TRƯỜNG ĐẠI HỌC THỦY LỢI

**KHOA CÔNG NGHỆ THÔNG TIN**

A blue and white logo

Description automatically generated with low confidence

**BÀI TẬP LỚN**

HỌC PHẦN: HỌC MÁY

**ĐỀ TÀI: DỰ ĐOÁN BỆNH UNG THƯ VÚ**

Giáo viên hướng dẫn: Nguyễn Huy Đức

Sinh viên/nhóm sinh viên thực hiện:

1. Trần Thanh Thảo, lớp 62TH-NB

2. Vũ Thanh Sơn, lớp 62TH-NB

3. Nguyễn Hoàng Sơn, lớp 62TH-NB

**Hà Nội, năm 2022**

**MỤC LỤC**

[**Phần 1: Tổng quan** 3](#_Toc117444883)

[***1. Giới thiệu về học máy*** 3](#_Toc117444884)

[**1.1 - Lịch sử và vai trò của machine learning** 3](#_Toc117444885)

[**1.2 - Ưu điểm và hạn chế của Học có giám sát và không giám sát** 4](#_Toc117444886)

[***2. Trình bày phương pháp học máy được sử dụng trong bài tập lớn*** 5](#_Toc117444887)

[**2.1 Phương pháp k-fold cross validation** 5](#_Toc117444888)

[**2.2 Phương pháp Perceptron Learning Agorithm** 6](#_Toc117444889)

[***3.*** ***Trình bày bài toán*** 7](#_Toc117444890)

[**Phần 2: Thực nghiệm** 8](#_Toc117444891)

[***1. Mô tả tập dữ liệu*** 8](#_Toc117444892)

[***2. Mô tả cách giải bài toán bằng phương pháp học máy*** 9](#_Toc117444893)

[***3. Đánh giá mô hình*** 9](#_Toc117444894)

[***4. Mô tả các chức năng của chương trình*** 9](#_Toc117444895)

[***Kết luận*** 10](#_Toc117444896)

[**Tài liệu tham khảo** 10](#_Toc117444897)

# **Phần 1: Tổng quan**

## ***1. Giới thiệu về học máy***

### **1.1 - Lịch sử và vai trò của machine learning**

* Khái niệm machine learning

Machine learning (tạm dịch: học máy) có khởi nguồn từ lĩnh vực trí tuệ nhân tạo (AI). Đây là giải pháp có khả năng khiến các ứng dụng, phần mềm trở nên thông minh hơn mà không cần thực hiện bất kỳ mã hóa nào. Thông qua dữ liệu đầu vào, Machine learning có thể dự đoán các giá trị đầu ra mới.

* Lịch sử hình thành 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
* Vai trò của học máy là tập trung phát triển chương trình máy tính truy cập dữ liệu và sử dụng chúng để tự học. Mục tiêu chính của hoạt động này là giúp máy tính tự hoạt động mà không cần sự hỗ trợ của con người. Thậm chí, nó đã len lỏi vào hầu hết các lĩnh vực của đời sống mà có thể chúng ta không nhận ra.
* Ứng dụng nổi bật:
* Xe tự hành của Google và Tesla
* hệ thống tự tag khuôn mặt trong ảnh của Facebook
* trợ lý ảo Siri của Apple
* hệ thống gợi ý sản phẩm của Amazon
* hệ thống gợi ý phim của Netflix
* hệ thống dịch đa ngôn ngữ Google Translate v.v.v
* Machine learning được xem là một phần quan trọng thuộc lĩnh vực khoa học dữ liệu. Chúng sử dụng các phương pháp thống kê, thuật toán để phân loại, dự đoán và khám phá những thông tin quan trọng của dữ liệu. Nhờ những chuỗi thông tin này, người dùng nhanh chóng đưa ra các quyết định trong các hoạt động kinh doanh của mình. Vì vậy, Machine learning chính là giải pháp lý tưởng giúp doanh nghiệp tác động đến chỉ số tăng trưởng doanh thu.

### **1.2 - Ưu điểm và hạn chế của Học có giám sát và không giám sát**

|  |  |  |
| --- | --- | --- |
|  | ***Học có giám sát*** | ***Học không giám sát*** |
| **Ưu điểm** | * 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. * giúp giải quyết nhiều loại vấn đề tính toán trong thế giới thực. * Tập dữ liệu **có gán nhãn** * **Độ chính xác cao** * **Phương pháp đơn giản** | * có thể được coi là **tự học** trong đó thuật toán có thể tìm thấy các mẫu chưa biết trước đây trong tập dữ liệu không có bất kỳ loại nhãn nào. * giúp mô hình hóa các hàm mật độ xác suất, tìm kiếm sự bất thường trong dữ liệu và hơn thế nữa. |
| **Hạn chế** | * 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. | * Không có cách nào để có được cách thức hoặc phương pháp dữ liệu được sắp xếp vì tập dữ liệu **không được gắn nhãn**. * Có thể **kém chính xác hơn** vì dữ liệu đầu vào không được con người biết và dán nhãn bởi con người chế tạo máy làm việc đó. * Thông tin thu được bởi thuật toán có thể không phải lúc nào cũng tương ứng với lớp đầu ra * Tính toán **phức tạp** |

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

### **2.1 Phương pháp k-fold cross validation**

Cross validation là một phương pháp thống kê được sử dụng để ước lượng hiệu quả của các mô hình học máy. Nó thường được sử dụng để so sánh và chọn ra mô hình tốt nhất cho một bài toán. Kỹ thuật này dễ hiểu, dễ thực hiện và cho ra các ước lượng tin cậy hơn so với các phương pháp khác. Trong bài này ta sẽ tìm hiểu về kỹ thuật cross validation với k-fold.

* *K-fold cross validation*

Tham số quan trọng trong kỹ thuật này là k, đại diện cho số nhóm mà dữ liệu sẽ được chia ra. Vì lý do đó, nó được mang tên k-fold cross-validation. Khi giá trị của k được lựa chọn, người ta sử dụng trực tiếp giá trị đó trong tên của phương pháp đánh giá. Ví dụ với k=10, phương pháp sẽ mang tên 10-fold cross-validation.

Kỹ thuật này thường bao gồm các bước như sau:

1. Xáo trộn dataset một cách ngẫu nhiên
2. Chia dataset thành k nhóm
3. Với mỗi nhóm:

3.1 Sử dụng nhóm hiện tại để đánh giá hiệu quả mô hình

3.2 Các nhóm còn lại được sử dụng để huấn luyện mô hình

3.3 Huấn luyện mô hình

3.4 Đánh giá và lưu lại kết quả

4. Lựa chọn mô hình có kết quả tốt nhất

*- Cấu hình k*

Giá trị k là thông số quan trọng để có thể đánh giá chính xác mô hình, vậy thì lựa chọn thông số này như thế nào?

Ba chiến thuật phổ biến để lựa chọn k:

Đại diện: Giá trị của k được chọn để mỗi tập train/test đủ lớn, có thể đại diện về mặt thống kê cho dataset chứa nó.

k=10: Giá trị của k được gán cố định bằng 10, một giá trị thường được sử dụng và được chứng minh là cho sai số nhỏ, phương sai thấp (thông qua thực nghiệm).

k=n: Giá trị của k được gán cố định bằng n, với n là kích thước của dataset, như vậy mỗi mãu sẽ được sử dụng để đánh giá mô hình một lần. Cách tiếp cận này còn có tên leave-one-out cross-validation.

Giá trị k=10 là một cấu hình rất phổ biến. Nên sử dụng giá trị này nếu như gặp khó khăn trong việc lựa chọn một giá trị thích hợp cho bài toán. Ngoài ra, cũng nên chọn giá trị của k sao cho có thể chia đều các mẫu vào các nhóm.

### **2.2 Phương pháp Perceptron Learning Agorithm**

* Ý tưởng cơ bản của thuật toán:

Xuất phát từ một nghiệm dự đoán nào đó,

Qua mỗi vòng lặp, nghiệm sẽ được cập nhật tới một ví trí tốt hơn

Cập nhật dựa trên việc giảm giá trị của một hàm mất mát nào đó.

Ta có: Ma trận X = là ma trận chứa các điểm dữ liệu

Ma trận Y =

nếu thuộc class 1

nếu thuộc class 2

Giả sử, tại một thời điểm, ta tìm được đường boundary là đường phẳng có phương trình:

Hay với W =

X =

Ta thấy, nếu w là một nghiệm của bài toán Perceptron, với một điểm dữ liệu mới x chưa được gán nhãn, ta có thể xác định class của nó bằng phép toán như sau:

trong đó, sgn là hàm xác định dấu, với giả sử rằng sgn(0)=1.

* Xét hàm mất mát sau đây:

Với là tập hợp các điểm bị misclassifed (tập hợp này thay đổi theo w).

Nhận xét rằng khi một điểm misclassified nằm càng xa boundary thì giá trị sẽ càng lớn, nghĩa là sự sai lệch càng lớn. Giá trị nhỏ nhất của hàm mất mát này cũng bằng 0 nếu không có điểm nào bị misclassifed.

Với một điểm dữ liệu bị misclassified, hàm mất mát trở thành:

Đạo hàm tương ứng:

Vậy quy tắc cập nhật là:

với η là learning rate được chọn bằng 1. Ta có một quy tắc cập nhật rất gọn là: .

Nói cách khác, với mỗi điểm xi bị misclassifed, ta chỉ cần nhân điểm đó với nhãn yi của nó, lấy kết quả cộng vào w ta sẽ được w mới.

* Tóm tắt thuật toán Perceptron:
* Chọn ngẫu nhiên một vector hệ số w với các phần tử gần 0.
* Duyệt ngẫu nhiên qua từng điểm dữ liệu
* Nếu được phân lớp đúng, chúng ta không làm gì
* Ngược lại, cập nhật w theo công thức: .
* Kiểm tra xem có bao nhiêu điểm bị misclassifed. Nếu không còn điểm nào, dừng thuật toán. Nếu còn, quay lại bước 2.

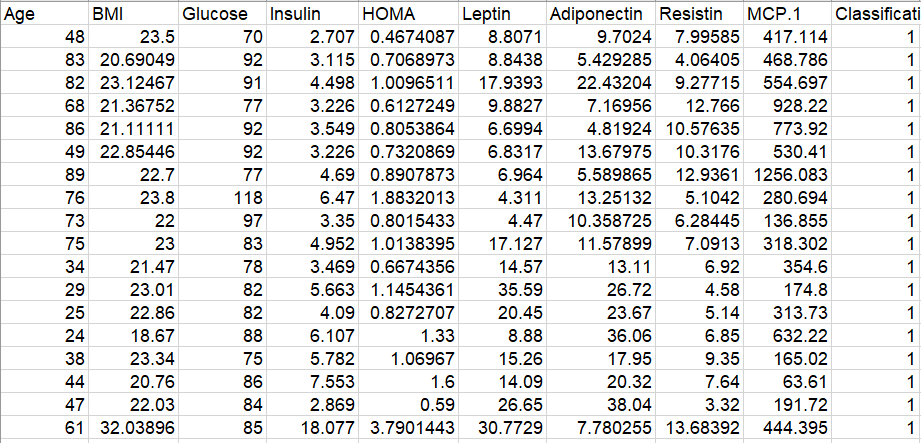
## ***Trình bày bài toán***

* Mô tả bài toán
* Tên bài toán: Dự đoán bệnh ung thư vú
* Mô tả: Cho biết sự hiện diễn có hoặc không có của ung thư vú. Các yếu tố dự báo là dữ liệu nhân trắc học và các thông số có thể được thu thập trong phân tích máu định kỳ.
* Input: tập dữ liệu gồm các thuộc tính và đã được gán nhãn
* Ouput: kết quả dự đoán một người có mắc bệnh ung thư vú hay không
* Tóm tắt công việc thực hiện của bài toán:
* Đọc dữ liệu từ file
* Chuẩn hóa dữ liệu
* Chia dữ liệu theo tỷ lệ 70% train, 30% test, không xáo trộn.
* Sử dụng các thuật toán để giải quyết bài toán:
  + Perceptron
  + K –fold cross validation
* Đánh giá mô hình dựa trên các độ đo: accuracy, precission, recall, F1-score

# **Phần 2: Thực nghiệm**

# ***1.Mô tả tập dữ liệu***

* Dữ liệu có 117 vector gồm các thuộc tính và nhãn



* Mô tả ma trận dữ liệu (X):
* Gồm 117 dòng và 9 cột
* Mỗi cột tương ứng với một thuộc tính lần lượt là: Age, BMI, Glucose, Insulin, HOMA, Leptin, adiponectin, resistin, MCP-1

|  |  |
| --- | --- |
| **Tên thuộc tính** | **Ý nghĩa** |
| Age | Tuổi tác |
| BMI | chỉ số khối cơ thể (kg / m2) |
| Glucose | lượng đường glucozo (mg / dL) |
| Insulin | hormone từ các tế bào đảo tụy ở tuyến tụy tiết ra |
| HOMA | chỉ số chỉ ra mức độ kháng insulin |
| Leptin | hormone được sản xuất bởi các tế bào mỡ của cơ thể |
| Adiponectin | Nồng độ Adiponectin (µg / mL) |
| Resistin | Hormone Từ Mô Mỡ (ng / mL) |
| MCP-1 | chất điều hòa sinh trưởng có vai trò ức chế quá trình sinh tổng hợp của ethylene. (pg / dL) |

* Mô tả nhãn lớp (Y): gồm 117 dòng và 1 cột tương ứng là Classification

|  |  |
| --- | --- |
| **Tên nhãn** | **Ý nghĩa** |
| Classification | 1: Người khỏe mạnh  2: mắc bệnh ung thư vú |

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

* Phương pháp CROSS-VALIDATION
* Bước 1: Chia toàn bộ tập dữ liệu thành k phần (phương pháp k-fold cross validation).
* Bước 2: Chọn ngẫu nhiên k-1 phần làm training data, 1 phần còn lại làm test data. Sử dụng phương pháp học máy đã lựa chọn trên tập training data và test data để xây dựng và đánh giá mô hình. Bước 2 này được làm k lần.
* Bước 3: Chọn mô hình có (train error + validation error) là nhỏ nhất.

Trong 3 lần trên, lần nào có (train error + validation error) nhỏ nhất thì mô hình huấn luyện của lần đó được chọn làm mô hình dự đoán cho dữ liệu mới.

* Dùng phương pháp Perceptron để xây dựng mô hình trên tập dữ liệu mới

## ***Đánh giá mô hình***

Dùng tập test data để đánh giá chất lượng của mô hình được lựa chọn theo các độ đo:

- Accuracy: 60%

- Precision: ~ 61%

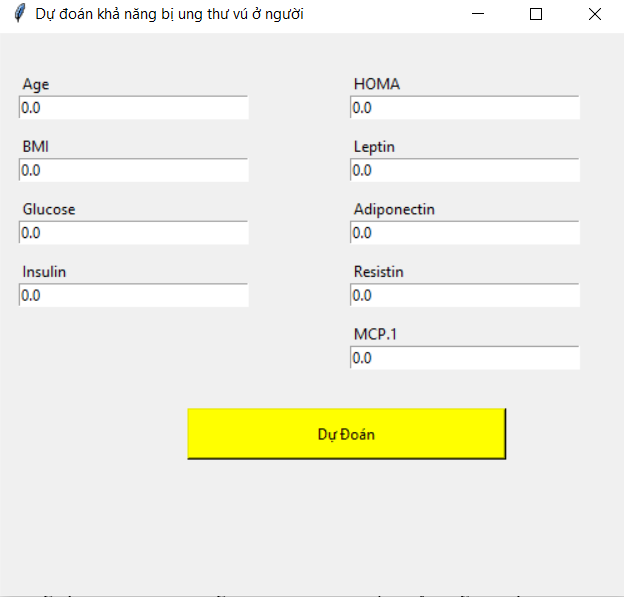
- Recall: 60%

- F1 – scored: 57%

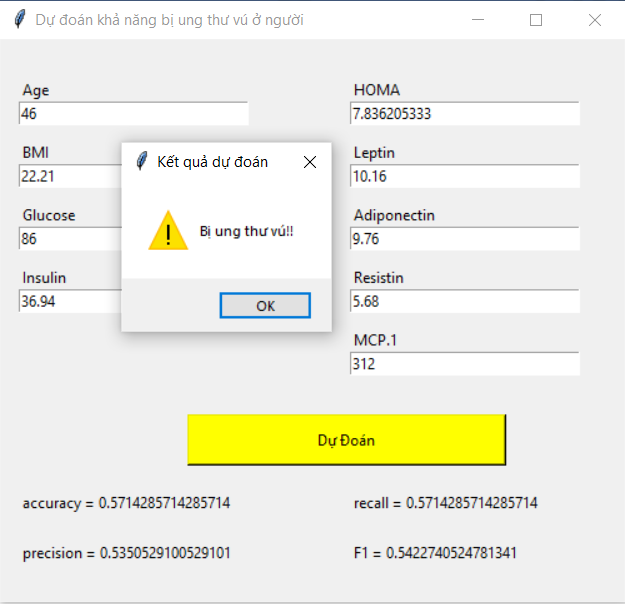
## 

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

- Hình ảnh form giao diện dự đoán:



- Hình ảnh giao diện sau khi dự đoán:



## ***Kết luận***

Qua bài làm về đề tài: Ứng dụng phương pháp Perceptron Learning Agorithm và K-fold cross vallidation để dự đoán người mắc bệnh ung thư vú, nhóm chúng em hi vọng mô hình này có thể phát triển hơn nữa. Đồng thời thông qua bài làm giúp nhóm em có những kiến thức về môn học máy cũng như áp dụng kiến thức đó vào phân tích các bài toán thực tế.

* Tìm ra tỷ lệ dự đoán đúng, tỉ lệ dự đoán sai.
* Cho ra kết quả dự đoán dữ liệu về người có khả năng mắc bệnh ung thư vú thông qua các chỉ số

Mặc dù đã có nhiều cố gắng trong suốt quá trình làm bài tập, nhưng kiến thức còn hạn chế, cũng như chưa có kinh nghiệm trong thực tế nên không thể tránh khỏi những sai sót. Vì vậy, nhóm em mong thầy, cô góp ý để có thể chỉnh sửa, hoàn thiện bài làm tốt hơn.

## **Tài liệu tham khảo**

1. Blog của Tino group: <https://tino.org/vi/machine-learning-la-gi/>
2. Blog: <https://thatim.com/phuong-phap-hoc-khong-giam-sat-thuong-duoc-su-dung-trong-bai-toan-nao>
3. Machine learning cơ bản: <https://machinelearningcoban.com/>
4. UCI: <https://archive.ics.uci.edu/ml/index.php>
5. Slide bài giảng