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: Tìm hiểu, triển khai thuật toán ID3 và**

**phân tích thành phần chính PCA cho bài toán**

**“dự đoán chất lượng sữa”.**

Giáo viên hướng dẫn: Nguyễn Thị Kim Ngân

Nhóm sinh viên thực hiện: Nhóm 6

1. Nguyễn Thị Mây, lớp 62TH1

2. Nguyễn Thị Tươi, lớp 62TH1

3. Nguyễn Hồng Thương, lớp 62TH1

**Hà Nội, ngày 25 tháng 10 năm 2022**

Mục lục

[**Phần 1: Tổng quan** 3](#_Toc117693595)

[1. Giới thiệu về học máy 3](#_Toc117693596)

[a) Lịch sử và vai trò của machine learning. 3](#_Toc117693597)

[b) Ưu điểm và hạn chế của Học có giám sát và không giám sát. 3](#_Toc117693598)

[2. Trình bày phương pháp học máy được sử dụng trong bài tập lớn 4](#_Toc117693599)

[a) Thuật toán ID3. 4](#_Toc117693600)

[b) Phương pháp thành phần chính (PCA) 5](#_Toc117693601)

[3.Trình bày bài toán. 6](#_Toc117693602)

[**Phần 2: Thực nghiệm** 6](#_Toc117693603)

[1. Mô tả tập dữ liệu 6](#_Toc117693604)

[2. Mô tả cách giải bài toán bằng phương pháp học máy 8](#_Toc117693605)

[3. Đánh giá mô hình 8](#_Toc117693606)

[4. Mô tả các chức năng của chương trình 8](#_Toc117693607)

[5. Kết luận 9](#_Toc117693608)

[6. Tài liệu tham khảo 9](#_Toc117693609)

# Phần 1: Tổng quan

## 1. Giới thiệu về học máy

1. Lịch sử và vai trò của machine learning.

* Machine learning là 1 ứng dụng của trí tuệ nhân tạo cung cấp cho các hệ thống khả năng tự học dựa trên dữ liệu mà không cần lập trình rõ ràng, máy có thể tự dự đoán hoặc đưa ra quyết định mà không cần lập trình.
* Các ứng dụng phổ biến của machine learning:
  + Tự động phân loại
  + Ứng dụng trong các mạng xã hội.
  + Nhận diện hình ảnh.
  + Chăm sóc sức khỏe.
  + Tài chính.

1. Ưu điểm và hạn chế của Học có giám sát và không giám sát.

|  |  |  |
| --- | --- | --- |
|  | **Supervised learning**  **(học có giám sát)** | **Unsupervised learning**  **(học không có giám sát)** |
| Khái niệm | Là cách huấn luyện một mô hình trong đó dữ liệu học có đầu vào và đầu ra tương ứng đầu vào đó. Mô hình được huấn luyện bằng cách giảm thiểu sai số lỗi **(loss)** của các dự đoán tại các vòng lặp huấn luyện. | Huấn luyện một mô hình trong đó dữ liệu học chỉ bao gồm đầu vào mà không có đầu ra. Mô hình sẽ được huấn luyện cách để tìm cấu trúc hoặc mối quan hệ giữa các đầu vào. |
| Ưu điểm | - Biết được đầu ra của 1 dữ liệu mới dựa trên các mẫu dữ liệu đã được gán nhãn từ trước.  -Giải quyết các vấn đề tính toán trong thế giới thực.  - Giúp tối ưu hóa các tiêu chí hiệu suất với sự trợ giúp của kinh nghiệm. | - xử lý dữ liệu không có nhãn. |
| Nhược điểm | - chỉ tốt với dữ liệu nhỏ.  - Tốn nhiều thời gian tính toán.  - Ranh giới quyết định có thể bị giới hạn nếu tập huấn luyện của bạn không có ví dụ mà bạn muốn có trong một lớp. | - không biết câu trả lời chính xác cho mỗi dữ liệu đầu vào.  - Tính toán phức tạp.  - Ít chính xác. |

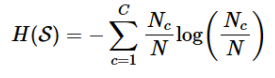
## 2. Trình bày phương pháp học máy được sử dụng trong bài tập lớn

### a) Thuật toán ID3.

* Mục đích của phương pháp: Tạo ra cây quyết định bằng cách sử dụng hàm “entropy” làm thước đo.
* Input: Là một tập dữ liệu đã được gán nhãn.
* Output: là cây quyết định
* Cách thực hiện:

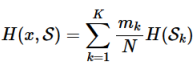
**B1:** tính entropy trên toàn bộ tập S

* + H(S) : entropy trên toàn bộ tập dữ liệu S
  + S: tập dữ liệu huấn luyện(gồm cả x và y)
  + N: tổng số mẫu dữ liệu huấn luyện
  + Nc: số các mẫu dữ liệu trong tập S đc gán nhãn c
  + C: tổng số nhãn khác nhau trong tập S
  + c: chỉ số của nhãn dữ liệu



**B2:** Tính entropy của thuộc tính x trên tập S

* + x: là thuộc tính đang xét
  + K: số giá trị của thuộc tính đó
  + mk: số mẫu có thuộc tính x và có giá trị là k
  + Sk: tập các mẫu sao cho thuộc tính x có giá trị là k
  + H(Sk): entropy trên tập dữ liệu trong S có thuộc tính là x và có giá trị k



**B3:** tính *information gain* trên thuộc tính x



* + Thay vì tìm thuộc tính có information gain cao nhất thì ta tìm thuộc tính có entropy nhỏ nhất

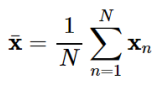


**B4:** Chia dữ liệu vào các nút con tương ứng với các giá trị của thuộc tính được chọn, thực hiện lặp tìm ra thuộc tính tốt nhất trong tập dữ liệu mới.

* **Điều kiện dừng:**
  + TH1: nếu tất cả các mẫu trong node con thuộc cùng 1 lớp C (entropy=0) thì node đó đc gán nhãn C
  + TH2: nếu node con là rỗng (không có dữ liệu) thì node đó sẽ đc gán = nhãn phổ biến nhất trong tập S
  + TH3: nếu không còn thuộc tính nào để phân chia (các thuộc tính đều đc xét hết) -> node lá đó sẽ đc gán = nhãn phổ biến nhất
* Đánh giá mô hình:
  + Ưu điểm: cây phân loại thực hiện tốt trong thực tế
  + Nhược điểm: không bị giới hạn, các cây riêng lẻ dễ bị quá mức.

### b) Phương pháp thành phần chính (PCA)

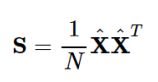
* Mục đích của phương pháp: Tìm một hệ trực chuẩn mới sao cho trong hệ này, các thành phần quan trọng nhất nằm trong K thành phần đầu tiên.
* Input: tập thuộc tính, số thành phần chính muốn giữ lại.
* Output: tập thuộc tính mới sau khi loại bỏ các thuộc tính ít quan trọng.
* Cách thực hiện:
  + B1: Tính vector kỳ vọng của toàn bộ dữ liệu:



* + B2: Trừ mỗi điểm dữ liệu đi vector kỳ vọng của toàn bộ dữ liệu:



* + B3: Tính ma trận hiệp phương sai S:



* + B4: Tính các trị riêng và vector riêng của S , sắp xếp chúng theo thứ tự giảm dần của trị riêng.
  + B5: Chọn 𝐾 vector riêng ứng với 𝐾 trị riêng lớn nhất để xây dựng ma trận 𝐔𝐾 có các cột tạo thành một hệ trực giao. 𝐾 vectors này là các thành phần chính.
  + B6: Chiếu dữ liệu ban đầu đã chuẩn hoá xuống không gian con tìm được.
  + B7: Dữ liệu mới chính là toạ độ của các điểm dữ liệu trên không gian mới.



## 3.Trình bày bài toán.

* Tên bài toán: Dự đoán chất lượng sữa
* Mục đích của bài toán: dự đoán chất lượng sữa khi biết các thông tin liên quan.
* Input:

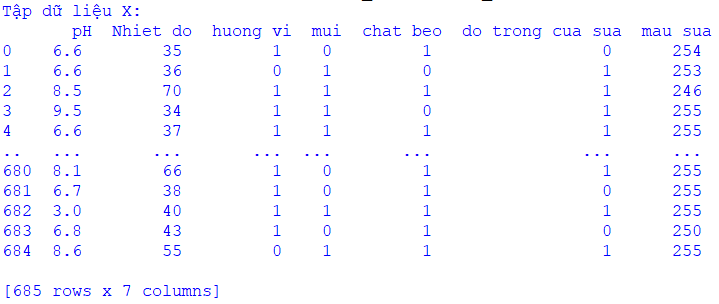
1. pH
2. Nhiệt độ
3. Hương vị
4. mùi
5. chất béo
6. Độ trong của sữa
7. Màu sữa

* Output: Chất lượng (high/low)
* Tóm tắt công việc thực hiện của bài toán:
  + Đọc tập dữ liệu của bài toán
  + Chia dữ liệu thành 2 phần: 70% cho tập train, 30% cho tập test
  + Sử dụng phương pháp phân tích thành phần chính (PCA) kết hợp với thuật toán ID3 để tìm mô hình tốt cho bài toán phân lớp.
  + Tạo giao diện form cho phép người dùng dự đoán nhãn. Trên form sử dụng textbox để nhập thông tin từ các trường. Thao tác click sẽ hiển thị kết quả dự đoán của các mô hình

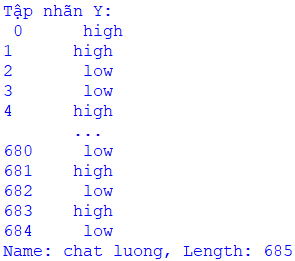
# Phần 2: Thực nghiệm

1. **Mô tả tập dữ liệu**

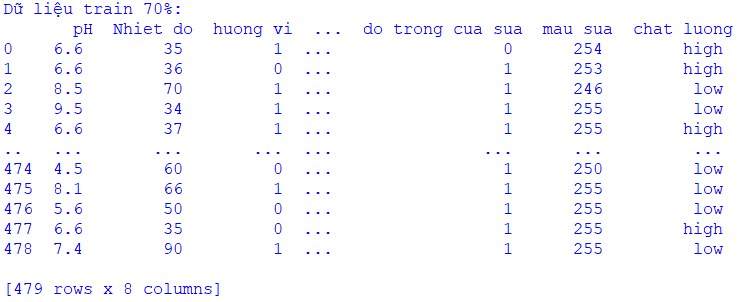
* Số lượng vector dữ liệu(số lượng mẫu dữ liệu): 685
* Mỗi vector dữ liệu gồm 7 thông tin: pH, Nhiệt độ, Hương vị, mùi, chất béo, Độ trong của sữa, Màu sữa.
* Nhãn lớp của dữ liệu là chất lượng sữa, mỗi vector dữ liệu sẽ tương ứng với 1 giá trị chất lượng sữa (high/low).
* Tập dữ liệu X [7,685] là 1 ma trận dữ liệu đầu vào với mỗi hàng của X là 1 vector dữ liệu gồm các thông tin: pH, Nhiệt độ, Hương vị, mùi, chất béo, Độ trong của sữa, Màu sữa.



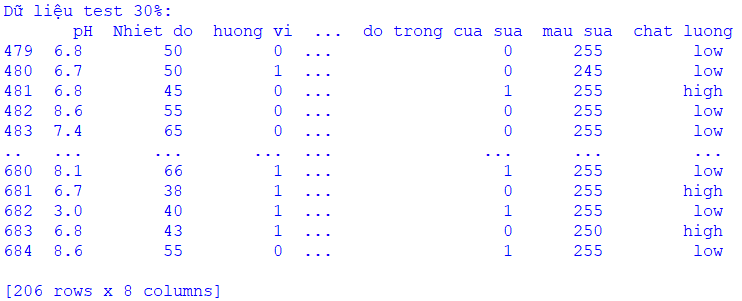
* Nhãn lớp Y là 1 vector có số dòng là 685 và có 1 cột là “chất lượng”.



* Chia tập dữ liệu thành 2 phần: 70% dùng để huấn luyện mô hình, 30% dùng để kiểm tra sự phù hợp của mô hình.
  + 70% Dữ liệu sd để huấn luyện:



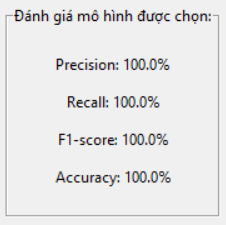
* + 30% Dữ liệu sd để kiểm tra sự phù hợp của mô hình:



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

* Bước 1: Dùng phương pháp Phân tích thành phần chính để lựa chọn tập các thuộc tính tốt nhất cho bài toán. Từ tập training data và test data ban đầu, sử dụng các thành phần chính tốt nhất đã chọn để tạo ra tập training data và test data mới.
* Bước 2: Dùng phương pháp học máy ID3 để xây dựng mô hình và đánh giá mô hình trên tập dữ liệu mới.

1. **Đánh giá mô hình**



* Mô hình tốt do có tỉ lệ dự đoán đúng là 100%.

1. **Mô tả các chức năng của chương trình**

* Chức năng của chương trình: dự đoán chất lượng sữa dựa trên các thông tin do người dùng nhập vào gồm 7 thông tin: pH, Nhiệt độ, Hương vị, mùi, chất béo, Độ trong của sữa, Màu sữa.



1. **Kết luận**

* Thực hiện bài toán dự đoán trên thuật toán Cây quyết định(ID3) và sử dụng phương pháp phân tích thành phần chính (PCA).

1. **Tài liệu tham khảo**

* <https://www.kaggle.com/datasets/yrohit199/milk-quality>
* <https://machinelearningcoban.com/2016/12/28/linearregression/>