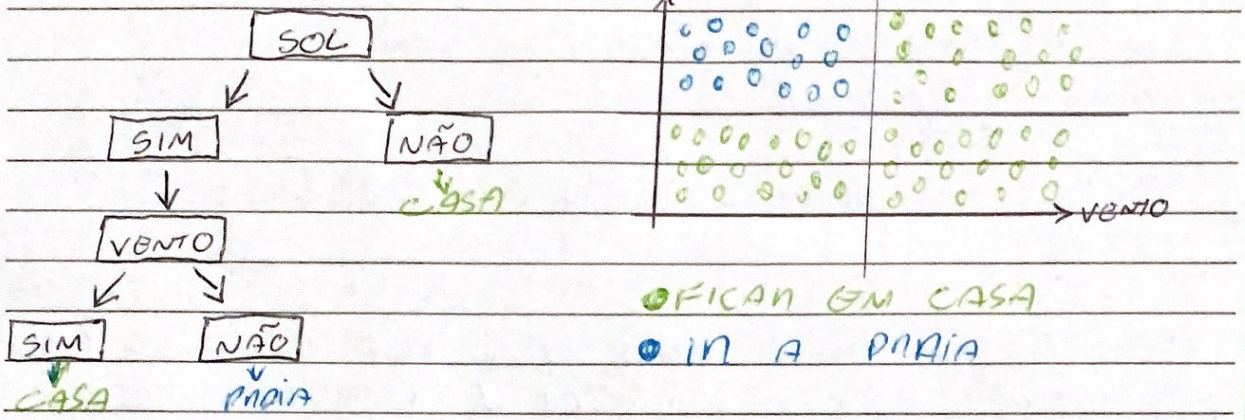
S T Q Q S S D

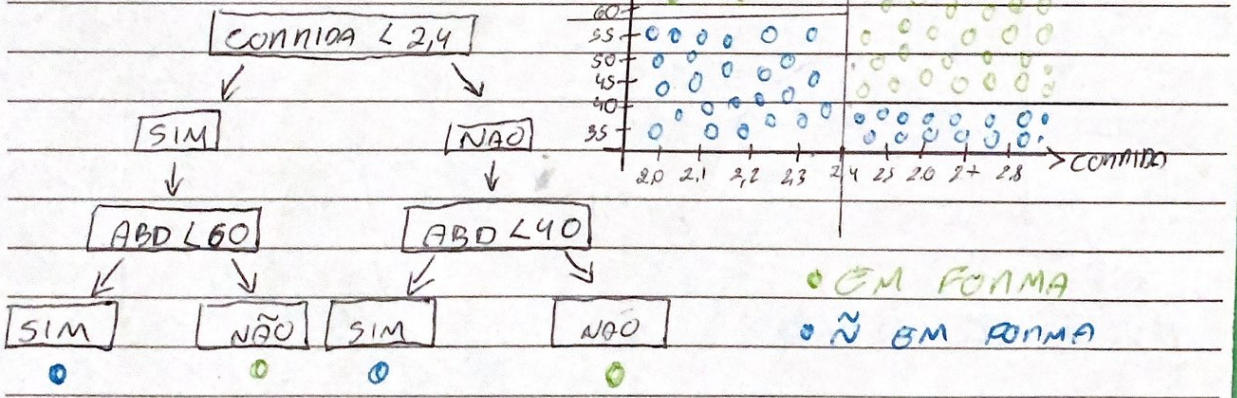
__/__/__

DECISION TREES - PARTE 1: ENTROPIA

EXEMPLO 1:



EXEMPLO 2:



GANHO INFORMAÇÃO = $ENTROPIA_{PAI} - \sum PESO_{FILHO} \cdot ENTROPIA_{FILHO}$

$ENTROPIA = -\sum_{i} P_i \log_2 P_i$, $PESO = \frac{Nº\ AMOSTRAS\ FILHO}{Nº\ AMOSTRAS\ PAI}$

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

spiral

__/__/__

S T Q Q S S D

$P_S = 3/6 = 1/2$ ENTROPIA$_D = -(1/2 \log_2 1/2 + 1/2 \log_2 1/2)$

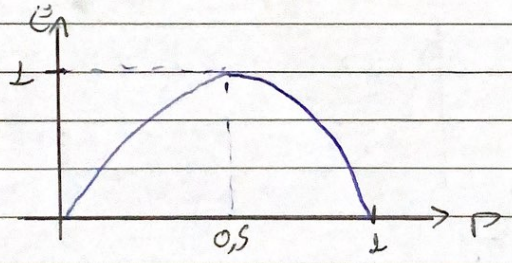
$P_N = 3/6 = 1/2$ $= 1$

CASO A PROPORÇÃO DE SIM PARA NÃO FOSSE DE 5 PARA 1, TERÍAMOS:

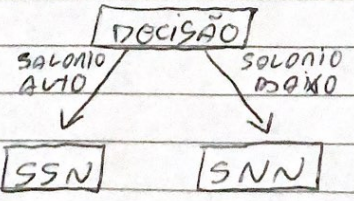
$P_S = 5/6$ ENTROPIA$_D = -(5/6 \log_2 5/6 + 1/6 \log_2 1/6)$

$P_N = 1/6$ $= 0,65$

OBS. QUANTO MAIS OS DADOS ESTÃO BEM DIVIDIDOS (BEM NO MEIO), MAIS PRÓXIMO DE 1 SERÁ A ENTROPIA, ENQUANTO QUE MAIS DISPERSOS OU DESBALANCEADOS, MAIS PRÓXIMO DE ZERO SERÁ A ENTROPIA



#VARIÁVEL SALÁRIO



▷ RAMO/FILHO SALÁRIO ALTO:

- $P_S = 2/3$
- $P_N = 1/3$
- $E = -(2/3 \log_2 2/3 + 1/3 \log_2 1/3)$ $= 0,92$
- PESO $= 3/6 = 1/2$

$GI = 1 - (1/2 \cdot 0,92 + 1/2 \cdot 0,92)$ $= 0,08$

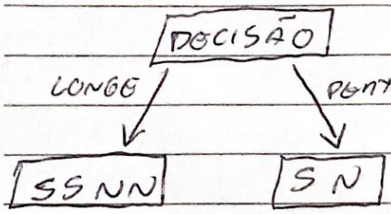
▷ RAMO/FILHO SALÁRIO BAIXO

- $P_S = 1/3$
- $P_N = 2/3$
- $E = 0,92$
- PESO $= 3/6 = 1/2$

S T Q Q S S D

__/__/__

VARIÁVEL LOCALIZAÇÃO

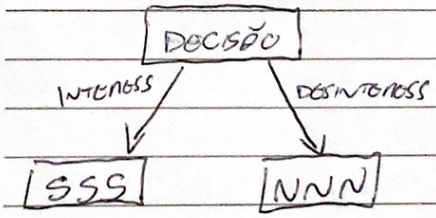


▷ RAMO LONGE
- $P_S = 2/4 = 1/2$
- $P_{CSO} = 4/6$
- $P_N = 2/4 = 1/2$
- $G = 1$

▷ RAMO PERTO
- $P_S = 1/2$
- $P_{OSO} = 2/6$
- $P_N = 1/2$
- $G = 1$

$$GI = 1 - (4/6 \cdot 1 + 2/6 \cdot 1)$$
$$= 0$$

VARIÁVEL FUNÇÃO



▷ RAMO INTERESSANTE
- $G = 0$

▷ RAMO DESINTERESSANTE
- $G = 0$

$$GI = 1$$

OBS. A VARIÁVEL QUE TIVER O MAIOR GANHO DE INFORMAÇÃO SERÁ A PRIMEIRA VARIÁVEL A SER UTILIZADA PARA A CRIAÇÃO DA ÁRVORE DE DECISÃO.

OBS. NO CASO ESPECÍFICO DA VARIÁVEL FUNÇÃO, COMO O GI FOI 1, ESSA ÁRVORE JÁ SERIA NOSSA ÁRVORE FINAL

OBS. DEPOIS DE ESCOLHER A VARIÁVEL COM MAIOR GI, A GENTE CONTINUA CONSTRUINDO NOSSA ÁRVORE DE DECISÃO COMO FIZEMOS ATÉ AQUI, PODEM CONSIDERANDO ESSES NÓS DE PERGUNTA COMO SENDO OS PRÓXIMOS PONTOS DE PARTIDA PARA CALCULAR O RESTANTE DA ÁRVORE

spiral