Classificação de Solicitações de Crédito com Algoritmos Supervisionados:
Uma Abordagem com SVM e Árvores de Decisão

Dayane Freitas Felipe Ramos

CENTRO DE INFORMÁTICA

UNIVERSIDADE FEDERAL DA PARAÍBA

João Pessoa, PB

2025

Dayane Freitas Felipe Ramos

Classificação de Solicitações de Crédito com Algoritmos Supervisionados: Uma Abordagem com SVM e Árvore de Decisão

Artigo apresentado ao curso
Ciência de Dados e Inteligência Artificial
do Centro de Informática, da Universidade Federal da Paraíba,
como requisito para a obtenção do grau de bacharel.

Orientador: Tiago Maritan Ugolino de Araújo

Maio de 2025

Catalogação na publicação
Seção de Catalogação e Classificação

R175c Ramos, Dayane Freitas Felipe.
Classificação de solicitações de crédito com algoritmos supervisionados: uma abordagem com SVM e árvores de decisão / Dayane Freitas Felipe Ramos. - João Pessoa, 2025.
22 f. : il.

Orientação: Tiago Araújo.
Coorientação: Gilberto Filho.
TCC (Graduação) - UFPB/CI.

1. Aprendizado de máquina. 2. Árvores de decisão. 3. SVM. I. Araújo, Tiago. II. Gilberto Filho. III. Título.

UFPB/CI CDU 004.8

Elaborado por MICHELLE DE KASSIA FONSECA BARBOSA - CRB-738

CENTRO DE INFORMÁTICA

UNIVERSIDADE FEDERAL DA PARAÍBA

Trabalho de Conclusão de Curso de Ciência de Dados e Inteligência Artificial intitulado ***Classificação de Solicitações de Crédito com Algoritmos Supervisionados: Uma Abordagem com SVM e Árvore de Decisão*** de autoria de Dayane Freitas Felipe Ramos, aprovada pela banca examinadora constituída pelos seguintes professores:

Prof. Dr. Tiago Maritan Ugolino de Araújo
Universidade Federal da Paraíba (UFPB)

Prof. Dr. Gilberto Farias de Souza Filho
Universidade Federal da Paraíba (UFPB)

Prof. Dr. Thaís Gaudencio do Rêgo
Universidade Federal da Paraíba (UFPB)

Prof. Dr. Moises Dantas dos Santos
Universidade Federal da Paraíba (UFPB)

João Pessoa, 12 de maio de 2025

Centro de Informática, Universidade Federal da Paraíba
Rua dos Escoteiros, Mangabeira VII, João Pessoa, Paraíba, Brasil CEP: 58058-600
Fone: +55 (83) 3216 7093 / Fax: +55 (83) 3216 7117

CLASSIFICAÇÃO DE SOLICITAÇÕES DE CRÉDITO COM ALGORITMOS SUPERVISIONADOS: UMA ABORDAGEM COM SVM E ÁRVORE DE DECISÃO.

Dayane Freitas Felipe Ramos[1], Tiago Maritan Ugolino de Araújo[1], Gilberto Farias de Souza Filho[1]

Centro de Informática – Universidade Federal da Paraíba (UFPB)
João Pessoa – PB – Brasil

dayane.ramos@academico.ufpb.br, tiago@ci.ufpb.br, gilberto@ci.ufpb.br

Abstract. *This work proposes the use of Machine Learning techniques to analyze credit applications from clients of a risk management institution. Two approaches were developed: one using Decision Tree and the other using Support Vector Machine (SVM). Using a real dataset with behavioral and financial attributes, the objective was to predict whether a credit application would be approved or rejected. The results indicate better performance for the Decision Tree (accuracy 79.3%, recall 0.83) compared to the SVM (accuracy 73.2%, recall 0.57), with emphasis on its interpretability. The study also analyzed the impact of removing redundant features and the interpretability capacity of the models. Both approaches proved to be effective, with the Decision Tree standing out for its interpretability and the SVM for its stability.*

Resumo. *Este trabalho propõe o uso de técnicas de Aprendizado de Máquina (Machine Learning) para analisar solicitações de crédito de clientes de uma instituição de gestão de risco. Foram desenvolvidas duas abordagens: uma com Árvore de Decisão e outra com Máquinas de Vetores de Suporte (do inglês, Support Vector Machine - SVM). Utilizando uma base real com atributos comportamentais e financeiros, o objetivo foi prever se uma solicitação seria aprovada ou reprovada. Os resultados indicam melhor desempenho da Árvore de Decisão (acurácia 79,3%, recall 0,83) frente à SVM (acurácia 73,2%, recall 0,57), com destaque para sua interpretabilidade. Também se analisou o impacto da remoção de atributos redundantes e a capacidade interpretativa dos modelos. Os resultados mostram que ambas as abordagens são eficazes, com destaque para a interpretabilidade da árvore e a estabilidade da SVM.*

1. Introdução

A concessão de crédito é uma atividade essencial para o funcionamento do sistema financeiro, impulsionando o consumo, viabilizando investimentos e promovendo o crescimento econômico. No entanto, essa prática envolve riscos significativos, especialmente no contexto brasileiro, onde os índices de inadimplência se mantêm historicamente elevados. De acordo com levantamento do Serviço de Proteção ao Crédito - SPC Brasil (2024) e da Confederação Nacional de Dirigentes Lojistas CNDL (2024), cerca de 41,79% da população adulta brasileira encontrava-se inadimplente, o que corresponde a mais de 66 milhões de pessoas com alguma pendência financeira registrada nos órgãos de proteção ao crédito (CNDL/SPC, 2024). Esse cenário impõe um desafio

crítico às instituições financeiras, que precisam balancear a ampliação do acesso ao crédito, com mecanismos robustos de controle de risco.

Tradicionalmente, os modelos de avaliação de crédito baseiam-se em critérios manuais, como o histórico de pagamentos, renda declarada e consultas a cadastros restritivos. No entanto, tais métodos, embora amplamente utilizados, carecem de capacidade preditiva quando comparados a abordagens baseadas em dados e Aprendizado de Máquina. A crescente digitalização das operações financeiras tem proporcionado o acúmulo de grandes volumes de dados comportamentais, transacionais e operacionais, o que abre espaço para técnicas mais avançadas de classificação e predição de risco de crédito. Nesse cenário, algoritmos de Aprendizado de Máquina têm se destacado como alternativas eficazes para modelar o comportamento dos clientes e prever a probabilidade de inadimplência com maior precisão e adaptabilidade (Lessmann et al., 2015; Baesens et al., 2003).

Entre os algoritmos supervisionados, destacam-se as Máquinas de Vetores de Suporte (*Support Vector Machines* — SVM) e a Árvore de Decisão. Ambos têm sido amplamente aplicados no domínio da classificação de crédito, devido à sua robustez frente a conjuntos de dados de alta dimensionalidade e capacidade de modelar relações complexas entre variáveis. O SVM, fundamentado na Teoria de Vapnik-Chervonenkis (Vapnik, 1995), constrói hiperplanos ótimos de separação entre classes, buscando maximizar as margens entre os vetores de suporte, o que o torna altamente eficaz em cenários com grande variabilidade. Já as Árvores de Decisão, são conhecidas pela facilidade de interpretação, por traduzirem decisões em regras simples do tipo "se-então", o que as torna particularmente atrativas para aplicações reguladas, como é o caso do setor financeiro (Hand & Henley, 1997; Breiman et al., 1984).

Apesar da vasta aplicação de tais algoritmos em diversos contextos, ainda existe uma lacuna relevante quanto à sua análise comparativa utilizando dados reais e recentes do mercado brasileiro de crédito, sobretudo no que diz respeito à integração de variáveis comportamentais extraídas de sistemas internos de decisão. A ausência de padronização na coleta e análise dessas informações e a escassez de estudos que tratem da interpretabilidade dos modelos, no contexto regulatório brasileiro (como a Resolução BCB nº 4.893/2021 e a Lei Geral de Proteção de Dados — LGPD), reforçam a necessidade de investigações aplicadas, com base em dados reais e cenários operacionais.

Este trabalho tem como objetivo o desenvolvimento de uma solução baseada em ML, para a análise de solicitações de crédito de clientes de uma instituição financeira. Neste sentido, foram desenvolvidos dois modelos baseados nos algoritmos SVM e Árvore de Decisão, na tarefa de classificação de solicitações de crédito, utilizando dados reais extraídos da plataforma Capacitor, sistema interno de uma empresa especializada em gestão de risco e análise de crédito. A base de dados é composta por variáveis como score interno, percentual de comprometimento mensal, limite de crédito, classificação de risco, histórico de atrasos e presença de restrições. Os dados utilizados foram extraídos a partir de *logs* de execução do sistema Capacitor, os quais registram, em formato JSON, os detalhes de cada análise de crédito processada. Através da extração de caminhos hierárquicos (*paths*) em arquivos no formato JSON oriundos da plataforma — sistema interno utilizado para armazenar *logs* de decisão de crédito —, as variáveis de interesse foram mapeadas e estruturadas em colunas. Cada caminho corresponde a uma sequência de chaves, que aponta diretamente para um campo específico dentro da estrutura.

Os modelos foram avaliados com base em métricas de desempenho como acurácia, precisão, *recall* e f1-*score*, além da análise da matriz de confusão e da

interpretabilidade das decisões. Esta abordagem permite não apenas avaliar a eficácia preditiva dos algoritmos, mas também examinar sua viabilidade prática no contexto de sistemas de decisão automatizados, alinhados às exigências legais e operacionais do setor bancário e financeiro. Como resultado, o estudo propõe uma reflexão crítica sobre o uso de técnicas de aprendizado supervisionado na concessão de crédito, oferecendo subsídios técnicos para profissionais da área financeira, cientistas de dados e gestores interessados em incorporar soluções inteligentes e interpretáveis aos seus fluxos de análise de risco.

A principal contribuição deste trabalho reside na aplicação de modelos supervisionados sobre dados comportamentais reais, com ênfase na comparação entre desempenho e interpretabilidade, elementos cada vez mais valorizados em processos regulados. Ao trazer à tona os potenciais e limitações de cada abordagem, o estudo busca contribuir com o avanço das práticas de crédito baseadas em evidências, promovendo decisões mais justas, eficientes e alinhadas com as diretrizes de proteção ao consumidor e à estabilidade do sistema financeiro.

A escolha dos algoritmos SVM e Árvore de Decisão se justifica por suas características complementares: enquanto o SVM oferece desempenho mesmo em conjuntos de dados com alta dimensionalidade e desbalanceamento, a Árvore de Decisão se destaca pela interpretabilidade e transparência nas decisões, aspectos que são cruciais como os de concessão de crédito. Essa combinação permite explorar tanto a performance preditiva quanto a explicabilidade dos modelos.

1. Estrutura do Documento

Este artigo é dividido em seis seções principais, que têm o objetivo de conduzir o leitor desde o contexto geral do problema, até as conclusões que surgem da aplicação prática dos modelos. Começamos com uma introdução ao tema, onde destacamos a relevância da análise de crédito no Brasil e o papel dos algoritmos de aprendizado supervisionado, como uma alternativa para melhorar a precisão e a transparência nas decisões. Na sequência, a seção de fundamentação teórica explora os conceitos essenciais sobre concessão de crédito, as métricas de avaliação e os princípios dos algoritmos utilizados, oferecendo o embasamento necessário para entender a metodologia do estudo.

A seção dedicada aos trabalhos relacionados revisa pesquisas anteriores que aplicaram SVM e Árvores de Decisão em problemas de classificação similares, discutindo suas abordagens, limitações e contribuições, com o intuito de posicionar o presente estudo dentro do corpo de conhecimento existente. Na sequência, a seção de metodologia detalha o processo de extração, preparação e modelagem dos dados, descrevendo os critérios adotados para avaliação e comparação dos algoritmos. A seção de resultados e discussões apresenta os experimentos realizados, os indicadores obtidos e as implicações práticas das descobertas, destacando os pontos fortes e limitações de cada abordagem. Por fim, as considerações finais sintetizam as principais contribuições do estudo, indicam possíveis desdobramentos futuros e sugerem caminhos para novas investigações na área de análise de risco com suporte de inteligência artificial.

2. Fundamentação Teórica

A análise de crédito é um processo crítico no contexto financeiro, sendo responsável por determinar a viabilidade de concessão de recursos a indivíduos ou empresas, com base em sua capacidade presumida de pagamento. Historicamente fundamentada em modelos estatísticos tradicionais, essa prática tem evoluído rapidamente com o avanço das técnicas

de ciência de dados e Aprendizado de Máquina. O objetivo da análise de crédito é prever, com o máximo de acurácia possível, o risco de inadimplência associado a uma determinada solicitação, equilibrando critérios como rentabilidade, segurança e conformidade regulatória [Baesens et al., 2003].

Em instituições financeiras e *fintechs*, os modelos preditivos aplicados à concessão de crédito costumam se apoiar em um conjunto de variáveis que representam características socioeconômicas, histórico de comportamento financeiro e indicadores derivados de bases de dados externas. A seleção, tratamento e padronização dessas variáveis é uma etapa fundamental para o desempenho dos modelos, sendo que a incorporação de dados comportamentais — como atrasos médios, proporção de renda comprometida ou histórico de negativação — tem se mostrado cada vez mais relevante (LESSMANN et al., 2015). A utilização de bases estruturadas a partir de registros, como a adotada neste estudo, amplia a profundidade da análise, ao permitir que *logs* detalhados de decisões anteriores sejam transformados em variáveis explicativas.

Entre os algoritmos de aprendizado supervisionado, dois se destacam pela robustez e aplicabilidade na tarefa de classificação binária: a SVM e as Árvores de Decisão. O SVM é um método não paramétrico que busca encontrar um hiperplano ótimo capaz de separar as classes, de forma maximamente distante entre os vetores de suporte (VAPNIK, 1995). Essa técnica é eficaz, mesmo em cenários com dados de alta dimensionalidade e classes desbalanceadas, especialmente quando associada ao uso de funções kernel, como a RBF (*Radial Basis Function*), que permite mapear os dados para espaços de maior dimensão e resolver problemas não linearmente separáveis (BURGES, 1998). A regularização por meio do parâmetro C, que controla o grau de penalização para erros de classificação: valores menores de C permitem que o modelo erre mais durante o treinamento para encontrar uma separação mais simples e generalizável; enquanto valores maiores forçam o modelo a acertar mais os dados de treino, o ajuste de γ\gammaγ no kernel RBF são estratégias centrais para controlar a complexidade do modelo e evitar sobreajuste (*overfitting*).

As Árvores de Decisão são estruturas hierárquicas de regras do tipo "se-então", capazes de representar o processo de tomada de decisão de forma transparente e interpretável. No contexto da análise de crédito, sua principal vantagem está na legibilidade dos critérios utilizados para aprovação ou reprovação de propostas, algo que é especialmente importante diante de exigências legais como a Resolução BCB n.º 4.893/2021, que trata da explicabilidade de modelos automatizados, e da LGPD. A árvore é construída por divisões sucessivas do conjunto de dados, com base na maximização do ganho de informação ou redução da impureza (Gini, entropia), sendo possível aplicar técnicas de poda, como o *Minimal Cost-Complexity Pruning*, para reduzir a complexidade do modelo e mitigar o sobreajuste (BREIMAN et al., 1984).

A análise de importância das variáveis (*feature importance*) é outro elemento essencial no processo de modelagem supervisionada. Essa análise permite identificar quais atributos exercem maior influência sobre o resultado predito, auxiliando tanto na interpretação dos modelos, quanto na sua otimização. A matriz de correlação, nesse sentido, oferece uma primeira aproximação estatística da relação entre as variáveis independentes e a variável alvo, permitindo selecionar atributos altamente relevantes e eliminar redundâncias que poderiam comprometer a performance dos algoritmos. Visualizações como *pairplots* e *heatmaps* são ferramentas complementares que revelam padrões, interações e colinearidades presentes no conjunto de dados (MÜLLER; GUIDO, 2016).

A avaliação dos modelos se dá por meio de métricas clássicas da estatística aplicada à classificação, como a acurácia, a precisão, a sensibilidade (*recall)* e o *f1-score*, além da matriz de confusão. Cada uma dessas métricas fornece uma dimensão específica do desempenho preditivo: a acurácia indica a proporção total de acertos, enquanto a precisão e o *recall* avaliam, respectivamente, a confiabilidade e a abrangência das previsões em cada classe. O *f1-score*, por sua vez, sintetiza esses dois indicadores em uma única métrica harmônica, especialmente útil em cenários desbalanceados — como é comum no crédito, onde a proporção de aprovações tende a ser maior que a de reprovações (FREITAS et al., 2020). As fórmulas utilizadas para o cálculo dessas métricas (POWERS, 2011) são apresentadas a seguir:

- **Acurácia**: Proporção total de acertos (positivos e negativos) entre todas as previsões realizadas.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN}$$

- **Precisão**: Proporção de exemplos classificados como positivos que realmente são positivos.

$$Precisão = \frac{VP}{VP + FP}$$

- ***Recall*** (ou Sensibilidade): Proporção de exemplos positivos corretamente identificados.

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

- **F1-*Score***: média harmônica entre precisão e *recall*, útil especialmente quando há desbalanceamento entre as classes.

$$F1 = 2 \cdot \frac{Precisão + Recall}{Precisão + Recall}$$

Em que:

- VP = Verdadeiros positivos.
- VN = Verdadeiros negativos.
- FP = Falsos positivos.
- FN = Falsos negativos.

A otimização dos hiperparâmetros dos modelos, especialmente no SVM e nas Árvores de Decisão, é comumente realizada por meio de técnicas como o *Grid Search* com validação cruzada. Essa abordagem permite testar sistematicamente diferentes combinações de parâmetros e selecionar aquela que apresenta melhor desempenho médio nos subconjuntos de validação, garantindo maior generalização do modelo final. No caso do SVM, a escolha dos parâmetros CCC e γ\gammaγ influencia diretamente a capacidade do modelo de lidar com *outliers* e margens de separação estreitas. Já nas Árvores de Decisão, o parâmetro α\alpha do *cost-complexity pruning* define o grau de complexidade aceitável, controlando o número de folhas e a profundidade da árvore (SCIKIT-LEARN DEVELOPERS, 2024).

A etapa de otimização dos modelos preditivos foi conduzida por meio da técnica de validação cruzada *k-fold*, com $k = 5$, a qual divide o conjunto de treinamento em cinco subconjuntos de tamanhos semelhantes. Em cada iteração, quatro subconjuntos são usados para treinamento e um para validação, repetindo-se o processo até que todos tenham sido utilizados para validação uma vez. Essa técnica promove maior robustez estatística e reduz a variabilidade dos resultados. A implementação foi realizada com a classe *GridSearchCV*, da biblioteca *scikit-learn*, que executa uma busca exaustiva por combinações de hiperparâmetros. No modelo SVM, foram ajustados os parâmetros C (coeficiente de penalização), *gamma* (coeficiente do kernel RBF) e o tipo de *kernel* (linear e RBF). Já na Árvore de Decisão, o parâmetro *ccp_alpha* foi ajustado para controlar a complexidade da árvore, e gerada após a poda baseada em custo-complexidade.

Dessa forma, o alinhamento entre essas técnicas, neste contexto, é essencial para a construção de sistemas de decisão automatizada eficazes, éticos e auditáveis no setor financeiro.

3. Trabalhos Relacionados

A análise de crédito e a previsão de inadimplência têm sido temas recorrentes na literatura científica, com crescente interesse pelo uso de algoritmos de Aprendizado de Máquina, em substituição aos métodos estatísticos tradicionais. Para esta seção, foi conduzida uma pesquisa bibliográfica exploratória no repositório *Google Scholar*, utilizando strings como "credit scoring machine learning", "SVM credit risk", "decision tree loan approval", entre outras. Foram selecionados três trabalhos com critérios de inclusão baseados na aderência temática (foco em análise de crédito), uso explícito de algoritmos supervisionados, comparáveis ao escopo deste estudo (SVM e Árvores de Decisão) e publicação em veículos científicos reconhecidos.

O primeiro estudo analisado é o trabalho de **Torvekar e Game (2019)**, intitulado *Predictive Analysis of Credit Score for Credit Card Defaulters*. Neste estudo, os autores exploram a aplicação de algoritmos como Random Forest, SVM, Naive Bayes e Regressão Logística, para prever inadimplência em duas bases de dados públicas: o *Statlog Australian Credit* e o *Default of Credit Card Clients*. As métricas de avaliação utilizadas incluem acurácia, precisão e *recall*, tanto nas ferramentas WEKA quanto KNIME. Os resultados apontam desempenho superior do Random Forest em ambos os conjuntos de dados, seguido pela regressão logística e SVM. O estudo destaca a importância da personalização do modelo conforme o perfil da base de dados, mas não aprofunda aspectos como interpretabilidade, nem aborda técnicas de ajuste fino de parâmetro do modelo (*tuning)* avançado. Em contraste, o presente trabalho aplica *tuning* via validação cruzada, análise de correlação e ajusta pesos para lidar com o desbalanceamento, além de utilizar dados reais do setor financeiro nacional.

O segundo trabalho, de **Peixoto, Araki e Centeno (2016)**, aplica os algoritmos SVM e Árvores de Decisão para classificação de imagens de alta resolução do sensor ADS-40. Embora o domínio de aplicação seja diferente — sensoriamento remoto — o estudo apresenta contribuições relevantes sobre a aplicabilidade dos dois modelos, destacando as vantagens do SVM em cenários com grande variabilidade e a maior interpretabilidade das árvores. O experimento incluiu o uso de imagens RGB e infravermelhas, além de perfilamento a laser, e utilizou o algoritmo J48 para a árvore de decisão e SVM com kernel RBF. Os autores destacam que o SVM apresentou maior exatidão em termos de classificação, enquanto a árvore teve melhor desempenho

computacional. Este artigo é útil por reforçar a aplicabilidade dos dois algoritmos em contextos complexos.

O terceiro artigo, de **Deepika et al. (2022)**, aborda a *Implementation of Credit Card Fraud Detection Using Random Forest Algorithm*. Embora o foco principal seja a detecção de fraude — e não exatamente inadimplência — o estudo compara o desempenho de SVM, Random Forest e Naive Bayes, utilizando um conjunto de dados com mais de 284 mil transações. O algoritmo Random Forest demonstrou desempenho superior, seguido do SVM. Os autores utilizaram como base as métricas de acurácia, precisão, *recall* e f1-*score*, analisando separadamente os resultados sobre transações genuínas e fraudulentas. Apesar de relevante metodologicamente, o estudo limita-se à aplicação direta de classificadores padrão, sem incorporar técnicas de pré-processamento, engenharia de atributos ou análise de interpretabilidade dos modelos. Isso contrasta com a presente pesquisa, que enfatiza o uso de *feature importance* (importância das variáveis), ajuste de pesos em classes críticas e poda baseada em custo-complexidade.

Comparando os estudos revisados com a proposta deste trabalho, destacam-se três diferenciais principais: (i) o uso de dados reais internos com forte componente comportamental e histórico, obtidos de sistemas operacionais de crédito; (ii) o aprofundamento na explicação dos critérios decisórios, por meio da análise de importância de atributos e poda de árvores; e (iii) a aplicação de validação cruzada e *tuning* sistemático dos hiperparâmetros dos modelos, aumentando a robustez das conclusões. Enquanto a literatura explora majoritariamente dados públicos com foco em acurácia bruta, este trabalho visa também à explicabilidade e aplicabilidade real dos modelos, no processo decisório de concessão de crédito no Brasil.

4. Metodologia e Desenvolvimento

O desenvolvimento deste estudo foi estruturado em etapas sequenciais, iniciando-se com a preparação da base de dados, seguida da modelagem supervisionada com algoritmos de classificação e da aplicação de técnicas de ajuste de parâmetros e controle de complexidade. A arquitetura geral da solução foi implementada em linguagem Python, com apoio das bibliotecas *pandas, numpy, matplotlib, seaborn e scikit-learn*, as quais proporcionaram funcionalidades completas para tratamento de dados, modelagem preditiva, avaliação estatística e visualização.

Esse processo ocorre de acordo com o fluxo detalhado, apresentado na Figura 1, que ilustra a sequência metodológica desde a importação dos dados até a avaliação dos modelos.

Figura 1: Fluxograma da metodologia. Fonte: Elaboração própria pelo Figma.

A base de dados utilizada foi extraída de um sistema interno de análise de crédito, contendo registros de propostas analisadas por meio de variáveis comportamentais e financeiras dos solicitantes. O arquivo, em formato CSV, foi carregado utilizando a biblioteca *pandas*. O conjunto de atributos incluía campos como gênero, score interno, classificação de risco, percentual de comprometimento da renda, valor limite estimado de crédito e presença de restrições cadastrais. A variável-alvo, denominada resultado, indicava se a solicitação foi aprovada ou reprovada.

Para garantir maior transparência e reprodutibilidade na modelagem preditiva, a Tabela 1 apresenta a descrição detalhada dos atributos utilizados na base de dados. Cada atributo foi extraído e transformado em formato tabular para posterior análise. As variáveis englobam tanto características demográficas e comportamentais dos solicitantes, quanto informações históricas de crédito, sendo devidamente tratadas e codificadas durante a etapa de pré-processamento.

Atributo	**Tipo**	**Descrição**
IDADE	Numérico	Idade do solicitante no momento da proposta.
GÊNERO	Categórico	Gênero do solicitante: Masculino ou Feminino.
FILIAL	Categórico	Identificação da filial onde a proposta foi registrada.
SCORE_INTERNO	Numérico	Score interno de crédito gerado pelo sistema da empresa.
RENDA	Numérico	Renda mensal declarada do solicitante.
PERCENTUAL_LIMITE_MENSAL	Numérico	Percentual da renda comprometida com o limite de crédito solicitado.
LIMITE_TOTAL	Numérico	Valor total do limite de crédito solicitado.
CLASSIFICAÇÃO	Categórico	Classificação de risco atribuída ao solicitante.
DATA	Data	Data em que a proposta foi registrada.
ATRASO_MEDIO_DIAS	Numérico	Quantidade média de dias de atraso em contratos anteriores (se houver).
RESTRIÇÕES	Booleano	Indica se o cliente possui pendências ou registros negativos.
RESULTADO	Categórico	Resultado final da proposta.

Tabela 1: Descrição dos Atributos

Inicialmente, foram realizadas etapas de limpeza e pré-processamento dos dados. Variáveis categóricas como "GÊNERO" foram transformadas em valores binários, enquanto atributos nominais como "CLASSIFICAÇÃO" foram codificados numericamente com *LabelEncoder,* que é uma ferramenta do *scikit-learn* usada para converter variáveis categóricas em uma única coluna numérica, atribuindo m número inteiro único. A variável "RESTRIÇÕES" foi tratada como um booleano, assumindo valor 1, em casos de inadimplência ou pendências financeiras. Registros incompletos foram removidos, e a variável temporal "DATA" foi descartada, por não contribuir para a análise preditiva.

Em seguida, o conjunto de dados foi dividido entre variáveis preditoras (*X*) e variável-alvo (*y*). A separação entre treino e teste foi realizada utilizando a função *train_test_split* da biblioteca *scikit-learn,* com proporção de 70% para treino e 30% para teste. Como etapa de normalização, foi aplicada a padronização (*StandardScaler*) sobre os atributos numéricos, a fim de garantir que todas as variáveis contribuíssem de forma equilibrada para o processo de aprendizado dos modelos.

Durante a análise exploratória inicial, identificou-se que a base de dados apresentava desbalanceamento entre as classes, com predominância de propostas aprovadas. Essa característica é comum em cenários reais de crédito, nos quais a maior parte das solicitações tende a ser aprovada. Para evitar que esse desbalanceamento comprometesse o aprendizado dos modelos, foram adotadas estratégias específicas: no caso do SVM, foi utilizado o parâmetro *class_weight* com pesos ajustados, atribuindo maior penalização aos erros da classe "reprovado". Com intuito de equilibrar a influência das classes durante o treinamento, sem necessidade de alterar a distribuição original dos dados.

Antes da modelagem, foi realizada uma análise exploratória para compreensão das relações entre as variáveis. A matriz de correlação foi utilizada para identificar possíveis colinearidades e redundâncias, auxiliando na seleção de atributos mais representativos para o treinamento dos algoritmos. Visualizações gráficas como *heatmap* e *pairplot* também foram aplicadas para identificar padrões de separabilidade entre as classes da variável-alvo.

O primeiro algoritmo empregado foi a Árvore de Decisão (*DecisionTreeClassifier*), com critério de impureza do tipo Gini. Após o treinamento inicial, foi adotada a técnica de *Minimal Cost-Complexity Pruning,* com base no caminho de poda obtido por meio da função *cost_complexity_pruning_path.* Foram geradas diversas árvores com diferentes níveis de complexidade *(ccp_alpha*) e, com apoio de gráficos de validação, selecionou-se o modelo com melhor relação entre desempenho e simplicidade estrutural. A profundidade da árvore, o número de nós e a impureza total foram monitorados ao longo do processo, com apoio de visualizações em *matplotlib*.

O segundo algoritmo empregado foi o SVM, configurado inicialmente com kernel linear e pesos diferenciados entre as classes (*class_weight*= {0: 1,5, 1: 1}), de forma a mitigar o impacto do desbalanceamento da classe "reprovação". Posteriormente, aplicou-se a técnica de validação cruzada *k-fold* com $k = 5$, utilizando a classe *GridSearchCV* da biblioteca *scikit-learn* para realizar a otimização dos hiperparâmetros *C*, *gamma* e *kernel*. O conjunto de treinamento foi dividido em cinco partes iguais; em cada rodada, quatro partes foram usadas para treinamento e uma para teste, até que todas as partes tivessem sido utilizadas como teste uma vez. A média dos resultados obtidos nas cinco iterações foi utilizada para selecionar a combinação de parâmetros com melhor desempenho médio, garantindo a robustez e a capacidade de generalização do modelo. A validação cruzada

foi conduzida com *GridSearchCV*, que utiliza, por padrão, *k-fold* estratificado para tarefas de classificação. Nesse método, o conjunto de dados é dividido em *k* partes (ou *folds*) de tamanho semelhante. Em cada iteração, *k - 1* partes são utilizadas para o treinamento e 1 parte é reservada para validação. O processo se repete *k* vezes, até que todos os subconjuntos tenham sido utilizados como validação uma vez. A versão estratificada garante que a proporção entre as classes da variável alvo (aprovado/reprovado) seja mantida em cada uma dessas divisões, o que é especialmente importante em bases desbalanceadas. Dessa forma, evita-se que um *fold* contenha majoritariamente exemplos de apenas uma classe, o que comprometeria a avaliação realista do modelo. No presente estudo, foi adotado $k = 5$ para o modelo SVM e $k = 7$ para a Árvore de Decisão.

Adicionalmente, foi aplicada uma segunda etapa de otimização para o modelo de árvore utilizando *GridSearchCV* sobre os valores de *ccp_alpha*, combinando ajuste fino com validação cruzada (7-*fold*). Essa abordagem permitiu reduzir o risco de sobreajuste e identificar a estrutura ideal para a árvore com base em desempenho médio.

Todo o processo de modelagem e validação foi realizado de forma sistemática e reprodutível, utilizando controle de aleatoriedade por meio do parâmetro *random_state*, garantindo a consistência das partições de dados e dos resultados. A metodologia adotada buscou garantir tanto a robustez estatística dos modelos, quanto sua viabilidade prática para aplicação em sistemas para decisão de crédito.

5. Resultado e Discussões

A análise dos modelos preditivos propostos foi conduzida com base na avaliação quantitativa dos resultados obtidos por meio de métricas clássicas de classificação, bem como na interpretação visual e crítica dos padrões identificados no conjunto de dados. Os modelos SVM e Árvore de Decisão foram comparados em termos de desempenho, capacidade de generalização, sensibilidade à classe minoritária e interpretabilidade, buscando-se uma abordagem que conciliasse a precisão com transparência decisória.

5.1 Análise Exploratória e Seleção de Variáveis

A Figura 2 apresenta a matriz de correlação, os atributos FILIAL, GENERO, CLASSIFICAÇÃO e LIMITE_TOTAL foram descartados por apresentarem correlações próximas de zero ou comportamento redundante em relação a outras variáveis mais representativas. Por exemplo, a variável CLASSIFICAÇÃO apresentou correlação de (-0.07) com a variável alvo, sendo considerada estatisticamente irrelevante para fins preditivos.

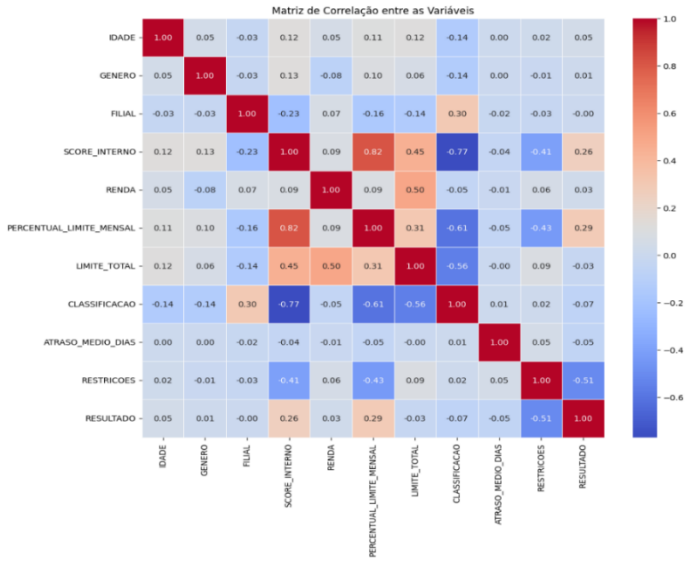

Figura 2: Matriz de Correlação

5.2 Resultados do Modelo de Árvore de Decisão

A árvore de decisão foi treinada inicialmente sem poda, alcançando média de 98,4% no conjunto de treino (*Ein* = 0,0162) e 79,3% no conjunto de teste (*Eout* = 0.2072), o que já indicava sinais de sobreajuste. O desempenho do modelo nesse cenário pode ser observado na Tabela 2, que apresenta as métricas de avaliação, com destaque para o *recall* da classe "reprovado", que atingiu 84%.

Além disso, a Figura 3 ilustra a estrutura inicial da árvore gerada nesse primeiro treinamento, evidenciando a elevada profundidade e a grande quantidade de nós, características típicas de modelos com alto risco de sobreajuste quando não é aplicado controle de complexidade.

	precisão	*recall*	*f1-score*	suporte
0	0,77	0,84	0,81	4743
1	0,82	0,74	0,78	4464
acurácia			0,79	9207
macro agv	0,80	0,79	0,79	9207
weighted avg	0,79	0,79	0,79	9207

Tabela 2: Resultados do modelo de Árvore de Decisão treinado sem técnica de poda. Métricas de desempenho obtidas antes da aplicação de estratégias de controle de complexidade. Fonte: Gerado em Python via biblioteca *scikit-learn.*

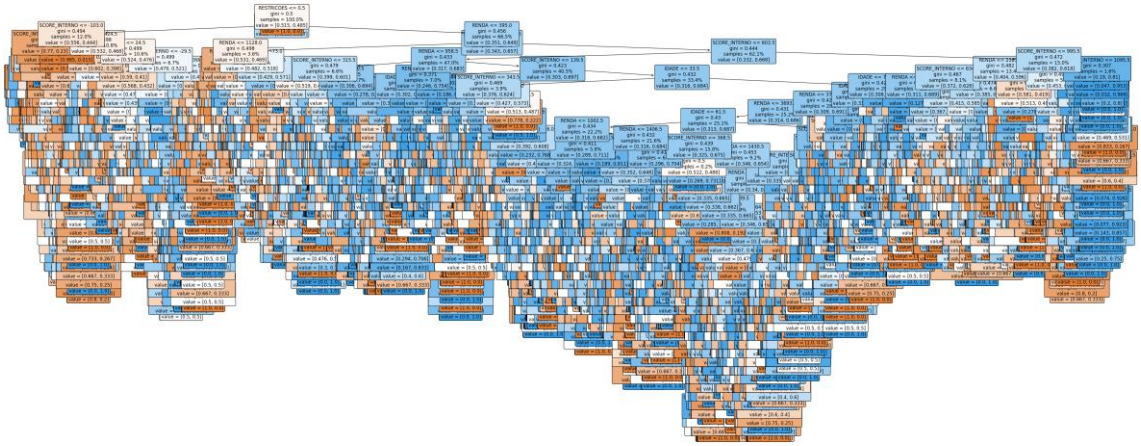

Figura 3: Estrutura inicial da Árvore de Decisão gerada antes da aplicação de poda, evidenciando a elevada profundidade e quantidade de nós. Fonte: Gerado em Python via *sklearn.tree.plot_tree*.

No entanto, a complexidade da estrutura — elevada profundidade e número de nós — sugeriu ocorrência de *overfitting*. Para mitigar esse problema, foi aplicada a técnica *de Minimal Cost-Complexity Pruning*. Foram testados 3.586 valores distintos de ccp_alpha, variando entre 0,000 e 0,128. Esses valores foram aplicados progressivamente para gerar árvores com diferentes níveis de complexidade. A Figura 4 apresenta a relação entre o Valor de Alpha e a Impureza Total das folhas (Impureza VS Alpha Efetivo), juntamente com a relação entre Número de Nós VS Alpha e Profundidade da Árvore VS Alpha. Essas visualizações permitiram compreender como a complexidade da árvore é reduzida, à medida que aumenta o valor de *ccp_alpha*. O ponto de equilíbrio identificado foi *ccp_alpha* = 0,002, que resultou em um bom compromisso entre simplicidade estrutural e desempenho preditivo.

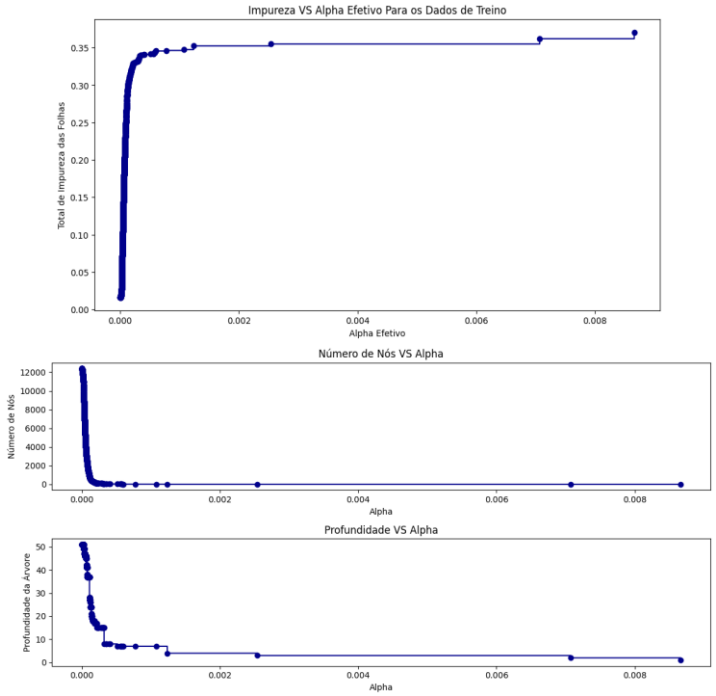

Figura 4: Curvas de impureza, número de nós e profundidade em função do *alpha* no processo de poda, evidenciando o impacto da complexidade na estrutura da árvore. Fonte: Gerado em Python via biblioteca *scikit-learn.*

A árvore sem regularização apresentou acurácia de 98,4% no conjunto de treino (*Ein* = 0.0162) e 79,3% no conjunto de teste (*Eout* = 0,2072). Após a regularização com *ccp_alpha* = 0,002, a acurácia caiu para 72,3% no treino (*Ein* = 0,2766) e 71,8% no teste (*Eout* = 0,2824). Essa perda de desempenho é explicada pelo fato de que, ao aplicar a poda, diversos ramos da árvore que estavam especializados em capturar padrões muito específicos dos dados de treino foram eliminados. Embora esses ramos tenham contribuído para uma alta acurácia no treino, eles não generalizavam bem para novos dados.

A Figura 5 representa a árvore redesenhada da versão final da árvore de decisão após regularização com *ccp_alpha* = 0.002. Nela, os nós de decisão exibem diretamente os atributos analisados (como restrições, idade, renda e score_interno) e os limiares utilizados nas divisões. A leitura da árvore permite observar que a presença de restrições cadastrais foi o fator mais decisivo para reprovação imediata, com 21,6% das amostras classificadas como "negado" logo na raiz. Para os demais, os caminhos de decisão incluem critérios como idade inferior a 29,5 anos e renda inferior a R$ 395,00, indicando que o modelo prioriza perfis com menor risco comportamental e maior capacidade de pagamento. Os nós finais (folhas) foram substituídos por rótulos explícitos – "NEGADO" ou "APROVADO".

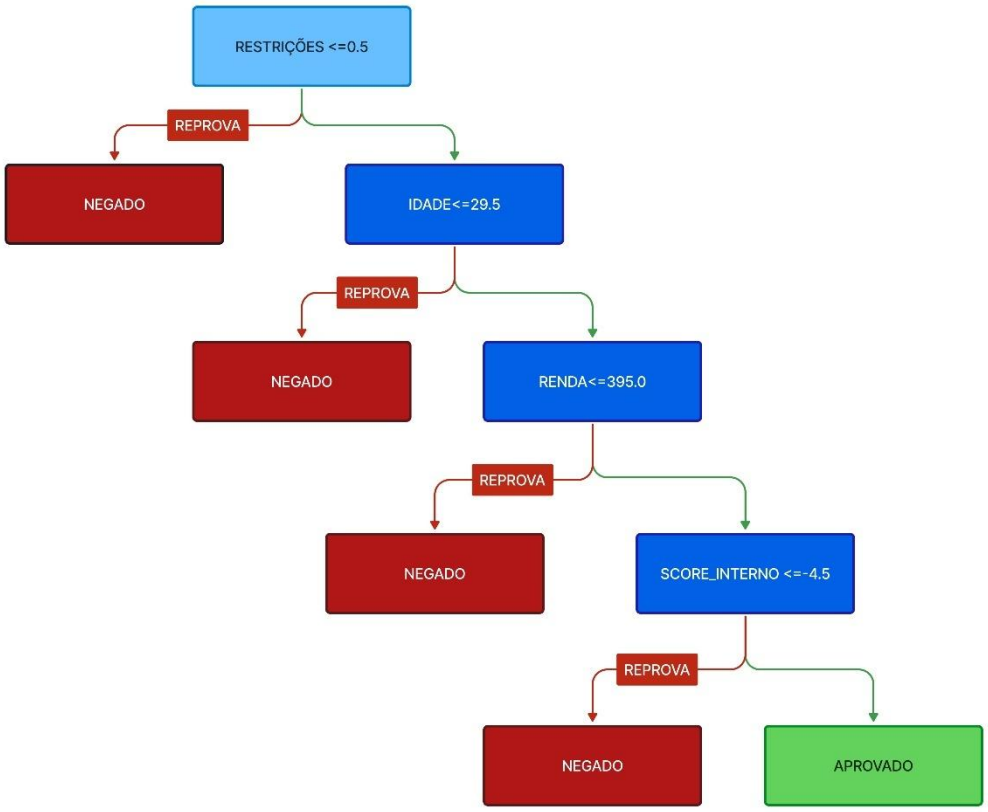

Figura 5: Fluxo simplificado da Árvore de Decisão final após aplicação de poda e ajuste de *ccp_alpha*.
Fonte: Elaboração própria utilizando a ferramenta Figma utilizei *ccp_alpha* do *scikit-learn* em Python.

Após a escolha preliminar de *ccp_alpha* = 0.002 com base na análise visual, o modelo foi refinado por meio de *GridSearchCV*, com validação cruzada em 7 grupos, para identificar o valor ótimo de *ccp_alpha* entre as 3.586 variações geradas. O modelo

final resultante obteve acurácia de 97,3% no treino (*Ein* = 0,0270) e 79,3% no teste (*Eout* = 0,2069), igualando o desempenho inicial, mas com uma estrutura de árvore muito mais interpretável. As métricas detalhadas estão apresentadas na Tabela 3.

	precisão	*recall*	*f1-score*	suporte
0	0,78	0,83	0,81	4743
1	0,81	0,75	0,78	4464
acurácia			0.79	9207
macro agv	0,79	0,79	0,79	9207
weighted avg	0,79	0,79	0,79	9207

Tabela 3: Desempenho do modelo de Árvore de Decisão após o ajuste dos parâmetros via *GridSearchCV*. Fonte: Elaboração própria com base nos resultados obtidos com validação cruzada utilizando *scikit-learn* em Python.

5.3 Resultados do Modelo SVM

O modelo SVM foi inicialmente treinado com hiperparâmetros padrão (*kernel rbf*, *C*=1, *gamma*='scale'), apresentando acurácia de 71,8% no conjunto de treino (*Ein* = 0,2816) e 71,5% no conjunto de teste (*Eout* = 0,2848). A matriz de confusão correspondente está representada na Figura 6.

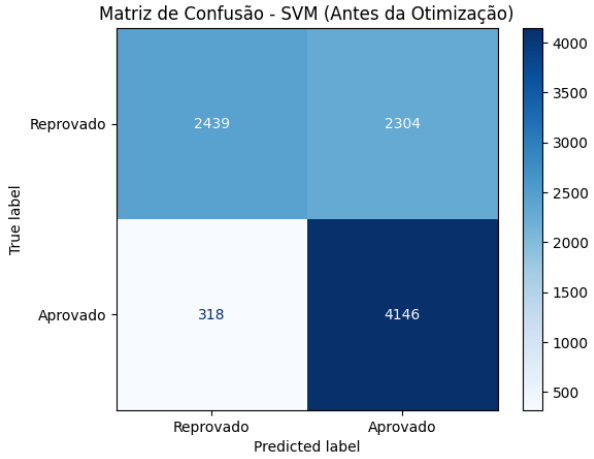

Figura 6: Matriz de Confusão antes da Otimização

A classe "Aprovado" foi classificada com precisão de 64% e *recall* de 93%, indicando forte sensibilidade para identificar corretamente os aprovados, mas baixa precisão na classe "Reprovado", sugerindo que o modelo apresentava viés favorável à classe majoritária. Os resultados completos antes da otimização dos hiperparâmetros estão apresentados na Tabela 4.

	precisão	*recall*	*f1-score*	suporte
0	0,88	0,51	0,65	4743
1	0,64	0,93	0,76	4464
acurácia			0,72	9207
macro agv	0,76	0,72	0,71	9207
weighted avg	0,77	0,72	0,70	9207

Tabela 4:Relatório de Classificação antes da otimização dos hiperparâmetros. Fonte: Elaboração própria com base nos resultados obtidos utilizando *scikit-learn* em Python.

Em seguida, aplicou-se a técnica *GridSearchCV* com validação cruzada 5-*fold* para otimizar os hiperparâmetros *C* e *gamma*, dentro de um intervalo de busca definido (*C* = [0,1, 1, 5, 10] e *gamma* = [0,01, 0,05, 0,1, 1]). O melhor modelo encontrado foi SVC (*C*=10, *gamma*=1, *kernel*='rbf').

Com os hiperparâmetros ajustados, a acurácia do modelo subiu para 73,8% no conjunto de treino (*Ein* = 0,2610) e 73,3% no teste (Eout = 0,2670), conforme mostra a Tabela 5. O desempenho geral também melhorou em termos de equilíbrio entre precisão e recall nas duas classes, especialmente para a classe "Reprovado", cuja taxa de acerto aumentou de 51% para 57%. Esse comportamento pode ser visto pela Figura 7, que apresenta a matriz de confusão após a otimização dos hiperparâmetros.

	precisão	*recall*	*f1-score*	suporte
0	0,86	0,57	0,69	4743
1	0,67	0,90	0,77	4464
acurácia			0,72	9207
macro agv	0,76	0,74	0,73	9207
weighted avg	0,77	0,73	0,73	9207

Tabela 5: Relatório de Classificação depois da otimização dos hiperparâmetros. Fonte: Elaboração própria com base nos resultados obtidos utilizando *scikit-learn* em Python.

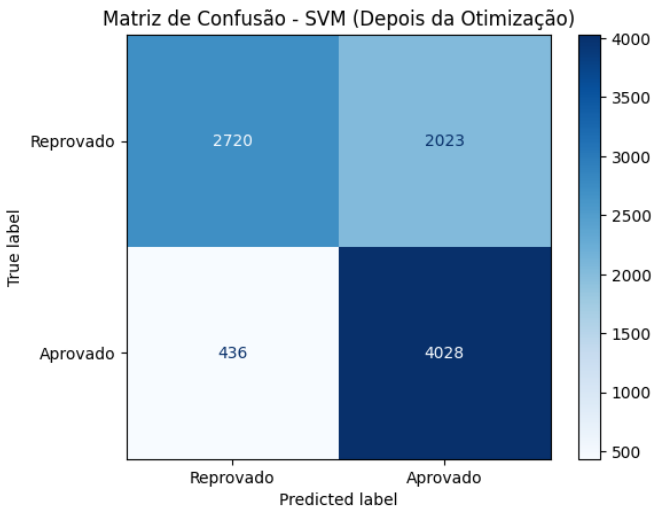

Figura 7: Matriz de Confusão depois da Otimização.

Esse ganho de desempenho evidencia a importância do ajuste de hiperparâmetros para maximizar a capacidade de generalização do SVM, reduzindo o viés e otimizando a fronteira de separação entre as classes. Embora o SVM não ofereça a mesma interpretabilidade das Árvores de Decisão, seu desempenho estável o torna uma alternativa robusta para cenários com maior complexidade de padrões e dados mais densos.

As Figuras 6 e 7 apresentam, respectivamente, as matrizes de confusão antes e depois da otimização. Pode-se notar que, após o *tuning*, o número de falsos negativos e falsos positivos foi reduzido, resultando em maior equilíbrio nas predições.

Com base nesses resultados, conclui-se que o SVM, quando otimizado, supera a performance inicial e se aproxima dos resultados obtidos pela Árvore de Decisão, oferecendo boa capacidade preditiva.

5.4 Comparação e Discussão Crítica

A análise comparativa entre o SVM e a Árvore de Decisão otimizados revela diferenças significativas em termos de desempenho preditivo, capacidade de generalização, sensibilidade às classes, e interpretabilidade, todos aspectos cruciais no contexto de análise de crédito. A Tabela 6 sintetiza as principais métricas de ambos os modelos, fornecendo uma visão geral para a discussão.

Modelo	Acurácia Treino	Acurácia Teste	*Ein*	*Eout*	*Recall* (Reprovado)	*f1-score* (Reprovado)	*f1-score* (Aprovado)
Árvore de Decisão	0,9730	0,7931	0,0270	0,2069	0,83	0,81	0,78
SVM	0,7389	0,7329	0,2611	0,2671	0,57	0,69	0,77

Tabela 6:Comparação de Desempenho entre SVM e Árvore de Decisão Otimizados. Fonte: Elaboração própria de acordo com as métricas obtidas.

A Árvore de Decisão se destacou pela capacidade de explicação das decisões, permitindo visualizar claramente as regras geradas (como no fluxo da Figura 5). Em contrapartida, o SVM demonstrou maior estabilidade ao longo das iterações de otimização, respondendo bem ao *tuning* de hiperparâmetros, e apresentou ligeira vantagem no equilíbrio entre as métricas das classes. Entretanto, sua estrutura torna o modelo menos intuitivo. Em síntese, a escolha entre os modelos depende do contexto de aplicação, se o foco for interpretabilidade e auditoria, a Árvore de Decisão é preferível. Se o foco for estabilidade e o SVM pode ser a melhor alternativa.

6. Considerações Finais

Este trabalho apresentou uma análise comparativa entre os algoritmos SVM e Árvores de Decisão aplicados à classificação de solicitações de crédito, utilizando um conjunto de dados reais extraído de um sistema interno de análise comportamental de risco. A partir da estruturação e pré-processamento das variáveis, foram aplicadas técnicas de

aprendizado supervisionado com foco em desempenho preditivo, interpretabilidade e robustez, alinhadas às necessidades operacionais e regulatórias do setor financeiro.

Como proposta para trabalhos futuros, sugere-se a implementação de técnicas de ensemble, como *Random Forest*, *XGBoost* ou *Voting Classifier*, com o objetivo de combinar os pontos fortes de diferentes algoritmos e potencializar o desempenho preditivo. A combinação de modelos pode contribuir para maior estabilidade, redução de viés e melhoria da generalização, especialmente em contextos com dados desbalanceados e variáveis com impacto diverso no processo decisório. Sugere-se, ainda, como continuidade deste estudo, a adaptação da abordagem para problemas de regressão, possibilitando prever variáveis numéricas relacionadas ao risco de crédito, como a probabilidade estimada de inadimplência ou o valor ideal de limite concedido.

A metodologia envolveu a preparação criteriosa dos dados, aplicação de técnicas exploratórias para seleção de atributos, treinamento inicial dos modelos, além da utilização de validação cruzada e ajuste de hiperparâmetros por meio de *GridSearchCV*. Adicionalmente, na Árvore de Decisão, foi empregada a técnica de poda baseada em custo-complexidade, promovendo o controle da sobreajuste e a simplificação do modelo. Os resultados obtidos demonstraram que ambos os algoritmos são eficazes para a tarefa de classificação.

7. Referências

BAESENS, Bart; SETIONO, Rudy; MUES, Christophe; VANTHIENEN, Jan. Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science*, v. 49, n. 3, 2003.

BREIMAN, Leo; FRIEDMAN, Jerome H.; OLSHEN, Richard A.; STONE, Charles J. *Classification and Regression Trees*. Belmont: Wadsworth, 1984.

BURGES, Christopher J. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, v. 2, n. 2, 1998.

DEEPIKA; RANI, Dr. Suman; RANI, Manju. Implementation of Credit Card Fraud Detection Using Random Forest Algorithm. *Journal of Pharmaceutical Negative Results*, v. 13, n. 3, 2022.

FREITAS, Flávia S.; SILVA, Felipe A.; LOPES, Ricardo H. Análise comparativa de algoritmos de classificação aplicados ao crédito bancário. *Revista de Sistemas e Computação*, v. 10, n. 1, 2020.

LESSMANN, Stefan; BAESENS, Bart; SEOW, Hsin-Vonn; THI, Lyn C. Benchmarking classification models for credit scoring: A survey and evaluation. *European Journal of Operational Research*, v. 247, n. 1, 2015. MÜLLER, Andreas C.; GUIDO, Sarah. *Introduction to Machine Learning with Python: A Guide for Data Scientists*. O'Reilly Media, 2016.

PEIXOTO, Ana Clara; ARAKI, Nelson; CENTENO, Jéssica. Utilização do algoritmo SVM (Support Vector Machine) e Árvores de Decisão para Classificação de Imagens de Alta Resolução do Sensor ADS-40. *Revista Brasileira de Cartografia*, v. 68, n. 6, 2016.

SCIKIT-LEARN DEVELOPERS. Decision Tree Pruning. Disponível em: https://scikit-learn.org/stable/modules/tree.html#cost-complexity-pruning. Acesso em: 25 abr. 2025.

TORVEKAR, Amruta; GAME, H. Predictive Analysis of Credit Score for Credit Card Defaulters. *International Journal of Recent Technology and Engineering (IJRTE)*, v. 8, n. 4, 2019.

VAPNIK, Vladimir N. *The Nature of Statistical Learning Theory*. New York: Springer, 1995.

Powers, D. M. W. (2011). *Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation*. Journal of Machine Learning Technologies