S T Q Q S S D

__/__/__

Decision Trees - Parte 2: Índice GINI

$$GINI = 1 - \sum_{i=1}^{C} Pi^2$$

VarA	VarB	Classe
1	13	1
0	-2	1
0	27	1
1	9	1
0	67	0
0	45	0
1	21	0
0	50	0

▷ Variável A

VarA = 1:

Classe 1: 2/3

Classe 0: 1/3

$GINI = 1 - ((2/3)^2 + (1/3)^2)$

$= 0,44$

VarA = 0:

Classe 1: 2/5

Classe 0: 3/5

$GINI = 1 - ((2/5)^2 + (3/5)^2)$

$= 0,48$

$$GINI_{VarA} = \frac{3}{8} \cdot 0,44 + \frac{5}{8} \cdot 0,48 = 0,46$$

▷ Variável B

VarB < 27:

Classe 1: 3/4

Classe 0: 1/4

$GINI = 1 - ((3/4)^2 + (1/4)^2)$

$= 0,37$

spiral

__/__/__

S T Q Q S S D

#VonB >=27:

CLASSE 1: 1/4 GINI = 0,37

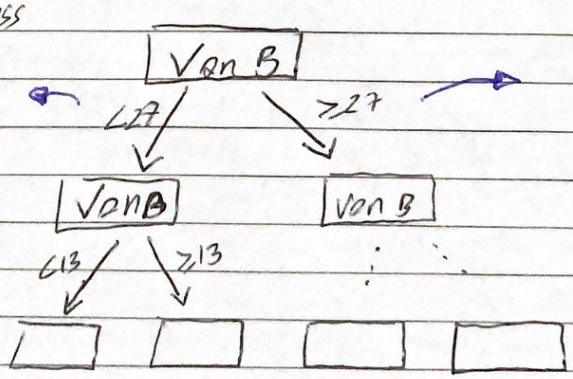
CLASSE 0: 3/4

$$GINI_{VonB} = \frac{4}{8} \cdot 0,37 + \frac{4}{8} \cdot 0,37 = 0,37$$

OBS. PORA COMEÇAR A CRIAR NOSSA ÁRVORE DE DECISÃO DEVEMOS ESCOLHER A VARIÁVEL COM O MENOR INDICE GINI

OBS. EM UM CASO HIPOTÉTICO EM QUE OS DADOS ESTÃO BEM DIVIDIDOS (METADE E METADE), O INDICE GINI É ALTO. QUANTO MAIS DIVIDIDO OS DADOS ESTÃO, MAIOR O INDICE GINI, ENQUANTO QUE QUANTO MAIS SEPARADOS OS DADOS ESTIVEREM, MENOR O INDICE GINI, SENDO QUE ESTE ÚLTIMO CASO É O CASO QUE MAIS VAI EXPLICAR OS DADOS.

VarA	VarB	Class
1	13	1
0	-2	1
1	9	1
1	21	0



VarA	VarB	Class
0	27	1
0	67	0
0	45	0
0	50	

spiral