Simpósio de Excelência em Gestão e Tecnologia
XXIISEGeT
Economia Circular: Desafios para construir uma sociedade sustentável
Realização:

Aprendizado de máquinas na previsão e análise de rotatividade de clientes

Fernando Akira Namioka
fernando.akira@unesp.br
UNESP

Jose Roberto Dale Luche
dale.luche@unesp.br
UNESP

Claudia Regina de Freitas
claudia.freitas@unesp.br
UNESP

Resumo: A análise de dados desempenha um papel cada vez mais estratégico no setor bancário, especialmente no que diz respeito à previsão do churn de clientes. Nesse cenário, o uso de técnicas de aprendizado de máquina se mostra promissor, fornecendo recursos avançados para identificar e reduzir a evasão de clientes. Este estudo, de natureza aplicada e quantitativa, teve como objetivo analisar a performance de diferentes algoritmos de aprendizado de máquina na previsão do churn bancário. Foi conduzido um estudo de caso exploratório, com dados coletados e processados em ambiente laboratorial controlado. Foram comparados diversos modelos, incluindo Naive Bayes, SVM, Regressão Logística, KNN, Árvore de Decisão, Random Forest e XGBoost, avaliados a partir de métricas como acurácia, precisão e recall, utilizando a técnica de validação cruzada k-fold. Os resultados revelaram que o modelo XGBoost apresentou o melhor desempenho, consolidando-se como a alternativa mais eficaz para prever o churn de clientes bancários. Esses achados contribuem para aprimorar estratégias de retenção de clientes no setor financeiro.

Palavras Chave: Aprendizado de Máqui - Previsão de churn - Análise de Dados - Técnicas de Classifi

1. INTRODUÇÃO

Em um mercado altamente competitivo, manter os clientes se tornou uma prioridade essencial para os bancos. A perda de clientes, ou churn, é um grande problema para essas instituições, pois perder clientes afeta negativamente tanto a receita quanto a estabilidade financeira. Pesquisas mostram que conquistar um novo cliente pode ser até cinco vezes mais caro do que manter um já existente, destacando a necessidade de estratégias eficazes para manter os clientes (Geiler et al., 2022; Reichheld & Schefter, 2000).

Nesse cenário, o aprendizado de máquinas (Machine Learning) se destaca como uma ferramenta poderosa para prever e analisar o churn de clientes. Com a capacidade de analisar grandes volumes de dados, identificar padrões complexos e prever comportamentos futuros com alta precisão, os algoritmos de aprendizado de máquinas permitem que os bancos antecipem quais clientes estão mais propensos a sair. Com essas previsões, os bancos podem tomar medidas proativas para melhorar a satisfação e a fidelidade dos clientes, reduzindo a taxa de churn (Huang et al., 2017; Verbeke et al., 2012).

O aprendizado de máquinas tem sido eficaz em diversas áreas do setor financeiro, como na detecção de fraudes, análise de crédito e personalização de serviços (Ngai et al., 2011). Aplicar essas técnicas na previsão de churn não só melhora a capacidade de previsão, mas também permite uma segmentação mais precisa dos clientes. Isso facilita a implementação de estratégias de marketing direcionadas (Lemmens & Croux, 2006), permitindo que os bancos identifiquem rapidamente os clientes em risco e adotem ações específicas para mantê-los, como ofertas personalizadas ou melhorias no atendimento ao cliente.

Além disso, a previsão de churn utilizando aprendizado de máquinas pode ser integrada com sistemas de CRM (Customer Relationship Management), proporcionando uma visão completa do cliente e permitindo uma abordagem mais personalizada e eficaz (Hadden et al., 2007). Estudos mostram que integrar sistemas de CRM com técnicas avançadas de análise de dados pode aumentar significativamente a retenção de clientes e melhorar a lucratividade da instituição (Ngai et al., 2009).

Este estudo tem como objetivo avaliar a eficácia de diferentes técnicas de aprendizado de máquina na previsão de churn de clientes bancários. Para isso, foi utilizada uma base de dados pública e sintética, aplicando e comparando diversos modelos de aprendizado supervisionado, como Naive Bayes, SVM, Regressão Logística, KNN, Árvore de Decisão, Random Forest e XGBoost. Cada modelo foi avaliado com base em métricas de desempenho como acurácia, precisão, recall e F1-score, além da análise de falsos positivos e negativos.

Este trabalho está organizado da seguinte forma: na Seção 2, são apresentados os conceitos fundamentais de aprendizado de máquinas, técnicas de tratamento de dados, classificação e validação de modelos. A Seção 3 descreve a base de dados utilizada, as ferramentas empregadas e os procedimentos adotados para a realização do estudo. A Seção 4 detalha a seleção e exploração da base de dados, o tratamento dos dados, a análise de correlação, a definição dos parâmetros dos modelos e a execução dos experimentos. Na Seção 5, os resultados são apresentados e discutidos, comparando a eficácia dos diferentes modelos de aprendizado de máquina na previsão de churn. Finalmente, a Seção 6 resume os principais achados do estudo, suas implicações práticas e sugestões para trabalhos futuros.

2. REFERENCIAL TEÓRICO

Nesta seção, são explorados os conceitos fundamentais de aprendizado de máquinas, tratamento de dados, técnicas de classificação e a importância da matriz de confusão na avaliação de modelos de classificação.

2.1 APRENDIZADO DE MÁQUINAS

A Inteligência Artificial (IA) é uma área da ciência da computação dedicada à criação de sistemas que exibem comportamentos inteligentes. O aprendizado de máquinas (Machine Learning, ML) é um subcampo da IA que se concentra no desenvolvimento de algoritmos capazes de aprender com dados e fazer previsões ou tomar decisões sem serem explicitamente programados para isso. Segundo Mitchell (1997), aprendizado de máquinas é "o estudo de algoritmos que melhoram automaticamente através da experiência".

O aprendizado de máquinas pode ser dividido em três categorias principais: supervisionado, não supervisionado e reforço. No aprendizado supervisionado os algoritmos são treinados com um conjunto de dados rotulado, aprendendo a mapear entradas para saídas. No aprendizado não supervisionado, os algoritmos analisam e agrupam dados não rotulados com base em similaridades e diferenças. No aprendizado por reforço os algoritmos aprendem a tomar decisões por tentativa e erro, recebendo feedback sobre suas ações (Alpaydin, 2020).

Com aplicações em diversos setores, o aprendizado de máquinas tem transformado processos e práticas em negócios, saúde, educação, finanças, entre outros. Na área financeira, é utilizado para modelar riscos, otimizar portfólios, detectar fraudes e melhorar a experiência do cliente. No setor de saúde, contribui para o diagnóstico precoce de doenças e a personalização de tratamentos. No varejo, aprimora a logística, a gestão de estoques e a personalização de recomendações de produtos (Jordan e Mitchell, 2015). A versatilidade do aprendizado de máquinas demonstra sua importância e impacto potencial em vários aspectos da sociedade e da economia.

A Figura 1 ilustra de forma esquemática as diversas categorias de aprendizado de máquina e suas aplicações correspondentes. Em cada contexto de aplicação, os algoritmos de aprendizado de máquina necessitam de uma abordagem analítica específica. Compreender o problema em questão é fundamental para selecionar e aplicar a técnica mais apropriada e eficiente, garantindo a otimização do desempenho e a precisão dos resultados.

Figura 1: Tipos de Aprendizagem de Máquina
Fonte: Maia (2020)

Apesar dos avanços significativos, o aprendizado de máquinas enfrenta desafios como o risco de *overfitting* (tendência do modelo), a necessidade de grandes quantidades de dados rotulados de qualidade e questões de interpretabilidade dos modelos. A interpretabilidade é especialmente importante em setores regulados, como o bancário, onde as decisões dos modelos precisam ser explicáveis (Ribeiro et al., 2016).

2.2 TRATAMENTO DOS DADOS

O tratamento de dados é um processo crítico no desenvolvimento de modelos de aprendizado de máquinas. Esta etapa envolve a identificação e correção de erros, inconsistências e imprecisões nos dados. A qualidade dos dados influencia diretamente a eficácia dos modelos de ML, impactando desde a acurácia das previsões até a confiança dos usuários nos resultados gerados (Yen, 2023).

Existem diversas técnicas de tratamento de dados, com o objetivo de melhorar a qualidade dos dados, aprimorar o desempenho dos modelos de ML, economizar tempo e recursos, e aumentar a confiança nos resultados. Algumas dessas técnicas incluem:

- Manipulação de Valores Ausentes: Técnicas como imputação de média ou uso de algoritmos para preencher valores ausentes.
- Tratamento de Dados Duplicados: Identificação e remoção de registros duplicados para evitar redundância.
- Correção de Formatos e Padronização: Ajuste de formatos de dados (como datas e números) e padronização de categorias e escalas.
- Detecção e Remoção de Outliers: Utilização de técnicas estatísticas ou modelos de ML para identificar e tratar outliers.

- Redução de Dimensionalidade: Técnicas como Análise de Componentes Principais (PCA) para reduzir a complexidade dos dados.

2.3 SEPARAÇÃO E VALIDAÇÃO DOS DADOS DE TREINO E TESTE

Para desenvolver um modelo de aprendizado de máquina, é fundamental segmentar o conjunto de dados (dataset) em duas partes distintas: dados de treino e dados de teste. Os dados de treino são usados para construir e refinar o modelo, permitindo que ele aprenda e se adapte às características dos dados. Já os dados de teste, que devem ser independentes dos dados de treino, têm a função de avaliar a eficácia e a precisão do modelo já treinado. Esta separação é essencial para garantir que o modelo seja capaz de generalizar suas previsões ou decisões para novos dados que não foram vistos durante o processo de treinamento. Uma separação comum do dataset é a divisão em 30% em dados de teste e 70% em dados de treino (Leite, 2020).

2.3.1 MÉTODO DE VALIDAÇÃO CRUZADA K-FOLD

O k-fold cross-validation é uma técnica robusta para validar a eficácia de modelos de aprendizado de máquinas. Nesta abordagem, o conjunto de dados é dividido em 'k' subconjuntos (ou 'folds'). O modelo é treinado e testado 'k' vezes, utilizando cada vez um fold diferente como conjunto de teste e os demais para treinamento.

Kohavi (1995) demonstra que essa técnica reduz a variância associada à avaliação do modelo, fornecendo uma estimativa mais confiável de seu desempenho. O método de validação cruzada k-fold é conveniente pois cada observação no conjunto de dados é utilizada tanto para treinamento quanto para teste, garantindo uma utilização eficiente dos dados. No entanto, escolher o número apropriado de 'folds' é fator determinante. Um número excessivo pode aumentar a carga de processamento computacional, enquanto um número insuficiente pode não refletir adequadamente a variabilidade dos dados (Molinaro et al., 2005).

2.4 APRENDIZADO DE MÁQUINAS SUPERVISIONADO

O aprendizado de máquinas supervisionado é uma abordagem central na inteligência artificial onde o modelo é treinado em dados rotulados para realizar tarefas específicas. Existem dois tipos principais de tarefas em aprendizado supervisionado: classificação e regressão. Na classificação, o objetivo é categorizar dados em classes pré-definidas, como na identificação de clientes propensos a churn em um banco (Miric et al., 2023). Já na regressão, o foco é prever uma variável contínua, como o valor de uma casa ou a temperatura de um dia (Panigrahi et al., 2023).

Ambos os tipos requerem um conjunto de dados rotulado, onde as entradas (features) estão diretamente associadas a saídas conhecidas (labels ou targets). Este processo possibilita que os modelos aprendam a relação entre entradas e saídas e façam previsões precisas para novos dados (Sem et al, 2020). A eficácia de diferentes algoritmos de aprendizado supervisionado em tarefas de classificação e regressão tem sido amplamente estudada, demonstrando a versatilidade e a capacidade de adaptação desses modelos a diversos tipos de problemas (Caruana & Niculescu-Mizil, 2006).

2.4.1 MODELOS DE APRENDIZADO POR CLASSIFICAÇÃO

Diversos modelos são empregados em tarefas de classificação no aprendizado supervisionado, incluindo:

- Naive Bayes: Um modelo baseado em probabilidade que aplica o teorema de Bayes com a suposição de independência entre as características. É eficaz em grandes conjuntos de

dados e frequentemente utilizado em classificação de texto (Amin et al., 2023; Rish, 2001).

- SVM (Support Vector Machine): Utiliza hiperplanos em um espaço de alta dimensão para classificar dados. É eficaz para dados de dimensão elevada e conhecido por sua robustez (Cortes & Vapnik, 1995; Quek et al., 2023).
- Regressão Logística: Um modelo estatístico que estima a probabilidade de uma variável dependente binária. É amplamente utilizado devido à sua simplicidade e eficácia em problemas de classificação binária (Tran et al., 2023; Walker & Duncan, 1967).
- KNN (K-Nearest Neighbors): Um método que classifica cada ponto de dados com base na maioria dos votos de seus 'k' vizinhos mais próximos. É simples e eficaz, especialmente em conjuntos de dados onde as relações espaciais são importantes (Akakpo et al., 2023; Cover & Hart, 1967).
- Árvore de Decisão: Modelo que utiliza uma estrutura de árvore para tomar decisões baseadas em regras simples inferidas dos dados de treino. (Quinlan, 1986; Usman-Hamza et al., 2024).
- Random Forest: Um conjunto de árvores de decisão que melhora a estabilidade e a precisão da classificação, reduzindo o risco de overfitting (Al-Sultan & Al-Baltah, 2024; Breiman, 2001).
- XGBoost (Extreme Gradient Boosting): Um algoritmo de boosting que otimiza modelos de árvores de decisão de forma gradativa, sendo eficaz em muitas tarefas de classificação com destaque para sua velocidade e performance (Chen & Guestrin, 2016; Liu et al., 2024).

2.5 MATRIZ DE CONFUSÃO

A matriz de confusão é uma ferramenta poderosa no aprendizado de máquinas para avaliar o desempenho de algoritmos de classificação. Ela permite visualizar a performance do modelo ao comparar os valores reais (ou observados) com os valores previstos pelo modelo. Fawcett (2006) descreve a matriz de confusão como um meio de entender não apenas a eficácia geral do modelo, mas também as formas específicas em que ele pode estar acertando ou errando.

		Cliente deixou o Banco?	
		Real	
		sim	não
Previsão	sim	Verdadeiro Positivo	Falso Negativo
	não	Falso Positivo	Verdadeiro Negativo

Figura 2: Matriz de Confusão

Uma matriz de confusão (ver Figura 2) é composta por quatro elementos fundamentais: verdadeiros positivos (VP), falsos positivos (FP), verdadeiros negativos (VN) e falsos negativos (FN) (Figura 2). Stehman (1997) explica que VP e VN representam as classificações corretas, enquanto FP e FN são os erros de classificação.

A partir desses valores, é possível calcular várias métricas de desempenho, como:

- Precisão: $= \frac{VP}{VP+FP}$ proporção de previsões positivas corretas em relação ao total de previsões positivas feitas pelo modelo.

- Recall (Sensibilidade): $= \frac{VP}{VP+FN}$ proporção de verdadeiros positivos identificados em relação ao total de casos positivos reais.
- Acurácia: $= \frac{VP+VN}{VP+FN+FP+VN}$ medida mais direta e compreensível do desempenho de um modelo. Calcula a proporção de previsões corretas em relação ao total de previsões.

A aplicação da matriz de confusão vai além da mera avaliação quantitativa. Ela fornece ideias sobre o tipo de erro que um modelo está cometendo, o que é importante para ajustes e melhorias. Powers (2011) argumenta que a compreensão detalhada fornecida pela matriz de confusão é essencial para aprimorar modelos de classificação, especialmente em contextos onde diferentes tipos de erros têm consequências distintas.

3. MATERIAIS E MÉTODOS

Nesta seção, descrevem-se a base de dados utilizada, as ferramentas empregadas e os procedimentos adotados para a realização do estudo.

3.1 BASE DE DADOS

Para o estudo da previsão de churn de clientes bancários, foi selecionada uma base de dados da plataforma Kaggle, proposta por Shruti Iyer (2023) contendo informações de 10.000 clientes. Esta base foi escolhida devido à sua abrangência e relevância para o fenômeno em estudo. Os dados são sintéticos, mas representativos de situações reais, oferecendo um panorama detalhado do comportamento e das características dos clientes bancários, elementos fundamentais para a análise de churn.

A base de dados compreende uma série de variáveis críticas para a compreensão do churn de clientes em instituições bancárias. Cada registro na base representa um cliente e inclui informações como idade, saldo, pontuação de crédito, renda, gênero, país, posse de cartão de crédito, status de cliente ativo, quantidade de produtos adquiridos, sobrenome do cliente, tempo de permanência no banco e a rotulação de churn.

3.2 Tratamento da Base de Dados

3.2.1 Importação

A importação dos dados foi a primeira etapa da análise. A base de dados foi inicialmente importada como um arquivo CSV (Valores Separados por Vírgula) em uma pasta do Google Drive. Esta abordagem ofereceu uma integração direta e eficiente com as ferramentas do Google utilizadas no projeto, como o Google Colab.

A utilização do Google Drive como repositório de dados apresentou vantagens em relação à acessibilidade e segurança dos dados. O armazenamento em nuvem permitiu que os dados estivessem disponíveis e seguros, acessíveis a partir de qualquer local com conexão à internet. Após o armazenamento dos dados, a base foi lida e manipulada utilizando a biblioteca Pandas no Python por meio do ambiente de desenvolvimento Google Colab.

3.2.2 Pré-processamento

A primeira etapa de pré-processamento envolveu a transformação de variáveis categóricas nominais em variáveis categóricas ordinais. Especificamente, as colunas referentes

ao gênero e ao país dos clientes foram convertidas em números. Esta transformação facilitou o uso em modelos de aprendizado de máquina que requerem dados de entrada numéricos. Por exemplo, o gênero foi codificado como 1 para feminino e 0 para masculino, enquanto os países foram representados com 1 para França, 2 para Alemanha e 3 para Espanha.

3.2.3 Bases de Treino e Teste

Após o pré-processamento, a base de dados foi dividida em conjuntos de treino e teste. Adotou-se a proporção de 70% dos dados para a base de treino e 30% para a base de teste. Esta distribuição busca um equilíbrio entre ter dados suficientes para treinar os modelos adequadamente e possuir uma quantidade significativa de dados não vistos para testar e validar o desempenho dos modelos. A divisão dos dados foi realizada de forma aleatória para garantir que ambos os conjuntos fossem representativos do conjunto de dados total.

4. EXPERIMENTOS

Nesta seção, detalharemos os experimentos realizados para prever o churn de clientes bancários utilizando diferentes técnicas de aprendizado de máquina. Descreveremos a seleção e exploração da base de dados, o tratamento dos dados, a análise de correlação, a definição dos parâmetros dos modelos, e finalmente, a execução dos experimentos.

4.1 ANÁLISE DE CORRELAÇÃO

Para analisar a relevância dos atributos previsores em relação ao resultado de churn, bem como a inter-relação entre os próprios atributos previsores, foi realizada uma análise de correlação. Esta análise envolveu a construção de uma matriz de correlação. O resultado desta análise é apresentado na Tabela 1.

Tabela 1: Matriz de Correlação

	País	**Sexo**	**Idade**	**Fidelidade**	**Saldo**	**Qtd de Produtos**	**Possui Cartão**	**Membro Ativo**	**Renda**	**Churn**
País	1,000	-0,005	0,023	0,004	0,063	0,004	-0,009	0,007	-0,005	0,036
Sexo	-0,005	1,000	0,028	-0,015	-0,007	0,022	-0,006	-0,023	0,011	0,107
Idade	0,023	0,028	1,000	-0,010	0,022	-0,031	-0,012	0,085	-0,015	0,285
Fidelidade	0,004	-0,015	-0,010	1,000	-0,017	0,013	0,023	-0,028	-0,015	-0,014
Saldo	0,063	-0,007	0,022	-0,017	1,000	-0,276	-0,011	0,006	0,006	0,106
Qtd.	0,004	0,022	-0,031	0,013	-0,276	1,000	0,003	0,014	-0,014	-0,048
Cartão	-0,009	-0,006	-0,012	0,023	-0,011	0,003	1,000	-0,012	0,010	-0,007
Ativo	0,007	-0,023	0,085	-0,028	0,006	0,014	-0,012	1,000	-0,006	-0,156
Renda	-0,005	0,011	-0,015	-0,015	0,006	-0,014	0,010	-0,006	1,000	0,003
Churn	0,036	0,107	0,285	-0,014	0,106	-0,048	-0,007	-0,156	0,003	1,000

A matriz de correlação revelou alguns insights importantes:

- Relativa Importância dos Atributos: Alguns atributos, como idade e status de cliente ativo, mostraram-se mais influentes na predição do churn em comparação com outros, como país e gênero.

- Complexidade na Predição de Churn: O resultado do churn de clientes parece ser uma composição de todas as variáveis preditoras, indicando que a decisão de um cliente de deixar o banco é multifatorial e não pode ser atribuída a um único fator.

4.2 PARÂMETROS DOS MODELOS

Os parâmetros são configurações que definem como o modelo irá operar e aprender a partir dos dados. A definição dos parâmetros dos modelos impacta diretamente a acurácia e a performance de cada modelo. Neste estudo, optou-se por utilizar os parâmetros padrões para cada um dos modelos de aprendizado de máquina aplicados. Os modelos utilizados foram Naive Bayes, SVM, Regressão Logística, KNN, Árvore de Decisão, Random Forest e XGBoost.

Aplicou-se também a técnica de validação cruzada k-fold em cada modelo de aprendizado de máquina, com a execução de 30 folds. Este método envolveu a repetição do processo de treinamento e teste 30 vezes, utilizando diferentes subconjuntos de dados a cada vez. A acurácia final reportada para cada modelo é a média das acurácias obtidas nas 30 execuções, proporcionando uma medida confiável e abrangente do desempenho de cada modelo na previsão do churn de clientes bancários.

4.3 EXECUÇÃO DOS EXPERIMENTOS

Para cada modelo de aprendizado de máquina, os seguintes passos foram executados:

1. Treinamento do Modelo: Cada modelo foi treinado utilizando a base de dados de treino.
2. Validação Cruzada: A técnica de validação cruzada k-fold foi aplicada para avaliar a performance do modelo.
3. Avaliação do Desempenho: As métricas de acurácia, precisão, recall e F1-score foram calculadas para cada modelo, proporcionando uma avaliação abrangente de seu desempenho.

Para a execução dos experimentos, foi utilizado o pseudocódigo na Figura 3.

```
# Importação das bibliotecas necessárias
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from xgboost import XGBClassifier

# Carregamento da base de dados
dados = pd.read_csv('caminho/para/dados.csv')

# Pré-processamento dos dados
dados['Genero'] = dados['Genero'].map({'Feminino': 1, 'Masculino': 0})
dados['Pais'] = dados['Pais'].map({'Franca': 1, 'Alemanha': 2, 'Espanha': 3})

# Divisão em treino e teste
X = dados.drop('Churn', axis=1)
y = dados['Churn']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Definição dos modelos
modelos = {
'Naive Bayes': GaussianNB(),
'SVM': SVC(),
'Logistic Regression': LogisticRegression(),
'KNN': KNeighborsClassifier(),
'Decision Tree': DecisionTreeClassifier(),
'Random Forest': RandomForestClassifier(),
'XGBoost': XGBClassifier()
}

# Execução dos experimentos
resultados = {}
for nome, modelo in modelos.items():
modelo.fit(X_train, y_train)
acuracia = cross_val_score(modelo, X_train, y_train, cv=30, scoring='accuracy').mean()
resultados[nome] = acuracia

# Exibição dos resultados
print(resultados)
```

Figura 3: Pseudocódigo para execução dos experimentos

5. RESULTADOS E DISCUSSÃO

Nesta seção, são apresentados e discutidos os resultados obtidos a partir da aplicação de diferentes modelos de aprendizado de máquina na previsão de churn de clientes bancários. A avaliação dos modelos foi baseada em métricas de desempenho, como acurácia, precisão, recall e F1-score, além da análise de falsos positivos e negativos.

5.1 MATRIZES DE CONFUSÃO

As matrizes de confusão foram ferramentas-chave na avaliação dos modelos de aprendizado de máquina neste estudo, fornecendo informações detalhadas sobre o desempenho dos modelos em termos de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

O conjunto de teste, composto por 3.000 clientes, registra 2.379 retenções (clientes que permaneceram no banco) e 621 churns (clientes que saíram do banco). Com base nesses dados, cada modelo de aprendizado de máquina testado neste estudo apresentou na Tabela 2 os resultados na identificação desses churns e retenções.

Tabela 2: Resultados

Métrica/Modelo	Naive Bayes	SVM	Logistic Regression	KNN	Decision Tree	Random Forest	XGBoost
Acurácia Média	0,8142	0,8552	0,8100	0,8318	0,8395	0,8413	0,8594
Falsos Positivos	215	76	101	116	299	7	93
Verdadeiros Positivos	250	277	142	226	369	128	311
Verdadeiros Negativos	2.164	2.303	2.278	2.263	2.080	2.372	2.286
Falsos Negativos	371	344	479	395	252	493	310

A acurácia final de cada modelo, calculada como a média das acurácias obtidas nos 30 folds da validação cruzada, é um indicador importante do desempenho de cada modelo.

A análise dos falsos positivos fornece uma visão adicional sobre os erros cometidos por cada modelo, ajudando a entender melhor suas limitações e pontos fortes.

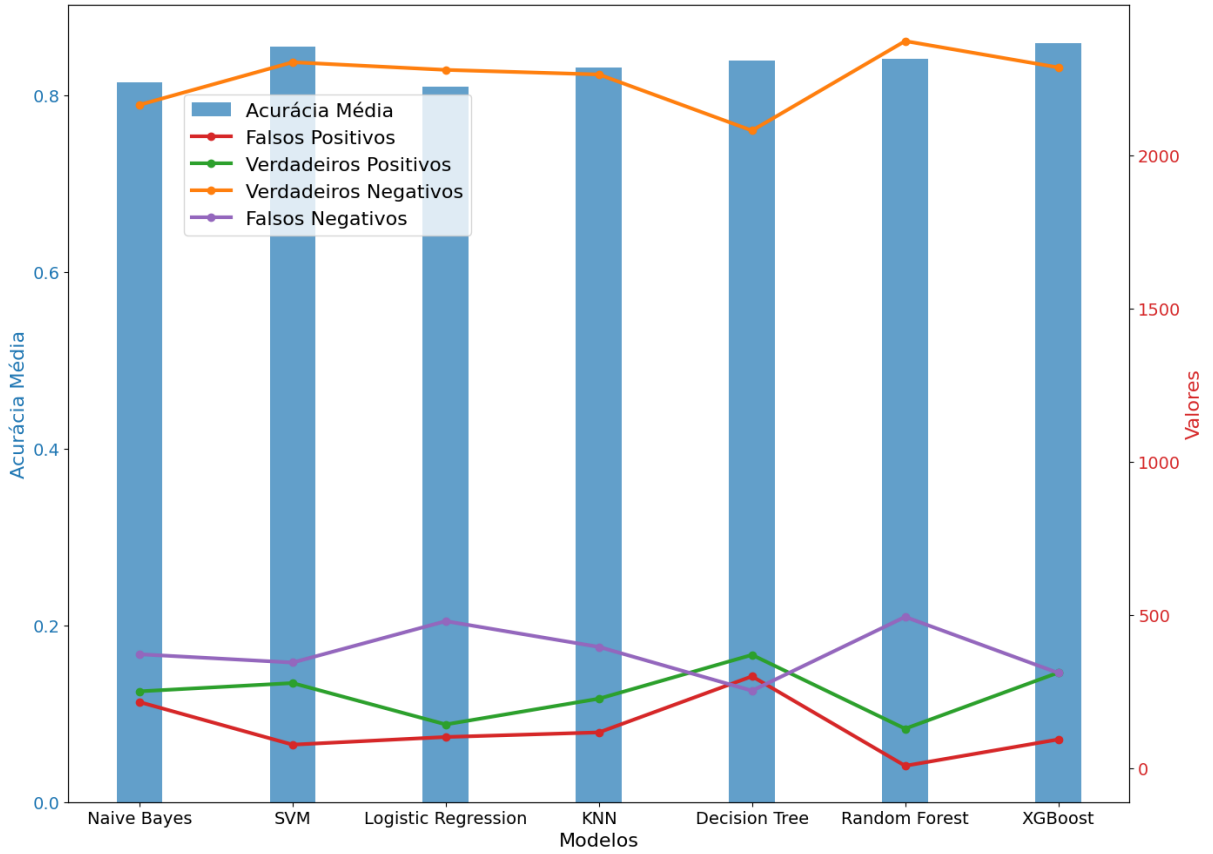

Figura 4: Gráfico de Desempenho de Modelos na Previsão de churn

O gráfico apresenta uma comparação detalhada do desempenho de diferentes modelos de aprendizado de máquina na previsão de churn de clientes bancários. No eixo x, estão listados os sete modelos analisados: Naive Bayes, SVM, Regressão Logística, KNN, Árvore de Decisão, Random Forest e XGBoost. A acurácia média de cada modelo é representada por barras azuis, com os valores exibidos no eixo y esquerdo. Além disso, as linhas coloridas indicam as contagens de falsos positivos, verdadeiros positivos, verdadeiros negativos e falsos negativos para cada modelo, conforme indicado na legenda. As linhas vermelhas, verdes, laranjas e roxas representam, respectivamente, os falsos positivos, verdadeiros positivos, verdadeiros negativos e falsos negativos, com os valores correspondentes no eixo y direito. Por meio do gráfico, pode-se confirmar mais uma vez que o modelo XGBoost apresenta a maior acurácia média, enquanto o modelo Random Forest tem a menor quantidade de falsos positivos, destacando-se entre os modelos analisados.

5.2 AVALIAÇÃO DOS RESULTADOS

A análise dos resultados revela que, embora todos os modelos de aprendizado de máquina testados tenham apresentado níveis de acurácia relativamente bons, existem diferenças significativas entre eles:

- XGBoost destacou-se como o modelo mais eficaz para a previsão de churn de clientes bancários, apresentando a maior acurácia média (0,8594). Sua capacidade de gerir a complexidade e as não linearidades dos dados contribuiu significativamente para seu desempenho superior.

- SVM e Random Forest também apresentaram desempenhos sólidos, com acurácias médias de 0,8552 e 0,8413, respectivamente. O Random Forest teve o menor número de falsos positivos (7), indicando uma excelente capacidade de generalização.
- Decision Tree apresentou o maior número de falsos positivos (299), sugerindo uma maior tendência a overfitting.
- Logistic Regression apresentou a menor acurácia média (0,8100), possivelmente devido à sua natureza linear, que pode não capturar adequadamente as complexidades dos dados de churn bancário.

Os resultados indicam que técnicas avançadas de aprendizado de máquina, como XGBoost, podem oferecer melhorias significativas na previsão de churn de clientes bancários. A escolha do modelo adequado é importante e deve considerar não apenas a acurácia, mas também outros fatores, como a interpretação dos resultados e a capacidade de generalização.

A análise detalhada dos falsos positivos e negativos fornece insights valiosos sobre os tipos de erros cometidos pelos modelos, permitindo ajustes e melhorias. A capacidade de prever com precisão o churn de clientes pode ajudar instituições bancárias a desenvolver estratégias mais eficazes de retenção de clientes, personalizar serviços e melhorar a satisfação do cliente.

6. CONCLUSÕES

Os resultados demonstram a eficácia de diferentes técnicas de aprendizado de máquina na previsão de churn de clientes bancários. A análise comparativa revelou que os modelos de Random Forest, XGBoost e SVM apresentaram desempenho superior em termos de acurácia, precisão e recall.

Os achados corroboram com o estudo realizado por Geiler et al. (2022), onde foi feita uma análise abrangente de métodos de aprendizado de máquina para previsão de churn. O estudo de Geiler et al. avaliou o comportamento de onze métodos de aprendizado supervisionado e semi-supervisionado em conjunto com sete abordagens de amostragem em dezesseis conjuntos de dados publicamente disponíveis relacionados ao churn. Eles concluíram que técnicas como Random Forest e XGBoost, combinadas com métodos de reamostragem, são altamente eficazes na previsão de churn, especialmente em cenários com dados desbalanceados. Essa conclusão é consistente com os resultados obtidos em nosso estudo, que também destacou a eficácia de Random Forest e XGBoost.

Além disso, Geiler et al. enfatizam a importância de uma abordagem de ensemble para melhorar a robustez e a precisão das previsões de churn. Em nosso estudo, observamos que a combinação de técnicas de reamostragem com modelos de ensemble, como Random Forest e XGBoost, resultou em melhorias significativas no desempenho preditivo, confirmando as recomendações práticas propostas por Geiler et al. (2022).

Os resultados deste estudo oferecem implicações práticas valiosas para instituições financeiras que buscam melhorar suas estratégias de retenção de clientes. A implementação de modelos de aprendizado de máquina para prever churn permite uma abordagem proativa na identificação e retenção de clientes em risco, resultando em benefícios econômicos significativos devido à redução na taxa de rotatividade de clientes.

Para futuras pesquisas, sugere-se a exploração de técnicas avançadas de deep learning e a integração de sistemas de CRM com métodos de aprendizado de máquina para personalização mais eficaz das estratégias de retenção de clientes. Além disso, a aplicação

dessas técnicas em diferentes setores, além do bancário, pode fornecer insights adicionais sobre a generalização e adaptabilidade dos modelos preditivos de churn.

7. REFERÊNCIAS

AL-SULTAN, S. Y.; AL-BALTAH, I. A. An Improved Random Forest Algorithm (ERFA) Utilizing an Unbalanced and Balanced Dataset to Predict Customer Churn in the Banking Sector. IEEE Access, 2024.

AMIN, A.; ADNAN, A.; ANWAR, S. An adaptive learning approach for customer churn prediction in the telecommunication industry using evolutionary computation and Naïve Bayes. Applied Soft Computing, v. 137, 2023, p. 110103.

BEACH, R.; MUHLEMANN, A. P.; PRICE, D. H. R.; PATERSON, A.; SHARP, J. A. A review of manufacturing Flexibility. European Journal of Operational Research, v. 122, 2000, pp. 41-57.

BREIMAN, L. Random Forests. Machine Learning, v. 45, n. 1, 2001, pp. 5-32.

CARUANA, R.; NICULESCU-MIZIL, A. An Empirical Comparison of Supervised Learning Algorithms. Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 161-168.

CHEN, T.; GUESTRIN, C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785-794.

COVER, T. M.; HART, P. E. Nearest neighbor pattern classification. IEEE Transactions on Information Theory, v. 13, n. 1, 1967, pp. 21-27.

CORTES, C.; VAPNIK, V. Support-Vector Networks. Machine Learning, v. 20, n. 3, 1995, pp. 273-297.

FAWCETT, T. An Introduction to ROC Analysis. Pattern Recognition Letters, v. 27, 2006, pp. 861-874.

GEILER, L.; AFFELDT, S.; NADIF, M. A survey on machine learning methods for churn prediction. International Journal of Data Science and Analytics, v. 14, n. 3, 2022, pp. 217-242.

HADDEN, J.; TIWARI, A.; ROY, R.; RUTA, D. Computer assisted customer churn management: State-of-the-art and future trends. Computers & Operations Research, v. 34, n. 10, 2007, pp. 2902-2917.

HUANG, B.; KECHADI, M. T.; BUCKLEY, B. Customer churn prediction in telecommunications. Expert Systems with Applications, v. 39, n. 1, 2017, pp. 1414-1425.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. Science, v. 349, n. 6245, 2015, pp. 255-260.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, 1995, pp. 1137-1145.

LEMMENS, A.; CROUX, C. Bagging and boosting classification trees to predict churn. Journal of Marketing Research, v. 43, n. 2, 2006, pp. 276-286.

LEITE, E. R. Machine Learning: Técnicas e Aplicações. Editora XYZ, 2020.

LIU, Z.; JIANG, P.; DE BOCK, K. W.; WANG, J.; ZHANG, L.; NIU, X. Extreme gradient boosting trees with efficient Bayesian optimization for profit-driven customer churn prediction. Technological Forecasting and Social Change, v. 198, 2024, p. 122945.

MAIA, B. Tipos de Aprendizado de Máquina #3. 2020. Disponível em: https://dev.to/beatrizmaiads/tipos-de-aprendizado-de-maquina-3-5d66, Acesso em 12/08/2023.

MIRIC, M.; JIA, N.; HUANG, K. G. Using supervised machine learning for large-scale classification in management research: The case for identifying artificial intelligence patents. Strategic Management Journal, v. 44, n. 2, 2023, pp. 491-519.

MITCHELL, T. M. Machine Learning. McGraw-Hill, 1997.

MOLINARO, A. M.; SIMON, R.; PIEDMONTE, M. Estimating prediction error in classification using the unbalanced bootstrap. Statistics in Medicine, v. 24, 2005, pp. 2289-2305.

NGAI, E. W. T.; HU, Y.; WONG, Y. H.; CHEN, Y.; SUN, X. The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, v. 50, n. 3, 2011, pp. 559-569.

NGAI, E. W. T.; XIU, L.; CHAU, D. C. K. Application of data mining techniques in customer relationship management: A literature review and classification. Expert Systems with Applications, v. 36, n. 2, 2009, pp. 2592-2602.

PANIGRAHI, B.; KATHALA, K. C. R.; SUJATHA, M. A machine learning-based comparative approach to predict the crop yield using supervised learning with regression models. Procedia Computer Science, v. 218, 2023, pp. 2684-2693.

POWERS, D. M. W. Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies, v. 2, 2011, pp. 37-63.

QUEK, J. Y. V.; LIM, J. S. Y.; WONG, K. M.; LOW, J. X.; CHEONG, K. W.; SEAH, M. H. Customer churn prediction through attribute selection analysis and support vector machine. Journal of Telecommunications and the Digital Economy, v. 11, n. 3, 2023, pp. 180-194.

REICHHELD, F. F.; SCHEFTER, P. E-loyalty: Your secret weapon on the web. Harvard Business Review, v. 78, n. 4, 2000, pp. 105-113.

RIBEIRO, M. T.; SINGH, S.; GUESTRIN, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135-1144.

RISH, I. An empirical study of the naive Bayes classifier. IJCAI 2001 Workshop on Empirical Methods in AI, 2001, pp. 41-46.

SEN, P. C.; HAJRA, M.; GHOSH, M. Supervised classification algorithms in machine learning: A survey and review. In: Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018. Springer Singapore, 2020, pp. 99-111.

SHRUTI IYER. Churn Modelling. 2023. Disponível em: https://www.kaggle.com/datasets/shrutimechlearn/churn-modelling. Acesso em 04/08/203.

STEHMAN, S. V. Selecting and interpreting measures of thematic classification accuracy. Remote Sensing of Environment, v. 62, n. 1, 1997, pp. 77-89.

TRAN, H.; LE, N.; NGUYEN, V. H. Customer churn prediction in the banking sector using machine learning-based classification models. Interdisciplinary Journal of Information, Knowledge & Management, v. 18, 2023.

USMAN-HAMZA, F. E.; ABOBAKAR, A. T.; KADAR, I. M.; et al. Empirical analysis of tree-based classification models for customer churn prediction. Scientific African, v. 23, 2024, p. e02054.

VERBEKE, W.; DEJAEGER, K.; MARTENS, D.; HUR, J.; BAESENS, B. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research, v. 218, n. 1, 2012, pp. 211-229.

WALKER, S. H.; DUNCAN, D. B. Estimation of the probability of an event as a function of several independent variables. Biometrika, v. 54, n. 1-2, 1967, pp. 167-179.

YEN, G. G. Machine Learning and Applications. Springer, 2023.