A Hybrid Clustering-Based Approach for Renewal Opportunity Prioritization in Payroll-Deducted Loan Portfolios

Edmagno do N. Lins[1], Ivan Zichtl Santos[1], Paulo Ribeiro Lins Júnior[1], Igor Barbosa da Costa[1]

Instituto Federal da Paraíba – IFPB
Programa de Pós Graduação em Tecnologia da Informação – PPGTI
edmagno.lins@academico.ifpb.edu.br, ivan.zichtl@academico.ifpb.edu.br,
paulo.lins@ifpb.edu.br, igor.costa@ifpb.edu.br

Abstract. The identification of renewal opportunities in payroll-deducted loan portfolios is an important task for financial institutions, as it helps sales teams prioritize contracts with greater operational potential for refinancing. In practice, however, historical labels indicating effective renewal conversions are often unavailable, inconsistent, or unreliable, limiting the direct use of traditional supervised learning approaches. This paper proposes a hybrid data mining approach for renewal opportunity prioritization in payroll-deducted loan portfolios. The method integrates information from loan contracts, payroll margin records, and liquidation history to build an analytical dataset. K-Means clustering is applied to identify contractual profiles, and the cluster with the highest operational adherence to renewal opportunities is used to generate pseudo-labels. Based on these pseudo-labels, a Random Forest model is used as an auxiliary supervised model to produce an operational prioritization score. Experiments were conducted as a pilot study using a real-world dataset with 739 contractual instances from a small municipal agreement. The results indicate that the selected target cluster is characterized by intermediate remaining term, relevant paid percentage, and low available margin. The proposed approach does not estimate actual renewal conversion probabilities, but provides an interpretable decision-support artifact for prioritizing renewal opportunities in scenarios where explicit conversion labels are not available.

CCS Concepts: • **Information systems** → **Data mining**; *Decision support systems*; • **Computing methodologies** → **Cluster analysis**; *Supervised learning by classification*.

Keywords: clustering, data mining, payroll-deducted loans, pseudo-labeling, random forest, renewal opportunity prioritization

1. INTRODUÇÃO

Os empréstimos consignados constituem uma modalidade de crédito caracterizada pelo desconto direto das parcelas em folha de pagamento ou benefício previdenciário. Por essa garantia de pagamento, são amplamente utilizados por aposentados, pensionistas do INSS, servidores públicos e trabalhadores formais, apresentando menor risco de inadimplência para instituições financeiras e, em geral, taxas de juros inferiores às de outras modalidades de crédito [Banco Central do Brasil 2018; Pinheiro 2020; Gomes et al. 2024]. Estudos sobre o tema apontam que o crédito consignado ampliou o acesso ao crédito e apresentou crescimento relevante no sistema financeiro brasileiro, inclusive em períodos de crise [Banco Central do Brasil 2018; Coelho et al. 2012; Schuh et al. 2017].

Nesse contexto, a renovação ou refinanciamento de contratos representa uma oportunidade estratégica para instituições financeiras e empresas processadoras de crédito. Diferentemente da prospecção de novos clientes, a renovação ocorre sobre uma carteira já conhecida, com histórico contratual, informações de margem consignável e registros de liquidações anteriores. Entretanto, os contratos podem estar em diferentes estágios de maturidade, com variações de prazo restante, valor de parcela, margem disponível e histórico de renegociação, tornando difícil a identificação sistemática de oportunidades relevantes.

Do ponto de vista de descoberta de conhecimento em bases de dados, o desafio consiste em identificar padrões contratuais e financeiros associados a maior potencial operacional de renovação, especialmente quando não há rótulos históricos explícitos, consistentes ou confiáveis sobre a efetiva conversão de contratos em renovações. Nesses cenários, abordagens supervisionadas tradicionais são limitadas, pois dependem de uma variável-alvo previamente observada. Técnicas não supervisionadas, por outro lado, podem auxiliar na segmentação da carteira e na identificação de perfis contratuais úteis para a construção de indicadores de priorização.

Diante desse contexto, este trabalho propõe uma abordagem híbrida para identificação e priorização de oportunidades de renovação em carteiras de crédito consignado. A proposta integra dados de consignações, margem consignável e liquidações, utiliza clusterização para identificar perfis de contratos e emprega pseudo-rotulagem para treinar um modelo supervisionado auxiliar capaz de gerar um *score* operacional de priorização. Esse *score* não é tratado como probabilidade real observada de conversão, mas como medida de aderência ao perfil contratual identificado como favorável à renovação.

O estudo foi conduzido como um projeto piloto a partir de dados reais de um convênio municipal de pequeno porte, totalizando 739 instâncias contratuais. Esse recorte foi escolhido intencionalmente para avaliar a viabilidade metodológica da abordagem em um ambiente operacional controlado, com dados previamente tratados e estrutura conhecida. Por se tratar de uma base de pequeno porte e origem única, os resultados devem ser interpretados como evidências iniciais de aplicabilidade, uma vez que a distribuição dos clusters e a heurística de seleção do cluster-alvo podem refletir características específicas do convênio analisado. Assim, a replicação em carteiras de maior porte e maior heterogeneidade é necessária para avaliar a generalização da abordagem.

A questão central investigada é: *como identificar e priorizar, em bases históricas de crédito consignado, contratos com maior aderência a perfis operacionais de renovação, mesmo na ausência de rótulos históricos explícitos de conversão?*

As principais contribuições deste trabalho são: (i) a construção de uma base analítica integrada a partir de diferentes fontes de dados relacionadas ao crédito consignado; (ii) a identificação de perfis contratuais por meio de clusterização; e (iii) a geração de um *score* operacional de priorização com base em pseudo-rotulagem e Random Forest, interpretado como aderência ao perfil-alvo e não como probabilidade real de renovação.

O restante do artigo está organizado da seguinte forma. A Seção 2 apresenta os trabalhos relacionados, a Seção 3 descreve os procedimentos metodológicos, a Seção 4 apresenta os resultados e discussões, e a Seção 5 reúne as considerações finais.

2. TRABALHOS RELACIONADOS

Técnicas de aprendizado de máquina têm sido amplamente utilizadas em aplicações financeiras, especialmente em problemas de análise de crédito, predição de inadimplência, avaliação de risco e apoio à decisão. Revisões recentes sobre *credit scoring* indicam o uso recorrente de algoritmos como Regressão Logística, Árvores de Decisão, Random Forest, Gradient Boosting, XGBoost e redes neurais para modelar risco e comportamento financeiro a partir de dados cadastrais, contratuais e transacionais [Ayari et al. 2026]. Embora esses estudos estejam geralmente voltados à concessão de crédito ou à previsão de inadimplência, eles demonstram a relevância de métodos de mineração de dados para apoiar decisões em carteiras financeiras.

Além da modelagem supervisionada, abordagens baseadas em segmentação têm sido exploradas para lidar com a heterogeneidade de clientes e contratos. Bosker et al. [Bosker et al. 2025] investigaram modelos de risco de crédito considerando estruturas de agrupamento, enquanto Bao et al. [Bao et al. 2019] propuseram a integração de algoritmos supervisionados e não supervisionados para avaliação de risco. Métodos como K-Means também têm sido empregados para segmentar clientes e identificar

perfis de comportamento financeiro ou de consumo [Tabianan et al. 2022; Aliyev et al. 2020]. Essas abordagens são relevantes para o presente trabalho por mostrarem que a segmentação pode apoiar a interpretação de bases heterogêneas e a caracterização de perfis operacionais.

A combinação entre segmentação e classificação também tem sido investigada em tarefas financeiras. Boughaci et al. [Boughaci et al. 2021] exploraram o uso de segmentação associada a modelos de classificação em problemas de *credit scoring* e predição de falência. Em cenários nos quais não há rótulos históricos explícitos para o evento de interesse, técnicas de aprendizado semissupervisionado e pseudo-rotulagem oferecem uma alternativa para a construção de modelos a partir de rótulos auxiliares ou aproximados [van Engelen and Hoos 2020; Fredriksson et al. 2025].

Apesar desses avanços, a maior parte dos estudos financeiros concentra-se em problemas supervisionados clássicos, como concessão de crédito, risco de inadimplência, falência ou abandono de clientes [He and Ding 2024]. O problema investigado neste trabalho é distinto: busca-se identificar contratos com maior aderência a perfis operacionais de renovação em crédito consignado, em um cenário no qual não há rótulos históricos confiáveis de conversão. Assim, este estudo se posiciona na interseção entre segmentação de contratos, pseudo-rotulagem e apoio à decisão operacional, utilizando a Random Forest como modelo auxiliar para transformar o perfil identificado pela clusterização em um *score* de priorização.

3. PROCEDIMENTOS METODOLÓGICOS

A metodologia proposta foi organizada em quatro macroetapas: (i) integração e pré-processamento das bases de dados; (ii) construção da base analítica e engenharia de atributos; (iii) segmentação não supervisionada dos contratos; e (iv) pseudo-rotulagem e geração de um *score* operacional de priorização.

3.1 Construção da Base Analítica e Engenharia de Atributos

A base de dados utilizada neste trabalho corresponde a um projeto piloto desenvolvido a partir de contratos de crédito consignado vinculados a um convênio municipal de pequeno porte. A escolha desse recorte foi intencional, com o objetivo de avaliar a metodologia proposta em um ambiente operacional controlado, utilizando dados reais e previamente tratados. Os dados foram utilizados de forma analítica e anonimizada, sem exposição de informações pessoais identificáveis.

A base analítica foi construída a partir da integração de três fontes de dados: a base de margem consignável, contendo 899 registros; a base analítica de consignações, contendo 739 registros; e a base de liquidações, contendo 2.835 registros. Essas fontes representam, respectivamente, informações sobre margem disponível para contratação de crédito, dados contratuais das operações consignadas e histórico de liquidações. O período coberto compreende operações registradas entre fevereiro de 2022 e maio de 2025.

A base analítica de consignações foi adotada como referência principal da integração, preservando a unidade de análise do estudo. As demais fontes foram utilizadas para enriquecer essa base com informações de margem e histórico de liquidações. Após procedimentos de limpeza, padronização, conversão tipológica e consolidação dos registros, o processo resultou em uma base final com 739 instâncias contratuais.

A engenharia de atributos teve como objetivo representar cada instância contratual por variáveis associadas ao ciclo de vida do contrato e à disponibilidade de margem. Para a etapa de modelagem, foram selecionadas quatro variáveis principais: *valor da parcela*, *prazo total*, *prazo restante* e *margem disponível*. Essas variáveis foram utilizadas tanto na clusterização quanto no treinamento do modelo Random Forest.

Variáveis derivadas, como percentual pago, foram preservadas para interpretação dos perfis e para o cálculo da heurística de seleção do cluster-alvo, mas excluídas do conjunto de entrada do K-Means e do Random Forest. Essa decisão evita redundância estrutural entre atributos, uma vez que o percentual pago é derivado do prazo total e do prazo restante. Incluí-lo na modelagem poderia atribuir peso duplicado à dimensão temporal do contrato, afetando as distâncias euclidianas no K-Means e a interpretação da importância dos atributos no Random Forest.

Antes da aplicação do K-Means, as variáveis numéricas utilizadas na clusterização foram padronizadas por Z-score, reduzindo o efeito de diferenças de escala entre atributos monetários e temporais.

3.2 Clusterização e Definição do Perfil-Alvo

A segmentação dos contratos foi realizada por meio do algoritmo K-Means. Foram testadas diferentes configurações de número de clusters, e a escolha de $k = 4$ foi apoiada pela análise conjunta do Método do Cotovelo e do Coeficiente de Silhueta, buscando equilibrar coesão interna, separação entre grupos e interpretabilidade operacional. O modelo final foi ajustado com os parâmetros `n_clusters = 4`, `init = k-means++`, `n_init = 'auto'` e `random_state = 42`. Os demais parâmetros permaneceram com os valores padrão da implementação `KMeans` utilizada.

Após a aplicação do K-Means, os clusters foram analisados a partir de seus valores médios nas variáveis utilizadas na modelagem e em atributos auxiliares de interpretação. A definição do cluster-alvo foi realizada por meio de uma heurística operacional baseada em dois componentes: razão de pagamento e pressão de margem.

Para cada cluster c, a razão de pagamento foi definida como:

$$RP_c = 1 - \frac{PR_c}{PT_c}$$

em que RP_c representa a razão de pagamento média do cluster, PR_c representa o prazo restante médio e PT_c representa o prazo total médio. A pressão de margem foi definida como:

$$PM_c = \frac{1}{MD_c + 1}$$

em que PM_c representa a pressão de margem do cluster e MD_c representa a margem disponível média. A partir dessas duas medidas, foi calculado um *score* heurístico para cada cluster:

$$H_c = RP_c \times PM_c$$

O cluster-alvo foi definido como aquele que apresentou o maior valor de H_c:

$$c^* = \arg\max_c H_c$$

A heurística utiliza o percentual pago de forma agregada por cluster, aplicando-se sobre os valores médios e não sobre instâncias individuais. Dessa forma, sua utilização ocorre apenas na seleção do perfil-alvo e não introduz vazamento de informação no treinamento do modelo auxiliar.

Essa regra prioriza grupos com maior percentual pago e menor margem disponível. Portanto, o cluster selecionado não representa uma classe real observada de renovação, mas um perfil contratual considerado operacionalmente aderente a oportunidades de renovação.

3.3 Pseudo-Rotulagem e Modelo Supervisionado Auxiliar

Após a definição do cluster-alvo, foi aplicada uma estratégia de pseudo-rotulagem. As instâncias pertencentes ao cluster de interesse foram rotuladas como classe positiva, enquanto as instâncias pertencentes aos demais clusters foram rotuladas como classe negativa. Esses pseudo-rótulos foram utilizados para treinar um modelo Random Forest.

O modelo Random Forest foi treinado com as mesmas quatro variáveis utilizadas na clusterização: valor da parcela, prazo total, prazo restante e margem disponível. Os hiperparâmetros definidos foram: `n_estimators = 150`, `max_depth = 6` e `random_state = 42`. Os demais hiperparâmetros permaneceram com os valores padrão do `RandomForestClassifier`.

A Random Forest foi utilizada como modelo supervisionado auxiliar, treinado a partir dos pseudo-rótulos derivados da clusterização. Seu objetivo não foi descobrir uma classe real de renovação, inexistente na base analisada, nem substituir a etapa não supervisionada. A função do modelo foi operacionalizar o perfil identificado pelo K-Means em um *score* contínuo de aderência ao cluster-alvo, permitindo ordenar contratos e analisar a importância relativa dos atributos utilizados.

Como não havia rótulos reais de conversão em renovação, as métricas calculadas para o Random Forest não devem ser interpretadas como avaliação preditiva de renovações reais. Para verificar apenas a consistência entre a etapa não supervisionada e o modelo auxiliar, foi realizada uma validação cruzada estratificada com cinco partições sobre o conjunto pseudo-rotulado. Essa avaliação mede a fidelidade do modelo aos pseudo-rótulos, isto é, sua capacidade de reproduzir a separação entre o cluster-alvo e os demais clusters.

Ressalta-se que, nesta etapa, o objetivo não é demonstrar superioridade em relação a regras de negócio ou medidas diretas de distância aos centroides, mas avaliar a viabilidade de transformar um perfil operacional obtido por clusterização em um indicador contínuo e interpretável de priorização.

4. RESULTADOS E DISCUSSÕES

Esta seção apresenta os resultados obtidos com a aplicação da abordagem proposta. Inicialmente, discute-se a escolha do número de clusters. Em seguida, são analisados os perfis contratuais identificados, a definição do cluster-alvo para pseudo-rotulagem, a fidelidade do modelo auxiliar aos pseudo-rótulos e a geração do *score* operacional de priorização.

4.1 Validação da Clusterização e Perfil dos Clusters

A definição do número de agrupamentos foi apoiada pela análise conjunta do Método do Cotovelo e do Coeficiente de Silhueta. O Método do Cotovelo indicou redução acentuada da variância intra-cluster até $k = 4$, com ganhos marginais menores para valores superiores. De forma complementar, o Coeficiente de Silhueta indicou que a configuração com quatro clusters apresenta equilíbrio adequado entre coesão interna, separação entre grupos e interpretabilidade operacional. A Figura 1 apresenta os resultados dessas métricas.

A aplicação do K-Means resultou na segmentação da base em quatro grupos, totalizando 739 instâncias contratuais. A Tabela I apresenta o perfil médio dos clusters nas variáveis de modelagem e em atributos auxiliares de interpretação. A coluna *Score médio de priorização* antecipa o resultado da etapa seguinte: trata-se da saída média do modelo auxiliar associada à classe pseudo-positiva, apresentada de forma consolidada para facilitar a leitura comparativa entre grupos. Sua derivação é detalhada na Seção 4.2.

Os resultados indicam que os clusters representam diferentes estágios do ciclo de vida contratual. O Cluster 0 reúne contratos com prazo restante médio de aproximadamente 35 meses, percentual

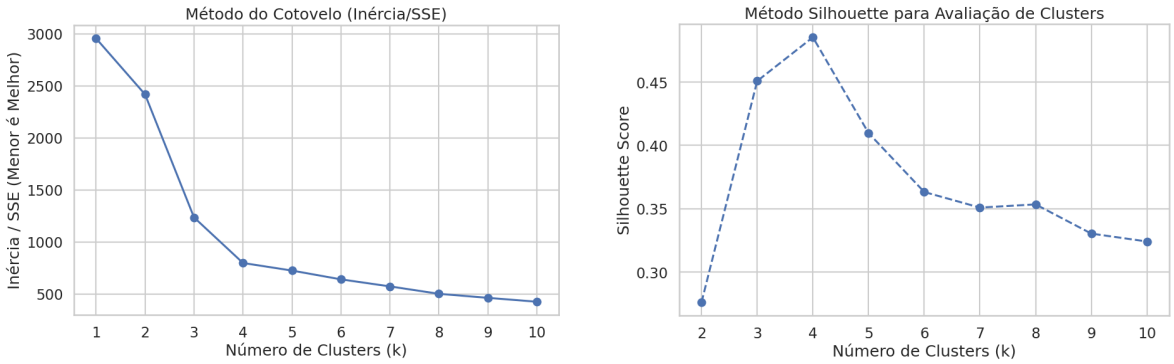

Fig. 1. Validação da clusterização por Método do Cotovelo e Coeficiente de Silhueta.

Table I. Perfil médio dos clusters obtidos pelo algoritmo de agrupamento.

Variável	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**
Valor da parcela (R$)	381,81	266,96	520,45	1459,00
Prazo total	63,12	122,36	57,10	113,93
Prazo restante	35,08	109,97	33,27	102,22
Margem disponível (R$)	9,07	3,99	56,20	7,34
Percentual pago	0,43	0,10	0,42	0,11
Score médio de priorização	0,97	0,00	0,01	0,02
Quantidade de instâncias	**168**	**388**	**92**	**91**

pago médio de 43% e margem disponível média de R$ 9,07. Esse perfil sugere contratos com maturidade suficiente para uma possível renovação, mas com baixa margem livre para novas contratações independentes.

O Cluster 2 apresenta percentual pago semelhante ao Cluster 0, em torno de 42%, mas possui margem disponível média mais elevada, de R$ 56,20. Essa diferença indica que o percentual pago, isoladamente, não é suficiente para definir o perfil de oportunidade, sendo necessário considerar também a restrição de margem. Já os Clusters 1 e 3 apresentam prazos restantes superiores a 100 meses e percentuais pagos próximos ou inferiores a 11%, caracterizando contratos menos maduros no ciclo de amortização.

A análise dos centroides e da heurística operacional indicou o Cluster 0 como grupo com maior aderência ao perfil operacional de renovação. Com base nessa análise, as instâncias pertencentes ao Cluster 0 foram utilizadas como classe positiva na etapa de pseudo-rotulagem, enquanto as instâncias dos demais clusters foram utilizadas como classe negativa. Assim, o target de treino não representa uma renovação efetivamente observada, mas sim a aderência ao perfil contratual definido como operacionalmente relevante para priorização.

4.2 Score de Priorização e Importância dos Atributos

Para avaliar se a Random Forest foi capaz de reproduzir a separação definida pela clusterização, aplicou-se validação cruzada estratificada com cinco partições sobre o conjunto pseudo-rotulado. O modelo obteve acurácia média de 99,30% (± 0,43%), F1-Score médio de 98,64% (± 0,85%) e ROC-AUC médio de 99,92% (± 0,13%). Esses resultados indicam que as variáveis utilizadas foram suficientes para reproduzir de forma consistente a fronteira entre o cluster-alvo e os demais grupos no recorte analisado. No entanto, essas métricas devem ser interpretadas como medidas de fidelidade aos pseudo-rótulos, e não como desempenho preditivo sobre renovações reais.

A partir dos pseudo-rótulos gerados pela clusterização, o modelo Random Forest produziu um *score* contínuo de aderência ao perfil operacional de renovação. Esse *score* foi utilizado para ordenar as instâncias contratuais e gerar um ranking de oportunidades para apoio à decisão operacional. A análise dos contratos mais bem posicionados no ranking mostrou coerência com o perfil esperado: os maiores *scores* concentram-se em contratos com prazos restantes intermediários, margens disponíveis reduzidas e percentuais pagos compatíveis com a janela operacional de renovação.

A análise de importância dos atributos do Random Forest permitiu avaliar quais variáveis mais contribuíram para reproduzir a separação entre o cluster-alvo e os demais grupos no conjunto pseudo-rotulado. Conforme apresentado na Figura 2, o atributo **prazo_restante** apresentou a maior importância relativa, com 37,9%, seguido por **margem_disponivel**, com 31,7%. Em conjunto, essas duas variáveis representaram aproximadamente 69,6% da importância atribuída pelo modelo.

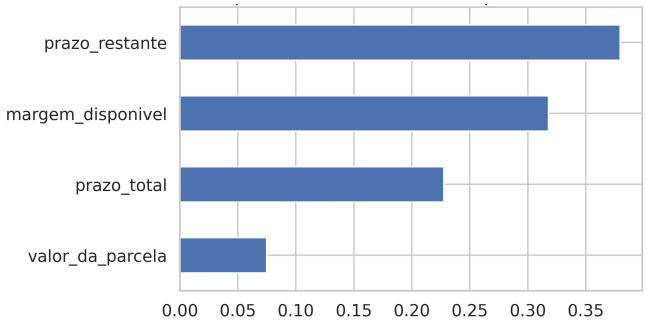

Fig. 2. Importância dos atributos no modelo Random Forest auxiliar.

Esse resultado deve ser interpretado como evidência de consistência interna do pipeline, e não como identificação de fatores causais ou preditivos de renovações reais. Como os pseudo-rótulos foram derivados da clusterização e o cluster-alvo foi selecionado por uma heurística baseada em maturidade contratual e pressão de margem, é esperado que variáveis relacionadas ao ciclo de vida do contrato e à margem disponível tenham maior peso no modelo auxiliar. Assim, a importância dos atributos indica principalmente quais variáveis contribuíram para reproduzir a separação operacional definida pelo próprio método. O prazo restante indica o estágio de maturidade da operação, enquanto a margem disponível sinaliza a possibilidade ou restrição para novas contratações. É importante destacar que o *score* apresentado não deve ser interpretado como probabilidade real de conversão em renovação. Como o modelo foi treinado a partir de pseudo-rótulos derivados da clusterização, o *score* representa a proximidade da instância em relação ao perfil operacional definido como favorável à renovação.

5. CONSIDERAÇÕES FINAIS

Este trabalho apresentou uma abordagem híbrida para identificação e priorização de oportunidades de renovação em carteiras de crédito consignado. A proposta integrou dados de consignações, margem consignável e liquidações para construir uma base analítica de contratos, combinando clusterização não supervisionada, pseudo-rotulagem e modelagem supervisionada auxiliar.

Os resultados indicaram que o Cluster 0 apresentou maior aderência ao perfil operacional de renovação, reunindo contratos com prazo restante intermediário, percentual pago relevante e baixa margem disponível. A partir desse grupo, foram gerados pseudo-rótulos utilizados no treinamento de um modelo Random Forest, responsável por produzir um *score* de priorização. A avaliação por validação cruzada indicou que o modelo reproduziu de forma consistente os pseudo-rótulos derivados da clusterização, enquanto a análise de importância dos atributos indicou que *prazo_restante* e

margem_ disponivel foram as variáveis mais relevantes para reproduzir a separação entre o cluster-alvo e os demais grupos no conjunto pseudo-rotulado.

A principal contribuição do trabalho está na construção de um artefato de apoio à decisão para cenários nos quais não há rótulos históricos explícitos e confiáveis de conversão em renovação. O *score* gerado não deve ser interpretado como probabilidade real de renovação, mas como uma medida de aderência ao perfil operacional identificado pela clusterização.

Como limitações, destacam-se o caráter piloto da base utilizada, restrita a um convênio municipal de pequeno porte, e o uso de pseudo-rótulos derivados da própria estrutura de agrupamento. Embora a validação cruzada indique que a Random Forest reproduziu de forma consistente os pseudo-rótulos, essa avaliação não substitui uma validação com conversões reais de renovação. Além disso, o estudo não compara, nesta etapa, o ranking gerado com baselines baseados em regras simples de negócio, como ordenação direta por percentual pago, margem disponível ou distância ao centroide do cluster-alvo. Como trabalhos futuros, pretende-se validar o ranking em cenário operacional, expandir a base para carteiras de maior porte e comparar o pipeline proposto com regras de negócio e medidas diretas de distância aos centroides. Essas extensões permitirão avaliar a estabilidade dos perfis identificados e a utilidade do *score* em diferentes contextos de crédito consignado.

Agradecimentos

Os autores agradecem ao Instituto Federal da Paraíba (IFPB) e ao Programa de Pós-Graduação em Tecnologia da Informação (PPGTI) pelo apoio institucional.

REFERENCES

Aliyev, M., Ahmadov, E., Gadirli, H., Mammadova, A., and Alasgarov, E. Segmenting bank customers via rfm model and unsupervised machine learning, 2020. arXiv preprint arXiv:2008.08662.

Ayari, H., Guetari, R., and Kraïem, N. Machine learning powered financial credit scoring: a systematic literature review. *Artificial Intelligence Review* 59 (13), 2026. Published online: 18 November 2025.

Banco Central do Brasil. Empréstimo consignado: características, acesso e uso. Relatório de Cidadania Financeira, 2018. Acesso em: 28 jun. 2026.

Bao, W., Ning, L., and Yue, K. Integration of unsupervised and supervised machine learning algorithms for credit risk assessment. *Expert Systems with Applications* vol. 128, pp. 301–315, 2019.

Bosker, J., Gürtler, M., and Zöllner, M. Machine learning-based variable selection for clustered credit risk modeling. *Journal of Business Economics* vol. 95, pp. 617–652, 2025. Published online: 03 December 2024.

Boughaci, D., Alkhawaldeh, A. A. K., Jaber, J. J., and Hamadneh, N. Classification with segmentation for credit scoring and bankruptcy prediction. *Empirical Economics* 61 (3): 1281–1309, 2021.

Coelho, C. A., De Mello, J. M. P., and Funchal, B. The brazilian payroll lending experiment. *The Review of Economics and Statistics* 94 (4): 925–934, 2012.

Fredriksson, T., Bosch, J., and Olsson, H. H. An empirical evaluation of deep semi-supervised learning. *International Journal of Data Science and Analytics* 20 (4): 4127–4148, 2025.

Gomes, R. d. S., Oliveira, E. R. d., Santos, G. C. d., and Gonçalves, R. R. O ambiente socioeconômico influencia o uso de crédito consignado? *REDECA, Revista Eletrônica do Departamento de Ciências Contábeis & Departamento de Atuária e Métodos Quantitativos* vol. 11, pp. e66978, 2024.

He, C. and Ding, C. H. Q. A novel classification algorithm for customer churn prediction based on hybrid ensemble-fusion model. *Scientific Reports* 14 (20179), 2024.

Pinheiro, M. A. P. Empréstimos bancários consignados de duas ou mais instituições financeiras: uma perspectiva à luz do direito do consumidor como direito fundamental. *Revista da Defensoria Pública da União* (13): 21–33, 2020.

Schuh, A. B., Coronel, D. A., and Bender Filho, R. Payroll loans and its relationship with the aggregate economic activity (2004–2014). *RAM. Revista de Administração Mackenzie* 18 (1): 148–173, 2017.

Tabianan, K., Velu, S., and Ravi, V. K-means clustering approach for intelligent customer segmentation using customer purchase behavior data. *Sustainability* 14 (12): 7243, 2022.

van Engelen, J. E. and Hoos, H. H. A survey on semi-supervised learning. *Machine Learning* vol. 109, pp. 373–440, 2020.