**ĐỀ TÀI: DÙNG CÂY QUYẾT ĐỊNH DỰ BÁO THỜI TIẾT**

# 

Contents

[PHẦN MỞ ĐẦU 3](#_Toc60335689)

[CHƯƠNG I. TỔNG QUAN VỀ DỰ BÁO THỜI TIẾT VÀ LÝ THUYẾT CÂY QUYẾT ĐỊNH 4](#_Toc60335690)

[1. GIỚI THIỆU 4](#_Toc60335691)

[2. DÙNG CÂY QUYẾT ĐỊNH DỰ BÁO THỜI TIẾT. 5](#_Toc60335692)

[2.1. Tên Đề Tài 5](#_Toc60335693)

[2.2. Mục Đích Đề Tài 5](#_Toc60335694)

[2.3. Phát biểu bài toán 5](#_Toc60335695)

[CHƯƠNG II. TÌM HIỂU VỀ CÂY QUYẾT ĐỊNH 6](#_Toc60335696)

[1. Các định nghĩa 6](#_Toc60335697)

[2. Các kiểu cây quyết định 7](#_Toc60335698)

[a. Cây hồi quy (*Regression tree*) 7](#_Toc60335699)

[b. Cây phân loại (*Classification tree*) 7](#_Toc60335700)

[3. Uư/ Nhược điểm của thuật toán cây quyết định 7](#_Toc60335701)

[CHƯƠNG III: PHẦN MỀM WEKA, NGUỒN DỮ LIỆU VÀ GIAO DIỆN DỰ BÁO THỜI TIẾT 9](#_Toc60335713)

# PHẦN MỞ ĐẦU

Con người đã cố gắng dự báo thời tiết trong nhiều thiên niên kỷ. Năm 650 BC, người [Babylon](https://vi.wikipedia.org/wiki/Babylon) tiên đoán thời tiết dựa vào hình dạng của các đám mây cũng như dựa vào [chiêm tinh học](https://vi.wikipedia.org/wiki/Chi%C3%AAm_tinh_h%E1%BB%8Dc). Khoảng 340 BC, [Aristotle](https://vi.wikipedia.org/wiki/Aristoteles) miêu tả một phần thời tiết trong cuốn [Meteorologica](https://vi.wikipedia.org/w/index.php?title=Meteorology_(Aristotle)&action=edit&redlink=1). Sau đó, [Theophrastus](https://vi.wikipedia.org/wiki/Theophrastos) đã biên soạn một cuốn sách về dự báo thời tiết, gọi là Sách của dấu hiệu. Tục ngữ về dự báo thời tiết của [người Trung Hoa](https://vi.wikipedia.org/wiki/Ng%C6%B0%E1%BB%9Di_Hoa) có lẽ có từ khoảng 300 BC,[[4]](https://vi.wikipedia.org/wiki/D%E1%BB%B1_b%C3%A1o_th%E1%BB%9Di_ti%E1%BA%BFt#cite_note-4) mà cũng là vào cùng khoảng thời gian những nhà thiên văn cổ đại Ấn Độ phát triển các phương pháp dự báo thời tiết. Năm 904 AD, cuốn Nông nghiệp Nabatean của Ibn Wahshiyya nói về dự báo thời tiết liên quan tới sự thay đổi [khí quyển](https://vi.wikipedia.org/wiki/Kh%C3%AD_quy%E1%BB%83n) và dấu hiệu dự báo liên quan tới thay đổi từ các [hành tinh](https://vi.wikipedia.org/wiki/H%C3%A0nh_tinh) và các [ngôi sao](https://vi.wikipedia.org/wiki/Sao); dấu hiệu của [mưa](https://vi.wikipedia.org/wiki/M%C6%B0a) dựa trên quan sát các [pha Mặt Trăng](https://vi.wikipedia.org/wiki/Pha_M%E1%BA%B7t_Tr%C4%83ng); và dự báo thời tiết dựa trên sự chuyển động của [gió](https://vi.wikipedia.org/wiki/Gi%C3%B3).

Phương pháp dự báo thời tiết của người cổ đại thường dựa trên quan sát một phần của những sự kiện, và do vậy chỉ nhận ra được một phần của thời tiết sẽ xảy ra như thế nào. Ví dụ, nếu quan sát thấy khi [Mặt Trời lặn](https://vi.wikipedia.org/wiki/M%E1%BA%B7t_Tr%E1%BB%9Di_l%E1%BA%B7n) có màu đỏ, thì ngày hôm sau thời tiết khá thuận lợi. Kinh nghiệm như thế được tích lũy qua nhiều thế hệ tựu lại thành [tục ngữ về thời tiết](https://vi.wikipedia.org/w/index.php?title=T%E1%BB%A5c_ng%E1%BB%AF_v%E1%BB%81_th%E1%BB%9Di_ti%E1%BA%BFt&action=edit&redlink=1). Tuy nhiên không phải mọi tiên đoán này đều tin cậy được, và nhiều câu trong dân gian về thời tiết đã không còn thỏa mãn đối với các thử nghiệm thống kê nghiêm ngặt.

Cho đến khi phát minh ra [điện báo](https://vi.wikipedia.org/wiki/%C4%90i%E1%BB%87n_b%C3%A1o) năm 1835 thì thời kỳ hiện đại của dự báo thời tiết mới bắt đầu. Trước thời điểm này, thực tế không thể truyền tin từ nơi xa về trạng thái hiện tại của thời tiết nhanh hơn một chuyến tàu hơi nước (và tàu chạy bằng hơi nước cũng mới được phát minh ở thời đó). Cuối những năm 1840, điện báo đã cho phép báo cáo điều kiện thời tiết trên một vùng rộng lớn một cách tức thời, cho phép dự báo từ những hiểu biết về điều kiện thời tiết ngoài việc thu thập số liệu từ hướng gió. Hai nhà khoa học khai sinh ra khoa học dự báo thời tiết là [Francis Beaufort](https://vi.wikipedia.org/w/index.php?title=Francis_Beaufort&action=edit&redlink=1) (tên tuổi ông gắn với [thang Beaufort](https://vi.wikipedia.org/wiki/Thang_s%E1%BB%A9c_gi%C3%B3_Beaufort)) và người bảo trợ của ông là [Robert Fitzroy](https://vi.wikipedia.org/w/index.php?title=Robert_Fitzroy&action=edit&redlink=1) (người phát triển [áp kế](https://vi.wikipedia.org/wiki/%C3%81p_k%E1%BA%BF) Fitzroy). Cả hai đều có ảnh hưởng đến [hải quân](https://vi.wikipedia.org/wiki/H%E1%BA%A3i_qu%C3%A2n) và [chính phủ](https://vi.wikipedia.org/wiki/Ch%C3%ADnh_ph%E1%BB%A7) [Vương quốc Anh](https://vi.wikipedia.org/wiki/V%C6%B0%C6%A1ng_qu%E1%BB%91c_Anh), mặc dù ở thời điểm đó có những nhạo báng trên báo chí về công việc của hai ông, nghiên cứu của họ lại đạt được niềm tin khoa học, và được [Hải quân Hoàng gia](https://vi.wikipedia.org/wiki/H%E1%BA%A3i_qu%C3%A2n_Ho%C3%A0ng_gia_Anh) chấp nhận, từ đó trở thành những kiến thức cơ sở cho dự báo thời tiết ngày nay. Để truyền đạt thông tin (dữ liệu đo đạc và quan sát tình hình thời tiết hiện tại) một cách chính xác, cần thiết phải có từ vựng tiêu chuẩn miêu tả những khái niệm liên quan đến thời tiết (như mây, gió...); đối với mây cho tới những năm 1890 đã có những bộ atlas hoàn thiện mô tả nó.

# CHƯƠNG I. TỔNG QUAN VỀ DỰ BÁO THỜI TIẾT VÀ LÝ THUYẾT CÂY QUYẾT ĐỊNH

## GIỚI THIỆU

**Dự báo thời tiết** là một ngành ứng dụng của khoa học và công nghệ để tiên đoán trạng thái và vị trí của bầu khí quyển trong tương lai gần. Loài người đã nỗ lực dự báo thời tiết một cách không chính thức từ nhiều thiên niên kỳ trước, và việc dự báo thời tiết một cách chính thức bắt đầu từ thế kỷ mười chín. Công tác dự báo thời tiết được thực hiện bằng cách thu thập số liệu về trạng thái hiện tại của bầu khí quyển và áp dụng những hiểu biết khoa học về các quá trình khí quyển để tiên đoán sự tiến triển của khí quyển.

Nỗ lực dự báo của con người chủ yếu dựa trên cơ sở về sự thay đổi của áp suất khí quyển, điều kiện hiện tại của thời tiết, và điều kiện bầu trời, các mo hình dự báo được sử dụng để dự báo trong tương lai. Do bản chất hỗn loạn của khí quyển nên cần phải có những siêu máy tính để giải các phương trình mô tả bầu khí quyển. Những sai số trong việc đo đạc các số liệu đầu vào và sự hiểu biết chưa hoàn thiện về các hoạt động của khí quyển đã làm cho công tác dự báo trở lên ít chính xác trên nhiều địa điểm trong cùng một khoảng thời gian và khi thời gian dự báo tăng lên. Việc sử dụng kết hợp và liên ứng các mô giúp giảm thiểu sai số và chọn ra được kết quả khả quan nhất.

Cảnh báo thời tiết là dự báo quan trọng bởi vì nó cung cấp thông tin nhằm bảo vệ cuộc sống con người cũng như tài sản và các hoạt động ngoài trời. Dự báo về nhiệt độ và lượng mưa là quan trọng trong nông nghiệp, giao thông,...

## DÙNG CÂY QUYẾT ĐỊNH DỰ BÁO THỜI TIẾT.

### Tên Đề Tài

Dùng cây quyết định dự báo thời tiết

### Mục Đích Đề Tài

Giúp cho chúng ta có thể biết được tình hình thời tiết trong ngày thông qua các yếu tố thời tiết nào đó được đo đạc cùng thời điểm. Qua đó chúng ta có thể chủ động hơn với kế hoạch hằng ngày

### Phát biểu bài toán

Bài toán là một hệ chuyên gia nhằm dự báo một số kiểu thời tiết thông thường như: mứa nắng, mát mẻ,râm mát. Với mong muốn có thê dự báo cho chúng ta biết được các kiểu thời tiết để chúng ta có thể chủ động hơn với các hoạt động hằng ngày.Để làm được điều đó thì hệ thống cần một cơ sở tri thức đầy đủ về các kiểu thời tiết nêu trên , tương ứng đó thì hệ thống cần có một ứng dụng để sử dụng các dữ liệu trong bộ nhớ, móc xích chúng lại với nhau, để đauw ra kết luận cuối cùng. Với việc sử dụng cây quyết định, chúng ta có thể đưa ra những lập luận chính xác để đạt được mục đích yêu cầ

# CHƯƠNG II. 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[1]. 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 [1]. 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*)

## Uư/ 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.

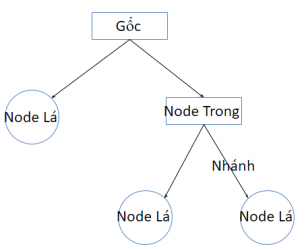
# Cây quyết định hay gặp vấn đề [overfitting](https://trituenhantao.io/tu-dien-thuat-ngu/overfitting/)

**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.

# CHƯƠNG III: NGUỒN DỮ LIỆU VÀ GIAO DIỆN DỰ BÁO THỜI TIẾT

Giao diện

