Dataset utilizado: e-Commerce (Walmart) Sales Dataset (‘https://www.kaggle.com/datasets/devarajv88/walmart-sales-dataset/data’)

Este conjunto de dados se concentra nas transações dos clientes da renomada rede de lojas Walmart, capturando informações detalhadas sobre dados demográficos do usuário, categorias de produtos e valores de compra. Inclui identificadores exclusivos de usuários e produtos, sexo, faixas etárias, ocupação (disfarçada), categorias de cidade, duração da permanência na cidade atual, estado civil, categorias de produtos (disfarçadas) e valores das compras.

A análise desse conjunto de dados oferece informações valiosas sobre a base de clientes e o comportamento de compra na rede Walmart. Ele revela detalhes sobre dados demográficos dos clientes, preferências de produtos e padrões de gastos. Além de permitir entender vários aspectos das operações do Walmart, como estratégias de marketing, segmentação de clientes e demandas de produtos. Ele pode aprimorar a tomada de decisões estratégicas em áreas como gerenciamento de estoque, marketing direcionado e gerenciamento de relacionamento com o cliente.

O conjunto de dados contém as seguintes colunas:

User_ID: ID do usuário
Product_ID: ID do produto
Gender: Sexo do usuário
Age: Idade em caixas
Occupation: Ocupação(Mascarado)
City_Category: Categoria da cidade (A, B, C)
Stay_In_Current_City_Years: Número de anos de permanência na cidade atual
Marital_Status: Estado civil

Product_Category: Categoria de produto (mascarada)
Purchase: Valor da compra

Antes de aplicar qualquer modelo, é fundamental entender a estrutura dos dados, identificar valores nulos e realizar as transformações necessárias, ou seja, precisamos realizar a coleta e preparação dos dados.

Posteriormente, seguimos para análise exploratória de dados, onde conseguimos identificar que:

* Os maiores consumidores da rede são do sexo masculino (Gráfico 1)

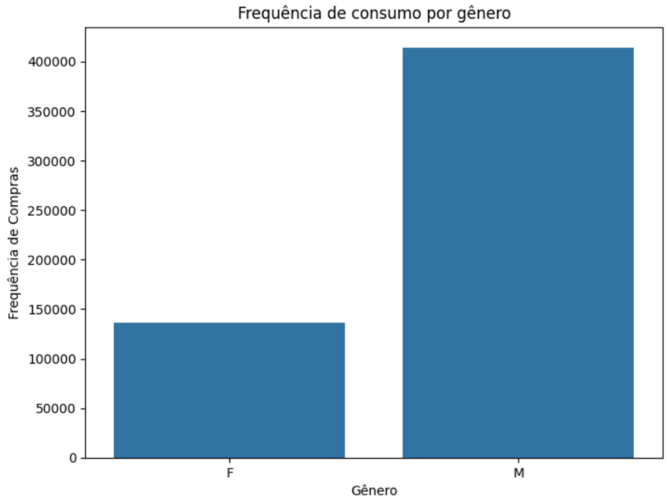

Gráfico 1

* Os consumidores com maior frequência de consumo estão localizados na faixa etária de 26 a 35 anos, seguidos pelas faixas de 36 a 45 anos e 18 a 25 anos, respectivamente. (Gráfico 2)

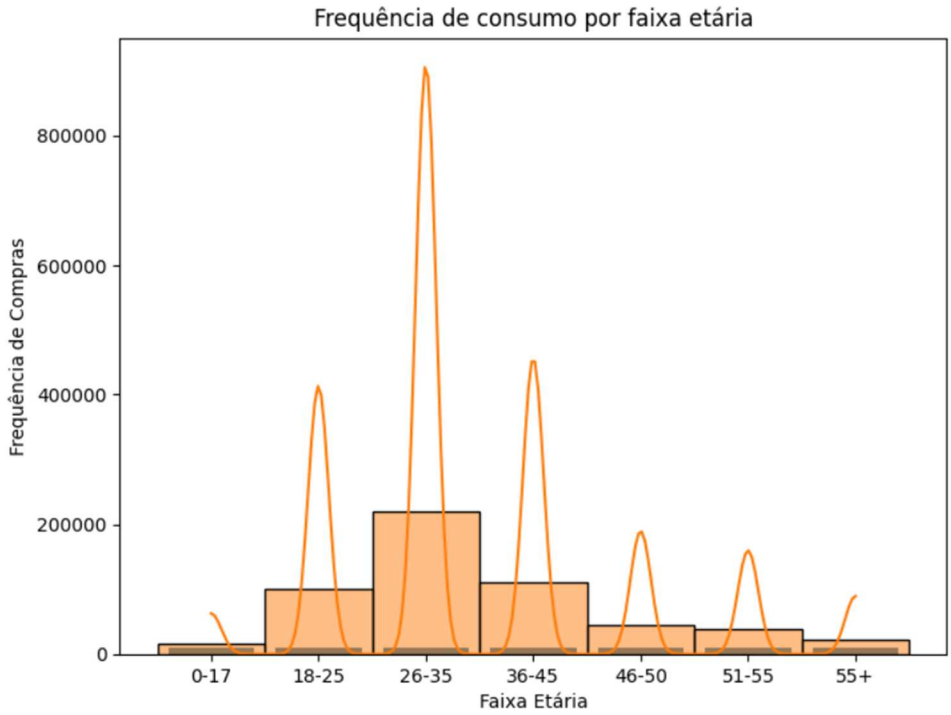

Gráfico 2

* O valor das compras é um pouco maior para o sexo masculino. (Gráfico 3)

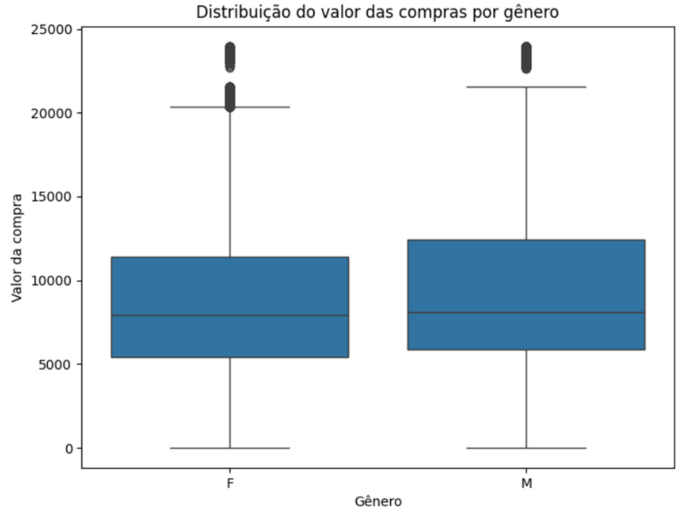

Gráfico 3

* Os consumidores da cidade C tem maior frequência de compras do que as demais cidades. (Gráfico 4)

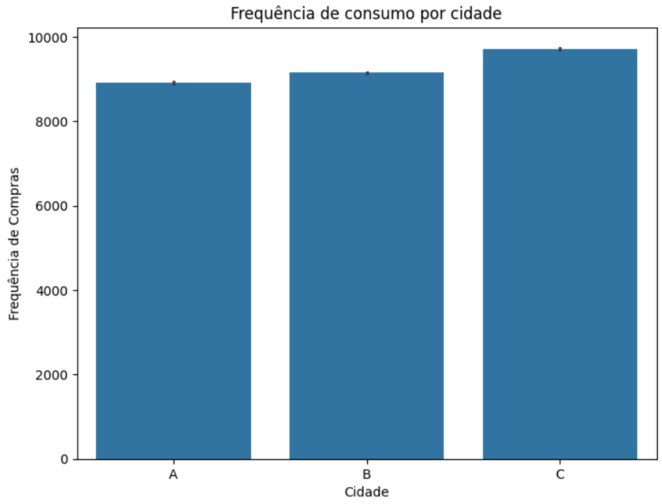

Gráfico 4

* Relação entre as categorias dos produtos e o valor das compras (Gráfico 5)

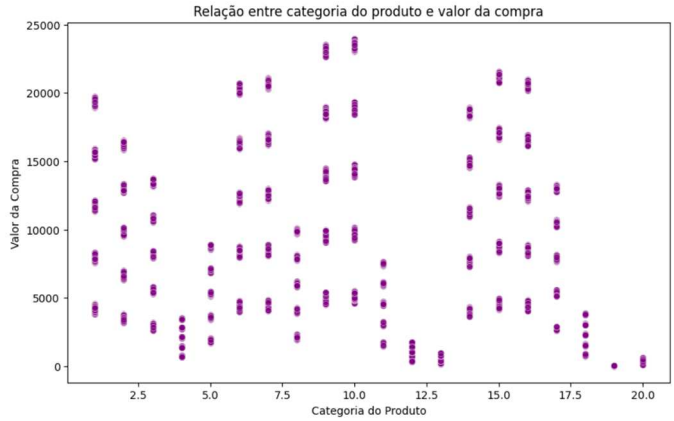

Gráfico 5

* Verificamos que não há correlação significativas entre as variáveis numéricas do conjunto de dados. (Gráfico 6)

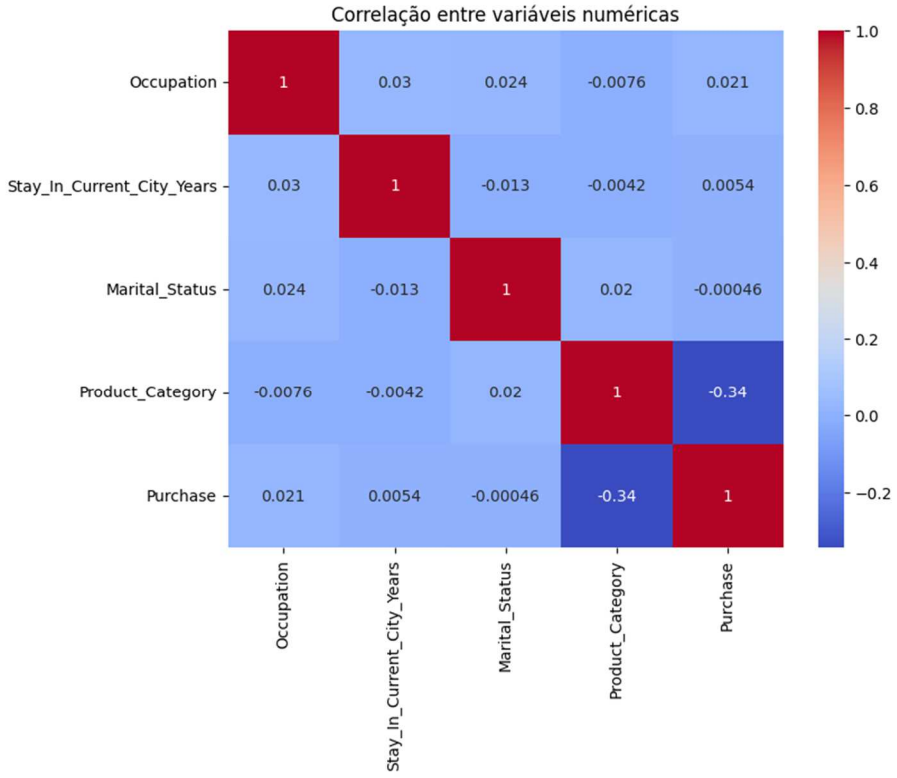

Gráfico 6

* Os consumidores solteiros tendem a comprar vezes mais que os casados. (Gráfico 7)

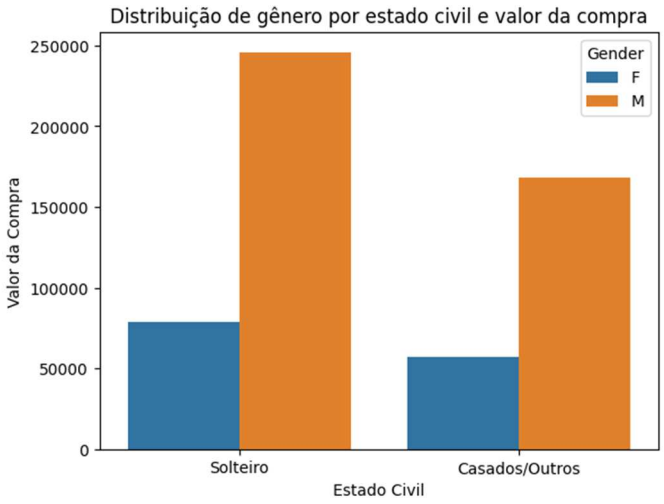

Gráfico 7

Foi realizada o pré-processamento para Machine Learning através da separação das variáveis numérica e categóricas, suas transformações e normalização. Em seguida iniciado os treinamentos e teste dos modelos de regressão, de modo a realizar uma análise exploratória e modelagem preditiva para prever o valor de compras (Purchase) com base nas informações dos clientes e produtos.

Sobre os resultados encontrados, verificamos que o R^2 Score através da regressão baseada na árvore de decisão indica que o modelo consegue explicar 64,7% da variância dos valores de compras (Purchase), o que indica um desempenho razoável, mas que pode ser aprimorado.

Neste sentido, foi incluído no código um comparativo de resultado entre 04 tipo de regressão, a regressão linear, a Árvore de decisão, a Radom Forest (Floresta Aleatória) e o XGBoost. E conforme

resultados abaixo detalhados, podemos concluir a melhor escolha é o modelo XGBoost, por trazer resultados mais preciso, mas frente a pouca variação, a regressão linear, inclusive por se tratar de um modelo mais simples e fácil de interpretar, seria uma boa opção de Machine Learning, para o estudo.

MODELO	**MAE**	**MSE**	**RMSE**	**R^2**
Linear Regression	**0.367991**	**0.237016**	**0.486843**	**0.633759**
Decision Tree	0.383556	0.252912	0.502904	0.609197
Random Forest	0.382936	0.250814	0.500813	0.612439
XGBoost	**0.364908**	**0.233593**	**0.483314**	**0.639049**