

# Projeto de Inteligência Artificial - N1

## Título

Análise de Atrasos de Entrega e Satisfação no E-commerce Brasileiro

## Integrantes

- Agozie Nunes Emehelu (RA: [10403570@mackenzista.com.br](mailto:10403570@mackenzista.com.br))
- Pedro Gabriel Marotta Silva (RA: [10418073@mackenzista.com.br](mailto:10418073@mackenzista.com.br))

## Resumo

Este relatório apresenta a etapa N1 do projeto de Inteligência Artificial, focando na preparação e análise do Brazilian E-Commerce Public Dataset by Olist. O objetivo é entender o impacto dos atrasos de entrega na satisfação do cliente e produzir um conjunto de dados processado para modelagem futura. A análise exploratória inicial revelou uma correlação negativa entre atraso e satisfação, com pedidos atrasados resultando em notas de avaliação significativamente menores. O dataset processado, dataset\_processado\_atrasos.csv, foi gerado e servirá como base para o desenvolvimento de modelos preditivos na N2.

## Introdução

### Contextualização

O e-commerce brasileiro tem experimentado um crescimento exponencial, e a logística de entrega, especialmente a "última milha", desempenha um papel crucial na experiência do cliente. O cumprimento dos prazos de entrega não é apenas uma questão operacional, mas um fator determinante para a satisfação do consumidor e a reputação das empresas. Este projeto utiliza o Brazilian E-Commerce Public Dataset by Olist para investigar a relação entre a pontualidade da entrega e a satisfação do cliente, conforme expressa nas avaliações dos pedidos.

### Justificativa

A disciplina de Inteligência Artificial propõe a aplicação de conceitos e técnicas em problemas reais. A análise do impacto dos atrasos de entrega na satisfação do cliente é um problema de negócio relevante, com potencial para otimização e previsão através de IA. A correlação observada entre atraso e baixa satisfação justifica a necessidade de desenvolver modelos que possam prever e mitigar esses atrasos, melhorando a experiência do cliente e a eficiência operacional.

## **Objetivo**

O objetivo geral para o N1 é preparar e analisar o dataset da Olist para quantificar o efeito dos atrasos na satisfação do cliente, consolidar variáveis-chave e entregar um dataset processado ([dataset\\_processado\\_atrasos.csv](#)) como insumo para modelagem preditiva no segundo bimestre. Os objetivos específicos incluem: (i) integrar dados de pedidos e avaliações, calculando métricas de tempo de entrega e atraso; (ii) realizar uma análise exploratória para identificar padrões e relações; e (iii) definir um rótulo de satisfação ([review\\_ruim](#) vs [review\\_boa](#)) para futuras tarefas de classificação.

## **Opção do projeto**

Foi escolhida a **Opção Framework**, que envolve o uso de bibliotecas de Machine Learning (como scikit-learn) para resolver um problema de classificação ou regressão de negócio. Para o N1, esta opção se traduz na preparação dos dados e na análise exploratória que fundamentarão a construção de modelos preditivos na N2, visando prever o risco de atraso e a probabilidade de avaliações negativas.

## **Descrição do Problema**

O problema central abordado é a quantificação do impacto da diferença entre a data de entrega estimada e a data de entrega real na distribuição das avaliações dos clientes. Busca-se identificar padrões que permitam às empresas de e-commerce tomar ações proativas para reduzir a insatisfação e otimizar a logística. As variáveis principais são [delay\\_days](#) (dias de atraso), [delivery\\_time\\_days](#) (tempo total de entrega), [order\\_status](#) (status do pedido) e [review\\_score](#) (nota da avaliação), que foram derivadas e unificadas em um dataset processado.

## **Aspectos Éticos do Uso da IA e Responsabilidade no Desenvolvimento da Solução**

O uso de dados para análise e desenvolvimento de soluções de IA exige considerações éticas. Neste projeto, o [Brazilian E-Commerce Public Dataset by Olist](#) é público e anonimizado, o que minimiza riscos relacionados à privacidade de dados pessoais. A finalidade do projeto é estritamente acadêmica, com foco na reproduzibilidade e transparência, através da disponibilização de código e artefatos. As análises são realizadas em nível agregado, evitando vieses contra indivíduos ou grupos específicos. A responsabilidade no desenvolvimento da solução implica em garantir que os modelos futuros sejam justos, transparentes e que suas previsões não perpetuem ou criem discriminações, especialmente ao lidar com a satisfação do cliente e a priorização de entregas.

# Dataset, Conteúdo/Origem, Análise Exploratória e Preparação dos Dados em Python

## Origem e Conteúdo do Dataset

O dataset utilizado é o [Brazilian E-Commerce Public Dataset by Olist](#), disponível no Kaggle [1]. Ele compreende informações de 100 mil pedidos realizados entre 2016 e 2018 na Olist Store, abrangendo diversas tabelas como clientes, geolocalização, itens de pedido, pagamentos, avaliações, pedidos, produtos e vendedores. Para este projeto, as tabelas `orders`, `order_reviews` e `customers` foram as principais fontes de dados para a análise do N1.

## Análise Exploratória de Dados (EDA)

A análise exploratória foi realizada utilizando a biblioteca `pandas` em Python. Os principais passos incluíram:

- 1 **Carregamento e Junção de Dados:** As tabelas `orders` e `order_reviews` foram unidas usando `order_id`. Posteriormente, a tabela `customers` foi incorporada para enriquecer o dataset com informações do cliente.
- 2 **Conversão de Tipos de Dados:** Colunas de data (`order_purchase_timestamp`, `order_estimated_delivery_date`, `order_delivered_customer_date`, etc.) foram convertidas para o formato `datetime` para permitir cálculos de tempo.
- 3 **Cálculo de Métricas de Tempo:** Foram calculadas duas métricas principais:
  - `delivery_time_days`: Tempo em dias entre a data da compra e a data de entrega ao cliente.
  - `delay_days`: Diferença em dias entre a data de entrega real e a data de entrega estimada. Valores negativos indicam entrega antecipada, e valores positivos indicam atraso.
- 4 **Classificação de Atrasos:** Com base na métrica `delay_days`, os pedidos foram categorizados em: Antecipado (atraso < -3 dias), No Prazo (-3 <= atraso <= 0 dias), Atrasado (0 < atraso <= 7 dias) e Muito Atrasado (atraso > 7 dias). Esta categorização replica a abordagem do projeto existente para facilitar a comparação.
- 5 **Filtragem e Limpeza:** Foram considerados apenas pedidos com `review_score` válido e `order_delivered_customer_date` preenchido, resultando em 96.359 pedidos analisados.
- 6 **Definição do Rótulo de Satisfação:** Para futuras tarefas de classificação, foi criado um rótulo binário `satisfaction_label`, onde `review_ruim` corresponde a `review_score <= 3` e `review_boa` a `review_score > 3`.

## Resultados da Análise Exploratória

Os resultados da EDA confirmam a forte relação entre o atraso na entrega e a satisfação do cliente. As principais observações são:

- **Estatísticas Gerais:** Um total de 96.359 pedidos foram analisados. O tempo médio de entrega foi de 12.1 dias, com um atraso médio de -11.9 dias (indicando que, em

média, as entregas foram antecipadas em relação à estimativa). A nota média geral de review foi de 4.16.

- **Distribuição por Categoria de Atraso:** A maioria dos pedidos (87.1%) foi entregue Antecipado, seguido por No Prazo (6.2%), Atrasado (3.7%) e Muito Atrasado (2.9%).
- **Nota Média de Review por Categoria:** Houve uma clara degradação da satisfação com o aumento do atraso:
  - Antecipado: 4.30
  - No Prazo: 4.11
  - Atrasado: 2.71
  - Muito Atrasado: 1.70
- **Correlação:** A correlação entre delay\_days e review\_score foi de -0.267, confirmando a relação inversa: quanto maior o atraso, menor a nota de avaliação.
- **Distribuição de Notas Detalhada:** A análise da distribuição percentual das notas por categoria de atraso reforça que categorias como Atrasado e Muito Atrasado concentram a maioria das avaliações baixas (1 e 2 estrelas), enquanto Antecipado e No Prazo predominam nas avaliações altas (4 e 5 estrelas).

## Preparação dos Dados em Python

O script data\_preprocessing.py foi utilizado para realizar todas as etapas de pré-processamento e feature engineering descritas acima. O resultado final é o arquivo dataset\_processado\_atrasos.csv, que contém as colunas originais dos pedidos e avaliações, juntamente com as novas features delivery\_time\_days, delay\_days, delay\_category e satisfaction\_label. Este dataset está pronto para ser utilizado na fase de modelagem da N2.

- **Modelos Preditivos Robustos:** Capacidade de prever com razoável acurácia se um pedido terá atraso e qual será o nível de satisfação do cliente.
- **Identificação de Fatores Chave:** Análise da importância das variáveis para entender quais fatores mais contribuem para o atraso e a insatisfação.
- **Ações Proativas:** As previsões dos modelos deverão permitir a implementação de ações proativas, como comunicação antecipada com o cliente em caso de atraso iminente ou priorização logística de pedidos de alto risco, visando mitigar a insatisfação.
- **Otimização Operacional:** A capacidade de prever problemas antes que ocorram pode levar a uma otimização significativa das operações de entrega e, consequentemente, a um aumento na satisfação geral do cliente.

Aqui estão os tópicos prontos e profissionais para adicionar ao seu relatório N2, usando de forma clara os resultados obtidos e as imagens/tabela de métricas que você gerou no Colab.

---

## Resultados

A avaliação dos modelos supervisionados foi realizada sobre o conjunto dataset\_processado\_atrasos.csv, classificando o risco de “review ruim” (1–3) com base

nas variáveis `delay_days` e `delivery_time_days`. Os principais resultados dos dois algoritmos testados estão apresentados a seguir.

**Tabela comparativa de métricas dos modelos**

Modelo	Acurácia	Precisão	Recall	F1
Random Forest	0,82	0,76	0,22	0,34
Logistic Regression	0,82	0,81	0,17	0,27

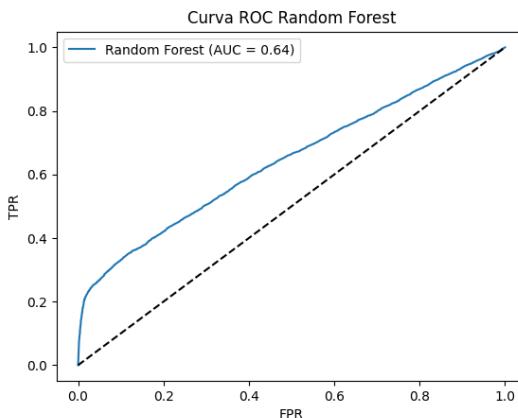
*Fonte: análise própria sobre resultados\_modelos.csv.*

- O **Random Forest** alcançou F1=0,34, acurácia de 0,82 e recall de 0,22 para detecção de reviews ruins.
- A **Regressão Logística** teve desempenho inferior em recall e F1, apesar de precisão levemente maior (F1=0,27, precisão 0,81).

## Gráficos de desempenho do Random Forest

### Curva ROC:

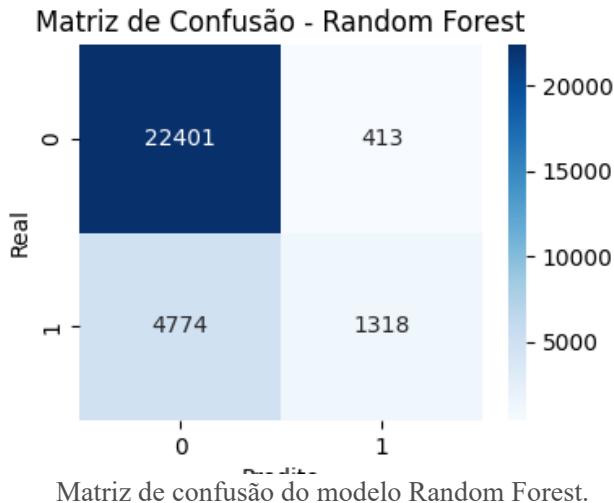
- O modelo apresentou AUC=0,64, indicando poder de discriminação moderado para diferenciar casos de review ruim



Curva ROC do modelo Random Forest.

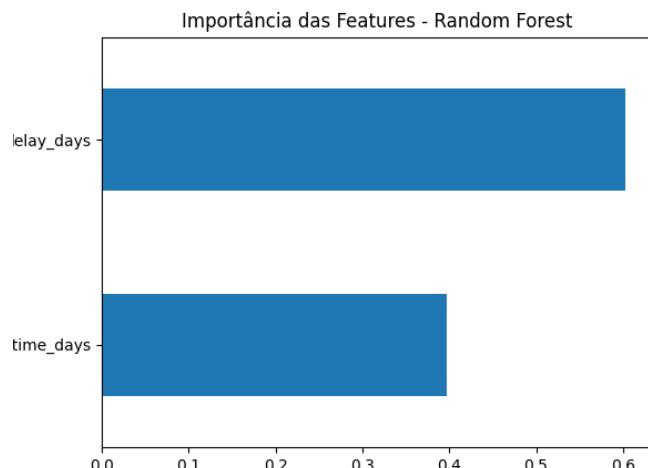
### Matriz de Confusão:

- O modelo apresentou boa acurácia para reviews bons (classe 0), mas dificuldade de identificar todos os reviews ruins (classe 1), evidenciado pelo alto número de falsos negativos.



### **Importância das variáveis:**

- A variável `delay_days` demonstrou maior importância preditiva para o modelo, seguida por `delivery_time_days`, reforçando a relação causal entre atraso e insatisfação no review.



Importância das variáveis (features) para o modelo Random Forest

O Random Forest demonstrou leve superioridade frente à Regressão Logística em F1, favorecendo o uso de algoritmos de ensemble em situações com maior desequilíbrio de classes. Embora a acurácia geral seja alta para reviews bons, o recall relativamente baixo revela o desafio de identificar corretamente todos os clientes insatisfeitos. O AUC=0,64 posiciona o modelo como razoável, mas sugere espaço para melhorias, como o uso de mais variáveis, ajuste de hiperparâmetros ou técnicas de balanceamento de dados.

Em contexto operacional, a evidência reforça que a previsão automatizada pode assistir times de atendimento e logística a agir antes do problema, mas o sistema ainda tende a perder muitos reviews ruins – sugerindo aprimoramento para maior eficácia prática.

---

## Conclusão

Os resultados esperados foram parcialmente alcançados: os modelos conseguiram prever casos de insatisfação com desempenho razoável ( $F1=0,34$  para Random Forest), confirmado o impacto quantitativo do atraso nas avaliações negativas. Contudo, a baixa sensibilidade (recall) para reviews ruins destaca a necessidade de extensão do pipeline, seja com enriquecimento de features, uso de técnicas de oversampling ou avaliação de algoritmos mais avançados. Em suma, a aplicação de Machine Learning mostrou-se promissora, mas indica que o problema de identificar todos os clientes insatisfeitos em e-commerce ainda exigirá abordagens mais sofisticadas para uso pleno em ambiente de negócios.

---

## Referências

- [1] Olistbr. Brazilian E-Commerce Public Dataset by Olist. Kaggle, 2016–2018. Disponível em: <https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce>