UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

MILENA HAMERSKI

PREDIÇÃO DE INADIMPLÊNCIA NO PÓS-CONCESSÃO DE CRÉDITO: APRENDIZADO DE MÁQUINA COM DADOS DO SISTEMA DE INFORMAÇÕES DE CRÉDITO (SCR)

GUARAPUAVA

2026

MILENA HAMERSKI

PREDIÇÃO DE INADIMPLÊNCIA NO PÓS-CONCESSÃO DE CRÉDITO: APRENDIZADO DE MÁQUINA COM DADOS DO SISTEMA DE INFORMAÇÕES DE CRÉDITO (SCR)

Post-Disbursement Credit Default Prediction: Machine Learning Using Data from the Brazilian Credit Information System (SCR)

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. Me. Dênis Lucas Silva

Coorientador: Profª. Drª Kelly Lais Wiggers

GUARAPUAVA

2026

MILENA HAMERSKI

PREDIÇÃO DE INADIMPLÊNCIA NO PÓS-CONCESSÃO DE CRÉDITO: APRENDIZADO DE MÁQUINA COM DADOS DO SISTEMA DE INFORMAÇÕES DE CRÉDITO (SCR)

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.

Data de aprovação: 02/julho/2026

Prof. Dênis Lucas Silva
Mestre
Universidade Tecnológica Federal do Paraná

Prof. Eleandro Maschio
Doutor
Universidade Tecnológica Federal do Paraná

Prof. Renata Luiza Stange Carneiro Gomes
Doutora
Universidade Tecnológica Federal do Paraná

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.

Ao meu orientador, Prof. Me. Dênis Lucas Silva, agradeço por ter aceitado conduzir a etapa final deste projeto e por toda a disponibilidade, atenção e contribuição durante sua realização.

À minha coorientadora, Prof.ª Dra. Kelly Lais Wiggers, minha profunda gratidão por todo o apoio, dedicação e orientação ao longo deste trabalho. 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.

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, transformação de variáveis, redução da alta cardinalidade categórica e avaliação de estratégias de balanceamento de classes. 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 as métricas de Acurácia, F1-Score e ROC AUC. Os resultados demonstraram que a estratégia de representação das variáveis categóricas exerceu maior influência no desempenho dos modelos do que as técnicas de balanceamento avaliadas, destacando-se o algoritmo *Random Forest* aplicado ao cenário de agrupamento por frequência da variável submodalidade, que obteve o melhor desempenho entre as configurações analisadas. Como complemento ao estudo, foi desenvolvido um Produto Mínimo Viável (MVP), composto por uma interface web integrada a uma API, para demonstrar uma possível aplicação prática do modelo preditivo. 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 Brazilian 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 they have been granted. To conduct the study, the dataset underwent several preprocessing stages, including missing value treatment, feature transformation, reduction of high-cardinality categorical variables, and the evaluation of class balancing strategies. 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 showed that the representation strategy adopted for categorical variables had a greater impact on model performance than the evaluated class balancing techniques, with the Random Forest algorithm applied to the frequency-based grouping scenario for the *submodalidade* variable achieving the best overall performance. As a complement to the study, a Minimum Viable Product (MVP), consisting of a web interface integrated with an API, was developed to demonstrate a possible practical application of the predictive model. It is concluded that the proposed approach can contribute to post-credit granting monitoring and credit risk analysis, supporting 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

AM	Aprendizado de Máquina
CMN	Conselho Monetário Nacional
CNN	Convolutional Neural Network
FN	Falso Negativo
FP	Falso Positivo
IPCA	Índice Nacional de Preços ao Consumidor Amplo
MD	Mineração de Dados
PIB	Produto Interno Bruto
RNN	Recurrent Neural Network
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)
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo

SUMÁRIO

1 INTRODUÇÃO

1.1 Considerações iniciais

Nos últimos anos, o mercado de crédito brasileiro enfrentou um cenário de elevada instabilidade econômica, impulsionado pelos efeitos da pandemia da COVID-19 e por seus desdobramentos sobre a atividade econômica. Oscilações no Produto Interno Bruto (PIB), na inflação medida pelo Índice Nacional de Preços ao Consumidor Amplo (IPCA), na taxa de desemprego e na taxa básica de juros (Sistema Especial de Liquidação e Custódia (taxa básica de juros da economia brasileira) (SELIC)) afetaram diretamente a capacidade de pagamento das famílias, contribuindo para o aumento dos índices de inadimplência. Estudos recentes demonstram que a inadimplência apresenta forte relação com essas variáveis macroeconômicas, evidenciando a necessidade de mecanismos mais eficientes para identificação e monitoramento do risco de crédito (MAIA; RAMíREZ; CUTINO, 2025).

Esse cenário é refletido nos indicadores do mercado brasileiro. De acordo com o Mapa da Inadimplência e Negociação de Dívidas da Serasa (Serasa Experian, 2025), o número de brasileiros inadimplentes passou de 72,54 milhões em maio de 2024 para 78,8 milhões em agosto de 2025, representando aproximadamente metade da população adulta do país. Entre as principais origens das dívidas destacam-se aquelas relacionadas a bancos e cartões de crédito (27,3%), contas básicas de consumo (20,8%) e financeiras (19,5%), evidenciando a relevância do desenvolvimento de métodos capazes de apoiar a análise e o gerenciamento do risco de crédito.

Nesse contexto, a Inteligência Artificial e, em especial, o Aprendizado de Máquina (Aprendizado de Máquina (AM)) vêm sendo amplamente aplicados ao setor financeiro devido à capacidade de identificar padrões complexos em grandes volumes de dados e gerar modelos preditivos para diferentes problemas de negócio. Estudos recentes demonstram que a integração dessas técnicas com análise de grandes volumes de dados (*big data*[1]), Mineração de Dados (MD) e métodos estatísticos possibilita o desenvolvimento de modelos capazes de apoiar a avaliação do risco de crédito e a identificação de padrões associados à inadimplência (XIANYU; HAI, 2023).

Embora grande parte das pesquisas concentre seus esforços na etapa de concessão do crédito, o período pós-concessão também representa um desafio estratégico para as instituições financeiras. Após a liberação dos recursos, torna-se necessário acompanhar continuamente o comportamento das operações, monitorar a evolução do risco da carteira e identificar clientes com maior probabilidade de inadimplência. A identificação antecipada dessas situações pos-

[1] O termo *big data* refere-se ao conjunto de técnicas e ferramentas utilizadas para coletar, armazenar, processar e analisar grandes volumes de dados, estruturados ou não, visando à extração de informações relevantes.

sibilita direcionar ações preventivas, apoiar estratégias de recuperação de crédito e otimizar a alocação de recursos destinados ao gerenciamento do risco.

Nesse cenário, modelos preditivos baseados em aprendizado de máquina podem atuar como ferramentas de apoio ao monitoramento da carteira de crédito, identificando padrões dificilmente perceptíveis por métodos convencionais e contribuindo para reduzir subjetividades e aumentar a consistência das análises. Diante desse contexto, este trabalho propõe desenvolver e avaliar modelos de aprendizado de máquina para predição de inadimplência no período pós-concessão de crédito, investigando seu potencial para identificar operações com maior propensão ao aumento do risco e apoiar a tomada de decisão em instituições financeiras.

1.2 Objetivos

1.2.1 Objetivo geral

Avaliar o desempenho de modelos de aprendizado de máquina na predição de inadimplência no período pós-concessão de operações de crédito de pessoas físicas, utilizando dados históricos do Sistema de Informações de Crédito (SCR).

1.2.2 Objetivos específicos

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

- Compreender o ciclo de crédito e o período pós-concessão sob a perspectiva da modelagem de risco de crédito e definição do evento de inadimplência.
- Investigar e selecionar técnicas de aprendizado de máquina supervisionado aplicáveis à predição de inadimplência em operações de crédito de pessoas físicas.
- Preparar e tratar os dados do Sistema de Informações de Crédito (SCR), incluindo limpeza, tratamento de valores ausentes e engenharia de atributos para modelagem preditiva.
- Treinar, ajustar e avaliar modelos de aprendizado de máquina supervisionado para predição de inadimplência, utilizando técnicas de validação e métricas adequadas de desempenho.
- Comparar o desempenho dos modelos por meio de métricas de avaliação, identificando suas capacidades de generalização e limitações na predição de inadimplência.

1.3 Justificativa

A inadimplência no período pós-concessão de crédito representa um dos principais desafios na gestão de risco das instituições financeiras, uma vez que corresponde à fase em que o comportamento de pagamento do tomador se manifesta após a liberação do crédito. Diferentemente da etapa de concessão, na qual o risco é estimado com base em informações históricas e cadastrais, o período pós-concessão permite observar a evolução do comportamento financeiro ao longo do tempo, quando o risco de crédito se materializa de forma mais evidente.

No contexto do sistema financeiro, o crescimento do volume de operações de crédito e a maior complexidade das carteiras tornam cada vez mais relevante a adoção de abordagens analíticas capazes de apoiar a compreensão do comportamento de inadimplência. Nesse cenário, técnicas de aprendizado de máquina têm sido amplamente exploradas por sua capacidade de identificar padrões complexos em grandes volumes de dados, especialmente em situações nas quais há relações não lineares entre variáveis associadas ao risco de crédito.

A utilização de bases de dados abertas, como aquelas derivadas do Sistema de Informações de Crédito (SCR), viabiliza a análise de padrões de comportamento de crédito em escala populacional, contribuindo para estudos reprodutíveis e orientados a evidências empíricas. Embora essas bases não apresentem o mesmo nível de granularidade de informações observadas em ambientes internos de instituições financeiras, elas oferecem uma visão agregada e consistente das operações de crédito ao longo do tempo, favorecendo a investigação de padrões associados à inadimplência.

Nesse contexto, a aplicação de modelos preditivos possibilita a transformação de dados históricos de crédito em informações analíticas que auxiliam na identificação de padrões relacionados ao risco de inadimplência. Essa abordagem também contribui para a estruturação da análise de risco, reduzindo parcialmente a dependência de avaliações subjetivas e permitindo a exploração sistemática de relações entre variáveis observáveis e o comportamento de pagamento.

Além disso, tais modelos favorecem a padronização e a consistência das análises de risco de crédito, especialmente em atividades relacionadas ao monitoramento de carteiras e à identificação de perfis com maior propensão à inadimplência. Isso se torna particularmente relevante em contextos com grande volume de dados, nos quais abordagens tradicionais podem apresentar limitações em termos de escalabilidade e uniformidade.

Do ponto de vista econômico e social, a inadimplência impacta tanto a sustentabilidade das instituições financeiras quanto o acesso ao crédito por parte dos clientes. Dessa forma, o uso de dados públicos aliados a técnicas de aprendizado de máquina contribui para uma melhor compreensão do fenômeno da inadimplência e para o desenvolvimento de análises mais estruturadas no contexto de risco de crédito.

2 REFERENCIAL TEÓRICO

2.1 Ciclo de Crédito e Mecanismos de Controle

O ciclo de crédito compreende um conjunto de etapas sequenciais que vão desde a definição das políticas internas das instituições financeiras até a liquidação efetiva das operações. Inicialmente, a instituição estabelece sua política de crédito, composta por diretrizes e estratégias voltadas à mitigação do risco de crédito, definido como a possibilidade de perdas financeiras decorrentes do não cumprimento das obrigações contratuais por parte do tomador de recursos (OLIVEIRA, 2024).

Na etapa de concessão, realiza-se uma análise detalhada do perfil do tomador em relação às condições vigentes. Após a liberação dos recursos, inicia-se a fase de acompanhamento, na qual as operações são monitoradas continuamente para permitir a identificação precoce da deterioração do risco e a aplicação de ajustes contratuais tempestivos. Por fim, a fase de cobrança e liquidação envolve estratégias de resolução de pendências, como renegociações e ações judiciais de recuperação, visando minimizar as perdas financeiras (OLIVEIRA, 2024).

No contexto nacional, o Sistema de Informações de Crédito (SCR), gerido pelo Banco Central do Brasil (BCB), atua como um dos principais mecanismos de apoio à gestão do risco de crédito. O SCR consolida dados detalhados sobre operações de crédito ativas enviadas pelas instituições financeiras, abrangendo tanto pessoas físicas quanto jurídicas. Diferentemente de bureaus de crédito privados de registro de inadimplentes, o SCR possui uma finalidade predominantemente prudencial e regulatória, servindo ao regulador como instrumento de supervisão macroprudencial da estabilidade do sistema financeiro.

O Banco Central do Brasil divulga mensalmente estatísticas agregadas derivadas dessa base, com uma defasagem aproximada de 30 dias (Conselho Monetário Nacional; Banco Central do Brasil, 2022). Esses indicadores macroeconômicos segmentados (como carteira ativa, inadimplência e ativo problemático) apoiam o gerenciamento de risco e servem de subsídio para decisões estratégicas de alocação de crédito.

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

2.2 Mineração de Dados

A mineração de dados (MD), no contexto da análise de risco 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). No âmbito financeiro, essas técnicas viabilizam a detecção precoce de fraudes, a classificação de risco e a identificação de anomalias que escapam às ferramentas estatísticas tradicionais.

O processo típico de mineração de dados envolve um ciclo contínuo estruturado em etapas que englobam desde a coleta e integração de fontes heterogêneas, passando pelo pré-processamento, pela seleção e engenharia de atributos relevantes, até a modelagem preditiva e a validação de seus outputs. Cada uma dessas fases é crítica para assegurar a consistência dos modelos, permitindo que a análise de risco de crédito pós-concessão se sustente em evidências empíricas reprodutíveis e robustas. A natureza iterativa e cíclica desse processo de descoberta é representada no ciclo de vida dos dados (Figura 1), no qual a avaliação dos resultados retroalimenta a base de dados para refinar continuamente a qualidade das previsões e o monitoramento do risco.

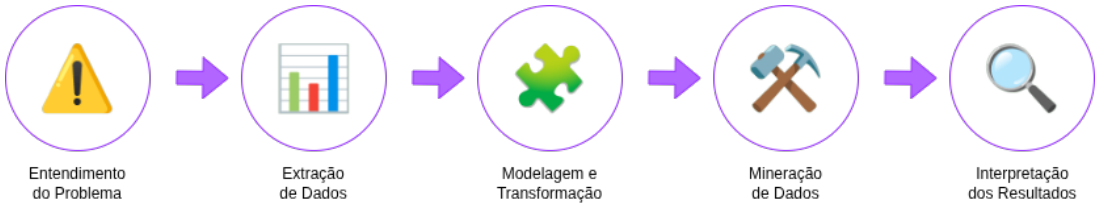

Figura 1 – Etapas do processo de Mineração de Dados. Adaptado de DBM Sistemas

2.2.1 Pré-processamento de Dados

Antes da modelagem por meio de algoritmos de aprendizado de máquina, é indispensável executar etapas de pré-processamento para higienizar e estruturar os dados brutos. Essa

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

preparação envolve o tratamento de inconsistências lógicas, a imputação de valores ausentes, a mitigação de ruído e a identificação de *outliers*[2], além de ajustes de escala e variância por meio de técnicas de padronização ou normalização estatística.

Adicionalmente, os datasets reais de concessão de crédito enfrentam o desafio do desbalanceamento de classes, caracterizado por uma quantidade significativamente maior de registros de adimplentes (classe majoritária) em detrimento de casos de inadimplência (classe minoritária). Essa disparidade estrutural induz vieses nos classificadores, que tendem a apresentar alta acurácia global à custa de uma ineficácia na detecção do verdadeiro risco (OLIVEIRA, 2024). Para balancear a distribuição de classes, empregam-se técnicas de reamostragem, divididas em:

- **Subamostragem (*Undersampling*):** reduz de forma aleatória ou heurística o número de amostras da classe majoritária, visando equilibrar a proporção com a classe minoritária.
- **Sobreamostragem (*Oversampling*):** expande o volume de amostras da classe minoritária, seja pela replicação de dados ou pela geração de instâncias sintéticas.

Entre os métodos sintéticos de sobreamostragem, destaca-se o Synthetic Minority Oversampling Technique (Técnica Sintética de Sobreamostragem da Classe Minoritária) (SMOTE) (*Synthetic Minority Over-sampling Technique*). Em vez de replicar observações existentes, o algoritmo identifica os k-vizinhos mais próximos de cada amostra da classe minoritária e gera sinteticamente novos pontos de dados distribuídos aleatoriamente ao longo dos segmentos de reta que unem essas instâncias vizinhas no espaço multidimensional (CAETANO, 2024).

O cuidado crítico na implementação dessas técnicas reside na estrita prevenção do vazamento de dados (*Data Leakage*). O balanceamento com o SMOTE deve ocorrer exclusivamente sobre as partições de treinamento. Se o balanceamento for aplicado ingenuamente à base de dados completa antes do particionamento entre treino e teste, as instâncias sintéticas geradas carregarão informações da partição de teste que vazarão para o treinamento. Esse erro metodológico gera métricas de validação ilusoriamente infladas e resulta em modelos ineficazes no ambiente produtivo real.

É fundamental destacar que o SMOTE e demais técnicas de sobreamostragem nunca devem, sob hipótese alguma, ser aplicados à base de dados completa antes da divisão entre os conjuntos de treinamento e teste. O correto procedimento metodológico exige que a geração de instâncias sintéticas ocorra exclusivamente sobre o conjunto de treinamento já isolado (e, no caso de validação cruzada, estritamente dentro de cada dobra de treino). Conforme alertado por Demircioğlu (2024), a aplicação de métodos de sobreamostragem previamente à validação cruzada induz um alto viés otimista (*high bias*) nas métricas de avaliação, mascarando a real capacidade de generalização do classificador e comprometendo a integridade dos testes de validação.

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

2.2.2 Engenharia de Atributos

A engenharia de atributos (*feature engineering*) consiste no processo de criação, seleção ou transformação de variáveis com o objetivo de melhorar a representação dos dados para os algoritmos de aprendizado de máquina. A qualidade dessa representação influencia diretamente a capacidade dos modelos em identificar padrões e estabelecer relações entre os atributos preditores e a variável-alvo (KELLEHER; NAMEE; D'ARCY, 2020).

Segundo Kelleher, Namee e D'Arcy (2020), a etapa de exploração dos dados permite compreender características importantes das variáveis, como distribuição, variabilidade e presença de atributos categóricos com elevada cardinalidade. Essas características orientam a definição de transformações que tornam os dados mais adequados ao processo de aprendizagem.

No contexto de dados tabulares, uma das abordagens mais utilizadas para representar variáveis categóricas é o *One-Hot Encoding*, que converte cada categoria em uma variável binária. Entretanto, quando aplicado diretamente a atributos com grande número de categorias distintas, esse método pode aumentar significativamente a dimensionalidade do conjunto de dados, produzindo matrizes esparsas e impactando negativamente o desempenho de alguns algoritmos de aprendizado de máquina.

Dessa forma, a literatura recomenda que a etapa de engenharia de atributos considere estratégias de transformação e consolidação das variáveis antes da codificação, buscando reduzir a complexidade dos dados e preservar as informações mais relevantes para a modelagem (KELLEHER; NAMEE; D'ARCY, 2020).

2.3 Aprendizado de Máquina

O aprendizado de máquina (AM) é um subcampo da Inteligência Artificial focado no desenvolvimento de sistemas computacionais que extraem padrões de dados históricos de forma autônoma para aprimorar suas previsões de forma iterativa. Essa tecnologia fundamenta-se na modelagem matemática de algoritmos de otimização que generalizam padrões a partir de instâncias conhecidas para rotular instâncias futuras e não vistas. No âmbito analítico do risco de crédito, a modelagem preditiva apoia as decisões financeiras ao identificar padrões de inadimplência não lineares e complexos em bases estruturadas (HAN; PEI; TONG, 2022).

Os paradigmas de aprendizado de máquina diferenciam-se principalmente pelo formato de treinamento e pela presença ou ausência de rótulos de supervisão nos dados, dividindo-se nas seguintes categorias:

- **Aprendizado Supervisionado:** utiliza conjuntos de dados onde cada observação possui um rótulo conhecido[3], sendo a abordagem padrão para problemas de classificação binária de risco (adimplente versus inadimplente).
- **Aprendizado Não Supervisionado:** opera com dados sem rotulação prévia, identificando estruturas de agrupamento ou comportamentos atípicos intrínsecos à distribuição dos dados, com aplicações comuns na segmentação comportamental de clientes.
- **Aprendizado por Reforço:** baseia-se no aprendizado contínuo de um agente interativo por meio de feedbacks de recompensa ou penalização associados a tomadas de decisão sequenciais.

2.3.1 Algoritmos de Aprendizado de Máquina

No desenvolvimento de modelos de predição de inadimplência, o aprendizado supervisionado desempenha um papel central ao aprender a relação entre os atributos da base de dados e a variável-alvo que representa a ocorrência de atraso. Segundo Kelleher, Namee e D'Arcy (2020), esse processo envolve duas etapas principais: o treinamento do modelo com dados históricos e sua posterior validação em dados não utilizados durante o treinamento. Essa abordagem reduz a dependência de análises manuais e de regras heurísticas, sendo especialmente adequada para problemas com grande volume de dados e múltiplas variáveis.

Estudos empíricos aplicados ao risco de crédito atestam a necessidade de testar múltiplas hipóteses de modelagem e arquiteturas algorítmicas. Conforme demonstrado por Xianyu e Hai (2023), modelos lineares simples, algoritmos baseados em distâncias métricas e estruturas ensemble (como Random Forest e redes neurais profundas) respondem de maneiras distintas a distorções nos dados e ao desbalanceamento de classes, exigindo uma seleção empírica rigorosa das técnicas de modelagem. Com base nesses referenciais, as principais abordagens de classificação aplicadas a bases tabulares estão consolidadas na Tabela 1, detalhando suas características de funcionamento e as bases técnico-teóricas que guiam o treinamento de seus estimadores.

2.3.2 Estratégias de Validação de Modelos

Na construção de modelos de aprendizado de máquina, a capacidade de generalização para dados inéditos constitui o principal critério de validação. Para estimar o desempenho dos classificadores sem incorrer em sobreajuste, a base de dados original é dividida em dois subconjuntos distintos: o conjunto de treino, destinado ao ajuste dos parâmetros internos do

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

Categoria e Algoritmo	Descrição Técnico-Teórica
Modelos Lineares: Regressão Logística	Modelo estatístico para classificação binária que utiliza a função logística (sigmoide) para estimar a probabilidade de ocorrência da classe-alvo.
Baseados em Distância: K-Nearest Neighbors (KNN)	Algoritmo não paramétrico baseado em instâncias que classifica uma amostra de acordo com a classe predominante entre seus k vizinhos mais próximos.
Baseados em Particionamento: Árvore de Decisão (DT)	Modelo não paramétrico que particiona recursivamente os dados por meio de critérios de pureza, gerando regras de decisão de fácil interpretação.
Ensemble de *Bagging*: Random Forest (RF)	Conjunto de múltiplas árvores de decisão construídas por *bagging* e amostragem aleatória de atributos, reduzindo variância e o risco de *overfitting*.
Otimização de Margem: Máquinas de Vetores de Suporte (SVM)	Algoritmo que busca o hiperplano de separação com maior margem entre as classes, podendo utilizar funções *kernel* para problemas não lineares.
Ensemble de *Boosting*: XGBoost e AdaBoost	Modelos de *boosting* que constroem preditores de forma sequencial, buscando corrigir os erros das iterações anteriores.
Redes Neurais Artificiais: MLP, CNN e RNN	Modelos de aprendizado profundo compostos por camadas de neurônios artificiais, capazes de aprender padrões complexos em diferentes tipos de dados.
Agrupamento Não Supervisionado: Clustering K-means	Algoritmo de agrupamento que particiona os dados em k grupos, minimizando a distância entre as amostras e o centroide de cada grupo.

Tabela 1 – Principais algoritmos de Aprendizado de Máquina.

estimador, e o conjunto de teste, mantido sob isolamento estrito para servir como simulador do comportamento do modelo em um cenário produtivo real (CAETANO, 2024).

Para cenários com forte desbalanceamento e dados altamente dimensionais, adota-se a validação cruzada do tipo *k-fold* para evitar vieses de seleção. Esse protocolo divide a partição de treino em k subconjuntos homogêneos (dobras), alternando iterativamente o treinamento em $k-1$ partições e a validação na partição residual restante. A aplicação integrada de etapas de escala e balanceamento (como o SMOTE) exige o uso de pipelines computacionais (*Pipelines*) para assegurar que essas operações estatísticas sejam computadas estritamente sob as dobras de treino correspondentes em cada iteração de validação, neutralizando o risco de vazamento de informações e a superestimação do desempenho preditivo nos testes (OLIVEIRA, 2024; DEMIRCIOĞLU, 2024).

Associado ao processo de modelagem, realiza-se a otimização de hiperparâmetros por meio de algoritmos de busca exaustiva em grade (*Grid Search*). Essa técnica avalia recursivamente combinações discretas de hiperparâmetros definidos pelo cientista (como limites de profundidade de árvores de decisão ou coeficientes de regularização de modelos lineares) por meio de validação cruzada, identificando as parametrizações que garantam a maximização do poder preditivo e a resiliência estatística do estimador (KELLEHER; NAMEE; D'ARCY, 2020).

2.3.3 Métricas de Avaliação

A avaliação rigorosa do desempenho preditivo exige a verificação de desvios como o *overfitting* (sobreajuste), fenômeno no qual o estimador minimiza a função de perda nos dados de treino por meio da memorização de padrões específicos do ruído, mas falha em generalizar suas previsões contra dados de teste. A Figura 2 ilustra esse comportamento e evidencia a divergência entre a separação ótima e a fronteira excessivamente complexa de um estimador sobreajustado.

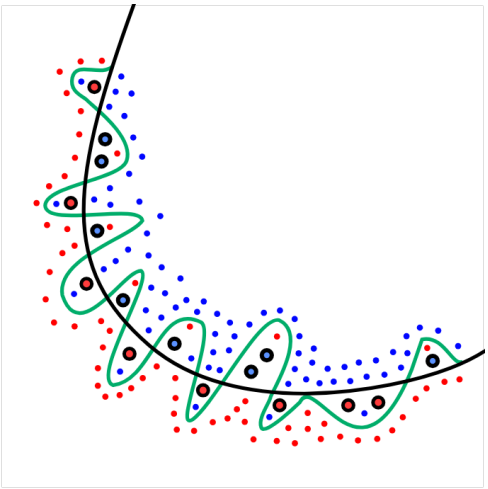

Figura 2 – Representação do fenômeno de *overfitting*. Fonte: Wikipédia.

Para mitigar diagnósticos equivocados em bases de dados de crédito desbalanceadas, a análise do desempenho estatístico apoia-se em múltiplas métricas derivadas da matriz de confusão, composta pelas frequências de verdadeiros positivos (Verdadeiro Positivo (VP)), verdadeiros negativos (Verdadeiro Negativo (VN)), falsos positivos (Falso Positivo (FP)) e falsos negativos (Falso Negativo (FN)). Enquanto a acurácia clássica mensura a proporção simples de acertos em relação ao total, seu indicador isolado mascara comportamentos de classificação ineficientes em classes minoritárias. Desse modo, adotam-se métricas específicas para avaliar a qualidade e a sensibilidade dos classificadores:

- **Sensibilidade (Recall):** Mede a capacidade do classificador de identificar corretamente as instâncias positivas da classe-alvo (inadimplentes), calculada pela razão entre verdadeiros positivos e a soma de verdadeiros positivos e falsos negativos.

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

- **Precisão (Precision):** Avalia a confiabilidade das previsões da classe positiva, expressando o percentual de acertos dentre todos os casos rotulados pelo modelo como

pertencentes à classe de interesse.

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

- **F1-score:** Média harmônica ponderada entre precisão e sensibilidade, fornecendo uma métrica integrada de estabilidade do modelo aplicável a cenários assimétricos.

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

- **ROC AUC:** Área sob a curva de sensibilidade em relação ao percentual de falsos positivos (FPR) em múltiplos limiares de probabilidade decisórios. O indicador resume graficamente o desempenho do classificador (Figura **??**); valores próximos a 1 representam excelente poder de discriminação de classes, enquanto o patamar de 0,5 corresponde ao comportamento estatístico de uma classificação aleatória.

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

A curva ROC e a métrica AUC avaliam a capacidade discriminatória do classificador de forma independente do limiar de decisão (Google Developers, 2026). Como ilustrado na Figura 3, quanto mais próxima a curva estiver do canto superior esquerdo, melhor tende a ser o desempenho discriminatório do modelo. Em problemas de classificação desbalanceada, a combinação entre Acurácia, F1-Score e ROC AUC fornece uma avaliação mais abrangente do desempenho preditivo (YANG; KHORSHIDI; AICKELIN, 2024).

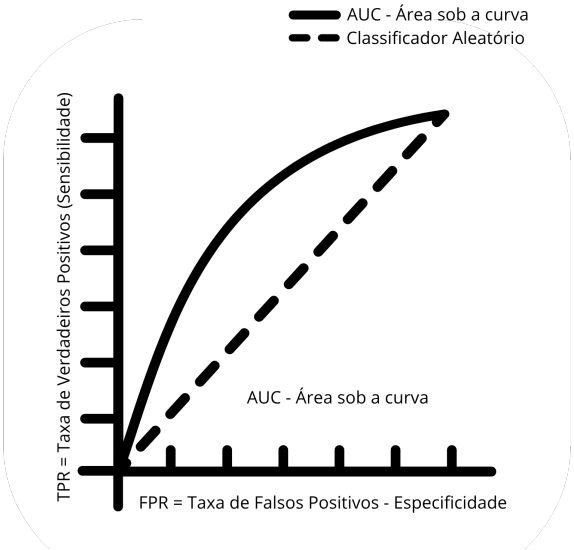

Figura 3 – Exemplo de curva ROC. Fonte: Elaboração própria.

3 TRABALHOS RELACIONADOS

Diversos estudos exploram a modelagem preditiva em ambientes bancários, avaliando estratégias analíticas para classificar o comportamento de tomadores de crédito. Essas referências fornecem subsídios conceituais e empíricos essenciais para orientar o desenvolvimento metodológico deste projeto de modelagem de risco pós-concessão.

3.1 Inadimplência e Crédito

A modelagem de crédito sob a perspectiva macroeconômica é analisada por Maia, Ramírez e Cutino (2025), que investigam a relação entre indicadores como inflação (IPCA), taxa de juros básica (SELIC) e o Produto Interno Bruto (PIB) frente às taxas de inadimplência brasileiras. O estudo confirma que a inadimplência apresenta um comportamento pró-cíclico sensível às oscilações e à instabilidade da economia real, ressaltando o valor preditivo da incorporação de dados externos na política de gestão de risco das instituições.

Sob a abordagem de dados estruturados individuais, Oliveira (2024) investiga modelos preditivos baseados em aprendizado de máquina aplicados ao crédito de pessoas físicas, integrando informações cadastrais com indicadores de recebimento de programas assistenciais federais. O estudo evidencia que a consolidação de informações socioeconômicas alternativas atua como redutor de assimetria informativa e melhora a precisão na concessão de crédito pessoal.

Além disso, a referida metodologia fundamenta-se no uso de bases de dados abertas governamentais para permitir a auditabilidade e a transparência do modelo, o que fornece as diretrizes para a estrutura de dados públicos do SCR utilizada nesta pesquisa. Esses estudos complementares reforçam a importância econômica e o impacto social do desenvolvimento de modelos mais estruturados para avaliar o risco de inadimplência tanto na concessão inicial quanto no monitoramento pós-concessão.

3.2 Seleção de Algoritmos

O delineamento experimental e a escolha das técnicas de aprendizado de máquina contam com referências importantes na literatura aplicada. O trabalho de Oliveira (2024) detalha a aplicação de pipelines de pré-processamento estruturados e a avaliação de algoritmos como Regressão Logística e Random Forest com foco na análise de variáveis socioeconômicas.

Em termos de comparação de desempenho e comportamento de redes neurais, Xianyu e Hai (2023) desenvolvem uma arquitetura preditiva comparando modelos clássicos a redes recorrentes (Recurrent Neural Network (RNN)) e convolucionais (Convolutional Neural Network

(CNN)) em problemas de crédito, ressaltando a influência crítica do ajuste sistemático de hiperparâmetros na estabilização dos estimadores.

Adicionalmente, Caetano (2024) apresenta em seu estudo em Estatística um pipeline analítico focado no balanceamento de classes desbalanceadas para detecção de inadimplência, evidenciando as vantagens comparativas de técnicas de reamostragem supervisionada. A convergência metodológica observada nesses trabalhos orienta a seleção de estimadores lineares, baseados em árvores e ensembles, além da adoção de protocolos consistentes de teste para validação dos resultados obtidos.

3.3 Síntese Comparativa dos Trabalhos

A Tabela 2 consolida as principais características dos estudos da literatura, destacando os tipos de bases de dados mapeadas, os principais algoritmos avaliados e o enfoque metodológico definido em cada pesquisa.

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
Demircioğlu (2024)	Dados reais (radiômica) e simulados	SMOTE, RF, SVM, Regressão Logística	Viés decorrente do oversampling antes da validação cruzada (data leakage)

Tabela 2 – Comparação entre trabalhos relacionados.

A convergência dos trabalhos destaca a consolidação do aprendizado de máquina para modelagem de classificação complexa, bem como a necessidade de rigor metodológico ao lidar com classes desbalanceadas e evitar vazamento de dados. A presente pesquisa diferencia-se das abordagens tradicionais ao concentrar a investigação sob o período pós-concessão de crédito a partir dos dados do SCR, atuando de maneira complementar aos modelos preditivos focados exclusivamente na etapa inicial de aprovação cadastral.

4 MATERIAIS E MÉTODOS

Este capítulo descreve o fluxo metodológico adotado no projeto, detalhando a fonte dos dados empíricos, as ferramentas tecnológicas empregadas e as etapas da experimentação. A condução da pesquisa baseou-se em um fluxo estruturado de aprendizado de máquina, com o intuito de treinar e avaliar modelos preditivos de risco de crédito, bem como viabilizar uma aplicação prática dos resultados obtidos.

4.1 Materiais

4.1.1 Base de Dados

Para a condução do experimento, foram utilizados dados públicos referentes a operações de crédito disponibilizados por meio do Sistema de Informações de Crédito (SCR), mantido pelo Banco Central do Brasil. A escolha do SCR justifica-se pela disponibilidade de dados reais e altamente dimensionais, atributos essenciais para a construção fidedigna de modelos preditivos de risco de crédito em larga escala.

A extração inicial abrangeu os registros governamentais do sistema. No entanto, o escopo deste estudo limitou-se ao segmento de Pessoas Físicas (PF). Assim, foi aplicado um filtro programático na base de dados para excluir operações vinculadas a Pessoas Jurídicas (PJ), resultando em um conjunto de dados focado unicamente no perfil de endividamento individual. Esta base caracterizou-se pela presença de atributos como modalidade de empréstimo, porte da ocupação do tomador, faixas temporais de vencimento e quantidade de operações ativas.

Os atributos originais selecionados para compor o espaço de características do modelo estão apresentados na Tabela 3.

4.1.2 Ferramentas e Tecnologias

Para a execução e orquestração do projeto, os seguintes recursos computacionais foram empregados, selecionados e organizados de acordo com sua função específica no fluxo de trabalho:

- **Linguagem e Ambiente**: O desenvolvimento ocorreu integralmente na linguagem Python, com o suporte de ambientes virtuais isolados (*venv*) para o gerenciamento de pacotes e garantia da reprodutibilidade.
- **Manipulação e Processamento de Dados**: A estruturação tabular, a limpeza e o pré-processamento de atributos basearam-se nas bibliotecas *Pandas* e *NumPy*.

Variável	Descrição do Atributo no SCR
uf	Estado da federação do tomador de crédito (categórica).
cnae_ocupacao	Ocupação principal ou setor econômico do tomador (categórica).
porte	Porte do cliente baseado em faixas de rendimento (categórica).
modalidade	Categoria macro da operação de crédito concedida (categórica).
submodalidade	Subdivisão específica do produto de crédito contratado (categórica).
numero_de_operacoes	Quantidade de contratos ativos na carteira do cliente (numérica).
a_vencer_ate_90_dias	Saldo a vencer com prazo de até 90 dias (numérica, monetária).
a_vencer_de_91_ate_360_dias	Saldo a vencer entre 91 e 360 dias (numérica, monetária).
a_vencer_de_361_ate_1080_dias	Saldo a vencer entre 361 e 1080 dias (numérica, monetária).
a_vencer_de_1081_ate_1800_dias	Saldo a vencer entre 1081 e 1800 dias (numérica, monetária).
a_vencer_de_1801_ate_5400_dias	Saldo a vencer entre 1801 e 5400 dias (numérica, monetária).
a_vencer_acima_de_5400_dias	Saldo a vencer superior a 5400 dias (numérica, monetária).
carteira_a_vencer	Somatório total dos saldos ainda não vencidos (numérica, monetária).
carteira_vencida	Valor total dos saldos com pagamento em atraso (numérica, monetária).

Tabela 3 – Atributos selecionados para a modelagem preditiva pós-concessão.

- **Aprendizado de Máquina**: A integração das etapas de balanceamento de classes (*imbalanced-learn*), padronização matemática e modelagem preditiva ocorreu através da biblioteca *Scikit-learn*. O algoritmo de *Gradient Boosting* baseou-se no pacote *XGBoost*.
- **Desenvolvimento da Aplicação (MVP)**: A interface interativa de predição foi implementada utilizando o *framework* de retaguarda (*backend*) *FastAPI*, executado pelo servidor *Uvicorn*. A exibição visual do painel gerencial baseou-se em templates *Jinja2* acoplados ao HTML e CSS.

4.2 Métodos

A metodologia adotada neste trabalho foi estruturada como um fluxo de aprendizado de máquina composto por duas etapas principais: a fase offline, responsável pela preparação dos

dados e construção dos modelos preditivos, e a fase online, destinada à aplicação e avaliação dos modelos obtidos.

Na fase offline, foram realizados os procedimentos de preparação dos dados históricos, incluindo limpeza, transformação de atributos, engenharia de características e definição dos conjuntos utilizados nos experimentos. Posteriormente, diferentes algoritmos de aprendizado de máquina foram treinados e comparados, resultando na seleção do modelo preditivo com melhor desempenho.

Na fase online, o modelo selecionado foi aplicado a novos dados para geração de predições e análise do seu desempenho. Essa etapa também serviu como base para o desenvolvimento complementar do Produto Mínimo Viável (MVP), utilizado para demonstrar visualmente a aplicação do modelo em um cenário de monitoramento de risco de crédito.

A Figura 4 apresenta o fluxo geral dos experimentos realizados, contemplando as etapas de preparação, modelagem, aplicação e avaliação dos modelos.

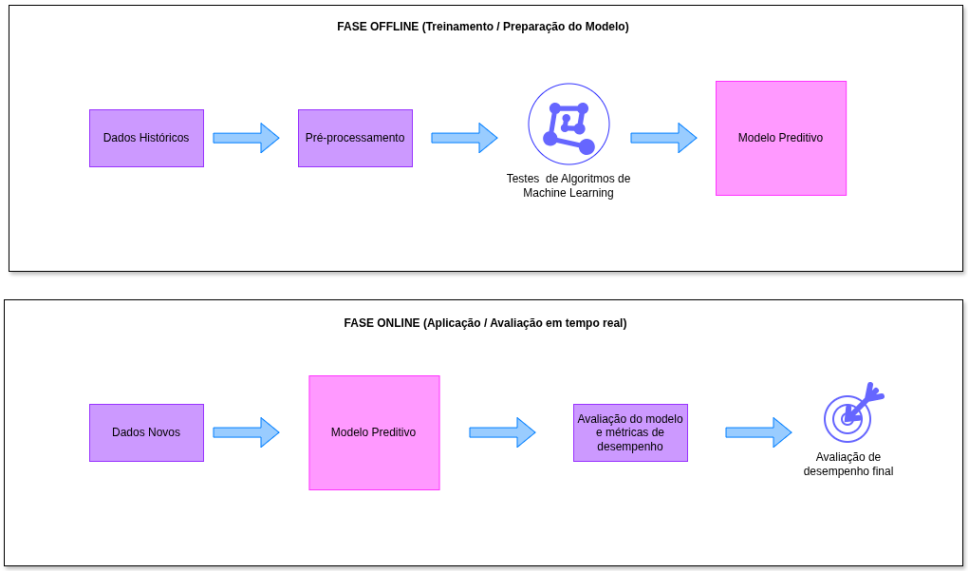

Figura 4 – Fluxo metodológico dos experimentos. Fonte: Autoria própria.

4.2.1 Pré-processamento dos Dados

O tratamento dos dados brutos foi realizado por meio de uma sequência de etapas de preparação e transformação, com o objetivo de adequar a base aos algoritmos de classificação binária. As principais etapas aplicadas foram:

- **Definição da Variável Alvo**: A variável relacionada à carteira vencida foi transformada em uma variável binária, representando a ocorrência ou não da inadimplência. Registros com valores de atraso superiores a zero foram classificados como pertencentes à classe positiva (1), enquanto operações sem valores vencidos receberam a classe negativa (0).

- **Tratamento de Inconsistências**: Os atributos financeiros originalmente armazenados como texto, contendo vírgulas como separador decimal, foram convertidos para o formato numérico adequado. Além disso, inconsistências identificadas em variáveis quantitativas, como valores negativos no número de operações, foram corrigidas utilizando a mediana dos valores válidos da população. Para preservar a informação sobre a ausência ou correção do dado original, também foi criada uma variável indicadora (*flag*).
- **Tratamento de Valores Ausentes**: Para variáveis categóricas com valores ausentes, optou-se pela manutenção dos registros e pelo preenchimento das lacunas com uma categoria específica denominada “Indisponível”, evitando a perda de informações durante o processo de preparação da base.
- **Transformação de Variáveis Categóricas**: As variáveis qualitativas foram convertidas para uma representação numérica por meio da técnica de *One-Hot Encoding*. Esse processo transformou cada categoria em uma variável binária (*dummy*), permitindo sua utilização pelos algoritmos de aprendizado de máquina empregados.

4.2.2 Engenharia de Atributos

Uma atenção metodológica específica foi dedicada ao tratamento da variável relacionada à submodalidade da operação de crédito, devido à sua elevada cardinalidade. A utilização direta dessa característica por meio da técnica de *One-Hot Encoding* poderia resultar em uma representação excessivamente esparsa dos dados, aumentando a dimensionalidade da base e potencialmente prejudicando o desempenho e a convergência dos modelos avaliados.

Para lidar com essa limitação e, ao mesmo tempo, analisar a relevância preditiva dessa informação, foram definidos três cenários experimentais distintos de engenharia de atributos. O primeiro consistiu na remoção completa da variável de submodalidade do conjunto de dados. O segundo realizou um agrupamento baseado na frequência de ocorrência das categorias, no qual submodalidades com menos de mil registros foram consolidadas em uma única categoria denominada “Outros”. Por fim, o terceiro cenário utilizou uma abordagem baseada em regras de negócio, agrupando manualmente as descrições originais em nove categorias semânticas e financeiras mais abrangentes, como “Cartão de Crédito”, “Financiamento Habitacional” e “Microcrédito”.

A comparação entre esses cenários permitiu avaliar o impacto de diferentes estratégias de representação dessa variável, buscando identificar a abordagem capaz de preservar maior quantidade de informação relevante sem comprometer a eficiência computacional dos modelos.

4.2.3 Algoritmos e Pipelines de Modelagem

A estrutura de modelagem foi organizada utilizando a classe de *Pipeline* disponibilizada pelo Scikit-learn, juntamente com a implementação de pipelines do pacote *imblearn.pipeline* para integração das etapas de balanceamento dos dados. Essa abordagem permitiu consolidar, em um único fluxo de processamento, as etapas de transformação das variáveis, aplicação das técnicas de balanceamento e treinamento dos modelos de classificação.

Com o objetivo de avaliar diferentes abordagens preditivas e identificar quais algoritmos apresentavam melhor desempenho para a tarefa de previsão de inadimplência, foram implementados e comparados seis modelos de classificação:

- Regressão Logística;
- Árvore de Decisão;
- *K-Nearest Neighbors* (KNN);
- *Random Forest*;
- Máquinas de Vetores de Suporte (SVM);
- *XGBoost*.

A seleção desses algoritmos buscou contemplar diferentes famílias de métodos de aprendizado de máquina, incluindo modelos lineares, baseados em árvores de decisão, métodos orientados à distância e técnicas de *ensemble learning*, permitindo uma comparação mais abrangente do comportamento dos modelos diante das características da base analisada.

4.2.4 Estratégia de Treinamento e Validação

Após a conclusão das etapas de pré-processamento, os dados foram divididos aleatoriamente em conjuntos de treinamento e teste, utilizando uma estratégia estratificada para preservar a proporção original das classes em ambas as partições.

Considerando o elevado desbalanceamento entre as classes da variável alvo, foi aplicada a técnica SMOTE (*Synthetic Minority Over-sampling Technique*) como estratégia de balanceamento durante o treinamento dos modelos. Para evitar a ocorrência de vazamento de dados (*data leakage*), a geração de amostras sintéticas não foi realizada antes da separação dos dados. O SMOTE foi aplicado exclusivamente dentro das partições de treinamento durante as etapas da validação cruzada, garantindo que as informações utilizadas para criar novas amostras fossem provenientes apenas dos dados disponíveis em cada etapa de treinamento.

A otimização dos hiperparâmetros dos modelos foi realizada por meio da técnica *GridSearchCV*, utilizando validação cruzada estratificada com 5 partições (*folds*). Esse procedimento

permitiu comparar diferentes configurações dos algoritmos e selecionar as melhores combinações de parâmetros sem utilizar informações provenientes do conjunto de teste, que permaneceu reservado para a avaliação final do desempenho preditivo.

4.2.5 Métricas de Avaliação

A avaliação dos modelos foi realizada com base nas métricas de desempenho apresentadas na Seção 2.3.3 do Capítulo 2. Para cada algoritmo avaliado, foram consideradas as métricas de Acurácia, F1-Score e ROC AUC, permitindo uma análise complementar do desempenho obtido nos diferentes cenários experimentais.

Durante o processo de otimização dos hiperparâmetros por meio do *GridSearchCV*, a métrica ROC AUC (*Receiver Operating Characteristic - Area Under the Curve*) foi adotada como função de avaliação (*scoring*), por apresentar uma medida consistente da capacidade de discriminação entre as classes em problemas de classificação desbalanceada, independentemente de um limiar específico de decisão.

Na comparação dos resultados finais, entretanto, foi atribuído maior peso ao F1-Score, por representar um melhor equilíbrio entre precisão e sensibilidade na identificação de operações inadimplentes. A métrica ROC AUC foi utilizada de forma complementar para avaliar o poder discriminatório dos classificadores, enquanto a Acurácia foi considerada como uma medida auxiliar do desempenho global dos modelos.

4.2.6 Desenvolvimento do Produto Mínimo Viável (MVP)

Como etapa complementar da pesquisa, foi desenvolvido um Produto Mínimo Viável (MVP) com o propósito de demonstrar uma possível aplicação prática do modelo preditivo obtido. Destaca-se que o desenvolvimento da aplicação não constitui o foco principal deste trabalho, cujo objetivo central está relacionado à construção, avaliação e comparação dos modelos de aprendizado de máquina.

O MVP foi desenvolvido como uma ferramenta demonstrativa capaz de utilizar o modelo treinado, previamente armazenado no formato *.joblib*, para realizar novas predições a partir de dados fornecidos pelo usuário. A aplicação disponibiliza uma interface para preenchimento das informações de entrada e apresenta como resultado a probabilidade estimada de inadimplência calculada pelo modelo selecionado.

5 EXPERIMENTOS

O delineamento experimental deste trabalho tem como objetivo avaliar o desempenho preditivo de diferentes algoritmos de aprendizado de máquina na tarefa de classificação de inadimplência, utilizando como base histórica os dados pós-concessão do Sistema de Informações de Crédito (SCR). Adicionalmente, busca-se investigar o impacto de diferentes estratégias de engenharia de atributos sobre a variável categórica de alta cardinalidade `submodalidade`.

Para tanto, os experimentos foram estruturados em etapas sistemáticas: preparação estrutural e limpeza dos dados brutos; formulação de cenários preditivos com base na representação de atributos; definição do protocolo de validação cruzada livre de vazamento de dados; e, por fim, a busca pelo espaço ótimo de hiperparâmetros por meio de uma estratégia de busca em grade (*Grid Search*).

5.1 Preparação e Transformação dos Dados

A extração inicial do SCR totalizava 310.504 registros. Visando adequar o escopo do projeto à previsão de inadimplência para o segmento de varejo individual, aplicou-se um filtro no atributo `cliente` para isolar unicamente registros associados a Pessoas Físicas (PF). Esse procedimento resultou em uma base consolidada de 182.631 amostras, representando uma contração de 41,18% do volume bruto original.

Após a seleção dos atributos primários, a base foi submetida a um conjunto de transformações e tratamentos de inconsistências:

1. **Definição da Variável Alvo**: O atributo de saída (alvo) foi derivado da coluna `carteira_vencida`. Os valores originais em formato de texto foram normalizados (substituição de vírgula por ponto) e convertidos para ponto flutuante. A variável alvo de inadimplência foi definida binariamente como:

$$y = \begin{cases} 1, & \text{se } \texttt{carteira\_vencida} > 0 \\ 0, & \text{se } \texttt{carteira\_vencida} \leq 0 \end{cases}$$

2. **Tratamento de Dados Ausentes**: Valores nulos na variável `porte` foram preenchidos com o rótulo categórico "Indisponível". Semelhantemente, registros nulos em `submodalidade` receberam o rótulo "desconhecido". Essa estratégia visa mitigar o viés gerado pela exclusão arbitrária de observações incompletas.

3. **Saneamento do Histórico de Operações**: O atributo `numero_de_operacoes` continha registros rotulados textualmente como ≤ 15, os quais foram mapeados para o valor numérico inteiro 15. Registros com contadores negativos (< 0) foram conside-

rados inconsistências do sistema e redefinidos como valores ausentes (*missing*). Para preservar a informação da inconsistência, gerou-se uma variável indicadora auxiliar (*flag*) binária (`operacoes_missing` $\in \{0, 1\}$). A imputação do valor numérico faltante na variável principal foi efetuada utilizando-se a mediana da distribuição amostral.

4. **Padronização Financeira e Evitação de Divisão por Zero**: Os valores monetários da carteira a vencer foram convertidos de formato textual para numérico de ponto flutuante. Com o objetivo de evitar erros computacionais (como indeterminação em divisões matemáticas em engenharia de atributos subsequente), todos os registros que continham o valor zero na variável consolidada `carteira_a_vencer` foram ajustados para o valor unitário 1.

5. **Codificação Categórica**: Para permitir o processamento dos dados por algoritmos não baseados unicamente em árvores, as variáveis categóricas remanescentes (`uf`, `cnae_ocupacao`, `porte` e `modalidade`) foram mapeadas para variáveis binárias (*dummies*) por meio da codificação *One-Hot Encoding*.

Para garantir que as etapas de balanceamento estatístico e a padronização das escalas numéricas (*StandardScaler*) não comprometessem a validação, esses métodos foram aplicados dinamicamente via pipeline computacional apenas nos conjuntos de treino internos de cada dobra de validação cruzada, prevenindo o vazamento de informações.

5.2 Cenários Experimentais da Variável Submodalidade

O atributo categórico `submodalidade` identifica a natureza específica da linha de crédito. Contudo, a variável apresenta uma distribuição característica do tipo *long tail* (cauda longa), em que poucas submodalidades agrupam a grande maioria dos contratos ativos, enquanto dezenas de outras registram frequências insignificantes de uso, como ilustrado na Figura 5.

A aplicação do *One-Hot Encoding* sobre a totalidade de categorias brutas elevaria substancialmente a esparsidade e a cardinalidade da matriz de covariáveis. A fim de avaliar o impacto preditivo dessa variável e determinar a melhor estratégia de agregação teórica, foram projetados e executados três cenários metodológicos:

- **Cenário 1 – Sem Submodalidade (Baseline)**: Consiste na exclusão completa da variável `submodalidade` do espaço de características, mantendo apenas a variável macro `modalidade` e os atributos demográficos e financeiros originais. Estabelece a linha de base empírica para avaliar se o detalhamento da linha de crédito é, de fato, relevante na predição de risco de inadimplência. A codificação gera 62 atributos preditivos.

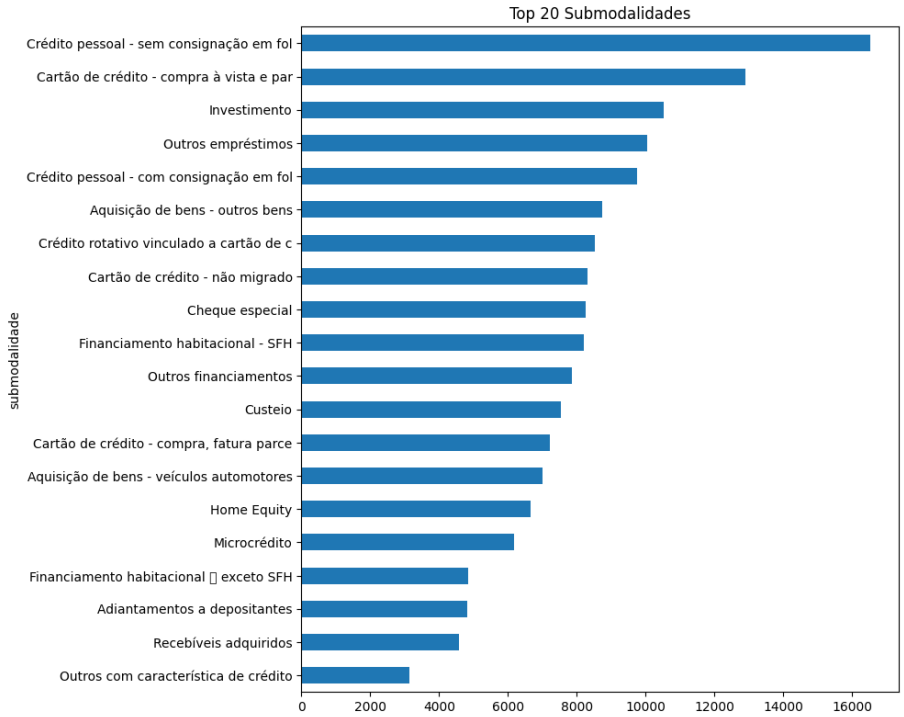

Figura 5 – Distribuição empírica das vinte submodalidades mais frequentes na base PF.

- **Cenário 2 – Submodalidade Agrupada por Frequência**: Concentra em uma única categoria residual rotulada como “Outros” todas as submodalidades com contagem histórica inferior a um limiar estatístico de 1.000 ocorrências na base de dados. Essa abordagem de agregação de categorias raras preserva a identidade de submodalidades estatisticamente relevantes, reduzindo o ruído analítico de categorias esparsas. A codificação resulta em 92 atributos preditivos.

- **Cenário 3 – Submodalidade Agrupada por Regras de Negócio (*Engineered*)**: Realiza o agrupamento semântico manual das categorias a partir do conhecimento de domínio de mercado de crédito. Baseando-se em mapeamentos estruturados por expressões regulares (conforme detalhado na Tabela 4), as submodalidades foram fundidas em dez classes conceituais homogêneas, mantendo a interpretabilidade gerencial dos produtos financeiros. A codificação resulta em 71 atributos preditivos.

A Tabela 5 consolida as características estruturais e as dimensões dos três conjuntos de dados analisados nos experimentos.

5.3 Configuração do Pipeline e Treinamento

O fluxo metodológico adotado para a execução dos experimentos é apresentado na Figura 6.

Categoria Consolidada	Terminologias Mapeadas da Base SCR
Cartão de Crédito	Termos correlacionados a "cartão de crédito", "fatura de cartão"ou "compra parcelada emitente".
Crédito Pessoal	Linhas de "crédito pessoal", "consignado em folha"e "sem consignação".
Financiamento Habitacional	Contratos vinculados a "habitacional", "SFH", "imobiliário"ou "home equity".
Financiamento de Veículos	Contratos associados a "veículos automotores"e operações de "arrendamento".
Crédito Rural	Operações agrícolas de "custeio", "comercialização"ou "agroindustriais".
Capital de Giro	Contratos de "capital de giro"ou "conta garantida".
Microcrédito	Contratos de "microcrédito".
Cheque Especial	Contratos de "cheque especial"e limites emergenciais.
Aquisição de Bens	Contratos finalistas de "aquisição de bens"em geral.
Outros	Operações não enquadradas semanticamente nos grupos anteriores.

Tabela 4 – Critério de mapeamento semântico para o Cenário 3 (*submodalidade_engineered*).

Cenário Experimental	Volume de Amostras	Preditores	Colunas Totais
1 - Sem Submodalidade	182.631	62	63
2 - Agr. por Frequência	182.631	92	93
3 - Regra de Negócio	182.631	71	72

Tabela 5 – Dimensões comparativas dos conjuntos de dados gerados pelos cenários.

Em todos os cenários experimentais, o conjunto de dados foi dividido de forma estratificada em 70% para treinamento e 30% para teste, preservando a proporção das classes em ambas as partições.

A seleção dos melhores hiperparâmetros foi realizada por meio de um pipeline de aprendizado de máquina implementado com a classe `Pipeline` da biblioteca `imbalanced-learn`, integrado ao algoritmo de busca em grade (`GridSearchCV`). O pipeline executa, de forma sequencial, as seguintes etapas:

1. Padronização dos atributos numéricos utilizando o `StandardScaler`;
2. Aplicação opcional da técnica de sobreamostragem sintética SMOTE;
3. Treinamento do algoritmo de classificação.

Para possibilitar a comparação entre as abordagens com e sem balanceamento, a aplicação do SMOTE foi incluída como um hiperparâmetro do pipeline, assumindo os valores lógicos `True` e `False`. Quando desabilitado, foi utilizada a transformação *passthrough*, permitindo que o pipeline prosseguisse sem realizar a etapa de sobreamostragem.

Como o SMOTE foi incorporado ao pipeline, a geração das amostras sintéticas ocorreu exclusivamente sobre os dados de treinamento de cada uma das cinco dobras da validação cruzada estratificada (*5-Fold Stratified Cross-Validation*). Dessa forma, os conjuntos de validação e o conjunto de teste permaneceram inalterados, evitando vazamento de dados (*data leakage*) e

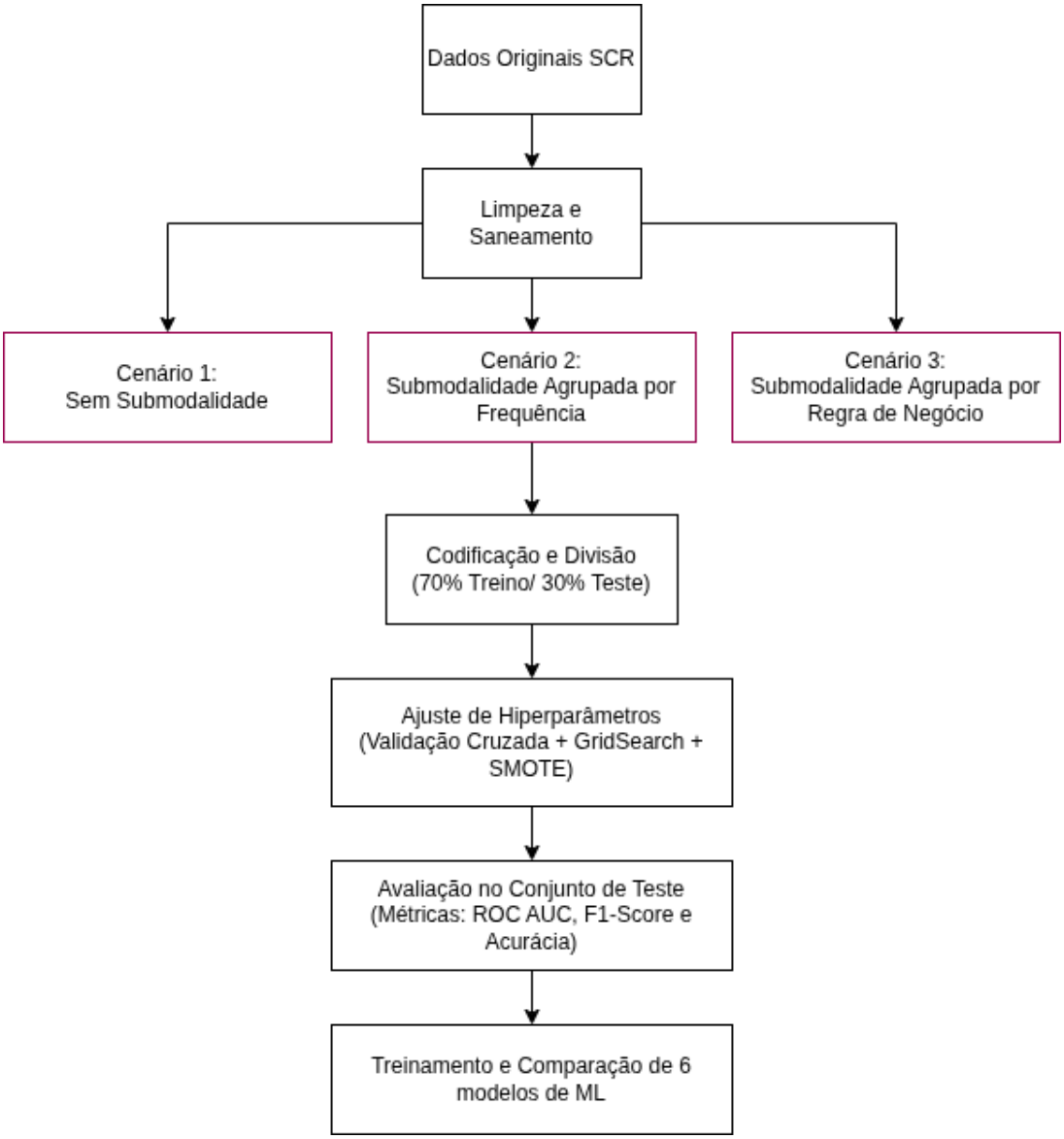

Figura 6 – Fluxo metodológico das etapas dos experimentos.

garantindo uma avaliação consistente do desempenho dos modelos. A melhor combinação de hiperparâmetros foi selecionada com base na métrica ROC AUC durante o processo de busca em grade.

5.4 Hiperparâmetros e Espaço de Busca dos Algoritmos

As experimentações contemplaram a avaliação de seis algoritmos de aprendizado supervisionado, buscando abranger diferentes hipóteses de fronteiras de decisão (lineares, baseadas em partições de espaço e comitês aditivos e bootstrap).

Os parâmetros de controle avaliados e seus respectivos intervalos de busca (*grid*) foram selecionados com base no suporte computacional fornecido pela biblioteca *Scikit-learn* (Scikit-

learn Developers, 2025) e pelo algoritmo de otimização *XGBoost*. As grades experimentais definidas para o *Grid Search* em cada modelo são detalhadas na Tabela 6.

Algoritmo Classificador	**Espaço de Busca de Hiperparâmetros (*Grid Search*)**
Regressão Logística	Regularização L2 ($C \in \{0.01, 0.1, 1, 10\}$); Otimizador de coeficientes (*solver* = "lbfgs").
Árvore de Decisão	Profundidade máxima (*max_depth* $\in \{10, 20, 30\}$); Amostras mínimas para divisão (*min_samples_split* $\in \{2, 5, 10\}$); Amostras mínimas por folha (*min_samples_leaf* $\in \{1, 2, 4\}$); Critério de impureza (*criterion* $\in \{$"gini", "entropy"$\}$).
K-Nearest Neighbors (KNN)	Número de vizinhos próximos (*n_neighbors* $\in \{3, 5, 7, 9\}$); Função de peso (*weights* $\in \{$"uniform", "distance"$\}$); Métrica de distância (*metric* $\in \{$"euclidean", "manhattan"$\}$).
Random Forest	Número de estimadores (*n_estimators* $\in \{100, 200\}$); Amostras mínimas para divisão (*min_samples_split* $\in \{2, 5\}$); Amostras mínimas por folha (*min_samples_leaf* $\in \{1, 2\}$).
SVM (LinearSVC)	Regularização L2 ($C \in \{0.1, 1, 10\}$); Perda por hinge quadrática (*loss* = "squared_hinge").
XGBoost	Número de estimadores (*n_estimators* $\in \{200, 400\}$); Taxa de aprendizado (*learning_rate* $\in \{0.05, 0.1\}$); Profundidade máxima (*max_depth* $\in \{3, 5\}$); Amostragem de linhas (*subsample* $\in \{0.8, 1.0\}$); Amostragem de atributos (*colsample_bytree* $\in \{0.8, 1.0\}$).

Tabela 6 – Grade de hiperparâmetros utilizada na busca exaustiva dos modelos.

Após a identificação da melhor parametrização combinatória para cada algoritmo baseada na média obtida na validação cruzada do conjunto de treino, os pipelines candidatos foram retreinados utilizando o conjunto de treinamento completo e avaliados estatisticamente contra o conjunto de teste de dados não vistos.

6 RESULTADOS E DISCUSSÃO

Este capítulo apresenta a análise comparativa dos resultados obtidos nos experimentos computacionais. O desempenho dos estimadores foi mensurado no conjunto de teste independente por meio das métricas F1-Score, ROC AUC e Acurácia. A ROC AUC foi empregada durante a otimização dos hiperparâmetros, enquanto o F1-Score foi adotado como principal critério para a seleção e comparação dos modelos finais, sendo complementado pelas métricas de ROC AUC e Acurácia.

6.1 Resultados do Cenário 1: Sem Submodalidade (Baseline)

Abaixo são expostos os desempenhos dos estimadores no cenário de referência, no qual a variável `submodalidade` foi excluída da modelagem.

6.1.1 Validação Cruzada

A Tabela 7 apresenta a média de ROC AUC obtida na etapa de validação cruzada, ordenada de forma decrescente.

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

Tabela 7 – Resultados da validação cruzada no Cenário 1 (Sem Submodalidade).

Os modelos baseados em comitês de decisão (Random Forest e XGBoost) lideraram o treinamento com valores médios de ROC AUC superiores a 0,92. Os estimadores lineares (Regressão Logística e SVM) e o classificador baseado em distância (KNN) apresentaram desempenho inferior, registrando médias de validação abaixo de 0,78.

6.1.2 Conjunto de Teste

A Tabela 8 consolida os desempenhos finais observados após a projeção dos pipelines contra o conjunto de teste independente, ordenados de forma decrescente pelo F1-Score.

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

Tabela 8 – Desempenho no conjunto de teste independente (Sem Submodalidade).

A Random Forest sem aplicação de SMOTE obteve o maior F1-Score (0,8726) e ROC AUC (0,9288) deste cenário, seguida pelo XGBoost sem SMOTE (0,8628). Os classificadores lineares e de distância mantiveram F1-Scores em patamares baixos, evidenciando menor capacidade de separação sob a ausência de detalhamento da linha de crédito.

6.2 Resultados do Cenário 2: Submodalidade Agrupada por Frequência

Esta seção detalha os resultados referentes ao cenário em que a variável de submodalidade foi mantida sob agrupamento por frequência estatística.

6.2.1 Validação Cruzada

As médias de ROC AUC obtidas na validação cruzada para os estimadores neste cenário estão apresentadas na Tabela 9, ordenadas de forma decrescente.

A introdução da variável agrupada por frequência gerou uma elevação generalizada nas médias de validação em relação ao baseline, permitindo que a Regressão Logística e o SVM superassem a marca de 0,82 de ROC AUC. A Random Forest sem SMOTE registrou a maior ROC AUC de treinamento (0,9443).

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

Tabela 9 – Resultados da validação cruzada no Cenário 2 (Submodalidade Agrupada).

6.2.2 Conjunto de Teste

As métricas finais obtidas no conjunto de teste independente estão consolidadas na Tabela 10, ordenadas de forma decrescente pelo F1-Score.

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

Tabela 10 – Desempenho no conjunto de teste independente (Submodalidade Agrupada).

A Random Forest sem SMOTE alcançou o F1-Score mais elevado (0,8879). Nota-se um avanço expressivo de aproximadamente 4 pontos percentuais no F1-Score dos estimadores lineares (Regressão Logística e SVM) em relação ao baseline, evidenciando que o detalhamento da submodalidade agrupada adicionou capacidade explicativa relevante ao espaço de características.

6.3 Resultados do Cenário 3: Submodalidade Engineered

Os resultados referentes ao cenário de agrupamento semântico baseado em regras de negócio são apresentados a seguir.

6.3.1 Validação Cruzada

As médias de ROC AUC obtidas na validação cruzada para os estimadores neste cenário estão apresentadas na Tabela 11, ordenadas de forma decrescente.

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

Tabela 11 – Resultados da validação cruzada no Cenário 3 (Submodalidade Engineered).

Os modelos de comitê de decisão baseados em árvores mantiveram-se na liderança (Random Forest com 0,9400 e XGBoost com 0,9376), porém registrando médias de validação ligeiramente inferiores em relação ao Cenário 2.

6.3.2 Conjunto de Teste

A Tabela 12 apresenta as métricas computadas no conjunto de teste independente, ordenadas de forma decrescente pelo F1-Score.

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

Tabela 12 – Desempenho no conjunto de teste independente (Submodalidade Engineered) ordenado por F1-Score.

A Random Forest sem SMOTE obteve o maior F1-Score do cenário (0,8824). Em comparação ao agrupamento por frequência (Cenário 2), a abordagem baseada em regras de negócio

resultou em perdas preditivas marginais para a maioria dos estimadores de melhor desempenho.

6.4 Comparação Geral dos Cenários

A Tabela 13 resume o desempenho dos melhores pipelines identificados em cada cenário experimental, ordenados pelo F1-Score.

Cenário Experimental	Modelo	F1-Score	ROC AUC	Acurácia
Cenário 2 (Submodalidade agrupada)	Random Forest	**0,8879**	**0,9444**	**0,8694**
Cenário 3 (Submodalidade engineered)	Random Forest	0,8824	0,9399	0,8631
Cenário 1 (Sem submodalidade)	Random Forest	0,8726	0,9288	0,8519

Tabela 13 – Desempenho comparativo dos melhores modelos obtidos em cada cenário (Ordenado por F1-Score).

O modelo Random Forest consolidou-se como o melhor estimador nos três cenários de análise. A comparação atesta que a inclusão estruturada da submodalidade (Cenário 2) proporcionou uma elevação de 1,53 pontos percentuais no F1-Score em relação ao baseline desprovido da variável. A Figura 7 ilustra a distribuição dos F1-Scores obtidos pelos estimadores.

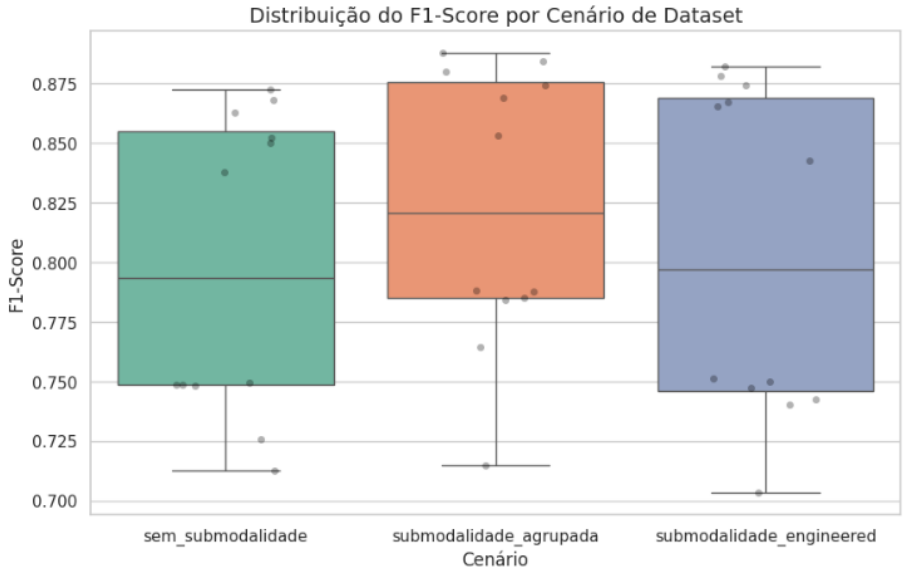

Figura 7 – Distribuição dos valores de F1-Score obtidos em cada cenário experimental.

A distribuição visual confirma que o Cenário 2 (Agrupado por Frequência) promoveu um deslocamento geral positivo nas métricas de desempenho. A Tabela 14 resume os melhores valores de F1-Score por algoritmo em cada cenário.

A abordagem por frequência (Cenário 2) obteve o maior F1-Score para todas as arquiteturas avaliadas. O classificador KNN registrou a menor performance em todas as rodadas, o que está associado à sensibilidade do cálculo das distâncias sob a presença de alta dimensionalidade gerada após a binarização das variáveis qualitativas.

Modelo Classificador	Cenário 2 (Agrupada)	Cenário 3 (Engineered)	Cenário 1 (Baseline)
Random Forest	**0,8879**	0,8824	0,8726
XGBoost	**0,8800**	0,8745	0,8628
Decision Tree	**0,8690**	0,8656	0,8522
Logistic Regression	**0,7884**	0,7501	0,7498
SVM (LinearSVC)	**0,7852**	0,7426	0,7486
KNN	**0,7645**	0,7514	0,7259

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

6.5 Impacto da Aplicação do SMOTE

Os experimentos revelaram que a aplicação do SMOTE não resultou em benefícios preditivos práticos para os modelos avaliados. A Tabela 15 compara o comportamento do estimador Random Forest sob as duas configurações de balanceamento nos três cenários de análise.

Cenário Experimental	SMOTE	F1-Score	ROC AUC	Acurácia
Cenário 2 (Submodalidade agrupada)	Não	**0,8879**	**0,9444**	**0,8694**
Cenário 2 (Submodalidade agrupada)	Sim	0,8846	0,9432	0,8672
Cenário 3 (Submodalidade engineered)	Não	**0,8824**	**0,9399**	**0,8631**
Cenário 3 (Submodalidade engineered)	Sim	0,8784	0,9385	0,8604
Cenário 1 (Sem submodalidade)	Não	**0,8726**	**0,9288**	**0,8519**
Cenário 1 (Sem submodalidade)	Sim	0,8684	0,9274	0,8495

Tabela 15 – Comparação de impacto da técnica SMOTE no estimador Random Forest.

A desativação do SMOTE promoveu melhorias marginais consistentes em todos os cenários. A hipótese para esse comportamento fundamenta-se em dois aspectos principais:

- **Representatividade Absoluta da Classe Minoritária**: A base histórica de treinamento continha um volume absoluto de observações reais de inadimplentes estatisticamente representativo (aproximadamente 25% da base total de 182.631 registros). Algoritmos de ensemble baseados em árvores possuem mecanismos naturais que toleram desbalanceamentos sob grandes volumes absolutos de amostras reais.

- **Geração de Ruído em Espaços Binarizados**: A aplicação do SMOTE em um espaço contendo características binarizadas via One-Hot Encoding provoca a projeção de instâncias sintéticas logicamente inconsistentes ou posicionadas na zona de transição entre as classes. Esse fenômeno induz a sobreposição de classes (*class overlap*), gerando perturbações no limite de separação estatística e prejudicando levemente a capacidade de generalização do estimador.

6.6 Ranking Geral dos Experimentos

A Tabela 16 apresenta a consolidação dos dez experimentos de melhor desempenho obtidos dentre as 36 configurações testadas, ordenados pelo F1-Score no conjunto de teste independente.

Posição	Algoritmo Classificador	Cenário Experimental	SMOTE	F1-Score
1	Random Forest	Cenário 2 (Agrupada)	Não	**0,8879**
2	Random Forest	Cenário 2 (Agrupada)	Sim	0,8846
3	Random Forest	Cenário 3 (Engineered)	Não	0,8824
4	XGBoost	Cenário 2 (Agrupada)	Não	0,8800
5	Random Forest	Cenário 3 (Engineered)	Sim	0,8784
6	XGBoost	Cenário 3 (Engineered)	Não	0,8745
7	XGBoost	Cenário 2 (Agrupada)	Sim	0,8742
8	Random Forest	Cenário 1 (Sem submodalidade)	Não	0,8726
9	Decision Tree	Cenário 2 (Agrupada)	Não	0,8690
10	Random Forest	Cenário 1 (Sem submodalidade)	Sim	0,8684

Tabela 16 – Ranking dos dez melhores experimentos consolidados pelo F1-Score.

A hegemonia dos comitês baseados em árvores é atestada pelas posições do ranking, com as variantes de Random Forest e XGBoost ocupando as sete primeiras posições. A Figura 8 apresenta graficamente a comparação entre esses melhores modelos.

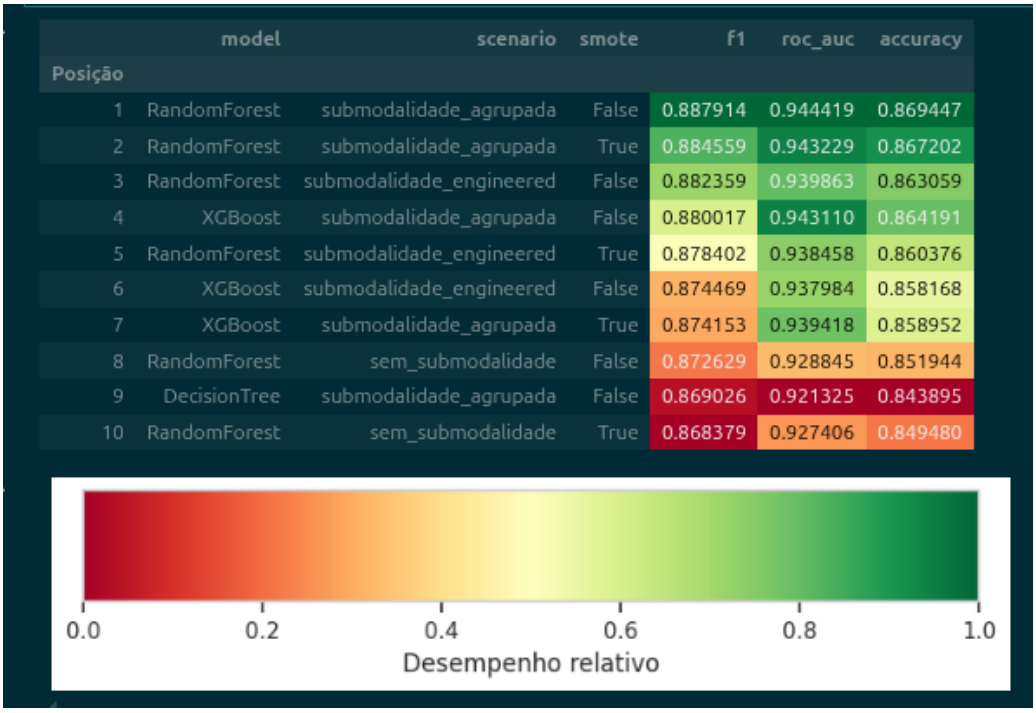

Figura 8 – Visualização comparativa dos dez melhores resultados pelo F1-Score.

6.7 Melhor Modelo Encontrado e Demonstração do Protótipo

O melhor desempenho global do estudo foi obtido pelo algoritmo Random Forest no Cenário 2 (Submodalidade Agrupada por Frequência) sem a aplicação da técnica SMOTE, registrando F1-Score de 0,8879, ROC AUC de 0,9444 e Acurácia de 0,8694 no conjunto de teste independente. O modelo apresentou o melhor equilíbrio entre capacidade de identificação das operações inadimplentes e desempenho geral dentre todas as configurações experimentais avaliadas.

A Tabela 17 resume os principais indicadores de desempenho obtidos pelo modelo selecionado.

Métrica de Avaliação	Desempenho no Teste
F1-Score (Métrica Principal)	**0,8879**
ROC AUC (Auxiliar)	0,9444
Acurácia (Auxiliar)	0,8694

Tabela 17 – Indicadores de desempenho do modelo Random Forest eleito.

Os hiperparâmetros selecionados pela busca em grade aplicada ao pipeline final estão descritos na Tabela 18. Observa-se que a configuração vencedora não utilizou balanceamento artificial de classes, indicando que, para o conjunto de dados analisado, a aplicação do SMOTE não proporcionou ganhos adicionais de desempenho.

Hiperparâmetro	Valor Otimizado Selecionado
Número de árvores na floresta (*n_estimators*)	200
Amostras mínimas para divisão (*min_samples_split*)	5
Amostras mínimas por folha (*min_samples_leaf*)	1
Balanceamento artificial (SMOTE)	Desativado (*passthrough*)

Tabela 18 – Parametrização final do modelo de melhor desempenho.

A Figura 9 apresenta a comparação das curvas ROC dos classificadores avaliados no Cenário 2 (Submodalidade Agrupada por Frequência). O comportamento observado reforça o melhor desempenho das abordagens baseadas em métodos de conjunto, especialmente Random Forest e XGBoost, quando comparadas aos modelos lineares ou baseados em distância geométrica.

Após a seleção do modelo final, o classificador treinado foi armazenado por meio da serialização dos arquivos `best_model.joblib` e `model_columns.joblib`, contendo, respectivamente, o modelo ajustado e a estrutura das variáveis utilizadas durante o treinamento. Esses arquivos permitiram disponibilizar o modelo para inferência em uma aplicação web demonstrativa desenvolvida com FastAPI.

O protótipo desenvolvido teve como objetivo demonstrar uma possibilidade de utilização prática do modelo preditivo em um contexto de apoio à decisão pós-concessão de crédito. A aplicação permite a inserção manual das variáveis preditivas utilizadas pelo classificador, como

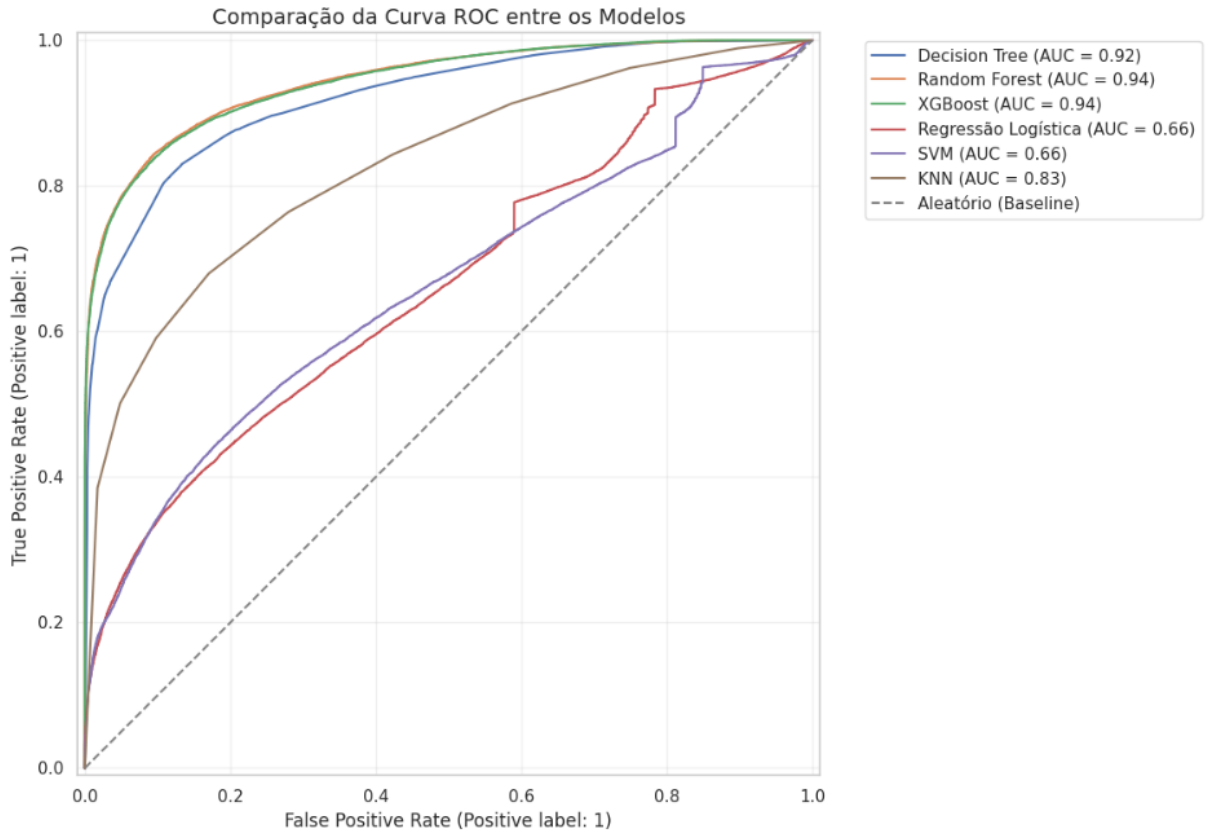

Figura 9 – Comparação das curvas ROC dos modelos avaliados no cenário de submodalidade agrupada.

UF, ocupação, porte de renda, modalidade de crédito e quantidade de operações, retornando a probabilidade estimada de inadimplência para a operação analisada.

A Figura 10 apresenta a tela de simulação individual, na qual são informadas características cadastrais e operacionais da operação de crédito. Após o envio dos dados, o modelo realiza a inferência e apresenta a probabilidade estimada de inadimplência, permitindo uma interpretação direta do nível de risco identificado.

Além da análise individual, foi desenvolvida uma visão consolidada da carteira de crédito para demonstrar o potencial de utilização do modelo em atividades de monitoramento pós-concessão. A Figura 11 apresenta o painel gerencial contendo a distribuição dos associados classificados por nível de risco (Baixo, Médio e Alto), permitindo uma visualização agregada da carteira analisada.

A demonstração apresentada possui caráter experimental e tem como finalidade evidenciar a integração entre o modelo preditivo desenvolvido e uma interface de apoio à interpretação dos resultados. Dessa forma, o protótipo não representa uma solução operacional completa, mas uma demonstração da aplicabilidade prática da abordagem proposta.

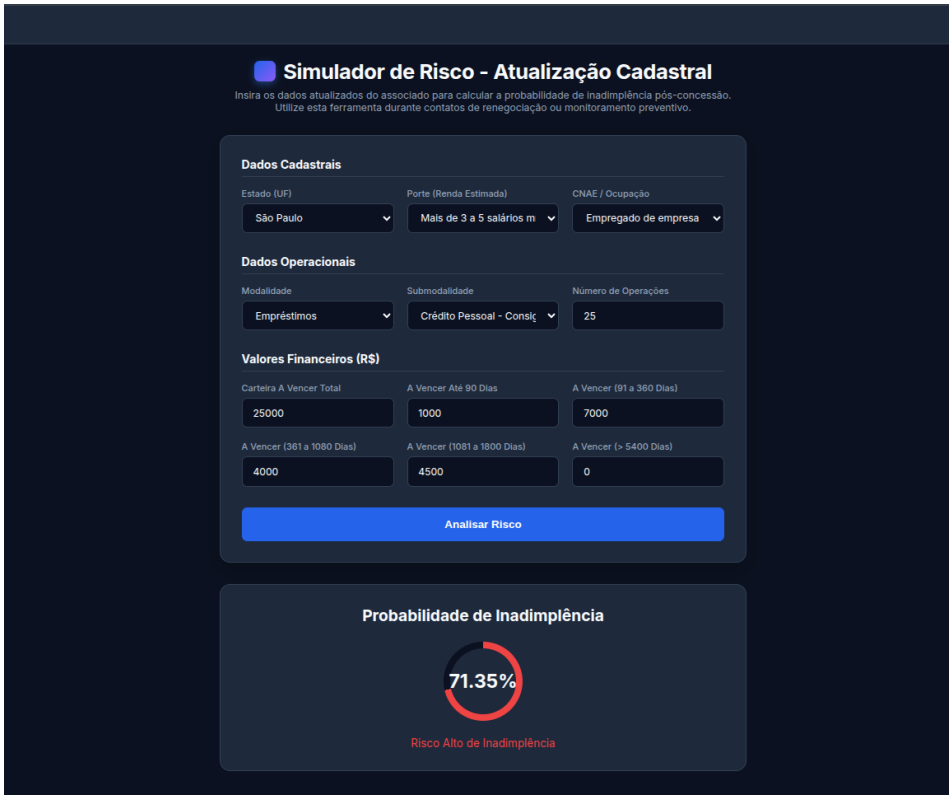

Figura 10 – Interface de simulação individual para análise de risco de inadimplência.

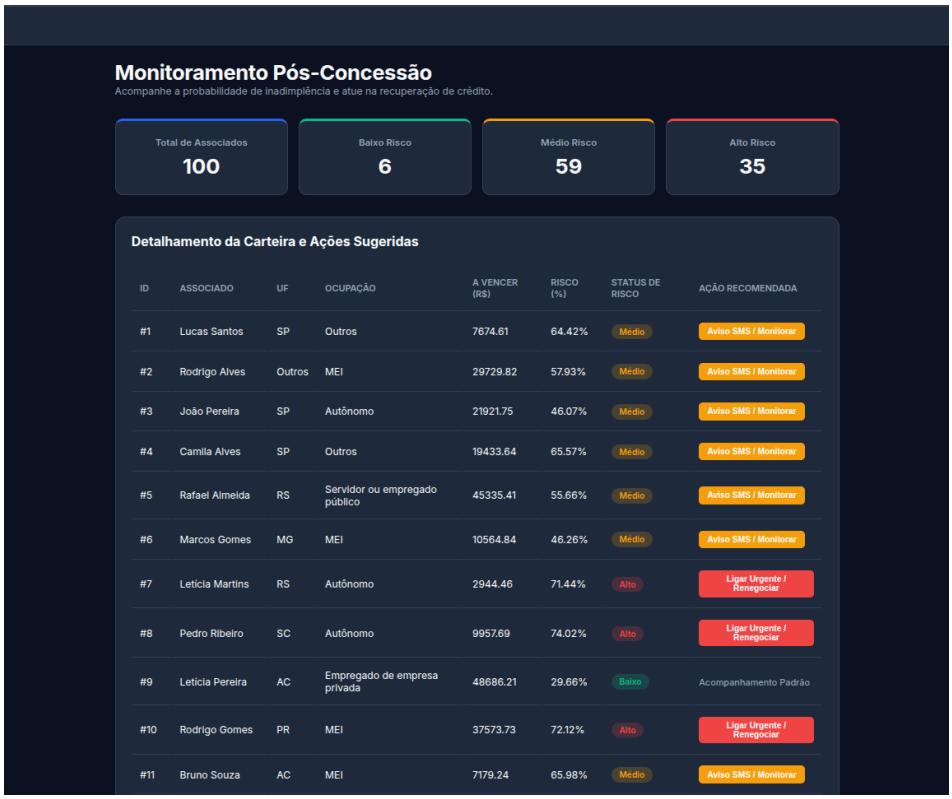

Figura 11 – Dashboard de monitoramento da carteira de crédito por nível de risco.

6.8 Discussão Geral dos Resultados

A análise integrada dos experimentos realizados indica que a estratégia de representação dos atributos categóricos apresentou maior influência no desempenho dos classificadores avaliados do que a aplicação das técnicas de balanceamento de classes consideradas. Os resultados obtidos demonstram que a forma de transformação das variáveis categóricas pode alterar significativamente a capacidade dos modelos em identificar padrões associados à inadimplência.

O Cenário 2, baseado no agrupamento por frequência, apresentou desempenho superior ao Cenário 3, que utilizou um agrupamento orientado por regras de negócio. Uma possível explicação para esse comportamento está relacionada à preservação das características estatísticas das categorias originais. Embora o agrupamento semântico permita uma interpretação mais próxima do conhecimento de negócio, essa transformação pode reduzir a granularidade dos dados ao reunir categorias que, apesar de apresentarem similaridade conceitual, podem possuir comportamentos estatísticos distintos em relação ao risco de inadimplência. Por outro lado, a abordagem baseada em frequência manteve maior separação entre categorias com representatividade estatística suficiente, permitindo que os classificadores explorassem padrões específicos presentes na distribuição original dos dados. Esses resultados sugerem que, para a base analisada, a preservação da variabilidade estatística da variável *submodalidade* foi mais relevante para o desempenho preditivo do que um agrupamento fundamentado exclusivamente em critérios semânticos.

Em relação às técnicas de balanceamento, observou-se que a aplicação do SMOTE não proporcionou ganhos consistentes de desempenho e, em alguns cenários, resultou em pequenas reduções nas métricas obtidas. Uma possível explicação é que algoritmos baseados em árvores, como o *Random Forest* e o *XGBoost*, já apresentam boa capacidade de lidar com moderados desbalanceamentos de classe. Nesses casos, a geração de amostras sintéticas pode ter introduzido maior variabilidade nos dados de treinamento sem acrescentar informações relevantes para a identificação dos padrões de inadimplência, refletindo em um desempenho ligeiramente inferior quando os modelos foram avaliados sobre dados reais não modificados.

Outro aspecto relevante refere-se à ausência de variáveis macroeconômicas no conjunto de atributos utilizados. A base empregada neste estudo foi obtida a partir de dados públicos do Sistema de Informações de Crédito (SCR), representando um recorte estático do período analisado. Dessa forma, indicadores econômicos dinâmicos, como taxa básica de juros, inflação e mercado de trabalho, não foram incorporados ao modelo, pois a estrutura disponível não apresentava uma dimensão temporal adequada para capturar seus efeitos sobre o comportamento de crédito. A inclusão dessas variáveis representa uma possibilidade de extensão da pesquisa em cenários que utilizem séries históricas mais longas.

Além disso, modelos preditivos aplicados ao risco de crédito estão sujeitos à degradação de desempenho ao longo do tempo em decorrência de mudanças no comportamento dos con-

sumidores, alterações econômicas e modificações nas características das carteiras analisadas. Embora este trabalho não tenha realizado uma avaliação temporal de degradação do modelo (*model drift*), aplicações em ambientes produtivos devem considerar mecanismos de monitoramento contínuo e estratégias periódicas de atualização. Nesse contexto, a disponibilidade recorrente de novos dados do SCR possibilita a investigação de estratégias de retreinamento periódico, cuja frequência deve ser definida conforme a estabilidade observada do modelo e as necessidades operacionais da instituição.

Quanto à aplicabilidade prática, destaca-se que modelos desenvolvidos exclusivamente com dados públicos apresentam limitações quando comparados aos ambientes reais de instituições financeiras, que normalmente possuem informações internas adicionais, como histórico de relacionamento, comportamento de pagamento, utilização de produtos e variáveis transacionais. Dessa forma, uma possibilidade de evolução consiste na utilização do modelo desenvolvido como uma camada complementar de informação de risco, em que a probabilidade de inadimplência estimada a partir dos dados públicos possa ser incorporada a modelos internos já existentes, combinando uma visão agregada do mercado de crédito com características específicas dos clientes da instituição.

De forma geral, os resultados obtidos indicam que a escolha da representação dos atributos, da estratégia de balanceamento e do algoritmo de classificação possui impacto relevante no desempenho preditivo para o problema analisado. O modelo *Random Forest* aplicado ao Cenário 2 apresentou o melhor resultado entre as configurações avaliadas, alcançando F1-Score de 0,8879 e ROC AUC de 0,9444. Entretanto, as conclusões obtidas permanecem restritas ao conjunto de dados utilizado, aos cenários experimentais definidos e às configurações de modelagem avaliadas neste trabalho.

7 CONSIDERAÇÕES FINAIS

Este trabalho apresentou uma abordagem baseada em aprendizado de máquina para a previsão de inadimplência no período pós-concessão de crédito, utilizando dados públicos do Sistema de Informações de Crédito (SCR) do Banco Central do Brasil. Os resultados demonstraram a viabilidade da utilização de técnicas de classificação para identificar padrões associados ao risco de inadimplência e apoiar processos de monitoramento de carteiras de crédito.

Os resultados obtidos permitiram responder ao problema de pesquisa proposto, evidenciando que a representação das variáveis categóricas exerce influência significativa no desempenho dos modelos preditivos. O melhor desempenho foi obtido pelo algoritmo *Random Forest* no Cenário 2 (Submodalidade Agrupada por Frequência), sem aplicação de SMOTE, alcançando F1-Score de 0,8879, ROC AUC de 0,9444 e Acurácia de 0,8694 no conjunto de teste.

Como contribuição prática, foi desenvolvido um Produto Mínimo Viável (MVP) integrado a uma API em FastAPI, demonstrando a possibilidade de utilização do modelo preditivo em um fluxo automatizado de inferência. Embora desenvolvido para fins de demonstração, o MVP evidencia a viabilidade de integração do modelo a aplicações de apoio à decisão em risco de crédito.

Entre as limitações do estudo destacam-se a utilização de uma base pública com características estáticas, a ausência de informações comportamentais internas das instituições financeiras e a não incorporação de variáveis macroeconômicas. Além disso, a avaliação do modelo foi realizada em ambiente offline, sem análise contínua de desempenho após a implantação.

Como trabalhos futuros, recomenda-se avaliar o comportamento do modelo ao longo do tempo, monitorando possíveis ocorrências de *data drift* e *concept drift* (KELLEHER; NAMEE; D'ARCY, 2020), bem como investigar estratégias de retreinamento periódico e a integração de dados internos das instituições financeiras, incorporando informações cadastrais, comportamentais e transacionais para ampliar o poder preditivo dos modelos.

Por fim, visando favorecer a transparência e a reprodutibilidade da pesquisa, o código-fonte desenvolvido foi disponibilizado em repositório público na plataforma GitHub, conforme referência apresentada neste trabalho (HAMERSKI, 2026). Espera-se que os resultados obtidos contribuam para futuras pesquisas sobre análise preditiva de inadimplência utilizando dados públicos e para o desenvolvimento de soluções de apoio à gestão do risco de crédito.

REFERÊNCIAS

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.

DEMIRCIOĞLU, A. Applying oversampling before cross-validation will lead to high bias in radiomics. **Scientific Reports**, v. 14, n. 1, p. 11563, 2024. Disponível em: https://doi.org/10.1038/s41598-024-62585-z.

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.

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.

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.

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.

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/.

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]
```