Phân tích thăm dò bộ dữ liệu Concrete Compressive Strength*

Phan Vỹ Hào[1], Nguyễn Hoài Bảo[2], Trần Xuân Phú[3], Trần Đình Nam[4], and TS. Đỗ Trọng Hợp[5]

Khoa Khoa học và Kỹ thuật Thông tin
ĐH Công nghệ Thông tin - ĐH Quốc Gia HCM
TP. Hồ Chí Minh, Việt Nam
19520524@gm.uit.edu.vn[1], 19520405@gm.uit.edu.vn[2],
19520843@gm.uit.edu.vn[3], 19520758@gm.uit.edu.vn[4], hopdt@uit.edu.vn[5]

Tóm tắt nội dung Trong bài báo cáo này, chúng tôi sẽ trình bày về quá trình phân tích về bộ dữ liệu, mối tương quan giữa các thuộc tính để có thể chọn ra các yếu tố tốt nhất nhằm xây dựng mô hình hồi quy dự đoán cường độ nén của bê tông cường độ cao dựa trên bộ dữ liệu *Concrete Compressive Strength* được lấy từ Kho lưu trữ học máy của UCI. Ngoài ra, chúng tôi cũng sẽ trình bày chi tiết quy trình chọn lọc thuộc tính, xây dựng mô hình và cải tiến để đạt kết quả tốt nhất, cũng như cách chúng tôi sử dụng các biện pháp trực quan lẫn thống kê để đánh giá mô hình.

1 Giới thiệu đồ án

Bê tông cường độ cao hay High-Performance Concrete (HPC) là một khái niệm mới trong ngành công nghiệp bê tông xây dựng. Để tạo ra bê tông cường độ cao không chỉ cần ba nguyên liệu cơ bản tạo nên bê tông thông thường (xi măng, nước, cốt liệu thô và mịn) mà còn cần vật liệu kết tính như tro bay (fly ash), xỉ lò cao (blast furnace slag) và phụ gia như chất siêu dẻo (superplasticizer). Vì những lí do trên, việc dự đoán độ cứng của bê tông cường độ cao rất khó. [7]

Định luật Abrams về sự đối nghịch giữa tỉ lệ nước-xi măng và độ cứng của bê tông ra đời năm 1918 được cho là một phát kiến vĩ đại và hữu dụng nhất trong lịch sử công nghệ sản xuất bê tông. Theo định luật này, khi tỉ lệ nước giảm thì độ cứng của bê tông tăng lên và ngược lại. Tức, khi so sánh các loại bê tông khác nhau (như bê tông thường và bê tông cường độ cao) chỉ dựa trên tỉ lệ nước-xi măng mà không cần quan tâm đến thành phần. Một vài nghiên cứu cho thấy độ cứng của bê tông được xác định không chỉ dựa trên tỉ lệ nước-xi măng mà còn bởi các thành phần khác.

Mặc dù định luật Abrams vẫn đúng trong hầu hết trường hợp nhưng vẫn có những trường hợp ngoại lệ được ghi nhận vì công thức chỉ đề cập tới các loại bê tông thông thường mà thành phần không có vật liệu kết dính như tro bay hay

* Giảng viên hướng dẫn: TS. Đỗ Trọng Hợp

xỉ lò cao. Đây là một thiếu sót vì với sự ra đời và ứng dụng rộng rãi của bê tông cường độ cao (HPC) hay bê tông cường độ siêu cao (UHPC), việc hiểu rõ thành phần rồi từ đó dự đoán được độ cứng của bê tông là cần thiết nếu muốn tối ưu hóa, tạo ra loại bê tông mới trong tương lai. [7]

Độ cứng đặc trưng của bê tông được xác định dựa trên khả năng chịu lực khi nó đã đông đặc 28 ngày. Như vậy, việc tính toán sẽ tốn một khoảng thời gian dài và để đẩy nhanh tiến độ nghiên cứu thì việc dự đoán độ cứng của bê tông chỉ với thông tin các thành phần cơ bản của nó là rất cần thiết - đây là lúc học máy được áp dụng. Sử dụng bộ dữ liệu đã chọn, ta huấn luyện mô hình học máy. Sau huấn luyện, bằng việc đưa vào thông tin của các thuộc tính tạo nên loại bê tông, ta có thể dự đoán độ cứng với sai số cho phép.

Về mục tiêu của đồ án:

- Nắm được cách xử lí dữ liệu và phân tích trực quan.
- Sử dụng ANOVA để phân tích định lượng, tìm các yếu tố quan trọng ảnh hưởng đến yếu tố đầu ra.
- Xây dựng nhiều mô hình dự đoán dựa trên các yếu tố vừa lọc được nhằm tìm ra mô hình cho kết quả tốt nhất.

Về nội dung của đồ án lần này, chúng tôi sẽ thực hiện EDA (Exploratory Data Analysis) hay còn gọi là *Phân tích dữ liệu thăm dò* trên bộ dữ liệu *Concrete Compressive Strength Data Set* [7]. Phân tích dữ liệu thăm dò gồm 2 bước chính là phân tích trực quan và phân tích định lượng.

Đối với phân tích trực quan, chúng tôi tiến hành:

- Biểu diễn dữ liệu dưới dạng biểu đồ hộp, biểu đồ cột và biểu đồ hexbin.
- Quan sát biểu đồ, nhận xét sự phân bố các điểm dữ liệu, đặt ra giả thuyết và kết luận.
- Sử dụng ma trận tương quan để xem xét mối quan hệ, sự tương tác giữa các thuộc tính với nhau, từ đó loại bỏ những thuộc tính có mức ảnh hưởng rất thấp đến thấp. Các thuộc tính còn lại sẽ được xem xét cẩn thận ở bước phân tích định lượng.

Với phân tích định lượng:

- Dùng ANOVA để phân tích, tìm ra các thuộc tính ảnh hưởng đến kết quả đầu ra.
- Có những thuộc tính khi đứng riêng lẻ sẽ không có ý nghĩa nên phải thử kết hợp các thuộc tính với nhau theo từng nhóm.
- Sau mỗi lần loại bỏ thuộc tính, chúng tôi sẽ tiến hành phân tích ANOVA với các thuộc tính còn lại.

Bước cuối cùng, chúng tôi sẽ xây dựng và huấn luyện nhiều mô hình hồi quy với đầu ra là độ chịu lực của bê tông và so sánh chúng với nhau để rút ra đc mô hình phù hợp nhất bằng cánh sử dụng các thuộc tính đã phân tích kĩ lưỡng.

Tiếp sau đây, ở phần 2 chúng tôi sẽ mô tả chi tiết hơn về bộ dữ liệu. Nắm được các thuộc tính của bộ dữ liệu, chúng tôi tiến hành phân tích trực quan ở phần 3 để tìm xem ảnh hưởng của từng yếu tố cũng như lọc ra các thuộc tính

không cần thiết. Ở phần 4, chúng tôi tiến hành phân tích định lượng với các yếu tố còn lại nhằm tìm ra những yếu tố thực sự ảnh hưởng tới yếu tố đầu ra. Từ kết quả sàng lọc ở hai phần 3 và 4, chúng tôi tiến hành huấn luyện mô hình hồi quy với các yếu tố mà chúng tôi cho là có ảnh hưởng, đánh giá kết quả của mô hình và chọn ra mô hình tốt nhất. Cuối cùng, xem xét lại toàn bộ đồ án nhằm đúc kết những thành tựu mà nhóm đã đạt được, những khó khăn mà nhóm gặp phải trong quá trình phát triển đồ án cũng như hướng phát triển của đồ án sau này.

2 Bộ dữ liệu

2.1 Nguồn gốc bộ dữ liệu

Bộ dữ liệu *Concrete Compressive Strength* được thực hiện bởi giáo sư I-Cheng Yeh thuộc trường Đại học Trung Hoa và được công bố cộng đồng vào ngày 3 tháng 8 năm 2007 [8].

2.2 Mô tả bộ dữ liệu

Bảng 1 trình bày tóm tắt các thông tin cơ bản về bộ dữ liệu của nhóm. Ngoài ra, chúng tôi cũng thực hiện viết tắt tên cho các thuộc tính (trình bày ở bảng 2) nhằm thống kê dễ dàng hơn trên môi trường `R`, mà ở các phần sau, chúng tôi sẽ trình bày kết quả và số liệu tương ứng với tên viết tắt.

Thông tin	**Nội dung**			
Tên bộ dữ liệu	Concrete Compressive Strength Data Set			
Số thuộc tính	9 Trong đó có 8 thuộc tính đầu vào, 1 thuộc tính đầu ra.			
Số lượng điểm dữ liệu	1030			
Thông tin tên các thuộc tính	STT	Tên thuộc tính	Ý nghĩa	Đơn vị
	1	Cement	Xi măng	kg/m^3
	2	Blast Furnace Slag	Xỉ lò cao nghiền mịn	kg/m^3
	3	Fly Ash	Tro bay	kg/m^3
	4	Water	Nước	kg/m^3
	5	Superplasticizer	Chất siêu hóa dẻo	kg/m^3
	6	Coarse Aggregate	Cốt liệu thô	kg/m^3
	7	Fine Aggregate	Cốt liệu mịn	kg/m^3
	8	Age	Thời gian kể từ lúc trộn	ngày
	9	Concrete Compressive Strength	Cường độ nén bê tông	MPa
Tác giả	Giáo sư I-Cheng Yeh Khoa Quản lý Thông tin Đại học Trung Hoa - Đài Loan			

Bảng 1: Bảng mô tả tóm tắt bộ dữ liệu.

STT	Thuộc tính gốc	Tên viết tắt
1	Cement	Cement
2	Blast Furnace Slag	Blast
3	Fly Ash	FlyAsh
4	Water	Water
5	Superplasticizer	SupPlas
6	Coarse Aggregate	CoarseAgg
7	Fine Aggregate	FineAgg
8	Age	age
9	Concrete compressive strength	strength

Bảng 2: Danh mục viết tắt các thuộc tính trong thống kê R.

2.3 Thống kê mô tả dữ liệu

Ở bảng 3, chúng tôi thực hiện thống kê mô tả cho bộ dữ liệu này, nhằm có cái nhìn tổng quát trước khi đi vào phân tích từng thành phần yếu tố. Trong đó:

- **n**: Số lượng điểm dữ liệu hợp lệ (không NA hoặc Null).
- **mean**: Giá trị trung bình của thuộc tính, được tính bằng tổng giá trị chia cho số lượng phần tử của tập dữ liệu.
- **sd**: Độ lệch chuẩn của thuộc tính, thể hiện sự biến thiên của dữ liệu. Với x là mỗi điểm dữ liệu trong tập dữ liệu, độ lệch chuẩn được tính bằng công thức: $\sqrt{\frac{\sum|x-mean|^2}{n}}$.
- **median**: Điểm trung vị, còn gọi là điểm bách phân vị 50%, nơi chia tập dữ liệu thành hai phần bằng nhau.
- **trimmed**: Viết tắt của trimmed mean, là giá trị trung bình được điều chỉnh sau khi loại bỏ một tỉ lệ phần trăm nhỏ của giá trị lớn nhất và nhỏ nhất. Giá trị này tránh được tình trạng ảnh hưởng tiêu cực của giá trị ngoại lệ trong tập dữ liệu.
- **mad**: viết tắt của *Median Absolute Deviation* (trung vị của độ lệch tuyệt đối), là một thước đo chính xác hơn median, hạn chế được ảnh hưởng của giá trị ngoại lệ.
- **min**: Giá trị nhỏ nhất của tập dữ liệu.
- **max**: Giá trị lớn nhất của tập dữ liệu.
- **range**: Khoảng cách giữa giá trị **min** và **max**.
- **skew**: Độ lệch thể hiện sự không đối xứng trong phân phối xác suất của một biến.
  - $-0.5 \leq skew \leq 0.5$: Dữ liệu tương đối đối xứng (phân phối chuẩn).
  - $-1 \leq skew \leq -0.5 \mid 0.5 \leq skew \leq 1$: Dữ liệu có độ lệch vừa phải.
  - $skew \notin (-1; 1)$: Dữ liệu có độ lệch rất lớn.
- **kurtosis**: Độ nhọn. Giá trị này dùng để quan sát vùng đuôi và các giá trị cực trị của một phân bố.
  - $Kurtosis = 3$ (Mesokurtic): Các giá trị cực trị phân bố theo phân phối chuẩn, không có ngoại lệ.

  - $Kurtosis > 3$ (Leptokurtic): Các cực trị mỏng và cao bất thường, vùng đuôi ít biến đổi.
  - $Kurtosis < 3$ (Platykurtic): Đỉnh phẳng và phân tán lớn.
- **se**: Viết tắt cho *Standard Error* (sai số chuẩn).

	Cement	**Blast**	**FlyAsh**	**Water**	**SupPlas**	**CoarseAgg**	**FineAgg**	**age**	**strength**
n	1030	1030	1030	1030	1030	1030	1030	1030	1030
mean	281.17	73.90	54.19	181.57	6.2	972.92	773.58	45.66	35.82
sd	104.51	86.28	64	21.36	5.97	77.75	80.18	63.17	16.71
median	272.9	22	0	185	6.35	968	779.51	28	34.44
trimmed	273.47	62.43	46.85	181.19	5.56	973.49	776.41	32.53	34.96
mad	117.72	32.62	0	19.27	7.87	68.64	67.44	31.13	16.2
min	102	0	0	121.75	0	801	594	1	2.33
max	540	359.4	200.1	247	32.2	1145	992.6	365	82.6
range	438	359.4	200.1	125.25	32.2	344	398.6	364	80.27
skew	0.51	0.8	0.54	0.07	0.91	-0.04	-0.25	3.26	0.42
kurtosis	-0.53	-0.52	-1.33	0.11	1.39	-0.61	-0.11	12.07	-0.32
se	3.26	2.69	1.99	0.67	0.19	2.42	2.5	1.97	0.52

Bảng 3: Thống kê mô tả các thuộc tính.

3 Phân tích trực quan dữ liệu

3.1 Biểu đồ hộp

Vai trò và mục đích Theo [1], biểu đồ hộp (box plot) là một công cụ tiền xử lý phổ biến đối với các dữ liệu định tính. Nó giúp chúng ta quan sát được vùng trung tâm và biến thiên của dữ liệu, đặc biệt là để phát hiện sự khác biệt về biến thiên dữ liệu trong các giữa các nhóm khác nhau.

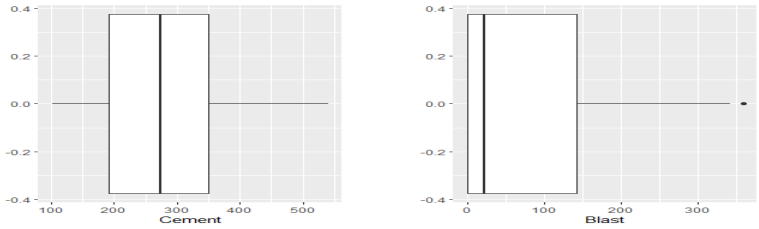
Loại biểu đồ này cung cấp các thông tin sau:

- Điểm trung vị (bách phân vị 50%) của dữ liệu được đánh dấu bằng một ký hiệu hoặc đường thẳng chính giữa hộp.
- Khoảng cách từ bách phân vị 25% đến bách phân vị 75% được gói vào một hộp vuông.
- Hai chiếc râu bên ngoài hộp là những điểm nằm trong dữ liệu nhưng không phải ngoại lệ.
- Những điểm dữ liệu ngoại lệ được thể hiện bằng những đốm tròn nhỏ phía ngoài chiếc râu.

Những người làm dữ liệu thường dựa vào loại biểu đồ này để khai phá những câu hỏi sau:

- Yếu tố này có vai trò đối với kết quả không?
- Các nhóm con có sự khác biệt về trung tâm của dữ liệu không?
- Các nhóm con có sự khác biệt về biến thiên không?
- Có giá trị ngoại lệ nào không?

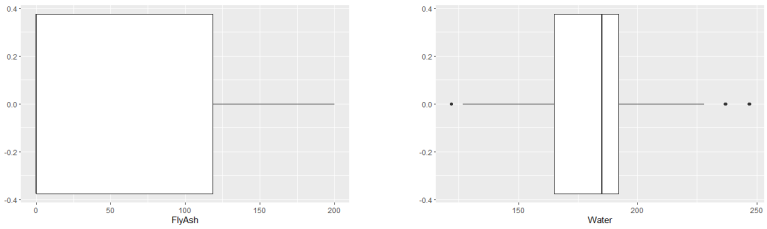
Phân tích dữ liệu với biểu đồ hộp



(a) Yếu tố Cement. (b) Yếu tố Blast Furnace Slag.

Hình 1: Biểu đồ hộp với hai yếu tố Cement và Blast Furnace Slag.

Hình 1 thể hiện biểu đồ hộp với với hai thuộc tính Cement và Blast Furnace Slag. Yếu tố Cement có phân bố dữ liệu hướng về chính giữa, trong khi yếu tố Blast Furnace Slag có độ lệch rất lớn khi phân bố tập trung về gần giá trị 0, và có giá trị ngoại lệ.

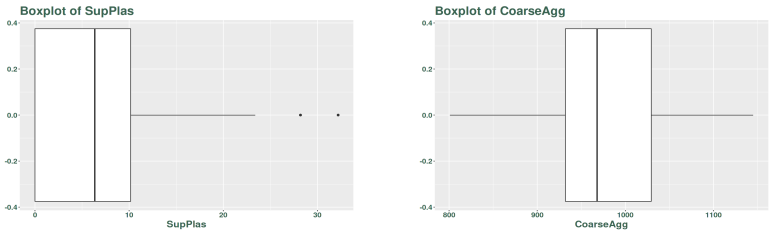


(a) Yếu tố FlyAsh. (b) Yếu tố Water.

Hình 2: Biểu đồ hộp với hai yếu tố FlyAsh và Water.

Hình 2a cho ta thấy trung vị của yếu tố FlyAsh bằng 0, tức phần lớn điểm dữ liệu mang giá trị 0, phân bố nghiên về phía bên và không có điểm ngoại lệ.

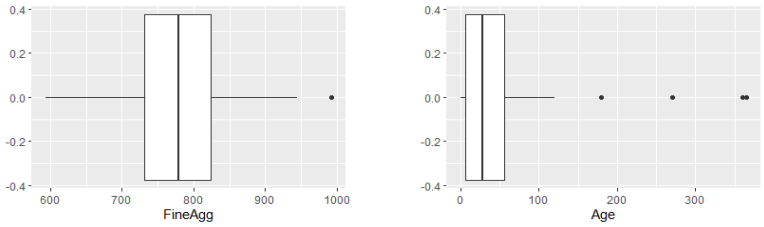
Hình 2b thể hiện yếu tố Water. Phân bố dữ liệu tuy không mất cân bằng như FlyAsh nhưng không đồng đều, có các điểm ngoại lệ ở hai đầu biểu đồ.



(a) Yếu tố SupPlas. (b) Yếu tố CoarseAgg.

Hình 3: Biểu đồ cột với hai yếu tố SupPlas và CoarseAgg.

Hình 3 là biểu đồ hộp của hai yếu tố SupPlas và CoarseAgg, cả hai biểu đồ này đều cho thấy rằng sự phân bố dữ liệu của hai thuộc tính hay không được cân đối, tuy nhiên ở yếu tố SupPlas, ta thấy được có một vài điểm ngoại lệ với giá trị chênh lệch khá lớn so với các giá trị chung.



(a) Yếu tố Fine Aggregate. (b) Yếu tố Age.

Hình 4: Biểu đồ hộp với hai yếu tố Fine Aggregate và Age.

Hình 4a có giá trị trung bình nằm trong khoảng 780 và các điểm dữ liệu trải đều sang 2 bên, ít điểm ngoại lệ.

Hình 4b cho thấy dữ liệu tập trung rất nhiều trong khoảng giá trị từ 0 đến 50, và có xuất hiện nhiều điểm ngoại lệ nằm ở bên phải hộp.

3.2 Biểu đồ cột

Vai trò và mục đích

Theo [1], biểu đồ cột (histogram) có mục đích chính là tóm tắt phân bố của dữ liệu bằng hình ảnh đối với dữ liệu đơn biến. Biểu đồ cột cho ta thấy các thông tin:

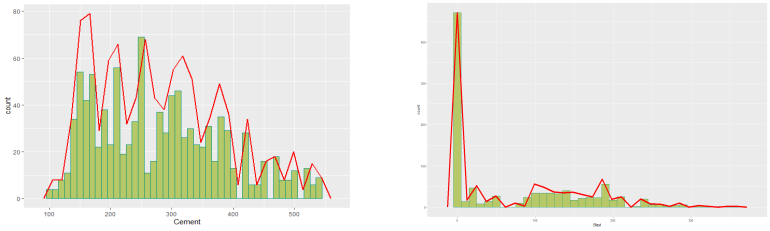
1. Trung tâm và độ trải rộng (spread) của dữ liệu.
2. Độ lệch (skewness) của dữ liệu.
3. Các điểm ngoại lệ (outlier).

4. Những điểm phổ biến (mode) của dữ liệu.

Biểu đồ dạng này có thể giúp trả lời các câu hỏi nghiên cứu sau:

- Dữ liệu có dạng phân bố nào?
- Dữ liệu có giá trị nào và trải rộng đến đâu?
- Dữ liệu có phân phối chuẩn hay lệch?
- Có bất kì điểm ngoại lệ nào trong dữ liệu không?

Phân tích trực quan với biểu đồ cột

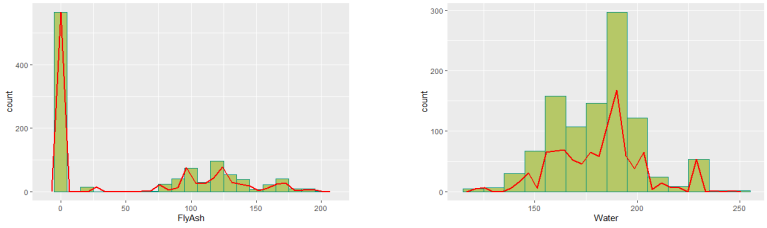


(a) Yếu tố Cement. (b) Yếu tố Blast Furnace Slag.

Hình 5: Biểu đồ cột với hai yếu tố Cement và Blast Furnace Slag.

Qua quan sát ở hình 5, yếu tố Cement ở hình 5a có độ trải rộng và phân bố không lệch về một phía, do đó mà không có ngoại lệ.

Yếu tố Blast Furnace Slag cực kỳ mất cân bằng khi giá trị 0 chiếm số lượng áp đảo so với các giá trị khác, tuy vậy dữ liệu lại trải rất rộng đến lớn hơn 300 nên chắc chắn yếu tố này chứa đựng nhiều điểm dữ liệu ngoại lệ. Qua đó có thể đặt ra một giả thuyết rằng yếu tố Blast Furnace Slag này có ít ảnh hưởng đến biến độc lập đầu ra của mô hình.

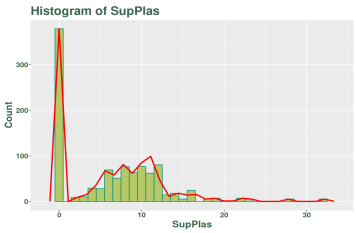


(a) Yếu tố FlyAsh. (b) Yếu tố Water.

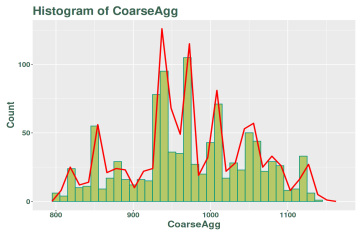
Hình 6: Biểu đồ cột với hai yếu tố FlyAsh và Water.

Hình 6a cho thấy yếu tố FlyAsh có số lượng điểm dữ liệu mang giá trị 0 áp đảo (hơn 500) so với các giá trị khác (từ 0-100 điểm). Chính vì sự mất cân bằng này, trong nhiều trường hợp dù giá trị đầu ra có thay đổi thì FlyAsh vẫn luôn bằng 0. Điều này đặt ra giả thuyết yếu tố FlyAsh không ảnh hưởng đến kết quả *strength*

Hình 6b cho thấy, trong yếu tố Water các điểm dữ liệu phân bố tập trung trong khoảng từ 150-200, ngoài ra dữ liệu phân bố cực kỳ thưa thớt ở hai đầu (số lượng điểm gần như bằng 0 ở một số giá trị) dẫn đến việc sẽ tồn tại các điểm ngoại lệ.



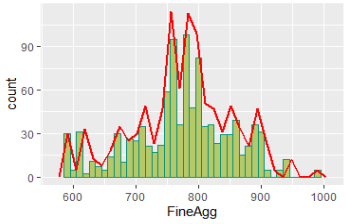
(a) Yếu tố SupPlas.



(b) Yếu tố CoarseAgg.

Hình 7: Biểu đồ cột với hai yếu tố SupPlas và CoarseAgg.

Ở hình 7, ở biểu đồ cột của yếu tố SupPlas ta thấy được số điểm dữ liệu có giá trị 0 chiếm số lượng khá lớn, chính vì thế khiến cho sự phân bố của dữ liệu ở yếu tố này mất cân bằng, ngoài ra yếu tố này còn tồn tại một số ít điểm có giá trị từ 25 đến 30 khiến cho yếu tố này xuất hiện một vài điểm ngoại lệ, từ đó ta có thể đặt ra giả thuyết rằng yếu tố SupPlas ít ảnh hưởng đến giá trị đầu ra của mô hình. Ở biểu đồ của yếu tố CoarseAgg ta thấy dữ liệu của yếu tố này đều xuất hiện ở mỗi giá trị, ở giá trị từ 930 - 1000 chiếm phần lớn điểm dữ liệu và ở phía biên của biểu đồ cũng tồn tại số ít điểm dữ liệu nên yếu tố không xuất hiện điểm ngoại lệ.



(a) Yếu tố Fine Aggregate.

(b) Yếu tố Age.

Hình 8: Biểu đồ cột với hai yếu tố Fine Aggregate và Age.

Hình 8a cho thấy lượng dữ liệu lớn tập trung trong khoảng 650-800 và không lệch, có dạng phân phối chuẩn và dàn đều trong khoảng giá trị nên ít điểm ngoại lệ.

Hình 8b bị lệch phải với lượng lớn dữ liệu tập trung trong khoảng từ 0-100, trong khi đó miền giá trị trải rộng đến giá trị hơn 300 nên có nhiều điểm ngoại lệ ở miền bên phải dữ liệu.

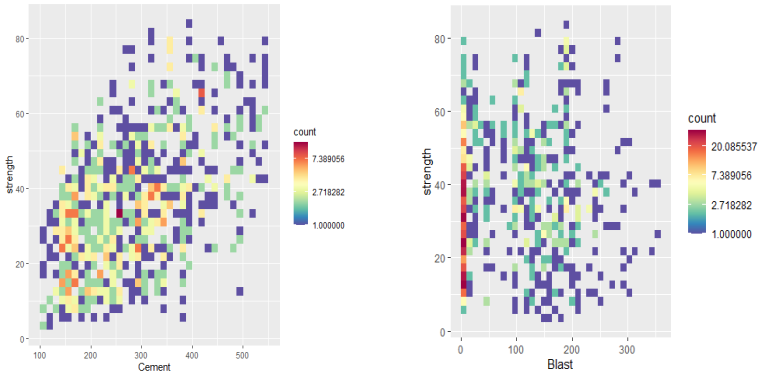
3.3 Biểu đồ Hexbin

Vai trò và mục đích

Quan sát từng thuộc tính riêng lẻ sẽ khó nhận ra sự tương tác của các yếu tố, vì thế mà chúng tôi sử dụng thêm một loại biểu đồ nữa: Hexbin. Loại biểu đồ này rất hiệu quả trong việc thể hiện sự phân bố dữ liệu tương quan giữa hai yếu tố. Đây là một dạng biểu đồ đặc biệt của Heatmap.

Biểu đồ này cho ta biết về tần suất xuất hiện của một điểm dữ liệu. Nếu màu càng đậm thì tần suất xuất hiện càng cao.

Phân tích trực quan với biểu đồ Hexbin



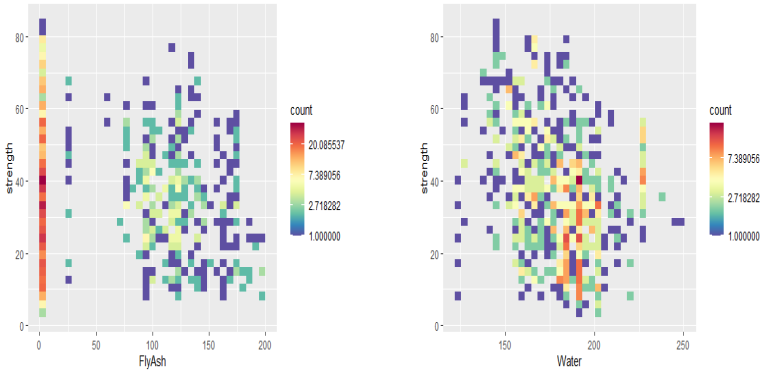
(a) Yếu tố Cement. (b) Yếu tố Blast Furnace Slag.

Hình 9: Biểu đồ hexbin với hai yếu tố Cement và Blast Furnace Slag.

Hình 9 thể hiện tần suất xuất hiện của các điểm dữ liệu trong mối quan hệ giữa biến đầu ra *Concrete compressive strength* với hai yếu tố Cement và Blast Furnace Slag. Hình 9a thể hiện các điểm dữ liệu có phân bố tập trung vào phía giá trị nhỏ. Ta có thể đặt ra một giả thuyết ban đầu rằng khối lượng xi măng (Cement) có tương quan thuận với cường độ nén của bê tông, xi măng càng nhiều dẫn đến bê tông càng nén chặt.

Hình 9b thể hiện các điểm dữ liệu có tần suất xuất hiện hoàn toàn ngẫu nhiên. Đặc biệt là khi yếu tố Blast bằng 0 thì biến đầu ra vẫn có giá trị trải

rộng trong khoảng giá trị của nó. Một nhận định ban đầu có thể đặt ra là yếu tố Blast Furnace Slag không có tương quan với sức nén của bê tông.

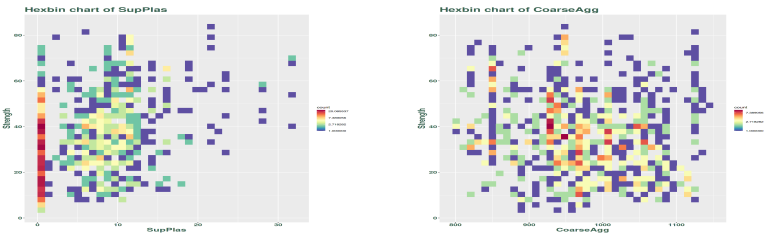


(a) Yếu tố FlyAsh. (b) Yếu tố Water.

Hình 10: Biểu đồ hexbin với hai yếu tố FlyAsh và Water.

Nhìn vào hình 10a , khi giá trị của đầu ra strength biến thiên, nó biến thiên theo chiều dọc của giá trị 0 của thuộc tính FlyAsh. Căn cứ vào điều này, ta có thể củng cố giả thuyết đã đặt ra ở hình 6a rằng yếu tố FlyAsh không có tương quan đến sức nén của bê tông.

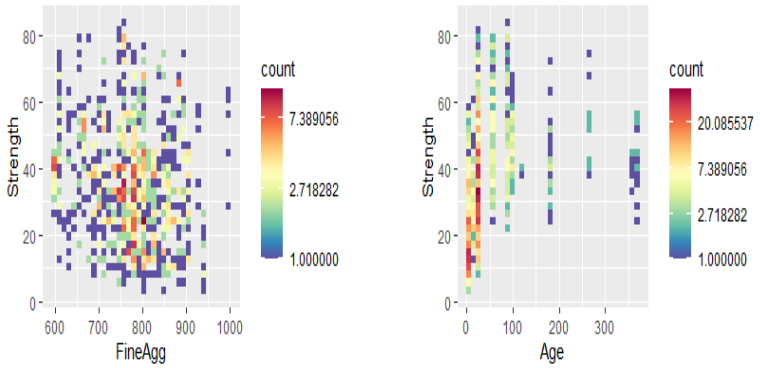
Nhìn hình 10b, tần suất xuất hiện của các điểm dữ liệu bên ngoài khoảng 125-235 là rất ít, việc này sẽ tạo ra các điểm ngoại lệ. Các điểm còn lại phân bố tương đối đều, riêng vùng gần giá trị trung bình của cả hai yếu tố Water lẫn strength xuất hiện dày đặc các điểm dữ liệu. Từ đó ta có được giả thuyết rằng khối lượng nước (Water) có tương quan với cường độ nén của xi măng và khi khối lượng nước đạt giá trị trung bình của nó, cường độ nén của xi măng cũng đạt mức trung bình.



(a) Yếu tố SupPlas. (b) Yếu tố CoarseAgg.

Hình 11: Biểu đồ hexbin với hai yếu tố SupPlas và CoarseAgg.

Hình 11 là biểu đồ dạng hexbin của hai yếu tố SupPlas và CoarseAgg, ở biểu đồ SupPlas khi giá trị SupPlas bằng 0 thì giá trị đầu ra strength biến thiên theo chiều dọc của giá trị SupPlas và khi giá trị nằm trong khoảng 25 - 30 thì tần suất xuất hiện của các điểm này là rất thấp, điều này khiến cho giả thuyết yếu tố SupPlas không ảnh hưởng đến giá trị đầu ra càng được củng cố. Ở biểu đồ CoarseAgg, các điểm có tần suất xuất hiện cao chủ yếu nằm ở phía giữa của 2 yếu tố, ta có thể đưa ra một giả thuyết từ điều này là cốt liệu thô có tương quan với độ nén của xi măng, khi cốt liệu thô được sử dụng ở mức trung bình thì cường độ nén của xi măng cũng đạt mức trung bình.



(a) Yếu tố FineAgg. (b) Yếu tố Age.

Hình 12: Biểu đồ hexbin với hai yếu tố Fine Aggregate và Age.

Hình 12a cho thấy các điểm dữ liệu tập trung nhiều trong khoảng Age từ 700-850 và Strength nằm trong khoảng từ 10-50. Ngoài ra, các điểm dữ liệu phân bố rải rác không theo một quy luật tuyến tính nào nên có thể đặt ra giả thiết yếu tố Fine Aggregate không có tương quan đến sức nén của bê tông.

Hình 12b xuất hiện các đường dọc thẳng đứng do đây là giá trị rời rạc, nằm trong khoảng từ 0-300, trong đó tập trung nhiều trong khoảng từ 0-100, mức Age từ 100 trở lên chỉ tập trung nhiều ở một số điểm nhất định và là các điểm ngoại lệ. Ngoài ra các đường dọc đều trải rất rộng lên khoảng giá trị của đầu ra.

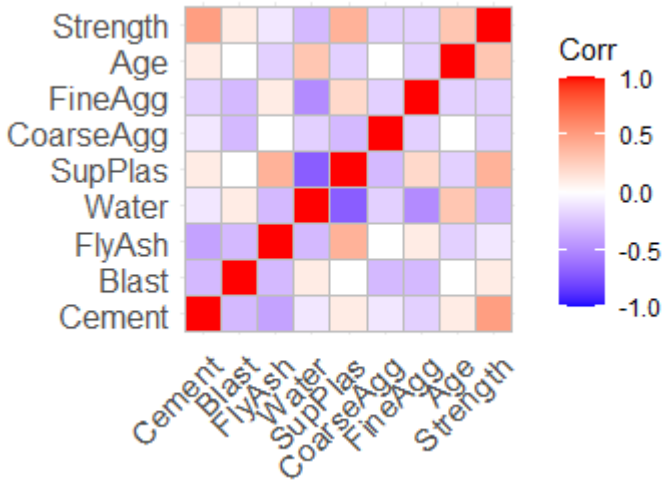
3.4 Ma trận tương quan

Vai trò và mục đích

Theo [2], Correlation Matrix là một bảng biểu diễn hệ số tương quan giữa tất cả các cặp biến trong bảng và là một công cụ hiệu quả để tổng hợp dữ liệu cũng như phát hiện và biểu diễn patterns trong bộ dữ liệu. Biểu đồ này cho ta biết được hệ số tương quan của từng cặp thuộc tính với gam màu nóng đại diện cho tương quan thuận và gam màu lạnh đại diện cho tương quan nghịch. Màu

càng đậm chứng tỏ hệ số tương quan càng tiến về 1 (đối với gam màu nóng) hoặc tiến về -1 (đối với gam màu lạnh).

Phân tích trực quan với Correlation Matrix



Hình 13: Correlation Matrix

Từ hình 13 ta rút ra một số nhận xét sau:

- Biến *Concrete Compressive Strength* có tương quan tương đối đối với 3 biến Cement, Water và Superplasticizer (SupPlas), trong đó biến Cement và Superplasticizer tương quan thuận và biến Water tương quan nghịch.
- 2 biến Blast Furnace Slag (Blast) và FlyAsh có tương quan gần như bằng 0, điều đó cho thấy 2 biến này gần như không ảnh hưởng nhiều đến kết quả đầu ra của bộ dữ liệu.
- Các biến còn lại cho thấy tương quan yếu.

4 Phân tích định lượng

4.1 Chọn lọc thuộc tính bằng giải thuật Rừng Ngẫu nhiên

Một thách thức đặt ra là số lượng thuộc tính rất nhiều, trong khi đó việc dựa vào phân tích trực quan sẽ không đảm bảo được tính khách quan và chính xác của việc chọn thuộc tính. Chúng tôi quyết định tiến hành áp dụng Học máy trong việc lựa chọn thuộc tính. Thuật toán được sử dụng là **Recursive Feature Elimination** (RFE - Loại bỏ thuộc tính hồi quy).

RFE thuộc dạng wrapper-style, bởi nó sử dụng thuật toán Random Forest (Rừng ngẫu nhiên) trong lõi để giúp đánh giá hiệu quả của cách lựa chọn thuộc tính. Nó bắt đầu bằng việc sử dụng toàn bộ thuộc tính, sau đó loại bỏ một cách hiệu quả dần dần để đạt được số lượng mong muốn các thuộc tính quan trọng. Quá trình tính toán hiệu quả này đã giúp nó trở nên phổ biến trong số những phương pháp chọn lọc thuộc tính.

Chúng tôi tiến hành sử dụng hàm `rfe` xây dựng sẵn trong gói `caret`. Mô hình được chạy với 10 vòng lặp và cho kết quả với 5 thuộc tính có ý nghĩa nhất trong số 8 thuộc tính đầu vào ban đầu, bao gồm: **age, Cement, Water, FineAgg, Blast**.

Recursive feature selection Outer resampling method: Cross-Validated (10 fold) Resampling performance over subset size:							
Variables	RMSE	Rsquared	MAE	RMSESD	RsquaredSD	MAESD	Selected
1	12.934	0.4074	10.283	1.0022	0.07863	0.8389	
2	8.991	0.7164	7.081	0.5243	0.03848	0.3382	
3	7.295	0.8245	5.599	0.6302	0.03035	0.3942	
4	6.323	0.8818	4.852	0.5448	0.01514	0.3576	
5	6.231	0.8935	4.872	0.5853	0.02256	0.3223	
6	4.923	0.9207	3.585	0.6064	0.01767	0.3271	*
7	5.063	0.9198	3.673	0.6949	0.02011	0.3856	
8	5.169	0.9189	3.813	0.6970	0.02189	0.3738	
The top 5 variables (out of 6): age, Cement, Water, FineAgg, Blast							

Bảng 4: Kết quả của thuật toán RFE.

Từ quan sát kết quả ở bảng 4, thuộc tính **Blast** có lẽ gây bất ngờ nhất vì chúng tôi đã cho rằng đây là thuộc tính có khả năng cao cần được loại bỏ sau khi phân tích trực quan. Tuy nhiên, khi sử dụng `rfe` cốt lõi là thuật toán `Random Forest`, thuộc tính **Blast** mang lại kết quả đáng kể. Điều này cho thấy Học máy có thể đóng vai trò quan trọng trong phân tích thuộc tính, không kém gì các phương pháp phân tích và trực quan dữ liệu khác.

4.2 Thử nghiệm ANOVA

ANOVA lần 1

Index		Df	Sum Sq	Mean Sq	F value	Pr(>F)	
1	age	1	31061	31061	366.004	<2e-16	***
2	Cement	1	64105	64105	755.387	<2e-16	***
3	Water	1	34597	34597	407.679	<2e-16	***
4	FineAgg	1	11712	11712	138.006	<2e-16	***
5	Blast	1	20891	20891	246.175	<2e-16	***
6	age:Cement	1	1966	1966	23.166	1.72e-06	***
7	age:Water	1	10706	10706	126.151	<2e-16	***
8	Cement:Water	1	19	19	0.219	0.639655	
9	age:FineAgg	1	5532	5532	65.187	1.95e-15	***
10	Cement:FineAgg	1	255	255	2.999	0.083621	.
11	Water:FineAgg	1	529	529	6.232	0.012707	*
12	age:Blast	1	9	9	0.101	0.751053	
13	Cement:Blast	1	2150	2150	25.338	5.71e-07	***
14	Water:Blast	1	149	149	1.753	0.185842	
15	FineAgg:Blast	1	2017	2017	23.773	1.26e-06	***
16	age:Cement:Water	1	161	161	1.902	0.168202	
17	age:Cement:FineAgg	1	2186	2186	25.756	4.62e-07	***
18	age:Water:FineAgg	1	182	182	2.149	0.142971	
19	Cement:Water:FineAgg	1	1458	1458	17.175	3.70e-05	***
20	age:Cement:Blast	1	134	134	1.584	0.208537	
21	age:Water:Blast	1	5653	5653	66.614	9.89e-16	***
22	Cement:Water:Blast	1	1079	1079	12.710	0.000381	***
23	age:FineAgg:Blast	1	120	120	1.418	0.234045	
24	Cement:FineAgg:Blast	1	18	18	0.210	0.647162	
25	Water:FineAgg:Blast	1	2818	2818	33.205	1.10e-08	***
26	age:Cement:Water:FineAgg	1	7	7	0.081	0.776380	
27	age:Cement:Water:Blast	1	569	569	6.708	0.009737	**
28	age:Cement:FineAgg:Blast	1	267	267	3.150	0.076222	.
29	age:Water:FineAgg:Blast	1	1686	1686	19.864	9.26e-06	***
30	Cement:Water:FineAgg:Blast	1	421	421	4.959	0.026174	*
31	age:Cement:Water:FineAgg:Blast	1	22	22	0.257	0.612328	
32	Residuals	998	84695	85			
	Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1						

Bảng 5: Bảng kết quả ANOVA lần 1 của tương tác giữa 5 thuộc tính đối với đầu ra.

Ta đưa ra được 2 giả thuyết không như sau:

- H0: Mean của các thuộc tính riêng lẻ không có ý nghĩa thống kê.
- H0: Không có sự tương tác giữa các thuộc tính.

Nhìn vào bảng 5, ta thấy:

- 5 thuộc tính riêng lẻ: age, Cement, Water, FineAgg, Blast đều có ý nghĩa thống kê với $P < 0.05$ (-10.56 < 0.05). Điều này chứng tỏ sử dụng chọn lọc thuộc tích bằng giải thuật Rừng Ngẫu nhiên đem lại kết quả đáng tin cậy.

- 6 tương tác giữa 2 thuộc tính age:Cement, age:Water, age:FineAgg, Water:FineAgg, Cement:Blast và FineAgg:Blast đều có ý nghĩa thống kê với P < 0.05.
- 4 tương tác giữa 2 thuộc tính không có ý nghĩa thống kê bao gồm Cement:Water, Cement:FineAgg, age:Blast và Water:Blast.
- 5 tương tác giữa 3 thuộc tính gồm age:Cement:FineAgg, Cement:Water:FineAgg, age:Water:Blast, Cement:Water:Blast và Water:FineAgg:Blast đều mang ý nghĩa thống kê với P < 0.05.
- 5 tương tác giữa 3 thuộc tính không mang ý nghĩa thống kê với P > 0.05 bao gồm age:Cement:Water, age:Water:FineAgg, age:Cement:Blast, age:FineAgg:Blast và Cement:FineAgg:Blast.
- 2 tương tác age:Cement:Water:FineAgg và age:Cement:FineAgg:Blast không có ý nghĩa thống kê với P lần lượt là 0.77638 và 0.076222, lớn hơn 0.05.
- 2 tương tác age:Cement:Water:Blast và Cement:Water:FineAgg:Blast có ý nghĩa thống kê với P < 0.05.
- Tương tác age:Cement:Water:FineAgg:Blast không có ý nghĩa thống kê với P > 0.05 do đó ta không sử dụng tương tác này trong việc xây dựng mô hình.

ANOVA lần 2

Index		Df	Sum Sq	Mean Sq	F value	Pr(F)	
1	age	1	31061	31061	335.565	<2e-16	***
2	Cement	1	64105	64105	692.564	<2e-16	***
3	Water	1	34597	34597	373.774	<2e-16	***
4	FineAgg	1	11712	11712	126.528	<2e-16	***
5	Blast	1	20891	20891	225.702	<2e-16	***
6	age:Cement	1	1966	1966	21.240	4.57e-06	***
7	age:Water	1	10706	10706	115.659	<2e-16	***
8	age:FineAgg	1	5543	5543	59.887	2.43e-14	***
9	Water:FineAgg	1	696	696	7.516	0.00622	**
10	Cement:Blast	1	2109	2109	22.786	2.08e-06	***
11	FineAgg:Blast	1	1655	1655	17.884	2.56e-05	***
12	age:Cement:FineAgg	1	1821	1821	19.672	1.02e-05	***
12	Cement:Water:FineAgg	1	7	7	0.079	0.77930	
14	age:Water:Blast	1	379	379	4.091	0.04338	*
15	Cement:Water:Blast	1	266	266	2.874	0.09032	.
16	Water:FineAgg:Blast	1	48	48	0.513	0.47388	
17	age:Cement:Water:Blast	1	125	125	1.352	0.24519	
18	Cement:Water:FineAgg:Blast	1	812	812	8.778	0.00312	**
19	age:Water:FineAgg:Blast	1	5185	5185	56.019	1.56e-13	***
20	Residuals	1010	93488	93			
	Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1						

Bảng 6: Bảng kết quả ANOVA lần 2 của tương tác giữa 5 thuộc tính đối với đầu ra.

Nhìn vào bảng 6, ta thấy:

- Ở lần thử nghiệm ANOVA thứ 2 này, chúng tôi chỉ chọn các thuộc tính riêng lẻ và các tương tác giữa chúng mà có ý nghĩa thống kê ở lần thử nghiệm ANOVA thứ 1 để tiến hành thử nghiệm.
- 4 tương tác không có ý nghĩa thống kê bao gồm Cement:Water:FineAgg, Cement:Water:Blast, Water:FineAgg:Blast, age:Cement:Water:Blast với giá trị P lần lượt là 0.77930, 0.09032, 0.47388, 0.24519 (đều lớn hơn 0.05)
- Các thuộc tính riêng lẻ và tương tác còn lại đều có ý nghĩa thống kê.

ANOVA lần 3

Index		Df	Sum Sq	Mean Sq	F value	Pr(F)	
1	age	1	31061	31061	330.053	<2e-16	***
2	Cement	1	64105	64105	681.190	<2e-16	***
3	Water	1	34597	34597	367.635	<2e-16	***
4	FineAgg	1	11712	11712	124.450	<2e-16	***
5	Blast	1	20891	20891	221.995	<2e-16	***
6	age:Cement	1	1966	1966	20.891	5.46e-06	***
7	age:Water	1	10706	10706	113.760	<2e-16	***
8	age:FineAgg	1	5543	5543	58.903	3.88e-14	***
9	Water:FineAgg	1	696	696	7.393	0.00666	**
10	Cement:Blast	1	2109	2109	22.411	2.51e-06	***
11	FineAgg:Blast	1	1655	1655	17.590	2.98e-05	***
12	age:Cement:FineAgg	1	1821	1821	19.349	1.20e-05	***
13	age:Water:Blast	1	348	348	3.697	0.05479	.
14	Cement:Water:FineAgg:Blast	1	704	704	7.483	0.00634	**
15	age:Water:FineAgg:Blast	1	3833	3833	40.727	2.66e-10	***
16	Residuals	1014	95425	94			
	Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1						

Bảng 7: Bảng kết quả ANOVA lần 3 của tương tác giữa 5 thuộc tính đối với đầu ra.

Nhìn vào bảng 7, ta thấy:

- Ở lần thử nghiệm ANOVA thứ 3 này, chúng tôi loại những tương tác giữa các thuộc tính không có ý nghĩa thống kê ở lần thử nghiệm ANOVA thứ 2 và lấy phần còn lại để tiến hành thử nghiệm.
- 1 tương tác không có ý nghĩa thống kê bao gồm age:Water:Blast với giá trị P = 0.05479 (lớn hơn 0.05).
- Các thuộc tính riêng lẻ và tương tác còn lại đều có ý nghĩa thống kê.

ANOVA lần 4

Index		Df	Sum Sq	Mean Sq	F value	Pr(F)	
1	age	1	31061	31061	316.405	<2e-16	***
2	Cement	1	64105	64105	653.022	<2e-16	***
3	Water	1	34597	34597	352.433	<2e-16	***
4	FineAgg	1	11712	11712	119.304	<2e-16	***
5	Blast	1	20891	20891	212.815	<2e-16	***
6	age:Cement	1	1966	1966	20.027	8.50e-06	***
7	age:Water	1	10706	10706	109.056	<2e-16	***
8	age:FineAgg	1	5543	5543	56.468	1.25e-13	***
9	Water:FineAgg	1	696	696	7.087	0.00789	**
10	Cement:Blast	1	2109	2109	21.485	4.03e-06	***
11	FineAgg:Blast	1	1655	1655	16.863	4.34e-05	***
12	age:Cement:FineAgg	1	1821	1821	18.549	1.82e-05	***
13	Cement:Water:FineAgg:Blast	1	616	616	6.273	0.01242	*
14	age:Water:FineAgg:Blast	1	55	55	0.558	0.45510	
15	Residuals	1015	99640	98			
	Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1						

Bảng 8: Bảng kết quả ANOVA lần 4 của tương tác giữa 5 thuộc tính đối với đầu ra.

Nhìn vào bảng 8, ta thấy:

- Tương tự như lần thử nghiệm ANOVA thứ 2 và thứ 3, chúng tôi loại những tương tác giữa các thuộc tính không có ý nghĩa thống kê và lấy phần còn lại để tiến hành thử nghiệm.
- 1 tương tác không có ý nghĩa thống kê bao gồm age:Water:FineAgg:Blast với giá trị P = 0.45510 (lớn hơn 0.05).
- Các thuộc tính riêng lẻ và tương tác còn lại đều có ý nghĩa thống kê.

ANOVA lần 5

Nhìn vào bảng 9, ta thấy ở lần thử nghiệm ANOVA thứ 5 này, tất cả các thuộc tính cơ bản và tương tác giữa chúng đều có ý nghĩa thống kê với giá trị P $>$ 0.05.

Index		Df	Sum Sq	Mean Sq	F value	Pr(F)	
1	age	1	31061	31061	316.543	<2e-16	***
2	Cement	1	64105	64105	653.306	<2e-16	***
3	Water	1	34597	34597	352.586	<2e-16	***
4	FineAgg	1	11712	11712	119.356	<2e-16	***
5	Blast	1	20891	20891	212.908	<2e-16	***
6	age:Cement	1	1966	1966	20.036	8.46e-06	***
7	age:Water	1	10706	10706	109.103	<2e-16	***
8	age:FineAgg	1	5543	5543	56.492	1.24e-13	***
9	Water:FineAgg	1	696	696	7.090	0.00787	**
10	Cement:Blast	1	2109	2109	21.494	4.01e-06	***
11	FineAgg:Blast	1	1655	1655	16.870	4.32e-05	***
12	age:Cement:FineAgg	1	1821	1821	18.557	1.81e-05	***
13	Cement:Water:FineAgg:Blast	1	616	616	6.275	0.01240	*
14	Residuals	1016	99695	98			
	Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1						

Bảng 9: Bảng kết quả ANOVA lần 5 của tương tác giữa 5 thuộc tính đối với đầu ra.

4.3 Tổng kết

Sau phân tích, chúng tôi quyết định đưa vào mô hình Hồi Quy các yếu tố và tương tác sau:

1. age
2. Cement
3. Water
4. FineAgg
5. Blast
6. age:Cement
7. age:Water
8. age:FineAgg
9. Water:FineAgg
10. Cement:Blast
11. FineAgg:Blast
12. age:Cement:FineAgg
13. Cement:Water:FineAgg:Blast

5 Thực nghiệm mô hình Hồi quy

5.1 Cơ sở lý thuyết về các độ đo đánh giá

R-Squared (hay R^2) là một độ đo thống kê thường được sử dụng trong các mô hình hồi quy nhằm xác định tỷ lệ phương sai của biến phụ thuộc được giải thích bằng biến độc lập. Hay nói cách khác, R-Squared cho thấy mức độ phù hợp của dữ liệu so với mô hình và nó có giá trị trong khoảng từ 0 đến 1, càng tiến về 1 thì mô hình càng tốt. [3]

Công thức tính R-Squared: [3]

$$R^2 = \frac{SS_{regression}}{SS_{total}}$$

Trong đó:

- $SS_{Regression}$ (Regression Sum of Squares): tổng các độ lệch bình phương giải thích từ hồi quy).
- SS_{total} (Total Sum of Squares): tổng các độ lệch bình phương toàn bộ.

Tuy nhiên, R-Squared có 2 nhược điểm lớn: [4]

- Càng thêm nhiều thuộc tính thì R-Squared càng tăng, nó không bao giờ giảm cả. Chính vì vậy, một mô hình với nhiều thuộc tính trông có vẻ phù hợp nếu chỉ dựa vào R-Squared.
- Nếu một mô hình có nhiều thuộc tính và chứa nhiều đa thức bậc cao, nó sẽ bao gồm cả những điểm nhiễu trong bộ dữ liệu dẫn đến mô hình bị overfitting (R-Squared cao nhưng khả năng dự đoán chính xác thấp).

Chính vì hai nhược điểm trên, ta sẽ sử dụng R-Squared hiệu chỉnh (adjust R-Squared) như một phương án tối ưu hơn. R-Squared hiệu chỉnh là phiên bản cải tiến của R-Squared, nó chỉ tăng khi thuộc tính thêm vào có ý nghĩa và giảm nếu thuộc tính thêm vào không có ý nghĩa. R-squared hiệu chỉnh có thể âm và luôn nhỏ hơn R-Squared. Đây chính là lí do các nhà nghiên cứu thường sử dụng nó thay cho R-Squared. [4]

Công thức tính R-Squared hiệu chỉnh: [6]

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

Trong đó:

- n: số lượng mẫu quan sát.
- k: số tham số của mô hình.

5.2 Xây dựng mô hình Hồi quy

Mô hình Linear Regression sử dụng bộ tham số cơ bản

Phương trình hồi quy tổng quát của mô hình:

$$\begin{aligned} strength =& \beta_0 + \beta_1 \times \text{age} + \beta_2 \times \text{Cement} + \\ & \beta_3 \times Water + \beta_4 \times FineAgg + \beta_5 \times Blast \end{aligned} \tag{1}$$

Tiếp theo, chúng tôi đặt giả thuyết như sau:

- **H0:** Từng thuộc tính không ảnh hưởng đến kết quả đầu ra.
- **H1:** Từng thuộc tính có ảnh hưởng đến kết quả đầu ra.

Coefficients:						
Index		Estimate	Std. Error	t value	Pr(>|t|)	
1	(Intercept)	79.562273	7.968050	9.985	<2e-16	***
2	age	0.107689	0.006294	17.110	<2e-16	***
3	Cement	0.079711	0.004153	19.192	<2e-16	***
4	Water	-0.335463	0.021329	-15.728	<2e-16	***
5	FineAgg	-0.019013	0.006081	-3.127	0.00183	**
6	Blast	0.057694	0.005038	11.452	<2e-16	***
7	Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
8	Residual standard error: 11.03 on 820 degrees of freedom Multiple R-squared: 0.5668, Adjusted R-squared: 0.5642 F-statistic: 214.6 on 5 and 820 DF, p-value: <2.2e-16					

Bảng 10: Summary Linear Regression với tham số cơ bản

Kết quả cho thấy cả 5 thuộc tính đều có $P < 0.05 =>$ H1 đúng đối với tất cả các thuộc tính. Phương trình hồi quy đạt được là:

$$\begin{aligned} strength =& 79.562273 + 0.107689 \times \text{age} + 0.079711 \times \text{Cement} \\ & - 0.335463 \times Water - 0.019013 \times FineAgg + 0.057694 \times Blast \end{aligned} \tag{2}$$

R-Squared hiệu chỉnh của mô hình tham số cơ bản: 0.5642 (*)

Mô hình hồi quy có tương tác

Thông qua việc thử nghiệm ANOVA, chúng tôi đã lọc ra được các thuộc tính và mối quan hệ có ý nghĩa thống kê. Dựa vào đó ta có được mô hình hồi quy tổng quát như sau:

$$\begin{aligned} strength =& \beta_0 + \beta_1 \times age + \beta_2 \times Cement + \beta_3 \times Water + \beta_4 \times FineAgg + \beta_5 \times Blast \\ & + \beta_6 \times age \times Cement + \beta_7 \times age \times Water + \beta_8 \times age \times FineAgg \\ & + \beta_9 \times Water \times FineAgg + \beta_{10} \times Cement \times Blast + \beta_{11} \times FineAgg \times Blast \\ & + \beta_{12} \times age \times Cement \times FineAgg \\ & + \beta_{13} \times Cement \times Water \times FineAgg \times Blast \end{aligned} \tag{3}$$

Tiếp theo, chúng tôi tiến hành chạy t-test để kiểm tra từng biến cụ thể với giả thuyết như sau:

- **H0:** Từng thuộc tính và tương tác có ảnh hưởng đến kết quả đầu ra.
- **H1:** Từng thuộc tính và tương tác không ảnh hưởng đến kết quả đầu ra.

Theo bảng 11, ta thấy các thuộc tính `age`, `Cement`, `Blast` và các tương tác `age:Cement`, `age:Water`, `age:FineAgg`, `FineAgg:Blast` và `age:Cement:FineAgg` có ảnh hưởng tới kết quả đầu ra với $P < 0.05$. Theo đó, phương trình hồi quy

	Coefficients:					
Index		Estimate	Std. Error	t value	Pr(>|t|)	
1	(Intercept)	1.626e+01	2.910e+01	0.559	0.57639	
2	age	2.159e+00	2.118e-01	10.192	<2e-16	***
3	Cement	7.960e-02	5.848e-03	13.613	<2e-16	***
4	Water	2.517e-02	1.580e-01	0.159	0.87346	
5	FineAgg	1.650e-02	3.540e-02	0.466	0.64131	
6	Blast	-1.542e-01	5.214e-02	-2.957	0.00320	**
7	age:Cement	-2.195e-03	3.871e-04	-5.670	1.99e-08	***
8	age:Water	-4.250e-03	3.627e-04	-11.717	<2e-16	***
9	age:FineAgg	-1.417e-03	2.145e-04	-6.608	7.04e-11	***
10	Water:FineAgg	-2.242e-04	1.967e-04	-1.140	0.25465	
11	Cement:Blast	-8.122e-05	1.783e-04	-0.455	0.64892	
12	FineAgg:Blast	1.961e-04	6.604e-05	2.969	0.00307	**
13	age:Cement:FineAgg	2.397e-06	5.394e-07	4.444	1.01e-05	***
14	Cement:Water:FineAgg:Blast	2.340e-09	1.428e-09	1.639	0.10157	
15	Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					
16	Residual standard error: 9.911 on 812 degrees of freedom Multiple R-squared: 0.6538, Adjusted R-squared: 0.6483 F-statistic: 118 on 13 and 812 DF, p-value: <2.2e-16					

Bảng 11: Bảng t-test từng tương tác

đạt được là:

$$\begin{aligned} strength =& 2.159 \times age + 7.960e-02 \times Cement - 1.542e-01 \times Blast - 2.195e-03 \times age \times Cement \\ & - 4.250e-03 \times age \times Water - 1.417e-03 \times age \times FineAgg \\ & + 1.961e-04 \times FineAgg \times Blast \\ & + 2.340e-09 \times Cement \times Water \times FineAgg \times Blast \end{aligned} \tag{4}$$

R-squared hiệu chỉnh của mô hình có các tương tác: 0.6483 (**).

Từ (*) và (**) ta thấy R-squared hiệu chỉnh của mô hình có tương tác cao hơn so với của mô hình tham số cơ bản.

–> Mô hình có tham số cơ bản và các tương tác của nó tốt hơn.

Residual standard error: 9.911.

Tuy nhiên mô hình có *residual standard error* tương đối cao, có thể nói mô hình chưa thực sự phù hợp với bộ dữ liệu. Chính vì thế, chúng tôi sẽ tìm hiểu nguyên nhân dẫn đến residual standard error cao và cách khắc phục ở các phần sau.

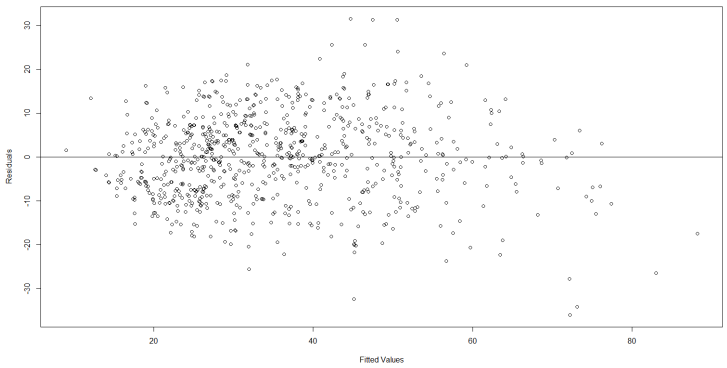
5.3 Hiệp phương sai không đồng nhất (Heteroscedasticity)

Một mô hình mắc phải tình trạng hiệp phương sai không đồng nhất (HP-SKDN) sẽ làm cho kết quả hồi quy trở nên không đáng tin cậy và không thể được xem là một mô hình chính xác.

Cụ thể hơn trong phân tích hồi quy, tình trạng HPSKDN là khái niệm nói về phần dư (residuals) hoặc lỗi (error). Khi đó, phương sai của các sai số trong một mô hình hồi quy không ổn định đối với tất cả các điểm dự đoán. Để cho kết quả mô hình có độ tin cậy cao, phương sai của phần dư trong tất cả các điểm dự đoán cần phải ổn định [5, 9].

Để kiểm định tính chất HPSKDN, chúng tôi đặt ra hai giả thuyết sau:

- **H0:** Hiệp phương sai đồng nhất. (Phần dư của mô hình được phân phối với phương sai như nhau.)
- **H1:** Hiệp phương sai không đồng nhất. (Phần dư của mô hình được phân phối với phương sai tán xạ.)



Hình 14: Biểu đồ tương quan giữa phần dư với giá trị thực.

Trước hết, về mặt trực quan, chúng tôi hiện thực biểu đồ tương quan giữa phần dư (residuals) và giá trị khớp (fitted values) (hình 14). Chúng ta thấy rằng các phần dư được phân bố rải rác với độ phân tán ngày càng ngẫu nhiên và bất quy tắc đối với những giá trị lớn. Do đó có thể tạm cho rằng mô hình đã gặp phải tình trạng HPSKDN.

Ngoài ra về mặt thống kê, nhằm đạt được kiểm định chính xác hơn, chúng tôi sử dụng *Breusch-Pagan test* để kiểm định hai giả thuyết trên.

studentized Breusch-Pagan test
data: relation
BP = 101.24, df = 13, p-value = 9.546e-16

Bảng 12: Kết quả kiểm định Breusch-Pagan test.

Từ kết quả quan sát được từ *Breusch-Pagan test* được thể hiện ở bảng 12, ta nhận thấy `p-value < 0.05`. Do đó bác bỏ giả thuyết H0, mô hình đã xây dựng tồn tại tình trạng HPSKDN.

Để giải quyết tình trạng này, chúng tôi sẽ sử dụng mô hình **Weighted Least Squares Regression**.

5.4 Xây dựng mô hình Weighted Least Squares Regression

Mô hình này có khả năng cải thiện tình trạng HPSKDN nhờ vào việc đánh trọng số nhiều hơn vào những điểm quan sát có phương sai của phần dư thấp, qua đó giúp cho mô hình đạt được độ chính xác cao hơn đối với tất cả các điểm quan sát [9].

Index	Coefficients:					
		Estimate	Std. Error	t value	Pr(>|t|)	
1	(Intercept)	-3.188e+01	3.061e+01	-1.041	0.298019	
2	age	2.257e+00	2.201e-01	10.254	<2e-16	***
3	Cement	8.365e-02	5.742e-03	14.570	<2e-16	***
4	Water	2.609e-01	1.646e-01	1.585	0.113388	
5	FineAgg	7.101e-02	3.738e-02	1.900	0.057838	.
6	Blast	-1.519e-01	5.105e-02	-2.974	0.003022	**
7	age:Cement	-2.043e-03	4.081e-04	-5.007	6.80e-07	***
8	age:Water	-4.726e-03	4.029e-04	-11.730	<2e-16	***
9	age:FineAgg	-1.354e-03	2.200e-04	-6.155	1.18e-09	***
10	Water:FineAgg	-5.045e-04	2.044e-04	-2.468	0.013775	*
11	Cement:Blast	-1.370e-05	1.897e-04	-0.072	0.942453	
12	FineAgg:Blast	2.050e-04	6.477e-05	3.165	0.001609	**
13	age:Cement:FineAgg	2.060e-06	5.698e-07	3.615	0.000319	***
14	Cement:Water:FineAgg:Blast	1.561e-09	1.509e-09	1.034	0.301286	
15	Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					
16	Residual standard error: 1.237 on 812 degrees of freedom Multiple R-squared: 0.6536, Adjusted R-squared: 0.6481 F-statistic: 117.9 on 13 and 812 DF, p-value: <2.2e-16					

Bảng 13: Summary kết quả từ mô hình Weighted Least Squares Regression.

Bảng 13 mô tả kết quả cho mô hình **WLSR**. Ta thấy các thuộc tính `age`, `Cement`, `Blast` và các tương tác `age:Cement`, `age:Water`, `age:FineAgg`, `Water:FineAgg`, `FineAgg:Blast`, `age:Cement:FineAgg` có ảnh hưởng tới kết quả đầu ra với $P < 0.05$. Chúng ta rút ra được công thức đại diện tốt nhất cho mô hình như sau:

$$\begin{aligned} strength =& 2.257 \times age + 8.365 \times 10^{-2} \times Cement \\ & - 1.519 \times 10^{-1} \times Blast \\ & - 2.043 \times 10^{-3} \times age \times Cement \\ & - 4.726 \times 10^{-3} \times age \times Water \\ & - 1.354 \times 10^{-3} \times age \times FineAgg \\ & - 5.045 \times 10^{-4} \times Water \times FineAgg \\ & + 2.050 \times 10^{-4} \times FineAgg \times Blast \\ & + 2.060 \times 10^{-6} \times age \times Cement \times FineAgg \end{aligned} \tag{5}$$

R-Squared hiệu chỉnh: 0.6536
Residual standard error: 1.237

Trong bài toán trên, giá trị R-Squared hiệu chỉnh là 65.36%. Như vậy các biến độc lập và tương tác của nó chỉ giải thích được 65.36 % sự biến thiên của biến phụ thuộc. Phần còn lại được giải thích bởi các biến ngoài mô hình và sai số ngẫu nhiên.

Từ kết quả của *residual standard error* cho ta thấy, vấn đề hiệp phương sai không đồng nhất đã được giải quyết, tuy R-Squared hiệu chỉnh có giảm nhưng không đáng kể, bù lại độ tin cậy của mô hình tăng lên.

Chúng tôi đi đến kết luận rằng đây là mô hình tốt nhất dành cho bộ dữ liệu của nhóm.

5.5 Xử lí outliers sử dụng độ đo Leverage

Leverage là độ đo khoảng cách giữa các giá trị biến độc lập của một quan sát so với các giá trị của các quan sát khác. Khi được đo bằng các giá trị dự đoán, nó là khoảng cách được chuẩn hóa đến giá trị trung bình của các yếu tố dự báo. Có thể nói, khi giá trị Leverage của một điểm dữ liệu cao thì có thể xem như điểm dữ liệu đó là outliers.

Công thức của độ đo leverage:

$$h_{ii} = \frac{\delta \hat{y}_i}{\delta y_i}$$

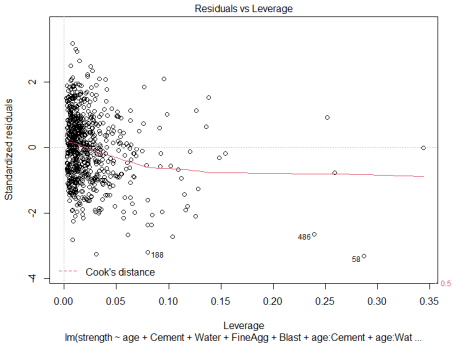
Trong đó:

- $\delta \hat{y}_i$: Giá trị dự đoán thứ i của y.
- δy_i: Giá trị thứ i của y.
- h_{ii}: Giá trị Leverage thứ i.

Nhìn vào hình 15, ta có thể thấy điểm dữ liệu có giá trị leverage trên 0.05 là outliers nên chúng tôi quyết định bỏ những điểm dữ liệu đó đi.

Sau đó chúng tôi xây dựng mô hình Regression dựa trên bộ dữ liệu mới:

Chỉ số R-Square hiệu chỉnh của mô hình mới đạt 0.6846, tăng từ 0.8463 so với mô hình chưa xử lí outliers. Điều đó cho thấy việc xử lí outliers cho ra kết quả khả quan.



H

Hình 15: Caption

Index	Coefficients:					
		Estimate	Std. Error	t value	Pr(>|t|)	
1	(Intercept)	1.833e+01	3.459e+01	0.530	0.596261	
2	age	1.992e+00	3.020e-01	6.596	7.9e-11	***
3	Cement	8.452e-02	6.127e-03	13.796	<2e-16	***
4	Water	-1.234e-02	1.890e-01	-0.065	0.947972	
5	FineAgg	1.437e-02	4.237e-02	0.339	0.734634	
6	Blast	-6.503e-02	5.513e-02	-1.179	0.238571	
7	age:Cement	-1.834e-03	6.684e-04	-2.744	0.006217	**
8	age:Water	-4.334e-03	4.728e-04	-9.167	<2e-16	***
9	age:FineAgg	-1.135e-03	3.309e-04	-3.429	0.000639	***
10	Water:FineAgg	-2.026e-04	2.356e-04	-0.860	0.390157	
11	Cement:Blast	-1.944e-04	1.859e-04	-1.046	0.295883	
12	FineAgg:Blast	9.916e-05	7.020e-05	1.412	0.158225	
13	age:Cement:FineAgg	1.988e-06	8.751e-07	2.272	0.023375	*
14	Cement:Water:FineAgg:Blast	2.881e-09	1.482e-09	1.944	0.052274	.
15	Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					
16	Residual standard error: 9.417 on 766 degrees of freedom Multiple R-squared: 0.6898, Adjusted R-squared: 0.6846 F-statistic: 131 on 13 and 766 DF, p-value: <2.2e-16					

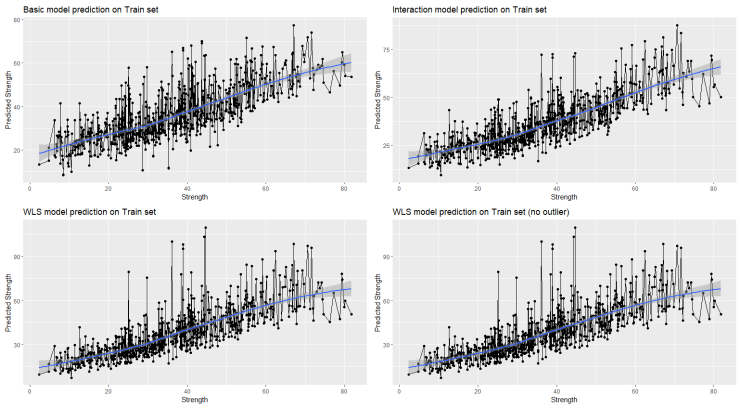
Bảng 14: WLS có xử lý Outliers

Phương trình hồi quy của mô hình là:

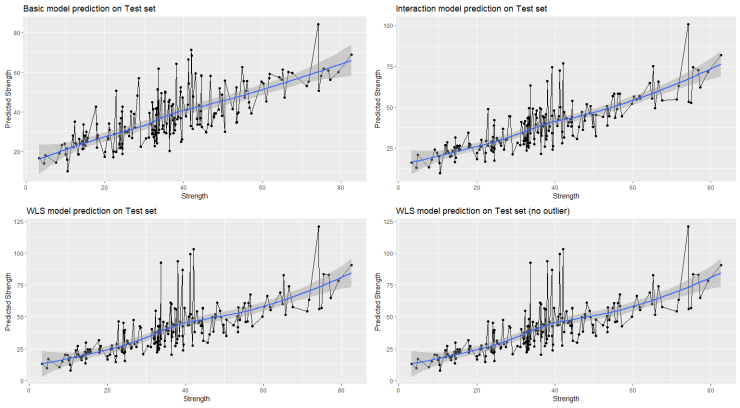
$$\begin{aligned} strength =& 1.992 \times age + 8.452 \times 10^{-2} \times Cement \\ & - 1.834 \times 10^{-3} \times age \times Cement \\ & - 4.334 \times 10^{-3} \times age \times Water \\ & - 1.135 \times 10^{-3} \times age \times FineAgg \\ & + 1.988 \times 10^{-6} \times age \times Cement \times FineAgg \end{aligned} \tag{6}$$

5.6 Đánh giá mô hình

Kết quả dự đoán Sau đây ở hình 16 và hình 17 là những biểu đồ chúng tôi trực quan hóa kết quả khi sử dụng các mô hình đã tạo để dự đoán trên hai bộ dữ liệu huấn luyện (train set) và bộ dữ liệu kiểm thử (test set).



Hình 16: Kết quả dự đoán và thực tế trên tập huấn luyện.



Hình 17: Kết quả dự đoán và thực tế trên tập kiểm tra.

- Nếu mô hình chính xác tuyệt đối, các điểm dữ liệu màu đen sẽ nằm trên đường thẳng màu xanh nước biển được xây dựng từ mô hình hồi quy
- Nếu mô hình không phù hợp, các chấm đen sẽ xuất hiện 1 cách ngẫu nhiên không theo hướng của đường thẳng.

– Trên bộ dữ liệu huấn luyện (hình 16), ta thấy tất cả mô hình đều cho ra các điểm dự đoán biến thiên mạnh so với đường màu xanh. Trong đó, mô hình cơ bản (trái trên) và mô hình có các biến tương tác (phải trên) nhìn chung biến thiên ít hơn và có dải giá trị của trục y ngắn hơn hai mô hình dưới.
– Trên bộ kiểm tra (hình 17), chúng ta thấy một đặc điểm nổi bật rằng mô hình với bộ tham số cơ bản (trái trên), và mô hình WLS (trái dưới), cho kết quả tốt, biến thiên không quá mạnh so với đường màu xanh. Trong khi đó, mô hình WLS nếu được huấn luyện trên bộ dữ liệu đã xử lý ngoại lệ (phải dưới) thì không có sự thay đổi tích cực nào so với mô hình WLS ban đầu. Còn mô hình có tương tác (phải trên), có sự biến thiên mạnh và không đáng tin cậy.

	MAE trên bộ huấn luyện (*)	**MAE trên bộ kiểm tra (**)**	****/***
Mô hình cơ bản	8.758806	9.297172	1.06
Mô hình tương tác	7.78715	7.984923	1.03
Mô hình WLS	7.904266	8.76696	1.11
WLS trên dữ liệu đã xử lý	7.904266	8.76696	1.11

Bảng 15: Độ đo Mean Absolute Error trên bốn mô hình.

Về cách đánh giá độ đo, giả sử độ đo **MAE** là 0.4, miền giá trị của kết quả đầu ra là 100 thì đây là mô hình tốt. Ngược lại, nếu miền giá trị của kết quả đầu ra là 2 thì đây là mô hình không tốt. Nói cách khác, các độ đo như **MAE** càng nhỏ so với miền giá trị của kết quả đầu ra thì càng tốt.

Khoảng giá trị của thuộc tính đầu ra là 80.267. Vậy độ đo **MAE** trên tất cả bốn mô hình đều thấp hơn khoảng 11% so với khoảng giá trị, chứng tỏ mô hình tốt. Tuy vậy, mô hình Weight Least Square của chúng tôi mang lại sự cải thiện đáng kể. Vậy có thể kết luận rằng mô hình *Weight Least Square* là tốt nhất.

Giả thuyết 1: Mean của phần dư là 0

Khi một mô hình có phần dư trung bình xấp xỉ bằng 0, có nghĩa là mô hình đạt được độ tin cậy cao và các dự đoán gần khớp với giả trị thực.

	Mean of residuals
Mô hình cơ bản	3874776×10^{-17}
Mô hình có tương tác	-1066634×10^{-15}
Weighted Least Square	-0.04673711
WLS trên dữ liệu xử lý ngoại lệ	1692275×10^{-15}

Bảng 16: Mean phần dư của 4 mô hình

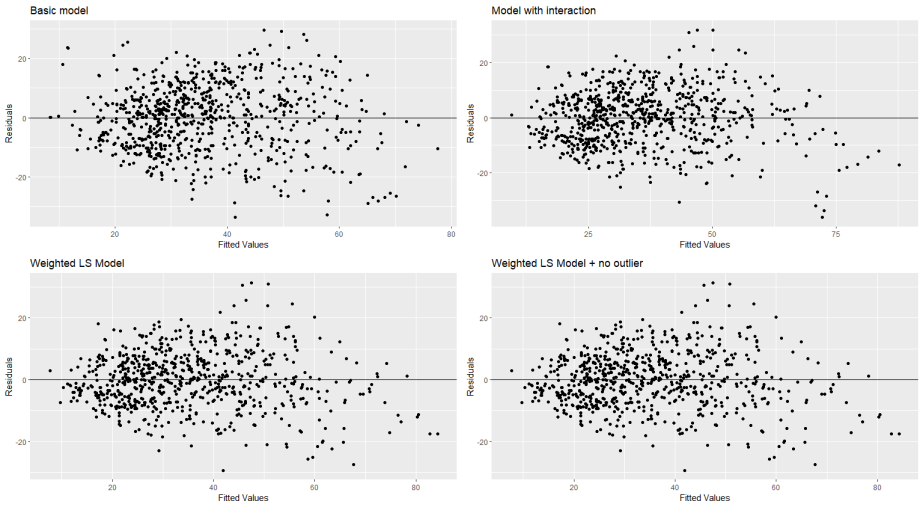
Theo bảng 16:

– Mô hình Linear cơ bản không bao gồm các tương tác, Linear bao gồm cả các tương tác và Weighted Least Square đã bỏ outlier đạt độ chính xác đáng tin cậy khi đưa vào dự đoán thực tế vì *mean* của chúng gần như bằng 0
– Mô hình Weighted Least Square tuy không đạt độ chính xác như ba mô hình trên nhưng mean vẫn rất thấp và xấp xỉ 0 nên ta cũng có thể kết luận rằng mô hình này có thể đưa ra kết quả dự đoán tốt.

Giả thuyết 2: Mô hình có hiệp phương sai đồng nhất

Giờ đây chúng ta cần phải kiểm định lại mô hình một lần nữa để đảm bảo mô hình có tính HPSDN. Chỉ khi đó, kết quả mô hình mới đáng tin cậy trên toàn bộ mọi điểm dự đoán. Hiện tại chúng tôi đặt ra hai giả thuyết như sau:

– **H0:** Mô hình có tính hiệp phương sai đồng nhất.
– **H1:** Mô hình có tính hiệp phương sai không đồng nhất.



Hình 18: Phân bố của phương sai của phần dư (lỗi) của bốn mô hình.

Biểu đồ hình 18 mô tả sự phân số của phương sai của lỗi (error). Nhìn chung thì mô hình cơ bản (trái trên), mô hình WLSR (trái dưới) và mô hình WLSR trên dữ liệu đã xử lý (phải dưới) có phân bố lỗi tương đồng với nhau trên và dưới đường thẳng ngang hơn là mô hình có tương tác, nghĩa là chúng có thể có hiệp phương sai đồng đều. Đối với mô hình có tương tác (phải trên), phân bố không đồng đều và bất quy tắc đối với những giá trị lớn.

Chúng tôi sẽ dùng *Breusch-Pagan test* để kiểm định chính xác giả thuyết này.

Mô hình cơ bản	Mô hình có tương tác	Mô hình WLSR	Mô hình WLSR + loại bỏ outliers
studentized Breusch-Pagan test data: relation_basic BP = 85.058, df = 5, p-value < 2.2e-16	studentized Breusch-Pagan test data: relation BP = 98.371, df = 13, p-value = 3.429e-15	studentized Breusch-Pagan test data: wls_model BP = 5.4479, df = 13, p-value = 0.964	studentized Breusch-Pagan test data: wls_model BP = 5.4479, df = 13, p-value = 0.964

Bảng 17: Breusch-Pagan test trên tất cả mô hình đã xây dựng.

Bảng 17 mô tả kết quả của phép *Breusch-Pagan test* trên bộ dữ liệu Test mà nhóm đã chia trước đó và không sử dụng nó trong quá trình xây dựng mô hình.

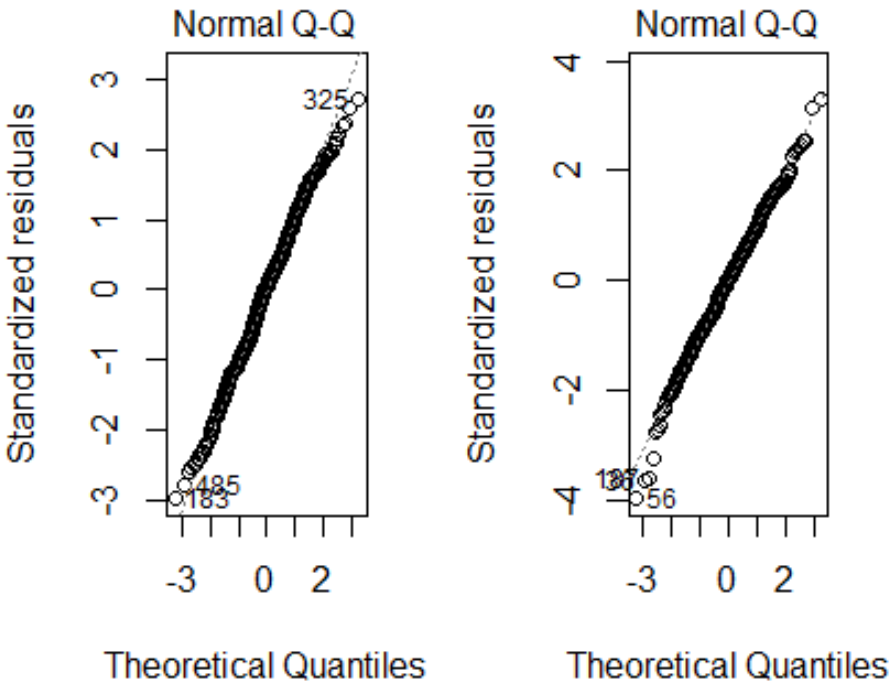
- Bác bỏ H0 đối với hai mô hình cơ bản (chỉ gồm 5 yếu tố) và mô hình hồi quy có tương tác. Vậy hai mô hình này không thỏa mãn được giả thuyết thứ 2.

- Không thể bác bỏ H0 đối với mô hình *Weighted Least Square Regression* và mô hình *Weighted Least Square Regression* được xây dựng từ bộ huấn luyện đã qua xử lý ngoại lệ bởi kết quả $p - value > 0.05$.

- Vậy mô hình *Weighted Least Square Regression* dù được huấn luyện trên tập nào cũng mang tính chất Hiệp phương sai đồng nhất.

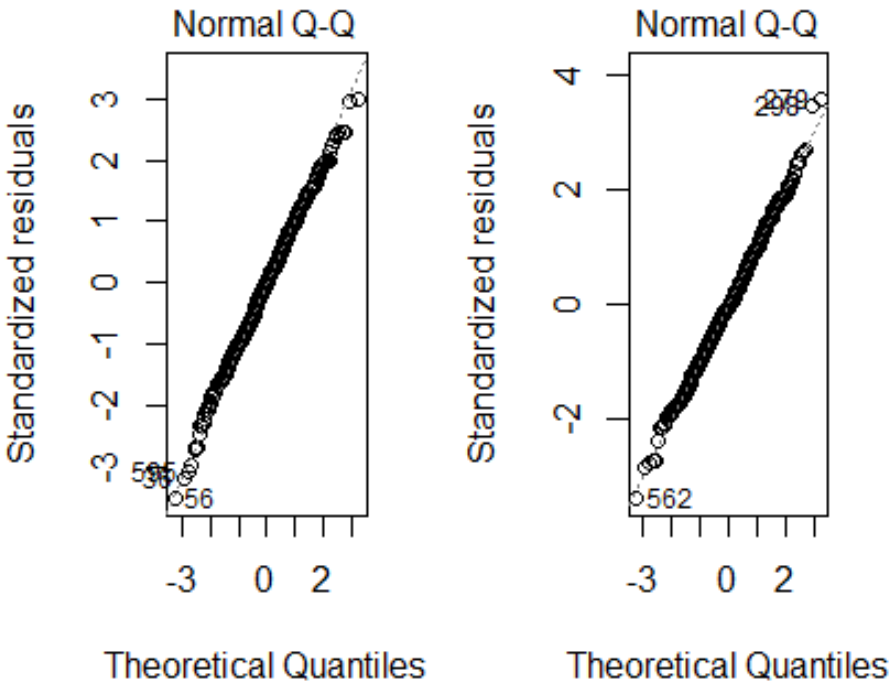
Giả thuyết 3: Tính tiêu chuẩn của phần dư

Biểu đồ Q-Q thường được áp dụng khi muốn kiểm định phân phối phần dư của một mô hình có tuân theo phân phối chuẩn hay không. Nếu các giá trị trên biểu đồ Q-Q có xu hướng nằm trên trục góc 45 độ, thì có nghĩa là phân phối của phần dư tuân theo phân phối chuẩn.



Hình 19: Phân phối chuẩn của phần dư của mô hình linear cơ bản (trái) và mô hình linear có tương tác (phải).

Như biểu đồ hình 19, ta thấy các điểm đều nằm trên một đường thẳng trừ một vài ngoại lệ ở hai đầu, mô hình linear có tương tác có *Standardlized residuals* cao hơn 1 đơn vị so với mô hình linear cơ bản nhưng nhìn chung phần dư ở cả hai mô hình đều tuân theo phân phối chuẩn nên giả thuyết đặt ra cho hai mô hình này là đúng.

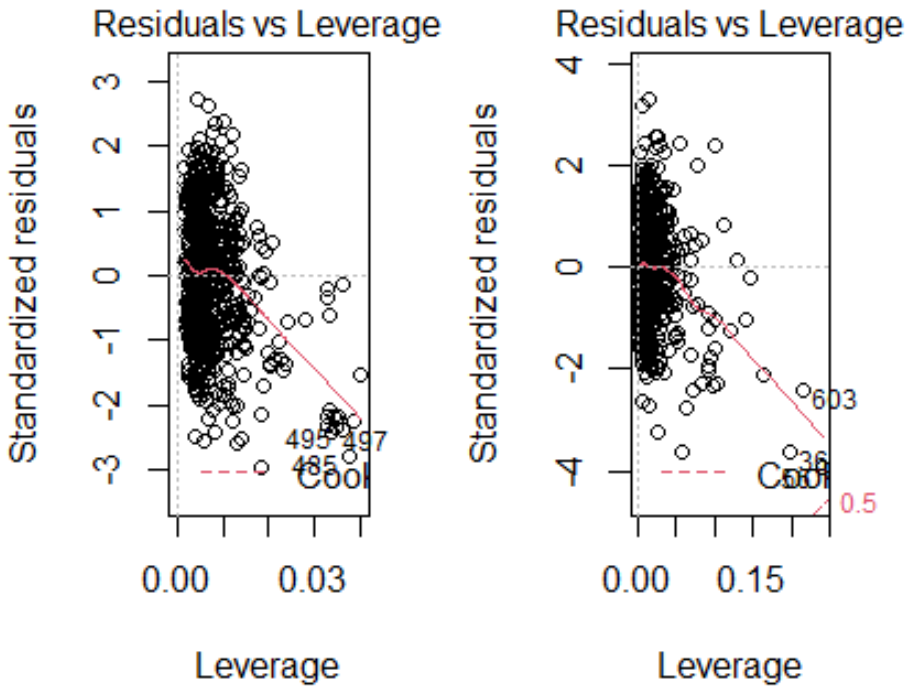


Hình 20: Phân phối chuẩn của phần dư của mô hình wls chưa xử lí outlier (trái) và mô hình wls đã xử lí outlier (phải).

Nhìn vào biểu đồ hình 20, ta thấy ở mô hình wls chưa xử lí outlier và cả mô hình wls đã xử lí outlier các điểm đều nằm trên một đường thẳng kể cả ở hai đầu, có thể nói phần dư ở mô hình wls cũng có tính phân phối chuẩn và mô hình này tốt hơn hai mô hình linear vì gần như không có điểm lệch khỏi đường thẳng. Giả thuyết đặt ra cũng hoàn toàn đúng với hai mô hình wls này.

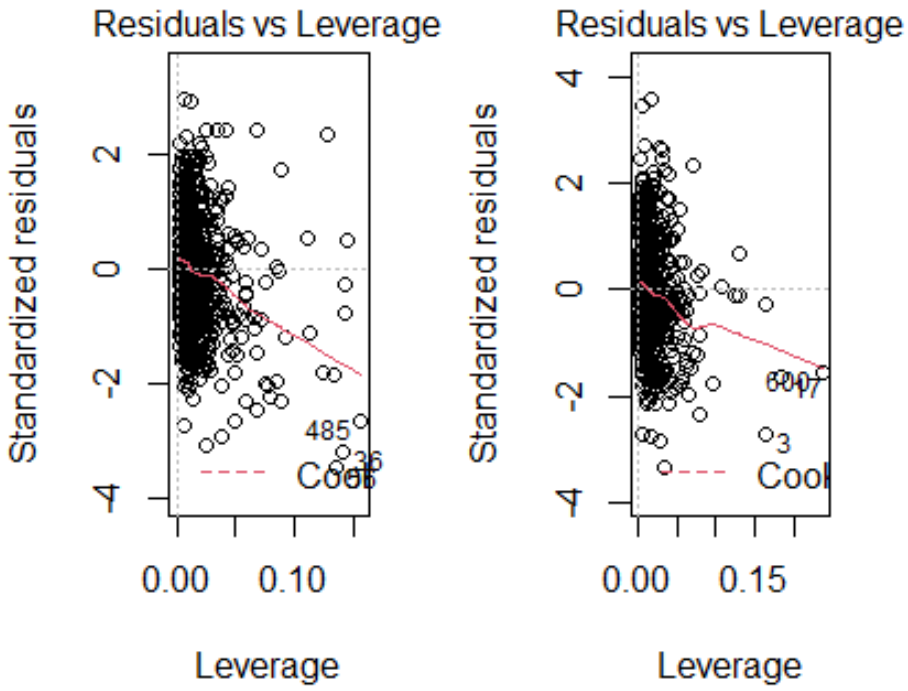
Giả thuyết 4: Có tồn tại điểm ngoại lệ ảnh hưởng lớn đến mô hình

Mô hình được đánh giá là bị ảnh hưởng bởi các điểm ngoại lệ khi standardized residuals vượt mức +/-3 thậm chí rất tệ nếu đạt tới +/-4.



Hình 21: Điểm ngoại lệ của mô hình linear cơ bản (trái) và mô hình linear có tương tác (phải).

Biểu đồ hình 21 cho thấy mô hình linear cơ bản có standardized residuals của điểm xa nhất (#485) nằm trong mức +/-3 – tức mô hình không bị ảnh hưởng nhiều bởi điểm ngoại lệ và áp dụng thực tế sẽ cho kết quả tốt. Còn với mô hình linear có tương tác, standardized residuals vượt mức an toàn +/-3 (điểm #36) và thậm chí chạm tới mức +/-4. Như vậy, giả thuyết đặt ra đúng với mô hình linear có tương tác và không đúng với mô hình linear cơ bản.



Hình 22: Điểm ngoại lệ của mô hình wls chưa xử lí outlier (trái) và mô hình wls đã xử lí outlier (phải).

Nhìn biểu đồ hình 22, cả hai mô hình có cùng mức cooks'D và đều tồn tại các điểm ngoại lệ chạm mức Standardlized residuals +/-4. Vậy giả thuyết đặt ra cũng hoàn toàn đúng với hai mô hình wls này dù đã xử lí outlier hay chưa.

6 Kết luận

6.1 Những thành quả đạt được

Thông qua quá trình phân tích định lượng bằng cách sử dụng cả hai phương pháp là sử dụng thuật toán RFE và kiểm định ANOVA, chúng tôi đã có thể chọn ra được những thuộc tính quan trọng để đưa vào huấn luyện mô hình như sau:

1. Thuộc tính **Age**.

2. Thuộc tính **Cement**.
3. Thuộc tính **Water**.
4. Thuộc tính **Fine Aggregate**.
5. Thuộc tính **Blast Furnace Slag**.

Một vài khó khăn đã xuất hiện trong quá trình thực hiện đồ án này, đầu tiên là việc số lượng thuộc tính của bộ dữ liệu khá nhiều, gây khó khăn cho việc chọn lọc các tương tác tiềm năng, tuy nhiên điều đó được chúng tôi giải quyết bằng cách sử dụng giải thuật RFE nhằm loại đi những thuộc tính không có ý nghĩa thống kê từ đó giảm đi số lượng thuộc tính cần phải phân tích, phương pháp này mang lại kết quả khá tốt khi đã loại bỏ được 3 thuộc tính không mang lại nhiều lợi ích cho việc xây dụng mô hình. Ngoài việc số lượng thuộc tính khá lớn, khi xây dựng mô hình chúng tôi đã gặp phải một thử thách khác chính là mô hình gặp phải tình trạng hiệp phương sai không đồng nhất, điều này sẽ khiến mô hình trở nên không đáng tin cậy. Để giải quyết điều này, chúng tôi đã sử dụng mô hình Weighted Least Squares Regression nhằm cải tiến mô hình. Ngoài ra việc xử lí outlier trên bộ dữ liệu huấn luyện không cải thiện độ chính xác của mô hình.

Mô hình hồi quy mà nhóm chúng tôi đề xuất có độ chính xác cao và độ tin cậy tốt bằng những con số rất cụ thể đã được trình bày ở 5.6, và kiểm định qua bốn giả thuyết khác nhau.

Tuy nhiên, công tác tiền xử lý dữ liệu đối vơi ngoại lệ đóng vai trò rất mờ nhạt trong xây dựng mô hình khi không thực sự cải thiện mô hình khác biệt so với mô hình WLS trước đó.

6.2 Hướng phát triển trong tương lai

Tuy có độ chính xác và tin cậy cao, mô hình vẫn còn bị ảnh hưởng bởi một số điểm ngoại lệ trong dữ liệu. Điều này là do nhóm hướng đến việc phân tích thăm dò là chính, tránh việc chỉnh sửa dữ liệu.

Trong tương lai nếu đồ án được phát triển thêm, chúng tôi sẽ tiền xử lý dữ liệu nhiều hơn, cũng như cải tiến mô hình mạnh mẽ và chính xác cao.

7 Tài liệu tham khảo

[1] J. M. Chambers. *Graphical methods for data analysis*. CRC Press, 2018.
[2] Corporate Finance Institute". Correlation Matrix, 08 2019.
[3] Corporate Finance Institute". R-Squared, 06 2020.
[4] J. Frost. Multiple regression analysis: Use adjusted r-squared and predicted r-squared to include the correct number of variables. *Minitab Blog*, 13(6), 2013.
[5] J. Frost. Heteroscedasticity in Regression Analysis, 03 2019.
[6] S. Glen. Adjusted r2/adjusted r-squared: What is it used for. *Retrieved from Statistics How To: https://www. statisticshowto. datasciencecentral. com/adjusted-r2*, 2013.

[7] I.-C. Yeh. Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete research*, 28(12):1797–1808, 1998.
[8] I.-C. Yeh. Concrete Compressive Strength Data Set, 2007-08-03.
[9] Z. Zach. How to Perform Weighted Least Squares Regression in R, 12 2020.