### 2.2.1.1. Thuật toán ID3

Giờ chúng ta tìm hiểu cách thức hoạt động của thuật toán cây quyết định thông qua thuật toán đơn giản ID3.

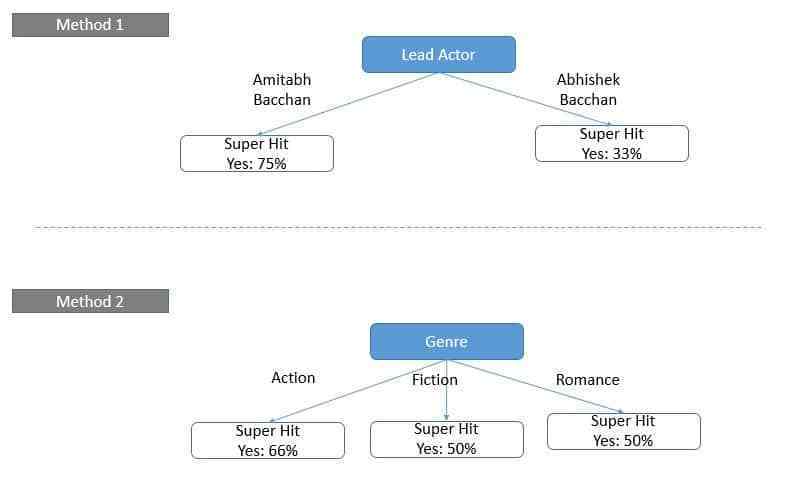
ID3 là viết tắt của Iterative Dichotomizer3 và được đặt tên như vậy vì thuật toán lặp lại (lặp đi lặp lại) phân đôi (chia) các đặc tính thành hai hoặc nhiều nhóm ở mỗi bước. ID3 là thuật toán do Ross Quinlan phát minh, được sử dụng để tạo cây quyết định từ tập dữ liệu và là thuật toán thuật toán phổ biến nhất được sử dụng để xây dựng cây. ID3 sử dụng [Entropy](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/) và Information Gain để xây dựng một cây quyết định.

Ta xét ví dụ :

Bạn muốn xem xét sự thành công của một bộ phim thông qua hai yếu tố: diễn viên chính của phim và thể loại phim:

| Lead Actor | Genre | Hit(Y/N) |
| --- | --- | --- |
| Amitabh Bacchan | Action | Yes |
| Amitabh Bacchan | Fiction | Yes |
| Amitabh Bacchan | Romance | No |
| Amitabh Bacchan | Action | Yes |
| Abhishek Bacchan | Action | No |
| Abhishek Bacchan | Fiction | No |
| Abhishek Bacchan | Romance | Yes |

Giả sử, bạn muốn xác định độ thành công của bộ phim chỉ trên 1 yếu tố, bạn sẽ có hai cách thực hiện sau: qua diễn viên chính của phim và qua thể loại phim.



Hình 2.2. Hình ảnh minh họa sơ đồ xác định độ thành công của bộ phim

Qua sơ đồ, ta có thể thấy rõ ràng ràng, với phương pháp thứ nhất, ta phân loại được rõ ràng, trong khi phương pháp thứ hai, ta có một kết quả lộn xộn hơn. Và tương tự, cây quyết định sẽ thực hiện như trên khi thực hiện việc chọn các biến.

* Entropy**trong Cây quyết định (Decision Tree)**

Entropy là thuật ngữ thuộc Nhiệt động lực học, là thước đo của sự biến đổi, hỗn loạn hoặc ngẫu nhiên. Năm 1948, Shannon đã mở rộng khái niệm Entropy sang lĩnh vực nghiên cứu, thống kê với công thức như sau:

Với một phân phối xác suất của một biến rời rạc x có thể nhận n giá trị khác nhau x1,x2,…,xn.

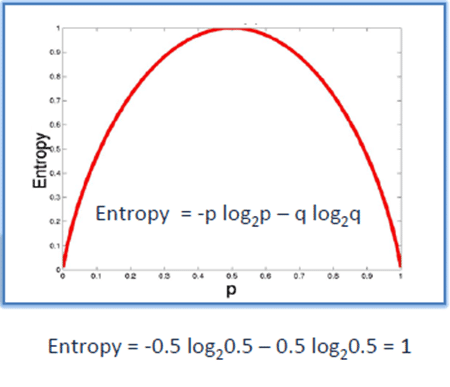
Giả sử rằng xác suất để x nhận các giá trị này là pi=p(x=xi).

Ký hiệu phân phối này là p = (p1 ,p2 ,…,pn). [Entropy](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/" \t "_blank) của phân phối này được định nghĩa là:

= -

Giả sử bạn tung một đồng xu, entropy sẽ được tính như sau:

H = -(0.5 + 0.5 )



Hình 2.3. Hình ảnh minh họa hàm Entropy

Hình vẽ trên biểu diễn sự thay đổi của hàm [entropy](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/" \t "_blank). Ta có thể thấy rằng, entropy đạt tối đa khi xác suất xảy ra của hai lớp bằng nhau.

* P tinh khiết: pi = 0 hoặc pi = 1
* P vẩn đục: pi = 0.5, khi đó hàm Entropy đạt đỉnh cao nhất.
* Information Gain **trong Cây quyết định (Decision Tree)**

Information Gain dựa trên sự giảm của hàm [Entropy](https://trituenhantao.io/tu-dien-thuat-ngu/entropy/" \t "_blank) khi tập dữ liệu được phân chia trên một thuộc tính. Để xây dựng một cây quyết định, ta phải tìm tất cả thuộc tính trả về Infomation gain cao nhất.

Để xác định các nút trong mô hình cây quyết định, ta thực hiện tính Infomation Gain tại mỗi nút theo trình tự sau:

* **Bước 1**: Tính toán hệ số Entropy của biến mục tiêu S có N phần tử với Nc phần tử thuộc lớp C cho trước:

**H(S) =** -

* **Bước 2**: Tính hàm số Entropy tại mỗi thuộc tính: với thuộc tính x, các điểm dữ liệu trong S được chia ra K child node S1, S2, …, SK với số điểm trong mỗi child node lần lượt là m1, m2 ,…, mK , ta có:

H(x,S) =

* **Bước 3**: Chỉ số Gain Information được tính bằng:

G(x,S) = H(S) – H(x,S)