UNIVERSIDADE FEDERAL DE ALAGOAS

CENTRO DE TECNOLOGIA

ENGENHARIA CIVIL

JOÃO GABRIEL ROCHA VANDERLEI

ANÁLISE DE DADOS COMERCIAIS PARA GESTÃO DE PERDAS DE ÁGUA E DETECÇÃO DE FRAUDES NO MUNICÍPIO DE VITÓRIA DE SANTO ANTÃO

Maceió

2023

JOÃO GABRIEL ROCHA VANDERLEI

ANÁLISE DE DADOS COMERCIAIS PARA GESTÃO DE PERDAS DE ÁGUA E DETECÇÃO DE FRAUDES NO MUNICÍPIO DE VITÓRIA DE SANTO ANTÃO

Trabalho de Conclusão de Curso apresentado ao curso de Engenharia Civil, do Centro de Tecnologia da Universidade Federal de Alagoas, como requisito parcial para obtenção do título de Bacharel em Engenharia Civil.

Orientador: Professor Dr. Marllus Gustavo Ferreira Passos das Neves

Maceió

2023

Catalogação na Fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecário: Marcelino de Carvalho Freitas Neto – CRB-4 – 1767

V235a Vanderlei, João Gabriel Rocha.
Análise de dados comerciais para gestão de perdas de água e detecção de fraudes no município de Vitória de Santo Antão / João Gabriel Rocha Vanderlei. – Maceió, 2023.
91 f. : il., grafs. e tabs. color.

Orientador: Marllus Gustavo Ferreira Passos das Neves.
Monografia (Trabalho de conclusão de curso em Engenharia Civil) – Universidade Federal de Alagoas. Centro de Tecnologia. Maceió, 2023.

Bibliografia: f. 89-90.
Anexos: f. 91.

1. Abastecimento de água. 2. Perdas de água. 3. Aprendizagem de máquina. I. Título.

CDU: 628.1

AGRADECIMENTOS

Dedico este trabalho de conclusão de curso, bem como todo o meu percurso académico até aqui no curso de Engenharia Civil, e expresso os meus sinceros agradecimentos:

Primeiramente, gostaria de expressar minha profunda gratidão aos meus pais, José Elder e Josefa Elzany, que, desde o início da minha trajetória de estudante, batalharam e se dedicaram ao máximo para que eu tivesse as condições necessárias para seguir os meus objetivos e alcançar os meus sonhos. Desde sempre foi fornecido apoio constante em todas as situações e momentos em que precisei, guiando-me diariamente no caminho que gostaria de seguir.

A minha irmã, Bárbara, que está sempre próxima e me anima todos os dias, demonstrando o amor de irmã sincero e verdadeiro. Também apoiando em todos os momentos necessários, principalmente no presente momento em que estudamos na mesma universidade, com a troca de experiências e apoio mútuo nas mais diversas situações.

À minha namorada, Brenda, por seu constante apoio e incentivo desde que estamos juntos. Ela tem sido minha fonte de conforto e motivação durante os momentos difíceis, ouvindo minhas reclamações e sempre me motivando. Sua companhia em minha vida tem sido um ponto de recarga, renovando as minhas energias para enfrentar a rotina do dia a dia.

Aos meus amigos de curso, por estarmos juntos nos momentos bons até os mais difíceis durante a trajetória de toda a graduação, tornando a rotina na Universidade mais acolhedora. Sou grato por cada risada, cada conversa e cada memória que construímos ao longo desses anos. Em especial à Débora e Heitor, por toda a parceria durante o curso e apoio no desenvolvimento deste trabalho.

Ao meu orientador Marllus Gustavo e à coorientadora Daysy, pelo encorajamento e direcionamento ao longo deste trabalho. A orientação foi fundamental para o desenvolvimento de todo o Trabalho de Conclusão de Curso. Ao meu coordenador D'Angellis e a Giovanna por toda a paciência, dedicação e direcionamento na área no saneamento.

Agradeço à COMPESA, em especial a Rodolfo e Flávio, pela autorização no uso dos dados para a elaboração deste trabalho.

Aos meus colegas de estágio, Alexandra, Alexandre, Luiz, Pedro e John, que compartilham comigo o dia a dia de escritório. Além de contribuir para tornar a jornada de mais agradável e produtiva, a constante troca de ideias agregou tanto para o meu crescimento quanto para o desenvolvimento desse trabalho.

À Universidade Federal e Alagoas e ao Centro de Tecnologia por fornecerem um ensino de qualidade, gratuito e acessível, correspondendo a boa parte da minha rotina e me acolhendo como uma segunda casa durante o tempo de graduação.

Por último, quero agradecer a todo e qualquer tipo de apoio de qualquer pessoa que me ajudou a trilhar o caminho percorrido até os dias atuais.

RESUMO

Devido ao progressivo aumento do índice de perdas nos sistemas de abastecimento de água (SAA) no Brasil, as perdas aparentes tem se tornado cada vez mais impactantes no desempenho e faturamento das concessionárias vigentes. As perdas aparentes, em síntese, englobam o volume consumido, mas não registrado pela companhia prestadora do serviço sem ser faturada – causada por erros de leitura nos hidrômetros, fraudes, ligações clandestina e falhas no cadastro comercial. Com isso, faz-se importante a realização de diagnósticos no setor comercial com o objetivo de identificar, controlar e reduzir tais empecilhos. Para tanto, analisou-se o banco de dados do cadastro comercial da cidade de Vitória de Santo Antão, através dos dados disponibilizados pela Companhia Pernambucana de Saneamento (COMPESA), realizando uma síntese comercial do município a partir da estratificação dos dados disponíveis das ligações existentes. Além disso, avaliou-se, também, os dados mensais de medição e variação de consumo de água por categoria de ligação, vida útil atual do parque de hidrômetros e faixas de consumo nas economias. Por conseguinte, ressaltado o fato de que as fraudes são atos criminosos e que devem ser combatidos, destacou-se a importância de analisar as anormalidades nas ligações referentes ao consumo de água. Dessa forma, foi evidenciada a usabilidade de algoritmos de aprendizagem de máquina como alternativa para criação de modelos para identificação de possíveis fraudes derivadas das anormalidades de consumo de água.

Palavras-chave: Sistemas de abastecimento de água; Perdas aparentes; Aprendizagem de máquina.

ABSTRACT

Due to the progressive increase in the rate of losses in water supply systems (SAA) in Brazil, apparent losses have become increasingly impacting on the performance and billing of current concessionaires. Apparent losses, in short, encompass the volume consumed, but not recorded by the company providing the service without being billed - caused by reading errors in water meters, fraud, clandestine connections and failures in the commercial register. Thus, it is important to carry out diagnostics in the commercial sector in order to identify, control and reduce such obstacles. Therefore, we seek to analyze the commercial registry database of the city of Vitória de Santo Antão, through the data provided by Companhia Pernambucana de Saneamento (COMPESA), performing a commercial synthesis of the city from the stratification of the available data of the existing connections. In addition, the monthly measurement data and variation of water consumption by connection category, current useful life of the water meter park and consumption ranges in the economies are also evaluated. Therefore, highlighting the fact that frauds are criminal acts and that they must be fought, the importance of analyzing the abnormalities in the connections referring to water consumption is highlighted. Thus, the usability of machine learning algorithms stands out as an alternative for creating models to identify possible fraud arising from water consumption abnormalities.

Keywords: Water supply systems; Apparent losses; Machine learning.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

ABNT	Associação Brasileira de Normas Técnicas
COMPESA	Companhia Pernambucana de Saneamento
DMC	Distrito de Medição e Controle
F	Faturamento
FN	Falso Negativo
FP	Falso Positivo
HD	Hidrômetro
IDM	Índice de Desemprenho da Medição
IQR	Amplitude Interquartil
IWA	*International Water Association*
KNN	*k-Nearest Neighbors*
NAN	*Not a Number*
PPP	Parceria Público-Privada
Q1	Primeiro Quartil
Q3	Terceiro Quartil
SAA	Sistema de Abastecimento de Água
SNIS	Sistema Nacional de Informações sobre Saneamento
TN	Verdadeiro Negativo
TP	Verdadeiro Positivo
UFAL	Universidade Federal de Alagoas
VC	Volume Consumido
VF	Volume Faturado

SUMÁRIO

1 INTRODUÇÃO

O ser humano desde os primórdios sempre dependeu da água com quantidade e qualidades adequadas para o seu desenvolvimento e sobrevivência – constituindo um recurso indispensável no que se refere à vida animal e vegetal. Concomitantemente a isso, a concepção de sistemas de abastecimento de água tornou-se equitativamente essencial, designados ao atendimento de povoados ou grandes municípios com o objetivo de produzir e distribuir água a uma determinada população, atendendo, na maioria das vezes, o padrão de qualidade, transporte e fornecimento após a sua retirada da natureza (BRASIL, 2006).

É de senso comum a recorrente problemática relacionada às perdas de água nos sistemas públicos de abastecimento de água em regiões urbanas no Brasil, caracterizando-se, hodiernamente, como um dos principais desafios das companhias vigentes no país, que, quando negligenciado, prejudica a eficiência operacional, afeta a distribuição de água e desequilibra a oferta e a demanda, comprometendo, assim, a constância do fornecimento da água de qualidade e, consequentemente, a universalização almejada pelos administradores públicos que tem como objetivo atender a demanda da população crescente.

No Brasil, de acordo com o último diagnóstico de água e esgoto do Sistema Nacional de Informação sobre Saneamento (SNIS), os índices de perda na distribuição de água têm sofrido um aumento constante desde 2015 até chegar ao valor de 40,1% no ano de 2020. Este índice influencia no uso da água de cada brasileiro atendido pelas diversas companhias existentes no país, agravado pelo crescimento desordenado sem planejamento das cidades e ausência de manutenção dos sistemas de abastecimento implantados (SNIS, 2020).

Por outro lado, tem-se que da captação até a distribuição da água tratada para o consumidor sempre ocorrerão perdas – destacando-se as decorrentes de erros operacionais, gestão ineficiente das companhias e ausência de manutenção nas tubulações do sistema. Dado o exposto, as companhias de saneamento distinguem as perdas em real e aparente (TSUTIYA, 2006).

Em virtude da problemática apresentada, este trabalho busca a realizar uma análise dos dados comerciais de consumo de uma companhia real, verificando os fatores que influenciam na redução da perda aparente e suas anormalidades associadas.

2 OBJETIVOS

O objetivo geral é realizar uma análise dos dados comerciais, identificando as principais causas das perdas aparentes, aplicando as práticas de redução de perdas aparentes estudadas e desenvolver um modelo para identificação de fraudes.

2.1 Objetivos Específicos

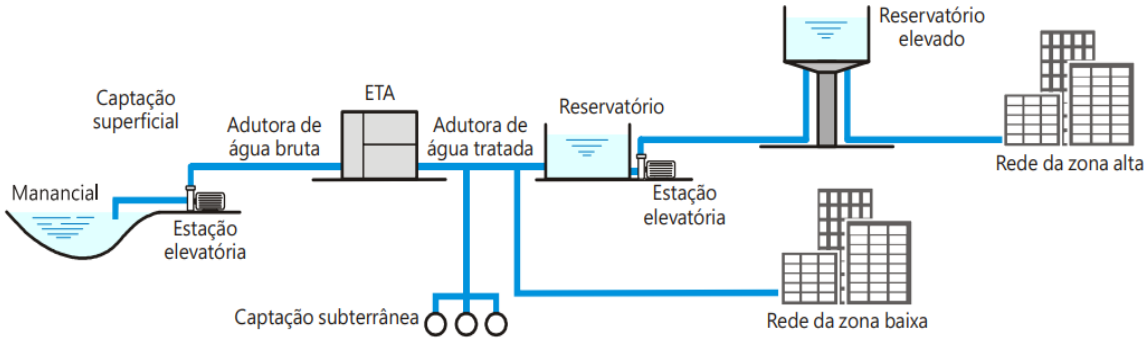
- Realizar uma síntese comercial do município a partir da estratificação dos dados disponíveis das ligações existentes para o diagnóstico das perdas aparentes;
- Analisar os dados mensais de medição e variação de consumo de água: por categoria de economia, vida útil atual do parque de hidrômetros e medição por faixa de consumo nas economias;
- Analisar anormalidades nas ligações no que se refere ao consumo de água para a proposição do método para identificação de fraudes;
- Avaliar a precisão de cada um dos modelos propostos no que se refere à capacidade de identificar as fraudes decorrentes das anormalidades de consumo de água.

3 FUNDAMENTAÇÃO TEÓRICA

3.1 Sistemas de abastecimento de água

Os Sistemas de abastecimento de água (SAA) podem ser descritos por um agrupamento de instalações, equipamentos, serviços e encargos que cujo desígnio final converge no transporte de água às diversas comunidades de uma região – zonas urbanas e rurais – priorizando a continuidade, segurança e potabilidade. Convenciona-se, normalmente, um sistema padrão composto pelas seguintes unidades respectivamente ordenadas: captação, estação elevatória, estação de tratamento de água, adução de água, reservarão e setor de distribuição (BEZERRA e CHEUNG, 2013).

Figura 1 – Componentes de um sistema de abastecimento.



Fonte: (BEZERRA e CHEUNG, 2013).

Tem-se como requisito para que um SAA atenda às necessidades da população a proteção de seus consumidores no que se refere à saúde humana vinculada a água, priorizando sempre a eficiência e a sustentabilidade. Dessa forma, a entidade gestora responsável, da governança até as fases operacionais, deve buscar compromisso com os objetivos de desenvolvimento sustentável (ODS) em todos os seus setores e consumidores. Sendo assim, qualquer que seja a companhia, sendo pública ou privada, deve entregar, essencialmente, como produto para seus clientes uma água com quantidade, qualidade, pressão e constância ininterrupta (BEZERRA e CHEUNG, 2013).

Para a concepção de um SAA, é de suma importância a realização de diversos estudos de caracterização do local de implantação, análises para o atendimento das diversas diretrizes e parâmetros para que atue de maneira eficiente em concerne com

o projeto desenvolvido. Sendo realizada durante a etapa incipiente, recomenda-se um estudo de concepção precedido de um diagnóstico técnico e ambiental ou, também, um Plano Diretor da área de atividade (TSUTIYA, 2006).

Na conjuntura atual, após a aprovação do novo marco legal do saneamento, houve um avanço no modelo de Parceria Público-Privada (PPP) que causou a atração em massa de investidores para o setor de abastecimento de água e saneamento no geral. Além disso, diversas empresas privadas têm surgido com um grande potencial para oferecer soluções adequadas a diferentes casos, oferecendo diagnósticos precisos e esperados para a concepção de SAA.

3.1.1 Balanço hídrico

A *International Water Association* (IWA) propôs um método relevante para classificar e padronizar a terminologia voltada para o estudo de perdas em SAA, o Balanço Hídrico, difundindo e amplamente utilizado na hodiernidade. Trata-se de uma aplicação Top-Down que avalia as perdas através do volume de entrada subtraído o volume de água de fato consumido (BEZERRA e CHEUNG, 2013).

Além disso, pode-se definir o balanço hídrico como um “balanço de massa” realizado a partir de dados anuais, tanto do setor comercial quanto operacional de determinado cadastro disponibilizado pela concessionaria estudada. A modelagem pode ser realizada para um setor de abastecimento, Distrito de Medição e Controle (DMC), ou, até mesmo, um SAA e suas partes operadas. Nessa conjuntura, necessita-se de um cálculo de perdas fundamentado para que viabilizem taxa de acertos confiável (AESBE, 2015).

Figura 2 – Balanço Hídrico proposto pela IWA para SAA

<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 rowspan="2">Consumo autorizado não faturado</td><td>Consumo não faturado medido (usos próprios, caminhão pipa etc.)</td><td rowspan="5">Água não faturada</td></tr>
<tr><td>Consumo não faturado não medido (combate a incêndios, favelas etc.)</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).

3.1.2 Perdas de água em sistemas de abastecimento

Visto que desde a captação até a distribuição de água sempre haverá perdas, as companhias buscam atingir um determinado nível de perdas considerado aceitável economicamente e operacionalmente. Concomitantemente, um SAA projetado corretamente e gerenciado de maneira adequada, demonstrará ótima performance, assim como um índice de perdas reduzido. Para a estimativas de perdas em um SAA, faz-se necessário o indicador percentual que relaciona o volume disponibilizado que será distribuído – volume macromedido –, com o volume obtido a partir da leitura dos hidrômetros – volume micromedido (BEZERRA e CHEUNG, 2013).

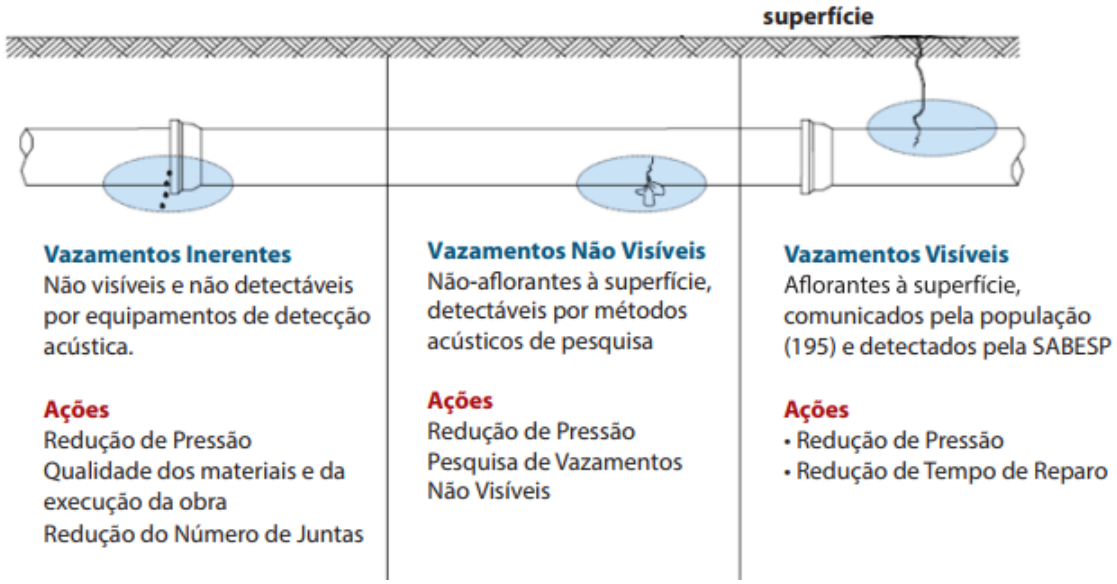
Uma vez amenizado o índice de perdas, tem-se que a companhia apresentará um aumento no faturamento e consequentemente na performance econômica que retornará positivamente para os consumidores em forma de redução de tarifas no consumo de água. (TSUTIYA, 2006).

Identificados os dois tipos de perdas distinguidos pelas companhias, faz-se preciso pontuar de início que as perdas reais e aparentes possuem origens díspares.

As perdas reais caracterizam-se pela perda do próprio produto, ou seja, a água efetivamente perdida, em qualquer um dos processos dentro do SAA, levando em

conta, até mesmo, sem estar tratada. Mantêm-se diretamente associadas ao custo operacional para a distribuição, tratamento e produção da água, ou seja, quanto maior for o desempenho concomitante à otimização, menor será o custo marginal da água. Em síntese, a perda real ou perda física pode ser definida como o volume de água que foi produzido, porém, devido a vazamentos e extravasamento de reservatórios, não contemplou o consumidor final que deveria ser abastecido (SNSA, 2018). É possível observar na Figura 3 os tipos de vazamentos e medidas combate as perdas reais.

Figura 3 – Tipos de vazamentos e ações de combate a perdas reais.

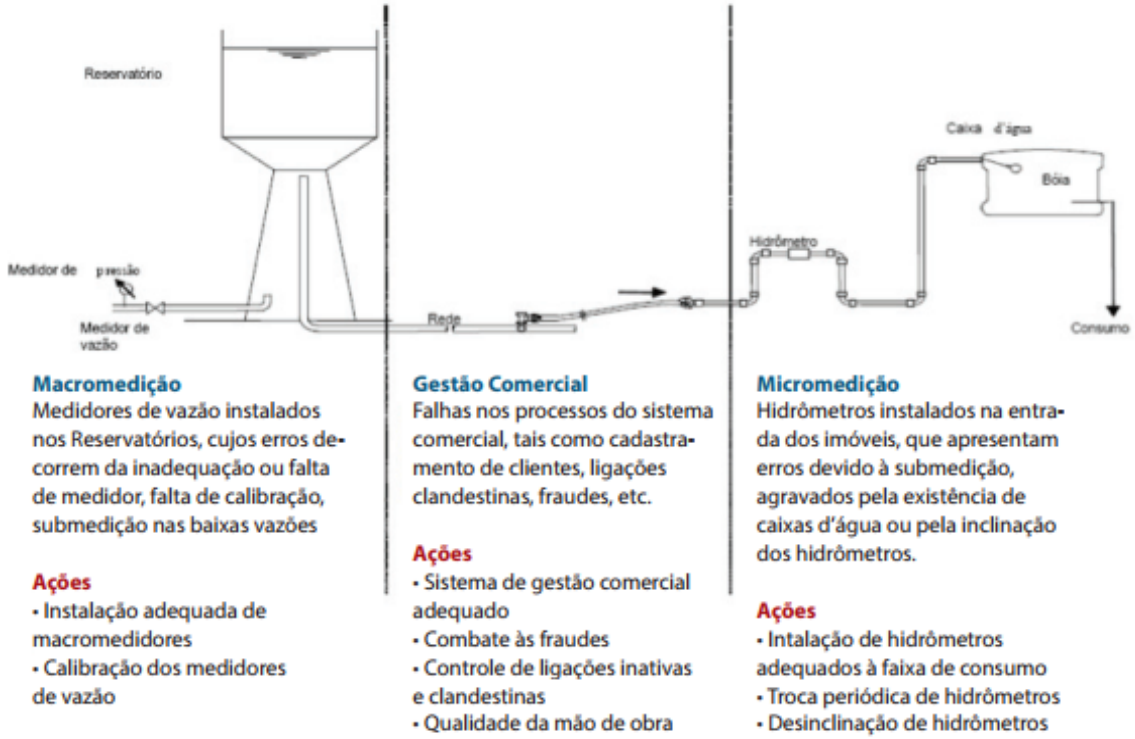


Fonte: (BÁGGIO, 2014).

Por outro lado, as perdas aparentes, conhecidas, também, por perdas não-físicas ou perdas comerciais, representem o volume de água consumido que não foi registrado pela companhia e os prestadores de serviço de saneamento. Dessa forma, a água é distribuída, entretanto não é faturada pela companhia, fato assíduo devido ao erro de leitura dos hidrômetros (HDs), fraudes, ligações clandestinas (*by-pass*) e falhas no cadastro comercial (registro de inatividade nas ligações efetivamente ativas). Essas perdas influenciam diretamente no faturamento da companhia, pois quanto mais eficiente e precisa a micromedição, assim como o cadastro comercial, maior será o faturamento como um todo. Além disso, possuem a mesma natureza de consumos e dependem pouco das pressões média – diferente das perdas reais -, principalmente

em sistemas de preservação intradomiciliares, cujas pressões a que estão os pontos de consumo são normalizadas (SNSA, 2018).

Figura 4 – Tipos e ações de combate a perdas aparentes



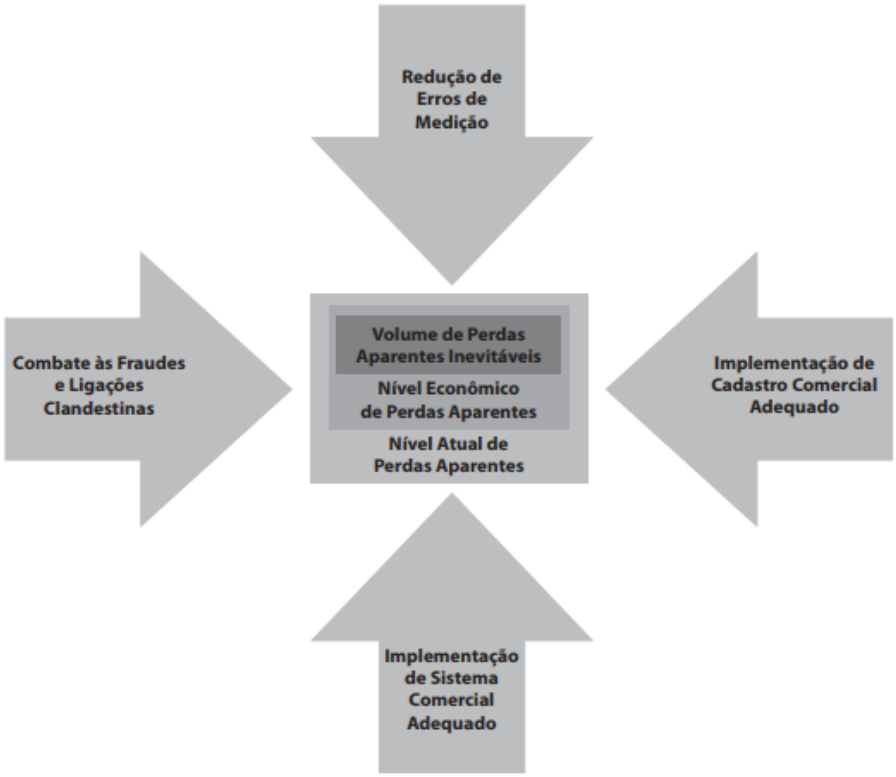
Fonte: (BÁGGIO, 2014).

3.1.3 Redução e controle de perdas aparentes

Primeiramente, deve-se estabelecer que a perda aparente zero é praticamente impossível de ser atingida em um SAA. Uma vez que existem perdas inevitáveis na medição do volume entregue ao consumidor pela companhia – advindo de erros ou imprecisão no cadastro comercial –, pelos erros dos medidores e, por fim, por fraudes nos HDs e ligações. Portanto, faz-se necessário trabalhar com um nível de perdas aparentes inevitáveis considerado aceitável (SNSA, 2018).

Segundo SNSA (2018), dado que a eficiência de uma companhia no setor supracitado possui uma relação intrínseca a gestão e desempenho das atividades do prestador de serviço, as perdas não físicas devem ser minimizadas a níveis coerentes as condições econômicas e específicas da região de atuação de maneira a atingir os valores admissíveis.

Figura 5 – Cruz de ataque às perdas aparentes.



Fonte: (BÁGGIO, 2014).

Nessa conjuntura, as perdas comerciais devem sempre ser avaliadas pelo valor de varejo da água (tarifa), devido ao fato de que, conforme a perda real, também ocasiona em um valor de retorno adicional que encarece o repasse para os consumidores contribuintes. Em síntese, as perdas comerciais podem ocorrer através de erros de medição e leitura, erros cadastrais ou na contabilização da água consumida e pelo consumo não autorizado.

Segundo Bezerra e Cheung (2013), dentre as principais causas, destacam-se: ligações clandestinas, roubo ou uso ilegal – como a utilização de água em válvulas de descarga de adutoras –, fraudes nos HDs (violado ou invertido), violação do lacre de proteção, erros de leituras nos HDs e falhas no cadastro comercial como um todo – causado pela desatualização, ligações não cadastradas ou ligações inativas cadastradas como ativas.

Nesse sentido, necessitam-se tecnologias adequadas de medição de água, educação ambiental, comunicação social, controle de fraudes e um sistema funcional de atualização cadastral para que haja a redução e controle do tipo específico de

perda em questão no Brasil – país dotado de disparidades sociais e crescimento urbano ausente de planejamento (SNSA, 2018).

Visto que as intervenções voltadas para a redução das perdas aparentes necessitam de um investimento consideravelmente menor que o voltado para redução de perdas reais e não requerem grandes manutenções ou reparos operacionais, tem-se que as propostas de intervenções de performance para esse setor apresentam grandes potenciais de incremento em volume consumido, faturado e faturamento de água.

Conforme Bezerra e Cheung (2013), tem-se como as principais medidas práticas de intervenção para o controle das perdas comerciais as seguintes alternativas:

- Melhoria na macromedição e micromedição do SAA;
- Adequação dos HDs de acordo com os perfis de consumo dos usuários, priorizando, principalmente, os grandes consumidores (consumo acima de 50 m^3/s);
- Fiscalização e correção na instalação dos HDs;
- Setorização da leitura dos HDs supervisionada por projetistas e operadores que prestam serviço à companhia;
- Substituição periódica e otimizada de hidrômetros levando em conta o desempenho da medição em função do tempo de instalação;
- Melhorias no procedimento de identificação e combate a fraudes em ligações ativas e inativas;
- Buscar 100% de hidrometração, ou seja, regularizar ligações que não possuem HDs;
- Tornar a gestão do sistema comercial mais eficientes através da realização de diagnósticos.

3.1.4 Erros de medição (submedição)

Segundo Bezerra e Cheung (2013), os erros de medição e faturação do consumo de água deve-se, sobretudo, à idade de conservação, dimensionamento e erro de leitura dos contadores, ou, até mesmo, devido à falta de acesso aos mesmos. Quanto à faturação, evidenciam-se os erros no registro e tratamento dos dados assim como na estimativa dos consumos.

Do mesmo modo, tem-se que a redução dos erros de medição é alcançada por meio de uma gestão eficaz do parque de HDs. A ausência dessa gerência desencadeia uma queda progressiva da receita da companhia e uma inviabilização no cálculo das perdas aparentes devido à falta de precisão. Concomitantemente, um SAA sem hidrometração ou com medições imprecisas que ocasionem uma incerteza nos dados obtidos não torna possível a realização de um Balanço Hídrico relevante. Portanto, visa-se a um diagnóstico comercial seguido de uma campanha de substituição periódica aliado a programas eficientes de manutenção e de dimensionamento adequado com a compra de contadores com certificados do Inmetro – Instituto Nacional de Metrologia, Qualidade e Tecnologia (BEZERRA e CHEUNG, 2013).

Salienta-se ainda que o hidrômetro é um aparelho que apresenta desgaste e imprecisão com o passar dos anos, havendo cerca de 50% de índice de submedição em HDs com mais de 10 anos de uso após a sua instalação, impactando fortemente a receita da companhia. Assim sendo, para alcançar o nível ótimo de perdas – cerca de 5% – prioriza-se a fiscalização tipo de medição no hidrômetro, idade do medidor do hidrômetro, permanência da exatidão ao longo do tempo, instalação do medidor (atentando-se a posição de montagem) e registro de ocorrências –paralisação, bloqueio, embaçamento e fraudes (BEZERRA e CHEUNG, 2013).

3.1.5 Hidrômetros (HDs)

Em uma companhia de abastecimento, tem-se que os medidores – conhecidos como macromedidores – são responsáveis por medir o volume de água oferecido aos seus consumidores. Além disso, todo o sistema de micromedição, incluindo suas respectivas funções, devem estar inseridos na conjuntura do setor comercial da empresa (COELHO, 2009).

Persiste a concepção de que o hidrômetro é um aparelho ideal de medida, porém a realidade não é essa. Trata-se de um medidor com limitação no registro exato do volume de água transpassado e é limitado por várias condições de utilizador – principalmente pelo tempo de utilização de suas peças a partir da data de instalação do mesmo (COELHO, 2009).

Outrossim, as perdas por imprecisão dos HDs contribuem para uma parcela significativa das perdas aparentes, tornando importante o papel da concessionária de

identificar, quantificar e resolver as suas causas, minimizando o impacto na medição de volume de água consumido.

Figura 6 – Hidrômetro cujo mostrador encontra-se ilegível

Fonte: (BEZERRA e CHEUNG, 2013).

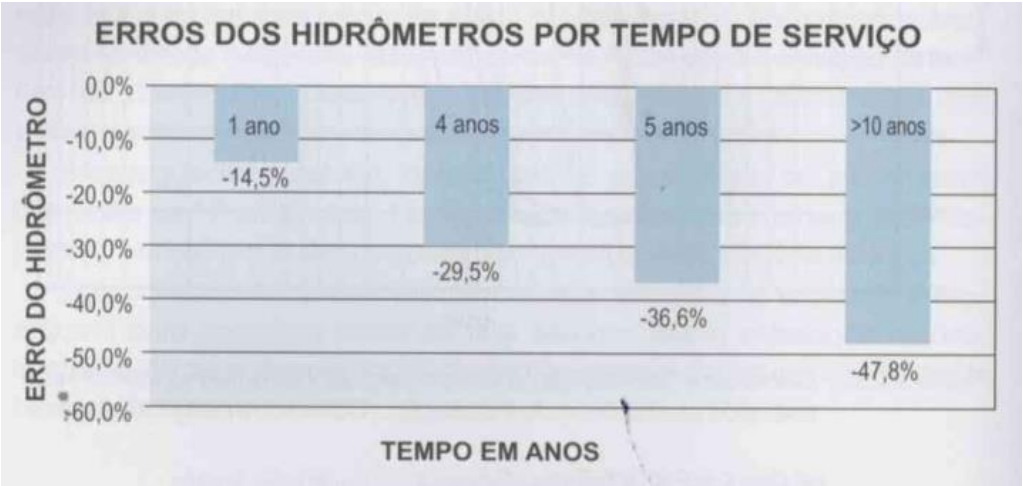
Segundo Coelho (2009), companhias que não possuem um programa sistemático de manutenção preventiva dos HDs apresentam perdas por submedição em torno de 10 a 30%. Apresentando, dessa forma, uma perda de faturamento de água anual gigantesca quando comparado aos valores que poderiam ser gastos em uma sistemática planejada de manutenção preventiva e corretiva para o parque de HDs manter-se confiável e eficiente.

Devido à disponibilidade quase inexistente de informação a partir de estudos realizados de casos reais no que se refere ao desempenho dos HDs instalados ao longo do tempo, faz-se importante ressaltar o estudo realizado por meio de ensaios de laboratório pelo engenheiro Maurício Fourniol (FOURNIOL, 2004). Identificou-se – nos HDs velocimétricos de 3 m³/h de vazão máxima estudados – 14,5% de perda média no primeiro ano e 47,8% nos HDs com 10 ou mais anos de instalação.

Sendo assim, todas as concessionárias, públicas ou privadas, devem manter, indubitavelmente, uma boa gestão no seu parque de hidrômetros para a manutenção dos volumes de submedição. Visto que no Brasil, opera-se, na maioria das vezes, através de reservatórios domiciliares, o volume submedido possui uma grande tendência de progressão. Além disso, os medidores velocimétricos de classe B,

amplamente difundidos no País, apresentam pouca eficiência em baixa vazão e tendem a perder boa parte da sua capacidade ao longo dos anos (AESBE, 2015).

Figura 7 – Erros de medição de hidrômetros domiciliares velocimétricos em uma rede de distribuição de Maceió - AL



Fonte: (BEZERRA e CHEUNG, 2013).

Para garantir a confiabilidade das medições e a justa cobrança do consumo de água, recomenda-se e, muitas concessionárias de serviços de abastecimento de água têm adotado, a prática da substituição de hidrômetros a partir de um intervalo de 5 anos, dentro do panorama da companhia (BRK AMBIENTAL, 2021).

3.1.6 Combate às fraudes e uso não autorizado

Indubitavelmente, as fraudes devem ser combatidas de maneira constante pela companhia de saneamento responsável, visto que qualquer sinal de descaso em relação à fiscalização, maior será a chance de surgir possíveis fraudadores. Faz-se importante ressaltar que as fraudes são atos criminosos passíveis de arcar com todos os seus respectivos procedimentos jurídicos e processuais de maneira a conceber todas as penalidades cabíveis em lei. Dessa forma, a educação voltada para a conscientização da população e consumidores deve ser vista como prioridade pela companhia (TSUTIYA, 2006).

Segundo SNSA (2018), enquadram-se como uso não autorizado:

- As ligações clandestinas realizando furto na rede de distribuição sem o usuário estar devidamente cadastrado no banco de dados comerciais da companhia;

- Fraudes nas ligações de clientes que já estão cadastrados no banco de dados, por meio de violação do hidrômetro ou ligação direta antes da leitura do contador;
- Falhas no cadastro comercial da companhia, destacando-se ligações de água cortadas ou suprimidas que estão enquadradas como inativas associado ao fato dos clientes não respeitarem a inibição do fornecimento previamente prestado. Além dos HDs e ligações ausentes de registro no banco de dados comercial.

É importante ressaltar que as ligações de caráter clandestino ocorrem, tanto em empreendimentos comerciais e de alto padrão, quanto áreas dotadas de um poder aquisitivo menor, ambas as situações são favorecidas nos casos em que a fiscalização é baixa ou inexistente.

Entretanto, há uma maior frequência em áreas cujas habitações são subnormais ausentes de um SAA formal, visto que, na maioria das vezes, trata-se de ocupações irregulares e áreas invadidas, não registradas ou ocupáveis por lei. Nos HDs, as irregularidades mais predominantes são as violações do medidor inviabilizando o seu funcionamento correto e, também, as inversões do próprio hidrômetro. Já nos ramais, a fraude mais corriqueira é o desvio de água antes do medidor em si (SNSA, 2018).

Em síntese, tem-se que os tipos de fraude e seus conceitos no que se referem a identificação, podem ser observados na Tabela 1.

Tabela 1 – Critérios para seleção de casos suspeitos de uso não autorizados

Tipos de Fraude	Conceitos
Não-ativos: ligações cortadas, suprimidas e suspensas	Ligações cortadas e suprimidas, quando não fiscalizadas, podem levar a religações indevidas
Código de fraudes originados na leitura	-Suspeita de hidrômetro parado; -Hidrômetro não localizado, desaparecido, violado danificado ou invertido; *-By pass* -Violação do selo do cavalete do hidrômetro; -Leitura igual, sem consumo; -Leitura menor que a anterior.
Consumo zero sem ocorrência de leitura	Ligações sem consumo registrado nos últimos 3 meses

Tipos de Fraude	Conceitos
Consumo baixo sem ocorrência de leitura	Ligações com consumo >0 até 2 m³ nos últimos 3 meses ou, caso o total seja muito grande, reduzir até 1 m³ e priorizar imóveis de alto padrão construtivo.
Forte redução de consumo	Por exemplo, três reduções consecutivas no consumo
Existência de piscina e sem poço com consumo < 10 m³/mês	Imóveis de todas as categorias com ligação ativa de água, com piscina e sem poço com consumo médio menor que 10 m³/economia/mês
Cliente não-ativo com poço	Imóveis de todas as categorias com ligação não ativa de água e com poço

Fonte: (SNSA, 2018).

3.1.7 Aperfeiçoamento no setor comercial

Levando em consideração a redução das perdas aparentes, as companhias devem possuir um sistema de informação e controle adequado para o gerenciamento do setor comercial como um todo – englobando todos os requerimentos de ligações, cadastro comercial após solicitado e programas voltados para leitura, contabilização do consumo, faturamento e execução sínteses relatórios (TSUTIYA, 2006).

Por conseguinte, o cadastro de novas ligações de água e a identificação de ligações não cadastradas, clandestinas ou suprimidas reativadas irregularmente, deve ser ágil e eficiente de modo a evitar o detrimento do respectivo faturamento e o consumo de volume de água não aferido. Dessa forma, faz-se necessária uma atualização constante do cadastro comercial através telemetria e leitura manual com funcionários capacitados (leituristas) (TSUTIYA, 2006).

Faz-se relevante ressaltar, também, a existência do processo de atualização da categoria do imóvel em que se encontra registrado no cadastro comercial, visto que há diferenças nas estruturas tarifárias para cada companhia em função do estabelecimento ser residencial, comercial, industrial ou público. Frequentemente são notificados casos em que não são comunicados a alteração de uso ou quantidade de economias cujo aumento da tarifa é iminente. Destacam-se os casos em que edificações residenciais se convertem para comerciais e de imóveis que sofreram subdivisões em diferentes ocupações cadastrados como apenas uma economia.

As perdas comerciais apresentam um grande vínculo com o cliente consumidor da companhia responsável pelo seu abastecimento. Concomitantemente, os problemas nos HDs e fraudes estão diretamente associados ao cliente pertencente

que deve responder diretamente por qualquer irregularidade – identificada através de denúncias ou análises do banco de dados comercial (TSUTIYA, 2006).

3.2 Ciência de dados

A extração de informações úteis de grandes conjuntos de dados por meio de técnicas estatísticas e de computação é a base da ciência de dados, uma área interdisciplinar que busca gerar insights e conhecimentos capazes de embasar decisões informadas e fundamentadas em fatos (GREAT LEARNING TEAM, 2023).

O advento da ciência de dados ocorre em um contexto em que empresas e organizações têm acesso a volumes crescentes de dados sobre clientes, processos e produtos. Por meio de algoritmos de aprendizado de máquina e outras técnicas avançadas, cientistas de dados conseguem identificar padrões ocultos nos dados e prever tendências futuras (GREAT LEARNING TEAM, 2023).

Para a análise de dados ser bem-sucedida, é fundamental que o profissional possua um conjunto de habilidades que englobam conhecimentos em estatística, programação, visualização de dados e tomada de decisão, além de uma compreensão profunda do contexto em que os dados estão inseridos. Para isso, é necessário seguir algumas fases essenciais, dentre elas a coleta de dados, a limpeza de dados, a análise exploratória de dados, a modelagem de dados e a comunicação dos resultados. Seguir essas etapas é crucial para garantir uma análise precisa e confiável dos dados e uma aplicação eficaz dos resultados obtidos (GREAT LEARNING TEAM, 2023).

De acordo com Géron (2019), a coleta de dados é uma das fases mais críticas do processo de ciência de dados. É essencial que os dados sejam representativos e validados para garantir a confiabilidade das análises. Os dados são então armazenados em bancos de dados ou planilhas para fácil acesso. A limpeza de dados é uma etapa importante que visa identificar e corrigir erros ou inconsistências nos dados coletados, garantindo que os dados estejam prontos para a análise.

Uma vez que os dados estão limpos, a análise exploratória de dados é realizada para identificar padrões, tendências e relações entre os dados. Técnicas de visualização de dados são frequentemente usadas nesta fase para ajudar a identificar esses padrões. Em seguida, a modelagem de dados é aplicada, utilizando algoritmos de aprendizado de máquina e outras técnicas estatísticas para realizar previsões ou

identificar possíveis cenários futuros. Essa fase é crucial para a tomada de decisão informada e embasada em fatos.

Por fim, a comunicação dos resultados é essencial para garantir que as análises realizadas sejam compreendidas e aplicadas corretamente. Isso envolve a criação de relatórios e apresentações que comuniquem os insights gerados a partir dos dados. Em última análise, o sucesso do projeto de ciência de dados é determinado pela qualidade da comunicação dos resultados e pela aplicação efetiva das descobertas (GÉRON, 2019).

Além disso, faz-se de grande importância para a ciência e análise de dados, conceituar a definição de um *outlier* para se trabalhar com uma grande base de informações com diversas, variáveis, linhas e colunas. Tem-se que um outlier é o valor considerado muito distante dos demais valores dentro de uma base de dados. Apesar disso, o valor ser um outlier não o caracteriza como inválido ou errado. Porém, na maioria das vezes, trata-se do resultado de erros de dados ou leituras errôneas de determinado medidos. Dessa forma, quando advindos de dados ruins, a média aritmética, por exemplo, que sofre influência de valores extremos, terá uma estimativa ruim influenciada por esse valor. Nesse sentido, devem ser identificados e dignos de uma maior atenção, tornando-se importantes para a detecção de fraudes (BRUCE e BRUCE, 2019).

Existem várias técnicas para identificar outliers, que variam de métodos estatísticos simples a técnicas mais avançadas de análise de dados. Faz-se relevante destacar os boxplots que se tratam de uma representação gráfica que permite visualizar a distribuição dos dados de forma eficiente. Os valores considerados outliers são representados como pontos isolados que se encontram fora dos limites superior e inferior do boxplot. Essa abordagem gráfica é uma maneira útil de identificar valores extremos em diferentes grupos de dados, facilitando a comparação entre eles.

O boxplot é composto por um retângulo que representa a área entre o primeiro quartil (Q1) e o terceiro quartil (Q3) dos dados, e uma linha vertical que indica a mediana. O primeiro quartil, Q1, é o valor que divide os dados em 25% abaixo e 75% acima desse valor. O terceiro quartil, Q3, é o valor que divide os dados em 75% abaixo e 25% acima desse valor.

Além disso, ressalta-se o conceito de o conceito de amplitude interquartil (IQR), essencial para a construção do boxplot, pois ele representa a amplitude dos dados

centrais e ajuda a identificar possíveis outliers. Outrossim, o IQR é uma medida de dispersão importante na estatística descritiva e auxilia na análise dos dados

No entanto, em alguns casos, pode haver valores que são naturalmente limitados a um mínimo positivo, como em dados de contagem ou de taxa. Nestes casos, o método do IQR pode não ser apropriado para detectar outliers, pois o limite inferior (Q1 - 1,5IQR) pode ser menor que zero. Para resolver esse problema, uma abordagem alternativa é utilizar o limite inferior de Q1 - 3IQR, o que permite que o valor mínimo seja zero.

3.2.1 Aprendizado de Máquina

A área da inteligência artificial que se dedica a permitir que os computadores aprendam a partir de dados sem serem explicitamente programados é conhecida como aprendizado de máquina. Conforme Géron (2019), essa técnica é utilizada para encontrar padrões em dados e, a partir desses padrões, fazer previsões ou tomar decisões. O aprendizado de máquina é dividido em três categorias: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço.

Em engenharia e ciência de dados, é fundamental compreender conceitos como modelo, calibração e validação de modelo, e prognóstico ou predição. De acordo com James et al. (2013), um modelo é uma representação simplificada de um sistema ou fenômeno que permite a previsão de comportamentos futuros. A calibração de um modelo é o processo de ajustar seus parâmetros para que ele possa fazer previsões mais precisas. Já a validação de modelo é o processo de avaliar o desempenho do modelo em dados não utilizados para sua construção. Por fim, a predição é a capacidade de um modelo de fazer previsões precisas em dados desconhecidos.

No contexto da máquina, o aprendizado é necessário para que ela possa identificar padrões nos dados e fazer previsões ou tomar decisões mais precisas. Goodfellow et al. (2016) destacam que os algoritmos de aprendizado de máquina são responsáveis por identificar padrões nos dados e utilizá-los para fazer previsões. Assim, o aprendizado de máquina é fundamental para permitir que as máquinas realizem tarefas complexas que exigem capacidades cognitivas, como reconhecimento de fala, reconhecimento de imagem e tradução de idiomas.

Conforme explica Géron (2019), o aprendizado supervisionado é uma tarefa em que o modelo é treinado para mapear entradas e saídas com base em um conjunto de exemplos rotulados. Em outras palavras, o modelo é alimentado com dados de entrada previamente rotulados pelo ser humano para aprender a prever a saída correta para novos dados de entrada. Já no aprendizado não supervisionado, o objetivo é identificar padrões, estruturas e relações em dados não rotulados. Nessa abordagem, o modelo é treinado para encontrar padrões e agrupamentos em dados sem a necessidade de rótulos pré-definidos pelo ser humano.

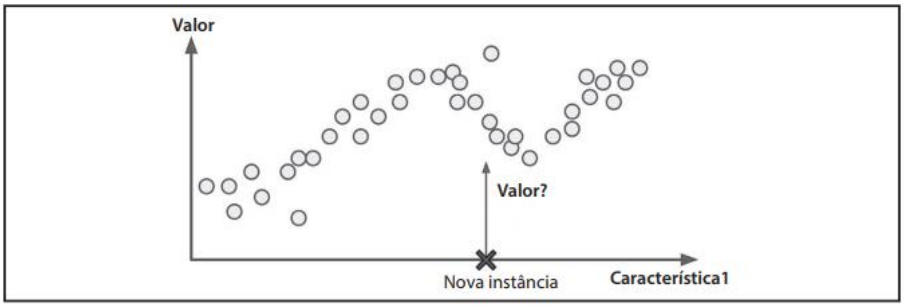
Em síntese, o aprendizado de máquina é uma técnica que permite que as máquinas aprendam a partir de dados e possam realizar tarefas complexas. É importante compreender os conceitos de modelo, calibração e validação de modelo, e prognóstico ou predição na engenharia e ciência de dados. A máquina precisa aprender para identificar padrões nos dados e fazer previsões precisas.

No aprendizado supervisionado, tem-se que os dados de treinamento fornecidos ao algoritmo já precedem com as soluções desejadas na base de dados, ou seja, os rótulos. Dessa forma, torna-se possível realizar a classificação ou prever um alvo de valor numérico a partir de outras características correlacionadas a uma variável alvo, descrita como regressão (GÉRON, 2019).

Desse modo, uma instância, ou melhor, um único ponto de dados ou observação completa composta de várias variáveis, alimenta um determinado modelo de aprendizagem (GÉRON, 2019). A Figura 8 representa a utilização de uma instância em um modelo de regressão

Além disso, é necessário que haja intervenção humana para fornecer dados rotulados aos modelos, enquanto no aprendizado não supervisionado, essa interferência não é necessária. Isso ocorre porque os modelos do aprendizado supervisionado são treinados com dados rotulados, enquanto os modelos do aprendizado não supervisionado são treinados com dados não rotulados. Nesse caso, os algoritmos utilizados encontram padrões e estruturas nos dados por conta própria (GÉRON, 2019).

Figura 8 – Instância em um modelo de regressão

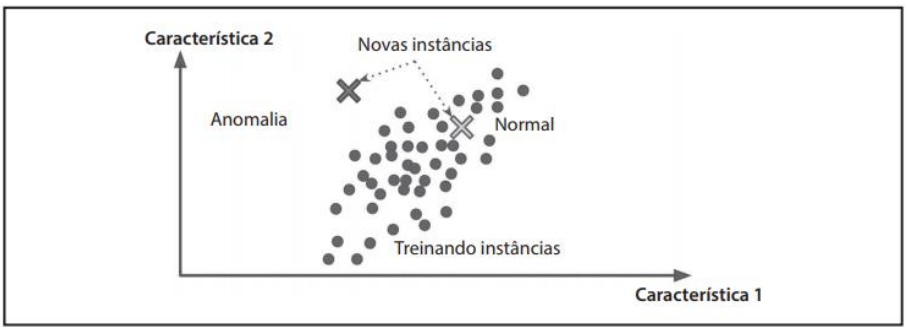


Fonte: (GÉRON, 2019).

O cenário descrito como não supervisionado é definido por seus dados de treinamento não serem rotulados, ou seja, o sistema tenta aprender sem orientação. Uma importante função da atividade não supervisionada trata-se da detecção de anomalias – como, por exemplo, a detecção de volumes consumidos de água incomuns para evitar fraudes ou remover automaticamente outliers de uma determinada base de dados antes de disponibilizá-lo para outro algoritmo de aprendizado. Nessa conjuntura, treina-se o sistema com entradas normais e, quando se identifica uma nova entrada, classifica-se em normal ou anomalia (GÉRON, 2019).

O procedimento descrito utilizando o método não supervisionado para detecção de anomalias pode ser exemplificado visualmente pela Figura 9.

Figura 9 – Detecção de anomalia



Fonte: (GÉRON, 2019).

Em síntese, o aprendizado supervisionado requer a intervenção humana para rotular os dados utilizados no treinamento do modelo, enquanto no aprendizado não supervisionado essa intervenção não é necessária, tornando o processo mais autônomo.

3.2.2 K Vizinhos Mais Próximos (KNN)

O K Vizinhos Mais Próximos ou *k-Nearest Neighbors* (KNN) é um dos algoritmos mais simples voltados à previsão e à classificação, visto que não existem modelos a serem ajustados. Faz parte da família de algoritmos conhecida como aprendizado embasado em instâncias (*instance-based learning*), visto que não existem parâmetros a serem aprendidos, e a classificação é baseada na distância até algumas amostras (k) de treinamento (HARRISON, 2020).

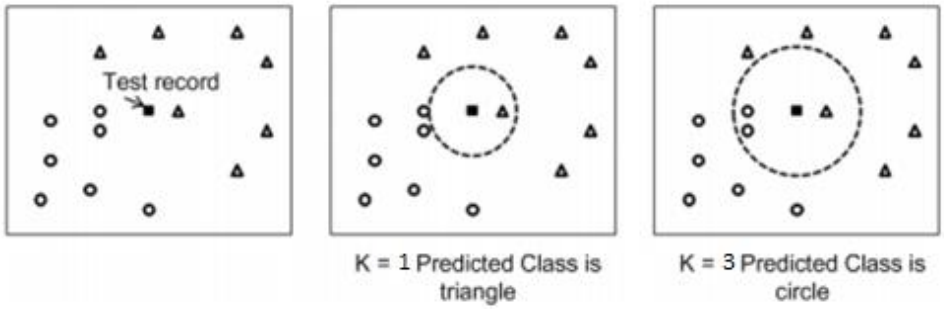
O modelo supracitado pressupõe que a distância em relação as amostras são suficientes para a realização de uma inferência que determine a sua classificação, sem existir nenhuma pressuposição sobre os dados subjacentes ou respectivas distribuições. Tem-se como principal dificuldade a determinação do valor aprovado de k e o fato de que a dimensionalidade pode atrapalhar as métricas de distância (HARRISON, 2020).

Por exemplo, trabalhando com um conjunto de dados de treinamento que consiste em várias formas geométricas, incluindo quadrados, triângulos e círculos. Para tanto, cada forma geométrica é representada por suas características, como comprimento dos lados, área, coordenadas dos vértices, entre outros. Além disso, cada forma possui uma classe associada: quadrado, triângulo ou círculo.

Para utilizar o KNN, primeiro precisamos determinar o valor do parâmetro k, que representa o número de vizinhos mais próximos a serem considerados para a classificação. Em seguida, quando recebemos um novo exemplo, um quadrado no nosso caso, o algoritmo calcula a distância entre esse exemplo e todos os exemplos do conjunto de treinamento. A distância pode ser medida usando diferentes métricas, como a distância euclidiana (HARRISON, 2020).

Depois de calcular as distâncias, o KNN seleciona os k exemplos do conjunto de treinamento que estão mais próximos do exemplo de teste. Esses k vizinhos mais próximos têm seus rótulos de classe considerados para determinar a classe do exemplo de teste. Por exemplo, se definirmos k = 3, o algoritmo identificará os três exemplos mais próximos do quadrado de teste. Se dois desses exemplos forem círculos e um for triangulo, o KNN classificará o quadrado como pertencente à classe que possui a maioria dos vizinhos mais próximos, ou seja, a classe "círculo", da mesma forma que é representado na Figura 10.

Figura 10 – Classificação através do k vizinhos mais próximos



Fonte: Adaptado de (KOTU e DESHPANDE, 2019).

3.2.3 Árvore de decisão

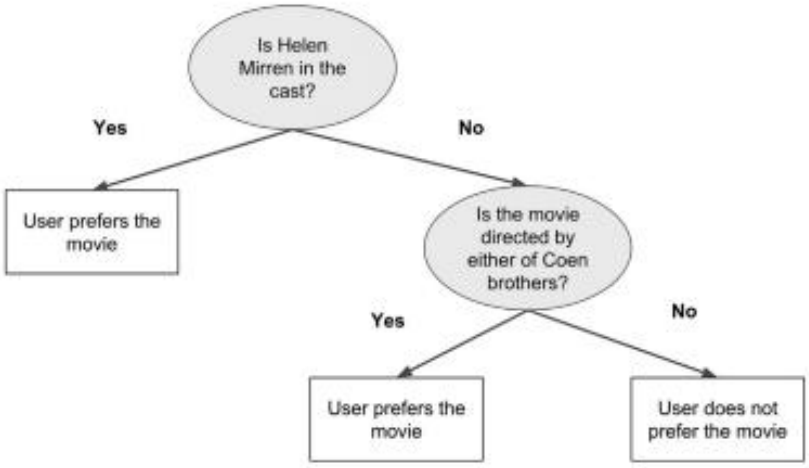
Conceitua-se que os modelos de árvore como efetivos e bastante populares, desenvolvidos em 1984 por Leo Breiman e outros autores. Constituem o alicerce das ferramentas de modelagem preditiva mais utilizadas e efetivas, bastante difundidas na ciência de dados como um todo (BRUCE e BRUCE, 2019).

Ademais, as árvores de decisão ou *decision trees* aderem tanto à classificação quanto à predição. O algoritmo avalia explicadamente, em cada nível da árvore, várias separações nos atributos e a divisão que apresentar menor erro (impureza) será escolhida (HARRISON, 2020).

Em síntese, define-se um modelo de árvore como um conjunto de regras “se-então-senão” que apresentam uma grande facilidade de entendimento e, consequentemente, implementação. Para tanto, tem-se que as árvores de decisão são compostas por diferentes nós e ramos na árvore. Cada nó representa uma característica a ser considerada, enquanto os ramos representam diferentes resultados ou opções possíveis (BRUCE e BRUCE, 2019).

Por exemplo, a Figura 11 explicita uma árvore de decisão cujas perguntas específicas da árvore de decisão para determinar a preferência dos usuários em relação a um filme. É levado em consideração se uma determinada atriz está no elenco e se o filme é dirigido por um ator específico, totalizando dois nós (duas perguntas) e três ramos na árvore de decisão em questão.

Figura 11 – Exemplo de diagrama de uma arvore de decisão



Fonte: (KOTU e DESHPANDE, 2019).

Sendo assim, essa árvore de decisão pode ser utilizada como um modelo simples para entender a preferência dos usuários com base nas informações fornecidas. No entanto, é importante observar que essas são perguntas específicas para este exemplo e podem não refletir todas as variáveis e nuances que influenciam a preferência do público em relação a um filme.

3.2.4 Floresta aleatória

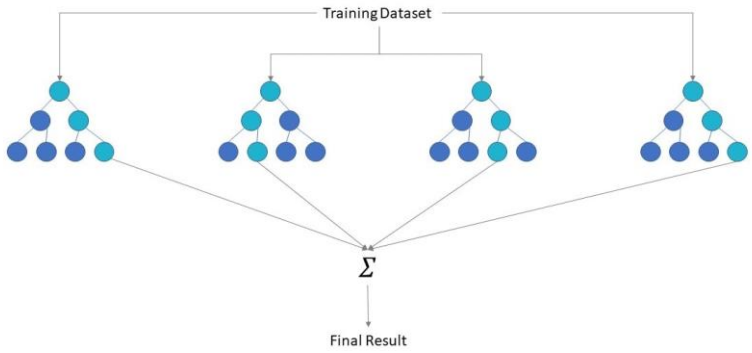
Visto que as árvores de decisão são relevantes devido ao fato de serem explicáveis, assim como apresentado na Figura 11, mas apresentam uma grande tendência de superadequação ou *overfitting* – modelo que se ajusta demasiadamente bem ao conjunto de dados previamente trabalhado, porém se torna ineficiente para previsão de novos resultados. Dessa forma, uma floresta aleatória ou *random forest* abdica da facilidade de explicação do modelo em troca de uma maior tendencia de generalização no que se refere a sua execução (HARRISON, 2020).

Basicamente o algoritmo da floresta aleatória é um método que realiza previsões mais precisas do que uma árvore simples, porém as regras de decisão intuitivas da árvore simples são perdidas. Apresenta, também, como resultado a medida de importância de variável que classifica as preditoras em função da sua contribuição para a precisão do modelo elaborado (BRUCE e BRUCE, 2019).

Em síntese o algoritmo cria uma "floresta" de árvores de decisão, onde cada árvore é treinada com diferentes amostras e características aleatórias. Ao combinar

as previsões das árvores individuais, tem-se uma previsão mais robusta e geralmente mais precisa assim como representado na Figura 12.

Figura 12 – Diagrama de classificação da floresta aleatória



Fonte: (IBM, 2020).

3.2.5 Avaliando modelos de classificação

Dentre as métricas de classificação dos modelos – comumente utilizados quando vários algoritmos diferentes são utilizados na modelagem preditiva – destaca-se a matriz de confusão. Trata-se de uma tabela que mostra o número de previsões corretas e incorretas categorizadas por tipo de resposta (BRUCE e BRUCE, 2019).

Trata-se de uma tabela de coincidência que separa as decisões tomadas pelo classificador, evidenciando, explicitamente, como uma classe está sendo confundida com outra, permitindo que os diferentes erros possam ser tratados de maneira separada. Na diagonal principal, encontram-se as contagens das decisões corretas, já os erros do classificador são os falsos positivos (dados negativos classificados como positivos) e os falsos negativos (dados positivos classificados como negativos) (PROVOST e FAWCETT, 2016).

Utiliza-se, comumente, a matriz de confusão com dimensão 2x2 (“nxn”, sendo “n” a quantidade de classes). Na Tabela 2 pode-se observar a matriz de confusão que será utilizada no trabalho em questão.

Tabela 2 – Matriz de confusão

		Classes Previstas	
		Positivo (1)	Negativo (0)
Classes Reais	Positivo (1)	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Negativo (0)	Falso Positivo (FP)	Verdadeiro Negativo (TN)

Fonte: (BRUCE e BRUCE, 2019).

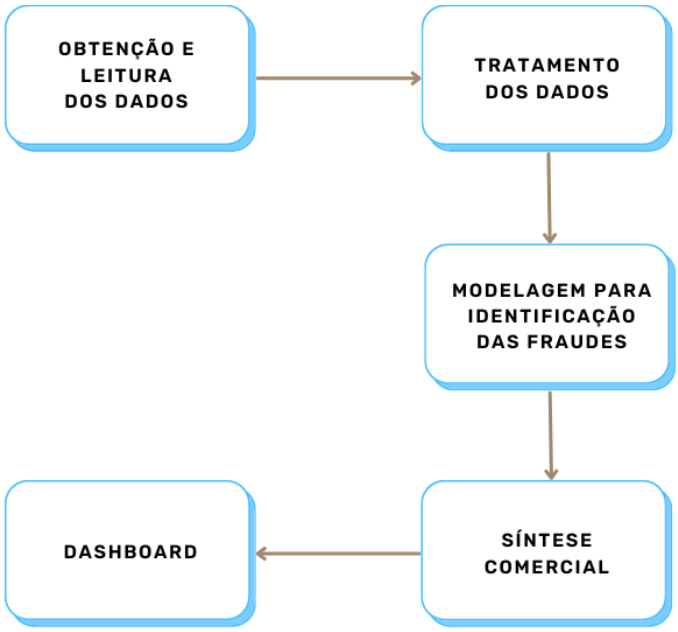
Uma vez identificados cada uma das categorias presentes na matriz de confusão, faz-se importante definir a acurácia do modelo, ou seja, a porcentagem de classificações corretas. Dessa forma, a acurácia pode ser calculada por meio da equação (1).

$$\textbf{acurácia}\ (\%) = \frac{\textbf{Predições corretas}}{\textbf{Todas as predições}} = \frac{\textbf{TP} + \textbf{TN}}{\textbf{TP} + \textbf{FP} + \textbf{TN} + \textbf{FN}} \qquad (1)$$

4 MÉTODOS

A metodologia deste trabalho divide-se em quatro etapas: Obtenção, leitura, identificação e tratamento dos dados; Modelagem para identificação de anormalidades e indícios de prováveis fraudes; Síntese comercial das perdas aparentes; e, por fim, *Dashboard* de resultados. A metodologia utilizada neste trabalho é descrita pelo fluxograma da Figura 13.

Figura 13 – Fluxograma representativo da metodologia utilizada.

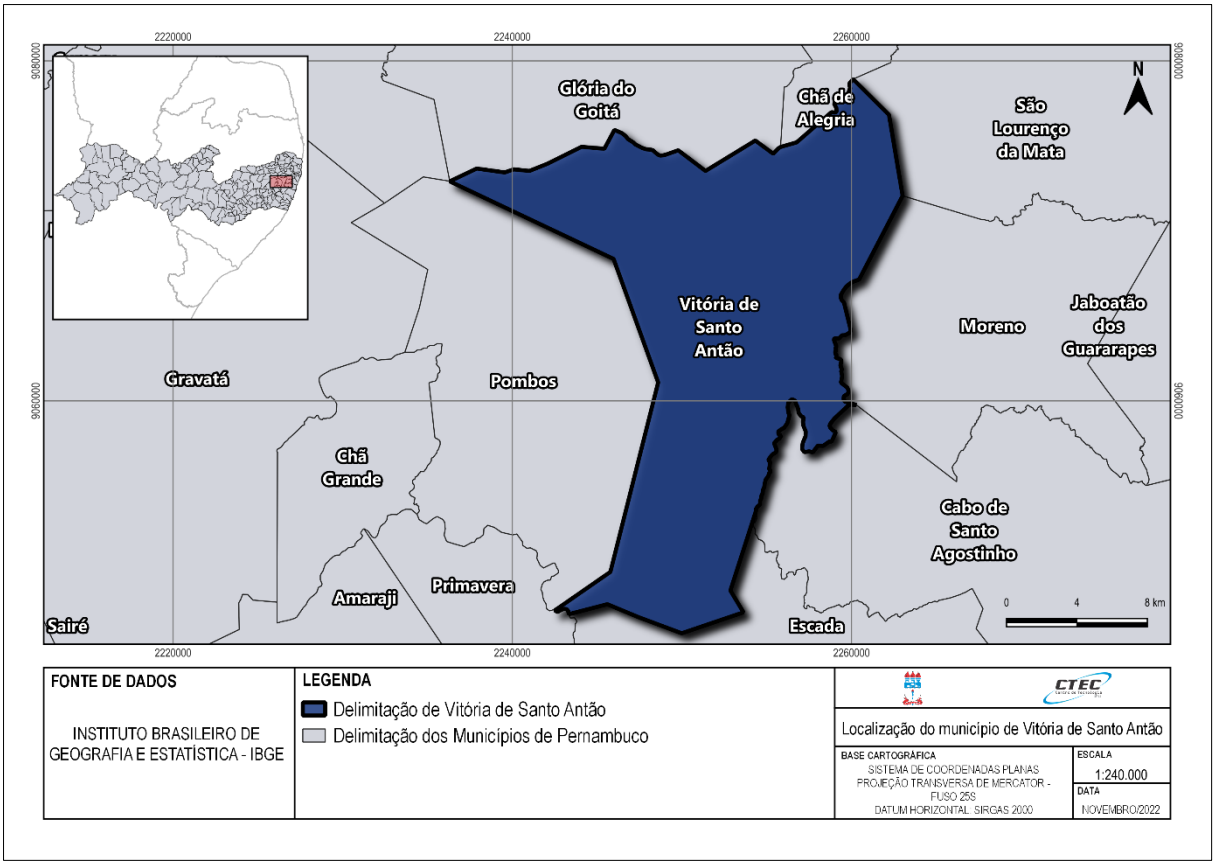


Fonte: (Autor, 2023).

4.1 Obtenção, leitura, identificação e tratamento dos dados

Inicialmente, foram utilizados os dados fornecidos pela Companhia Pernambucana de Saneamento (COMPESA), atual concessionária do fornecimento de água de Pernambuco. Dentre os arquivos disponibilizados, tem-se o cadastro comercial do Lote 4, contendo as informações de duas cidades, dentre elas, Vitória de Santo Antão, foco deste trabalho.

Figura 14 – Mapa de localização da cidade de Vitória de Santo Antão – PE.



Fonte: (Autor, 2023).

A obtenção dos dados foi realizada por meio da disponibilização de três planilhas eletrônicas (formato xlsx) conforme apresentado na seguinte listagem de arquivos imprescindíveis para o estudo identificados na Tabela 3. Já na Tabela 4 é possível observar cada informação por coluna nos dados disponibilizados por meio dos arquivos.

Tabela 3 – Planilhas obtidas da concessionária

Arquivo xlsx	Dados comerciais por matrícula
Lote 4 - Dados Comerciais 5_1.xlsx	Dados dos meses de março/2020 até junho/2020
Lote 4 - Dados Comerciais 5_2.xlsx	Dados dos meses de julho/2020 até outubro/2020
Lote 4 - Dados Comerciais 5_3.xlsx	Dados dos meses de novembro/2020 até fevereiro/2021

Fonte: (Autor, 2022).

Tabela 4 – Sumário de dados obtidos da concessionária.

Dados	Formato
Categoria da ligação	Residencial, comercial, industrial ou pública
Situação da ligação de água	Ligada, cortada, suprimida, factível, entre outras

Dados	Formato
Data de instalação do HD	Dia/mês/ano
Número de economias	Quantidade
Série de no mínimo 12 meses de volume consumido	m³
Volume faturado	m³
Anormalidade de consumo	Tipo de anormalidade
Anormalidade de leitura	Tipo de anormalidade

Fonte: (Autor, 2022).

Sendo a categoria a classificação da economia em função da natureza de ocupação. Já a economia pode ser definida como imóveis com ocupação única ou subdivisão independente, identificáveis ou comprováveis pela finalidade de sua ocupação legal, e que possuam instalação privada ou comum para abastecimento de água e/ou coleta de esgoto (COMPESA, 2020).

A situação da ligação de água em cada imóvel presente no banco de dados comercial pode ser explicitada de acordo com os significados da Tabela 5:

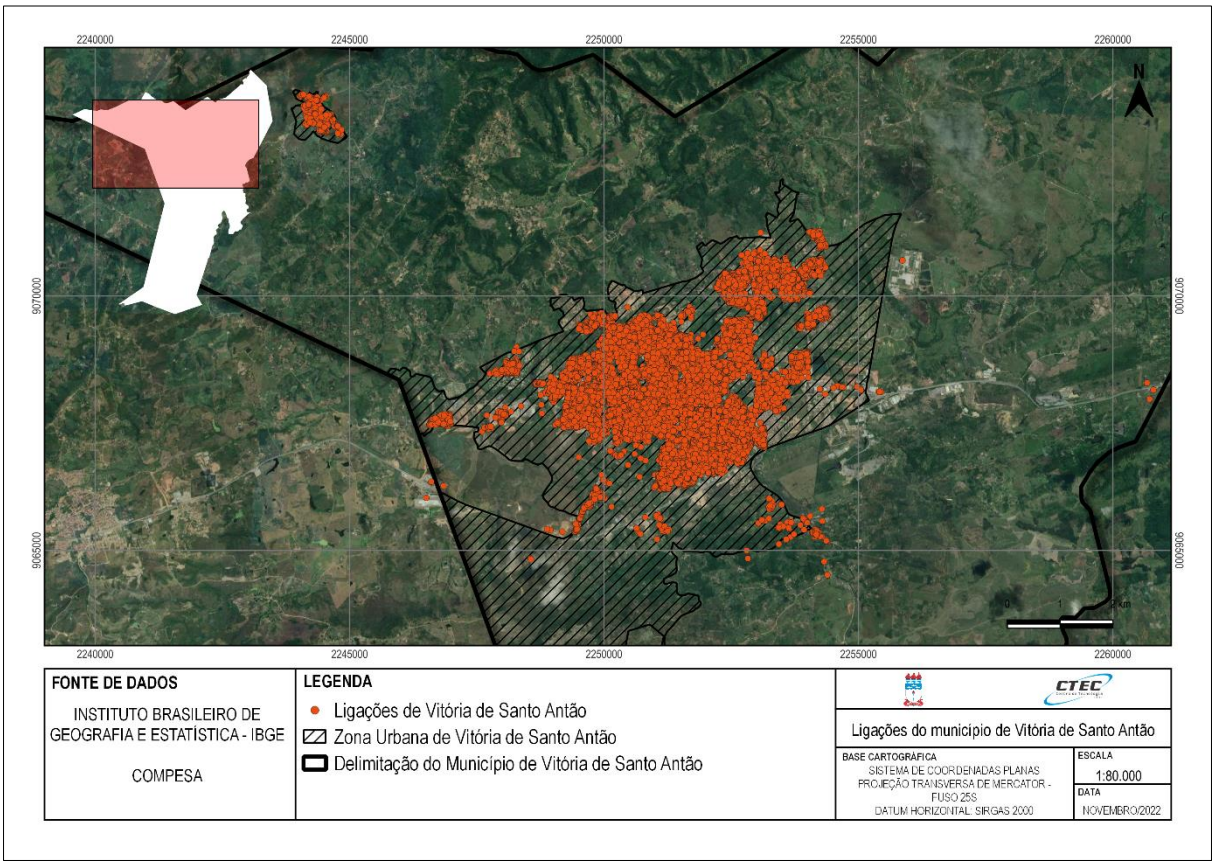
Tabela 5 – Significado de cada uma das informações de situação da ligação de água.

Situação da ligação de água	Significado
LIGADO	Imóveis que possuem ramal predial de água conectado à rede de distribuição de água, podendo possuir ou não medição do consumo de água
FACTÍVEL	Imóveis ou terrenos vagos situados em área urbana servida por rede de distribuição de água, mas que não possuem ramais prediais interligados à rede
POTENCIAL	Imóveis ou terrenos vagos situados em área não servida por rede de distribuição de água
CORTADO	Imóveis que sofreram interrupção ou desligamento dos serviços da companhia, por meio de instalação de dispositivo supressor ou outro meio
CORTADO A PEDIDO	Imóveis que sofreram interrupção ou desligamento dos serviços pela companhia, a pedido do usuário, por meio de instalação de dispositivo supressor ou outro meio
SUPRIMIDO	Imóveis que sofreram interrupção ou desligamento definitivo dos serviços, por meio de retirada das instalações entre o ponto de conexão e a rede pública, com a suspensão da emissão de faturas
SUPRIMIDO A PEDIDO	Imóveis que sofreram interrupção ou desligamento definitivo dos serviços, a pedido do usuário, por meio de retirada das instalações entre o ponto de conexão e a rede pública, com a suspensão da emissão de faturas

Fonte: (COMPESA, 2020).

Por conseguinte, é possível observar no mapa da Figura 15, todas as ligações únicas obtidas que foram disponibilizadas pela companhia

Figura 15 – Mapa das ligações da cidade de Vitória de Santo Antão – PE.



Fonte: (Autor, 2022).

No decorrer da realização do TCC, utilizaram-se diversas ferramentas que foram essenciais para o desenvolvimento do código e acompanhamento do processo de elaboração do trabalho. Dentre elas, destacam-se o Anaconda e o *Jupyter Notebook*.

O *Anaconda* é uma distribuição de *Python* que oferece várias bibliotecas e ferramentas pré-instaladas para ciência de dados. Essa ferramenta simplificou o processo de instalação e gerenciamento das bibliotecas necessárias para o desenvolvimento do código do TCC. Além disso, o *Anaconda* possibilitou a criação de ambientes virtuais que garantiram a compatibilidade das bibliotecas e simplificaram a administração de diferentes projetos.

Por outro lado, o *Jupyter Notebook* é uma aplicação web que permite a criação de documentos interativos que combinam código, texto explicativo e visualizações gráficas. Esse software foi utilizado para elaborar o código do TCC, viabilizando a escrita e a execução do código em blocos interativos, assim como a visualização dos

resultados obtidos em cada etapa do projeto. Além disso, o *Jupyter Notebook* possibilitou a exportação do documento em diversos formatos, o que facilitou a apresentação dos resultados e a documentação do trabalho.

Também foi utilizado o *Google Colaboratory*, ferramenta escolhida para permitir o acompanhamento online do código e a sua disponibilização. O Google Colaboratory é uma plataforma gratuita que permite a criação e o compartilhamento de *notebooks Jupyter* na nuvem, sem a necessidade de instalação de software no computador local. Dessa forma, o *Google Colaboratory* tornou possível o acompanhamento online do código em tempo real, bem como a sua disponibilização para os orientadores e demais interessados em acessar o trabalho.

Caso haja interesse em acompanhar o progresso da codificação do TCC é possível acessá-la por meio Google Colaboratory, disponível através do link: Análise de dados comerciais e identificação de fraudes (Google Colaboratory). O arquivo Jupyter Notebook encontra-se nesse repositório e apresenta todo o código desenvolvido para a elaboração do TCC, possibilitando a visualização dos resultados obtidos em cada etapa do projeto. Após acessar o link, basta entrar na sua conta, instalar o aplicativo do *Google Colaboratory* e abrir o arquivo enviado com o aplicativo conectado.

Prosseguindo com o código, foi criado um *datraframe* – estrutura de dados tabular bidimensional – para cada uma das planilhas, com o objetivo de armazenar assim como tornar possível o processamento, por meio da linguagem de programação *Python* e suas bibliotecas *Pandas* e *Numpy*, ferramentas essas de uso gratuito e aberto, evitando custos para a elaboração do diagnóstico, além de oferecerem estruturas para manipulação de grandes tabelas numéricas, séries temporais, suporte para vetores multidimensionais com operações básicas, tratar dados ausentes, agregar e tratar dados. Ademais, tornam-se alternativas às planilhas eletrônicas comumente utilizadas que apresentam limitações nas quantidades de linhas e colunas, apresentando uma queda no processamento ou desempenho quando esse número é testado.

Em sequência, utilizando as próprias funções da biblioteca Pandas, tornou-se possível realizar o tratamento dos dados disponibilizados com o objetivo de tornar mais prática a extração de cada informação a depender do tipo do dado em cada coluna do *dataframe*.

Após isso, fez-se possível a extração de informações gerais como o número total de ligações, economias, faturamento, volume consumido, média de idade de instalação do parque de HDs e identificar ligações com e sem HD.

Concomitantemente, foi possível realizar a aplicação do conceito de outliers para a identificando os valores discrepantes em cada coluna na base de dados e variáveis que não tragam informações úteis para a resolução do problema, assim como para a criação do modelo realizada na próxima etapa. Como, por exemplo, ligações industriais que realmente possuam um valor muito alto de consumo de água que destoe completamente do restante do cadastro comercial.

Quando se analisam os valores de consumo de água ao longo de 12 meses, é importante identificar os outliers para compreender as características dos dados. A distância interquartil (IQR), que é a diferença entre os quartis Q1 e Q3, é uma das maneiras de se identificar outliers. No entanto, quando os valores de consumo apresentam uma restrição natural de não poderem ser menores que 0, como no caso dos dados de consumo de água, a utilização da regra geral de 1.5*IQR pode subestimar a presença de outliers, já que essa regra foi criada para dados sem essa restrição.

Nesses casos, é mais indicado utilizar a regra de 3*IQR, que é mais adequada a esse tipo de dados e é menos suscetível a falsos negativos. Essa regra amplia a distância interquartil, permitindo uma maior tolerância para valores que estejam distantes do intervalo interquartil.

Desse modo, foi possível realizar a criação de recursos adicionais a partir dos valores de cada matrícula para uma posterior relevância na identificação de fraudes.

4.2 Modelagem para identificação de anormalidades e indícios de prováveis fraudes

Para a identificação de indícios de prováveis fraudes utilizou-se o conceito de *outliers* e Aprendizado de Máquina visto que se trata de uma problemática complexa sem uma solução exata ao utilizar-se abordagens tradicionais, associado ao fato de envolver uma grande quantidade de dados para a realização do estudo.

Em seguida, foi realizada uma análise exploratória referente aos dados prontos e previamente tratados da etapa anterior, observando como cada variável e recurso

adicional criado se relaciona e como cada uma delas influencia na caracterização da anormalidade.

Procedimento que ocorreu por meio da criação da matriz de correlação de cada um dos dados. Por padrão nas funções para calcular as matrizes de correlação do pacote *Pandas* em *Python* é utilizado o coeficiente de correlação de Pearson – ferramenta importante para identificar padrões e tendências nos dados e pode ser utilizado para fazer previsões e inferências sobre a relação entre as variáveis.

Segundo Triola (2018), o coeficiente de correlação de Pearson é uma medida estatística que avalia a relação linear entre duas variáveis continuas e varia entre -1 e +1. Valores próximos de +1 indicam uma correlação positiva perfeita, enquanto valores próximos de -1 indicam uma correlação negativa perfeita. Quando o coeficiente é próximo de 0, isso indica uma ausência de correlação entre as variáveis.

Outrossim, fez-se necessário o pré-processamento dos dados antes da modelagem. Nessa conjuntura, para colunas categóricas de texto (*strings*) na base de dados, deve-se transformá-los em variáveis *dummy* – assumem o valor de 0 ou 1 empregue para classificar dados em categorias mutualmente exclusivas. Uma vez que os parâmetros mais significativos foram identificados e transformados para valores numéricos, será possível a utilização dos algoritmos para o modelo de aprendizagem de máquina.

Não obstante, para uma análise posterior, as variáveis foram normalizadas com objetivo de escalar os seus valores para um intervalo específico, geralmente [0,1], com o objetivo de remover as diferenças de escalas das variáveis. A normalização pode ser necessária porque muitos algoritmos de aprendizagem de máquina são sensíveis à escala dos dados de entrada. Nesse sentido, variáveis com escalas diferentes podem ter um impacto desigual no modelo e podem levar a resultados incorretos ou enviesados.

A criação dos modelos de aprendizagem de máquina – KNN, árvore de decisão e floresta aleatória –, e todos os outros procedimentos atrelados a ele, ocorreram por meio da biblioteca *Scikit-Learn* de código aberto referência na disponibilidade de algoritmos para essa vertente, tornando-se uma alternativa prática, eficiente e gratuita para utilização, amplamente buscada para os estudos de aprendizado de máquina por meio do *Python*.

Com isso, foi possível dividir os parâmetros em dados de treino e teste para serem importados como entrada nos algoritmos de aprendizagem de máquina. A

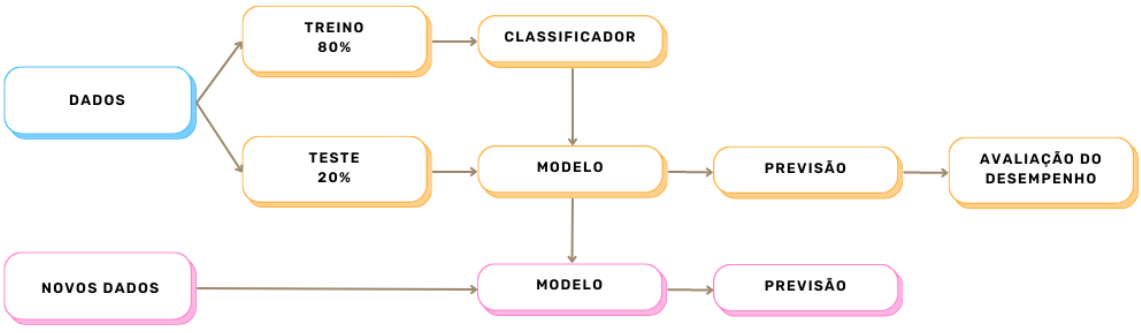
proporção dos conjuntos de dados a serem incluídos para treinamento do modelo foi de 80% como entrada na função que realiza essa divisão aleatoriamente e, consequentemente 20% dos dados foram utilizados testar o modelo criado.

Em resumo, os dados de teste são usados para validar o conhecimento do modelo, simulando como ele responderia a novos dados de entrada. Isso permite avaliar posteriormente o desempenho do modelo. Além disso, a divisão evita vieses na avaliação do modelo. Uma vez que utilizando os mesmos dados para treinar e testar o modelo, tem-se um desempenho superestimado visto que ele já teria visto esses dados durante o treinamento. Desse modo, a dividir os dados em treino e teste é importante para avaliar o desempenho e garantir uma avaliação imparcial.

Após isso, foi possível observar os resultados e comparar, através da matriz de confusão e a porcentagem de acurácia, qual modelo possui maior precisão e melhor se adequa a finalidade proposta de identificar as prováveis fraudes. Quando necessário, foi feito o uso da identificação dos outliers para uma posterior melhoria nas variáveis da base de dados com o objetivo de obter uma melhor acurácia. Com isso, uma vez que o modelo já foi criado, tornou-se possível realizar o teste com novos dados ou dados fictícios para ver como o modelo classifica cada ligação.

A Figura 16 ilustra a criação dos modelos supervisionados realizada na etapa em questão.

Figura 16 – Criação dos modelos de aprendizagem supervisionado.



Fonte: (Autor, 2023).

É importante destacar que os modelos de aprendizagem supervisionada desenvolvidos foram criados sem a utilização de hiperparâmetros. Caracterizados como parâmetros externos definidos antes do treinamento e que influenciam o

comportamento e o desempenho do modelo. Eles diferem dos parâmetros do modelo, que são ajustados pelo próprio algoritmo de aprendizagem durante o treinamento.

4.3 Análise comercial das perdas aparentes

Para a síntese comercial foram realizados cálculos e manipulações envolvendo as informações disponibilizadas pela companhia – através das bibliotecas *Pandas* e *Numpy*. Obtendo-se, assim, o volume consumido (VC), volume faturado (VF) e faturamento (F) por economia de cada ligação. Por conseguinte, os dados foram agrupados e estratificados em faixas de consumo, exibindo, também, os resultados em valores percentuais integrais e acumulados. Não obstante, foram estratificadas as ligações ativas por idade de instalação do hidrômetro, por categoria e por marca de hidrômetro.

Em seguida, fez-se uso das idades de instalação dos hidrômetros e do Índice de Desempenho da Medição (IDM). Índice este que associa o desempenho do hidrômetro levantado em laboratório de determinados hidrômetros de um parque estudado com o perfil de consumo da população, conforme descritas as recomendações e procedimentos da norma ABNT NBR 15.538 – Medidores de água potável – Ensaios para avaliação de eficiência (AESBE, 2015). Com essas informações obtidas para cada um dos HDs, tornou-se possível realizar a estimativa de submedição no parque de hidrômetros.

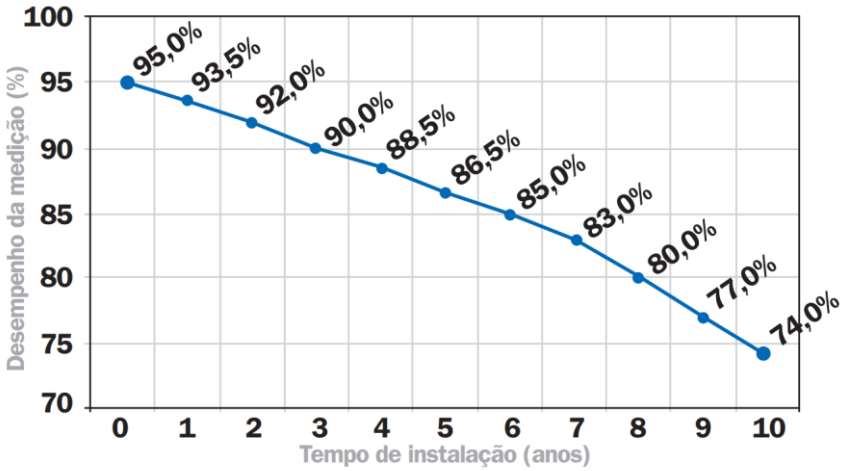
Durante o processo de análise do desempenho de um parque de hidrômetros, foi utilizado um conjunto de ferramentas disponibilizadas pela AESBE (2015) para estimar a submedição das ligações. Essas ferramentas incluem a equação (2) de submedição e a curva-padrão de desempenho da medição em função da idade de instalação do hidrômetro exemplificada na Figura 17.

$$\boldsymbol{volume\ submedido} = \frac{\boldsymbol{volume\ micromedido}}{\boldsymbol{IDM}} - \boldsymbol{volume\ micromedido} \qquad (2)$$

A equação de submedição é uma importante fórmula que permite estimar a submedição de um hidrômetro a partir da idade de sua instalação, considerando diversos fatores que podem afetar o seu desempenho, como o desgaste de peças e a sedimentação de resíduos. Com isso, é possível identificar possíveis problemas de

submedição em um parque de hidrômetros e realizar medidas para melhorar a precisão das medições realizadas. Trata-se, portanto, de uma ferramenta valiosa para garantir a qualidade dos serviços de abastecimento de água.

Figura 17 – Desempenho da medição em função do tempo de instalação.



Fonte: (AESBE, 2015).

A curva-padrão de desempenho da medição em relação à idade de instalação do hidrômetro apresentada na Figura 17 é uma ferramenta gráfica que possibilita uma visualização clara do comportamento do desempenho dos hidrômetros ao longo do tempo. Essa curva apresenta o percentual de submedição em relação à idade de instalação do hidrômetro, permitindo a identificação de tendências ou anomalias no desempenho dos equipamentos.

Desse modo, cada ano apresentado anteriormente está associado a um IDM, caracterizando a porcentagem de água consumida que o hidrômetro consegue medir. Trata-se de uma ferramenta útil para analisar o desempenho de hidrômetros e tomar medidas para melhorar a precisão das medições realizadas.

Com isso, tornou-se possível criar uma função para calcular o IDM para cada uma das colunas a partir da idade de cada um dos hidrômetros já obtidos. Utilizou-se os valores de desempenho da medição da Figura 17 até dez anos de instalação, e, para valores maiores que dez anos, utilizou-se o último valor de IDM do gráfico (74%)

Por conseguinte, obteve-se o volume corrigido – valor calculado através do quociente da equação (2) do volume consumido pelo IDM – e, por fim realizou-se a subtração pelo próprio volume consumo para obter-se o volume submedido de cada uma das ligações, realizando assim uma análise essencial para garantir a precisão

das medições e evitar prejuízos financeiros para a empresa responsável pelo fornecimento de água.

Todos os resultados supracitados estarão dispostos, conjuntamente, de maneira visual, através da representação dos gráficos de:

- Ligações, quantidade de economias e situação das ligações por perfil e categoria;
- Volume consumido médio (m³/ligação), volume faturado médio (m³/ligação) e faturamento médio (R$/economia) perfil e categoria;
- Média das idades das ligações;
- Estratificações das ligações ativas;
- Curva ABC do volume consumido e faturado por ligação ativa;
- Quantidade de anormalidades de consumo no HD.

Desse modo, apresenta-se do impacto das perdas aparentes sobre o sistema e receita financeira da companhia, além de analisar a situação do parque de hidrômetros. Estes gráficos foram desenvolvidos e plotados pelas bibliotecas *Seaborn* e *Plotly*, de extrema valia no que se refere a visualização de dados com *Python*, dotadas de vários tipos de gráficos interativos, tornando-se essenciais para as inferências estatísticas e demonstração dos resultados dos dados trabalhados.

4.4 *Dashboard* de resultados

Em síntese, a maioria os gráficos e resultados obtidos nas outras etapas foram organizados em um *dashboard* no formato de um relatório, desenvolvido por meio do framework *Dash* – originado da mesma empresa desenvolvedora da biblioteca *Plotly* – de código aberto utilizado para a criação rápida de aplicativos em *Python* para exposição de dados. Por fim, o *dashboard* foi colocado no ar como uma aplicação web com o link disponibilizado para o leitor.

5 RESULTADOS

5.1 Preparação dos dados

Uma vez criado cada *dataframe* para sua respectiva planilha, concatenou-se os três *dataframes* de dados totalizando 2.348.976 linhas e 19 colunas de informação bruta possuindo a matrícula como índice que pode ser observado na Figura 18. Matrículas, a priori, repetidas e englobando as cidades de Caruaru e Vitória de Santo Antão, porém representando meses diferentes de registro de informação, tornando necessário tratamento e filtragem dos dados.

Figura 18 – *Dataframe* final criado a partir de todas as planilhas disponibilizadas.

Matricula	Gerencia	Localidade	Municipio	Perfil	Categoria Principal	Qtd Economias	Condominio?	Situacao Lig Agua	Lig Agua Micromed?	Dt Inst Hidrometro	AM Referencia	Faturamento Esgoto
53553435	GNR AGRESTE CENTRAL	ALTO DO MOURA	CARUARU	NaN	RESIDENCIAL	1	NAO	NaN	NAO	NaT	NaN	NaN
53553524	GNR AGRESTE CENTRAL	ALTO DO MOURA	CARUARU	NaN	RESIDENCIAL	1	NAO	NaN	NAO	NaT	NaN	NaN
53553540	GNR AGRESTE CENTRAL	ALTO DO MOURA	CARUARU	NORMAL	RESIDENCIAL	1	NAO	LIGADO	SIM	2016-08-02	202006.0	0.0
53553540	GNR AGRESTE CENTRAL	ALTO DO MOURA	CARUARU	NORMAL	RESIDENCIAL	1	NAO	LIGADO	SIM	2016-08-02	202005.0	0.0
53553540	GNR AGRESTE CENTRAL	ALTO DO MOURA	CARUARU	NORMAL	RESIDENCIAL	1	NAO	LIGADO	SIM	2016-08-02	202004.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...
109418751	GNR MATA SUL	VITORIA DE SANTO ANTAO	VITORIA DE SANTO ANTAO	NORMAL	RESIDENCIAL	1	NAO	LIGADO	SIM	2021-02-03	202101.0	0.0
109419561	GNR MATA SUL	VITORIA DE SANTO ANTAO	VITORIA DE SANTO ANTAO	TARIFA SOCIAL	RESIDENCIAL	1	NAO	LIGADO	NAO	NaT	202102.0	0.0
109419561	GNR MATA SUL	VITORIA DE SANTO ANTAO	VITORIA DE SANTO ANTAO	NORMAL	RESIDENCIAL	1	NAO	LIGADO	SIM	2021-02-03	202101.0	0.0
109426622	GNR MATA SUL	VITORIA DE SANTO ANTAO	VITORIA DE SANTO ANTAO	NORMAL	RESIDENCIAL	1	NAO	LIGADO	SIM	2021-02-10	202102.0	0.0
109428250	GNR MATA SUL	VITORIA DE SANTO ANTAO	VITORIA DE SANTO ANTAO	NORMAL	RESIDENCIAL	1	NAO	LIGADO	SIM	2021-02-12	202102.0	0.0

2348976 rows × 19 columns

Fonte: (Autor, 2023).

Para tanto, foram concatenados três *dataframes* distintos que possuem informações por mês, organizados em linhas para cada uma das matrículas. Nessa conjuntura, fez-se necessário pivotar em *dataframes* distintos as informações de volume consumido (VC), volume faturado (VF) e anormalidades, organizando-as em 12 colunas.

Cada coluna representando um registro mensal, de forma a tornar os dados mais trabalháveis e levar em consideração a série de dados de um ano para as

análises posteriores. A estrutura obtida em cada um deles é representada pela Figura 19 e contém todas as matrículas do banco de dados disponibilizado.

Figura 19 – *Dataframe* pivotado designado para trabalhar com os valores de volume consumido ao longo de 12 meses.

	Vol Consumido											
AM Referencia	202003.0	202004.0	202005.0	202006.0	202007.0	202008.0	202009.0	202010.0	202011.0	202012.0	202101.0	202102.0
Matricula												
6150420	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6150438	NaN	10.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN	6.0	7.0	4.0
6150497	NaN	0.0	NaN	NaN	0.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6150519	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6150527	19.0	NaN	NaN	NaN	NaN	NaN	14.0	17.0	15.0	14.0	26.0	23.0

Fonte: (Autor, 2023).

De maneira semelhante, também se criou um *dataframe* pivotado contendo as informações com registro de 12 meses de anormalidades, que serão utilizados posteriormente na etapa de análise das anormalidades.

5.2 Análise exploratória e criação de recursos

Para a análise dos dados, separou-se um *dataframe* final, seguindo a estrutura de colunas apresentada na Figura 18, com apenas os dados da cidade de Vitória de Santo Antão para o último mês de registro do banco de dados (fevereiro de 2021), visto que possui as últimas e mais recentes informações de perfil, categoria, quantidade de economias e situação da ligação. Foram contabilizadas 31.718 linhas que representam cada uma das ligações com matrículas únicas para a cidade analisada.

Foram criados *dataframes* para analisar o consumo e as anormalidades no intervalo de 12 meses, com mesma estrutura após ter as informações pivotadas assim como apresentado na Figura 19, apenas para analisar e criar os recursos. Após isso cada um dos recursos foi associado por sua respectiva matrícula e adicionados no *dataframe* final de 31.718 linhas.

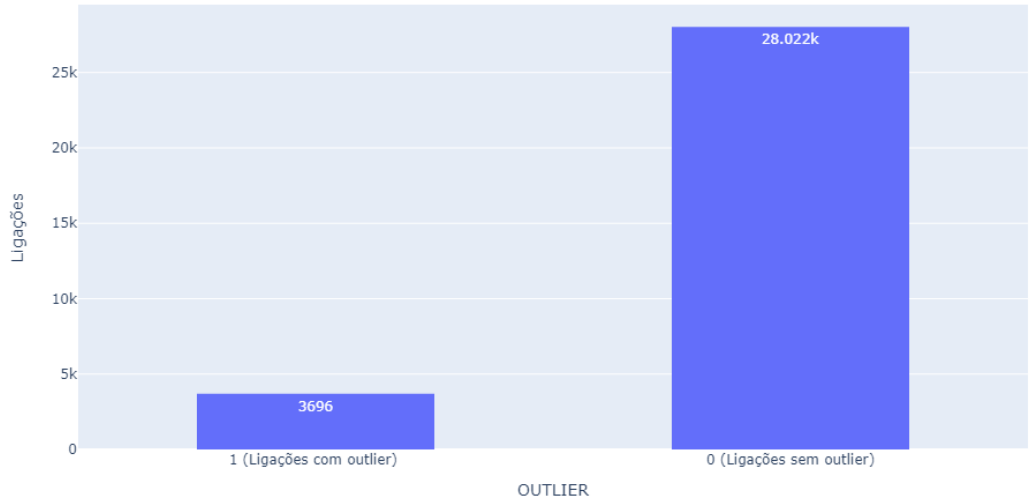
5.2.1 Análise dos volumes e características do HD

Analisando o consumo das ligações, foi criada uma função para identificar as ligações que possuem *outliers*, percorrendo todas as linhas do dataframe da Figura

19 para cada uma das ligações em função dos valores de consumo de 12 meses de cada coluna. Para isso, a função armazena em uma nova coluna o valor 1 caso a condição seja atendida e 0 caso não ocorra. Utilizou-se da regra de 3*IQR como limite superior, ampliando o IQR, visto que o limite mínimo não pode ser zero.

A quantidade de ligações totais com outlier nos 12 meses de consumo pode ser observado na Figura 20.

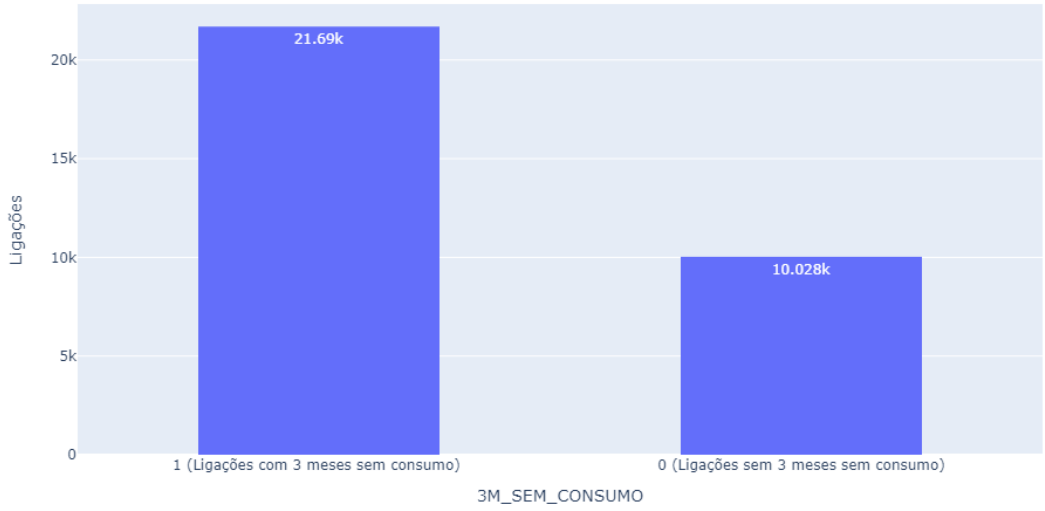
Figura 20 – Quantidade de ligações com outlier nos 12 meses de consumo.



Fonte: (Autor, 2023).

Em seguida, após identificado os *outliers*, foram calculados o VC médio e VF médio de cada uma dessas ligações, sendo criada uma nova coluna para cada um deles. Continuando a análise do VC, também se fez relevante criar uma função para identificar a ocorrência de três meses consecutivos com consumo igual 0 m³ ou consumo não registrado representado pela variável *NaN* (Not a Number). Realizando a mesma categorização binária explicitada na identificação dos *outliers*, armazenando, assim, esse novo recurso em uma coluna do *dataframe* de volume consumido. A quantidade de ligações totais com outlier nos 12 meses de consumo pode ser observado na Figura 21.

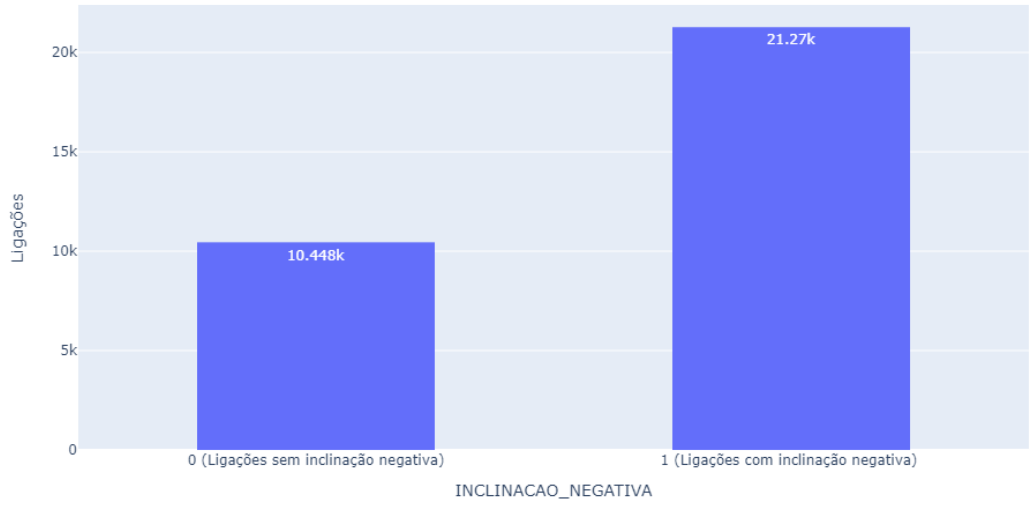
Figura 21 – Quantidade de ligações com três meses consecutivos de consumo igual a 0 m³.



Fonte: (Autor, 2023).

Por fim, calculou-se a inclinação da linha de tendência dos valores de VC de cada uma das ligações para os 12 meses de registro, com o objetivo de identificar as ligações cujo consumo apresenta um decréscimo. Recurso esse que também foi armazenado em uma nova coluna que identifica quando cada ligação possui inclinação negativa ou não. A quantidade de ligações totais com outlier nos 12 meses de consumo pode ser observado na Figura 22.

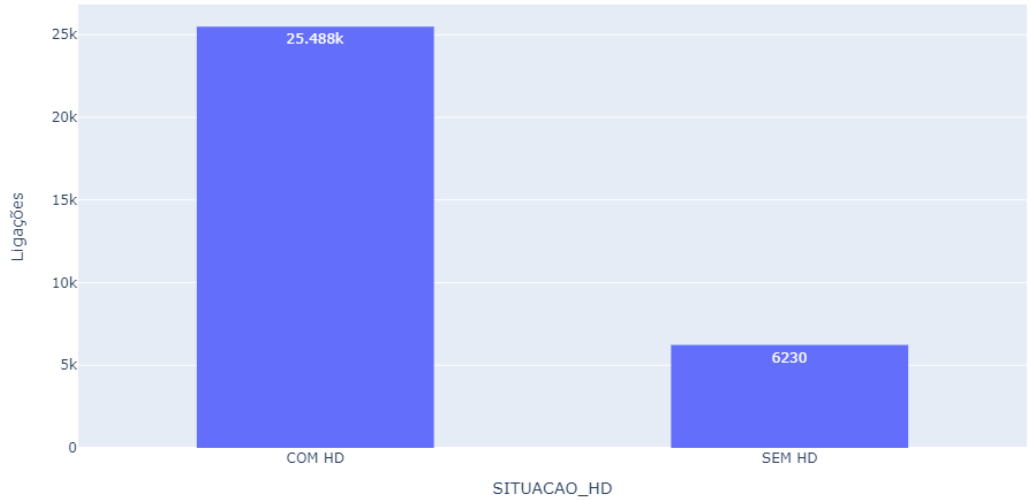
Figura 22 – Quantidade de ligações com inclinação negativa.



Fonte: (Autor, 2023).

Ademais, foi criou-se a coluna “SITUACAO_HD”, identificando se a ligação possui hidrômetro, através de uma checagem da condição de existência da data da instalação do HD registrada na coluna “Dt Inst Hidrometro” já existente no banco de dados comercial. A quantidade de ligações com e sem hidrômetro pode ser observada na Figura 23.

Figura 23 – Quantidade de ligações totais com e sem HD.

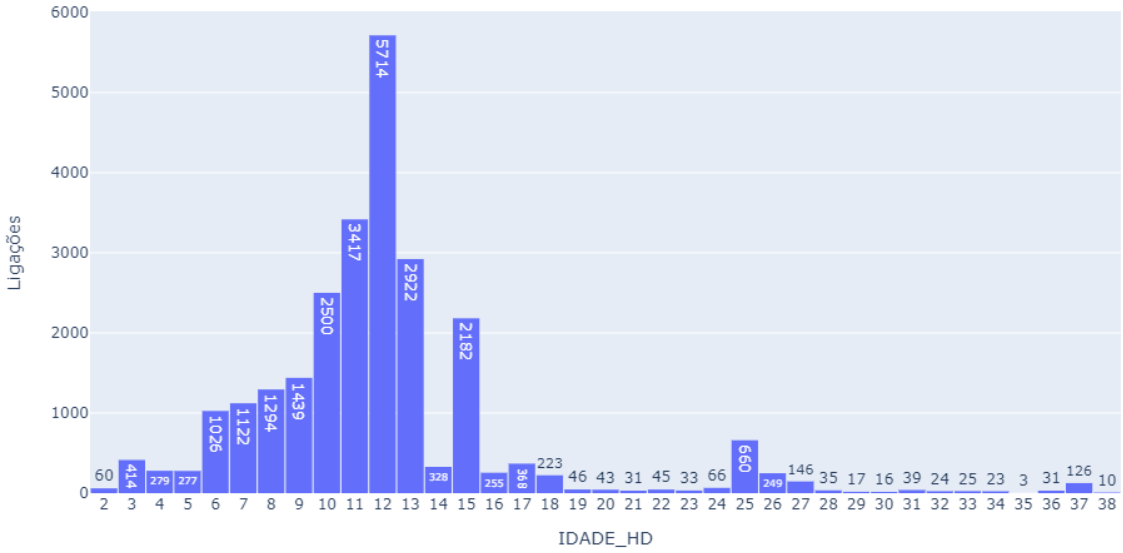


Fonte: (Autor, 2023).

Por último, foram mantidas apenas as ligações com HD (25.488 ligações), de modo a tornar possível o cálculo de suas respectivas idades. Para tanto, foi criada a coluna “IDADE_HD” que recebeu a informação de idade atual do HD em anos, para cada ligação a partir da coluna com a data de instalação presente na base de dados.

A idade de todas as matrículas pode ser observada na Figura 24. Evidencia-se que a maior concentração dos hidrômetros está entre os 10 e 15 anos de idade.

Figura 24 – Quantidade de ligações por idade do HD.



Fonte: (Autor, 2023).

Cada um desses recursos criados foi adicionado no *dataframe* final supracitado, representado pela Figura 18, em cada uma das suas matrículas únicas, com o objetivo de tornar possível a geração de cada quantitativo. Os recursos adicionais criados até o momento para todas as ligações estão representados, em síntese, na Tabela 6.

Tabela 6 – Resumo dos recursos adicionais criados.

Colunas	Informação
VOL_CONSUMIDO_MEDIO	Valor do VC médio.
VOL_FATURADO_MEDIO	Valor do VF médio.
OUTLIER	1 para ligações com outlier nos valores de consumo e 0 para ligações sem.
3M_SEM_CONSUMO	1 para ligações com três meses consecutivos sem consumo e 0 para ligações sem.
INCLINACAO_NEGATIVA	1 para ligações com inclinação negativa e 0 para ligações sem.
SITUACAO_HD	Ligações com ou sem HD
IDADE_HD	Idade do HD

Fonte: (Autor, 2023).

Cada um desses recursos criados são relevantes e por si só já descrevem características das ligações que podem apresentar uma maior probabilidade de fraude, assim como apresentado na Tabela 1. Desse modo, dado a relevância das informações, os recursos serão avaliados e posteriormente utilizados para ajudar o modelo a identificar as fraudes.

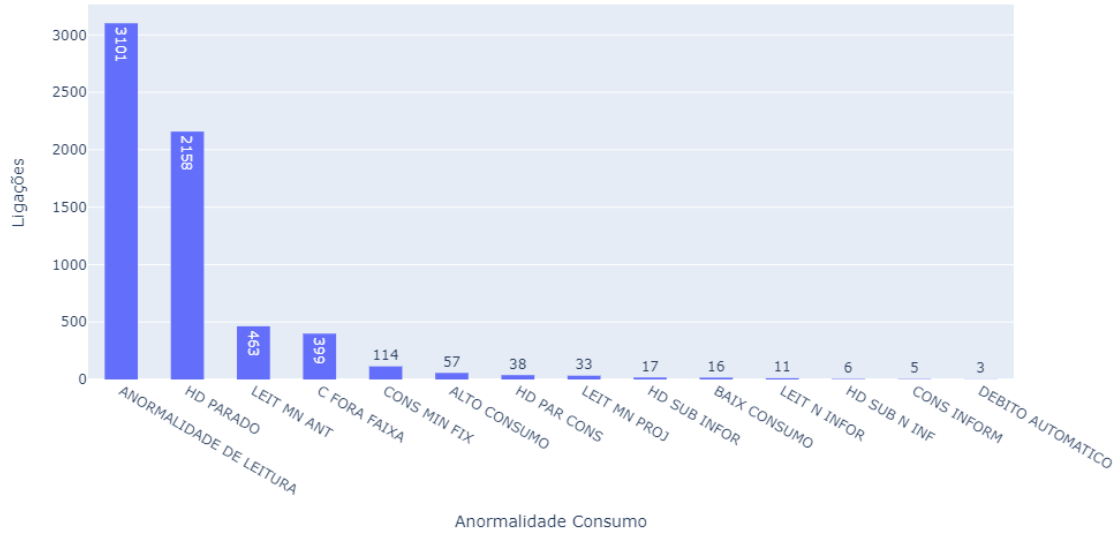
5.2.2 Análise das anormalidades

Existe, na base de dados, a coluna “Anormalidade Consumo” identificando cada tipo de anormalidade relacionada ao consumo presente nas ligações quando há ocorrência.

Primeiramente, analisou-se o último mês de modo a observar como cada uma das anormalidades se distribui e quais foram as mais recentes identificadas de modo a selecionar quais delas apresentam um maior indicativo de fraude. Na Figura 25 é possível observar o gráfico da quantidade de ligações por anormalidade de consumo registradas no último mês.

Destaca-se a anormalidade de leitura do consumo e HD parado. A anormalidade de leitura presente na coluna de anormalidades de consumo generaliza algumas das anormalidades presentes na coluna “Anormalidade Leitura”. Entretanto, não é regra que a anormalidade de leitura, qualquer que seja ela, será, também, concomitantemente registrada na coluna de anormalidade de consumo.

Figura 25 – Quantidade de ligações por anormalidade de consumo no mês de fevereiro de 2021.

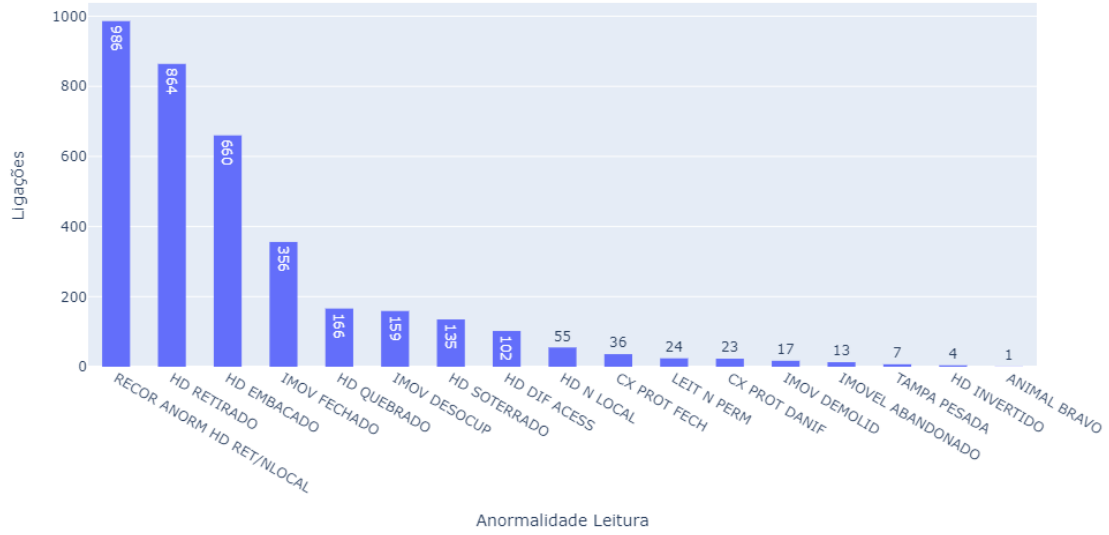


Fonte: (Autor, 2023).

Desse modo, ambas as colunas de anormalidade de consumo e leitura devem ser levadas em consideração, principalmente quanto a sua frequência dentro do intervalo do registro mensal de 12 meses de cada ligação. Além disso, é importante ressaltar que essas informações foram obtidas em campo pelos leituristas.

Na Figura 26, faz-se possível observar o gráfico da quantidade de ligações por anormalidade de leitura registradas no último mês, destacando-se o HD retirado e sua recorrência.

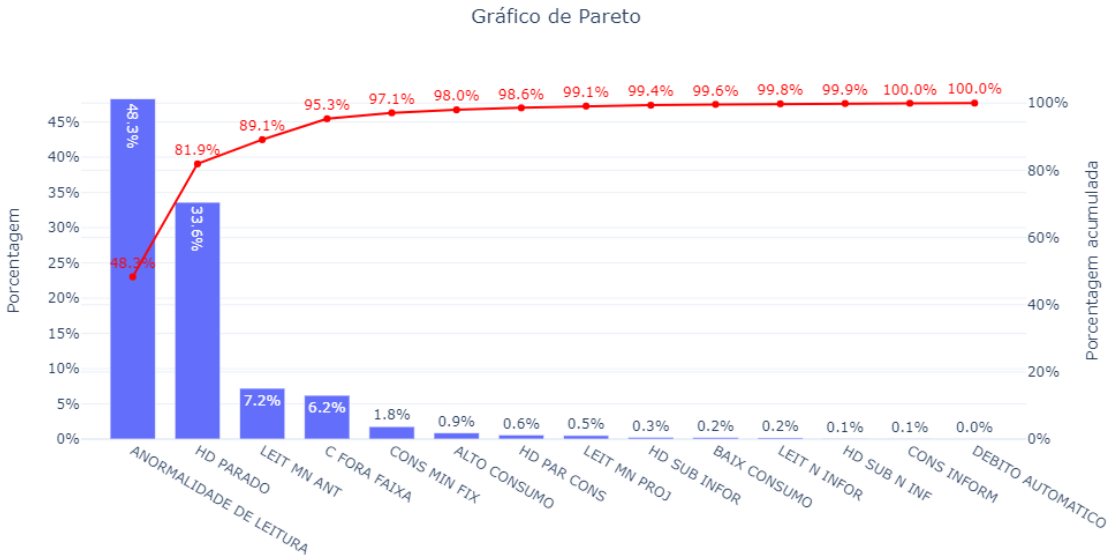
Figura 26 – Quantidade de ligações por anormalidade de leitura no mês de fevereiro de 2021.



Fonte: (Autor, 2023).

Sabendo que o gráfico de Pareto permite analisar a frequência de ocorrência de diferentes itens em um conjunto de dados, ordenando-os da maior para a menor frequência, tornando-se uma ferramenta útil para identificar os principais problemas, falhas ou oportunidades em um determinado processo ou sistema, fez-se relevante demonstrar, através da Figura 27, a porcentagem acumulada de cada uma das anormalidades de consumo.

Figura 27 – Gráfico de Pareto da quantidade de ligações por anormalidade de consumo no mês de fevereiro de 2021.



Fonte: (Autor, 2023).

Dentre as anormalidades de consumo e leitura apresentadas, faz-se relevante destacar aquelas principais que indicam uma maior probabilidade de ocorrência de fraude, sendo elas explicitadas, respectivamente na Tabela 7 e Tabela 8.

Tabela 7 – Anormalidades de consumo que indicam uma maior probabilidade de ocorrência de fraude.

Anormalidades de consumo	Abreviação
Hidrômetro parado	HD PARADO
Leitura menor que a anterior	LEIT MN ANT
Baixo consumo	BAIX CONSUMO
Leitura menor que a projetada	LEIT MN PROJ

Fonte: (Autor, 2023).

Tabela 8 – Anormalidades de leitura que indicam uma maior probabilidade de ocorrência de fraude.

Anormalidades de leitura	Abreviação
Hidrômetro não localizado	HD PARADO
Hidrômetro retirado	LEIT MN ANT
Recorrente anormalidade – hidrômetro retirado/não localizado	RECOR ANORM HD RET/NLOCAL
Hidrômetro invertido	HD INVERTIDO
Caixa de proteção danificada	CX PROT DANIF
Hidrômetro quebrado	HD QUEBRADO

Fonte: (Autor, 2023).

Por conseguinte, uma vez destacadas as principais anormalidades a serem consideradas, foi possível analisar como um todo, levando em consideração os 12 meses de registro de anormalidades dos *dataframes* criados anteriormente, o principal indicativo de fraude da base de dados.

Para tanto, foi criada uma função para identificar quando ocorrem três meses seguidos de registro de alguma das quatro principais anormalidades de consumo da Tabela 7 que indicam uma maior probabilidade de fraude.

De maneira análoga, criou-se outra função para identificar quando ocorrem três meses seguidos de registro de alguma das seis principais anormalidades de leitura da Tabela 8 que indicam uma maior probabilidade de fraude, ou seja, será o indicativo de fraude utilizado como rótulo posteriormente como entrada de cada um dos modelos.

Com objetivo de tornar mais visual o registro de anormalidades mensal, evidencia-se, na Figura 28, o *dataframe* de anormalidades criado previamente cujas análises de maior probabilidade de fraude foram feitas.

Figura 28 – *Dataframe* de anormalidades representando o registro mensal de anormalidades por matrícula.

	Anormalidade Consumo_202003.0	Anormalidade Consumo_202004.0	Anormalidade Consumo_202005.0	Anormalidade Consumo_202006.0	Anormalidade Consumo_202007.0	Anormalidade Consumo_202008.0	Anormalidade Consumo_202009.0	Anc Consumc
Matricula								
6150411	ANORMALIDADE DE LEITURA	ANORMALIDADE DE LEITURA	ANORMALIDADE DE LEITURA	ANORMALIDADE DE LEITURA	ANORMALIDADE DE LEITURA	ANORMALIDADE DE LEITURA	ANORMALIDADE DE LEITURA	ANOR DI
6150420	None	None	None	None	None	None	None	
6150438	LEIT MN ANT	None	ANORMALIDADE DE LEITURA	ANORMALIDADE DE LEITURA	ANORMALIDADE DE LEITURA	ANORMALIDADE DE LEITURA	ANORMALIDADE DE LEITURA	ANOR DI
6150446	None	None	None	None	None	None	None	
6150454	ANORMALIDADE DE LEITURA	ANORMALIDADE DE LEITURA	ANORMALIDADE DE LEITURA	ANORMALIDADE DE LEITURA	ANORMALIDADE DE LEITURA	ANORMALIDADE DE LEITURA	ANORMALIDADE DE LEITURA	ANOR DI

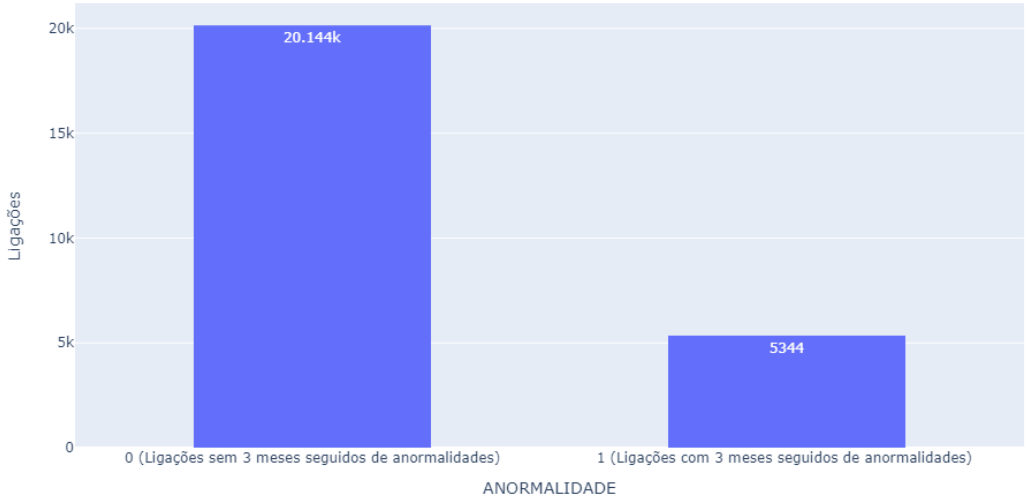
Fonte: (Autor, 2023).

Com isso, tornou-se possível criar a coluna “ANORMALIDADE” no *dataframe* final que identifica quando alguma ligação atendeu alguma das duas condições descritas anteriormente. Em outras palavras, caso a ligação possua três meses seguidos de registro de alguma das anormalidades de consumo ou três meses de registro de alguma das anormalidades de leitura, armazenou-se o valor 1 e, caso não ocorra, foi retornado o valor 0 na coluna.

O novo recurso criado que caracteriza a condição de provável fraude foi adicionado no *dataframe* final da Figura 18 em cada uma das suas matrículas únicas na forma de coluna.

Sendo assim, o estudo de fraudes será direcionado para a condição ter ou não essa anormalidade. Representando, assim, o rótulo de cada uma das ligações para a aplicação dos modelos supervisionados. Desse modo, torna-se possível observar, na Figura 29, a quantidade de ligações com três meses seguidos de anormalidade e sem três meses seguidos de anormalidade.

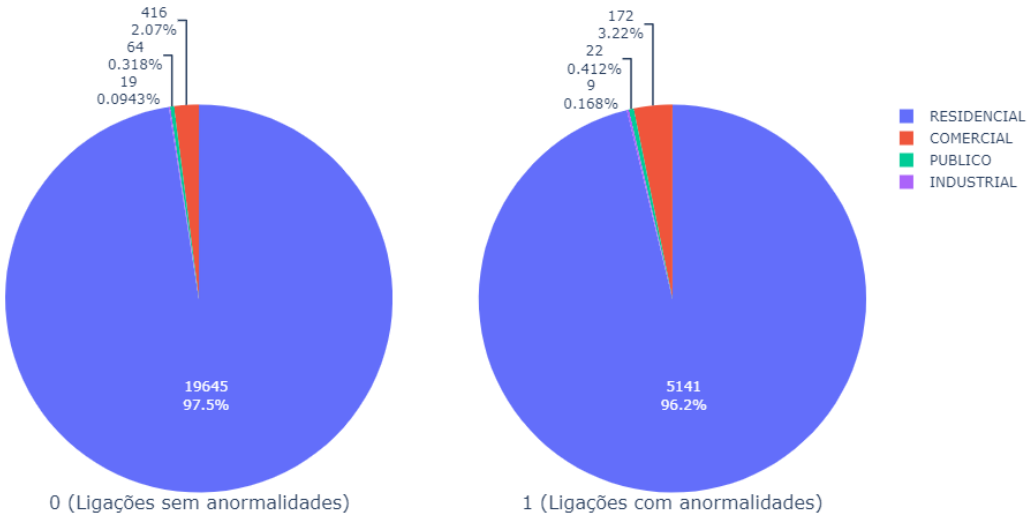
Figura 29 – Quantidade de ligações com e sem anormalidade



Fonte: (Autor, 2023).

Concomitantemente, a quantidade de ligações com e sem anormalidade por categoria pode ser observada na Figura 30.

Figura 30 – Quantidade de ligações com anormalidade e sem anormalidade por categoria



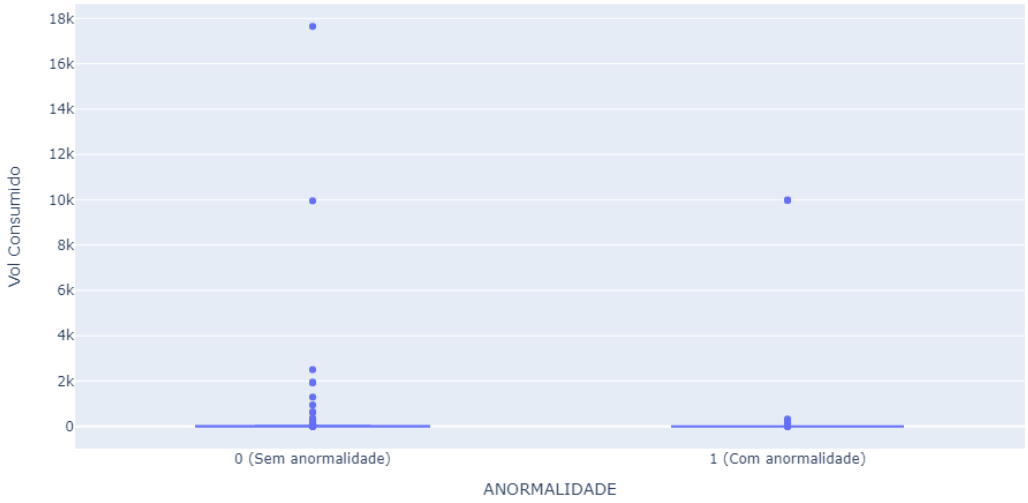
Fonte: (Autor, 2023).

Uma vez definido o rótulo de anormalidade, plotaram-se os *boxplots* com o objetivo de observar como se comporta o padrão da condição de anormalidade em relação a outras variáveis das ligações. Os gráficos com os boxplots de anormalidades de leitura do consumo em relação ao VC, VF, quantidade de economias e idade dos HDs podem ser vistos nas figuras abaixo.

Nota-se que nas Figura 31 e Figura 32, as duas primeiras variáveis analisadas, VC e VF, alguns valores *outliers* existem nessas variáveis, destoando tanto do limite superior ao ponto de achatar todo o *bloxplot* durante a plotagem.

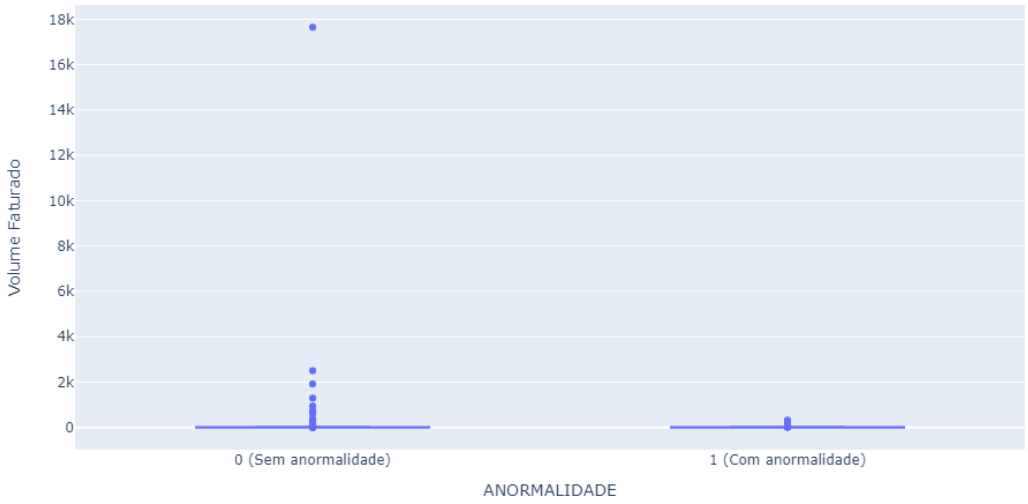
Nos *boxplots* em que não há anormalidade de leitura, evidencia-se a presença de alguns valores elevados de volume consumido e faturado. O valor mais extremo, de 17.649 m^3, corresponde a uma ligação industrial cujo volume consumido e faturado destoa bastante das outras categorias de ligações e por algumas ligações residenciais cujo hidrômetro está invertido. Além disso, os valores de volume consumido médio e volume faturado médio apresentam comportamento semelhante aos apresentados.

Figura 31 – Boxplots de volume consumido em relação à existência de anormalidade.



Fonte: (Autor, 2023).

Figura 32 – Boxplots de volume faturado em relação à existência de anormalidade.

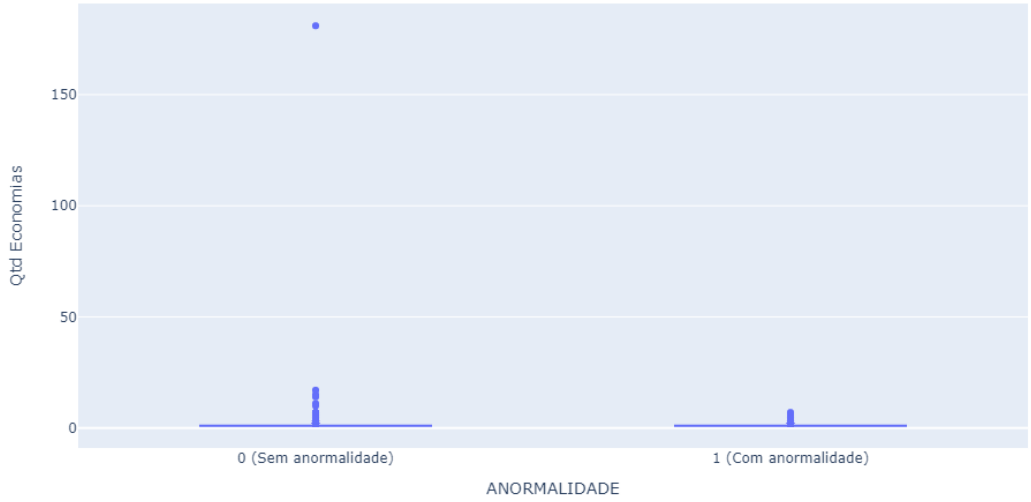


Fonte: (Autor, 2023).

Na Figura 33 percebe-se a existência de uma ligação cuja quantidade de economias (181 economias) destoa de maneira abrupta das demais amostras, ocasionada devido ao fato de ser uma ligação com perfil coorporativo, apesar de ser residencial.

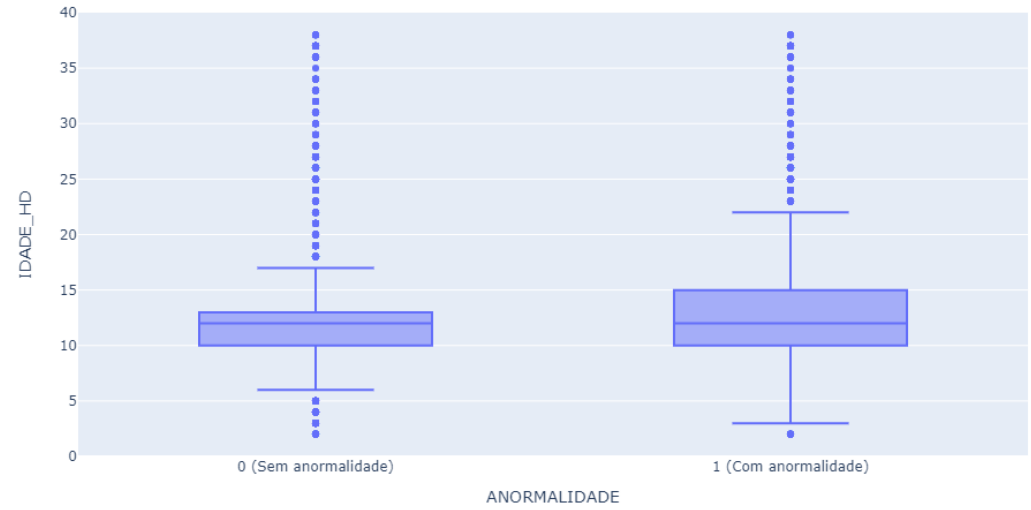
De modo geral, na Figura 34, nota-se uma distribuição relativamente distinta para as duas classes, fato que irá contribuir positivamente para o modelo de aprendizado de máquina realizado posteriormente. Apesar da existência de *outliers*, seus valores não destoam tanto do limite superior comparado as outras variáveis discorridas.

Figura 33 – Boxplots de quantidade de economias em relação à existência de anormalidade.



Fonte: (Autor, 2023).

Figura 34 – Boxplots de idade dos HDs em relação à existência de anormalidade.



Fonte: (Autor, 2023).

Nesse sentido, dependendo da correlação de cada uma dessas variáveis e do resultado da modelagem, será avaliada a necessidade de uma remoção dos outliers cujos valores são incoerentes ou que possuem uma frequência quase que única de modo a evitar que atrapalhem a criação do modelo de aprendizagem de máquina.

5.3 Pré-processamento

Prosseguindo para o pré-processamento dos dados, foram selecionadas as colunas a serem utilizadas para a criação do modelo. Dentre elas, foram substituídos

os valores ausentes da coluna “Vol Consumido” e “VOL_CONSUMIDO_MEDIO” por 0 para tornar possível seus respectivos usos – uma vez que só são aceitos apenas valores numéricos no modelo.

Por conseguinte, no que se refere as colunas com valores categóricos de texto, foram transformadas em variáveis *dummy*. As colunas “Perfil”, “Categoria Principal” e “Situacao Lig Agua” foram substituídas pelo número de colunas correspondente a quantidade de cada uma de seus valores categóricos de texto.

As colunas ou variáveis que foram utilizadas no modelo de aprendizagem de máquina podem ser observadas na Figura 35. Dentre elas, já é possível observar as variáveis *dummy* que foram criadas anteriormente. Além disso, destaca-se a quantidade de valores ausentes em cada uma das colunas, visto que é imprescindível que para a criação do modelo que não exista nenhum valor ausente.

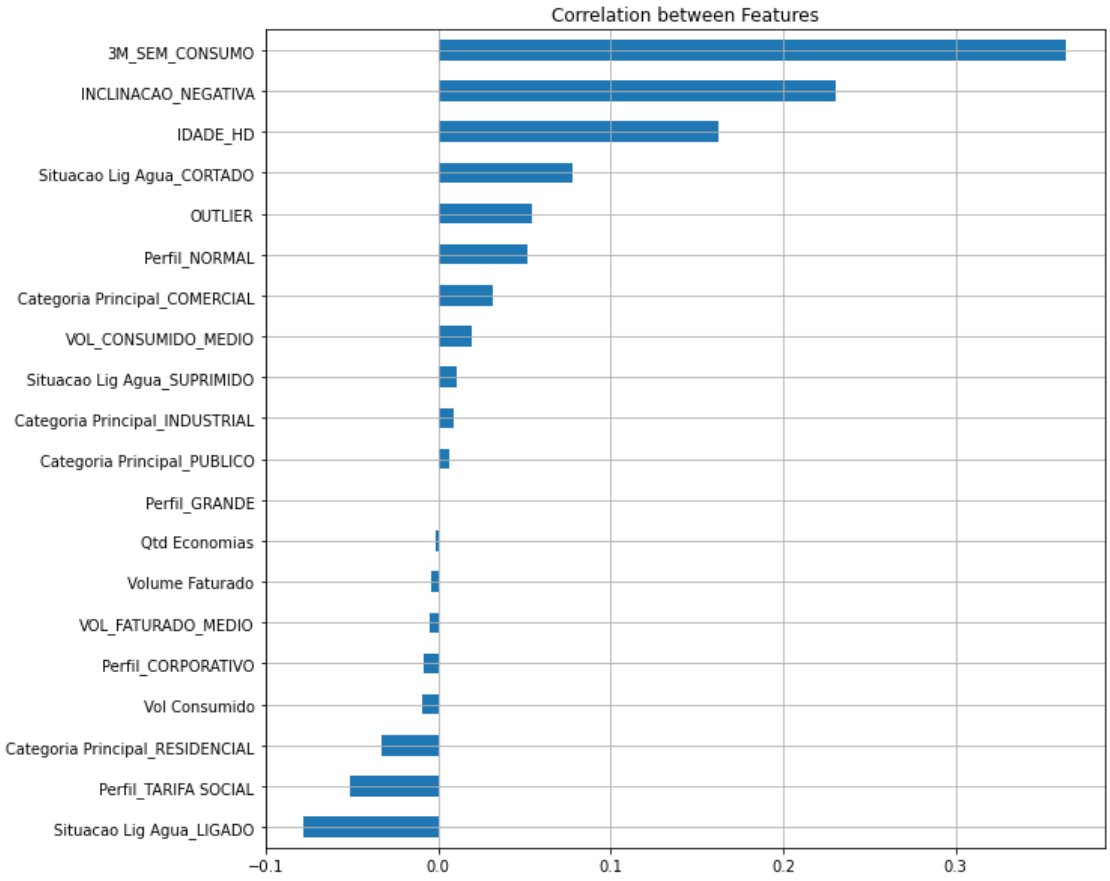
Figura 35 – Variáveis utilizadas no modelo de aprendizagem de máquina e suas respectivas quantidades de valores ausentes.

Variaveis	Valores Ausentes
Qtd Economias	0.0
Vol Consumido	0.0
Volume Faturado	0.0
VOL_CONSUMIDO_MEDIO	0.0
VOL_FATURADO_MEDIO	0.0
OUTLIER	0.0
3M_SEM_CONSUMO	0.0
INCLINACAO_NEGATIVA	0.0
ANORMALIDADE	0.0
IDADE_HD	0.0
Perfil_CORPORATIVO	0.0
Perfil_GRANDE	0.0
Perfil_NORMAL	0.0
Perfil_TARIFA SOCIAL	0.0
Categoria Principal_COMERCIAL	0.0
Categoria Principal_INDUSTRIAL	0.0
Categoria Principal_PUBLICO	0.0
Categoria Principal_RESIDENCIAL	0.0
Situacao Lig Agua_CORTADO	0.0
Situacao Lig Agua_LIGADO	0.0
Situacao Lig Agua_SUPRIMIDO	0.0

Fonte: (Autor, 2023).

Com isso, tornou-se possível estudar a correlação, através do coeficiente de Pearson, de cada uma das variáveis em relação à condição de anormalidade destacada, apresentando grande indicativo e probabilidade de fraude, assim como representado no gráfico da Figura 36.

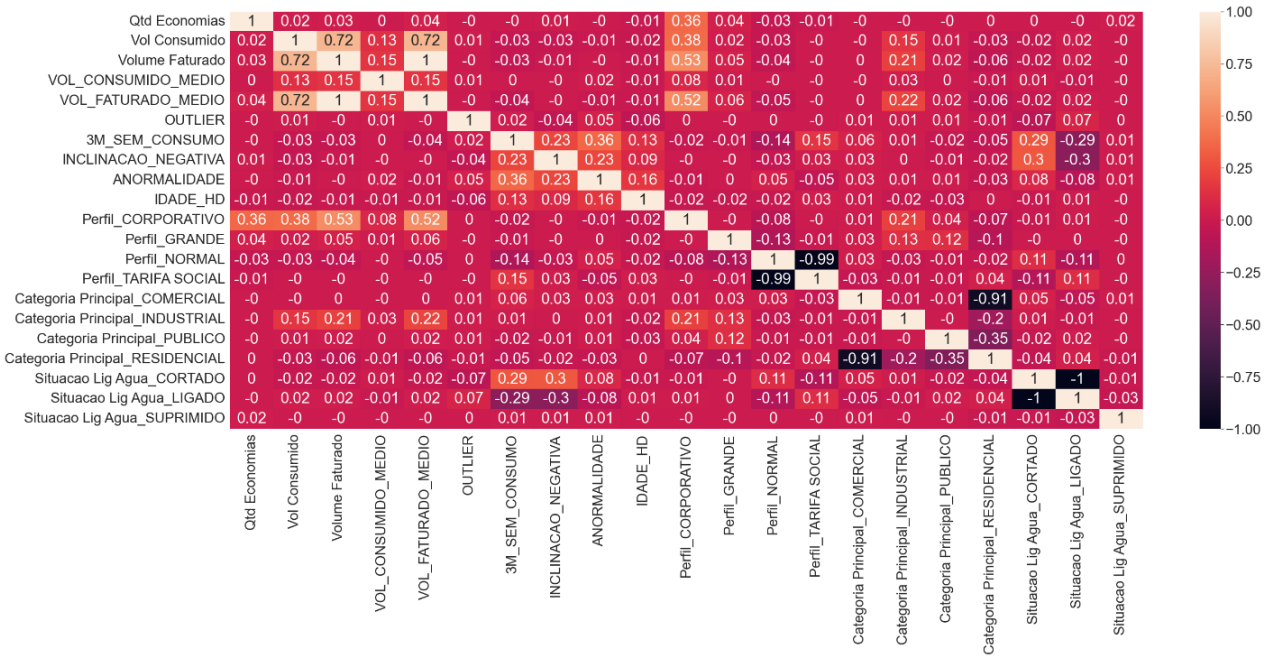
Figura 36 – Gráfico de correlação das variáveis em relação a anormalidade.



Fonte: (Autor, 2023).

Não obstante, também foi plotada a matriz de correlação, que mostra a correlação de cada uma das variáveis selecionadas entre si, que pode ser observada na Figura 37. Percebe-se que “3M_SEMCONSUMO”, “INCLINACAO_NEGATIVA”, “IDADE_HD”, “Situacao Lig Agua_CORTADO” e “OUTLIER” possuem os valores mais significativos. Todas as variáveis apresentadas, incluindo as de menor correlação, foram utilizadas para a criação dos modelos.

Figura 37 – Matriz de correlação de todas as variáveis selecionadas.



Fonte: (Autor, 2023).

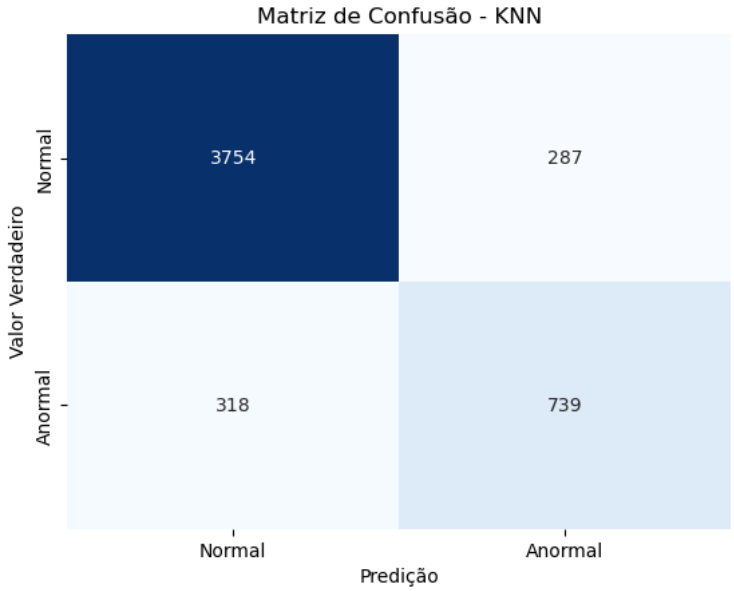
Quanto as variáveis de maior correlação, tem-se que a “3M_SEMCONSUMO” associa-se bem com a condição de provável fraude estudada a partir dos três meses consecutivos de anormalidade de consumo ou leitura. Entretanto, quanto as variáveis “INCLINACAO_NEGATIVA” e “IDADE_HD”, é importante esclarecer que, quando associadas, também podem caracterizar uma submedição dos hidrômetros e não necessariamente fraude.

5.4 Processamento e resultado dos modelos

Dessa forma, fazendo-se uso das bibliotecas destacadas na metodologia foram criadas cada uma das funções para executar os modelos a partir da divisão dos dados em treino e teste, tornando possível obter as matrizes de confusão para os modelos K Vizinhos Mais Próximos, Árvore aleatória e Floresta aleatória, que podem ser observados nas imagens abaixo.

O modelo K Vizinhos Mais Próximos alcançou uma acurácia na validação de 88,13% após seu processamento. Para uma análise mais detalhada, foi gerada uma matriz de confusão, apresentada na Figura 38. Essa matriz revela que para a classe 0 ou normal, o modelo previu corretamente 3.754 registros, porém errou em 287 registros. Já para a classe 1 ou anormal, o modelo obteve 739 previsões corretas, mas cometeu erros em 318 registros.

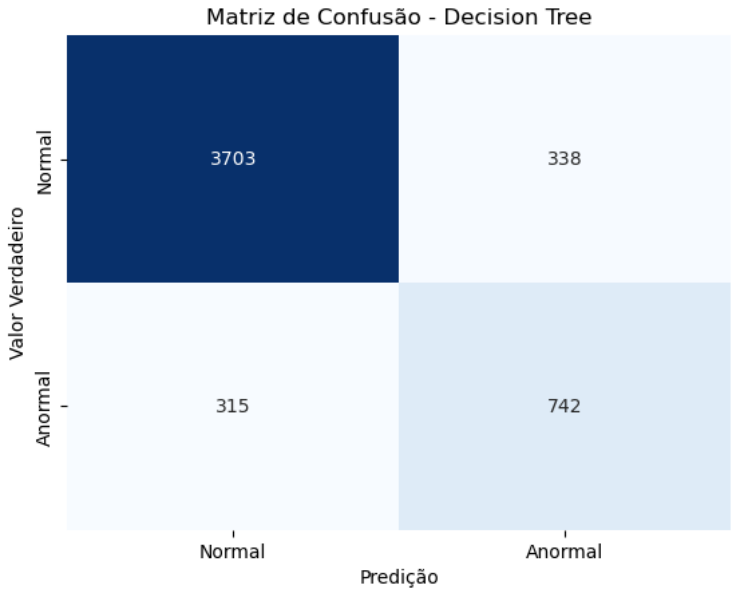
Figura 38 – Matriz de confusão para o modelo K Vizinhos Mais Próximos.



Fonte: (Autor, 2023).

O modelo Árvore de Decisão alcançou uma acurácia na validação de 87,19% após seu processamento. Para uma análise mais detalhada, foi gerada uma matriz de confusão, apresentada na Figura 39. Essa matriz revela que para a classe 0 ou normal, o modelo previu corretamente 3.703 registros, porém errou em 338 registros. Já para a classe 1 ou anormal, o modelo obteve 742 previsões corretas, mas cometeu erros em 315 registros. Portanto, o modelo se saiu melhor do que o anterior para classificar os casos da classe 1 e pior em classificar os casos da classe 0.

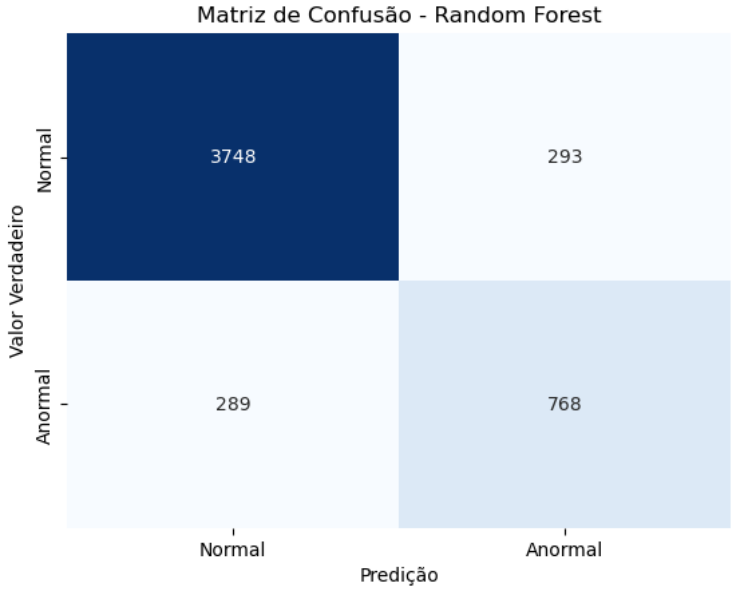
Figura 39 – Matriz de confusão para o modelo Árvore de Decisão.



Fonte: (Autor, 2023).

O modelo Floresta Aleatória alcançou uma acurácia na validação de 88,58% após seu processamento. Para uma análise mais detalhada, foi gerada uma matriz de confusão, apresentada na Figura 40. Essa matriz revela que para a classe 0 ou normal, o modelo previu corretamente 3.748 registros, porém errou em 293 registros. Já para a classe 1 ou anormal, o modelo obteve 768 previsões corretas, mas cometeu erros em 289 registros. Consequentemente, o modelo foi o que se saiu melhor em classificar os casos da classe 1 e da classe 0.

Figura 40 – Matriz de confusão para o modelo Floresta Aleatória.



Fonte: (Autor, 2023).

Cada uma das matrizes de confusão apresentadas possui a estrutura descrita na Tabela 2, seguido por um gradiente de cores para estilizar proporção numérica de valores identificados como TP, FN, FP e TN. Por conseguinte, o modelo Floresta Aleatória foi o que melhor performou, dentre os modelos analisados, com 88,58% acurácia.

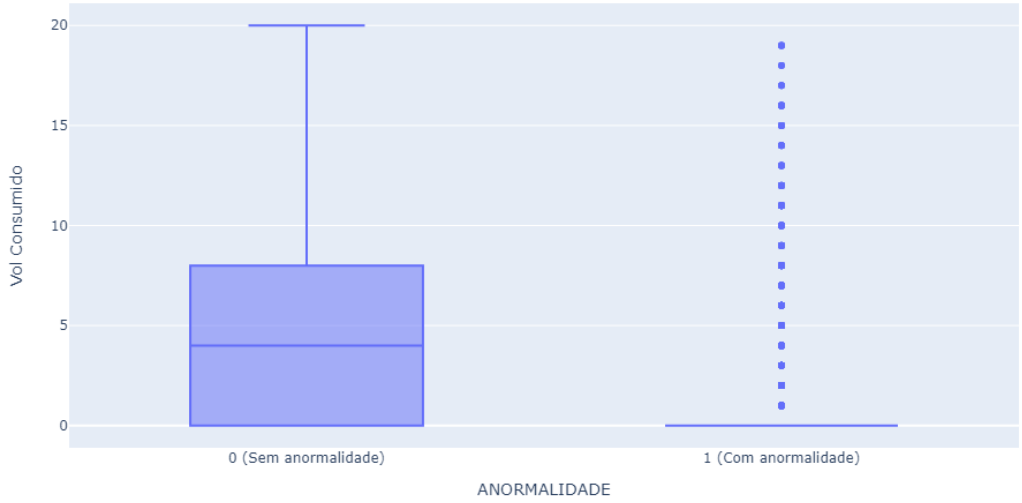
Nessa conjuntura, obteve-se um resultado relativamente semelhante para cada um dos três modelos, após a utilização dos dados bem tratados e limpos, sem valores ausentes ou variáveis categóricas de texto. Além disso, o modelo de aprendizagem de máquina Floresta Aleatória obteve a maior acurácia e, consequentemente, é o mais preciso dentre os três estudados.

5.5 Reprocessamento dos dados

Apesar dos resultados obtidos anteriormente já terem alcançado uma acurácia relativamente alta, ainda podem ser realizadas melhorias na base de dados. Diante desse cenário, foram removidas as ligações consideradas *outliers* de todas as colunas numéricas que não representam condições binárias ou variáveis *dummy*. Visto que tais ligações podem estar afetando o modelo e levando a conclusões erradas, apesar de elas representarem valores reais de ligações cadastradas e não dados errôneos.

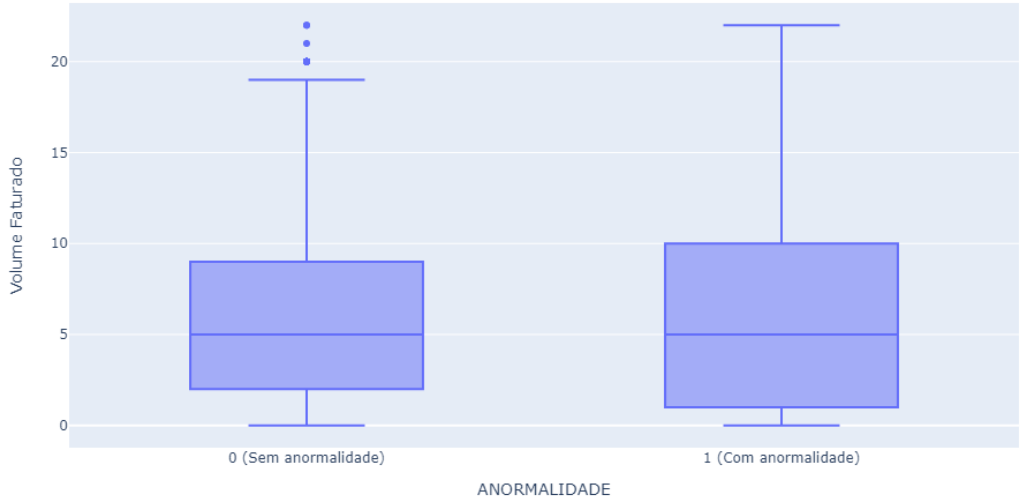
Foram removidas, no total, 4.576 matrículas que possuíam *outliers*. Os novos boxplots das variáveis numéricas sem outliers estratificados pela existência de anormalidade podem ser observados a seguir:

Figura 41 – Boxplots de volume consumido em relação à existência de anormalidade após a remoção dos outliers de volume consumido.



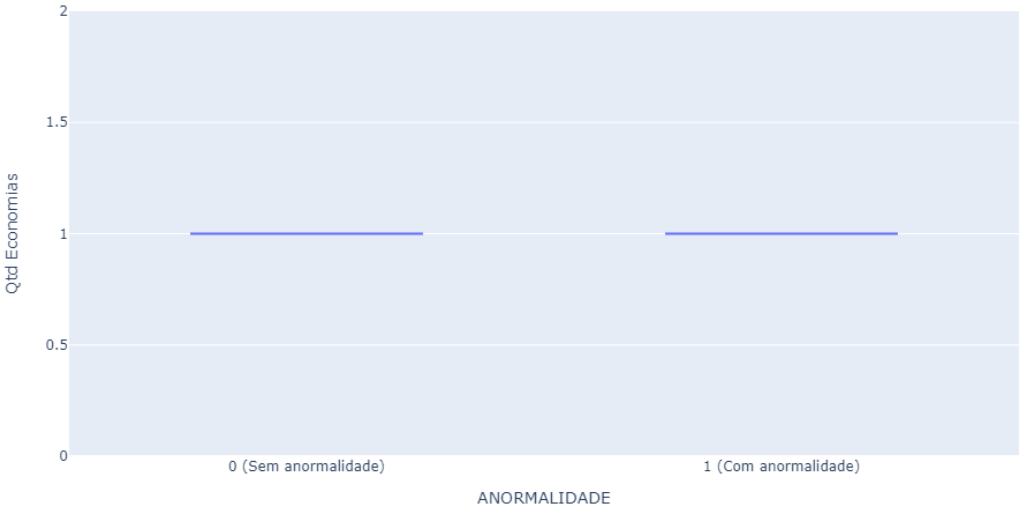
Fonte: (Autor, 2023).

Figura 42 – Boxplots de volume faturado em relação à existência de anormalidade após a remoção dos outliers de volume faturado.



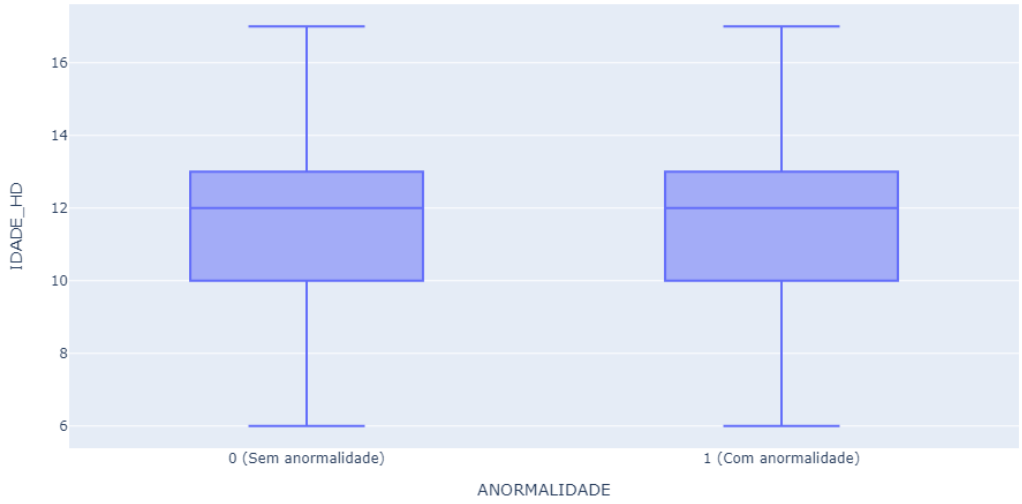
Fonte: (Autor, 2023).

Figura 43 – Boxplots de quantidade de economias em relação à existência de anormalidade após a remoção dos outliers da quantidade de economias.



Fonte: (Autor, 2023).

Figura 44 – Boxplots de quantidade de economias em relação à existência de anormalidade após a remoção dos outliers da quantidade de economias.



Fonte: (Autor, 2023).

Por fim, antes de realizar o processamento dos dados de entrada do modelo de aprendizagem, realizou-se a normalização de todas as variáveis destacadas na Figura 36 e Figura 37. Não houve necessidade de normalizar as variáveis *dummy* e os recursos criados visto que representam condições no formato binário já incluso no intervalo [0,1].

Após o reprocessamento e criação das novas matrizes de confusão, foram calculadas, novamente, as acurácias de cada um dos modelos trabalhados: KNN obteve 89,14% de acurácia, Árvore de Decisão atingiu 88,76% de acurácia e, por fim, Floresta Aleatória alcançou o maior valor de dentre os modelos analisados com 90,36% acurácia.

Uma vez que o modelo foi treinado, é possível realizar a identificação de fraudes em novas ligações, visto que o modelo supervisionado aprende a generalizar a partir dos exemplos de treinamento.

Todavia, ele não memoriza os dados de treinamento, mas sim aprende a reconhecer padrões gerais que são indicativos de fraudes. Portanto, mesmo que novas ligações fraudulentas não tenham sido vistas durante o treinamento, o modelo pode identificá-las com base nas características e nos padrões aprendidos.

Entretanto, é importante ressaltar que a identificação de novas fraudes por meio do modelo supervisionado não é infalível. O modelo pode exibir falsos positivos

(ou seja, classificar entrada legítima como fraude) ou falsos negativos (classificar fraude como legítima).

Portanto, é importante que o modelo seja constantemente atualizado e reavaliado à medida que novos dados são disponibilizados para melhorar sua precisão e garantir uma detecção de fraude mais eficaz. Além disso, é importante combinar a detecção automatizada – através dos algoritmos destacados –, com a análise humana para tomadas de decisão mais precisas e mitigação de erros.

5.6 Síntese comercial

Por meio de filtragens do *dataframe*, através de funções da biblioteca pandas para cada uma das demandas apresentadas, obtiveram-se os dados gerais dispostos na Tabela 9 do cadastro comercial do município de Vitória de Santo Antão referentes ao último mês dos dados disponibilizados (fevereiro de 2021).

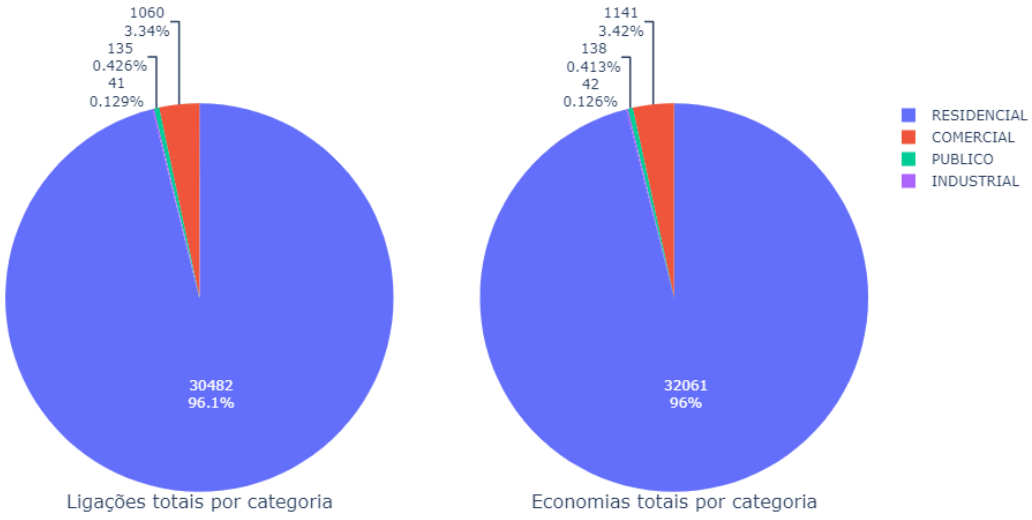
Tabela 9 – Dados gerais do cadastro comercial de Vitória de Santo Antão (02/2021).

Dados Gerais	Valor	Unidade
Ligações totais	31.718	ligações
Economias totais	33.382	economias
Ligações ativas	24.551	ligações
Economias ativas	25.720	economias
Volume consumido (Lig. totais)	188.078	m³/mês
Volume consumido (Lig. ativas)	175.549	m³/mês
Volume faturado (Lig. totais)	231.235	m³/mês
Volume faturado (Lig. ativas)	207.325	m³/mês
Média de idade de instalação do parque de hidrômetros (Lig. ativas)	12,04	anos

Fonte: (Autor, 2023).

De maneira a tornar a visualização mais clara e objetiva, foram plotados diversos gráficos, utilizando a biblioteca *plotly*, discorrendo mais profundamente cada um dos dados gerais apresentados. A Figura 45 apresenta, respectivamente, os percentuais da distribuição das ligações e economias nas diferentes categorias de uso da economia principal, que variam de acordo com a sua natureza de ocupação.

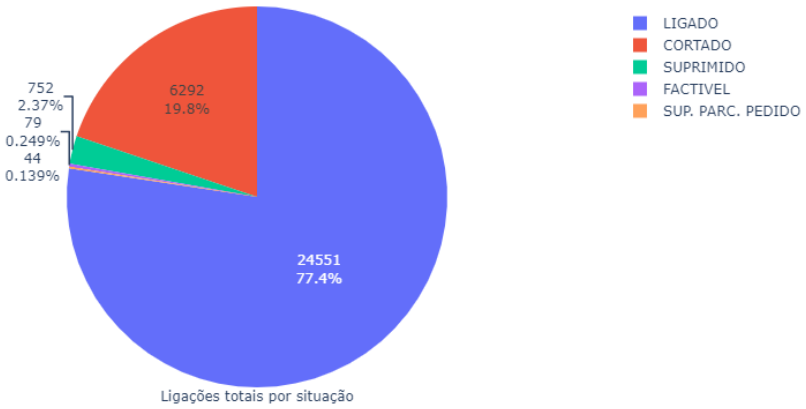
Figura 45 – Ligações e economias totais por categoria.



Fonte: (Autor, 2023).

As categorias comercial, público e industrial apresentam cerca de 4% da parcela total de ligações e economias totais por categoria. Obteve-se, também o percentual de ligações por situação da ligação de água representado pela Figura 46.

Figura 46 – Ligações totais por situação.



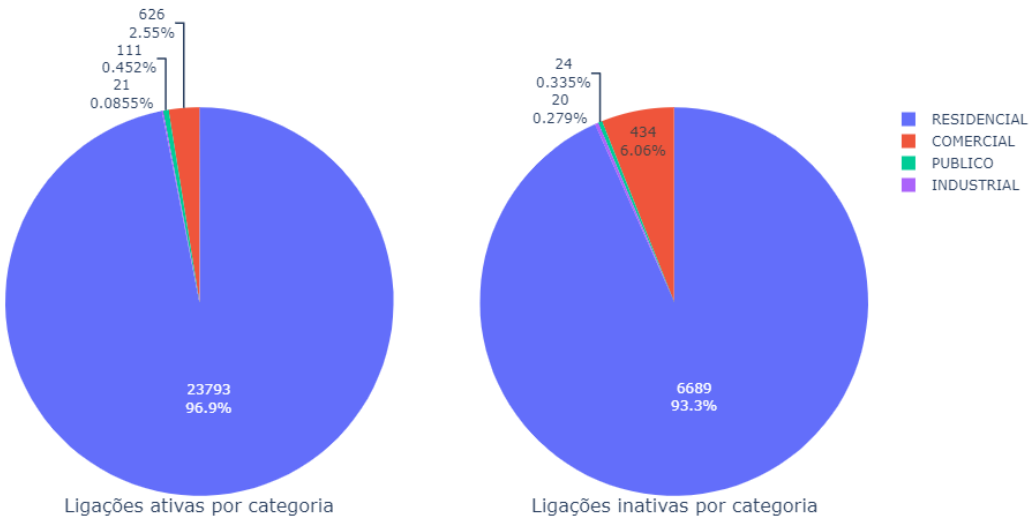
Fonte: (Autor, 2023).

Cerca de 20% das ligações totais apresentam a situação como cortada. Número esse que não deve ser ignorado, uma vez que uma ligação cortada não pode estar consumindo água, ou seja, um condicionante para fraude.

Com isso, separaram-se as ligações entre dois principais tipos de situação, sendo ativas quando possuírem o valor “LIGADO” na coluna “Situacao Lig Agua” do *dataframe* e inativas para os demais valores “CORTADO”, “SUPRIMIDO”, “SUP. PARC. PEDIDO (SUPRIMIDO PARCIALMENTE A PEDIDO)” e “FACTÍVEL” da mesma coluna. A Figura 47 expõe o percentual ligações ativas por categoria e o percentual de ligações inativas por categoria respectivamente.

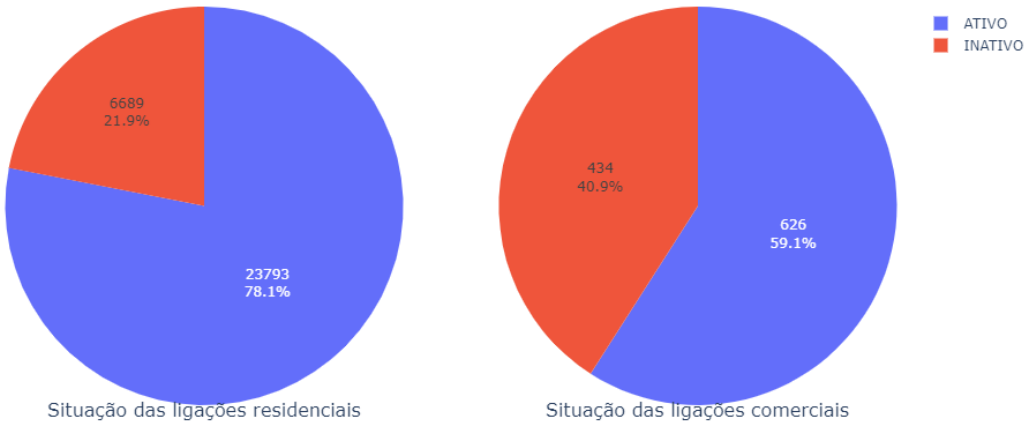
Por conseguinte, extraiu-se a quantidade de ligações do universo por situação para cada um dos tipos de categorias existentes. Os percentuais de situação das ligações residenciais e situação das ligações comercias podem ser observados na Figura 48, enquanto que os percentuais de situação das ligações públicas e situação das ligações industriais podem ser observados na Figura 49.

Figura 47 – Ligações ativas por categoria e ligações inativas por categoria.



Fonte: (Autor, 2023).

Figura 48 – Situação das ligações residências e situação das ligações comerciais.

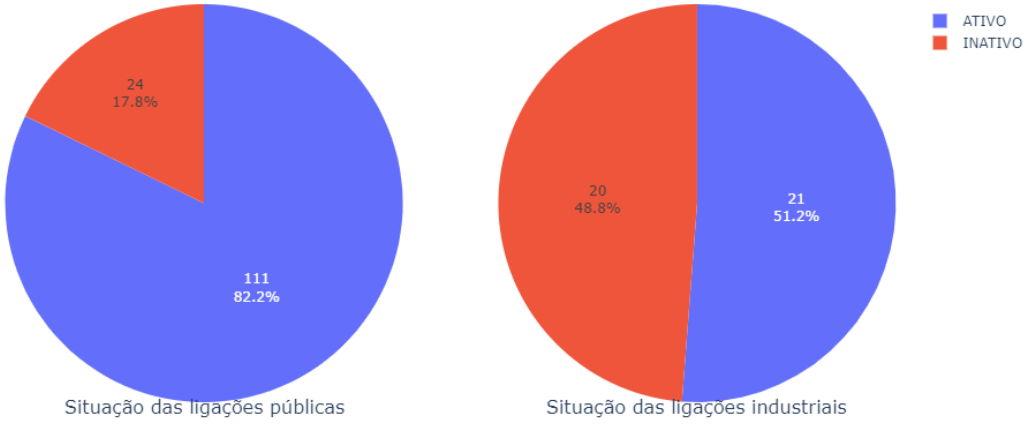


Fonte: (Autor, 2023).

O percentual de ligações cuja situação é ativa sempre se sobressai em todas as quatro categorias apresentadas. Entretanto, para as ligações industriais e comerciais há um grande aumento nas ligações com situação inativa.

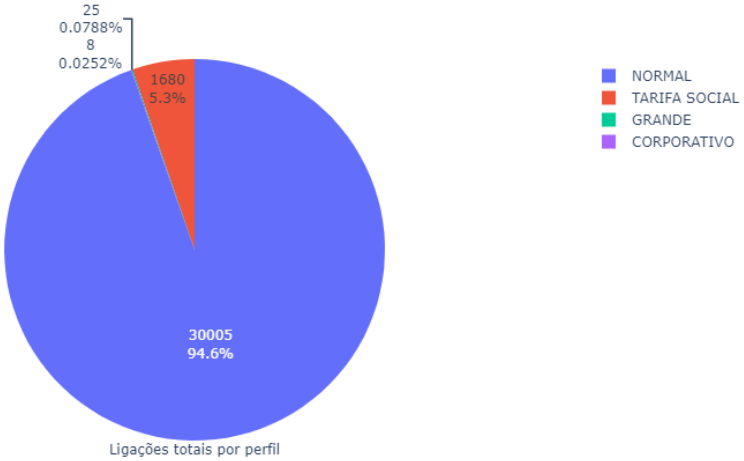
Também foram obtidas as quantidades de ligações totais por cada um dos perfis existentes e, posteriormente, analisada para ligações ativas residenciais e não residências. Os percentuais de ligações totais por perfil podem ser observados na Figura 50

Figura 49 – Situação das ligações públicas e situação das ligações industriais.



Fonte: (Autor, 2023).

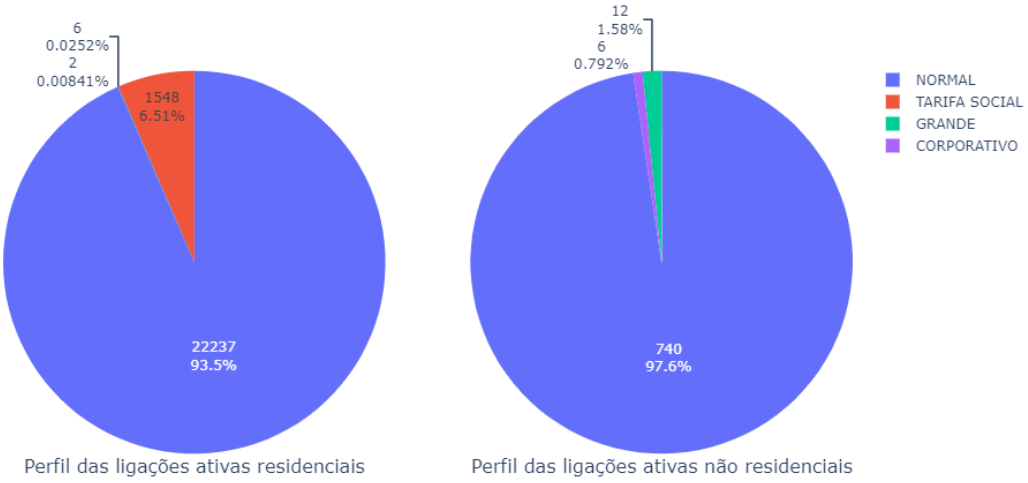
Figura 50 – Ligações totais por perfil.



Fonte: (Autor, 2023).

Concomitantemente, os percentuais de ligações ativas residências e não residenciais por perfil podem ser observados na Figura 51.

Figura 51 – Perfil das ligações ativas residências e perfil das ligações ativas não residenciais.



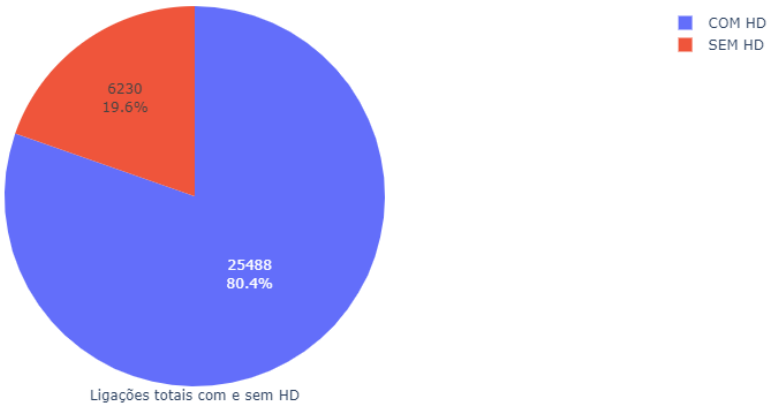
Fonte: (Autor, 2023).

Grande parte das ligações que apresentam tarifa social são residenciais, visto que se trata de uma política pública que busca garantir o acesso à água para as famílias em situação de vulnerabilidade socioeconômica. A tarifa social é um desconto aplicado na conta de água dessas famílias, com o objetivo de tornar o consumo mais acessível.

Uma vez que já foram obtidas as ligações com e sem HD durante a etapa de tratamento dos dados, tornou-se possível obter os seus valores percentuais para

ligações totais e os valores de categoria para cada uma dessas condições. O percentual de ligações totais com e sem HD pode ser observado na Figura 52.

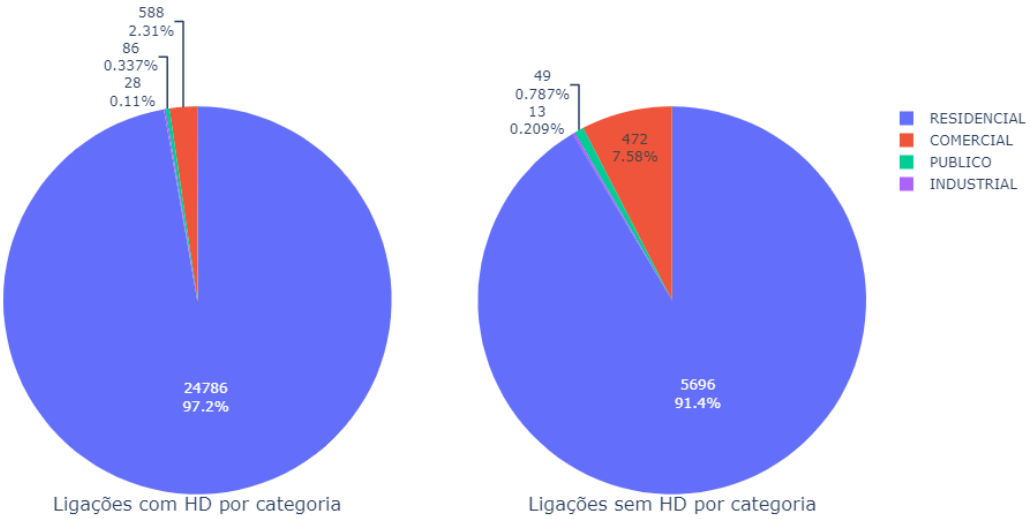
Figura 52 – Ligações totais com e sem HD.



Fonte: (Autor, 2023).

Não obstante, os percentuais de ligações com HD por categoria e sem HD por categoria podem ser observados na Figura 53.

Figura 53 – Ligações com HD por categoria e Ligações sem HD por categoria.

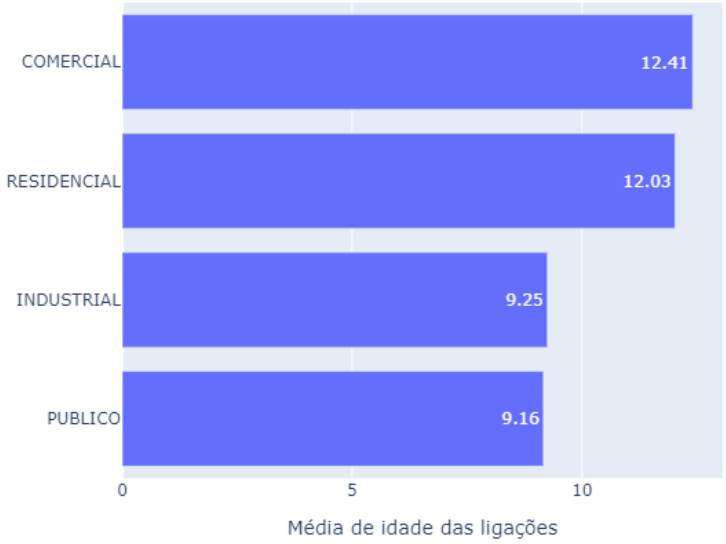


Fonte: (Autor, 2023).

As ligações comerciais, públicas e industriais apresentam juntas cerca de 10% das ligações sem HD. Além disso, foram calculadas as médias das idades de

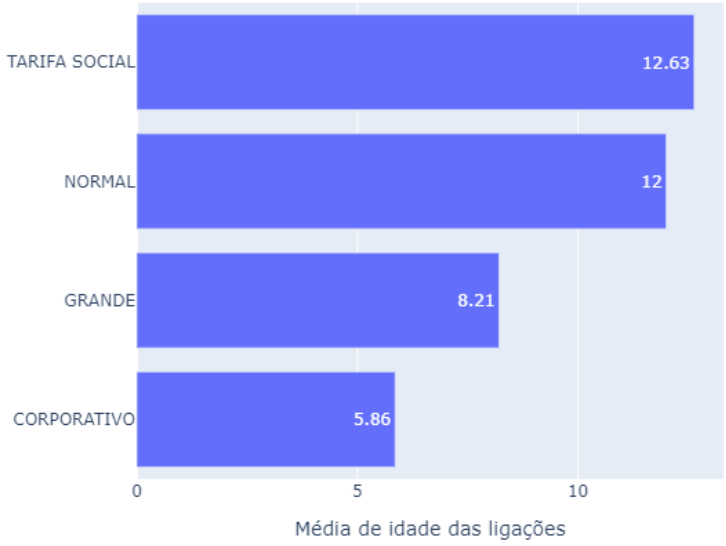
instalação dos HDs das ligações ativas por categoria e perfil, apresentadas, respectivamente na Figura 54 e na Figura 55.

Figura 54 – Média de idade das ligações por categoria.



Fonte: (Autor, 2023).

Figura 55 – Média de idade das ligações por perfil.

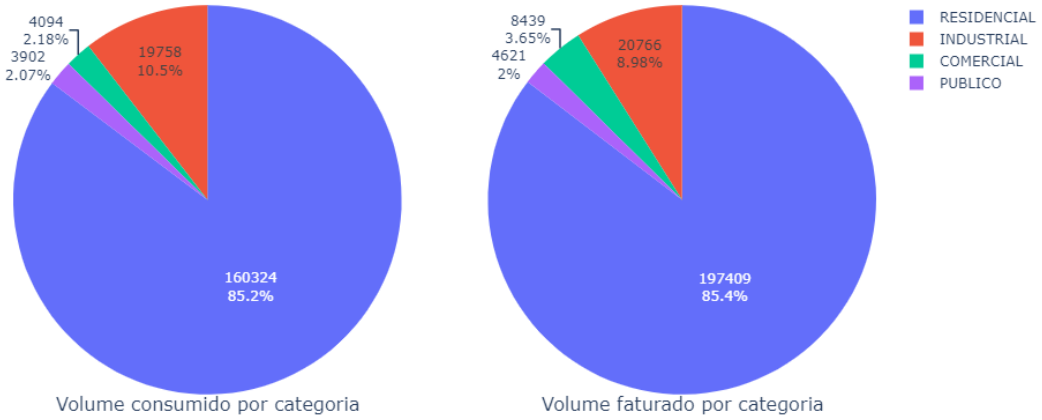


Fonte: (Autor, 2023).

As ligações de categoria comercial e industrial, assim como perfil normal e de tarifa social, apresentam HDs com uma média de 12 anos de idade de instalação. Devido a isso, grande parte dos HDs existentes do parque já ultrapassaram o prazo de substituição de 5 anos. Contribuindo de maneira efetiva para a submedição devido ao desgaste do medidor.

Por fim, de modo a observar como os volumes se distribuem, foram calculados os percentuais de volume consumido e faturado por categoria, explicitados na Figura 56.

Figura 56 – Volume consumido por categoria e volume faturado por categoria.

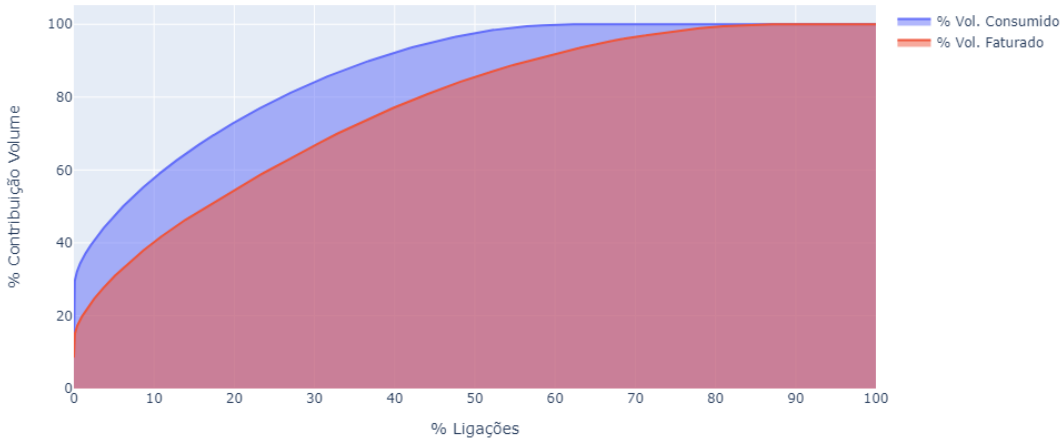


Fonte: (Autor, 2023).

As ligações industriais apresentam cerca de 10% do volume consumido mesmo com uma quantidade menor de ligações, devido ao alto consumo comparado a ligações residenciais.

Na Figura 57, que representa a curva ABC das ligações ativas com a contribuição do volume consumido e faturado, torna-se possível notar que os dois volumes crescem, inicialmente, de maneira análoga. Entretanto, o volume consumido atinge os 100% muito antes do volume faturado. Fato que pode ser explicado pela enorme quantidade de ligações com um consumo de 0 m³, porém possuem um volume faturado mesmo sendo mínimo.

Figura 57 – Curva ABC do volume consumido e faturado por ligação ativa



Fonte: (Autor, 2023).

Devido ao fato de que a estratificação dos dados é de extrema valia para a realização de comparações entre grupos de informações, tornando possível ter-se uma análise mais completa e abrangente dos dados, fez-se relevante estratificar as ligações do *dataframe*.

Criou-se uma função condicional para classificar cada uma das ligações totais em faixas de consumo a partir da coluna “Vol Consumido”. Foram consideradas apenas as ligações que apresentam dados de volume consumido. Porém, as ligações cuja informação é ausente foram identificadas com a variável “S/D” para representar a condição de estarem sem dados.

Dessa forma, de modo a elaborar um resultado semelhante as tabelas dinâmicas do Excel, também conhecidas como *pivot tables*, utilizou-se uma função da própria biblioteca pandas para criar *pivot tables* a partir do dataframe e agrupar cada uma das colunas de informações desejadas na tabela de resultados por faixa de consumo, atribuindo um somatório para o volume consumido, volume faturado, quantidade de ligações, e efetuada a média da idade de hidrômetros.

As Figuras a seguir apresentam as tabelas de resumo e a estratificação das ligações ativas por faixa de volume consumido, assim como o percentual da contribuição da quantidade de Ligações (Lig.), Volume consumido (VC) e Volume faturado (VF). Além disso, também se fez relevante expor o percentual acumulado por faixa de consumo.

a) [0-10]: refere-se às ligações com consumo entre 0 e 10 m³/mês;

b) (10-20]: refere-se às ligações com consumo entre 10 e 20 m³/mês;

c) (20-50]: refere-se às ligações com consumo entre 30 e 100 m³/mês;

d) (50-100]: refere-se às ligações com consumo entre 50 e 100 m³/mês;

e) (100-150]: refere-se às ligações com consumo entre 100 e 150 m³/mês;

f) (150-200]: refere-se às ligações com consumo entre 150 e 200 m³/mês;

g) (200-300]: refere-se às ligações com consumo entre 200 e 300 m³/mês;

h) (300-500]: refere-se às ligações com consumo entre 300 e 500 m³/mês;

i) (500-1000]: refere-se às ligações com consumo entre 500 e 1.000 m³/mês;

j) (1000-20000]: refere-se às ligações com consumo entre 1.000 e 20.000 m³/mês;

k) S/D: refere-se às ligações sem dados de consumo.

Figura 58 – Estratificação das ligações ativas por faixa de consumo.

FAIXA_CONSUMO	Média da Idade	Volume Consumido (VC)	Volume Faturado (VF)	Quantidade de Ligações
a) [0-10]	11.74	57 265 m³	74 656 m³	13 861
b) (10-20]	10.98	43 751 m³	43 653 m³	3 125
c) (20-50]	10.44	18 730 m³	18 331 m³	685
d) (50-100]	8.80	3 259 m³	2 561 m³	49
e) (100-150]	9.45	1 325 m³	1 238 m³	11
f) (150-200]	8.00	875 m³	648 m³	5
g) (200-300]	8.83	1 520 m³	941 m³	6
h) (300-500]	9.25	1 336 m³	1 424 m³	4
i) (500-1000]	3.67	2 203 m³	2 203 m³	3
j) (1000-20000]	7.14	45 285 m³	23 434 m³	7
k) S/D	15.36	0 m³	38 236 m³	6 795
TOTAL	12.04	175 549 m³	207 325 m³	24 551

Fonte: (Autor, 2023).

Os maiores valores de VC, VF e quantidade de ligações está concentrado na faixa de consumo de 0 a 10 m³/mês. Além disso a faixa em questão também possui a segunda maior média de idade de instalação dos HDs. Entretanto, há uma grande quantidade de ligações que não há informação de volume consumido, ligações essas antigas vistos que possuem a maior média na idade de instalação dos HDs.

Figura 59 – Estratificação das ligações ativas por faixa de consumo.

FAIXA_CONSUMO	%VC	%VF	%Ligações
a) [0-10]	32.62 %	36.01 %	56.46 %
b) (10-20]	24.92 %	21.06 %	12.73 %
c) (20-50]	10.67 %	8.84 %	2.79 %
d) (50-100]	1.86 %	1.24 %	0.20 %
e) (100-150]	0.75 %	0.60 %	0.04 %
f) (150-200]	0.50 %	0.31 %	0.02 %
g) (200-300]	0.87 %	0.45 %	0.02 %
h) (300-500]	0.76 %	0.69 %	0.02 %
i) (500-1000]	1.25 %	1.06 %	0.01 %
j) (1000-20000]	25.80 %	11.30 %	0.03 %
k) S/D	0.00 %	18.44 %	27.68 %
TOTAL	100.00 %	100.00 %	100.00 %

Fonte: (Autor, 2023).

As ligações que se enquadram dentro da faixa de consumo de 0 a 10 m³/mês correspondem a 56,46% das ligações ativas. Além disso, 25,80% do VC total está concentrado na maior faixa de consumo, ocasionado, provavelmente pelas ligações industriais. Ainda assim, 27,68% das ligações não possuem informação de volume consumido na base de dados, número muito elevado de dados ausentes.

Figura 60 – Percentual acumulado da estratificação por faixa de consumo.

FAIXA_CONSUMO	%VC	%VF	%Ligações
a) [0-10]	32.62 %	36.01 %	56.46 %
b) (10-20]	57.54 %	57.06 %	69.19 %
c) (20-50]	68.21 %	65.91 %	71.98 %
d) (50-100]	70.07 %	67.14 %	72.18 %
e) (100-150]	70.82 %	67.74 %	72.22 %
f) (150-200]	71.32 %	68.05 %	72.24 %
g) (200-300]	72.19 %	68.51 %	72.27 %
h) (300-500]	72.95 %	69.19 %	72.28 %
i) (500-1000]	74.20 %	70.25 %	72.29 %
j) (1000-20000]	100.00 %	81.56 %	72.32 %
k) S/D	100.00 %	100.00 %	100.00 %

Fonte: (Autor, 2023).

Percebe pela Figura 60 que as duas primeiras faixas de consumo (0 a 10 m³/mês e 10 a 20 m³/mês) apresentam cerca de 70% de VC e ligações ativas analisadas.

Ademais, na Figura 61 torna-se possível observar o volume submedido total para as ligações ativas, representando um total de 49.183 m³ de volume submedido em função da idade do parque de hidrômetros estudado. Dessa forma, o parque de hidrômetros apresentou uma submedição de 28% em relação ao volume consumido para as ligações estudadas.

Figura 61 – Estratificação dos volumes consumidos e submedidos por faixa de consumo.

FAIXA_CONSUMO	Média da Idade	Volume Consumido (VC)	Volume Corrigido	Volume Submedido	Quantidade de Ligações
a) [0-10]	11.74	57 265 m³	75 587 m³	18 322 m³	13 861
b) (10-20]	10.98	43 751 m³	57 420 m³	13 669 m³	3 125
c) (20-50]	10.44	18 730 m³	24 323 m³	5 593 m³	685
d) (50-100]	8.80	3 259 m³	4 107 m³	848 m³	49
e) (100-150]	9.45	1 325 m³	1 701 m³	376 m³	11
f) (150-200]	8.00	875 m³	1 097 m³	222 m³	5
g) (200-300]	8.83	1 520 m³	1 949 m³	429 m³	6
h) (300-500]	9.25	1 336 m³	1 731 m³	395 m³	4
i) (500-1000]	3.67	2 203 m³	2 468 m³	265 m³	3
j) (1000-20000]	7.14	45 285 m³	54 350 m³	9 065 m³	7
k) S/D	15.36	0 m³	0 m³	0 m³	6 795
TOTAL	12.04	175 549 m³	224 732 m³	49 183 m³	24 551

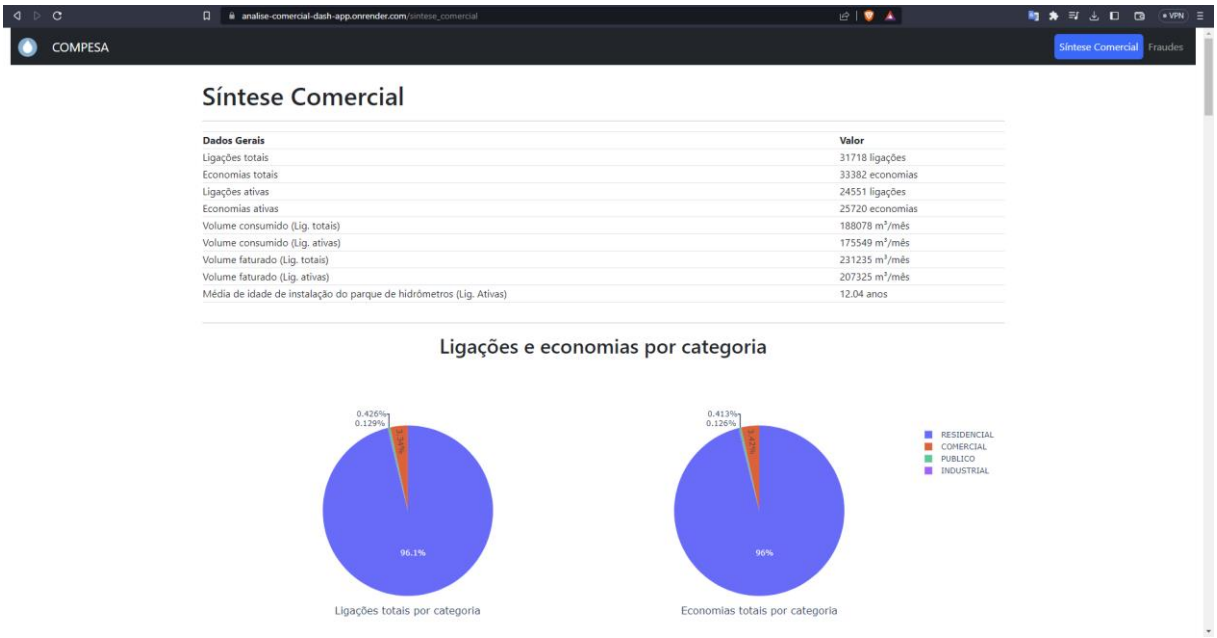
Fonte: (Autor, 2023).

De maneira resumida, pode-se realizar o cálculo rápido da submedição estimada através da idade média de todo o parque de hidrômetros total e do volume consumido total. Uma vez que a média de idade encontrada foi de 12 anos e observando o gráfico da Figura 17, percebe-se que para HDs com mais de 10 anos, convém adotar-se pelo menos 26% de submedição geral para o volume consumido total.

5.7 Dashboard

Por fim, dashboard criado sintetizando os resultados obtidos pode ser observado na Figura 62 e acessado através do link Dashboard: Relatório Análise Comercial.

Figura 62 – Dashboard com síntese dos resultados.



Fonte: (Autor, 2023).

6 CONCLUSÕES

Primeiramente, vale destacar que a contribuição da comunidade de código aberto foi essencial para o andamento deste trabalho, pois incluiu várias abordagens adotando bibliotecas previamente desenvolvidas por terceiros e amplamente utilizadas por uma grande variedade de indivíduos.

Ao longo deste estudo, foi evidenciado que cada etapa do processo de análise de dados foi essencial para garantir resultados confiáveis e significativos, os resultados obtidos neste trabalho demonstram que a preparação dos dados foi um passo fundamental para viabilizar a análise exploratória e a criação de recursos adicionais para a análise dos dados de consumo de água.

A preparação dos dados permitiu a criação de um *dataframe* final com ligações organizadas por matrícula, possibilitando a análise de dados de uma grande quantidade de linhas que seria inviável por métodos convencionais.

A fase de análise exploratória dos dados revelou importantes esclarecimentos que permitem identificar tendências, padrões e relações entre as variáveis em estudo. Esta análise aprofundada permitiu uma melhor compreensão do cenário sob investigação e apoiou as conclusões e recomendações propostas, gerando insights valiosos para a organização. Além disso, foram identificadas 5.344 ligações com três meses de anormalidade de consumo ou leitura, utilizadas para o rótulo de provável fraude.

O pré-processamento dos dados foi crucial para tratar os valores ausentes e transformar as variáveis categóricas em variáveis *dummy*, possibilitando o uso adequado dos dados no modelo. A análise de correlação, realizada por meio do coeficiente de Pearson e da matriz de correlação, permitiu identificar as variáveis mais relevantes para a condição de anormalidade que representa o rótulo de provável fraude em estudo. Sendo a conjuntura de três meses sem consumo o indicador mais relevante.

Em seguida, o processamento e os resultados dos modelos mostraram que a acurácia obtida foi relativamente alta para os três modelos estudados (K Vizinhos Mais Próximos, Árvore de Decisão e Floresta Aleatória), após o uso de dados bem tratados e limpos. Destaca-se que o modelo de Floresta Aleatória obteve a maior acurácia, se tratando do mais preciso dentre os três modelos analisados.

Entretanto, o reprocessamento dos dados, com a remoção dos outliers identificados nas colunas numéricas e a normalização das variáveis relevantes, resultou em uma melhoria quase que irrelevante dos resultados, com um pequeno aumento na acurácia dos modelos. Dessa forma, faz-se importante destacar que o reprocessamento dos dados não foi satisfatório ou não foi necessário. Visto que não é vantajosa a retirada de dados, ou melhor, ligações com informações reais, apesar de considerados *outliers*, para um aumento desprezível na acurácia.

Desse modo, evidencia-se a necessidade de realizar utilizar diferentes métricas que avaliam a precisão e outras abordagens para melhoria do desempenho nos modelos supervisionados, como, por exemplo, o ajuste dos hiperparâmetros que controlam o comportamento de cada modelo.

Os resultados obtidos destacam a importância da preparação adequada dos dados e da análise exploratória para identificar padrões e características relevantes nos dados de consumo de água, possibilitando insights importantes para a gestão e planejamento do cadastro comercial dos sistemas de abastecimento de água das cidades.

Além disso, realizou-se uma síntese comercial dos dados gerais do cadastro comercial do município de Vitória de Santo Antão, referentes ao último mês dos dados disponibilizados (fevereiro de 2021). Por meio de filtragens do *dataframe* e utilização de funções da biblioteca pandas, foram obtidos dados relevantes sobre as ligações e economias totais, ligações e economias ativas, volume consumido e volume faturado, entre outros.

A visualização dos dados foi facilitada pela criação de gráficos utilizando a biblioteca plotly, o que permitiu uma análise mais aprofundada de cada uma das variáveis estudadas. Os gráficos apresentaram informações sobre a distribuição das ligações e economias nas diferentes categorias de uso da economia principal, assim como a situação das ligações de água (ativa ou inativa) e a presença ou ausência de hidrômetro (HD) nas diferentes categorias.

Quanto ao cálculo da submedição, foram alcançados resultados satisfatórios uma vez que a partir da estimativa de volume submedido individualmente para cada ligação, obteve-se um percentual coerente de 28% de submedição, ratificado pela submedição geral para a idade média do parque de hidrômetros. Condicionado, desse modo, pela grande quantidade de hidrômetros com idade média elevada do parque.

Fato esse que também remete a alta correlação identificada da idade dos HDs com a condição de fraude.

Por fim, é importante explicitar a limitação dos dados trabalhados visto que a série histórica disponibilizada no banco de dados foi de apenas um ano. Quanto maior a quantidade de dados disponibilizados, melhor condicionada será a análise. Concomitantemente, destaca-se o impacto da pandemia e da quarentena em relação aos dados apresentados, visto que durante esse período, na maioria das capitais, os leituristas não visitaram as casas e medição foi realizada a partir de valores médios.

Em síntese, os resultados obtidos neste estudo contribuem para uma melhor compreensão e visualização da situação do cadastro comercial de Vitória de Santo Antão, podendo ser utilizados como base para a tomada de decisões relacionadas à gestão do abastecimento de água no município. Recomenda-se que os gestores e responsáveis pela gestão do sistema de abastecimento de água utilizem esses resultados para aprimorar as estratégias de controle, monitoramento e planejamento do serviço, visando a melhoria da eficiência e qualidade do abastecimento de água para a população.

7 SUGESTÕES DE TRABALHOS FUTUROS

Estudar e utilizar o aprendizado de máquina não supervisionado de modo a avaliar sua eficiência em identificar as anormalidades do banco de dados comercial.

Utilizar outros parâmetros para avaliar o desempenho do modelo como acurácia, precisão, recall, F1-score, entre outras.

Elaborar metodologias para a melhora da precisão dos modelos de aprendizagem de máquina supervisionado estudados no presente trabalho. Concentrando-se no ajuste dos hiperparâmetros de cada modelo para se obter uma melhor performance.

Quanto ao aperfeiçoamento da análise comercial, sugere-se a continuidade da análise dos dados comerciais, buscando identificar possíveis padrões ou tendências ao longo do tempo, e a realização de análises mais detalhadas sobre as ligações inativas, como as causas para a inatividade e possíveis ações para reduzir esse problema.

Não obstante, uma outra abordagem interessante referente a análise comercial é estudar o volume perdido por essas ligações fraudulentas em relação ao comportamento do consumo das ligações regulares.

Além disso, pode-se realizar análises comparativas com outros municípios ou regiões, visando identificar boas práticas e estratégias que possam ser aplicadas em Vitória de Santo Antão.

REFERÊNCIAS

AESBE. Associação Brasileira das Empresas Estaduais de Saneamento. Série Balanço Hídrico. **Guia prático de procedimentos para estimativa de submedição no parque de hidrômetros.** 1ª edição. Brasília – DF. 2015.

BÁGGIO, M. A. Redução de Perdas em Sistemas de Abastecimento de Água. 2ª ed. Brasília: FUNASA, 2014. Disponível em: http://www.funasa.gov.br/site/wp-content/files_mf/reducao_de_perdas_em_saa74.pdf. Acesso em: 30 set. 2022.

BEZERRA, S. T. M.; CHEUNG, P. B. **Perdas de Água: Tecnologias de Controle**. João Pessoa: Editora da UFPB, 2013.

BRASIL. FUNASA. **Manual de Saneamento**. 4ª ed. rev. Brasília: FUNASA, 2006

BRK AMBIENTAL. Notícias: BRK Ambiental investe na modernização de hidrômetros. 2021. Disponível em: https://www.brkambiental.com.br/noticias/brk-ambiental-investe-na-modernizacao-de-hidrometros. Acesso em: 29 out. 2022.

BRUCE, P.; BRUCE A. **Estatística Prática para Cientistas de Dados: 50 conceitos essenciais**. 1. ed. Rio de Janeiro: Alta Books, 2019.

COELHO, A. C. **Micromedição em sistemas de abastecimento de água**: Livro editora universitária UFPB, 2009

COMPESA. Companhia Pernambucana de Saneamento. CADASTRO TÉCNICO LOTE 04. Dados Vetoriais. Caruaru – PE. 2020. COMPESA. Companhia Pernambucana de Saneamento. CADASTRO TÉCNICO LOTE 04. Dados Vetoriais. Vitória de Santo Antão – PE. 2020.

FOURNIOL, M. 2004. Avaliação do Parque de Hidrômetros. Maceió, AL (apud COELHO, 2009)

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

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep Learning**. MIT Press, 2016.

HARRISON, M. *Machine Learning* – Guia de Referência Rápida. [S.l.]: Novatec, 2020.

IBM. Random Forest. 2020. Disponível em: <https://www.ibm.com/cloud/learn/random-forest>. Acesso em: 29 out. 2022.

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. **An Introduction to Statistical Learning: with Applications in R**. Springer, 2017.
KOTU, V.; DESHPANDE, B. *Data Science*: *Concepts and Practice*. [S.l.]: Elsevier Science, 2019.

GREAT LEARNING TEAM**. What is Data Science? Beginner's Guide**. My Great Learning, 21 dez. 2022. Disponível em: https://www.mygreatlearning.com/blog/what-is-data-science/. Acesso em: 28 fev. 2023.

PROVOST, F.; FAWCETT, T. Data Science para Negócios. [S.l.]: Alta Books, 2016.

SNIS. Secretaria Nacional de Saneamento. **Diagnóstico dos Serviços de Água e Esgoto.** Brasília: Ministério do Desenvolvimento, 2020. Disponível em: http://www.snis.gov.br/downloads/diagnosticos/ae/2020/DIAGNOSTICO_TEMATICO_VISAO_GERAL_AE_SNIS_2021.pdf. Acesso em: 30 set. 2022.

SNSA. Sistema Nacional de Saneamento Ambiental. Perdas Aparentes. **AÇÕES DE ASSISTÊNCIA TÉCNICA EM REDUÇÃO E CONTROLE DE PERDAS DE ÁGUA E USO EFICIENTE DE ENERGIA ELÉTRICA.** 2018.

TRIOLA, M. F. **Introdução à Estatística**. 13. ed. Rio de Janeiro: LTC, 2018.

TSUTIYA, M. T. **Abastecimento de Água.** São Paulo: Departamento de Engenharia Hidráulica e Sanitária da Escola Politécnica da USP, 2006.

ANEXO 1 – AUTORIZAÇÃO PARA USO DOS DADOS

COMPESA -

Processo nº 0060500614.000117/2022-80

Despacho: 242

Destinatário: GPD

Considerando SEI nº 30232603 e SEI nº 30254221, apresento concordância no que se refere ao fornecimento de dados para pesquisa de trabalho de conclusão de curso.

Atenciosamente,

FLÁVIO COUTINHO CAVALCANTE

Diretor de Negócios e Eficiência

Documento assinado eletronicamente por **Flavio Coutinho Cavalcante**, em 09/11/2022, às 10:24, conforme horário oficial de Recife, com fundamento no art. 10º, do Decreto nº 45.157, de 23 de outubro de 2017.

A autenticidade deste documento pode ser conferida no site http://sei.pe.gov.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **30326614** e o código CRC **C6BF7023**.

COMPANHIA PERNAMBUCANA DE SANEAMENTO

Av. Cruz Cabugá, 1387, - Bairro Santo Amaro, Recife/PE - CEP 50040-000, Telefone:

Criado por rodrigofalcao, versão 4 por rodrigofalcao em 09/11/2022 10:24:04.