

# Previsão de Churn Bancário Utilizando Modelos de Machine Learning Clássico

## 1. INTRODUÇÃO, PROBLEMA E JUSTIFICATIVA

### 1.1. CONTEXTUALIZAÇÃO DO PROBLEMA REAL

- 1.1.1. **Domínio de Aplicação:** O projeto aborda o setor bancário, com foco no comportamento do cliente e nas estratégias de retenção adotadas pelas instituições financeiras. Esse domínio envolve a análise de dados demográficos, financeiros e comportamentais para entender por que clientes deixam a instituição e quais fatores influenciam a decisão de cancelamento
- 1.1.2. **Situação Atual:** Atualmente, o processo de identificação de clientes propensos ao cancelamento é realizado de forma limitada. A maioria dos bancos utiliza regras fixas, como faixas de pontuação ou histórico de movimentação, além de consultas manuais feitas por equipes de relacionamento. Esse método apresenta baixa precisão, pois não considera relações mais profundas entre as variáveis nem identifica padrões complexos ao longo do tempo. Como consequência, muitas ações de retenção são tomadas **tarde demais** ou direcionadas para **clientes que não estão em risco**, gerando desperdício de recursos, perda de receita e impacto negativo na satisfação geral.
- 1.1.3. **Base Teórica:** Diversas pesquisas confirmam que o churn bancário é um problema crítico e recorrente. Verbeke et al. (2012) mostram que prever cancelamento em serviços financeiros aumenta significativamente a eficácia das ações de retenção. Além disso, Dahiya (2020) destaca que estratégias tradicionais são insuficientes para lidar com o volume e a complexidade dos dados bancários modernos, reforçando a necessidade de abordagens baseadas em Machine Learning. Esses estudos comprovam que o problema é relevante, atual e amplamente discutido na literatura acadêmica.

### 1.2. DEFINIÇÃO DO PROBLEMA DE MACHINE LEARNING:

- 1.2.1. **Definição Técnica:** Trata-se de um problema de classificação binária, em que o objetivo é treinar um modelo para prever a probabilidade de um cliente cancelar seu contrato de acordo com as informações disponíveis.

- 1.2.2. **Variável Alvo (Target):** A variável dependente a ser prevista é “Cancelou”, em que 1 representa o cancelamento do cliente e 0 representa o não cancelamento.
- 1.2.3. **Dataset Escolhido:** O conjunto de dados escolhido foi “Bank Customers Churn”, disponível no repositório Kaggle. (<https://www.kaggle.com/datasets/santoshd3/bank-customers?select=Churn+Modeling.csv>)

### 1.3. JUSTIFICATIVA E RELEVÂNCIA

- 1.3.1. **Impacto:** O modelo poderá auxiliar a empresa a perder menos clientes, o que se traduz em uma perda menor de receita. Com esta solução, a empresa saberá com mais certeza quais clientes devem receber mais atenção.
- 1.2.2. **Relevância:** O uso de Machine Learning é importante para resolver este problema porque permite identificar padrões complexos e não lineares no comportamento dos clientes, algo que métodos tradicionais como regras fixas, análises manuais ou modelos estatísticos simples não conseguem capturar com a mesma precisão. Além disso, modelos de ML conseguem aprender continuamente com novos dados, tornando-as ferramentas escaláveis e adaptáveis ao ambiente competitivo do setor bancário. Com isso, a instituição passa a prever cancelamentos de forma antecipada, reduzindo perdas financeiras e permitindo ações de retenção mais assertivas e personalizadas.

### 1.3. OBJETIVOS

#### 1.3.1. OBJETIVO GERAL

- 1.2.1.1. Desenvolver e comparar modelos de Machine Learning Clássico para prever o risco de cancelamento de clientes de um banco, utilizando um conjunto de dados públicos.

#### 1.3.2. OBJETIVOS ESPECÍFICOS

##### 1.3.2.1. Lista de passos técnicos e sequenciais:

- a) Obter, inspecionar e estruturar o dataset real e pertinente ao problema.
- b) Realizar a Análise Exploratória de Dados (EDA) para diagnosticar a qualidade dos dados.
- c) Implementar o pipeline de Pré-processamento e Engenharia de Atributos (Feature Engineering).

- d) Treinar e otimizar dois (2) modelos de Machine Learning Clássico para comparação, um modelo de Regressão Logística, e o outro sendo de Árvore de Decisão.
- e) Comparar o desempenho dos modelos utilizando métricas adequadas e a base de testes.
- f) Analisar a contribuição dos atributos (Feature Importance), se for aplicável ao modelo escolhido.

#### **1.4. FUNDAMENTAÇÃO TEÓRICA**

##### **1.4.2.1. REVISÃO BIBLIOGRÁFICA DO DOMÍNIO**

###### **a) CONCEITOS CHAVE DO DOMÍNIO:**

Cancelamento (Churn): É quando o cliente deixa de utilizar os serviços do banco, encerrando sua conta ou migrando para outra instituição. É um indicador importante porque representa perda direta de receita e aumento do custo com aquisição de novos clientes.

Pontuação de Crédito (Credit Score): É uma nota que indica o risco financeiro do cliente, baseado em seu histórico de pagamentos e comportamento de crédito. Quanto maior a pontuação, menor o risco de inadimplência.

Balanço (Balance): É o valor total de dinheiro que o cliente mantém em sua conta. Esse indicador reflete o nível de relacionamento e pode influenciar na chance de cancelamento.

**b) ESTATÍSTICAS E FATOS:** O cancelamento de clientes no setor bancário é um problema global significativo. Pesquisas mostram que: A taxa média internacional de churn bancário varia entre 15% e 25% ao ano (Dahiya, 2020). O custo para adquirir um novo cliente pode ser até 5 a 7 vezes maior do que o custo para reter um cliente já existente (Bain & Company, 2020). A perda de clientes pode gerar redução direta no faturamento anual, uma vez que bancos dependem de taxas, tarifas e uso contínuo dos serviços. No Brasil, o aumento de bancos digitais elevou a taxa de migração de clientes para instituições com tarifas menores e serviços mais ágeis,

intensificando a competição e a necessidade de estratégias de retenção. Estudos acadêmicos apontam que um aumento de apenas 5% na retenção pode gerar aumento de até 25% no lucro de uma instituição financeira (Reichheld & Sasser, 1990)

**c) SOLUÇÕES NÃO-ML EXISTENTES:** Os métodos tradicionais, sem uso de Machine Learning para essa tarefa, envolvem análise manual dos dados, métodos estatísticos simples e pesquisa com os clientes, que necessitam de muito esforço e podem não gerar resultados satisfatórios e confiáveis. (OWOLABI; UCHE; ADENIKEN; EFIJEMUE; ATTAKORAH; EMI-JOHNSON; HINNEH, 2024)

### 1.3.3.2. REFERENCIAL TEÓRICO EM MACHINE LEARNING (O que deve conter - Foco nos Algoritmos):

#### a) **MODELO A MODELO B:**

a) **Regressão Logística** A Regressão Logística é um algoritmo de classificação que modela a probabilidade de uma instância pertencer a uma determinada classe. Apesar do nome "regressão", trata-se de um modelo de classificação binária que utiliza a função sigmoide para transformar saídas lineares em probabilidades entre 0 e 1. O modelo calcula uma combinação linear das features e aplica a função logística:  $P(y=1|x) = 1 / (1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)})$ . O treinamento é realizado através da maximização da verossimilhança, frequentemente utilizando métodos como descida de gradiente. A Regressão Logística é amplamente utilizada por sua interpretabilidade, velocidade e eficácia em problemas linearmente separáveis.

b) **Árvore de Decisão:** A Árvore de Decisão é um algoritmo não paramétrico que particiona recursivamente o espaço de features com base em regras de decisão. Cada nó interno representa um teste sobre uma feature, cada ramo representa um resultado do teste e cada folha representa uma classe. O algoritmo utiliza medidas como Ganho de Informação, Índice Gini ou Erro de Classificação para selecionar as melhores features para divisão. Árvores de Decisão são intuitivas, não requerem normalização de dados e podem capturar relações não lineares. No entanto, são propensas a overfitting, o que pode ser

mitigado através de poda (pruning) ou utilização em ensembles como Random Forest.

- b) **MODELO DE ENSEMBLE (Se usar):** Se você usar uma técnica de conjunto (como Random Forest), explique a ideia de combinar modelos e por que isso ajuda a melhorar o resultado.

## **1.5. METODOLOGIA E DESENVOLVIMENTO (O Pipeline de ML)** Esta seção deve ser um manual técnico detalhado.

### **1.5.1. ANÁLISE EXPLORATÓRIA DE DADOS (EDA) (O que deve conter):**

- a) O dataset utilizado possui 10.000 instâncias e 14 colunas originais, reduzidas para 11 após remoção de colunas irrelevantes (RowNumber e Surname). As variáveis incluem tipos numéricos inteiros (CreditScore, Age, Tenure), floats (Balance, EstimatedSalary) e categóricas (Geography, Gender). O diagnóstico revelou ausência de valores nulos, porém identificou-se significativo desbalanceamento de classes: apenas 20,37% dos clientes cancelaram (classe positiva), enquanto 79,63% permaneceram ativos. Foram gerados histogramas para distribuição de idades, boxplots para análise de outliers em salário estimado e gráficos de barras para composição por gênero e país. A correlação entre variáveis foi analisada através de matriz de calor, revelando que idade (Age) apresenta maior correlação com o cancelamento.

### **1.5.2. PRÉ-PROCESSAMENTO E FEATURE ENGINEERING (O que deve conter):**

- a) A divisão dos dados seguiu proporção 70% treino e 30% teste, estratificada pela variável alvo para preservar a distribuição original das classes. Esta estratégia é crucial para garantir que modelos sejam avaliados em dados representativos. Como não há valores nulos, não foi necessário tratamento específico. Para variáveis categóricas, aplicou-se One-Hot Encoding em "Geography" (3 países) e Ordinal Encoding em "Gender" (binária). Variáveis numéricas foram padronizadas utilizando StandardScaler, que centraliza os dados na média zero com desvio padrão unitário, adequado para algoritmos sensíveis à escala como Regressão Logística. Em engenharia de

atributos, criou-se a feature "RendaPorProduto" (EstimatedSalary/NumOfProducts) para capturar valorização do cliente por produto contratado.

### 1.5.3. IMPLEMENTAÇÃO, TREINAMENTO E OTIMIZAÇÃO DOS MODELOS (A e B) (O que deve conter):

- a) A implementação utilizou a biblioteca scikit-learn no ambiente Google Colab. Para ambos os modelos, construiu-se um pipeline integrando pré-processamento e algoritmo. A otimização de hiperparâmetros foi realizada com GridSearchCV e validação cruzada de 5 folds. Na Regressão Logística, testou-se regularização L1/L2 com C variando [0.01, 0.1, 1, 10, 100]. Para Árvore de Decisão, explorou-se max\_depth [3, 5, 10, None], min\_samples\_split [2, 5, 10] e criterion ["gini", "entropy"]. O balanceamento de classes foi tratado com parâmetro class\_weight='balanced' e oversampling via SMOTE apenas no conjunto de treino.

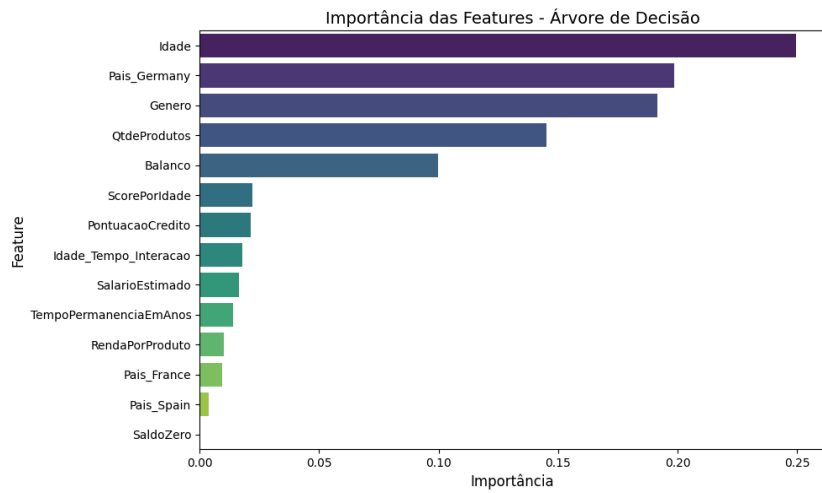
### 1.5.4. AVALIAÇÃO E ANÁLISE DE RESULTADOS (O que deve conter):

- a) **Comparação Quantitativa:**

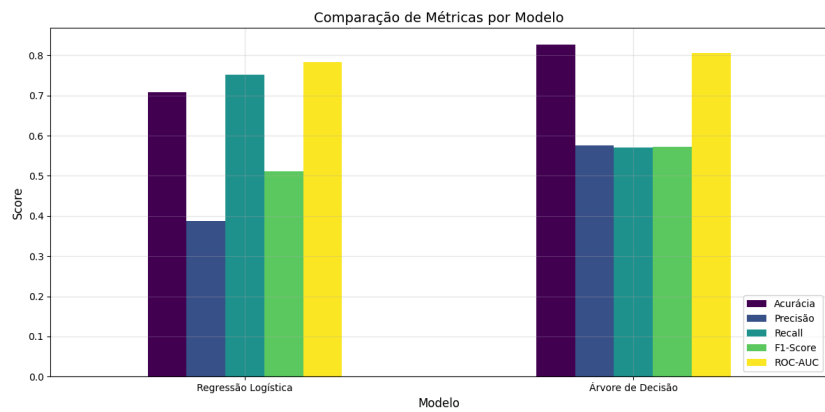
| Modelo              | Acurácia | Precisão | Recall | F1-Score | ROC-AUC |
|---------------------|----------|----------|--------|----------|---------|
| Regressão Logística | 70,77%   | 38,77%   | 75,12% | 51,14%   | 78,27%  |
| Árvore de Decisão   | 82,67%   | 57,50%   | 57,12% | 57,31%   | 80,58%  |

A partir da tabela, é possível perceber que a Árvore de Decisão excedeu em grande parte das métricas, com exceção do Recall.

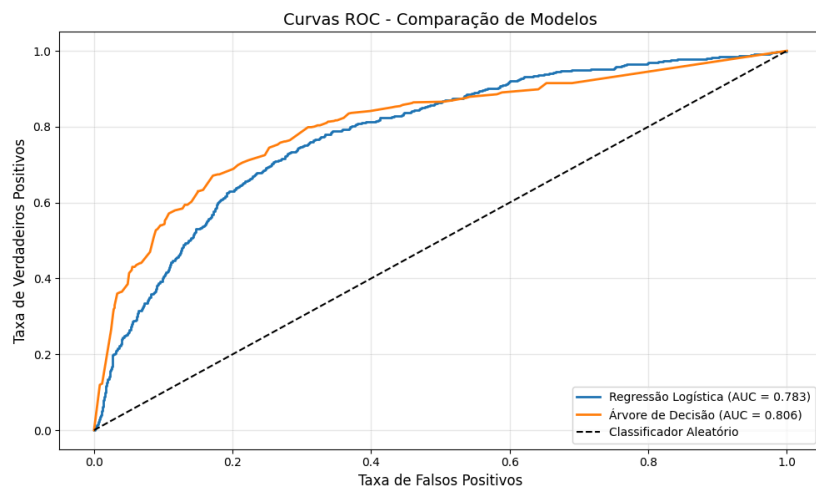
- b) **Justificativa das Métricas:** No contexto de churn bancário, o Recall é prioritário, pois identificar corretamente clientes que cancelaram (verdadeiros positivos) é mais crucial do que evitar falsos positivos. Um falso negativo (cliente que cancela não identificado) representa perda de receita, enquanto um falso positivo gera apenas custo de retenção. Portanto, privilegiamos modelos com maior Recall, mesmo com trade-off em Precisão.
- c) **Visualizações:**



Importância das Features - Árvore de Decisão



Comparação de métricas por modelo



Curvas ROC - Comparação de Modelos

d) **Análise Crítica:** A Árvore de Decisão apresentou desempenho superior em diversas métricas, com exceção do Recall (75,12% na Regressão Logística vs. 57,12% na Árvore de Decisão). Este resultado sugere que relações não lineares e interações entre features são relevantes para prever churn. A Regressão Logística, embora menos precisa, oferece maior interpretabilidade através dos coeficientes. A Árvore de Decisão também permite análise de importância de features, revelando "Idade", "SalarioEstimado" e "País" como preditores mais influentes. O trade-off entre desempenho e interpretabilidade deve considerar o objetivo do negócio: para ações automatizadas de retenção, prioriza-se desempenho; para análises estratégicas, interpretabilidade.

## 2. REFERÊNCIAS

### 2.1. Churn Bancário e Contexto

ANDERSON, R. *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford: Oxford University Press, 2007.

VERBEKE, W.; MARTENS, D.; BAESENS, B. Social network analysis for customer churn prediction. *Applied Soft Computing*, v. 14, p. 431–446, 2014.

DOI: <https://doi.org/10.1016/j.asoc.2013.08.015>

KOTLER, P.; KELLER, K. L. *Administração de Marketing*. São Paulo: Pearson, 2012.

Owolabi, O. S., Uche, P. C., Adeniken, N. T., Efijemue, O., Attakorah, S., Emi-Johnson, O. G. and Hinneh, E. (2024) Comparative Analysis of Machine Learning Models for Customer Churn Prediction in the U.S. Banking and Financial Services: Economic Impact and Industry-Specific Insights. *Journal of Data Analysis and Information Processing*, 12, 388-418. doi: [10.4236/jdaip.2024.123021](https://doi.org/10.4236/jdaip.2024.123021)

### 2.2. Conceitos: Churn, Balance, Credit Score

FICO. *What is a credit score?* Disponível em: <https://www.fico.com>

KOTU, V.; DESHPANDE, B. *Data Science: Concepts and Practice*. Burlington: Morgan Kaufmann, 2019.

LEMMENS, A.; GUPTA, S. Managing churn to maximize profits. *Harvard Business Review*, 2020. Disponível em: <https://hbr.org>

### 2.3. Fundamentação em Machine Learning

MITCHELL, T. M. *Machine Learning*. New York: McGraw-Hill, 1997.

MURPHY, K. P. *Machine Learning: A Probabilistic Perspective*. Cambridge: MIT Press, 2012.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer, 2009.

### 2.4. Algoritmos Utilizados

Regressão Logística:

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. *Applied Logistic Regression*. Hoboken: Wiley, 2013.

Árvore de Decisão:

QUINLAN, J. R. Induction of decision trees. *Machine Learning*, v. 1, n. 1, p. 81–106, 1986.

BREIMAN, L. et al. *Classification and Regression Trees*. Boca Raton: CRC Press, 2017.

## 2.5. Pipeline e Pré-processamento

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. Burlington: Morgan Kaufmann, 2011.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

Disponível em: <http://jmlr.org/papers/v12/pedregosa11a.html>

## 2.6. Dataset Utilizado

KAGGLE. *Churn Modelling Dataset*. 2016.

Disponível em: <https://www.kaggle.com/datasets>

## 2.7. Métricas e Avaliação

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, v. 45, n. 4, p. 427–437, 2009.

FAWCETT, T. An introduction to ROC analysis. *Pattern Recognition Letters*, v. 27, p. 861–874, 2006.