

# Sử dụng mô hình học máy để phân loại giai đoạn xơ gan

Đào Mạnh Đức - 22000085

Tạ Đăng Đức - 22000088

Nguyễn Huy Hoàng - 22000095

Khoa Toán - Cơ - Tin học

Trường Đại học Khoa học Tự nhiên

Đại học Quốc gia Hà Nội

*Giảng viên hướng dẫn:* TS. Cao Văn Chung

Hà Nội, tháng 4 năm 2025

# Nội dung trình bày

- 1 Giới thiệu
- 2 Dữ liệu và Tiền xử lý
- 3 Phân tích và trực quan hóa dữ liệu
- 4 Giảm chiều dữ liệu
- 5 Phương pháp
  - Phương pháp phân cụm
  - Phương pháp phân loại
  - Phương pháp hồi quy

# Đặt vấn đề (*đã trình bày*)

- **Xơ gan** là hậu quả của tổn thương gan kéo dài, dẫn đến mô sẹo và suy giảm chức năng gan.
- **Phân loại giai đoạn bệnh** giúp xác định phác đồ điều trị phù hợp, cải thiện tiên lượng cho bệnh nhân.
- **Vấn đề đặt ra:** Việc đánh giá dựa trên chỉ số lâm sàng có thể chưa tối ưu và còn phụ thuộc vào kinh nghiệm bác sĩ.
- **Hướng tiếp cận:** Ứng dụng thử nghiệm học máy để hỗ trợ phân loại giai đoạn bệnh dựa trên dữ liệu thực tế.

# Mục tiêu và phạm vi nghiên cứu (đã trình bày)

- Xây dựng và đánh giá các mô hình học máy **phân loại giai đoạn xơ gan** dựa trên đặc trưng lâm sàng và xét nghiệm.
- Thực hiện tiền xử lý dữ liệu: phát hiện ngoại lai, chuẩn hóa, biến đổi đặc trưng.
- So sánh hiệu quả các mô hình (KNN, MLP, SVM) qua các chỉ số đánh giá phù hợp.
- Đánh giá và lựa chọn mô hình có hiệu quả cao nhất trong bài toán cụ thể.

# Mô tả tập dữ liệu (*đã trình bày*)

- **Nguồn gốc:** Nghiên cứu lâm sàng về PBC (Mayo Clinic, 1974-1984).
- **Kích thước:** 25,000 bản ghi (sau mở rộng), 418 bệnh nhân gốc.
- **Đặc trưng:** 19 thuộc tính (định lượng & định tính).
- **Biến mục tiêu:** Stage - Giai đoạn xơ gan (1, 2, hoặc 3).

# Mô tả tập dữ liệu (đã trình bày)

STT	Tên trường	Kiểu dữ liệu	Mô tả
1	N_Days	Numeric	Số ngày từ khi đăng ký đến khi xảy ra một trong ba sự kiện: tử vong, ghép gan hoặc kết thúc nghiên cứu
2	Status	Nominal	Tình trạng bệnh nhân: C (còn sống), CL (đã được ghép gan), D (đã tử vong)
3	Drug	Nominal	Loại thuốc điều trị: D-penicillamine hoặc giả dược (placebo)
4	Age	Numeric	Tuổi bệnh nhân (tính theo ngày)
5	Sex	Nominal	Giới tính: M (nam), F (nữ)
6	Ascites	Nominal	Có báng bụng: N (Không), Y (Có)
7	Hepatomegaly	Nominal	Có gan to: N (Không), Y (Có)
8	Spiders	Nominal	Có giãn mao mạch hình nhện: N (Không), Y (Có)
9	Edema	Nominal	Phù: N (không phù), S (phù nhẹ), Y (phù không đáp ứng thuốc)

# Mô tả tập dữ liệu (đã trình bày)

STT	Tên trường	Kiểu dữ liệu	Mô tả
1	Bilirubin	Numeric	Nồng độ bilirubin huyết thanh (mg/dl)
2	Cholesterol	Numeric	Nồng độ cholesterol trong huyết thanh (mg/dl)
3	Albumin	Numeric	Nồng độ albumin (g/dl)
4	Copper	Numeric	Lượng đồng trong nước tiểu ( $\mu\text{g}$ /ngày)
5	Alk_Phos	Numeric	Nồng độ phosphatase kiềm (U/lít)
6	SGOT	Numeric	Nồng độ SGOT (U/ml)
7	Tryglicerides	Numeric	Nồng độ triglyceride trong huyết thanh (mg/dl)
8	Platelets	Numeric	Số lượng tiểu cầu (nghìn/ $\text{ml}^3$ )
9	Prothrombin	Numeric	Thời gian prothrombin (giây)
10	Stage	Ordinal	Giai đoạn mô học của bệnh (1, 2 hoặc 3)

# Đọc dữ liệu (đã trình bày)

	N_Days	Status	Drug	Age	Sex	Ascites	Hepatomegaly	Spiders	Edema	Bilirubin	Cholesterol	Albumin	Copper	Alk_Phos	SGOT	Tryglicerides	Platelets	Prothrombin	Stage	
0	2221	C	Placebo	18499	F	N		Y	N	N	0.5	149.000000	4.04	227.0	598.0	52.70	57.000000	256.0	9.9	1
1	1230	C	Placebo	19724	M	Y		N	Y	N	0.5	219.000000	3.93	22.0	663.0	45.00	75.000000	220.0	10.8	2
2	4184	C	Placebo	11839	F	N		N	N	N	0.5	320.000000	3.54	51.0	1243.0	122.45	80.000000	225.0	10.0	2
3	2090	D	Placebo	16467	F	N		N	N	N	0.7	255.000000	3.74	23.0	1024.0	77.50	58.000000	151.0	10.2	2
4	2105	D	Placebo	21699	F	N		Y	N	N	1.9	486.000000	3.54	74.0	1052.0	108.50	109.000000	151.0	11.5	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
24995	3584	D	D-penicillamine	23612	F	N		N	N	N	0.8	231.000000	3.87	173.0	9009.8	127.71	96.000000	295.0	11.0	2
24996	3584	D	D-penicillamine	23612	F	N		N	N	N	0.8	231.000000	3.87	173.0	9009.8	127.71	96.000000	295.0	11.0	2
24997	971	D	D-penicillamine	16736	F	N		Y	Y	Y	5.1	369.510563	3.23	18.0	790.0	179.80	124.702128	104.0	13.0	3
24998	3707	C	D-penicillamine	16990	F	N		Y	N	N	0.8	315.000000	4.24	13.0	1637.0	170.50	70.000000	426.0	10.9	2
24999	3707	C	D-penicillamine	16990	F	N		Y	N	N	0.8	315.000000	4.24	13.0	1637.0	170.50	70.000000	426.0	10.9	2

25000 rows × 19 columns

## Hình: Mẫu dữ liệu



# Data Cleaning (đã trình bày)

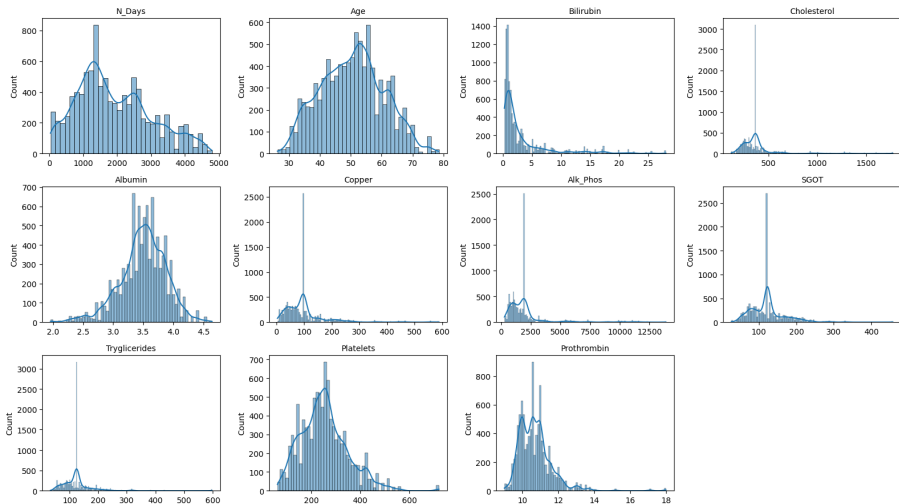
- **Kiểm tra giá trị thiếu:** Mẫu dữ liệu không có giá trị thiếu
- **Kiểm tra trùng lặp dữ liệu:** Bộ dữ liệu gồm có 25,000 mẫu dữ liệu và sau khi kiểm tra thì có 15,361 mẫu bị trùng lặp. Sau khi thực hiện loại bỏ trùng lặp thì bộ dữ liệu còn 9,639 mẫu.
- **Chuyển đổi dữ liệu:**
  - Biến nhị phân như *Ascites*, *Hepatomegaly*, *Spiders* được ánh xạ: "N"  $\rightarrow$  0, "Y"  $\rightarrow$  1.
  - Biến Sex: "F"  $\rightarrow$  0, "M"  $\rightarrow$  1.
  - Biến Drug: "Placebo"  $\rightarrow$  0, "D-penicillamine"  $\rightarrow$  1.
  - Các biến nhiều mức như *Edema*, *Status* được xử lý bằng One-hot Encoding.

# Phân tích các tham số thống kê (đã trình bày)

	N_Days	Age	Bilirubin	Cholesterol	Albumin	Copper	Alk_Phos	SGOT	Tryglicerides	Platelets	Prothrombin	Stage
count	9639.000000	9639.000000	9639.000000	9639.000000	9639.000000	9639.000000	9639.000000	9639.000000	9639.000000	9639.000000	9639.000000	9639.000000
mean	1910.982571	18429.717606	3.228571	371.706706	3.496118	97.027569	1973.572709	122.317487	123.587337	253.787605	10.713328	2.029152
std	1093.620373	3693.953156	4.512278	197.824339	0.382319	73.108854	1827.063380	47.653515	55.206301	95.740700	0.922026	0.809956
min	41.000000	9598.000000	0.300000	120.000000	1.960000	4.000000	289.000000	26.350000	33.000000	62.000000	9.000000	1.000000
25%	1103.000000	15628.000000	0.800000	271.000000	3.290000	51.000000	1031.000000	89.900000	93.000000	188.000000	10.000000	1.000000
50%	1690.000000	18628.000000	1.300000	369.510563	3.520000	97.648387	1713.000000	122.556346	124.702128	249.000000	10.600000	2.000000
75%	2598.000000	20819.000000	3.300000	369.510563	3.760000	102.000000	1982.655769	134.850000	125.000000	307.000000	11.100000	3.000000
max	4795.000000	28650.000000	28.000000	1775.000000	4.640000	588.000000	13862.400000	457.250000	598.000000	721.000000	18.000000	3.000000

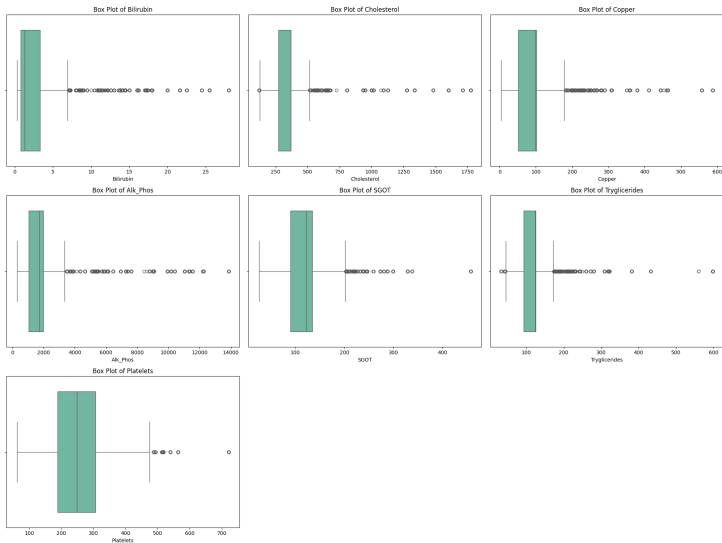
Hình: Các tham số thống kê

# Kiểm tra phân phối của dữ liệu (đã trình bày)



**Hình:** Biểu đồ histogram và đường KDE mô tả phân bố của các thuộc tính liên tục

# Phát hiện outliers (đã trình bày)



Hình: Biểu đồ Box plot

# Phương pháp IQR (đã trình bày)

## Các bước thực hiện:

- 1 Tính các giá trị tứ phân vị và IQR:
  - $Q1$ : Phân vị thứ nhất (25% dữ liệu đầu).
  - $Q3$ : Phân vị thứ ba (75% dữ liệu đầu).
  - $IQR = Q3 - Q1$ : Khoảng tứ phân vị.

- 2 Xác định khoảng giá trị hợp lý:

$$\text{Lower bound} = Q1 - 1.5 \times IQR, \quad \text{Upper bound} = Q3 + 1.5 \times IQR$$

- 3 Phát hiện ngoại lệ: Các giá trị nằm ngoài khoảng  $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$  được coi là ngoại lệ.

# Xử lý ngoại lệ (*đã trình bày*)

## Phương pháp xử lý:

- **Cắt (Clipping):** Đưa giá trị vượt ngưỡng về ngưỡng giới hạn.
- **Thay thế bằng trung vị:** Thay các giá trị ngoại lệ bằng giá trị trung vị để giữ xu hướng chính của dữ liệu.
- **Loại bỏ:** Loại bỏ các mẫu chứa ngoại lệ, áp dụng khi dữ liệu đủ lớn.

# Chuẩn hóa Dữ liệu (đã trình bày)

- Mục tiêu: Đưa các thuộc tính liên tục về cùng một thang đo để:
  - Loại bỏ ảnh hưởng của đơn vị và độ lớn khác nhau.
  - Tăng hiệu quả huấn luyện mô hình học máy.
- Sử dụng phương pháp **StandardScaler**:

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

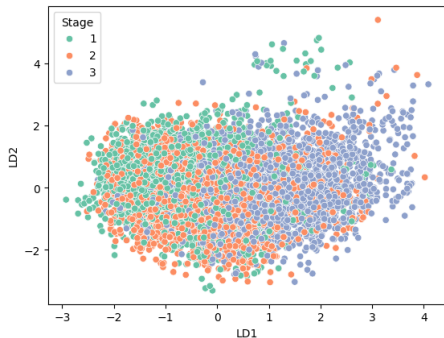
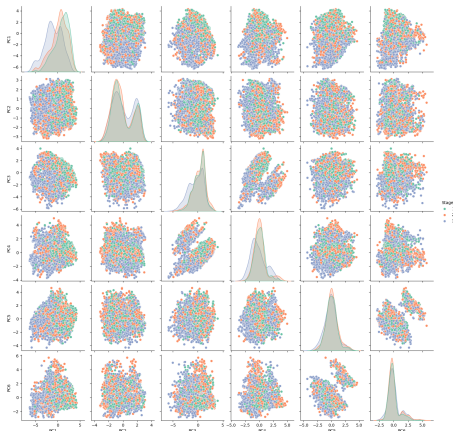
Trong đó:

- $x$ : giá trị ban đầu
- $\mu$ : trung bình
- $\sigma$ : độ lệch chuẩn

# Giảm chiều và trực quan hóa dữ liệu

- Giúp trực quan hóa và đánh giá khả năng phân tách dữ liệu theo giai đoạn xơ gan.

PCA scatter plots theo cặp 2 thành phần





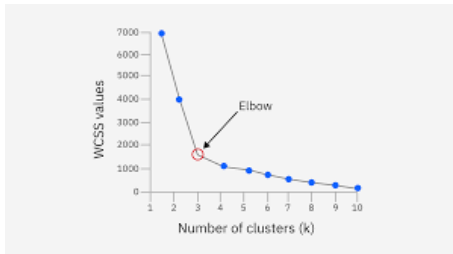
# Thuật toán K-means Clustering (đã trình bày)

- **K-means** là thuật toán học không giám sát dùng để phân cụm dữ liệu thành  $K$  nhóm dựa trên mức độ tương đồng.
- Dữ liệu đầu vào không có nhãn (label). Mỗi điểm dữ liệu chỉ thuộc một cụm duy nhất.
- Mục tiêu: Tìm  $K$  trung tâm cụm (centroids) và gán mỗi điểm vào cụm sao cho **tổng khoảng cách bình phương đến trung tâm cụm là nhỏ nhất**.

$$\min \sum_{j=1}^K \sum_{x_i \in S_j} \|x_i - c_j\|^2$$

# Phương pháp Elbow để chọn $K$ tối ưu (đã trình bày)

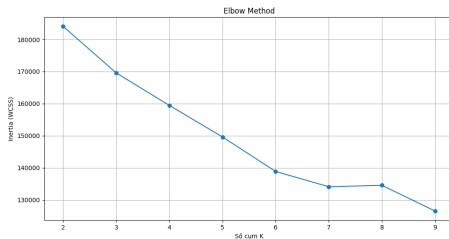
- Chạy K-means với các giá trị  $K$  khác nhau.
- Tính tổng bình phương sai số trong cụm (WCSS) cho mỗi  $K$ .
- Vẽ biểu đồ WCSS theo  $K$ .
- **$K$  tối ưu** là tại “góc khuỷu tay” – nơi WCSS bắt đầu giảm chậm lại.



# Ứng dụng vào dữ liệu

## Tìm $K$ :

Thu được kết quả sau từ thực nghiệm:

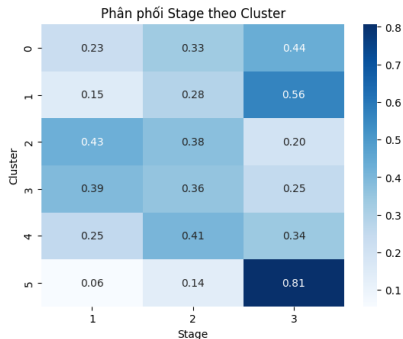


Hình: Biểu đồ kết quả phương pháp Elbow

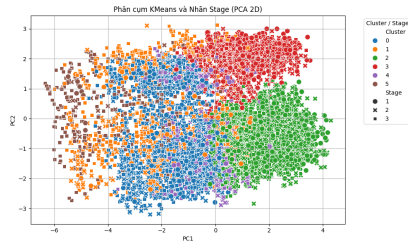
**Nhận xét:** Giá trị  $K = 6$  được lựa chọn làm số cụm tối ưu để áp dụng thuật toán KMeans Clustering trong các bước tiếp theo.

# Ứng dụng vào dữ liệu

## Kết quả thu được:



(a) Bảng phân phối Stage - Cluster



(b) Scatter plot phân phối

**Silhouette Score:** 0.1299  $\rightarrow$  Dữ liệu không đủ mạnh để phân cụm

# Các phương pháp phân loại (*đã trình bày*)

## KNN (K-Nearest Neighbors)

- Phi tham số, lazy learning. Phân loại dựa trên *biểu quyết đa số* của  $K$  láng giềng gần nhất.
- Đơn giản, dễ hiểu, không giả định về dữ liệu.
- Cần chuẩn hóa dữ liệu, chọn  $K$  phù hợp, chi phí dự đoán cao.

# Kết quả thực nghiệm KNN

Dữ liệu	Test size	Acc Train	Acc Test	Precision	Recall	F1-score
Dữ liệu gốc	0.2	0.9912	0.7894	0.79	0.79	0.79
	0.3	0.9918	0.7994	0.80	0.80	0.80
	0.4	0.9939	0.7884	0.79	0.79	0.79
Dữ liệu giảm chiều	0.2	0.9912	0.7173	0.72	0.72	0.72
	0.3	0.9918	0.7230	0.72	0.72	0.72
	0.4	0.9939	0.7041	0.70	0.70	0.70

# Các phương pháp phân loại (*đã trình bày*)

## MLP (Multi-Layer Perceptron)

- Mạng nơ-ron truyền thẳng, học biểu diễn *phi tuyến* qua các lớp ẩn và thuật toán lan truyền ngược (backpropagation).
- Linh hoạt, mạnh mẽ, nền tảng của Deep Learning.
- Nhiều siêu tham số cần tinh chỉnh, dễ overfitting, cần dữ liệu lớn.

# Kết quả thực nghiệm MLP

Dữ liệu	Test size	Acc Train	Acc Test	Precision	Recall	F1-score
Dữ liệu gốc	0.2	0.8688	0.7007	0.70	0.70	0.70
	0.3	0.8711	0.7168	0.72	0.72	0.72
	0.4	0.9092	0.6948	0.69	0.69	0.69
Dữ liệu giảm chiều	0.2	0.8248	0.6955	0.69	0.69	0.69
	0.3	0.8359	0.6974	0.70	0.70	0.70
	0.4	0.8439	0.6914	0.69	0.69	0.69



## Multi-class SVM (Support Vector Machine)

- Tìm *siêu phẳng tối ưu* phân tách các lớp sao cho khoảng cách từ điểm gần nhất của mỗi lớp đến siêu phẳng (*margin*) là lớn nhất có thể, dựa trên các véc-tơ hỗ trợ (support vectors).

# Kết quả thực nghiệm SVM (4:1)

Model	Train Acc	Test Acc	Precision	Recall	F1-score
<b>Dữ liệu gốc</b>					
Hard/Soft Margin	0.5653	0.5555	0.5500	0.5572	0.5489
Multi-class	0.5660	0.5493	0.5463	0.5491	0.5401
Polynomial Kernel	0.8611	0.7158	0.7178	0.7147	0.7158
RBF Kernel	0.9803	0.7426	0.7409	0.7415	0.7412
Sigmoid Kernel	0.5636	0.5539	0.5487	0.5558	0.5472
<b>Dữ liệu giảm chiều (PCA)</b>					
Hard/Soft Margin	0.5575	0.5420	0.5389	0.5436	0.5375
Multi-class	0.5506	0.5405	0.5330	0.5413	0.5311
Polynomial Kernel	0.8690	0.6883	0.6911	0.6868	0.6883
RBF Kernel	0.8937	0.7272	0.7261	0.7278	0.7269
Sigmoid Kernel	0.5570	0.5420	0.5389	0.5436	0.5376

# Kết quả thực nghiệm SVM (7:3)

Model	Train Acc	Test Acc	Precision	Recall	F1-score
<b>Dữ liệu gốc</b>					
Hard/Soft Margin	0.5610	0.5533	0.5498	0.5551	0.5463
Multi-class	0.5598	0.5558	0.5477	0.5557	0.5450
Polynomial Kernel	0.8724	0.7133	0.7123	0.7116	0.7120
RBF Kernel	0.9803	0.7317	0.7310	0.7307	0.7308
Sigmoid Kernel	0.5611	0.5536	0.5502	0.5554	0.5467
<b>Dữ liệu giảm chiều (PCA)</b>					
Hard/Soft Margin	0.5539	0.5415	0.5366	0.5431	0.5348
Multi-class	0.5514	0.5450	0.5373	0.5455	0.5350
Polynomial Kernel	0.7992	0.6822	0.6841	0.6806	0.6819
RBF Kernel	0.8963	0.7244	0.7251	0.7236	0.7242
Sigmoid Kernel	0.5514	0.5432	0.5392	0.5447	0.5373

# Kết quả thực nghiệm SVM (3:2)

Model	Train Acc	Test Acc	Precision	Recall	F1-score
<b>Dữ liệu gốc</b>					
Hard/Soft Margin	0.5629	0.5511	0.5483	0.5532	0.5436
Multi-class	0.5570	0.5609	0.5544	0.5618	0.5451
Polynomial Kernel	0.9321	0.6999	0.6974	0.6970	0.6973
RBF Kernel	0.9435	0.7326	0.7321	0.7317	0.7319
Sigmoid Kernel	0.5629	0.5511	0.5484	0.5533	0.5435
<b>Dữ liệu giảm chiều (PCA)</b>					
Hard/Soft Margin	0.5513	0.5423	0.5369	0.5437	0.5353
Multi-class	0.5535	0.5458	0.5401	0.5487	0.5358
Polynomial Kernel	0.8793	0.6712	0.6755	0.6702	0.6717
RBF Kernel	0.9030	0.7098	0.7107	0.7091	0.7099
Sigmoid Kernel	0.5514	0.5384	0.5344	0.5402	0.5325

# Các phương pháp hồi quy

## Linear Regression

Là mô hình học có giám sát, đơn giản và tuyến tính. Mô hình tuyến tính dự đoán  $y$  từ các biến đầu vào bằng tổng hợp tuyến tính:

$\hat{y} = \mathbf{w}^T \mathbf{x}$ . Tham số  $\mathbf{w}$  được ước lượng bằng cách tối thiểu hóa tổng bình phương sai số giữa giá trị dự đoán và thực tế.

## Random Forest

Là mô hình học có giám sát thuộc nhóm ensemble (rừng cây quyết định). Mô hình tổng hợp nhiều cây quyết định huấn luyện trên các tập dữ liệu con và đặc trưng ngẫu nhiên. Dự đoán là trung bình các cây. Giảm phương sai, kháng nhiễu tốt và đánh giá được tầm quan trọng của đặc trưng.

# Kết quả thực nghiệm

Model	Test Size	Train MSE	Test MSE	Train MAE	Test MAE	Train $R^2$	Test $R^2$
Linear	0.2	0.0993	0.0997	0.2542	0.2543	0.319	0.369
Linear (PCA)	0.2	0.1111	0.1100	0.2733	0.2715	0.308	0.303
RF	0.2	0.0036	0.0351	0.0422	0.1259	0.978	0.778
RF (PCA)	0.2	0.0105	0.0658	0.0764	0.1909	0.934	0.583
Linear	0.3	0.1007	0.0985	0.2562	0.2527	0.378	0.380
Linear (PCA)	0.3	0.1126	0.1085	0.2767	0.2702	0.304	0.317
RF	0.3	0.0041	0.0360	0.0452	0.1289	0.975	0.773
RF (PCA)	0.3	0.0115	0.0658	0.0806	0.1918	0.929	0.586
Linear	0.4	0.1082	0.1089	0.2652	0.2673	0.361	0.356
Linear (PCA)	0.4	0.1201	0.1208	0.2849	0.2868	0.290	0.287
RF	0.4	0.0047	0.0460	0.0486	0.1467	0.972	0.729
RF (PCA)	0.4	0.0126	0.0774	0.0846	0.2100	0.925	0.543


Cảm ơn thầy và các bạn đã lắng  
nghe!


# Tài liệu tham khảo và Nguồn dữ liệu I

 Mayo Clinic Primary Biliary Cirrhosis Data:

Có sẵn trên Kaggle: <https://byvn.net/4noN>

 Slide bài giảng - TS. Cao Văn Chung

 Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *JMLR*, 12, 2825-2830.  
<https://scikit-learn.org/>

 Machine Learning Cơ Bản.  
<https://machinelearningcoban.com/>

 Deep AI KhanhBlog.  
<https://phamdinhhkhanh.github.io/deepai-book/intro.html>