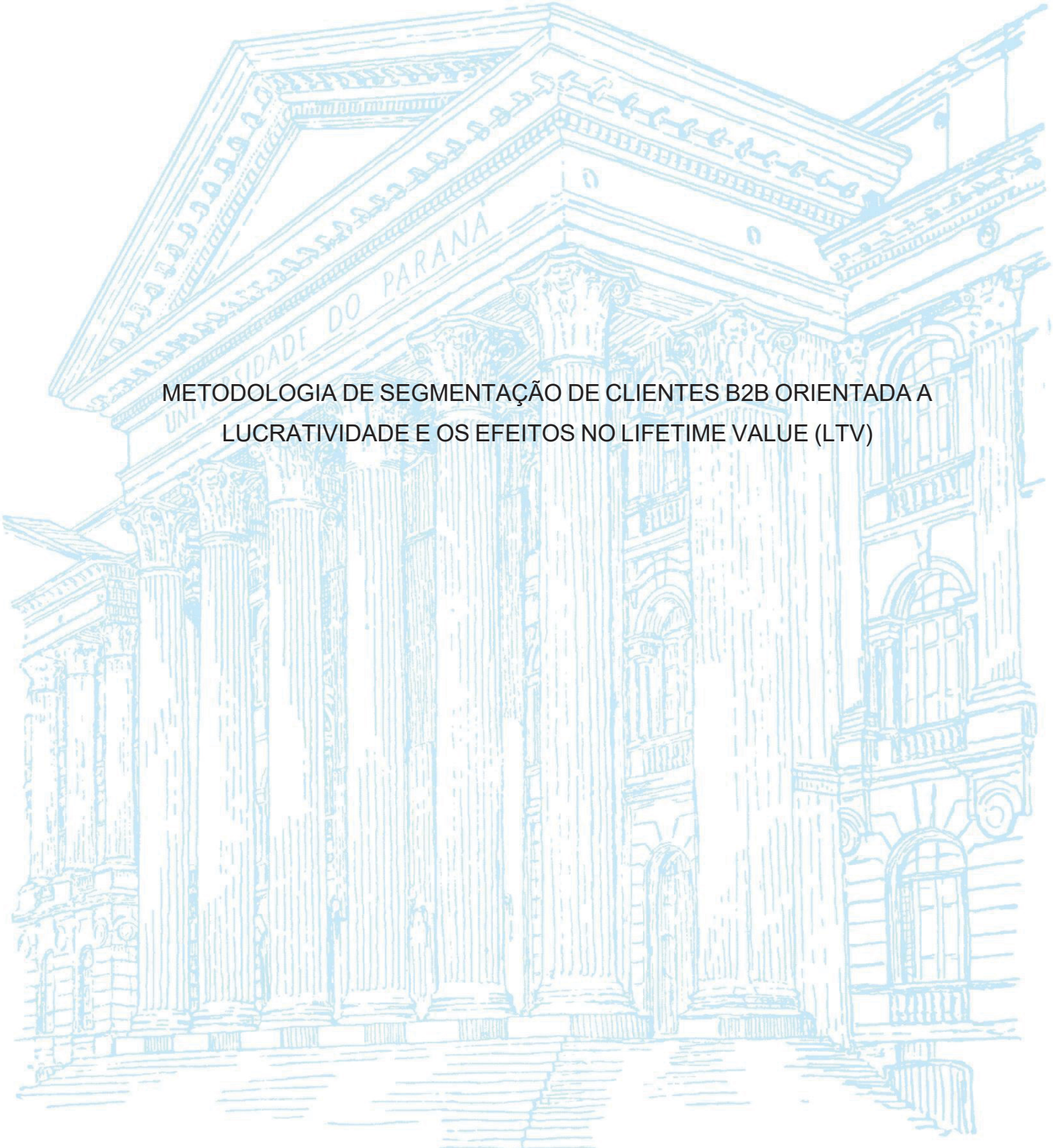
UNIVERSIDADE FEDERAL DO PARANÁ

HELLEN EUNICE DA SILVA SOMAVILLA




# METODOLOGIA DE SEGMENTAÇÃO DE CLIENTES B2B ORIENTADA A LUCRATIVIDADE E OS EFEITOS NO LIFETIME VALUE (LTV)

CURITIBA

2025

HELLEN EUNICE DA SILVA SOMAVILLA

# METODOLOGIA DE SEGMENTAÇÃO DE CLIENTES B2B ORIENTADA A LUCRATIVIDADE E OS EFEITOS NO LIFETIME VALUE (LTV)

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Gestão de Organizações, Liderança e Decisão “PPGOLD”, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Mestre.

Orientador: Cassius Tadeu Scarpin

CURITIBA

2025

DADOS INTERNACIONAIS DE CATALOGAÇÃO NA PUBLICAÇÃO (CIP)
UNIVERSIDADE FEDERAL DO PARANÁ
SISTEMA DE BIBLIOTECAS – BIBLIOTECA DE CIÊNCIAS SOCIAIS APLICADAS

Somavilla, Hellen Eunice da Silva
Metodologia de segmentação de clientes B2B orientada a lucratividade e os efeitos no *lifetime value* (LTV)/ Hellen Eunice da Silva Somavilla. – 2025.
1 recurso on-line: PDF.

Dissertação (mestrado) - Universidade Federal do Paraná, Setor de Ciências Sociais Aplicadas, Programa de Pós-Graduação em Gestão de Organizações, Liderança e Decisão.
Orientador: Cassius Tadeu Scarpin.

1. Marketing. 2. *B2B marketing (Business to business marketing)*. 3. Clientes – fidelização. 4. Lucros. I. Scarpin, Cassius Tadeu. II. Universidade Federal do Paraná. Setor de Ciências Sociais Aplicadas. Programa de Pós-Graduação em Gestão de Organizações, Liderança e Decisão III. Título.

Bibliotecária Lívia Rezende Ladeia – CRB – 9/2199

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIÊNCIAS SOCIAIS E APLICADAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO GESTÃO DE ORGANIZAÇÕES, LIDERANÇA E DECISÃO - 40001016172P9

# TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação GESTÃO DE ORGANIZAÇÕES, LIDERANÇA E DECISÃO da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **HELLEN EUNICE DA SILVA SOMAVILLA**, intitulada: **METODOLOGIA DE SEGMENTAÇÃO DE CLIENTES B2B ORIENTADA A LUCRATIVIDADE E OS EFEITOS NO LIFETIME VALUE (LTV)**, sob orientação do Prof. Dr. CASSIUS TADEU SCARPIN, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.

A outorga do título de mestra está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 16 de Dezembro de 2025.

Assinatura Eletrônica
17/12/2025 09:54:55.0
CASSIUS TADEU SCARPIN
Presidente da Banca Examinadora

Assinatura Eletrônica
17/12/2025 11:02:56.0
CLAUDIMAR PEREIRA DA VEIGA
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
17/12/2025 15:12:34.0
KELLEN DAYELLE ENDLER
Avaliador Externo (PONTIFÍCIA UNIVERSIDADE CATÓLICA DO PARANÁ)

Assinatura Eletrônica
17/12/2025 12:36:53.0
DAVID GABRIEL DE BARROS FRANCO
Avaliador Externo (UNIVERSIDADE FEDERAL DO NORTE DO TOCANTINS)

Aos pilares da minha vida: minhas queridas filhas e esposo, Manoella, Martinna e Romério, cujo amor e alegria diários são minha fonte de inspiração e força; ao meu orientador professor Cassius pelo encorajamento nos momentos difíceis e aos amigos acolhedores que agiram em cada momento de dúvida e celebração. Cada um de vocês foi peça chave nessa conquista, pela importância do amor, do apoio e da amizade. Dedico esta dissertação a vocês, com toda a minha gratidão.

## AGRADECIMENTOS

Agradeço primeiramente a Deus, por me conceder força, saúde e sabedoria durante toda essa jornada acadêmica.

Me desafio constantemente a resolver problemas a partir de diferentes perspectivas, até encontrar soluções que sejam eficientes e aplicáveis. Fui formada por pais amorosos e guiados por valores que ultrapassavam a busca por riqueza material. Carrego a marca da perda e a experiência de sobreviver a uma doença que quase interrompeu meu caminho durante o mestrado, elementos que fortaleceram minha resiliência e redefiniram meu propósito.

Esta dissertação representa, assim, mais do que um trabalho acadêmico. Ela simboliza a ressignificação de vivências intensas e a reconciliação entre passado e presente, reafirmando minha identidade como alguém que persevera, aprende e transforma desafios em significado.

Dedico este trabalho às minhas filhas, cuja força e curiosidade renovam diariamente meu compromisso com o conhecimento e com o futuro. Ao meu companheiro, Romério Somavilla, sou grata pela presença constante e pela paciência nos inúmeros dias e noites acompanhados de exercícios, livros e artigos espalhados.

Manifesto meu reconhecimento à minha líder e mentora, Josiane Boing, pela confiança depositada em projetos que marcaram minha carreira, entre eles a aplicação prática desta metodologia em uma empresa de destaque no setor de tecnologia.

Agradeço as amigas Anne Cene e Leticya Mira pela parceria, sustentaram madrugadas de estudo e reflexão, tornando o percurso mais leve.

Estendo meus agradecimentos aos professores e à Universidade Federal do Paraná, por promoverem um ambiente transformador, capaz de ampliar horizontes, estimular o pensamento crítico e fomentar a realização de sonhos por meio de desafios que elevam e fortalecem.

Ao meu orientador, Professor Cassius Tadeu Scarpin, pela dedicação, orientação e suas valiosas contribuições.

Aos professores do PPGOLD, que contribuíram significativamente para minha formação, compartilhando conhecimentos que levarei para toda a vida.

Só sei que nada sei.

PLATÃO, (*Apologia de Sócrates*, 2000)

O homem não é nada além daquilo que a educação faz dele.

Kant (2004)

## RESUMO

As organizações, atuando em mercados cada vez mais limitados ou saturados, enfrentam constantemente o desafio de fidelizar seus clientes e têm um processo de venda complexo para a progressão do negócio, seja na expansão das frentes já existentes ou na criação de novas. Essas organizações buscam impedir que seus clientes existentes se tornem inativos, enquanto orientam na priorização de novas empresas a serem prospectadas, alocando recursos em seus orçamentos para sustentar o portfólio e atender a requisitos estratégicos. Esta pesquisa visa realizar um estudo analítico sobre a importância da metodologia de segmentação de clientes orientada à lucratividade e sobre seus efeitos no *Lifetime Value* (*LTV)*. Trata-se da aplicação da descoberta de conhecimento em bancos de dados (KDD - *Knowledge Discovery in Databases*). Como limitações, o estudo contemplará o cenário de empresas do segmento financeiro nacional, dependentes do fator de inadimplência e atuantes no modelo B2B (*Business to Business*), utilizando informações públicas do cadastro de pessoas jurídicas. Por meio dos resultados obtidos, verifica-se que a metodologia segmentada de clientes orientada ao LTV contribui para as inferências sobre investimentos estruturais, para a formação das metas de conversão dos negócios, para o tempo de vida dos clientes e para a definição de projetos comerciais, como expansão ou abertura de novas frentes de campanhas de marketing. Isso auxilia na compreensão dos limites das expectativas quanto ao grau de assertividade das projeções. O método, portanto, abre espaço para proposições de melhoria relacionadas às variáveis CAC (*Customer Acquisition Cost*) e *Churn*, que compõem a formulação matemática do LTV. Diante disso, este trabalho propõe uma metodologia de segmentação de clientes baseada em variáveis de lucratividade, com o uso de algoritmos de aprendizado de máquina, visando aumentar a assertividade da priorização comercial e apoiar a tomada de decisão estratégica com base em dados. Para obter resultados mais consistentes em retenção de clientes, alocação eficiente de recursos e previsão do comportamento futuro dos leads e clientes atuais, a proposta busca integrar técnicas de *clusterização*, análise fatorial e modelos preditivos, alinhadas ao processo de KDD, possibilitando inferências mais precisas e personalizadas sobre o valor de cada cliente ao longo do tempo



**Palavras-chave:** Estratégia Comercial, Inteligência de Mercado, Lucratividade e/ou Indicadores de Lucratividade, Filtros de Colaborativos, Cluster e Negócios.

## ABSTRACT

Organizations operating in increasingly constrained or saturated markets constantly face the challenge of retaining their customers and managing a complex sales process necessary for business progression, whether by expanding existing fronts or creating new ones. These organizations strive to prevent current customers from becoming inactive while prioritizing new prospects, allocating budget resources to sustain the portfolio and meet strategic requirements. This research aims to conduct an analytical study on the importance of profitability-oriented customer segmentation methodology and its effects on *Lifetime Value* (LTV*)*. It involves the application of *Knowledge Discovery in Databases* (KDD). As a limitation, the study will focus on companies in the national financial sector that are dependent on default rates and operate under the B2B (Business to Business) model, using publicly available data related to corporate registration. The results show that a segmented customer approach oriented toward LTV contributes to inferences about structural investments, setting business conversion goals, determining customer lifetime, and defining commercial projects such as expansion or the launch of new *marketing* campaigns. This helps to better understand the limitations of expectations regarding the accuracy of projections. Therefore, the method introduces propositions aimed at improving variables such as CAC and *Churn*, which are part of the mathematical formulation of LTV. In this context, the study proposes a customer segmentation methodology based on profitability variables, using machine learning algorithms to increase the accuracy of commercial prioritization and support *data-driven* strategic decision-making. To achieve better results in customer retention, efficient resource allocation, and forecasting future behavior of leads and current clients, the proposed approach integrates clustering techniques, factor analysis, and predictive models aligned with the KDD process, enabling more precise and personalized insights into each customer's lifetime value.



**Keywords:** Commercial Strategy, Market Intelligence, Profitability and/or Profitability Indicators, Collaborative Filtering, Clustering, Business.

## ÍNDICE DE FIGURAS

## LISTA DE SIGLAS

| | |
|---|---|
| B2B | *Business to Business* (Negócios para Negócios) |
| B2C | *Business to Consumer* (Negócios para Consumidores) |
| CAC | *Customer Acquisition Cost* (Custo de Aquisição de Cliente) |
| CRM | *Customer Relationship Management* (Gestão de Relacionamento com o Cliente) |
| KDD | *Knowledge Discovery in Databases* (Descoberta de Conhecimento em Bancos de Dados) |
| LTV | *Lifetime Value* (Valor do Tempo de Vida do Cliente) |
| LTR | *Lifetime Retention* (Tempo de Retenção do Cliente) |
| PCA | *Principal Component Analysis* (Análise de Componentes Principais) |
| KPI | *Key Performance Indicator* (Indicador-Chave de Desempenho) |
| ROI | *Return on Investment* (Retorno sobre Investimento) |
| ABM | *Account-Based Marketing* (*Marketing* Baseado em Contas) |
| ESG | *Environmental, Social, and Governance* (Ambiental, Social e Governança) |
| RNN | *Recurrent Neural Network* (Rede Neural Recorrente) |
| Seq2Seq | *Sequence to Sequence* (Sequência para Sequência) |
| AHP/MCDA | *Analytic Hierarchy Process/Multi-Criteria Decision Analysis* (Processo |

| | |
|---|---|
| | de Hierarquia Analítica/Análise de Decisão Multicritério) |
| CDAP | *Cross-Domain Adaptive Framework* (Estrutura Adaptativa entre Domínios) |
| CX | *Customer Experience* (Experiência do Cliente) |
| VP | Verdadeiros Positivos |
| FP | Falsos Positivos |
| VN | Verdadeiros Negativos |
| FN | Falsos Negativos |
| SSE | *Sum of Squared Errors* (Soma dos Erros Quadrados) |
| IQR | *Interquartile Range* (Intervalo Interquartil) |
| ANOVA | *Analysis of Variance* (Análise de Variância) |
| DBSCAN | *Density-Based Spatial Clustering of Applications with Noise* (Agrupamento Espacial Baseado em Densidade para Aplicações com Ruído) |

**SUMÁRIO**

# 1 INTRODUÇÃO

No atual cenário corporativo, empresas inseridas em mercados saturados enfrentam desafios significativos para manter sua base de clientes ativa e, ao mesmo tempo, expandir suas operações comerciais em um ambiente altamente competitivo. Com a intensificação da concorrência e a crescente exigência dos consumidores, torna-se cada vez mais difícil preservar o engajamento e a lealdade do público-alvo apenas com abordagens tradicionais. Nesse contexto, destaca-se a necessidade de estratégias bem estruturadas que sustentem o relacionamento com os clientes já conquistados, evitando sua inatividade e possível evasão. Para tanto, muitas organizações alocam recursos significativos em seus orçamentos, não apenas para viabilizar campanhas de fidelização, mas também para garantir o cumprimento de metas comerciais e de objetivos estratégicos. Tais metas, por sua vez, costumam ser fundamentadas em projeções de longo prazo, frequentemente amparadas em inferências heurísticas e planejamentos orientados por experiências anteriores e tendências de mercado.

É nesse ambiente desafiador que emergem metodologias e técnicas orientadas por dados, com o propósito de auxiliar gestores na identificação das necessidades reais de seus clientes e prospects. O termo "prospects", amplamente utilizado no meio empresarial, refere-se a indivíduos ou empresas que ainda não realizaram uma compra, mas que apresentam perfil compatível com os critérios do público-alvo da organização, configurando-se como potenciais clientes. No presente trabalho, o termo será mantido em sua forma original em inglês, respeitando seu uso consagrado no contexto comercial e de *marketing*. Com o auxílio dessas ferramentas, os gestores podem identificar padrões comportamentais e desenhar ofertas personalizadas que atendam simultaneamente às expectativas dos clientes e às metas de rentabilidade das empresas.

Entre as métricas que se destacam nas estratégias comerciais contemporâneas, merece destaque o *Lifetime Value* (LTV), ou valor do tempo de vida do cliente. Trata-se de uma métrica que monitora a rentabilidade de cada cliente ao longo de seu relacionamento com a empresa.

Segundo Olnén (2022), o LTV representa o montante total de receita que um cliente pode gerar, sendo especialmente útil para mensurar o sucesso das estratégias

de retenção e orientar investimentos em ações comerciais específicas. Quando bem utilizado, o LTV permite que empresas identifiquem os clientes com maior potencial de retorno, otimizando os esforços de fidelização e personalização de serviços.

Complementando essa visão, Wu et al. (2023) destacam que a aplicação do LTV tem impactos diretos na ampliação da margem de lucro, pois orienta a criação de ofertas mais assertivas, ações proativas de relacionamento e intervenções estratégicas voltadas para a retenção. Além disso, permite um gerenciamento mais inteligente de clientes de baixa geração de receita, promovendo o redirecionamento de recursos para segmentos mais rentáveis e viabilizando o planejamento de futuras oportunidades comerciais com base no valor acumulado de cada perfil de cliente.

Dessa forma, esta pesquisa propõe-se a realizar um estudo com base em dados públicos de mercado, oriundos de bases de dados amplas e abertas, comumente associadas ao conceito de Big Data. A proposta metodológica foi concebida pela autora com base em variáveis escolhidas de forma heurística e estratégica, considerando as particularidades do produto ou serviço de interesse — geralmente estruturado sob a forma de campanhas de *marketing* direcionadas ou perfis ideais de clientes. O objetivo central é demonstrar a relevância da segmentação inteligente de *leads*, priorizando estrategicamente as ações de prospecção e o gerenciamento cotidiano da área comercial, com foco na maximização da rentabilidade e na eficiência operacional.

Nesse cenário, os sistemas de recomendação ganham protagonismo como ferramentas essenciais para impulsionar as vendas e refinar as estratégias de *marketing*. Tais sistemas atuam tanto na atração de novos clientes quanto na fidelização dos já existentes, proporcionando experiências mais personalizadas e, consequentemente, mais eficazes. A filtragem colaborativa, uma das técnicas mais consolidadas nesse campo, vem sendo constantemente aprimorada por meio da integração com diversas abordagens analíticas, ampliando sua capacidade de gerar recomendações relevantes e contextualizadas.

Entre essas abordagens, destaca-se a proposta deste estudo, que consiste na integração entre o processo de Descoberta de Conhecimento em Bancos de Dados (*Knowledge Discovery in Databases* – KDD) e os sistemas de recomendação colaborativos. O KDD, ao explorar grandes volumes de dados e extrair padrões relevantes, permite a obtenção de insghits mais profundos e personalizadas, elevando

significativamente a qualidade das decisões comerciais baseadas em dados. Essa sinergia entre sistemas inteligentes e mineração de dados representa um avanço importante na busca por estratégias comerciais mais embasadas e preditivas.

Conforme Fayyad et al. (1996), o KDD desempenha papel estratégico ao transformar dados brutos em conhecimento aplicável, permitindo que decisões importantes sejam tomadas com base em informações robustas, estruturadas e alinhadas aos objetivos organizacionais. Os autores reforçam que o uso do KDD em sistemas de apoio à decisão comercial não apenas aumenta a eficiência analítica, mas também fortalece a capacidade das empresas de responder de forma ágil e fundamentada às dinâmicas do mercado.

## 1.1 DELIMITAÇÃO DO TEMA

O processo de tomada de decisão do planejamento comercial baseia-se em duas etapas fundamentais e interdependentes: uma tática e outra estratégica. Na etapa tática, predomina uma abordagem analítica e racional, baseada em dados concretos, indicadores de desempenho e cálculos numéricos que sustentam decisões objetivas. Essa fase é orientada por métricas quantificáveis como faturamento, margem de contribuição, taxa de conversão, entre outras. No entanto, embora a racionalidade seja o eixo principal, não se descarta a presença de inferências pontuais derivadas de situações excepcionais, como alterações políticas internas, sazonalidades específicas ou ocorrências de outliers — isto é, registros que fogem ao padrão estatístico, mas que podem sinalizar oportunidades ou ameaças relevantes ao planejamento. Tais exceções, embora menos frequentes, são consideradas por sua capacidade de alterar os rumos táticos, mesmo quando não previstas pelos modelos matemáticos tradicionais.

Na etapa estratégica, por sua vez, o foco se desloca para uma visão mais holística e de longo prazo. Aqui, o papel dos gestores de alto escalão se torna mais proeminente, pois são eles que, baseando-se em sua vivência, conhecimento acumulado do setor e leitura do ambiente externo, contribuem com interpretações e julgamentos subjetivos. As ações, derivadas de experiências anteriores ou da sensibilidade diante de sinais do mercado, são fundamentais para orientar decisões

que extrapolam a objetividade dos números, permitindo um direcionamento mais robusto das metas comerciais e dos investimentos futuros. Essa combinação entre análise empírica e intuição estratégica busca alinhar a empresa às transformações do mercado, promovendo um crescimento sustentável e planejado.

Desta forma, a presente pesquisa está relacionada, principalmente, à etapa tática. Visa a gerar uma metodologia que define quais os *leads* e clientes, do universo mapeado previamente, necessitam de priorização de ações de relacionamento ou prospecção. Em decorrência da metodologia proposta, a ser apresentada ao longo do trabalho, a aplicação de um método de segmentação de clientes, com a adoção de variáveis de lucratividade para alavancagem da estratégia comercial torna-se uma possibilidade real e prática para aumentar a produtividade da área comercial.

Previsões fundamentadas em indicadores de lucratividade, como o LTV, não apenas ajudam o planejamento financeiro da empresa, mas também contribuem para melhores decisões de *marketing* e orientam o gerenciamento de relacionamento com o cliente através do sistema de *Customer Relationship Management* (CRM) (WANG et al., 2019).

## 1.2 OBJETIVOS

### 1.2.1 Objetivo geral

Desenvolver uma metodologia de segmentação de clientes que permita, de forma sistematizada e baseada em dados, identificar características específicas e recorrentes de cada grupo, a partir de variáveis comerciais relevantes. A proposta visa estruturar a definição de segmentos com base em critérios quantitativos e qualitativos, considerando aspectos operacionais e financeiros dos clientes. Além disso, objetiva-se incorporar à metodologia variáveis diretamente relacionadas à lucratividade, como o Faturamento e o Custo de Aquisição de Clientes ou *Customer Acquisition Cost* (CAC), a fim de gerar inferências consistentes e aplicáveis ao indicador do LTV. A abordagem busca oferecer suporte técnico e estratégico à operação comercial, permitindo decisões mais assertivas sobre prospecção, retenção e priorização de

contas no ambiente B2B, com foco no aumento da rentabilidade e na otimização do relacionamento com os clientes ao longo do tempo.

### 1.2.2 Objetivos específicos

a) Identificar, a partir da literatura especializada, os principais critérios utilizados na segmentação de *leads* e clientes no ambiente B2B, considerando tanto abordagens tradicionais quanto modelos contemporâneos de *marketing* orientado por dados, de forma a compreender como diferentes variáveis (demográficas, comportamentais, financeiras e relacionais) influenciam a categorização de perfis de empresas e tomadores de decisão.

b) Analisar o conceito do LTV e sua utilização como métrica orientadora em estratégias de segmentação de clientes, com ênfase em sua aplicabilidade prática para estimar o potencial de receita futura, orientar investimentos comerciais, priorizar contas estratégicas e subsidiar decisões sobre retenção, upsell e alocação de recursos.

c) Investigar modelos teóricos de pontuação de *leads* com base em variáveis relacionadas à rentabilidade e ao ciclo de vida do cliente, buscando compreender como sistemas de classificação podem apoiar a definição de prioridades de prospecção e engajamento em contextos empresariais com orçamentos limitados e metas de alta conversão.

d) Revisar os fundamentos do processo de descoberta de conhecimento em bases de dados e suas aplicações na organização e interpretação de dados, analisando cada etapa do fluxo — desde a seleção, pré-processamento e mineração até a avaliação e visualização de padrões — com vistas à geração de ações sustentáveis para uso estratégico no ambiente comercial.

e) Examinar, à luz de estudos existentes, a influência de percepções gerenciais e experiências de mercado na tomada de decisão estratégica em contextos comerciais, compreendendo de que maneira fatores subjetivos e heurísticos interagem com os dados quantitativos para formar estratégias híbridas, que conciliam análise baseada em evidências com a intuição executiva.

## 1.3 JUSTIFICATIVA

A presente dissertação justifica-se pelo interesse em aprofundar a discussão acadêmica sobre metodologias de segmentação de clientes no contexto B2B, com base em métricas de lucratividade, como o LTV. Em mercados cada vez mais competitivos e orientados por dados, a capacidade de identificar os clientes mais valiosos e direcionar esforços de forma estratégica tornou-se um diferencial crítico para empresas que buscam maximizar o retorno sobre seus investimentos comerciais. Nesse sentido, a segmentação orientada por valor tem sido amplamente destacada em estudos recentes como uma abordagem eficaz para subsidiar decisões tanto estratégicas quanto operacionais, especialmente em setores que demandam racionalização de recursos e gestão otimizada de carteiras de clientes (WU et al., 2023; WANG et al., 2019).

A análise da literatura especializada evidencia que as práticas de segmentação com apoio de algoritmos e modelos preditivos vêm se consolidando como importantes ferramentas de suporte técnico à tomada de decisão. Em particular, a integração desses métodos ao KDD amplia significativamente a capacidade de transformar grandes volumes de dados brutos em informações relevantes e acionáveis. Essa abordagem, conforme discutido por Fayyad et al. (1996), Han, Kamber e Pei (2011), permite a construção de modelos analíticos robustos, com potencial para revelar padrões ocultos no comportamento de clientes e apoiar estratégias comerciais baseadas em evidências.

Além disso, a relevância da presente pesquisa também se justifica pela necessidade de compreender, em profundidade, como métricas como o CAC e o *churn* (taxa de evasão de clientes) impactam diretamente na modelagem do LTV, influenciando a priorização de ações comerciais e a alocação eficiente de recursos. Tais indicadores, quando utilizados de forma integrada à segmentação de clientes, permitem não apenas projetar o valor futuro das contas existentes, mas também identificar os perfis que representam maior risco ou menor retorno, otimizando a performance da área de vendas e relacionamento.

Nesse contexto, esta dissertação propõe a estruturação teórica de uma metodologia que considere tanto dados objetivos (quantitativos, típicos da etapa tática

do planejamento) quanto percepções subjetivas e gerenciais (qualitativas, típicas da etapa estratégica), conforme referenciado por autores como Kanchanapoom e Chongwatpol (2022). A proposta busca, assim, refletir a realidade híbrida da gestão comercial, que combina métricas precisas com a experiência acumulada dos gestores no trato com o mercado.

Trata-se, portanto, de um estudo de caráter exploratório, baseado em fontes secundárias e fundamentado em uma ampla revisão de literatura científica nacional e internacional. Ao abordar a segmentação de clientes orientada à rentabilidade, no escopo do modelo B2B, espera-se que esta pesquisa contribua de forma relevante para o avanço do debate metodológico na área de *marketing* analítico e inteligência comercial, fornecendo subsídios para práticas mais eficientes, sustentáveis e alinhadas às exigências do mercado contemporâneo.

## 1.4 CONTRIBUIÇÕES

O trabalho propõe uma metodologia sistematizada para segmentação de clientes no ambiente B2B. Essa abordagem permite agrupar clientes com características de rentabilidade similares, algo que vai além da segmentação demográfica ou do comportamento tradicional. Trata-se de um avanço metodológico relevante, pois combina dados objetivos com técnicas de ciência de dados, aprimorando a precisão da tomada de decisão comercial.

O estudo aplica algoritmos de aprendizado de máquina, como o KMeans para clusterização e uma sequência práticas estatísticas como correlação e PCA. Essa aplicação prática evidencia como ferramentas estatísticas e computacionais podem ser incorporadas ao planejamento tático comercial, tornando-o mais assertivo, baseado em evidências e dados reais de mercado e de relacionamento.

Ao incorporar o LTV como eixo central da metodologia de segmentação, a pesquisa amplia o uso estratégico dessa métrica, tradicionalmente subutilizada. O LTV é aqui relacionado não só ao potencial de receita, mas também à definição de metas, ao direcionamento de recursos e à criação de campanhas de *marketing* mais eficazes, ajudando as empresas a priorizar clientes de maior valor futuro.

A pesquisa oferece uma ferramenta concreta de apoio à priorização de clientes e *leads*, o que é altamente relevante para empresas que operam no modelo B2B, em que o ciclo de vendas costuma ser longo e os custos de aquisição são elevados. A proposta metodológica ajuda a identificar clientes estratégicos, reduzindo o risco de *churn* e promovendo maior retorno sobre o investimento comercial.

Ao demonstrar que é possível tomar decisões mais inteligentes com base em análises orientadas por dados, o trabalho incentiva empresas — especialmente aquelas de menor maturidade em ciência de dados — a adotar modelos de KDD. Isso representa uma contribuição significativa para a transformação digital e cultural nas organizações.

A autora reconhece os limites da metodologia proposta, como a dependência de dados públicos secundários e a necessidade de atualização periódica dos modelos. Isso adiciona realismo e responsabilidade científica, indicando que, embora a metodologia seja promissora, sua eficácia depende da manutenção, adaptação e capacitação contínua das equipes envolvidas.

A dissertação abre caminho para novos estudos que podem ampliar ou refinar a metodologia com uso de outras variáveis, algoritmos mais sofisticados (como redes neurais ou aprendizado supervisionado), ou aplicação em outros setores além do financeiro. Também fornece uma base para a criação de sistemas inteligentes de recomendação B2B, baseados em valor preditivo e comportamento de clientes.

## 1.5 ORGANIZAÇÃO DO TRABALHO

O trabalho foi organizado para garantir uma leitura clara e didática, guiando o leitor desde a contextualização do problema até os resultados e suas aplicações práticas. Ele está estruturado em capítulos que se complementam, oferecendo uma visão completa da pesquisa sobre segmentação de clientes com técnicas de ciência de dados. A seguir, um resumo de cada parte:

a) capítulo 1 – Introdução: apresenta o contexto, o problema de pesquisa, objetivos, justificativa, limitações e a estrutura do trabalho, com foco na importância da segmentação para empresas B2B.

b) capítulo 2 – Revisão da Literatura: reúne conceitos teóricos de *marketing*, ciência de dados e inteligência de negócios. Aborda o processo de KDD, técnicas de mineração de dados, segmentação no mercado B2B e métricas como CAC e LTV.

c) capítulo 3 – Trabalhos Correlatos: analisa estudos semelhantes, destacando metodologias e resultados, o que reforça a relevância do tema e aponta lacunas na literatura.

d) capítulo 4 – Metodologia: detalha os procedimentos práticos da pesquisa, incluindo a preparação dos dados, aplicação do algoritmo KMeans, métricas simuladas e ferramentas utilizadas, como Python e bibliotecas específicas.

e) capítulo 5 – Resultados e Discussão: apresenta e interpreta os resultados da clusterização, relacionando-os aos objetivos da pesquisa e propondo estratégias práticas para cada segmento de cliente identificado.

f) conclusão: revisa os objetivos alcançados, destaca as contribuições do estudo e propõe sugestões para pesquisas futuras, incluindo o uso de dados mais robustos e técnicas mais avançadas.

# 2 REVISÃO DA LITERATURA

Este capítulo apresenta uma revisão aprofundada dos principais conceitos, modelos e ferramentas que sustentam a proposta deste trabalho, com ênfase nas áreas de *marketing* orientado por dados, segmentação de clientes e análise preditiva. Inicialmente, são abordados os fundamentos do KDD, seguido das etapas de tratamento, transformação e mineração de dados. Em seguida, explora-se a literatura sobre segmentação de clientes, destacando as particularidades e os desafios do ambiente B2B. Por fim, são discutidas métricas de desempenho comercial como do LTV e o CAC, essenciais para a construção de modelos estratégicos baseados em lucratividade. Essa revisão tem por objetivo sustentar teoricamente a metodologia proposta e posicionar a pesquisa no estado da arte das práticas contemporâneas em inteligência de mercado.

## 2.1 PROCESSO DE *KNOWLEDGE DISCOVERY IN DATABASE (KDD)*

A descoberta de conhecimento em bases de dados, conhecida pelo termo em inglês *Knowledge Discovery in Databases (KDD)*, é um processo sistemático e interdisciplinar voltado à extração de informações úteis e de conhecimento relevante a partir de grandes volumes de dados. Esse processo compreende uma série de etapas interligadas, que vão desde a seleção e pré-processamento dos dados até a mineração propriamente dita e a posterior interpretação dos padrões extraídos. Conforme apontam Han et al. (2011), as fases iniciais do KDD incluem a limpeza, integração, seleção e transformação dos dados, que antecedem a aplicação de algoritmos de mineração voltados à identificação de padrões significativos.

O crescimento exponencial na geração e armazenamento de dados em diversas áreas do conhecimento tem ampliado significativamente a relevância do KDD. A capacidade de transformar dados brutos em conhecimento estratégico torna-se essencial para a tomada de decisões orientadas por dados. Nesse cenário, destaca-se a necessidade de métodos eficazes para lidar com a complexidade, o volume e a variabilidade das informações disponíveis. A precisão e a qualidade dos dados tornam-se, assim, elementos centrais para o sucesso do processo de

descoberta. Dados incompletos, inconsistentes ou irrelevantes podem comprometer diretamente os resultados obtidos, levando a interpretações errôneas ou a descobertas ineficazes (HAN et al., 2011).

Dessa forma, a preparação dos dados é considerada uma etapa crítica no processo de KDD. Essa preparação envolve atividades como a limpeza de inconsistências, a normalização para uniformizar os formatos e a transformação dos dados em estruturas adequadas para análise. Segundo Han et al. (2011), a eficácia dos algoritmos de mineração de dados está intrinsecamente ligada à qualidade dos dados que recebem como entrada. Portanto, uma preparação meticulosa contribui significativamente para garantir que os padrões extraídos sejam confiáveis, coerentes e, sobretudo, úteis no contexto de aplicação.

Um exemplo prático da aplicação bem-sucedida das técnicas de KDD encontra-se no estudo conduzido por Ekstrand et al. (2010), que aborda sistemas de recomendação baseados em filtragem colaborativa. Nesse estudo, os autores demonstram como a análise de grandes volumes de dados sobre o comportamento e as preferências dos usuários pode ser utilizada para gerar sugestões personalizadas em plataformas interativas. Essa abordagem não apenas melhora a experiência do usuário, mas também otimiza a eficácia dos sistemas de recomendação, ressaltando o valor do KDD na personalização e na relevância das informações apresentadas.

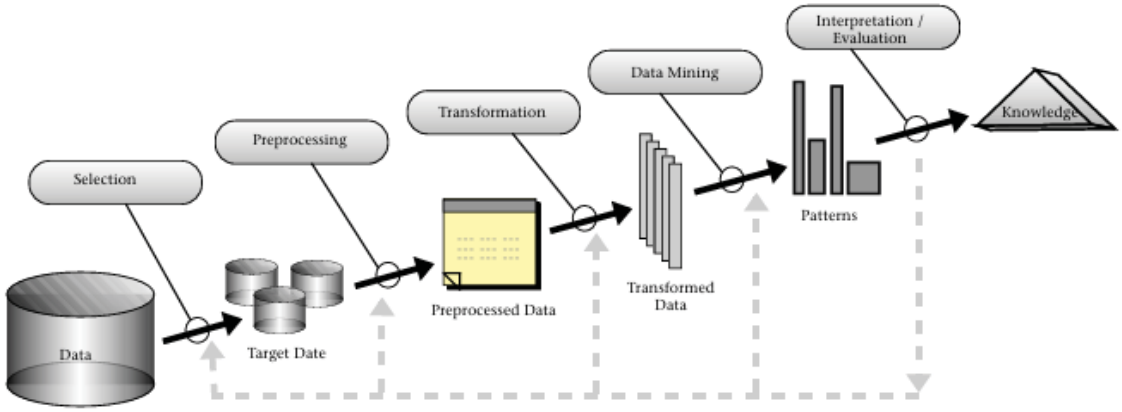
Complementando essa perspectiva, Fayyad et al. (1996) definem o KDD como um campo interdisciplinar cuja finalidade é extrair conhecimento útil a partir de grandes conjuntos de dados. Os autores descrevem o processo como composto por várias etapas fundamentais, incluindo a seleção, a limpeza, o enriquecimento e a transformação dos dados, seguidas da aplicação de algoritmos de mineração para a identificação de padrões relevantes. Eles ainda destacam que a importância crescente do KDD está diretamente relacionada ao avanço da tecnologia e à consequente ampliação do volume de dados disponíveisl para análise, o que impõe a necessidade de métodos analíticos robustos e eficientes.

A compreensão detalhada do processo de KDD revela não apenas a complexidade técnica envolvida na manipulação e análise de grandes volumes de dados, mas também a necessidade de uma abordagem sistemática e bem estruturada. Conforme ilustrado na figura 01, o KDD é um processo interativo,

composto por diversas etapas interdependentes que se iniciam pela seleção dos dados e se estendem até a descoberta e validação de padrões. Cada uma dessas etapas contribui para refinar e preparar os dados, aumentando progressivamente sua qualidade e potencial analítico.

A figura 01 não apenas delimita as fases do KDD, mas também evidencia a conexão dinâmica entre elas, sugerindo que o processo de descoberta de conhecimento é cíclico e adaptável. Isso implica que, a cada iteração, os dados podem ser reavaliados e ajustados com base nos resultados anteriores, promovendo uma melhoria contínua na qualidade da análise. Tal abordagem é indispensável em contextos em que a precisão e a relevância das informações extraídas são determinantes para o sucesso de projetos analíticos.

**Figura 1 - Uma visão geral das etapas que compõem o processo KDD**




Fonte: Fayyad et. al (1996)

Diante do crescente volume e da diversidade de dados disponíveis, a aplicação do KDD torna-se uma estratégia essencial para organizações que desejam transformar seus dados em ações. Por meio de uma análise cuidadosa e estruturada, é possível converter grandes quantidades de dados brutos em conhecimento significativo, capaz de embasar decisões mais informadas, estratégicas e alinhadas aos objetivos organizacionais. Nesse sentido, a figura 01 cumpre um papel duplo: além de representar visualmente as fases do processo de KDD, atua também como um guia conceitual que evidencia a importância da interdependência e do rigor em cada uma das etapas envolvidas na descoberta de conhecimento em bases de dados.

### 2.1.1 Estruturação dos dados

A etapa de tratamento de dados no processo de descoberta de conhecimento em bases de dados (KDD) representa uma fase crítica para assegurar a confiabilidade, integridade e usabilidade dos dados a serem utilizados nas etapas subsequentes de análise. Esta fase visa preparar os dados de modo a permitir que os algoritmos de mineração operem com máxima eficiência e precisão. Para isso, torna-se necessário realizar uma série de procedimentos, como a identificação e correção de ruídos, a resolução de inconsistências e a normalização dos dados.

Entre as tarefas mais comuns do tratamento de dados destacam-se a padronização de formatos, o preenchimento ou remoção de valores ausentes, a detecção e tratamento de outliers, bem como a transformação de variáveis categóricas em representações numéricas adequadas para os modelos analíticos. Esses procedimentos são fundamentais não apenas para garantir a qualidade dos dados, mas também para aumentar a acurácia dos modelos, reduzir o tempo de processamento computacional e evitar distorções nos resultados que possam comprometer a interpretação e aplicação prática dos padrões descobertos.

Nesse contexto, Cheng e Chen (2009) enfatizam que o tratamento e o pré-processamento dos dados são fatores determinantes para o desempenho de algoritmos de agrupamento, especialmente em aplicações voltadas a sistemas de CRM. Segundo os autores, a eficácia desses algoritmos depende diretamente da qualidade dos dados de entrada, uma vez que a presença de ruídos, valores extremos ou variáveis mal representadas pode comprometer a formação de clusters coesos e semanticamente relevantes. Assim, o sucesso da segmentação de clientes e, por consequência, das estratégias de *marketing* orientadas por dados, está intrinsecamente ligado à minuciosidade do pré-processamento realizado.

Portanto, a etapa de tratamento de dados não deve ser encarada como uma simples etapa preparatória, mas como uma fase estratégica que influencia diretamente a qualidade do conhecimento extraído e sua utilidade na tomada de decisão. A negligência nesta etapa pode comprometer todo o processo de KDD, enquanto a execução cuidadosa contribui para gerar resultados mais robustos, interpretáveis e acionáveis.

#### 2.1.1.1 Eliminação de ruídos e inconsistências

Remover o excesso de informações, ruídos e inconsistências representa uma etapa fundamental na preparação dos dados no KDD, uma vez que dados imprecisos ou de baixa qualidade podem comprometer significativamente os resultados obtidos nas etapas subsequentes de mineração e análise. A presença de valores duplicados, erros de entrada, lacunas ou informações incoerentes tende a distorcer os padrões e correlações descobertos, levando a decisões equivocadas e interpretações falhas. Além disso, quanto maior o volume de dados e mais diversas as fontes envolvidas, maior a complexidade dos problemas de qualidade que podem surgir, o que exige metodologias mais robustas para sua resolução.

De acordo com Han, Kamber e Pei (2011), o pré-processamento dos dados é composto por diversas técnicas, incluindo o preenchimento de valores ausentes, a suavização de ruídos, a correção de inconsistências, a detecção e remoção de *outliers*, além da padronização e transformação de atributos. Essas etapas são essenciais para garantir a integridade, a completude e a utilidade do conjunto de dados antes de sua exploração analítica. A negligência nessa fase pode comprometer a construção de modelos de aprendizado, gerar viés nos resultados e dificultar a replicabilidade dos experimentos analíticos, afetando diretamente a tomada de decisões.

Entre os métodos mais comuns de suavização, destacam-se os filtros de média e de mediana, que substituem valores individuais por médias ou medianas calculadas com base em seus vizinhos mais próximos, reduzindo, assim a variabilidade aleatória. Já a suavização por binning agrupa os dados em intervalos (*bins*) e ajusta os valores com base em estatísticas internas de cada intervalo, promovendo homogeneidade local. A suavização por regressão, por sua vez, ajusta uma função matemática — linear ou não linear — aos dados, permitindo a identificação e atenuação de tendências ou flutuações acentuadas. Essas técnicas são particularmente úteis em conjuntos de dados altamente variáveis, como séries temporais financeiras, dados de sensores ou registros de comportamento de usuários.

A detecção de *outliers* é outra etapa crítica, pois esses valores atípicos podem interferir negativamente nos resultados dos modelos estatísticos e algoritmos de aprendizado de máquina. Métodos estatísticos convencionais, baseados em medidas

de tendência central e dispersão — como média e desvio padrão — são frequentemente utilizados para identificar e remover esses pontos anômalos. Contudo, técnicas mais avançadas, como o algoritmo DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*), proposto por Ester et al. (1996), têm se mostrado eficazes para detectar *outliers* em grandes volumes de dados multidimensionais, pois consideram a densidade local de pontos ao invés de simples critérios globais. A combinação de métodos estatísticos e algoritmos de aprendizado não supervisionado é, em muitos casos, recomendada para garantir maior precisão na detecção desses casos extremos.

Além disso, inconsistências nos dados, muitas vezes resultantes da fusão de diferentes bases de dados, erros de digitação ou atualizações mal-conduzidas, devem ser tratadas por meio de inspeções manuais, validações cruzadas e aplicação de regras de integridade baseadas no domínio dos dados. Isso inclui a verificação de chaves primárias, da integridade referencial, dos formatos esperados e de padrões semânticos consistentes. O uso de ferramentas automatizadas de *data cleaning* também tem ganhado espaço, especialmente em contextos que envolvem grandes volumes de dados e necessidade de escalabilidade. Ferramentas como Talend, Trifacta e Apache Nifi vêm sendo amplamente adotadas para automatizar processos de limpeza, enriquecimento e integração de dados em pipelines modernos de engenharia de dados.

Complementando esse panorama, Do e Batzoglou (2008), em estudos voltados para bioinformática, destacam a importância da normalização e padronização como elementos fundamentais para garantir a comparabilidade entre conjuntos de dados heterogêneos. Essas técnicas ajustam os dados para uma escala comum, evitando que atributos de magnitudes distintas dominem o processo de mineração. A normalização é especialmente importante em algoritmos baseados em distância, como *K-Means*, redes neurais e *Support Vector Machines* (SVM), nos quais variáveis em escalas distintas podem afetar desproporcionalmente a formação de padrões e decisões de agrupamento.

No campo dos sistemas de recomendação, Ekstrand, Riedl e Konstan (2010) demonstram que abordagens sofisticadas de filtragem de dados — incluindo técnicas baseadas em conteúdo e colaborativas — dependem fortemente da qualidade e da coerência dos dados de entrada. Uma base mal preparada pode gerar

recomendações irrelevantes ou enviesadas, prejudicando a experiência do usuário e reduzindo a eficácia dos sistemas inteligentes. Nesses cenários, a etapa de tratamento e pré-processamento influencia diretamente a confiabilidade dos sistemas e sua aceitação por parte dos usuários finais.

Assim, a manipulação criteriosa e sistemática das informações torna-se essencial não apenas para garantir a acurácia dos modelos e interpretações derivadas da mineração de dados, mas também para assegurar a reprodutibilidade dos resultados e sua aplicabilidade em contextos reais. A qualidade dos dados está diretamente relacionada à capacidade da organização de extrair conhecimento confiável, relevante e acionável, contribuindo de forma decisiva para a orientação estratégica e a geração de vantagem competitiva sustentada. O investimento em boas práticas de preparação de dados deve ser encarado como parte fundamental da cultura analítica das empresas, impactando diretamente os resultados de curto e longo prazo.

Em síntese, a preparação adequada dos dados não é uma etapa acessória, mas sim uma condição *sine qua non* para o sucesso de qualquer projeto baseado em análise de dados. No contexto corporativo, essa etapa representa a base sobre a qual serão construídas as estratégias analíticas e preditivas, justificando o investimento em processos e ferramentas que assegurem a excelência na gestão da informação. A negligência nesse estágio pode comprometer toda a cadeia de valor analítico, enquanto sua execução cuidadosa abre caminho para decisões mais eficazes, processos mais eficientes e inovação orientada por dados.

#### 2.1.1.2 Correção de valores ausentes

O correto tratamento de dados ausentes é uma etapa essencial no processo de preparação de dados, pois assegura a qualidade, consistência e integridade do conjunto de dados. Dados faltantes, se não tratados adequadamente, podem comprometer a validade das análises estatísticas, distorcer resultados e, consequentemente, impactar negativamente a eficácia dos modelos de mineração de dados. Assim, é fundamental adotar abordagens criteriosas e fundamentadas para a identificação, análise e tratamento dessas lacunas, garantindo que as inferências

obtidas sejam confiáveis e representativas. A negligência nesse aspecto compromete não apenas a robustez dos modelos, mas também a confiança nas decisões baseadas em dados, o que é especialmente crítico em ambientes empresariais e científicos.

A primeira etapa crítica consiste na identificação da existência e da distribuição dos valores ausentes no banco de dados. Isso envolve a quantificação do volume de dados faltantes por variável e a localização exata de onde ocorrem essas ausências. Essa análise inicial fornece uma visão geral da extensão do problema e auxilia na definição de estratégias apropriadas de imputação ou eliminação de registros, quando necessário. Ferramentas como mapas de calor e gráficos de dispersão podem ser utilizadas para visualizar as lacunas de forma clara, permitindo um diagnóstico mais assertivo e facilitando a comunicação com stakeholders.

Posteriormente, torna-se imprescindível realizar uma análise do padrão de ocorrência dos dados ausentes. Essa análise visa determinar se os dados estão ausentes completamente ao acaso (*Missing Completely at Random* – MCAR), ausentes ao acaso (*Missing at Random* – MAR) ou ausentes de forma não aleatória (*Not Missing at Random* – NMAR). A identificação desse padrão é decisiva para a seleção da técnica de tratamento mais eficaz. Por exemplo, se os dados estão ausentes de forma sistemática, isso pode refletir vieses no processo de coleta, falhas de instrumentação, ou ainda uma relação estrutural com outras variáveis do conjunto de dados. Compreender o mecanismo de ausência permite reduzir o risco de interpretações equivocadas, além de melhorar a precisão dos modelos preditivos.

Riedl e Konstan (2011), ao analisarem os impactos dos valores ausentes em sistemas de recomendação, ressaltam que a ausência de dados pode afetar diretamente a capacidade de personalização desses sistemas. Os autores destacam que dados incompletos reduzem a acurácia das recomendações, além de comprometer a robustez e a equidade dos algoritmos, especialmente em abordagens colaborativas, que dependem fortemente da completude das interações entre usuários e itens. Isso evidencia como o tratamento adequado de dados ausentes não é apenas uma questão técnica, mas uma necessidade funcional para garantir a performance e confiabilidade de sistemas inteligentes.

Entre as estratégias mais empregadas para lidar com dados ausentes, destacam-se:

a) Imputação por média, mediana ou moda: utilizada em situações de baixa complexidade, onde os valores ausentes são substituídos por estatísticas simples de tendência central. É uma abordagem eficiente em bases com pequenas proporções de ausência e pouca variabilidade.

b) Imputação por regressão: quando há uma relação identificável entre a variável ausente e outras variáveis do conjunto, é possível estimar os valores ausentes com base em modelos de regressão linear ou múltipla, mantendo maior coerência estatística entre os atributos.

c) Técnicas de aprendizado de máquina: como *k-Nearest Neighbors* (k-NN) e redes neurais, também são aplicadas para imputar valores com base em padrões complexos de similaridade ou aprendizado supervisionado, sendo úteis para bases com estrutura multidimensional e interdependência entre variáveis.

d) Eliminação de registros ou variáveis: adotada quando a quantidade de dados ausentes é suficientemente pequena para não comprometer a integridade do conjunto, ou quando a variável não possui relevância significativa para o objetivo do modelo. Trata-se de uma solução prática, mas deve ser aplicada com cautela para evitar perda de informação relevante.

e) Modelos múltiplos de imputação: como o *Multiple Imputation by Chained Equations* (MICE), considerados mais sofisticados, permitem a geração de múltiplos conjuntos imputados, incorporando a variabilidade e incerteza associadas ao processo. Essa abordagem melhora a validade estatística das análises posteriores, especialmente em estudos inferenciais.

A escolha da técnica de tratamento mais adequada deve considerar não apenas a proporção de dados faltantes, mas também o contexto analítico, a estrutura das variáveis e o impacto potencial sobre os resultados. Um tratamento inadequado pode introduzir viés, mascarar relações reais ou gerar interpretações enganosas. Por isso, é importante realizar testes comparativos entre métodos de imputação e avaliar os efeitos em métricas de desempenho dos modelos subsequentes.

Ademais, o tratamento de dados ausentes deve ser documentado de forma transparente, para garantir a rastreabilidade e reprodutibilidade dos resultados. Em ambientes corporativos, onde decisões estratégicas são tomadas com base em

análises preditivas, negligenciar essa etapa pode acarretar prejuízos operacionais e financeiros significativos. A adoção de uma política de governança de dados, com protocolos claros para tratamento de lacunas, torna-se um diferencial competitivo e de conformidade.

Portanto, a gestão criteriosa de dados faltantes é uma prática indispensável no ciclo de vida da ciência de dados. Quando bem executada, ela assegura a fidelidade das análises, potencializa a acurácia dos modelos e contribui para a geração de conhecimento de alto valor agregado, alinhado às metas organizacionais e à realidade dos negócios. Trata-se de uma etapa que, embora muitas vezes invisível aos olhos do usuário final, sustenta toda a credibilidade e aplicabilidade dos resultados analíticos, sendo essencial para o sucesso de qualquer projeto orientado por dados.

#### 2.1.1.3 Normalização dos dados

A normalização dos dados constitui uma etapa essencial no processo de preparação de dados, assegurando que os atributos estejam expressos em escalas compatíveis e adequadas para posterior análise estatística e modelagem computacional. Este procedimento visa padronizar os valores dos atributos numéricos, de modo que todos tenham igual influência sobre os algoritmos de mineração de dados e aprendizado de máquina. Sem essa padronização, variáveis com escalas numericamente mais amplas podem dominar o processo de análise, elevando a resultados enviesados e interpretações equivocadas.

Esse problema é especialmente crítico em algoritmos que dependem de métricas de distância, como a distância euclidiana, empregada em métodos de clusterização (ex.: *K-Means*) e de classificação (ex.: k-NN). Por exemplo, em um conjunto de dados que inclui variáveis como faturamento anual (em milhões) e número de funcionários (em dezenas), a variável de maior escala tenderá a influenciar desproporcionalmente os resultados caso não haja reescalonamento adequado. A normalização, nesse caso, garante que cada atributo contribua de forma equitativa no cálculo das distâncias, preservando a integridade analítica do modelo.

Entre as principais motivações para normalizar os dados, destacam-se:

a) A eliminação de unidades heterogêneas, que é crucial em contextos em que variáveis são expressas em unidades diferentes (por exemplo, metros, reais, porcentagens). Essa heterogeneidade, se não tratada, compromete a comparabilidade entre os atributos.

b) A melhoria da convergência de algoritmos de otimização, como o gradiente descendente, utilizado em redes neurais e regressão logística, que tende a alcançar soluções ótimas mais rapidamente quando os dados estão em escalas semelhantes.

c) O balanceamento entre variáveis, garantindo que nenhuma variável domine o modelo apenas por apresentar valores numéricos mais elevados.

A ausência da normalização pode impactar diretamente a eficácia dos modelos, levando a previsões imprecisas, instabilidade nos parâmetros estatísticos e dificuldade na extração de conhecimento útil, o que compromete a qualidade das decisões estratégicas baseadas em dados.

No contexto de modelos de gestão de receita, por exemplo, Cao et al. (2023) reforçam a importância da normalização como etapa fundamental para a aplicação correta de modelos logísticos multinominais, amplamente utilizados para prever comportamentos de compra. A normalização, segundo os autores, não apenas facilita a modelagem e interpretação, como também aumenta a robustez e a precisão das estimativas, especialmente em bases de dados heterogêneas e de alta variabilidade.

Conforme discutido por Han et al. (2011), várias técnicas podem ser aplicadas para normalizar os dados, sendo escolhidas de acordo com as características específicas da base de dados e os objetivos da análise. Entre as abordagens mais comuns, destacam-se:

a) Escalonamento min-max: redimensiona os valores para um intervalo pré-definido, geralmente [0, 1]. É apropriado quando os dados não apresentam valores extremos significativos, pois outliers podem distorcer o resultado do reescalonamento.

b) Padronização Z-score: transforma os dados para que tenham média zero e desvio padrão um, tornando-os compatíveis com algoritmos que assumem distribuição normal dos dados. É uma técnica amplamente utilizada quando se espera simetria estatística.

c) Normalização pelo máximo absoluto: Reescala os dados com base no valor absoluto máximo, de modo que todos os valores estejam entre -1 e 1. Essa abordagem é útil para dados esparsos, frequentemente encontrados em aplicações de aprendizado profundo e processamento de linguagem natural.

d) Escalonamento robusto (*Robust Scaler*): Baseado na mediana e no intervalo interquartil (IQR), essa técnica é especialmente eficaz na presença de outliers, pois é menos sensível a valores extremos. É indicada para bases de dados reais, onde a presença de anomalias é comum e a robustez estatística é desejável.

Cada uma dessas técnicas apresenta vantagens e limitações específicas, e a escolha apropriada depende tanto das propriedades estatísticas da base de dados quanto do modelo analítico a ser utilizado. Ignorar essa etapa pode não apenas reduzir a performance computacional do modelo, mas também comprometer seriamente a qualidade e confiabilidade das inferências realizadas a partir dos dados.

Em síntese, a normalização é uma etapa estratégica e indispensável no ciclo de vida da ciência de dados. Ao assegurar uma base de dados homogênea, balanceada e escalonada, ela melhora o desempenho dos modelos analíticos, evita distorções nas análises e fortalece a precisão das previsões e a interpretação dos resultados obtidos. Dessa forma, contribui diretamente para o sucesso das iniciativas de mineração de dados, descoberta de conhecimento e tomada de decisão baseada em evidências.

#### 2.1.1.4 Outliers (Boxplot)

Os valores atípicos, também conhecidos como outliers, são observações que se desviam significativamente do padrão geral de um conjunto de dados. Esses valores extremos podem surgir por diferentes razões, como erros de mensuração, falhas na entrada de dados, flutuações experimentais ou, em muitos casos, características genuínas e relevantes que refletem fenômenos incomuns ou exceções significativas (WASSERMAN, 2020). Embora os outliers possam representar ruídos que distorcem a análise, também podem fornecer informações importantes quando contextualizados corretamente. Sua presença, portanto, deve ser avaliada com

cautela, considerando não apenas a natureza estatística da anomalia, mas também sua possível relevância no domínio de aplicação.

A presença de outliers tem impacto direto na qualidade das análises estatísticas e na confiabilidade dos modelos preditivos. Eles podem influenciar métricas de tendência central (como a média) e de dispersão (como o desvio padrão), enviesar modelos de regressão e comprometer o desempenho de algoritmos de aprendizado de máquina. Em particular, Cao et al. (2023) demonstram que a remoção ou o tratamento adequado de outliers é essencial em modelos de previsão de demanda e gestão de receita que combinam estruturas de demanda independentes com modelos *logit multinomial*. O estudo destaca como a presença de outliers pode levar à subestimação ou à superestimação da demanda, impactando negativamente a eficiência das decisões estratégicas, como a precificação e a alocaçãos de recursos. Isso mostra que ignorar a existência de valores extremos pode acarretar consequências graves na prática, especialmente em contextos em que a acurácia dos dados é vital para decisões operacionais.

Para a detecção de outliers, uma das ferramentas gráficas mais eficientes e amplamente utilizadas é o boxplot (ou gráfico de caixa e bigodes). Esse gráfico oferece uma representação visual da distribuição dos dados com base em medidas-resumo — como os quartis, a mediana, o mínimo e o máximo — e permite a identificação objetiva de valores discrepantes. O intervalo interquartil (IQR), calculado como a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1), é usado como base para definir os limites dos chamados “bigodes” do boxplot. Valores que se encontram fora do intervalo compreendido entre Q1 – 1,5×IQR e Q3 + 1,5×IQR são considerados potenciais outliers e geralmente são destacados no gráfico como pontos individuais (KRISHNAMURTHY; DESHPANDE, 2022; ZUUR; IENO; ELPHICK, 2019). A simplicidade e clareza visual do boxplot o tornam especialmente valioso para análises exploratórias iniciais, permitindo decisões rápidas quanto à necessidade de intervenções mais aprofundadas nos dados.

A análise gráfica por meio de boxplots é especialmente útil na fase de exploração de dados do processo de Knowledge Discovery in Databases (KDD), pois permite uma rápida identificação de anomalias antes da aplicação de técnicas mais robustas de modelagem. Em contextos de negócios, saúde, engenharia ou ciências sociais, os outliers podem representar tanto riscos analíticos quanto oportunidades de

descoberta, dependendo da forma como são interpretados. Em alguns casos, esses valores extremos podem indicar mudanças importantes no comportamento dos dados, revelando tendências emergentes ou eventos de alto impacto que merecem atenção especial.

Adicionalmente, métodos estatísticos como o teste de *Grubbs*, o *z-score* padronizado e técnicas de clusterização (como *DBSCAN*) também são amplamente utilizados para detectar outliers em grandes volumes de dados. O *DBSCAN*, por exemplo, identifica pontos que não pertencem a regiões de alta densidade, sendo eficaz na detecção de outliers em conjuntos de dados multidimensionais. Com o avanço da ciência de dados, têm-se adotado abordagens híbridas que combinam estatística clássica com algoritmos de aprendizado de máquina para classificar, ponderar e até mesmo corrigir ou imputar valores discrepantes com maior grau de confiabilidade (HAN et al., 2011). Essas estratégias visam não apenas identificar, mas também integrar de forma inteligente os outliers ao processo analítico, seja por meio de exclusão justificada ou por meio da adaptação dos modelos para lidar com essas variações. Dessa forma, a gestão criteriosa de outliers torna-se uma etapa estratégica para garantir resultados analíticos mais robustos, coerentes e aplicáveis em diferentes domínios.

#### 2.1.1.4.1 Estrutura e componentes de um boxplot

O boxplot, também conhecido como gráfico de caixa e bigodes (box-and-whisker plot), é uma ferramenta gráfica amplamente utilizada na estatística exploratória para representar de forma sintética a distribuição de um conjunto de dados. Sua principal função é apresentar visualmente cinco medidas-resumo fundamentais: valor mínimo, primeiro quartil (Q1), mediana (Q2), terceiro quartil (Q3) e valor máximo, permitindo observar a dispersão, a simetria e a presença de possíveis valores atípicos (outliers) em um conjunto de dados (PAGANO; GAUVREAU, 2018).

Essa técnica oferece uma visão clara da variabilidade dos dados e da densidade em torno dos quartis, sendo particularmente eficaz na comparação entre distribuições distintas ou na identificação de assimetrias e desvios. A seguir, detalham-se os principais elementos constituintes do boxplot:

a) Mediana (Q2): representada por uma linha horizontal localizada dentro da caixa, a mediana corresponde ao segundo quartil, ou seja, o ponto que separa os 50% inferiores dos 50% superiores dos dados. É uma medida robusta de tendência central, menos sensível a valores extremos do que a média aritmética (FIELD; MILES; FIELD, 2017).

b) Caixa (Q1 a Q3): a estrutura retangular do gráfico compreende a faixa entre o primeiro quartil (Q1, 25% dos dados) e o terceiro quartil (Q3, 75% dos dados). Esse intervalo, denominado amplitude interquartil (IQR – *Interquartile Range*), representa os 50% centrais dos dados, excluindo os extremos inferiores e superiores. A IQR é uma medida fundamental de dispersão, eficaz para caracterizar a variabilidade dos dados sem a influência de outliers (WASSERMAN, 2020; WICKHAM; GROLEMUND, 2017).

c) Bigodes (Whiskers): os bigodes se estendem a partir das extremidades da caixa até os limites inferiores e superiores, definidos como 1,5 vezes a IQR abaixo de Q1 e acima de Q3. Valores dentro desses limites são considerados "normais" na distribuição dos dados (KRISHNAMURTHY; DESHPANDE, 2022). A extensão dos bigodes ajuda a visualizar a cauda da distribuição e identificar a assimetria.

d) Valores atípicos (Outliers): dados que se encontram fora dos limites dos bigodes são identificados como outliers e geralmente são representados por círculos, asteriscos ou outros símbolos. Esses valores podem indicar erros de medição, registros incorretos ou fenômenos reais fora da distribuição esperada, e merecem investigação especial, pois podem influenciar de forma significativa análises estatísticas e decisões baseadas em dados (ZUUR; IENO; ELPHICK, 2019).

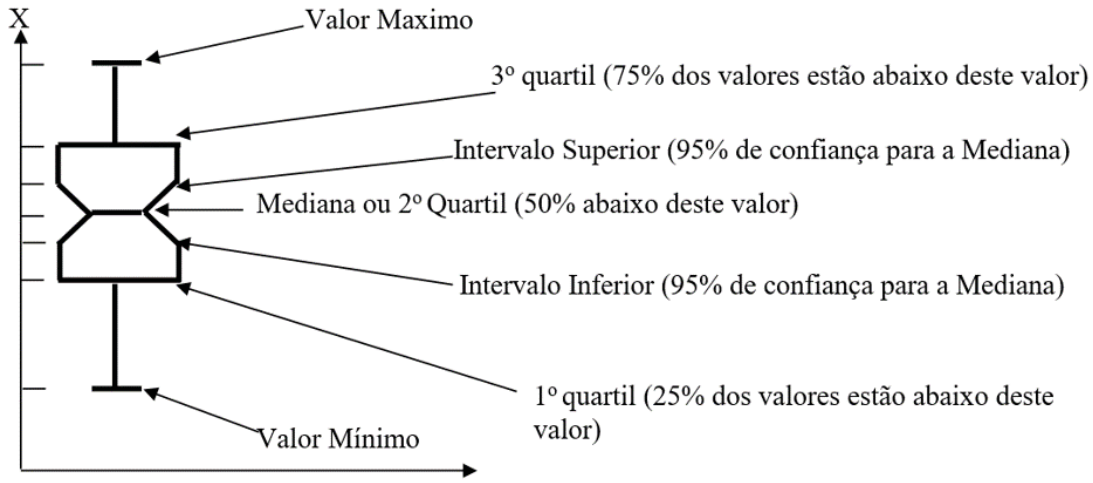
e) Valores mínimos e máximos (dentro dos limites): Os extremos inferiores e superiores que ainda se encontram dentro dos limites definidos pelos bigodes representam os menores e maiores valores considerados regulares na distribuição. Eles delimitam a "cauda" do conjunto de dados sem incluir os pontos considerados atípicos (KABACOFF, 2021).

Além de sua simplicidade visual, o boxplot é especialmente útil em contextos comparativos, como quando se deseja analisar diferentes grupos ou categorias de

uma variável. Ao permitir a visualização simultânea de mediana, dispersão e simetria, o gráfico de caixa e bigodes torna-se uma ferramenta indispensável em análises exploratórias, diagnósticos estatísticos e em aplicações que envolvem a limpeza e validação de dados.

A figura 2 demonstra como as informações são visualmente dispostas:

**Figura 2 – Boxplot**




Fonte: Adaptado pelo autor, 2025.

Os boxplots são ferramentas eficazes para a identificação de outliers, análise de distribuição e comparação entre grupos de dados. Neles, os outliers são facilmente identificados como pontos fora dos bigodes, que facilita a visualização de anomalias. Além disso, o boxplot permite uma rápida visualização da distribuição dos dados, evidenciando a presença de simetria ou assimetria. Ao comparar boxplots de diferentes grupos, é possível identificar diferenças significativas na distribuição e na presença de outliers entre os grupos.

Ao usar boxplots, é possível visualizar claramente como os outliers podem influenciar a distribuição dos dados. Por exemplo, a presença de outliers pode distorcer a média, puxando-a para cima ou para baixo. Outliers também aumentam a variabilidade aparente dos dados, refletida pelo comprimento dos bigodes do boxplot. Identificar e tratar outliers pode levar a decisões mais informadas e precisas, melhorando a qualidade das análises e previsões.

### 2.1.2 Transformação de dados

A transformação de dados compreende um conjunto de processos essenciais destinados a converter dados brutos em formatos adequados às exigências das etapas subsequentes de análise e modelagem. Entre as técnicas mais utilizadas destacam-se a normalização, a padronização, a discretização, a codificação de variáveis categóricas e a transformação logarítmica, dentre outras.

Essas metodologias são cruciais para garantir a compatibilidade dos dados com os algoritmos de mineração e de aprendizado de máquina, promovendo ganhos expressivos em deesempenho computacional, robustez e acurácia dos modelos resultantes.

Particularmente, quando os atributos apresentam escalas ou unidades distintas, a aplicação correta da transformação torna-se imprescindível, visto que muitos algoritmos baseados em medidas de distância ou gradientes são sensíveis à magnitude dos valores. Assim, a transformação de dados configura-se como uma etapa crítica no ciclo de vida da análise, prevenindo vieses e facilitando a convergência e generalização dos modelos.

Paralelamente, a análise multivariada representa um campo estatístico dedicado à observação e à interpretação simultânea de múltiplas variáveis dependentes, possibilitando uma compreensão mais holística e realista de fenômenos complexos caracterizados por inter-relações e dependências mútuas. Amplamente empregada em áreas diversas, como estudos ambientais, geográficos e socioeconômicos, essa abordagem estatística visa não necessariamente soluções otimizadas isoladamente, mas sim a construção de representações interpretáveis e coerentes dos sistemas analisados (Nijkamp, 1999). Métodos consagrados como a análise de componentes principais (PCA), a análise fatorial, a análise de agrupamentos (*cluster analysis*) e a análise discriminante figuram entre as principais ferramentas para a identificação de padrões latentes, redução dimensional e segmentação de conjuntos multivariados.

No contexto do método proposto para segmentação de clientes B2B, a análise multivariada assume papel estratégico ao viabilizar a exploração estruturada e interpretável de grandes volumes de dados heterogêneos, que englobam variáveis

como faturamento, porte empresarial, tempo de relacionamento e volume de compras. Essa abordagem permite identificar agrupamentos naturais (clusters) de clientes com perfis similares, facilitando a visualização e compreensão dos segmentos emergentes.

Além disso, a combinação entre a análise multivariada e a transformação adequada dos dados fortalece a fundamentação estatística e comercial das estratégias de segmentação, sustentando decisões baseadas em evidências concretas. Dessa forma, contribui decisivamente para a formulação de ações personalizadas, a priorização eficiente de recursos e a otimização do relacionamento com diferentes perfis dentro do portfólio de clientes.

#### 2.1.2.1 Método de *clusterização*

A análise de agrupamentos (ou *cluster analysis*) refere-se a um conjunto de técnicas estatísticas que têm como objetivo principal agrupar objetos ou observações com base em suas semelhanças e diferenças, buscando identificar estruturas naturais nos dados. A ideia central é formar grupos — chamados de clusters — de forma que os elementos pertencentes a um mesmo grupo apresentem alta similaridade entre si, enquanto os grupos diferentes sejam, idealmente, o mais distintos possível uns dos outros. Essa similaridade geralmente é medida por meio de distâncias matemáticas, como a distância euclidiana ou a de Manhattan, aplicadas sobre as variáveis disponíveis (MALHOTRA, 2006).

Diferentemente de métodos supervisionados, a análise de agrupamentos não pressupõe uma variável-alvo ou categorias pré-definidas. Ela é uma técnica descritiva, utilizada quando não se conhece, a priori, a estrutura do conjunto de dados, permitindo descobrir padrões latentes e segmentos relevantes sem interferência de suposições. Por isso, não se faz distinção entre variáveis independentes e dependentes: todas são consideradas na definição das semelhanças. Como aponta Hair et al. (2009), essa característica torna o método especialmente valioso em pesquisas exploratórias, onde o objetivo é revelar agrupamentos naturais de indivíduos, objetos ou empresas a partir de um grande volume de dados multivariados.

Esse tipo de análise mostra-se extremamente útil em contextos em que o número de observações é elevado, tornando impraticável a análise individual de cada

elemento. A partir da formação dos clusters, é possível reduzir a complexidade dos dados, facilitando tanto a visualização quanto a interpretação de tendências e comportamentos semelhantes entre os elementos agrupados. Dessa forma, a análise de agrupamentos é frequentemente empregada em áreas como *marketing*, biologia, psicologia, geografia e ciência de dados — por exemplo, na segmentação de clientes, na classificação de espécies, no agrupamento de regiões geográficas ou na redução de dimensionalidade para aprendizado de máquina.

Na análise de agrupamentos, não há conhecimento prévio sobre o número, tamanho ou as características dos grupos a serem formados. Os algoritmos assumem que os dados falarão por si, e os clusters são obtidos com base em medidas matemáticas de proximidade, sem qualquer rótulo externo. Por isso, trata-se de uma técnica não supervisionada, voltada à descoberta de padrões ocultos nos dados (HAIR et al., 2009).

O processo de clusterização pode ser dividido em duas etapas fundamentais: (1) a estimação das medidas de similaridade ou dissimilaridade entre os objetos, e (2) a aplicação de um algoritmo de agrupamento, que utilizará essas medidas para formar os grupos. Existem diversas técnicas para conduzir essa análise, e a escolha da abordagem mais adequada depende do tipo de dados, do objetivo do estudo e da quantidade de informações disponíveis.

Segundo Hair et al. (2009), as técnicas de agrupamento podem ser classificadas em dois grandes grupos:

a) Abordagem hierárquica: caracteriza-se pela construção de uma estrutura em forma de árvore (dendrograma), a partir de fusão sucessiva (aglomeração) ou divisão recursiva (divisiva) dos elementos. Inicialmente, cada observação é tratada como um grupo separado, e os grupos são combinados com base em critérios de proximidade, formando novos grupos em níveis hierárquicos até que todos estejam reunidos. O dendrograma resultante mostra visualmente as distâncias entre os agrupamentos formados, permitindo ao analista decidir, a posteriori, o número mais adequado de clusters, com base em saltos significativos nas distâncias.
b) Abordagem não hierárquica: ao contrário da hierárquica, essa abordagem exige que o número de clusters seja definido previamente pelo pesquisador. O

algoritmo mais conhecido desta categoria é o *K-Means*, que busca particionar os dados em k grupos distintos, minimizando a variância *intra-cluster* e maximizando a variância entre os clusters. A técnica é eficiente e amplamente utilizada, sobretudo em contextos com grandes volumes de dados, onde o custo computacional da abordagem hierárquica se torna inviável.

Ambas as abordagens possuem vantagens e limitações, e muitas vezes são utilizadas de forma complementar. Por exemplo, a análise hierárquica pode ser empregada inicialmente para estimar um número apropriado de clusters, que então é refinado por meio do *K-Means* ou de outro método não hierárquico. A escolha criteriosa da abordagem e da métrica de similaridade é fundamental para garantir agrupamentos coerentes e interpretáveis, especialmente em aplicações como a segmentação de clientes B2B, onde decisões estratégicas serão tomadas com base nos perfis identificados.

#### 2.1.2.1.1 *Elbow*

O método do cotovelo (*Elbow Method*) é uma técnica visual amplamente utilizada na análise de agrupamentos (clustering) para determinar o número ideal de clusters, especialmente no contexto do algoritmo *K-Means*. O principal objetivo desse método é identificar um ponto ótimo na curva que representa a relação entre o número de clusters e a qualidade da segmentação, evitando problemas comuns como o subajuste (*underfitting*), onde poucos clusters não capturam adequadamente a heterogeneidade dos dados, e o superajuste (*overfitting*), que ocorre quando clusters excessivos fragmentam desnecessariamente os grupos, prejudicando a interpretabilidade e a generalização do modelo.

O método baseia-se na análise da Soma dos Erros Quadrados (*Sum of Squared Errors* – SSE), também chamada de inércia total, que mensura a soma das distâncias quadráticas entre os pontos e os centróides de seus respectivos clusters. À medida que o número de clusters kkk aumenta, a SSE diminui de forma monotônica, pois os dados são particionados em grupos menores e mais homogêneos, o que reduz a distância média dos pontos ao centróide de cada *cluster*. Inicialmente, essa redução é acentuada, já que a divisão dos dados em poucos clusters gera grandes

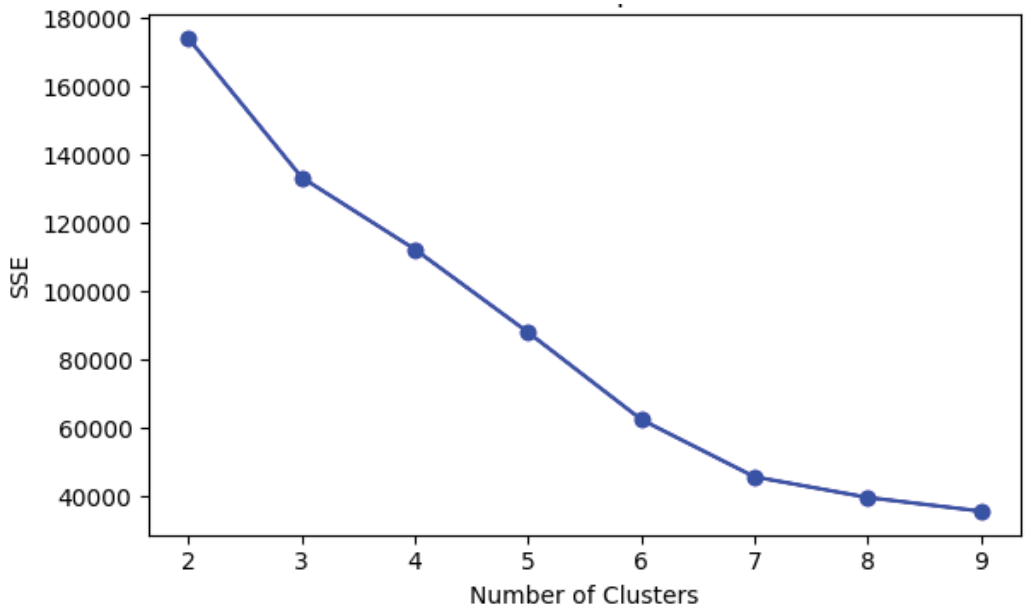
agrupamentos heterogêneos; portanto, a criação de novos clusters melhora significativamente a coesão interna.

Entretanto, após certo valor crítico de kkk, a redução da SSE torna-se menos significativa, pois os clusters já são suficientemente detalhados para representar as estruturas subjacentes dos dados. Neste estágio, o acréscimo de novos clusters produz ganhos marginais mínimos na homogeneidade, ao custo de aumentar a complexidade do modelo. O gráfico do número de clusters versus SSE, assim, forma uma curva com um formato característico semelhante a um cotovelo, cujo ponto de inflexão é interpretado como o número ideal de clusters a ser utilizado.

Na figura 3**,** observa-se a curva da SSE em função do número de clusters. Inicialmente, a SSE apresenta uma queda expressiva ao passar de 2 para 3 clusters, refletindo uma melhora substancial na coesão dos grupos. Conforme mais clusters são adicionados, a SSE continua a decrescer, porém em ritmo desacelerado. A partir de k=4k = 4k=4, nota-se uma diminuição marginal no ritmo de queda da SSE, configurando visualmente o “cotovelo” da curva — o ponto onde os ganhos em coesão são insuficientes para justificar a maior complexidade do modelo.

A identificação desse ponto é fundamental para o equilíbrio entre simplicidade e eficácia**.** Optar por um número de clusters inferior pode resultar em grupos demasiadamente heterogêneos, comprometendo a representatividade dos perfis. Por outro lado, um número excessivo de clusters pode gerar uma segmentação super fragmentada, dificultando a interpretação dos resultados e a aplicação prática das conclusões.

**Figura 3 - Exemplificação Gráfica *Elbow*.**




Fonte: Adaptado pelo autor, 2025.

Assim, com base na análise gráfica apresentada, a escolha de k=4k = 4k=4 revela-se adequada para o conjunto de dados em questão, assegurando uma segmentação representativa das estruturas latentes subjacentes e alinhada aos objetivos analíticos do estudo. Esse número promove uma divisão equilibrada, que capta a diversidade dos dados sem sacrificar a interpretabilidade e a robustez do modelo.

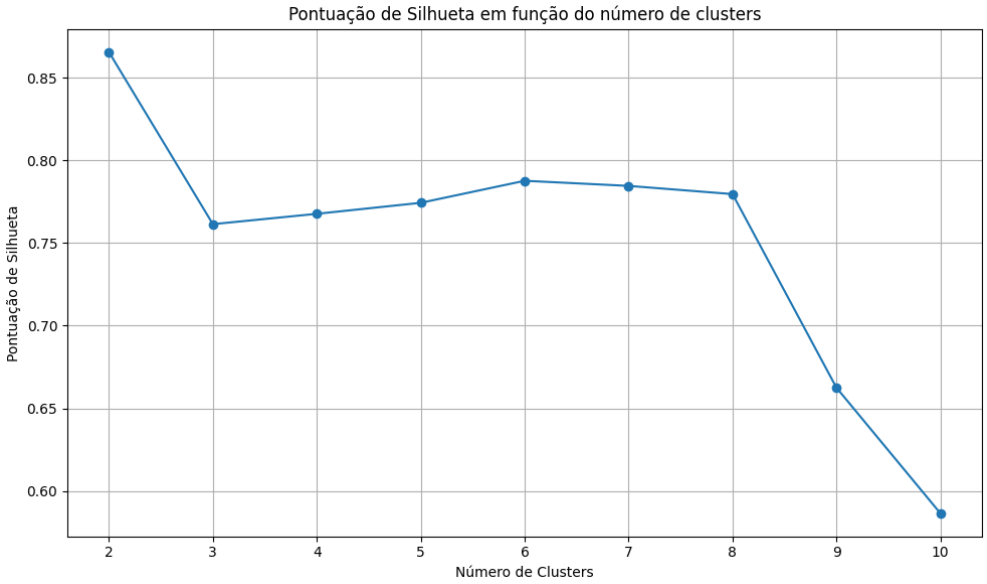
Além do aspecto visual, recomenda-se complementar a decisão do número ideal de clusters com outras métricas quantitativas, como o coeficiente de silhueta, que avalia a separação entre os grupos, com métodos estatísticos baseados em validação cruzada e na estabilidade dos clusters. Essa abordagem integrada fortalece a confiabilidade da segmentação e permite escolhas mais informadas e justificadas no processo analítico.

2.1.2.1.2 Silhouette

Para complementar a análise do número ideal de clusters, a Pontuação de Silhueta é utilizada como uma métrica que avalia a qualidade dos agrupamentos com base na coesão interna e na separação entre os grupos. Diferentemente do Método

do Cotovelo, que foca na redução do erro interno, a Silhueta oferece uma visão do quão bem definidos e distintos estão os clusters formados, auxiliando na validação da segmentação obtida.

**Figura 4 – Gráfico da Pontuação de Silhueta em função do número de clusters**




Fonte: Adaptado pelo autor, 2025.

Na figura 4, apresenta-se o gráfico da Pontuação de Silhueta em função do número de clusters, que é uma das métricas mais importantes para avaliar a qualidade dos agrupamentos obtidos por técnicas de clusterização. A pontuação de Silhueta mede a consistência interna dos clusters ao combinar a coesão dos elementos dentro de cada grupo e a separação entre os grupos distintos, possibilitando uma avaliação quantitativa da qualidade do particionamento.

O gráfico evidencia que a pontuação atinge seu valor máximo, superior a 0,85, quando o número de clusters é igual a 2. Esse resultado indica que, nessa configuração, os grupos apresentam forte coesão interna — ou seja, os elementos pertencentes ao mesmo cluster são altamente similares — e uma clara separação em relação aos elementos dos demais clusters. Tal cenário sugere que o particionamento em dois grupos fornece uma segmentação robusta e facilmente interpretável, tornando-a altamente recomendada para aplicações práticas.

Ao aumentar o número de clusters para 3 ou mais, observa-se uma queda significativa na pontuação de Silhueta, que se estabiliza em valores entre 0,76 e 0,79 até aproximadamente 8 clusters. Esta faixa indica uma qualidade moderada, refletindo que os clusters criados possuem sobreposição ou pouca distinção clara entre eles, o que pode dificultar a interpretação dos grupos e comprometer a utilidade da segmentação em contextos reais.

Quando o número de clusters ultrapassa esse ponto, especialmente a partir de 9 ou 10 grupos, a pontuação diminui de forma mais acentuada, sinalizando que a qualidade da segmentação é severamente comprometida. Esse comportamento sugere que a divisão adicional cria grupos artificiais ou muito fragmentados, que provavelmente não representam padrões reais ou úteis dentro do conjunto de dados.

Ao confrontar essa análise com os resultados obtidos pelo Método do Cotovelo (figura 03), identifica-se uma divergência metodológica significativa. Enquanto o Método do Cotovelo, baseado na minimização da soma dos erros quadrados (SSE), indica que a escolha de 4 clusters poderia ser adequada por equilibrar homogeneidade e complexidade, a métrica de Silhueta prioriza a qualidade da separação entre os grupos, apontando claramente que o agrupamento com apenas 2 clusters oferece a melhor segmentação em termos estatísticos e interpretativos.

Essa divergência não deve ser interpretada como uma contradição, mas sim como um indicativo da complexidade inerente à análise de agrupamentos, especialmente quando os dados possuem estruturas intrincadas, com sobreposição ou ausência de fronteiras claras entre grupos. Portanto, a decisão final sobre o número de clusters deve considerar múltiplas perspectivas: o embasamento em diferentes métricas quantitativas, o conhecimento de domínio do problema, os objetivos específicos da segmentação e a viabilidade prática de implementação e interpretação dos resultados.

Em resumo, a análise da Pontuação de Silhueta se mostra uma ferramenta valiosa para validar tanto visual quanto estatisticamente a qualidade dos agrupamentos. No presente estudo, essa métrica reforça a recomendação de segmentar o conjunto de dados em 2 clusters, proporcionando um modelo mais coeso, distinto e interpretável, que pode ser aplicado com maior segurança para suportar decisões estratégicas.

#### 2.1.2.2 *K-Means*

O método *K-Means* é uma técnica de análise não hierárquica amplamente utilizada em projetos de mineração de dados e ciência de dados por sua simplicidade, eficiência computacional e facilidade de interpretação dos resultados. Ao receber um número pré-definido de agrupamentos (k), o algoritmo tem como objetivo particionar os dados em k clusters distintos, de forma que cada observação pertença ao grupo cujo centroide (ponto central do cluster) esteja mais próximo. Esse processo é repetido sucessivamente até que o modelo atinja um estado estável, ou seja, até que as observações deixem de mudar de grupo entre as iterações (HAIR, 2009; HAN et al., 2011).

Inicialmente, os centroides são definidos aleatoriamente no espaço de atributos. Em seguida, cada observação é associada ao cluster mais próximo com base em uma métrica de distância – normalmente, a distância Euclidiana. Após essa etapa de alocação, os centroides de cada grupo são recalculados como a média aritmética das observações pertencentes ao respectivo cluster. Essa realocação dos centroides resulta, então, em uma nova redistribuição dos dados. O processo se repete de forma iterativa: a cada ciclo, os dados são reagrupados em torno dos centroides atualizados, e os centroides são novamente recalculados com base nas novas composições dos clusters.

Esse procedimento iterativo continua até que o algoritmo atinja um ponto de convergência, ou seja, quando a composição dos clusters deixa de variar entre as iterações consecutivas. Em algumas implementações, um critério de parada adicional pode ser estabelecido com base em um número máximo de iterações ou em um limiar mínimo de variação entre os centroides.

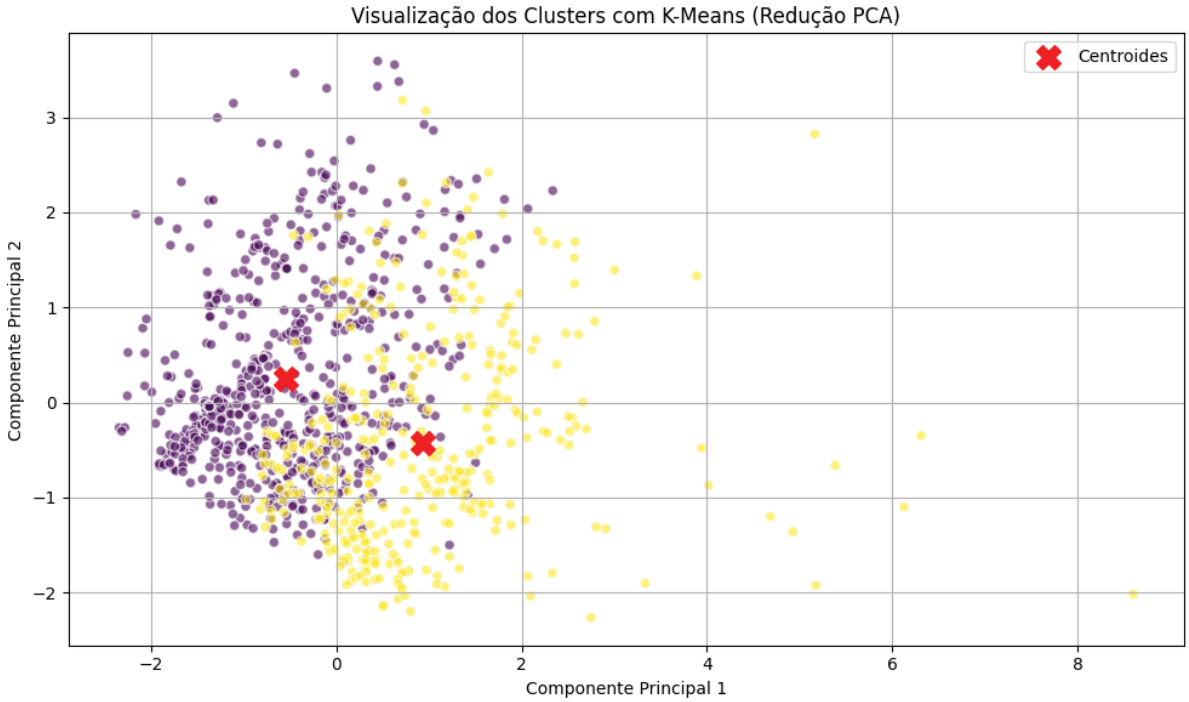
O objetivo central do *K-Means* é minimizar a variância *intra-cluster*, o que significa reduzir a soma das distâncias quadradas entre os pontos e seus respectivos centroides. Com isso, o método busca maximizar a coesão interna de cada grupo e a separação entre os diferentes *clusters*, resultando em agrupamentos mais homogêneos internamente e bem distintos entre si. Essa característica o torna

particularmente eficaz em contextos de segmentação de mercado, análise comportamental e agrupamento de padrões de consumo, entre outros.

Essa abordagem foi empregada neste trabalho como técnica principal para realizar a segmentação dos dados, proporcionando uma maneira objetiva e estatisticamente fundamentada de agrupar os clientes B2B com base em suas características multivariadas.

A figura 5 apresenta a visualização dos agrupamentos obtidos com o algoritmo *K-Means*, utilizando redução de dimensionalidade por Análise de Componentes Principais (PCA):

**Figura 5 - Visualização dos clusters gerados pelo *K-Means* com redução PCA**




Fonte: Adaptado pelo autor, 2025.

### 2.1.2.3 Análise fatorial

A análise fatorial configura-se como uma das técnicas estatísticas multivariadas mais relevantes quando o objetivo é compreender a estrutura latente de um conjunto de variáveis inter-relacionadas. Sua aplicação tem como finalidade principal reduzir a

dimensionalidade dos dados por meio da identificação de um número reduzido de fatores subjacentes – não observáveis diretamente – que, juntos, explicam a maior parte da variabilidade comum existente entre os indicadores analisados. Trata-se, portanto, de uma ferramenta que permite sintetizar informações complexas em estruturas mais manejáveis e interpretáveis, facilitando tanto a visualização quanto a compreensão dos fenômenos estudados.

De acordo com Höppner et al. (2018), a análise fatorial é amplamente empregada para revelar padrões ocultos nos dados, agrupando variáveis com alto grau de correlação em torno de fatores comuns que representam dimensões latentes de um fenômeno estudado. Esse agrupamento é particularmente útil em contextos nos quais se busca entender o comportamento de consumidores, identificar segmentos de mercado ou estudar fenômenos sociais e organizacionais em que múltiplas variáveis podem estar relacionadas a construtos teóricos mais amplos.

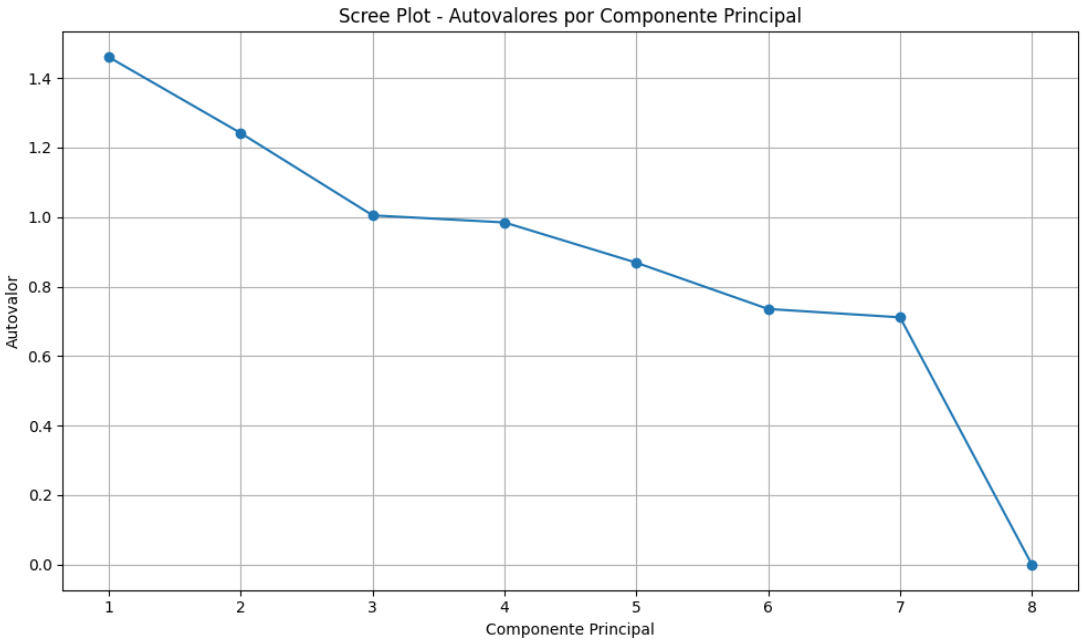
O processo metodológico inicia-se com a construção da matriz de correlação entre as variáveis observadas, a partir da qual se procede à extração dos fatores. Entre os métodos mais utilizados para essa extração destaca-se a Análise de Componentes Principais (PCA), conforme descrito por Thompson (2004), que permite decompor a variância total dos dados em componentes independentes. Cada fator extraído é associado a um autovalor (*eigenvalue*), que representa a quantidade de variância explicada por aquele fator específico. Além disso, são obtidos autovetores (*eigenvectors*) que indicam as cargas fatoriais, isto é, o grau de correlação entre cada variável observada e os fatores latentes.

Ao considerar essas cargas, é possível interpretar os fatores como combinações lineares das variáveis originais, sendo que apenas os fatores com variância significativa (geralmente com autovalor superior a 1) são mantidos para análise. Assim, determina-se o número ideal de fatores que conseguem explicar uma parcela substancial da variância total do modelo, ao mesmo tempo em que se evita a inclusão de fatores espúrios ou pouco representativos. Quanto maior o número de fatores retidos, maior será a capacidade explicativa da análise, porém menor será a simplificação do modelo – por isso, é fundamental buscar um equilíbrio entre parsimônia e poder explicativo.

Uma ferramenta complementar que contribui para essa decisão é o Scree Plot, ou gráfico de autovalores. Essa representação visual permite avaliar o ponto de inflexão da curva, indicando quantos fatores devem ser considerados relevantes. Esse ponto, geralmente associado ao critério de Kaiser (que considera autovalores maiores que 1), marca a transição entre fatores significativos e fatores com contribuição marginal para a explicação da variância.

A seguir, apresenta-se a figura 6, que ilustra o Scree Plot gerado a partir da análise dos dados do presente estudo:

**Figura 6 - S*cree Plot* – Distribuição dos Autovalores por Componente Principal**




Fonte: Adaptado pelo autor, 2025.

Na figura 6, observa-se que os dois primeiros componentes principais apresentam autovalores superiores a 1, indicando que explicam uma parcela significativa da variância total do modelo. A partir do terceiro componente, os autovalores diminuem gradativamente, evidenciando uma inclinação menos acentuada na curva. Esse comportamento sugere a presença de um ponto de inflexão entre o segundo e o terceiro fator, o que reforça a ideia de que os dois primeiros fatores são os mais relevantes para explicar os dados.

A interpretação adequada desse gráfico auxilia na seleção de um modelo mais parcimonioso, evitando tanto a subextração quanto a superextração de fatores. A escolha final do número de componentes a serem mantidos deve considerar não apenas os critérios estatísticos, como o valor dos autovalores e o percentual de variância explicada, mas também o conhecimento teórico do pesquisador e a aplicabilidade prática dos fatores no contexto do estudo. Dessa forma, o Scree Plot constitui uma ferramenta valiosa para a validação empírica da estrutura fatorial adotada.

Nesse sentido, a análise fatorial desempenha um papel estratégico na redução da complexidade dos dados, viabilizando interpretações mais robustas e direcionadas. Sua utilidade se estende a diversos campos da pesquisa acadêmica e aplicada, incluindo psicometria, *marketing*, educação, ciências sociais e comportamento do consumidor, sendo considerada uma técnica essencial no arsenal metodológico da estatística multivariada.

#### 2.1.2.4 Análise de componentes principais (PCA)

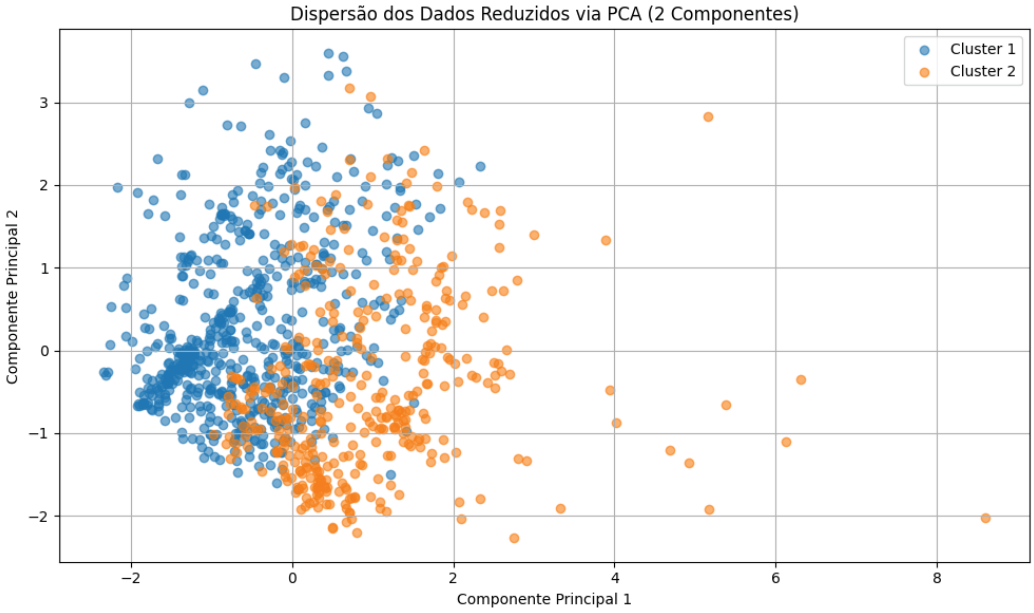
A Análise de Componentes Principais (PCA – *Principal Component Analysis*) é uma das técnicas estatísticas multivariadas mais consagradas para a redução da dimensionalidade dos dados. Sua principal finalidade consiste em transformar um conjunto possivelmente grande de variáveis inter-relacionadas em um novo conjunto, menor e composto por variáveis não correlacionadas – os chamados componentes principais. Esses componentes são combinações lineares das variáveis originais, construídas de modo a reter o máximo possível da variância total dos dados em um número mínimo de dimensões. Dessa forma, a PCA permite preservar a essência da informação contida nos dados originais, ao mesmo tempo em que elimina redundâncias e simplifica as estruturas.

De acordo com Hair et al. (2009), a aplicação da PCA inicia-se com a padronização das variáveis (quando possuem escalas diferentes), seguida pela construção da matriz de covariância entre os atributos. A partir dessa matriz, são calculados os autovalores (que indicam a quantidade de variância explicada por cada componente) e os autovetores (que definem as direções principais da variabilidade nos dados). Os componentes principais são então ordenados com base na variância

que explicam, sendo o primeiro componente aquele que representa a maior variabilidade dos dados, o segundo componente representa a maior variância residual ortogonal ao primeiro, e assim sucessivamente.

A figura 7, apresentada a seguir, ilustra o *Scree Plot* gerado a partir da decomposição PCA do conjunto de dados analisado neste estudo. Observa-se que os dois primeiros componentes explicam uma fração significativa da variância total, o que sugere que a maior parte da informação contida nas variáveis originais pode ser representada de forma eficiente em apenas duas dimensões. Este tipo de visualização é particularmente útil na definição do número ideal de componentes a serem retidos, pois destaca o ponto de inflexão (ou "joelho") onde o acréscimo de novos componentes passa a representar ganhos marginais na variância explicada.

**Figura 7 – Scree Plot dos Autovalores por Componente Principal**




Fonte: Adaptado pelo autor, 2025.

A PCA é especialmente útil em cenários com muitos variáveis que podem dificultar análises ou visualizações diretas. Ao condensar essas variáveis em poucos componentes, é possível gerar gráficos de dispersão bidimensionais ou tridimensionais que revelam padrões, agrupamentos ou outliers nos dados. Além disso, ao eliminar dimensões com variância muito baixa (frequentemente associadas

a ruídos), a PCA melhora o desempenho de modelos computacionais, como algoritmos de clusterização, classificação e regras de associação, favorecendo maior acurácia e menor sobreajuste.

A utilidade prática da PCA também é observada em estudos como o de Cumps et al. (2009), que empregaram essa técnica na etapa de pré-processamento para otimizar a indução de regras com o algoritmo AntMiner+. O objetivo era extrair regras compreensíveis sobre o alinhamento estratégico entre negócios e Tecnologias da Informação e Comunicação (TIC), a pertir de um extenso conjunto de dados provenientes de 641 organizações. Nesse contexto, a PCA foi fundamental para reduzir a complexidade dos dados sem comprometer a capacidade explicativa dos modelos gerados, evidenciando sua relevância como etapa preparatória na análise de dados em ambientes corporativos e acadêmicos.

Além de atuar como ferramenta de redução de dimensionalidade, a PCA desempenha um papel crítico na identificação de multicolinearidade entre variáveis, na priorização de atributos relevantes e na obtenção de *ações* estruturais sobre o conjunto de dados. Sua aplicabilidade é transversal a diversas áreas do conhecimento, como finanças, biologia, *marketing*, engenharia, ciência de dados e ciências sociais, consolidando-se como um recurso metodológico de grande valor em estudos quantitativos.

### 2.1.3 Data mining para classificação

Diante da crescente demanda por análise de grandes volumes de dados e da rápida evolução das tecnologias de inteligência artificial, a aplicação de técnicas analíticas avançadas tornou-se não apenas relevante, mas indispensável para a extração de pontuações estratégicas e a tomada de decisões baseadas em evidências. Em um cenário em que a complexidade e a variedade dos dados crescem exponencialmente, métodos como árvores de decisão, redes neurais artificiais e algoritmos genéticos têm se consolidado como ferramentas centrais no campo do aprendizado de máquina - *machine learning* e da mineração de dados - *data mining*.

Essas técnicas vêm revolucionando a forma como as organizações e pesquisadores tratam os dados, permitindo o reconhecimento de padrões ocultos, a

antecipação de comportamentos futuros e a automação de processos decisórios. Particularmente, os algoritmos de classificação destacam-se por sua capacidade de categorizar observações com base em características previamente identificadas, o que é essencial em contextos que envolvem diagnóstico, segmentação, previsão de *churn*, recomendação de produtos, entre outros. sendo particularmente relevantes em estudos sobre inteligência de negócios (HAN; KAMBER; PEI, 2012; TAN; STEINBACH; KUMAR, 2019).

#### 2.1.3.1 Algoritmos de classificação

Entre essas abordagens, as árvores de decisão têm como principal atrativo a sua interpretabilidade: os modelos gerados por esse método são compostos por regras simples, estruturadas de forma hierárquica, que permitem ao analista compreender o racional por trás de cada decisão ou classificação. Essa transparência é especialmente valiosa em áreas que exigem rastreabilidade e explicações claras, como o setor financeiro e a área da saúde. (BREIMAN, 2001)

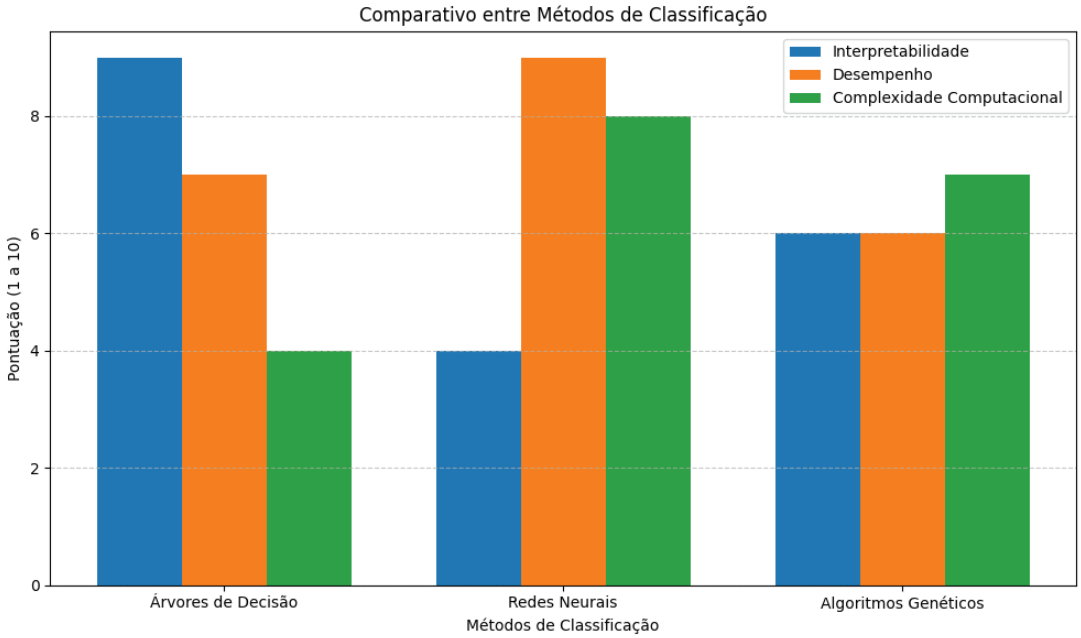
As redes neurais artificiais, por sua vez, inspiradas no funcionamento do cérebro humano, apresentam notável capacidade de capturar relações não lineares entre variáveis, sendo amplamente utilizadas em tarefas que envolvem reconhecimento de padrões complexos, classificação multi classe, predição contínua e identificação de anomalias. Embora exijam maior poder computacional e apresentem menor interpretabilidade em comparação com as árvores de decisão, seu desempenho preditivo em grandes bases de dados é frequentemente superior, especialmente quando ajustadas por meio de técnicas de regularização e otimização. (DOMINGOS, 2015)

Os algoritmos genéticos, por fim, representam uma classe de métodos inspirados nos princípios da seleção natural e da evolução biológica. Sua principal aplicação no contexto de mineração de dados está na otimização de modelos preditivos, onde são empregados para selecionar subconjuntos ideais de variáveis, ajustar hiper parâmetros e descobrir regras de classificação de alta qualidade. Um exemplo notório de aplicação é a técnica AntMiner+, um algoritmo baseado em colônia de formigas e princípios evolutivos, utilizado para a indução de regras interpretáveis e

a previsão de *churn* de clientes, combinando eficiência e inteligibilidade. (CUMPS et al., 2009)

Neste trabalho, cada uma dessas abordagens será discutida em profundidade, com foco na aplicação prática e na análise comparativa dos resultados obtidos. As árvores de decisão serão exploradas como ferramenta explicativa e interpretável para regras de classificação; as redes neurais artificiais serão implementadas como modelo preditivo de maior complexidade e poder de generalização; e os algoritmos genéticos serão utilizados como mecanismo de otimização e extração de conhecimento, com destaque para sua flexibilidade em problemas de múltiplos objetivos e espaços de busca extensos. Essa diversidade metodológica permitirá uma análise abrangente do problema proposto, considerando tanto o desempenho quanto a explicabilidade das soluções.

**Figura 8 – *C*omparação entre abordagens de classificação: interpretabilidade, desempenho preditivo e complexidade computacional.**




Fonte: Adaptado pelo autor, a partir de Breiman (2001), Domingos (2015) e Cumps et al. (2009)

A Figura 8 apresenta uma síntese conceitual derivada dos autores discutidos neste capítulo, considerando os critérios de interpretabilidade, desempenho preditivo e complexidade computacional. Trata-se de uma representação gráfica, de caráter

descritivo e comparativo, construída a partir das contribuições de Breiman (2001), Domingos (2015) e Cumps et al. (2009).

Nota-se que as árvores de decisão se destacam pela alta transparência e facilidade de explicação, enquanto as redes neurais apresentam maior desempenho em predições complexas, porém com menor explicabilidade. Já os algoritmos genéticos ocupam uma posição intermediária, oferecendo boa capacidade de otimização com interpretabilidade moderada, o que os torna atrativos em contextos híbridos. Essa visualização contribui para fundamentar a escolha metodológica conforme os objetivos específicos do estudo e as restrições do domínio de aplicação.

#### 2.1.3.2 Decision tree - árvores de decisão para regras de classificação

A técnica da árvore de decisão é amplamente utilizada no aprendizado supervisionado para tarefas de classificação e regressão, destacando-se como uma das abordagens mais intuitivas e explicáveis dentro do campo da ciência de dados. Ela opera segmentando iterativamente o espaço de atributos, criando partições que visam maximizar a pureza dos subconjuntos resultantes em relação à variável-alvo. Em outras palavras, o algoritmo constrói uma estrutura hierárquica em formato de árvore, onde cada nó interno representa uma decisão baseada em uma variável, e os nós-folha indicam a predição final (classe ou valor).

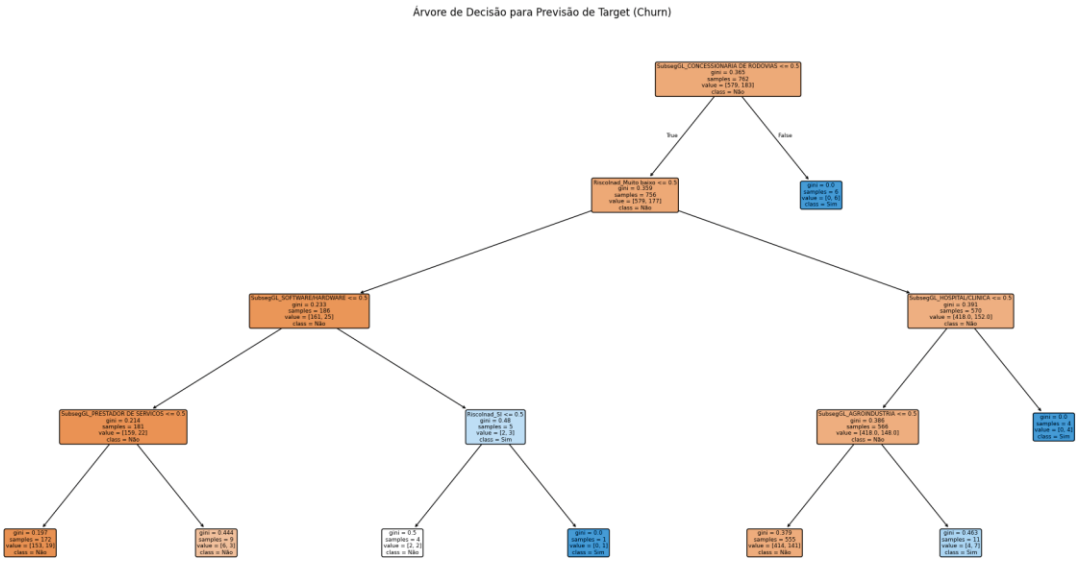
Uma das grandes vantagens das árvores de decisão é sua capacidade interpretativa, pois o modelo resultante pode ser facilmente visualizado e compreendido até mesmo por especialistas não técnicos. Essa característica é especialmente valiosa em domínios como o *marketing*, a saúde e o setor financeiro, onde a transparência na tomada de decisão é tão importante quanto a precisão dos modelos (LAROCHELLE et al., 2022).

Ao empregar o algoritmo da árvore de decisão em contextos comerciais, como na segmentação de clientes ou previsão de *churn*, torna-se viável estimar a probabilidade de um cliente adquirir ou abandonar um produto ou serviço com base em seu perfil sociodemográfico, comportamental ou histórico de consumo. A estrutura hierárquica da árvore facilita a identificação dos fatores mais relevantes para cada

decisão, fornecendo *insights* práticos e acionáveis para gestores e analistas (MOLNAR, 2022).

Segundo Cumps et al. (2009), técnicas de mineração de dados fundamentadas em árvores de decisão demonstram elevada eficiência na classificação de bases complexas e heterogêneas, permitindo a inferência de regras claras e concisas diretamente a partir dos dados coletados. Um exemplo recente da eficácia dessa abordagem é apresentado no estudo de Chen et al. (2021), publicado no *Journal of Marketing Analytics*, que utilizou árvores de decisão combinadas com análise fatorial para prever rotatividade de clientes em empresas de telecomunicações. Os autores conseguiram identificar os principais drivers de retenção de clientes, como tempo de contrato, volume de reclamações e pacotes promocionais, com elevada acurácia e interpretabilidade.

**Figura 9 – Exemplo esquemático de uma árvore de decisão simulando a classificação de clientes com base em perfil e comportamento.**




Fonte: Adaptado pelo autor, 2025.

Além disso, o uso de algoritmos de árvore de decisão, quando combinado com métodos de explicação como os valores SHAP (SHapley Additive exPlanations), conforme proposto por Lundberg et al. (2020), pode aprimorar significativamente a capacidade de interpretação dos modelos, elucidando o impacto individual de cada

variável em cada predição. Esse nível de explicabilidade é altamente valorizado em aplicações empresariais, sobretudo em contextos regulados, como o setor bancário. Como ressalta Domingos (2015), embora árvores de decisão não sejam uma solução universal para todos os problemas, sua robustez, simplicidade e eficiência computacional as tornam ferramentas de primeira escolha em muitas aplicações do mundo real.

## 2 SEGMENTAÇÃO DE CLIENTES

No contexto empresarial contemporâneo, marcado por mercados altamente competitivos, dinâmicos e saturados, as organizações enfrentam desafios cada vez mais complexos para fidelizar clientes, manter uma base de consumidores engajada e otimizar processos de venda que exigem personalização e agilidade (KOTLER e KELLER, 2006). Nesse cenário, a segmentação de clientes se revela como uma metodologia estratégica imprescindível para a formulação de ações de *marketing* mais eficazes, o desenvolvimento de campanhas direcionadas, a melhoria da retenção e, sobretudo, a maximização do valor do tempo de vida do cliente, indicador que mensura o retorno financeiro potencial que cada cliente pode gerar ao longo de sua relação com a empresa (KUMAR 2018).

Para garantir o sucesso dessa abordagem, é essencial a aplicação de técnicas analíticas e estatísticas robustas que assegurem a qualidade e a relevância dos segmentos formados. Isso inclui a avaliação criteriosa da pureza e uniformidade interna dos grupos, de modo a garantir que os clientes agrupados compartilhem características e comportamentos similares; a clara diferenciação entre os segmentos, para que cada grupo represente um perfil distinto e estrategicamente útil (MCDONALD e DUNBAR, 2012). A aplicação rigorosa de testes de hipóteses para validar estatisticamente as diferenças observadas; além da análise das correlações entre os segmentos e variáveis de negócio relevantes, como ticket médio, frequência de compra e canais de aquisição. Tais análises são fundamentais não só para validar a robustez dos agrupamentos, mas também para transformar dados em ações práticas que embasem decisões estratégicas e operacionais da organização (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996).

Este estudo aborda, de maneira sistemática, os múltiplos aspectos interligados da segmentação de clientes. Inicialmente, foca-se na avaliação da coesão interna dos segmentos, garantindo que cada grupo apresente alta homogeneidade, o que facilita a compreensão de perfis de clientes e a personalização de estratégias. Posteriormente, explora-se a diferenciação entre grupos, assegurando que os segmentos sejam suficientemente distintos para justificar ações de *marketing* diferenciadas e específicas, potencializando o impacto das campanhas e a eficácia do relacionamento com o cliente (THOMAS, 2016).

Além disso, os testes de hipóteses desempenham papel crucial ao fornecer uma base estatística para comparar segmentos, confirmando que as diferenças observadas são significativas e não fruto do acaso. Complementarmente, a análise de correlação entre segmentos e indicadores de desempenho do negócio permite identificar quais grupos apresentam maior potencial para contribuir com o crescimento e a rentabilidade da empresa, direcionando recursos para iniciativas com maior retorno sobre investimento (ROI) (FARRIS et al, 2020).

Por fim, o estudo aprofunda-se na análise dos indicadores financeiros centrais para a gestão de clientes, como o CAC e o LTV. A compreensão detalhada desses conceitos possibilita avaliar a rentabilidade e a viabilidade econômica dos segmentos, orientando a definição de estratégias que promovam crescimento sustentável e equilíbrio financeiro (BERGER e NARS, 1998). Essa visão integrada entre análise estatística, segmentação estratégica e métricas financeiras configura-se como um diferencial competitivo, permitindo às organizações não apenas responder às demandas atuais do mercado, mas também antecipar oportunidades, otimizar investimentos e fortalecer o relacionamento com diferentes perfis de clientes (PAYNE e FROW, 2017).

### 2.2.1 Pureza e uniformidade dos segmentos

A pureza e a uniformidade são critérios fundamentais para avaliar a consistência interna e a qualidade dos segmentos ou grupos formados em um conjunto de dados, especialmente no contexto da segmentação de clientes. A pureza refere-se à proporção de membros dentro de um segmento que compartilham uma

mesma característica-alvo, evidenciando a predominância de uma categoria específica — seja um perfil demográfico, comportamento de consumo ou faixa etária. Em outras palavras, um segmento puro indica que a maior parte dos elementos pertence a uma única classe bem definida, o que facilita a interpretação e aplicação prática do grupo.

Por outro lado, a uniformidade analisa a homogeneidade das características internas do segmento em múltiplas dimensões. Ela mede o grau de similaridade entre os membros do grupo considerando diferentes atributos simultaneamente, o que confere uma visão mais ampla da coesão do segmento. Uma alta uniformidade significa que os clientes dentro do grupo compartilham um conjunto de características semelhantes, reforçando a robustez e a validade do agrupamento.

A avaliação combinada da pureza e da uniformidade é essencial para verificar se os agrupamentos gerados possuem coesão interna suficiente para serem úteis do ponto de vista analítico e estratégico. Grupos homogêneos, que apresentam alta pureza e uniformidade, indicam que as necessidades e comportamentos dos clientes são mais alinhados, aumentando a probabilidade de sucesso das estratégias personalizadas, como campanhas de *marketing* direcionadas, ofertas segmentadas e planos de fidelização específicos.

Para mensurar esses critérios, ferramentas computacionais baseadas em Python são frequentemente utilizadas, com o apoio de bibliotecas como scikit-learn, numpy e pandas, que possibilitam cálculos quantitativos precisos e análises estatísticas detalhadas. No processo de avaliação, diferentes cenários são possíveis:

a) Alta pureza: quando a maioria dos membros do segmento compartilha a mesma característica dominante, indicando uma forte coesão interna e maior previsibilidade do comportamento dos clientes, o que facilita a formulação de ações específicas e eficazes.
b) Baixa pureza: quando o grupo apresenta uma mistura significativa de características distintas, revelando uma ligação interna fraca e alta heterogeneidade, o que pode sugerir a necessidade de revisitar os critérios de segmentação ou realizar um reagrupamento para aprimorar a consistência dos grupos.

A uniformidade pode ser medida por meio da Entropia, um conceito extraído da Teoria da Informação, que quantifica o grau de incerteza ou desordem em um conjunto de dados. Na análise dos segmentos, a entropia assume um papel crucial:

a) Baixa entropia: indica alta uniformidade, significando que os membros do grupo são bastante semelhantes, com pouca variação interna, o que fortalece a utilidade do segmento para ações direcionadas e específicas.
b) Alta entropia: reflete grande diversidade dentro do segmento, sugerindo baixa uniformidade e dificultando a aplicação de estratégias padronizadas, pois o comportamento dos membros é mais disperso.

Em suma, segmentos considerados puros e uniformes — com alta pureza e baixa entropia — são preferíveis, pois oferecem maior clareza interpretativa e permitem o desenvolvimento de ações de *marketing* e relacionamento mais eficazes e direcionadas. Esses grupos fornecem uma base sólida para a personalização, o que é fundamental para a maximização do valor do cliente e a otimização dos recursos da organização.

Exemplificando essa abordagem, Dahana et al. (2019) investigaram a pureza e uniformidade dos segmentos baseando-se em características relacionadas ao estilo de vida e comportamento de compra, evidenciando que grupos bem definidos aprimoram significativamente a precisão das previsões do LTV. Já Verbeke et al. (2011) destacam o uso de algoritmos genéticos, como o AntMiner+, para garantir a coesão interna dos grupos, promovendo agrupamentos fundamentados em regras claras e interpretáveis, o que facilita a aplicação prática dos resultados por profissionais de *marketing* e analistas de dados.

### 2.2.2 Diferenciação entre segmento

A distinção entre os grupos é fundamental para garantir que sejam exclusivos e abrangentes ao mesmo tempo. Essa distinção clara assegura que cada grupo represente um perfil específico e não se sobreponha a outros, o que facilita a alocação eficiente de recursos e a definição de mensagens e ações customizadas. Uma clara separação entre os grupos facilita a criação de estratégias de *marketing* direcionadas, tornando-as mais eficazes e eficientes, já que permite identificar necessidades,

comportamentos e preferências de forma mais precisa. Essa diferenciação torna possível a personalização de campanhas, ofertas, canais de comunicação e até mesmo o desenvolvimento de produtos específicos para determinados segmentos.

A diferenciação entre grupos pode ser medida por meio de diferentes métricas e métodos estatísticos, que auxiliam na quantificação da distância, variação e significância entre os agrupamentos. Alguns deles são amplamente utilizados na literatura e na prática analítica:

a) Distância entre centros dos agrupamentos (Centroides)
b) Análise de variância (ANOVA)
c) Teste de hipóteses (Testes T)

A distância entre centros dos agrupamentos (Centroides) é uma maneira direta e intuitiva de avaliar a diferenciação entre grupos, calculando a distância euclidiana (ou outras métricas, como Manhattan ou Mahalanobis) entre os centroides (pontos médios) dos agrupamentos. Quanto maior for essa distância, maior será a diferenciação entre os grupos em termos das variáveis consideradas na segmentação, ou seja:

a) Alta distância: sinaliza uma grande diferenciação entre os grupos, sugerindo que os agrupamentos são bem separados no espaço de atributos e possuem características únicas, o que favorece o uso prático dos segmentos.
b) Baixa distância: indica uma baixa diferenciação, mostrando que os agrupamentos são semelhantes e podem não representar grupos distintos de forma clara, exigindo, possivelmente, revisão dos critérios de segmentação.

A Análise de Variância (ANOVA) é uma técnica estatística utilizada para comparar as médias de várias amostras (ou grupos) e verificar se pelo menos uma delas difere significativamente das outras. Esse método é essencial para validar se as diferenças observadas entre os grupos são estatisticamente significativas ou se podem ser atribuídas ao acaso. Os principais resultados da ANOVA são:

a) Um f-estatístico alto e um valor-p baixo (geralmente < 0,05) indicam que pelo menos uma média de grupo difere significativamente das outras, sugerindo distinção real entre os grupos.

b) Um f-estatístico baixo e um valor-p alto sugerem que não há evidências suficientes para afirmar que as médias são diferentes, enfraquecendo a validade da segmentação.

Os testes de hipótese (Testes T) são utilizados para comparar as médias de dois grupos e verificar se são estatisticamente diferentes entre si. Esses testes são úteis em análises par-a-par, especialmente quando o número de grupos é pequeno. Para comparações múltiplas, a ANOVA é mais apropriada, embora os testes T continuem úteis em análises pontuais. As interpretações são similares:

a) Um t-estatístico alto e um valor-p baixo apontam que as médias dos dois grupos são significativamente diferentes.

b) Um t-estatístico baixo e um valor-p alto indicam que não há diferença estatisticamente significativa entre as médias dos grupos comparados.

De acordo com estudos de Dahana et al. (2019), a diferenciação entre grupos foi feita com base em características comportamentais e psicográficas dos consumidores, resultando em segmentos que apresentaram diferenças significativas em métricas como o LTV, permitindo estratégias específicas e mais rentáveis para cada perfil. Da mesma forma, Cumps et al. (2009) empregaram técnicas baseadas em algoritmos genéticos para estruturar grupos claramente distintos, assegurando que cada agrupamento tivesse identidade própria e viabilizando a implementação de estratégias comerciais mais direcionadas, eficazes e alinhadas aos objetivos do negócio.

### 2.2.3 Matriz de confusão

A matriz de confusão é uma ferramenta essencial e amplamente utilizada para avaliar o desempenho de modelos de classificação, como os empregados para prever o LTV ou a probabilidade de *churn* (rotatividade). Ela organiza, em forma tabular, os resultados das previsões feitas pelo modelo em relação aos valores reais conhecidos, permitindo a identificação clara de acertos e erros cometidos pelo algoritmo. Essa visualização facilita o diagnóstico de problemas, como desbalanceamento de classes ou viés de previsão.

A matriz apresenta os seguintes elementos fundamentais:

a) Verdadeiros positivos (VP): instâncias positivas corretamente classificadas como positivas;

b) Falsos positivos (FP): instâncias negativas incorretamente classificadas como positivas;

c) Verdadeiros negativos (VN): instâncias negativas corretamente classificadas como negativas;

d) Falsos negativos (FN): instâncias positivas incorretamente classificadas como negativas.

A partir de uma matriz de confusão, diversas métricas de desempenho podem ser extraídas para fornecer uma análise detalhada da performance do classificador, permitindo avaliar sua eficácia sob diferentes perspectivas. As principais métricas incluem:

a) Acurácia: a proporção total de previsões corretas (VP + VN) sobre o total de amostras analisadas. Representa uma visão geral do desempenho do modelo, mas pode ser enganosa em casos de classes desbalanceadas.

1

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

b) Precisão (ou valor preditivo positivo): mede a proporção de verdadeiros positivos entre todos os casos classificados como positivos. Indica o quão confiável é o modelo quando prevê uma classe positiva.

2

$$Precisão = \frac{VP}{VP + FP}$$

c) Recall (ou sensibilidade/revocação): mede a capacidade do modelo em identificar corretamente todas as instâncias positivas reais. É especialmente importante em contextos onde a omissão de positivos é crítica, como retenção de clientes com alto LTV.

3

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

d) F1-score: combina precisão e recall em uma única métrica, calculando a média harmônica entre elas. É útil quando há necessidade de balancear ambas as métricas, especialmente em cenários com dados desbalanceados.

4

$$F1 - Score = \frac{2 * (Precisão * Recall)}{Precisão + Recall}$$

e) Especificidade: mede a capacidade do modelo de identificar corretamente as instâncias negativas, ou seja, quantos verdadeiros negativos foram corretamente detectados entre todos os casos realmente negativos.

5

$$Especificidade = \frac{VN}{VN + FP}$$

Essas métricas são especialmente valiosas em contextos empresariais, pois permitem entender o desempenho de modelos que classificam clientes com diferentes potenciais de retorno (LTV alto, médio ou baixo), bem como antecipar clientes propensos à evasão.

No estudo de Zhang et al. (2022), a matriz de confusão foi empregada como ferramenta central para avaliar a precisão das previsões do LTV, evidenciando a efetividade do método proposto na classificação correta dos clientes com maior potencial de receita. Os autores demonstraram que o uso de métricas derivadas da matriz ajudou a aprimorar os modelos por meio de ajustes finos nos parâmetros e balanceamento entre classes.

Similarmente, Verbeke et al. (2011) também utilizam a matriz de confusão para avaliar a acurácia dos modelos preditivos, especialmente na previsão da rotatividade de clientes (*churn*), ressaltando a importância dessa ferramenta na validação e no refinamento contínuo dos modelos de previsão. A análise detalhada dos erros de classificação permitiu aos autores identificar padrões ocultos e ajustar algoritmos para

melhorar a sensibilidade a classes minoritárias, que são de alta relevância estratégica para o negócio.

### 2.2.4 Testes e hipóteses

Os testes de hipóteses são amplamente utilizados para comparar grupos e determinar se as diferenças observadas nas características dos clientes possuem significância estatística. Essas análises permitem inferir, com base em dados amostrais, se as variações entre os grupos são reais ou se poderiam ter ocorrido por acaso, contribuindo para decisões mais fundamentadas e confiáveis. Ao aplicar testes de hipóteses no contexto de segmentação, evita-se a adoção de estratégias baseadas em padrões espúrios ou interpretações subjetivas, promovendo maior rigor na análise dos dados.

Entre os testes mais comuns estão o teste t de *Student*, para comparação entre dois grupos, e a ANOVA, para múltiplos grupos. Ambos são úteis para comparar médias de variáveis como frequência de compra, valor médio gasto ou engajamento digital. A interpretação dos resultados é feita com base no valor-p, que representa a probabilidade de se observar uma diferença tão extrema quanto a verificada, caso a hipótese nula (de que não há diferença) seja verdadeira. Assim:

a) Um valor-p baixo (geralmente menor que 0,05) leva à rejeição da hipótese nula, indicando que a diferença observada é estatisticamente significativa.
b) Um valor-p alto sugere que não há evidência suficiente para afirmar que os grupos diferem significativamente.

No estudo conduzido por Zhang et al. (2022), foram empregados testes de hipóteses para comparar diferentes grupos de clientes com base em atributos demográficos e comportamentais. A análise estatística confirmou que as variações entre os grupos não eram aleatórias, validando a importância dessas diferenças para a segmentação e, consequentemente, para a definição de estratégias de *marketing* mais direcionadas e personalizadas. O uso criterioso de testes permitiu identificar quais variáveis mais influenciam o valor do tempo de vida do cliente e como diferentes perfis se comportam ao longo do ciclo de relacionamento com a empresa.

Adicionalmente, Verhoeven et al. (2023) empregaram essas análises em um contexto voltado à gestão de receitas, avaliando a eficácia de diversas estratégias aplicadas a grupos distintos de clientes. Os testes de hipóteses foram essenciais para verificar quais abordagens apresentaram diferenças significativas nos resultados obtidos, evidenciando a utilidade desses testes na validação empírica das práticas adotadas na segmentação e no planejamento de campanhas personalizadas. Os resultados reforçaram que estratégias baseadas em análises estatisticamente embasadas tendem a apresentar maior retorno e previsibilidade, o que é vital para a tomada de decisões em ambientes competitivos.

### 2.2.5 Análise de correlação entre segmentos e variáveis de negócios

A análise de correlação investiga as relações estatísticas entre os grupos de clientes e indicadores empresariais relevantes, como receita, frequência de compras, ticket médio, taxa de recompra e, especialmente, o valor do tempo de vida do cliente. Esse tipo de análise é fundamental para identificar quais segmentos contribuem de forma mais significativa para o desempenho do negócio, possibilitando uma visão estratégica baseada em dados e não apenas em suposições.

A correlação é geralmente quantificada por meio de coeficientes como o coeficiente de correlação de Pearson, que mede a força e direção de uma relação linear entre duas variáveis. Valores próximos de +1 indicam correlação positiva forte, valores próximos de -1 indicam correlação negativa forte, e valores próximos de 0 indicam ausência de correlação linear significativa. Essa métrica é essencial para avaliar o impacto potencial de diferentes segmentos nas variáveis de negócio, guiando decisões como a alocação de orçamento, a personalização de ofertas e a definição de prioridades comerciais.

No estudo realizado por Zhang et al. (2022), uma análise correlacional foi conduzida com o intuito de examinar a relação entre os grupos de clientes e o LTV, revelando que determinados grupos apresentavam correlações mais expressivas com altos valores de LTV. Isso permitiu à equipe identificar quais perfis de clientes mereciam maior atenção quanto à retenção e ao investimento em *marketing*,

reforçando a importância de priorizar segmentos com maior potencial de retorno financeiro.

A importância da análise de correlação na identificação de padrões e na compreensão do impacto dos diferentes segmentos nas variáveis de negócios é amplamente reconhecida na literatura de *marketing* e gestão. Segundo Malhotra (2018), essa ferramenta estatística oferece aos gestores uma visão analítica sobre como as características dos segmentos de mercado se relacionam com indicadores de desempenho, auxiliando na tomada de decisões estratégicas baseadas em evidências concretas.

Ao permitir a identificação de relações ocultas entre os perfis de clientes e os resultados da empresa, a análise de correlação apoia o direcionamento mais preciso de recursos e esforços para os grupos de clientes mais rentáveis, engajados ou promissores. Com isso, as organizações conseguem otimizar campanhas, melhorar a alocação de orçamento e aprimorar a performance de vendas, *marketing* e atendimento, garantindo que as decisões estejam alinhadas com o comportamento real do mercado.

### 2.2.6 Interpretação dos resultados e indicadores

A interpretação dos resultados constitui uma etapa crítica dentro do ciclo de desenvolvimento de modelos analíticos e preditivos, especialmente no contexto da segmentação de clientes e previsão de métricas de valor, como o LTV. Trata-se da fase em que os resultados quantitativos obtidos ao longo do processo de modelagem são transformados em informações qualitativas, compreensíveis e aplicáveis ao contexto organizacional, com o objetivo de embasar decisões estratégicas fundamentadas em dados.

A eficácia de um modelo é comumente avaliada por meio de indicadores de desempenho, como acurácia, precisão, recall e F1-*score*, os quais oferecem diferentes perspectivas sobre a qualidade das previsões. A seleção e interpretação adequadas desses indicadores são essenciais para entender não apenas se o modelo funciona, mas como e em que situações ele apresenta melhores desempenhos.

a) A acurácia, por exemplo, mede a proporção de previsões corretas entre todas as realizadas, sendo uma métrica intuitiva e útil em contextos em que as

classes estão balanceadas. No entanto, em situações com desbalanceamento de classes, essa métrica pode mascarar o desempenho real do modelo.

b) A precisão indica a proporção de verdadeiros positivos entre todas as predições positivas feitas pelo modelo, sendo especialmente relevante em cenários em que falsos positivos devem ser minimizados, como em campanhas de retenção de clientes.

c) O recall (ou sensibilidade) mede a capacidade do modelo de identificar corretamente todos os casos positivos, o que é crucial quando o custo de perder instâncias positivas (como clientes de alto valor que estão prestes a *churnar*) é elevado.

d) A pontuação F1 combina precisão e recall em uma média harmônica, balanceando ambas as métricas em um único valor. Essa medida é particularmente valiosa em contextos de classes desbalanceadas, como frequentemente ocorre em análises de *churn*, detecção de fraudes ou segmentações com grupos de baixa representatividade.

No estudo de Dahana et al. (2019), por exemplo, a interpretação dos resultados foi conduzida com base em uma análise integrada desses indicadores. Os autores demonstraram que os modelos aplicados à segmentação de clientes e à estimativa do LTV apresentaram níveis elevados de F1-score e recall, evidenciando sua robustez e confiabilidade, especialmente no que diz respeito à capacidade de identificar clientes de alto valor potencial. Essa abordagem reforça a importância de considerar múltiplas métricas para compreender os pontos fortes e limitações do modelo de maneira abrangente.

Além da análise técnica dos resultados, a compreensão contextual dos achados é essencial para garantir que as ações que serão geradas tenham aplicabilidade prática no ambiente de negócios. Verbeke et al. (2011) argumentam que a simples obtenção de métricas estatisticamente satisfatórias não garante a utilidade dos modelos, sendo fundamental realizar uma interpretação aprofundada e orientada ao negócio. Isso inclui compreender como os segmentos identificados se relacionam com as estratégias comerciais da empresa, quais variáveis influenciam significativamente o comportamento dos clientes, e quais ações podem ser derivadas

diretamente das previsões realizadas, como campanhas direcionadas, melhorias no atendimento, ou políticas de fidelização.

Dessa forma, a etapa de interpretação atua como um elo entre a modelagem analítica e a ação gerencial, transformando resultados técnicos em conhecimento aplicado que contribui efetivamente para a melhoria dos processos decisórios e para a maximização do valor gerado pela análise de dados.

### 2.2.7 CAC e LTV

O valor do tempo de vida do cliente, é uma métrica central nas estratégias de *marketing* orientadas por dados e desempenha papel estratégico nas decisões comerciais de médio e longo prazo. No contexto da inteligência analítica aplicada à gestão B2B, o LTV permite estimar, com base em dados históricos e comportamentais, o valor econômico total que um cliente pode gerar ao longo do relacionamento com a empresa. Essa métrica transcende a simples mensuração da receita pontual, incorporando aspectos como frequência de compra, recorrência de relacionamento, potencial de indicação e engajamento comercial, o que a torna fundamental para orientar ações de retenção, fidelização e priorização de contas estratégicas.

Diante da crescente complexidade dos mercados e da alta competitividade do ambiente B2B, torna-se cada vez mais necessário compreender não apenas o perfil atual dos clientes, mas também o potencial de valor que representam ao longo do tempo. O LTV, nesse sentido, se consolida como uma variável crítica para o planejamento comercial, oferecendo suporte à definição de metas, à alocação eficiente de recursos e ao desenvolvimento de campanhas mais precisas e rentáveis. Estudos recentes (Pollak, 2021; Zhang et al., 2022; Li et al., 2022; Afiniti, 2022; Su et al., 2023) reforçam essa importância, posicionando o LTV como um dos principais indicadores para sustentar decisões em contextos voláteis e altamente dinâmicos.

Além disso, a utilização integrada do LTV com o CAC possibilita uma avaliação mais acurada da sustentabilidade financeira das estratégias adotadas. A razão LTV/CAC constitui-se em um indicador decisivo para verificar a viabilidade de campanhas e investimentos em novos clientes. Quando esse índice é superior a 1,

indica-se que o valor gerado supera o investimento realizado, contribuindo para um crescimento saudável e sustentado. Caso contrário, sinaliza-se a necessidade de revisão das abordagens comerciais, sob risco de prejuízos cumulativos à margem de contribuição.

No âmbito da segmentação de clientes, o LTV potencializa análises mais refinadas, permitindo identificar quais grupos de clientes oferecem maior retorno ao longo do tempo. Essa abordagem, quando aliada a técnicas de clusterização e modelos preditivos — como os adotados neste trabalho —, viabiliza o desenvolvimento de estratégias personalizadas de atendimento, precificação e retenção. Com isso, aumenta-se a precisão na definição de prioridades comerciais, melhora-se o direcionamento das ações de *marketing* e fortalece-se a eficiência operacional da equipe comercial.

Em síntese, o LTV não apenas contribui para a racionalização dos recursos empresariais, mas também sustenta a construção de um modelo de gestão comercial orientado à rentabilidade. Sua aplicação integrada à metodologia proposta neste trabalho reforça a importância de decisões baseadas em dados para a manutenção e expansão de relacionamentos comerciais duradouros e lucrativos no ambiente B2B.

#### 2.2.7.1 CAC: *Customer Acquisition Cost*

O CAC é um indicador-chave que mede os gastos totais com *marketing* e vendas realizados com o objetivo de conquistar novos clientes. Trata-se, portanto, de uma estimativa do investimento médio necessário para converter um lead em cliente ativo, incluindo ações diretas e indiretas que influenciam o processo de decisão do consumidor. Como destacado por Wu et al. (2023), o CAC tem papel central nas análises de desempenho comercial e sustentabilidade financeira de empresas orientadas por dados.

Na concepção de Burelli (2019), a maioria das empresas aloca uma parte significativa de sua receita nas áreas de *marketing* e vendas, com a expectativa de retorno na forma de expansão de base de clientes e aumento de receita. Nesse sentido, é crucial que as organizações realizem uma análise detalhada sobre o montante investido em canais específicos (como mídia paga, inbound *marketing*,

feiras, equipes comerciais, entre outros) e o número de clientes efetivamente captados por meio de cada um deles. Essa análise é fundamental para identificar os canais mais eficientes e lucrativos, otimizando os esforços comerciais e maximizando o retorno sobre investimento — como também é ressaltado por Pollak (2021).

Em conformidade com essa perspectiva, Afiniti (2022) destaca que a aquisição de um novo cliente frequentemente requer um investimento inicial elevado, que não se limita à comunicação e publicidade do produto ou serviço, mas também envolve custos operacionais com equipes de vendas, ferramentas de CRM, estrutura de atendimento e treinamentos. Esse esforço financeiro visa estruturar e escalar o negócio, especialmente em mercados altamente competitivos ou em fases de expansão acelerada. Como resultado, a aquisição de clientes pode representar uma das maiores despesas operacionais de uma organização, podendo, em cenários extremos, ultrapassar 50% do faturamento bruto, especialmente em startups ou empresas em estágio inicial.

Dada a materialidade do investimento em aquisição, o acompanhamento rigoroso e contínuo do CAC torna-se essencial para uma gestão orientada por indicadores. Este KPI permite que líderes de vendas, analistas de *marketing* e executivos de alto escalão, como CEOs e CFOs, tenham uma visão clara do crescimento atual do negócio e da viabilidade econômica desse crescimento no médio e longo prazo. Ele ainda possibilita identificar gargalos, desperdícios e oportunidades de melhoria nos processos comerciais e de comunicação.

Nessa perspectiva, o cálculo do CAC pode ser representado pela seguinte fórmula (6):

6

$$CAC = \frac{Cm + v}{Nc}$$

Onde:

a) CAC é o Custo de Aquisição de Cliente
b) Cm+v é o custo total de *marketing* e vendas para a aquisição de clientes (investimentos)
c) Nc é o número de novos clientes adquiridos.

Essa fórmula, embora de aplicação conceitualmente simples, exige cuidado na obtenção dos dados. O desafio prático recai sobre a atribuição precisa dos investimentos aos canais corretos, bem como a correta contabilização dos clientes originados em função desses gastos, especialmente em contextos com múltiplos pontos de contato, vendas indiretas ou ciclos longos de conversão.

É fundamental ressaltar que o CAC não inclui custos fixos de produção ou despesas administrativas, tampouco investimentos em pesquisa e desenvolvimento, suporte técnico, jurídico ou financeiro. Ele deve incluir exclusivamente os custos relacionados às áreas de vendas e *marketing*, tais como salários de equipes comerciais, mídia paga, comissões, plataformas de automação, eventos, e até mesmo custos incorridos com *leads* que não converteram em clientes, já que fazem parte do custo médio de aquisição.

Assim, para a sustentabilidade de um modelo de negócios, o custo de aquisição de clientes não pode ser superior ao valor que esse cliente gera para a organização ao longo de seu relacionamento. Como enfatizado por Li et al. (2022), a relação LTV/CAC deve idealmente ser superior a 3:1, indicando que o valor gerado por um cliente supera amplamente o custo de aquisição, garantindo rentabilidade e escalabilidade ao modelo comercial.

#### 2.2.7.2 LTV: *O valor no tempo de vida dos clientes*

O LTV, conforme já abordado anteriormente (seção 2.2.8), refere-se ao valor financeiro total que um cliente gera para a empresa ao longo de todo o seu relacionamento com a marca. Essa métrica projeta, com base em dados históricos e estimativas futuras, o montante líquido que a organização pode esperar obter de um cliente individual até o término do vínculo comercial.

Olnén (2022) complementa essa definição ao destacar que o LTV representa o lucro médio gerado pelo cliente no período analisado, já considerando os custos variáveis associados ao seu ciclo de vida, como atendimento, suporte, *marketing* de retenção e operação logística. Isso reforça a importância de tratar o LTV não apenas como um indicativo de receita, mas como uma medida direta de rentabilidade por cliente.

De forma mais precisa, o LTV pode ser definido como a receita líquida total esperada pela empresa ao longo de todo o tempo em que o cliente se mantiver ativo, ou seja, subtraídos os custos diretamente atribuíveis ao atendimento de suas necessidades. Dessa maneira, ele permite avaliar a viabilidade econômica de estratégias de aquisição, fidelização e desenvolvimento de relacionamento com diferentes segmentos de clientes.

Segundo a abordagem proposta por Zhang et al. (2022), o cálculo do LTV deve considerar, essencialmente, três fatores fundamentais:

a) Margem de contribuição: corresponde à receita anual gerada pelo cliente, descontadas as despesas operacionais diretas envolvidas em seu atendimento. Reflete o lucro líquido obtido com o cliente em cada período.

b) Taxa de retenção (retention rate): representa o percentual de clientes que permanecem ativos de um período para o outro, sendo crucial para estimar a duração média do relacionamento e, por consequência, o valor total gerado.

c) Taxa de desconto: expressa o custo de capital da empresa ou o valor do dinheiro no tempo. É aplicada para converter os fluxos de caixa futuros gerados pelo cliente em valor presente, permitindo uma avaliação realista da rentabilidade futura.

Além da estimativa do valor monetário, o tempo de vida do cliente (*Lifetime*, ou LTR – *Lifetime Retention*) também é uma variável importante. A seguir, é apresentada a fórmula para o cálculo do Lifespan (L), baseado na *churn rate*:

7

$$LTR = L = \frac{1}{C}$$

Onde:

a) L é o tempo de vida útil esperado do cliente (em períodos, como anos ou meses);

b) C é a *churn rate*, ou taxa de evasão dos clientes no período.

A fórmula da taxa de *churn* é:

8

$$C = \frac{P}{I}$$

Em que:

a) P representa o número de clientes perdidos no período;
b) I é o número de clientes ativos no início do período.

Substituindo essa expressão na fórmula do Lifespan, temos:

9

$$LTR = L = \frac{I}{\frac{P}{I}}$$

Ou seja, a fórmula final simplificada torna-se:

10

$$LTR = L = \frac{I}{P}$$

Esse cálculo fornece uma estimativa direta da longevidade média dos clientes, com base na proporção entre os clientes retidos e os perdidos. Quanto menor a taxa de *churn*, maior o tempo de vida do cliente, refletindo um relacionamento mais estável e duradouro, com maior potencial de geração de receita.

A partir da compreensão dessas fórmulas e dos conceitos de LTV e LTR, verifica-se que o cálculo dessas métricas é fundamental para entender a viabilidade do negócio, sua capacidade de gerar valor sustentável e sua eficiência na alocação de recursos em *marketing* e vendas. Elas permitem antecipar retornos, definir prioridades e orientar decisões estratégicas baseadas em dados.

Entretanto, é importante ressaltar que nenhuma métrica, quando analisada isoladamente, é capaz de oferecer uma compreensão completa do cenário de negócios. A análise conjunta do LTV, do LTR e do CAC proporcionam uma visão mais holística e acionável, permitindo avaliar o equilíbrio entre aquisição, retenção e

rentabilidade. Somente com essa perspectiva integrada é possível garantir a sustentabilidade e o crescimento saudável da base de clientes ao longo do tempo.

## 2.3 SEGMENTAÇÃO DE CLIENTES B2B

A segmentação de clientes no ambiente B2B (*Business to Business*) constitui uma prática estratégica fundamental para organizações que almejam direcionar de forma mais eficaz seus recursos de *marketing*, vendas e atendimento, sobretudo em mercados de alta competitividade e com estruturas de decisão complexas. Diferentemente do contexto B2C (*Business to Consumer*), em que a segmentação costuma se basear em critérios demográficos, psicográficos e comportamentais de consumidores individuais, o B2B apresenta desafios adicionais, exigindo abordagens multidimensionais e profundamente analíticas.

Essas abordagens precisam considerar, entre outros fatores:

a) O potencial de lucratividade de cada cliente empresarial;
b) A previsibilidade do relacionamento a longo prazo;
c) O grau de alinhamento estratégico entre as soluções ofertadas e as necessidades do cliente;
d) E indicadores quantitativos fundamentais, o LTV e o CAC, que oferecem uma visão financeira do relacionamento (KOTLER; KELLER, 2016).

A análise segmentada da base de clientes, quando orientada por dados e fundamentada em modelos analíticos preditivos ou classificatórios, permite identificar perfis empresariais com maior propensão a gerar retorno financeiro contínuo. Essa abordagem baseada em dados favorece decisões como:

a) A priorização de esforços comerciais em contas de alto valor;
b) A personalização de ofertas de produtos, preços ou serviços conforme as demandas do segmento;
c) E a reavaliação da alocação orçamentária em canais de *marketing*, prospecção e suporte, de modo a otimizar o uso dos recursos disponíveis.

O ambiente B2B é marcado por características específicas que aumentam sua complexidade:

a) Ciclos de venda mais longos e imprevisíveis;
b) Envolvimento de múltiplos tomadores de decisão (ex: áreas técnica, financeira e jurídica);
c) Negociações altamente personalizadas e por vezes consultivas;
d) Contratos de valor elevado e prazos longos;
e) Menor volume de transações, mas com maior impacto unitário na receita.

Diante desse cenário, adotar uma estratégia de segmentação robusta e baseada em valor não é apenas recomendável, mas imperativo para o sucesso organizacional e a sustentabilidade das ações comerciais no médio e longo prazo. De acordo com Kumar (2018), empresas que adotam práticas de segmentação baseadas no valor do cliente apresentam resultados superiores em rentabilidade e fidelização, além de reduzirem significativamente os custos com aquisição e retenção — reflexo direto da maior assertividade nas ações.

Ademais, a segmentação possibilita a personalização das comunicações, produtos, serviços e propostas de valor, adaptando-os às necessidades, dores e objetivos específicos de cada grupo ou vertical de clientes. Essa customização orientada por dados não apenas melhora a experiência do cliente (*Customer Experience* – CX), como também impulsiona os índices de retenção e reduz a taxa de evasão (*churn*).

Pollak (2021) demonstra que a eficácia de ações de *marketing* personalizadas pode ser ampliada em até 30% quando são apoiadas por modelos de segmentação baseados em dados históricos e comportamento preditivo, destacando o papel da inteligência comercial na formulação de estratégias centradas no cliente.

Assim, a segmentação no B2B transcende o papel de agrupamento estático de contas e se posiciona como uma ferramenta dinâmica de gestão estratégica, capaz de transformar dados em conhecimento e conhecimento em vantagem competitiva sustentável.

### 2.3.1 Critérios relevantes para segmentação B2B

A segmentação de clientes no ambiente B2B (*Business to Business*) exige uma abordagem criteriosa e multifacetada, dada a complexidade e especificidade das relações comerciais entre empresas. A literatura especializada aponta diversos critérios que podem ser empregados nesse processo, cuja escolha está intrinsecamente ligada aos objetivos estratégicos do negócio, à natureza do produto ou serviço oferecido, e ao grau de maturidade analítica da organização. De forma geral, esses critérios podem ser organizados em três grandes categorias principais: financeiros, comportamentais e estratégicos.

Critérios Financeiros abrangem aspectos ligados ao desempenho econômico e capacidade financeira dos clientes corporativos. Exemplos incluem o faturamento anual, a margem de lucro, o tamanho da empresa (quantidade de colaboradores ou capital investido), o volume de compras e o histórico de pagamentos. Esses indicadores são essenciais para entender o potencial de investimento e o valor comercial de cada cliente, além de auxiliar na priorização de esforços e recursos para os segmentos com maior retorno esperado.

Critérios Comportamentais focam nas interações e nos padrões observados ao longo da jornada do cliente, incluindo frequência e volume de compras, lealdade à marca, canais de compra preferidos, tempo de relacionamento com a empresa e respostas a campanhas de *marketing*. Esses fatores fornecem ações importantes sobre o comportamento real dos clientes, permitindo identificar segmentos com diferentes níveis de engajamento, propensão à recompra e abertura para ofertas personalizadas.

Critérios Estratégicos envolvem características que refletem a importância e o alinhamento do cliente com os objetivos de longo prazo da empresa. Entre eles, destacam-se o grau de influência no mercado, o potencial para parcerias estratégicas, a sinergia tecnológica, o perfil de inovação e a maturidade digital. Esses critérios ajudam a segmentar clientes não apenas pelo valor imediato, mas também pelo papel que desempenham no ecossistema de negócios, possibilitando a construção de relacionamentos duradouros e colaborativos.

A escolha e a combinação desses critérios devem ser orientadas por uma análise cuidadosa das necessidades específicas do negócio, da disponibilidade e qualidade dos dados e da capacidade analítica da empresa. Quando bem aplicados,

esses critérios viabilizam a criação de segmentos robustos, relevantes e acionáveis, que servem como base para estratégias comerciais mais eficazes, campanhas de *marketing* direcionadas e uma gestão de relacionamento mais estratégica e personalizada no contexto B2B.

#### 2.3.1.1 Critérios financeiros

A literatura especializada identifica uma gama abrangente de critérios que podem ser utilizados no processo de segmentação de clientes no contexto B2B (*Business to Business*). A escolha desses critérios depende, em grande parte, dos objetivos estratégicos do negócio, do tipo de produto ou serviço oferecido, da dinâmica do mercado de atuação e do nível de maturidade analítica e tecnológica da empresa. À medida que as organizações avançam em seus processos de transformação digital e coleta de dados, torna-se possível aplicar segmentações mais refinadas e propor ações.

Esses critérios podem ser agrupados, de forma geral, em três grandes categorias:

a) Critérios financeiros: consideram variáveis quantitativas que indicam a rentabilidade, risco e potencial econômico do cliente. Exemplos incluem:
    a. Faturamento anual da empresa-cliente;
    b. *Ticket* médio das compras realizadas;
    c. CAC;
    d. LTV;
    e. Margem de contribuição;
    f. Volume de compras recorrentes.

Esses indicadores permitem priorizar contas com maior retorno financeiro esperado e avaliar a viabilidade econômica de estratégias específicas para cada grupo.

b) Critérios comportamentais: avaliam como o cliente interage com a empresa, seus hábitos de compra, frequência de relacionamento e respostas a campanhas comerciais ou de *marketing*. Incluem:

    a. Histórico de interações com canais de vendas (online ou presencial);
    b. Participação em programas de fidelidade ou eventos corporativos;
    c. Nível de engajamento com conteúdos digitais (*e-mails, webinars, e-books*);
    d. Tempo médio entre as compras (*buying cycle*);
    e. Velocidade de resposta em negociações.

Esses dados ajudam a identificar o nível de maturidade da conta, seu potencial de crescimento e o tipo de abordagem comercial mais eficaz.

c) Critérios estratégicos: envolvem a adequação do cliente ao posicionamento da empresa e seu alinhamento com a proposta de valor, visão de futuro ou até objetivos ESG (Ambiental, Social e Governança). Exemplos:

    a. Setor de atuação (ex: saúde, varejo, manufatura);
    b. Modelo de negócio (B2B, B2C, B2B, etc);
    c. Grau de sinergia tecnológica ou operacional com o portfólio atual;
    d. Localização geográfica e potencial de expansão regional;
    e. Potencial de parceria estratégica ou co-desenvolvimento de soluções.

Esses critérios são fundamentais para selecionar contas-chave (*key accounts*), definir nichos prioritários ou estruturar abordagens de vendas complexas, como o *Account-Based Marketing* (ABM).

A correta combinação entre essas dimensões permite que a segmentação B2B vá além da classificação superficial dos clientes, promovendo uma visão mais holística e orientada a resultados. Empresas que integram esses critérios de maneira sistemática conseguem priorizar oportunidades de maior valor, otimizar a alocação de recursos comerciais e personalizar suas estratégias de relacionamento com maior precisão.

#### 2.3.1.2 Critérios comportamentais

Os critérios comportamentais analisam o histórico de interação entre a empresa e seus clientes, oferecendo insumos valiosos para a personalização de estratégias de *marketing*, vendas e atendimento. Ao contrário dos critérios puramente financeiros, que focam na rentabilidade passada ou projetada, os critérios comportamentais permitem avaliar o grau de engajamento, maturidade e responsividade do cliente ao longo do tempo, fornecendo uma visão mais rica sobre o relacionamento estabelecido com a organização.

Entre os principais exemplos de critérios comportamentais aplicáveis à segmentação B2B, destacam-se:

a) Frequência de compras e recorrência de pedidos: identifica padrões de consumo regulares ou sazonais, úteis para prever demandas e antecipar ofertas;
b) Tempo médio de relacionamento com a empresa: mede a longevidade da parceria comercial, o que pode estar correlacionado a confiança, retenção e potencial de venda sobre a base de clientes chamada de *upselling*;
c) Engajamento com canais de comunicação e suporte técnico: avalia o envolvimento do cliente com e-mails, chamadas, reuniões, abertura de chamados e uso de portais de autoatendimento;
d) Respostas a campanhas de *marketing* anteriores: inclui taxas de abertura de e-mails, cliques em links, participação em eventos e conversões registradas em campanhas específicas (DAHANA et al., 2019).

Esses dados são, em geral, extraídos de ferramentas do tipo CRM, como *Salesforce*, *HubSpot* ou *Microsoft Dynamics*, bem como de plataformas de automação de *marketing* (ex: RD *Station*, *Mailchimp*, *ActiveCampaign*). A análise conjunta dessas informações permite construir perfis de comportamento longitudinal, com destaque para mudanças no padrão de consumo, queda no engajamento ou sinais de *churn* iminente, possibilitando ações preventivas.

Além disso, esses critérios comportamentais conferem dinamismo à segmentação, pois possibilitam que os segmentos evoluam com o tempo — um conceito alinhado à segmentação preditiva e aos princípios de *Customer Success*

(CS). Quando aplicados de forma consistente, eles permitem à empresa desenvolver estratégias mais precisas e oportunas, como campanhas de reativação de clientes inativos, ofertas específicas baseadas em comportamento recente, e até mesmo fluxos automatizados de nutrição e fidelização.

#### 2.3.1.3 Critérios estratégicos

Os critérios estratégicos avaliam o potencial de um cliente para contribuir com o crescimento futuro da empresa fornecedora, indo além da rentabilidade imediata e considerando aspectos como sinergia de longo prazo, valor estratégico da parceria e possibilidade de coevolução comercial. Diferenciam-se dos critérios financeiros e comportamentais por enfatizarem a perspectiva de alinhamento estrutural e estratégico entre as partes, especialmente relevante no contexto B2B, onde as relações tendem a ser mais duradouras e complexas.

São exemplos típicos desses critérios:

a) Potencial de expansão da conta (*upsell/cross-sell*): refere-se à capacidade de aumentar o volume de negócios com o cliente ao oferecer produtos complementares (*cross-sell*) ou *upgrades,* ou seja, adicionais no mesmo produto (upsell), ampliando o valor da conta ao longo do tempo;
b) Aderência aos produtos ou serviços ofertados: mede o grau de compatibilidade entre as soluções da empresa fornecedora e as necessidades atuais e futuras da empresa cliente;
c) Sinergia cultural e estratégica entre as empresas: considera afinidades em termos de valores corporativos, estilo de gestão, visão de futuro e práticas comerciais, fatores que facilitam a construção de parcerias sólidas e duradouras;
d) Posicionamento da empresa cliente dentro de seu próprio mercado: avalia se o cliente é líder, referência ou inovador em seu segmento, o que pode gerar efeitos indiretos positivos como credibilidade, visibilidade e influência no setor (KANCHANAPOOM; CHONGWATPOL, 2022).

Esses critérios, embora mais qualitativos por natureza, podem — e devem — ser operacionalizados de forma sistemática, a partir de escalas de avaliação interna, checklists padronizados, entrevistas com executivos da área comercial e painéis de validação entre áreas técnicas e estratégicas. Empresas mais maduras podem empregar métodos como análise multicritério (AHP/MCDA) ou modelos de *scoring* ponderado para atribuir pesos a esses critérios e classificá-los de forma consistente em sistemas de CRM ou plataformas de *account planning*.

Além disso, os critérios estratégicos são frequentemente utilizados na definição de *Key Accounts* (Contas Chave), ABM (*Account-Based Marketing*) e planejamentos de parcerias estratégicas, por permitirem identificar clientes que, mesmo não sendo os mais rentáveis no curto prazo, oferecem elevado potencial de valor estratégico e institucional para a empresa fornecedora — seja pelo potencial de co-inovação, pela abertura de novos mercados ou pela influência que exercem no setor.

### 2.3.2 Técnicas quantitativas para segmentação B2B

Com a digitalização dos processos empresariais e o crescimento exponencial do volume e da variedade de dados disponíveis, surgiram metodologias mais robustas, escaláveis e automatizadas para a segmentação de clientes. A incorporação de técnicas de ciência de dados e, em especial, de machine learning, revolucionou a forma como as empresas identificam e compreendem seus públicos-alvo, permitindo o agrupamento de clientes com base em padrões ocultos que muitas vezes não são perceptíveis por métodos tradicionais ou análises univariadas.

#### 2.3.2.1 Clusterização

Técnicas de clusterização (ou agrupamento não supervisionado) são amplamente utilizadas na criação de segmentos homogêneos de clientes, com base em similaridades de comportamento, características transacionais ou atributos demográficos. Entre os algoritmos mais populares, destacam-se:

a) *K-Means*: eficaz na formação de clusters com base na distância euclidiana entre variáveis previamente normalizadas, sendo especialmente útil em bases

de dados estruturadas com grande volume de observações. Sua simplicidade e velocidade de execução o tornam adequado para aplicações em tempo real e dashboards interativos (HAN; KAMBER; PEI, 2011).

b) *DBSCAN* (*Density-Based Spatial Clustering of Applications with Noise*): permite identificar clusters de forma flexível, com base na densidade de pontos, sendo eficaz na detecção de outliers e em situações em que os clusters não têm formato esférico.

c) *Hierarchical Clustering*: constrói uma árvore de agrupamentos (dendrograma), útil para análises exploratórias, especialmente quando o número ideal de clusters não é conhecido previamente.

#### 2.3.2.2 Modelos supervisionados

Modelos de aprendizado supervisionado são indicados quando o objetivo é prever variáveis de interesse, como LTV, *churn* (evasão) ou propensão de compra. Estes modelos aprendem com dados rotulados históricos e produzem classificações ou regressões com base em novos dados. Destacam-se:

a) *Random forest*: modelo baseado em árvores de decisão, altamente robusto e interpretável, adequado para previsão de *churn* e pontuação de clientes por risco;

b) *Gradient boosting machines* (GBM, XGBoost, LightGBM): técnicas poderosas que combinam vários modelos fracos para formar um preditor forte, com excelente desempenho preditivo;

c) Redes neurais artificiais (RNA): recomendadas quando há uma alta complexidade não-linear entre as variáveis, sendo capazes de capturar padrões sofisticados, especialmente em grandes bases.

Estudos como o de Bauer e Jannach (2021) evidenciam que o uso desses modelos supervisionados em estratégias de segmentação preditiva eleva significativamente a acurácia das decisões comerciais, sobretudo em campanhas de retenção e recomendação.

#### 2.3.2.3 Análise fatorial

A análise fatorial é uma técnica estatística que permite reduzir a dimensionalidade de bases de dados com muitas variáveis correlacionadas, facilitando a interpretação dos dados e a identificação de fatores latentes que influenciam o comportamento dos clientes. Por meio dela, é possível agrupar variáveis que representam dimensões comuns, como sensibilidade a preço, grau de digitalização ou nível de interação com a marca. A análise fatorial é especialmente útil em estudos de comportamento organizacional e pesquisas B2B com grande número de atributos qualitativos (HAIR et al., 2009).

#### 2.3.2.3 Processos KDD e CRISP-DM

Para garantir que a segmentação seja realizada de forma estruturada e alinhada aos objetivos organizacionais, é recomendada a adoção de metodologias consolidadas de mineração de dados, como:

a) KDD: define um processo sistemático que inclui seleção, pré-processamento, transformação, mineração de dados e interpretação dos resultados (FAYYAD et al., 1996);
b) CRISP-DM (*Cross Industry Standard Process for Data Mining*): modelo de referência amplamente utilizado na indústria, que organiza o processo de ciência de dados em seis fases interdependentes: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implantação.

A adoção desses modelos metodológicos assegura que o projeto de segmentação seja consistente, replicável e orientado a resultados, promovendo integração entre áreas técnicas e de negócio e contribuindo para uma governança analítica mais madura.

### 2.3.3 Desafios atuais e perspectivas futuras

Embora as técnicas analíticas e ferramentas de modelagem estejam cada vez mais acessíveis e democratizadas, a segmentação B2B ainda enfrenta obstáculos significativos que comprometem a eficácia dos modelos implementados e a escalabilidade das estratégias geradas a partir deles. Entre os principais desafios, destacam-se:

a) Fragmentação dos dados entre diferentes sistemas legados (ERP, CRM, BI), que dificulta a obtenção de uma visão única e consolidada do cliente, comprometendo a consistência das análises;

b) Baixa qualidade, incompletude ou desatualização dos dados, fatores que afetam diretamente os resultados da modelagem preditiva e aumentam o risco de viés e inferências incorretas;

c) Falta de integração entre as áreas de *marketing*, vendas e tecnologia da informação, o que impede a implantação eficaz de estratégias baseadas em dados e dificulta a governança analítica organizacional;

d) Mudanças rápidas no comportamento dos clientes, especialmente em cenários de incerteza ou crise, como observado durante e após a pandemia de COVID-19, que exigem modelos mais ágeis, adaptativos e sensíveis ao tempo (LI et al., 2022).

Como resposta a esses desafios, observa-se uma tendência crescente à adoção de sistemas de segmentação dinâmica, baseados em inteligência artificial e análise em tempo real. Essas soluções buscam substituir os modelos estáticos e rígidos por abordagens adaptativas, que acompanham o ciclo de vida do cliente em tempo contínuo. O uso de algoritmos de deep learning, redes neurais convolucionais e técnicas de análise de sentimentos aplicadas a interações textuais (como e-mails, chats, transcrições de reuniões virtuais e chamadas telefônicas) tem possibilitado uma visão mais rica e preditiva da jornada do cliente B2B, permitindo intervenções mais precisas e tempestivas (SU et al., 2023; HUANG; RUST, 2020).

A evolução da segmentação de clientes no ambiente B2B acompanha essa transformação: passou-se de abordagens empíricas e intuitivas, baseadas em julgamento de especialistas ou histórico comercial, para modelos matematicamente fundamentados e orientados por dados, com validação estatística e capacidade de generalização. Nesse novo paradigma, a utilização combinada de métricas financeiras

como o LTV e o CAC, associada a algoritmos de machine learning supervisionados e não supervisionados, permite a construção de segmentos altamente eficazes na maximização do valor do cliente, com benefícios diretos em rentabilidade, fidelização e ROI de campanhas.

A abordagem *data-driven*, portanto, não apenas amplia o conhecimento sobre os clientes, como também potencializa a personalização de ofertas, a eficiência operacional e a competitividade das organizações. Em vez de se basear em segmentações fixas, ela permite modelos responsivos e continuamente atualizados, ajustando-se conforme os dados comportamentais, contextuais e mercadológicos evoluem.

Dessa forma, a compreensão aprofundada dos critérios de segmentação, o domínio das técnicas de análise de dados e, sobretudo, a integração entre áreas estratégicas como *marketing*, vendas, TI e inteligência de mercado tornam-se pilares fundamentais para o sucesso das estratégias comerciais no cenário B2B contemporâneo. Esse alinhamento é indispensável para garantir que as ideias geradas pela análise de dados se traduzam em ações efetivas e orientadas a resultados.

Estudos futuros devem considerar, entre outras vertentes promissoras, a evolução dos modelos de inteligência artificial generativa, com potencial para criar perfis sintéticos, simular jornadas de clientes e gerar conteúdos personalizados em escala, bem como a integração de dados não estruturados — como voz, texto livre, imagens e vídeos — nos modelos preditivos de valor e comportamento. Essa integração representa um novo patamar de sofisticação analítica, com impacto direto na assertividade, automação e personalização da segmentação B2B.

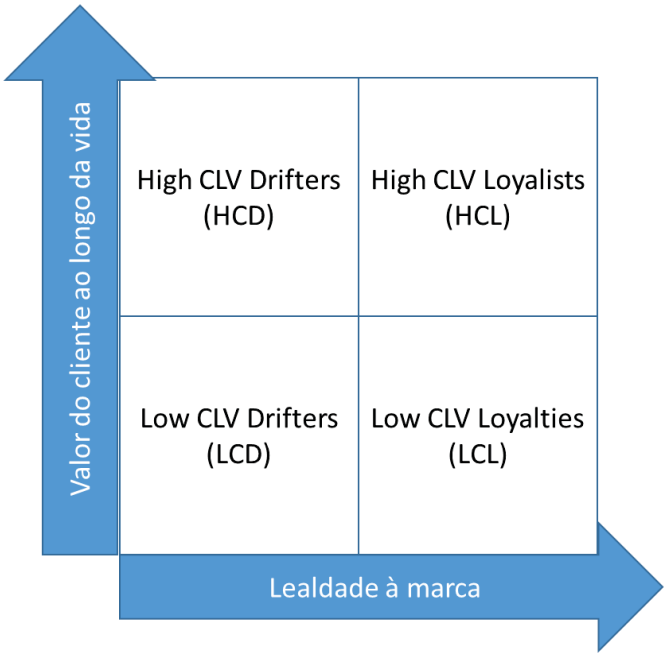
# 3 TRABALHOS CORRELATOS

A segmentação de clientes e a gestão do valor que cada cliente representa ao longo do tempo são temas centrais na literatura contemporânea de *marketing* estratégico e gestão de relacionamento. Essas práticas surgem como respostas fundamentais à necessidade de alocar recursos de forma eficiente em mercados cada vez mais saturados, competitivos e orientados por dados. Nesse contexto, Kotler e Keller (2006) argumentam que as empresas não devem tentar satisfazer indiscriminadamente todos os consumidores, mas sim concentrar seus esforços nos que demonstram maior potencial de retorno financeiro. Para esses autores, um cliente lucrativo é definido como “uma pessoa, família ou empresa cujas receitas ao longo da vida excedem, em um valor aceitável, os custos da empresa para atrair, vender e atender esse cliente”. A partir dessa concepção, emerge o conceito de LTV — ou Valor do Tempo de Vida do Cliente — como uma métrica-chave para orientar decisões estratégicas relacionadas à aquisição, retenção e expansão do relacionamento com os clientes. O LTV permite quantificar o valor econômico de longo prazo gerado por cada cliente, funcionando como um guia para decisões mais inteligentes sobre quais perfis merecem investimentos contínuos e quais podem ser despriorizados.

A capacidade de identificar, prever e gerenciar clientes de alto valor tornou-se, atualmente, um dos pilares das estratégias empresariais focadas em performance e fidelização sustentável. O LTV consolidou-se como uma métrica central não só para o planejamento de ações comerciais e de *marketing*, mas também como critério essencial para segmentação preditiva, orçamentação estratégica, alocação eficiente de recursos multicanal e projeção de retorno sobre investimento (ROI). Sua importância crescente é respaldada por diversos estudos contemporâneos, como os de Pollak (2021), Zhang et al. (2022), Li et al. (2022), Afiniti (2022) e Su et al. (2023). Esses pesquisadores demonstram que o LTV sintetiza de maneira integrada os benefícios econômicos gerados pelas interações dos clientes com a organização ao longo de todo o ciclo de vida, permitindo que as empresas tomem decisões mais assertivas tanto no nível operacional quanto no tático e estratégico. Em ambientes digitais e *omnichannel*, caracterizados pela volatilidade e distribuição dispersa do comportamento do consumidor, a modelagem do LTV torna-se ainda mais crítica para antecipar *churn*, identificar oportunidades de *upsell* e definir prioridades de atendimento.

No campo aplicado, Kanchanapoom e Chongwatpol (2022) apresentam um modelo de segmentação orientado pelo LTV no setor de medicina complementar e alternativa — um mercado marcado pela importância de relacionamentos de longo prazo e construção gradual de confiança. Conforme a figura 10, os autores propõem a divisão dos clientes em quatro segmentos distintos, que combinam critérios de valor e lealdade à marca. Este modelo visa identificar os segmentos com maior potencial futuro, servindo como instrumento preditivo para direcionar decisões comerciais e de *marketing*. A classificação segmenta os clientes em: (i) alto valor e alta lealdade, onde a recomendação é fortalecer e preservar o vínculo; (ii) alto valor e baixa lealdade, em que o foco deve ser aumentar a fidelização; (iii) baixo valor e alta lealdade, nos quais é recomendada a maximização do retorno, com possível descontinuação futura; e (iv) baixo valor e baixa lealdade, que podem ser alvo de desvinculação gradual. Essa abordagem permite uma alocação de recursos mais eficiente, alinhada ao retorno potencial previsto, enfatizando a importância de estratégias direcionadas, personalizadas e sustentáveis de relacionamento com o cliente.

**Figura 10 - Proposta de segmentação de clientes baseado no valor ao longo da sua vida e na sua lealdade à marca**




Fonte: Adaptado pelo autor a partir de Kanchanapoom e Chongwatpol (2022).

O estudo de Afiniti (2022) reforça a centralidade do LTV na tomada de decisões gerenciais, especialmente em setores com modelos contratuais de relacionamento com clientes. Os autores argumentam que uma estimativa acurada do valor do tempo

de vida dos clientes é essencial para o alinhamento entre investimentos em aquisição e o retorno financeiro projetado ao longo do tempo. Para esse fim, propõem um modelo flexível de riscos proporcionais, que permite incorporar a probabilidade de *churn* (evasão) como variável-chave no cálculo do LTV. A abordagem parte do pressuposto de que a organização possui um modelo de *churn* minimamente calibrado, cuja integração à modelagem de LTV permite calcular, com maior precisão, o tempo esperado de permanência de um cliente, ponderando esse tempo pelos lucros esperados em cada período. Isso torna o modelo particularmente adequado para ambientes com relações contratuais explícitas, como telecomunicações, seguros ou assinaturas de serviços digitais, onde os fluxos de receita são previsíveis, mas dependem criticamente da retenção de clientes.

Complementarmente, Su et al. (2023) enfrentam um dos principais desafios relacionados à modelagem do LTV em ambientes não contratuais e altamente dinâmicos, como plataformas de publicidade online. Nesses contextos, os dados de consumo por usuário tendem a ser escassos, fragmentados ou inconsistentes dentro de um único domínio de análise. Como alternativa, os autores propõem uma estrutura adaptativa entre domínios, denominada CDAF (*Cross-Domain Adaptive Framework*), que permite a transferência de aprendizado de um domínio com dados abundantes (por exemplo, uma plataforma digital consolidada) para outro domínio com dados mais limitados (como uma plataforma emergente). O método proposto busca mitigar dois problemas simultâneos: (i) a escassez de dados históricos de consumo e (ii) o desalinhamento estatístico entre os domínios fonte e alvo. Para isso, a CDAF adota uma arquitetura que aprende padrões gerais de LTV em plataformas relacionadas, preservando a generalização e ajustando as distribuições para o novo domínio. Essa estratégia permite realizar predições mais robustas, mesmo em ambientes onde a informação direta sobre o comportamento dos usuários ainda está em formação, destacando-se como um exemplo promissor de transferência de aprendizado (*transfer learning*) no campo de modelagem de valor de cliente.

Na mesma linha de enfrentamento das limitações dos modelos tradicionais, Zhang et al. (2022) destacam que o LTV, ao mensurar a contribuição econômica de longo prazo de clientes ao longo de relacionamentos contínuos com produtos ou serviços, pode fornecer insumos decisivos para a definição de estratégias de entrega de valor. No entanto, os autores argumentam que as abordagens atuais enfrentam dois entraves significativos: por um lado, a incapacidade de modelar adequadamente

relações temporais e não lineares; por outro, a ausência de soluções computacionalmente viáveis para grandes volumes de dados. Em resposta, Zhang e colaboradores propõem um modelo geral de LTV, que supera a fragmentação das abordagens anteriores ao integrar aspectos de longo prazo em vez de se limitar a estimativas baseadas em cliques ou compras recentes. Para alcançar esse objetivo, os autores implementam uma solução de programação dinâmica rápida, baseada em um método de bisseção mutado e na hipótese de experimentação sem memória, o que permite acelerar o processo de otimização dos parâmetros envolvidos na projeção do LTV. Essa proposta se mostra particularmente eficaz para aplicações em ambientes digitais e plataformas de serviços contínuos, onde o comportamento do cliente é complexo e a avaliação de seu valor futuro exige uma abordagem preditiva mais sofisticada e adaptável.

Pollak (2021) explora um dos principais desafios enfrentados pelas empresas ao prever o LTV de clientes em contextos não contratuais, nos quais a relação com o consumidor é descontinuada ou intermitente. Nesse tipo de ambiente, onde não há garantias explícitas de continuidade da relação comercial, a estimativa do valor do tempo de vida dos clientes torna-se dependente essencialmente de padrões históricos de compra. Com isso, a previsão exige um modelo que consiga inferir comportamentos futuros a partir de dados passados. O autor realiza uma comparação entre dois métodos: o primeiro, baseado no modelo estatístico conhecido como “compre até morrer” (*Buy-Till-You-Die Model*), que utiliza dados transacionais anteriores para modelar a propensão de recompra até a "morte" do cliente (i.e., inatividade); o segundo, uma rede neural artificial, aplicada ao mesmo conjunto de dados. A análise realizada oferece resultados quantitativos e qualitativos que comparam a precisão, a robustez e a aplicabilidade prática de ambas as abordagens. Como conclusão, Pollak propõe diretrizes práticas para que gestores de *marketing* escolham o modelo mais adequado, a depender do tipo de dado disponível, da complexidade do domínio e do objetivo estratégico da organização.

Em complemento, Bauer e Jannach (2021) propõem um conjunto de técnicas baseadas em inteligência artificial que visam elevar a precisão das previsões de LTV em contextos altamente dinâmicos, como o comércio eletrônico e plataformas digitais. Dentre as inovações destacadas, encontra-se o uso de redes neurais recorrentes (RNNs), capazes de capturar dependências temporais nas interações entre clientes e produtos, o que permite uma modelagem mais realista do comportamento sequencial

de compra. Além disso, os autores empregam modelos de atenção (*attention models*), que aumentam a capacidade da rede em focar seletivamente em eventos relevantes da sequência de interação, melhorando a capacidade preditiva. Para lidar com a qualidade variável dos dados, é sugerido um pré-processamento avançado, que inclui tratamento de valores ausentes, normalização e codificação apropriada de variáveis categóricas. Um dos diferenciais do estudo é a introdução de modelos Seq2Seq (*Sequence to Sequence*), comumente utilizados em tarefas como tradução automática, mas aqui aplicados para mapear a sequência completa de interações cliente-produto ao longo do tempo. Por fim, Bauer e Jannach propõem uma arquitetura híbrida, combinando modelos baseados em características (*feature-based*) com modelos sequenciais, de modo a explorar as vantagens de ambas as abordagens, mitigando as limitações associadas a soluções isoladas. Essa proposta representa um avanço no campo da modelagem preditiva, especialmente no que diz respeito à complexidade comportamental dos consumidores digitais.

No mesmo escopo, Li et al. (2022) enfrentam o desafio da previsão do LTV em ambientes de altíssima escala, como plataformas digitais com bilhões de usuários. Nesse cenário, a modelagem tradicional torna-se inviável devido à diversidade de perfis de usuários, à alta variabilidade dos dados e à necessidade de predições em tempo real. A solução proposta pelos autores foi aplicada em uma empresa de tecnologia chinesa de grande porte, utilizando um arcabouço robusto de ciência de dados e aprendizado de máquina, que inclui algoritmos de previsão baseados em séries temporais, machine learning supervisionado, processamento em tempo real e inteligência artificial adaptativa. A proposta se destaca pela capacidade de processar grandes volumes de dados de forma eficiente e responsiva, permitindo que as predições de LTV sejam atualizadas dinamicamente conforme o comportamento do usuário evolui. Essa abordagem demonstra que, além da sofisticação algorítmica, é imprescindível escalabilidade e integração com sistemas operacionais de negócio, para garantir que as previsões de valor de cliente possam ser utilizadas de forma prática, ágil e alinhada às necessidades de mercado.

Olnén (2022) destaca que a precisão na estimativa do LTV é um fator determinante para organizações que desejam otimizar suas estratégias de relacionamento com o cliente e, ao mesmo tempo, maximizar a rentabilidade no longo prazo. Para alcançar esse objetivo, o autor emprega técnicas avançadas de aprendizado de máquina, com ênfase em redes neurais profundas (deep learning),

que se mostram particularmente eficazes na captura das complexidades e nuances comportamentais dos consumidores. Esses modelos computacionais têm a capacidade de aprender, a partir de grandes volumes de dados históricos, padrões sutis em variáveis como a frequência de compras, o valor transacional médio e o nível de engajamento com os serviços prestados, permitindo uma previsão mais acurada do valor futuro de cada cliente.

Ainda segundo Olnén (2022), uma compreensão aprofundada do LTV previsto capacita as empresas não apenas a avaliar o retorno sobre o investimento (ROI) em campanhas de *marketing*, mas também a tomar decisões estratégicas quanto ao valor de mercado da própria empresa, especialmente em contextos de aquisição, fusão ou abertura de capital. Além disso, ao classificar os clientes com base em seu LTV estimado, os profissionais de *marketing* podem realocar de forma mais eficiente os recursos destinados à aquisição, retenção ou estratégias de *upsell* e *cross-sell*. O autor também chama atenção para uma característica estatística frequentemente presente nas distribuições de LTV: a cauda pesada, isto é, a presença de poucos clientes que geram valores muito elevados, contrastando com a maioria que gera menor retorno. Diante disso, o desempenho dos modelos preditivos é avaliado por dois critérios principais: discriminação, que avalia a capacidade de distinguir clientes de alto e baixo valor, e calibração, que mede a proximidade entre os valores previstos e os valores observados. Contudo, o processo de ponderação entre essas métricas, segundo Olnén, é oneroso e sujeito a vieses, dado que exige análise manual. Com base nos experimentos relatados, o autor infere que a discriminação tende a receber 19 vezes mais peso que a calibração durante a avaliação dos modelos, indicando uma priorização prática da capacidade de segmentação sobre a exatidão absoluta das previsões. O autor também observa que há uma lacuna na literatura quanto ao efeito do aumento do horizonte temporal de dados históricos sobre a precisão preditiva dos modelos, o que sugere um campo promissor para futuras investigações.

No mesmo escopo de aplicação ao varejo digital, Jasek et al. (2019) argumentam que a escolha de um modelo LTV apropriado é uma etapa crucial para empresas que buscam implementar uma abordagem gerencial baseada em valor do cliente em suas plataformas de e-commerce B2C. O contexto do varejo online impõe pressupostos e desafios específicos, como a natureza não contratual do relacionamento com os clientes, a recorrência imprevisível das compras e a variabilidade no comportamento de consumo ao longo do tempo. Os autores

conduzem uma análise comparativa entre onze diferentes modelos probabilísticos de previsão de LTV, avaliando tanto o desempenho estatístico quanto a capacidade preditiva em cenários reais de comércio eletrônico. Os resultados obtidos evidenciam que, embora existam diversas abordagens teóricas para a previsão do LTV, alguns modelos são claramente superiores quando aplicados a ambientes de alta complexidade e dinamismo, como o varejo digital. A pesquisa reforça, assim, a necessidade de adequação contextual na escolha do modelo, tendo em vista as características operacionais do negócio, os tipos de dados disponíveis e os objetivos estratégicos da organização. Em última análise, o estudo de Jasek et al. ressalta que o entendimento profundo do valor do cliente é essencial para sustentar decisões comerciais assertivas, promover a eficiência operacional e garantir vantagem competitiva sustentável no ambiente digital contemporâneo.

Win e Bo (2020) enfatizam que a segmentação de clientes com base no LTV configura uma prática essencial no *marketing* contemporâneo, especialmente em ambientes digitais competitivos. Ao possibilitar a identificação e a priorização de grupos de clientes segundo seu valor financeiro estimado ao longo do tempo, essa abordagem permite que as empresas otimizem seus investimentos em aquisição, retenção e fidelização. Os autores aplicam o algoritmo Random Forest, um modelo de aprendizado de máquina supervisionado, com o objetivo de prever a classe de LTV dos clientes em um horizonte de um ano. Os resultados obtidos demonstram que esse tipo de técnica é eficaz para orientar decisões estratégicas no CRM, permitindo que o varejista direcione seus recursos para clientes com maior potencial de retorno, aumentando a eficiência operacional e maximizando o valor agregado. O estudo reforça, assim, a viabilidade e a aplicabilidade prática de métodos preditivos baseados em machine learning na formulação de estratégias de *marketing* no contexto digital.

Na mesma direção, Dahana et al. (2019) abordam o LTV como uma métrica crítica para a construção de estratégias de *marketing* eficazes, especialmente em setores em rápida transformação, como o varejo de moda online. Em sua proposta metodológica, os autores desenvolvem um modelo de classe latente que considera a frequência de compra, a duração do ciclo de vida do cliente e o valor médio das transações como variáveis determinantes para inferir o LTV em diferentes segmentos de mercado. O estudo introduz uma dimensão inovadora ao incorporar padrões de estilo de vida como variável explicativa da heterogeneidade do LTV entre segmentos, demonstrando que fatores comportamentais e psicográficos podem ter impacto

substancial sobre o valor de longo prazo gerado pelos clientes. Ao aplicar o modelo a um conjunto de dados reais de transações e perfis comportamentais de consumidores em uma plataforma de moda, os autores demonstram a capacidade preditiva do modelo proposto, ampliando as possibilidades de segmentação inteligente e customização de campanhas de *marketing*. De forma complementar, os próprios autores definem o LTV como o valor total esperado que a empresa pode obter de um único cliente ao longo de toda a duração do relacionamento, considerando a receita líquida e os custos variáveis associados ao atendimento desse cliente, o que alinha a métrica tanto à visão financeira quanto à perspectiva estratégica da organização.

No contexto da indústria de jogos digitais, Burelli (2019) oferece uma contribuição relevante ao destacar os desafios e oportunidades na modelagem preditiva do comportamento dos jogadores, especialmente em modelos de negócios orientados a serviços, como os jogos *Free to Play* (F2P). Nesse tipo de modelo, a ausência de barreiras iniciais de pagamento e a grande variação no comportamento de engajamento e de gastos tornam a previsão de receitas futuras altamente complexa. O autor argumenta que, diante dessa volatilidade, torna-se essencial dispor de modelos preditivos robustos, capazes de fornecer suporte às decisões relacionadas à aquisição de usuários, personalização de experiências *in-game* e otimização de recursos de desenvolvimento e operação. O artigo ressalta que, para que estratégias eficazes sejam implementadas, é necessário entender não apenas as escolhas passadas dos jogadores, mas também antecipar possíveis trajetórias de comportamento, utilizando dados históricos e técnicas avançadas de *data science* e de aprendizado de máquina. Nesse sentido, a modelagem do LTV em jogos digitais não apenas amplia o entendimento sobre a economia do jogador, mas também possibilita a definição de estratégias mais sustentáveis e orientadas por dados para monetização e retenção.

Wu et al. (2023) exploram os desafios da previsão do LTV em contextos onde a escassez de eventos de consumo e a alta variabilidade dos dados impõem barreiras significativas à precisão das estimativas. Essa realidade é especialmente comum em aplicativos centrados no cliente, nos quais a interação pode ser esporádica e os dados disponíveis são ruidosos ou incompletos. Os autores criticam os métodos tradicionais que treinam preditores de LTV com base em uma única visão dos dados, argumentando que essa abordagem tende a extrair conhecimento de forma limitada e potencialmente enviesada. Para superar tais limitações, propuseram uma estrutura de

multivisualização contrastiva, projetada como uma solução *plug-and-play* (PnP), compatível com diferentes arquiteturas de modelos (*backbones*). Essa estrutura integra múltiplos regressores de LTV heterogêneos, que trazem conhecimentos complementares, resultando em maior robustez e precisão na estimativa do valor do cliente. Além disso, a utilização do aprendizado contrastivo permite capturar relações latentes entre amostras semelhantes, mitigando a dependência da abundância de dados rotulados e reforçando a capacidade do modelo em generalizar padrões úteis.

No mesmo eixo de inovação metodológica, Wang et al. (2019) propõem uma abordagem estatística para a modelagem do LTV que leva em consideração tanto a probabilidade de *churn* (rotatividade) quanto a distribuição assimétrica dos dados de valor, frequentemente observada em mercados com clientes de alto e baixo valor extremo. A proposta metodológica baseia-se em uma mistura entre massa de ponto zero e distribuição log-normal, resultando na chamada distribuição log-normal inflada de zero (ZILN). Tal modelagem é especialmente eficaz para capturar a natureza de "cauda pesada" dos dados de LTV, ao mesmo tempo em que quantifica a incerteza nas previsões pontuais, o que é fundamental para a tomada de decisões estratégicas sob risco. Os autores validam o modelo tanto em modelos lineares tradicionais quanto em redes neurais profundas (DNNs), evidenciando sua flexibilidade e adaptabilidade a diferentes contextos de aplicação. Para avaliação da performance preditiva, são utilizados o coeficiente de Gini normalizado, que mede a capacidade discriminativa do modelo, e gráficos de decil, que avaliam a calibração das previsões. Os resultados empíricos obtidos a partir de dois conjuntos de dados reais demonstram a eficácia do modelo ZILN para diferentes aplicações comerciais e níveis de granularidade nos dados.

Por fim, Cao et al. (2023) abordam a previsão do comportamento do consumidor e sua interseção com a otimização de sortimento, ampliando a aplicação de modelos preditivos de valor para além da estimativa do LTV. Os autores investigam a escolha do cliente a partir de uma mistura de modelos de demanda, que combina a demanda independente com o modelo de logit multinomial, refletindo a realidade de mercados nos quais diferentes segmentos de clientes seguem padrões de decisão distintos. Nesse contexto, cada produto do portfólio possui uma receita esperada associada, e o objetivo do modelo é encontrar o sortimento ótimo — ou seja, a combinação de produtos que maximiza a receita esperada de um cliente. A proposta metodológica mostra que esse problema pode ser resolvido de forma eficiente por

meio da formulação e resolução de um programa linear, tornando a abordagem viável do ponto de vista computacional. Um dos principais achados do estudo é que o tamanho ideal do sortimento cresce proporcionalmente ao tamanho relativo do segmento de clientes que se comporta conforme o modelo de demanda independente, implicando que diferentes perfis de comportamento exigem estratégias diferenciadas de oferta de produtos para a maximização de valor.

## 4 METODOLOGIA

Este trabalho propõe o desenvolvimento de uma metodologia para segmentação de clientes no contexto B2B, com ênfase em empresas que atuam no setor de serviços de cobrança, embora seus princípios e técnicas sejam igualmente aplicáveis a outros setores intensivos em relacionamento com clientes, como segmentos de tecnologia, consultoria, engenharia e indústria de base. A base conceitual da proposta está ancorada na clássica visão de Kotler e Keller (2012),

segundo a qual a segmentação é um dos pilares fundamentais para a efetividade das estratégias de *marketing*. Para os autores, o conhecimento aprofundado do perfil dos clientes e sua organização em grupos coerentes permite a alocação mais racional de recursos, a personalização de ofertas e a maximização do retorno sobre os investimentos em vendas e relacionamento.

A metodologia sugerida fundamenta-se em técnicas de análise de dados e algoritmos de clusterização, com o objetivo de agrupar os clientes com base em variáveis quantitativas e qualitativas, como nível de faturamento, porte organizacional (número de funcionários), tempo de relacionamento com a empresa e indicadores de desempenho. Essa abordagem se alinha às etapas do KDD descrito por Fayyad et al. (1996).

A utilização da clusterização como técnica de segmentação permite que se identifiquem padrões ocultos no comportamento dos clientes, fornecendo à equipe comercial subsídios para tomada de decisão mais precisa quanto às estratégias de abordagem, retenção e reativação de clientes. Tal como sugerem Tan, Steinbach e Kumar (2019), a identificação de agrupamentos homogêneos a partir de dados históricos melhora substancialmente a capacidade preditiva das ações comerciais, permitindo um direcionamento mais assertivo dos recursos e maior aderência entre o perfil do cliente e a proposta de valor da empresa.

Com isso, a proposta metodológica contribui para otimizar indicadores centrais da gestão comercial, como a redução do CAC e o aumento do LTV, compreendido como o valor total gerado por um cliente ao longo do seu ciclo de vida com a empresa (KOTLER; KELLER, 2012; OLIVEIRA, 2018). Dessa forma, a segmentação baseada em dados reais, ancorada no ciclo do KDD, não apenas confere maior objetividade ao processo de gestão de clientes, como também reforça a cultura *data-driven* na tomada de decisões estratégicas. Ao final do processo, a pesquisa evidencia como o uso de algoritmos de agrupamento pode representar uma poderosa ferramenta de apoio à gestão comercial e à inteligência de mercado, com impactos diretos sobre a rentabilidade, a fidelização de clientes e o posicionamento competitivo da organização no ambiente B2B.

## 4.1 PROCEDIMENTOS METODOLÓGICOS

A metodologia adotada neste trabalho tem como referência o KDD, sendo estruturada de forma sequencial e interativa, com foco na extração de conhecimento útil a partir de grandes volumes de dados.

A primeira etapa, de seleção e coleta de dados, contempla o uso de bases secundárias, públicas e acessíveis, que contém informações cadastrais e financeiras de empresas brasileiras que atuam em relações comerciais do tipo *Business to Business* (B2B). Os critérios para inclusão dos dados baseiam-se em atributos relevantes para análise de rentabilidade e relacionamento comercial, tais como faturamento anual, número de funcionários, tempo de operação no mercado e histórico de inadimplência.

Na etapa seguinte, realiza-se o tratamento e pré-processamento dos dados, essencial para garantir a qualidade e a integridade da base a ser analisada. Serão aplicadas técnicas como identificação e remoção de ruídos, inconsistências e valores ausentes, com o apoio de métodos de imputação estatística e exclusão criteriosa de registros inválidos. A normalização das variáveis será feita por meio de escalonamento Min-Max e padronização por Z-*Score*, assegurando homogeneidade nas escalas numéricas. Adicionalmente, outliers serão detectados e tratados com base na análise gráfica de boxplots e nos limites estatísticos da amplitude interquartil (IQR), de forma a garantir a robustez dos modelos subsequentes.

Posteriormente, serão realizadas a transformação e redução de dimensionalidade, com o objetivo de condensar as variáveis mais relevantes e eliminar redundâncias sem perda significativa de informação. Serão aplicadas técnicas com a análise de correlação e Análise de Componentes Principais (PCA), que permitem extrair fatores latentes e otimizar a performance computacional dos algoritmos empregados nas etapas posteriores.

A segmentação dos clientes será conduzida por meio da técnica de clusterização não supervisionada *K-Means*, escolhida por sua eficiência computacional e simplicidade interpretativa. A definição do número ideal de clusters será realizada com base em critérios objetivos, como o Método do Cotovelo (*Elbow Method*) e a Pontuação de Silhueta (*Silhouette Score*), de modo a garantir a formação de grupos internamente homogêneos e externamente distintos. A segmentação

resultante permitirá identificar perfis comerciais com similaridades estruturais, facilitando a definição de estratégias personalizadas.

Na sequência, serão aplicados modelos de classificação e predição, para estimar o LTV e a probabilidade de *churn* de cada cliente. Para isso, serão empregados algoritmos supervisionados como Árvores de Decisão, Random Forest, Redes Neurais Artificiais e Algoritmos Genéticos do tipo *AntMiner+*. A escolha dessas técnicas justifica-se por sua capacidade de capturar padrões complexos, mesmo em contextos com alta dimensionalidade, além de apresentarem boa interpretabilidade e desempenho preditivo comprovado na literatura.

A validação dos segmentos formados será realizada com base em métricas quantitativas e qualitativas. Avaliar-se-á a pureza interna dos clusters, isto é, a proporção de membros que compartilham características predominantes, bem como a diferenciação externa entre os grupos, por meio da distância euclidiana entre os centroides e de testes estatísticos como ANOVA e Testes T. Além disso, será conduzida uma análise de correlação entre os segmentos formados e indicadores de negócio relevantes, como LTV, CAC e taxa de *churn*.

Por fim, a etapa de interpretação dos resultados buscará traduzir os achados analíticos em insumos estratégicos para a gestão comercial. A partir da caracterização dos segmentos, será possível propor ações de relacionamento prioritárias, otimizar a alocação de recursos de *marketing* e estruturar campanhas de retenção ou prospecção, considerando o potencial de rentabilidade e o risco associado a cada grupo de clientes. Assim, espera-se demonstrar como uma abordagem orientada por dados pode impulsionar a eficiência e a eficácia das estratégias de segmentação no contexto B2B

A Figura 11, apresenta o esquema das etapas que compõem a aplicação dessa metodologia no contexto empresarial.

**Figura 11 – Esquema da Metodologia de Segmentação de Clientes**

| Etapas | Atividades | Descrição | Objetivo | Resultados Esperados |
|---|---|---|---|---|
| **Coleta dos Dados** | **Coleta e Organização dos Dados** | Reunião de dados transacionais B2B provenientes de sistemas internos (ERP, CRM e bases históricas), incluindo receita, frequência de compra, tempo de relacionamento, custos de aquisição (CAC) e variáveis comportamentais dos clientes. | Estruturar uma base de dados consistente e representativa do comportamento dos clientes. | Base de dados consolidada e organizada para análise. |
| **Pré-processamento** | **Tratamento e Preparação dos Dados** | Realização de limpeza dos dados, padronização das variáveis, tratamento de valores ausentes e outliers, seleção das variáveis relevantes e construção da métrica de Lifetime Value (LTV). | Garantir qualidade, comparabilidade e confiabilidade dos dados utilizados na análise. | Dados preparados e métrica de LTV definida. |
| **Análise Estatística** | **Análise e Mineração de Dados** | Aplicação de análise exploratória dos dados e técnicas estatísticas, seguida da utilização de métodos de clusterização para identificação de padrões e agrupamentos dos clientes. | Identificar padrões de comportamento e estruturas latentes na base de clientes. | Identificação de agrupamentos preliminares de clientes. |
| **Mineração de Dados** | **Segmentação de Clientes Orientada ao LTV** | Formação e caracterização dos segmentos de clientes com base no LTV, análise comparativa entre os clusters e avaliação da relação entre LTV e CAC. | Classificar os clientes segundo seu potencial de valor ao longo do tempo. | Segmentos de clientes claramente definidos e caracterizados. |
| **Simulação das Métricas de Negócio** | **Inferências Analíticas** | Interpretação dos segmentos identificados com foco em suporte à tomada de decisão gerencial, priorização de investimentos e definição de metas comerciais. | Gerar inferências analíticas a partir dos resultados da segmentação. | Subsídios analíticos para decisões estratégicas. |
| | **Aplicação Gerencial e Ciclo de Aprendizado** | Utilização dos segmentos para orientar ações comerciais e de marketing, acompanhamento dos resultados e retroalimentação do modelo com novos dados. | Tornar a metodologia replicável e aplicável em contextos empresariais. | Metodologia operacional e interativa para gestão de clientes. |

Fonte: Adaptado pelo autor, 2025.

### 4.1.1 Coleta dos Dados

Os dados externos que foram utilizados são provenientes de fontes brasileiras de reconhecimento público, como o Cadastro Nacional da Pessoa Jurídica – CNPJ, mantido pela Receita Federal do Brasil – RFB, registros das Juntas Comerciais, dados tornados públicos pela Comissão de Valores Mobiliários – CVM e estatísticas públicas geradas por entidades como o Instituto Brasileiro de Geografia e Estatística – IBGE.

Estas bases, apesar de não fornecerem individualmente informações financeiras abrangentes ou integradas para todas as empresas, representam a espinha dorsal da infraestrutura pública de informação sobre pessoas jurídicas do país. Na prática, esses dados são coletados, processados, tratados, integrados e enriquecidos de forma contínua por empresas de tecnologia especializada em inteligência de mercado, estas fornecem os dados com objetivo de auxiliar na construção de estratégias que apoiam a tomada de decisão ou para estudos pontuais dos seus clientes conforme demanda.

### 4.1.2 Pré-processamento

Após a escolha da base de dados, foi necessário realizar o pré-processamento, etapa considerada essencial em qualquer projeto de mineração de dados e que influencia diretamente a qualidade dos modelos e das interpretações subsequentes (FAYYAD et al., 1996; HAN; KAMBER; PEI, 2012). O pré-processamento tem como objetivo central preparar os dados de forma a garantir que os algoritmos de análise operem sobre um conjunto coerente, livre de ruídos e inconsistências, maximizando a confiabilidade dos resultados. A primeira atividade conduzida nesse processo foi a etapa de limpeza dos dados, que consistiu na verificação detalhada da qualidade dos registros, bem como na identificação e eliminação de inconsistências, erros de digitação, duplicatas e valores ausentes. Conforme alertam Han, Kamber e Pei (2012), a presença de dados incompletos ou imprecisos pode comprometer profundamente o desempenho de modelos preditivos e de agrupamento, levando a interpretações equivocadas e à tomada de decisões inadequadas. Durante essa análise, observou-se que diversas linhas da base não apresentavam valores preenchidos em campos considerados críticos, especialmente o Faturamento Presumido (*FatPres*) e a Quantidade de Funcionários, variáveis fundamentais para o processo de clusterização. Por essa razão, optou-se pela exclusão de todos os registros incompletos nessas variáveis, de modo a assegurar a integridade e a consistência da análise posterior.

A segunda etapa consistiu na transformação dos dados, mais especificamente na aplicação de uma função logarítmica sobre os valores de faturamento. Essa técnica é amplamente utilizada em estudos que lidam com variáveis financeiras, dada a frequência de distribuições assimétricas e a presença de outliers severos — ou seja, empresas cujos faturamentos são excepcionalmente altos em comparação com a média da amostra (TAN; STEINBACH; KUMAR, 2019). Ao aplicar o logaritmo, reduz-se a amplitude dos valores, comprimindo as escalas e permitindo uma análise mais homogênea. Essa transformação é particularmente útil em algoritmos de agrupamento baseados em distância, como o *K-Means*, pois evita que empresas muito grandes exerçam influência desproporcional na definição dos centroides dos *clusters* (HAN; KAMBER; PEI, 2012). Trata-se, portanto, de uma etapa crucial para garantir que a

segmentação reflita padrões reais de similaridade e não apenas diferenças de ordem de magnitude.

Além disso, foi incorporada uma nova variável ao conjunto de dados, denominada Desempenho do Cliente. Essa variável não estava presente originalmente na base, mas foi simulada com valores entre 1 e 10, com o intuito de ilustrar o potencial analítico da introdução de métricas qualitativas na segmentação de clientes. Conforme salientam Kotler e Keller (2012), a avaliação do desempenho dos clientes deve considerar não apenas aspectos financeiros, mas também comportamentais e relacionais, como a regularidade nos pagamentos, engajamento com os serviços e feedbacks operacionais. Embora a métrica de desempenho utilizada neste estudo tenha caráter ilustrativo, sua inclusão representa uma boa prática na modelagem orientada ao cliente, permitindo a construção de estratégias mais personalizadas e eficientes. Com essas ações de limpeza, transformação e criação de variáveis adicionais, a base de dados passou a apresentar melhores condições para a aplicação de técnicas de clusterização, assegurando maior robustez estatística, coerência analítica e aplicabilidade prática aos resultados obtidos.

### 4.1.3 Análise estatística

Depois da etapa de preparação da base de dados, foi conduzida uma análise estatística exploratória com o objetivo de compreender, em maior profundidade, as características dos dados disponíveis antes da aplicação dos métodos de agrupamento. Conforme argumentam Han, Kamber e Pei (2012), a análise exploratória constitui uma etapa indispensável no processo de mineração de dados, pois permite identificar padrões, inconsistências, tendências e valores atípicos que podem comprometer a integridade dos resultados extraídos pelas técnicas posteriores.

O primeiro passo consistiu no cálculo de medidas estatísticas descritivas — como média, valor mínimo, máximo e amplitude — com ênfase nas variáveis de maior relevância para o estudo, a saber: faturamento presumido e quantidade de funcionários. Como apontam Fayyad et al. (1996), a obtenção de resumos estatísticos é uma prática fundamental para lidar com grandes volumes de dados, uma vez que facilita a interpretação inicial e auxilia na identificação de assimetrias e possíveis

distorções. A análise revelou que, embora a maior parte das empresas apresente faturamento em faixas intermediárias, algumas registram valores extremamente elevados, chegando a cifras bilionárias. Essa disparidade justifica a aplicação da transformação logarítmica ao faturamento, procedimento amplamente adotado em estudos financeiros com o intuito de minimizar a influência de *outliers* e promover uma distribuição mais equilibrada dos dados (TAN; STEINBACH; KUMAR, 2019).

Quanto à Quantidade de Funcionários, verificou-se uma heterogeneidade significativa entre as empresas analisadas, com registros que variam de microestruturas operacionais a grandes corporações. Essa diversidade reforça a necessidade de considerar múltiplos atributos no processo de segmentação, como defendido por Kotler e Keller (2012), que enfatizam a importância de reconhecer a pluralidade de perfis no ambiente B2B, onde diferentes portes organizacionais implicam necessidades e comportamentos comerciais distintos.

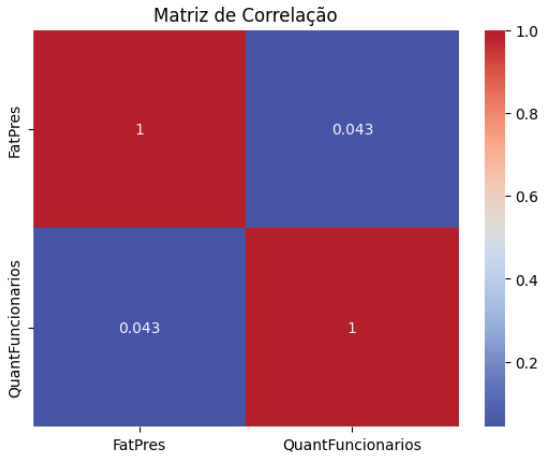
Para aprofundar a análise, foi examinada a correlação entre o faturamento e o porte das empresas, representado pela quantidade de colaboradores. Utilizou-se, para isso, o coeficiente de correlação de Pearson, ferramenta estatística indicada por Han, Kamber e Pei (2012) para avaliar a intensidade da associação linear entre variáveis numéricas. O valor obtido, próximo de 0,043, revelou uma correlação praticamente nula entre essas variáveis, indicando que o número de funcionários não é, por si só, um preditor direto do faturamento empresarial. Esse achado está em consonância com a realidade do mercado B2B, onde empresas enxutas em termos de pessoal — como firmas de consultoria ou tecnologia — podem apresentar faturamentos elevados, ao passo que organizações com grande número de funcionários, como prestadoras de serviços operacionais, podem operar com margens de receita mais modestas.

Para ilustrar visualmente essa constatação, foi construída uma matriz de correlação (figura 1), conforme recomendação de Tan, Steinbach e Kumar (2019). Essa ferramenta permite a representação gráfica da força de relação entre pares de variáveis. Na diagonal principal, observa-se sempre o valor 1, que representa a autocorrelação de cada variável consigo mesma. Fora da diagonal, o valor de 0,043 entre Faturamento e Funcionários reforça visualmente a ausência de relação direta entre essas variáveis.

Essa visualização evidencia que as variáveis analisadas não possuem dependência linear, o que é um indicativo importante para o processo de clusterização. Isso demonstra que não é possível, nem prudente, assumir que uma empresa com maior número de funcionários necessariamente gera mais receita — ou vice-versa. Tal constatação justifica a decisão metodológica de manter ambas as variáveis na modelagem, visto que cada uma oferece uma dimensão analítica distinta, agregando valor à identificação de padrões e à construção dos grupos.

Ao utilizar essas variáveis em conjunto no processo de agrupamento, o algoritmo pode captar nuances específicas do perfil organizacional dos clientes. Por exemplo, é possível identificar clusters compostos por empresas de alta receita e estrutura reduzida, como startups de base tecnológica, bem como grupos formados por organizações com muitos funcionários, mas com faturamento relativamente menor, como empresas do setor de serviços operacionais ou intensivos em mão de obra.

**Figura 12 — Matriz de correlação entre Faturamento e Quantidade de Funcionários**




Fonte: Adaptado pelo autor, 2025.

A integração de variáveis que capturam diferentes aspectos do perfil empresarial potencializa a eficácia da segmentação, tornando-a mais robusta e alinhada às exigências do mercado. Essa abordagem também se mostra coerente com os princípios defendidos por Kotler e Keller (2012), que argumentam que estratégias comerciais bem-sucedidas no ambiente B2B exigem uma compreensão holística do comportamento do cliente. Ao evitar reducionismos e considerar a

complexidade dos dados, aumenta-se a probabilidade de gerar grupos mais coerentes e úteis para a definição de ações estratégicas de vendas, prospecção e relacionamento.

### 4.1.4 Mineração de dados

Com a base de dados devidamente limpa, transformada e explorada estatisticamente, foi possível avançar para a etapa de mineração de dados, considerada uma das fases mais importantes, conforme proposto por Fayyad et al. (1996). Esta fase é responsável por extrair padrões úteis e estruturados a partir de grandes volumes de dados, sendo particularmente relevante em contextos empresariais que visam gerar inteligência competitiva. No escopo desta pesquisa, a mineração de dados tem como finalidade identificar padrões ocultos no perfil dos clientes empresariais, possibilitando a criação de estratégias mais personalizadas e eficazes de relacionamento comercial, *marketing* e vendas. Para isso, recorre-se ao uso de técnicas de agrupamento — também chamadas de *clustering*.

Dentre os diversos algoritmos de agrupamento disponíveis, optou-se pelo uso do *KMeans*, amplamente reconhecido na literatura por sua eficácia na segmentação de dados numéricos contínuos, simplicidade conceitual e rapidez de execução (TAN; STEINBACH; KUMAR, 2019). A ausência da necessidade de rótulos prévios torna o *KMeans* especialmente adequado para ambientes em que os dados não foram previamente classificados — como é o caso de muitas bases comerciais reais — permitindo a descoberta de estruturas latentes com autonomia.

Além do seu rigor matemático, o *KMeans* se destaca por ser um dos métodos mais acessíveis em termos computacionais, podendo ser executado com eficiência mesmo em bases de grande porte. Isso o torna uma ferramenta altamente viável para ser utilizada por equipes comerciais e de *marketing* que, muitas vezes, não dispõem de suporte técnico contínuo. Outro diferencial relevante está na clareza dos seus resultados, que facilita a interpretação e aplicação prática dos clusters identificados, característica essencial quando se busca utilizar a análise de dados como ferramenta de apoio à tomada de decisão estratégica (KOTLER; KELLER, 2012).

Durante o delineamento metodológico deste trabalho, outras técnicas também foram consideradas, a fim de assegurar que a escolha do algoritmo mais adequado

fosse pautada em critérios de coerência com os objetivos da pesquisa. A Análise Fatorial, por exemplo, é frequentemente utilizada para redução de dimensionalidade, agrupando variáveis correlacionadas em componentes principais e facilitando a visualização e interpretação de grandes conjuntos de dados (HAIR et al., 2009). Contudo, como este estudo concentrou-se em um número propositalmente reduzido de variáveis — especificamente o Faturamento e a Quantidade de Funcionários — optou-se por não aplicar métodos de redução, preservando a interpretação direta dos clusters gerados a partir dessas variáveis brutas.

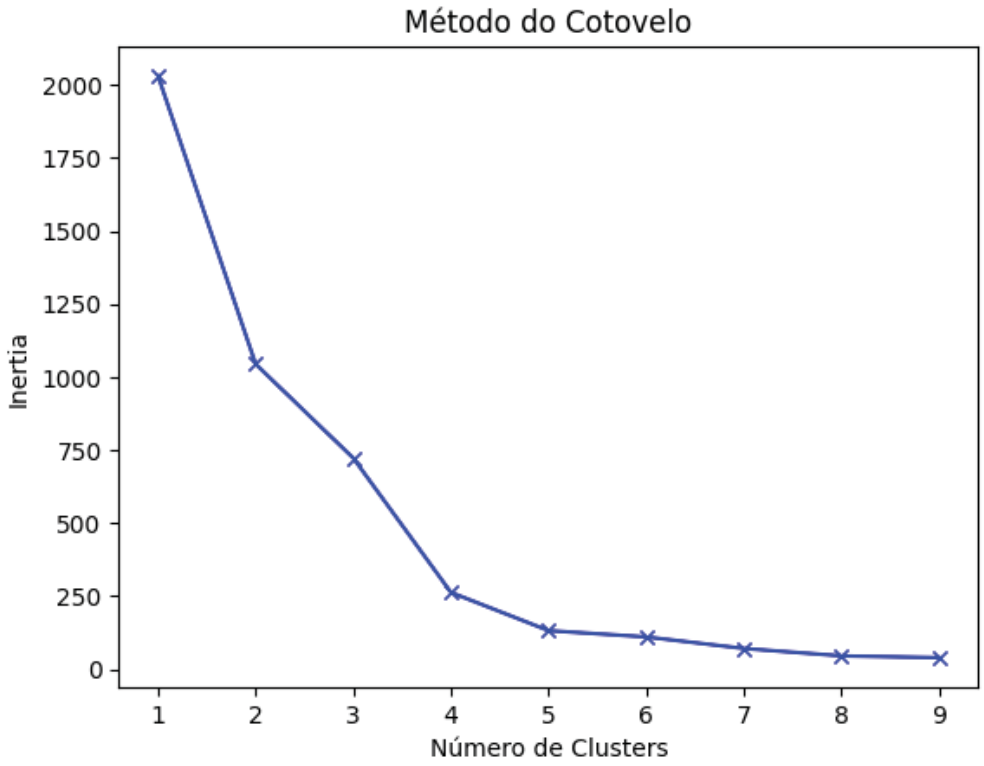
De forma semelhante, algoritmos supervisionados como Árvore de Decisão e *Random Forest* são frequentemente utilizados em tarefas preditivas, nas quais há um atributo de interesse (variável-alvo) conhecido e rotulado (HAN; KAMBER; PEI, 2012). Porém, como a intenção deste trabalho não é prever um resultado específico, mas sim descobrir padrões naturais de agrupamento entre empresas, esses métodos supervisionados não se mostraram apropriados ao problema em questão. Avaliaram-se ainda abordagens mais sofisticadas, como os algoritmos inspirados em inteligência de enxames — por exemplo, a Otimização por Colônia de Formigas (*Ant Colony Optimization* – ACO) — que, apesar de sua eficácia em contextos de alta complexidade, apresentam custo computacional elevado e exigem parametrização cuidadosa para obter soluções estáveis. Tais requisitos dificultam sua aplicação em ambientes empresariais rotineiros, em que a simplicidade operacional e a rapidez na obtenção de *insights* são características decisivas (TAN; STEINBACH; KUMAR, 2019).

Dentro desses aspectos, a escolha do algoritmo *K-Means* demonstrou-se a mais adequada tanto do ponto de vista técnico quanto do prático, conciliando rigor analítico, velocidade de processamento e usabilidade. Essa decisão metodológica está em consonância com a proposta deste estudo, que busca entregar uma solução robusta e aplicável para segmentação de clientes no contexto B2B, promovendo uma análise baseada em dados com potencial de impacto direto nas estratégias comerciais da organização.

Diante das comparações realizadas entre diferentes abordagens, o algoritmo *K-Means* foi definitivamente escolhido como a técnica central de agrupamento para este trabalho, devido ao seu equilíbrio entre simplicidade operacional, eficiência computacional, qualidade dos agrupamentos gerados e clareza dos resultados. Para

determinar o número ótimo de clusters a ser utilizado no algoritmo, foi aplicado o método do cotovelo (*Elbow Method*), amplamente recomendado por Han, Kamber e Pei (2012) como uma das formas mais eficazes de validar a quantidade de agrupamentos em cenários não supervisionados. Essa técnica consiste em calcular a soma das distâncias quadráticas dentro dos clusters (inércia intra-cluster) para diferentes valores de kkk e observar em qual ponto o ganho marginal na redução dessa inércia se torna pouco expressivo — formando um “cotovelo” na curva, o que indica o número ideal de clusters para balancear qualidade da segmentação e parcimônia interpretativa.

**Figura 13 — Gráfico do método do cotovelo para definição do número de clusters (dados brutos).**




Fonte: Adaptado pelo autor, 2025.

A análise inicial foi realizada utilizando os dados de Faturamento na forma original, sem transformações. A figura 13 apresenta o gráfico gerado nessa etapa, em que se nota uma queda acentuada nos primeiros valores de kkk, sinalizando que o algoritmo é eficaz em reduzir a variabilidade dentro dos grupos à medida que mais clusters são adicionados. No entanto, observa-se também que, a partir de determinado ponto, essa taxa de redução desacelera consideravelmente, indicando que a adição de novos agrupamentos não oferece ganhos substanciais na compactação dos dados. Esse comportamento é típico em bases com alta variabilidade interna (HAN; KAMBER; PEI, 2012), como aquelas compostas por registros financeiros empresariais, e reforça a aplicabilidade do método do cotovelo como instrumento diagnóstico.

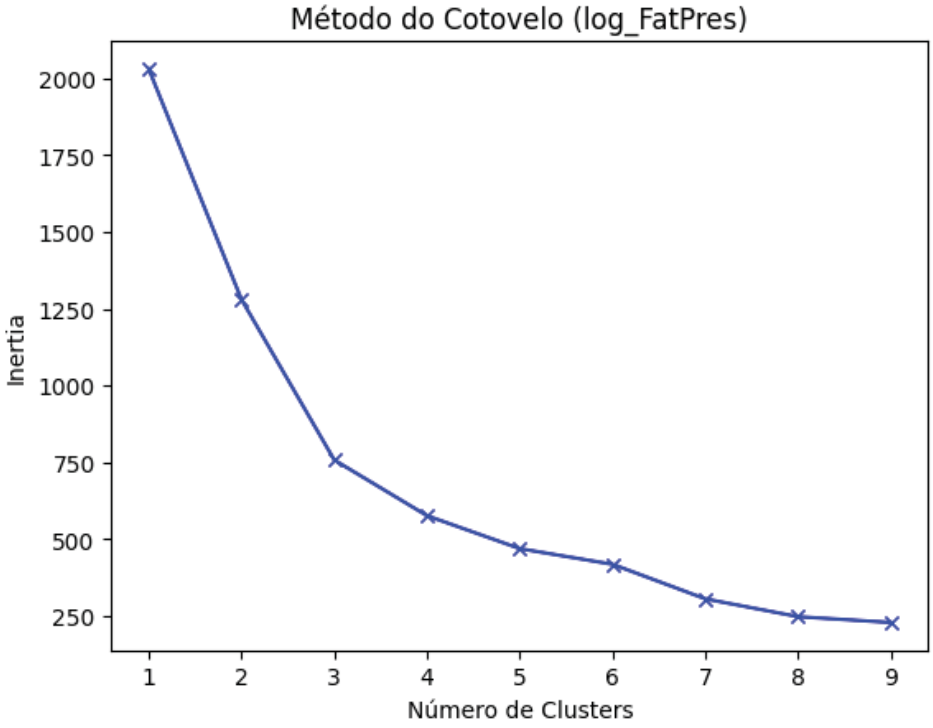
Apesar disso, um obstáculo importante emergiu nessa primeira análise: a presença de valores extremos de faturamento (*outliers*), bastante comuns em bases do tipo B2B, especialmente quando há empresas de grande porte inseridas no mesmo conjunto que pequenas e médias organizações. Esses valores fora da curva distorcem a distribuição e afetam diretamente a métrica de inércia, fazendo com que o gráfico do cotovelo perca definição e dificulte a visualização do ponto de inflexão exato. Conforme descrevem Han, Kamber e Pei (2012), esse fenômeno é recorrente em bases com ampla dispersão numérica e exige estratégias de tratamento específico, como normalização ou transformação de escala.

Como solução, foi adotada a transformação logarítmica da variável Faturamento, prática consagrada na literatura estatística para lidar com distribuições assimétricas e escalas amplas, especialmente em dados financeiros (TAN; STEINBACH; KUMAR, 2019). Ao aplicar o logaritmo, as diferenças entre os valores se comprimem, reduzindo a influência de outliers e equilibrando a contribuição dos dados para o cálculo da inércia. A reaplicação do método do cotovelo com a nova variável transformada está ilustrada na figura 14. Observa-se, neste novo gráfico, que a curva se torna mais suave e o ponto de inflexão mais nítido, permitindo identificar com maior segurança o número ótimo de clusters a ser utilizado. Essa transformação, portanto, não apenas melhora a qualidade estatística da análise como também fortalece sua robustez metodológica, eliminando ruídos causados por distorções extremas na escala de Faturamento.

Complementarmente, para reforçar a escolha do número de agrupamentos e validar visualmente a coerência dos clusters gerados, foram elaborados gráficos de dispersão, conforme recomendação de Han, Kamber e Pei (2012), que destacam a importância da visualização como recurso para validar padrões de agrupamento e comunicar resultados de maneira acessível a públicos não técnicos. A figura 13 apresenta o gráfico de dispersão elaborado com os dados originais de Faturamento versus Quantidade de Funcionários, e nela é possível perceber uma forte concentração de pontos em uma faixa estreita, com alguns registros distantes, à direita do plano, representando empresas de altíssimo faturamento. Essa compressão compromete a clareza da visualização e pode obscurecer os agrupamentos reais.

Em resposta a essa limitação, foi gerado um novo gráfico de dispersão, agora com o Faturamento transformado logaritmicamente, conforme apresentado na figura 5. Essa modificação melhora substancialmente a distribuição visual dos dados, permitindo observar com mais nitidez como os pontos se organizam no espaço bidimensional e, consequentemente, como os *clusters* se definem. Essa abordagem confirma, de maneira empírica e visual, que a transformação logarítmica não apenas aprimora os resultados do método do cotovelo, mas também potencializa a capacidade do *K-Means* de formar grupos mais bem definidos, coerentes e aderentes à realidade mercadológica. Ao reduzir os efeitos dos extremos, a análise torna-se mais representativa da distribuição da maioria das empresas da base, possibilitando a formulação de estratégias comerciais mais precisas e contextualizadas.

**Figura 14 — Gráfico do método do cotovelo com transformação logarítmica do Faturamento.**




Fonte: Adaptado pelo autor, 2025.

A análise do gráfico do método do cotovelo após a aplicação da transformação logarítmica na variável Faturamento revelou uma mudança significativa no comportamento da curva. Ao suavizar a distribuição dos dados, a transformação eliminou distorções causadas por valores extremamente elevados de algumas empresas, que anteriormente exerciam influência desproporcional sobre os cálculos de inércia *intra-cluster*. Como ressaltam Tan, Steinbach e Kumar (2019), esse tipo de transformação é altamente recomendado em contextos de análise financeira, pois permite uma melhor estabilização da variância e viabiliza uma interpretação mais precisa da estrutura latente dos dados. Como resultado, o ponto de inflexão — que indica a quantidade ótima de agrupamentos — tornou-se mais nítido, facilitando sua identificação e, consequentemente, aumentando a confiabilidade do modelo de segmentação (HAN; KAMBER; PEI, 2012).

Paralelamente à análise numérica proporcionada pelo método do cotovelo, recorreu-se a gráficos de dispersão como técnica complementar de validação visual dos agrupamentos. Essa prática é incentivada por Han, Kamber e Pei (2012), que reconhecem a importância das representações gráficas como ferramentas essenciais para avaliar a coesão e a separabilidade dos clusters formados, sobretudo quando o

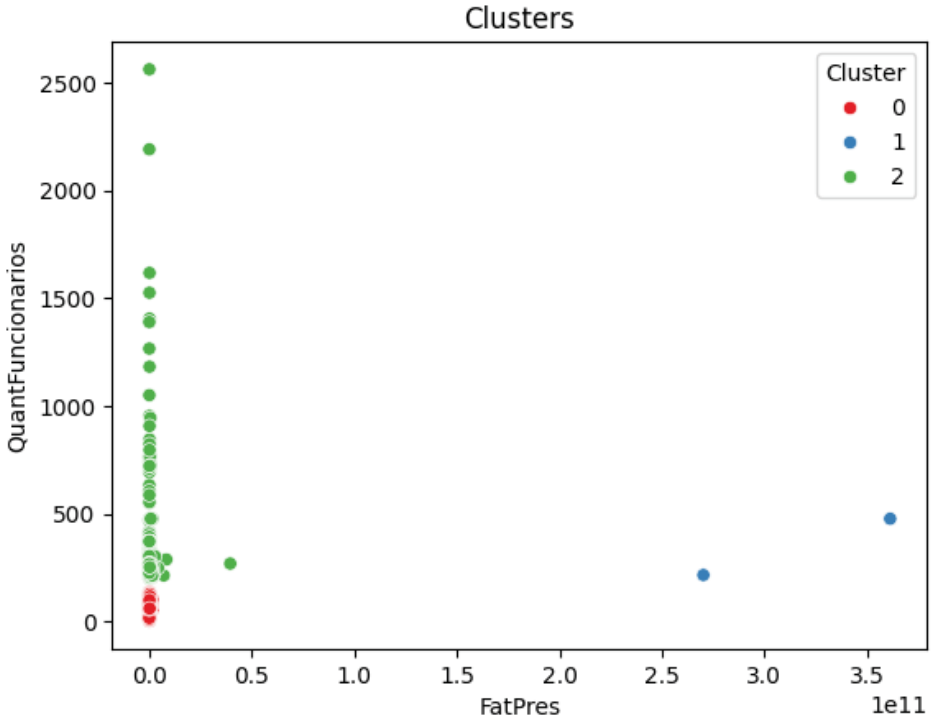
objetivo é comunicar os achados a públicos diversos, incluindo gestores e tomadores de decisão não especializados em ciência de dados. Os gráficos de dispersão facilitam a observação intuitiva dos padrões de distribuição, mostrando como os registros se posicionam em relação às variáveis principais — neste estudo, Faturamento e Quantidade de Funcionários.

A figura 14 apresenta o gráfico de dispersão construído com os dados de faturamento em sua escala original. Nota-se uma alta concentração de pontos próximos à origem do plano cartesiano, o que indica que a maioria das empresas possui faturamentos relativamente baixos. No entanto, observa-se também a presença de pontos isolados e muito distantes no eixo horizontal, correspondentes a empresas com faturamentos excepcionalmente elevados. Essa disparidade gera uma compressão visual dos dados, dificultando a identificação clara dos agrupamentos e comprometendo a análise visual da distribuição dos clientes.

Para contornar essa limitação e aprimorar a qualidade da visualização, foi gerado um novo gráfico de dispersão com o Faturamento transformado logaritmicamente, conforme orientações metodológicas de Tan, Steinbach e Kumar (2019). A figura 5 exibe os resultados dessa abordagem, evidenciando uma distribuição muito mais homogênea dos dados no espaço bidimensional. Com a compressão da escala, os pontos passam a se posicionar de forma mais equilibrada, permitindo visualizar com maior nitidez os contornos de cada cluster. Essa clareza reforça a qualidade do agrupamento gerado pelo algoritmo *KMeans*, que agora opera sobre uma base de dados menos assimétrica e mais representativa da realidade mercadológica.

Portanto, a aplicação da transformação logarítmica, tanto na análise do método do cotovelo quanto na visualização por dispersão, revelou-se uma estratégia metodológica eficaz para lidar com a natureza desigual dos dados financeiros empresariais. A melhora na definição dos clusters não apenas contribui para a robustez da modelagem, mas também facilita sua aplicação prática, permitindo que os resultados gerados orientem decisões comerciais mais precisas e segmentadas.

**Figura 15 — Distribuição dos clusters considerando Faturamento original e Quantidade de Funcionários.**




Fonte: Adaptado pelo autor, 2025.

Ao analisar a figura 15, observa-se uma forte concentração de pontos próximos ao valor zero no eixo de Faturamento, com apenas um pequeno número de empresas posicionadas mais à direita do gráfico, evidenciando valores de faturamento consideravelmente elevados. Esse tipo de distribuição desigual é característico de bases de dados empresariais, especialmente em contextos B2B, nos quais um número reduzido de grandes contas concentra a maior parte da receita da empresa, enquanto a maioria dos clientes possui faturamentos mais modestos (HAN; KAMBER; PEI, 2012). Essa assimetria severa compromete a utilidade do gráfico de dispersão original, pois os dados da maior parte dos clientes ficam comprimidos em uma faixa muito estreita, dificultando a distinção de perfis e a visualização de possíveis agrupamentos.

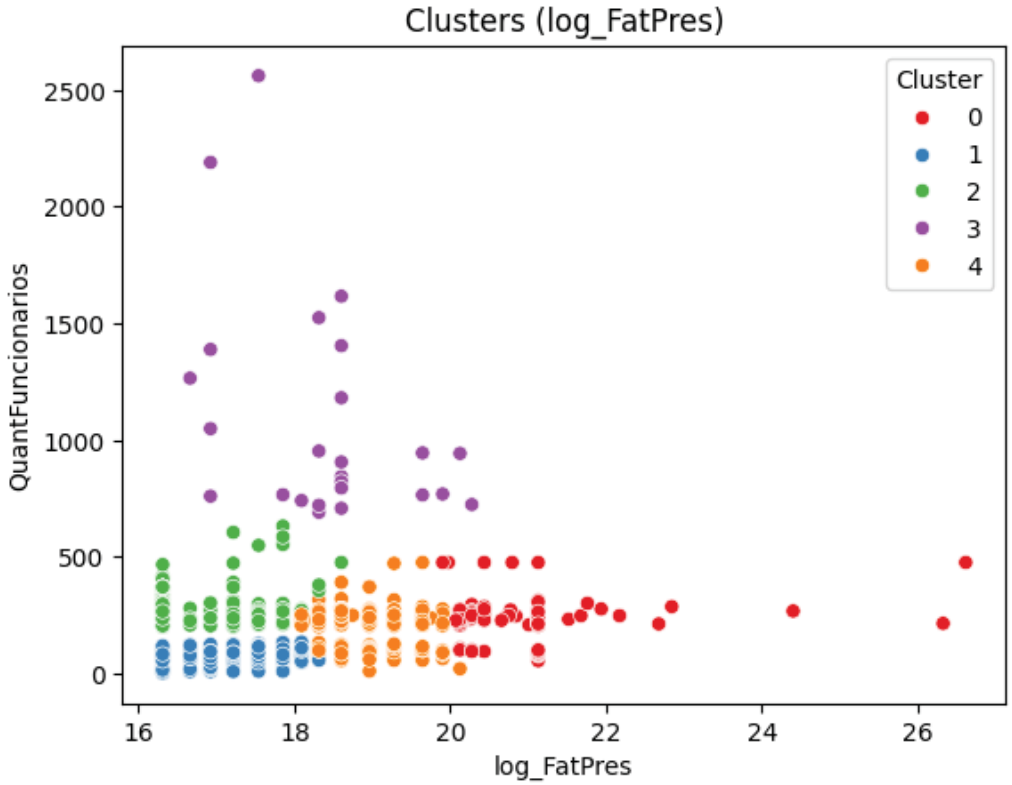
Para mitigar essa distorção e permitir uma análise mais clara e representativa, foi aplicada uma transformação logarítmica na variável faturamento. Tal estratégia é amplamente recomendada na literatura especializada como forma eficaz de lidar com variáveis altamente assimétricas e de atenuar a influência de outliers (TAN; STEINBACH; KUMAR, 2019). A figura 5, que apresenta o gráfico de dispersão com o Faturamento já transformado, evidencia uma distribuição mais equilibrada ao longo

do eixo horizontal. Os pontos agora se espalham de maneira mais homogênea, o que facilita não apenas a percepção visual dos clusters, mas também melhora os cálculos de distância realizados pelo algoritmo *KMeans*, resultando em agrupamentos mais coerentes e consistentes com a realidade de mercado.

Além dos ganhos técnicos, a transformação logarítmica também contribui para a clareza da comunicação dos resultados, especialmente quando apresentados a públicos não técnicos. O gráfico com a escala ajustada oferece uma representação visual mais acessível e intuitiva, permitindo que gestores e tomadores de decisão compreendam facilmente as justificativas adotadas nas etapas de pré-processamento, como defendido por Han, Kamber e Pei (2012). Essa abordagem favorece a aceitação prática do modelo e fortalece sua aplicação no ambiente corporativo.

Portanto, a comparação entre os gráficos de dispersão com o Faturamento em escala original (figura 16) e transformada (figura 15) comprova que o uso do logaritmo foi uma decisão metodológica essencial para aprimorar a qualidade da clusterização. A transformação permitiu que o algoritmo detectasse padrões mais representativos da diversidade empresarial, ao mesmo tempo que facilitou a visualização e interpretação dos dados. Com isso, os objetivos da segmentação — identificar grupos de clientes mais precisos, úteis e alinhados às estratégias comerciais — foram alcançados com maior eficácia.

**Figura 16 — Distribuição dos clusters considerando Faturamento transformado em log e Quantidade de Funcionários.**




Fonte: Adaptado pelo autor, 2025.

A análise dos gráficos de dispersão comprova, de forma clara e objetiva, que a combinação do algoritmo *K-Means* com a transformação logarítmica da variável Faturamento constitui uma estratégia altamente eficaz para a segmentação de clientes em grupos mais homogêneos. Essa abordagem equilibra a distribuição dos dados e revela padrões que seriam mascarados por valores extremos, conforme ressaltam Han, Kamber e Pei (2012) e Tan, Steinbach e Kumar (2019). Ao suavizar as discrepâncias provocadas por grandes outliers, a transformação permite ao algoritmo formar clusters mais representativos da realidade empresarial.

A aplicação conjunta dessas técnicas assegura que o agrupamento final reflita com maior fidelidade a diversidade dos perfis de clientes, oferecendo uma leitura mais justa tanto para pequenas empresas quanto para grandes contas estratégicas. Como destacam Kotler e Keller (2012), a clareza na definição dos segmentos é essencial para que as áreas comerciais e de *marketing* possam alinhar suas ações ao potencial de cada grupo, promovendo maior eficácia nas estratégias de prospecção, relacionamento e fidelização.

Adicionalmente, a visualização gráfica dos clusters favorece a compreensão dos resultados por gestores e equipes operacionais que não possuem formação técnica, tornando a segmentação uma ferramenta acessível e prática no apoio ao planejamento comercial. Essa acessibilidade permite, por exemplo, a personalização de ofertas e a alocação mais inteligente de recursos, garantindo que o esforço comercial seja concentrado nos clusters com maior potencial de receita e valor de relacionamento ao longo do tempo. Dessa forma, consolida-se uma atuação orientada por dados, alinhada à estratégia de negócios e voltada à maximização do retorno sobre os investimentos realizados (KOTLER; KELLER, 2012).

### 4.1.5 Simulação de métricas de negócio

Para complementar a análise técnica dos clusters e estabelecer uma conexão direta e prática entre a segmentação de clientes e os indicadores estratégicos fundamentais para a gestão comercial em ambientes B2B, este estudo realizou uma simulação aplicada de duas métricas amplamente reconhecidas e utilizadas no contexto corporativo: o CAC e o LTV. Essas métricas são essenciais para a compreensão da eficiência dos investimentos comerciais e para a formulação de estratégias que maximizem o retorno sobre o capital aplicado, conforme destacado por Kotler e Keller (2012).

O LTV representa o montante financeiro estimado que uma empresa pode gerar ao longo de todo o relacionamento com um cliente, sendo uma métrica crucial para avaliar a lucratividade potencial de contas individuais ou segmentos específicos. Considerando a ausência de dados históricos detalhados, como duração exata do relacionamento ou taxas de *churn*, optou-se por uma abordagem prática e simplificada para a estimativa do LTV. Neste estudo, o LTV foi estimado como 120% do faturamento atual de cada cliente, simulando cenários comuns e realistas do mercado B2B, que envolvem renovações contratuais, vendas adicionais (*upsell*) e vendas cruzadas (*cross-sell*). Essa metodologia está alinhada às orientações de Stone e Woodcock (2014), que recomendam a adaptação do cálculo de LTV às características e limitações das bases de dados disponíveis.

É importante ressaltar que, idealmente, o LTV deve incorporar fatores dinâmicos como a duração do ciclo de vida do cliente, frequência e recorrência de

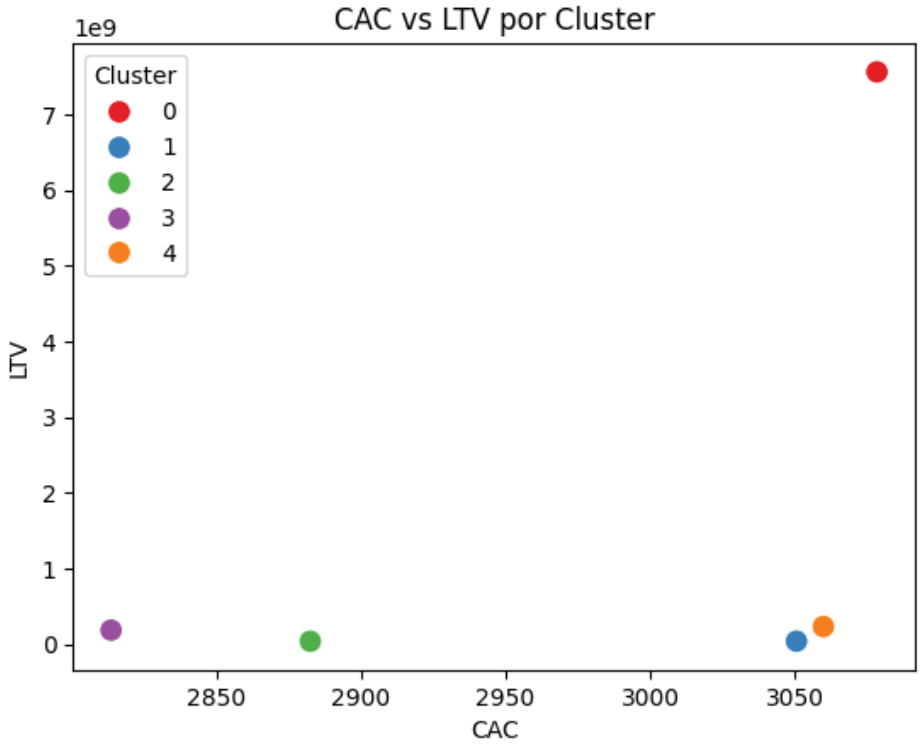
compras, e comportamento de fidelização. No entanto, dada a restrição de dados históricos detalhados, a simulação percentual adotada aqui possibilitou comparações realistas e consistentes entre os clusters formados, oferecendo uma perspectiva relativa e prática sobre o valor potencial de cada grupo.

Por sua vez, o CAC foi estimado com base em uma média representativa dos custos envolvidos na aquisição de cada cliente. Essa média contemplou despesas típicas do contexto B2B, como investimentos em *marketing*, deslocamentos, salários e comissões da equipe comercial, bem como custos operacionais associados à preparação e apresentação de propostas comerciais. Tal estimativa está em conformidade com a definição de Kotler e Keller (2012), que enfatizam que o CAC deve refletir o conjunto de investimentos necessários para converter um prospect em cliente efetivo, especialmente em processos de vendas consultivas e complexas, características marcantes do ambiente B2B.

Com essas duas métricas simuladas para cada cliente da base, foi possível calcular as médias de LTV e CAC por cluster, criando assim um panorama comparativo detalhado da rentabilidade relativa de cada segmento. Essa análise comparativa é fundamental para identificar quais clusters apresentam o equilíbrio mais favorável entre o custo de aquisição e o retorno financeiro esperado, subsidiando decisões estratégicas de alocação de recursos, otimização do funil comercial e priorização dos esforços de *marketing* e vendas (Stone; Woodcock, 2014).

A figura 16 ilustra um gráfico de dispersão que posiciona cada cluster de acordo com seus valores médios de CAC e LTV. Cada ponto representa um cluster distinto, permitindo uma visualização clara e imediata dos grupos mais atrativos — caracterizados por um alto LTV associado a um CAC controlado — e daqueles com baixo retorno financeiro combinado a custos de aquisição elevados, que indicam possíveis ineficiências e desperdícios operacionais.

**Figura 17 — Relação entre CAC e LTV médios por cluster.**




Fonte: Adaptado pelo autor, 2025.

Observa-se, na figura 16, que determinados clusters se destacam por apresentar um LTV médio elevado, mesmo mantendo CACs próximos ou abaixo da média geral. Essa constatação valida a premissa de Kotler e Keller (2012) de que clientes de alto valor, embora possam demandar investimentos iniciais maiores, compensam amplamente esses custos por meio de ciclos de compra mais longos, maior fidelidade e ticket médio elevado. Esses clusters configuram-se como contas estratégicas prioritárias, que justificam a implementação de ações intensivas e personalizadas de relacionamento, incluindo atendimento dedicado, consultorias especializadas e propostas sob medida.

Em contrapartida, a análise também evidenciou clusters com LTV relativamente baixo, mas que apresentam CACs similares aos grupos mais rentáveis. Conforme alertam Stone e Woodcock (2014), essa situação representa um risco operacional significativo: a alocação de recursos comerciais em clientes de baixa rentabilidade pode gerar sobrecarga da força de vendas, reduzir a eficiência operacional e comprometer o retorno global dos investimentos comerciais. Tal cenário reforça a importância de diferenciar os níveis de atendimento e investimento segundo o potencial econômico real de cada cluster.

Essa leitura prática e visual do gráfico permite a definição de critérios objetivos para priorização e alocação de esforços comerciais. *Clusters* com alto LTV e CAC sob controle devem ser acompanhados por executivos de contas especializados, programas de fidelização e estratégias de *upsell* e *cross-sell* estruturadas. Já os grupos menos rentáveis demandam modelos de atendimento escaláveis, automatizados e de baixo custo, como canais digitais, suporte remoto e propostas padronizadas, garantindo cobertura comercial eficiente sem comprometer a rentabilidade.

Em síntese, a análise conjunta de LTV e CAC não apenas valida a qualidade técnica da clusterização, mas sobretudo traduz os resultados em recomendações estratégicas de aplicação imediata e prática. Essa integração entre mineração de dados e gestão comercial assegura maior racionalidade e eficiência na alocação de recursos, aprimora o retorno sobre investimentos e contribui diretamente para o aumento sustentável da lucratividade da base de clientes (Stone; Woodcock, 2014; Kotler; Keller, 2012).

## 4.2 FERRAMENTAS E *SOFTWARES*

As etapas de processamento, análise e modelagem dos dados serão implementadas por meio da linguagem de programação *Python*, amplamente adotada em projetos de ciência de dados devido à sua versatilidade, robustez e vasta gama de bibliotecas especializadas. A manipulação de dados tabulares e estruturas matriciais será realizada com o suporte das bibliotecas Pandas e *NumPy*, permitindo uma organização eficiente dos dados e facilitando operações estatísticas, agregações e transformações. Para a identificação e visualização de valores ausentes, será utilizada a biblioteca *Missingno*, que oferece representações gráficas intuitivas para apoiar decisões sobre imputações ou exclusões.

Na etapa de modelagem preditiva e segmentação, o *framework Scikit-learn* desempenhará papel central, sendo responsável pela aplicação de algoritmos de clusterização (como *K-Means*), classificação supervisionada (como *Random Forest* e Árvores de Decisão), além de ferramentas de pré-processamento (normalização, escalonamento e codificação) e validação cruzada de modelos. Para a redução de dimensionalidade e análise fatorial, será empregada a biblioteca *FactorAnalyzer*, que

permite a extração de componentes principais e a avaliação da adequação das variáveis aos fatores latentes.

A visualização gráfica dos resultados será conduzida com o auxílio das bibliotecas *Matplotlib* e *Seaborn*, que oferecem recursos avançados para gerar gráficos de dispersão, boxplots, mapas de calor, histogramas e outras representações úteis para análise exploratória e apresentação dos achados. Por fim, a modelagem preditiva baseada em algoritmos genéticos será realizada por meio do *AntMiner+*, técnica que integra princípios de inteligência coletiva e evolução computacional, proporcionando classificações interpretáveis e eficazes, especialmente em cenários de regras de decisão complexas.

## 4.3 LIMITAÇÕES METODOLÓGICAS

O presente estudo apresenta algumas limitações que devem ser consideradas na interpretação e generalização dos resultados. A principal restrição está relacionada à natureza secundária dos dados utilizados, os quais, apesar de representativos, podem não abranger com exatidão todas as particularidades do mercado em análise. Essa limitação compromete, em certa medida, a profundidade da inferência sobre comportamentos específicos ou dinâmicas emergentes em determinados nichos. Além disso, a aplicabilidade prática dos resultados obtidos está condicionada à relativa estabilidade dos padrões históricos de comportamento das empresas, o que pode ser impactado por mudanças econômicas, variações setoriais ou transformações estruturais nas estratégias de consumo e relacionamento entre empresas. Outro fator que merece destaque é a complexidade interpretativa de alguns modelos analíticos empregados, como as redes neurais artificiais, cuja natureza de caixa-preta pode dificultar a explicação dos critérios de segmentação para gestores não técnicos. Essa característica pode gerar resistência organizacional à adoção de abordagens baseadas em ciência de dados, especialmente em ambientes empresariais mais tradicionais, nos quais a tomada de decisão ainda se baseia fortemente em heurísticas e experiências acumuladas.

A base de dados utilizada na pesquisa é denominada *base_leads_ok.xlsx*, a qual contém registros reais de empresas que mantêm ou mantiveram algum tipo de relação comercial com uma prestadora de serviços de cobrança. Conforme

argumentam Han, Kamber e Pei (2012), a qualidade, consistência e relevância dos dados de entrada são determinantes para o êxito de qualquer projeto de mineração de dados, justificando a escolha de uma base já consolidada, revisada e organizada. Essa escolha também está em consonância com as orientações de Fayyad et al. (1996) sobre a importância da preparação de dados no ciclo KDD, etapa crítica que antecede a análise propriamente dita. A base em questão reúne aproximadamente 1.000 registros, sendo que cada linha representa uma empresa única. Os dados incluem tanto informações cadastrais básicas, como razão social e setor de atuação, quanto variáveis de maior relevância analítica, como o Faturamento Presumido *(*FatPres*)* — indicador do volume financeiro movimentado pela empresa — e a *Quantidade de Funcionários*, que permite inferir o porte organizacional.

Complementarmente, a base contempla variáveis auxiliares, como segmento de mercado, localização geográfica, risco de inadimplência e status de atividade, que foram utilizadas de forma exploratória na etapa inicial do estudo. Essas informações adicionais contribuíram para verificar a consistência dos clusters gerados, bem como a sua capacidade de representar grupos economicamente e operacionalmente distintos. Em linha com as contribuições de Kotler e Keller (2012), compreender o tamanho, a complexidade e o potencial de consumo de cada cliente é um passo fundamental para qualquer iniciativa de segmentação orientada a resultados, especialmente no contexto B2B, onde os volumes transacionais e os ciclos de relacionamento tendem a ser mais longos. Ao optar por uma base realista e alinhada ao mercado de atuação da empresa em questão, o estudo assegura maior aplicabilidade dos seus achados. Dessa forma, os clusters resultantes da análise podem ser efetivamente utilizados como subsídio à atuação da equipe comercial, permitindo a personalização de estratégias de prospecção, abordagem e retenção de acordo com o perfil identificado de cada segmento (KOTLER; KELLER, 2012; OLIVEIRA, 2018).

## 5 RESULTADOS E DISCUSSÃO

A implementação da metodologia proposta de segmentação de clientes B2B, fundamentada em variáveis de lucratividade, possibilitou a identificação de padrões relevantes nos dados analisados, viabilizando uma classificação mais estratégica da

base de clientes. Após o pré-processamento, normalização e transformação dos dados, aplicou-se o algoritmo de clusterização *K-Means*, que revelou a existência de quatro agrupamentos distintos (clusters) como estrutura ótima de segmentação.

A definição do número ideal de clusters foi embasada no método do cotovelo, que indicou redução significativa da variância *intra-cluster* até o quarto agrupamento. Embora a pontuação de silhueta mais alta tenha sido observada em uma configuração com dois clusters (valor de 0,85), optou-se por manter quatro agrupamentos a fim de preservar um nível maior de granularidade analítica e interpretabilidade gerencial, permitindo ações mais direcionadas por perfil de cliente.

O processo de segmentação teve como base variáveis quantitativas relacionadas ao desempenho comercial, tais como faturamento, CAC, LTV e número de funcionários, variáveis estas previamente validadas em termos de correlação e relevância estatística. A análise resultante permitiu agrupar empresas com características similares em termos de comportamento de compra, retorno financeiro e complexidade de relacionamento.

A Tabela a seguir sintetiza os principais atributos médios de cada grupo identificado, proporcionando uma visão comparativa dos perfis de cliente formados.

O Cluster 1 reúne empresas com alto faturamento, elevado número de funcionários, baixo CAC e alto LTV, configurando o grupo mais atrativo do ponto de vista estratégico. Esses clientes não apenas geram elevada receita ao longo do tempo, como também demandam menor esforço de aquisição e manutenção. Representam, portanto, ativos valiosos a serem preservados, sendo altamente recomendável a aplicação de estratégias de fidelização, personalização de relacionamento e expansão de serviços. Esses clientes concentram maior valor financeiro acumulado e devem ser foco de atenção contínua da equipe de contas estratégicas.

**Figura 18 – Características dos clusters identificados**

| Cluster | Faturamento (R$) | CAC Médio (R$) | LTV Médio (R$) | Nº Funcionários | Interpretação Estratégica |
|---|---|---|---|---|---|
| 1 | Alto | Baixo | Alto | Elevado | Clientes estratégicos prioritários |
| 2 | Médio | Médio-Alto | Médio-Alto | Médio | Potencial para upsell |
| 3 | Baixo | Alto | Baixo | Baixo | Baixa atratividade e alto risco |
| 4 | Variável | Variável | Médio | Médio-Alto | Clientes em observação |

Fonte: Adaptado pelo autor, 2025.

O Cluster 2, por sua vez, é composto por empresas de faturamento médio e engajamento elevado, mesmo com um CAC levemente superior à média. O LTV também se apresenta em nível satisfatório, sinalizando que esse grupo tem forte potencial de crescimento por meio de estratégias de *upsell* e *cross-sell*. A proximidade com o perfil ideal de cliente sugere que intervenções relativamente simples — como ajustes em propostas comerciais, ofertas direcionadas ou acompanhamento consultivo — podem acelerar o aumento do valor gerado por esse segmento.

O *Cluster* 3 é considerado o grupo menos vantajoso, reunindo empresas com baixo faturamento, baixo número de funcionários, elevado CAC e LTV reduzido. Esse segmento representa um risco de baixa rentabilidade, já que os custos de aquisição e manutenção superam ou se aproximam perigosamente da receita gerada ao longo do tempo. Recomenda-se a revisão da estratégia de atendimento a esse grupo, com análise de viabilidade para redirecionamento de recursos ou aplicação de modelos operacionais mais enxutos, como automação de processos, autoatendimento ou despriorização comercial.

Já o Cluster 4 apresentou características mais heterogêneas, com variações expressivas em seus indicadores de performance. Algumas empresas desse grupo demonstraram comportamento próximo ao Cluster 2, enquanto outras se assemelham ao *Cluster* 3. Esse agrupamento foi classificado como "clientes em observação", exigindo monitoramento contínuo e coleta de dados adicionais para melhor definição de seu papel estratégico. A presença de empresas promissoras dentro deste cluster sugere a oportunidade de ações pontuais de qualificação e análise de comportamento ao longo do tempo, com potencial de migração para segmentos superiores.

Durante a etapa de análise estatística, identificou-se uma correlação positiva relevante entre o faturamento das empresas e o número de funcionários, o que validou a escolha dessas variáveis como componentes centrais da segmentação. Essa associação é coerente com o contexto B2B, onde organizações maiores tendem a apresentar maior capacidade de consumo e possibilidade de estabelecer contratos de maior valor agregado.

Essa evidência reforça a hipótese de que indicadores estruturais, como porte e faturamento, são variáveis-chave para prever comportamento futuro e valor potencial

dos clientes. Isso também respalda a ideia de que modelos preditivos baseados em atributos financeiros e operacionais podem ser mais eficazes do que abordagens exclusivamente demográficas ou setoriais.

Um dos principais diferenciais da metodologia aplicada foi a possibilidade de avaliar a razão entre LTV e CAC por cluster. Essa métrica, amplamente reconhecida como indicador de viabilidade financeira da carteira, serviu como filtro de priorização estratégica. Os clusters em que o LTV superava o CAC em mais de três vezes foram considerados altamente rentáveis, justificando investimento contínuo em ações de relacionamento e desenvolvimento.

Por outro lado, clusters em que a razão LTV/CAC foi inferior a 1,5 indicaram possível desequilíbrio entre esforço e retorno, requerendo revisão de canais de aquisição, abordagem comercial, ou mesmo questionamento da manutenção desses clientes na base ativa. A análise permitiu ainda estimar a eficiência dos recursos da área comercial, evidenciando onde há maior retorno por real investido.

Essas análises fornecem insumos diretos para a reconfiguração das carteiras, definição de metas por segmento e elaboração de campanhas personalizadas, baseadas em perfil, histórico e comportamento transacional.

Com base nos padrões identificados, as seguintes recomendações estratégicas são sugeridas para a área comercial:

a) Priorizar o relacionamento com os clientes do *Cluster* 1, por meio de estratégias de fidelização, ofertas exclusivas e atendimento consultivo personalizado. A retenção desses clientes tem impacto direto e relevante sobre a lucratividade global da organização.

b) Investir em desenvolvimento de contas no Cluster 2, com foco em *upsell*, *cross-sell* e estímulo ao aumento do ticket médio, utilizando estratégias de nurturing, propostas escaláveis e ofertas modulares.

c) Reduzir o CAC no Cluster 3, ou revisar a pertinência da manutenção desses clientes na carteira ativa, avaliando o potencial de automação ou migração para canais de relacionamento de menor custo.

d) Monitorar o Cluster 4 com apoio de dashboards e métricas de engajamento, a fim de identificar contas com potencial de ascensão e separar aquelas com tendência à evasão.

Essas recomendações visam maximizar o retorno sobre os investimentos comerciais e alinhar os esforços da equipe à geração de valor de longo prazo, de forma sustentável e embasada em evidências.

Apesar da robustez metodológica e dos resultados promissores, a pesquisa apresenta limitações importantes. A principal delas refere-se à dependência de dados públicos secundários, obtidos de cadastros oficiais de pessoas jurídicas, que podem não refletir integralmente a atualidade dos relacionamentos comerciais ou mudanças recentes nos padrões de consumo.

Adicionalmente, a eficácia dos modelos depende de sua atualização periódica, especialmente em contextos de mercado voláteis ou altamente dinâmicos. Mudanças em políticas comerciais, variações macroeconômicas ou transformações no perfil dos consumidores podem impactar significativamente os padrões anteriormente identificados.

Por fim, vale destacar que a implementação da metodologia requer capacitação técnica mínima, tanto para o uso de ferramentas analíticas quanto para a interpretação dos resultados, o que pode representar uma barreira inicial para empresas com baixa maturidade digital ou analítica.

## 5.2 DISCUSSÃO

Os resultados obtidos ao longo deste estudo não apenas evidenciaram a viabilidade técnica da segmentação de clientes orientada por dados, como também apresentaram implicações práticas significativas para a gestão comercial B2B. A proposta metodológica, centrada na integração de técnicas de ciência de dados com métricas de valor do cliente, revelou-se uma abordagem eficaz para enfrentar os desafios contemporâneos relacionados à alocação de recursos comerciais, priorização de contas e otimização da rentabilidade por cliente.

A primeira grande contribuição refere-se à validação da metodologia de segmentação desenvolvida. Utiliza algoritmos de aprendizado de máquina não supervisionado, como o *K-Means* aliado à técnica de Análise de Componentes Principais (PCA), a abordagem adotada demonstrou robustez na formação de agrupamentos homogêneos e semanticamente coerentes. A análise do método do cotovelo indicou que quatro clusters representavam um ponto ótimo entre

complexidade e interpretabilidade dos dados, permitindo uma segmentação mais granular sem comprometer a coesão dos grupos. Embora o coeficiente de silhueta tenha atingido valor máximo (0,85) para uma configuração com dois clusters — sugerindo alta coesão interna —, optou-se por quatro agrupamentos justamente para proporcionar maior riqueza de ideias e ações estratégicas diferenciadas por perfil.

Do ponto de vista prático, essa decisão metodológica revela-se fundamental, pois permite que empresas classifiquem seus clientes com base em dados objetivos, como faturamento, CAC e LTV, reduzindo a subjetividade inerente às decisões baseadas unicamente em experiência ou feeling dos gestores. Ao estruturar essa segmentação com base em variáveis quantitativas validadas estatisticamente, a organização pode atuar de forma mais precisa e fundamentada na definição de prioridades e na elaboração de estratégias específicas para cada grupo identificado.

Outro ponto de destaque na discussão é a relevância da relação entre o LTV e o CAC como métrica-chave para o direcionamento estratégico da atuação comercial. A análise da razão entre essas duas variáveis ao longo dos clusters revelou padrões consistentes que sustentam a alocação mais eficiente de investimentos. Clientes classificados nos clusters com LTV elevado e CAC reduzido mostraram-se altamente rentáveis, o que justifica o foco em ações de retenção, fidelização e expansão do relacionamento, por meio de estratégias de *upsell* e *cross-sell*. Já nos casos em que o CAC superava o LTV, foi possível identificar perfis de clientes cuja manutenção ativa no portfólio representava uma ineficiência financeira. Essa situação levanta importantes reflexões sobre a necessidade de revisão das estratégias de aquisição, dos canais utilizados e do próprio fit comercial desses *leads* com a proposta de valor da empresa.

Empresas com razão LTV/CAC acima de 3 podem ser consideradas sustentáveis e estrategicamente vantajosas, enquanto segmentos com razão inferior a 1 podem representar não apenas baixa lucratividade, mas também potenciais gargalos operacionais e desperdício de recursos. Essa análise proporciona uma visão mais racional e objetiva da carteira, substituindo abordagens lineares ou generalistas por um modelo de decisão orientado por valor.

A aplicabilidade da metodologia em diferentes setores também se apresenta como um ponto relevante a ser discutido. Embora o estudo tenha sido conduzido com base em dados do setor financeiro nacional, a abordagem é perfeitamente adaptável

a outros mercados B2B, como tecnologia, telecomunicações, logística e até mesmo em áreas menos tradicionais como a medicina alternativa, conforme exemplificado por estudos anteriores. A estrutura da metodologia permite que variáveis específicas de cada setor — como ticket médio, taxa de *churn*, tempo médio de contrato ou indicadores de satisfação — sejam incorporadas ao modelo, personalizando a análise conforme as particularidades do negócio. Isso amplia o escopo de aplicação e torna o framework proposto uma ferramenta versátil e escalável para diferentes contextos organizacionais.

Contudo, algumas limitações importantes devem ser reconhecidas. A principal delas é a dependência de dados públicos secundários, oriundos de cadastros de pessoas jurídicas, os quais, embora acessíveis, podem apresentar atrasos na atualização e baixa granularidade comportamental. Isso limita a capacidade preditiva dos modelos em contextos que exigem agilidade e atualizações constantes, especialmente em cenários pós-pandêmicos ou de instabilidade econômica. Além disso, a análise concentrou-se exclusivamente em variáveis quantitativas, o que pode deixar de fora aspectos qualitativos relevantes, como nível de satisfação do cliente, engajamento com a marca ou histórico de interações, que certamente enriqueceriam as inferências obtidas.

Outro desafio identificado é a necessidade de atualização periódica dos clusters e dos modelos analíticos, uma vez que o comportamento dos clientes, os ciclos econômicos e as estratégias organizacionais estão em constante transformação. A adoção da metodologia em ambientes reais, portanto, exige integração com sistemas internos da empresa, como CRM, ERP e plataformas de automação de *marketing*, possibilitando a automação da segmentação e o monitoramento em tempo real de indicadores-chave.

Nesse sentido, abrem-se diversas possibilidades para pesquisas futuras. Uma evolução natural do modelo seria a incorporação de técnicas mais avançadas de inteligência artificial, como redes neurais recorrentes (RNNs) ou modelos de previsão baseados em Transformers, para estimar o LTV de forma dinâmica e contextualizada. Além disso, a combinação entre algoritmos de clusterização e modelos explicativos, como árvores de decisão (*Decision Trees*), pode oferecer maior interpretabilidade dos grupos formados, tornando os resultados ainda mais acessíveis para gestores não técnicos.

Outra frente promissora de expansão está na análise de dados não estruturados, como *text mining* de *e-mails*, reclamações, interações em SAC ou avaliações em redes sociais, o que poderia revelar padrões comportamentais e emocionais ainda mais profundos. Esses dados, integrados com os modelos quantitativos atuais, dariam origem a uma segmentação mais holística e orientada à experiência do cliente (*Customer Experience* – CX).

Por fim, a integração prática da metodologia com ferramentas de mercado, como Salesforce, SAP ou plataformas de *Business Intelligence*, permitiria a criação de dashboards interativos e automatizados, capazes de atualizar as segmentações em tempo real, enviar alertas sobre mudanças de perfil e sugerir ações comerciais personalizadas por cluster.

# 6 CONCLUSÕES

Este estudo teve como objetivo propor, desenvolver e validar uma metodologia de segmentação de clientes B2B orientada à lucratividade, ancorada em princípios da ciência de dados e guiada por métricas estratégicas amplamente reconhecidas no contexto de gestão de relacionamento com o cliente, como o LTV e o CAC. A partir da combinação do procedimento sugerido pelo KDD e adição de algoritmos de machine learning não supervisionado, notadamente o *K-Means* e a Análise de Componentes Principais (PCA), foi possível estruturar uma abordagem robusta, replicável e adaptável à realidade de diferentes setores.

A metodologia proposta apresentou resultados altamente satisfatórios no contexto analisado, evidenciando que é possível agrupar clientes de forma coerente e estratégica a partir de dados objetivos. Ao identificar quatro clusters distintos, a segmentação permitiu diferenciar perfis de clientes com base em sua rentabilidade, engajamento e potencial de crescimento. Essa granularidade revelou-se extremamente útil para as decisões comerciais, já que possibilitou não apenas a priorização de contas com alto potencial de retorno (LTV elevado e CAC reduzido), mas também a identificação de grupos que representam alto custo e baixo retorno, exigindo reavaliação da abordagem comercial.

A razão LTV/CAC, utilizada como indicador central na análise, proporcionou um critério objetivo e estratégico para classificar os clientes em função de seu valor financeiro ao longo do tempo em relação ao esforço necessário para adquiri-los e mantê-los. A presença de clusters com LTV/CAC superior a 3 indicou contas altamente rentáveis, que devem ser foco de retenção, fidelização e expansão. Por outro lado, a existência de segmentos com LTV/CAC inferior a 1 acendeu um sinal de alerta sobre a eficiência dos investimentos em aquisição, revelando possíveis desequilíbrios na distribuição de recursos e esforços comerciais.

A metodologia também se mostrou flexível e adaptável a diferentes realidades organizacionais. Embora tenha sido aplicada a um conjunto de dados oriundos do setor financeiro B2B, a estrutura do modelo permite sua transposição para setores diversos, como tecnologia, telecomunicações, educação corporativa, logística ou serviços profissionais. Essa adaptabilidade decorre do fato de que o modelo pode ser ajustado por meio da substituição ou adição de variáveis específicas ao contexto de

cada setor, como ticket médio, índice de *churn*, tempo médio de contrato ou frequência de recompra. Além disso, as técnicas estatísticas utilizadas são suficientemente consolidadas e interpretáveis para aplicação em diferentes graus de maturidade analítica organizacional.

No campo prático, o modelo proposto representa uma ferramenta concreta de apoio à decisão gerencial, pois pode ser facilmente integrado a sistemas de CRM (como Salesforce, *Dynamics* 365 ou *Zoho*) e plataformas de *Business Intelligence* (como Power BI ou Tableau), gerando oportunidades e ações em tempo real. Essa capacidade de integração operacional transforma a metodologia de um instrumento analítico em uma solução estratégica com potencial de automação, contribuindo diretamente para a eficiência da força comercial e a maximização do retorno sobre o investimento (ROI) em *marketing* e vendas.

Comparativamente aos métodos tradicionais de segmentação — frequentemente baseados em categorias subjetivas, demográficas ou por setor de atuação —, a abordagem proposta introduz uma nova lógica orientada por valor, onde o comportamento financeiro e o ciclo de vida do cliente são os elementos centrais para a tomada de decisão. Isso representa uma quebra de paradigma importante, pois substitui suposições intuitivas por dados empíricos, estruturados e analisáveis, criando uma base sólida para decisões mais racionais e eficazes.

O modelo oferece a vantagem da dinamicidade, ou seja, a possibilidade de atualização periódica das segmentações à medida que novos dados são inseridos nos sistemas corporativos. Tal característica é fundamental em contextos voláteis, como o mercado pós-pandemia ou setores sujeitos a transformações tecnológicas constantes. A capacidade de reclassificar clientes conforme seu comportamento evolui ao longo do tempo permite que a empresa se antecipe a riscos de *churn*, identifique oportunidades emergentes e adapte suas estratégias de maneira contínua.

As contribuições deste estudo são múltiplas e relevantes tanto para a academia quanto para o mundo corporativo. No âmbito acadêmico, o trabalho avança ao propor uma integração efetiva entre técnicas de KDD, algoritmos de clustering e métricas estratégicas de *marketing* relacional, oferecendo uma contribuição metodológica aplicada ao campo da gestão de clientes em mercados B2B — uma área ainda carente de modelos preditivos eficazes. A operacionalização de conceitos como LTV e CAC

em uma estrutura analítica concreta, baseada em dados reais, representa uma inovação metodológica que pode ser replicada e refinada por futuros pesquisadores.

Do ponto de vista empresarial, a metodologia representa um instrumento imediatamente aplicável, com potencial de transformar a maneira como as organizações gerenciam suas carteiras de clientes. Sua utilização permite aumentar a eficiência operacional, reduzir desperdícios com contas de baixo retorno, melhorar a alocação de recursos comerciais e potencializar o valor gerado por clientes de alta atratividade. A segmentação orientada à lucratividade não apenas melhora o desempenho financeiro direto, como também eleva o nível de maturidade analítica da empresa, aproximando-a de uma cultura *data-driven*.

Como toda proposta metodológica, esta pesquisa possui limitações que precisam ser consideradas. A principal delas refere-se à utilização de dados secundários e públicos, que embora acessíveis, podem carecer de atualizações frequentes ou não captar nuances subjetivas do relacionamento com o cliente. Além disso, o modelo não contemplou variáveis qualitativas, como nível de satisfação, percepção de valor ou grau de fidelidade, que certamente enriqueceriam a análise caso estivessem disponíveis. Há também o desafio de garantir que os dados internos das empresas estejam estruturados, limpos e disponíveis em tempo hábil, o que nem sempre é uma realidade em organizações com baixa maturidade digital.

Diante dessas limitações, sugerem-se caminhos para pesquisas futuras que ampliem e aprofundem o modelo aqui desenvolvido. Entre eles, destaca-se o uso de inteligência artificial avançada, como redes neurais, modelos preditivos de aprendizado profundo (deep learning) e sistemas híbridos que combinem clustering com algoritmos supervisionados. A inclusão de dados não estruturados, como interações em canais de atendimento, e-mails, reviews e redes sociais, por meio de técnicas de processamento de linguagem natural (NLP) e *text mining*, também representa uma fronteira promissora. Além disso, a integração plena do modelo com plataformas de CRM e automação de *marketing* poderia permitir a segmentação automática e a geração de ideias e ações em tempo real, aumentando significativamente o impacto da metodologia na performance organizacional.

Em síntese, a presente pesquisa comprova que a segmentação de clientes B2B baseada em dados e orientada à lucratividade é não apenas viável, mas essencial em ambientes de negócios cada vez mais dinâmicos, competitivos e orientados por

performance. Ao transformar dados em conhecimento e conhecimento em ação, a metodologia aqui apresentada oferece um novo paradigma de gestão de relacionamento com clientes, pautado pela inteligência analítica, pela personalização estratégica e pela maximização de valor ao longo do tempo.

## REFERÊNCIAS BIBLIOGRÁFICAS

AFINITI, V. P. A new approach to proportional hazards modeling for estimating customer lifetime value. 2022.

BARAN, R. J.; GALKA, R. J.; STRUNK, D. P. CRM: the foundations of contemporary *marketing* strategy. Londres: Routledge, 2013.

BAUER, J.; JANNACH, D. Improved customer lifetime value prediction with sequence-to-sequence learning and feature-based models. *ACM Transactions on Knowledge Discovery from Data*, v. 15, n. 5, 2021.

BERGER, P. D.; NASR, N. I. Customer lifetime value: *marketing* models and applications. *Journal of Interactive Marketing*, v. 12, n. 1, p. 17-30, Winter 1998.

BREIMAN, L. Random forests. *Machine Learning*. Dordrecht: Springer, 2001.

BURELLI, P. Predicting customer lifetime value in free-to-play games. 2019.

CAO, Y.; RUSMEVICHIENTONG, P.; TOPALOGLU, H. Revenue management under a mixture of independent demand and multinomial logit models. *Operations Research*, v. 71, n. 2, p. 603–625, 2023.

CHENG, H.; CHEN, Y. Classification of the risk levels of heart disease using a hybrid data mining approach. In: *Proceedings of the International Multiconference of Engineers and Computer Scientists*, v. 1, 2009.

CUMPS, B. et al. Inferring comprehensible business ICT alignment rules. *Information & Management*, v. 46, n. 2, p. 116-124, 2009. DOI: 10.1016/j.im.2008.05.005.

DAHANA, W. D.; MIWA, Y.; MORISADA, M. Linking lifestyle to customer lifetime value: an exploratory study in an online fashion retail market. *Journal of Business Research*, v. 99, p. 319–331, 2019.

DO, C. B.; BATZOGLOU, S. What is the expectation maximization algorithm? *Nature Biotechnology*, v. 26, n. 8, p. 897-899, 2008.

DOMINGOS, P. The master algorithm. Basic Books, 2015.

EKSTRAND, M. D.; RIEDL, J. T.; KONSTAN, J. A. Collaborative filtering recommender systems. *Foundations and Trends in Human–Computer Interaction*, v. 4, n. 2, p. 81-173, 2010.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. Portland: AAAI Press, 1996. p. 226–231.

FARRIS, P. W. et al. *Marketing* metrics: the definitive guide to measuring *marketing* performance. Londres: Pearson, 2020.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to *knowledge discovery in databases*. *AI Magazine*, v. 17, n. 3, p. 37-54, 1996.

FIELD, A.; MILES, J.; FIELD, Z. Discovering statistics using R. 2. ed. London: Sage, 2017.

HAIR, J. F. et al. Multivariate data analysis. 7. ed. Upper Saddle River, NJ: Prentice Hall, 2009.

HAN, J.; KAMBER, M.; PEI, J. Data mining: concepts and techniques. 3. ed. Waltham: Morgan Kaufmann, 2011.

HARRIS, C. R. et al. Array programming with NumPy. *Nature*, v. 585, n. 7825, p. 357-362, 2020.

HÖPPNER, S. et al. Profit driven decision trees for *churn* prediction. *European Journal of Operational Research*, 2018. Disponível em: https://www.elsevier.com/locate/ejor. Acesso em: 21 jul. 2024.

HUANG, M.; RUST, R. T. Engaged to a robot? The role of AI in service. *Journal of Service Research*, v. 23, p. 97-113, 2020.

HUNTER, J. D. Matplotlib: a 2D graphics environment. *Computing in Science & Engineering*, v. 9, n. 3, p. 90-95, 2007.

JASEK, P. et al. Comparative analysis of selected probabilistic customer lifetime value models in online shopping. *Journal of Business Economics and Management*, v. 20, n. 3, p. 398–423, 2019.

KABACOFF, R. R in action: data analysis and graphics with R. 3. ed. Shelter Island: Manning, 2021.

KANCHANAPOOM, K.; CHONGWATPOL, J. Integrated customer lifetime value (CLV) and customer migration model to improve customer segmentation. *Journal of Marketing Analytics*, 2022. Disponível em: https://link.springer.com/article/10.1057/s41270-022-00158-7. Acesso em: 22 jan. 2024.

KANCHANAPOOM, K.; CHONGWATPOL, J. Integrated customer lifetime value models to support *marketing* decisions in the complementary and alternative medicine industry. *Benchmarking*, 2023.

KANT, Immanuel. Reflexões sobre a educação. 3a Ed. São Paulo: Editora Nacional, 2004.

KELLER, K. L. Strategic brand management: building, measuring, and managing brand equity. Londres: Pearson, 2014.

KOTLER, P.; KELLER, K. L. Administração de *marketing*. 12. ed. São Paulo: Pearson Prentice Hall, 2006.

KRISHNAMURTHY, R.; DESHPANDE, P. Data visualization with Python. 2. ed. Birmingham: Packt, 2022.

KUMAR, A. et al. Customer lifetime value prediction: using machine learning to forecast CLV and enhance customer relationship management. In: *7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*. IEEE, 2023.

KUMAR, V. Managing customers for profit: strategies to increase profits and build loyalty. Philadelphia: Wharton School Publishing, 2018.

KUMAR, V.; DIXIT, A.; JAVALGI, R. G.; DASS, M. Relationship *marketing* in the digital age: concepts, practices, and perspectives. *Journal of Marketing Management*, v. 36, p. 216-244, 2020.

LAROCHELLE, H. et al. Interpretable machine learning: decision trees and beyond. MIT Press, 2022.

LI, K. et al. Billion-user customer lifetime value prediction: an industrial-scale solution from Kuaishou. In: *Proceedings of the International Conference on Information and Knowledge Management*. Association for Computing Machinery, 2022. p. 3243–3251.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. *NeurIPS*, 2020.

MALHOTRA, N. K. *Marketing* research: an applied orientation. 5. ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.

MATPLOTLIB. Documentação oficial do Matplotlib. Disponível em: https://matplotlib.org/. Acesso em: 16 jul. 2024.

MCDONALD, M.; DUNBAR, I. Market segmentation: how to do it and how to profit from it. John Wiley & Sons, 2012.

MCKINNEY, W. Python for data analysis. 3. ed. O'Reilly, 2022.

MCKINNEY, W. Data structures for statistical computing in Python. In: *Proceedings of the 9th Python in Science Conference*, 2010.

MINTZBERG, H.; AHLSTRAND, B.; LAMPEL, J. Safari de estratégia: um roteiro pela selva do planejamento estratégico. 2. ed. Porto Alegre: Bookman, 2010.

MISSINGNO. Repositório oficial do Missingno no GitHub. Disponível em: https://github.com/ResidentMario/missingno. Acesso em: 16 jul. 2024.

MOLNAR, C. Interpretable machine learning. 2. ed. 2022. Disponível em: https://christophm.github.io/interpretable-ml-book/.

NATIONAL ACADEMIES OF SCIENCES. Data science for undergraduates: consensus study report. Washington: The National Academies Press, 2021.

NIJKAMP, P. Multivariate analysis in practice: the application of statistical methods. Berlin: Springer-Verlag, 1999.

NUMPY. Documentação oficial do NumPy. Disponível em: https://numpy.org/. Acesso em: 16 jul. 2024.

OLIVEIRA, D. P. R. Planejamento estratégico: conceitos, metodologia e práticas. 34. ed. São Paulo: Atlas, 2018.

OLNÉN, J. Customer lifetime value: maximizing profitability through customer loyalty. Business Insights Press, 2022.

PAGANO, M.; GAUVREAU, K. Principles of biostatistics. 2. ed. Boca Raton: CRC Press, 2018.

PANDAS. Documentação oficial do Pandas. Disponível em: https://pandas.pydata.org/. Acesso em: 16 jul. 2024.

PAYNE, A.; FROW, P. Strategic customer management: integrating relationship *marketing* and CRM. Cambridge: Cambridge University Press, 2017.

PLATÃO. Apologia de Sócrates. Tradução de Carlos Alberto Nunes. Belém: EDUFPA, 2000.

PEDREGOSA, F. et al. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

POLLAK, Z. Predicting customer lifetime value – e-commerce use case. 2021.

POLLAK, Z. Deep learning applications in customer lifetime value prediction. *Data Science Journal*, v. 20, 2021.

QUINLAN, J. R. Induction of decision trees. *Machine Learning*, v. 1, n. 1, p. 81–106, 1986.

REZAEINIA, S. M.; RAHMANI, R. Recommender system based on customer segmentation (RSCS). *Kybernetes*, v. 45, n. 6, p. 946–961, 2016.

RIEDL, J.; KONSTAN, J. A. Human–Computer Interaction Handbook: fundamentals, evolving technologies, and emerging applications. In: JACKO, J. A. (ed.). 3. ed. Boca Raton: CRC Press, 2011.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. *Nature*, v. 323, n. 6088, p. 533–536, 1986.

RUST, R. T.; LEMON, K. N.; ZEITHAML, V. A. Return on *marketing*: using customer equity to focus *marketing* strategy. *Journal of Marketing*, v. 68, n. 1, p. 109–127, 2004.

SCIKIT-LEARN. Documentação oficial do Scikit-learn. Disponível em: https://scikit-learn.org/stable/. Acesso em: 16 jul. 2024.

SEABORN. Documentação oficial do Seaborn. Disponível em: https://seaborn.pydata.org/. Acesso em: 16 jul. 2024.

STONE, M. et al. SCHEMA: information on *marketing* and customer engagement performance – reality versus dreams. *The Bottom Line*, 2019 (Accepted). DOI: 10.1108/BL-02-2019-0065.

SU, H. et al. Cross-domain adaptative learning for online advertisement customer lifetime value prediction. 2023.

TAN, P. N.; STEINBACH, M.; KUMAR, V. Introduction to data mining. 2. ed. Harlow: Pearson, 2019.

THOMAS, R. J. Multistage market segmentation: an exploration of B2B segment alignment. *Journal of Business and Industrial Marketing*, v. 31, n. 7, p. 821–834, 2016.

THOMPSON, B. Exploratory and confirmatory factor analysis: understanding concepts and applications. Washington, DC: American Psychological Association, 2004.

TIMES HIGHER EDUCATION. World university rankings 2023: data science and analytics. 2023. Disponível em: https://www.timeshighereducation.com.

VANDERPLAS, J. Python data science handbook: essential tools for working with data. 1. ed. Sebastopol: O'Reilly Media, 2016.

VERBEKE, W.; MARTENS, D.; BAESENS, B. Building comprehensible customer *churn* prediction models with advanced rule induction techniques. *Expert Systems with Applications*, v. 38, n. 3, p. 2354–2364, 2011.

VERHOEVEN, D.; PESCH, T.; CAO, Y. Utilizing genetic algorithms for revenue management optimization. *Journal of Revenue and Pricing Management*, v. 22, n. 3, p. 245–265, 2023.

WANG, X.; LIU, T.; MIAO, J. A deep probabilistic model for customer lifetime value prediction. 2019. Disponível em: http://arxiv.org/abs/1912.07753.

WASKOM, M. et al. Missingno: a missing data visualization suite. 2020. Disponível em: https://github.com/ResidentMario/missingno.

WASKOM, M. L. et al. Seaborn: statistical data visualization. *Journal of Open Source Software*, v. 5, n. 51, p. 3021, 2020. DOI: 10.21105/joss.03021.

WASSERMAN, L. All of statistics: a concise course in statistical inference. 2. ed. New York: Springer, 2020.

WICKHAM, H.; GROLEMUND, G. R for data science. Sebastopol: O'Reilly, 2017.

WIN, T. T.; BO, K. S. Predicting customer class using customer lifetime value with random forest algorithm. In: *International Conference on Advanced Information Technologies (ICAIT)*. IEEE, 2020. p. 236–241.

WU, C. et al. Contrastive multi-view framework for customer lifetime value prediction. *Proceedings of the ACM Web Conference*, p. 2400–2408, 2023.

XIE, Y. et al. Customer *churn* prediction using improved balanced random forests. *Expert Systems with Applications*, v. 120, p. 239–250, 2019. DOI: 10.1016/j.eswa.2018.11.030.

ZHANG, Z.; ZHAO, Y.; HUZHANG, G. Exploit customer lifetime value with memoryless experiments. 2022. Disponível em: http://arxiv.org/abs/2201.06254.

ZUUR, A. F.; IENO, E. N.; ELPHICK, C. S. A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, v. 10, n. 1, p. 170–181, 2019.

# GLOSSÁRIO

1.**B2B (*Business to Business*)**

- **Definição:** Modelo de negócios em que as transações comerciais ocorrem entre empresas, em contraste com o B2C (*Business to Consumer*), que envolve vendas diretas ao consumidor final.
- **Contexto:** Utilizado para descrever empresas que fornecem produtos ou serviços para outras empresas, como no caso de segmentação de clientes corporativos.

2.**Big Data**

- **Definição:** Conjuntos de dados extremamente volumosos e complexos, que exigem ferramentas avançadas para captura, armazenamento, análise e visualização.
- **Contexto:** Empregado na pesquisa para análise de dados públicos e descoberta de padrões em grandes volumes de informações.

3.**CAC (*Customer Acquisition Cost*)**

- **Definição:** Custo total para adquirir um novo cliente, incluindo despesas com *marketing*, vendas e outras atividades relacionadas.
- **Contexto:** Utilizado como métrica para avaliar a eficiência de estratégias de prospecção e comparar com o LTV.

4.***Churn***

- **Definição:** Taxa de evasão de clientes, representando a porcentagem de clientes que deixam de utilizar um produto ou serviço em um determinado período.
- **Contexto:** Fundamental para calcular o tempo de retenção do cliente (LTR) e o LTV.

5.**Clusterização**

- **Definição:** Técnica de aprendizado não supervisionado que agrupa dados com base em similaridades, como o algoritmo *K-Means*.
- **Contexto:** Aplicada para segmentar clientes em grupos homogêneos com características comuns.

6.**CRM (*Customer Relationship Management*)**

- **Definição:** Estratégias, tecnologias e práticas para gerenciar interações com clientes, visando melhorar relacionamentos e retenção.
- **Contexto:** Mencionado como ferramenta para implementar ações de segmentação e fidelização.

7.**KDD (*Knowledge Discovery in Databases*)**

- **Definição:** Processo sistemático para extrair conhecimento útil de grandes volumes de dados, envolvendo etapas como pré-processamento, mineração e interpretação.
- **Contexto:** Base metodológica para a segmentação de clientes proposta no estudo.

8.**LTV**

- **Definição:** Valor total estimado que um cliente gera para uma empresa ao longo de todo o relacionamento.
- **Contexto:** Métrica central para orientar estratégias de retenção, priorização de clientes e alocação de recursos.

9.**Machine Learning**

- **Definição:** Campo da inteligência artificial que utiliza algoritmos para permitir que sistemas aprendam com dados e melhorem desempenho sem programação explícita.
- **Contexto:** Empregado em técnicas como clusterização, árvores de decisão e redes neurais para segmentação e previsão de LTV.

10.**Outliers**

- **Definição:** Valores atípicos em um conjunto de dados que se desviam significativamente da distribuição normal.
- **Contexto:** Identificados e tratados durante o pré-processamento para evitar distorções nas análises.

11.**Segmentação de Clientes**

- **Definição:** Divisão da base de clientes em grupos com características semelhantes para direcionar estratégias personalizadas.

- **Contexto:** Objetivo principal da pesquisa, utilizando critérios como lucratividade, comportamento e perfil estratégico.

12. **Valores Ausentes**

- **Definição:** Dados faltantes em um conjunto, que podem ser tratados por técnicas como imputação ou exclusão.
- **Contexto:** Parte do pré-processamento para garantir a qualidade dos dados analisados.

# ANEXOS

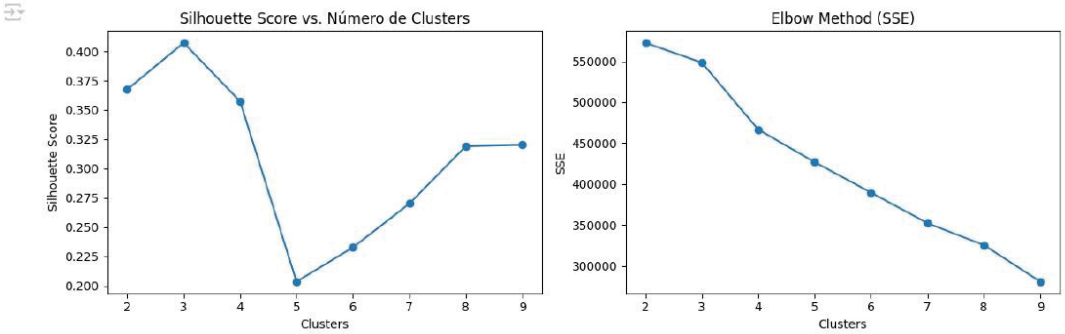


```
import pandas as pd
import numpy as np
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler, LabelEncoder
from sklearn.metrics import silhouette_score
from sklearn import tree
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.tree import export_graphviz
import graphviz


np.random.seed(42)


source = "/content/drive/MyDrive/Base_ListaLeads_55M_vs2.xlsx"
df = pd.read_excel(source, sheet_name="Cred", skiprows=1)


# Preenche valores ausentes
for col in ['Exportador', 'Importador']:
    df[col] = df[col].fillna("Não")

for col in ['Microregião', 'Mesoregião']:
    df[col] = df[col].fillna("SI")


# Remove linhas com dados críticos ausentes
df.dropna(subset=['FatPres', 'QuantFuncionarios', '%Rec'], inplace=True)


# Converte CodGr para binário
df['CodGr'] = df['CodGr'].apply(lambda x: 1 if x != 0 else 0)


# Codifica variáveis categóricas
label_cols = ['SegmentoGL', 'NívelAtiv', 'Porte', 'Situação', 'Exportador', 'Importador']
for col in label_cols:
    df[col] = LabelEncoder().fit_transform(df[col].astype(str))


# Seleciona colunas para clusterização
colunas_cluster = ['FatPres', 'FatR$', '%Rec', 'QuantFuncionarios',
                   'CodGr', 'PD_QtdNegAbe', 'PD_Perd',
                   'SegmentoGL', 'NívelAtiv', 'Porte', 'Situação', 'Exportador', 'Importador']



# Escala os dados
scaler = StandardScaler()
df_scaled = scaler.fit_transform(df[colunas_cluster])


# Cria DataFrame escalado
df_scaled_df = pd.DataFrame(df_scaled, columns=colunas_cluster)

# Exibe primeiras linhas para validação
print(df_scaled_df.head())
```

```
     FatPres     FatR$     %Rec  QuantFuncionarios     CodGr  PD_QtdNegAbe  \
0 -0.041971 -0.027498 -0.06458          -0.168083 -0.095115     -0.234790
1 -0.038676 -0.027498 -0.06458          -0.609404 -0.095115     -0.234790
2 -0.041971 -0.027498 -0.06458          -0.438145 -0.095115     -0.234790
3  0.194598 -0.027498 -0.06458           0.457669 -0.095115     -0.234790
4  0.063149 -0.027498 -0.06458           0.839708 -0.095115      3.631028

    PD_Perd  SegmentoGL  NívelAtiv     Porte  Situação  Exportador  Importador
0 -0.445985   -0.737714   0.075751  1.106231       0.0   -0.391563   -0.469478
1 -0.445985    0.650100   5.095660  1.106231       0.0   -0.391563   -0.469478
2 -0.445985    0.650100   0.075751  1.106231       0.0   -0.391563   -0.469478
3 -0.445985   -0.341196   0.075751 -0.903970       0.0   -0.391563   -0.469478
4  0.366199   -0.935973   0.075751 -0.903970       0.0   -0.391563   -0.469478
```

```
# Avaliação de número ótimo de clusters
silhouette_scores = []
sse = []

cluster_range = range(2, 10)
for n_clusters in cluster_range:
    kmeans = KMeans(n_clusters=n_clusters, random_state=42, n_init='auto')
    kmeans.fit(df_scaled_df)

    silhouette_scores.append(silhouette_score(df_scaled_df, kmeans.labels_))
```

```
    sse.append(kmeans.inertia_)


# Plota Silhouette Score e SSE (Elbow)
plt.figure(figsize=(12, 4))

plt.subplot(1, 2, 1)
plt.plot(cluster_range, silhouette_scores, marker='o')
plt.title('Silhouette Score vs. Número de Clusters')
plt.xlabel('Clusters')
plt.ylabel('Silhouette Score')

plt.subplot(1, 2, 2)
plt.plot(cluster_range, sse, marker='o')
plt.title('Elbow Method (SSE)')
plt.xlabel('Clusters')
plt.ylabel('SSE')

plt.tight_layout()
plt.show()
```




```
# CÓPIA DO DATAFRAME FILTRADO PARA TRATAMENTO
df_trat = df[['SegmentoGL', 'NívelAtiv', 'QuantFuncionarios', 'FatPres']].copy()

# CONVERSÃO DE CATEGÓRICOS PARA NÚMEROS
label_encoder = LabelEncoder()
df_trat['SegmentoGL'] = label_encoder.fit_transform(df_trat['SegmentoGL'])
df_trat['NívelAtiv'] = label_encoder.fit_transform(df_trat['NívelAtiv'])


# ESCALONA OS DADOS
scaler = StandardScaler()
df_trat_scaled = scaler.fit_transform(df_trat)


# CLUSTERIZAÇÃO COM KMEANS
quantidade_cluster = 5
k_means = KMeans(init="k-means++", n_clusters=quantidade_cluster, n_init=100, random_state=42)
k_means.fit(df_trat_scaled)
k_means_labels = k_means.labels_


# Adiciona os clusters ao dataframe original
df_cluster = df.copy()
df_cluster["cluster"] = k_means_labels


# VISUALIZAÇÃO: GRÁFICO DE DISPERSÃO POR CLUSTER
import matplotlib.patches as mpatches
plt.figure(figsize=(10, 6))
```

```
<Figure size 1000x600 with 0 Axes>
<Figure size 1000x600 with 0 Axes>
```

```
# Verifica se as colunas existem antes de tentar plotar
x_col = 'Valor Cadastro Ativos'
```

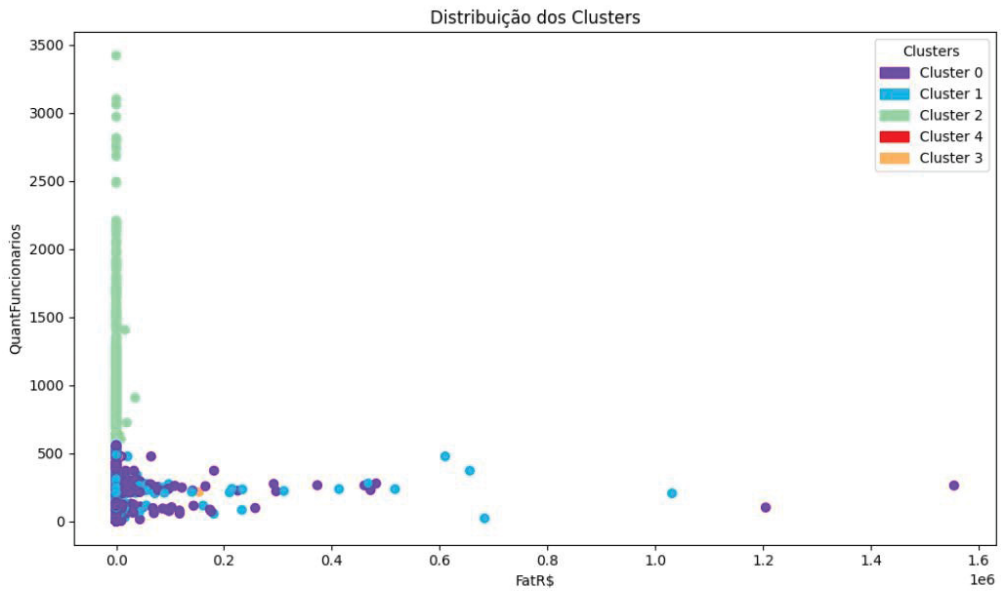
```
y_col = 'Qtde. Titulos Ativos'

if x_col in df.columns and y_col in df.columns:
    scatter = plt.scatter(df_cluster[x_col], df_cluster[y_col], c=df_cluster["cluster"], cmap="rainbow")
    legend_labels = df_cluster["cluster"].unique()
    legend_handles = [mpatches.Patch(color=scatter.cmap(scatter.norm(c)), label=f'Cluster {c}') for c in legend_labels]
    plt.legend(handles=legend_handles, title="Clusters")
    plt.xlabel(x_col)
    plt.ylabel(y_col)
    plt.title("Distribuição dos Clusters")
    plt.tight_layout()
    plt.show()
else:
    print(f"As colunas '{x_col}' e/ou '{y_col}' não estão disponíveis no dataframe.")
```

```
As colunas 'Valor Cadastro Ativos' e/ou 'Qtde. Titulos Ativos' não estão disponíveis no dataframe.
```

```
print(df.columns.tolist())
```

```
gAbe', 'PD_Perd', 'PD_Prior', 'FxFat', 'FatPres', 'Porte', 'Situação', 'NívelAtiv', 'RiscoInad', 'CNAEs Secundários', 'Cidade', 'Est
```

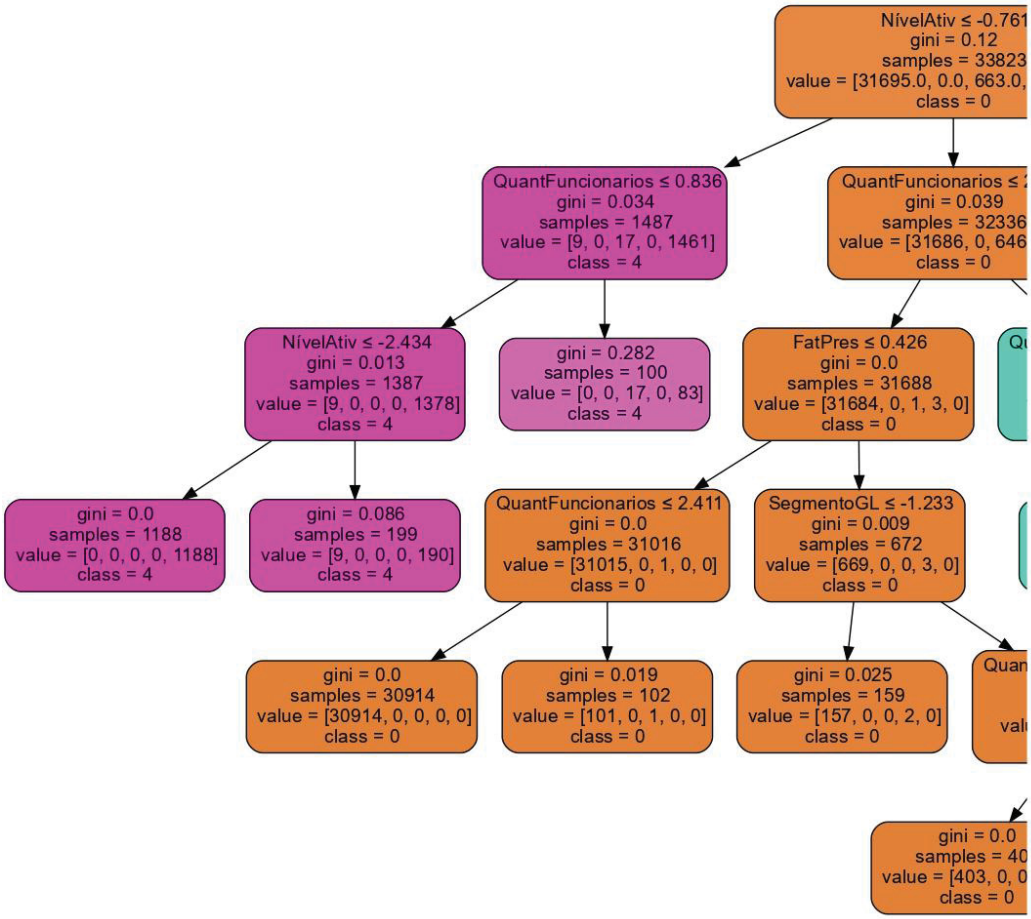
```
x_col = 'FatR$'
y_col = 'QuantFuncionarios'

plt.figure(figsize=(10, 6))
scatter = plt.scatter(df_cluster[x_col], df_cluster[y_col], c=df_cluster["cluster"], cmap="rainbow")
legend_labels = df_cluster["cluster"].unique()
legend_handles = [mpatches.Patch(color=scatter.cmap(scatter.norm(c)), label=f'Cluster {c}') for c in legend_labels]
plt.legend(handles=legend_handles, title="Clusters")
plt.xlabel(x_col)
plt.ylabel(y_col)
plt.title("Distribuição dos Clusters")
plt.tight_layout()
plt.show()
```




```
from sklearn.tree import DecisionTreeClassifier

# Treina a árvore com os dados tratados e os rótulos de cluster
tree_model = DecisionTreeClassifier(
    max_depth=7, min_samples_split=10, min_samples_leaf=100, random_state=42
)
tree_model.fit(df_trat_scaled, k_means_labels)
```

```
DecisionTreeClassifier
DecisionTreeClassifier(max_depth=7, min_samples_leaf=100, min_samples_split=10,
                       random_state=42)
```

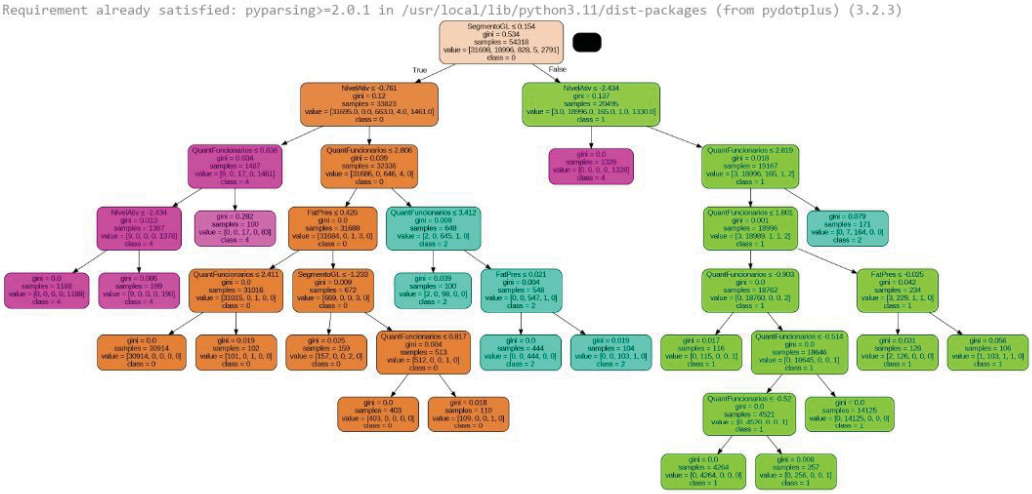
```
from sklearn.tree import export_graphviz
import graphviz

# Gera o código DOT da árvore
dot_data = export_graphviz(
    tree_model,
    out_file=None,
    feature_names=df_trat.columns,
    class_names=[str(i) for i in np.unique(k_means_labels)],
    filled=True,
    rounded=True,
    special_characters=True
)

# Renderiza no notebook
graphviz.Source(dot_data)
```




```
!pip install graphviz
!pip install pydotplus

import pydotplus
from IPython.display import Image
from sklearn.tree import export_graphviz

# Gera o arquivo .dot da árvore
export_graphviz(
    tree_model,
    out_file="delivery_tree.dot",
    feature_names=df_trat.columns,
    class_names=[str(i) for i in np.unique(k_means_labels)],
    filled=True,
```

```
    rounded=True,
    special_characters=True
)

# Converte o .dot em imagem
graph = pydotplus.graph_from_dot_file("delivery_tree.dot")
graph.write_png("delivery_tree.png")

# Mostra a imagem no notebook
Image(graph.create_png())
```

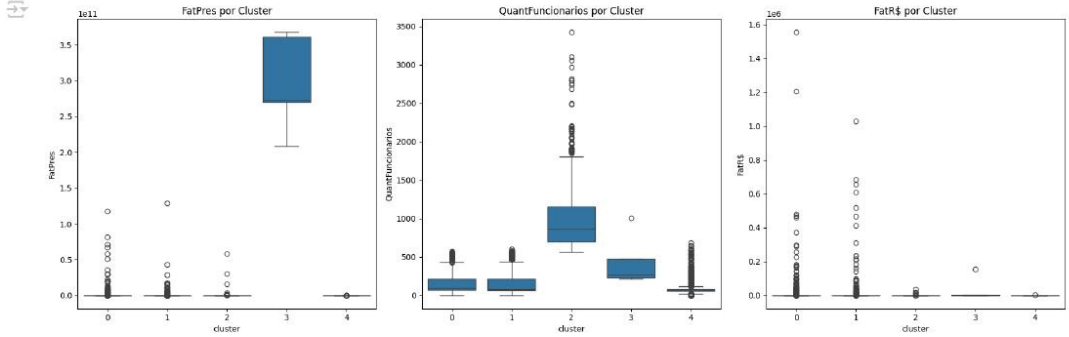
```
Requirement already satisfied: graphviz in /usr/local/lib/python3.11/dist-packages (0.21)
Requirement already satisfied: pydotplus in /usr/local/lib/python3.11/dist-packages (2.0.2)
Requirement already satisfied: pyparsing>=2.0.1 in /usr/local/lib/python3.11/dist-packages (from pydotplus) (3.2.3)
```

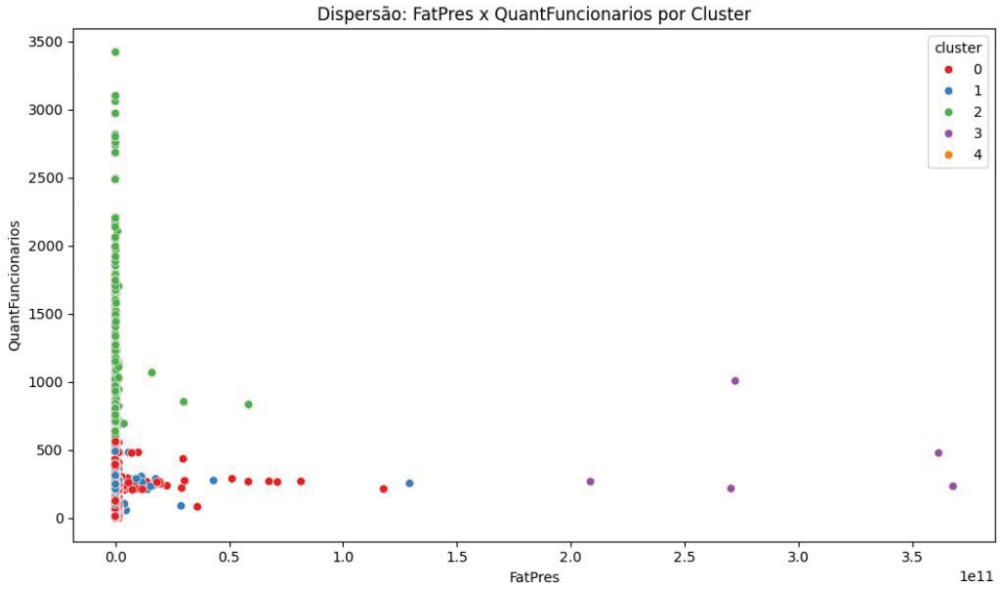


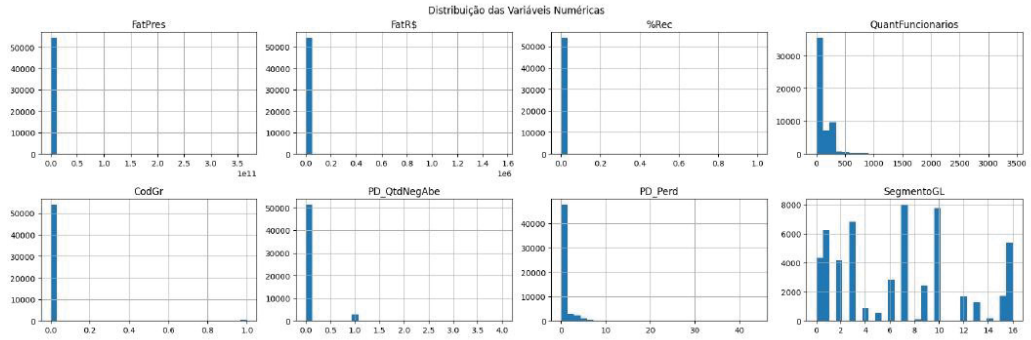

```
from google.colab import files
files.download("delivery_tree.png")
```

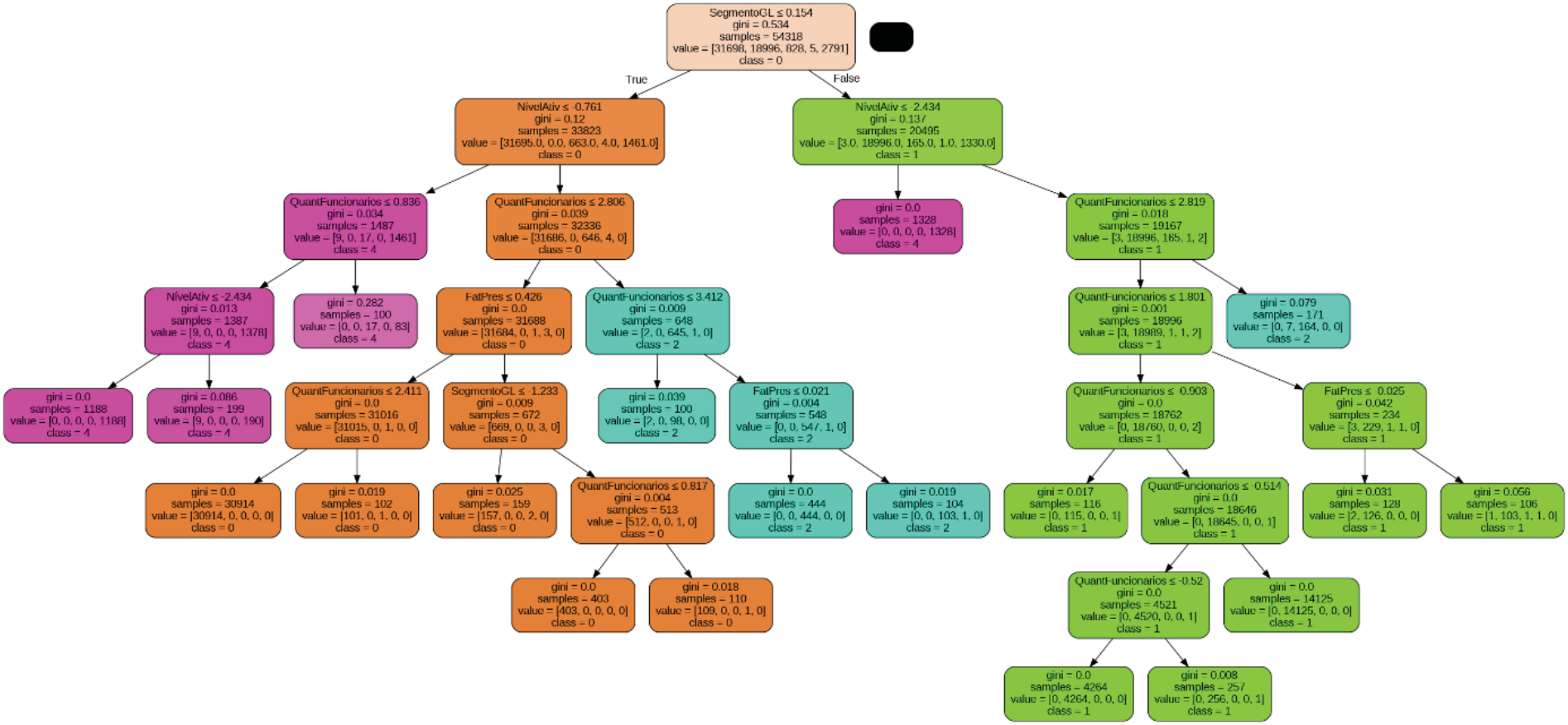
```

# Gráfico 1: Boxplot por cluster para FatPres, QuantFuncionarios e FatR$
fig, axes = plt.subplots(1, 3, figsize=(18, 6))
sns.boxplot(data=df_cluster, x='cluster', y='FatPres', ax=axes[0])
axes[0].set_title('FatPres por Cluster')
sns.boxplot(data=df_cluster, x='cluster', y='QuantFuncionarios', ax=axes[1])
axes[1].set_title('QuantFuncionarios por Cluster')
sns.boxplot(data=df_cluster, x='cluster', y='FatR$', ax=axes[2])
axes[2].set_title('FatR$ por Cluster')
plt.tight_layout()
plt.show()

# Gráfico 2: Dispersão entre FatPres e QuantFuncionarios por cluster
plt.figure(figsize=(10, 6))
sns.scatterplot(data=df_cluster, x='FatPres', y='QuantFuncionarios', hue='cluster', palette='Set1')
plt.title('Dispersão: FatPres x QuantFuncionarios por Cluster')
plt.tight_layout()
```

```
    plt.show()

    # Gráfico 3: Histograma das variáveis numéricas
    df_cluster[colunas_cluster].hist(figsize=(18, 12), bins=30)
    plt.suptitle('Distribuição das Variáveis Numéricas')
    plt.tight_layout()
    plt.show()

    # Gráfico 4: Heatmap de correlação
    plt.figure(figsize=(14, 10))
    sns.heatmap(df_cluster[colunas_cluster].corr(), annot=True, cmap="coolwarm", fmt=".2f")
    plt.title('Correlação entre Variáveis')
    plt.tight_layout()
    plt.show()
```


FatPres por Cluster
QuantFuncionarios por Cluster
FatR$ por Cluster
cluster


Dispersão: FatPres x QuantFuncionarios por Cluster
cluster
0
1
2
3
4
QuantFuncionarios
FatPres
1e11


Distribuição das Variáveis Numéricas
FatPres
FatR$
%Rec
QuantFuncionarios
CodGr
PD_QtdNegAbe
PD_Perd
SegmentoGL


SegmentoGL ≤ 0.154
gini = 0.534
samples = 54318
value = [31698, 18996, 828, 5, 2791]
class = 0
True
False
NivelAtiv ≤ -0.761
gini = 0.12
samples = 33823
value = [31695.0, 0.0, 663.0, 4.0, 1461.0]
class = 0
NivelAtiv ≤ -2.434
gini = 0.137
samples = 20495
value = [3.0, 18996.0, 165.0, 1.0, 1330.0]
class = 1
QuantFuncionarios ≤ 0.836
gini = 0.034
samples = 1487
value = [9, 0, 17, 0, 1461]
class = 4
QuantFuncionarios ≤ 2.806
gini = 0.039
samples = 32336
value = [31686, 0, 646, 4, 0]
class = 0
gini = 0.0
samples = 1328
value = [0, 0, 0, 0, 1328]
class = 4
QuantFuncionarios ≤ 2.819
gini = 0.018
samples = 19167
value = [3, 18996, 165, 1, 2]
class = 1
NivelAtiv ≤ -2.434
gini = 0.013
samples = 1387
value = [9, 0, 0, 0, 1378]
class = 4
gini = 0.282
samples = 100
value = [0, 0, 17, 0, 83]
class = 4
FatPres ≤ 0.426
gini = 0.0
samples = 31688
value = [31684, 0, 1, 3, 0]
class = 0
QuantFuncionarios ≤ 3.412
gini = 0.009
samples = 648
value = [2, 0, 645, 1, 0]
class = 2
QuantFuncionarios ≤ 1.801
gini = 0.001
samples = 18996
value = [3, 18989, 1, 1, 2]
class = 1
gini = 0.079
samples = 171
value = [0, 7, 164, 0, 0]
class = 2
gini = 0.0
samples = 1188
value = [0, 0, 0, 0, 1188]
class = 4
gini = 0.086
samples = 199
value = [9, 0, 0, 0, 190]
class = 4
QuantFuncionarios ≤ 2.411
gini = 0.0
samples = 31016
value = [31015, 0, 1, 0, 0]
class = 0
SegmentoGL ≤ -1.233
gini = 0.009
samples = 672
value = [669, 0, 0, 3, 0]
class = 0
gini = 0.039
samples = 100
value = [2, 0, 98, 0, 0]
class = 2
FatPres ≤ 0.021
gini = 0.004
samples = 548
value = [0, 0, 547, 1, 0]
class = 2
QuantFuncionarios ≤ 0.903
gini = 0.0
samples = 18762
value = [0, 18760, 0, 0, 2]
class = 1
FatPres ≤ 0.025
gini = 0.042
samples = 234
value = [3, 229, 1, 1, 0]
class = 1
gini = 0.0
samples = 30914
value = [30914, 0, 0, 0, 0]
class = 0
gini = 0.019
samples = 102
value = [101, 0, 1, 0, 0]
class = 0
gini = 0.025
samples = 159
value = [157, 0, 0, 2, 0]
class = 0
QuantFuncionarios ≤ 0.817
gini = 0.004
samples = 513
value = [512, 0, 0, 1, 0]
class = 0
gini = 0.0
samples = 444
value = [0, 0, 444, 0, 0]
class = 2
gini = 0.019
samples = 104
value = [0, 0, 103, 1, 0]
class = 2
gini = 0.017
samples = 116
value = [0, 115, 0, 0, 1]
class = 1
QuantFuncionarios ≤ 0.514
gini = 0.0
samples = 18646
value = [0, 18645, 0, 0, 1]
class = 1
gini = 0.031
samples = 128
value = [2, 126, 0, 0, 0]
class = 1
gini = 0.056
samples = 106
value = [1, 103, 1, 1, 0]
class = 1
gini = 0.0
samples = 403
value = [403, 0, 0, 0, 0]
class = 0
gini = 0.018
samples = 110
value = [109, 0, 0, 1, 0]
class = 0
QuantFuncionarios ≤ 0.52
gini = 0.0
samples = 4521
value = [0, 4520, 0, 0, 1]
class = 1
gini = 0.0
samples = 14125
value = [0, 14125, 0, 0, 0]
class = 1
gini = 0.0
samples = 4264
value = [0, 4264, 0, 0, 0]
class = 1
gini = 0.008
samples = 257
value = [0, 256, 0, 0, 1]
class = 1