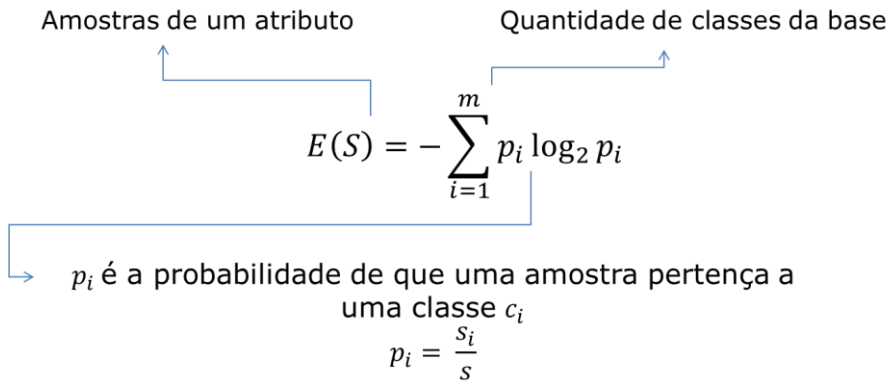
IMPACTA

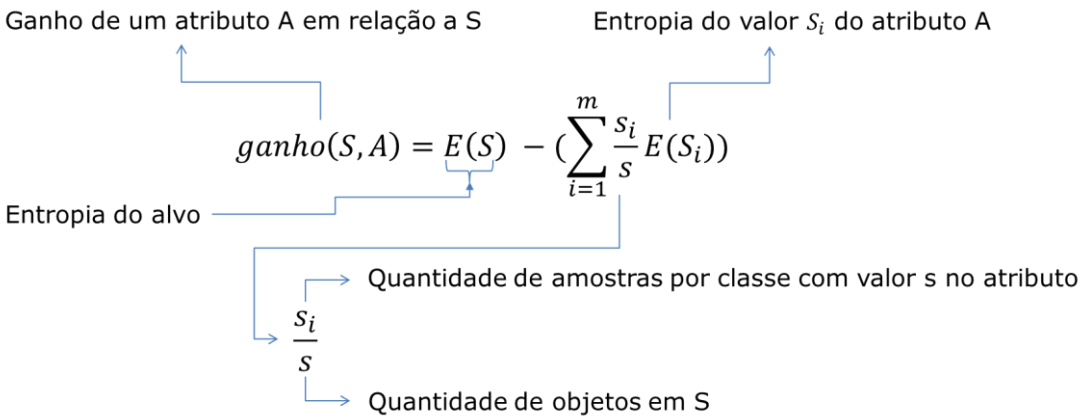
CÁLCULO DA ENTROPIA

Em nossa aula sobre árvore de decisão, usamos a seguinte base de dados como exemplo para construção da árvore:

Aspecto	Temperatura	Umidade	Vento	Jogar
Sol	Quente	Alta	Não	Não
Sol	Quente	Alta	Sim	Não
Nublado	Quente	Alta	Não	Sim
Chuva	Branda	Alta	Não	Sim
Chuva	Fria	Normal	Não	Sim
Chuva	Fria	Normal	Sim	Não
Nublado	Fria	Normal	Sim	Sim
Sol	Branda	Alta	Não	Não
Sol	Fria	Normal	Não	Sim
Chuva	Branda	Normal	Não	Sim
Sol	Branda	Normal	Sim	Sim
Nublado	Branda	Alta	Sim	Sim
Nublado	Quente	Normal	Não	Sim
Chuva	Branda	Alta	Sim	Não

O cálculo da entropia é obtido da seguinte forma:





Vale lembrar que os termos amostra, objeto ou ainda instância são equivalentes para se referir a uma linha de uma base de dados.

Então, a entropia da nossa resposta esperada (alvo) é:

$S = 14\ amostras$

$S_{sim} = 9$

$S_{não} = 5$

$E(9,5) = -(\frac{9}{14}\log_2\frac{9}{14} + \frac{5}{14}\log_2\frac{5}{14})$

$E(9,5) = 0.940\ bits$

Aspecto (Sol, Nublado, Ensolarado)

Sol ocorre 5 vezes na base em que 2 vezes a reposta esperada é **jogar** e 3 **não jogar**:

$Sol\ (S_{sim} = 2, S_{não} = 3, S = 5)$

Nublado ocorre 4 vezes em todas as ocorrências a resposta é **não jogar**:

$Nublado\ (S_{sim} = 4, S_{não} = 0, S = 4)$

Chuvoso ocorre 5 vezes na base em que 3 vezes a reposta é **jogar** e 2 **não jogar**:

$Chuva\ (S_{sim} = 3, S_{não} = 2, S = 5)$

Agora, calculamos a entropia de cada um dos valores do aspecto:

$$E(S) = -\sum_{i=1}^{m} p_i \log_2 p_i$$

$$Sol = -(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}) = 0.971$$

$$Nublado = -(\frac{4}{4}\log_2\frac{4}{4} + \frac{0}{4}\log_2\frac{0}{4}) = 0$$

$$Chuva = -(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}) = 0.971$$

Temos, então, que o ganho de informação fornecido pelo atributo **Aspecto** é:

$$ganho(S,\ Aspecto) = 0.940 - (0.971 * \frac{5}{14} + 0 * \frac{4}{14} + 0.971 * \frac{5}{14}) = 0.246$$

Fazendo o cálculo para os demais atributos obtemos:

$$ganho(S,\ Temperatura) = 0.940 - (1 * \frac{4}{14} + 0{,}918 * \frac{6}{14} + 0.811 * \frac{4}{14}) = 0.029$$

$$ganho(S,\ Umidade) = 0.940 + (0.985 * \frac{7}{14} - 0.592 * \frac{7}{14}) = 0.152$$

$$ganho(S,\ Vento) = 0.940 + (0.811 * \frac{8}{14} - 1 * \frac{6}{14}) = 0.048$$

Nosso objetivo é encontrar o atributo com maior ganho de informação, por isso o escolhido é o atributo **Aspecto**.