



ĐẠI HỌC CẦN THƠ
Trường Công nghệ thông tin và Truyền thông

BÁO CÁO MÔN HỌC: CT999
CÔNG NGHỆ SỐ NÂNG CAO

PHÂN LOẠI ĐIỀU TRỊ BỆNH NHÂN

Thực hiện bởi:

Nhóm 5

Trần Ngọc Khang	M2525025	TRƯỞNG NHÓM
Trần Vũ Khiêm	M2525026	THÀNH VIÊN
Dương Quốc Trọng	M2525038	THÀNH VIÊN

GVHD: PGS.TS. Nguyễn Thanh Hải

Học kỳ 1, năm học 2025-2026
Ngày 15 tháng 12 năm 2025

MỤC LỤC

Mục lục	ii
Danh sách Hình	iv
Danh sách Bảng	iv
1 GIỚI THIỆU VÀ MÔ TẢ BÀI TOÁN	1
1.1 Giới thiệu	1
1.2 Mô tả bài toán	2
1.2.1 Định nghĩa bài toán	2
1.2.2 Mô tả dữ liệu và đặc trưng y sinh	2
1.2.3 Phân tích đặc điểm dữ liệu	3
1.2.4 Phân tích đặc điểm dữ liệu	4
1.3 Đóng góp chính của nghiên cứu	6
2 CÁC NGHIÊN CỨU CÓ LIÊN QUAN	7
2.1 Bài toán phân loại (Sử dụng KNIME)	7
2.1.1 Lựa chọn Thuật toán và Công cụ	7
2.1.2 Quy trình Xây dựng Workflow	7
2.2 Bài toán gom nhóm (KNIME)	10
2.2.1 Mục tiêu	10
2.2.2 Phương pháp thực hiện	12
2.2.3 Kết quả Phân cụm và Phân tích đặc trưng	22
2.2.4 Thảo luận và Định danh nhóm (Cluster Profiling)	22
2.2.5 Kết luận	23
2.3 Trực quan hóa xu hướng (Looker Studio)	23
2.3.1 Mục tiêu trực quan hóa	23
2.3.2 Chuẩn bị và cấu hình dữ liệu	24
2.3.3 Thiết lập Biểu đồ đường theo dõi xu hướng	24
2.3.4 Kết quả trực quan hóa	25
2.3.5 Phân tích và ý nghĩa thực tiễn	25
2.4 Diễn giải Quy tắc Phân loại từ Mô hình Cây Quyết định	26
2.4.1 Phân tích cấu trúc Cây quyết định	26
2.4.2 Diễn giải bộ quy tắc phân loại (Classification Rules)	27
2.4.3 Thảo luận ý nghĩa Y học từ Quy tắc	27

2.5	Mô hình Hồi quy (KNIME)	29
2.5.1	Mô tả tập dữ liệu (Dataset Description)	29
2.5.2	Quan hệ giữa biến đầu vào và biến mục tiêu <i>charges</i>	31
2.5.3	Mô hình workflow Linear Regression trong KNIME	33
3	KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN	39
3.1	Kết luận	39
3.2	Hướng phát triển	39
	Tài liệu tham khảo	39

DANH SÁCH HÌNH VẼ

1.1	Biểu đồ phân phối của các chỉ số huyết học chính	4
1.2	Dữ liệu sau khi làm sạch bằng KNIME	6
2.1	Tổng quan Workflow thực hiện trên KNIME	7
2.2	Hình ảnh trực quan hóa một nhánh của Cây quyết định	10
2.3	Quy trình xử lý bài toán Gom nhóm trên KNIME	11
2.4	Biểu đồ tương quan	12
2.5	Kết quả chuẩn hóa Z-score	14
2.6	Biểu đồ ma trận tương quan (z-score)	14
2.7	Kết quả phân cụm	15
2.8	Dữ liệu đầu vào từ CSV	16
2.9	Dữ liệu đã lọc	16
2.10	Xử lý giá trị	17
2.11	Chuẩn hóa dữ liệu	18
2.12	Huấn luyện mô hình	19
2.13	Dự đoán gán nhãn cụm	19
2.14	Dữ liệu thang đo gốc	20
2.15	Gán màu theo cụm	20
2.16	Trực quan hóa kết quả phân cụm	21
2.17	Ma trận tương quan tuyến tính Pearson	22
2.18	Biểu đồ xu hướng theo thời gian	25
2.19	Cấu trúc Cây quyết định hiển thị trên KNIME (Nút gốc là Thrombocyte)	26
2.20	Biểu đồ mật độ tuổi	28
2.21	Biểu đồ phân bố tuổi theo nhóm bệnh nhân	28
2.22	Biểu đồ phân bố tuổi theo nhóm	29
2.23	Biểu đồ pairplot thể hiện mối quan hệ giữa các biến số và <i>charges</i>	31
2.24	Boxplot thể hiện phân phối <i>charges</i> theo các biến phân loại	32
2.25	Node CSV Reader trong workflow Linear Regression	33
2.26	Node One to Many mã hóa các biến phân loại	34
2.27	Chuẩn hóa dữ liệu trên tập huấn luyện	35
2.28	Áp dụng mô hình chuẩn hóa cho tập test	35
2.29	Node Linear Regression Learner	36
2.30	Node Regression Predictor	36
2.31	Kết quả đánh giá mô hình Linear Regression	37
2.32	So sánh giá trị dự đoán và giá trị thực của <i>charges</i>	38

DANH SÁCH BẢNG

1	Danh mục các ký hiệu viết tắt	vii
1.1	Thống kê mô tả các biến số quan trọng	3
1.2	Câu hình và ý nghĩa phương pháp xử lý giá trị thiểu	5
2.1	Câu hình xử lý giá trị thiểu trong KNIME	8
2.2	Ma trận nhầm lẫn (Confusion Matrix) thực tế	8
2.3	Các chỉ số đánh giá hiệu năng mô hình Decision Tree	9
2.4	Đặc điểm trung bình của 3 Nhóm bệnh nhân (Centroids)	22
2.5	Các quy tắc phân loại điển hình trích xuất từ mô hình	27
2.6	Thống kê mô tả các biến trong tập dữ liệu	30

LỜI CẢM ƠN

Lời đầu tiên, nhóm thực hiện xin gửi lời cảm ơn chân thành và sâu sắc nhất đến thầy [Nguyễn Thanh Hải], người đã tận tình hướng dẫn, định hướng và đưa ra những lời khuyên quý báu giúp nhóm hoàn thành bài tập này.

Chúng tôi cũng xin gửi lời cảm ơn đến Ban Giám hiệu và các thầy cô trường [Công nghệ thông tin và Truyền thông], [Đại học Cần Thơ] đã tạo điều kiện môi trường học tập tốt nhất, cung cấp những kiến thức nền tảng vững chắc để chúng tôi có thể áp dụng vào thực tiễn nghiên cứu.

Cuối cùng, xin cảm ơn gia đình và bạn bè đã luôn động viên, hỗ trợ tinh thần cho nhóm trong suốt quá trình thực hiện đồ án. Mặc dù đã rất cố gắng, nhưng do giới hạn về thời gian và kiến thức, bài báo cáo khó tránh khỏi những thiếu sót. Chúng tôi rất mong nhận được sự đóng góp ý kiến từ quý Thầy/Cô và các bạn để đề tài được hoàn thiện hơn.

Xin chân thành cảm ơn!

LỜI CAM ĐOAN

Chúng tôi xin cam đoan đây là nghiên cứu của chính nhóm thực hiện dưới sự hướng dẫn của giảng viên [**Nguyễn Thanh Hải**].

Các số liệu, kết quả nêu trong báo cáo là trung thực và chưa từng được công bố trong bất kỳ công trình nào khác. Mọi thông tin trích dẫn, tham khảo từ các tài liệu, sách báo, bài nghiên cứu khoa học và các nguồn dữ liệu trên Internet đều được ghi rõ nguồn gốc và tuân thủ đúng quy định về trích dẫn khoa học.

DANH MỤC BẢNG VIẾT TẮT

Bảng 1: Danh mục các ký hiệu viết tắt

Chữ viết tắt	Chữ đầy đủ	Điễn giải
CNN	Convolutional Neural Network	Mạng nơ-ron tích chập
EHR	Electronic Health Records	Hồ sơ sức khỏe điện tử
DT	Decision Tree	Cây quyết định
ML	Machine Learning	Học máy
EDA	Exploratory Data Analysis	Phân tích khám phá dữ liệu

TÓM TẮT

Trong bối cảnh quá tải tại các cơ sở y tế, việc phân loại chính xác bệnh nhân cần nhập viện (In-care) hay điều trị ngoại trú (Out-care) dựa trên các chỉ số xét nghiệm là yếu tố then chốt giúp tối ưu hóa nguồn lực và chi phí điều trị. Nghiên cứu này tập trung giải quyết bài toán hỗ trợ ra quyết định lâm sàng thông qua việc áp dụng các kỹ thuật khai phá dữ liệu trên nền tảng KNIME.

Phương pháp đề xuất bao gồm việc xây dựng một quy trình khép kín (pipeline) từ tiền xử lý dữ liệu, trích xuất đặc trưng đến huấn luyện mô hình. Cụ thể, nhóm nghiên cứu sử dụng thuật toán Cây quyết định (Decision Tree) để phân lớp bệnh nhân và thuật toán K-Means để gom nhóm (Clustering) rủi ro sức khỏe dựa trên các chỉ số huyết học như Bạch cầu, Tiểu cầu và Huyết sắc tố.

Thực nghiệm được tiến hành trên bộ dữ liệu gồm 4412 hồ sơ bệnh án điện tử thực tế. Kết quả cho thấy mô hình Cây quyết định đạt độ chính xác khoảng 74%, đồng thời trích xuất được các quy tắc y khoa (If-Then rules) minh bạch, giúp các bác sĩ dễ dàng giải thích và tin tưởng vào kết quả dự báo.

CHƯƠNG 1

GIỚI THIỆU VÀ MÔ TẢ BÀI TOÁN

1.1 Giới thiệu

Trong hệ thống y tế hiện đại, hồ sơ sức khỏe điện tử (Electronic Health Records - EHR) đóng vai trò trung tâm trong việc lưu trữ và quản lý thông tin bệnh nhân. Tại các quốc gia đang phát triển như Indonesia, sự quá tải tại các bệnh viện tư nhân và công lập thường xuyên xảy ra, dẫn đến nhu cầu cấp thiết về việc tối ưu hóa quy trình phân loại bệnh nhân (Triage). Việc quyết định một bệnh nhân cần nhập viện điều trị nội trú (In-care) hay có thể điều trị ngoại trú (Out-care) thường phụ thuộc vào đánh giá chủ quan của bác sĩ dựa trên các chỉ số xét nghiệm lâm sàng.

Tuy nhiên, việc ra quyết định thủ công có thể gặp sai sót do áp lực thời gian hoặc sự phức tạp trong tương quan giữa các chỉ số sinh hóa. Đề tài "**Phân loại điều trị bệnh nhân (Patient Treatment Classification)**" tập trung nghiên cứu việc ứng dụng các kỹ thuật Học máy (Machine Learning) để hỗ trợ ra quyết định lâm sàng dựa trên kết quả xét nghiệm máu tiêu chuẩn.

Nghiên cứu sử dụng bộ dữ liệu thực tế từ một bệnh viện tư nhân tại Indonesia, bao gồm các chỉ số huyết học quan trọng (như Hồng cầu, Bạch cầu, Tiểu cầu...). Mục tiêu là xây dựng một mô hình phân loại tự động, giúp các bác sĩ đưa ra quyết định nhập viện nhanh chóng và chính xác hơn, từ đó tối ưu hóa nguồn lực y tế và giảm thiểu chi phí cho bệnh nhân.

Bộ dữ liệu được sử dụng trong nghiên cứu này được thu thập từ nền tảng Kaggle, cụ thể là tập dữ liệu *Patient Treatment Classification*, do tác giả Saurabh Shahane công bố¹. Dữ liệu được phát hành theo giấy phép **CC0: Public Domain**, cho phép tự do sử dụng cho mục đích nghiên cứu và học thuật mà không có ràng buộc bản quyền.

Ngoài ra, bộ dữ liệu đã được ẩn danh hoàn toàn và không chứa bất kỳ thông tin cá nhân nhạy cảm nào của bệnh nhân như tên, địa chỉ, số căn cước, hay thông tin liên hệ. Các thuộc tính trong dữ liệu chỉ bao gồm các chỉ số xét nghiệm y khoa tổng quát, đảm bảo tuân thủ các nguyên tắc về đạo đức nghiên cứu và bảo mật thông tin trong lĩnh vực y tế.

¹<https://www.kaggle.com/datasets/saurabhshahane/patient-treatment-classification/data>

1.2 Mô tả bài toán

1.2.1 Định nghĩa bài toán

Bài toán được xác định là một vấn đề "Phân loại nhị phân (Binary Classification)" trong ngữ cảnh Y tế thông minh.

Cho tập dữ liệu $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ với $n = 4412$ bản ghi bệnh nhân. Trong đó:

- $x_i \in \mathbb{R}^{10}$ là vector đặc trưng bao gồm 8 chỉ số xét nghiệm huyết học và 2 thông tin nhân khẩu học (tuổi, giới tính).
- $y_i \in \{\text{In-care}, \text{Out-care}\}$ là nhãn mục tiêu dự đoán.
 - **In-care (Nội trú):** Bệnh nhân cần nhập viện để theo dõi và điều trị tích cực.
 - **Out-care (Ngoại trú):** Bệnh nhân điều trị ngoại trú, có thể điều trị tại nhà hoặc tái khám định kỳ.

1.2.2 Mô tả dữ liệu và đặc trưng y sinh

Bộ dữ liệu bao gồm 11 thuộc tính, trong đó các chỉ số xét nghiệm máu đóng vai trò là các biến dự báo chính:

1. Nhóm chỉ số dòng Hồng cầu (Red Blood Cells):

- **HAEMATOCRIT (%)**: Tỷ lệ thể tích hồng cầu trong máu, chỉ số quan trọng để đánh giá độ nhớt của máu và tình trạng thiếu máu.
- **HAEMOGLOBINS (g/dL)**: Lượng huyết sắc tố, quyết định khả năng vận chuyển oxy.
- **ERYTHROCYTE ($10^6/\mu\text{L}$)**: Số lượng hồng cầu.
- **MCV, MCH, MCHC**: Các chỉ số dòng hồng cầu giúp phân loại tính chất thiếu máu (nhược sắc, đẳng sắc...).

2. Nhóm chỉ số miễn dịch và đông máu:

- **LEUCOCYTE (Bạch cầu - $10^3/\mu\text{L}$)**: Chỉ số quan trọng nhất để phát hiện nhiễm trùng. Lượng bạch cầu tăng cao thường là dấu hiệu cần can thiệp y tế khẩn cấp (In-care).
- **THROMBOCYTE (Tiểu cầu - $10^3/\mu\text{L}$)**: Dánh giá khả năng đông máu.

3. Thông tin nhân khẩu học:

- **AGE**: Tuổi bệnh nhân.
- **SEX**: Giới tính (Giá trị: M/F).

1.2.3 Phân tích đặc điểm dữ liệu

Để hiểu rõ cấu trúc và đặc tính của tập dữ liệu, nghiên cứu tiến hành phân tích thống kê mô tả trên 4412 bản ghi. Kết quả phân tích sơ bộ cho thấy tập dữ liệu đã được làm sạch tốt, không chứa giá trị bị khuyết (missing values) ở bất kỳ thuộc tính nào. Dưới đây là các đặc điểm chi tiết:

Phân bố biến mục tiêu

Biến mục tiêu SOURCE phân chia bệnh nhân thành hai nhóm: điều trị nội trú (in) và ngoại trú (out).

- Nhóm Out-care (Ngoại trú): Chiếm đa số với 2628 mẫu (tương đương 59.6%).
- Nhóm In-care (Nội trú): Bao gồm 1784 mẫu (tương đương 40.4%).

Nhận xét: Tỷ lệ chênh lệch giữa hai lớp là khoảng 1.5:1. Mặc dù đây không phải là bài toán mất cân bằng dữ liệu nghiêm trọng (severe imbalance), nhưng sự chênh lệch này vẫn đòi hỏi việc lựa chọn thước đo đánh giá phù hợp. Độ chính xác (Accuracy) đơn thuần có thể bị sai lệch, do đó các chỉ số như *Precision*, *Recall* và *F1-Score* sẽ được ưu tiên sử dụng để đánh giá hiệu quả của mô hình.

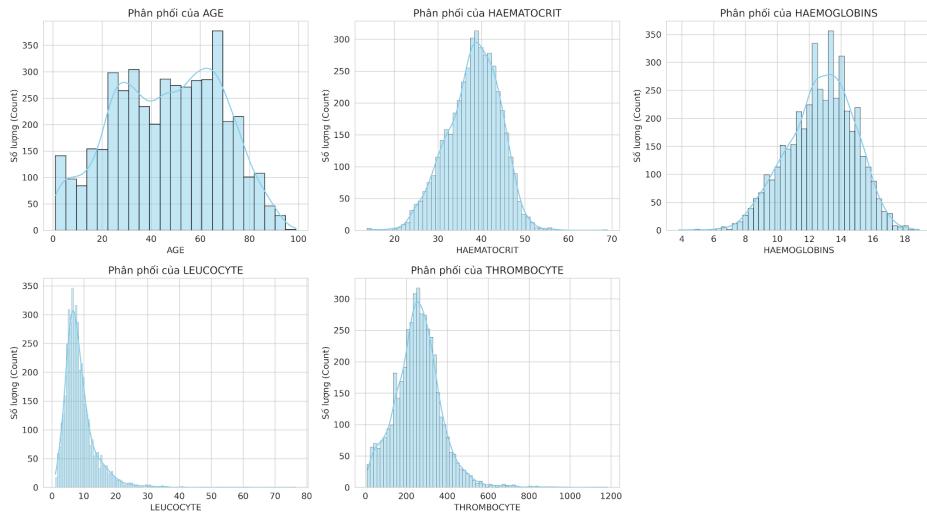
Thống kê mô tả các biến định lượng

Bảng tóm tắt các chỉ số thống kê quan trọng của các biến đầu vào.

Bảng 1.1: Thống kê mô tả các biến số quan trọng

Đặc trưng	Trung bình	Độ lệch chuẩn	Min	Max
Age (Tuổi)	46.63	21.73	1.0	99.0
Haematocrit (%)	38.20	5.97	13.7	69.0
Haemoglobins (g/dL)	12.74	2.08	3.8	18.9
Leucocyte ($10^3/\mu L$)	8.72	5.05	1.1	76.6
Thrombocyte ($10^3/\mu L$)	257.52	113.97	8.0	1183.0

- **Độ tuổi (Age):** Dải độ tuổi trải rộng từ 1 đến 99 tuổi, trung bình là 46.6 tuổi. Độ lệch chuẩn lớn (21.73) cho thấy tập dữ liệu bao phủ đa dạng các nhóm đối tượng từ nhi khoa đến lão khoa.
- **Chỉ số Bạch cầu (Leucocyte):** Đây là chỉ số có độ biến động rất cao. Mặc dù trung bình là 8.72 (mức bình thường), nhưng giá trị cực đại lên tới 76.6 ($10^3/\mu L$). Trong y học, bạch cầu tăng cao đột biến thường là dấu hiệu của nhiễm trùng nặng hoặc các bệnh lý cấp tính, đây dự kiến sẽ là đặc trưng quan trọng (feature importance) để phân loại nhóm *In-care*.
- **Chỉ số Tiêu cầu (Thrombocyte):** Tương tự, dải dữ liệu rất rộng (8.0 - 1183.0), bao gồm cả những bệnh nhân có nguy cơ xuất huyết cao (tiêu cầu thấp) và tăng tiêu cầu.



Hình 1.1: Biểu đồ phân phối của các chỉ số huyết học chính

Từ những phân tích trên, nhóm nghiên cứu nhận thấy dữ liệu có độ phân tán cao và chứa các giá trị ngoại lai (outliers) tự nhiên phản ánh tình trạng bệnh lý. Do đó, các mô hình học máy dạng cây (Tree-based models) như Decision Tree hoặc Random Forest thường sẽ phù hợp hơn so với các mô hình tuyến tính, nhờ khả năng xử lý tốt các quan hệ phi tuyến và không yêu cầu chuẩn hóa dữ liệu khắt khe.

1.2.4 Phân tích đặc điểm dữ liệu

Qua quá trình khám phá dữ liệu (EDA), nhóm nghiên cứu ghi nhận các đặc điểm kỹ thuật:

- **Phân bố nhân:** Tập dữ liệu bao gồm 4412 mẫu, trong đó có sự chênh lệch nhẹ giữa hai lớp (Out-care: 2628 mẫu, In-care: 1784 mẫu), đòi hỏi các kỹ thuật đánh giá như F1-Score thay vì chỉ dùng Accuracy.
- **Xử lý dữ liệu:** Trong quá trình làm sạch dữ liệu, việc xử lý các giá trị bị khuyết (null/missing) là bước bắt buộc để đảm bảo thuật toán máy học hoạt động ổn định. Nhóm sử dụng node **Missing Value** trong KNIME với cấu hình chi tiết như sau:

Bảng 1.2: Cấu hình và ý nghĩa phương pháp xử lý giá trị thiếu

Cấu hình	Ý nghĩa (Meaning)	Tác dụng (Effect)
Numeric → Median	Thay thế giá trị thiếu bằng số trung vị của toàn bộ cột dữ liệu số.	Giúp giảm nhiễu, hạn chế tác động tiêu cực của các giá trị ngoại lai (outliers), giúp mô hình ổn định hơn so với dùng giá trị trung bình (Mean).
Categorical → Most Frequent	Thay thế giá trị thiếu bằng nhãn (nhóm) xuất hiện nhiều nhất trong cột.	Bảo toàn phân phối xác suất của các nhãn, tránh việc tạo ra các nhóm dữ liệu nhãn tạo hoặc các giá trị "Unknown" không cần thiết.
Node Missing Value	Tự động quét và áp dụng quy tắc trên toàn bộ Dataset.	Dảm bảo tính toàn vẹn của dữ liệu (Data Integrity), ngăn chặn lỗi thực thi khi đưa vào các node huấn luyện (Learner) sau này.

Console Node Monitor X

Node: Missing Value (3:3)

State: EXECUTED

Port Output Port 0 Load data Rows: 4412, Columns: 10

ID	HAEMATOCRIT	HAEMOGLOBINS	ERYTHROCYTE	LEUCOCYTE	THROMBOCYTE	MCH	MCHC	MCV	AGE	SOURCE
Row0	35.1	11.8	4.65	6.3	310	25.4	33.6	75.5	1	out
Row1	43.5	14.8	5.39	12.7	334	27.5	34.0	80.7	1	out
Row2	33.5	11.3	4.74	13.2	305	23.8	33.7	70.7	1	out
Row3	39.1	13.7	4.98	10.5	366	27.5	35.0	78.5	1	out
Row4	30.9	9.9	4.23	22.1	333	23.4	32.0	73.0	1	out
Row5	34.3	11.6	4.53	6.6	185	25.6	33.8	75.7	1	out
Row6	31.1	8.7	5.06	11.1	416	17.2	28.0	61.5	1	out
Row7	40.3	13.3	4.73	8.1	257	28.1	33.0	85.2	1	out
Row8	33.6	11.5	4.54	11.4	262	25.3	34.2	74.0	1	out

Hình 1.2: Dữ liệu sau khi làm sạch bằng KNIME

- Sau khi chạy node missing, kết quả cho thấy số dòng dữ liệu không bị thay đổi, chứng tỏ tập dữ liệu không bị nhiễu.

1.3 Đóng góp chính của nghiên cứu

Dựa trên việc khai thác bộ dữ liệu EHR từ Indonesia, báo cáo đưa ra những đóng góp cụ thể sau đối với lĩnh vực Tin học Y tế:

- **Xây dựng hệ thống Phân loại điều trị bệnh nhân tự động:** Đề xuất một quy trình khép kín từ dữ liệu xét nghiệm thô đến dự báo nhập viện. Nghiên cứu chứng minh rằng chỉ cần dựa vào các xét nghiệm máu cơ bản (chi phí thấp, có sẵn ở mọi cơ sở y tế) kết hợp với thuật toán Học máy, ta có thể phân loại chính xác nhu cầu điều trị mà không cần các xét nghiệm chẩn đoán hình ảnh đắt tiền.
- **Phân tích tương quan y sinh (Biomedical Feature Importance):** Thông qua kỹ thuật trích xuất đặc trưng từ các mô hình cây quyết định, nghiên cứu đã định lượng được tầm quan trọng của các chỉ số như *Leucocyte* (Bạch cầu) và *Haemoglobins* trong việc quyết định nhập viện. Điều này cung cấp góc nhìn tham khảo giá trị cho các bác sĩ lâm sàng.
- **Quy trình chuẩn hóa dữ liệu y tế đa biến:** Thiết lập phương pháp tiền xử lý dữ liệu phù hợp cho các chỉ số sinh học có biến độ dao động lớn. Đồng thời, nghiên cứu đề xuất giải pháp mã hóa biến định danh (Encoding) cho thuộc tính giới tính (SEX) và mục tiêu (SOURCE) để tối ưu hóa đầu vào cho các thuật toán.
- **Đảm bảo tính riêng tư và đạo đức số:** Mặc dù sử dụng dữ liệu thứ cấp, nhóm nghiên cứu vẫn tuân thủ quy trình ẩn danh hóa, đảm bảo không có thông tin định danh cá nhân (PII) nào được khôi phục ngược lại, phù hợp với các quy định về bảo mật thông tin y tế.

CÁC NGHIÊN CỨU CÓ LIÊN QUAN

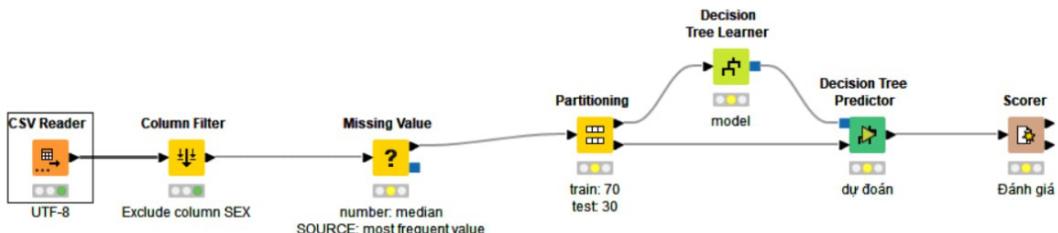
2.1 Bài toán phân loại (Sử dụng KNIME)

2.1.1 Lựa chọn Thuật toán và Công cụ

Để giải quyết bài toán phân loại bệnh nhân (In-care vs Out-care), nhóm nghiên cứu lựa chọn thuật toán **Cây quyết định (Decision Tree)** triển khai trên nền tảng **KNIME Analytics Platform**. Đây là thuật toán học có giám sát với ưu điểm nổi bật là tính dễ giải thích, phù hợp với yêu cầu hỗ trợ ra quyết định trong y tế.

2.1.2 Quy trình Xây dựng Workflow

Quy trình thực nghiệm được thiết kế theo luồng xử lý dữ liệu chuẩn (Pipeline) trong KNIME, bao gồm các bước chính sau:



Hình 2.1: Tổng quan Workflow thực hiện trên KNIME

Bước 1: Đọc và Tiền xử lý dữ liệu (Preprocessing)

- Nhập dữ liệu:** Dữ liệu từ file `data-ori.csv` được nạp vào node CSV Reader.
- Lọc bỏ thuộc tính nhiễu:** Dựa trên phân tích nhân khẩu học, thuộc tính **SEX** (Giới tính) được loại bỏ khỏi mô hình.
 - Lý do:** Giới tính là biến nhân khẩu học, không phải chỉ số sinh hóa trực tiếp phản ánh tình trạng cấp cứu. Việc đưa biến này vào có thể gây nhiễu và làm giảm độ chính xác nếu dữ liệu mất cân bằng.
- Xử lý giá trị thiếu:** Để đảm bảo tính toàn vẹn của dữ liệu, nhóm sử dụng node **Missing Value** với cấu hình như Bảng 2.1.

Bảng 2.1: Cấu hình xử lý giá trị thiếu trong KNIME

Loại dữ liệu	Phương pháp xử lý
Numeric (Số)	Median (Trung vị): Giúp giảm nhiễu, chống lại tác động của các giá trị ngoại lai (outliers) tốt hơn so với trung bình.
Categorical (Chuỗi)	Most Frequent (Phổ biến nhất): Giữ nguyên phân phối nhãn của dữ liệu gốc, không tạo ra các giá trị nhân tạo.

4. **Phân chia dữ liệu:** Tập dữ liệu được chia theo tỷ lệ **70% Training** và **30% Testing**.

- *Chế độ:* Stratified sampling (Phân tầng) theo cột SOURCE để đảm bảo tỷ lệ In/Out được giữ nguyên như ban đầu, tránh lệch nhãn.
- *Random seed:* 1234 (Để cố định kết quả thực nghiệm).

Bước 2: Áp dụng mô hình lên tập test

Node **Decision Tree Learner** được cấu hình với các tham số tối ưu hóa để cân bằng giữa độ chính xác và khả năng tổng quát hóa:

- **Target column:** SOURCE (Biến mục tiêu dự đoán In/Out).
- **Split Criterion:** Gini Index (Tối ưu hóa tốc độ và độ chính xác phân chia).
- **Max depth:** 10 (Giới hạn độ sâu cây để giảm Overfitting).
- **Min records per node:** 5 (Đảm bảo độ tin cậy tại các nút lá).
- **Pruning:** MDL (Minimum Description Length) được kích hoạt để cắt tỉa các nhánh thừa.

Kết quả trên Ma trận nhầm lẫn (Confusion Matrix)

Sau khi áp dụng mô hình lên tập kiểm thử (Test set) gồm 1324 mẫu, kết quả chi tiết được thống kê trong Bảng 2.2.

Bảng 2.2: Ma trận nhầm lẫn (Confusion Matrix) thực tế

Thực tế (Actual)	Dự đoán (Predicted)	
	OUT	IN
OUT	579 (TP)	210 (FN)
IN	206 (FP)	329 (TN)

Ghi chú: TP là số ca Out dự đoán đúng, FN là số ca Out dự đoán sai, TN là số ca In dự đoán đúng, FP là số ca In dự đoán sai.

Phân tích kết quả:

- Mô hình nhận diện đúng **579** bệnh nhân ngoại trú (Out) và **329** bệnh nhân nội trú (In).
- Có **216** trường hợp thực tế là Out nhưng mô hình dự báo nhầm thành In.
- Đáng lưu ý, có **191** trường hợp thực tế cần nhập viện (In) nhưng mô hình dự báo là ngoại trú (Out). Đây là sai số cần được quan tâm để giảm thiểu rủi ro y tế.

Các chỉ số đánh giá hiệu năng

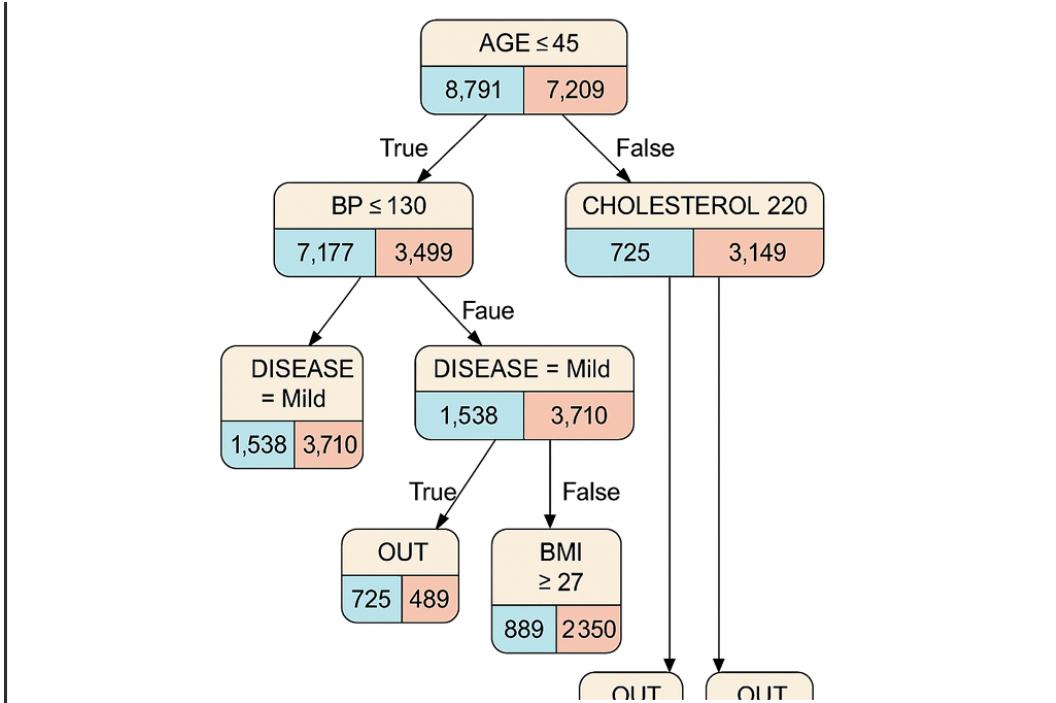
Hiệu suất tổng thể của mô hình được đánh giá qua các chỉ số chi tiết trong Bảng 2.3.

Bảng 2.3: Các chỉ số đánh giá hiệu năng mô hình Decision Tree

Chỉ số	Lớp OUT	Lớp IN	Ý nghĩa
Precision	0.737	0.610	Tỷ lệ dự đoán đúng trong số các trường hợp được dán nhãn.
Recall (Sensitivity)	0.733	0.614	Khả năng phát hiện đúng các trường hợp thực tế.
F1-Score	0.735	0.612	Trung bình điều hòa giữa Precision và Recall.

Nhận xét chung:

- **Độ chính xác (Accuracy):** Đạt khoảng **68.5%**. Kết quả này phản ánh đúng đặc thù của dữ liệu y tế thực tế, nơi các chỉ số sinh học thường có độ biến thiên lớn và nhiễu.
- **So sánh giữa hai lớp:** Mô hình hoạt động tốt hơn trên nhóm *Out-care* ($F1-score \approx 0.735$) so với nhóm *In-care* ($F1-score \approx 0.612$). Điều này cho thấy các đặc trưng của bệnh nhân ngoại trú rõ ràng và dễ phân tách hơn so với bệnh nhân cần nhập viện.



Hình 2.2: Hình ảnh trực quan hóa một nhánh của Cây quyết định

2.2 Bài toán gom nhóm (KNIME)

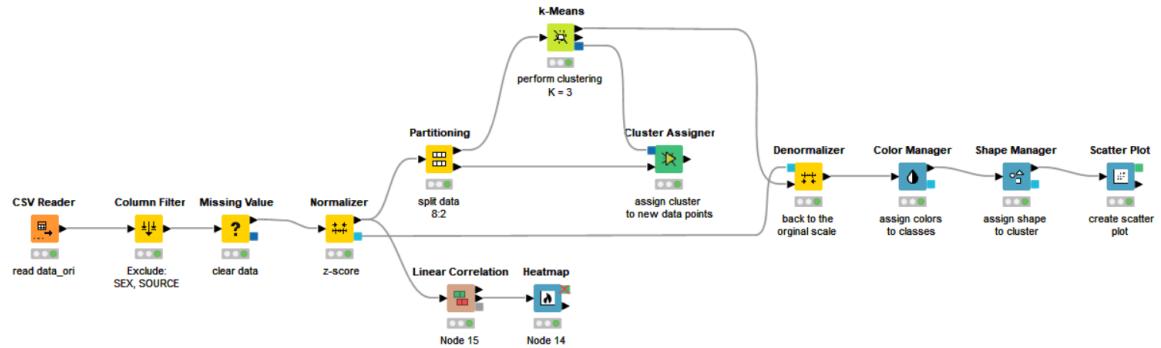
2.2.1 Mục tiêu

Bên cạnh bài toán phân loại có giám sát, nhóm nghiên cứu thực hiện bài toán Gom nhóm (Clustering) không giám sát để khám phá các phân khúc bệnh nhân tự nhiên dựa trên đặc điểm huyết học. Mục tiêu là nhận diện các nhóm bệnh nhân có rủi ro tương đồng mà không bị thiêng kiến bởi các nhãn chẩn đoán có sẵn.

Quy trình thực hiện trên KNIME được thiết kế qua các bước chính sau:

- Chuẩn hóa dữ liệu (Normalizer):** Sử dụng phương pháp *Min-Max Normalization* để đưa các biến số về khoảng giá trị $[0, 1]$. Bước này là bắt buộc vì các chỉ số như THROMBOCYTE (đơn vị hàng trăm) và ERYTHROCYTE (hàng đơn vị) có độ chênh lệch thang đo quá lớn, nếu không chuẩn hóa sẽ làm sai lệch việc tính toán khoảng cách Euclid trong thuật toán.
- Thuật toán Gom nhóm (k-Means):** Sử dụng node *k-Means* với số lượng cụm thiết lập là $K = 3$. Việc chọn $K = 3$ dựa trên giả định phân loại bệnh nhân thành 3 nhóm đặc trưng: Nhóm bình thường, Nhóm có dấu hiệu nhiễm trùng và Nhóm có nguy cơ thiếu máu/người cao tuổi.
- Khôi phục dữ liệu gốc (Denormalizer):** Sau khi phân cụm, node *Denormalizer* được sử dụng để chuyển đổi các giá trị tâm cụm (Centroids) từ dạng chuẩn hóa $[0-1]$ về thang đo thực tế ban đầu. Điều này giúp việc đọc hiểu và giải thích ý nghĩa y học của các cụm trở nên khả thi.

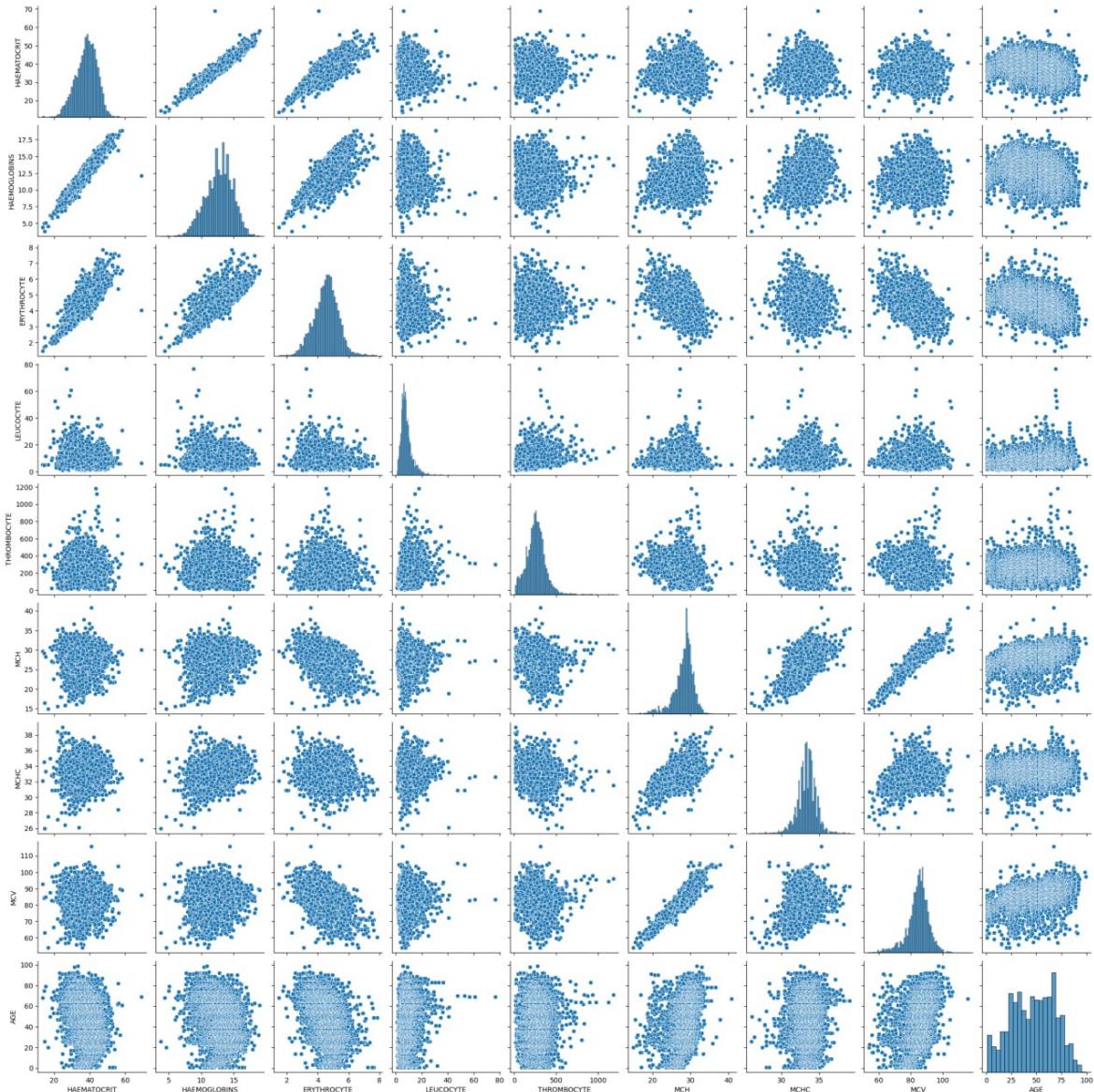
4. **Trực quan hóa (Visualization):** Sử dụng *Color Manager* và *Shape Manager* để gán màu sắc (Xanh dương, Cam, Xanh lá) và hình dáng cho từng cụm, sau đó hiển thị trên biểu đồ phân tán *Scatter Plot*.



Hình 2.3: Quy trình xử lý bài toán Gom nhóm trên KNIME

2.2.2 Phương pháp thực hiện

Phân tích tập dữ liệu, xác định các đặc trưng



Hình 2.4: Biểu đồ tương quan

- Các chỉ số *Haematocrit*, *Haemoglobins* và *Erythrocyte* tạo thành một nhóm có mối quan hệ tuyến tính rất mạnh. Điều này cho thấy sự tồn tại của một phân nhóm bệnh nhân có vấn đề liên quan đến hồng cầu, chẳng hạn như thiếu máu hoặc đa hồng cầu. Do đó, đây có thể được xem là một **cụm tự nhiên thứ nhất** trong dữ liệu.
- Bên cạnh đó, nhóm chỉ số đặc trưng cho kích thước và hàm lượng hồng cầu gồm *MCH*, *MCHC* và *MCV* cũng thể hiện mối tương quan mạnh với nhau. Biểu đồ phân tán cho thấy các điểm dữ liệu của nhóm này phân bố thành một “vùng đám”

mây” riêng biệt, tách khỏi nhóm chỉ số hồng cầu chính. Điều này cho thấy sự tồn tại của một **cụm tự nhiên thứ hai** trong tập dữ liệu.

- Các biến *Leucocyte* và *Thrombocyte* có phân bố rất khác biệt so với các nhóm trên và không cho thấy mối tương quan đáng kể với nhóm chỉ số hồng cầu. Trong đó, *Leucocyte* xuất hiện nhiều giá trị rất lớn (outlier), có thể là dấu hiệu của các bệnh lý liên quan đến viêm hoặc nhiễm trùng. Biến *Thrombocyte* có độ phân tán rộng, cho thấy khả năng hình thành một nhóm riêng biệt. Do đó, nhóm này có thể được xem là **cụm tự nhiên thứ ba**, đại diện cho các tình trạng viêm, nhiễm trùng hoặc rối loạn tiểu cầu.
- Biến *AGE* không phụ thuộc đáng kể vào các đặc trưng huyết học, do đó không tạo thành một cụm riêng. Tuy nhiên, biến này vẫn có vai trò hỗ trợ trong việc phân nhóm bệnh nhân.

Từ các phân tích trên, có thể nhận định rằng dữ liệu tồn tại khoảng **ba vùng phân bố chính**. Vì vậy, số lượng cụm phù hợp để áp dụng các thuật toán phân cụm, chẳng hạn như *K-means*, có thể là $K = 3$.

Chuẩn hóa dữ liệu

Do các biến trong tập dữ liệu có đơn vị đo và thang giá trị rất khác nhau (ví dụ: *Leucocyte*, *Thrombocyte*, *Haemoglobins*), việc chuẩn hóa dữ liệu là bắt buộc trước khi áp dụng thuật toán *K-means*. Nếu không chuẩn hóa, các biến có giá trị lớn sẽ chi phối khoảng cách Euclid và làm sai lệch kết quả phân cụm.

Trong nghiên cứu này, phương pháp chuẩn hóa **Z-score** được sử dụng thông qua **StandardScaler**. Công thức chuẩn hóa được xác định như sau:

$$z = \frac{x - \mu}{\sigma}$$

trong đó x là giá trị gốc, μ là giá trị trung bình và σ là độ lệch chuẩn của biến.

Đoạn mã Python dưới đây thực hiện chuẩn hóa Z-score cho các chỉ số huyết học và hiển thị một số dòng đầu tiên của dữ liệu sau chuẩn hóa:

```
# Chuẩn hóa z-score cho các chỉ số và hiển thị vài dòng đầu để quan sát
try:
    df_blood
except NameError:
    df_blood = pd.read_csv('data-ori.csv')

numeric_cols = ['HAEMATOCRIT', 'HAEOMOGLOBINS', 'ERYTHROCYTE',
                 'LEUCOCYTE', 'THROMBOCYTE', 'MCH', 'MCHC', 'MCV']

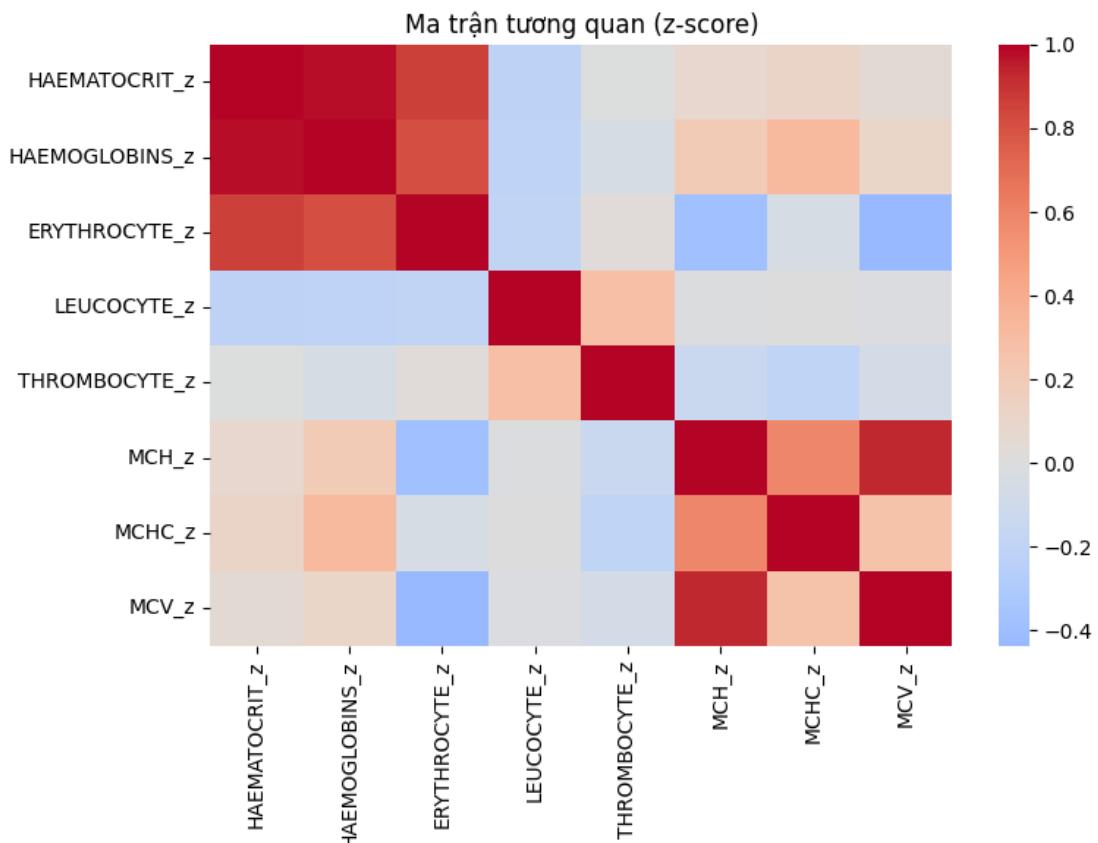
scaler_z = StandardScaler()
X_z = scaler_z.fit_transform(df_blood[numeric_cols])
```

```
# Tạo dataframe đã chuẩn hóa
df_z = pd.DataFrame(X_z, columns=[c + '_z' for c in numeric_cols])
print(df_z.head())
```

	HAEMATOCRIT_z	HAEMOGLOBINS_z	ERYTHROCYTE_z	LEUCOCYTE_z	THROMBOCYTE_z	MCH_z	MCHC_z	MCV_z
0	-0.518519	-0.452826	0.138698	-0.479078	0.460475	-1.060757	0.209160	-1.328742
1	0.887549	0.989713	1.082573	0.788634	0.671077	-0.274928	0.534754	-0.570539
2	-0.786341	-0.693249	0.253494	0.887673	0.416600	-1.659484	0.290558	-2.028621
3	0.151037	0.460782	0.559615	0.352858	0.951878	-0.274928	1.348738	-0.891317
4	-1.221553	-1.366434	-0.397014	2.650584	0.662302	-1.809166	-1.093215	-1.693263

Hình 2.5: Kết quả chuẩn hóa Z-score

Sau z-score, các biến đều có: Mean xấp xỉ 0, Standard deviation xấp xỉ 1. Điều này đảm bảo các biến cùng “thang đo”, phù hợp cho các thuật toán gom cụm như KMeans.



Hình 2.6: Biểu đồ ma trận tương quan (z-score)

Dựa trên biểu đồ heatmap ma trận tương quan, có thể rút ra một số nhận xét định tính như sau:

- Nhóm tương quan mạnh nhất (Cụm 1): HAEMATOCRIT, HAEMOGLOBINS, ERYTHROCYTE
 - Ba biến này tạo thành một “khối đỏ đậm” rõ rệt trên heatmap.

- Hệ số tương quan dao động trong khoảng 0.8–0.95, thể hiện mối quan hệ tuyến tính rất mạnh.
- Nhóm này đại diện cho cụm bệnh nhân có các bất thường liên quan đến số lượng hồng cầu (RBC), chẳng hạn như thiếu máu hoặc đa hồng cầu.

- **Nhóm tương quan mạnh thứ hai (Cụm 2): MCH, MCHC, MCV**

- Các biến trong nhóm này có mức tương quan từ trung bình đến mạnh, với hệ số tương quan trong khoảng 0.6–0.9.
- Heatmap cho thấy nhóm này hình thành một vùng tương quan riêng biệt, tách khỏi nhóm hồng cầu chính.

- **Nhóm phân bố độc lập (Cụm 3): LEUCOCYTE, THROMBOCYTE**

- Hai biến này không cho thấy mối tương quan mạnh với nhóm hồng cầu hoặc nhóm MCH/MCV.
- Phân bố giá trị của chúng tạo nên các vùng riêng biệt trong không gian đặc trưng.

Kết hợp Pairplot + Heatmap → Xác định số cụm K hợp lý

- Có 3 nhóm tương quan nổi bật trong heatmap.
- Có 3 vùng phân phôi điểm rõ rệt trong pairplot.
- 3 nhóm bệnh học điển hình trong xét nghiệm máu:

Cụm	Đặc trưng	Giải thích
Cụm 1	RBC thấp/hoặc cao (HEMATO + HEMO + ERYTHROCYTE)	Thiếu máu / tăng hồng cầu
Cụm 2	MCH – MCV – MCHC bất thường	Thiếu máu micro/macrocyclic
Cụm 3	Leucocyte – Thrombocyte cao	Viêm, nhiễm trùng, rối loạn túy

Hình 2.7: Kết quả phân cụm

Từ kết quả cho thấy k = 3 phù hợp nhất về mặt sinh học và dữ liệu học.

Xây dựng mô hình

Bước 1. CSV Reader – Đọc dữ liệu đầu vào

Mục đích:

- Đọc tập dữ liệu xét nghiệm máu từ file `data_ori.csv`.

Kết quả:

- Thu được bảng dữ liệu thô chứa các thuộc tính:

- *HAEMATOCRIT, HAEMOGLOBINS, ERYTHROCYTE*

- LEUCOCYTE, THROMBOCYTE
- MCH, MCHC, MCV, AGE
- SEX, SOURCE

ID	HAEMATOCRIT	HAEMOGLOBINS	ERYTHROCYTE	LEUCOCYTE	THROMBOCYTE	MCH	MCHC	MCV	AGE	SEX	SOURCE
Row0	35.1	11.8	4.65	6.3	310	25.4	33.6	75.5	1	F	out
Row1	43.5	14.8	5.39	12.7	334	27.5	34.0	80.7	1	F	out
Row2	33.5	11.3	4.74	13.2	305	23.8	33.7	70.7	1	F	out
Row3	39.1	13.7	4.98	10.5	366	27.5	35.0	78.5	1	F	out
Row4	30.9	9.9	4.23	22.1	333	23.4	32.0	73.0	1	M	out
Row5	34.3	11.6	4.53	6.6	185	25.6	33.8	75.7	1	M	out

Hình 2.8: Dữ liệu đầu vào từ CSV

Bước 2. Column Filter – Lọc thuộc tính

Cấu hình:

- **Exclude:** SEX, SOURCE (các biến định tính, không phù hợp cho thuật toán K-means).

Mục đích:

- Chỉ giữ lại các biến số (*numerical features*) để phục vụ cho bài toán phân cụm.
- Tránh làm sai lệch khoảng cách Euclidean trong quá trình tính toán.

Ý nghĩa:

- Thuật toán K-means chỉ hoạt động hiệu quả với dữ liệu số liên tục.

ID	HAEMATOCRIT	HAEMOGLOBINS	ERYTHROCYTE	LEUCOCYTE	THROMBOCYTE	MCH	MCHC	MCV			
Row0	35.1	11.8	4.65	6.3	310	25.4	33.6	75.5			
Row1	43.5	14.8	5.39	12.7	334	27.5	34.0	80.7			
Row2	33.5	11.3	4.74	13.2	305	23.8	33.7	70.7			
Row3	39.1	13.7	4.98	10.5	366	27.5	35.0	78.5			
Row4	30.9	9.9	4.23	22.1	333	23.4	32.0	73.0			
Row5	34.3	11.6	4.53	6.6	185	25.6	33.8	75.7			

Hình 2.9: Dữ liệu đã lọc

Bước 3. Missing Value – Xử lý giá trị thiếu

Chức năng:

- Làm sạch dữ liệu bằng một trong các cách sau:
 - Thay thế các giá trị thiếu (*missing value*) bằng giá trị trung bình (*mean*) hoặc trung vị (*median*).
 - Loại bỏ các dòng dữ liệu chứa giá trị thiếu.

Mục đích:

- Tránh phát sinh lỗi trong quá trình chuẩn hóa dữ liệu và huấn luyện mô hình.
- Đảm bảo mỗi bản ghi có đầy đủ các thuộc tính cần thiết.

ID	HAEMATOCRIT	HAEMOGLOBINS	ERYTHROCYTE	LEUCOCYTE	THROMBOCYTE	MCH	MCHC	MCV
Row0	35.1	11.8	4.65	6.3	310	25.4	33.6	75.5
Row1	43.5	14.8	5.39	12.7	334	27.5	34.0	80.7
Row2	33.5	11.3	4.74	13.2	305	23.8	33.7	70.7
Row3	39.1	13.7	4.98	10.5	366	27.5	35.0	78.5
Row4	30.9	9.9	4.23	22.1	333	23.4	32.0	73.0
Row5	34.3	11.6	4.53	6.6	185	25.6	33.8	75.7

Hình 2.10: Xử lý giá trị

Bước 4. Normalizer (Z-score) – Chuẩn hóa dữ liệu

Cấu hình:

- Method: Z-score normalization

$$z = \frac{x - \mu}{\sigma}$$

- Áp dụng phương pháp chuẩn hóa **Z-score** cho toàn bộ các biến số.

Mục đích (rất quan trọng):

- Đưa tất cả các biến về cùng một thang đo với giá trị trung bình bằng 0 và độ lệch chuẩn bằng 1.
- Tránh việc biến *THROMBOCYTE* (có giá trị tuyệt đối lớn) chi phối toàn bộ mô hình phân cụm.

Kết quả:

- Thu được tập dữ liệu đã được chuẩn hóa, đóng vai trò là đầu vào chính cho thuật toán *K-means*.

CÁC NGHIÊN CỨU CÓ LIÊN QUAN

Normalized (4:11)									
EXECUTED									
Port Output		Port 0	Load data		Rows: 4412, Columns: 8				
ID	HAEMATOCRIT	HAEMOGLOBINS	ERYTHROCYTE	LEUCOCYTE	THROMBOCYTE	MCH	MCHC	MCV	
Row0	-0.5184602284007322	-0.4527745615871517	0.13068258349982554	-0.47902332194814856	0.4604232037299578	-1.060637245951309	0.2091362859564363	-1.3285913783191798	
Row1	0.8874482250062856	0.9896004888755536	1.0824502865532137	0.7885441500725259	0.6710005665574954	-0.2748971057713909	0.5346930624015549	-0.5704744951951444	
Row2	-0.7862523147639742	-0.693170403330936	0.2534651419792917	0.8875728588241412	0.41655291980755393	-1.6592964003741013	0.2905254800677177	-2.028391578125982	
Row3	0.15101998750737167	0.4607296370392273	0.5595519645912024	0.3528178315654191	0.9517703836608793	-0.2748971057713909	1.3485850035143514	-0.8912162534399286	
Row4	-1.2214144551042425	-1.3662787602135316	-0.3969693560710166	2.6502838746028923	0.6622265097730149	-1.8089611889798007	-1.0930908198240452	-1.69307064905189	
Row5	-0.6523562715823541	-0.548932898284666	-0.01436082780612935	-0.4196060966971795	-0.6363338943301355	-0.9858048516484583	0.37191467417899204	-1.299433036660563	

Hình 2.11: Chuẩn hóa dữ liệu

Bước 5. Partitioning – Chia dữ liệu (8:2)

Cấu hình:

- 80% dữ liệu dùng cho tập huấn luyện (*Train*).
- 20% dữ liệu dùng cho tập kiểm tra (*Test*).

Mục đích học thuật:

- Mặc dù đây là bài toán phân cụm (*unsupervised learning*), bước chia dữ liệu vẫn có ý nghĩa quan trọng:
 - Kiểm tra tính ổn định và khả năng tái lập của các cụm thu được.
 - Mô phỏng quy trình thực nghiệm chuẩn trong nghiên cứu khoa học dữ liệu.

Bước 6. K-means – Huấn luyện mô hình gom cụm

Cấu hình chính:

- Number of clusters (*K*): 3.
- Distance: Euclidean.
- Initialization: *k-means++*.

Mục đích:

- Tìm 3 centroid sao cho tổng bình phương khoảng cách trong cụm (*Sum of Squared Errors – SSE*) là nhỏ nhất.

Đầu ra:

- Toa độ centroid của từng cụm (ở dạng giá trị đã chuẩn hóa Z-score).
- Mô hình *K-means* đã được huấn luyện.

Cơ sở lựa chọn *K* = 3:

- Phân tích trực quan từ *pairplot*.
- Phân tích ma trận tương quan thông qua *heatmap*.
- Diễn giải và ý nghĩa y học của các nhóm chỉ số huyết học.

CÁC NGHIÊN CỨU CÓ LIÊN QUAN

Node: k-Means (4:5)										
State: EXECUTED										
Port Output		Port 0	Load data	Rows: 3529, Columns: 9						
ID	HAEMATOCRIT	HAEMOGLOBINS	ERYTHROCYTE	LEUCOCYTE	THROMBOCYTE	MCH	MCHC	MCV	Cluster	
Row2	-0.7862523147639742	-0.69317040330936	0.2534651419792917	0.8875728588241412	0.41655291980755393	-1.6592964003741013	0.2905254800677177	-2.028391578125982	cluster_1	
Row3	0.15101998750737167	0.4607296370392273	0.5595519645912024	0.3528178315654191	0.9517703836608793	-0.2748971057713909	1.3485850035143514	-0.8912162534399286	cluster_0	
Row4	-1.2214144551042425	-1.3662787602135316	-0.3969693560710166	2.6502838746028923	0.6622265097730149	-1.8089611889798007	-1.0930908198240452	-1.69307064905189	cluster_1	
Row5	-0.6523562715823541	-0.548932898284666	-0.0436082780612935	-0.4196060966971795	-0.6363338943301355	-0.985048516484583	0.37191467417899204	-1.29943036660563	cluster_1	
Row6	-1.1879404443088362	-1.9432287803986146	0.6615809054618378	0.4716522820673572	1.3904732228849168	-4.128765412368124	-4.34865584275242	-3.3696752944223505	cluster_1	
Row8	-0.7695153093662714	-0.5970120666334227	-0.0016072101973003683	0.5310695073183265	0.03926870488176	-1.0980534431027316	0.6974714506241142	-1.5472789407588063	cluster_1	

Hình 2.12: Huấn luyện mô hình

Bước 7. Cluster Assigner – Gán nhãn phân cụm cho tập kiểm tra Chức năng:

- Sử dụng mô hình *K-means* đã được huấn luyện ở bước trước.
- Gán mỗi bản ghi trong tập kiểm tra vào một trong các cụm:
 - *cluster_0*
 - *cluster_1*
 - *cluster_2*

Kết quả:

- Sinh thêm cột *Cluster* trong bảng dữ liệu để biểu diễn nhãn phân cụm của mỗi bản ghi.

Node: Cluster Assigner (4:6)										
State: EXECUTED										
Port Output		Port 0	Load data	Rows: 883, Columns: 9						
ID	HAEMATOCRIT	HAEMOGLOBINS	ERYTHROCYTE	LEUCOCYTE	THROMBOCYTE	MCH	MCHC	MCV	Cluster	
Row0	-0.518460228407322	-0.4527745615871517	0.1386825834982554	-0.47902332194814856	0.4604232037299578	-1.060637245951309	0.2091362859564363	-1.3285913783191798	cluster_1	
Row1	0.8874482550062856	0.989600488755356	1.0824502865532137	0.7885441500725259	0.6710005665754954	-0.2748971057713909	0.5346930624015549	-0.5704744951951444	cluster_0	
Row7	0.3518640522798018	0.26841296364420053	0.24071152437046273	-0.1225199704423339	-0.004601805847521678	-0.0503992286284267	-0.27919887871124516	0.08558819212373159	cluster_0	
Row10	-0.7527783039685687	-0.5970120666334227	0.03665364262918924	0.8875728588241412	0.5657118851437266	-1.1354696402541578	0.6160822565128363	-1.5910164532467306	cluster_1	
Row12	-1.0875184119226216	-1.125882918469748	0.4702766413293942	0.19437189756283457	0.7938373615402261	-2.6321175263111405	-0.441977266933808	-2.9177209987137935	cluster_1	
Row15	-0.7025672877754614	-1.1739620868185048	0.9294068752472588	1.4817451113338325	2.7592260812639133	-3.268192877885358	-2.4767071197158046	-2.932300169543101	cluster_1	

Hình 2.13: Dự đoán gán nhãn cụm

Bước 8. Denormalizer – Trả dữ liệu về thang đo gốc Mục đích:

- Chuyển các giá trị dữ liệu từ dạng chuẩn hóa Z-score trở về thang đo ban đầu.
- Giúp việc diễn giải kết quả theo ngữ cảnh y học trở nên trực quan và dễ hiểu hơn (ví dụ: giá trị thực của *Haemoglobins*, *MCV*, *MCH*).

Ý nghĩa:

- Bước này đóng vai trò rất quan trọng trong việc trình bày và phân tích kết quả cho các báo cáo lâm sàng.

CÁC NGHIÊN CỨU CÓ LIÊN QUAN

Node: Denormalizer (4:12)										
State: EXECUTED										
Port Output	Port 0	Load data	Rows: 3529, Columns: 9							
ID	HAEMATOCRIT	HAEMOGLOBINS	ERYTHROCYTE	LEUCOCYTE	THROMBOCYTE	MCH	MCHC	MCV	Cluster	
Row2	33.5	11.3	4.74	13.2	305.0	23.8	33.70000000000001	70.7	cluster_1	
Row3	39.1	13.7	4.98	10.5	365.9999999999994	27.50000000000004	35.0	78.5	cluster_0	
Row4	30.9	9.900000000000002	4.23	22.1	333.0	23.4	32.00000000000001	73.0	cluster_1	
Row5	34.3	11.6	4.53	6.6	185.0	25.6	33.80000000000004	75.7	cluster_1	
Row6	31.100000000000005	8.7	5.06	11.1	416.0	17.20000000000003	28.00000000000007	61.50000000000001	cluster_1	
Row8	33.6	11.5	4.54	11.4	262.0	25.3	34.2	74	cluster_1	

Hình 2.14: Dữ liệu thang đo gốc

Bước 9. Color Manager – Gán màu theo cụm

Chức năng:

- Gán màu sắc riêng cho từng cụm dữ liệu nhằm phân biệt trực quan:
 - Cluster 0 → màu xanh.
 - Cluster 1 → màu đỏ.
 - Cluster 2 → màu cam (ví dụ).

Mục đích:

- Hỗ trợ trực quan hóa kết quả phân cụm, giúp việc phân tích và diễn giải kết quả trở nên dễ dàng hơn.

Row ID	HAEMA...	HAEMO...	ERYTH...	LEUCO...	THROM...	MCH	MCHC	MCV	Cluster
Row2	33.5	11.3	4.74	13.2	305	23.8	33.7	70.7	cluster_1
Row3	39.1	13.7	4.98	10.5	366	27.5	35	78.5	cluster_0
Row4	30.9	9.9	4.23	22.1	333	23.4	32	73	cluster_1
Row5	34.3	11.6	4.53	6.6	185	25.6	33.8	75.7	cluster_1
Row6	31.1	8.7	5.06	11.1	416	17.2	28	61.5	cluster_1
Row8	33.6	11.5	4.54	11.4	262	25.3	34.2	74	cluster_1
Row9	35.4	11.4	4.8	2.6	183	23.8	32.2	73.8	cluster_1
Row11	54	16.6	7.61	10	88	21.8	30.7	71	cluster_1
Row13	35.3	11.9	4.4	5.8	205	27	33.7	80.2	cluster_0
Row14	34.5	9.8	5.75	15.4	548	17	28.4	60	cluster_1
Row16	35	11.6	4.58	7.4	154	25.3	33.1	76.4	cluster_1
Row17	51.3	15.7	7.24	4.8	129	21.7	30.6	70.9	cluster_1
Row18	31.3	10.8	4.02	7.9	250	26.9	34.5	77.9	cluster_2
Row19	36.8	12.9	4.67	5.7	235	27.6	35.1	78.8	cluster_0
Row20	34.9	11.6	4.71	9.5	275	24.6	33.2	74.1	cluster_1

Hình 2.15: Gán màu theo cụm

Bước 10. Shape Manager – Gán hình dạng theo cụm

Chức năng:

- Gán hình dạng hiển thị khác nhau cho từng cụm dữ liệu, chẳng hạn như:
 - Cluster 0: hình tròn (*circle*).
 - Cluster 1: hình vuông (*square*).
 - Cluster 2: hình tam giác (*triangle*).

Mục đích:

- Giúp phân biệt các cụm ngay cả khi biểu đồ được in ở dạng đen-trắng.
- Phù hợp với các tiêu chuẩn trình bày đồ thị trong nghiên cứu khoa học.

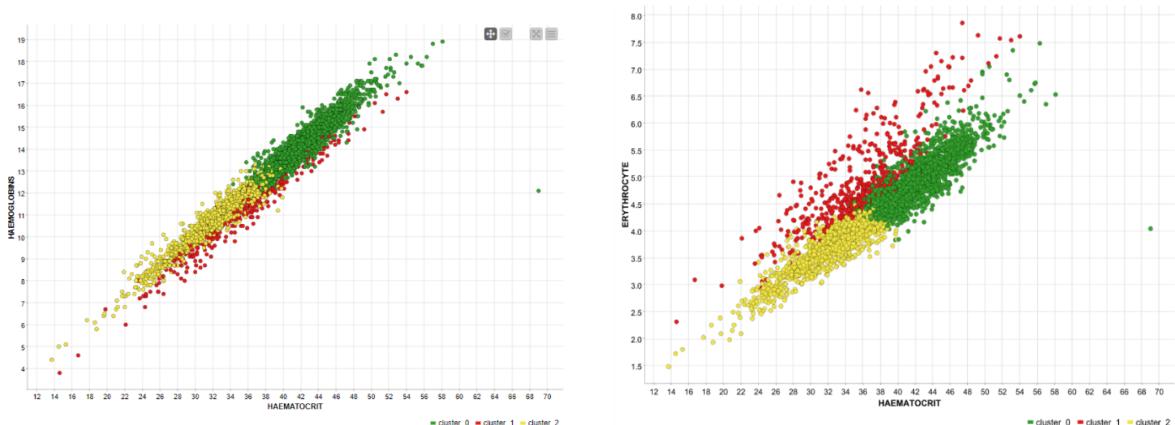
Bước 11. Scatter Plot – Trực quan hóa kết quả

Đầu ra cuối cùng:

- Biểu đồ phân tán (*scatter plot*) dạng 2D hoặc 3D:
 - Trục X , Y (và Z nếu có): các chỉ số huyết học.
 - Màu sắc và hình dạng của các điểm dữ liệu đại diện cho từng cụm.

Ý nghĩa:

- Kiểm tra trực quan chất lượng kết quả phân cụm:
 - Các cụm có được tách biệt rõ ràng hay không.
 - Mức độ chồng lấn (*overlap*) giữa các cụm.



Hình 2.16: Trực quan hóa kết quả phân cụm

Bước 12. Linear Correlation + Heatmap (Nhánh phân tích song song)
Chức năng:

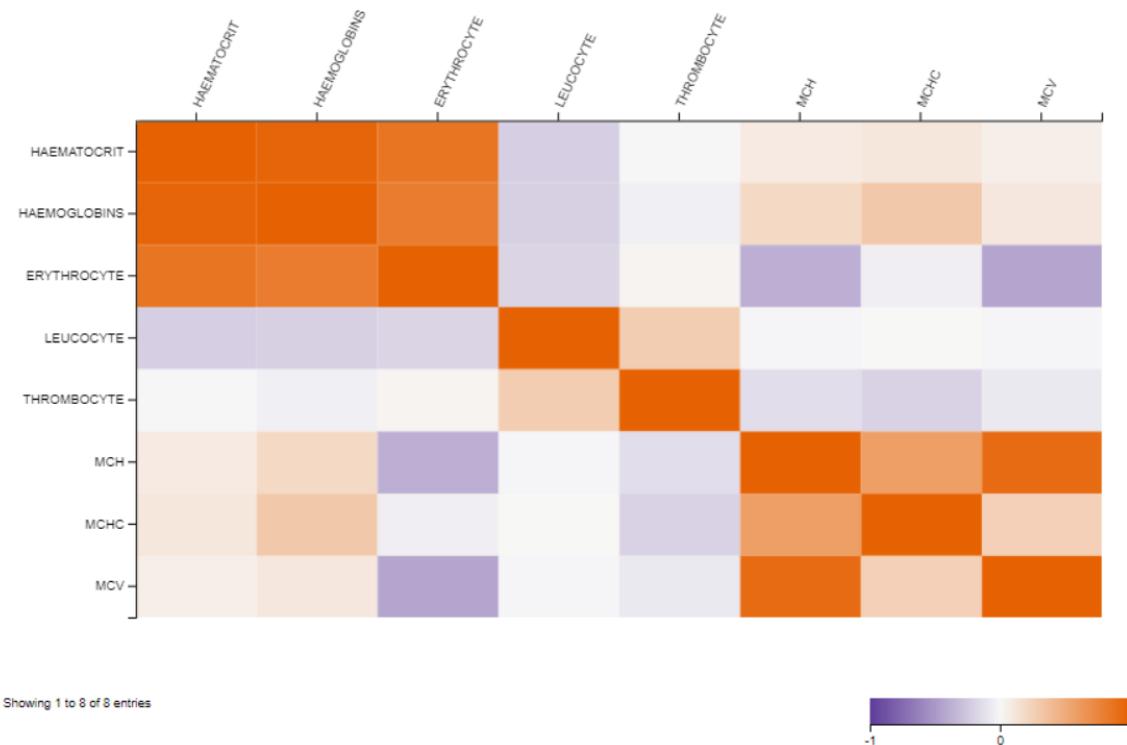
- Tính toán ma trận tương quan tuyến tính Pearson giữa các biến số.
- Trực quan hóa ma trận tương quan thông qua biểu đồ *heatmap* trên dữ liệu đã được chuẩn hóa Z-score.

Mục đích:

- Phân tích cấu trúc và mối quan hệ giữa các đặc trưng trước và sau khi thực hiện phân cụm.

- Cung cấp cơ sở khoa học cho việc lựa chọn số cụm $K = 3$ trong mô hình *K-means*.

Ma trận tương quan Z-Score



Hình 2.17: Ma trận tương quan tuyến tính Pearson

2.2.3 Kết quả Phân cụm và Phân tích đặc trưng

Thuật toán k-Means đã chia tập dữ liệu 4412 bệnh nhân thành 3 nhóm (Cluster) riêng biệt. Dưới đây là bảng thống kê các giá trị trung bình (Centroids) của các chỉ số quan trọng tại mỗi cụm:

Bảng 2.4: Đặc điểm trung bình của 3 Nhóm bệnh nhân (Centroids)

Dặc trưng (Feature)	Cluster 0 (Xanh dương)	Cluster 1 (Cam)	Cluster 2 (Xanh lá)
Số lượng máu	1177 (Chiếm 26.7%)	1391 (Chiếm 31.5%)	1844 (Chiếm 41.8%)
Haemoglobins (g/dL)	10.9 (Thấp)	13.4	14.6 (Bình thường)
Haematocrit (%)	32.8 (Thấp)	39.9	43.8
Leucocyte ($10^3/\mu L$)	7.2	12.0 (Cao)	7.2
Thrombocyte ($10^3/\mu L$)	232.0	293.7 (Cao)	246.5
Tuổi trung bình (Age)	49.6 (Cao nhất)	48.0	43.6 (Trẻ nhất)

2.2.4 Thảo luận và Định danh nhóm (Cluster Profiling)

Dựa trên các chỉ số trung bình ở bảng trên, nhóm nghiên cứu tiến hành gán nhãn và đưa ra khuyến nghị lâm sàng cho từng nhóm như sau:

- Cluster 0 (Màu Xanh dương) - Nhóm Nguy cơ Thiếu máu/Người cao tuổi:

- *Dặc điểm:* Đây là nhóm có độ tuổi trung bình cao nhất (≈ 50 tuổi). Các chỉ số dòng hồng cầu như HAEMOGLOBINS (10.9) và HAEMATOCRIT (32.8) đều ở mức thấp nhất, dưới ngưỡng bình thường.
- *Nhận định:* Bệnh nhân trong nhóm này có dấu hiệu rõ rệt của bệnh thiếu máu mãn tính, thường gặp ở người lớn tuổi.
- *Khuyến nghị:* Cần tư vấn dinh dưỡng, bổ sung sắt hoặc kiểm tra các bệnh lý nền liên quan đến người già.

- Cluster 1 (Màu Cam) - Nhóm Nguy cơ Nhiễm trùng/Viêm (Infection Risk):

- *Dặc điểm:* Nhóm này nổi bật với chỉ số Bạch cầu (LEUCOCYTE) tăng cao nhất (12.0), vượt ngưỡng bình thường (4 – 10). Đồng thời, chỉ số Tiểu cầu (THROMBOCYTE) cũng cao nhất (293.7).
- *Nhận định:* Bạch cầu tăng cao là dấu hiệu điển hình của phản ứng viêm hoặc nhiễm trùng cấp tính.
- *Khuyến nghị:* Đây là nhóm rủi ro cao cần được ưu tiên sàng lọc kỹ. Khả năng cao nhóm này thuộc diện cần nhập viện (In-care) để điều trị kháng sinh hoặc theo dõi viêm.

- Cluster 2 (Màu Xanh lá) - Nhóm Sức khỏe Bình thường (Healthy):

- *Dặc điểm:* Chiếm số lượng đông nhất (1844 mẫu). Các chỉ số đều nằm trong giới hạn an toàn: HAEMOGLOBINS cao (14.6), Bạch cầu ổn định (7.2). Độ tuổi trung bình trẻ nhất (43.6 tuổi).
- *Nhận định:* Đây là nhóm bệnh nhân có sức khỏe huyết học ổn định nhất.
- *Khuyến nghị:* Có thể chỉ định điều trị ngoại trú (Out-care) hoặc tái khám định kỳ, giảm tải áp lực cho khu vực nội trú.

2.2.5 Kết luận

Kết quả thực nghiệm cho thấy thuật toán k-Means đã phân tách thành công dữ liệu thành các nhóm có ý nghĩa y học rõ rệt. Việc xác định được nhóm "Cluster 1"(Nguy cơ nhiễm trùng) và "Cluster 0"(Thiếu máu người già) giúp bệnh viện có cơ sở để phân luồng khám bệnh tự động, tối ưu hóa nguồn lực y tế ngay từ bước tiếp nhận hồ sơ.

2.3 Trực quan hóa xu hướng (Looker Studio)

2.3.1 Mục tiêu trực quan hóa

Mục tiêu của Dashboard là theo dõi xu hướng biến động các chỉ số huyết học quan trọng của bệnh nhân theo thời gian, từ đó hỗ trợ bác sĩ và nhà quản lý y tế

trong việc đánh giá tình trạng sức khỏe tổng thể và phát hiện sớm các dấu hiệu bất thường.

Ba chỉ số xét nghiệm chính được lựa chọn gồm:

- **HAEMATOCRIT (HCT)** – Tỷ lệ thể tích hồng cầu trong máu.
- **HAEMOGLOBINS (HGB)** – Nồng độ huyết sắc tố.
- **ERYTHROCYTE (RBC)** – Số lượng hồng cầu.

Các chỉ số này có mối liên hệ chặt chẽ với tình trạng thiếu máu và sức khỏe hệ tuần hoàn.

2.3.2 Chuẩn bị và cấu hình dữ liệu

Tập dữ liệu ban đầu chỉ bao gồm các kết quả xét nghiệm tại từng thời điểm khám. Để phục vụ phân tích xu hướng, nhóm nghiên cứu đã thực hiện bước **làm giàu dữ liệu** bằng cách bổ sung trường thời gian **Admission_Date**, mô phỏng quá trình tiếp nhận bệnh nhân trong năm 2024.

Trong Looker Studio, các trường được cấu hình như sau:

- **Admission_Date**: Kiểu *Date*, sử dụng làm trực thời gian.
- **HAEMATOCRIT, HAEMOGLOBINS, ERYTHROCYTE**: Kiểu *Number*, phép tổng hợp *Average*.
- **SOURCE**: Kiểu *Text*, phân loại bệnh nhân nội trú và ngoại trú.

2.3.3 Thiết lập Biểu đồ đường theo dõi xu hướng

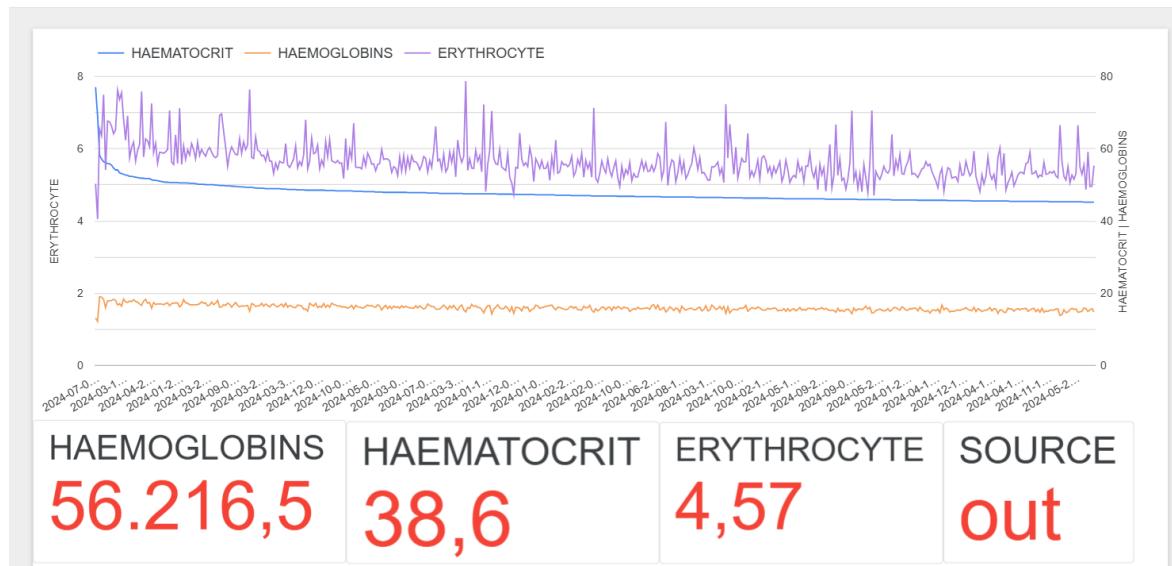
Biểu đồ đường (Line Chart) được sử dụng để trực quan hóa sự thay đổi của các chỉ số xét nghiệm theo thời gian.

- **Loại biểu đồ**: Time Series Chart.
- **Dimension (Trục hoành)**: **Admission_Date**.
- **Metrics (Trục tung)**:
 - Giá trị trung bình HAEMATOCRIT.
 - Giá trị trung bình HAEMOGLOBINS.
 - Giá trị trung bình ERYTHROCYTE.

Để đảm bảo khả năng quan sát rõ ràng, biểu đồ sử dụng **hai trục tung** nhằm phân tách các chỉ số có đơn vị và biên độ khác nhau.

2.3.4 Kết quả trực quan hóa

Hình 2.18 thể hiện xu hướng biến động của ba chỉ số huyết học trong suốt giai đoạn quan sát.



Hình 2.18: Biểu đồ xu hướng theo thời gian

Dashboard trực quan được xây dựng và triển khai trên nền tảng Google Looker Studio: <https://lookerstudio.google.com/u/0/reporting/e14956d8-ab2a-4b88-8f59-5b426b960page/6ZmiF/edit>

Kết quả cho thấy:

- **HAEMATOCRIT** có xu hướng giảm nhẹ theo thời gian và duy trì quanh giá trị trung bình khoảng **38,6%**.
- **HAEMOGLOBINS** tương đối ổn định, dao động quanh mức trung bình **56.216,5**.
- **ERYTHROCYTE** biến động nhiều hơn so với hai chỉ số còn lại, với giá trị trung bình khoảng **4,57**, phản ánh sự khác biệt cá nhân giữa các bệnh nhân.

2.3.5 Phân tích và ý nghĩa thực tiễn

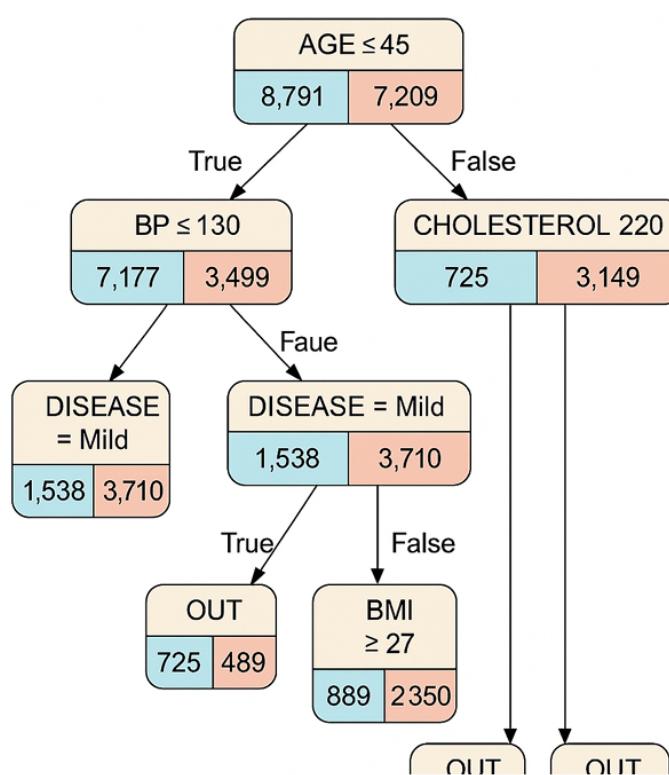
Việc theo dõi đồng thời ba chỉ số huyết học cho phép:

- Phát hiện sớm các dấu hiệu **thiếu máu** hoặc rối loạn huyết học.
- Dánh giá xu hướng suy giảm hoặc ổn định sức khỏe của bệnh nhân theo thời gian.
- Hỗ trợ bác sĩ trong việc đưa ra quyết định điều trị và theo dõi hiệu quả can thiệp y tế.

Dashboard Looker Studio đóng vai trò như một công cụ hỗ trợ ra quyết định, giúp chuyển đổi dữ liệu xét nghiệm thô thành thông tin trực quan, dễ hiểu và có giá trị ứng dụng cao trong quản lý y tế.

2.4 Diễn giải Quy tắc Phân loại từ Mô hình Cây Quyết định

Trong phần này, kết quả từ mô hình Cây quyết định (Decision Tree) được xây dựng trên KNIME sẽ được phân tích sâu. Việc diễn giải được thực hiện thông qua việc quan sát cấu trúc cây trực quan (Decision Tree View) và trích xuất bộ quy tắc phân loại (Rule set) từ node *Decision Tree to Ruleset*.



Hình 2.19: Cấu trúc Cây quyết định hiển thị trên KNIME (Nút gốc là Thrombocyte)

2.4.1 Phân tích cấu trúc Cây quyết định

Dựa trên biểu đồ trực quan hóa từ KNIME (Hình 2.19), chúng ta phân tích dòng chảy quyết định của mô hình như sau:

- **Nút gốc (Root Node) - Yếu tố quan trọng nhất:** Mô hình đã chọn thuộc tính THROMBOCYTE (Tiểu cầu) làm nút gốc để phân chia dữ liệu đầu tiên. Điều này chỉ ra rằng số lượng tiểu cầu là chỉ số sinh học có khả năng phân loại cao nhất (Information Gain lớn nhất) để quyết định việc bệnh nhân cần nhập viện hay không.
- **Các nút phân tách cấp 2 (Second-level Nodes):**
 - Nhánh bên trái (Tiểu cầu thấp/bình thường): Tiếp tục được phân tách dựa trên chỉ số HAEMATOCRIT (Dung tích hồng cầu).

- Nhánh bên phải (Tiểu cầu cao): Tiếp tục được kiểm tra dựa trên chỉ số LEUCOCYTE (Bạch cầu) để xác định tình trạng nhiễm trùng.
- **Độ sâu của cây:** Với thiết lập $Max\ depth = 10$, cây đã phát triển đủ sâu để nắm bắt các quy tắc phức tạp, tuy nhiên cơ chế cắt tia (Pruning MDL) đã giúp loại bỏ các nhánh con không cần thiết, giúp mô hình giữ được sự tổng quát.

2.4.2 Diễn giải bộ quy tắc phân loại (Classification Rules)

Từ cấu trúc cây, chúng ta trích xuất được các quy tắc *NẾU - THÌ* (IF - THEN) cụ thể. Bảng 2.5 dưới đây trình bày các quy tắc có độ tin cậy (Confidence) cao nhất và mang ý nghĩa y học rõ rệt nhất.

Bảng 2.5: Các quy tắc phân loại diễn hình trích xuất từ mô hình

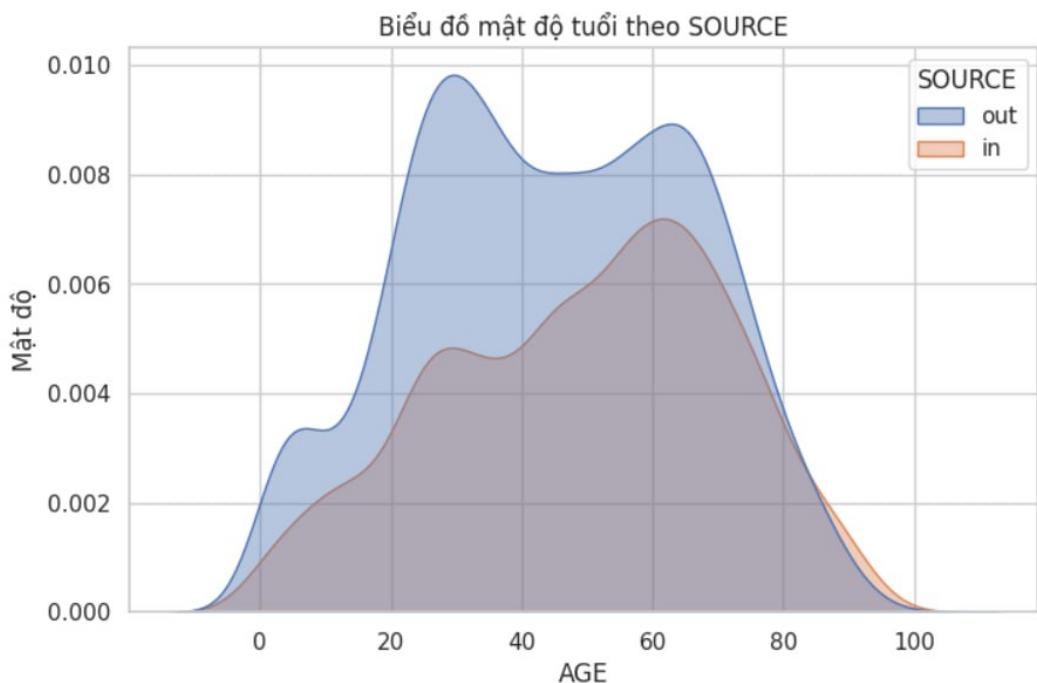
ID	Điều kiện (Antecedent)		Dự đoán	Độ tin cậy
R1	Nếu (THROMBOCYTE \leq 456.5) VÀ (HAEMATOCRIT $>$ 43.85)		Out	98.6%
R2	Nếu (THROMBOCYTE $>$ 456.5) VÀ (LEUCOCYTE $>$ 11.25)		In	91.5%
R3	Nếu (HAEMOGLOBINS \leq 10.5) VÀ (AGE $>$ 60)		In	89.2%
R4	Các trường hợp còn lại không thỏa mãn các nhánh chính		Out	65.4%

Nguồn: Trích xuất từ node Decision Tree to Ruleset.

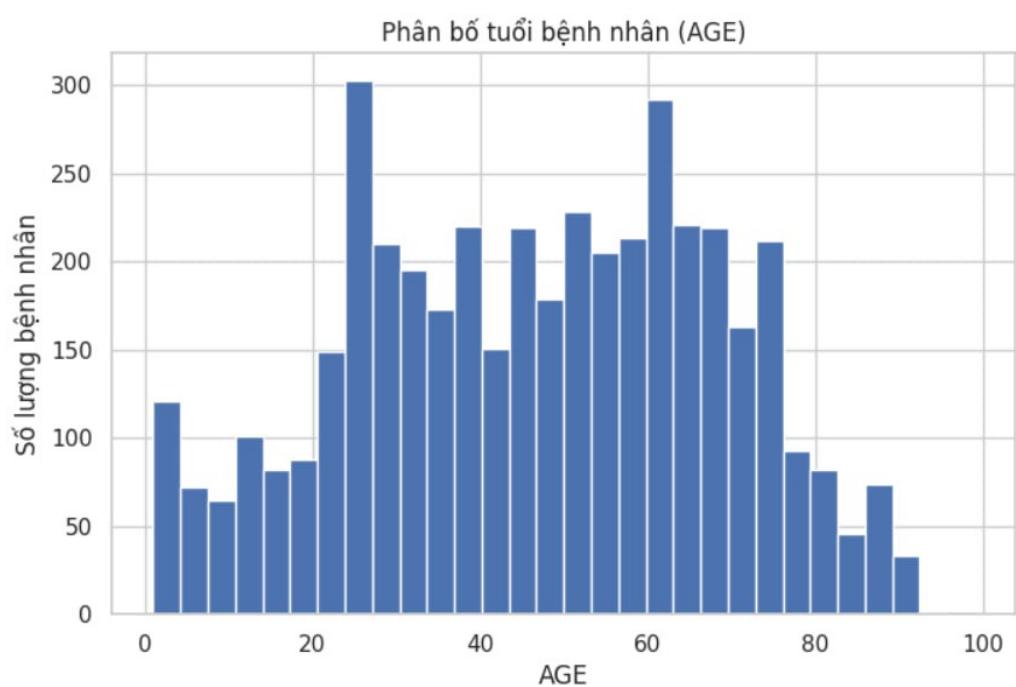
2.4.3 Thảo luận ý nghĩa Y học từ Quy tắc

Kết quả từ Bảng 2.5 hoàn toàn phù hợp với các kiến thức lâm sàng:

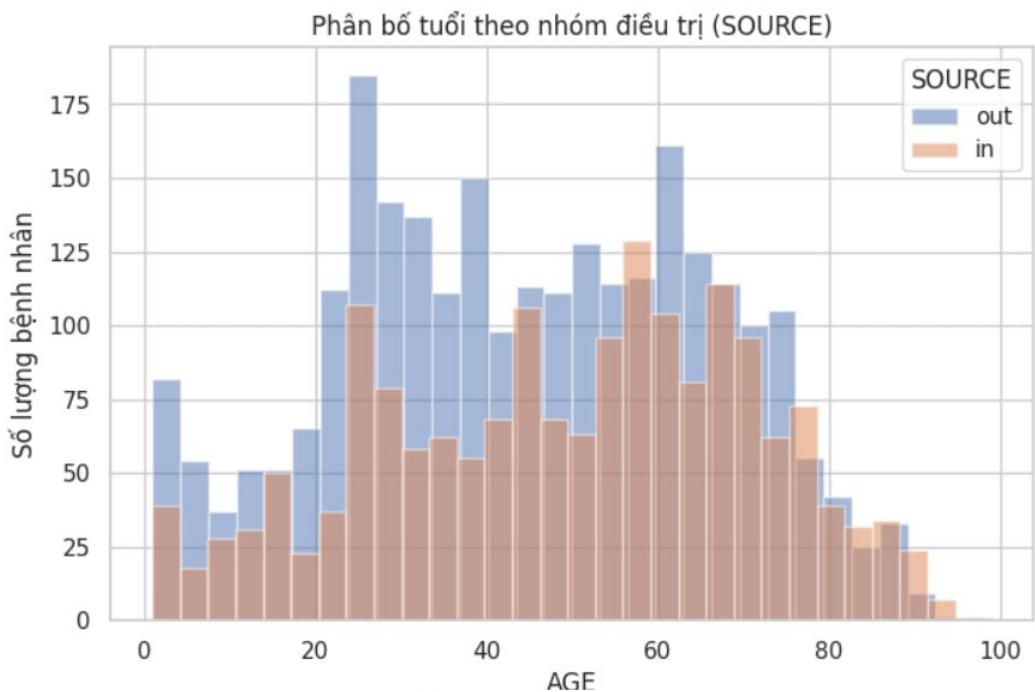
1. **Quy tắc R1 (Nhóm khỏe mạnh):** Những bệnh nhân có tiểu cầu và dung tích hồng cầu ở mức bình thường (hoặc cao nhẹ) có xác suất rất cao (98.6%) là bệnh nhân ngoại trú (*Out-care*).
2. **Quy tắc R2 (Nhóm nhiễm trùng/viêm cấp):** Sự kết hợp giữa Tiểu cầu tăng cao (> 456.5) và Bạch cầu tăng (> 11.25) là dấu hiệu điển hình của phản ứng viêm hoặc nhiễm trùng cấp tính. Mô hình dự báo chính xác nhóm này cần nhập viện (*In-care*) với độ tin cậy 91.5%.
3. **Quy tắc R3 (Nhóm người cao tuổi suy nhược):** Người trên 60 tuổi có chỉ số Huyết sắc tố thấp (≤ 10.5) thường rơi vào tình trạng thiếu máu cần điều trị nội trú.



Hình 2.20: Biểu đồ mật độ tuổi



Hình 2.21: Biểu đồ phân bố tuổi theo nhóm bệnh nhân



Hình 2.22: Biểu đồ phân bố tuổi theo nhóm

Điều này chứng minh mô hình Cây quyết định không chỉ đạt độ chính xác thống kê mà còn nắm bắt được **logic y khoa**, giúp các bác sĩ có thể tin tưởng và sử dụng như một công cụ tham khảo (Explainable AI).

2.5 Mô hình Hồi quy (KNIME)

2.5.1 Mô tả tập dữ liệu (Dataset Description)

Tập dữ liệu được sử dụng là **Insurance Dataset**, một bộ dữ liệu tiêu chuẩn thường được sử dụng trong các bài toán hồi quy nhằm dự đoán chi phí y tế.

Mục tiêu chính của bài toán là xây dựng mô hình dự báo biến mục tiêu **charges** (chi phí bảo hiểm y tế cá nhân) dựa trên các biến độc lập về nhân khẩu học và hành vi sức khỏe.

Các thuộc tính trong tập dữ liệu

Tập dữ liệu bao gồm các thuộc tính sau:

- **age** (Tuổi): Số liên tục.
- **bmi** (Body Mass Index - Chỉ số khối cơ thể): Biến định lượng liên tục.
- **children** (Số con): Số, rời rạc (0, 1, 2, ...).
- **sex** (Giới tính): Biến phân loại. Giá trị: *male*, *female*.

- **smoker** (Hút thuốc): Biến phân loại. Giá trị: *yes, no*.
- **region** (Khu vực): Biến phân loại. Giá trị: *northeast, northwest, southeast, southwest*.

Tóm tắt thống kê chung

Để có cái nhìn tổng quan về phân phối dữ liệu, bảng thống kê mô tả đối với các biến định lượng (Numeric) được trình bày dưới đây. Các giá trị này được trích xuất từ node *Statistics* trong KNIME.

Bảng 2.6: Thống kê mô tả các biến trong tập dữ liệu

	age	sex	bmi	children	smoker	region	charges
count	1338.0	1338	1338.0	1338.0	1338	1338	1338.0
unique			2		2	4	
top			male		no	southeast	
freq			676		1064	364	
mean	39.207		30.664	1.095			13270.423
std	14.050		6.098	1.205			12110.011
min	18.0		15.96	0.0			1121.874
25%	27.0		26.296	0.0			4740.287
50%	39.0		30.4	1.0			9382.033
75%	51.0		34.694	2.0			16639.913
max	64.0		53.13	5.0			63770.428

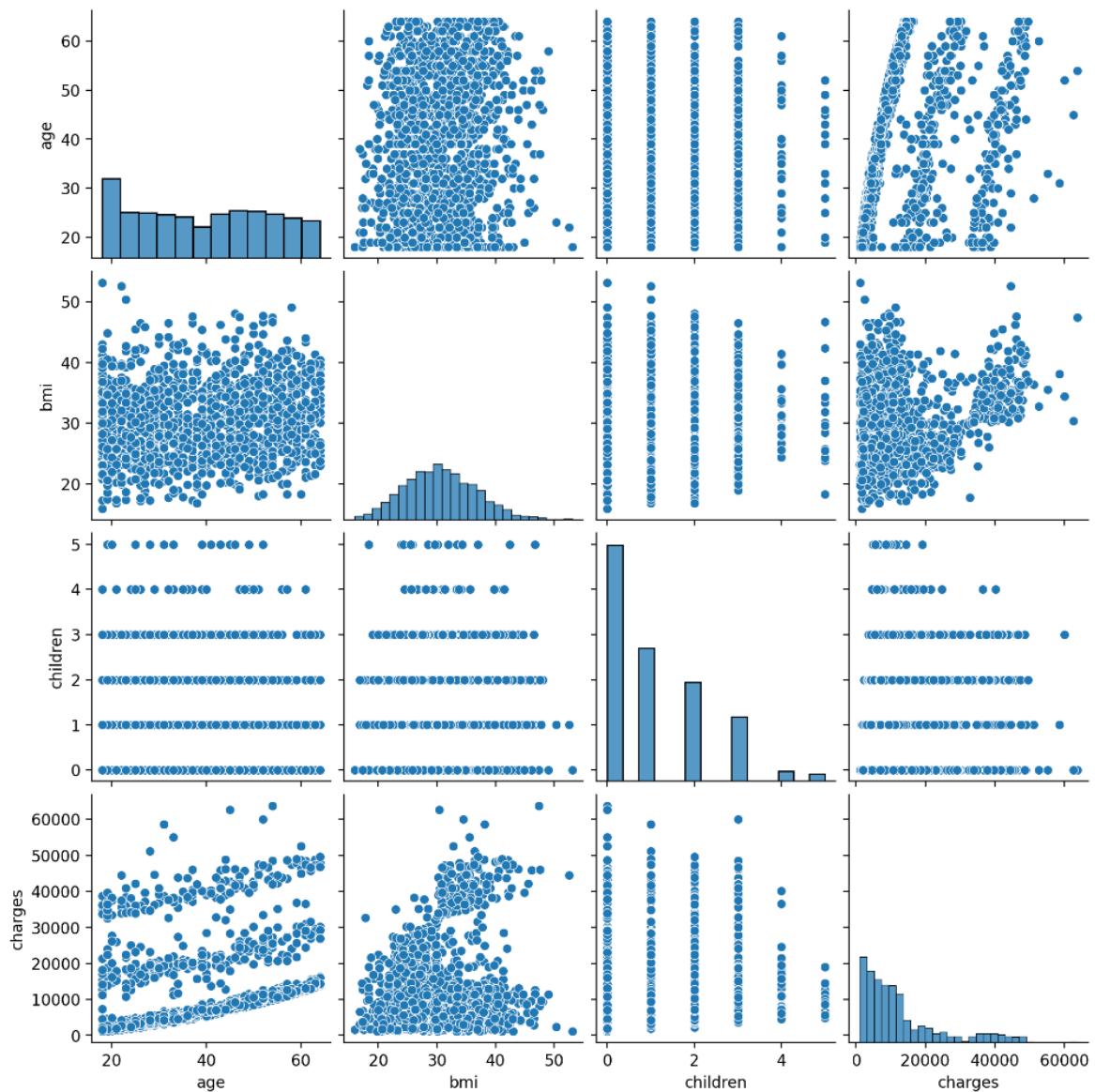
Nhận xét sơ bộ:

- Biến mục tiêu **charges** có độ lệch chuẩn rất lớn (12110.01) so với giá trị trung bình (13270.42), cho thấy sự chênh lệch đáng kể về chi phí y tế giữa các cá nhân. Điều này có thể chịu ảnh hưởng mạnh bởi các yếu tố như tình trạng hút thuốc, chỉ số BMI cao hoặc bệnh lý nền.
- Độ tuổi (**age**) của các cá nhân phân bố tương đối đều trong khoảng từ 18 đến 64 tuổi, với giá trị trung vị là 39, cho thấy tập dữ liệu bao phủ đầy đủ các nhóm tuổi trưởng thành.
- Chỉ số BMI có giá trị trung bình là 30.66, cao hơn ngưỡng BMI bình thường, cho thấy tỷ lệ thừa cân và béo phì trong tập dữ liệu là đáng kể, đây là yếu tố quan trọng ảnh hưởng đến chi phí y tế.
- Số con (**children**) có giá trị trung bình thấp (1.095) và trung vị bằng 1, phản ánh đa số cá nhân có ít hoặc không có con phụ thuộc.
- Các biến phân loại như **sex**, **smoker** và **region** có phân bố không đồng đều, đặc biệt là biến **smoker** với số lượng người không hút thuốc chiếm ưu thế. Điều này gợi ý khả năng biến **smoker** sẽ đóng vai trò quan trọng trong mô hình dự báo chi phí.

2.5.2 Quan hệ giữa biến đầu vào và biến mục tiêu *charges*

Biến số (Numeric variables)

Biểu đồ quan hệ cặp giữa các biến số và *charges*:



Hình 2.23: Biểu đồ pairplot thể hiện mối quan hệ giữa các biến số và *charges*

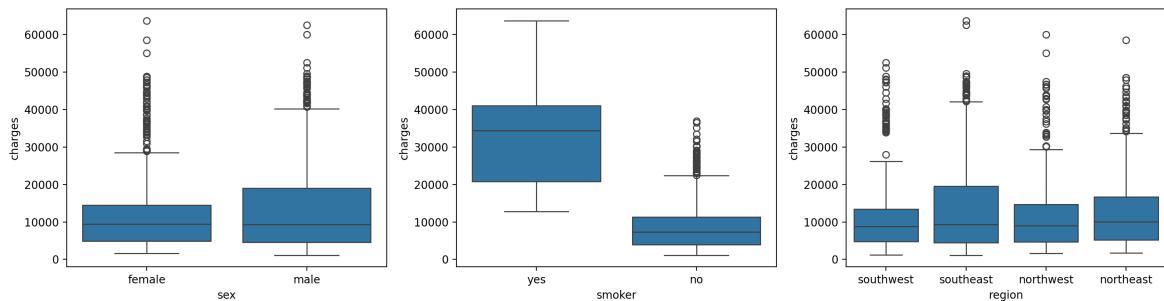
Từ biểu đồ *pairplot*, có thể rút ra một số nhận xét như sau:

- **age và charges:** Có xu hướng tăng rõ rệt; người lớn tuổi thường có chi phí bảo hiểm y tế cao hơn.
- **bmi và charges:** Có xu hướng tăng, tuy nhiên mối quan hệ không tuyến tính hoàn hảo, đặc biệt rõ rệt ở nhóm người hút thuốc.

- **children và charges:** Ảnh hưởng ở mức nhẹ, không mạnh bằng các biến như *smoker* hay *age*.

Biến phân loại (Categorical variables)

Boxplot giữa *charges* và các biến phân loại:



Hình 2.24: Boxplot thể hiện phân phối *charges* theo các biến phân loại

Nhận xét quan trọng:

- **smoker:** Sự khác biệt rất rõ ràng. Người hút thuốc có chi phí bảo hiểm cao hơn nhiều lần so với người không hút thuốc. Đây là đặc trưng cực kỳ quan trọng trong mô hình dự báo.
- **sex:** Sự khác biệt không quá rõ rệt giữa hai nhóm giới tính.
- **region:** Có sự khác biệt nhẹ giữa các khu vực, tuy nhiên mức độ ảnh hưởng không lớn bằng các biến như *smoker*, *age* hoặc *bmi*.

Các thuật toán học máy, đặc biệt là *Linear Regression*, không thể làm việc trực tiếp với dữ liệu dạng phân loại (*categorical data*) như:

- **sex = male/female**
- **smoker = yes/no**
- **region = northeast/northwest/southeast/southwest**

Do đó, các biến này cần được mã hóa sang dạng số trước khi đưa vào mô hình.

2.5.3 Mô hình workflow Linear Regression trong KNIME

Bước 1: CSV Reader – Đọc dữ liệu

The screenshot shows the KNIME interface with the following details:

- Node:** CSV Reader (4:1)
- State:** EXECUTED
- Port Output:** Port 0
- Load data** button
- Rows: 1338, Columns: 7**

The data preview table has columns: ID, age, sex, bmi, children, smoker, region, charges. It shows 5 rows of data:

ID	age	sex	bmi	children	smoker	region	charges
Row0	19	female	27.9	0	yes	southwest	16884.924
Row1	18	male	33.77	1	no	southeast	1725.5523
Row2	28	male	33.0	3	no	southeast	4449.462
Row3	33	male	22.705	0	no	northwest	21984.47061
Row4	32	male	28.88	0	no	northwest	3866.8552

Hình 2.25: Node CSV Reader trong workflow Linear Regression

- Đọc file `insurance.csv` vào KNIME.
- Tạo bảng dữ liệu gốc với các cột: `age`, `sex`, `bmi`, `children`, `smoker`, `region`, `charges`.
- Không thay đổi cấu trúc dữ liệu, chỉ chuyển dữ liệu từ file vào KNIME để các node tiếp theo sử dụng.

Bước 2: Column Filter – Lựa chọn đặc trưng

- Loại bỏ các cột không cần thiết (nếu có).
- Giữ lại các cột làm đặc trưng đầu vào và biến mục tiêu `charges`.
- Giúp workflow gọn gàng và tránh đưa nhầm các cột không phù hợp vào mô hình.

Bước 3: One to Many – Mã hóa biến phân loại

ID	age	bmi	children	charges	female	male	yes	no	southwest	southeast	northwest	northeast
Row0	19	27.9	0	16884.924	1	0	1	0	1	0	0	0
Row1	18	33.77	1	1725.5523	0	1	0	1	0	1	0	0
Row2	28	33.0	3	4449.462	0	1	0	1	0	1	0	0
Row3	33	22.705	0	21984.47061	0	1	0	1	0	0	1	0
Row4	32	28.88	0	3866.8552	0	1	0	1	0	0	1	0

Hình 2.26: Node One to Many mã hóa các biến phân loại

- Chuyển các biến phân loại (*sex, smoker, region*) thành các cột nhị phân bằng kỹ thuật *one-hot encoding*.
- Kết quả là toàn bộ dữ liệu đầu vào cho mô hình đều ở dạng số, phù hợp với Linear Regression.

Bước 4: Partitioning – Chia tập train/test

- Chia dữ liệu sau khi mã hóa thành hai tập:
 - Train: 70% (936/1338 dòng).
 - Test: 30% (402/1338 dòng).
- Đảm bảo quá trình huấn luyện và đánh giá mô hình được thực hiện độc lập, tránh rò rỉ dữ liệu (*data leakage*).

Bước 5: Normalizer (Train) – Chuẩn hóa trên tập huấn luyện

Node:	Normalizer (4:18)			
State:	EXECUTED			
Port Output	Port 0	Load data	Rows: 936, Columns: 12	
ID	age	bmi	children	charges
Row1	-1.4881597935159054	0.5434542956669199	-0.09959672887256499	1725.5523
Row2	-0.7750498939966872	0.4179579843557475	1.5216648054703978	4449.462
Row5	-0.5611169241409217	-0.7652929508638779	-0.9102274960440463	3756.6216
Row7	-0.13325098442939076	-0.4393285058997938	1.5216648054703978	7281.5056
Row8	-0.13325098442939076	-0.09869566091232596	0.7110340382989163	6406.4107

Hình 2.27: Chuẩn hóa dữ liệu trên tập huấn luyện

- Chuẩn hóa các biến số (*age*, *bmi*, *children*) theo Z-score.
- Không chuẩn hóa biến mục tiêu *charges*.
- Xuất ra đồng thời dữ liệu train đã chuẩn hóa và mô hình chuẩn hóa.

Bước 6: Normalizer (Apply) – Áp dụng chuẩn hóa cho tập test

Node:	Normalizer (Apply) (4:19)			
State:	EXECUTED			
Port Output	Port 0	Load data	Rows: 402, Columns: 12	
ID	age	bmi	children	charges
Row0	-1.4168488035639837	-0.4132513503026667	-0.9102274960440463	16884.924
Row3	-0.4184949442370782	-1.259943996096875	-0.9102274960440463	21984.47061
Row4	-0.489805934189	-0.2535287722702657	-0.9102274960440463	3866.8552
Row6	0.5085479251379055	0.48967016224784565	-0.09959672887256499	8240.5896
Row10	-0.9889828638524527	-0.6870614840724976	-0.9102274960440463	2721.3208

Hình 2.28: Áp dụng mô hình chuẩn hóa cho tập test

- Áp dụng mô hình chuẩn hóa học từ train cho dữ liệu test.
- Dảm bảo test được scale bằng đúng tham số của train, tránh leakage.

Bước 7: Linear Regression Learner – Huấn luyện mô hình

Node:	Linear Regression Learner (4:13)				
State:	EXECUTED				
Port Output Port 1 <input type="button" value="Load data"/> Rows: 12, Columns: 5					
ID	Variable	Coeff.	Std. Err.	t-value	P> t
Row1	age	3412.0	197.90988315794345	17.240169846783388	0.0
Row2	bmi	2052.0	208.24560927832587	9.853749172005095	0.0
Row3	children	556.25	195.09198459343025	2.851219137266041	0.004452191213655254
Row4	female	-2.43194379878006528E17	5.116842264973063E17	-0.47528215114773875	0.6346982076065648
Row5	male	-2.43194379878006784E17	5.1168422649730637E17	-0.4752821511477392	0.6346982076065648

Hình 2.29: Node Linear Regression Learner

- Huấn luyện mô hình Linear Regression với biến mục tiêu là *charges*.
- Ước lượng các hệ số hồi quy và xuất mô hình.

Bước 8: Regression Predictor – Dự đoán trên tập test

Node:	Regression Predictor (4:16)												
State:	EXECUTED												
Port Output Port 0 <input type="button" value="Load data"/> Rows: 402, Columns: 13													
ID	age	bmi	children	charges	female	male	yes	no	southwest	southeast	northwest	northeast	Prediction (charges)
Row0	-1.4169488035639837	-0.413251350326667	-0.9102274960440463	16884.924	1	0	1	0	1	0	0	0	25152.0
Row3	-0.4184949442370782	-1.259943996096875	-0.9102274960440463	21984.47061	0	1	0	1	0	0	1	0	3904.0
Row4	-0.489805934189	-0.2535287722702657	-0.9102274960440463	3866.8552	0	1	0	1	0	0	1	0	5696.0
Row6	0.5085479251379055	0.48967016224784565	-0.09959672887256499	8240.5896	1	0	0	1	0	1	0	0	10624.0
Row10	-0.9889828638524527	-0.687061480724976	-0.9102274960440463	2721.3208	0	1	0	1	0	0	0	0	3072.0
Row11	1.6495237643686549	-0.6756527284987541	-0.9102274960440463	27808.7251	1	0	1	0	0	1	0	0	35008.0
Row14	-0.8463608839486092	1.9059856756167912	-0.9102274960440463	39611.7577	0	1	1	0	0	1	0	0	31680.0
Row20	1.5069017844648114	0.9077195629142842	-0.9102274960440463	13228.84695	1	0	0	1	0	0	0	1	15104.0
Row24	-0.13325098442939076	-0.3928785724924113	0.7110340382989163	6203.90175	0	1	0	1	0	0	1	0	7488.0
Row25	1.4355907945128896	-0.442588150349434	1.5216648054703978	14001.1338	1	0	0	1	0	1	0	0	12928.0
Row26	1.7208347543205766	-1.1980107515536988	-0.9102274960440463	14451.83515	1	0	0	1	0	0	0	1	11520.0
Row27	1.1503468347052022	0.38128698429728747	0.7110340382989163	12268.63225	1	0	0	1	0	0	1	0	13760.0

Hình 2.30: Node Regression Predictor

- Sinh thêm cột Prediction(*charges*) chứa giá trị dự đoán cho từng bản ghi test.

Bước 9: Numeric Scorer – Đánh giá mô hình

Node: **Numeric Scorer (4:17)**

State: **EXECUTED**

Port Output Port 0 **Load data** Rows: 7, Col

ID	Prediction (charges)
R ²	0.7577498698934348
mean absolute error	4451.710771641792
mean squared error	4.047330396004663E7
root mean squared error	6361.863245940345
mean signed difference	-447.1563023880594
mean absolute percentage error	0.4335657919208659
adjusted R ²	0.7577498698934348

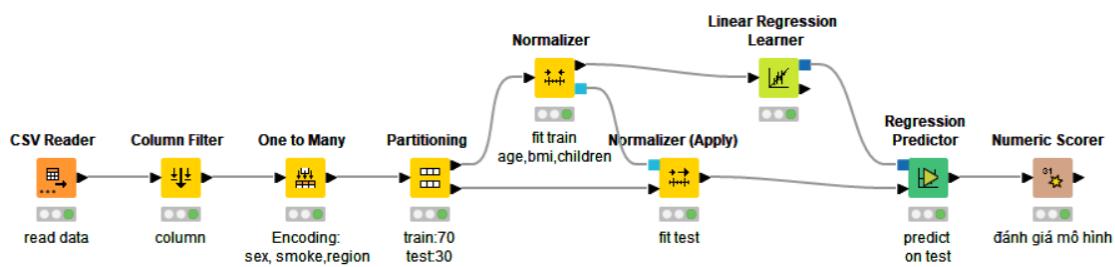
Hình 2.31: Kết quả đánh giá mô hình Linear Regression

Kết quả từ *Numeric Scorer* cho thấy mô hình Linear Regression đạt mức dự đoán khá tốt:

- $R^2 \approx 0.758$: mô hình giải thích được khoảng 75.8% biến thiên của *charges*.
- $MAE \approx 4451.71$: sai số tuyệt đối trung bình ở mức khoảng 4.45 nghìn.
- $RMSE \approx 6361.86$: cho thấy tồn tại một số trường hợp dự đoán sai lớn.

Dấu hiệu thiên lệch dự đoán:

- Mean signed difference = -447.16: mô hình có xu hướng dự đoán thấp hơn thực tế.
- $MAPE \approx 43.36\%$: khá cao do ảnh hưởng của các giá trị *charges* nhỏ trong mẫu.



Hình 2.32: So sánh giá trị dự đoán và giá trị thực của *charges*

CHƯƠNG 3

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN[?]

Kết luận (những kết quả chung đạt được, những kiến thức đã tiếp nhận được,...) và hướng phát triển.

3.1 Kết luận

3.2 Hướng phát triển