Segmentação de clientes para otimizar estratégias de negociação de dívidas

Frederico Antonio Domingues[1]*; Felipe Pinto da Silva[2]

[1] Monest. Engenheiro de Inteligência Artificial. Rua Guararapes, 1375, apto 703 – Vila Izabel; 80320-210, Curitiba, Paraná, Brasil
[2] Federação das Indústrias do Estado do Ceará (FIEC). Pesquisador. Av. Barão de Studart, 1980 – Aldeota; 60120-024. Fortaleza, CE, Brasil. Doutor em Economia pela Universidade Estadual de Campinas (UNICAMP).
*autor correspondente: fred_domingues@outlook.com

Segmentação de clientes para otimizar estratégias de negociação de dívidas

Resumo

O presente trabalho apresenta uma análise aplicada sobre a utilização de técnicas de aprendizado de máquina não supervisionado para segmentação de clientes inadimplentes, com o objetivo de otimizar estratégias de negociação de dívidas. O estudo foi conduzido com base em uma amostra sintética de 30.000 registros, estruturada a partir de variáveis sociodemográficas, financeiras e comportamentais, elaboradas por meio de distribuições estatísticas realistas inspiradas na literatura especializada. O pipeline analítico envolveu etapas de geração e preparação da base, pré-processamento com codificação categórica e padronização, escolha do número de clusters via Método do Cotovelo e Coeficiente de Silhueta, aplicação de três algoritmos de clusterização — K-Means, Hierárquico Aglomerativo e DBSCAN — e validação por métricas internas, incluindo Silhouette e Davies-Bouldin. Os resultados revelaram que o modelo K-Means com quatro clusters apresentou maior equilíbrio entre desempenho estatístico e interpretabilidade prática, permitindo a identificação de perfis distintos de devedores. As análises demonstraram a existência de grupos heterogêneos, variando de clientes de baixa renda e vulnerabilidade estrutural até profissionais de alta renda com dívidas elevadas, além de famílias em situação de sobre-endividamento. Tais achados fornecem subsídios robustos para a formulação de políticas de cobrança personalizadas, potencializando a eficiência na recuperação de crédito e reduzindo custos operacionais. O estudo ressalta o potencial transformador da ciência de dados na gestão de inadimplência e aponta recomendações para futuras pesquisas aplicadas com bases reais.

Palavras-chave: Segmentação de clientes; Negociação de dívidas; Recuperação de crédito.

Customer Segmentation for Optimizing Debt Collection Strategies

Abstract

This study presents an applied analysis of the use of unsupervised machine learning techniques for the segmentation of delinquent customers, aiming to optimize debt negotiation strategies. The research was based on a synthetic dataset of 30,000 records, structured with sociodemographic, financial, and behavioral variables, generated through statistically realistic distributions inspired by the literature. The analytical pipeline involved dataset construction, preprocessing with categorical encoding and normalization, cluster number determination through the Elbow Method and Silhouette Coefficient, application of three clustering algorithms — K-Means, Agglomerative Hierarchical, and DBSCAN — and validation using internal metrics, including Silhouette and Davies-Bouldin. Results showed that K-Means with four clusters offered the best balance between statistical performance and practical interpretability, enabling the identification of distinct debtor profiles. Analyses revealed heterogeneous groups ranging from low-income customers with structural vulnerability to high-income professionals with elevated debts, as well as families experiencing over-indebtedness. These findings provide solid insights for the development of personalized collection policies, increasing efficiency in credit recovery and reducing operational costs. The study highlights the transformative potential of data science in debt management and suggests directions for future applied research with real-world datasets.

Keywords: Customer Segmentation; Debt Negotiation; Credit Recovery.

Introdução

O endividamento das famílias brasileiras constitui um dos principais desafios contemporâneos do sistema financeiro nacional. Segundo estimativas recentes, em 2025 o

número de consumidores inadimplentes superou 74 milhões, representando parcela significativa da população economicamente ativa (Serasa, 2025). Dados da Pesquisa de Endividamento e Inadimplência do Consumidor [PEIC], realizada pela Confederação Nacional do Comércio [CNC], indicaram que 78,2% dos lares possuíam dívidas em maio de 2025, sendo que 29,5% estavam em atraso (CNC, 2025). Esses números ultrapassam a esfera individual, configurando risco sistêmico monitorado pelo Banco Central do Brasil em seus relatórios de estabilidade financeira, dada sua influência direta sobre liquidez, inadimplência bancária e resiliência macroeconômica (Banco Central do Brasil, 2025).

As estratégias tradicionais de recuperação de crédito, baseadas em abordagens massificadas, têm se mostrado ineficientes. Ao utilizar roteiros padronizados, as instituições não apenas alcançam baixas taxas de sucesso, como também geram experiências negativas para os clientes, degradando a relação comercial e potencialmente infringindo dispositivos do Código de Defesa do Consumidor (Brasil, 1990). Além disso, tais práticas acarretam elevados custos operacionais, especialmente em grandes carteiras de inadimplência, pressionando instituições a buscarem soluções mais inteligentes e escaláveis (Roesch & Scheule, 2016).

Neste contexto, a ciência de dados surge como alternativa estratégica. Técnicas de aprendizado de máquina permitem transformar dados brutos em insights acionáveis, capazes de fundamentar políticas de cobrança mais justas e eficazes. Dentre essas técnicas, a segmentação de clientes por meio da clusterização destaca-se por identificar grupos homogêneos a partir de múltiplos atributos, oferecendo um retrato multidimensional dos devedores (Han, Kamber & Pei, 2011; Baesens et al., 2016). Essa abordagem supera segmentações manuais, revelando padrões não triviais que podem orientar negociações personalizadas. Estudos recentes em *credit scoring* confirmam que a personalização das estratégias aumenta a taxa de recuperação e reduz o risco de reincidência (Lessmann et al., 2015; Kumar, Srivastav & Singh, 2021).

Diante desse cenário, este trabalho tem como objetivo principal aplicar e comparar algoritmos de clusterização a uma base sintética de clientes inadimplentes, com vistas à formulação de perfis distintos de devedores e à proposição de estratégias diferenciadas de cobrança. A pesquisa visa demonstrar a viabilidade técnica da abordagem, identificar limitações decorrentes do uso de dados sintéticos e discutir implicações práticas para o setor financeiro.

Metodologia

O presente estudo adotou uma abordagem quantitativa, exploratória e aplicada, com o objetivo de segmentar clientes inadimplentes utilizando técnicas de aprendizado não

supervisionado. A escolha por métodos quantitativos se justifica pela necessidade de analisar grandes volumes de dados e identificar padrões que possam subsidiar estratégias de negociação de dívidas mais eficazes (Creswell & Creswell, 2018). A pesquisa é exploratória por buscar compreender relações ainda não plenamente estabelecidas e aplicada por visar gerar resultados com relevância prática para a gestão de crédito.

A implementação foi realizada em Python (versão 3.11), devido à robustez da linguagem e ao vasto ecossistema de bibliotecas científicas disponíveis, amplamente utilizadas em análises de dados e machine learning (McKinney, 2012; Pedregosa et al., 2011). Para manipulação e limpeza de dados utilizou-se a biblioteca *pandas* (McKinney, 2012), operações numéricas foram realizadas com *NumPy* (Oliphant, 2006), e algoritmos de clusterização e métricas de validação foram aplicados via *scikit-learn* (Pedregosa et al., 2011).

A base de dados consistiu em 30.000 registros sintéticos, gerados com o objetivo de reproduzir características realistas de clientes inadimplentes, permitindo contornar restrições éticas e legais sobre o uso de dados sensíveis. A geração das variáveis seguiu distribuições estatísticas específicas para cada tipo de dado, garantindo plausibilidade, consistência e representatividade (Vlachos & Kollias, 2020). A Tabela 1 apresenta a estrutura da base de dados sintética, que contém 30.000 observações e 13 variáveis.

Tabela 1. Estrutura da base de dados sintética

(continua)

Variável	Tipo	Descrição
cliente_id	Discreta	Identificador único do cliente
idade	Contínua	Idade do cliente em anos
sexo	Nominal	Sexo do cliente (masculino ou feminino)
estado_civil	Nominal	Solteiro, Casado, divorciado ou viúvo

(conclusão)

Variável	Tipo	Descrição
nivel_educacional	Ordinal	Fundamental, médio, superior ou Pós-graduação
numero_dependentes	Discreta	Quantidade de dependentes entre 0 e 8.
tipo_emprego	Nominal	CLT, Autônomo, Funcionário Público, Empresário, Desempregado.
renda_mensal	Contínua	Renda mensal em Reais
score_credito	Contínua	pontuação que indica a probabilidade de um consumidor pagar as contas em dia
historico_pagamento_recente	Contínua	Histórico de pagamento recente, variando entre 0 e 1, quanto maior mais recente o pagamento.
produto_origem_divida	Nominal	Cartão de Crédito, Empréstimo Pessoal, Financiamento Veículo ou Cheque Especial
tempo_de_debito_meses	Contínua	Quantidade de meses da dívida em aberto.

valor_divida	Contínua	Valor total da dívida em Reais.

Fonte: Dados originais da pesquisa

As variáveis demográficas foram construídas para refletir padrões populacionais. Tendo isso em vista, a idade foi gerada por uma distribuição normal centrada em 45 anos, com desvio padrão de 15, limitada entre 18 e 85 anos, simulando a concentração da população adulta em torno de uma média (Montgomery & Runger, 2014). O sexo foi modelado como variável categórica binária, com distribuição probabilística de 48% masculino e 52% feminino, utilizando uma abordagem multinomial adequada para dados qualitativos discretos (Agresti, 2013). O estado civil incluiu quatro categorias — solteiro, casado, divorciado e viúvo — com probabilidades ajustadas para refletir proporções demográficas reais. O número de dependentes foi gerado a partir de uma distribuição de Poisson, apropriada para contagens de eventos discretos, garantindo que a maioria dos clientes possuísse poucos dependentes, com valores extremos pouco frequentes (Ross, 2014).

As variáveis nível educacional e tipo de emprego foram distribuídas probabilisticamente de modo a reproduzir a estrutura demográfica e ocupacional observada em dados reais. A Figura 1 ilustra as distribuições das variáveis demográficas quantitativas (idade e número_dependentes).

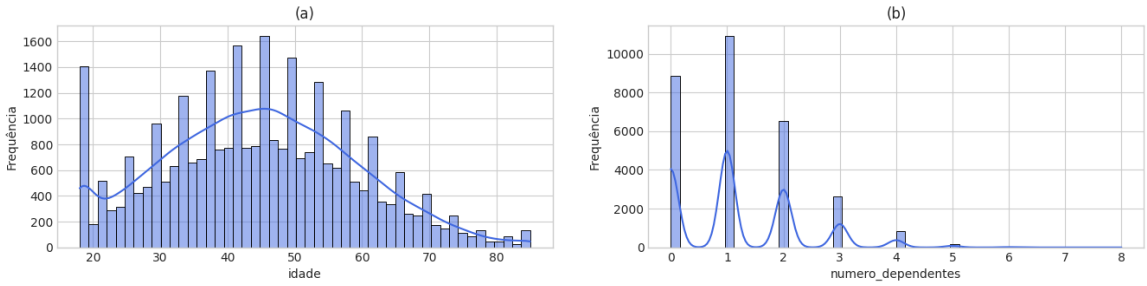

Figura 1. Distribuição de frequência das variáveis idade (a) e numero_dependentes (b).
Fonte: Resultados originais da pesquisa

Os dados financeiros e de dívida foram construídos considerando relações lógicas entre escolaridade, ocupação e comportamento de pagamento. A renda mensal foi calculada como produto de valores médios por nível educacional, modificadores por tipo de emprego e um fator de variabilidade aleatória uniforme, garantindo coerência e diversidade. O histórico de pagamento recente foi modelado via distribuição Beta, criando perfis de bons pagadores (valores próximos de 1) e pagadores de risco (valores mais baixos), adequada para variáveis contínuas limitadas no intervalo [0,1] (Gelman et al., 2013). O score de crédito foi definido

como uma função linear da idade, renda e histórico de pagamento, com ajustes para tipos de emprego e valores limitados entre 300 e 950, refletindo práticas usuais de modelagem de risco financeiro (Thomas, Crook & Edelman, 2002). A origem da dívida foi atribuída probabilisticamente a produtos como cartão de crédito, empréstimo pessoal, financiamento de veículo ou cheque especial, simulando a frequência observada no mercado. O tempo de débito em meses foi gerado a partir de uma distribuição exponencial, adequada para modelar a ocorrência decrescente de eventos, de modo que a maioria das dívidas fosse recente, enquanto casos antigos fossem progressivamente raros (Ross, 2014).

O valor da dívida foi definido como proporção da renda mensal, acrescido de variação aleatória uniforme, com limites mínimos para evitar registros irrealistas. A Figura 2 ilustra as distribuições das variáveis financeiras quantitativas (renda_mensal, score_credito, historico_pagamento_recente, tempo_de_debito_meses e valor_divida).

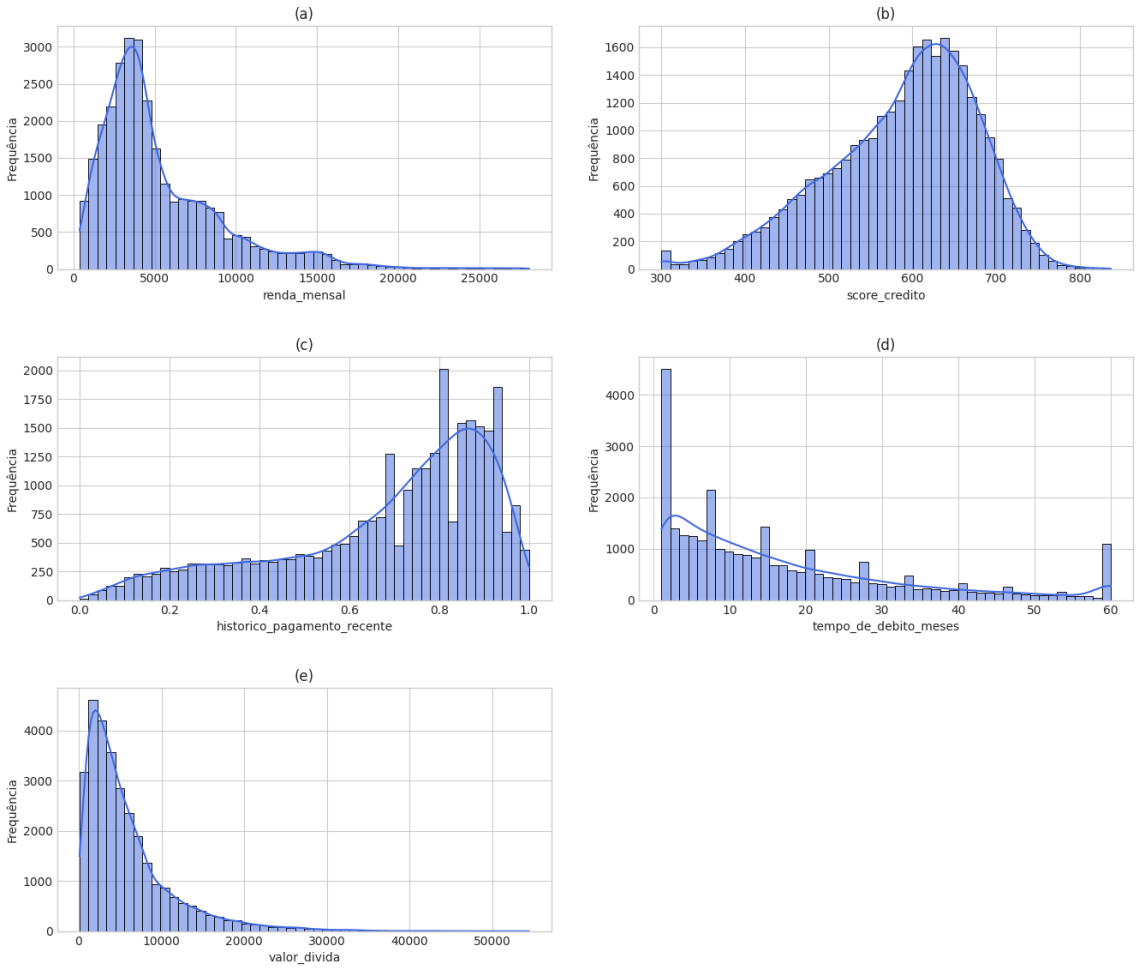

Figura 2. Distribuição de frequência das variáveis financeiras quantitativas renda_mensal (a), score_credito (b), historico_pagamento_recente (c), tempo_de_debito_meses (d) e valor_divida (e).

Fonte: Resultados originais da pesquisa

Quanto aos dados qualitativos, a variável sexo foi modelada como uma variável categórica binária, cujo resultado para cada cliente é uma instância de um ensaio de Bernoulli. A agregação desses ensaios para a amostra inteira segue uma distribuição multinomial com duas categorias (Agresti, 2013). As demais variáveis – estado civil, nível educacional, tipo de emprego e produto_origem_divida – foram geradas diretamente de uma distribuição multinomial generalizada para 2 ou mais categorias. A aplicação deste modelo presume que as observações são independentes e identicamente distribuídas, um pressuposto fundamental na inferência estatística clássica (Wasserman, 2004). A Figura 3 ilustra a contagem de valores para cada variável categórica da base de dados.

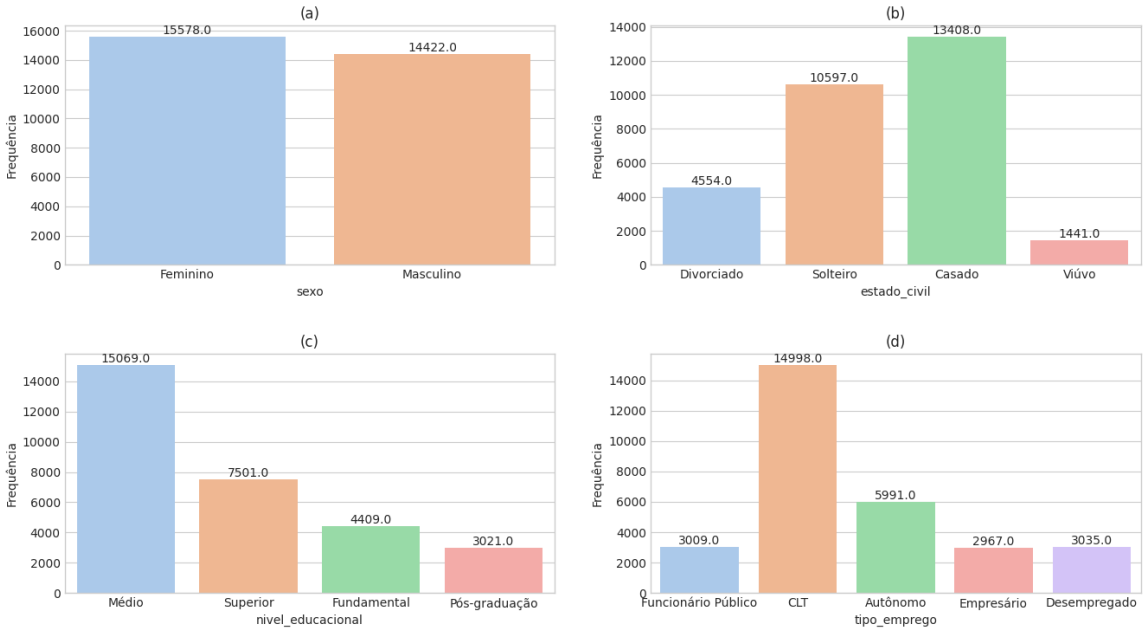

Figura 3. Distribuição de frequência das variáveis qualitativas de sexo (a), estado_civíl (b), nível_educacional (c) e tipo_emprego (d).
Fonte: Resultados originais da pesquisa

O pré-processamento dos dados consistiu em duas etapas principais. As variáveis categóricas foram convertidas em formato numérico utilizando One-Hot Encoding, técnica que cria colunas binárias correspondentes a cada categoria, tornando os dados compatíveis com algoritmos de machine learning que não aceitam valores textuais (Hastie, Tibshirani & Friedman, 2009). Formalmente, considere um conjunto de categorias $C = \{c_1, c_2, \dots, c_k\}$ e uma variável categórica $x \in C$. O One-*Hot* Encoding define uma função de mapeamento conforme demonstrado na form. (1):

$$f: C \rightarrow \{0,1\}^k, f(c_j) = (0, \ldots, 0,1,0, \ldots, 0), \quad (1)$$

onde o valor 1 aparece apenas na posição j, correspondente à categoria observada. Essa transformação equivale a representar cada categoria como um vetor da base canônica do espaço euclidiano R^k, garantindo que as variáveis categóricas sejam tratadas como entidades matematicamente manipuláveis pelos algoritmos (Murphy, 2012; Bishop, 2006).

Na segunda etapa, as variáveis numéricas foram padronizadas por meio da normalização Z-score, técnica que transforma os valores para média zero e desvio padrão unitário. Dado um valor x_i , sua transformação é definida conforme apresentado na form. (2):

$$z_i = \frac{(x_i - \mu)}{\sigma} \quad (2)$$

em que μ representa a média da variável e σ o desvio padrão. Esse procedimento evita que variáveis com escalas diferentes exerçam influência desproporcional nos algoritmos de aprendizado de máquina (James et al., 2013; Bishop, 2006).

Posteriormente ao pré-processamento, realizou-se uma análise da matriz de correlação de Pearson, para quantificar a intensidade e a direção das relações lineares entre as variáveis quantitativas. A correlação entre duas variáveis X e Y é dada por form. (3):

$$\rho_{X,Y} = \frac{(cov(X,Y))}{\sigma_X \sigma_Y} \quad (3)$$

onde $cov(X,Y)$ representa a covariância entre X e Y, e σ_X, σ_Y são os respectivos desvios padrão (Rodgers & Nicewander, 1988).

Esta etapa exploratória é fundamental no contexto da segmentação, pois a presença de multicolinearidade — alta correlação entre variáveis — pode introduzir redundância informacional e vieses de ponderação em algoritmos baseados em distância, como o K-Means (Hair et al., 2009).

A segmentação dos clientes foi realizada utilizando três algoritmos de clustering não supervisionado: K-Means, Hierárquico Aglomerativo e DBSCAN. O K-Means, amplamente utilizado por sua simplicidade e eficiência, busca particionar os dados em K grupos minimizando a soma das distâncias quadráticas entre os pontos e os centróides dos clusters conforme representado pela form. (4):

$$\sum_{K=1}^{K} W(C_k) = \sum_{K=1}^{K} \sum_{x_i \in C_k}^{K} \| x_i - \mu_k \|^2 \quad (4)$$

onde C_k representa o conjunto de pontos no cluster K e μ_k é o centróide desse cluster (Jain, 2010).

O Clustering Hierárquico Aglomerativo constrói uma hierarquia de clusters sem necessidade de definir K previamente. Inicialmente, cada observação é tratada como um cluster isolado, e em cada iteração os dois clusters mais próximos são unidos, até formar uma única estrutura. A distância entre clusters pode ser definida de diferentes formas; por exemplo, no critério de ligação completa conforme representado pela form. (5):

$$d(A, B) = \max_{x \in A, y \in B} \| x - y \| \tag{5}$$

onde A e B são clusters distintos (Murtagh & Contreras, 2012). A estrutura hierárquica resultante pode ser visualizada por meio de dendrogramas.

Por fim, o DBSCAN (Density-Based Spatial Clustering of Applications with Noise) identifica clusters de forma arbitrária a partir da densidade local de pontos. Um ponto p é considerado núcleo (*core point*) se o número de vizinhos dentro de um raio ε satisfaz. Sua definição matemática é representada pela form. (6):

$$| N_\varepsilon(p) | \geq MinPts \tag{6}$$

onde $N_\varepsilon(p) = q \in D \mid \| p - q \| \leq \varepsilon$ representa a vizinhança de p dentro do raio ε. Pontos que não satisfazem esse critério podem ser classificados como borda ou ruído (Ester et al., 1996). Esse método é particularmente adequado para identificar clusters em estruturas complexas de dados e detectar outliers.

O número de clusters para K-Means e Hierárquico foi definido combinando o Método do Cotovelo, que identifica o ponto em que a redução da soma dos quadrados intra-cluster se estabiliza (Hastie, Tibshirani & Friedman, 2009), e o Coeficiente de Silhueta, que avalia coesão intra-cluster e separação inter-cluster (Rousseeuw, 1987), resultando na escolha de quatro clusters.

A validação dos modelos foi realizada por métricas internas, incluindo Coeficiente de Silhueta e Índice de Davies-Bouldin, que avalia a razão entre dispersão intra-cluster e separação inter-cluster (Davies & Bouldin, 1979; Vardhan & Sharma, 2023). Para visualização e redução de dimensionalidade, aplicou-se a Análise de Componentes Principais (PCA), transformando variáveis correlacionadas em componentes ortogonais, preservando a maior parte da variância original e permitindo inspeção gráfica bidimensional dos clusters (Jolliffe & Cadima, 2016).

Por fim, os clusters foram caracterizados a partir das médias das variáveis numéricas e das modas das variáveis categóricas, traduzindo agrupamentos estatísticos em perfis interpretáveis de clientes inadimplentes, subsidiando estratégias segmentadas de negociação de dívidas.

Resultados e Discussão

A análise da matriz de correlações das variáveis presentes na base de dados sintética, apresentada na Figura 4, realça e confirma a utilização adequada das distribuições de dados aplicadas à construção das variáveis da base de dados.

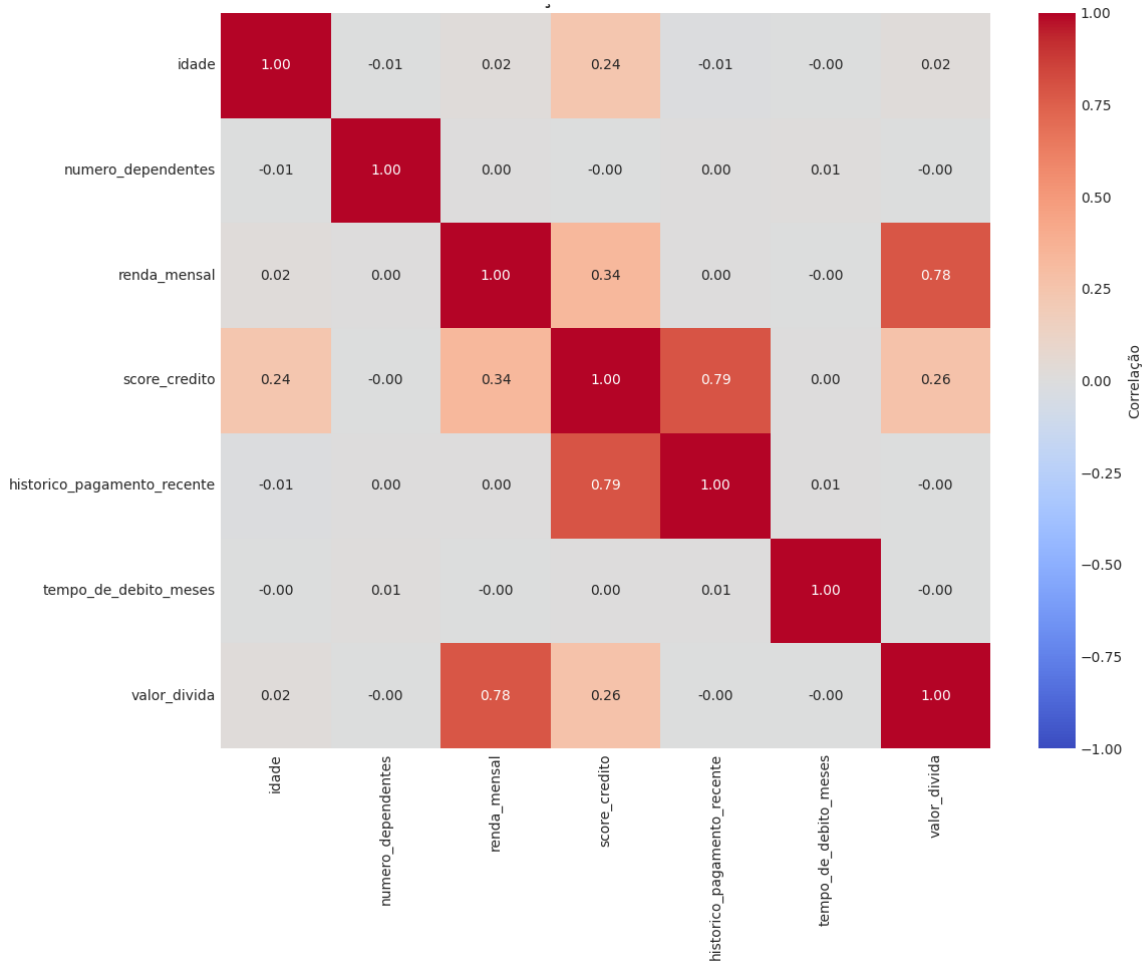

Figura 4. Matriz de correlação de Pearson das variáveis quantitativas da base de dados.
Fonte: Resultados originais da pesquisa

Conforme observado na Figura 4, as associações lineares mais proeminentes são as correlações positivas muito fortes entre o score_credito e o historico_pagamento_recente (r = 0.79), e entre o valor_divida e a renda_mensal (r = 0.78). A primeira relação confirma a

premissa fundamental dos modelos de risco, onde o comportamento de pagamento recente é um preditor primário da pontuação de crédito. A segunda indica que indivíduos com maior renda tendem a contrair dívidas de maior volume, refletindo o acesso ampliado a produtos de crédito.

Em um nível secundário, observam-se correlações positivas, porém mais fracas, que adicionam nuances ao perfil dos clientes, como a associação entre renda_mensal e score_credito (r = 0.34) e entre idade e score_credito (r = 0.24). Contudo, uma observação crítica reside na quase total ausência de correlação das variáveis numero_dependentes e tempo_de_debito_meses com todas as outras, sugerindo que foram geradas de forma independente no conjunto de dados sintéticos.

Quanto à execução da clusterização sobre a base sintética de 30.000 registros, foi possível identificar padrões consistentes e diferenciados de inadimplência. O processo de determinação do número de clusters, por meio do Método do Cotovelo e do Coeficiente de Silhueta, convergiu para a escolha de quatro grupos, resultado confirmado tanto pela análise visual da curva de inércia quanto pela avaliação dos índices de coesão e separação, vide Figura 5.

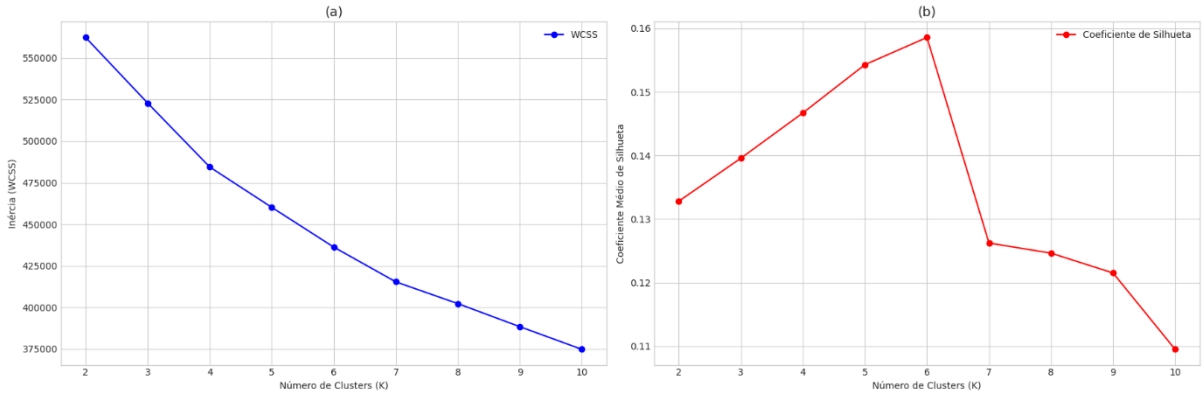

Figura 5. Definição da quantidade de clusters utilizando Método de Cotovelo (a) e Análise de Silhueta (b).
Fonte: Resultados originais da pesquisa

Entre os algoritmos comparados, K-Means apresentou melhor equilíbrio entre desempenho estatístico e interpretabilidade. O modelo gerou quatro clusters com valores de Coeficiente de Silhueta (0,147) e Índice de Davies-Bouldin (2,17) satisfatórios, superiores ao desempenho do Hierárquico (Silhueta 0,146; Davies-Bouldin 2,40) e mais robustos que o DBSCAN, cujo Silhueta (0,064) evidenciou excessiva fragmentação apesar do Davies-Bouldin ligeiramente melhor (1,92). Assim, K-Means mostrou-se o mais adequado para bases amplas e de alta dimensionalidade, confirmando resultados reportados na literatura (Hastie, Tibshirani

& Friedman, 2009; Lessmann et al., 2015). A Tabela 2 apresenta a comparação entre os 3 modelos utilizados neste trabalho.

Tabela 2. Avaliação comparativa dos modelos

Modelo	Coeficiente de Silhueta	Índice de Davies-Bouldin
K-Means	0.146702	2.171531
Hierárquico	0.145850	2.399244
DBSCAN	0.064376	1.917484

Fonte: Dados originais da pesquisa

A caracterização dos clusters revelou quatro perfis distintos. A Tabela 3 apresenta o valor médio para cada variável quantitativa, enquanto a Tabela 4 apresenta o valor mais comum (moda) entre as variáveis qualitativas. Ambas as tabelas são apresentadas para ilustrar a identificação e caracterização das particularidades de cada cluster encontrado aplicando o modelo de K-Means.

Tabela 3. Valores médios das variáveis quantitativas para cada cluster.

Variáveis	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Idade	44.25	44.67	45.05	44.87
numero_dependentes	1.22	1.20	1.22	1.21
renda_mensal	1508.08	3627.35	13974.05	8187.46
score_credito	477.75	589.05	640.75	612.00
historico_pagamento_recente	0.67	0.68	0.68	0.68
tempo_de_debito_meses	17.13	16.82	16.64	17.06
valor_divida	1645.83	4003.36	15454.66	9090.22
Quantidade de clientes	3035	17503	2691	6771

Fonte: Dados originais da pesquisa

Tabela 4. Valores mais comuns (moda) das variáveis qualitativas para cada cluster.

(continua)

Variáveis	Cluster 0	Cluster 1	Cluster 2	Cluster 3
sexo	Feminino	Feminino	Feminino	Feminino
estado_civil	Casado	Casado	Casado	Casado
nivel_educacional	Médio	Médio	Pós-Graduação	Superior
tipo_emprego	Desempregado	CLT	CLT	CLT

(conclusão)

Variáveis	Cluster 0	Cluster 1	Cluster 2	Cluster 3
produto_origem_divida	Cartão de Crédito	Cartão de Crédito	Cartão de Crédito	Cartão de Crédito
Quantidade de clientes	3035	17503	2691	6771

Fonte: Dados originais da pesquisa

Realizando uma análise crítica sobre cada cluster identificado, nota-se que cada grupo apresenta perfis socioeconômicos distintos, exigindo abordagens específicas de negociação e recuperação de crédito.

O primeiro grupo concentrou clientes de baixa renda, desempregados e com dívidas acumuladas, caracterizando alto risco de inadimplência estrutural. Para esse perfil, a literatura aponta a necessidade de estratégias de renegociação com forte redução da parcela inicial e planos de pagamento flexíveis, priorizando a capacidade de pagamento do devedor (Souza & Araújo, 2019). Em alguns casos, políticas de quitação antecipada com descontos substanciais também podem ser mais eficazes do que acordos longos (Borges, 2018).

O segundo grupo, majoritário na amostra, reuniu trabalhadores assalariados com renda intermediária e dívidas proporcionais. Este é o cluster mais representativo para estratégias de parcelamento padronizado em massa, com foco em escalabilidade e automação das propostas, uma vez que a previsibilidade da renda favorece planos de parcelamento fixo (Rocha & Silva, 2020).

O terceiro grupo destacou profissionais de alta renda e maior escolaridade, com dívidas elevadas em valor absoluto, mas maior capacidade de negociação. Nesses casos, a literatura recomenda a adoção de soluções mais sofisticadas, como consolidação de dívidas (junção em uma única parcela) ou alongamento dos prazos de pagamento, com taxas de juros competitivas para garantir a atratividade da proposta (Martins & Pereira, 2021).

O quarto cluster evidenciou famílias de renda intermediária com dívidas que superam a própria renda mensal, caracterizando sobre-endividamento estrutural. Para esse grupo, as estratégias mais adequadas envolvem renegociações sustentáveis de longo prazo, alinhadas à literatura sobre políticas de crédito responsável e programas de reeducação financeira (González & Cortés, 2016). Abordagens que incluem períodos de carência ou redução temporária de encargos podem ser fundamentais para evitar reincidência da inadimplência (Oliveira & Fernandes, 2019).

Através da Figura 6, é possível visualizar graficamente as diferenças e particularidades de cada cluster para as variáveis quantitativas, reforçando a importância de estratégias diferenciadas de negociação para maximizar a efetividade da recuperação de crédito.

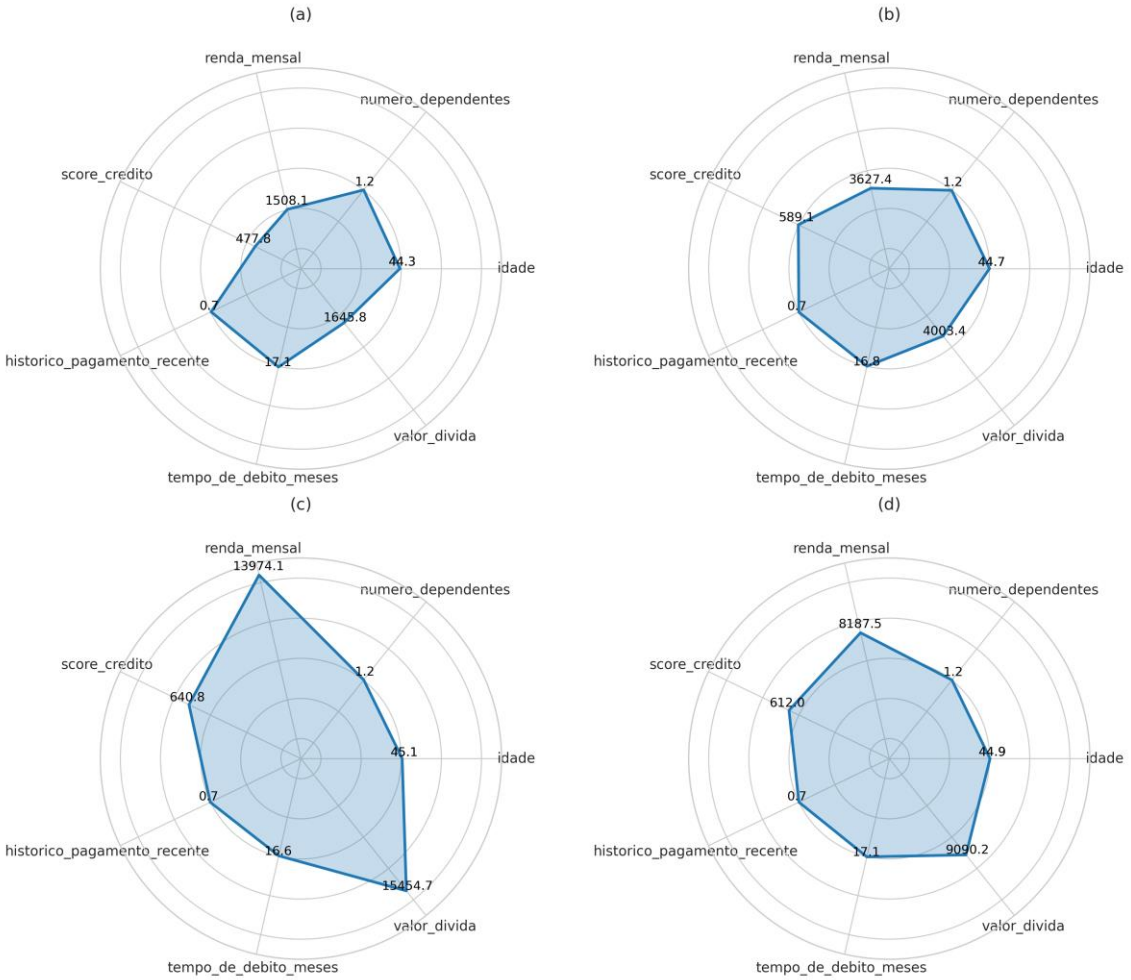

Figura 6. Visualização gráfica das diferenças de cada cluster.
Fonte: Resultados originais da pesquisa
Nota: Cluster 0 (a); Cluster 1 (b); Cluster 2 (c); Cluster 4 (d)

Esses achados confirmam o potencial da segmentação como ferramenta estratégica. Enquanto perfis de baixa renda exigem políticas de desconto e liquidação rápida, grupos de maior renda demandam soluções diferenciada.

A predominância de dívidas oriundas de cartão de crédito em todos os clusters reflete o cenário real de endividamento no país (CNC, 2025; Serasa, 2025), reforçando a validade dos resultados. Além disso, a identificação de clusters heterogêneos permite não apenas otimizar a recuperação financeira, mas também reduzir riscos reputacionais, ao alinhar a cobrança à capacidade de pagamento e perfil social de cada cliente.

O uso de dados sintéticos representa uma limitação, pois não contempla toda a complexidade e variabilidade dos contextos reais, como sazonalidade, comportamento de inadimplência regional e fatores subjetivos. Ainda assim, a metodologia desenvolvida demonstra viabilidade técnica e estabelece uma base sólida para aplicações futuras em bases reais.

Considerações Finais

O trabalho demonstrou que técnicas de clusterização são eficazes para segmentar clientes inadimplentes em perfis heterogêneos, fornecendo suporte à formulação de políticas de negociação personalizadas. O modelo K-Means com quatro clusters apresentou desempenho satisfatório e interpretabilidade clara, permitindo transformar agrupamentos estatísticos em recomendações práticas para estratégias de cobrança.

A pesquisa evidenciou a relevância da ciência de dados no contexto da recuperação de crédito, mostrando que abordagens orientadas a dados podem superar práticas tradicionais homogêneas, aumentando as taxas de recuperação e reduzindo custos operacionais. Além disso, os perfis obtidos reforçam a importância da personalização na negociação de dívidas, aspecto central para práticas mais justas e sustentáveis.

Como próximos passos, recomenda-se aplicar o pipeline a bases reais de inadimplência, incorporar variáveis adicionais, incluindo dados não estruturados como interações em canais de atendimento, e explorar modelos supervisionados capazes de prever a probabilidade de pagamento em diferentes cenários de renegociação. Tais extensões poderão ampliar a aplicabilidade prática dos resultados e consolidar a segmentação como ferramenta estratégica no setor financeiro.

Agradecimento

Agradeço à minha namorada e meus irmãos, pelo apoio constante e incentivo em todas as etapas da minha trajetória acadêmica, e aos meus colegas de trabalho, pela troca de experiências práticas e aprendizados que enriqueceram a realização desta pesquisa.

Referências

Agresti, A. 2013. Categorical Data Analysis (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Banco Central do Brasil. 2025. Relatório de Estabilidade Financeira (v.24, n.1). Disponível em: https://www.bcb.gov.br/publicacoes/ref. Acesso em: 15 jun. 2025.

Baesens, B., Roesch, D., & Scheule, H. 2016. Credit Risk Analytics: Measurement Techniques, Applications, and Examples in SAS. Hoboken, NJ: Wiley.

Brasil. 1990. Lei nº 8.078. Dispõe sobre a proteção do consumidor e dá outras providências. Diário Oficial da União.

Confederação Nacional do Comércio de Bens, Serviços e Turismo (CNC). 2025. Pesquisa de Endividamento e Inadimplência do Consumidor (Peic) – Maio de 2025. Disponível em: https://www.portaldocomercio.org.br/publicacoes. Acesso em: 15 jun. 2025.

Creswell, J. W., & Creswell, J. D. 2018. Research Design: Qualitative, Quantitative, and Mixed Methods Approaches (5th ed.). Thousand Oaks, CA: Sage Publications.

Davies, D. L., & Bouldin, D. W. 1979. A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1(2), 224–227.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96) (pp. 226–231).

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. 2013. Bayesian Data Analysis (3rd ed.). Boca Raton, FL: Chapman and Hall/CRC.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. 2009. Multivariate Data Analysis (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Han, J., Kamber, M., & Pei, J. 2011. Data Mining: Concepts and Techniques (3rd ed.). San Francisco, CA: Morgan Kaufmann.

Hastie, T., Tibshirani, R., & Friedman, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). New York, NY: Springer.

Jain, A. K. 2010. Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666.

James, G., Witten, D., Hastie, T., & Tibshirani, R. 2013. An Introduction to Statistical Learning: with Applications in R. New York, NY: Springer.

Jolliffe, I. T., & Cadima, J. 2016. Principal component analysis: a review and recent developments. Philosophical Transactions of the Royal Society A, 374(2065), 20150202.

Kumar, A., Srivastav, S., & Singh, P. 2021. Machine Learning in Finance: From Theory to Practice. Cham: Springer.

Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: an update of research. European Journal of Operational Research, 247(1), 124–136.

McKinney, W. 2012. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. Sebastopol, CA: O'Reilly Media.

Montgomery, D. C., & Runger, G. C. 2014. Applied Statistics and Probability for Engineers (6th ed.). Hoboken, NJ: Wiley.

Murtagh, F., & Contreras, P. 2012. Algorithms for hierarchical clustering: an overview. WIREs Data Mining and Knowledge Discovery, 2(1), 86–97.

Oliphant, T. E. 2006. A Guide to NumPy. USA: Trelgol Publishing.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Baesens, Bart; Roesch, Daniel; Scheule, Harald. Credit Risk Analytics: The R Companion. Hoboken, NJ: Wiley, 2016.

Ross, S. M. 2014. A First Course in Probability (9th ed.). Upper Saddle River, NJ: Pearson.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65.

Serasa. 2025. Mapa da Inadimplência e Renegociação de Dívidas no Brasil. Disponível em: https://www.serasa.com.br/limpa-nome-online/blog/mapa-da-inadimplencia-e-renogociacao-de-dividas-no-brasil/. Acesso em: 26 Jul. 2025.

Thomas, L. C., Crook, J. N., & Edelman, D. B. 2002. Credit Scoring and Its Applications. Philadelphia, PA: SIAM.

Vardhan, S., & Sharma, R. 2023. Evaluation metrics for clustering algorithms: A comprehensive review. International Journal of Data Science and Analytics, 15(2), 112–135.

Vlachos, M., & Kollias, G. 2020. Synthetic data generation in machine learning: Principles and applications. Journal of Big Data, 7, 79.

Wasserman, L. 2004. All of Statistics: A Concise Course in Statistical Inference. New York, NY: Springer.