

**TRƯỜNG ĐẠI HỌC ĐIỆN LỰC**  
**KHOA CÔNG NGHỆ THÔNG TIN**



**ĐẠI HỌC ĐIỆN LỰC**  
**ELECTRIC POWER UNIVERSITY**

**BÁO CÁO CHUYÊN ĐỀ HỌC PHẦN**  
**MÔN HỌC: KHAI PHÁ DỮ LIỆU**

**Đề tài:**

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

<b>Sinh viên thực hiện</b>	<b>: NGUYỄN DUY ĐẠT</b> <b>NGUYỄN ĐỨC THÀNH AN</b> <b>TRẦN ANH DUY</b>
<b>Giảng viên hướng dẫn</b>	<b>: TS. NGUYỄN THỊ THANH TÂN</b>
<b>Ngành</b>	<b>: CÔNG NGHỆ THÔNG TIN</b>
<b>Chuyên ngành</b>	<b>: CÔNG NGHỆ PHẦN MỀM</b>
<b>Lớp</b>	<b>: D14CNPM7</b>
<b>Khóa</b>	<b>: 2019-2023</b>

**PHIẾU CHẤM ĐIỂM**

STT	Họ và tên sinh viên	Nội dung thực hiện	Điểm	Chữ ký
1	Nguyễn Duy Đạt MSV: 19810310532			
2	Nguyễn Đức Thành An MSV: 19810310482			
3	Trần Anh Duy MSV: 19810310499			

Họ và tên giảng viên	Chữ ký	Ghi chú
Giảng viên chấm 1:		
Giảng viên chấm 2:		

# MỤC LỤC

<b>PHIẾU CHẤM ĐIỂM .....</b>	<b>.....</b>
<b>LỜI MỞ ĐẦU.....</b>	<b>.....</b>
<b>PHẦN 1: TÓM TẮT VỀ ĐỀ TÀI.....</b>	<b>1</b>
1.1    Đặt vấn đề.....	1
1.2    Cơ sở hình thành đề tài.....	1
1.3    Mục tiêu đề tài.....	2
1.4    Đối tượng và phương pháp nghiên cứu.....	3
1.5    Ý nghĩa đề tài .....	3
<b>PHẦN 2. THUẬT TOÁN XGBOOST.....</b>	<b>4</b>
2.1    Giới thiệu.....	4
2.2    Thuật toán XGBoost.....	4
2.3    Cây quyết định kết hợp (decision tree ensembles) .....	9
2.4    Các bước thuật toán XGBoost.....	10
2.5    Thuật toán XGBoost.....	11
2.6    Ưu điểm và nhược điểm của thuật toán XGBoost .....	15
<b>PHẦN 3: ỨNG DỤNG .....</b>	<b>16</b>
3.1    Các trường dữ liệu có trong dataset: .....	16
3.2    Đánh giá dữ liệu .....	17
3.3    Kết quả.....	19
<b>KẾT LUẬN.....</b>	<b>20</b>
<b>TÀI LIỆU THAM KHẢO .....</b>	<b>21</b>

# DANH MỤC HÌNH ẢNH

Hình 1. 1 Hình minh họa	2
Hình 2. 1: Phương pháp bagging	5
Hình 2. 2 Phương pháp boosting	6
Hình 2. 3 Model 1	7
Hình 2. 4 Model 2	7
Hình 2. 5 Model 3	8
Hình 2. 6 Hình cây quyết định theo tuổi	10
Hình 2. 7 Hình hai cây quyết định kết hợp	10
Hình 2. 8 Mô hình các bước thuật toán XGBoost	11
Hình 2. 9 Hình ví dụ thuật toán	11
Hình 2. 10 Hình ví dụ cây quyết định	12
Hình 2. 11 Hình ví dụ learning rate	14
Hình 3. 1 Hình biểu nhiệt của missing data	17
Hình 3. 2 Hình biểu đồ tổng quan dữ liệu	18
Hình 3. 3 Hình tổng quan thông số khách hàng	19
Hình 3. 4 Hình biểu đồ người trả được nợ theo tuổi	19
Hình 3. 5 Hình dữ liệu đầu vào	19
Hình 3. 6 Hình kết quả dự đoán	19

## LỜI MỞ ĐẦU

Trong thời đại ngày nay, yếu tố quyết định thành công trong mọi lĩnh vực luôn gắn liền với việc nắm bắt, thống kê và khai thác thông tin hiệu quả. Dữ liệu ngày càng lớn nên việc tìm ra những thông tin tiềm ẩn trong chúng càng khó khăn hơn.

Khai phá tri thức là một lĩnh vực nghiên cứu mới, mở ra một thời kỳ trong việc tìm ra thông tin hữu ích. Nhiệm vụ cơ bản của lĩnh vực này là khai phá tri thức trong cơ sở dữ liệu, khai phá dữ liệu trong cơ sở dữ liệu không phải là một hệ thống phân tích tự động mà là một quá trình tương tác thường xuyên giữa con người với cơ sở dữ liệu được sự trợ giúp của nhiều phương pháp và công cụ tin học.

Chúng em xin bày tỏ sự biết ơn sâu sắc của mình tới TS Nguyễn Thị Thanh Tân, người đã trực tiếp hướng dẫn, chỉ bảo tận tình, cung cấp tài liệu và phương pháp nghiên cứu khoa học để chúng em hoàn thành bản luận văn này. Chúng em xin gửi lời cảm ơn tới các thầy cô giáo đã dạy dỗ trong quá trình chúng em theo học tại trường Đại học Điện Lực.

Trong suốt quá trình nghiên cứu, mặc dù đã hết sức cố gắng nhưng chắc chắn bài luận không tránh khỏi những thiếu sót, rất mong quý thầy cô góp ý để bài báo cáo kết thúc học phần môn học “Khai phá dữ liệu” được hoàn chỉnh hơn.

Chúng em xin chân thành cảm ơn!

# PHẦN 1: TÓM TẮT VỀ ĐỀ TÀI

## 1.1 Đặt vấn đề

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 bảng biểu và khai phá dữ liệu.

Khai phá dữ liệu là quá trình phân loại, sắp xếp các tập hợp dữ liệu lớn để xác định các mẫu và thiết lập các mối liên hệ nhằm giải quyết các vấn đề nhờ phân tích dữ liệu. Các MCU khai phá dữ liệu cho phép các doanh nghiệp có thể dự đoán được xu hướng tương lai. Khai phá dữ liệu là một kỹ thuật dựa trên nền tảng của nhiều lý thuyết như xác suất, 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. Nguồn dữ liệu người dùng ngân hàng rất lớn, nếu áp dụng khai phá dữ liệu trong lĩnh vực này sẽ mang lại nhiều ý nghĩa cho ngành kinh tế. Nó sẽ cung cấp những thông tin quý giá nhằm hỗ trợ trong việc giải quyết các vấn đề mà ngân hàng đang gặp phải trong quá trình.

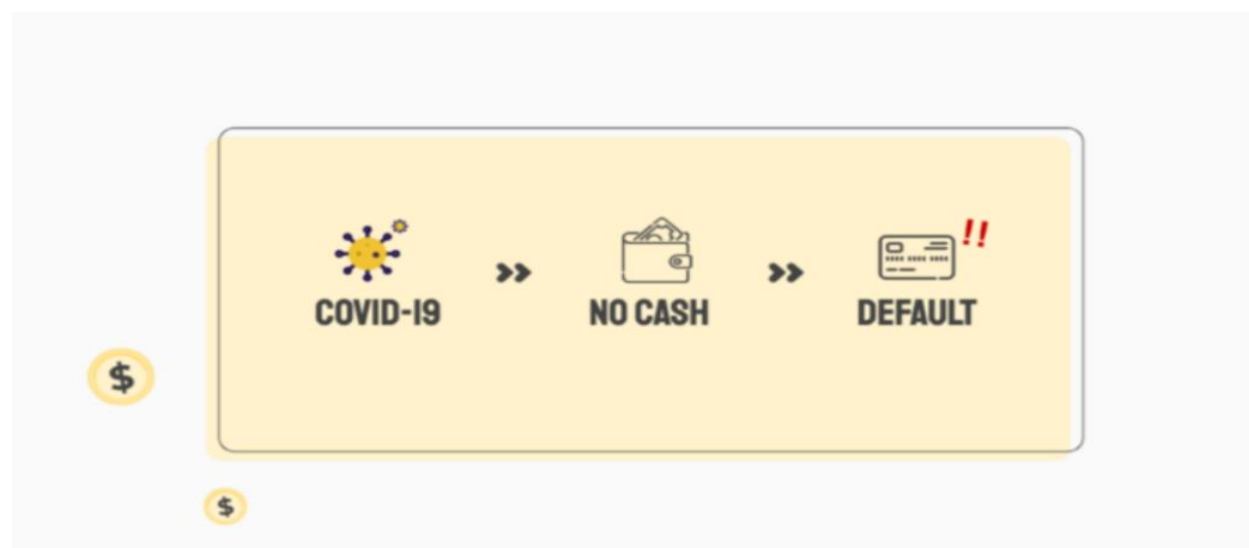
## 1.2 Cơ sở hình thành đề tài

JPMorgan Chase, Wells Fargo, Bank of America, Citigroup và Goldman Sachs đã tăng tổng số tiền dành cho các khoản nợ xấu lên gần 20 tỷ USD trong quý đầu tiên, báo cáo thu nhập được công bố trong hai ngày qua cho thấy.

Bank of America và Citigroup hôm thứ Tư cho biết lợi nhuận của họ giảm hơn 40% trong quý đầu tiên do cả hai đều dành hàng tỷ USD cho các khoản nợ khó đòi. Trước đó một ngày, JPMorgan Chase và Wells Fargo báo cáo lợi nhuận thậm chí còn sụt giảm mạnh hơn do các ngân hàng này cũng dành ra một khoản tiền lớn để bù lỗ cho các khoản vay. Doanh số bán lẻ đã giảm 8,7% trong tháng 3, mức giảm hàng tháng tồi tệ nhất được ghi nhận. Chi tiêu của người tiêu dùng chiếm khoảng 70% tổng sản phẩm quốc nội của Hoa Kỳ, vì vậy sự sụt giảm như vậy là đặc biệt khó khăn. Câu hỏi duy nhất là mức độ nghiêm trọng như thế nào: Tổng sản phẩm quốc nội quý II dự kiến sẽ giảm từ 30% đến 40% và tỷ lệ thất nghiệp có thể tăng cao tới 25%.

Do COVID-19, nhiều người đã mất việc làm, dẫn đến những người thiếu tiền mặt và mặc định thanh toán bằng thẻ tín dụng của họ. Mọi thứ trở nên tồi tệ đến mức các công ty thẻ tín dụng, chẳng hạn như JP Morgan và Citigroup, phải dành thêm một khoản dự phòng để bù đắp tổn thất do thẻ tín dụng bị vỡ nợ. Bây giờ, đây là một trường hợp rất nghiêm trọng, không thường xuyên xảy ra (tôi thực sự hy vọng là không).

Mọi người không có khả năng thanh toán cho các hóa đơn thẻ tín dụng của họ có thể do các trường hợp khác nhau. Tuy nhiên, khi đó là hành vi cố ý, nghĩa là khách hàng không có kế hoạch trả lại tiền cho ngân hàng, thì hành vi đó sẽ bị coi là gian lận. Dù bằng cách nào, điều này cũng tiềm ẩn rủi ro rất lớn đối với các công ty phát hành thẻ tín dụng và chúng ta cần phải tìm cách để đánh dấu họ.



*Hình 1. 1 Hình minh họa*

Để giải quyết vấn đề này, chúng tôi có thể dự đoán các tài khoản vỡ nợ tiềm năng dựa trên các thuộc tính nhất định. Ý tưởng là càng phát hiện sớm các tài khoản vỡ nợ tiềm ẩn, thì tổn thất mà chúng ta phải gánh chịu càng thấp. Mặt khác, chúng tôi có thể chủ động bằng cách cung cấp các thủ thuật cho khách hàng để tránh vỡ nợ. Điều này không chỉ bảo vệ khách hàng của chúng tôi mà còn giảm thiểu rủi ro và tổn thất có thể xảy ra.

### **1.3 Mục tiêu đề tài**

Đề tài tập trung vào nghiên cứu kỹ thuật cây quyết định trong khai phá dữ liệu, từ đó nắm bắt được những giải thuật làm tiền đề cho nghiên cứu và xây dựng ứng dụng cụ thể. Sau khi phân

tích đặc điểm của dữ liệu thu nhập được và lựa chọn giải thuật phù hợp với dữ liệu, việc xây dựng và đánh giá chất lượng, độ hiệu quả của hệ thống cũng là mục tiêu chính của đề tài. Để dự đoán liệu khách hàng có vỡ nợ khi thanh toán thẻ tín dụng của họ vào tháng tới hay không.

#### **1.4 Đối tượng và phương pháp nghiên cứu**

Đề tài tập trung vào nghiên cứu kỹ thuật cây quyết định và thuật toán XGBoost trong khai phá dữ liệu vào việc phân tích cơ sở dữ liệu tỷ lệ người dùng vỡ nợ ( hay còn gọi là không trả được số tiền đã nợ trong ngân hàng). Bộ dữ liệu, được lấy từ Kho lưu trữ Máy học UC Irvine, chứa thông tin về các khách hàng sử dụng thẻ tín dụng ở Đài Loan. Nó bao gồm 30.000 quan sát và 24 tính đặc điểm của người dùng.

#### **1.5 Ý nghĩa đề tài**

Với sự trợ giúp của máy tính, đề tài đóng góp một biện pháp thực hiện hỗ trợ các ngân hàng hiện nay, đánh giá được người dùng sử dụng thẻ tín dụng. Kết quả, Kinh nghiệm thu được khi thực hiện đề tài này sẽ giúp các ngân hàng phát hiện ra các người dùng gian lận hoặc không thể trả được nợ, đồng thời mong muốn những người đang công tác trong lĩnh vực về kinh tế, ngân hàng và Khoa học máy tính ngồi lại với nhau để tìm ra những giải pháp tốt hơn để giải quyết trong vấn đề bằng cách kết hợp giữa 2 lĩnh vực kinh tế và khoa học máy tính.



## PHẦN 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 bảng biểu và khai phá dữ liệu.

Khai phá dữ liệu là một kỹ thuật dựa trên nền tảng của nhiều lý thuyết như xác suất, 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 đích phân lớp (classification) và hồi quy (regression) dữ liệu.

### 2.2 Thuật toán XGBoost

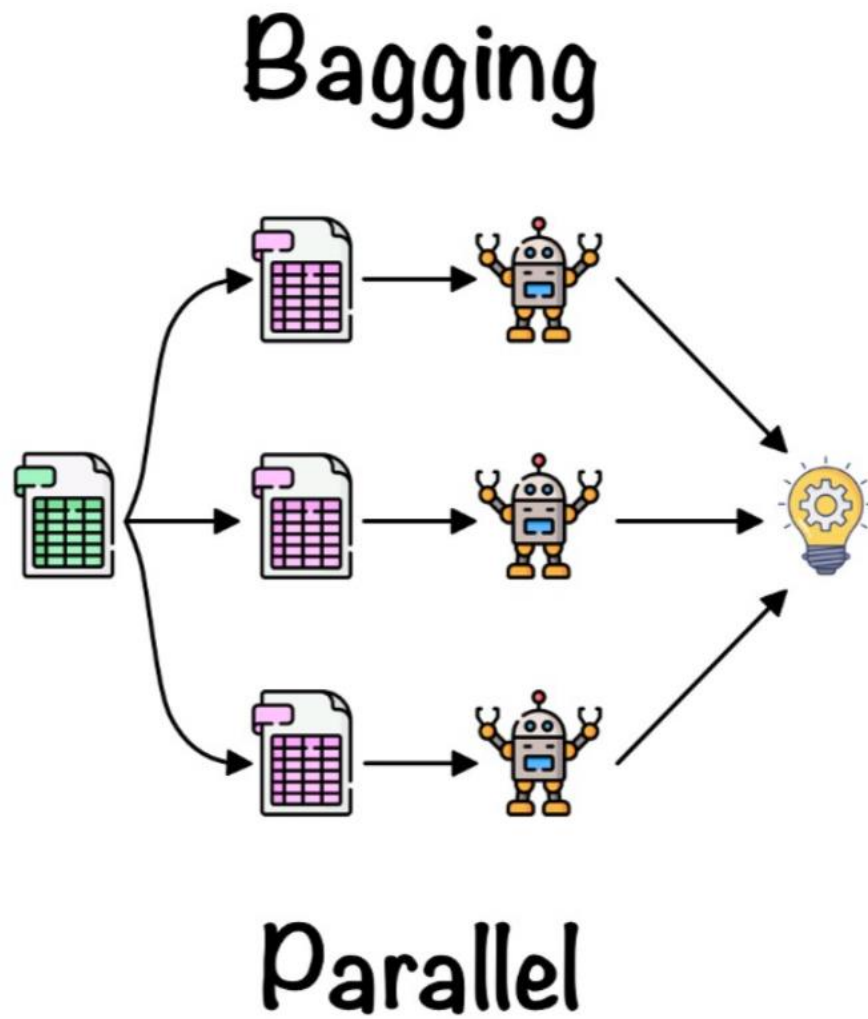
#### 2.2.1 Ensemble learning là gì?

- 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:

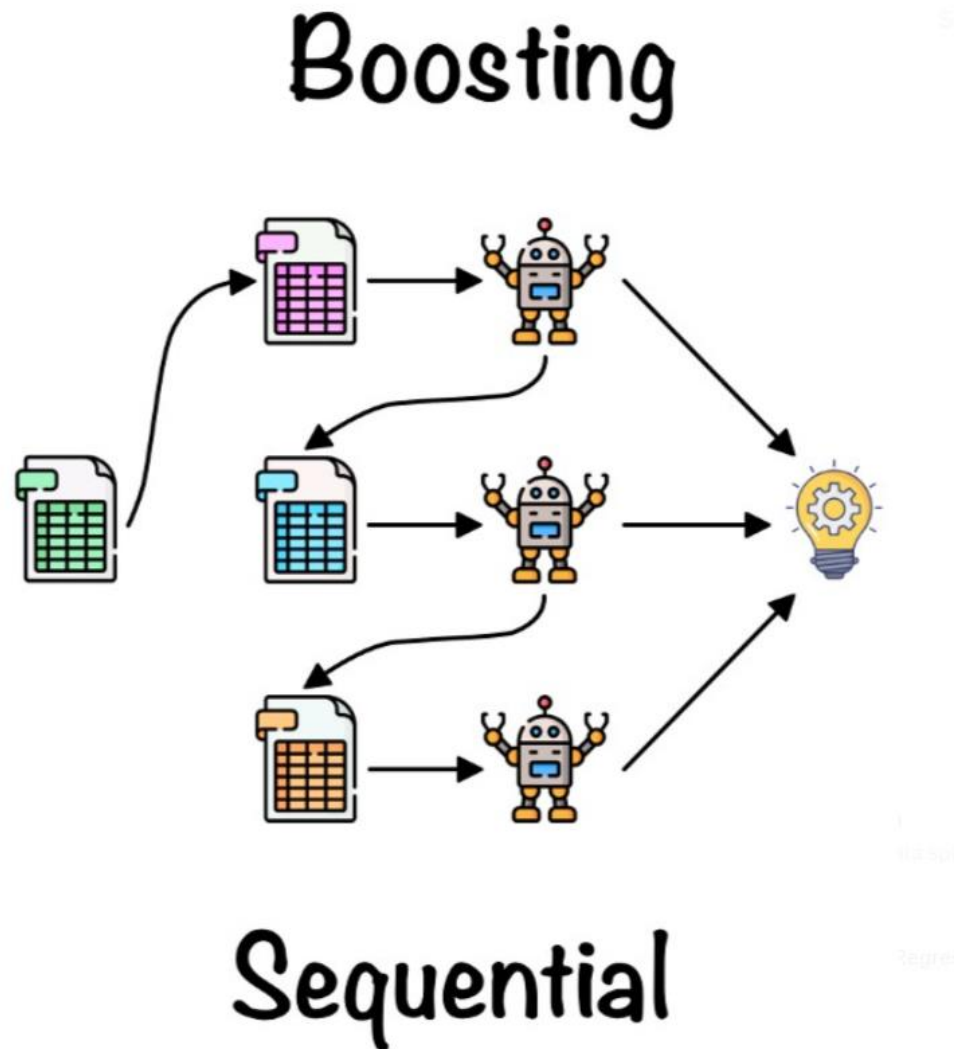
- Bằng cách Tổng hợp các phiếu bầu từ một nhóm các chuyên gia, mỗi chuyên gia sẽ mang lại kinh nghiệm và nền tảng của riêng mình để giải quyết vấn đề dẫn đến kết quả tốt hơn.
- Thuật toán Ensemble sẽ sử dụng phương pháp Bagging và Boosting để kết hợp các mô hình yếu lại với nhau.

### 2.2.2 Phương pháp Bagging



Hình 2. 1: Phương pháp bagging

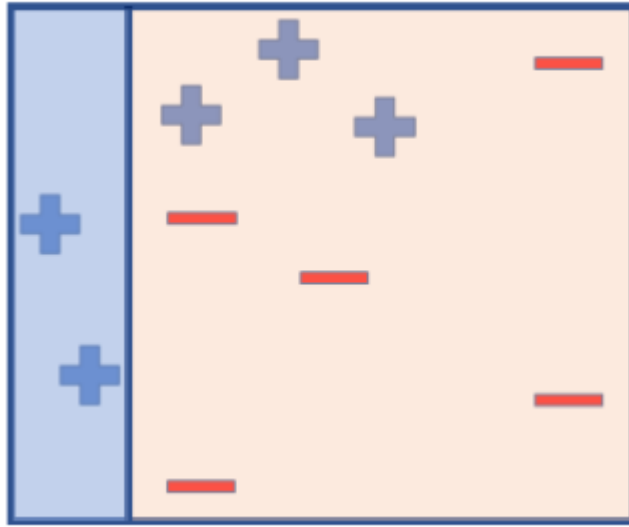
### 2.2.3 Phương pháp Boosting



Hình 2. 2 Phương pháp boosting

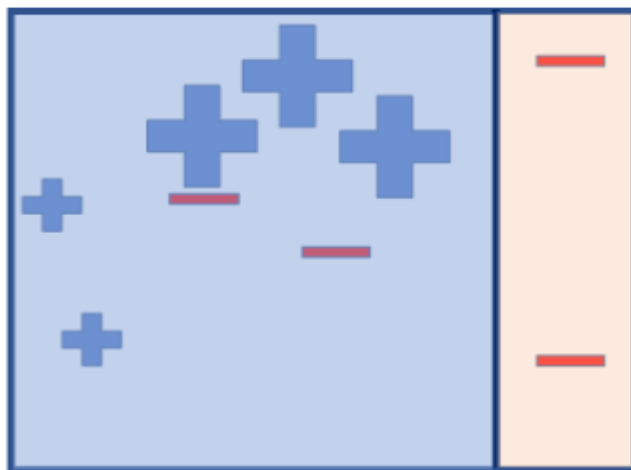
- Boosting là một kỹ thuật hoạt động bằng cách đào tạo (training) các mô hình yếu theo cách tuần tự.
- Mỗi model đang cố gắng học hỏi từ model yếu trước đó và trở nên tốt hơn trong việc đưa ra dự đoán.
- Thuật toán lặp lại cho đến khi số lượng mô hình tối đa được tạo hoặc cho đến khi mô hình cung cấp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

Ví dụ về boosting:



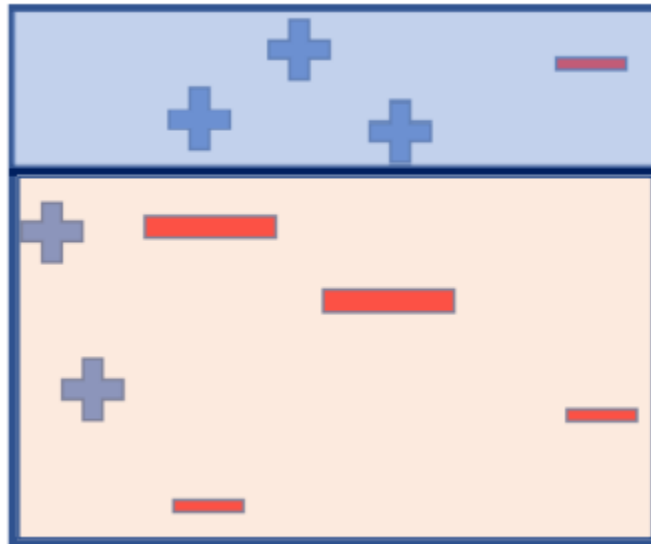
Hình 2. 3 Model 1

- 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 mẫu (+)



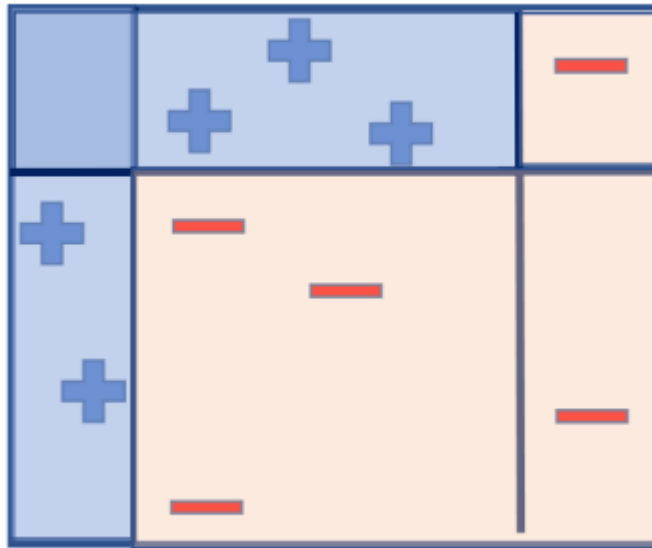
Hình 2. 4 Model 2

- 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 được phân 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.



- 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 (decision tree ensembles)

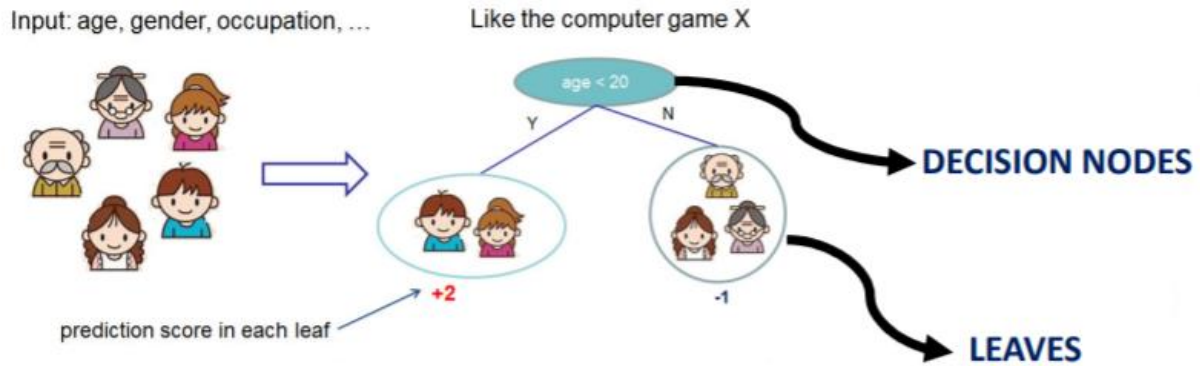
Cây quyết định là kỹ thuật Máy học giám sát (Supervised learning) trong đó dữ liệu được phân chia theo một điều kiện / tham số nhất định.

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?

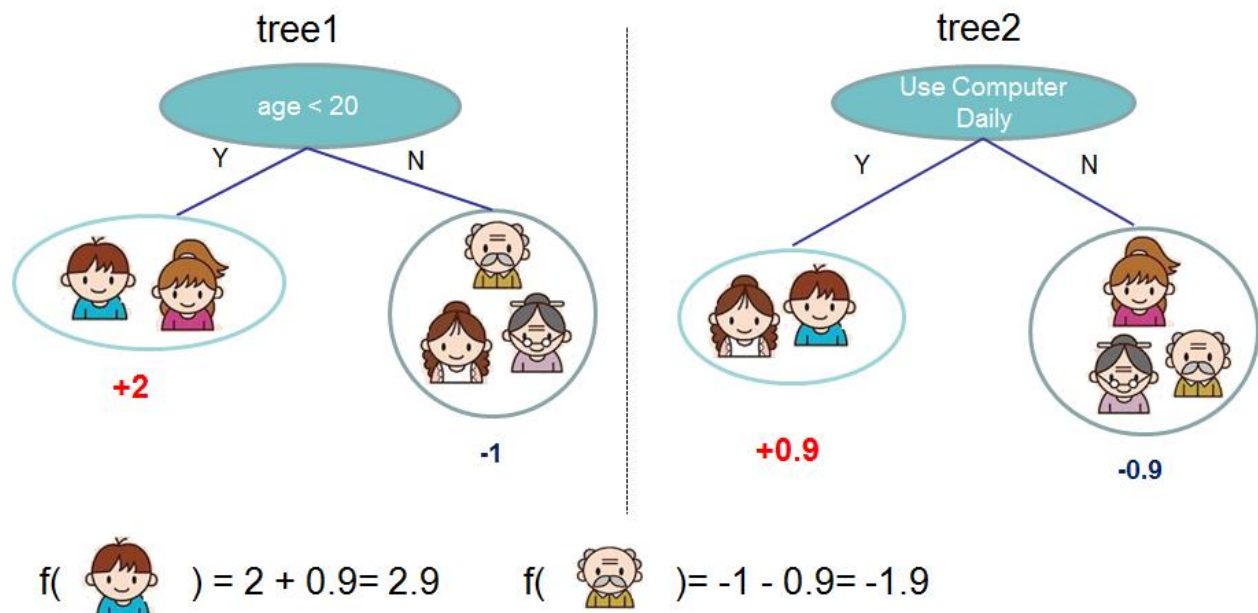
Các thành viên sẽ được chia đến các lá và có các số điểm cho trước nhất định



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

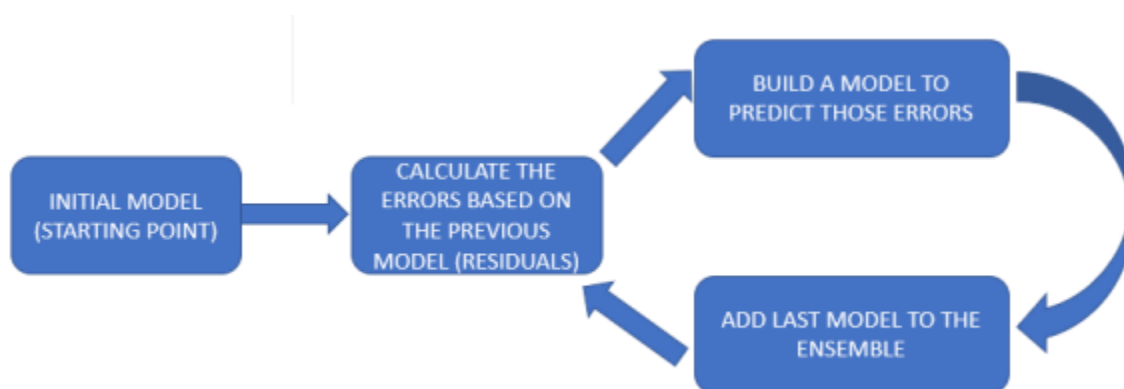
Điểm dự đoán của từng cây riêng lẻ được tổng hợp lên để có được điểm số cuối cùng.

## 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:

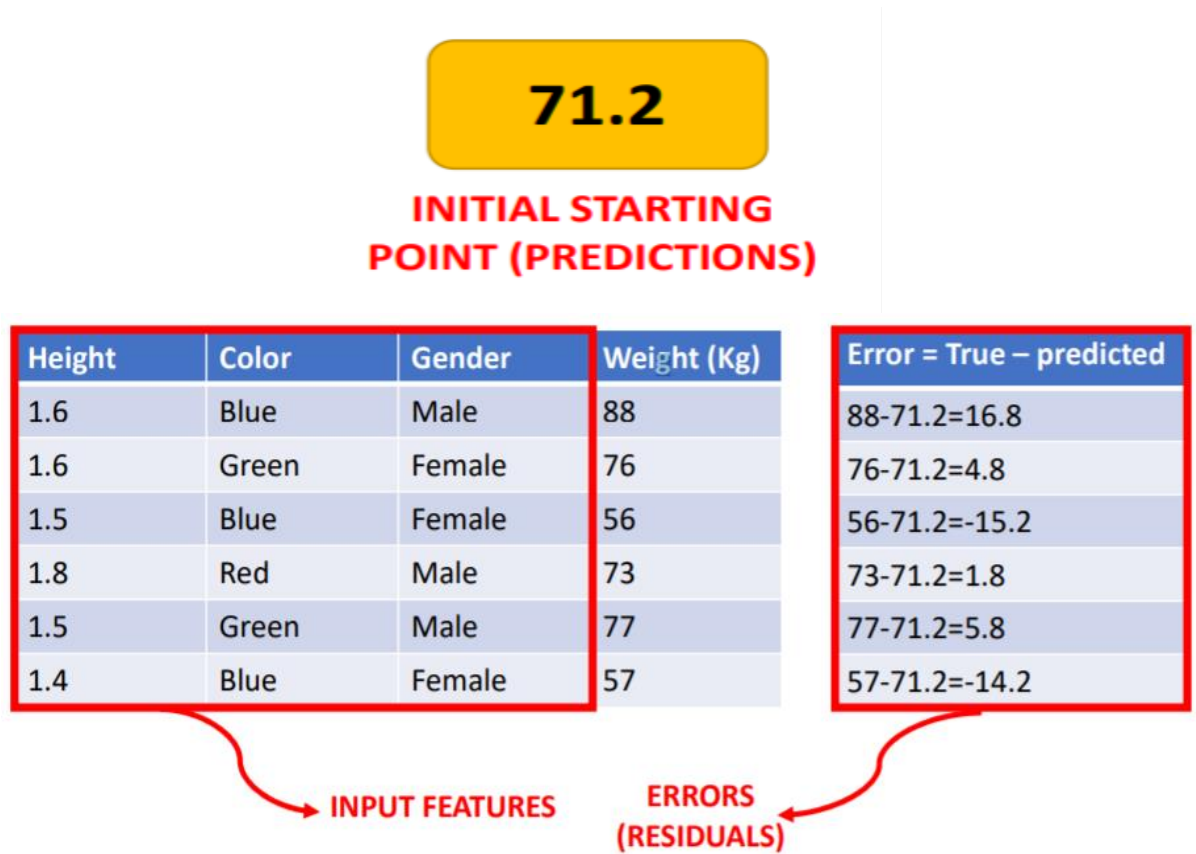
Height	Color	Gender	Weight (Kg)
1.6	Blue	Male	88
1.6	Green	Female	76
1.5	Blue	Female	56
1.8	Red	Male	73
1.5	Green	Male	77
1.4	Blue	Female	57

→ INPUT FEATURES
→ VARIABLE TO BE PREDICTED

Hình 2. 9 Hình ví dụ thuật toán

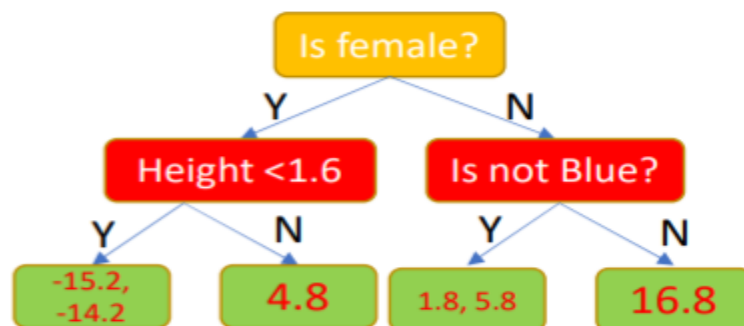


Giả sử rằng các dự đoán ban đầu của mô hình (điểm bắt đầu) là Weight trung bình là 71,2



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).



Hình 2. 10 Hình ví dụ cây quyết định

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

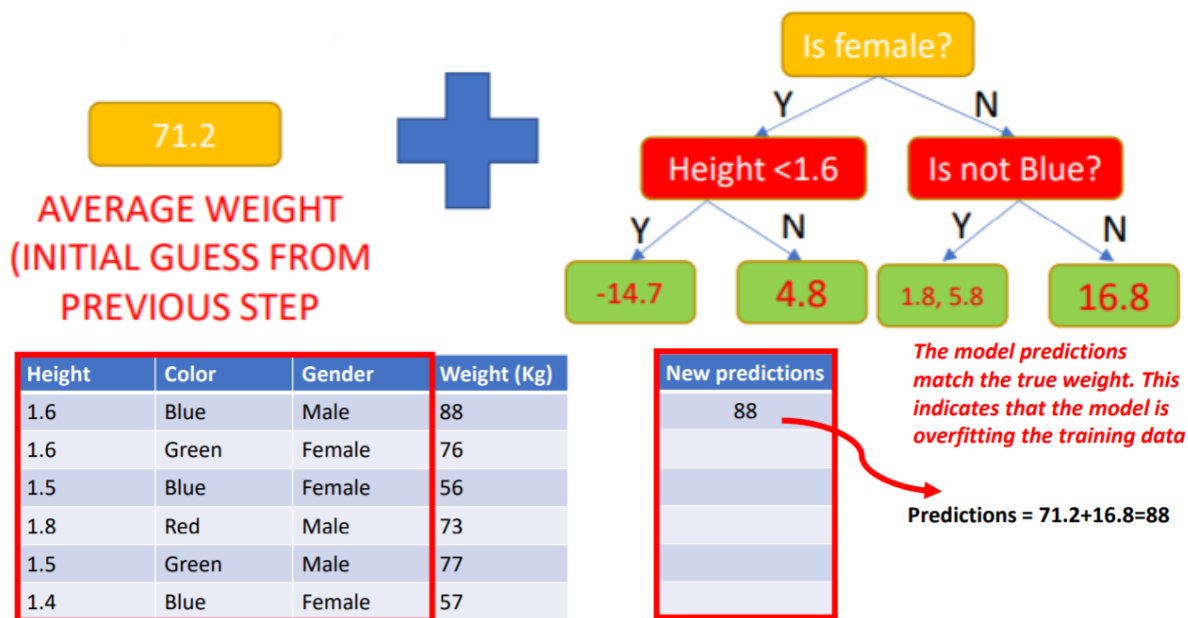
Thay thế các giá trị bằng giá trị trung bình được hiển thị bên dưới.



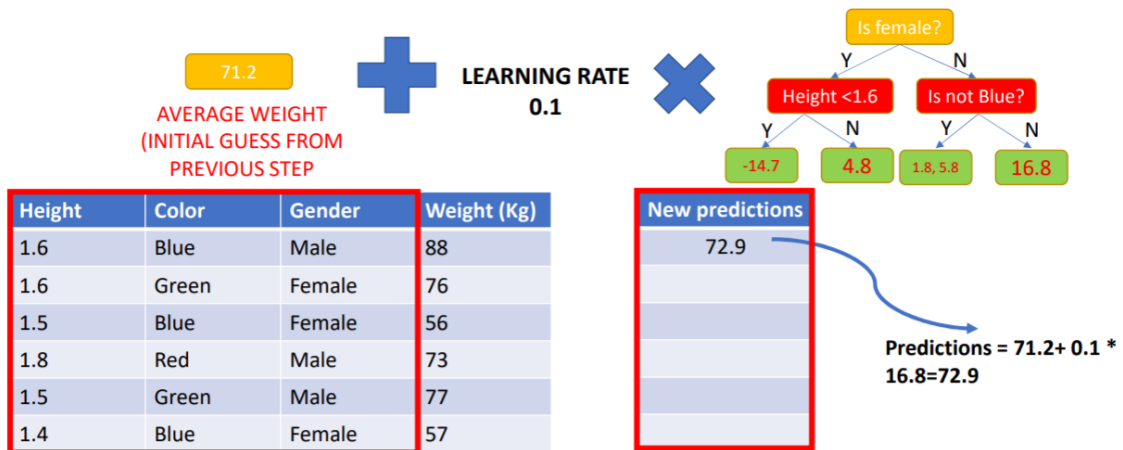
$$Average_1 = (-15.2 - 14.2)/2 = -14.7$$

$$Average_2 = (1.8 + 5.8)/2 = 3.8$$

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



Thêm learning rate để giải quyết vấn đề overfitting

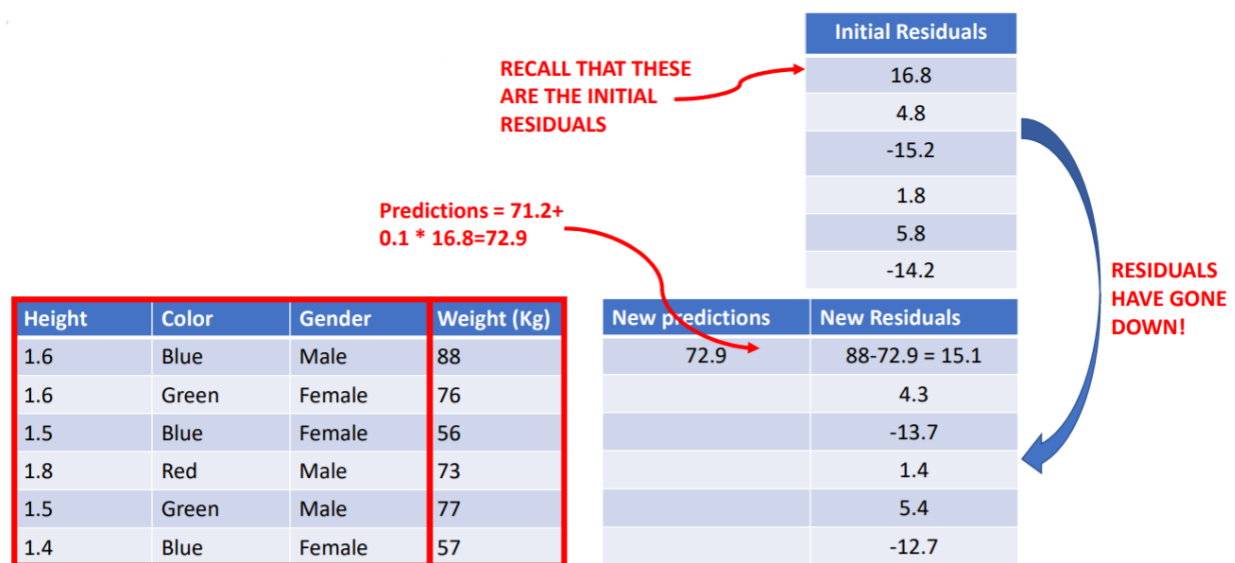


Hình 2. 11 Hình ví dụ learing rate

Tham số này được sử dụng cho mục đích mở rộng quá trình đào tạo 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 đạt 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

## PHẦN 3: ỨNG DỤNG

### 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)

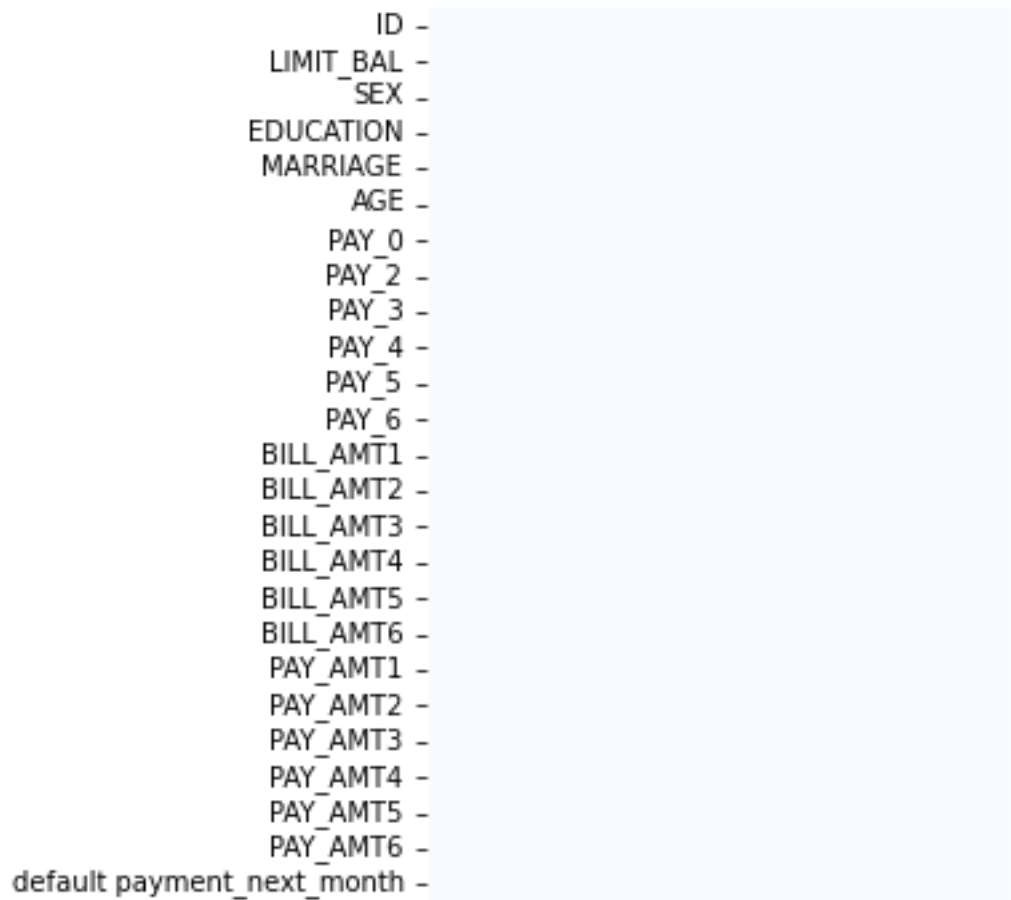
PAY\_AMT5: Số tiền đã thanh toán vào tháng 5 năm 2005 (Đô-la Đài Loan)

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

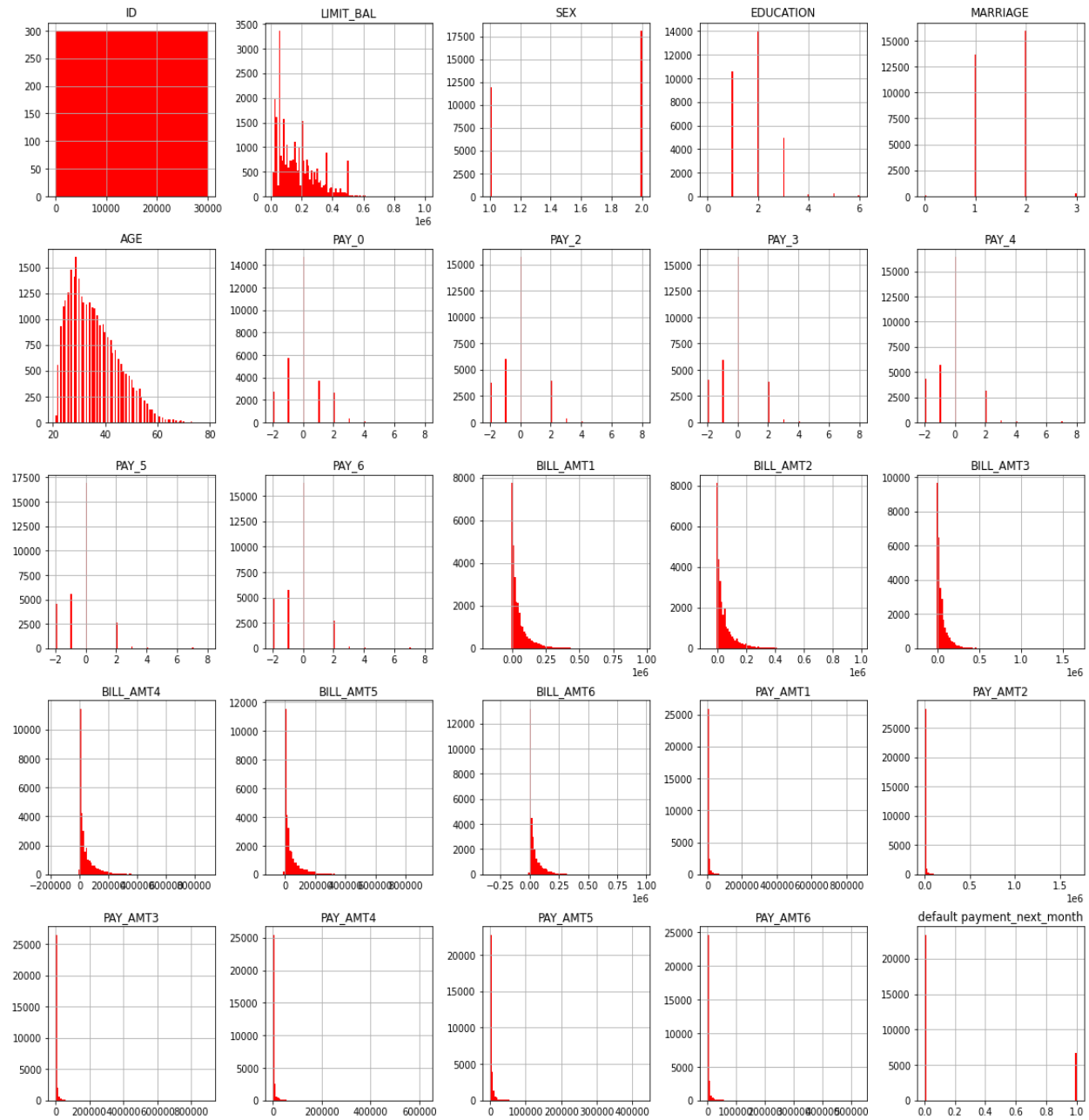
- Kiểm tra missing data:



The figure is a heatmap visualization showing the distribution of missing data across various variables. The variables listed on the left are: ID, LIMIT\_BAL, SEX, EDUCATION, MARRIAGE, AGE, PAY\_0, PAY\_2, PAY\_3, PAY\_4, PAY\_5, PAY\_6, BILL\_AMT1, BILL\_AMT2, BILL\_AMT3, BILL\_AMT4, BILL\_AMT5, BILL\_AMT6, PAY\_AMT1, PAY\_AMT2, PAY\_AMT3, PAY\_AMT4, PAY\_AMT5, PAY\_AMT6, and default payment\_next\_month. The heatmap area to the right of these labels is mostly light blue, indicating that there is no missing data for these variables. The variable 'default payment\_next\_month' at the bottom shows a distinct pattern of missing data, represented by a darker blue area.

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

- Tổng quan dữ liệu



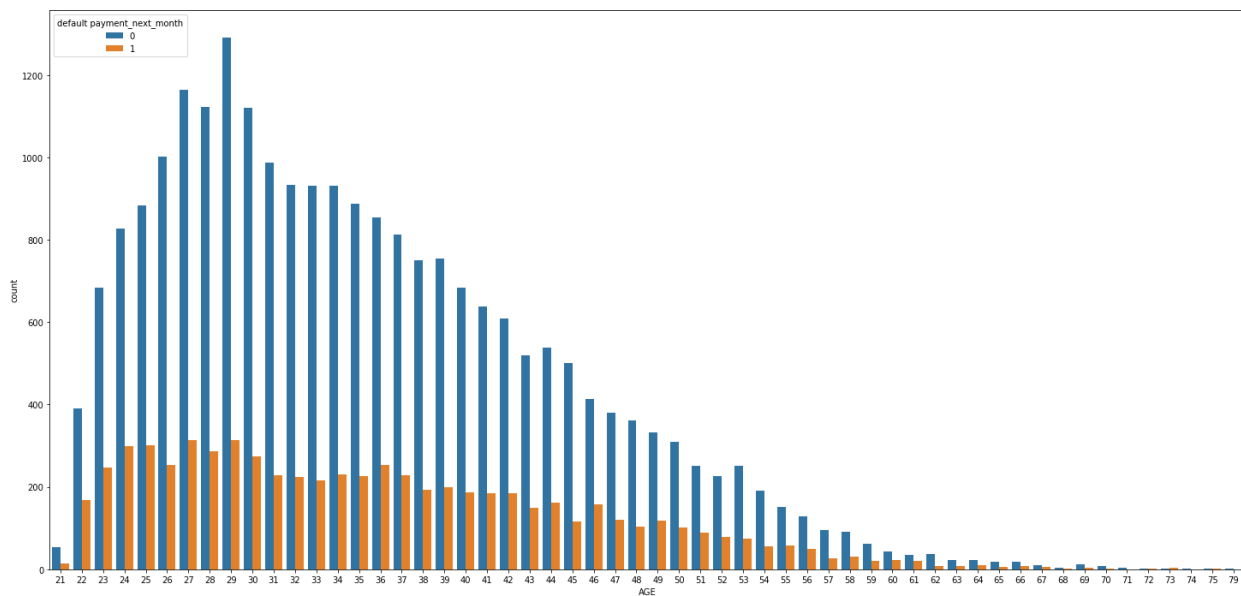
Hình 3. 2 Hình biểu đồ tổng quan dữ liệu

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

Tổng số khách hàng = 30000  
 Số lượng khách hàng vỡ nợ = 6636  
 Phần trăm khách hàng đã vỡ nợ = 22.12 %  
 Số lượng khách hàng không vỡ nợ = 23364  
 Phần trăm khách hàng không vỡ nợ = 77.88000000000001 %

Hình 3. 3 Hình tổng quan thông số 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:

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
LIMIT_BAL	AGE	PAY_0	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6
500000	21	9	2	2	2	-1	1	5142	11	689	124	42	0	20	689	32	234	21	0

Hình 3. 5 Hình dữ liệu đầu vào

Chúng em đã dự đoán được kết quả như sau:

Người này vỡ nợ

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 dùng, ...

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 những hạn chế và nâng cấp thêm đáp ứng nhu cầu thực tế.

## TÀI LIỆU THAM KHẢO

- [1]. Các tài liệu tham khảo của cô Nguyễn Thị Thanh Tân
- [2]. <https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset>
- [3]. <https://viblo.asia/p/gradient-boosting-tat-tan-tat-ve-thuat-toan-manh-me-nhat-trong-machine-learning-YWOZrN7vZQ0>
- [4]. <https://pandas.pydata.org/docs/reference>
- [5]. <https://xgboost.readthedocs.io/en/latest/tutorials/model.html>