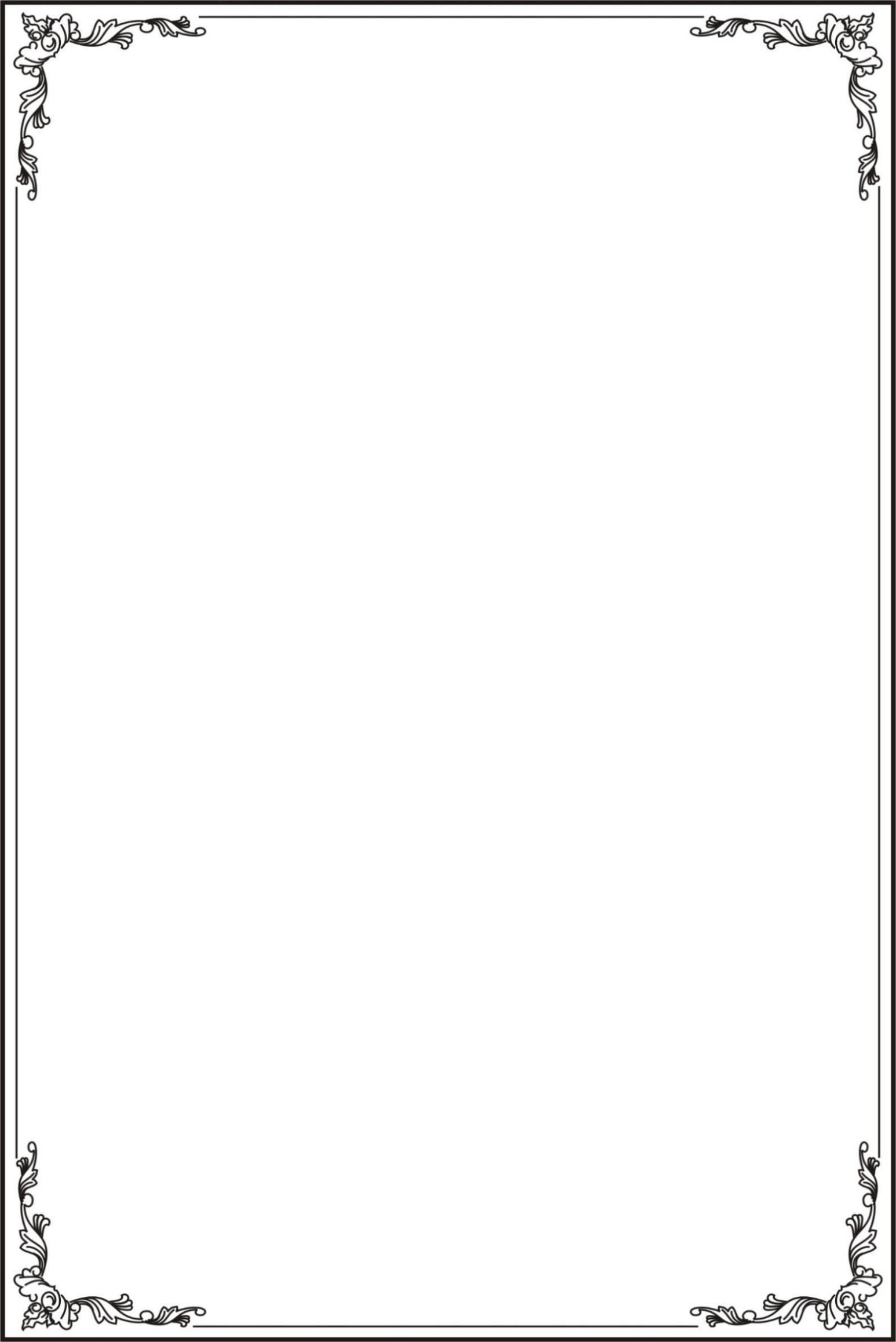
**TRƯỜNG ĐẠI HỌC CNTT & TRUYỀN THÔNG**



**KHOA CÔNG NGHỆ THÔNG TIN**

**======\*\*\*======**



**BÁO CÁO BTL THUỘC HỌC PHẦN :**

**TRÍ TUỆ NHÂN TẠO**

**ĐỀ TÀI: SỬ DỤNG THUẬT TOÁN ID3 ĐỂ ĐƯA RA QUYẾT ĐỊNH NÊN ĐI CÂU CÁ HAY KHÔNG**

|  |  |
| --- | --- |
| GVHD: | Ths. Mai Thanh Hồng |
| Nhóm - Lớp: | 14 - 2025IT6121002 |
| Thành viên: | Nguyễn Đức Anh - 2023602367 |
|  | Nguyễn Trung Kiên - 2023602180 |
|  | Nguyễn Tùng Lâm - 2023602428 |
|  | Đỗ Hải Lâm - 2023602406 |
|  |  |

**Hà Nội, 2025**

# LỜI CẢM ƠN

Để hoàn thành bài tập lớn này, chúng em - Nhóm 14 - xin gửi lời cảm ơn chân thành đến Cô **Mai Thanh Hồng** đã tận tâm giảng dạy, hướng dẫn, định hướng và truyền đạt cho chúng em những kiến thức quý báu trong môn Trí tuệ nhân tạo. Những kiến thức mà cô truyền đạt không chỉ là nền tảng vững chắc cho đề tài này mà còn là hành trang quý báu giúp chúng em tự tin hơn trong quá trình học tập và phát triển sau này.

Chúng em cũng xin gửi lời cảm ơn chân thành đến các bạn trong lớp, những người đã hỗ trợ, động viên và tạo điều kiện thuận lợi để nhóm hoàn thành bài báo cáo đúng tiến độ và đạt chất lượng.

Kính chúc Cô luôn dồi dào sức khỏe và gặt hái nhiều thành công trong sự nghiệp giảng dạy. Chúc các bạn học tập tốt và đạt được nhiều thành tích trong chặng đường phía trước.

**Chúng em xin chân thành cảm ơn!**

**Nhóm thực hiện**

**Nhóm 14**

# MỤC LỤC

[LỜI CẢM ƠN 2](#_Toc216635253)

[MỤC LỤC 3](#_Toc216635254)

[DANH MỤC HÌNH VẼ 4](#_Toc216635255)

[DANH MỤC BẢNG BIỂU 5](#_Toc216635256)

[MỞ ĐẦU 6](#_Toc216635257)

[1. Lý do chọn chủ đề/đề tài 6](#_Toc216635258)

[2. Mục tiêu nghiên cứu 6](#_Toc216635259)

[3. Đối tượng và phạm vi nghiên cứu 7](#_Toc216635260)

[4. Kết quả mong muốn đạt được của đề tài 8](#_Toc216635261)

[CHƯƠNG 1. PHÂN LỚP DỮ LIỆU VÀ THUẬT TOÁN ID3 9](#_Toc216635262)

[1.1. Phân lớp dữ liệu 9](#_Toc216635263)

[1.2. Phân lớp dựa trên cây quyết định 10](#_Toc216635264)

[1.2.1. Cây quyết định 10](#_Toc216635265)

[1.2.2. Độ lợi thông tin 11](#_Toc216635266)

[1.2.3. Thuật toán ID3 12](#_Toc216635267)

[1.2.4. Đánh giá thuật toán ID3 17](#_Toc216635268)

[CHƯƠNG 2. Phát biểu và giải quyết bài toán 19](#_Toc216635269)

[2.1. Mô tả bài toán 19](#_Toc216635270)

[2.2. Mục tiêu 19](#_Toc216635271)

[2.3. Phạm vi áp dụng 19](#_Toc216635272)

[2.4. Tập thuộc tính 19](#_Toc216635273)

[2.5. Tập dữ liệu huấn luyện 20](#_Toc216635274)

[2.5.1. Nguồn xây dựng bộ dữ liệu. 20](#_Toc216635275)

[2.5.2. Tập dữ liệu 23](#_Toc216635276)

[CHƯƠNG 3. CÀI ĐẶT THUẬT TOÁN 31](#_Toc216635277)

[3.1. Cài đặt thuật toán 31](#_Toc216635278)

[3.2. Kết luận và hướng phát triển 43](#_Toc216635279)

[3.2.1. Đánh giá mô hình 43](#_Toc216635280)

[3.2.2. Kết luận 44](#_Toc216635281)

[3.2.2. Hướng phát triển 45](#_Toc216635282)

[TÀI LIỆU THAM KHẢO 45](#_Toc216635283)

# DANH MỤC HÌNH VẼ

[Hình 1.1: Xây dựng mô hình phân lớp 9](#_Toc216526973)

[Hình 1.2: Đánh giá và áp dụng phân lớp cho tập dữ liệu mới 10](#_Toc216526974)

[Hình 1.3: Cây quyết định 16](#_Toc216526975)

[Hình 3.1: Thư viện cần import và tệp dữ liệu 30](#_Toc216527035)

[Hình 3.2: Hàm entropy 30](#_Toc216527036)

[Hình 3.3: Hàm tính độ lợi thông tin của 1 thuộc tính 31](#_Toc216527037)

[Hình 3.4: Xây dựng cây quyết định 33](#_Toc216527038)

[Hình 3.5: Hàm in cây quyết định 35](#_Toc216527039)

[Hình 3.6: Hàm main ( hàm chính ) 35](#_Toc216527040)

[Hình 3.7: Kết quả chạy code ( ảnh 1) 36](#_Toc216527041)

[Hình 3.8: Kết quả chạy code ( ảnh 2) 37](#_Toc216527042)

[Hình 3.9: Kết quả chạy code ( ảnh 3) 38](#_Toc216527043)

[Hình 3.10: Vẽ cây quyết định 41](#_Toc216527046)

[Hình 3.11: Giao diện của website và kết quả dự đoán 42](#_Toc216527047)

# DANH MỤC BẢNG BIỂU

[Bảng 1.2: Tập dữ liệu huấn luyện 14](#_Toc216510101)

[Bảng 2.1: Tập dữ liệu sử dụng 22](#_Toc216510115)

# MỞ ĐẦU

## Lý do chọn chủ đề/đề tài

Trong bối cảnh Trí tuệ nhân tạo ngày càng được ứng dụng rộng rãi vào đời sống, việc khai thác các thuật toán học máy để hỗ trợ ra quyết định trở nên ngày càng phổ biến và cần thiết. Thuật toán ID3 – một phương pháp xây dựng cây quyết định dựa trên entropy và information gain – là một trong những thuật toán nền tảng, đơn giản nhưng mang lại hiệu quả cao.

Việc lựa chọn đề tài “Ứng dụng thuật toán ID3 để quyết định có nên đi câu cá hay không” xuất phát từ mong muốn hiểu rõ cách thức ID3 phân tích dữ liệu và đưa ra quyết định dựa trên các yếu tố ảnh hưởng trong đời sống thực tế như thời tiết, gió, nhiệt độ hay mức độ ẩm. Đây là một tình huống gần gũi, dễ mô phỏng và minh họa rõ ràng quá trình xây dựng cây quyết định.

Thông qua đề tài này, nhóm hướng đến việc hiểu sâu hơn về cách tính entropy, information gain và quy trình tạo cây quyết định. Minh họa tính ứng dụng của ID3 trong các bài toán dự đoán đơn giản nhưng thiết thực và rèn luyện kỹ năng phân tích dữ liệu, triển khai thuật toán và đánh giá kết quả mô phỏng.

Với đề tài mang tính trực quan và thực tế này, quá trình nghiên cứu trở nên dễ tiếp cận hơn, đồng thời giúp làm rõ những giá trị cốt lõi của thuật toán ID3 trong lĩnh vực học máy.

## Mục tiêu nghiên cứu

* Nắm vững khái niệm entropy, information gain và quy trình xây dựng cây quyết định.
* Hiểu rõ ưu điểm, hạn chế và phạm vi áp dụng của ID3 trong học máy.
* Xác định các thuộc tính ảnh hưởng đến quyết định (thời tiết, sức gió, nhiệt độ, độ ẩm,…).
* Tiền xử lý, chuẩn hóa và đánh giá chất lượng dữ liệu.
* Triển khai thuật toán dựa trên tập dữ liệu đã thu thập.
* Trực quan hóa cấu trúc cây nhằm thể hiện rõ logic ra quyết định.
* Kiểm tra mức độ chính xác và khả năng dự đoán của cây quyết định tạo ra.
* So sánh kết quả với trực giác hoặc các phương pháp dự đoán đơn giản khác.
* Phân tích mức độ áp dụng của mô hình trong đời sống hàng ngày.
* Gợi ý hướng phát triển hoặc mở rộng thuật toán cho các bài toán ra quyết định khác.

## Đối tượng và phạm vi nghiên cứu

* Đối tượng nghiên cứu:
  + Thuật toán ID3 và cơ chế xây dựng cây quyết định dựa trên entropy và information gain.
  + Tập dữ liệu mô tả điều kiện môi trường ảnh hưởng đến quyết định đi câu cá, bao gồm các thuộc tính như: thời tiết, nhiệt độ, sức gió, độ ẩm, tình trạng mặt nước, v.v.
  + Mối quan hệ giữa các thuộc tính trong dữ liệu và kết quả cuối cùng (có đi câu cá hay không).
* Phạm vi nghiên cứu:
  + Đề tài chỉ tập trung ứng dụng một thuật toán duy nhất là ID3 để xây dựng mô hình dự đoán.
  + Dữ liệu được sử dụng mang tính mô phỏng hoặc dữ liệu mẫu quy mô nhỏ, không mở rộng sang dữ liệu thực tế phức tạp hay nhiều nguồn khác nhau.
  + Phân tích dừng lại ở mức xây dựng và đánh giá mô hình cơ bản, không mở rộng sang các biến thể của cây quyết định như C4.5, CART hay các kỹ thuật ensemble khác.
  + Phạm vi ứng dụng giới hạn trong bài toán ra quyết định đơn lẻ: dự đoán xem điều kiện hiện tại có phù hợp để đi câu cá hay không, không nghiên cứu sâu về các hoạt động thủy sản hay dự báo môi trường nâng cao.

## Kết quả mong muốn đạt được của đề tài

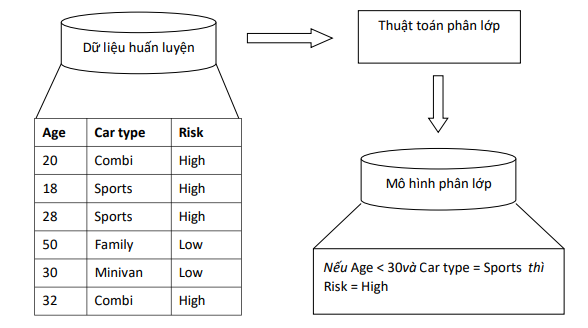
* Xây dựng được mô hình cây quyết định dựa trên thuật toán ID3, thể hiện rõ các nhánh lựa chọn và điều kiện dẫn đến quyết định cuối cùng có đi câu cá hay không.
* Trình bày đầy đủ quá trình tính toán entropy và information gain, qua đó minh họa chi tiết cách thuật toán ID3 lựa chọn thuộc tính phân chia dữ liệu.
* Tạo ra mô hình dự đoán có độ chính xác chấp nhận được khi kiểm thử trên tập dữ liệu mẫu, chứng minh thuật toán ID3 có khả năng áp dụng hiệu quả cho các bài toán ra quyết định đơn giản.
* Hiểu rõ mối quan hệ giữa các thuộc tính môi trường (như thời tiết, gió, độ ẩm…) và kết quả dự đoán, đồng thời rút ra được các yếu tố có ảnh hưởng mạnh nhất đến quyết định đi câu cá.
* Xây dựng được cây quyết định trực quan hóa, giúp người đọc dễ dàng hình dung logic hoạt động của mô hình.
* Đề xuất hướng phát triển hoặc mở rộng cho các nghiên cứu tiếp theo, chẳng hạn sử dụng tập dữ liệu lớn hơn, thử nghiệm các thuật toán ra quyết định khác hoặc so sánh hiệu quả giữa các phương pháp.

# CHƯƠNG 1. PHÂN LỚP DỮ LIỆU VÀ THUẬT TOÁN ID3

## Phân lớp dữ liệu

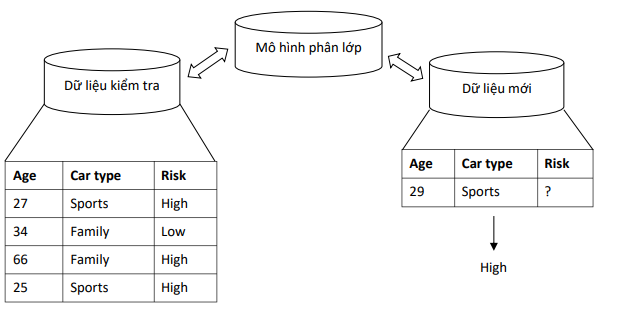
Phân lớp dữ liệu (data classification) là xếp đối tượng dữ liệu vào một trong các lớp đã được xác định trước. Phân lớp dữ liệu gồm hai bước là học và phân lớp:

* Bước học (giai đoạn huấn luyện): nhằm xây dựng một mô hình mô tả một tập các lớp dữ liệu đã định trước.Tập hợp các bộ dữ liệu được dùng để xây dựng mô hình được gọi là tập dữ liệu huấn luyện. Mô hình được biểu diễn dưới dạng một tập các luật phân lớp, công thức toán học, ...



Hình 1.1: Xây dựng mô hình phân lớp

* Bước phân lớp: bước này dùng mô hình đã xây dựng ở bước trước để phân lớp (gán nhãn) cho những đối tượng chưa được xếp lớp. Trước khi vận hành mô hình cần đánh giá độ chính xác của mô hình bằng cách so sánh kết quả của các mẫu kiểm tra với kết quả phân lớp của mô hình. Độ chính xác là phần trăm của số mẫu kiểm tra được phân lớp đúng. Tập dữ liệu kiểm tra và tập dữ liệu huấn luyện độc lập với nhau.



Hình 1.2: Đánh giá và áp dụng phân lớp cho tập dữ liệu mới

Phân lớp dữ liệu nhằm rút trích các mô hình mô tả các lớp dữ liệu hoặc dự đoán xu hướng dữ liệu. Một số giải thuật phân lớp:

* Phân lớp với cây quyết định (decision tree)
* Phân lớp Bayesian (Bayesian classifier)
* Phân lớp với mạng nơron (neural)
* Phân lớp với k láng giềng gần nhất (k – nearest neighbor)
* Phân lớp với suy diễn dựa trên tình huống (case – based reasoning)
* Phân lớp dựa trên thuật giải di truyền (genetic algorithms)
* Phân lớp với lý thuyết tập thô (rough sets), - Phân lớp với lý thuyết tập mờ (fuzzy sets), …

## Phân lớp dựa trên cây quyết định

### Cây quyết định

Trong lĩnh vực máy học, cây quyết định là một kiểu mô hình dự báo (predictive model) nghĩa là một ánh xạ từ các quan sát về một sự vật/ hiện tượng tới các kết luận về giá trị mục tiêu của sự vật/ hiện tượng.

Quá trình học trên cây quyết định được dựa trên tập dữ liệu huấn luyện có dạng bảng, ở đó các cột là các thuộc tính điều kiện và một thuộc tính mục tiêu (quyết định), mỗi dòng sẽ được gán nhãn là giá trị của thuộc tính mục tiêu. Chẳng hạn với bảng dữ liệu, các thuộc tính “Age” và “Salary” là thuộc tính điều kiện còn thuộc tính mục tiêu là “Class”; các dòng được gán nhãn “Good” hoặc “Bad” là giá trị của thuộc tính “Class”.

Cây quyết định là một công cụ khá phổ biến trong nhiều dạng ứng dụng, với cơ chế rút trích các luật dạng if – then dựa trên tập dữ liệu mẫu. Sau đây là định nghĩa cây quyết định.

Cây quyết định là dạng đặc biệt của cấu trúc cây nhằm hỗ trợ cho tiến trình ra quyết định. Cây gồm các thành phần:

* Nút trong: biểu diễn thuộc tính
* Nhánh: biểu diễn giá trị cụ thể của thuộc tính đó
* Nút lá: đại diện cho giá trị dự đoán của thuộc tính mục tiêu

### Độ lợi thông tin

Với bảng dữ liệu huấn luyện có các thuộc tính điều kiện A1, A2, …, An, vấn đề đặt ra là trong số thuộc tính này ta sẽ lựa chọn thuộc tính nào làm gốc của cây. Để lựa chọn thuộc tính làm gốc của cây ta cần đưa ra tiêu chuẩn để đánh giá các thuộc tính này, một trong các tiêu chuẩn đó là độ lợi thông tin (information gain) của thuộc tính. Các độ đo dựa trên lý thuyết thông tin đều sử dụng đến một khái niệm đó là entropy.

Entropy là đại lượng đặc trưng cho mức độ hỗn loạn của các phần tử trong quần thể S

Entropy của tập S được tính theo công thức:

Entropy(S) =

Với:

S: tập dữ liệu mẫu;

bt: tổng số phần tử có trong phân hoạch;

bj: tổng số phần tử có thuộc tính A với giá trị của thuộc tính là j;

bji: tổng số phần tử của thuộc tính A có giá trị j ứng với thuộc tính mục tiêu có giá trị i.

* Nếu Entropy (S) = 0 thì S đồng nhất (phân tách hoàn toàn)
* Nếu Entropy (S) = 1 thì số phần tử thuộc lớp ci (i = 1, 2,…, n) bằng nhau
* Nếu Entropy (S) ∈ (0;1) thì số phần tử thuộc các lớp ci khác nhau Độ lợi thông tin (information gain) là đại lượng được dùng để chọn thuộc tính nhằm phân chia tập dữ liệu huấn luyện.

Độ lợi thông tin của thuộc tính A trong tập mẫu S, ký hiệu Gain (S, A) và được xác định qua công thức: Gain (S, A) = Entropy (S) – Info (S, A)

Trong đó: Info(S, A) = ∑( \* Entropy())

: tập con của S mà ở đó thuộc tính A có giá trị v;

|S|: số mẫu trong S.

Chẳng hạn S là tập hợp của 14 mẫu trong đó có 9 mẫu thuộc lớp Yes và 5 mẫu thuộc lớp No. Khi đó Entropy (S) = - (9/14).log2(9/14) - (5/14).log2(5/14) = 0,940

Nhận xét: Entropy (S) là hằng số đối với tất cả thuộc tính. Mức độ hỗn loạn của các phân hoạch do thuộc tính A tạo ra (Entropy ()) được mong đợi là có giá trị càng nhỏ càng tốt (ít hỗn loạn).

### Thuật toán ID3

Thuật toán ID3 (Iterative DiChaudomiser 3) được J. Ross Quinlan phát triển tại đại học Sydney năm 1975. ID3 dựa trên thuật giải CLS (Concept Learning System). ID3 là thuật giải học có giám sát (supervised learning algorithm), tạo cây quyết định dựa trên tập dữ liệu có sẵn. Cây kết quả được sử dụng để phân lớp cho mẫu mới chưa được gán nhãn.

Thuật toán ID3 tạo cây quyết định dựa trên độ lợi thông tin thu được từ tập dữ liệu huấn luyện. Thuộc tính được chọn là thuộc tính có độ lợi thông tin lớn nhất. Thuật toán ID3 thường được sử dụng với các thuộc tính định danh (nominal attributes) để phân lớp với dữ liệu đầy đủ.

Thuật toán này tạo cây quyết định theo chiến lược phân tích từ trên xuống (Top – Down), bắt đầu từ một tập các đối tượng và các thuộc tính của nó. Tại mỗi nút của cây một thuộc tính được kiểm tra, kết quả của phép kiểm tra này được sử dụng để phân chia tập đối tượng theo kết quả kiểm tra trên. Quá trình này được thực hiện một cách đệ quy cho tới khi tập đối tượng trong cây con được sinh ra thuần nhất (các mẫu có cùng giá trị của thuộc tính mục tiêu) theo một tiêu chí phân lớp nào đó.

Cho X là tập các vectơ đặc trưng (không gian đặc trưng)

X = {x1, x2,…, xn} trong đó xi = (xi1, xi2,…, xim) với i = 1, 2, …, n

C là tập các giá trị phân lớp

c: X → C thực hiện gán nhãn cho phần tử trong X

D = {(x1, c(x1)), (x2, c(x2),…, (xn, c(xn)} ⊆ X × C

Nhiệm vụ: Dựa trên D, hãy xây dựng cây quyết định T xấp xỉ c. ID3(D, Atributes, Target) {

Tạo nút t cho cây,

* Nếu tất cả các mẫu trong D đều thuộc lớp dương thì trả về cây có nút gốc t và được gán nhãn “+”,
* Nếu tất cả các mẫu trong D đều thuộc lớp âm thì trả về cây có nút gốc t và được gán nhãn “-”,
* Đặt t là nhãn phổ biến nhất của thuộc tính mục tiêu trong D, - Nếu Atributes rỗng thì trả về cây có một nút gốc trỏ bởi t,
* Trường hợp khác:
* Đặt A\* là thuộc tính phân lớp tốt nhất trong D,
* Thuộc tính quyết định của t là A\* .
* Với mỗi giá trị a của A\* thực hiện:
* Bổ sung nhánh mới dưới t ứng với A\* = “a”,
* Đặt D\_a là tập con của D chứa các mẫu mà A\* = “a”,
* Nếu D\_a rỗng thì dưới nhánh mới này bổ sung nút lá với nhãn phổ biến nhất trong D.
* Ngược lại, dưới nhánh mới này bổ sung cây con ID3

(D\_a, Attributes − {A\*}, Target). - Trả về t.

}

Sau đây là mã giả cho thuật giải ID3:

ID3(D, Attributes, Target) {

1. t = createNode()

2. IF ∀(x, c(x)) ∈D: c(x) = 1 THEN label(t) = ‘+’ return(t) ENDIF

3. IF ∀(x, c(x)) ∈ D: c(x) = 0 THEN label(t) = ‘-’ return(t) ENDIF

4. label(t) = mostCommonClass(D, Target)

5. IF Attributes = ∅ THEN return(t) ENDIF

6. A\* = argmaxA∈Attributes(Gain(D, A))

7. FOR EACH a∈A\* DO

Da = {(x, c(x)) ∈ D: x| \* A = a}

IF Da = ∅ THEN

t ’ = createNode()

label(t’ ) = mostCommonClass(D, Target)

createEdge(t, a, t’ )

ELSE

createEdget(t, a, ID3(Da, Attributes – {A\*}, Target))

ENDIF

ENDDO

8. return(t)}

Để minh họa cho thuật toán này ta xét tập dữ liệu huấn luyện trong bảng sau:

Bảng 1.2: Tập dữ liệu huấn luyện

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Outlook | Temperature | Humidity | Wind | WaterClarity | Time | Play |
| Âm u | Nóng | Vừa | Không gió | Trong | Khác | Có |
| Âm u | Nóng | Thấp | Gió nhẹ | Trong | Tối | Có |
| Âm u | Lạnh | Thấp | Gió nhẹ | Trong | Tối | Có |
| Mưa nhẹ | Nóng | Vừa | Gió mạnh | Quá đục | Chiều muộn | Không |
| Mát mẻ | Lạnh | Thấp | Gió nhẹ | Trong | Sáng sớm | Có |
| Mưa lớn | Nóng | Vừa | Gió nhẹ | Quá đục | Chiều muộn | Không |
| Mát mẻ | Lạnh | Vừa | Gió mạnh | Trong | Tối | Có |
| Mát mẻ | Vừa | Vừa | Gió nhẹ | Trong | Sáng sớm | Có |
| Mưa lớn | Vừa | Cao | Không gió | Quá đục | Sáng sớm | Không |
| Mát mẻ | Lạnh | Thấp | Gió nhẹ | Hơi đục | Khác | Có |
| Mưa lớn | Lạnh | Thấp | Gió nhẹ | Quá đục | Sáng sớm | Không |
| Âm u | Vừa | Cao | Gió nhẹ | Trong | Khác | Có |
| Nắng gắt | Lạnh | Thấp | Không gió | Hơi đục | Sáng sớm | Có |
| Mưa lớn | Vừa | Thấp | Gió mạnh | Quá đục | Chiều muộn | Không |

Bây giờ ta tính độ lợi thông tin cho từng thuộc tính, sau đó chọn thuộc tính có độ lợi thông tin lớn nhất làm gốc của cây quyết định. Xét thuộc tính “Outlook”:

* Entropy(S)= - (9/13).log2(9/13) - (4/13).log2(4/13) = 0,889
* Entropy() = 0
* Entropy() = 0
* Entropy() = 0
* Entropy() = - (1/2).log2(1/2) - (1/2).log2(1/2) = 1,00
* Entropy() = - (1/4).log2(1/4) - (3/4).log2(3/4) = 0,811
* Gain(S, Outlook) = Entropy(S) - (2/13).Entropy() - (4/13).Entropy() = 0,889 – (0,154 + 0,250) = 0,485

Tương tự như vậy với các thuộc tính điều kiện còn lại

* Gain(S, Time) = 0,286
* Gain(S, WaterClarity) = 0,158
* Gain(S, Wind) = 0,114
* Gain(S, Humidity) = 0,109
* Gain(S, Temperature) = 0,096

Trong các thuộc tính điều kiện trên ta thấy “Outlook” có độ lợi thông tin lớn nhất nên thuộc tính này sẽ được chọn làm gốc của cây.

Trong nhánh Outlook = Mưa nhẹ, = {2 mẫu} = {(Có,Không)}

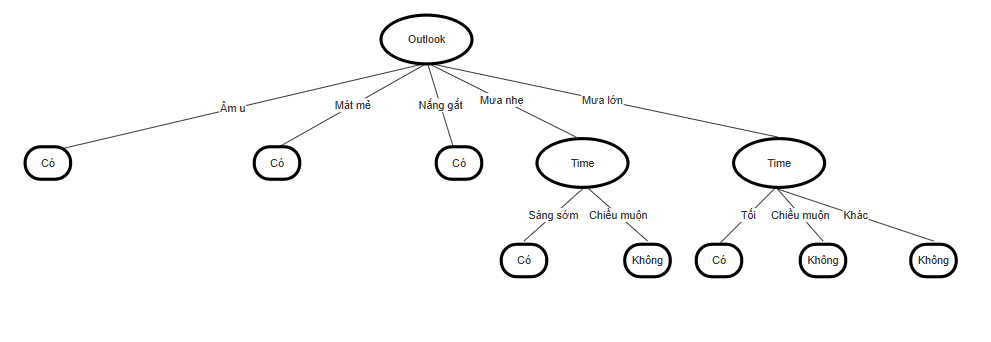
Gain(, Time) = 1,00

Do Gain đạt giá trị lơn nhất nên ta chọn Time làm thuộc tính tách kế tiếp với 2 nhánh:

Nếu Outlook = Mưa nhẹ AND Time = Sáng sớm → Play = Có

Nếu Outlook = Mưa nhẹ AND Time = Chiều muộn → Play = Không

Tương tự như vậy với nhánh Outlook = “Mưa lớn” ta cũng tính được độ lợi thông tin cho các thuộc tính còn lại và chọn “Time” làm gốc. Cuối cùng ta có cây quyết định:



Hình 1.3: Cây quyết định

Từ cây quyết định (hình ) ta có tập luật sau:

* Rule 1: IF Outlook = Âm u THEN Play = Có
* Rule 2: IF Outlook = Mát mẻ THEN Play = Có
* Rule 3: IF Outlook = Nắng gắt THEN Play = Có
* Rule 4: IF Outlook = Mưa nhẹ AND Time = Sáng sớm THEN Play = Có
* Rule 5: IF Outlook = Mưa nhẹ AND Time = Chiều muộn THEN Play = Không
* Rule 6: IF Outlook = Mưa lớn AND Time = Tối THEN Play = Có
* Rule 7: IF Outlook = Mưa lớn AND Time = Sáng sớm THEN Play = Không
* Rule 8: IF Outlook = Mưa lớn AND Time = Khác THEN Play = Không
* Rule 9: IF Outlook = Mưa lớn AND Time = Chiều muộn THEN Play = Không

### Đánh giá thuật toán ID3

Thuật toán ID3 là một trong những thuật toán tiêu biểu dùng để xây dựng cây quyết định trong bài toán phân lớp dữ liệu. Thông qua việc sử dụng độ đo Entropy và Information Gain, ID3 cho phép lựa chọn thuộc tính phân chia dữ liệu một cách có hệ thống và logic. Tuy nhiên, bên cạnh những ưu điểm nổi bật, thuật toán này cũng tồn tại một số hạn chế nhất định.

Ưu điểm, ID3 có cấu trúc mô hình đơn giản, dễ hiểu và dễ diễn giải. Kết quả của thuật toán là một cây quyết định với các quy tắc rõ ràng dạng “nếu – thì”, giúp người dùng dễ dàng theo dõi và giải thích quá trình ra quyết định. Ngoài ra, thuật toán có tốc độ huấn luyện nhanh khi áp dụng trên các tập dữ liệu có kích thước nhỏ hoặc trung bình, đặc biệt phù hợp cho mục đích học tập và minh họa nguyên lý phân lớp dữ liệu.

Tuy nhiên, thuật toán ID3 cũng bộc lộ nhiều hạn chế khi áp dụng vào các bài toán thực tế. Trước hết, ID3 chỉ xử lý hiệu quả với các thuộc tính rời rạc, trong khi các thuộc tính liên tục cần phải được rời rạc hóa trước khi đưa vào mô hình. Bên cạnh đó, thuật toán có xu hướng ưu tiên các thuộc tính có nhiều giá trị khác nhau do Information Gain dễ đạt giá trị cao, dẫn đến hiện tượng thiên lệch trong quá trình lựa chọn thuộc tính. Ngoài ra, ID3 không có cơ chế cắt tỉa cây (pruning), khiến mô hình dễ bị overfitting, đặc biệt khi dữ liệu có nhiễu hoặc số lượng mẫu huấn luyện hạn chế.

Từ những phân tích trên có thể thấy rằng, mặc dù ID3 có vai trò quan trọng trong việc đặt nền tảng cho các thuật toán cây quyết định, nhưng thuật toán này chưa thực sự phù hợp cho các bài toán dữ liệu lớn và phức tạp. Trong thực tế, các thuật toán cải tiến như C4.5 hoặc CART thường được ưu tiên sử dụng do khả năng xử lý thuộc tính liên tục tốt hơn và có cơ chế cắt tỉa giúp tăng khả năng tổng quát hóa của mô hình.

# CHƯƠNG 2. Phát biểu và giải quyết bài toán

## Mô tả bài toán

Câu cá là một hoạt động giải trí phổ biến tại miền Bắc Việt Nam. Tuy nhiên, hiệu quả của việc câu cá chịu ảnh hưởng mạnh mẽ bởi nhiều yếu tố môi trường như thời tiết, nhiệt độ, gió, độ đục nước và thời gian trong ngày. Người đi câu thường dựa trên kinh nghiệm cá nhân để quyết định có nên đi câu hay không, dẫn đến tính chủ quan và kết quả không ổn định.

Trong bối cảnh đó, việc ứng dụng các kỹ thuật học máy – đặc biệt là thuật toán ID3 – để xây dựng mô hình dự đoán “Có nên đi câu cá nước ngọt hay không?” có ý nghĩa thực tiễn:

* Giảm sự phụ thuộc vào kinh nghiệm cá nhân.
* Chuẩn hóa việc đánh giá điều kiện câu cá.
* Hỗ trợ người mới đi câu đưa ra quyết định hợp lý.
* Tận dụng dữ liệu thực tế để học các mối quan hệ phức tạp giữa các yếu tố môi trường.

## Mục tiêu

Xây dựng thuật toán đưa ra quyết định có nên đi câu cá nước hay không bằng thuật toán id3 dựa trên các điều kiện thực tế. Bài toán chỉ áp dụng khi câu cá nước ngọt ở miền Bắc Việt Nam.

Cây quyết định sẽ dựa trên các thuộc tính môi trường – thời tiết có ảnh hưởng trực tiếp đến hành vi ăn của cá nước ngọt. Bao gồm: Thời tiết (Outlook), Nhiệt độ (Temperature), Độ ẩm (Humidity), Tốc độ gió (Wind), Độ đục của nước (WaterClarity), thời gian (Time)

## Phạm vi áp dụng

* Phạm vi bài toán giới hạn :
  + Khu vực: Miền Bắc Việt Nam
  + Môi trường: hồ, sông , ao nước ngọt
  + Các loài cá: cá trắm cỏ, cá chép, cá trôi, cá rô phi, cá rô đồng, cá trê

## Tập thuộc tính

Bài toán sử dụng sáu thuộc tính đầu vào, được lựa chọn dựa trên các yếu tố môi trường – thời tiết có ảnh hưởng trực tiếp đến hành vi kiếm ăn của cá nước ngọt tại miền Bắc Việt Nam.

Thuộc tính đầu tiên là Thời tiết (Outlook), biểu thị tình trạng bầu trời và lượng mưa, gồm bốn giá trị: Nắng gắt, Mát mẻ,Âm u, Mưa nhẹ và Mưa lớn.

Thuộc tính thứ hai là Nhiệt độ (Temperature), được phân thành ba mức Lạnh, Vừa (20–28°C) và Nóng, phản ánh mức độ ảnh hưởng của nhiệt độ môi trường lên khả năng cá nổi lên kiếm mồi.

Thuộc tính thứ ba là Độ ẩm (Humidity), gồm ba mức Thấp, Vừa (40–70%) và Cao, giúp mô tả sự ổn định của thời tiết và không khí – yếu tố thường liên quan đến mức độ hoạt động của cá.

Thuộc tính thứ tư là Gió (Wind), bao gồm Không gió, Gió nhẹ (5–20 km/h) và Gió mạnh; tốc độ gió quyết định sự dao động của mặt nước, từ đó ảnh hưởng tới khả năng cá quan sát và tiếp cận mồi. Thuộc tính thứ năm,

Độ đục của nước (WaterClarity), có ba mức Trong, Hơi đục (20–50 cm) và Quá đục; độ trong của nước là yếu tố quan trọng vì cá nước ngọt dựa nhiều vào khả năng nhìn thấy mồi.

Cuối cùng, thuộc tính Thời gian (Time) quy định thời điểm trong ngày, gồm Sáng sớm (5:00–10:00), Chiều muộn (16:00–19:00), Tối (19:00–24:00) và Khác (những khoảng thời gian còn lại), vì từng loài cá thường có thời điểm hoạt động mạnh hơn trong ngày.

## Tập dữ liệu huấn luyện

### Nguồn xây dựng bộ dữ liệu.

Bài toán chỉ áp dụng cho một số loài cá phổ biến ở khu vực miền Bắc Việt Nam, do đó bộ dữ liệu được xây dựng dựa trên 6 loài cá nước ngọt phổ biến nhất, bao gồm: cá Trắm Cỏ, cá Chép, cá Trôi, cá Rô Phi, cá Rô Đồng và cá Trê.

Dữ liệu được tổng hợp từ các nguồn nghiên cứu thủy sản trong nước và quốc tế, phản ánh mối quan hệ giữa điều kiện môi trường, thời tiết và tập tính ăn mồi của từng loài cá.

#### 1. Cá Trắm Cỏ (Ctenopharyngodon idella)

**Môi trường ưa thích**

* Nhiệt độ nước: 20 – 32°C (tối ưu 25–30°C)
* Độ trong của nước: nước trong hoặc hơi đục, ít bùn
* Dòng chảy/gió: nước tĩnh hoặc chảy nhẹ
* Độ ẩm không khí: 60% – 85%
* Thời tiết thích hợp: mát, ít mưa, không có gió mạnh
* Thời điểm ăn mạnh: sáng sớm và chiều mát

**Đặc điểm**

* Chủ yếu ăn rong, cỏ tươi và các loại thực vật thủy sinh như bèo tấm
* Hoạt động kiếm ăn mạnh khi môi trường nước ổn định

**Nguồn tham khảo**

* Viện Nghiên cứu Nuôi trồng Thủy sản I (RIA1)
* FAO – Species Fact Sheet: Grass Carp

#### 2. Cá Chép (Cyprinus carpio)

**Môi trường ưa thích**

* Nhiệt độ nước: 17 – 28°C (tối ưu 20–25°C)
* Độ trong của nước: hơi đục, có lớp bùn mỏng
* Dòng chảy: ao hồ tĩnh, kênh mương có dòng chảy chậm
* Gió: gió nhẹ, tránh gió mạnh
* Thời tiết thích hợp: mát, âm u nhẹ, mưa phùn
* Thời điểm ăn mạnh: sáng sớm và chiều muộn

**Đặc điểm**

* Là loài cá ăn tạp, thức ăn bao gồm giun, ấu trùng côn trùng, bột bắp và khoai

**Nguồn tham khảo**

* Giáo trình Thủy sinh học – Trường Đại học Nông Lâm
* FAO – Carp Culture Technical Guide

#### 3. Cá Trôi (Cirrhinus molitorella)

**Môi trường ưa thích**

* Nhiệt độ nước: 22 – 32°C
* Độ trong của nước: từ hơi đục đến đục
* Dòng chảy/gió: nước tĩnh, ít gió
* Độ ẩm không khí: 70% – 90%
* Thời tiết thích hợp: nóng ẩm, mưa nhỏ, trời âm u
* Thời điểm ăn mạnh: buổi sáng và buổi chiều

**Đặc điểm**

* Thức ăn chủ yếu là rong mềm, bèo tấm và cám

**Nguồn tham khảo**

* Viện Nghiên cứu Nuôi trồng Thủy sản I (RIA1) – Cá trôi Ấn
* Sách Kỹ thuật nuôi cá truyền thống

#### 4. Cá Rô Phi (Oreochromis niloticus)

**Môi trường ưa thích**

* Nhiệt độ nước: 25 – 35°C (có thể chịu lạnh đến khoảng 18°C)
* Độ trong của nước: hơi đục đến đục, nước nhiều phù sa
* Dòng chảy: nước tĩnh hoặc chảy nhẹ
* Gió: gió nhẹ đến vừa
* Thời tiết thích hợp: nắng nhẹ, nhiệt độ ấm
* Thời điểm ăn mạnh: sáng và chiều

**Đặc điểm**

* Là loài ăn tạp mạnh, sinh trưởng tốt và dễ nuôi

**Nguồn tham khảo**

* FAO – Tilapia Culture Guide
* Viện Nghiên cứu Nuôi trồng Thủy sản I – Tilapia Studies

#### 5. Cá Rô Đồng (Anabas testudineus)

**Môi trường ưa thích**

* Nhiệt độ nước: 20 – 34°C
* Độ trong của nước: đục, nhiều bùn
* Gió: gió nhẹ
* Thời tiết thích hợp: nóng ẩm
* Thời điểm ăn mạnh: chiều và tối

**Đặc điểm**

* Ăn các loại động vật nhỏ, có khả năng thích nghi tốt với môi trường nước kém chất lượng

**Nguồn tham khảo**

* Bộ Nông nghiệp và Phát triển Nông thôn – Kỹ thuật nuôi cá Rô đồng
* FAO – Inland Fish Catalogue

#### 6. Cá Trê (Clarias spp.)

**Môi trường ưa thích**

* Nhiệt độ nước: 25 – 32°C
* Độ trong của nước: rất đục, nhiều bùn
* Gió: không ảnh hưởng nhiều
* Thời tiết thích hợp: nóng, ẩm
* Thời điểm ăn mạnh: chiều tối đến ban đêm

**Đặc điểm**

* Là loài cá ăn tạp, đặc biệt thích các loại mồi tanh như tim, gan và cá vụn

**Nguồn tham khảo**

* Giáo trình Nuôi cá nước ngọt – Trường Đại học Nông Lâm
* FAO – Catfish Culture Notes

### Tập dữ liệu

Bảng 2.1: Tập dữ liệu sử dụng

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Outlook | Temperature | Humidity | Wind | WaterClarity | Time | Play |
| Âm u | Nóng | Vừa | Không gió | Trong | Khác | Có |
| Âm u | Nóng | Thấp | Gió nhẹ | Trong | Tối | Có |
| Âm u | Lạnh | Thấp | Gió nhẹ | Trong | Tối | Có |
| Mưa nhẹ | Nóng | Vừa | Gió mạnh | Quá đục | Chiều muộn | Không |
| Mát mẻ | Lạnh | Thấp | Gió nhẹ | Trong | Sáng sớm | Có |
| Mưa lớn | Nóng | Vừa | Gió nhẹ | Quá đục | Chiều muộn | Không |
| Mát mẻ | Lạnh | Vừa | Gió mạnh | Trong | Tối | Có |
| Mát mẻ | Vừa | Vừa | Gió nhẹ | Trong | Sáng sớm | Có |
| Mưa lớn | Vừa | Cao | Không gió | Quá đục | Sáng sớm | Không |
| Mát mẻ | Lạnh | Thấp | Gió nhẹ | Hơi đục | Khác | Có |
| Mưa lớn | Lạnh | Thấp | Gió nhẹ | Quá đục | Sáng sớm | Không |
| Âm u | Vừa | Cao | Gió nhẹ | Trong | Khác | Có |
| Nắng gắt | Lạnh | Thấp | Không gió | Hơi đục | Sáng sớm | Có |
| Mưa lớn | Vừa | Thấp | Gió mạnh | Quá đục | Chiều muộn | Không |
| Mát mẻ | Nóng | Cao | Không gió | Quá đục | Tối | Không |
| Mưa nhẹ | Vừa | Thấp | Không gió | Trong | Tối | Có |
| Âm u | Lạnh | Vừa | Không gió | Quá đục | Sáng sớm | Không |
| Mưa nhẹ | Lạnh | Thấp | Không gió | Quá đục | Sáng sớm | Không |
| Mưa nhẹ | Nóng | Thấp | Không gió | Hơi đục | Chiều muộn | Có |
| Âm u | Lạnh | Vừa | Gió mạnh | Quá đục | Chiều muộn | Không |
| Mưa nhẹ | Lạnh | Cao | Gió mạnh | Trong | Tối | Không |
| Mưa nhẹ | Vừa | Thấp | Gió nhẹ | Hơi đục | Tối | Có |
| Mưa lớn | Nóng | Vừa | Gió mạnh | Hơi đục | Sáng sớm | Không |
| Mát mẻ | Vừa | Vừa | Không gió | Hơi đục | Sáng sớm | Có |
| Âm u | Vừa | Cao | Gió mạnh | Hơi đục | Khác | Không |
| Mát mẻ | Lạnh | Vừa | Không gió | Hơi đục | Tối | Có |
| Nắng gắt | Nóng | Cao | Không gió | Hơi đục | Tối | Không |
| Mưa nhẹ | Lạnh | Cao | Không gió | Quá đục | Sáng sớm | Không |
| Mưa nhẹ | Lạnh | Vừa | Gió nhẹ | Hơi đục | Sáng sớm | Có |
| Mát mẻ | Vừa | Vừa | Gió nhẹ | Hơi đục | Chiều muộn | Có |
| Âm u | Lạnh | Cao | Gió mạnh | Trong | Tối | Không |
| Mưa lớn | Nóng | Cao | Gió nhẹ | Hơi đục | Tối | Không |
| Âm u | Lạnh | Vừa | Gió nhẹ | Quá đục | Tối | Không |
| Nắng gắt | Vừa | Cao | Gió nhẹ | Quá đục | Sáng sớm | Không |
| Mưa lớn | Lạnh | Vừa | Gió nhẹ | Trong | Chiều muộn | Không |
| Âm u | Vừa | Thấp | Không gió | Hơi đục | Chiều muộn | Có |
| Mưa lớn | Lạnh | Cao | Gió nhẹ | Quá đục | Tối | Không |
| Nắng gắt | Lạnh | Vừa | Không gió | Trong | Tối | Có |
| Mưa nhẹ | Vừa | Vừa | Không gió | Quá đục | Chiều muộn | Không |
| Mưa lớn | Vừa | Thấp | Gió mạnh | Quá đục | Tối | Không |
| Nắng gắt | Nóng | Thấp | Không gió | Trong | Sáng sớm | Có |
| Nắng gắt | Vừa | Cao | Gió nhẹ | Trong | Chiều muộn | Có |
| Âm u | Lạnh | Vừa | Gió nhẹ | Trong | Khác | Có |
| Mát mẻ | Lạnh | Cao | Gió mạnh | Trong | Khác | Không |
| Âm u | Vừa | Thấp | Không gió | Trong | Sáng sớm | Có |
| Nắng gắt | Lạnh | Thấp | Không gió | Hơi đục | Khác | Có |
| Mưa lớn | Lạnh | Cao | Gió nhẹ | Hơi đục | Khác | Không |
| Mát mẻ | Lạnh | Vừa | Gió nhẹ | Trong | Chiều muộn | Có |
| Nắng gắt | Lạnh | Cao | Gió nhẹ | Hơi đục | Tối | Có |
| Âm u | Vừa | Cao | Gió mạnh | Quá đục | Sáng sớm | Không |
| Nắng gắt | Lạnh | Cao | Gió nhẹ | Trong | Chiều muộn | Có |
| Âm u | Vừa | Vừa | Không gió | Trong | Tối | Có |
| Mưa lớn | Lạnh | Vừa | Gió nhẹ | Hơi đục | Sáng sớm | Không |
| Âm u | Vừa | Cao | Gió nhẹ | Trong | Sáng sớm | Có |
| Nắng gắt | Lạnh | Vừa | Gió nhẹ | Hơi đục | Tối | Có |
| Mưa nhẹ | Nóng | Vừa | Gió nhẹ | Quá đục | Sáng sớm | Không |
| Mát mẻ | Vừa | Cao | Không gió | Quá đục | Sáng sớm | Không |
| Mưa nhẹ | Nóng | Vừa | Gió nhẹ | Hơi đục | Tối | Có |
| Âm u | Lạnh | Cao | Gió nhẹ | Hơi đục | Chiều muộn | Có |
| Mát mẻ | Nóng | Thấp | Gió mạnh | Hơi đục | Chiều muộn | Không |
| Nắng gắt | Nóng | Vừa | Không gió | Trong | Chiều muộn | Có |
| Âm u | Nóng | Vừa | Không gió | Hơi đục | Khác | Có |
| Nắng gắt | Vừa | Vừa | Gió mạnh | Hơi đục | Sáng sớm | Không |
| Mát mẻ | Lạnh | Cao | Gió mạnh | Quá đục | Khác | Không |
| Mưa nhẹ | Lạnh | Thấp | Gió nhẹ | Quá đục | Tối | Không |
| Mát mẻ | Nóng | Thấp | Gió nhẹ | Trong | Khác | Có |
| Nắng gắt | Lạnh | Cao | Gió mạnh | Trong | Chiều muộn | Không |
| Âm u | Nóng | Vừa | Gió nhẹ | Hơi đục | Khác | Có |
| Mưa lớn | Nóng | Vừa | Gió nhẹ | Hơi đục | Chiều muộn | Không |
| Mát mẻ | Nóng | Thấp | Gió nhẹ | Quá đục | Tối | Không |
| Mát mẻ | Vừa | Cao | Gió mạnh | Hơi đục | Sáng sớm | Không |
| Mưa nhẹ | Vừa | Vừa | Gió mạnh | Hơi đục | Khác | Không |
| Mát mẻ | Vừa | Thấp | Gió mạnh | Trong | Tối | Có |
| Mưa nhẹ | Vừa | Cao | Gió mạnh | Quá đục | Khác | Không |
| Mát mẻ | Vừa | Cao | Không gió | Hơi đục | Sáng sớm | Có |
| Nắng gắt | Lạnh | Thấp | Không gió | Trong | Chiều muộn | Có |
| Mát mẻ | Nóng | Vừa | Không gió | Hơi đục | Chiều muộn | Có |
| Âm u | Vừa | Cao | Gió nhẹ | Hơi đục | Chiều muộn | Có |
| Mưa nhẹ | Nóng | Vừa | Gió nhẹ | Trong | Sáng sớm | Có |
| Nắng gắt | Lạnh | Thấp | Gió mạnh | Trong | Sáng sớm | Có |
| Âm u | Lạnh | Vừa | Gió nhẹ | Trong | Tối | Có |
| Mưa lớn | Lạnh | Vừa | Không gió | Trong | Tối | Không |
| Mát mẻ | Vừa | Cao | Không gió | Hơi đục | Chiều muộn | Có |
| Âm u | Lạnh | Cao | Không gió | Hơi đục | Khác | Có |
| Mưa nhẹ | Vừa | Thấp | Gió mạnh | Hơi đục | Tối | Không |
| Mưa nhẹ | Nóng | Thấp | Gió mạnh | Hơi đục | Chiều muộn | Không |
| Mưa nhẹ | Vừa | Vừa | Không gió | Hơi đục | Khác | Có |
| Mưa nhẹ | Vừa | Vừa | Gió nhẹ | Hơi đục | Tối | Có |
| Mưa nhẹ | Nóng | Cao | Không gió | Quá đục | Chiều muộn | Không |
| Mưa lớn | Nóng | Vừa | Gió nhẹ | Trong | Sáng sớm | Không |
| Mát mẻ | Nóng | Cao | Gió nhẹ | Quá đục | Tối | Không |
| Mưa nhẹ | Vừa | Cao | Không gió | Trong | Chiều muộn | Có |
| Mưa lớn | Nóng | Cao | Gió mạnh | Quá đục | Tối | Không |
| Âm u | Nóng | Cao | Không gió | Trong | Tối | Không |
| Mát mẻ | Lạnh | Thấp | Gió mạnh | Trong | Khác | Có |
| Nắng gắt | Lạnh | Thấp | Gió nhẹ | Hơi đục | Khác | Có |
| Mưa nhẹ | Nóng | Cao | Không gió | Hơi đục | Chiều muộn | Không |
| Mưa lớn | Vừa | Vừa | Gió mạnh | Trong | Sáng sớm | Không |
| Mát mẻ | Nóng | Vừa | Gió nhẹ | Hơi đục | Chiều muộn | Có |
| Âm u | Nóng | Thấp | Gió nhẹ | Hơi đục | Khác | Có |
| Âm u | Lạnh | Vừa | Gió nhẹ | Hơi đục | Chiều muộn | Có |
| Mát mẻ | Lạnh | Cao | Gió nhẹ | Quá đục | Chiều muộn | Không |
| Nắng gắt | Nóng | Cao | Không gió | Trong | Chiều muộn | Không |
| Mưa nhẹ | Lạnh | Thấp | Gió nhẹ | Trong | Chiều muộn | Có |
| Mưa lớn | Lạnh | Thấp | Gió nhẹ | Quá đục | Chiều muộn | Không |
| Nắng gắt | Lạnh | Vừa | Gió mạnh | Trong | Sáng sớm | Có |
| Âm u | Vừa | Cao | Gió nhẹ | Trong | Chiều muộn | Có |
| Mưa nhẹ | Lạnh | Thấp | Gió mạnh | Hơi đục | Chiều muộn | Không |
| Nắng gắt | Vừa | Vừa | Gió nhẹ | Hơi đục | Tối | Có |
| Nắng gắt | Nóng | Cao | Gió mạnh | Trong | Sáng sớm | Không |
| Mưa nhẹ | Vừa | Thấp | Gió mạnh | Trong | Chiều muộn | Có |
| Nắng gắt | Nóng | Thấp | Gió mạnh | Trong | Tối | Không |
| Mưa lớn | Lạnh | Cao | Gió mạnh | Trong | Tối | Không |
| Nắng gắt | Lạnh | Vừa | Không gió | Hơi đục | Chiều muộn | Có |
| Âm u | Vừa | Vừa | Không gió | Hơi đục | Tối | Có |
| Nắng gắt | Vừa | Thấp | Gió nhẹ | Trong | Sáng sớm | Có |
| Âm u | Lạnh | Thấp | Không gió | Quá đục | Khác | Không |
| Âm u | Nóng | Thấp | Không gió | Trong | Khác | Có |
| Mưa lớn | Lạnh | Thấp | Gió mạnh | Trong | Chiều muộn | Không |
| Âm u | Lạnh | Thấp | Gió mạnh | Trong | Sáng sớm | Có |
| Mưa lớn | Nóng | Thấp | Không gió | Trong | Tối | Không |
| Mưa nhẹ | Nóng | Cao | Không gió | Trong | Khác | Không |
| Mưa nhẹ | Nóng | Cao | Gió mạnh | Quá đục | Chiều muộn | Không |
| Mưa nhẹ | Vừa | Vừa | Gió mạnh | Quá đục | Tối | Không |
| Mát mẻ | Vừa | Thấp | Gió mạnh | Trong | Sáng sớm | Có |
| Mát mẻ | Vừa | Thấp | Không gió | Hơi đục | Sáng sớm | Có |
| Mưa lớn | Vừa | Vừa | Không gió | Quá đục | Chiều muộn | Không |
| Âm u | Vừa | Thấp | Gió nhẹ | Trong | Tối | Có |
| Nắng gắt | Nóng | Vừa | Gió nhẹ | Trong | Tối | Có |
| Âm u | Lạnh | Cao | Không gió | Trong | Chiều muộn | Có |
| Nắng gắt | Vừa | Vừa | Gió mạnh | Hơi đục | Chiều muộn | Không |
| Mưa lớn | Nóng | Cao | Gió nhẹ | Hơi đục | Khác | Không |
| Mưa lớn | Vừa | Cao | Gió nhẹ | Quá đục | Tối | Không |
| Nắng gắt | Nóng | Thấp | Gió mạnh | Hơi đục | Chiều muộn | Không |
| Nắng gắt | Lạnh | Thấp | Gió mạnh | Trong | Khác | Có |
| Mưa lớn | Nóng | Cao | Gió nhẹ | Trong | Sáng sớm | Không |
| Nắng gắt | Nóng | Thấp | Không gió | Hơi đục | Chiều muộn | Có |
| Mát mẻ | Lạnh | Vừa | Gió mạnh | Hơi đục | Khác | Không |
| Nắng gắt | Nóng | Thấp | Gió nhẹ | Trong | Chiều muộn | Có |
| Mưa lớn | Lạnh | Vừa | Gió nhẹ | Trong | Khác | Không |
| Mưa nhẹ | Nóng | Cao | Gió mạnh | Hơi đục | Chiều muộn | Không |
| Mát mẻ | Lạnh | Cao | Không gió | Hơi đục | Tối | Có |
| Nắng gắt | Vừa | Vừa | Gió nhẹ | Quá đục | Khác | Không |
| Mưa nhẹ | Nóng | Vừa | Gió mạnh | Trong | Khác | Không |
| Nắng gắt | Vừa | Thấp | Không gió | Quá đục | Khác | Không |
| Âm u | Lạnh | Vừa | Gió nhẹ | Hơi đục | Khác | Có |
| Mưa nhẹ | Vừa | Vừa | Gió mạnh | Trong | Sáng sớm | Có |
| Mưa lớn | Nóng | Thấp | Gió nhẹ | Trong | Chiều muộn | Không |
| Âm u | Lạnh | Thấp | Không gió | Hơi đục | Chiều muộn | Có |
| Nắng gắt | Vừa | Vừa | Gió nhẹ | Quá đục | Chiều muộn | Không |

### Tập dữ liệu test

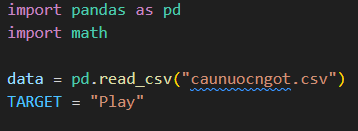
Bảng 2.2: Bảng dữ liệu test

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Outlook | Temperature | Humidity | Wind | WaterClarity | Time | Play |
| Mát mẻ | Vừa | Vừa | Gió nhẹ | Trong | Sáng sớm | Có |
| Mát mẻ | Vừa | Thấp | Gió nhẹ | Hơi đục | Chiều muộn | Có |
| Âm u | Vừa | Cao | Không gió | Trong | Sáng sớm | Có |
| Âm u | Nóng | Vừa | Gió nhẹ | Hơi đục | Chiều muộn | Có |
| Mưa nhẹ | Vừa | Cao | Gió nhẹ | Hơi đục | Sáng sớm | Có |
| Mưa nhẹ | Nóng | Vừa | Gió nhẹ | Trong | Chiều muộn | Có |
| Nắng gắt | Nóng | Thấp | Gió nhẹ | Trong | Sáng sớm | Có |
| Nắng gắt | Vừa | Vừa | Không gió | Hơi đục | Chiều muộn | Có |
| Âm u | Vừa | Vừa | Không gió | Hơi đục | Khác | Có |
| Mát mẻ | Lạnh | Vừa | Gió nhẹ | Trong | Chiều muộn | Có |
| Mưa lớn | Vừa | Cao | Gió mạnh | Quá đục | Tối | Không |
| Mưa lớn | Nóng | Vừa | Gió mạnh | Hơi đục | Chiều muộn | Không |
| Nắng gắt | Nóng | Cao | Gió mạnh | Trong | Tối | Không |
| Âm u | Lạnh | Cao | Không gió | Quá đục | Tối | Không |
| Mưa nhẹ | Lạnh | Cao | Gió mạnh | Hơi đục | Tối | Không |
| Mát mẻ | Lạnh | Cao | Gió mạnh | Trong | Khác | Không |
| Nắng gắt | Nóng | Vừa | Gió mạnh | Hơi đục | Tối | Không |
| Mưa lớn | Nóng | Cao | Không gió | Quá đục | Chiều muộn | Không |
| Âm u | Lạnh | Vừa | Gió mạnh | Quá đục | Sáng sớm | Không |
| Mưa nhẹ | Vừa | Cao | Không gió | Quá đục | Khác | Không |
| Mát mẻ | Vừa | Thấp | Không gió | Hơi đục | Sáng sớm | Có |
| Âm u | Vừa | Vừa | Gió nhẹ | Trong | Chiều muộn | Có |
| Mưa nhẹ | Vừa | Vừa | Gió nhẹ | Hơi đục | Chiều muộn | Có |
| Nắng gắt | Nóng | Thấp | Không gió | Trong | Sáng sớm | Có |
| Mát mẻ | Nóng | Vừa | Gió nhẹ | Hơi đục | Khác | Có |
| Âm u | Vừa | Thấp | Không gió | Hơi đục | Chiều muộn | Có |
| Mưa nhẹ | Nóng | Vừa | Không gió | Hơi đục | Sáng sớm | Có |
| Nắng gắt | Vừa | Vừa | Gió nhẹ | Hơi đục | Chiều muộn | Có |
| Mát mẻ | Vừa | Vừa | Gió nhẹ | Trong | Sáng sớm | Có |
| Âm u | Vừa | Cao | Gió nhẹ | Hơi đục | Chiều muộn | Có |
| Mưa lớn | Lạnh | Cao | Gió mạnh | Trong | Tối | Không |
| Nắng gắt | Nóng | Cao | Không gió | Hơi đục | Tối | Không |
| Âm u | Lạnh | Thấp | Gió mạnh | Quá đục | Khác | Không |
| Mưa nhẹ | Lạnh | Cao | Gió mạnh | Trong | Tối | Không |
| Mát mẻ | Lạnh | Cao | Không gió | Quá đục | Tối | Không |
| Nắng gắt | Vừa | Cao | Gió mạnh | Hơi đục | Tối | Không |
| Mưa lớn | Nóng | Vừa | Gió mạnh | Trong | Khác | Không |
| Âm u | Lạnh | Cao | Gió mạnh | Hơi đục | Tối | Không |
| Mưa nhẹ | Vừa | Cao | Gió mạnh | Quá đục | Chiều muộn | Không |
| Nắng gắt | Nóng | Vừa | Gió mạnh | Trong | Tối | Không |

# CHƯƠNG 3. CÀI ĐẶT THUẬT TOÁN

## 3.1. Cài đặt thuật toán

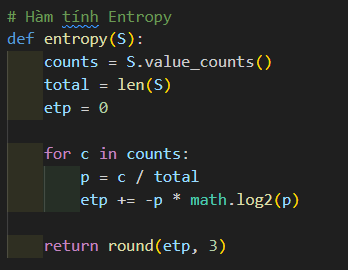
**Bước 1:** Import thư viện và tập dữ liệu:



Hình 3.1: Thư viện cần import và tệp dữ liệu

* pandas (pd): thư viện dùng để đọc, xử lý và phân tích dữ liệu dạng bảng (DataFrame).
* math: cung cấp các hàm toán học cơ bản như log, sqrt, pow…, thường dùng trong các thuật toán như ID3.
* data: biến chứa toàn bộ tập dữ liệu đầu vào phục vụ cho việc huấn luyện và xây dựng cây quyết định.
* TARGET = "Play": xác định cột mục tiêu (nhãn cần dự đoán) trong tập dữ liệu, dùng làm kết quả đầu ra của thuật toán ID3.

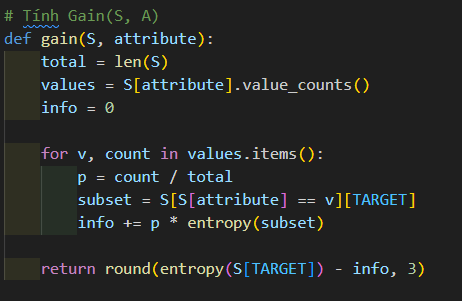
**Bước 2**: Xây dựng hàm tính entropy:



Hình 3.2: Hàm entropy

* .value\_counts() là một phương thức trong thư viện pandas (Python), dùng để đếm số lần xuất hiện của mỗi giá trị trong một cột (Series) hoặc trong dữ liệu.
* Math.log2(x) là một phương thức tính log2(x).
* Round(x,number) là một phương thức chỉ lấy bao nhiêu số sau dấu phẩy của số x.

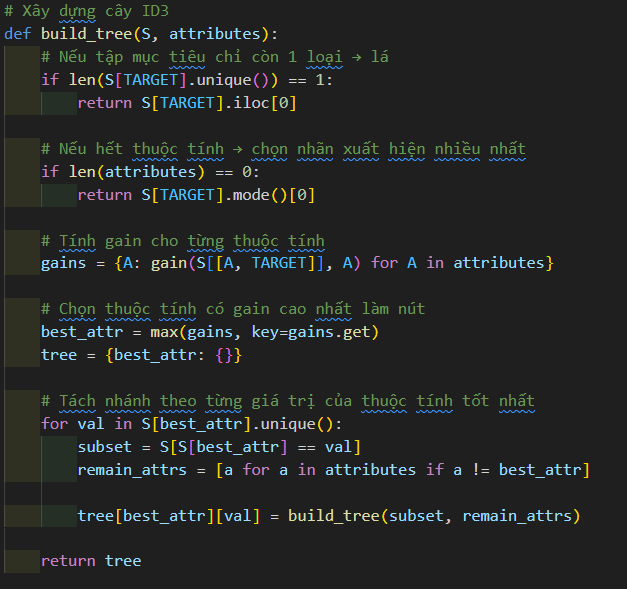
**Bước 3**: Xây dựng hàm tính độ lợi thông tin của 1 thuộc tính:



Hình 3.3: Hàm tính độ lợi thông tin của 1 thuộc tính

* .items() là một phương thức trong Python, thường dùng với dictionary, pandas Series hoặc DataFrame, để lấy từng cặp (key, value) trong quá trình duyệt dữ liệu.
* .value\_counts() là một phương thức trong thư viện pandas (Python), dùng để đếm số lần xuất hiện của mỗi giá trị trong một cột (Series) hoặc trong dữ liệu.
* len(S): dùng để lấy tổng số mẫu (số dòng) trong tập dữ liệu S. Ví dụ: nếu bảng dữ liệu có 14 dòng thì len(S) = 14.
* S[attribute].value\_counts() (pandas): đếm số lần xuất hiện của mỗi giá trị trong cột attribute. Ví dụ: cột Outlook có Sunny = 5, Rain = 4, Overcast = 5.
* values.items(): dùng để duyệt từng cặp giá trị và số lần xuất hiện tương ứng trong attribute. Ví dụ: (Sunny, 5), (Rain, 4), (Overcast, 5).
* p = count / total: tính tỷ lệ số mẫu của một giá trị attribute so với toàn bộ tập dữ liệu. Ví dụ: Sunny có 5 mẫu trên 14 mẫu thì p = 5/14.
* S[S[attribute] == v][TARGET]: lọc ra các dòng có giá trị attribute bằng v, sau đó lấy cột TARGET để thu được tập nhãn tương ứng. Ví dụ: với Outlook = Sunny thì lấy các nhãn PlayTennis của những dòng có Sunny.
* entropy(subset): tính entropy của tập con sau khi chia theo giá trị v của attribute, phản ánh mức độ hỗn loạn của nhãn trong nhánh đó.
* info += p \* entropy(subset): cộng dồn entropy của từng nhánh sau khi nhân với trọng số p, tạo thành entropy trung bình sau khi chia theo attribute.
* entropy(S[TARGET]): entropy ban đầu của toàn bộ tập dữ liệu trước khi chia, chỉ xét cột nhãn TARGET.
* entropy(S[TARGET]) − info: chính là Information Gain, biểu thị mức độ giảm entropy khi chia dữ liệu theo thuộc tính attribute.
* round(..., 3): làm tròn giá trị Gain đến 3 chữ số thập phân để thuận tiện cho việc trình bày và so sánh.

**Bước 4**: Xây dựng cây quyết định:



Hình 3.4: Xây dựng cây quyết định

* .unique() là một phương thức trong pandas, dùng để lấy ra các giá trị khác nhau (không trùng lặp) trong một cột dữ liệu (Series).
* .iloc[0] là một phương thức trong pandas, dùng để lấy giá trị đầu tiên của dữ liệu.
* Cách hoạt động:

+) Hàm nhận hai tham số S và attributes. S là tập dữ liệu sẽ xét, attributes là tập thuộc tính sẽ xét.

+) if len(S[TARGET].unique()) == 1: return S[TARGET].iloc[0]

Nếu Tập mục tiêu chỉ còn một giá trị Có hoặc Không thì dữ liệu đã thuần nhất, nút hiện tại trở thành lá của cây; trả về giá trị dự đoán (lá).

+) if len(attributes) == 0: return S[TARGET].mode()[0]

Nếu Tập thuộc tính(attributes) đã hết nhưng dữ liệu vẫn chưa thuần nhất thì trả về Giá trị xuất hiện nhiều nhất trong cột TARGET.

+) gains = {A: gain(S[[A, TARGET]], A) for A in attributes}

Tính entropy của từng thuộc tính trong tập attributes và lưu vào dict.

Ví dụ: attributes[“Outlook”, “Wind”, “Temperature”]

Sau khi tính được gain thì lưu vào dict: gains={“Outlook”: 0.2, “Wind”: 0.1, “Temperature”: 0.2}

+) best\_attr = max(gains, key=gains.get)

Gán thuộc tính có gain cao nhất vào biến best\_attr.

+) tree = {best\_attr: {}}

Tạo cây với biến best\_attr là nút.

+)  for val in S[best\_attr].unique():

        subset = S[S[best\_attr] == val]

        remain\_attrs = [a for a in attributes if a != best\_attr]

        tree[best\_attr][val] = build\_tree(subset, remain\_attrs)

return tree

Duyệt từng giá trị (gán bởi biến val) có trong nút. Ví dụ nút “Outlook” sẽ có các giá trị như: [“Mưa nhẹ”, “Nắng”, “Mưa lớn”].

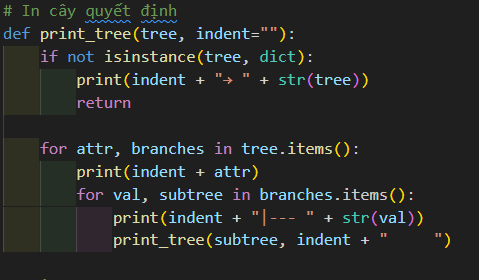
subset: tập con tương ứng của có giá trị (val) trong tập dữ liệu đang xét.

Remain\_attrs: tập dữ liệu đang xét(S) sau khi loại bỏ nút hiện tại.

tree[best\_attr][val] = build\_tree(subset, remain\_attrs): Nhánh val của nút best\_attr tiếp tục tìm nhánh con bằng cách gọi lại hàm build\_tree(subset, remain\_attrs) tức tìm cây tiếp theo với tập dữ liệu subset và tập thuộc tính remain\_attrs.

Sau khi tìm được lá cuối cùng hàm return tree trả về cây quyết định cuối cùng.

**Bước 5**: hàm in cây:

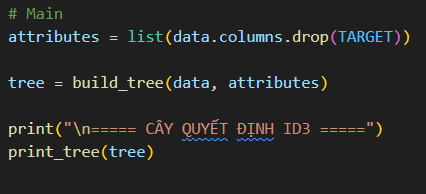


Hình 3.5: Hàm in cây quyết định

Nếu dict hiện tại không chứa 1 dict con, tức là nút hiện tại không chứa nút con( chỉ chứa lá) thì in lá và kết thúc nhánh hiện tại.

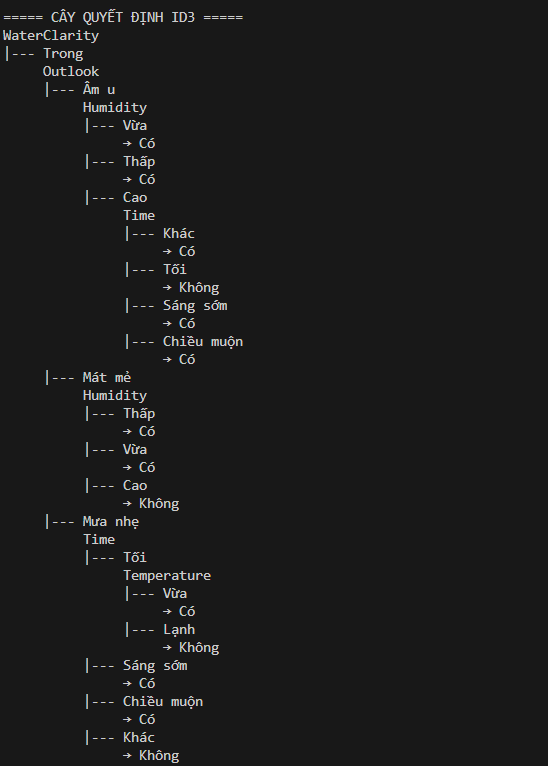
Nếu vẫn chứa nhánh con thì duyệt từng giá trị để gọi hàm tìm nhánh tiếp.

**Bước 6**: Hàm chính:

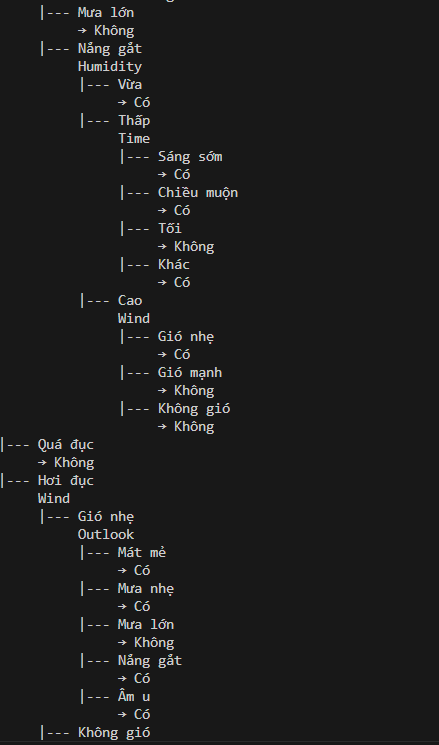


Hình 3.6: Hàm main ( hàm chính )

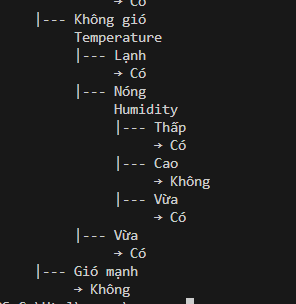
Khi chạy ta thu được:



Hình 3.7: Kết quả chạy code ( ảnh 1)



Hình 3.8: Kết quả chạy code ( ảnh 2)

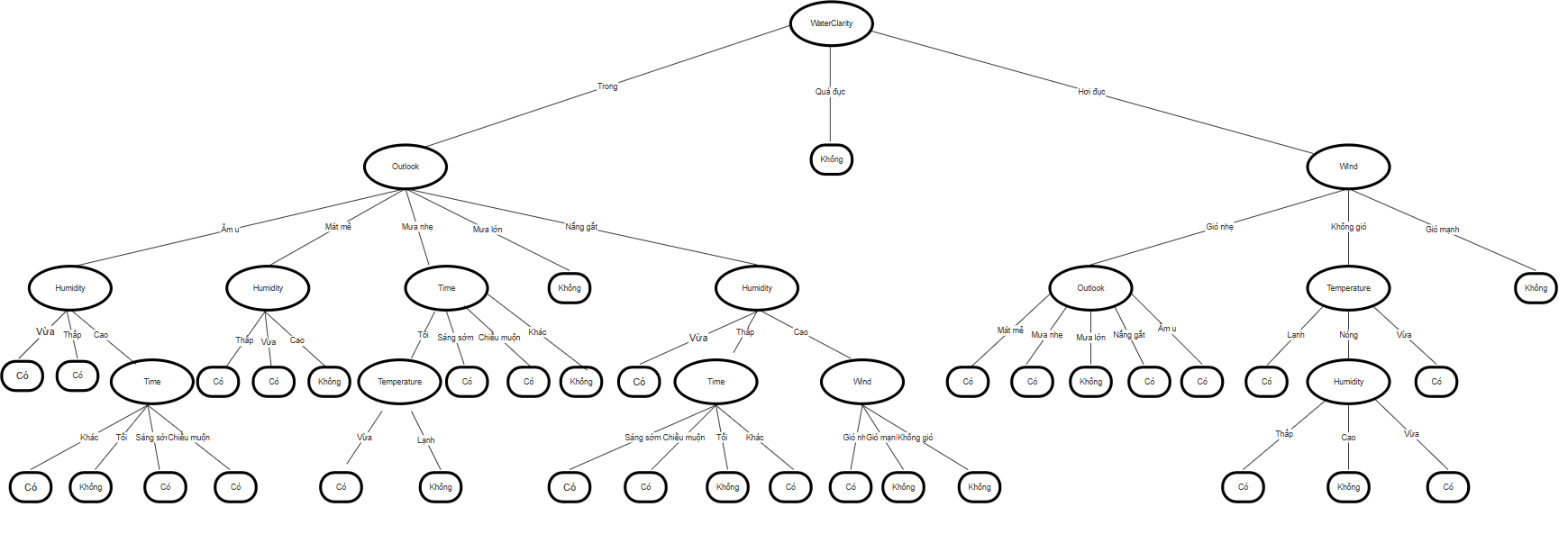


Hình 3.9: Kết quả chạy code ( ảnh 3)

**Bước 7**: Sau khi có được cây phân lớp, ta tiến hành hành xây dựng các bộ luật để phân lớp cho dữ liệu mới.

* Rule 1: If WaterClarity = "Trong" and Outlook = "Âm u" and Humidity = "Vừa" Then Play = "Có"
* Rule 2: If WaterClarity = "Trong" and Outlook = "Âm u" and Humidity = "Thấp" Then Play = "Có"
* Rule 3: If WaterClarity = "Trong" and Outlook = "Âm u" and Humidity = "Cao" and Time = "Tối" Then Play = "Không"
* Rule 4: If WaterClarity = "Trong" and Outlook = "Âm u" and Humidity = "Cao" and Time = "Sáng sớm" Then Play = "Có"
* Rule 5: If WaterClarity = "Trong" and Outlook = "Âm u" and Humidity = "Cao" and Time = "Chiều muộn" Then Play = "Có"
* Rule 6: If WaterClarity = "Trong" and Outlook = "Âm u" and Humidity = "Cao" and Time = "Khác" Then Play = "Có"
* Rule 7: If WaterClarity = "Trong" and Outlook = "Mát mẻ" and Humidity = "Thấp" Then Play = "Có"
* Rule 8: If WaterClarity = "Trong" and Outlook = "Mát mẻ" and Humidity = "Vừa" Then Play = "Có"
* Rule 9: If WaterClarity = "Trong" and Outlook = "Mát mẻ" and Humidity = "Cao" Then Play = "Không"
* Rule 10: If WaterClarity = "Trong" and Outlook = "Mưa nhẹ" and Time = "Tối" and Temperature = "Vừa" Then Play = "Có"
* Rule 11: If WaterClarity = "Trong" and Outlook = "Mưa nhẹ" and Time = "Tối" and Temperature = "Lạnh" Then Play = "Không"
* Rule 12: If WaterClarity = "Trong" and Outlook = "Mưa nhẹ" and Time = "Sáng sớm" Then Play = "Có"
* Rule 13: If WaterClarity = "Trong" and Outlook = "Mưa nhẹ" and Time = "Chiều muộn" Then Play = "Có"
* Rule 14: If WaterClarity = "Trong" and Outlook = "Mưa nhẹ" and Time = "Khác" Then Play = "Không"
* Rule 15: If WaterClarity = "Trong" and Outlook = "Mưa lớn" Then Play = "Không"
* Rule 16: If WaterClarity = "Trong" and Outlook = "Nắng gắt" and Humidity = "Vừa" Then Play = "Có"
* Rule 17: If WaterClarity = "Trong" and Outlook = "Nắng gắt" and Humidity = "Thấp" and Time = "Sáng sớm" Then Play = "Có"
* Rule 18: If WaterClarity = "Trong" and Outlook = "Nắng gắt" and Humidity = "Thấp" and Time = "Chiều muộn" Then Play = "Có"
* Rule 19: If WaterClarity = "Trong" and Outlook = "Nắng gắt" and Humidity = "Thấp" and Time = "Tối" Then Play = "Không"
* Rule 20: If WaterClarity = "Trong" and Outlook = "Nắng gắt" and Humidity = "Thấp" and Time = "Khác" Then Play = "Có"
* Rule 21: If WaterClarity = "Trong" and Outlook = "Nắng gắt" and Humidity = "Cao" and Wind = "Gió nhẹ" Then Play = "Có"
* Rule 22: If WaterClarity = "Trong" and Outlook = "Nắng gắt" and Humidity = "Cao" and Wind = "Gió mạnh" Then Play = "Không"
* Rule 23: If WaterClarity = "Trong" and Outlook = "Nắng gắt" and Humidity = "Cao" and Wind = "Không gió" Then Play = "Không"
* Rule 24: If WaterClarity = "Quá đục" Then Play = "Không"
* Rule 25: If WaterClarity = "Hơi đục" and Wind = "Gió nhẹ" and Outlook = "Mát mẻ" Then Play = "Có"
* Rule 26: If WaterClarity = "Hơi đục" and Wind = "Gió nhẹ" and Outlook = "Mưa nhẹ" Then Play = "Có"
* Rule 27: If WaterClarity = "Hơi đục" and Wind = "Gió nhẹ" and Outlook = "Mưa lớn" Then Play = "Không"
* Rule 28: If WaterClarity = "Hơi đục" and Wind = "Gió nhẹ" and Outlook = "Nắng gắt" Then Play = "Có"
* Rule 29: If WaterClarity = "Hơi đục" and Wind = "Gió nhẹ" and Outlook = "Âm u" Then Play = "Có"
* Rule 30: If WaterClarity = "Hơi đục" and Wind = "Không gió" and Temperature = "Lạnh" Then Play = "Có"
* Rule 31: If WaterClarity = "Hơi đục" and Wind = "Không gió" and Temperature = "Nóng" and Humidity = "Thấp" Then Play = "Có"
* Rule 32: If WaterClarity = "Hơi đục" and Wind = "Không gió" and Temperature = "Nóng" and Humidity = "Cao" Then Play = "Không"
* Rule 33: If WaterClarity = "Hơi đục" and Wind = "Không gió" and Temperature = "Nóng" and Humidity = "Vừa" Then Play = "Có"
* Rule 34: If WaterClarity = "Hơi đục" and Wind = "Không gió" and Temperature = "Vừa" Then Play = "Có"
* Rule 35: If WaterClarity = "Hơi đục" and Wind = "Gió mạnh" Then Play = "Không"

**Bước 8**: Vẽ cây quyết định



Hình 3.100: Vẽ cây quyết định

**Bước 9:** Tối ưu lại bộ luật để ứng dụng xây dựng website

* Rule 1: If WaterClarity = "Trong" and Outlook = "Âm u" and Humidity = "Cao" and Time = "Tối" Then Play = "Không"
* Rule 2: If WaterClarity = "Trong" and Outlook = "Mát mẻ" and Humidity = "Cao" Then Play = "Không"
* Rule 3: If WaterClarity = "Trong" and Outlook = "Mưa nhẹ" and Time = "Tối" and Temperature = "Lạnh" Then Play = "Không"
* Rule 4: If WaterClarity = "Trong" and Outlook = "Mưa nhẹ" and Time = "Khác" Then Play = "Không"
* Rule 5: If WaterClarity = "Trong" and Outlook = "Mưa lớn" Then Play = "Không"
* Rule 6: If WaterClarity = "Trong" and Outlook = "Nắng gắt" and Humidity = "Thấp" and Time = "Tối" Then Play = "Không"
* Rule 7: If WaterClarity = "Trong" and Outlook = "Nắng gắt" and Humidity = "Cao" and Wind ≠ "Gió nhẹ" Then Play = "Không"
* Rule 8: If WaterClarity = "Quá đục" Then Play = "Không"
* Rule 9: If WaterClarity = "Hơi đục" and Wind = "Gió nhẹ" and Outlook = "Mưa lớn" Then Play = "Không"
* Rule 10: If WaterClarity = "Hơi đục" and Wind = "Không gió" and Temperature = "Nóng" and Humidity = "Cao" Then Play = "Không"
* Rule 11: If WaterClarity = "Hơi đục" and Wind = "Gió mạnh" Then Play = "Không"

**Bước 10:** Xây dựng hàm dự đoán để đánh giá dữ liệu trên tập test

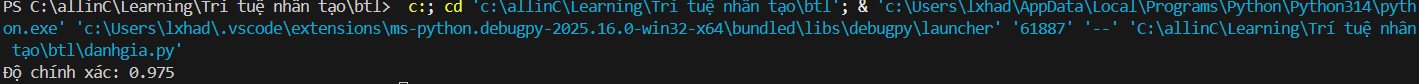




A screen shot of a computer program

AI-generated content may be incorrect.

Hình 3.11: Hàm dự đoán, đánh giá



Hình 3.12: Kết quả đánh giá

**Bước 11**: Xây dựng website

Phần bảng bên phải cho phép người dùng nhập các thuộc tính liên quan đến hoạt động đi câu cá như nhiệt độ, độ ẩm, độ trong của nước,… Dựa trên các thông tin này, hệ thống sẽ tiến hành dự đoán và trả về kết quả sau khi người dùng nhấn nút “Dự đoán”.

A screenshot of a phone

AI-generated content may be incorrect.

Hình 3.13: Giao diện website và kết quả dự đoán

## 3.2. Kết luận và hướng phát triển

### 3.2.1. Đánh giá mô hình

#### 3.2.1.1. Kết quả đạt được

Mô hình cây quyết định ID3 đã được xây dựng thành công nhằm dự đoán khả năng đi câu cá (Có / Không) dựa trên các thuộc tính môi trường như: nhiệt độ, độ ẩm, gió, độ trong của nước, thời điểm và tình trạng thời tiết.

Kết quả thực nghiệm cho thấy:

* Mô hình hoạt động ổn định trên tập dữ liệu nhỏ, các thuộc tính đều ở dạng rời rạc.
* Cây quyết định sinh ra tập luật rõ rang mặc dù tập luật có hơi lớn(35 mẫu), tuy nhiên có thể chuyển trực tiếp thành các câu lệnh điều kiện (if–else) trong chương trình.
* Các quyết định “Không đi câu” thường xuất phát từ những điều kiện môi trường bất lợi (nước quá đục, gió mạnh, mưa lớn…), phù hợp với thực tế.

#### 3.2.1.2. Ưu điểm của mô hình ID3

* Dễ hiểu và dễ giải thích: mỗi nhánh trong cây tương ứng với một điều kiện cụ thể, thuận tiện cho việc diễn giải kết quả.
* Triển khai đơn giản: mô hình có thể được cài đặt nhanh chóng bằng các cấu trúc điều kiện, không yêu cầu thư viện phức tạp.
* Phù hợp cho mục đích học tập: giúp người học nắm rõ nguyên lý xây dựng cây quyết định và cách sinh tập luật.
* Hiệu quả với dữ liệu rời rạc: ID3 xử lý tốt các thuộc tính phân loại như “Nắng gắt”, “Mưa nhẹ”, “Gió mạnh”,…

#### 3.2.1.3. Hạn chế của mô hình

* Tập luật sinh ra nhiều và cồng kềnh: cây quyết định ban đầu tạo ra tới 35 luật, gây khó khăn cho việc quản lý và bảo trì.
* Dễ bị overfitting: mô hình học sát dữ liệu huấn luyện, kém tổng quát khi gặp dữ liệu mới.
* Không xử lý tốt dữ liệu liên tục: các giá trị số (nhiệt độ, độ ẩm, gió) phải được rời rạc hóa thủ công.
* Nhạy cảm với nhiễu dữ liệu: chỉ cần thay đổi nhỏ trong tập huấn luyện có thể làm cấu trúc cây thay đổi đáng kể.

#### 3.2.1.4. Đánh giá việc rút gọn tập luật

Trong quá trình triển khai, tập luật đã được rút gọn chỉ giữ các luật dẫn đến kết quả “Không”, các trường hợp còn lại mặc định là “Có”.

Cách tiếp cận này mang lại:

* Giảm đáng kể độ phức tạp của mã nguồn.
* Giữ được các quyết định quan trọng (điều kiện không nên đi câu).
* Tăng khả năng đọc hiểu và bảo trì chương trình.

Tuy nhiên, việc rút gọn có thể làm giảm độ chính xác trong một số trường hợp biên, do bỏ qua các luật chi tiết dẫn đến kết quả “Có”.

### 3.2.2. Kết luận

Ưu điểm, trong phạm vi bài tập lớn, mô hình cây quyết định được xây dựng dựa trên thuật toán ID3 đã đạt được mục tiêu đề ra là phân lớp dữ liệu một cách trực quan và dễ hiểu. Thông qua việc sử dụng độ đo Entropy và Information Gain, mô hình có khả năng lựa chọn thuộc tính phù hợp để phân chia dữ liệu, từ đó đưa ra các quyết định phân lớp rõ ràng.

Kết quả thực nghiệm cho thấy mô hình hoạt động hiệu quả trên tập dữ liệu có kích thước nhỏ và các thuộc tính rời rạc. Cấu trúc cây quyết định cho phép dễ dàng theo dõi quá trình ra quyết định và giải thích kết quả, điều này đặc biệt có ý nghĩa trong các bài toán cần tính minh bạch và khả năng diễn giải mô hình.

Nhược điểm, mô hình vẫn tồn tại một số hạn chế như xu hướng overfitting, khả năng xử lý kém với dữ liệu nhiễu và chưa tối ưu khi làm việc với các thuộc tính liên tục. Đặc biệt khi giá trị của các thuộc tính càng tăng thì độ phức tạp của mô hình cũng gia tăng làm cho kích thước cây quyết định lớn hơn và hiệu quả của thuật toán không được cao. Sau khi có cây thì việc xây dựng tập luật để phân lớp cho tập mẫu tiếp theo rất phức tạp.

Nhìn chung, thuật toán ID3 phù hợp cho mục đích học tập và nghiên cứu cơ bản, giúp người học nắm vững nguyên lý xây dựng cây quyết định. Tuy chưa đáp ứng tốt các yêu cầu của bài toán dữ liệu lớn trong thực tế, mô hình vẫn hoàn thành tốt vai trò minh họa và đạt được yêu cầu của bài tập lớn.

## 3.3. Hướng phát triển

Trong thời gian tới, mô hình có thể được cải tiến và phát triển theo nhiều hướng nhằm nâng cao hiệu quả và khả năng ứng dụng thực tế. Trước hết, có thể áp dụng các thuật toán cây quyết định cải tiến như C4.5 hoặc CART để khắc phục hạn chế của ID3, đặc biệt là vấn đề thiên lệch thuộc tính và overfitting. Việc bổ sung cơ chế cắt tỉa cây sẽ giúp mô hình có khả năng tổng quát hóa tốt hơn.

Bên cạnh đó, mô hình có thể được mở rộng để xử lý các thuộc tính liên tục bằng cách tự động xác định ngưỡng phân chia phù hợp, đồng thời bổ sung các kỹ thuật tiền xử lý dữ liệu như xử lý giá trị thiếu và giảm nhiễu. Ngoài ra, việc đánh giá mô hình bằng các chỉ số như độ chính xác (Accuracy), độ thu hồi (Recall) và độ chính xác dự đoán (Precision) trên nhiều tập dữ liệu khác nhau cũng là một hướng phát triển quan trọng.

Cuối cùng, mô hình có thể được tích hợp vào các hệ thống ứng dụng thực tế hoặc kết hợp với các phương pháp học máy khác như Random Forest hay Boosting để nâng cao độ chính xác và tính ổn định của kết quả phân lớp.

# TÀI LIỆU THAM KHẢO

[1]. Trần Hùng Cường, Nguyễn Phương Nga, 2014, *Giáo trình trí tuệ nhân*

*tạo*, Trường Đại học Công nghiệp Hà Nội, NXB thống kê.

[2]. https://scikit-learn.org/stable/modules/tree.html