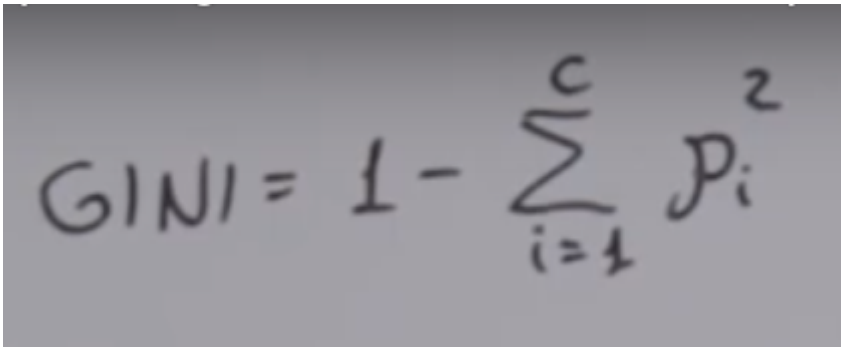
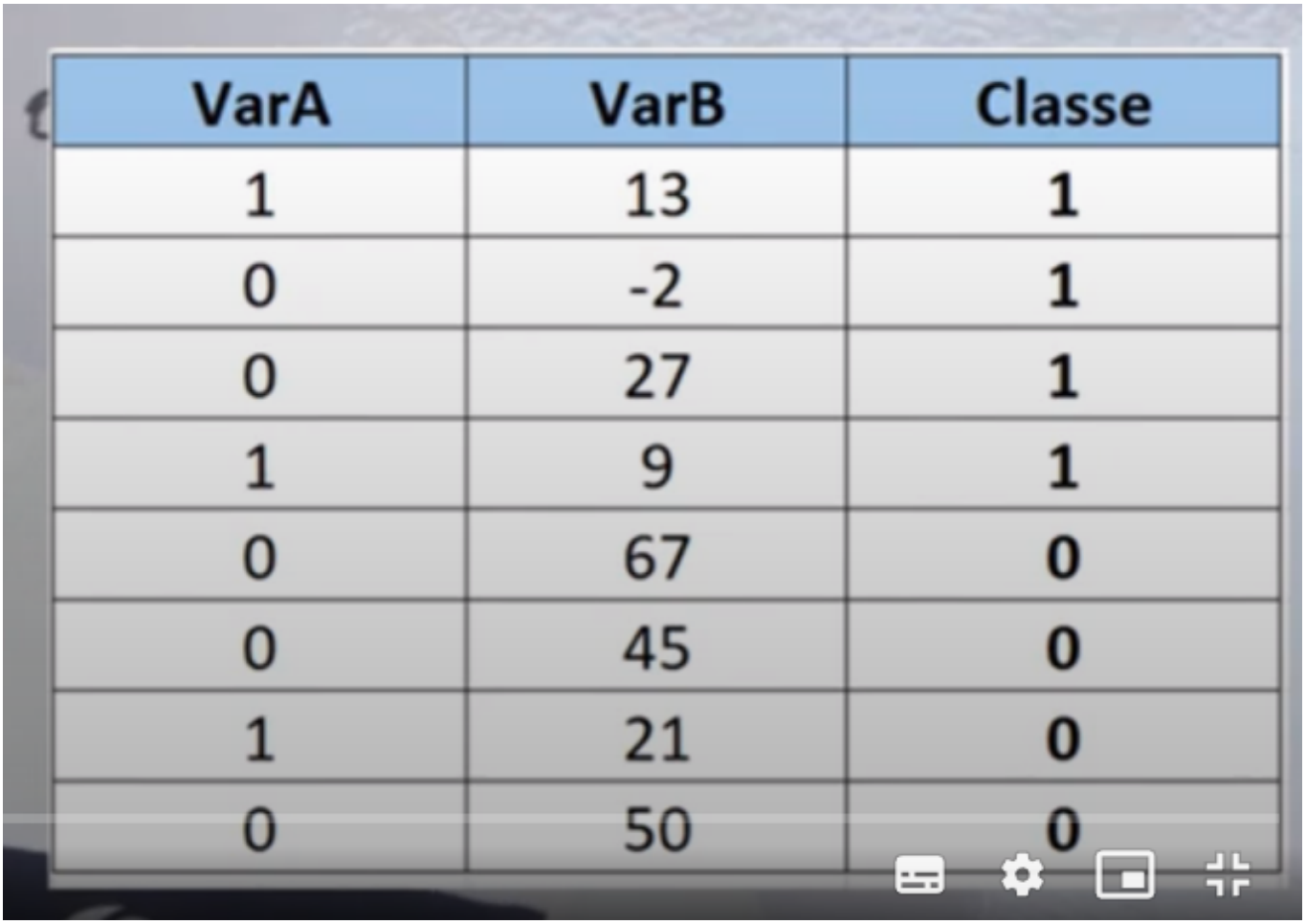
Índice Gini para classificação

Assim como a entropia, podemos utilizar o índice de Gini para criar uma árvore de decisão para classificação. Veremos abaixo um exemplo de como criar a árvore:

Fórmula de Gini:

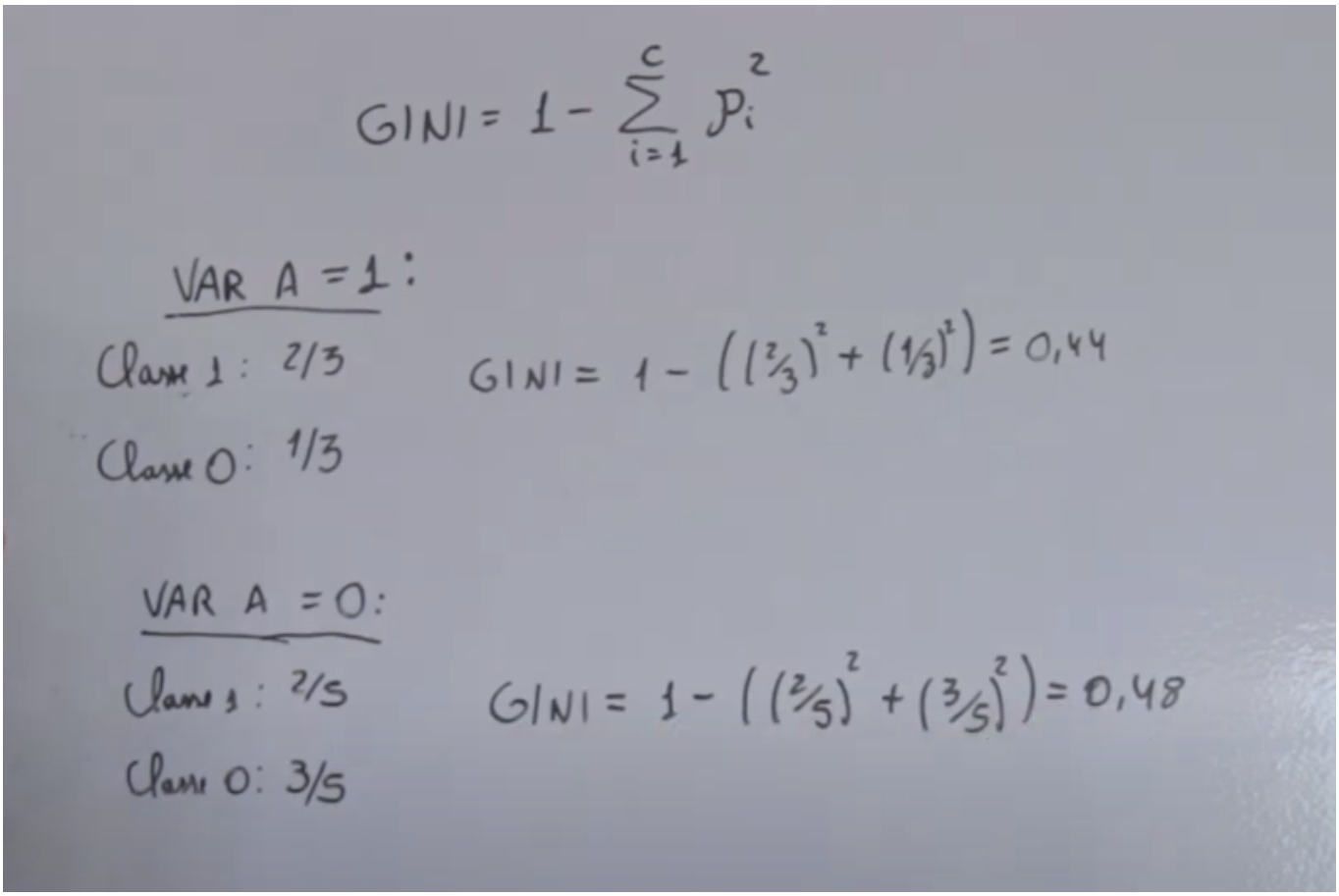


Exemplo prático de classificação:



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

Calcular o índice Gini para variável A:



Primeiro deve-se separar entre variável A = 1 e A = 0, após isso se contabiliza quantas ocorrências que a variável A era igual 1 e da classe 1 e 0, no caso
Classe 1 = 2/3 e Classe 0 = 1/3

Também fazemos isso quando a variável A = 0, somamos a ocorrência das duas classes quando a variável é igual a 0 e temos:

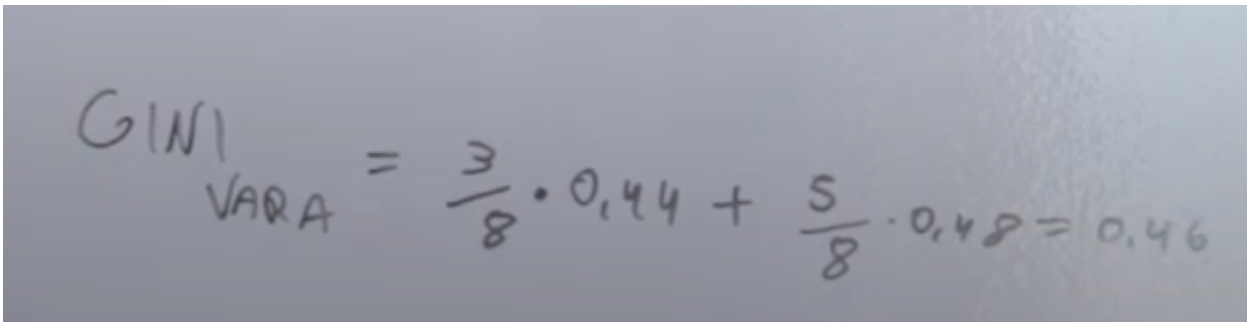
Classe 1 = 2 / 5 e Classe 0 = 3 / 5

Feito isso, basta aplicar a fórmula de Gini:

Var A = 1:
Gini = 1 - ((2/3)2 + (1/3)2) = 0,44

Var A = 0:
Gini = 1 - ((2/5)2 + (3/5)2) = 0,48

Por fim para calcular o índice Gini da variável A como um todo, temos que fazer uma média ponderada considerando o total de amostras, no caso 8:

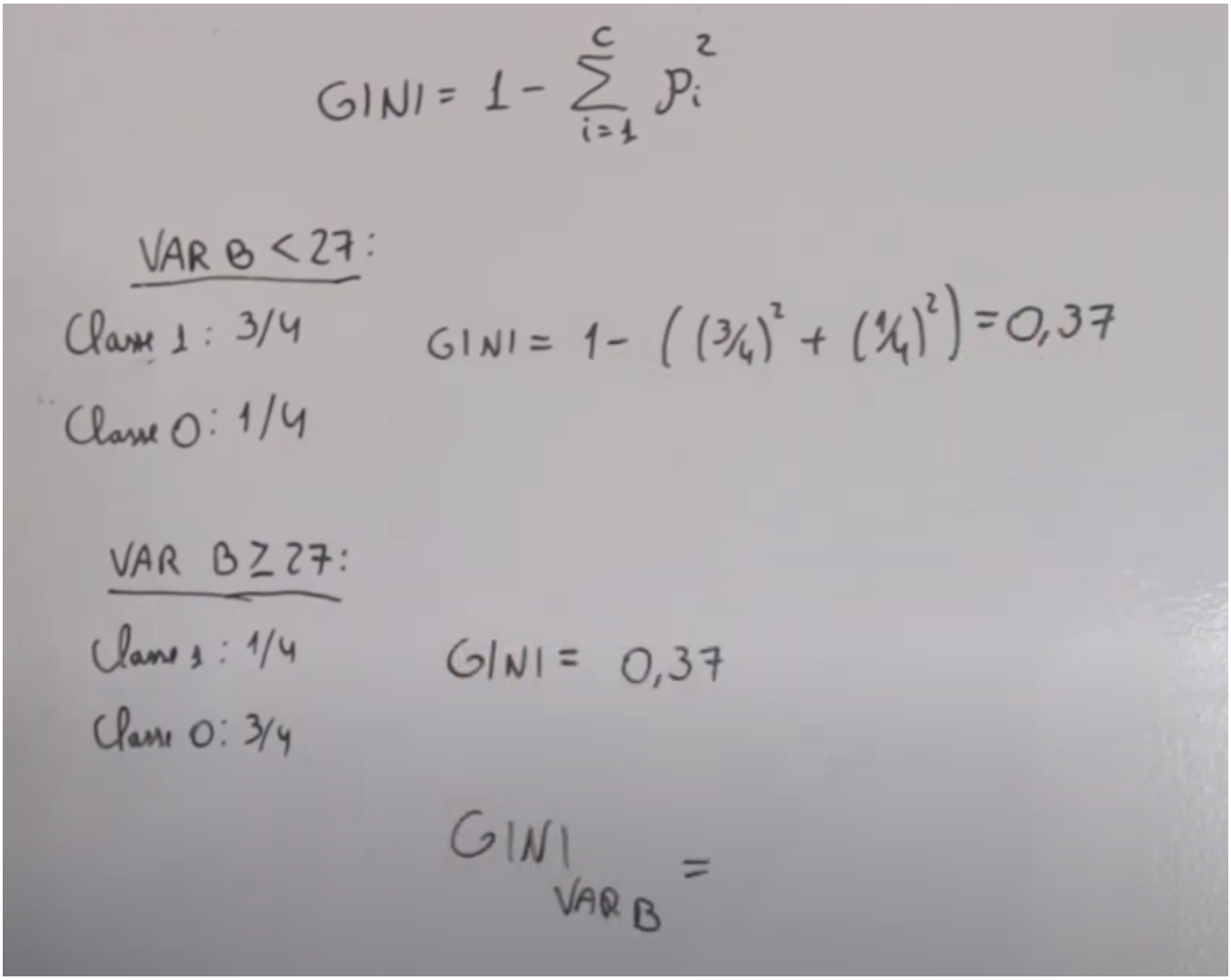


3/8 são quando a variável A = 1 e 5 / 8 quando igual a 0, basta multiplicar cada uma pelo índice correspondente anterior e somar os resultados = 0,46

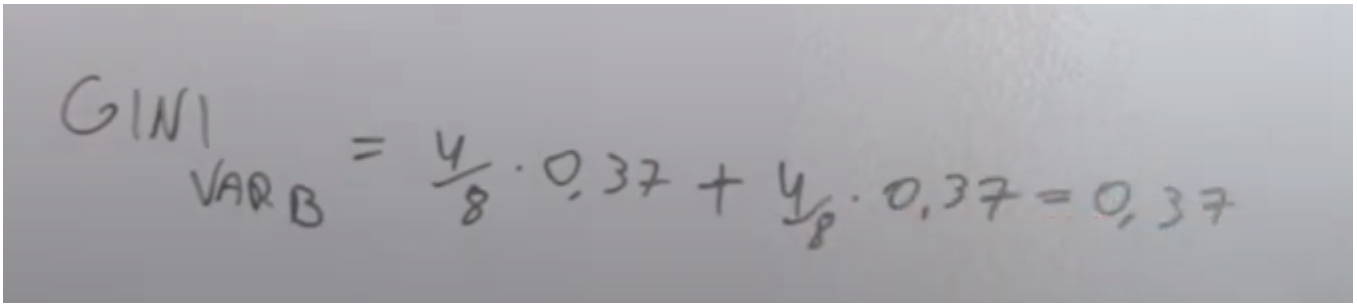
Calcular o índice Gini para variável B:

Como a variável B não é binária, temos que encontrar um valor intermediário para separar os valores em duas partes, nesse caso foi escolhido o 27:

Seguindo a mesma lógica, podemos observar os valores abaixo de 27 e maiores que 27:

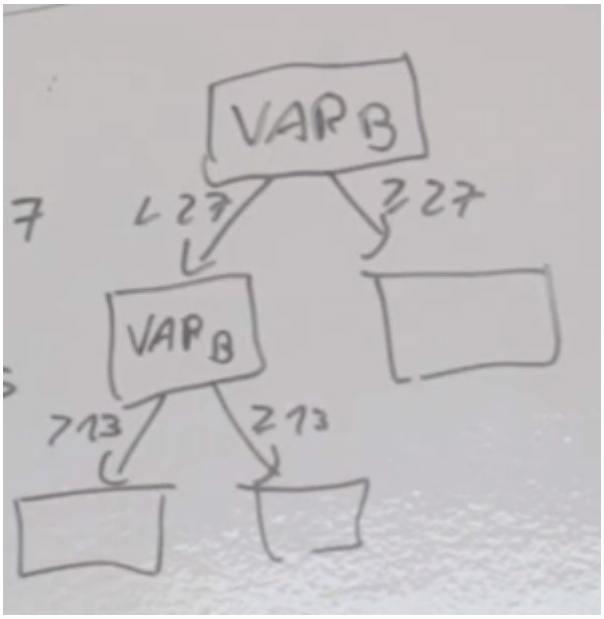


Resultado:



Após calcular o índice Gini das duas variáveis, escolhemos o menor valor para ser a raíz da árvore e não o maior como na entropia. Isso se deve, pois quanto maior o índice, mais os dados estarão divididos ao meio, ou seja, da mesma forma que quanto maior desordem, os dados estarão 50% divididos de um lado e do outro, sendo o pior cenário para separar os nós da árvore.

Feito isso temos o nó raíz da árvore, e para calcular o restante, basta continuar com a condição da variável B, e verificar todas as ocorrências menores que 27 ou maiores para montar novos índices para variável A e B, e montar o segundo nó, e assim sucessivamente da forma abaixo:



Veja que como escolhemos a B ela é o primeiro nó, após isso podemos decidir que o menor índice no lado esquerdo onde são valores < 27 também é da variável B e pra isso temos que selecionar outro valor de quebra, no caso maiores ou menores que 13 e assim sucessivamente, caso fosse a variável A, bastava dividir por valores 0 ou 1 e continuar a árvore. Isso é feito até dividirmos todos os dados ou um limite imposto pelo algoritmo. Vale dizer que quando dividimos o primeiro nó em 2, uma parte dos dados vai pra direita e esquerda, então cada quebra o número diminuí até não ter mais ninguém para dividir.