**TRƯỜNG ĐẠI HỌC ĐIỆN LỰC**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO CHUYÊN ĐỀ HỌC PHẦN**

**KHAI PHÁ DỮ LIỆU**

**SỬ DỤNG CÂY QUYẾT ĐỊNH ĐỂ DỰ ĐOÁN NƠI SẢN XUẤT CỦA CÁC Ô TÔ BỊ HƯ HỎNG DO TAI NẠN**

|  |  |
| --- | --- |
|  |  |
| |  |  | | --- | --- | | **Sinh viên thực hiện** | **:LẠI THỊ TÂM**  **VŨ KHÁNH HUYỀN**  **PHẠM LAN ANH** | | **Giảng viên hướng dẫn** | **: VŨ VĂN ĐỊNH** | | | **Ngành** | **: CÔNG NGHỆ THÔNG TIN** | | | **Chuyên ngành** | **: CÔNG NGHỆ PHẦN MỀM** | | | **Lớp** | **: D16CNPM4** | | | **Khóa** | **: 2021-2026** | | |  |

**Hà Nội, tháng 5 năm 2024**

**PHIẾU CHẤM ĐIỂM**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Họ và Tên** | **Nội dung thực hiện** | **Điểm** | **Chữ ký** |
| 1 | Lại Thị Tâm | Báo cáo + tìm hiểu cây quyết định |  |  |
| 2 | Vũ Khánh Huyền | Báo cáo + tìm hiểu cây quyết định |  |  |
| 3 | Phạm Lan Anh | Báo cáo + tìm hiểu cây quyết định |  |  |

|  |  |  |
| --- | --- | --- |
| **Họ và tên** | **Chữ ký** | **Ghi chú** |
| Giảng viên chấm 1: |  |  |
| Giảng viên chấm 2: |  |  |

**MỤC LỤC**

LỜI MỞ ĐẦU

[CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU 2](#_Toc163201178)

[1.1 Khái niệm chung: 2](#_Toc163201179)

[1.1.1 Khái niệm cơ bản: 2](#_Toc163201180)

[1.1.2 Các bước trong quá trình khai phá 2](#_Toc163201181)

[1.1.3 Ứng dụng của khai phá dữ liệu 4](#_Toc163201182)

[1.2 Tiền xử lý dữ liệu 4](#_Toc163201183)

[1.2.1 Dữ liệu 4](#_Toc163201184)

[1.2.2 Làm sạch dữ liệu (data cleaning) 4](#_Toc163201185)

[1.2.3 Tích hợp dữ liệu (data integration) 5](#_Toc163201186)

[1.2.4 Biến đổi dữ liệu (data transformation) 6](#_Toc163201187)

[1.2.5. Thu giảm dữ liệu (data reduction) 6](#_Toc163201188)

[1.3 Phương pháp khai phá 7](#_Toc163201189)

[1.3.1 Phân loại 7](#_Toc163201190)

[1.3.2. Hồi quy 7](#_Toc163201191)

[1.3.3 Luật kết hợp 8](#_Toc163201192)

[1.3.4 Phân cụm 8](#_Toc163201193)

[CHƯƠNG 2: GIẢI QUYẾT BÀI TOÁN CÂY QUYẾT ĐỊNH 9](#_Toc163201194)

[2.1. Tổng quan về thuật toán cây quyết định 9](#_Toc163201195)

[2.1.1. Giới thiệu thuật toán 9](#_Toc163201196)

[2.1.2. Cấu trúc của thuật toán 11](#_Toc163201197)

[2.2. Các thuật toán liên quan đến cây quyết định 11](#_Toc163201198)

[2.2.1. ID3 (Iterative Dichotomiser 3) Algorithm 11](#_Toc163201199)

[2.2.2. ID3 (Examples, Target\_Attribute, Attributes) 11](#_Toc163201200)

[2.2.3. C4.5 Algorithm 13](#_Toc163201201)

[2.3. Ứng dụng của thuật toán 16](#_Toc163201202)

[2.4. Ưu và nhược điểm của thuật toán 16](#_Toc163201203)

[2.4.1. Ưu điểm: 16](#_Toc163201204)

[2.4.2. Nhược điểm: 17](#_Toc163201205)

[CHƯƠNG 3: ỨNG DỤNG THUẬT TOÁN 19](#_Toc163201206)

[3.1 Thuật toán phân lớp cây quyết định dự đoán nơi sản xuất của các ô tô bị hư hỏng do tai nạn 19](#_Toc163201207)

[3.1.1. Phát biểu bài toán: 19](#_Toc163201208)

[3.1.2 Yêu cầu bài toán: 19](#_Toc163201209)

[3.1.3. Xây dựng bộ dữ liệu: 19](#_Toc163201210)

[KẾT LUẬN 26](#_Toc163201211)

[TÀI LIỆU THAM KHẢO 27](#_Toc163201212)

**DANH MỤC HÌNH ẢNH**

[Hình 1. 1 Quá trình khai phá tri thức 3](#_Toc163201270)

[Hình 1. 2 Minh họa về cây quyết định 10](#_Toc163201271)

**LỜI MỞ ĐẦU**

Trong nhiều năm qua, cùng với sự phát triển của công nghệ thông tin và ứng dụng của công nghệ thông tin trong nhiều lĩnh vực của đời sống xã hội, thì lượng dữ liệu được các cơ quan thu thập và lưu trữ ngày một nhiều lên. Người ta lưu trữ những dữ liệu này vì cho rằng nó ẩn chứa những giá trị nhất định nào đó. Tuy nhiên theo thống kê thì chỉ có một lượng nhỏ của những dữ liệu này (khoảng từ 5% đến 10%) là luôn được phân tích, số còn lại họ không biết sẽ phải làm gì và có thể làm gì với những dữ liệu này, nhưng họ vẫn tiếp tục thu thập và lưu trữ vì hy vọng những dữ liệu này sẽ cung cấp cho họ những thông tin quý giá một cách nhanh chóng để đưa ra những quyết định kịp thời vào một lúc nào đó.

Chính vì vậy, các phương pháp quản trị và khai thác cơ sở dữ liệu truyền thống ngày càng không đáp ứng được thực tế đã làm phát triển một khuynh hướng kỹ thuật mới đó là Kỹ thuật phát hiện tri thức và khai phá dữ liệu (KDD - Knowledge Discovery and Data Mining). Kỹ thuật phát hiện tri thức và khai phá dữ liệu đã và đang được nghiên cứu, ứng dụng trong nhiều lĩnh vực khác nhau trên thế giới, tại Việt Nam kỹ thuật này còn tương đối mới mẻ tuy nhiên cũng đang được nghiên cứu và bắt đầu đưa vào một số ứng dụng thực tế.

Vì vậy, hiện nay ở nước ta vấn đề phát hiện tri thức và khai phá dữ liệu đang thu hút được sự quan tâm của nhiều người và nhiều công ty phát triển ứng dụng công nghệ thông tin.

# CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

## 1.1 Khái niệm chung:

### 1.1.1 Khái niệm cơ bản:

"Data Mining" là "quá trình khai phá", trích xuất, khai thác và sử dụng những dữ liệu có giá trị tiềm ẩn từ bên trong lượng lớn dữ liệu được lưu trữ trong các cơ sở dữ liệu (CSDL), kho dữ liệu, trung tâm dữ liệu lớn hơn là Big Data dựa trên kĩ thuật như mạng nơ ron, lí thuyết tập thô, tập mờ, biểu diễn tri thức. Data Mining là một công đoạn trong hoạt động “làm sạch” dữ liệu giúp cho dữ liệu được truyền dẫn một cách nhanh nhất.

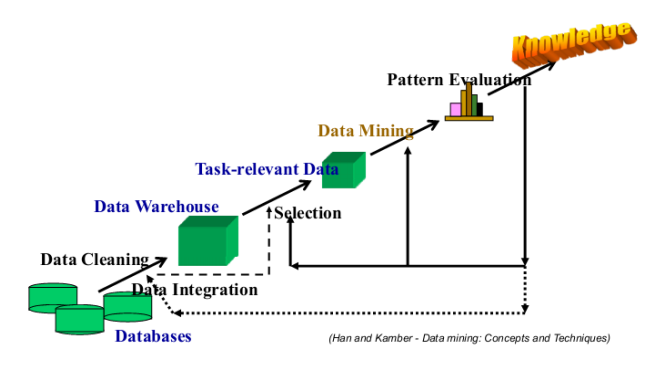
         Quá trình khai phá dữ liệu là một quá trình phức tạp bao gồm kho dữ liệu chuyên sâu cũng như các công nghệ tính toán. Hơn nữa, Data Mining không chỉ giới hạn trong việc trích xuất dữ liệu mà còn được sử dụng để chuyển đổi, làm sạch, tích hợp dữ liệu và phân tích mẫu.Có nhiều tham số quan trọng khác nhau trong Data Mining, chẳng hạn như quy tắc kết hợp, phân loại, phân cụm và dự báo. Một số tính năng chính của Data Mining:

* Dự đoán các mẫu dựa trên xu hướng trong dữ liệu.
* Tính toán dự đoán kết quả
* Tạo thông tin phản hồi để phân tích
* Tập trung vào cơ sở dữ liệu lớn hơn.
* Phân cụm dữ liệu trực quan

### 1.1.2 *Các Bước trong Data mining:*

Các bước quan trọng khi Data Mining bao gồm:

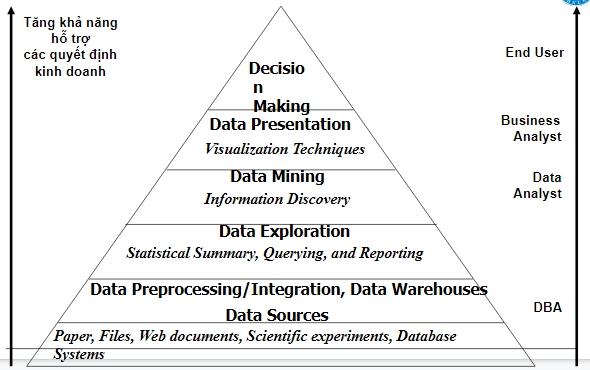
* Bước 1: Làm sạch dữ liệu – Trong bước này, dữ liệu được làm sạch sao cho không có dữ liệu tạp hay bất thường trong dữ liệu.
* Bước 2: Tích hợp dữ liệu – Trong quá trình tích hợp dữ liệu, nhiều nguồn dữ liệu sẽ kết hợp lại thành một.
* Bước 3: Lựa chọn dữ liệu – Trong bước này, dữ liệu được trích xuất từ cơ sở dữ liệu.
* Bước 4: Chuyển đổi dữ liệu – Trong bước này, dữ liệu sẽ được chuyển đổi để thực hiện phân tích tóm tắt cũng như các hoạt động tổng hợp.
* Bước 5: Khai phá dữ liệu – Trong bước này, chúng tôi trích xuất dữ liệu hữu ích từ nhóm dữ liệu hiện có.
* Bước 6: Đánh giá mẫu – Chúng tôi phân tích một số mẫu có trong dữ liệu.
* Bước 7: Trình bày thông tin – Trong bước cuối cùng, thông tin sẽ được thể hiện dưới dạng cây, bảng, biểu đồ và ma trận.
* Đây là một góc nhìn từ hệ thống cơ sở dữ liệu điển hình.
* Khai phá dữ liệu đóng một vai trò thiết yếu trong quá trình khai phá trí thức.



Hình 1. 1 Mô hình mô tả quá trình khai phá dữ liệu

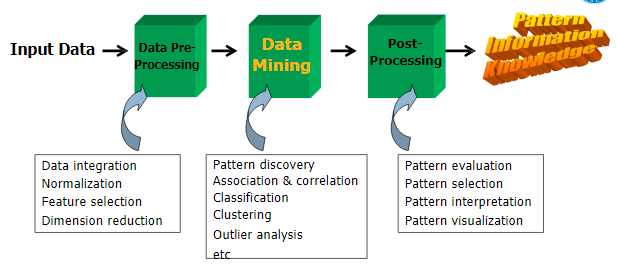
*Quá trình khai phá trí thức là một chuỗi lặp:*

* Data cleaning (làm sạch dữ liệu)
* Data integration (tích hợp dữ liệu)
* Data selection (chọn lựa dữ liệu)
* Data transformation (biến đổi dữ liệu)
* Data mining (khai phá dữ liệu)
* Pattern evaluation (đánh giá mẫu)
* Knowledge presentation (biểu diễn tri thức)
* Quá trình khai phá dữ liệu trong kinh doanh thông minh



*Hình 1.2:Mô hình mô tả quá trình khai phá trong kinh doanh*

- Quá trình khai phá dữ liệu từ góc nhìn máy học thống kê



*Hình 1.3:Mô hình mô tả quá trinh khai phá từ máy học thống kê*

## 1.2. Kiến trúc của một hệ thống khai phá dữ liệu

Khai phá dữ liệu là quá trình rút trích thông tin bổ ích từ những kho dữ liệu lớn, là quá trình chính trong khai phá trí thức từ cơ sở dữ liệu.

Kiến trúc của một hệ thống khai phá dữ liệu có các thành phần như:

### C:\Users\Pham Quynh\OneDrive\Hình ảnh\Cuộn phim\KPDL\Capture.PNG

*Hình 1.4: Khám phá trí thức trong cơ sở dữ liệu điển hình*

* Cơ sở dữ liệu, kho dữ liệu hoặc lưu trữ thông tin khác: Đây là một hay các tập cơ sở dữ liệu, các kho dữ liệu, các trang tính hay các dạng khác của thông tin được lưu trữ. Các kỹ thuật làm sạch dữ liệu và tích hợp dữ liệu có thể được thực hiện.
* Máy chủ cơ sở dữ liệu (Database or Warehouse Server): Máy chủ có trách nhiệm lấy những dữ liệu thích hợp dựa trên những yêu cầu khám phá của người dùng.
* Cơ sở trí thức (Knowledge-base): Đây là miền tri thức dùng để tìm kiếm hay đánh giá độ quan trọng của các mẫu kết quả thu được. Trí thức này có thể bao gồm một sự phân cấp khái niệm dùng để tổ chức các thuộc tính hay các giá trị thuộc tính ở các mức trừu tượng khác nhau.
* Máy khai phá dữ liệu (Data mining engine): Là một hệ thống khai phá dữ liệu cần phải có một tập các modul chức năng để thực hiện công việc, chẳng hạn như kết hợp, phân lớp, phân cụm.
* Modul đánh giá mẫu (Pattern evaluation): Bộ phận tương tác với các modul khai phá dữ liệu để tập trung vào việc duyệt tìm các mẫu đáng được quan tâm. Nó có thể dùng các ngưỡng về độ quan tâm để lọc mẫu đã khám phá được. Cũng có thể modul đánh giá mẫu được tích hợp vào modul khai phá dữ liệu, tùy theo các cài đặt của phương pháp khai phá dữ liệu được dùng.
* Giao diện đồ họa người dùng (Graphical user interface): Bộ phận này cho phép người dùng giao tiếp với hệ thống khai phá dữ liệu. Thông qua giao diện này người dùng tương tác với hệ thống bằng cách đặc tả một yêu cầu khai phá hay một nhiệm vụ, cung cấp thông tin trợ giúp cho việc tìm kiếm và thực hiện khia phá thăm dò trên các kết quả khai phá trung gian. Ngoài ra bộ phận này còn cho phép người dùng xem các lược đồ cơ sở dữ liệu, lược đồ kho dữ liệu, các đánh giá mẫu và hiển thị các mẫu trong các khuôn dạng khác nhau.

## 1.3.  Chức năng khai phá dữ liệu:

Các kỹ thuật khai phá dữ liệu thực hiện 2 chức năng chính:

* Chức năng mô tả: mô tả về các tính chất hoặc các đặc tính chung của dữ liệu trong cơ sở dữ liệu, các kỹ thuật này gồm có: phân cụm (Clustering), tổng hợp (Summarization), trực quan hóa (Visualization), phân tích sự phát triển và độ lệnh (Evolution and deviation analysis), phân tích luật kết hợp (Association rules),...
* Chức năng dự đoán: đưa ra các dự đoán dựa vào các suy diễn trên dữ liệu hiện thời, các kỹ thuật này gồm có: phân lớp (Classification), hồi quy (Regression), cây quyết định (Decision tree), ....

### 1.3.1. Phân lớp dữ liệu (Classification)

* Là dạng phân tích dữ liệu nhằm rút trích các mô hình mô tả các lớp dữ liệu hoặc dự đoán xu hướng dữ liệu.
* Quá trình gồm hai bước:
* Bước học (giai đoạn huấn luyện): xây dựng bộ phân lớp (classifier) bằng việc phân tích/học tập huấn luyện.
* Bước phân lớp (classification): phân lớp dữ liệu/đối tượng mới nếu độ chính xác của bộ phân lớp được đánh giá là có thể chấp nhận được (acceptable).
* Các thuật toán phân lớp dữ liệu
* Phân lớp với cây quyết định (decision tree)
* Phân lớp với Naïve Bayesian
* Phân lớp với k phần tử gần nhất (k-nearest neighbor)
* Phân lớp với máy vector hỗ trợ (SVM)
* Phân lớp với mạng neural (neural network)
* Phân lớp dựa trên tiến hoá gen (genetic algorithms)
* Phân lớp với lý thuyết tập thô, tập mờ (rough sets)
* Phân lớp với lý thuyết tập mờ (fuzzy sets) ...

### 1.3.2. Hồi quy (Regression)

Là việc học một hàm ánh xạ từ một mẫu dữ liệu thành một biến dự đoán có giá trị thực. Nhiệm vụ của hồi quy tương tự như phân lớp, điểm khác nhau chính là ở chỗ thuộc tính để dự báo là liên tục chứ không phải là rời rạc. Việc dự báo các giá trị số thường được làm bởi các phương pháp thống kê cổ điển, chẳng hạn như hồi quy tuyến tính. Tuy nhiên, phưong pháp mô hình hóa cũng được sử dụng, ví dụ: cây quyết định.

### *1.3.3. Phân cụm dữ liệu (Clustering)*

* Là quá trình phân nhóm/cụm dữ liệu/đối tượng vào các lớp/cụm.
* Các đối tượng trong cùng một cụm tương tự với nhau hơn so với đối tượng ở cụm khác.
* Các yêu cầu của phân cụm dữ liệu:
* Có thể tương thích, hiệu quả với dữ liệu lớn, số chiều lớn.
* Có khả năng xử lý các dữ liệu khác nhau.
* Có khả năng khám phá các cụm với các dạng bất kỳ.
* Khả năng thích nghi với dữ liệu nhiễu.
* Ít nhạy cảm với thứ tự của các dữ liệu vào.
* Phân cụm rằng buộc.
* Dễ hiểu và dễ sử dụng.
* Phân loại các phương pháp phân cụm:
* Phân hoạch (partitioning): phân hoạch tập dữ liệu n phần tử thành k cụm.
* Phân cấp (hierarchical): xây dựng phân cấp các cụm trên cơ sở các đối tượng dữ liệu đang xem xét.
* Dựa trên mật độ (density-based): dựa trên hàm mật độ, số đối tượng lân cận của đối tượng dữ liệu.
* Dựa trên lưới (grid-based): dựa trên dữ liệu nhiều chiều, chủ yếu áp dụng cho lớp dữ liệu không gian.
* Dựa trên mô hình (model-based): một mô hình giả thuyết được đưa ra cho mỗi cụm; sau đó hiệu chỉnh các thông số để mô hình phù hợp với cụm dữ liệu/đối tượng nhất.

### 1.3.4. Khai phá luật kết hợp (Association rules)

Chẳng hạn như có luật: âm nhac, thể thao => thiếu nhi, nghĩa là những người mua sách âm nhạc và thể thao thì cũng mua sách thiếu nhi. Lúc đó ta sẽ quan tâm đến số lượng trường hợp khách hàng thỏa mãn luật này trong cơ sở dữ liệu hay độ hỗ trợ (Support) cho luật này. Độ hỗ trợ cho luật chính là phần trăm số bản ghi có cả sách âm nhạc, thể thao và thiếu nhi hay tất cả những người thích cả ba loại sách nói trên. Tuy nhiên, giá trị độ hỗ trợ là không đủ, có thể có trường hợp ta có một nhóm tương đối những người đọc cả ba loại trên nhƣng lại có một nhóm với lực lượng lớn hơn những người thích sách thể thao, âm nhạc mà không thích sách thiếu nhi. Trong trường hợp này tính kết hợp rất yếu mặc dù độ hỗ trợ tương đối cao, như vậy chúng ta cần thêm một độ đo thứ hai đó là độ tin cậy (confidence). Độ tin cậy chính là phần trăm các bản ghi có sách thiếu nhi trong số các bản ghi có sách âm nhạc và thể thao.

### 1.3.5. Phân tích sự phát triển và độ lệch

Nhiệm vụ này tập trung vào khám phá hầu hết sự thay đổi có nghĩa dưới dạng độ đo đã biết trước hoặc giá trị chuẩn, phát hiện độ lệch đáng kể giữa nội dung của tập con dữ liệu thực và nội dung mong đợi. Hai mô hình độ lệch hay dùng là lệch theo thời gian hay lệch theo nhóm. Độ lệch theo thời gian là sự thay đổi có ý nghĩa của dữ liệu theo thời gian. Độ lệch theo nhóm là sự khác nhau giữa dữ liệu trong hai tập con dữ liệu, ở đây tính cả trường hợp tập con dữ liệu này thuộc tập con kia, nghĩa là xác định dữ liệu trong một nhóm con của đối tượng có khác đáng kể so với toàn bộ đối tượng không? Theo cách này, sai sót dữ liệu hay sai lệch so với giá trị thông thường được phát hiện.

Vì những nhiệm vụ này yêu cầu số lượng và các dạng thông tin rất khác nhau nên chúng thường ảnh hưởng đến việc thiết kế và chọn phương pháp khai phá dữ liệu khác nhau. Ví dụ như phương pháp cây quyết định tạo ra được một mô tả phân biệt được các mẫu giữa các lớp nhưng không có tính chất và đặc điểm của lớp.

# CHƯƠNG 2: GIẢI QUYẾT BÀI TOÁN CÂY QUYẾT ĐỊNH

## 2.1. Tổng quan về thuật toán cây quyết định

### 2.1.1. Giới thiệu thuật toán

Trong lĩnh vực [máy học](https://vi.wikipedia.org/wiki/M%C3%A1y_h%E1%BB%8Dc), 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](https://vi.wikipedia.org/wiki/Khai_ph%C3%A1_d%E1%BB%AF_li%E1%BB%87u). 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](https://vi.wikipedia.org/wiki/T%E1%BA%ADp_h%E1%BB%A3p) 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](https://vi.wikipedia.org/wiki/%C4%90%E1%BB%87_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](https://vi.wikipedia.org/w/index.php?title=R%E1%BB%ABng_ng%E1%BA%ABu_nhi%C3%AAn&action=edit&redlink=1) (*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](https://vi.wikipedia.org/wiki/X%C3%A1c_su%E1%BA%A5t_c%C3%B3_%C4%91i%E1%BB%81u_ki%E1%BB%87n).

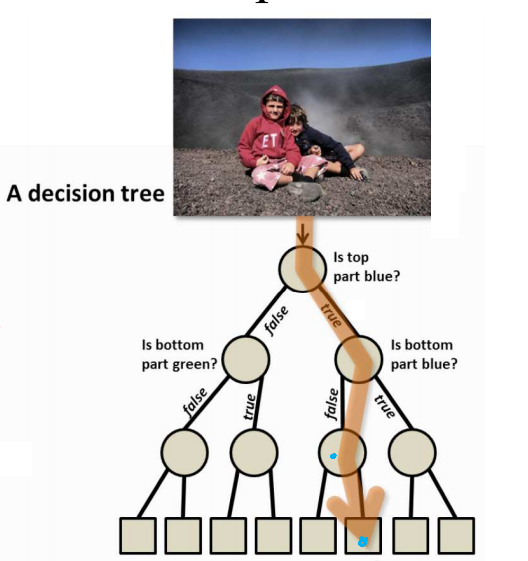
Cây quyết định có thể được mô tả như là sự kết hợp của các kỹ thuật toán học và tính toán nhằm hỗ trợ việc mô tả, phân loại và tổng quát hóa một tập dữ liệu cho trước.

Dữ liệu được cho dưới dạng các bản ghi có dạng:

*(x, y) = (x1, x2, x3..., xk, y)*

Biến phụ thuộc (*dependant variable*) *y* là biến mà chúng ta cần tìm hiểu, phân loại hay tổng quát hóa. *x1*, *x2*, *x3*... là các biến sẽ giúp ta thực hiện công việc đó.

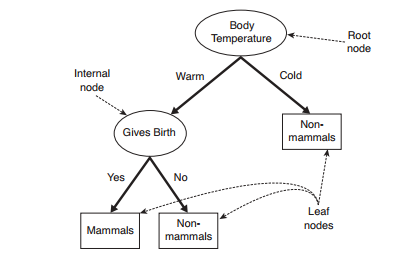
Tóm lại, cây quyết định (Decision Tree) là một cây phân cấp có cấu trúc được dùng để phân lớp các đối tượng dựa vào dãy các luật (series of rules). Khi cho dữ liệu về các đối tượng gồm các thuộc tính cùng với lớp (classes) của nó, cây quyết định sẽ sinh ra các luật để dự đoán lớp của các đối tượng chưa biết (unseen data).



Hình 1. 2 Minh họa về cây quyết định

### 2.1.2. Cấu trúc của thuật toán

Decision Trees gồm 3 phần chính: 1 node gốc (root node), những node lá (leaf nodes) và các nhánh của nó (branches). Node gốc là điểm bắt đầu của cây quyết định và cả hai node gốc và node chứa câu hỏi hoặc tiêu chí để được trả lời. Nhánh biểu diễn các kết quả của kiểm tra trên nút. Ví dụ câu hỏi ở node đầu tiên yêu cầu câu trả lời là “yes” hoặc là “no” thì sẽ có 1 node con chịu trách nhiệm cho phản hồi là “yes”, 1 node là “no”.



## 2.2. Các thuật toán liên quan đến cây quyết định

## 2.3. Ứng dụng của thuật toán

Trong [lý thuyết quyết định](https://vi.wikipedia.org/wiki/L%C3%BD_thuy%E1%BA%BFt_quy%E1%BA%BFt_%C4%91%E1%BB%8Bnh) (chẳng hạn [quản lý rủi ro](https://vi.wikipedia.org/wiki/Qu%E1%BA%A3n_l%C3%BD_r%E1%BB%A7i_ro)), một cây quyết định (decision tree) là một [đồ thị](https://vi.wikipedia.org/wiki/L%C3%BD_thuy%E1%BA%BFt_%C4%91%E1%BB%93_th%E1%BB%8B) của các quyết định và các hậu quả có thể của nó (bao gồm rủi ro và hao phí tài nguyên). Cây quyết định được sử dụng để xây dựng một [kế hoạch](https://vi.wikipedia.org/wiki/K%E1%BA%BF_ho%E1%BA%A1ch) nhằm đạt được [mục tiêu](https://vi.wikipedia.org/wiki/M%E1%BB%A5c_ti%C3%AAu) mong muốn. Các cây quyết định được dùng để hỗ trợ quá trình ra quyết định. Cây quyết định là một dạng đặc biệt của [cấu trúc cây](https://vi.wikipedia.org/wiki/C%E1%BA%A5u_tr%C3%BAc_c%C3%A2y).

## 2.4. Ưu và nhược điểm của thuật toán

### 2.4.1. Ưu điểm:

So với các phương pháp [khai phá dữ liệu](https://vi.wikipedia.org/wiki/Khai_ph%C3%A1_d%E1%BB%AF_li%E1%BB%87u) khác, cây quyết định là phương pháp có một số ưu điểm:

* Cây quyết định dễ hiểu. Người ta có thể hiểu mô hình cây quyết định sau khi được giải thích ngắn.
* Việc chuẩn bị dữ liệu cho một cây quyết định là cơ bản hoặc không cần thiết. Các kỹ thuật khác thường đòi hỏi [chuẩn hóa dữ liệu](https://vi.wikipedia.org/wiki/Chu%E1%BA%A9n_h%C3%B3a_d%E1%BB%AF_li%E1%BB%87u), cần tạo các biến phụ (dummy variable) và loại bỏ các giá trị rỗng.
* Cây quyết định có thể xử lý cả dữ liệu có giá trị bằng số và dữ liệu có giá trị là tên thể loại. Các kỹ thuật khác thường chuyên để phân tích các bộ dữ liệu chỉ gồm một loại biến. Chẳng hạn, các luật quan hệ chỉ có thể dùng cho các biến tên, trong khi [mạng nơ-ron](https://vi.wikipedia.org/wiki/M%E1%BA%A1ng_n%C6%A1-ron) chỉ có thể dùng cho các biến có giá trị bằng số.
* Cây quyết định là một mô hình hộp trắng. Nếu có thể quan sát một tình huống cho trước trong một mô hình, thì có thể dễ dàng giải thích điều kiện đó bằng logic Boolean. Mạng nơ-ron là một ví dụ về mô hình hộp đen, do lời giải thích cho kết quả quá phức tạp để có thể hiểu được.
* Có thể thẩm định một mô hình bằng các kiểm tra thống kê. Điều này làm cho ta có thể tin tưởng vào mô hình.
* Cây quyết định có thể xử lý tốt một lượng dữ liệu lớn trong thời gian ngắn. Có thể dùng [máy tính cá nhân](https://vi.wikipedia.org/wiki/M%C3%A1y_t%C3%ADnh_c%C3%A1_nh%C3%A2n) để phân tích các lượng dữ liệu lớn trong một thời gian đủ ngắn để cho phép các nhà chiến lược đưa ra quyết định dựa trên phân tích của cây quyết định.

### 2.4.2. Nhược điểm:

Nhược điểm của cây quyết định - khó giải quyết được những vấn đề có dữ liệu phụ thuộc thời gian liên tục - dễ xảy ra lỗi khi có quá nhiều lớp chi phí tính toán để xây dựng mô hình cây quyết định cao.

* Hiệu của phân lớp của cây quyết định (Series of Rules) phụ thuộc rất lớn vào training data. Chẳng hạn cây quyết định được tạo ra bởi chỉ giới hạn 10 samples training data trong ví dụ trên thì hiệu quả ứng dụng cây quyết định để dự đoán các trường hợp khác là không cao (thường training data phải đủ lớn và tin cậy). Cây quyết định là một phương pháp phân lớp rất hiệu quả và dễ hiểu. Tuy nhiên có một số chú ý khi sử dụng cây quyết định trong xây dựng các mô hình phân lớp như sau:
* Có rất nhiều thuật toán phân lớp như ID3, J48, C4.5, CART (Classification and Regression Tree),… Việc chọn thuật toán nào để có hiệu quả phân lớp cao tùy thuộc vào rất nhiều yếu tố, trong đó cấu trúc dữ liệu ảnh hưởng rất lớn đến kết quả của các thuật toán. Chẳng hạn như thuật toán ID3 và CART cho hiệu quả phân lớp rất cao đối với các trường dữ liệu số (quantitative value) trong khi đó các thuật toán như J48, C4.5 có hiệu quả hơn đối với các dữ liệu Qualititive value (ordinal, Binary, nominal).

# CHƯƠNG 3: ỨNG DỤNG THUẬT TOÁN

## 3.1 Thuật toán phân lớp cây quyết định dự đoán nơi sản xuất của các ô tô bị hư hỏng do tai nạn

### 3.1.1. Phát biểu bài toán:

Bài toán dự đoán nơi sản xuất của loại xe là đưa ra bộ thông tin của các loại xe từ số mile đi được trong 1 đơn vị gallon, loại xi lanh, số inch khối, thời gian sản xuất,... sau đó dự đoán nơi sản xuất đó.

- Input: Thông tin của loại xe

- Output: Dự đoán được nơi sản xuất của loại xe đó.

### 3.1.2 Yêu cầu bài toán:

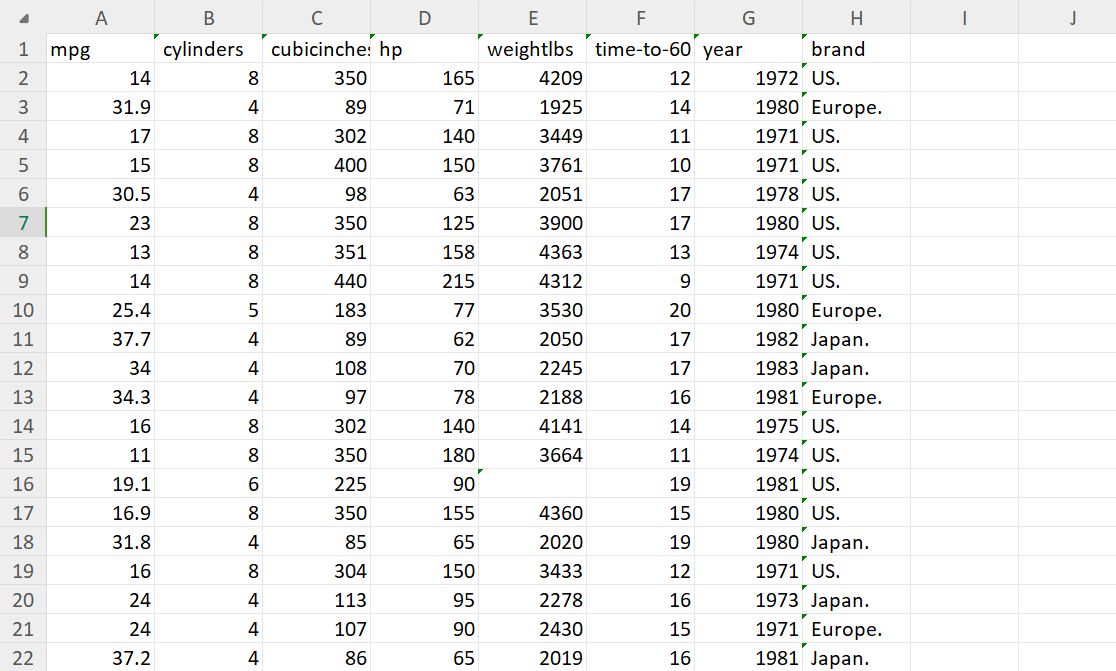
- Lấy dữ liệu từ bộ dữ liệu cung cấp

- Phân tích bộ dữ liệu

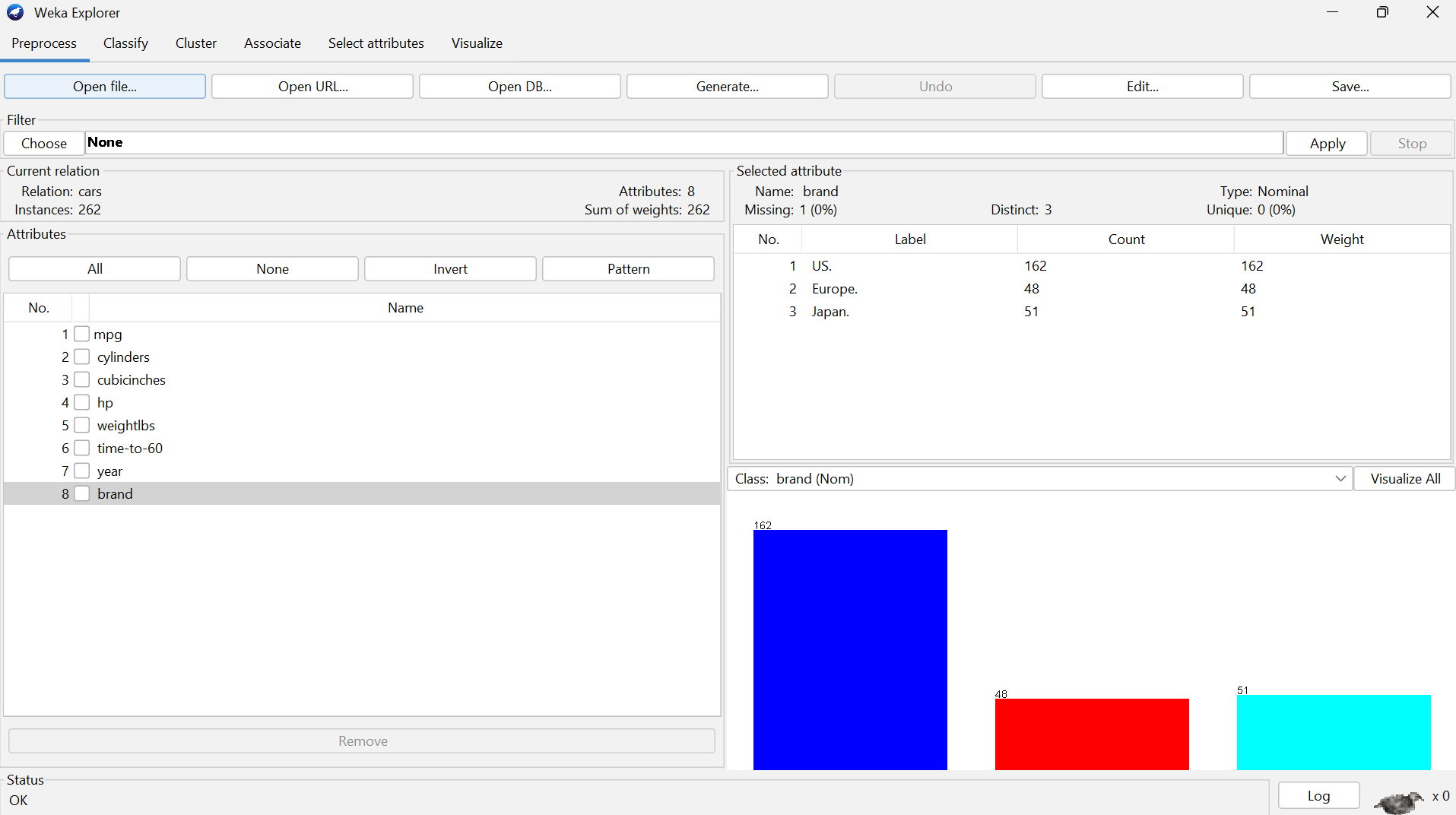
- Chuẩn đoán nơi sản xuất của các loại xe

### 3.1.3. Xây dựng bộ dữ liệu:

Link bộ dữ liệu: <https://www.kaggle.com/abineshkumark/carsdata>



Ta nạp bộ dữ liệu cars.csv vào weka:



Ta thấy có 261 bộ dữ liệu với 8 thuộc tính:

- Mpg: mức tiết kiệm nhiên liệu

- Cylinders : xi lanh

- Cubicinches : Đơn vị thể tích khối

- Hp: Mã lực

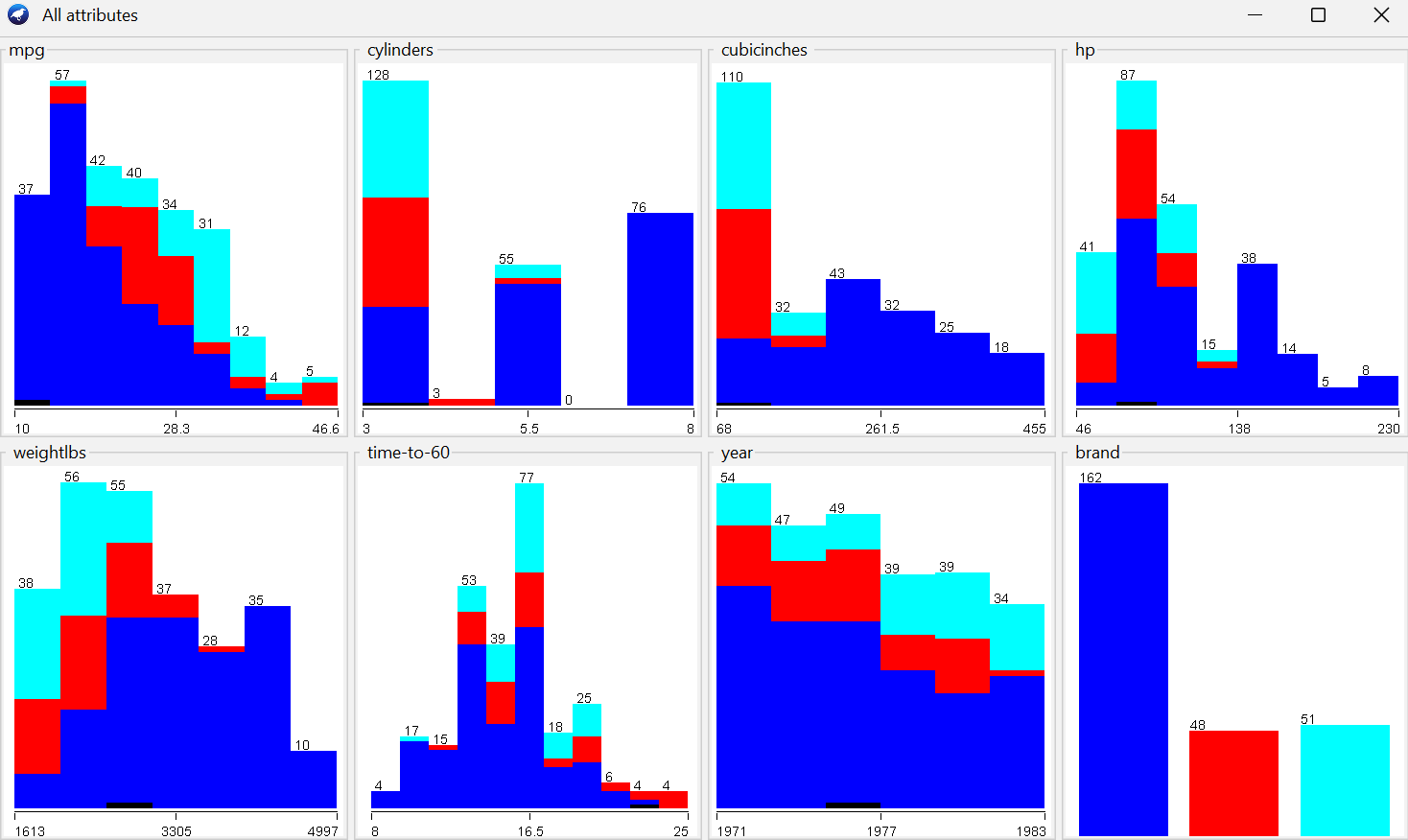
- Weightlbs: trọng lượng

- Time-to-60: thời gian 60

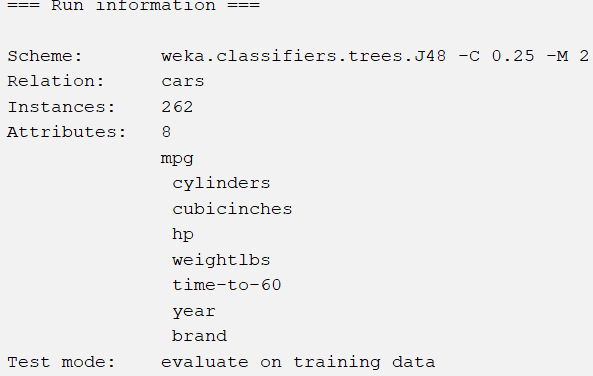
- Year: năm sx

- Brand: nơi sản xuất

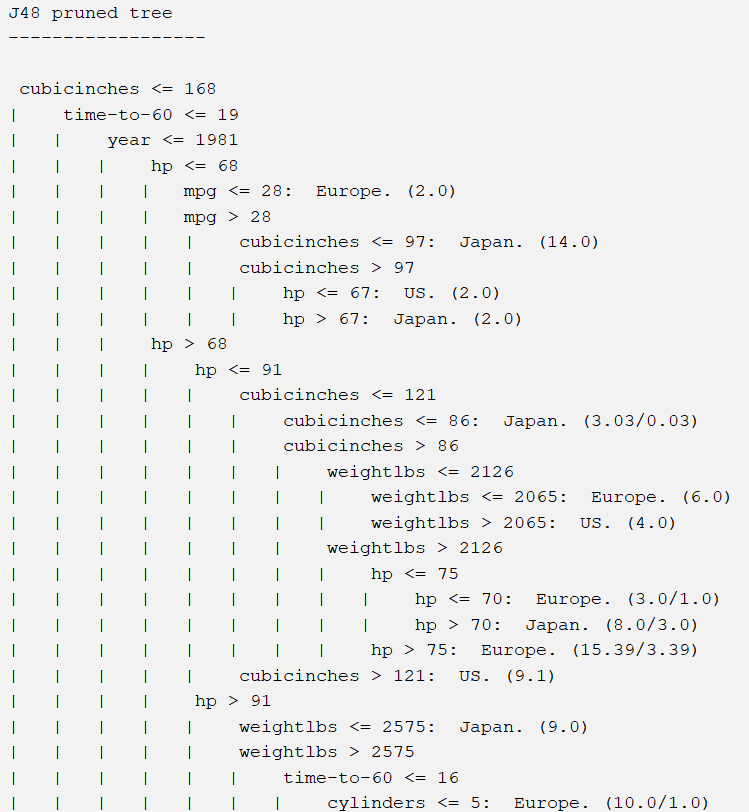
Các thuộc tính:

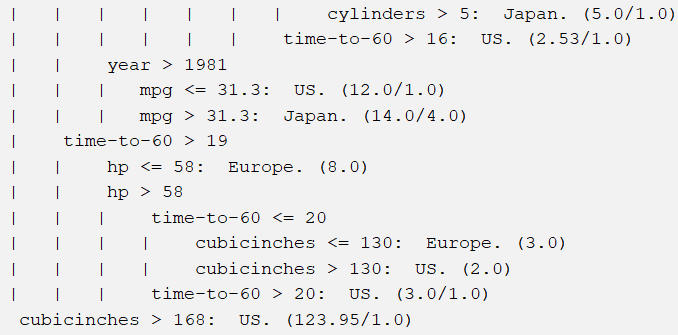


Chạy thông tin:

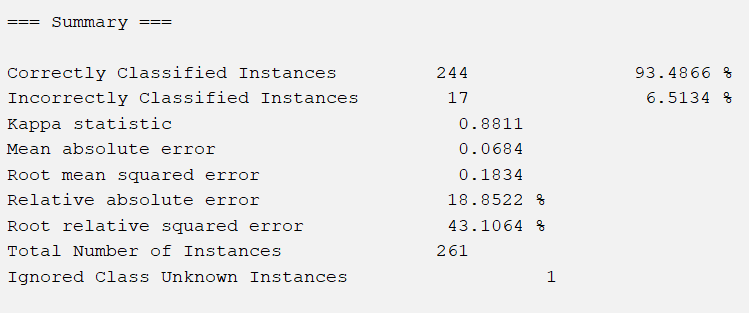


Phân tích dữ liệu

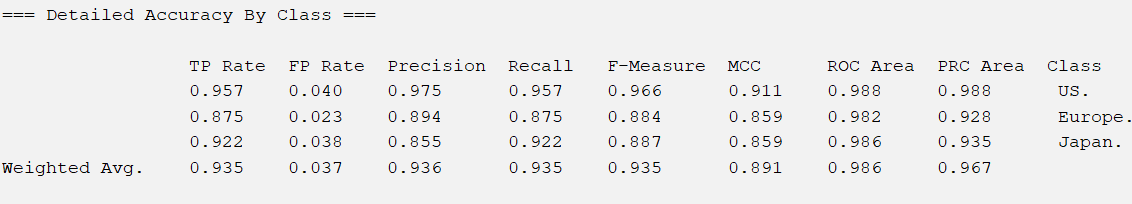




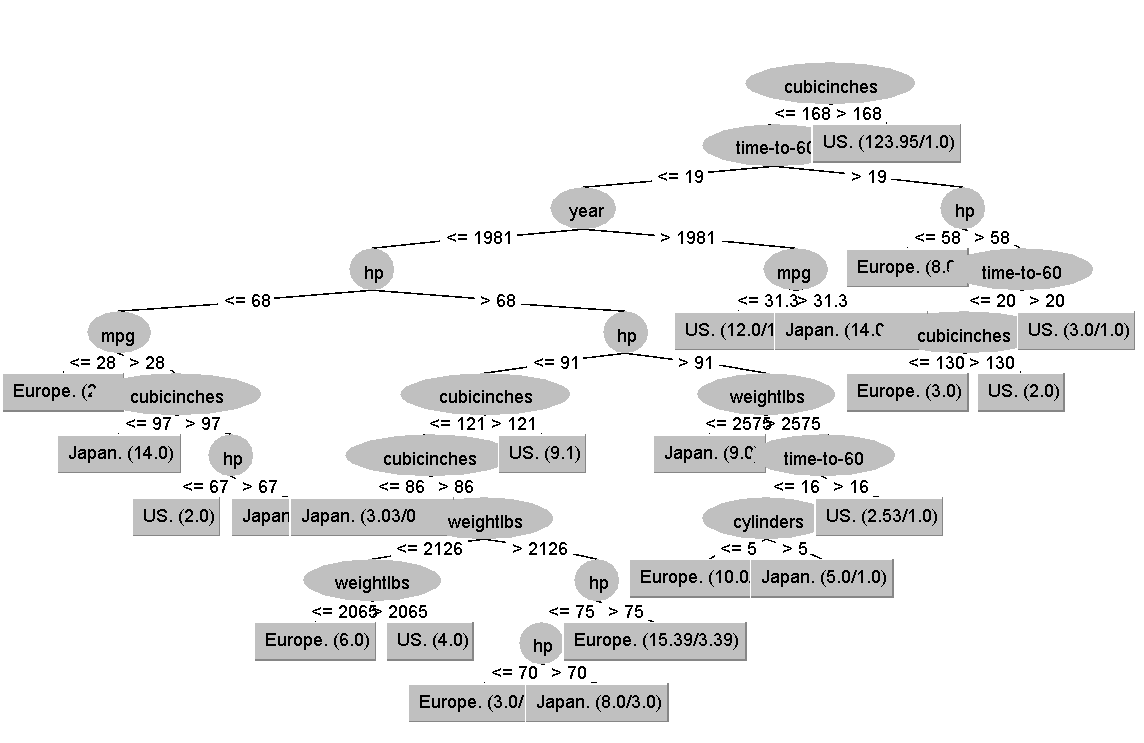
Tổng hợp:



Độ chính xác phân tích theo lớp:



Kết quả:



- Nhìn vào hình ta thấy có thể phân biệt một cách dễ dàng.

- Nơi sản xuất của các xe được suy ra từ các thuộc tính có thể nhìn vào hình giải thích dễ dàng.

- Mang lại hiệu quả về mặt tìm kiếm nơi sản xuất.

- Đánh giá đúng về loại xe đó.

# KẾT LUẬN

# TÀI LIỆU THAM KHẢO