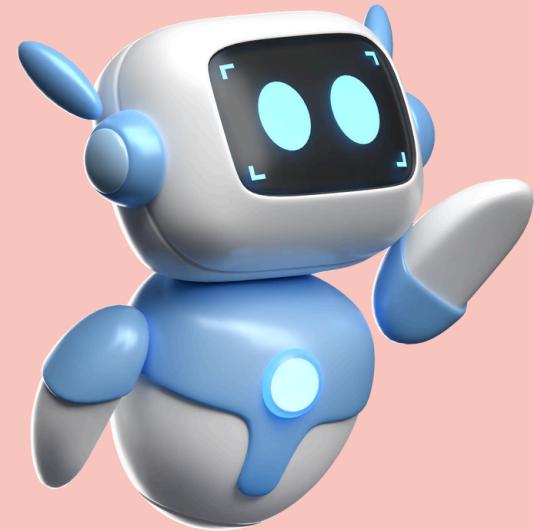


CHUẨN ĐOÁN VÀ PHÂN CẤP BỆNH TIM HỖ TRỢ QUYẾT ĐỊNH LÂM SÀNG

Môn: Khai Phá Dữ Liệu

Giảng viên hướng dẫn: TS.Nguyễn Ngọc Thủy

BỐ CỤC:



1.Giới Thiệu

2.Cách tiếp cận và phương pháp thực hiện

3.Thử nghiệm và đánh giá



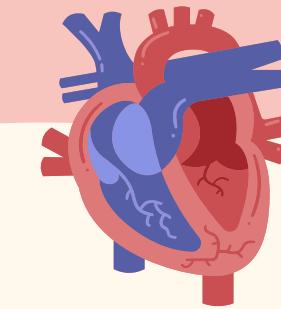
1.GIỚI THIỆU:

+) Bệnh tim là một trong những nguyên nhân gây tử vong hàng đầu trên thế giới. Việc dự đoán sớm bệnh tim giúp tăng khả năng điều trị và cải thiện sức khỏe cộng đồng

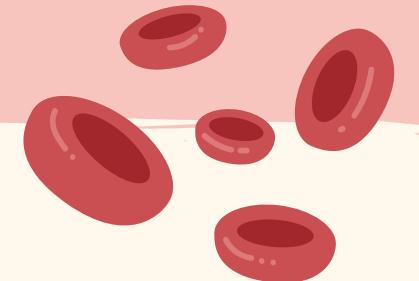
+) Ứng dụng trí tuệ nhân tạo (AI) và học máy(ML) mang lại cơ hội lớn cho việc chẩn đoán y tế. Đề tài này nghiên cứu việc áp dụng các mô hình học máy để dự đoán nguy cơ liên quan đến bệnh tim bao gồm (dự đoán bệnh tim, và dự đoán mức độ bệnh tim) dựa trên dữ liệu y tế.



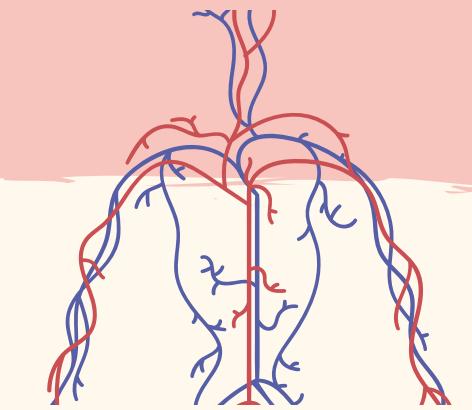
2.Cách tiếp cận và phương pháp thực hiện



Thu thập dữ liệu về (nhịp tim , huyết áp ,cholesterolon..) của mọi người bao gồm cả người bị bệnh tim và không bị bệnh tim



Xử lý dữ liệu mà ta thu thập được để thuật toán ta lựa chọn hoạt động ổn định



Lựa chọn thuật toán phù hợp với bài toán ta đặt ra

2.1.Thu thập dữ liệu về (nhịp tim, huyết áp, cholesterol..) của mọi người bao gồm cả người bị bệnh tim và không bị bệnh tim.

Phân tích tập dữ liệu

- Tập dữ liệu trên gồm 500 bệnh nhân trong đó có 300 người mắc bệnh tim 200 người không mắc bệnh tim.
- Tập dữ liệu trên bao gồm có 6 thuộc tính và một nhãn.
- Ta có thể nhận thấy tập dữ liệu của ta không quá lớn vậy tập dữ liệu này có thể hoạt động tốt với một số mô hình như (SVM, logistic regression..)

Age	Gender	BloodPres	Cholesterol	HeartRate	QuantumP	HeartDisease
68	1	105	191	107	8.362241	1
58	0	97	249	89	9.249002	0
44	0	93	190	82	7.942542	1
72	1	93	183	101	6.495155	1
37	0	145	166	103	7.6539	1
50	1	114	271	73	8.631604	0
68	1	156	225	73	7.559545	1
48	0	156	236	61	9.152103	0
52	0	116	266	114	9.146932	0
40	1	121	255	96	9.6838	0
40	1	139	235	64	8.732313	0
53	1	150	176	97	8.984835	0
65	0	140	206	104	7.545538	1
69	1	108	180	100	6.799794	1
53	1	110	283	113	9.190135	0
32	1	94	247	119	9.322586	0
51	0	171	161	106	8.316819	1

2.2. Lựa chọn thuật toán phù hợp với bài toán ta đặt ra

Các thuật toán đã học

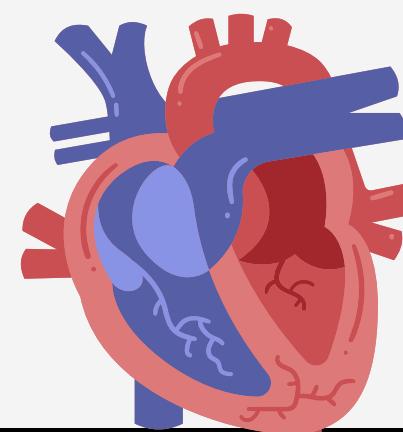
- linear regression
- logistic regression
- SVM
- K-Nearest Neighbors (KNN)
- Decision Tree
- Random Forest
- Naive Bayes
- ...

Thuật toán lựa chọn

- Ở đây thuật toán mà chúng tôi chọn để dự đoán bệnh tim là **thuật toán SVM**. Lý do:
 - + Thuật toán này phân loại tốt cho dữ liệu nhị phân.
 - + Hiệu quả cao với dữ liệu không tuyến tính(nhưng vấn đề liên quan đến sức khoẻ thường có dữ liệu không tuyến tính).
 - + Ít bị overfitting.
 - + Hoạt động tốt với tập dữ liệu không quá lớn.

Còn với dự đoán mức độ bệnh tim thì chúng tôi kỹ thuật **stacking 2 thuật toán(catboost và LightGBM)**. Lý do:

- + Đảm bảo dữ liệu sạch, chính xác để phân tích tốt hơn.
- + Giúp học nhanh và tìm ra mối liên hệ mới trong dữ liệu.
- + Tìm ra các thông tin chủ chốt để tập trung cải thiện.

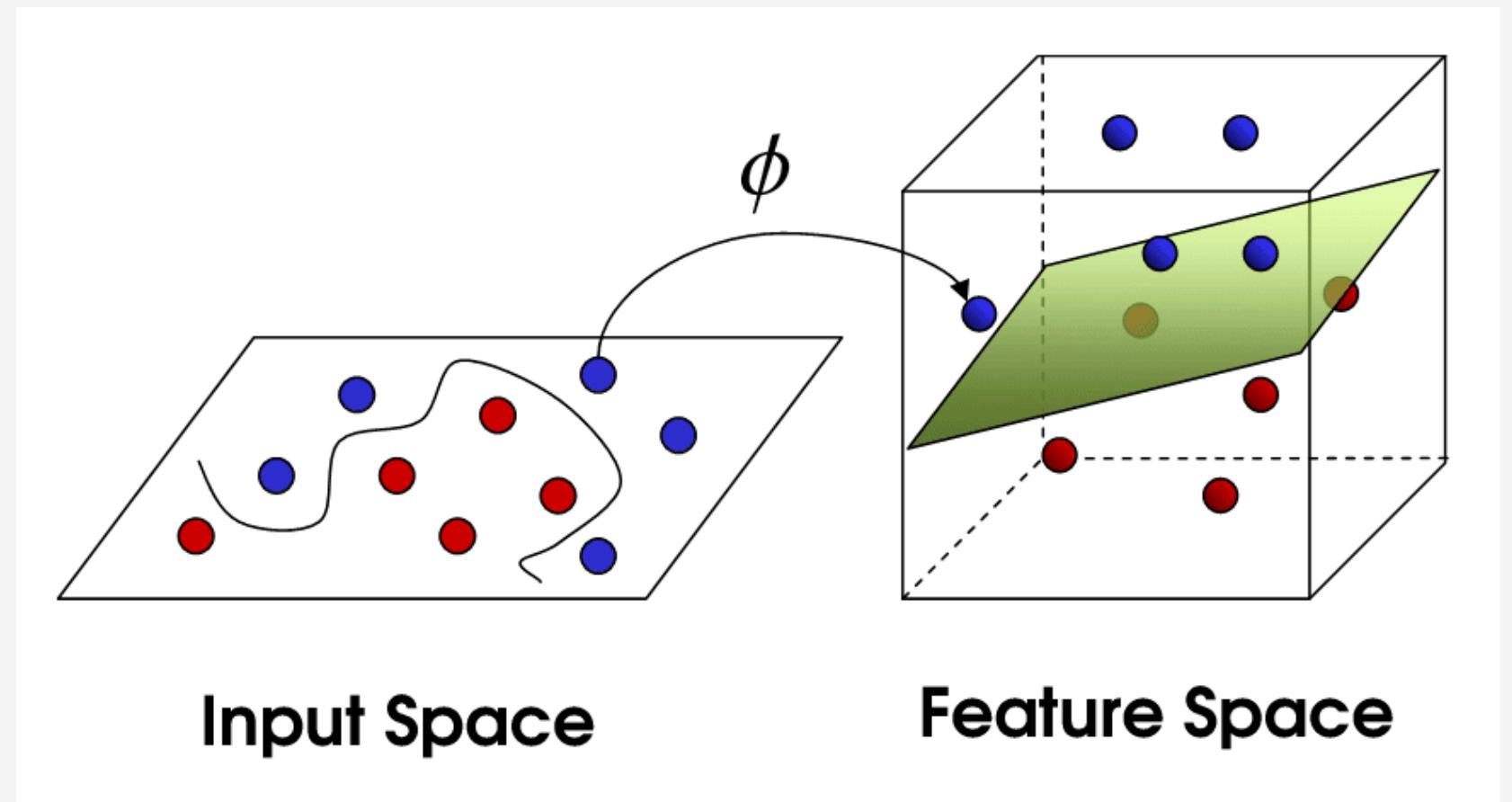
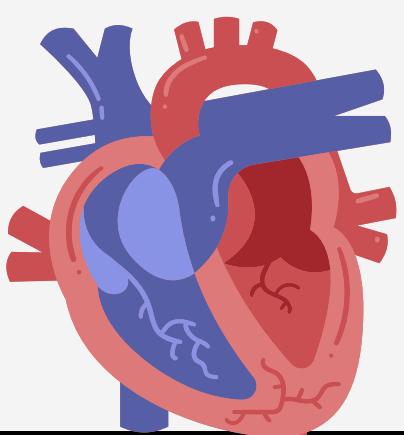


Chuẩn đoán bệnh tim

2.2.1. Giới thiệu thuật toán SVM:

SVM (Support Vector Machine) là một thuật toán học máy có giám sát, dùng để phân loại và đôi khi hồi quy. Mục tiêu của SVM là tìm một siêu phẳng tối ưu để phân tách các lớp dữ liệu, sao cho khoảng cách giữa siêu phẳng và các điểm gần nhất của mỗi lớp là lớn nhất.

Khi dữ liệu không phân tách tuyến tính được, SVM sử dụng các hàm kernel để ánh xạ dữ liệu sang không gian cao hơn, giúp phân tách dễ dàng hơn.



LIKE



2.2.2.Cách thức hoạt động của thuật toán SVM:

1. Nhận dữ liệu	Có các điểm với nhãn 0 hoặc 1
2. Tìm đường phân chia	Là đường/siêu phẳng tốt nhất, xa cả hai lớp
3. Dùng support vectors	Để xác định đường phân chia
4. Phân loại dữ liệu mới	Dựa vào phía của đường phân chia mà điểm đó nằm

Sau khi SVM huấn luyện:

Siêu phẳng phân chia hai lớp (giữa 0 và 1).

Support vectors các điểm quyết định siêu phẳng.

Các hệ số (w) và độ lệch (b) trong phương trình của siêu phẳng:

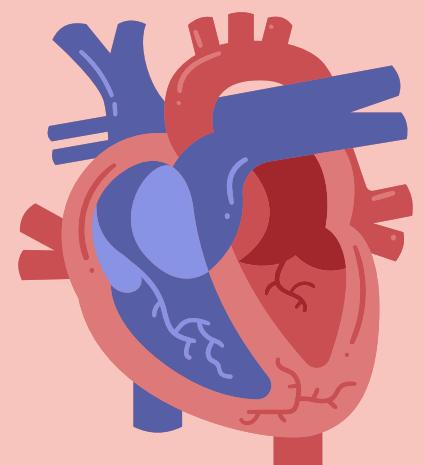
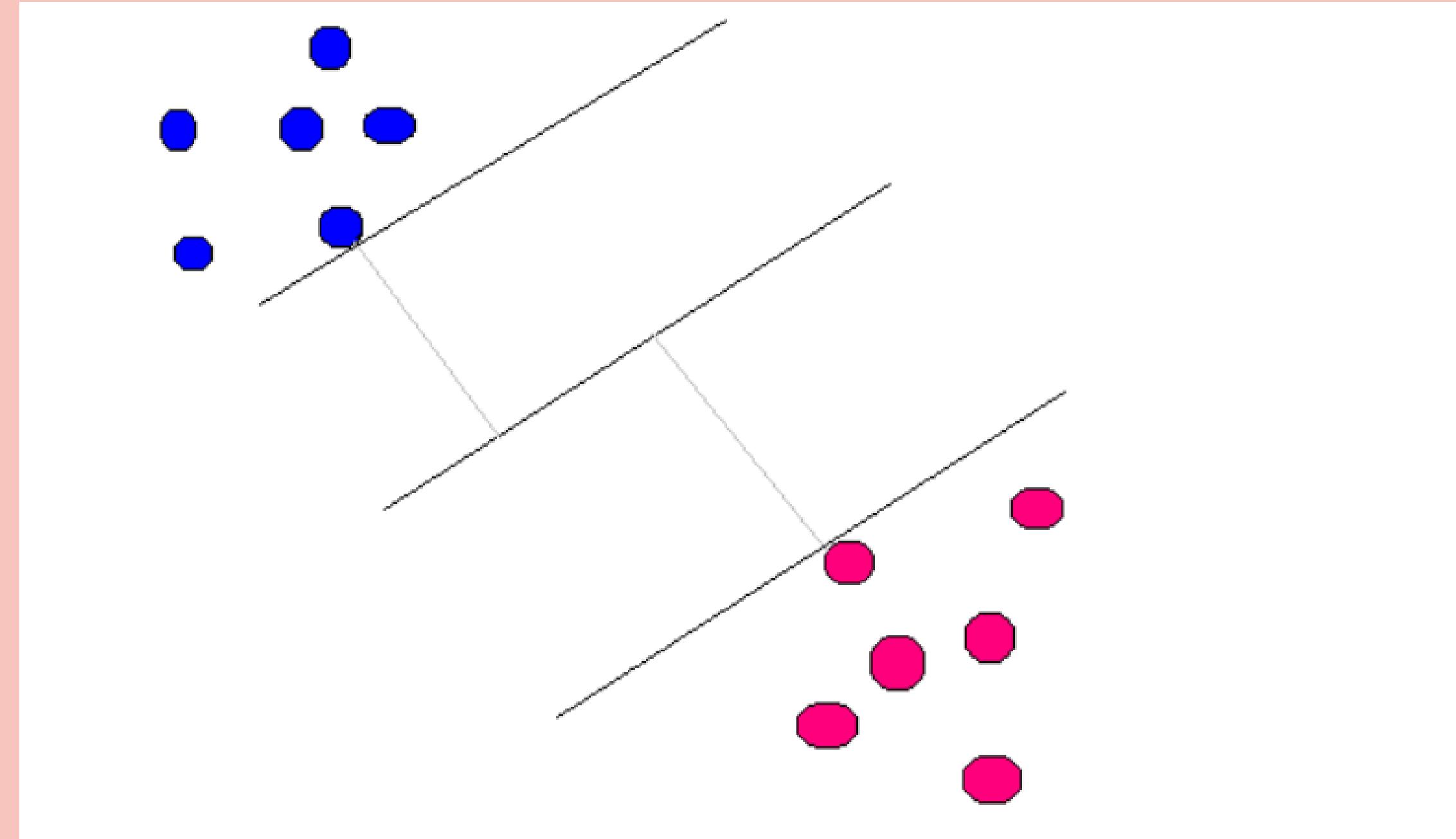
$$f(x) = w \cdot x + b$$

- Với một điểm dữ liệu mới x , SVM tính giá trị $f(x)$:
 - Nếu $f(x) > 0 \rightarrow$ dự đoán là 1
 - Nếu $f(x) < 0 \rightarrow$ dự đoán là 0

LIKE



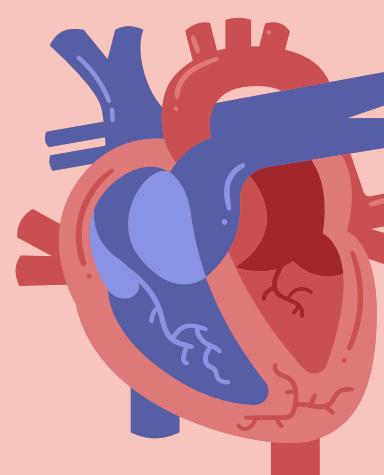
Mô tả cách hoạt động của SVM



2.2.3.Thử nghiệm và đánh giá Thuật toán SVM

```
Best Parameters: {'max_depth': None, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 200}
Accuracy: 0.92
Precision: 0.9642857142857143
Recall: 0.9
F1 Score: 0.9310344827586207
Confusion Matrix:
[[38  2]
 [ 6 54]]
ROC-AUC: 0.9658333333333333
Matthews Correlation Coefficient (MCC): 0.8388884110737054
Log Loss: 0.22494453569005737
Cohen's Kappa: 0.8360655737704918

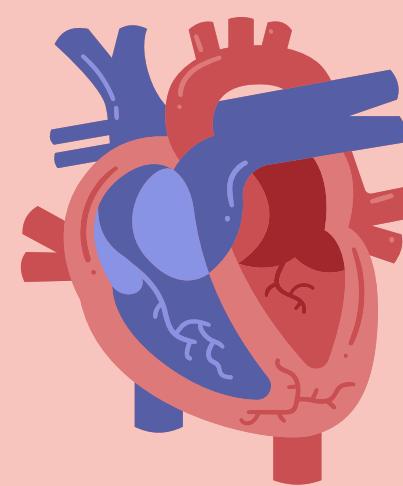
Feature Importance:
      Feature  Importance
5  QuantumPatternFeature  0.693444
3          Cholesterol  0.133224
0              Age  0.105020
4          HeartRate  0.032173
2        BloodPressure  0.030858
1            Gender  0.005280
```



Đánh giá thuật toán:



Accuracy	Dự đoán đúng trên tổng số	92%
Precision	Đoán đúng bệnh / tổng số đoán là bệnh	96.4%
Recall	Đoán đúng bệnh / tổng số người thực sự bị bệnh	90%
F1 Score	Trung bình hài hòa Precision và Recall	93.1%
ROC-AUC	Khả năng phân biệt hai lớp qua các ngưỡng dự đoán	96.58%



LIKE



2.3.Xử lý dữ liệu mà ta thu thập được để thuật toán ta lựa chọn hoạt động ổn định

Ban đầu chúng tôi sẽ dự đoán bệnh tim và thuật toán chúng tôi dùng để dự đoán bệnh tim là SVM.

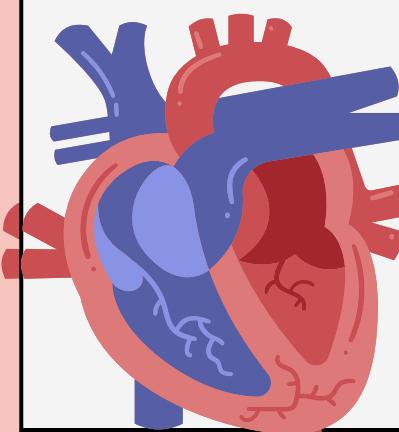
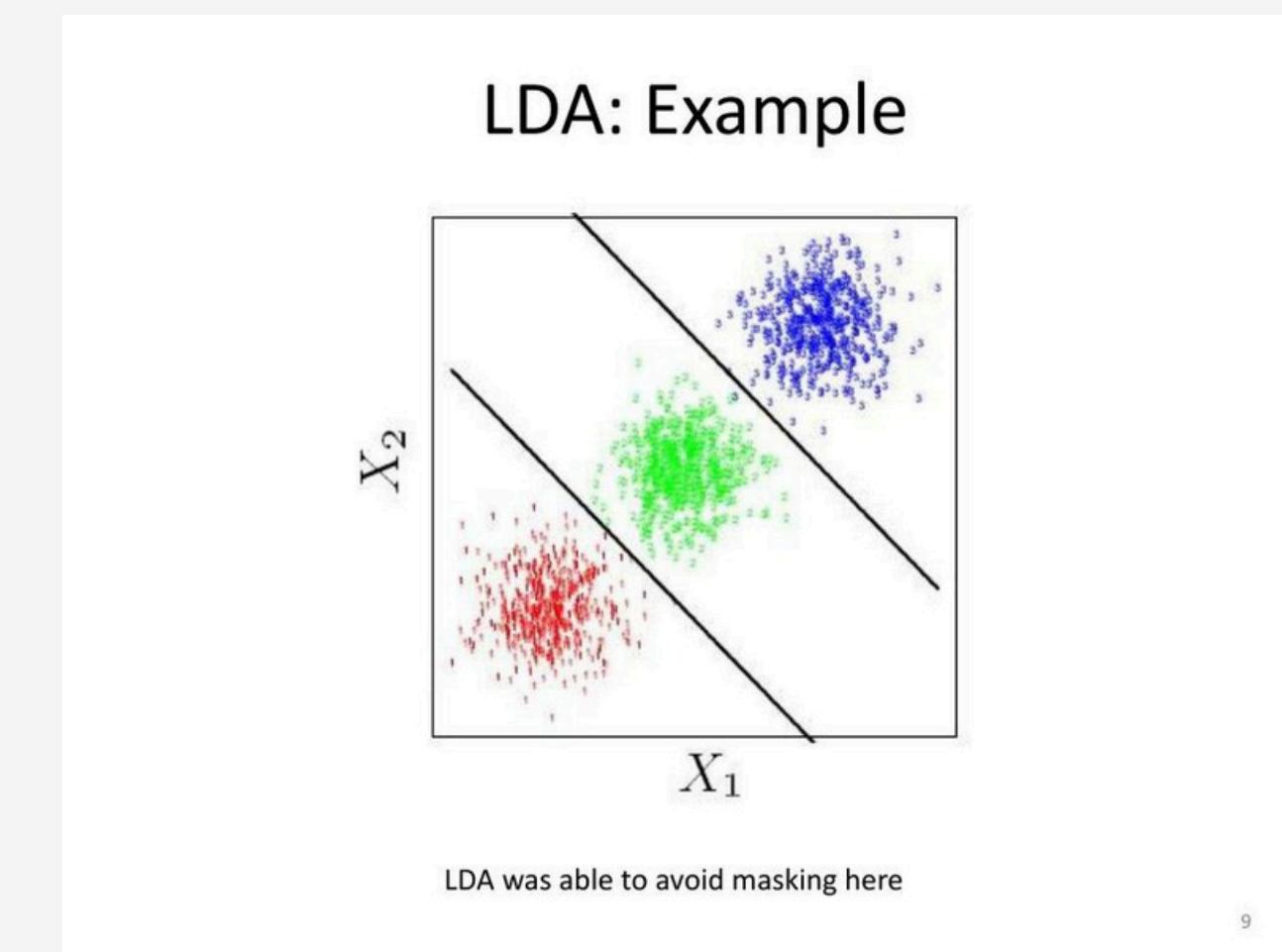
Nhưng thật sự SVM không hoạt động tốt với dữ liệu đa chiều . Vậy chúng ta sẽ cố gắng giảm chiều để thuật toán SVM hoạt động tốt.

Chúng tôi sẽ lựa chọn thuật toán LDA để giảm chiều dựa trên nhãn để cố gắng tăng độ chính xác của thuật toán SVM

2.4.Giới thiệu thuật toán LDA

Linear Discriminant Analysis (LDA) là kỹ thuật mà không chỉ giảm chiều mà còn giúp phân loại.

- Cách thức hoạt động: LDA tìm ra các đường phân cách giữa các lớp khác nhau trong dữ liệu. Nó cố gắng tối đa hóa phương sai giữa các lớp và giảm thiểu phương sai trong lớp.
- Ưu điểm: Thích hợp cho các bài toán phân loại.



2.4.1.Cách thức hoạt động của giải thuật LDA

Bước 1: Tính vectơ trung bình

- Tính vectơ trung bình toàn cục
- Tính vectơ trung bình của từng lớp

$$\mu = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\mu_k = \frac{1}{n_k} \sum_{x_i \in C_k} x_i$$

Bước 2: Tính ma trận phương sai

- Tính ma trận phương sai trong lớp
- Tính ma trận phương sai giữa các lớp

$$S_w = \sum_{k=1}^2 \sum_{x_i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^T$$

$$S_b = \sum_{k=1}^2 n_k (\mu_k - \mu)(\mu_k - \mu)^T$$

2.4.1.Cách thức hoạt động của giải thuật LDA

Bước 3: Tính ma trận chuyển đổi (Tìm W tối ưu):

+) Hàm mục tiêu của LDA

$$J(W) = \frac{W^T S_b W}{W^T S_w W}$$

Trong đó:

S_b là ma trận phương sai giữa các lớp (between-class scatter matrix).

S_w là ma trận phương sai trong cùng lớp (within-class scatter matrix).

W là ma trận chuyển đổi cần tìm.

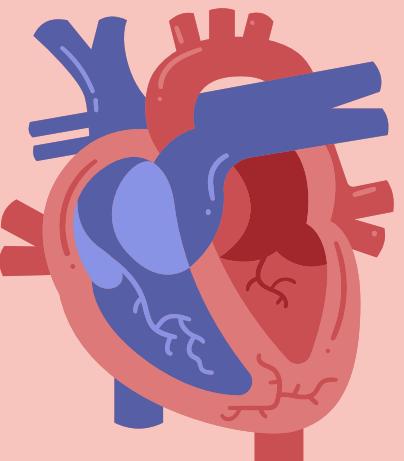
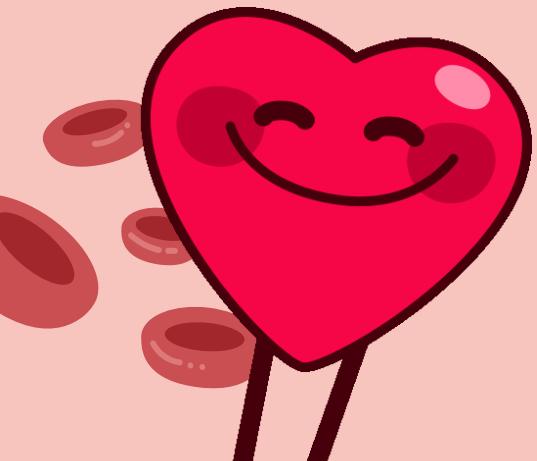
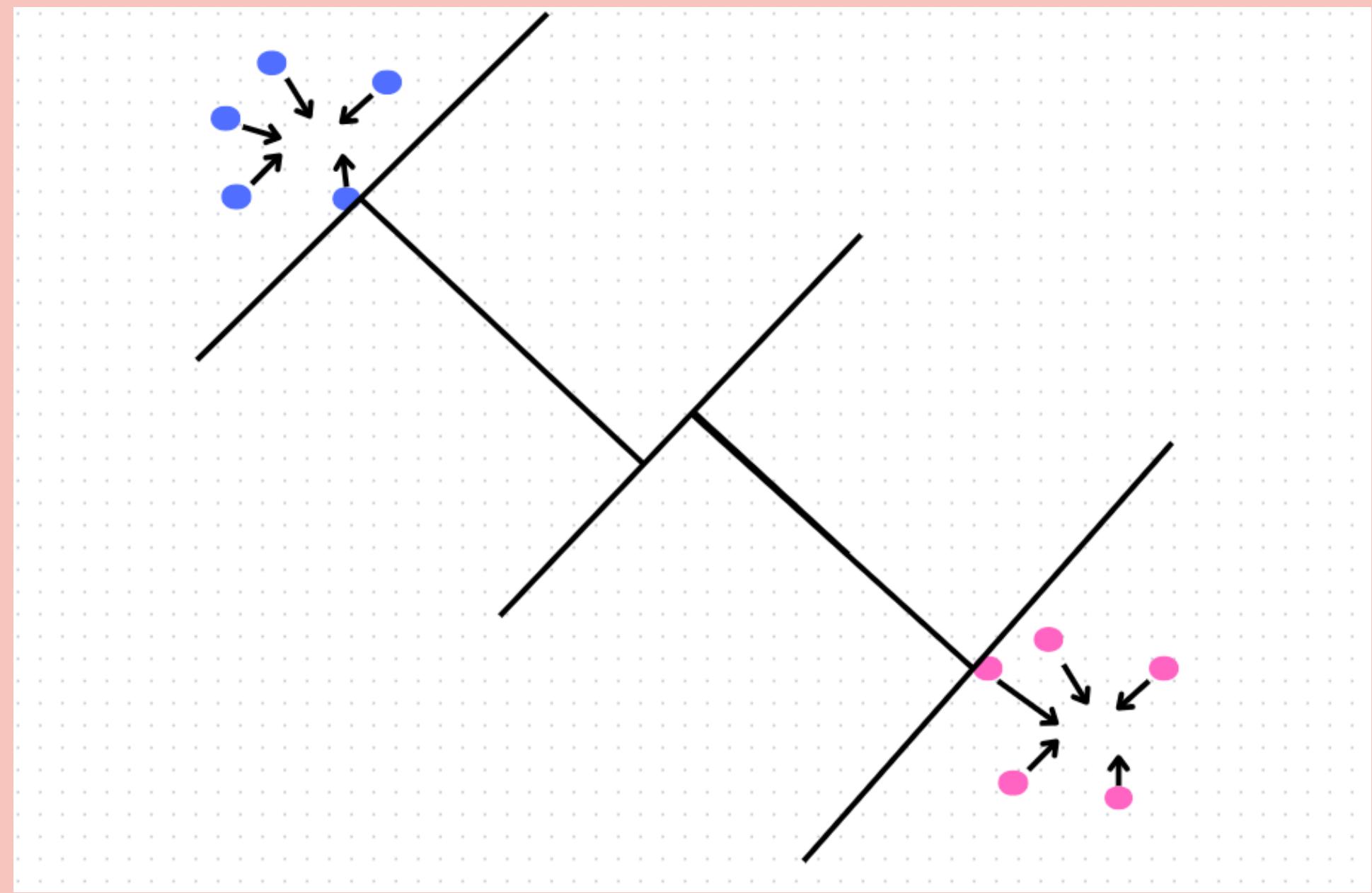
+) Để tìm W tối ưu, LDA chuyển bài toán tối ưu hóa này thành một bài toán giá trị riêng (eigenvalue problem). Cụ thể:

$$S_w^{-1} S_b v = \lambda v$$

Bước 4.Giảm chiều :

+) Quá trình chiếu dữ liệu gốc X xuống không gian mới bằng cách nhân với ma trận W .

2.4.2. Mô tả thuật toán SVM sau khi dùng LDA để giảm chiều

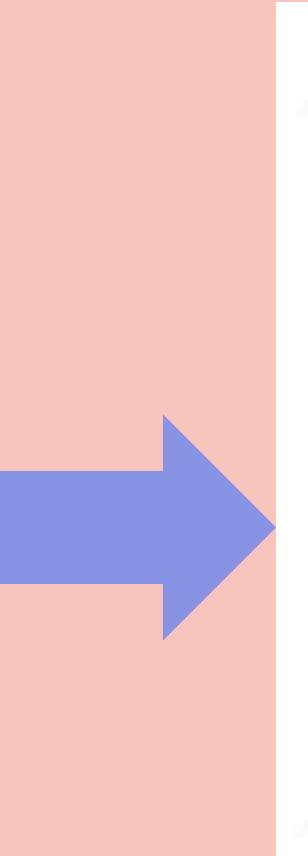


2.4.2. Kết quả sau khi sử dụng LDA :

Sau khi áp dụng LDA để giảm chiều dữ liệu từ sáu đặc trưng(có nghĩa là 6 chiều) xuống còn một đặc trưng (có nghĩa là 1 chiều), tôi thấy rằng việc sử dụng thuật toán SVM trên không gian đặc trưng đã được tối ưu hóa này đã giúp tăng đáng kể độ chính xác của mô hình. Cụ thể, độ chính xác (accuracy) đã cải thiện, chứng minh rằng LDA không chỉ giảm nhiễu mà còn tăng khả năng phân tách lớp, hỗ trợ SVM hoạt động hiệu quả hơn trên bài toán dự đoán bệnh tim."

```
Best Parameters: {'max_depth': None, 'min_samples_leaf': 4, 'min_samples_split': 10, 'n_estimators': 200}
Accuracy: 0.92
Precision: 0.9642857142857143
Recall: 0.9
F1 Score: 0.9310344827586207
Confusion Matrix:
[[38  2]
 [ 6 54]]
ROC-AUC: 0.9658333333333333
Matthews Correlation Coefficient (MCC): 0.8388884110737054
Log Loss: 0.22494453569005737
Cohen's Kappa: 0.8360655737704918

Feature Importance:
      Feature  Importance
5   QuantumPatternFeature  0.693444
3        Cholesterol  0.133224
0            Age  0.105020
4        HeartRate  0.032173
2       BloodPressure  0.030858
1           Gender  0.005280
```



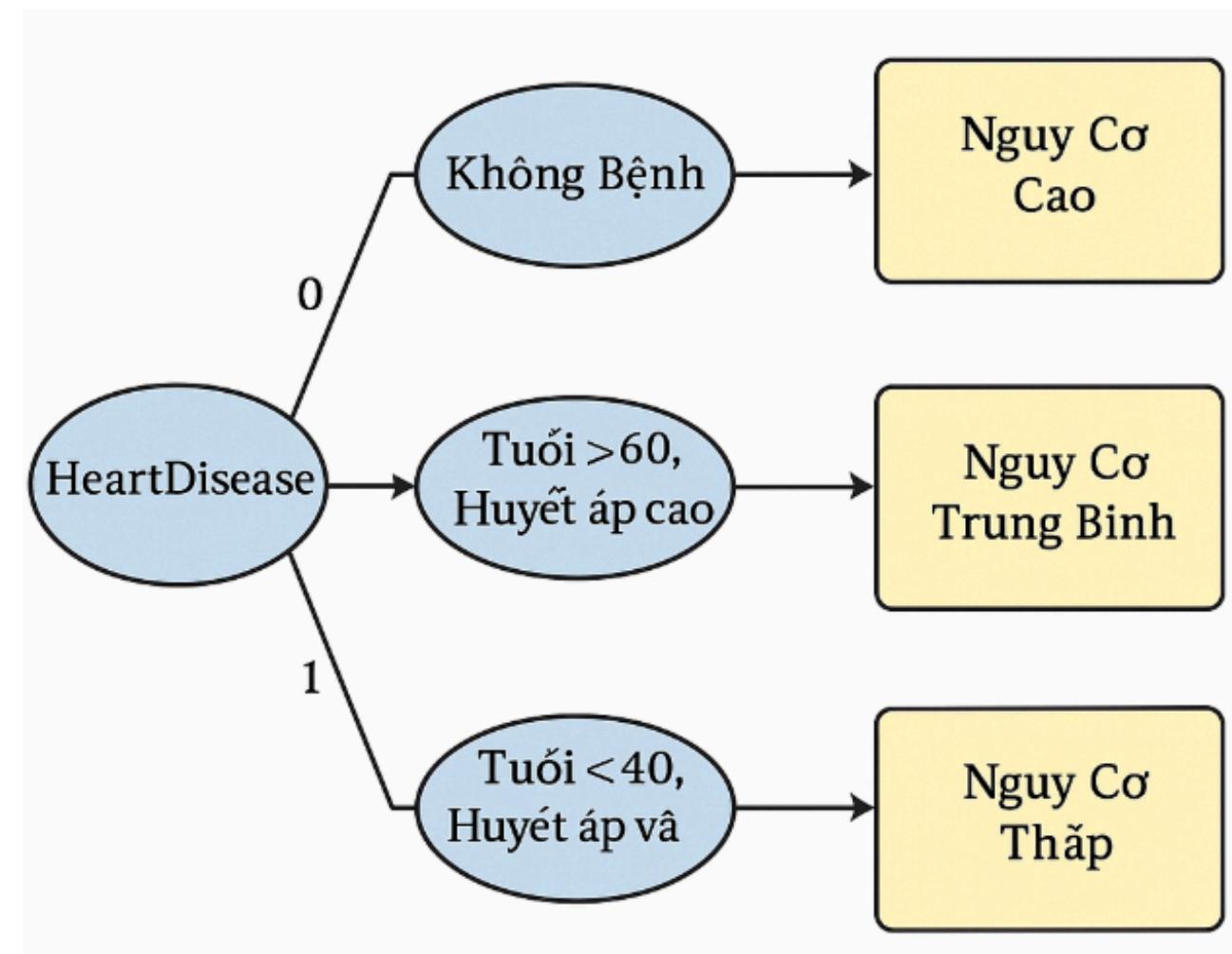
```
Best Parameters: {'C': 10, 'degree': 2, 'gamma': 'scale', 'kernel': 'rbf'}
Accuracy: 0.94
Precision: 0.95
Recall: 0.95
F1 Score: 0.9500000000000001

Classification Report:
precision    recall  f1-score   support
No Disease     0.93     0.93     0.93      40
Has Disease     0.95     0.95     0.95      60
accuracy          0.94     0.94     0.94     100
macro avg       0.94     0.94     0.94     100
weighted avg    0.94     0.94     0.94     100
```

2.4.3. Tiền xử lí dữ liệu(chuẩn đoán mức độ bệnh tim):

1.Kỹ Thuật Nhãn (Label Engineering)

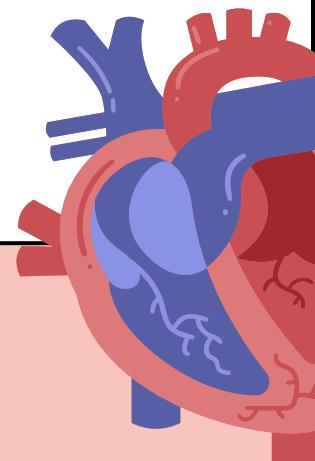
Kỹ Thuật Nhãn (Label Engineering) trong học máy là quá trình thiết kế, tạo ra và quản lý nhãn (labels) cho dữ liệu phù hợp để huấn luyện mô hình.



2.Kỹ Thuật Đặc Trưng (Feature Engineering)

Kỹ thuật đặc trưng (Feature Engineering) là quá trình biến đổi dữ liệu thô thành các đặc trưng mới giúp mô hình học máy dự đoán tốt hơn. Ví dụ:

- BP_Cholesterol (Huyết Áp × Cholesterol): Giúp mô hình nhận ra rằng khi cả huyết áp và cholesterol đều cao, nguy cơ bệnh tăng.
- Age_BP (Tuổi × Huyết Áp): Giúp mô hình học rằng huyết áp cao ở người lớn tuổi nguy hiểm hơn so với người trẻ.





2.4.4. Tiền xử lý dữ liệu(chuẩn đoán mức độ bệnh tim):

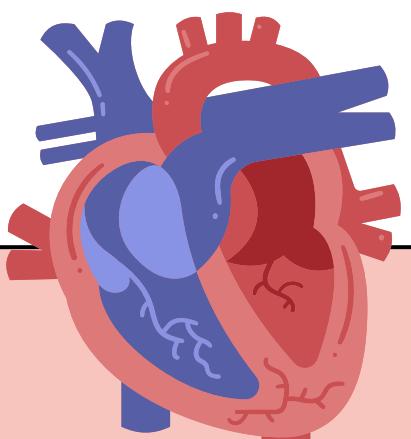
3. Lựa Chọn Đặc Trưng (Feature Selection)

- Lựa chọn đặc trưng giúp đơn giản hóa mô hình, giảm thời gian huấn luyện và tăng độ chính xác.
- SelectKBest (dùng f_classif, ANOVA F-test) chọn 7 đặc trưng phân biệt rõ các mức độ bệnh tim (DiseaseLevel) dựa trên F-score.
- F-score cao cho thấy đặc trưng có khả năng phân tách tốt.
- Phương pháp này giảm nguy cơ quá khớp và giữ lại các đặc trưng quan trọng.

4. Chuẩn Hóa(Standardization)

Chuẩn hóa (Standardization) là một kỹ thuật tiền xử lý dữ liệu, đưa các đặc trưng về cùng một thang đo với trung bình bằng 0 và độ lệch chuẩn bằng 1.

- Chuẩn hóa 7 đặc trưng (như BP_Cholesterol, Age_BP) chỉ dùng thông số từ tập huấn luyện để tránh rò rỉ dữ liệu.
- Đảm bảo các đặc trưng đồng nhất, nâng cao độ chính xác dự đoán DiseaseLevel.



2.4.4. Sử dụng thuật toán CatBoost và LightGBM (chuẩn đoán mức độ bệnh tim)

CatBoost là thuật toán boosting trên cây quyết định của Yandex. Điểm mạnh: xử lý tốt đặc trưng phân loại mà không cần one-hot encoding.

LightGBM là thuật toán boosting do Microsoft phát triển, nổi bật với tốc độ huấn luyện nhanh, tiêu thụ ít bộ nhớ và hiệu quả cao trên tập dữ liệu lớn. Nó dùng chiến lược xây cây theo chiều sâu (leaf-wise), giúp mô hình chính xác hơn và hỗ trợ xử lý đặc trưng phân loại hiệu quả.

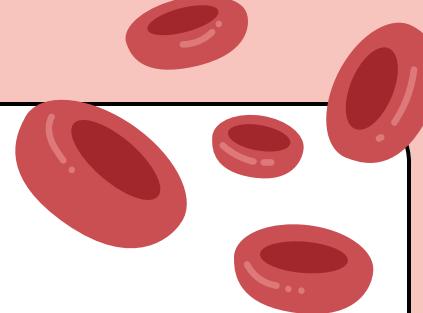
Độ chính xác của CatBoost: 0.9854

	precision	recall	f1-score	support
None	0.98	0.97	0.98	129
Mild	0.97	0.99	0.98	115
Moderate	1.00	0.98	0.99	116
Severe	0.98	1.00	0.99	120
accuracy			0.99	480
macro avg	0.99	0.99	0.99	480
weighted avg	0.99	0.99	0.99	480

Độ chính xác của LightGBM: 0.9812

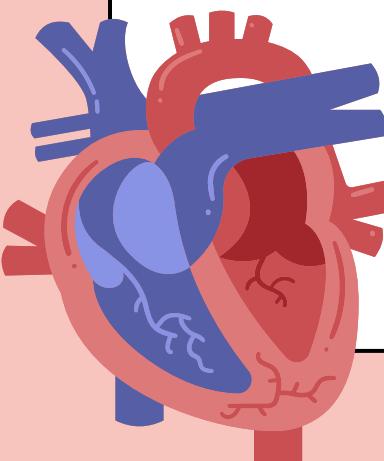
	precision	recall	f1-score	support
None	0.98	0.96	0.97	129
Mild	0.97	0.99	0.98	115
Moderate	1.00	0.97	0.99	116
Severe	0.97	1.00	0.98	120
accuracy				480
macro avg	0.98	0.98	0.98	480
weighted avg	0.98	0.98	0.98	480

Chuẩn đoán mức độ bệnh tim

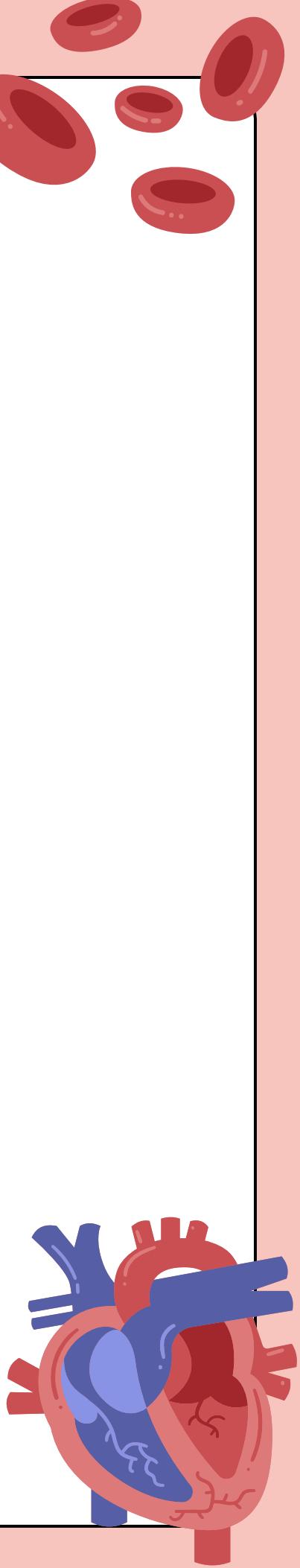
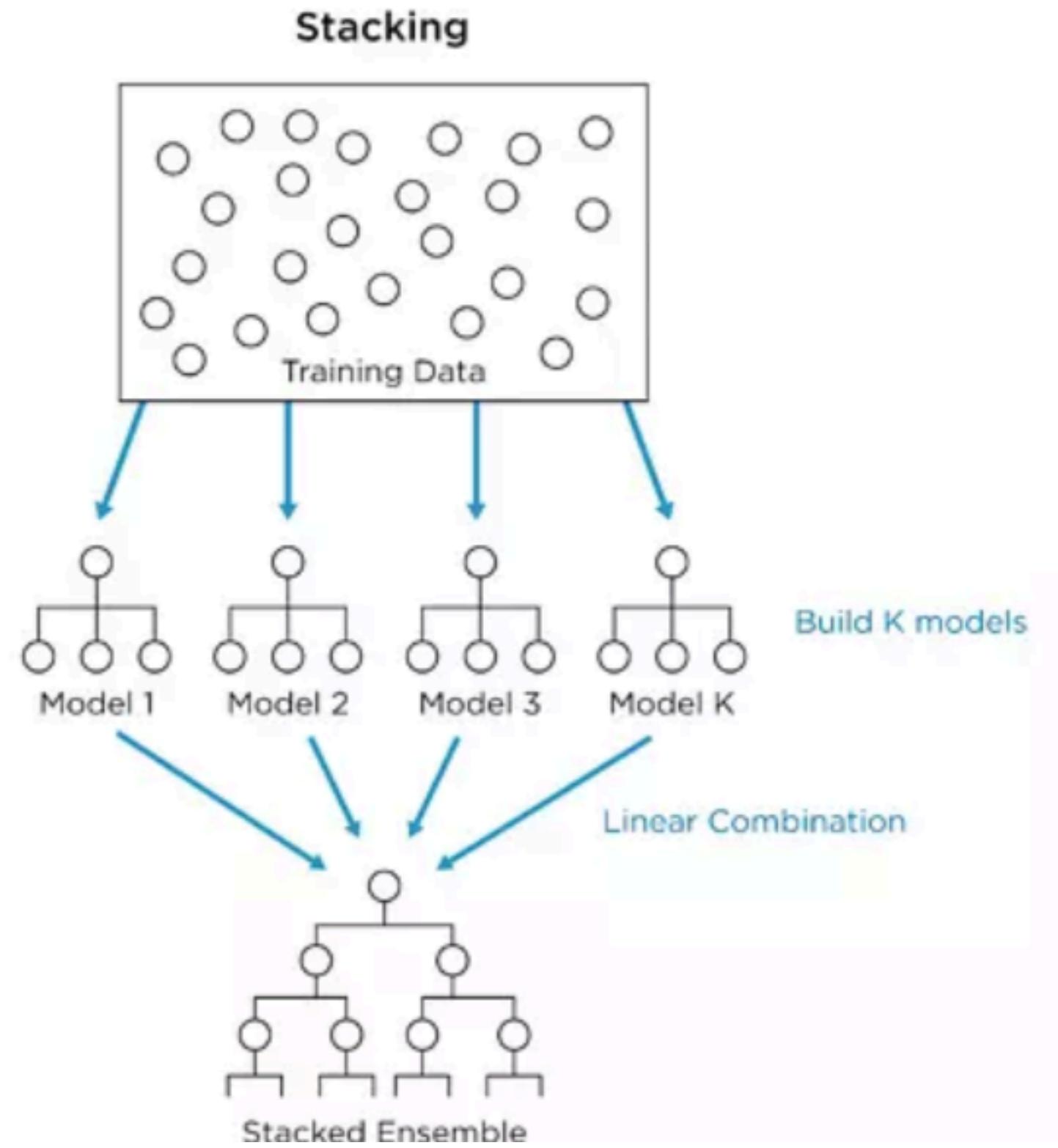


2.5. Giới thiệu kĩ thuật Stacking:

- Stacking là kĩ thuật ensemble, kết hợp nhiều mô hình để tăng độ chính xác.
- Cách hoạt động: “Các mô hình (base models) tạo dự đoán, rồi một meta-model học cách kết hợp chúng để cho kết quả tốt nhất.”
- Minh họa: “Giống như bạn hỏi ý kiến nhiều chuyên gia, rồi dùng kinh nghiệm để chọn ý kiến đúng nhất.”
- Lợi ích: Tận dụng thế mạnh của các mô hình khác nhau, giảm sai số.



Phương thức hoạt động:



Kết quả của mô hình kết hợp 2 thuật toán Catboost và LightGBM bằng Stacking:

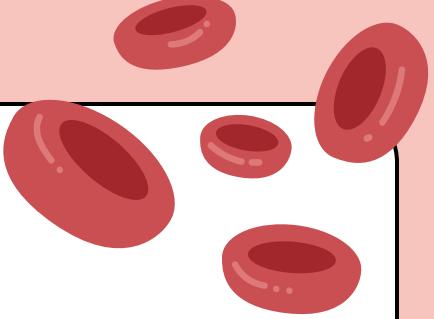
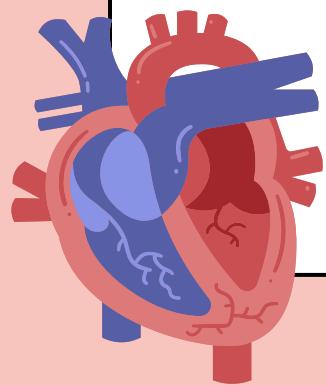
Độ chính xác của mô hình Stacking: 0.9875

Báo cáo phân loại:

	precision	recall	f1-score	support
None	0.98	0.98	0.98	129
Mild	0.98	0.99	0.99	115
Moderate	1.00	0.98	0.99	116
Severe	0.98	1.00	0.99	120
accuracy			0.99	480
macro avg	0.99	0.99	0.99	480
weighted avg	0.99	0.99	0.99	480

3.Kết luận:

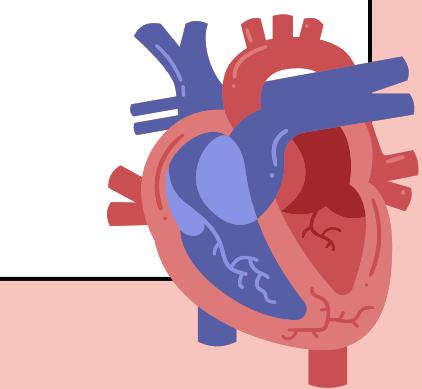
Trong bài tập nhóm môn Khai phá dữ liệu, nhóm em đã xây dựng mô hình phân loại mức độ bệnh từ dữ liệu y tế qua các bước: tiền xử lý, trực quan hóa, chọn và huấn luyện mô hình, đánh giá, và cải tiến. Ban đầu, em thử nghiệm các thuật toán SVM, Random Forest, KNN, sau đó dùng các mô hình mạnh như CatBoost, và LightGBM. Kết hợp CatBoost và LightGBM bằng kỹ thuật stacking đơn giản giúp đạt độ chính xác 98.75%, cải thiện precision, recall, f1-score trên cả 4 lớp bệnh, đồng thời đảm bảo mô hình ổn định và cân bằng – rất quan trọng trong y tế. Bài tập này giúp chúng em nắm vững quy trình khai phá dữ liệu, đánh giá mô hình toàn diện, và ứng dụng thực tiễn của học máy hiện đại.



Hạn chế:

Mặc dù mô hình đã đạt được kết quả rất tốt, nhận thấy vẫn còn một số hạn chế nhất định:

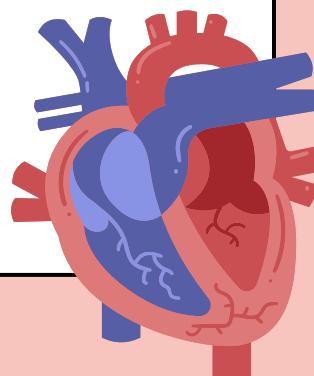
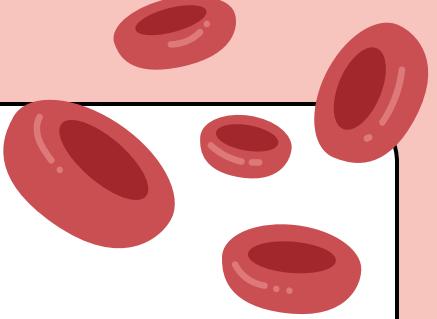
- Tập dữ liệu còn giới hạn về quy mô: Số lượng mẫu không quá lớn nên khả năng khái quát hóa cho dữ liệu thực tế vẫn còn là dấu hỏi.
- Chưa xử lý sâu về mất cân bằng dữ liệu: Một số lớp như "Severe" có số lượng ít hơn so với các lớp khác, dễ dẫn đến hiện tượng học lệch nếu kiểm soát tốt.
- Stacking mới dừng ở mức đơn giản: Đây mới chỉ dùng trung bình kết quả dự đoán của hai mô hình, chưa kết hợp theo cách có trọng số hoặc thông qua meta-model mạnh hơn.



Định hướng phát triển:

Trong tương lai, em mong muốn:

- Mở rộng và làm sạch dữ liệu, kết hợp kỹ thuật xử lý mất cân bằng như SMOTE, ADASYN.
- Làm thêm các model sử dụng mô hình deep learning như MLP hoặc CNN (nếu có dữ liệu hình ảnh) để mở rộng hướng nghiên cứu.



THANK'S
FOR
WATCHING