Árvore de Decisão

Quem sou eu?

Mathias Cesar

Graduado em Processamento de Dados pela UEG;

Especialização em Banco de Dados com ênfase em Ciências de Dados pela UFG;

Aluno da Escola Livre de IA;

Aluno do Mestrado em Ciência da Computação (Especial)

GOIAN **ROCK'IN ROLL**

DECISIONTREE

DECISIONTREE

1 – Teoria

DECISIONTREE

DECISIONTREE

CARACTERÍSTICAS

Método Preditivo

Algoritmo SUPERVISIONADO

CLASSIFICAÇÃO E REGRESSÃO

Existem muitos tipos de algoritmos de Aprendizado de Árvore de Decisão

Exemplos: ID3, C4.5, CART.....

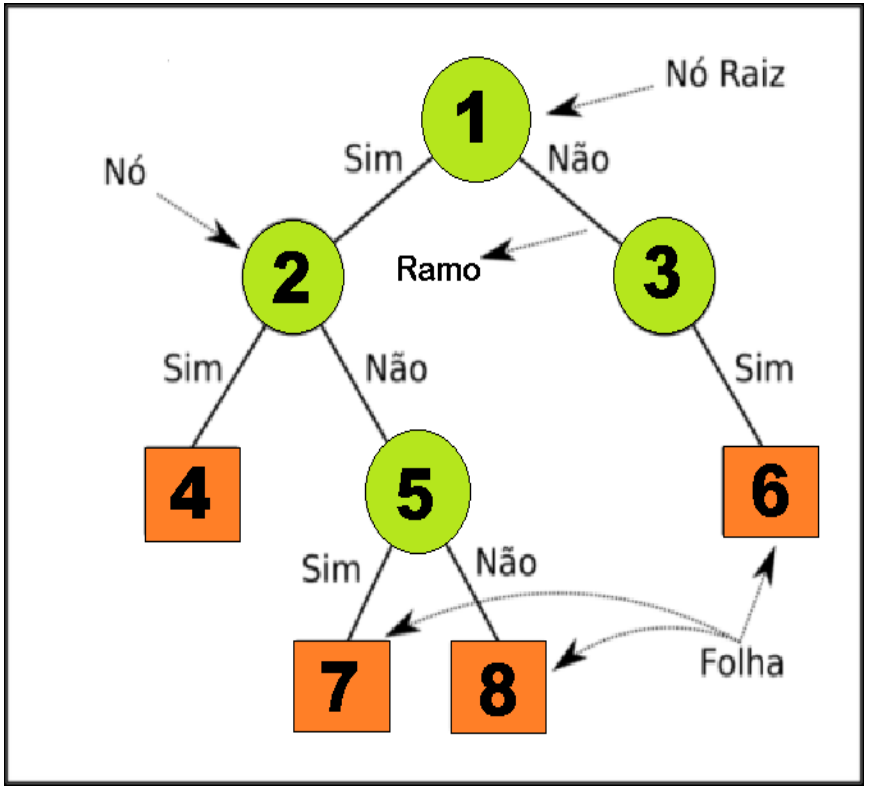
DECISIONTREE

REPRESENTAÇÃO

NÓ – responsável por testar o atributo;

RAMO – corresponde ao valor do atributo;

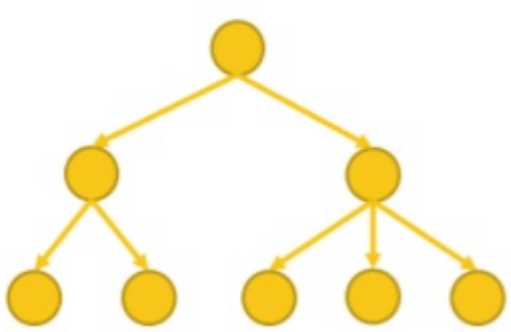
FOLHA – atribui o resultado final ou seja, a classificação ou a regressão)



DECISIONTREE – ID3

A principal característica do algoritmo ID3, é que ele APRENDE, construindo árvore de cima para baixo, no modelo (TOP-DOWN)

(IF – THEN)



DECISIONTREE – ID3

A grande questão é?

Quem será o atributo raiz?

DECISIONTREE – ID3

Será aquele que tiver o

MELHOR/MAIOR

GANHO DE INFORMAÇÃO

DECISIONTREE

ENTROPIA

É uma medida de aleatoriedade (impureza) de uma variável. Usada para estimar a aleatoriedade da variável a prever.

Fórmula:

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

DECISIONTREE

GANHO DE INFORMAÇÃO

Permite criar um índice para medir qual o melhor atributo para ser o primeiro Node da árvore.

Fórmula:

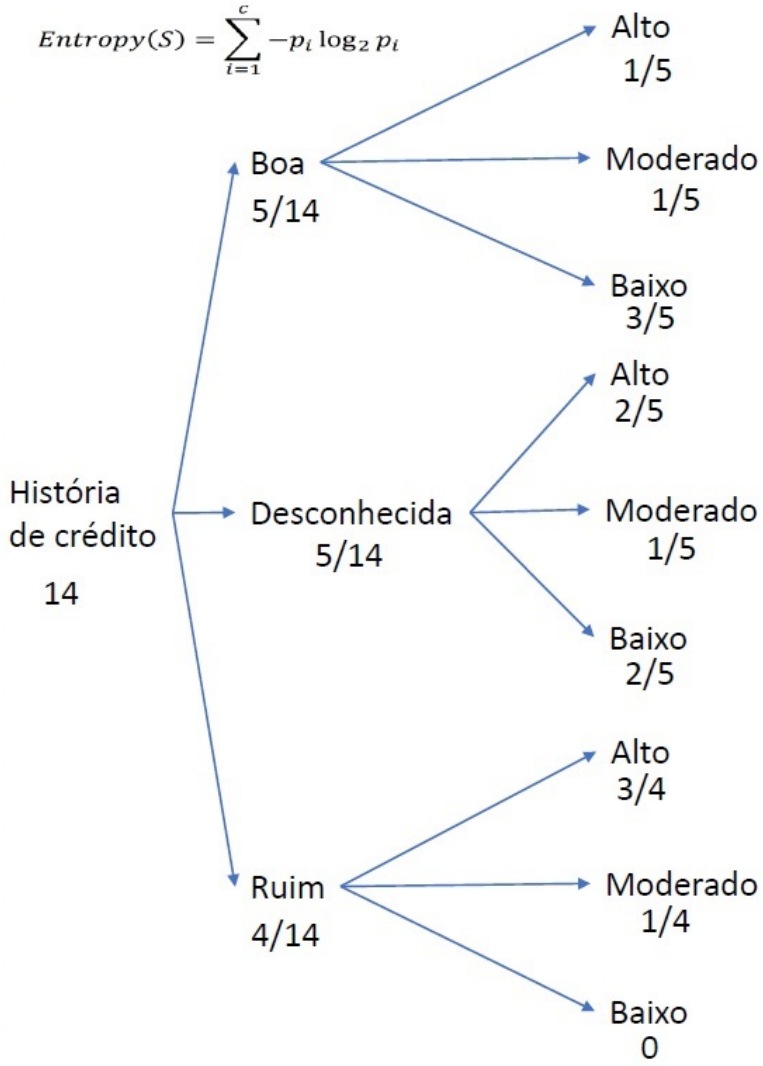
$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_{v)}$$

Base de dados Risco de Crédito

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 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

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



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

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

E(s) = -3/4 * log(3/4; 2) – 1/4 * log(1/4; 2) – 0 * log(0; 2) = **0.81**

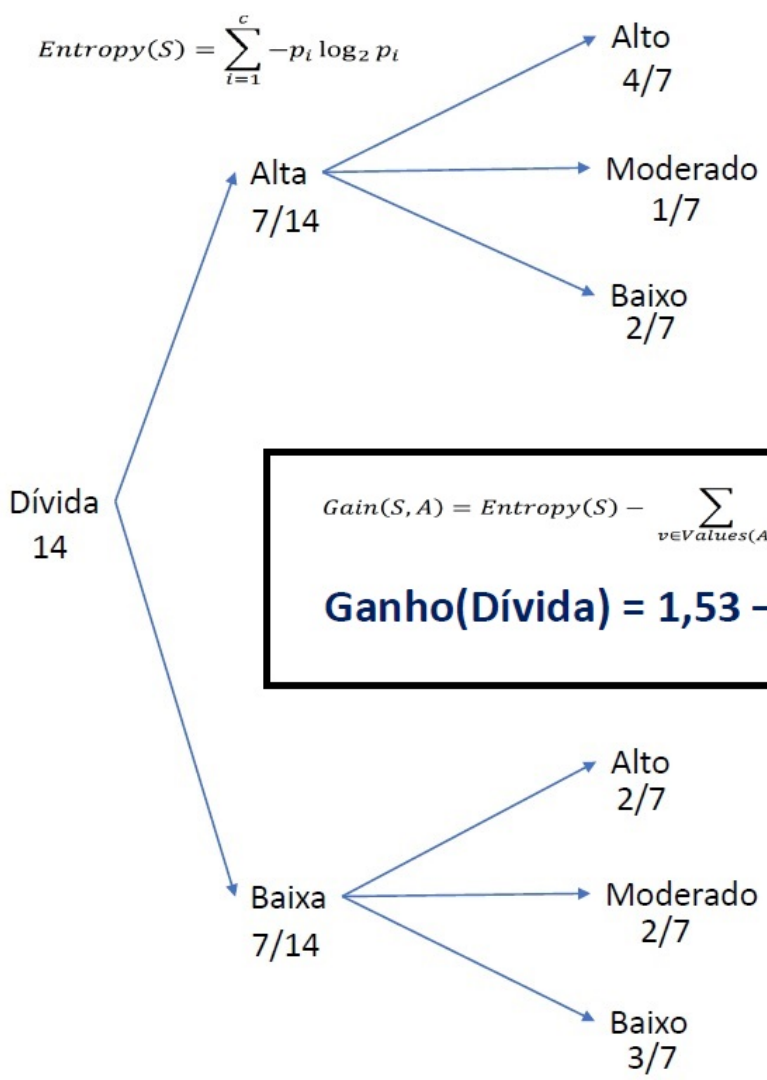
GANHO DE INFORMAÇÃO

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

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

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

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



E(s) = -4/7 * log(4/7; 2) – 1/7 * log(1/7; 2) – 2/7 * log(2/7; 2) = **1,38**

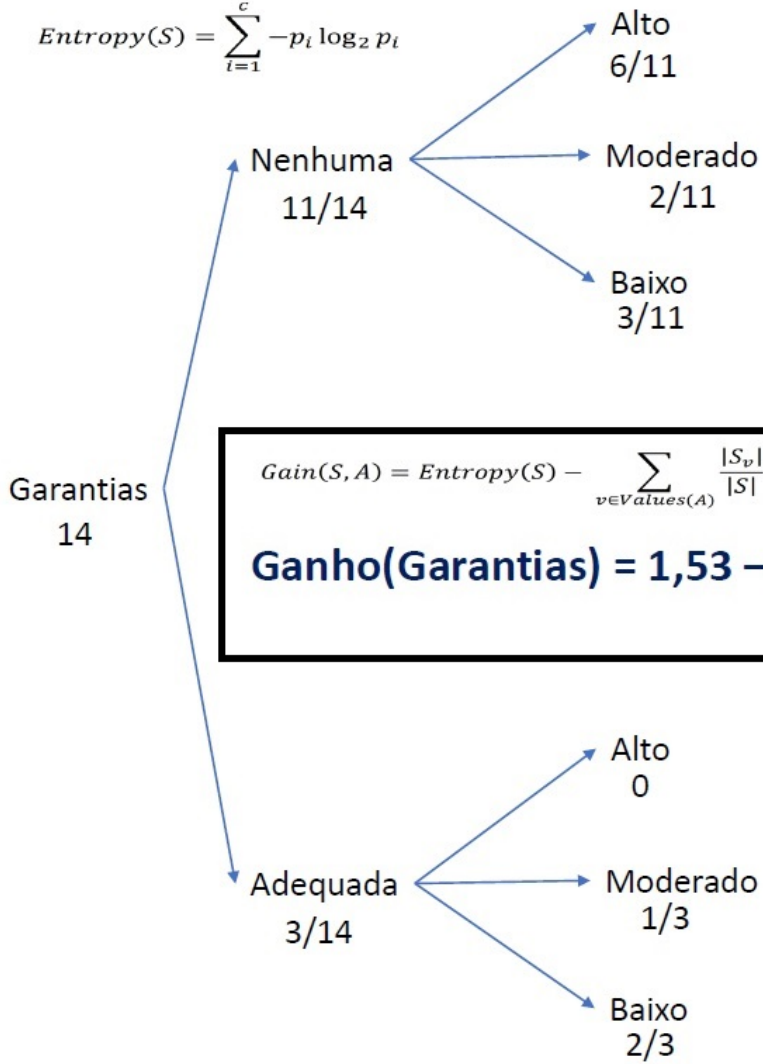
E(s) = -2/7 * log(2/7; 2) – 2/7 * log(2/7; 2) – 3/7 * log(3/7; 2) = **1,56**

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

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

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



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

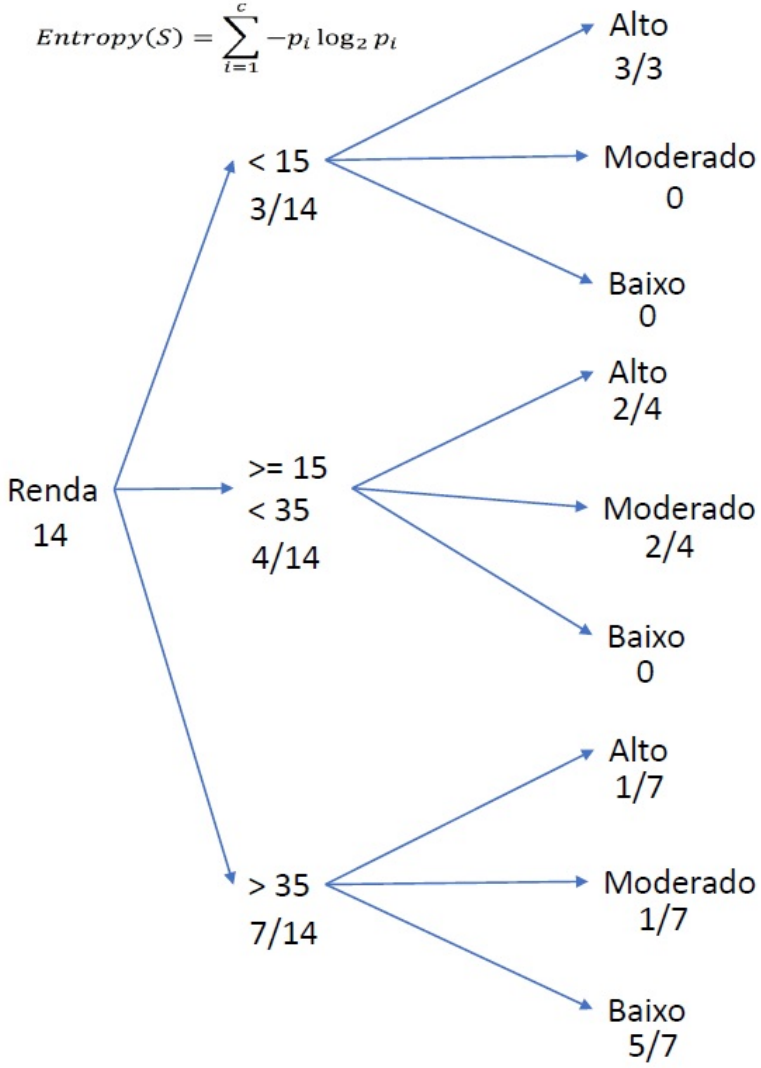
$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

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

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

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

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



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**

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

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

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

GANHO DE INFORMAÇÃO

Renda:	**0,66**
História:	0,26
Garantias:	0,20
Dívida:	0,06

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

< 15

>= 15
< 35

> 35

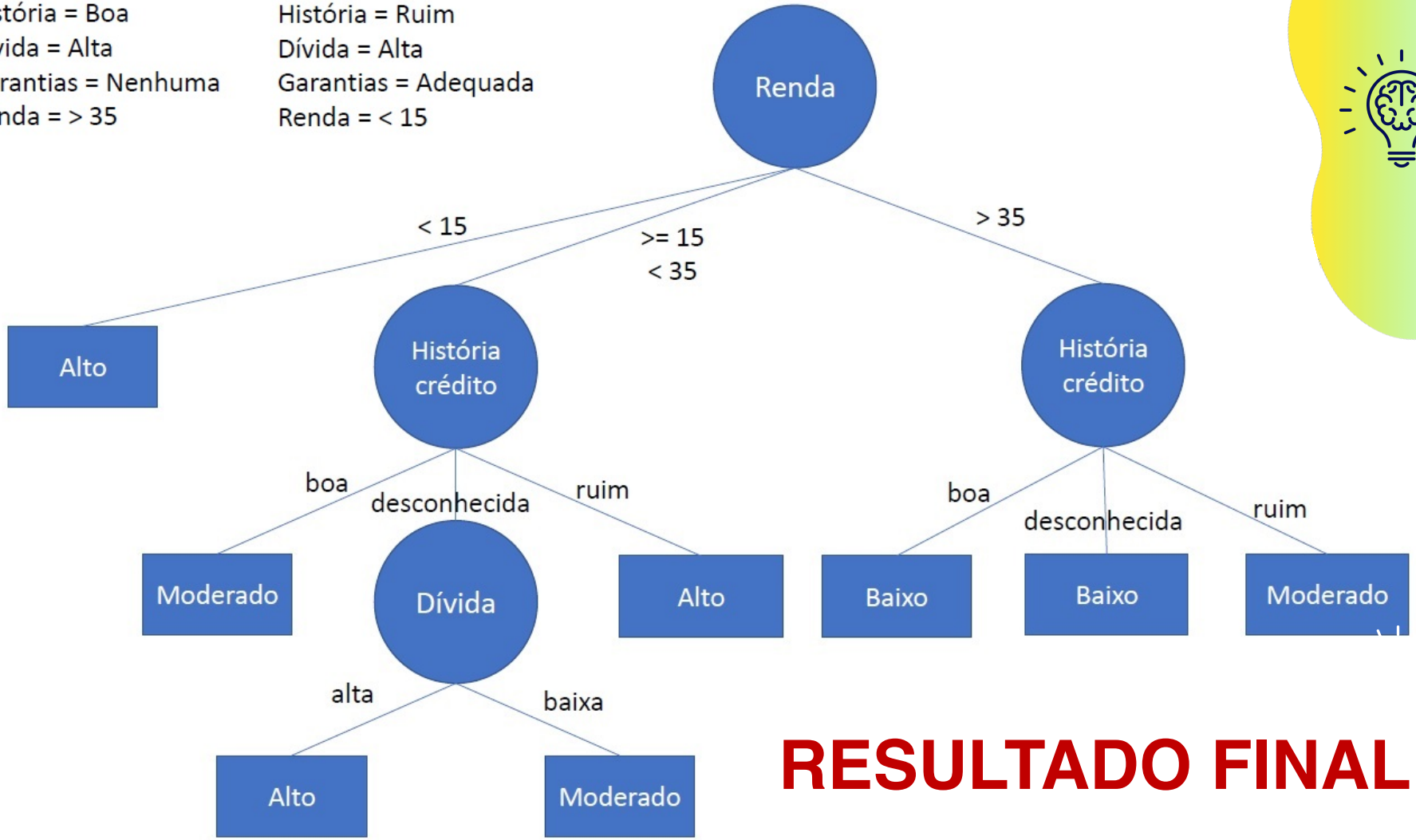
História do crédito	Dívida	Garantias	Renda anual	Risco
Ruim	Alta	Nenhuma	< 15.000	Alto
Ruim	Baixa	Nenhuma	< 15.000	Alto
Boa	Alta	Nenhuma	< 15.000	Alto

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

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

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



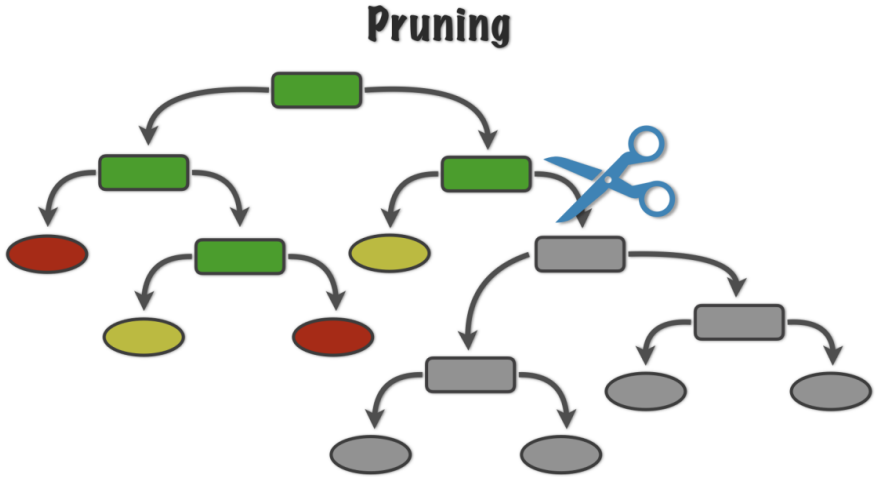
RESULTADO FINAL

DECISIONTREE

PODA DA ÁRVORE

É uma forma de controle da profundidade da árvore,

afim de garantir que não haja o overfitting.



DecisionTree

Árvore de Decisão

RANDOM FOREST

Ensemble Learning - **é uma técnica de aprendizado de máquina que combina o resultado de múltiplos modelos em busca de produzir um melhor modelo preditivo.**

Classificação – **Usa o voto da maioria (classificação) para dar o resultado final.**

Regressão – **Usa a média.**

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

Porque é Random?

Escolhe de forma aleatória K os atributos para comparação da métrica de pureza/impureza

K = 3

Árvores = 3

RANDOM FOREST

1

História
Dívida
Garantias

2

Renda
Dívida
Garantias

3

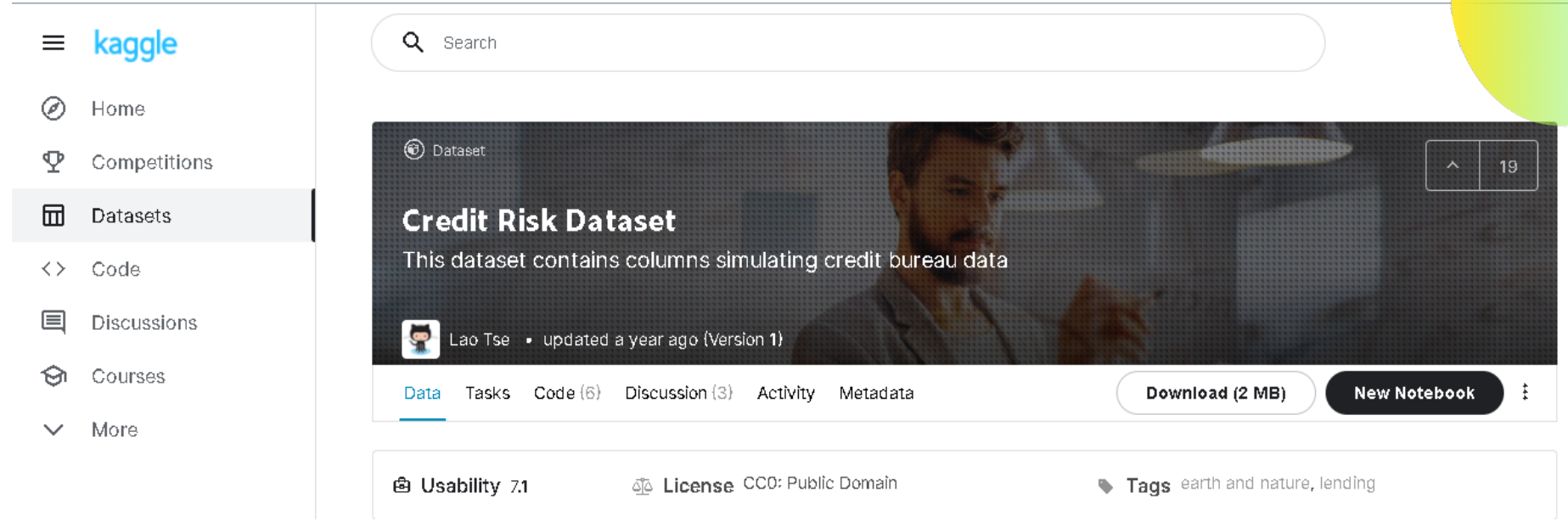
Renda
História
Dívida

RANDOM FOREST

Árvore de Decisão

Base de Dados

https://www.kaggle.com/laotse/credit-risk-dataset



CONTATOS

Email:

mathiascesar@discente.ufg.br

Linkedin:

www.linkedin.com/in/mathias-cesar-2941081a4/

Mathias Cesar

GOIAN

O

UFG
UNIVERSIDADE FEDERAL DE GOIÁS

ROCK'IN ROLL