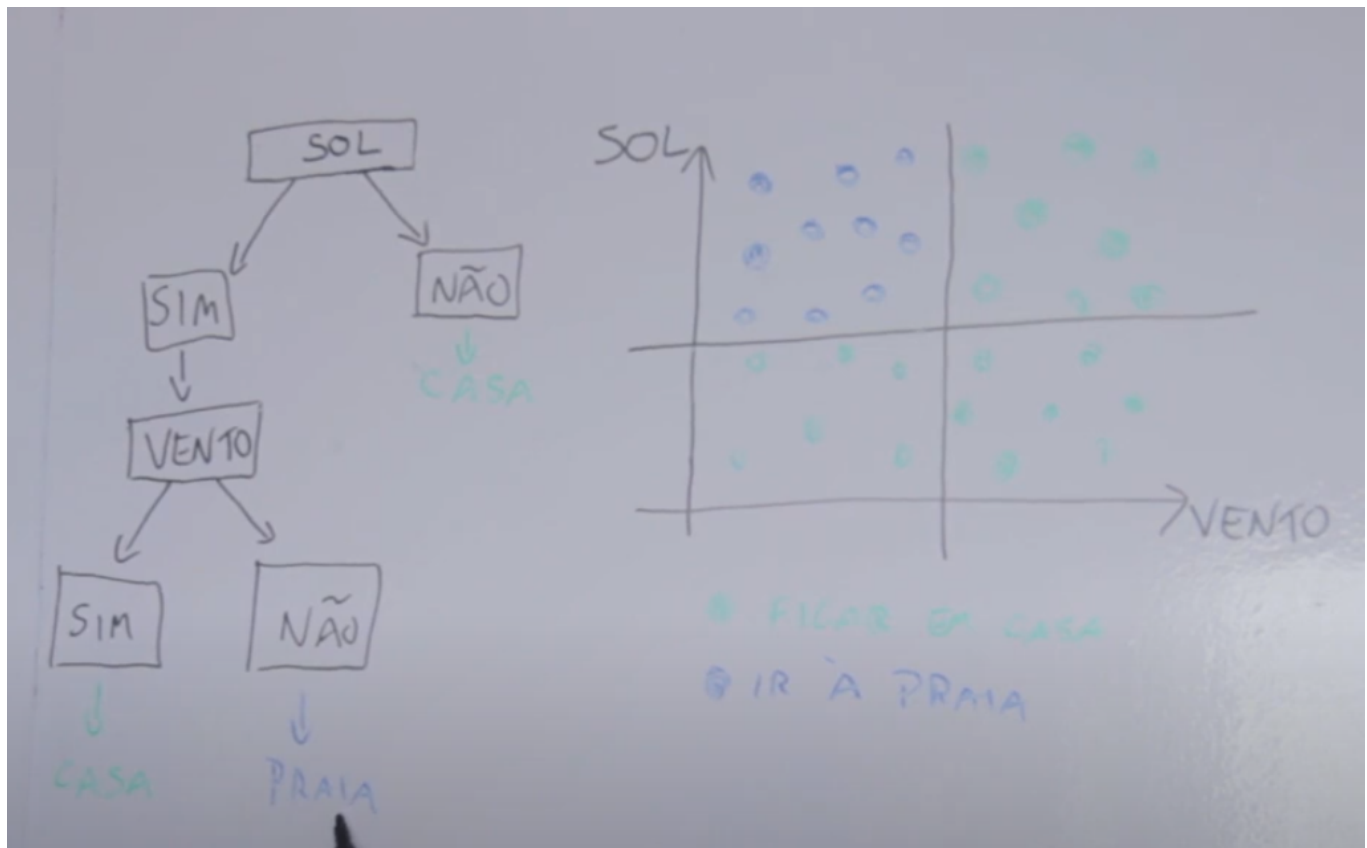


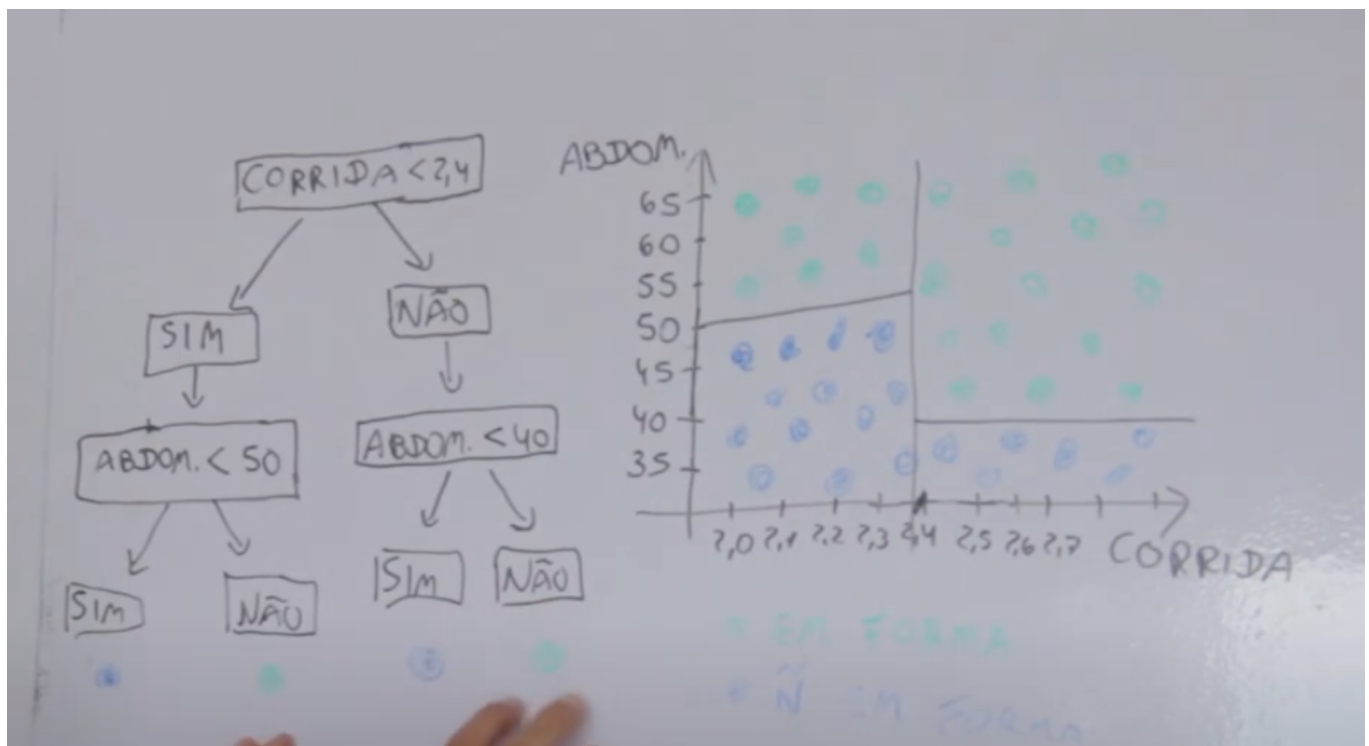
# Entropia para classificação

Exemplo simples de uma árvore de decisão:



O problema acima é uma árvore de decisão baseada em momentos que as pessoas vão para a praia ou ficam em casa. Dividindo os pontos em 4 quadrantes, podemos ver que quando tem muito sol, e pouco vento, as pessoas tendem mais a ir para a praia. Podemos então montar uma árvore de decisão para esse problema. Se não há sol, normalmente ficam em casa, se há sol, temos que olhar a variável vento. Caso tenha muito vento as pessoas ficam em casa, e caso não tenha vão à praia. A árvore é dividida em nós e raízes, os nós são as variáveis e as raízes o resultado final encontrado.

Outro exemplo para árvore de decisão utilizando números como condicionais para os nós, seria o caso abaixo:



Podemos olhar os km que uma pessoa corre em 12 minutos e a quantidade de abdominais para definir se ela está ou não em forma. É possível observar que com 3 retas foi possível dividir os dados de forma que consigamos prever as classes com facilidade apenas caminhando nas condicionais da árvore.

## Ganho de informação

$$\text{GANHO INFORMAÇÃO} = \text{ENTROPIA}_{\text{PAI}} - \sum \text{PESO}_{\text{FILHO}} \cdot \text{ENTROPIA}_{\text{FILHO}}$$

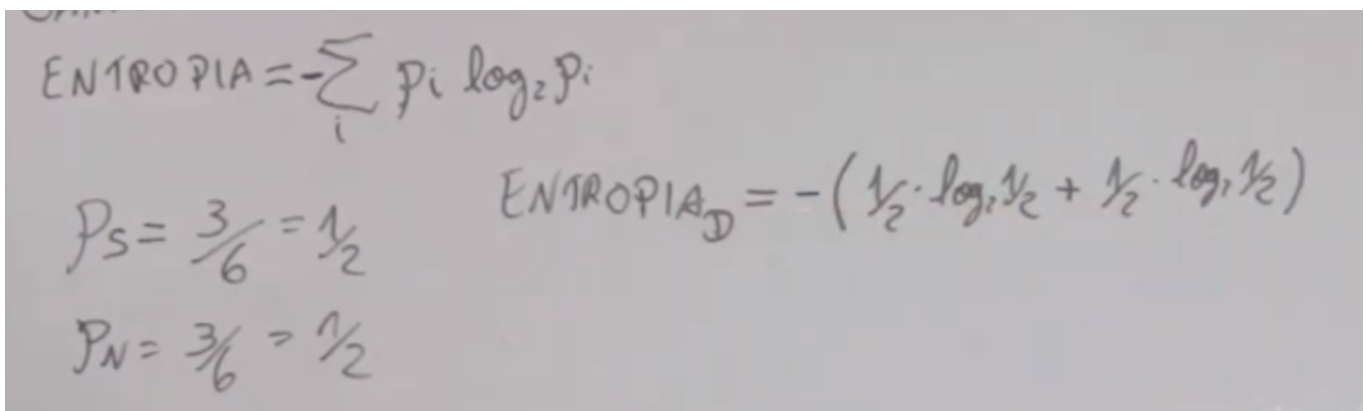
## Entropia

$$\text{ENTROPIA} = \sum_i p_i \log_2 p_i$$

## Utilizando as fórmulas na prática:

Dataset abaixo se deve a decisão final de aceitar o emprego ou não, baseado nas features:

Salário	Localização	Função	Decisão
alto	longe	interessante	SIM
baixo	perto	desinteressante	NÃO
baixo	longe	interessante	SIM
alto	longe	desinteressante	NÃO
alto	perto	interessante	SIM
baixo	longe	desinteressante	NÃO



$$ENTROPIA = -\sum_i p_i \log_2 p_i$$

$$P_S = \frac{3}{6} = \frac{1}{2}$$

$$P_N = \frac{3}{6} = \frac{1}{2}$$

$$ENTROPIA_D = -\left(\frac{1}{2} \cdot \log_2 \frac{1}{2} + \frac{1}{2} \cdot \log_2 \frac{1}{2}\right)$$

Podemos calcular a entropia primeiro encontrando as probabilidades de cada decisão do dataset, no caso de Sim ou Não.

$$P_{sim} = 3 / 6 = 1 / 2$$

$$P_{não} = 3 / 6 = 1 / 2$$

Agora basta inserir na fórmula do somatório que resulta na entropia:

$$Entropia = - (1 / 2 \log_2 1/2 + 1 / 2 \log_2 1/2 )$$

$$Entropia = - (1 / 2 (-1) + 1 / 2 (-1))$$

$$Entropia = 1$$

Outro exemplo, considerando que temos 5 decisões Sim e apenas 1 sendo não:

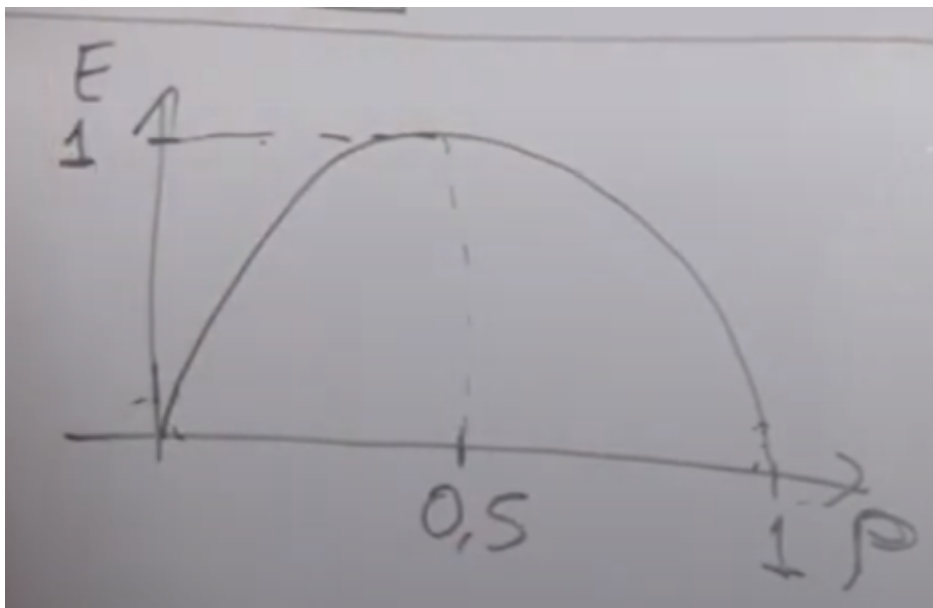
$$p_s = \frac{5}{6}$$

$$p_r = \frac{1}{6}$$

$$E = - \left( \frac{5}{6} \cdot \log_2 \frac{5}{6} + \frac{1}{6} \cdot \log_2 \frac{1}{6} \right)$$

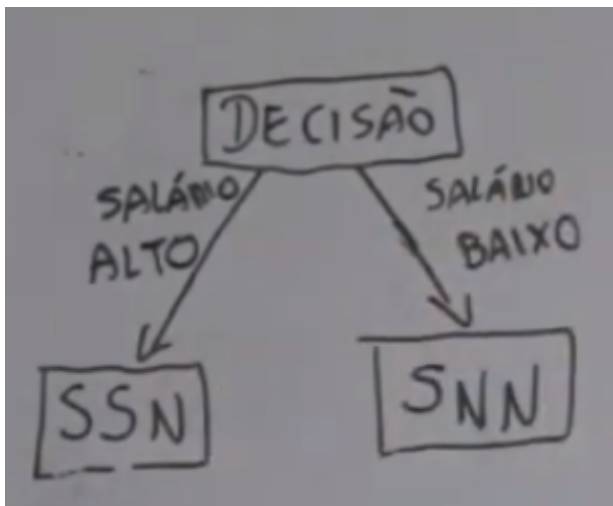
$$\underline{E = 0,65}$$

É importante salientar que quanto mais balanceado estiver os dados de decisão, maior a entropia, chegando no ponto 1. Caso os dados estejam desbalanceados a entropia tende a 0. Quanto mais desordenação nos dados (balanceamento), ou seja, difícil de dividir os dados, maior será a entropia, e quanto mais desbalanceado, ou seja, ordenados, a entropia é menor. Podemos ver isso no gráfico abaixo:



## Calculando o ganho de informação para a variável salário

Observando somente a variável salário, temos o seguinte:



Se o salário for Alto, temos 2 decisões sim e uma não, e para o salário baixo, apenas uma Sim e 2 não.

A partir disso devemos voltar para a fórmula do ganho de informação:

$$\text{GANHO INFORMAÇÃO} = \text{ENTROPIA}_{\text{PAI}} - \sum \text{PESO}_{\text{FILHO}} \cdot \text{ENTROPIA}_{\text{FILHO}}$$

Devemos calcular a entropia do pai, que seria a raiz da árvore, no caso a variável salário, e também a entropia dos filhos (Salário Alto e Baixo).

Portanto utilizando a fórmula de entropia nos dois casos:

RAMO SAL. ALTO:

$$P_S = \frac{2}{3}$$

$$P_N = \frac{1}{3}$$

$$E = - \left( \frac{2}{3} \log_2 \frac{2}{3} + \frac{1}{3} \log_2 \frac{1}{3} \right)$$

$$E = 0,92$$

RAMO SAL. BAIXO:

$$P_S = \frac{1}{3}$$

$$P_N = \frac{2}{3}$$

$$E = 0,92$$

Com a entropia dos filhos calculada, basta calcular o peso de cada um dos filhos com a seguinte fórmula:

$$\text{PESO} = \frac{N^{\circ} \text{ AMOSERAS FILHO}}{N^{\circ} \text{ AMOSERAS PAI}}$$

$$\text{Peso1} = 3 / 6 = 1 / 2$$

$$\text{Peso2} = 3 / 6 = 1 / 2$$

Portanto o ganho de informação da variável Salário é:

$$GI = 1 - \left( \frac{1}{2} \cdot 0,92 + \frac{1}{2} \cdot 0,92 \right)$$

$$GI = 0,08$$

Após isso temos que calcular o ganho de informação de todas as outras variáveis, e a qual tiver mais ganho, deverá ser escolhida para ser a raiz da árvore que irá separar os dados iniciais.

## Calcular ganho de informação para Localização:

$$\text{GANHO INFORMAÇÃO} = \text{ENTROPIA}_{\text{PAI}} - \sum \text{PESO}_{\text{FILHO}} \cdot \text{ENTROPIA}_{\text{FILHO}}$$

$$\text{ENTROPIA} = - \sum p_i \log_2 p_i \quad \left| \quad \text{PESO} = \frac{N^{\circ} \text{ AMOSTRAS FILHO}}{N^{\circ} \text{ AMOSTRAS PAI}} \right.$$

```

graph TD
    A[DECISÃO] -- LONGE --> B[SSNN]
    A -- PERTO --> C[SN]
        
```

RAMO LONGE:

$P_S = \frac{2}{4} = \frac{1}{2}$        $\text{PESO} = \frac{4}{6}$   
 $P_N = \frac{2}{4} = \frac{1}{2}$   
 $E = 1$

RAMO PERTO:

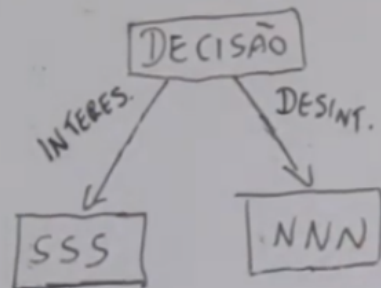
$P_S = \frac{1}{2}$        $\text{PESO} = \frac{2}{6}$   
 $P_N = \frac{1}{2}$   
 $E = 1$

$GI = 1 - \left( \frac{4}{6} \cdot 1 + \frac{2}{6} \cdot 1 \right)$   
 $GI = 0$

## Calcular ganho de informação para Função:

$$\text{GANHO INFORMAÇÃO} = \text{ENTROPIA}_{\text{PAI}} - \sum \text{PESO}_{\text{FILHO}} \cdot \text{ENTROPIA}_{\text{FILHO}}$$

$$\text{ENTROPIA} = - \sum_i p_i \log_2 p_i \quad \left| \quad \text{PESO} = \frac{N^{\circ} \text{ AMOSTRAS FILHO}}{N^{\circ} \text{ AMOSTRAS PAI}} \right.$$



$$GI = 1$$

RAMO INTERESSANTE :

$$E = 0$$

RAMO DESINTERESSANTE :

$$E = 0$$

Interessante observar que como não tem divisão de classes entre Interessado e Desinteressado, a entropia acaba sendo 0 e zerando o somatório dos filhos, o que garante que há 100% de ganho de informação. Portanto essa variável será a escolhida para ser a raiz da árvore por ter o maior ganho.

Outra observação é que só com essa variável já é possível definir se a decisão será sim ou não, pois ela divide igualmente as decisões de sim e não, ou seja, a árvore de decisão já está pronta.

Em um outro exemplo, onde as decisões não são 100% divididas já nos primeiros nós, a lógica é continuar executando o cálculo de entropia do pai e filho, porém os novos nós são considerados os pais e os filhos serão calculados posteriormente, recursivamente até um ponto de parada da árvore.