

**HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY**

**SCHOL OF INFORMATION TECHNOLOGY AND COMMUNICATION**

**-----🙞🙜🕮🙞🙜-----**



**DSS PROJECT REPORT**

**Project name**: Career counseling system for students

at Hust

**Supervisor Assoc. Prof. Phạm Văn Hải**

**Student names :** **Group 8**

|  |  |
| --- | --- |
| **Nguyễn Đức Minh** | **20176042** |
| **Nguyễn Trí Vinh** | **20176066** |
| **Hoàng Ngọc Nhất** | **20176048** |

***Hà Nôi, 1 / 2022***

***Hà Nội, tháng 10 năm 2012***

Mục Lục

[**I.** **Thực trạng** 4](#_Toc94283851)

[**II.** **Tổng về phần mềm** 4](#_Toc94283852)

[**III.** **Tìm hiểu về cây quyết định** 5](#_Toc94283853)

[1. Các định nghĩa 5](#_Toc94283854)

[2. Các kiểu cây quyết định 6](#_Toc94283855)

[a. Cây hồi quy (*Regression tree*) 6](#_Toc94283856)

[b. Cây phân loại (*Classification tree*) 6](#_Toc94283857)

[3. Ưu/ Nhược điểm của thuật toán cây quyết định 6](#_Toc94283858)

[Ưu điểm: 6](#_Toc94283859)

[Cây quyết định là một thuật toán đơn giản và phổ biến. Thuật toán này được sử dụng rộng rãi bới những lợi ích của nó: 6](#_Toc94283860)

[Nhược điểm: 6](#_Toc94283861)

[Kèm với đó, cây quyết định cũng có những nhược điểm cụ thể: 6](#_Toc94283862)

[ Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của bạn. Thạm chí, với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình cây quyết định có thể thay đổi hoàn toàn. 6](#_Toc94283863)

[ Cây quyết định hay gặp vấn đề overfitting 6](#_Toc94283864)

[**Thuật toán cây quyết định ID3 hoặc C4.5** 7](#_Toc94283865)

[**1.** **Thuật toán cây quyết định** 7](#_Toc94283866)

[**2.** **Xây dựng cây quyết định gồm 2 bước:** 7](#_Toc94283867)

[**3.** **Đánh giá thuật toán cây quyết định trong lĩnh vực khai phá dữ liệu** 8](#_Toc94283868)

[**4.** **Bài Toán Tổng Quát** 9](#_Toc94283869)

[**5.** **Ví dụ cụ thể** 10](#_Toc94283870)

[**Bảng data** 10](#_Toc94283871)

[**Thuật toán:** 13](#_Toc94283872)

[Bước 1: Xét các thuộc tính: 13](#_Toc94283873)

[**a.** **Xét thuộc tính ngành nghề:** 13](#_Toc94283874)

[b. Xét thuộc tính học lực 13](#_Toc94283875)

[c. Xét thuộc tính sở thích 14](#_Toc94283876)

[d. Xét thuộc tính năng khiếu 14](#_Toc94283877)

[e. Xét thuộc tính kỹ năng mềm 14](#_Toc94283878)

[f. Xét thuộc tính ngôn ngữ 15](#_Toc94283879)

[**Bước 2: Lấy IG của thuộc tính lớn nhất** 15](#_Toc94283880)

[a. Xét ngành Trí tuệ nhân tạo 15](#_Toc94283881)

[b. Xét ngành Lập trình web 17](#_Toc94283882)

[c. Xét ngành Lập trình game 18](#_Toc94283883)

[d. Xét ngành Tư vấn viên 20](#_Toc94283884)

[e. Xét lập trình viên An Ninh Mạng 20](#_Toc94283885)

[f. Xét Lập trình ứng dụng điện thoại 22](#_Toc94283886)

[g. Xét Thiết kế Game 24](#_Toc94283887)

[h. Xét Tester 24](#_Toc94283888)

[i. Xét Kỹ sư cầu nối 25](#_Toc94283889)

[Bước 3: Tổng hợp thành cây quyết định 26](#_Toc94283890)

[**IV.** **Tài Liệu Tham Khảo** 27](#_Toc94283891)

1. **Thực trạng**

Chọn ngành nghề là một nhiệm vụ quan trọng, chi phối phần lớn suy nghĩ và hoạt động của các bạn sinh viên Đại Học. Tuy nhiên trong quá trình học tập cũng như trải nghiệm công việc thực tế vẫn còn hạn chế. Hầu hết các trường Đại học đều có tư duy là học xong sẽ bắt đầu đi tìm việc cho nên quá trình chuẩn bị thông tin, kiến thức cho công tác chọn ngành nghề phù hợp chưa được tốt. Mặt khác, nhu cầu công việc hàng năm thay đổi liên tục, việc thiếu chủ động tìm hiểu cũng như cân bằng năng lực bản thân với công việc mình mong muốn có thể dẫn đến tư duy ngành nghề không phù hợp. Khi sinh viên có những băn khoăn khác nhau về sự phù hợp nghề, về thị trường lao động, khả năng phát triển của nghề đã chọn trong tương lai... thì hầu như không có thời gian để hỏi, hoặc không được giải thích một cách thoả đáng. Chính vì vậy đã dẫn đến tình trạng thừa thầy thiếu thợ và hàng năm có rất nhiều sinh viên tốt nghiệp ở nhiều ngành nghề khác nhau nhưng các em rất khó tham gia vào thị trường lao động , một phần do sinh viên không đáp ứng được yêu cầu của nghề nghiệp, một phần do ngành nghề của các em chọn không có trong nhu cầu của thị trường lao động.

1. **Tổng về phần mềm**

* Phần mềm: sử dụng thuật toán C4.5 làm thuật toán chính cho quá trình xử lý dữ liệu.
* Xây dựng 1 bộ dữ liệu mẫu để training cho phần mềm
* Người dùng nhập thông tin vào để phần mềm ra trợ giúp quyết định và học từ phản hồi người dùng

\* Input:

- Tiêu chí:

+ Học lực: Giỏi, Khá, Trung Bình

+ Sở thích: chơi game, xem phim, lướt web, đọc sách, khoa học

+ Năng khiếu: Vẽ, hát

+ Kỹ năng mềm: giao tiếp, thuyết trình, lãnh đạo, thiết kế và quản trị

+ Ngôn ngữ lập trình: Java, C, JavaScript, Python

\* Output:

Nghề phù hợp

1. **Tìm hiểu về cây quyết định**

## Các định nghĩa

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. Mỗi một nút trong (*internal node*) tương ứng với một biến; đường nối giữa nó với nút con của nó thể hiện một giá trị cụ thể cho biến đó. Mỗi nút lá đại diện cho giá trị dự đoán của biến mục tiêu, cho trước các giá trị của các biến được biểu diễn bởi đường đi từ nút gốc tới nút lá đó. Kỹ thuật học máy dùng trong cây quyết định được gọi là học bằng cây quyết định, hay chỉ gọi với cái tên ngắn gọn là cây quyết định.

Học bằng cây quyết định cũng là một phương pháp thông dụng trong khai phá dữ liệu. Khi đó, cây quyết định mô tả một cấu trúc cây, trong đó, các lá đại diện cho các phân loại còn cành đại diện cho các kết hợp của các thuộc tính dẫn tới phân loại đó. Một cây quyết định có thể được học bằng cách chia tập hợp nguồn thành các tập con dựa theo một kiểm tra giá trị thuộc tính. Quá trình này được lặp lại một cách đệ quy cho mỗi tập con dẫn xuất. Quá trình đệ quy hoàn thành khi không thể tiếp tục thực hiện việc chia tách được nữa, hay khi một phân loại đơn có thể áp dụng cho từng phần tử của tập con dẫn xuất. Một bộ phân loại rừng ngẫu nhiên (*random forest*) sử dụng một số cây quyết định để có thể cải thiện tỉ lệ phân loại.

Cây quyết định cũng là một phương tiện có tính mô tả dành cho việc tính toán các xác suất có điều kiện.

Cây quyết định có thể được mô tả như là sự kết hợp của các kỹ thuật toán học và tính toán nhằm hỗ trợ việc mô tả, phân loại và tổng quát hóa một tập dữ liệu cho trước.

## Các kiểu cây quyết định

### Cây hồi quy (*Regression tree*)

### Cây phân loại (*Classification tree*)

## Ưu/ Nhược điểm của thuật toán cây quyết định

# Ưu điểm:

# Cây quyết định là một thuật toán đơn giản và phổ biến. Thuật toán này được sử dụng rộng rãi bới những lợi ích của nó:

* Mô hình sinh ra các quy tắc dễ hiểu cho người đọc, tạo ra bộ luật với mỗi nhánh lá là một luật của cây.
* Dữ liệu đầu vào có thể là là dữ liệu missing, không cần chuẩn hóa hoặc tạo biến giả
* Có thể làm việc với cả dữ liệu số và dữ liệu phân loại
* Có thể xác thực mô hình bằng cách sử dụng các kiểm tra thống kê
* Có khả năng là việc với dữ liệu lớn

# Nhược điểm:

# Kèm với đó, cây quyết định cũng có những nhược điểm cụ thể:

# Mô hình cây quyết định phụ thuộc rất lớn vào dữ liệu của bạn. Thạm chí, với một sự thay đổi nhỏ trong bộ dữ liệu, cấu trúc mô hình cây quyết định có thể thay đổi hoàn toàn.

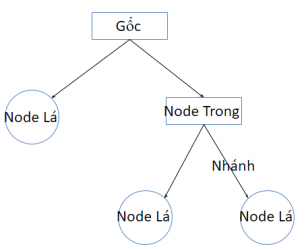
# Cây quyết định hay gặp vấn đề [overfitting](https://trituenhantao.io/tu-dien-thuat-ngu/overfitting/)

**Thuật toán cây quyết định ID3 hoặc C4.5**

1. **Thuật toán cây quyết định**

Thuật toán cây quyết định cho ra kết quả là một tập luật của những dữ liệu huấn luyện có thuộc tính. Cây quyết định là một công cụ phổ biến trong khai phá và phân lớp dữ liệu. Đặc điểm của cây quyết định: là một cây có cấu trúc, trong đó:

* Root (Gốc): Là nút trên cùng của cây.
* Node trong: nút trung gian trên một thuộc tính đơn (hình Oval).
* Nhánh: Biểu diễn các kết quả của kiểm tra trên nút
* Node lá: Biểu diễn lớp hay sự phân phối lớp (hình vuông hoặc chữ nhật)



1. **Xây dựng cây quyết định gồm 2 bước:**

* **Phát triển cây quyết định:** đi từ gốc, đến các nhánh, phát triển quy nạp theo hình thức chia để trị.
* **Bước 1.** Chọn thuộc tính “tốt” nhất bằng một độ đo đã định trước
* **Bước 2.** Phát triển cây bằng việc thêm các nhánh tương ứng với từng giá trị của thuộc tính đã chọn
* **Bước 3.** Sắp xếp, phân chia tập dữ liệu đào tạo tới node con
* **Bước 4.** Nếu các ví dụ được phân lớp rõ ràng thì dừng.

**Ngược lại:** lặp lại bước 1 tới bước 4 cho từng node con

* **Cắt tỉa cây:** nhằm đơn giản hóa, khái quát hóa cây, tăng độ chính xác

1. **Đánh giá thuật toán cây quyết định trong lĩnh vực khai phá dữ liệu**

**Thuận lợi:**

* Quá trình xây dựng cây quyết định không dùng kiến thức về lĩnh vực dữ liệu đang nghiên cứu hoặc thông số đầu vào nào.
* Kết quả của quá trình huấn luyện (học) được biểu diễn dưới dạng cây nên dễ hiểu và gần gũi với con người.
* Nhìn chung, các giải thuật cây quyết định cho kết quả có độ chính xác khá cao.

**Khó khăn:**

* Đối với các tập dữ liệu có nhiều thuộc tính thì cây quyết định sẽ lớn (về chiều sâu cả chiều ngang), vì vậy làm giảm độ dễ hiểu.
* Việc xếp hạng các thuộc tính để phân nhánh dựa vào lần phân nhánh trước đó và bỏ qua sự phụ thuộc lẫn nhau giữa các thuộc tính.
* Khi dùng độ lợi thông tin (Information Gain) để xác định thuộc tính rẽ nhánh, các thuộc tính có nhiều giá trị thường được ưu tiên chọn.

|  |  |
| --- | --- |
| Pseudocode:  ·        Kiểm tra case cơ bản  ·        Với mỗi thuộc tính A tìm thông tin nhờ việc tách thuộc tính A  ·        Chọn a\_best là thuộc tính mà độ đo lựa chọn thuộc tính “tốt nhất”  ·        Dùng a\_best làm thuộc tính cho node chia cắt cây.  ·        Đệ quy trên các danh sách phụ được tạo ra bởi việc phân chia theo a\_best, và thêm các node này như là con của node | (1)ComputerClassFrequency(T); (2) if OneClass or FewCases return a leaf; Create a decision node N; (3) ForEach Attribute A ComputeGain(A); (4)N.test=AttributeWithBestGain; (5) if (N.test is continuous) find Threshold; (6) ForEach T’ in the splitting of T (7) If ( T’ is Empty ) Child of N is a leaf else (8) Child of N=FormTree(T’); (9) ComputeErrors of N; return N |

1. **Bài Toán Tổng Quát**

Hai độ đo được sử dụng trong C4.5 là information gain và gain ratio. RF(Cj,S) biểu diễn tần xuất (Relative Frequency) các case trong S thuộc về lớp Cj.

R

Với |Sj| là kích thước tập các case có giá trị phân lớp là Cj. |S| là kích thước tập dữ liệu đào tạo.

Chỉ số thông tin cần thiết cho sự phân lớp: I(S) với S là tập cần xét sự phân phối lớp được tính bằng:

* Độ đo lựa chọn thuộc tính “tốt nhất”:

Sau khi S được phân chia thành các tập con S1, S2,…, St bởi test B thì information gain được tính bằng:

* + *Information gain: Test B chia* S={S1,S2,…,St)
* Test B sẽ được chọn nếu có G(S, B) đạt giá trị lớn nhất.

Tuy nhiên có một vấn đề khi sử dụng G(S,B) ưu tiên test có số lượng lớn kết quả, ví dụ G(S,B) đạt cực đại với test mà từng Si chỉ chứa một case đơn. Tiêu chuẩn gain ratio giải quyết được vấn đề này bằng việc đưa vào thông tin tiềm năng (potential information) của bản thân mỗi phân hoạch.

*Thông tin tiềm năng (potential information) của bản thân mỗi phân*

*hoạch:*

Gain ratio = G(S, B) / P(S, B) lớn nhất => chọn test B

Trong mô hình phân lớp C4.5 , có thể dùng một trong hai loại chỉ số Information Gain hay Gain ratio để xác định thuộc tính tốt nhất. Trong đó Gain ratio là lựa chọn mặc định.

1. **Ví dụ cụ thể**

**ÁP DỤNG THUẬT TOÁN CÂY QUYẾT ĐỊNH**

*Trong bài toán hỗ trợ ra quyết định chọn nghề nghiệp cho sinh viên*

**Bảng data**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| STT | Ngành nghề | Học lực | Sở thích | Năng khiếu | Kỹ năng mềm | Ngôn ngữ lập trình | SPDecision |
| 1 | Trí tuệ nhân tạo | G | Khoa học | Không | Phân tích | Python | Yes |
| 2 | Trí tuệ nhân tạo | G | Không | Không | Thiết kế | Python | Yes |
| 3 | Trí tuệ nhân tạo | K | Khoa học | Không | Thiết kế | Python | No |
| 4 | Trí tuệ nhân tạo | TB | Không | Không | Thuyết trình | Python | No |
| 5 | Lập trình web | G | Lướt Web | Có | Thiết kế | Javascript | Yes |
| 6 | Lập trình web | K | Lướt Web | Có | Thiết kế | Javascript | Yes |
| 7 | Lập trình web | TB | Không | Không | Thiết kế | Javascript | No |
| 8 | Lập trình game | G | Chơi Game | Không | Thiết kế | Java | Yes |
| 9 | Lập trình game | K | Chơi Game | Không | Thiết kế | Java | Yes |
| 10 | Lập trình game | TB | Lướt web | Không | Thiết kế | Java | No |
| 11 | Tư vấn viên | K | Lướt Web | Không | Thuyết trình | C | Yes |
| 12 | Tư vấn viên | TB | Lướt Web | Không | Giao tiếp | C | Yes |
| 13 | Lập trình viên an ninh mạng | G | Khoa học | Có | Thiết Kế | C | Yes |
| 14 | Lập trình viên an ninh mạng | K | Khoa học | Có | Thiết kê | C | No |
| 15 | Lập trình viên an ninh mạng | TB | Khoa học | Có | Không | C | No |
| 16 | Lập trình ứng dụng điện thoại | G | Không | Không | Thuyết trình | Javascript | No |
| 17 | Lập trình ứng dụng điện thoại | K | Khoa học | Có | Phân tích | Javascript | Yes |
| 18 | Lập trình ứng dụng điện thoại | TB | Khoa học | Không | Không | C | No |
| 19 | Thiết kế Game | G | Chơi game | Có | Không | Java | Yes |
| 20 | Thiết kế game | K | Chơi game | có | Thiết kế | C | Yes |
| 21 | Thiết kế Game | TB | Chơi game | Có | Phân tích | C | Yes |
| 22 | Tester | G | Không | Không | Phân tích | Python | No |
| 23 | Tester | K | Không | Không | Không | Python | No |
| 24 | Tester | TB | Chơi game | không | không | C | No |
| 25 | Kỹ sư cầu nối | G | Không | Không | Giao tiếp | Java | No |
| 26 | Kỹ sư cầu nối | K | Lướt web | Có | Giao tiếp | C | Yes |
| 27 | Kỹ sư cầu nối | TB | Lướt web | Có | Giao tiếp | Java | Yes |

**Thuật toán:**

Đặt Yes: Y (có 15 kết quả)

Đặt No: N (có 12 kết quả)

Ta có entropy(D) = - = 0.991

Bước 1: Xét các thuộc tính:

1. **Xét thuộc tính ngành nghề:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Ngành nghề** | **Yi** | **Ni** | **I(Yi,Ni)** |
| Trí tuệ nhân tạo | 2 | 2 | 1 |
| Lập trình web | 2 | 1 | 0,918 |
| Lập trình game | 2 | 1 | 1 |
| Tư vấn viên | 2 | 0 | 0 |
| Lập trình viên an ninh mạng | 1 | 2 | 0,918 |
| Lập trình ứng dụng điện thoại | 1 | 2 | 0,918 |
| Thiết kế Game | 3 | 0 | 0 |
| Tester | 0 | 3 | 0 |
| Kỹ sư cầu nối | 2 | 1 | 0,918 |

Remainder (Ngành nghề) D = = 0,667

* IG(Ngành nghề) = Entropy(D) - Remainder (Ngành nghề) D = 0.991– 0.533 = **0.324**

1. Xét thuộc tính học lực

|  |  |  |  |
| --- | --- | --- | --- |
| **Học lực** | **Yi** | **Ni** | **I(Yi,Ni)** |
| G | 6 | 3 | 0,918 |
| K | 6 | 3 | 0,918 |
| TB | 3 | 6 | 0,918 |

Remainder (Học lực) D = = 0,918

* IG(Ngành nghề) = Entropy(D) - Remainder (Học lực) D = 0.991– 0.918= 0.**073**

1. Xét thuộc tính sở thích

|  |  |  |  |
| --- | --- | --- | --- |
| **Sở thích** | **Yi** | **Ni** | **I(Yi,Ni)** |
| Khoa học | **3** | **4** | **0,985** |
| Không | **1** | **6** | **0,592** |
| Lướt web | 6 | 1 | 0,592 |
| Chơi game | 5 | 1 | 0,65 |

Remainder (Sở thích) D = = 0,707

* IG(Sở thích) = Entropy(D) - Remainder (Sở thích) D = 0.991– 0.707 = 0.**284**

1. Xét thuộc tính năng khiếu

|  |  |  |  |
| --- | --- | --- | --- |
| Năng khiếu | Yi | Ni | I(Yi,Ni) |
| Có | 9 | 2 | 0,684 |
| Không | 6 | 10 | 0,954 |

Remainder (Năng khiếu) D = = 0,844

* IG(Năng khiếu) = Entropy(D) - Remainder (Năng khiếu) D = 0.991– 0.844= **0.147**

1. Xét thuộc tính kỹ năng mềm

|  |  |  |  |
| --- | --- | --- | --- |
| Kĩ năng | Yi | Ni | I(Yi,Ni) |
| Phân tích | 3 | 1 | 0,811 |
| Thiết kế | 7 | 4 | 0,946 |
| Thuyết trình | 1 | 2 | 0,918 |
| Không | 1 | 4 | 0,722 |
| Giao tiếp | 3 | 1 | 0,811 |

Remainder (Kĩ năng) D = = 0,861

* IG(Kĩ năng) = Entropy(D) - Remainder (Kĩ năng) D = 0.991– 0.861= **0.13**

1. Xét thuộc tính ngôn ngữ

|  |  |  |  |
| --- | --- | --- | --- |
| Ngôn ngữ | Yi | Ni | I(Yi,Ni) |
| Python | 2 | 4 | 0,918 |
| Javascript | 3 | 2 | 0.971 |
| Java | 4 | 2 | 0,918 |
| C | 6 | 4 | 0,971 |

Remainder (Ngôn ngữ) D = = 0,948

* IG(ngôn ngữ) = Entropy(D) - Remainder (Kĩ năng) D = 0.991– 0.948= **0.043**

**Bước 2: Lấy IG của thuộc tính lớn nhất**

Ta có: IG(ngành nghề) = **0,324 là lớn nhất**

Có 9 giá trị ngành nghề tất cả nên ta sẽ xét ở từng trường hợp

1. Xét ngành Trí tuệ nhân tạo

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| STT | Ngành nghề | Học lực | Sở thích | Năng khiếu | Kỹ năng mềm | Ngôn ngữ lập trình | SPDecision |
| 1 | Trí tuệ nhân tạo | G | Khoa học | Không | Phân tích | Python | Yes |
| 2 | Trí tuệ nhân tạo | G | Không | Không | Thiết kế | Python | Yes |
| 3 | Trí tuệ nhân tạo | K | Khoa học | Không | Thiết kế | Python | No |
| 4 | Trí tuệ nhân tạo | TB | Không | Không | Thuyết trình | Python | No |

Ta có Entropi(D1) = - =1

* Xét thuộc tính học lực

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| G | 2 | 2 | 1 |
| K | 0 | 1 | 0 |
| TB | 0 | 1 | 0 |

Remainder (Học lực) D1 = = 0,148

* IG(Học lực) = Entropy(D1) - Remainder (Học lực) D1 = 1– 0.148= **0.852**
* Xét thuộc tính sở thích

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| Khoa học | 1 | 1 | 1 |
| Không | 1 | 1 | 1 |

Remainder (Sở thích) D1 = = 0,148

* IG(sở thích) = Entropy(D1) - Remainder (sở thích) D1 = 1– 0.148 = **0.852**
* Xét thuộc tính năng khiếu

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| Không | 2 | 2 | 1 |

Remainder (Năng khiếu) D1 = = 0,148

* IG(Năng khiếu) = Entropy(D1) - Remainder (Năng khiếu) D1 = 1– 0.148 = **0.852**
* Xét thuộc tính kĩ năng mềm

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| Phân tích | 1 | 0 | 0 |
| Thiết kế | 1 | 1 | 1 |
| Không | 0 | 1 | 0 |

Remainder (Kĩ năng) D1 = = 0,074

* IG(Kĩ năng)D1 = Entropy(D1) - Remainder (Kỹ năng) D1 = 1-0,074 = **0.926**
* Xét thuộc tính ngôn ngữ lập trình

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| Python | 2 | 2 | 1 |

Remainder (Ngôn ngữ) D1 = = 0,148

* IG(Ngôn ngữ)D1 = Entropy(D1) - Remainder (Ngôn ngữ) D1 = 1-0,148 = **0.852**

1. Xét ngành Lập trình web

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| STT | Ngành nghề | Học lực | Sở thích | Năng khiếu | Kỹ năng mềm | Ngôn ngữ lập trình | SPDecision |
| 5 | Lập trình web | G | Lướt Web | Có | Thiết kế | Javascript | Yes |
| 6 | Lập trình web | K | Lướt Web | Có | Thiết kế | Javascript | Yes |
| 7 | Lập trình web | TB | Không | Không | Thiết kế | Javascript | No |

Ta có Entropi(D2) = - =0,918

* Xét thuộc tính học lực

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| G | 1 | 0 | 0 |
| K | 1 | 0 | 0 |
| TB | 0 | 1 | 0 |

Remainder (Học lực) D2= = 0

* IG(Học lực)D2 = Entropy(D2) - Remainder (Học lực) D2 = 0,918– 0= **0.918**
* Xét thuộc tính sở thích

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| Lướt web | 2 | 0 | 0 |
| Không | 0 | 1 | 0 |

Remainder (Sở thích) D2 = = 0

* IG(sở thích)D2 = Entropy(D2) - Remainder (sở thích) D2 = 0,918 = **0.918**
* Xét thuộc tính năng khiếu

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| Không | 0 | 1 | 0 |
| có | 2 | 0 | 0 |

Remainder (Năng khiếu) D2 = = 0

* IG(Năng khiếu)D2 = Entropy(D2) - Remainder (Năng khiếu) D2 = 0,918– 0 = **0.918**
* Xét thuộc tính kĩ năng mềm

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| Thiết kế | 2 | 1 | 0,918 |

Remainder (Kĩ năng) D2 = = 0,102

* IG(Kĩ năng)D2 = Entropy(D2) - Remainder (Kỹ năng) D2 = 0,918-0,102 = **0.816**
* Xét thuộc tính ngôn ngữ lập trình

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| Javascript | 2 | 1 | 0,918 |

Remainder (Ngôn ngữ) D2 = = 0,102

* IG(Ngôn ngữ)D2 = Entropy(D2) - Remainder (Ngôn ngữ) D2 = 0,918-0,102 = **0.816**

1. Xét ngành Lập trình game

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| STT | Ngành nghề | Học lực | Sở thích | Năng khiếu | Kỹ năng mềm | Ngôn ngữ lập trình | SPDecision |
| 8 | Lập trình game | G | Chơi Game | Không | Thiết kế | Java | Yes |
| 9 | Lập trình game | K | Chơi Game | Không | Thiết kế | Java | Yes |
| 10 | Lập trình game | TB | Lướt web | Không | Thiết kế | Java | No |

Ta có Entropi(D3) = - =0,918

* Xét thuộc tính học lực

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| G | 1 | 0 | 0 |
| K | 1 | 0 | 0 |
| TB | 0 | 1 | 0 |

Remainder (Học lực) D3= = 0

* IG(Học lực)D3 = Entropy(D3) - Remainder (Học lực) D3 = 0,918– 0= **0.918**
* Xét thuộc tính sở thích

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| Lướt web | 0 | 1 | 0 |
| Chơi game | 2 | 0 | 0 |

Remainder (Sở thích) D3 = = 0

* IG(sở thích)D3 = Entropy(D3) - Remainder (sở thích) D3 = 0,918 = **0.918**
* Xét thuộc tính năng khiếu

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| Không | 2 | 1 | 0,918 |

Remainder (Năng khiếu) D3 = = 0,102

* IG(Năng khiếu)D3 = Entropy(D3) - Remainder (Năng khiếu) D3 = 0,918– 0,102 = **0.816**
* Xét thuộc tính kĩ năng mềm

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| Thiết kế | 2 | 1 | 0,918 |

Remainder (Kĩ năng) D3 = = 0,102

* IG(Kĩ năng)D3 = Entropy(D3) - Remainder (Kỹ năng) D3 = 0,918-0,102 = **0.816**
* Xét thuộc tính ngôn ngữ lập trình

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| Java | 2 | 1 | 0,918 |

Remainder (Ngôn ngữ) D3 = = 0,102

* IG(Ngôn ngữ)D3 = Entropy(D3) - Remainder (Ngôn ngữ) D3 = 0,918-0,102 = **0.816**

1. Xét ngành Tư vấn viên

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| STT | Ngành nghề | Học lực | Sở thích | Năng khiếu | Kỹ năng mềm | Ngôn ngữ lập trình | SPDecision |
| 11 | Tư vấn viên | K | Lướt Web | Không | Thuyết trình | C | Yes |
| 12 | Tư vấn viên | TB | Lướt Web | Không | Giao tiếp | C | Yes |

Ta có Entropi(D4) = =0

1. Xét lập trình viên An Ninh Mạng

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| STT | Ngành nghề | Học lực | Sở thích | Năng khiếu | Kỹ năng mềm | Ngôn ngữ lập trình | SPDecision |
| 13 | Lập trình viên an ninh mạng | G | Khoa học | Có | Thiết Kế | C | Yes |
| 14 | Lập trình viên an ninh mạng | K | Khoa học | Có | Thiết kê | C | No |
| 15 | Lập trình viên an ninh mạng | TB | Khoa học | Có | Không | C | No |

Ta có Entropi(D5) = - =0,918

* Xét thuộc tính học lực

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| G | 1 | 0 | 0 |
| K | 0 | 1 | 0 |
| TB | 0 | 1 | 0 |

Remainder (Học lực) D5= = 0

* IG(Học lực)D5 = Entropy(D5) - Remainder (Học lực) D5 = 0,918– 0= **0.918**
* Xét thuộc tính sở thích

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| Khoa Học | 1 | 2 | 0,918 |

Remainder (Sở thích) D5 = = 0,102

* IG(sở thích)D5 = Entropy(D5) - Remainder (sở thích) D5 = 0,918 – 0,102 = **0.816**
* Xét thuộc tính năng khiếu

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| có | 1 | 2 | 0,918 |

Remainder (Năng khiếu) D5 = = 0,102

* IG(Năng khiếu)D5 = Entropy(D5) - Remainder (Năng khiếu) D5= 0,918– 0,102 = **0.816**
* Xét thuộc tính kĩ năng mềm

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| Thiết kế | 1 | 1 | 1 |
| Không | 0 | 1 | 0 |

Remainder (Kĩ năng) D5 = = 0,074

* IG(Kĩ năng)D5 = Entropy(D5) - Remainder (Kỹ năng) D5 = 0,918-0,074 = **0.844**
* Xét thuộc tính ngôn ngữ lập trình

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| C | 1 | 2 | 0,918 |

Remainder (Ngôn ngữ) D5 = = 0,102

* IG(Ngôn ngữ)D5 = Entropy(D5) - Remainder (Ngôn ngữ) D5 = 0,918-0,102 = **0.816**

1. Xét Lập trình ứng dụng điện thoại

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| STT | Ngành nghề | Học lực | Sở thích | Năng khiếu | Kỹ năng mềm | Ngôn ngữ lập trình | SPDecision |
| 16 | Lập trình ứng dụng điện thoại | G | Không | Không | Thuyết trình | Javascript | No |
| 17 | Lập trình ứng dụng điện thoại | K | Khoa học | Có | Phân tích | Javascript | Yes |
| 18 | Lập trình ứng dụng điện thoại | TB | Khoa học | Không | Không | C | No |

Ta có Entropi(D6) = - =0,918

* Xét thuộc tính học lực

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| G | 0 | 1 | 0 |
| K | 1 | 0 | 0 |
| TB | 0 | 1 | 0 |

Remainder (Học lực) D6= = 0

* IG(Học lực)D6 = Entropy(D6) - Remainder (Học lực) D6 = 0,918– 0= **0.918**
* Xét thuộc tính sở thích

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| Khoa Học | 1 | 1 | 1 |
| Không | 0 | 1 | 0 |

Remainder (Sở thích) D6 = = 0,074

* IG(sở thích)D6 = Entropy(D6) - Remainder (sở thích) D6 = 0,918 – 0,074 = **0.844**
* Xét thuộc tính năng khiếu

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| có | 1 | 0 | 0 |
| Không | 0 | 2 | 0 |

Remainder (Năng khiếu) D6 = = 0

* IG(Năng khiếu)D6 = Entropy(D6) - Remainder (Năng khiếu) D6= 0,918– 0 = **0.918**
* Xét thuộc tính kĩ năng mềm

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| Thuyết trình | 0 | 1 | 0 |
| Không | 0 | 1 | 0 |
| Phân tích | 1 | 0 | 0 |

Remainder (Kĩ năng) D6 = 0

* IG(Kĩ năng)D6 = Entropy(D6) - Remainder (Kỹ năng) D6 = 0,918-0 = **0.918**
* Xét thuộc tính ngôn ngữ lập trình

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| C | 0 | 1 | 0 |
| Javascript | 1 | 1 | 1 |

Remainder (Ngôn ngữ) D6= = 0,074

* IG(Ngôn ngữ)D6 = Entropy(D6) - Remainder (Ngôn ngữ) D6 = 0,918-0,074 = **0.844**

1. Xét Thiết kế Game

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| STT | Ngành nghề | Học lực | Sở thích | Năng khiếu | Kỹ năng mềm | Ngôn ngữ lập trình | SPDecision |
| 19 | Thiết kế Game | G | Chơi game | Có | Không | Java | Yes |
| 20 | Thiết kế game | K | Chơi game | có | Thiết kế | C | Yes |
| 21 | Thiết kế Game | TB | Chơi game | Có | Phân tích | C | Yes |

Ta có Entropi(D7) = =0

1. Xét Tester

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| STT | Ngành nghề | Học lực | Sở thích | Năng khiếu | Kỹ năng mềm | Ngôn ngữ lập trình | SPDecision |
| 22 | Tester | G | Không | Không | Phân tích | Python | No |
| 23 | Tester | K | Không | Không | Không | Python | No |
| 24 | Tester | TB | Chơi game | không | không | C | No |

Ta có Entropi(D8) = =0

1. Xét Kỹ sư cầu nối

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| STT | Ngành nghề | Học lực | Sở thích | Năng khiếu | Kỹ năng mềm | Ngôn ngữ lập trình | SPDecision |
| 25 | Kỹ sư cầu nối | G | Không | Không | Giao tiếp | Java | No |
| 26 | Kỹ sư cầu nối | K | Lướt web | Có | Giao tiếp | C | Yes |
| 27 | Kỹ sư cầu nối | TB | Lướt web | Có | Giao tiếp | Java | Yes |

Ta có Entropi(D9) = - =0,918

* Xét thuộc tính học lực

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| G | 0 | 1 | 0 |
| K | 1 | 0 | 0 |
| TB | 1 | 0 | 0 |

Remainder (Học lực) D9= = 0

* IG(Học lực)D9 = Entropy(D9) - Remainder (Học lực) D9 = 0,918– 0= **0.918**
* Xét thuộc tính sở thích

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| Lướt web | 2 | 0 | 0 |
| Không | 0 | 1 | 0 |

Remainder (Sở thích) D9 = 0

* IG(sở thích)D9 = Entropy(D9) - Remainder (sở thích) D9 = 0,918 – 0= **0.918**
* Xét thuộc tính năng khiếu

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| có | 2 | 0 | 0 |
| Không | 0 | 1 | 0 |

Remainder (Năng khiếu) D9 = = 0

* IG(Năng khiếu)D9 = Entropy(D9) - Remainder (Năng khiếu) D9= 0,918– 0 = **0.918**
* Xét thuộc tính kĩ năng mềm

|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| Phân tích | 2 | 1 | 0,918 |

Remainder (Kĩ năng) D9 = =0,102

* IG(Kĩ năng)D9= Entropy(D9) - Remainder (Kỹ năng) D9 = 0,918-0,102 = **0.816**
* Xét thuộc tính ngôn ngữ lập trình

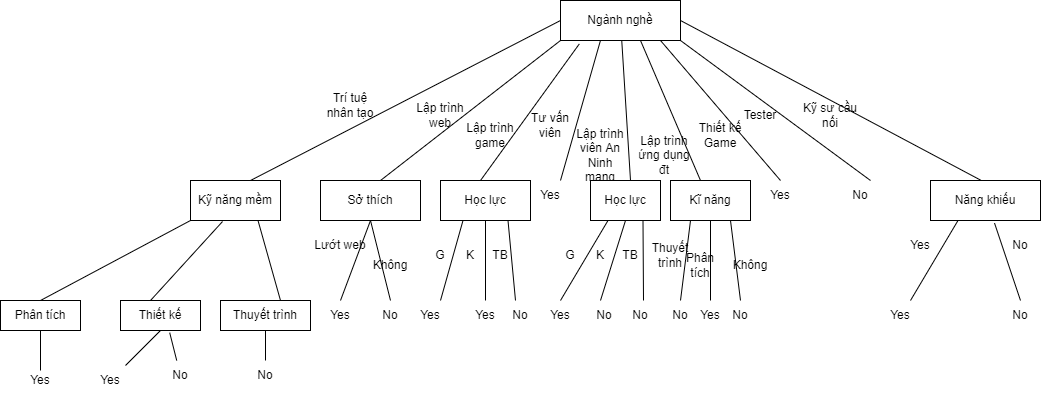
|  |  |  |  |
| --- | --- | --- | --- |
|  | Yi | Ni | I(Yi,Ni) |
| C | 1 | 0 | 0 |
| Java | 1 | 1 | 1 |

Remainder (Ngôn ngữ) D9= = 0,074

* IG(Ngôn ngữ)D9 = Entropy(D9) - Remainder (Ngôn ngữ) D9 = 0,918-0,074 = **0.844**

Bước 3: Tổng hợp thành cây quyết định

Sau khi tổng hợp và lấy ra các thuộc tính có giá trị lớn nhất ta tổng hợp được cây quyết định:



1. **Tài Liệu Tham Khảo**

**1.**[**https://trituenhantao.io/kien-thuc/decision-tree/**](https://trituenhantao.io/kien-thuc/decision-tree/)

**2.** [**https://sinhvientot.net/thuat-toan-cay-quyet-dinh-c45/**](https://sinhvientot.net/thuat-toan-cay-quyet-dinh-c45/)

**3.** [**https://www.programmersought.com/article/55246223413/**](https://www.programmersought.com/article/55246223413/)

**4.** [**https://programmerall.com/article/3460838855/**](https://programmerall.com/article/3460838855/)