**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

🙤🙧🟍🙥🙦



**PHÂN TÍCH VÀ DỰ ĐOÁN**

**NGUY CƠ NHỒI MÁU CƠ TIM**

|  |  |  |  |
| --- | --- | --- | --- |
| **Nhóm 21** | | | |
| Sinh viên thực hiện: | | | |
| STT | Họ tên | MSSV | Ngành |
| 1 | Nguyễn H. Cát Thy | 21522665 | KHDL |

**TP. HỒ CHÍ MINH – 12/2024**

# GIỚI THIỆU

Đề tài này nhằm phân tích và dự đoán nguy cơ nhồi máu cơ tim dựa trên bộ dữ liệu "Heart Attack Analysis & Prediction Dataset" từ Kaggle. Mục tiêu chính của đề tài là khám phá các yếu tố nguy cơ như độ tuổi, giới tính, huyết áp, cholesterol, và loại cơn đau thắt ngực để đánh giá khả năng mắc phải căn bệnh này. Để thực hiện, nhóm sử dụng các công cụ như Pandas và Numpy để xử lý dữ liệu, kết hợp với Matplotlib và Seaborn để trực quan hóa kết quả. Các mô hình học máy như Logistic Regression, K-Nearest Neighbors, Decision Tree, Random Forest, và XGBoost được áp dụng để dự đoán nguy cơ nhồi máu cơ tim. Kết quả đạt được là việc xây dựng được các mô hình phân loại chính xác, từ đó giúp đưa ra dự đoán về khả năng mắc bệnh của từng cá nhân.

Bộ dữ liệu phân tích được tham khảo từ Kaggle [1], xử lý để thực hiện đề tài này. Đề tài là do nhóm tự thiết kế, không dựa trên bất kỳ đề tài nào khác và chỉ được sử dụng để phục vụ môn học DS105.

# MÔ TẢ BỘ DỮ LIỆU

Bộ dữ liệu "Heart Attack Analysis & Prediction Dataset" từ Kaggle chứa thông tin về các bệnh nhân và nguy cơ mắc nhồi máu cơ tim. Bộ dữ liệu này được sử dụng để phân tích các yếu tố ảnh hưởng đến khả năng bị nhồi máu cơ tim và dự đoán nguy cơ mắc bệnh dựa trên các đặc trưng của bệnh nhân. Bộ dữ liệu bao gồm nhiều đặc trưng liên quan đến sức khỏe như độ tuổi, huyết áp, cholesterol, và tình trạng cơn đau thắt ngực, với mục tiêu xác định mối quan hệ giữa các yếu tố này và nguy cơ nhồi máu cơ tim.

Mô tả các đặc trưng (Features):

|  |  |  |  |
| --- | --- | --- | --- |
| STT | Tên thuộc tính | Kiểu dữ liệu | Nội dung thuộc tính |
| 1 | age | int64 | Tuổi của bệnh nhân |
| 2 | sex | int64 | Giới tính của bệnh nhân |
| 3 | cp | int64 | Loại cơn đau thắt ngực |
| 4 | trtbps | int64 | Huyết áp khi nghỉ (mmHg) |
| 5 | chol | int64 | Mức cholesterol (mg/dL) |
| 6 | fbs | int64 | Đường huyết lúc đói |
| 7 | restecg | int64 | Kết quả điện tâm đồ khi nghỉ |
| 8 | thalachh | int64 | Nhịp tim tối đa đạt được |
| 9 | exng | int64 | Cơn đau thắt ngực do gắng sức |
| 10 | oldpeak | float64 | Thể hiện độ chênh lệch ST [2] (ST depression) đo được trên điện tâm đồ khi nghỉ ngơi và sau khi tập thể dục |
| 11 | slp | int64 | Độ dốc của đoạn ST [2] (Slope of the peak exercise ST segment) trên điện tâm đồ sau khi thực hiện bài kiểm tra gắng sức. |
| 12 | caa | int64 | Số lượng mạch máu lớn bị tắc nghẽn |
| 13 | thall | int64 | Kết quả kiểm tra mức độ máu lưu thông trong tim |
| 14 | output | int64 | Mục tiêu dự đoán |

*Bảng 1. Mô tả chi tiết các thuộc tính trong bộ dữ liệu*

Trong đó:

* Giá trị của biến cp thể hiện các loại cơn đau thắt ngực của bệnh nhân chia thành 4 loại: 0: đau thắt ngực điển hình, 1: đau thắt ngực bất thường, 2: đau ngực nhưng không đau thắt, 3: không có triệu chứng.
* Biến fbs thể hiện lượng đường huyết lúc đói 1: lớn hơn 120 mg/dL, 0: nhỏ hơn hoặc bằng 120 mg/dL
* Biến restecg thể hiện kết quả điện tâm đồ khi nghỉ 0: bình thường, 1: bất thường sóng ST-T [2], 2: tăng gánh thất trái
* Giá trị của biến output thể hiện nguy cơ bị nhồi máu cơ tim: 0: không có nguy cơ nhồi máu cơ tim, 1: có nguy cơ nhồi máu cơ tim.

Bộ dữ liệu nhóm đã chọn bao gồm 303 mẫu và 14 biến. Trong đó, có 9 biến phân loại và 5 biến số. Bộ dữ liệu không có giá trị khuyết và chỉ có 1 giá trị trùng lặp. Cái nhìn sơ bộ về bộ dữ liệu như sau:

A graph with numbers and a bar

Description automatically generated with medium confidence A screenshot of a computer

Description automatically generated

Hình . Phân phối nhãn của bộ dữ liệu Hình . Giá trị khuyết của bộ dữ liệu

# PHƯƠNG PHÁP PHÂN TÍCH

Kế hoạch phân tích dữ liệu của nhóm là kết quả của việc khai phá và nghiên cứu dữ liệu từ đó đưa ra quy trình phân tích dữ liệu phù hợp cho bộ dữ liệu được chọn. Phương pháp của nhóm được tóm tắt trong sơ đồ sau:

A diagram of a diagram

Description automatically generated

Hình . Quy trình PTDL

Mô tả:

Đầu tiên, nhóm sẽ chọn ra bộ dữ liệu phù hợp để phục vụ cho đồ án môn học dựa trên các thông tin cơ bản và chung nhất của bộ dữ liệu. Sau đó, nhóm sẽ tiến hành xác định và định nghĩa bài toán dựa trên bộ dữ liệu đã chọn. Bộ dữ liệu tiếp đến sẽ được nhóm khai phá và thực hiện các bước tiền xử lý như là làm sạch, xóa giá trị trùng lặp. Sau khi đã giải quyết các vấn đề cơ bản của bộ dữ liệu, nhóm sẽ tiếp tục phân tích thăm dò để có cái nhìn sâu sắc hơn về bộ dữ liệu. Từ đó, hiểu rõ về các yếu tố làm tăng nguy cơ bị nhồi máu cơ tim. Nếu có vấn đề trong bộ dữ liệu mà sẽ ảnh hưởng đến kết quả của mô hình dự đoán như giá trị ngoại lai thì nhóm sẽ quay lại bước tiền xử lý lần nữa và sau đó xây dựng các mô hình dự đoán và độ đo đánh giá thích hợp. Với kết quả dự đoán đúng và sai được tính ở bước trước, nhóm sẽ đánh giá sơ bộ về các thuộc tính ảnh hưởng đến kết quả trong bộ dữ liệu.

# PHÂN TÍCH THĂM DÒ CHUYÊN SÂU

Bộ dữ liệu của nhóm chỉ có giá trị số nên sau khi xử lý dữ liệu trùng lặp, nhóm tiến hành phân tích thăm dò bộ dữ liệu như sau:

## Khám phá mối quan hệ giữa các biến số và phân loại với biến mục tiêu:

A chart of numbers and a red line

Description automatically generated with medium confidence

Hình . Biểu đồ tương quan giữa các thuộc tính

Nhóm quy ước thang đo của độ tương quan: thấp: 0 → 0.1, trung bình: 0.1 → 0.4, cao: 0.4 → 1. Có thể thấy ở biểu đồ trên các thuộc tính cp, thalachh, exng và oldpeak là các biến có mối tương quan cao với output và có thể là các yếu tố quan trọng để xác định nguy cơ nhồi máu cơ tim vì thế nhóm sẽ tập trung tìm hiểu các thuộc tính này.

## Old peak có mối liên hệ với nguy cơ nhồi máu cơ tim không?

A diagram of a graph

Description automatically generated with medium confidence

Hình . Mối quan hệ giữa old peak và nguy cơ nhồi máu cơ tim

Dựa trên biểu đồ thể hiện mối quan hệ giữa old peak (mức độ giảm ST trên điện tâm đồ khi gắng sức) và nguy cơ nhồi máu cơ tim, ta có thể nhận xét như sau:

* Nhóm không có nguy cơ nhồi máu cơ tim (output = 0) có giá trị oldpeak lớn hơn, cho thấy mức độ giảm ST khi gắng sức cao hơn. Ngược lại, nhóm có nguy cơ nhồi máu cơ tim (output = 1) có oldpeak thấp hơn, cho thấy mức giảm ST ít đáng kể hơn.
* Giải thích: Oldpeak phản ánh sự thiếu máu cơ tim cục bộ khi gắng sức. Một mức giảm ST lớn (giá trị oldpeak cao) có thể liên quan đến tình trạng thiếu máu cơ tim ổn định nhưng không phải lúc nào cũng dẫn đến nhồi máu cơ tim cấp tính. Nhóm có nguy cơ nhồi máu cơ tim (output = 1) có oldpeak thấp có thể vì bệnh nhân ở tình trạng nghiêm trọng hơn hoặc không thể tạo ra đáp ứng gắng sức đầy đủ để biểu hiện giảm ST rõ ràng.

## Các loại cơn đau thắt ngực có liên quan đến nhau hoặc khả năng bị nhồi máu cơ tim không?

A graph with red and grey bars

Description automatically generated

Hình . Mối quan hệ giữa các loại cơn đau ngực và Nguy cơ Nhồi máu cơ tim

Nhận xét:

* Loại 0 (đau thắt ngực điển hình): Đây là loại cơn đau thắt ngực phổ biến nhất trong bộ dữ liệu. Có sự phân bố không đều giữa các nhóm nguy cơ (output), với số lượng người không có nguy cơ nhồi máu cơ tim cao hơn rất nhiều so với những người có nguy cơ.
* Loại 1 (đau thắt ngực bất thường): Đây là loại cơn đau thắt ngực ít phổ biến hơn, và số lượng người có nguy cơ nhồi máu cơ tim lại cao hơn so với nhóm không có nguy cơ.
* Loại 2 (đau nhưng không đau thắt): Đây là loại cơn đau thắt ngực có sự chênh lệch đáng kể giữa nhóm có và không có nguy cơ. Nhóm có nguy cơ nhồi máu cơ tim chiếm tỷ lệ lớn hơn.
* Loại 3 (Không có triệu chứng): Cơn đau thắt ngực loại 3 có rất ít người trong bộ dữ liệu, với tỷ lệ giữa nhóm không có nguy cơ và có nguy cơ là gần như bằng nhau.

## Cơn đau thắt ngực do gắng sức có liên quan đến nhau hoặc khả năng bị nhồi máu cơ tim không?

A graph with a red and grey rectangular bar

Description automatically generated with medium confidence

Hình . Mối quan hệ giữa cơn đau thắt ngực do gắng sức và Nguy cơ Nhồi máu cơ tim

**Nhận xét:**

Nguy cơ nhồi máu cơ tim chủ yếu tập trung ở những người không có cơn đau thắt ngực do gắng sức. Trong khi đó, những người có cơn đau thắt ngực do gắng sức thường có tỉ lệ không mắc nhồi máu cơ tim cao hơn. Điều này là do nhồi máu cơ tim không phải lúc nào cũng có dấu hiệu rõ ràng, những bệnh nhân có nguy cơ nhồi máu cơ tim cao có thể đã gặp các vấn đề nghiêm trọng về động mạch vành hoặc tim mạch, khiến cơn đau thắt ngực không xuất hiện dù tim bị tổn thương.

## Nhịp tim tối đa đạt được có đóng vai trò như một yếu tố nguy cơ đối với nhồi máu cơ tim không?

A diagram of a graph

Description automatically generated with medium confidence

Hình . Quan hệ giữa nhịp tim tối đa đạt được với nguy cơ nhồi máu cơ tim

**Nhận xét:**

Nhóm có nguy cơ nhồi máu cơ tim có nhịp tim tối đa đạt được trung bình và trung vị cao hơn so với nhóm không có nguy cơ. Điều này có thể rằng ở những bệnh nhân có nguy cơ nhồi máu cơ tim, nhịp tim tối đa đạt được trong quá trình hoạt động có xu hướng cao hơn.

# XÂY DỰNG MÔ HÌNH ĐÁNH GIÁ VÀ KẾT QUẢ

## Tiền xử lý dữ liệu:

Sau các bước phân tích, thăm dò dữ liệu, nhóm nhận thấy dữ liệu có nhiều giá trị ngoại lai cần được xử lý trước khi xây dựng mô hình đánh giá. Nhóm quyết định chọn xử lý các giá trị ngoại lai sử dụng phương pháp IQR (Interquartile Range) [3]. Sau đó dùng MinmaxScaler [4] để đưa dữ liệu về cùng một khoảng [0, 1].

## Các mô hình lựa chọn:

* Logistic Regression
* K Neighbors
* SVC
* Decision Tree
* Random Forest
* Gradient Boosting
* XGBoost

## Kết quả

|  |  |  |  |
| --- | --- | --- | --- |
| Mô hình | Accuracy | Recall | Precision |
| Logistic Regression | 85.25% | 86.49% | 88.89% |
| K Neighbors | 78.69% | 81.08% | 83.33% |
| SVC | 80.33% | 78.38% | 87.88% |
| Decision Tree | 73.77% | 72.97% | 89.66% |
| Random Forest | 83.61% | 81.08% | 87.88% |
| Gradient Boosting | 73.77% | 70.27% | 83.87% |
| XGBoost | 81.97% | 78.38% | 90.62% |

*Bảng 2. Kết quả của các mô hình dự đoán*

Nhóm sử dụng các độ đo là Accuracy, Recall và Precision và có kết quả dự đoán như trên. Bộ dữ liệu "Heart Attack Analysis & Prediction Dataset" có sự phân bổ tương đối cân bằng giữa hai nhãn mục tiêu (có nguy cơ và không có nguy cơ nhồi máu cơ tim), nhưng vẫn có một số chênh lệch nhất định. Điều này giúp các mô hình như Logistic Regression và Random Forest hoạt động tốt do chúng xử lý tốt trên các bộ dữ liệu cân bằng. Ngược lại, các mô hình như K Neighbors hoạt động kém hơn vì không tận dụng được mối quan hệ phi tuyến và bị ảnh hưởng bởi các khoảng cách trong không gian dữ liệu. Và với kích thước nhỏ chỉ 303 mẫu, các mô hình phức tạp như XGBoost và Gradient Boosting khó có thể phát huy hết tiềm năng, dẫn đến hiệu suất chỉ ở mức trung bình. Những mô hình này thường cần dữ liệu lớn hơn để huấn luyện và điều chỉnh tốt hơn.

# KẾT LUẬN

Đề tài đã thành công trong việc phân tích và dự đoán nguy cơ nhồi máu cơ tim dựa trên bộ dữ liệu "Heart Attack Analysis & Prediction Dataset". Qua quá trình tiền xử lý dữ liệu như làm sạch, loại bỏ giá trị trùng lặp và xử lý ngoại lai bằng phương pháp IQR, nhóm đã chuẩn bị một bộ dữ liệu sạch, sẵn sàng cho việc huấn luyện mô hình.

Nhóm đã áp dụng các mô hình học máy như Logistic Regression, K-Nearest Neighbors, SVC, Decision Tree, Random Forest, Gradient Boosting và XGBoost. Trong đó, Logistic Regression đạt hiệu suất cao nhất với độ chính xác 85.2%, recall 86.5% và precision 88.9%. Các mô hình Random Forest và Gradient Boosting cũng thể hiện kết quả ấn tượng với độ chính xác trên 80%.

Kết quả phân tích thăm dò dữ liệu đã xác định các đặc trưng quan trọng như oldpeak, nhịp tim tối đa (thalachh), loại cơn đau ngực (cp) và tình trạng cơn đau do gắng sức (exng) có mối liên hệ chặt chẽ với nguy cơ nhồi máu cơ tim. Nhóm đã xây dựng được một quy trình khoa học để phân tích, đánh giá và dự đoán nguy cơ mắc bệnh, từ đó góp phần hỗ trợ phát hiện sớm và kịp thời can thiệp cho các bệnh nhân.

**TÀI LIỆU THAM KHẢO**

Heart Attack Analysis & Prediction Dataset. Link: <https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset/data> (10/12/2024)

Đoạn ST trong kết quả điện tâm đồ là gì? Link: <https://www.vinmec.com/vie/bai-viet/doan-st-trong-ket-qua-dien-tam-do-la-gi-vi?utm_source=chatgpt.com> (10/12/2024)

Interquartile Range in Statistics. Link: <https://www.geeksforgeeks.org/interquartile-range/> (12/12/2024)

MinMaxScaler. Link: <https://scikit-learn.org/1.5/modules/generated/sklearn.preprocessing.MinMaxScaler.html> (12/12/2024)

**PHỤ LỤC PHÂN CÔNG NHIỆM VỤ**

|  |  |  |
| --- | --- | --- |
| **STT** | **Thành viên** | **Nhiệm vụ** |
| 1 | Nguyễn H. Cát Thy | Tìm bộ dữ liệu phù hợp  Khám phá dữ liệu  Tiền xử lý  Xây dựng mô hình dự đoán  Đánh giá kết quả |