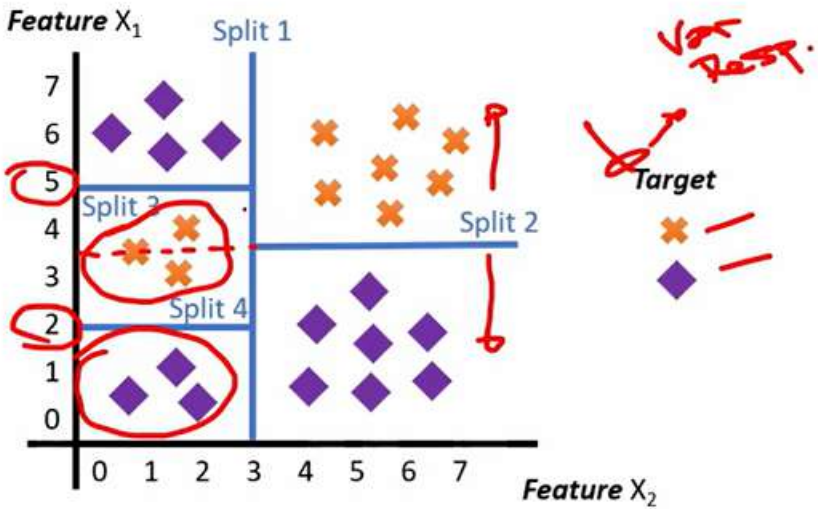
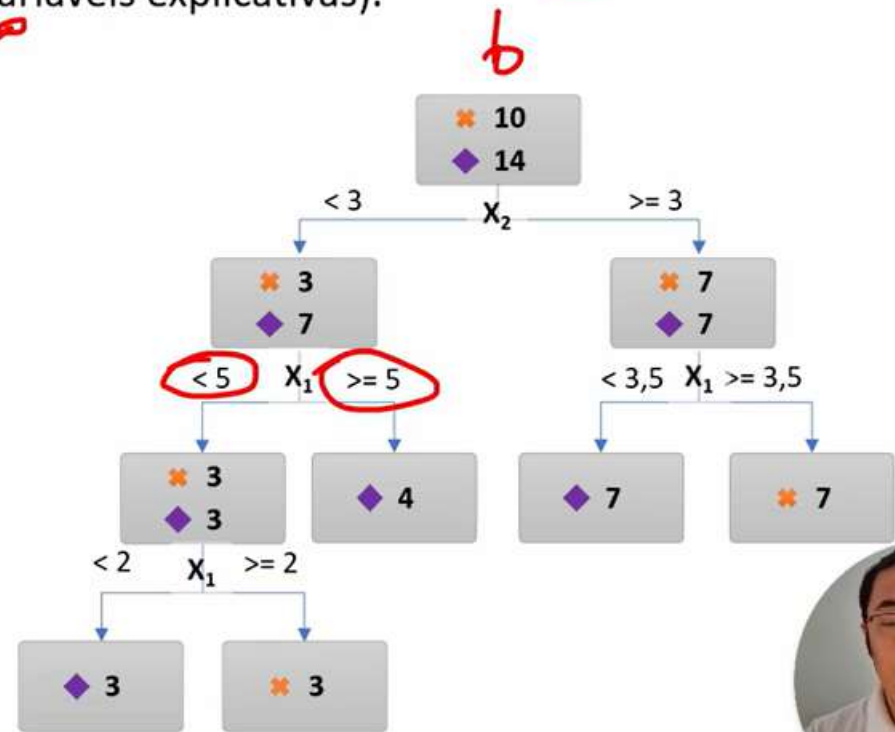
Árvores de Decisão

Intuição

Preditiva.ai

Uma técnica **muito versátil** e bastante utilizada para o desenvolvimento de modelos é a das **Árvores de Decisão**. Neste método, as **observações** são sucessivamente **divididas em grupos** (*splits*) de acordo com **suas características,** representadas pelas ***features*** (variáveis explicativas).



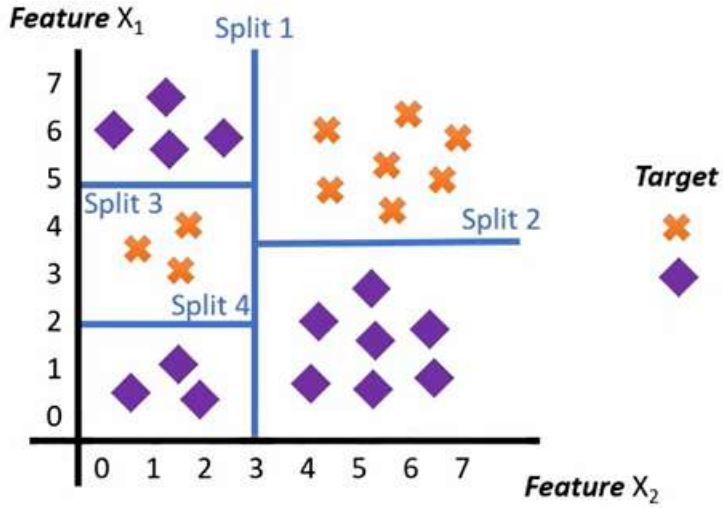


Árvores de Decisão

Intuição

Preditiva.ai

Uma técnica **muito versátil** e bastante utilizada para o desenvolvimento de modelos é a das **Árvores de Decisão**. Neste método, as **observações** são sucessivamente **divididas em grupos** (*splits*) de acordo com **suas características,** representadas pelas ***features*** (variáveis explicativas).



Dessa forma, as observações com **características parecidas** ficam em um **mesmo grupo**, enquanto observações com **características distintas** ficam em **grupos diferentes**.

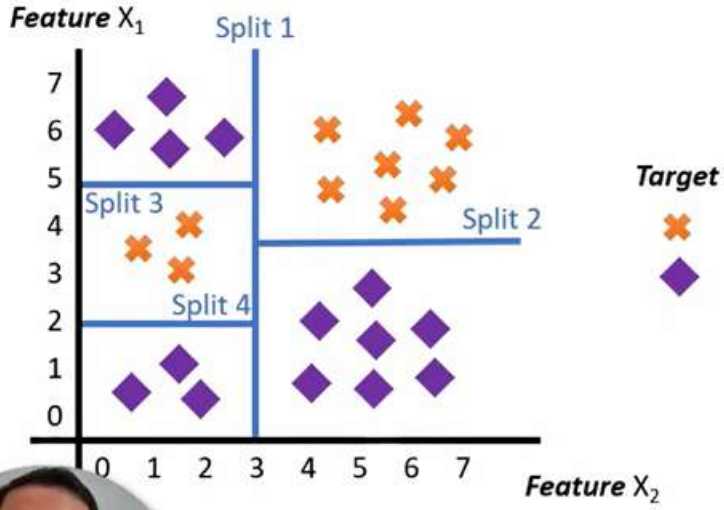
Por ser uma técnica de **aprendizado supervisionado**, a forma como as observações são divididas busca exatamente **minimizar o erro** nas estimativas do modelo em relação ao ***target*** (variável resposta).

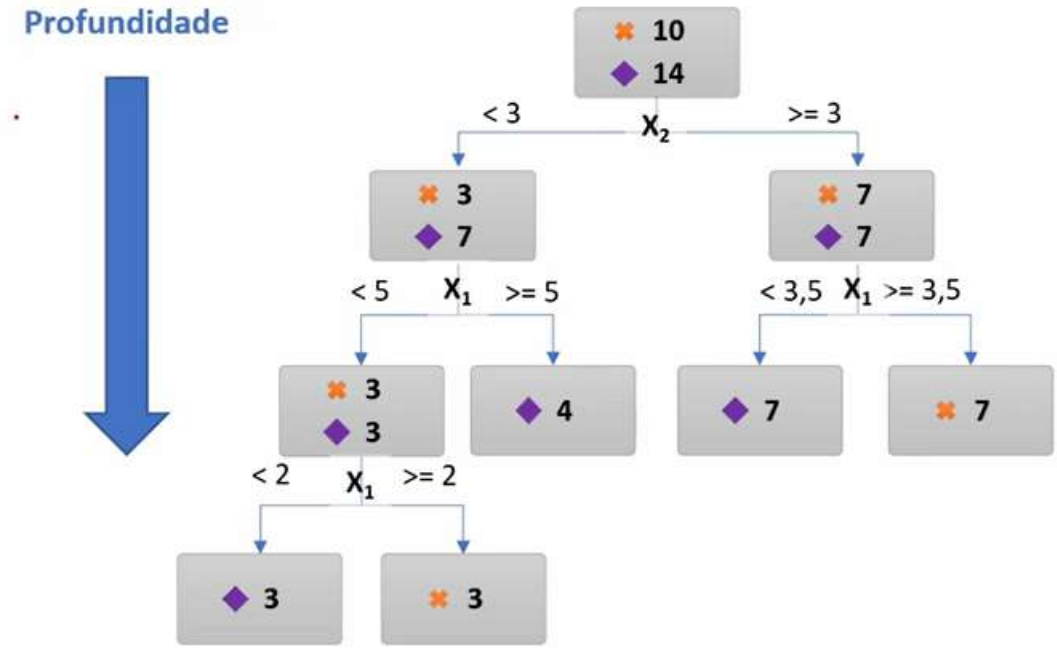
Árvores de Decisão

Intuição

Preditiva.ai

A **estrutura** das **Árvores de Decisão** podem influenciar fortemente o **desempenho do modelo** desenvolvido. Mais adiante veremos como escolher os **hiperparâmetros** que definem essa estrutura de forma a obter a **melhor combinação** para cada conjunto de dados.



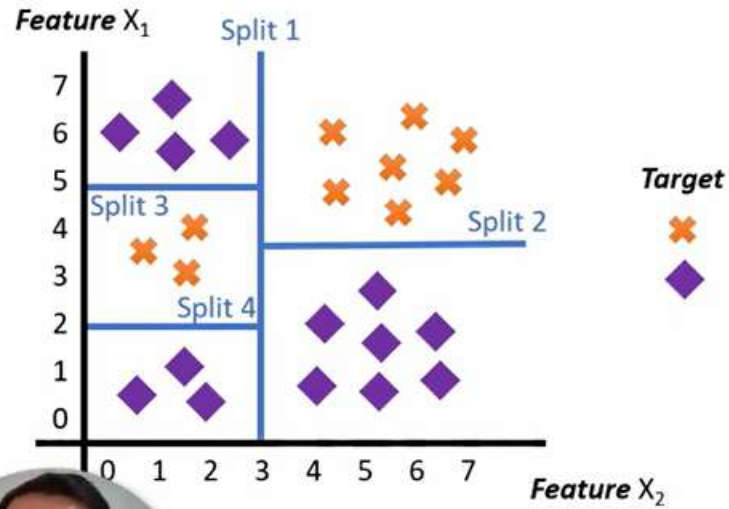


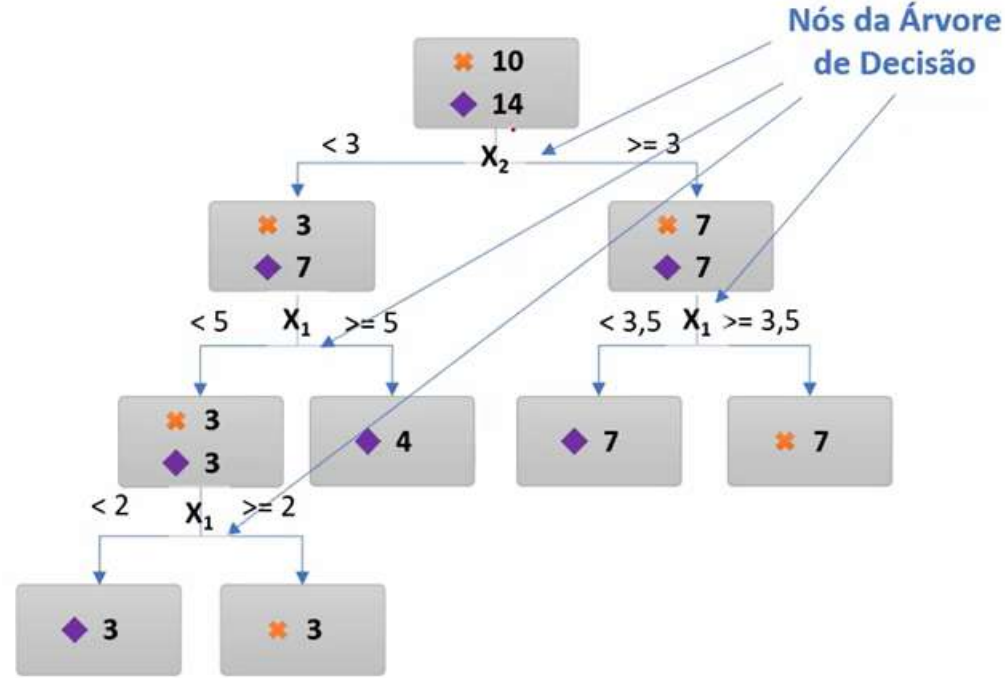
Árvores de Decisão

Intuição

Preditiva.ai

A **estrutura** das **Árvores de Decisão** podem influenciar fortemente o **desempenho do modelo** desenvolvido. Mais adiante veremos como escolher os **hiperparâmetros** que definem essa estrutura de forma a obter a **melhor combinação** para cada conjunto de dados.



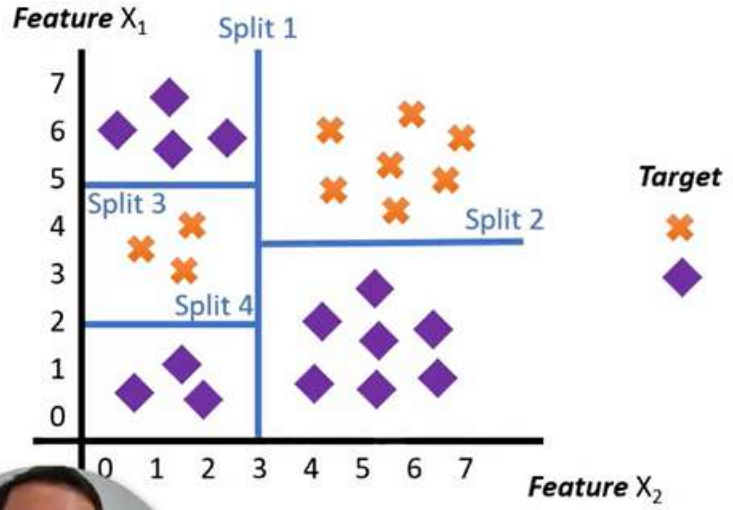


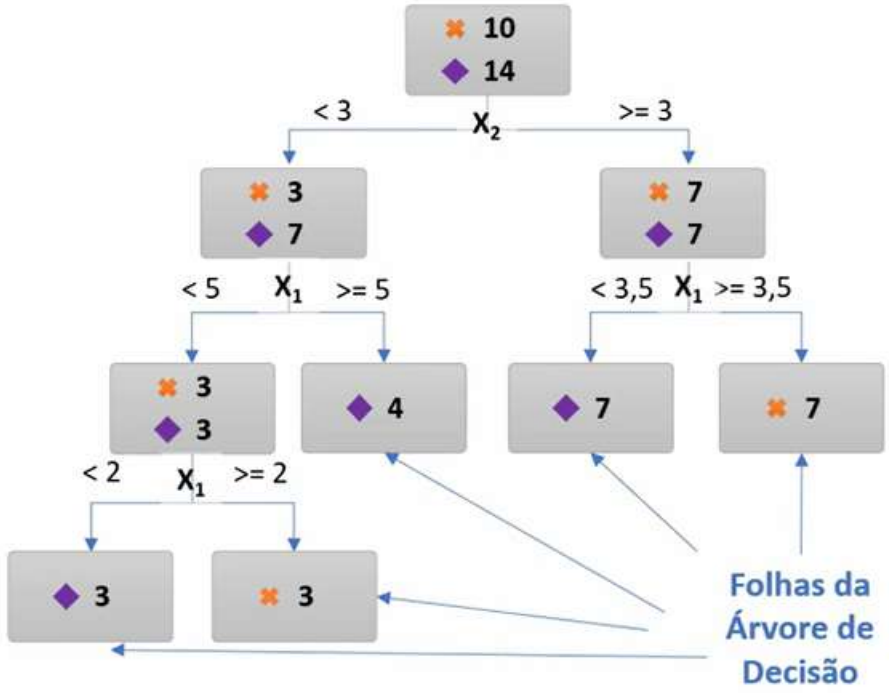
Árvores de Decisão

Intuição

Preditiva.ai

A **estrutura** das **Árvores de Decisão** podem influenciar fortemente o **desempenho do modelo** desenvolvido. Mais adiante veremos como escolher os **hiperparâmetros** que definem essa estrutura de forma a obter a **melhor combinação** para cada conjunto de dados.





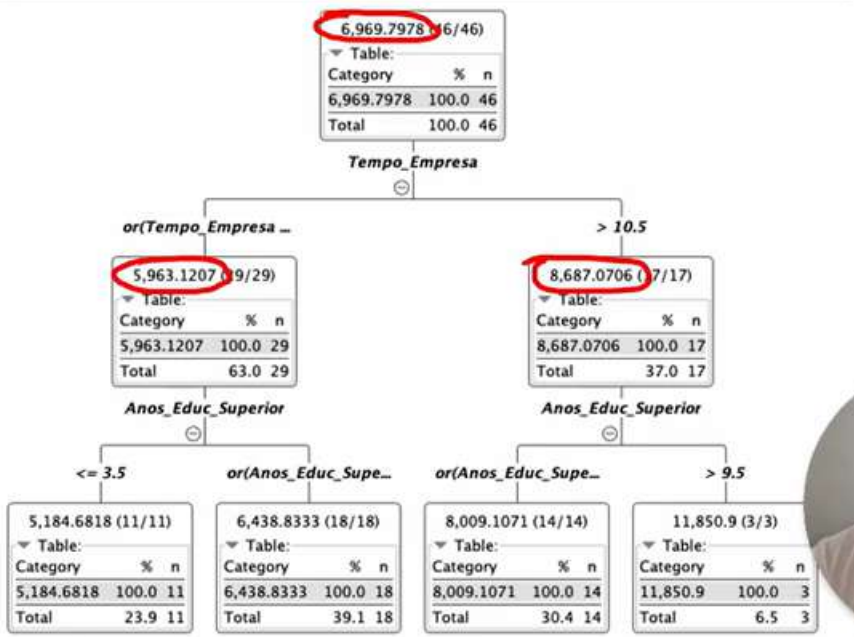
Árvores de Decisão

Intuição

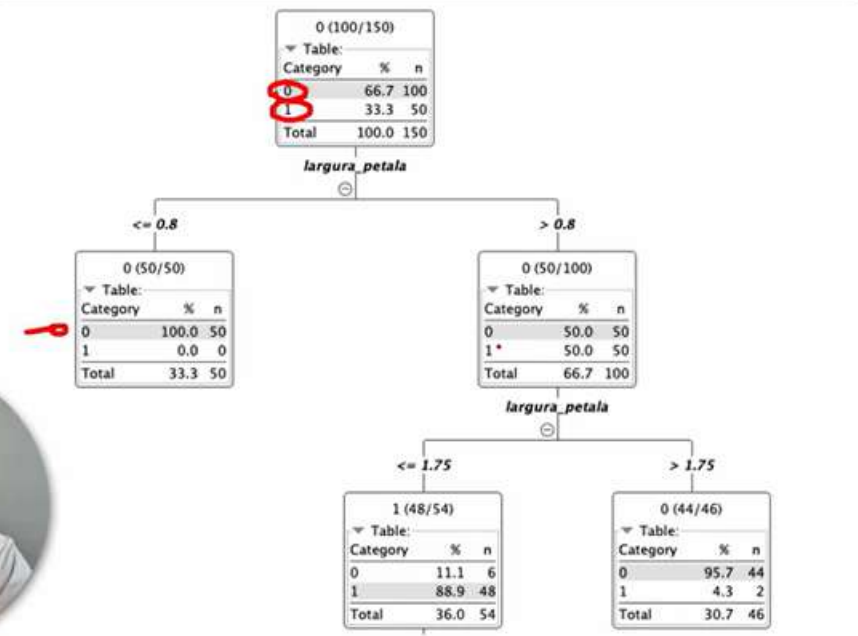
Preditiva.ai

Uma característica que demonstra muito bem a versatilidade das **Árvores de Decisão** é que podemos utilizar essa técnica para tarefas de **Regressão** (*target* quantitativo) ou **Classificação** (*target* qualitativo), utilizando ***features*** quantitativas e/ou qualitativas.

REGRESSÃO



CLASSIFICAÇÃO



Árvores de Decisão

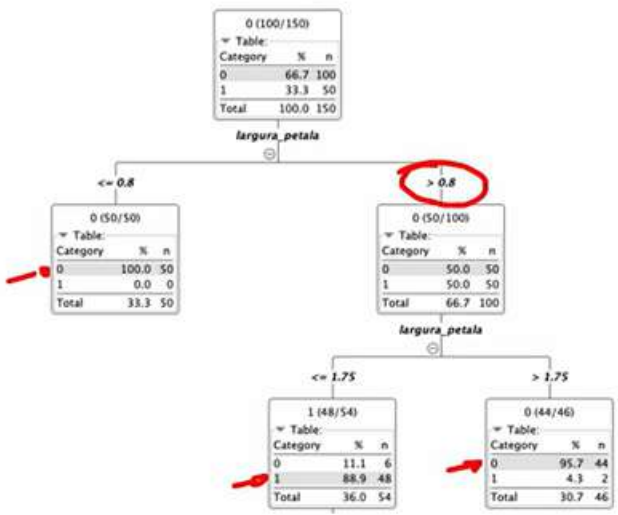
Intuição

Preditiva.ai

Apesar de serem uma técnica bastante simples, as **Árvores de Decisão** apresentam algumas **características muito positivas** para o desenvolvimento de modelos.

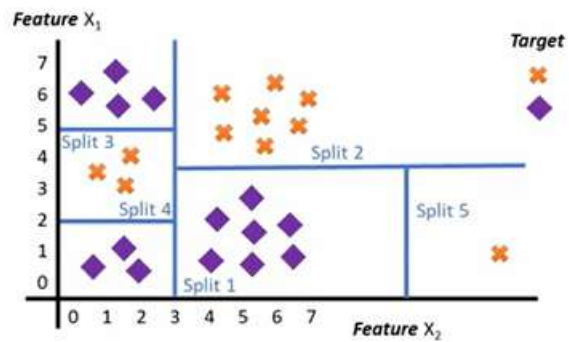
SIMPLES DE ENTENDER E INTERPRETAR

As regras resultantes das divisões podem ser facilmente entendidas e o modelo facilmente **interpretável**.



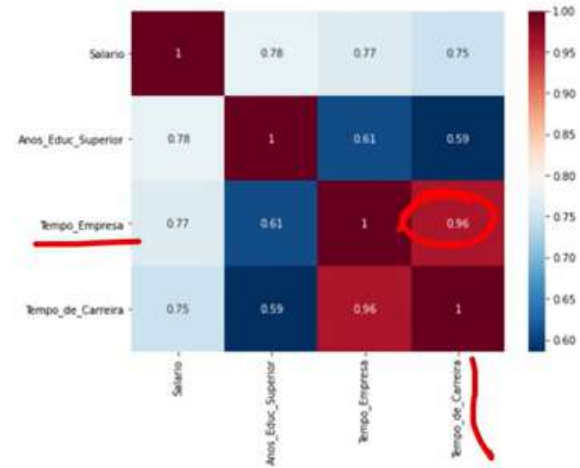
REQUER POUCO DATA PREP

Observações com **valores extremos (*outliers*) ou *missing values*** podem ser simplesmente separadas das demais observações.



ROBUSTA A MULTICOLINEARIDADE

As ***features*** são selecionadas **individualmente** em cada etapa do **algoritmo de aprendizagem**.



Árvores de Decisão

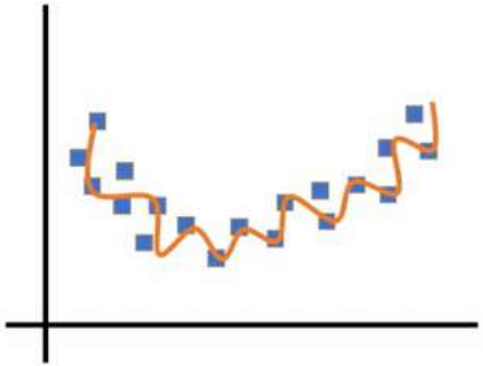
Intuição

Preditiva.ai

Mas assim como outras técnicas, as **Árvores de Decisão** também possuem **limitações** que podem comprometer a qualidade dos modelos desenvolvidos.

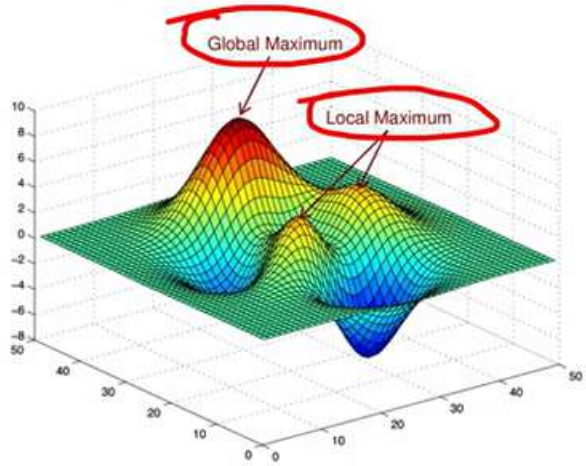
OVERFITTING

O modelo desenvolvido pode ser bastante complexo, fazendo com que ele **perca capacidade de generalização**.



APRENDIZADO POR HEURÍSTICA

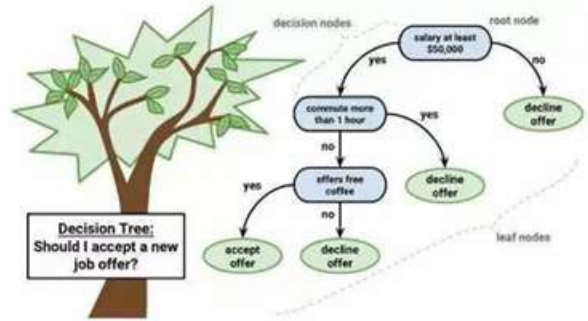
Ainda é **computacionalmente inviável** testar todas as possíveis divisões de cada *feature*.



Fonte: https://www.researchgate.net/figure/11-Illustration-of-local-optimum-and-global-optimum_fig16_306558608

INSTÁVEL COM POUCOS DADOS

Mudança em poucas observações pode gerar **alteração das divisões** e modificar o resultado de forma expressiva.



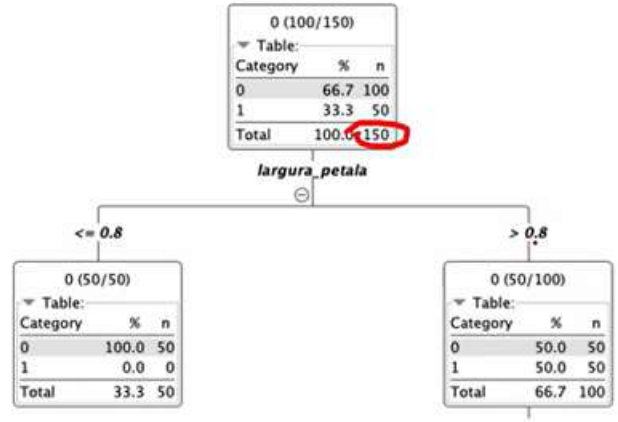
Fonte: https://www.analyticsvidhya.com/blog/2021/05/25-questions-to-test-your-skills-on-decision-trees

Árvores de Decisão

Classificação

Preditiva.ai

Em tarefas de **classificação** o objetivo do algoritmo de **aprendizagem** é **maximizar a separação** entre as diferentes classes do *target* utilizando as *features* disponíveis.



Neste exemplo temos 2 tipos de flores:

- 1: **Orquídeas**
- 0: **Não orquídeas**

A 1ª divisão foi utilizando a *feature* **largura_pétala**. As flores com a **largura_petala <= 0.8** cm são todas "**não orquídeas**". Logo, a probabilidade de uma flor nesse grupo ser "**não orquídeas**" é de **100%**.

Árvores de Decisão

Classificação

Preditiva.ai

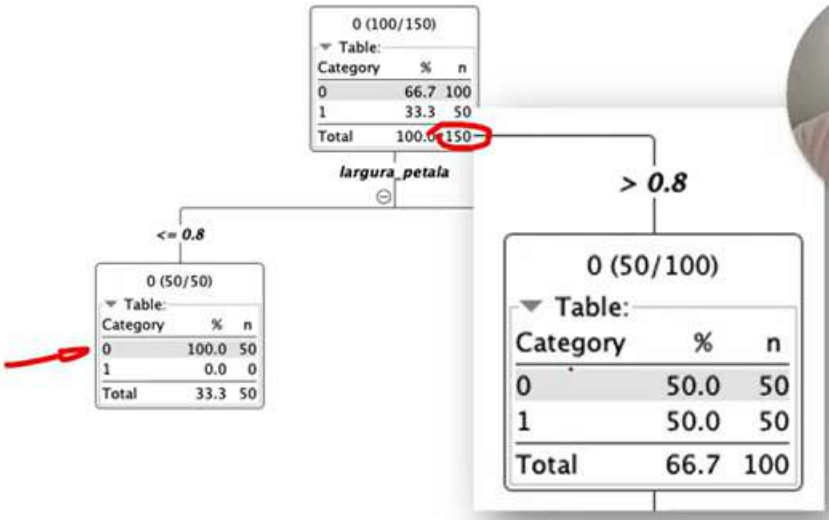
Em tarefas de **classificação** o objetivo do algoritmo de **aprendizagem** é **maximizar a separação** entre as diferentes classes do *target* utilizando as *features* disponíveis.

Neste exemplo temos 2 tipos de flores:

- 1: **Orquídeas**
- 0: **Não orquídeas**

A 1ª divisão foi utilizando a *feature* **largura_pétala**. As flores com a **largura_petala <= 0.8** cm são todas "**não orquídeas**". Logo, a probabilidade de uma flor nesse grupo ser "**não orquídeas**" é de **100%**.

Já as flores com a **largura_petala > 0.8** estão misturadas entre "**orquídeas**" e "**não orquídeas**" na proporção de **50%** cada. Ou seja, a probabilidade de uma flor nesse grupo ser "**orquídeas**" é de **50%**.



Árvores de Decisão
Classificação

Preditiva.ai

Em tarefas de **classificação** o objetivo do algoritmo de **aprendizagem** é **maximizar a separação** entre as diferentes classes do *target* utilizando as *features* disponíveis.

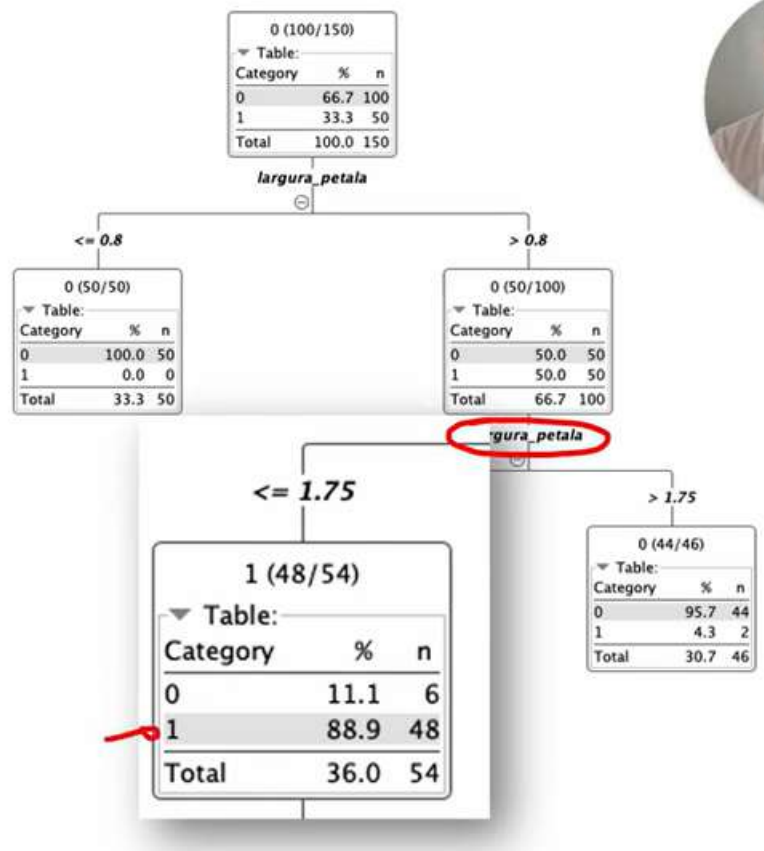
Neste exemplo temos 2 tipos de flores:

- 1: **Orquídeas**
- 0: **Não orquídeas**

Na 2ª divisão foi utilizada novamente a feature **largura_petala**.

O grupo de flores com a **largura_petala <= 1.75** cm tem uma concentração maior de "**orquídeas**", e a probabilidade de uma flor nesse grupo ser "**orquídea**" é de **88,9%**.

Já o grupo com a **largura_petala > 1.75** tem uma concentração maior de "**não orquídeas**". Nesse grupo a probabilidade de uma flor ser "**não orquídeas**" é de **95,7%**.



Árvores de Decisão

Algoritmo CART - *Classification and Regression Trees*

Preditiva.ai

Algoritmo CART para Classificação

1. Seleção de um par (**x** , **s**) sendo **x** uma das ***features*** disponíveis na base de dados e **s** um valor para **divisão** usando essa ***feature***;
2. Cálculo da **função de erro** conforme abaixo:
   - $F\ (x, s) = \frac{qte\ de\ observs_{nó\ da\ esquerda}}{qte\ de\ observs\ do\ nó\ superior} * Impureza_{nó\ da\ esquerda} + \frac{qte\ de\ observs_{nó\ da\ direita}}{qte\ de\ observs\ do\ nó\ superior} * Impureza_{nó\ da\ direita}$
3. Seleção do par (**x** , **s**) que **minimiza** a função de erro acima;
4. **Divisão** utilizando o par selecionado e **repetição do processo** até que algum **critério de parada** seja atingido, por exemplo: máxima profundidade, número mínimo de observações por nó, entre outras.

Como medida de **Impureza**, podemos usar as medidas **Gini** ou **Entropia**. Como as duas produzem resultados muito parecidos, a maior parte das bibliotecas escolhem a medida **Gini** por ser mais rápida em ser calculada.

$$Gini = 1 - \sum_{k=1}^{n} {p_{i,k}}^2$$, sendo que $p_{i,k}$ é a razão da classe "**k**" pelo total de observações do nó "**i**".

Salvamento Automático — Pasta2 - Excel — Pesquisar (Alt+G) — Fernando Itano

Arquivo Página Inicial Inserir Desenhar Layout da Página Fórmulas Dados Revisão Exibir Desenvolvedor Ajuda Power Pivot — Comentários — Compartilhar

Salvar Valores Valores e Formatação de Número Fórmulas Formatação Colar Especial Novo

G18 =1-G16

$$Gini = 1 - \sum_{k=1}^{n} {p_{i,k}}^2$$, sendo que $p_{i,k}$ é a razão da classe "**k**" pelo total de observações do nó "**i**".

Classe	Nº Obs
0	100
1	50

≤ 0,8 > 0,8 log. pet

Classe	Nº Obs
0	100
1	0

Classe	Nº Obs
0	50
0	50

p1,0	1,00	p2,0	0,25
p1,1	0,00	p2,1	0,25
	1,00		0,50

Impureza Esquerda	**0,00**	**Impureza Direita**	**0,50**

Planilha1

Pronto Calcula