**UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ**

**MILENA HAMERSKI**

# ANÁLISE E PREDIÇÃO DE INADIMPLÊNCIA PARA APOIO À DECISÃO NO PÓS-CONCESSÃO DE CRÉDITO

**GUARAPUAVA**

**2026**

**MILENA HAMERSKI**

# ANÁLISE E PREDIÇÃO DE INADIMPLÊNCIA PARA APOIO À DECISÃO NO PÓS-CONCESSÃO DE CRÉDITO

## Default Analysis and Prediction for Decision Support in Post-Credit Granting

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

Coorientador: Profª. Me. Dênis Lucas Silva

**GUARAPUAVA**

**2026**

## AGRADECIMENTOS

A conclusão deste trabalho representa também o encerramento de uma etapa muito importante da minha vida. Por isso, gostaria de expressar minha sincera gratidão a todas as pessoas que, de diferentes formas, contribuíram para que eu chegasse até aqui.

Agradeço aos meus pais, Edson Hamerski e Angela Hamerski, pelo amor, apoio incondicional e por sempre acreditarem em mim. Obrigada por valorizarem a educação e por terem me ensinado, desde cedo, a importância do conhecimento. Sou grata por todos os esforços, renúncias e investimentos feitos ao longo da minha vida para que eu tivesse oportunidades que me permitiram chegar até aqui. Esta conquista também é de vocês.

À minha irmã, Flavia Hamerski, agradeço pela companhia, pelo carinho e por estar presente ao longo dessa jornada, compartilhando momentos de alegria e também os desafios que fizeram parte deste caminho.

À minha orientadora, Prof.ª Dra. Kelly Lais Wiggers, minha profunda gratidão por todo o apoio, dedicação e orientação. Mesmo diante de importantes mudanças em sua trajetória profissional e de um momento tão especial em sua vida, sua gestação, permaneceu comprometida com este trabalho até o final, contribuindo de forma decisiva para sua realização.

Ao meu coorientador, Prof. Me. Dênis Lucas Silva, agradeço por ter aceitado dar continuidade à orientação deste projeto e por toda a disponibilidade, atenção e contribuição durante sua etapa final.

Aos meus amigos, que me ouviram nos momentos de cansaço, ansiedade e preocupação, agradeço pela paciência, pelo apoio e pelas palavras de incentivo. Compartilhar os desafios desta fase tornou o caminho mais leve e possível.

Agradeço também à Universidade Tecnológica Federal do Paraná (UTFPR), instituição que me proporcionou uma formação de excelência. Tenho muito orgulho de ter estudado em uma universidade pública, gratuita e de qualidade, que transforma vidas por meio da educação.

Meu agradecimento a todos os professores que fizeram parte da minha trajetória acadêmica, compartilhando conhecimentos, experiências e ensinamentos que vão muito além da sala de aula. Estendo minha gratidão aos servidores da universidade, que desempenham um papel essencial para o funcionamento da instituição e para a formação de seus estudantes.

Por fim, agradeço a todos que, de alguma forma, contribuíram para esta conquista. Cada palavra de incentivo, cada gesto de apoio e cada aprendizado fizeram parte desta caminhada e estarão sempre presentes na minha memória.

## RESUMO

Este trabalho apresenta uma abordagem voltada à análise e à predição de inadimplência no período pós-concessão de crédito, utilizando técnicas de aprendizado de máquina aplicadas a dados públicos do Sistema de Informações de Crédito (SCR). O objetivo consiste em investigar a capacidade de modelos preditivos de identificar padrões associados ao risco de inadimplência, contribuindo para o acompanhamento de operações de crédito após sua concessão. Para a realização do estudo, a base de dados foi submetida a etapas de pré-processamento, incluindo tratamento de valores ausentes, redução de alta cardinalidade categórica, balanceamento de classes e transformação de variáveis. Na etapa de modelagem, foram treinados e avaliados diferentes algoritmos de classificação, entre eles Regressão Logística, Árvore de Decisão, KNN, Random Forest, SVM e XGBoost, utilizando métricas como Acurácia, F1-Score e ROC AUC. Os resultados demonstraram que os modelos foram capazes de identificar padrões relacionados à inadimplência, apresentando desempenho satisfatório para a tarefa de classificação e evidenciando o potencial das técnicas de aprendizado de máquina na análise de risco de crédito. Como complemento ao estudo, foi desenvolvida uma interface web simplificada para ilustrar uma possível aplicação prática dos modelos gerados. Conclui-se que a abordagem proposta pode contribuir para atividades de monitoramento e análise de risco no contexto pós-concessão de crédito, fornecendo subsídios para a identificação antecipada de operações com maior probabilidade de inadimplência.



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

**ABSTRACT**

This study presents an approach focused on default analysis and prediction in the post-credit granting period, using machine learning techniques applied to public data from the Credit Information System (SCR). The objective is to investigate the ability of predictive models to identify patterns associated with default risk, contributing to the monitoring of credit operations after their approval. To conduct the study, the dataset underwent several preprocessing stages, including missing value treatment, reduction of high-cardinality categorical variables, class balancing, and feature transformation. During the modeling phase, different classification algorithms were trained and evaluated, including Logistic Regression, Decision Tree, K-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (SVM), and XGBoost, using Accuracy, F1-Score, and ROC AUC as evaluation metrics. The results demonstrated that the models were capable of identifying patterns related to default, achieving satisfactory performance in the classification task and highlighting the potential of machine learning techniques for credit risk analysis. As a complement to the study, a simplified web interface was developed to illustrate a possible practical application of the generated models. It is concluded that the proposed approach can contribute to monitoring and risk analysis activities in the post-credit granting context, providing support for the early identification of operations with a higher probability of default.

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

## LISTA DE FIGURAS

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

### Siglas

| | |
|---|---|
| ADASYN | Adaptive Synthetic Sampling (Técnica de Amostragem Sintética Adaptativa para a Classe Minoritária) |
| AM | Aprendizado de Máquina |
| BCB | Banco Central do Brasil |
| CMN | Conselho Monetário Nacional |
| CNN | Convolutional Neural Network |
| FN | Falso Negativo |
| FP | Falso Positivo |
| IPCA | Índice Nacional de Preços ao Consumidor Amplo |
| LGPD | Lei Geral de Proteção de Dados Pessoais |
| MD | Mineração de Dados |
| PIB | Produto Interno Bruto |
| RNN | Recurrent Neural Network |
| SCR | Sistema de Informações de Crédito |
| SELIC | Sistema Especial de Liquidação e Custódia (taxa básica de juros da economia brasileira) |
| SMOTE | Synthetic Minority Over-sampling Technique (Técnica Sintética de Sobreamostragem da Classe Minoritária) |
| SPC | Serviço de Proteção ao Crédito (SPC Brasil) |
| SVM | Support Vector Machine |
| TCC | Trabalho de Conclusão de Curso |
| VN | Verdadeiro Negativo |
| VP | Verdadeiro Positivo |

**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), no Índice Nacional de Preços ao Consumidor Amplo (IPCA), no desemprego e na taxa básica de juros do Sistema Especial de Liquidação e Custódia (taxa básica de juros da economia brasileira) (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 (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 relevância de abordagens analíticas voltadas à avaliação e compreensão 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 compreender de forma mais estruturada os fatores associados ao risco de crédito.

A inteligência artificial pode ser entendida como o conjunto de técnicas computacionais capazes de aprender padrões a partir de dados e realizar previsões ou inferências 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 (MD) e Aprendizado de Máquina (AM) possibilita construir modelos preditivos para avaliação do risco de crédito, com potencial de maior consistência em relação a abordagens tradicionais (XIANYU; HAI, 2023). Essa aplicação reforça o uso de

[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 análise.

modelos estatísticos e computacionais como ferramentas para análise de risco no contexto de crédito.

O interesse pela aplicação da inteligência artificial nesse cenário decorre do potencial dessas técnicas em transformar grandes volumes de dados em análises consistentes capazes de apoiar a compreensão do risco de crédito no período pós-concessão. Atualmente, muitas análises ainda dependem fortemente do julgamento humano do analista, o que pode envolver subjetividade e variações entre avaliadores. Nesse contexto, os modelos preditivos atuam como ferramentas complementares, contribuindo para uma análise mais estruturada e baseada em dados.

A escolha pelo período pós-concessão decorre das características da base de dados utilizada neste estudo. Os dados analisados são provenientes de registros históricos de operações de crédito já contratadas e reportadas ao Sistema de Informações de Crédito (SCR), representando situações em que a concessão do crédito já ocorreu. Dessa forma, o foco da pesquisa não está na decisão de aprovar ou rejeitar uma solicitação de crédito, mas sim na análise do comportamento das operações ao longo do tempo e na identificação de padrões associados à inadimplência após a concessão. Espera-se que os conhecimentos obtidos possam futuramente contribuir para o desenvolvimento de ferramentas de apoio ao monitoramento da carteira, à priorização de estratégias de acompanhamento e recuperação de crédito e à identificação antecipada de operações com maior propensão ao aumento do risco, possibilitando intervenções mais tempestivas por parte das instituições financeiras.

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 no pós-concessão de crédito, contribuindo para análises mais assertivas e consistentes. Essa abordagem dialoga com a necessidade de maior eficiência e confiabilidade na análise de risco em um ambiente financeiro que demanda constante adaptação às transformações tecnológicas.

## 1.2 Objetivos

### 1.2.1 Objetivo geral

Desenvolver e avaliar modelos de aprendizado de máquina para análise e predição de inadimplência no período pós-concessão de crédito para Pessoa Física, utilizando dados financeiros históricos como suporte à tomada de decisão.

### 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 predição de inadimplência e suas aplicações no contexto de risco de crédito.
- Selecionar e preparar um conjunto de dados financeiros, realizando limpeza, transformação e tratamento das variáveis relevantes para a modelagem.
- Treinar e avaliar modelos de classificação para predição de inadimplência, utilizando métricas adequadas para análise de desempenho.
- Comparar os resultados obtidos pelos modelos, identificando suas capacidades e limitações na identificação de padrões associados à inadimplência.

## 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 analítica e estratégica.

Informações como principalidade[2], faixa de risco, capacidade de pagamento e existência de avais, entre outras, constituem insumos relevantes para observar tendências, compreender o comportamento dos clientes após a liberação do crédito e apoiar a análise de risco no período pós-concessão.

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 mitigação de risco de crédito.

A execução completa do processo de análise de dados, envolvendo limpeza, integração, seleção, transformação, mineração de dados e avaliação de padrões, evidencia o potencial dessas técnicas para apoiar a análise de risco de crédito, contribuindo para a compreensão

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

de comportamentos associados à inadimplência e para a geração de insumos analíticos que auxiliem a tomada de decisão no contexto pós-concessão.

A crescente demanda por maior agilidade e padronização nas atividades relacionadas ao monitoramento da carteira torna a adoção de técnicas analíticas ainda mais pertinente. Mesmo utilizando dados públicos, abordagens baseadas em aprendizado de máquina contribuem para reduzir subjetividade, acelerar o tempo de resposta e aumentar a consistência das análises. Além disso, modelos preditivos podem auxiliar na detecção precoce de indícios de maior probabilidade de inadimplência, favorecendo que estratégias de acompanhamento e recuperação sejam direcionadas de maneira mais eficiente e tempestiva.

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. Nesse contexto, o uso de modelos analíticos baseados em dados abertos pode contribuir para uma gestão de risco mais eficiente e preventiva, promovendo maior estabilidade e confiabilidade no ambiente de crédito.

# 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 Serviço de Proteção ao Crédito (SPC Brasil) (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 Pessoais (LGPD).

Além desses birôs privados, existe o Sistema de Informações de Crédito (SCR), gerido pelo Banco Central do Brasil (BCB). 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, incluindo empréstimos, repasses interfinanceiros, coobrigações e limites de crédito (Conselho Monetário Nacional; Banco Central do Brasil, 2022). O principal objetivo do SCR é permitir o monitoramento prudencial das instituições financeiras, ajudando o BCB 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 (Conselho Monetário Nacional; 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 (Sistema de Informações de Crédito), 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 (Conselho Monetário Nacional; Banco Central do Brasil, 2022).

## 2.2 Mineração de Dados

A mineração de dados (MD), no contexto deste trabalho aplicado à análise de inadimplência no período pós-concessão de crédito, 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 (AM), 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, no contexto da análise de risco de crédito, 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 análises mais seguras e embasadas. 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.

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

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

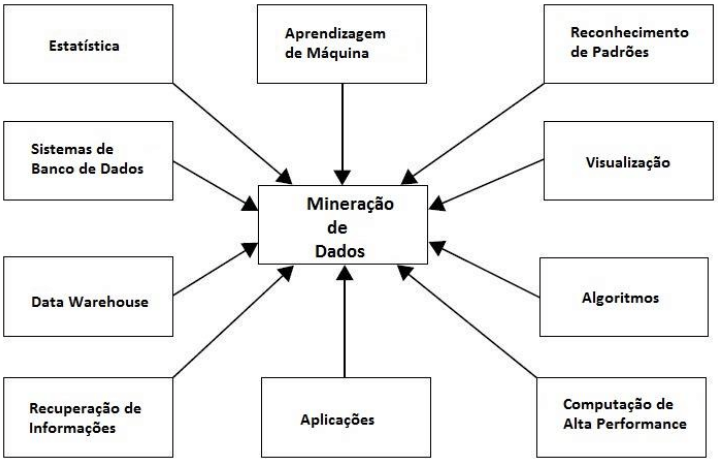



**Figura 1 – Multidisciplinaridade da Mineração de Dados. Fonte: Research Gate**

O processo de MD, no contexto de problemas como a predição de inadimplência, 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 AM, 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 na análise de risco de crédito e na compreensão de padrões associados ao comportamento de inadimplência.

Vale ressaltar que esse processo é cíclico, ou seja, existe um ciclo de vida dos dados. No contexto da análise de crédito, é fundamental que os dados sejam continuamente atualizados e refinados, de modo a manter a relevância e a precisão das análises e previsões geradas. A Figura 2 apresenta uma visão geral das principais etapas envolvidas no processo de mineração de dados, desde a obtenção e preparação dos dados até a geração de conhecimento para apoio à análise de risco.


**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, no contexto da predição de inadimplência, é essencial garantir a qualidade e a representatividade dos dados. O pré-processamento consiste em um conjunto de técnicas voltadas à preparação dos dados, removendo inconsistências, valores ausentes, ruídos e *outliers*[2], 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 a 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 problemas de análise de crédito, é o desbalanceamento de classes. Esse desbalanceamento 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 por meio da replicação de dados existentes ou da geração de novas instâncias sintéticas. Entre as técnicas mais utilizadas estão Synthetic Minority Over-sampling Technique (Técnica Sintética de Sobreamostragem da Classe Minoritária) (SMOTE) e Adaptive Synthetic Sampling (Técnica de Amostragem Sintética Adaptativa para a Classe Minoritária) (ADASYN), que geram exemplos artificiais com base nas características das amostras existentes (CAETANO, 2024).

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

A escolha da técnica de balanceamento deve considerar as características da base de dados e o tipo de problema de predição de inadimplência. 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

[2] Valores atípicos que se distanciam significativamente do comportamento geral dos dados.

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 (2024) empregou 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 na identificação de risco de crédito.

## 2.3 Aprendizado de Máquina

O aprendizado de máquina (AM), no contexto deste trabalho, é um campo 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 apoiar a análise de risco com base em informações históricas. 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 modelos se adaptem a novos dados e aprimorem 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, sendo amplamente aplicados em problemas de classificação como a predição de inadimplência:

- **Aprendizado Supervisionado:** utiliza dados rotulados[3], sendo amplamente utilizado em problemas de classificação de crédito, como a identificação de clientes adimplentes ou inadimplentes.
- **Aprendizado Não Supervisionado:** trabalha com dados não rotulados, buscando identificar padrões ou agrupamentos ocultos. No contexto da análise de crédito, pode ser utilizado para segmentação de clientes ou detecção de comportamentos atípicos.
- **Aprendizado por Reforço:** baseia-se na interação de um agente com o ambiente, em que o modelo aprende por meio de recompensas e penalidades, sendo mais comum em problemas de decisão sequencial e otimização de estratégias.

### 2.3.1 Algoritmos de Aprendizado de Máquina

No contexto da predição de inadimplência, o aprendizado de máquina desempenha um papel central na construção de modelos capazes de identificar padrões relevantes a partir de dados históricos e apoiar a análise de risco. O livro de Kelleher, Namee e D'Arcy (2020)

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

oferece uma fundamentação teórica sobre o funcionamento do aprendizado supervisionado, descrevendo-o como um processo automatizado capaz de aprender a relação entre atributos 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 destaca que, em problemas como análise de crédito, a complexidade dos dados torna inviável a criação manual de regras, reforçando a necessidade de algoritmos capazes de lidar com múltiplas variáveis e grandes volumes de informação.

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 em diferentes cenários, destacando a relevância da escolha adequada dos algoritmos para problemas de classificação no contexto de risco de crédito.

Com base nesses referenciais, tanto teóricos quanto aplicados, foram identificados os algoritmos mais comuns utilizados em problemas de análise de crédito e predição de inadimplência, 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.

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

Na construção de modelos de AM aplicados à predição de inadimplência, 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 históricos de crédito.
- **Conjunto de teste:** utilizado para avaliar o desempenho do modelo em dados independentes, não vistos durante o treinamento, garantindo uma avaliação mais realista 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 do comportamento do modelo em dados de crédito não observados.

| Algoritmo | Descrição |
|---|---|
| **Convolutional Neural Network (CNN) e Recurrent Neural Network (RNN)** | Modelos inspirados no funcionamento do cérebro humano, capazes de aprender padrões complexos e generalizar para novas situações. No contexto de análise de crédito, podem ser aplicados em dados sequenciais ou não estruturados, embora não sejam os mais comuns em bases tabulares de inadimplência. |
| **Regressão Linear e Logística** | A regressão linear identifica relações entre variáveis; a regressão logística é amplamente utilizada em problemas de classificação binária, como a previsão de inadimplência. |
| **Decision Tree (DT)** | Estrutura em árvore que divide os dados em ramos baseados em regras de decisão, sendo amplamente utilizada em problemas de crédito devido à sua interpretabilidade. |
| **Random Forest (RF)** | Conjunto de árvores de decisão aplicando técnicas de *ensemble*, reduzindo *overfitting* e aumentando a robustez, sendo amplamente utilizado em problemas de risco de crédito. |
| **Support Vector Machine (SVM)** | Busca o hiperplano ótimo que separa classes, podendo utilizar funções *kernel* para lidar com dados não lineares, sendo aplicada em problemas de classificação como inadimplência. |
| **XGBoost** | Algoritmo de *boosting* baseado em árvores de decisão, eficiente e escalável para grandes volumes de dados, amplamente utilizado em modelos preditivos de risco de crédito. |
| **AdaBoost** | Combina múltiplos classificadores fracos em um classificador forte, atribuindo maior peso a exemplos classificados incorretamente, sendo aplicado em tarefas de classificação binária como inadimplência. |
| **Clustering K-means** | Técnica não supervisionada que agrupa dados com base na similaridade, auxiliando na identificação de perfis de clientes e padrões de comportamento em análise de crédito. |

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

Em problemas de classificação, especialmente na predição de inadimplência, 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, podem ser aplicadas antes da divisão para reduzir efeitos de desbalanceamento (CAETANO, 2024).

Além da divisão simples em treino e teste, uma abordagem amplamente utilizada para avaliação mais robusta de modelos de aprendizado de máquina é a validação cruzada. Esse método consiste em dividir o conjunto de dados em múltiplos subconjuntos (folds), permitindo que o modelo seja treinado e testado em diferentes combinações de dados. Dessa forma, obtém-se uma estimativa mais estável e confiável do desempenho do modelo na predição de inadimplência, reduzindo o risco de dependência de uma única divisão dos dados.

A validação cruzada do tipo *k-fold*, por exemplo, divide os dados em $k$ partes iguais, treinando o modelo em $k-1$ folds e validando em um fold diferente a cada iteração. De acordo com Caetano (2024), essa abordagem, comumente utilizando $k = 10$, permite uma avaliação mais consistente do desempenho do modelo, mitigando o risco de sobreajuste aos dados de treinamento e aumentando a confiabilidade dos resultados obtidos (CAETANO, 2024).

No contexto deste trabalho, a validação cruzada é utilizada como estratégia complementar à divisão em treino e teste, contribuindo para uma avaliação mais robusta dos modelos de aprendizado de máquina aplicados à análise de risco de crédito. Essa abordagem permite comparar o desempenho dos algoritmos sob diferentes particionamentos dos dados, aumentando a confiabilidade das métricas obtidas e reduzindo a variabilidade dos resultados.

### 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:

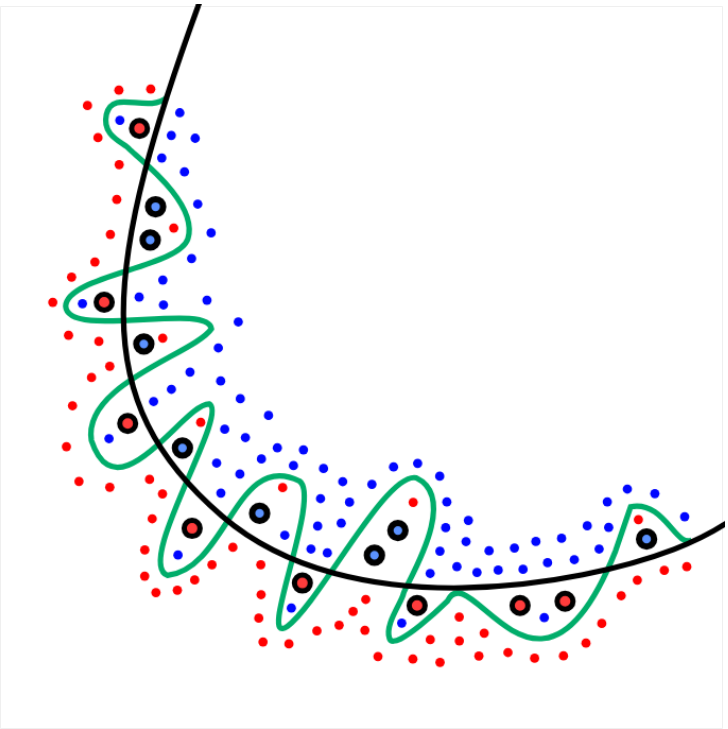


**Figura 3 – Representação do fenômeno de *overfitting*. Fonte: Wikipédia.**

- As bolinhas vermelhas representam os dados de treino, utilizados pelo modelo para aprendizado.
- As bolinhas azuis com contorno preto representam os dados de teste, utilizados para avaliação da generalização.

- A linha verde representa um modelo que sofreu *overfitting*, ajustando-se excessivamente aos dados de treino e apresentando pior desempenho no teste.
- A linha preta representa um modelo mais equilibrado, com melhor capacidade de generalização.

Para medir o desempenho de um modelo de classificação, são utilizadas métricas adequadas ao tipo de problema. No caso da previsão de inadimplência, além da acurácia, são utilizadas métricas que consideram o desempenho em relação à classe de interesse, especialmente em cenários de desbalanceamento.

A acurácia é definida como a proporção de previsões corretas em relação ao total de observações:

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

Embora seja amplamente utilizada, a acurácia pode ser enganosa em bases desbalanceadas, pois um modelo pode obter alto desempenho apenas ao prever corretamente a classe majoritária.

Nesse contexto, são utilizadas métricas baseadas na matriz de confusão, composta por verdadeiros positivos (Verdadeiro Positivo (VP)), verdadeiros negativos (Verdadeiro Negativo (VN)), falsos positivos (Falso Positivo (FP)) e falsos negativos (Falso Negativo (FN)).

As principais métricas utilizadas são:

- **Sensibilidade (Recall):** mede a proporção de casos positivos corretamente identificados pelo modelo.

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

- **Precisão (Precision):** indica a proporção de previsões positivas que estão corretas.

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

- **F1-score:** média harmônica entre precisão e recall.

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- **ROC AUC (Receiver Operating Characteristic – Area Under the Curve):** avalia a capacidade do modelo em distinguir entre as classes ao longo de diferentes limiares de decisão. A curva ROC é construída a partir da taxa de verdadeiros positivos (TPR) e da taxa de falsos positivos (FPR). Quanto mais próximo de 1 o valor da AUC, maior

a capacidade discriminatória do modelo, enquanto valores próximos de 0,5 indicam desempenho equivalente a uma classificação aleatória.

$$TPR = \frac{VP}{VP + FN} \qquad FPR = \frac{FP}{FP + VN}$$

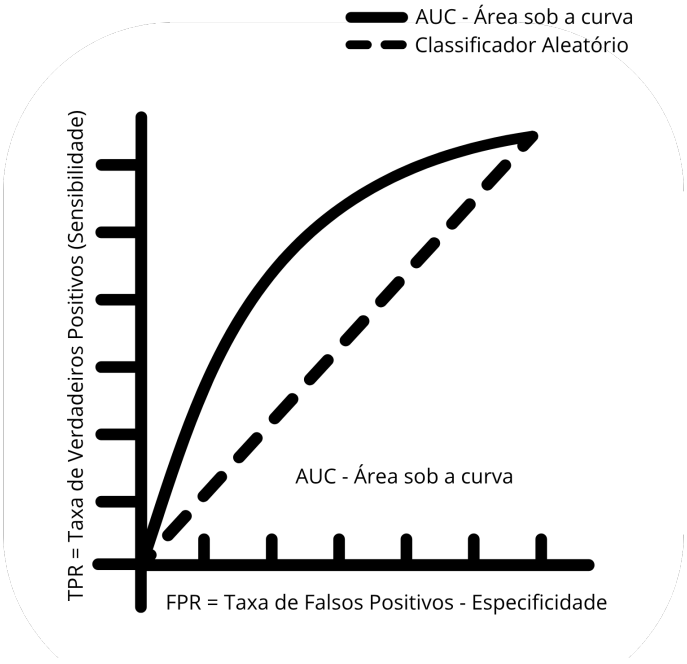



**Figura 4 – Exemplo de curva ROC. Fonte: Elaboração própria.**

A Figura 4 apresenta uma representação da curva ROC. O eixo vertical corresponde à taxa de verdadeiros positivos (TPR), também conhecida como sensibilidade, enquanto o eixo horizontal representa a taxa de falsos positivos (FPR). Cada ponto da curva corresponde a um limiar de decisão diferente utilizado pelo classificador (Google Developers, 2026).

A interpretação da curva ocorre da seguinte forma: quanto mais próxima ela estiver do canto superior esquerdo do gráfico, melhor será o desempenho do modelo, pois isso indica uma elevada taxa de acertos com uma baixa taxa de erros. Em contrapartida, uma curva próxima da diagonal principal representa um classificador com desempenho semelhante ao aleatório, sem capacidade significativa de distinção entre as classes (Google Developers, 2026).

A área sob a curva, conhecida como AUC (*Area Under the Curve*), resume o desempenho do modelo em um único valor numérico. Valores próximos de 1 indicam excelente capacidade de separação entre as classes, enquanto valores próximos de 0,5 sugerem desempenho equivalente a uma classificação aleatória (Google Developers, 2026). Dessa forma, a métrica ROC AUC é amplamente utilizada na avaliação de modelos de classificação, especialmente em problemas com desbalanceamento de classes.

Estudos recentes indicam que, em problemas com desbalanceamento de classes, a utilização conjunta de múltiplas métricas é necessária para uma avaliação mais adequada do desempenho dos modelos (YANG; KHORSHIDI; AICKELIN, 2024).

Dessa forma, a combinação entre acurácia, F1-score e ROC AUC permite uma avaliação consistente da capacidade preditiva dos modelos no contexto deste estudo.

# 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 neste trabalho de predição de inadimplência no contexto pós-concessão de crédito.

## 3.1 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 instabilidade econômica e de aumento das taxas de juros. Essa conclusão destaca a importância de acompanhar variáveis macroeconômicas na formulação de políticas de crédito e no desenvolvimento de modelos preditivos de risco, como o proposto neste trabalho.

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 com a inclusão de perfis historicamente menos favorecidos no sistema financeiro. 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 faz uso de dados abertos disponibilizados por instituições públicas, o que reforça a transparência das informações e possibilita a reprodutibilidade dos experimentos. Além disso, a metodologia adotada serve como referência para a estruturação deste estudo, especialmente nas etapas de pré-processamento, tratamento de desbalanceamento e avaliação de modelos de aprendizado de máquina aplicados à inadimplência.

Esses trabalhos relacionados são fundamentais para compreender o contexto em que este estudo se insere, tanto sob a perspectiva econômica quanto sob a perspectiva metodo-

lógica. Em conjunto, eles reforçam a importância de modelos preditivos capazes de apoiar a análise de risco de crédito no período pós-concessão.

## 3.2 Seleção de Algoritmos

Alguns trabalhos prévios foram fundamentais para embasar tanto a fundamentação teórica quanto a definição das técnicas aplicadas neste projeto de predição de inadimplência.

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 pela forma estruturada de condução da análise preditiva aplicada ao risco de crédito.

Outro estudo relevante, conduzido por Xianyu e Hai (2023), abordou a previsão de inadimplência corporativa utilizando algoritmos como Random Forest, Regressão Logística, CNN e RNN. Os autores destacaram a importância da divisão dos dados em treino e teste, além do ajuste de hiperparâmetros para melhorar o desempenho dos modelos. Os resultados indicaram variações significativas entre os algoritmos, reforçando que a escolha do modelo impacta diretamente a qualidade da previsão de risco.

O trabalho de Caetano (2024), desenvolvido no contexto de um Trabalho de Conclusão de Curso (TCC) em Estatística, também forneceu uma base metodológica relevante, uma vez que trata de um problema semelhante de classificação de clientes adimplentes e inadimplentes. A estrutura do pipeline analítico, o tratamento do desbalanceamento e a comparação entre modelos supervisionados foram especialmente úteis como referência para este estudo.

De modo geral, os trabalhos analisados apresentam convergência metodológica, iniciando pela análise exploratória dos dados, passando pela preparação e seleção de atributos, e avançando para o treinamento e avaliação de modelos supervisionados. Também é recorrente a preocupação com o desbalanceamento das classes e com a escolha de métricas adequadas para avaliação do desempenho.

A síntese dessas abordagens permite identificar boas práticas que fundamentam a construção do modelo proposto neste trabalho, especialmente no que se refere ao pré-processamento, seleção de algoritmos e validação de desempenho em problemas de predição de inadimplência.

## 3.3 Síntese Comparativa dos Trabalhos

A Tabela 2 apresenta uma síntese comparativa dos principais trabalhos relacionados, destacando o tipo de dados utilizados, os algoritmos aplicados e o foco principal de cada estudo.

| Trabalho | Dados | Algoritmos | Foco |
|---|---|---|---|
| Oliveira (2024) | Dados públicos governamentais | Regressão Logística, RF | Crédito pessoal e análise socioeconômica |
| Xianyu et al. (2023) | Dados corporativos | RF, Regressão Logística, CNN, RNN | Inadimplência corporativa |
| Caetano (2024) | Base de crédito simulada / acadêmica | RF, DT, SMOTE/ADASYN | Classificação de crédito e balanceamento |

**Tabela 2 – Comparação entre trabalhos relacionados.**

A partir da análise comparativa, observa-se que os estudos apresentam forte convergência na utilização de técnicas de aprendizado de máquina aplicadas ao problema de risco de crédito, variando principalmente quanto ao tipo de dado e ao nível de complexidade das abordagens adotadas.

Este trabalho diferencia-se por concentrar a análise no contexto pós-concessão de crédito, utilizando dados públicos do Sistema de Informações de Crédito (SCR), com foco na identificação de padrões associados ao comportamento de inadimplência após a liberação do crédito. Essa abordagem busca contribuir para uma visão complementar às análises tradicionalmente concentradas na etapa de concessão inicial.

# 4 MATERIAIS E MÉTODOS

Este capítulo apresenta os materiais e os procedimentos metodológicos adotados para o desenvolvimento da pesquisa. Inicialmente, descrevem-se a base de dados e as ferramentas computacionais empregadas. Em seguida, apresentam-se as técnicas utilizadas para preparação dos dados, modelagem, treinamento e avaliação dos modelos de aprendizado de máquina.

## 4.1 Materiais

### 4.1.1 Base de dados

A base de dados utilizada neste trabalho é proveniente do Sistema de Informações de Crédito (SCR), mantido pelo Banco Central do Brasil (Banco Central do Brasil, 2025). O SCR constitui o principal sistema nacional de registro das operações de crédito realizadas por instituições financeiras autorizadas a operar no país, reunindo informações relacionadas a contratos, modalidades de crédito, perfil dos tomadores e indicadores associados ao risco das operações.

Os dados são disponibilizados periodicamente por meio do Portal de Dados Abertos do Banco Central em formato tabular, possibilitando sua utilização em estudos estatísticos, econômicos e computacionais relacionados à análise de crédito e ao comportamento financeiro dos clientes.

Neste estudo, foram considerados exclusivamente registros de pessoas físicas, uma vez que o objetivo da pesquisa consiste na construção de modelos para previsão de inadimplência nesse segmento. A base analisada contém aproximadamente 180 mil registros após filtragem, representando operações de crédito observadas em diferentes períodos.

Cada registro da base corresponde a um conjunto de informações agregadas sobre operações de crédito associadas a um cliente. Dessa forma, a estrutura dos dados é organizada em formato tabular, contendo atributos que representam características do tomador e das operações de crédito.

Entre os atributos disponíveis encontram-se variáveis relacionadas ao perfil do cliente, como ocupação e porte, além de variáveis associadas às operações, como modalidade e submodalidade de crédito, indexador, quantidade de operações e indicadores financeiros.

No conjunto de variáveis numéricas, destacam-se informações como valores de carteira ativa, montantes a vencer, valores vencidos, carteira inadimplida arrastada e ativo problemático. Esses atributos fornecem uma visão quantitativa do comportamento das operações de crédito ao longo do tempo.

Os atributos presentes na base possuem diferentes naturezas, incluindo variáveis categóricas nominais, ordinais e variáveis numéricas contínuas. Essa heterogeneidade exige a

aplicação de técnicas de pré-processamento específicas antes da utilização de algoritmos de aprendizado de máquina.

A combinação entre variáveis comportamentais e financeiras torna a base adequada para o desenvolvimento de modelos preditivos de inadimplência, permitindo a exploração de padrões associados ao risco de crédito no período pós-concessão.

### 4.1.2 Ferramentas e tecnologias

As ferramentas utilizadas neste trabalho compõem o ecossistema Python para ciência de dados e aprendizado de máquina, sendo empregadas em todas as etapas do pipeline experimental, desde a preparação dos dados até a avaliação dos modelos.

### 4.1.3 Ferramentas e tecnologias

As ferramentas utilizadas neste trabalho compõem o ecossistema Python para ciência de dados e aprendizado de máquina, sendo empregadas em todas as etapas do pipeline experimental, desde a preparação dos dados até a avaliação dos modelos.

- **Python**: linguagem de programação utilizada como base para o desenvolvimento de todo o pipeline de análise e modelagem (Python Software Foundation, 2025).
- **Polars**: biblioteca utilizada para manipulação e processamento eficiente de dados tabulares, especialmente em operações de leitura, filtragem e transformação de grandes volumes de dados (Polars Developers, 2025).
- **Pandas**: biblioteca utilizada para manipulação, limpeza e estruturação de dados durante as etapas de pré-processamento e análise exploratória (Pandas Developers, 2025).
- **Scikit-learn**: biblioteca empregada na implementação, treinamento, validação cruzada, otimização de hiperparâmetros e avaliação dos modelos de aprendizado de máquina (Scikit-learn Developers, 2025).
- **XGBoost**: biblioteca utilizada para implementação do algoritmo *Extreme Gradient Boosting*, aplicado aos experimentos de classificação supervisionada (XGBoost Developers, 2025).
- **NumPy**: biblioteca utilizada para operações matemáticas, manipulação de matrizes e suporte computacional às etapas de modelagem (NumPy Developers, 2025).
- **Matplotlib**: biblioteca utilizada para geração de gráficos e visualizações dos resultados experimentais (Matplotlib Developers, 2025).

- **Seaborn**: biblioteca utilizada para visualização estatística e apoio à análise exploratória dos dados (Seaborn Developers, 2025).
- **Jupyter Notebook**: ambiente interativo utilizado para desenvolvimento, execução e documentação dos experimentos computacionais (Project Jupyter, 2025).
- **Git**: sistema de controle de versão utilizado para gerenciamento e rastreamento das alterações realizadas no código-fonte do projeto (Git Project, 2025).
- **GitHub**: plataforma utilizada para hospedagem e versionamento do código-fonte desenvolvido neste trabalho (GitHub, Inc., 2025).
- **Repositório do projeto**: o código-fonte, scripts de análise e artefatos desenvolvidos neste trabalho foram disponibilizados publicamente para fins de transparência e reprodutibilidade (HAMERSKI, 2026).

## 4.2 Métodos

Esta seção apresenta o procedimento metodológico adotado para a construção e avaliação dos modelos de aprendizado de máquina aplicados à previsão de inadimplência no contexto pós-concessão de crédito. O processo, ilustrado na Figura 5, foi estruturado de forma a garantir a reprodutibilidade dos experimentos, contemplando etapas de preparação dos dados, modelagem e avaliação dos resultados.

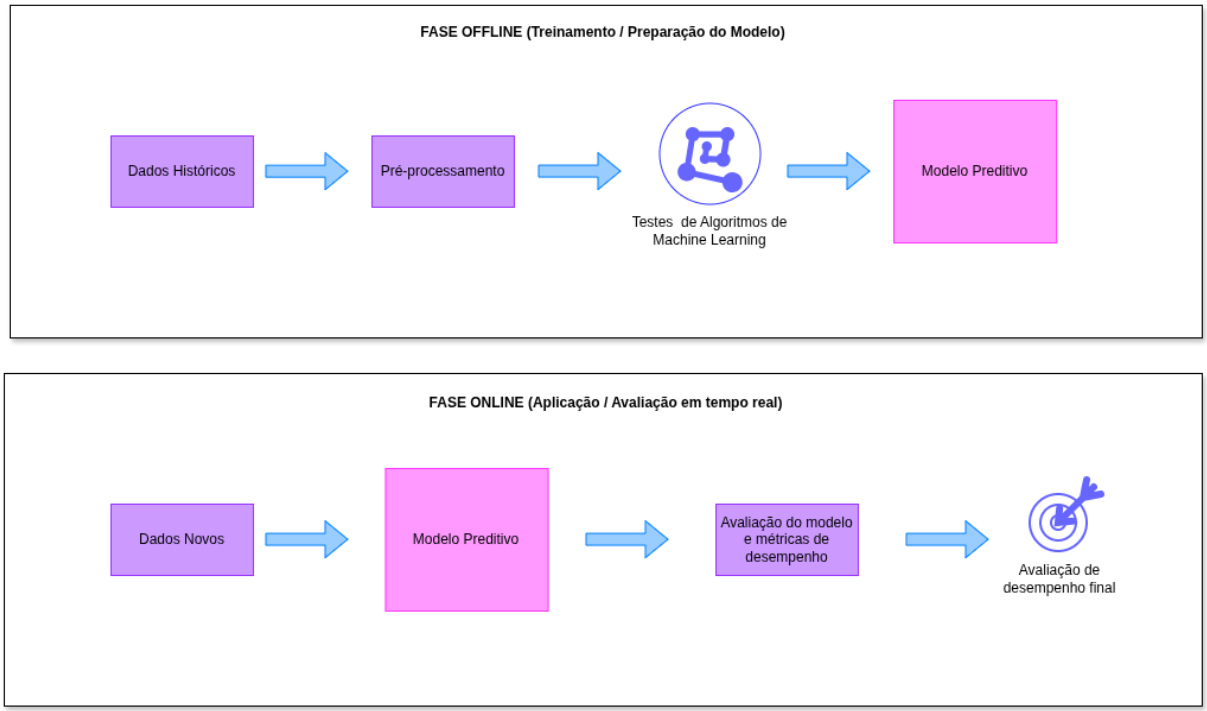



**Figura 5 – Fluxo geral de treinamento e avaliação dos modelos no contexto da análise de risco de crédito. Fonte: Autoria própria.**

O presente estudo adota uma abordagem baseada em aprendizado de máquina supervisionado para a construção de modelos preditivos de inadimplência em operações de crédito. O objetivo central consiste em identificar padrões presentes em dados históricos do Sistema de Informações de Crédito (SCR), permitindo distinguir clientes adimplentes e inadimplentes a partir de atributos relacionados ao perfil dos tomadores e às características das operações de crédito no período pós-concessão.

O processo metodológico não se limita a uma sequência técnica isolada, mas é estruturado como um fluxo integrado de análise de dados aplicado ao problema de risco de crédito. Inicialmente, os dados passam por etapas de limpeza, integração e padronização, garantindo consistência para sua utilização em algoritmos de aprendizado de máquina.

Em seguida, as variáveis categóricas são transformadas em representações numéricas, possibilitando sua utilização pelos modelos computacionais. Essa etapa é essencial no contexto da análise de crédito, uma vez que grande parte das informações disponíveis no SCR e em bases financeiras contém atributos qualitativos relacionados ao perfil dos clientes e das operações.

Após essa preparação, os dados são organizados em conjuntos de treinamento e teste, permitindo a construção e avaliação dos modelos de forma controlada e reprodutível. Essa separação é fundamental para simular o comportamento dos algoritmos em dados de crédito não observados, aproximando a avaliação do cenário real de análise de risco.

Durante o desenvolvimento, foram aplicadas técnicas de aprendizado supervisionado e diferentes algoritmos de classificação amplamente utilizados na literatura de análise de crédito. A definição final dos modelos, bem como sua parametrização e comparação, é detalhada na seção de experimentos.

O fluxo metodológico também considera a necessidade de evitar problemas como sobreajuste e vazamento de dados, especialmente em cenários envolvendo bases desbalanceadas e múltiplas transformações. Dessa forma, as etapas de pré-processamento e modelagem são estruturadas de maneira integrada, garantindo consistência entre as fases de treinamento e avaliação dos modelos.

Por fim, o desempenho dos modelos é avaliado com base em métricas apropriadas para problemas de classificação binária, permitindo a análise comparativa da capacidade preditiva dos diferentes algoritmos no contexto de risco de crédito e suporte à identificação de inadimplência no período pós-concessão.

### 4.2.1 Pré-processamento dos dados

A etapa de pré-processamento teve como objetivo limpar, transformar e estruturar a base de dados bruta utilizada neste estudo, viabilizando a aplicação dos algoritmos de aprendizado de máquina no contexto da previsão de inadimplência. Todos os procedimentos foram implementados em Python, compondo o pipeline analítico dos experimentos.

Inicialmente, a base foi filtrada para conter apenas registros de Pessoas Físicas (PF), uma vez que o foco da análise está na modelagem do risco de crédito nesse segmento. Em seguida, definiu-se a variável-alvo a partir da coluna `carteira_vencida`, transformada em uma variável binária: valores maiores que zero foram classificados como inadimplentes (classe 1) e os demais como adimplentes (classe 0), caracterizando um problema de classificação binária.

Para garantir a consistência dos atributos numéricos, realizou-se a padronização de formatos, incluindo a substituição de vírgulas por pontos em variáveis financeiras, permitindo sua correta interpretação como valores do tipo *float*.

O tratamento de valores ausentes e inconsistentes foi realizado de forma específica por variável:

- Nas variáveis categóricas `porte` e `submodalidade`, valores ausentes foram substituídos por "Indisponível"e "desconhecido", respectivamente.
- Na variável `numero_de_operacoes`, valores negativos foram tratados como inconsistências, sendo registrados por meio de uma variável indicadora (`operacoes_missing`) e posteriormente imputados pela mediana da distribuição.

Devido à alta cardinalidade da variável categórica `submodalidade`, foram definidos três cenários experimentais para avaliação de seu impacto no desempenho dos modelos: (I) remoção da variável; (II) agrupamento por frequência, mantendo categorias com mais de 1.000 ocorrências e agrupando as demais em "Outros"; e (III) agrupamento baseado em regras de negócio, com consolidação das submodalidades em macrogrupos como crédito pessoal, cartão de crédito e financiamento habitacional.

As variáveis categóricas remanescentes foram transformadas em variáveis numéricas por meio da técnica de *One-Hot Encoding*, que representa cada categoria como uma variável binária (0 ou 1).

Por fim, os dados foram divididos em conjuntos de treino (70%) e teste (30%), utilizando amostragem estratificada para preservar a proporção das classes. O tratamento de desbalanceamento por meio da técnica SMOTE (Synthetic Minority Over-sampling Technique) (CAETANO, 2024; YANG; KHORSHIDI; AICKELIN, 2024) foi aplicado exclusivamente no conjunto de treinamento dentro dos fluxos de validação, evitando vazamento de informação entre treino e teste.

### 4.2.2 Modelagem

A etapa de modelagem foi estruturada por meio da utilização de *pipelines* computacionais, integrando de forma sequencial as etapas de transformação dos dados e o treinamento dos modelos de aprendizado de máquina. Essa abordagem permitiu organizar o fluxo experimental de maneira modular, garantindo consistência entre as operações aplicadas aos dados.

Os modelos supervisionados foram treinados com o objetivo de aprender a relação entre as variáveis explicativas e a variável-alvo definida como inadimplência, caracterizando o problema como uma tarefa de classificação binária.

A utilização de *pipelines* assegurou que todas as transformações aplicadas aos dados de treinamento fossem reproduzidas de forma idêntica nos dados de teste, preservando a integridade do processo experimental e evitando inconsistências entre as fases de treinamento e avaliação.

Além disso, essa estrutura foi fundamental para evitar problemas de vazamento de dados (*data leakage*), especialmente em etapas que dependem da distribuição dos dados, como normalização, imputação de valores e técnicas de balanceamento.

Dessa forma, o pipeline de modelagem permitiu a padronização do processo de treinamento dos algoritmos avaliados, servindo como base para a comparação consistente entre os diferentes modelos de aprendizado de máquina utilizados neste estudo.

### 4.2.3 Estratégia de aprendizado supervisionado

O problema de previsão de inadimplência foi formulado como uma tarefa de aprendizado supervisionado do tipo classificação binária, na qual o objetivo consiste em estimar a ocorrência ou não de inadimplência a partir de variáveis explicativas relacionadas ao perfil dos clientes e às características das operações de crédito.

Nesse contexto, cada observação da base de dados é representada por um vetor de atributos preditivos associado a um rótulo de classe, permitindo que os algoritmos aprendam padrões a partir de dados históricos e utilizem essas relações para realizar previsões sobre novos registros não observados durante o treinamento.

A adoção dessa abordagem é particularmente adequada ao problema em análise, uma vez que o risco de crédito envolve relações complexas entre variáveis financeiras, comportamentais e operacionais, muitas vezes não lineares, que podem ser capturadas por modelos de aprendizado supervisionado.

Um aspecto relevante neste cenário é o desbalanceamento entre classes, dado que a proporção de casos de inadimplência é significativamente menor em relação aos casos de adimplência (CAETANO, 2024; BERRADA; BARRAMOU; ALAMI, 2024). Esse fator impacta diretamente o processo de aprendizado dos modelos, podendo levar a viés na predição da classe majoritária, o que torna necessária a adoção de estratégias específicas de tratamento e avaliação.

Dessa forma, o aprendizado supervisionado foi utilizado como base metodológica para a construção dos modelos preditivos, permitindo a análise sistemática do comportamento dos algoritmos no contexto de análise de risco de crédito.

### 4.2.4 Algoritmos de aprendizado de máquina

Para a construção dos modelos neste estudo, foram considerados diferentes algoritmos de aprendizado de máquina supervisionado amplamente utilizados em problemas de classificação (OLIVEIRA, 2024). A utilização de múltiplas abordagens tem como objetivo permitir a comparação de diferentes famílias de modelos no contexto da previsão de inadimplência, abrangendo métodos lineares, baseados em instâncias, baseados em árvores e técnicas de ensemble.

Foram avaliados os seguintes algoritmos:

A Regressão Logística, que modela a probabilidade de ocorrência de inadimplência por meio de uma combinação linear das variáveis explicativas, sendo amplamente utilizada como modelo de referência em problemas de classificação binária.

As Árvores de Decisão, que realizam particionamentos sucessivos do espaço de atributos com base em critérios de impureza, permitindo a construção de modelos interpretáveis e capazes de capturar relações não lineares.

O K-Vizinhos Mais Próximos (KNN), um método baseado em instâncias que classifica novas observações a partir da similaridade com exemplos presentes na base de treinamento.

O Random Forest, um método ensemble baseado na agregação de múltiplas árvores de decisão treinadas de forma independente, o que contribui para a redução da variância e aumento da estabilidade do modelo.

As Máquinas de Vetores de Suporte (SVM), que buscam encontrar um hiperplano ótimo de separação entre as classes, maximizando a margem entre as observações de diferentes categorias.

O XGBoost, um algoritmo de boosting baseado na construção sequencial de modelos, no qual cada nova árvore é treinada para corrigir os erros das anteriores, resultando em alto desempenho em dados estruturados.

A combinação desses algoritmos permite uma análise comparativa abrangente entre diferentes estratégias de aprendizado supervisionado aplicadas ao problema de risco de crédito.

### 4.2.5 Métricas de avaliação

A avaliação dos modelos foi conduzida por meio de métricas apropriadas para problemas de classificação binária, considerando especialmente o desbalanceamento das classes presente no contexto de análise de crédito.

A acurácia foi utilizada como métrica de referência geral, por medir a proporção total de acertos do modelo. No entanto, devido à sua limitação em cenários desbalanceados, ela não foi utilizada de forma isolada na análise dos resultados.

O F1-Score foi adotado como métrica complementar por equilibrar precisão e recall, sendo especialmente relevante em problemas nos quais a identificação correta da classe minoritária é importante.

A métrica ROC AUC foi utilizada como principal critério de comparação entre os modelos, por avaliar a capacidade discriminativa independentemente do limiar de decisão e apresentar menor sensibilidade ao desbalanceamento entre classes.

A utilização conjunta dessas métricas permite uma avaliação mais consistente do desempenho dos modelos, possibilitando uma comparação equilibrada entre as diferentes abordagens testadas neste estudo.

# 5 EXPERIMENTOS

O objetivo dos experimentos consiste em avaliar a capacidade de diferentes algoritmos de aprendizado de máquina em prever situações de inadimplência a partir dos dados disponibilizados pelo Sistema de Informações de Crédito (SCR). Para isso, foi desenvolvido um conjunto estruturado de experimentos capaz de investigar tanto o comportamento dos classificadores quanto o impacto de diferentes estratégias de representação dos atributos utilizados durante a modelagem.

A condução experimental foi organizada em quatro etapas principais. Inicialmente realizou-se a preparação e transformação dos dados, visando adequar a base às técnicas de aprendizado supervisionado. Em seguida, foram definidos cenários experimentais distintos para análise da variável submodalidade, atributo considerado potencialmente relevante para a caracterização das operações de crédito. Posteriormente, os modelos foram treinados utilizando diferentes estratégias de balanceamento e validação. Por fim, os resultados produzidos foram comparados por meio de métricas de desempenho adequadas ao problema de classificação binária.

## 5.1 Preparação e Transformação dos Dados

A base de dados original obtida a partir do Sistema de Informações de Crédito (SCR) continha 310.504 registros. Como o objetivo deste estudo consiste na previsão de inadimplência para pessoas físicas, foi aplicado um filtro sobre o atributo referente ao tipo de cliente, mantendo exclusivamente registros classificados como Pessoa Física (PF). Após essa etapa, a base passou a conter 182.631 registros, correspondendo a uma redução de 41,18% em relação ao volume inicial.

A partir da seleção dos registros, foi realizada uma etapa de preparação dos dados com o objetivo de garantir consistência, qualidade e adequação às técnicas de aprendizado de máquina empregadas nos experimentos.

Para a construção dos modelos, foi inicialmente selecionado um conjunto de variáveis provenientes do SCR relacionadas ao perfil do cliente e às características das operações de crédito. Essas variáveis correspondem aos atributos originais da base de dados, antes das etapas de transformação, agrupamento de categorias e criação dos diferentes cenários experimentais descritos na Seção **??**. As variáveis selecionadas estão apresentadas na Tabela 3. O código completo empregado na seleção e preparação dessas variáveis encontra-se disponível no Apêndice correspondente.

Inicialmente, foi definida a variável-alvo utilizada na tarefa de classificação. A coluna `carteira_vencida`, originalmente armazenada em formato textual, foi convertida para representação numérica. Em seguida, foi criada uma variável binária, na qual registros com valor

| Variável | Descrição |
|---|---|
| uf | Unidade federativa associada ao cliente. |
| cliente | Tipo de cliente registrado na operação. |
| cnae_ocupacao | Classificação CNAE ou ocupação principal do cliente. |
| porte | Porte do cliente informado na base. |
| modalidade | Modalidade da operação de crédito. |
| submodalidade | Submodalidade da operação de crédito. |
| numero_de_operacoes | Quantidade de operações de crédito associadas ao registro. |
| a_vencer_ate_90_dias | Valor da carteira a vencer em até 90 dias. |
| a_vencer_de_91_ate_360_dias | Valor da carteira a vencer entre 91 e 360 dias. |
| a_vencer_de_361_ate_1080_dias | Valor da carteira a vencer entre 361 e 1080 dias. |
| a_vencer_de_1081_ate_1800_dias | Valor da carteira a vencer entre 1081 e 1800 dias. |
| a_vencer_de_1801_ate_5400_dias | Valor da carteira a vencer entre 1801 e 5400 dias. |
| a_vencer_acima_de_5400_dias | Valor da carteira a vencer acima de 5400 dias. |
| carteira_a_vencer | Valor total da carteira a vencer. |
| carteira_vencida | Variável utilizada para construção da variável-alvo do problema. |

**Tabela 3 – Variáveis originais selecionadas para a modelagem antes das etapas de transformação e construção dos cenários experimentais.**

superior a zero foram classificados como inadimplentes (classe 1), enquanto registros com valor igual a zero foram classificados como adimplentes (classe 0).

Também foram realizados procedimentos de tratamento de valores ausentes. Na variável porte, os registros sem informação foram substituídos pelo rótulo *"Indisponível"*. Para a variável submodalidade, os valores ausentes foram substituídos pelo rótulo *"desconhecido"*. Esse procedimento permitiu preservar informações potencialmente relevantes sem a necessidade de descarte de observações.

A variável numero_de_operacoes exigiu um tratamento específico devido à presença de inconsistências nos dados originais. Registros contendo a expressão textual $\leq 15$ foram convertidos para o valor numérico 15. Valores negativos foram considerados inconsistências e substituídos por valores ausentes. Adicionalmente, foi criada uma variável indicadora para registrar a ocorrência dessas ausências, e os valores faltantes resultantes foram imputados utilizando a mediana da distribuição do atributo.

As variáveis financeiras também passaram por procedimentos de padronização e conversão de tipos, incluindo a substituição de separadores decimais e a transformação para formato numérico. Para evitar problemas computacionais em operações matemáticas posteriores, valores iguais a zero na variável carteira_a_vencer foram substituídos por 1 quando utilizados como denominadores em cálculos derivados.

Por fim, as variáveis categóricas utilizadas pelos modelos foram transformadas por meio da técnica *One-Hot Encoding*, que converte categorias textuais em variáveis binárias, permitindo sua utilização pelos algoritmos de aprendizado de máquina avaliados neste trabalho.

As etapas de balanceamento de classes e padronização de atributos não foram aplicadas globalmente à base de dados. Esses procedimentos foram executados exclusivamente dentro das dobras de treinamento durante a validação cruzada, evitando vazamento de informação (*data leakage*) e garantindo maior confiabilidade na avaliação dos modelos.

A Figura 6 apresenta uma visão geral do pipeline experimental adotado neste estudo. Após a etapa de limpeza e preparação dos dados, foram construídos três cenários distintos para o tratamento da variável `submodalidade`: remoção da variável, agrupamento por frequência e agrupamento baseado em regras de negócio. Em seguida, cada cenário passou pelas etapas de codificação dos atributos, divisão em conjuntos de treinamento e teste, otimização de hiperparâmetros por meio de validação cruzada e *Grid Search*, treinamento dos modelos e avaliação final utilizando métricas de classificação.

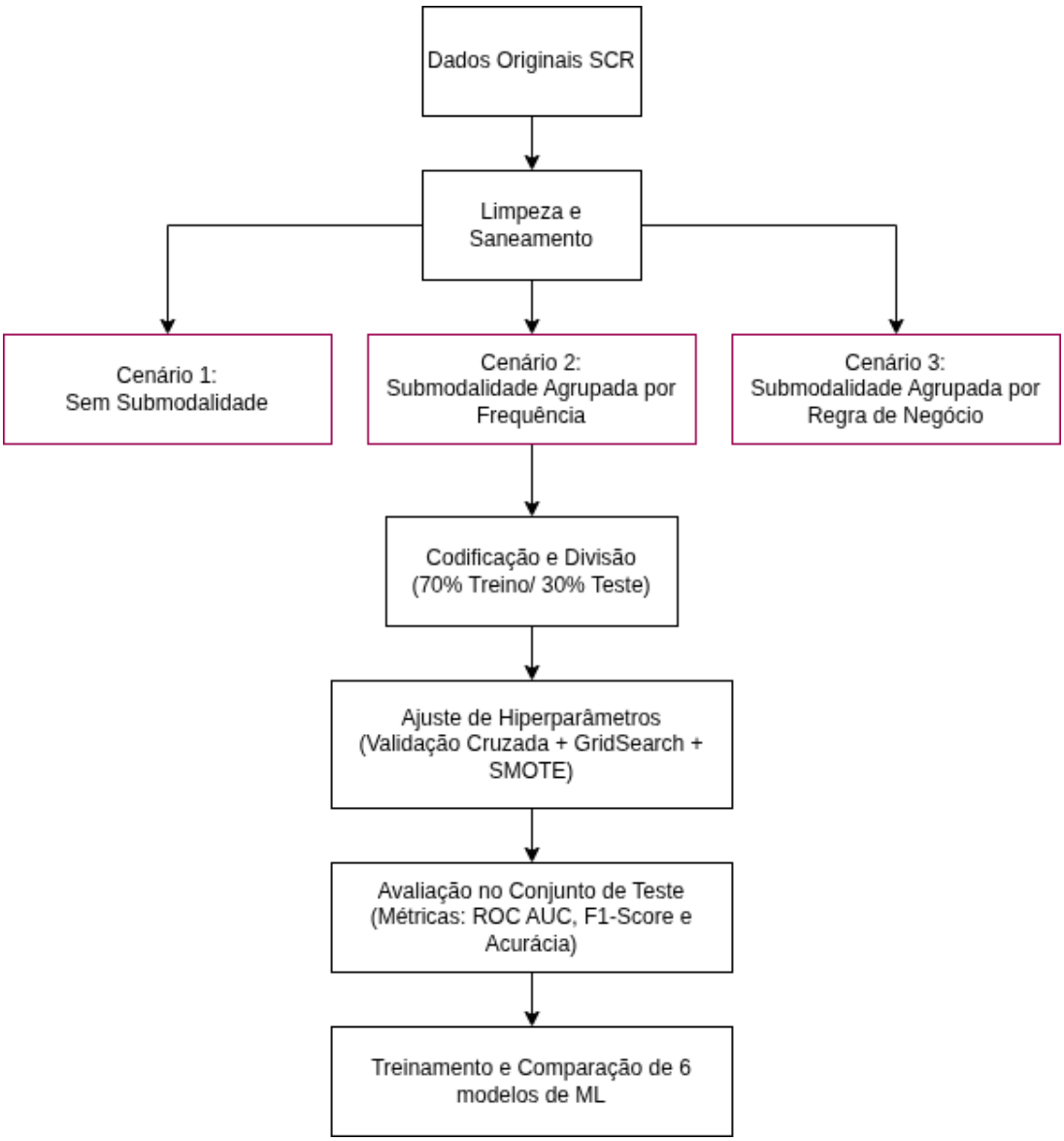



**Figura 6 – Fluxo metodológico do pipeline experimental. Fonte: Autoria própria.**

A variável *submodalidade* representa o tipo específico de operação de crédito contratada pelo cliente e constitui um dos atributos potencialmente mais relevantes para a previsão de

inadimplência. Diferentemente da variável *modalidade*, que corresponde a uma classificação mais ampla das operações de crédito definida pelo SCR, a *submodalidade* fornece um nível adicional de detalhamento sobre o produto financeiro utilizado pelo cliente.

Durante análises preliminares, observou-se que a elevada quantidade de categorias distintas presentes nesse atributo aumentava significativamente a dimensionalidade da base após a aplicação do processo de codificação categórica *One-Hot Encoding*[1]. Esse comportamento resultava em maior complexidade computacional durante o treinamento dos modelos e potencial redução da capacidade de generalização, especialmente em algoritmos mais sensíveis à alta dimensionalidade.

Uma análise exploratória da variável revelou uma distribuição do tipo *long tail*[2] (cauda longa), caracterizada pela concentração da maior parte dos registros em poucas categorias e pela existência de diversas categorias com baixa frequência de ocorrência. Em alguns casos, determinadas submodalidades possuíam apenas algumas dezenas de registros, tornando sua contribuição estatística limitada para o processo de aprendizagem.

Considerando esse cenário, foram definidos três conjuntos de dados distintos para avaliar o impacto da representação da variável *submodalidade* no desempenho dos modelos de classificação.

É importante destacar que os três cenários mantiveram exatamente a mesma quantidade de registros. Como a diferença entre eles está exclusivamente na forma de tratamento da variável *submodalidade*, todos os conjuntos de dados permaneceram com 182.631 registros. As diferenças ocorreram apenas na quantidade de atributos gerados após a aplicação do processo de codificação categórica.

- **Cenário 1 – Sem Submodalidade (Baseline)**

  Neste cenário, a variável *submodalidade* foi removida da base de dados, mantendo-se apenas a variável *modalidade* e os demais atributos originalmente disponíveis. O objetivo foi estabelecer uma linha de base para comparação, permitindo avaliar o desempenho dos modelos sem utilizar informações detalhadas sobre o tipo específico de operação de crédito.

  Esse cenário também possibilita verificar se a informação contida na variável *modalidade* já é suficiente para representar adequadamente o perfil das operações de crédito dos clientes.

[1] Método de transformação de atributos categóricos em representações numéricas binárias. Para cada categoria existente é criada uma nova coluna indicadora, permitindo que algoritmos de aprendizado de máquina processem variáveis textuais. Em atributos de alta cardinalidade, como a variável *submodalidade*, essa técnica pode aumentar significativamente o número de atributos do conjunto de dados.

[2] Distribuição na qual poucas categorias concentram grande parte dos registros, enquanto muitas outras categorias possuem poucas ocorrências individuais.

Após a aplicação do *One-Hot Encoding*, esse conjunto de dados resultou em 63 colunas, sendo 62 atributos preditivos e uma variável-alvo.

- **Cenário 2 – Submodalidade Agrupada por Frequência**

  Após a análise exploratória da distribuição das categorias, observou-se que grande parte das submodalidades possuía baixa representatividade estatística. Para reduzir o impacto dessas categorias raras, foram avaliados diferentes limiares de frequência, incluindo valores de 50, 100, 200, 500 e 1.000 ocorrências.

  Para cada limiar analisado foram observados o número de categorias preservadas e a proporção de registros mantidos sem agrupamento. O valor de 1.000 ocorrências apresentou o melhor equilíbrio entre redução da cardinalidade e preservação da informação disponível, mantendo aproximadamente 97% dos registros originais em categorias representativas.

  Dessa forma, todas as submodalidades com frequência inferior a 1.000 registros foram agrupadas em uma nova categoria denominada *"Outros"*, enquanto as demais permaneceram individualmente identificadas.

  Essa estratégia, conhecida na literatura como *rare category grouping*, busca reduzir ruídos provenientes de categorias pouco frequentes, diminuir a dimensionalidade dos dados e minimizar riscos de sobreajuste (*overfitting*) durante o treinamento dos modelos.

  Após a codificação categórica, esse cenário produziu um conjunto de dados com 93 colunas, sendo 92 atributos preditivos e uma variável-alvo.

- **Cenário 3 – Submodalidade Agrupada por Regras de Negócio (*Engineered*)**

  Embora o agrupamento por frequência reduza a cardinalidade da variável, ele preserva apenas critérios estatísticos e não considera relações semânticas entre os diferentes produtos financeiros.

  Dessa forma, foi desenvolvida uma terceira estratégia baseada em conhecimento de domínio sobre operações de crédito. Nesse cenário, as submodalidades originais foram agrupadas em categorias mais amplas definidas a partir da natureza financeira dos produtos oferecidos pelas instituições.

  A construção desses agrupamentos foi realizada utilizando regras baseadas em expressões regulares e análise das descrições das submodalidades presentes na base. O objetivo foi criar categorias intermediárias entre a modalidade original do SCR e as submodalidades específicas, preservando parte da informação de negócio sem manter a elevada granularidade observada na variável original.

  Como resultado, as submodalidades foram consolidadas em dez grupos representativos: Cartão de Crédito, Crédito Pessoal, Financiamento Habitacional, Financiamento

de Veículos, Crédito Rural, Capital de Giro, Microcrédito, Cheque Especial, Aquisição de Bens e Outros.

Essa estratégia procura combinar interpretabilidade e redução de complexidade, permitindo que operações semelhantes compartilhem uma mesma representação durante o processo de modelagem.

Após a aplicação do *One-Hot Encoding*, esse conjunto de dados resultou em 72 colunas, sendo 71 atributos preditivos e uma variável-alvo.

A Figura 7 apresenta as vinte submodalidades mais frequentes da base analisada, evidenciando a concentração dos registros em um conjunto reduzido de categorias e a existência de uma longa cauda composta por categorias menos frequentes.

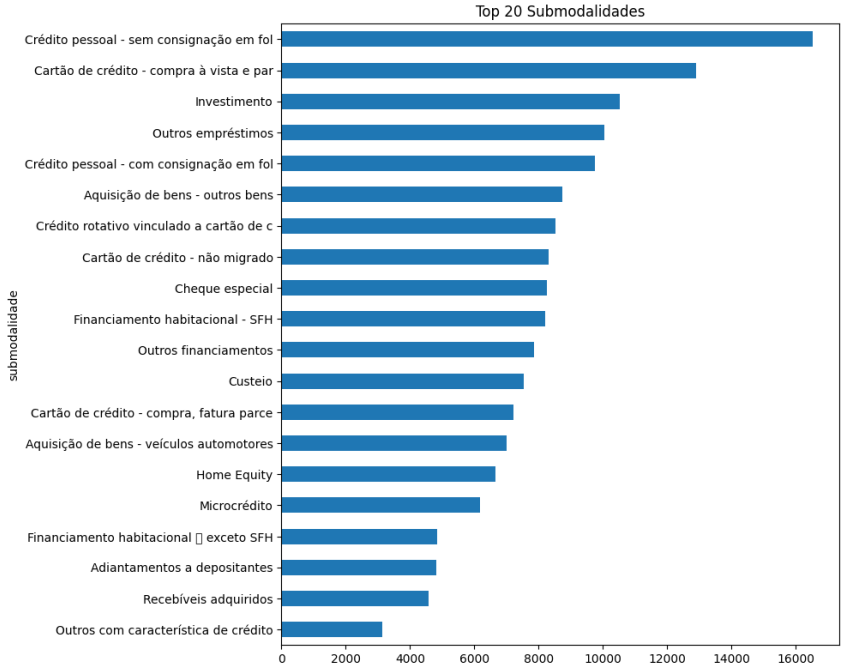



**Figura 7 – Vinte submodalidades mais frequentes da base de dados.**

A Tabela 4 apresenta os agrupamentos definidos para o cenário baseado em regras de negócio.

A Tabela 5 resume as características dos três conjuntos de dados utilizados nos experimentos.

Observa-se que a estratégia baseada em agrupamento por frequência produziu a maior dimensionalidade dentre os cenários avaliados, enquanto a remoção completa da variável submodalidade resultou na representação mais simples. Já o agrupamento baseado em regras de negócio posicionou-se como uma alternativa intermediária, preservando parte da informação semântica das operações de crédito ao mesmo tempo em que reduziu a quantidade de atributos gerados.

Essa comparação permitiu avaliar experimentalmente o impacto da granularidade da variável *submodalidade* sobre o desempenho dos modelos de previsão de inadimplência.

| Grupo | Critério de agrupamento |
|---|---|
| Cartão de Crédito | Submodalidades relacionadas a cartões de crédito e faturas. |
| Crédito Pessoal | Operações de empréstimo pessoal consignado e não consignado. |
| Financiamento Habitacional | Operações destinadas à aquisição, construção ou garantia de imóveis. |
| Financiamento de Veículos | Operações destinadas à aquisição ou arrendamento de veículos. |
| Crédito Rural | Operações de custeio, investimento e comercialização do setor agropecuário. |
| Capital de Giro | Operações voltadas ao financiamento das atividades correntes de empresas. |
| Microcrédito | Operações classificadas como microcrédito. |
| Cheque Especial | Operações de adiantamento a depositantes e limite especial em conta corrente. |
| Aquisição de Bens | Operações destinadas à aquisição de bens de consumo ou equipamentos. |
| Outros | Submodalidades não enquadradas nos grupos anteriores. |

**Tabela 4 – Agrupamento das submodalidades por regras de negócio utilizado no cenário *submodalidade_engineered*.**

**Tabela 5 – Comparação entre os cenários experimentais.**

| Cenário | Registros | Preditores | Colunas Totais |
|---|---|---|---|
| Sem Submodalidade | 182.631 | 62 | 63 |
| Submodalidade Agrupada | 182.631 | 92 | 93 |
| Submodalidade Engineered | 182.631 | 71 | 72 |

## 5.2 Configuração Experimental

Após a definição dos três cenários experimentais, os conjuntos de dados foram preparados para o treinamento e avaliação dos modelos de aprendizado de máquina. Inicialmente, as variáveis categóricas presentes em cada cenário foram transformadas para representação numérica por meio da técnica de *One-Hot Encoding*, na qual cada categoria é convertida em uma variável binária própria. Essa etapa permitiu que os algoritmos de classificação processassem adequadamente os atributos categóricos da base de dados.

Em seguida, os dados foram divididos em conjuntos de treinamento e teste utilizando uma estratégia estratificada, de forma a preservar a proporção original das classes de adimplentes e inadimplentes em ambas as partições. Foram destinados 70% dos registros para treinamento dos modelos e 30% para avaliação final de desempenho. Essa divisão foi realizada uma única vez e mantida para todos os experimentos, garantindo comparabilidade entre os resultados obtidos nos diferentes cenários e algoritmos avaliados.

Considerando o desbalanceamento natural existente na base de dados, também foi avaliada a utilização da técnica *Synthetic Minority Over-sampling Technique* (SMOTE). O método gera exemplos sintéticos da classe minoritária a partir da interpolação entre observações vizinhas, buscando equilibrar a distribuição das classes durante o treinamento. Para evitar vazamento de dados (*data leakage*), a aplicação do SMOTE foi realizada exclusivamente sobre os

dados de treinamento e apenas no interior das dobras da validação cruzada, não sendo aplicada aos conjuntos de validação nem ao conjunto de teste.

O processo de seleção dos melhores hiperparâmetros foi conduzido por meio da técnica de *Grid Search*. Nessa abordagem, diferentes combinações de hiperparâmetros previamente definidas são avaliadas sistematicamente durante o treinamento dos modelos. Para cada combinação testada, foi utilizada validação cruzada estratificada com cinco dobras (*Stratified K-Fold*), permitindo estimar de forma mais robusta a capacidade de generalização dos classificadores.

A métrica adotada para seleção dos melhores hiperparâmetros foi a área sob a curva ROC (ROC AUC), por apresentar menor sensibilidade ao desbalanceamento das classes e avaliar a capacidade discriminatória dos modelos independentemente de um limiar específico de classificação. Após a identificação da melhor configuração para cada algoritmo e cenário experimental, o modelo correspondente foi retreinado utilizando todo o conjunto de treinamento e posteriormente avaliado sobre o conjunto de teste independente.

Dessa forma, cada combinação entre cenário experimental, algoritmo de classificação e estratégia de balanceamento foi submetida ao mesmo protocolo de treinamento, otimização e avaliação, permitindo comparações consistentes entre os resultados obtidos.

## 5.3 Algoritmos Avaliados

Foram avaliados seis algoritmos de aprendizado de máquina amplamente utilizados em problemas de classificação supervisionada: Regressão Logística, Árvore de Decisão, K-Vizinhos Mais Próximos (KNN), Floresta Aleatória (*Random Forest*), Máquinas de Vetores de Suporte (SVM) e XGBoost (*Extreme Gradient Boosting*). A seleção desses algoritmos buscou contemplar diferentes paradigmas de aprendizado, incluindo modelos lineares, baseados em instâncias, baseados em árvores de decisão e métodos de ensemble.

A otimização dos modelos foi realizada por meio da técnica de *Grid Search*, que consiste na avaliação sistemática de combinações previamente definidas de hiperparâmetros. Para todos os algoritmos foi utilizada validação cruzada estratificada com cinco dobras, tendo como critério de seleção a métrica ROC AUC média obtida durante o processo de validação.

Os intervalos de hiperparâmetros avaliados foram definidos com base nas recomendações da documentação oficial das bibliotecas utilizadas, especialmente o Scikit-learn (Scikit-learn Developers, 2025), bem como em valores amplamente empregados em aplicações de classificação supervisionada. A seleção desses intervalos buscou contemplar diferentes níveis de complexidade dos modelos e permitir uma exploração representativa do espaço de busca, mantendo ao mesmo tempo a viabilidade computacional dos experimentos.

Além dos hiperparâmetros específicos de cada algoritmo, também foi incluído no processo de busca um parâmetro responsável por controlar a utilização da técnica SMOTE (*Synthetic Minority Over-sampling Technique*). Dessa forma, cada combinação de hiperparâmetros foi avaliada tanto com balanceamento quanto sem balanceamento da classe minoritária. Para evi-

tar vazamento de dados (*data leakage*), a aplicação do SMOTE ocorreu exclusivamente dentro das dobras de treinamento da validação cruzada, por meio de um pipeline implementado com a biblioteca *Imbalanced-Learn*.

A Regressão Logística foi utilizada como modelo linear de referência para a tarefa de classificação. Nesse algoritmo, foi avaliado o parâmetro de regularização (C), responsável por controlar o equilíbrio entre ajuste aos dados e capacidade de generalização do modelo. Foram testados valores entre 0,01 e 10, permitindo analisar diferentes níveis de penalização dos coeficientes.

A Árvore de Decisão foi incluída por sua elevada interpretabilidade e capacidade de modelar relações não lineares entre as variáveis. Os experimentos investigaram diferentes profundidades máximas da árvore, bem como critérios de pré-poda relacionados à quantidade mínima de amostras necessárias para divisão de nós e formação de folhas. Também foram comparados os critérios de impureza Gini e Entropia para seleção das divisões.

O algoritmo K-Vizinhos Mais Próximos (KNN) foi empregado para representar métodos baseados em similaridade entre observações. Foram avaliados diferentes números de vizinhos, formas de ponderação da influência dos vizinhos e métricas de distância, permitindo verificar qual configuração melhor capturava a estrutura dos dados.

A Floresta Aleatória (*Random Forest*) foi utilizada como representante dos métodos de ensemble baseados em bagging. Os experimentos variaram o número de árvores componentes da floresta e os parâmetros de controle de crescimento das árvores individuais, buscando equilibrar capacidade preditiva e generalização.

Para as Máquinas de Vetores de Suporte (SVM), optou-se pela implementação *LinearSVC*, recomendada para bases de dados com grande quantidade de observações. A otimização concentrou-se no parâmetro (C), responsável por controlar o compromisso entre maximização da margem de separação e penalização dos erros de classificação.

Por fim, foi avaliado o algoritmo XGBoost (*Extreme Gradient Boosting*), amplamente reconhecido por seu elevado desempenho em problemas de classificação tabular. Foram investigadas diferentes combinações envolvendo quantidade de árvores, taxa de aprendizado, profundidade máxima das árvores e proporções de amostragem de linhas e colunas, parâmetros fundamentais para o controle do processo de aprendizado e mitigação de sobreajuste.

A Tabela 6 apresenta os hiperparâmetros e respectivos valores avaliados durante o processo de otimização.

**Tabela 6 – Hiperparâmetros avaliados durante o processo de Grid Search.**

| **Algoritmo** | **Hiperparâmetros avaliados** |
|---|---|
| Regressão Logística | $C \in \{0.01, 0.1, 1, 10\}$, solver = *lbfgs*. |
| Árvore de Decisão | max_depth $\in \{10, 20, 30\}$;<br>min_samples_split $\in \{2, 5, 10\}$;<br>min_samples_leaf $\in \{1, 2, 4\}$;<br>criterion $\in \{$*gini*, *entropy*$\}$. |
| KNN | n_neighbors $\in \{3, 5, 7, 9\}$;<br>weights $\in \{$*uniform*, *distance*$\}$;<br>metric $\in \{$*euclidean*, *manhattan*$\}$. |
| Random Forest | n_estimators $\in \{100, 200\}$;<br>min_samples_split $\in \{2, 5\}$;<br>min_samples_leaf $\in \{1, 2\}$. |
| SVM (LinearSVC) | $C \in \{0.1, 1, 10\}$. |
| XGBoost | n_estimators $\in \{200, 400\}$;<br>learning_rate $\in \{0.05, 0.1\}$;<br>max_depth $\in \{3, 5\}$;<br>subsample $\in \{0.8, 1.0\}$;<br>colsample_bytree $\in \{0.8, 1.0\}$. |

# 6 RESULTADOS

Conforme descrito no Capítulo 4, foram conduzidos experimentos com o objetivo de avaliar o impacto de diferentes estratégias de representação da variável *submodalidade* na tarefa de predição de inadimplência. Para isso, foram considerados três cenários experimentais independentes:

- **Sem submodalidade**: cenário de referência (*baseline*), no qual a variável foi removida da base de dados;
- **Submodalidade agrupada**: cenário em que categorias pouco frequentes foram agrupadas na categoria *Outros*, reduzindo a cardinalidade do atributo;
- **Submodalidade engineered**: cenário baseado em regras de negócio, no qual as submodalidades foram consolidadas em grupos representativos de produtos financeiros.

Em cada cenário foram avaliados seis algoritmos de aprendizado de máquina: Decision Tree, K-Nearest Neighbors (KNN), Logistic Regression, Random Forest, Support Vector Machine (SVM) e XGBoost.

Além da comparação entre os cenários de representação da variável submodalidade, os experimentos também avaliaram o impacto da técnica de balanceamento SMOTE. Dessa forma, cada algoritmo foi treinado e avaliado em duas configurações distintas: com e sem aplicação do método de sobreamostragem.

A seleção dos hiperparâmetros foi realizada por meio de validação cruzada estratificada em cinco partições (*Stratified K-Fold*), utilizando busca em grade (*Grid Search*) sobre o conjunto de treinamento. Posteriormente, os melhores modelos encontrados foram avaliados em um conjunto de teste independente, permitindo medir sua capacidade de generalização em dados não observados durante o treinamento.

O desempenho dos classificadores foi analisado utilizando as métricas ROC-AUC, F1-Score e Acurácia. Considerando o caráter desbalanceado do problema de classificação, o F1-Score foi adotado como principal métrica de comparação entre os experimentos, uma vez que considera simultaneamente os efeitos da precisão e da revocação.

Ao todo, foram executados 36 experimentos, resultantes da combinação entre os seis algoritmos avaliados, os três cenários de dados e as duas configurações relacionadas ao uso do SMOTE. As seções seguintes apresentam os resultados obtidos em cada cenário e uma comparação consolidada entre os modelos.

## 6.1 Resultados do Cenário Sem Submodalidade

Esta seção apresenta os resultados obtidos no cenário de referência (*baseline*), no qual a variável *submodalidade* foi removida da base de dados. O objetivo deste experimento foi ava-

liar a capacidade dos modelos em prever a inadimplência utilizando apenas os demais atributos disponíveis, permitindo posteriormente comparar os efeitos das diferentes estratégias de representação da variável de submodalidade.

Para cada algoritmo foram avaliadas duas configurações distintas: com e sem aplicação do SMOTE. Inicialmente são apresentados os resultados obtidos durante a etapa de validação cruzada utilizada para seleção de hiperparâmetros. Em seguida, são discutidos os resultados finais observados no conjunto de teste.

### 6.1.1 Validação Cruzada

A Tabela 7 apresenta os melhores resultados de ROC-AUC obtidos durante o processo de validação cruzada para cada algoritmo avaliado.

| **Modelo** | **SMOTE** | **ROC-AUC (CV)** |
|---|---|---|
| Decision Tree | Sim | 0,9048 |
| Decision Tree | Não | 0,9082 |
| KNN | Sim | 0,7095 |
| KNN | Não | 0,7141 |
| Logistic Regression | Sim | 0,7764 |
| Logistic Regression | Não | 0,7752 |
| Random Forest | Sim | 0,9273 |
| Random Forest | Não | **0,9292** |
| SVM | Sim | 0,7674 |
| SVM | Não | 0,7664 |
| XGBoost | Sim | 0,9192 |
| XGBoost | Não | 0,9261 |

**Tabela 7 – Resultados da validação cruzada no cenário sem submodalidade**

Observa-se que os melhores resultados de validação cruzada foram obtidos pelos modelos baseados em árvores, especialmente Random Forest e XGBoost, que alcançaram valores de ROC-AUC superiores a 0,91. Em contrapartida, os algoritmos KNN, Logistic Regression e SVM apresentaram desempenho inferior, indicando menor capacidade de separação entre as classes utilizando apenas os atributos disponíveis neste cenário.

Além disso, nota-se que a aplicação do SMOTE não produziu ganhos relevantes durante a etapa de validação cruzada. Em alguns casos, especialmente nos modelos baseados em árvores, os melhores resultados foram observados sem a utilização da técnica de sobreamostragem.

### 6.1.2 Conjunto de Teste

Após a seleção dos melhores hiperparâmetros, os modelos foram avaliados em um conjunto de teste independente. A Tabela 8 apresenta os resultados obtidos para as métricas ROC-AUC, F1-Score e Acurácia.

**Tabela 8 – Resultados no conjunto de teste para o cenário sem submodalidade**

| Modelo | SMOTE | ROC-AUC | F1-Score | Acurácia |
|---|---|---|---|---|
| Decision Tree | Sim | 0,9035 | 0,8377 | 0,8180 |
| Decision Tree | Não | 0,9072 | 0,8522 | 0,8277 |
| KNN | Sim | 0,7205 | 0,7130 | 0,6700 |
| KNN | Não | 0,7268 | 0,7259 | 0,6768 |
| Logistic Regression | Sim | 0,7737 | 0,7486 | 0,7107 |
| Logistic Regression | Não | 0,7727 | 0,7498 | 0,7110 |
| Random Forest | Sim | 0,9274 | 0,8684 | 0,8495 |
| Random Forest | Não | **0,9288** | **0,8726** | **0,8519** |
| SVM | Sim | 0,7650 | 0,7485 | 0,7089 |
| SVM | Não | 0,7641 | 0,7486 | 0,7086 |
| XGBoost | Sim | 0,9194 | 0,8503 | 0,8337 |
| XGBoost | Não | 0,9269 | 0,8628 | 0,8458 |

Os resultados obtidos no conjunto de teste confirmam o comportamento observado durante a validação cruzada. O modelo Random Forest apresentou o melhor desempenho geral, alcançando ROC-AUC de 0,9288, F1-Score de 0,8726 e acurácia de 0,8519 na configuração sem aplicação do SMOTE.

O segundo melhor resultado foi obtido pelo XGBoost, que atingiu ROC-AUC de 0,9269, F1-Score de 0,8628 e acurácia de 0,8458. A proximidade entre os resultados desses dois algoritmos sugere que métodos baseados em *ensembles*[1] de árvores possuem elevada capacidade para modelar relações complexas presentes nos dados de crédito.

Os modelos baseados em uma única árvore de decisão apresentaram desempenho intermediário, enquanto Logistic Regression, SVM e KNN obtiveram resultados significativamente inferiores em todas as métricas avaliadas.

### 6.1.3 Discussão dos Resultados

Os resultados obtidos demonstram que os algoritmos baseados em árvores de decisão apresentaram desempenho superior no cenário sem utilização da variável submodalidade. Em especial, Random Forest e XGBoost alcançaram os maiores valores de ROC-AUC e F1-Score, indicando elevada capacidade de discriminação entre clientes adimplentes e inadimplentes.

A comparação entre as configurações com e sem aplicação do SMOTE evidencia que a técnica de balanceamento não trouxe ganhos consistentes para os modelos avaliados. Em praticamente todos os algoritmos, os melhores resultados foram obtidos sem a utilização da sobreamostragem. Esse comportamento sugere que o conjunto de dados já apresentava informações suficientes para caracterizar a classe minoritária, tornando desnecessária a geração de exemplos sintéticos.

[1] Técnica de aprendizado de máquina que combina múltiplos modelos preditivos com o objetivo de obter melhor desempenho e capacidade de generalização do que modelos individuais.

Também é possível observar que modelos lineares, como Logistic Regression, e algoritmos baseados em distância, como KNN, apresentaram desempenho inferior quando comparados aos métodos baseados em árvores. Isso indica que as relações existentes entre os atributos e a variável alvo provavelmente envolvem padrões não lineares, melhor capturados por algoritmos mais flexíveis.

Os resultados apresentados nesta seção estabelecem uma linha de base para os experimentos subsequentes. Nas próximas seções será analisado se a inclusão da variável submodalidade, por meio das estratégias de agrupamento e engenharia de atributos propostas, é capaz de produzir melhorias adicionais no desempenho dos modelos.

## 6.2 Resultados do Cenário Submodalidade Agrupada

Esta seção apresenta os resultados obtidos no cenário em que a variável *submodalidade* foi mantida, porém submetida a um tratamento de redução de cardinalidade. Categorias com baixa frequência (menos de 1.000 ocorrências) foram agrupadas sob o rótulo "Outros", visando preservar a informação da variável ao mesmo tempo em que se reduz o ruído estatístico gerado por submodalidades extremamente raras.

Assim como no cenário de referência, foram avaliadas configurações com e sem a aplicação da técnica SMOTE para lidar com o desbalanceamento das classes.

### 6.2.1 Validação Cruzada

A Tabela 9 detalha os melhores desempenhos de ROC-AUC alcançados durante o processo de validação cruzada para a otimização dos hiperparâmetros.

| **Modelo** | **SMOTE** | **ROC-AUC (CV)** |
|---|---|---|
| Decision Tree | Sim | 0,9174 |
| Decision Tree | Não | 0,9217 |
| KNN | Sim | 0,7538 |
| KNN | Não | 0,7670 |
| Logistic Regression | Sim | 0,8330 |
| Logistic Regression | Não | 0,8332 |
| Random Forest | Sim | 0,9430 |
| Random Forest | Não | **0,9443** |
| SVM | Sim | 0,8247 |
| SVM | Não | 0,8247 |
| XGBoost | Sim | 0,9394 |
| XGBoost | Não | 0,9428 |

**Tabela 9 – Resultados da validação cruzada no cenário submodalidade agrupada**

A análise inicial revela um salto perceptível no desempenho geral de todos os modelos quando comparado ao cenário sem a variável (*baseline*). Os algoritmos baseados em árvores continuam liderando as métricas, com o Random Forest alcançando um ROC-AUC expressivo

de 0,9443. É notável, também, o ganho de capacidade preditiva dos modelos não baseados em árvores; a Regressão Logística, por exemplo, saltou de valores em torno de 0,77 para 0,83. Novamente, a aplicação da sobreamostragem com SMOTE não demonstrou vantagem clara durante o treinamento.

### 6.2.2 Conjunto de Teste

A Tabela 10 reporta o desempenho final dos modelos ajustados quando avaliados sobre o conjunto de teste independente.

**Tabela 10 – Resultados no conjunto de teste para o cenário submodalidade agrupada**

| Modelo | SMOTE | ROC-AUC | F1-Score | Acurácia |
|---|---|---|---|---|
| Decision Tree | Sim | 0,9149 | 0,8534 | 0,8317 |
| Decision Tree | Não | 0,9213 | 0,8690 | 0,8439 |
| KNN | Sim | 0,7523 | 0,7149 | 0,6809 |
| KNN | Não | 0,7657 | 0,7645 | 0,7110 |
| Logistic Regression | Sim | 0,8315 | 0,7880 | 0,7530 |
| Logistic Regression | Não | 0,8323 | 0,7884 | 0,7536 |
| Random Forest | Sim | 0,9432 | 0,8846 | 0,8672 |
| Random Forest | Não | **0,9444** | **0,8879** | **0,8694** |
| SVM | Sim | 0,8239 | 0,7844 | 0,7469 |
| SVM | Não | 0,8241 | 0,7852 | 0,7475 |
| XGBoost | Sim | 0,9394 | 0,8742 | 0,8590 |
| XGBoost | Não | 0,9431 | 0,8800 | 0,8642 |

A superioridade do Random Forest se manteve confirmada no conjunto de teste (ROC-AUC de 0,9444, F1-Score de 0,8879), consolidando-se como a abordagem mais robusta sem a utilização do balanceamento de classes. O XGBoost acompanhou de perto o topo do *ranking*, registrando métricas bastante similares. A inclusão da *submodalidade* agrupada refletiu positivamente nas métricas de classificação global.

### 6.2.3 Discussão dos Resultados

Os resultados apurados neste cenário atestam que a variável *submodalidade* possui uma expressiva importância no mapeamento da predisposição de um cliente ao *default*. Diferentemente do cenário anterior em que a variável fora ocultada, o modelo passou a dispor de um forte preditor sobre o tipo de crédito concedido.

A estratégia aplicada para tratar a alta cardinalidade — consolidando fatias esparsas de submodalidades raras na categoria "Outros"— provou-se altamente efetiva. Essa técnica impediu a explosão de dimensionalidade durante a transformação *One-Hot Encoding* (que prejudicaria as distâncias no KNN e a velocidade das árvores), mas garantiu que a essência do comportamento de crédito predominante fosse assimilada matematicamente pelos modelos.

A estabilidade em relação ao SMOTE corrobora a tese de que o ganho de informações através de boas técnicas de representação de dados categóricos (*Feature Engineering*) é frequentemente superior a técnicas de amostragem sintética quando há um grande volume absoluto de exemplos disponíveis para ambas as classes reais.

## 6.3 Resultados do Cenário com Submodalidade Engineered

Esta seção apresenta os resultados obtidos no cenário em que a variável *submodalidade* foi transformada por meio de regras de negócio, agrupando categorias em conjuntos representativos de produtos financeiros. O objetivo deste experimento foi avaliar se uma representação mais semântica da informação poderia contribuir para o desempenho dos modelos de predição de inadimplência.

Assim como nos demais cenários, cada algoritmo foi avaliado em duas configurações: com e sem aplicação do método SMOTE. Inicialmente são apresentados os resultados da validação cruzada e, posteriormente, os resultados obtidos no conjunto de teste.

### 6.3.1 Validação Cruzada

A Tabela 11 apresenta os melhores resultados de ROC-AUC obtidos durante o processo de validação cruzada.

| Modelo | SMOTE | ROC-AUC (CV) |
|---|---|---|
| Decision Tree | Sim | 0,9158 |
| Decision Tree | Não | 0,9204 |
| KNN | Sim | 0,7226 |
| KNN | Não | 0,7336 |
| Logistic Regression | Sim | 0,7960 |
| Logistic Regression | Não | 0,7962 |
| Random Forest | Sim | 0,9383 |
| Random Forest | Não | **0,9400** |
| SVM | Sim | 0,7877 |
| SVM | Não | 0,7877 |
| XGBoost | Sim | 0,9333 |
| XGBoost | Não | 0,9376 |

**Tabela 11 – Resultados da validação cruzada no cenário com submodalidade engineered**

Os resultados obtidos durante a validação cruzada indicam que os algoritmos baseados em ensembles de árvores continuam apresentando o melhor desempenho. O Random Forest obteve o maior ROC-AUC, seguido pelo XGBoost. Em comparação ao cenário sem submodalidade, observa-se um aumento nos valores de ROC-AUC para praticamente todos os modelos, sugerindo que a representação engineered da variável adicionou informação relevante ao processo de classificação.

Além disso, assim como observado anteriormente, a aplicação do SMOTE não proporcionou melhorias significativas durante a etapa de validação cruzada.

### 6.3.2 Conjunto de Teste

Após a seleção dos melhores hiperparâmetros, os modelos foram avaliados em um conjunto de teste independente. A Tabela 12 apresenta os resultados obtidos.

**Tabela 12 – Resultados no conjunto de teste para o cenário com submodalidade engineered**

| Modelo | SMOTE | ROC-AUC | F1-Score | Acurácia |
|---|---|---|---|---|
| Decision Tree | Sim | 0,9141 | 0,8426 | 0,8238 |
| Decision Tree | Não | 0,9206 | 0,8656 | 0,8424 |
| KNN | Sim | 0,7220 | 0,7036 | 0,6642 |
| KNN | Não | 0,7322 | 0,7514 | 0,6919 |
| Logistic Regression | Sim | 0,7955 | 0,7501 | 0,7179 |
| Logistic Regression | Não | 0,7956 | 0,7475 | 0,7152 |
| Random Forest | Sim | 0,9385 | 0,8784 | 0,8604 |
| Random Forest | Não | **0,9399** | **0,8824** | **0,8631** |
| SVM | Sim | 0,7874 | 0,7426 | 0,7086 |
| SVM | Não | 0,7875 | 0,7406 | 0,7063 |
| XGBoost | Sim | 0,9332 | 0,8674 | 0,8517 |
| XGBoost | Não | 0,9380 | 0,8745 | 0,8582 |

Os resultados obtidos no conjunto de teste confirmam as observações realizadas durante a validação cruzada. O modelo Random Forest apresentou novamente o melhor desempenho geral, alcançando ROC-AUC de 0,9399, F1-Score de 0,8824 e acurácia de 0,8631 na configuração sem utilização do SMOTE.

O XGBoost apresentou desempenho bastante próximo, alcançando ROC-AUC de 0,9380 e F1-Score de 0,8745. Ambos os algoritmos superaram os resultados observados no cenário sem submodalidade, indicando que a representação engineered da variável forneceu informações adicionais úteis para a identificação de padrões de inadimplência.

Os modelos Logistic Regression e SVM apresentaram melhorias moderadas em relação ao cenário de referência, enquanto o KNN continuou apresentando os menores valores entre os algoritmos avaliados.

### 6.3.3 Discussão dos Resultados

Os resultados obtidos demonstram que a estratégia de engenharia da variável submodalidade contribuiu positivamente para o desempenho dos modelos. Em comparação ao cenário sem submodalidade, observam-se aumentos consistentes nos valores de ROC-AUC e F1-Score para a maioria dos algoritmos avaliados.

O Random Forest permaneceu como o modelo de melhor desempenho, alcançando os maiores valores em todas as métricas analisadas. O XGBoost apresentou resultados muito

próximos, reforçando a adequação de métodos baseados em ensembles de árvores para problemas de análise de risco de crédito.

Outro aspecto relevante é que a representação engineered apresentou desempenho superior ao cenário de referência e permitiu ganhos mesmo para algoritmos tradicionalmente menos robustos, como Logistic Regression e SVM. Isso sugere que o agrupamento baseado em regras de negócio conseguiu capturar características relevantes dos diferentes produtos financeiros, tornando a informação mais útil para os modelos.

Por fim, assim como observado nos demais cenários, a aplicação do SMOTE não resultou em melhorias consistentes. Em praticamente todos os algoritmos, os melhores resultados foram obtidos sem a utilização da técnica de sobreamostragem, indicando que a qualidade da representação dos atributos teve maior impacto no desempenho do que o balanceamento artificial das classes.

## 6.4 Comparação Geral dos Cenários

Após a avaliação individual dos três cenários experimentais, torna-se possível analisar o impacto das diferentes estratégias de representação da variável *submodalidade* sobre o desempenho dos modelos de predição de inadimplência.

A Tabela 13 apresenta os melhores resultados obtidos em cada cenário, considerando o modelo com maior F1-Score no conjunto de teste.

| Cenário | Modelo | ROC-AUC | F1-Score | Acurácia |
|---|---|---|---|---|
| Sem submodalidade | Random Forest | 0,9288 | 0,8726 | 0,8519 |
| Submodalidade agrupada | Random Forest | **0,9444** | **0,8879** | **0,8694** |
| Submodalidade engineered | Random Forest | 0,9399 | 0,8824 | 0,8631 |

**Tabela 13 – Comparação dos melhores modelos em cada cenário**

Os resultados demonstram que a inclusão da variável *submodalidade* contribuiu positivamente para o desempenho dos modelos de classificação. Em todos os cenários, o algoritmo Random Forest apresentou os melhores resultados, evidenciando sua robustez para modelar relações complexas presentes nos dados de crédito.

Comparando os cenários, observa-se que a estratégia de agrupamento das submodalidades produziu o melhor desempenho geral, alcançando ROC-AUC de 0,9444, F1-Score de 0,8879 e acurácia de 0,8694. Em relação ao cenário sem utilização da variável, houve um aumento de aproximadamente 1,53 pontos percentuais no F1-Score, indicando melhora na capacidade de identificação dos clientes inadimplentes.

O cenário baseado em *feature engineering* também apresentou desempenho superior ao cenário de referência, alcançando F1-Score de 0,8824. Entretanto, seus resultados ficaram ligeiramente abaixo daqueles observados na abordagem de agrupamento.

Além da comparação dos melhores modelos, é importante analisar o comportamento geral dos experimentos realizados em cada cenário. A Figura 8 apresenta a distribuição dos

valores de F1-Score obtidos pelos diferentes algoritmos avaliados. Observa-se que o cenário de submodalidade agrupada não apenas apresentou o melhor resultado individual, mas também concentrou os maiores valores medianos de F1-Score e uma distribuição mais elevada em comparação aos demais cenários. Esse comportamento indica que os ganhos observados não ficaram restritos a um único modelo, mas ocorreram de forma consistente entre diferentes algoritmos.

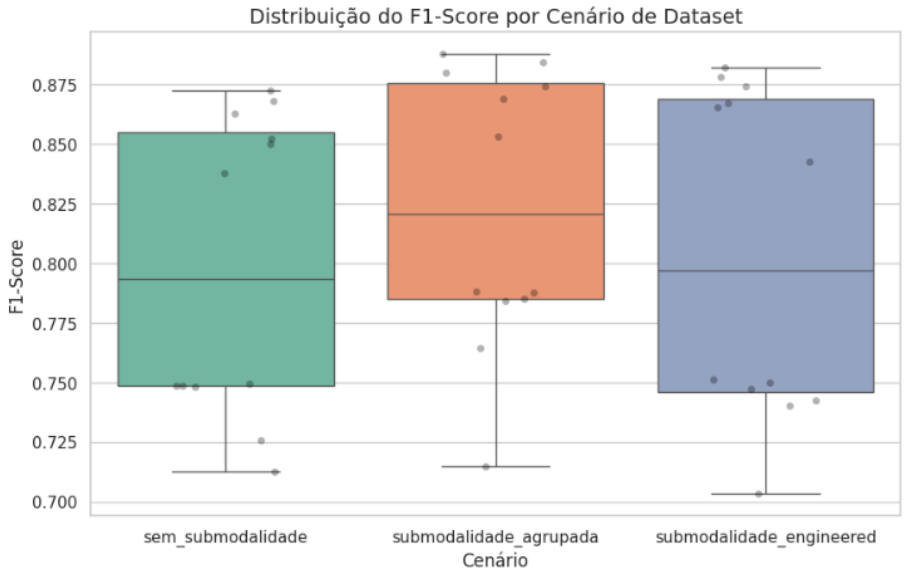



**Figura 8 – Distribuição dos valores de F1-Score obtidos em cada cenário experimental**

Por outro lado, os cenários sem submodalidade e com submodalidade engineered apresentaram distribuições semelhantes, embora a abordagem baseada em regras de negócio tenha demonstrado uma leve vantagem em relação ao cenário de referência. Esses resultados reforçam que a inclusão da variável *submodalidade* contribui para o aumento da capacidade preditiva dos modelos, especialmente quando representada por meio de agrupamentos capazes de preservar informações relevantes e reduzir a fragmentação dos dados.

Além da comparação entre os cenários, é possível observar o comportamento específico de cada algoritmo. A Tabela 14 apresenta os melhores resultados de F1-Score obtidos por cada modelo em cada cenário.

**Tabela 14 – Melhores valores de F1-Score por algoritmo em cada cenário**

| Algoritmo | Sem Submodalidade | Agrupada | Engineered |
|---|---|---|---|
| Decision Tree | 0,8522 | **0,8690** | 0,8656 |
| KNN | 0,7259 | **0,7645** | 0,7514 |
| Logistic Regression | 0,7498 | **0,7884** | 0,7501 |
| Random Forest | 0,8726 | **0,8879** | 0,8824 |
| SVM | 0,7486 | **0,7852** | 0,7426 |
| XGBoost | 0,8628 | **0,8800** | 0,8745 |

Os resultados evidenciam que a estratégia de agrupamento das submodalidades produziu ganhos consistentes para todos os algoritmos avaliados. Esse comportamento sugere que a

redução da cardinalidade do atributo permitiu preservar informações relevantes sobre os produtos financeiros ao mesmo tempo em que reduziu o ruído decorrente da existência de categorias excessivamente específicas ou pouco representativas.

Já a abordagem baseada em regras de negócio também apresentou ganhos em relação ao cenário sem submodalidade, especialmente para os modelos baseados em árvores. Contudo, para algoritmos lineares e baseados em distância, como Logistic Regression, SVM e KNN, os benefícios foram mais modestos.

Outro resultado importante observado ao longo dos experimentos refere-se ao uso do SMOTE. Em praticamente todos os cenários e algoritmos, as melhores métricas foram obtidas sem a aplicação da técnica de sobreamostragem. Isso sugere que o conjunto de dados original já apresentava características suficientes para que os modelos aprendessem padrões discriminativos sem a necessidade de geração de amostras sintéticas.

De forma geral, os resultados indicam que a variável *submodalidade* possui relevância para a tarefa de predição de inadimplência. Entre as estratégias avaliadas, a abordagem de agrupamento apresentou o melhor equilíbrio entre simplicidade de implementação e desempenho preditivo, tornando-se a alternativa mais adequada para utilização em ambientes reais de análise de crédito.

Por fim, observa-se que os melhores resultados globais do estudo foram obtidos pelo modelo Random Forest no cenário de submodalidade agrupada, alcançando ROC-AUC de 0,9444, F1-Score de 0,8879 e acurácia de 0,8694. Esses resultados demonstram que a combinação entre modelos de ensemble baseados em árvores e uma representação adequada da variável de submodalidade pode contribuir significativamente para o aumento da capacidade preditiva em problemas de classificação de inadimplência.

## 6.5 Impacto da Aplicação do SMOTE

Além da avaliação das diferentes estratégias de representação da variável *submodalidade*, este trabalho também investigou o impacto da aplicação da técnica SMOTE (*Synthetic Minority Over-sampling Technique*) sobre o desempenho dos modelos de classificação.

O objetivo do SMOTE é reduzir os efeitos do desbalanceamento entre as classes por meio da geração de exemplos sintéticos da classe minoritária. Em problemas de crédito, essa abordagem é frequentemente utilizada para aumentar a capacidade dos modelos em identificar clientes inadimplentes, reduzindo o viés em direção à classe majoritária.

Para avaliar sua contribuição, todos os algoritmos foram treinados e testados em duas configurações distintas: com e sem aplicação da técnica. A análise foi realizada nos três cenários experimentais considerados neste estudo.

A Tabela 15 apresenta os resultados obtidos pelo algoritmo Random Forest, que foi o modelo de melhor desempenho em todos os cenários avaliados.

**Tabela 15 – Impacto do SMOTE nos resultados do Random Forest**

| Cenário | SMOTE | ROC-AUC | F1-Score | Acurácia |
|---|---|---|---|---|
| Sem submodalidade | Sim | 0,9274 | 0,8684 | 0,8495 |
| Sem submodalidade | Não | **0,9288** | **0,8726** | **0,8519** |
| Submodalidade agrupada | Sim | 0,9432 | 0,8846 | 0,8672 |
| Submodalidade agrupada | Não | **0,9444** | **0,8879** | **0,8694** |
| Submodalidade engineered | Sim | 0,9385 | 0,8784 | 0,8604 |
| Submodalidade engineered | Não | **0,9399** | **0,8824** | **0,8631** |

Observa-se que, em todos os cenários, os melhores resultados foram obtidos sem a utilização do SMOTE. Embora as diferenças observadas sejam relativamente pequenas, elas ocorreram de forma consistente para as três métricas avaliadas.

Esse comportamento também foi identificado nos demais algoritmos analisados. Em geral, a aplicação da técnica produziu ganhos pouco expressivos ou reduções marginais de desempenho quando comparada aos modelos treinados utilizando apenas os dados originais.

Uma possível explicação para esse resultado está relacionada às características do conjunto de dados utilizado. Apesar da existência de desbalanceamento entre as classes, a quantidade de observações disponíveis mostrou-se suficiente para que os algoritmos identificassem padrões relevantes associados à inadimplência sem a necessidade de gerar exemplos sintéticos adicionais.

Outro fator importante refere-se ao desempenho dos modelos baseados em árvores. Algoritmos como Random Forest, XGBoost e Decision Tree possuem mecanismos internos capazes de lidar relativamente bem com distribuições desbalanceadas, especialmente quando há quantidade adequada de exemplos da classe minoritária durante o treinamento.

Além disso, a geração de amostras sintéticas pode introduzir instâncias artificiais que não representam perfeitamente a distribuição real dos dados. Em situações nas quais os atributos já apresentam boa capacidade discriminativa, a inclusão dessas observações pode aumentar o ruído do conjunto de treinamento e prejudicar levemente a capacidade de generalização dos modelos.

Os resultados obtidos neste estudo indicam que a utilização do SMOTE não trouxe benefícios significativos para a tarefa de predição de inadimplência considerada. Em todos os cenários avaliados, os modelos sem aplicação da técnica apresentaram desempenho igual ou superior aos seus equivalentes balanceados.

Dessa forma, conclui-se que o principal fator responsável pelos ganhos observados ao longo dos experimentos foi a estratégia de representação da variável *submodalidade*, e não a aplicação de técnicas de sobreamostragem. Consequentemente, para o conjunto de dados utilizado neste trabalho, o treinamento com os dados originais mostrou-se a alternativa mais adequada, produzindo modelos mais simples e com melhor capacidade de generalização.

## 6.6 Ranking Geral dos Experimentos

Com o objetivo de consolidar os resultados obtidos ao longo dos experimentos, esta seção apresenta um ranking dos modelos avaliados considerando a métrica F1-Score no conjunto de teste. Conforme discutido anteriormente, essa métrica foi adotada como principal critério de comparação por equilibrar precisão e revocação, sendo especialmente adequada para problemas de classificação com classes desbalanceadas.

A Tabela 16 apresenta os dez melhores resultados obtidos entre todos os 36 experimentos realizados.

| Pos. | Modelo | Cenário | SMOTE | F1-Score |
|---|---|---|---|---|
| 1 | Random Forest | Submodalidade agrupada | Não | 0,8879 |
| 2 | Random Forest | Submodalidade engineered | Não | 0,8824 |
| 3 | XGBoost | Submodalidade agrupada | Não | 0,8800 |
| 4 | Random Forest | Submodalidade agrupada | Sim | 0,8846 |
| 5 | XGBoost | Submodalidade engineered | Não | 0,8745 |
| 6 | Random Forest | Sem submodalidade | Não | 0,8726 |
| 7 | Decision Tree | Submodalidade agrupada | Não | 0,8690 |
| 8 | Random Forest | Submodalidade engineered | Sim | 0,8784 |
| 9 | Random Forest | Sem submodalidade | Sim | 0,8684 |
| 10 | XGBoost | Submodalidade agrupada | Sim | 0,8742 |

**Tabela 16 – Ranking dos melhores experimentos com base no F1-Score**

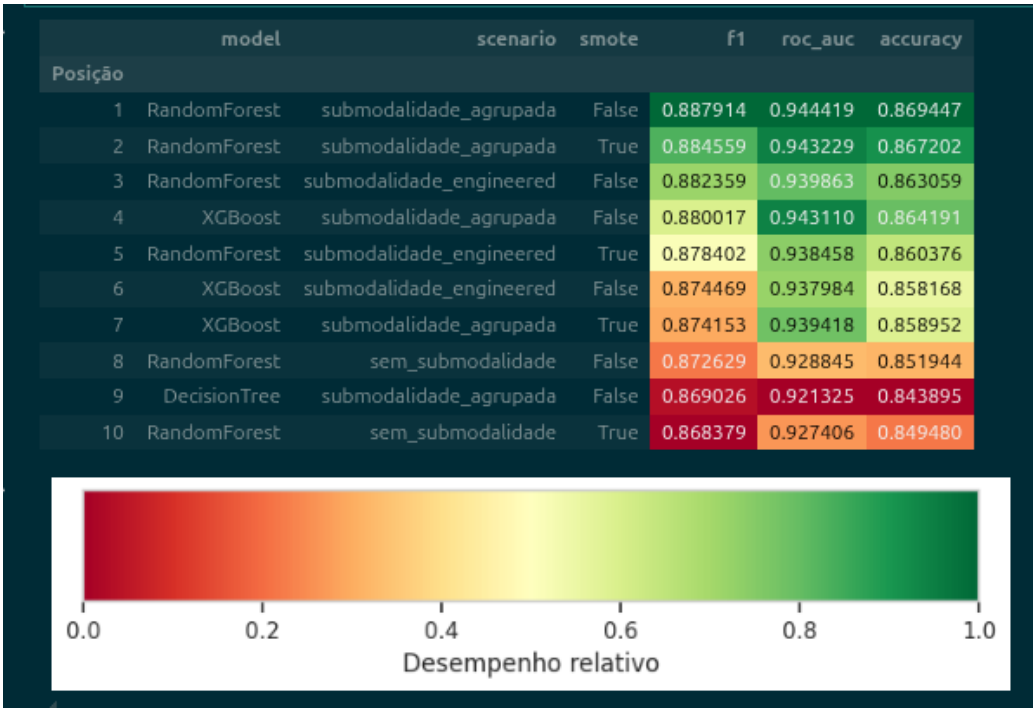
Os resultados evidenciam a predominância dos modelos baseados em árvores de decisão entre os experimentos de melhor desempenho. Das dez primeiras posições do ranking, oito foram ocupadas por Random Forest ou XGBoost, reforçando a capacidade desses algoritmos em capturar padrões complexos presentes nos dados de crédito.

Observa-se também que os cenários que incorporaram informações de submodalidade ocuparam as primeiras posições da classificação. Em especial, a abordagem de submodalidade agrupada apresentou os melhores resultados globais, ocupando a primeira colocação com o modelo Random Forest sem aplicação do SMOTE.

Outro aspecto relevante é que a maioria dos experimentos melhor posicionados foi obtida sem a utilização de técnicas de sobreamostragem. Esse comportamento está alinhado com as análises apresentadas na Seção 6.5, indicando que a inclusão de exemplos sintéticos não trouxe benefícios significativos para o conjunto de dados utilizado.

A Figura 9 apresenta uma visualização dos dez melhores experimentos classificados de acordo com o F1-Score.

De forma geral, os resultados confirmam que a inclusão da variável *submodalidade* contribuiu para o aumento da capacidade preditiva dos modelos e que os algoritmos de ensemble baseados em árvores apresentaram o melhor desempenho para a tarefa de predição de inadimplência. O melhor resultado global foi obtido pelo modelo Random Forest no cenário de submodalidade agrupada sem aplicação do SMOTE, alcançando F1-Score de 0,8879, ROC-AUC de 0,9444 e acurácia de 0,8694.




| Posição | model | scenario | smote | f1 | roc_auc | accuracy |
|---|---|---|---|---|---|---|
| 1 | RandomForest | submodalidade_agrupada | False | 0.887914 | 0.944419 | 0.869447 |
| 2 | RandomForest | submodalidade_agrupada | True | 0.884559 | 0.943229 | 0.867202 |
| 3 | RandomForest | submodalidade_engineered | False | 0.882359 | 0.939863 | 0.863059 |
| 4 | XGBoost | submodalidade_agrupada | False | 0.880017 | 0.943110 | 0.864191 |
| 5 | RandomForest | submodalidade_engineered | True | 0.878402 | 0.938458 | 0.860376 |
| 6 | XGBoost | submodalidade_engineered | False | 0.874469 | 0.937984 | 0.858168 |
| 7 | XGBoost | submodalidade_agrupada | True | 0.874153 | 0.939418 | 0.858952 |
| 8 | RandomForest | sem_submodalidade | False | 0.872629 | 0.928845 | 0.851944 |
| 9 | DecisionTree | submodalidade_agrupada | False | 0.869026 | 0.921325 | 0.843895 |
| 10 | RandomForest | sem_submodalidade | True | 0.868379 | 0.927406 | 0.849480 |

**Figura 9 – Ranking dos dez melhores experimentos considerando o F1-Score**

## 6.7 Melhor Modelo Encontrado

Entre os 36 experimentos realizados neste trabalho, o melhor desempenho foi obtido pelo algoritmo Random Forest no cenário de submodalidade agrupada sem aplicação da técnica SMOTE. Esse modelo alcançou os maiores valores simultaneamente para as métricas ROC-AUC, F1-Score e acurácia, destacando-se como a solução mais eficaz para a tarefa de predição de inadimplência considerada neste estudo.

A Tabela 17 apresenta um resumo dos resultados obtidos pelo modelo selecionado.

| Métrica | Valor |
|---|---|
| ROC-AUC | 0,9444 |
| F1-Score | 0,8879 |
| Acurácia | 0,8694 |

**Tabela 17 – Desempenho do melhor modelo encontrado**

O desempenho obtido demonstra elevada capacidade de discriminação entre clientes adimplentes e inadimplentes, evidenciada pelo valor de ROC-AUC superior a 0,94. Além disso, o F1-Score de 0,8879 indica um bom equilíbrio entre precisão e revocação, característica particularmente importante em aplicações de análise de crédito, nas quais erros de classificação podem gerar impactos financeiros relevantes.

A escolha do cenário de submodalidade agrupada mostrou-se fundamental para a obtenção desse resultado. O agrupamento das categorias menos frequentes permitiu reduzir a cardinalidade da variável sem eliminar informações relevantes sobre os diferentes tipos de ope-

rações de crédito. Como consequência, os modelos puderam explorar de forma mais eficiente os padrões presentes nos dados, aumentando sua capacidade preditiva.

Outro aspecto importante é que o melhor modelo foi obtido sem a utilização do SMOTE. Esse resultado reforça as conclusões apresentadas anteriormente, indicando que a geração de exemplos sintéticos não trouxe benefícios significativos para o conjunto de dados analisado. Pelo contrário, os melhores desempenhos foram observados quando os algoritmos foram treinados utilizando apenas os dados originais.

Em relação à configuração do algoritmo, o modelo Random Forest foi treinado utilizando os hiperparâmetros apresentados na Tabela 18.

| **Hiperparâmetro** | **Valor** |
|---|---|
| Número de árvores (*n_estimators*) | 200 |
| *min_samples_split* | 5 |
| *min_samples_leaf* | 1 |
| SMOTE | Não aplicado |

**Tabela 18 – Hiperparâmetros do melhor modelo**

Os resultados obtidos demonstram que a combinação entre uma representação adequada da variável *submodalidade* e algoritmos de ensemble baseados em árvores constitui uma estratégia eficaz para a predição de inadimplência. Dessa forma, o modelo Random Forest com submodalidade agrupada foi selecionado como o modelo final deste trabalho, servindo como principal evidência experimental para responder ao problema de pesquisa proposto.

## 6.8 Discussão dos Resultados

Os experimentos realizados permitiram avaliar tanto a influência da variável *submodalidade* quanto o impacto da aplicação do SMOTE na tarefa de predição de inadimplência. A análise conjunta dos resultados evidencia que a forma de representação dos dados exerceu influência mais significativa sobre o desempenho dos modelos do que a utilização de técnicas de balanceamento.

No cenário de referência, em que a variável *submodalidade* foi removida da base de dados, os algoritmos baseados em árvores já apresentaram desempenho elevado, especialmente Random Forest e XGBoost. Esse resultado sugere que os demais atributos disponíveis no conjunto de dados possuem capacidade relevante para explicar os padrões associados à inadimplência.

Entretanto, a inclusão da variável *submodalidade* proporcionou ganhos adicionais de desempenho. Tanto a estratégia de agrupamento quanto a abordagem baseada em regras de negócio apresentaram resultados superiores ao cenário sem submodalidade, indicando que informações relacionadas ao tipo de operação de crédito contribuem para a diferenciação entre clientes adimplentes e inadimplentes.

Entre as estratégias avaliadas, o cenário de submodalidade agrupada apresentou os melhores resultados globais. Uma possível explicação para esse comportamento é que o agrupamento reduziu a cardinalidade da variável ao mesmo tempo em que preservou informações relevantes sobre os produtos financeiros. Dessa forma, tornou-se possível diminuir a fragmentação dos dados causada por categorias pouco frequentes, reduzindo o risco de sobreajuste e favorecendo a capacidade de generalização dos modelos.

A abordagem de submodalidade *engineered* também produziu melhorias em relação ao cenário de referência, confirmando que a incorporação de conhecimento de domínio pode ser uma estratégia eficaz para enriquecer a representação dos dados. Contudo, os resultados obtidos ficaram ligeiramente abaixo daqueles observados na estratégia de agrupamento, sugerindo que algumas informações discriminativas podem ter sido perdidas durante o processo de consolidação das categorias.

Outro aspecto relevante observado ao longo dos experimentos foi a superioridade consistente dos algoritmos baseados em árvores de decisão. Random Forest e XGBoost apresentaram os maiores valores de ROC-AUC, F1-Score e acurácia entre os modelos avaliados. Esse comportamento está alinhado com a literatura da área de análise de crédito, que frequentemente destaca a capacidade desses algoritmos em modelar relações não lineares, interações entre atributos e padrões complexos presentes em bases de dados financeiras.

Os resultados obtidos evidenciam que os algoritmos Random Forest e XGBoost apresentaram desempenho superior aos demais modelos avaliados. Essa superioridade pode ser observada tanto nas métricas quantitativas quanto na análise gráfica das curvas ROC, apresentada na Figura 10. Além disso, os melhores resultados foram obtidos nos cenários que incorporaram informações de submodalidade, especialmente na abordagem de submodalidade agrupada.

A Figura 10 apresenta uma comparação das curvas ROC dos principais modelos avaliados utilizando o cenário de submodalidade agrupada, que apresentou o melhor desempenho geral. Observa-se que os modelos Random Forest e XGBoost mantiveram curvas mais próximas do canto superior esquerdo do gráfico, evidenciando maior capacidade de distinguir corretamente clientes adimplentes e inadimplentes em diferentes limiares de decisão. Em contrapartida, modelos como KNN, SVM e Regressão Logística apresentaram curvas mais próximas da linha de classificação aleatória, indicando menor poder discriminatório.

A análise gráfica reforça os resultados quantitativos apresentados anteriormente, demonstrando que os modelos que obtiveram maiores valores de ROC-AUC também apresentaram melhor separação entre as classes ao longo de toda a curva ROC.

Por outro lado, algoritmos lineares e baseados em distância, como Regressão Logística, SVM e KNN, apresentaram desempenho inferior em todos os cenários avaliados. Embora esses modelos sejam amplamente utilizados em tarefas de classificação, os resultados sugerem que a estrutura dos dados analisados exige métodos capazes de capturar relações mais complexas entre as variáveis.

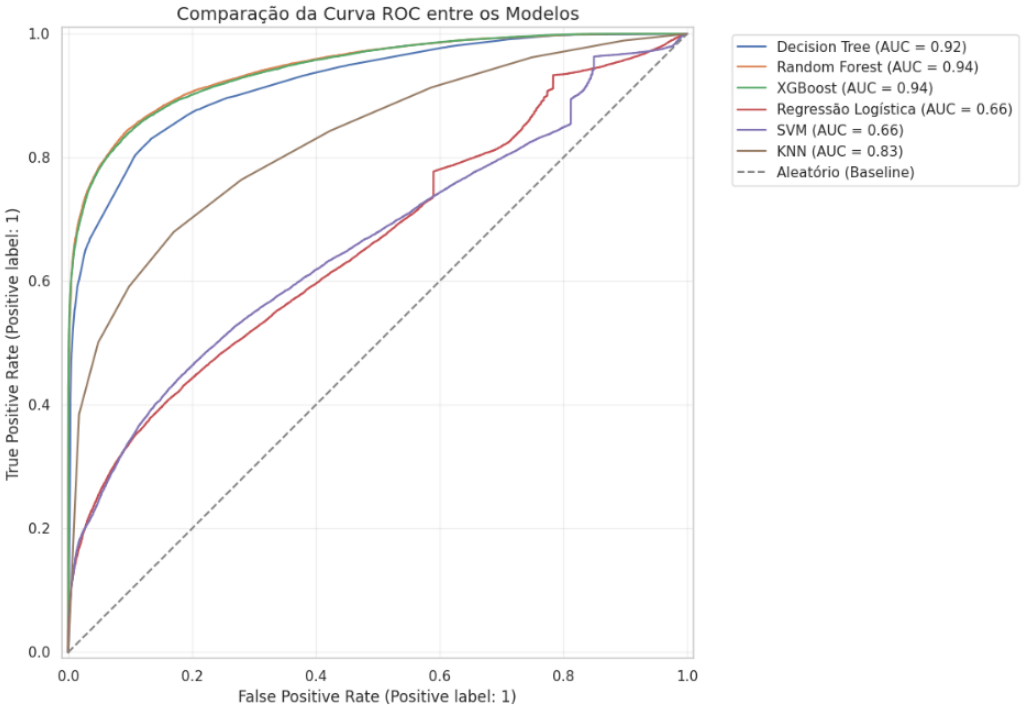



**Figura 10 – Comparação das curvas ROC dos modelos avaliados no cenário de submodalidade agrupada.**

Em relação ao uso do SMOTE, os resultados indicaram que a técnica não produziu ganhos significativos de desempenho. Em praticamente todos os cenários, os melhores resultados foram obtidos sem a aplicação da sobreamostragem. Esse comportamento sugere que o conjunto de dados utilizado já continha informações suficientes para caracterizar adequadamente a classe minoritária, reduzindo a necessidade de geração de exemplos sintéticos.

De forma geral, os resultados obtidos respondem positivamente ao problema de pesquisa proposto neste trabalho. A inclusão da variável *submodalidade* mostrou-se capaz de melhorar a capacidade preditiva dos modelos de aprendizado de máquina, especialmente quando representada por meio da estratégia de agrupamento. Além disso, os experimentos demonstraram que algoritmos de ensemble baseados em árvores constituem a alternativa mais adequada para a tarefa de predição de inadimplência no contexto analisado.

Por fim, destaca-se que o melhor modelo identificado foi o Random Forest treinado no cenário de submodalidade agrupada sem aplicação do SMOTE, alcançando ROC-AUC de 0,9444, F1-Score de 0,8879 e acurácia de 0,8694. Esses resultados demonstram que a combinação entre uma representação apropriada dos dados e algoritmos robustos de aprendizado de máquina pode contribuir significativamente para a construção de sistemas de apoio à decisão no contexto de análise de crédito e gestão de risco.

# 7 CONSIDERAÇÕES FINAIS

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

A relevância do estudo se confirmou diante da crescente adoção de inteligência artificial no setor financeiro, especialmente em atividades relacionadas ao monitoramento contínuo de carteiras e à melhoria da gestão de risco. Embora soluções desse tipo sejam mais consolidadas em instituições de grande porte, observou-se que ainda existe espaço para abordagens mais acessíveis, transparentes e adaptáveis a diferentes realidades organizacionais. Nesse sentido, o trabalho contribuiu ao apresentar uma aplicação prática dessas técnicas no contexto pós-concessão, com foco exploratório e educacional.

Com a realização dos experimentos e testes, foi possível validar a viabilidade da abordagem proposta, evidenciando o potencial dos modelos de aprendizado de máquina na identificação preliminar de sinais de risco. Os resultados obtidos indicam que a metodologia adotada pode funcionar como um apoio complementar às práticas tradicionais de análise, especialmente na priorização de casos e no acompanhamento da carteira ao longo do tempo.

Apesar disso, algumas limitações devem ser consideradas, como as restrições inerentes ao uso de bases de dados públicas e a necessidade de simplificação do escopo para garantir consistência metodológica. Ainda assim, os resultados reforçam que mesmo dados abertos podem fornecer informações relevantes quando tratados com técnicas adequadas de análise.

Além disso, embora a proposta envolva uma abordagem analítica e experimental baseada em modelos de aprendizado de máquina, este trabalho não contempla o desenvolvimento de um sistema completo para uso em ambiente produtivo. O escopo concentrou-se na construção, avaliação e análise de modelos preditivos em contexto acadêmico. Entretanto, com o objetivo de ilustrar uma possível aplicação prática dos resultados obtidos, foi desenvolvida uma interface simples de caráter demonstrativo, disponibilizada juntamente com o código-fonte do projeto. Essa interface permite visualizar informações relacionadas aos modelos desenvolvidos e exemplifica como uma solução desse tipo poderia ser utilizada por usuários em atividades de acompanhamento e análise de risco de crédito. Ressalta-se, contudo, que o protótipo possui finalidade exclusivamente ilustrativa e não foi concebido para operação em ambiente real.

Também é importante destacar que a validação realizada neste estudo ocorreu exclusivamente em ambiente offline, utilizando dados históricos previamente coletados. Dessa forma, aspectos relacionados ao monitoramento contínuo do desempenho do modelo, à identificação de degradação ao longo do tempo e à definição de estratégias de retreinamento periódico não foram contemplados, constituindo oportunidades para pesquisas futuras.

Embora o estudo tenha sido conduzido exclusivamente com dados públicos e em ambiente acadêmico, a metodologia desenvolvida demonstra um caminho viável para a aplicação de técnicas de aprendizado de máquina no acompanhamento pós-concessão de crédito em instituições financeiras. Nesse sentido, os resultados obtidos não devem ser interpretados apenas como um exercício experimental, mas também como uma referência metodológica que poderá orientar futuras iniciativas utilizando bases de dados mais abrangentes, incluindo informações internas e variáveis adicionais que não estavam disponíveis neste trabalho. Além disso, o desenvolvimento desta pesquisa proporcionou uma compreensão mais aprofundada das etapas envolvidas na construção de modelos preditivos para análise de crédito, contribuindo para a consolidação do conhecimento adquirido ao longo da graduação e fornecendo direcionamentos para trabalhos futuros e aplicações em contextos reais.

O código-fonte desenvolvido neste trabalho está disponível em um repositório no GitHub, devidamente referenciado nas referências bibliográficas (HAMERSKI, 2026), com o objetivo de garantir transparência, reprodutibilidade dos experimentos e possibilitar a continuidade ou extensão da pesquisa. O repositório também inclui a interface demonstrativa desenvolvida como complemento ao estudo.

Por fim, conclui-se que o estudo atingiu seu objetivo ao demonstrar uma aplicação prática de aprendizado de máquina no contexto pós-concessão de crédito, servindo como base para futuras extensões. Entre as possibilidades de trabalhos futuros, destacam-se a utilização de bases mais robustas, o aprimoramento dos modelos aplicados, a integração da solução em uma interface interativa mais completa para visualização e apoio à análise de risco, bem como a realização de validações periódicas utilizando dados mais recentes. Essa continuidade permitiria acompanhar eventuais alterações no comportamento da carteira de crédito, monitorar a degradação do desempenho dos modelos ao longo do tempo e definir estratégias adequadas de atualização e retreinamento.

# 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.

BERRADA, I. R.; BARRAMOU, F.; ALAMI, O. B. Towards a machine learning-based model for corporate loan default prediction. **International Journal of Advanced Computer Science and Applications**, v. 15, n. 3, 2024. Disponível em: http://dx.doi.org/10.14569/IJACSA.2024.0150357.

CAETANO, T. M. TCC, **Algoritmos de aprendizado de máquina no estudo da inadimplência em uma instituição financeira**. 2024. Disponível em: https://repositorio.ufu.br/handle/123456789/41843.

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

Git Project. **Git**. 2025. https://git-scm.com/. Acesso em: 15 nov. 2025.

GitHub, Inc. **GitHub**. 2025. https://github.com/. Acesso em: 15 nov. 2025.

Google Developers. **Classificação: ROC e AUC | Machine Learning Crash Course**. 2026. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc?hl=pt-br. Acesso em: 21 jun. 2026.

HAMERSKI, M. **Análise exploratória e modelagem de dados para inadimplência**. 2026. https://github.com/milenahamerski/analise-exploratoria-scr. Repositório GitHub. Acesso em: 14 jun. 2026.

HAN, J.; PEI, J.; TONG, H. **Data Mining: Concepts and Techniques**. [*S.l.*]: Morgan Kaufmann, 2022.

KELLEHER, J.; NAMEE, B.; D'ARCY, A. **Fundamentals of Machine Learning for Predictive Data Analytics**. MIT Press, 2020. 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, p. e5138, 2025. Disponível em: https://ojs.revistadelos.com/ojs/index.php/delos/article/view/5138.

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

NumPy Developers. **NumPy Documentation**. 2025. https://numpy.org/. Acesso em: 15 nov. 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 Developers. **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 Developers. **Polars: DataFrames for the New Era**. 2025. https://pola.rs/. Acesso em: 15 de Novembro de 2025.

Project Jupyter. **Project Jupyter**. 2025. https://jupyter.org/. Acesso em: 15 nov. 2025.

Python Software Foundation. **Python Documentation**. 2025. https://www.python.org/. Acesso em: 15 nov. 2025.

Scikit-learn Developers. **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 Developers. **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/.

Serasa Experian. **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/.

XGBoost Developers. **XGBoost Documentation**. 2025. https://xgboost.readthedocs.io/. Acesso em: 15 nov. 2025.

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**, 2024. Disponível em: https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2024.1430245.

# APÊNDICE A – Códigos de Pré-processamento dos Dados

Este apêndice apresenta os principais arquivos utilizados na etapa de pré-processamento dos dados. O objetivo é documentar a estrutura do pipeline empregado para carregamento, limpeza, transformação e particionamento da base utilizada nos experimentos.

## A.1 Arquivo _load_data.py

O arquivo `_load_data.py` é responsável pelo carregamento inicial da base de dados e pela seleção das colunas que serão utilizadas nas etapas posteriores.

```
import pandas as pd

def load_data(path):
    df = pd.read_csv(path, sep=";")

    df = df[[
        "uf", "cliente", "cnae_ocupacao",
        "porte", "modalidade", "submodalidade",
        "carteira_vencida",
        "numero_de_operacoes",
        "a_vencer_ate_90_dias",
        "a_vencer_de_91_ate_360_dias",
        "a_vencer_de_361_ate_1080_dias",
        "a_vencer_de_1081_ate_1800_dias",
        "a_vencer_de_1801_ate_5400_dias",
        "a_vencer_acima_de_5400_dias",
        "carteira_a_vencer"
    ]]

    return df
```

## A.2 Arquivo _cleaning.py

O arquivo `_cleaning.py` concentra as regras de limpeza dos dados, conversões de tipos (como strings monetárias para float), tratamento de dados faltantes e a definição da variável alvo (*target*) do modelo de inadimplência.

```
import numpy as np
import pandas as pd

def clean_data(df):
    df = df[df["cliente"] == "PF"].copy()

    df["carteira_vencida"] = (
        df["carteira_vencida"]
```

```
        .astype(str)
        .str.replace(",", ".", regex=False)
        .astype(float)
    )
    df["carteira_vencida"] = (df["carteira_vencida"] > 0).astype(int)

    df["porte"] = df["porte"].fillna("Indisponível")
    df["submodalidade"] = df["submodalidade"].fillna("desconhecido")

    # Conversão de valores monetários de string (vírgula) para float
    cols_valores = [
        "a_vencer_ate_90_dias",
        "a_vencer_de_91_ate_360_dias",
        "a_vencer_de_361_ate_1080_dias",
        "a_vencer_de_1081_ate_1800_dias",
        "a_vencer_de_1801_ate_5400_dias",
        "a_vencer_acima_de_5400_dias",
        "carteira_a_vencer"
    ]

    for col in cols_valores:
        df[col] = (
            df[col].astype(str)
            .str.replace(",", ".", regex=False)
        )
        df[col] = df[col].astype(float)

    df["carteira_a_vencer"] = df["carteira_a_vencer"].replace(0, 1)

    # operações
    df["numero_de_operacoes"] = df["numero_de_operacoes"].replace("<= 15", 15)
    df["numero_de_operacoes"] = pd.to_numeric(df["numero_de_operacoes"], errors="
    coerce")

    df["operacoes_missing"] = (df["numero_de_operacoes"] < 0).astype(int)
    df.loc[df["numero_de_operacoes"] < 0, "numero_de_operacoes"] = np.nan

    df["numero_de_operacoes"] = df["numero_de_operacoes"].fillna(
        df["numero_de_operacoes"].median()
    )

    return df
```

## A.3 Arquivo _scenarios.py

Este arquivo contém a lógica para testar diferentes representações da variável `submodalidade`, permitindo avaliar qual abordagem (remoção, agrupamento por frequência ou categorização de negócios) gera os melhores resultados.

```
def _agrupar_submodalidade(df, threshold=1000):
    freq = df["submodalidade"].value_counts()
    validas = freq[freq > threshold].index

    df["submodalidade"] = df["submodalidade"].apply(
        lambda x: x if x in validas else "Outros"
    )
    return df

def _group_submodalidade_engineered(df):
    def get_group(sub):
        s = str(sub).lower().strip()

        # 1. Cartão de Crédito
        if "cartão de crédito" in s or "cartao de credito" in s \
           or "fatura de cartão" in s or "fatura de cartao" in s:
            return "Cartão de Crédito"

        # 2. Crédito Pessoal
        elif "crédito pessoal" in s or "credito pessoal" in s:
            return "Crédito Pessoal"

        # 3. Financiamento Habitacional / Imobiliário
        elif "habitacional" in s or "imobiliário" in s \
             or "imobiliario" in s or "home equity" in s:
            return "Financiamento Habitacional"

        # 4. Financiamento de Veículos / Arrendamento
        elif "veículos automotores" in s or "veiculos automores" in s \
             or "veículos autom." in s or "veiculos autom." in s \
             or "arrendamento" in s:
            return "Financiamento de Veículos"

        # 5. Crédito Rural / Agroindustrial
        elif "custeio" in s or "comercialização" in s \
             or "comercializacao" in s or "agroindustriais" in s:
            return "Crédito Rural"

        # 6. Capital de Giro
        elif "capital de giro" in s or "conta garantida" in s:
```

```
            return "Capital de Giro"

        # 7. Microcrédito
        elif "microcrédito" in s or "microcredito" in s:
            return "Microcrédito"

        # 8. Cheque Especial
        elif "cheque especial" in s:
            return "Cheque Especial"

        # 9. Aquisição de Bens
        elif "aquisição de bens" in s or "aquisicao de bens" in s:
            return "Aquisição de Bens"

        # 10. Outros / Não classificados
        else:
            return "Outros"

    df["submodalidade_grupo"] = df["submodalidade"].apply(get_group)
    return df

def apply_scenario(df, scenario):

    if scenario == "sem_submodalidade":
        return df.drop("submodalidade", axis=1)

    elif scenario == "submodalidade_agrupada":
        return _agrupar_submodalidade(df)

    elif scenario == "submodalidade_engineered":
        df = _group_submodalidade_engineered(df)
        return df.drop("submodalidade", axis=1)

    else:
        raise ValueError("Cenário inválido")
```

## A.4 Arquivo _split.py

O arquivo `_split.py` prepara a base final, aplicando as codificações necessárias (*One-Hot Encoding*) e particionando os dados em conjuntos de treino e teste.

```
from sklearn.model_selection import train_test_split
import pandas as pd

def split_data(df, use_smote=False):
```

```
5
6       X = df.drop(["carteira_vencida", "cliente"], axis=1)
7       y = df["carteira_vencida"]
8
9       X = pd.get_dummies(X, drop_first=True)
10      X = X.apply(pd.to_numeric)
11
12      X_train, X_test, y_train, y_test = train_test_split(
13          X, y,
14          test_size=0.3,
15          random_state=42,
16          stratify=y
17      )
18
19      # Nota: O SMOTE foi removido daqui para evitar vazamento de dados (data leakage)
20      # durante a validação cruzada. Agora o SMOTE deve ser aplicado via Pipeline
21      # nos notebooks de modelagem.
22
23      return X_train, X_test, y_train, y_test
```

### A.5 Arquivo main_preprocessing.py

O arquivo orquestrador `main_preprocessing.py` unifica todos os módulos do *pipeline* descritos anteriormente. Ele recebe o caminho da base bruta e um cenário, executando as funções de forma sequencial.

```
1  from preprocessing._load_data import load_data
2  from preprocessing._cleaning import clean_data
3  from preprocessing._scenarios import apply_scenario
4  from preprocessing._split import split_data
5
6  def load_and_preprocess(path, scenario, use_smote=True):
7
8      df = load_data(path)
9      df = clean_data(df)
10     df = apply_scenario(df, scenario)
11
12     return split_data(df, use_smote)
```

# APÊNDICE B – Treinamento e Otimização de Hiperparâmetros

Este apêndice documenta as rotinas desenvolvidas para o treinamento e a otimização dos modelos preditivos.

Para maximizar a capacidade de generalização dos algoritmos, foi implementada uma busca exaustiva de hiperparâmetros aliada à validação cruzada (*Cross-Validation*). O trecho de código a seguir ilustra o padrão arquitetural adotado neste estudo, utilizando como exemplo um classificador baseado em Árvore de Decisão (*Decision Tree*).

Para mitigar o vazamento de dados (*data leakage*) entre as partições de treino e validação, o balanceamento de classes sintético gerado pelo método SMOTE foi estritamente encapsulado dentro do `Pipeline` da biblioteca *Imbalanced-Learn*. Adicionalmente, a presença do balanceamento (SMOTE) foi tratada como um próprio hiperparâmetro (passando o valor nulo `'passthrough'`), o que permitiu avaliar durante o *GridSearch* o desempenho do modelo com e sem a geração sintética de dados da classe minoritária.

## B.1 Exemplo de Rotina com GridSearchCV e Pipeline

```
from sklearn.model_selection import GridSearchCV
from imblearn.pipeline import Pipeline
from imblearn.over_sampling import SMOTE
from sklearn.tree import DecisionTreeClassifier

# Definicao do Pipeline integrando SMOTE e o classificador
pipeline_dt = Pipeline([
    ('smote', SMOTE(random_state=42)),
    ('dt', DecisionTreeClassifier(random_state=42))
])

# Parametros para busca: 'passthrough' instrui a execucao sem o SMOTE
param_grid_dt = {
    'smote': ['passthrough', SMOTE(random_state=42)],
    'dt__criterion': ['gini', 'entropy'],
    'dt__max_depth': [None, 5, 10, 20],
    'dt__min_samples_split': [2, 5, 10],
    'dt__min_samples_leaf': [1, 2, 4]
}

# Configuracao do GridSearchCV (5 folds) com otimizacao voltada ao ROC AUC
grid_dt = GridSearchCV(
    estimator=pipeline_dt,
    param_grid=param_grid_dt,
    cv=5,
    scoring="roc_auc",
    n_jobs=-1
)
```

```
29
30 # Execucao do treinamento
31 grid_dt.fit(X_train, y_train)
32
33 # Extracao do melhor modelo
34 best_dt = grid_dt.best_estimator_
35
36 # Predicao nas amostras de teste invisiveis ao modelo durante o treinamento
37 y_pred = best_dt.predict(X_test)
38 y_proba = best_dt.predict_proba(X_test)[:, 1]
```