

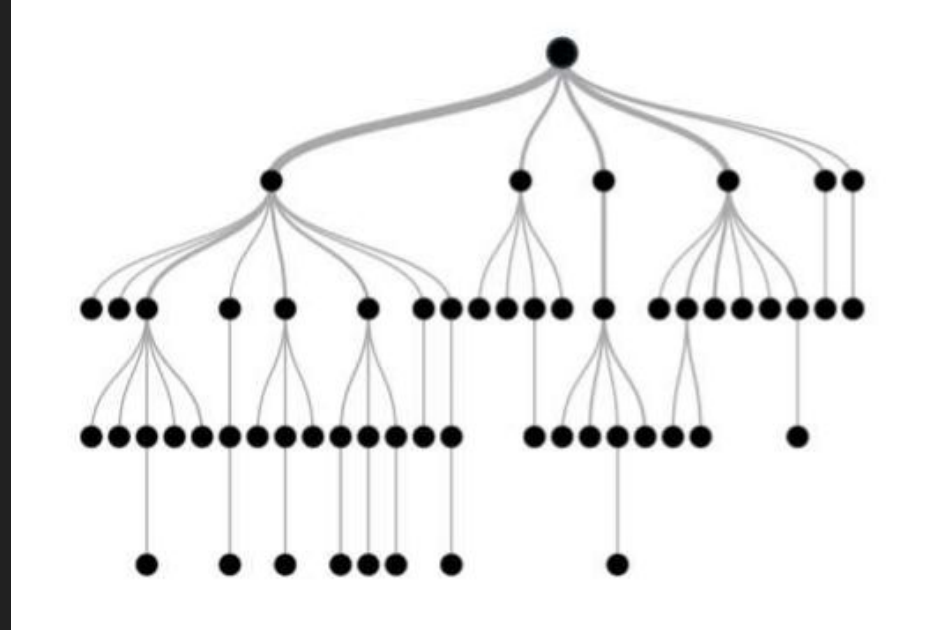
# Data Science do ZERO

Capítulo 06 - Machine Learning

**Árvores de Decisão**

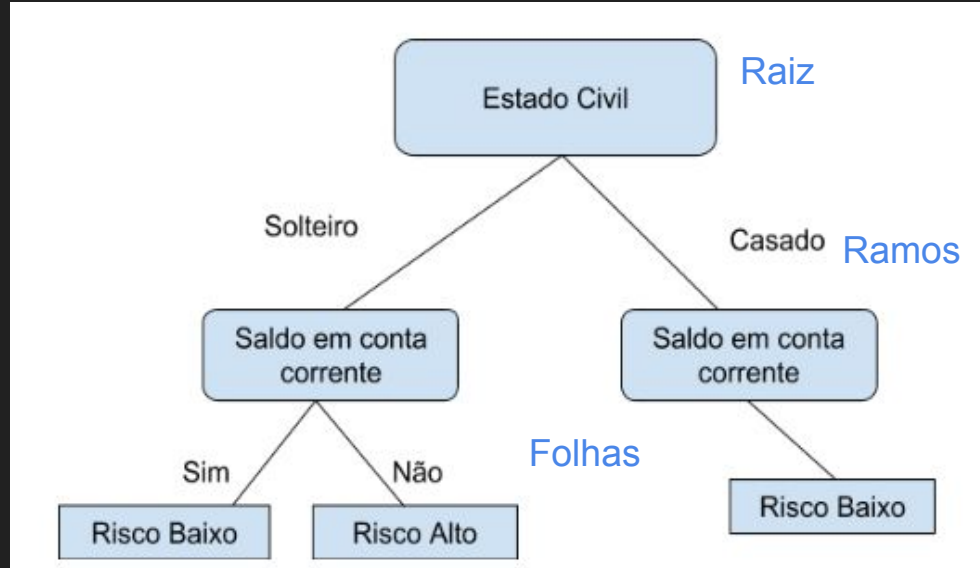
# Árvores de Decisão

- Algoritmo de **Machine Learning** Supervisionado utilizado para Classificação ou regressão.
- Consiste na representação em forma de árvore.
- As árvores possuem 3 nós, por exemplo: raiz, ramos e folhas.
- A raiz e os ramos são pontos de checagem.
- Ao percorrer cada nó o algoritmo toma decisões.
- As folhas são os resultados finais.



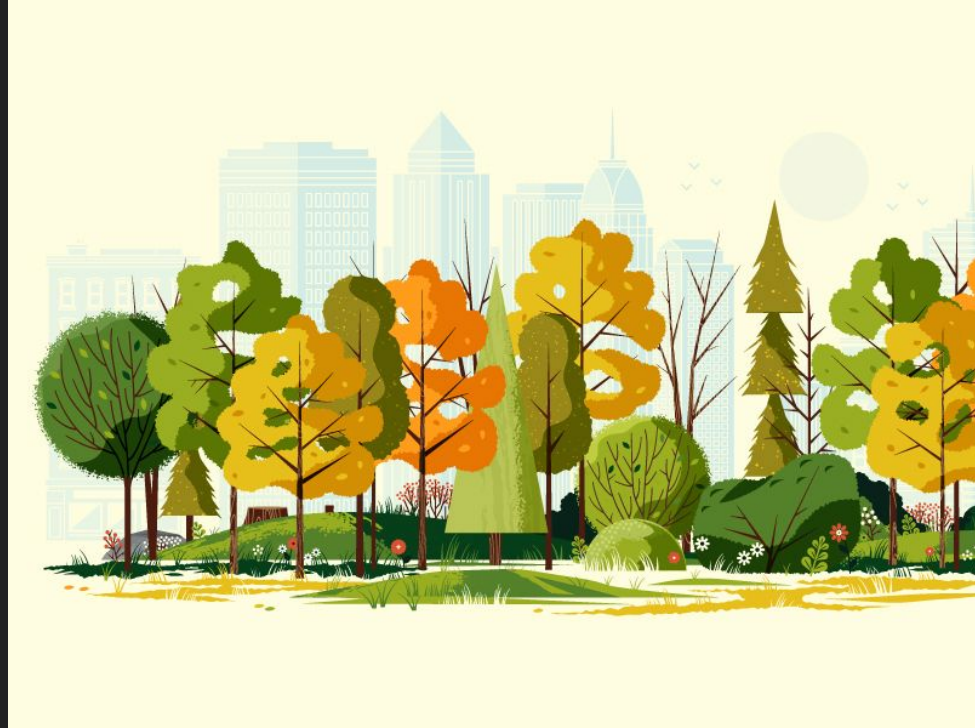
# Árvores de Decisão

- As árvores são construídas a partir da indução de regras
- Para cada regra são feitas decisões que ditam a estrutura da árvore.
- Por isso o nome árvore de decisão.
- Veja no exemplo as raízes, ramos e folhas da árvore a seguir.
- Perceba que os valores dos atributos são decisões a serem tomadas.



# Árvores de Decisão

- Por que estudar árvores de decisão?



# Árvores de Decisão

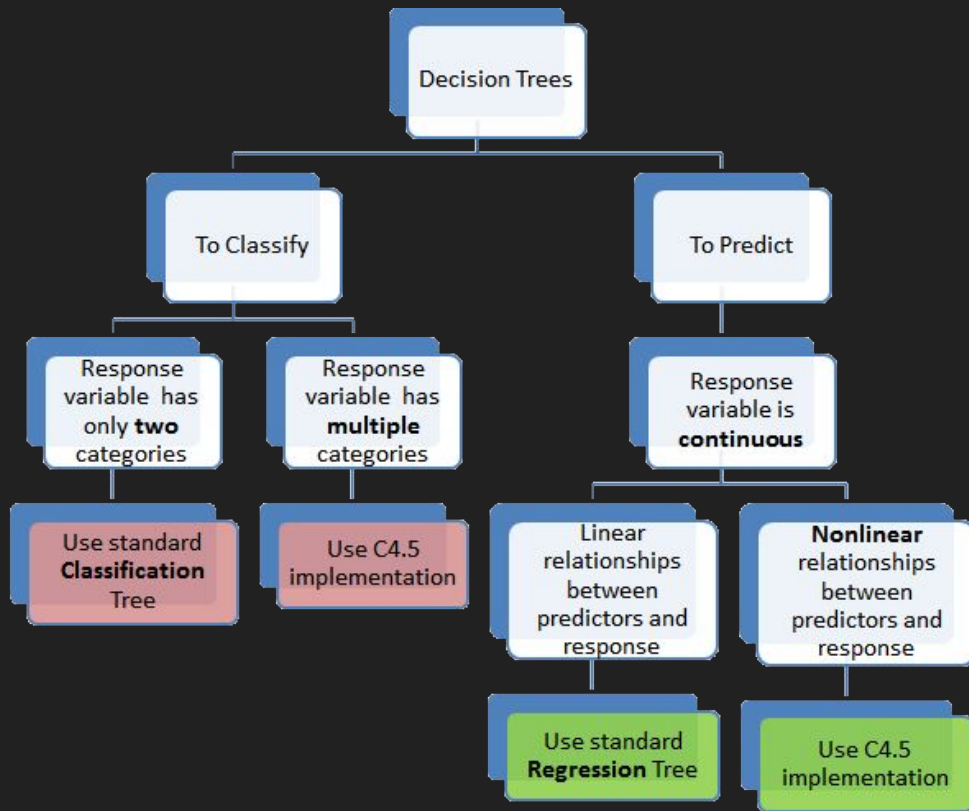
- Algumas vantagens

- Fácil entendimento

- Através da representação gráfica das árvores fica fácil o entendimento por parte de usuários sem conhecimento em analytics.

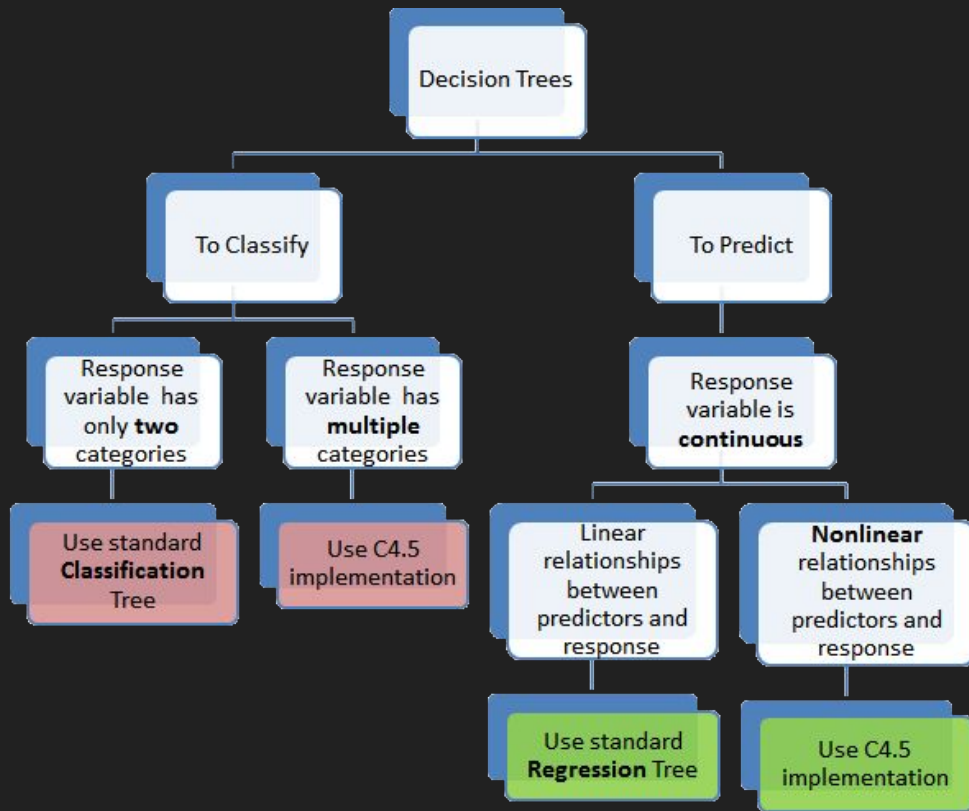
- Viabiliza a exploração dos dados

- Durante a exploração dos dados é possível utilizar modelos baseados em árvores de decisão com o objetivo de identificar features mais relevantes e seus relacionamentos.



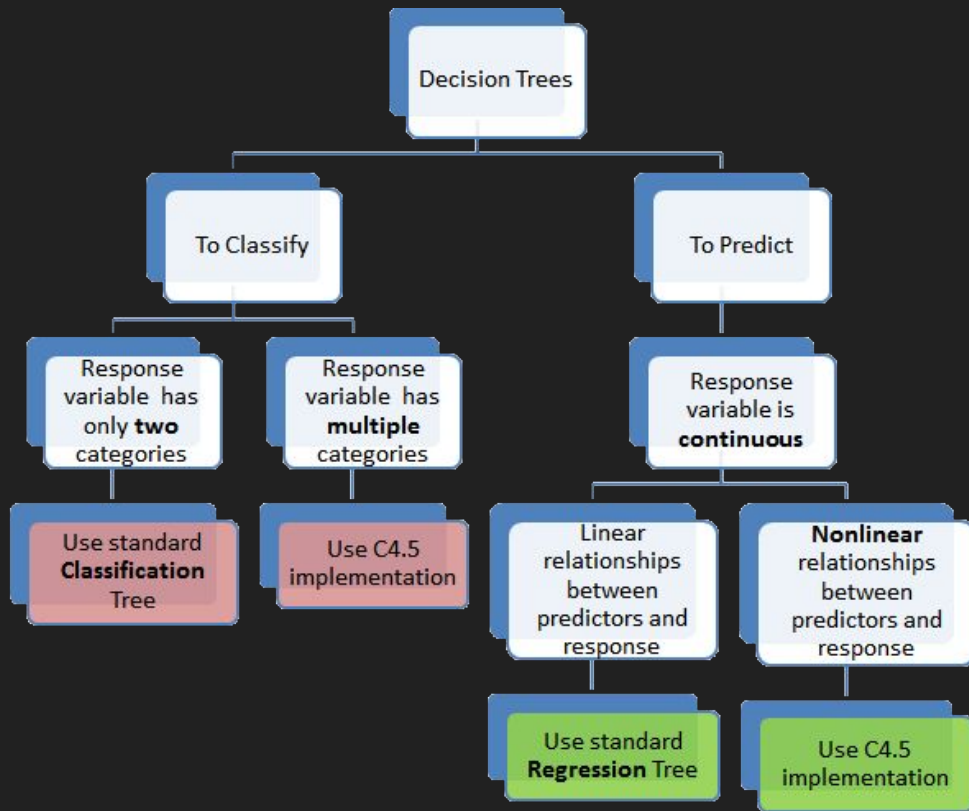
# Árvores de Decisão

- **Suporta tipos de dados distintos:**
  - Podem trabalhar com dados categóricos e numéricos.
- **Não exigem pré-processamento dos dados como modelos baseados em distância.**
- **Mais robusta contra outliers**
  - Devido a sua arquitetura baseada em divisão dos dados de acordo com o ganho de informação. Valores outliers não influenciam tanto para a construção do modelo.



# Árvores de Decisão

- Algumas desvantagens
- Mais propenso a Overfitting:
  - Devido a facilidade para se ajustar aos dados a árvore pode facilmente sofrer overfitting.



# Árvores de Decisão

- Como as árvores são construídas:
  - Quais os atributos que são usados como raiz e ramos da árvore?
  - Como escolher os atributos com mais seletividade?
  - Quando a árvore deve parar de crescer?
  - Como escolher os atributos com mais seletividade?





# Medidas para divisão dos dados

- Índice Gini
- Ganho de Informação
- Redução de Variância



# Como evitar o overfitting?

- Limitar o crescimento da árvore.
  - Uma árvore de decisão pode se ajustar tanto aos dados e ter um ramo para cada valor único do nível folha



# Como evitar o overfitting?

- Profundidade máxima da árvore e valor máximo de features a considerar para a divisão
- Valor mínimo de amostras em um atributo a considerar para a divisão
- Valor máximo de níveis folha



# Como evitar o overfitting?

- **Fazer a poda da árvore**
  - **Pré-poda**
    - Verificação de ganho de informação de um determinado atributo durante a etapa de construção da árvore.



# Como evitar o overfitting?

- **Fazer a poda da árvore**
  - **Pós-poda**
    - Após a construção da árvore, ramos são selecionados e após sofrerem a poda a acurácia é calculada para medir a eficácia do modelo.



# Modelos baseados em Árvores vs Modelos Lineares

- Relacionamento forte entre variáveis dependentes e independente; **modelo linear**
- Relacionamento fraco ou alta complexidade; **modelo baseado em árvore**
- Necessário para compreensão e geração de informação; **modelo baseado em árvore**.



Hands on!