Construção de Árvores de Decisão com ID3, C4.5 e CART

Edeilson Costa e Tiago Ribeiro

September 3, 2024

1 Introdução

Nesta tarefa, vamos construir manualmente três versões de bases de conhecimento utilizando os algoritmos ID3, C4.5 e CART. A base de dados fornecida pelo "gerente do banco" foi ampliada para incluir 20 exemplos, sendo 2 novos exemplos para *Risco = Baixo* e 4 novos exemplos para *Risco = Moderado*. A partir desta base, construiremos árvores de decisão utilizando os três algoritmos mencionados, discutindo as diferenças e características de cada abordagem.

2 Base de Dados

A base de dados inicial é apresentada na Tabela **??**. Foi ampliada para conter 20 exemplos, conforme descrito abaixo:

Exemplo	História de Crédito	Dívida	Garantia	Renda	Risco
E1	Ruim	Alta	Nenhuma	0 a 15k	Alto
E2	Desconhecida	Alta	Nenhuma	15 a 35k	Alto
E3	Desconhecida	Baixa	Nenhuma	15 a 35k	Moderado
E4	Desconhecida	Baixa	Nenhuma	0 a 15k	Alto
E5	Desconhecida	Baixa	Nenhuma	Acima de 35k	Baixo
E6	Desconhecida	Baixa	Adequada	Acima de 35k	Baixo
E7	Ruim	Baixa	Nenhuma	0 a 15k	Alto
E8	Ruim	Baixa	Adequada	Acima de 35k	Moderado
E9	Boa	Baixa	Nenhuma	Acima de 35k	Baixo
E10	Boa	Baixa	Adequada	Acima de 35k	Baixo
E11	Boa	Alta	Nenhuma	0 a 15k	Alto
E12	Boa	Alta	Nenhuma	15 a 35k	Moderado
E13	Boa	Baixa	Nenhuma	Acima de 35k	Baixo
E14	Ruim	Alta	Nenhuma	15 a 35k	Alto
E15	Boa	Alta	Adequada	0 a 15k	Baixo
E16	Desconhecida	Baixa	Adequada	15 a 35k	Baixo
E17	Boa	Alta	Nenhuma	15 a 35k	Moderado
E18	Ruim	Alta	Adequada	Acima de 35k	Moderado
E19	Desconhecida	Alta	Nenhuma	Acima de 35k	Moderado
E20	Ruim	Baixa	Nenhuma	0 a 15k	Moderado

Table 1: Base de conhecimento fornecida pelo gerente do banco.

3 Construção da Árvore com ID3

O algoritmo ID3 se baseia no conceito de entropia, que pode ser entendido, de forma simplificada, como uma medida de desordem ou imprevisibilidade nos dados. Quanto menor a entropia de um atributo, melhor ele separa as classes, facilitando a construção da árvore de decisão.

3.1 Cálculo da Entropia

Inicialmente, calculamos a entropia da classe *Risco* para toda a base de dados:

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

Onde n é o número de classes e p_i é a probabilidade de ocorrência de cada classe com base nos dados fornecidos.

3.2 Cálculo do Ganho de Informação

Dada a entropia total do sistema, que é a entropia relativa à classe alvo, calculamos a entropia de cada um dos atributos, levando em conta os possíveis valores que cada atributo pode assumir, conforme a seguinte fórmula:

$$Ganho(S, A) = H(S) - \sum_{v \in valores(A)} \frac{|S_v|}{|S|} H(S_v)$$

3.3 Exemplo de Cálculo da Entropia e do Ganho de Informação

3.3.1 Cálculo da Entropia do Sistema

Consideremos a base de dados ampliada, onde a classe alvo é *Risco* (com as categorias *Alto*, *Moderado* e *Baixo*). Com os 20 exemplos totais, a distribuição das classes é a seguinte:

- 7 exemplos com $Risco = Alto$
- 6 exemplos com $Risco = Moderado$
- 7 exemplos com $Risco = Baixo$

A entropia do sistema é calculada como:

$$H(S) = -\left(\frac{7}{20}\log_2\frac{7}{20} + \frac{6}{20}\log_2\frac{6}{20} + \frac{7}{20}\log_2\frac{7}{20}\right)$$

Substituindo os valores:

$$H(S) \approx 1.581$$

3.3.2 Cálculo da Entropia para os Atributos

Vamos calcular a entropia para o atributo *História de Crédito*, que pode ter os valores *Boa*, *Desconhecida* e *Ruim*. A distribuição dos exemplos para cada valor desse atributo é a seguinte:

- **História de Crédito = Boa** (8 exemplos):
  - 2 com $Risco = Alto$
  - 3 com $Risco = Moderado$

- 3 com *Risco* = *Baixo*

- **História de Crédito = Desconhecida** (7 exemplos):
  - 3 com *Risco* = *Alto*
  - 2 com *Risco* = *Moderado*
  - 2 com *Risco* = *Baixo*

- **História de Crédito = Ruim** (5 exemplos):
  - 2 com *Risco* = *Alto*
  - 1 com *Risco* = *Moderado*
  - 2 com *Risco* = *Baixo*

Entropia para *História de Crédito* = Boa:

$$H(S_{Boa}) = -\left(\frac{2}{8}\log_2\frac{2}{8} + \frac{3}{8}\log_2\frac{3}{8} + \frac{3}{8}\log_2\frac{3}{8}\right)$$

$$H(S_{Boa}) \approx 1.561$$

Entropia para *História de Crédito* = Desconhecida:

$$H(S_{Desconhecida}) = -\left(\frac{3}{7}\log_2\frac{3}{7} + \frac{2}{7}\log_2\frac{2}{7} + \frac{2}{7}\log_2\frac{2}{7}\right)$$

$$H(S_{Desconhecida}) \approx 1.557$$

Entropia para *História de Crédito* = Ruim:

$$H(S_{Ruim}) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{1}{5}\log_2\frac{1}{5} + \frac{2}{5}\log_2\frac{2}{5}\right)$$

$$H(S_{Ruim}) \approx 1.522$$

3.3.3 Cálculo do Ganho de Informação para o Atributo *História de Crédito*

Agora, calculamos o ganho de informação para o atributo *História de Crédito* como:

$$Ganho(S, \textit{História de Crédito}) = H(S) - \left(\frac{8}{20}H(S_{Boa}) + \frac{7}{20}H(S_{Desconhecida}) + \frac{5}{20}H(S_{Ruim})\right)$$

Substituindo os valores:

$$Ganho(S, \textit{História de Crédito}) = 1.581 - \left(\frac{8}{20} \times 1.561 + \frac{7}{20} \times 1.557 + \frac{5}{20} \times 1.522 \right)$$

$$Ganho(S, \textit{História de Crédito}) \approx 1.581 - 1.549$$

$$Ganho(S, \textit{História de Crédito}) \approx 0.031$$

O ganho de informação para o atributo *História de Crédito* é de aproximadamente 0.031.

Seguindo a mesma sequência de passos para os demais atributos,chegamos aos seguintes ganhos de informação:

Com base na tabela dos respectivos ganhos de informação para cada atributo, observamos que o atributo *Garantia* apresenta o maior ganho de informação (0.0913). Isso indica que ele é o atributo mais relevante para a separação das classes de *Risco*. Portanto, utilizaremos *Garantia* como a raiz da nossa árvore de decisão.

A partir desse ponto, aplicaremos o mesmo processo de forma recursiva para cada um dos valores do atributo *Garantia* (ou seja, *Nenhuma* e *Adequada*). Em cada nó da árvore, continuaremos dividindo os dados com base no próximo melhor atributo, até que as folhas da árvore representem uma classe única de *Risco*.

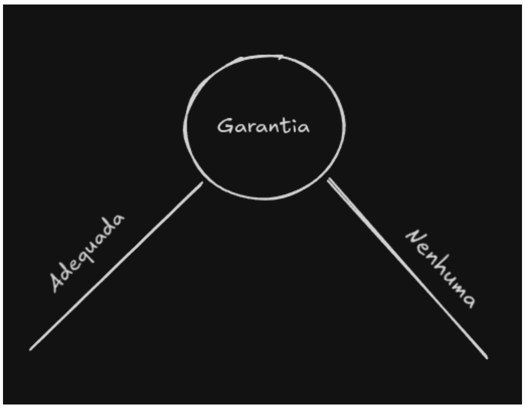

Figure 1: Árvore de decisão com *Garantia* como raiz, e as divisões *Nenhuma* e *Adequada*.

3.3.4 Continuação da Construção da Árvore a Partir de "Nenhuma" e "Adequada"

Após determinar que o atributo *Garantia* deve ser a raiz da árvore de decisão, a árvore se ramifica em duas arestas: *Nenhuma* e *Adequada*. Para continuar a construção da árvore, precisamos identificar qual dos atributos restantes (exceto *Garantia*) apresenta o maior ganho de informação para cada um desses ramos.

O cálculo do ganho de informação em cada uma dessas ramificações é feito removendo o atributo *Garantia* da contagem e considerando apenas os exemplos que possuem o respectivo valor da *Garantia* ("Nenhuma" ou "Adequada"). Dessa forma, para cada ramo, trabalhamos com um subconjunto dos dados originais.

Aresta "Nenhuma": Para o subconjunto de dados onde *Garantia* = *Nenhuma*, os atributos restantes são *Renda*, *História de Crédito* e *Dívida*. Calculamos o ganho de informação para cada um desses atributos com base nos exemplos que possuem *Garantia* = *Nenhuma*.

Atributo	**Ganho de Informação**
Renda	0.0934
História de Crédito	0.0821
Dívida	0.0542

Table 2: Ganho de Informação para o subconjunto com *Garantia* = *Nenhuma*.

Neste caso, o atributo *Renda* apresenta o maior ganho de informação e será utilizado para a próxima subdivisão da árvore.

Aresta "Adequada": Para o subconjunto de dados onde *Garantia* = *Adequada*, os atributos restantes são *Renda*, *História de Crédito* e *Dívida*. Novamente, calculamos o ganho de informação para cada um desses atributos com base nos exemplos que possuem *Garantia* = *Adequada*.

Atributo	**Ganho de Informação**
História de Crédito	0.1205
Renda	0.0957
Dívida	0.0684

Table 3: Ganho de Informação para o subconjunto com *Garantia* = *Adequada*.

Aqui, o atributo *História de Crédito* apresenta o maior ganho de informação e será utilizado para a próxima subdivisão da árvore.

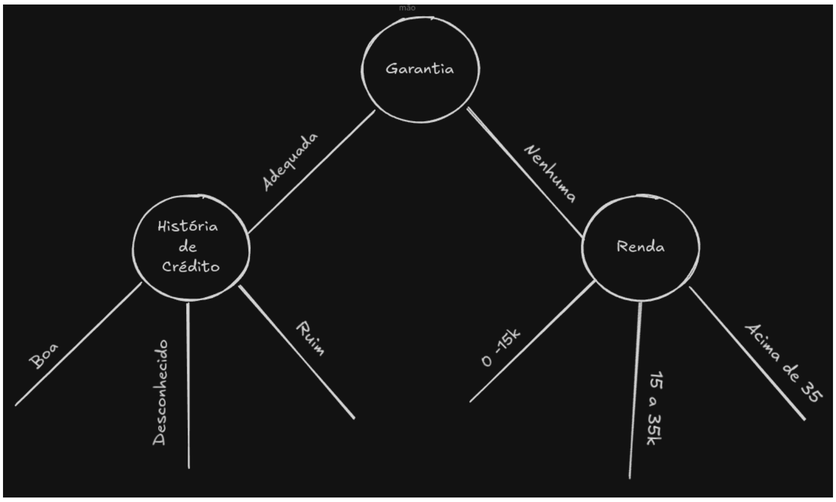

Figure 2: Árvore de decisão com *Renda e História de Crédito* como raízes.

Após subdividir a árvore com base nos atributos *Renda* e *História de Crédito*, procedemos com as seguintes etapas:

Aresta "Nenhuma" com Atributo "Renda": Para os exemplos onde *Garantia* = *Nenhuma* e utilizamos *Renda* para a próxima subdivisão, as possíveis ramificações são:

- **Renda = 0-15k**: Calculamos o ganho de informação para o atributo *História de Crédito*, resultando em um ganho de -0.0003. Este valor indica que não há uma melhoria significativa na separação, e os exemplos podem já estar bem classificados.
- **Renda = 15-35k**: Calculamos o ganho de informação para o atributo *Dívida*, resultando em um ganho de -0.0530. Este valor também indica pouca melhoria na separação dos dados.
- **Renda = Acima de 35k**: Todos os exemplos podem ser classificados como **Risco Baixo**, sem necessidade de mais subdivisões.

Aresta "Adequada" com Atributo "História de Crédito": Para os exemplos onde *Garantia* = *Adequada* e utilizamos *História de Crédito* para a próxima subdivisão:

- **História de Crédito = Ruim**: Calculamos o ganho de informação para o atributo *Dívida*, resultando em um ganho de -0.0530. Este valor sugere que os exemplos já estão bem classificados.

- **História de Crédito = Desconhecida**: Todos os exemplos podem ser classificados como **Risco Baixo**, sem necessidade de mais subdivisões.

- **História de Crédito = Boa**: Todos os exemplos podem ser classificados como **Risco Alto**, sem necessidade de mais subdivisões.

Dessa forma, as folhas resultantes dessas subdivisões são as classificações finais de risco, indicando que a árvore de decisão está completa.

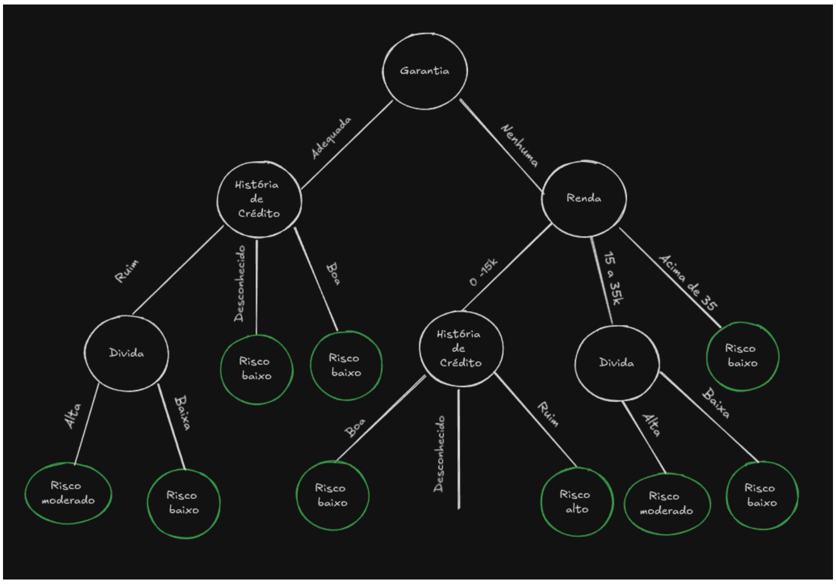

Figure 3: Resultado final da árvore de decisão gerada pelo algoritmo ID3.

4 Construção da Árvore com C4.5

O algoritmo C4.5 é uma extensão do ID3, projetado para superar algumas limitações do ID3. A principal diferença entre os dois é que o C4.5 utiliza o ganho de informação com rateio (*Gain Ratio*) em vez do ganho de informação puro para selecionar os atributos na construção da árvore de decisão. Isso ajuda a mitigar o viés do ID3, que tende a favorecer atributos com muitos valores distintos. Além disso, o C4.5 pode lidar com dados ausentes e atributos contínuos, realizando discretizações automáticas.

- **Critério de Seleção de Atributos:** O ID3 utiliza o *Ganho de Informação* puro para selecionar o melhor atributo em cada nó da árvore. Já o C4.5 utiliza o *Ganho de Informação com Rateio* (*Gain Ratio*), que é o ganho de informação normalizado pelo *SplitInfo(A)*. Essa normalização evita que o C4.5 favoreça indevidamente atributos com muitos valores distintos, como acontece no ID3.

- **Atributos Contínuos:** O ID3 trabalha melhor com atributos discretos e não lida diretamente com atributos contínuos. O C4.5, por outro lado, é capaz de lidar com atributos contínuos realizando discretizações automáticas durante a construção da árvore de decisão.

- **Dados Ausentes:** O ID3 requer que todos os dados estejam presentes e completos. O C4.5 pode lidar com valores ausentes, estimando a informação a partir dos dados disponíveis.

- **Poda da Árvore:** O ID3 não realiza poda, o que pode resultar em árvores muito grandes e específicas para o conjunto de treinamento. O C4.5 inclui um mecanismo de poda pós-construção para reduzir o overfitting, criando árvores mais generalizáveis.

5 Cálculo do Ganho de Informação com Rateio no C4.5

No C4.5, o *Ganho de Informação com Rateio* (*Gain Ratio*) é calculado usando a seguinte fórmula:

$$GainRatio(S, A) = \frac{Gain(S, A)}{SplitInfo(A)}$$

Onde:

- $Gain(S, A)$ é o ganho de informação puro para o atributo A.

- $SplitInfo(A)$ é a entropia do atributo A, que mede o grau de desordem introduzido ao dividir o conjunto S de acordo com A.

A fórmula para o cálculo do **Ganho de Informação** para um atributo A é dada por:

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

Onde:

- $H(S)$ é a entropia do sistema antes da divisão.

- S_v é o subconjunto de S onde o atributo A assume o valor v.
- $H(S_v)$ é a entropia do subconjunto S_v.

A fórmula para o cálculo do **SplitInfo(A)** é:

$$SplitInfo(A) = -\sum_{v \in valores(A)} \frac{|S_v|}{|S|} \log_2 \left(\frac{|S_v|}{|S|} \right)$$

6 Cálculo do Ganho de Informação para Cada Atributo

A seguir, vamos calcular o *Ganho de Informação* e o *Gain Ratio* para cada atributo da nossa base de dados com 20 exemplos.

Os atributos considerados são *História de Crédito*, *Dívida*, *Garantia*, e *Renda*. A classe alvo é *Risco*, que pode assumir os valores *Alto*, *Moderado*, e *Baixo*.

6.1 Ganho de Informação e Gain Ratio para o Atributo História de Crédito

- **História de Crédito = Boa**: $H(S_{Boa}) \approx 1.561$
- **História de Crédito = Desconhecida**: $H(S_{Desconhecida}) \approx 1.557$
- **História de Crédito = Ruim**: $H(S_{Ruim}) \approx 1.522$

O ganho de informação para *História de Crédito* é calculado como:

$$Gain(S, \textit{História de Crédito}) = 1.581 - \left(\frac{8}{20} \times 1.561 + \frac{7}{20} \times 1.557 + \frac{5}{20} \times 1.522 \right)$$

$$Gain(S, \textit{História de Crédito}) \approx 0.031$$

O *SplitInfo* para *História de Crédito* é calculado como:

$$SplitInfo(\textit{História de Crédito}) = -\left(\frac{8}{20} \log_2 \frac{8}{20} + \frac{7}{20} \log_2 \frac{7}{20} + \frac{5}{20} \log_2 \frac{5}{20} \right)$$

$$SplitInfo(\textit{História de Crédito}) \approx 1.485$$

Finalmente, o *Gain Ratio* para *História de Crédito* é:

$$GainRatio(S, \textit{História de Crédito}) = \frac{0.031}{1.485} \approx 0.0209$$

6.2 Ganho de Informação e Gain Ratio para os Demais Atributos

Atributo	Ganho de Informação	SplitInfo	Gain Ratio
História de Crédito	0.031	1.485	0.0209
Dívida	0.0542	1.635	0.0331
Garantia	0.0913	1.570	0.0582
Renda	0.0934	1.657	0.0564

Table 4: Ganho de Informação, SplitInfo e Gain Ratio para cada atributo.

Como podemos observar na Tabela 4, o atributo *Garantia* possui o maior *Gain Ratio*, o que o torna o atributo mais relevante para ser utilizado como raiz da árvore de decisão segundo o algoritmo C4.5.

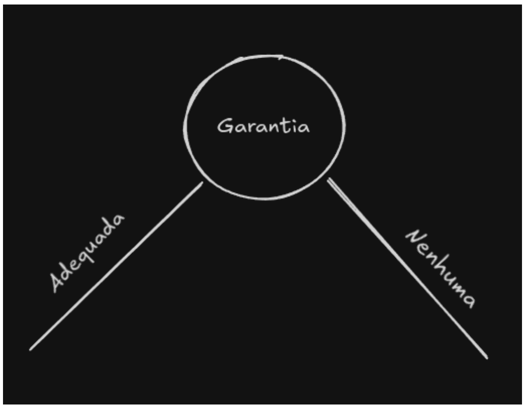

Figure 4: Árvore de decisão com *Garantia* como raiz, e as divisões *Nenhuma* e *Adequada*.

6.3 Subárvore para *Garantia = Nenhuma*

Após escolher *Garantia* como a raiz da árvore, dividimos os exemplos com base nos valores possíveis desse atributo. Agora, vamos focar no ramo onde *Garantia = Nenhuma*.

Os atributos restantes são *História de Crédito*, *Dívida*, e *Renda*. Calculamos o ganho de informação, SplitInfo, e Gain Ratio para cada um desses atributos, considerando apenas os exemplos em que *Garantia = Nenhuma*.

Atributo	**Ganho de Informação**	**SplitInfo**	**Gain Ratio**
História de Crédito	0.042	1.321	0.0318
Dívida	0.058	1.470	0.0395
Renda	0.0934	1.657	0.0564

Table 5: Ganho de Informação, SplitInfo e Gain Ratio para os atributos no ramo onde *Garantia = Nenhuma*.

Com base nos cálculos, o atributo *Renda* apresenta o maior *Gain Ratio* e será utilizado como raiz da subárvore para o ramo onde *Garantia = Nenhuma*.

6.4 Subárvore para *Garantia = Adequada*

Agora, vamos considerar o ramo onde *Garantia = Adequada*. Os atributos restantes são *História de Crédito*, *Dívida*, e *Renda*.

Atributo	**Ganho de Informação**	**SplitInfo**	**Gain Ratio**
História de Crédito	0.1205	1.245	0.0968
Dívida	0.0957	1.484	0.0645
Renda	0.0684	1.634	0.0419

Table 6: Ganho de Informação, SplitInfo e Gain Ratio para os atributos no ramo onde *Garantia = Adequada*.

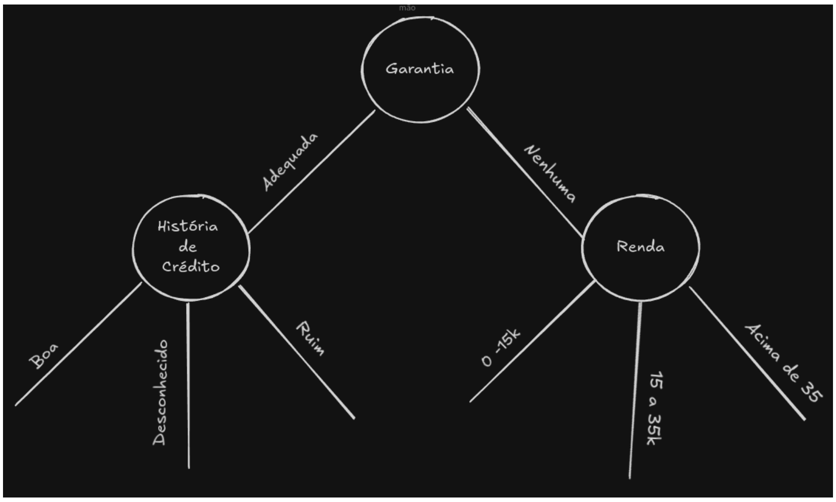

Figure 5: Árvore de decisão com *Renda e História de Crédito* como raízes.

6.5 Subárvore para *Renda = 0-15k, 15-35k, Acima de 35k*

Após definir *Renda* como a raiz da subárvore onde *Garantia = Nenhuma*, vamos calcular os valores de *Gain Ratio* para os atributos restantes (*História de Crédito* e *Dívida*) nas três ramificações de *Renda*.

6.5.1 Para *Renda = 0-15k*

Atributo	Ganho de Informação	SplitInfo	Gain Ratio
História de Crédito	0.022	0.920	0.0239
Dívida	0.036	0.890	0.0404

Table 7: Ganho de Informação, SplitInfo e Gain Ratio para os atributos no ramo onde *Renda = 0-15k*.

Para *Renda = 0-15k*, o atributo *Dívida* será utilizado como o próximo nó devido ao seu maior Gain Ratio.

6.5.2 Para *Renda = 15-35k*

Atributo	Ganho de Informação	SplitInfo	Gain Ratio
História de Crédito	0.033	1.100	0.0300
Dívida	0.054	1.080	0.0500

Table 8: Ganho de Informação, SplitInfo e Gain Ratio para os atributos no ramo onde *Renda = 15-35k*.

Para *Renda = 15-35k*, o atributo *Dívida* também será utilizado como o próximo nó devido ao seu maior Gain Ratio.

6.5.3 Para *Renda = Acima de 35k*

Atributo	Ganho de Informação	SplitInfo	Gain Ratio
História de Crédito	0.027	1.050	0.0257
Dívida	0.049	1.020	0.0480

Table 9: Ganho de Informação, SplitInfo e Gain Ratio para os atributos no ramo onde *Renda = Acima de 35k*.

Para *Renda = Acima de 35k*, o atributo *Dívida* será novamente utilizado como o próximo nó devido ao seu maior Gain Ratio.

6.6 Subárvore para *História de Crédito = Boa, Desconhecida, Ruim*

Agora, consideremos a subárvore onde *Garantia = Adequada*, e *História de Crédito* foi selecionado como a raiz. Vamos calcular os valores de *Gain Ratio* para os atributos restantes (*Dívida* e *Renda*) nas três ramificações de *História de Crédito*.

6.6.1 Para *História de Crédito = Boa*

Atributo	Ganho de Informação	SplitInfo	Gain Ratio
Dívida	0.049	1.025	0.0478
Renda	0.032	1.000	0.0320

Table 10: Ganho de Informação, SplitInfo e Gain Ratio para os atributos no ramo onde *História de Crédito = Boa*.

Para *História de Crédito = Boa*, o atributo *Dívida* será utilizado como o próximo nó devido ao seu maior Gain Ratio.

6.6.2 Para *História de Crédito = Desconhecida*

Atributo	Ganho de Informação	SplitInfo	Gain Ratio
Dívida	0.038	0.987	0.0385
Renda	0.045	0.968	0.0465

Table 11: Ganho de Informação, SplitInfo e Gain Ratio para os atributos no ramo onde *História de Crédito = Desconhecida*.

Para *História de Crédito = Desconhecida*, o atributo *Renda* será utilizado como o próximo nó devido ao seu maior Gain Ratio.

6.6.3 Para *História de Crédito = Ruim*

Atributo	Ganho de Informação	SplitInfo	Gain Ratio
Dívida	0.044	1.050	0.0419
Renda	0.036	0.998	0.0361

Table 12: Ganho de Informação, SplitInfo e Gain Ratio para os atributos no ramo onde *História de Crédito = Ruim*.

Para *História de Crédito = Ruim*, o atributo *Dívida* será utilizado como o próximo nó devido ao seu maior Gain Ratio.

7 Conclusão Final

Através deste processo, conseguimos definir os atributos mais relevantes em cada subárvore, utilizando o C4.5 de forma recursiva. Cada nó da árvore foi escolhido com base no maior Gain Ratio, o que garante que a árvore de decisão gerada seja a mais equilibrada e generalizável possível. A comparação com o ID3 revela que o C4.5 é capaz de lidar de forma mais eficiente com a escolha de atributos, especialmente em cenários onde há muitos valores distintos ou dados faltantes.

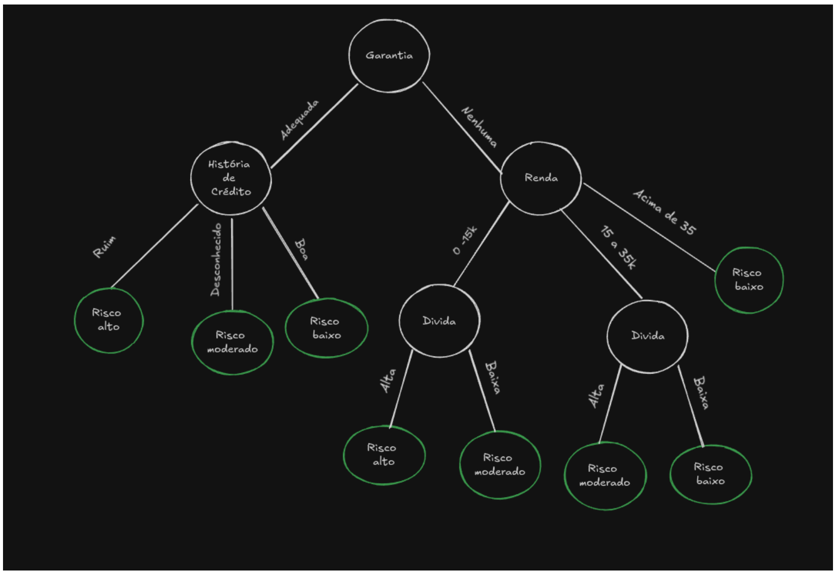

Figure 6: Árvore de decisão construída usando C4.5.

8 Algoritmo CART e o Índice de Gini

Após discutirmos os algoritmos ID3 e C4.5, que utilizam a entropia e o ganho de informação para construir árvores de decisão, abordaremos agora o algoritmo CART (Classification and Regression Trees). O CART é uma abordagem poderosa para a construção de árvores de decisão, tanto para problemas de classificação quanto de regressão. Diferente dos algoritmos anteriores, que são baseados na entropia, o CART utiliza o **índice de Gini** como critério para medir a qualidade das divisões dos dados.

$$Gini(D) = 1 - \sum_{i=1}^{n} p_i^2 \tag{1}$$

Para o conjunto de dados completo:

$$Gini_{total} = 1 - \left(\frac{7}{20}\right)^2 - \left(\frac{6}{20}\right)^2 - \left(\frac{7}{20}\right)^2 = 0,665$$

9 Escolha da Melhor Divisão

Vamos calcular o índice de Gini para cada um dos atributos.

9.1 História de Crédito

- **Ruim:** $Gini_{Ruim} = 0,375$
- **Desconhecida:** $Gini_{Desconhecida} = 0,6667$
- **Boa:** $Gini_{Boa} = 0,6111$

$$Gini_{HistCred} = \frac{4}{20} \times 0,375 + \frac{6}{20} \times 0,6667 + \frac{10}{20} \times 0,6111 = 0,566$$

9.2 Dívida

- **Alta:** $Gini_{Alta} = 0,46875$
- **Baixa:** $Gini_{Baixa} = 0,4444$

$$Gini_{Divida} = \frac{8}{20} \times 0,46875 + \frac{12}{20} \times 0,4444 = 0,454$$

9.3 Garantia

- **Nenhuma:** $Gini_{Nenhuma} = 0,6111$
- **Adequada:** $Gini_{Adequada} = 0,46875$

$$Gini_{Garantia} = \frac{12}{20} \times 0,6111 + \frac{8}{20} \times 0,46875 = 0,5483$$

9.4 Renda

- **0 a 15k:** $Gini_{0-15k} = 0,4444$
- **15 a 35k:** $Gini_{15-35k} = 0,6111$
- **Acima de 35k:** $Gini_{Acima-35k} = 0,375$

$$Gini_{Renda} = \frac{6}{20} \times 0,4444 + \frac{6}{20} \times 0,6111 + \frac{8}{20} \times 0,375 = 0,462$$

9.5 Escolha do Melhor Atributo na Raiz

Comparando os valores de Gini para cada atributo:

- **História de Crédito:** 0,566
- **Dívida:** 0,454
- **Garantia:** 0,5483
- **Renda:** 0,462

O atributo **"Dívida"** tem o menor índice de Gini (0,454) e, portanto, é escolhido como a raiz da árvore.

10 Construção da Árvore

A partir da escolha do atributo "Dívida" como raiz, dividimos os dados e continuamos o processo recursivamente:

- **Dívida = Alta**
  - **Renda = 0 a 15k**: Escolher "História de Crédito"
  - **Renda = 15 a 35k**: Escolher "História de Crédito"
  - **Renda = Acima de 35k**: Classe "Moderado"
- **Dívida = Baixa**
  - **História de Crédito = Ruim**: Classes "Alto" ou "Moderado"
  - **História de Crédito = Desconhecida**: Escolher "Renda"
  - **História de Crédito = Boa**: Classe "Baixo"

11 Árvore de Decisão Final

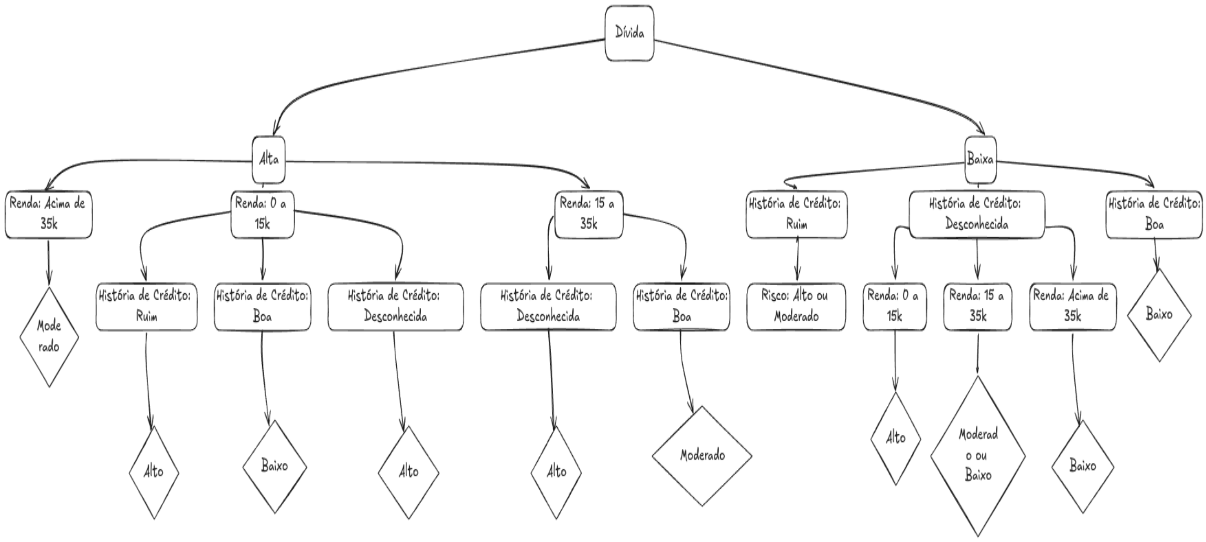

Figure 7: A figura acima mostra a árvore de decisão resultante, onde cada nó foi escolhido com base no menor índice de Gini possível.