Aula 19 - Árvore de Decisão - Decision Tree

Profa. Gabrielly Queiroz

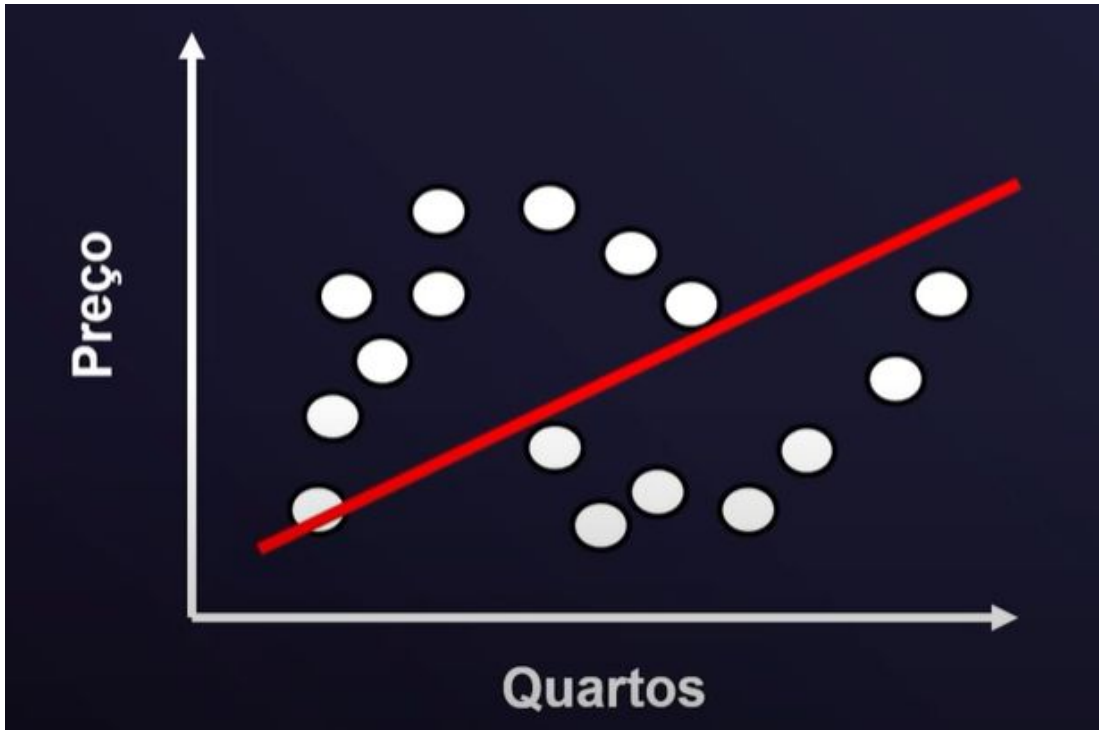
Árvores de Decisão

- As árvores de decisão foram criadas para resolver problemas onde é necessário tomar decisões claras e objetivas com base em dados. Elas simulam o processo humano de tomada de decisão, como responder a uma série de perguntas **sim/não** até chegar a uma conclusão.
- "Devo aprovar ou não o crédito de um cliente?" Renda do cliente. Histórico de pagamento. Dívida atual.
- Antes das árvores de decisão, a abordagem seria usar **um único modelo matemático**, como a regressão linear ou logística. Esses modelos são bons, mas podem sofrer com o **problema de viés e variância**, que afeta sua precisão.
- **Viés:** simplicidade excessiva. Um modelo com alto viés pode ser **muito rígido** e não captar a complexidade dos dados. Exemplo: Usar uma regra simples como "Se a renda for maior que X, aprove o crédito" pode ignorar outros fatores importantes.
- **Variância: excesso de ajuste** do modelo. Um modelo com alta variância tenta **memorizar os dados** em vez de generalizar, ou seja, ele funciona bem no treinamento, mas falha com novos dados. Exemplo: Criar uma regra extremamente específica para cada cliente nos dados históricos pode levar a decisões confusas para novos clientes.

Árvores de Decisão

Regressão Linear:

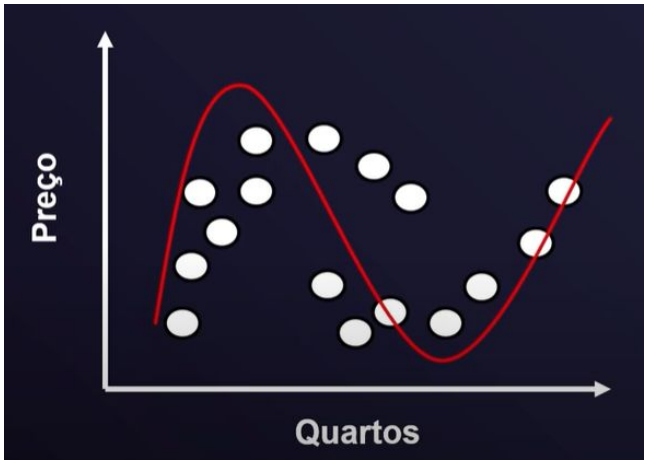
- Alto Viés e Baixa variância

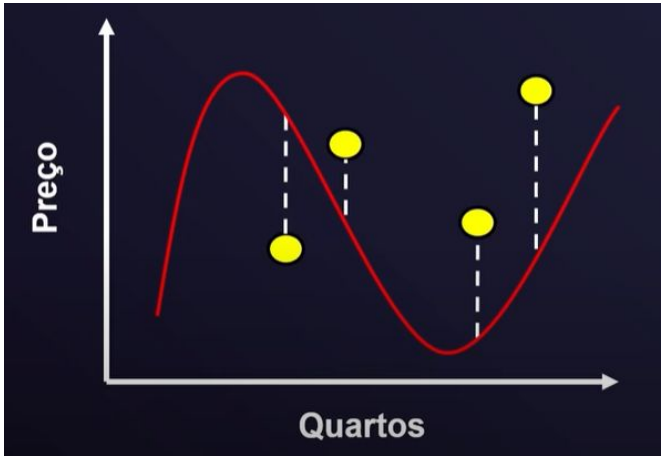


Árvores de Decisão

Regressão Polinomial:

- Baixo Viés e Alta Variância





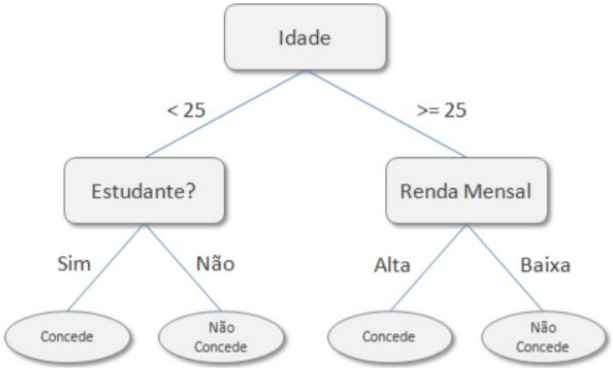
Árvores de Decisão

Uma árvore de decisão é um modelo de aprendizado supervisionado que divide os dados em subgrupos com base em perguntas ou condições

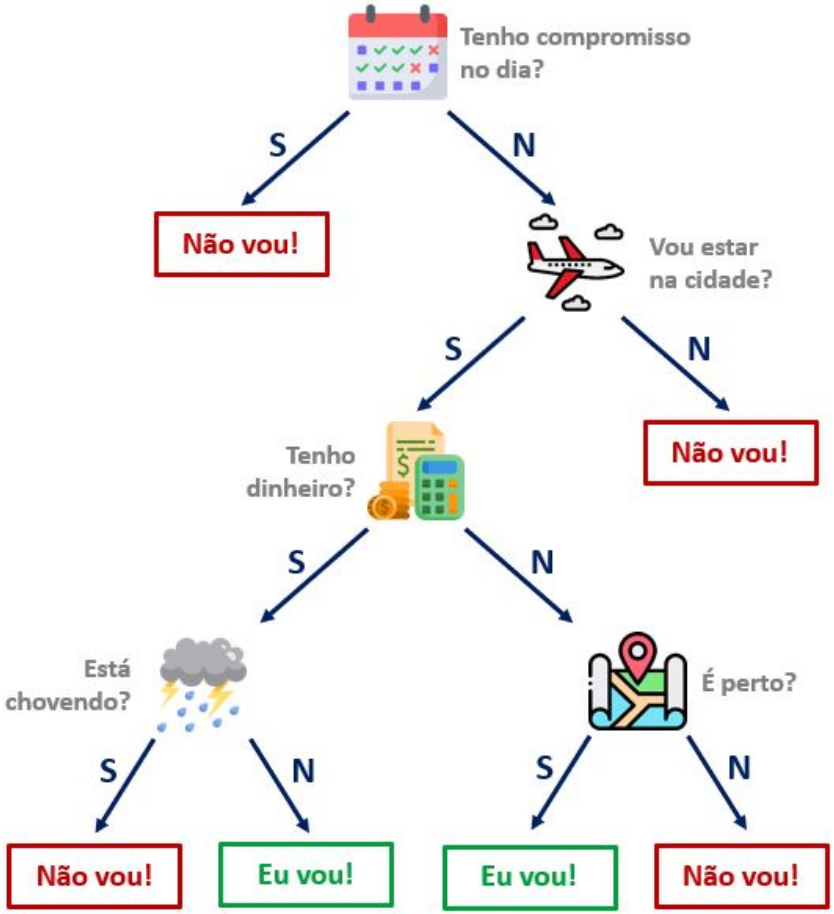
As árvores de decisão foram criadas para encontrar um equilíbrio entre **viés e variância**:

1. **Regras gerais no topo:** A árvore começa simples, com uma divisão ampla, como **"Idade < 25?"**, que captura padrões gerais e reduz o **viés**.
2. **Regras específicas em níveis inferiores:** As divisões se tornam mais refinadas, como **"Estudante?"** ou **"Renda Alta?"**, garantindo que as decisões considerem detalhes relevantes sem serem excessivamente complexas, reduzindo a **variância**.

Esse processo hierárquico evita decisões simplistas (alto viés) ou específicas demais (alta variância), criando um modelo equilibrado e eficaz.



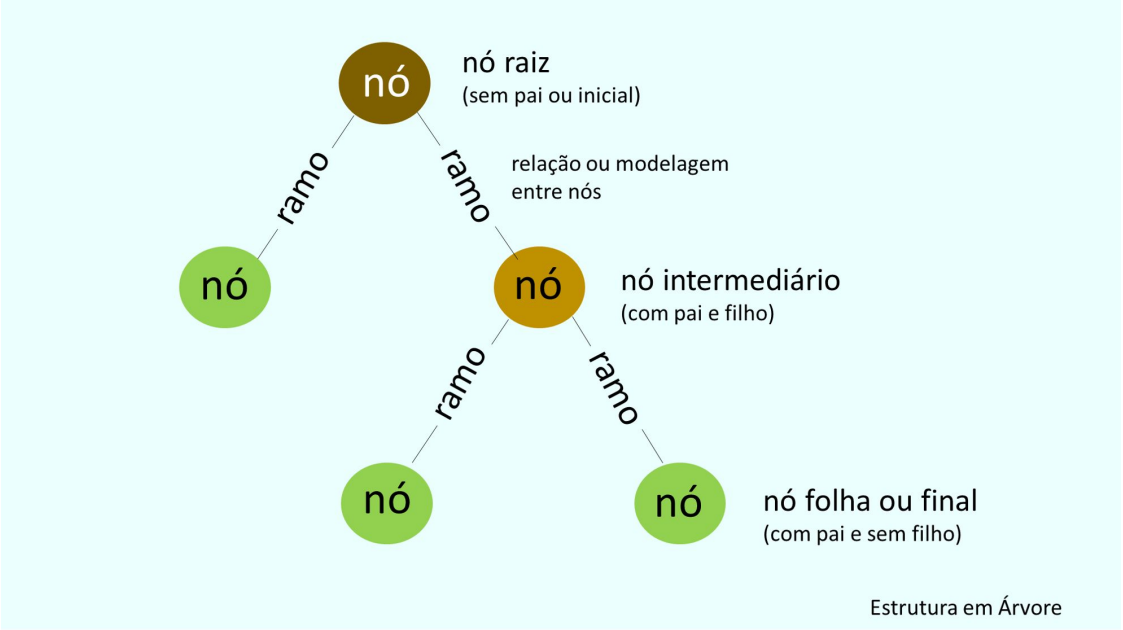
Árvores de Decisão



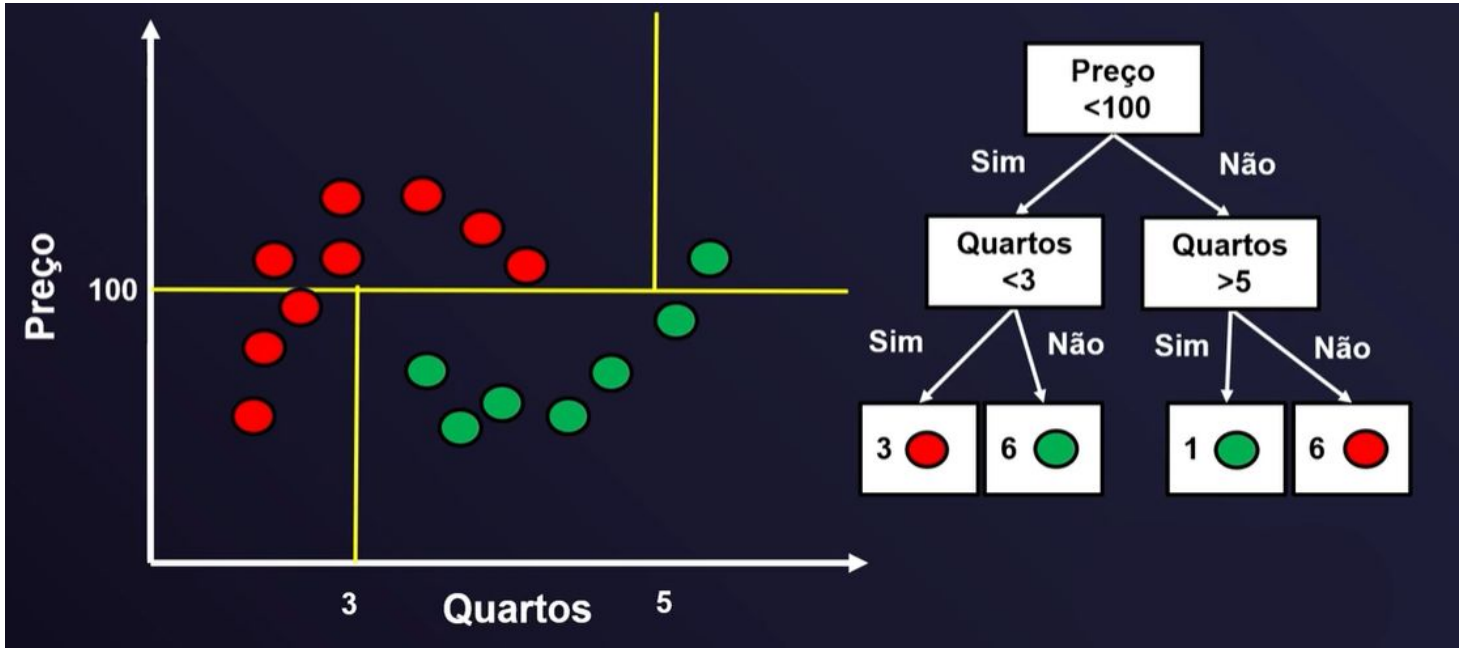
Árvores de Decisão

Uma **árvore de decisão** é uma **estrutura hierárquica** usada para tomar decisões com base em regras. Ela é composta por:

- **Nó Raiz (Root):** primeiro ponto de decisão.
- **Nós Internos:** representam testes lógicos sobre atributos.
- **Folhas (Leaf Nodes):** representam decisões finais (classes ou valores).



Árvores de Decisão

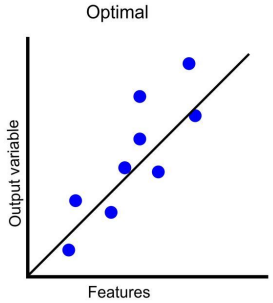


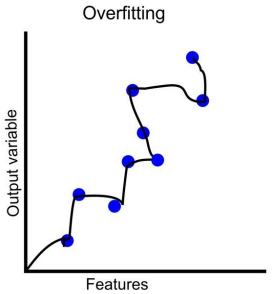
Árvores de Decisão

Desvantagens:

Tamanho excessivo - Árvore pode crescer muito e levar ao overfitting.

Overfitting é quando o modelo **decora os dados de treino**, mas não consegue prever bem em situações novas.





Árvores de Decisão

A **árvore de decisão por si só é apenas uma estrutura de representação** (uma forma de organizar regras).

O que caracteriza o **Aprendizado de Máquina (Machine Learning)** é o **processo automático de escolher as melhores regras** para construir essa árvore com base em dados.

Árvore de Decisão

Algoritmo	Tipo	Critério de Divisão	Pontos-Chave
ID3	Classificação	Ganho de Informação (Entropia)	Simples, sem poda, não lida com valores contínuos. Base teórica para outros.
C4.5	Classificação	Razão de Ganho	Evolução do ID3. Suporta valores contínuos e faltantes. Faz poda.
CART	Classificação e Regressão	Gini (classificação), MSE (regressão)	Usa apenas divisões binárias. Muito eficiente. Base do Scikit-learn.
CHAID	Classificação	Teste Qui-quadrado	Usa estatística para ramificação múltipla. Comum em pesquisas e análises sociais.

Critérios de Divisão de Nós

Índice de Gini

Gini = 0 → Conjunto **puro**: todos os exemplos pertencem à mesma classe.

Gini próximo de 1 → Conjunto **impuro**: os exemplos estão bem distribuídos entre várias classes.

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

n é o número de classes possíveis,

p_i é a **proporção de elementos da classe** i no conjunto S.

Índice de Gini

Suponha um conjunto S com 10 exemplos:

- 6 da classe **Sim**
- 4 da classe **Não**

Cálculo:

$$p_{\text{Sim}} = \frac{6}{10} = 0.6, \quad p_{\text{Não}} = \frac{4}{10} = 0.4$$

$$\text{Gini}(S) = 1 - (0.6^2 + 0.4^2) = 1 - (0.36 + 0.16) = 1 - 0.52 = 0.48$$

Esse valor indica que o conjunto ainda está um pouco misturado (não puro).

Como funciona uma Árvore de Decisão

Seleciona-se o **atributo mais informativo** para dividir os dados.

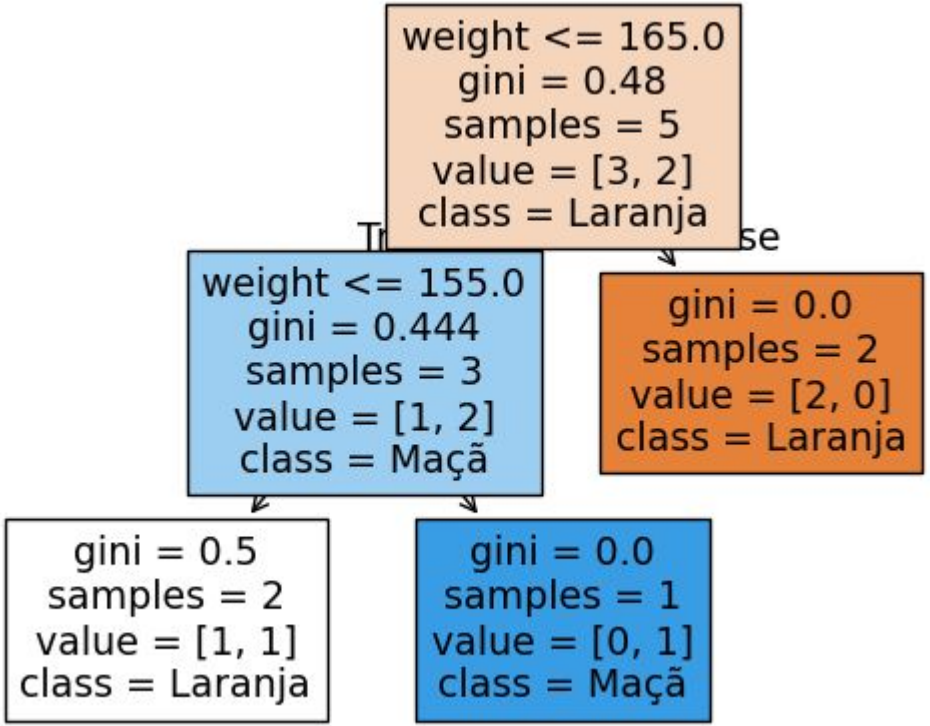
Os dados são separados em **subconjuntos** com base nessa divisão.

O processo se repete **recursivamente** para cada subconjunto.

Quando não há mais ganhos significativos ou os dados estão puros, o nó vira **folha**.

CART

Peso das frutas em gramas e Textura.



Resumo

Algoritmo	Tipo	Uso Principal	Vantagens	Desvantagens
Regressão Linear	Supervisado (Regressão)	Predizer valores contínuos (ex.: preço, peso).	Simples, fácil de interpretar; bom para relações lineares.	Fraco para relações não lineares; sensível a outliers.
Regressão Logística	Supervisado (Classificação)	Classificar em duas ou mais categorias (ex.: spam/não spam).	Interpretação probabilística; funciona bem para problemas lineares.	Fraco para relações complexas; sensível a dados desbalanceados.
K-means (Clusterização)	Não Supervisado	Agrupar dados com características semelhantes (ex.: segmentação de clientes).	Simples, rápido para grandes dados; útil para clusters bem definidos.	Sensível à escolha de kkk; pode convergir para soluções locais.
Árvores de Decisão	Supervisado (Classificação/Regressão)	Tomar decisões com base em perguntas hierárquicas (ex.: aprovar crédito).	Fácil de interpretar; trabalha bem com dados categóricos e contínuos.	Propensa ao overfitting; pode criar árvores muito complexas sem poda.

Colab

https://colab.research.google.com/drive/1oAgC8NnWSBkscDbDPjeN0hJ7yUN1myDk?usp=sharing

Atividade

Escolha uma base de dados (do Scikit-learn, do Kaggle ou um arquivo CSV). Faça o carregamento dos dados, selecione colunas que sirvam como atributos de entrada (X) e uma coluna como rótulo (y). Em seguida, divida os dados em treino e teste, treine um modelo de árvore de decisão com o Scikit-learn, avalie o desempenho do modelo (acurácia) e faça pelo menos três previsões para novos dados simulados. Mostre também a visualização da árvore.