UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

MILENA HAMERSKI

SISTEMA DE APOIO À DECISÃO PÓS-CONCESSÃO DE CRÉDITO: ANÁLISE E PREDIÇÃO DE INADIMPLÊNCIA

GUARAPUAVA

2025

MILENA HAMERSKI

SISTEMA DE APOIO À DECISÃO PÓS-CONCESSÃO DE CRÉDITO: ANÁLISE E PREDIÇÃO DE INADIMPLÊNCIA

Decision Support System for Post-Credit Evaluation: Default Analysis and Prediction

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Tecnologia Em Sistemas Para Internet do Curso Superior De Tecnologia Em Sistemas Para Internet da Universidade Tecnológica Federal do Paraná.

Orientador: Profª. Drª Kelly Lais Wiggers

GUARAPUAVA

2025

RESUMO

Este trabalho propõe o desenvolvimento de um protótipo de sistema preditivo voltado ao acompanhamento pós-concessão de crédito, com o objetivo de identificar antecipadamente padrões de risco de inadimplência ao longo do ciclo de crédito. A motivação para sua realização decorre da necessidade de aprimorar o monitoramento contínuo da carteira, utilizando de forma mais eficiente os dados históricos disponíveis. A metodologia prevista inclui a preparação e tratamento do conjunto de dados, envolvendo etapas como limpeza, padronização e seleção de variáveis relevantes para o processo preditivo. Posteriormente, serão aplicados algoritmos de aprendizado de máquina para modelagem do risco, adotando métricas reconhecidas para avaliar o desempenho dos modelos. O protótipo a ser desenvolvido terá como foco a aplicação e avaliação de modelos de aprendizado de máquina para estimar o risco de inadimplência no período pós-concessão de crédito. O objetivo é verificar se esses modelos apresentam desempenho adequado para apoiar esse tipo de análise, considerando as métricas selecionadas para avaliação.

Palavras-chave: aprendizado de maquina; pos-concess; risco de credito; inadimplencia; modelagem preditiva.

ABSTRACT

This work proposes the development of a predictive system prototype aimed at the post-loan monitoring of credit, with the objective of identifying early patterns of default risk throughout the credit lifecycle. The motivation for its development stems from the need to enhance continuous portfolio monitoring by making more efficient use of available historical data. The planned methodology includes the preparation and processing of the dataset, involving steps such as cleaning, standardization, and the selection of variables relevant to the predictive task. Subsequently, machine learning algorithms will be applied for risk modeling, adopting recognized metrics to evaluate model performance. The prototype to be developed will focus on the application and assessment of machine learning models to estimate default risk in the post-loan period. The goal is to verify whether these models achieve adequate performance to support this type of analysis, based on the selected evaluation metrics.

Keywords: machine learning; post-loan monitoring; credit risk; default; predictive modeling.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Considerações iniciais

Nos últimos anos, o Brasil atravessou um cenário macroeconômico atípico, marcado por forte instabilidade decorrente da pandemia da COVID-19 e suas consequências sobre a atividade econômica. Oscilações expressivas no Produto Interno Bruto (PIB), na inflação (IPCA), no desemprego e na taxa básica de juros (SELIC) impactaram diretamente a capacidade de pagamento das famílias, refletindo-se nos níveis de inadimplência. Estudos recentes confirmam que os níveis de inadimplência respondem com sensibilidade às variações da Selic e do desemprego, sugerindo a necessidade de maior prudência na concessão de crédito em cenários macroeconômicos instáveis (MAIA; RAMíREZ; CUTINO, 2025).

Esse contexto macroeconômico encontra correspondência nos dados práticos do mercado de crédito brasileiro. De acordo com o Mapa da Inadimplência e Negociação de Dívidas do Serasa (EXPERIAN, 2025), o número de brasileiros inadimplentes saltou de 72,54 milhões em maio de 2024 para 78,8 milhões em Agosto de 2025, representando quase metade da população. Entre os principais tipos de dívidas, destacam-se aquelas vinculadas a bancos e cartões de crédito (27,3%), contas essenciais como água, luz e gás (20,8%) e dívidas com financeiras (19,5%). Esses números evidenciam a fragilidade financeira de grande parte da população e a urgência de mecanismos mais eficazes para previsão e mitigação do risco de crédito.

A escolha do tema é resultante de experiências e questionamentos ao longo do percurso acadêmico e profissional, evidenciando uma afinidade pessoal com a junção entre tecnologia e finanças. A área de crédito e inadimplência apresenta-se como um campo de relevância prática e social, especialmente diante do contexto atual, em que instituições financeiras buscam soluções mais eficazes para reduzir riscos e sustentar a confiança de seus associados.

A inteligência artificial pode ser entendida como o conjunto de técnicas computacionais capazes de aprender padrões a partir de dados e realizar previsões ou recomendações com base nesse aprendizado. No campo financeiro, essas ferramentas permitem identificar correlações complexas que muitas vezes não são percebidas por métodos tradicionais de análise. Estudos recentes demonstram que a integração de tecnologias como análise de grandes volumes de dados (*big data*[1]), mineração de dados e aprendizado de máquina possibilita construir modelos preditivos robustos para avaliar o risco de crédito corporativo, superando abordagens tradicionais em precisão e confiabilidade (XIANYU; HAI, 2023). Essa aplicação abre espaço para sistemas que apoiam decisões estratégicas e preventivas, contribuindo para a gestão de riscos e a redução de inadimplência.

[1] O termo *big data* se refere ao conjunto de técnicas e ferramentas utilizadas para coletar, armazenar, processar e analisar grandes volumes de dados, estruturados ou não, com o objetivo de extrair informações relevantes para a tomada de decisão.

O interesse pela aplicação da inteligência artificial nesse cenário decorre do potencial que essas ferramentas possuem para transformar grandes volumes de dados em análises consistentes, capazes de apoiar processos de tomada de decisão mais embasados por parte do analista de crédito. Atualmente, as decisões permanecem baseadas, em grande parte, no julgamento humano do analista, sujeito à subjetividade, lentidão e vieses involuntários. A tecnologia, nesse sentido, potencialmente se apresenta como uma aliada para complementar a análise do crédito, contribuindo com novas perspectivas e maior precisão.

A relevância do estudo também se apoia no impacto direto que a inadimplência gera na saúde financeira de instituições e de seus clientes. Ao considerar o uso de modelos preditivos, busca-se explorar a possibilidade de antecipar cenários de risco e, assim, apoiar a construção de estratégias mais assertivas. Essa abordagem dialoga com a necessidade de soluções compatíveis com as demandas do setor em um ambiente financeiro que, cada vez mais, demanda eficiência, confiança e adaptação constante às transformações tecnológicas.

1.2 Objetivos

1.2.1 Objetivo geral

Desenvolver um sistema de apoio à decisão pós-concessão de crédito para Pessoa Física, capaz de prever a probabilidade de inadimplência a partir da análise de dados financeiros.

1.2.2 Objetivos específicos

Com base no objetivo geral deste trabalho, os objetivos específicos são:

- Analisar o ciclo de crédito e suas etapas, destacando como o risco de inadimplência se manifesta no período pós-concessão.
- Investigar técnicas de aprendizado de máquina aplicadas à previsão de inadimplência e suas aplicações preditivas.
- Selecionar e preparar um conjunto de dados financeiros, realizando limpeza, normalização e tratamento das variáveis relevantes para a construção do modelo.
- Desenvolver e treinar um modelo preditivo de inadimplência, utilizando métricas apropriadas para avaliar seu desempenho.
- Validar, interpretar e aplicar os resultados obtidos, analisando a precisão do modelo e sua utilidade como ferramenta de apoio à decisão em instituições financeiras.

1.3 Justificativa

O tema deste trabalho surge da observação do funcionamento da esteira de crédito, abrangendo tanto a análise quanto o acompanhamento pós-concessão, em uma instituição financeira de médio porte localizada no sul do Brasil. A vivência desses processos e as discussões recorrentes sobre o aumento da inadimplência evidenciaram a existência de um grande volume de dados valiosos que, apesar de registrados rotineiramente, ainda são pouco explorados de maneira estratégica. Informações como principalidade[2], faixa de risco, capacidade de pagamento (CAPAG) e existência de avais, entre outras, constituem insumos relevantes para observar tendências, compreender o comportamento dos clientes após a liberação do crédito e apoiar decisões relacionadas ao acompanhamento da carteira. Considerando a necessidade de transparência e a possibilidade de divulgação integral dos resultados, este estudo utiliza exclusivamente dados públicos, permitindo demonstrar o potencial analítico dessas informações sem expor dados internos, sensíveis ou proprietários.

A relevância do tema se reforça diante da possibilidade de transformar registros históricos em conhecimento útil por meio do aprendizado de máquina. Mesmo com o uso de bases públicas, análises desse tipo podem auxiliar na investigação de comportamentos associados ao período pós-concessão e na identificação de fatores ligados ao aumento da inadimplência, além de indicar oportunidades para aprimorar políticas de acompanhamento e estratégias de recuperação de crédito. A execução completa do processo de análise de dados — envolvendo limpeza, integração, seleção, transformação, mineração e avaliação de padrões — evidencia o potencial dessas técnicas para apoiar decisões estratégicas e apontar caminhos de modernização tecnológica que futuramente podem ser aplicados em ambientes financeiros com bases internas mais abrangentes.

A crescente demanda por maior agilidade e padronização nas atividades relacionadas ao monitoramento da carteira torna a adoção de técnicas automatizadas ainda mais pertinente. Mesmo utilizando dados públicos, análises automatizadas contribuem para reduzir subjetividade, acelerar o tempo de resposta e aumentar a eficiência. Além disso, modelos preditivos podem auxiliar na detecção precoce de indícios de maior probabilidade de inadimplência, favorecendo que equipes de cobrança e recuperação direcionem seus esforços de maneira mais estratégica e tempestiva. Assim, o uso de algoritmos tende a complementar, e não substituir, a atuação humana, fortalecendo o acompanhamento contínuo da carteira.

Do ponto de vista social e econômico, a inadimplência afeta não apenas as instituições financeiras, mas também seus clientes, podendo gerar restrições e dificultar o acesso a crédito. Sistemas preditivos baseados em dados abertos podem oferecer benefícios para a gestão de risco e para a proteção dos usuários, contribuindo para maior estabilidade e confiança no ambiente financeiro.

[2] A principalidade é o grau de importância que uma instituição financeira tem na vida de seus clientes.

2 REFERENCIAL TEÓRICO

2.1 Ciclo de Crédito e Mecanismos de Controle

O ciclo de crédito passa por fases estruturadas que conduzem desde a definição de políticas até a conclusão da operação de crédito. Inicialmente, a instituição financeira estabelece sua política de crédito, composta por estratégias e regras voltadas à redução da exposição a possíveis riscos de crédito. O risco de crédito refere-se à possibilidade de ocorrência de perdas financeiras decorrentes do não cumprimento das obrigações pelo tomador ou contraparte, incluindo situações de inadimplência, redução de ganhos ou custos relacionados à recuperação de créditos problemáticos (OLIVEIRA, 2024).

A fase de concessão do crédito consiste na compatibilização entre o perfil do tomador e as condições estipuladas pela instituição, enquanto a etapa de acompanhamento e manutenção permite monitorar continuamente a saúde financeira das operações e identificar oportunidades de ajustes. Por fim, a etapa de cobrança e liquidação visa à conclusão do contrato, utilizando estratégias como renegociação, descontos ou outras medidas para reduzir perdas financeiras (OLIVEIRA, 2024).

Para aumentar a assertividade na gestão de crédito, diversas instituições adotam mecanismos que alinham suas políticas internas às condições macroeconômicas e ao comportamento financeiro dos clientes. No Brasil, uma dessas ferramentas é o bureau de crédito. A palavra bureau, de origem francesa, refere-se a bancos de dados financeiros que auxiliam empresas na decisão de conceder crédito. Os bureaus aplicam grades de pontuação (*credit scoring*) para avaliar a saúde financeira dos clientes, permitindo análises mais precisas do risco de crédito. Na prática, quando um consumidor atrasa o pagamento de uma dívida ou assume um novo compromisso financeiro, essa informação é registrada pela empresa credora e reportada ao bureau. Quando a dívida é quitada, o registro é atualizado ou removido, refletindo fielmente o comportamento financeiro do cliente. No Brasil, os principais birôs de crédito são a Serasa, a Boa Vista, o SPC Brasil e a Quod (SERASA, 2024), que atuam de forma privada e comercial, conforme as diretrizes da Lei do Cadastro Positivo e da Lei Geral de Proteção de Dados (LGPD).

Além desses birôs privados, existe o Sistema de Informações de Crédito (SCR), gerido pelo Banco Central do Brasil. Diferentemente dos bureaus, o SCR não é uma empresa de análise de crédito, mas um sistema público que compila informações detalhadas sobre operações de crédito de pessoas físicas e jurídicas cujo risco direto exceda R$ 200,00, incluindo empréstimos, repasses interfinanceiros, coobrigações e limites de crédito (Banco Central do Brasil, 2022). O principal objetivo do SCR é permitir o monitoramento prudencial das instituições financeiras, ajudando o Banco Central a identificar operações atípicas ou de alto risco e a adotar medidas preventivas para preservar a estabilidade do sistema financeiro. Além disso, o SCR fornece às instituições uma base de dados estruturada que contribui para decisões de crédito

mais seguras, respeitando a privacidade do cliente, que deve autorizar expressamente o acesso às suas informações (Banco Central do Brasil, 2022).

> De acordo com a Resolução nº 5.037, de 29/09/2022, do Conselho Monetário Nacional (CMN), são consideradas operações de crédito, para efeitos do SCR, empréstimos, financiamentos, adiantamentos, arrendamento mercantil, prestação de aval, fiança, coobrigação, créditos baixados como prejuízo, operações com instrumentos de pagamento pós-pagos, entre outras modalidades, sendo necessário que o acesso às informações do sistema por instituições financeiras dependa da autorização expressa do cliente (Banco Central do Brasil, 2022).

2.2 Mineração de Dados

A Mineração de Dados (MD) é o processo de descoberta de padrões, relações e informações úteis em grandes volumes de dados, permitindo a extração de conhecimento significativo a partir de dados brutos. Diferentemente de apenas armazenar ou organizar informações, o objetivo da MD é extrair conhecimento que possa apoiar decisões estratégicas, previsões de comportamento e análises complexas que auxiliem gestores, cientistas e analistas em diversos setores, incluindo finanças, saúde, marketing, indústria e tecnologia da informação. Por ser um domínio de estudo fortemente guiado pelas necessidades de uma aplicação ou problema real, a MD incorporou diversas técnicas de outras áreas, como estatística, aprendizado de máquina, reconhecimento de padrões, bancos de dados e *data warehouses*[1], além de algoritmos especializados que permitem desde a análise descritiva até a prescritiva (HAN; PEI; TONG, 2022).

Além disso, a MD desempenha um papel crucial na identificação de tendências e comportamentos que não seriam facilmente perceptíveis por métodos tradicionais de análise de dados. Por exemplo, em instituições financeiras, a MD permite a detecção precoce de riscos de crédito, fraudes e anomalias, contribuindo diretamente para a tomada de decisão mais segura e embasada. Em empresas de comércio eletrônico, auxilia na recomendação de produtos, segmentação de clientes e otimização de campanhas de marketing, aumentando a eficiência operacional e a competitividade. A versatilidade da MD também se estende a setores como saúde, onde pode ser aplicada para identificar padrões de doenças, prever surtos e auxiliar em diagnósticos, ou em logística, para otimizar rotas, estoques e recursos.

Sendo um campo tão multidisciplinar (Figura 1), a MD contribui significativamente para o sucesso das análises e das aplicações práticas, proporcionando um diferencial competitivo importante para organizações que dependem de dados para decisões estratégicas (HAN; PEI; TONG, 2022).

[1] O termo “data warehouses” refere-se a sistemas especializados para armazenamento e análise de grandes volumes de dados.

O processo de (MD) envolve diversas etapas que transformam dados brutos em conhecimento útil. Geralmente, esse processo inclui a coleta e integração de dados provenientes de múltiplas fontes, a limpeza e o pré-processamento para corrigir inconsistências e lidar com valores ausentes, a seleção de atributos relevantes para análise, a aplicação de técnicas de mineração, que podem variar desde simples estatísticas descritivas até algoritmos complexos de aprendizado de máquina, e, por fim, a interpretação dos resultados para gerar descobertas acionáveis.

Cada uma dessas etapas é essencial para garantir a qualidade, confiabilidade e utilidade das descobertas, permitindo que os conhecimentos obtidos sejam aplicados de forma eficaz na tomada de decisões estratégicas e operacionais em diferentes contextos e setores.

Vale ressaltar que esse processo é cíclico, ou seja, existe um ciclo de vida dos dados. É fundamental que os dados sejam continuamente atualizados e refinados, de modo a manter a relevância e a precisão das análises e previsões geradas.

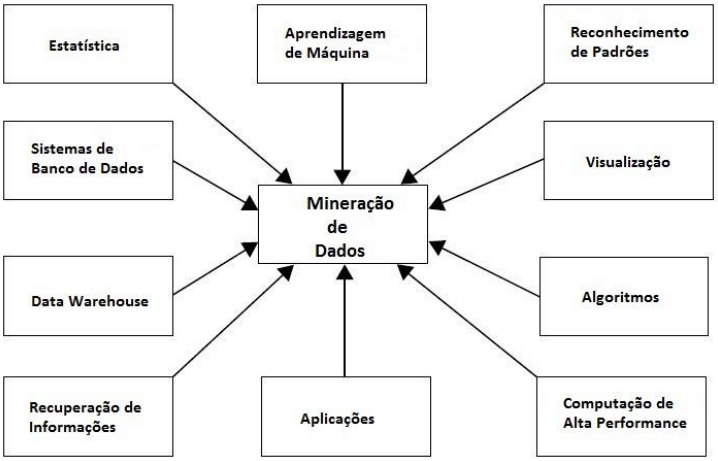

Figura 1 – Multidisciplinaridade da Mineração de Dados. Fonte: Research Gate

Figura 2 – Etapas do processo de Mineração de Dados. Adaptado de DBM Sistemas

2.2.1 Pré-processamento de Dados

Antes da aplicação de qualquer modelo de aprendizado de máquina, é essencial garantir a qualidade e representatividade dos dados. O pré-processamento consiste em um conjunto de técnicas voltadas à preparação dos dados, removendo inconsistências, valores ausentes, ruídos e *outliers*, além de ajustar a escala e a distribuição das variáveis.

O tratamento de *outliers*, por exemplo, visa identificar e lidar com valores que se distanciam significativamente do comportamento geral da amostra, os quais podem distorcer o desempenho do modelo. Esses pontos podem ser removidos ou ajustados, dependendo do contexto e da natureza do dado. Outras etapas comuns incluem a normalização e padronização, que tornam as variáveis comparáveis entre si, evitando que atributos com escalas maiores dominem o processo de aprendizado.

Um dos desafios mais recorrentes em bases reais, especialmente em análise de crédito, é o desbalanceamento de classes. Esse desbalanceamento entre classes é comum em bases de dados de crédito, nas quais geralmente há uma proporção significativamente maior de registros pertencentes à classe majoritária, como clientes adimplentes, em comparação à classe minoritária, que representa os casos de inadimplência. Essa diferença pode introduzir viés no modelo, levando-o a favorecer as classes mais frequentes e comprometendo sua capacidade de generalização (OLIVEIRA, 2024).

Para lidar com esse problema, utilizam-se técnicas de reamostragem, divididas em duas categorias principais:

- **Sobreamostragem (*Oversampling*):** aumenta o número de amostras da classe minoritária, replicando dados existentes ou gerando novas instâncias sintéticas, como no método SMOTE (*Synthetic Minority Over-sampling Technique*) ou ADASYN (*Adaptive Synthetic Sampling*), que criam exemplos artificiais a partir dos vizinhos mais próximos (CAETANO, 2024).

- **Subamostragem (*Undersampling*):** reduz o número de amostras da classe majoritária, equilibrando a proporção entre as classes.

A escolha da técnica de balanceamento deve considerar as características da base de dados e o tipo de problema. Trabalhos recentes reforçam a importância dessa etapa, mostrando que o reequilíbrio entre as classes pode melhorar significativamente o desempenho e a estabilidade dos modelos preditivos. Por exemplo, Xianyu et al. (2023) aplicaram oversampling aleatório em um conjunto de dados desbalanceado, expandindo a base original de 4.293 para 8.100 amostras, o que contribuiu para uma melhor representação da classe minoritária (XIANYU; HAI, 2023). De forma semelhante, Caetano et al. (2024) empregaram os métodos SMOTE e ADASYN para gerar amostras sintéticas em cenários de previsão de inadimplência, obtendo resultados mais equilibrados e robustos (CAETANO, 2024).

A aplicação adequada dessas técnicas permite que o modelo aprenda de forma equilibrada os padrões das classes minoritária e majoritária, resultando em uma performance mais robusta e menos enviesada.

2.3 Aprendizado de Máquina

O Aprendizado de Máquina (AM) é um campo dentro da Inteligência Artificial que busca compreender como os computadores podem aprender ou melhorar seu desempenho a partir de dados. Por meio de algoritmos específicos, os sistemas conseguem identificar padrões, realizar previsões e tomar decisões com base em informações históricas ou em tempo real. Essa abordagem tem se mostrado essencial em diversas áreas, incluindo análise de crédito, detecção de fraudes, reconhecimento de padrões e sistemas de recomendação, permitindo que os computadores se tornem capazes de adaptar-se a novos dados e aprimorar continuamente suas previsões.

A seguir, são apresentados os principais tipos de AM (HAN; PEI; TONG, 2022), que se diferenciam pela forma como os dados e os rótulos são utilizados no processo de treinamento:

- **Aprendizado Supervisionado:** utiliza dados rotulados[2], em que o sistema aprende a associar entradas a saídas conhecidas.
- **Aprendizado Não Supervisionado:** trabalha com dados não rotulados, buscando identificar padrões ou agrupamentos ocultos. É útil para detectar anomalias ou fazer clusterização (divisão dos dados em grupos com base em similaridade).
- **Aprendizado por Reforço:** baseia-se na interação de um agente com o ambiente, em que o modelo aprende por meio de recompensas e penalidades.

2.3.1 Algoritmos de Aprendizado de Máquina

O aprendizado de máquina desempenha um papel central na previsão de risco e na construção de modelos de inadimplência, pois permite extrair padrões relevantes a partir de dados históricos e produzir decisões automatizadas baseadas nessas relações. O livro de Kelleher, Namee e D'Arcy (2020) oferece uma fundamentação teórica densa sobre o funcionamento do aprendizado supervisionado, descrevendo-o como um processo automatizado capaz de aprender a relação entre um conjunto de atributos descritivos e uma variável-alvo a partir de instâncias passadas. Segundo o autor, esse processo ocorre em duas etapas principais: a aprendizagem do modelo com base em exemplos históricos e a utilização desse modelo para realizar previsões em novos dados. O autor também ilustra, por meio de exemplos relacionados ao crédito, como a complexidade dos conjuntos de dados atuais torna inviável a criação manual de regras e reforça a necessidade de algoritmos capazes de lidar com múltiplas variáveis e grandes volumes de informação.

[2] Dados rotulados são aqueles em que cada entrada possui uma saída ou classe conhecida, usada para treinar o modelo.

Complementando essa visão teórica, estudos empíricos reforçam a aplicabilidade prática desses métodos na previsão de inadimplência. O trabalho de Xianyu e Hai (2023), por exemplo, desenvolveu um modelo de previsão de inadimplência corporativa utilizando técnicas de *big data* e comparando algoritmos como *Random Forest*, Regressão Logística e Redes Neurais. Os autores analisaram não apenas o desempenho, mas também a confiabilidade dos modelos ao longo de diferentes cenários, destacando a relevância da escolha adequada dos algoritmos para problemas de classificação binária no contexto financeiro.

Com base nesses referenciais, tanto teóricos quanto aplicados, foram identificados os algoritmos mais comuns utilizados em sistemas de análise de crédito, incluindo métodos supervisionados de classificação e regressão, além de técnicas não supervisionadas quando aplicáveis. Esses algoritmos estão sistematizados na Tabela 1, que apresenta os principais algoritmos de aprendizado de máquina empregados em estudos de risco de crédito, acompanhados de uma breve descrição de suas características essenciais.

Algoritmo	Descrição
Redes Neurais	Modelos inspirados no funcionamento do cérebro humano, capazes de aprender padrões complexos e generalizar para novas situações.
Regressão Linear e Logística	Regressão linear identifica correlações entre variáveis; regressão logística é usada para classificação binária, como adimplente/inadimplente.
Decision Tree (DT)	Estrutura em árvore que divide os dados em ramos baseados em regras de decisão, facilitando a interpretação.
Random Forest (RF)	Conjunto de árvores de decisão aplicando *ensemble*, reduzindo overfitting e aumentando a robustez.
Support Vector Machine (SVM)	Busca o hiperplano ótimo que separa classes, podendo usar funções kernel para dados não lineares.
XGBoost	Algoritmo de *boosting* baseado em árvores de decisão, eficiente e escalável para grandes volumes de dados.
AdaBoost	Combina múltiplos classificadores fracos em um classificador forte, aplicando maior peso a exemplos mal classificados.
Clustering K-means	Técnica não supervisionada que agrupa dados com base na similaridade, auxiliando na identificação de padrões subjacentes.

Tabela 1 – Principais algoritmos de AM utilizados em sistemas preditivos de crédito.

2.3.2 Separação da base de dados em conjuntos de treino e teste

Na construção de modelos de aprendizado de máquina, é fundamental avaliar a capacidade de generalização do modelo para novos dados. Para isso, a base de dados é tipicamente dividida em dois subconjuntos:

- **Conjunto de treino:** utilizado para ajustar os parâmetros do modelo e aprender os padrões presentes nos dados.

- **Conjunto de teste:** utilizado para avaliar o desempenho do modelo em dados independentes, não vistos durante o treinamento, garantindo uma avaliação imparcial da capacidade preditiva.

As proporções mais comuns para essa divisão são 60:40, 70:30 ou 80:20, sendo que a maior fração é destinada ao conjunto de treino (CAETANO, 2024). Essa escolha busca equilibrar a necessidade de um conjunto de treino suficientemente grande para aprendizagem com a necessidade de um conjunto de teste representativo para avaliação.

Em problemas de classificação, é importante garantir que a distribuição das classes seja preservada em ambos os conjuntos, evitando viés e distorções na avaliação. Técnicas de balanceamento de classes, como o *oversampling* ou SMOTE (*Synthetic Minority Over-sampling Technique*), podem ser aplicadas antes da divisão para reduzir efeitos de desbalanceamento (CAETANO, 2024).

Além disso, a utilização de validação cruzada permite testar o modelo em diferentes subconjuntos do conjunto de treino, aumentando a confiabilidade das métricas de desempenho e reduzindo o risco de *overfitting* (conceito detalhado na Seção 2.3.3).

Métricas como acurácia, *recall* e F1-score são comumente utilizadas para avaliar modelos de classificação, fornecendo uma visão abrangente da capacidade do modelo em identificar corretamente instâncias das diferentes classes.

2.3.3 Métricas de Avaliação

A avaliação de modelos de aprendizado de máquina é uma etapa essencial para verificar sua capacidade de generalização e desempenho. Um dos primeiros aspectos a serem observados durante essa análise é o *overfitting*. Esse fenômeno ocorre quando o modelo apresenta excelente desempenho nos dados de treino, mas resultados insatisfatórios nos dados de teste. Em outras palavras, o modelo memoriza os exemplos de treinamento em vez de aprender padrões gerais, comprometendo sua capacidade de fazer previsões sobre novos dados (OLIVEIRA, 2024).

A Figura 3 ilustra esse comportamento:

- As bolinhas vermelhas representam os dados de treino, ou seja, os exemplos que o modelo utilizou para aprender.
- As bolinhas azuis com contorno preto representam os dados de teste, exemplos que o modelo não viu durante o treino, usados para avaliar sua capacidade de generalização.
- A linha verde mostra um modelo que sofreu overfitting: ele se ajusta exatamente aos dados de treino, acertando quase todas as bolinhas vermelhas, mas erra nos dados de teste (bolinhas azuis).

- A linha preta representa um modelo regularizado (mais equilibrado): ele não acompanha todos os detalhes do treino e comete alguns erros nas bolinhas vermelhas, mas consegue prever melhor os dados de teste, mostrando que aprendeu padrões que funcionam para novos casos.

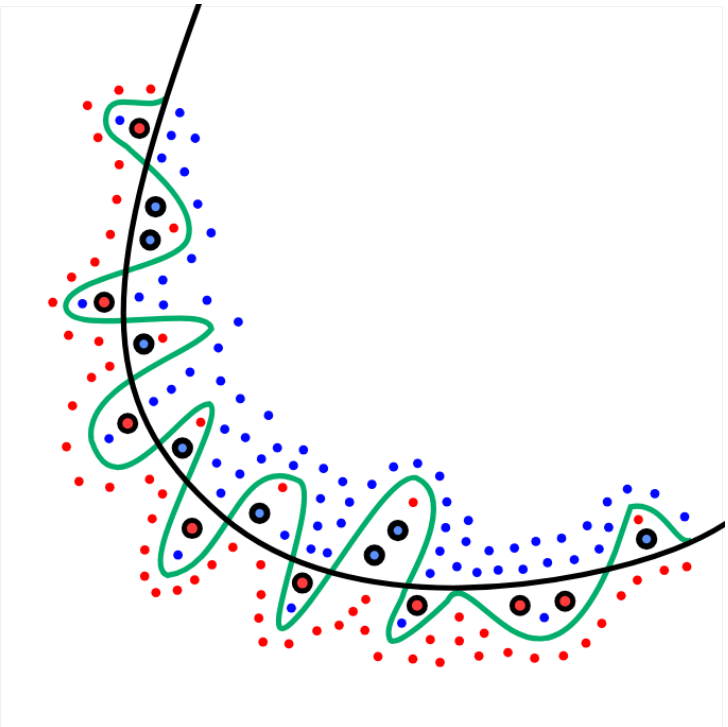

Figura 3 – Representação do fenômeno de *overfitting* Fonte: Wikipédia.

Para medir a performance de um modelo, é necessário utilizar métricas adequadas ao tipo de problema. No caso de problemas de classificação, como a previsão de inadimplência, a métrica mais simples e intuitiva é a acurácia, definida como a razão entre o número de previsões corretas e o total de observações:

$$\text{Acurácia} = \frac{\text{número de previsões corretas}}{\text{total de previsões}}$$

Por exemplo, se um modelo classifica corretamente 48 de 50 observações, sua acurácia é de 96%. Embora seja uma métrica bastante utilizada, a acurácia nem sempre é a melhor forma de avaliar um modelo, especialmente em conjuntos de dados desbalanceados. Considere, por exemplo, um problema de detecção de fraudes em que apenas uma pequena fração das transações é realmente fraudulenta. Um modelo que classifique quase todas as transações como legítimas pode apresentar uma acurácia alta, mesmo sendo ineficaz para o propósito principal. Em um cenário hipotético com 30 transações, sendo 3 fraudulentas, um modelo que acerte 28 casos legítimos e erre 2 das fraudes teria uma acurácia de 93,3%, mas deixaria de identificar a maioria das fraudes, o que é indesejável.

Nesses casos, métricas adicionais devem ser consideradas, como:

- **Sensibilidade (*Recall*):** mede a capacidade do modelo de identificar corretamente as instâncias positivas (ex.: inadimplentes). É a proporção de verdadeiros positivos entre todas as instâncias realmente positivas.

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

onde VP são os verdadeiros positivos e FN os falsos negativos.

- **Precisão (*Precision*):** avalia quantas das previsões positivas feitas pelo modelo estão corretas. É especialmente útil quando o custo de um falso positivo é alto.

$$\text{Precisão} = \frac{VP}{VP + FP}$$

- ***F1-Score:*** combina precisão e sensibilidade em uma única métrica harmônica, equilibrando os dois aspectos.

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

Conjuntos de dados desbalanceados exigem atenção especial na avaliação de modelos, pois a presença de classes minoritárias pode distorcer a interpretação de métricas como a acurácia. Estudos recentes mostram que a classificação multiclasse em cenários de forte desequilíbrio apresenta desafios adicionais, sobretudo no processo de reamostragem e na análise do desempenho, o que reforça a importância de escolher métricas adequadas ao problema (YANG; KHORSHIDI; AICKELIN, 2024). Estratégias de *oversampling*, incluindo abordagens métricas, adaptativas, estruturais e híbridas, podem melhorar a representação das classes minoritárias e, com isso, aumentar a confiabilidade de métricas como sensibilidade, precisão e *F1-score*.

Assim, a escolha adequada das métricas é fundamental para garantir que o modelo preditivo realmente cumpra seu propósito de identificar corretamente os casos de maior risco, considerando os desafios impostos pelo desequilíbrio de classes.

3 TRABALHOS RELACIONADOS

Diversos estudos têm explorado o uso do aprendizado de máquina em instituições financeiras, especialmente nas áreas de análise de crédito e previsão de inadimplência. Essas pesquisas servem como base conceitual e metodológica para o desenvolvimento deste projeto, contribuindo para a definição das técnicas e algoritmos empregados.

3.0.1 Inadimplência e Crédito

O estudo de Maia, Ramírez e Cutino (2025) foi uma das principais referências para compreender o cenário macroeconômico pós-pandemia e o aumento dos índices de inadimplência no Brasil. Os autores analisam a relação entre indicadores como IPCA, taxa SELIC e PIB mensal e o comportamento da inadimplência no período de 2021 a 2024, evidenciando que o fenômeno é sensível às oscilações da economia.

O trabalho demonstra que a inadimplência é um processo pró-cíclico, isto é, tende a crescer em momentos de expansão econômica e de aumento das taxas de juros. Essa conclusão destaca a importância de acompanhar variáveis macroeconômicas na formulação de políticas de crédito e na construção de modelos preditivos. Para este projeto, a pesquisa de Maia, Ramírez e Cutino (2025) fornece uma base teórica importante para compreender possíveis fatores que impactam a análise de risco de crédito, independentemente de sua eventual utilização como variáveis no modelo.

Outro estudo de destaque é o de Oliveira (2024), que propõe a utilização de aprendizado de máquina em um modelo de concessão de crédito pessoal, incorporando dados do Auxílio Emergencial concedido pelo Governo Federal entre abril e dezembro de 2020. O foco da pesquisa é avaliar como a inclusão dessas informações socioeconômicas pode impactar a precisão e a robustez dos modelos de previsão de risco de crédito.

A relevância do trabalho de Oliveira (2024) está na sua preocupação em atender um público de baixa renda, que frequentemente enfrenta barreiras para obter crédito em instituições tradicionais. Ao cruzar dados financeiros com variáveis sociais, o estudo conseguiu aprimorar significativamente o desempenho do modelo, demonstrando que o contexto econômico e social exerce influência direta sobre a capacidade de pagamento dos clientes.

Assim como no trabalho dela, este projeto também fará uso de dados abertos disponibilizados pelo governo, pela ampla quantidade de registros existentes e pela facilidade de acesso, o que reforça a transparência das informações e possibilita a reprodutibilidade e continuidade de estudos futuros. Além disso, a metodologia adotada pela autora, serve como referência para a estruturação deste estudo, especialmente nas etapas de pré-processamento, correção de desbalanceamento e avaliação dos modelos.

Esses trabalhos relacionados têm sido fundamentais para direcionar as escolhas metodológicas e técnicas deste projeto, tanto na seleção dos algoritmos quanto na definição do

tipo de dados a serem utilizados. Ao combinar perspectivas macroeconômicas e sociais, busca-se aqui desenvolver um modelo preditivo mais abrangente e contextualizado, capaz de auxiliar instituições financeiras na identificação de padrões de inadimplência com maior precisão.

3.0.2 Seleção de Algoritmos

Alguns trabalhos prévios foram fundamentais para embasar tanto a fundamentação teórica quanto a definição das técnicas a serem aplicadas neste projeto. O estudo de Oliveira (2024) utilizou dados abertos do governo brasileiro e apresentou um fluxo completo de pré-processamento, seleção de atributos e avaliação de modelos supervisionados, fazendo uso de métricas como acurácia, recall e F1-score. Esse trabalho serviu como referência inicial, especialmente pela forma estruturada de condução da análise preditiva.

Outro estudo relevante, conduzido por Xianyu e Hai (2023), abordou a previsão de inadimplência corporativa utilizando algoritmos como Random Forest, Regressão Logística, Redes Neurais Convolucionais (CNN) e Redes Neurais Recorrentes (RNN). Os autores destacaram a importância da divisão dos dados em treino e teste, além da necessidade de ajuste criterioso dos hiperparâmetros para aprimorar o desempenho. Os resultados indicaram que, embora o Random Forest apresentasse bom desempenho e a Regressão Logística fosse mais limitada, as RNNs obtiveram os melhores resultados em acurácia, *recall* e *F1-score*. O estudo também realizou análises de sensibilidade, avaliando a estabilidade dos modelos frente a alterações na base de dados, reforçando a importância de verificar a consistência das soluções propostas.

O trabalho de Caetano (2024), desenvolvido no contexto de um TCC em Estatística, também forneceu uma base metodológica relevante, uma vez que tratava de um problema semelhante, envolvendo classificação de clientes em adimplentes e inadimplentes por meio de algoritmos supervisionados de aprendizado de máquina. A estrutura do pipeline analítico, a preocupação com o tratamento do desbalanceamento e a avaliação comparativa de modelos foram especialmente úteis para o desenvolvimento deste estudo.

De modo geral, os trabalhos analisados apresentaram forte convergência metodológica: iniciam pela análise exploratória dos dados, seguem pela definição e preparação dos atributos, avançam para a seleção de algoritmos, frequentemente optando por modelos amplamente utilizados e com flexibilidade para manipulação de variáveis, e então realizam o treinamento e validação em fases offline e online, utilizando bases de treino e teste. Todos também reforçam a necessidade de tratar adequadamente conjuntos desbalanceados e a importância de selecionar algoritmos e métricas coerentes com as características dos dados e com os objetivos da predição.

A síntese desses estudos permitiu identificar boas práticas na escolha de algoritmos, nas etapas de pré-processamento e na condução da avaliação, servindo como referência para estruturar as etapas de processamento conceitual deste projeto e fundamentar a construção de modelos preditivos confiáveis e bem alinhados ao problema investigado.

4 MATERIAIS E MÉTODOS

Este capítulo descreve os recursos utilizados e os procedimentos previstos para a análise dos dados e o desenvolvimento do estudo. Na subseção Materiais, apresentam-se a base de dados, os atributos considerados e as ferramentas utilizadas. Na subseção Métodos, descrevem-se as etapas que compõem o processo de pré-processamento, modelagem, treinamento e avaliação dos modelos de aprendizado de máquina.

4.1 Materiais

4.1.1 Base de dados

O estudo considera uma base de dados pública que reúne informações de crédito de pessoas físicas e jurídicas, com o objetivo de identificar padrões associados à inadimplência. O recorte analisado corresponde a um mês de 2025, disponibilizado pelo portal de dados abertos do Banco Central do Brasil (Banco Central do Brasil, 2025).

Com base em uma análise preliminar da estrutura dos arquivos, estima-se que o segmento de pessoas físicas (PF) contenha aproximadamente 145 mil registros. Esse valor oferece uma referência inicial para compreensão do volume e da granularidade dos dados, podendo variar após a consolidação final da base. O conjunto disponibiliza informações referentes a diferentes perfis de tomadores, modalidades de crédito e limites contratados.

Cada registro reúne atributos relacionados ao perfil do cliente e às operações de crédito, incluindo ocupação, porte (faixa de rendimento), modalidade contratada, indexador, quantidade de operações e valores distribuídos em diferentes faixas de prazo. Também estão presentes indicadores associados ao risco, como saldo vencido acima de 15 dias, carteira ativa, carteira inadimplida arrastada e ativo problemático. A diversidade desses elementos permite explorar relações relevantes para a construção de modelos de previsão de inadimplência.

Além disso, o Banco Central do Brasil publica mensalmente, por meio do Sistema de Informações de Créditos (SCR), arquivos contendo dados agregados das operações de crédito no país. Os relatórios incluem informações segmentadas por tipo de cliente (PF/PJ), modalidade, porte, natureza da ocupação, origem dos recursos e indexador, com divulgação realizada com defasagem aproximada de 60 dias. Esse conjunto de dados forma a base estruturante utilizada no estudo.

4.1.2 Ferramentas e tecnologias

As ferramentas listadas a seguir representam as tecnologias que estão sendo consideradas neste momento para a manipulação, exploração e modelagem dos dados. No entanto,

conforme o desenvolvimento do projeto avançar, outras bibliotecas ou recursos poderão ser incorporados, caso se mostrem necessários ou mais adequados às etapas futuras.

- **Polars**: Biblioteca open-source para manipulação de dados com foco em desempenho. Segundo o site oficial, é uma das soluções mais rápidas para processamento em uma única máquina, e possui uma API tipada, bem estruturada, expressiva e de fácil uso (POLARS, 2025).
- **Pandas**: Biblioteca amplamente utilizada para análise de dados em Python, buscando fornecer uma base de alto nível para análises práticas em cenários reais (PANDAS, 2025). No contexto deste estudo, pode complementar o Polars em operações estatísticas, transformações específicas e integração com outras bibliotecas científicas.
- **Scikit-learn**: Biblioteca de código aberto amplamente utilizada para aprendizado de máquina em Python. Oferece suporte a técnicas supervisionadas e não supervisionadas, além de ferramentas para ajuste de modelos, pré-processamento de dados, seleção de modelos, avaliação e diversas utilidades associadas ao ciclo completo de modelagem (SCIKIT-LEARN, 2025).
- **Matplotlib**: Biblioteca abrangente para criação de visualizações estáticas, animadas e interativas em Python (MATPLOTLIB, 2025).
- **Seaborn**: Biblioteca de visualização construída sobre o Matplotlib, oferecendo uma interface de alto nível para a criação de gráficos estatísticos mais informativos e esteticamente refinados (SEABORN, 2025).

4.2 Métodos

4.2.1 Pré-processamento dos dados

O processo de pré-processamento inclui um conjunto de etapas voltadas à preparação da base para aplicação de técnicas de aprendizado de máquina. Entre as operações que podem vir a ser consideradas, destacam-se:

- **Tratamento de valores ausentes e outliers:** os dados passam por verificação de valores nulos, inconsistentes ou extremos. Estratégias de imputação ou remoção podem ser aplicadas conforme a análise exploratória.
- **Análise exploratória e visualização:** são obtidas estatísticas descritivas, distribuições de variáveis e proporções das classes, permitindo identificar padrões preliminares e possíveis desbalanceamentos.

- **Classificação binária:** cada registro recebe uma indicação de inadimplência ou não, de modo a possibilitar o uso de algoritmos supervisionados.
- **Balanceamento de dados:** técnicas como undersampling, oversampling ou SMOTE (*Synthetic Minority Over-sampling Technique*) podem ser empregadas para lidar com desequilíbrios entre classes.
- **Divisão em conjuntos de treino e teste:** os dados são organizados em subconjuntos destinados ao treinamento e à avaliação dos modelos, reduzindo riscos de *overfitting*.

4.2.1.1 Balanceamento da base de dados

A distribuição entre clientes adimplentes e inadimplentes tende a ser desbalanceada, o que pode influenciar negativamente o desempenho dos modelos. Para amenizar esse viés, o estudo considera diferentes estratégias de balanceamento, incluindo *undersampling*, *oversampling* e técnicas baseadas em geração sintética, como *SMOTE* (CAETANO, 2024). A escolha da abordagem mais adequada depende dos resultados observados durante a fase de experimentação.

4.2.2 Treinamento e validação do modelo

Modelos supervisionados de aprendizado de máquina são considerados para o problema de previsão de inadimplência. O processo metodológico envolve a divisão dos dados em conjuntos de treino e teste (Figura 4) e a experimentação com diferentes algoritmos, de modo a identificar configurações potencialmente adequadas ao contexto do estudo. Conforme ilustrado na figura, o fluxo de desenvolvimento é organizado em duas etapas complementares: a fase *offline*, dedicada à construção inicial do modelo, e a fase *online*, associada à sua aplicação contínua.

Na fase offline, o modelo é estruturado a partir de dados históricos, passando por etapas de pré-processamento que incluem padronização de variáveis, seleção de atributos, tratamento de valores ausentes e técnicas de balanceamento. Em seguida, diferentes algoritmos de aprendizado de máquina podem ser avaliados sob distintas combinações de parâmetros, permitindo identificar configurações que apresentem desempenho satisfatório nas métricas adotadas. Esse processo possui caráter iterativo, uma vez que ajustes sucessivos podem ser realizados conforme os resultados intermediários indicarem necessidade de refinamento.

A fase *online*, por sua vez, corresponde ao momento em que o modelo é aplicado a novos dados, mantendo a estrutura utilizada na fase de treinamento. As previsões geradas são então monitoradas por meio de métricas de desempenho, possibilitando acompanhar a estabilidade do modelo ao longo do tempo e identificar eventuais degradações em sua capacidade preditiva. Caso alterações significativas sejam observadas, seja no perfil dos dados, seja na

qualidade das previsões, o modelo poderá retornar à fase offline para ajustes ou readequações metodológicas.

Essa divisão entre etapas *offline* e *online* contribui para garantir tanto a qualidade inicial do modelo quanto sua capacidade de adaptação às mudanças que possam ocorrer nos padrões de comportamento da carteira de crédito.

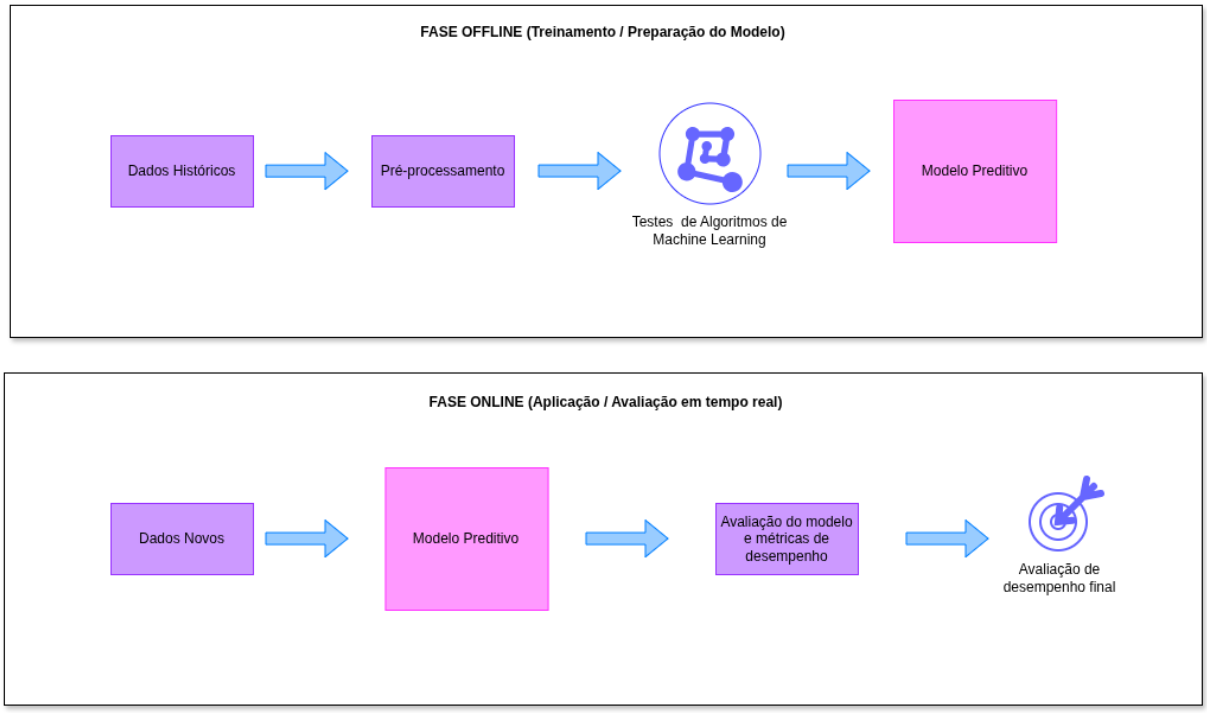

Figura 4 – Fluxo geral de treinamento e aplicação do modelo. Fonte: Autoria própria.

4.2.3 Métricas de avaliação

A definição das métricas de avaliação a serem utilizadas neste trabalho ainda depende da escolha final dos algoritmos de aprendizado de máquina que serão adotados, uma vez que diferentes técnicas podem demandar indicadores específicos para mensurar adequadamente seu desempenho. De modo geral, métricas como Acurácia, Precisão, Revocação (*Recall*), *F1-Score* e AUC-ROC são amplamente empregadas em estudos que envolvem classificação, pois permitem avaliar tanto a capacidade de separação entre as classes quanto o equilíbrio entre erros do tipo I e II (OLIVEIRA, 2024). Contudo, a seleção definitiva dessas métricas será realizada após a determinação dos modelos utilizados, garantindo maior adequação metodológica entre as técnicas escolhidas e os critérios de avaliação aplicados.

5 RESULTADOS PARCIAIS

Nesta seção, são apresentados os primeiros avanços obtidos ao longo do desenvolvimento do trabalho. A Tabela 2 descreve os principais atributos inicialmente considerados na análise preditiva, incluindo suas categorias e possíveis valores. A compreensão desses elementos é fundamental para contextualizar a estrutura preliminar dos modelos de aprendizado de máquina e orientar a investigação de padrões no período pós-concessão.

Conforme apresentado na Metodologia, o estudo utilizará uma base pública disponibilizada mensalmente pelo Banco Central do Brasil, contendo informações detalhadas das operações de crédito registradas no Sistema de Informações de Créditos (SCR). Embora os arquivos incluam dados de pessoas físicas (PF) e pessoas jurídicas (PJ), optou-se por concentrar a análise no segmento de pessoas físicas, a fim de manter um escopo metodologicamente controlado e coerente com os objetivos do estudo.

Neste momento, estão sendo considerados apenas os registros referentes a um mês de 2025, que incluem informações associadas ao perfil do cliente, às características das operações contratadas e a indicadores financeiros vinculados ao risco de crédito. Entre as variáveis disponíveis, encontram-se unidade da federação, ocupação, faixa de rendimento (porte), modalidade de crédito, indexador, número de operações e faixas de valores a vencer, além de indicadores como saldo vencido acima de 15 dias, carteira ativa, carteira inadimplida arrastada e ativo problemático.

As análises iniciais da base têm permitido identificar quais atributos tendem a apresentar maior potencial para compor os modelos preditivos, considerando critérios como completude, relevância e aderência ao objetivo da pesquisa. A partir dessa exploração preliminar, os atributos listados a seguir configuram, até o momento, os principais candidatos a serem utilizados nos experimentos, podendo ser ajustados conforme o avanço das etapas subsequentes.

Tabela 2 – Atributos utilizados para treinamento do modelo de aprendizado de máquina

Atributo	Categorias / Valores possíveis
Ocupação	PF - Aposentado/pensionista, PF - Autônomo, PF - Empregado de empresa privada, PF - Empregado de entidades sem fins lucrativos, PF - Empresário, PF - MEI, PF - Outros, PF - Servidor ou empregado público.
Porte	PF - Até 1 salário mínimo, PF - Mais de 1 a 2 salários mínimos, PF - Mais de 2 a 3 salários mínimos, PF - Mais de 3 a 5 salários mínimos, PF - Mais de 5 a 10 salários mínimos, PF - Mais de 10 a 20 salários mínimos, PF - Acima de 20 salários mínimos, PF - Sem rendimento, PF - Indisponível.
Modalidade	PF - Cartão de crédito, PF - Empréstimo com consignação em folha, PF - Habitacional, PF - Outros créditos, PF - Rural e agroindustrial, PF - Veículos.
Vencido Acima de 15 dias	Variável alvo do modelo. Registros com valor maior que 0 são codificados como 1 (inadimplentes), enquanto valores iguais a 0 permanecem como 0 (adimplentes).

5.1 Cronograma de atividades para o TCC II

O cronograma a seguir organiza as etapas previstas para o desenvolvimento do TCC II, considerando o retorno das atividades acadêmicas e a necessidade de avançar de forma gradual sobre coleta, modelagem, avaliação e redação final. Embora possa ser ajustado ao longo do semestre, ele estabelece uma linha de trabalho clara para garantir a conclusão de todas as fases técnicas e acadêmicas dentro do prazo oficial de defesa.

Tabela 3 – Cronograma de atividades para o TCC II

Período	Atividades
Nov–Dez/2025 (já realizadas)	Revisão bibliográfica complementar; exploração inicial dos dados públicos; identificação preliminar dos atributos relevantes; ajustes estruturais no texto produzido no TCC I.
Março	Consolidação e documentação da base de dados; limpeza, padronização e integração das variáveis; definição final dos modelos e métricas de avaliação.
Abril	Implementação dos modelos de aprendizado de máquina; experimentos iniciais; ajustes de hiperparâmetros; comparação sistemática entre algoritmos.
Maio	Avaliação final dos modelos; construção de visualizações, tabelas e descrições quantitativas; interpretação dos resultados alinhada ao problema de pesquisa.
Junho	Redação aprofundada da análise dos resultados; revisão metodológica; ajustes de coerência e clareza; normalização conforme ABNT.
Julho	Entrega da versão final; preparação da apresentação de defesa; realização da defesa do TCC II.

6 CONSIDERAÇÕES FINAIS

Este trabalho de conclusão de curso busca desenvolver uma abordagem analítica voltada ao período pós-concessão de crédito, utilizando técnicas de aprendizado de máquina aplicadas a bases de dados públicas. O objetivo é demonstrar, em um contexto acadêmico, como registros históricos podem ser explorados para identificar sinais relacionados ao aumento da inadimplência e apoiar estratégias de acompanhamento e recuperação.

A relevância do estudo se evidencia diante da crescente adoção de inteligência artificial no setor financeiro, especialmente para o monitoramento contínuo da carteira e aprimoramento da gestão de risco. Embora soluções semelhantes sejam mais frequentes em instituições de grande porte, ainda há espaço para metodologias acessíveis e transparentes que possam ser adaptadas a diferentes realidades organizacionais. Dessa forma, o trabalho busca oferecer um direcionamento inicial, ilustrando o potencial dessas técnicas no contexto pós-concessão.

Espera-se que o desenvolvimento resulte em um protótipo analítico capaz de apoiar a priorização de casos e a identificação preliminar de risco ao longo do ciclo pós-concessão, funcionando como um complemento às práticas já existentes. Embora o foco principal desta etapa esteja na modelagem e na análise, abre-se a possibilidade de, caso o cronograma permita, incorporar uma interface demonstrativa para fins de visualização e apresentação dos resultados. A intenção não é substituir processos, mas demonstrar caminhos possíveis para modernização e uso estratégico de informações disponíveis.

Entre as limitações previstas, destacam-se as características dos dados públicos utilizados e a necessidade de manter um escopo controlado para garantir coerência metodológica. Ainda assim, acredita-se que o estudo pode gerar contribuições relevantes ao evidenciar como bases abertas podem apoiar análises importantes no acompanhamento da carteira, além de servir como referência acadêmica e ponto de partida para desenvolvimentos futuros.

REFERÊNCIAS

Banco Central do Brasil. **Sistema de Informações de Crédito - Dados Abertos**. 2025. Acesso em: 13 out. 2025. Disponível em: https://dadosabertos.bcb.gov.br/dataset/scr_data.

Banco Central do Brasil, C. . **Resolução CMN n.º5.037, de 29 de setembro de 2022 – Dispõe sobre o Sistema de Informações de Crédito (SCR)**. 2022. Disponível em: https://www.bcb.gov.br/estabilidadefinanceira/exibenormativo?tipo=Resolu%C3%A7%C3%A3o%20CMN&numero=5037. Acesso em: 23 out. 2025.

CAETANO, T. M. Trabalho de Conclusão de Curso (TCC), **Algoritmos de aprendizado de máquina no estudo da inadimplência em uma instituição financeira**. Universidade Federal de Uberlândia, 2024. Disponível em: https://repositorio.ufu.br/handle/123456789/41843.

EXPERIAN, S. **Mapa da Inadimplência e Negociação de Dívidas no Brasil**. 2025. Acesso em: 17 ago. 2025. Disponível em: https://www.serasa.com.br/limpa-nome-online/blog/mapa-da-inadimplencia-e-renogociacao-de-dividas-no-brasil/.

HAN, J.; PEI, J.; TONG, H. **Data Mining: Concepts and Techniques**. Morgan Kaufmann, 2022. (The Morgan Kaufmann Series in Data Management Systems). ISBN 9780128117613. Disponível em: https://books.google.com.br/books?id=NR1oEAAAQBAJ.

KELLEHER, J.; NAMEE, B.; D'ARCY, A. **Fundamentals of Machine Learning for Predictive Data Analytics, second edition: Algorithms, Worked Examples, and Case Studies**. MIT Press, 2020. ISBN 9780262361101. Disponível em: https://books.google.com.br/books?id=UM_tDwAAQBAJ.

MAIA, A. C.; RAMíREZ, Y. S.; CUTINO, Y. P. Influência dos indicadores macroeconômicos na inadimplência de pessoas físicas no brasil. **REVISTA DELOS**, v. 18, n. 67, p. e5138, maio 2025. Disponível em: https://ojs.revistadelos.com/ojs/index.php/delos/article/view/5138.

MATPLOTLIB. **Matplotlib: Visualization with Python**. 2025. Disponível em: https://matplotlib.org/stable/. Acesso em: 15 de novembro de 2025.

OLIVEIRA, L. T. d. S. d. **O uso de técnicas de aprendizado de máquina para concessão de crédito: um estudo a partir do portal de dados abertos brasileiro**. 2024. Dissertação (Dissertação de Mestrado) — Universidade de Brasília, 2024. Disponível em: http://repositorio.unb.br/handle/10482/52675.

PANDAS. **Pandas Documentation: Getting Started Overview**. 2025. Disponível em: https://pandas.pydata.org/docs/getting_started/overview.html. Acesso em: 15 de novembro de 2025.

POLARS. **Polars: DataFrames for the New Era**. 2025. https://pola.rs/. Acesso em: 15 de Novembro de 2025.

SCIKIT-LEARN. **Scikit-learn: Machine Learning in Python – Getting Started**. 2025. Disponível em: https://scikit-learn.org/stable/getting_started.html. Acesso em: 15 de novembro de 2025.

SEABORN. **Seaborn: Statistical Data Visualization**. 2025. Disponível em: https://seaborn.pydata.org/. Acesso em: 15 de novembro de 2025.

SERASA. **O que é birô de crédito e como esse tipo de empresa funciona**. 2024. Blog Serasa. Acesso em: 23 out. 2025. Disponível em: https://www.serasa.com.br/score/blog/o-que-e-biro-de-credito-e-como-esse-tipo-de-empresa-funciona/.

XIANYU, Q.; HAI, M. Research on default prediction model of corporate credit risk based on big data analysis algorithm. **Procedia Computer Science**, v. 221, p. 300–307, 2023. ISSN 1877-0509. Tenth International Conference on Information Technology and Quantitative Management (ITQM 2023). Disponível em: https://www.sciencedirect.com/science/article/pii/S1877050923007378.

YANG, Y.; KHORSHIDI, H. A.; AICKELIN, U. A review on over-sampling techniques in classification of multi-class imbalanced datasets: insights for medical problems. **Frontiers in Digital Health**, Volume 6 - 2024, 2024. ISSN 2673-253X. Disponível em: https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2024.1430245.