹ thuật làm sạch dữ liệu và tích hợp dữ liệu có thể được thực hiện.
* Máy chủ cơ sở dữ liệu (Database or Warehouse Server): Máy chủ có trách nhiệm lấy những dữ liệu thích hợp dựa trên những yêu cầu khám phá của người dùng.
* Cơ sở trí thức (Knowledge-base): Đây là miền tri thức dùng để tìm kiếm hay đánh giá độ quan trọng của các mẫu kết quả thu được. Trí thức này có thể bao gồm một sự phân cấp khái niệm dùng để tổ chức các thuộc tính hay các giá trị thuộc tính ở các mức trừu tượng khác nhau.
* Máy khai phá dữ liệu (Data mining engine): Là một hệ thống khai phá dữ liệu cần phải có một tập các modul chức năng để thực hiện công việc, chẳng hạn như kết hợp, phân lớp, phân cụm.
* Modul đánh giá mẫu (Pattern evaluation): Bộ phận tương tác với các modul khai phá dữ liệu để tập trung vào việc duyệt tìm các mẫu đáng được quan tâm. Nó có thể dùng các ngưỡng về độ quan tâm để lọc mẫu đã khám phá được. Cũng có thể modul đánh giá mẫu được tích hợp vào modul khai phá dữ liệu, tùy theo các cài đặt của phương pháp khai phá dữ liệu được dùng.
* Giao diện đồ họa người dùng (Graphical user interface): Bộ phận này cho phép người dùng giao tiếp với hệ thống khai phá dữ liệu. Thông qua giao diện này người dùng tương tác với hệ thống bằng cách đặc tả một yêu cầu khai phá hay một nhiệm vụ, cung cấp thông tin trợ giúp cho việc tìm kiếm và thực hiện khia phá thăm dò trên các kết quả khai phá trung gian. Ngoài ra bộ phận này còn cho phép người dùng xem các lược đồ cơ sở dữ liệu, lược đồ kho dữ liệu, các đánh giá mẫu và hiển thị các mẫu trong các khuôn dạng khác nhau.

## 1.3.  Chức năng khai phá dữ liệu:

Các kỹ thuật khai phá dữ liệu thực hiện 2 chức năng chính:

* Chức năng mô tả: mô tả về các tính chất hoặc các đặc tính chung của dữ liệu trong cơ sở dữ liệu, các kỹ thuật này gồm có: phân cụm (Clustering), tổng hợp (Summarization), trực quan hóa (Visualization), phân tích sự phát triển và độ lệnh (Evolution and deviation analysis), phân tích luật kết hợp (Association rules),...
* Chức năng dự đoán: đưa ra các dự đoán dựa vào các suy diễn trên dữ liệu hiện thời, các kỹ thuật này gồm có: phân lớp (Classification), hồi quy (Regression), cây quyết định (Decision tree), ....

### 1.3.1. Phân lớp dữ liệu (Classification)

* Là dạng phân tích dữ liệu nhằm rút trích các mô hình mô tả các lớp dữ liệu hoặc dự đoán xu hướng dữ liệu.
* Quá trình gồm hai bước:
* Bước học (giai đoạn huấn luyện): xây dựng bộ phân lớp (classifier) bằng việc phân tích/học tập huấn luyện.
* Bước phân lớp (classification): phân lớp dữ liệu/đối tượng mới nếu độ chính xác của bộ phân lớp được đánh giá là có thể chấp nhận được (acceptable).
* Các thuật toán phân lớp dữ liệu
* Phân lớp với cây quyết định (decision tree)
* Phân lớp với Naïve Bayesian
* Phân lớp với k phần tử gần nhất (k-nearest neighbor)
* Phân lớp với máy vector hỗ trợ (SVM)
* Phân lớp với mạng neural (neural network)
* Phân lớp dựa trên tiến hoá gen (genetic algorithms)
* Phân lớp với lý thuyết tập thô, tập mờ (rough sets)
* Phân lớp với lý thuyết tập mờ (fuzzy sets) ...

### 1.3.2. Hồi quy (Regression)

Là việc học một hàm ánh xạ từ một mẫu dữ liệu thành một biến dự đoán có giá trị thực. Nhiệm vụ của hồi quy tương tự như phân lớp, điểm khác nhau chính là ở chỗ thuộc tính để dự báo là liên tục chứ không phải là rời rạc. Việc dự báo các giá trị số thường được làm bởi các phương pháp thống kê cổ điển, chẳng hạn như hồi quy tuyến tính. Tuy nhiên, phưong pháp mô hình hóa cũng được sử dụng, ví dụ: cây quyết định.

### *1.3.3. Phân cụm dữ liệu (Clustering)*

* Là quá trình phân nhóm/cụm dữ liệu/đối tượng vào các lớp/cụm.
* Các đối tượng trong cùng một cụm tương tự với nhau hơn so với đối tượng ở cụm khác.
* Các yêu cầu của phân cụm dữ liệu:
* Có thể tương thích, hiệu quả với dữ liệu lớn, số chiều lớn.
* Có khả năng xử lý các dữ liệu khác nhau.
* Có khả năng khám phá các cụm với các dạng bất kỳ.
* Khả năng thích nghi với dữ liệu nhiễu.
* Ít nhạy cảm với thứ tự của các dữ liệu vào.
* Phân cụm rằng buộc.
* Dễ hiểu và dễ sử dụng.
* Phân loại các phương pháp phân cụm:
* Phân hoạch (partitioning): phân hoạch tập dữ liệu n phần tử thành k cụm.
* Phân cấp (hierarchical): xây dựng phân cấp các cụm trên cơ sở các đối tượng dữ liệu đang xem xét.
* Dựa trên mật độ (density-based): dựa trên hàm mật độ, số đối tượng lân cận của đối tượng dữ liệu.
* Dựa trên lưới (grid-based): dựa trên dữ liệu nhiều chiều, chủ yếu áp dụng cho lớp dữ liệu không gian.
* Dựa trên mô hình (model-based): một mô hình giả thuyết được đưa ra cho mỗi cụm; sau đó hiệu chỉnh các thông số để mô hình phù hợp với cụm dữ liệu/đối tượng nhất.

### 1.3.4. Khai phá luật kết hợp (Association rules)

Chẳng hạn như có luật: âm nhac, thể thao => thiếu nhi, nghĩa là những người mua sách âm nhạc và thể thao thì cũng mua sách thiếu nhi. Lúc đó ta sẽ quan tâm đến số lượng trường hợp khách hàng thỏa mãn luật này trong cơ sở dữ liệu hay độ hỗ trợ (Support) cho luật này. Độ hỗ trợ cho luật chính là phần trăm số bản ghi có cả sách âm nhạc, thể thao và thiếu nhi hay tất cả những người thích cả ba loại sách nói trên. Tuy nhiên, giá trị độ hỗ trợ là không đủ, có thể có trường hợp ta có một nhóm tương đối những người đọc cả ba loại trên nhƣng lại có một nhóm với lực lượng lớn hơn những người thích sách thể thao, âm nhạc mà không thích sách thiếu nhi. Trong trường hợp này tính kết hợp rất yếu mặc dù độ hỗ trợ tương đối cao, như vậy chúng ta cần thêm một độ đo thứ hai đó là độ tin cậy (confidence). Độ tin cậy chính là phần trăm các bản ghi có sách thiếu nhi trong số các bản ghi có sách âm nhạc và thể thao.

### 1.3.5. Phân tích sự phát triển và độ lệch

Nhiệm vụ này tập trung vào khám phá hầu hết sự thay đổi có nghĩa dưới dạng độ đo đã biết trước hoặc giá trị chuẩn, phát hiện độ lệch đáng kể giữa nội dung của tập con dữ liệu thực và nội dung mong đợi. Hai mô hình độ lệch hay dùng là lệch theo thời gian hay lệch theo nhóm. Độ lệch theo thời gian là sự thay đổi có ý nghĩa của dữ liệu theo thời gian. Độ lệch theo nhóm là sự khác nhau giữa dữ liệu trong hai tập con dữ liệu, ở đây tính cả trường hợp tập con dữ liệu này thuộc tập con kia, nghĩa là xác định dữ liệu trong một nhóm con của đối tượng có khác đáng kể so với toàn bộ đối tượng không? Theo cách này, sai sót dữ liệu hay sai lệch so với giá trị thông thường được phát hiện.

Vì những nhiệm vụ này yêu cầu số lượng và các dạng thông tin rất khác nhau nên chúng thường ảnh hưởng đến việc thiết kế và chọn phương pháp khai phá dữ liệu khác nhau. Ví dụ như phương pháp cây quyết định tạo ra được một mô tả phân biệt được các mẫu giữa các lớp nhưng không có tính chất và đặc điểm của lớp.

# CHƯƠNG 2 THUẬT TOÁN XGBOOST

## 2.1 Giới thiệu

Ứng dụng công nghệ thông tin vào việc lưu trữ và xử lý thông tin ngày nay được áp dụng hầu hết trong lĩnh vực, điều này đã tạo ra một lượng lớn dữ liệu được lưu trữ với kích thước tăng lên không ngừng. Đây chính là điều kiện tốt cho việc khai thác kho dữ liệu để đem lại tri thức có ích với các công cụ truy vấn, lập việc khai thác kho dữ liệu để đem lại tri thức có ích với các công cụ truy vấn, lập bẳng biểu và khai phá dữ liệu.

Khai phá dữ liệu là một kỹ thuật dựa trên nền tảng của nhiều lý thuyết như xác xuất, thống kê, máy học nhằm tìm kiếm các tri thức tiềm ẩn trong các kho dữ liệu có kích thước lớn mà người dùng khó có thể nhận biết bằng những kỹ thuật thông thường.

Trong lĩnh vực học máy thuật XGBoost là một thuật toán dựa trên mô hình cây quyết định. Thuật toán Xgboost để giải quyết các bài toán học giám sát (supervised learning) nó cho độ chính xác khá cao bên cạnh các mô hình học máy khác. Thuật toán XGBoost được phát triển để sử dụng nhằm mục đính phân lớp (classification) và hồi quy (regression) dữ liệu.

## 2.2 Thuật toán XGBoost

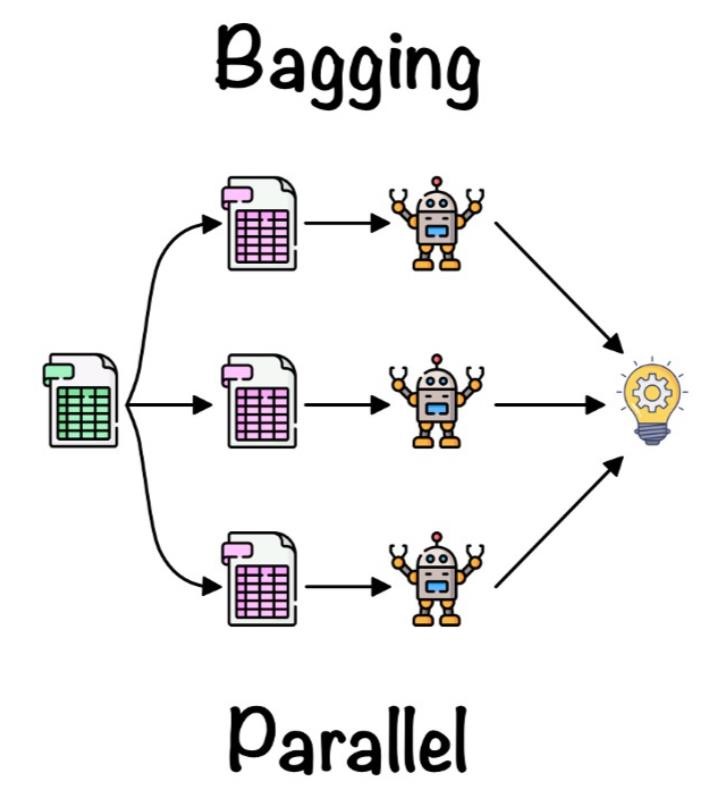
**2.2.1 Ensemble learning là gì?**

* Ensemble learning là một phương pháp với tư tưởng là “Thay vì cố gắng xây dựng một mô hình tốt duy nhất, chúng ta sẽ xây dựng một họ các mô hình yếu hơn một chút, nhưng khi kết hợp các mô hình lại sẽ thu được một mô hình còn vượt trội hơn**”.**

Ví dụ:

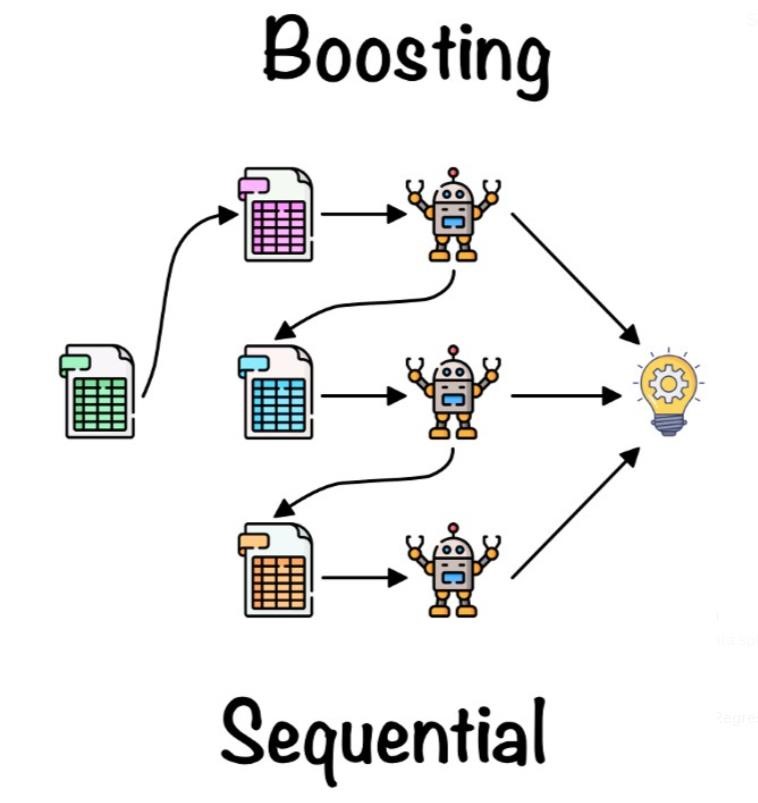
* Bằng cách Tổng hợp các phiếu bầu từ một nhóm các chuyên gia, mỗi chuyên gia sẽ mang lại kinh nghiệm và nền tảng của riêng mình để giải quyết vấn đề dẫn đến kết quả tốt hơn.
* Thuật toánEnsemble sẽ sử dụng phương pháp Baggingvà Boosting để kết hợp các mô hình yếu lại với nhau.

**2.2.2 Phương pháp Bagging**



### *Hình 2. 1: Phương pháp bagging*

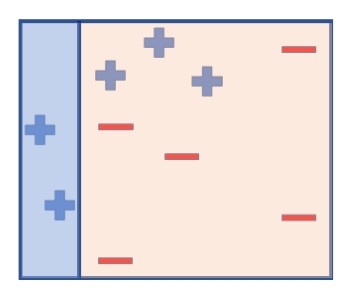
**2.2.3 Phương pháp Boosting**



### *Hình 2. 2 Phương pháp boosting*

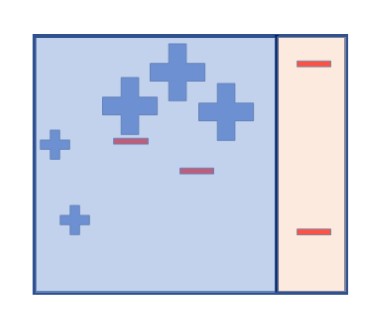
* Boosting là một kỹ thuật hoạt động bằng cách đào tạo (training) các mô hình yếu theo cách tuần tự.
* Mỗi model đang cố gắng học hỏi từ model yếu trước đó và trở nên tốt hơn trong việc đưa ra dự đoán.
* Thuật toán lặp lại cho đến khi số lượng mô hình tối đa được tạo hoặc cho đến khi mô hình cung cấp các dự đoán tốt

Ví dụ về boosting:



### *Hình 2. 3 Model 1*

* Model 1 hoạt động bằng cách cố gắng phân loại hai lớp (+) và (-) với đường thẳng đứng
* Model 1 đã gán trọng số bằng nhau cho tất cả các điểm dữ liệu vì nó không có kiến thức hoặc kinh nghiệm trước đó
* Model 1 đã phân loại sai 3 mấu (+)

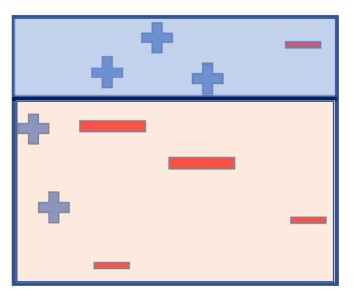


### *Hình 2. 4 Model 2*

* Model 2 học hỏi từ những sai lầm của Model 1 và gán thêm trọng số cho các điểm dữ liệu được phân loại sai (3 dấu +) như thể hiện trong hình bên trên
* Vì vậy, Model 2 vẽ một đường phân cách dọc và lần này “đảm bảo” phân loại đúng các

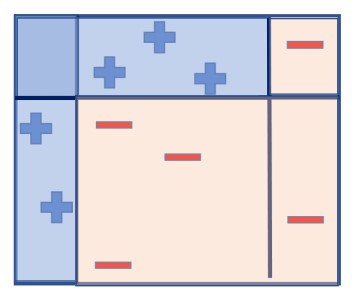
điểm này!

* Model 2 đã làm rất tốt việc phân loại chính xác các điểm có trọng số cao hơn nhưng trong quá trình này, nó đã phân loại sai 2 mẫu (-) màu đỏ



### *Hình 2. 5 Model 3*

* Model 3 học hỏi từ những sai lầm của Model 2 đó và gán thêm trọng số cho các điểm dữ liệu đượcphân loại sai (2 -) như trong hình bên dưới:
  + Vì vậy, mô hình số 3 vẽ một đường phân cách ngang và lần này “đảm bảo” phân loại đúng các điểm này!
  + Mô hình đã làm rất tốt việc phân loại chính xác các điểm có trọng số cao hơn nhưng trong quá trình này, nó đã phân loại sai hai mẫu (+) màu xanh lam.



*Hình 2.6 Model 3*

* Mô hình số 4 kết hợp tất cả các sai lầm từ tất cả các mô hình yếu này để xây dựng một mô hình mạnh hơn nhiều phân loại chính xác tất cả các điểm dữ liệu.

**2.3 Cây quyết định kết hợp (decsion tree ensembles)**

Cây quyết định là kỹ thuật Máy học giám sát (Supervised learning) trong đó dữ liệu được phân chia theo một điều kiện / tham số nhất định.

Cây gồm các nút quyết định và các lá.

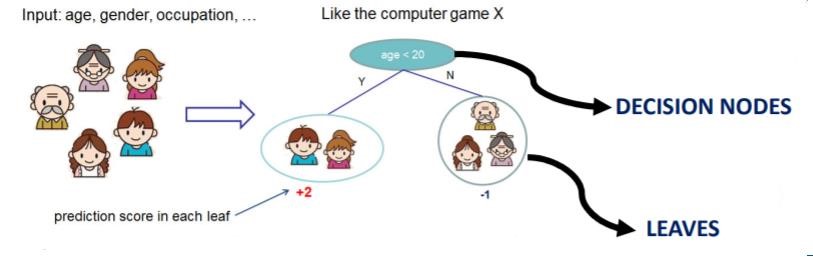
* Lá là quyết định hoặc kết quả cuối cùng.
* Các nút quyết định là nơi dữ liệu được phân chia dựa trên một thuộc tính
* Mô hình cây kết hợp bao gồm phân loại (classification) và cây truy hồi (regression trees) –

CART

- Là hướng đi của XGBoost

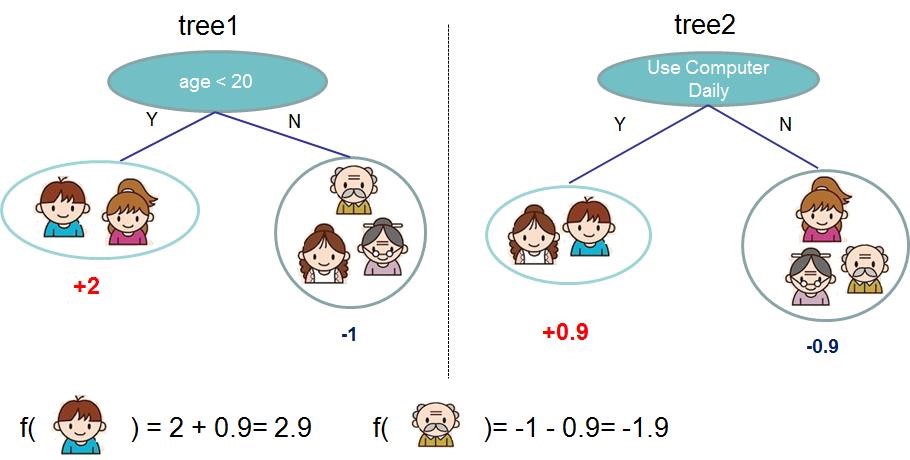
Ví dụ: Dùng CART để phân loại xem một thành viên trong gia đình có thích chơi game X trên máy tính không?

Các thành viên sẽ được chia đến các lá và có các số điểm cho trước nhất định



### *Hình 2. 6 Hình cây quyết định theo tuổi*

Thông thường, một cây đơn lẻ không đủ mạnh để sử dụng trong thực tế. Những gì thực sự được sử dụng là mô hình tổng hợp, tổng hợp dự đoán của nhiều cây với nhau. Mô hình Cây quyết định kết hợp sẽ được kết hợp từ nhiều cây như hình dưới.



*Hình 2. 7 Hình hai cây quyết định kết hợp*

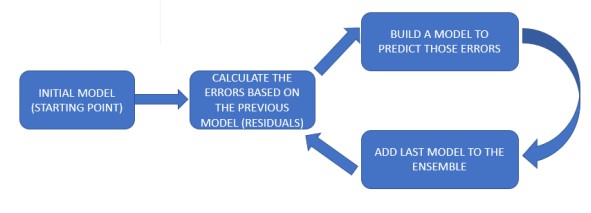
Điểm dự đoán của từng cây riêng lẻ được tổng hợp lên để có được điểm số cuối cùng.

## 2.4 Các bước thuật toán XGBoost

XGBoost liên tục xây dựng các mô hình mới và kết hợp chúng thành một mô hình tổng thể

* Ban đầu xây dựng mô hình đầu tiên và tính toán sai số cho mỗi lần quan sát trong tập dữ liệu
* Sau đó, Xây dựng một mô hình mới để dự đoán những phần dư đó (lỗi)
* Sau đó, Thêm dự đoán từ mô hình này vào nhóm các mô hình

XGboost vượt trội hơn so với thuật gradient boosting vì nó cung cấp sự cân bằng tốt giữa bias và variance (gradient boosting chỉ được tối ưu hóa cho variance có overfit dữ liệu đào tạo trong khi Xgboost cung cấp các thuật ngữ chính quy có thể cải thiện tổng quát hóa mô hình).



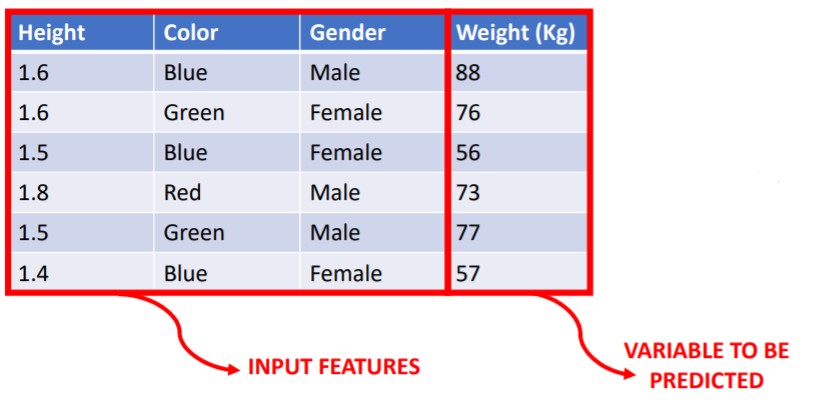
*Hình 2. 8 Mô hình các bước thuật toán XGBoost*

## 2.5 Thuật toán XGBoost

XGBoost hoạt động bằng cách xây dựng một cây dựa trên lỗi (phần còn lại) từ cây trước đó.

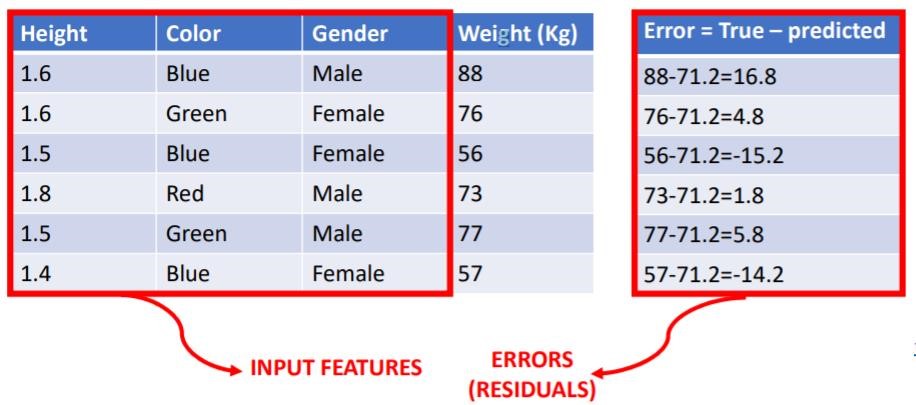
XGBoost chia thành các cây và sau đó thêm các dự đoán từ cây mới vào các dự đoán từ các cây trước đó

Ví dụ:



### *Hình 2. 9 Hình ví dụ thuật toán*

Giả sử rằng các dự đoán ban đầu của mô hình (điểm bắt đầu) là Weight trung bình là 71,2



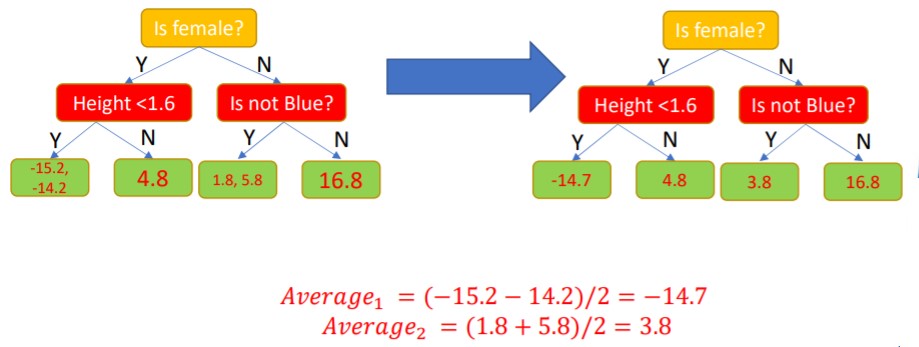
*2.10 Hình ví dụ minh họa thuật toán*

XGBoost xây dựng một cây dựa trên lỗi từ cây đầu tiên.

Cây được xây dựng bằng cách giả định rằng các đặc điểm (chiều cao, màu sắc và giới tính) dự đoán phần còn lại (cột mới mà chúng ta vừa tạo).

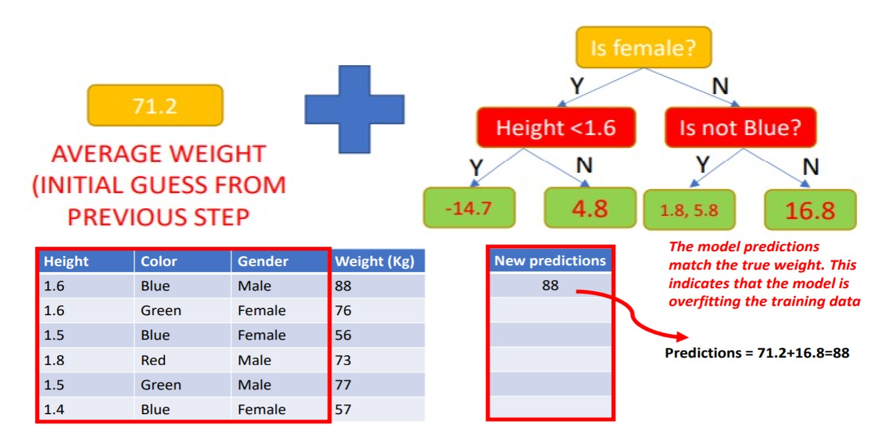
- Lưu ý rằng số lượng lá được giới hạn ở 4 trong ví dụ này để đơn giản

Thay thế các giá trị bằng giá trị trung bình được hiển thị bên dưới.



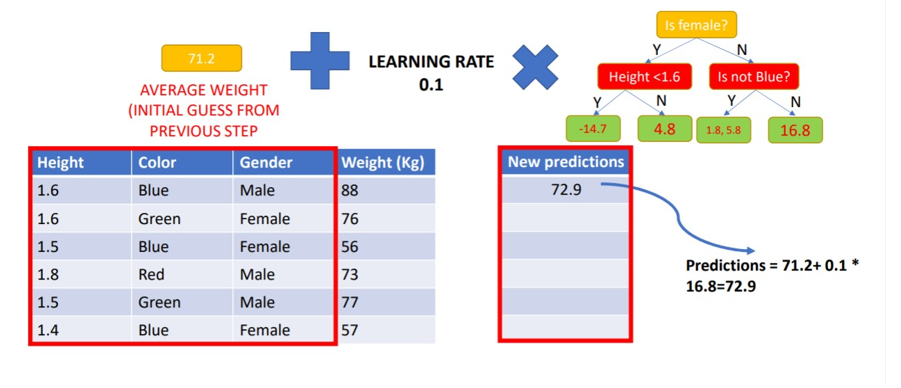
*Hình 2.11 Hình ví dụ về cây quyết định 1*

Bây giờ chúng ta đã xây dựng một cây, hãy kết hợp các dự đoán trước đó với cây mới để tạo ra các dự đoán mới

\

*Hình 2.12 Hình Ví dụ minh họa về cây quyết định 2*

Thêm learning rate để giải quyết vấn đề overfitting

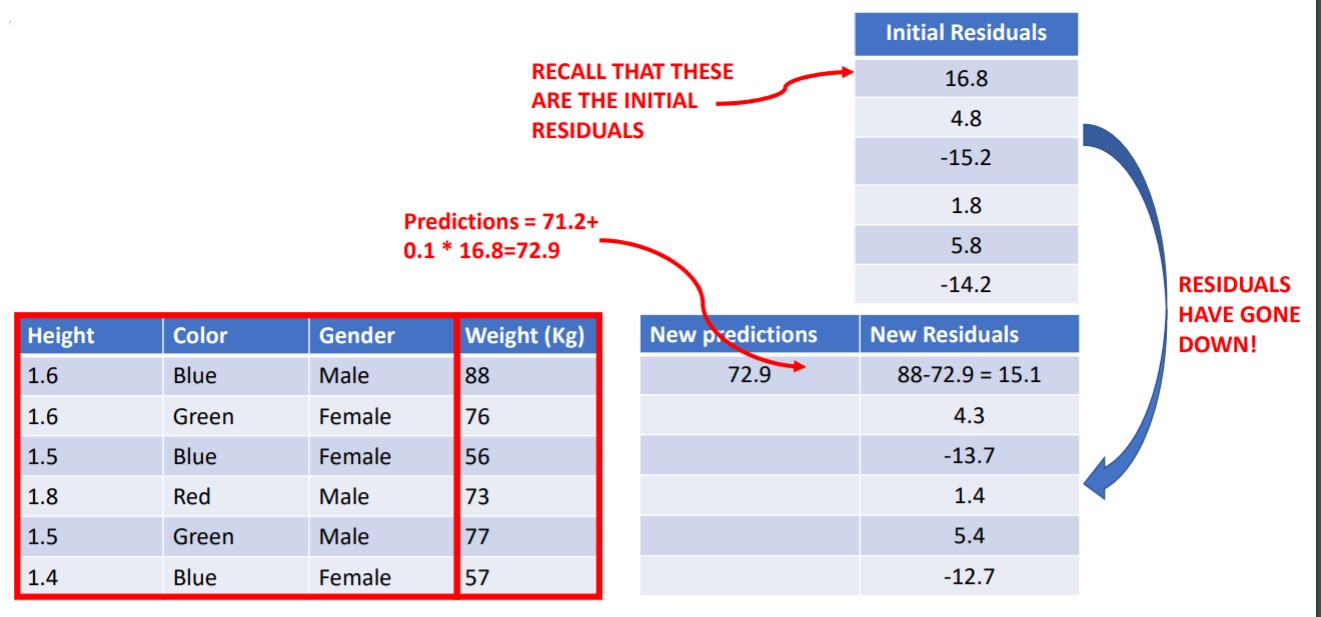


*Hình 2.12 Hình Ví dụ minh họa về cây quyết định 3*

Tham số này được sử dụng cho mục đích mở rộng quá trình đào bằng bằng cách điều chỉnh thông tin mới được thêm vào từ cây mới.

Việc thêm tham số learning rate và mở rộng quá trình đào tạo thì giúp tiến gần hơn đến giá trị thực.

Bằng cách thực hiện các bước nhỏ hơn, mô hình dẫn đến các dự đoán tốt hơn trên tập dữ liệu thử nghiệm (phương sai thấp).



Tiếp tục cho đến khi đặt tới số cây giới hạn.

**2.6 Ưu điểm và nhược điểm của thuật toán XGBoost**

Ưu điểm:

* Không cần thực hiện bất kì quy trình cân bằng nào
* Có thể làm việc tốt với những dữ liệu bị thất lạc
* Phân loại và hồi quy khá tốt
* Hiệu quả về tính toán và đưa ra các dự đoán nhanh Nhược điểm:
* Cần điều chỉnh rộng
* Tốn nhiều thời gian để học

# CHƯƠNG 3: ỨNG DỤNG CỦA THUẬT TOÁN

## 3.1 Các trường dữ liệu có trong dataset:

ID: ID của khách hàng

LIMIT\_BAL: Số dư trong thẻ tín dụng (đơn vị tiền là Đô la Đài Loan)

SEX: Giới tính (1=nam, 2=nữ)

EDUCATION: Trình độ học vấn (1 = cao hơn đại học, 2 = đại học, 3 = cấp 3, 4 = khác, 5 = không rõ, 6 = không rõ)

MARRIAGE: Tình trạng hôn nhân (1 = đã kết hôn, 2 = độc thân, 3 = khác)

AGE: Tuổi

PAY\_0: Trạng thái thanh toán vào tháng 9 năm 2005 (-1 = đã thanh toán, 1 = trễ 1 tháng, 2 = trễ 2 tháng, …)

PAY\_2: Trạng thái thanh toán vào tháng 8 năm 2005 (Giá trị giống như trên)

PAY\_3: Trạng thái thanh toán vào tháng 7 năm 2005 (Giá trị giống như trên)

PAY\_4: Trạng thái thanh toán vào tháng 6 năm 2005 (Giá trị giống như trên)

PAY\_5: Trạng thái thanh toán vào tháng 5 năm 2005 (Giá trị giống như trên)

PAY\_6: Trạng thái thanh toán vào tháng 4 năm 2005 (Giá trị giống như trên)

BILL\_AMT1: Sao kê của tháng 9 năm 2005 (Đô-la Đài Loan)

BILL\_AMT2: Sao kê của tháng 8 năm 2005 (Đô-la Đài Loan)

BILL\_AMT3: Sao kê của tháng 7 năm 2005 (Đô-la Đài Loan)

BILL\_AMT4: Sao kê của tháng 6 năm 2005 (Đô-la Đài Loan)

BILL\_AMT5: Sao kê của tháng 5 năm 2005 (Đô-la Đài Loan)

BILL\_AMT6: Sao kê của tháng 4 năm 2005 (Đô-la Đài Loan)

PAY\_AMT1: Số tiền đã thanh toán vào tháng 9 năm 2005 (Đô-la Đài Loan)

PAY\_AMT2: Số tiền đã thanh toán vào tháng 8 năm 2005 (Đô-la Đài Loan)

PAY\_AMT3: Số tiền đã thanh toán vào tháng 7 năm 2005 (Đô-la Đài Loan)

PAY\_AMT4: Số tiền đã thanh toán vào tháng 6 năm 2005 (Đô-la Đài Loan)

PAY\_AMT5: Số tiền đã thanh toán vào tháng 5 năm 2005 (Đô-la Đài Loan)

PAY\_AMT6: Số tiền đã thanh toán vào tháng 4 năm 2005 (Đô-la Đài Loan)

default.payment.next.month: Vỡ nợ (1 = đúng, 0 = sai)

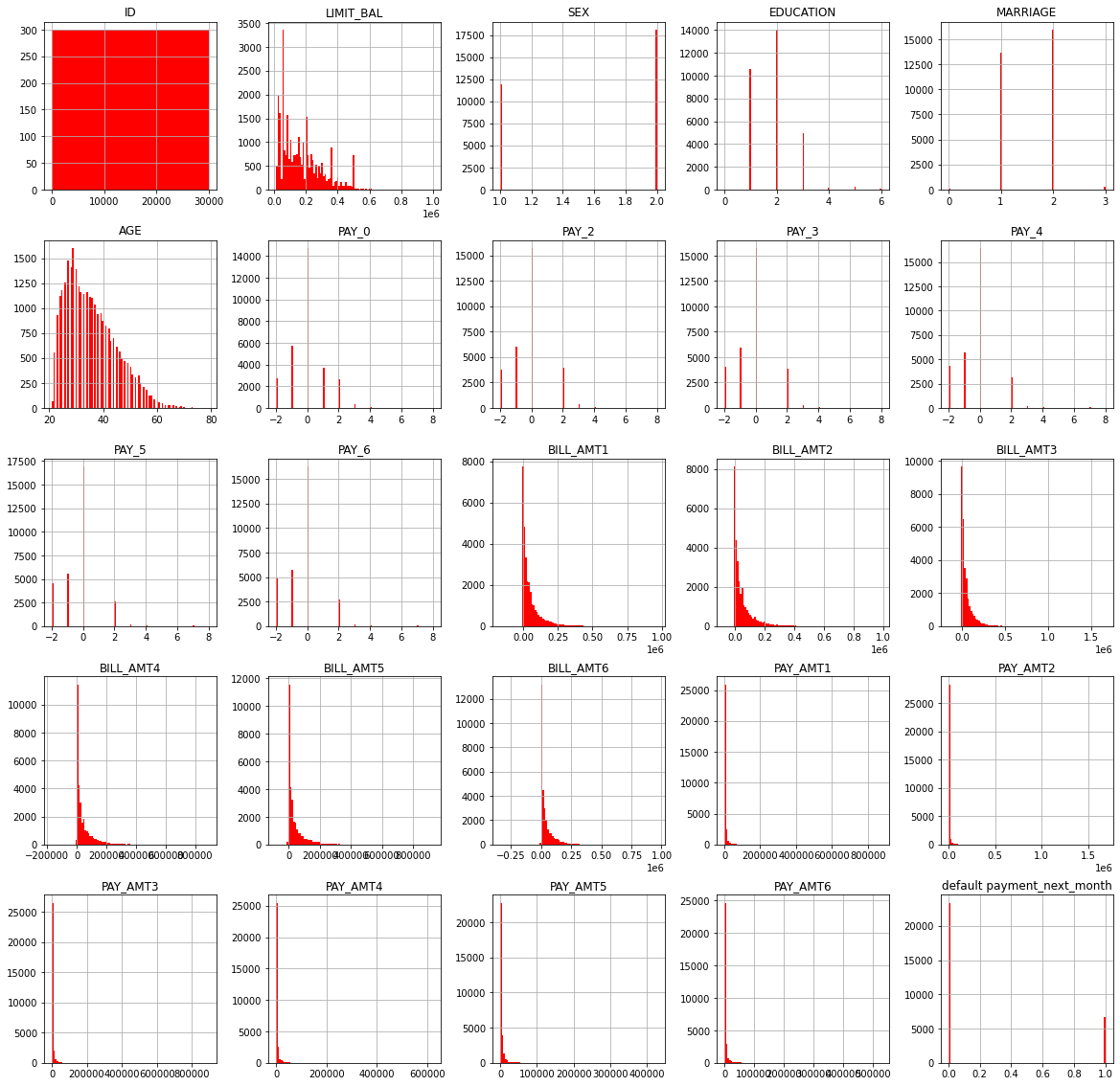
## 3.2 Đánh giá dữ liệu

- Kiểm tra missing data:



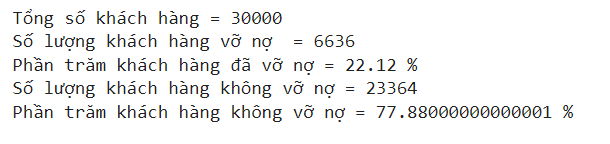
### *Hình 3. 1 Hình biểu nhiệt của missing data*

- Tổng quan dữ liệu



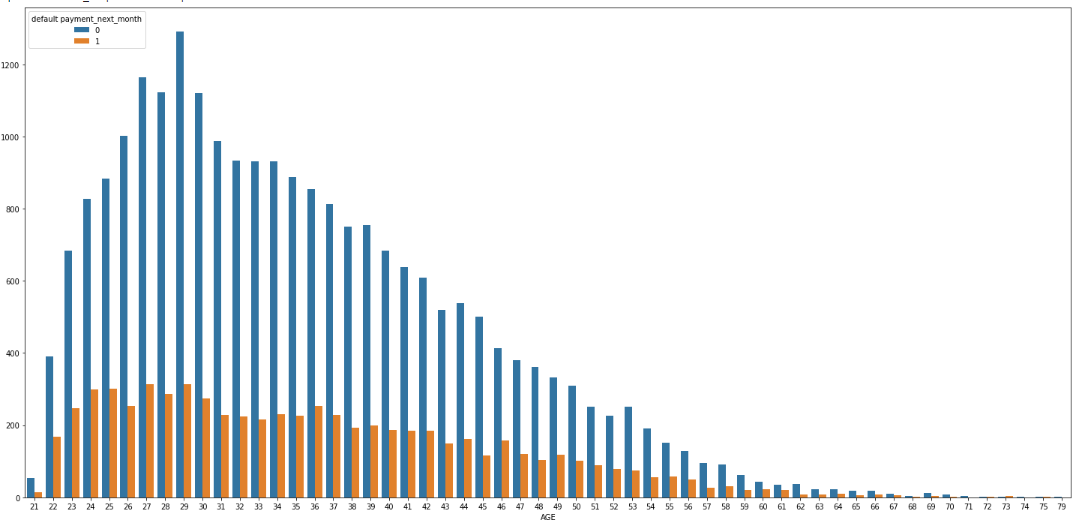
### *Hình 3. 2 Hình biểu đồ tổng quan dữ liệu*

* Danh sách bao nhiêu người vỡ nợ và không vỡ nợ:



*Hình 3.3 Hình tổng quan về thông số của khách hàng*

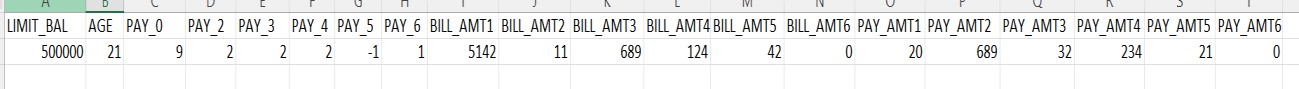
* Biểu đồ số lượng người trả được nợ dựa theo tuổi:



*Hình 3. 4 Hình biểu đồ người trả được nợ theo tuổi*

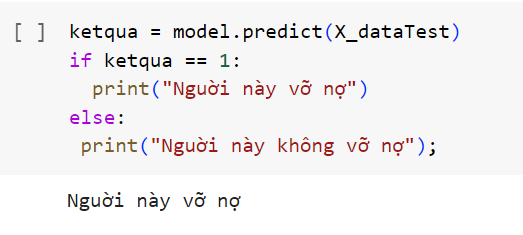
## 3.3 Kết quả

Với dữ liệu đầu vào cho trước:



### *Hình 3. 5 Hình dữ liệu đầu vào*

Chúng em đã dự đoán được kết quả như sau:



*Hình 3. 6 Hình kết quả dự đoán*

# KẾT LUẬN

Sau thời gian thực hiện, chúng em đã thực hiện được một số kết quả sau:

* Giải quyết được bài toán đặt ra.
* Tìm hiểu về thuật toán XGBoost.
* Xử lý cũng như nắm bắt được thông tin cơ bản của tệp dữ liệu.

Bài toán của bọn em vẫn còn những hạn chế như chưa có giao diện người dụng, ...

Với sự nỗ lực của chúng em và sự giúp đỡ, chỉ bảo tận tình của cô Nguyễn Thị Thanh Tân hướng dẫn, cuối cùng em cũng hoàn thành xong đề tài. Tuy vậy, với những thuận lợi và khó khăn trong quá trình làm việc, bài làm về cơ bản đã hoàn thành nhưng không thể tránh khỏi sai sót.

Trong thời gian tới, chúng em sửa đổi nhưng hạn chế và nâng cấp thêm đáp ứng nhu cầu thực tế.

# TÀI LIỆU THAM KHẢO

[1]. <https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset>

[2]. [https://viblo.asia/p/gradient-boosting-tat-tan-tat-ve-thuat-toan-manh-me-nhat-trong-machinelearning-YWOZrN7vZQ0](https://viblo.asia/p/gradient-boosting-tat-tan-tat-ve-thuat-toan-manh-me-nhat-trong-machine-learning-YWOZrN7vZQ0)

[3]. <https://pandas.pydata.org/docs/reference>

[4]. <https://xgboost.readthedocs.io/en/latest/tutorials/model.html>