UNIVERSIDADE FEDERAL DE ALAGOAS

CAMPUS A. C. SIMÕES

CENTRO DE TECNOLOGIA

BACHARELADO EM ENGENHARIA CIVIL

JÚLIO CÉSAR SILVA

DETECÇÃO DE CONSUMOS NÃO AUTORIZADOS EM SISTEMAS DE ABASTECIMENTO DE ÁGUA UTILIZANDO CIÊNCIA DE DADOS

MACEIÓ-AL

2025

JÚLIO CÉSAR SILVA

DETECÇÃO DE CONSUMOS NÃO AUTORIZADOS EM SISTEMAS DE ABASTECIMENTO DE ÁGUA UTILIZANDO CIÊNCIA DE DADOS

Plano de Trabalho de Conclusão de Curso apresentado ao Colegiado do curso de Engenharia Civil da Universidade Federal de Alagoas, como requisito parcial à obtenção do título de Bacharel em Engenharia Civil.

Orientadora: Professora Dra. Daysy Lira Oliveira Cavalcanti.

Coorientador: Me. Mahelvson Bazilio Chaves.

MACEIÓ-AL

2025

Folha de Aprovação

JÚLIO CÉSAR SILVA

DETECÇÃO DE CONSUMOS NÃO AUTORIZADOS EM SISTEMAS DE ABASTECIMENTO DE ÁGUA UTILIZANDO CIÊNCIA DE DADOS

Plano de Trabalho de Conclusão de Curso apresentado ao Colegiado do curso de Engenharia Civil da Universidade Federal de Alagoas, como requisito parcial à obtenção do título de Bacharel em Engenharia Civil.

Banca Examinadora:

Orientadora: Professora Dra. Daysy Lira Oliveira Cavalcanti.

Examinador: Professor Dr. Vladimir Caramori Borges de Souza

Examinador: João Gabriel Rocha Vanderlei

RESUMO

Desde a década de 1980, o consumo de água tem aumentado cerca de 1% ao ano, impulsionado pelo crescimento populacional, desenvolvimento socioeconômico e mudanças nos padrões de uso, o que reforça a necessidade de uma gestão sustentável dos recursos hídricos. Nesse contexto, as perdas de água ao longo das diferentes etapas de um sistema de abastecimento impactam diretamente o acesso à água, comprometendo a qualidade dos serviços prestados pelas companhias de abastecimento. Esses sistemas, inevitavelmente sujeitos a perdas, têm sua eficiência e controle relacionados a tais falhas. Posto isso, este trabalho tem como objetivo, por meio da análise do comportamento de consumo dos usuários, detectar irregularidades que geram perdas, aumentando assim a eficiência dos sistemas de abastecimento de água. Para isso, aplicou-se o método desenvolvido por Ghamkhar et al., (2023) o qual utiliza o algoritmo de aprendizado de máquina não supervisionado DBSCAN. Como característica essencial para lidar com a alta dimensionalidade e a baixa resolução dos dados, utilizando apenas os registros de consumo dos usuários, aplica-se a complexidade de Lempel-Ziv. Essa métrica permite avaliar a variabilidade e a imprevisibilidade das séries temporais, contribuindo para a identificação de padrões atípicos. O modelo foi aplicado a dados históricos de consumo de uma concessionária brasileira e demonstrou ser capaz de identificar com clareza padrões anormais, mesmo em cenários com dados de baixa resolução e qualidade limitada. Seis tipos distintos de padrões irregulares foram identificados, com destaque para 3.009 matrículas classificadas como possivelmente irregulares. A ausência de dados rotulados e de informações como status da matrícula e categoria do usuário limitou a validação formal e análises mais direcionadas. Ainda, não foi possível estimar o impacto ambiental das irregularidades detectadas, uma vez que os consumos irregulares apresentaram, em média, volumes superiores aos normais. Ainda assim, os resultados evidenciam o potencial do modelo como ferramenta de apoio à gestão operacional e à priorização de inspeções em campo.

Palavra-Chave: Perdas aparentes; Detecção de irregularidades; Aprendizagem de máquina; Series Temporais.

ABSTRACT

Since the 1980s, global water consumption has increased by approximately 1% per year, driven by population growth, socioeconomic development, and changing usage patterns, highlighting the need for sustainable water resource management. In this context, water losses throughout the different stages of a supply system directly affect access to water and compromise the quality of services provided by utilities. These systems, inherently subject to losses, have their efficiency and control closely linked to such failures. Given this scenario, the objective of this work is to detect irregularities that contribute to water losses through the analysis of user consumption behavior, thereby increasing the efficiency of water supply systems. To achieve this, the method proposed by Ghamkhar et al. (2023) was applied, which uses the unsupervised machine learning algorithm DBSCAN. As an essential feature for handling the high dimensionality and low resolution of the data—relying solely on consumption records—this study uses the Lempel-Ziv Complexity metric. This measure assesses the variability and unpredictability of time series, supporting the identification of atypical patterns. The model was applied to historical consumption data from a Brazilian water utility and proved capable of clearly identifying anomalous patterns, even in scenarios with low-resolution and poor-quality data. Six distinct types of anomalies were identified, highlighting 3,009 connections classified as potentially irregular. The lack of labeled data and key information, such as the connection's status and user category, limited formal validation and more targeted analyses. Furthermore, it was not possible to estimate the environmental impact of the detected anomalies, as the anomalous consumption values were, on average, higher than the regular ones. Nevertheless, the results demonstrate the model's potential as a support tool for operational management and for prioritizing field inspections.

Keywords: Apparent losses; Anomaly detection; Machine learning; Time series.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
SAA	Sistema de Abastecimento de Água
NMSB	Novo Marco do Saneamento Básico
SINISA	Sistema Nacional de Informações em Saneamento Básico
IWA	International Water Association
IDC	International Data Corporation
ML	Aprendizado de Máquina
AED	Análise Exploratória de Dados
DBSCAN	Density Based Spatial Clustering of Applicantion With Noise
TSLF	Fator de Comprimento da Serie Temporal
LZC	Complexidade de Lempel-Ziv
KNN	K-ésimo vizinho mais próximo
INMETRO	Instituto Nacional de Metrologia, Qualidade e Tecnologia

LISTA DE SÍMBOLOS

C	Matriz de Consumo
D	Matriz de Características Extraídas
L_u	Numero de Unidades de Consumo Distintas Registradas
L_t	Comprimento Total da Serie Temporal
C_i	Serie Temporal de Entrada
X	Sequencia Finita
S	Sequencia Simbólica
$W(S)$	Conjunto de todas as subsequências
$c(n)$	Valor Correspondente do LZC
N_{LZC}	Normalização dos valores de LZC
$u(n)$	Limite Superior do LZC
$MinPts$	Número Minimo de Pontos de um Conjunto
ε	Raio de distancia

SUMÁRIO

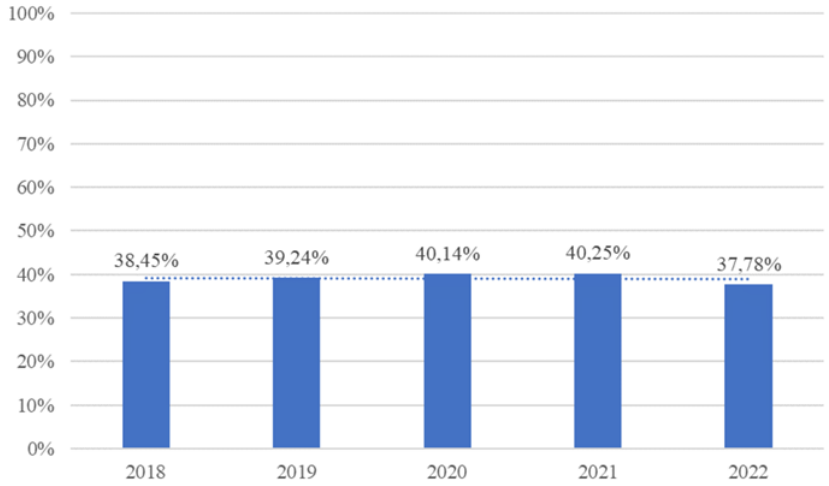
1 INTRODUÇÃO

Desde a década de 1980, o consumo da água tem aumentado em 1% ao ano, devido ao crescimento populacional, ao desenvolvimento socioeconômico e às mudanças nos padrões de consumo (NAÇÕES UNIDAS, 2024). Diante desse cenário, torna-se imprescindível uma gestão sustentável dos recursos hídricos, a fim de assegurar sua disponibilidade em quantidade e qualidade adequadas (GHAMKHAR et al., 2023).

Para alcançar esses objetivos, os gestores hídricos adotam diversas estratégias, sendo o controle de perdas nos sistemas de abastecimento de água (SAA) um aspecto essencial no planejamento e na concepção de infraestruturas urbanas sustentáveis (GHAMKHAR et al., 2023). As perdas em um SAA referem-se à diferença entre o volume de água produzido e o efetivamente consumido pelos usuários. Os sistemas estão inevitavelmente sujeitos a perdas, as quais se correlacionam com sua eficiência e controle. A questão das perdas impacta diretamente o acesso à água e a qualidade dos serviços prestados pelas companhias, que enfrentam o desafio de atender a uma demanda crescente ao mesmo tempo em que buscam captar menores volumes, visando aumentar a eficiência dos sistemas de abastecimento, reduzindo custos e minimizando impactos ambientais (TRATA BRASIL, 2024; BEZERRA e CHEUNG, 2013).

Segundo a atualização do Marco Legal do Saneamento Básico, promovida pela Lei nº 14.026, de 15 de julho de 2020 — que alterou dispositivos da Lei nº 11.445/2007 —, as diretrizes nacionais preveem a redução do índice de perdas na distribuição de água para até 25% até o ano de 2033. A Figura 1 ilustra a evolução das perdas na distribuição de água no Brasil ao longo dos últimos cinco anos de dados disponíveis (2018 a 2022), com base em informações fornecidas pelo Sistema Nacional de Informações em Saneamento Básico (SINISA).

Figura 1: Histórico recente das perdas na distribuição de água no Brasil



Fonte: TRATA BRASIL (2024)

Conforme apresentado na Figura 1, os índices de perdas de água no Brasil apresentaram um crescimento contínuo entre 2018 e 2021, aumentando de 38,45% para 40,25%. Em 2022, observou-se uma leve redução para 37,78%. Embora esse valor tenha apresentado uma discreta melhora no último ano, ele ainda é significativamente superior à meta estabelecida pelo marco regulatório. Ao comparar os indicadores de perdas de água no Brasil com os padrões internacionais, observa-se uma grande disparidade, uma vez que a média nos países desenvolvidos é de 15% (TRATA BRASIL, 2024).

De acordo com a International Water Association (IWA), as perdas podem ser classificadas como reais (físicas) ou aparentes (comerciais). As perdas reais decorrem de vazamentos que ocorrem ao longo das diferentes etapas de produção, como captação, tratamento, armazenamento e distribuição. Por sua vez, as perdas aparentes correspondem aos volumes de água consumidos, mas não faturados, resultantes de ligações clandestinas, fraudes nos hidrômetros, erros nas leituras dos hidrômetros e falhas no cadastro comercial (TRATA BRASIL, 2024).

Um metro cúbico de água que se perde na tubulação envolve os custos relacionados à sua produção e distribuição. Já os custos do consumo não registrado são equivalentes ao do último metro cúbico utilizado pelo usuário, o que os torna, portanto, mais elevados (GHAMKHAR et al., 2023). Posto isso, as perdas aparentes impactam diretamente a receita das companhias de saneamento, uma vez que correspondem a volumes produzidos e consumidos, mas não faturados. Consequentemente, um elevado nível de perdas aparentes reduz a capacidade financeira dos prestadores, limitando os recursos disponíveis para ampliar a oferta, melhorar a qualidade dos serviços ou realizar os investimentos necessários para a manutenção e reposição da infraestrutura (TRATA BRASIL, 2024). Portanto, é de extrema importância que as companhias de saneamento consigam mapear as perdas aparentes.

Entretanto, mapear as perdas aparentes apresenta um desafio significativo devido ao vasto volume de dados que as concessionárias precisam gerenciar, aliado à baixa qualidade frequente dessas informações. A presença de dados incompletos, inconsistentes ou desatualizados agrava a complexidade do processo, combinado à diversidade de fontes de informação e à variabilidade nos padrões de consumo, dificultando a identificação manual precisa de irregularidades no consumo (GHAMKHAR et al., 2023).

Neste cenário, a utilização da Ciência de Dados torna-se imprescindível, pois possibilita a coleta e análise de dados provenientes de diversas fontes, com o objetivo de apoiar a tomada

de decisões, utilizando, em geral, grandes volumes de dados de maneira sistematizada. Mais do que uma ferramenta isolada, a Ciência de Dados constitui um conjunto abrangente de métodos e técnicas voltados à análise orientada por dados (ESCOVEDO & KOSHIYAMA, 2020).

Diante da problemática apresentada, este trabalho tem como objetivo detectar irregularidades no comportamento de consumo dos usuários que contribuem para perdas no sistema de abastecimento de água, visando aumentar a eficiência operacional. Para isso, foi aplicado o método proposto por Ghamkhar et al. (2023), que aprimora o algoritmo de aprendizado de máquina não supervisionado DBSCAN incorporando características como Complexidade de Lempel-Ziv (LZC) e Fator de Comprimento de Séries Temporais (TSLF) para uma detecção mais eficaz de irregularidades no consumo.

2 OBJETIVOS

2.1 Objetivo Geral

Identificar padrões irregulares em séries temporais de consumo de água de uma companhia brasileira de abastecimento, por meio de técnicas de aprendizado de máquina não supervisionado, com foco em métodos de agrupamento e análise de complexidade temporal, visando contribuir para a redução de perdas aparentes e a otimização da gestão operacional.

2.2 Objetivos Específicos

- Detectar padrões comportamentais atípicos no consumo de água com base em agrupamento por densidade (DBSCAN) e medidas de complexidade temporal (TSLF e LZC).

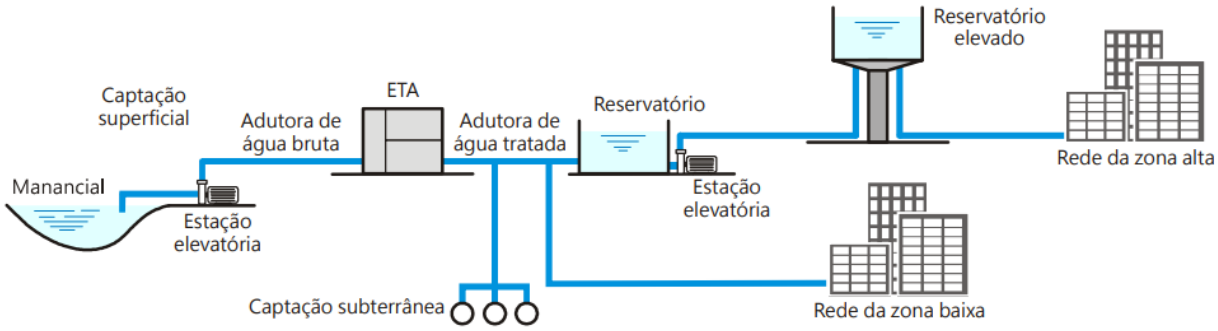
3 FUNDAMENTAÇÃO TEÓRICA

3.1 Sistema de Abastecimento de Água

Os Sistemas de Abastecimento de Água (SAA) são infraestruturas responsáveis pela captação, tratamento, armazenamento e distribuição de água para os núcleos urbanos, assegurando regularidade, segurança e qualidade. De acordo com a Lei Federal nº 11.445/2007, regulamentada pelo Decreto nº 7.217, de 21 de junho de 2010, os serviços de abastecimento devem ser prestados com eficiência e sustentabilidade econômica. Para que os SAA's desempenhem adequadamente sua função de proteger os consumidores contra riscos à saúde humana, é essencial que sua concepção, implantação, operação e manutenção sejam projetadas com rigor técnico (BEZERRA e CHEUNG, 2013).

Um sistema convencional de abastecimento de água é formado pelas unidades de captação, estação elevatória, estação de tratamento, adução, reservação e setor de distribuição, conforme ilustrado na Figura 2. A disposição dessas unidades varia conforme as características específicas de cada projeto (BEZERRA e CHEUNG, 2013).

Figura 2 : Componentes de um sistema de abastecimento.



Fonte: BEZERRA e CHEUNG (2013)

3.1.1 Balanço Hídrico

O balanço hídrico, por definição, consiste na quantificação das entradas e saídas de água em um sistema. Em um SAA, o balanço hídrico pode ser conceituado como um 'balanço de massa', elaborado com base em dados anuais dos setores comercial e operacional de uma determinada concessionária. Trata-se de uma abordagem estruturada para avaliar os componentes dos fluxos e usos da água no sistema, gerando diversos indicadores de desempenho que permitem o monitoramento das ações técnicas e operacionais (VANDERLEI, 2023; TSUTYIA, 2006). Assim, o balanço hídrico configura-se como uma ferramenta indispensável para a auditoria da água.

Posto isso, a International Water Associaton (IWA) desenvolveu um método para classificar, padronizar e uniformizar a terminologia relacionada ao tema. Esse método envolve uma matriz que esquematiza os processos pelos quais a água passa, desde sua entrada no sistema até a chegada ao consumidor final, sendo conhecido como Balanço Hídrico, conforme ilustrado na Figura 3. Internacionalmente, o consagrado Balanço Hídrico é referido como uma aplicação Top-Down, pois inicia a avaliação das perdas através do cálculo do balanço hídrico pelo volume de água que entra no sistema subtraído do volume de água efetivamente consumido (BEZERRA e CHEUNG, 2013; TRATA BRASIL, 2024).

Figura 3: Balanço Hídrico proposta pela IWA para sistemas de abastecimento de água.

<table>
<tr><td rowspan="6">Volume de entrada no sistema</td><td rowspan="3">Consumo autorizado</td><td rowspan="2">Consumo autorizado faturado</td><td>Consumo faturado medido (incluir água exportada)</td><td rowspan="2">Água faturada</td></tr>
<tr><td>Consumo faturado não medido (estimados)</td></tr>
<tr><td>Consumo autorizado não faturado</td><td>Consumo não faturado medido (usos próprios, caminhão pipa etc.)
Consumo não faturado não medido (combate a incêndios, favelas etc.)</td><td rowspan="5">Água não faturada</td></tr>
<tr><td rowspan="3">Perda de água</td><td rowspan="2">Perdas aparentes</td><td>Uso não autorizado (fraudes e falhas de cadastro)</td></tr>
<tr><td>Erros de medição (micromedição)</td></tr>
<tr><td rowspan="3">Perdas reais</td><td>Vazamentos nas adutoras e/ou redes de distribuição</td></tr>
<tr><td>Vazamentos e extravasamentos em reservatórios</td></tr>
<tr><td>Vazamentos em ramais prediais (a montante do ponto de medição)</td></tr>
</table>

Fonte: BEZERRA e CHEUNG (2013)

De acordo com as diretrizes da IWA, o ponto de partida para a análise é o volume total de água que ingressa no sistema, incluindo a água importada. Durante o processo de distribuição, a água pode ser categorizada como consumo autorizado ou perda. O consumo autorizado refere-se ao volume de água fornecido a clientes autorizados, sendo classificado como faturado ou não faturado e, ainda, subdividido em medido e não medido. Por outro lado, as perdas correspondem ao volume de água produzido que não é faturado pela concessionária, podendo ser classificadas em aparentes ou reais. Conforme ilustrado na Figura 3, a perda é calculada como a diferença entre o volume total de entrada no sistema e o consumo autorizado (TRATA BRASIL, 2024).

3.1.2 Perdas de água em sistemas de abastecimento

As perdas em um SAA correspondem à diferença entre o volume de água produzido, medido por meio da macromedição que registra o total de entrada no sistema de distribuição, e o volume efetivamente consumido pelos usuários, registrado pela micromedição por meio dos hidrômetros. Tais sistemas, inevitavelmente, apresentam perdas que impactam diretamente sua eficiência e capacidade de controle. Embora seja inviável eliminar completamente essas perdas, o volume observado na maioria dos sistemas brasileiros excede significativamente os níveis aceitáveis, dificultando a obtenção de um equilíbrio satisfatório do ponto de vista econômico e operacional. Nesse contexto, o método desenvolvido pela IWA tornou-se uma referência no Brasil para a análise e controle das perdas (BEZERRA e CHEUNG, 2013; TRATA BRASIL, 2024).

De acordo com a IWA, as perdas podem ser classificadas como reais (físicas) ou aparentes (comerciais).

3.1.2.1 Perdas Reais

As perdas reais decorrem de vazamentos que ocorrem ao longo das diferentes etapas de produção, como captação, tratamento, armazenamento e distribuição, estando geralmente associadas à falta de ações preventivas de controle de vazamentos, além de problemas operacionais e de manutenção (TRATA BRASIL, 2024; SILVA, 2021). A Tabela 1 apresenta as principais causas dessas perdas e as respectivas magnitudes em cada etapa do processo de produção.

Tabela 1: Origem e magnitude de perdas reais

Subsistemas	Origens	Magnitudes
Adução de Água Bruta	Vazamento nas tubulações Limpeza do poço de sucção	Variável, em função do estado das tubulações e da eficiência operacional
Tratamento	Vazamentos estruturais Lavagem de filtros Descarga de lodo	Significativa, em função do estado das tubulações e da eficiência operacional
Reserva	Vazamento estruturais Extravasamentos Limpeza	Variável, em função do estado das tubulações e da eficiência operacional
Adução de Água Tratada	Vazamentos nas tubulações Limpeza do poço de sucção Descargas	Variável, em função do estado das tubulações e da eficiência operacional
Distribuição	Vazamentos na rede Vazamentos em ramais Descargas	Significativa, em função do estado das tubulações e principalmente das pressões

Fonte: Adaptado de TRATA BRASIL (2024)

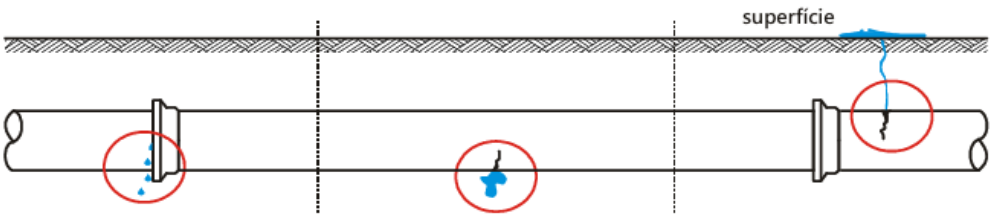
Os vazamentos são classificados em três tipos: inerentes, não visíveis e visíveis. Os vazamentos inerentes são aqueles resultantes de pequenos orifícios nas tubulações, juntas de conexões e peças especiais de controle, entre outros componentes do sistema. Esses vazamentos, por serem imperceptíveis, geralmente não são detectáveis por métodos convencionais, mas, em conjunto, podem contribuir significativamente para as perdas no sistema (BEZERRA e CHEUNG, 2013).

Os vazamentos não visíveis, por sua vez, não afloram à superfície e demandam o uso de equipamentos específicos, como detectores acústicos, para sua identificação. Apresentam vazões moderadas, e sua duração depende diretamente da frequência com que as concessionárias realizam inspeções regulares. No contexto brasileiro, a ausência de campanhas sistemáticas para a detecção desses vazamentos faz com que representem uma parcela expressiva das perdas reais nos sistemas de abastecimento (BEZERRA e CHEUNG, 2013).

Por fim, os vazamentos visíveis decorrem, em sua maioria, de rupturas em tubulações e peças do sistema, ocasionadas por fatores como sobrepressões, sobrecargas excessivas, defeitos estruturais, recalques do terreno ou cargas elevadas sobre o pavimento. Esses vazamentos apresentam vazões elevadas e, diferentemente dos demais, são prontamente identificáveis, pois se manifestam de forma evidente e imediata após sua ocorrência (BEZERRA e CHEUNG, 2013).

Contudo, a gestão torna-se crucial para o desenvolvimento dos sistemas e a redução das perdas. Nesse viés, faz-se necessário implementar ações para o controle de vazamentos onde destaca-se a gestão eficiente da pressão, incorporando a redução de pressão e a minimização de flutuações, apresentando resultados mais eficazes do que a simples diminuição da pressão da rede. Essa abordagem se mostra significativamente mais eficiente do que a simples diminuição da pressão na rede, pois não apenas reduz a probabilidade de rompimentos, como também minimiza os impactos negativos associados às oscilações de pressão, que são um fator preponderante para a ocorrência de vazamentos e rupturas (BEZERRA e CHEUNG, 2013). A Figura 4 apresenta os tipos de vazamentos mencionados e as respectivas principais ações para o controle de vazamentos.

Figura 4: Síntese das ações para o controle de perdas reais.



Vazamentos Inerentes	Vazamentos Não Visíveis	Vazamentos Visíveis
Não visíveis e não detectáveis por equipamentos de detecção acústica.	Não aflorantes à superfície, detectáveis por métodos acústicos de pesquisa.	Aflorantes à superfície, comunicados pela população.
Principais Ações		
- Redução de pressão. - Qualidade dos materiais e da execução da obra. - Redução do número de juntas.	- Redução de pressão. - Qualidade dos materiais e da execução da obra. - Pesquisa de vazamentos não visíveis.	- Redução de pressão. - Qualidade dos materiais e da execução da obra. - Diminuição do tempo de reparo.

Fonte: BEZERRA e CHEUNG (2013) apud TARDELLI (2005)

3.1.2.2 Perdas aparentes

As perdas aparentes referem-se aos volumes de água consumidos, porém não autorizados nem faturados, decorrentes de ligações clandestinas, fraudes nos hidrômetros, erros nas leituras desses dispositivos e falhas no cadastro comercial (TRATA BRASIL, 2024). A Tabela 2 apresenta as principais causas dessas perdas e suas respectivas magnitudes, que podem ser significativas, dependendo de certos fatores.

Tabela 2: Origem e magnitude de perdas aparentes.

Origens	Magnitudes
Ligações clandestinas / irregulares	Podem ser significativas, dependendo de: i. Procedimentos cadastrais e de faturamento; ii. Manutenção preventiva; iii. Adequação de hidrômetros; iv. Monitoramento do sistema.
Ligações sem hidrômetros	
Hidrômetros parados	
Hidrômetros que subestimam o volume consumido	
Ligações inativas reabertas	
Erros de leitura	
Número de economias errado	

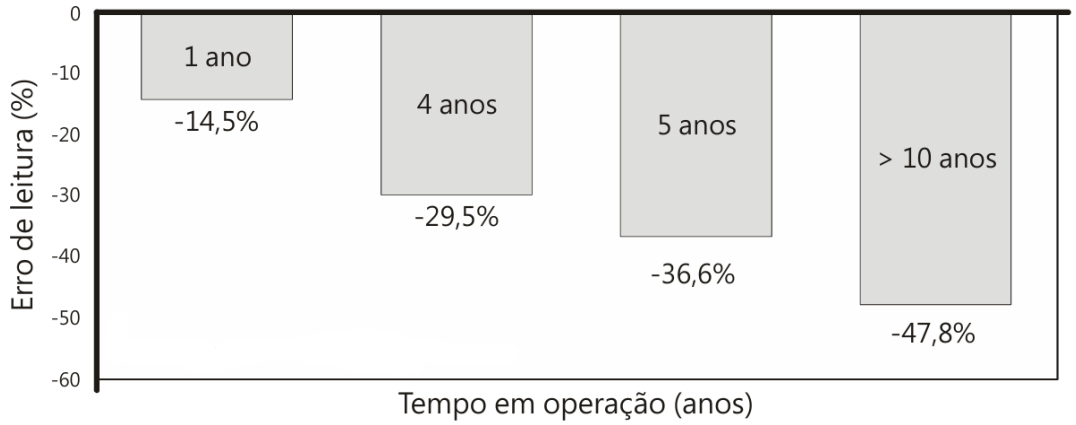
Fonte: Adaptado de TRATA BRASIL (2024)

Contudo, as perdas aparentes ocorrem principalmente de três formas: erros de medição, erros cadastrais ou na contabilidade do consumo de água e consumo não autorizado. O erro de medição pode ocorrer por diversos fatores, sendo os principais: fraudes, tempo de operação dos

medidores, dimensionamento inadequado, qualidade da água e do medidor, instalação incorreta e falhas na manutenção. Entre esses fatores, destaca-se o dimensionamento inadequado, que está relacionado à ampla diversidade de medidores disponíveis no mercado, os quais variam em tamanho, tipo e classe. Essa diversidade, aliada à grande variação nos padrões de consumo nos sistemas urbanos, dificulta a correta especificação dos medidores pelos profissionais responsáveis (BEZERRA e CHEUNG, 2013).

Outro aspecto relevante é o tempo de operação dos hidrômetros, que tende a comprometer sua precisão devido ao desgaste ao longo dos anos. Esse fator tem impacto direto na qualidade das medições realizadas (BEZERRA e CHEUNG, 2013). Para ilustrar essa questão, a Figura 5 apresenta os resultados de um estudo conduzido pelo engenheiro Maurício Fourniol (FOURNIOL, 2004). No estudo, foram realizados ensaios de laboratório para avaliar os erros de hidrômetros velocímetricos com vazão máxima de 3 m³/h. Os resultados indicaram uma perda média de 14,5% após o primeiro ano de uso, aumentando para 47,8% em hidrômetros com idade igual ou superior a 10 anos.

Figura 5: Erros de medição de hidrômetros resultantes dos ensaios de hidrômetros domiciliares velocimétricos em uma rede de distribuição de Maceió - AL.



Fonte: BEZERRA e CHEUNG (2013) apud FOURNIOL (2004)

Por sua vez, os erros cadastrais ou na contabilidade do consumo de água, podem ser atribuídos a diversos fatores, destacando-se: erro de leitura dos contadores, erro humano na análise dos dados e do faturamento, desatualização do cadastro dos clientes e mudança no banco de dados das empresas. Entre esses fatores, destaca-se o erro de leitura dos contadores onde leituristas inexperientes ou despreparados leem os medidores incorretamente ou registram os dados erroneamente, como ao posicionar uma vírgula no lugar errado. Adicionalmente, medidores com visores sujos, defeituosos ou localizados em áreas de difícil acesso também contribuem para a ocorrência de leituras equivocadas (BEZERRA e CHEUNG, 2013).

Figura 6: Hidrômetro com o mostrador ilegível

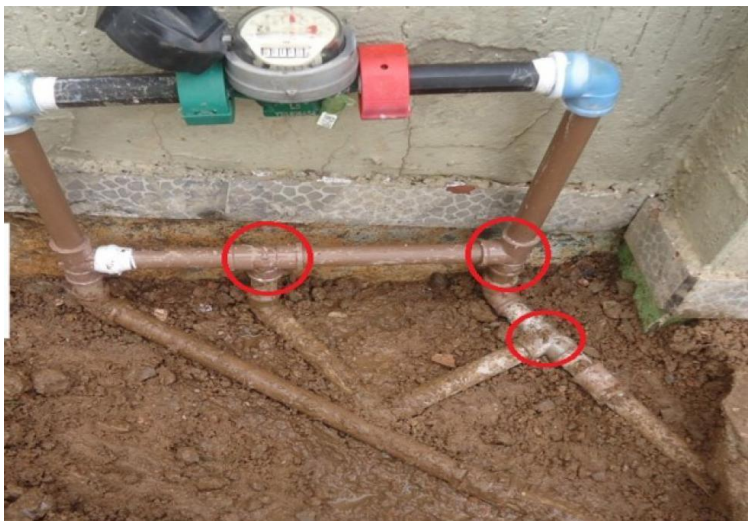
Fonte: BEZERRA e CHEUNG (2013)

Tais erros impactam diretamente os bancos de dados das empresas, comprometendo a integridade do sistema de informações. Quando essa integridade é afetada por negligência ou falhas, a avaliação dos programas de controle de perdas se torna prejudicada. Consequentemente, os dados de demanda de água, fundamentais para a modelagem de sistemas de distribuição e para estudos de planejamento, também são distorcidos, comprometendo a eficácia das decisões estratégicas (BEZERRA e CHEUNG, 2013).

Por fim, o consumo não autorizado refere-se a qualquer forma de manipulação intencional destinada a distorcer o real volume de água consumido, caracterizando-se como uma tentativa deliberada de obtenção de vantagens indevidas em relação à concessionária. As práticas mais recorrentes incluem ligações clandestinas e o uso fraudulento de água proveniente de descargas e hidrantes. No caso das ligações clandestinas, estas são realizadas por meio da instalação de um bypass — popularmente conhecido como "gato" — que desvia a água antes de sua passagem pelo hidrômetro, impedindo sua medição, como ilustrado na Figura 7. Outra prática comum é a religação do fornecimento de água sem a devida autorização da companhia, agravando ainda mais os prejuízos operacionais (BEZERRA e CHEUNG, 2013).

É importante ressaltar que as fraudes devem ser combatidas de maneira contínua pela companhia de saneamento, uma vez que qualquer indício de negligência na fiscalização aumenta significativamente a probabilidade do surgimento de novos fraudadores (VANDERLEI, 2023).

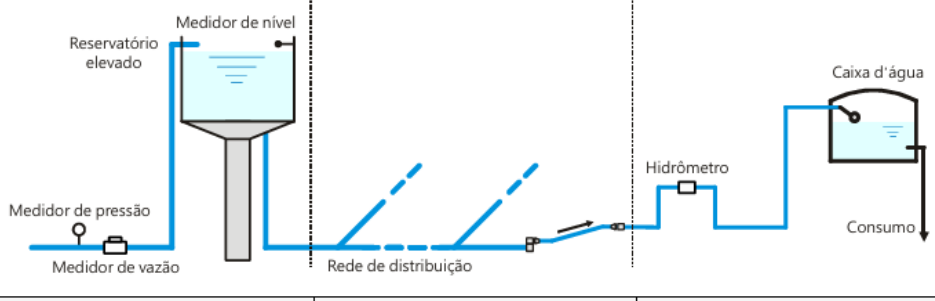
Figura 7: Ligação clandestina por by-pass.



Fonte: GAZETA DIGITAL (2018)

Com o objetivo de sintetizar e organizar de forma clara as principais ações voltadas ao controle e à redução de perdas aparentes, a Figura 8 apresenta um esquema ilustrativo que detalha as medidas necessárias para a efetiva gestão dessas perdas.

Figura 8: Síntese das ações para o controle de perdas aparentes.



Macromedição	Gestão Comercial	Micromedição
Medidores de vazão instalados nos reservatórios e setores das redes hidráulicas, cujos erros decorrem da falta ou inadequação do medidor, da falta de calibração, da submedição etc.	Falhas nos processos do sistema comercial, tais como erro no cadastramento de clientes, ligações clandestinas, fraudes etc.	Hidrômetros instalados na entrada dos imóveis que apresentam erros devido à submedição, agravados pela existência de caixas d'água ou pela má instalação dos hidrômetros.
Principais Ações		
- Instalação adequada dos macromedidores. - Calibração dos medidores de vazão.	- Sistema de gestão comercial adequado. - Combate às fraudes. - Controle de ligações inativas e clandestinas. - Qualidade da mão de obra.	- Adoção de hidrômetros adequados à faixa de consumo. - Troca periódica de hidrômetros. - Instalação correta de hidrômetros.

Fonte: BEZERRA e CHEUNG (2013) apud TARDELLI (2005)

É importante destacar que, independentemente do nível de perdas em que o sistema se enquadre, a continuidade das ações de controle é fundamental para a manutenção de índices satisfatórios. Dessa forma, o combate aos vazamentos deve ser encarado como uma prática constante e indispensável (BEZERRA e CHEUNG, 2013).

Em comparação com as perdas reais, as perdas aparentes causam um impacto financeiro significativamente maior no orçamento das companhias. Isso se deve ao fato de que a água perdida na tubulação acarreta custos relacionados à sua produção e distribuição. Por outro lado, os custos associados ao consumo não registrado equivalem ao valor do último metro cúbico consumido pelo usuário, o que os torna proporcionalmente mais elevados (GHAMKHAR et al., 2023).

3.2 Ciência de Dados

Nos últimos quinze anos, a intensificação dos investimentos em infraestrutura de negócios tem propiciado a coleta massiva de dados corporativos, culminando em um exponencial crescimento do volume de dados gerados diariamente (PROVOST e FAWCETT, 2016). O cenário mundial atual é marcado pela criação e expansão contínua de inúmeras bases de dados, em uma velocidade exponencial. De acordo com a International Data Corporation (IDC), estima-se que, em 2025, a quantidade de dados produzidos globalmente atingirá cerca de 163 zettabytes. A maioria desses dados é gerada por aplicativos e websites de redes sociais, bancos, e-commerce e sistemas internos das empresas (ESCOVEDO & KOSHIYAMA, 2020). Empresas e organizações enfrentam desafios não apenas em relação ao armazenamento, mas também ao processamento e à interpretação desses dados, com o objetivo de extrair informações valiosas e aplicá-las na tomada de decisões estratégicas (PROVOST e FAWCETT, 2016).

Nesse contexto, a Ciência de Dados se consolida como uma área fundamental. O conceito de Ciência de Dados refere-se à coleta de dados provenientes de diversas fontes para fins de análise, com o objetivo de apoiar a tomada de decisões, utilizando, em geral, grandes volumes de dados de maneira sistematizada. Portanto, Ciência de Dados não é uma ferramenta isolada, mas um conjunto de métodos destinados a embasar decisões com base em dados (ESCOVEDO & KOSHIYAMA, 2020). Suas aplicações abrangem desde otimizações empresariais, marketing e saúde pública, até a detecção de irregularidades, como será abordado neste trabalho.

Portanto, para processar e extrair informações úteis a partir desses dados, é essencial automatizar diversas etapas de coleta, processamento e análise, visando à tomada de decisões, uma vez que, devido ao grande volume de dados disponível, a execução manual dessas tarefas torna-se inviável. Nesse contexto, surge o Aprendizado de Máquina (ML), que visa identificar padrões e gerar insights a partir dos dados de forma automática e contínua. Com o uso de algoritmos inteligentes, o ML permite que sistemas aprendam com os dados históricos, ajustem-se às novas informações e aprimorem suas previsões e classificações, sem a necessidade de reprogramação constante, contribuindo diretamente para decisões mais assertivas e ágeis (ESCOVEDO & KOSHIYAMA, 2020).

3.2.1 Qualidade dos dados

No contexto da Ciência de Dados, um dos principais desafios enfrentados por analistas e cientistas de dados é a baixa qualidade dos dados. Dados de baixa qualidade incluem informações incompletas, inconsistentes, desatualizadas ou imprecisas. A má qualidade dos dados pode levar a conclusões e decisões equivocadas, com impactos negativos significativos para uma empresa ou organização. Isso ocorre porque erros nos dados tendem a se propagar ao longo das etapas de processamento e modelagem, afetando a precisão das conclusões. Por essa razão, as empresas devem adotar políticas e procedimentos rigorosos para garantir a qualidade dos dados (DATA SCIENCE ACADEMY, 2023)

Nesse contexto, surge o conceito de Qualidade dos Dados, que se refere ao grau de excelência dos dados e à adequação de seu uso para atender aos requisitos específicos de uma empresa ou organização. Portanto, a qualidade dos dados é fundamental para a tomada de decisões baseadas em dados, análises, ML e outras aplicações que dependem de informações precisas e confiáveis (DATA SCIENCE ACADEMY, 2023).

3.2.2 Pré-processamento dos dados

Para trabalhar com Ciência de Dados, é necessário dispor de dados, e quanto maior a quantidade de dados (desde que sejam de qualidade), mais fácil será identificar padrões ou encontrar fórmulas matemáticas que os expliquem. Esses dados podem ser provenientes de diversas fontes, estruturadas ou não, como planilhas, documentos, imagens, bancos de dados, entre outros. Na prática, muitos desses dados apresentam baixa qualidade, o que gera a necessidade de pré-processamento (ESCOVEDO & KOSHIYAMA, 2020).

O pré-processamento de dados é o conjunto de atividades que visa converter dados brutos em dados preparados, ou seja, em formatos úteis e eficientes. Geralmente, essa etapa inclui processos como a limpeza, tratamento de valores ausentes, normalização e redução de dimensionalidade (PINHEIRO, 2021). Entretanto, no presente trabalho, nem todas essas etapas foram aplicadas integralmente devido à metodologia empregada.

Sem um conjunto de dados adequadamente preparado, as análises e a modelagem tornam-se pouco confiáveis. Portanto, o pré-processamento é uma etapa fundamental que precede a realização de análises e modelagens, assegurando a qualidade e a confiabilidade das conclusões obtidas (PINHEIRO, 2021).

3.2.2.1 Analise exploratória (AED)

Com uma base de dados definida, é fundamental compreender as suas características e o comportamento de determinadas variáveis para obter os melhores resultados possíveis na aplicação de algoritmos de ML. Sem um entendimento adequado dos dados, corre-se o risco de aplicar modelos incorretos ou fazer suposições equivocadas (ESCOVEDO & KOSHIYAMA, 2020).

Para lidar com esses desafios, a Análise Exploratória de Dados (AED) desempenha um papel fundamental ao permitir a identificação de padrões e tendências nos dados. Seu objetivo principal é sintetizar as principais características do conjunto de dados, utilizando métodos estatísticos descritivos e visualizações gráficas para facilitar a interpretação e auxiliar na tomada de decisões (ESCOVEDO & KOSHIYAMA, 2020).

A estatística descritiva organiza, resume e descreve os dados sem extrapolar conclusões para além da amostra analisada. Essa abordagem permite obter uma visão geral das informações disponíveis, identificar padrões, verificar relações entre variáveis e compreender o comportamento de cada uma delas. Para isso, são utilizadas estatísticas sumarizadas, como média, mediana, desvio padrão, valores mínimos e máximos, entre outras métricas, possibilitando uma melhor compreensão da distribuição e variabilidade das variáveis analisadas (DNC, 2024).

3.2.2.2 Limpeza dos dados

Os dados originais da base podem conter muitas partes irrelevantes ou ausentes, não estando adequados para a aplicação de técnicas de ML. Nesse contexto, a preparação dos dados é essencial, pois envolve o processo de transformar dados brutos em um formato apropriado

para análise. Esse processo pode incluir o tratamento e/ou preenchimento de dados ausentes, a redução de ruídos, a identificação e remoção de valores aberrantes, bem como a resolução de inconsistências, garantindo que os dados estejam no formato correto para a aplicação (GOMES, 2019).

3.2.2.2.1 Tratamento dos dados ausentes

A ocorrência de valores ausentes está frequentemente associada à baixa qualidade dos dados e pode ser resultado de erros ou lacunas no processo de coleta manual de informações. Entre as principais causas, destacam-se falhas de digitação, mau funcionamento de dispositivos de aquisição de dados e omissão intencional por parte da fonte de informação, entre outros fatores (PINHEIRO, 2021).

O tratamento ou preenchimento desses dados ausentes pode ser realizado por diferentes abordagens, tais como:

- Remoção dos registros que contenham atributos nulos (GOMES, 2019);
- Preenchimento dos valores ausentes por meio de uma constante, média, mediana, moda do atributo ou um valor aleatório (GOMES, 2019);
- Representação dos dados ausentes pela sigla NaN (Not a Number).

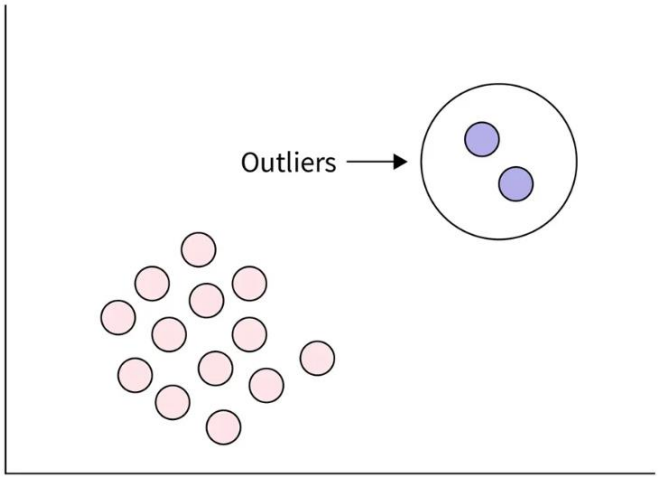
É fundamental destacar que cada tipo de dado pode demandar uma estratégia específica para o tratamento de valores ausentes. Nenhuma técnica isolada é suficiente para solucionar todas as inconsistências causadas pela ausência de dados (PINHEIRO, 2021).

3.2.2.2.2 Ruídos

Os ruídos, também denominados ‘outliers’, são valores que apresentam variações significativas em relação aos demais valores observados para um mesmo conjunto de dados (PINHEIRO, 2021), ilustrado na Figura 9.

No entanto, um dado ruidoso não deve ser automaticamente considerado inválido ou incorreto, pois sua interpretação depende do contexto da análise. Caso esses ruídos não sejam tratados adequadamente, algoritmos de ML podem sofrer interferências, resultando em inconsistências nos resultados, comprometendo a precisão das previsões e reduzindo a confiabilidade do modelo (GOMES, 2024).

Figura 9: Exemplo de dados ruidosos.



Fonte: GOMES (2024)

Para mitigar esses efeitos, diversas técnicas podem ser empregadas, tais como o método Z-Score, gráficos Box Plot, DBSCAN, entre outras abordagens estatísticas e de aprendizado de máquina. A escolha do método mais apropriado depende da natureza dos dados e do objetivo da análise (PINHEIRO, 2021). No presente trabalho, a detecção de ruídos desempenha um papel fundamental, uma vez que esses valores podem indicar possíveis irregularidades no consumo de água.

3.2.2.2.3 Resolução de inconsistências

A resolução de inconsistências é um processo fundamental na gestão de bases de dados, consistindo na identificação, correção e padronização de erros ou discrepâncias que possam comprometer a integridade das informações. Essas inconsistências podem surgir devido a erros humanos, falhas em sistemas de coleta de dados ou problemas na integração de múltiplas fontes, resultando em impactos negativos na qualidade das análises e na tomada de decisões (SRINIVANSAN et al., 2024).

Para mitigar esses problemas, torna-se fundamental identificar os tipos e as causas das inconsistências nos dados. Para isso, diversas técnicas e ferramentas de análise podem ser empregadas, tais como estatísticas descritivas, criação de perfis de dados, visualização gráfica e validação de registros, permitindo uma inspeção detalhada e estruturada do conjunto de dados (SRINIVANSAN et al., 2024).

Dentre as principais inconsistências observadas em bases de dados, destacam-se:

- **Valores duplicados**: registros idênticos inseridos mais de uma vez, o que pode distorcer análises estatísticas e afetar a confiabilidade dos resultados (JUNIOR, 2021);
- **Erros tipográficos**: falhas na digitação que geram variações indesejadas nos dados, dificultando sua agregação e comparação (JUNIOR, 2021);
- **Falta de padronização**: inconsistências decorrentes da representação de um mesmo dado de diferentes formas, prejudicando a uniformidade da base (JUNIOR, 2021);
- **Diferença de unidades de medida**: registros de um mesmo atributo em unidades distintas podem resultar em erros na conversão e interpretação dos dados (ESCOVEDO & KOSHIYAMA, 2020);
- **Valores fora do intervalo esperado**: determinadas variáveis possuem limites predefinidos e, quando os dados apresentam valores extremos ou incoerentes, podem indicar erros de entrada ou possíveis anomalias que necessitam de investigação (ESCOVEDO & KOSHIYAMA, 2020).

A adoção de métodos eficientes para a detecção e correção dessas inconsistências é essencial para garantir a qualidade dos dados, aprimorar a confiabilidade das análises e viabilizar a aplicação de técnicas avançadas.

3.2.2.3 Transformação de dados

Empresas que lidam com grandes volumes de dados, especialmente aquelas que operam no setor de abastecimento de água, frequentemente enfrentam o desafio de extrair valor significativo dessas informações de maneira eficiente. A presença de atributos com escalas distintas pode comprometer a análise e dificultar a identificação de padrões relevantes. Além disso, a alta dimensionalidade dos dados pode aumentar a complexidade computacional e reduzir a eficácia dos modelos analíticos (GOMES, 2019; HAYES & DOWNIE, 2024).

Diante desse cenário, torna-se essencial a transformação dos dados, um processo que visa converter dados brutos em um formato mais adequado para análise e modelagem. Essa etapa assegura a compatibilidade com os sistemas de destino, aprimora a usabilidade e melhora a qualidade das informações, sendo um componente fundamental no gerenciamento de dados. Além disso, desempenha um papel crucial na organização, análise e armazenamento de dados, garantindo maior eficiência nos processos analíticos (HAYES & DOWNIE, 2024).

Diversas técnicas podem ser empregadas para a transformação de dados, incluindo redução da dimensionalidade, normalização, padronização, transformação logarítmica e codificação de variáveis categóricas. No entanto, neste trabalho, será dada ênfase à redução de dimensionalidade e à normalização, as quais serão abordadas nos tópicos subsequentes.

3.2.2.3.1 Redução dos dados

O processamento de grandes quantidades de informação pode resultar em sobrecarga computacional, aumento do tempo de processamento e complexidade desnecessária nos modelos de análise, comprometendo a identificação de padrões e a detecção de ruídos. Esse fenômeno é frequentemente denominado de 'maldição da dimensionalidade' (GHAMKHAR et al., 2023). Nesse contexto, é importante destacar que a dimensionalidade diz respeito ao número de atributos (colunas) de um banco de dados, e não à quantidade de registros (linhas).

Nesse cenário, a redução de dados surge como uma solução eficaz, uma vez que busca diminuir o volume e a complexidade dos dados sem comprometer a integridade e a representatividade da informação original, permitindo o armazenamento de grandes quantidades de dados de forma mais eficiente. Além disso, a redução de dados também contribui para evitar o fenômeno conhecido como overfitting, que ocorre quando um algoritmo se ajusta excessivamente aos dados de treinamento, resultando em um modelo incapaz de fazer previsões ou conclusões precisas ao ser aplicado a novos dados (IBM, 2024).

As principais técnicas de redução de dados incluem agregação de cubo de dados, seleção de subconjuntos de atributos, redução da numerosidade e redução da dimensionalidade. Neste trabalho, será utilizada a técnica de redução da dimensionalidade.

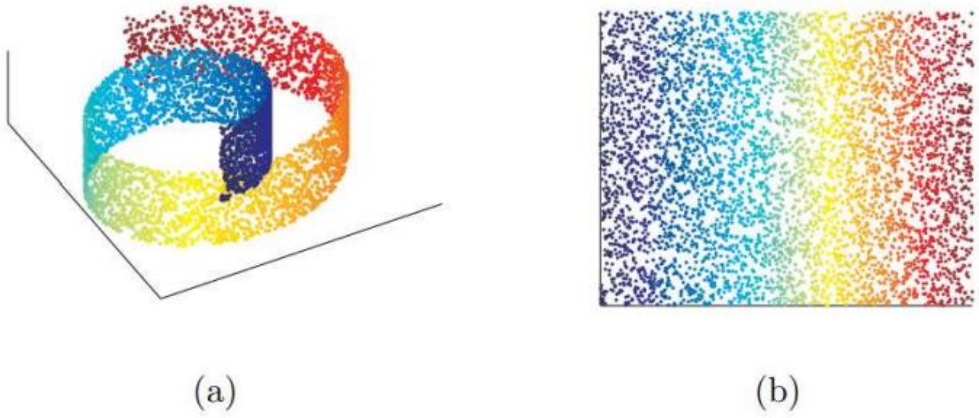
A redução da dimensionalidade é uma técnica utilizada para simplificar conjuntos de dados que possuem muitas dimensões (ou características) a serem inseridas em um modelo de treino, reduzindo o número de dimensões necessárias para descrever adequadamente os dados, enquanto ainda se capturam as propriedades significativas dos dados originais. Isso envolve a remoção de dimensões irrelevantes ou redundantes, com o objetivo de criar um modelo com um número reduzido de dimensões (MUREL & KAVLAKOGLU, 2024).

Além de acelerar o treinamento, a redução da dimensionalidade também é extremamente útil para a visualização de dados. Conjuntos de dados com quatro ou mais dimensões não permitem uma visualização direta. Reduzir o número de dimensões para duas (ou três) permite a plotagem de um conjunto de dados de alta dimensionalidade em um gráfico,

muitas vezes proporcionando insights valiosos por meio da detecção visual de padrões. Assim, a redução da dimensionalidade possibilita a representação de conjuntos de dados com três ou mais dimensões em somente duas (GERON, 2019; GOUVEIA, 2022). A Figura 10 ilustra o processo de redução da dimensionalidade, em que a Figura 10 (a) representa o conjunto de dados originais em três dimensões, enquanto a Figura 10 (b) demonstra a redução para duas dimensões.

Nessa representação, os dados são projetados em um espaço de menor dimensionalidade, preservando suas principais características e relações, mas com menos complexidade.

Figura 10: Redução da dimensionalidade.



Fonte: MOHRI et al (2018)

Para alcançar a redução da dimensionalidade, existem duas abordagens principais: a primeira consiste na seleção das características mais relevantes do conjunto de dados original, e a segunda, na extração de novas características que capturem as mesmas informações dos dados originais, transformando-as em um conjunto de variáveis de menor dimensão, sem perder a essência dos dados (GHAMKHAR et al., 2023). Embora os métodos de redução da dimensionalidade sejam diferentes na operação, todos eles transformam espaços de alta dimensão em espaços de baixa dimensão por meio de extração ou combinação de variáveis (MUREL & KAVLAKOGLU, 2024).

3.2.2.3.2 Normalização

A normalização é uma técnica de transformação de dados que transforma os valores das variáveis em uma escala comum, sem distorcer as diferenças nas faixas de valores entre os atributos (PINHEIRO, 2021). Esse processo é fundamental no pré-processamento de dados, especialmente para algoritmos de ML baseados em distância, como o DBSCAN.

A presença de escalas distintas entre os atributos pode comprometer a análise, uma vez que variáveis com valores absolutos mais elevados tendem a exercer maior influência nos cálculos de similaridade. Isso pode impactar negativamente a identificação de padrões e anomalias, tornando a normalização uma etapa essencial para garantir resultados mais equilibrados e representativos no treinamento dos modelos (ESCOVEDO & KOSHIYAMA, 2020).

Entre os principais métodos de normalização, destacam-se:

- **Escalonamento Min-Max:** Redimensiona os dados para um intervalo específico, geralmente entre 0 e 1, garantindo que o menor valor do conjunto seja transformado em 0 e o maior em 1 (PINHEIRO, 2021).
- **Padronização:** Transforma os valores para uma distribuição com média zero e desvio padrão unitário, permitindo comparações entre atributos com escalas diferentes (PINHEIRO, 2021).
- **Escalonamento Robusto:** Utiliza a distância interqualítica e apresenta maior poder de discriminação dos dados, podendo ignorar o outlier (PINHEIRO, 2021).

3.2.3 Aprendizado de máquina (ML)

Como mencionado anteriormente, o ML é um processo no qual algoritmos são capazes de aprender a partir de dados, identificar padrões e tomar decisões, visando detectar padrões e gerar insights de forma automática e contínua. Em outras palavras, o ML permite que computadores aprendam com os dados, sem que seja necessário programar explicitamente essa tarefa. A capacidade de um sistema 'aprender' e melhorar automaticamente com a experiência torna o ML uma ferramenta poderosa para solucionar problemas complexos que seriam difíceis ou impossíveis de serem resolvidos manualmente por seres humanos (ESCOVEDO & KOSHIYAMA, 2020).

Segundo Costa & Pyres (2024), de modo geral, o desempenho de um modelo de ML está diretamente relacionado ao tipo de algoritmo utilizado. Existem três tipos principais de algoritmos:

- **Classificação**: O objetivo do algoritmo de classificação é identificar a categoria à qual uma amostra pertence, retornando a classe prevista para cada caso.

- **Regressão**: O algoritmo de regressão visa prever um valor numérico com base nas características da amostra, resultando em um valor numérico estimado.
- **Agrupamento**: O algoritmo de agrupamento tem como finalidade agrupar amostras semelhantes, criando grupos com base nas características compartilhadas pelas amostras.

Além disso, de acordo com Costa & Pyres (2024), os modelos de ML podem ser classificados com base na presença de rótulos nos dados, que indicam a classificação desejada para um determinado conjunto de informações. Embora existam diversos algoritmos de ML, eles podem ser agrupados em quatro categorias principais:

- **Aprendizado Supervisionado**: Nesse tipo de aprendizado, a máquina recebe como entrada um conjunto de dados rotulados, ou seja, com a saída esperada previamente definida, para treinar o modelo. Após o treinamento, o modelo é capaz de realizar previsões para novos dados não rotulados. Esse tipo de aprendizado é amplamente utilizado em problemas de classificação e regressão.
- **Aprendizado Não Supervisionado**: Diferentemente do aprendizado supervisionado, o aprendizado não supervisionado não utiliza dados rotulados. O objetivo é identificar padrões e estruturas ocultas nos dados, sendo útil para a exploração inicial de dados e para revelar relacionamentos desconhecidos.
- **Aprendizado Semissupervisionado**: Essa técnica combina elementos do aprendizado supervisionado e do não supervisionado. Nessa abordagem, apenas uma parte dos dados é rotulada, enquanto o restante permanece sem rótulo. Com essa informação parcial, o modelo aprende padrões a partir dos dados rotulados e utiliza os dados não rotulados para generalizar esses padrões para novas instâncias.
- **Aprendizado por Reforço**: O aprendizado por reforço baseia-se em feedbacks obtidos por meio de tentativa e erro. O sistema atua em um ambiente e ajusta suas ações para maximizar a recompensa ao longo do tempo. Essa abordagem não depende de dados rotulados para aprender, mas se apoia em um processo contínuo de interação e aprimoramento.

3.2.4 DBSCAN

O DBSCAN, abreviação do termo ‘Density Based Spatial Clustering of Application with Noise’ (Agrupamento Espacial Baseada em Densidade de Aplicações com Ruído) é um algoritmo de aprendizado de máquina não supervisionado, proposto por Ester et al. (1996). Esse algoritmo identifica conjuntos (ou agrupamentos) em dados de entrada com base em suas densidades locais, sendo altamente eficaz para detectar conjuntos de formas arbitrárias e para reconhecer e separar ruídos nos dados, o que o torna uma ferramenta robusta para análise de dados complexos (GHAMKHAR et al., 2023).

O DBSCAN utiliza dois parâmetros fundamentais: o raio de distância (ε), que define a distância máxima para que dois pontos sejam considerados parte de um mesmo conjunto, e o número mínimo de pontos (MinPts), que estabelece a quantidade mínima de vizinhos necessária para garantir a densidade do conjunto. Um valor muito pequeno para ε pode resultar em uma grande quantidade de pontos sendo classificados como ruído, enquanto um valor excessivamente alto pode gerar conjuntos amplos e generalizados ou, em casos extremos, agrupar todas as instâncias em um único grupo. Da mesma forma, para o parâmetro MinPts, valores altos tendem a aumentar a quantidade de ruídos, enquanto valores muito baixos podem resultar em múltiplos agrupamentos de baixa representatividade. Assim, a escolha adequada desses parâmetros é essencial para o desempenho eficiente do algoritmo DBSCAN (MONTEIRO & CARL, 2020; ESTER et al., 1996).

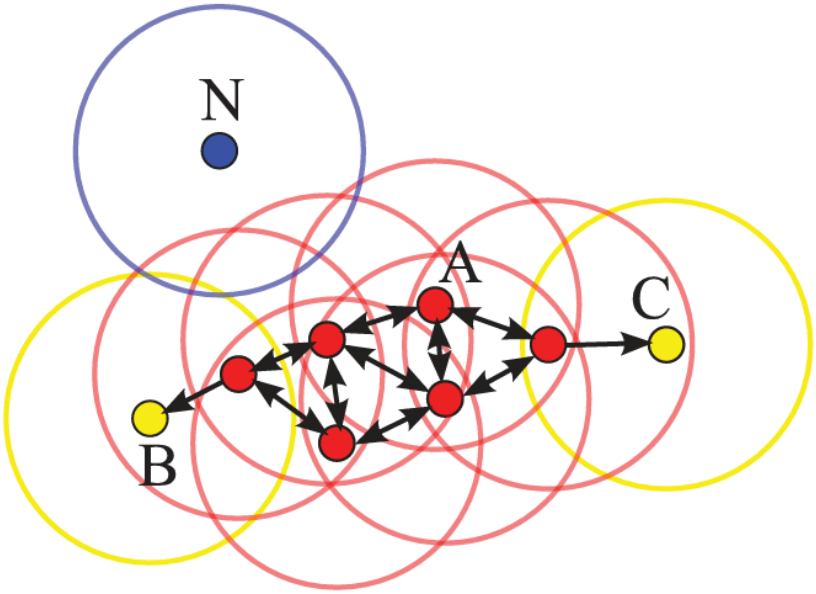
De acordo com MONTEIRO & CARL (2020), o algoritmo DBSCAN categoriza os dados em conjuntos com base na similaridade de seus vizinhos, definindo três tipos principais de pontos:

- **Ponto central**: Um ponto é classificado como central se possui MinPts pontos ou mais dentro de um raio ε de distância.
- **Ponto de fronteira**: Um ponto é considerado de fronteira quando não possui MinPts pontos dentro de um raio ε, mas está localizado dentro do raio ε de um ponto central.
- **ruído**: Um ponto é classificado como ruído se não está dentro do raio ε de um ponto central nem possui MinPts pontos dentro de seu próprio raio ε.

É importante ressaltar que os pontos de ruídos também são referidos como anomalias ou irregularidades, e este trabalho e projetado para detecta-los.

Com os tipos de classificação definidos, o algoritmo inicia escolhendo um ponto arbitrário e classificando-o como núcleo, borda ou ruído. Se o ponto for identificado como núcleo, todos os pontos em seu raio de distância (ε) são incluídos no mesmo conjunto, podendo ocorrer reclassificações. Após classificar todos os pontos na região do núcleo, o processo retorna ao primeiro passo, e o algoritmo continua até que todos os pontos estejam devidamente classificados (MONTEIRO & CARL, 2020). A teoria subjacente a esse algoritmo é exemplificada na Figura 11, onde se utiliza MinPts = 4.

Figura 11: Exemplo do funcionamento do algoritmo DBSCAN



Fonte: SHUBERT et al (2017)

Ao analisar a figura acima, observa-se que o ponto A e os demais pontos em vermelho são classificados como pontos centrais, pois a área ao redor de cada um deles, em um raio ε, contém pelo menos 4 pontos (incluindo o próprio ponto). Como todos esses pontos são mutuamente acessíveis, eles formam um único conjunto. Os pontos B e C, apesar de não serem centrais, são acessíveis a partir de A (por meio de outros pontos centrais) e, assim, também integram o conjunto. O ponto N, por outro lado, é classificado como ruído, pois não é um ponto central nem diretamente acessível (SHUBERT et al, 2017).

3.3 Séries temporais

Como mencionado anteriormente, o volume de dados gerados diariamente pelas empresas tem crescido exponencialmente nos últimos anos, sendo coletados e armazenados com o objetivo de extrair informações relevantes para a tomada de decisão. Para isso, é essencial definir técnicas eficientes que auxiliem nesse processo, permitindo análises mais

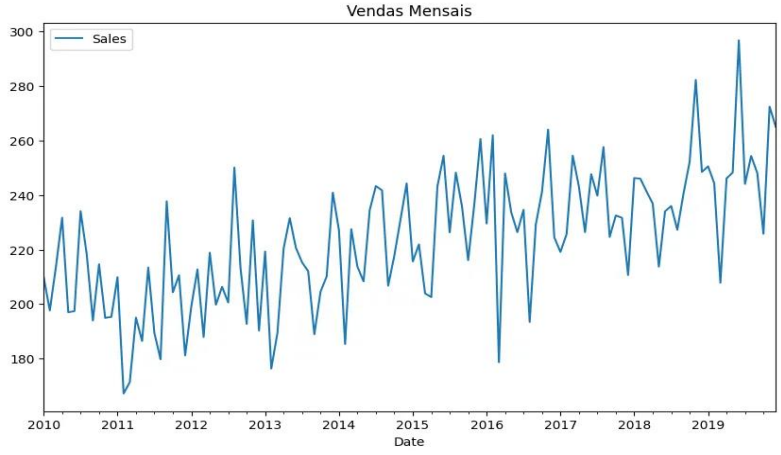
precisas e estratégicas. Dentre os diversos tipos de dados, aqueles registrados ao longo do tempo formam uma estrutura conhecida como série temporal.

Conforme Penteado (2024), séries temporais são conjuntos de dados coletados ao longo do tempo, nos quais as observações são registradas em intervalos regulares. Cada observação corresponde a um momento a um momento no tempo, possibilitando a análise da evolução de uma variável ao longo do tempo (FARIA, 2024). Segundo Morettin e Toloi (2006), o estudo dessas séries pode ser motivado pelo interesse em investigar o mecanismo gerador da série temporal, realizar previsões de valores futuros, descrever o comportamento observado ou identificar periodicidades relevantes nos dados. Um exemplo prático desse tipo de análise pode ser encontrado no estudo do consumo de água, onde as observações representam o volume consumido mensalmente ao longo do tempo, permitindo a investigação de padrões e variações no comportamento do consumo.

A previsibilidade de um evento depende de diversos fatores, como a compreensão dos elementos que o influenciam, a disponibilidade de dados históricos, a semelhança entre padrões passados e futuros e até mesmo o impacto das previsões no próprio acontecimento. Portanto, ao analisar uma série temporal, é fundamental considerar a presença e a interação desses fatores (FARIA, 2024).

A Figura 12 exibe a representação gráfica de uma série temporal, baseada nas vendas mensais de um determinado item em uma loja. No eixo das ordenadas, está representada a quantidade de vendas, enquanto no eixo das abscissas, encontra-se o período correspondente às vendas realizadas.

Figura 12: Representação gráfica de uma série temporal



Fonte: GUIMARÃES (2024)

De modo geral, a análise de séries temporais é comumente categorizada em duas abordagens principais: algoritmos de previsão e algoritmos de detecção de eventos. Os algoritmos de previsão têm como objetivo estimar o valor de uma determinada observação em um instante futuro, com base nos valores previamente registrados. Já os algoritmos de detecção de eventos buscam identificar padrões ou anomalias nos dados, permitindo a identificação de comportamentos atípicos ao longo do tempo (VIGNOLI, 2021).

3.3.1 Comportamento de uma série temporal

Segundo Mazzi (2018), o comportamento de uma série temporal reflete como os dados evoluem ao longo do tempo, podendo seguir padrões regulares ou apresentar variações inesperadas. Compreender esses padrões é essencial para análises precisas e tomadas de decisão informadas.

3.3.1.1 Componentes

Um dos principais objetivos da análise de séries temporais é decompor a série em um conjunto de componentes não observáveis (latentes), os quais podem estar associados a diferentes tipos de variações temporais (MAZZI, 2018). Ou seja, esse processo busca identificar e distinguir as diversas mudanças que ocorrem nos dados ao longo do tempo.

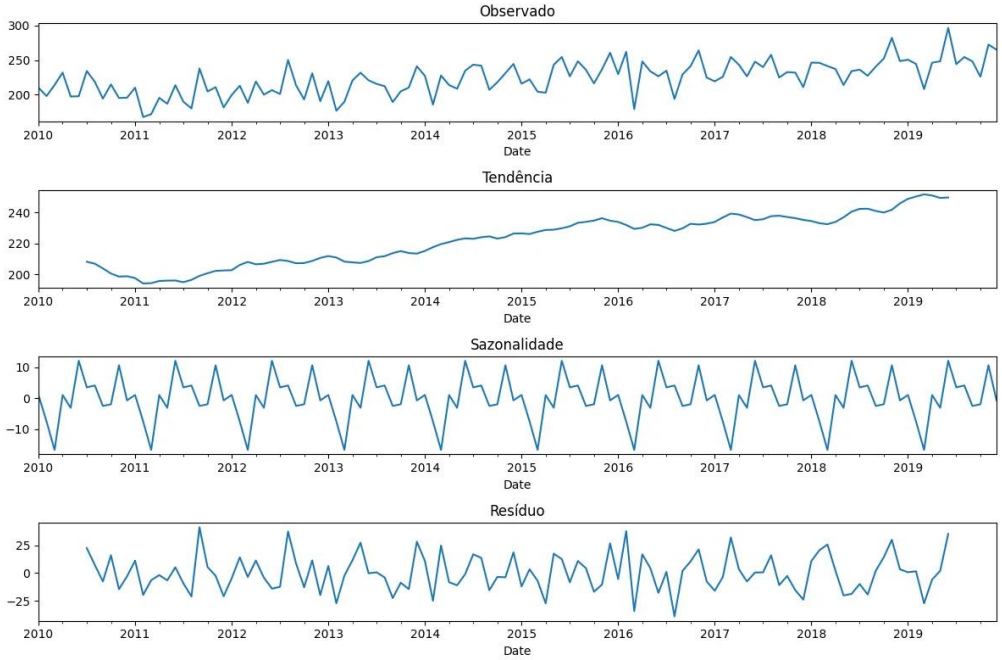
A análise de séries temporais geralmente envolve a decomposição dos dados em quatro componentes principais:

- **Tendência:** Representa o grau de variação da média dos dados da série ao longo do tempo, podendo indicar tendências de crescimento, declínio ou estabilidade. Essa variação pode ser identificada tanto por meio da análise visual quanto por métodos estatísticos, como médias móveis, regressão linear ou decomposição de séries temporais (FARIA, 2024);
- **sazonalidade:** Caracteriza-se pela ocorrência de padrões regulares de flutuação nos dados em períodos específicos, como mensais ou anuais, apresentando um comportamento cíclico e previsível. Assim como a tendência, essa variação pode ser identificada visualmente por meio de um gráfico de série temporal ou analisada por métodos estatísticos, como a autocorrelação e a decomposição de séries temporais (FARIA, 2024);

- **ciclicidade:** Trata-se de flutuações que ocorrem em intervalos de tempo mais longos e em uma escala temporal mais ampla do que a sazonalidade. Essas variações podem ser influenciadas por fatores econômicos, políticos ou tecnológicos (FARIA, 2024);
- **ruído:** Compreendem flutuações aleatórias que não são explicadas pelos componentes de tendência e sazonalidade. Essas variações podem resultar de fatores externos desconhecidos ou de imprecisões na medição (GUIMARAES, 2024).

A Figura 13 exibe a representação gráfica da decomposição da série temporal da Figura 12, construída a partir das vendas mensais de um determinado item em uma loja.

Figura 13: Decomposição da série temporal.



Fonte: GUIMARÃES (2024)

O primeiro gráfico mostrado na Figura 13, denominado 'Observado', representa a série temporal original. Os gráficos seguintes correspondem à decomposição dessa série. O segundo gráfico exibe a tendência, o terceiro representa a sazonalidade e o quarto apresenta os resíduos, que correspondem às variações aleatórias não explicadas pela tendência e sazonalidade.

A decomposição da série temporal permite uma análise mais detalhada dos dados, facilitando a compreensão dos fatores que influenciam a variação dos valores ao longo do tempo. Com essa abordagem, é possível isolar padrões e identificar tendências e sazonalidades, facilitando o desenvolvimento de estratégias e a tomada de decisões fundamentadas (GUIMARÃES, 2024).

3.3.2 Detecção de eventos

No contexto da análise de dados de séries temporais, é comum identificar mudanças significativas no comportamento da série em determinados pontos ou intervalos de tempo. Essas alterações, frequentemente, correspondem à ocorrência de eventos. Um evento detectado em séries temporais pode representar um fenômeno de relevância específica dentro de um determinado domínio de conhecimento, manifestando-se por meio de picos, variações de nível, mudanças abruptas, entre outros (CEFET-RJ, 2025).

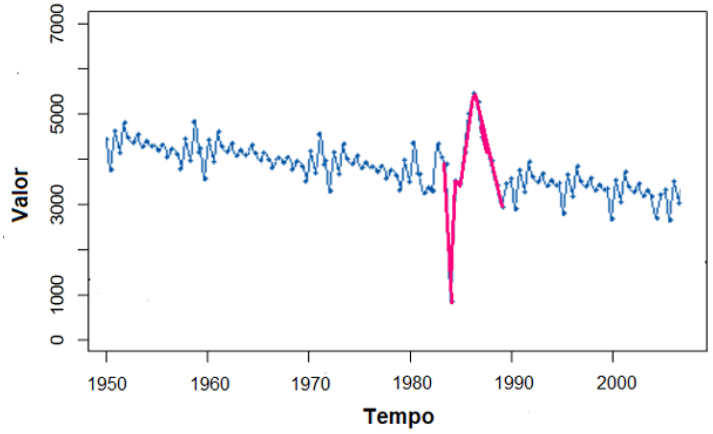
Um aspecto relevante na identificação de eventos em séries temporais é a determinação do período em que ocorrem, pois geralmente se manifestam como anomalias ou pontos de mudança.

3.3.2.1 Anomalias

As anomalias são observações discrepantes que divergem do comportamento esperado da série temporal. Essas discrepâncias podem resultar de erros de medição, mudanças estruturais ou eventos inesperados, como crises econômicas ou falhas em sistemas (MAZZI, 2018).

A partir da classificação proposta por Chandola et al., (2009) as anomalias em séries temporais podem ser categorizadas em três tipos principais: pontuais, contextuais e coletivas. As Anomalias pontuais correspondem às instâncias isoladas que diferem significativamente do padrão normal da série. Um exemplo clássico é uma transação financeira com um valor excepcionalmente alto, que pode indicar uma atividade incomum (CHANDOLA et al., 2009). A Figura 14 ilustra um exemplo de serie temporal contendo anomalias pontuais nos últimos pontos de dados entre 1980 e 1990.

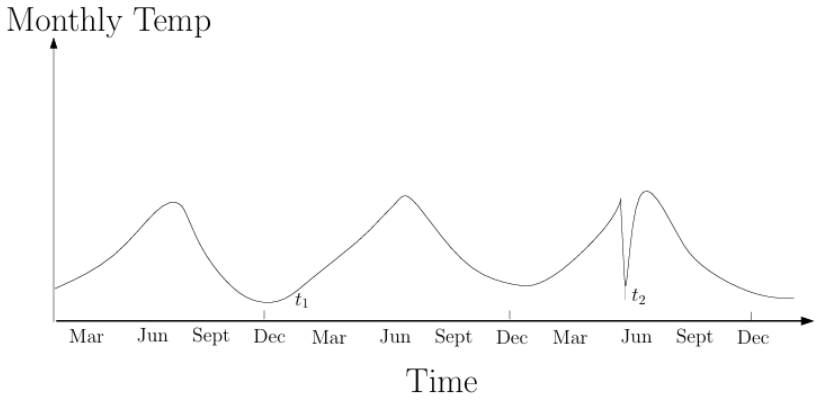
Figura 14: Exemplo de serie temporal com anomalias pontuais.



Fonte: VIGNOLI (2021)

As Anomalias contextuais, por sua vez, ocorrem quando um dado é considerado anômalo apenas dentro de um contexto específico (CHANDOLA et al., 2009). A Figura 15 apresenta um exemplo de anomalia contextual em uma série temporal de temperatura, onde uma temperatura pode ser normal durante o inverno (ponto t_1), mas, no verão, a mesma temperatura (ponto t_2) seria considerada uma anomalia.

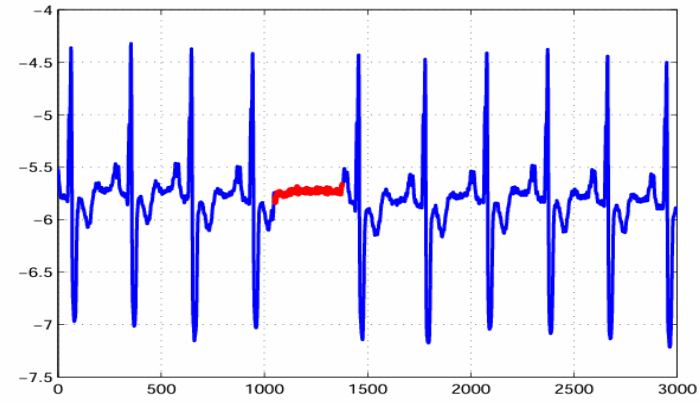
Figura 15: Exemplo de serie temporal com anomalias contextuais.



Fonte: CHANDOLA et al. (2009)

Por fim, as anomalias coletivas manifestam-se quando um grupo de pontos de dados, analisado em conjunto, apresenta um comportamento incomum, embora, individualmente, os pontos não sejam considerados anômalos (CHANDOLA et al., 2009). A Figura 16 apresenta um exemplo desse tipo de anomalia utilizando um eletrocardiograma humano, onde a região destacada representa uma anomalia, pois apresenta valores baixos por um período anormalmente longo, correspondente a uma contração atrial prematura.

Figura 16: Exemplo de serie temporal com anomalias coletivas.



Fonte: CHANDOLA et al. (2009)

A detecção de anomalias em séries temporais é essencial para identificar alterações, entender suas causas e impactos, sendo aplicável em diversos domínios. A premissa por trás das técnicas de detecção é a criação de um modelo probabilístico, estatístico ou algorítmico que

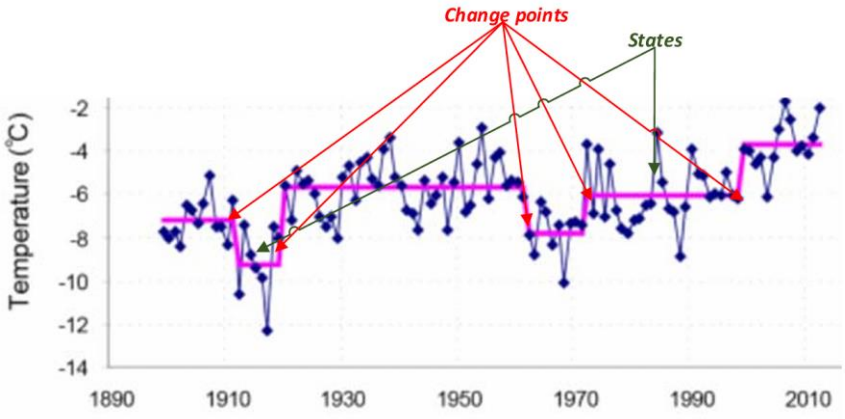
caracterize os dados normais. Os desvios em relação a esse modelo são utilizados para identificar valores discrepantes, ou seja, as anomalias (VIGNOLI, 2021).

3.3.2.2 Pontos de mudança

Um ponto de mudança refere-se a um instante no tempo em que as propriedades estatísticas de uma série temporal sofrem uma modificação abrupta. Esses pontos podem ocorrer devido a eventos externos ou mudanças internas na distribuição dos dados. A identificação de um ponto de mudança é essencial para detectar variações que podem impactar significativamente a operação de um sistema (BASSEVILE & NIKIFOROV, 1993).

As mudanças podem se manifestar em diversas características do processo, sendo classificadas com base nas alterações dessas características, como a média, a variância ou outros parâmetros estatísticos relevantes (BASSEVILE & NIKIFOROV, 1993). A Figura 17 apresenta um exemplo de série temporal contendo múltiplos pontos de mudança. Os dados representam as tendências anuais de temperatura em Spitsbergen no período de 1899 a 2010 (AMINIKHANGHAHI & COOK, 2016).

Figura 17: Exemplo de serie temporal com pontos de mudança.



Fonte: AMINIKHANGHAHI & COOK (2016)

A Figura 17 evidencia que, durante esse intervalo, o clima de Spitsbergen passou por seis regimes distintos. Cada um desses segmentos da série temporal é denominado estado, correspondendo a períodos em que os parâmetros que regem o processo permanecem inalterados. A transição entre dois estados distintos é marcada por um ponto de mudança. Portanto, o principal objetivo da detecção de pontos de mudança é identificar essas transições entre estados.

3.3.3 Padrões de consumo

No âmbito da análise de séries temporais aplicadas a Sistemas de Abastecimento de Água (SAA), os padrões de consumo representam sequências temporais que descrevem o comportamento do uso da água por parte dos usuários. A identificação e a compreensão desses padrões permitem o estabelecimento de perfis típicos de consumo, os quais servem como referência para a detecção de irregularidades, possibilitando ações corretivas, a alocação mais eficiente de recursos e o aprimoramento da gestão operacional do sistema.

Processos em conformidade com o controle apresentam variabilidade natural e fatores recorrentes de baixa magnitude. Por outro lado, processos fora de controle evidenciam fatores atípicos, como atividades fraudulentas ou comportamentos irregularidades. De acordo com Ghamkhar et al. (2023), diversos fatores podem levar ao consumo anormal, tais como:

- Consumo ilegal;
- Utilização por usuários não autorizados
- Consumo nulo por usuários regularmente cadastrados;
- Existência de usuários não reportados ou clandestinos;
- Alterações no padrão de consumo decorrentes de mudanças no número de ocupantes ou na duração da estadia;
- Erros de medição, como falhas nos equipamentos de leitura ou imprecisões no registro dos dados;
- Uso de fontes alternativas de abastecimento, como poços ou reservatórios particulares;
- Modificações na estrutura ou configuração da conexão de serviço.

A fim de detectar irregularidades, as companhias de saneamento realizam o monitoramento contínuo dos dados de consumo, geralmente com base em registros históricos de longo prazo. Por meio da análise dessas séries temporais, torna-se possível identificar padrões atípicos, como variações abruptas seguidas de estabilização, picos de consumo fora do esperado, quedas acentuadas e subsequente estabilização, registros prolongados de consumo zero, entre outros comportamentos fora da normalidade.

Quando identificadas tais irregularidades, procedimentos operacionais são acionados, incluindo a realização de inspeções em campo para averiguar a origem da irregularidade. Caso seja confirmada a ocorrência de atividade fraudulenta ou não conforme, a concessionária pode aplicar medidas corretivas, como o desligamento do fornecimento, a aplicação de penalidades

administrativas e a atualização do cadastro do usuário. Após a regularização e eventual reconexão, uma nova série temporal de consumo é iniciada para fins de monitoramento e controle subsequente.

3.3.4 Extração das características

Além dos métodos tradicionais de análise de séries temporais, é possível empregar métricas que avaliam a complexidade e a variabilidade dos dados ao longo do tempo. Entre essas métricas, destacam-se o Fator de Comprimento de Série Temporal (TSLF) e a Complexidade de Lempel-Ziv (LZC).

3.3.4.1 Fator de Comprimento de Série Temporal (TSLF)

O TSLF é uma das características extraídas das séries temporais para representar a regularidade de uma ocorrência ao longo do tempo. O cálculo do TSLF é baseado na proporção entre o tamanho da série com valores únicos registrados e o tamanho total da série. Ou seja, é definido como a razão entre o número de pontos de consumo distintas registrados em uma série e o número total de pontos possíveis ao longo do período analisado (GHAMKHAR et al., 2023). A formulação deste cálculo encontra-se na Equação 1:

$$TSLF = \frac{L_u}{L_t} \tag{1}$$

Em que L_u representa o número de unidades de consumo distintas registradas na série temporal e L_t correspondem ao comprimento total da série temporal.

Portanto, essa métrica assume o valor 1 quando há consumo em todos os meses e os valores são distintos entre si. No entanto, a presença de períodos sem registro de consumo ou a repetição de valores ao longo do tempo fará com que o TSLF resulte em um valor inferior a 1, refletindo a falta de variabilidade e regularidade nos dados temporais (GHAMKHAR et al., 2023).

3.3.4.2 Complexidade de Lempel-Ziv (LZC)

O LZC é uma métrica derivada dos trabalhos de Lempel e Ziv (1976), que mede a complexidade ou a "aleatoriedade" de uma sequência finita de dados. Esse conceito é amplamente utilizado em diversas áreas, incluindo compressão de dados, biologia computacional e análise de séries temporais.

O método supracitado consiste em quantificar a complexidade de uma sequência com base na contagem do número de padrões distintos dentro dela. A métrica é obtida através de um processo iterativo no qual um novo padrão é registrado toda vez que uma subsequência inédita é identificada. Esse processo reflete a taxa de crescimento da complexidade da sequência, sendo maior para sequências mais aleatórias e menor para sequências com padrões altamente repetitivos (Lempel e Ziv, 1976).

Segundo Ghamkhar et al., (2023), dada uma sequência finita como a série temporal de entrada $X = \{C_i | C_i \in R, 1 \leq i \leq n\}$, onde C_i corresponde à série temporal de entrada, o cálculo da LZC envolve quatro etapas essenciais.

1) Matematicamente, a sequência finita X é convertida em uma sequência simbólica $S = \{s_i\}_1^n$, composta por "zeros" e "uns". Esse processo simplifica a série, substituindo valores numéricos contínuos por um formato padronizado, o que facilita a identificação de padrões. Para realizar essa conversão, cada ponto da sequência é comparado a um valor de limiar T, conforme ilustrado na Equação 2:

$$s_i = \begin{cases} 0 & Min \leq x_i \leq T \\ 1 & T < x_i \leq Max \end{cases} \quad (2)$$

Em que Min, T e Max correspondem, respectivamente, ao mínimo, o limiar e ao máximo da série temporal de entrada. Para determinar o valor do limiar, a série temporal completa é dividida uniformemente em K intervalos de tamanho igual entre os valores Min e Max utilizando-se símbolos correspondentes $\{0,1, \dots, k-1\}$. Essa divisão é representada na Equação 3:

$$s_i = \begin{cases} 0 \quad , & Min \leq x_i \leq T \\ 1 \quad , & T < x_i \leq 2T \\ \vdots & \vdots \\ k-1, & (k-1)T < x_i \leq Max \end{cases} \quad (3)$$

Dessa forma, uma série de valores numéricos se transforma em uma sequência de zeros e uns, algo como 0, 1, 1, 0, 1, ... Esse processo pode ser estendido para criar sequências com mais símbolos, dividindo o intervalo de valores em várias faixas (ou bins) ao invés de apenas duas.

2) A partir da sequência resultante S, identificam-se os padrões distintos ao analisar a sequência da esquerda para a direita. Para ilustrar, considere que $S_{i,j}$ representa uma

subsequência de S, começando na posição i e terminando na posição j; o conjunto de todas as subsequências é denotado por $W(S)$, como ilustrado na Equação 4.

$$W(S) = \{S_{i,j}, 1 \leq i \leq n \,\&\, j \geq i\} \tag{4}$$

3) Após identificar todos os padrões únicos na sequência resultante S, conta-se todos os padrões diferentes que foram encontrados. Esse valor corresponde ao LZC da sequência, denotado de $c(n)$. O processo citado acima, começa com a análise da sequencia S da esquerda para a direita, com um conjunto vazio $W(S) = \{\}$ que armazenará todas as subsequências distintas encontradas. Em cada passo, uma subsequência $S_{i,j}$ é comparada com as subsequências presentes em $W(S)$. Caso $S_{i,j}$ já estiver em $W(S)$, acrescenta-se 1 ao valor de j. Caso contrário, a subsequência $S_{i,j}$ é inserida em $W(S)$, indicando um novo padrão. Em seguida, os índices são ajustados para $i = i + j$ e $j = 1$. Esse processo continua até que $i + j \leq n$, onde n é o comprimento total de S.
4) Para possibilitar a comparação do índice de complexidade entre séries temporais de diferentes tamanhos, a LZC pode ser normalizada utilizando uma medida baseada no número de símbolos únicos (α) presentes na sequência de entrada. Lempel & Ziv (1976) demonstraram que a complexidade LZC possui um limite superior aplicável a séries de qualquer comprimento, sendo este limite calculado conforme a Equação 5:

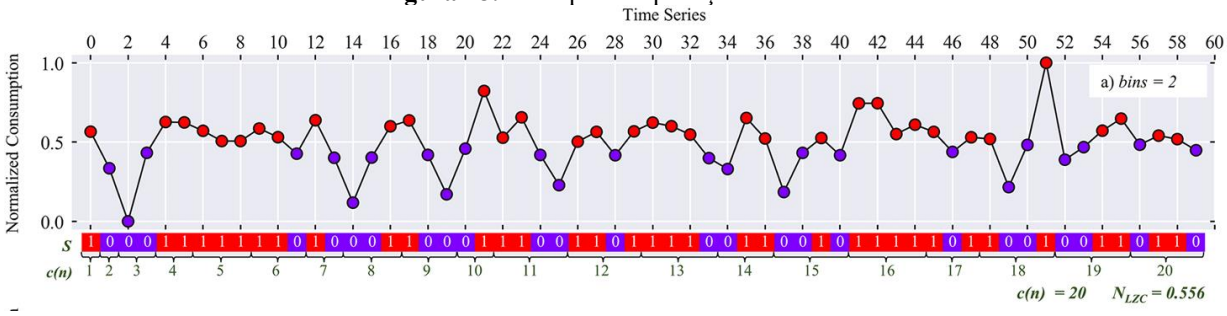
$$u(n) < \frac{n}{(1 - (\varepsilon_n)\log_\alpha(n)}\ , \varepsilon_n = 2\frac{1 + \log_\alpha \log_\alpha(\alpha n)}{\log_\alpha(n)} \tag{5}$$

Assim, para obter os valores normalizados de LZC (N_{LZC}), em função do comprimento reduzido da série temporal, calcula-se a razão entre o valor de LZC e o valor de seu limite superior, conforme demonstrado na Equação 6:

$$N_{LZC} = \frac{c(n)}{u(n)} \approx (\frac{c(n)}{n})(1 - \varepsilon_n)\log_\alpha(n) \tag{6}$$

A Figura 18 ilustra um exemplo da aplicação do LZC em uma série temporal normalizada, onde é exibido os valores de consumo ao longo do tempo. A discretização dos dados foi realizada utilizando dois intervalos (bins = 2), convertendo os valores contínuos da série em uma sequência binária.

Figura 18: Exemplo da aplicação do LZC.



Fonte: GHAMKHAR et al. (2023)

A Figura 18 exibe os valores de consumo ao longo do tempo, onde os pontos em vermelho e roxo representam os valores discretizados da série. Na parte inferior, observa-se a representação simbólica da sequência (S), onde os valores foram transformados em 0s e 1s com base em um limiar de separação. A métrica $c(n)$, indicada na imagem, representa o número de padrões distintos identificados ao longo da sequência, enquanto o valor normalizado $N_{LZC} = 0{,}556$ permite a comparação com outras séries temporais de diferentes tamanhos.

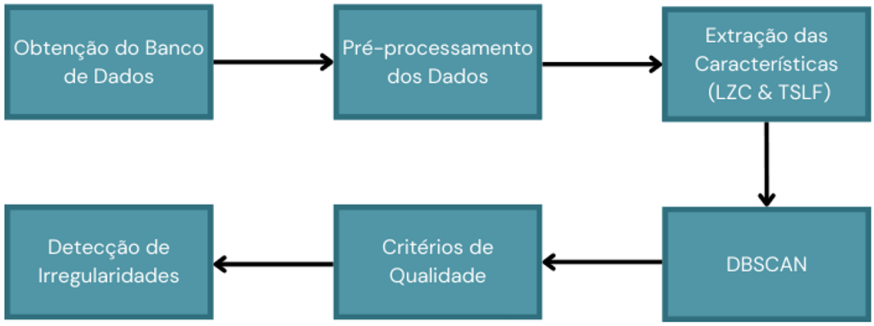
Essa abordagem possibilita a identificação de padrões de consumo e eventuais irregularidades, uma vez que séries mais estruturadas e previsíveis apresentam menores valores de LZC, enquanto padrões irregulares tendem a gerar valores mais elevados, os quais podem sinalizar comportamentos irregulares.

4 METODOLOGIA

A metodologia utilizada neste estudo baseia-se na abordagem descrita por Ghamkhar et al., (2023), cujo trabalho apresentou uma abordagem eficiente para a análise de detecção de irregularidades, demonstrando um elevado nível de precisão na identificação de comportamentos fraudulentos em dados de consumo. A estrutura metodológica proposta no referido artigo foi adaptada para a realidade da presente pesquisa, visando garantir a aplicabilidade dos métodos em uma nova base de dados.

Posto isso, a metodologia deste trabalho divide-se em algumas etapas como: Conjunto de dados, pré-processamento dos dados, características extraídas do conjunto de dados, DBSCAN, critérios de qualidade e detecção de irregularidades. Dessa forma, a metodologia utilizada neste trabalho é descrita pelo fluxograma da Figura 19.

Figura 19: Fluxograma representativo da metodologia utilizada



Fonte: Autor (2025)

4.1 Conjunto de dados

A princípio, utilizou-se dados fornecidos pela companhia de serviços de saneamento básico ENOPS Engenharia SA, que presta serviços com foco em redução de perdas e na otimização operacional. Dentre os arquivos disponibilizados, encontra-se o cadastro comercial de um setor de abastecimento de uma cidade brasileira. Considerando que o consumo não autorizado é uma questão social, os dados foram recebidos sob a condição de anonimato. Como medida de proteção à privacidade e aos interesses da ENOPS, o setor será o referenciado como Setor 1.

Posto isso, foram obtidos dados comerciais referentes ao consumo de água por ligação, por meio de planilhas eletrônicas no formato xlsx, organizadas por ano. Foram disponibilizados dados anuais referentes ao período de janeiro de 2018 a dezembro de 2022, totalizando cinco planilhas, com informações referentes a aproximadamente 106,5 mil matrículas. Cada planilha

reúne informações referentes aos usuários, incluindo o consumo medido, a data da leitura, o código do hidrômetro e a data de sua instalação. A Tabela 3 apresenta essas informações, juntamente com o formato dos dados.

Tabela 3: Estrutura dos dados obtidos.

Dados	Formato	Unidade
Matricula	Código	-
Setor de Abastecimento	Código	-
Volume Consumido	Número Inteiro	m^3
Data da Leitura	Data	Dia/Mês/Ano
Número de série do Hidrômetro	Código	-
Data de Instalação do Hidrômetro	Data	Dia/Mês/Ano

Fonte: Autor (2025)

É relevante destacar que os dados fornecidos não incluem informações sobre a categoria da ligação, a qual classifica os consumidores conforme o tipo de uso da água, como residencial, comercial, industrial ou público. Além disso, verificou-se a ausência de informações sobre a situação da ligação, que indica se o imóvel está ativo, inativo, cortado ou suprimido.

Adicionalmente, parte do período analisado coincide com o surto da Covid-19 no Brasil, durante o qual muitas pessoas permaneceram em casa, resultando em um possível aumento no consumo de água. Esse fator pode ter impacto direto na análise das séries temporais de consumo, influenciando a identificação de padrões e possíveis irregularidades, uma vez que mudanças no comportamento dos usuários podem ter ocasionado desvios atípicos nos dados.

Além disso, foi disponibilizada uma planilha com informações sobre as matrículas com potencial de fraude. Esta planilha inclui dados como setor de abastecimento, bairro e quantidade de meses associados a cada matrícula considerada fraudulenta, com informações referentes a aproximadamente 2424 matrículas. Imóveis com três ocorrências de status 'desocupado', 'demolido' ou 'fechado' não foram considerados nesta planilha. A Tabela 4 apresenta essas informações, juntamente com o formato dos dados de cada coluna.

Tabela 4: Estrutura da planilha com os possíveis fraudadores.

Dados	Formato	Unidade
Matrícula	Código	-
Setor de Abastecimento	Código	-
Nome do Bairro	Texto	-
Período de associação	Número	Meses
Data de Instalação do Hidrômetro	Data	Dia/Mês/Ano

Fonte: Autor (2025)

Com base nos dados comerciais obtidos, foram excluídos os usuários com registros de consumo em menos dois quintos do período analisado, equivalente a 24 meses, por serem considerados usuários novos e, provavelmente, não apresentarem medidores defeituosos ou adulterados, uma vez que foram conectados recentemente à rede.

4.2 Pré-processamento de dados

Após a obtenção dos dados, é necessário realizar diversas considerações para preparar o conjunto de dados de maneira adequada ao uso em algoritmos de ML, visando aprimorar o desempenho dos modelos e aumentar a confiabilidade dos resultados. Nesse contexto, a limpeza de dados é uma etapa fundamental, que consiste em transformar dados brutos em um formato apropriado para análise.

Um dos principais aspectos da limpeza de dados é o preenchimento de valores ausentes. Os dados de consumo de água frequentemente contêm valores ausentes ou incorretos, comumente associados a problemas de qualidade na coleta manual, resultando em erros ou lacunas. Esse tipo de medição, sujeita a falhas humanas, aumenta a probabilidade de omissões e inconsistências, o que leva a registros incompletos na base de dados (GHAMKHAR et al., 2023). Assim, nos casos em que um ou mais meses de dados estejam faltando, utiliza-se o valor NaN, que indica dados não disponíveis ou não registrados, sugerindo que estão ausentes ou foram omitidos.

Outra etapa importante no pré-processamento é a normalização, uma técnica de transformação de dados que transforma os valores das variáveis em uma escala comum, sem distorcer as diferenças nas faixas de valores entre os atributos. É relevante destacar que essa técnica é aplicada a cada atributo individualmente. Para garantir consistência, métodos de normalização baseados em características, como o escalonamento Min-Max, são geralmente empregados. Esse método define valores mínimos e máximos para cada atributo dentro do

intervalo de 0 a 1 (PINHEIRO, 2021), sendo a equação da normalização Min-Max descrita como:

$$x' = \frac{x - min\ (x)}{max(x) - min\ (x)} \tag{7}$$

Em que x' representa o valor normalizado, x é o valor original, e $min\ (x)$ e $max(x)$ correspondem, respectivamente, ao valor mínimo e ao valor máximo dentro do conjunto de dados considerado.

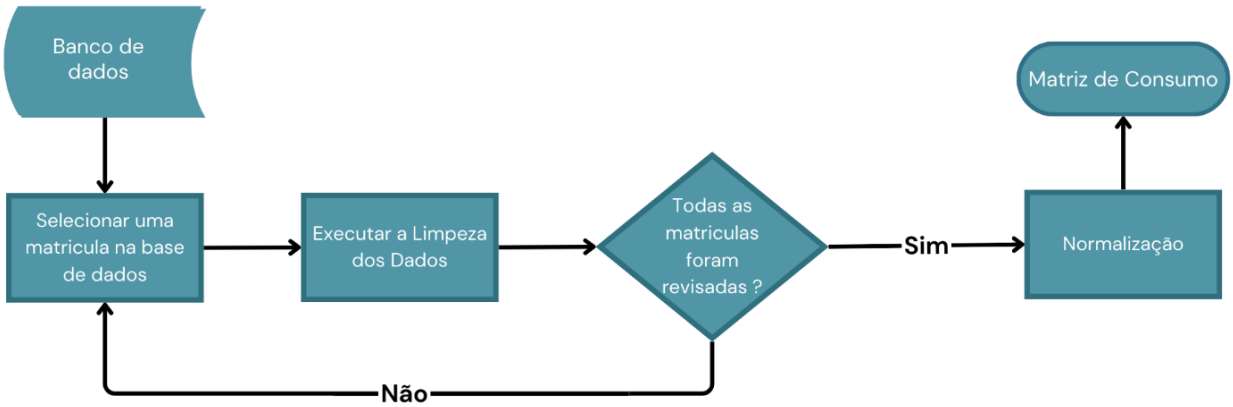
Após a etapa de limpeza, as séries temporais de consumo devem ser normalizadas para adequação ao algoritmo de ML. Esse processo gera a matriz de consumo, conforme ilustrado na Equação 8:

$$C = \begin{Bmatrix} C_{0,0} & \cdots & C_{0,i} \\ \vdots & \ddots & \vdots \\ C_{j,0} & \dots & C_{j,i} \end{Bmatrix} \begin{matrix} i \in \{0, \dots, n\} \\ , j \in \{0, \dots, m\} \end{matrix} \tag{8}$$

Em que $C_{j,i}$ é um elemento na linha j e coluna i da matriz que representa o consumo do j-esimo mês e do i-esimo consumidor, m é o número total de leituras mensais do medidor, e n é o número total de matrículas de usuários. Vale ressaltar que m equivale a 60 meses, tendo em vista que este é o período analisado (GHAMKHAR et al., 2023).

Para ilustrar melhor as etapas do pré-processamento de dados, a Figura 20 apresenta o fluxograma representativo da metodologia desta seção.

Figura 20: Fluxograma representativo da metodologia utilizada para o pré-processamento dos dados.



Fonte: Adaptado de GHAMKHAR et al., (2023)

4.3 Redução da dimensionalidade

Conforme mencionado no tópico 3.2.2.3.1, a redução da dimensionalidade é obtida por meio de técnicas específicas que buscam simplificar a estrutura dos dados sem perda

significativa de informações relevantes. Para lidar com essa questão, optou-se inicialmente pela aplicação da técnica de extração de características. Isso envolve a transformação dos dados originais em um conjunto de variáveis de menor dimensão que ainda mantem a essência dos dados originais, permitindo que os algoritmos de aprendizado de máquina trabalhem de forma mais eficiente e com menos interferência de ruido.

O procedimento de extração de características consiste em codificar os dados originais, de dimensão m, em um conjunto de características de dimensão d, resultando em um conjunto de dados com d dimensões. Esse processo resume de forma eficaz as principais características dos dados brutos, ao mesmo tempo que reduz a dimensionalidade, o que pode contribuir para um processamento mais rápido e diminuir o risco de overfitting na maioria dos casos (GHAMKHAR et al., 2023).

É importante mencionar que o algoritmo de ML utilizado percorre os pontos do conjunto de dados apenas uma vez e precisa calcular a distância entre cada par de pontos no conjunto de dados. Dessa forma, a complexidade computacional será, para todo o algoritmo, da ordem de $O(n^2 * D)$, onde n representa o numero total de pontos e D é o custo de cálculo das distâncias. No entanto, quando estruturas de indices eficientes sao empregadas e a dimensao dos dados é baixa ($d \leq 5$), essa complexidade pode ser reduzida para $O(n * \log n)$.

Portanto, para otimizar esse processo, este trabalho adota cinco características baseadas na complexidade das séries temporais, reduzindo o número de dimensões e, consequentemente, o custo computacional do algoritmo (SHUBERT et al, 2017; GHAMKHAR et al., 2023).

4.4 Características extraídas do conjunto de dados

Para facilitar a análise e reduzir a complexidade dos dados, foram extraídas características específicas das séries temporais, essenciais para o treinamento do modelo de ML. As séries temporais representam conjuntos de dados organizados cronologicamente, permitindo a análise de padrões e tendências ao longo do tempo. Neste trabalho, utilizam-se as características do TSLF e da LZC para a detecção de irregularidades, sendo ambas recomendadas como indicadores confiáveis para conjuntos de dados com baixa resolução, como as leituras de medidores de água (GHAMKHAR et al., 2023). As subseções a seguir fornecem uma explicação detalhada de cada uma dessas características.

No algoritmo de ML, os dados de entrada são organizados em um conjunto com *n* instâncias ou usuários, cada um descrito por cinco dimensões ou características específicas.

Essas dimensões são compostas pelas seguintes características: uma associada ao TSLF e quatro referentes ao LZC. Com base nessas características, foi gerada a matriz representada na Equação 9.

$$D = \begin{Bmatrix} f_{0,1} & f_{0,2} & f_{0,3} & f_{0,4} & f_{0,5} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ f_{j,1} & f_{j,2} & f_{j,3} & f_{j,4} & f_{j,5} \end{Bmatrix}, j \in \{0, \dots, n\} \quad (9)$$

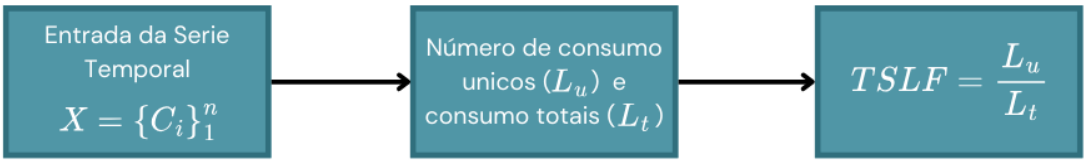
Em que $f_{j,1}$ representa a característica associada ao TSLF e $f_{j,2}$, $f_{j,3}$, $f_{j,4}$, $f_{j,5}$ correspondem as características relacionadas ao LZC.

4.4.1 Fator de comprimento da série temporal (TSLF)

Como mencionado no tópico 3.3.4.1, o TSLF é uma das características extraídas das séries temporais para representar a regularidade de uma ocorrência ao longo do tempo. Sua utilização, neste trabalho, visa superar inconsistências e lacunas nas séries temporais, especialmente em períodos com ausência de consumo registrado (GHAMKHAR et al., 2023).

Para ilustrar melhor essa etapa de obtenção do TSLF, a Figura 21 apresenta o fluxograma representativo do algoritmo utilizado nesta subseção.

Figura 21: Fluxograma representativo do algoritmo do TSLF.



Fonte: Adaptado de GHAMKHAR et al., (2023)

4.4.2 Complexidade de Lempel-Ziv (LZC)

Conforme mencionado no tópico 3.3.4.2, LZC é uma métrica de complexidade utilizada para avaliar a variabilidade e a imprevisibilidade de uma série temporal. Neste trabalho, sua aplicação visa identificar padrões complexos a partir da análise do consumo de água, permitindo a detecção de comportamentos atípicos que podem indicar irregularidades, como variações incomuns no consumo (GHAMKHAR et al., 2023).

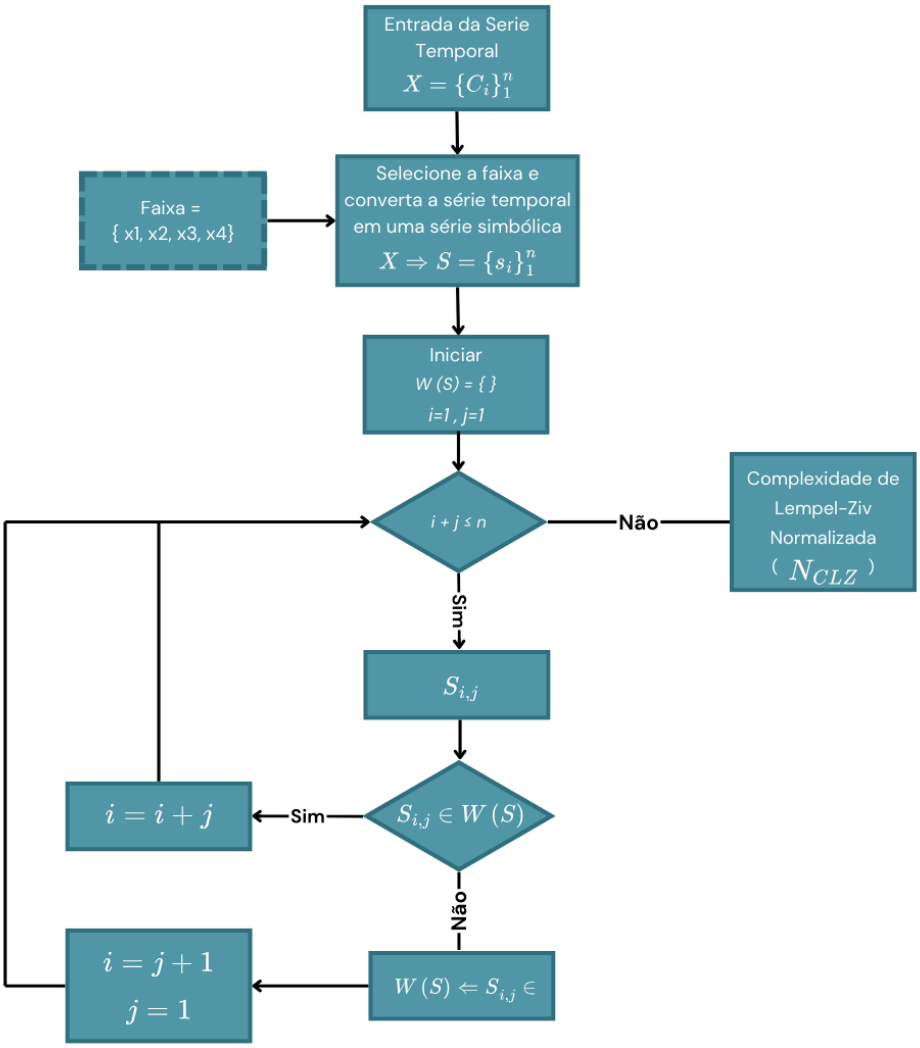
Para a geração das características da LZC, a série temporal de cada usuário é segmentada em faixas, onde a quantidade de intervalos de tamanhos iguais é definida com base em experimentação. Esse processo é repetido para quatro faixas distintas (x_1, x_2, x_3 e x_4), cada uma

com características próprias, permitindo uma análise mais refinada das variações na série temporal.

Essa segmentação tem como objetivo extrair características de cada intervalo, proporcionando informações adicionais sobre o comportamento da série. Cada uma dessas quatro características, conforme a explicação do método, assume um valor diferente dependendo da leitura de consumo na série temporal e do número de intervalos alocados, o que leva a um valor que indica a LZC de cada série temporal. Além disso, o valor das características selecionadas varia com base em como foram selecionadas (GHAMKHAR et al., 2023).

Para ilustrar melhor essa etapa de obtenção das características do LZC, a Figura 22 apresenta o fluxograma representativo do algoritmo utilizado nesta subseção.

Figura 22: Fluxograma representativo da obtenção das características do LZC.



Fonte: Adaptado de GHAMKHAR et al., (2023)

4.5 DBSCAN

Conforme abordado no tópico 3.2.4, a implementação do algoritmo DBSCAN requer a definição de dois parâmetros: $MinPts$ e ε. Para estimar o valor de $MinPts$, de acordo com a recomendação de Sander et al. (1998), multiplica-se o número de dimensões do conjunto de dados por dois, conforme demonstrado na Equação 10:

$$MinPts = 2dim \tag{10}$$

Em que dim corresponde ao número de dimensões do conjunto de dados. De acordo com Ghamkhar et al., (2023), o aumento desse parâmetro melhora o desempenho do algoritmo em conjuntos de dados de alta dimensionalidade, com dados redundantes e/ou elevado volume de ruídos.

Para estimar o valor do segundo parâmetro, utilizou-se um método baseado na distância do vizinho mais próximo K (KNN). Esse método consiste, primeiramente, em calcular a distância do K-ésimo vizinho mais próximo para cada ponto do conjunto de dados, sendo o valor de K determinado conforme a Equação 11.

$$K = MinPts - 1 \tag{11}$$

Em seguida, as distâncias são organizadas em ordem crescente, e um gráfico é gerado em que o eixo Y representa as distâncias para o K-ésimo vizinho mais próximo, enquanto o eixo X indica a posição dos pontos na lista ordenada. No gráfico resultante, a curva geralmente exibe uma transição clara entre uma região inicial mais plana e uma região com uma inclinação mais acentuada, formando um ponto de inflexão, conhecido como "ponto de cotovelo". Esse ponto de inflexão representa o valor de ε mais apropriado para o algoritmo DBSCAN, pois marca a fronteira entre pontos próximos e pontos mais distantes, considerados ruído. Assim, o valor de ε definido a partir do ponto de cotovelo permite capturar a estrutura dos conjuntos, ao mesmo tempo em que elimina pontos isolados, aumentando a precisão do modelo na detecção de irregularidades (SANDER et al., 1998; GHAMKHAR et al., 2023).

Portanto, após a definição dos parâmetros, conforme Ester (1996), para um conjunto de dados $D = \{p_i | p_i \in R, 1 \leq i \leq n\}$, onde n é o numero de pontos/instancias no conjunto de dados, para um ponto p qualquer, calcula-se o conjunto $N_\varepsilon(p)$, que representa todos os pontos no conjunto de dados D que estão a uma distância menor ou igual a ε de p, ou seja, $N_\varepsilon(p) = \{q \in D | dist(p,q) \leq \varepsilon, \mathrm{p} \neq \mathrm{q}\}$. Esse conjunto de pontos vizinhos a p é essencial para determina a densidade ao redor de p. Se $N_\varepsilon(p)$ contiver pelo menos $MinPts$, ou seja, $|N_\varepsilon(p)| \geq MinPts$,

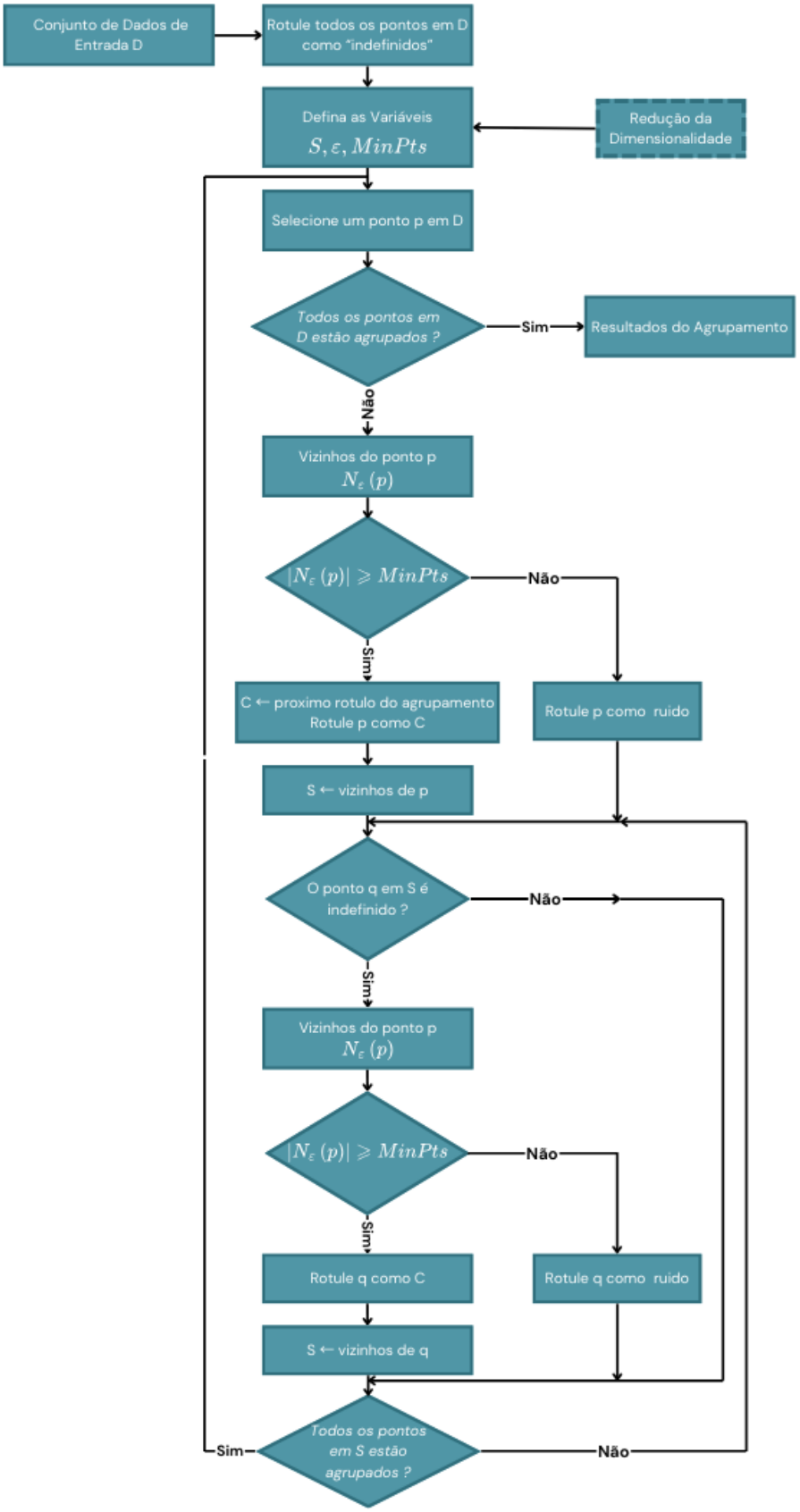
então p é classificado como ponto central e pode servir como um nucleo para a expansao de um conjunto.

Por outro lado, um ponto que não é central, mas está dentro da vizinhança ε de um ponto central, é classificado como ponto de fronteira, enquanto pontos que não atendem a nenhum desses critérios são classificados como ruídos ($ruido = \{p \in D | \forall i: p \notin D_i$).

O processo de formação do conjunto começa a partir de um ponto central p. A partir de p, o algoritmo verifica todos os pontos em $N_\varepsilon(p)$ e os adiciona ao conjunto se forem pontos centrais ou de fronteira. Esse processo é repetido para cada ponto vizinho até que todos os pontos densamente conectados sejam incluídos no conjunto, formando assim um agrupamento coeso.

Para ilustrar esse processo, a Figura 23 apresenta cada etapa detalhada, desde a seleção inicial dos pontos, passando pela formação dos conjuntos, até a identificação dos pontos classificados como ruído. Essa visualização auxilia na compreensão do funcionamento do algoritmo, evidenciando o agrupamento de pontos densamente conectados e a separação dos elementos fora desses agrupamentos.

Figura 23: Fluxograma representativo do código do DBSCAN.



Fonte: Adaptado de GHAMKHAR et al., (2023)

4.6 Critérios de qualidade

Uma vez que os parâmetros do DBSCAN foram estabelecidos e o algoritmo implementado, dois sinais de alerta básicos são utilizados para avaliar a qualidade de resultados obtidos. Esses sinais incluem o número de pontos de ruído e o tamanho do maior conjunto.

A taxa aceitável de dados classificados como ruído geralmente varia de 1% a 30%, enquanto o tamanho do maior conjunto desempenha um papel crucial na determinação se um conjunto de dados degenerou, dependendo do tipo de dados e de sua dimensionalidade (SHUBERT et al, 2017).

4.7 Ferramentas utilizadas

Para o desenvolvimento dessas atividades, utilizou-se a linguagem de programação Python 3.7, amplamente utilizado por sua sintaxe simples e pelo oferecimento uma plataforma robusta de bibliotecas especializadas, facilitando o processamento, a manipulação e a análise de grandes volumes de informações (CARVALHO, 2024).

Uma das principais vantagens da utilização da linguagem Python está na possibilidade de escrever códigos de forma compacta e intuitiva, o que favorece um desenvolvimento mais ágil e com menor propensão a erros. Além disso, a linguagem permite integrar todas as etapas do processo analítico, desde a leitura e o tratamento dos dados até a aplicação de modelos, por meio de blocos de código claros e objetivos.

Em termos simples, bibliotecas em Python são coleções de módulos e funções que oferecem funcionalidades específicas. Entre as bibliotecas usadas neste estudo, destacam-se:

- **Pandas**: utilizada para a manipulação e análise de dados, possibilitando a leitura de arquivos, organização de dados em estruturas tabulares e realização de operações estatísticas e de filtragem.
- **NumPy**: essencial para operações numéricas e manipulação de arrays multidimensionais, servindo como base para diversas outras bibliotecas.
- **Matplotlib**: utilizadas para a criação de gráficos e visualizações que auxiliam na compreensão dos padrões e tendências presentes nos dados.
- **Scikit-learn**: biblioteca fundamental para a aplicação de algoritmos de aprendizado de máquina, incluindo funções para treinamento, avaliação de modelos e pré-processamento de dados.

- **Kneed**: empregada na identificação automática do "ponto de cotovelo" em curvas, auxiliando na escolha de parâmetros para algoritmos baseados em agrupamento.

Caso haja interesse, o arquivo com o código-fonte desenvolvido para este trabalho pode ser acessado através do link: Analise e detecção de consumos não autorizados. Embora os dados utilizados não possam ser disponibilizados por questões de confidencialidade, o código permite que os leitores analisem a coerência das etapas de processamento e modelagem aplicadas, bem como compreendam a lógica adotada na construção das análises.

5 RESULTADOS

Neste capítulo, são apresentados os resultados obtidos por meio da aplicação do procedimento metodológico descrito no capítulo anterior, bem como as discussões relacionadas às análises realizadas. Inicialmente, realiza-se o pré-processamento, abordando a caracterização estatística das informações coletadas e, em seguida, preparando os dados para a aplicação do algoritmo. Posteriormente, determinam-se os valores dos parâmetros empregados na extração de características e na configuração do algoritmo. Por fim, são apresentadas as métricas de desempenho do modelo e a interpretação dos resultados obtidos.

5.1 Pré-processamento

5.1.1 Exploração e tratamento inicial dos dados

Após a obtenção dos dados por meio de planilhas no formato .xlsx fornecidas pela ENOPS Engenharia S.A., realizou-se a análise exploratória, etapa fundamental para compreender as características do conjunto de dados utilizado no estudo. Inicialmente, essa análise foi conduzida para cada planilha, cada uma contendo dados anuais da companhia. Procedimentos idênticos foram aplicados a todas as planilhas, a fim de garantir a uniformidade na extração de informações.

Os dados fornecidos pela empresa não correspondem ao conjunto de dados brutos, mas sim a um conjunto resumido, no qual não foi possível distinguir algumas categorias nem identificar com precisão o status atual das matrículas durante o período de coleta, devido à ausência de determinadas informações resultante da condição de anonimato mencionada anteriormente.

Posto isso, os dados obtidos são estruturados, seguindo uma organização rígida e bem definida, composta por tabelas formadas por linhas e colunas que obedecem a regras específicas e identificam suas características sobre determinado assunto. As colunas representam os diferentes atributos ou variáveis do conjunto de dados, sendo formadas de acordo com a Tabela 3, que apresenta a estrutura de colunas de um único mês. Cada coluna possui um tipo de dado bem definido, como valores numéricos, categóricos ou alfanuméricos, refletindo a natureza das informações registradas. Um conjunto idêntico de colunas foi repetido para cada um dos 12 meses do ano, resultando em uma estrutura replicada mensalmente. Dessa forma, cada grupo

de colunas caracteriza os dados de um mês específico, como identificadores de matrícula, setor comercial, volume consumido, datas de leitura e valores de consumo registrados.

Contudo, como resultado final, obteve-se um total de cinco planilhas nesse formato, cada uma variando na quantidade de linhas. Cada linha representa os dados de uma única matrícula durante todo o período anual, consolidando as informações de um consumidor específico para os 12 meses analisados. A quantidade de linhas difere entre as planilhas devido ao número total de matrículas registradas em cada conjunto de dados, como detalhado na Tabela 5, que apresenta o ano correspondente de cada planilha, além da quantidade de linhas e colunas de cada uma.

Tabela 5: Informações gerais sobre os dados obtidos.

Ano	Quantidade de Linhas	Quantidade de Colunas
2018	100.445	85
2019	102.893	85
2020	103.836	85
2021	105.190	85
2022	108.079	85

Fonte: Autor (2025)

Nesse viés, inicialmente verificou-se a procedência dos diferentes atributos e variáveis do conjunto de todas as planilhas, conforme exemplificado na Figura 24, buscando compreender detalhadamente suas características e origem, a fim de possibilitar uma melhor interpretação e tratamento ao longo do estudo.

Figura 24: Tipologia dos dados das colunas

```
#Verificando a tipologia de cada coluna do banco
banco2018.dtypes.head(10)
```

```
INSCRICAO                              int64
COMPETENCIA_JAN_2018                 float64
SETOR_ABASTECIMENTO_JAN_2018          object
SETOR_COMERCIAL_JAN_2018             float64
COD_HIDROMETRO_JAN_2018               object
DATA_INSTALACAO_HIDROMETRO_JAN_2018   object
DATA_LEITURA_JAN_2018                 object
CONSUMO_MEDIDO_JAN_2018              float64
COMPETENCIA_FEV_2018                 float64
SETOR_ABASTECIMENTO_FEV_2018          object
dtype: object
```

Fonte: Autor (2025)

Durante esse processo, foram constatadas inconsistências na tipologia de colunas presentes em diferentes planilhas, como os referentes ao setor de abastecimento, a data de

leitura e a data de instalação do hidrômetro, que foram inicialmente classificadas pela biblioteca Pandas como 'object', um tipo genérico utilizado para armazenar textos ou diferentes tipos de dados misturados. Essa classificação pode comprometer análises quantitativas ou operações baseadas em datas, motivo pelo qual foi necessário realizar o tratamento adequado dos dados. Assim, os campos mencionados foram convertidos, respectivamente, para os tipos numérico e data, a fim de garantir a consistência do conjunto de dados e viabilizar análises futuras.

Em sequência, devido ao grande volume de dados, realizou-se uma análise para verificar a existência de matrículas duplicadas, o que poderia gerar informações conflitantes, resultando em análises enviesadas e comprometendo o desempenho do algoritmo. Entretanto, observou-se que nenhuma planilha continha matrículas duplicadas, demonstrando uma certa organização na base de dados fornecida.

Posteriormente, realizou-se a filtragem das colunas que continham os dados de consumo mensal dos usuários, com o objetivo de isolar as informações relevantes para a análise, eliminando aquelas que não apresentavam relação direta com os objetivos do estudo. Na sequência, procedeu-se à renomeação das colunas selecionadas, com a finalidade de padronizar a nomenclatura e, assim, facilitar tanto a manipulação quanto a interpretação dos dados ao longo do processo analítico.

A partir das etapas anteriores, foi possível realizar a verificação da qualidade dos dados, por meio da análise voltada à identificação de possíveis inconsistências nos registros de consumo mensal. Para isso, foram calculadas, para cada mês de cada ano analisado, as porcentagens de valores iguais a zero e de valores ausentes, ambos considerados potenciais indicadores de dados inválidos ou não registrados de forma adequada.

A Tabela 6 apresenta, de forma consolidada, os resultados obtidos para os anos analisados além de uma nova coluna contendo a média mensal das porcentagens de inconsistências. Cada seção da imagem refere-se a um ano distinto, permitindo observar a evolução e a distribuição das inconsistências ao longo do tempo.

Tabela 6: Evolução dos percentuais de valores nulos e ausentes

	01.2018	02.2018	03.2018	04.2018	05.2018	06.2018	07.2018	08.2018	09.2018	10.2018	11.2018	12.2018	Média
Porcentagem 0	14.8%	15.2%	15.8%	16.2%	16.2%	16.4%	15.6%	14.6%	14.7%	14.7%	14.7%	15.0%	15.3%
Porcentagem NaN	3.7%	3.4%	2.9%	2.4%	1.9%	1.7%	1.7%	1.7%	1.8%	1.7%	1.6%	1.5%	2.2%
Total	18.4%	18.6%	18.8%	18.6%	18.1%	18.0%	17.3%	16.4%	16.5%	16.4%	16.4%	16.5%	17.5%

	01.2019	02.2019	03.2019	04.2019	05.2019	06.2019	07.2019	08.2019	09.2019	10.2019	11.2019	12.2019	Média
Porcentagem 0	13.4%	13.8%	14.1%	14.8%	14.5%	15.5%	15.4%	15.9%	17.2%	16.9%	15.4%	15.6%	15.2%
Porcentagem NaN	3.7%	3.7%	3.6%	3.5%	3.4%	3.3%	3.1%	3.1%	3.1%	3.0%	3.0%	1.3%	3.1%
Total	17.2%	17.4%	17.7%	18.3%	17.9%	18.8%	18.5%	18.9%	20.2%	19.9%	18.3%	16.9%	18.3%
	01.2020	**02.2020**	**03.2020**	**04.2020**	**05.2020**	**06.2020**	**07.2020**	**08.2020**	**09.2020**	**10.2020**	**11.2020**	**12.2020**	**Média**
Porcentagem 0	15.2%	15.4%	15.9%	16.2%	17.3%	17.2%	16.2%	15.6%	15.1%	14.9%	14.8%	14.7%	15.7%
Porcentagem NaN	2.0%	1.8%	1.7%	1.5%	1.4%	1.3%	1.2%	1.1%	1.0%	1.0%	1.0%	1.0%	1.3%
Total	17.2%	17.2%	17.6%	17.6%	18.6%	18.5%	17.4%	16.7%	16.1%	15.9%	15.7%	15.7%	17.0%
	01.2021	**02.2021**	**03.2021**	**04.2021**	**05.2021**	**06.2021**	**07.2021**	**08.2021**	**09.2021**	**10.2021**	**11.2021**	**12.2021**	**Média**
Porcentagem 0	14.7%	14.7%	14.8%	15.0%	14.9%	14.5%	14.4%	14.2%	14.2%	14.2%	14.2%	14.2%	14.5%
Porcentagem NaN	2.1%	2.0%	2.2%	2.3%	2.2%	2.0%	2.4%	2.4%	2.3%	2.4%	2.4%	2.5%	2.3%
Total	16.8%	16.7%	17.1%	17.3%	17.2%	16.5%	16.8%	16.6%	16.5%	16.6%	16.6%	16.7%	16.8%
	01.2022	**02.2022**	**03.2022**	**04.2022**	**05.2022**	**06.2022**	**07.2022**	**08.2022**	**09.2022**	**10.2022**	**11.2022**	**12.2022**	**Média**
Porcentagem 0	13.9%	14.1%	14.1%	14.5%	14.6%	14.5%	14.6%	14.4%	13.8%	13.9%	14.7%	15.0%	14.3%
Porcentagem NaN	5.2%	5.1%	5.1%	5.1%	5.1%	5.1%	5.1%	5.0%	4.8%	4.7%	1.0%	0.8%	4.3%
Total	19.1%	19.3%	19.2%	19.6%	19.7%	19.6%	19.6%	19.4%	18.6%	18.6%	15.6%	15.8%	18.7%

Fonte: Autor (2025)

Observa-se que, em todos os anos, há variação nas proporções de valores nulos e zerados entre os meses. Em média, os valores ausentes ficaram em torno de 2,64% ao mês, enquanto os valores iguais a zero oscilaram em torno de 14,98%. Portanto, a média dos valores iguais a zero é relativamente elevada, especialmente e considerarmos que se trata de registros mensais de consumo de água. Por outro lado, embora a média de valores ausentes seja mais baixa, merece atenção, sobretudo no ano de 2022, quando esse índice atingiu 4,0%, praticamente o dobro da média observada nos anos anteriores. No total, o percentual médio de dados com algum tipo de irregularidade alcança 17,66%, o que indica que quase um quinto dos registros mensais de consumo apresenta inconsistências que podem comprometer a qualidade das análises subsequentes.

No âmbito das companhias de saneamento, é comum o entendimento de que ligações que apresentam consumo igual a zero de forma constante, ou com elevada frequência ao longo de um determinado período, constituem potenciais matrículas com indícios de irregularidades operacionais ou fraudes. Diferentemente de abordagens tradicionais que descartam dados faltantes ou nulos, neste trabalho tais registros são tratados como elementos relevantes para a análise. Isso porque a ocorrência de valores ausentes ou iguais a zero pode estar associada a comportamentos irregulares no consumo de água, como possíveis fraudes, leituras incorretas, interrupções no fornecimento ou inatividade dos hidrômetros. Portanto, mapear essas proporções ao longo do tempo permite identificar padrões suspeitos, contribuindo para a melhoria da eficiência operacional do sistema de abastecimento.

Na etapa seguinte, foi gerado um resumo estatístico mensal dos dados, contemplando métricas como contagem de registros, média, desvio padrão, valores mínimos, máximos e quartis, que indicam a dispersão e a tendência central do consumo. A Tabela 7 apresenta essas estatísticas aplicadas ao consumo mensal de cada conjunto de dados analisado.

Tabela 7: Descrição estatística de cada mês do banco de dados

	01.2018	02.2018	03.2018	04.2018	05.2018	06.2018	07.2018	08.2018	09.2018	10.2018	11.2018	12.2018
count	96758.00	96986.00	97488.00	98000.00	98526.00	98758.00	98739.00	98693.00	98644.00	98726.00	98826.00	98891.00
mean	12.04	11.34	11.01	11.48	11.86	11.46	12.20	9.54	10.23	10.84	11.05	9.75
std	266.50	280.88	294.99	307.64	358.28	333.38	364.36	53.66	98.03	294.95	202.17	55.17
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	3.00	3.00	2.00	2.00	2.00	2.00	2.00	3.00	3.00	3.00	3.00	3.00
50%	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00
75%	12.00	11.00	11.00	11.00	11.00	11.00	11.00	11.00	11.00	11.00	12.00	11.00
max	54324.00	57509.00	60318.00	63272.00	66139.00	68710.00	71290.00	6287.00	21868.00	88490.00	60933.00	6291.00
	01.2019	**02.2019**	**03.2019**	**04.2019**	**05.2019**	**06.2019**	**07.2019**	**08.2019**	**09.2019**	**10.2019**	**11.2019**	**12.2019**
count	99044.00	99127.00	99206.0	99296.00	99414.00	99493.00	99720.00	99751.00	99732.00	99829.00	99850.00	101561.00
mean	13.26	11.20	9.3	9.60	9.89	9.07	9.04	8.85	7.87	7.94	10.53	9.74
std	80.55	67.71	48.2	52.85	57.04	53.82	54.99	58.98	32.19	34.09	61.01	55.86
min	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	3.00	3.00	3.0	3.00	3.00	2.00	2.00	2.00	2.00	2.00	3.00	2.00
50%	7.00	7.00	7.0	7.00	7.00	7.00	7.00	7.00	6.00	6.00	7.00	7.00
75%	12.00	12.00	11.0	11.00	12.00	11.00	11.00	11.00	10.00	10.00	12.00	11.00
max	3920.00	4738.00	5318.0	5128.00	6653.00	6984.00	8053.00	6988.00	2601.00	3030.00	7429.00	5883.00
	01.2020	**02.2020**	**03.2020**	**04.2020**	**05.2020**	**06.2020**	**07.2020**	**08.2020**	**09.2020**	**10.2020**	**11.2020**	**12.2020**
count	101750.00	101970.00	102096.00	102324.00	102430.00	102477.0	102564.00	102677.00	102758.00	102802.00	102831.00	102821.00
mean	10.25	9.93	9.36	10.38	9.88	10.0	19.47	10.13	10.74	10.50	10.90	10.83
std	55.05	54.37	54.32	59.20	55.30	57.4	1600.24	60.97	61.70	64.48	63.93	67.00
min	0.00	0.00	0.00	0.00	0.00	0.0	0.00	0.00	0.00	0.00	0.00	0.00
25%	3.00	2.00	2.00	2.00	2.00	2.0	2.00	3.00	3.00	3.00	3.00	3.00
50%	7.00	7.00	7.00	7.00	7.00	7.0	7.00	7.00	8.00	8.00	8.00	8.00
75%	12.00	11.00	11.00	12.00	12.00	12.0	12.00	12.00	12.00	12.00	12.00	12.00
max	6574.00	6540.00	6980.00	7570.00	6558.00	6986.0	390110.00	7919.00	7493.00	9969.00	7108.00	9285.00
	01.2021	**02.2021**	**03.2021**	**04.2021**	**05.2021**	**06.2021**	**07.2021**	**08.2021**	**09.2021**	**10.2021**	**11.2021**	**12.2021**
count	103012.00	103057.00	102834.00	102730.00	102859.00	103135.00	102693.00	102716.00	102764.00	102705.00	102677.00	102597.00
mean	10.78	10.20	10.19	10.34	9.96	10.22	9.82	9.98	10.18	10.13	10.14	10.37
std	59.22	120.59	79.72	62.97	61.06	55.87	55.66	60.47	56.26	59.69	59.13	58.08
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
50%	8.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00
75%	12.00	11.00	12.00	12.00	11.00	12.00	11.00	11.00	12.00	12.00	12.00	12.00
max	6783.00	33330.00	14985.00	8977.00	8366.00	6561.00	6666.00	7638.00	6877.00	6526.00	6495.00	6743.00

	01.2022	02.2022	03.2022	04.2022	05.2022	06.2022	07.2022	08.2022	09.2022	10.2022	11.2022	12.2022
count	102504.00	102520.00	102549.00	102552.00	102585.00	102612.00	102618.00	102720.00	102841.00	102990.00	107050.00	107215.00
mean	10.26	9.62	9.97	9.16	9.62	9.73	9.66	9.74	10.55	10.30	10.60	10.39
std	57.81	54.84	61.95	53.91	57.04	62.33	59.70	62.43	62.41	62.18	83.71	59.26
min	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	3.00	3.00	3.00	2.00	2.00	2.00	2.00	2.00	3.00	3.00	3.00	3.00
50%	7.00	7.00	7.00	6.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00	7.00
75%	12.00	11.00	11.00	10.00	11.00	11.00	11.00	11.00	12.00	12.00	12.00	12.00
max	6791.00	6500.00	8694.00	5470.00	6037.00	7736.00	6114.00	6045.00	6541.00	6781.00	18746.00	6499.00

Fonte: Autor (2025)

A análise estatística descritiva dos dados de consumo de água, compreendendo o período de janeiro de 2018 a dezembro de 2022, revelou um padrão relativamente estável ao longo do tempo, especialmente no que se refere à média mensal de consumo. De modo geral, os valores médios situaram-se entre 9 e 12 m³, com variações sazonais discretas. Observa-se um leve aumento nos meses de janeiro e julho, possivelmente associados ao período de férias escolares e variações climáticas, que influenciam diretamente o comportamento dos consumidores.

Os valores dos quartis (Q1, Q2 e Q3) reforçam essa tendência de estabilidade: o primeiro quartil variou entre 2 e 3 m³, a mediana manteve-se em 7 m³ e o terceiro quartil situou-se entre 10 e 12 m³. Isso demonstra que a maioria dos consumidores apresenta padrões regulares de consumo, com pouca variabilidade entre os meses analisados.

Durante os anos de 2020 e 2021, período correspondente à pandemia de COVID-19, era esperado que as medidas de isolamento social e a maior permanência das pessoas em casa resultassem em um aumento significativo no consumo de água. No entanto, a análise detalhada dos dados sugere que esse aumento não foi consistente. O pico observado na média de consumo em julho de 2020 (superior a 19 m³) está fortemente associado a um valor máximo atípico registrado neste mês, que ultrapassou 390 mil m³. Esse valor destoa drasticamente da série histórica e dos demais registros, indicando a provável ocorrência de um erro de leitura, digitação ou medição.

Ao se analisar os valores máximos de consumo, identificam-se registros significativamente superior à média nos primeiros anos da série histórica. Esses picos, que são drasticamente reduzidos nos anos posteriores, podem estar associados à operação de uma unidade industrial anteriormente conectada ao sistema de abastecimento. A posterior desconexão dessa unidade consumidora explicaria a ausência de valores tão elevados nos anos seguintes, sugerindo uma possível mudança estrutural no perfil de consumo da região analisada.

Do ponto de vista estatístico, esses registros podem ser considerados ruídos, uma vez que destoam fortemente do comportamento predominante entre os demais consumidores. A teoria dos dados com presença industrial justifica essa interpretação, pois consumidores industriais tendem a apresentar padrões de consumo mais elevados e menos previsíveis, o que afeta diretamente os parâmetros descritivos da série, como a média e, principalmente, o desvio padrão. Tal influência é evidenciada nos meses em que o desvio padrão atinge valores significativamente elevados, refletindo a maior dispersão dos dados em função desses consumidores atípicos.

Após a análise individual das planilhas, realizou-se à integração dos dados de consumo provenientes dos cinco arquivos, utilizando o número de matrícula como chave primária para a unificação das informações. No entanto, foram excluídos os usuários que apresentavam registros de consumo em menos de dois quintos do período analisado, equivalente a 24 meses, por serem considerados novos usuários e, provavelmente, não apresentarem medidores defeituosos ou adulterados, uma vez que foram conectados recentemente à rede. Após essa filtragem, totalizaram-se 106.497 matrículas, conforme ilustrado na Tabela 8.

Tabela 8: Dados de consumo integrados em uma base de dados.

matricula	01.2018	02.2018	03.2018	04.2018	05.2018	06.2018	07.2018	08.2018	09.2018	10.2018	11.2018	12.2018	01.2019	02.2019	03.2019	04.2019
2020	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0	2.0	0.0	2.0	1.0	11.0	8.0	1.0	10.0
23124	4.0	6.0	5.0	7.0	7.0	6.0	5.0	4.0	6.0	7.0	6.0	7.0	0.0	4.0	8.0	23.0
24511	3.0	2.0	3.0	2.0	2.0	1.0	2.0	2.0	2.0	3.0	13.0	8.0	1.0	1.0	0.0	10.0
28355	0.0	1.0	0.0	1.0	0.0	2.0	5.0	4.0	5.0	3.0	5.0	4.0	11.0	18.0	9.0	10.0
32506	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	14.0	13.0	13.0	17.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
76037460	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
76037495	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
76037517	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
90811208	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	9.0	2.0	3.0
97233587	9.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	15.0	10.0	8.0	12.0

106497 rows × 60 columns

Fonte: Autor (2025)

Com as planilhas unificadas, realizou-se uma nova verificação da qualidade dos dados, a fim de identificar eventuais inconsistências introduzidas durante o processo de unificação. A Tabela 9 ilustra uma amostra representativa desses valores atualizados.

Tabela 9: Proporção de valores ausentes e zeros na base de dados após a unificação

	01.2018	02.2018	03.2018	04.2018	05.2018	06.2018	07.2018	08.2018	09.2018	10.2018	11.2018	12.2018	01.2019	02.2019	03.2019
Porcentagem 0	13.9%	14.3%	14.9%	15.3%	15.3%	15.4%	14.7%	13.8%	13.8%	13.8%	13.9%	14.1%	13.0%	13.3%	13.7%
Porcentagem NaN	9.1%	8.9%	8.5%	8.0%	7.5%	7.3%	7.3%	7.3%	7.4%	7.3%	7.2%	7.1%	7.0%	6.9%	6.8%
Total	23.1%	23.2%	23.4%	23.2%	22.7%	22.7%	22.0%	21.1%	21.2%	21.1%	21.1%	21.3%	20.0%	20.2%	20.5%

Fonte: Autor (2025)

Nessa etapa, observou-se um aumento na quantidade de valores ausentes e uma constância no número de registros com valor igual a zero. No entanto, a porcentagem desse último apresentou uma leve diminuição em relação ao total, em virtude do aumento do tamanho da amostra após a integração das planilhas. Em média, os registros com valores iguais a zero representaram aproximadamente 14,5% dos registros mensais, o que corresponde a uma redução de 0,48% em relação à base original. Por outro lado, os valores ausentes apresentaram uma média de 5,8%, indicando um aumento de 3,16% em comparação com a base anterior.

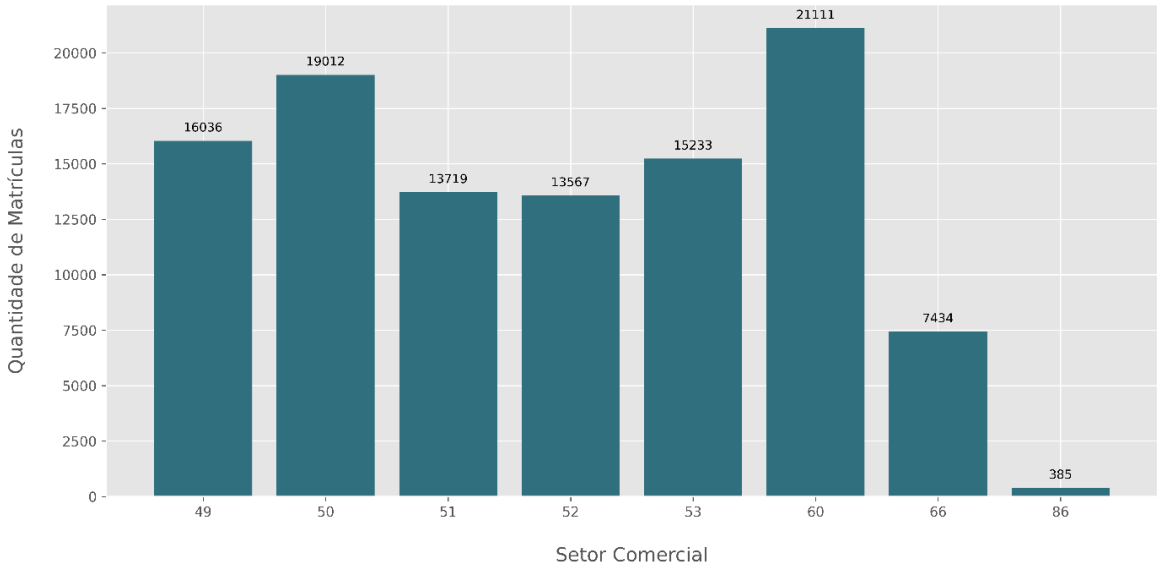
Tal comportamento era esperado, uma vez que a base integrada passou a contemplar matrículas cadastradas ao longo do período analisado e que, portanto, não possuíam histórico de consumo anterior à sua entrada no sistema. Como resultado, os meses anteriores ao cadastramento dessas unidades apresentaram lacunas ou registros com valor zero. Adicionalmente, identificou-se que algumas matrículas foram descontinuadas no banco de dados ao longo do tempo, possivelmente em razão de desligamentos do sistema, alterações cadastrais ou substituições de usuários. Esses fatores também contribuíram para o aumento da quantidade de dados faltantes.

Ainda com base na base de dados integrada, foi realizada uma análise da quantidade de matrículas presentes em cada setor definido pela setorização comercial. Através dessa análise, é possível obter uma visão geral sobre como as matrículas estão distribuídas pelos setores comerciais, o que pode ser útil para a tomada de decisões, como a alocação de investimentos em infraestrutura, campanhas de conscientização, ou otimização dos serviços prestados. Além disso, essa análise contribui para uma compreensão mais aprofundada do comportamento de consumo por setor, dispensando a necessidade de dados geoespaciais, ao utilizar a classificação numérica existente para caracterizar cada setor.

Para garantir a consistência da análise, utilizou-se a última informação de setorização conhecida para cada matrícula. Essa abordagem foi adotada com o intuito de evitar distorções causadas por eventuais mudanças nos setores ao longo do tempo, assegurando que a análise

refletisse a configuração atual, sem a interferência de alterações anteriores. Assim, cada matrícula foi atribuída ao setor mais recente, garantindo uma análise precisa e alinhada com o estado atual da setorização. A Figura 25 ilustra graficamente os resultados dessa análise, apresentando a distribuição das matrículas por setor comercial com base na setorização mais recente disponível.

Figura 25: Matrículas por setor comercial



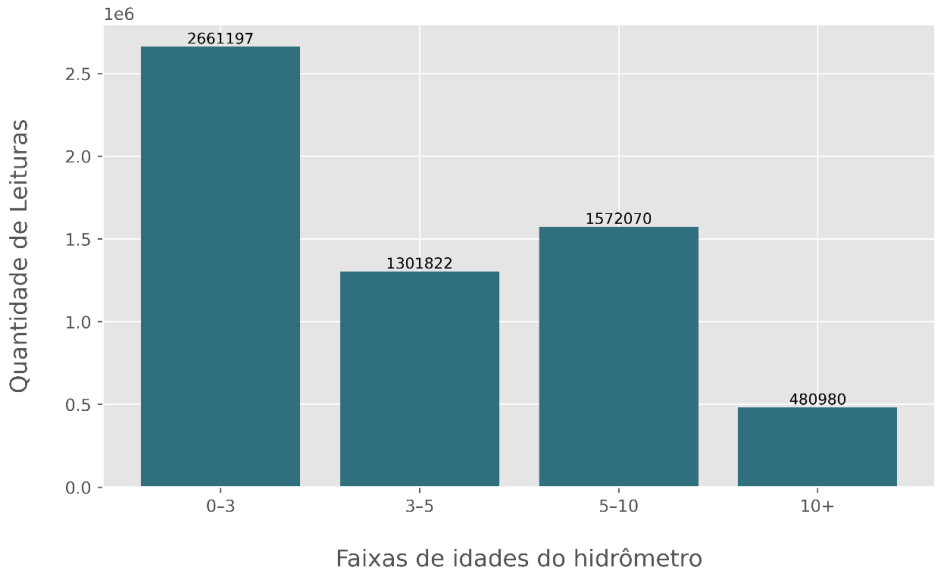
Fonte: Autor (2025)

Observa-se que a distribuição das matrículas entre os setores comerciais é bastante heterogênea. O setor 60 concentra o maior número de matrículas, totalizando 21.111 registros, seguido pelos setores 50 e 49, com 19.012 e 16.036 matrículas, respectivamente. Em contrapartida, os setores 66 e 86 apresentam os menores quantitativos, com 7.434 e apenas 385 matrículas.

Por fim, foi realizada uma análise da idade dos hidrômetros em relação ao momento de cada medição de consumo. A análise da idade média dos hidrômetros contribui para identificar possíveis necessidades de substituição ou manutenção preventiva, além de fornecer subsídios para entender eventuais desvios nos padrões de consumo que possam estar relacionados à obsolescência dos dispositivos de medição. Ressalta-se que todas as matrículas consideradas na análise possuíam hidrômetros disponíveis.

Para cada registro de medição, a idade do hidrômetro foi calculada com base na diferença entre a data da medição e a data de instalação do equipamento. Em seguida, os valores obtidos foram agrupados em faixas, com o objetivo de facilitar a visualização da distribuição geral da idade dos hidrômetros em relação a cada aferição, conforme ilustrado na Figura 26.

Figura 26: Faixas de Idades dos Hidrômetros com base em cada aferição



Fonte: Autor (2025)

Observa-se que a maioria das medições foi realizada em hidrômetros com até 3 anos de uso, representando 44,23% do total. Em seguida, destaca-se a faixa etária entre 3 e 5 anos, com 21,64% das medições. No entanto, chama atenção o fato de que 26,13% das medições ocorreram em hidrômetros com idade entre 5 e 10 anos, e 8% em equipamentos com mais de 10 anos de uso. Esses percentuais são particularmente relevantes diante da recomendação do Instituto Nacional de Metrologia, Qualidade e Tecnologia (INMETRO), Portaria 246/00, que orienta a substituição dos hidrômetros a cada 5 anos. Diante disso, constata-se que 34,13% das medições foram realizadas em equipamentos fora do prazo recomendado, o que pode indicar perdas médias de leitura superiores a 36,6%, conforme apontado por Fournioul (2004).

5.1.2 Preparação dos dados para modelagem

Após a realização da análise descritiva e do tratamento dos dados, iniciou-se o processo de preparação dos dados para a modelagem. Nesse contexto, o banco de dados foi segmentado em dois conjuntos, com o intuito de otimizar o desempenho do modelo e mitigar a formação de agrupamentos inconsistentes, os quais poderiam surgir devido à grande variação no histórico de consumo entre os usuários.

Para essa segmentação, consideraram-se todos os valores de consumo diferentes de *NaN* como registros válidos, sendo contabilizadas para determinar a quantidade total de leituras ao longo do período analisado. Dessa forma, o primeiro conjunto é composto por usuários que apresentaram dados disponíveis em, no mínimo, 80% dos 60 meses analisados, ou seja, entre

48 e 60 meses. Já o segundo conjunto abrange os usuários com dados disponíveis em menos de 80% desse período, correspondendo a até 47 meses.

Nessas condições, o primeiro conjunto apresentou 94438 matrículas, correspondendo a 88,7% do banco de dados, enquanto o segundo conjunto é composto por 12.059 matrículas, representando os 11,3% restantes. A visualização desses dois conjuntos está apresentada na Tabela 10, a qual exibe, respectivamente, os dados do Conjunto 1 (a) e do Conjunto 2 (b).

Tabela 10: Matrizes de consumo mensais para os dois conjuntos analisados: a) Conjunto 1, b) Conjunto 2

	01.2018	02.2018	03.2018	04.2018	05.2018	06.2018	07.2018	08.2018	09.2018	10.2018	11.2018	12.2018	01.2019	02.2019	03.2019	04.2019
matricula																
2020	1.0	1.0	1.0	1.0	1.0	1.0	1.0	2.0	2.0	0.0	2.0	1.0	11.0	8.0	1.0	10.0
23124	4.0	6.0	5.0	7.0	7.0	6.0	5.0	4.0	6.0	7.0	6.0	7.0	0.0	4.0	8.0	23.0
24511	3.0	2.0	3.0	2.0	2.0	1.0	2.0	2.0	2.0	3.0	13.0	8.0	1.0	1.0	0.0	10.0
28355	0.0	1.0	0.0	1.0	0.0	2.0	5.0	4.0	5.0	3.0	5.0	4.0	11.0	18.0	9.0	10.0
32506	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	14.0	13.0	13.0	17.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
63858495	17.0	20.0	19.0	21.0	9.0	0.0	14.0	13.0	18.0	15.0	26.0	20.0	1.0	15.0	0.0	9.0
63859017	5.0	14.0	12.0	12.0	12.0	11.0	11.0	11.0	12.0	11.0	10.0	8.0	5.0	4.0	4.0	10.0
66511500	6.0	6.0	5.0	13.0	11.0	13.0	6.0	9.0	10.0	9.0	15.0	9.0	3.0	0.0	8.0	0.0
90811208	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	4.0	9.0	2.0	3.0
97233587	9.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	15.0	10.0	8.0	12.0

94438 rows × 61 columns

	01.2018	02.2018	03.2018	04.2018	05.2018	06.2018	07.2018	08.2018	09.2018	10.2018	11.2018	12.2018	01.2019	02.2019	03.2019	04.2019
matricula																
352799	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11.0	9.0	3.0	13.0
356700	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	14.0	17.0	4.0	24.0
381853	10.0	14.0	13.0	10.0	13.0	7.0	3.0	1.0	1.0	2.0	0.0	0.0	16.0	10.0	7.0	16.0
461342	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	11.0	6.0	9.0	5.0
548227	4.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.0	7.0	0.0	3.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
76037444	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
76037452	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
76037460	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
76037495	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
76037517	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

12059 rows × 61 columns

Fonte: Autor (2025)

Portanto, ao realizar o agrupamento separadamente nesses subconjuntos, tornou-se possível lidar de forma mais eficaz com a disparidade na disponibilidade de dados, especialmente no caso dos usuários com registros menos completos.

Em seguida, foi realizado o processo de normalização em ambos os conjuntos de dados, utilizando o método de escalonamento Min-Max. Essa técnica transformou os valores das variáveis para uma escala comum, preservando as diferenças relativas entre os atributos. Como

resultado, os valores de cada variável passaram a variar dentro do intervalo de 0 a 1, garantindo uniformidade na escala sem comprometer a integridade das informações.

Após o processo de normalização, iniciou-se a construção da matriz de consumo, que serve como dado de entrada para o algoritmo de extração das caracteristicas. Nessa matriz, as colunas representam as matrículas e as linhas correspondem aos meses do período analisado, totalizando 60 registros temporais. Como resultado, obteve-se uma matriz com 60 linhas por 94.438 colunas para o Conjunto 1, e uma matriz com 60 linhas por 12.059 colunas para o Conjunto 2, conforme ilustrado na Tabela 11.

Tabela 11: consumo Matrizes de consumo normalizadas: a) Conjunto 1, b) Conjunto 2

matricula	2020	23124	24511	28355	32506	44008	58734	79707	82422	99392	138371	141585	141615	152420	152790	163848
01.2018	0.05	0.173913	0.200000	0.000000	0.000000	0.062500	0.260870	0.578947	0.750000	0.652174	0.538462	0.25	0.086957	0.6250	0.090909	0.153846
02.2018	0.05	0.260870	0.133333	0.055556	0.000000	0.020833	0.260870	0.473684	0.583333	0.521739	0.538462	0.25	0.130435	0.2500	0.010101	0.384615
03.2018	0.05	0.217391	0.200000	0.000000	0.000000	0.020833	0.304348	0.526316	0.541667	0.565217	0.692308	0.40	0.086957	0.0000	0.030303	0.000000
04.2018	0.05	0.304348	0.133333	0.055556	0.000000	0.020833	0.521739	0.473684	0.583333	0.521739	0.692308	0.30	0.086957	0.0000	0.101010	0.000000
05.2018	0.05	0.304348	0.133333	0.000000	0.000000	0.062500	1.000000	0.631579	0.583333	0.521739	0.615385	0.25	0.130435	0.0000	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
08.2022	0.20	0.043478	0.000000	0.111111	0.043478	0.145833	0.260870	0.684211	0.291667	0.608696	0.923077	0.60	0.260870	0.5000	0.010101	0.384615
09.2022	0.30	0.086957	0.000000	0.055556	0.043478	0.187500	0.260870	0.684211	0.333333	0.739130	0.846154	1.00	0.260870	0.6250	0.020202	0.538462
10.2022	0.25	0.043478	0.066667	0.000000	0.043478	0.187500	0.304348	0.578947	0.333333	0.695652	0.692308	0.70	0.217391	0.5625	0.010101	0.461538
11.2022	0.25	0.043478	0.000000	0.055556	0.043478	0.229167	0.304348	0.684211	0.250000	0.695652	0.923077	0.65	0.260870	0.5000	0.020202	0.692308
12.2022	0.25	0.043478	0.000000	0.000000	0.043478	0.187500	0.217391	0.578947	0.375000	0.739130	0.692308	0.65	0.217391	0.6250	0.000000	0.384615

60 rows × 94438 columns

matricula	352799	356700	381853	461342	548227	560170	560359	560391	593150	594024	594032	594393	596442	597287	598143	615439
01.2018	0.0	0.0	0.370370	0.0	0.285714	0.210526	0.0	0.0	0.666667	0.0	NaN	NaN	0.000000	0.0	0.0	0.230769
02.2018	0.0	0.0	0.518519	0.0	0.071429	0.578947	0.0	0.0	0.500000	0.0	NaN	NaN	0.000000	0.0	0.0	0.230769
03.2018	0.0	0.0	0.481481	0.0	0.000000	0.526316	0.0	0.0	0.541667	0.0	NaN	NaN	0.000000	0.0	0.0	0.153846
04.2018	0.0	0.0	0.370370	0.0	0.000000	0.421053	0.0	0.0	0.708333	0.0	NaN	NaN	0.000000	NaN	0.0	0.230769
05.2018	0.0	0.0	0.481481	0.0	0.000000	0.578947	0.0	0.0	0.666667	0.0	NaN	NaN	0.000000	NaN	0.0	0.153846
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
08.2022	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.363636	0.40	0.866667	NaN	NaN	NaN
09.2022	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.363636	0.50	1.000000	NaN	NaN	NaN
10.2022	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.363636	0.55	0.933333	NaN	NaN	NaN
11.2022	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.454545	0.35	1.000000	NaN	NaN	NaN
12.2022	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.363636	0.50	0.800000	NaN	NaN	NaN

60 rows × 12059 columns

Fonte: Autor (2025)

5.2 Definição dos parâmetros de entrada

5.2.1 Extração das características para o LZC

Para a seleção dos quatro valores ideais correspondentes às quatro faixas distintas (x_1, x_2, x_3 e x_4), foram realizados testes com um conjunto de matrículas selecionadas aleatoriamente, com o intuito de identificar distintos níveis de complexidade nas séries

temporais de consumo de água, com base no LZC. Dessa forma, diversos valores foram avaliados, conforme ilustrado na Tabela 12.

Tabela 12: Valores do NLZC para as diferentes faixas.

	matricula	NLZC_2	NLZC_4	NLZC_9	NLZC_20	NLZC_30	NLZC_40	NLZC_50	NLZC_60	NLZC_70	NLZC_80	NLZC_90	NLZC_100	NLZC_110	NLZC_120
0	352799	0.803668	0.589784	0.672472	0.755887	0.773856	0.786134	0.795256	0.802422	0.808273	0.813188	0.817405	0.821084	0.824339	0.827249
1	356700	1.067608	0.719722	0.754630	0.775862	0.790006	0.799970	0.807481	0.813434	0.818323	0.822448	0.825999	0.829106	0.831861	0.834329
2	381853	0.942118	0.690551	0.720028	0.871346	1.080817	1.032032	1.108272	1.117523	1.058911	1.064905	1.136933	1.074552	1.078534	1.149728
3	461342	0.577939	0.453915	0.520658	0.590679	0.608684	0.620696	0.629516	0.636394	0.641982	0.646658	0.650658	0.654141	0.657215	0.659961
4	548227	0.750591	0.599760	0.638769	0.804516	0.819865	0.830618	0.838702	0.845099	0.850347	0.854771	0.858578	0.861907	0.864857	0.867500

Fonte: Autor (2025)

Com base na análise da Tabela 12, definiram-se os valores das faixas como 2 (LZC padrão), 4, 9 e 120, uma vez que cada uma dessas faixas apresenta variações distintas dependendo da leitura de consumo na série temporal e do número de intervalos alocados, resultando em valores que representam a LZC de cada série temporal. Ademais, o valor das características extraídas varia de acordo com o critério de seleção. Por exemplo, as características obtidas dentro de determinados tamanhos de intervalo, como o intervalo de 80 a 119 neste estudo, são semelhantes às extraídas com 120 intervalos, não fornecendo informações adicionais ao modelo.

5.2.2 Parâmetros DBSCAN

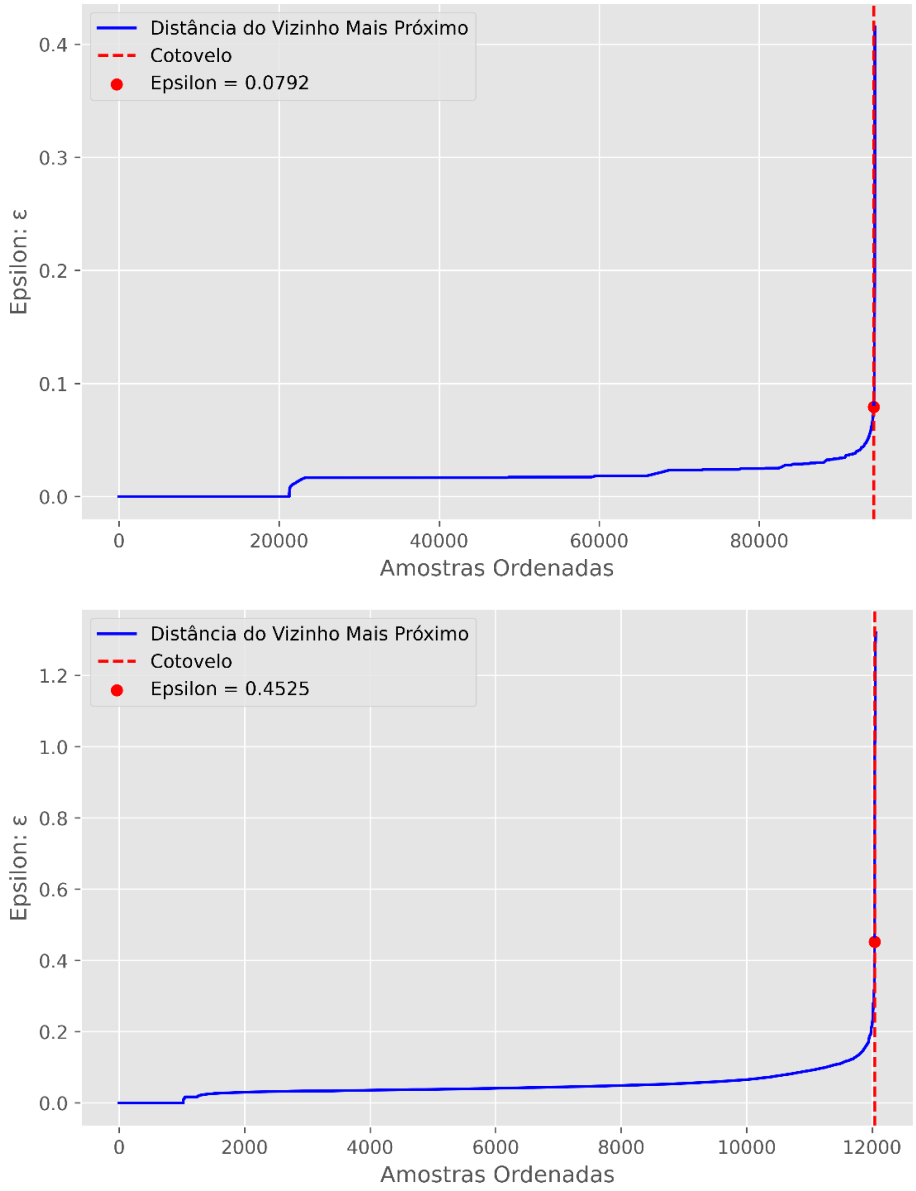
Conforme abordado anteriormente, a implementação do algoritmo DBSCAN requer a definição de dois parâmetros: $MinPts$ e ε. Para estimar o valor de $MinPts$, foi utilizado a Equação 10, considerando a o valor da variável dim igual a 5, correspondente às cinco dimensões da matriz de consumo. Com base nesse cálculo, foi determinado o valor de *MinPts* igual a 10.

Com a variável acima estabelecida, procedeu-se à estimativa do segundo parâmetro necessário para a aplicação do algoritmo. Para isso, inicialmente, utilizou-se o método KNN. O valor de K foi calculado por meio da Equação 11, considerando $MinPts$ igual a 10, conforme mencionado anteriormente. A partir desse cálculo, obteve-se K igual a 9, o que implica na consideração da distância até o nono vizinho mais próximo para cada ponto do conjunto de dados. Posteriormente, essas distâncias foram organizadas em ordem crescente.

Em sequência, para a definição do ponto de cotovelo, procedeu-se por meio da biblioteca *Kneed*, conforme descrito na metodologia proposta por Ghamkhar et al. (2023). Esse

processo resultou na geração da Figura 27 , onde o eixo Y representa as distâncias até o nono vizinho mais próximo, enquanto o eixo X indica a posição dos pontos na lista ordenada.

Figura 27: Determinação do ponto de cotovelo sem aplicação de suavização: a) Conjunto 1, b) Conjunto 2

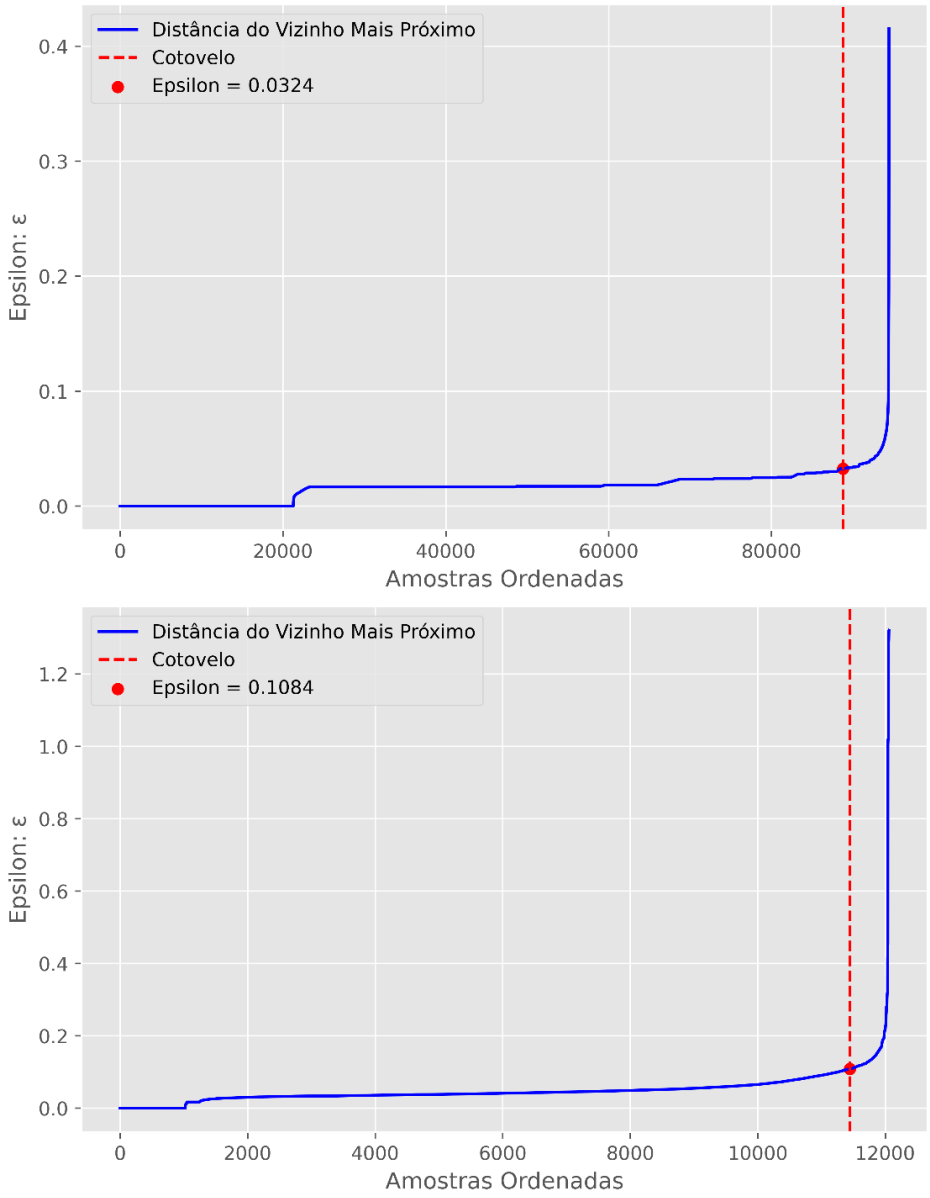


Fonte: Autor (2025)

Ao analisar as figuras, observa-se que a definição do ponto de cotovelo foi incorreta, uma vez que o valor selecionado se encontra acima das recomendações apresentadas no Tópico 4.5. O ponto ideal deveria situar-se em uma região da curva que geralmente apresenta uma transição nítida entre um trecho inicial mais plano e outro com inclinação significativamente acentuada. Essa imprecisão decorre das características específicas dos dados analisados, como a ausência de uma inflexão acentuada e a suavidade da curva, o que comprometeu a eficácia da biblioteca na identificação do ponto ideal, resultando na escolha de valores inadequados para o parâmetro ε em ambos os conjuntos.

Diante dessa limitação, tornou-se necessária a realização de uma análise visual complementar, com o objetivo de identificar um valor mais apropriado para o parâmetro ε, assegurando maior precisão na identificação dos agrupamentos. Para isso, foi necessário ajustar o parâmetro de suavização fornecido pela própria biblioteca *Kneed*, o qual influencia diretamente na sensibilidade da detecção do ponto de inflexão. Esse ajuste foi realizado de forma empírica, por meio da variação incremental do valor de suavização e da análise visual do gráfico gerado, até que fosse possível identificar um "joelho" que representasse adequadamente a transição esperada nos dados. Como resultado, obteve-se os valores ilustrados na Figura 28.

Figura 28: Determinação do ponto de cotovelo com a aplicação da suavização: a) Conjunto 1, b) Conjunto 2



Fonte: Autor (2025)

Portanto, após uma análise visual minuciosa dos gráficos, foram selecionados os pontos de inflexão mais condizentes com o comportamento esperado da curva. Como resultado,

obteve-se o valor do parâmetro ε igual a 0,0324 para o Conjunto 1 e 0,1084 para o Conjunto 2, conforma demonstrado na Figura 28(a) e Figura 28(b) respectivamente.

5.3 Resultados do modelo

Após a definição dos parâmetros de entrada, as características TSLF e LZC foram extraídas de ambos os conjuntos, conforme descrito na metodologia. Em seguida, o algoritmo DBSCAN foi aplicado ao treinamento dos conjuntos de dados Conjunto 1 e Conjunto 2. Os resultados obtidos a partir desse processo são apresentados nos subcapítulos seguintes.

5.3.1 Conjunto 1

Conforme abordado anteriormente, os valores dos parâmetros obtidos para o Conjunto 1 foram $\varepsilon = 0{,}0324$ e $MinPts = 10$, resultando na formação de 3 agrupamentos e 2.622 ruídos, correspondendo a 2,78% do total de usuários.

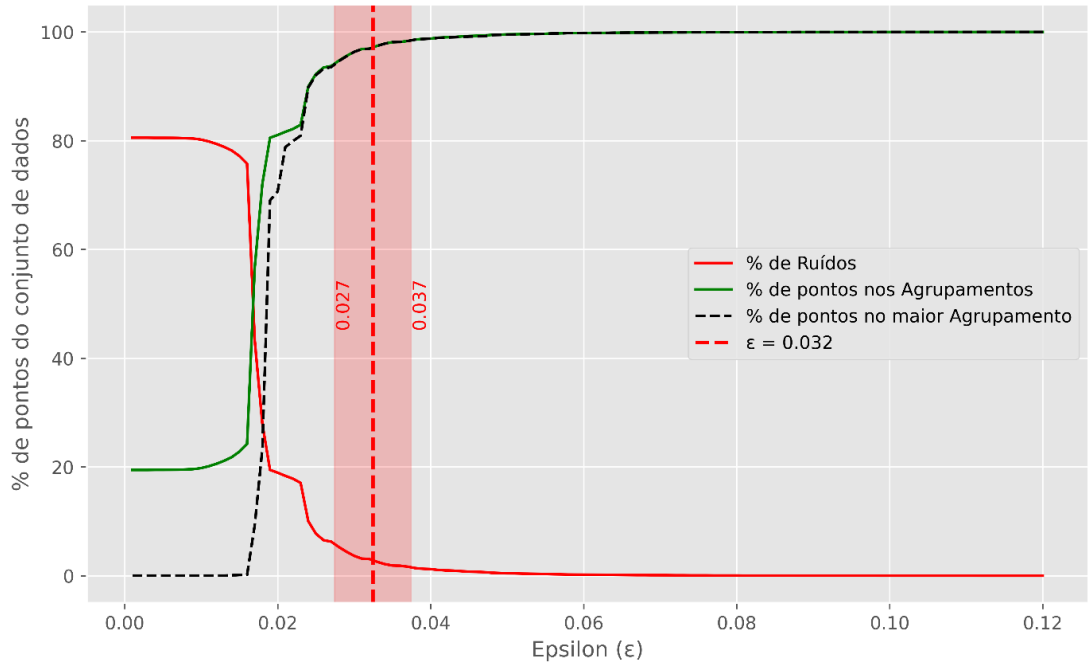
A avaliação do resultado, foi realizada com base nos critérios previamente estabelecidos na metodologia, os quais consideram a proporção de ruídos e o tamanho do maior agrupamento, que deve ser relativamente constante, sem apresentar grande discrepância em relação aos demais. Diante disso, verifica-se que o primeiro critério foi atendido, uma vez que a proporção de ruídos obtida foi de 2,78%, estando dentro do intervalo recomendado de 1% a 30%. Para a verificação do segundo critério, bem como para análises complementares, foram testados diversos valores do parâmetro ε. Os agrupamentos resultantes estão apresentados na Tabela 13. Com base nesses dados, foi elaborada a Figura 29, com o intuito de proporcionar uma visualização mais clara dos tamanhos dos agrupamentos em função da variação de ε.

Tabela 13: Variação dos agrupamentos em função dos valores de ε do Conjunto 1

	ε	Quantidade de Agrupamentos	% de Pontos dentro dos Agrupamentos	Ruidos	% de Ruidos	Pontos do Maior Agrupamento	% do Maior Agrupamento
0	0.001	1282	19.45	76071	80.55	38	0.04
1	0.002	1282	19.45	76070	80.55	38	0.04
2	0.003	1282	19.45	76067	80.55	38	0.04
3	0.004	1283	19.47	76047	80.53	39	0.04
4	0.005	1283	19.47	76047	80.53	39	0.04
...	...	...	...	...	...	...	...
115	0.116	1	99.99	11	0.01	94427	99.99
116	0.117	1	99.99	11	0.01	94427	99.99
117	0.118	1	99.99	11	0.01	94427	99.99
118	0.119	1	99.99	11	0.01	94427	99.99
119	0.120	1	99.99	11	0.01	94427	99.99

Fonte: Autor (2025)

Figura 29: Variação da proporção de ruídos e tamanho dos agrupamentos em função de ε para o Conjunto 1



Fonte: Autor (2025)

A Figura 29 apresenta a influência do parâmetro ε na aplicação do algoritmo DBSCAN, considerando três variáveis principais: a porcentagem de pontos identificados como ruídos, a porcentagem de pontos atribuídos aos agrupamentos e a porcentagem de pontos pertencentes ao maior agrupamento. Observa-se que o tamanho do maior agrupamento permanece relativamente constante nas proximidades do valor de ε previamente definido, atendendo, portanto, ao segundo critério estabelecido.

Para valores reduzidos de ε, uma parcela significativa dos dados é classificada como ruído, o que indica a ausência de densidade suficiente para a formação de agrupamentos. Com o aumento de ε, verifica-se uma queda acentuada na quantidade de ruídos e um crescimento significativo na proporção de pontos atribuídos aos agrupamentos, sinalizando o início da formação de agrupamentos consistentes. A linha preta tracejada indica que, a partir de valores de ε superiores a 0,024, um único agrupamento tende a concentrar a maior parte dos dados, o que pode caracterizar sobreajuste e comprometer a capacidade do modelo em segmentar padrões distintos de comportamento.

A faixa destacada em vermelho no gráfico delimita o intervalo de valores de ε compreendido entre 0,027 e 0,037, escolhido para permitir uma análise mais refinada em torno do ponto previamente definido, representado pela linha tracejada. Entretando, a faixa coincide com a região considerada ideal para o valor de ε, na qual se observa um equilíbrio entre a redução dos ruídos e a manutenção de uma estrutura de agrupamentos heterogênea, favorecendo

a identificação de comportamentos irregulares. Portanto, a área sombreada representa os intervalos viáveis para ε. A fim de investigar com maior precisão os valores próximos ao ponto de cotovelo, os dados da Tabela 13 foram filtrados, resultando nos valores ilustrados na Tabela 14.

Tabela 14: Variação dos agrupamentos nos valores de ε em torno do cotovelo no Conjunto 1

	ε	Quantidade de Agrupamentos	% de Pontos dentro dos Agrupamentos	Ruídos	% de Ruídos	Pontos do Maior Agrupamento	% do Maior Agrupamento
25	0.026	19	93.46	6180	6.54	88052	93.24
26	0.027	17	93.73	5925	6.27	88324	93.53
27	0.028	10	94.72	4989	5.28	89372	94.64
28	0.029	7	95.59	4163	4.41	90210	95.52
29	0.030	2	96.34	3459	3.66	90969	96.33
30	0.031	3	96.85	2972	3.15	91446	96.83
31	0.032	3	96.93	2898	3.07	91520	96.91
32	0.033	2	97.44	2417	2.56	92012	97.43
33	0.034	3	97.86	2021	2.14	92400	97.84
34	0.035	1	98.12	1778	1.88	92660	98.12
35	0.036	1	98.16	1741	1.84	92697	98.16
36	0.037	1	98.33	1577	1.67	92861	98.33
37	0.038	2	98.60	1320	1.40	93107	98.59
38	0.039	2	98.72	1207	1.28	93207	98.70

Fonte: Autor (2025)

Observa-se que, à medida que o valor de ε se aproxima de 0,032, ocorre uma transição significativa no comportamento do algoritmo. Há uma redução na quantidade de ruídos e um aumento no número de pontos alocados em agrupamentos, o que indica que o algoritmo começa a identificar estruturas mais amplas nos dados, incorporando pontos anteriormente considerados isolados. Como consequência, diversos agrupamentos menores tendem a se fundir, o que reduz o número total de agrupamentos distintos. Com isso, um ou poucos agrupamentos passam a concentrar a maior parte dos dados, o que pode sinalizar o início de um possível sobreajuste, no qual o algoritmo perde sensibilidade para identificar subdivisões relevantes no conjunto. Essa condição é confirmada pela diferença praticamente inexistente entre a porcentagem total de pontos alocados em agrupamentos e a proporção concentrada no maior agrupamento.

Além disso, é possível observar que, com a redução contínua de ε, o número de ruídos aumenta progressivamente. No entanto, os ruídos previamente identificados são mantidos, o que mostra que o algoritmo é cumulativo na detecção de ruídos, reforçando a importância de uma escolha cuidadosa do parâmetro.

Dessa forma, observa-se que os valores de ε situados nas proximidades do ponto de cotovelo apresentam resultados relativamente semelhantes. Assim, o valor selecionado com base nesse ponto pode não corresponder necessariamente à configuração que maximiza a

eficiência do algoritmo, sobretudo na ausência de um conjunto de dados previamente rotulado, o que inviabiliza a aplicação de técnicas de validação supervisionada. Cabe destacar que muitas concessionárias de abastecimento não dispõem de dados rotulados, devido à recente aquisição dos contratos de concessão, o que resulta na falta de um histórico consistente para subsidiar esse tipo de análise.

Considerando essa limitação, Ghamkhar et al. (2023) destacam que a definição dos parâmetros mais adequados para esse tipo de conjunto de dados depende, essencialmente, de dois fatores: a quantidade de usuários que a companhia pretende inspecionar e a natureza dos padrões representados por cada agrupamento.

Em relação ao primeiro fator, a porcentagem de medidores e conexões que poderão ser efetivamente inspecionados pela equipe técnica da companhia está diretamente associada ao número de ruídos detectados. Com base na Tabela 14, e considerando a similaridade nos resultados obtidos ao redor do ponto de cotovelo, recomenda-se a adoção de $\varepsilon = 0{,}037$ (com 1,67% de ruídos) para cenários com restrições orçamentárias, e $\varepsilon = 0{,}026$ (com 6,54% de ruídos) quando houver maior disponibilidade de recursos para a realização de inspeções adicionais

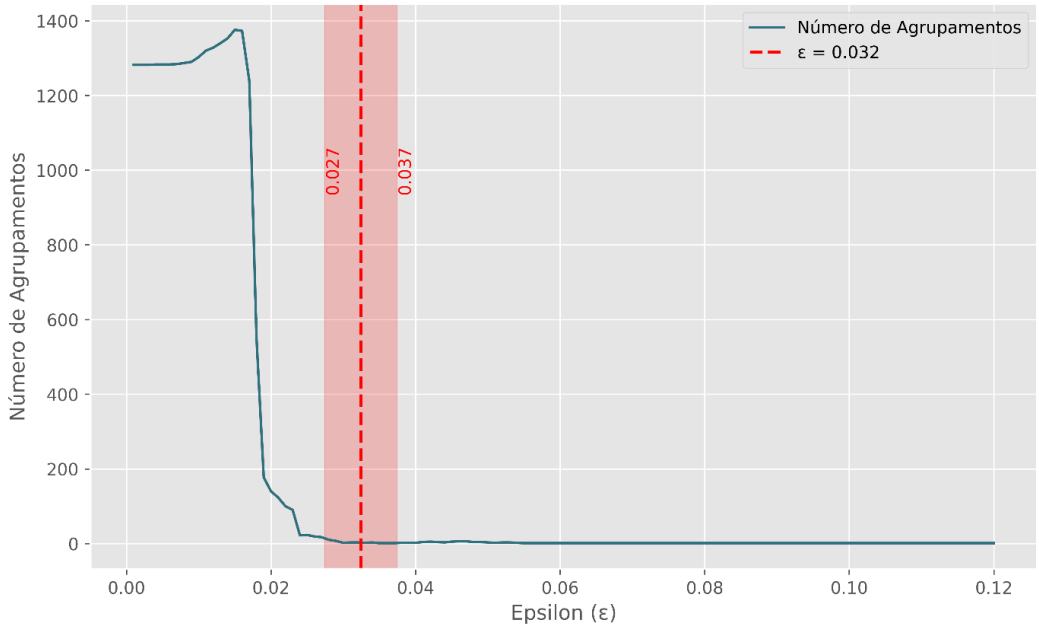
Em relação ao segundo fator, conforme descrito por Ghamkhar et al. (2023), é esperado que o primeiro e maior agrupamento seja predominantemente composto por usuários com padrões de consumo regulares, classificados como “normais”. À medida que a quantidade de ruídos aumenta, observa-se uma redução no número de usuários pertencentes a esse agrupamento principal. O segundo agrupamento, denominado “zeros”, é formado por usuários cuja série temporal de consumo é inteiramente composta por valores nulos ou apresenta poucos registros diferentes de zero, mantendo, em geral, uma quantidade de membros praticamente constante. Os demais usuários podem ser distribuídos em um ou mais agrupamentos adicionais, sendo que qualquer usuário não associado a um agrupamento específico é classificado como ruído.

Conforme mencionado anteriormente, com a utilização do ponto de cotovelo para definição do parâmetro ε, obteve-se a formação de três agrupamentos distintos. Ao analisar os elementos presentes nesses agrupamentos, foi possível identificar o agrupamento correspondente ao perfil de consumo “normal”. No entanto, não foi identificado um agrupamento isolado compatível com o chamado “zeros”. A ausência desse agrupamento pode indicar que os usuários com esse perfil foram incorporados a outros agrupamentos, possivelmente por apresentarem padrões semelhantes aos de consumo atípico, o que resultou

em sua fusão com grupos heterogêneos ou na sua classificação como ruídos. Assim, os dois agrupamentos restantes podem conter registros com comportamentos irregulares distintos, agrupados com base em suas particularidades de densidade e proximidade no espaço de características.

Posto isso, elaborou-se a Figura 30, que ilustra a quantidade de agrupamentos resultantes a partir da variação dos valores do parâmetro ε, com o objetivo de proporcionar uma visualização mais clara desse comportamento

Figura 30: Variação da proporção de número de agrupamentos em função do parâmetro ε do Conjunto 1



Fonte: Autor (2025)

Observa-se que, para valores de ε inferiores a aproximadamente 0,023, o algoritmo forma uma quantidade excessiva de agrupamentos, indicando alta sensibilidade a pequenas variações do parâmetro ε a pequenas variações de densidade no espaço de características. Nesse cenário, registros com pequenas diferenças entre si são separados em diferentes agrupamentos, gerando uma fragmentação excessiva da base de dados. À medida que ε aumenta, observa-se uma redução abrupta no número de agrupamentos, refletindo a fusão de grupos muito específicos em conjuntos mais coesos.

Portanto, diante da análise da variação do número de agrupamentos em função do parâmetro ε, conclui-se que valores ligeiramente inferiores, situados entre 0,024 e 0,032, apresentam potencial para uma segmentação mais sensível e eficaz. Essa faixa preserva uma quantidade adequada de agrupamentos e pode favorecer a identificação de grupos menos densos, como os usuários com consumo nulo.

5.3.2 Conjunto 2

Conforme abordado anteriormente, os valores dos parâmetros obtidos para o Conjunto 2 foram $\varepsilon = 0{,}1084$ e $MinPts = 10$, resultando na formação de 29 agrupamentos e 387 ruídos, correspondendo a 3,21% do total de usuários.

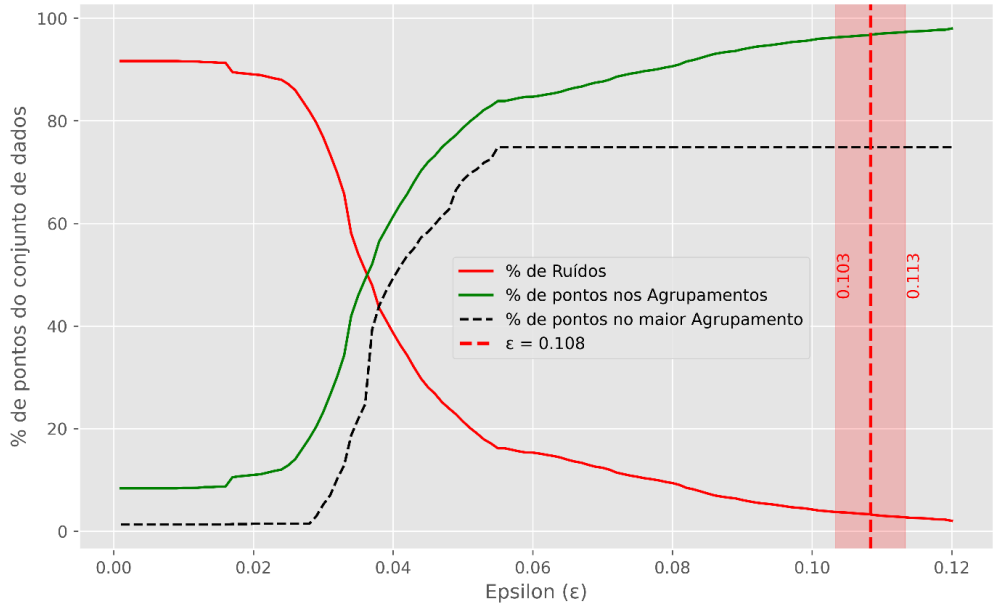
Para fins de validação, o primeiro critério foi atendido, uma vez que a proporção de ruídos permaneceu dentro do intervalo considerado aceitável. No que se refere à verificação do segundo critério, bem como a análises complementares, os agrupamentos gerados para diferentes valores de ε estão apresentados na Tabela 15. Adicionalmente, a Figura 31 foi elaborada com o objetivo de facilitar a visualização da distribuição dos tamanhos dos agrupamentos em função da variação desse parâmetro.

Tabela 15: Variação dos agrupamentos em função dos valores de ε do Conjunto 2

	ε	Quantidade de Agrupamentos	% de Pontos dentro dos Agrupamentos	Ruídos	% de Ruídos	Pontos do Maior Agrupamento	% do Maior Agrupamento
0	0.001	19	8.35	11052	91.65	156	1.29
1	0.002	19	8.35	11052	91.65	156	1.29
2	0.003	19	8.35	11052	91.65	156	1.29
3	0.004	19	8.35	11052	91.65	156	1.29
4	0.005	19	8.35	11052	91.65	156	1.29
...	...	...	...	...	...	...	...
115	0.116	28	97.50	302	2.50	9027	74.86
116	0.117	28	97.60	290	2.40	9027	74.86
117	0.118	28	97.73	274	2.27	9027	74.86
118	0.119	27	97.74	272	2.26	9027	74.86
119	0.120	28	98.00	241	2.00	9027	74.86

Fonte: Autor (2025)

Figura 31: Variação da proporção de ruídos e tamanho dos agrupamentos em função de ε para o Conjunto 2



Fonte: Autor (2025)

A partir da análise da Figura 31, observa-se que o tamanho do maior agrupamento permanece relativamente constante nas proximidades do valor de ε previamente definido, atendendo, assim, ao segundo critério estabelecido.

Assim como no Conjunto 1, valores baixos de ε resultam em alta proporção de ruídos devido à baixa densidade. Com o aumento de ε, há redução dos ruídos e aumento dos pontos agrupados, indicando a formação de agrupamentos consistentes. A linha preta tracejada indica que, a partir de valores de ε superiores a aproximadamente 0,055, o maior agrupamento se estabiliza, concentrando cerca de 75% dos dados do conjunto. Esse comportamento evidencia que o agrupameno dominante já está consolidado, e aumentos adicionais no valor de ε apenas reduzem a quantidade de ruídos, que passam a compor e expandir agrupamentos menores.

Diferentemente do Conjunto 1, não é possível afirmar que a faixa destacada em vermelho no gráfico corresponde à região ideal para os valores do parâmetro ε, uma vez que a estabilidade dos agrupamentos é observada apenas a partir de ε = 0,055. Dentro desse intervalo, verifica-se uma redução expressiva na proporção de ruídos, acompanhada por um aumento gradual na concentração de pontos nos agrupamentos, sendo o maior agrupamento mantido constante ao longo desse trecho. Com o objetivo de aprofundar a análise na vizinhança do ponto de cotovelo, os dados apresentados na Tabela 15 foram filtrados, resultando nas informações detalhadas na Tabela 16.

Tabela 16: Variação dos agrupamentos nos valores de ε em torno do cotovelo no Conjunto 2

	ε	Quantidade de Agrupamentos	% de Pontos dentro dos Agrupamentos	Ruídos	% de Ruídos	Pontos do Maior Agrupamento	% do Maior Agrupamento
98	0.099	27	95.57	534	4.43	9027	74.86
99	0.100	28	95.78	509	4.22	9027	74.86
100	0.101	28	96.00	482	4.00	9027	74.86
101	0.102	28	96.10	470	3.90	9027	74.86
102	0.103	28	96.23	455	3.77	9027	74.86
103	0.104	28	96.33	443	3.67	9027	74.86
104	0.105	28	96.39	435	3.61	9027	74.86
105	0.106	28	96.51	421	3.49	9027	74.86
106	0.107	28	96.62	408	3.38	9027	74.86
107	0.108	28	96.68	400	3.32	9027	74.86
108	0.109	29	96.88	376	3.12	9027	74.86
109	0.110	29	97.01	360	2.99	9027	74.86
110	0.111	29	97.11	349	2.89	9027	74.86
111	0.112	29	97.18	340	2.82	9027	74.86
112	0.113	29	97.27	329	2.73	9027	74.86
113	0.114	29	97.40	314	2.60	9027	74.86
114	0.115	29	97.45	307	2.55	9027	74.86
115	0.116	28	97.50	302	2.50	9027	74.86
116	0.117	28	97.60	290	2.40	9027	74.86
117	0.118	28	97.73	274	2.27	9027	74.86

Fonte: Autor (2025)

De modo semelhante ao observado no Conjunto 1, verifica-se que, à medida que o valor de ε se aproxima do ponto de cotovelo previamente definido, ocorre uma redução na proporção de ruídos e um aumento na densidade dos agrupamentos, indicando que o algoritmo começa a integrar pontos anteriormente isolados a estruturas existentes.

Em teoria, esse comportamento poderia levar à fusão de agrupamentos menores, concentrando os dados em poucos grupos e sugerindo um possível sobreajuste, com consequente perda de sensibilidade na identificação de subdivisões relevantes. Entretanto, essa condição não se confirma, uma vez que permanece praticamente estável a quantidade de agrupamentos, ao mesmo tempo em que o maior agrupamento mantém sua proporção constante e a inclusão progressiva de pontos ocorre de forma equilibrada. Isso sugere que, mesmo com a redução contínua dos ruídos e o aumento da densidade, o modelo preserva sua capacidade de segmentação relevante, sem aglutinar excessivamente os dados em um único agrupamento dominante.

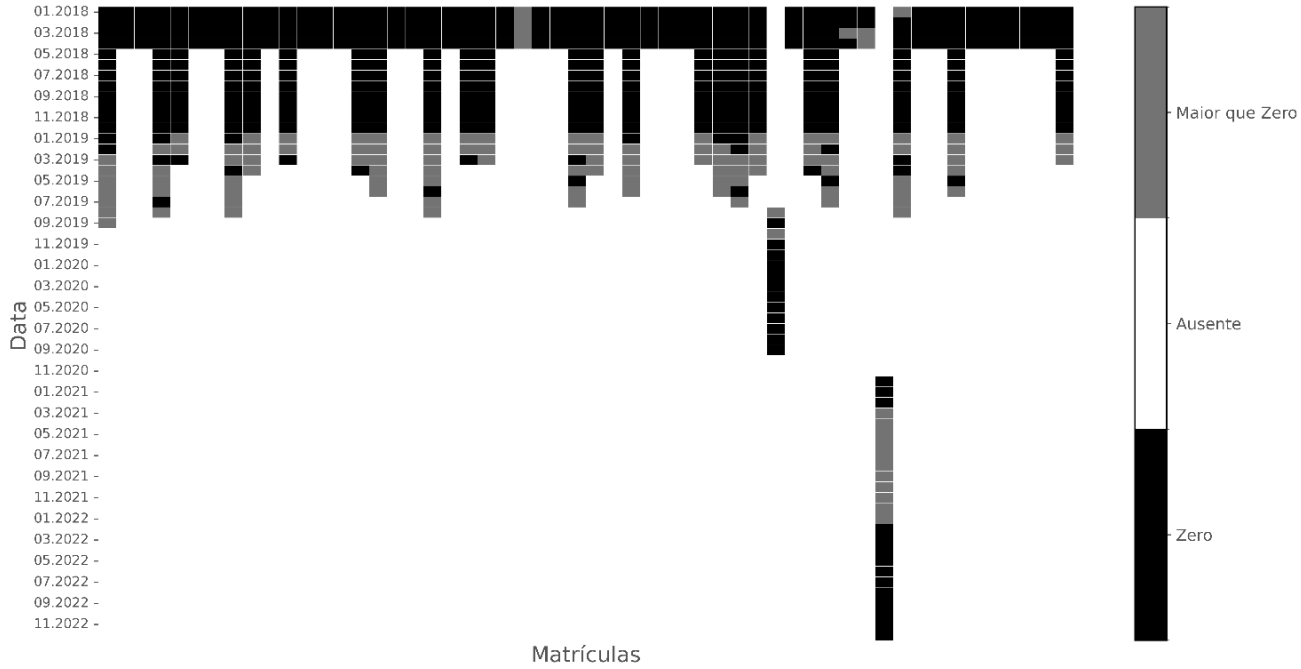
Além disso, à medida que o valor de ε é progressivamente reduzido, observa-se um aumento contínuo na quantidade de ruídos, mantendo-se aqueles previamente identificados, o que evidencia o comportamento cumulativo do algoritmo na detecção dessas ocorrências. Nota-se, ainda, que os valores de ε próximos ao ponto de cotovelo produzem resultados relativamente semelhantes, sobretudo após esse ponto. Portanto, o valor definido como ponto de cotovelo pode não representar, necessariamente, a configuração que maximiza a eficiência do algoritmo.

Levando em consideração essa limitação e a recomendação de Ghamkhar et al. (2023) quanto a esse aspecto, tem-se para o primeiro fator, com base na Tabela 16 e na similaridade dos resultados obtidos nas proximidades do ponto de cotovelo, a recomendação da adoção de $\varepsilon = 0{,}116$ (com 2,50% de ruídos) em cenários com restrições orçamentárias, e $\varepsilon = 0{,}091$ (com 5,80% de ruídos) quando houver maior disponibilidade de recursos para a realização de inspeções adicionais.

Em relação ao segundo fator, conforme mencionado anteriormente, com a utilização do ponto de cotovelo para definição do parâmetro ε, obteve-se a formação de 29 agrupamentos distintos. Diferentemente do Conjunto 1, a análise dos elementos presentes nos agrupamentos permitiu identificar agrupamentos correspondentes tanto ao perfil de consumo “normal” quanto ao perfil de consumo “zeros”. Ressalta-se que ambos os perfis apresentaram uma elevada proporção de dados ausentes. Em alguns agrupamentos, observam-se exclusivamente valores nulos ou ausentes, indicando registros praticamente vazios. Em outros, há predominância de

zeros e valores nulos embora, pontualmente, possam ocorrer valores distintos de zero. Esses agrupamentos podem ser melhor visualizados na Figura 32, na qual o eixo *y* representa as datas e o eixo *x* corresponde às matrículas. Os valores iguais a zero estão representados na cor preta, os valores diferentes de zero em cinza, e os valores ausentes na cor branca.

Figura 32: Agrupamentos com predominância de registros zerados ou nulos.

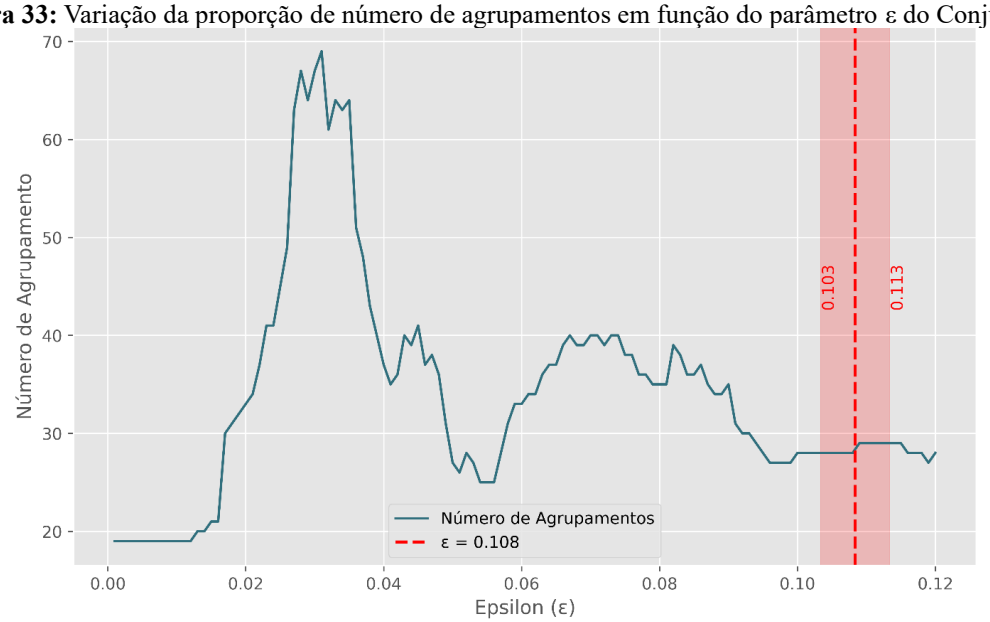


Fonte: Autor (2025)

De modo geral, tais agrupamentos foram formados com base na quantidade de valores nulos e de zeros presentes em cada matrícula, refletindo diferentes padrões de sequência. A depender da quantidade e da distribuição desses padrões, os registros foram classificados em grupos distintos. Dessa forma, percebe-se que o processo de agrupamentos foi eficaz em separar as séries com base em seus padrões de preenchimento, mesmo sem a necessidade de rótulos prévios. Essa segmentação fornece uma base valiosa para análises posteriores, permitindo, por exemplo, a identificação de registros que devem ser desconsiderados, tratados ou investigados individualmente por apresentarem irregularidades ou inconsistências recorrentes.

Posto isso, elaborou-se a Figura 33, que ilustra a quantidade de agrupamentos resultantes a partir da variação dos valores do parâmetro ε, com o objetivo de proporcionar uma visualização mais clara desse comportamento

Figura 33: Variação da proporção de número de agrupamentos em função do parâmetro ε do Conjunto 2



Fonte: Autor (2025)

De modo semelhante ao comportamento observado no Conjunto 1, valores baixos de ε resultam em uma fragmentação excessiva dos dados, com a formação de numerosos agrupamentos devido à elevada sensibilidade do algoritmo a pequenas variações de densidade. À medida que ε é incrementado, o gráfico apresenta oscilações, evidenciando que o algoritmo ainda não consegue distinguir padrões relevantes sem segmentar excessivamente os dados. Nas proximidades do ponto de cotovelo, observa-se a estabilização do número de agrupamentos em torno de 28, indicando que o modelo atinge uma configuração mais estável e generalista, com a formação de agrupamentos mais amplos e menos suscetíveis a variações menores entre os registros.

Portanto, a partir da análise da variação do número de agrupamentos em função do parâmetro ε, conclui-se que valores situados nas proximidades do ponto de cotovelo tendem a oferecer os melhores resultados, por promoverem um equilíbrio entre a fragmentação excessiva dos dados e a proporção de ruídos. Ressalta-se, ainda, que além da análise dos ruídos, é fundamental investigar os agrupamentos individualmente, uma vez que podem conter irregularidades ou inconsistências recorrentes.

5.4 Padrões de consumo

A partir dos resultados obtidos pelo modelo de agrupamento DBSCAN, identificou-se um conjunto de registros classificados como ruído, totalizando 3.009 matrículas que não se ajustaram a nenhum dos agrupamentos formados. Tais registros indicam padrões de consumo que se desviam significativamente do comportamento predominante da base de dados.

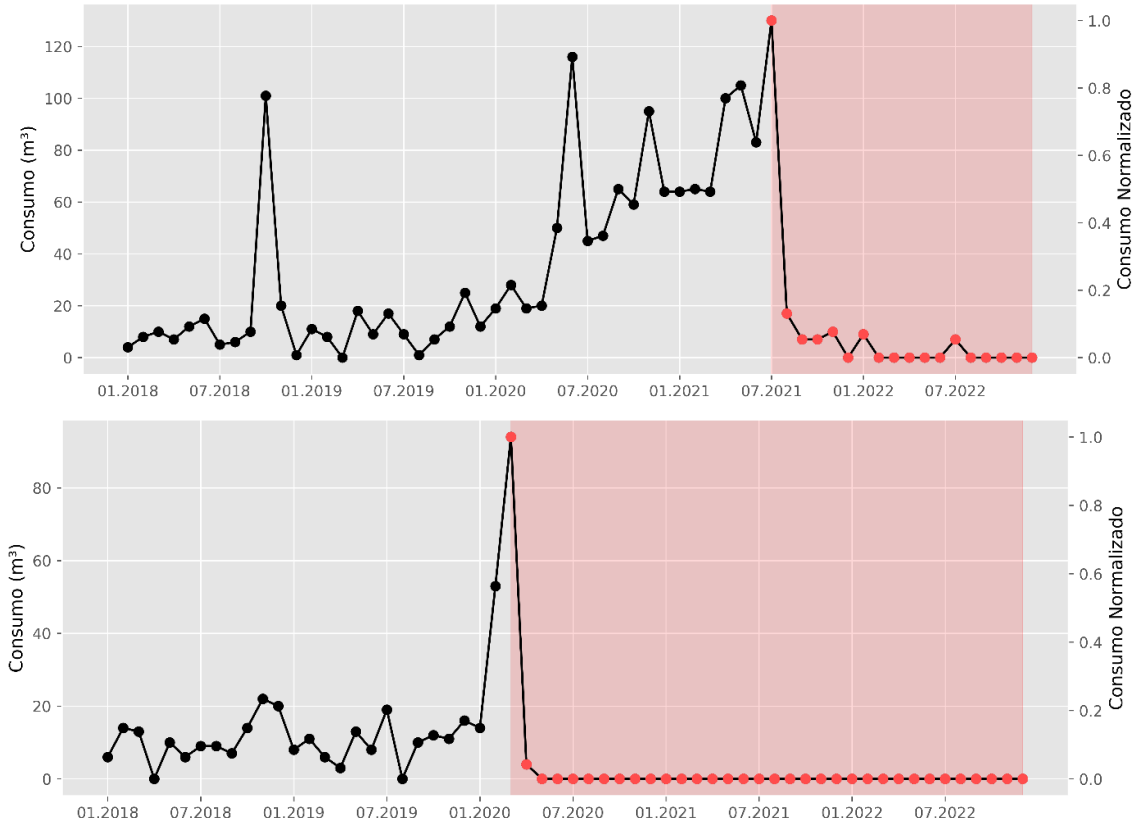
Conforme discutido no tópico 3.3.3, esses padrões atípicos podem decorrer de diversos fatores, incluindo irregularidades operacionais, falhas de medição ou práticas fraudulentas.

Diante desse cenário, realizou-se uma análise detalhada das séries temporais associadas a essas matrículas, com o objetivo de verificar a presença de padrões irregulares semelhantes aos descritos por Ghamkhar et al. (2023). Essa análise foi conduzida manualmente e de forma visual, por meio da inspeção individual dos gráficos de consumo, permitindo uma avaliação criteriosa de cada caso. Como resultado, foi possível identificar os seis padrões irregulares propostos pelos autores, a saber: (i) queda súbita seguida de estabilização; (ii) aumento súbito seguido de estabilização; (iii) pico(s) pontual(is) de consumo; (iv) consumo consistentemente baixo; (v) queda gradual seguida de estabilização; e (vi) registros com consumo zerado.

5.4.1 Queda súbita seguida de estabilização

Casos exemplares de consumo de água e séries temporais normalizadas representando o primeiro padrão são apresentados na Figura 34. Observa-se que esses casos mantêm um comportamento regular até o momento em que ocorre uma queda abrupta no consumo, a partir da qual os valores se estabilizam em um patamar significativamente inferior ao anterior.

Figura 34: Padrão 1 – queda súbita seguida de estabilização



Fonte: Autor (2025)

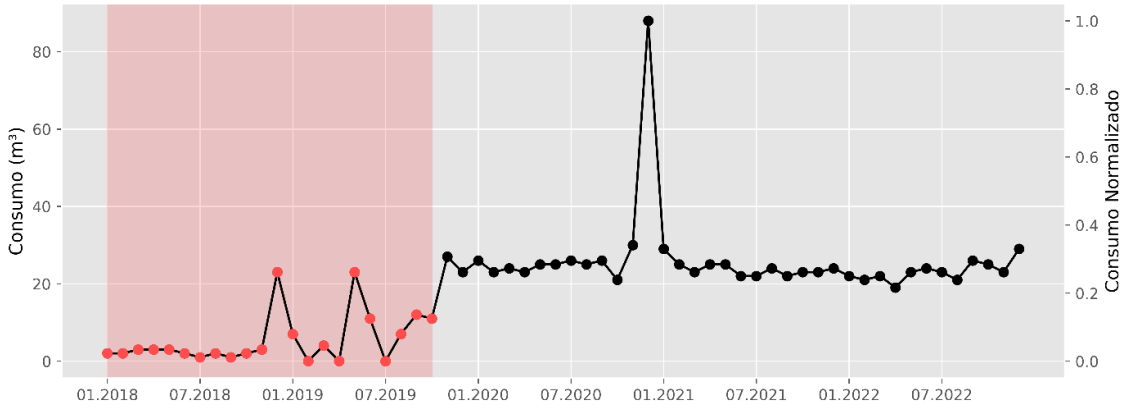
Do ponto de vista da detecção de eventos, o comportamento da Figura 34 (a) e Figura 34 (b) evidencia a ocorrência de um ponto de mudança abrupto, marcado por uma descontinuidade clara entre os padrões anteriores e posteriores ao evento. Essa transição, destacada pela área sombreada nas figuras, caracteriza a passagem entre dois estados distintos do processo de consumo. Na Figura 45 (a), observa-se uma tendência crescente nos valores de consumo até o ponto de mudança, seguido por uma queda abrupta, após a qual os valores passam a oscilar entre zero e valores muito próximos de zero. Já na Figura 45 (b), identifica-se uma tendência de estabilização anterior ao ponto de ruptura, sendo sucedida por uma queda acentuada, com o consumo permanecendo zerado ao longo do tempo.

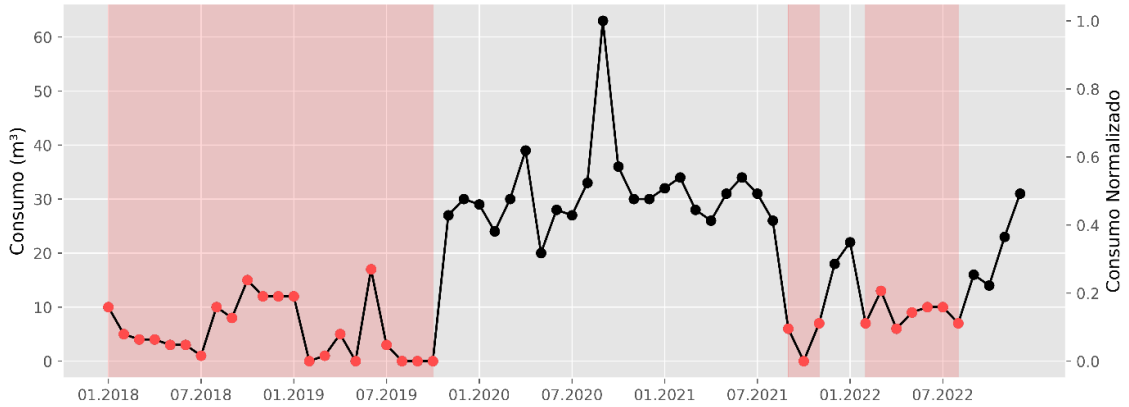
Apesar da ausência de comunicação formal por parte dos usuários nesses casos, é possível que a mudança esteja relacionada a ações intencionais, como adulteração de medidores ou instalação de desvios na conexão de entrada de água. Tais alterações, caso realizadas discretamente, podem escapar à detecção por agentes de campo, especialmente em imóveis com medidores de difícil acesso ou em áreas de baixa fiscalização (GHAMKHAR et al., 2023).

5.4.2 Aumento súbito seguido de estabilização

A Figura 35 apresenta exemplos que caracterizam o segundo padrão identificado. Diferentemente do padrão anterior, este consiste em um aumento súbito no consumo, seguido por um período de estabilização em níveis superiores ao regime inicial.

Figura 35: Padrão 2 – aumento súbita seguida de estabilização





Fonte: Autor (2025)

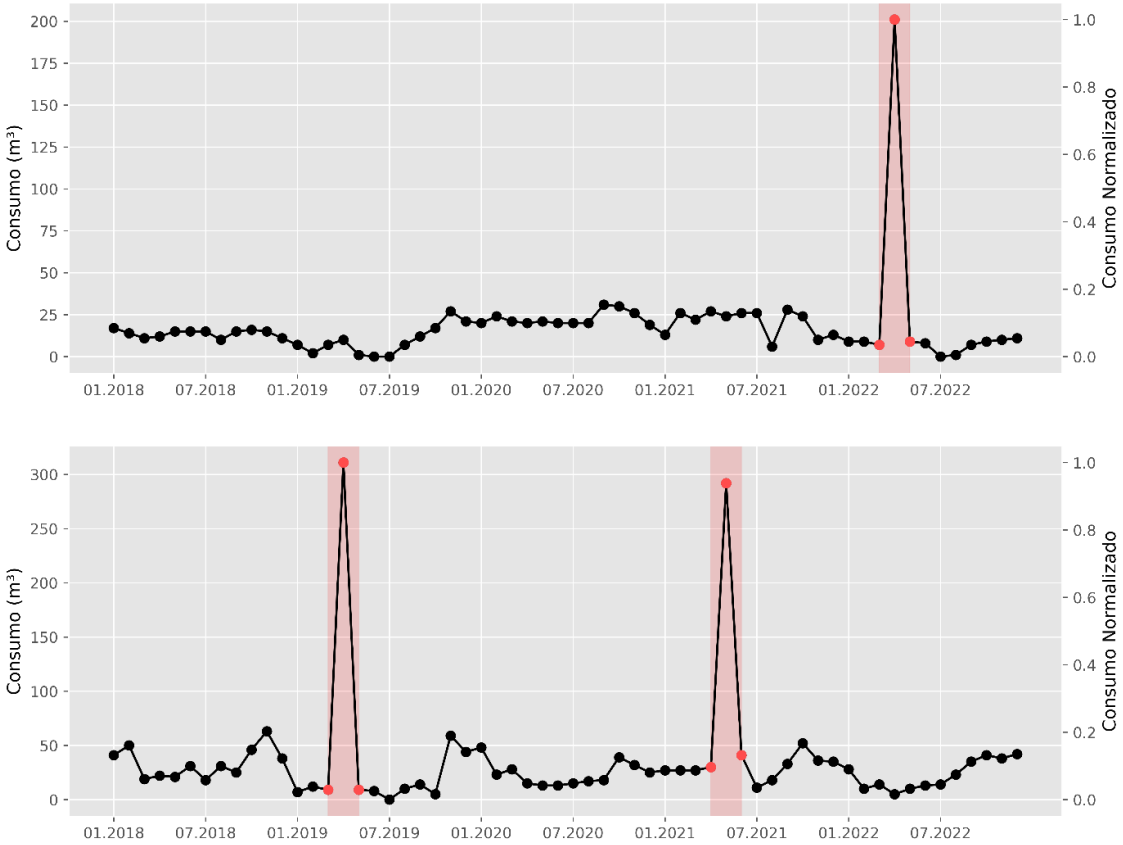
As Figura 35 (a) e Figura 35 (b), apresentam a ocorrência de um ponto de mudança abrupto, caracterizado por um aumento significativo entre os padrões de consumo anteriores e posteriores ao evento, destacado pela área sombreada. Na Figura 46(a), observa-se um consumo relativamente estável em níveis baixos até meados de 2019, momento em que ocorre um aumento repentino, seguido de estabilização. Já na Figura 46(b), identifica-se um padrão semelhante, marcando a transição entre dois regimes distintos de consumo. Contudo, nesta última figura, observa-se, no início e no final de 2022, uma tendência de crescimento no consumo.

Usuários pertencentes a essa categoria apresentam um ou mais episódios de aumentos repentinos seguidos de estabilização no consumo. Esse comportamento contrasta com o padrão anteriormente analisado e pode estar relacionado a fatores como o aumento no número de ocupantes do imóvel, troca de inquilinos, início de atividades comerciais ou locações temporárias em unidades residenciais. Em determinados casos, o aumento pode estar vinculado à reativação do uso após longos períodos de inatividade, conforme ilustrado nas séries temporais apresentadas. Ademais, há indícios de que tais alterações possam decorrer de ações intencionais, como tentativas de mascarar desvios ou justificar variações de consumo por meio da geração de padrões artificialmente inflados (GHAMKHAR et al., 2023).

5.4.3 Pico(s) pontual(is) de consumo

No terceiro padrão, exemplificado na Figura 36, destacam-se picos de consumo que surgem de forma pontual ao longo da série temporal. Tais elevações isoladas indicam comportamentos atípicos em comparação à tendência predominante.

Figura 36: Padrão 3 – pico(s)s pontual(is) de consumo



Fonte: Autor (2025)

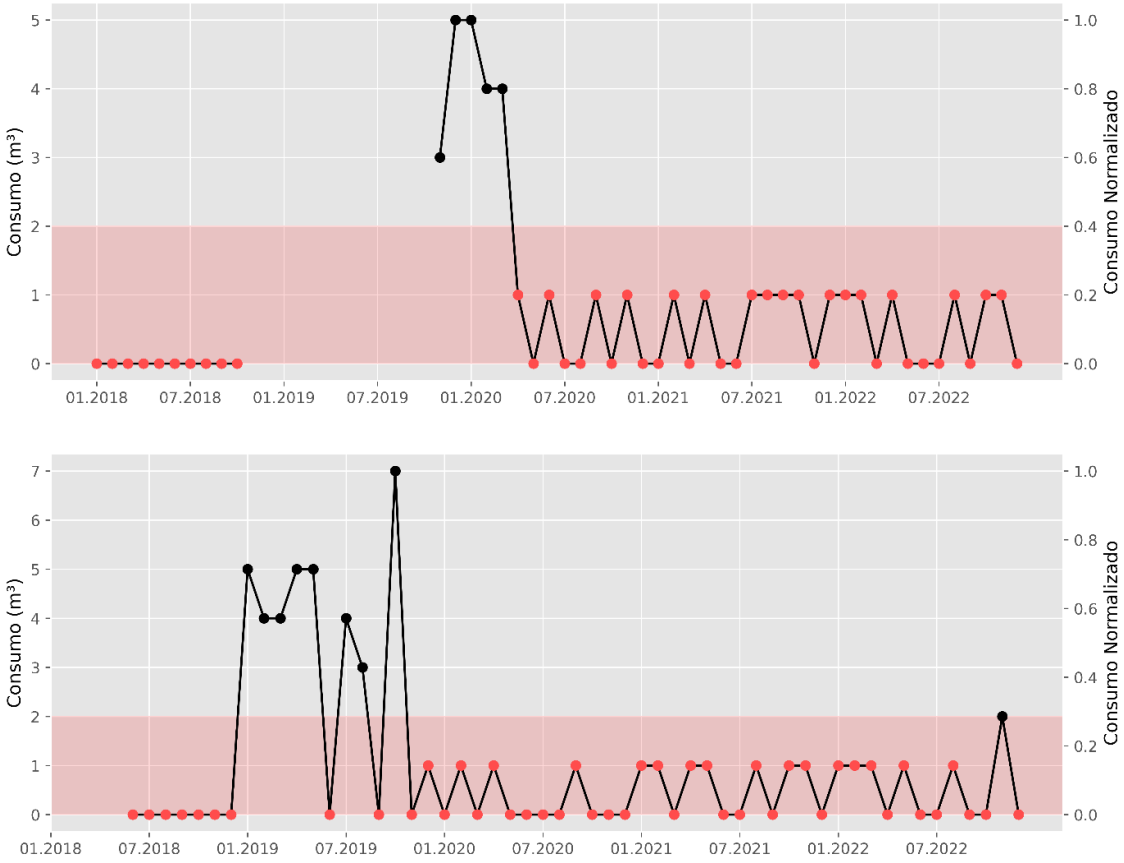
Do ponto de vista da detecção de eventos, as Figura 36 (a) e Figura 36 (b) evidenciam anomalias pontuais, ou seja, valores que destoam significativamente do padrão local da série. As áreas sombreadas destacam os intervalos em que ocorrem esses desvios, reforçando a interpretação de que se trata de episódios isolados e não de mudanças estruturais no comportamento de consumo. Esses picos são isolados e contrastam fortemente com o comportamento regular e estável predominante nas demais observações.

Entre as possíveis causas para esse tipo de padrão estão falhas operacionais, como erros de leitura cometidos por agentes ao registrar os dados dos hidrômetros, problemas de transmissão de dados para o sistema da concessionária, ou até mesmo vazamentos temporários não reportados. Também é plausível que esses picos estejam associados ao uso indevido da água domestica para atividades comerciais ou de construções (GHAMKHAR et al., 2023).

5.4.4 Consumo consistentemente baixo

Representações gráficas do quarto padrão podem ser observadas na Figura 37. Nesse caso, verifica-se um comportamento de consumo consistentemente baixo durante praticamente todo o período analisado, sem variações significativas.

Figura 37: Padrão 4 – consumo consistentemente baixo



Fonte: Autor (2025)

As Figura 37 (a) e Figura 37 (b) apresentam séries temporais nas quais os volumes de água consumidos permanecem sistematicamente em níveis muito baixos ao longo de todo o período analisado. Em grande parte dessas séries, nota-se a ausência de variações expressivas e a predominância de valores próximos de zero em extensos trechos, conforme destacado pelas áreas sombreadas.

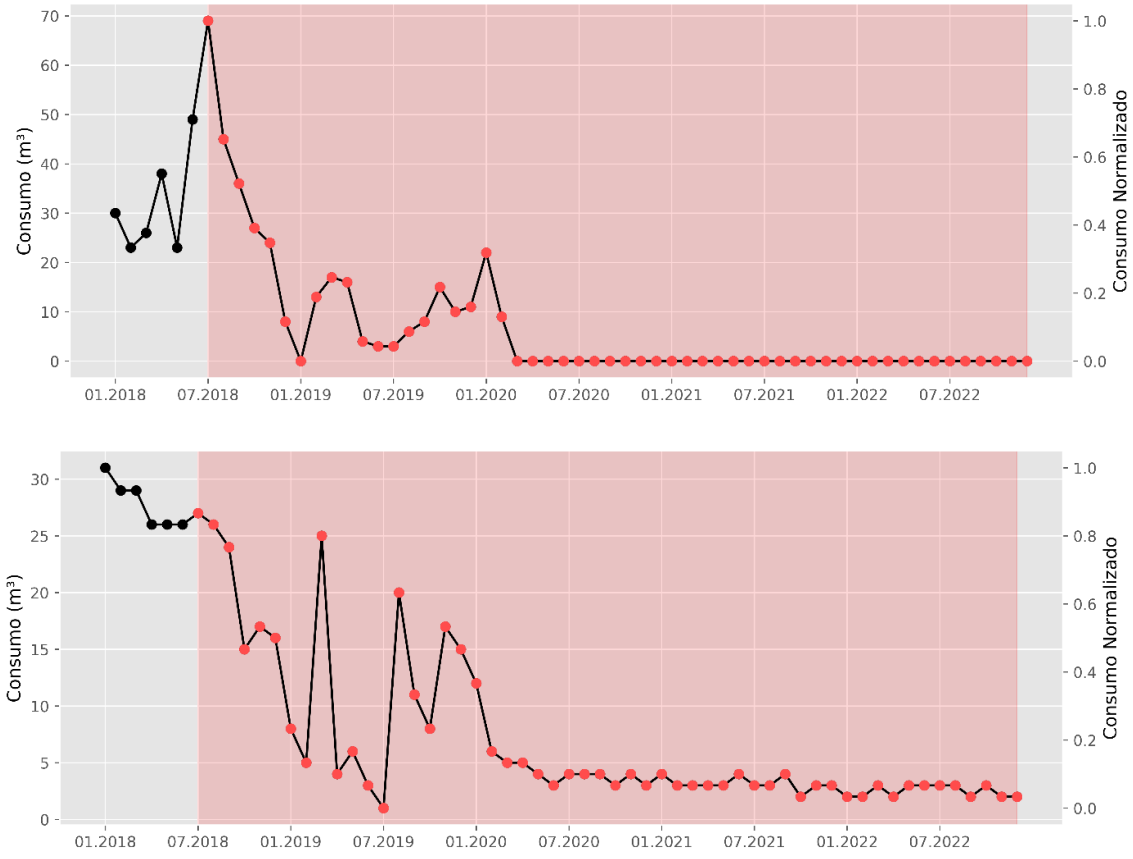
Embora tal comportamento possa refletir situações legítimas, como imóveis pouco habitados, desocupados ou utilizados apenas esporadicamente, ele também pode indicar sinais de submedição, perdas ou interferências nos dispositivos de medição. De modo geral, as concessionárias estabelecem um valor médio de consumo mensal com base em características regionais e culturais. Quando os dados observados apresentam consumos persistentemente inferiores a esse valor de referência, surgem indícios de irregularidade, especialmente na

ausência de comunicação oficial que justifique tal condição. Ademais, o uso de hidrômetros antigos ou descalibrados pode contribuir para a recorrência de leituras subestimadas ao longo do tempo (GHAMKHAR et al., 2023).

5.4.5 Queda gradual seguida de estabilização

O quinto padrão, apresentado na Figura 38, evidencia uma trajetória de queda gradual no consumo, a qual culmina em um novo patamar de estabilização. Essa transição é menos abrupta em comparação aos padrões anteriores, sugerindo uma alteração progressiva no perfil de uso.

Figura 38: Padrão 5 – queda gradual seguida de estabilização



Fonte: Autor (2025)

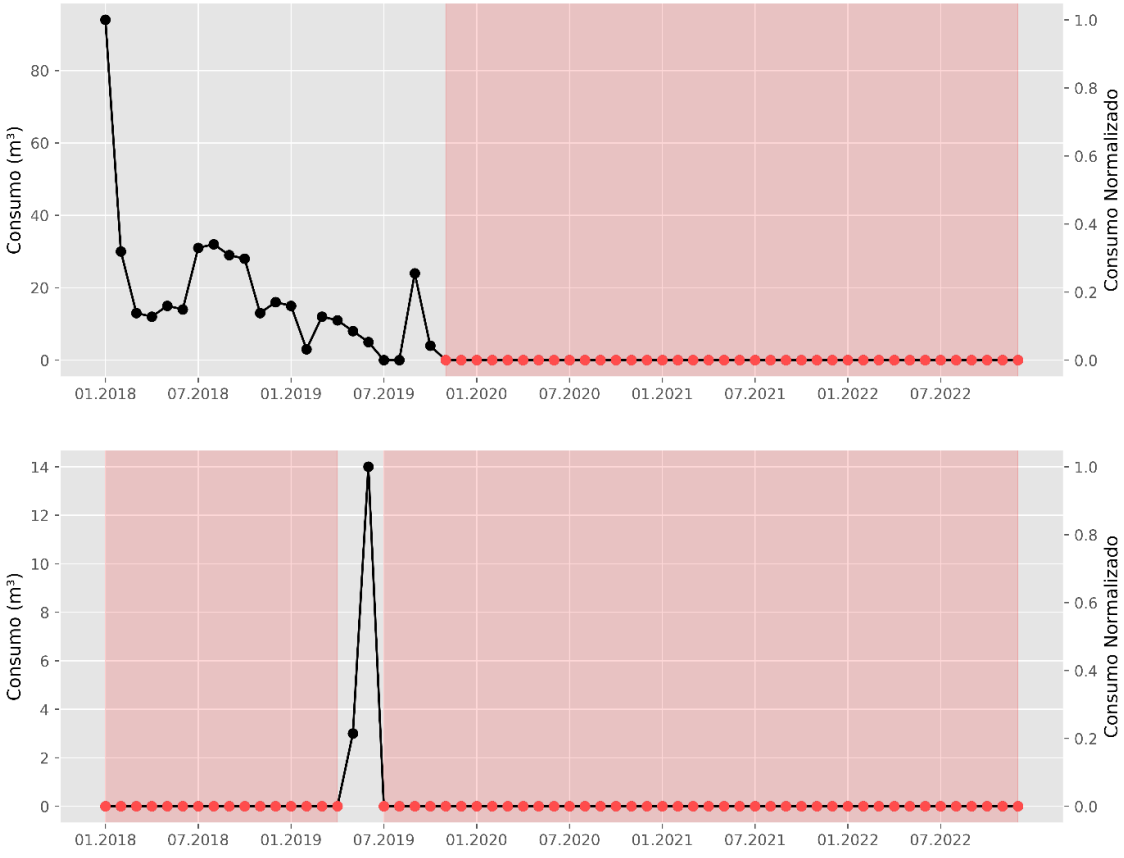
Do ponto de vista da detecção de eventos, as Figura 38 (a) e Figura 38 (b) revelam uma tendência descendente consistente, caracterizada por uma queda gradual no consumo, seguida de estabilização em níveis próximos ou iguais a zero, conforme destacado pelas áreas sombreadas. Diferentemente do Padrão 1, a transição neste caso ocorre de maneira mais sutil e distribuída ao longo de múltiplos períodos. Tal característica pode refletir tanto mudanças legítimas no comportamento do usuário quanto tentativas deliberadas de ocultar práticas fraudulentas.

Entre as hipóteses plausíveis para esse padrão, destacam-se o uso de hidrômetros antigos com desgaste progressivo, que passam a registrar volumes inferiores ao real, ou estratégias de fraude mais cautelosas, nas quais o usuário reduz gradualmente o consumo com o intuito de evitar alertas automáticos ou investigações. Essa redução sistemática pode preceder a instalação de desvios ou adulterações permanentes, tornando o processo menos perceptível à fiscalização (GHAMKHAR et al., 2023).

5.4.6 Registros com consumo zerado

O último padrão identificado, ilustrado na Figura 39, refere-se a séries temporais com longos intervalos de tempo apresentando registros contínuos de consumo igual a zero. Esse comportamento é característico de um grupo específico de usuários agrupados no chamado agrupamentos "zeros", cujos consumos, ao longo de meses consecutivos, permanecem invariavelmente nulos, como demonstrado na Figura 32. No entanto, a figura apresentada refere-se apenas às matrículas classificadas como ruído. Isso ocorreu porque o Conjunto 1 não conseguiu formar esse agrupamento específico, resultando na exclusão dessas matrículas de qualquer agrupamento definido.

Figura 39: Padrão 6 – consumos zerados



Fonte: Autor (2025)

As Figura 39 (a) e Figura 39 (b) representam claramente a ausência de variabilidade nas séries temporais em um certo período, com registros zerados de forma sistemática. As áreas sombreadas marcam os períodos onde essa condição se mantém. Sua identificação é essencial para que medidas corretivas e vistorias em campo sejam priorizadas, visando restaurar a confiabilidade do sistema de medição.

Embora esse padrão possa, em certos casos, estar associado à desocupação temporária do imóvel, desligamentos não formalizados ou situações operacionais legítimas, sua persistência ao longo de períodos prolongados levanta suspeitas sobre o real funcionamento do sistema de medição, indicando consumo não autorizado, adulteração do medidor ou desvios na calibração do equipamento. Essa situação classifica seu consumo como um processo fora de controle (GHAMKHAR et al., 2023).

5.5 Analise das matrículas classificadas como ruído

Buscando entender como falhas na medição e no registro de dados podem contribuir para a presença de ruídos, analisou-se a ocorrência de valores ausentes e nulos nas séries temporais de consumo, assim como a idade dos hidrômetros utilizados nas matrículas. Essa investigação permite avaliar até que ponto falhas na coleta de dados ou a degradação dos equipamentos podem estar contribuindo para a geração de padrões inconsistentes ou atípicos, impactando diretamente a performance dos modelos de detecção de irregularidades.

Posto isso, a Tabela 17 apresenta uma amostra representativa desses valores, na qual são demonstradas as médias mensais das porcentagens de inconsistências das matrículas classificadas como ruído.

Tabela 17: Proporção de valores ausentes e zeros das matrículas classificadas como ruídos.

	01.2018	02.2018	03.2018	04.2018	05.2018	06.2018	07.2018	08.2018	09.2018	10.2018	11.2018	12.2018	01.2019	02.2019	03.2019	04.2019
Porcentagem 0	11.4%	11.8%	12.2%	12.7%	12.1%	12.2%	11.1%	10.3%	10.3%	10.7%	10.7%	11.4%	12.2%	13.2%	13.6%	13.9%
Porcentagem NaN	13.7%	13.6%	13.2%	12.6%	12.1%	12.0%	12.4%	12.6%	13.0%	13.1%	12.8%	11.5%	10.0%	10.0%	11.7%	11.6%
Total	25.1%	25.4%	25.4%	25.3%	24.2%	24.2%	23.5%	22.8%	23.3%	23.7%	23.5%	22.9%	22.2%	23.1%	25.3%	25.5%

Fonte: Autor (2025)

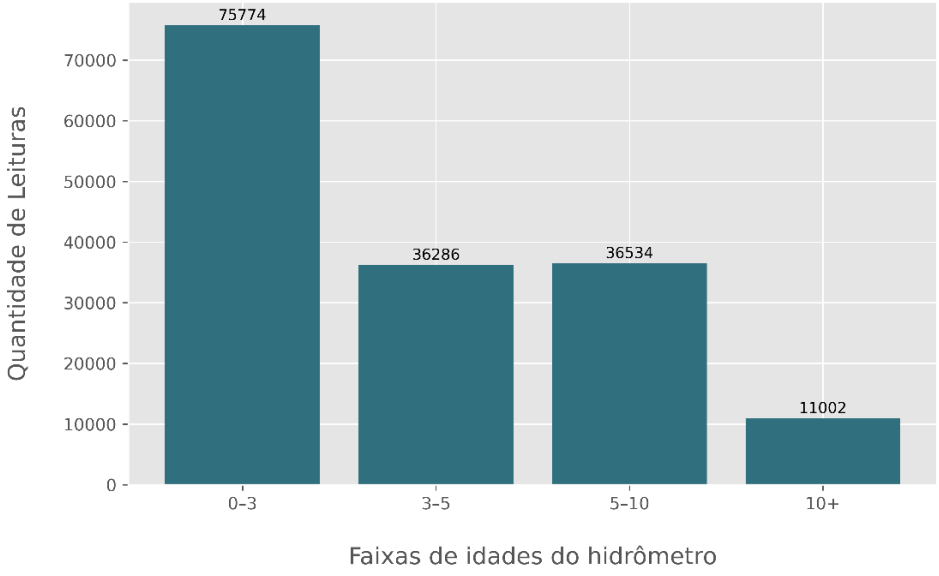
A análise dos dados revela variações nas proporções de valores nulos e zerados ao longo dos meses. Em média, os valores ausentes corresponderam a aproximadamente 11,5% por mês, enquanto os valores iguais a zero oscilaram em torno de 12%, totalizando 23,5%. Esse percentual indica que quase um quarto dos registros mensais de consumo apresenta inconsistências.

Ao comparar as proporções de valores ausentes e zerados das matrículas classificadas como ruído com aquelas obtidas após o processo de unificação, conforme apresentado na Tabela 9, observa-se uma leve redução de 2,5% nos valores nulos, enquanto os valores ausentes apresentaram um aumento de 5,7%, praticamente o dobro.

A variação nas proporções de valores nulos e ausentes entre os dados brutos e as matrículas classificadas como ruído pode ser explicada, em parte, pela realocação de registros inconsistentes para o agrupamento "zeros", evitando sua classificação direta como ruído e, assim, reduzindo os valores nulos. Por outro lado, o aumento na proporção de valores ausentes entre as matrículas classificadas como ruído evidencia que esse tipo de inconsistência exerce influência significativa na identificação desses casos.

Com relação à idade dos hidrômetros das matrículas classificadas como ruído, as idades de cada medição foram agrupadas em faixas, conforme ilustrado na Figura 40.

Figura 40: Faixas de Idades dos Hidrômetros das matrículas classificadas como ruídos.



Fonte: Autor (2025)

Observa-se que as medições realizadas com hidrômetros de até 5 anos correspondem a aproximadamente 70% das matrículas classificadas como ruído, enquanto aquelas com mais de 5 anos representam cerca de 30%. Ao comparar esses dados com os apresentados na Figura 21, verifica-se que os valores são bastante semelhantes, com variações pouco significativas. Com isso, é possível concluir que a idade dos hidrômetros, isoladamente, não exerce influência determinante na classificação de uma matrícula como ruído, sugerindo que outros fatores, como falhas operacionais ou inconsistências nos dados de consumo, desempenham um papel mais relevante nesse processo.

5.6 Comparação com registros de suspeitas

Com o intuito de avaliar a coerência dos resultados obtidos pelo modelo proposto neste trabalho, realizou-se uma comparação com um conjunto de registros fornecido pela Enops. Essa planilha, contém cerca de 2.424 matrículas identificadas como suspeitas com base em métodos estatísticos aplicados internamente pela empresa.

É importante destacar que os dados fornecidos não representam fraudes confirmadas, mas sim irregularidades detectadas por análises estatísticas que consideram desvios de comportamento no consumo, tendências incomuns e outros indicadores operacionais. De forma similar, o modelo desenvolvido neste trabalho, baseado no algoritmo DBSCAN e complementado por métricas como o TSLF e o LZC, também tem como objetivo identificar padrões atípicos nas séries temporais de consumo, sem atribuir rótulo definitivo de fraude.

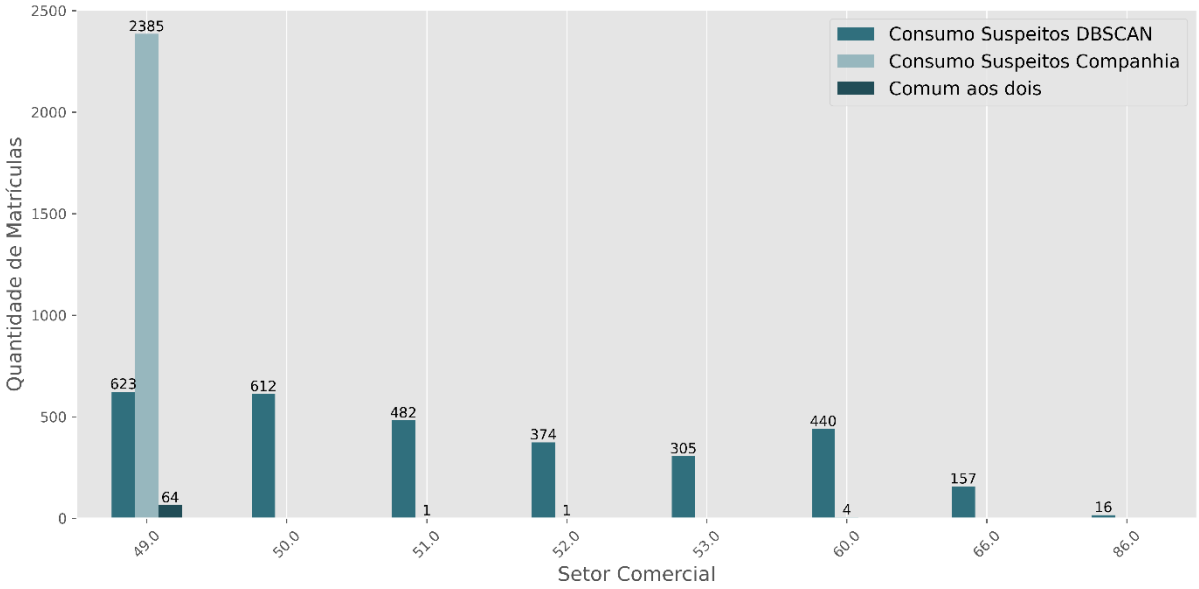
Ao aplicar o modelo ao banco de dados e isolar os registros classificados como irregulares, foram identificadas aproximadamente 3.009 matrículas com comportamento discrepante, além das matrículas pertencentes aos agrupamentos formados no Conjunto 2, que também devem ser analisadas por apresentarem padrões incomuns, ainda que densamente agrupados. Esses casos, embora não rotulados como ruído, podem representar desvios em relação ao consumo considerado normal pela concessionária.

A comparação entre os dois registros de suspeitas, revelou uma interseção de 64 matrículas, revelando uma coincidência relativamente pequena frente ao total de detecções. Para investigar com maior profundidade essa baixa sobreposição, foi realizada uma análise detalhada da distribuição das matrículas em relação aos agrupamentos identificados pelo DBSCAN e aos setores de abastecimento.

No que se refere à análise dos agrupamentos, observou-se que 2.282 das matrículas apontadas como suspeitas pela concessionária estavam inseridas nos agrupamentos classificados como "normais" pelo modelo, enquanto o restante foi distribuído entre agrupamentos periféricos ou rotulado como ruído.

Adicionalmente, para a análise segmentada por setor de abastecimento, foi elaborada a Figura 41, a qual ilustra a distribuição das possíveis fraudes entre os diferentes setores comerciais, com o objetivo de proporcionar uma visualização mais clara da dispersão das matrículas classificadas como suspeitas nos registros identificados pelo modelo proposto e os fornecidos pela concessionária.

Figura 41: Distribuição de matrículas anômalas por setor comercial nos dois registros



Fonte: Autor (2025)

Analisando a Figura 41, nota-se que os consumos suspeitos gerados pelo algoritmo, tem uma distribuição mais homogênea em relação aos setores comerciais, com destaque para o setor 49, que concentra 623 das 3.009 matrículas anômalas. Por outro lado, os registros fornecidos pela concessionária estão fortemente concentrados nesse mesmo setor, totalizando 2.385 matrículas, o que representa aproximadamente 15% do total de 16.036 matrículas do setor comercial, conforme indicado na Figura 30. Essa concentração sugere que a planilha fornecida pela companhia se refere exclusivamente ao setor 49, o que explica a ausência de suspeitas em outros setores. Ressalta-se ainda que a interseção entre os dois conjuntos ocorre integralmente no setor 49.

A identificação de aproximadamente 15% de matrículas suspeitas no setor 49 representa um valor expressivamente elevado, sobretudo considerando seu universo de mais de 16 mil unidades consumidoras, o que equivale a cerca de uma em cada sete matrículas apresentando indícios de comportamento atípico.

Análises anteriores mostraram que 2.282 dessas matrículas, sinalizadas pela concessionária, foram alocadas pelo DBSCAN no maior agrupamento, interpretado pelo modelo como comportamento "normal". Supondo que essas matrículas de fato representem fraudes, o resultado sugere que irregularidades recorrentes e padronizadas, por ocorrerem com alta densidade, podem ter sido absorvidas como parte do padrão dominante, escapando à classificação como ruído.

Como o DBSCAN define irregularidades com base na densidade de vizinhança (ε e MinPts), fraudes concentradas podem não ser detectadas com parâmetros conservadores. Nesse contexto, a redução do valor de ε no Conjunto 1 poderia aumentar a sensibilidade do modelo a variações internas aos agrupamentos, ampliando sua capacidade de identificar padrões irregulares concentrados, ainda que com maior risco de falsos positivos.

6 CONCLUSÕES

Este trabalho propôs a aplicação do método desenvolvido por Ghamkhar et al. (2023), que consiste na detecção de padrões anormais em sistemas de abastecimento de água por meio da análise do comportamento de consumo dos usuários.

Mesmo diante das limitações observadas na base de dados, o modelo demonstrou capacidade eficaz na identificação de padrões irregulares de consumo, classificando 3.009 matrículas como ruídos e segmentando agrupamentos com comportamentos compatíveis com possíveis irregularidades. Um dos principais diferenciais da abordagem adotada é sua flexibilidade, permitindo a calibragem dos parâmetros de detecção de acordo com diferentes contextos operacionais. Dessa forma, o modelo pode ser ajustado tanto para cenários com restrições orçamentárias — priorizando casos mais críticos — quanto para situações em que há maior disponibilidade de recursos, possibilitando a ampliação do número de inspeções em campo. Esses resultados evidenciam o potencial da abordagem para auxiliar concessionárias na priorização de inspeções em campo, na identificação de ligações suspeitas e na reavaliação de perfis de consumo que destoam significativamente dos padrões históricos.

A partir do modelo aplicado, foi possível identificar de forma clara um conjunto significativo de padrões irregulares de consumo de água, que se desviam do comportamento típico da maioria dos usuários. A análise das séries temporais rotuladas como ruído possibilitou a categorização de seis padrões distintos de irregulares, conforme a tipologia proposta por Ghamkhar et al. (2023):

1. Queda súbita seguida de estabilização
2. Aumento súbito seguido de estabilização
3. Picos pontuais de consumo
4. Consumo consistentemente baixo
5. Queda gradual seguida de estabilização
6. Consumo zerado

Esses padrões foram validados por meio de inspeção visual das séries temporais, concentrando-se, em sua maioria, em matrículas com alta incidência de inconsistências nos dados, como longos períodos sem leitura registrada ou medições zeradas, indicando possíveis falhas operacionais ou irregularidades no consumo.

A identificação e categorização dos padrões possibilitam que as concessionárias direcionem suas ações de forma mais estratégica, priorizando fiscalizações conforme o tipo de irregularidade e seu potencial impacto. Desse modo, os padrões detectados não apenas demonstram a eficácia do método adotado, mas também oferecem uma base sólida para o desenvolvimento de sistemas automatizados de monitoramento voltados à redução de perdas e ao aprimoramento da eficiência operacional.

Com isso, a detecção automatizada de comportamentos atípicos viabiliza a priorização de inspeções em campo e a aplicação de medidas corretivas de forma proativa, otimizando a alocação de recursos humanos e financeiros. Essa abordagem contribui diretamente para o fortalecimento da gestão da distribuição de água, permitindo a identificação precoce de fraudes, falhas de medição e demais irregularidades, mesmo em cenários com dados de baixa resolução (Ghamkhar et al., 2023).

Por fim, constatou-se que cerca de 34% dos hidrômetros possuíam mais de cinco anos de uso no momento da leitura, condição essa, não recomendada pelo INMETRO, em razão da imprecisão nas medições. Tal situação contribui significativamente para as perdas aparentes, uma vez que hidrômetros antigos tendem a subestimar o volume efetivamente consumido, conforme apontado por Fournioul (2004). Diante disso, reforça-se a importância da substituição periódica desses dispositivos como medida complementar às estratégias automatizadas de detecção de irregularidades, visando à melhoria da qualidade dos dados e à eficácia das ações de controle de perdas.

6.1 Limitações da pesquisa

Apesar dos resultados promissores obtidos com a aplicação do modelo proposto, esta pesquisa apresenta algumas limitações que impactaram diretamente a profundidade das análises e a validação dos achados.

A principal limitação diz respeito à ausência de dados rotulados no conjunto analisado. A falta de informações previamente classificadas como “normais” ou “anômalas” inviabilizou a validação quantitativa dos resultados obtidos pelo modelo, bem como o ajuste fino dos hiperparâmetros do algoritmo DBSCAN.

Outro fator que restringiu as análises foi a ausência de atributos relevantes na tabela fornecida pela concessionária. O banco de dados não incluía informações sobre o status das matrículas — se ativas, desligadas ou temporariamente inativas — nem sobre sua categoria

(residencial, comercial, industrial etc.). Essas informações teriam permitido segmentações mais precisas e contextualizadas dos padrões de consumo, além de facilitar a identificação de comportamentos irregulares em grupos específicos.

Por fim, não foi possível estimar os impactos relacionados à sustentabilidade a partir dos padrões irregulares identificados. Isso ocorreu porque, diferentemente do esperado, os consumos classificados como irregulares apresentaram, em média, volumes superiores aos dos consumos considerados normais. Tal resultado inviabilizou a aplicação direta da abordagem de Ghamkhar et al. (2023) para estimativas de economia de água ou redução de perdas, limitando, assim, uma avaliação mais aprofundada sobre o potencial do método em contribuir para a sustentabilidade dos sistemas de abastecimento.

REFERÊNCIAS

AMINIKHANGHAHI, Samaneh; COOK, Diane J. *A Survey of Methods for Time Series Change Point Detection*. Washington State University, Pullman, WA. 08 de setembro de 2016. Disponível em: < A survey of methods for time series change point detection | Knowledge and Information Systems>. Acesso em: 25/02/2025.

BASSEVILE, Michèle; NIKIFOROV, Igor. **Detection of Abrupt Changes: Theory and Application.** Prentice-Hall, 1993

BEZERRA, Saulo de Tarso Marques; CHEUNG, Peter Batista. **Perdas de Água: Tecnologias de Controle**. João Pessoa: Editora da UFPB, 2013.

CARVALHO, Caroline. **O que é Python? — um guia completo para iniciar nessa linguagem de programação**. Alura, 4 de setembro de 2024. Disponível em:< O que é Python: um guia para iniciar nessa linguagem | Alura>. Acesso em: 25/10/2024.

CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA DO RIO DE JANEIRO. **Detecção de eventos.** Cefet-RJ**.** Disponível em: < Detecção de eventos – DAL >. Acesso em: 25/02/ 2025.

CHANDOLA, Varun; BANERJEE, Arindam; KUMAR, Vipin. **Anomaly Detection : A Survey**. 30 de julho de 2009. Disponível em: <https://doi.org/10.1145/1541880.154188>. Acesso em: 25/02/2025.

COSTA, Mirla; PYRES, Leticia. **Direto ao ponto: o que é Machine Learning com exemplos reais**. 19 de janeiro de 2024. Disponível em: <O que é Machine Learning? | Alura>. Acesso em: 12/09/2024.

DATA SCIENCE ACADEMY. **As 6 Dimensões da Qualidade de Dados (Data Quality).** 26 de novembro de 2023. Disponível em:< As 6 Dimensões da Qualidade de Dados (Data Quality) - Data Science Academy>. Acesso em: 05/10/2024.

DNC. **Estatística Descritiva: o que é, como funciona e** exemplos. 15 de abril de 2024. Disponível em: < Estatística Descritiva: o que é, como funciona e exemplos - Blog DNC>. Acesso em: 20/01/2025.

ESCOVEDO, Tatiana; KOSHIYAMA, Adriano. **Introdução a Data Science — Algoritmos de Machine Learning e métodos de análise**. São Paulo, Ed. Casa do Código, 2020.

ESTER, Martin; KRIEGEL, Hans-Peter; SANDER, Jorg; XU, Xiaowei. **A density-based algorithm for discovering clusters in large spatial databases with noise**. 1996. Disponível em:< https://doi.org/ 10.5555/3001460.3001507>. Acesso em: 20/10/2024.

FARIA, Rayssa. **Entendendo Séries Temporais: Ferramentas para Tomada de Decisão.** Aquare.la, 5 de junho de 2024. Disponível em: < Entendendo Séries Temporais - Aquarela Analytics >. Acesso em: 24/02/2025.

FOURNIOL, M. 2004. Avaliação do Parque de Hidrômetros. Maceió, AL (apud BEZERRA e CHEUNG, 2013).

GAZETA DIGITAL**. Furtos tiram 23,4 milhões de litros de água de Cuiabá por ano**. *Gazeta Digital.* Cuiaba, 21 set. 2018. Disponível em: <Furtos tiram 23,4 milhões de litros de água de Cuiabá por ano | Gazeta Digital>. Acesso em: 10/01/2025

GÉRON, A. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow**. [S.l.]:Alta Books, 2019.

GHAMKHAR, Hani; GHAZIZADEH, Mohammadreza; MOHAJERI, Seyed; MOSLEHI, Iman; YOUSEFI, Ehsan. **An unsupervised method to exploit low-resolution water meter data for detecting end-users with abnormal consumption: Employing the DBSCAN and time series complexity.** *Sustainable Cities and Society*, *94.* 2023. Disponível em: <https://doi.org/10.1016/j.scs.2023.104516>. Acesso em: 01/08/2024.

GOMES, Pedro. **Conheça as Etapas do Pré-Processamento de dados.** DataGeeks, 13 de dezembro de 2019. Disponível em: <Pré-Processamento de Dados: Técnicas e Etapas de Tratamento> . Acesso em: 01/10/2024.

GOMES, Pedro. **Outliers: O que são Pontos Fora da Curva e Como Tratá-los?.** DataGeeks, 13 de dezembro de 2024. Disponível em: <Outliers: O que são Pontos Fora da Curva e Como Tratá-los?. Acesso em: 20/01/2025.

GOUVEIA, Cristiano. **TÉCNICAS DE APRENDIZADO DE MÁQUINA APLICADAS À PREDIÇÃO DE VAZAMENTOS EM RAMAIS DE REDES DE DISTRIBUIÇÃO DE ÁGUA**. Monografia – Mestrado em Tecnologia Ambiental e Recursos Hídricos, Universidade de Brasília. Brasília, p134.2022.

GUIMARAES, Nara. **Introdução às Séries Temporais.** Medium, 16 de junho de 2024. Disponível em: <Introdução às Séries Temporais. As séries temporais estão presentes em… | by Nara Guimaraes | Medium>. Acesso em: 25/10/2024

HAYES, Molly; DOWNIE, Amanda. **O que é transformação de dados?.** IBM, 19 de junho de 2024. Disponível em: < O que é transformação de dados? | IBM>. Acesso em: 10/01/2025.

IBM. **O que é overfitting?**. 2024. Disponível em: <O que é overfitting? | IBM>. Acesso em: 05/10/2024.

INSTITUTO TRATA BRASIL. ESTUDO DE PERDAS DE ÁGUA DO INSTITUTO TRATA BRASIL DE 2022 (SINISA 2022): **DESAFIOS NA EFICIÊNCIA DO SANEAMENTO BÁSICO NO BRASIL**. TRATABRASIL.ORG, 2024. Disponível em: <Estudo-da-GO-Associados-Perdas-de-Agua-de-2024-V2.pdf (tratabrasil.org.br) >. Acesso em 2024.

JUNIOR, Jose. **Data Cleaning.** 19 de jan. de 2021. Disponível em: < Data Cleaning | LinkedIn>. Acesso em: 20/01/2025.

LEMPEL, A.; ZIV, J. **On the complexity of finite sequences.** IEEE Transactions on Information Theory, 1976. Disponível em: <https://doi.org/10.1109/TIT.1976.1055501>. Acesso em 20/10/2024.

MAZZI, Gian Luigi (Ed.). **Handbook on Seasonal Adjustment**. Luxembourg: Publications Office of the European Union, 2018. ISBN 978-92-79-80170-9. Disponível em:< http://ec.europa.eu/eurostat/about/policies/copyright >. Acesso em: 25/02/2025.

MOHRI, Mehryar.; ROSTMIZADEH, Afshin.; TALWALKAR, Ameet. **Foundations of Machine Learning**, MIT Press, second edition, 2018.

MONTEIRO, Gabriel; CARL, Hugo. **Entendendo DBSCAN.** Medium, 23 de novembro de 2020. Disponível em:< Entendendo DBSCAN | por Gabriel Monteiro e Hugo Carl | Medium>. Acesso em: 25/10/2024

MORETTIN, P. A.; TOLOI, C. M. Análise de séries temporais, 2ª edição rev. e ampl. São Paulo: Edgard Blucher, 2006.

MUREL, Jacob; KAVLAKOGLU, Eda. **O que é redução da dimensionalidade?**. IBM, 05 de janeiro de 2024. Disponível em: <O que é redução da dimensionalidade? | IBM>. Acesso em: 01/10/2024.

NAÇOES UNIDAS. Água. **Nações Unidas,** 2021. Disponível em: < Água - Nações Unidas - ONU Portugal> . Acesso em: 01/10/2024.

PENTEADO, Karoline. **Um pouco sobre séries temporais e suas aplicações**. Alura, 17 de julho de 2024. Disponível em:< Um pouco sobre séries temporais e suas aplicações | Alura>. Acesso em: 26/10/2024.

PEREIRA, Guilherme. **Tudo que você precisa saber sobre Python para Análise de dados**. Cubos Academy, 14 de novembro 2023. Disponível em: <Tudo que você precisa saber: Python para Análise de dados>. Acesso em: 25/10/2024.

PINHEIRO, Nina. **Pré-processamento de dados com Python.** Data Hackers, 07 de agosto de 2021. Disponível em:< Pré-processamento de dados com Python | by Nina Maria Pinheiro | Data Hackers | Medium>. Acesso em: 01/10/2024

PROVOST, Foster; FAWCETT, Tom. **Data Science para Negócios: O que Você Precisa Saber Sobre Mineração de Dados e Pensamento Analítico de Dados**. Rio de Janeiro: Editora Alta Books, 2016.

SANDER, Jorg; ESTER, Martin; KRIEGEL, Hans-Pet; XU, Xiaowei. **Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications.** Data Mining and Knowledge Discovery, 2, 169–194. 1998. Disponível em: < https://doi.org/10.1023/A:1009745219419>. Acesso em: 20/10/2024.

SCHUBERT, Erich; SANDER, Jorg; ESTER, Martin; KRIEGEL, Hans-Peter; XU, Xiaowei. **DBSCAN Revisited, Revisited: WhyandHowYouShould(Still) Use DBSCAN**. ACM Transactions on Database Systems, 19 de julho de 2017. Disponível em: < https://doi.org/10.1145/3068335>. Acesso em: 01/10/2024.

SILVA, Kairo. **ANÁLISE ESTATÍSTICA PARA DETECÇÃO DE CONSUMO NÃO AUTORIZADO NA MICROMEDIÇÃO DO SISTEMA DE ABASTECIMENTO HÍDRICO DO MUNICÍPIO DE JOINVILLE.** Monografia – Mestrado em Engenharia Civil, Universidade do Estado de Santa Catarina. Joinville, p151.2021.

SRINIVANSAN, Srivatsan; MAHARAJ, Sahir; RICCIO, Donato. **Quais são as melhores maneiras de lidar com inconsistências de dados na limpeza de dados para Ciência de Dados?**. 10 de junho de 2024. Disponível em: < Quais são as melhores maneiras de lidar com inconsistências de dados na limpeza de dados para Ciência de Dados?> Acesso em: 10/01/2025.

TSUTIYA, Milton Tomoyuki. **Abastecimento de Água**. 3ª edição. Departamento de Engenharia Hidráulica e Sanitária da Escola Politécnica da Universidade de São Paulo. 643p. São Paulo, 2006.

VANDERLEI, João. **ANÁLISE DE DADOS COMERCIAIS PARA GESTÃO DE PERDAS DE ÁGUA E DETECÇÃO DE FRAUDES NO MUNICÍPIO DE VITÓRIA DE SANTO ANTÃO.** Monografia – Faculdade de Engenharia Civil, Universidade federal de alagoas. Maceió, p.91.2023.

VIGNOLI, Luciana. **ANALISE COMPARATIVA DE METODOS PARA DETECCAO DE EVENTOS EM SERIES TEMPORAIS.** Dissertação – Programa de Pós-graduação em Ciência da Computação – Centro Federal de Educação Tecnológica Celso Suckow da Fonseca. Rio de Janeiro, p.89.2021.