Análise de dados para determinação de clientes em potencial

Álvaro C. Negromonte, Júlio César B. da Silva, Luiz *Felipe* S. Lustosa, *Vinícius de S. Rodrigues*
(acn3, jcb3, lfsl, vsr)@cin.ufpe.br

Centro de Informática, Universidade Federal de Pernambuco, Recife, Brasil

***Abstract*—Este projeto tem como objetivo classificar o perfil de possíveis clientes de um produto de investimento, distinguindo aqueles com alta probabilidade de compra daqueles com baixa probabilidade. Utilizando o classificador probabilístico Ingênuo de Bayes, implementado em linguagem Python, a análise explora características demográficas, comportamentais e financeiras para identificar padrões relevantes. O projeto utiliza um conjunto de dados real de uma campanha de marketing bancário e busca otimizar estratégias para futuras ações. O modelo proposto é desenvolvido com ferramentas como Pandas e Matplotlib, enfatizando a aplicabilidade prática e a capacidade de adaptação às transformações sociais.**

***Index Terms*— Dados, cliente ideal, classificador probabilístico, Naive Bayes, Teorema de Bayes**

I. Objetivos

Nosso objetivo principal neste projeto é classificar o perfil de possíveis clientes de um produto de investimento, diferenciando entre aqueles que têm alta probabilidade de compra e aqueles que não têm. Por meio de uma análise exploratória detalhada, buscamos identificar a correlação entre os diversos atributos disponíveis, como características demográficas, comportamentais e financeiras, e a decisão final do cliente. Essa abordagem permitirá não apenas melhorar a compreensão do perfil do cliente ideal, mas também otimizar estratégias para futuras campanhas.

II. Justificativa

A escolha da proposta de análise de dados justifica-se pela compreensão, por parte do grupo, da similaridade do problema proposto com aqueles encontrados no mercado de trabalho. Além disso, interpretamos que o problema em questão pode ser adequadamente analisado por meio do Classificador Ingênuo de Bayes.

Outrossim, esse método é particularmente adequado, pois suporta modelos que lidam com dados binários e categóricos, permitindo sua combinação em um mesmo projeto. Ademais, o treinamento incremental oferecido pelo classificador possibilita atualizações contínuas no banco de dados, promovendo a adaptação das análises ao longo do tempo e em consonância com as transformações socias.

III. Base de dados

Nesta seção apresentaremos o conjunto de dados selecionado e descrever suas características.

A base de dados [5] consiste numa campanha do marketing de um banco para vender um produto de investimentos chamado “título”. Nela, há os seguintes atributos:

1) Colunas numéricas (int64):

1.1: Saldo_Conta_Corrente: saldo atual do cliente na conta corrente.
1.2: Qte_de_Ligações_Feitas: número de ligações feitas para o cliente durante a campanha.
1.3: Idade: idade do cliente.

2) Colunas categóricas (object):

2.1: Cliente_Comprou_o_Titulo?: indica se o cliente comprou o produto de investimento (Sim/Não).

2.2: Profissão: profissão do cliente (e.g., Operário, Administrador, Aposentado).

2.3: Estado_Civil: estado civil do cliente (e.g., Solteiro, Casado).

2.4: Formação: nível de escolaridade do cliente (e.g., Ensino Médio, Ensino Superior).

2.5: Cliente_Devedor?: indica se o cliente tem uma dívida registrada (Sim/Não).

2.6: Tem_Hipoteca?: indica se o cliente possui uma hipoteca (Sim/Não).

2.7: Tem_Emprestimo?: indica se o cliente possui um empréstimo (Sim/Não).

IV. ANÁLISE EXPLORATÓRIA DE DADOS

Nesta seção, descreveremos a análise de dados que foi realizada, a qual dividimos em três etapas principais: Entendimento dos dados, Preparação dos dados (tratamento) e Análise exploratória.

Na primeira etapa supracitada, realizamos uma análise superficial dos dados descobrindo o tamanho da tabela, como os dados estão organizados nela e os dados estatísticos das colunas numéricas. Na segunda etapa, nós verificamos se há dados nulos, incorretos ou faltantes a fim de limpar a base para tornar a análise mais precisa. Assim, eliminamos uma coluna com todos os dados faltantes.

Na etapa de Análise exploratória de fato, visamos a exploração e visualização dos dados para identificar padrões e tendências. Para isso, usamos técnicas estatísticas e análises descritivas que nos ajudaram a entender as relações entre variáveis e validar hipóteses

Buscando encontrar padrões, fizemos as seguintes análises e chegamos a alguns insights.

A. Quantitativo de compradores x Hipoteca

Observamos que clientes sem hipoteca apresentam uma probabilidade significativamente maior de realizar compras, tanto em termos absolutos quanto em porcentagem em relação àqueles que possuem hipoteca. Portanto, é recomendável direcionar campanhas de marketing especificamente para esse segmento de clientes sem hipoteca, a fim de maximizar as oportunidades de vendas e engajamento.

B. Quantitativo de compradores x Quantidade de ligações feitas

Notamos que, a partir da sexta ligação, o número de conversões diminui consideravelmente. Isso indica um desperdício de recursos, já que os profissionais envolvidos nas ligações estão investindo tempo em clientes que apresentam baixa probabilidade de conversão. Assim, recomenda-se estabelecer um limite de cerca de seis ligações por cliente, permitindo um direcionamento mais eficaz da equipe de vendas e priorizando o contato com clientes que estão mais propensos a realizar a compra nas primeiras interações.

C. Quantitativo de compradores x Estado Civil

A partir do gráfico, notamos que a taxa de não compra entre os casados é significativamente alta. Assim, é recomendado explorar por que tantos casados não compram. Isso pode ser devido a prioridades financeiras (como compra de imóveis, educação de filhos) ou falta de conhecimento sobre investimentos. Estratégias de marketing poderiam ser adaptadas para abordar essas preocupações e destacar como o investimento pode complementar seu planejamento financeiro familiar.

Além disso, notamos que os solteiros têm uma boa taxa de compra, o que indica que têm maior disposição para assumir riscos financeiros ou estão mais abertos a investir. Assim, esse grupo pode ser um alvo promissor para campanhas que enfatizam a construção de patrimônio e a independência financeira. O marketing pode explorar a ideia de que investimentos são uma forma de alcançar objetivos pessoais, como viajar, adquirir bens ou garantir uma aposentadoria confortável.

Por fim, notamos que a taxa de compra e não compra dos divorciados é equilibrada, o que indica que esse grupo pode estar em transição ou reavaliando suas situações financeiras após a separação. Assim, campanhas que abordam a reestruturação financeira após um divórcio podem ressoar bem com os divorciados. Oferecer informações sobre como os investimentos podem ajudar a construir uma nova segurança financeira ou proporcionar estabilidade durante períodos de mudança pode ser atraente.

D. *Quantitativo de compradores x Profissão*

A partir dessa análise, concluímos que há perfis profissionais com maior interesse de compra do título ou com maiores condições financeiras para comprá-lo. São eles: administrador, técnico, operário, aposentados, serviços gerais e estudante. Assim, essas profissões com alta taxa de compra indicam grupos prioritários para campanhas futuras, pois já demonstram interesse no produto. Além disso, é

recomendado personalizar essas campanhas para profissões específicas com mensagens que estejam de acordo com seus interesses e estilos de vida para aumentar a taxa de conversão, já que algumas dessas profissões também têm uma alta quantidade de não compradores.

Realizando uma análise mais específica da classe mais compradora, os administradores, nota-se que representam um dos públicos-alvo mais significativos para o banco. Como administradores tendem a ter maior estabilidade financeira e conhecimento sobre produtos financeiros, sua propensão a compra é mais consistente. Assim, as campanhas para eles poderiam destacar o produto de investimento como parte de uma estratégia de crescimento financeiro, possivelmente oferecendo pacotes de investimento voltados para profissionais que buscam diversificação e crescimento

E. *Quantitativo de compradores x Devedores*

A partir dessa análise, concluímos que a diferença entre clientes devedores e não devedores demonstra que a estabilidade financeira é um fator imprescindível na decisão de comprar o título. Assim, é recomendado que as campanhas devem priorizar clientes não devedores, que estão mais preparados financeiramente e têm maior probabilidade de comprar o título, enquanto campanhas educativas podem ser uma estratégia de suporte para clientes devedores visando um investimento no título em longo prazo.

F. Quantitativo de compradores x Tem ou não empréstimo

A partir dessa análise, notamos que a presença ou ausência de empréstimos é um fator considerável na propensão dos clientes para investimentos. Assim, é recomendado que campanhas de investimento devem priorizar clientes sem empréstimos, enquanto os clientes com empréstimos poderiam ser beneficiados por conteúdos educativos e estratégias de planejamento financeiro. Dessa forma, o banco pode maximizar o retorno de suas campanhas, engajando clientes mais preparados para investir e, ao mesmo tempo, apoiando o planejamento financeiro daqueles que ainda não estão prontos.

G. Quantitativo de compradores x Saldo

A partir dessa análise, concluímos que a faixa de saldo desempenha um papel importante na decisão de compra. Assim, são recomendadas que sejam realizadas campanhas focadas em clientes com saldos moderados, visto que eles têm mais chances de sucesso, enquanto aqueles com saldo alto podem exigir produtos financeiros mais robustos para captar seu interesse, ampliando o engajamento entre diferentes perfis financeiros.

H. Quantitativo de compradores x Formação

A partir dessa análise, concluímos que a maior propensão de compra entre clientes com ensino médio completo e ensino superior completo indica que um nível educacional mais alto pode estar associado a uma maior conscientização sobre a importância de investimentos para o futuro financeiro. Assim, campanhas de marketing podem se beneficiar ao promover o título para esse grupo, focando em como o investimento se encaixa em um plano financeiro de longo prazo e em objetivos de vida importantes e reforçando a importância do planejamento financeiro e os benefícios de investir desde cedo.

Além disso, notamos que a menor propensão de compra entre clientes com ensino fundamental sugere que parte do público pode ter menos conhecimento ou confiança em investimentos, o que aponta para uma oportunidade de capacitar clientes com menos familiaridade sobre o tema, aumentando o entendimento sobre os benefícios do título e, consequentemente, atraindo-os para o investimento.

Assim, uma abordagem educacional, com conteúdos que expliquem os conceitos básicos de investimento e seu impacto no futuro financeiro, poderia ajudar a expandir o mercado. Realizar workshops, disponibilizar materiais informativos e utilizar redes sociais para simplificar o tema pode atrair clientes de diferentes perfis educacionais e estimular o interesse em investir.

I. *Quantitativo de compradores x Idade*

A partir dessa análise, concluímos que clientes da faixa etária de 25 a 40 anos estão entre os maiores compradores. Essa faixa etária representa um grupo que geralmente busca consolidar a estabilidade financeira e planejar o futuro. Esse grupo está em um

estágio de vida onde o interesse em produtos de investimento é elevado, seja para o crescimento patrimonial, aposentadoria, ou projetos de longo prazo. Nesse sentido, é recomendado que campanhas futuras destaquem como o produto de investimento contribui para a segurança financeira e o crescimento pessoal (algo que é valorizado por essa faixa etária), além de promover o investimento como o início de uma relação financeira com o banco.

J. *Quantitativo de compradores na faixa etária de 25 a 40 anos x Formação*

A partir dessa análise, notamos que, embora não sejam considerados perfis tradicionais de investidores, como foi analisado anteriormente, estudantes (71,1%) e desempregados (60,1%) apresentam uma alta propensão de compra, talvez por estarem em busca de crescimento financeiro e segurança para o futuro. Assim, são recomendadas campanhas direcionadas para esses grupos, o que pode enfatizar como os investimentos podem ajudar a garantir segurança financeira e oportunidades para crescimento patrimonial, mesmo para quem está em início de carreira ou em transição.

Após essa fase de análises de tendências, realizamos umas análises mais específicas a fim de chegar no perfil de cliente ideal (ICP) baseada nos clientes administradores (cuja profissão constitui a maior parte dos compradores): faixa etária mais propensa a compra, saldo corrente dos compradores, hipoteca, devedor, empréstimo e quantidade de ligações feitas.

A partir dessas análises específicas em relação aos administradores, notamos que o cliente ICP é um administrador com faixa etária de 25 a 40 anos, cuja maioria tem um saldo de até R$ 15.000, não possuem hipoteca, não é devedora, não tem empréstimos e com baixa quantidade de ligações para compra do título. Assim, o perfil identificado pode ajudar a moldar futuras campanhas de marketing. Além disso, focar em administradores entre 25 e 40 anos com um saldo corrente modesto pode ser uma abordagem eficaz. Isso pode incluir ofertas personalizadas que abordem suas necessidades específicas de investimento. Dado que muitos administradores têm um perfil financeiro conservador (sem dívidas ou hipotecas), iniciativas de educação financeira podem ser implementadas para orientá-los sobre a importância de investir, como maximizar seu saldo e explorar opções de investimentos que se alinhem a seus objetivos financeiros, incentivando, consequentemente, a compra do título. Por fim, com base na correlação entre a quantidade de ligações e a conversão de vendas, as equipes de vendas e marketing devem considerar estratégias para otimizar o número de contatos e a qualidade das interações. Por exemplo, personalizar as abordagens de comunicação e utilizar insights dos clientes pode aumentar a eficácia das ligações.

V. CLASSIFICADOR INGÊNUO DE BAYES

A. Naive Bayes

O Naive Bayes é um algoritmo de aprendizado de máquina baseado no Teorema de Bayes, que calcula a probabilidade de uma classe com base nas características fornecidas. Ele é chamado de "naive" (ingênuo) porque assume que todas as variáveis preditoras (features) são independentes entre si, o que raramente acontece na prática, mas ainda assim funciona bem em muitos problemas.

O Teorema de Bayes é definido como:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Sendo que:

- $P(Y|X)$ → Probabilidade de a amostra pertencer à classe Y dado os atributos X (probabilidade posterior);
- $P(X|Y)$ → Probabilidade dos atributos X ocorrerem dado que a amostra pertence à classe Y (verossimilhança);
- $P(Y)$ → Probabilidade da classe Y ocorrer sem considerar X (probabilidade a priori);
- $P(X)$ → Probabilidade de observar os atributos X (evidência, um valor constante para todas as classes).

O Naive Bayes escolhe a classe Y que maximiza a probabilidade posterior P(Y|X).

B. O Gaussian Naive Bayes (GNB)

O Gaussian Naive Bayes (GNB) é uma variação do Naive Bayes que assume que as variáveis numéricas seguem uma distribuição normal (gaussiana) dentro de cada classe.

A distribuição normal é definida por:

$$p(x_i|y_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(x_i-\mu_j)^2}{2\sigma_j^2}}$$

Onde:

- Xi → Valor da variável X;
- μY → Média da variável X dentro da classe Y;
- (σ^2) Y → Variância da variável X dentro da classe Y.

Na prática, o GNB funciona em duas etapas: treinamento e Predição. No treinamento para cada variável preditora Xi e para cada classe Y, o modelo estima uma média (μY) e o desvio padrão(σY). Outrossim, na predição para um novo exemplo, o modelo calcula a probabilidade de ele pertencer a cada classe usando a fórmula de Bayes. A classe com maior probabilidade posterior P(Y | X) é escolhida.

A equipe optou por seguir esse modelo devido a própria natureza do problema proposto, cujo objetivo da aplicação é prever se um cliente comprará ou não um título financeiro com base em suas características (idade, saldo bancário, número de ligações recebidas, etc.). Esse é um problema de classificação binária, onde o modelo precisa decidir entre duas classes possíveis: Classe 1 (o cliente comprou o título financeiro) e Classe 0 (o cliente não comprou o título financeiro).

Ademais, o Naive Bayes é eficiente em cenários com poucos dados, pois utiliza probabilidades baseadas em estimativas simples (média e variância). Em aplicações de marketing bancário, muitas vezes há dados limitados sobre clientes interessados em determinado produto financeiro, tornando o GNB uma opção viável. Além disso, o GNB assume que os dados seguem uma distribuição normal dentro de cada classe. Isso é válido para algumas variáveis como: idade, saldo da conta corrente e a quantidade de ligações recebidas.

C. Implementação do Classificador

O código produzido é constituído pelas seguintes etapas: bibliotecas utilizadas, tratamento dos dados, treinamento do modelo GNB, previsão de classe, e medição da importância das variáveis.

1) Bibliotecas utilizadas

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from sklearn.naive_bayes import GaussianNB
from sklearn.metrics import classification_report
import numpy as np
```

I) A biblioteca pandas é utilizada para manipular e estruturar os dados em DataFrames, que são tabelas semelhantes ao Excel.

II) O scikit-learn (sklearn) é uma biblioteca fundamental para aprendizado de máquina no Python;

1) train_test_split(): divide os dados em conjunto de treino e conjunto de teste;
2) LabelEncoder(): converte variáveis categóricas (como "Sim"/"Não" ou "Masculino"/"Feminino") em números, pois algoritmos de Machine Learning só trabalham com valores numéricos;
3) GaussianNB(): é um classificador baseado no Teorema de Bayes, que assume que os dados seguem uma distribuição normal (Gaussiana).
4) classification_report(): gera métricas de desempenho do modelo, incluindo precisão, recall e F1-score;
5) NumPy: é uma biblioteca para cálculo numérico e manipulação de arrays.

2) Tratamento dos dados

```python
# Carregar os dados corretamente (removendo índices extras)
df = pd.read_excel("bank_marketing.xlsx", header=1)

# Remover a coluna "Unnamed: 0" caso exista
df.drop(columns=["Unnamed: 0"], errors="ignore", inplace=True)

# Renomear a coluna-alvo corretamente
df.rename(columns={"Cliente_Comprou_o_Titulo?": "Alvo"}, inplace=True)

# Converter colunas numéricas corretamente
num_cols = ["Idade", "Saldo_Conta_Corrente", "Qte_de_Ligações_Feitas"]
for col in num_cols:
    df[col] = pd.to_numeric(df[col], errors="coerce")

# Transformar variáveis categóricas
label_encoders = {}
for col in df.select_dtypes(include=["object"]).columns:
    le = LabelEncoder()
    df[col] = le.fit_transform(df[col])
    label_encoders[col] = le

# Separar variáveis independentes e dependentes
X = df.drop(columns=["Alvo"])
y = df["Alvo"]

# Dividir os dados em treino e teste
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Nesse trecho do código, ocorre a leitura do arquivo que será utilizado tanto para treinar o modelo quanto para testá-lo. Durante essa etapa, é feito um ajuste para considerar a linha 1 como cabeçalho. Em seguida, remove-se a coluna “Unnamed: 0”, que pode surgir devido à formatação do arquivo “bank_marketing.xlsx”.

Além disso, a coluna “Cliente_Comprou_o_Titulo?” é renomeada para “Alvo” para tornar o código mais curto e legível. Também é realizada a conversão das colunas numéricas para garantir que os valores sejam manipulados corretamente, pois originalmente estavam em formato de texto.

As variáveis categóricas são transformadas em valores numéricos para que possam ser processadas pelo modelo de aprendizado de máquina.

Após esse tratamento, ocorre a separação das variáveis independentes (X) e dependentes (y). X recebe todas as colunas exceto a variável-alvo “Alvo” (dados de entrada), enquanto y recebe apenas essa variável-alvo (o que queremos prever).

Por fim, os dados são divididos em dois conjuntos: 80% para treinamento e 20% para teste, garantindo que o modelo possa ser avaliado corretamente.

3) Treinamento do Modelo

```python
# Treinar o modelo
model = GaussianNB()
model.fit(X_train, y_train)
```

O comando model = GaussianNB() cria um modelo de classificação utilizando a classe GaussianNB() da biblioteca scikit-learn. Em seguida, o método model.fit() treina o modelo com os dados de treinamento, permitindo que ele aprenda as probabilidades condicionais dos atributos em relação às classes-alvo.

Durante o treinamento, o modelo calcula estatísticas essenciais, como a média e a variância de cada variável (feature) dentro de cada classe (0 ou 1, no caso de um problema de classificação binária).

Para cada variável presente em X_train, o modelo estima a média e a variância dentro de cada classe da variável-alvo y_train. Posteriormente, aplica o Teorema de Bayes para calcular a probabilidade de cada classe e, com base nesses valores, atribui a classe mais provável a cada nova observação.

3.1) Previsão da classe

```python
# Fazer previsões
y_pred = model.predict(X_test)
```

Esse código gera previsões utilizando o modelo Naive Bayes (GaussianNB) previamente treinado.Quanto à entrada:

I) X_test contém os dados de teste, ou seja, novas observações sem os rótulos reais (y_test) que o modelo ainda não viu.

II) Cada linha de X_test representa um novo cliente, e cada coluna contém uma variável preditiva (exemplo: idade, saldo da conta, número de ligações feitas etc.).

3.2) Processo interno do modelo:

O modelo calcula a probabilidade de cada

classe (por exemplo, comprou (1) ou não comprou (0)). Além disso, utiliza o Teorema de Bayes, juntamente com a distribuição Gaussiana, para estimar a probabilidade de cada instância pertencer a uma determinada classe. Por fim, classe com a maior probabilidade a posteriori é atribuída como previsão final.

3.3) Saída (y_pred):

y_pred será um vetor contendo as previsões do modelo para cada linha de X_test.

3.4) Importância das Variáveis

```
# Importância das variáveis (baseada na média das probabilidades)
feature_importance = np.abs(model.theta_[1] - model.theta_[0])
```

Esse código calcula a importância das variáveis (features) no modelo Gaussian Naive Bayes (GNB) com base na diferença das médias das distribuições das classes.

No Gaussian Naive Bayes, cada variável preditiva (feature) é modelada como uma distribuição Gaussiana (normal) separada para cada classe da variável-alvo.

O atributo model.theta_ contém as médias (μ\muμ) estimadas de cada variável para cada classe: model.theta_[0] (vetor com as médias das features quando a classe-alvo é 0) e model.theta_[1] (vetor com as médias das features quando a classe-alvo é 1).

O código calcula a diferença absoluta entre as médias das features nas duas classes (1 e 0). Quanto maior essa diferença, maior a influência da variável na separação entre as classes. E se uma variável tem médias muito próximas para ambas as classes, seu impacto na classificação é baixo.

O resultado (feature_importance) é um vetor, onde cada valor representa a importância relativa de uma variável dentro do conjunto de dados.

3.5) Resto do código

Ocorre a soma total das importâncias (importance_sum). Em seguida, os valores são normalizados para que representem percentuais (%) e somem exatamente 100%. Além disso, o último valor (importance_percentage[-1]) é ajustado para compensar eventuais erros numéricos no arredondamento, garantindo que a soma final seja precisamente 100%.

Também há a ordenação das variáveis por importância, onde são combinadas as colunas de X, suas importâncias normalizadas (%) e os valores médios para a classe 1 (model.theta_[1]). As variáveis são organizadas da mais importante para a menos importante, utilizando a importância percentual como critério de ordenação.

Após isso, é gerado um ranking das variáveis por importância, destacando o valor mais influente, que corresponde à média da variável entre os clientes que compraram o título (Alvo = 1). Caso a variável seja categórica, o código converte o número de volta para seu nome original utilizando label_encoders.

Por fim, ocorre a criação do perfil do cliente ideal, filtrando apenas os clientes que compraram (Alvo = 1). Para cada variável (col): se for categórica, obtém o valor mais frequente (mode()[0]) e converte de volta para texto utilizando label_encoder; Se for numérica, calcula a média dos valores.

O objetivo dessa etapa é criar um perfil do cliente mais propenso a comprar, auxiliando na segmentação e otimização de estratégias.

VI. Experimentos

Foram realizados 3 tipos de experimentos, são eles: comparação com outros modelos, diferentes proporções de treino e teste e teste de normalização das variáveis. A fim de que, possamos avaliar o comportamento do modelo.

Nesta seção, será debatido os aspectos técnicos de cada experimento.

A. Comparação entre modelos distintos

Após o tratamento dos dados, três algoritmos diferentes são utilizados para análise: Naive Bayes, Random Forest (que utiliza múltiplas árvores de decisão para melhorar a precisão) e Support Vector Machine (que encontra um hiperplano ótimo para separar as classes). Cada um desses modelos gera previsões ligeiramente distintas, que serão analisadas

posteriormente.

Cada modelo foi treinado com os mesmos dados e avaliado por meio das seguintes métricas:

I) Precisão (Precision): Mede a qualidade das previsões positivas.

II) Recall: Avalia a capacidade do modelo de identificar corretamente os casos positivos.

III) F1-score: Representa a média harmônica entre Precisão e Recall, proporcionando um equilíbrio entre ambas as métricas.

B. Diferença de comportamento em proporções distintas de treino e teste

O experimento avalia diferentes proporções de dados para treino e teste, utilizando as seguintes divisões: 50% treino / 50% teste, 60% treino / 40% teste, 70% treino / 30% teste, 80% treino / 20% teste e 90% treino / 10% teste

Para cada proporção, os dados são divididos com train_test_split(), garantindo que o conjunto de teste sempre contenha os dados não utilizados no treinamento.

Em cada divisão, o modelo Naive Bayes é treinado com nb_model.fit(X_train, y_train). Em seguida, ele realiza previsões sobre os dados de teste (y_pred_nb = nb_model.predict(X_test)) e sua acurácia é calculada com accuracy_score(y_test, y_pred_nb).

A acurácia representa a porcentagem de previsões corretas em relação ao total de exemplos testados. Esse valor é armazenado para cada proporção de treino.

Os resultados são organizados em um DataFrame (results_df), exibidos na tela e visualizados graficamente com seaborn.lineplot(). O gráfico auxilia na identificação de como a acurácia do modelo varia conforme mais dados são utilizados no treinamento.

C. Teste de diferentes Métodos de Normalização

O objetivo do experimento é comparar três abordagens de normalização e seu impacto no desempenho do modelo Gaussian Naive Bayes (GaussianNB):

I) Sem Normalização: Os dados são utilizados no formato original.

II) StandardScaler: Subtrai a média e divide pelo desvio padrão, resultando em dados com média 0 e desvio padrão 1.

III) MinMaxScaler: Transforma os valores para um intervalo entre 0 e 1, preservando a distribuição original.

Para cada abordagem, tanto os dados de treino (X_train) quanto os de teste (X_test) passam pela mesma transformação.

Em seguida, o modelo Gaussian Naive Bayes é treinado e avaliado em cada conjunto de dados, realizando previsões sobre os dados de teste (y_pred) e medindo a acurácia com accuracy_score(), indicando a proporção de previsões corretas.

Os resultados são organizados e apresentados na forma de tabela, exibindo a acurácia para cada método de normalização, e na forma de gráfico de barras, ilustrando visualmente o impacto de cada abordagem no desempenho do modelo.

VII. ANÁLISE DOS RESULTADOS

Dividiremos esse tópico em análises de 4 subtópicos principais.

A. Implementação do Classificador Ingênuo de Bayes

Apesar da presença de várias variáveis no modelo, a variável Saldo_Conta_Corrente domina a classificação, com 99,4% de importância. Isso pode representar um problema, pois o classificador Naïve Bayes assume independência entre as variáveis, e o forte peso de uma única variável pode estar prejudicando sua capacidade de generalização. Esse fator pode explicar a acurácia relativamente baixa (60%) e o desbalanceamento na recall das classes.

Além disso, o impacto excessivo do Saldo_Conta_Corrente faz com que o perfil ideal de cliente pareça estar mais alinhado com a capacidade financeira do que com outros atributos, já que variáveis como idade, profissão e estado civil têm pouca influência na decisão. Esse comportamento

sugere que o modelo pode estar subaproveitando informações potencialmente relevantes, limitando sua performance geral.

Para reduzir a dependência excessiva do Saldo_Conta_Corrente e melhorar a acurácia do modelo, pode-se testar classificadores mais flexíveis, como Random Forest ou SVM, que não assumem independência entre as variáveis. Além disso, pode-se explorar técnicas de normalização ou seleção de atributos para equilibrar a importância das variáveis e melhorar a generalização do modelo.

B. Comparação de Modelos

O modelo Random Forest apresentou o melhor desempenho geral entre os três, com métricas em torno de 0.62 em precisão, recall e F1-score. O Naive Bayes teve resultados próximos ao Random Forest em precisão e recall, mas foi ligeiramente inferior no F1-score. Já o SVM teve o pior desempenho, o que pode estar relacionado à sua sensibilidade à escala dos dados ou à natureza específica do conjunto de dados utilizado. Concluímos, portanto, que o Random Forest é o modelo mais equilibrado entre as métricas avaliadas, enquanto o Naive Bayes permanece competitivo, com a vantagem de ser mais simples e leve computacionalmente. O SVM, por outro lado, mostrou-se menos eficaz nesse cenário específico.

C. *Diferentes proporções de treino e teste*

A acurácia do Naive Bayes manteve-se estável mesmo com o aumento gradual da proporção dos dados usados para treinamento (de 50% a 90%). Esse comportamento indica que o modelo atinge seu desempenho máximo com uma quantidade moderada de dados, característica comum a algoritmos mais simples, que não se beneficiam significativamente de grandes volumes de dados de treino.

D. Teste de normalização das variáveis

O melhor desempenho foi observado sem a aplicação de normalização, atingindo uma acurácia em torno de 0.60. As duas formas de normalização testadas resultaram em queda de desempenho, o que sugere que a distribuição original dos dados já estava adequada aos algoritmos utilizados. Isso também indica que, neste caso, a normalização pode ter prejudicado a modelagem ao distorcer relações importantes entre as variáveis.

VIII. CONCLUSÕES E DISCUSSÕES

Este projeto teve como objetivo analisar o perfil dos clientes que adquiriram um título financeiro durante a campanha de marketing do banco, utilizando técnicas de estatística e aprendizado de máquina para identificar padrões relevantes. A partir da análise exploratória dos dados, foram levantados insights estratégicos que podem contribuir para a otimização de campanhas futuras.

Ao analisar os resultados, percebe-se que fatores como saldo bancário, endividamento, idade, nível de formação, estado civil e profissão influenciam diretamente na decisão de compra do título. Clientes com saldo de até R$ 15.000, não-devedores, sem hipoteca ou empréstimos e pertencentes às profissões de administração, mostraram maior propensão à compra. Além disso, a faixa etária entre 25 e 40 anos apresentou uma taxa significativa de conversão, sugerindo que esse público tem maior interesse em investimentos financeiros. Também é válido ressaltar a importância das ligações, as quais, se usadas estrategicamente, com uma boa adequação aos clientes, são uma ótima ferramenta de conversão.

A implementação do classificador Naive Bayes permitiu a construção de um modelo preditivo para identificar clientes com maior probabilidade de compra. No entanto, observou-se que a variável Saldo_Conta_Corrente dominou a classificação, com 99,4% de importância, indicando um possível viés no modelo. Esse fator impactou sua acurácia, que ficou em aproximadamente 60%, e reduziu a capacidade do modelo de generalizar para diferentes perfis de clientes. Para contornar essa limitação, sugere-se explorar modelos mais flexíveis, como Random Forest, que apresentou melhor desempenho geral, ou realizar ajustes no pré-processamento dos dados, como a normalização ou a seleção de atributos mais equilibrada.

A comparação entre diferentes modelos de aprendizado de máquina mostrou que o Random Forest obteve os melhores resultados em termos de precisão, recall e F1-score, tornando-se uma alternativa viável para futuras aplicações. Já o SVM apresentou o pior desempenho, possivelmente devido à sensibilidade à escala dos dados. Além disso, o

experimento com diferentes proporções de treino e teste revelou que o Naive Bayes não obteve ganhos significativos ao aumentar a quantidade de dados de treinamento, indicando que sua performance é bastante estável.

Por fim, chegamos aos testes com normalização, os quais demonstraram que a acurácia do modelo foi melhor sem a aplicação de transformações nos dados, sugerindo que a distribuição original já era adequada ao problema. Esse resultado reforça a importância de avaliar cuidadosamente a necessidade de normalização em diferentes contextos, com o intuito de termos a melhor perfomance e acurácea do modelo.

Diante dos resultados obtidos, recomenda-se que futuras campanhas de marketing foquem nos segmentos de clientes identificados como mais propensos à compra, além de considerar abordagens mais personalizadas com base em seus perfis específicos. No aspecto técnico, a exploração de modelos mais robustos (e.g. Random Foreste e ANN's) e a otimização do pré-processamento dos dados podem contribuir para aprimorar a capacidade preditiva das análises, aumentando a eficiência das estratégias comerciais do banco e garantindo resultados ainda melhores.

IX. *DIVISÃO DE TRABALHO ENTRE A EQUIPE*

Em relação à Proposta do Projeto, a primeira parte do projeto, após entrarmos em consenso quanto ao tema e database do nosso projeto, dividimos o trabalho da seguinte forma: o integrante Álvaro Cavalcante se responsabilizou pela parte de título, autores e objetivos; os integrantes Júlio César e Vinícius Sousa se responsabilizaram em conjunto pela parte de metodologia e referências; por fim, o integrante Luiz Lustosa se responsabilizou pela parte de justificativa e montagem do cronograma seguido pela equipe.

Já em relação à segunda parte do projeto, no que diz respeito ao Relatório Final, o integrante Vinícius Sousa se responsabilizou pela análise do database, pela implementação da análise exploratória de dados e pelos tópicos dessas duas partes do relatório final do projeto. O integrante Luiz Lustosa se responsabilizou pela implementação do Classificador Ingênuo de Bayes e pelos experimentos e pelos tópicos dessas duas partes do relatório final do projeto. O integrante Júlio César ajudou Luiz na implementação dos algoritmos e também se responsabilizou pelo tópico de Análise dos Resultados no relatório final. Por fim, o integrante Álvaro Cavalcante se responsabilizou pela análise geral do projeto e pelo tópico de Conclusão e Discussões no relatório final.

Por fim, dividimos a realização dos slides da apresentação e a apresentação em si de modo que cada um se responsabilizou pelas suas respectivas partes do relatório final.

X. REFERÊNCIAS

[1]https://frons.com.br/blog/engenharia/teorema-de bayes-o-que-e-aplicacoes-e-como-usar-i-cae treinamentos/

[2] https://pt.wikipedia.org/wiki/Teorema_de_Bayes

[3] https://blog.somostera.com/data-science/naive bayes#:~:text=O%20classificador%20Naive%20Bayes% 20%C3%A9,features%20s%C3%A3o%20independentes %20entre%20si.

[4] https://www.ibm.com/br-pt/topics/naive-bayes

[5]https://docs.google.com/spreadsheets/d/1O6Wft _JJbr cnFdfgEH9JKezzsdIVvILC/edit?gid=1230927564# gid= 1230927564

[6]https://medium.com/@nicolasfaleiros/classificadornaive-bayes-e-o-teorema-de-bayes-84a76c177 93