

**BỘ GIÁO DỤC VÀ ĐÀO TẠO**

**ĐẠI HỌC ĐÀ NẴNG**

**VŨ CA GIÁP**

**XÂY DỰNG HỆ TRỢ GIÚP RA QUYẾT ĐỊNH  
TRONG CÔNG TÁC TUYỂN SINH ĐẠI HỌC  
HỆ TỪ XA TẠI VIỆN ĐẠI HỌC MỞ HÀ NỘI**

**Chuyên ngành: KHOA HỌC MÁY TÍNH**

**Mã số: 60.48.01**

**TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT**

**Đà Nẵng - Năm 2012**

Công trình được hoàn thành tại  
**ĐẠI HỌC ĐÀ NẴNG**

Người hướng dẫn khoa học: **PGS.TS. Phan Huy Khánh**

Phản biện 1: **TS. NGUYỄN NGỌC CHÂU**

Phản biện 2: **GS.TS. NGUYỄN THANH THỦY**

Luận văn được bảo vệ trước Hội đồng chấm Luận văn tốt nghiệp thạc sĩ kỹ thuật họp tại Đại học Đà Nẵng vào ngày 04 tháng 03 năm 2012.

*Có thể tìm hiểu luận văn tại:*

- Trung tâm Thông tin – Học liệu, Đại học Đà Nẵng
- Trung tâm Học liệu, Đại học Đà Nẵng

## MỞ ĐẦU

### 1. Lý do chọn đề tài

Hiện nay, kinh tế thế giới đang chuyển nhanh sang kinh tế tri thức, với sự đóng góp của các ngành có chất lượng tri thức cao như: công nghệ thông tin, công nghệ sinh học, công nghệ nano,... Trong xu thế hội nhập với thế giới, Việt Nam không thể đứng ngoài dòng chảy của hướng phát triển này, với thực tế là nhu cầu nhân lực của các thành phần kinh tế trong cả nước và nhu cầu học tập của lực lượng lao động rất lớn (*phụ lục 1*) [22].

Tuy nhiên, trong hoàn cảnh kinh tế đất nước còn nhiều khó khăn, mặc dù đã ưu tiên trích một tỷ lệ ngân sách quốc gia khá cao cho giáo dục đào tạo, nhưng vẫn không thể đáp ứng được nhu cầu về cơ sở vật chất, trang thiết bị, giáo trình,... cho các trường đại học, cao đẳng,... Vì vậy chủ trương xã hội hóa giáo dục để huy động mọi nguồn lực cho giáo dục, đào tạo là một hướng đi tất yếu của Nhà nước ta trong thời điểm hiện nay.

Trong bối cảnh đó, Viện Đại học Mở Hà Nội (VDHMHN) đã được thành lập ngày 03/11/1993 theo quyết định 535/TTg của Thủ tướng Chính phủ. Với nhiệm vụ chính trị là liên kết với các trường Đại học, Cao đẳng và các Trung tâm Giáo dục thường xuyên tại các tỉnh, thành phố trên cả nước để đào tạo và phát triển đại học hệ Từ xa, nhằm đào tạo nguồn nhân lực có chất lượng phục vụ cho sự nghiệp phát triển kinh tế xã hội, đặc biệt là sự nghiệp Công nghiệp hóa - Hiện đại hóa của đất nước trong giai đoạn hiện nay và lâu dài.

Qua quá trình phát triển, đến nay nhà trường đã tuyển sinh cho hệ từ xa được 85.759 học viên (*phục lục 3a*), số học viên đã tốt nghiệp 23.741 (*phục lục 3b*), số học viên đang theo học 41.928 (*phụ lục 2*). Để đạt được những thành tựu này là nhờ sự quan tâm, chỉ đạo sát sao của Bộ Giáo dục và Đào tạo, sự phấn đấu không ngừng của tập thể lãnh

đạo, giáo viên, cán bộ công nhân viên Viện Đại học Mở Hà Nội trong suốt những năm tháng qua.

Những thành tựu mà nhà trường đạt được là rất đáng khích lệ. Tuy nhiên, vẫn còn đó những tồn tại cần khắc phục trong công tác quản lý, đào tạo và nhất là công tác tuyển sinh. Qua số liệu thống kê cho thấy, số lượng tuyển sinh của VĐHMHN từ năm 2007 đến năm 2010 ngày càng giảm dần (*phụ lục 3a*) [8].

Để giải quyết vấn đề khó khăn trong công tác tuyển sinh, ngoài việc đảm bảo chất lượng đào tạo, mở rộng mã ngành thì VĐHMHN cần phải tăng cường đầu tư cho việc quảng bá hình ảnh nhà trường, nhằm tăng cường tính cạnh tranh đối với các trường có tuyển sinh hệ từ xa (*phụ lục 2*).

Ngoài ra, nhà trường cần chú trọng phát triển liên kết đào tạo và mở rộng địa bàn tuyển sinh. Tính đến thời điểm này, VĐHMHN đã thiết lập 68 trạm đào tạo từ xa đóng trên địa bàn của trên 30 tỉnh/thành phố, tạo thành mạng lưới trải dài từ Tây Bắc đến Tây Nguyên và Nam Bộ nhằm phục vụ nhu cầu học tập đông đảo của người dân từ thành thị đến nông thôn, từ miền núi đến hải đảo (*phụ lục 4*) [24].

Tuy nhiên, kết quả của việc đầu tư này vẫn chưa đạt được như mong muốn, số lượng các đợt tuyển sinh thất bại (không đủ số lượng học viên để mở lớp) tại các địa bàn liên kết đào tạo vẫn còn ở mức cao (*phụ lục 5*), dẫn đến những hậu quả không như mong muốn là uy tín, thương hiệu của nhà trường bị ảnh hưởng, kinh phí đầu tư cho công tác tuyển sinh tốn kém (*phụ lục 6*), ảnh hưởng không nhỏ đến cơ hội học tập của người dân.

Qua nghiên cứu đánh giá, nhà trường đã rút ra một số nguyên nhân chính dẫn đến việc tuyển sinh thất bại trong thời gian qua:

Ngành tuyển sinh chưa phù hợp với nhu cầu của người học; vùng miền, đặc thù phát triển kinh tế-xã hội tại địa phương.

Chưa có biện pháp nghiên cứu, đánh giá mức độ cạnh tranh của các trường

Khả năng thống kê, điều tra và đánh giá số liệu liên quan đến quá trình đào tạo nói chung và tuyển sinh nói riêng của nhà trường, của các địa phương liên kết trong những năm qua còn hạn chế. Do đó, chưa có cái nhìn bao quát, xuyên suốt dẫn đến việc quyết định địa điểm, ngành tuyển sinh chưa phù hợp,... đây là nguyên nhân chính của vấn đề thất bại trong công tác tuyển sinh.

Chưa ứng dụng CNTT để trợ giúp cho công tác ra quyết định trong khâu xác định ngành nghề khi tuyển sinh.

Để góp phần tăng cường chất lượng quyết định trong việc chọn ngành nghề tuyển sinh, thì việc ứng dụng công nghệ thông tin vào công tác này là tất yếu vì:

Môi trường ứng dụng CNTT đang phát triển mạnh tại VDHMHN.

Các dữ liệu về người học (nhu cầu, tài chính, năng lực,...), định hướng phát triển kinh tế - xã hội, ngành nghề ưu tiên phát triển tại địa bàn tuyển sinh,... rất nhiều và đa dạng. Vì vậy, nếu không ứng dụng CNTT, cán bộ chuyên trách tuyển sinh của nhà trường và của địa phương khó lòng nắm vững, tổng hợp để có những thông tin hữu ích phục vụ ra quyết định đúng đắn, kịp thời.

Chính vì những lý do trên, tôi quyết định chọn đề tài **“Xây dựng hệ trợ giúp ra quyết định trong công tác tuyển sinh Đại học hệ Từ xa tại Viện Đại học Mở Hà Nội”** nhằm hỗ trợ cho phòng đào tạo, ban giám hiệu nhà trường và lãnh đạo các đơn vị liên kết có một cách đánh giá bao quát về hiệu quả tuyển sinh, đào tạo trong những năm qua. Từ đó có biện pháp khắc phục các mặt còn hạn chế nhằm đưa ra những chiến lược phát triển hợp lý trong thời gian đến. Ngoài ra, hệ thống còn giúp đưa ra các quyết định nhanh chóng, kịp thời khi chọn ngành nghề

tuyển sinh, góp phần giảm thiểu số lượng các đợt tuyển sinh thất bại của nhà trường đã từng xảy ra trong thời gian qua.

## **2. Mục tiêu và nhiệm vụ nghiên cứu**

Mục tiêu mà đề tài hướng đến là xây dựng và áp dụng có hiệu quả cho việc trợ giúp ra quyết định trong công tác tuyển sinh đại học hệ Từ xa tại Viện Đại học Mở Hà Nội. Để thực hiện mục đích ý tưởng đề ra cần nghiên cứu và tiến hành triển khai các nội dung như sau:

Tìm hiểu, phân tích thực trạng công tác tuyển sinh của nhà trường và các đơn vị liên kết để đề ra giải pháp hợp lý trong việc xây dựng và triển khai hệ thống.

Nghiên cứu các thuật toán liên quan đến cây quyết định. Phân tích, đánh giá và triển khai áp dụng thuật toán C4.5 để ứng dụng trong công tác tuyển sinh.

Áp dụng cơ sở lý thuyết nền tảng để xây dựng và triển khai ứng dụng.

## **3. Đối tượng và phạm vi nghiên cứu**

Từ yêu cầu của đề tài, ta xác định được đối tượng và phạm vi nghiên cứu của đề tài cụ thể như sau:

### ***Đối tượng nghiên cứu:***

Học viên đang theo học hoặc đã tốt nghiệp ra trường.

Nhu cầu xã hội đối với các ngành mà nhà trường tuyển sinh

Quy trình tuyển sinh đại học hệ Từ xa.

Sự phát triển về quy mô đào tạo hệ từ xa của các trường khác.

Các vấn đề về lý thuyết trợ giúp quyết định để đề ra giải pháp ứng dụng vào hệ thống trợ giúp quyết định phục vụ công tác tuyển sinh.

### ***Phạm vi nghiên cứu:***

Số liệu thống kê về công tác tuyển sinh, đào tạo của nhà trường trong 10 năm gần đây.

Số liệu thống kê về tình hình học viên sau khi tốt nghiệp.

Nghiên cứu về đặc thù vùng miền, các thành phần kinh tế-xã hội đối với các địa phương tuyển sinh.

Nghiên cứu lý thuyết trợ giúp quyết định, xây dựng ứng dụng dựa trên chủ đề là tên đề tài đã được đặt.

#### **4. Phương pháp nghiên cứu**

Để ứng dụng mục tiêu và nhiệm vụ của luận văn, tôi kết hợp hai phương pháp nghiên cứu, đó là:

##### ***Phương pháp nghiên cứu lý thuyết:***

Nghiên cứu tài liệu, ngôn ngữ và các công nghệ có liên quan.

Tổng hợp, thu thập các tài liệu về công tác tuyển sinh, đào tạo đại học hệ Từ xa của nhà trường.

##### ***Phương pháp nghiên cứu thực nghiệm:***

Phân tích yêu cầu thực tế của bài toán và áp dụng các thuật toán có liên quan để trợ giúp việc lập trình, xây dựng ứng dụng.

Thống kê, phân tích các số liệu thực tế trong công tác tuyển sinh.

Kiểm tra, thử nghiệm và đưa ra nhận xét, đánh giá kết quả đạt được.

#### **5. Kết quả đạt được**

Đề xuất được giải pháp kỹ thuật, xây dựng được chương trình thực hiện việc trợ giúp quyết định trong công tác tuyển sinh.

Hệ thống đơn giản, dễ sử dụng và có tính linh hoạt.

Đưa ra quyết định một cách nhanh chóng, tối ưu và có giá trị cho người sử dụng.

#### **6. Ý nghĩa khoa học và thực tiễn**

##### ***Về mặt khoa học***

Đề tài sẽ đưa ra một phương thức ứng dụng cây quyết định trong công tuyển sinh, tạo tiền đề cho những nghiên cứu ứng dụng trong công tác tuyển sinh sau này.

### ***Về mặt thực tiễn***

Đề tài sẽ ứng dụng các công cụ, ngôn ngữ lập trình để xây dựng hệ thống trợ giúp công tác tuyển sinh.

Sản phẩm sẽ là hệ thống phục vụ đắc lực, kịp thời và có độ chính xác cao cho phòng Đào tạo, ban giám hiệu nhà trường, cho lãnh đạo các đơn vị liên kết.

Triển khai hệ thống tại Viện Đại học Mở Hà Nội và các đơn vị liên kết nhằm phục vụ tốt yêu cầu trong công tác tuyển sinh.

## **7. Bố cục luận văn**

Sau phần mở đầu, giới thiệu..., nội dung chính của luận văn được chia thành 3 chương như sau:

Chương 1, trình bày cơ sở lý thuyết làm nền tảng để xây dựng ứng dụng, bao gồm: Hệ trợ giúp quyết định, các mô hình toán học thường dùng trong các bài toán ra quyết định. Cây quyết định và giải thuật C4.5 xây dựng cây quyết định.

Chương 2, tìm hiểu, giới thiệu và phân tích thực trạng công tác tuyển sinh tại nhà trường, nêu những vấn đề hạn chế và đề xuất giải pháp khắc phục, đó là giải pháp ứng dụng cây quyết định để giải quyết bài toán đặt ra.

Chương 3, trình bày chi tiết về mô hình kiến trúc tổng thể của hệ thống và phương pháp xây dựng ứng dụng. Tiến hành kịch bản thử nghiệm trên số liệu thực tế, sau đó đánh giá kết quả đạt được và khả năng triển khai ứng dụng trên toàn hệ thống.

Cuối cùng là phần đánh giá, kết luận và hướng phát triển của đề tài.

## **CHƯƠNG 1**

### **CƠ SỞ LÝ THUYẾT**

#### **1.1. HỆ TRỢ GIÚP QUYẾT ĐỊNH**

##### **1.1.1. Mở đầu**

##### **1.1.2. Khái niệm hệ trợ giúp quyết định**

##### **1.1.3. Quá trình ra quyết định**

##### **1.1.4. Các thành phần của hệ trợ giúp quyết định**

#### **1.2. RA QUYẾT ĐỊNH TRONG QUẢN LÝ**

##### **1.2.1. Mở đầu**

##### **1.2.2. Các phương pháp ra quyết định trong quản lý**

##### *1.2.2.1. Ra quyết định theo cấu trúc của vấn đề*

##### *1.2.2.2. Ra quyết định theo tính chất của vấn đề*

##### *1.2.2.3. Ra quyết định trong điều kiện rủi ro*

##### *1.2.2.4. Ra quyết định đa yếu tố*

##### **1.2.3. Các bước của quá trình ra quyết định**

##### **1.2.4. Bài toán ra quyết định**

#### **1.3. CÂY QUYẾT ĐỊNH**

##### **1.3.1. Giới thiệu chung**

##### **1.3.2. Phân lớp dữ liệu dựa trên các kiểu cây quyết định**

##### **1.3.3. Giải thuật cơ bản xây dựng cây quyết định**

##### **1.3.4. Chọn thuật toán C4.5 xây dựng cây quyết định**

#### **1.4. THUẬT TOÁN C4.5**

##### **1.4.1. Giới thiệu**

##### **1.4.2. Giải thuật C4.5 xây dựng cây quyết định từ trên xuống**

1.4.2.1. Thuật toán C4.5

**Function** *xay\_dung\_cay*(*T*)

```
{  
  1. <Tính toán tần suất các giá trị trong các lớp của T>;  
  2. If <Kiểm tra các mẫu, nếu thuộc cùng một lớp hoặc có  
    rất ít mẫu khác lớp> Then <Trả về 1 nút lá>  
    Else <Tạo một nút quyết định N>;  
  3. For <Với mỗi thuộc tính A> Do <Tính giá trị  
    Gain(A)>;  
  4. <Tại nút N, thực hiện việc kiểm tra để chọn ra thuộc  
    tính có giá trị Gain tốt nhất (lớn nhất). Gọi N.test  
    là thuộc tính có Gain lớn nhất>;  
  5. If <Nếu N.test là thuộc tính liên tục> Then <Tìm  
    ngưỡng cho phép tách của N.test>;  
  6. For <Với mỗi tập con T' được tách ra từ tập T> Do  
    (T' được tách ra theo quy tắc:  
      - Nếu N.test là thuộc tính liên tục tách theo  
        ngưỡng ở bước 5.  
      - Nếu N.test là thuộc tính phân loại rời rạc  
        tách theo các giá trị của thuộc tính này.  
    )  
  7. { If <Kiểm tra, nếu T' rỗng> Then  
      <Gán nút con này của nút N là nút lá>;  
      Else  
  8. <Gán nút con này là nút được trả về bằng cách gọi đệ  
    qui lại đối với hàm xay_dung_cay(T'), với tập  
    T'>;  
  }  
  9. <Tính toán các lỗi của nút N>;  
      <Trả về nút N>;  
}
```

1.4.2.2. Đánh giá độ phức tạp của thuật toán C4.5

### 1.4.2.3. Chọn thuộc tính phân loại tốt nhất

### 1.4.2.4. Entropy đo tính thuần nhất của tập ví dụ

Khái niệm entropy của một tập S được định nghĩa trong lý thuyết thông tin là số lượng mong đợi các bit cần thiết để mã hóa thông tin về lớp của một thành viên rút ra một cách ngẫu nhiên từ tập S. Trong trường hợp tối ưu, mã có độ dài ngắn nhất. Theo lý thuyết thông tin, mã có độ dài tối ưu là mã gán  $-\log_2 p$  bits cho thông điệp có xác suất là p [7].

Trong trường hợp S là tập ví dụ, thì thành viên của S là một ví dụ, mỗi ví dụ thuộc một lớp hay có một giá trị phân loại.

Entropy có giá trị nằm trong khoảng [0..1].

Entropy(S) = 0: tập S chỉ toàn ví dụ thuộc cùng một loại, hay S là thuần nhất.

Entropy(S) = 1: tập ví dụ S có các ví dụ thuộc các loại khác nhau với độ pha trộn là cao nhất.

$0 < \text{Entropy}(S) < 1$ : tập ví dụ S có số lượng ví dụ thuộc các loại khác nhau là không bằng nhau.

Để đơn giản ta xét trường hợp các ví dụ của S chỉ thuộc loại âm (-) hoặc dương (+).

Cho trước:

Tập S là tập dữ liệu huấn luyện, trong đó thuộc tính phân loại có hai giá trị, giả sử là âm (-) và dương (+).

Trong đó:

$p_+$  là phần các ví dụ dương trong tập S.

$p_-$  là phần các ví dụ âm trong tập S.

Khi đó, entropy đo độ pha trộn của tập S theo công thức sau:

$$\text{Entropy}(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

Một cách tổng quát hơn, nếu các ví dụ của tập S thuộc nhiều hơn hai loại, giả sử là có c giá trị phân loại thì công thức entropy tổng quát là:

$$\text{Entropy}(S) \equiv \sum_{i=1}^c -p_i \log_2 p_i$$

*1.4.2.5. Lượng thông tin thu được đo mức độ giảm Entropy mong đợi*

Entropy là một số đo đo độ pha trộn của một tập ví dụ, bây giờ chúng ta sẽ định nghĩa một phép đo hiệu suất phân loại các ví dụ của một thuộc tính. Phép đo này gọi là lượng thông tin thu được (hay độ lợi thông tin), nó đơn giản là lượng giảm entropy mong đợi gây ra bởi việc phân chia các ví dụ theo thuộc tính này.

Một cách chính xác hơn,  $\text{Gain}(S, A)$  của thuộc tính  $A$ , trên tập  $S$ , được định nghĩa như sau:

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{Value}(A)} \frac{|S_v|}{|S|} \text{Entropy}(S_v)$$

Giá trị  $\text{Value}(A)$  là tập các giá trị có thể cho thuộc tính  $A$ , và  $S_v$  là tập con của  $S$  mà  $A$  nhận giá trị  $v$ .

*1.4.2.6. Tỷ suất lợi ích Gain Ratio*

Khái niệm độ lợi thông tin  $\text{Gain}$  có xu hướng ưu tiên các thuộc tính có số lượng lớn các giá trị. Nếu thuộc tính  $D$  có giá trị riêng biệt cho mỗi bản ghi (thuộc tính *Ngày* ở bảng dữ liệu trên), thì  $\text{Entropy}(S, D) = 0$ , như vậy  $\text{Gain}(S, D)$  sẽ đạt giá trị cực đại. Rõ ràng, một phân vùng như vậy thì việc phân loại là vô ích.

Thuật toán C4.5, một cải tiến của ID3, mở rộng cách tính Information Gain thành Gain Ratio để cố gắng khắc phục sự thiên lệch.

Gain Ratio được xác định bởi công thức sau:

$$\text{GainRatio}(S, A) = \frac{\text{Gain}(S, A)}{\text{SplitInformation}(S, A)}$$

Với  $\text{SplitInformation}(S, A)$  chính là thông tin do phân tách của  $A$  trên cơ sở giá trị của thuộc tính phân loại  $S$ . Công thức tính như sau:

$$\text{SplitInformation}(S, A) = - \sum_{i \in \text{Value}(S)} \frac{|A_i|}{|A|} \log_2 \frac{|A_i|}{|A|}$$

Trong đó:

*Value(S)* là tập các giá trị của thuộc tính *S*

*A<sub>i</sub>* là tập con của tập *A* ứng với thuộc tính *S* = giá trị là *v<sub>i</sub>*

**1.4.3. Phương pháp đánh giá mức độ hiệu quả**

**1.4.4. Chuyển cây về dạng luật**

**1.4.5. Ứng dụng tập luật**

## CHƯƠNG 2

### PHÂN TÍCH THỰC TRẠNG CÔNG TÁC TUYỂN SINH TẠI VIỆN ĐẠI HỌC MỞ HÀ NỘI

**2.1. GIỚI THIỆU VỀ VIỆN ĐẠI HỌC MỞ HÀ NỘI**

**2.2. CÔNG TÁC TUYỂN SINH ĐẠI HỌC HỆ TỪ XA**

**2.2.1. Quy trình tuyển sinh**

**2.2.2. Mô hình tuyển sinh**

**2.2.3. Mục tiêu tuyển sinh**

**2.2.4. Thực trạng tuyển sinh**

**2.2.5. Vấn đề trợ giúp quyết định**

**2.2.6. Giải pháp xây dựng hệ thống trợ giúp quyết định.**

**2.3. ỨNG DỤNG CÂY QUYẾT ĐỊNH**

**2.3.1. Lượng hóa dữ liệu để đưa vào thuật toán C4.5**

**2.3.2. Phân tích dữ liệu**

Qua phân tích ở phần thực trạng công tác tuyển sinh, ta thấy có nhiều yếu tố ảnh hưởng đến công tác tuyển sinh. Tuy nhiên, chúng ta chú trọng phân tích những yếu tố chính sau:

**Chuyên ngành đào tạo (CNDT):** Các ngành nhà trường có tuyển sinh

**Nhu cầu xã hội (NCXH):** Số liệu điều tra nhu cầu của xã hội đối với các ngành dự kiến tuyển sinh.

**Tỷ lệ bỏ học (TLBH):** Số liệu thống kê tỷ lệ bỏ học của học viên qua các năm đào tạo.

**Cạnh tranh (CT):** Có hay không có sự cạnh tranh trong công tác tuyển sinh của các trường khác.

Các yếu tố trên chính là tập thuộc tính, dựa vào tập thuộc tính này để dự đoán giá trị cho thuộc tính đích **Quyết định**.

Để thuận tiện trong việc huấn luyện cây quyết định, đối với các thuộc tính có các giá trị so sánh, ta tiến hành rời rạc hóa và ký hiệu hóa các giá trị của chúng.

*Nhu cầu xã hội:* Nếu nhu cầu xã hội  $< 80$  thì ký hiệu là 80-; nếu  $80 \leq$  nhu cầu xã hội  $\leq 100$ , ký hiệu 80..100; nếu  $100 <$  nhu cầu xã hội  $\leq 120$ , ký hiệu 101..120; và  $> 120$  ký hiệu là 120+.

*Tỷ lệ bỏ học:* 32- (dưới 32%); 32..42 (từ 32% đến 42%); 42+ (trên 42%).

*Cạnh tranh:* Nếu trên cùng một địa bàn và tuyển sinh tại cùng một đơn vị liên kết mà có từ hai trường trở lên cùng tuyển sinh một ngành thì ta đánh giá mức độ cạnh tranh là cao (CAO), ngược lại là thấp (THAP).

Từ các số liệu ở phụ lục 12, phụ lục 15, ta lập bảng giả định về công tác tuyển sinh của các ngành như sau:

**Bảng 2.3. Bảng dữ liệu rút gọn**

| TT | NCDT | NCXH    | TLBH (%) | CT   | QD  |
|----|------|---------|----------|------|-----|
| 1  | QTKD | 80-     | 32-      | THAP | NO  |
| 2  | KT   | 120+    | 42+      | THAP | YES |
| 3  | LKT  | 120+    | 42+      | THAP | YES |
| 4  | QTKD | 80..100 | 32-      | CAO  | YES |
| 5  | KT   | 120+    | 32-      | CAO  | YES |
| 6  | LKT  | 120+    | 32-      | CAO  | YES |

| TT | NCDT | NCXH     | TLBH (%) | CT   | QD  |
|----|------|----------|----------|------|-----|
| 7  | QTKD | 80..100  | 32-      | THAP | YES |
| 8  | KT   | 120+     | 32..42   | CAO  | YES |
| 9  | LKT  | 120+     | 32..42   | CAO  | YES |
| 10 | QTKD | 80..100  | 32..42   | CAO  | NO  |
| 11 | KT   | 120+     | 42+      | CAO  | Y/N |
| 12 | LKT  | 120+     | 42+      | CAO  | Y/N |
| 13 | QTKD | 80..100  | 32..42   | THAP | Y/N |
| 14 | KT   | 101..120 | 32-      | CAO  | Y/N |
| 15 | LKT  | 101..120 | 32-      | CAO  | Y/N |
| 16 | QTKD | 80..100  | 42+      | CAO  | NO  |
| 17 | KT   | 101..120 | 32..42   | CAO  | Y/N |
| 18 | LKT  | 101..120 | 32..42   | CAO  | Y/N |
| 19 | QTKD | 80..100  | 42+      | THAP | NO  |
| 20 | KT   | 101..120 | 42+      | CAO  | NO  |
| 21 | LKT  | 101..120 | 42+      | CAO  | NO  |
| 22 | QTKD | 101..120 | 32-      | CAO  | Y/N |
| 23 | KT   | 101..120 | 32-      | THAP | YES |
| 24 | LKT  | 101..120 | 32-      | THAP | YES |
| 25 | QTKD | 101..120 | 32..42   | CAO  | Y/N |
| 26 | KT   | 101..120 | 32..42   | THAP | YES |
| 27 | LKT  | 101..120 | 32..42   | THAP | YES |
| 28 | QTKD | 101..120 | 42+      | CAO  | NO  |
| 29 | KT   | 101..120 | 42+      | THAP | Y/N |
| 30 | LKT  | 101..120 | 42+      | THAP | Y/N |
| 31 | QTKD | 101..120 | 32-      | THAP | YES |
| 32 | KT   | 80..100  | 32-      | CAO  | YES |
| 33 | LKT  | 80..100  | 32-      | CAO  | YES |
| 34 | QTKD | 101..120 | 32..42   | THAP | YES |
| 35 | KT   | 80..100  | 32..42   | CAO  | NO  |
| 36 | LKT  | 80..100  | 32..42   | CAO  | NO  |
| 37 | QTKD | 101..120 | 42+      | THAP | Y/N |

| TT | NCDT | NCXH    | TLBH (%) | CT   | QD  |
|----|------|---------|----------|------|-----|
| 38 | KT   | 80..100 | 32..42   | THAP | Y/N |
| 39 | LKT  | 80..100 | 32..42   | THAP | Y/N |
| 40 | QTKD | 120+    | 32-      | CAO  | YES |
| 41 | KT   | 80..100 | 42+      | THAP | NO  |
| 42 | LKT  | 80..100 | 42+      | THAP | NO  |
| 43 | QTKD | 120+    | 32..42   | CAO  | YES |
| 44 | KT   | 80-     | 32-      | THAP | NO  |
| 45 | LKT  | 80-     | 32-      | THAP | NO  |
| 46 | QTKD | 120+    | 42+      | CAO  | Y/N |
| 47 | KT   | 80-     | 32..42   | CAO  | NO  |
| 48 | LKT  | 80-     | 32..42   | CAO  | NO  |
| 49 | QTKD | 120+    | 42+      | THAP | YES |
| 50 | LKT  | 80..100 | 42+      | THAP | NO  |

Từ ví dụ trên, áp dụng thuật thuật toán C4.5 cho tập dữ liệu huấn luyện để xây dựng cây quyết định cho bài toán tuyển sinh.

### 2.3.3. Triển khai giải thuật C4.5 xây dựng cây quyết định

Để xây dựng cây quyết định ta phải xác định nút gốc để phân tách cây. Thuộc tính có độ lợi thông tin lớn nhất sẽ được chọn làm nút gốc.

Gọi S là tập thuộc tính đích. Có tất cả 50 ví dụ, trong đó YES xuất hiện trong tập thuộc tính đích 19 lần, NO 16 lần và Y/N 15 lần.

Áp dụng công thức tính Entropy, ta có:

$$\text{Entropy}(S) = -(19/50)\log_2(19/50) - (16/50)\log_2(16/50) - (15/50)\log_2(15/50) = 1.58$$

Đối với thuộc tính “*Chuyên ngành đào tạo*”, ta tính Entropy của các tập con S được chia bởi các giá trị của thuộc tính “*Chuyên ngành đào tạo*” như bảng sau:

**Bảng 2.4. Bảng Entropy(S) phân theo CNDT**

| CNDT | Số lượng | QD  |    |     | Entropy |
|------|----------|-----|----|-----|---------|
|      |          | YES | NO | Y/N |         |
| QTKD | 17       | 7   | 5  | 5   | 1.57    |
| KT   | 16       | 6   | 5  | 5   | 1.58    |
| LKT  | 17       | 6   | 6  | 5   | 1.58    |

Entropy của S đối với thuộc tính “*Chuyên ngành đào tạo*” là:

$$\text{Entropy}(S, \text{CNDT}) = (17/50) \times 1.57 + (16/50) \times 1.58 + (17/50) \times 1.58 = 1.58$$

Độ lợi thông tin tương ứng là:

$$\text{Gain}(S, \text{CNDT}) = \text{Entropy}(S) - \text{Entropy}(S, \text{CNDT}) = 1.58 - 1.58 = 0.00$$

Tỷ suất lợi ích Gain Ratio:

$$\text{SplitInfor}(S, \text{CNDT}) = - (17/50) \times \log_2(17/50) - (16/50) \times \log_2(16/50) - (17/50) \times \log_2(17/50) = 1.58$$

$$\text{GainRatio}(S, \text{CNDT}) = \text{Gain}(S, \text{CNDT}) / \text{SplitInfor}(S, \text{CNDT}) = 0.00 / 1.58 = 0.00$$

Một cách tương tự, ta tính độ lợi thông tin Gain và tỉ suất lợi ích Gain Ratio của các thuộc tính còn lại

**Bảng 2.5. Độ lợi thông tin của thuộc tính “*Nhu cầu xã hội*”**

| NHCH     | Số lượng | QD  |    |     | Entropy |
|----------|----------|-----|----|-----|---------|
|          |          | YES | NO | Y/N |         |
| 80-      | 5        | 0   | 5  | 0   | 0.00    |
| 80..100  | 15       | 4   | 8  | 3   | 1.46    |
| 101..120 | 18       | 6   | 3  | 9   | 1.46    |
| 120+     | 12       | 9   | 0  | 3   | 0.00    |

**Bảng 2.6. Độ lợi thông tin của thuộc tính “Tỷ lệ bỏ học”**

| TLBH   | Số lượng | QD  |    |     | Entropy |
|--------|----------|-----|----|-----|---------|
|        |          | YES | NO | Y/N |         |
| 32-    | 16       | 10  | 3  | 3   | 1.33    |
| 32..42 | 17       | 6   | 5  | 6   | 1.58    |
| 42+    | 17       | 3   | 8  | 6   | 1.48    |

**Bảng 2.7. Độ lợi thông tin của thuộc tính “Cạnh tranh”**

| TLBH | Số lượng | QD  |    |     | Entropy |
|------|----------|-----|----|-----|---------|
|      |          | YES | NO | Y/N |         |
| CAO  | 27       | 9   | 9  | 9   | 1.58    |
| THAP | 23       | 10  | 7  | 6   | 1.55    |

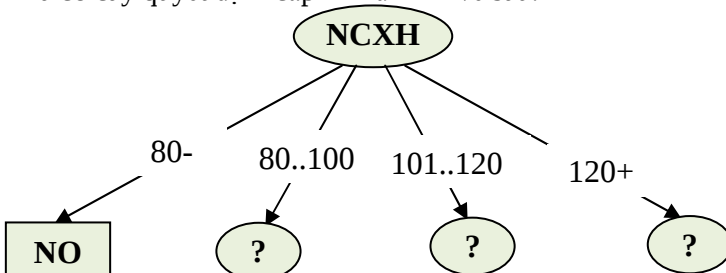
Tiếp theo, ta so sánh kết quả tính GainRatio của các thuộc tính

**Bảng 2.8. Bảng so sánh kết quả tính GainRatio của các thuộc tính**

| Thuộc tính | Gain | SplitInfor | GainRatio   |
|------------|------|------------|-------------|
| CNDT       | 0.00 | 1.58       | 0.00        |
| NCXH       | 0.62 | 1.88       | <b>0.33</b> |
| TLBH       | 0.11 | 1.58       | 0.07        |
| CT         | 0.01 | 1.00       | 0.01        |

Ta nhận thấy  $\text{GainRatio}(S, \text{NCXH}) = 0.33$  đạt giá trị lớn nhất, do đó thuộc tính nhu cầu xã hội có khả năng phân loại tốt nhất. Chính vì vậy ta sẽ chọn thuộc tính này làm nút gốc phân tách cây.

Ta có cây quyết định cấp 1 như hình vẽ sau:



Hình 2.3. Cây quyết định cấp 1

Ứng với giá trị NCXH = 80- có Entropy = 0. Do đó, tại nhánh này sẽ là nút có phân lớp là NO.

Đối với các nhánh còn lại (NCXH = 80..100; NCXH = 101..120; NCXH = 120+), ta tiếp tục vận dụng thuật toán để tìm thuộc tính làm nút của cây.

### **Xét nhánh NCXH =80..100**

**Bảng 2.9. Bảng dữ liệu trường hợp NCXH = 80..100**

| TT | NCDT | NCXH    | TLBH   | CT   | QD  |
|----|------|---------|--------|------|-----|
| 1  | QTKD | 80..100 | 32-    | CAO  | YES |
| 2  | QTKD | 80..100 | 32-    | THAP | YES |
| 3  | QTKD | 80..100 | 32..42 | CAO  | NO  |
| 4  | QTKD | 80..100 | 32..42 | THAP | Y/N |
| 5  | QTKD | 80..100 | 42+    | CAO  | NO  |
| 6  | QTKD | 80..100 | 42+    | THAP | NO  |
| 7  | KT   | 80..100 | 32-    | CAO  | YES |
| 8  | LKT  | 80..100 | 32-    | CAO  | YES |
| 9  | KT   | 80..100 | 32..42 | CAO  | NO  |
| 10 | LKT  | 80..100 | 32..42 | CAO  | NO  |
| 11 | KT   | 80..100 | 32..42 | THAP | Y/N |
| 12 | LKT  | 80..100 | 32..42 | THAP | Y/N |
| 13 | KT   | 80..100 | 42+    | THAP | NO  |
| 14 | LKT  | 80..100 | 42+    | THAP | NO  |
| 15 | LKT  | 80..100 | 42+    | THAP | NO  |

Gọi S1 là tập thuộc tính đích. Áp dụng công thức tính Entropy ta có:

$$\text{Entropy}(S1) = -(3/15) \times \log_2(3/15) - (9/15) \times \log_2(9/15) - (3/15) \times \log_2(3/15) = \mathbf{1.37}$$

Tiếp theo ta lần lượt tính Gain của các thuộc tính như đã tính ở phần trên, kết quả như sau:

**Bảng 2.10. Độ lợi thông tin của thuộc tính “CNDT”**

| CNDT | Số lượng | QD  |    |     | Entropy (i) | Entropy (CNDT) | Gain (S1,CNDT) |
|------|----------|-----|----|-----|-------------|----------------|----------------|
|      |          | YES | NO | Y/N |             |                |                |
| QTKD | 6        | 2   | 3  | 1   | 1.46        | 1.44           | 0.01           |
| KT   | 4        | 1   | 2  | 1   | 1.50        |                |                |
| LKT  | 5        | 1   | 3  | 1   | 1.37        |                |                |

**Bảng 2.11. Độ lợi thông tin của thuộc tính “TLBH”**

| TLBH   | Số lượng | QD  |    |     | Entropy (i) | Entropy (TLBH) | Gain (S1,TLBH) |
|--------|----------|-----|----|-----|-------------|----------------|----------------|
|        |          | YES | NO | Y/N |             |                |                |
| 32-    | 4        | 4   | 0  | 0   | 0.00        | 0.00           | 0.93           |
| 32..42 | 6        | 0   | 3  | 3   | 0.00        |                |                |
| 42+    | 5        | 0   | 5  | 0   | 0.00        |                |                |

**Bảng 2.12. Độ lợi thông tin của thuộc tính “CT”**

| CT   | Số lượng | QD  |    |     | Entropy (i) | Entropy (CT) | Gain (S1,CT) |
|------|----------|-----|----|-----|-------------|--------------|--------------|
|      |          | YES | NO | Y/N |             |              |              |
| CAO  | 7        | 3   | 4  | 0   | 0.00        | 0.75         | 0.71         |
| THAP | 8        | 1   | 4  | 3   | 1.41        |              |              |

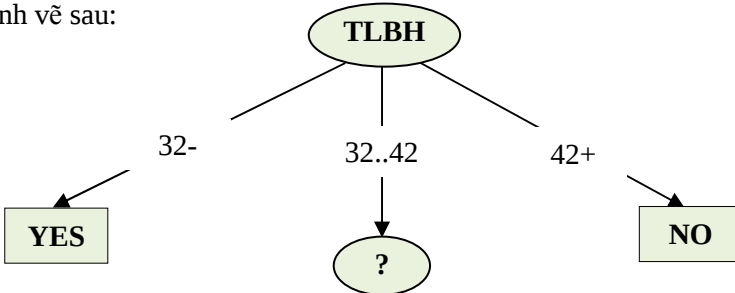
Tiếp theo ta tính SplitInfor và GainRatio của các thuộc tính, kết quả như sau:

**Bảng 2.13. Kết quả tính SplitInfor và GainRatio**

| Thuộc tính | Gain | SplitInfor | GainRatio   |
|------------|------|------------|-------------|
| CNDT       | 0.02 | 1.57       | 0.01        |
| TLBH       | 1.46 | 1.57       | <b>0.93</b> |
| CT         | 0.71 | 1.00       | 0.71        |

Qua bảng so sánh kết quả trên, ta thấy GainRatio(S1,TLBH) = 0.93 đạt giá trị lớn nhất, do đó tại nhánh này ta chọn thuộc tính TLBH để làm nút phân tách cây.

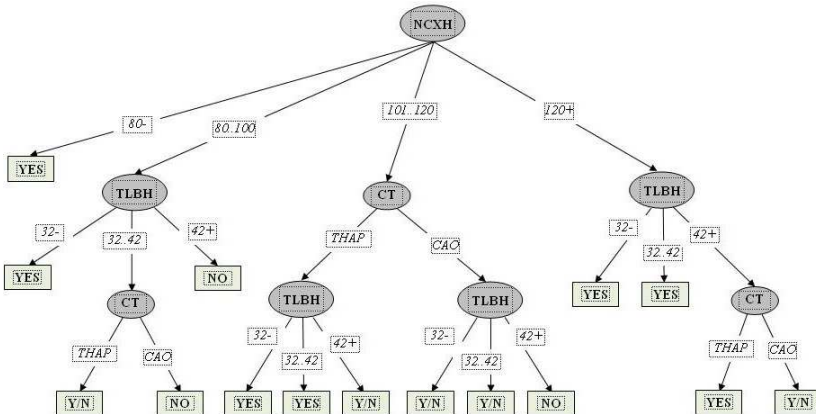
Ta có cây quyết định cấp ứng với nhánh NCXH = 80..100 như hình vẽ sau:



**Hình 2.4. Cây quyết định ứng với nhánh NCXH = 80..100**

Qua bảng 2.10 ta thấy, ứng với tất các giá trị TLBH = 32- và TLBH = 42+ có entropy = 0, do đó, tại các nhánh này sẽ là nút lá với phân lớp là YES (TLBH = 32) và NO (TLBH = 42+).

Tương tự như cách tính ở trên ta xét cho các nhánh còn lại. Cuối cùng ta xây dựng được cây quyết định hoàn chỉnh như hình 2.5.



**Hình 2.4. Cây quyết định hoàn chỉnh**

### 2.3.4. Tạo luật từ cây quyết định

## **CHƯƠNG 3**

### **XÂY DỰNG ỨNG DỤNG THỬ NGHIỆM**

#### **3.1. XÂY DỰNG ỨNG DỤNG TRỢ GIÚP QUYẾT ĐỊNH**

##### **3.1.1. Chức năng của hệ thống**

##### **3.1.2. Kiến trúc tổng thể của hệ thống**

##### **3.1.3. Đặc tả chi tiết các thành phần**

##### **3.1.4. Cài đặt chương trình**

#### **3.2. THỬ NGHIỆM ỨNG DỤNG**

##### **3.2.1. Huấn luyện cây quyết định**

##### **3.2.2. Áp dụng cây quyết định**

#### **3.3. ĐÁNH GIÁ KẾT QUẢ THỰC HIỆN**

##### **3.3.1. Về ứng dụng hệ thống**

Sau khi đưa hệ thống trợ giúp quyết định trong công tác tuyển sinh vào thử nghiệm với số dữ liệu của 56 lớp đã tuyển sinh và đào tạo (tại 3 địa phương Đà Nẵng, Quảng Nam và TP. Huế), chúng tôi rút ra được những nhận xét, đánh giá một cách tương đối về hệ thống như sau:

Hệ thống được triển khai cài đặt, thử nghiệm dễ dàng, dễ sử dụng.

Hệ thống hoạt động ổn định, hiển thị kết quả nhanh chóng và rất thuận tiện.

Với số lượng dữ liệu thử nghiệm, hệ thống đã xử lý và cho ra được những quyết định có độ chính xác cao, phù hợp với tính toán thực tế.

##### **3.3.2. Về hiệu quả trong công tác tuyển sinh.**

Với số liệu thống kê của 55 lớp khi đưa vào hệ thống để phân tích và đưa ra kết quả, hệ thống đã phát hiện ra 38 lớp tuyển sinh và đào tạo đạt hiệu quả (chiếm 69.09%) và 17 lớp (lỗi) tuyển sinh, đưa vào đào tạo không đạt hiệu quả về mặt kinh tế (chiếm 30.91%).

**Bảng 3.1. Bảng thống kê nguyên nhân tuyển sinh không hiệu quả**

| TT               | Nguyên nhân                                    | Số lượng  | Tỷ lệ    |
|------------------|------------------------------------------------|-----------|----------|
| 1                | Có sự cạnh tranh                               | 1         | 5.88%    |
| 2                | Nhu cầu xã hội thấp hơn mức học viên tối thiểu | 3         | 17.65%   |
| 3                | Tỷ lệ học viên bỏ học cao                      | 13        | 76.47%   |
| <b>TỔNG CỘNG</b> |                                                | <b>17</b> | <b>-</b> |

Có thể thấy rằng, nguyên nhân tuyển sinh không đạt hiệu quả chủ yếu do tỷ lệ học viên bỏ học quá cao (chiếm 76.47%). Ngày nay, công tác tuyển sinh ngày càng gặp nhiều khó khăn vì lý do người học có nhiều sự lựa chọn: trường đào tạo, cấp độ đào tạo, ngành nghề đào tạo,... Tuyển sinh được học viên đã khó, giữ được sĩ số học viên ổn định trong suốt quá trình đào tạo lại càng khó hơn.

Vì vậy, ngoài việc triển khai thật tốt và có hiệu quả trong công tác tuyển sinh, cần phải có biện pháp nhằm giải quyết bài toán giảm tỷ lệ bỏ học của học viên xuống mức tối thiểu có thể chấp nhận được, nhằm nâng cao hiệu quả của công tác tuyển sinh của nhà trường.

Trong quá trình thực hiện luận văn, chúng tôi đã phân tích số liệu, các yếu tố liên quan đến công tác tuyển sinh như đã nêu ở phần “*thực trạng tuyển sinh*”. Qua kinh nghiệm công tác, chúng tôi có thể đánh giá và chỉ ra nguyên nhân những lần tuyển sinh không đạt hiệu quả của nhà trường trong thời gian qua như sau:

Đầu tiên, nhà trường chưa chú trọng đến công tác điều tra, thống kê về nhu cầu xã hội và sự cạnh tranh của các trường đối với các ngành, địa bàn tuyển sinh.

Tiếp đến, công tác tuyển sinh tại địa phương hoàn toàn giao cho các đơn vị liên kết tự làm mà không có sự phối hợp, giám sát từ các phòng ban chức năng của nhà trường.

Hơn nữa, công tác tư vấn tuyển sinh chưa được chú trọng. Công tác này do các cán bộ tại địa phương liên kết đảm trách mà không có sự tập huấn bài bản.

Ngoài ra, nhà trường chưa có sự tính toán cụ thể về các khoản chi phí phải trả đối với các lớp mở tại các địa phương để đưa ra số lượng học viên tối thiểu có thể mở lớp. Do đó, đã có những quyết định mở lớp chưa phù hợp, số lượng học viên đăng ký thấp hơn số học viên tối thiểu nhưng vẫn đưa vào đào tạo (*phụ lục 11*).

Cuối cùng, nhà trường chưa có biện pháp hữu hiệu nhằm hạn chế tỷ lệ bỏ học của học viên. Ngay cả đối với những lớp đào tạo đạt hiệu quả về mặt kinh tế nhưng tỷ lệ bỏ học vẫn cao, dẫn đến ảnh hưởng đến kết quả chung của công tác tuyển sinh.

Tỷ lệ bỏ học của học viên cao, có thể do những lý do sau:

Thời gian đào tạo một khóa học của nhà trường kéo dài 5 năm, trong khi các trường khác chỉ có 4 năm. (*phụ lục 13*)

Học từ xa, học viên lấy việc tự học thông qua giáo trình là chủ yếu, tuy nhiên việc tổ chức, ra đề thi để đánh giá chất lượng vẫn chưa sát với giáo trình, bài giảng... Do đó, kết quả thi của học viên không đạt yêu cầu chiếm tỷ lệ cao. Có những học viên thi đi, thi lại đến bảy, tám lần vẫn chưa đạt. (*phụ lục 21*)

Việc xét miễn môn học để tạo điều kiện cho học viên rút ngắn thời gian đào tạo còn nhiều bất cập, chưa hợp tình, hợp lý.

Công tác quản lý, theo dõi lớp học chưa được quan tâm đúng mức, dẫn đến việc nhắc nhở, động viên và giải quyết các tồn đọng, thắc mắc của học viên chưa kịp thời, thỏa đáng.

Trên đây là những kết quả đạt được khi thực hiện luận văn, kết quả này giúp chúng tôi có thể đánh giá một cách bao quát về quá trình tuyển sinh cũng như đào tạo của Viện Đại học Mở Hà Nội trong thời gian qua. Chúng tôi sẽ nêu những vấn đề còn bất cập ở trên với ban giám hiệu nhà trường để có những biện pháp khắc phục. Đồng thời đưa

hệ thống trợ giúp quyết định trong công tác tuyển sinh vào sử dụng nhằm giúp cho phòng đào tạo, ban giám hiệu, lãnh đạo các đơn vị liên kết có một công cụ hỗ trợ hữu ích, nhanh chóng, hiệu quả trong việc quyết định công tác tuyển sinh sau này.

## KẾT LUẬN

### 1. Kết quả đạt được

#### *Về mặt khoa học*

Luận văn đã tiến hành phân tích, tìm hiểu được quy trình tuyển sinh của Viện Đại học Mở Hà Nội. Phát hiện ra những vấn đề còn hạn chế đề xuất với ban giám hiệu nhằm có những phương án khắc phục để nâng cao hiệu quả trong công tác tuyển sinh.

Nắm được các phương pháp và các mô hình toán học, áp dụng để giải quyết yêu cầu luận văn đã đặt ra.

Nghiên cứu và vận dụng giải thuật C4.5 để xây dựng mô hình dự đoán bằng cây quyết định.

#### *Về mặt thực tiễn*

Luận văn đã nêu được giải pháp kỹ thuật để xây dựng hệ thống trợ giúp quyết định trong công tác tuyển sinh.

Xây dựng được ứng dụng có khả năng phân tích tốt các dữ liệu về công tác tuyển sinh của nhà trường trong những năm qua.

Việc kết hợp lý thuyết về hệ trợ giúp quyết định và phân tích dữ liệu bằng cây quyết định là rất cần thiết, nó giúp giảm thiểu đáng kể thời gian tính toán trong công tác tuyển sinh.

Hệ thống có thể giúp cho phòng đào tạo, ban giám hiệu nhà trường và lãnh đạo các đơn vị liên kết ra quyết định một cách kịp thời, khoa học, tránh được các tình huống quyết định theo cảm tính nhằm hạn chế các trường hợp tuyển sinh không hiệu quả dẫn đến thiệt hại về kinh tế, lãng phí thời gian và tiền bạc của người học.

Có thể nói, đây là một công cụ hữu ích nhằm cung cấp cho nhà trường có thêm một giải pháp hỗ trợ về công tác tuyển sinh sau này.

## **2. Hạn chế**

Hệ thống hiện tại chỉ xử lý dữ liệu được lưu trữ bằng các tập tin Excel, chưa kết nối và truy xuất dữ liệu trực tiếp đến hệ quản trị cơ sở dữ liệu SQL Server của nhà trường, của các đơn vị liên kết. Do đó cần một khoảng thời gian để chuyển đổi dữ liệu từ SQL Server sang Excel.

## **3. Hướng phát triển**

Cần thử nghiệm hệ thống với khối lượng dữ liệu lớn để đánh giá lại độ tin cậy của cây quyết định công tác tuyển sinh.

Phối hợp với các phòng ban chức năng, các đơn vị liên kết để tính toán, xây dựng hoàn thiện hơn tập dữ liệu huấn luyện nhằm đưa hệ thống vào sử dụng có hiệu quả cho tất cả các địa phương liên kết.

Nghiên cứu vận dụng thuật toán C5.0, là một cải tiến của C4.5, để giảm thiểu tỉ lệ lỗi, nâng cao hiệu suất và tăng cường độ tối ưu cho ứng dụng.

Tiếp tục phát triển, nâng cấp hệ thống, hỗ trợ kết nối, truy xuất và xử lý dữ liệu trực tiếp với hệ quản trị CSDL SQL Server đồng thời mở rộng sang các hệ quản trị CSDL khác.