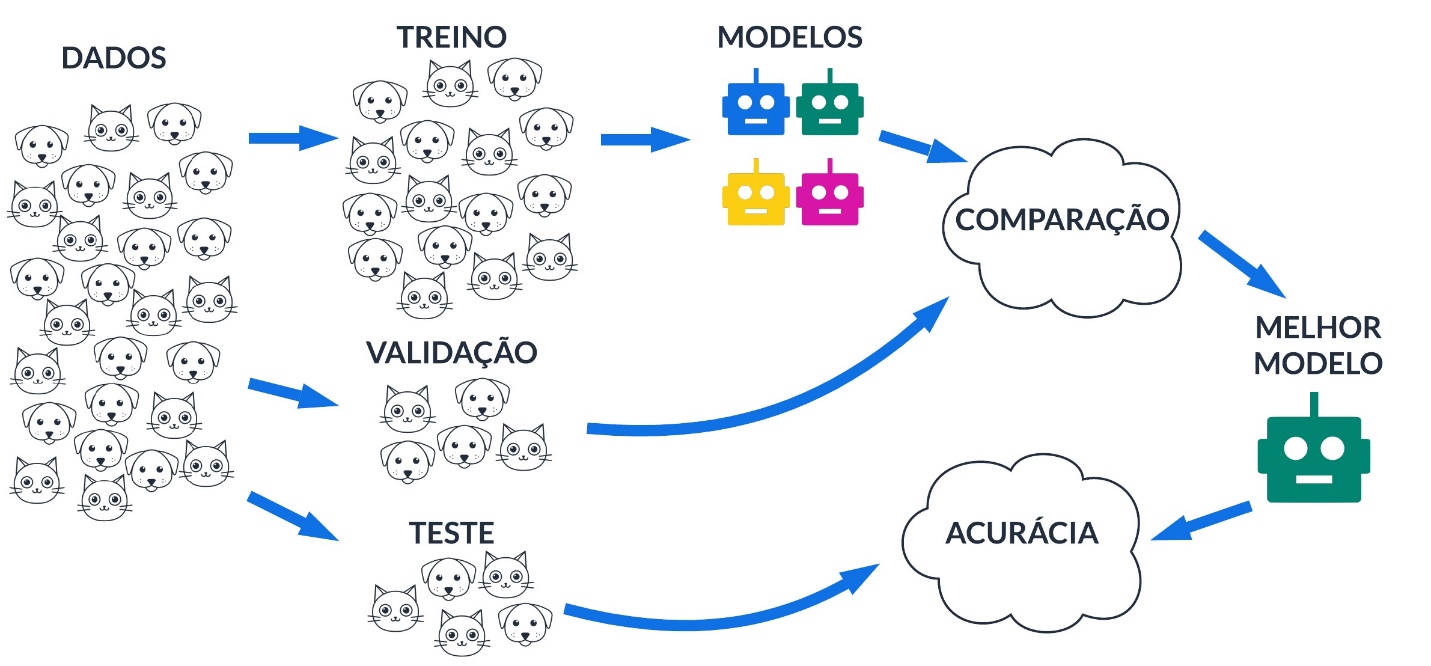
**1) Estratégia para criação de uma base de dados para treinamento, teste e validação**

**Proporções Sugeridas:**

* **Treinamento:** 70%
* **Teste:** 20%
* **Validação:** 10%

**Justificativa:** No desenvolvimento de modelos de aprendizado de máquina, é fundamental dividir os dados em conjuntos específicos para treinamento, teste e validação. Esta divisão é crucial para garantir que o modelo seja treinado de forma eficaz, avaliado corretamente e que seus parâmetros sejam ajustados para otimizar a performance.

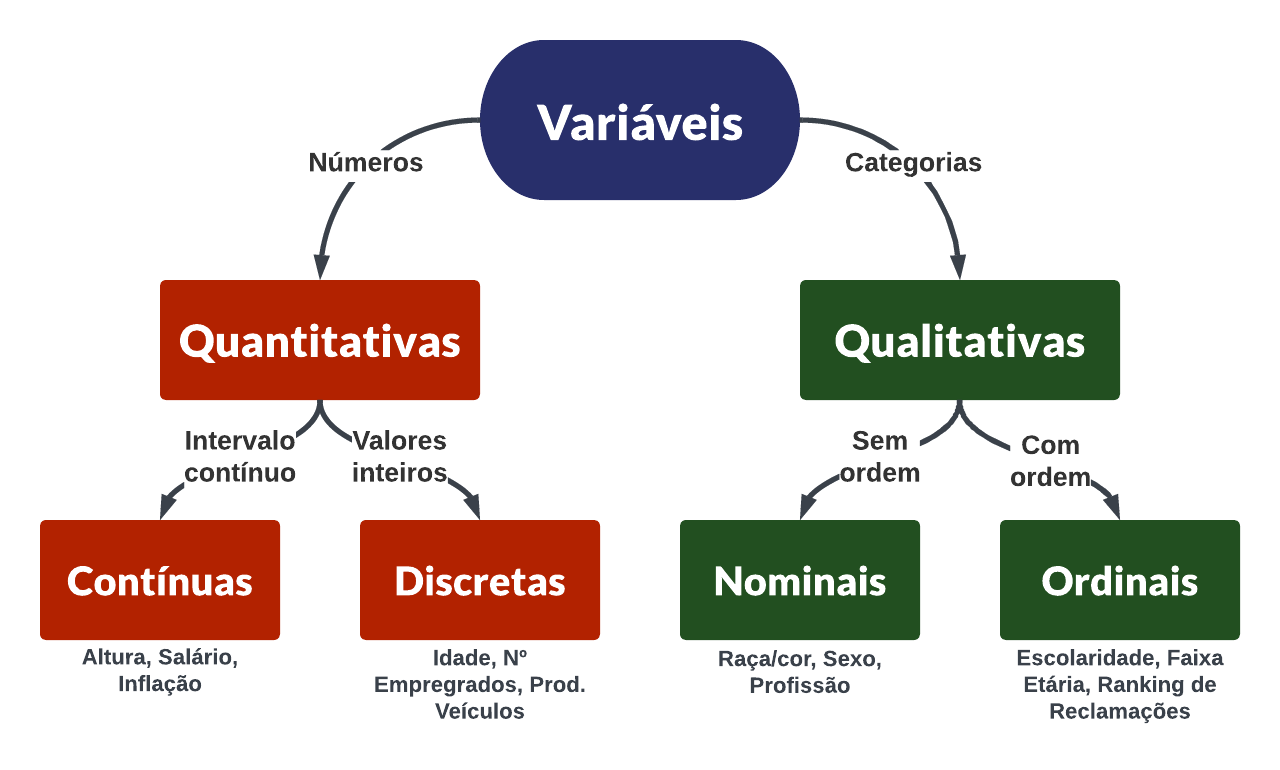
* **Treinamento (70%):** A maior parte dos dados deve ser destinada ao treinamento do modelo. Isso ocorre porque é necessário que o modelo tenha uma quantidade substancial de dados para identificar padrões, reconhecer correlações e aprender as regras subjacentes que governam o comportamento das variáveis.
* **Teste (20%):** O conjunto de teste é utilizado para avaliar o desempenho do modelo em dados que não foram previamente vistos. Esta avaliação é essencial para verificar a capacidade do modelo de generalizar seu aprendizado para novos dados, garantindo assim a sua aplicabilidade em cenários reais.
* **Validação (10%):** Durante o processo de treinamento, o conjunto de validação é empregado para ajustar os hiper parâmetros do modelo. Este ajuste é necessário para evitar o fenômeno conhecido como overfitting, onde o modelo se ajusta excessivamente aos dados de treinamento, perdendo assim sua capacidade de generalizar.



**2) Diferença entre variável categórica e numérica**

As variáveis em um conjunto de dados podem ser classificadas, de forma geral, em dois tipos principais: numéricas e categóricas.

* **Variáveis Numéricas:** São aquelas que assumem valores quantitativos e podem ser medidas de maneira precisa. Estas variáveis permitem a realização de operações matemáticas, como soma e média. Exemplos de variáveis numéricas presentes na base de dados incluem "valor", "quantidade", "valor\_total\_bruto", "valor\_comissao" e "lucro\_liquido". Essas variáveis são fundamentais para análises que envolvem quantificação, como cálculos de médias, dispersão e outras estatísticas descritivas.
* **Variáveis Categóricas:** São aquelas que representam categorias ou grupos. Estas variáveis podem ser nominais, onde não há uma ordem implícita entre os valores (como "estado" e "produto"), ou ordinais, onde os valores têm uma ordem natural. No contexto da base de dados analisada, variáveis como "cod\_pedido", "regiao\_pais", "produto", "data", "estado", "formapagto", "centro\_distribuicao", "responsavelpedido" e "categoriaprod" são categóricas. Essas variáveis são usadas para classificar os dados em grupos ou categorias, permitindo a realização de análises de frequência e proporção.



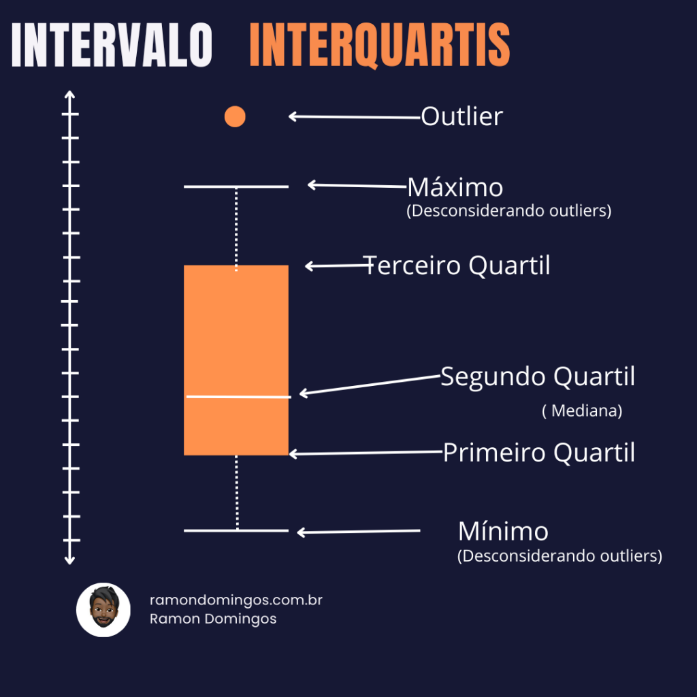
**3) Importância do código do pedido**

O **código do pedido (cod\_pedido)** desempenha um papel crucial na organização e análise dos dados. Como um identificador único para cada transação, ele permite rastrear individualmente cada pedido, assegurando que todas as informações associadas a uma transação específica sejam corretamente vinculadas. Em análises preditivas, a utilização do código do pedido é essencial para manter a integridade dos dados, evitando duplicidades ou associações incorretas que poderiam comprometer a acurácia dos resultados. Por exemplo, na previsão de padrões de compra ou na análise de comportamento de clientes, a identificação precisa de cada pedido permite uma análise mais detalhada e assertiva.

**4) Identificação de Outliers**

**Estratégia para Identificação de Outliers:** Outliers são valores que se desviam significativamente dos outros dados de um conjunto. A identificação desses valores é um passo importante, pois outliers podem indicar erros de entrada de dados ou condições excepcionais que precisam ser analisadas separadamente.

* **Análise de Dispersão (Boxplot):** Um método eficiente para a identificação de outliers é a utilização de boxplots, que visualizam a distribuição dos dados e destacam valores que estão fora dos limites esperados. Valores situados a mais de 1,5 vezes o intervalo interquartil (IQR) acima do terceiro quartil ou abaixo do primeiro quartil são considerados outliers e merecem atenção especial.
* **Análise Estatística:** Outra abordagem comum é o cálculo do z-score para cada ponto de dados, que indica quantos desvios padrão um valor está distante da média. Valores com z-scores muito altos ou baixos são identificados como outliers. As variáveis "valor", "quantidade", "valor\_total\_bruto", "valor\_comissao" e "lucro\_liquido" são especialmente suscetíveis a apresentar outliers, dada a natureza dos dados que podem incluir transações atípicas ou erros de registro.



Identificação de Outliers na Prática:

(Arquivo anexado)

Texto

Descrição gerada automaticamente

A estratégia utilizada para identificar os registros outliers nas colunas numéricas relevantes seguiu uma abordagem baseada no **Interquartile Range (IQR)**, que é uma técnica robusta e amplamente utilizada para detectar valores atípicos (outliers) em conjuntos de dados. Aqui está uma descrição detalhada da estratégia:

**1. Identificação das Colunas Relevantes**

* Primeiramente, selecionamos as colunas relevantes do DataFrame que contêm dados numéricos e que são de interesse para a análise de outliers. Neste caso, as colunas selecionadas foram: valor, quantidade, valor\_total\_bruto, valor\_comissao, e lucro\_liquido.

**2. Verificação e Conversão para Tipo Numérico**

* Antes de proceder com a identificação de outliers, verificamos se essas colunas estavam no tipo de dado correto (numérico). Caso as colunas estivessem em outro tipo (como object), realizamos a conversão para tipo numérico usando a função pd.to\_numeric(). Essa conversão é essencial para garantir que as operações matemáticas necessárias para identificar outliers possam ser realizadas corretamente.

**3. Cálculo dos Outliers Usando IQR**

* Utilizamos o **Interquartile Range (IQR)**, que é uma medida estatística da dispersão dos dados, para identificar outliers:
  + **Q1 (Primeiro Quartil)**: O valor que delimita os 25% menores valores do conjunto de dados.
  + **Q3 (Terceiro Quartil)**: O valor que delimita os 25% maiores valores do conjunto de dados.
  + **IQR**: A diferença entre Q3 e Q1, representando a amplitude interquartil.
  + **Limites Inferior e Superior**: Calculamos os limites inferior (Q1 - 1.5 \* IQR) e superior (Q3 + 1.5 \* IQR). Qualquer valor fora desse intervalo foi considerado um outlier.

**4. Contagem de Outliers**

* Para cada coluna relevante, calculamos quantos valores estavam abaixo do limite inferior ou acima do limite superior, e essa contagem foi armazenada em um dicionário.

**5. Visualização Gráfica**

* Para facilitar a interpretação dos resultados, criamos um gráfico de barras que exibe a quantidade de outliers em cada coluna relevante. Esse gráfico permite uma visualização clara da distribuição dos outliers e ajuda a identificar rapidamente quais colunas têm maior incidência de valores atípicos.

**Conclusão**

* A estratégia baseada no IQR é eficaz porque é menos influenciada por outliers do que outras medidas, como a média e o desvio padrão. Isso torna o IQR uma ferramenta confiável para detectar outliers em conjuntos de dados, especialmente quando há a possibilidade de valores extremos influenciar os resultados.

**5) Melhor estratégia para tratar os outliers**

Uma vez identificados os outliers, torna-se necessário decidir a melhor abordagem para tratá-los. A escolha da estratégia adequada depende do contexto dos dados e dos objetivos específicos da análise.

**Opções:**

* **Remoção:** Se os outliers forem identificados como erros ou valores irrelevantes para o objetivo da análise, a remoção desses pontos de dados pode ser a opção mais apropriada. A exclusão de outliers ajuda a evitar distorções nos resultados das análises.
* **Transformação:** Em casos em que os outliers são parte natural da distribuição, mas possuem um impacto desproporcional, a aplicação de transformações matemáticas, como logaritmos ou raízes quadradas, pode ser uma solução eficaz para atenuar a influência desses valores sem removê-los do conjunto de dados.
* **Truncamento:** Quando os outliers representam valores extremos, mas válidos, pode ser preferível limitar (truncar) esses valores ao percentil máximo permitido, como o 95º percentil, preservando assim a integridade dos dados sem permitir que valores excessivamente altos ou baixos distorçam as análises.

**Melhor Estratégia:** A escolha da estratégia mais adequada para tratar outliers deve ser baseada no entendimento profundo do contexto dos dados e dos objetivos da análise. Se os outliers são considerados erros, a **remoção** é recomendada. Se os outliers são representações legítimas de fenômenos raros, o **truncamento** pode ser a melhor abordagem. Em cenários onde a variação é esperada e natural, uma **transformação** pode ajudar a suavizar a distribuição dos dados, permitindo uma análise mais equilibrada e robusta.