**Dữ liệu dạng bảng => Goal là dữ liệu label hoặc phân lớp dc**

**Index Attr1 Attr2 … Attrn (Text, Number, Label, Range)**

**1 A B1 E**

**2 E1 B2**

**3 B3**

**… B4**

**N**

**M1 M Mn**

**(P1,P2,P3)**

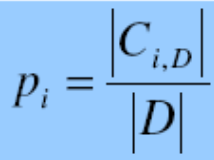
**Cây quyết định: phân lớp như nào ra sao? chọn nốt nào là gốc?**

**Trong Attr2 có phân hoạch của goal => tìm thuộc thính trung gian có phân hoạch gần giống nhất => Attr2 (độ lợi thông tin)**

**Số m phân hoạch là do ta quyết định => heuristic**

**3 nhóm C1, C2, C3 có xác xuất P1,P2,P3 => có bao nhiêu nhóm mang dữ liệu C1 và chia cho tập dữ liệu**

**Dữ liệu 1 cột Attr Goal: D => Chia D thành m nhóm con Ci (I chạy từ 1 đến m)**

**xác xuất mỗi nhóm con xuất hiện**

A mathematical equation with a blue background

Description automatically generated**Entropy**

**Log là scaling function**

**Chọn Attr trung gian chia thược tính trên đó, độ hỗn loạn của dữ liệu gốc trừ độ hỗn loạn của trung gian (trung gian càng thuần nhất càng tiến đến 0 => tính phân hóa dễ nhất do thuần nhất có thể)**

**=> Gain lớn nhất đặt lên đầu**

**=> có bao nhiêu giá trị => bấy nhiêu nhánh => over spilt => Dẫn đến Gain Ratio (hạn chế cột nào chia quá nhiều nhánh)**