Nhóm: 19

Tên thành viên:

- Lê Viết Lâm Quang (20520290)
- Hồ Thanh Tịnh (20520813)
- Nguyễn Mỹ Hạnh (21522820)

BÁO CÁO TIẾN ĐỘ THỰC HIỆN ĐỒ ÁN MÔN HỌC

Tên đồ án: Angiographic coronary heart disease prediction using Machine Learning
(Dự đoán bệnh nhân mắc bệnh động mạch vành sử dụng mô hình máy học)

Bối cảnh:

- Coronary heart disease (còn được gọi là bệnh mạch vành) là tình trạng bệnh lý mà lớp trầm tích mỡ, cholesterol và các tạp chất khác bám vào thành hạch của động mạch vành (động mạch nuôi tim), hạn chế lưu lượng máu đến tim từ đó gây ra các triệu chứng như đau thắt ngực và khó thở, dẫn đến các vấn đề nghiêm trọng như cơn đau tim, tim đập nhanh, suy tim hoặc khả năng gây chết người.

- Bệnh mạch vành là dạng bệnh tim mạch phổ biến nhất ở người lớn. Theo Tổ chức Y tế Thế giới (WHO), bệnh lý này là nguyên nhân gây tử vong hàng đầu trên toàn cầu, chiếm khoảng 18 triệu trường hợp tử vong mỗi năm. Riêng ở Việt Nam, tính đến năm 2017, số lượng bệnh nhân mắc bệnh mạch vành đã tăng gấp đôi so với năm 1990, đạt gần 1 triệu trường hợp hàng năm.

Phát biểu bài toán:

- Với bối cảnh trên, việc áp dụng các mô hình máy học để dự đoán bệnh nhân có mắc bệnh mạch vành hay không thông qua các dữ liệu y học thu thập được sẽ giúp cải thiện quá trình điều trị bệnh lý này.

- Nhóm sử dụng bộ dataset từ đường link:
https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

Bộ dữ liệu:

- Bộ dữ liệu gồm 11 thuộc tính là gồm các dữ liệu chung như tuổi, giới tính và các dữ liệu y khoa như: bệnh lý đau ngực, huyết áp, cholesterol, … và nhãn là loại nhị phân với 2 trường hợp là có bệnh hoặc không. Bộ dataset đã được đánh nhãn sẵn, nên đây là bài toán Supervised Learning.

Tập dữ liệu Heart Disease của UCI, là một bộ dữ liệu được thu thập số liệu trực tiếp từ bệnh nhân đang trong quá trình điều trị bệnh hở mạch vành bằng cách chụp mạch vành bằng tia X (Angiography) từ những viện, trung tâm liên quan đến lĩnh vực y học lớn trên thế giới. Tập dữ liệu Heart Disease của UIC đã được xây dựng từ năm 1988, được tổng hợp bởi 4 nguồn cơ sở dữ liệu khác nhau:

- Hungarian Institute of Cardiology. Budapest: Andras Janosi, M.D.
- V.A. Medical Center, Long Beach and Cleveland Clinic Foundation: Robert Detrano, M.D., Ph.D.
- University Hospital, Zurich, Switzerland: William Steinbrunn, M.D.
- University Hospital, Basel, Switzerland: Matthias Pfisterer, M.D.

Về bộ dữ liệu nhóm thu thập được trên Kaggle, bộ dữ liệu này được tạo bằng cách kết hợp các bộ dữ liệu khác nhau đã có sẵn một cách độc lập. Bộ dữ liệu là sự tổng hợp của 5 nguồn dữ liệu (4 trong đó là từ UCI dataset) với hơn 11 đặc trưng phổ biến khiến cho nó trở thành bộ dữ liệu bệnh tim phục vụ cho mục đích nghiên cứu, cũng như là cho các phương pháp dự đoán bệnh tim kết hợp với mô hình học máy. Năm nguồn dữ liệu được thu thập gồm:

- Cleveland: 303 quan sát
- Hungarian: 294 quan sát
- Switzerland: 123 quan sát
- Long Beach VA: 200 quan sát
- Stalog (Heart) Data Set: 270 quan sát
  - <u>Trong đó:</u>
    Tổng số: 1190 quan sát
    Bị trùng lặp: 272 quan sát
    Còn lại: 918 quan sát

Kế hoạch thực hiện:

- Nhóm gồm 3 thành viên nên mỗi người sẽ áp dụng một mô hình máy học để giải quyết bài toán, cụ thể là:

- Lê Viết Lâm Quang: Adaptive Boost
- Hồ Thanh Tịnh: Artificial Neural Network (ANN)
- Nguyễn Mỹ Hạnh: Logistic Regression

- Phương pháp đánh giá:

- Metrics gồm: accuracy, f1-score, precision, recall
- Dùng phương pháp K-Fold Cross Validation để chia dữ liệu thành các tập train, test khác nhau (k = 5, test size = 30%). Từ đó kết quả cần tìm là giá trị trung bình của các metrics trên sau 5 lần chạy.

- Mốc thời gian:

- 20/5/2023: Thời hạn hoàn thành nội dung mô hình mỗi thành viên chọn (gồm nội dung lý thuyết và code hoàn chỉnh phần mình).
- 21/5/2023 – 10/6/2023:
    - Tổng hợp các phương pháp thành một bài báo cáo hoàn chỉnh, với phân công như sau:
        - Phần abstract, phần giới thiệu bài toán, phần tổng kết, edit lại file .docx hoàn chỉnh. (Tịnh)
        - Phân tích và xử lý bộ dữ liệu. (Quang)
    - Làm slide thuyết trình. (Hạnh)