|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **BỘ CÔNG NGHIỆP**  **TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP HÀ NỘI**  **KHOA CÔNG NGHỆ THÔNG TIN**  **======\*\*\*======**  **Logo%20HaUI%20ban%20chuan**  **BÀI TẬP LỚN HỆ CHUYỂN GIA**  **XÂY DỰNG HỆ CHUYỂN GIA DỰ ĐOÁN TĂNG TRƯỞNG GIÁ VÀNG**   |  |  | | --- | --- | | **Giáo viên HD:** | **Th.S Lê Thị Thủy** | | **Lớp :**  **Nhóm :**  **Thành viên:** | **ĐH – HTTT1 – K11**  **14**  **Đỗ Mạnh Quang – 1141260081** | |  | **Trần Thị Thảo Nguyên – 1141260048** | |  |  | |  |  | |  |  | |  |  |   *Hà nội, Năm 2019* |

# LỜI CAM ĐOAN

Nhóm em bao gồm Đỗ Mạnh Quang và Trần Thị Thảo Nguyên, sinh viên lớp Hệ Thống Thông Tin 1 K11 xin cam đoan bài tập lớn này được viết bởi nhóm em dưới sự hướng dẫn của Cô giáo, Th.S Lê Thị Thủy. Tất cả các kết quả đạt được trong bài tập lớn này là quá trình tìm hiểu, nghiên cứu của riêng nhóm em. Trong toàn bộ nội dung của bài tập lớn, những điều được trình bày là kết quả của nhóm em hoặc là được tổng hợp từ nhiều nguồn tài liệu khác. Các tài liệu tham khảo đều có xuất xứ rõ ràng và được trích dẫn hợp pháp.

Nhóm em xin hoàn toàn chịu trách nhiệm và chịu mọi hình thức kỷ luật theo quy định cho lời cam đoan của mình.

Hà Nội, ngày ..09.. tháng ..08.. năm 2019

Người cam đoan Người cam đoan

Quang Nguyên

Đỗ Mạnh Quang Trần Thị Thảo Nguyên

# LỜI CẢM ƠN

Lời đầu tiên, nhóm em xin bày tỏ sự cảm ơn chân thành đối với Cô giáo Th.S Lê Thị Thủy – giáo viên hướng dẫn trực tiếp của nhóm em. Cô Thủy đã cho nhóm em những gợi ý và chỉ dẫn quý báu trong quá trình nghiên cứu và hoàn thiện bài tập lớn.

Nhóm em cũng xin gửi lời cảm ơn tới các thầy cô trong khoa Công nghệ thông tin, trường Đại học Công Nghiệp Hà Nội đã hướng dẫn, chỉ bảo và tạo điều kiện cho chúng em học tập và nghiên cứu tại trường trong suốt thời gian qua.

Mặc dù đã cố gắng hoàn thành bài tập lớn nhưng chắc chắn sẽ không tránh khỏi những sai sót, nhóm em kính mong nhận được sự thông cảm và chỉ bảo của cô và các bạn.

Nhóm em xin chân thành cảm ơn!

# MỤC LỤC

Contents

[LỜI CAM ĐOAN 2](#_Toc16286008)

[LỜI CẢM ƠN 3](#_Toc16286009)

[MỤC LỤC 4](#_Toc16286010)

[CHƯƠNG 1 : Tổng quan hệ chuyên gia 6](#_Toc16286011)

[1.1 TỔNG QUAN VỀ HỆ CHUYÊN GIA 6](#_Toc16286012)

[1.1.1 HỆ CHUYÊN GIA LÀ GÌ? 6](#_Toc16286013)

[1.1.2 ĐẶC TRƯNG VÀ ƯU ĐIỂM CỦA HỆ CHUYÊN GIA 7](#_Toc16286014)

[1.1.3 CÁC LĨNH VỰC ỨNG DỤNG CỦA HỆ CHUYÊN GIA 8](#_Toc16286015)

[1.1.4 CẤU TRÚC CỦA HỆ CHUYÊN GIA 8](#_Toc16286016)

[1.2 CƠ SỞ TRI THỨC 10](#_Toc16286017)

[1.2.1 PHÂN BIỆT TRI THỨC VÀ DỮ LIỆU 10](#_Toc16286018)

[1.2.2 PHÂN LOẠI TRI THỨC 12](#_Toc16286019)

[1.2.2.1 TRI THỨC MÔ TẢ 12](#_Toc16286020)

[1.2.2.2 TRI THỨC THỦ TỤC 12](#_Toc16286021)

[1.2.2.3 TRI THỨC ĐIỀU KHIỂN 13](#_Toc16286022)

[1.2.3 CÁC CẤP ĐỘ TRI THỨC 13](#_Toc16286023)

[1.2.3.1 TRI THỨC ĐỘNG PHỤ THUỘC VÀO TÌNH HUỐNG,KHÔNG GIAN VÀ THỜI GIAN 13](#_Toc16286024)

[1.2.3.2 TRI THỨC BẤT ĐỊNH, TRI THỨC KHÔNG ĐẦY ĐỦ 14](#_Toc16286025)

[1.2.4 CÁC PHƯƠNG PHÁP BIỂU DIỄN TRI THỨC 14](#_Toc16286026)

[1.2.4.1 BIỂU DIỄN TRI THỨC BẰNG LUẬT SINH 15](#_Toc16286027)

[1.2.4.2 BIỂU DIỄN TRI THỨC BẰNG LOGIC 16](#_Toc16286028)

[1.2.4.3 BIỂU DIỄN TRI THỨC BẰNG MẠNG NGỮ NGHĨA 17](#_Toc16286029)

[1.2.4.4 BIỂU DIỄN TRI THỨC BẰNG NGÔN NGỮ NHÂN TẠO 18](#_Toc16286030)

[1.2.4.5 BIỂU DIỄN BẰNG BỘ BA LIÊN HỢP O.A.V 18](#_Toc16286031)

[1.2.4.6 BIỂU DIỄN TRI THỨC BẰNG FRAME 19](#_Toc16286032)

[1.3 MÔ TƠ SUY DIỄN 20](#_Toc16286033)

[1.3.1 CƠ CHẾ SUY DIỄN 20](#_Toc16286034)

[1.3.1.1 SUY DIỄN TIẾN 20](#_Toc16286035)

[1.3.1.2 SUY DIỄN LÙI 21](#_Toc16286036)

[1.3.1.3 CƠ CHẾ HỖN HỢP 22](#_Toc16286037)

[1.3.2 CƠ CHẾ ĐIỀU KHIỂN 22](#_Toc16286038)

[1.3.2.1 CHỌN HƯỚNG SUY DIỄN 22](#_Toc16286039)

[1.3.2.2 GIẢI QUYẾT CÁC VẤN ĐỀ CẠNH TRANH 23](#_Toc16286040)

[CHƯƠNG 2 : CÁC THUẬT TOÁN ÁP DỤNG TRONG BÀI TOÁN 26](#_Toc16286041)

[2.1 THUẬT TOÁN ID3 26](#_Toc16286042)

[2.1.1 TƯ TƯỞNG THUẬT TOÁN ID3 26](#_Toc16286043)

[2.1.2 HÀM SỐ ENTROPY 27](#_Toc16286044)

[2.1.3 THUẬT TOÁN ID3 28](#_Toc16286045)

[2.2 VECTO ĐẶC TRƯNG 29](#_Toc16286046)

[2.3 ĐỘ ĐO HỖN LOẠN 29](#_Toc16286047)

[CHƯƠNG 3 : XÂY DỰNG ỨNG DỤNG 30](#_Toc16286048)

[3.1 PHÂN TÍCH BÀI TOÁN 30](#_Toc16286049)

[3.1.1 Nghiên Cứu Các Yếu Tố Ảnh Hướng Đến Giá Vàng 30](#_Toc16286050)

[3.1.2 Bảng Thông Kê Số Liệu 30](#_Toc16286051)

[3.1.3 Sử dụng ID3 giải quyết bài toán với 20 dòng dữ liệu 34](#_Toc16286052)

[3.2 Kết quả cài đặt thuật toán 43](#_Toc16286053)

[CHƯƠNG 4 : Tài liệu tham khảo 46](#_Toc16286054)

# CHƯƠNG 1 : Tổng quan hệ chuyên gia

## TỔNG QUAN VỀ HỆ CHUYÊN GIA

## HỆ CHUYÊN GIA LÀ GÌ?

* Hệ chuyên gia là một hệ thống chương trình máy tính chứa các thông tin, tri thức và các quá trình suy luận về một lĩnh vực cụ thể nào đó để giải quyết các vấn đề khó hoặc hóc búa đòi hỏi sự tinh thông đầy đủ của các chuyên gia con người đối với các giải pháp của họ. Nói một cách khác hệ chuyên gia là dựa trên tri thức của các chuyên gia con người giỏi nhất trong lĩnh vực quan tâm.
* Tri thức của hệ chuyên gia bao gồm các sự kiện và luật. Các sự kiện được cấu thành bởi nhiều thông tin, được thu thập rộng rãi, công khai và được sự đồng tình của các chuyên gia con người trong lĩnh vực. Các luật biểu thị sự quyết đoán chuyên môn của các chuyên gia trong lĩnh vực.
* Mức độ hiệu quả của một hệ chuyên gia phụ thuộc vào kích thước và chất lượng của cơ sở tri thức mà hệ đó có được.
* Mỗi hệ chuyên gia chỉ đặc trưng cho một lĩnh vực, vấn đề nào đó như y học, tài chính, khoa học hay công nghệ, vv…, mà không phải là cho bất cứ một lĩnh vực vấn đề nào.

Ví dụ: Hệ chuyên gia về lĩnh vực y học để phát hiện các căn bệnh lây nhiễm sẽ có nhiều chi thức về một số triệu chứng lây bệnh, lĩnh vực tri thức y học bao gồm các căn bệnh, triệu chứng và chữa trị.

* Hoạt động của một hệ chuyên gia dựa trên tri thức được minh họa như sau:

Người dùng (user)

Hệ thống giao tiếp với người dùng

(user interface)

Cơ sở tri thức (knowledge base)

Máy suy diễn (inference engine)

## ĐẶC TRƯNG VÀ ƯU ĐIỂM CỦA HỆ CHUYÊN GIA

* Hệ chuyên gia có 4 đặc trưng cơ bản:
  + - * Hiệu quả cao: khả năng trả lời với mức độ tinh thông bằng hoặc cao hơn so với chuyên gia con người trong cùng lĩnh vực.
      * Thời gian trả lời thỏa đáng: Thời gian trả lời bằng hoặc nhanh hơn chuyên gia con người để đi đến cùng 1 quyết định – là 1 hệ thời gian thực.
      * Độ tin cậy cao: Không xảy ra sự cố hoặc giảm sút độ tin cậy khi sử dụng.
      * Dễ hiểu: Giải thích các bước suy luận 1 cách dễ hiểu và nhất quán.
* Những ưu điểm của hệ chuyên gia:
* Ngày càng phổ cập: được phát triển không ngừng với hiệu quả sử dụng không thể phủ nhận.
* Giá thành hạ.
* Giảm rủi ro: giúp con người tránh được môi trường rủi ro, nguy hiểm.
* Tính sẵn sàng: có thể hoạt động bất cứ lúc nào.
* Đa lĩnh vực: về nhiều lĩnh vực khác nhau.
* Tin cậy: luôn đảm bảo độ tin cậy khi khai thác.
* Khả năng giải thích: có thể diễn giải các câu trả lời 1 cách rõ ràng và chi tiết.
* Trả lời nhanh: trả lời thời gian thực, khách quan.
* Tính ổn định, suy luận lôgic và đầy đủ.
* Như một người hướng dẫn thông minh.
* Có thể truy cập như một cơ sở dữ liệu thông minh.

## CÁC LĨNH VỰC ỨNG DỤNG CỦA HỆ CHUYÊN GIA

Tính đến thời điểm này, hàng trăm hệ chuyên gia đã được xây dựng và báo cáo thường xuyên trong các tạp chí, sách báo và hội thảo khoa học. Ngoài ra còn có các hệ chuyên gia được sử dụng trong các công ty, các tổ chức quân sự mà không được công bố vì lý do bảo mật. Một số ứng dụng của hệ chuyên gia như:

* Chuẩn đoán: lập luận dựa trên các sự kiện quan sát được.
* Hướng dẫn: dạy học thông minh để SV có thể hỏi các câu hỏi vì sao (why?), như thế nào (how?) và cái gì xảy ra nếu (what if?) giống như hỏi thầy giáo.
* Giải thích: giải thích các dữ liệu thu nhận được.
* Theo dõi kiểm tra: So sánh dữ liệu thu được với dữ liệu chuyên môn để đánh giá hiệu quả.
* Lập kế hoạch: lập kế hoạch sản xuất theo yêu cầu.
* Dự đoán: dự đoán hậu quả nếu một tình huống xảy ra.
* Chữa trị: Đưa ra cách giải quyết một vấn đề.
* Điều khiển: điều khiển 1 quá trình bao gồm: diễn giải, chuẩn đoán, kiểm tra, lập kế hoạch, dự đoán và chữa trị.

## CẤU TRÚC CỦA HỆ CHUYÊN GIA

Một hệ chuyên gia kiểu mẫu gồm các phần cơ bản sau:

Bộ giải thích

Giao diện người dùng

Cơ sở tri thức (các luật)

Máy suy diễn

Bộ nhớ làm việc

Bộ thu nhận tri thức

Lịch công việc

Máy suy diễn

Lịch công việc

Cơ sở tri thức (các luật)

Bộ nhớ làm việc

Bộ thu nhận tri thức

* + - Cơ sở tri thức: gồm các phần tử (đơn vị) tri thức, thường là các luật, được tổ chức như 1 CSDL.
    - Máy suy diễn: suy luận dựa trên luật; xem xét những luật nào thỏa mãn các sự kiện, các đối tượng; chọn luật có ưu tiên cao nhất.
    - Lịch công việc (agenda): danh sách các luật ưu tiên do máy suy diễn tạo ra thỏa mãn các sự kiện, đối tượng có mặt trong bộ nhớ làm việc.
    - Bộ nhớ làm việc (working memory): chứa các sự kiện phục vụ cho các luật.
    - Bộ giải thích (explanation facility): giải thích cách suy luận của hệ thống cho người sử dụng.
    - Bộ thu nhận tri thức: Cho phép người dùng bổ sung tri thức vào hệ thống một cách tự động thay vì biểu diễn một cách tường minh. Đây là bộ phận thường có của nhiều HCG.
    - Giao diện người dùng: là nơi người dùng và HCG trao đổi với nhau.

Để thực hiện được các công việc của các thành phần trên trong cấu trúc hệ chuyên gia phải có một hệ điều khiển và quản lý việc học tập, tích lũy tri thức cho lĩnh vực heeh đảm nhận gọi là “Hệ quản trị cơ sở tri thức”. Hệ quản trị cơ sở tri thức thực chất là quản lý và điều khiển công việc của bộ thu nạp tri thức, bộ giải thích, mô tơ suy diễn. Nó phải đảm bảo các yêu cầu:

* Giảm dư thừa tri thức, dữ liệu
* Tính nhất quán và phi mâu thuẫn của tri thức
* Tính toàn vẹn và an toàn
* Giải quyết các vấn đề cạnh tranh
* Chuyển đổi tri thức
* Ngôn ngữ xử lý tri thức

## CƠ SỞ TRI THỨC

## PHÂN BIỆT TRI THỨC VÀ DỮ LIỆU

Chúng ta có thể dựa vào một số đặc trưng sau để phân biệt quy ước tri thức và dữ liệu:

* Khả năng tự giải thích nội dung: Dữ liệu đưa vào máy tính không tự giải thích nổi, đôi khi được mã hóa cho ngắn gọn để dễ cài đặt trong máy. Chỉ có người lập trình đó mới có thể hiểu được nội dung, ý nghĩa của dữ liệu, nhưng tri thức có thể tự giải thích nội dung của mình với người sử dụng bất kỳ.
* Tính cấu trúc: Một trong những đặc tính cơ bản của hoạt động nhận thức của con người đối với thế giwois xung quanh là khả năng phân tích cấu trúc của các đối tượng. Tri thức được đưa vào máy cũng cần có khả năng tạo ra được một sự phân cấp giữa các khái niệm và mối quan hệ giữa chúng.
* Tính liên hệ: Ngoài các quan hệ về cấu trúc trong mỗi tri thức (khái niệm, quá trình, hiện tượng, sự kiện) giữa các đơn vị tri thức còn có nhiều mối liên hệ khác (không gian, thời gian, nhân quả,…). Một số nghiên cứu đã chỉ ra số các liên hệ cơ bản giữa các sự kiện xấp xỉ 200 lần. Một cơ sở tri thức được kết hợp với số liên hệ cơ bản này có thể mô tả và biểu diễn được hầu hết mọi vấn đề mà chúng ta quan tâm.
* Tính chủ động:
* Như chúng ta đã thấy, dữ liệu có vai trò bị động vì nó phụ thuộc vào sự khai thác của chương trình cụ thể.
* Trong xã hội loài người khi hoạt động bất kỳ ở đâu và ở trong lĩnh vực nào thì con người bao giờ cũng bị điều khiển bằng chính tri thức (vốn hiểu biết) của mình. Nhờ có tri thức mà con người đã hình thành mục tiêu và các hành vi để thực hiện mục tiêu đó. Quá trình này luôn đi kèm với sự bổ sung tri thức và khắc phục sự mâu thuẫn giữa các tri thức để đi đến hoàn thiện dần cơ sở tri thức trong mỗi người.
* Đối với các tri thức biểu diễn trong máy cũng vậy, chúng chủ động hướng người sử dụng biết khai thác tri thức. Đó chính là quá trình kích hoạt tri thức được thể hiện trong các hệ chuyên gia được xây dựng trên các cơ sở tri thức biểu diễn ở mức độ cao có khả năng tiếp nhận, tinh chế, tự hoàn thiện ngay trong quá trình hoạt động của hệ. Tính chủ động của tri thức còn thể hiện sinh động thông qua các ngôn ngữ lập trình trí tuệ nhân tạo nư Lisp, Prolog…ở đó không còn có sự phân biệt rõ ràng giữa dữ liệu và thủ tục.

## PHÂN LOẠI TRI THỨC

Tri thức tồn tại dưới 2 dạng cơ bản:

* Tri thức định lượng
* Tri thức định tính

Tri thức định lượng thường gắn với các loại kinh nghiệm khác nhau. Ở đây chúng ta xét về tri thức định tính.

Tri thức định tính được chia làm 3 loại:

* Tri thức mô tả
* Tri thức thủ tục
* Tri thức điều khiển

## TRI THỨC MÔ TẢ

Cho những thông tin về một sự kiện, hiện tượng hay quá trình mà không đưa ra thông tin về cấu trúc bên trong cũng như phương pháp sử dụng bên trong của tri thức đó.

Ví dụ: Khẳng định “Việt Nam là đất nước tươi đẹp”. Đây là một khẳng định bất biến, không phụ thuộc vào tình huống, không gian và thời gian. Các tri thức phụ thuộc không gian và thời gian đòi hỏi những mô hình biểu diễn đặc biệt, cho phép thể hiện các tương quan giữa các sự kiện, quá trình không gian và thời gian.

Ngoài ra các tri thưc mô tả còn cho phép mô tả các mối liên hệ, các ràng buộc giữa các đối tượng, các sự kiện và các quá trình. Ví dụ “Tôi muốn mua bút” miêu tả mối quan hệ giữa đối tượng “tôi” và “bút” thông qua quan hệ “muốn mua”.

## TRI THỨC THỦ TỤC

Cho ta những phương pháp cấu trúc tri thức, ghép nối và suy diễn các tri thức mới từ những tri thức đã có. Các tri thức loại này tạo nên cơ sở của kỹ nghệ xử lý tri thức.

Một số thủ tục tri thức cơ bản:

* Tổng hợp tri thức: Suy diễn, Quy diễn, quy nạp.
* Học tự động: 2 cách suy diễn logic thường được sử dụng trong các hệ thống là:
* Modus Ponens

Nghĩa là nếu A đúng, A suy ra B thì B cũng đúng

* Modus Tollens

Nghĩa là nếu B sai, A suy ra B thì A cũng sai

## TRI THỨC ĐIỀU KHIỂN

Dùng để điều khiển, phối hợp các nguồn tri thức thủ tục và tri thức mô tả khác nhau.

## CÁC CẤP ĐỘ TRI THỨC

## TRI THỨC ĐỘNG PHỤ THUỘC VÀO TÌNH HUỐNG,KHÔNG GIAN VÀ THỜI GIAN

Các tri thức mô tả, tri thức thủ tục, tri thức điều khiển không phụ thuộc vào yếu tố không gian, thời gian được gọi là tri thức tĩnh. Các tri thức loại này tạo nên phàn lõi trong các cơ cấu tri thức. Nguồn các cơ cấu tri thức này thường phát sinh từ các tài liệu chuyên môn, các nguyên lý chung của khoa học. Ví dụ: “Nếu một đường thẳng vuông góc với một trong hai đường thẳng song song thì nó vuông góc với đường thẳng còn lại”.

Tuy vậy có những tri thức còn phụ thuộc vào lich sử, thông qua tham số thời gian và không gian có thể xuất hiện tường minh hoặc không tường minh trong các phát biểu. Chẳng hạn phát biểu: “Việt Nam không phải là thành viên của tổ chức WTO” chỉ đúng ở thời điểm trước năm 2008, còn hiện nay Việt Nam đã gia nhập vào tổ chức WTO. Chính yếu tố đó mà quá trình suy diễn trong các cơ sở tri thức được phụ thuộc không gian, thời gian có thể giao hoán hay không giao hoán bộ phận, đơn điệu hay không đơn điệu.

## TRI THỨC BẤT ĐỊNH, TRI THỨC KHÔNG ĐẦY ĐỦ

Trong nhiều trường hợp các tri thức có thể đúng hoặc sai. Tuy vậy trong thực tế ta gặp phải các phát biểu không phải lúc nào cũng xác định được chúng đúng hay sai. Ví dụ: “Trời có thể mưa”, trong trường hợp này không thể quyết định 100% là trời mưa hay không mưa. Các tri thức không chính xác là các mệnh đề phát biểu mà giá trị chân lý của chúng không thể chỉ ra một cách chính xác, tương ứng với thang đo quy ước. Ví dụ: “Anh ta cao khoảng 1m70”.

Cũng có thể xuất hiện các tri thức không đầy đủ trong các phát biểu, các mô tả. Ví dụ: “Thông thường nếu như anh ta đi thì chị ấy cũng đi”, đây là phát biểu bất định, song chỉ có tác dụng nếu biết được một chút về sự kiện “anh ta có đến hay không”.

Nói chung, các tri thức bất định, không chính xác và không đầy đủ xuất hiện là do các phát biểu, người ta sử dụng các yếu tố ngôn ngữ không rõ ràng như: có lẽ, có thể, khoảng, nói chung…Một trong những cách tiếp cận để xử lý các loại tri thức trên là sử dụng cách tiếp cận tri thức mờ. Các lý thuyết lập luận xấp xỉ đã và đang được quan tâm, nghiên cứu rất nhiều.

## CÁC PHƯƠNG PHÁP BIỂU DIỄN TRI THỨC

Bằng nhiều cách khác nhau:

* Các luật sinh (sản xuất)
* Lôgic
* Mạng ngữ nghĩa
* Ngôn ngữ nhân tạo
* Bộ ba Đối tượng-Thuộc tính-Giá trị (O-A-V: Object-Attribute-Value)
* Frame …

## BIỂU DIỄN TRI THỨC BẰNG LUẬT SINH

* Hầu hết HCG hiện nay dùng luật sinh biểu diễn tri thức, do:
* Tính đơn thể: các luật có tính độc lập, có thể mở rộng một cách dễ dàng;
* Khả năng giải thích: cách biểu diễn dùng các luật IF THEN, cho phép dễ dàng giải thích cách suy luận;
* Tương tự quá trình nhận thức của con người trong giải quyết vấn đề.
* Các dạng luật sinh:
* Dạng 1: IF <điều kiện> THEN <hành động>
* Dạng 2: IF <điều kiện> THEN <kết luận> DO <hành động>
* Luật có thể có tên, phần <điều kiện> còn gọi là vế trái của luật, tiền đề, mẫu so khớp; phần sau THEN là phần kết luận hay hệ quả.
* Ví dụ: Các luật giao thông:
  + IF <đèn đỏ sáng> THEN dừng
  + IF đèn xanh sáng THEN đi
* **Ưu điểm và nhược điểm của biểu diễn tri thức bằng luật sinh**
* **Ưu điểm:**

Biểu diễn tri thức bằng luật đặc biệt hữu hiệu trong những tình huống hệ thống cần đưa ra những hành động dựa vào những sự kiện có thể quan sát được. Nó có những ưu điểm chính yếu sau đây :

* Các luật rất dễ hiểu nên có thể dễ dàng dùng để trao đổi với người dùng (vì nó là một trong những dạng tự nhiên của ngôn ngữ).
* Có thể dễ dàng xây dựng được cơ chế suy luận và giải thích từ các luật.
* Việc hiệu chỉnh và bảo trì hệ thống là tương đối dễ dàng.
* Có thể cải tiến dễ dàng để tích hợp các luật mờ.
* Các luật thường ít phụ thuộc vào nhau.
* **Nhược điểm**
* Các tri thức phức tạp đôi lúc đòi hỏi quá nhiều (hàng ngàn) luật sinh. Điều này sẽ làm nảy sinh nhiều vấn đề liên quan đến tốc độ lẫn quản trị hệ thống.
* Thống kê cho thấy, người xây dựng hệ thống trí tuệ nhân tạo thích sử dụng luật sinh hơn tất cả phương pháp khác (dễ hiểu, dễ cài đặt) nên họ thường tìm mọi cách để biểu diễn tri thức bằng luật sinh cho dù có phương pháp khác thích hợp hơn! Đây là nhược điểm mang tính chủ quan của con người.
* Cơ sở tri thức luật sinh lớn sẽ làm giới hạn khả năng tìm kiếm của chương trình điều khiển. Nhiều hệ thống gặp khó khăn trong việc đánh giá các hệ dựa trên luật sinh cũng như gặp khó khăn khi suy luận trên luật sinh.

## BIỂU DIỄN TRI THỨC BẰNG LOGIC

Dựa vào các khái niệm cơ bản về logic mệnh đề và logic vị từ, với một số bài toán, các trạng thái được mô tả qua các biểu thức logic. Khi đó bài toán được phát biểu lại dưới dạng:

A.Chứng minh: Từ suy ra một trong các kết luận: .

Ở đây , là các biểu thức logic (mệnh đề hoặc vị từ)

B.Tìm phép gán cho các biến tự do sao cho từ suy ra một trong các kết luận ,…,.

* Cơ sở tri thức bằng logic mệnh đề:

Cơ sở tri thức gồm 2 phần:

* Các sự kiện
* Các luật
* Cơ sở tri thức biểu diễn bằng logic vị từ:

Cơ sở tri thức được cấu tạo bởi 2 phần:

* Tập các sự kiện F
* Tập các luật R
* **Ưu điểm và nhược điểm của biểu diễn tri thức bằng logic**
* **Ưu điểm**
* Là ngôn ngữ biểu diễn kiểu mô tả.
* Có khả năng suy diễn đối với các cơ chế quên thuộc: Pronens & Tollens.
* Khá trực quan với người sử dụng.
* Khá gần gũi về cú pháp với các lệnh lập trình logic, chẳng hạn như Prolog.
* Có thể dùng để mô tả cấu trúc mô hình và xử lý động mô hình.
* Có thể kiểm tra tính mâu thuẫn trong cơ sở tri thức.
* Tính mô-đun cao, do vậy các tri thức có thể thêm bớt sửa đổi khá độc lập với nhau và các cơ chế suy diễn
* **Nhược điểm**
* Mức độ hình thức hóa cao, dẫn tới khó hiểu ngữ nghĩa của các vị từ khi xét chương trình.
* Năng xuất xử lý thấp. Một trong những khó khăn cơ bản của quá trình suy diễn là cơ chế hợp và suy diễn vét cạn.
* Do các tri thức được biểu diễn nhờ các vị từ, nên ưu thế sử dụng cấu trúc dữ liệu không được khai thác triệt để.

## BIỂU DIỄN TRI THỨC BẰNG MẠNG NGỮ NGHĨA

Sử dụng đồ thị gồm các nút (node) và các cung (arc) nối các nút. Nút thể hiện các đối tượng, thuộc tính của đối tượng và giá trị thuộc tính; cung thể hiện mối quan hệ giữa các nút.

Ví dụ: Tri thức “Sẻ là một loài chim; chim có cánh và biết bay” được biểu diễn bằng mạng ngữ nghĩa sau:

là

có

biết

* **Ưu điểm và nhược điểm của biểu diễn tri thức bằng mạng ngữ nghĩa**
* **Ưu điểm**
* Cho phép biểu diễn một cách trực quan các sự kiện và mối quan hệ giữa chúng
* Tính mô-đun cao, theo nghĩa các tri thức thêm vào hoàn toàn độc lập với các tri thức cũ
* Là ngôn ngữ biểu diễn dạng mô tả
* Có thể áp dụng một số cơ chế trên mạng: Cơ chế truyền và thừa hưởng thông tin giữa các đối tượng.
* **Nhược điểm**
* Không có một phương pháp suy diễn chung cho mọi loại mạng ngữ nghĩa
* Khó kiểm soát được quá trình cập nhật tri thức, dễ dẫn đến mâu thuẫn trong cơ sở tri thức.

## BIỂU DIỄN TRI THỨC BẰNG NGÔN NGỮ NHÂN TẠO

Hiện nay các HCG đối thoại với người sử dụng bằng ngôn ngữ tự nhiên không nhiều.

Các HCG thường dùng 1 ngôn ngữ nhân tạo để biểu diễn tri thức.

## BIỂU DIỄN BẰNG BỘ BA LIÊN HỢP O.A.V

Biểu diễn tri thức bằng bộ ba liên hợp O.A.V là sử dụng bộ ba “Đối tượng”- “Thuộc tính”- “Giá trị” (Object-Attribute-Value) để chỉ ra rằng đối tượng với thuộc tính đã cho nào đó có một giá trị nào đó.

* **Ưu điểm và nhược điểm của biểu diễn tri thức bằng bộ ba liên hợp OAV**
* **Ưu điểm**
* Cho phép biểu diễn các đối tượng một cách trực quan
* Tính mô-đun tương đối cao
* Là ngôn ngữ biểu diễn dạng mô tả
* Cho phép diễn đạt tường minh các luật suy diễn
* **Nhược điểm**
* Mang những nhược điểm của mạng ngữ nghĩa. Ngoài ra khi sử dụng phương pháp này, các quan hệ, liên kết giữa các đối tượng không thể biểu diễn một cách tường minh.

## BIỂU DIỄN TRI THỨC BẰNG FRAME

**Frame** là một cấu trúc dữ liệu chứa đựng tất cả những tri thức liên quan đến một đối tượng cụ thể nào đó. Frames có liên hệ chặt chẽ đến khái niệm hướng đối tượng (thực ra frame là nguồn gốc của lập trình hướng đối tượng). Ngược lại với các phương pháp biểu diễn tri thức đã được đề cập đến, frame "đóng gói" toàn bộ một đối tượng, tình huống hoặc cả một vấn đề phức tạp thành một thực thể duy nhất có cấu trúc. Một frame bao hàm trong nó một khối lượng tương đối lớn tri thức về một đối tượng, sự kiện, vị trí, tình huống hoặc những yếu tố khác. Do đó, frame có thể giúp ta mô tả khá chi tiết một đối tượng.

**Frame** là một cấu trúc dữ liệu chứa đựng tất cả những tri thức liên quan đến một đối tượng cụ thể nào đó. Frames có liên hệ chặt chẽ đến khái niệm hướng đối tượng (thực ra frame là nguồn gốc của lập trình hướng đối tượng). Ngược lại với các phương pháp biểu diễn tri thức đã được đề cập đến, frame "đóng gói" toàn bộ một đối tượng, tình huống hoặc cả một vấn đề phức tạp thành một thực thể duy nhất có cấu trúc. Một frame bao hàm trong nó một khối lượng tương đối lớn tri thức về một đối tượng, sự kiện, vị trí, tình huống hoặc những yếu tố khác. Do đó, frame có thể giúp ta mô tả khá chi tiết một đối tượng.

* **Ưu điểm và nhược điểm của biểu diễn tri thức bằng FRAME**
* **Ưu điểm**
* Đáp ứng tất cả các yêu cầu về biểu diễn tri thức.
* Cho phép người sử dụng khá tự do khi biểu diễn tri thức.
* FRAME không chỉ sử dụng để mô tả tri thức mà còn dùng để thể hiện các thuật toán suy dẫn.
* Tận dụng được những điểm mạnh của biểu diễn thủ tục và mô tả.
* Qúa trình xử lý trên FRAME độc lập theo nghĩa kế thừa thông tin, không nhât thiết phải tuần tự.
* **Nhược điểm**
* Phương pháp biểu diễn quá phức tạp và cồng kềnh
* Phương pháp FRAME tiện lợi đối với kỹ sư xử lý tri thức cũng như người sử dụng có trình độ cao, nhưng lại là sự quá tải đối với những người sử dụng thông thường.
* Các giá trị của slot có thể gán qua thực hiện các thủ tục, điều này làm cho việc thu nạp và cập nhật tri thức trở nên phức tạp và làm khả năng mềm dẻo phù hợp với những thay đổi của môi trường bên ngoài bị giảm xuống.
* Do cấu trúc của FRAME nên khi biểu diễn cần phải sử dụng các biện pháp khá cầu kỳ. Vì vậy làm mất đi tính trực quan trong phương pháp biểu diễn.
* Đối với các bài toán phức tạp thì việc mô tả và điều khiển hệ thống sử dụng FRAME sẽ phức tạp lên rất nhiều so với phương pháp biểu diễn khác.

## MÔ TƠ SUY DIỄN

## CƠ CHẾ SUY DIỄN

## SUY DIỄN TIẾN

Suy diễn tiến là lập luận từ các sự kiện, sự việc để rút ra các kết luận. Ví dụ: nếu trời mưa trước khi ra khỏi nhà (sự kiện) thì phải lấy áo mưa (kết luận).

Trong phương pháp này, người sử dụng cung cấp các sự kiện cho hệ chuyên gia để hệ thống (máy suy diễn) tìm ra cách rút ra kết luận có thể. Kết luận được xem là những thuộc tính có thể được gán giá trị. Trong số những kết luận này có thể có những kết luận làm người sử dụng quan tâm, một số khác không nói lên điều gì, một số khác có thể vắng mặt.

Các sự kiện thường có dạng: Attribute = Value

Lần lượt các sự kiện trong cơ sở tri thức được chọn và hệ thống xem xét tất cả các luật mà các sự kiện này xuất hiện như là tiền đề. Theo nguyên tắc lập luận trên, hệ thống sẽ lấy ra những luật thỏa mãn. Sau khi gán giá trị cho các thuộc tính thuộc kết luận tương ứng, người ta nói rằng các sự kiện đã được thỏa mãn. Các thuộc tính được gán giá trị sẽ là một phần của kết quả. Sau khi đã được xem xét, kết quả này được xuất ra cho người sử dụng dùng.

## SUY DIỄN LÙI

Phương pháp suy diễn lùi tiến hành các lập luận theo chiều ngược lại (đối với các phương pháp suy diễn tiến). Từ một giả thuyết như là một kết luận, hệ thống đưa ra một tình huống trả lời gồm các sự kiện là cơ sở của giả thuyết đã cho này.

Ví dụ: nếu ai đó vào nhà mà cầm áo mưa và quần áo bị ướt thì giả thuyết này là trời mưa. Để củng cố giả thuyết này, ta hỏi người đó xem có phải trời mưa không? Nếu người đó trả lời có thì giả thuyết trời mưa là đúng và trở thành một sự kiện. Nghĩa là trời mưa nên phải cầm áo mưa và quần áo bị ướt.

Suy diễn lùi là cho phép nhận được giá trị của một thuộc tính. Đó là câu trả lời cho câu hỏi “giá trị của thuộc tính A là bao nhiêu ?” với A là một đích.

Để xác định giá trị của A, cần có các nguồn thông tin. Những nguồn này có thể là những câu hỏi hoặc có thể là những luật. Căn cứ vào các câu hỏi, hệ thống nhận được một cách trực tiếp từ người sử dụng những giá trị của thuộc tính liên quan. Căn cứ vào các luật, hệ thống suy diễn có thể tìm ra giá trị sẽ là kết luận của một trong số các kết luận có thể của thuộc tính liên quan, ....

Ý tưởng của thuật toán suy diễn lùi như sau: Với mỗi thuộc tính đã cho, người ta định nghĩa nguồn của nó:

* Nếu thuộc tính xuất hiện như là tiền đề của một luật (phần đầu của luật) thì nguồn sẽ thu gọn thành một câu hỏi.
* Nếu thuộc tính xuất hiện như là hậu quả của một luật (phần cuối của luật) thì nguồn sẽ là các luật mà trong đó thuộc tính là kết luận.
* Nếu thuộc tính là trung gian, xuất hiện đồng thời như là tiền đề và như là kết luận, khi đó nguồn có thể là các luật, hoặc có thể là các câu hỏi mà chưa được nêu ra.

Nếu mỗi lần với câu hỏi đã cho, người sử dụng trả lời hợp lệ, giá trị trả lời này sẽ được gán cho thuộc tính và xem như thành công. Nếu nguồn là các luật hệ thống này sẽ lấy lần lượt các luật mà thuộc tính đích xuất hiện như kết luận, để có thể tìm giá trị các thuộc tính thuộc tiền đề. Nếu các luật thỏa mãn, thuộc tính kết luận sẽ được ghi nhận.

## CƠ CHẾ HỖN HỢP

Sử dụng kết hợp cả hai phương pháp suy diễn trên.

## CƠ CHẾ ĐIỀU KHIỂN

## CHỌN HƯỚNG SUY DIỄN

Cho fgt = # GT GT: tập các sự kiện ban đầu

fkl = #KL KL: tập các sự kiện kết quả

ftrước = max #{ r R/r có thể áp dụng cho 1 tập con F nào đó }

= max # lọc (F,R)

fsau = max #{ r R/r có cung một sự kiện ở vế phải }

(\*) Các luật Heuristic:

Luật 1: nếu fsau < ftrước thì chọn suy diễn tiến

Luật 2: nếu fsau > ftrước thì chọn suy diễn lùi

Luật 3: nếu fsau = ftrước và fgt < fkl thì chọn suy diễn tiến.

Luật 4: nếu fsau = ftrước và fgt > fkl thì chọn suy diễn lùi.

Luật 5: nếu fsau = ftrước và fgt = fkl người thiết kế có thể chọn 1 trong 2 phương pháp suy diễn để sử dụng.

## GIẢI QUYẾT CÁC VẤN ĐỀ CẠNH TRANH

1. Cạnh tranh trong suy diễn tiến

Tình huống cạnh tranh xảy ra khi và chỉ khi tồn tại F và r1, r2 R mà:

r1: left1 q1, r2: left2 q2, left1 F, left2 F ⬄ # lọc (F,R) 2

ta có lọc ({a},R) = {r1,r2}

??? Đặt vấn đề: Làm thế nào để chọn 1 luật r trong số các luật có thể áp dụng được bằng lọc (F,R) ?

* Giải pháp 1: Tổ chức các luật có thể sử dụng được như một hàng đợi.
* Giải pháp 2: Tổ chức các luật có thể sử dụng theo xếp chồng.
* Giải pháp 3: Sử dụng Heuristic.

Đối với mỗi r R, bằng kỹ thuật Heuristic ta đánh giá liên hệ hàm ước lượng h trong KL với 1 phần vế phải của luật r, r: left q

H(r,KL) = h(q,KL)

Nguyên tắc: Luật r: left q sẽ được chọn khi và chỉ khi h(q,KL) min/max

* Giải pháp 4: Thực hiện sắp xếp thứ tự các sự kiện (Đồ thị FPG - Fact Precedence Graph)

Cho tập luật R và mỗi sự kiện của R là một nút, các luật là các dây cung trong đồ thị FPG.

Giải pháp 4’: sử dụng đồ thi VÀ/ HOẶC

Mỗi luật r: p1^p2^...^pn q tương đương với một cụm cung “VÀ”

* Giải pháp 5: Sử dụng đồ thị thứ tự luật (RPG – Rule Precedence Graph)

Mỗi luật rj là sắp xếp thứ tự với luật ri (ký hiệu là rj ri) nếu và chỉ nếu tồn tại một sự kiện f sao cho:

rj: left f; ri: ...f... q

Một luật r được gọi là “khởi đầu – initial” nếu và chỉ nếu:

r: left q và left GT INITIAL

Mỗi luật r được gọi là “kết thúc – final” nếu và chỉ nếu:

r: left q và left GT FINAL

Biểu diễn trong đồ thị RPG:

Mỗi luật khởi đầu được coi là “áp dụng”: APP = {INITIAL}.

Cho App(r) điểm – vào(r) = {r’} tồn tại một đỉnh r’ r trong RPG;

r: left q; mỗi r’ App(r) có thể áp dụng

nếu left {q’/r’ App(r) } GT thì luật r cũng có thể được áp dụng.

Suy diễn trong đồ thị RPG như sau:

* Chọn một luật trong APPLICABLE.
* Thực hiện luật đó.

Chú ý: đối với suy diễn theo chiều rộng APP = hàng đợi, suy diễn theo chiều sâu APP là xếp chồng.

(\*) Một số kinh nghiệm (Heuristics):

* h(r,FINALS) = h(r) # điểm – ra(r) = #{r’/ tồn tại một cung r r’ trong RPG}

**Luật được chọn ⬄ h(r) = # điểm – ra(r) max**

* h(r,FINALS) = #

**Luật được chọn ⬄ h(r,FINALS) min**

* h(r,FINALS) = # {p/ p: r FINALS} (số lượng các luật đi từ luật r đến FINALS trong đồ thị RPG)

**Luật được chọn ⬄ h(r,FINALS) max**

1. Cạnh tranh trong suy diễn lùi

* Cạnh tranh trong suy diễn lùi xảy ra khi và chỉ khi với một sự kiện f nào đó tồn tại ít nhất 2 luật r1, r2:

r1: left1 f và r2: left2 f

* Định nghĩa hàm tìm thấy:

Tìm\_thấy (f) = {r/ r: left f}

* Câu hỏi đặt ra: Làm thế nào để chọn một luật sao cho r Tìm\_thấy (f)
* Giải pháp 1:

Nếu ri, rj Tìm\_thấy (f) và ri, rj không được sử dụng nữa và i < j thì ri được chọn.

* Giải pháp 2: Sử dụng đồ thị VÀ/ HOẶC và đồ thị FPG

(\*) Một vài phương pháp kinh nghiệm (Heuristic)

* Xét luật r : left q, với mỗi sự kiện f độ\_dài (f, GT) = độ dài đường đi ngăn nhất từ GT đến f.

h(r,GT) = max (độ\_dài (f,GT) / f left)

**Luật được chọn ⬄ h(r1,GT) min**

* h(r,GT) = # left / \* r: left q \*/

**Luật được chọn ⬄ # left min**

* Hàm độ\_sâu (f,GT) = 0 nếu f GT

Max { max (độ\_sâu(q) + 1)} r: left q, q left

**Luật được chọn ⬄ max {độ\_sâu(q)} min**

* Hàm h(r,GT) = f left

**Luật được chọn ⬄ h(r,GT) min**

1. Việc giới hạn

Thao tác lọc thường tốn thời gian để tính toán nên phải lướt qua tập luật R để tìm ra tập có thể áp dụng được APPLICABLE. Để làm nhẹ con số thử nghiệm, người ta quyết định chỉ lọc một phần các luật hoặc sự kiện được đưa vào. Đó là giới hạn phải làm trước khi lọc.

# CHƯƠNG 2 : CÁC THUẬT TOÁN ÁP DỤNG TRONG BÀI TOÁN

## THUẬT TOÁN ID3

## TƯ TƯỞNG THUẬT TOÁN ID3

Trong ID3, chúng ta cần xác định thứ tự của thuộc tính cần được xem xét tại mỗi bước. Với các bài toán có nhiều thuộc tính và mỗi thuộc tính có nhiều giá trị khác nhau, việc tìm được nghiệm tối ưu thường là không khả thi. Thay vào đó, một phương pháp đơn giản thường được sử dụng là tại mỗi bước, một thuộc tính tốt nhất sẽ được chọn ra dựa trên một tiêu chuẩn nào đó (chúng ta sẽ bàn sớm). Với mỗi thuộc tính được chọn, ta chia dữ liệu vào các child node tương ứng với các giá trị của thuộc tính đó rồi tiếp tục áp dụng phương pháp này cho mỗi child node. Việc chọn ra thuộc tính tốt nhất ở mỗi bước như thế này được gọi là cách chọn greedy (tham lam). Cách chọn này có thể không phải là tối ưu, nhưng trực giác cho chúng ta thấy rằng cách làm này sẽ gần với cách làm tối ưu. Ngoài ra, cách làm này khiến cho bài toán cần giải quyết trở nên đơn giản hơn.

Sau mỗi câu hỏi, dữ liệu được phân chia vào từng child node tương ứng với các câu trả lời cho câu hỏi đó. Câu hỏi ở đây chính là một thuộc tính, câu trả lời chính là giá trị của thuộc tính đó. Để đánh giá chất lượng của một cách phân chia, chúng ta cần đi tìm một phép đo.

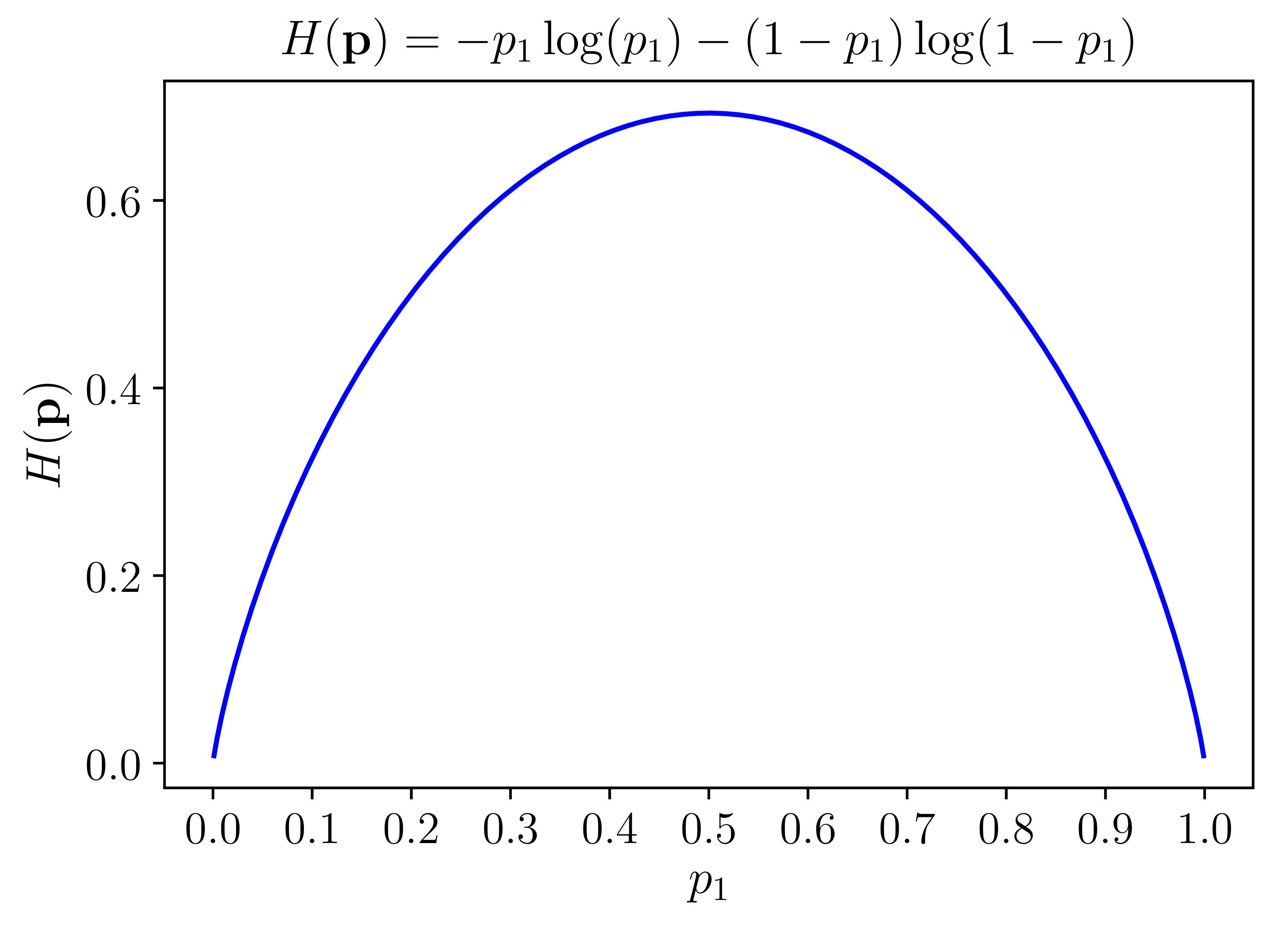
Trước hết, thế nào là một phép phân chia tốt? Bằng trực giác, một phép phân chia là tốt nhất nếu dữ liệu trong mỗi child node hoàn toàn thuộc vào một class–khi đó child node này có thể được coi là một leaf node, tức ta không cần phân chia thêm nữa. Nếu dữ liệu trong các child node vẫn lẫn vào nhau theo tỉ lệ lớn, ta coi rằng phép phân chia đó chưa thực sự tốt. Từ nhận xét này, ta cần có một hàm số đo độ tinh khiết (purity), hoặc độ vẩn đục (impurity) của một phép phân chia. Hàm số này sẽ cho giá trị thấp nhất nếu dữ liệu trong mỗi child node nằm trong cùng một class (tinh khiết nhất), và cho giá trị cao nếu mỗi child node có chứa dữ liệu thuộc nhiều class khác nhau.

Một hàm số có các đặc điểm này và được dùng nhiều trong lý thuyết thông tin là hàm entropy.

## HÀM SỐ ENTROPY

Cho một phân phối xác suất của một biến rời rạc xx có thể nhận nn giá trị khác nhau x1,x2,…,xnx1,x2,…,xn. Giả sử rằng xác suất để xx nhận các giá trị này là pi=p(x=xi)pi=p(x=xi) với 0≤pi≤1,∑ni=1pi=10≤pi≤1,∑i=1npi=1. Ký hiệu phân phối này là p=(p1,p2,…,pn)p=(p1,p2,…,pn). Entropy của phân phối này được định nghĩa là :

H(p) = − n∑i=1 pi log(pi) (1)



Đồ thị của hàm entropy với n=2

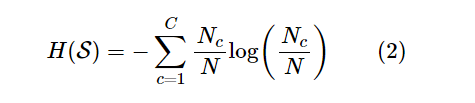
Tổng quát lên với n>2n>2, hàm entropy đạt giá trị nhỏ nhất nếu có một giá trị pi=1pi=1, đạt giá trị lớn nhất nếu tất cả các pipi bằng nhau ((việc này có thể được chứng minh bằng [phương pháp nhân tử Lagrange](https://machinelearningcoban.com/2017/04/02/duality/#--phuong-phap-nhan-tu-lagrange)).

Những tính chất này của hàm entropy khiến nó được sử dụng trong việc đo độ vẩn đục của một phép phân chia của ID3. Vì lý do này, ID3 còn được gọi là entropy-based decision tree.

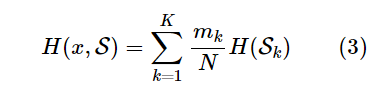
## THUẬT TOÁN ID3

Trong ID3, tổng có trọng số của entropy tại các leaf-node sau khi xây dựng decision tree được coi là hàm mất mát của decision tree đó. Các trọng số ở đây tỉ lệ với số điểm dữ liệu được phân vào mỗi node. Công việc của ID3 là tìm các cách phân chia hợp lý (thứ tự chọn thuộc tính hợp lý) sao cho hàm mất mát cuối cùng đạt giá trị càng nhỏ càng tốt. Như đã đề cập, việc này đạt được bằng cách chọn ra thuộc tính sao cho nếu dùng thuộc tính đó để phân chia, entropy tại mỗi bước giảm đi một lượng lớn nhất. Bài toán xây dựng một decision tree bằng ID3 có thể chia thành các bài toán nhỏ, trong mỗi bài toán, ta chỉ cần chọn ra thuộc tính giúp cho việc phân chia đạt kết quả tốt nhất. Mỗi bài toán nhỏ này tương ứng với việc phân chia dữ liệu trong một non-leaf node. Chúng ta sẽ xây dựng phương pháp tính toán dựa trên mỗi node này.

Xét một bài toán với C class khác nhau. Giả sử ta đang làm việc với một non-leaf node với các điểm dữ liệu tạo thành một tập SS với số phần tử là |S|=N. Giả sử thêm rằng trong số N điểm dữ liệu này, Nc,c=1,2,…, C điểm thuộc vào class c. Xác suất để mỗi điểm dữ liệu rơi vào một class c được xấp xỉ bằng Nc/N (maximum likelihood estimation). Như vậy, entropy tại node này được tính bởi:



Tiếp theo, giả sử thuộc tính được chọn là x. Dựa trên x, các điểm dữ liệu trong S được phân ra thành KKchild node S1,S2,…,SK với số điểm trong mỗi child node lần lượt là m1,m2,…,mK. Ta định nghĩa :



là tổng có trọng số entroy của mỗi child node–được tính tương tự như (2). Việc lấy trọng số này là quan trọng vì các node thường có số lượng điểm khác nhau.

Tiếp theo, ta định nghĩa information gain dựa trên thuộc tính x:

G(x,S)=H(S)−H(x,S)

Trong ID3, tại mỗi node, thuộc tính được chọn được xác định dựa trên:



tức thuộc tính khiến cho information gain đạt giá trị lớn nhất.

## VECTO ĐẶC TRƯNG

Quinlan quyết định thuộc tính phân hoạch bằng cách xây dựng các vecto đặc trưng cho mỗi giá trị của từng thuộc tính dẫn xuất và thuộc tính mục tiêu.

Với mỗi thuộc tính dẫn xuất A còn có thể sử dụng để phân hoạch, tính:

VA(j) = (T(j, r1), T(j, r2),…., T(j, r2) )

T(j, r1) = (tổng số phần tử trong phân hoạch có giá trị thuộc tính dãn xuất A là j và có giá trị thuộc tính mục tiêu là rj  ) / ( tổng số phần tử trong phân hoạch có giá trị thuộc tính dẫn xuất A là j).

Một vecto VA(j) được gọi là vecto đơn vị nếu nó chỉ có duy nhất một thành phần có giá trị 1 và những thành phần khác có giá trị 0.

Thuộc tính được chọn để phân hoạch là thuộc tính có nhiều vecto đơn vị nhất

## ĐỘ ĐO HỖN LOẠN

Thay vì phải xây dựng các vector đặc trưng như phương pháp của Quinlan, ứng với mỗi thuộc tính dẫn xuất ta chỉ cần tính ra độ đo hỗn loạn và lựa chọn thuộc tính nào có độ đo hỗn loại là thấp nhất. Công thức tính như sau :

TA = https://voer.edu.vn/file/36206

trong đó :

bt là tổng số phần tử có trong phân hoạch

bj là tổng số phần tử có thuộc tính dẫn xuất A có giá trị j.

bri : tổng số phần tử có thuộc tính dẫn xuất A có giá trị j và thuộc tính mục tiêu có giá trị i.

# CHƯƠNG 3 : XÂY DỰNG ỨNG DỤNG

## PHÂN TÍCH BÀI TOÁN

## Nghiên Cứu Các Yếu Tố Ảnh Hướng Đến Giá Vàng

**Quan hệ về nguồn cung vàng.**

Trước tiên, ta có thể nói rằng, vàng là một loại hàng hoá, một loại hàng hoá đặc biệt. Chính vì mang tính chất của một loại hàng hoá nên vàng cũng chịu tác động của quy luật cung cầu trên thị trường. Xét về nguồn cung của vàng, vàng được cung cấp bởi chủ yếu từ những nước có trữ lượng vàng lớn, sản lượng xuất khẩu lớn có tầm ảnh hưởng đến thị trường vàng thế giới: Nam Phi,Mỹ,Nga,Canada, Úc,… Theo dự báo, nguồn cung vàng cho thế giới sẽ giảm nhẹ khoảng 1,3% so với năm 2008 đạt mức 2.295 tấn(tương đương với 73,7 triệu ounce) nhưng vẫn đang cao hơn so với nhu cầu thực tế. Trong năm 2009, dự báo thế giới sẽ dư khoảng 300 tấn vàng (tương đương với 9,1 triệu ounce).

**Ảnh hưởng của USD đến giá vàng.**

Vì là một loại hàng hoá nên giá vàng cũng bị ảnh hưởng,tác động bởi các đồng tiền nội tệ và ngoại tệ mua nó mà ở đây chủ yếu là đồng USD. Chính vì thế, việc giao động đồng USD cũng ảnh hưởng không nhỏ tới giá vàng. Khi xem xét giá trị đồng USD, người ta thường đánh giá thông qua nền kinh tế Mỹ và những yếu tố chính được xem là “chỉ báo” phản ánh sức mạnh hay suy yếu của nền kinh tế Mỹ, đó là tình trạng thị trường nhà ở, thị trường lao động, thị trường tín dụng và thị trường vốn. Như ta đã biết, thị trường bất động sản là nơi khởi nguồn, là nơi bắt đầu cho cuộc đại suy thoái của kinh tế Mỹ. Sự suy giảm rồi kết quả là sụp đổ của thị trường bất động sản sẽ kéo theo sự giảm giá trị của đồng USD. Sự suy giảm giá trị của đồng tiền có sức ảnh hưởng lớn với nền kinh tế cũng kéo theo sự suy giảm về niềm tin nơi các nhà đầu tư về đồng tiền này. Các nhà đầu tư đang dần chuyển sang một loại tiền tệ khác có tính ổn định cao hơn và có tính an toàn hơn USD. Từ Euro,sang Bảng Anh, và thậm chí có thể là dùng vàng là đơn vị chuyển đổi hàng hoá.

## Bảng Thông Kê Số Liệu

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **STT** | **GIÁ TRỊ ĐỒNG $** | **LẠM PHÁT** | **NHU CẦU TIÊU THỤ** | **SẢN LƯỢNG KHAI THÁC** | **KẾT QUẢ** |
| 1 | TĂNG | TĂNG | THẤP | TB | CAO |
| 2 | TB | ỔN ĐỊNH | TB | CAO | THẤP |
| 3 | TĂNG | ỔN ĐỊNH | CAO | TB | CAO |
| 4 | TB | GIẢM | THẤP | CAO | THẤP |
| 5 | GIẢM | ỔN ĐỊNH | CAO | THẤP | THẤP |
| 6 | TB | GIẢM | THẤP | TB | THẤP |
| 7 | TĂNG | TĂNG | THẤP | THẤP | CAO |
| 8 | GIẢM | GIẢM | CAO | CAO | CAO |
| 9 | TB | GIẢM | THẤP | CAO | THẤP |
| 10 | GIẢM | TĂNG | CAO | TB | THẤP |
| 11 | TB | GIẢM | THẤP | CAO | CAO |
| 12 | GIẢM | GIẢM | THẤP | TB | THẤP |
| 13 | TB | TĂNG | TB | THẤP | THẤP |
| 14 | TĂNG | GIẢM | TB | THẤP | CAO |
| 15 | TĂNG | GIẢM | CAO | THẤP | CAO |
| 16 | TB | ỔN ĐỊNH | CAO | TB | CAO |
| 17 | GIẢM | TĂNG | THẤP | THẤP | THẤP |
| 18 | TB | GIẢM | CAO | CAO | CAO |
| 19 | TĂNG | GIẢM | CAO | THẤP | THẤP |
| 20 | GIẢM | ỔN ĐỊNH | CAO | TB | CAO |
| 21 | TĂNG | ỔN ĐỊNH | CAO | THẤP | THẤP |
| 22 | GIẢM | GIẢM | CAO | CAO | CAO |
| 23 | TB | TĂNG | THẤP | CAO | THẤP |
| 24 | TĂNG | GIẢM | CAO | TB | CAO |
| 25 | TĂNG | ỔN ĐỊNH | CAO | CAO | CAO |
| 26 | TB | GIẢM | TB | THẤP | THẤP |
| 27 | GIẢM | GIẢM | CAO | THẤP | THẤP |
| 28 | GIẢM | GIẢM | CAO | CAO | CAO |
| 29 | TB | ỔN ĐỊNH | CAO | TB | THẤP |
| 30 | GIẢM | TĂNG | THẤP | THẤP | THẤP |
| 31 | TĂNG | GIẢM | TB | CAO | CAO |
| 32 | TĂNG | GIẢM | TB | CAO | CAO |
| 33 | GIẢM | GIẢM | TB | THẤP | CAO |
| 34 | TB | TĂNG | THẤP | TB | THẤP |
| 35 | TĂNG | ỔN ĐỊNH | CAO | TB | THẤP |
| 36 | TĂNG | GIẢM | TB | THẤP | THẤP |
| 37 | TĂNG | GIẢM | TB | THẤP | CAO |
| 38 | TB | ỔN ĐỊNH | CAO | CAO | CAO |
| 39 | GIẢM | TĂNG | THẤP | CAO | THẤP |
| 40 | GIẢM | ỔN ĐỊNH | CAO | CAO | CAO |
| 41 | TĂNG | GIẢM | TB | TB | THẤP |
| 42 | GIẢM | ỔN ĐỊNH | CAO | TB | THẤP |
| 43 | TĂNG | ỔN ĐỊNH | CAO | TB | CAO |
| 44 | TĂNG | TĂNG | THẤP | THẤP | THẤP |
| 45 | GIẢM | GIẢM | TB | CAO | CAO |
| 46 | TB | GIẢM | TB | CAO | THẤP |
| 47 | GIẢM | TĂNG | THẤP | CAO | CAO |
| 48 | TĂNG | TĂNG | THẤP | THẤP | CAO |
| 49 | GIẢM | GIẢM | TB | TB | THẤP |
| 50 | TĂNG | GIẢM | CAO | TB | THẤP |
| 51 | TB | GIẢM | TB | CAO | CAO |
| 52 | TB | ỔN ĐỊNH | CAO | THẤP | THẤP |
| 53 | TĂNG | TĂNG | THẤP | CAO | THẤP |
| 54 | TĂNG | ỔN ĐỊNH | CAO | CAO | CAO |
| 55 | TB | GIẢM | CAO | TB | CAO |
| 56 | TB | ỔN ĐỊNH | CAO | TB | CAO |
| 57 | GIẢM | TĂNG | THẤP | TB | THẤP |
| 58 | GIẢM | GIẢM | CAO | TB | THẤP |
| 59 | TB | GIẢM | CAO | THẤP | THẤP |
| 60 | TĂNG | TĂNG | THẤP | TB | CAO |
| 61 | GIẢM | GIẢM | TB | THẤP | CAO |
| 62 | TB | TĂNG | THẤP | CAO | THẤP |
| 63 | TĂNG | TĂNG | THẤP | CAO | CAO |
| 64 | TB | TĂNG | THẤP | THẤP | THẤP |
| 65 | TB | ỔN ĐỊNH | CAO | CAO | THẤP |
| 66 | GIẢM | GIẢM | CAO | CAO | CAO |
| 67 | TĂNG | TĂNG | THẤP | TB | THẤP |
| 68 | TĂNG | ỔN ĐỊNH | CAO | THẤP | CAO |
| 69 | GIẢM | TĂNG | THẤP | CAO | THẤP |
| 70 | TB | ỔN ĐỊNH | CAO | CAO | CAO |
| 71 | TĂNG | GIẢM | CAO | CAO | CAO |
| 72 | GIẢM | GIẢM | CAO | THẤP | THẤP |
| 73 | TB | GIẢM | CAO | THẤP | THẤP |
| 74 | TĂNG | GIẢM | CAO | THẤP | CAO |
| 75 | TB | GIẢM | CAO | TB | CAO |
| 76 | GIẢM | ỔN ĐỊNH | CAO | THẤP | THẤP |
| 77 | GIẢM | GIẢM | CAO | TB | THẤP |
| 78 | TB | GIẢM | TB | CAO | CAO |
| 79 | TĂNG | GIẢM | TB | TB | CAO |
| 80 | TĂNG | TĂNG | THẤP | CAO | CAO |
| 81 | GIẢM | ỔN ĐỊNH | CAO | THẤP | THẤP |
| 82 | TĂNG | TĂNG | THẤP | THẤP | THẤP |
| 83 | GIẢM | TĂNG | THẤP | TB | THẤP |
| 84 | TĂNG | ỔN ĐỊNH | CAO | CAO | CAO |
| 85 | TB | GIẢM | CAO | CAO | CAO |
| 86 | GIẢM | ỔN ĐỊNH | CAO | TB | THẤP |
| 87 | TĂNG | ỔN ĐỊNH | CAO | TB | CAO |
| 88 | GIẢM | GIẢM | CAO | TB | THẤP |
| 89 | GIẢM | TĂNG | THẤP | TB | THẤP |
| 90 | TB | GIẢM | TB | TB | CAO |
| 91 | TĂNG | TĂNG | THẤP | TB | THẤP |
| 92 | TB | TĂNG | THẤP | CAO | CAO |
| 93 | GIẢM | ỔN ĐỊNH | CAO | CAO | THẤP |
| 94 | TB | GIẢM | CAO | CAO | CAO |
| 95 | TĂNG | GIẢM | TB | CAO | THẤP |
| 96 | TB | GIẢM | TB | TB | CAO |
| 97 | GIẢM | ỔN ĐỊNH | CAO | CAO | THẤP |
| 98 | TB | GIẢM | TB | THẤP | THẤP |
| 99 | GIẢM | GIẢM | TB | CAO | THẤP |
| 100 | TB | TĂNG | THẤP | TB | CAO |

## Sử dụng ID3 giải quyết bài toán với 20 dòng dữ liệu

**Tính Emtropy:**

Emtropy(S) = H(S) = -10/20 log2(10/20) – 10/20 log2 (10/20)

= 1

**Xét thuộc tính Giá trị đồng USD:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Giá trị đồng $** | **Lạm phát** | **Nhu cầu tiêu thụ** | **Sản lượng khai thác** | **Kết quả** |
| Tăng | Giảm | Thấp | Tb | Thấp |
| Tăng | Tăng | Thấp | Tb | Cao |
| Tăng | Ổn định | Cao | Tb | Cao |
| Tăng | Tăng | Thấp | Thấp | Cao |
| Tăng | Giảm | Tb | Thấp | Cao |
| Tăng | Giảm | Cao | Thấp | Cao |
| Tăng | Giảm | Cao | Thấp | Thấp |
|  |  |  |  |  |
| Tb | Ổn định | Tb | Cao | Thấp |
| Tb | Giảm | Thấp | Cao | Thấp |
| Tb | Giảm | Thấp | Tb | Thấp |
| Tb | Giảm | Thấp | Cao | Thấp |
| Tb | Giảm | Thấp | Cao | Cao |
| Tb | Tăng | Tb | Thấp | Thấp |
| Tb | Ổn định | Cao | Tb | Cao |
| Tb | Giảm | Cao | Cao | Cao |
|  |  |  |  |  |
| Giảm | Ổn định | Cao | Thấp | Thấp |
| Giảm | Giảm | Cao | Cao | Cao |
| Giảm | Tăng | Cao | Tb | Thấp |
| Giảm | Tăng | Thấp | Thấp | Thấp |
| Giảm | Ổn định | Cao | Tb | Cao |

H(tăng) = -5/7 log2 (5/7) – 2/7 log2 (2/7) = 0.863

H(trung bình) = -5/8 log2 (5/8) – 3/8 log2 (3/8) = 0.954

H(giảm) = -2/5 log2 (2/5) – 3/5 log2 (3/5) = 0.97

H(usd/S) = 7/20 \* H(tăng) + 8/20 \* H(trung bình) + 5/20 \* H(giảm)

= 0.926

Gain(usd) = Emtropy(S) – H(usd/S) = 1 – 0.926 = 0.074

**Xét thuộc tính lạm phát:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Giá trị đồng $** | **Lạm phát** | **Nhu cầu tiêu thụ** | **Sản lượng khai thác** | **Kết quả** |
| Tăng | Tăng | Thấp | Tb | Cao |
| Tăng | Tăng | Thấp | Thấp | Cao |
| Giảm | Tăng | Cao | Tb | Thấp |
| Tb | Tăng | Tb | Thấp | Thấp |
| Giảm | Tăng | Thấp | Thấp | Thấp |
|  |  |  |  |  |
| Tb | Ổn định | Tb | Cao | Thấp |
| Tăng | Ổn định | Cao | Tb | Cao |
| Giảm | Ổn định | Cao | Thấp | Thấp |
| Tb | Ổn định | Cao | Tb | Cao |
| Giảm | Ổn định | Cao | Tb | Cao |
|  |  |  |  |  |
| Tb | Giảm | Thấp | Cao | Thấp |
| Tb | Giảm | Thấp | Tb | Thấp |
| Giảm | Giảm | Cao | Cao | Cao |
| Tb | Giảm | Thấp | Cao | Thấp |
| Tb | Giảm | Thấp | Cao | Cao |
| Tăng | Giảm | Thấp | TB | Thấp |
| Tăng | Giảm | Tb | Thấp | Cao |
| Tăng | Giảm | Cao | Thấp | Cao |
| Tb | Giảm | Cao | Cao | Cao |
| Tăng | Giảm | Cao | Thấp | Thấp |

H(tăng) = -2/5 log2 (2/5) – 3/5 log2(3/5) = 0.97

H(ổn định) = -2/5 log2(2/5) – 3/5 log2(3/5) = 0.97

H(giảm) = -5/10 log2(5/10) – 5/10 log2(5/10) = 1

H(lamphat/S) = 5/20\*H(tăng) + 5/20\*H(ổn định) + 10/20\*H(giảm)

= 0.985

Gain(lamphat) = Emtropy(S) – H(lamphat/S) = 0.015

**Xét thuộc tính Nhu cầu tiêu thụ :**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Giá trị đồng $** | **Lạm phát** | **Nhu cầu tiêu thụ** | **Sản lượng khai thác** | **Kết quả** |
| Tăng | Ổn định | Cao | Tb | Cao |
| Giảm | Ổn định | Cao | Thấp | Thấp |
| Giảm | Giảm | Cao | Cao | Cao |
| Giảm | Tăng | Cao | Tb | Thấp |
| Tăng | Giảm | Cao | Thấp | Cao |
| Tb | Ổn định | Cao | Tb | Cao |
| Tb | Giảm | Cao | Cao | Cao |
| Tăng | Giảm | Cao | Thấp | Thấp |
| Giảm | Ổn định | Cao | Tb | Cao |
|  |  |  |  |  |
| Tăng | Tăng | Thấp | Tb | Cao |
| Tb | Giảm | Thấp | Cao | Thấp |
| Tb | Giảm | Thấp | Tb | Thấp |
| Tăng | Tăng | Thấp | Thấp | Cao |
| Tb | Giảm | Thấp | Cao | Thấp |
| Tb | Giảm | Thấp | Cao | Cao |
| Tăng | Giảm | Thấp | TB | Thấp |
| Giảm | Tăng | Thấp | Thấp | Thấp |
|  |  |  |  |  |
| Tb | Ổn định | Tb | Cao | Thấp |
| Tb | Tăng | Tb | Thấp | Thấp |
| Tăng | Giảm | Tb | Thấp | Cao |

H(cao) = 0.918 , H(trung bình) = 0.918 , H(thấp) = 0.954

H(nctt/S) = 0.932

Gain(nctt) = Emtropy(S) – H(nctt/S) = 0.068

**Xét thuộc tính sản lượn khai thác :**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Giá trị đồng $** | **Lạm phát** | **Nhu cầu tiêu thụ** | **Sản lượng khai thác** | **Kết quả** |
| Tb | Ổn định | Tb | Cao | Thấp |
| Tb | Giảm | Thấp | Cao | Thấp |
| Giảm | Giảm | Cao | Cao | Cao |
| Tb | Giảm | Thấp | Cao | Thấp |
| Tb | Giảm | Thấp | Cao | Cao |
| Tb | Giảm | Cao | Cao | Cao |
|  |  |  |  |  |
| Tăng | Tăng | Thấp | Tb | Cao |
| Tăng | Ổn định | Cao | Tb | Cao |
| Tb | Giảm | Thấp | Tb | Thấp |
| Giảm | Tăng | Cao | Tb | Thấp |
| Tăng | Giảm | Thấp | TB | Thấp |
| Tb | Ổn định | Cao | Tb | Cao |
| Giảm | Ổn định | Cao | Tb | Cao |
|  |  |  |  |  |
| Giảm | Ổn định | Cao | Thấp | Thấp |
| Tăng | Tăng | Thấp | Thấp | Cao |
| Tb | Tăng | Tb | Thấp | Thấp |
| Tăng | Giảm | Tb | Thấp | Cao |
| Tăng | Giảm | Cao | Thấp | Cao |
| Giảm | Tăng | Thấp | Thấp | Thấp |
| Tăng | Giảm | Cao | Thấp | Thấp |

H(cao) = 1, H(trung bình) = 0.985, H(thấp) = 0.985

H(slkt/S) = 0.989

Gain(slkt) = Emtropy(S) – H(slkt/S) = 0.01

Ta chọn MaxGain(value) là thuộc tính USD để phát triển cây.

**Ta có bảng dữ liệu mới :**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Giá trị đồng $** | **Lạm phát** | **nhu cầu tiêu thụ** | **Sản lượng khai thác** | **kết quả** |
| Tăng | Giảm | thấp | TB | thấp |
| tăng | tăng | thấp | TB | cao |
| tăng | ổn định | cao | TB | cao |
| tăng | tăng | thấp | thấp | cao |
| tăng | Giảm | TB | Thấp | cao |
| tăng | giảm | cao | thấp | cao |
| tăng | giảm | cao | thấp | thấp |
|  |  |  |  |  |
| TB | ổn định | TB | cao | thấp |
| TB | giảm | thấp | cao | thấp |
| TB | giảm | thấp | TB | thấp |
| TB | giảm | thấp | cao | thấp |
| TB | giảm | thấp | cao | cao |
| TB | tăng | TB | Thấp | thấp |
| TB | ổn định | cao | TB | cao |
| TB | giảm | cao | cao | cao |
|  |  |  |  |  |
| giảm | ổn định | cao | thấp | thấp |
| giảm | giảm | cao | cao | cao |
| giảm | tăng | cao | TB | thấp |
| giảm | tăng | thấp | thấp | thấp |
| giảm | ổn định | cao | TB | cao |

**Xét bảng tăng :**

**Xét thuộc tính lạm phát :**

|  |  |  |  |
| --- | --- | --- | --- |
| **Lạm phát** | **Nhu cầu tiêu thụ** | **Sản lượng khai thác** | **Kết quả** |
| Tăng | Thấp | Tb | Cao |
| Tăng | Thấp | Thấp | Cao |
|  |  |  |  |
| Ổn định | Cao | Tb | Cao |
|  |  |  |  |
| Giảm | Thấp | TB | Thấp |
| Giảm | TB | Thấp | Cao |
| Giảm | Cao | Thấp | Cao |
| Giảm | Cao | Thấp | Thấp |

H(tăng) = 0, H(ổn định) = 0, H(giảm) = 1

H(lamphat/S) = 0.571

**Xét thuộc tính nhu cầu tiêu thụ :**

|  |  |  |  |
| --- | --- | --- | --- |
| **Lạm phát** | **Nhu cầu tiêu thụ** | **Sản lượng khai thác** | **Kết quả** |
| Ổn định | Cao | Tb | Cao |
| Giảm | Cao | Thấp | Cao |
| Giảm | Cao | Thấp | Thấp |
|  |  |  |  |
| Giảm | TB | Thấp | Cao |
|  |  |  |  |
| Giảm | Thấp | TB | Thấp |
| Tăng | Thấp | Tb | Cao |
| Tăng | Thấp | Thấp | Cao |

H(cao) = 0.918, H(trung bình) = 0, H(thap) = 0.918

H(nctt/S) = 0.786

**Xét thuộc tính sản lượng khai thác :**

|  |  |  |  |
| --- | --- | --- | --- |
| **Lạm phát** | **Nhu cầu tiêu thụ** | **Sản lượng khai thác** | **Kết quả** |
| Giảm | Thấp | Tb | Thấp |
| Tăng | Thấp | Tb | Cao |
| Ổn định | Cao | Tb | Cao |
|  |  |  |  |
| Tăng | Thấp | Thấp | Cao |
| Giảm | Tb | Thấp | Cao |
| Giảm | Cao | Thấp | Cao |
| Giảm | Cao | Thấp | Thấp |

H(trung bình) = 0.918, H(thấp) = 0.811

H(slkt/S) = 0.857

Vì ta có :



Nên H(lamphat/S) là nhỏ nhất nên ta chọn thuộc tính lạm phát để phát triển cây.

**Ta có bảng dữ liệu mới :**

|  |  |  |  |
| --- | --- | --- | --- |
| **Lạm phát** | **Nhu cầu tiêu thụ** | **Sản lượng khai thác** | **Kết quả** |
| Tăng | Thấp | Tb | Cao |
| Tăng | Thấp | Thấp | Cao |
|  |  |  |  |
| Ổn định | Cao | Tb | Cao |
|  |  |  |  |
| Giảm | Thấp | Tb | Thấp |
| Giảm | Tb | Thấp | Cao |
| Giảm | Cao | Thấp | Cao |
| Giảm | Cao | Thấp | Thấp |

Do Lạm phát -> Tăng -> Cao, Lạm phát -> Ổn định -> Cao.

**Xét bảng giảm :**

**Xét thuộc tính nhu cầu tiêu thụ :**

|  |  |  |
| --- | --- | --- |
| **Nhu cầu tiêu thụ** | **Sản lượng khai thác** | **Kết quả** |
| Cao | Thấp | Cao |
| Cao | Thấp | Thấp |
|  |  |  |
| Tb | Thấp | Cao |
|  |  |  |
| Thấp | Tb | Thấp |

H(trung bình) = 0, H(thấp) = 0, H(cao) = 1, H(nctt/S) = 0.5

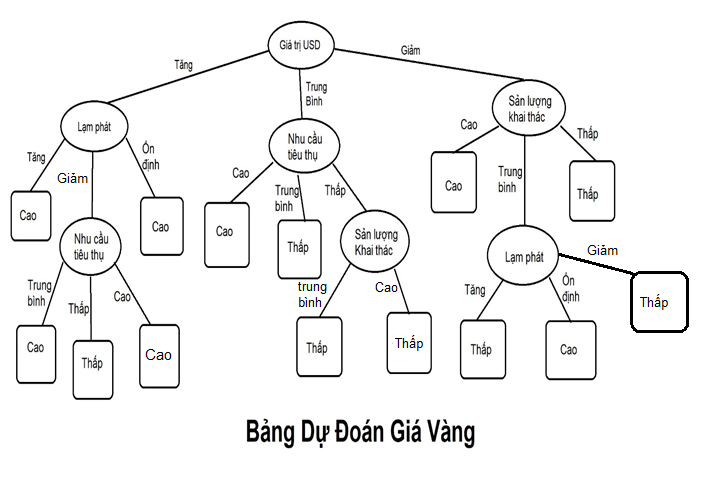
**Xét thuộc tính sản lượng khai thác :**

|  |  |  |
| --- | --- | --- |
| **Nhu cầu tiêu thụ** | **Sản lượng khai thác** | **Kết quả** |
| Thấp | Tb | Thấp |
|  |  |  |
| Tb | Thấp | Cao |
| Cao | Thấp | Cao |
| Cao | Thấp | Thấp |

H(trung bình) = 0, H(thấp) = 0.276, H(slkt/S) = 0.918

Vậy ta chọn Nhu cầu khai thác để xây dựng cây. Ta làm tương tự với bảng trung bình và giảm của thuộc tính giá trị đồng usd.

**Ta có cây quyết định :**

****

**Ta có tập luật :**

R1: if(Giá trị USD= tăng) and (Lạm phát=tăng) then giá vàng cao

R2: if(Giá trị USD= tăng) and (Lạm phát=ổn định) then giá vàng cao

R3: if(Giá trị USD= tăng) and (Lạm phát=giảm) and (NCTT=TB) then giá vàng cao

R4: if(Giá trị USD= tăng) and (Lạm phát=giảm) and (NCTT= Thấp) then giá vàng thấp

R5: if(Giá trị USD= tăng) and (Lạm phát=giảm) and (Nhu cầu thị trường = cao) then giá vàng cao

R6: if(Giá trị USD= TB) and (NCTT=cao) then giá vàng cao

R7: if(Giá trị USD= TB) and (NCTT=TB) then giá vàng thấp

R8: if(Giá trị USD= TB) and (NCTT=thấp) and (SLKT= trung bình) then giá vàng thấp

R9: if(Giá trị USD= TB) and (NCTT=thấp) and (SLKT= cao) then giá vàng thấp

R11: if(Giá trị USD= giảm) and (SLKT=cao) then giá vàng cao

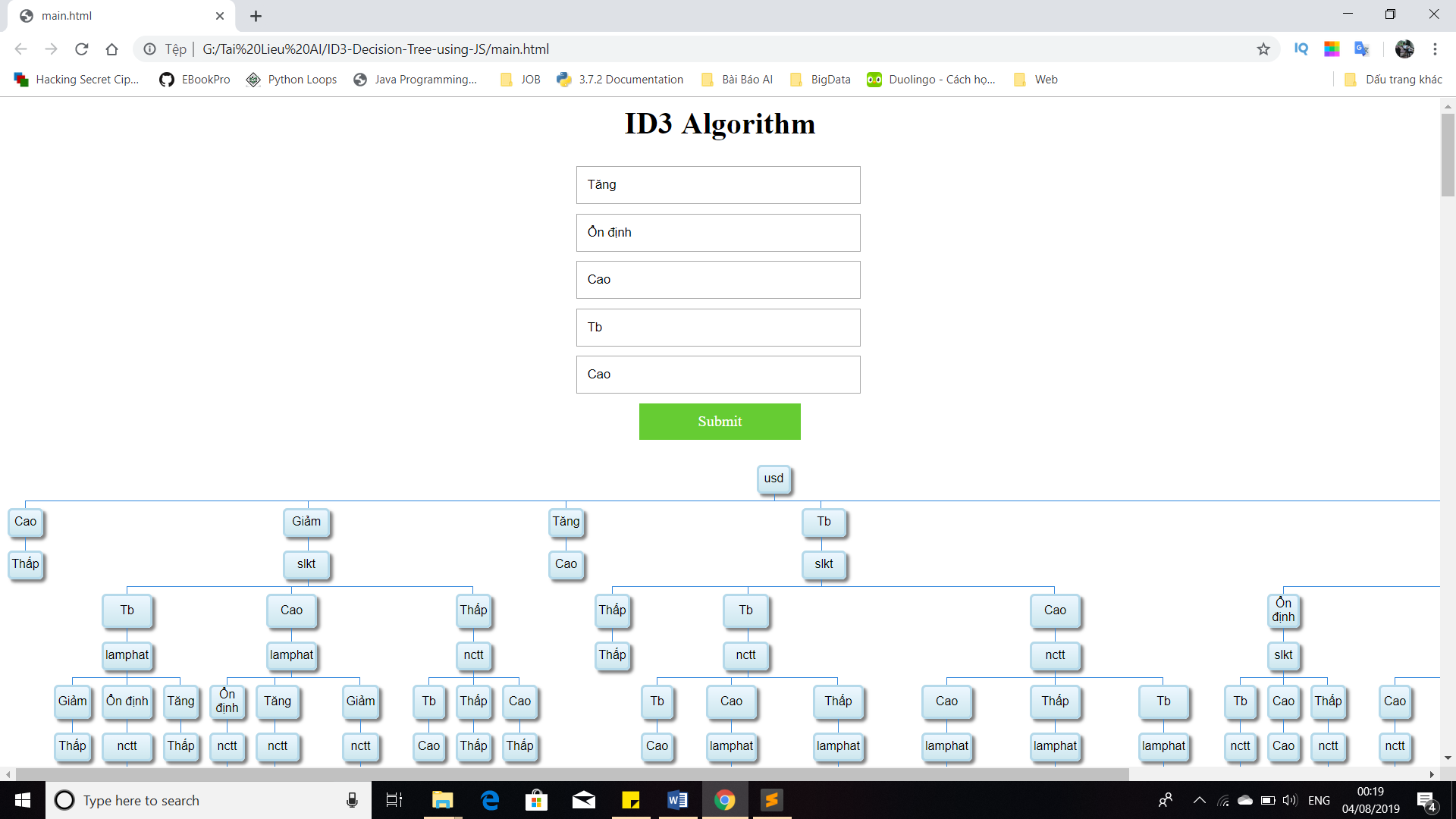
R12­: if(Giá trị USD= giảm) and (SLKT=thấp) then giá vàng thấp

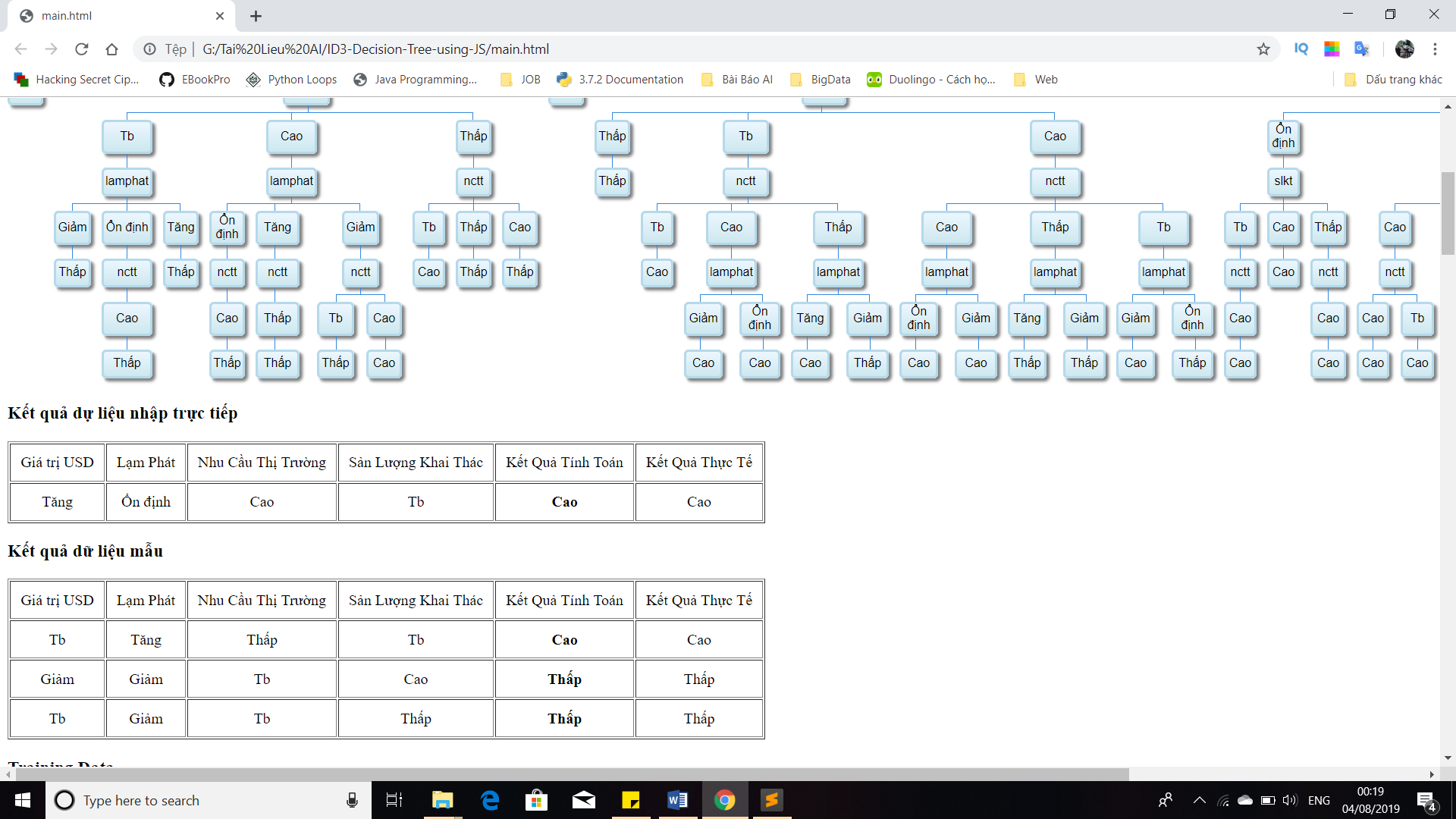
R13: if(Giá trị USD= giảm) and (SLKT=TB) and (Lạm phát=ổn định) then giá vàng cao

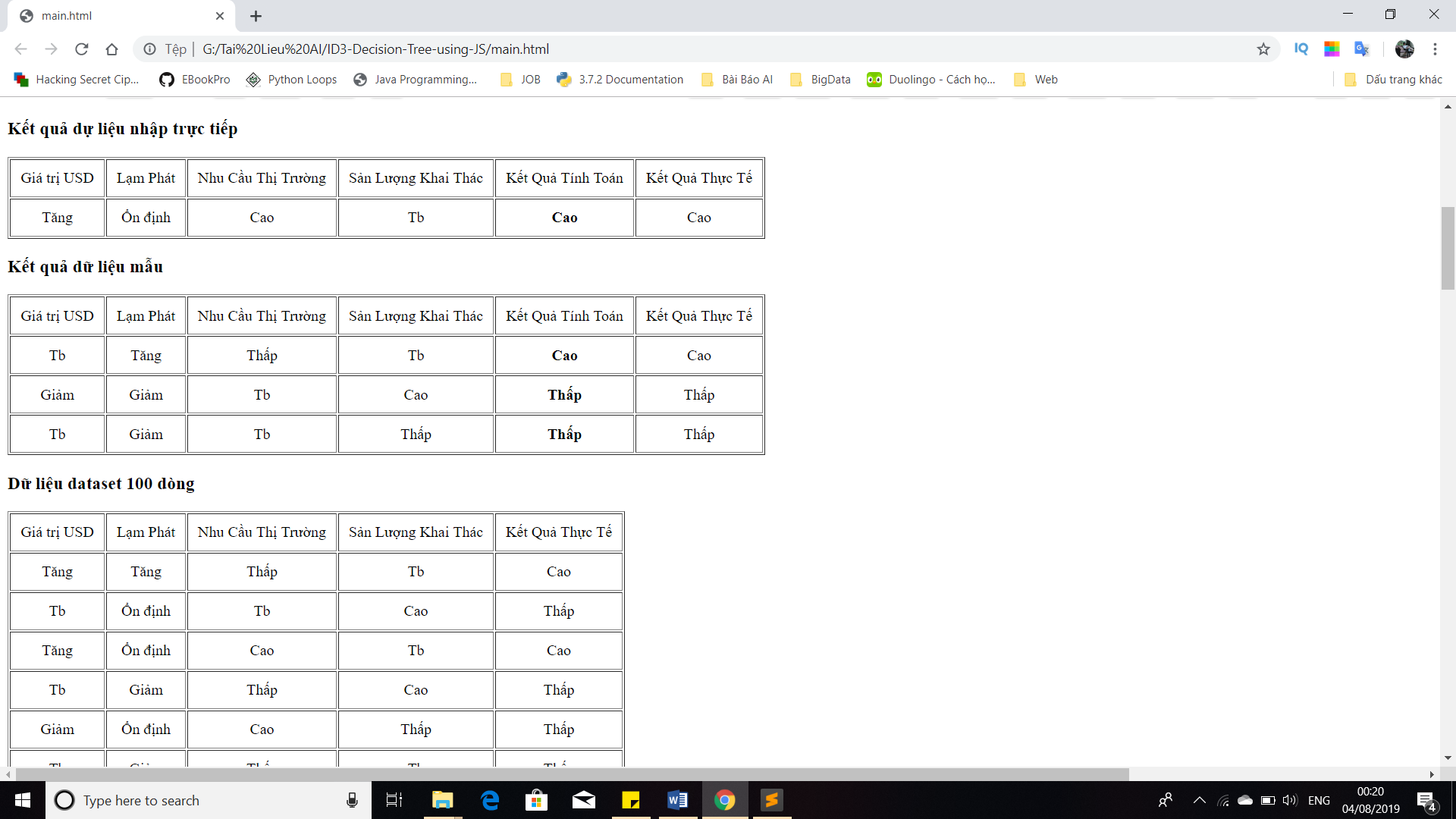
R14: if(Giá trị USD= giảm) and (SLKT=TB) and (Lạm phát=tăng) then giá vàng thấp

R15: if(Giá trị USD= giảm) and (SLKT=TB) and (Lạm phát=giảm) then giá vàng thấp

## Kết quả cài đặt thuật toán

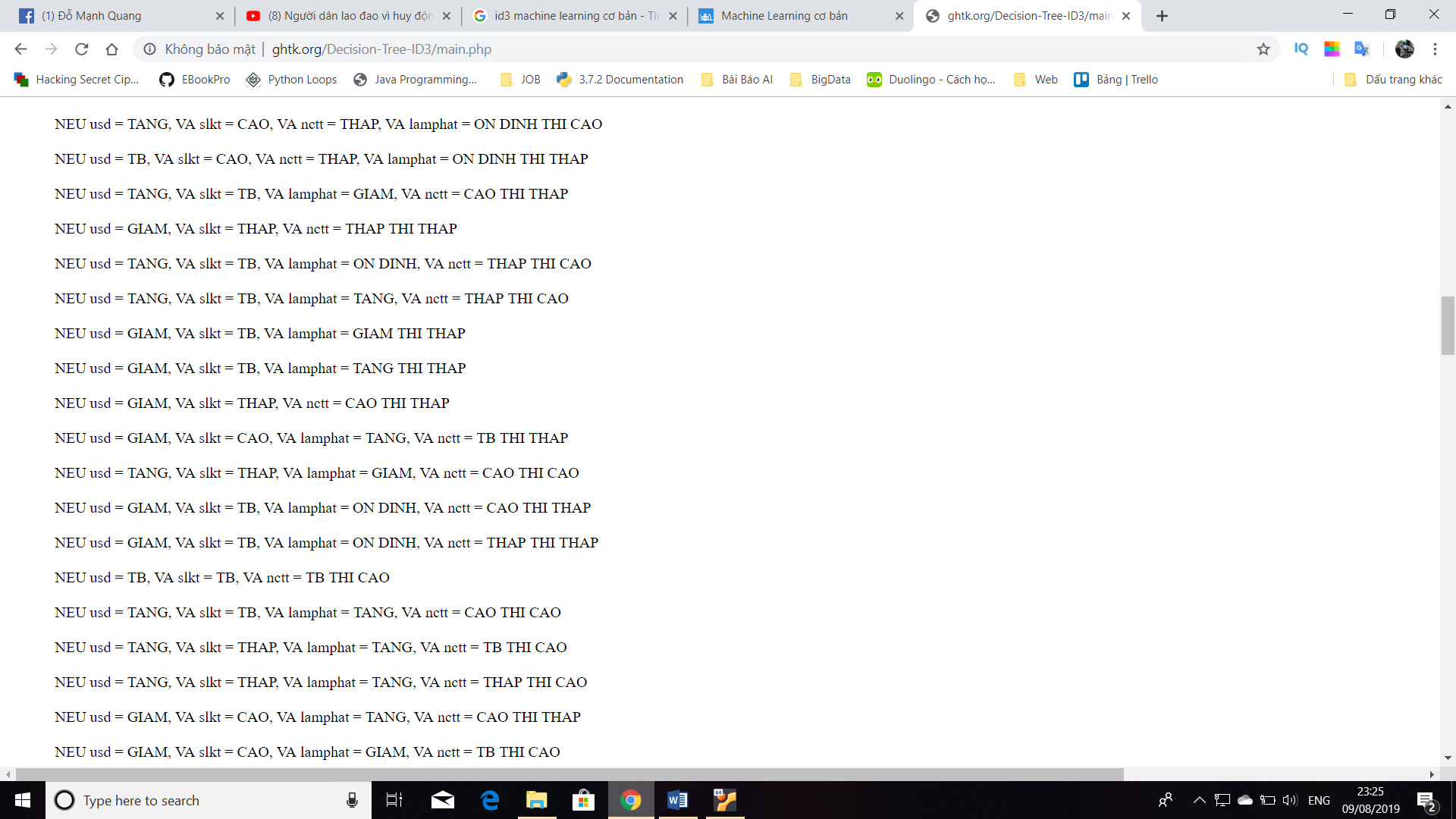














# CHƯƠNG 4 : Tài liệu tham khảo

<https://machinelearningcoban.com/2018/01/14/id3/>

<https://1upnote.me/post/2018/10/ds-ml-decision-tree-id3/>

<https://voer.edu.vn/m/hoc-bang-cach-xay-dung-cay-dinh-danh/41c26424>

<https://vi.wikipedia.org/wiki/C%C3%A2y_quy%E1%BA%BFt_%C4%91%E1%BB%8Bnh>

<https://en.wikipedia.org/wiki/ID3_algorithm>

<https://www.geeksforgeeks.org/decision-tree/>

<https://www.geeksforgeeks.org/decision-tree-introduction-example/>