TATIANI VIRISSIMO DOS SANTOS

ATIVIDADE INTEGRADORA

VISUALIZAÇÃO DE DADOS

Treze Tílias
2025

TATIANI VIRISSIMO DOS SANTOS

ATIVIDADE INTEGRADORA

VISUALIZAÇÃO DE DADOS

Orientador: Prof. Mario H. A. C. Adaniya

Treze Tílias
2025

RESUMO

A presente pesquisa trata-se de um estudo sobre o impacto do desbalanceamento de classes na acurácia de modelos preditivos, utilizando como base o dataset *Marketing Bank* da UCI. A pesquisa foi aplicada por meio da análise de um modelo de *Decision Tree Classifier*, observando seu desempenho na previsão das classes "Sim" e "Não".

Essa pesquisa tem como objetivo geral compreender os fatores que levaram o dataset a apresentar uma grande discrepância entre as classes e como esse desbalanceamento afetou a precisão do modelo, especialmente na previsão da classe minoritária ("Sim").

De acordo com o estudo bibliográfico desenvolvido, foi possível mostrar que datasets desbalanceados impactam diretamente o aprendizado supervisionado, favorecendo a predição da classe majoritária e prejudicando a identificação da classe minoritária. Esse problema é comum em diversas áreas e pode distorcer a performance dos modelos de machine learning.

Para o embasamento teórico, utilizou-se literatura acadêmica sobre aprendizado de máquina, desbalanceamento de classes e técnicas de balanceamento, além de experimentos práticos no dataset em questão.

Os métodos utilizados na pesquisa foram explicativos e descritivos, com abordagem quantitativa, analisando estatísticas do dataset e o impacto da distribuição dos dados na performance do modelo.

Por fim, a pesquisa constatou que o alto número de respostas "Não" no dataset *Marketing Bank* influenciou significativamente o modelo de *Decision Tree Classifier*, reduzindo sua capacidade de prever corretamente a classe "Sim". O desbalanceamento de classes pode ser mitigado por técnicas como, reamostragem dos dados, ajuste de pesos no modelo ou o uso de algoritmos mais robustos para esse tipo de problema.

Palavras-chave: Desbalanceamento de Classes; Aprendizado de Máquina; Classificação Preditiva; Marketing Bancário; Decision Tree; Inteligência Artificial; Modelagem de Dados; Acurácia de Modelos.

SUMÁRIO

1 INTRODUÇÃO

O presente estudo tem como objetivo analisar de forma aprofundada o banco de dados *Bank Marketing* da UCI, que se caracteriza por uma distribuição de classes extremamente desigual. Inicialmente, constatou-se que mais de 88% dos registros correspondem a respostas negativas—indicando que os clientes não aceitaram as ofertas dos bancos—enquanto apenas 11% refletem a aceitação das mesmas. Essa discrepância significativa levanta questionamentos acerca dos fatores subjacentes que possam ter contribuído para a formação de um dataset com resultados tão predominantemente negativos.

A relevância deste trabalho reside na necessidade de compreender, de forma sistemática e fundamentada, as possíveis causas desse desbalanceamento. A acurácia e a eficácia dos modelos de aprendizado de máquina estão intrinsecamente ligadas à qualidade e à distribuição dos dados utilizados para treinamento. Além disso, o trabalho propõe-se a utilizar diversas técnicas de visualização de dados, conforme abordadas em aulas, com o intuito de promover uma análise exploratória mais detalhada.

2.1 DESENVOLVIMENTO

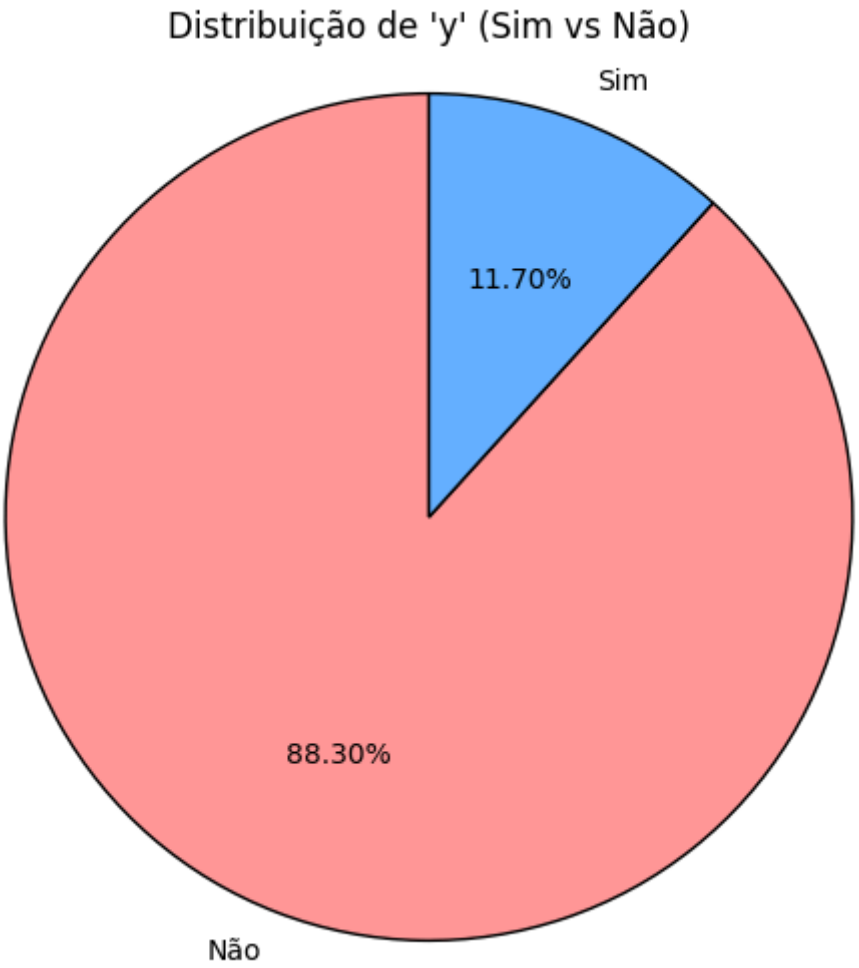
2.1.1 Escolhendo um banco de dados

O banco de dados *Bank Marketing* da UCI Machine Learning foi utilizado anteriormente no treinamento do modelo *Decision Tree Classifier*, com o objetivo de prever se um cliente aceitaria ou não a oferta do banco. Os resultados obtidos após o treinamento chamaram a atenção, pois o modelo apresentou um desempenho superior na previsão de respostas negativas, mas teve menor precisão ao prever respostas afirmativas. Esse comportamento levantou questionamentos sobre os fatores que poderiam estar impactando o aprendizado do modelo.

2.2.1 Entendendo os dados.

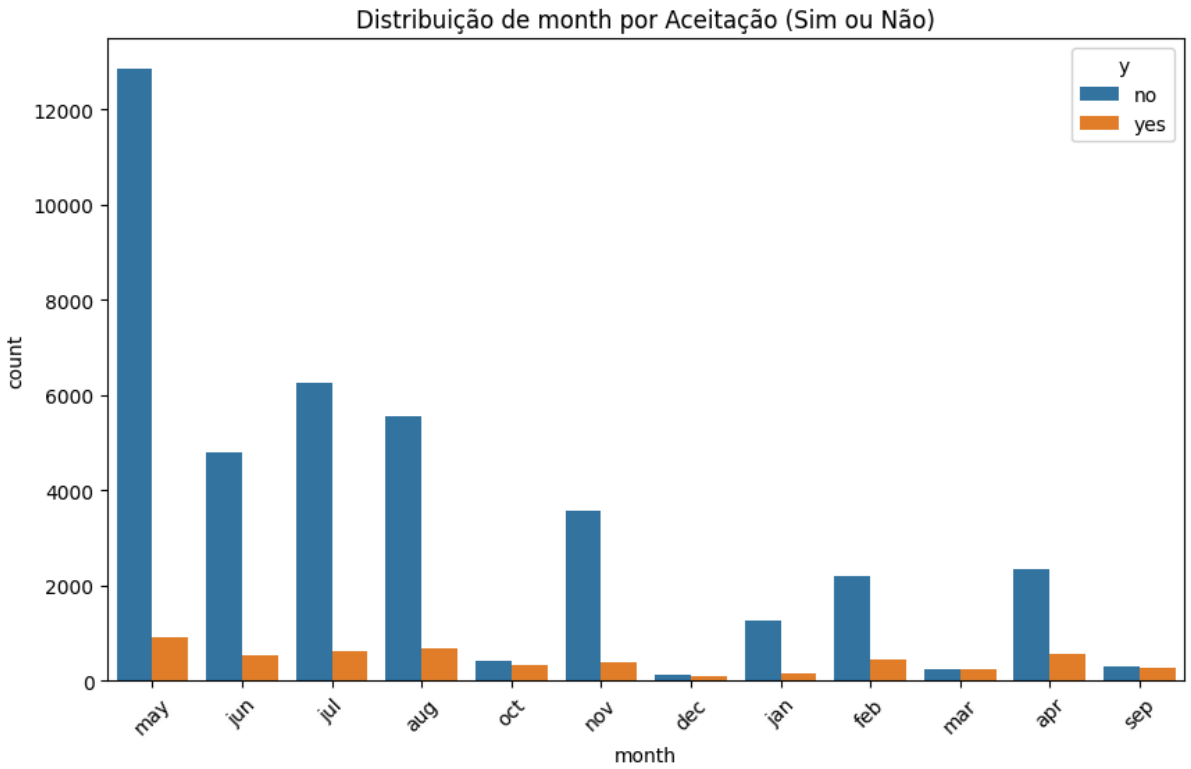
Os dados analisados consistem em uma tabela contendo diversas variáveis categóricas. A variável de interesse, denominada *Y*, representa o resultado da aceitação ou recusa da oferta pelo cliente.

A partir dessa variável, foi possível visualizar a distribuição das respostas "Sim" e "Não". Ao gerar um gráfico pizza, observou-se uma discrepância significativa entre os valores, conforme ilustrado na figura abaixo.



A partir dessa observação, foi possível identificar o fator que impactou o aprendizado do modelo *Decision Tree Classifier*. A distribuição dos dados revelou que 88% dos clientes recusaram a oferta do banco, enquanto apenas 11% a aceitaram.

Dando continuidade à análise exploratória, buscou-se compreender as razões para essa discrepância nos dados. A partir da observação de diferentes gráficos, não foi possível identificar, de forma precisa, um único fator responsável pela alta taxa de recusas. No entanto, alguns gráficos apresentaram informações relevantes, como a relação entre os meses do ano e as respostas dos clientes, evidenciando padrões que merecem investigação mais aprofundada.



Maio, por exemplo, apresentou um número significativamente maior de recusas, assim como os meses subsequentes. Esse fenômeno pode estar relacionado à maior frequência de ligações realizadas nesses períodos. No entanto, ainda não foi identificado um padrão significativo que explique, de forma conclusiva, os fatores responsáveis pela alta taxa de respostas negativas.

2.3.1 Análise Aprofundada

A partir de uma análise mais detalhada dos dados, foi possível compreender o contexto de maneira mais ampla. O conjunto de dados em questão abrange o período de 2008 a 2010 nos Estados Unidos. O ano de 2008, especificamente, foi marcado pela segunda pior crise financeira da história do país, conhecida como *Crise de 2008*. Esse evento teve início com o colapso do mercado imobiliário, que desencadeou impactos econômicos globais, afetando também diversos países da Europa.

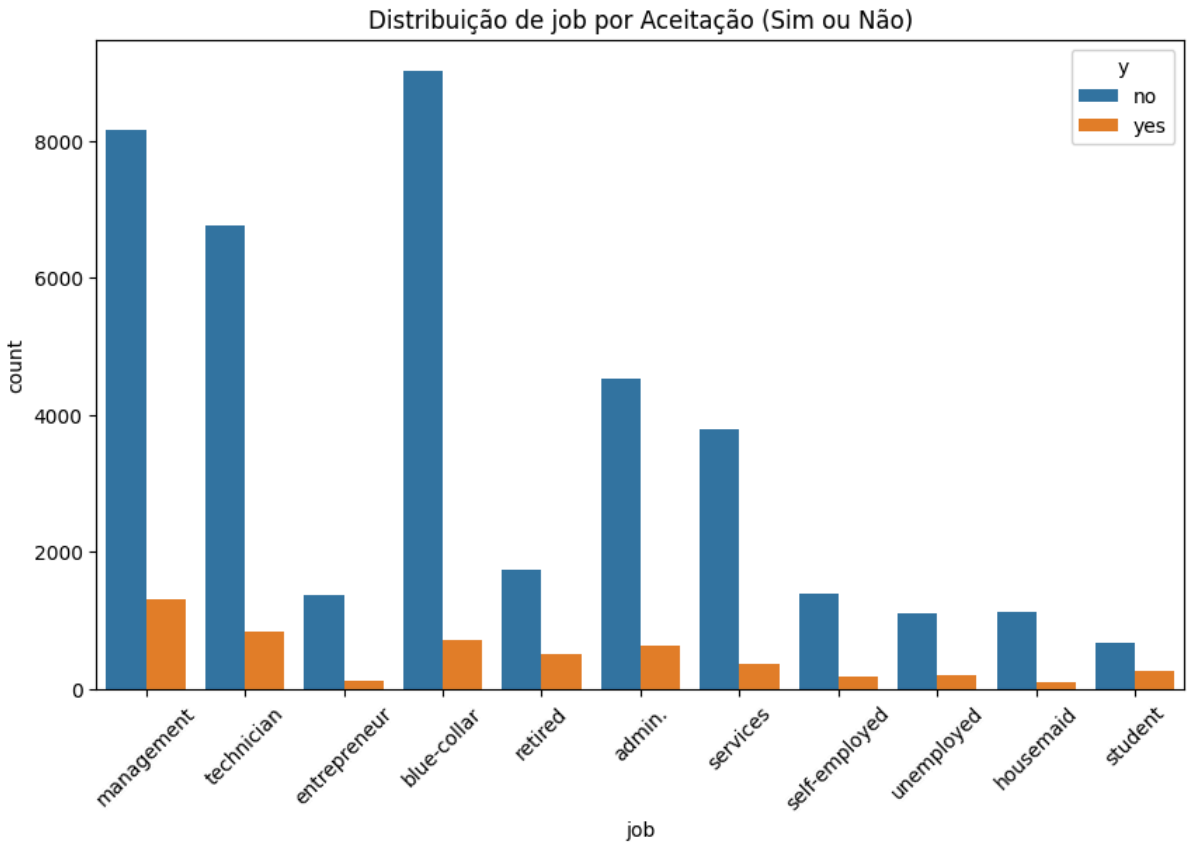
Para estabelecer a relação entre a *Crise de 2008* e os dados do banco de dados *Bank Marketing*, é necessário considerar o impacto global desse evento e sua origem em 15 de setembro de 2008. A crise teve início com a concessão excessiva de crédito à população, beneficiando,

principalmente, o setor imobiliário. Com o aumento da valorização dos imóveis devido à alta demanda e à oferta de crédito com juros reduzidos, muitas pessoas contraíram empréstimos superiores à sua capacidade de pagamento. Quando os preços dos imóveis atingiram níveis insustentáveis, um efeito de retrocesso ocorreu: a população não conseguiu mais arcar com os pagamentos, resultando em um grande número de endividados em todo o país. Como consequência, as bolsas de valores despencaram, levando governos de diversas nações a adotarem medidas emergenciais para conter a crise (*Stoodi*, 2023).

Com o aumento das taxas de juros por parte dos bancos, os custos dos empréstimos se elevaram, tornando o pagamento das parcelas ainda mais difícil para os devedores. Como resultado, diversas instituições financeiras enfrentaram problemas de liquidez, o que marcou o início da crise financeira (*Politize!*, 2023).

2.4.1 Reflexo nos Dados

A partir desse contexto, observou-se que algumas categorias de ocupação (*job*) apresentaram uma taxa de recusa significativamente maior do que outras, conforme ilustrado no gráfico abaixo.



A categoria *blue-collar* foi a que apresentou a maior taxa de recusas. Esses trabalhadores podem ter sido significativamente impactados pela *Crise de 2008*, uma vez que muitas fábricas e indústrias fecharam ou reduziram o número de vagas naquele período. Estima-se que milhões de pessoas perderam seus empregos, especialmente em setores como construção, manufatura e serviços financeiros (*Toro Investimentos*, 2023). Esse fator pode justificar a maior taxa de recusas observada nessa classe ocupacional.

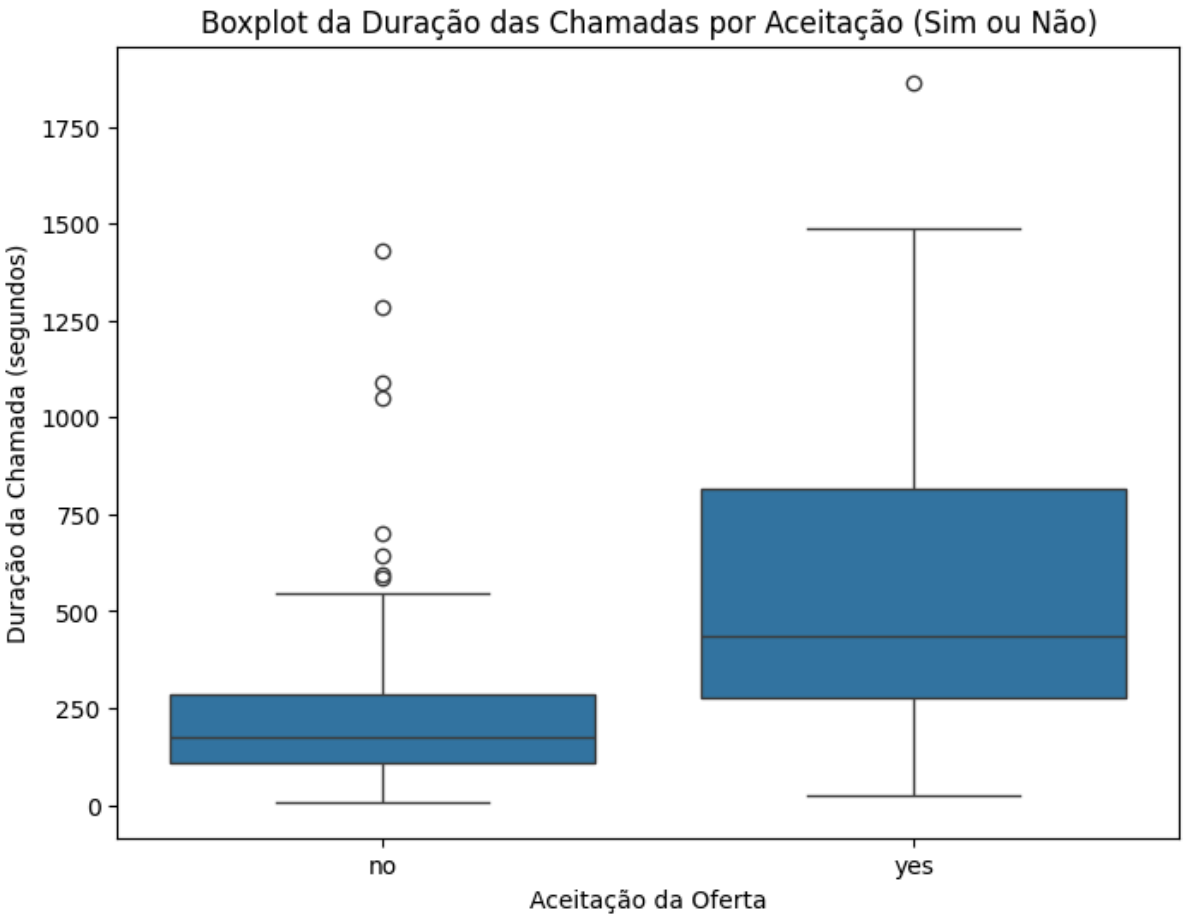
2.5.1 Impacto da Crise de 2008 nos dados.

Os valores observados no conjunto de dados tendem a ser predominantemente negativos devido à grave crise econômica enfrentada pelos Estados Unidos no período analisado, que abrange de maio de 2008 a novembro de 2010.

A taxa de respostas positivas corresponde a apenas 11% do total de respostas registradas, o que indica que poucos clientes aceitaram a

oferta bancária. No entanto, alguns fatores podem ser identificados entre os clientes que demonstraram maior propensão à aceitação.

Um dos padrões observados é o tempo de duração das chamadas. Clientes que permaneceram mais tempo na linha apresentaram uma maior taxa de aceitação em comparação àqueles que desligaram rapidamente, conforme demonstrado no gráfico abaixo.



2.6.1 Duração das Chamadas e a Aceitação da Oferta

Para compreender melhor a relação entre a duração das chamadas e a decisão dos clientes, foi analisada uma amostra de 150 registros. Os dados indicam que clientes que permanecem mais tempo na linha

apresentam maior propensão a aceitar a oferta, enquanto aqueles que encerram a chamada rapidamente tendem a recusá-la.

Esse comportamento sugere que a abordagem do atendente pode influenciar a decisão do cliente. Indivíduos que dedicam mais de quatro minutos à conversa demonstram maior interesse no conteúdo apresentado, o que pode indicar uma maior disposição para avaliar a oferta. Por outro lado, clientes que não estão interessados tendem a encerrar a chamada rapidamente, evitando futuras interações.

Dessa forma, um aspecto relevante para a estratégia do banco seria priorizar o retorno de chamadas para clientes que passaram mais tempo na linha antes de recusar a oferta. Esses clientes podem precisar de mais tempo para avaliar sua condição financeira e considerar a proposta, aumentando as chances de conversão em um contato futuro. Em contrapartida, insistir em clientes que recusam imediatamente pode ser menos eficaz, uma vez que demonstram desinteresse desde o início.

3 CONSIDERAÇÕES FINAIS

A crise financeira de 2008 foi um fator determinante nos resultados observados, sendo reconhecida como a segunda maior crise financeira dos Estados Unidos. A análise dos dados revelou que 88% das respostas em relação à aceitação da oferta do banco foram negativas, enquanto apenas 11% foram positivas. É possível observar que a população demonstrou cautela ao investir, uma vez que diversas instituições financeiras entraram em colapso, e setores como manufatura e construção foram severamente impactados, resultando em demissões em massa. O risco de alocar recursos financeiros nessas instituições, tornou-se elevado, especialmente para as classes economicamente menos privilegiadas, que foram as mais afetadas pela recessão.

Por outro lado, ainda que a recusa tenha sido predominante, a classe *Management* apresentou uma taxa de aceitação maior em comparação a outras categorias, ainda que a recusa tenha sido elevada. A experiência desse grupo com o mercado financeiro pode ter sido um fator que contribuiu para uma maior confiança na tomada de decisões, diferenciando-o das classes de menor renda, que adotaram uma postura mais conservadora na hora de investir.

Além disso, a relação entre a duração das chamadas e a aceitação das ofertas chamou a atenção. Os dados indicaram que clientes que permaneceram mais tempo na linha demonstraram maior predisposição à aceitação da proposta, enquanto aqueles que recusaram imediatamente apresentaram menor interesse. É possível que estratégias focadas em clientes que demonstram hesitação inicial, mas mantêm diálogo com os atendentes, possam aumentar a taxa de aceitação. O retorno de chamadas para esse perfil específico pode ser uma estratégia para captação de novos investidores.

REFERÊNCIAS

STOODI. **Crise de 2008: o que foi, causas e consequências**. Disponível em: https://blog.stoodi.com.br/blog/historia/crise-de-2008/. Acesso em: 25 mar. 2025.

POLITIZE! ***Crise financeira de 2008: entenda o que foi e suas consequências*.** Disponível em: https://www.politize.com.br/crise-financeira-de-2008/. Acesso em: 25 mar. 2025.