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 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 19 tháng 10 năm 2022**

**Phần 1: Tổng quan**

***1.Mô tả 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 Support Vector Machine, sử dụng PCA kết hợp với cây quyết định 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.

2***. Phương pháp học máy***

1. **ID3**

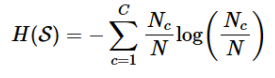
* ID3 là một thuật toán decision tree được áp dụng cho các bài toán classification mà tất cả các thuộc tính đều ở dạng categorical.
* Input: Là một tập dữ liệu đã được gán nhãn.
* Output: là cây quyết định
* Công việc thực hiện của bài toán:

+ Sử dụng hàm “entropy” làm thước đo trong khi tạo cây quyết định (decision tree) trong thuật toán ID3 (Classification and Regression Tree)

* Các bước thực hiện:

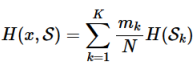
**B1:** tính entropy trên toàn bộ tập S

* + 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



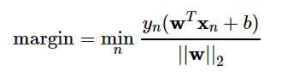
* + Thay vì tìm thuộc tính có information gain cao nhất thì ta tìm thuộc tính có entropy nhỏ nhất



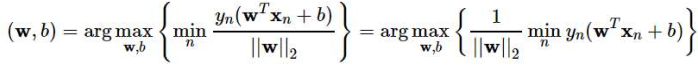
* **Điều kiện dừng:**
  + TH1: nếu tất cả các mẫu trong node con thuộc cùng 1 lớp C (entropy=0) thì node đó đc gán nhãn C
  + TH2: nếu node con là rỗng (không có dữ liệu) thì node đó sẽ đc gán = nhãn phổ biến nhất trong tập S
  + TH3: nếu không còn thuộc tính nào để phân chia (các thuộc tính đều đc xét hết) -> node lá đó sẽ đc gán = nhãn phổ biến nhất

1. **Support Vector Machine**

* Tìm ra siêu phẳng sao cho khoảng cách từ điểm gần nhất của mỗi lớp tới đường phân chia là như nhau(margin), và phải là lớn nhất.
* Input: Là một tập dữ liệu đã được gán nhãn.
* Output: là siêu phẳng sao cho margin bằng nhau và lớn nhất
* Các bước thực hiện:
  + B1: tính khoảng cách từ 1 điểm đến mặt phân chia, điểm nào có khoảng cách nhỏ nhất thì lấy làm margin.

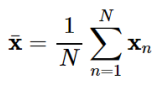


* + B2: Tìm ra siêu phẳng sao cho margin là lớn nhất <=> tìm w và b sao cho margin lớn nhất.



1. **PCA.**

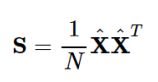
* Tìm một hệ trực chuẩn mới sao cho trong hệ này, các thành phần quan trọng nhất nằm trong K thành phần đầu tiên.
* Input: Là một tập dữ liệu có D chiều.
* Output: là 1 tập dữ liệu K chiều với K<D.
* Các bước thực hiện:
  + B1: Tính vector kỳ vọng của toàn bộ dữ liệu:



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



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



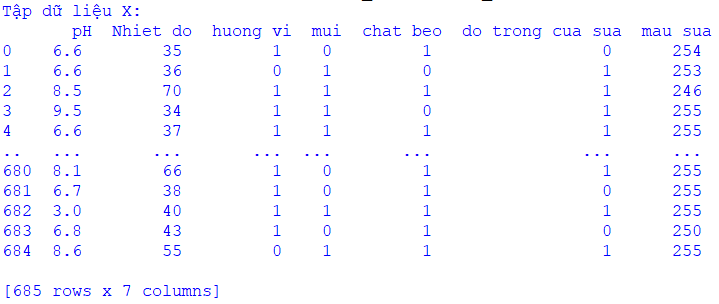
* + B4: Tính các trị riêng và vector riêng của S , sắp xếp chúng theo thứ tự giảm dần của trị riêng.
  + B5: Chọn 𝐾 vector riêng ứng với 𝐾 trị riêng lớn nhất để xây dựng ma trận 𝐔𝐾 có các cột tạo thành một hệ trực giao. 𝐾 vectors này là các thành phần chính.
  + B6: Chiếu dữ liệu ban đầu đã chuẩn hoá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



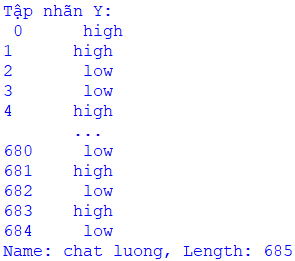
**Phần 2: Thực nghiệm**

***1.Mô tả tập dữ liệu của bài toán***

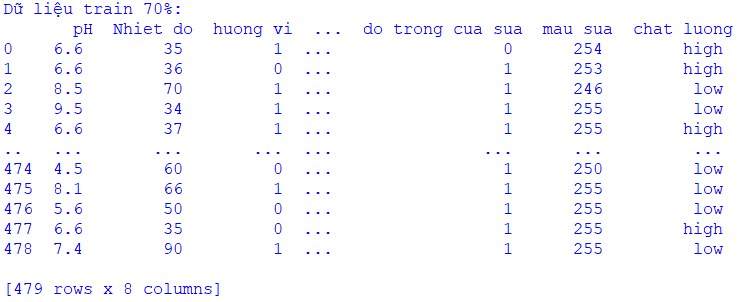
* Số lượng vector dữ liệu(số lượng mẫu dữ liệu): 685
* Mỗi vector dữ liệu gồm 7 thông tin: pH, Nhiệt độ, Hương vị, mùi, chất béo, Độ trong của sữa, Màu sữa.
* Nhãn lớp của dữ liệu là chất lượng sữa, mỗi vector dữ liệu sẽ tương ứng với 1 giá trị chất lượng sữa (high/low).
* Tập dữ liệu X [7,685] là 1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



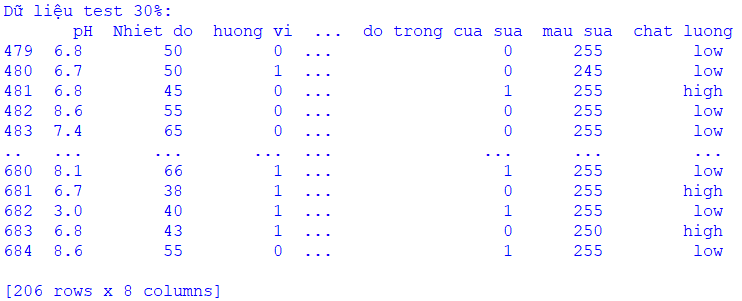
* Nhãn lớp Y là 1 vector có số dòng là 685 và có 1 cột là “chất lượng”.



* 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:

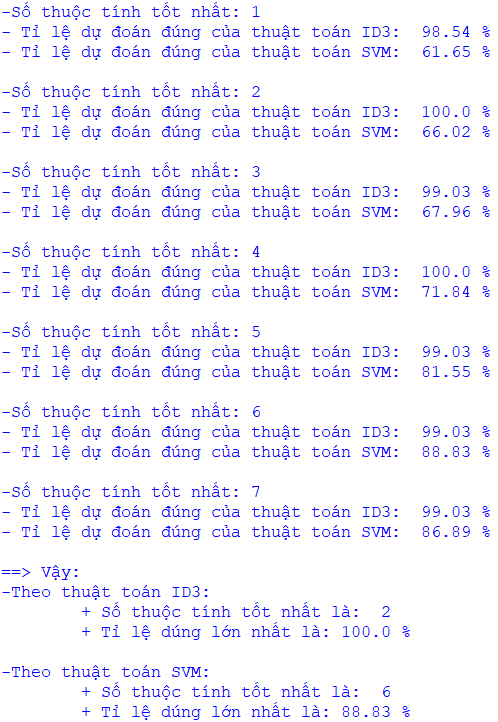


* + 30% Dữ liệu sd để kiểm tra sự phù hợp của mô hình:



***2. Phân tích kết quả của chương trình***

* So sánh chất lượng của mô hình phân lớp (tỷ lệ dự đoán đúng) khi thay đổi số lượng các thành phần chính.



**3. Kết luận**

* Thực hiện bài toán dự đoán trên thuật toán Support Vector Machine, Cây quyết định sử dụng phân tích thành phần chính (PCA)

**4. Tài liệu tham khảo**

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