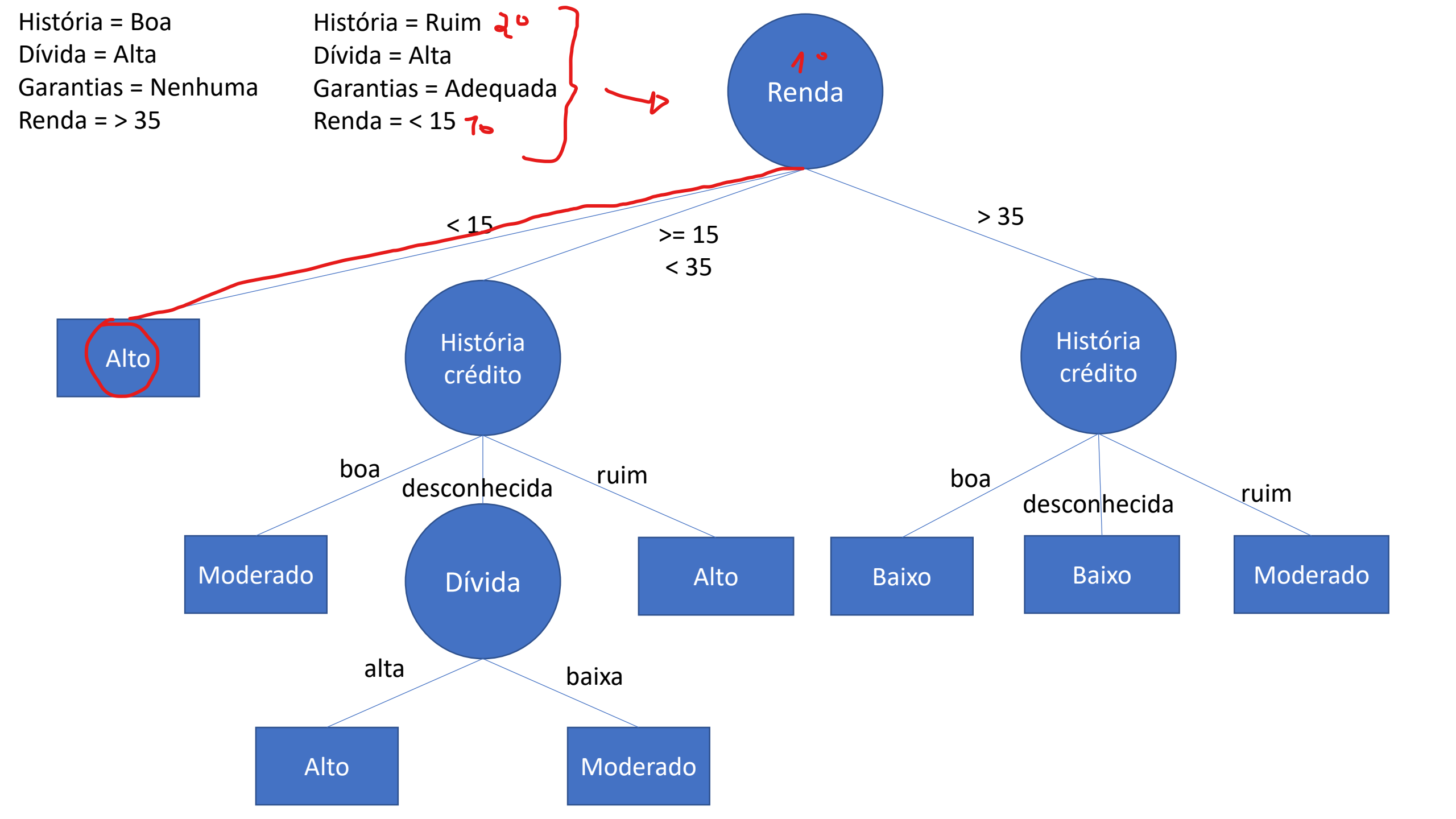
Árvores de decisão – Random Forest

Jones Granatyr

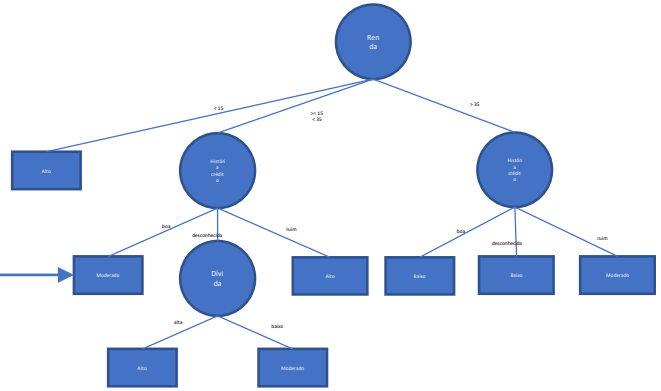
Base original

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
2º
1º
Renda
< 15
>= 15
< 35
> 35
Alto
História crédito
História crédito
boa
desconhecida
ruim
boa
desconhecida
ruim
Moderado
Dívida
Alto
Baixo
Baixo
Moderado
alta
baixa
Alto
Moderado

Risco do crédito	História do crédito			Dívida		Garantias		Renda anual		
	Boa 5	Desconhecida 1	Ruim 4	Alta 7	Baixa 7	Nenhuma 11	Adequada 3	< 15000 3	>= 15000 e <= 35000 4	> 35000 7
Alto 4/14	1/5	2/5	3/4	4/7	2/7	6/11	0	3/3	2/4	1/7
Moderado 2/14	1/5	1/5	1/4	1/7	2/7	2/11	1/3	0	2/4	1/7
Baixo 8/14	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 treinamento

Base teste

Somatório:
C = Quantidade dos valores probabilísticos
i = 1 quer dizer que de onde começa o somatório, no caso do primeiro C

Por exemplo se C = 2, faremos o cálculo desde o ínicio de i até chegar em C de 1 em 1

Entropia

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

Valores probabilísticos (3/10 por ex)

Ganho

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

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

3C

Alto = 6/14 1C

Moderado = 3/14 2C

Baixo = 5/14 3C

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

Pi de 1C

Pi de 1C

2C

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

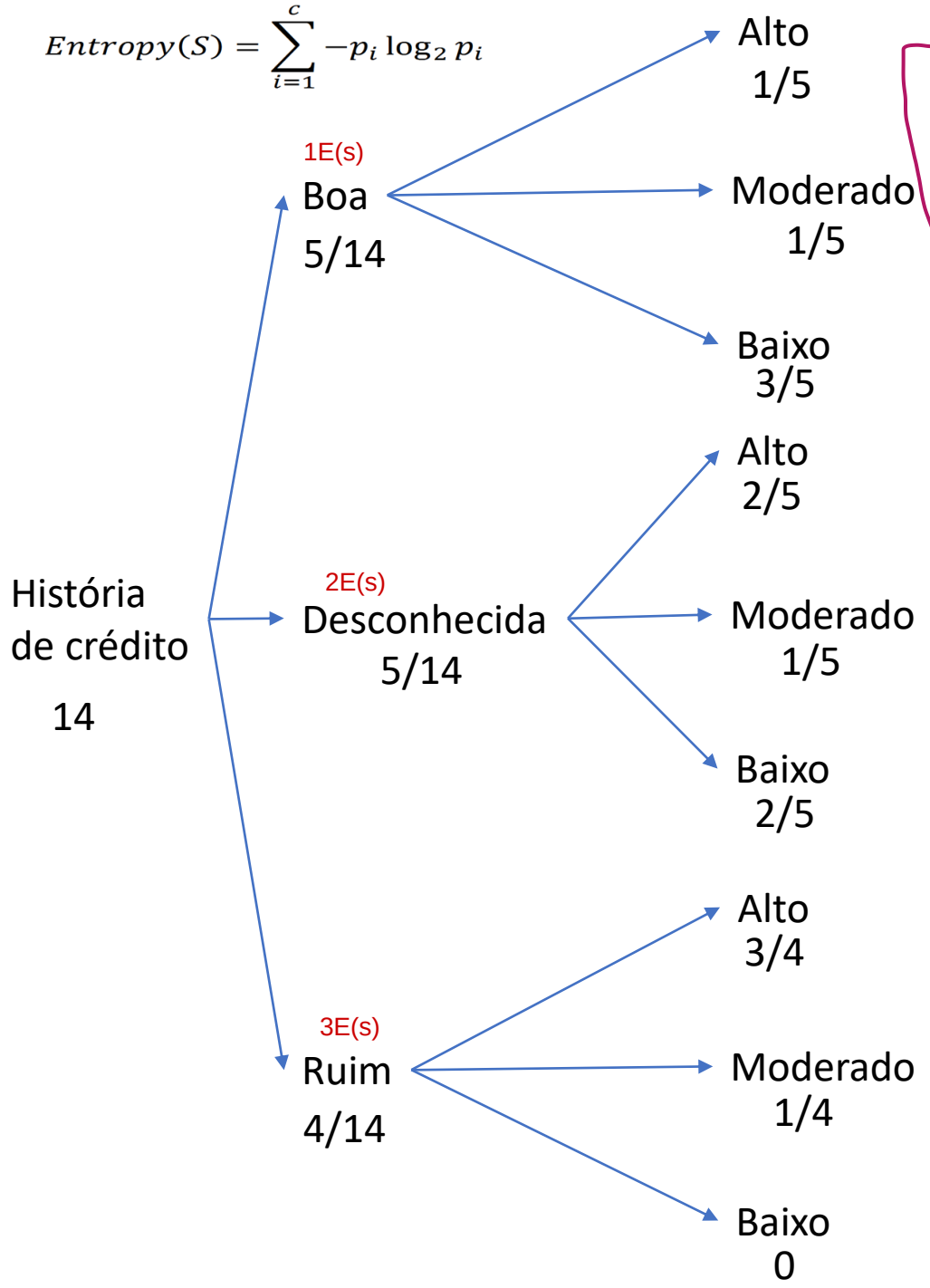
2c

3c

3c

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

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



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

ALTO moderado

E(s) = -1/5 * log(1/5; 2) – 1/5 * log(1/5; 2) – 3/5 * log(3/5; 2) = **1,37**

BAIXO

1E(s)

E(s) = -2/5 * log(2/5; 2) – 1/5 * log(1/5; 2) – 2/5 * log(2/5; 2) = **1,52**

2E(s)

Entropy(s)

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

S(Desconhecida 5/14) * sv(entropia do Desconhecida

S(Ruim) * sv(entropia do Ruim)

S(Boa 5/14) * sv (entropia do Boa)

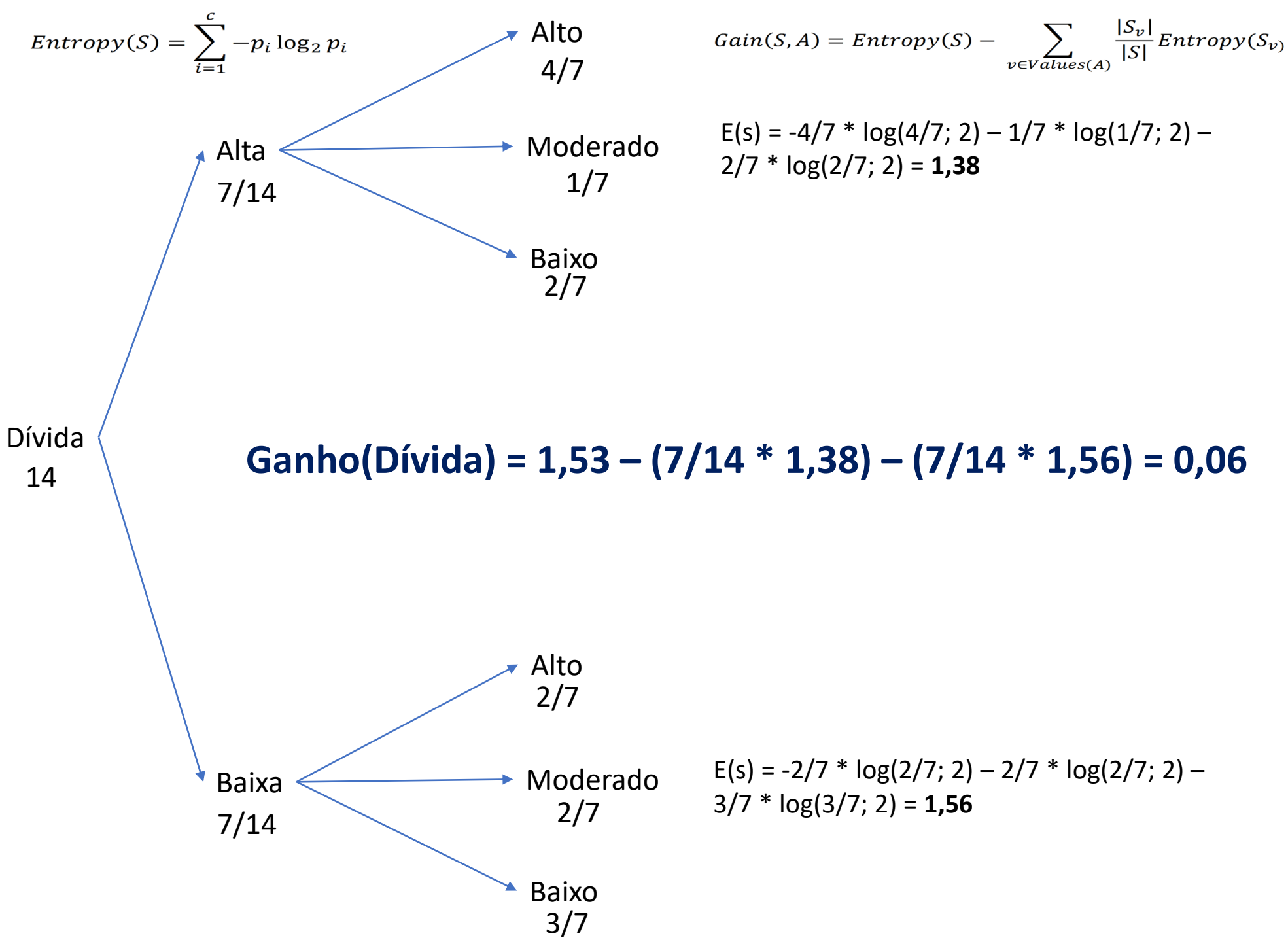
E(s) = -3/4 * log(3/4; 2) – 1/4 * log(1/4; 2) – 0 * log(0; 2) = **0,81**

3E(s)

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) = **1,38**

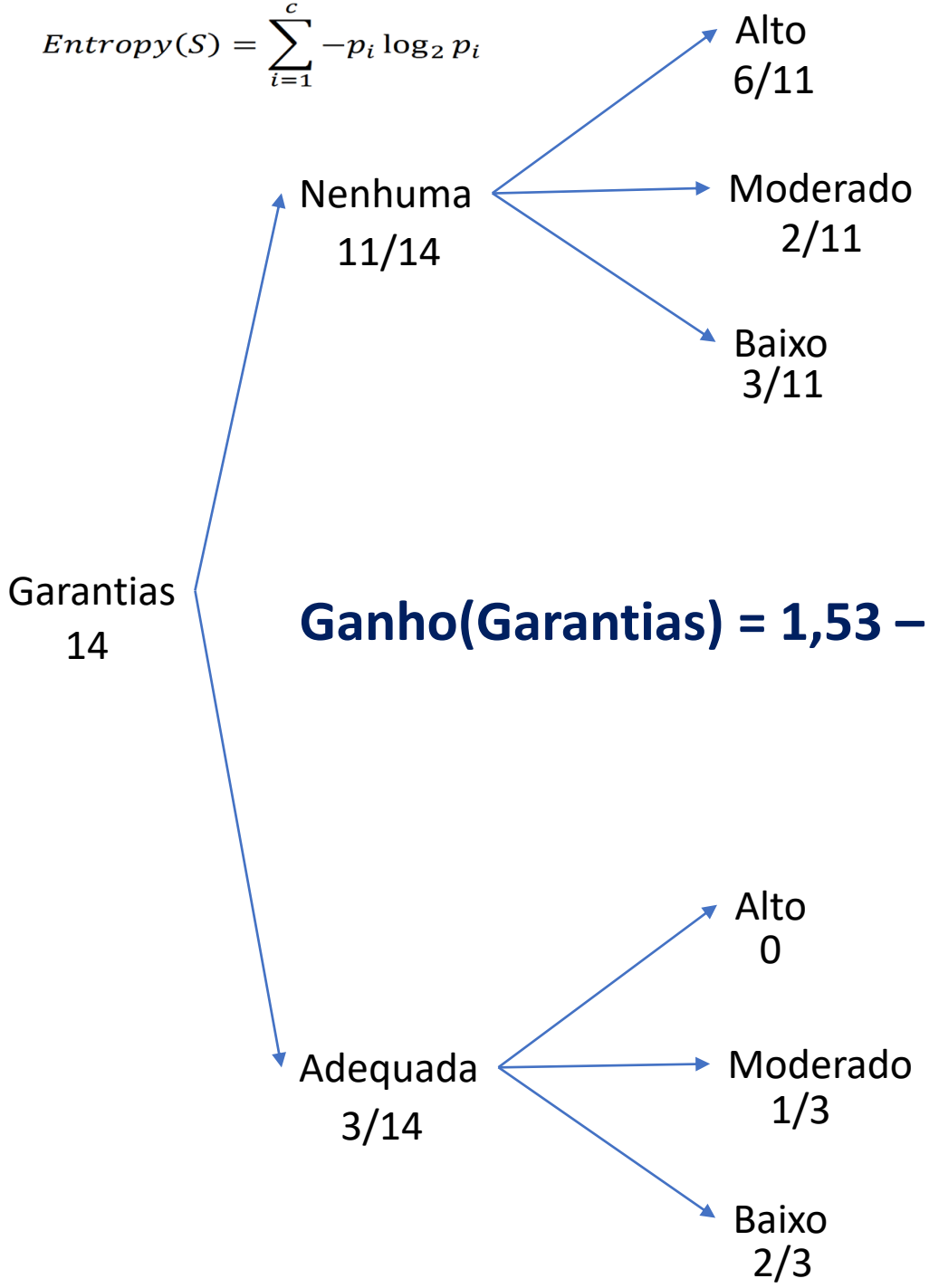
E(s) = -2/7 * log(2/7; 2) – 2/7 * log(2/7; 2) – 3/7 * log(3/7; 2) = **1,56**

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

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

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

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



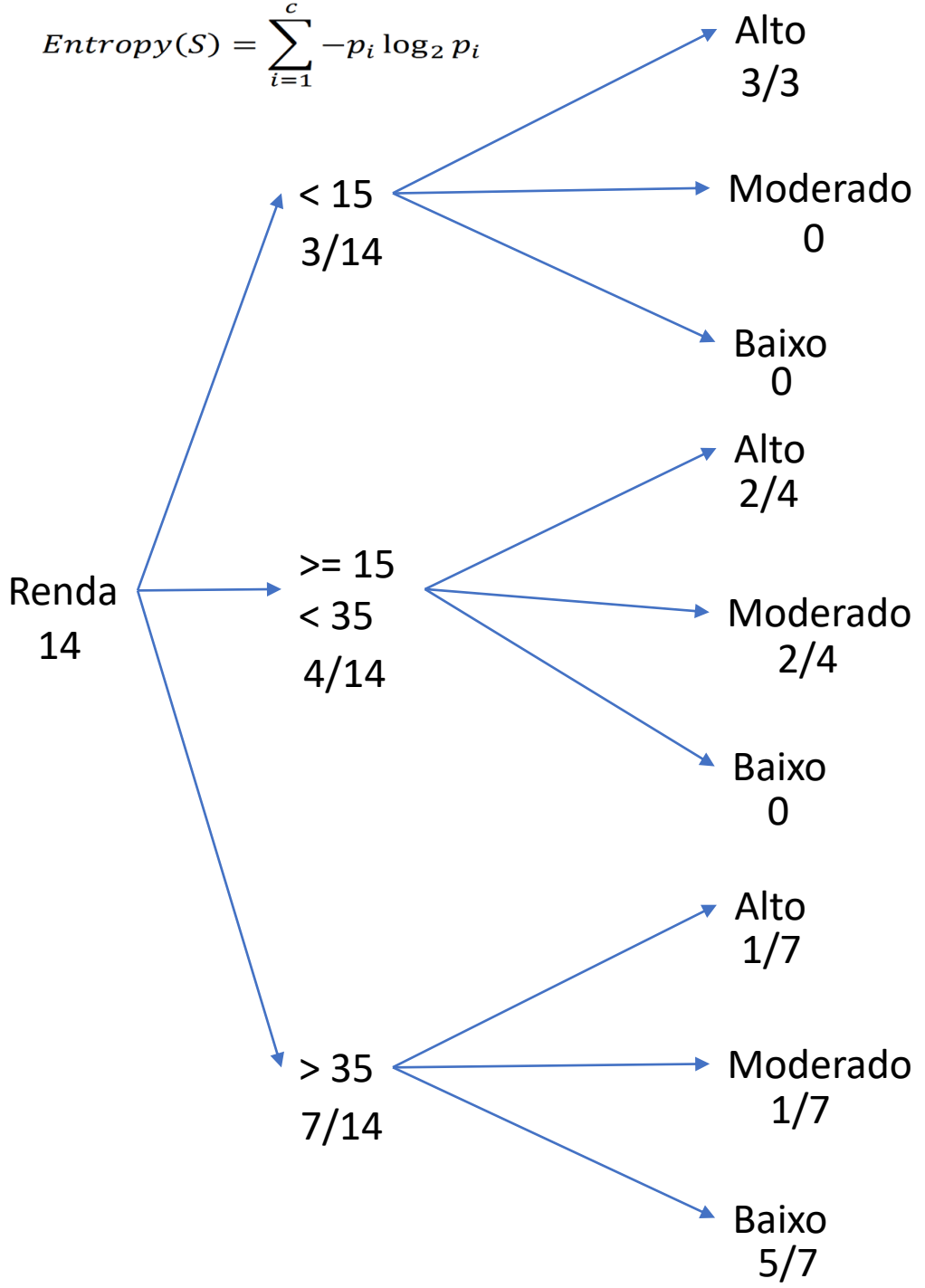
E(s) = -6/11 * log(6/11; 2) – 2/11 * log(2/11; 2) – 3/11 * log(3/11; 2) = **1,44**

E(s) = -0 * log(0; 2) – 1/3 * log(1/3; 2) – 2/3 * log(2/3; 2) = **0,92**

Ganho(Garantias) = 1,53 – (11/14 * 1,44) – (3/14 * 0,92) = 0,20

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

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



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

E(s) = -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

Com isso, ao definir o ganho de informação de cada atributo, conseguimos dizer qual é mais importante (o com maior número), e então realizar a árvore de decisão de acordo com a ordem, qual a "renda" será a raiz da árvore pois há o maior ganho

< 15

>= 15
< 35

> 35

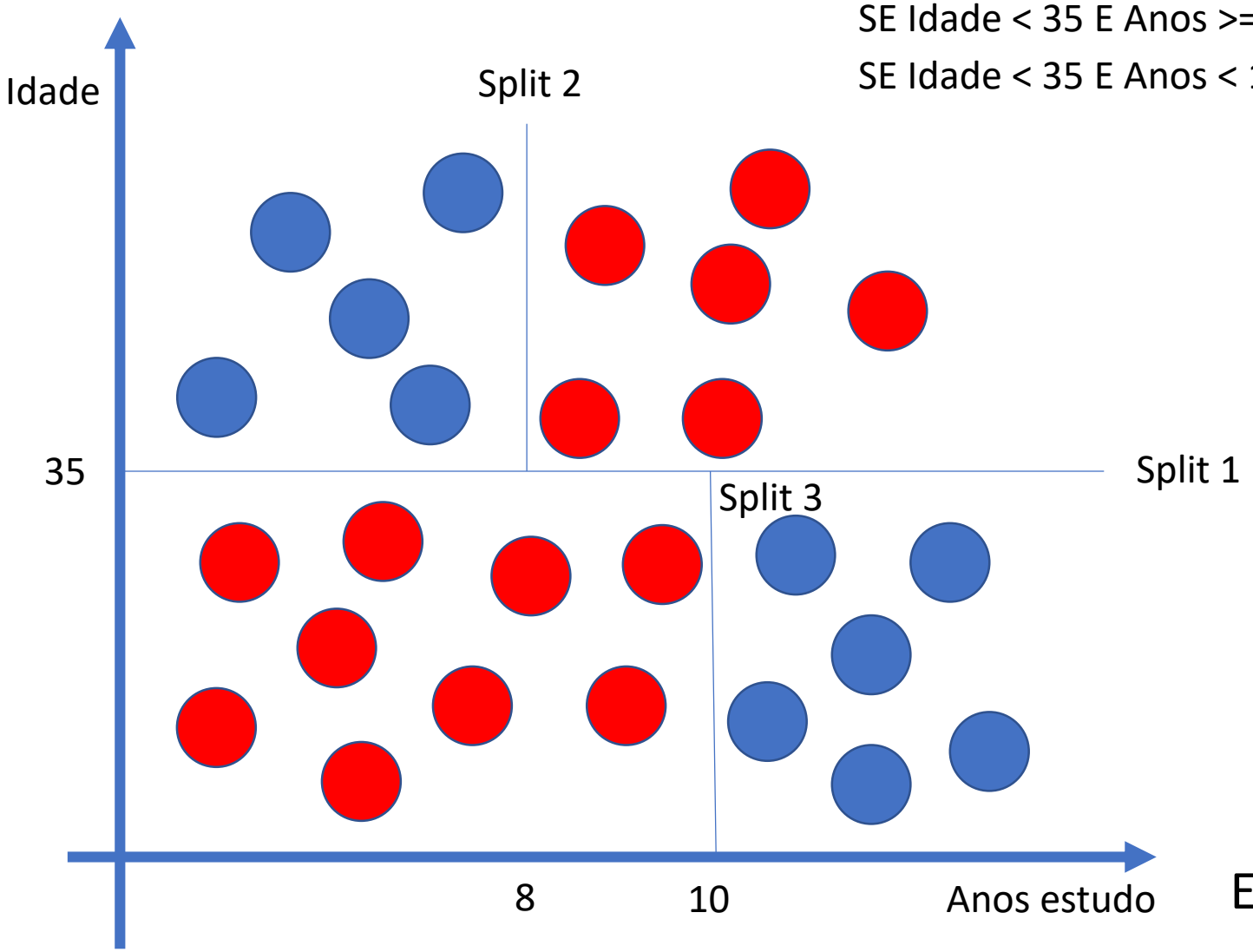
História do crédito	Dívida	Garantias	Renda anual	Risco
Ruim	Alta	Nenhuma	< 15.000	Alto
Ruim	Baixa	Nenhuma	< 15.000	Alto
Boa	Alta	Nenhuma	< 15.000	Alto

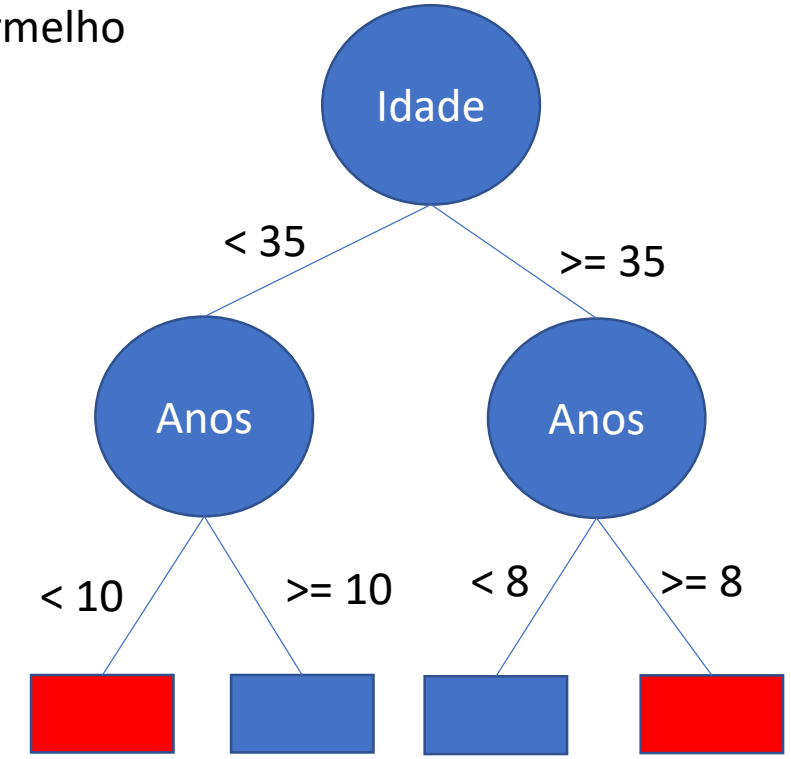
Refaremos os cálculos da entropia geral em cada ramo da árvore, de acordo com o valor da renda

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

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

SE Idade >= 35 E Anos >= 8 ENTÃO Vermelho
SE Idade >= 35 E Anos < 8 ENTÃO Azul
SE Idade < 35 E Anos >= 10 ENTÃO Azul
SE Idade < 35 E Anos < 10 ENTÃO Vermelho

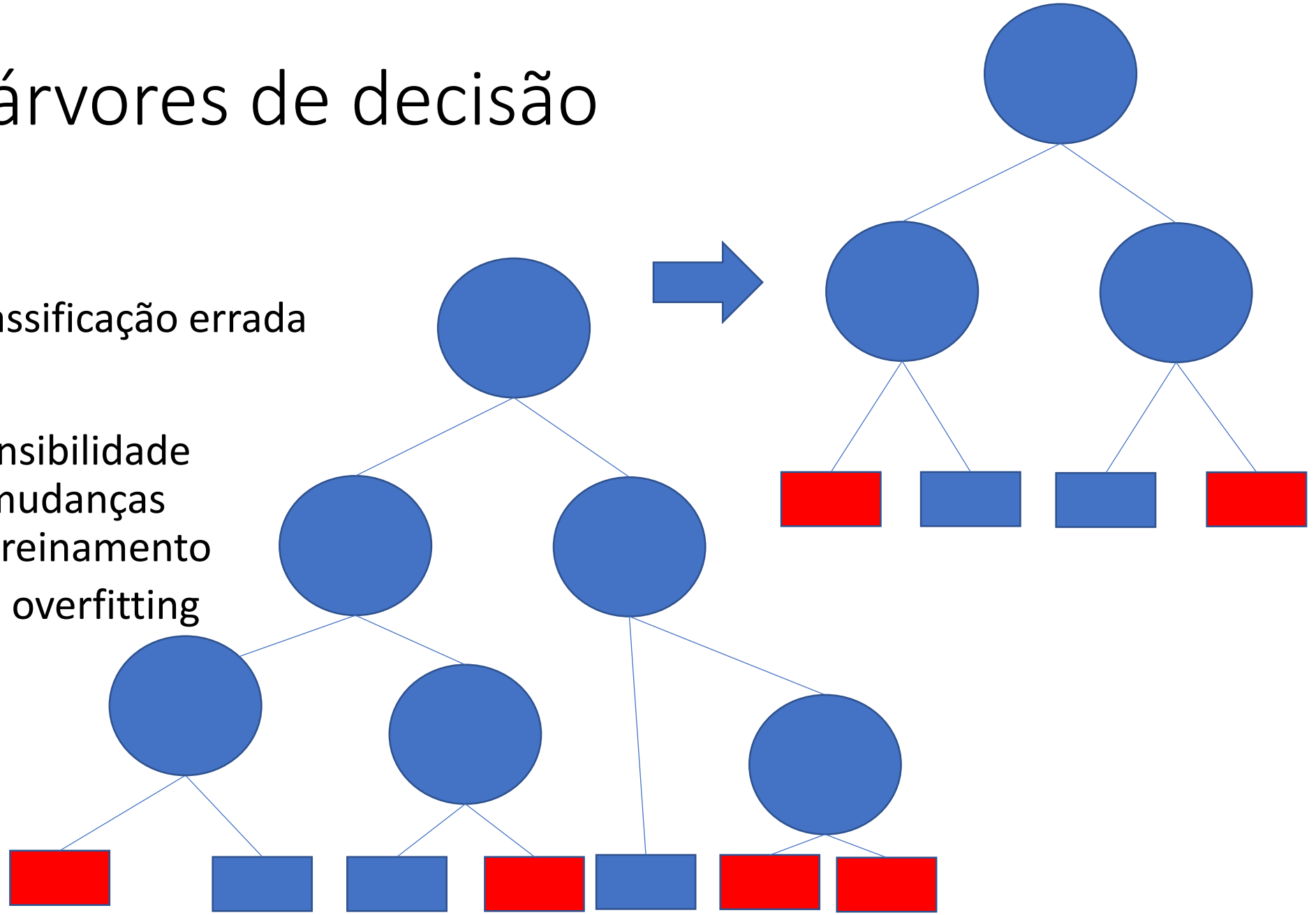




Encontrar o melhor conjunto de divisores

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

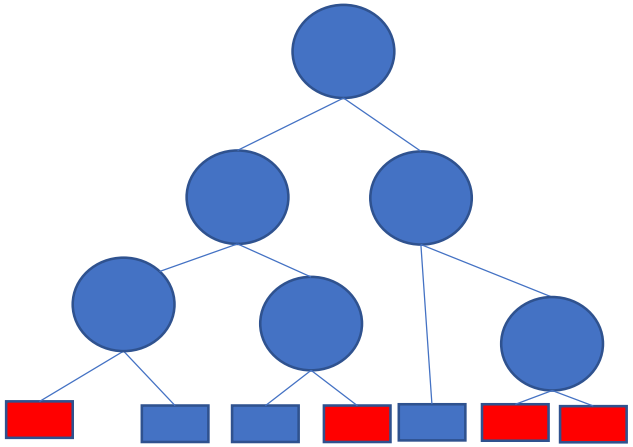
- Vantagens
  - Fácil interpretação
  - Não precisa normalização ou padronização
  - Rápido para classificar novos registros
- Desvantagens
  - Geração de árvores muito complexas
  - Pequenas mudanças nos dados pode mudar a árvore (poda pode ajudar)
  - Problema NP-completo para construir a árvore
- 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)

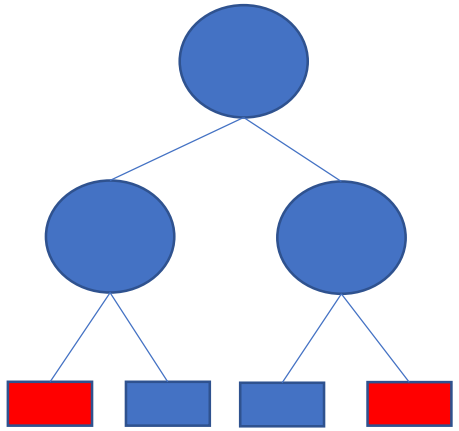
Random Forest

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
Ruim	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto

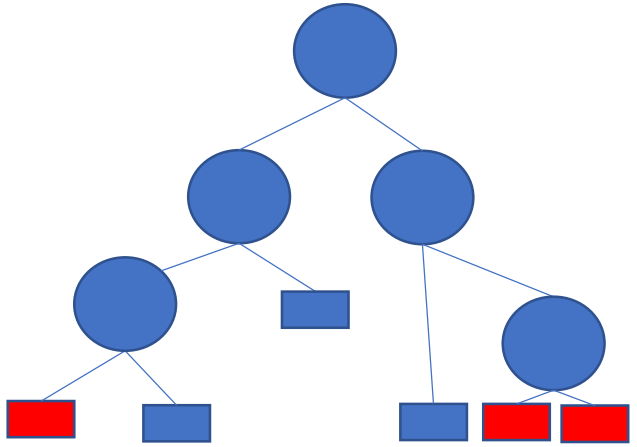
- Ensemble learning (aprendizagem em conjunto)
  - “Consultar diversos profissionais para tomar uma decisão”
  - Vários algoritmos juntos para construir um algoritmo mais “forte”
  - Usa a média (regressão) ou votos da maioria (classificação) para dar a resposta final



Risco = Alto



Risco = Baixo



Risco = Baixo

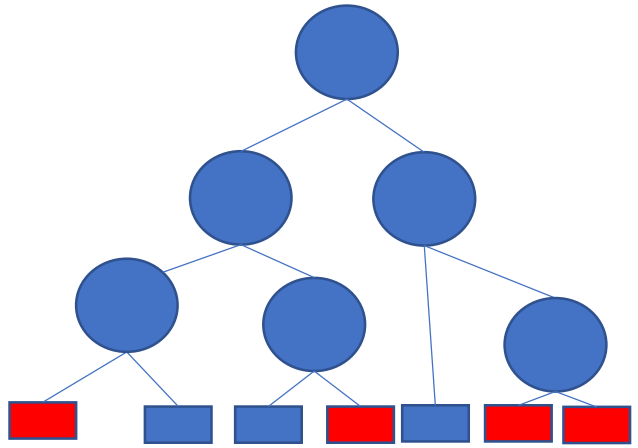
Random Forest

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	> 35.0000	Baixo
Ruim	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto

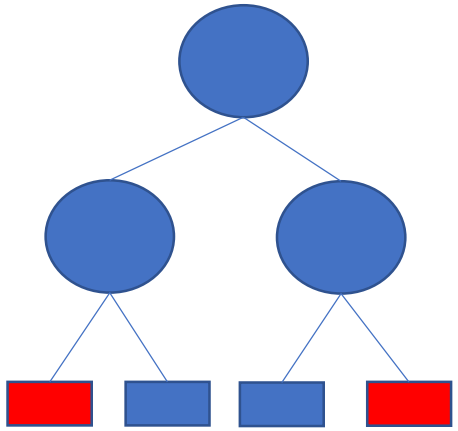
Escolhe de forma aleatória K atributos para comparação da métrica de pureza/impureza (impureza de gini/entropia)

K = 3

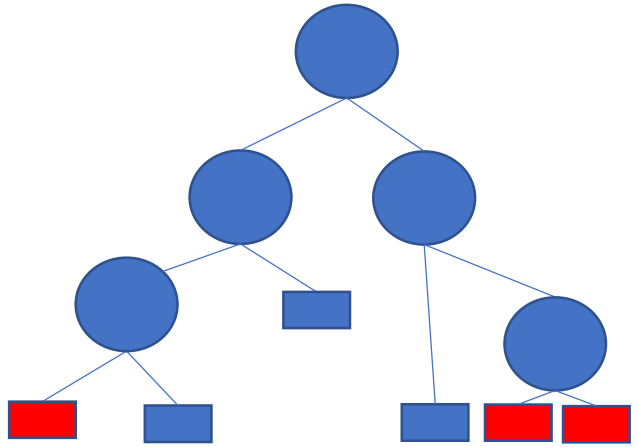
Árvores = 3



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