Árvore de Decisão - *Decision Tree*

Aplicação:

- Regressão
- Classificação com variáveis qualitativas

Vantagens:

- Intuitiva e fácil de entender.
- Grande poder de explicação
- Permite criar regras
- Classificação com variáveis qualitativas

Desvantagens:

- Menor acurácia no caso se regressões
- Pode ficar muito sensível a pequenas variações.
- Solução: *Bagging* e *Random Forest*

Decision Tree – Exemplo Análise de Crédito

Objetivo: Construir uma árvore de decisão para classificação de crédito

Variáveis de Classificação	Critério	Categoria
Credito Qualidade do Crédito	baixa probabilide de atraso ou default	bom
	alta probabilide de atraso ou default	ruim

Variáveis de Decisão	Critério	Categoria
Comprometimento Razão entre o valor da parcela e a renda líquida	ValorParcela/RendaLíquida ≤ 5%	Baixo
	5% ≤ ValorParcela/RendaLíquida < 10%	Médio
	10% ≤ ValorParcela/RendaLíquida	Alto
EstadoCivil Estado civil atual	solteiro	solteiro
	separado ou divorciado	divorciado
	casado ou união estável	casado
Historico Histórico de atrasos no pagamento de parcelas	1 ou mais atrasos superiores a 30 dias	atraso
	sem histórico de atrasos	pontual
CasaPropria Possui casa própria	possui casa própria em seu nome	sim
	não possui casa própria em seu nome	não

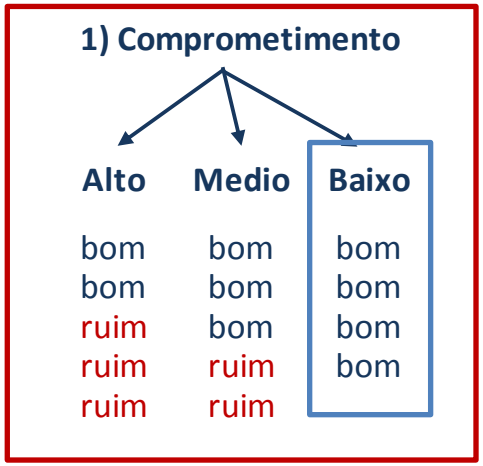
Decision Tree – Exemplo Análise de Crédito

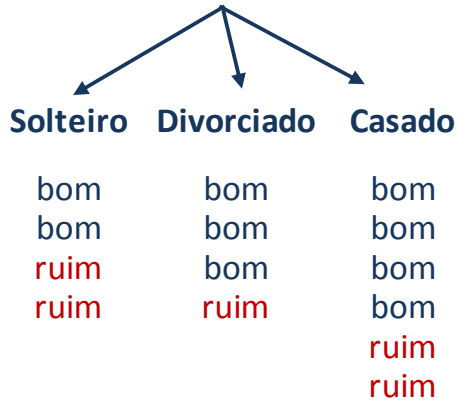
Base de Treinamento

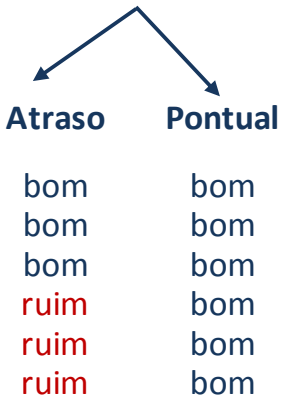
#	Comprometimento	EstadoCivil	Historico	CasaPropria	Credito
1	Alto	solteiro	atraso	sim	ruim
2	Alto	solteiro	atraso	não	ruim
3	Baixo	solteiro	atraso	sim	bom
4	Medio	casado	atraso	sim	bom
5	Medio	divorciado	pontual	sim	bom
6	Medio	divorciado	pontual	não	ruim
7	Baixo	divorciado	pontual	não	bom
8	Alto	casado	atraso	sim	ruim
9	Alto	divorciado	pontual	sim	bom
10	Medio	casado	pontual	sim	bom
11	Alto	casado	pontual	não	bom
12	Baixo	casado	atraso	não	bom
13	Baixo	solteiro	pontual	sim	bom
14	Medio	casado	atraso	não	ruim

Arquivos: R/databases/DecisionTress-CreditScore-data..xlsx

Decision Tree – Exemplo Análise de Crédito







bom	2	3	4
ruim	3	2	0
total	5	5	4

2	3	4
2	1	2
4	4	6

3	6
4	1
7	7

6	3
2	3
8	6

Qual árvore permite faze a classificação do crédito com ***menor incerteza***?

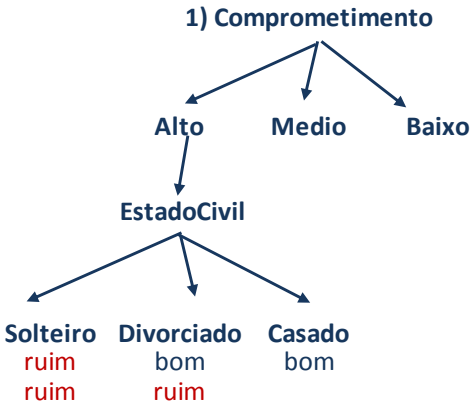
A árvore 1: se o *Comprometimento* da renda é *Baixo* é zero a ***a incerteza*** para classificar o crédito como *bom* é igual a zero.

A incerteza será tanto menor quanto maior a ***pureza*** do nó ou do grupo.

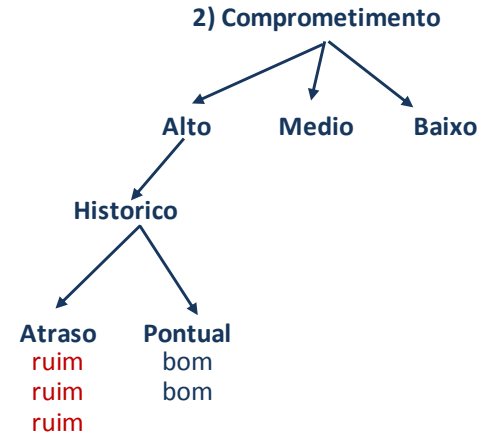
No nó *CasaPropria-Não* onde a incerteza é máxima já que 50% do crédito é *bom* ou *ruim*.

Decision Tree – Exemplo Análise de Crédito

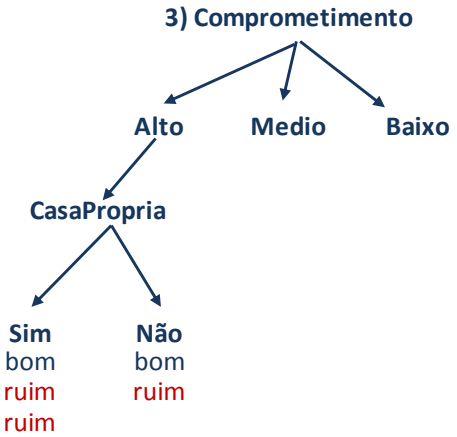
Para a classe *Comprometimento-Alto* podemos construir as seguintes árvores:



bom	0	1	1
ruim	2	1	0
total	2	2	1



bom	0	2
ruim	3	0
total	3	2



bom	1	1
ruim	2	1
total	3	2

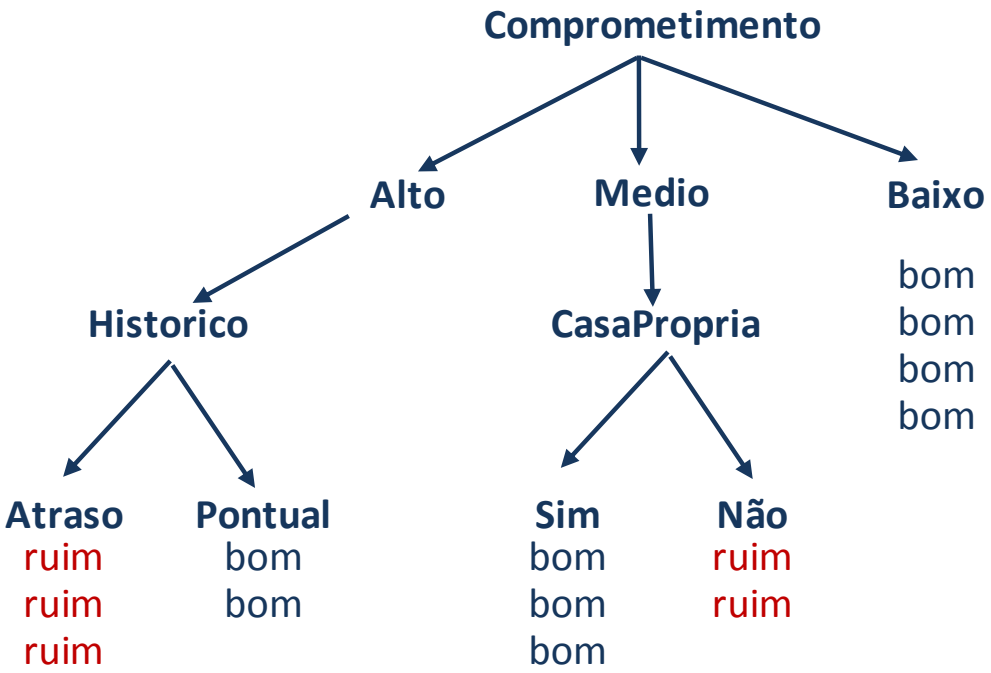
Qual árvore permite faze a classificação do crédito com ***menor incerteza***?

A árvore 2: conhecido o *Histórico* a incerteza em classificar o crédito como *bom* ou *ruim* é zero.

As classe *Atrasado* e *Pontual* são ***puras*** de cada classe.

Decision Tree – Exemplo Análise de Crédito

Seguindo o mesmo procedimento obtemos a árvore final:



Critérios de parada:

1. Todos os nós das extremidades da árvore (folhas) forem puros
2. Não houver mais divisões possíveis, ou seja, fim das variáveis
3. Quando a redução das incerteza com novos nós for igual a zero

Decision Tree – Medida de Incerteza

Características de uma medida de incerteza

1. A **incerteza é zero** quando para observada uma variável houver uma única classificação possível.

Ex: Dado *Comprometimento: Alto*

Historico: Atraso o crédito é ruim

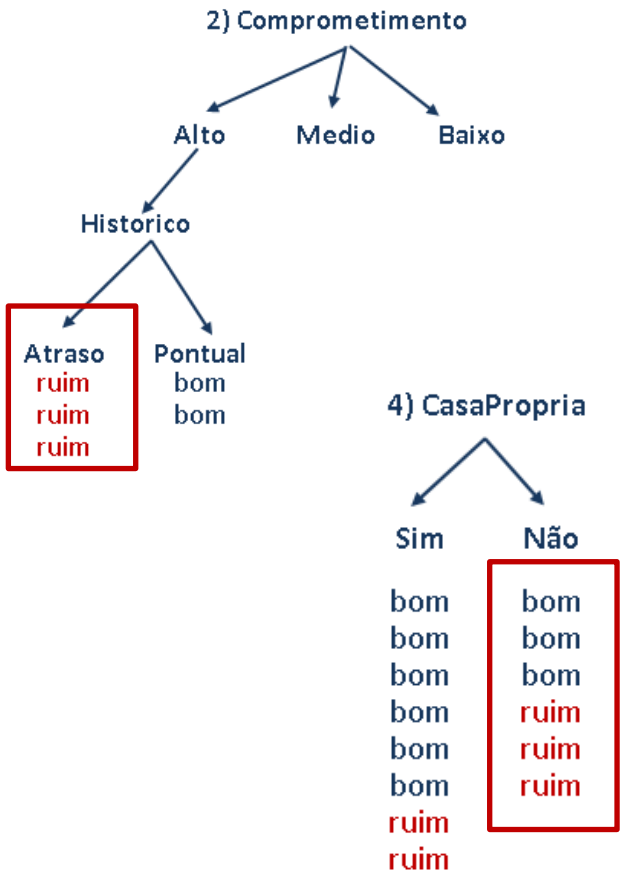
Historico: Pontual o crédito é bom

2. A **incerteza é máxima** quando para uma variável as classes estiverem igualmente distribuídas

Ex: *CasaPropria: Não* tem o mesmo número de créditos classificados como bom e ruim

3. A incerteza deve ser aditiva e permitir multiplas categorias:

Ex: incerteza (casado, não-casado) dever ser igual a incerteza (casado, solteiro, divorciado)

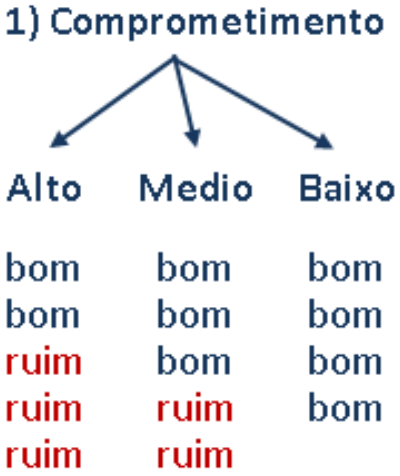


Decision Tree – Medida de Incerteza: Entropia

Entropia é a medida que utilizaremos para calcular a incerteza a cada nó da árvore

Em cada nó: Proporção da Classe $$p_i = \frac{\text{Número de Elementos da Classe}}{\text{Número de Elementos do Nó}}$$

Entropia do nó: $ent(p_1, p_2, ..., p_n) = -p_1.\log p_1 - p_2.\log p_2 - ...p_n.\log p_n$



Comprometimento Alto:

$$ent([2,3]) = -(\frac{2}{5}).\log(\frac{2}{5}) - (\frac{3}{5}).\log(\frac{3}{5}) = 0.97$$

Comprometimento Medio:

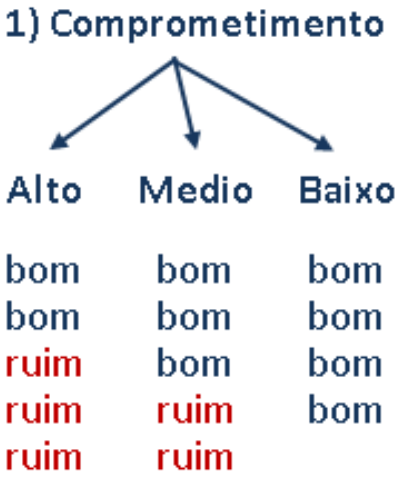
$$ent([3,2]) = -(\frac{3}{5}).\log(\frac{3}{5}) - (\frac{2}{5}).\log(\frac{2}{5}) = 0.97$$

Comprometimento Baixo:

$$ent([4,0]) = 0$$

Decision Tree – Medida de Incerteza: Entropia

Entropia da árvore = Média Ponderada da Entropia de cada grupo



Comprometimento Alto:

$$ent([2,3]) = -(\frac{2}{5}).\log(\frac{2}{5}) - (\frac{3}{5}).\log(\frac{3}{5}) = 0.97$$

Comprometimento Medio:

$$ent([3,2]) = -(\frac{3}{5}).\log(\frac{3}{5}) - (\frac{2}{5}).\log(\frac{2}{5}) = 0.97$$

Comprometimento Baixo:

$$ent(4,0) = 0$$

Para esta árvore:

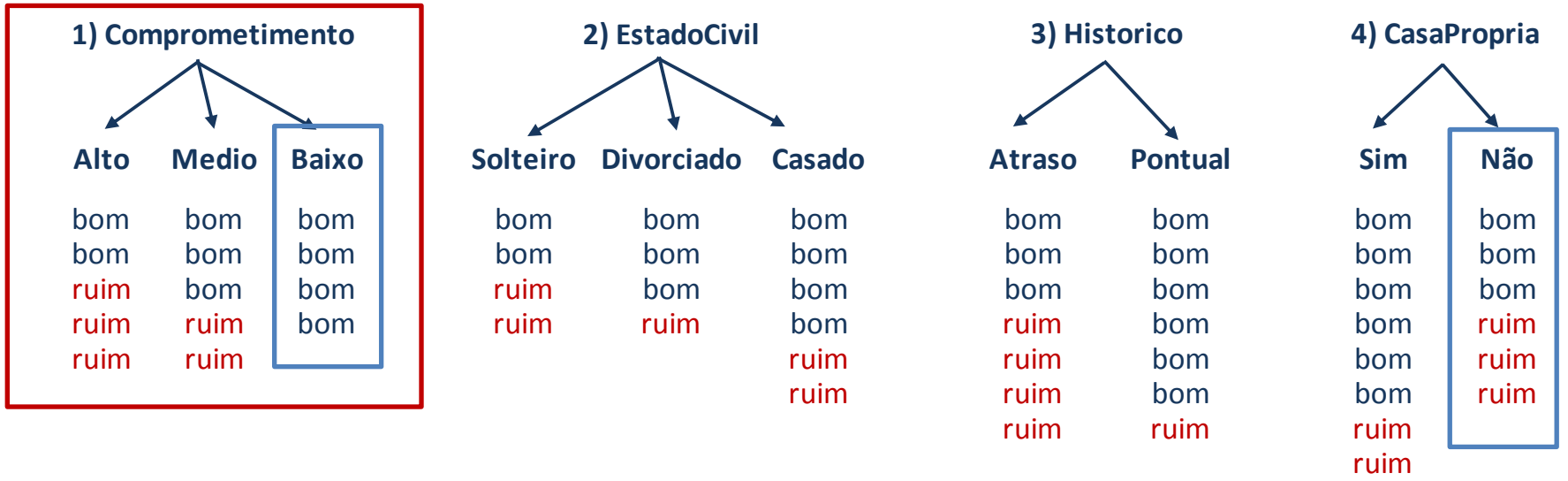
$$ent([2,3],[3,2],[4,0]) = \frac{5}{14} \times 0.97 + \frac{5}{14} \times 0.97 + \frac{4}{14} \times 0 = 0.693$$

Para situação inicial sem nenhuma divisão:

$$ent([9,5]) = -(\frac{9}{14}).\log(\frac{9}{14}) - (\frac{5}{14}).\log(\frac{5}{14}) = 0.94$$

Redução da incerteza: $ent([9,5]) - ent([2,3],[3,2],[4,0]) = 0.94 - 0.693 = 0.24$

Decision Tree – Exemplo Análise de Crédito



bom	2	3	4
ruim	3	2	0
total	5	5	4

2	3	4
2	1	2
4	4	6

3	6
4	1
7	7

6	3
2	3
8	6

Redução da incerteza:

1) $ent([9,5]) - ent([2,3],[3,2],[4,0]) = 0.24$

2) $ent([9,5]) - ent([2,2],[4,2],[3,1]) = 0.029$

3) $ent([9,5]) - ent([3,4],[6,1]) = 0.152$

4) $ent([9,5]) - ent([6,2],[3,3]) = 0.048$

Análise Exploratória de Dados

Melhorar a performance de árvores de decisão:

- *Bagging*
- *Random Forest*
  - Evitar árvores correlacionadas