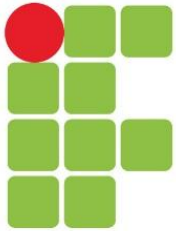


INSTITUTO FEDERAL
SÃO PAULO

# Aplicação de técnicas de ML e PLN na análise de sentimentos de comentários sobre produtos

Aluno: Daniel Vargas Shimamoto
CP3013391

INSTITUTO FEDERAL
SÃO PAULO

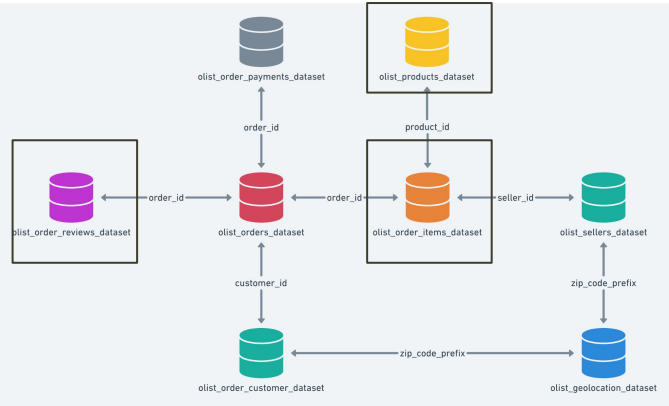
# ÍNDICE

INSTITUTO FEDERAL
SÃO PAULO

# INFORMAÇÕES GERAIS

**Dataset:**

- Olist Store Version 2 - 100 mil pedidos de vários marketplaces no Brasil entre 2016 e 2018
- Compra → Envio do produto → Recebimento/Vencimento da data de entrega → Pesquisa de Satisfação (Nota de 1 a 5 + Comentário)
- 8 datasets

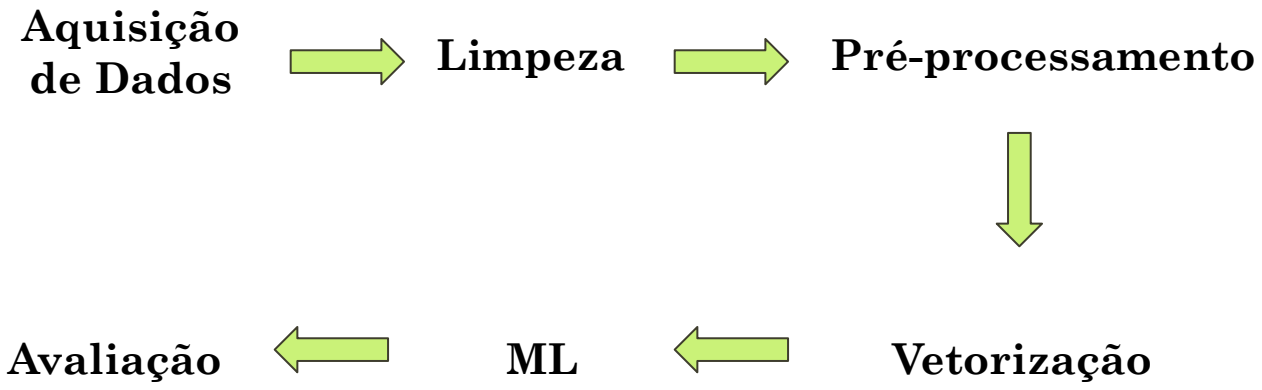



- Olist Order Reviews Dataset
  - Informações sobre satisfação
- Olist Order Items Dataset
  - Itens das ordens
- Olist Products Dataset
  - Informações dos produtos

INSTITUTO FEDERAL
SÃO PAULO

# OBJETIVO

- Com base nos comentários recebidos na pesquisa de satisfação, prever o sentimento (Negativo ou Positivo) utilizando modelos supervisionados.



INSTITUTO FEDERAL
SÃO PAULO

# AQUISIÇÃO DOS DADOS

- Datasets
  - Olist Order Reviews Dataset
  - Olist Order Items Dataset
  - Olist Products Dataset

- Colunas Finais
  - order_id, product_category_name, review_score, review_comment_message
  - Review_score → 1-3 Negativo; 4-5 Positivo (KIM et al., 2020)

INSTITUTO FEDERAL
SÃO PAULO

# LIMPEZA

- Remoção de avaliações sem comentários
- Remoção de comentários da mesma ordem do mesmo tipo de produto (mais de um produto por ordem)
- Quantidade final de dados 41.336
  - Negativos: 14.748 ~ 36%
  - Positivos: 26.588 ~ 64%

INSTITUTO FEDERAL
SÃO PAULO

# PRÉ PROCESSAMENTO

- Data: '([0-2][0-9]|(3)[0-1])(\/|\.)(((0)[0-9])|((1)[0-2]))(\/|\.)\d{2,4}'
  - Comentários: 258
  - Substituição: valorData

- Quebra de linha: '[\n\r]'
  - Comentários: 3890
  - Substituição:

- Dinheiro: '[R]{0,1}\$[ ]{0,}\d+(,|\.)\d+'
  - Comentários: 28
  - Substituição: valorDinheiro

- Numeração: '[0-9]+'
  - Comentários: 4196
  - Substituição: valorNumero

INSTITUTO FEDERAL
SÃO PAULO

# PRÉ PROCESSAMENTO

- **Pré processamento**
  - Normalização em minúsculo: Evitar que palavras iguais sejam tratadas de formas diferentes
  - Tratamento da palavra “Não”: A palavra "não" é muito importante para análise de comentários e pode aparecer de diversas formas em um texto. Por ser uma stopword ela será tratada para não perder essa informação
  - Remoção de caracteres especiais: Redução de ruídos
  - Remoção de acentos: Reduzir variação de erros gramaticais
  - Remoção de Stopwords: Palavras que adicionam ruído sem agregar informações

INSTITUTO FEDERAL
SÃO PAULO

# PRÉ PROCESSAMENTO

- **Stemming**
  - Reduzir a palavra ao seu radical
  - Representação uniforme (sem flexões)
  - Pode reduzir a palavra a uma classe gramatical incorreta

- nltk → RSLPStemmer()

| **Comentário Original** | | **Comentário pré processado** |
|---|---|---|
| **P**arab**éns** loj**as** lannist**er,** ador**ei** compr**ar pela I**nternet**,** segur**o e** pr**ático!** Parab**éns a** tod**os** fel**iz** **P**ásco**a** |  | parab loj lannist ador compr internet segur pra parab tod feliz pasco |

INSTITUTO FEDERAL
SÃO PAULO

# VETORIZAÇÃO

- **TFIDF**
  - Ponderação dos termos mais comuns usados em um documento em relação aos demais
  - sklearn.feature_extraction.text → TfidfVectorizer

$$TF = \frac{\text{Frequencia da palavra no documento}}{\text{Total de palavras no documento}}$$

$$IDF = \log\left(\frac{\text{Número total de documentos}}{\text{Número total de documentos com a palavra}}\right)$$

$$TF, IDF = TF * IDF$$

**Params**

- max_features: 300
- min_df: 7
- max_df: 0.8



OBS: Os dados foram separados em treino e teste antes de aplicar o TF-IDF

INSTITUTO FEDERAL
SÃO PAULO

# MACHINE LEARNING

- **Gaussian Naive Bayes**
  - Sem fine tuning
- **Regressão Logística**
  - Fine Tuning: penalty, C, class_weight
- **SGD Classifier**
  - Fine Tuning: penalty, alpha



*GridSearchCV com 5 folds

INSTITUTO FEDERAL
SÃO PAULO

# RESULTADOS

| Modelo | Acuracia Treino | Acuracia Teste | Tempo de treinamento (s) |
|---|---|---|---|
| Gaussian Naive Bayes | 0.834 | 0.833 | 2.33 |
| Logistic Regression | 0.887 | 0.888 | 35.15 |
| SGD Classifier | 0.888 | 0.887 | 106.04 |

- Os modelos possuem uma melhor acurácia nos dados com comentários positivos
  - Desbalanceamento das classes
- Regressão Logística e SGD tiveram resultados semelhantes
  - Regressão Logística possui um tempo de treinamento bem menor
- Próximos passos
  - Outras técnicas de vetorização (Bag of N-Grams)
  - Outros modelos de classificação (Ensemble Learning, redes neurais)

INSTITUTO FEDERAL
SÃO PAULO

# Referências


- KAGGLE. Brazilian E-Commerce Public Dataset by Olist. Version 2. Created by Francisco Magioli. Data Update: 2021/10/01. Disponível em: https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce. Acesso em: 24 jun. 2022.

- KIM, Y.; LEVY, J.; LIU, Y. Speech sentiment and customer satisfaction estimation in socialbot conversations. arXiv preprint arXiv:2008.12376. Disponível em: https://arxiv.org/pdf/2008.12376.pdf. Acesso em: 24 jun. 2022. 2020.

- GITHUB. OlistDataset. Disponível em: https://github.com/Shimad01/OlistDataset. Acesso em: 26 jun. 2022., 2022

INSTITUTO FEDERAL
SÃO PAULO

# OBRIGADO!