### **Documentação do Projeto: Análise de Dados e Modelo Preditivo de E-commerce**

* **Objetivo do Projeto:** Realizar uma análise completa de dados de uma plataforma de e-commerce, desde a criação de dashboards de BI até o desenvolvimento de um modelo de machine learning para prever a satisfação do cliente, e comunicar os resultados de forma eficaz.

### **Parte 1: Documentação do Processo no Power BI**

Esta seção detalha a construção do dashboard de Business Intelligence para monitoramento e análise dos indicadores de negócio.

#### **1.1. Extração e Carregamento de Dados (ETL)**

* **Fonte de Dados:** 9 arquivos no formato CSV .
* **Ferramenta:** Power Query (integrado ao Power BI Desktop).
* **Processo:**
  1. Os dados foram carregados no Power Query utilizando a função "Obter Dados" > "Texto/CSV".
  2. Cada arquivo foi importado como uma tabela separada.

#### **1.2. Transformação e Limpeza de Dados (Camada Semântica)**

As seguintes transformações foram aplicadas no Power Query para garantir a qualidade e consistência dos dados:

* **Tipagem de Colunas:**
  + **Datas:** Colunas como order\_purchase\_timestamp, order\_approved\_at, etc. (da tabela pedidos) foram convertidas para o tipo *Data/Hora*.
  + **Valores Decimais:** Colunas price e freight\_value (tabela itens\_pedidos) foram convertidas para *Número Decimal*.
  + **Valores Numéricos:** Colunas de medidas de produtos como product\_weight\_g, product\_length\_cm, etc. (tabela produtos) foram convertidas para *Número Inteiro*.
* **Renomeação de Colunas:** Colunas com nomes técnicos foram renomeadas para nomes amigáveis para facilitar a leitura nos relatórios (ex: customer\_city -> Cidade do Cliente).
* **Tradução de Dados:** A coluna product\_category\_name (tabela categorias\_produtos) foi traduzida do inglês para o português utilizando a função "Substituir Valores" para melhorar a experiência do usuário final.
* **Tratamento de Erros:** Verificações de valores nulos ou em branco foram realizadas, embora nenhuma ação de remoção massiva tenha sido necessária para os relatórios iniciais.

#### **1.3. Modelagem de Dados**

Após a limpeza, os dados foram carregados no Power BI e a modelagem foi ajustada na "Exibição de Modelo" para criar um **Esquema Estrela**.

* **Tabelas Fato (centrais):** itens\_pedidos e pedidos. Contêm os eventos e valores numéricos do negócio.
* **Tabelas Dimensão (circundantes):** clientes, produtos, vendedores, categorias\_produtos, pedidos\_pagamentos, avaliacoes\_pedidos. Contêm os atributos que descrevem os fatos.
* **Relacionamentos Criados:**
  + clientes[customer\_id] (1) -> pedidos[customer\_id] (\*)
  + pedidos[order\_id] (1) -> itens\_pedidos[order\_id] (\*)
  + pedidos[order\_id] (1) -> pedidos\_pagamentos[order\_id] (\*)
  + pedidos[order\_id] (1) -> avaliacoes\_pedidos[order\_id] (\*)
  + produtos[product\_id] (1) -> itens\_pedidos[product\_id] (\*)
  + vendedores[seller\_id] (1) -> itens\_pedidos[seller\_id] (\*)
  + categorias\_produtos[product\_category\_name] (1) -> produtos[product\_category\_name] (\*)

Todos os relacionamentos foram configurados como **ativos** (linha sólida) e com direção de filtro único da dimensão para a fato.

#### **1.4. Cálculos com DAX (Data Analysis Expressions)**

Medidas explícitas foram criadas para garantir cálculos consistentes e reutilizáveis:

* Faturamento Total = SUM(itens\_pedidos[price])
* Qtd Pedidos = DISTINCTCOUNT(pedidos[order\_id])
* Qtd Clientes = DISTINCTCOUNT(clientes[customer\_unique\_id])
* Ticket Médio = DIVIDE([Faturamento Total], [Qtd Pedidos])

#### **1.5. Visualização de Dados (Dashboards)**

Foram criadas 2 páginas de relatório para responder a diferentes perguntas de negócio:

* **Página "Visão Geral":**
  + **KPIs Principais (Cartões):** Faturamento Total, Quantidade de Pedidos, Quantidade de Clientes, Ticket Médio.
  + **Pedidos por Tipo de Pagamento (Gráfico de Pizza):** Analisa a preferência de pagamento dos clientes.
  + **Faturamento por categoria de produto(Tabela):** Analisa o faturamento total e por categoria de produto
  + **Seleção de ano(Segmentação de Dados):** Permite a seleção do ano a ser analisado.

Nesta página, os instrumentos visuais estão integrados para que se tenha uma análise mais detalhada sobre número de pedidos por categoria, ticket médio por categoria e faturamento total, bem como faturamento nos anos de 2016,2017 e 2018.

* **Página "Análise Geográfica e por Categoria":**
  + **Vendas por Estado (Tabela):** Mostra o faturamento por estado.
  + **Desempenho por Categoria/Produto (Matriz):** Tabela detalhada e agrupada que permite aprofundar a análise de vendas por categoria e, em seguida, por produto, bem como saber a quantidade de pedidos e o faturamento total de frete do produto.

### **Parte 2: Documentação do Processo no Python**

Esta seção documenta a análise exploratória e a construção do modelo de machine learning.

#### **2.1. Ambiente e Ferramentas**

* **Linguagem:** Python
* **Bibliotecas Principais:**
  + pandas: Para manipulação e estruturação dos dados.
  + sqlalchemy e sqlite3: Para criação e interação com o banco de dados SQL.
  + matplotlib e seaborn: Para visualização de dados.
  + scikit-learn: Para construção e avaliação do modelo de Machine Learning.
* **Ambiente de Desenvolvimento:** Jupyter Notebook.

#### **2.2. Extração e Armazenamento**

1. **Carga de Dados:** Os 9 arquivos CSV foram carregados em um dicionário de DataFrames do pandas, facilitando o acesso a cada tabela.
2. **Armazenamento em Banco de Dados:**
   * Um banco de dados **SQLite** (ecommerce.db) foi criado.
   * Cada DataFrame foi salvo como uma tabela no banco de dados usando o método df.to\_sql(). Esta etapa centraliza os dados e permite a execução de consultas SQL.
3. **Consulta SQL:** Demonstrou-se a capacidade de consultar o banco de dados via SQL para extrair insights, como o faturamento por estado, carregando o resultado diretamente em um novo DataFrame.

#### **2.3. Análise Exploratória de Dados (EDA)**

Foram respondidas perguntas de negócio chave através da manipulação de dados com pandas e visualizações:

* **Evolução do Faturamento Mensal:** Um gráfico de linha foi gerado para mostrar a tendência de vendas, identificando picos e sazonalidades.
* **Distribuição das Avaliações:** Um gráfico de contagem (countplot) revelou que a maioria das avaliações são positivas (nota 5), indicando uma base de clientes majoritariamente satisfeita.
* **Categorias Mais Vendidas:** Um gráfico de barras horizontais destacou as 10 categorias de produtos com maior volume de vendas.

#### **2.4. Modelo Preditivo: Previsão de Satisfação do Cliente**

* **Objetivo do Modelo:** Prever se um pedido receberá uma avaliação alta (nota 5) ou não, tratando o problema como uma **classificação binária**.
* **Engenharia de Features:**
  + **Variável Alvo:** Criou-se a coluna target\_alta\_satisfacao (1 para nota 5, 0 para notas 1-4).
  + **Features Criadas:** Foram desenvolvidas variáveis com maior poder preditivo:
    - tempo\_entrega: Diferença em dias entre a compra e a entrega.
    - diferenca\_estimativa\_entrega: Diferença em dias entre a entrega real e a data estimada.
    - proporcao\_frete\_preco: Razão entre o custo do frete e o preço do item.
* **Seleção e Preparação dos Dados:**
  + Foram selecionadas features numéricas para uma primeira versão do modelo.
  + Dados com valores ausentes nas colunas selecionadas foram removidos (dropna).
  + Os dados foram divididos em 80% para treino e 20% para teste (train\_test\_split), com estratificação para manter a proporção da variável alvo em ambos os conjuntos.
* **Treinamento do Modelo:**
  + **Algoritmo Escolhido:** RandomForestClassifier foi selecionado por sua robustez e capacidade de fornecer a importância das features.
  + O modelo foi treinado com o conjunto de treino (model.fit(X\_train, y\_train)).
* **Avaliação de Performance:**
  + O modelo foi avaliado no conjunto de testes, que ele nunca havia visto.
  + **Acurácia:** O modelo atingiu uma acurácia geral consistente, na casa de 85%.
  + **Matriz de Confusão:** A análise da matriz mostrou a capacidade do modelo em distinguir corretamente entre clientes satisfeitos e não satisfeitos.
  + **Relatório de Classificação:** Fornece métricas de precisão e recall para cada classe.
* **Principais Insights do Modelo:**
  + A análise da **importância das features** (feature\_importances\_) revelou que diferenca\_estimativa\_entrega e tempo\_entrega são, de longe, os fatores mais determinantes para a satisfação do cliente. Este é o principal insight acionável do modelo.