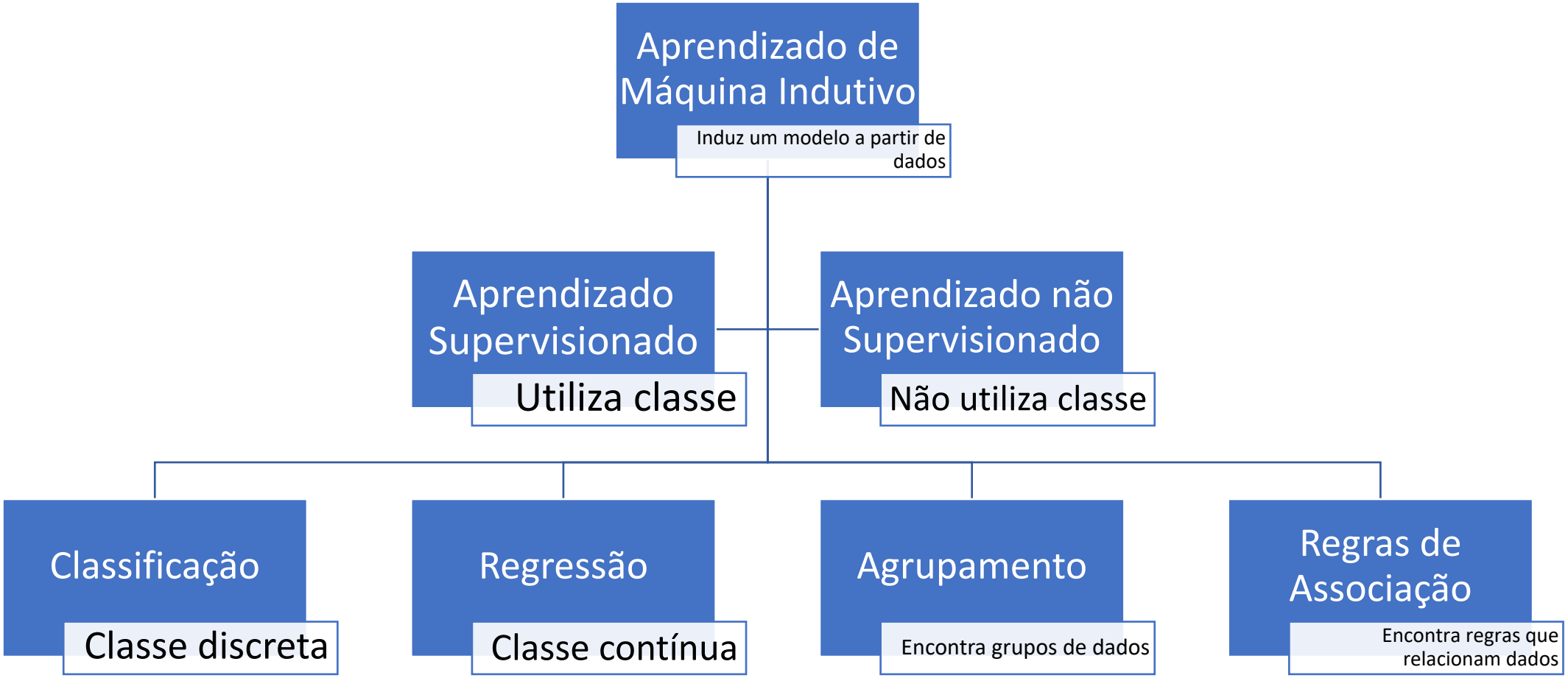
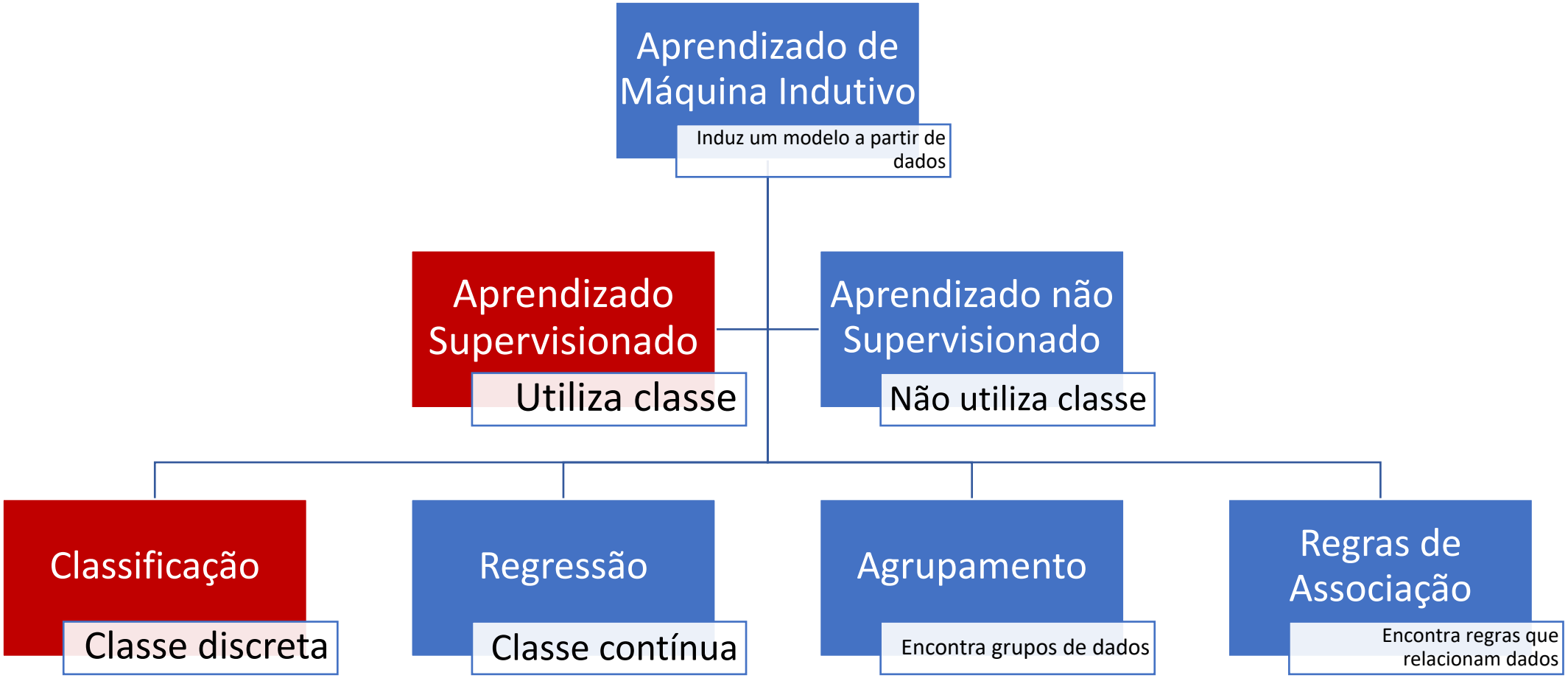
Aprendizado de Máquina
Aprendizado Supervisionado e Não Supervisionado
Aprendizado Supervisionado

Inteligência Artificial – 2020/1

Aprendizado de Máquina Supervisionado



Aprendizado de Máquina Supervisionado

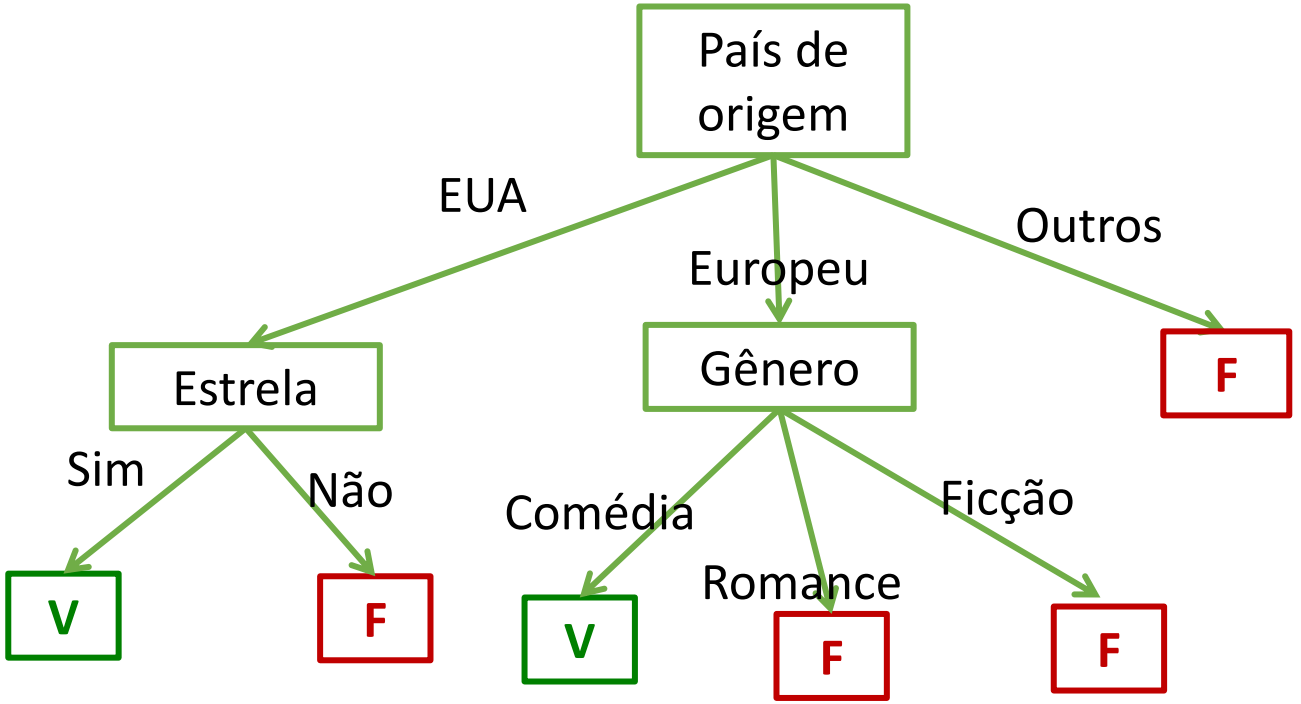


Conjunto de dados

Filme	País de Origem	Estrela	Gênero	Sucesso
Filme 1	Estados Unidos	Sim	Ficção científica	**V**
Filme 2	Estados Unidos	Não	Comédia	**F**
Filme 3	Estados Unidos	Sim	Comédia	**V**
Filme 4	Europeu	Não	Comédia	**V**
Filme 5	Europeu	Sim	Ficção científica	**F**
Filme 6	Europeu	Sim	Romance	**F**
Filme 7	Outros	Sim	Comédia	**F**
Filme 8	Outros	Não	Ficção científica	**F**
Filme 9	Europeu	Sim	Comédia	**V**
Filme 10	Estados Unidos	Sim	Comédia	**V**

Aprendizado de Máquina Supervisionado

- Árvore de Decisão
  - Estrutura de Árvore que resulta de um processo de aprendizado de máquina
  - Nós internos: atributos
  - Nós folha: classes



Aprendizado de Máquina Supervisionado

- Regras

O conhecimento representado em uma árvore de decisão pode ser representado, de forma equivalente, por regras:

Se País de Origem = EUA e

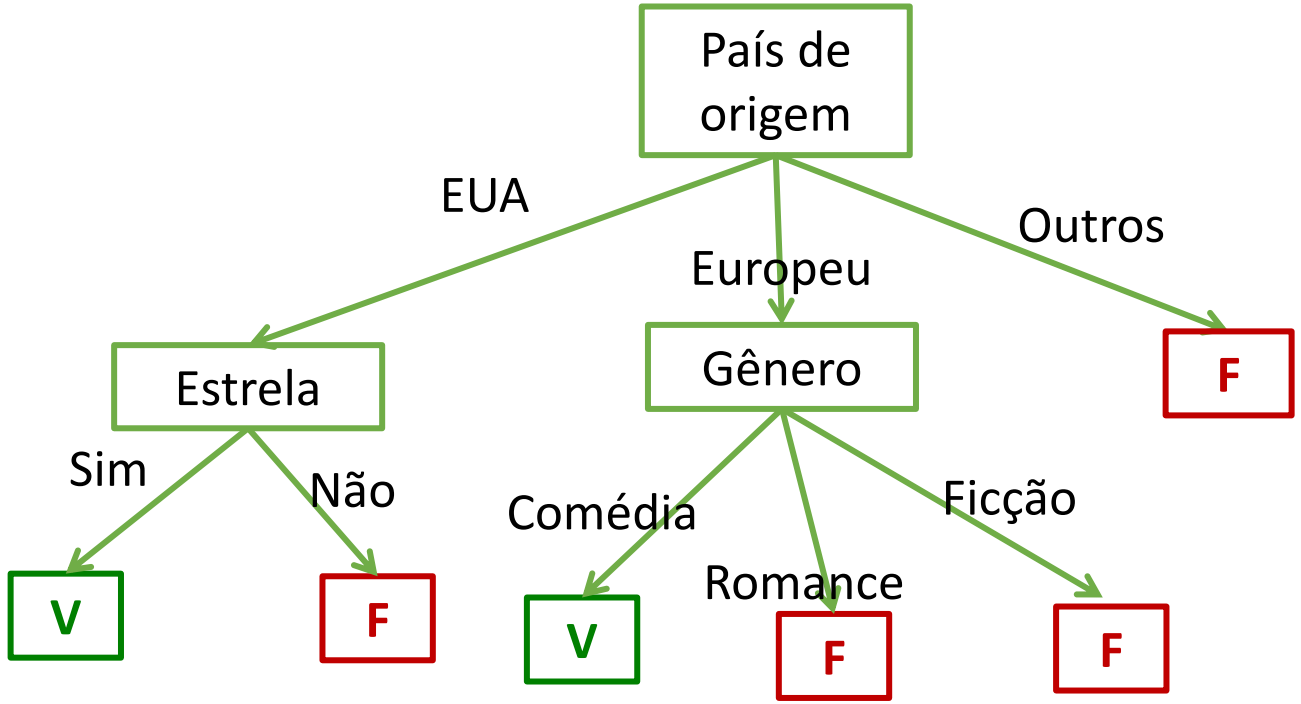
Estrela = Sim

Então Sucesso = **V**

Se País de Origem = Europeu e

Gênero = Comédia

Então Sucesso = **V**



Árvore de Decisão

Finalidade:

Classificar exemplos desconhecidos

Como construir?

Classificar o maior número de exemplos com a menor árvore possível.

Algoritmo Indutor de Árvore de Decisão:

Particiona recursivamente o conjunto de treinamento com base em um atributo selecionado até que os conjuntos obtidos com esse particionamento contenham dados de uma única classe.

Árvore de Decisão

- Algoritmo ID3
  - Proposto por Quinlan em 1980;
  - Trabalha apenas com atributos nominais;
  - Utiliza o ganho de informação como critério para selecionar os atributos;
  - Não utiliza pós-poda;
  - Não trata valores desconhecidos.

Árvore de Decisão

- Dados de treinamento e dados de teste
- No aprendizado de máquina supervisionado, o conjunto de dados é separado em conjunto de treinamento e conjunto de teste

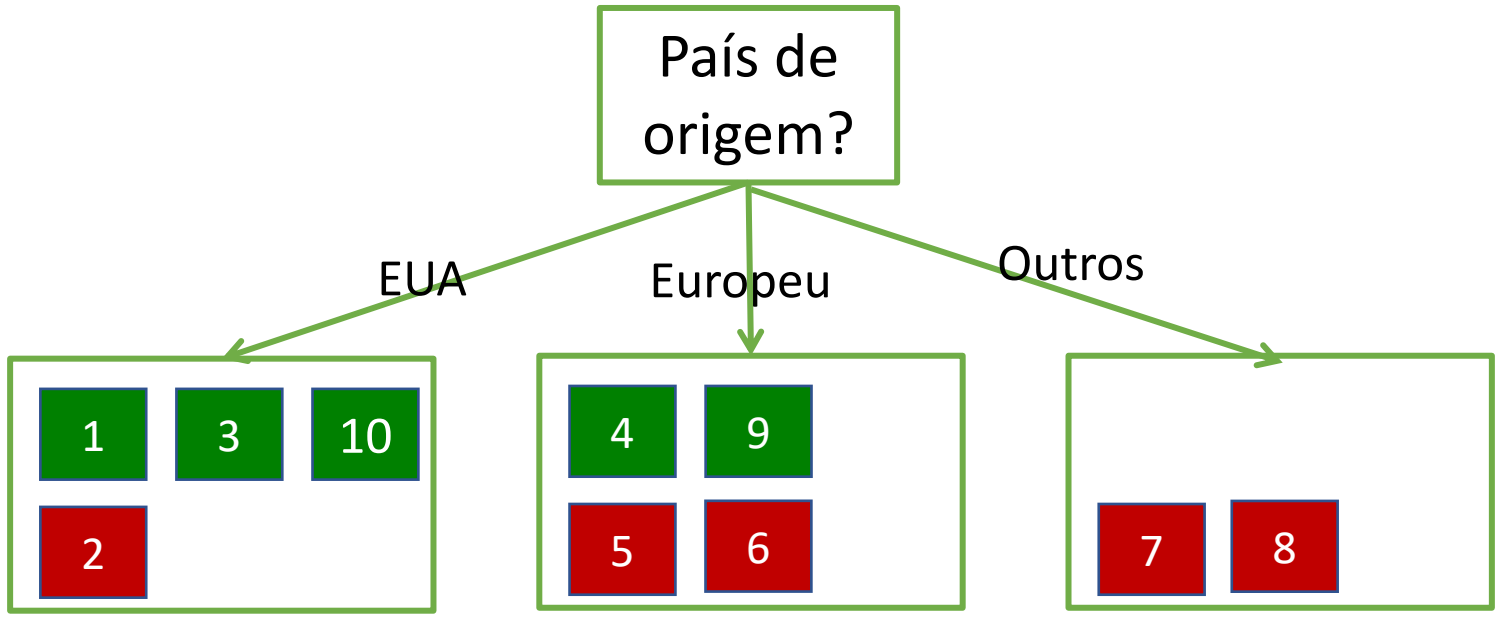
Filme	País de Origem	Estrela	Gênero	Sucesso
Filme 1	Estados Unidos	Sim	Ficção científica	V
Filme 2	Estados Unidos	Não	Comédia	F
Filme 3	Estados Unidos	Sim	Comédia	V
Filme 4	Europeu	Não	Comédia	V
Filme 5	Europeu	Sim	Ficção científica	F
Filme 6	Europeu	Sim	Romance	F
Filme 7	Outros	Sim	Comédia	F
Filme 8	Outros	Não	Ficção científica	F
Filme 9	Europeu	Sim	Comédia	V
Filme 10	Estados Unidos	Sim	Comédia	V
Filme 11	Estados Unidos	Não	Ficção científica	V
Filme 12	Estados Unidos	Sim	Romance	V
Filme 13	Outros	Não	Romance	F

Treinamento

Teste

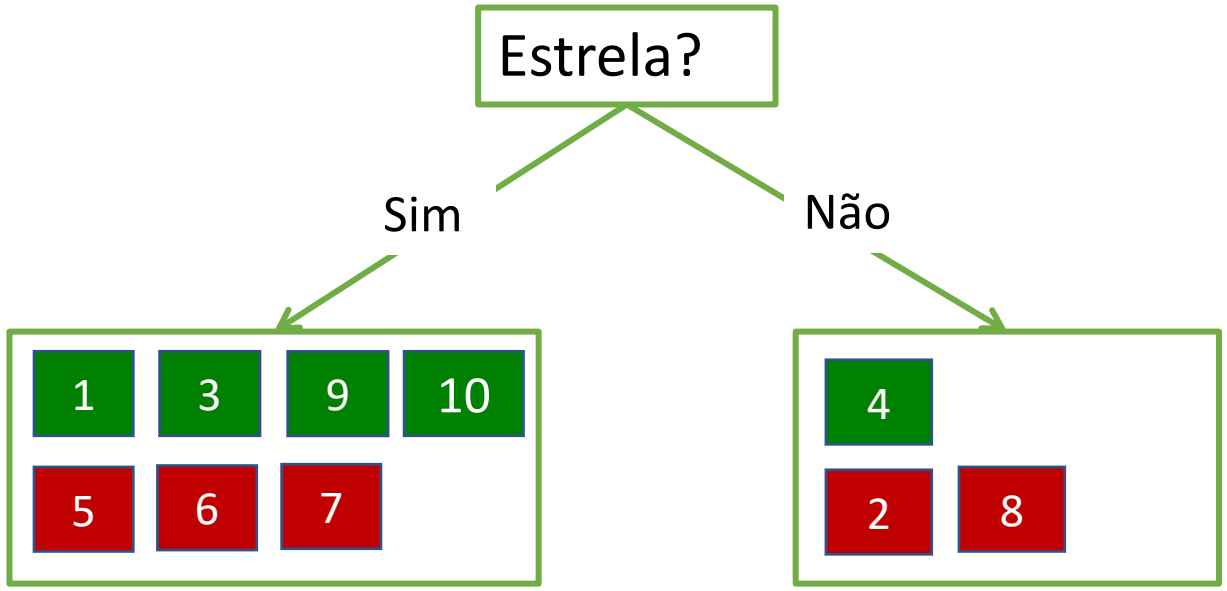
Árvore de Decisão

- Selecionar um atributo para dividir o conjunto de dados de treinamento
  - Caso fosse selecionado o atributo País de Origem:
  - Criação do nó País de origem
  - Criação de três arcos, um para cada valor: EUA, Europeu, Outros



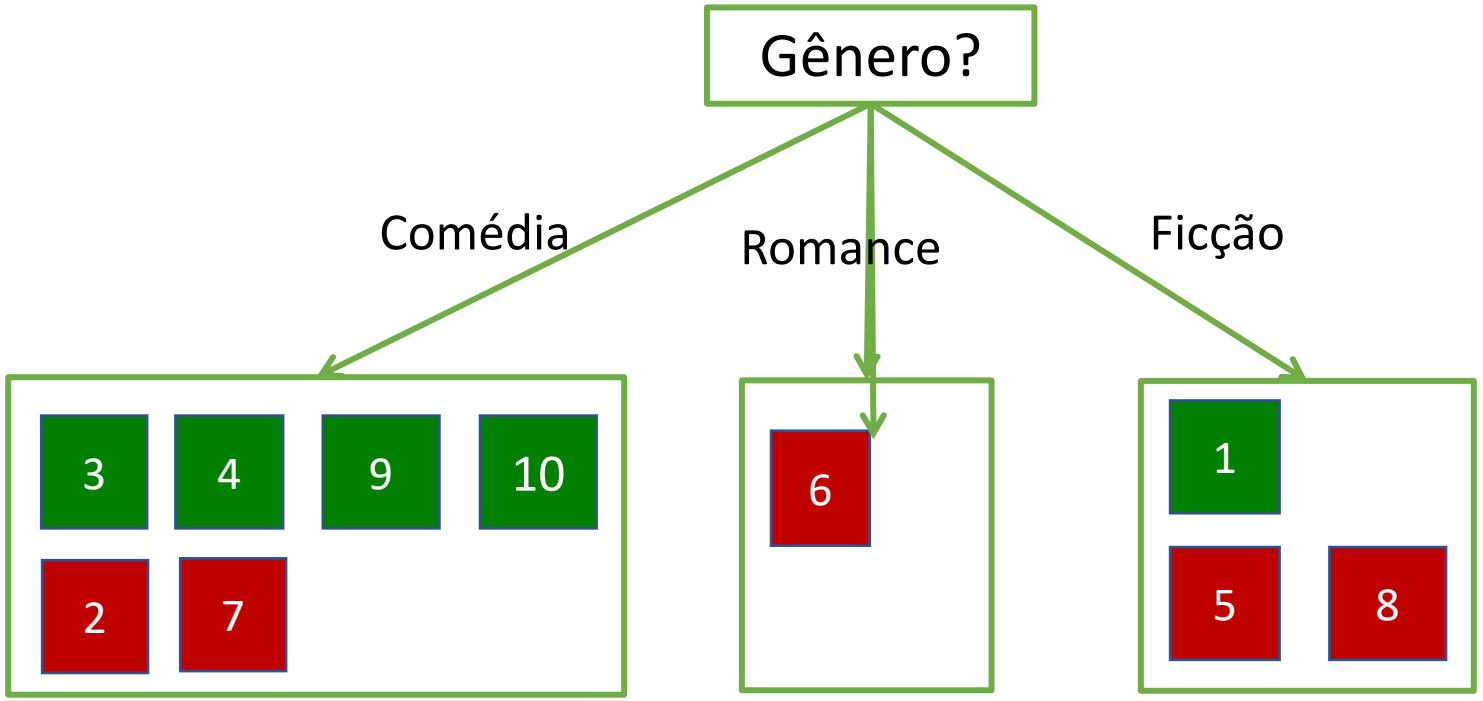
Árvore de Decisão

- Selecionar um atributo para dividir o conjunto de dados de treinamento
  - Caso fosse selecionado o atributo Estrela:
  - Criação do nó Estrela
  - Criação de dois arcos, um para cada valor: Sim, Não



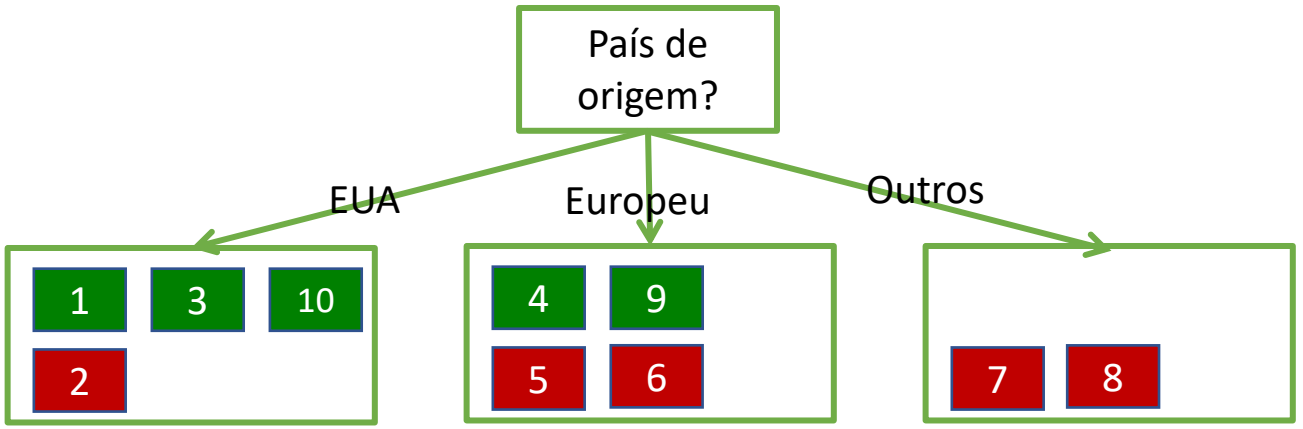
Árvore de Decisão

- Selecionar um atributo para dividir o conjunto de dados de treinamento
  - Caso fosse selecionado o atributo Gênero:
  - Criação do nó Gênero
  - Criação de três arcos, um para cada valor: Comédia, Romance, Ficção



Árvore de Decisão

- Qual atributo selecionar?
  - O atributo a ser selecionado deve ser aquele que melhor separa os dados de cada classe.
  - Para o atributo País de Origem:
  - O valor “Outros” só tem dados da classe **Não**; **(MUITO BOM!)**
  - O valor “EUA” tem a maioria dos dados da classe **Sim**; **(BOM!)**
  - O valor “Europeu” tem número igual de dados das duas classes. **(Não tão bom...)**



Árvore de Decisão

- Qual atributo selecionar?
- A estatística possui medidas que permitem avaliar as situações destacadas nos exemplos:

- Entropia: Medida de incerteza de uma variável randômica;

- Ganho de informação: redução em entropia

- O Algoritmo ID3 seleciona o atributo que oferece o maior ganho de informação.
  - Calcula a entropia do conjunto de dados
  - Calcula o ganho de informação de cada atributo e seleciona o que mais reduz a entropia

Cálculo de Entropia

Conjunto de dados S com duas classes: positivos e negativos

$$H(S) = -p_1 \log_2 p_1 - p_0 \log_2 p_0$$

$$p_1 = \frac{\textbf{positivos}}{\text{positivos + negativos}} \qquad p_0 = \frac{\textbf{negativos}}{\text{positivos + negativos}}$$

Valores de entropia para conjuntos com duas classes:

Próximos de 0 – entropia baixa, não há desordem nos dados

Próximos de 1 – entropia alta, há muita desordem na distribuição de classes

Cálculo de Entropia

- Conjunto de dados dos filmes

Filme	País de Origem	Estrela	Gênero	Sucesso
Filme 1	Estados Unidos	Sim	Ficção científica	**V**
Filme 2	Estados Unidos	Não	Comédia	**F**
Filme 3	Estados Unidos	Sim	Comédia	**V**
Filme 4	Europeu	Não	Comédia	**V**
Filme 5	Europeu	Sim	Ficção científica	**F**
Filme 6	Europeu	Sim	Romance	**F**
Filme 7	Outros	Sim	Comédia	**F**
Filme 8	Outros	Não	Ficção científica	**F**
Filme 9	Europeu	Sim	Comédia	**V**
Filme 10	Estados Unidos	Sim	Comédia	**V**

- Entropia do conjunto de dados original

$H(S) = -p_1 \log_2 p_1 – p_0 \log_2 p_0$

$p_1 = 5/10 = ½ = 0,5$

$p_0 = 5/10 = ½ = 0,5$

$H(S) = - 0,5 \log_2 0,5 – 0,5 \log_2 0,5 = \; - 0,5 (-1) – 0,5 (-1) = 0,5 + 0,5 = 1$

$H(S) = 1$

O conjunto tem entropia máxima: igual número de exemplos de cada classe.

Cálculo de Ganho de Informação

O ganho de informação de um atributo é a redução de entropia esperada com a escolha desse atributo.

No exemplo, o ganho de informação de cada atributo é dado por:

Ganho(**País de origem**) = H(S) – ERestante (**País de Origem**)
Ganho(**Estrela**) = H(S) – ERestante (**Estrela**)
Ganho(**Gênero**) = H(S) – ERestante (**Gênero**)

ERestante(A) – entropia esperada restante, depois de testar o atributo A

Selecionar o atributo que der o maior ganho de informação.

Cálculo de Ganho de Informação

Um atributo A com d valores distintos divide o conjunto de treinamento E em subconjuntos

$$E_1, E_2, ..., E_d.$$

Cada subconjunto E_k tem:

- p_k exemplos positivos e
- n_k exemplos negativos

Seguindo o ramo do atributo E_k, a entropia restante será

$$H(E_k) = -p_{1k} \log_2 p_{1k} - p_{0k} \log_2 p_{0k}$$

$$p_{1k} = \frac{pk}{pk + nk} \qquad p_{0k} = \frac{nk}{pk + nk}$$

Cálculo de Ganho de Informação

Calculando a entropia restante esperada para todos os valores do atributo A, temos a entropia restante esperada com a seleção do atributo A:

$$ERestante(A) = \sum_{k=1}^{d} \frac{p_k + n_k}{p + n} H(E_k)$$

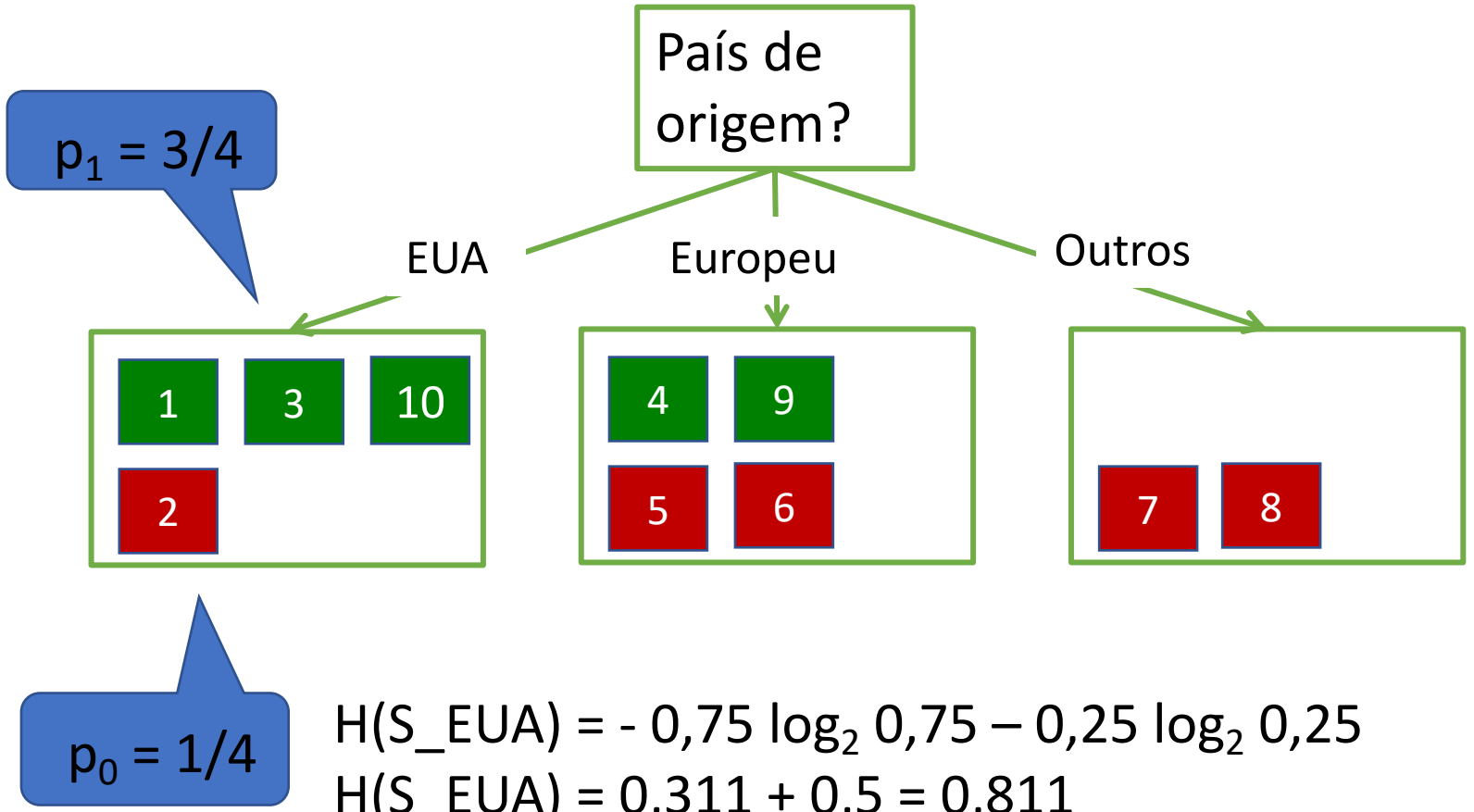
$$H(E_k) = -p_{1k} \log_2 p_{1k} - p_{0k} \log_2 p_{0k}$$

Ganho de informação obtido com a seleção do atributo A:

$$Ganho(A) = H(S) - ERestante(A)$$

Árvore de Decisão – Qual atributo selecionar?

Calcular a Entropia de País de origem = EUA

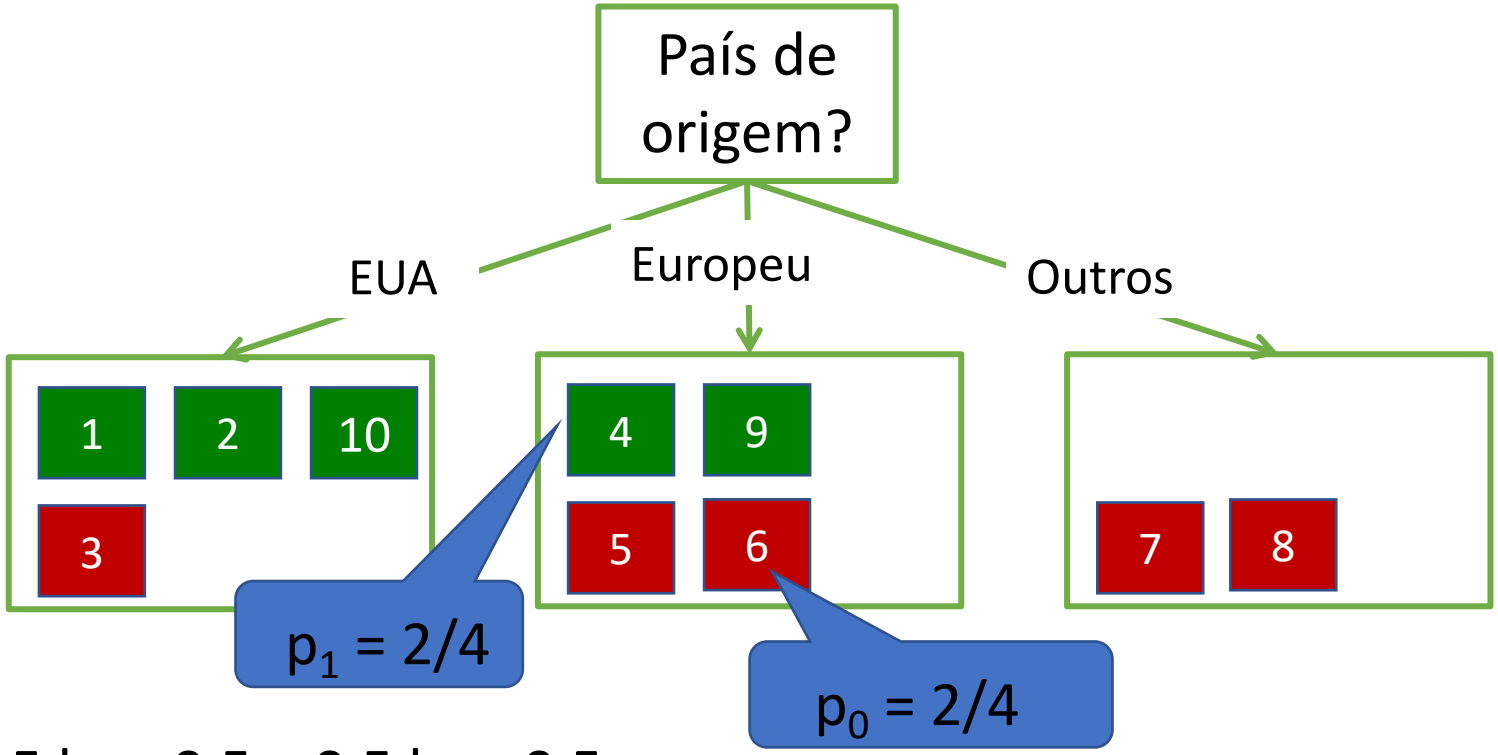


$H(S\_EUA) = - 0,75 \log_2 0,75 – 0,25 \log_2 0,25$

$H(S\_EUA) = 0,311 + 0,5 = 0,811$

Árvore de Decisão – Qual atributo selecionar?

Calcular a Entropia de País de origem = Europeu

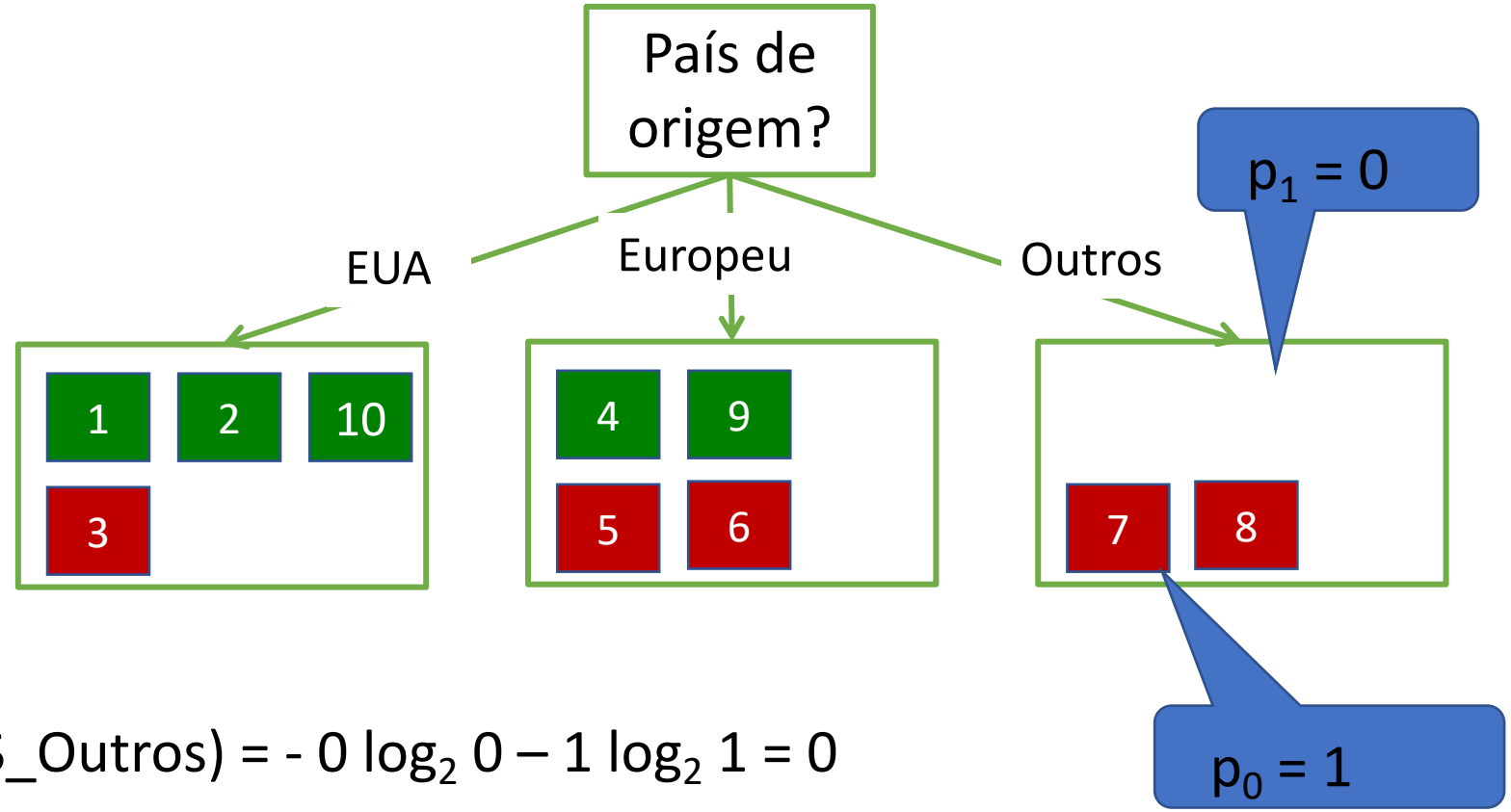


$H(S\_Europeu) = - 0,5 \log_2 0,5 – 0,5 \log_2 0,5$

$H(S\_Europeu) = 0,5 + 0,5 = 1$

Árvore de Decisão – Qual atributo selecionar?

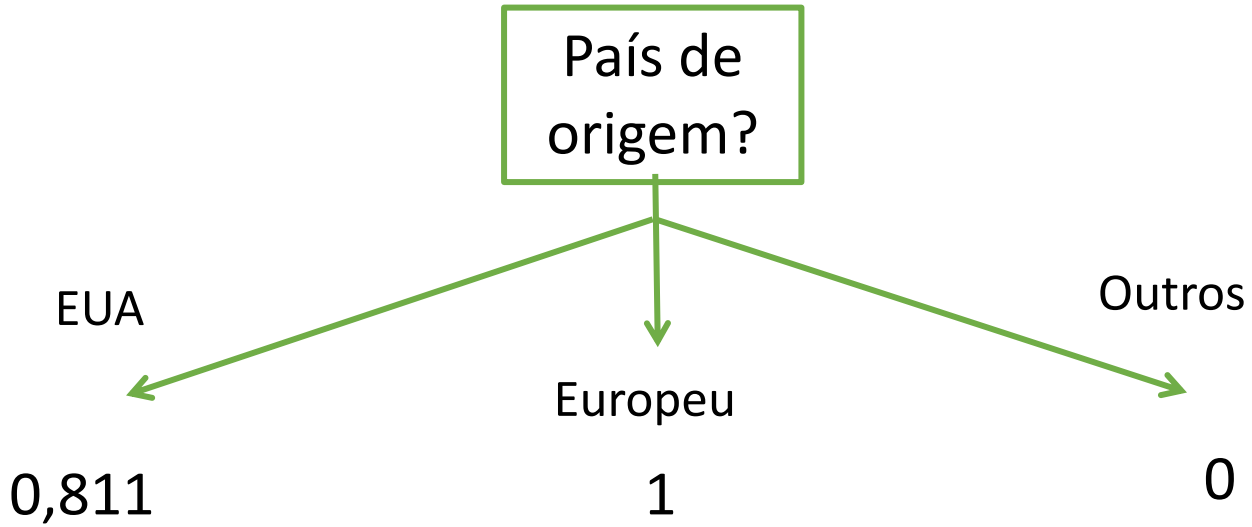
Calcular a Entropia de País de origem = Outros



$H(S\_Outros) = - 0 \log_2 0 - 1 \log_2 1 = 0$

Árvore de Decisão – Qual atributo selecionar?

Calcular a Entropia restante de País de Origem



ERestante(**País de Origem**) = (4/10) H(S_EUA) + (4/10) H(S_Europeu) + (2/10) H(S_Outros)

= 0,4 * 0,811+ 0,4 * 1+ 0,2 * 0 = 0,7244

Árvore de Decisão – Qual atributo selecionar?

Calcular o Ganho de Informação de País de Origem

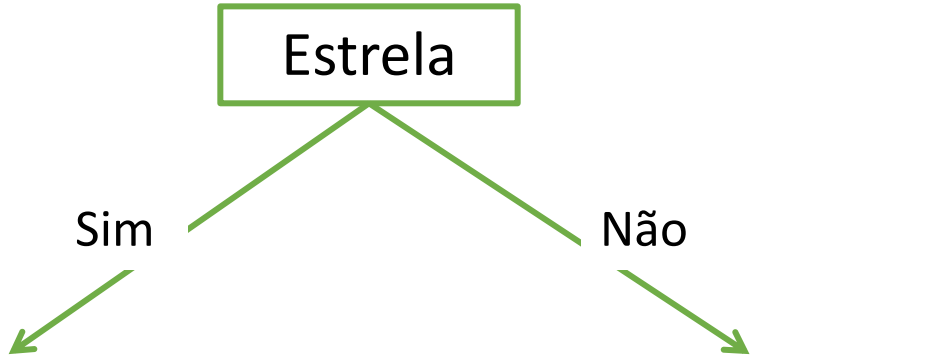
Ganho(**País de Origem**) = H(S) – ERestante (**País de Origem**)

= 1 - 0,7244

= 0,2756

Árvore de Decisão – Qual atributo selecionar?

Calcular Entropia Restante e Ganho de Informação de Estrela



ERestante(**Estrela**) = (7/10) H(S_Sim) + (3/10) H(S_Não)

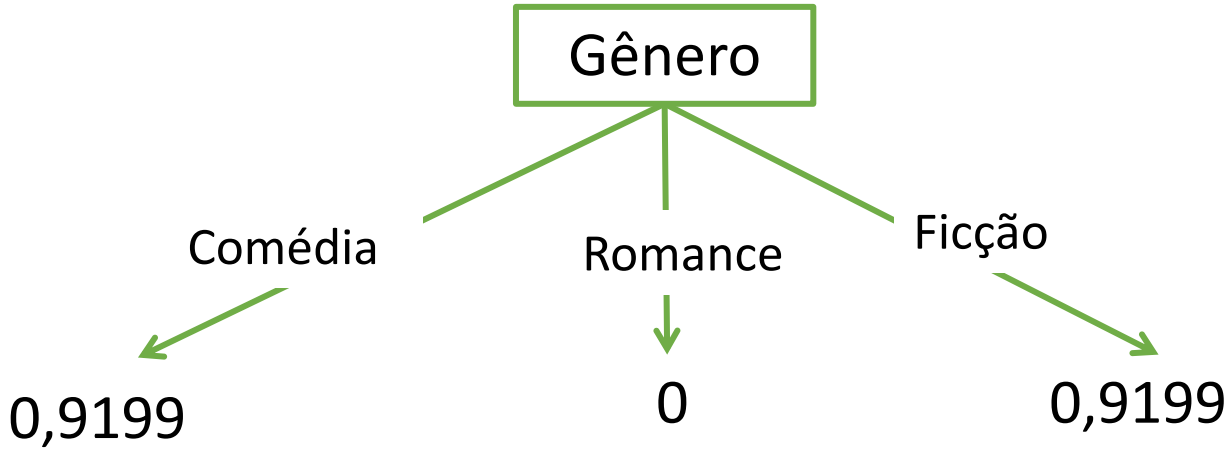
= 0,7 * 0,9852+ 0,3 * 0.9199

= 0,6896 + 0,2759 = 0,9656

Ganho(**Estrela**) = 1 - 0,9656 = 0,0344

Árvore de Decisão – Qual atributo selecionar?

Calcular Entropia Restante e Ganho de Informação de Gênero



ERestante (**Gênero**) = (6/10) H(S_Comédia) + (1/10) H(S_Romance) + (3/10) H(S_Ficção)

= 0,6 * 0,9199 + 0,1 * 0 + 0,3 * 0,9199

= 0,5519 + 0,2759 = 0,8279

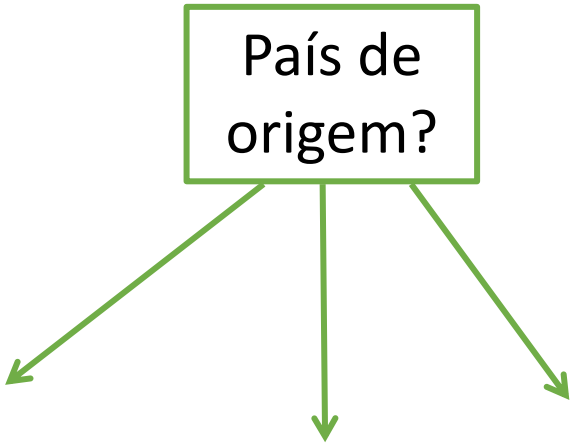
Ganho(**Gênero**) = 1 - 0,8279 = 0,1721

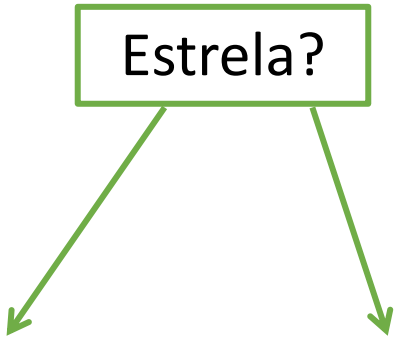
Árvore de Decisão – Qual atributo selecionar?

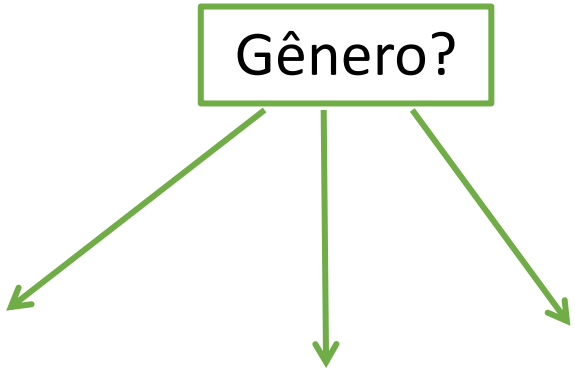
Ganho(**País de Origem**) = 1 - 0,7244 = 0,2756

Ganho(**Estrela**) = 1 - 0,9651 = 0,0344

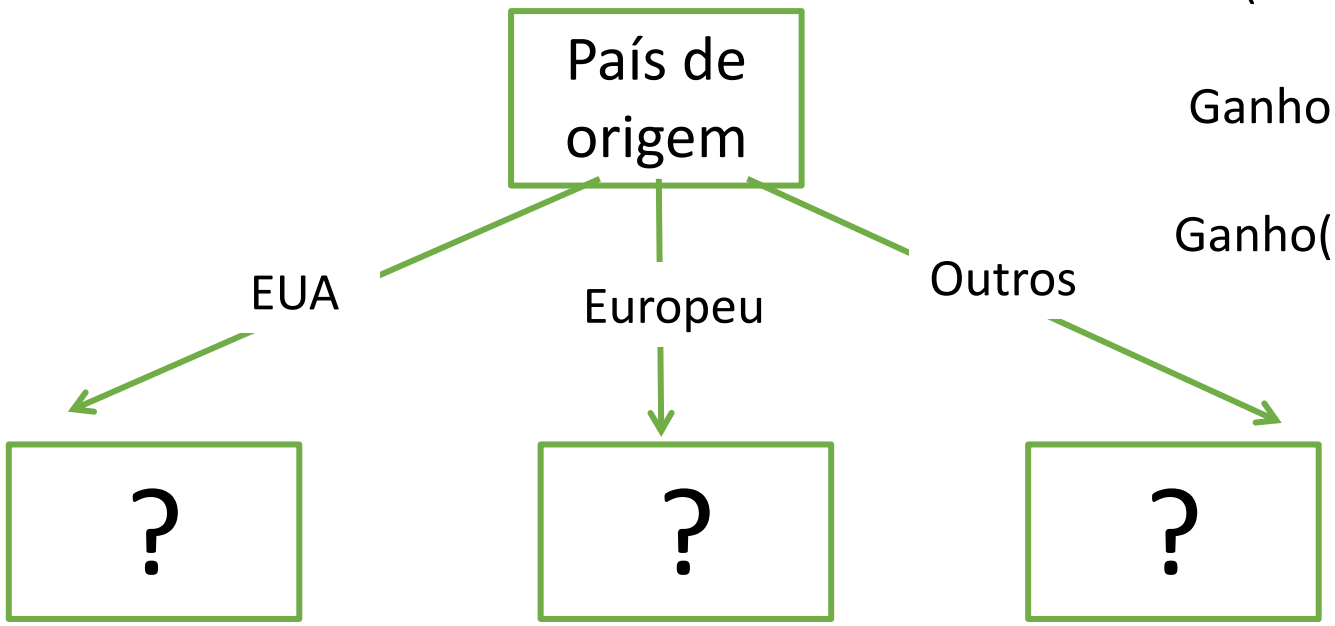
Ganho(**Gênero**) = 1 - 0,8264664 = 0,1721







Árvore de Decisão – Qual atributo selecionar?



Selecionado o maior

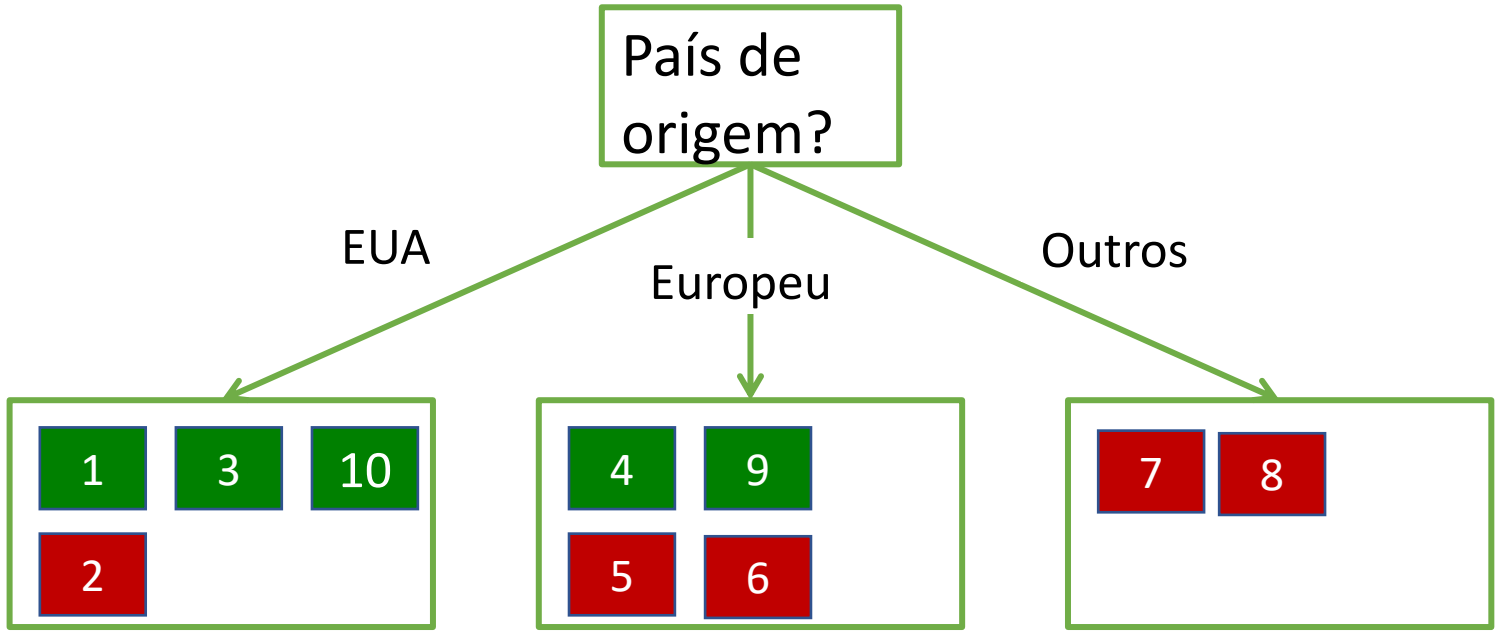
Ganho(**País de Origem**) = 1 - 0,7244 = 0,2756

Ganho(**Estrela**) = 1 - 0,9651 = 0,0349

Ganho(**Gênero**) = 1 - 0,8264664 = 0,17

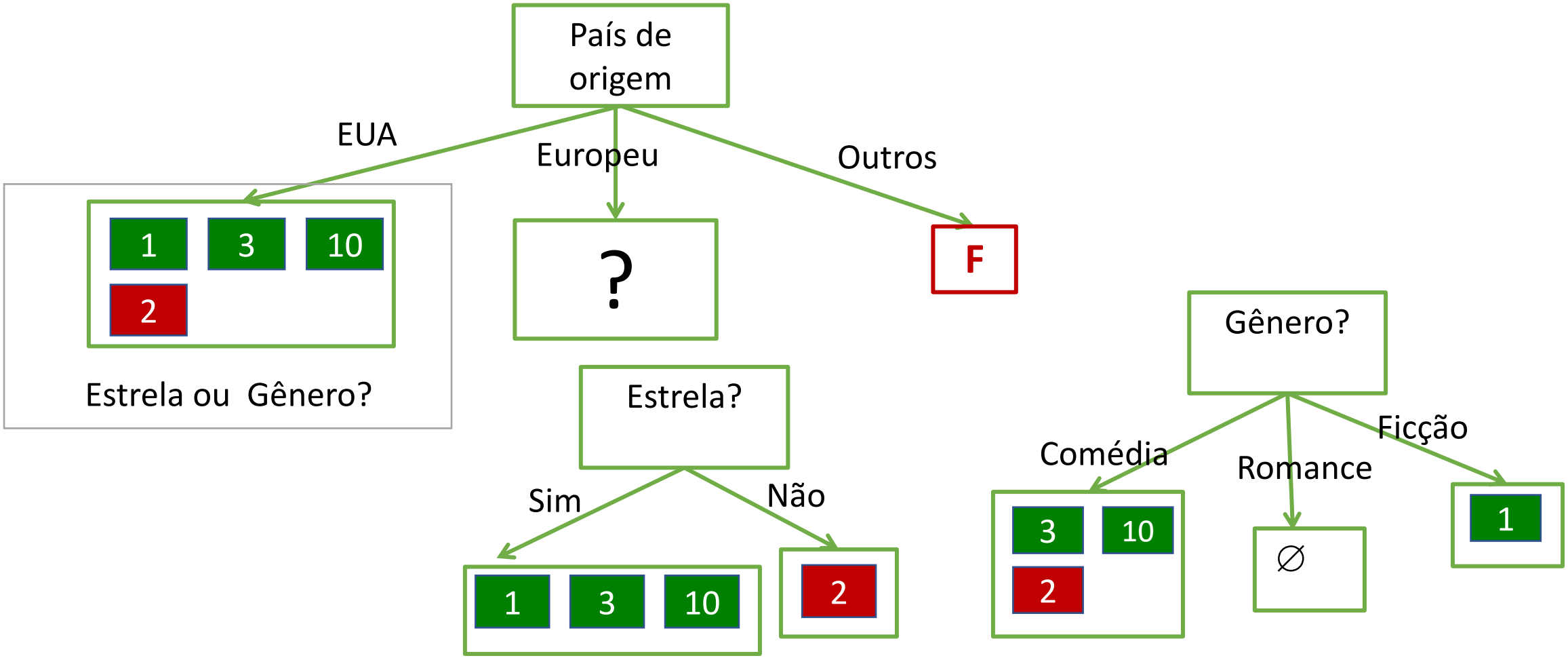
Árvore de Decisão – Próximos Passos

- O processo é repetido para todos os filhos criados pelo nó anterior.
  - Como o valor “Outros” leva a um subconjunto dos dados que tem só instâncias da classe **F**, esse nó se torna um nó folha, rotulado pela classe **F**.



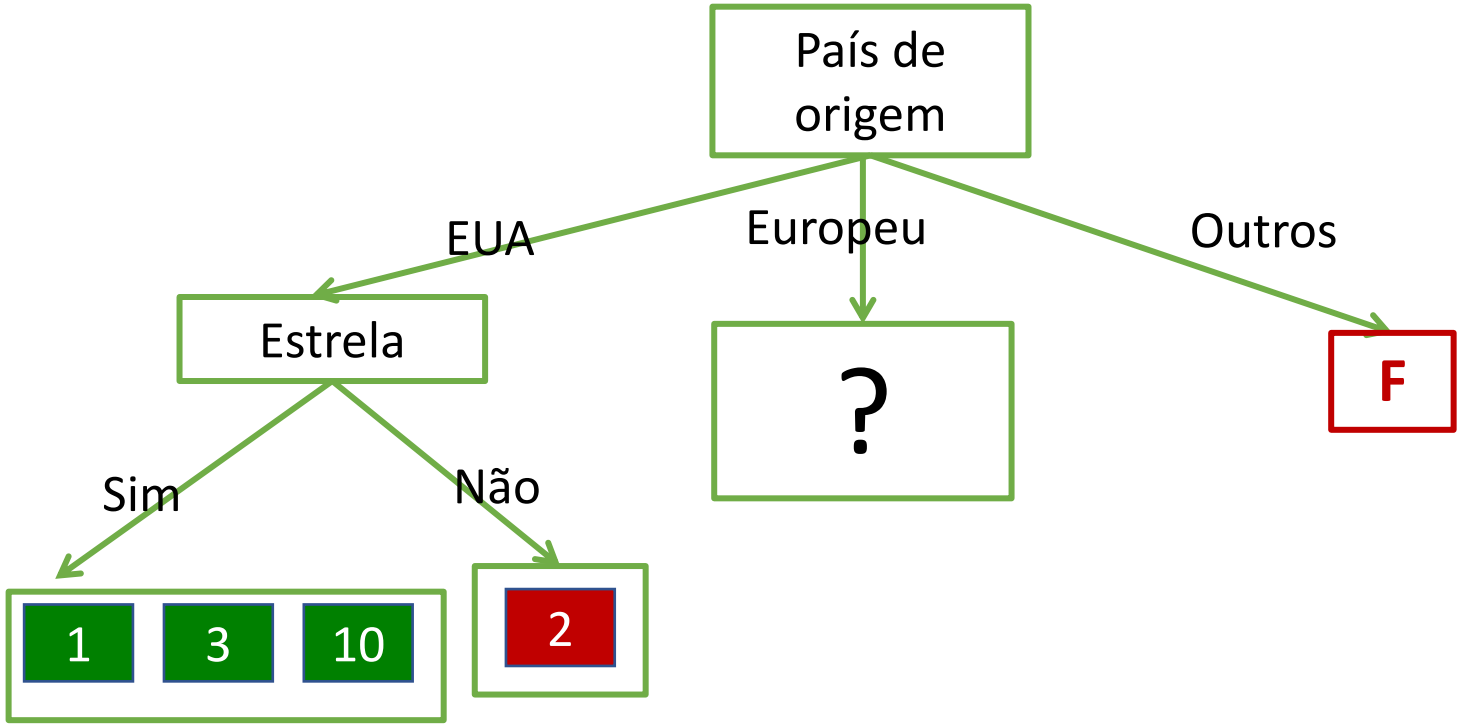
Árvore de Decisão – Próximos Passos

Calculando o ganho de informação de Estrela e Gênero para o nó do ramo País de Origem = EUA



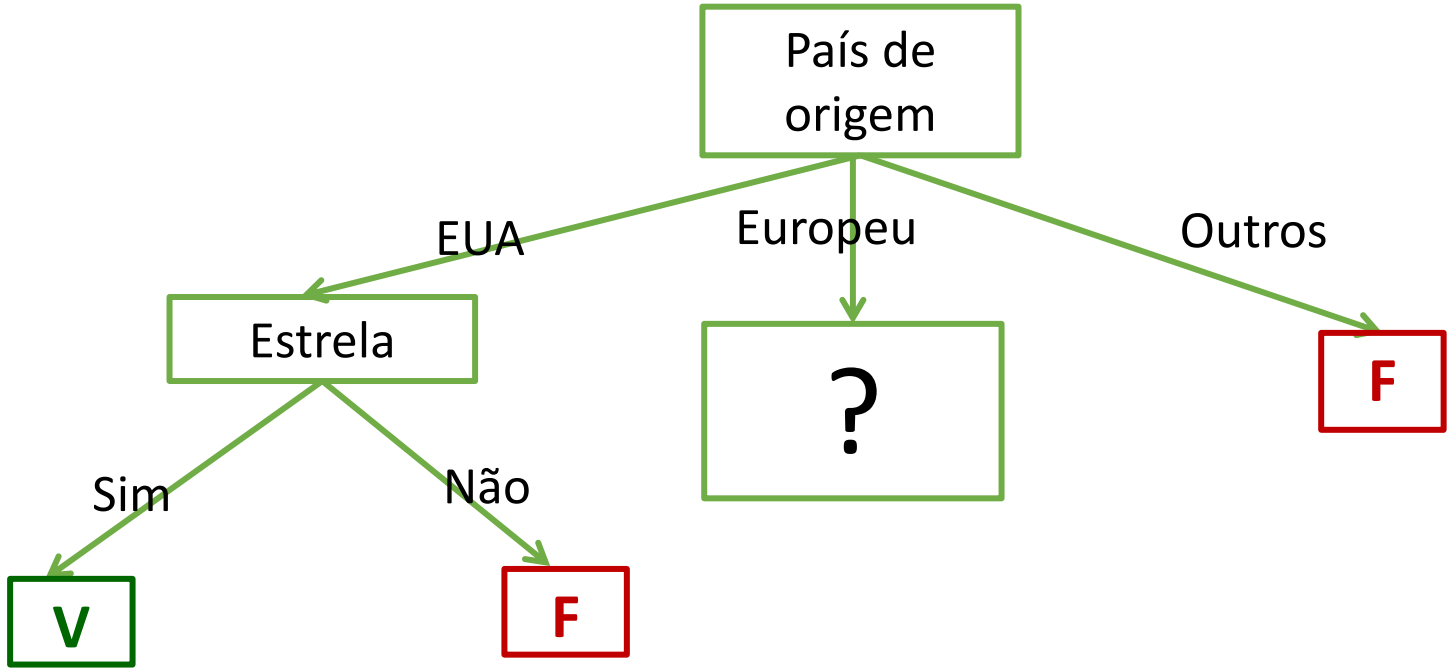
Árvore de Decisão – Próximos Passos

- O atributo Estrela tem o maior ganho de informação
- Cada um dos seus ramos leva a conjunto de dados da mesma classe
- Os nós do próximo nível já podem representar uma classe



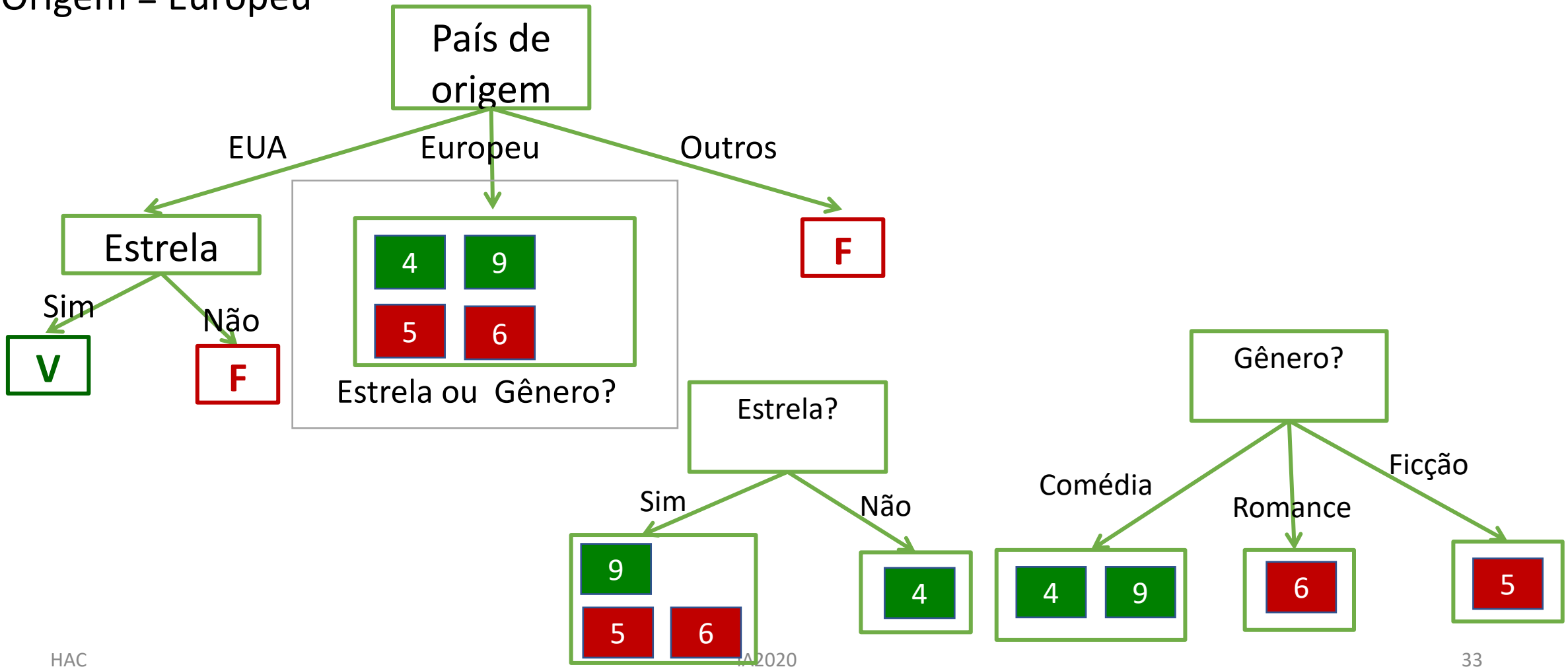
Árvore de Decisão – Próximos Passos

- O atributo Estrela tem o maior ganho de informação
- Cada um dos seus ramos leva a conjunto de dados da mesma classe
- Os nós do próximo nível já podem representar uma classe



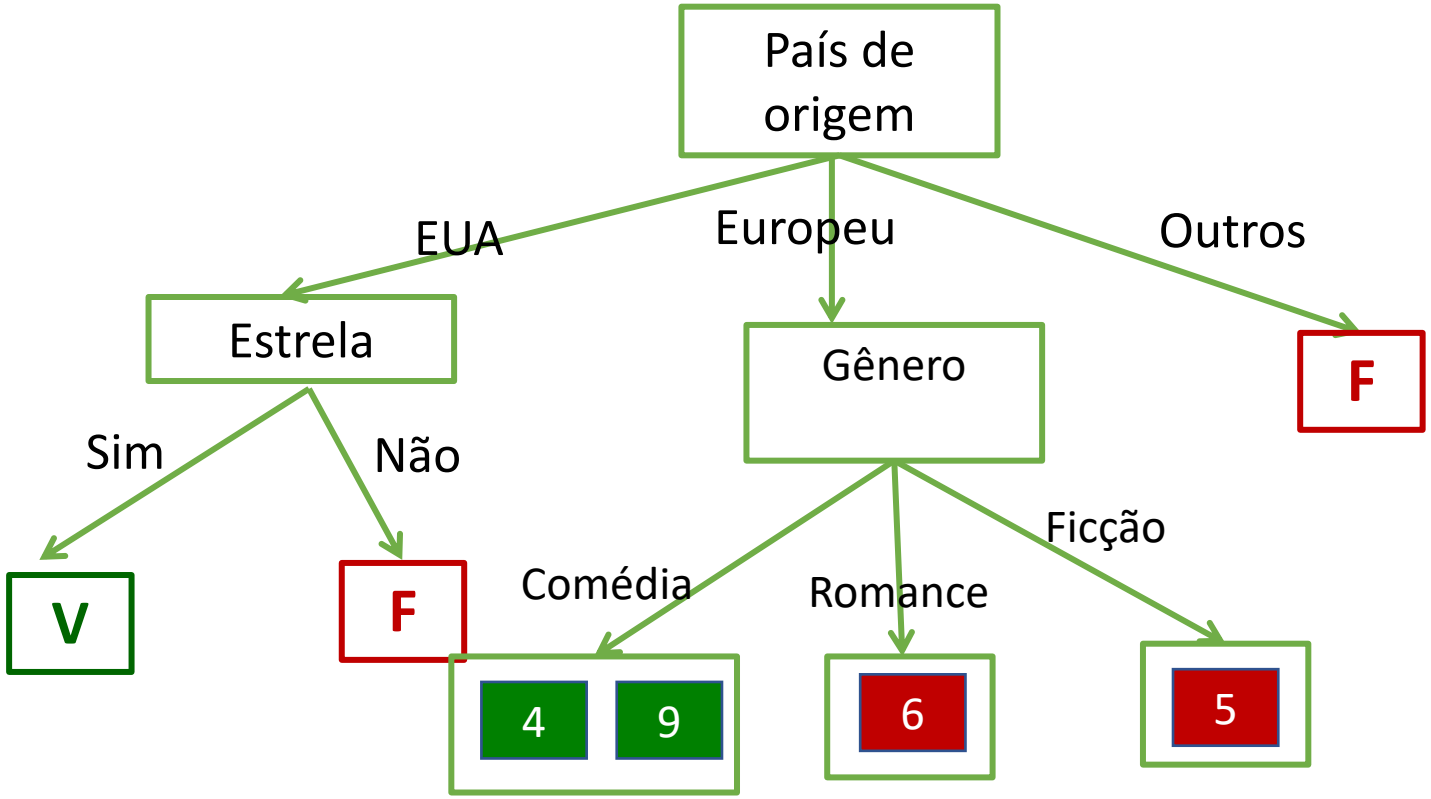
Árvore de Decisão – Próximos Passos

Calculando o ganho de informação de Estrela e Gênero para o nó do ramo País de Origem = Europeu



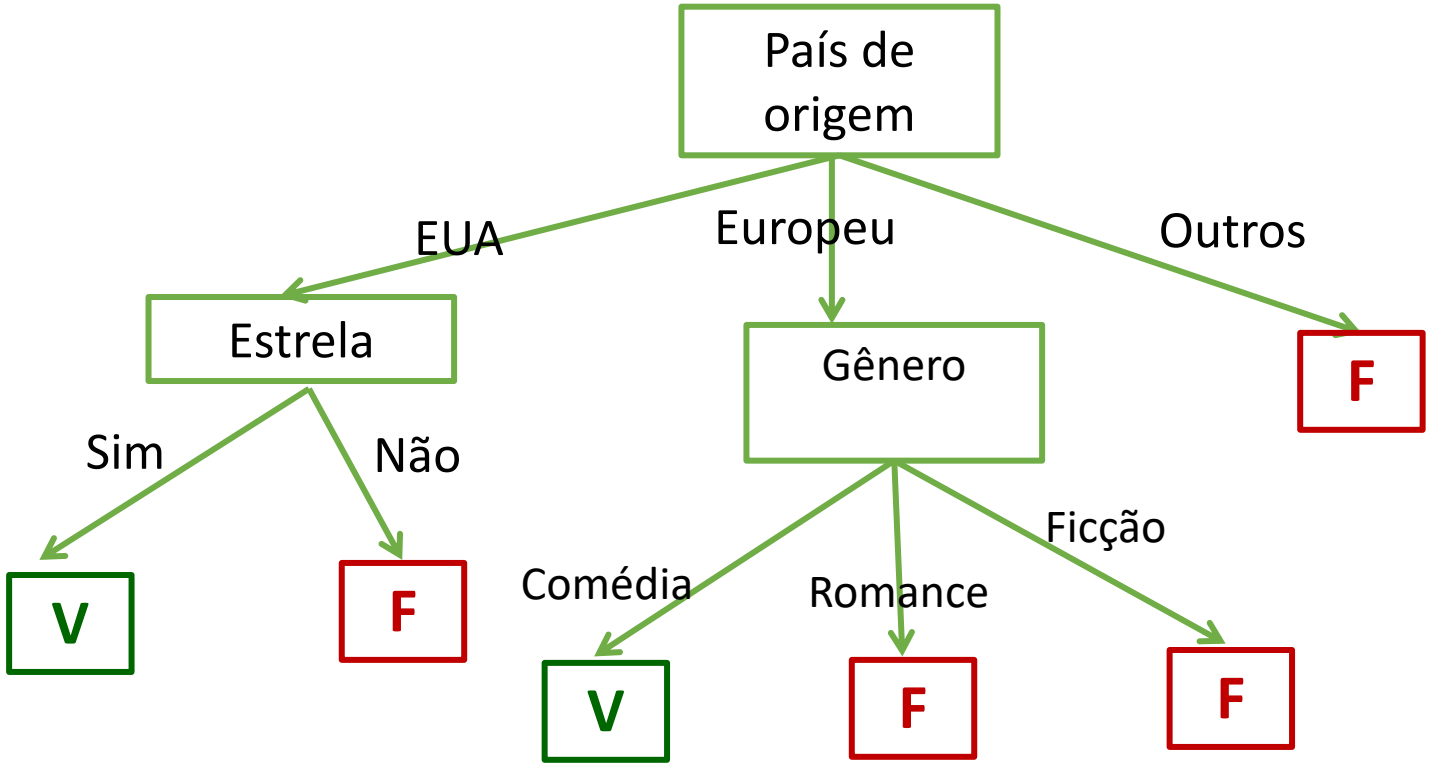
Árvore de Decisão – Próximos Passos

- O atributo Gênero tem o maior ganho de informação
- Cada um dos seus ramos leva a conjunto de dados da mesma classe
- Os nós do próximo nível já podem representar uma classe



Árvore de Decisão – Próximos Passos

- O atributo Gênero tem o maior ganho de informação
- Cada um dos seus ramos leva a conjunto de dados da mesma classe
- Os nós do próximo nível já podem representar uma classe



Árvore de Decisão – Teste

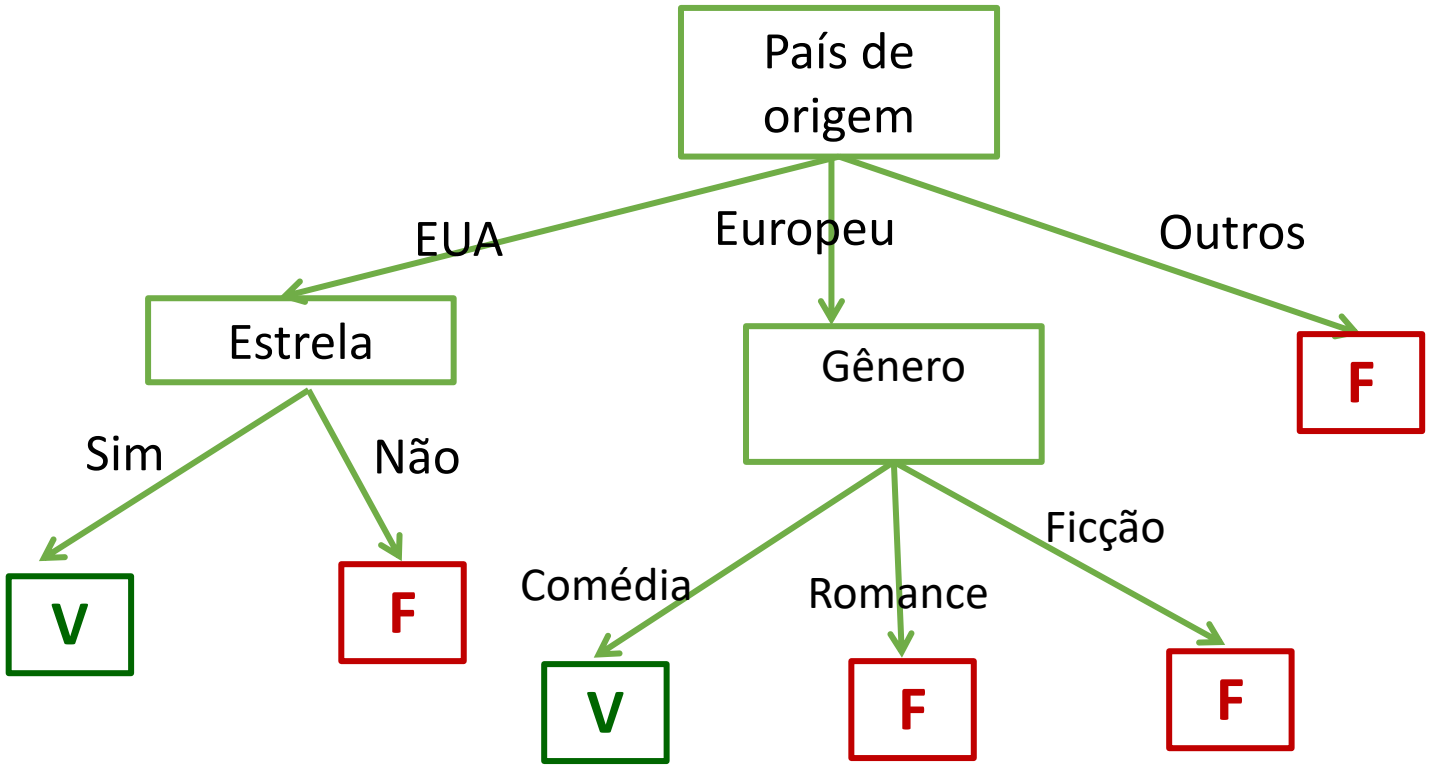
- Com a árvore de decisão pronta, exemplos de teste com classe conhecida são apresentados à arvore para testar o modelo aprendido

- Conjunto de dados de teste:

Filme	País de Origem	Estrela	Gênero	Sucesso
Filme 11	EUA	Não	Ficção	V
Filme 12	EUA	Sim	Romance	V
Filme 13	Outros	Não	Romance	F

- Predizer a possibilidade de sucesso do filme, em função de seus atributos

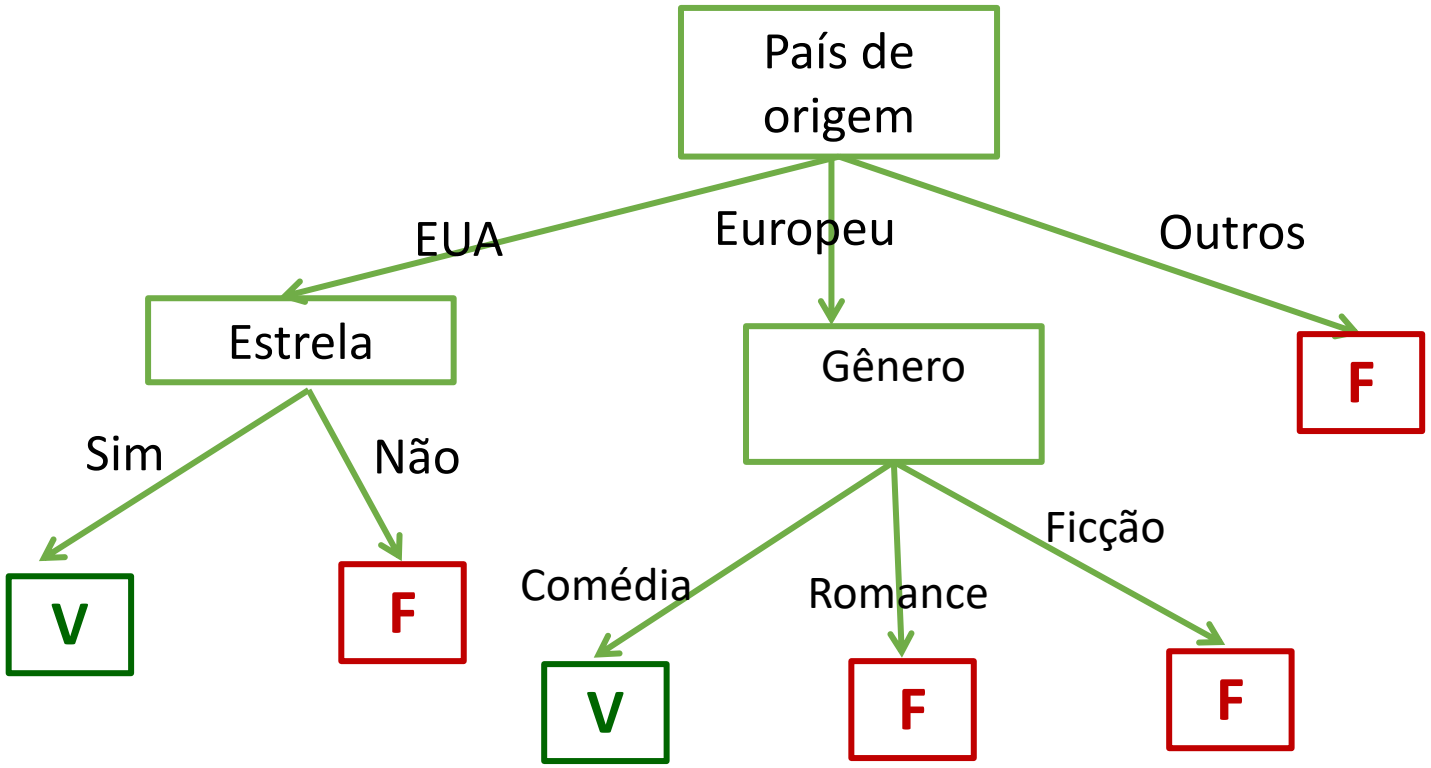
Árvore de Decisão – Teste



Filme	País de Origem	Estrela	Gênero	Classe Esperada	Classe Predita
Filme 11	EUA	Não	Ficção	V	**F**
Filme 12	EUA	Sim	Romance	V	**V**
Filme 13	Outros	Não	Romance	F	**F**

Árvore de Decisão – Teste

Acurácia= 2/3
Erro = 1/3



Filme	País de Origem	Estrela	Gênero	Classe Esperada	Classe Predita
Filme 11	EUA	Não	Ficção	V	F
Filme 12	EUA	Sim	Romance	V	V
Filme 13	Outros	Não	Romance	F	F

Cálculo de Entropia

- O cálculo de entropia pode ser generalizado quando o conjunto de dados possui mais que duas classes.

- Conjunto de dados S com c classes:

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

$$p_i = n_i / n$$

  - n_i – número de exemplos da classe i
  - n – número total de exemplos

Limitações do algoritmo ID3

- Trata apenas atributos categóricos
- Utiliza ganho de informação, que não leva em conta o número de arestas do nó
- Não trata valores desconhecidos
- Não utiliza poda, para evitar super ajuste

Algoritmo C4.5

- Apresenta melhorias com relação ao ID3:
- Trata valores categóricos e numéricos;
- Para atributos contínuos utiliza o teste simples com divisão binária;
- Trata valores desconhecidos;
- Utiliza razão de ganho para selecionar o atributo que melhor divide os exemplos;
- Apresenta um método de pós-poda.

Algoritmo C4.5 – Razão de ganho

- Problema do ganho de informação:
  - Dá preferência a atributos com número grande de valores;
  - Pode selecionar atributos irrelevantes, que tem um só exemplo para cada valor de atributo.
- Proposta de solução: razão de ganho
  - Modifica o ganho de informação para reduzir a tendência de favorecer atributos com muitas ramificações

Algoritmo C4.5 – Métodos de poda

- Arestas ou subárvores da árvore de decisão podem representar ruídos, erros ou exemplos específicos;
- Esse problema é chamado superajuste (overfitting);
- Significa que a árvore induzida está excessivamente ajustada ao conjunto de dados de treinamento e não aprendeu um conhecimento genérico
- A capacidade da árvore de classificar exemplos desconhecidos fica reduzida;
- É um fenômeno geral, podendo ocorrer em qualquer método de aprendizado.

Indutor de árvore de decisão (para atributos discretos)

função ARVORE (*exemplos, atributos, default*) retorna arvore

1. **se** não há exemplos **então retorne** valor default
2. **se** todos os exemplos tem a mesma classe **então retorne** a classe
3. *best* = escolha_atributo(*atributos, exemplos*);
4. *arvore* = nova arvore de decisão com atributo *best* na raiz
5. **para todo** valor v_i de *best* **faça**
6. *exemplos* $_i$ = {elementos de exemplos com *best* = v_i}
7. *subarvore* = ARVORE (exemplos $_i$, *atributos – best*, valor_maioria(*exemplos*))
8. adicione um ramo para *arvore* com rótulo v_i e subárvore *subarvore*
9. ***fim-para***
10. ***retorne*** *arvore*

Indutor de árvore de decisão (para atributos discretos)

- Valor default: valor atribuído ao nó folha quando nenhum exemplo chega até o nó. Pode ser o valor da classe majoritária (classe da maioria dos exemplos de treinamento).

- A função escolha_atributo(exemplo, atributos) utiliza alguma medida para escolher o atributo de divisão de um nó. No ID3 é o ganho de informação, no C4.5 é o razão de ganho.

- Número mínimo de exemplos nas folhas: é possível definir um número mínimo de exemplos para interromper o processo em um ramo. Se o número de exemplos estiver igual ou melhor que o mínimo, criar um nó folha com o valor default.