TRƯỜNG ĐẠI HỌC THỦY LỢI

**KHOA CÔNG NGHỆ THÔNG TIN**

**KHOA CÔNG NGHỆ THÔNG TIN**

**BÀI TẬP LỚN**

HỌC PHẦN: HỌC MÁY

**ĐỀ TÀI: DỰ ĐOÁN MUA BẢO HIỂM DU LỊCH**

Giáo viên hướng dẫn: Kim Ngân

Nhóm 5 thực hiện:

1. Bùi Quang Huy,Lớp 62PM2

2. Lê Duy Thái,Lớp 62PM2

**Hà Nội, năm 2022**



**Phần 1: Lý thuyết**

* Trình bày lý thuyết của các thuật toán ID3 và CART.

1. **Thuật toán ID3**

Cây quyết định là 1 mô hình có giám sát, có thể được áp dụng cho cả BT phân lớp (classification) và hồi quy (regression)

1. Input: Là 1 tập DL đã được gán nhãn
2. Output: Là 1 vecto dự đoán
3. Method: Sử dụng hàm “ entropy “ để đo độ tinh khiết ( purity ) / độ vẩn đục ( impurity ) của 1 phép phân chia

* B1: Tính entropy tại root node:

H (S) = -

* B2: Tính entropy tại child node:

H (x, S) = -

* B3: Định nghĩa “ information gain “ dựa trên thuộc tính x:

G (x, S) = H (S) – H (x, S)

* Tại mỗi node, thuộc tính được chọn được xác định dựa vào:

Text

Description automatically generated

1. Điều kiện dừng

* Tất cả các mẫu của nút thuộc cùng 1 lớp
* Không còn thuộc tính nào để phân chia
* Không còn mẫu nào

1. **Thuật toán CART**

CART có thể sử dụng để giải thích một biến phụ thuộc liên tục hoặc phân loại dưới dạng nhiều biến độc lập. CART sử dụng phương pháp phân vùng thường được gọi là “**chia để trị**”.

1. Input: Là 1 tập DL đã được gán nhãn
2. Output: Là 1 vecto dự đoán
3. Method: Sử dụng hàm “ Gini\_index “ để đo độ tinh khiết ( purity ) / độ vẩn đục ( impurity ) của 1 phép phân chia

* B1: Tính Gini:

Gini = 1 - = 1 -

* + - C: số lớp cần phân loại (số nhãn trong 1 thuộc tính)
    - N: là tổng số lượng phần tử ở Node đó
* B2: Tính Gini\_index:

Gini\_index = Gini(p) - . Gini(c­­­­k)

* + - Gini(p): chỉ số gini ở Node cha
    - K: số node con được tách ra
    - Gini(ck): chỉ số gini ở Node con thứ k
    - M: số phần tử ở Node p
    - mk : số phần tử ở Node con thứ k

mk = M

* B3: Chọn thuộc tính có hệ số Gini­split Nhỏ nhất:

Ginisplit = . Gini(c­­­­k)

1. Điều kiện dừng

* Tất cả các mẫu của nút thuộc cùng 1 lớp
* Không còn thuộc tính nào để phân chia
* Không còn mẫu nào

**Phần 2: Ứng dụng trong thực tế**

***1.Mô tả bài toán***

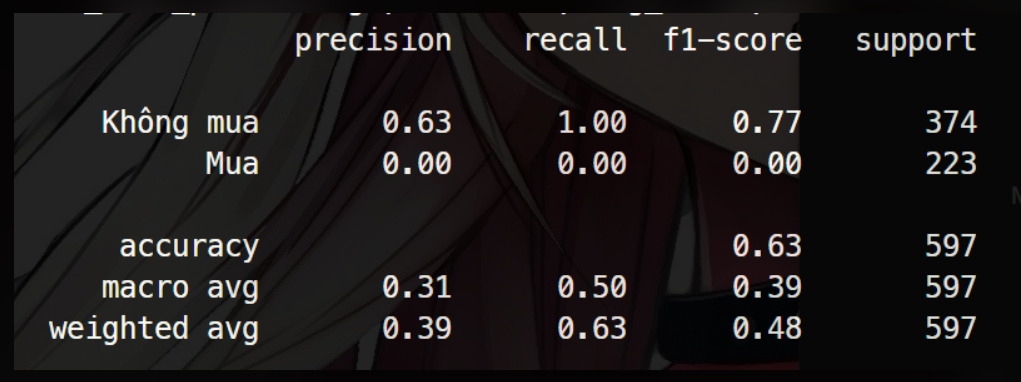
* Tên bài toán: Dự đoán mua bảo hiểm du lịch
* Mục đích của bài toán: Dự đoán mức mua bao hiểm du lịch
* Input: ma trận
* Output: cây quyết định,đường tuyến tính

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

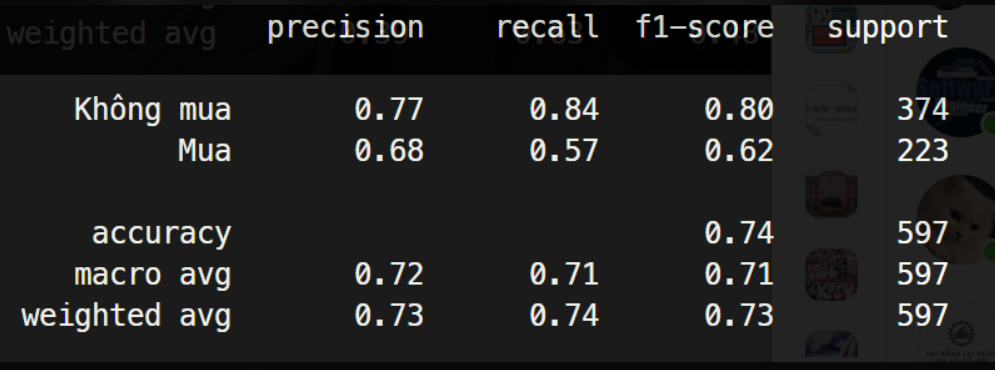
* Dữ liệu:
* **Age :** Tuổi khách hàng
* **EmploymentType :**Lĩnh vực khách hàng đang làm việc
* **GraduateOrNot :** khách hàng có tốt nghiệp đại học hay không
* **AnnualIncome:** Thu nhập hàng năm của khách hàng
* **FamilyMembers:** Số thành viên gia đình sống với khách hàng
* **ChronicDiseases:** Khách hàng có bất kỳ tình trạng mãn tính nào
* **FrequentFlyer:** Khách hàng đặt vé máy bay thường xuyên
* **EverTravelledAbroad:** Khách hàng đã từng đi du lịch nước ngoài chưa
* **TravelInsurance:** Khách Hàng Có Mua Bảo Hiểm Du Lịch Hay Không

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

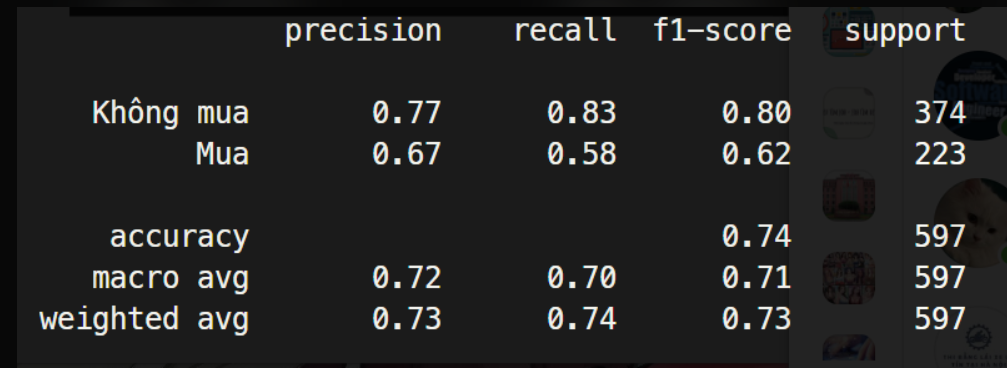
**Perceptron**

****

**--ID3**

****

**CART**

****

**Tài liệu tham khảo**

https://www.kaggle.com/datasets/tejashvi14/travel-insurance-prediction-data?fbclid=IwAR3mRjjfq4p0G-pd3snin-bcPxRY6e8fAB22bXsuvU8vwzufRsx7AzBW8MQ