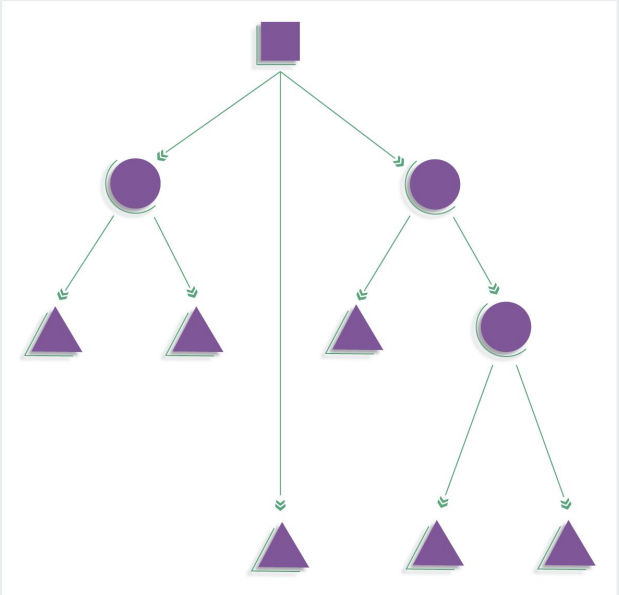
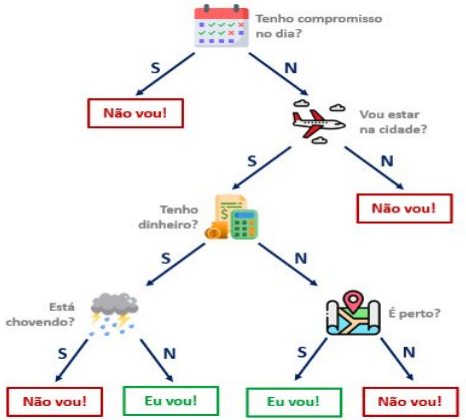
Árvores de Decisão

Vitor Oliveira e Gabriel Silva



Introdução

- Baseia-se na ideia de Árvores Binárias;
- Segue a ideia de Cima para Baixo, começando do primeiro nó (Raiz) e seguindo até o final da determinada ramificação (Folhas);
- Um nó folha indica uma classe;
- um nó de decisão contém um teste sobre o valor de um atributo, geralmente um teste com decisão de verdadeiro ou falso.

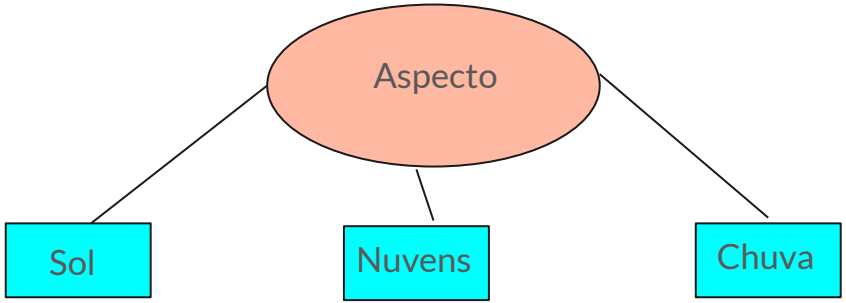


Algoritmo de C4.5

- Desenvolvido em 1993, sucesso do ID3 (*Interative Dichotomiser 3)*;
- Objetivo é a construção de uma árvore de decisão a partir de um conjunto de dados de treinamento;
- Utiliza da técnica de divisão e conquista, juntamente com o método guloso (decide um caminho e segue até o final daquele caminho, sem retornar)
- A seleção dos atributos é decidida através do *Ganho de informação (info gain), que será explorado a seguir.*

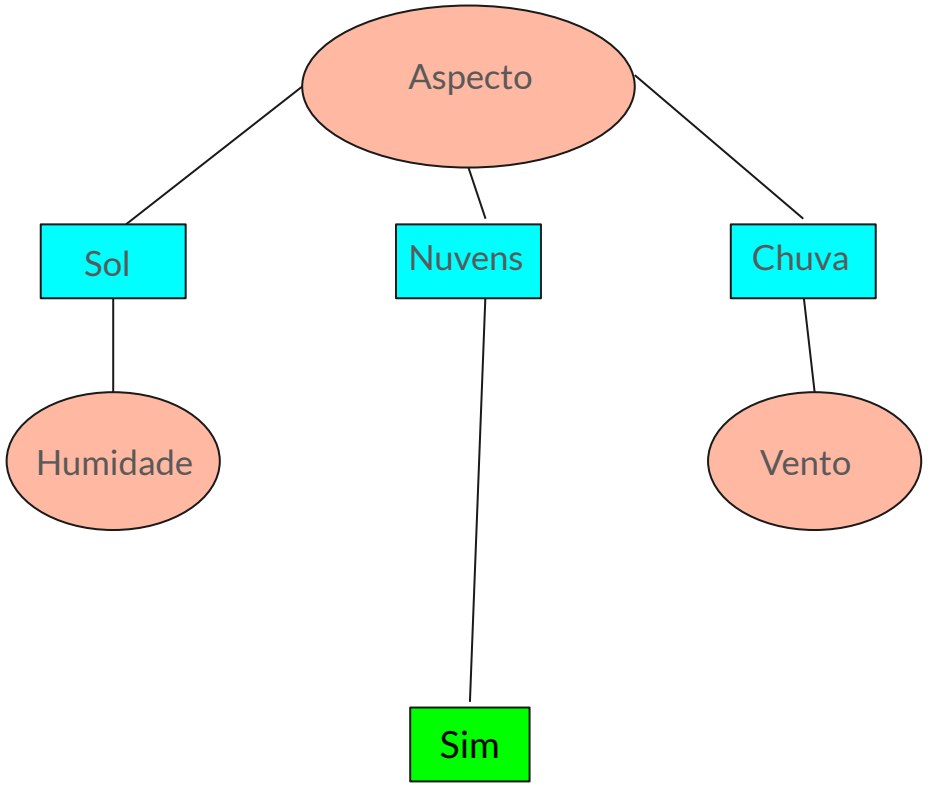
Indução de uma Árvore

Dia	Aspecto	Temp.	Humidade	Vento	Jogar Ténis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não



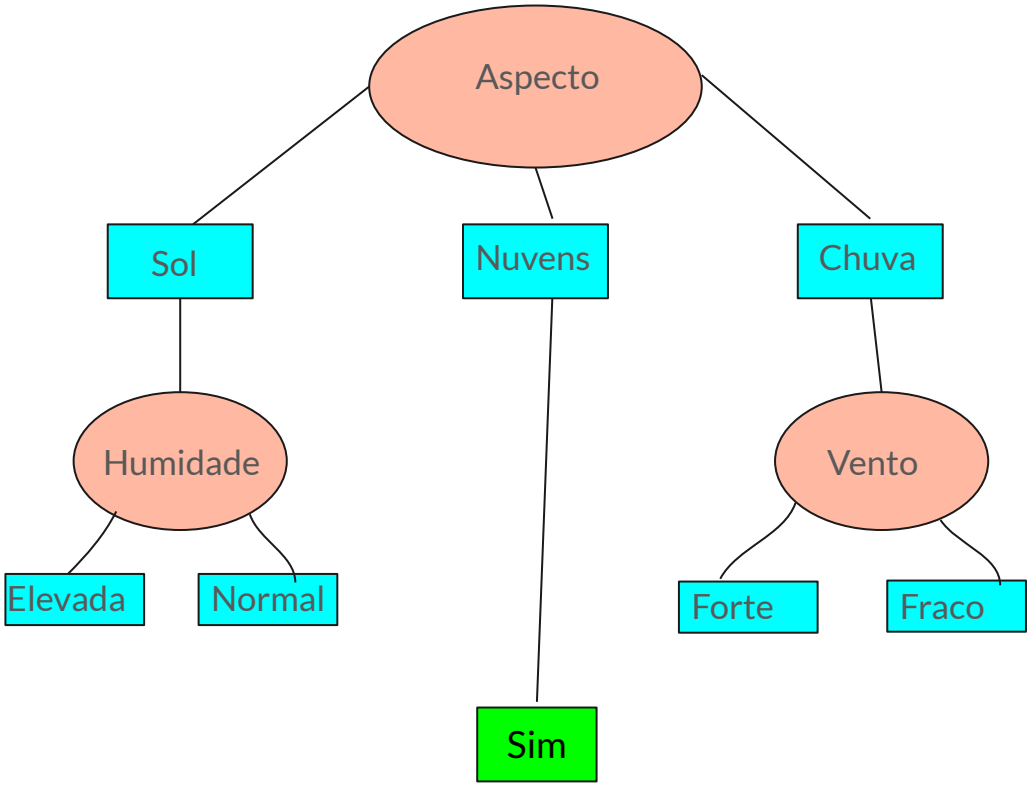
Indução de uma Árvore

Dia	Aspecto	Temp.	Humidade	Vento	Jogar Ténis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não



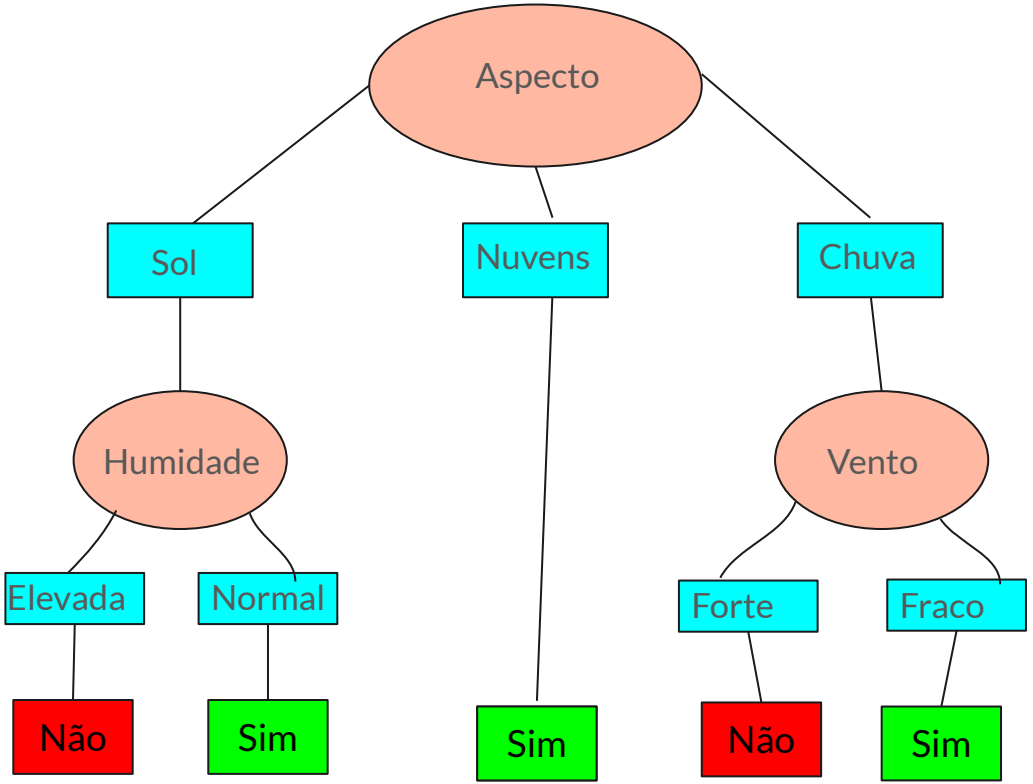
Indução de uma Árvore

Dia	Aspecto	Temp.	Humidade	Vento	Jogar Ténis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não



Indução de uma Árvore

Dia	Aspecto	Temp.	Humidade	Vento	Jogar Ténis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não



Critério de Entropia

- Mede o Grau de pureza de um conjunto
- Calcula a “falta de informação” dentro de um conjunto, ou nodo.

Dado um conjunto S, com instâncias pertencentes a classe *i*, com probabilidade p_i, temos:

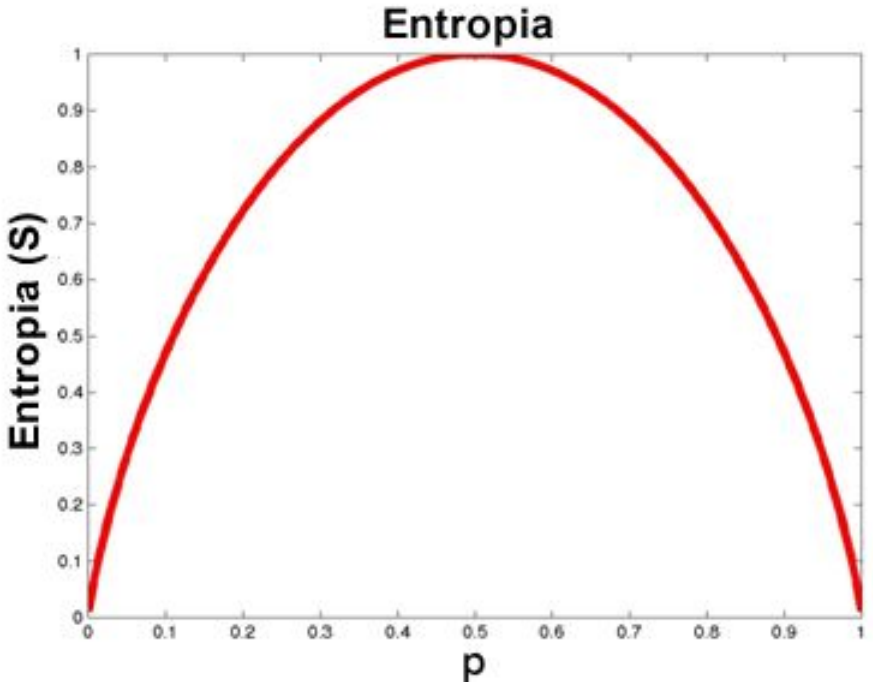
$$Entropia(S) = \sum p_i \log_2 p_i$$

- Na ideia de que as Árvores de Decisão tem resultados binários, essa equação pode ser alterada para a seguinte:

$$Entropia(S) = -p_+ \log_2 p_+ - p_- \log_2 p_-$$

Em que p- é a porção dos resultados negativos e p+ é a porção dos resultados positivos.

Critério de Entropia: Relação entre Entropia do Conjunto (S) com a probabilidade



Critério de Entropia: Ganho

- É a redução esperada da Entropia em relação á algum atributo;
- Ganho(S, A) é a redução esperada da entropia do conjunto S, através do atributo A.
- É utilizado para definir qual atributo será utilizado, tanto na raiz da árvore, quanto em seus nodos.

$$Ganho(S, A) = Entropia(S) - \sum_{v \in valores(A)} \frac{|S_v|}{|S|} \cdot Entropia(S_v)$$

Critério de Entropia: Exemplo

Dia	Aspecto	Temp.	Humidade	Vento	Jogar Ténis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não

S = [9+, 5-]

Entropia(S) = - (9/14) * log2 * (9/14) - (5/14) * log2 * (5/14)= 0.940

Humidade? -> Elevada = [3+,4-]; Normal [6+,1-]

Entropia(Elevada) = 0.985 e Entropia(Normal) = 0.592

Ganho(S,Humidade) = 0.940 - (7/14) * 0.985 - (7/14) * 0.592 = 0.151

Ganho(S,Humidade) = 0.151

Critério de Entropia: Exemplo

Dia	Aspecto	Temp.	Humidade	Vento	Jogar Ténis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não

Calculando o Ganho para cada atributo temos:

Ganho(S,Aspecto) = 0.247

Ganho(S,Humidade) = 0.151

Ganho(S,Temperatura) = 0.029

Ganho(S,Vento) = 0.048

Nestes valores, o que teve maior ganho foi o Aspecto, logo ele é o melhor escolhido para ser a raiz da árvore.

Para os próximos nodos , a decisão do atributo não terá a opção de escolher o Aspecto novamente até que todos os nodos tenham entropia nula.

Índice de Gini

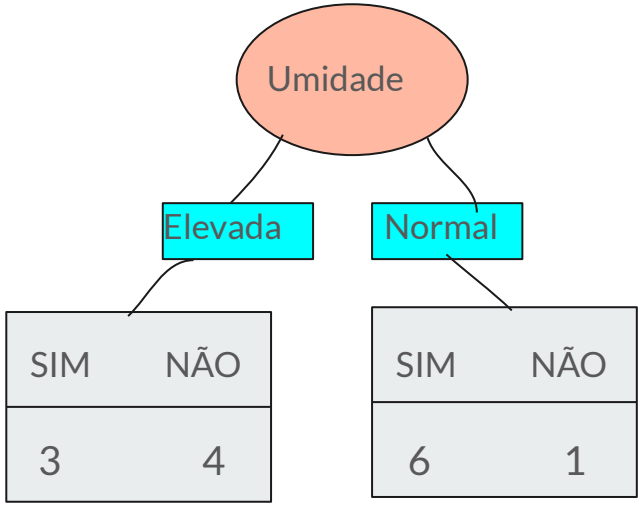
- Mede o grau de impureza dentro de um nodo;
- Impureza Máxima: classes têm a mesma distribuição, Impureza Mínima: apenas uma das classes existe no nodo.
- A Equação para o cálculo desse índice é a seguinte:

$$\text{Índice Gini} = 1 - \sum_{i=1}^{c} p_i^2$$

Em que c é a quantidade de classes e p_i é a frequência relativa a cada classe dentro do nodo.

Índice de Gini: Exemplo

Dia	Aspecto	Temp.	Humidade	Vento	Jogar Ténis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não



Para a categoria elevada, o seu resultado contém 3+ e 4-, nesse caso, o índice de Gini:
$1 - ((3 / 7)^2 + (4 / 7)^2) = 1 - 0.428^2 + 0.571^2 = 1 - 0,183 + 0.326 = 0.491$

Índice Gini(Elevada) = 0.491
Índice GIni(Normal) = 0. 245

Cálculo Gini total = (3+4/ 14)* 0.491 + (6+1/14) * 0.245 = 0.2455 + 0.1225 = 0.368

Índice Gini para Umidade = 0.368

Índice de Gini: Exemplo

Dia	Aspecto	Temp.	Humidade	Vento	Jogar Ténis
D1	Sol	Quente	Elevada	Fraco	Não
D2	Sol	Quente	Elevada	Forte	Não
D3	Nuvens	Quente	Elevada	Fraco	Sim
D4	Chuva	Ameno	Elevada	Fraco	Sim
D5	Chuva	Fresco	Normal	Fraco	Sim
D6	Chuva	Fresco	Normal	Forte	Não
D7	Nuvens	Fresco	Normal	Fraco	Sim
D8	Sol	Ameno	Elevada	Fraco	Não
D9	Sol	Fresco	Normal	Fraco	Sim
D10	Chuva	Ameno	Normal	Forte	Sim
D11	Sol	Ameno	Normal	Forte	Sim
D12	Nuvens	Ameno	Elevada	Forte	Sim
D13	Nuvens	Quente	Normal	Fraco	Sim
D14	Chuva	Ameno	Elevada	Forte	Não

Seguindo o mesmo raciocínio para as outras classes temos que:

Gini(Umidade) = 0.368

Gini(Aspecto) = 0.342

Gini(Vento) = 0.428

Gini(Temperatura) = 0.439

Com esses resultados a melhor classe para a raiz da árvore seria a com menor índice Gini, no caso, a classe de Aspecto.

Demonstração em Código