

# Árvores de decisão – Random Forest

Jones Granatyr

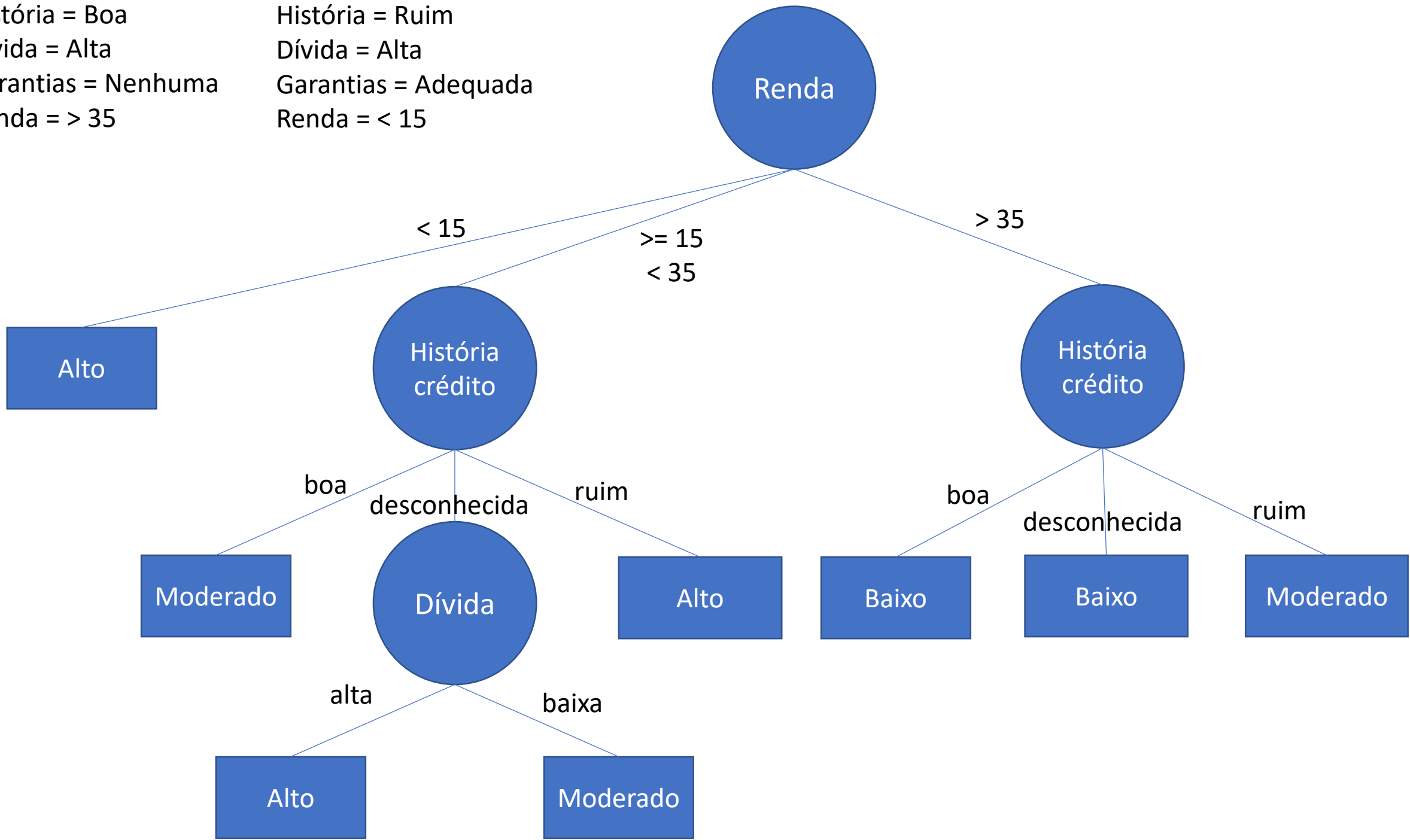


# Base original

História do crédito	Dívida	Garantias	Renda anual	Risco
Ruim	Alta	Nenhuma	< 15.000	Alto
Desconhecida	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto
Desconhecida	Baixa	Nenhuma	>= 15.000 a <= 35.000	Moderado
Desconhecida	Baixa	Nenhuma	> 35.000	Alto
Desconhecida	Baixa	Nenhuma	> 35.000	Baixo
Desconhecida	Baixa	Adequada	> 35.000	Baixo
Ruim	Baixa	Nenhuma	< 15.000	Alto
Ruim	Baixa	Adequada	> 35.000	Moderado
Boa	Baixa	Nenhuma	> 35.000	Baixo
Boa	Alta	Adequada	> 35.000	Baixo
Boa	Alta	Nenhuma	< 15.000	Alto
Boa	Alta	Nenhuma	>= 15.000 a <= 35.000	Moderado
Boa	Alta	Nenhuma	> 35.0000	Baixo

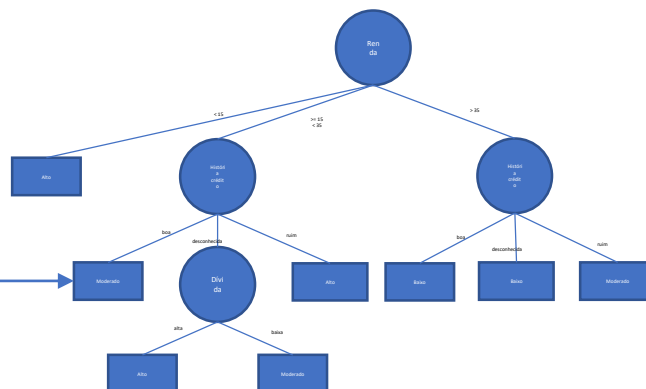
História = Boa  
Dívida = Alta  
Garantias = Nenhuma  
Renda = > 35

História = Ruim  
Dívida = Alta  
Garantias = Adequada  
Renda = < 15



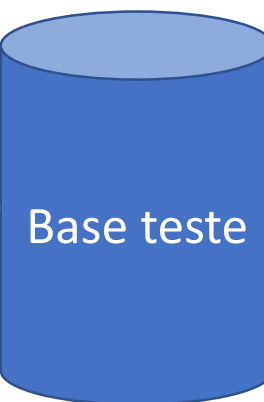
Risco de crédito	História do crédito			Dívida		Garantias		Renda anual		
	Baixa	Desenvolvida	Buena	Alta	Baixa	Intermedia	Alta	< 1500	1500 <= 3500	> 3500
Alta	1/5	2/5	3/4	4/7	2/7	6/11	0	3/3	2/4	1/7
Intermedia	1/5	1/5	1/4	1/7	2/7	2/11	1/3	0	2/4	1/7
Baixa	3/5	2/5	0	2/7	3/7	3/11	2/3	0	0	5/7

Naive bayes



Árvore de decisão

Registros  
% acerto



Base teste

Base  
treinamento

$$Entropy(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

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

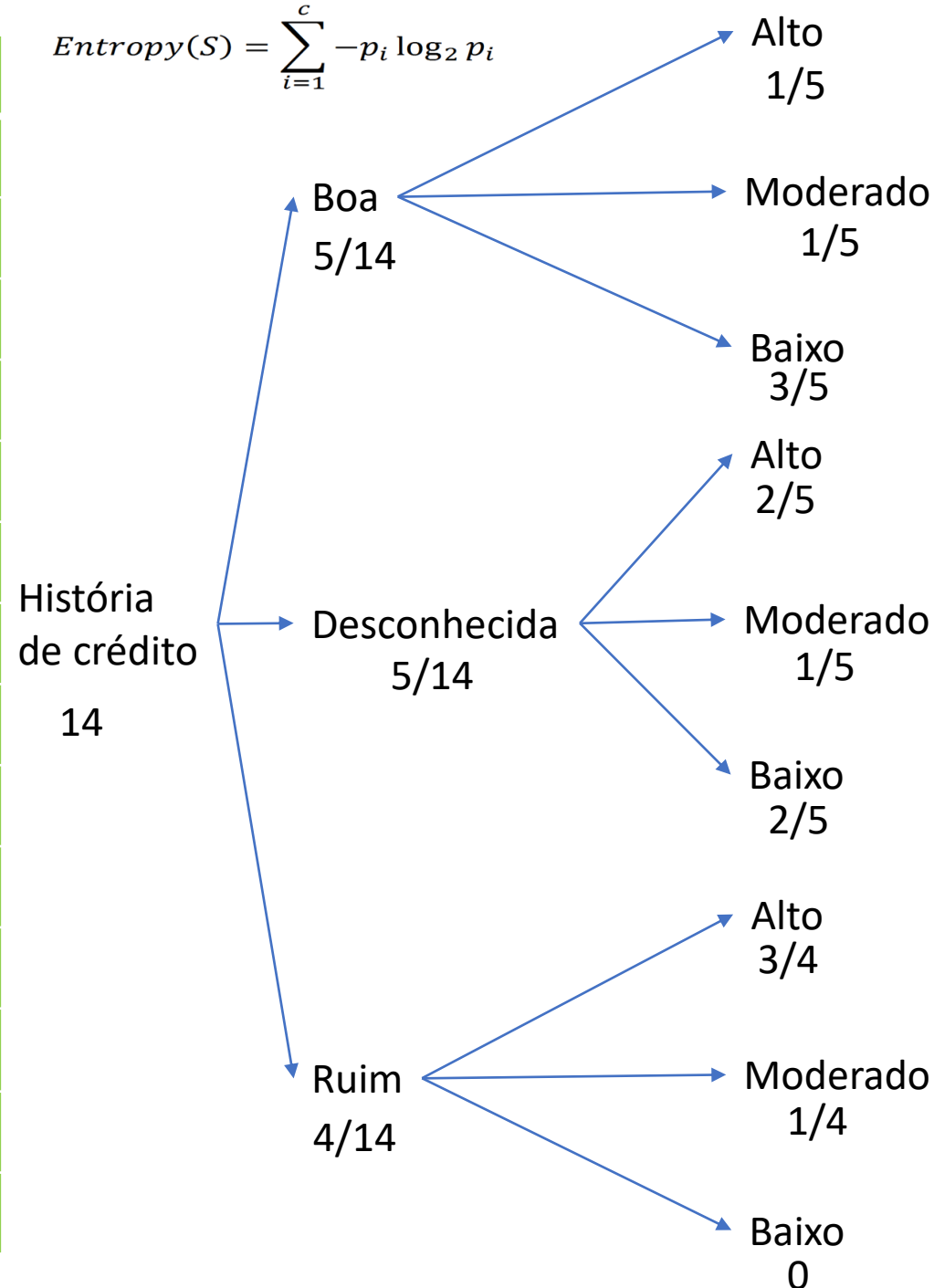
Risco
Alto
Alto
Moderado
Alto
Baixo
Baixo
Alto
Moderado
Baixo
Baixo
Alto
Moderado
Baixo
Alto

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

$$Entropy(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

$$E(s) = -6/14 * \log(6/14; 2) - 3/14 * \log(3/14; 2) - 5/14 * \log(5/14; 2) = \mathbf{1,53}$$

História do crédito	Risco
Ruim	Alto
Desconhecida	Alto
Desconhecida	Moderado
Desconhecida	Alto
Desconhecida	Baixo
Desconhecida	Baixo
Ruim	Alto
Ruim	Moderado
Boa	Baixo
Boa	Baixo
Boa	Alto
Boa	Moderado
Boa	Baixo
Ruim	Alto



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

$E(s) = -1/5 * \log(1/5; 2) - 1/5 * \log(1/5; 2) - 3/5 * \log(3/5; 2) = \mathbf{1,37}$

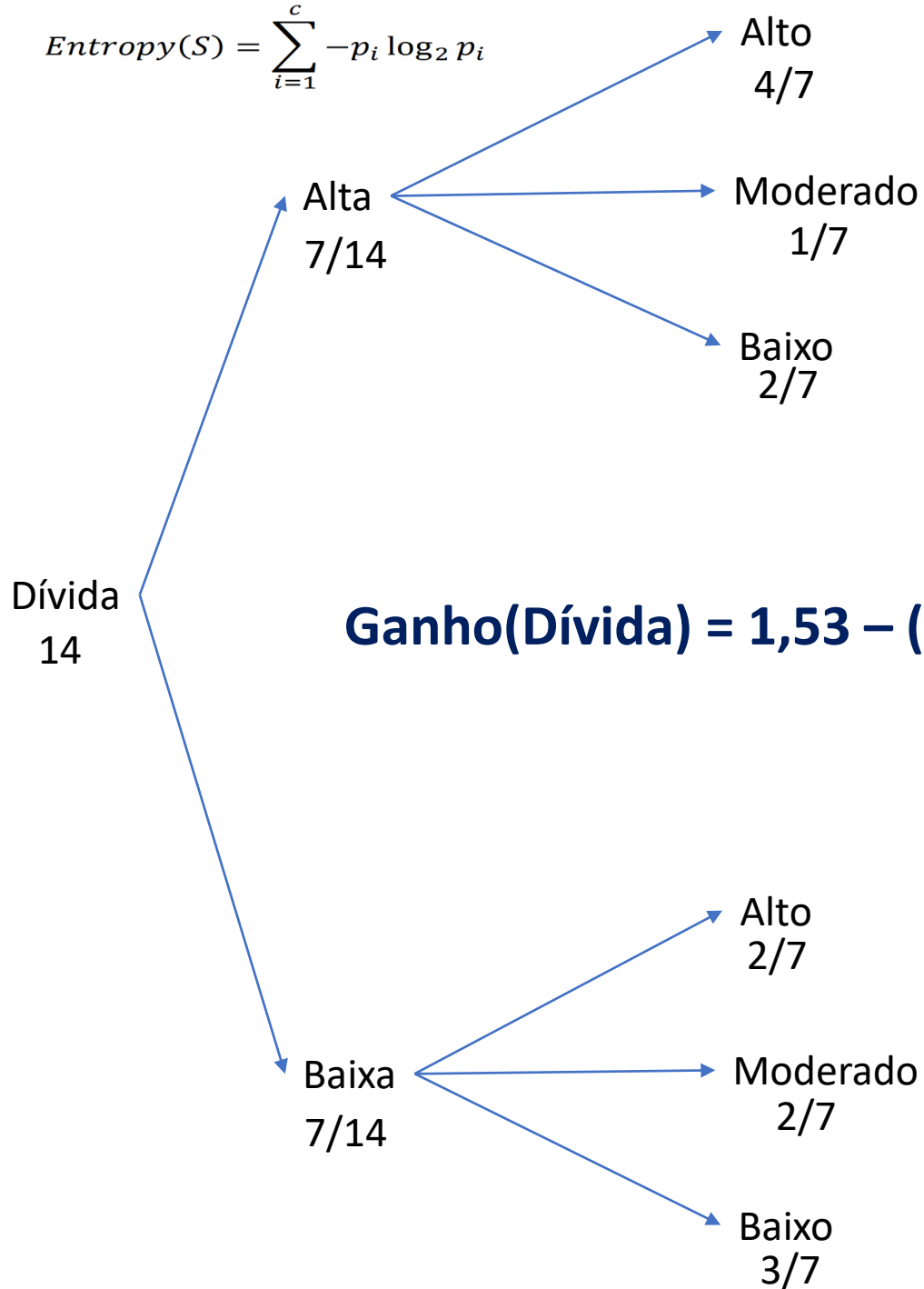
$E(s) = -2/5 * \log(2/5; 2) - 1/5 * \log(1/5; 2) - 2/5 * \log(2/5; 2) = \mathbf{1,52}$

**Ganho(História) = 1,53 – (5/14 \* 1,37) – (5/14 \* 1,52) – (4/14 \* 0,81) = 0,26**

$E(s) = -3/4 * \log(3/4; 2) - 1/4 * \log(1/4; 2) - 0 * \log(0; 2) = \mathbf{0,81}$

Dívida	Risco
Alta	Alto
Alta	Alto
Baixa	Moderado
Baixa	Alto
Baixa	Baixo
Baixa	Baixo
Baixa	Alto
Baixa	Moderado
Baixa	Baixo
Alta	Baixo
Alta	Alto
Alta	Moderado
Alta	Baixo
Alta	Alto

$$Entropy(S) = \sum_{i=1}^c -p_i \log_2 p_i$$



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

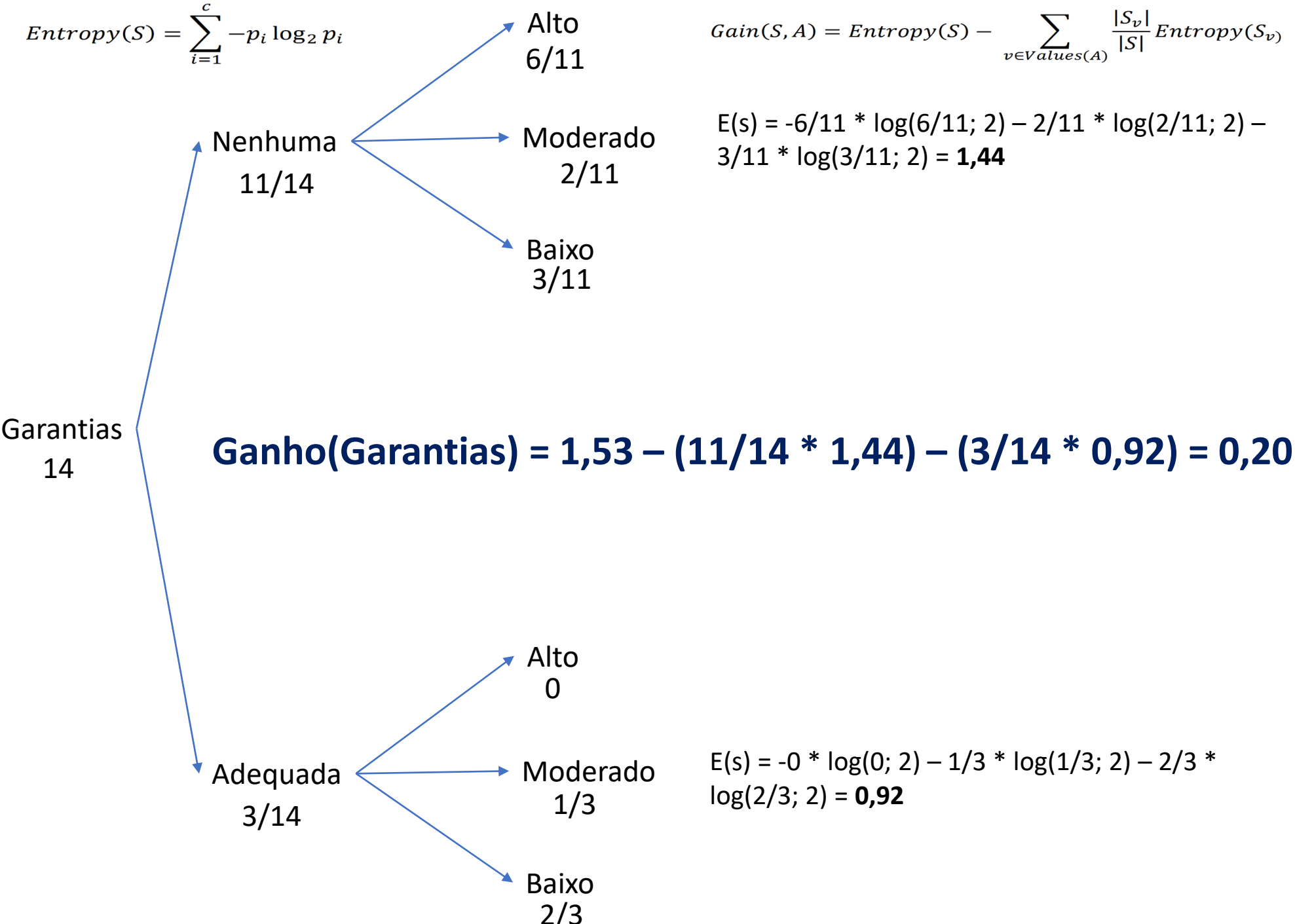
$$E(s) = -4/7 * \log(4/7; 2) - 1/7 * \log(1/7; 2) - 2/7 * \log(2/7; 2) = \mathbf{1,38}$$

**Ganho(Dívida) =  $1,53 - (7/14 * 1,38) - (7/14 * 1,56) = 0,06$**

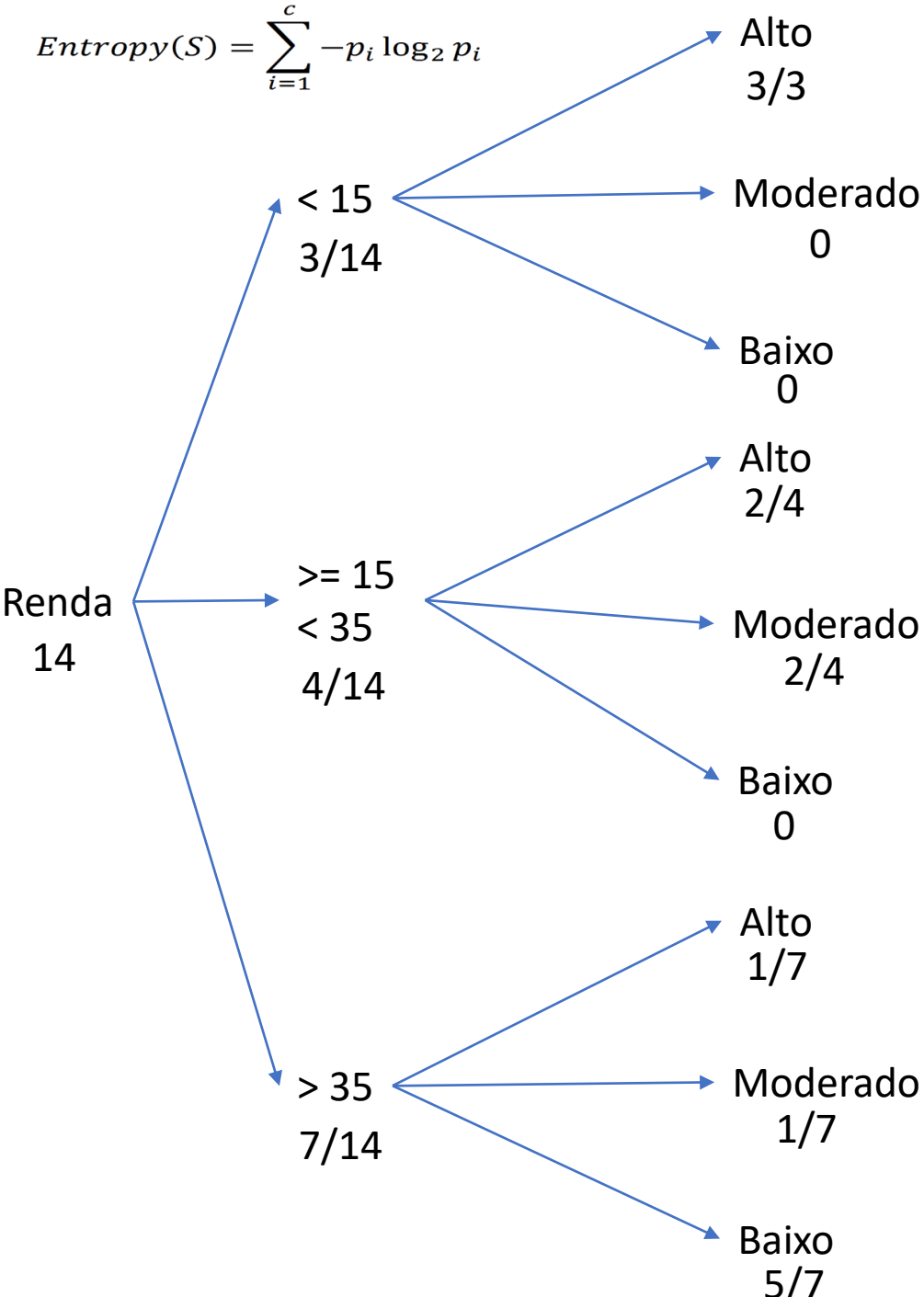
$$E(s) = -2/7 * \log(2/7; 2) - 2/7 * \log(2/7; 2) - 3/7 * \log(3/7; 2) = \mathbf{1,56}$$



Garantias	Risco
Nenhuma	Alto
Nenhuma	Alto
Nenhuma	Moderado
Nenhuma	Alto
Nenhuma	Baixo
Adequada	Baixo
Nenhuma	Alto
Adequada	Moderado
Nenhuma	Baixo
Adequada	Baixo
Nenhuma	Alto
Nenhuma	Moderado
Nenhuma	Baixo
Nenhuma	Alto



Renda anual	Risco
< 15.000	Alto
>= 15.000 a <= 35.000	Alto
>= 15.000 a <= 35.000	Moderado
> 35.000	Alto
> 35.000	Baixo
> 35.000	Baixo
< 15.000	Alto
> 35.000	Moderado
> 35.000	Baixo
> 35.000	Baixo
< 15.000	Alto
>= 15.000 a <= 35.000	Moderado
> 35.0000	Baixo
>= 15.000 a <= 35.000	Alto



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

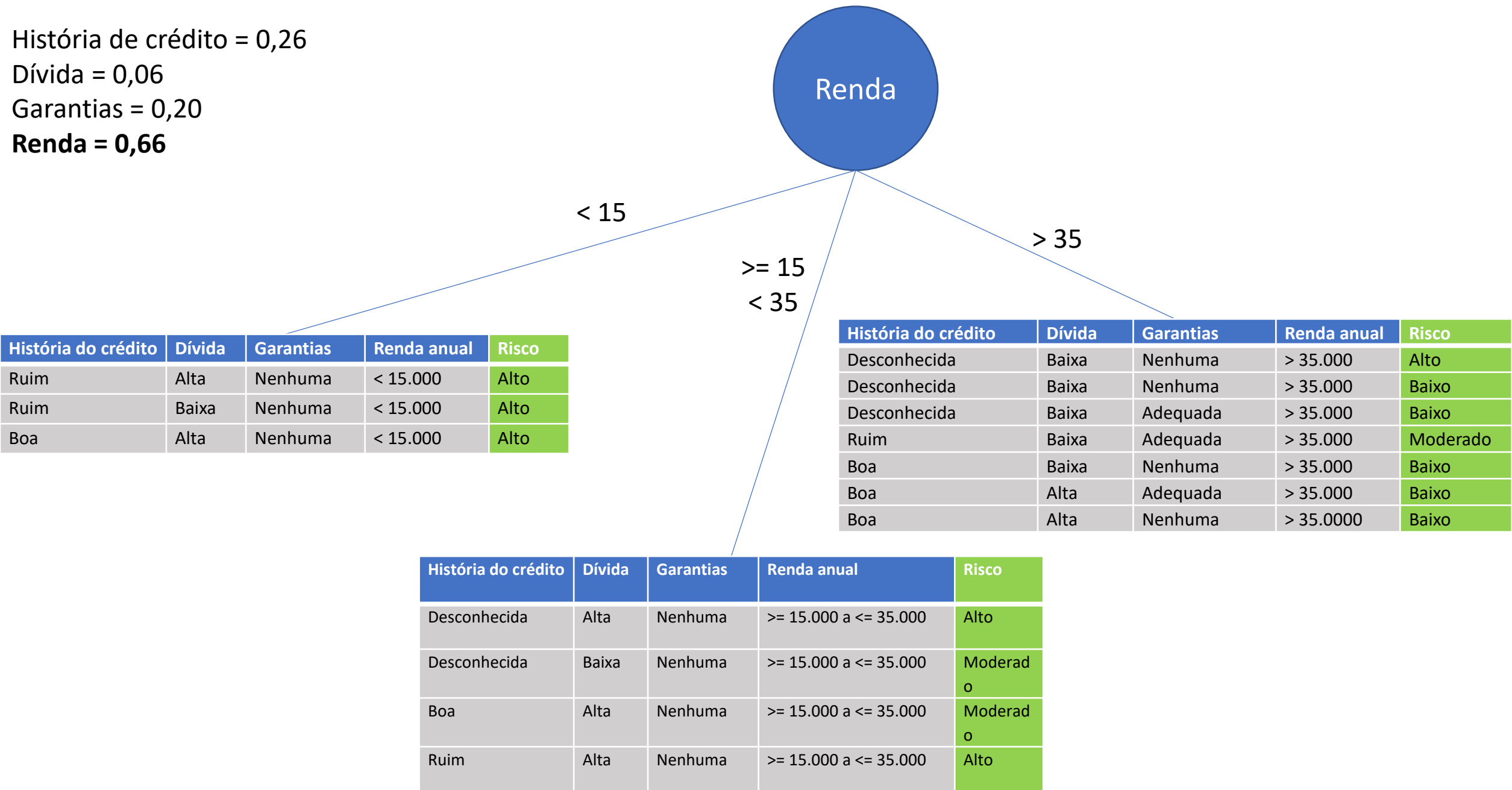
$E(s) = -3/3 * \log(3/3; 2) - 0 * \log(0; 2) - 0 * \log(0; 2) = 0,00$

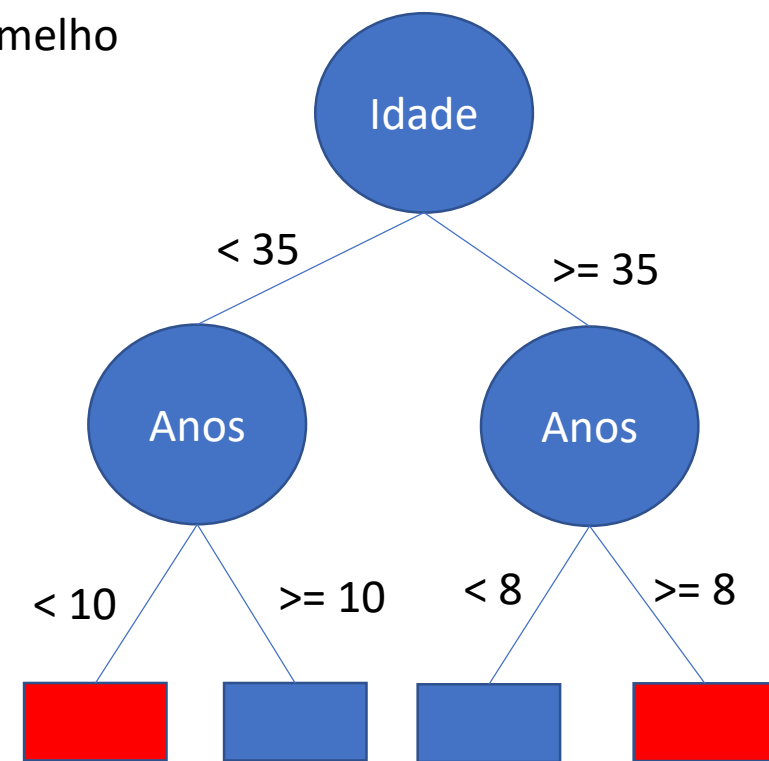
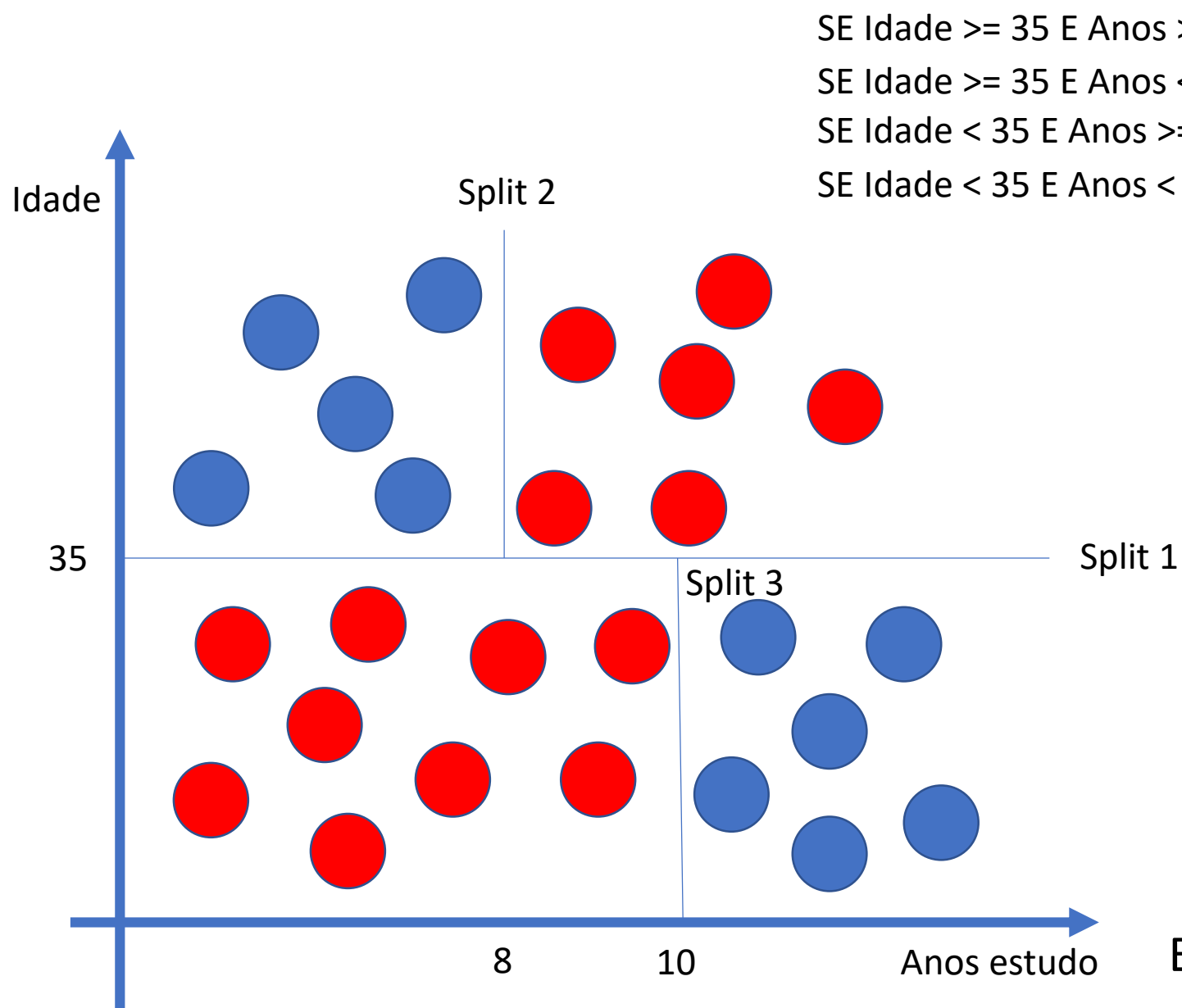
$E(s) = -2/4 * \log(2/4; 2) - 2/4 * \log(2/4; 2) - 0 * \log(0; 2) = 1,00$

**Ganho(Renda) = 1,53 – (3/14 \* 0,00) – (4/14 \* 1,00) – (7/14 \* 1,15) = 0,66**

$E(s) = -1/7 * \log(1/7; 2) - 1/7 * \log(1/7; 2) - 5/7 * \log(5/7; 2) = 1,15$

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





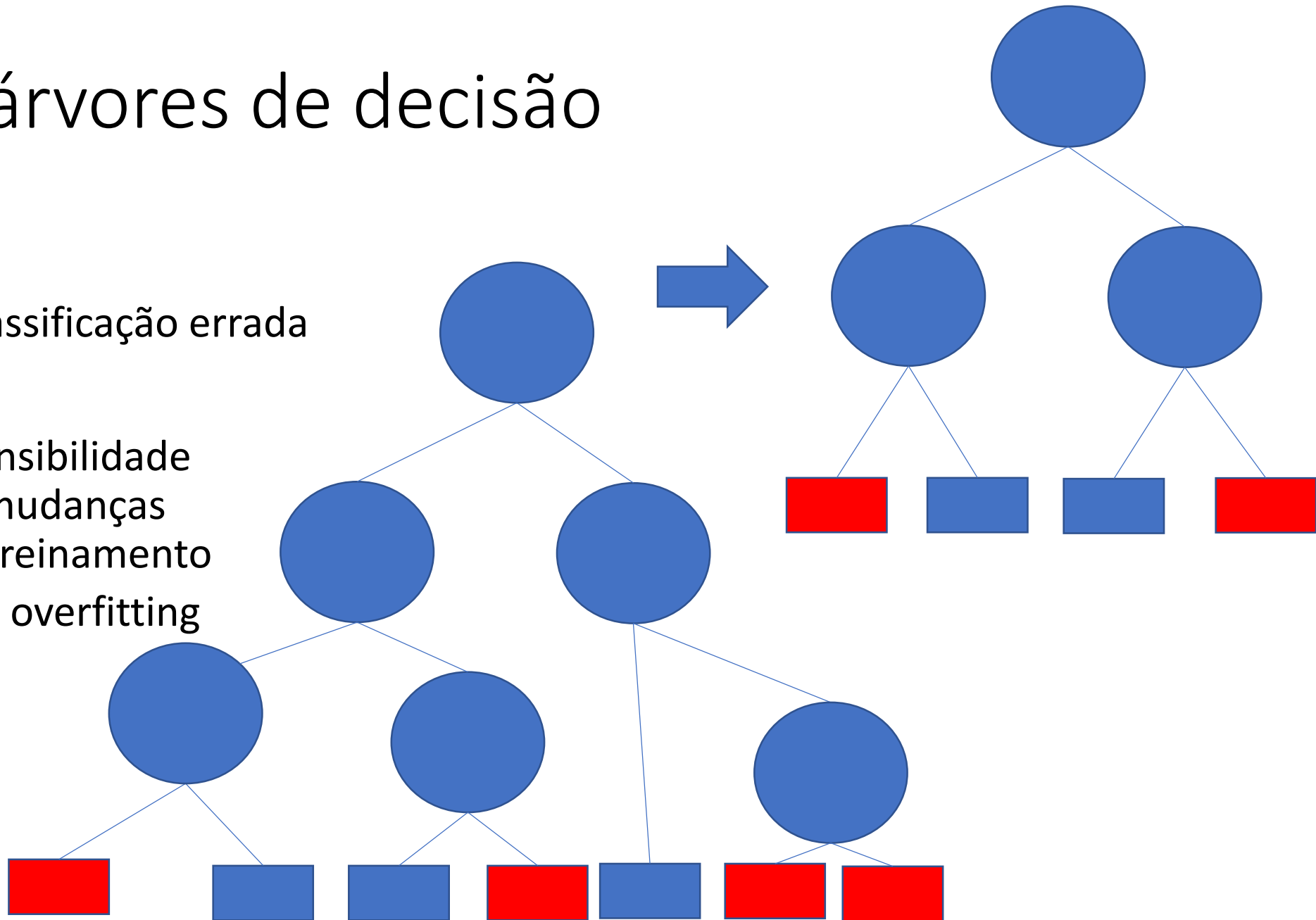
Encontrar o melhor conjunto de divisores





# Poda em árvores de decisão

- Bias (viés)
  - Erros por classificação errada
- Variância
  - Erros por sensibilidade pequena a mudanças na base de treinamento
  - Pode levar a overfitting



# Árvores de decisão

- Vantagens
  - Fácil interpretação
  - Não precisa normalização ou padronização
  - Rápido para classificar novos registros
- Desvantagens
  - Geração de árvores muito complexas
  - Pequenas mudanças nos dados pode mudar a árvore (poda pode ajudar)
  - Problema NP-completo para construir a árvore
- Eram muito populares em meados dos anos 90
- Upgrades como random forest (florestas randômicas) melhoram o desempenho (usado no Kinect da Microsoft)
- CART – classification and regression trees



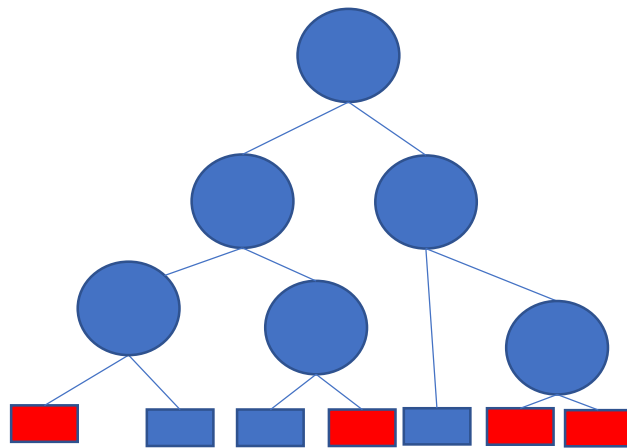
# Random Forest (floresta randômica)



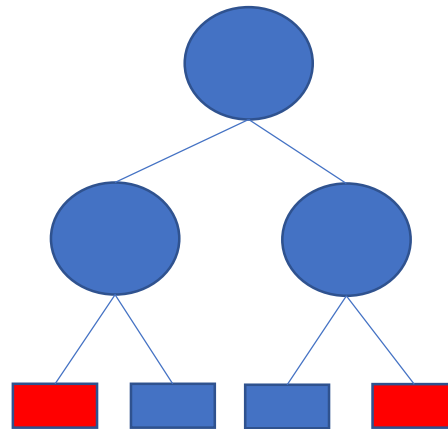


# Random Forest

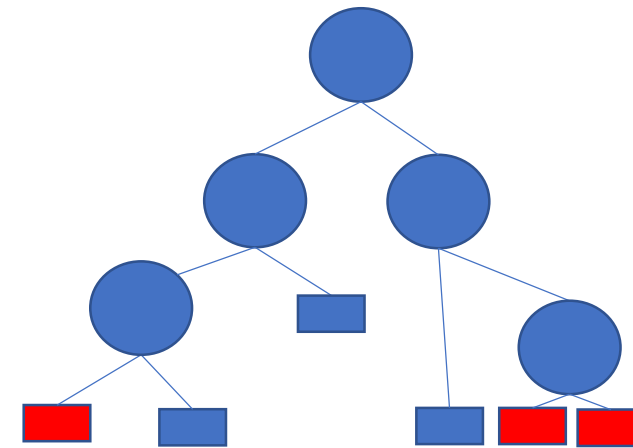
- Ensemble learning (aprendizagem em conjunto)
  - “Consultar diversos profissionais para tomar uma decisão”
  - Vários algoritmos juntos para construir um algoritmo mais “forte”
  - Usa a média (regressão) ou votos da maioria (classificação) para dar a resposta final



Risco = Alto



Risco = Baixo



Risco = Baixo

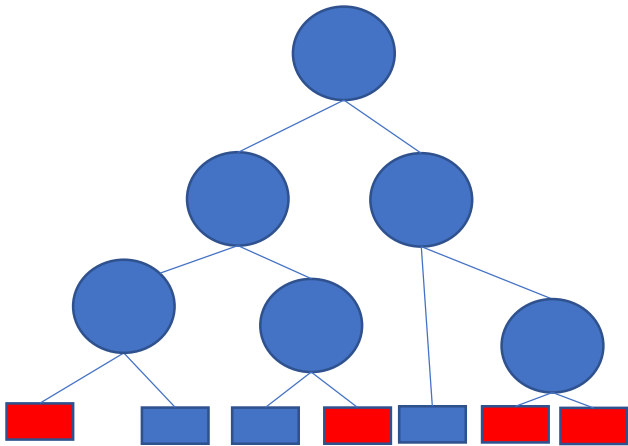
História do crédito	Dívida	Garantias	Renda anual	Risco
Ruim	Alta	Nenhuma	< 15.000	Alto
Desconhecida	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto
Desconhecida	Baixa	Nenhuma	>= 15.000 a <= 35.000	Moderado
Desconhecida	Baixa	Nenhuma	> 35.000	Alto
Desconhecida	Baixa	Nenhuma	> 35.000	Baixo
Desconhecida	Baixa	Adequada	> 35.000	Baixo
Ruim	Baixa	Nenhuma	< 15.000	Alto
Ruim	Baixa	Adequada	> 35.000	Moderado
Boa	Baixa	Nenhuma	> 35.000	Baixo
Boa	Alta	Adequada	> 35.000	Baixo
Boa	Alta	Nenhuma	< 15.000	Alto
Boa	Alta	Nenhuma	>= 15.000 a <= 35.000	Moderado
Boa	Alta	Nenhuma	> 35.000	Baixo
Ruim	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto

# Random Forest

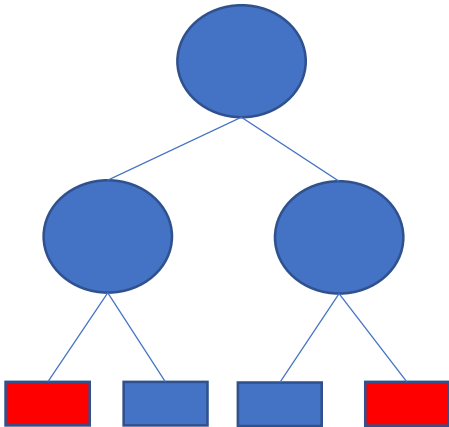
História do crédito	Dívida	Garantias	Renda anual	Risco
Ruim	Alta	Nenhuma	< 15.000	Alto
Desconhecida	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto
Desconhecida	Baixa	Nenhuma	>= 15.000 a <= 35.000	Moderado
Desconhecida	Baixa	Nenhuma	> 35.000	Alto
Desconhecida	Baixa	Nenhuma	> 35.000	Baixo
Desconhecida	Baixa	Adequada	> 35.000	Baixo
Ruim	Baixa	Nenhuma	< 15.000	Alto
Ruim	Baixa	Adequada	> 35.000	Moderado
Boa	Baixa	Nenhuma	> 35.000	Baixo
Boa	Alta	Adequada	> 35.000	Baixo
Boa	Alta	Nenhuma	< 15.000	Alto
Boa	Alta	Nenhuma	>= 15.000 a <= 35.000	Moderado
Boa	Alta	Nenhuma	> 35.000	Baixo
Ruim	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto

Escolhe de forma aleatória K atributos para comparação da métrica de pureza/impureza (impureza de gini/entropia)

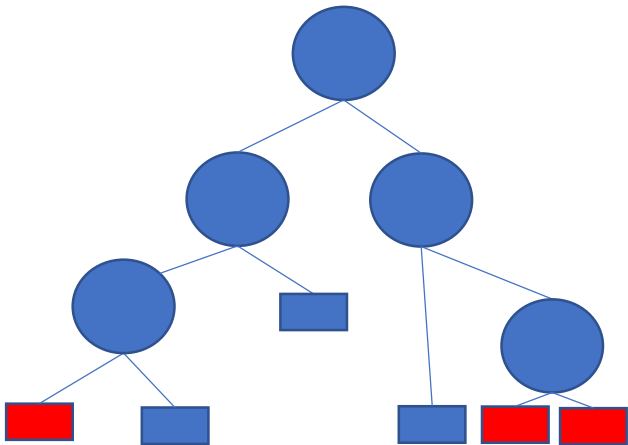
K = 3  
Árvores = 3



História de crédito  
Dívida  
Garantias



Renda  
Dívida  
Garantias



Renda  
História de crédito  
Dívida



Artigo: Real-Time Human Pose Recognition in Parts from Single Depth Images

# Conclusão

