TRƯỜNG ĐẠI HỌC THỦY LỢi

**KHOA CÔNG NGHỆ THÔNG TIN**

BÀI TẬP LỚN

HỌC PHẦN: HỌC MÁY

# ĐỀ TÀI 5: CART VÀ ID3

Giáo viên hướng dẫn: Nguyễn Huy Đức

Sinh viên/nhóm sinh viên thực hiện:

* Nguyễn Xuân Ngọc, lớp 62TH
* Bùi Thế Bảo, lớp 62TH
* Trần Trung Hiếu, lớp 62TH

**Hà Nội, năm 2022**

# Phần 1: Tổng quan bài toán

## 1. Giới thiệu về học máy

* Lịch sử của machine learning:
* Machine learning là thuật ngữ được đặt bởi Arthur Samuel vào năm 1959. Samuel là một IBMer người Mỹ kiêm nhà tiên phong trong lĩnh vực trí tuệ nhân tạo và máy tính chơi game. Năm 1960, thuật ngữ học máy phổ biến hơn thông qua cuốn sách của Nilsson, nội dung đề cập đến việc phân loại máy học.
* Vai trò của machine learning:
* Học máy là một ứng dụng của trí tuệ nhân tạo (AI) cung cấp cho các hệ thống khả năng tự động học hỏi và cải thiện từ kinh nghiệm mà không cần lập trình rõ ràng. Học máy tập trung vào việc phát triển các chương trình máy tính có thể truy cập dữ liệu và sử dụng nó để tự học
* Quá trình học bắt đầu bằng các quan sát hoặc dữ liệu. Ví dụ, để tìm kiếm các mẫu trong dữ liệu và đưa ra quyết định tốt hơn trong tương lai dựa trên các ví dụ mà chúng tôi cung cấp. Mục đích chính là cho phép các máy tính tự động học mà không cần sự can thiệp hay trợ giúp của con người và điều chỉnh các hành động tương ứng.
* Học có giám sát
* Ưu điểm:
* Học tập có giám sát cho phép thu thập dữ liệu và tạo ra dữ liệu đầu ra từ những kinh nghiệm trước đó.
* Giúp tối ưu hóa các tiêu chí hiệu suất với sự trợ giúp của kinh nghiệm.
* ML có giám sát giúp giải quyết nhiều loại vấn đề tính toán trong thế giới thực.
* Nhược điểm:
* Phân loại dữ liệu lớn có thể là một thách thức.
* Đào tạo cho việc học có giám sát cần rất nhiều thời gian tính toán, vì vậy, nó đòi hỏi rất nhiều thời gian.
* Học không có giám sát
* Ưu điểm:
* Nó cho phép mô hình tự hoạt động để phát hiện ra các mẫu và thông tin mà trước đó không bị phát hiện. Nó chủ yếu xử lý dữ liệu không có nhãn
* Nhược điểm
* Phương pháp tính toán phức tạp
* Kém chính xác

## 2. Trình bày phương pháp học máy được sử dụng trong bài tập lớn

- Input: age,anaemia,creatinine\_phosphokinase,diabetes,ejection\_fraction,high\_blood\_pressure,platelets,serum\_creatinine,serum\_sodium,sex,smoking,time

- Output: DEATH\_EVENT.

* Cách thực hiện:
* Đọc dữ liệu
* Tách cột trong dữ liệu đã cho
* Chia dữ liệu 70% train, 30% test
* Áp dụng thuật toán Classification and Regression Tree (CART)(Thư viện) -> vẽ cây -> tính độ chính xác của thuật toán CART
* Áp dụng thuật toán ID3(Thư viện)-> vẽ cây -> tính độ chính xác của thuật toán ID3

## 3.Trình bày bài toán

* Mô tả bài toán: Dựa vào input và output để vẽ cây, tính độ chính xác của thuật toán CART và ID3

- Input: age,anaemia,creatinine\_phosphokinase,diabetes,ejection\_fraction,high\_blood\_pressure,platelets,serum\_creatinine,serum\_sodium,sex,smoking,time

- Output: DEATH\_EVENT.

# Phần 2: Thực nghiệm

## 1. Trình bày bài toán

*-* Dữ liệu gồm những chiều thông tin:

* Input:

+ age: Tuổi

+ anaemia: giảm hồng cầu hoặc hemoglobin (boolean)

+ creatinine\_phosphokinase: mức độ của enzym CPK trong máu (mcg/L)

+ diabetes: là bệnh nhân bị tiểu đường (boolean)

+ ejection\_fraction: phần trăm máu rời khỏi tim mỗi lần co bóp (%)

+ high\_blood\_pressure: là bệnh nhân bị tăng huyết áp (boolean)

+ platelets: tiểu cầu trong máu (kiloplat tiểu cầu / mL)

+ serum\_creatinine: mức độ creatinine huyết thanh trong máu (mg / dL)

+ serum sodium: mức natri huyết thanh trong máu (mEq / L)

+ sex: nam hoặc nữ (boolean)

+ smoking: hút thuốc lá (boolean)

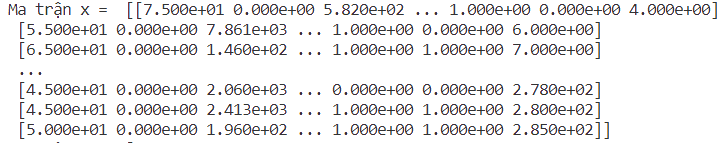
+ time: thời gian theo dõi (ngày)

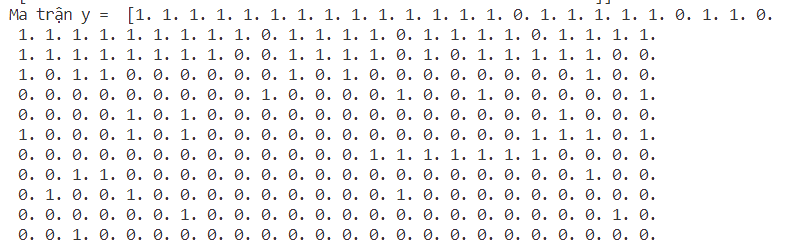
* Output:

+ DEATH\_EVENT: nếu bệnh nhân qua đời trong thời gian theo dõi (boolean)

* Mô tả ma trận dữ liệu (X), nhãn lớp (Y):

+Ma trận(X)



+Ma trận(Y)

* Chia tập dữ liệu thành 2 phần:
* Training data dùng để huấn luyện mô hình(70%).
* Test data dùng để kiểm tra sự phù hợp của mô hình(30%).

## 2. Mô tả cách giải bài toán bằng phương pháp học máy

* Bước chung cho 2 thuật toán ID3 và CART:
* Đọc dữ liệu liệu từ file csv
* Chia dữ liệu thành 2 phần
* Chia dữ liệu thành 70% training và 30% test
* Thêm thư viện vẽ
* Thuật toán CART:
* Áp dụng thuật toán CART
* Hiển thị cây thuật toán CART
* Tính độ chính xác của thuật toán CART
* Thuật toán ID3:
* Áp dụng thuật toán ID3
* Hiển thị cây thuật toán ID3
* Tính độ chính xác của thuật toán ID3

## 3. Đánh giá mô hình

### 3.1 Dùng tập test data để đánh giá mô hình theo độ đo của F1-score

1. Đối với ID3
   1. ******
2. Đối với thuật toán CART

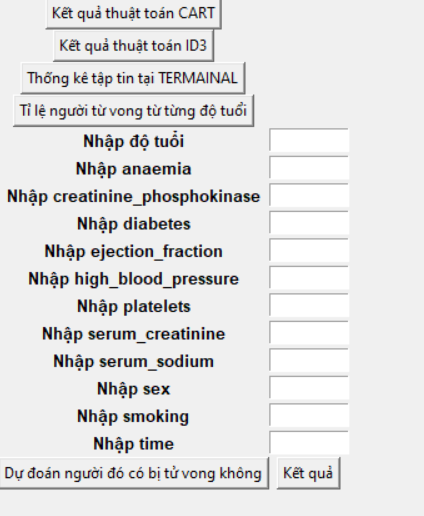
### 

### 3.2 Dùng tập test data để đánh giá mô hình theo Confusion matrix

1. Đối với ID3
   1. 
2. Đối với CART
   1. 

* Giải thích
  + Hàng thứ nhất là số lượng người sống dự đoán đúng
  + Hàng thứ hai là số lượng người chết dự đoán đúng
  + Cột thứ nhất là số lượng người sống thực tế
  + Cột thứ hai là số lượng người chết thực tế

## 4. Mô tả các chức năng của chương trình

* Chương trình có 5 nút sự kiện button(Click):
* button1: Vẽ cây của thuật toán CART và độ chính xác của thuật toán CART
* button2: Vẽ cây của thuật toán ID3 và độ chính xác của thuật toán ID3
* button3: Thống kế số liệu từ file csv
* button4: Thống kê tỷ lệ người tử vong ở từng độ tuổi
* button5: Khi click vào “Dự đoán người đố có vị tử vong hay không ” xẽ mở ra giao điện mới người nhập
* Giao điện khi chạy chương trình
* 

## Kết luận

* + Tóm lược các nội dung chính mà bài tập lớn làm được :
* Đọc file dữ liệu csv
* Chia dữ liệu ra làm 2 phần
* Vẽ được cây của thuật toán CART và ID3
* Tính được độ chính xác của thuật toán ID3 bằng F1-score và Confusion matrix
* Tính được độ chính xác của thuật toán CART bằng F1-score và Confusion matrix
* Dự đoán được từ dữ liệu người dùng nhập vào

## Tài liệu tham khảo

* + Các tài liệu được tham khảo trong báo cáo:
* [Dữ liệu](https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records)
* [Tài liệu tham khảo](https://machinelearningcoban.com/)