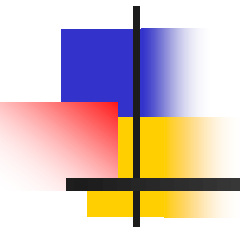


Arvores de Decisão

INTRODUÇÃO

- Árvores de Classificação, Árvores de Decisão e Regressão por Árvore são modelos estatísticos que a partir de uma variável resposta divide o conjunto de dados em grupos de acordo com valores específicos das variáveis explicativas.

- Cada divisão ou nó é determinado por um conjunto específico de valores de uma variável explicativa.

- Os resultados são apresentados em forma de dendograma ou árvore cujos nós identificam grupos de observações caracterizados por valores específicos de variável explicativa.

ARVORES DE DECISÃO

Predição de novos casos	→	Regras de Predição
Seleção de variáveis úteis	→	Busca de partições
Otimização da complexidade	→	Poda

As árvores de decisão são caracterizadas por **regras** organizadas **hierarquicamente**, com estrutura análoga a uma árvore.

ARVORES DE DECISÃO

Para selecionar variáveis explicativas úteis para prever a variável resposta, o Software usa um algoritmo de busca de divisões que considera:

- Uma regra de divisão dicotômica baseada em uma variável exploratória, e
- Uma medida de distância entre os grupos formados.

Quando a variável explicativa está em uma escala intervalar, a divisão dicotômica leva em conta todos os seus distintos valores ordenados como potenciais divisões

ARVORES DE DECISÃO

Sendo a variável resposta binária, tabelas 2x2 são formadas para cada uma das divisões dicotômicas dos valores distintos da variável explicativa, e calculam-se:

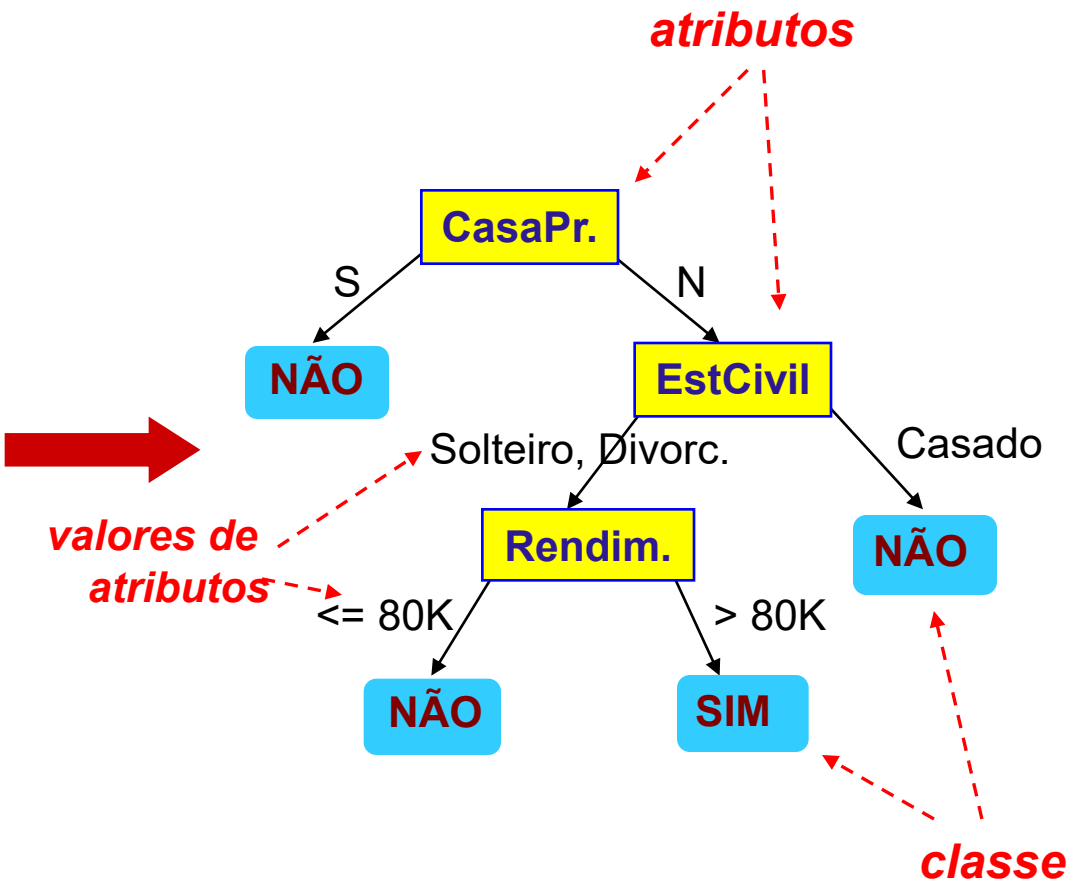
- a estatística Chi-quadrado de Pearson,
- o valor p associado e menos o logaritmo do valor p, que é chamado de logworth = - Log (valor p).

Valores grandes da estatística são indicações de grupos distintos e divisão promissora.

ARVORES DE DECISÃO

	categórico	categórico	continuo	classe
Id	Casa própria	EstCivil	Rendim.	Mau Pagador
1	S	Solteiro	125K	NÃO
2	N	Casado	100K	NÃO
3	N	Solteiro	70K	NÃO
4	S	Casado	120K	NÃO
5	N	Divorc.	95K	SIM
6	N	Casado	60K	NÃO
7	S	Divorc.	220K	NÃO
8	N	Solteiro	85K	SIM
9	N	Casado	75K	NÃO
10	N	Solteiro	90K	SIM

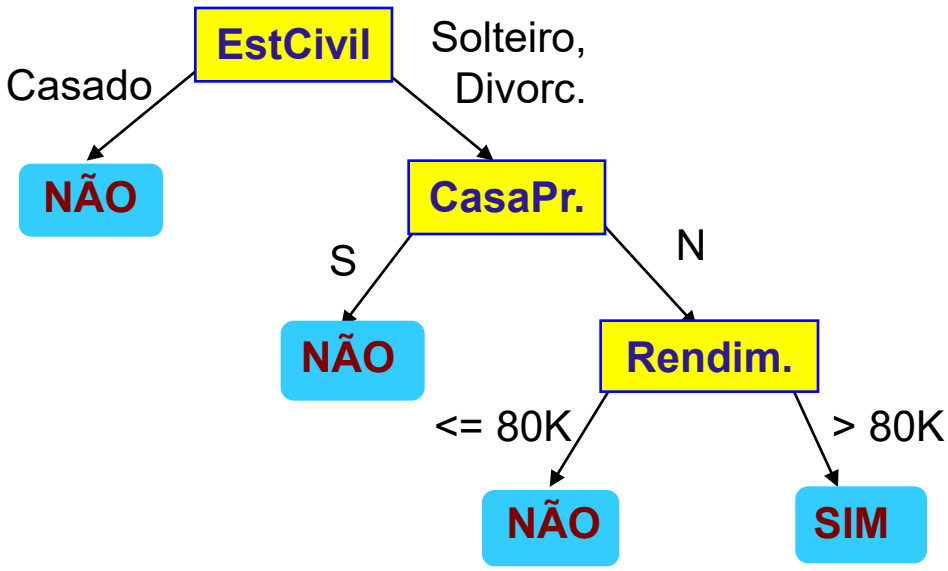
Dados de treinamento



Modelo: árvore de decisão

ARVORES DE DECISÃO

	categórico	categórico	continuo	classe
Id	Casa própria	EstCivil	Rendim.	Mau Pagador
1	S	Solteiro	125K	NÃO
2	N	Casado	100K	NÃO
3	N	Solteiro	70K	NÃO
4	S	Casado	120K	NÃO
5	N	Divorc.	95K	SIM
6	N	Casado	60K	NÃO
7	S	Divorc.	220K	NÃO
8	N	Solteiro	85K	SIM
9	N	Casado	75K	NÃO
10	N	Solteiro	90K	SIM



Pode haver mais de uma árvore para o mesmo conjunto de dados!!!

CRITÉRIO DE DIVISÃO: ENTROPIA

- Vem da Teoria da Informação
- Mede o grau de incerteza de um conjunto de exemplos
- Fórmula: $H(S) = - \Sigma\ p_i \log_2(p_i)$

Intuição:

- Nó puro (só uma classe) $\rightarrow$ Entropia = 0
- Classes balanceadas (50%-50%) $\rightarrow$ Entropia = 1

CRITÉRIO DE DIVISÃO: ÍNDICE DE GINI

- Mede a impureza do nó (probabilidade de erro de classificação)
- Fórmula: $G(S) = 1 - \Sigma p_i^2$

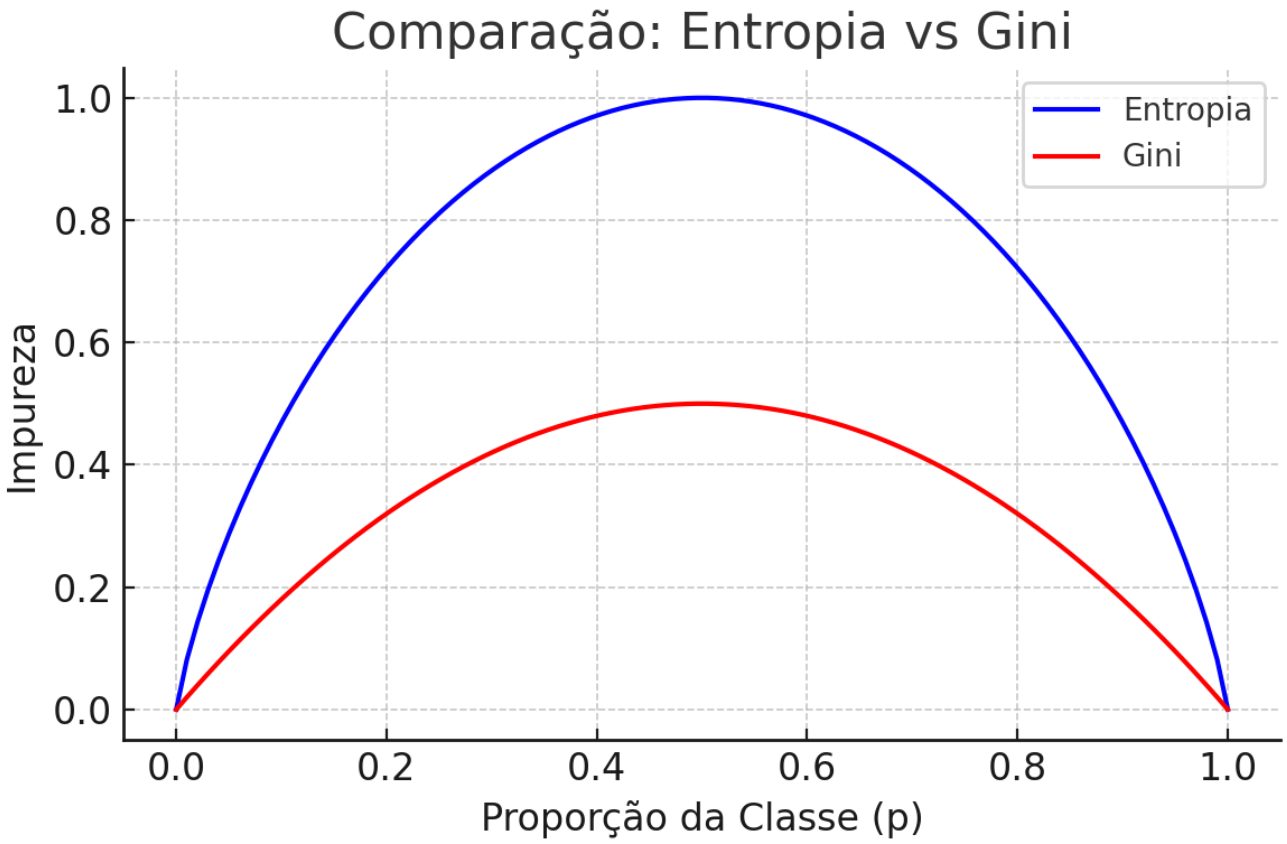
Intuição:

- – Nó puro (só uma classe) → Gini = 0
- – Classes balanceadas (50%-50%) → Gini = 0.5

Comparação:

- Ambos buscam nós mais puros
- Gini é mais rápido → usado como padrão no scikit-learn
- Entropia tende a gerar divisões mais equilibradas

COMPARAÇÃO VISUAL: ENTROPIA VS GINI



ARVORES DE DECISÃO

Além da regra de divisão dicotômica e da medida de grupos distintos, outros fatores também são considerados na formação de uma árvore:

- Grupos são formados com um número mínimo de elementos,
- As observações de variáveis explicativas com valores missing também são utilizadas na análise.
- O processo de partição é repetido para todas as variáveis explicativas no conjunto de treinamento.
- A árvore é formada com a aplicação do algoritmo de divisão em cada subgrupo formado.

MEDIDAS DE AJUSTE – MATRIZ DE CONFUSÃO

Como funciona:

- **Estrutura:**
  - É uma tabela que compara os valores reais (observados) com os valores previstos pelo modelo.
  - As linhas representam as classes reais, e as colunas representam as classes previstas.
  - Cada célula da matriz mostra o número de observações que se encaixam em uma determinada combinação de classe real e classe prevista.
- **Elementos principais:**
  - **Verdadeiros Positivos (VP):** O modelo previu corretamente a classe positiva.
  - **Verdadeiros Negativos (VN):** O modelo previu corretamente a classe negativa.
  - **Falsos Positivos (FP):** O modelo previu incorretamente a classe positiva (erro do tipo I).
  - **Falsos Negativos (FN):** O modelo previu incorretamente a classe negativa (erro do tipo II).

MEDIDAS DE AJUSTE – MATRIZ DE CONFUSÃO

		Valor Predito	
		Sim	**Não**
Real	**Sim**	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

MEDIDAS DE AJUSTE – MATRIZ DE CONFUSÃO

Exemplo seguradora (ocorrência de sinistro)

Observado	Predito		
	Ocorrência de Sinistro	Não-ocorrência de sinistro	Total
Ocorrência de Sinistro	25	7	32
Não-ocorrência de sinistro	5	163	168
Total	30	170	200

MEDIDAS DE AJUSTE – MATRIZ DE CONFUSÃO

ACURÁCIA

É a proporção de acertos do modelo: a capacidade da ferramenta acertar as previsões.

Recall = ACERTOS TOTAIS / TOTAL DE CASOS

= (VP+VN) / (VP + VN + FP + FN)

Observado	Predito		
	Ocorrência de Sinistro	Não-ocorrência de sinistro	Total
Ocorrência de Sinistro	25	7	32
Não-ocorrência de sinistro	5	163	168
Total	30	170	200

Recall= (25+163)/200 = 94% (acertos totais)

MEDIDAS DE AJUSTE – MATRIZ DE CONFUSÃO

QUAL O PROBLEMA DA ACURÁCIA?

MEDIDAS DE AJUSTE – MATRIZ DE CONFUSÃO

RECALL

É a proporção de verdadeiros positivos: a capacidade do sistema em predizer corretamente a condição.

Recall = ACERTOS POSITIVOS / TOTAL DE POSITIVOS

= VP / (VP + FN)

Observado	Predito		
	Ocorrência de Sinistro	Não-ocorrência de sinistro	Total
Ocorrência de Sinistro	25	7	32
Não-ocorrência de sinistro	5	163	168
Total	30	170	200

Recall= 25/32 = 78% (acertos no evento de interesse)

MEDIDAS DE AJUSTE – MATRIZ DE CONFUSÃO

PRECISION

A precisão é a razão entre o número de verdadeiros positivos e o número total de previsões positivas feitas pelo modelo (verdadeiros positivos e falsos positivos)..

SPEC = ACERTOS POSITIVOS / PREVISÃO TOTAL DE POSITIVOS

= VN / (VN + FP)

Observado	Predito		
	Ocorrência de Sinistro	Não-ocorrência de sinistro	Total
Ocorrência de Sinistro	25	7	32
Não-ocorrência de sinistro	5	163	168
Total	30	170	200

Precision = 25/30 = 83% (acertos na previsão)

MEDIDAS DE AJUSTE – MATRIZ DE CONFUSÃO

F1 - SCORE

O F1 Score é uma média harmônica entre precisão e recall. Veja abaixo as definições destes dois termos. Ela é muito boa quando você possui um dataset com classes desproporcionais.

Em geral, quanto maior o F1 score, melhor.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

F1 Score = 2* 83% *78%/(83%+78%) = 0,80

ARVORES DE DECISÃO

- PONTOS FORTES
- ✓ São de fácil compreensão humana;
- ✓ Obtém respostas bastante robustas na presença de problemas como a falta de dados;

- PONTOS FRACOS
- ✓ Podem ser extremamente suscetíveis a pequenas mudanças nos dados.
- ✓ Não conseguem trabalhar com dados que necessitam complexidade nas relações.

RANDOM FOREST

- Random Forest (Floresta aleatória) é uma técnica de aprendizagem de máquina que consiste em muitas árvores de decisão e produz a o resultado de saída como uma combinação de árvores individuais.
- O método combina a ideia de arvore e a seleção aleatória de recursos.

RANDOM FOREST

- Diferentemente do que acontece na criação de uma árvore de decisão simples, ao utilizar o Random Forest, o primeiro passo executado pelo algoritmo será selecionar aleatoriamente algumas amostras dos dados de treino, e não a sua totalidade.
- Com esta primeira seleção de amostras será construída a primeira árvore de decisão.
- No Random Forest, a definição da primeira variável não acontece com base em todas as variáveis disponíveis. O algoritmo irá escolher de maneira aleatória duas ou mais variáveis, e então realizar os cálculos com base nas amostras selecionadas, para definir qual dessas variáveis será utilizada no primeiro nó.

RANDOM FOREST

- Os dados de treinamento para uma árvore individual não são utilizados completamente, para a execução excluem alguns dos dados disponíveis. Os dados retidos do treinamento são chamados de amostra out of bag. Uma árvore individual usa apenas a amostra out of bag para formar previsões. Elas são mais confiáveis do que as previsões dos dados de treinamento.

- A média de árvores com diferentes amostras de treinamento reduz a dependência das previsões em uma amostra de treinamento específica. Aumentar o número de árvores não aumenta o risco de overfitting dos dados e pode reduzi-lo. No entanto, se as previsões de diferentes árvores estiverem correlacionadas, aumentar o número de árvores fará pouca ou nenhuma melhoria.