1º COLOCADO
CATEGORIA I – FINANCIAMENTO AO DESENVOLVIMENTO SUSTENTÁVEL, INCLUSIVO E INOVATIVO

Inteligência Artificial como Vetor de Inovação e Inclusão Financeira: *uma abordagem com aprendizado de máquina*

*Nicolas Philomeno Suhadolnik**
*Paulo Cesar Starke Junior**
*Eraldo Sérgio Barbosa da Silva***

* Banco Regional de Desenvolvimento do Extremo Sul (BRDE)
** Universidade Federal de Santa Catarina (UFSC)

DOI: doi.org/10.24328/2026/89481.12.1

1. Introdução

Avanços regulatórios e tecnológicos permitiram a emergência de novos modelos de intermediação financeira, capazes de contribuir para o aumento da concorrência e a redução do custo do crédito. No Brasil, o Banco Central vem promovendo uma série de iniciativas voltadas à modernização e aumento da eficiência do sistema financeiro, incluindo a implantação de uma moeda digital de banco central e o compartilhamento aberto de dados (Araujo, 2022). O conceito de sistema financeiro aberto (*Open Finance)* potencializa a inovação em instituições financeiras, permitindo que ofereçam produtos mais personalizados e inclusivos, fomentando a concorrência. Sem comprometer a estabilidade e a transparência, esse ambiente favorece a utilização de tecnologias financeiras de crédito como elemento de transformação e geração de valor.

Como resultado do uso intensivo de métodos computacionais e quantidades massivas de dados nos processos de decisão de crédito, os modelos de aprendizado de máquina (*machine learning*) passaram a ocupar posição de destaque (Louzada; Ara; Fernandes, 2016). Apesar do reconhecido potencial, há uma preocupação por parte dos reguladores em relação à forma como esses métodos são empregados e as implicações associadas à estabilidade financeira. Essas abordagens são frequentemente criticadas por apresentarem uma natureza de "caixa preta", que limitam ou impedem qualquer tentativa de interpretação de como as decisões são tomadas (Chakraborty; Joseph, 2017).

Este artigo investiga o potencial do aprendizado de máquina na modelagem do risco de crédito, com foco em sua aplicação em instituições financeiras do Sistema Nacional de Fomento (SNF), em particular, bancos de desenvolvimento e agências de fomento. Para tanto, realiza-se uma revisão sistemática das principais abordagens presentes na literatura recente e uma análise empírica com base em um conjunto ampliado de dados de uma instituição financeira. A partir da comparação entre diferentes algoritmos, observa-se que os modelos combinados apresentam desempenho preditivo superior de forma consistente em relação às técnicas tradicionais, evidenciando sua robustez e aplicabilidade no contexto financeiro.

Além desta introdução, o artigo está organizado da seguinte forma. Na Seção 2 são discutidas as dimensões do risco de crédito e as diferenças entre as abordagens tradicionais e os métodos de aprendizado de máquina. Na Seção

3 apresentam-se as características dos algoritmos e medidas de desempenho utilizadas. A Seção 4 dedica-se à análise exploratória e tratamento dos dados. Na Seção 5 são realizados os experimentos e análise dos resultados. Por fim, a Seção 6 apresenta as considerações finais.

2. Revisão de Literatura

2.1. Dimensões da Análise do Risco de Crédito

A decisão de crédito, historicamente fundamentada em julgamentos de analistas humanos, apresenta limitações relacionadas à inconsistência e à influência de preferências individuais (Abdou; Pointon, 2011). Com o crescimento do mercado de crédito e o avanço das tecnologias financeiras, métodos estatísticos e computacionais passaram a complementar a avaliação humana ou, em alguns casos, substituí-la..

Para instituições financeiras, a concessão de crédito exige a definição de uma medida de risco associada ao proponente. Uma abordagem amplamente difundida na literatura e na prática é a dos "5 Cs do crédito": Caráter, relacionado ao histórico de pendências e indicadores de confiabilidade; Capacidade, referente à disposição para cumprir obrigações financeiras, com ênfase no grau de endividamento; Condições, que abrangem fatores macroeconômicos externos ao controle do tomador; Capital, vinculado à estrutura patrimonial e à liquidez; e Colateral, correspondente às garantias oferecidas para mitigar o risco de inadimplência (Tirole, 2006; Bazarbash, 2019).

De modo geral, a avaliação do risco de crédito consiste em definir um indicador sintético a partir das informações disponíveis no momento da solicitação. Tendo em vista a natureza multidimensional do problema, não existe um número ótimo de atributos que devem ser utilizados na construção dos modelos (Hand; Henley, 1997). Apesar do grande volume de dados produzido atualmente, as limitações no acesso reforçam a relevância de iniciativas como o *Open Finance,* que viabiliza o compartilhamento de dados entre instituições e permite uma avaliação de risco de forma mais precisa e abrangente. Ao integrar dados sobre investimentos, seguros e histórico financeiro, essas soluções promovem maior personalização na oferta de crédito, ampliando a concorrência e a inclusão no sistema financeiro.

No contexto de transformação digital e ampliação do acesso ao crédito, é relevante considerar o papel das instituições financeiras que integram o Sistema Nacional de Fomento (SNF) no Brasil. Esse ecossistema é fundamental para reduzir desigualdades regionais e apoiar projetos com impacto socioambiental, especialmente em segmentos que enfrentam restrições ao crédito nos canais tradicionais. Nesse cenário, o compartilhamento de dados promovido por iniciativas como o *Open Finance* representa uma oportunidade relevante, ao reduzir assimetrias informacionais frente aos bancos comerciais e ampliar a efetividade da análise de crédito orientada por dados.

2.2. Modelos Tradicionais x Aprendizado de Máquina

Como observa Breiman (2001), o comprometimento excessivo com técnicas tradicionais tem levado ao desenvolvimento de teorias irrelevantes, conclusões questionáveis e impedido o tratamento adequado de uma série de problemas. Em particular, se o propósito for a utilização de quantidades massivas de dados para resolver problemas, é recomendável a adoção de um conjunto de ferramentas mais diversas, sendo os métodos de aprendizado de máquina um caminho efetivo (Varian, 2014).

A principal distinção entre métodos estatísticos/econométricos e aprendizado de máquina reside nos objetivos. Enquanto a econometria prioriza a estimativa precisa de parâmetros, construção de intervalos de confiança e inferência causal (Athey; Imbens, 2019), o aprendizado de máquina foca na criação de algoritmos preditivos e classificatórios com base em informações limitadas, sendo o desempenho fora da amostra, em geral, o critério central de avaliação (Bazarbash, 2019).

Apesar de frequentemente apresentarem desempenho preditivo superior, alguns modelos de aprendizado de máquina carecem de interpretabilidade, dificultando a compreensão dos mecanismos subjacentes aos resultados. Esse aspecto tem influenciado a forma de atuação das instituições financeiras, evidenciada, por exemplo, pela adoção de processos decisórios totalmente automatizados. Metodologicamente, técnicas de aprendizado supervisionado são particularmente adequadas para classificar tomadores em categorias como "bons" ou "maus pagadores".

Diversos estudos compararam o desempenho de modelos de aprendizado de máquina na avaliação do risco de crédito. A partir de uma revisão de 214 estudos, Abdou e Pointon (2011) destacam a inexistência de um método universalmente superior para seleção de variáveis, tamanho amostral, critérios de validação e desempenho na avaliação de risco de crédito. Embora técnicas avançadas, como redes neurais e modelos combinados, apresentem desempenho superior às abordagens tradicionais, como regressão logística, os autores observam que os resultados entre diferentes métodos tendem a ser similares.

Louzada, Ara e Fernandes (2016), em revisão sistemática de 187 artigos entre 1992 e 2015, identificam como predominante na literatura de risco de crédito os modelos de redes neurais, máquinas de vetores de suporte (SVM), lógica *fuzzy*, regressão linear, árvores de decisão, regressão logística e modelos combinados. Os trabalhos de Lessmann *et al.* (2015) e Dastile, Turgay e Moshe (2020) mostram que modelos combinados apresentam desempenho preditivo superior quando comparados com classificadores simples, especialmente os baseados em regressão logística.

Markov, Zinaida e Lapshin (2022), ao analisar 150 artigos entre 2016 e 2021, identificam uma tendência crescente na adoção de métodos mais sofisticados, como redes neurais, em detrimento de técnicas tradicionais, como regressão logística, com ganhos relevantes em desempenho preditivo. Por outro lado, Zhang e Yu (2024) destacam uma lacuna na literatura quanto ao tratamento de dados, ressaltando a influência das características dos dados na performance dos modelos e a necessidade de desenvolver abordagens adequadas para seleção de técnicas conforme o contexto. Assim, os resultados reforçam a preocupação com o desenvolvimento de novos modelos, mas também evidenciam a importância de considerar aspectos como seleção de variáveis e qualidade dos dados.

3. Materiais e Métodos

A escolha dos algoritmos e medidas de desempenho teve como base a revisão de uma série de estudos que apresentam objetivos e utilizam conjuntos de dados análogos aos deste trabalho. Em seguida, buscou-se identificar os algoritmos com melhor desempenho, considerando as medidas tradicionalmente utilizadas, conforme a Tabela 1.

TABELA 1

SUMÁRIO DE TRABALHOS RELACIONADOS

Artigo	Período	n	Atributos	Algoritmos	Medidas de Desempenho
Serrano-Cinca *et al.* (2015)	2008–2011	3.788	5	Regressão Logística	Acurácia Teste de Hosmer-Lemeshow R^2 de Nagelkerke
Malekipirbazari e Aksakalli (2015)	2012–2014	68.000	15	Regressão Logística K-Vizinhos Mais Próximos Florestas Aleatórias Máquinas de Vetores de Suporte	Acurácia Área sob a Curva (AUC) RMSE
Xia *et al.* (2020)	2011–2013	64.139	15	Regressão Logística Árvore de Decisão Florestas Aleatórias Redes Neurais Artificiais *Gradient Boosting* *XGBoost* *CatBoost*	Acurácia Taxa de Falso Positivo Taxa de Falso Negativo Área sob a Curva (AUC) Índice H (Hirsch)
Teply e Polena (2020)	2009–2013	212.252	23	Regressão Logística K-Vizinhos Mais Próximos Árvore de Decisão Florestas Aleatórias Naïve Bayes Análise Discriminante Linear Rede Bayesiana Máquinas de Vetores de Suporte Rede Neural Artificial	Acurácia Área sob a Curva (AUC) Teste KS Brier Score Índice de Gini Índice H (Hirsch)
Este Trabalho	2007–2018	1.305.402	18	Regressão Logística Árvore de Decisão K-Vizinhos Mais Próximos Máquinas de Vetores de Suporte Rede Neural Artificial Florestas Aleatórias *Extra Trees* *AdaBoost* *Gradient Boosting* *XGBoost*	Acurácia Precisão Sensibilidade F1 Área sob a Curva (AUC)

Fonte: Serrano-Cinca *et al.* (2015); Malekipirbazari e Aksakalli (2015); Xia *et al.* (2020); Teply e Polena (2020).

3.1. Algoritmos Selecionados

Esta seção apresenta uma visão geral dos algoritmos de aprendizado supervisionado selecionados. O escopo deste trabalho não inclui uma descrição exaustiva de cada método, entretanto, algumas referências apresentadas oferecem uma discussão mais detalhada.

Regressão Logística: A regressão logística é uma técnica ainda muito utilizada pelas instituições financeiras na tomada de decisões de crédito. Apesar de ser uma técnica relativamente simples, mostra-se adequada para lidar com problemas de classificação binária (Teply; Polena, 2020). Conforme Louzada, Ara e Fernandes (2016), o modelo é definido a partir de uma combinação linear entre o conjunto de variáveis independentes ou atributos $X = \{X_1,...,X_n\}$ e a transformação logística da variável dependente binária $Y \in \{y_1,y_2\}$. Assim, se considerarmos y_1 como a categoria de interesse da análise (por exemplo, inadimplente) o modelo pode ser representado como:

$$\log \frac{\pi}{1-\pi} = X\beta,$$

onde, $\pi = P(Y = y_1)$ e β é o vetor de coeficientes do modelo.

Árvores de Decisão: são relativamente fáceis de serem implementadas e apresentam maior grau de interpretabilidade dos resultados. São construídas a partir de regras de decisão organizadas na forma de arquitetura de árvores. O objetivo consiste em estabelecer uma sequência de condições do tipo se-então-senão cobrindo todas as combinações possíveis em uma estrutura hierárquica análoga a um fluxograma, onde as decisões posteriores dependem das anteriores e o resultado é obtido da sequência de todas as decisões desde o nó raiz até o nó terminal ou folha. A decisão de particionamento em cada nó é realizada de modo a maximizar uma medida de pureza, como o índice de Gini ou a entropia. As partições em cada nó são realizadas para garantir que indivíduos com taxas de inadimplência similares fiquem na mesma região. Para evitar o *overfitting*, deve ser estabelecido um limitador para o tamanho da árvore (Bazarbash, 2019).

Florestas Aleatórias: podem ser vistas como um método que combina várias árvores de decisão que se diferenciam de duas formas. Primeiro, cada árvore é construída a partir de uma subamostra (chamada de *bagging*) da amostra original. Em segundo lugar, as partições em cada nó são otimizadas em relação a um subconjunto aleatório dos atributos. Essas duas modificações provocam uma variação suficiente nas árvores geradas e implicam em uma suavização dos resultados que são obtidos a partir da média dos resultados de cada árvore. Em problemas de classificação, uma regra de maioria pode ser utilizada para determinar o resultado. Com isso, os modelos de florestas aleatórias possuem uma capacidade preditiva superior à observada para uma árvore isolada (Athey; Imbens, 2019).

K-Vizinhos Mais Próximos (KNN): considerado um dos mais simples e populares, o método dos K-Vizinhos Mais Próximos, no contexto de algoritmos de classificação, pode ser definido como:

$$h(x) = moda_{i \in \aleph_x} y_i,$$

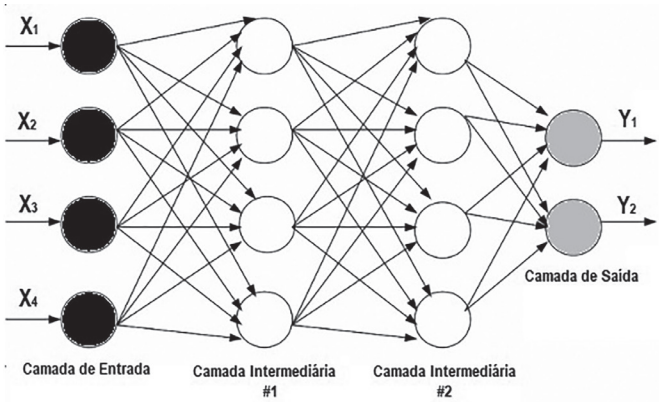
onde, $\aleph_x$ é o conjunto das k observações mais próximas de x. Assim, busca-se a classe mais frequente, y_i, observada para os k-vizinhos mais próximos de x, considerando uma medida específica de distância, como a distância euclidiana, aplicada ao conjunto de atributos de interesse (Izbicki; Santos, 2020).

Máquinas de Vetores de Suporte (SVM): em um problema de classificação binária, o objetivo do modelo SVM é encontrar a melhor função de classificação para separar os membros de cada uma das classes. A medida para estabelecer a "melhor" função de classificação pode ser obtida geometricamente. Para um conjunto de dados linearmente separável, uma função de classificação linear corresponde a um hiperplano de separação. Considerando que existem muitos hiperplanos lineares possíveis, o SVM garante que a melhor função de separação é encontrada maximizando a margem entre as duas classes. Geometricamente, a margem corresponde à menor distância entre o hiperplano e o ponto mais próximo de cada classe, também chamados de vetores de suporte (Wu *et al*., 2007).

Redes Neurais Artificiais: inspiradas na estrutura neural de organismos inteligentes que aprendem através da experiência, uma rede neural artificial é composta por vários nós de processamento ou neurônios. A arquitetura das redes neurais está organizada em camadas, conectadas através de uma estrutura de pesos relativos, como representado na Figura 1. Na primeira camada, ou camada de entrada, os atributos são utilizados no cálculo do valor dos nós que serão, em seguida, utilizados como entrada no cálculo dos nós da segunda camada. O processamento em cada nó ocorre por meio da aplicação de uma função de ativação, linear ou não linear. Uma rede neural pode possuir uma ou mais camadas intermediárias (*hidden layers*), além de uma camada de saída onde os resultados são obtidos. Na prática, podem ser utilizados modelos de redes neurais artificiais com dezenas de camadas, implicando em milhares ou milhões de parâmetros, o que pode exigir grande capacidade computacional. A principal vantagem das redes neurais está na sua flexibilidade para lidar com relações complexas em grandes conjuntos de dados. Por outro lado, as redes neurais artificiais estão associadas aos modelos "caixa preta", devido à dificul-

dade em interpretar como os resultados foram obtidos. Uma extensão dos modelos de redes neurais, conhecida como *deep learning*, é uma das áreas de aprendizado de máquina que tem recebido mais atenção nos últimos anos em função dos bons resultados obtidos em diversas aplicações (Bazarbach, 2019).

FIGURA 1
REDE NEURAL ARTIFICIAL



Fonte: Elaboração própria.

Modelos Combinados: existem diversos métodos para combinar modelos. Métodos de *boosting* podem ser utilizados para agregar um conjunto de classificadores de menor desempenho, de forma sequencial, de modo a obter um classificador com maior capacidade preditiva. De modo geral, o método de *boosting* consiste em treinar múltiplos modelos em sequência, sendo que a função de erro utilizada para treinar um modelo particular depende do desempenho dos modelos anteriores. Outra forma mais simples de combinar modelos é calcular a média das previsões de um conjunto de modelos individuais, como é feito no caso do *bagging*. Entre os métodos mais utilizados, destacam-se *AdaBoost* e *XGBoost* (Bishop, 2006; Xia *et al.*, 2020).

3.2. Medidas de Desempenho

Neste trabalho são utilizadas medidas de desempenho derivadas da matriz de confusão (*confusion matrix*) e área sob a curva (AUC). A matriz de confusão tem como objetivo comparar o resultado da classificação gerada pelo modelo com as verdadeiras classes observadas no conjunto de dados. No contexto da classifica-

ção binária de risco de crédito, um erro de classificação ocorre quando o modelo atribui uma classe ("inadimplente" ou "adimplente") a uma determinada empresa ou consumidor que diverge da classe verdadeira observada no conjunto de dados. Seguindo a definição de Louzada, Ara e Fernandes (2016) e Lindholm *et al.* (2021), a Tabela 2 apresenta a matriz de confusão, onde VP é o número de verdadeiros positivos, VN é o número de verdadeiros negativos, FP é o número de falsos positivos e FN é o número de falsos negativos, sendo $VP + VN + FP + FN = N$, onde N é o número total de observações do conjunto de dados.

TABELA 2
MATRIZ DE CONFUSÃO

	Previsto	
Observado	Inadimplente	Adimplente
Inadimplente	*VP*	*FN*
Adimplente	*FP*	*VN*

Fonte: Elaboração própria.

Acurácia: é uma medida de desempenho preditivo global que corresponde à proporção de classificações corretas realizadas pelo modelo. Apesar de ser muito utilizada, a acurácia pode ser inadequada para lidar com conjuntos de dados desbalanceados, favorecendo a classe majoritária.

$$Acurácia = \frac{VP + VN}{N}$$

Sensibilidade (*Recall*): também chamada de Taxa de Verdadeiros Positivos, corresponde à proporção de casos positivos ("inadimplentes") classificados corretamente pelo modelo em relação ao número total de verdadeiros positivos.

$$Sensibilidade = \frac{VP}{VP + FN}$$

Especificidade: conhecida também como Taxa de Verdadeiros Negativos, corresponde à proporção de casos negativos ("adimplentes") classificados corretamente pelo modelo em relação ao número total de verdadeiros negativos.

$$Especificidade = \frac{VN}{VN + FP}$$

Precisão: corresponde à proporção de casos positivos ("inadimplentes") classificados corretamente pelo modelo em relação ao número total de positivos.

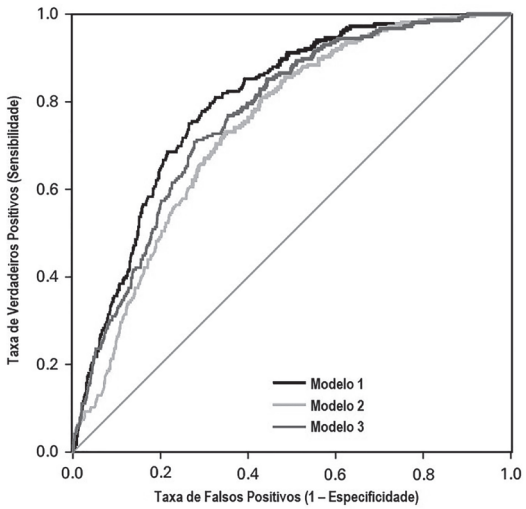
$$Precisão = \frac{VP}{VP + FP}$$

F1: permite obter uma medida sintética da Precisão e Sensibilidade através da sua média harmônica, bastante útil no caso de conjuntos de dados desbalanceados.

$$F1 = \frac{2 \times Precisão \times Sensibilidade}{Precisão + Sensibilidade}$$

A medida AUC, dada pela área sob a curva *Receiver Operating Characteristic* (ROC), também será utilizada. Geometricamente, a curva ROC é obtida plotando a medida de Taxa de Verdadeiros Positivos (Sensibilidade) no eixo y e a Taxa de Falsos Positivos (1 – Especificidade) no eixo x, considerando essa relação para diferentes pontos de corte na probabilidade estimada pelo modelo de classificação. Conforme a Figura 2, quanto mais próxima do canto superior esquerdo estiver a curva ROC, ou quanto maior o valor de AUC, melhor o desempenho do classificador.

FIGURA 2
CURVA ROC



Fonte: Elaboração própria.

4. Conjunto de Dados

4.1. Visão Geral e Tratamento Inicial

Um aspecto relevante dos trabalhos empíricos se refere à escolha do conjunto de dados. Diversos estudos utilizam dados disponíveis em repositórios públicos, como Australian Credit (AC) e German Credit (GC), ambos disponíveis no UCI Machine Learning Repository (Dua; Graff, 2017). Apesar da utilização frequente desses conjuntos de dados, o número reduzido de observações pode ser uma limitação importante em estudos cujo objetivo é a comparação do poder preditivo de diferentes classificadores.

O advento das FinTechs, incluindo plataformas de empréstimo *peer-to-peer* (P2P), ampliou significativamente a disponibilidade de conjuntos de dados (Berg *et al*., 2019). Embora os modelos utilizados sejam, em geral, desconhecidos, algumas empresas como a Lending Club, uma das plataformas pioneiras, permitem o acesso público aos dados. Assim, na escolha do conjunto de dados buscou-se, em primeiro lugar, ampliar o período da análise e o número de observações em relação aos estudos encontrados na literatura recente (Teply; Polena, 2020; Malekipirbazari; Aksakalli, 2015). Para isso, utilizou-se o conjunto de dados "*All Lending Club loan data*", disponível no repositório Kaggle (George, 2018) que, originalmente, reúne os empréstimos realizados pela Lending Club, no período de 2007–2018, com 2.260.701 observações (empréstimos) e 151 variáveis, com um total de 1,55 GB.

O conjunto de dados original contém as solicitações de empréstimo de pessoas físicas, no modelo de empréstimos P2P que foram aceitas pela plataforma Lending Club. As solicitações rejeitadas, ou seja, aquelas que não atenderam aos parâmetros da política de crédito da plataforma, não foram consideradas neste estudo. Os empréstimos coletivos, aqueles que possuem mais de um tomador, também foram removidos.

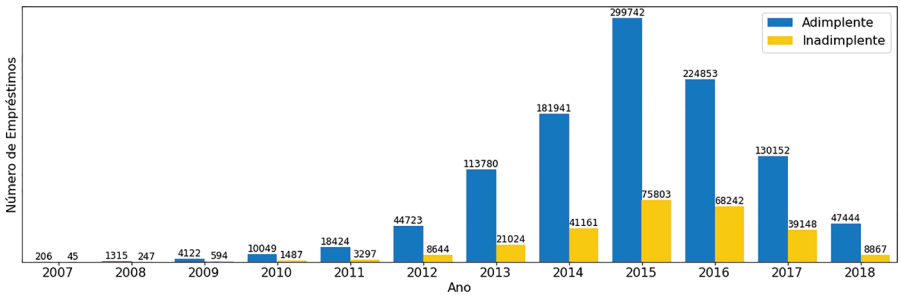
Para lidar com o problema de classificação supervisionada proposto, o atributo alvo "Situação do Empréstimo" foi transformado para representar um resultado binário. Nessa transformação, foram considerados apenas os empréstimos com situação *Fully Paid* (Adimplente) ou *Charged Off* (Inadimplente). Dessa forma, a variável passou a assumir o valor "0" para a situação adimplente e "1" para inadimplente. As demais situações, que incluem os empréstimos em andamento (sem atraso) e com atraso de até 120 dias, foram removidas. Com

isso, considerou-se apenas empréstimos que já atingiram seu prazo total e foram quitados ou aqueles inadimplentes com atraso superior a 120 dias.

O conjunto de dados resultante deste primeiro tratamento apresenta 1.076.751 (80,13%) empréstimos adimplentes e 268.559 (19,87%) inadimplentes. A Figura 3 mostra o número de empréstimos em cada classe ao longo do período analisado. A redução observada no número total de empréstimos a partir do ano de 2016 é resultado da remoção das operações em andamento, conforme tratamento descrito anteriormente para o atributo-alvo "Situação do Empréstimo". O desbalanceamento entre as classes "Adimplente" e "Inadimplente" será objeto de tratamento nas etapas posteriores.

FIGURA 3

NÚMERO DE EMPRÉSTIMOS POR ANO



Fonte: Elaboração própria.

4.2. Pré-Processamento dos Dados

Inicialmente, do total de 151 variáveis (atributos), 44 foram desconsideradas por apresentarem 50% ou mais de dados faltantes. Em seguida, após uma análise detalhada de cada variável, verificou-se que alguns atributos apresentavam redundância ou ausência de valor informacional para a classificação do risco de crédito. Isso ocorre, por exemplo, com variáveis que representam números identificadores da operação ou do tomador. Além disso, considerando que o objetivo deste trabalho é avaliar a capacidade preditiva de modelos de classificação de risco de crédito, foram incluídas apenas variáveis disponíveis no momento da solicitação do empréstimo pelo tomador. Assim, variáveis como as que identificam o histórico e programação de pagamentos do mutuário ou sua pontuação de crédito atualizada não foram consideradas.

Do total de 151 atributos disponíveis no conjunto de dados original, foram selecionados 18, além do atributo-alvo "Situação do Empréstimo". Na seleção dos atributos, buscou-se capturar aspectos vinculados às dimensões tradicionais do crédito, conforme discutido na Seção 2, além dos atributos que capturam as características gerais da solicitação. A Tabela 3 apresenta uma descrição de cada atributo selecionado.

TABELA 3

ATRIBUTOS SELECIONADOS

Atributo	Descrição	Tipo
Renda Anual	Declarada pelo tomador do empréstimo no momento da solicitação em dólares americanos	Numérico
Grau de Endividamento	Razão entre o valor dos pagamentos mensais de dívidas existentes e a renda mensal declarada	Numérico
Limite Comprometido	Razão entre o valor do crédito utilizado e o limite de crédito rotativo disponível (ex: cartões de crédito)	Numérico
Acesso ao Crédito	Número total de linhas de crédito abertas registradas no histórico de crédito do tomador	Numérico
Relacionamento Bancário	Número de linhas de crédito ativas para o tomador no momento da solicitação	Numérico
Histórico Bancário	Tempo, em anos, desde a abertura da primeira linha de crédito pelo tomador até o momento da solicitação	Categórico (valores possíveis: até 5 anos; 6–10 anos; 11–15 anos; 16–20 anos; +20 anos)
Pontuação de Crédito	Valor do limite inferior do intervalo de pontuação (FICO® *Score*) a que o tomador pertencia no momento da solicitação	Numérico
Pagamentos Atrasados	Indicador da existência de obrigações de pagamento com atraso superior a 30 dias nos últimos 2 anos	Categórico (valores possíveis: Sim ou Não)
Solicitações de Crédito	Número de solicitações de crédito nos últimos 6 meses, excluindo automóveis e hipotecas	Categórico (valores possíveis: nenhuma; 1; 2; 3 ou mais)
Pendências Cadastrais	Indicador da existência de pendências registradas no histórico cadastral do tomador	Categórico (valores possíveis: Sim ou Não)
Pendências Fiscais	Indicador da existência de pendências fiscais registradas no histórico do tomador	Categórico (valores possíveis: Sim ou Não)
Tempo de Trabalho	Tempo trabalhado pelo tomador, em anos, até o momento da solicitação	Categórico (valores possíveis: até 1 ano; 2–3 anos; 4–5 anos; 6–10 anos; +10 anos)
Tipo de Moradia	Situação de moradia do tomador no momento da solicitação	Categórico (valores possíveis: própria; hipotecada; alugada; outra)
Comprovação da Renda	Indicador de validação do valor ou da fonte de renda informada pelo tomador no momento da solicitação	Categórico (valores possíveis: verificada; não verificada; fonte verificada)
Valor do Empréstimo	Valor da solicitação de empréstimo em dólares americanos	Numérico

Atributo	Descrição	Tipo
Taxa de Juros do Empréstimo	Taxa de juros anual do empréstimo	Numérico
Prazo do Empréstimo	Prazo total do empréstimo em meses	Categórico (valores possíveis: 36 ou 60 meses)
Finalidade do Empréstimo	Finalidade ou objetivo do empréstimo selecionada pelo tomador no momento da solicitação	Categórico (valores possíveis: reestruturação de dívidas; cartão de crédito; reforma; compras; saúde; pequeno negócio; veículo; mudança; férias; imóvel; casamento; energia renovável; educação; outra)

Fonte: Elaboração própria.

A partir da identificação de valores discrepantes, faltantes e inconsistentes, foram realizadas transformações em alguns dos atributos selecionados. Para a variável "Renda Anual" foram identificados e removidos 13.288 (1,01%) valores discrepantes superiores a 2,5 desvios padrões da média. Para o "Grau de Endividamento", foi identificado e removido um valor negativo, tendo em vista que são esperados apenas valores positivos ou iguais a zero para este atributo. Foram identificados e removidos dois valores discrepantes na variável "Limite Comprometido", ambos superiores a 200, tendo como referência um valor médio de 52 e valor do terceiro quartil de 71 para a variável.

Para reduzir a quantidade de valores possíveis para a variável "Acesso ao Crédito", sem perda de informação, o valor máximo foi definido em 40. Assim, foram identificados e convertidos 1.199 (0,09%) valores superiores ao máximo estabelecido para este atributo. De forma análoga, para o atributo "Relacionamento Bancário", o valor máximo foi definido em 80 e, com isso, foram identificados e convertidos 1.281 (0,09%) valores superiores ao máximo estabelecido.

A variável categórica "Histórico Bancário" foi construída a partir do cálculo do número de anos decorridos desde a abertura da primeira linha de crédito pelo tomador até o momento da solicitação de crédito. Os atributos "Pagamentos Atrasados", "Pendências Cadastrais" e "Pendências Fiscais" foram transformados em variáveis binárias com valores possíveis "Sim" ou "Não". Considerando o número elevado de categorias no conjunto de dados original, as variáveis "Solicitações de Crédito" e "Tempo de Trabalho" foram discretizadas para reduzir o número de valores possíveis, sem perda de conteúdo informacional.

4.3. Análise Exploratória dos Dados

O conjunto de dados resultante do pré-processamento apresenta um total de 1.305.402 observações, sendo 1.045.072 (80,06%) empréstimos adimplentes e 260.330 (19,94%) inadimplentes. Do total de 18 atributos selecionados, sete são quantitativos, sendo cinco contínuos ("Renda Anual", "Grau de Endividamento", "Limite Comprometido", "Valor do Empréstimo" e "Taxa de Juros") e dois discretos ("Acesso ao Crédito" e "Relacionamento Bancário"). A Tabela 4 apresenta um sumário estatístico com as principais medidas-resumo para cada um dos atributos quantitativos selecionados.

TABELA 4

SUMÁRIO ESTATÍSTICO DOS ATRIBUTOS QUANTITATIVOS

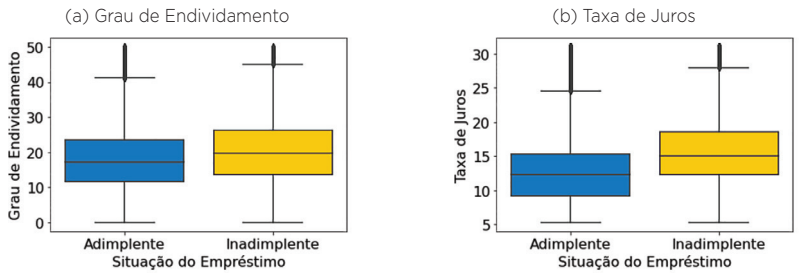
	Renda Anual	Grau de Endividamento	Limite Comprometido	Acesso ao Crédito	Relacionamento Bancário	Valor do Empréstimo	Taxa de Juros
Média	73,13 mil	18,1	51,85	11,57	24,93	14,21 mil	13,23
Desvio padrão	38,55 mil	8,35	24,44	5,42	11,91	8,57 mil	4,75
Mínimo	2 mil	0	0	0	2	0,5 mil	5,31
25%	46 mil	11,85	33,5	8	16	7,75 mil	9,75
50%	65 mil	17,61	52,2	11	23	12 mil	12,74
75%	90 mil	23,97	70,7	14	32	20 mil	15,99
Máximo	252,4 mil	49,96	193	40	80	40 mil	30,99

Fonte: Elaboração própria.

O grau de associação ou dependência entre os atributos quantitativos e a variável dependente "Situação do Empréstimo" podem ser exploradas, de forma preliminar, através da utilização de diagramas de caixa (*boxplot*), como mostra a Figura 4.

FIGURA 4

ASSOCIAÇÃO DE ATRIBUTOS QUANTITATIVOS COM "SITUAÇÃO DO EMPRÉSTIMO"



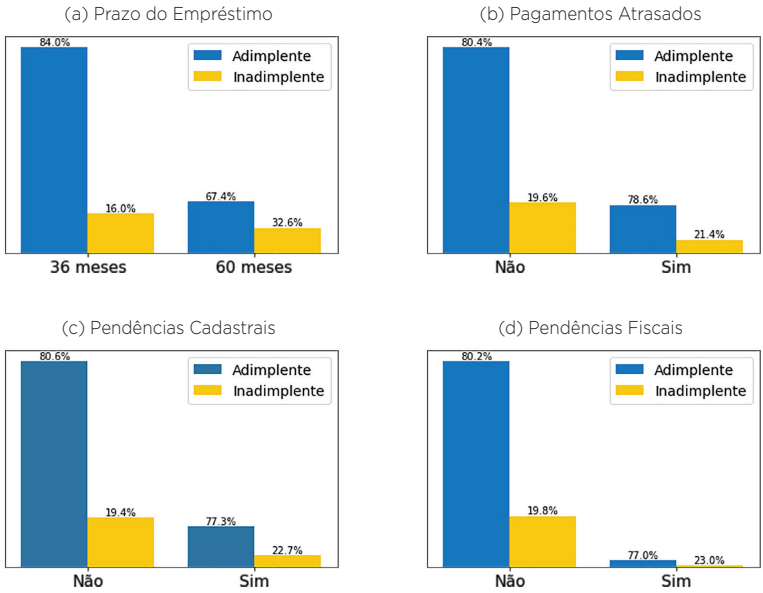
Fonte: Elaboração própria.

Pela análise visual inicial da Figura 4, é possível identificar uma associação positiva entre o "Grau de Endividamento" (Figura 4a) e a inadimplência, ou seja, maior grau de endividamento parece aumentar a chance de inadimplência. Da mesma forma, pela inspeção visual da variável "Taxa de Juros" (Figura 4b), é possível observar que maiores taxas de juros parecem estar associadas ao aumento da chance de inadimplência.

Pela análise visual das variáveis quantitativas "Renda Anual", "Limite Comprometido", "Acesso ao Crédito", "Relacionamento Bancário" e "Valor do Empréstimo" não foi possível identificar associação relevante com a variável "Situação do Empréstimo". Através das Figuras 5, 6 e 7 é possível analisar visualmente a relação entre os atributos qualitativos e a variável dependente "Situação do Crédito". A partir da inspeção visual da distribuição conjunta das variáveis, considerando cada classe do atributo-alvo "Situação do Empréstimo", é possível encontrar alguns indicativos de associação. Assim, não havendo dependência entre as variáveis, deve-se esperar as mesmas proporções para as classes "Adimplente" e "Inadimplente".

FIGURA 5

ASSOCIAÇÃO DE ATRIBUTOS QUALITATIVOS COM "SITUAÇÃO DO EMPRÉSTIMO"

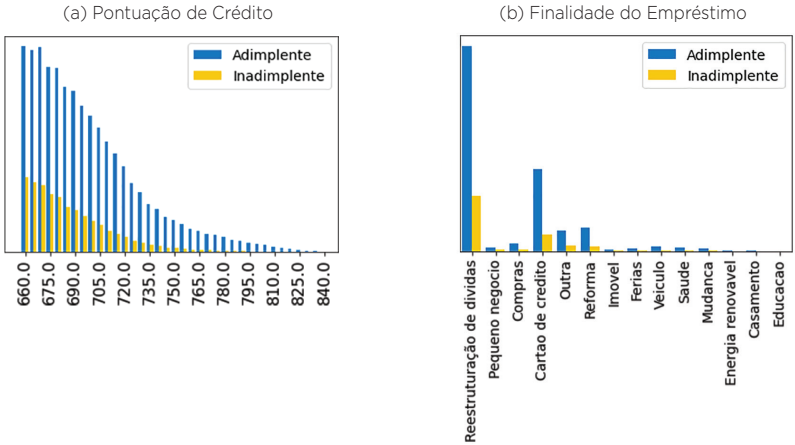


Fonte: Elaboração própria.

Em primeiro lugar, pela Figura 5, comparando as proporções de adimplentes e inadimplentes conforme o "Prazo do Empréstimo" (Figura 5a), é possível observar uma diferença significativa dentro de cada classe. Assim, um maior "Prazo do Empréstimo" (60 meses) parece estar associado a uma maior chance de inadimplência. Nota-se também que tomadores que apresentam "Pagamentos Atrasados" (Figura 5b), "Pendências Cadastrais" (Figura 5c) ou "Pendências Fiscais" (Figura 5d) estão associados a uma maior chance de inadimplência.

FIGURA 6

ASSOCIAÇÃO DE ATRIBUTOS QUALITATIVOS COM "SITUAÇÃO DO EMPRÉSTIMO"

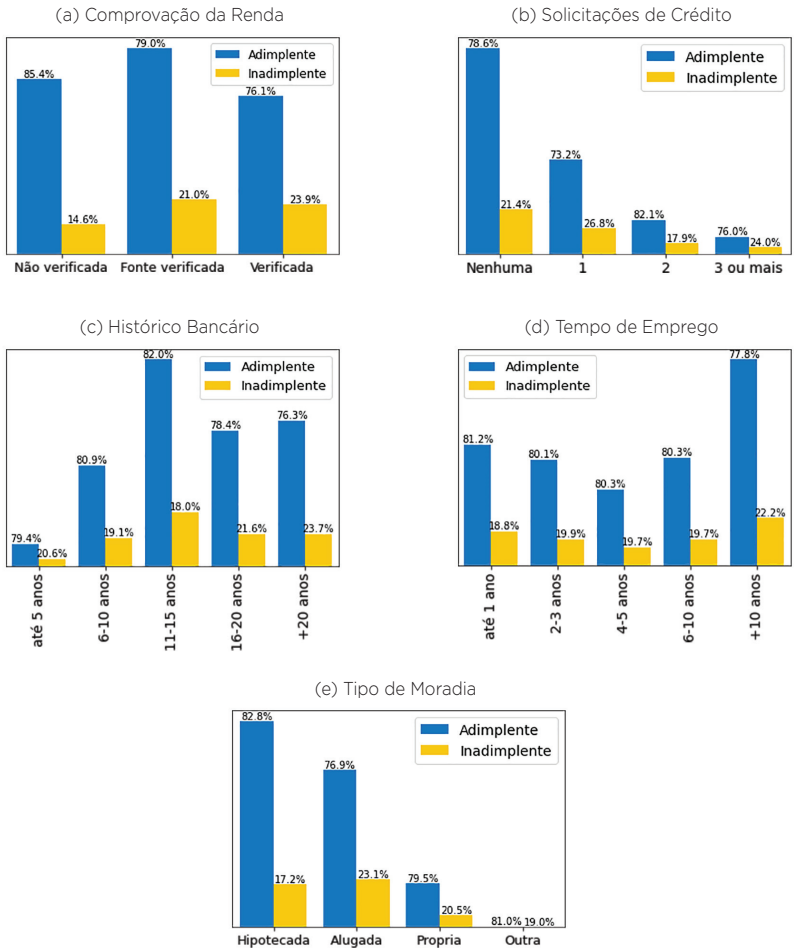


Fonte: Elaboração própria.

Os atributos "Pontuação de Crédito" e "Finalidade do Empréstimo" apresentam um número elevado de categorias, o que dificulta a análise visual. Entretanto, de acordo com os valores encontrados para a proporção de adimplentes e inadimplentes em cada categoria, é possível verificar que uma menor "Pontuação de Crédito" (Figura 6a) parece estar associada a uma maior chance de inadimplência. Com relação à "Finalidade do Empréstimo" (Figura 6b), os empréstimos que têm como objetivo "pequenos negócios" parecem implicar em maior chance de inadimplência.

FIGURA 7

ASSOCIAÇÃO DE ATRIBUTOS QUALITATIVOS COM "SITUAÇÃO DO EMPRÉSTIMO"



Fonte: Elaboração própria.

Por fim, vale destacar que a matriz de correlação calculada para as variáveis quantitativas selecionadas apresentou apenas um valor, em termos absolutos, superior a 0,5. A maior correlação encontrada foi entre as variáveis "Relacionamento Bancário" e "Acesso ao Crédito", com coeficiente igual a 0,7. De modo a evitar eventual perda de informação relevante para a classificação do risco de crédito, nenhum dos 18 atributos selecionados previamente foi removido do conjunto de dados.

5. Resultados

5.1. Desenho dos Experimentos

Com relação aos recursos computacionais, para a realização das simulações utilizou-se um computador Intel® Core™ i5-1035G1 CPU 1.19 GHz e 8 GB de memória RAM, utilizando o software JupyterLab®, versão 3.2.1, em linguagem de programação *Python*, versão 3.9.13. A Tabela 5 mostra a configuração inicial de hiperparâmetros utilizados nas simulações.

TABELA 5
ALGORITMOS E HIPERPARÂMETROS SELECIONADOS

Algoritmo	Hiperparâmetro
Regressão Logística (RL)	class_weight="balanced"
Árvore de Decisão (AD)	algorithm=CART, class_weight="balanced", max_depth=7, min_samples_leaf=0.01
K-Vizinhos Mais Próximos (KNN)	n_neighbors=11
Máquinas de Vetores de Suporte (SVM)	kernel="rbf", class_weight="balanced", max_iter=100,000
Rede Neural Artificial (RN)	hidden_layer_sizes=(8,4), activation="relu", solver="adam", learning_rate= 0.001("constant")
Florestas Aleatórias (FA)	class_weight="balanced", max_depth=7, min_samples_leaf=0.01
Extra Trees (ET)	class_weight="balanced", max_depth=7, min_samples_leaf=0.01
AdaBoost (AB)	algorithm= SAMME.R
Gradient Boosting (GB)	loss="log_loss", learning_rate=0.1, n_estimators=100, max_depth=3, min_samples_leaf=1
XGBoost (XGB)	booster=gbtree, learning_rate=0.3, gamma=0, alpha=0, min_child_weight=1, max_depth=6, sampling_method=uniform

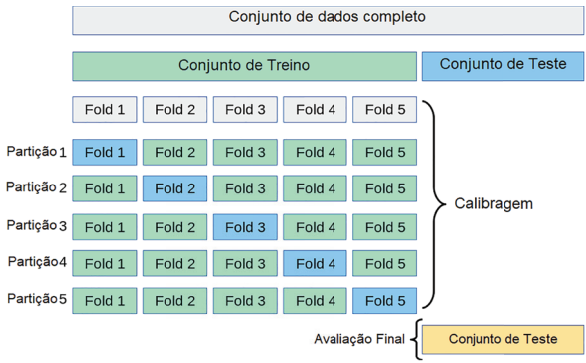
Fonte: Elaboração própria.

Em complemento ao tratamento dos dados descrito anteriormente, considerando a formatação de entrada exigida pelos algoritmos selecionados, todos os atributos qualitativos foram transformados em numéricos. Por fim, foi realizado o processo de normalização dos dados, de modo que cada atributo esteja no intervalo [0, 1], através do método *MinMaxScaler*, disponível na biblioteca *scikit-learn* versão 1.2.0.

Nas simulações, optou-se pela utilização do método de validação cruzada 5-*fold*, conforme sugerido por Malekipirbazari e Aksakalli (2015). Inicialmente, o conjunto de dados resultante da etapa de pré-processamento foi dividido em dois subconjuntos, sendo o primeiro de treino, contendo 80% do total de observações, e outro de teste com os 20% restantes. Em seguida, o subconjunto de treino foi particionado aleatoriamente em outros cinco subconjuntos (*folds*) de

mesmo tamanho. Com isso, na execução dos experimentos, cada um dos modelos foi treinado utilizando quatro subconjuntos resultantes do particionamento anterior e avaliado a partir do subconjunto que não foi utilizado para treinamento. Esse processo foi repetido cinco vezes, considerando um subconjunto diferente reservado para validação a cada execução. O subconjunto de teste, obtido do conjunto de dados pré-processado, foi reservado para utilização posterior na etapa de ajuste fino e avaliação do modelo final, conforme ilustrado na Figura 8.

FIGURA 8
VALIDAÇÃO CRUZADA



Fonte: Elaboração própria.

5.2. Resultados Gerais

Para avaliar o desempenho preditivo dos modelos, foram utilizadas as medidas de Acurácia, Precisão, Sensibilidade e F1, além da medida AUC, dada pela área sob a curva ROC. Nesta primeira abordagem, os modelos foram construídos sem ajuste fino de hiperparâmetros, ou seja, a partir da configuração tipicamente encontrada na literatura, com pequenas alterações para reduzir a chance de *overfitting* e limitar o tempo máximo de execução para alguns algoritmos. A Tabela 6 apresenta um sumário dos resultados de cada modelo nas simulações realizadas com o conjunto de dados completo após pré-processamento. Para cada uma das medidas de desempenho utilizadas, apresenta-se a média e o desvio padrão obtidos nas 5 iterações com validação cruzada.

TABELA 6

DESEMPENHO DOS MODELOS NO CONJUNTO DE DADOS COMPLETO[1]

Modelo	AUC	Acurácia	Precisão	Sensibilidade	F1	Tempo*
Regressão Logística (RL)	0.7087 [0.0019]	0.6568 [0.0013]	0.3205 [0.0026]	0.6403 [0.0037]	**0.4272 [0.0030]**	14
Árvore de Decisão (AD)	0.6998 [0.0018]	0.6167 [0.0076]	0.3003 [0.0039]	0.6896 [0.0104]	0.4183 [0.0027]	8
K-Vizinhos Mais Próximos (KNN)	0.6438 [0.0012]	0.7921 [0.0007]	0.4176 [0.0048]	0.1028 [0.0016]	0.1650 [0.0023]	7.806
Máquinas de Vetores de Suporte (SVM)	0.5488 [0.0078]	0.3759 [0.0546]	0.2012 [0.0035]	**0.7127 [0.0762]**	0.3130 [0.0061]	18.453
Rede Neural Artificial (RN)	0.7141 [0.0018]	0.8031 [0.0010]	**0.5640 [0.0115]**	0.0640 [0.0033]	0.1149 [0.0052]	43
Florestas Aleatórias (FA)	0.7032 [0.0012]	0.6283 [0.0022]	0.3058 [0.0025]	0.6773 [0.0030]	0.4214 [0.0026]	91
Extra Trees (ET)	0.6919 [0.0011]	0.6496 [0.0005]	0.3098 [0.0019]	0.6138 [0.0020]	0.4118 [0.0020]	69
AdaBoost (AB)	0.7087 [0.0013]	0.8023 [0.0010]	0.5439 [0.0063]	0.0657 [0.0052]	0.1172 [0.0082]	83
Gradient Boosting (GB)	0.7128 [0.0017]	0.8029 [0.0009]	0.5637 [0.0080]	0.0610 [0.0006]	0.1101 [0.0011]	351
XGBoost (XGB)	**0.7185 [0.0015]**	**0.8036 [0.0010]**	0.5549 [0.0081]	0.0857 [0.0019]	0.1484 [0.0031]	52

Fonte: Elaboração própria.

Como observado anteriormente, o conjunto de dados utilizado neste trabalho apresenta desbalanceamento entre as classes "Adimplente" (80%) e "Inadimplente" (20%). Nesse caso, pode-se verificar um desbalanceamento de aproximadamente 1:4 em favor da classe majoritária. Em alguns algoritmos de classificação, o desbalanceamento pode favorecer a classe majoritária e apresentar um desempenho preditivo inferior para a classe minoritária. Nesse contexto, a medida de Acurácia passa a ser um indicador limitado da performance dos modelos. Esse efeito pode ser observado na Tabela 6 através das medidas de Precisão, Sensibilidade e F1, em complemento à Acurácia.

Como forma de lidar com esse problema, adotou-se uma técnica de subamostragem, considerando que a quantidade de dados é suficientemente grande para realizarmos as simulações no conjunto de dados resultante. Através do método *RandomUnderSampler* da biblioteca *imblearn*, foram excluídas aleatoriamente observações da classe majoritária "Adimplente". O conjunto de dados

1 Desvio padrão apresentado entre colchetes. Destaque em negrito para o modelo com melhor desempenho em cada medida.

*Tempo médio, em segundos, por iteração.

após a reamostragem possui 416.528 observações igualmente divididas entre as classes "Adimplente" (50%) e "Inadimplente" (50%).

Com base nos mesmos modelos e hiperparâmetros utilizados nas simulações com o conjunto de dados completo, a Tabela 7 apresenta um sumário dos resultados obtidos por cada modelo nas simulações realizadas com o conjunto de dados balanceados. Conforme esperado, o processo de reamostragem contribuiu de forma significativa para a melhora do desempenho preditivo da classe minoritária, "Inadimplente", o que pode ser visto pelo aumento do valor das medidas Precisão, Sensibilidade e F1 na Tabela 7.

TABELA 7

DESEMPENHO DOS MODELOS NO CONJUNTO DE DADOS BALANCEADO[2]

Modelo	AUC	Acurácia	Precisão	Sensibilidade	F1	Tempo*
Regressão Logística (RL)	0.7076 [0.0015]	0.6499 [0.0017]	**0.6531 [0.0020]**	0.6393 [0.0031]	0.6461 [0.0024]	5,6
Árvore de Decisão (AD)	0.6988 [0.0013]	0.6431 [0.0016]	0.6366 [0.0031]	0.6670 [0.0139]	0.6513 [0.0053]	2,8
K-Vizinhos Mais Próximos (KNN)	0.6572 [0.0013]	0.6150 [0.0012]	0.6157 [0.0012]	0.6119 [0.0019]	0.6138 [0.0013]	1.327,4
Máquinas de Vetores de Suporte (SVM)	0.6111 [0.0117]	0.5335 [0.0362]	0.5309 [0.0378]	**0.8060 [0.1709]**	0.6279 [0.0353]	6.951,2
Rede Neural Artificial (RN)	0.7116 [0.0018]	0.6527 [0.0017]	0.6440 [0.0048]	0.6837 [0.0181]	**0.6631 [0.0063]**	33,7
Florestas Aleatórias (FA)	0.7026 [0.0008]	0.6459 [0.0015]	0.6390 [0.0025]	0.6709 [0.0039]	0.6545 [0.0019]	32,2
Extra Trees (ET)	0.6925 [0.0008]	0.6369 [0.0009]	0.6468 [0.0011]	0.6030 [0.0035]	0.6241 [0.0022]	26,3
AdaBoost (AB)	0.7078 [0.0012]	0.6500 [0.0011]	0.6441 [0.0014]	0.6704 [0.0030]	0.6570 [0.0019]	31
Gradient Boosting (GB)	0.7118 [0.0010]	0.6531 [0.0016]	0.6476 [0.0022]	0.6717 [0.0015]	0.6594 [0.0018]	130
XGBoost (XGB)	**0.7153 [0.0013]**	**0.6563 [0.0018]**	0.6507 [0.0022]	0.6749 [0.0019]	0.6626 [0.0020]	16,5

Fonte: Elaboração própria.

De modo geral, os modelos combinados baseados em métodos de *boosting*, utilizando árvores de decisão, *XGBoost* (XGB) e *Gradient boosting* (GB), apresentaram os melhores resultados, tanto no conjunto completo como no balanceado

2 Desvio padrão apresentado entre colchetes. Destaque em negrito para o modelo com melhor desempenho em cada medida.

*Tempo médio, em segundos, por iteração.

(reamostragem). Pela Tabela 8, vemos que o modelo XGB ficou classificado nas três primeiras posições em todas as medidas de desempenho preditivo.

TABELA 8

CLASSIFICAÇÃO DOS MODELOS NO CONJUNTO DE DADOS BALANCEADO[3]

Modelo	AUC	Acurácia	Precisão	Sensibilidade	F1
XGBoost (XGB)	1	1	2	3	2
Gradient boosting (GB)	2	2	3	4	3
Rede Neural Artificial (RN)	3	3	6	2	1
AdaBoost (ADA)	4	4	5	6	4
Regressão Logística (RL)	5	5	1	8	7
Florestas Aleatórias (FA)	6	6	7	5	5
Árvore de Decisão (AD)	7	7	8	7	6
Extra Trees (ET)	8	8	4	10	9
K-Vizinhos Mais Próximos (KNN)	9	9	9	9	10
Máquinas de Vetores de Suporte (SVM)	10	10	10	1	8

Fonte: Elaboração própria.

5.3. Otimização de Hiperparâmetros e Avaliação Final

A próxima etapa será dedicada à otimização dos hiperparâmetros, ou ajuste fino, além da avaliação do modelo final. Para isso, o modelo XGB foi escolhido por ter apresentado o melhor desempenho preditivo nas principais medidas analisadas entre todos os modelos previamente selecionados. O processo de otimização consiste em explorar o espaço de hiperparâmetros do modelo em busca de melhorar o desempenho preditivo de acordo com a medida de avaliação escolhida, que neste caso foi a AUC. Em seguida, a avaliação final do modelo será realizada utilizando o conjunto de teste resultante da divisão inicial do conjunto de dados pré-processado, que não foi utilizado em nenhuma etapa prévia, seja para treinamento ou validação dos modelos.

Para realizar o processo de otimização de hiperparâmetros, em geral, são utilizados os métodos *Grid Search* ou *Random Search*. O *Grid Search* consiste numa busca exaustiva pela melhor combinação de hiperparâmetros a partir de um conjunto de valores previamente definidos. No caso do *Random Search*, são realizadas combinações aleatórias de hiperparâmetros a partir do conjunto de valores previamente definidos, onde o número de iterações é explicitamente limitado. De acordo com Bergstra e Bengio (2012), processos de busca realiza-

3 Os números indicam a posição do modelo de acordo com cada medida de desempenho.

dos aleatoriamente, como o *Random Search*, são mais eficientes para otimização de hiperparâmetros em relação ao *Grid Search*, sendo capazes de encontrar modelos tão bons ou melhores com tempo de execução muito menor.

No processo de otimização dos hiperparâmetros, utilizou-se o método *RandomizedSearchCV*, disponível na biblioteca *scikit-learn*. A Tabela 9 apresenta o conjunto de hiperparâmetros e a grade de valores definidos para a busca aleatória. O valor ótimo de cada hiperparâmetro foi encontrado utilizando validação cruzada 5-*fold*, no conjunto de dados balanceado, considerando um limite de 200 iterações para as configurações possíveis. A partir dos hiperparâmetros encontrados no processo de otimização, o desempenho do modelo XGB foi avaliado no conjunto de testes.

TABELA 9

HIPERPARÂMETROS E VALORES

Hiperparâmetro	Grade de Valores	Valor Ótimo
max_depth	[3, 6, 8, 10, 12, 15, 20]	6
min_child_weight	[1, 3, 5, 7, 10]	1
gamma	[0, 0.0001, 0.001, 0.01, 0.1]	0.001
learning_rate	[0.01, 0.05, 0.1, 0.2, 0.3, 0.5]	0.3
alpha	[0, 0.0001, 0.001, 0.01, 0.1]	0.01
subsample	[0.1, 0.25, 0.5, 0.75, 1]	1
colsample_bytree	[0.1, 0.25, 0.5, 0.75, 1]	0.75
colsample_bylevel	[0.1, 0.25, 0.5, 0.75, 1]	0.75
colsample_bynode	[0.1, 0.25, 0.5, 0.75, 1]	1

Fonte: Elaboração própria.

A Tabela 10 apresenta um sumário das principais medidas de desempenho preditivo do modelo XGB, calculadas após o processo de otimização dos hiperparâmetros considerando o conjunto de dados de teste. De modo geral, o desempenho preditivo global do modelo, dado pela medida de Acurácia, foi de 0,65, ou seja, do total de classificações realizadas, 65% foram corretas. Analisando o desempenho por classe, a medida de Precisão para "Adimplente", que corresponde à proporção de casos negativos classificados corretamente pelo modelo em relação ao número total de negativos, revela uma taxa de acerto de 89%. Já para a classe "Inadimplente", é possível notar que o modelo teve um desempenho inferior, sendo que do total de classificações realizadas pelo modelo para esta classe, apenas 32% estavam corretas.

TABELA 10

DESEMPENHO DO MODELO *XGBOOST* COM OTIMIZAÇÃO DE HIPERPARÂMETROS

Classe	Precisão	Sensibilidade	F1	AUC
Adimplente (0)	0.8880	0.6393	0.7434	0.72
Inadimplente (1)	0.3151	0.6729	0.4292	0.72
Macro	0.6015	0.6561	0.5863	0.72
Balanceada	0.7746	0.6459	0.6812	-
Acurácia	-	-	0.6459	-

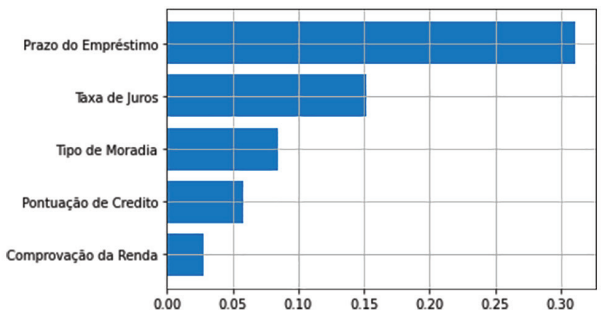
Fonte: Elaboração própria.

Pela medida de Sensibilidade, o desempenho para a classe "Inadimplente" mostra que o modelo teve uma Taxa de Verdadeiros Positivos de 67%, o que corresponde à proporção de casos positivos classificados corretamente em relação ao número total de verdadeiros positivos. Da mesma forma, olhando para a classe "Adimplente", o modelo teve uma Taxa de Verdadeiros Negativos de 64%. A medida AUC, calculada para as classes "Inadimplente" e "Adimplente", foi 0.72.

Uma das vantagens de utilizar algoritmos baseados em árvores de decisão, como é o caso do XGB, reside na possibilidade de calcular a contribuição, ou importância, de cada atributo para o desempenho do modelo. Grosso modo, a importância mede a contribuição de cada atributo na construção das árvores de decisão, ou seja, quanto maior o ganho médio de informação em todas as divisões em que um atributo é utilizado, maior é sua importância relativa. Como a importância é calculada explicitamente para cada atributo existente no conjunto de dados, é possível realizar uma ordenação por grau de importância, como mostra a Figura 9 para os 5 atributos mais relevantes. O "Prazo do Empréstimo" foi o atributo de maior importância relativa, com 31% do total, seguido da "Taxa de Juros" com 15%.

FIGURA 9

IMPORTÂNCIA RELATIVA DOS ATRIBUTOS



Fonte: Elaboração própria.

Os resultados deste trabalho corroboram Xia *et al.* (2020), Finlay (2011), Malekipirbazari e Aksakalli (2015), que encontraram desempenho preditivo superior dos métodos de *boosting* e *bagging* em aplicações empíricas de modelos de classificação de risco de crédito. Vale destacar que o custo computacional em termos de tempo de execução do modelo *XGBoost* foi relativamente baixo, sendo comparável aos modelos individuais como Regressão Logística (RL).

5.4. Implicações para o Sistema Nacional de Fomento

O Sistema Nacional de Fomento (SNF), formado por bancos de desenvolvimento, agências de fomento, bancos cooperativos, além da Finep e do Sebrae, ocupa posição estratégica no ecossistema financeiro brasileiro. Diferentemente das instituições financeiras convencionais, os objetivos dos bancos de desenvolvimento e agências de fomento não se restringem à maximização de resultados financeiros, mas estão orientados também à promoção do desenvolvimento econômico e social, à redução de desigualdades regionais e à viabilização de projetos que, em muitos casos, não encontrariam espaço no mercado de crédito privado. Essa natureza híbrida, que combina objetivos financeiros e finalidades públicas, gera especificidades que impactam diretamente no seu modo de atuação.

Em particular, os bancos de desenvolvimento, como o Banco Nacional de Desenvolvimento Econômico e Social (BNDES) e o Banco Regional de Desenvolvimento do Extremo Sul (BRDE), atuam com ênfase no financiamento de longo prazo para projetos estruturantes, especialmente em setores estratégicos como infraestrutura, inovação e sustentabilidade. Projetos dessa natureza exigem horizontes temporais mais longos, o que aumenta a exposição ao risco de inadimplência. Em linha com os resultados deste trabalho, quanto maior o prazo do financiamento, maior a probabilidade de ocorrência de eventos adversos, como crises econômicas, mudanças regulatórias, transformações tecnológicas ou mesmo alterações no perfil do tomador. Esse fator amplia o desafio para os modelos de análise de crédito, que precisam incorporar variáveis macroeconômicas, setoriais e institucionais, além de dados individuais dos tomadores.

De modo geral, a atuação das instituições financeiras que integram o SNF tem como característica o menor custo do crédito, seja em razão de margens reduzidas ou por incentivos governamentais e subsídios. Esse custo mais baixo pode, por um lado, contribuir para reduzir a inadimplência, ao aliviar o peso financeiro sobre os tomadores. Por outro lado, amplia o desafio da seleção cri-

teriosa, pois a atratividade das condições pode gerar excesso de demanda e incentivar a busca de crédito por agentes com diferentes perfis de risco. Nesse sentido, o aperfeiçoamento dos mecanismos de análise de crédito torna-se ainda mais relevante para preservar a sustentabilidade das instituições.

Podemos observar ainda uma distinção clara entre o crédito massificado de varejo e o crédito estruturado e não massificado. As agências de fomento operam principalmente com linhas de crédito direcionadas para micro e pequenas empresas, com destaque para microcrédito e capital de giro, muitas vezes em parceria com cooperativas de crédito, organizações sociais ou programas de microfinanças. Nessa modalidade, a análise de crédito com base em modelos de aprendizado de máquina mostra-se particularmente vantajosa, permitindo ganhos de escala, redução de custos operacionais, maior agilidade nos processos e ampliação da inclusão financeira. A adoção de ferramentas modernas de análises de crédito, associada ao compartilhamento de dados alternativos como histórico de pagamentos de serviços públicos ou movimentações digitais, pode reduzir significativamente a assimetria de informações e ampliar o alcance das políticas de fomento.

No âmbito do crédito estruturado, predominantemente conduzido por bancos de desenvolvimento e voltado à análise de projetos de maior porte e complexidade, a adoção de modelos avançados de avaliação de crédito revela-se igualmente promissora. Modelos preditivos podem ser utilizados para definir a remuneração da instituição financeira e as garantias necessárias conforme o risco identificado para o tomador, além de estabelecer parâmetros iniciais para a abertura de limites de crédito. Modelos de aprendizado de máquina também podem auxiliar na simulação de cenários, na avaliação de riscos setoriais e na identificação de externalidades positivas, como geração de empregos ou impacto ambiental, que são centrais para a missão dessas instituições.

Por fim, é importante destacar o papel que os novos modelos de análise de crédito baseados em aprendizado de máquina podem desempenhar na construção de um sistema de fomento mais eficiente. Essas ferramentas permitem que as instituições financeiras realizem o processo de análise e concessão de crédito com maior agilidade e menores custos, promovendo a inclusão financeira, especialmente entre micro e pequenas empresas (Bazarbash, 2019). Além disso, as classificações baseadas em aprendizado de máquina servem como uma ferramenta complementar valiosa para lidar com decisões de crédito complexas.

Nesses casos, a análise pode envolver modelos preditivos ampliados pela intervenção de analistas, que agregam à análise aspectos qualitativos relacionados ao empreendedorismo, à governança e à capacidade de execução dos projetos. Essa combinação entre tecnologia e julgamento humano reforça a missão das instituições financeiras de fomento na promoção do desenvolvimento sustentável e inclusivo, equilibrando eficiência técnica com sensibilidade social.

6. Conclusão

Este estudo demonstrou o potencial dos modelos de aprendizado de máquina na classificação de risco de crédito, especialmente em ambientes caracterizados por quantidades massivas de dados e complexidade analítica. A análise empírica, conduzida com dados reais de uma instituição financeira, evidenciou que os modelos combinados, com destaque para o *XGBoost*, superam consistentemente as abordagens tradicionais, como a regressão logística, em termos de desempenho preditivo. A aplicação de técnicas de reamostragem e otimização de hiperparâmetros contribuiu para mitigar os efeitos do desbalanceamento entre classes e aprimorar a capacidade de generalização dos modelos.

Os resultados obtidos neste estudo apresentam implicações relevantes para as instituições financeiras que integram o Sistema Nacional de Fomento (SNF). A superioridade dos modelos de aprendizado de máquina oferece uma alternativa promissora para aprimorar os processos de análise de crédito em instituições cuja atuação está orientada não apenas por resultados financeiros, mas também por objetivos sociais e de desenvolvimento regional. A capacidade desses modelos de lidar com grandes volumes de dados e identificar padrões complexos pode contribuir para a redução de assimetrias informacionais, ampliação da inclusão financeira e maior precisão na concessão de crédito, sobretudo em operações voltadas a micro e pequenas empresas. No contexto do crédito estruturado, os modelos preditivos podem auxiliar na definição de limites de crédito, garantias e remuneração conforme o perfil de risco dos tomadores, fortalecendo a sustentabilidade das operações e a missão pública dessas instituições.

Para aprofundar os resultados apresentados, recomenda-se que pesquisas futuras explorem a integração de variáveis macroeconômicas e temporais, visando capturar efeitos conjunturais sobre a inadimplência. Além disso, é perti-

nente o desenvolvimento de modelos explicáveis, que conciliem desempenho preditivo com transparência algorítmica, especialmente em contextos regulatórios. A utilização de dados alternativos, como histórico de pagamentos de serviços públicos e comportamento digital, representa uma via promissora para ampliar a inclusão financeira. Abordagens híbridas que combinem técnicas quantitativas com avaliação qualitativa, realizada por analistas humanos, também podem fortalecer a capacidade das instituições de fomento em decisões de crédito complexas. Por fim, estudos voltados à generalização dos modelos em diferentes contextos e à avaliação do impacto da automação sobre a equidade no acesso ao crédito são essenciais para consolidar o papel da inteligência artificial como vetor de inovação e transformação no sistema financeiro.

Referências

ABDOU, H. A.; POINTON, J. Credit scoring, statistical techniques and evaluation criteria: A review of the literature. **Intelligent Systems in Accounting, Finance and Management**, v. 18, n. 2–3, p. 59–88, 2011. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/isaf.325. Acesso em: 16 jun. 2022.

ARAUJO, F. Initial steps towards a central bank digital currency by the central bank of brazil. **BIS Papers**, n. 123, p. 31–37, 2022. Disponível em: https://www.bis.org/publ/bppdf/bispap123.pdf. Acesso em: 23 jun. 2022.

ATHEY, S.; IMBENS, G. W. Machine learning methods that economists should know about. **Annual Review of Economics**, v. 11, n. 1, p. 685–725, 2019. DOI: 10.1146/annurev-economics-080217-053433.

BAZARBASH, M. **Fintech in financial inclusio**n: Machine learning applications in assessing credit risk. IMF Working Papers, v. 2019, n. 109, 2019. Disponível em: https://www.elibrary.imf.org/view/journals/001/2019/109/001.2019.issue-109-en.xml. Acesso em: 24 maio 2022.

BERG, T.; BURG, V.; GOMBOVIĆ, A.; PURI, M. On the rise of fintechs: Credit scoring using digital footprints. **The Review of Financial Studies**, v. 33, n. 7, p. 2845–2897, 09 2019. DOI: 10.1093/rfs/hhz099.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. **Journal of Machine Learning Research**, v. 13, n. 10, p. 281–305, 2012. Disponível em: http://jmlr.org/papers/v13/bergstra12a.html. Acesso em: 10 jul. 2022.

BISHOP, C. M. **Pattern Recognition and Machine Learning (Information Science and Statistics)**. Berlin, Heidelberg: Springer-Verlag, 2006.

BREIMAN, L. Statistical modeling: The two cultures. **Statistical Science**, v. 16, n. 3, p. 199–215, 2001. Disponível em: http://www.jstor.org/stable/2676681. Acesso em: 20 jun. 2022.

CHAKRABORTY, C.; JOSEPH, A. Machine learning at central banks. **Bank of England Working Papers**, n. 674, 2017. Disponível em: https: //www.bankofengland.co.uk/working-paper/2017/machine-learning-at-central-banks. Acesso em: 10 mar. 2022.

Conselho Monetário Nacional (CMN). **Resolução nº 4.656, de 26 de abril de 2018**. Disponível em: https://www.bcb.gov.br/pre/normativos/busca/downloadNormativo.asp?arquivo=/Lists/Normativos/Attachments/50579/Res_4656_v1_O.pdf. Acesso em: 1 mar. 2022.

DASTILE, X.; TURGAY, C.; MOSHE, P. 2020. Statistical and machine learning models in credit scoring: A systematic literature survey. **Applied Soft Computing**, v. 91, 2020. DOI: 10.1016/j.asoc.2020.106263.

DUA, D.; GRAFF, C. **Uci machine learning repository**. 2017. Disponível em: http://archive.ics.uci.edu/ml. Acesso em: 14 abr. 2022.

FINLAY, S. Multiple classifier architectures and their application to credit risk assessment. **European Journal of Operational Research**, v. 210, n. 2, p. 368–378, 2011. ISSN 0377-2217. Disponível em: https://www.sciencedirect.com/science/article/pii/S0377221710006272. Acesso em: 17 dez. 2022.

GEORGE, N. **All lending club loan data**. 2018. Disponível em: https://www.kaggle.com/wordsforthewise/lending-club. Acesso em: 13 maio 2022.

HAND, D. J.; HENLEY, W. E. **Statistical classification methods in consumer credit scoring**: A review. **Journal of the Royal Statistical Society – Series A**, v. 160, n. 3, p. 523–541, 1997. Disponível em: http://www.jstor.org/stable/2983268. Acesso em: 11 jun. 2022.

IZBICKI, R.; SANTOS, T. M. dos. **Aprendizado de máquina**: uma abordagem estatística. 1ª edição, 2020. ISBN 978-65-00-02410-4.

LESSMANN, S.; BAESENS, B.; SEOW, H.; THOMAS, L. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. **European Journal of Operational Research**, v. 247, n. 1, p. 124–136, 2015. Disponível em: https://www.sciencedirect.com/science/article/pii/S0377221715004208. Acesso em: 14 jun. 2022.

LINDHOLM, A.; WAHLSTRÖM, N.; LINDSTEN, F.; SCHÖN, T. **Machine Learning**: A First Course for Engineers and Scientists. Reino Unido: Cambridge University Press, 2021. ISBN 9781108919371. Disponível em: https://books.google.com.br/books?id=nIWazgEACAAJ. Acesso em: 10 mar. 2022.

LOUZADA, F.; ARA, A.; FERNANDES, G. B. Classification methods applied to credit scoring: Systematic review and overall comparison. **Surveys in Operations Research and Management Science**, v. 21, n. 2, p. 117–134, 2016. Disponível em: https://www.sciencedirect.com/science/article/pii/S1876735416300101. Acesso em: 26 mar. 2022.

MALEKIPIRBAZARI, M.; AKSAKALLI, V. Risk assessment in social lending via random forests. **Expert Systems with Applications**, v. 42, n. 10, p. 4621–

4631, 2015. Disponível em: https://www.sciencedirect.com/science/article/pii/S0957417415000937. Acesso em: 18 jun. 2022.

MARKOV, A.; ZINAIDA, S; LAPSHIN, V. Credit scoring methods: Latest trends and points to consider. **The Journal of Finance and Data Science**, v. 8, p. 180–201, 2022. DOI: 10.1016/j.jfds.2022.07.002.

SERRANO-CINCA, C.; GUTIÉRREZ-NIETO, B.; LÓPEZ-PALACIOS, L. Determinants of default in p2p lending. **PLoS one**, v. 10, n. 10, p. e0139427, 2015.

TEPLY, P.; POLENA, M. Best classification algorithms in peer-to-peer lending. **The North American Journal of Economics and Finance**, v. 51, p. 100904, 2020. Disponível em: https://www.sciencedirect.com/science/article/pii/S1062940818302262. Acesso em: 3 abr. 2022.

TIROLE, J. **The theory of corporate finance**. Princeton: Princeton University Press, 2006.

VARIAN, H. R. Big data: New tricks for econometrics. **Journal of Economic Perspectives**, v. 28, n. 2, p. 3–28, 2014. Disponível em: https://www.aeaweb.org/articles?id=10.1257/jep.28.2.3. Acesso em: 13 jun. 2022.

VICENTE, J. Fintech disruption in Brazil: a study on the impact of open banking and instant payments in the brazilian financial landscape. **Social Impact Research Experience (SIRE)**, n. 86, 2020. Disponível em: https://repository.upenn.edu/sire/86. Acesso em: 11 jun. 2022.

WU, X.; KUMAR, V.; ROSS QUINLAN, J.; GHOSH, J.; YANG, Q.; MOTODA, H.; MCLACHLAN, G.; NG, A.; LIU, B.; YU, P.; ZHOU, Z.; STEINBACH, M.; HAND, D.; STEINBERG, D. Top 10 algorithms in data mining. **Knowledge and Information Systems**, v. 14, p. 1–37, 2007. DOI: 10.1007/s10115-007-0114-2.

XIA, Y.; HE, L.; LI, Y.; LIU, N; DING, Y. Predicting loan default in peer-to-peer lending using narrative data. **Journal of Forecasting**, v. 39, n. 2, p. 260–280, 2020. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/for.2625. Acesso em: 25 jun. 2022.

ZHANG, X. e YU, L. Consumer credit risk assessment: A review from the state-of-the-art classification algorithms, data traits, and learning methods. **Expert Systems with Applications**, v. 237-A, 2024. DOI: 10.1016/j.eswa.2023.121484.