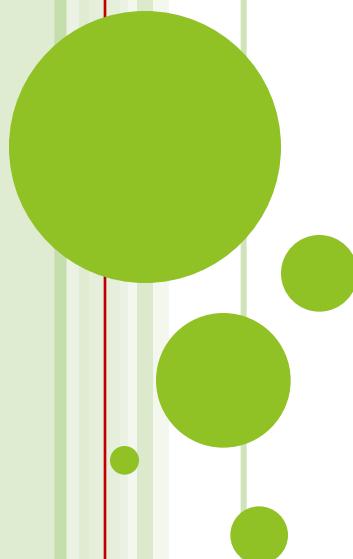


**INSTITUTO FEDERAL
SÃO PAULO**

MODELAGEM E AVALIAÇÃO DO DATASET “MARKETING CAMPAIGN”



Juliana Murakami

Tecnologia em Análise e Desenvolvimento de Sistemas

Disciplina: Introdução ao Aprendizado de Máquina

Docente: Everton Meyer da Silva

PROBLEMA PREDITIVO

- **O que será previsto?**
- Previsão binária da variável **Response** (0 = não aceitou, 1 = aceitou a campanha).
- **Por que isso é relevante?**
- Ajuda a empresa a direcionar campanhas para clientes com maior propensão de aceitar a oferta.
- Otimiza investimento em marketing.
- Reduz desperdício e aumenta taxa de conversão.



TÉCNICAS DE APRENDIZADO DE MÁQUINA

Modelos selecionados:

- ❑ **Regressão Logística**: modelo linear, probabilidade estimada pela função sigmoide.
- ❑ **Random Forest**: conjunto de múltiplas árvores de decisão treinadas de forma aleatória.
- ❑ **KNN**: classificação por proximidade com vizinhos mais próximos.

JUSTIFICATIVA DA ESCOLHA DOS MODELO

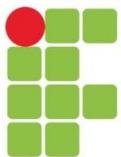
- Permitem comparação entre abordagens lineares e não lineares.
- Bons para dados tabulares.
- Possibilitam observar trade-offs entre acurácia, recall, capacidade de generalização e estabilidade.



PRÉ-PROCESSAMENTO (TRANSFORMAÇÕES)

- Remoção de registros com **Income** nulo.
- Engenharia de atributos:
 - **Age**,
 - **Total_Spent**,
 - **Children**.
- Remoção das colunas ID e outras não utilizadas.
- Separação entre **numéricas** e **categóricas**.
- StandardScaler** para numéricas.
- One-Hot Encoding** para categóricas.





DIVISÃO TREINO/TESTE

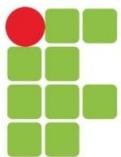
Configuração dos experimentos:

- ❑ Treino: **70%**
- ❑ Teste: **30%**
- ❑ **stratify = y** para manter proporções da classe.

Balanceamento:

- ❑ Uso de **class_weight='balanced'** para compensar desbalanceamento da classe 1.





PIPELINE COMPLETO

Pipeline integrado:

- Pré-processamento com ColumnTransformer
- Normalização
- Encoding
- Treinamento
- Predição
- Avaliação com métricas apropriadas





MÉTRICAS DE AVALIAÇÃO

- **Acurácia**
- **Precisão (1)**
- **Recall (1)**
- **F1-score**
- **ROC-AUC**

	Accuracy	Precision (1)	Recall (1)	F1 (1)	ROC-AUC
LogReg	0.829	0.461	0.83	0.593	0.907
RandomForest	0.878	0.806	0.25	0.382	0.874
KNN	0.872	0.636	0.35	0.452	0.792





VALIDAÇÃO CRUZADA

Configuração:

- Stratified K-Fold (5 folds)**
- Métrica: **ROC-AUC**

	ROC-AUC (mean)	ROC-AUC (std)
LogReg	0.879	0.020
RandomForest	0.873	0.018
KNN	0.773	0.035

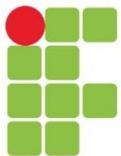


RESULTADOS OBTIDOS

Pipeline integrado:

- Pré-processamento com ColumnTransformer
- Normalização
- Encoding
- Treinamento
- Predição
- Avaliação com métricas apropriadas





CONCLUSÕES

- O modelo mais adequado é a **Regressão Logística**
- Apresenta o melhor equilíbrio entre recall e ROC-AUC
- Atende ao objetivo de identificar clientes com maior chance de aceitar a campanha



FONTES DE INFORMAÇÃO:

Dataset Kaggle:

<https://www.kaggle.com/datasets/rodsaldanha/marketing-campaign?resource=download>

Repositório Github: <https://github.com/JuMurak/machine-learning-marketing-campaign>

