Documentação - Árvore de decisão

Uma árvore de decisão é um modelo de machine learning usado para tomar decisões com base em perguntas sequenciais. Ela funciona como um fluxograma, onde cada nó interno representa uma decisão (ou condição) e cada folha indica um resultado.

Quando usar?

• Quando você precisa de explicações claras sobre como as decisões foram tomadas (como em diagnósticos médicos).

• Ideal para classificação (ex: "Este cliente vai cancelar a assinatura?") ou regressão (ex: "Qual será o valor de venda de um imóvel?").

Vantagens

• Fácil de interpretar: Dá para entender a lógica olhando a estrutura da árvore. • Funciona com dados mistos: Aceita dados numéricos e categóricos.

• Rápida para treinar: Principalmente com pequenas bases de dados.

Desvantagens

• Sensível a overfitting: Pode se ajustar demais aos dados de treino e ter baixo desempenho em novos dados.

• Menos precisa que outros modelos como Random Forest ou Redes Neurais em casos mais complexos.

Quando uma árvore de decisão é treinada, ela precisa decidir qual a melhor pergunta para separar os dados. A ideia é fazer as divisões que organizem os dados da maneira mais clara possível. Para isso, usamos métricas. As principais são:

Entropia

O cálculo da entropia em árvores de decisão vem da Teoria da Informação e é usado para ajudar o modelo a tomar decisões mais eficazes ao dividir dados.

Quanto mais "bagunçada" ou diversa a distribuição dos dados, maior é a entropia.

Em uma árvore de decisão, queremos reduzir essa incerteza ao máximo para fazer previsões mais precisas.

- Entropia baixa: Quando os dados estão mais organizados (por exemplo, quase todos os valores são iguais).

- Entropia alta: Quando os dados são variados e misturados (ex: metade "sim", metade "não").

Exemplo da Importância da Entropia

Se você tem um grupo com 10 aprovados e 10 reprovados e precisa tomar uma decisão, ele está "bagunçado" ***porque não há uma classe predominante.***

Ou seja, você não consegue dizer com confiança se um novo aluno nesse grupo provavelmente seria aprovado ou reprovado, **já que está 50/50. Isso é uma entropia alta – muita incerteza.**

Agora, imagine outro cenário: você tem dois grupos. No primeiro, 18 alunos aprovados e só 2 reprovados. No segundo, 1 aprovado e 19 reprovados.

Esses grupos estão muito mais organizados porque a maioria dos alunos em cada grupo pertence claramente a uma única classe (o primeiro grupo é quase todo de aprovados e o segundo quase todo de reprovados). **A entropia desses grupos é baixa, pois você tem pouca incerteza – se você precisar prever a situação de um novo aluno, é bem mais fácil dizer que, se ele cair no primeiro grupo, ele será aprovado, e, se cair no segundo grupo, será reprovado.**

Calculo da Entropia

A formula de cálculo da entropia funciona por meio da seguinte formula matemática:

$$\text{Entropia}(S) = -\sum_{i-1}^{n} p_i \log_2(p_i)$$

Para aplicar esse cálculo numa base de dados geral, é necessário fazer a contagem dos registros. Conforme abaixo.

Risco
Alto
Alto
Moderado
Alto
Baixo
Baixo
Alto
Moderado
Baixo
Baixo
Alto
Moderado
Baixo
Alto

Alto = 6/14
Moderado = 3/14
Baixo = 5/14

Feito isso, é necessário aplicar a formula, para descobrir a entropia geral da classe:

$$Entropia(S) = -\frac{6}{14} \cdot Log\left(\frac{6}{14}; 2\right) \cdot -\frac{3}{14} \cdot Log\left(\frac{3}{14}; 2\right) \cdot -\frac{5}{14} \cdot Log\left(\frac{5}{14}; 2\right)$$
$$= 1{,}53$$

Acima, notamos a entropia geral como 1,53.

Ganho de Informação

Para identificar qual atributo tem maior ganho de informação, precisamos calcular a entropia para cada um dos atributos que aparecem para chegar no valor de Risco, e aplicar a formula de ganho de informação.

Formula do ganho de informação

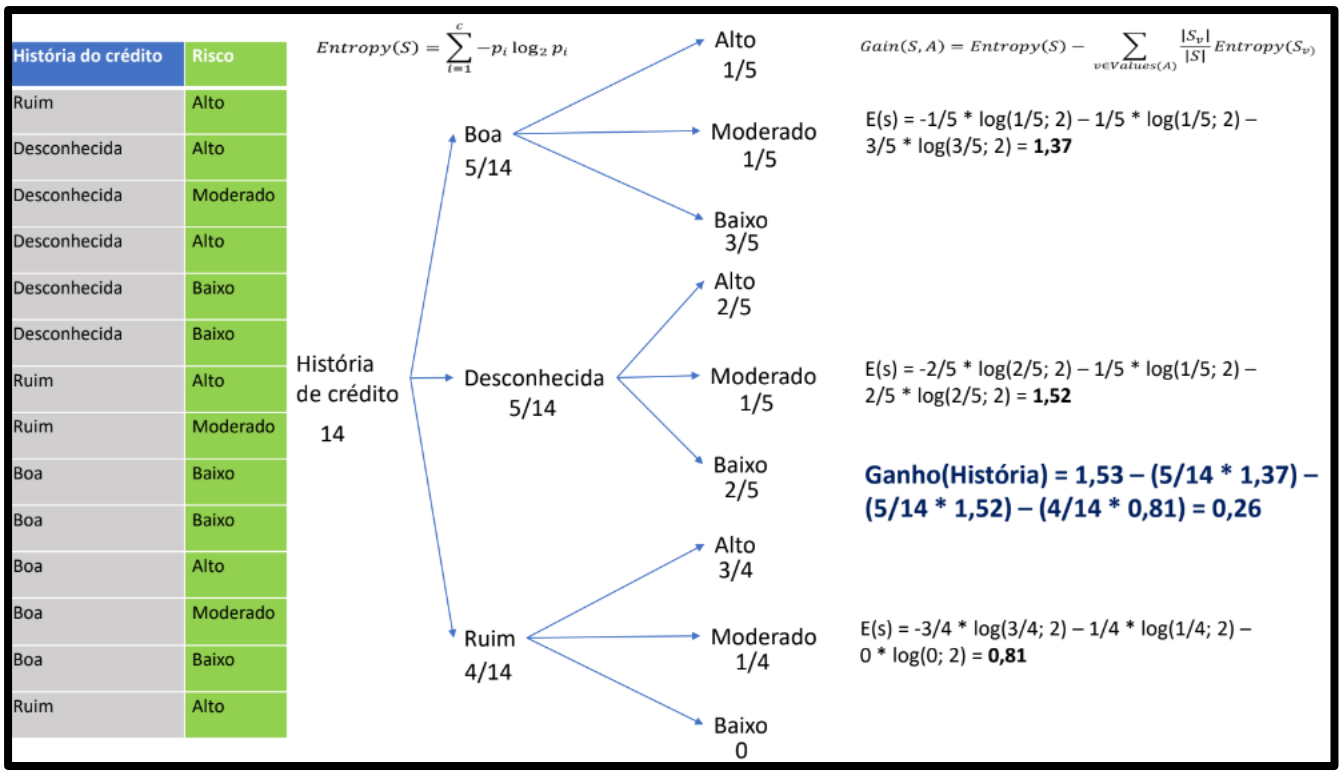
$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

Exemplo de aplicação:

Primeiro identificamos a entropia das classes dentro do atributo que estamos analisando.

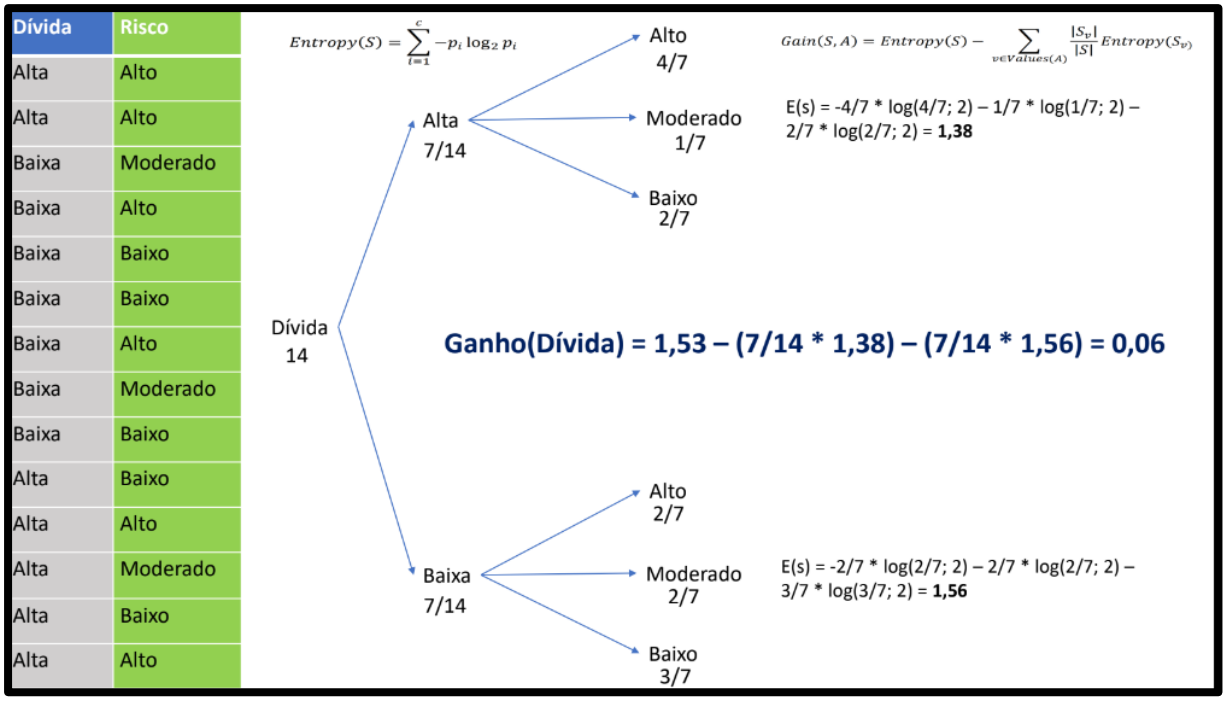
Exemplo: no caso de história de crédito, temos 3 classes: ***Boa, Desconhecida ou Ruim.***

Precisamos calcular a entropia para cada uma delas, e então aplicamos a formula.

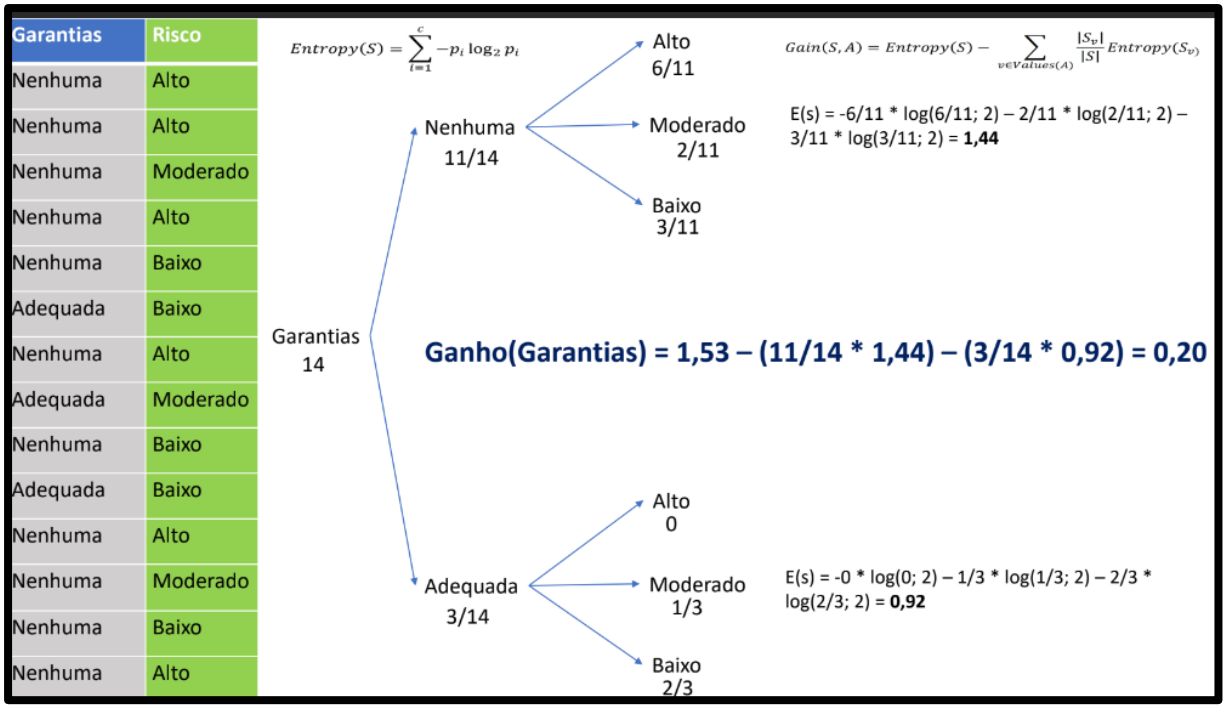


Aplicações para outros atributos

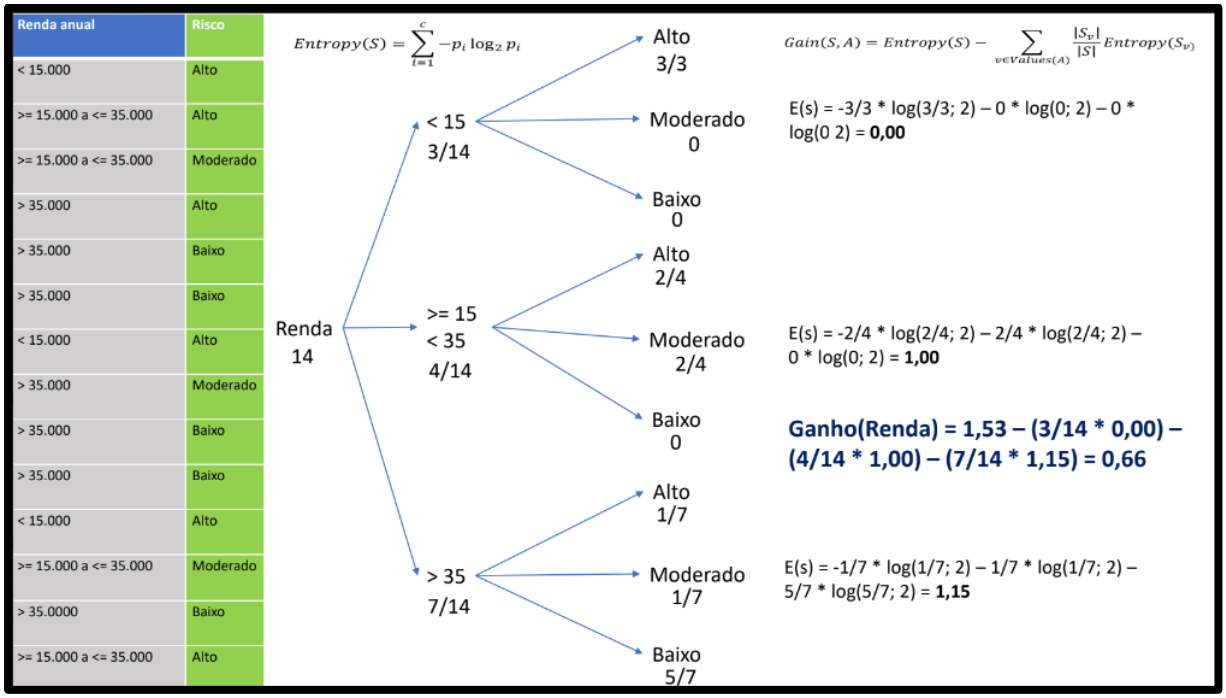
Atributo Divida



Atributo Garantia



Atributo Renda



Com esses cálculos realizados, conseguimos identificar os atributos que mais oferecem ganho de informações.

*Sendo eles**: Renda, História de Crédito e Garantias***

História de crédito = 0,26
Dívida = 0,06
Garantias = 0,20
Renda = 0,66