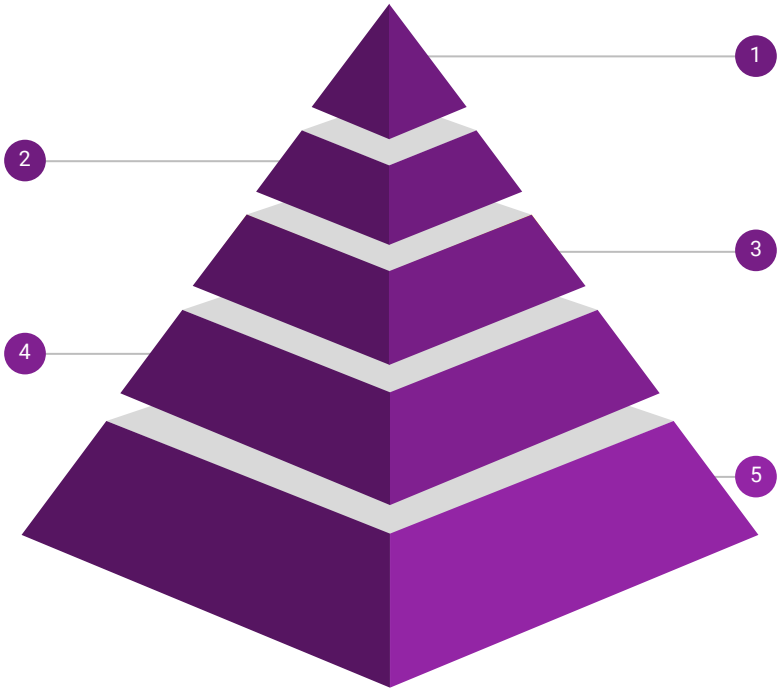
Heart Attack Prediction

DA44 Final Project
24/03/2024

Vo Huu Minh Chanh
Mentor: Mr. Tran Duc Trung

Nội dung



1 **Discovering**
Biz Understanding. Yêu cầu đề bài, vấn đề cần giải quyết

2 **Data Collection**
Thông tin về tập dữ liệu

3 **Preprocessing**
Data cleaning & Preparation

4 **Analyzing**
Exploratory Data Analysis
Diagnostic Analysis
Suggestion

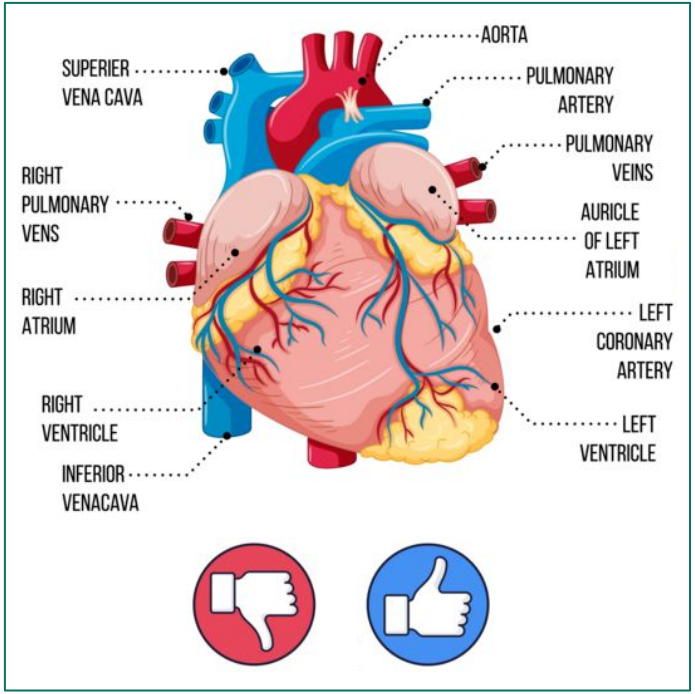
5 **Modeling**
Chọn lọc feature trên dataset, chọn model ML phù hợp. So sánh, kết luận và đề xuất

Discovering

Tổng quan

Sức khỏe tim mạch luôn là một chủ đề cần được quan tâm.

Các cơn đau tim hay nhồi máu cơ tim là một vấn đề sức khỏe toàn cầu quan trọng, đòi hỏi sự hiểu biết sâu sắc hơn về các dấu hiệu báo trước, cũng như các yếu tố giảm nhẹ rủi ro.



Đặt vấn đề

Bộ dữ liệu tập trung giải quyết duy nhất 1 vấn đề.

Điều này minh chứng cho nỗ lực chung nhằm nâng cao hiểu biết về sức khỏe tim mạch và chủ động mở đường cho một tương lai khỏe mạnh hơn.

TỶ LỆ MẮC BỆNH TIM

- **PHÒNG NGỪA**
- **KIỂM SOÁT**

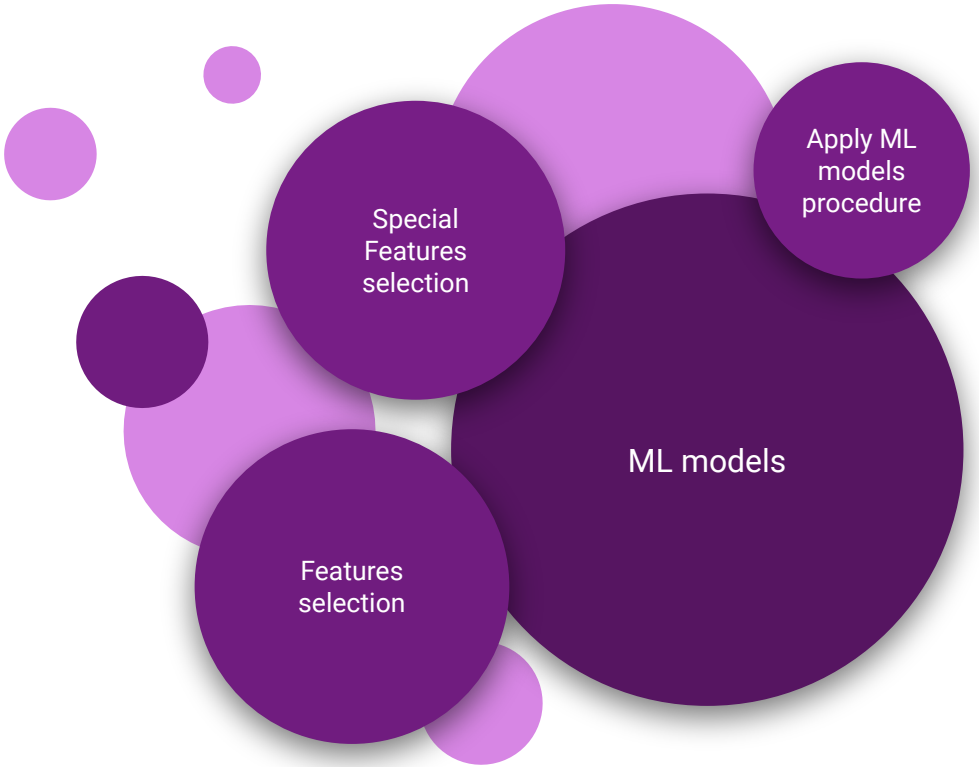
Tiếp cận vấn đề

Thực hiện các nhiệm vụ:

- Giải quyết bài toán **Binary Classification** về tỷ lệ mắc bệnh (Heart Attack Risk - HAR)
- Chọn lọc các thuộc tính (**features**) quan trọng ảnh hưởng đến HAR

  ** Các features đặc trưng (bệnh nhân tự đo/quan sát được, không cần thiết bị chuyên dụng) có đủ sức ảnh hưởng đến HAR hay không?*
- So sánh và chọn các model máy học (**ML models**) phù hợp với các features đã chọn.

  ** Quy trình xử lý mất cân bằng dữ liệu (resample) và chuẩn hóa (standardize) ảnh hưởng như thế nào đến kết quả?*



Data Collection & Preprocessing

Thông tin bộ dữ liệu

Dataset gồm

```
RangeIndex: 8763 entries, 0 to 8762
Data columns (total 26 columns):
```

- 26 fields thuộc tính (25 thuộc tính + 1 class Heart Attack Risk)

  Bao gồm cả thuộc tính phân loại và định lượng.

- 8763 records

Mô tả thông tin về bệnh lý và thói quen sinh hoạt của các bệnh nhân.

Patient ID	object
Age	int64
Sex	object
Cholesterol	int64
Blood Pressure	object
Heart Rate	int64
Diabetes	int64
Family History	int64
Smoking	int64
Obesity	int64
Alcohol Consumption	int64
Exercise Hours Per Week	float64
Diet	object
Previous Heart Problems	int64
Medication Use	int64
Stress Level	int64
Sedentary Hours Per Day	float64
Income	int64
BMI	float64
Triglycerides	int64
Physical Activity Days Per Week	int64
Sleep Hours Per Day	int64
Country	object
Continent	object
Hemisphere	object
Heart Attack Risk	int64

Tiền xử lý - Chất lượng dữ liệu

	Đánh giá
Tính đầy đủ	4/5
Nhất quán	5/5
Tuân thủ	4/5
Chính xác	4/5
Toàn vẹn	5/5
Kịp thời, phù hợp	-

Tiền xử lý

*Tính đầy đủ (Completeness)

Một số features không mang lại nhiều giá trị cho việc phân tích

Patient ID và các features về địa điểm: quốc tịch (Country), châu lục (Continent), bán cầu (Hemisphere)

→ **drop**

```
Category in Country is : ['Argentina' 'Canada' 'France' 'Thailand' 'Germany' 'Japan' 'Brazil'
 'South Africa' 'United States' 'Vietnam' 'China' 'Italy' 'Spain' 'India'
 'Nigeria' 'New Zealand' 'South Korea' 'Australia' 'Colombia'
 'United Kingdom']
Category in Continent is : ['South America' 'North America' 'Europe' 'Asia' 'Africa' 'Australia']
Category in Hemisphere is : ['Southern Hemisphere' 'Northern Hemisphere']
```

*Tính tuân thủ (Conformity)

Huyết áp (Blood Pressure) đang định dạng object

→ **Chuyển thành 2 features nhỏ là huyết áp tâm thu (BP_Systolic) và huyết áp tâm trương (BP_Diastolic) định dạng numeric**

Blood Pressure
158/88
165/93
174/99
163/100
91/88

```
4   Blood Pressure                8763 non-null   object
```

Tiền xử lý

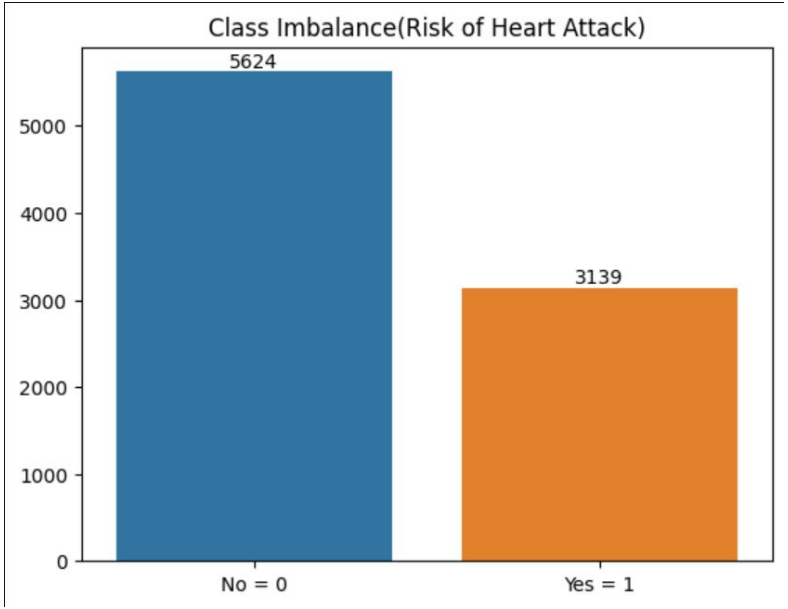
Mất cân bằng dữ liệu

Có sự mất cân bằng class dự đoán

- *Số ca có nguy cơ mắc bệnh là 3139 chiếm ~**36%** (Minority Class)*
- *Số ca không có nguy cơ là 5624 chiếm 64%, gấp 1.8 lần số ca có nguy cơ.*

→ Đánh giá mức độ: Nhẹ (Mild)

→ **Xử lý resampling khi áp dụng các ML models.**



Degree of imbalance	Proportion of Minority Class
Mild	20-40% of the data set
Moderate	1-20% of the data set
Extreme	<1% of the data set

Analyzing

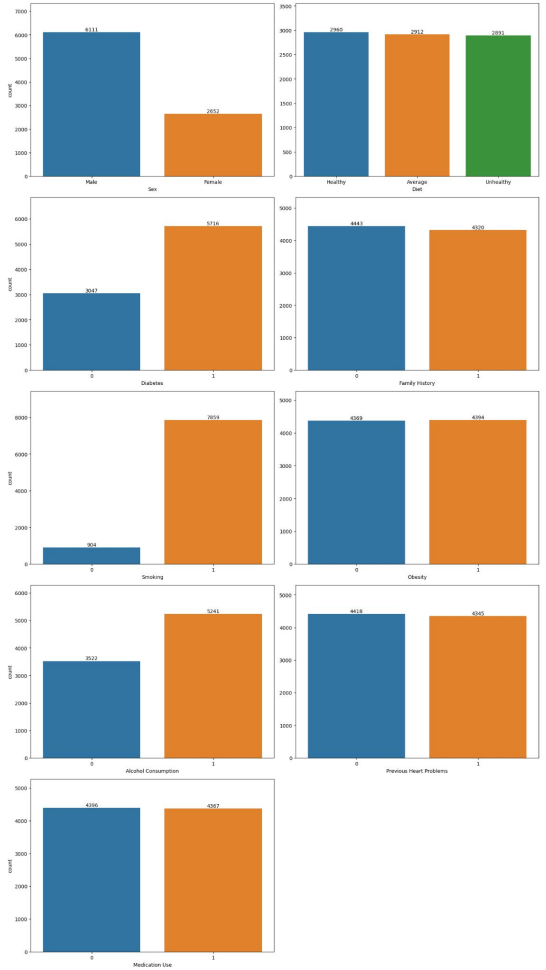
Insights

Các features phân loại - Số lượng

- **Giới tính (Sex) 6111/2652 (M/F)**
- Chế độ ăn (Diet) 2960/2912/2891 (Healthy/Avg/Unhealthy)
- **Đường huyết (Diabetes) 3047/5716 (0/1)**
- Tiền sử gia đình (Family History) 4443/4320 (0/1)
- **Hút thuốc (Smoking) 904/7859 (0/1)**
- Béo phì (Obesity) 4369/4394 (0/1)
- **Rượu/bia (Alcohol Consumption) 3522/5241 (0/1)**
- Tiền sử bệnh tim (Previous Heart Problems) 4418/ 4345 (0/1)
- Sử dụng thuốc (Medication Use) 4396/4367 (0/1)

Giới tính, Đường huyết, Hút thuốc và Rượu/bia có sự chênh lệch rõ ràng.

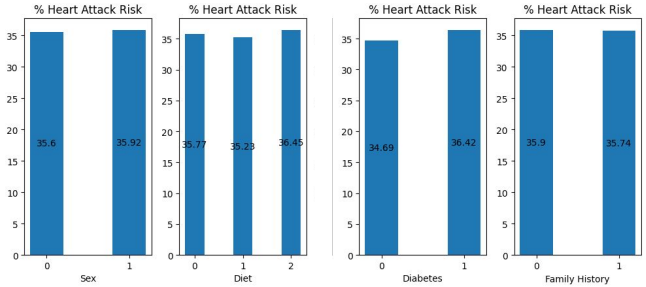
Hút thuốc chênh lệch rất mạnh (~89% có sử dụng)

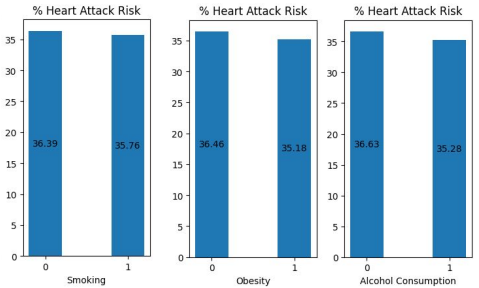


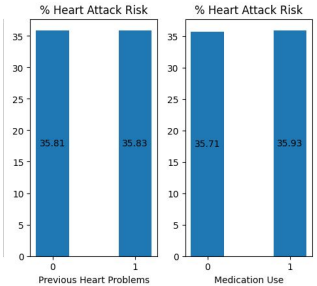
Insights

Các features phân loại - HAR

- Giới tính (Sex) 35.92/35.6 (M/F)
- **Chế độ ăn (Diet) 36.45/35.23/35.77 (Healthy/Avg/Unhealthy)**
- **Đường huyết (Diabetes) 34.69/36.42 (0/1)**
- Tiền sử gia đình (Family History) 35.9/35.74 (0/1)
- Hút thuốc (Smoking) 36.39/35.76 (0/1)
- **Béo phì (Obesity) 36.46/35.18 (0/1)**
- **Rượu/bia (Alcohol Consumption) 36.63/35.28 (0/1)**
- Tiền sử bệnh tim (Previous Heart Problems) 35.81.35.83 (0/1)
- Sử dụng thuốc (Medication Use) 35.71/35.93 (0/1)







Chế độ ăn, Đường huyết, Béo phì và Rượu/bia có sự chênh lệch.

➔ Chế độ ăn: Healthy và Unhealthy HAR cao, Avg là tốt nhất.

➔ Đường huyết: đường huyết cao nguy cơ tim mạch cao.

➔ Béo phì, Rượu/bia đi nghịch với HAR: không lành mạnh lại có HAR thấp hơn.

Insights

Các features phân loại

- Giới tính không ảnh hưởng nhiều dù tỷ lệ Nam nhiều hơn Nữ.
- **Chế độ ăn** các nhóm có số lượng tương đương nhau nhưng **Average** an toàn hơn.
- **Đường huyết** càng cao thì càng nguy hiểm.
- Tiền sử gia đình, Tiền sử bệnh tim, Sử dụng thuốc không ảnh hưởng nhiều, cả số lượng lẫn HAR đều tương đương.
- **Hút thuốc**, **Béo phì, Rượu/bia** đi ngược với HAR. Nhiều khả năng lúc trở thành "bệnh nhân" mới rèn luyện lối sống tích cực.

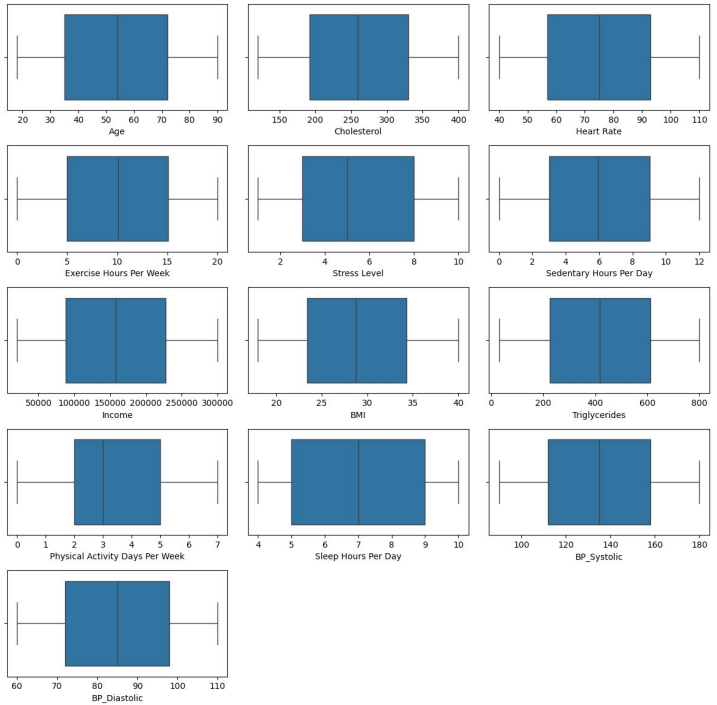
→ Do dataset chưa đủ chi tiết, một số features đang là câu hỏi Yes/No, cần thu thập chi tiết hơn.

(Hút thuốc trong bao lâu? Béo phì trong bao lâu, cấp độ? Sử dụng rượu/bia trong bao lâu?)

Insights

Các features định lượng - Phân vị

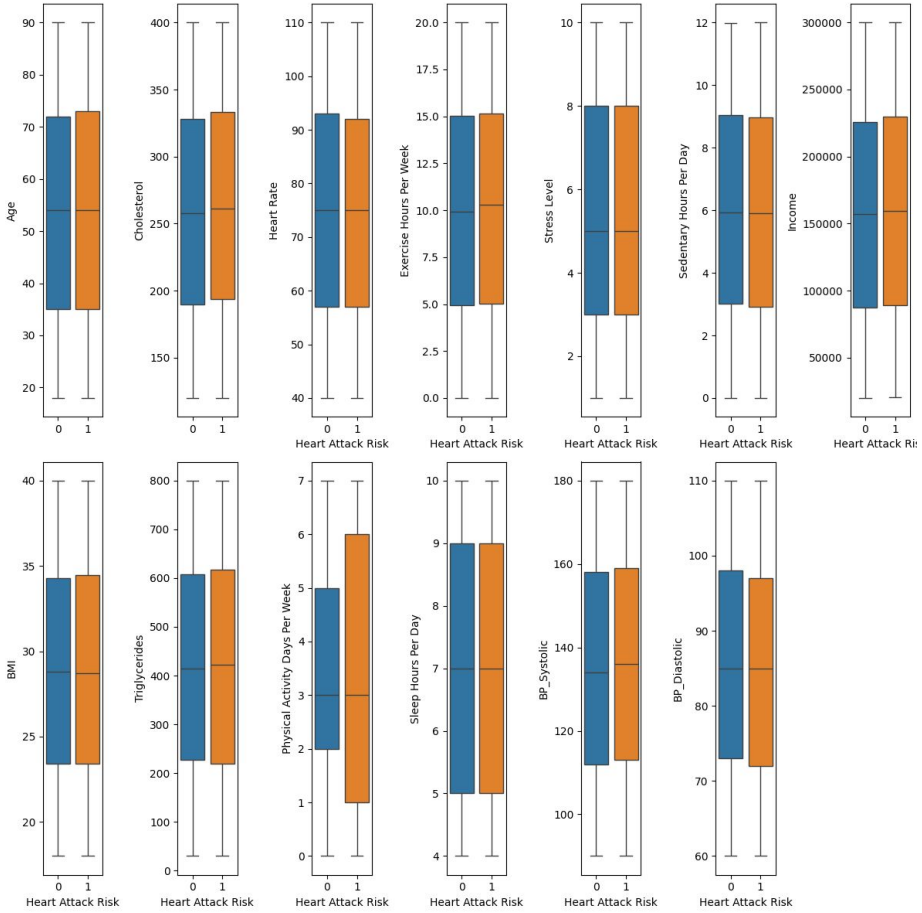
- Độ tuổi (Age)
- Cholesterol
- Nhịp tim (Heart Rate)
- Số giờ tập luyện/Tuần (Excercise Hours Per Week)
- Mức độ căng thẳng (Stress Level)
- Số giờ ít vận động/Ngày (Sedentary Hours Per Day)
- Thu nhập (Income)
- BMI
- Chất béo trung tính (Triglycerides)
- Số ngày hoạt động thể chất/Tuần (Physical Activity Days Per Week)
- Số giờ ngủ/Ngày (Sleep Hours Per Day)
- Huyết áp tâm thu (BP_Systolic)
- Huyết áp tâm trương (BP_Diastolic)



Insights

Các features định lượng - HAR

- Độ tuổi (Age)
- Cholesterol
- Nhịp tim (Heart Rate)
- Số giờ tập luyện/Tuần (Excercise Hours Per Week)
- Mức độ căng thẳng (Stress Level)
- Số giờ ít vận động/Ngày (Sedentary Hours Per Day)
- Thu nhập (Income)
- BMI
- Chất béo trung tính (Triglycerides)
- Số ngày hoạt động thể chất/Tuần (Physical Activity Days P
- Số giờ ngủ/Ngày (Sleep Hours Per Day)
- Huyết áp tâm thu (BP_Systolic)
- Huyết áp tâm trương (BP_Diastolic)



Insights

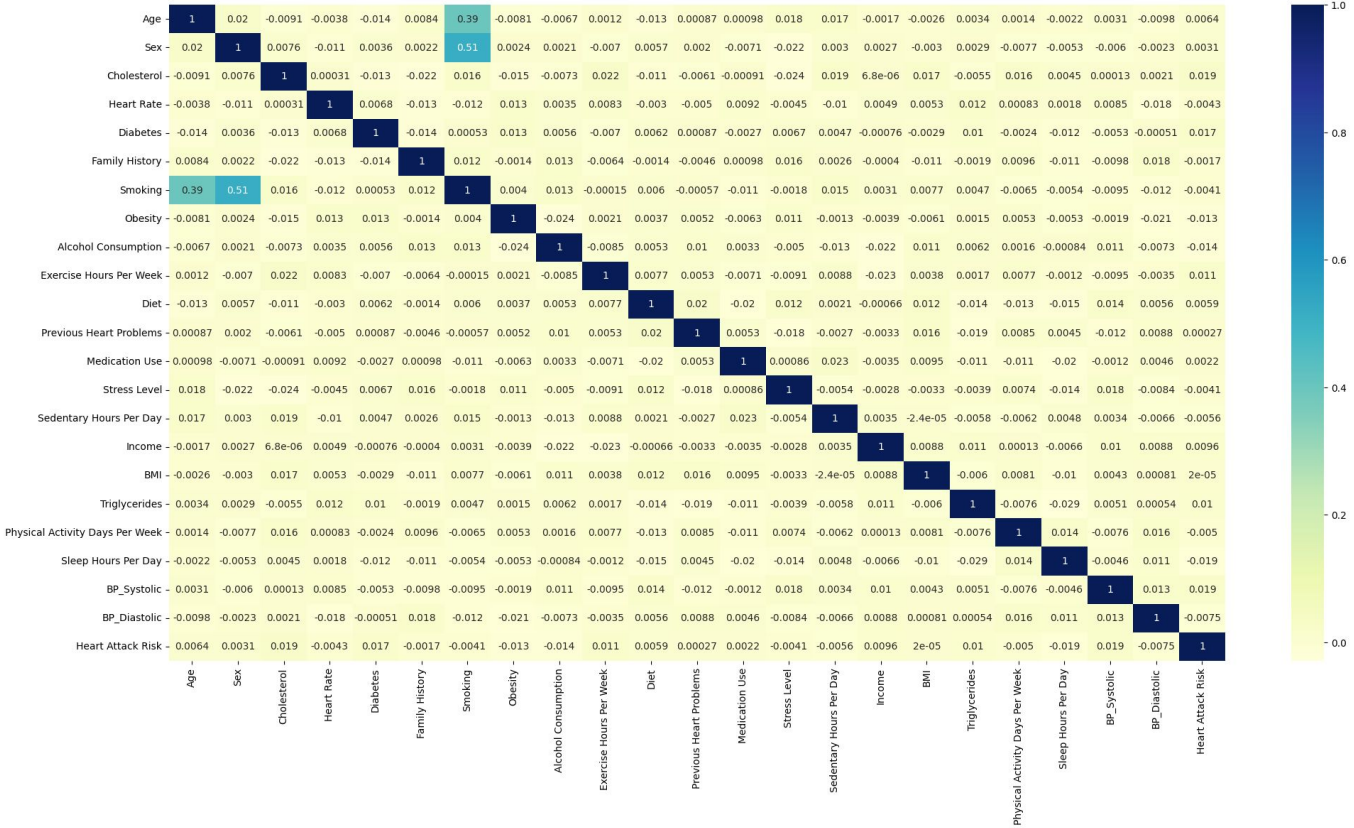
Các features định lượng

- Độ tuổi, Nhịp tim, Mức độ căng thẳng, BMI, Số giờ ngủ không ảnh hưởng nhiều.
- **Cholesterol, Huyết áp tâm thu** càng cao càng nguy hiểm.
- **Số giờ tập luyện/Tuần** cao cũng tăng nguy cơ (cần xem lại vì có liên quan Số ngày hoạt động thể chất/Tuần)
- **Số giờ ít vận động/Ngày** càng thấp càng có nguy cơ.
- Thu nhập cao cũng có tỷ lệ HAR cao nhưng không ảnh hưởng nhiều
  (do làm việc nhiều/thu nhập tốt nên khám bệnh nhiều hơn/...)
- **Chất béo trung tính, Số ngày hoạt động thể chất/Tuần** quá cao hoặc quá thấp cũng tăng tỷ lệ HAR.
- **Huyết áp tâm trương** càng thấp càng nguy hiểm

→ Kiểm soát chế độ dinh dưỡng, hàm lượng **Cholesterol, huyết áp**.
→ Sử dụng quan hệ giữa **Số giờ tập luyện, Số giờ ít vận động, Số ngày hoạt động thể chất, Số giờ ngủ** để cân chỉnh chế độ tập luyện.

➔ (Hoạt động thể chất trong tuần cần làm rõ ngày lao động và ngày tập luyện.
➔ Số ngày tập luyện và Số giờ tập luyện cần rõ ràng, điều độ.
➔ Số giờ ít vận động và số giờ ngủ cần tương ứng với các yếu tố trên, phân biệt ngày lao động và ngày tập luyện ...)

Tương quan các features



DISCOVERING | DATA COLLECTION | PREPROCESSING | **ANALYZING** | MODELING

Tương quan các features

Tương quan với HAR

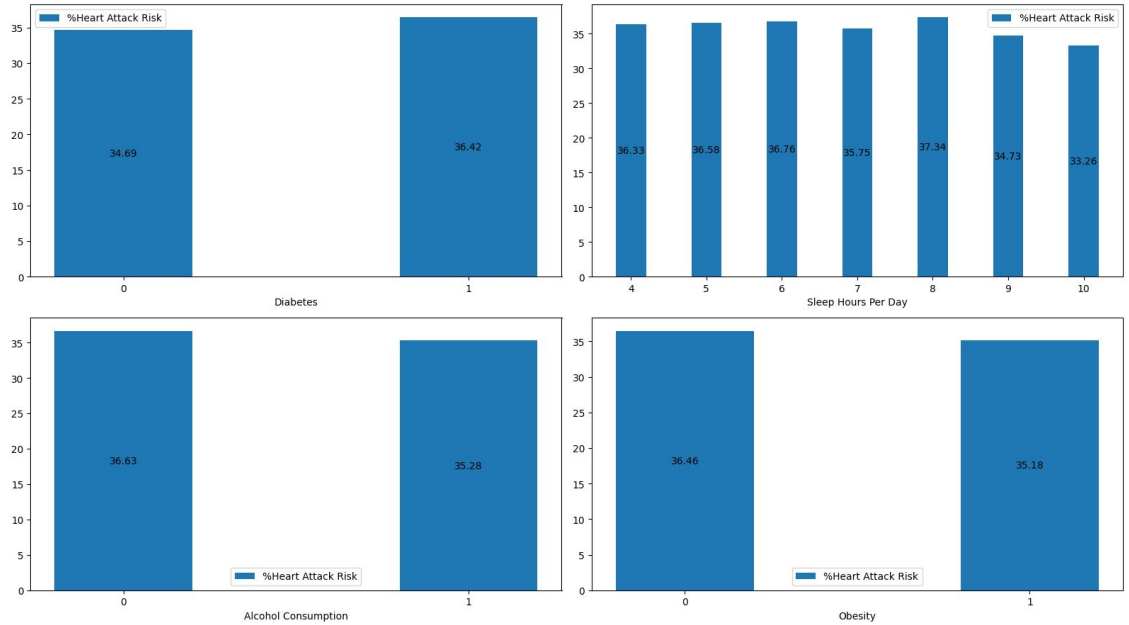
HAR bị ảnh hưởng nhiều bởi 8 yếu tố

- **Cholesterol #1, Huyết áp tâm thu #2, Đường huyết #4, Chất béo trung tính #8** (các features cần phải đo khám)
- **Số giờ ngủ/Ngày #3, Rượu/bia #5, Béo phì #6, Số giờ tập luyện/Tuần #7** (các features mỗi người có thể chủ động theo dõi)
- **Hút thuốc** không phải tác nhân trực tiếp ảnh hưởng HAR

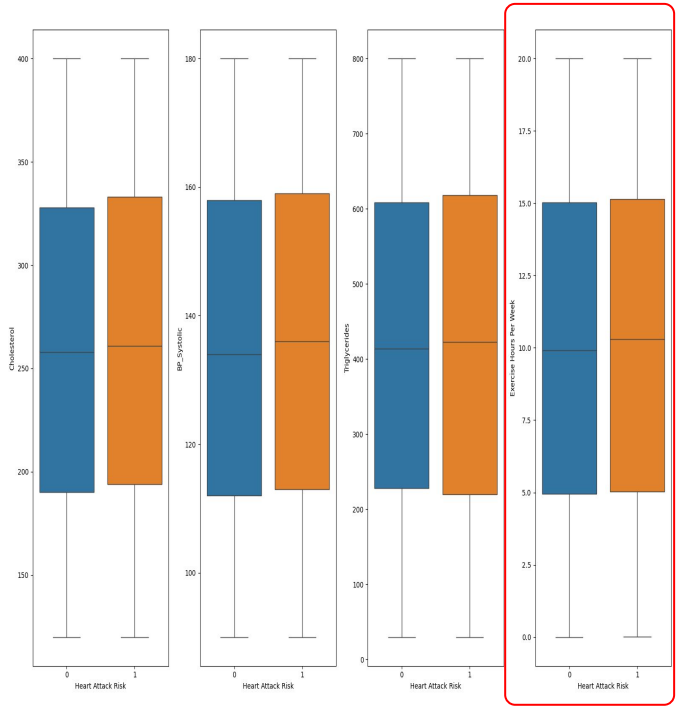
```
Heart Attack Risk                1.000000
Cholesterol                      0.019340
BP_Systolic                      0.018585
Sleep Hours Per Day             -0.018528
Diabetes                         0.017225
Alcohol Consumption             -0.013778
Obesity                         -0.013318
Exercise Hours Per Week          0.011133
Triglycerides                    0.010471
Income                           0.009628
BP_Diastolic                    -0.007509
Age                              0.006403
Diet                             0.005908
Sedentary Hours Per Day         -0.005613
Physical Activity Days Per Week -0.005014
Heart Rate                      -0.004251
Stress Level                    -0.004111
Smoking                         -0.004051
Sex                              0.003095
Medication Use                   0.002234
Family History                  -0.001652
Previous Heart Problems          0.000274
BMI                              0.000020
Name: Heart Attack Risk, dtype: float64
```

Visualize theo các features đề xuất

Diabetes, Sleep Hours Per Day,
Alcohol Consumption, Obesity

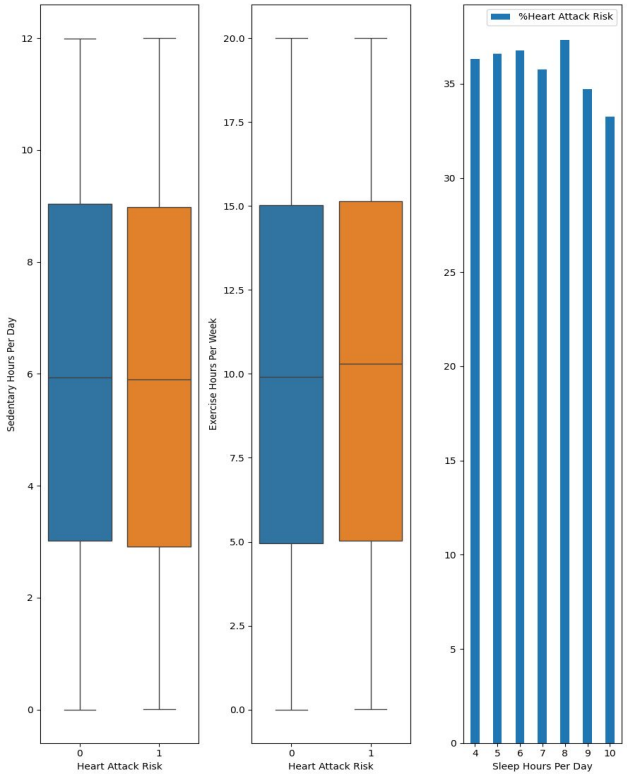


Cholesterol, BP_Systolic,
Triglycerides, Excercise Hours Per Week



DISCOVERING | DATA COLLECTION | PREPROCESSING | **ANALYZING** | MODELING

Kết luận



Tương tự EDA, và có thêm các insight:

- **Đường huyết và Béo phì** cho kết quả ngược nhau → Người có **đường huyết cao** nhưng lại **không béo phì** sẽ có tỷ lệ HAR cao hơn → Vấn đề chuyển hoá dinh dưỡng và tập luyện.
- **Số giờ ngủ** lý tưởng là **7, 9, 10 tiếng**. Ngủ **8 tiếng** nguy cơ cao #1 có thể nguyên nhân khác, tương tự lúc EDA (Ngủ đủ nhưng ít tập luyện hoặc tập luyện quá nhiều, số giờ ngủ không trùng với số giờ ít vận động/ngày...)

Modeling

Data preparing

** Ở phần Preprocessing đã thực hiện mã hoá Giới tính (Sex) và Chế độ ăn (Diet) thành numeric.*

- Tiến mã hoá Giới tính (one-hot encoding cho data không thứ tự) để apply các ML models

Features selection

Có 8 features giá trị Corr tốt → Thử top 10

- RFE (Recursive feature elimination) với Random Forest Classifier
- Correlation
- Các chỉ số thủ công (các chỉ số có thể tự theo dõi)

```
rf_estimator = RandomForestClassifier()
rfe = RFE(estimator=rf_estimator, n_features_to_select=top_features, step=1)
rfe = rfe.fit(X_rfe, y_rfe)
print('10 feature tốt nhất theo RFE:', X_rfe.columns[rfe.support_])
```

```
10 feature tốt nhất theo RFE: Index(['Age', 'Cholesterol', 'BP_Systolic', 'BP_Diastolic', 'Heart Rate',
       'Exercise Hours Per Week', 'Sedentary Hours Per Day', 'Income', 'BMI',
       'Triglycerides'],
      dtype='object')
```

RFE	'Age', 'Cholesterol', 'BP_Systolic', 'BP_Diastolic', 'Heart Rate', 'Exercise Hours Per Week', 'Sedentary Hours Per Day', 'Income', 'BMI', 'Triglycerides'
Correlation	'Cholesterol', 'BP_Systolic', 'Sleep Hours Per Day', 'Diabetes', 'Alcohol Consumption', 'Obesity', 'Exercise Hours Per Week', 'Triglycerides', 'Income', 'BP_Diastolic'
Thủ công	'Sleep Hours Per Day', 'Alcohol Consumption', 'Obesity', 'Exercise Hours Per Week'

Nhận xét: RFE chọn các features đi ngược với kết quả EDA và chẩn đoán

→ Chọn lọc features theo 3 kiểu **Tất cả, Top 10 Correlation và Top 4 thủ công**

ML models

Logistic Regression	• Giải quyết vấn đề liên quan đến tuyến tính hoặc phân loại nhị phân • Dễ apply và training.	• Dễ overfit ở những dataset high-dimension • Hiếm gặp trong thực tế
Decision Tree Classifier	• Giải quyết vấn đề phi tuyến tính • Làm việc trên dataset high-dimension • Dễ visualize và giải thích.	• Dễ overfit • Hay bị nhiễu • Tính toán phức tạp
Random Forest Classifier	• Giải quyết vấn đề overfit của DTC • Dùng được cho cả hồi quy và phân loại • Hoạt động tốt với giá trị phân loại và định lượng. • Hoạt động với cả dữ liệu khuyết và chưa chuẩn.	• Khó xác định tầm quan trọng của từng biến • Phức tạp, đòi hỏi sức mạnh tài nguyên và thời gian train.
Support Vector Machine	• Giải quyết vấn đề phi tuyến tính • Làm việc trên dataset high-dimension • Có thể làm việc trên dataset nhỏ	• Yêu cầu chọn đúng kernel **RBF (Radial Basis Function, phổ biến trong SVM classification)* • Không hiệu quả với dataset quá lớn
K Nearest Neighbors	• Có thể đưa ra dự đoán mà không cần train • Độ phức tạp tầm trung (O(n)) • Dùng được cho cả hồi quy và phân loại.	• Hay bị nhiễu với data không chuẩn • Yêu cầu chọn giá trị K đúng • Không hiệu quả với dataset quá lớn

So sánh các kiểu setup ML models

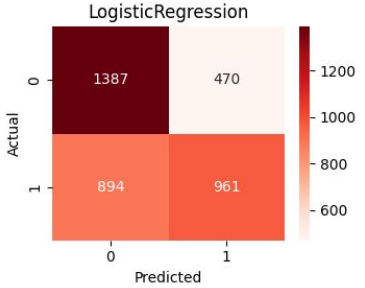
	All **Split → Standardize**	**All** **Resampling → Split**	**All** **Split → Resampling → Standardize**	**All** **Resampling → Split → Standardize**
LR (default)	*Acc: 0.64, Pre: 1.0*	*Acc: 0.5, Pre: 0.5*	*Acc: 0.56, Pre: 0.35*	*Acc: 0.61, Pre: 0.67*
DT (default)	*Acc: 0.52, Pre:0.33*	*Acc: 0.58, Pre: 0.58*	*Acc: 0.52, Pre: 0.35*	*Acc: 0.57, Pre: 0.57*
RF (random_state=42, n_estimators = 300, max_depth = 30, criterion = 'entropy')	***Acc: 0.64, Pre:0.39***	***Acc: 0.66, Pre: 0.72***	***Acc: 0.58, Pre: 0.33***	***Acc: 0.67, Pre: 0.72***
SVM (kernel = 'rbf)	*Acc: 0.64, Pre: 0.0*	*Acc: 0.48, Pre: 0.48*	*Acc: 0.57, Pre: 0.33*	*Acc: 0.66, Pre: 0.72*
KNN (k=10)	*Acc: 0.6, Pre: 0.31*	*Acc: 0.57, Pre: 0.58*	*Acc: 0.53, Pre: 0.34*	*Acc: 0.63, Pre: 0.63*

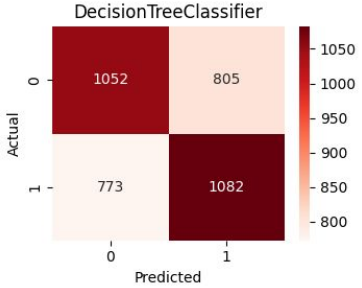
So sánh các kiểu setup ML models

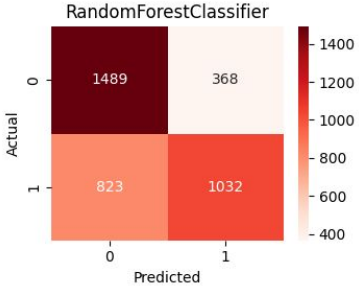
Các bước Resampling -> Split dataset -> Standardize cho kết quả tối ưu với tất cả model.

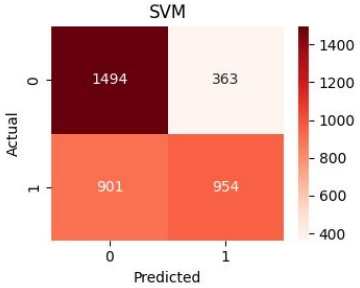
	All **Resampling → Split → Standardize**	**Top 10 Corr** **Resampling → Split → Standardize**	**Top 4 thủ công** **Resampling → Split → Standardize**
LR (default)	*Acc: 0.61, Pre: 0.67*	*Acc: 0.59, Pre: 0.58*	*Acc: 0.5, Pre: 0.5*
DT (default)	*Acc: 0.57, Pre: 0.57*	*Acc: 0.55, Pre: 0.55*	*Acc: 0.56, Pre: 0.57*
RF (random_state=42, n_estimators = 300, max_depth = 30, criterion = 'entropy')	***Acc: 0.67, Pre: 0.72***	***Acc: 0.62, Pre: 0.63***	***Acc: 0.58, Pre: 0.58***
SVM (kernel = 'rbf)	*Acc: 0.66, Pre: 0.72*	*Acc: 0.6, Pre: 0.61*	*Acc: 0.51, Pre: 0.51*
KNN (k=10)	*Acc: 0.63, Pre: 0.63*	*Acc: 0.57, Pre: 0.57*	*Acc: 0.56, Pre: 0.57*

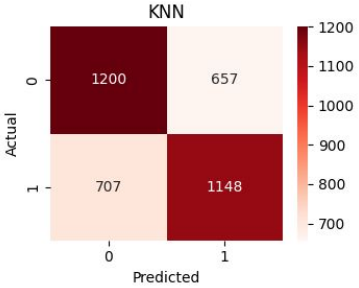
So sánh hiệu quả các model trước features selection

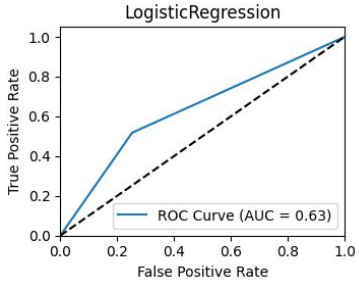




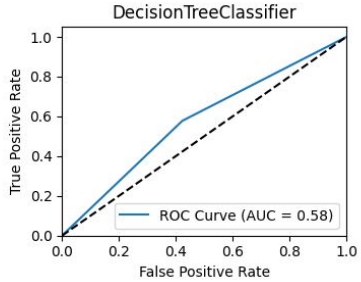




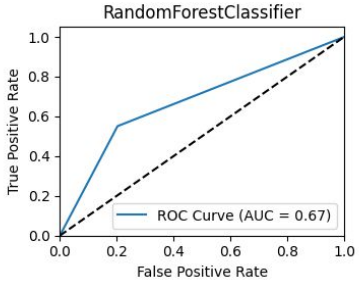




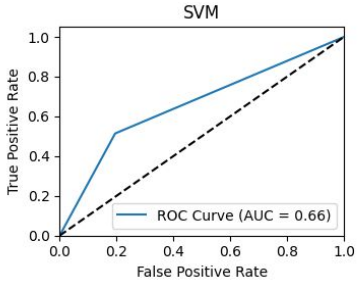
Acc: 0.61, Pre: 0.67



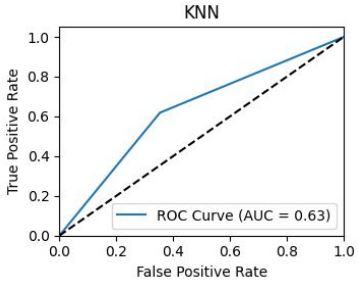
Acc: 0.57, Pre: 0.57



Acc: 0.67, Pre: 0.72



Acc: 0.66, Pre: 0.72



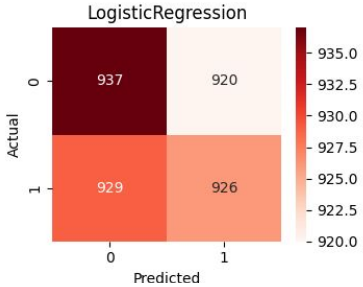
Acc: 0.63, Pre: 0.63

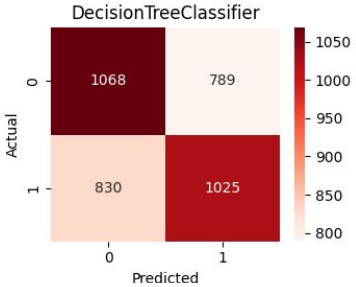
-Random Forest cho kết quả tốt nhất về nhiều chỉ số

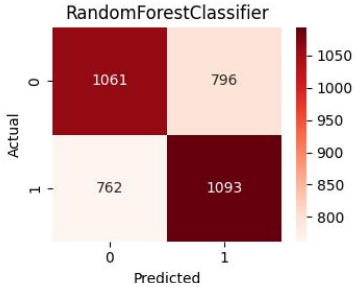
-Decision Tree dù độ chính xác chưa cao nhưng cho dự đoán tốt về mặt y khoa (FP>FN)

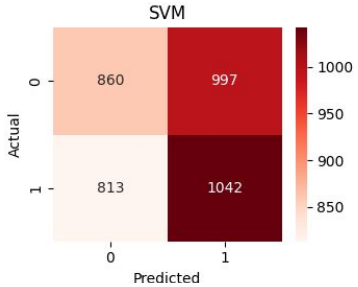
So sánh hiệu quả các model sau features selection

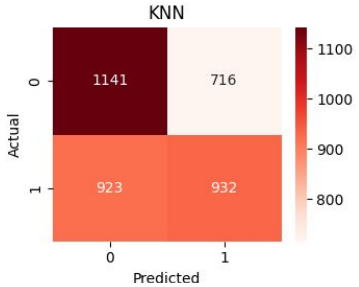
Top 4 thủ công

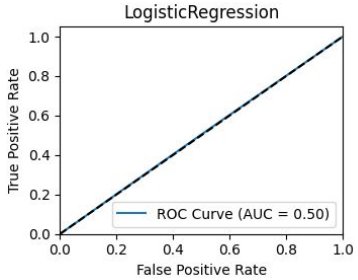


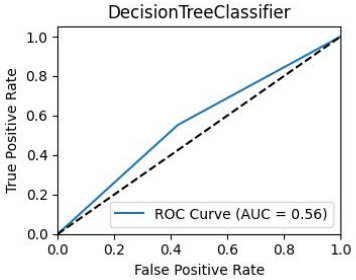


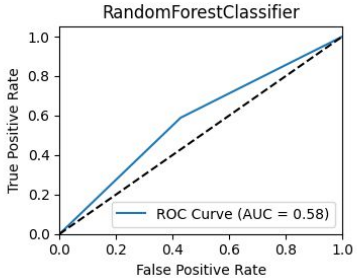


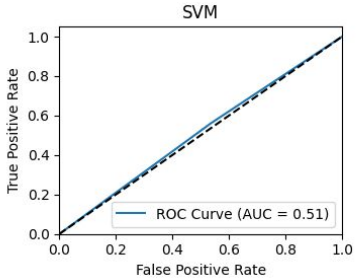


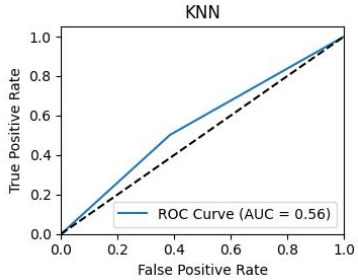












Acc: 0.5, Pre: 0.5 | Acc: 0.56, Pre: 0.57 | Acc: 0.58, Pre: 0.58 | Acc: 0.51, Pre: 0.51 | Acc: 0.56, Pre: 0.57

-Random Forest vẫn cho kết quả tốt nhất về nhiều chỉ số, và tốt trong cả trường hợp chọn lọc top 4 tính năng.
-Các features được chọn làm giảm độ chính xác, cân nhắc về việc đánh đổi giữa các yếu tố.
  + Top 10 features đầu vào (10/24 features, RF 67 -> 61%)
  + Top 4 features thủ công (4/24 features, RF 67 -> 58%).
-Trường hợp top 4 thủ công, model RF và SVM có chỉ số FP > FN, khá tốt với chẩn đoán y khoa.

Tổng kết

Tổng kết

- Với những ưu điểm vượt trội, model **Random Forest Classifier** mang lại hiệu quả tốt nhất cho chủ đề này.
- Chỉ cần 4 thông tin cơ bản **Số giờ ngủ/Ngày, Rượu/bia, Béo phì, Số giờ tập luyện/Tuần**. Người bình thường không cần thiết bị chuyên khoa vẫn thể đoán bản thân có nguy cơ mắc bệnh tim mạch hay không.

  → Ra quyết định nhanh trước khi tiến hành thăm khám.
- **Features selection** làm giảm đáng kể kết quả dự đoán của model.

  → Hầu hết thông tin của dataset sau khi preprocessing đều quý giá.

  Mỗi thông tin đều góp phần giúp chúng ta chủ động phòng tránh và kiểm soát các vấn đề tim mạch. Giúp bản thân chủ động điều chỉnh chế độ sinh hoạt, ăn uống và nghỉ dưỡng tốt hơn.

Phân bổ tiến độ

24/03/2024

DEMO

Dự kiến phát triển

Tìm thêm các dataset mới, có dữ liệu chuẩn và nhiều giá trị để phân tích hơn.

Điều chỉnh các hyperparameter ở Random Forest để tăng độ chính xác của mô hình.

Nguồn tham khảo

storytelling with data - Cole Nussbaumer Knaflic

stackoverflow.com

kaggle.com
https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset

matplotlib.org

Xin cảm ơn!

Mọi người đã chú ý theo dõi

Contact Me

+84 984 32 0841

chanhvokts@gmail.com