Pontifícia Universidade Católica
do Rio de Janeiro

Guilherme Vinícius Lima dos Anjos

Sistema Inteligente para Identificação de Suspeitos de Fraude no Consumo de Água

DISSERTAÇÃO DE MESTRADO

DEPARTAMENTO DE ENGENHARIA ELÉTRICA

Programa de Pós-Graduação em Engenharia Elétrica

Rio de Janeiro

Abril de 2022

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Guilherme Vinícius Lima dos Anjos

Sistema Inteligente para Identificação de Suspeitos de Fraude no Consumo de Água

Dissertação de Mestrado

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre pelo Programa de Pós-Graduação em Engenharia Elétrica, do Departamento de Engenharia Elétrica da PUC-Rio.

Orientador: Marley Maria B. Rebuzzi Vellasco

Coorientador: Karla Tereza Figueiredo Leite

Rio de Janeiro
Abril de 2022

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Guilherme Vinícius Lima dos Anjos

Sistema Inteligente para Identificação do Suspeito de Fraude no Consumo de Água

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-graduação em Engenharia Elétrica da PUC-Rio. Aprovada pela Comissão Examinadora abaixo:

Marley Maria Bernardes Rebuzzi Vellasco
Orientador(a)
Departamento de Engenharia Elétrica – PUC-Rio

Karla Tereza Figueiredo Leite
Co-Orientador(a)
UERJ

Paulo Ivson Netto Santos
Tecgraf – PUC-Rio

José Franco Machado do Amaral
UERJ

Harold Dias de Mello Junior
UERJ

Rio de Janeiro, 27 de abril de 2022

Guilherme Vinícius Lima dos Anjos

Graduou-se em Engenharia Mecatrônica pela Universidade Católica de Petrópolis em 2019. Atuou em conjunto com o Laboratório de Inteligência e Robótica Aplicada (LIRA) da PUC-Rio no desenvolvimento de ferramentas e métodos de Inteligência Artificial. É profissional na área de Ciência de Dados e trabalha pesquisando e aplicando novas metodologias da área de aprendizado de máquina.

Ficha Catalográfica

Anjos, Guilherme Vinícius Lima dos

Sistema inteligente para identificação de suspeitos de fraude no consumo de água / Guilherme Vinícius Lima dos Anjos; orientador: Marley Maria B. Rebuzzi Vellasco; coorientador: Karla Tereza Figueiredo Leite. – 2022.

131 f.: il. color.; 30 cm

Dissertação (mestrado)–Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, 2022.

Inclui bibliografia

1. Engenharia Elétrica – Teses. 2. Algoritmos evolucionários com inspiração quântica. 3. Comitê de classificadores. 4. Perdas aparentes. 5. Fraude. I. Vellasco, Marley M. B. R. (Marley Maria Bernardes Rebuzzi). II. Leite, Karla Tereza Figueiredo. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Elétrica. IV. Título.

CDD: 621.3

Para meus pais Elisabete e Marcos, minha irmã Carol e minha namorada Marianna, por todo apoio, confiança e motivação depositada em mim ao longo de todo o curso, e além dele.

Agradecimentos

A jornada acadêmica é um processo que é suavizado quando estamos rodeados de pessoas que dão o devido suporte e entidades que providenciam a estrutura necessária para tal.

Neste sentido, agradeço primeiramente, à toda minha família. Meus pais Marcos e Elisabete, e minha irmã Carol, por proporcionarem suporte e facilitar meu caminho até aqui.

A Marianna, companheira de longa data, agradeço a todo o amor, carinho, palavras e apoio desde o começo da graduação. Estou concluindo o mestrado e sem você essa equação não seria resolvida.

A minha professora orientadora, profª Marley Vellasco, e coorientadora profª Karla Figueiredo, agradeço pela amizade, pelas aulas, por todo apoio e suporte durante a confecção deste trabalho e toda jornada do meu mestrado. Tenham a certeza que foram responsáveis por fazerem meu trabalho e minha conclusão de curso mais gratificante.

Aos professores de graduação Giovane Quadrelli e Paulo Leite, por incentivarem a sequência de aprendizado acadêmico e por serem exemplares na minha graduação. Aproveito para agradecer a Universidade Católica de Petrópolis por fornecer toda estrutura necessária para a formação de jovens na cidade.

Aos colegas do Laboratório de Inteligência e Robótica Aplicada (LIRA) e do TecGraf que, de alguma forma, contribuíram na confecção deste trabalho.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Por isso agradeço a CAPES por todo apoio financeiro que pôde me dar estrutura para que eu pudesse me dedicar integralmente ao curso de mestrado.

Aos professores doutores que participaram da Comissão Examinadora.

Resumo

Dos Anjos, Guilherme Vinícius Lima; Vellasco, Marley M. B. Rebuzzi. **Sistema Inteligente para Identificação de Suspeitos de Fraude no Consumo de Água**. Rio de Janeiro, 2022. 131p. Dissertação de Mestrado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

Um dos maiores problemas de todas as empresas prestadoras de serviço de saneamento e distribuição de água é o de perdas oriundas de irregularidades (comerciais). Dentre os países com mais de 20 milhões de habitantes que mais sofrem desse tipo de perdas, o Brasil ocupa a 14º posição com 40% de perdas na distribuição. A Empresa A, estudo de caso deste trabalho, é uma companhia brasileira que atua no setor de saneamento e distribuição de água e, atua, principalmente, em 3 regiões, com valores de médias percentuais de perdas, em 2021, de 19%, 30% e 43%, respectivamente. Essas perdas são derivadas de muitos problemas, mas as principais são oriundas das fraudes nas ligações dos medidores de água, por exemplo: ligações clandestinas, *by-pass* e derivação de ramal. A principal forma de combater esse tipo de fraude é através de inspeções nos clientes. Geralmente utiliza-se um conjunto de heurísticas para identificar o suspeito de tal fraude ou irregularidade, porém esses métodos não retornam boas precisões. Na Empresa A, a precisão alcançada através das inspeções varia de 3% a 17% de região para região. Com isso, conclui-se que o procedimento não é eficaz. Sendo assim, o objetivo deste trabalho é desenvolver um sistema inteligente que possa identificar, com maior exatidão, o perfil de consumo do cliente que possui a fraude. O sistema desenvolvido é composto por duas metodologias baseadas em diversos algoritmos supervisionados de aprendizado de máquina. A primeira utiliza um filtro com intuito de agrupar os clientes com perfis similares. A segunda faz uso de um algoritmo evolutivo inspirado em computação quântica para a busca de hiperparâmetros e atributos. Além disso, ambas consideram comitês e exploram a utilização de variáveis históricas e exógenas pertinentes ao contexto. Os resultados obtidos mostraram-se superiores nas avaliações, quando comparadas aos verificados na Empresa A, alcançando até 44% de taxa de acerto.

Palavras-chave

Algoritmos Evolucionários com Inspiração Quântica; Comitê de Classificadores; Perdas Aparentes; Fraude.

Abstract

Dos Anjos, Guilherme Vinícius Lima; Vellasco, Marley M. B. Rebuzzi. **Intelligent System for the Identification of Fraud Suspects in Water Consumption**. Rio de Janeiro, 2022. 131p. Dissertação de Mestrado - Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

One of the biggest problems faced by all companies that provide sanitation and water distribution services is that of losses arising from (commercial) irregularities. Among the countries with more than 20 million inhabitants that suffer the most from this type of loss, Brazil occupies the 14th position with 40% of losses in distribution. Company A, the case study of this work, is a Brazilian company that operates in the sanitation and water distribution sector and operates mainly in 3 regions, with average percentage values of losses, in 2021, of 19%, 30 % and 43%, respectively. These losses derive from many problems, but the main ones arise from fraud in the connections of water meters, for example: clandestine connections, by-pass and branch derivation. The main way to combat this type of fraud is through customer inspections. Generally, a set of heuristics is used to identify the suspect of such fraud or irregularity, but these methods do not return good accuracy. At Company A, the accuracy achieved through inspections varies from 3% to 17% from region to region. Thus, it is concluded that the procedure is not effective. Therefore, the objective of this work is to develop an intelligent system that can identify, with greater accuracy, the consumption profile of the customer who has the fraud. The developed system is composed of two methodologies based on several supervised machine learning algorithms. The first uses a filter in order to group customers with similar profiles. The second makes use of an evolutionary algorithm inspired by quantum computing to search for hyperparameters and attributes. In addition, both consider committees and explore the use of historical and exogenous variables relevant to the context. The results obtained were superior in the evaluations, when compared to those verified in Company A, reaching up to 44% of success rate.

Keywords

Quantum-Inspired Evolutionary Algorithms; Classifiers Committee; Apperant Losses; Fraud.

Sumário

Lista de Figuras

Lista de Tabelas

1. Introdução

A água é um recurso natural fundamental para a sobrevivência. Aproximadamente 1,4 bilhões km cúbicos é o valor total estimado de água na Terra. Tendo 97,5% de sua composição sendo água salgada e o restante de água doce.

Apesar de ter sido declarado em 2010 pela Organização das Nações Unidas (ONU) que o acesso à água limpa e segura, e ao saneamento básico são direitos humanos fundamentais, ainda há muitos desafios acerca da questão do serviço de distribuição de água, no Brasil e no mundo.

No Brasil, o acesso à água é possível, em geral, por meio de distribuidoras que detém a concessão do serviço. A realização deste serviço envolve custos que vão desde a captação da água em mananciais, tratamento até a distribuição propriamente dita.

1.1. Motivação

Entre os muitos problemas enfrentados pelas empresas, públicas ou privadas, de distribuição de água, as perdas são as mais sensíveis.

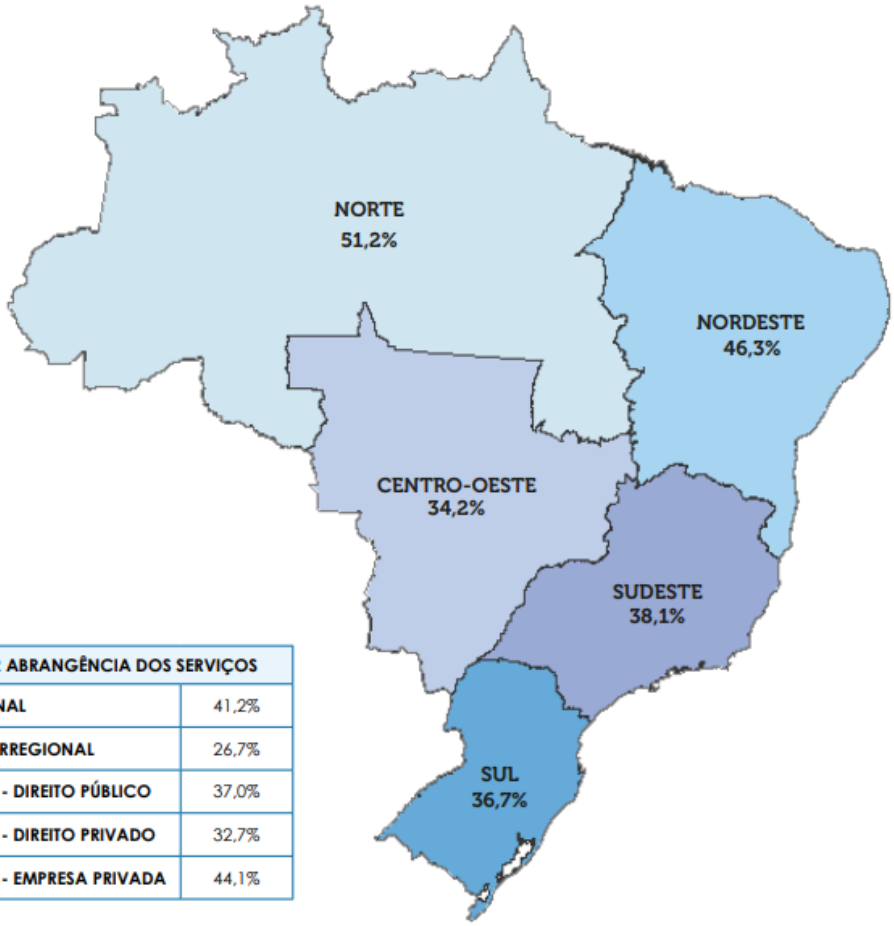
Dados do Sistema Nacional de Informação de Saneamento (SNIS) de 2020 apontam um índice de perdas na distribuição de água de 40,1% no Brasil. Ele apresenta crescimento contínuo após um período de estabilidade, entre 2012 e 2015, quando chegou a ficar abaixo de 37,0%. Em termos quantitativos, o índice significa que, a cada 100 litros disponibilizados pelos prestadores de serviços, apenas 59,9 são contabilizados como utilizados pelos consumidores (SNIS, 2021).

Nas Macrorregiões, os índices de perdas variam de 34,2%, na região Centro-Oeste a 51,2%, na Norte. Na abrangência do serviço, a amostra identifica perdas entre 26,7%, na prestação Microrregional, e 44,1%, na “Local - Empresa Privada” (SNIS, 2021), como pode ser observado na tabela da Figura 1. As macromedições são feitas em pontos da rede de distribuição, enquanto as micromedições ocorrem no ponto de atendimento ao usuário com hidrômetros.

Existem dois tipos de perdas: (1) perda aparente, quando a água consumida não é contabilizada (cobrada), devido a situações como ligações clandestinas (gatos), adulteração nas ligações e submedições (falta de calibragem nos hidrômetros);

e (2) perda real, quando há vazamentos em pontos das infraestruturas de distribuição.

As perdas reais recaem sobre os custos de produção e distribuição da água, enquanto as perdas aparentes atingem os custos de venda da água acrescidos dos custos da coleta de esgotos. Assim, as abordagens econômicas são diferentes entre esses dois tipos. Em outras palavras, o impacto econômico de uma pode ser mais significante que a outra, mesmo que o volume das perdas seja semelhante. Dessa forma, as perdas trazem impactos negativos para a sociedade, meio ambiente, receita das empresas e até mesmo aos investimentos necessários aos avanços do saneamento básico.



POR ABRANGÊNCIA DOS SERVIÇOS	
REGIONAL	41,2%
MICRORREGIONAL	26,7%
LOCAL - DIREITO PÚBLICO	37,0%
LOCAL - DIREITO PRIVADO	32,7%
LOCAL - EMPRESA PRIVADA	44,1%

Figura 1 - Índices de Perdas na Distribuição de Água.

Fonte: SNIS, 2021.

A partir da Figura 2, pode-se observar que há uma tendência de aumento no percentual de perdas ao longo dos últimos anos no Brasil. Sendo assim, é necessário utilizar a combinação de múltiplos métodos e ações que visem a melhoria da gestão, e aplicação de técnicas que possam começar a reduzir esses índices.

De forma geral, o método mais eficaz para uma empresa confirmar se há algum tipo de perda ou irregularidade – em relação às ligações clandestinas ou adulteradas – é utilizando heurísticas que, primeiro, indiquem a suspeita sobre esses clientes para que, posteriormente, a prestadora realize visitas técnicas ou inspeções. Quando as inspeções detectam fraudes ou irregularidades, o custo envolvido na inspeção pode ser coberto pela receita recuperada com a eliminação da fraude; entretanto, quando se realizam muitas inspeções que não detectam fraudes, essas podem agravar os prejuízos para as empresas distribuidoras e, consequentemente, para os consumidores.

Assim, a maioria das distribuidoras emprega métodos simples para direcionar suas inspeções, geralmente baseados no conhecimento de especialistas no assunto e no uso das informações contidas nas bases de dados das empresas.

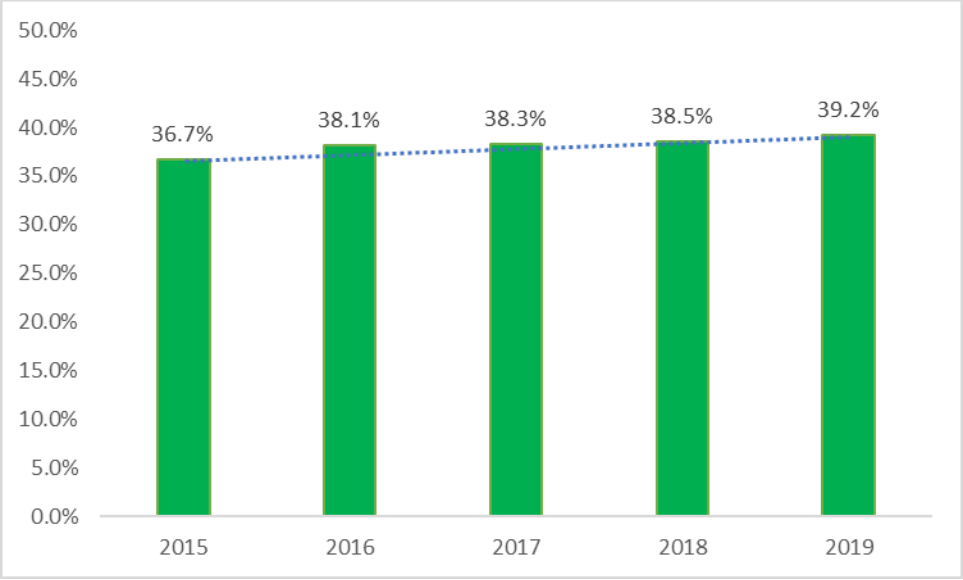


Figura 2 - Evolução do Índice de Perdas na Distribuição de Água.

Fonte: SNIS, 2021.

A Empresa A – estudo de caso deste trabalho – possui um conjunto de heurísticas que pode indicar se um dado cliente está cometendo alguma fraude ou irre-

gularidade. Esses clientes, nesse primeiro momento, são declarados, por meio dessas heurísticas, suspeitos, fazendo com que especialistas da empresa os visitem para confirmar, com base em evidências e sua expertise, se são, de fato, clientes fraudadores. A empresa atua em diversas regiões, porém, os dados disponibilizados envolvem apenas 3: Região G, Região P e Região T.

Através das heurísticas mencionadas acima, a empresa alcançou 9,2%, 16,8% e 3%, respectivamente em cada região, de percentual de acerto de clientes fraudadores ou irregulares entre setembro/2019 e fevereiro/2020. Pode-se concluir que o método não é eficaz.

Estudos de Al-Radaideh e Al-Zoubi (2018), Gopal e Balaji (2020) abordaram o problema utilizando duas técnicas: *Support Vector Machine* (SVM) e *K-Nearest Neighbor* (KNN). O primeiro utilizou dados de cadastro e consumo anual para detectar comportamentos anômalos dos clientes, enquanto o segundo, abordou o problema com o consumo histórico dos clientes e o que de fato foi faturado pelas empresas. Ambos alcançaram 70% de acurácia na detecção dos clientes fraudadores.

Sendo assim, propondo uma nova abordagem e, para melhorar a taxa de acerto da empresa, foi desenvolvido um sistema que reúne diversas técnicas de aprendizado de máqui*na* – e a combinação destas através dos comitês – que pode aprender os padrões de consumo dos clientes fraudadores e não-fraudadores, utilizando a grande base de dados de consumo mensal e também variáveis exógenas tais como: informações de precipitação, umidade, temperatura, etc.

1.2. Objetivos

O objetivo principal deste trabalho é investigar metodologias para o desenvolvimento de um sistema inteligente de apoio à decisão, que avalie uma gama maior de algoritmos e de aprendizado de máquina e técnicas de mineração de dados, visando melhorar a qualidade da detecção do cliente suspeito de fraude no consumo de água.

Outros objetivos secundários podem ser destacados:

- Avaliar algoritmos de aprendizado de máquina individualmente;

- Buscar a composição dos resultados destes em comitês de classificadores;

- Avaliar composição de variáveis construídas a partir do consumo dos clientes, além de avaliar a utilização de variáveis exógenas que pudessem agregar informações aos modelos;

- Explorar técnicas evolucionárias que busquem o melhor conjunto de variáveis e hiperparâmetros com objetivo de melhorar o desempenho do sistema, dada a grande quantidade de atributos e parâmetros relacionados que se deseja considerar;

- Investigar técnicas relacionadas à filtragem de registros para aumentar a confiabilidade da informação aprendida pelos modelos, pois, de forma geral, sabe-se que os serviços de distribuição energia, água, etc. possuem incerteza associada à rotulagem dos clientes (Ortega, 2008), o que ajuda a confundir os algoritmos durante o aprendizado;

- Comparar desempenho das metodologias aplicadas em um problema real e confrontá-los com os resultados da empresa;

- Combater de forma incisiva as irregularidades no consumo de água que, em geral, são responsáveis diretamente pelas perdas comerciais de distribuição de água das prestadoras de serviço, resultando em prejuízos exorbitantes de orçamento e desperdício de água.

1.3. Estrutura da Dissertação

O restante desse trabalho está dividido em mais seis capítulos.

O segundo capítulo descreve os conceitos básicos das perdas aparentes e perdas reais, assim como uma análise de prejuízo financeiro e hídrico, a partir destas perdas no Brasil e também nas regiões cobertas pela Empresa A.

O Capítulo 3 explica toda a teoria sobre Algoritmo Evolucionário com Inspiração Quântica (AEIQ), que é a base da busca por melhores parâmetros e atributos utilizado em uma das metodologias do sistema inteligente.

O Capítulo 4, por sua vez, explica detalhadamente como o sistema é construído e todo o processo iterativo da Metodologia que tem como base a utilização

de uma filtragem K-Means, denominada aqui de Metodologia por Filtragem, e outra baseada em um processo evolutivo quântico que será chamada de Metodologia Evolutiva. Assim como, quais variáveis são criadas e o tipo de limpeza que é feita na base de dados.

As análises do desempenho e os resultados obtidos com a aplicação do sistema sobre a base de dados disponibilizada pela empresa, são exibidos no Capítulo 5.

Por fim, o Capítulo 6 apresenta as conclusões e sugestões de trabalhos futuros a fim de aprimorar o sistema utilizado.

2. Perdas de Distribuição de Água

Um fator primordial na avaliação da distribuição de operadores de saneamento é o cálculo do volume de perdas de abastecimento de água diagnosticados através de indicadores competentes. Esses índices, quando em níveis altos e com tendências de crescimento, significam que há a necessidade de maiores esforços para aumentar a eficiência de planejamento, manutenção e de atividades operacionais e comerciais.

Em 2019, o índice de perdas de faturamento total no Brasil foi de 37,06%, um pouco melhor do que os 39,21% mensurados em 2017. Em contrapartida, o índice de perdas na distribuição, foi de 38,45%, apresentando piora em relação aos 38,29% encontrados em 2017 (ITB, 2021). Esses índices ainda estão muito longe do ideal visto que, segundo o SNIS, consideram-se municípios com padrão de excelência em perdas, aqueles que possuem indicadores inferiores a 25%.

Este capítulo descreve o prejuízo nacional e regional que estão atrelados a esses índices, bem como a teoria das perdas de distribuição de água.

2.1. Análise de Indicadores

Para apresentar os déficits e prejuízos que as perdas podem trazer em âmbitos regionais e nacionais, optou-se por utilizar índices percentuais e unitários baseados em volume que englobam os dois tipos de perdas mencionadas nas seções anteriores. Nesta dissertação três indicadores foram selecionados: um representando a parte de faturamento, outro indicando a perda na distribuição de água e um terceiro que mostra a quantidade em volume perdida por ligação (cliente registrado).

A seguir estão descritos os indicadores que serão utilizados neste trabalho.

2.1.1. Índice de Perdas de Faturamento Total – IPFT

O Índice de Perdas no Faturamento Total é um indicador que foi desenvolvido pelo Instituto Trata Brasil, e tem como base o IN013 – Índice de Perdas no Faturamento, definido pelo SNIS, e é expresso pela Equação (1).

$$IN013 = \frac{AG006 + AG018 - AG011 - AG024}{AG006 + AG018 - AG024} \times 100 \qquad (1)$$

Segundo o SNIS:

- AG006 – Volume de Água Produzido → é igual ao volume anual de água disponível para consumo;
- AG011 – Volume de Água Faturado → é definido pelo volume anual de água debitado ao total de economias (medidas e não medidas), para fins de faturamento;
- AG018 – Volume de Água Tratado Importado → representa o volume anual de água potável, previamente tratada, recebido de outros agentes fornecedores;
- AG024 – Volume de Serviço → é a soma dos volumes utilizados para atividades operacionais e especiais, e do volume de água recuperado.

O índice tem como objetivo avaliar o nível de água não faturada do sistema de abastecimento em termos percentuais. Também, apresenta uma visão sobre o quanto a empresa produz e não fatura. O IN013 tem como ponto negativo o fato de que as empresas definem diferentes tipos de volumes de serviço, o que leva a diferentes percentuais em regiões distintas. Além disso, pode não refletir o nível de eficiência da empresa dependendo da metodologia utilizada.

Porém, como informado anteriormente, o índice adotado neste trabalho foi o IPFT, o qual foi desenvolvido pelo ITB para um estudo de perdas. Tal diferença se dá pelo fato que o índice total não subtrai o termo AG024, como descrito na Equação (2).

$$IPFT = \frac{AG006 + AG018 - AG011}{AG006 + AG018} \times 100 \qquad (2)$$

Tal sugestão do Instituto, se dá pela distinção entre as informações fornecidas pelas prestadoras, e ainda segundo o ITB, espera-se que seja um volume irrisório, esses que correspondem aos processos de abastecimentos, tratamento de esgoto e fornecidos por caminhão-pipa. Além disso, leva em consideração os volumes de serviços perdidos. Sendo assim, esse índice se adequa melhor à informação do percentual de perda de faturamento de forma geral (ITB, 2021).

2.1.2. Índice de Perdas na Distribuição – IN049

Já o Índice de Perdas na Distribuição – IN049 – tem como objetivo medir percentualmente o volume de água efetivamente consumida em um sistema de abastecimento de água potável. Descrito pela Equação (3), tem como sua principal vantagem a análise do impacto das perdas reais e aparentes na distribuição em relação ao volume produzido. O ponto negativo da utilização desse índice se dá pela diferença que as empresas definem o volume proveniente de serviços especiais (AG024), por isso a utilização do mesmo pode trazer distorções. Além disso, os níveis de macromedições e micromedições de cada empresa pode prejudicar as comparações.

$$IN049 = \frac{AG006 + AG018 - AG010 - AG024}{AG006 + AG018 - AG024} \times 100 \qquad (3)$$

Todos os termos da equação já foram definidos anteriormente, com exceção do AG010 que, segundo o SNIS:

- AG010 – Volume de Água Consumido → corresponde ao volume anual de água consumido por todos os clientes, englobando todo o volume estimado para ligações sem hidrômetro (ou hidrômetro parado), além do volume micromedido e o volume de água tratado exportado para outros prestadores.

2.1.3. Índice de Perdas por Ligação – IN051

Este índice avalia unitariamente e efetivamente o volume de água perdido (L/dia/ligação). Assim como os outros índices, tem como ponto negativo o fato de cada empresa caracterizar, de forma individual, o volume de serviços (AG024). Além disso, a comparação entre cidades é inadequada, visto que cidades maiores, verticalizadas e com maior consumo por habitante, fornecerão índices maiores. Apesar disso, esse índice reflete muito bem a variação do nível de perdas por ligação e está expresso na Equação (4).

$$IN051 = \frac{AG006 + AG018 - AG010 - AG024}{AG002^{*}} \times \frac{1.000.000}{365} \qquad (4)$$

Todas as variáveis foram definidas anteriormente, com exceção da AG002, que, segundo o SNIS:

- AG002 – Quantidade de Ligações Ativas de Água → representa, literalmente, a quantidade de ligações ativas de água, providas ou não de hidrômetro, que estavam conectadas à rede de abastecimento de água. O asterisco indica que o valor é uma média aritmética dos valores de ligação.

2.2. Perdas de Água na Atualidade

Nesta seção serão apresentados os índices de perdas de água em âmbito mundial, nacional e regional.

2.2.1. Mundial

O intuito desta seção é apresentar o cenário de perdas em nível internacional. Essa análise tem como objetivo evidenciar a tendência geral, sem comparar diretamente, pois os indicadores podem ter diferentes origens e levar a distorções na análise. Porém, segundo o SNIS, esses são os indicadores que os órgãos utilizam para averiguar se uma determinada região está de acordo com os padrões ótimos estipulados pela *International Water Association* (IWA).

Segundo o estudo de 2021, do Instituto Trata Brasil, “a principal fonte de informações sobre água não faturada a nível mundial é a *International Benchmarking Network for Water and Sanitation Utilities* (IBNET). Vale destacar que a periodicidade dos dados disponíveis varia bastante entre os países, de tal modo que algumas observações datam de anos recentes, enquanto noutras os valores disponíveis mais atuais são referentes ao início dos anos 2000.”.

A Figura 3 apresenta os índices de perdas de água não faturadas em âmbito mundial. Esta análise é feita sobre os países que possuem mais de 20 milhões de habitantes e, portanto, representam uma maior parcela da população mundial.

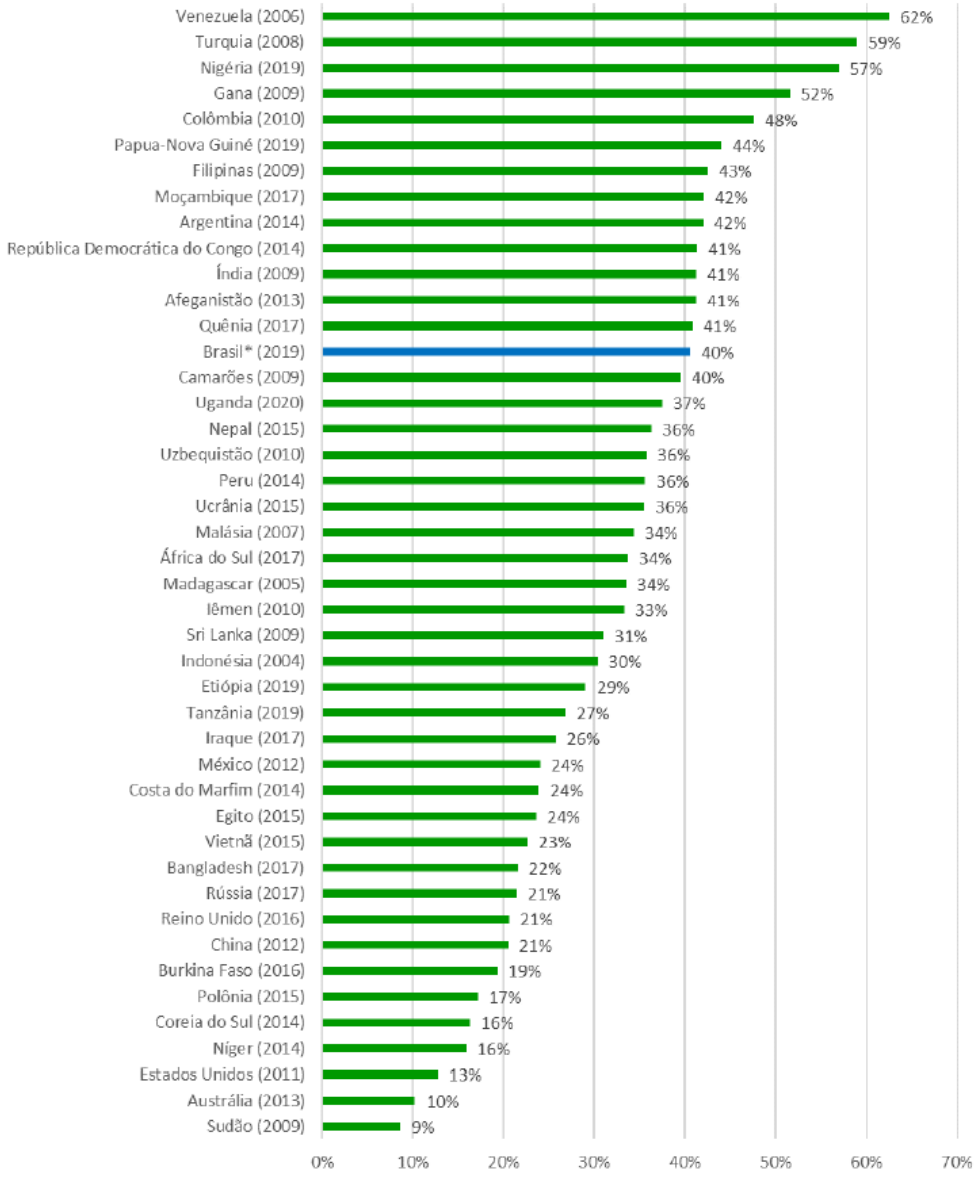


Figura 3 – Índices Internacionais de Perdas.

Fonte: ITB, 2021.

Conclui-se que o Brasil está bem atrás dos países desenvolvidos e, até mesmo dos países que tem similaridade na economia e desenvolvimento. Nos 44 países da amostra, o Brasil encontra-se na 31º colocação, o que o coloca em uma situação mais crítica.

2.2.2. Nacional

Como pôde ser visto anteriormente, o Brasil ainda precisa de muitos esforços e investimentos para contornar esse péssimo índice. Segundo o estudo de 2021, do Instituto Trata Brasil: “A média nacional das perdas de faturamento total em

2019 foi de 40,58%, 25 pontos percentuais acima da média dos países desenvolvidos, que é de 15%, e 5 pontos percentuais acima da média dos países em desenvolvimento, que é de 35%.”. Além disso, essas referências são bases de 2006, o que deixa a situação ainda mais crítica pois, com o decorrer do tempo, a tendência é que melhorias de tecnologia e investimento sejam feitas, ou seja, uma piora nos índices não é esperada.

A seguir estão apresentados os índices de perdas no faturamento (IN013 e IPFT) e perdas na distribuição (IN049) no Brasil de 2015 a 2019, respectivamente, na Figura 4 e Figura 5. A partir da análise destas, pode se concluir que, caso tenha ocorrido, não foi efetivo o esforço em reduzir a tendência de crescimento desses índices.

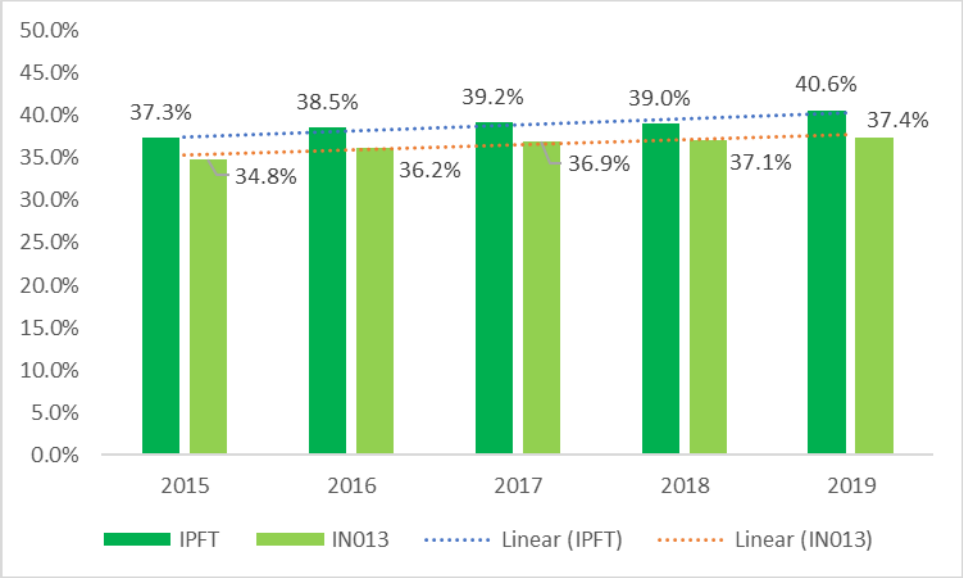


Figura 4 - Perdas no Faturamento (IPFT & IN013) - Brasil.

Fonte: ITB, 2021.

Isso fica ainda mais evidente no índice de perdas de distribuição, que no ano de 2020 chegou aos 40%. Ou seja, a tendência de crescimento está se mantendo, indicando a necessidade de implantação de, urgentemente, um método ou infraestrutura que possa erradicar essa “crise” hídrica.

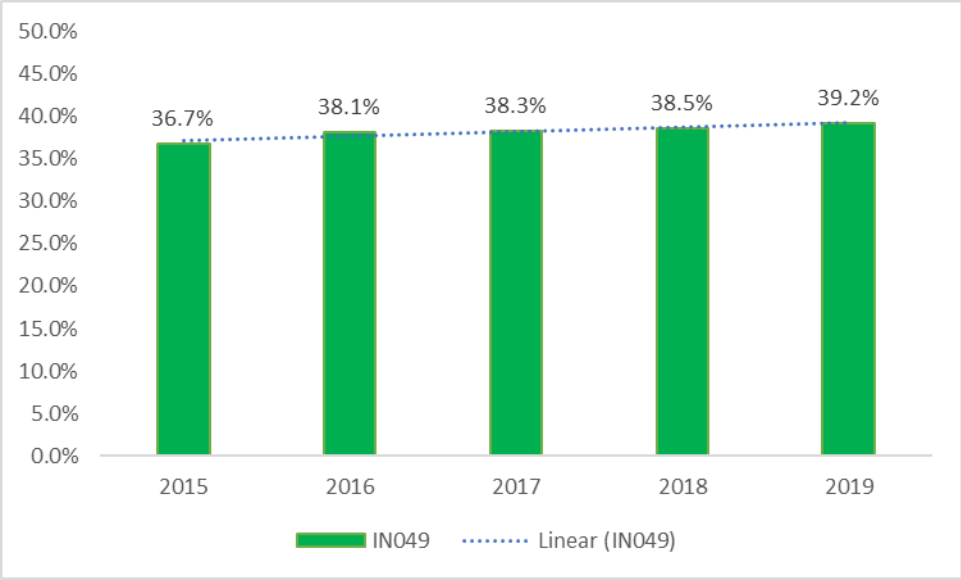


Figura 5 - Perdas na Distribuição (IN049) - Brasil.

Fonte: ITB, 2021.

2.2.3. Regional e Estadual

Nesta seção será apresentado de forma mais detalhada o problema de perdas no Brasil, observando em âmbito regional e estadual. A seguir, na Figura 6, Figura 7 e Figura 8, são apresentados os índices de perdas no faturamento, na distribuição e por ligação, respectivamente, dos últimos 5 anos disponíveis no SNIS à nível regional.

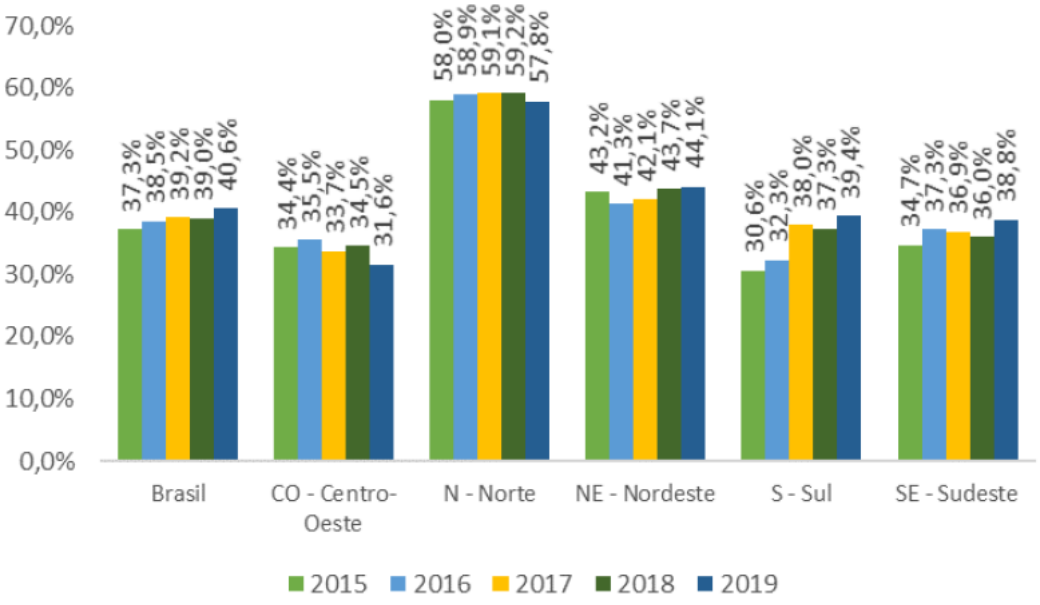


Figura 6 - Índice de Perdas no Faturamento por Região (IPFT de 2015-2019).

Fonte: ITB, 2021.

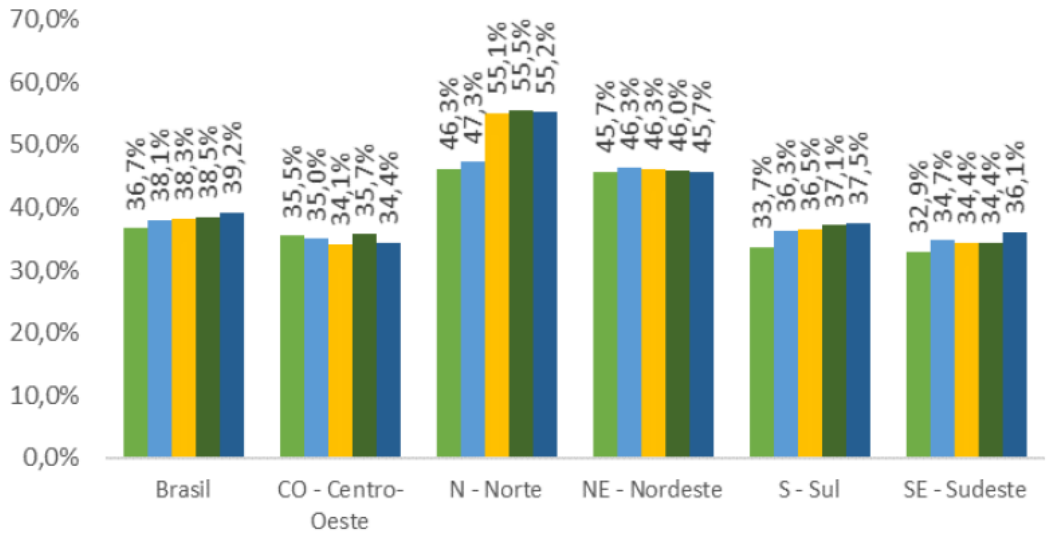


Figura 7 - Índice de Perdas na Distribuição por Região (IN049 de 2015 a 2019).

Fonte: ITB, 2021.

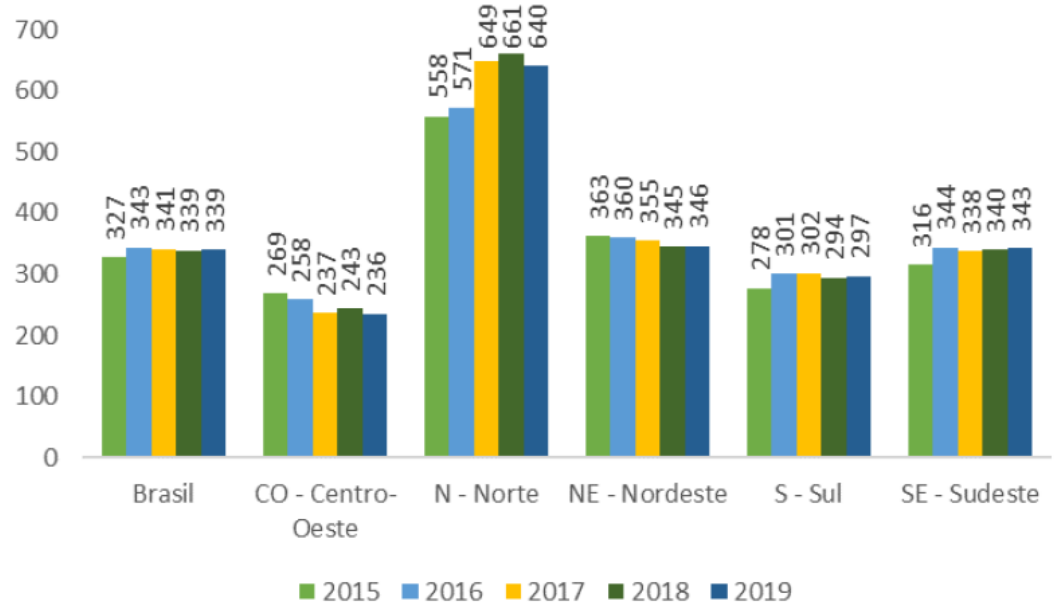


Figura 8 - Índice de Perdas por Ligação (L/ligação/dia)(IN051 de 2015 a 2019).

Fonte: ITB, 2021.

A partir das figuras pode ser observado que a região norte é que a mais sofre dessas perdas. Além disso, as regiões Norte e Nordeste juntas apresentam os piores índices de perdas no faturamento. Essas duas regiões provavelmente enfrentarão maiores desafios para reduzir estes índices de perda.

Além disso, é notável a estagnação ou acréscimo nos índices de todas as regiões, com exceção da região Centro-Oeste que conseguiu diminuir todos estes índices, mas ainda assim longe do ideal.

A partir da Figura 6, a região Sul ganhou destaque negativo por aumentar o índice de perdas no faturamento em pelo menos 9%, enquanto que o destaque positivo foi a região Centro-Oeste que conseguiu diminuir este mesmo índice em quase 3%. Na Figura 9, Figura 10 e Figura 11, são apresentados, respectivamente, os índices de perdas no faturamento, na distribuição e por ligação de cada Estado.

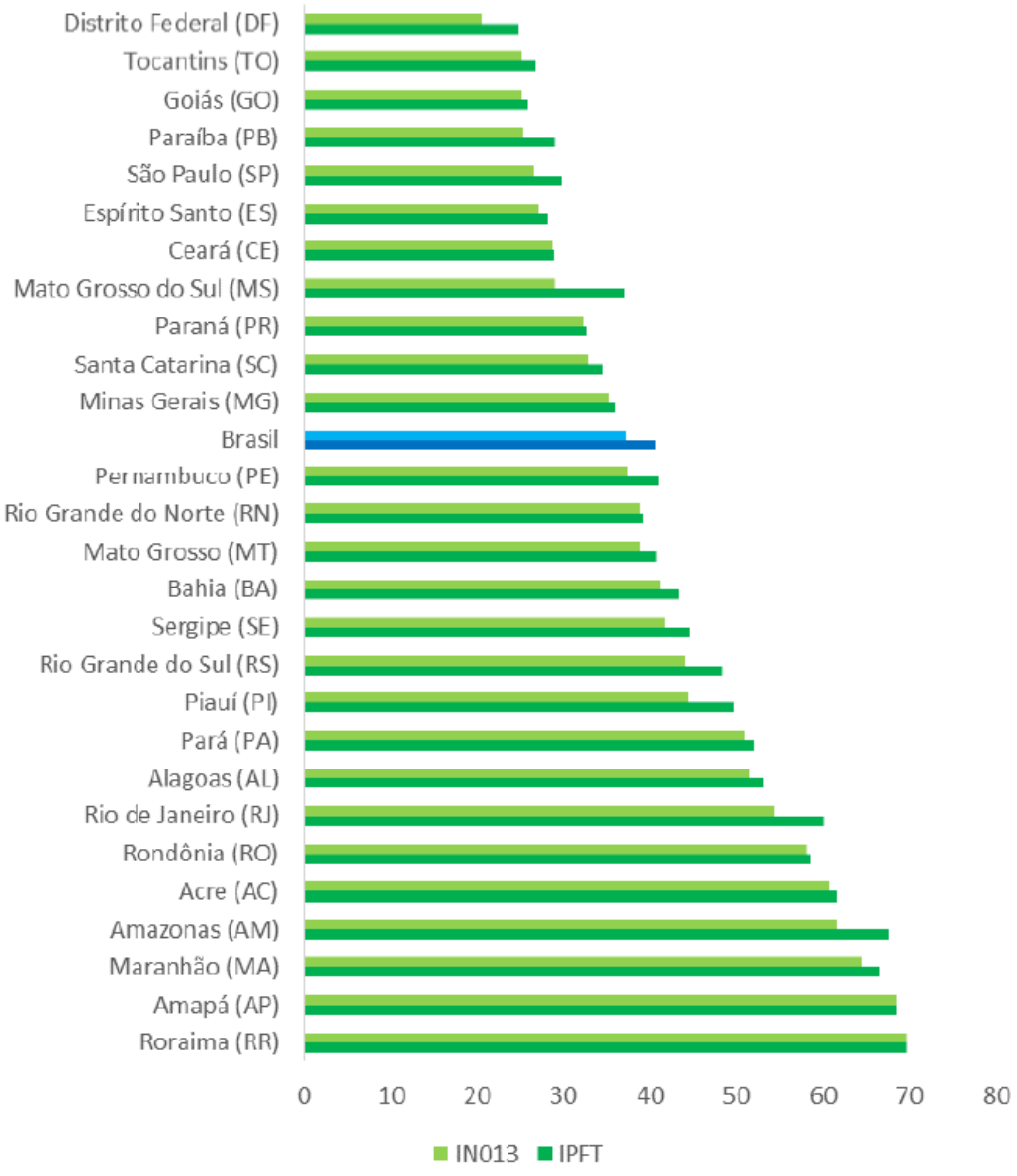


Figura 9 - Perdas no Faturamento por Estados (%) (IPFT de 2019).

Fonte: ITB, 2021.

Através da Figura 7 e Figura 8, conclui-se que a região Norte foi a que mais apresentou piora nos dois índices, com aumento de 9% nas perdas de distribuição e um aumento de quase 82 L/ligação/dia nas perdas por ligação. Já a região Centro-Oeste foi a que continuou apresentando maior melhora, com redução de 1% nas perdas de distribuição e redução de quase 33 L/ligação/dia.

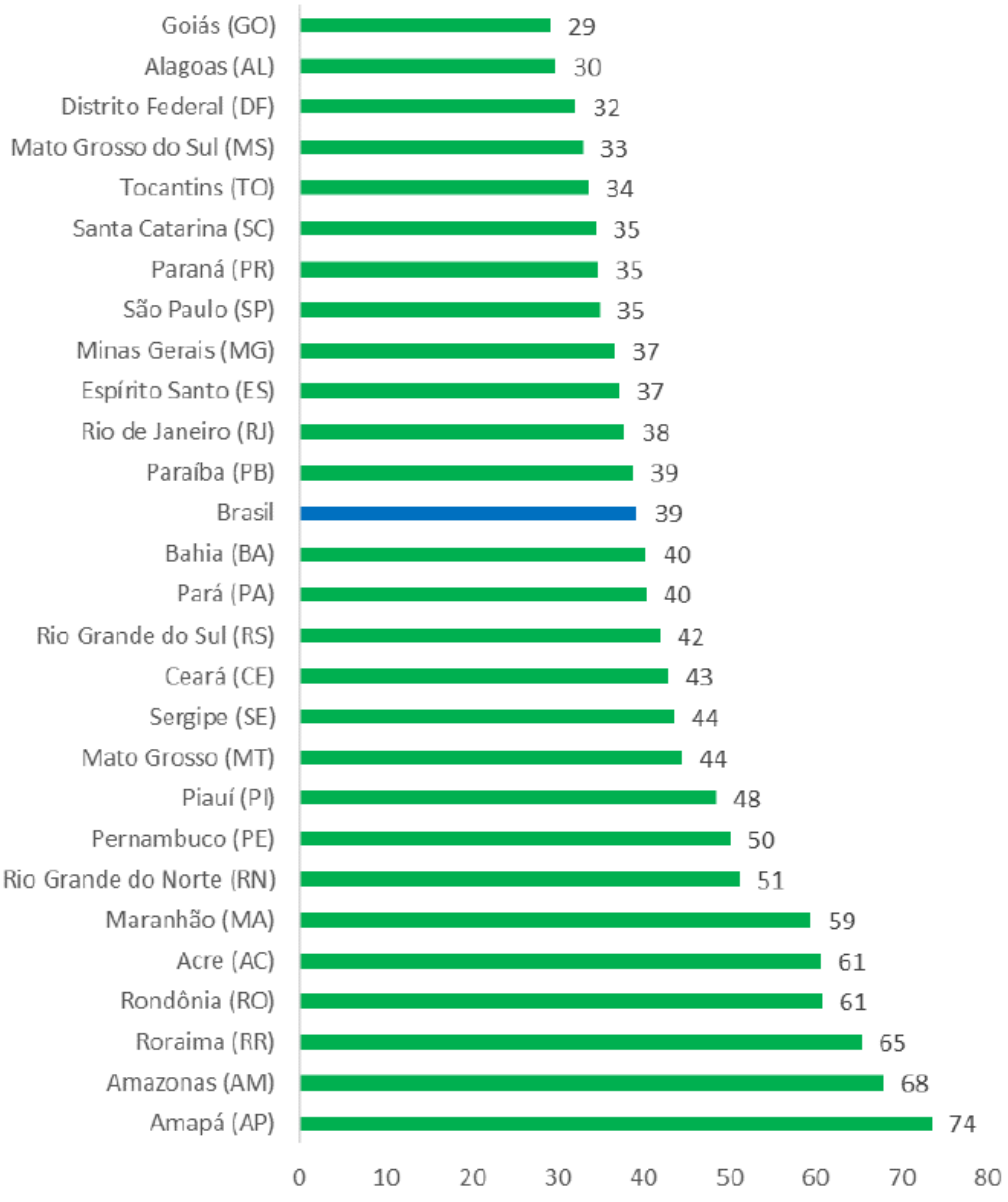


Figura 10 - Perdas na Distribuição por Estados (%) (IN049 de 2019).

Fonte: ITB, 2021.

Quando é feito um desmembramento das regiões e analisa-se os índices em âmbito estadual, pode-se observar que os índices por Estado acompanham o ordenamento das regiões, porém, há sempre algumas exceções.

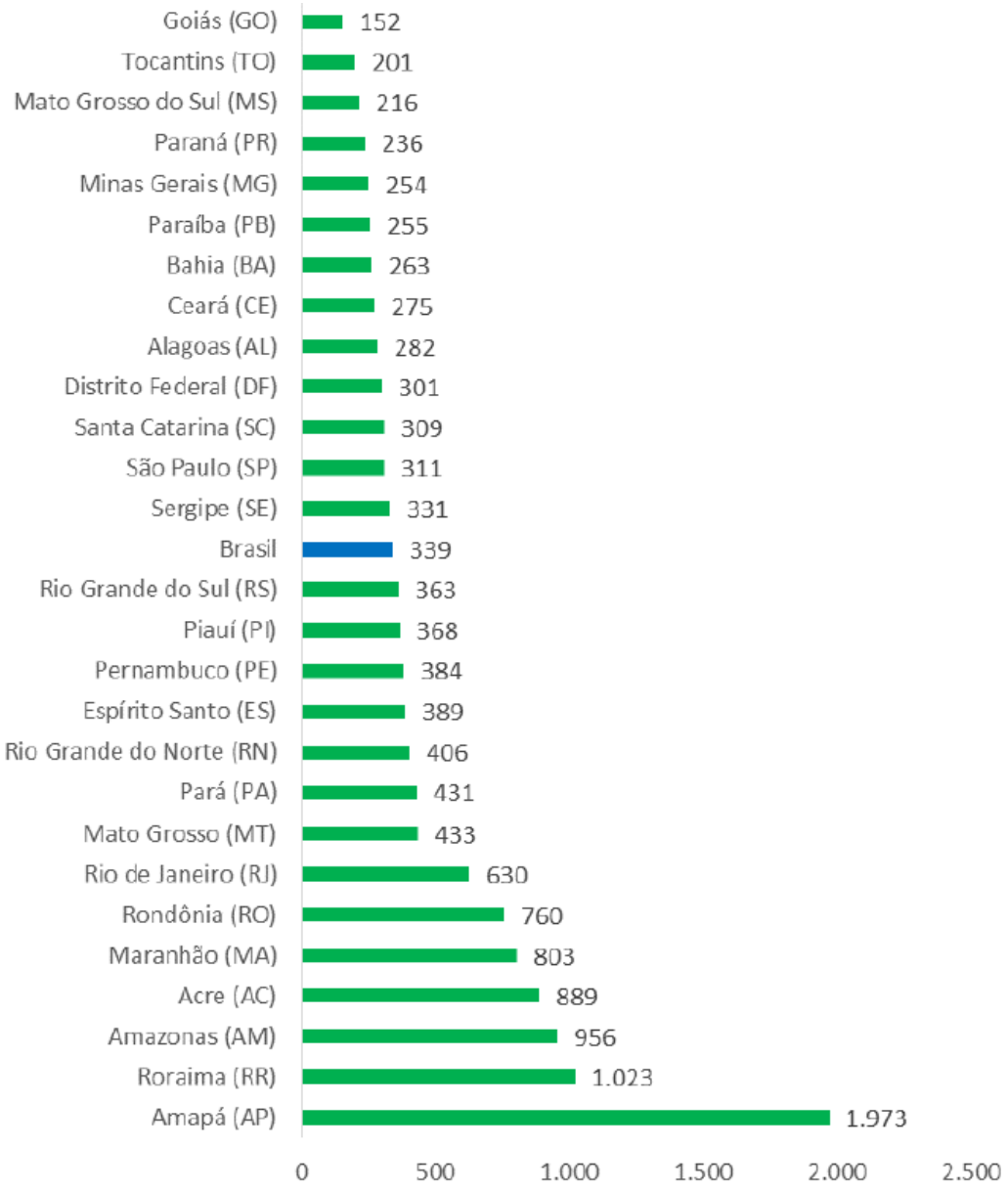


Figura 11 - Perdas por Ligação por Estados (%) (IN051 de 2019).

Fonte: ITB, 2021.

Há de se notar algumas peculiaridades, como por exemplo, a posição do Estado do Rio de Janeiro (RJ) que compõe a região Sudeste, a qual tem os melhores índices se comparado com outras regiões, perdendo apenas para o Centro-Oeste. Porém, o Estado tem índices que se equiparam a Estados do Norte e Nordeste, que são os piores no ranking.

Na Figura 9, é possível perceber que 11 dos 26 Estados do Brasil mais o Distrito Federal (DF), têm índices de perdas no faturamento maior ou igual a 50%. Isto torna a situação um tanto quanto preocupante.

No que diz respeito ao índice de perdas na distribuição, como pode ser visto na Figura 10, nove dos Estados da amostra tem indicativo próximo ou maior que 50%. Com grande destaque para o Amapá que tem 74% da sua distribuição de água perdida.

Por fim, o volume de água perdido pelos Estados está retratado na Figura 11. Segundo o SNIS, o padrão de excelência para perdas por ligação é 216 L/ligação/dia. Assim, a partir da figura, percebe-se que apenas Goiás, Tocantins e Mato Grosso do Sul compõem essa lista. Observa-se que o Estado do Rio de Janeiro apresenta um volume de perdas quase três vezes maior que o padrão de excelência e, de novo, o Amapá como destaque negativo, possui um índice de perda quase cinco vezes maior que o valor considerado de excelência.

2.3. Tipos de Perdas

Vazamentos, erros de medição, ligações clandestinas e consumos não autorizados, são alguns dos principais causadores de perdas de recursos hídricos em um processo de abastecimento de água por meio de redes de distribuição. Essas perdas significam prejuízo ao meio ambiente, à receita e aos custos de produção das empresas.

Uma rede de distribuição sem perdas é algo inviável em termos econômicos e técnicos, porém há um limite de excelência, como foi dito na seção anterior. Por isso, o volume de perdas de água constitui um índice importante com intuito de medir o desempenho dos prestadores de serviço em áreas como distribuição, planejamento, investimento e manutenção.

Antes da fundação da *International Water Association* (IWA) em 1998, a metodologia de avaliação dos prestadores de serviço era diferente entre países e empresas. A IWA padronizou o sistema de abastecimento, com um sistema denominado Balanço Hídrico. Esse sistema é basicamente uma matriz que estrutura os processos que a água pode passar desde o fornecimento até o cliente e hoje, já é comumente utilizado pelos reguladores e prestadores de serviços.

A partir da Tabela 1, percebe-se que o sistema começa com o volume de água produzido, podendo ser classificado como consumo autorizado ou como perda no processo de distribuição. O primeiro faz referência ao fornecimento hídrico aos clientes autorizados (medidos ou não). Já o segundo é a diferença entre o volume de entrada e o consumo autorizado.

Tabela 1 - Balanço Hídrico (IWA).

<table>
<tr><td rowspan="8">Água que entra no sistema (inclui água importada)</td><td rowspan="3">Consumo Autorizado</td><td rowspan="2">Consumo Autorizado Faturado</td><td>Consumo Faturado Medido (inclui água exportada)</td><td rowspan="2">Água Faturada</td></tr>
<tr><td>Consumo Faturado Não Medido (estimado)</td></tr>
<tr><td rowspan="2">Consumo Autorizado Não Faturado</td><td>Consumo Não Faturado medido (uso próprio, caminhão pipa, entre outros)</td><td rowspan="6">Água Não Faturada</td></tr>
<tr><td>Consumo Não Faturado Não Medido</td></tr>
<tr><td rowspan="5">Perdas de Água</td><td rowspan="2">Perdas Aparentes (Comerciais)</td><td>Uso Não Autorizado (fraudes e falhas de cadastro)</td></tr>
<tr><td>Erros de Medição (macro e micromedição)</td></tr>
<tr><td rowspan="3">Perdas Reais (Físicas)</td><td>Vazamentos e extravasamentos nos reservatórios (de adução e/ou distribuição)</td></tr>
<tr><td>Vazamentos nas adutoras e/ou redes (de distribuição)</td></tr>
<tr><td>Vazamentos nos ramais até o ponto de medição do cliente</td></tr>
</table>

Fonte: ITB, 2021.

O consumo autorizado é dividido em duas categorias, que por sua vez são separados em mais duas subcategorias:

I. Consumo Autorizado Faturado:

a. **Consumo faturado medido**: volume de água registrado no hidrômetro;
b. **Consumo faturado não medido**: representa o volume do consumo médio histórico ou, mínimo faturado quando se trata de lugares sem hidrômetro ou mal funcionamento do mesmo.

II. Consumo Autorizado Não Faturado:
a. **Consumo não faturado medido**: volume de água utilizado pela empresa para atividades operacionais especiais;
b. **Consumo não faturado não medido**: refere-se ao volume de água utilizado por razões sociais, como corpo de bombeiros.

As perdas reais (físicas) ou aparentes (comerciais) são classificadas pela IWA de acordo com sua natureza. A primeira representa a quantidade de água perdida durante as diferentes etapas de gerenciamento da água e está descrita no **Apêndice A**, enquanto que a segunda equivale ao volume de água consumido mas não autorizados nem faturados.

2.3.1. Perdas Aparentes (Comerciais)

As perdas aparentes ou comerciais representam a quantidade de água consumida que não é autorizada e muito menos faturada pela prestadora. Em outras palavras, são perdas motivadas por erros na medição do hidrômetro, fraudes, ligações clandestinas ou até mesmo falha no cadastro do cliente. A seguir são descritos os principais motivadores para as perdas comerciais:

I. Ligações clandestinas/irregulares;
II. Ligações sem hidrômetros;
III. Hidrômetros parados;
IV. Hidrômetros que subestimam o volume consumido;
V. Ligações inativas reabertas;
VI. Erros de leitura;
VII. Número de economias (quantidade de hidrômetros em um aglomerado, prédio, etc) errado.

Todos estes motivadores podem ter impacto significativo para a empresa dependendo dos procedimentos cadastrais e de faturamento, da manutenção preventiva, da adequação de hidrômetros e monitoramento do sistema. Além disso, destaca-se que o primeiro item (I) é encontrado com bastante frequência, seja em residências ou até mesmo em indústrias. Esse tipo de fraude consiste no impedimento total ou parcial da micromedição através de uma ligação adulterada. Basicamente são divididos em três tipos de adulteração:

➔ Derivação de ramal:

- O fraudador faz uma conexão antes do hidrômetro, fazendo com que a propriedade seja abastecida por esta nova conexão e seja parcialmente não medida (Figura 12).

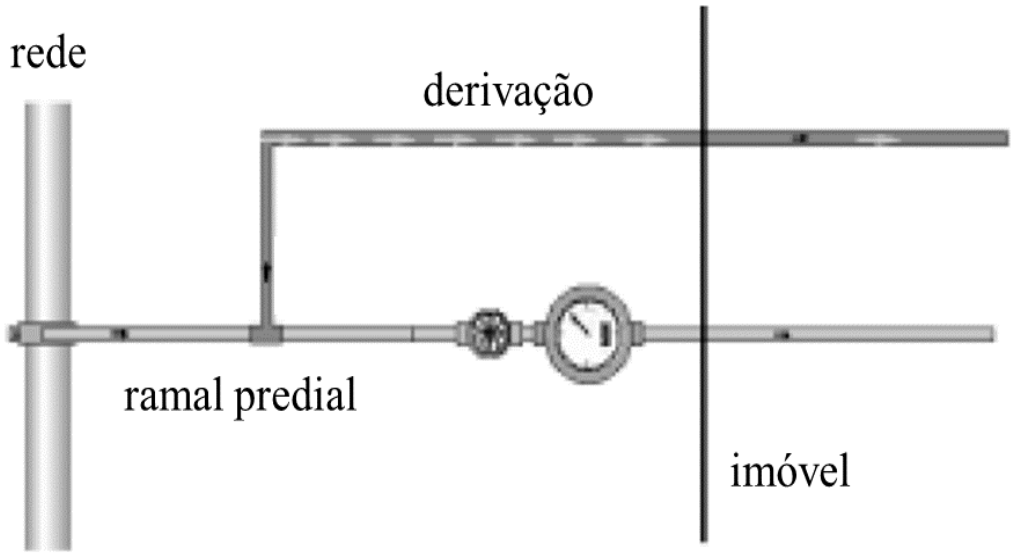


Figura 12 – Tipo de Irregularidade (Derivação de ramal).

Fonte: Modificado de Carvalho et al., 2004.

➔ *By-pass*:

- O fraudador faz a mesma conexão do exemplo anterior, porém com a diferença que essa nova conexão adulterada retorna para o ramal predial da propriedade, fazendo com que, nesse caso, o abastecimento seja totalmente sem medição (Figura 13).

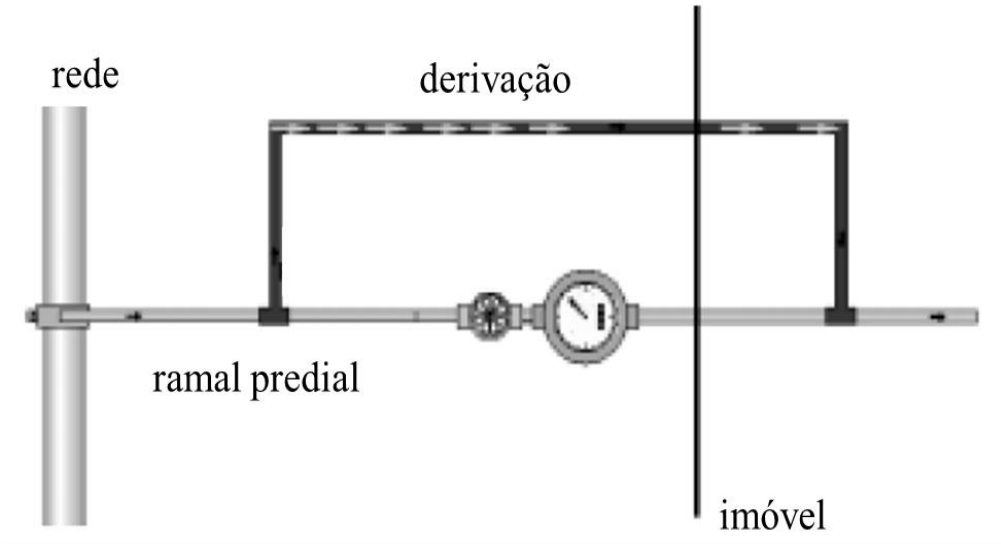


Figura 13 - Tipo de irregularidade (*By-pass*).

Fonte: Modificado de Carvalho et al., 2004.

➔ Ligação clandestina:

- O fraudador cria uma nova conexão a partir da rede de distribuição sem qualquer tipo de cadastro, registro ou contrato e, portanto, sem cobrança pelo seu consumo (Figura 14).

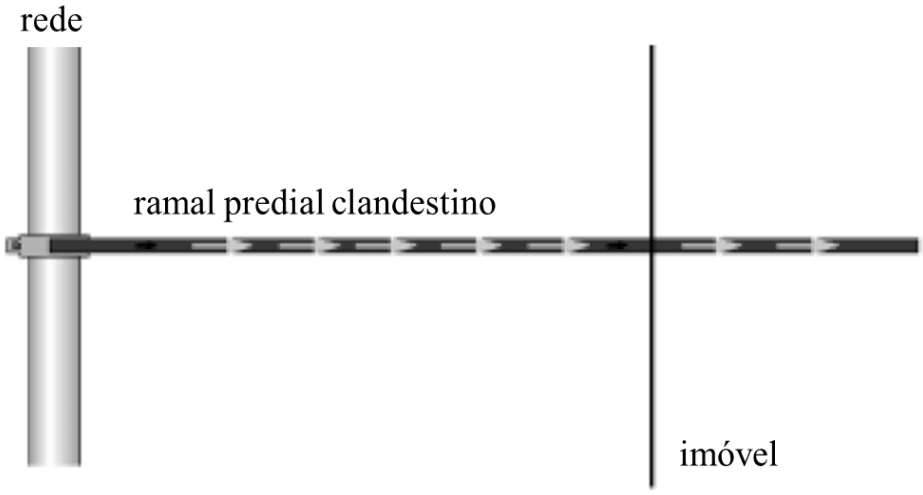


Figura 14 - Tipo de irregularidade (ligação clandestina).

Fonte: Modificado de Carvalho et al., 2004.

Pode-se concluir que as perdas aparentes têm impacto direto nas receitas das empresas, baseado no fato de que são consumidos volumes grandes de água que não são faturados. Isto faz com que os recursos disponíveis, que poderiam ser destinados à melhoria da infraestrutura e qualidade dos serviços, sejam drasticamente reduzidos.

2.4. Ganhos Hídricos e Econômicos com a Redução das Perdas

Para poder estimar os ganhos tanto hídricos quanto econômicos, precisa-se mensurar a quantidade de perdas de água (físicas e comerciais) no Brasil. Segundo o Instituto Trata Brasil, e utilizando dados do SNIS de 2019, o balanço hídrico do Brasil está apresentado na Tabela 2.

Tabela 2 - Balanço Hídrico no Brasil em 2019 (1.000 m³).

<table>
<tr><td rowspan="5">Água que entra no sistema (16.928.664)</td><td rowspan="2">Consumo autorizado faturado (10.058.746)</td><td>Consumo faturado medido (8.180.929)</td><td rowspan="2">Água faturada (10.058.746)</td></tr>
<tr><td>Consumo faturado não medido (1.877.817)</td></tr>
<tr><td colspan="2">Volume de serviços (862.693)</td><td rowspan="3">Água não faturada (6.869.919)</td></tr>
<tr><td colspan="2">Perdas comerciais (2.402.890)</td></tr>
<tr><td colspan="2">Perdas físicas (3.604.335)</td></tr>
</table>

Fonte: ITB, 2021.

De acordo com o Banco Mundial e a Tabela 4, pode-se considerar que de toda água não faturada em países desenvolvidos, 60% provém de perdas físicas e 40% das perdas comerciais sem contar com o volume de serviços (Liemberger, 2006).

> O volume perdido é suficiente para abastecer aproximadamente 63,1 milhões de brasileiros em um ano. Esta quantidade não somente equivale a pouco mais de 30% da população do país em 2019, como também corresponde a quase o dobro do número de habitantes sem acesso ao abastecimento de água nesse ano, cuja grandeza situa-se em torno de 33,2 milhões.
>
> Ao se admitir não a uma eliminação total das perdas, como no exercício acima, mas uma redução dos atuais 40,6% aos 25% pre-

> vistos em lei, o volume economizado seria da ordem de 2,2 bilhões de m³. Utilizando-se o mesmo consumo individual médio nacional empregado atualmente, isso equivale ao uso de aproximadamente 38,9 milhões de brasileiros em um ano, ou seja, quase 20% maior do que número de habitantes sem acesso ao abastecimento água em 2019. (ITB, 2021, p. 64)

Entrando no quesito econômico, pode-se estimar o impacto monetário que essas perdas trazem para as empresas e prestadoras, utilizando as tarifas:

- IN005 – Tarifa Média de Água: R$ 4,55 / m³ (SNIS, 2022);
- CMg Prod. Água – Custo marginal de produção de água: R$ 0,62 / m³ (SNIS, 2022).

Matematicamente estimando o impacto econômico das perdas, tem-se:

- $Impacto\ PF = Vol.PF\ (m^3) \times CMg\ Prod.Água\ (R\$/m^3)$
  - Impacto das perdas físicas é igual ao volume das perdas físicas multiplicado pelo custo marginal de produção de água.
- $Impacto\ PC = Vol.PC(m^3) \times IN005\ (R\$/m^3)$
  - Impacto das perdas comerciais é igual ao volume das perdas comerciais multiplicado pela tarifa média de água.
- $Impacto\ VS = Vol.AG024\ (m^3) \times CMg\ Prod.Água\ (R\$/m^3)$
  - Impacto do volume de serviços é igual ao volume de serviços multiplicado pelo custo marginal de produção de água.
- $Impacto\ Total = Impacto\ PF + Impacto\ PC + Impacto\ VS$
  - Já o impacto total é a soma dos três indicadores acima.

Como resultado, os valores obtidos estão detalhados na Tabela 3.

Tabela 3 - Impacto Econômico das Perdas no Brasil em 2019 (R$ 1.000).

Impacto PF	Impacto PC	Impacto VS	Impacto Total
2.234.687	10.933.149	534.869	13.702.705

Elaboração: Autor

Percebe-se que o impacto gerado por essas perdas, somado ao volume de serviços, dá um prejuízo estimado ao redor de 13,7 bilhões de reais, levando a uma situação bastante preocupante para as prestadoras.

Para tentar atingir bons índices no Brasil, o *Deutsche Gesellschaft für Internationale Zusammenarbeit GmbH* (GIZ), em conjunto com o Ministério das Cidades, elaborou um caderno temático chamado Projeto de Eficiência Energética em Sistemas de Abastecimento de Água (ProEESA) (FERREIRA et al., 2019). Esse plano tem como objetivo reduzir o índice de perdas em 10% até 2033 delineando estratégias e melhorando a eficiência energética no abastecimento de água. Dentre essas estratégias, podemos citar:

- Redução do consumo do usuário final: através da instalação de equipamentos mecânicos e eletrônicos nos banheiros para evitar desperdício e medidores individualizados nos condomínios;
- Redução de perdas de água nos sistemas de distribuição: revisão dos modelos tarifários, revisão das normas de construção associadas a redes de distribuição, etc;
- Melhoria da eficiência eletromecânica em sistemas de bombeamento de água nos sistemas de distribuição: através da normatividade regulatória que induza eficiência energética, entre outras estratégias;

Porém, outros esforços são necessários a fim de diminuir esse prejuízo que provém das perdas, principalmente, quando se trata de clientes irregulares e/ou fraudadores. Esses representam uma porcentagem de perdas comerciais, que sozinha, como visto acima, gera um impacto de quase 11 bilhões de reais.

2.5. Trabalhos Relacionados

O estudo de Al-Radaideh e Al-Zoubi (2018), propôs a utilização de detecção de clientes suspeitos de fraude utilizando a distância dos clientes irregulares para os clientes normais utilizando *Support Vector Machine* (SVM) e *K-Nearest Neiborgh* (KNN) sobre dados de cadastro e consumos por estações do ano. Os modelos foram aplicados em dados da Autoridade de Água da Jordânia (do inglês *Water Authority of Jordan* – WAJ). A área de concessão da empresa foi dividida em 10 regiões, e cada uma gerenciada por uma Unidade Organizacional Regional. As inspeções são feitas de forma aleatória aos clientes e, caso seja detectado roubo, o

mesmo é penalizado e registrado. Os dados são dispostos de 1990 até 2018, totalizando 16 milhões de registros e até 109 mil clientes. O valor de consumo era medido a cada três meses. A base dos clientes fraudadores (inspecionados aleatoriamente) foi concatenada com a base dos clientes normais e, em seguida, pré-processada com os critérios estabelecidos. Os algoritmos alcançaram acurácia em torno de 70% que supera, aparentemente, a taxa de acerto alcançado pela empresa que é 1%.

Já Gopal e Balaji (2020), também utilizaram as duas técnicas mencionadas anteriormente para separar esses clientes, na Índia. Porém, sobre os dados de consumo em confronto com a quantidade de água faturada, ou seja, o rótulo de fraudador é atribuído àquele cliente que a quantidade de água consumida fosse diferente do que a faturada. Ainda é feita uma análise gráfica no consumo desses clientes, o que ajuda a entender o comportamento e detectar a fraude de fato. O autor aponta que os algoritmos alcançaram uma acurácia de 70% e que melhorou as métricas alcançadas pela empresa (sem mencionar esse indicador). Outro ponto que não é mencionado pelo autor é a periodicidade dos dados coletados de consumo.

A partir disso, propõe-se o desenvolvimento de um sistema inteligente que aborda diversas técnicas de aprendizado de máquina, que possa aprender o comportamento de consumo dos clientes fraudadores e ajude na identificação dos mesmos. O sistema é composto por duas metodologias principais, e uma terceira que unifica as duas primeiras. Além disso, a métrica utilizada para avaliar o desempenho do sistema e dos algoritmos de forma geral, será a precisão ou taxa de acerto (detalhado mais a frente), pois se aproxima mais daquela utilizada pelas empresas distribuidoras de água.

As próximas seções irão detalhar todos os conceitos necessários, assim como o sistema inteligente em si.

3. Algoritmo Evolucionário com Inspiração Quântica (AEIQ)

Os problemas de otimização numérica são muito importantes, qualquer que seja a área com a qual estivermos lidando, como: mineração de dados, otimização de rota, minimização de custo, maximização de lucro, dentre outras.

Algoritmos Evolutivos (AEs) têm sido uma boa alternativa aos métodos clássicos de otimização, pois são facilmente adaptáveis a diferentes complexidades, além de lidar bem com ruídos ou problemas descontínuos, diferenciáveis ou multimodais (Back, 1997). Inspirado pela seleção natural de Darwin (1859), e pela genética molecular (Burian, 1996), os AEs definem uma otimização e metodologias de busca práticas e robustas. Quando comparados com algoritmos de otimização convencionais, os AEs providenciam uma abordagem geral para resolver problemas complexos, além de que suas capacidades de busca global, flexibilidades, performance robusta e adaptabilidade são considerados pontos fortes. Portanto, é uma área que está em constante evolução de pesquisa.

Hoje, existe uma diversa gama de algoritmos com diferentes características e pontos fortes. Os mais utilizados são: os algoritmos genéticos, programação genética, evolução diferencial, algoritmos culturais e programação evolutiva (Fogel et al., 1991, 1994, 1995, 1999). Apesar de solucionarem problemas de otimização de forma satisfatória, os algoritmos evolutivos, dependendo da complexidade, podem trazer problemas de desempenho, pois trata-se de algoritmos que avaliam muitas soluções por diversas vezes e isso pode ser custoso computacionalmente.

Dada a complexidade encontrada, nos últimos 30 anos, tem-se visto a aplicação de várias propriedades da física quântica a fim de construir uma nova gama de computadores, os computadores quânticos (Nielsen e Chuang, 2000). Diferentemente dos computadores clássicos, que lidam com dígitos binários (*bits*), computadores quânticos trabalham manipulando bits quânticos (*qubits*) que são a menor unidade de informação que pode ser armazenado em um computador quântico destes dois estados (Hey, 1999). Em vez dos estados comuns ‘0’ e ‘1’, um *qubit* pode também estar em uma superposição de dois estados, portanto uma partícula pode

efetivamente estar em muitos estados incompatíveis ao mesmo tempo (Nielsen e Chuang, 2000).

Entretanto, a limitação atualmente é a disponibilidade de máquinas robustas como essas. Assim, na década de 90, deu-se o início a uma pesquisa sobre os conceitos já existentes utilizando inspiração quântica que, nada mais é, utilizar os conceitos de física quântica como inspiração para o auxílio às ferramentas convencionais de computação, o que na prática pode potencializar seu desempenho.

Computação com inspiração quântica usa métodos computacionais baseados em conceitos e princípios da mecânica quântica, a fim de solucionar os vários problemas no contexto de um paradigma de computação clássica, tais como: Q-bits, superposição, portas lógicas quânticas e medições quânticas (Moore e Nayaranan, 1995). Este capítulo terá como foco os Algoritmos Evolucionários com Inspiração Quântica (QIEA's – *Quantum-Inspired Evolutionary Algorithms*)(Han e Kim, 2000). Esses algoritmos usam bits com inspiração quântica (Q-bits) para representar genótipos individuais; portas lógicas com inspiração quântica para operar sobre os Q-bits para gerar descendências; e usam os genótipos e fenótipos que são conectados por um processo de observação probabilístico. Em sistemas de mecânica quântica, o ato de observação faz com que a partícula quântica, que pode estar em mais de um estado simultâneo, seja observada em um único estado (Glassner, 2001).

Este capítulo irá explicar com detalhes os QIEAs, utilizando representação binária e representação real.

3.1. Algoritmos Evolucionários com Inspiração Quântica Utilizando Representação Binária

Foi proposto por Han e Kim (2000), um Algoritmo Evolucionário com Inspiração Quântica que utiliza representação binária (AEIQ-B). A proposta contemplava um algoritmo que era composto por um cromossomo, uma função de avaliação e uma dinâmica populacional. A representação desse cromossomo é feita de uma forma especial que simula um cromossomo formado por Q-bits. Cada Q-bit é composto por um par de números (α, β), como mostrado na Equação (5).

$$\begin{bmatrix} \alpha \\ \beta \end{bmatrix} \tag{5}$$

Onde $|\alpha|^2 + |\beta|^2 = 1$. Sendo que o valor de $|\alpha|^2$ dá a probabilidade de se encontrar o estado "0" e $|\beta|^2$ dá a probabilidade de que o Q-bit seja encontrado no estado "1". Um Q-bit pode estar no estado "1", no estado "0" ou em uma superposição linear dos dois estados.

A partir disso pode-se definir o indivíduo quântico formado por *m* Q-bits, que é definido pela Equação (6).

$$\begin{bmatrix} \alpha_1 | \alpha_2 | \cdots | \alpha_m \\ \beta_2 | \beta_2 | \cdots | \beta_m \end{bmatrix} \tag{6}$$

Onde $|\alpha_i|^2 + |\beta_i|^2 = 1, i = 1, 2, 3, 4, \ldots, m.$

Sendo assim, a representação por Q-bits tem a vantagem de poder representar uma superposição linear de estados. Utilizando como exemplo, na Equação (7), um sistema com 3 Q-bits e com 3 pares de amplitudes.

$$\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 1/2 \\ 1/\sqrt{2} & -1/\sqrt{2} & \sqrt{3}/2 \end{bmatrix} \tag{7}$$

Cada estado representado tem uma probabilidade de representar os estados equivalentes conforme a Tabela 4.

Tabela 4 - Probabilidade de observar cada possível estado do indivíduo quântico.

Estados	Probabilidade
$\|000\rangle$	$^1/_{16}$
$\|001\rangle$	$^3/_{16}$
$\|010\rangle$	$^1/_{16}$
$\|011\rangle$	$^3/_{16}$
$\|100\rangle$	$^1/_{16}$
$\|101\rangle$	$^3/_{16}$
$\|110\rangle$	$^1/_{16}$
$\|111\rangle$	$^3/_{16}$

Fonte: Elaborado pelo Autor.

De acordo com a Tabela 4, os estados do sistema podem ser representados conforme a Equação (8).

$$\frac{1}{4}|000\rangle + \frac{\sqrt{3}}{4}|001\rangle - \frac{1}{4}|010\rangle - \frac{\sqrt{3}}{4}|011\rangle + \frac{1}{4}|100\rangle + \frac{\sqrt{3}}{4}|101\rangle - \frac{1}{4}|110\rangle - \frac{\sqrt{3}}{4}|111\rangle \quad (8)$$

Cada termo representa a probabilidade de se encontrar os estados binários da equação.

O AEIQ-B é definido na Figura 15. O algoritmo é iniciado com um ou mais indivíduos quânticos, sendo esses inicializados de modo que os valores α_i e β_i (onde $i = 1, 2, 3, \ldots, m$; e m é o tamanho do indivíduo quântico) sejam todos iguais a $\frac{1}{\sqrt{2}}$. Fazendo com que todos os estados tenham a mesma probabilidade de serem observados nos estados "0" ou "1". Sabe-se que $Q(t)$ representa a população quântica, $P(t)$ representa a população clássica e $B(t)$ os melhores indivíduos da população clássica da geração t.


```
iniciar
    t ← 0;
    inicializa Q(t)
    gera P(t) observando estados de Q(t)
    avalia P(t)
    armazena as melhores soluções de P(t) em B(t)
    enquanto não ocorrer condição de parada
        t ← t + 1
        gera P(t) observando estados de Q(t − 1)
        avalia P(t)
        atualiza Q(t) usando q-gate
        armazena as melhores soluções de B(t − 1) e P(t) em B(t)
        armazena a melhor solução b de B(t)
    fim
fim
```

Figura 15 - Pseudocódigo do AEIQ-B.

Fonte: Abs da Cruz et al., 2006.

A atualização da população quântica, por sua vez, é feita através de um operador Q-gate, que é definido por uma matriz de rotação dada pela Equação (9).

$$U(\Delta\theta_i) = \begin{bmatrix} \cos(\Delta\theta_i) & -\sin(\Delta\theta_i) \\ \sin(\Delta\theta_i) & \cos(\Delta\theta_i) \end{bmatrix} \quad (9)$$

Cada coluna dos indivíduos quânticos será multiplicada pela matriz de rotação. Matematicamente, esta operação se dá pelo indivíduo quântico $Q(t) = \{(\alpha_1, \beta_1), (\alpha_2, \beta_2), \ldots , (\alpha_i, \beta_i)\}$ (onde $i = 1, 2, 3, \ldots, m$ e m é o tamanho do indivíduo quântico), esse indivíduo será atualizado de acordo com a Equação (10).

$$\begin{bmatrix} \alpha_i' \\ \beta_i' \end{bmatrix} = U(\Delta\theta_i) \begin{bmatrix} \alpha_i \\ \beta_i \end{bmatrix} \quad (10)$$

A matriz de rotação será capaz de modificar os valores de α_i e β_i, através de uma tabela que define os valores de $\Delta\theta$, a fim de aumentar a probabilidade de observação dos indivíduos com melhores avaliações. Em outras palavras, a representação desses estados é feita através de números complexos, e os eixos do círculo unitário elevados ao quadrado representam as probabilidades de os estados serem observados nos estados 0 ou 1. Portanto, a matriz rotação atualiza o vetor que aponta para esses eixos.

Por fim, ao longo das gerações, os melhores indivíduos gerados são armazenados em $B(t)$. A parte final do algoritmo é responsável por armazenar os melhores indivíduos gerados nas gerações anteriores com os melhores indivíduos da população atual (Han e Kim, 2002).

3.2. Algoritmos Evolucionários com Inspiração Quântica Utilizando Representação Real

A Figura 16 apresenta o pseudocódigo do Algoritmo Evolucionário com Inspiração Quântica utilizando representação com números reais (AEIQ–R).

O passo a passo do algoritmo está detalhado a seguir adentrando nos principais conceitos do modelo e como funciona.

```
iniciar
1.   t ← 1
2.   Gerar população quântica Q(t) com N indivíduos com G genes
3.   enquanto (t <= T)
4.       E(t) ← gerar indivíduos clássicos observando indivíduos quânticos
5.       se (t=1) então
6.           C(t) ← E(t)
7.       senão
8.           E(t) ← recombinação entre E(t) e C(t)
9.         avaliar E(t)
10.          C(t) ← K melhores indivíduos de [E(t) ∪ C(t)]
11.      fim se
12.      Q(t + 1) ← Atualiza Q(t) usando os N melhores indivíduos de C(t)
13.      t ← t + 1
14.  fim enquanto
fim
```

Figura 16 - Pseudocódigo do AEIQ–R.

Fonte: (Abs da Cruz et al., 2006)

3.2.1. População Quântica

Diferentemente dos algoritmos convencionais, nesse modelo, os que serão avaliados aqui, serão chamados de população clássica, os quais são gerados a partir da população quântica $Q(t)$. Essa população representa uma superposição de estados que são observados para criar a população clássica e então avaliados.

A população quântica é representada por um número de indivíduos quânticos. Cada indivíduo é formado por um número de genes, onde cada gene consiste em um par de valores ($g_{ij} = \mu_{ij}, \sigma_{ij}$) que representam, respectivamente, a média e a largura de um pulso quadrado. Esse pulso é utilizado pelo algoritmo para restringir um conjunto de possíveis valores observáveis dentro do domínio do problema que está sendo otimizado.

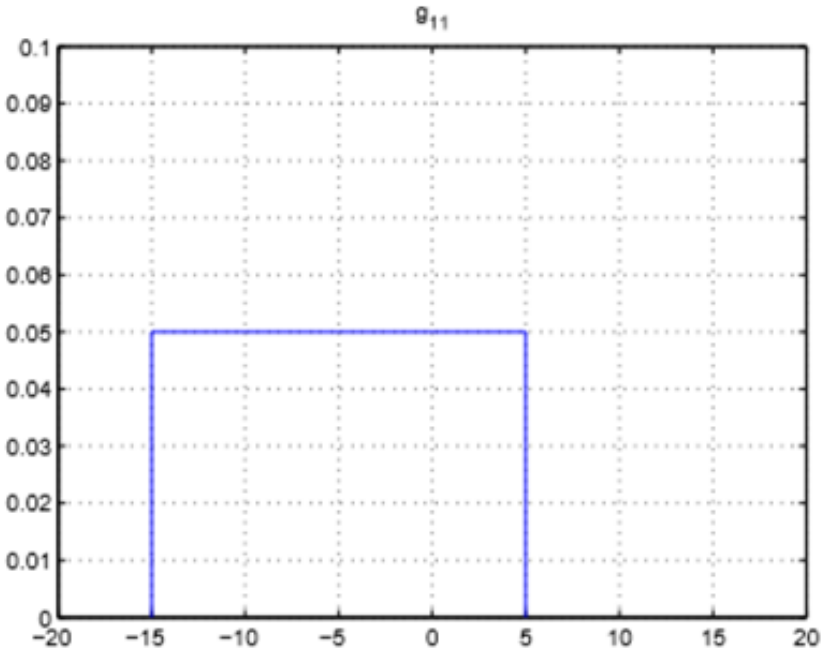


Figura 17 - Exemplo de um gene quântico ($g_{ij} = [-5,20]$).

Fonte: (Abs da Cruz et al., 2006).

Sendo assim, um indivíduo quântico representa um intervalo no campo de busca. A altura do pulso é calculada fazendo com que a área embaixo da curva seja igual a 1, pois representa a probabilidade do estado ser observado na largura do pulso, como mostrado na Equação (11).

$$h_{ij} * \sigma_{ij} = 1 \tag{11}$$

Sabendo que h_{ij} representa a altura do pulso e σ_{ij} a largura do pulso, para uma população com dois indivíduos, o pulso quadrado resultante está sendo representado na Figura 17.

O passo dois do pseudocódigo se dá pela criação da população quântica, que é feita gerando N indivíduos quânticos, com valor aleatório da média do pulso que esteja dentro do intervalo do domínio, e um valor de largura igual ao valor total do domínio.

3.2.2. Observação dos Indivíduos Quânticos

Após a criação da população quântica, o modelo entra no loop principal evolucionário. Este loop será executado pelo número estipulado de gerações T e é responsável por muitas tarefas.

A observação dos indivíduos quânticos é o passo primordial do algoritmo. Nesse passo o modelo cria os indivíduos clássicos a partir dos indivíduos quânticos. Em outras palavras, indivíduos cujos valores de genes são valores reais e dentro do domínio do problema. Esses genes quânticos, por se tratarem de uma função de pulso quadrado, representam a função densidade de probabilidade $p_{ij}(x)$ (FDP) usada pelo AEIQ–$\mathbb{R}$ para gerar os valores dos genes dos indivíduos clássicos. Sendo assim, a função de cada gene representa a densidade de probabilidade de se observar um determinado valor para o gene quântico, quando a superposição do mesmo for colapsada (Abs da Cruz et al., 2006).

Para cada gene é observado um número aleatório r no intervalo [0,1] e identifica-se o ponto x, dado a Equação (12).

$$P_{ij}(x) = \int_{-\infty}^{\infty} p_{ij}(t)\, dt \qquad (12)$$

Para exemplificar a geração da população clássica, considera-se uma população quântica formada por dois indivíduos, os quais possuem dois genes que têm, como função densidade de probabilidade, pulsos quadrados. Esses indivíduos estão detalhados na Tabela 5 e representados graficamente na Figura 18.

Tabela 5 - Exemplo de indivíduos quânticos que formam uma população quântica.

Indivíduo	Genes
q_1	$g_{11} = (\mu_{11} = -5, \sigma_{11} = 20)\ ;\ \ g_{12} = (\mu_{12} = 0, \sigma_{12} = 20)$
q_2	$g_{21} = (\mu_{21} = 5, \sigma_{21} = 20)\ ;\ \ g_{22} = (\mu_{22} = 5, \sigma_{22} = 20)$

Fonte: Abs da Cruz et al., 2006.

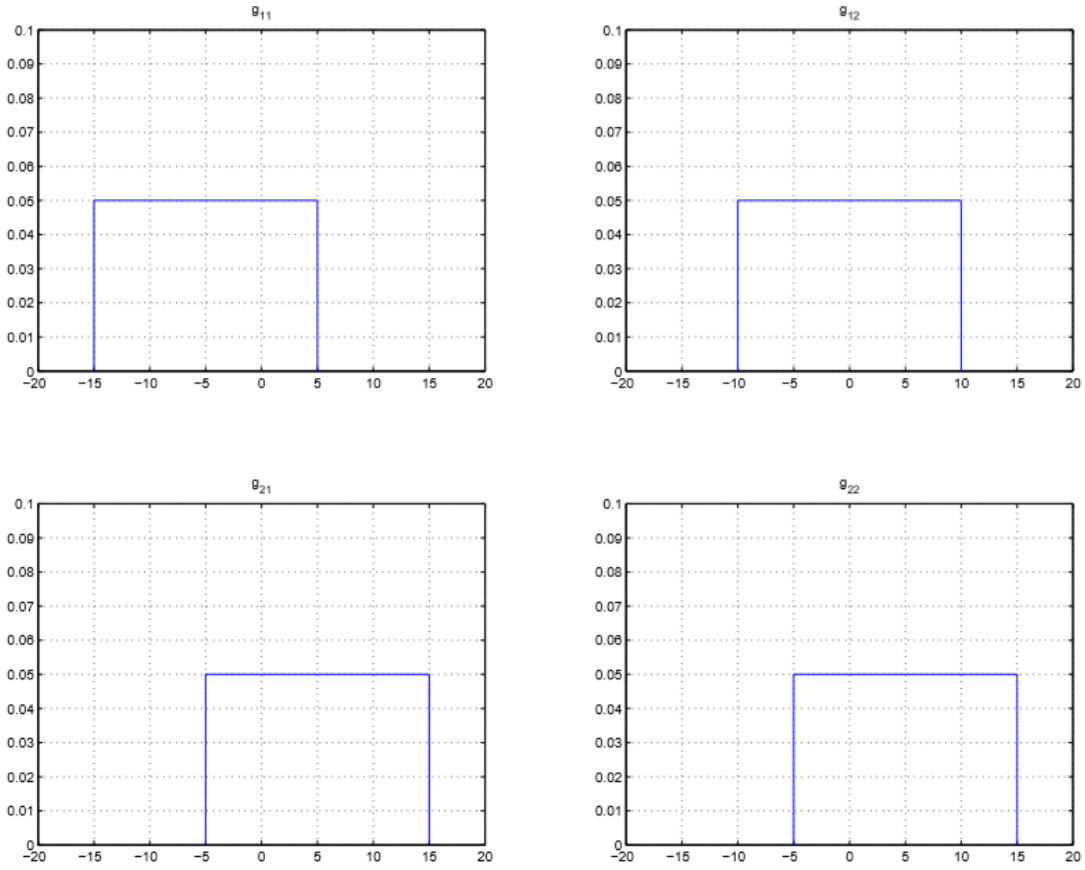


Figura 18 - Exemplos de genes quânticos utilizando função densidade de probabilidade de um pulso quadrado.

Fonte: Abs da Cruz et al., 2006.

Logo, a função $P_{ij}(x)$ (função cumulativa de probabilidade – FCP), pode ser representada por equações de reta, visto que as funções $p_{ij}(x)$ são constantes dentro dos intervalos e diferentes de 0. A Figura 19 mostra, graficamente, as FCPs relacionadas aos genes de exemplo.

Tendo as funções cumulativas dos genes quânticos é possível obter a população clássica. Dado que para cada gene, gera-se um número aleatório r_{ij} no intervalo [0,1] e, usando-se a equação de reta $y(x) = ax + b$, pode-se calcular o valor do gene clássico utilizando a Equação (13).

$$x_{mj} = r_{mj} * \sigma_{ij} + \left(\mu_{ij} - \frac{\sigma_{ij}}{2}\right) \qquad (13)$$

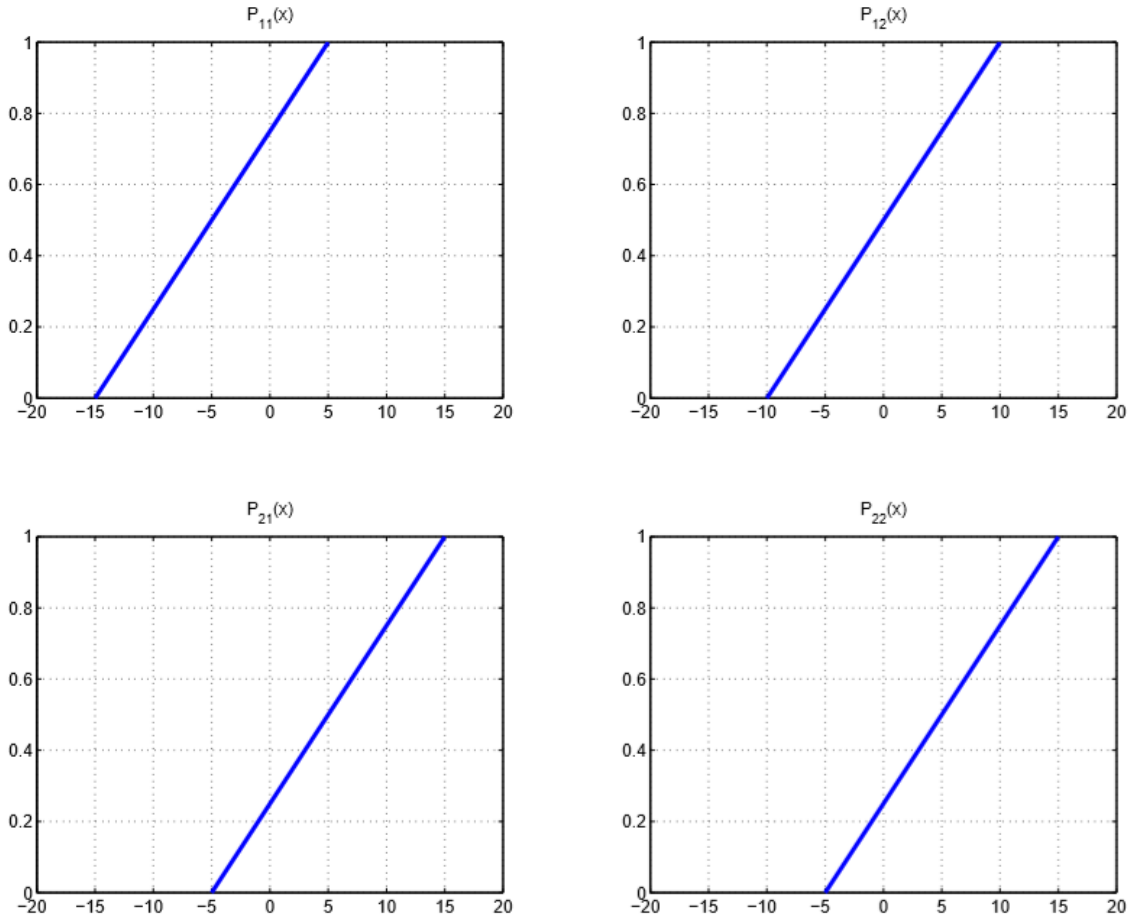


Figura 19 - Funções cumulativas de probabilidade resultante dos genes quânticos de exemplo .

Fonte: Abs da Cruz et al., 2006.

Onde x_{mj} representa o valor observado do indivíduo clássico m e do gene j, r_{mj} representa o valor do número aleatório gerado para aquele gene j do indivíduo m, enquanto que μ_{ij} e σ_{ij}, representam, respectivamente a posição do centro e a largura do pulso quadrado.

Uma vez gerada a população clássica $E(t)$, a partir das FCPs, e se não for a primeira geração do algoritmo – pois, na primeira geração uma população clássica temporária $C(t)$ é copiada da população clássica gerada $E(t)$ – é feita uma operação de *crossover* entre os indivíduos da população clássica gerada $E(t)$ e os indivíduos da população clássica temporária $C(t)$. O operador de *crossover* proposto está detalhado na Figura 20.

```
início
1.    para i = 1 até K
2.        seleciona o i-ésimo individuo e_i de E(t)
3.        seleciona o i-ésimo individuo c_i de C(t)
4.        para j = 1 até G
5.            r ← número aleatório no intervalo [0,1]
6.            se r < ξ
7.                e'_ij ← e_ij
8.            senão
9.                e'_ij ← c_ij
10.           fim se
11.       fim para
12.   fim para
fim
```

Figura 20 - Algoritmo de Crossover.

Fonte: Abs da Cruz et al., 2006.

Nesse algoritmo, K é o número de indivíduos na população clássica $C(t)$, G é o número de genes em cada indivíduo e o valor de ξ é a taxa de *crossover* utilizada durante o processo. Uma taxa igual a 1 irá copiar todos os valores dos genes no indivíduo criado, enquanto que uma taxa de 0 não irá modificar o indivíduo e nenhuma evolução será observada.

Após o uso do operador de *crossover*, a população temporária $E(t)$ precisa ser avaliada. Então, os K melhores indivíduos dessa população avaliada e da população $C(t)$ são selecionados, ordenados em um único conjunto e formarão uma nova população $C(t)$.

3.2.3. Atualização da População Quântica

No próximo passo, após a geração da população clássica, é necessário atualizar a população quântica. Esse processo é dividido em dois passos. O primeiro passo é modificar o valor do centro μ dos genes quânticos, fazendo com que o valor médio de cada gene seja igual ao valor do gene dos indivíduos clássicos ($\mu_{ij} = c_{ij}$). Onde μ_{ij} representa o valor do centro do gene j do indivíduo quântico i da população $Q(t)$, enquanto que c_{ij} o valor do gene j do indivíduo clássico i da população $C(t)$.

O segundo passo desse processo consiste em modificar a largura do pulso quadrado da função densidade de probabilidade dos genes quânticos. Essa alteração é feita em todos os genes dos indivíduos da população quântica. A heurística utilizada para tal modificação da largura é a regra do 1/5, que dá as seguintes condições: se no máximo 20% da população clássica criada tiver uma melhora de aptidão, a largura do pulso quadrado é reduzida; se essa taxa for maior que 20%, a largura do pulso é aumentada; se a quantidade de indivíduos que tiveram melhora da população for exatamente igual a 20%, não terá alteração na largura. Essa heurística pode ser representada pela Equação (14).

$$\sigma_{ij} = \begin{cases} \sigma_{ij}.\delta; & \varphi < 1/5 \\ \dfrac{\sigma_{ij}}{\delta}; & \varphi > 1/5 \\ \sigma_{ij}; & \varphi = 1/5 \end{cases} \tag{14}$$

Onde σ_{ij} é o valor da largura do gene j do indivíduo i da população quântica $Q(t)$, δ é um valor arbitrário no intervalo [0,1] e φ é a taxa da nova população que teve sua avaliação de aptidão melhorada.

Detrás de toda essa heurística, a ideia consiste em: se a quantidade de indivíduos da nova população, que teve sua aptidão melhorada, for maior que 1/5 da população, a largura do pulso deverá ser aumentada para encorajar uma busca global; se esses indivíduos representam menos do que 1/5 da nova população, a largura deverá ser reduzida para fazer mais buscas locais; se essa taxa for exatamente 1/5 da nova população a largura não sofrerá alterações, pois entende-se que 1/5 da população melhorada em relação à anterior é considerada uma taxa ideal. Essa atualização pode ser feita a cada geração, ou pode se definir um número Z de gerações (e.g a cada 5 gerações).

A Figura 21 representa um diagrama completo sobre o algoritmo da Figura 15, nele pode ser visto que a população quântica $Q(t)$ é composta de indivíduos quânticos que são representados por genes quânticos. A partir desta população, gera-se uma população clássica que, por sua vez, é avaliada e usada pra atualizar a população quântica em um loop iterativo de T gerações.

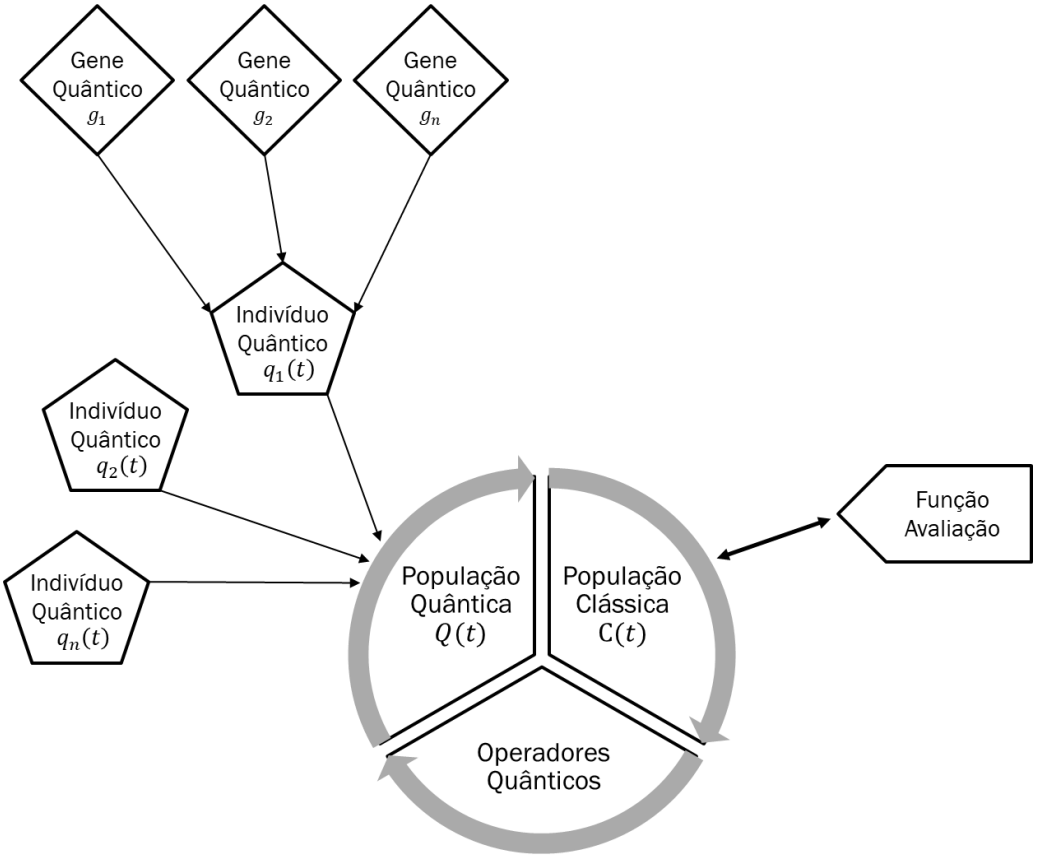


Figura 21 - Diagrama do algoritmo evolucionário com inspiração quântica.

Fonte: Modificado de Abs da Cruz et al., 2006.

3.3. Algoritmos Evolucionários com Inspiração Quântica Utilizando Representação Híbrida

O Algoritmo Evolucionário com Inspiração Quântica utilizando representação híbrida (AEIQ-BR) define um cromossomo que possui ambos tipos de genes quânticos: binário e real (Pinho et al., 2009). Esse cromossomo pode ser representado como:

$$q_j^t = \left[\left(q_j^t\right)_b \left(q_j^t\right)_r\right] = \left[\begin{pmatrix} \alpha_{j1}^t & \alpha_{j2}^t & \cdots & \alpha_{jk}^t \\ \beta_{j1}^t & \beta_{j2}^t & \cdots & \beta_{jk}^t \end{pmatrix} \begin{pmatrix} \mu_{j1}^t & \mu_{j2}^t & \cdots & \mu_{jk}^t \\ \sigma_{j1}^t & \sigma_{j2}^t & \cdots & \sigma_{jk}^t \end{pmatrix}\right] \quad (15)$$

Onde o primeiro termo do cromossomo representa a parte binária e o segundo termo a parte real.

O algoritmo completo está mostrado na Figura 22.

1. $t \leftarrow 0$;
2. criação da população quântica $Q(t)$ com representação híbrida;
3. **enquanto** $t < T$;
4. $t \leftarrow t + 1$
5. geração da população clássica $P(t)$ com representação híbrida observando $Q(t)$;
6. Avalia $P(t)$;
7. **se** $t = 1$ **então**;
8. $B(t) \leftarrow P(t)$;
9. **senão**;
10. $P(t) \leftarrow$ recombinação clássica entre $P(t) \cup B(t-1)$;
11. Avalia $P(t)$;
12. $B(t) \leftarrow$ melhores indivíduos de $P(t)$ $B(t-1)$;
13. $Q(t+1) \leftarrow$ atualiza $Q(t)$ usando os N melhores indivíduos de $C(t)$;
14. Atualiza a parte binária de $Q(t)$ usando os melhores indivíduos de $B(t)$ e um Q-gate;
15. Atualiza a parte real de $Q(t)$ usando os melhores indivíduos de $B(t)$ e um Q-crossover;
16. **fim se**;
17. **fim enquanto.**

Figura 22 - Pseudocódigo do algoritmo AEIQ-BR.

Fonte: Modificado de Pinho et al., 2009.

No passo 2, cada gene de $Q(t)$ é inicializado com probabilidades iguais para todos os estados. Pela parte binária, cada Q-bits é igual a $1/\sqrt{2}$. Já na parte real, μ e σ representam, respectivamente, o centro e a largura do intervalo do domínio do problema. Sabe-se que, inicialmente, q_j^0 é a superposição linear entre todos os possíveis estados, com probabilidades iguais de ocorrência.

A população clássica $P(t)$ é criada, no passo 5, baseada nos estados quânticos da população quântica $Q(t)$. Para cada Q-bit gera-se um número aleatório entre 0 e 1. Se esse número estiver entre 0 e α^2 então o bit clássico gerado é 0; senão o bit clássico será 1. Para a parte real, um número no intervalo $(\mu - \sigma)$ até $(\mu + \sigma)$ é aleatoriamente escolhido.

A avaliação do passo 6 consiste em uma função objetivo que depende do tipo de problema que o algoritmo está aplicado. No caso de problemas de classificação, a avaliação de cada individuo q_j^t considera o número total de registros corretamente classificados em uma classe. Portanto, a função de avaliação pode ser calculada pela Equação (16):

$$f(objetivo)_{classificação} = \frac{a_{ij} + \sum_{i=j>1}^{n} \frac{c_1 . a_i}{c_j}}{\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}} \quad (16)$$

Onde a_{ij} é o número de padrões da classe i classificados como classe j, n é o total de número de padrões, e c_j é o número total de padrões na classe j. c_1/c_j é um termo importante pois mantém todas as classes balanceadas in termos de tamanho do conjunto de treino.

A atualização do Q-bit da parte binária é feita pela média de um operador de rotação Q-gate (Han e Kin, 2000, 2002), definido pela Equação (17):

$$\begin{pmatrix} \alpha_{ij}^t + 1 \\ \beta_{ij}^t + 1 \end{pmatrix} = \begin{pmatrix} \cos(\Delta\theta) & -\text{sen}(\Delta\theta) \\ \text{sen}(\Delta\theta) & \cos(\Delta\theta) \end{pmatrix} \begin{pmatrix} \alpha_{ij}^t \\ \beta_{ij}^t \end{pmatrix} \quad (17)$$

$\Delta\theta$ é um parâmetro de rotação do ângulo que é dependente do problema que define quanto o vetor Q-bit será movido em direção às chances de os estados 0 ou 1 serem escolhidos na próxima geração.

Os indivíduos clássicos da geração anterior serão utilizados para atualizar os indivíduos quânticos dos genes da parte real q_{jk}^t. Considerando que g_{jk}^t é o k-ésimo gene do j-ésimo indivíduo clássico na geração t e μ_{jk}^t é o centro do k-ésimo gene do j-ésimo indivíduo quântico, a atualização de μ_{jk}^t ocorre de acordo com a Equação (18):

$$\mu_{jk}^{t+1} = \mu_{jk}^t + \left(g_{jk}^t - \mu_{jk}^t\right) u(0,1) \quad (18)$$

Onde $u(0,1)$ é um número aleatório no intervalo [0,1] escolhido a partir de uma distribuição normal.

A Figura 23 mostra o fluxograma do algoritmo evolucionário com inspiração quântica utilizando representação híbrida. A partir dela, observa-se que a população clássica $P(t)$ é gerada a partir da população quântica $Q(t)$ e, então, é combinada com a população clássica da geração anterior $B(t-1)$. Em seguida a população clássica recombinada é avaliada e os melhores indivíduos são armazenados $B(t)$ e, utilizados para atualizar a população quântica $Q(t)$ e a população clássica $B(t-1)$.

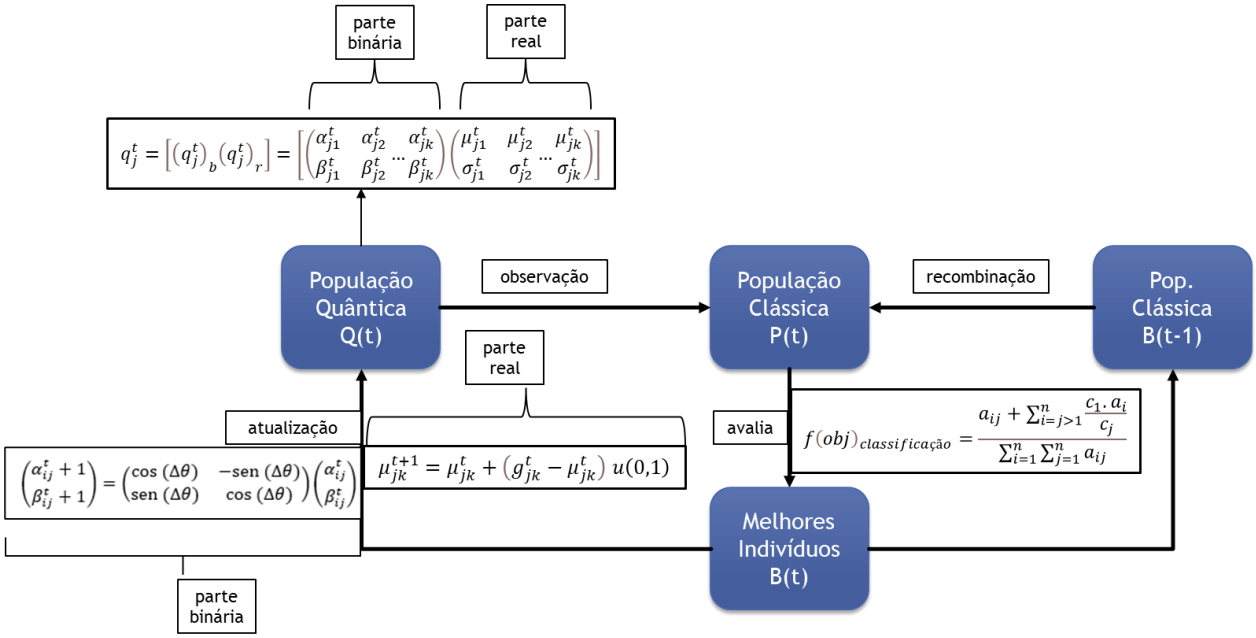


Figura 23 - Fluxograma AEIQ-BR.

A principal vantagem da utilização das representações real e híbrida, em relação à binária, é oferecer uma abordagem alternativa além de uma representação numérica mais direta, enquanto mantém as melhorias de performance relacionadas à inspiração quântica. Por isso, neste trabalho, será utilizado o AEIQ com representação híbrida, pois deseja-se utilizar o algoritmo para busca de parâmetros e atributos.

4. Sistema de Detecção de Suspeito de Fraude

Este capítulo é responsável por explicar e detalhar o sistema desenvolvido para detecção do suspeito de irregularidade/fraude no consumo de água.

O sistema de detecção é composto principalmente por duas metodologias (Figura 24). A primeira, denominada Metodologia por Filtragem (Metodologia F), é constituída por uma filtragem não-supervisionada responsável por selecionar os registros com maior probabilidade de serem identificados como Fraudadores. Em seguida é feita uma busca de hiperparâmetros utilizando *GridSearch.* Além disso, é formada por cinco algoritmos de classificação amplamente utilizados e conhecidos, são eles: *MultiLayer Perceptron* (MLP) (Rosenblatt, 1968), *Support Vector Machine* (SVM) (Vapnik, 2000), *Adaboost* (ADAB) (Schapire, 2013), *Random Forest* (RF) (Breiman, 2001) e *Xtreme Gradient Boost* (XGB) (Chen, 2016).

A segunda metodologia, denominada Metodologia Evolutiva (Metodologia E), utiliza uma abordagem diferente. Sendo composta principalmente por um Algoritmo Evolucionário com Inspiração Quântica utilizando representação híbrida (AEIQ-BR) que será utilizado para busca de parâmetros e atributos. Essa metodologia também utiliza, para classificação, os algoritmos: MLP, SVM, RF e XGB, com exceção do ADAB, que foi substituído pelo *Decision Tree* (DT) (Blockeel, 1998) pois, nas avaliações iniciais, não se mostrou promissor.

A Figura 24 apresenta o diagrama de blocos geral do sistema inteligente.

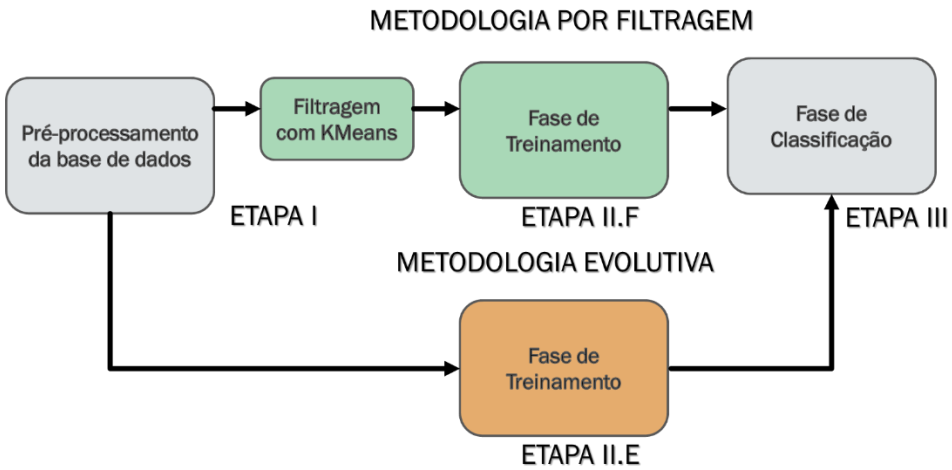


Figura 24 – Diagrama de Blocos do Sistema Inteligente.

Para as duas metodologias cada modelo fornece um resultado, porém, o sistema ainda fornece a possibilidade de combinar estes resultados utilizando um comitê de classificadores

Cada metodologia é constituída por, basicamente, três etapas: Pré-processamento da base de dados, Fase de Treinamento e Fase de Classificação. As duas metodologias compartilham o pré-processamento da Etapa I, com exceção da Filtragem com KMeans que é exclusiva para a Metodologia F. A Etapa II é diferente para as metodologias, por isso serão denominadas Etapa II.F e Etapa II.E. Já a Etapa III é similar para as duas metodologias.

Por fim, pretende-se juntar o que há de melhor nas duas metodologias e abordar uma terceira abordagem, chamada de Metodologia Completa. Nela utiliza-se tanto a filtragem (Metodologia F) de registros como o processo evolutivo (Metodologia E).

4.1. Metodologia por Filtragem

Na Etapa I, responsável pelo pré-processamento da base de dados, são realizados os processos de Extração, Limpeza, Criação de Variáveis, Normalização e Codificação, como pode ser visto na Figura 25. Essa etapa atua diretamente na construção da base de dados de treinamento e de teste, com foco em transformar a base de dados da empresa – composta por dados de cadastro, série histórica de consumo, inspeções, etc – em um *dataset* adequado para a utilização nos algoritmos de aprendizado de máquina.

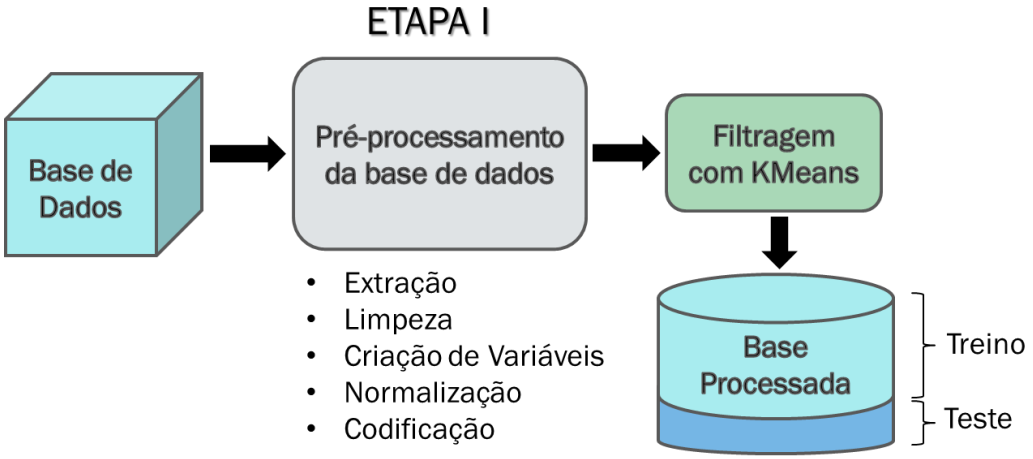


Figura 25 – Diagrama de Blocos Etapa I - Pré-Processamento da Base de Dados.

4.1.1. Etapa I – Pré-Processamento da Base de Dados

O sistema baseado em aprendizado de máquina visa identificar os clientes suspeitos de fraude e de irregularidade no consumo de água, a partir do consumo histórico proveniente da base de dados da empresa prestadora de serviço. Antes de entrar no detalhamento das etapas e processos, alguns pontos devem ser ressaltados.

O sistema usa os dados fornecidos para discriminar as classes Fraudador e Não Fraudador. Por isso, é desejável que esses dados sejam pré-processados a fim de que os algoritmos não tenham dificuldade em classificá-los. Sabendo disso:

- Foram considerados apenas clientes de categoria Residencial nesta dissertação. A separação desses clientes residenciais e não residenciais (comercial, industrial e público) é bastante oportuna pois eles possuem perfis de consumo distintos. Além disso, a distribuição de registros por categoria, por parte das empresas de saneamento, tende a ser bastante desbalanceada, sendo a grande maioria dos dados pertencente à categoria residencial.
- Foram consideradas apenas Irregularidades que afetam o consumo, pois o sistema fará a inferência considerando os dados de consumo histórico do cliente. Portanto, se houver alguma irregularidade que não afete o consumo, esta não poderá ser detectada pelos modelos.

Além desses, outros procedimentos são necessários para a obtenção de uma base de dados consistente. A seguir, estão descritos os processos utilizados na Etapa I, de Pré-Processamento de dados.

a. Extração e Limpeza dos Dados

A limpeza da base de dados é um procedimento necessário para qualquer trabalho envolvendo inferência. Responsável por excluir valores considerados absurdos ou discrepantes, geralmente oriundos de erros de preenchimento da base. Esses valores reduzem a qualidade do aprendizado e não melhoram a capacidade da inferência quando algo semelhante ocorre novamente. Assim, deve-se verificar suas ocorrências e realizar o tratamento adequado. Além disso, existem valores faltantes. Existem diversas técnicas que podem ajudar a preencher valores faltantes

em uma base de dados. Desprezá-los é a melhor forma de garantir que o resultado não seja enviesado por técnicas de preenchimento, porém há de se ressaltar que essa opção depende de uma grande abundância de registros. Como é o caso desta dissertação.

No processo de Extração todas as informações de interesse são extraídas da base fornecida pela empresa, como: consumo histórico mensal, consumo histórico faturado, ordens de serviço de interesse, quantidade de economias na ligação, número de ligação e etc. Nesse processo a base de dados original é processada por diversos filtros de interesse para que possa reunir apenas as informações que agreguem ao sistema.

O processo de Extração inicia-se desprezando os seguintes tipos de clientes, são eles:

- Registros cujos clientes não estão em situação ativa (variável que indica a situação do cliente: ativa ou diferente de ativa). Se o cliente não possui uma ligação ativa naquele mês, dificilmente terá uma medição coletada;
- Linhas que indicam o tipo de faturamento com valores diferentes dos normalmente utilizados;
- Clientes sem consumo registrados por hidrômetro, ou seja, possuem ligação com consumo fixo. Esses clientes não têm variação no perfil de consumo e, sendo assim, são registros ruins para qualquer tipo de inferência já que seus valores de consumo histórico não variam;
- Clientes com volume consumido igual a zero;
- Clientes com CPF em mais de uma ligação, fazendo com que não seja possível a detecção de mudança de cliente em um determinado imóvel ou economia, pois, pode se tratar de um imóvel alugado e com isso o consumo não estar relacionado ao responsável.

No próximo passo, filtrou-se a base considerando apenas as irregularidades identificadas e inspeções criadas, que fizessem alusão a algum tipo de alteração no consumo, ou seja, qualquer ordem de serviço emitida ou irregularidade apontada que pudesse afetar o consumo de água do cliente. Essas irregularidades são compostas pelos tipos: *by-pass*, violação no hidrômetro, hidrômetro invertido, hidrômetro danificado. Já as ordens de serviços criadas são aquelas requeridas com intuito

principal em vistoriar o cliente em busca de fraude ou algum tipo de irregularidade, e, como foi mencionado acima, é importante que esta vistoria tenha sido confirmada tanto positivamente (gerando a irregularidade) quanto negativamente (apontando que não há fraude).

O processo de Limpeza está atrelado ao processo de extração, dado que já foram retirados alguns registros que não possuíam valores de consumo ou estes eram iguais a zero.

Outro tratamento que foi feito na base de dados tem relação com os valores incongruentes, ou seja, valores considerados muito altos, que geralmente são oriundos do fato do medidor ter "zerado" naquele mês. Além desse, quando não há leitura no mês ou em um mês subsequente a outro com leitura. Os dois tipos de problemas foram resolvidos considerando o volume faturado, pois, se o volume medido é maior que o faturado, utilizou-se o volume faturado (ou cobrado) mais a diferença do crédito – variável que aponta se cliente tem algo a ser ressarcido da última leitura – do mês atual e o anterior.

b. Criação de Atributos

Com a base de dados extraída e limpa, os próximos passos foram construir os atributos de interesse e também atribuir uma classe aos registros.

O primeiro foi feito através de cálculos (média móvel, quantidade de meses abaixo de heurísticas, etc) utilizando valores da base de dados original, para que possam complementar as informações de consumo do cliente. São eles:

- Consumo histórico: 6 últimos meses;
- Média móvel de consumo: 3 meses, 6 meses e 12 meses;
- Quantidade de meses (em um ano) com consumo abaixo de 5 m^3;
- Quantidade de meses (em um ano) com consumo abaixo de 75% do consumo médio anual;
- Quantidade de meses (em um ano) com consumo abaixo de 50% do consumo médio anual.

O segundo foi basicamente atrelar o valor de saída aos registros baseados nas condições que já foram citadas anteriormente, uma vez que a base já está filtrada

apenas pelas vistorias que representam algum tipo de inspeção (ordens de serviço) e por irregularidades que confirmam o cliente fraudador. Sendo assim, os clientes considerados Fraudadores foram aqueles com inspeção/vistoriados e irregularidades apontadas e confirmadas, e os clientes Não Fraudadores foram aqueles com inspeção/vistoriados e sem irregularidade apontada.

Basicamente cada registro é classificado como acima, tendo como entrada as observações dos últimos 12 meses de consumo, a contar da data da inspeção. Em outras palavras, cada linha da base de dados é composta pelo histórico de consumo dos últimos 12 meses prévios ao mês da inspeção.

c. Normalização e Codificação da Base de Dados

A normalização dos valores numéricos e a codificação de atributos categóricos – para algoritmos cujas entradas necessariamente precisam ser numéricas, tais como as redes neurais – transformam os dados visando uniformizar as ordens de grandeza com intuito de evitar que atributos de maior ordem possuam "pesos" indiretamente associados.

Os atributos podem ser separados em variáveis quantitativas (numéricas) ou qualitativas (categóricas). As quantitativas naturalmente representam grandezas numéricas contínuas ou inteiras. As qualitativas podem ser ordinais ou nominais, sendo que para as ordinais é possível se estabelecer uma ordem (alto, médio ou baixo), diferentemente das nominais, em que não se pode associar qualquer tipo de ordem (caso da classificação do tipo de cliente: residencial, comercial, industrial, etc.).

Os dois tipos de normalização mais utilizados são mostrados nas Equações (19) e (20).

$$y = \frac{x - \mu}{\sigma} \tag{19}$$

$$y = \frac{x - min}{max - min} \tag{20}$$

Sabe-se que y representa o valor normalizado; x o valor original; μ e σ, são a média e o desvio padrão, respectivamente, dos valores considerados na normalização; e por fim, max e min, representam o valor máximo e o mínimo observado no atributo em questão.

Optou-se pela normalização da faixa de variação (mínimo e máximo). E a codificação normalmente utilizada em variáveis qualitativas, consiste em substituir o atributo contendo n diferentes categorias por n variáveis binárias que assumem valores 0 ou 1 (*one-hot*).

d. Filtragem Utilizando KMeans

A motivação dessa filtragem reside na dificuldade encontrada em diferenciar o que é um perfil normal de um fraudador. Os fraudadores, de uma forma geral, não possuem um padrão anômalo que facilite a sua identificação. Esta técnica foi utilizada separadamente para os dois perfis, fraudadores e não fraudadores.

Como exemplo do que foi mencionado acima, pode-se citar que um dos comportamentos considerados como fraudulento é o aumento ou a queda repentina do consumo de água, e esse também está presente em registros rotulados como não fraudadores. Desse modo, a utilização de um modelo sem uma seleção prévia dos usuários poderia resultar em mais casos falsos-positivos.

O módulo de filtragem possui duas etapas importantes. Primeiramente, é utilizado um algoritmo de agrupamento (k-Means), que gera k grupos utilizando os seguintes registros baseados nos atributos mencionados anteriormente.

De todos os grupos gerados, são selecionados aqueles que possuem uma proporção de fraudadores maior do que a proporção geral entre fraudadores e não fraudadores na base de dados.

A escolha do hiperparâmetro k é feita de tal modo que maximize a quantidade de fraudadores após a seleção dos grupos. É importante ressaltar que este agrupamento é aplicado posteriormente para a filtragem dos registros no conjunto de testes, ou seja, será considerado como fraudador os elementos do conjunto de teste que mais se aproximarem do centróide do grupo k escolhido a partir do treinamento.

Outra etapa importante do módulo de filtragem é a preparação dos registros para o treinamento do modelo. De fato, foi observado o problema de classes desbalanceadas, visto que, pode existir na base de dados uma quantidade muito maior de não fraudadores. Para amenizar esse problema para o aprendizado dos modelos, são utilizadas duas técnicas conhecidas: Subamostragem, que seleciona aleatoriamente elementos da classe majoritária e os remove até que a quantidade de elementos da classe majoritária e da minoritária possuam o mesmo valor. Essa técnica é usada para que a base de dados não fique desbalanceada para o treinamento dos modelos; e método Tomek Link que verifica e elimina o vizinho da classe majoritária mais próximo do elemento da classe minoritária. Essa técnica tem como vantagem fazer com que a separação entre as classes de fraudadores e não fraudadores seja melhor definida no conjunto de dados que está sendo tratado, ou seja, tem o potencial de aumentar a largura da fronteira entre as duas classes.

O *dataset* resultante é constituído das seguintes variáveis:

- Código de leitura (atributo que representa a descrição, através de códigos, da leitura do mês de referência – leitura normal, sem leitura, etc.);
- Consumo histórico: 6 últimos meses;
- Média móvel de consumo: 3 meses, 6 meses e 12 meses;
- Quantidade de meses (em um ano) com consumo abaixo de 5 m^3;
- Quantidade de meses (em um ano) com consumo abaixo de 75% do consumo médio anual;
- Quantidade de meses (em um ano) com consumo abaixo de 50% do consumo médio anual.

Após a etapa de filtragem, os dados selecionados foram utilizados para o treinamento dos modelos de classificação. Para a utilização em um conjunto de testes, a única etapa adicional necessária para a classificação foi a filtragem usando o k-Means.

4.1.2. Etapa II.F – Fase de Treinamento

Esta etapa é responsável por fazer o treinamento dos modelos a partir da base de dados gerada pela etapa anterior. Como pode ser visto na Figura 26.

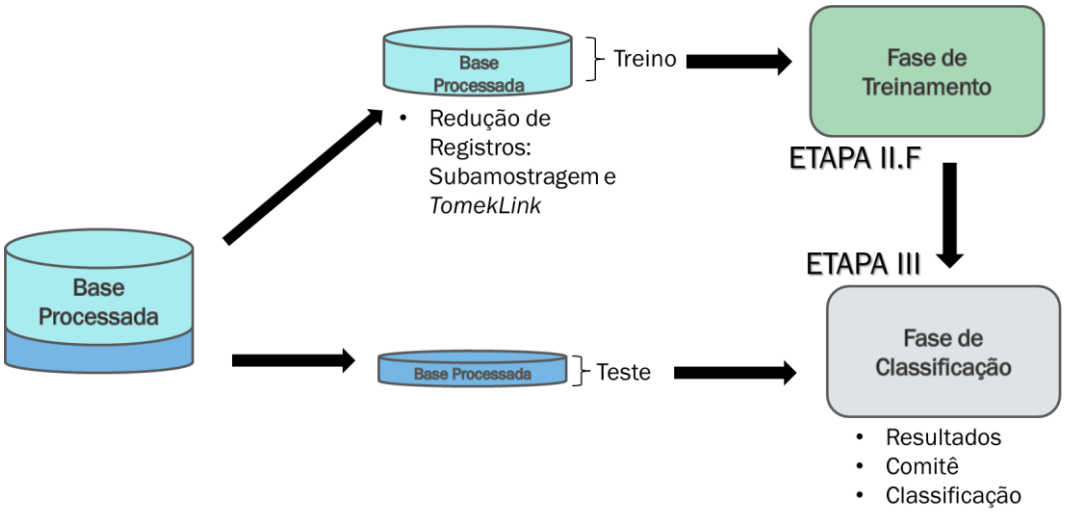


Figura 26 – Diagrama de Blocos Etapas II.F e III da Metodologia F (Fase de Treinamento e Classificação).

Antes do treinamento, como dito anteriormente, foi utilizado a técnica de busca de parâmetros que seriam utilizados para o treinamento desses algoritmos.

Foram testadas diferentes configurações de hiperparâmetros, através do método *GridSearch*. Para cada modelo, foram testados os seguintes parâmetros:

- *Random Forest* (RF):
  - Profundidade máxima: 3, 4, 5;
  - Características máxima: 2, 3, 4;
  - Qtd. de estimadores: 100, 300, 1000, 3000.
- *eXtreme Gradient Boost* (XGBoost):
  - Profundidade máxima: 3, 4, 5;
  - Gama: 0.5, 1.0, 1.5, 2, 5;
  - Peso mínimo: 1, 5, 10;
- *Adaboost* (ADAB):
  - Qtd. de estimadores: 300, 500, 700;
  - Taxa de aprendizado: 0.01, 0.1, 0.3, 0.5, 1.

- *Multilayer Perceptron* (MLP):
    - Neurônios na camada escondida: 10, 20, 50, 100;
    - Taxa de aprendizado: 0.001, 0.002, 0.005, 0.01.
- *Support Vector Machine* (SVM):
    - Kernel: linear, RBF;
    - C: 1, 2, 5, 10, 20.

Logo em seguida, todos os algoritmos tem seus conjuntos de hiperparâmetros ideais para a execução do treinamento. A próxima etapa é a de classificação do conjunto de teste.

4.1.3. Etapa III – Fase de Classificação

Nessa etapa inicia-se a fase de classificação e avaliação dos modelos. Com o resultado obtido nas etapas anteriores utiliza-se o conjunto de testes que foi inicialmente separado e nunca utilizado pelos algoritmos treinados. Uma vez que é executado o teste, tem-se as métricas individuais de cada modelo, as quais foram complementadas com a combinação dos mesmos através de um comitê de classificadores. Vale ressaltar que cada algoritmo tem como saída a probabilidade de dado cliente ser fraudador e, portanto, o comitê combina, por diferentes meios, as probabilidades desses algoritmos (detalhado a seguir).

Tendo as métricas alcançadas por cada modelo e o comitê, dependendo da necessidade, há a opção de se aplicar um limiar de decisão, que será explicado nas próximas seções.

A partir disso, o modelo classifica os clientes gerando uma lista ordenada com os clientes mais prováveis de estarem cometendo uma fraude baseado na precisão alcançada pelos testes.

a. Comitê de Classificadores e Limiar de Decisão

Como mencionado anteriormente, a partir da execução do conjunto de teste nos modelos treinados obtém-se as métricas e a precisão de cada modelo separadamente. O próximo passo foi combinar o resultado desses modelos utilizando um comitê de classificadores, ou seja, utilizar a decisão de cada modelo de forma conjunta. Dessa forma, foi possível tornar ainda mais robusta a confiança nas decisões resultantes para a classe fraudadora e não fraudadora, ao estabelecer um consenso entre os diferentes classificadores.

A utilização de Comitê de Classificadores – ou Ensemble – já é bastante explorada na literatura. Como tal, existem diversas abordagens que podem ser usadas na definição do comitê. Nesta dissertação foram aplicadas as abordagens a seguir: (Kuncheva, 2004):

- Voto Majoritário (VM) → Classifica os registros em função dos votos de cada classificador, ou seja, cada modelo tem peso de um voto e o registro será classificado com a maioria dos votos.
- Soma Ponderada (SP) → Classifica os registros mediante a soma ponderada das decisões dos modelos baseado em alguma métrica escolhida (precisão no caso deste trabalho).
- Fusão de Probabilidade (FProb) → Classifica os registros baseado no produto direto das probabilidades que cada modelo obtém de saída.

O sistema proposto também possui a capacidade de, a partir das probabilidades dos modelos, definir um limiar que aumenta a certeza dada pelo classificador quando infere os clientes. Com essa técnica, denominada Limiar de Decisão, é possível aumentar o percentual de acerto do sistema ao custo de reduzir a quantidade de inspeções que o mesmo indicaria. Em outras palavras, tomando como base o exemplo da Figura 27, supõe-se que o sistema indicou 1500 ligações com o limiar de decisão em 0,5 (50%); a empresa poderia optar por diminuir o número de inspeções a serem realizadas, aumentando o limiar de decisão (grau de certeza do modelo). No exemplo fictício, pode-se ter algo em torno de 900 inspeções com o limiar de decisão em 0,65 (65%). Essa variação no limiar resultaria em um aumento de 0,22 para algo em torno de 0,32, de taxa de acerto do modelo.

Essa técnica permite que a empresa tenha total liberdade de decidir qual caminho seguir baseado no resultado obtido pelos modelos, fazendo com que o sistema tenha um desempenho mais confiável.

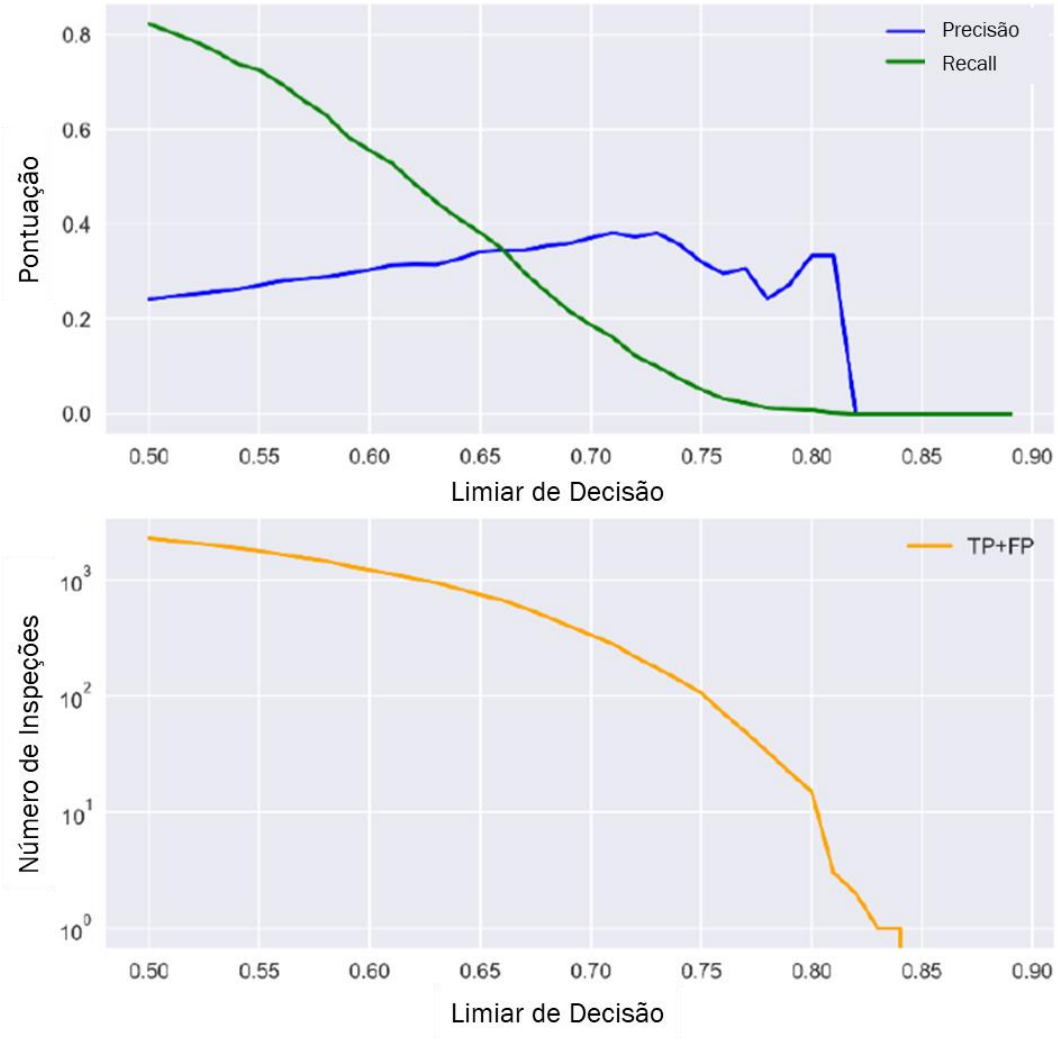


Figura 27 – Exemplo da variação da precisão em função do Limiar de Decisão.

4.2. Metodologia Evolutiva

Observou-se que a Metodologia F reduzia bastante a quantidade de registros presentes nos conjuntos de treino e teste, dependendo da região. Além disso, sabe-se que os grupos gerados pela Filtragem da Etapa I, apesar de não-supervisionada, é impulsionada quando se sabe a classe do cliente, o que difere de uma situação real. Portanto, na Metodologia E, optou-se por utilizar todos os dados provenientes

do Pré-Processamento da Etapa I, sem utilizar o processo de filtragem não-supervisionado.

4.2.1. Etapa I – Pré-Processamento da Base de Dados

A etapa I da Metodologia Evolutiva contém todos os pré-processamentos já descritos na seção 4.1.1, portanto, já foram apresentados. A única exceção é que a Metodologia E não possui o processo de filtragem de KMeans, como pode ser visto na Figura 28.

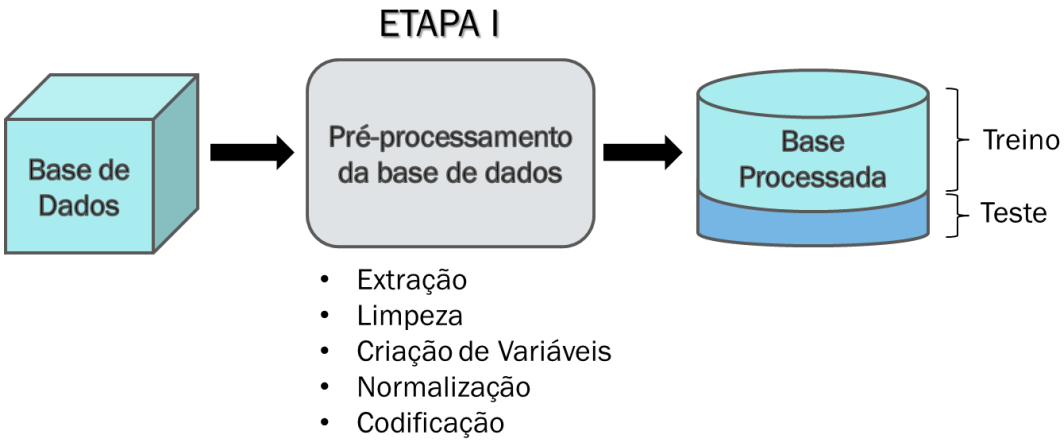


Figura 28 - Diagrama de Blocos Etapa I - Pré-Processamento da Base de Dados.

Além dos atributos utilizados na Metodologia F, optou-se por acrescentar, a partir da série temporal de consumo, mais alguns atributos que são combinações das informações dos consumos históricos observados no registro, descritos na Tabela 6. Também foram adicionadas algumas variáveis exógenas, ou seja, atributos que possam trazer informação externa, de outras fontes, à base de dados (clima, temperaturas, etc.). O objetivo é auxiliar o modelo a incluir características que são comumente encontradas em séries que possam destacar, por exemplo, sua sazonalidade.

Tabela 6 – Atributos Adicionados.

Série Temporal de 12 meses	Variáveis Exógenas mensais
Janela dos últimos 6 meses	Precipitação
Média dos últimos 3 meses	Umidade
Média dos últimos 6 meses	Temperatura Média
Média dos últimos 12 meses	Temperatura Mínima
Diferenças no consumo entre os valores dos últimos 3 meses	Temperatura Máxima
Meses com consumo abaixo de 5m³	Qtd. Feriados
Meses com consumo abaixo de 75% da média anual	
Meses com consumo abaixo de 50% da média anual	
Meses com consumo constante	
Maior aumento de consumo mensal	
Maior queda de consumo mensal	
Diferença entre o maior aumento e a maior queda	
Diferença entre o consumo mensal com o do ano anterior	
Sazonalidade	
Inclinação da linha de regressão calculada a partir dos valores da série de 12 meses	

A seguir os atributos estão descritos mais detalhadamente:

- Janela 6 meses → Leitura de cada um dos 6 meses que antecedem a identificação do cliente (mês 7, mês 8, mês 9, mês 10, mês 11, mês 12);
- Média dos últimos 3 (6 e 12) meses → média simples da janela de consumo;
- Diferença dos últimos 3 meses → são 3 atributos diferentes que representam a diferente entre o último e o penúltimo mês (mês 12- mês 11), entre o penúltimo e o antepenúltimo (mês 11- mês 10) e entre o último e o antepenúltimo (mês 12- mês 10);

- Meses com consumo abaixo → representa a quantidade de meses da série que ficaram abaixo das condições definidas (abaixo de 5m³, abaixo de 75% da média anual e abaixo de 50% da média anual);
- Meses com consumo constante → indica a quantidade de meses da série com consumo constante;
- Maior aumento, Maior queda e Diferença entre eles → estes atributos representam a maior variação positiva e negativa entre dois meses sucessivos, além da diferença entre esses indicadores;
- Diferença com o ano anterior → este indicador faz a diferença entre o consumo do mês de referência do ano atual e o ano anterior;
- Sazonalidade → é o atributo que indica literalmente em qual estação do ano se encontra o mês de referência;
- Inclinação da linha de regressão → aqui é feita uma regressão da série temporal de 12 meses de consumo e então calculada a inclinação dessa regressão;
- Precipitação → quantificação da chuva no determinado mês da região de interesse;
- Umidade → quantificação da umidade no determinado mês da região de interesse;
- Temperaturas média, mínima e máxima → atributos que indicam a temperatura alcançada no determinado mês da região;
- Quantidade de Feriados → indica quantos feriados no período da série de 12 meses.

O diagrama de blocos da Etapa II.2 pode ser vista na Figura 29.

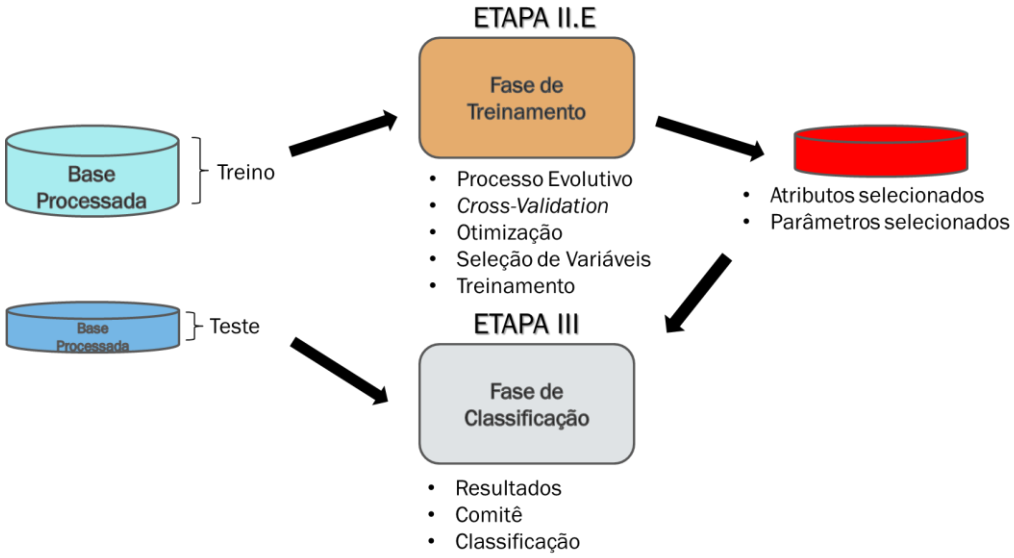


Figura 29 - Diagrama de Blocos Etapas II.2 e III da Metodologia E (Fase de Treinamento e Classificação).

4.2.2. Etapa II.E – Fase de Treinamento

A principal diferença da Metodologia E, se dá pelo desenvolvimento do Algoritmo Genético com Inspiração Quântica (AEIQ-BR) que foi implementado para melhorar a busca de parâmetros e de seleção de atributos, dado que a quantidade destes aumentou.

a. Processo Evolutivo e Função de Avaliação

Neste passo o AEIQ-BR e todo seu processo evolutivo foram criados. Em outras palavras, foi criada a população quântica formada por indivíduos quânticos. A população clássica é gerada através da observação da população quântica. Cada indivíduo, dessa nova população, representa uma solução do problema que, nesse caso, é o conjunto de parâmetros de cada algoritmo (mesmos algoritmos utilizados na Metodologia F, porém com DT substituindo o ADAB) e o conjunto de variáveis que se deseja usar para classificar as entradas em suspeito ou não de fraude.

A parametrização do algoritmo AEIQ-BR baseou-se nos estudos de Abs da Cruz e outros (2006), Pinho e outros (2009), e Ramos e outros (2016), adequando-

se os mesmos ao problema em questão. A Tabela 7 indica os parâmetros utilizados no AEIQ-BR.

Tabela 7 - Parâmetros do AEIQ-BR.

Quantidade de Indivíduos Quânticos	4
Número de Observações	2
Quantidade de Indivíduos Clássicos	8
Gerações	100
Taxa de Mutação	0.5
Mμ	0.95
Taxa de *Crossover*	0.8
Delta Theta	$0.001 * \pi$
Modelo	{MLP, SVM, DT, RF e XGBoost}
Métrica da Função Avaliação do Algoritmo	precisão

A quantidade de indivíduos representa o tamanho da população quântica e cada indivíduo representa uma solução, ou seja, quanto maior a população mais soluções são testadas a cada geração. O número de observações se dá pela quantidade de vezes que cada indivíduo quântico será testado com intuito de criar a população clássica.

As taxas de mutação e *crossover* representam a probabilidade de os indivíduos serem atualizados através da mutação do gene e o cruzamento de indivíduos selecionados pela roleta, respectivamente.

Já o Mμ é um parâmetro que representa a magnitude em que o centro dos genes quânticos é atualizado em relação às representações reais (ver Equação (13)). Já o *delta theta* representa a magnitude de atualização do ângulo (Equação (9)), que representa os valores binários.

A busca de parâmetros é feita para cada modelo de forma separada e a métrica precisão (*precision*), obtida na classificação do conjunto de validação é utilizada na aptidão de cada indivíduo. A opção pela métrica precisão será explicada com mais detalhes nas próximas seções.

Vale ressaltar que toda vez que o algoritmo avalia os indivíduos, ou seja, testa os valores dos genes de cada cromossomo, isso é feito por validação cruzada (*Cross-Validation*) com 5 *folds* e uma subamostragem (já descrita anteriormente) do conjunto de treino, resultando na média da precisão (métrica escolhida) dos 5-*folds*, criados considerando a subamostragem. Assim, esse procedimento avalia os modelos durante a evolução.

b. Busca de Parâmetros e Seleção de Variáveis

O algoritmo genético usou *steady-state*, ou seja, assegurou que os melhores indivíduos de cada geração continuariam na geração seguinte.

Uma vez que todo processo evolutivo foi finalizado, o resultado obtido foi o melhor indivíduo que representa os melhores parâmetros para aquele algoritmo especifico e o melhor conjunto de atributos. Os atributos utilizados foram descritos na seção 4.1.2.

Os conjuntos de hiperparâmetros de cada modelo utilizados na busca, estão descritos na Tabela 8.

Tabela 8 - Hiperparâmetros de cada algoritmo supervisionado.

MLP	Numéricos: • *Neurônios na camada escondida [2-20];* • *Alpha [0.001-0.1];* • *Tamanho do batch [200-1000];* • *Taxa de aprendizado inicial [0.0001-0.5];* • *Momento [0-1].* Categóricos: • *Função de Ativação da camada escondida (tanh, logistic, relu);* • *Otimizador (stochastic gradient descent, adam);* • *Taxa de aprendizado (constant, adaptative).*
SVM	Numéricos: • *C [1-10];* • *gama [0.001-0.01].* Categóricos: • *Kernel (linear, rbf).*
DT	Numéricos: • *Mínimo número de amostras por nó interno [2-50];*

	• *Mínimo número de amostras por folha[1-50];* • *Custo de complexidade [0-0.02]*. Categóricos: • *Critério de separação (gini, entropy).* • *Estratégia de separação (best, random).*
RF	Numéricos: • *Número de árvores [2-50];* • *Mínimo número de amostras por nó interno [2-50];* • *Mínimo número de amostras por folha[1-50];* • *Custo de complexidade [0-0.02]*. Categóricos: • *Critério de separação (gini, entropy).* • *Estratégia de separação (best, random).*
XGB	Numéricos: • *Número de rodadas de esforço [2-50];* • *Taxa de aprendizado [0.01-0.3];* • *Redução mínima de perda [0.0-0.4];* • *Profundidade máxima da árvore [2-10]*.

A Tabela 8 detalha os hiperparâmetros dos algoritmos assim como o intervalo de valores que cada um foi testado. Esses parâmetros estão divididos em valores categóricos e numéricos. Os valores numéricos são representados no algoritmo de forma real, enquanto os categóricos são representados de forma binária (tendo apenas a função de ativação das redes MLPs sido configuradas como binária *one-hot*), assim como os atributos.

De posse do resultado do processo evolutivo, o próximo passo foi treinar os modelos com o melhor conjunto de hiperparâmetros e atributos.

A Etapa III é a mesma descrita anteriormente e, portanto, já foi apresentada.

Por fim, a Metodologia Completa (Metodologia C), possui a filtragem utilizando K-Means para selecionar os registros de uma melhor forma, e também possui o processo evolutivo descrito anteriormente, a fim de buscar os melhores parâmetros e atributos utilizados.

A próxima seção detalhará as métricas utilizadas para analisar o desempenho do sistema frente às métricas alcançadas pela empresa.

4.3. Métricas de Análise de Desempenho

A fim de analisar qualquer resultado de classificação ou desempenho de um modelo de inferência, utiliza-se a matriz de confusão (Figura 30). Nela pode-se concluir que os dados contidos nas células indicam o número de exemplos que possuem referente classificação, sendo que as linhas representam os clientes inspecionados/situação real na base de dados da empresa. Enquanto que as colunas representam a classificação dada pelos modelos e/ou comitês.

A partir disso, pode-se calcular algumas métricas importantes. A acurácia (do inglês - *accuracy*) do modelo é obtida pela quantidade total de acertos pela quantidade total de dados avaliados. O *recall* apresenta a percentagem de amostras positivas classificadas corretamente sobre o total de amostras positivas (real), ou seja, a quantidade de exemplos positivos acertados em relação ao total de positivos (real).

		Valor Previsto (Predito pelo modelo no teste)	
		negativo	positivo
Valor real (Confirmado por análise)	negativo	**VN** Verdadeiro Negativo	**FP** Falso Positivo
	positivo	**FN** Falso Negativo	**VP** Verdadeiro Positivo

Figura 30 - Exemplo de matriz confusão.

Já a precisão (*precision*) define a porcentagem de exemplos que o modelo indicou como positivo e acertou, em relação ao total de exemplos indicados como positivo pelo modelo. Essa métrica se assemelha a métrica taxa de acerto aplicada pela empresa, visto que se baseia na quantidade de indicados corretamente como fraudadores pelo modelo sobre o total de clientes inspecionados.

Por isso a métrica utilizada para avaliar o desempenho do sistema perante ao da empresa será a métrica precisão, e será denominado taxa de acerto do sistema.

- Taxa de acerto do Sistema (tx_acert_{modelo}):

$$tx\_acert_{modelo} = \frac{VP}{VP + FP} \tag{21}$$

Sabe-se que VP representa a quantidade de clientes indicados corretamente como fraudadores pelo sistema e FP a quantidade de clientes indicados incorretamente como fraudadores pelo sistema.

- Taxa de acerto da Empresa ($tx\_acert_{empresa}$):

$$tx\_acert_{empresa} = \frac{clientes\ fraudadores\ confirmados}{todos\ os\ clientes\ inspecionados} \tag{22}$$

Portanto, a taxa de acerto da empresa é definida como o número total de clientes irregulares comprovados sobre o número total de clientes avaliados.

4.3.1. Teste Estatístico

Por fim, serão utilizados dois testes estatísticos com intuito de testar a hipótese nula de identidade dos diferentes *datasets* gerados pelos diferentes cenários (detalhados nas próximas seções).

O *Mann-Whitney U Test* é um teste de significância estatística não-paramétrico que possibilita determinar se duas amostras independentes – elementos das amostras provém de indivíduos distintos – possuem a mesma distribuição (Corder e Foreman, 2011). A suposição padrão ou hipótese nula é que não há diferença entre as distribuições das duas amostras. Rejeitar a Hipótese Nula (H_0) sugere que há uma provável diferença entre elas e, em um problema de classificação, pode-se garantir que um algoritmo é melhor que o outro.

Em alguns casos as amostras podem ser pareadas, ou dependentes. Neste caso, o *Wilcoxon Signed-Rank Test* é um teste estatístico não-paramétrico que é usado para comparar duas amostras dependentes. Amostras dependentes, em aprendizado de máquina, podem representar tanto o mesmo algoritmo em *datasets* diferentes ou algoritmos diferentes avaliando *datasets* nas mesmas condições de treino e teste (Corder e Foreman, 2011). Assim como o teste anterior, o teste de *Wilcoxon*,

tem como hipótese nula a identidade das duas amostras, ou seja, os dois conjuntos de dados possuem a mesma distribuição.

5. Estudo de Caso

Esse capítulo é responsável por apresentar toda a aplicação do sistema inteligente, descrito no capítulo anterior, em um caso real fornecido pela Empresa A que atua, dentre muitas regiões, principalmente na Região G, Região P e Região T.

Conforme a Tabela 9, pode-se observar os indicadores (IPFT, IN049 – 25% de índice ótimo – e IN051 – 316 L/ligação/dia de índice ótimo), apresentados no capítulo dois, e pode-se concluir que a Região G é a que mais se aproxima dos valores ótimos estipulados pelos órgãos responsáveis. Por outro lado, a Região T está bem distante desses valores, sendo considerada a região mais problemática nesse sentido. Já a Região P apresenta valores médios, visto que se trata de um conglomerado de municípios (diferente das outras regiões) e o único indicador que ficou acima do valor ótimo é o de perdas na distribuição.

Tabela 9 – Índices de Perdas das Cidades da Empresa A.

	IPFT (%)	IN049 (%)	IN051 (L/lig./dia)
Região G	23,3	19,3	114,1
Região P (médio)	11,3	30,6	165,3
Região T	36,3	43,9	314,8

Fonte: SNIS, 2022. Elaboração: Autor.

Neste capítulo são descritas todas as etapas e procedimentos do sistema aplicado diretamente em clientes inspecionados, destacando-se as Etapas de Pré-processamento, a de Classificação, e por fim, a indicação de suspeitos de fraude.

Um foco maior será dado à Metodologia E pois, devido à natureza das abordagens desenvolvidas, a Metodologia F desconsidera um número significativo de registros da base de dados e isso limitaria algumas das análises propostas.

5.1. Construção do Dataset Geral

Como dito anteriormente, o estudo foi feito baseado em um problema real da Empresa A, e a mesma disponibilizou os dados divididos em quatro grandes bases.

- Base de Informação dos Clientes (BIC): tabela preenchida com as informações de cadastro dos clientes, tais como: endereço, localização, bairro, cidade, número de ligação, quantidade de economias, categoria, se possui cisterna, tipo de medição, tipo de faturamento, etc.;
- Base de Leitura dos Clientes (BLC): tabela preenchida com todas as leituras de consumo dos clientes: volume consumido, volume faturado, status de leitura (variável binária que confirma se a leitura foi feita), código de leitura (descrição da leitura através de uma lista de códigos), etc.;
- Base de Inspeções Realizadas (BIR): tabela criada para registrar todas as inspeções feitas pela empresa nos clientes;
- Base de Fraude/Irregularidades Apontadas (BFA): tabela criada para registrar as fraudes ou irregularidades confirmadas a partir das inspeções realizadas.

Vale ressaltar que foram disponibilizadas quatro bases de dados (descritas acima) para cada uma das três regiões. Outro ponto importante é o fato de que apenas clientes inspecionados foram utilizados, dado que o problema será tratado com aprendizado supervisionado e, por isso, a necessidade da confirmação ou não de uma irregularidade/fraude nos clientes.

O próximo passo foi agrupar todas essas informações provenientes das bases originais em um novo *dataset*. Como pode ser visto na Figura 31, o conjunto de dados que contém todas as inspeções dos clientes (BIR) foi complementada com informações de consumo (BLC) e gerais dos clientes (BIC). Já a base BFA identifica os clientes que foram vistoriados e confirmados com irregularidade ou fraude. O resultado desta junção é a Base de Dados Gerais (BDG) com mais informações, em termos de atributos, porém com menor quantidade de registros, e com uma separação dos clientes: que foram detectados em irregularidade e os que a inspeção não retornou nenhuma identificação.

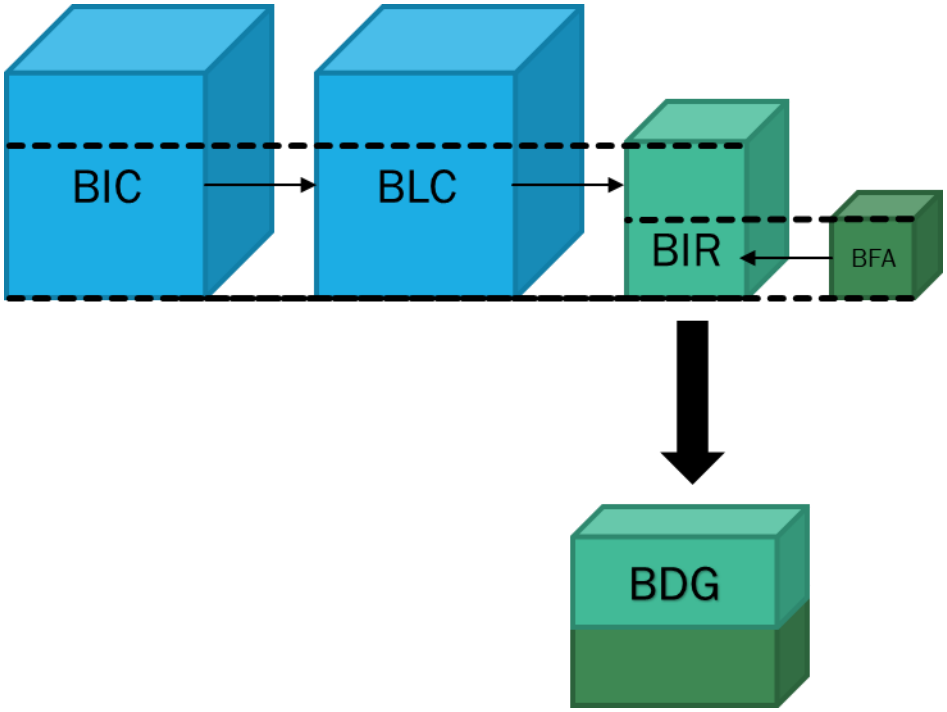


Figura 31 - Junção de informações das bases.

Essas bases são estruturadas de formas diferentes e cada uma possui um tamanho referente a esta estrutura, como pode ser visto na Tabela 10. A BIC é estruturada de forma que cada linha represente o cadastro de um consumidor. Já na BLC, cada linha representa uma leitura de consumo em um mês de cada cliente. Conclui-se que há clientes na BIC que não possuem leituras ou não aparecem na BLC. Já a BFA e a BIR são estruturadas de forma que cada linha representa uma irregularidade/fraude encontrada e uma inspeção realizada, respectivamente.

No *dataset* resultante, BDG, cada linha é a informação de leitura de consumo daquele mês específico em clientes que foram inspecionados e, somado a isso, informações de cadastro e se essa inspeção retornou uma irregularidade naquele mês.

Tabela 10 - Quantidade de Registros das Bases por Região.

Bases	BIC	BLC	BFA	BIR	BDG
Região G	45.925.883	21.168.933	147.165	3.003.479	14.245.907
Região P	14.493.005	9.729.442	58.571	2.145.482	7.270.988
Região T	18.963.816	15.146.083	47.061	2.048.722	6.583.795

5.2. Etapa I – Pré-Processamento da Base de Dados

Uma vez que a base de dados contém todas as informações necessárias e importantes provenientes das coletas de dados da Empresa, o próximo passo é pré-processar a base de forma que os modelos possam compreender os dados.

Os procedimentos descritos na seção **4.1.1 a** são executados nas três regiões.

Como mencionado anteriormente, essa etapa é responsável por desconsiderar todos os registros que não são de interesse, bem como as categorias. Além disso, a etapa realiza um tratamento nos valores de consumo dos clientes a fim de padronizar os dados incongruentes, faltantes e também que deveriam ser divididos pelo número de economias naquela devida ligação.

A outra parte da etapa visa criar atributos com os valores de consumo extraídos da base da empresa, combiná-los com atributos exógenos (clima, temperatura, umidade, etc) e, também atribuir classes de saída ou rótulos aos registros. Os atributos construídos e utilizados, pela Metodologia E, foram descritos no capítulo anterior e são mostrados na Tabela 6.

Cada linha desta nova base de dados, neste ponto, é composta pelo histórico de 12 meses de consumo de cada cliente que antecede uma inspeção, os atributos gerados por essa série histórica, os atributos exógenos relacionado a esse período de tempo e a classificação atribuída a cada um desses clientes. Se essa inspeção retornou uma irregularidade, o cliente é considerado Fraudador (*label* = 1). Se não retornou uma fraude, este cliente é considerado Não Fraudador (*label* = 0).

Após esta Etapa de Pré-Processamento (4.1.1), a Base Pré-Processada (BPP) está estruturada como mostrado na Tabela 11.

Sabe-se que a base processada possui apenas clientes inspecionados, e que essas inspeções são voltadas para suspeitas pautadas sobre o consumo dos clientes. Portanto, nota-se pela Tabela 11, que a razão Fraudador/Nº registros na BPP representa a taxa de acerto alcançada pela empresa durante as inspeções realizadas entre a data inicial e final dos dados coletados. Outro ponto importante a ser ressaltado, é que a Região T, possui uma taxa de acerto inferior quando comparada às outras

duas. Isso também é validado nos índices da Tabela 9, visto que se trata de uma região que perde bastante na distribuição de água.

As datas iniciais diferem para as Regiões G e P (junho/2016) e Região T (junho/2017).

Tabela 11 - Quantidade de registros das Bases Processadas por Região.

Bases	BDG	BPP	Fraudador	Data Inicial	Data Final
Região G	14.245.907	63.250	7.323	Jul/2017	Jul/2021
Região P	7.270.988	29.903	4.402	Jul/2017	Jul/2021
Região T	6.583.795	12.384	454	Jul/2018	Jul/2021

5.2.1. Definição dos Cenários e Treinamento

Antes de entrar na Fase de Treinamento, será feito uma definição e explicação de cada cenário de avaliação das metodologias que será utilizado para treinamento e análise.

A base de dados está disposta de Junho/2016 até julho/2021 para as Regiões G e P, e Junho/2017 até julho/2021 para a Região T. Deve ser destacado que, neste período, houve um evento pandêmico, o qual pode ter acarretado uma mudança de comportamento. A grande maioria da população passou a ficar em suas casas, fazendo com que o consumo fosse substancialmente maior a partir de tal período. Por isso, algumas análises foram feitas em cima disso. Primeiramente, o sistema foi aplicado em toda a base até um mês antes da pandemia, ou seja, de Junho/2016 (e 2017 para a Região T) até fevereiro/2020. Em seguida, foi analisado apenas acima dos dados de pandemia e depois adicionando os dados pré-pandêmicos de forma retroativa, sempre adicionando 3 meses antes do começo da COVID-19. Por fim, será analisado como o sistema se comporta em condições de funcionamento da prestadora de serviço, ou seja, o sistema é treinado até o mês n e colocado para uso nos meses $n+1$, $n+2$, até $n+m$, sendo m o mês previamente definido ou até que o modelo começasse a reduzir o desempenho esperado.

Portanto, as análises foram divididas nos seguintes cenários e, de forma geral, considera-se o mês de março/2020 o início da pandemia no Brasil:

i. Análise I:

- **a. Cenário A →** Todos os registros antes da pandemia da COVID-19; Dados de junho /2016 até fevereiro/2020. Janela de testes correspondente aos 6 últimos meses do período (setembro/2019 até fevereiro/2020);

ii. Análise II:

- **a. Cenário 1 →** Apenas registros após o começo da pandemia da COVID-19; Dados de março/2020 até julho/2021. Janela de testes restrita a apenas 1 mês (julho/2021) devido à quantidade de registros para treino;
- **b. Cenário 2 →** Registros correspondentes aos dados de três meses antes da pandemia + dados após o começo da pandemia. Dados de dezembro/2019 até julho/2021. Janela de testes de 1 mês para comparação com cenário anterior (julho/2021);
- **c. Cenário 3 →** Registros correspondentes aos dados de seis meses antes da pandemia + dados após o começo da pandemia. Dados de setembro/2019 até julho/2021. Janela de testes de 1 mês para comparação com cenário anterior (julho/2021);
- **d. Cenário 4 →** Registros correspondentes aos dados de nove meses antes da pandemia + dados após o começo da pandemia. Dados de junho/2019 até julho/2021. Janela de testes de 1 mês para comparação com cenário anterior (julho/2021);
- **e. Cenário 5 →** Registros correspondentes aos dados de 12 meses antes da pandemia + dados após o começo da pandemia. Dados de março/2019 até julho/2021. Janela de testes de 1 mês para comparação com cenário anterior (julho/2021);
- **f. Cenário 6 →** Registros correspondentes a todos os meses antes da pandemia + dados após o começo da pandemia. Dados de junho/2016 até julho/2021. Janela de testes de 1 mês para comparação com cenário anterior (julho/2021).

iii. Análise III:

a. **Cenário B** → Base inteira; Dados de junho/2016 até julho/2021. Janela de testes de 6 meses para testar performance do modelo em comparação indireta com Cenário A da Análise I (fevereiro/2021 até julho/2021);
b. **Cenário B-1** → Dados de junho/2016 até janeiro/2021 (mês n); Janela de testes subsequentes de 1 mês para simular dinâmica funcional da Empresa; Teste em fevereiro/2021 (mês $n + 1$);
c. **Cenário B-2** → Dados de junho/2016 até janeiro/2021 (mês n); Janela de testes subsequentes de 1 mês para simular dinâmica funcional da Empresa; Teste em março/2021 (mês $n + 2$);
d. **Cenário B-3** → Dados de junho/2016 até janeiro/2021 (mês n); Janela de testes subsequentes de 1 mês para simular dinâmica funcional da Empresa; Teste em abril/2021 (mês $n + 3$);
e. **Cenário B-4** → Dados de junho/2016 até janeiro/2021 (mês n); Janela de testes subsequentes de 1 mês para simular dinâmica funcional da Empresa; Teste em maio/2021 (mês $n + 4$);
f. **Cenário B-5** → Dados de junho/2016 até janeiro/2021 (mês n); Janela de testes subsequentes de 1 mês para simular dinâmica funcional da Empresa; Teste em junho/2021 (mês $n + 5$);

A Tabela 12, a seguir, resume os períodos de dados utilizados nos conjuntos de treino e teste para os diferentes cenários citados acima, bem como a quantidade de fraudadores em cada um deles.

A Análise I tem como intuito observar o desempenho do modelo das metodologias nos dados em que se há certeza de que não sofreram interferência da pandemia. Já a Análise II é feita, inicialmente, para os dados de pandemia apenas, e depois foi se adicionando, retroativamente, meses pré-pandemia para avaliar mudanças drásticas de desempenho do sistema. Por fim a Análise III é feita sob a perspectiva de funcionamento da empresa, isto é, o sistema é treinado até uma certa data, e passa a ser utilizado, em seguida, em todos os meses subsequentes.

Tabela 12 - Conjuntos de Treino e Teste nos Diferentes Cenários.

Cenários	Regiões	Treino	Teste

		Período	Fraude	Total	Período	Fraude	Total
CA	G	Jul/17 até Ago/19	5.050	39.472	Set/19 até Fev/20	1.092	10.629
	P		2.515	22.880		929	4.842
	T		95	4.124		67	1.822
C1	G	Abr/21 até Jun/21	480	2.404	Jul/21	173	1363
	P		211	322		87	130
	T		149	1.553		18	536
C2	G	Jan/21 até Jun/21	777	3.869	Jul/21	173	1363
	P		395	625		87	130
	T		180	2.701		18	536
C3	G	Out/20 até Jun/21	956	7.375	Jul/21	173	1363
	P		504	845		87	130
	T		217	3936		18	536
C4	G	Jul/20 até Jun/21	1.008	8.803	Jul/21	173	1363
	P		605	1.280		87	130
	T		264	4.956		18	536
C5	G	Abr/20 até Jun/21	1.008	10.791	Jul/21	173	1363
	P		755	1.641		87	130
	T		267	5.285		18	536
C6	G	Jul/17 até Jun/21	7.150	61.887	Jul/21	173	1363
	P		4.315	29.773		87	130
	T		436	11.848		18	536
CB	G	Jul/17 até Jan/21	6.495	57.907	Fev/21 até Jul/21	828	5.343
	P		3.971	29.232		431	671
	T		260	9.449		194	2.935
CB-1	G	Jul/17 até Jan/21	6.495	57.907	Fev/21	107	723
	P		3.971	29.232		74	109
	T		260	9.449		7	480
CB-2	G	Jul/17 até Jan/21	6.495	57.907	Mar/21	68	853
	P		3.971	29.232		59	110
	T		260	9.449		20	366
CB-3	G		6.495	57.907	Abr/21	172	710

	P	Jul/17 até Jan/21	3.971	29.232		62	103
	T		260	9.449		53	529
CB-4	G	Jul/17 até Jan/21	6.495	57.907	Mai/21	173	719
	P		3.971	29.232		77	108
	T		260	9.449		69	572
CB-5	G	Jul/17 até Jan/21	6.495	57.907	Jun/21	135	975
	P		3.971	29.232		72	111
	T		260	9.449		27	452

5.3. Etapa II.E – Metodologia E – Fase de Treinamento

Após a definição dos cenários, se dá início a Etapa II.E da Metodologia E – Fase de Treinamento.

Nesta etapa é criado todo processo evolutivo conforme descrito na seção 4.2.2, e se dá início à otimização dos hiperparâmetros e seleção de variáveis, utilizando os parâmetros da Tabela 7.

Vale ressaltar que nessa etapa é feita uma validação cruzada com 20% dos dados de treino escolhidos aleatoriamente, e o restante passa por uma subamostragem da classe minoritária para não haver viés do resultado para a classe majoritária. Essa execução é feita 10 vezes e o modelo que retornou o melhor resultado para validação é aplicado nos dados de teste.

Para cada Cenário estipulado existem três regiões, para cada região há cinco modelos e para cada modelo há um melhor conjunto de hiperparâmetros e variáveis encontrado por meio do processo evolutivo. Além disso, ainda se tem os resultados dos cinco modelos combinados pelos diferentes tipos de comitês (voto majoritário – VM –, soma ponderada – SP – e fusão de probabilidades – FProb). Isso nos levaria a, no caso da Análise II e III, oito resultados vezes x três regiões x quantidade de cenários igual a 288 resultados. Sendo assim, será mostrado neste capítulo apenas os parâmetros e atributos selecionados pelo melhor modelo de cada cenário, baseado na taxa de acerto alcançada, escolhidos a partir da performance obtida pelo

conjunto de validação. Porém, a precisão alcançada, no conjunto de validação, por todos os algoritmos em cada cenário será evidenciada no APÊNDICE A.

A seguir são mostrados os conjuntos de parâmetros e de atributos selecionados pelos melhores modelos em cada um dos cenários e dividido por análises. Vale ressaltar que, se o algoritmo obteve o melhor desempenho na validação, individualmente, será apresentado apenas ele. Enquanto que, se esse desempenho é oriundo do comitê, será mostrado ainda o melhor algoritmo, porém com a indicação de um (*) acompanhado do melhor comitê entre parêntesis:

i. Análise I: A seguir, na Tabela 13 e na Tabela 14, estão destacados os conjuntos de parâmetros e atributos, respectivamente, selecionados pela otimização nos melhores modelos do Cenário A.

Tabela 13 - Tabela de Parâmetros Selecionados pelos Melhores Modelos da Análise I.

	Melhor Modelo		
	G	P	T
	DT	RF	MLP* (VM)
CA	*MNSSIN: 40* *MNSPL: 12* *CPC: 0.0007* *SP: gini* *SS: best*	*NT: 49* *MNSSIN: 20* *MNSPL: 9* *CPC: 0.0007* *SP: entropy* *SS: best*	*HLS: 18* *ALPHA: 0.0209* *BS: 489* *ILR: 0.2996* *M: 0.3022* *HLAF: tanh* *OPT.: adam* *LR: constant*
Legenda	*MNSSIN: mínimo número de amostras por nó interno; MNSPL: número de amostras mínimas por folha; CPC: custo de redução de complexidade; SP: critério de separação; SS: estratégia de separação; NT: número de árvores; NBR: número de rodadas; LR: taxa de aprendizado; SMLR: redução mínima de perda pela separação; MTD: profundidade máxima da árvores; HLS: neurônios na camada escondida; BS: tamanho do batch; ILR: taxa de aprendizado inicial; M: momentum; HLAF: função de ativação da camada escondida; Opti: otimizador. *Maior acertividade alcançada pelo comitê.*		

Tabela 14 - Tabela de Atributos Selecionados pelos Melhores Modelos na Análise I.

Atributos	Cenários		
	CA		
Região	G	P	T
Precipitação		x	x
Umidade	x	x	x
Temperatura Média	x	x	x
Temperatura Máxima		x	
Temperatura Mínima		x	
Qtd. Feriados	x	x	x
Janela dos últimos 6 meses - Mês 1		x	
Janela dos últimos 6 meses - Mês 2		x	
Janela dos últimos 6 meses - Mês 3			
Janela dos últimos 6 meses - Mês 4	x		x
Janela dos últimos 6 meses - Mês 5			x
Janela dos últimos 6 meses - Mês 6	x	x	x
Média dos últimos 3 meses		x	x
Média dos últimos 6 meses			
Média dos últimos 12 meses			x
Diferença entre mês 1 e 2 da janela de 3 meses	x	x	x
Diferença entre mês 2 e 3 da janela de 3 meses	x	x	
Diferença entre mês 1 e 3 da janela de 3 meses		x	
Meses com consumo abaixo de 5m³	x	x	x
Meses com consumo abaixo de 75% da média anual	x	x	
Meses com consumo abaixo de 50% da média anual	x	x	
Maior aumento de consumo mensal	x		x
Maior queda de consumo mensal	x	x	
Diferença entre o maior aumento e a maior queda	x		x
Meses com consumo constante		x	
Diferença entre o consumo mensal com o do ano anterior	x	x	x
Inclinação da linha de regressão que descreve a série	x	x	
Sazonalidade - Verão		x	
Sazonalidade - Outono		x	x
Sazonalidade - Inverno	x		x
Sazonalidade - Primavera	x	x	x

ii. Análise II: A seguir, na Tabela 15 e na Tabela 16, estão destacados os conjuntos de parâmetros e atributos, respectivamente, selecionados pela otimização nos melhores modelos do Cenário 1 (C1), Cenário 2 (C2), Cenário 3 (C3), Cenário 4 (C4), Cenário 5 (C5) e Cenário 6 (C6).

Tabela 15 - Tabela de Parâmetros Selecionados pelos Melhores Modelos na Análise II.

	Melhor Modelo				Melhor Modelo		
	G	**P**	**T**		**G**	**P**	**T**
	DT	**RF**	**RF**		**XGB* (VM)**	**RF* (VM)**	**RF* (VM)**
C1	*MNSSIN: 32 MNSPL: 49 CPC: 0.0098 SP: gini SS: random*	*NT: 22 MNSSIN: 38 MNSPL: 2 CPC: 0.0184 SP: entropy SS: random*	*NT: 34 MNSSIN: 32 MNSPL: 15 CPC: 0.0114 SP: gini SS: best*	**C4**	*NBR: 42 LR: 0.2168 SMLR: 0.1901 MTD: 2*	*NT: 41 MNSSIN: 9 MNSPL: 18 CPC: 0.0086 SP: entropy SS: best*	*NT: 30 MNSSIN: 39 MNSPL: 26 CPC: 0.006 SP: gini SS: random*
	RF	**RF* (FProb)**	**XGB**		**XGB**	**RF**	**DT**
C2	*NT: 29 MNSSIN: 14 MNSPL: 26 CPC: 0.0186 SP: gini SS: random*	*NT: 30 MNSSIN: 19 MNSPL: 23 CPC: 0.0071 SP: entropy SS: best*	*NBR: 22 LR: 0.01 SMLR: 0.1537 MTD: 2*	**C5**	*NBR: 31 LR: 0.1824 SMLR: 0.1788 MTD: 2*	*NT: 25 MNSSIN: 28 MNSPL: 29 CPC: 0.012 SP: gini SS: random*	*MNSSIN: 20 MNSPL: 44 CPC: 0.0151 SP: gini SS: best*
	XGB	**XGB**	**DT* (FProb)**		**XGB* (VM)**	**XGB* (FProb)**	**RF* (SP)**
C3	*NBR: 20 LR: 0.2751 SMLR: 0.3338 MTD: 2*	*NBR: 32 LR: 0.2017 SMLR: 0.2032 MTD: 2*	*MNSSIN: 10 MNSPL: 42 CPC: 0.0174 SP: gini SS: best*	**C6**	*NBR: 36 LR: 0.1284 SMLR: 0.1487 MTD: 6*	*NBR: 4 LR: 0.0582 SMLR: 0.0482 MTD: 4*	*NT: 39 MNSSIN: 20 MNSPL: 4 CPC: 0.003 SP: entropy SS: random*
Legenda	*MNSSIN: mínimo número de amostras por nó interno; MNSPL: número de amostras mínimas por folha; CPC: custo de redução de complexidade; SP: critério de separação; SS: estratégia de separação; NT: número de árvores; NBR: número de rodadas; LR: taxa de aprendizado; SMLR: redução mínima de perda pela separação; MTD: profundidade máxima da árvores; HLS: neurônios na camada escondida; BS: tamanho do batch; ILR: taxa de aprendizado inicial; M: momentum; HLAF: função de ativação da camada escondida; Opti: otimizador. *Maior acertividade alcançada pelo comitê, porém melhor modelo dentre os outros.*						

Tabela 16 - Tabela de Atributos Selecionados pelos Melhores Modelos na Análise II.

Atributos	Cenários																	
	C1			C2			C3			C4			C5			C6		
Região	G	P	T	G	P	T	G	P	T	G	P	T	G	P	T	G	P	T
Precipitação	x			x			x	x		x	x			x	x	x		x
Umidade				x	x	x		x	x			x	x	x			x	x
Temperatura Média	x		x	x	x				x		x	x		x	x			
Temperatura Máxima	x			x	x	x	x		x	x	x		x			x		
Temperatura Mínima			x		x	x	x	x		x				x	x			
Qtd. Feriados			x	x			x	x		x			x					x
Janela dos últimos 6 meses - Mês 1			x	x	x	x	x	x	x						x	x	x	x
Janela dos últimos 6 meses - Mês 2	x			x		x	x		x			x		x				
Janela dos últimos 6 meses - Mês 3	x		x		x	x	x		x			x	x				x	
Janela dos últimos 6 meses - Mês 4		x		x	x	x	x	x	x						x	x	x	
Janela dos últimos 6 meses - Mês 5	x						x						x		x	x	x	
Janela dos últimos 6 meses - Mês 6			x	x		x	x		x			x	x	x				
Média dos últimos 3 meses				x	x					x	x	x						
Média dos últimos 6 meses			x	x		x			x		x	x		x		x		
Média dos últimos 12 meses	x	x		x	x		x	x		x			x	x		x	x	
Diferença entre mês 1 e 2 da janela de 3 meses			x		x	x	x		x	x			x	x	x		x	x
Diferença entre mês 2 e 3 da janela de 3 meses	x	x					x					x	x	x	x			
Diferença entre mês 1 e 3 da janela de 3 meses	x	x				x				x	x		x	x				x
Meses com consumo abaixo de 5m³	x	x	x	x	x		x	x	x	x			x	x	x	x		x
Meses com consumo abaixo de 75% da média anual				x	x	x	x	x	x	x	x	x	x		x	x		
Meses com consumo abaixo de 50% da média anual		x				x	x	x	x	x	x	x						
Maior aumento de consumo mensal	x	x			x			x	x		x	x	x	x				x
Maior queda de consumo mensal		x	x	x		x	x	x	x		x	x		x	x	x	x	x
Diferença entre o maior aumento e a maior queda		x			x	x		x		x	x			x	x		x	
Meses com consumo constante		x			x				x	x			x	x	x	x		x
Diferença entre o consumo mensal com o do ano anterior	x			x		x	x			x	x	x	x				x	x
Inclinação da linha de regressão que descreve a série	x			x			x	x	x					x		x		
Sazonalidade - Verão	x		x		x	x	x	x		x		x	x	x	x	x	x	x
Sazonalidade - Outono	x	x		x		x	x	x	x	x		x	x				x	
Sazonalidade - Inverno	x		x	x			x	x		x			x				x	
Sazonalidade - Primavera	x	x		x		x	x	x							x	x		x

A partir da Tabela 16, pode-se concluir que alguns atributos são menos importantes do que outros. Por exemplo, a quantidade de meses abaixo de 5m³ e a maior queda de consumo mensal foram selecionados 14 vezes em 18. Por outro lado, a média dos últimos três meses é um atributo pouco escolhido pelos modelos assim como a quantidade de feriados nessa análise.

Análise III: A seguir, na

iii. Tabela 17 e na Tabela 18, estão destacados os conjuntos de parâmetros e atributos, respectivamente, selecionados pela otimização nos melhores modelos do Cenário B (CB), Cenário B-1 (CB-1), Cenário B-2 (CB-2), Cenário B-3 (CB-3), Cenário B-4 (CB-4) e Cenário B-5 (CB-5). Vale ressaltar que essa análise é feita para a dinâmica funcional da empresa, ou seja, o modelo seria treinado até uma data (jan/21 neste exemplo) e testa a cada mês subsequente, sem alterar o treinamento do modelo. Por isso, se modelos iguais forem escolhidos para a mesma região terão os mesmos conjuntos de atributos e parâmetros. Depois de um tempo, caso a precisão entre em declínio, seria sugerido um novo treinamento do sistema. Esse declínio é evidência de que os perfis presentes estão se distanciando dos que foram usados para os ajustes dos algoritmos.

Tabela 17 - Tabela de Parâmetros Selecionados pelos Melhores Modelos na Análise III.

	Melhor Modelo				Melhor Modelo		
	G	P	T		G	P	T
	RF	MLP	DT		DT	MLP* (SP)	DT
CB	NT: 30 MNSSIN: 32 MNSPL: 18 CPC: 0 SP: gini SS: random	HLS: 15 ALPHA: 0.0418 BS: 690 ILR: 0.1258 M: 0.7282 HLAF: tanh OPT.: adam LR: adaptive	MNSSIN: 50 MNSPL: 45 CPC: 0.0093 SP: entropy SS: best	CB-3	MNSSIN: 11 MNSPL: 40 CPC: 0.0128 SP: entropy SS: best	HLS: 15 ALPHA: 0.0418 BS: 690 ILR: 0.1258 M: 0.7282 HLAF: tanh OPT.: adam LR: adaptive	MNSSIN: 50 MNSPL: 45 CPC: 0.0093 SP: entropy SS: best
	DT	XGB* (SP)	SVM		RF	RF	DT
CB-1	MNSSIN: 11 MNSPL: 40 CPC: 0.0128 SP: entropy SS: best	NBR: 16 LR: 0.0744 SMLR: 0.2502 MTD: 4	C: 8.4317 gamma: 0.0033 kernel: rbf	CB-4	NT: 30 MNSSIN: 32 MNSPL: 18 CPC: 0 SP: gini SS: random	NT: 36 MNSSIN: 47 MNSPL: 13 CPC: 0.0035 SP: gini SS: best	MNSSIN: 50 MNSPL: 45 CPC: 0.0093 SP: entropy SS: best
	DT	XGB* (VM)	DT		DT	DT	RF
CB-2	MNSSIN: 11 MNSPL: 40 CPC: 0.0128 SP: entropy SS: best	NBR: 16 LR: 0.0744 SMLR: 0.2502 MTD: 4	MNSSIN: 50 MNSPL: 45 CPC: 0.0093 SP: entropy SS: best	CB-5	MNSSIN: 11 MNSPL: 40 CPC: 0.0128 SP: entropy SS: best	MNSSIN: 21 MNSPL: 37 CPC: 0.0125 SP: gini SS: best	NT: 28 MNSSIN: 45 MNSPL: 3 CPC: 0.0091 SP: entropy SS: random
Legenda	MNSSIN: mínimo número de amostras por nó interno; MNSPL: número de amostras mínimas por folha; CPC: custo de redução de complexidade; SP: critério de separação; SS: estratégia de separação; NT: número de árvores; NBR: número de rodadas; LR: taxa de aprendizado; SMLR: redução mínima de perda pela separação; MTD: profundidade máxima da árvores; HLS: neurônios na camada escondida; BS: tamanho do batch; ILR: taxa de aprendizado inicial; M: momentum; HLAF: função de ativação da camada escondida; Opti: otimizador. *Maior acertividade alcançada pelo comitê, porém melhor modelo dentre os outros.						

Tabela 18 - Tabela de Atributos Selecionados pelos Melhores Modelos na Análise III.

Atributos	Cenários																	
	CB			CB-1			CB-2			CB-3			CB-4			CB-5		
Região	G	P	T	G	P	T	G	P	T	G	P	T	G	P	T	G	P	T
Precipitação			x						x			x			x			x
Umidade	x	x			x			x			x		x	x			x	x
Temperatura Média				x		x	x			x						x		
Temperatura Máxima				x			x			x				x		x		
Temperatura Mínima	x				x	x		x					x				x	x
Qtd. Feriados	x		x			x			x			x	x		x			x
Janela dos últimos 6 meses - Mês 1	x				x			x					x					x
Janela dos últimos 6 meses - Mês 2			x		x			x	x			x			x			
Janela dos últimos 6 meses - Mês 3	x	x				x					x		x	x			x	
Janela dos últimos 6 meses - Mês 4		x	x	x		x	x		x	x	x	x			x	x		
Janela dos últimos 6 meses - Mês 5	x			x	x	x	x	x		x			x	x		x		
Janela dos últimos 6 meses - Mês 6	x	x	x		x			x	x		x	x	x		x			
Média dos últimos 3 meses		x	x						x		x	x		x	x		x	x
Média dos últimos 6 meses	x	x									x		x	x			x	
Média dos últimos 12 meses	x	x		x	x	x	x	x		x	x		x			x	x	x
Diferença entre mês 1 e 2 da janela de 3 meses	x			x		x	x			x			x	x		x	x	
Diferença entre mês 2 e 3 da janela de 3 meses	x		x	x			x		x	x		x	x	x	x	x		
Diferença entre mês 1 e 3 da janela de 3 meses	x			x	x		x	x		x			x			x		x
Meses com consumo abaixo de 5m³	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	
Meses com consumo abaixo de 75% da média anual	x	x	x	x		x	x		x	x	x	x	x		x	x		x
Meses com consumo abaixo de 50% da média anual	x		x	x			x		x	x		x	x		x	x	x	
Maior aumento de consumo mensal		x		x	x	x	x	x		x	x			x		x		
Maior queda de consumo mensal	x		x	x	x		x	x	x	x		x	x	x	x	x	x	x
Diferença entre o maior aumento e a maior queda	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x
Meses com consumo constante		x	x	x	x		x	x	x	x	x	x		x	x	x		
Diferença entre o consumo mensal com o do ano anterior	x	x									x		x					
Inclinação da linha de regressão que descreve a série			x			x			x			x		x	x			x
Sazonalidade - Verão	x	x	x	x		x	x		x	x	x	x	x	x	x	x	x	x
Sazonalidade - Outono						x								x			x	x
Sazonalidade - Inverno			x						x			x			x		x	x
Sazonalidade - Primavera	x	x	x		x	x		x	x		x	x	x	x	x			

Na Tabela 18, mais uma vez, alguns atributos se repetem quando são escolhidos por quase todos os algoritmos através dos cenários. Quantidade de meses com consumo abaixo de 5m³, quantidade de meses com consumo abaixo de 75% da média anual, maior queda de consumo mensal e diferente entre maior aumento e maior queda, são alguns dos exemplos que foram escolhidos em quase todos os modelos de todas as regiões. Por outro lado, indicações de temperatura (mínima e média), de sazonalidade (outono e inverno) quase não foram selecionados.

Uma vez que todos os cenários estejam com seus conjuntos definidos, é feito o treinamento dos modelos com esses grupos de parâmetros e atributos.

Sobre as performances dos modelos escolhidos pelo conjunto de validação, pode-se observar que a maioria dos algoritmos "vencedores" são baseados em árvores de decisão. Isto pode nos dizer que um algoritmo baseado em separação de dados, como o SVM que separa os dados por uma função kernel, não tenha tido tanto sucesso, pois os dados talvez fossem muito "próximos" nesse espaço.

Em relação aos comitês, através da Tabela 15, observa-se que a utilização de comitês para a classificação foi bastante benéfica para o sistema, principalmente nos Cenários 4 e 6 com uma leve vantagem para o Comitê por Voto Majoritário.

5.4. Etapa II.F – Metodologia F – Fase de Treinamento

Esta metodologia não permite a análise de todos os cenários por se tratar de um método restritivo e, portanto, será aplicado a apenas alguns cenários, são eles: Cenário A da Análise I, Cenário 6 da Análise II e Cenário B da Análise III.

A Tabela 19 apresenta os conjuntos de treino e teste derivados da filtragem utilizada na Metodologia F para os Cenários apontados anteriormente.

Tabela 19 – Conjuntos de Treino e Teste por Cenário da Filtragem da Metodologia F.

Cenários	Regiões	Treino			Teste		
		Período	Fraude	Total	Período	Fraude	Total
CA	G	Jul/17 até Ago/19	2.023	13.371	Set/19 até Fev/20	417	3.688
	P		883	5.102		327	876
	T		48	825		33	364
C6	G	Jul/17 até Jun/21	6.215	47.695	Jul/21	138	855
	P		1.911	8.526		30	49
	T		212	2.743		9	117
CB	G	Jul/17 até Jan/21	3.868	29.703	Fev/21 até Jul/21	467	2.522
	P		2.152	10.612		199	326
	T		125	1.763		75	680

A partir dessa tabela e comparando com a Tabela 12, conclui-se que essa abordagem diminui consideravelmente o número de registros dos conjuntos. Isso tende a ter um melhor desempenho quando a região possui dados em abundância.

Os atributos utilizados na Metodologia F, foram descritos na seção 4.1.2. Já a busca de parâmetros, diferentemente da Metodologia E, foi utilizando o *GridSearch* e os conjuntos selecionados estão descritos na Tabela 20.

Tabela 20 - Tabela de Parrâmetros Selecionados na Metodologia F.

	Melhor Modelo		
	G	**P**	**T**
	RF	**SVM* (VM)**	**SVM**
CA	*MD: 3* *MF: 2* *NE: 300*	*C: 10* *kernel: rbf*	*C: 2* *kernel: rbf*
	SVM	**XGB**	**MLP* (VM)**
C6	*C: 10* *kernel: rbf*	*MD: 3* *gamma: 1* *MCW: 5*	*HLS: 100* *ILR: 0.005* *HLAF: relu*
	SVM	**RF**	**RF**
CB	*C: 20* *kernel: rbf*	*MD: 5* *MF: 4* *NE: 300*	*MD: 4* *MF: 4* *NE: 300*
Legenda	*MD: profundidade máxima; MF: características máximas; NE: número de estimadores; MCW: peso mínimo da folha; HLS: neurônios na camada escondida; ILR: taxa de aprendizado inicial; HLAF: função de ativação da camada escondida.* **Maior acertividade alcançada pelo comitê.*		

Logo em seguida, foi feita uma separação em 20% dos dados de treino para validação simples, e os 80% restantes dos dados de treino passaram pelas técnicas de subamostragem mencionadas na seção 4.1.2 em relação à classe minoritária para, mais uma vez, o treinamento dos modelos não serem afetados pelo desbalanceamento presente na base de dados.

5.5. Etapa III – Fase de Classificação

Após os modelos serem treinados com os conjuntos de parâmetros e validados, o próximo passo é utilizar esses modelos em dados ainda não vistos (conjunto

de teste) para poder validar a capacidade de generalização do sistema. A seguir são destacadas as matrizes confusão de todos os cenários, da Tabela 21 até a Tabela 26.

i. Tabela 22 apresentam os resultados do sistema inteligente para o Cenário A nas Metodologias F e E, respectivamente.

Tabela 21 - Sistema Inteligente x Clientes Inspecionados – Cenário A da Metodologia E.

				Clientes Inspecionados Pela Empresa A	
				Não Fraudador (0)	*Fraudador (1)*
Melhor Modelo	**CA**	**G**	*N. Fraud. (0)*	6114	555
			Fraud. (1)	612	126
		P	*N. Fraud. (0)*	2522	410
			Fraud. (1)	654	252
		T	*N. Fraud. (0)*	1011	25
			Fraud. (1)	85	9

Tabela 22 - Sistema Inteligente x Clientes Inspecionados – Cenário A da Metodologia F.

				Clientes Inspecionados Pela Empresa A	
				Não Fraudador (0)	*Fraudador (1)*
Melhor Modelo	**CA**	**G**	*N. Fraud. (0)*	2365	199
			Fraud. (1)	906	218
		P	*N. Fraud. (0)*	312	140
			Fraud. (1)	237	187
		T	*N. Fraud. (0)*	241	16
			Fraud. (1)	90	17

Tabela 23 - Sistema Inteligente x Clientes Inspecionados – Cenário 1 ao 6 da Metodologia E.

				Clientes Inspecionados Pela Empresa A	
				Não Fraudador (0)	*Fraudador (1)*
Melhor Modelo	**C1**	**G**	N. Fraud. (0)	1010	130
			Fraud. (1)	180	43
		P	*N. Fraud. (0)*	19	24
			Fraud. (1)	29	58
		T	*N. Fraud. (0)*	413	5
			Fraud. (1)	105	13
	C2	**G**	*N. Fraud. (0)*	930	115
			Fraud. (1)	260	58
		P	*N. Fraud. (0)*	30	49
			Fraud. (1)	13	38
		T	*N. Fraud. (0)*	454	10
			Fraud. (1)	64	8
	C3	**G**	*N. Fraud. (0)*	925	107
			Fraud. (1)	265	66
		P	*N. Fraud. (0)*	32	44
			Fraud. (1)	11	43
		T	*N. Fraud. (0)*	473	12
			Fraud. (1)	45	6
	C4	**G**	*N. Fraud. (0)*	1011	130
			Fraud. (1)	179	43
		P	*N. Fraud. (0)*	21	22
			Fraud. (1)	22	65
		T	*N. Fraud. (0)*	418	5
			Fraud. (1)	100	13
	C5	**G**	*N. Fraud. (0)*	1015	129
			Fraud. (1)	175	44
		P	*N. Fraud. (0)*	20	18
			Fraud. (1)	23	69
		T	*N. Fraud. (0)*	454	11
			Fraud. (1)	64	7
	C6	**G**	*N. Fraud. (0)*	1100	145
			Fraud. (1)	90	28
		P	*N. Fraud. (0)*	28	33
			Fraud. (1)	15	54
		T	*N. Fraud. (0)*	439	7
			Fraud. (1)	79	11

Tabela 24 - Sistema Inteligente x Clientes Inspecionados – Cenário 6 da Metodologia F.

				Clientes Inspecionados Pela Empresa A	
				Não Fraudador (0)	*Fraudador (1)*
Melhor Modelo	**C6**	**G**	*N. Fraud. (0)*	433	64
			Fraud. (1)	284	74
		P	*N. Fraud. (0)*	3	3
			Fraud. (1)	16	27
		T	*N. Fraud. (0)*	58	2
			Fraud. (1)	50	7

i. **Análise II**: Para a análise II, a Metodologia E consiste na variação dos Cenários 1 até 6 (Tabela 23), enquanto que a Metodologia F foi mostrada apenas para o Cenário 6 (Tabela 24). Pois, como mencionado anteriormente, a segunda abordagem consiste em uma introdutória à primeira.

ii. **Análise III**: Na terceira análise o intuito é testar a dinâmica que seria utilizada em uma distribuidora, ou seja, o sistema seria treinado até determinado mês e seria continuamente usado, repetidamente, por meses subsequentes até ser necessário um novo treinamento. A Tabela 25 mostra o sistema atuando com a Metodologia E para todos os cenários mencionados anteriormente, enquanto Tabela 26 representa os resultados obtidos pela Metodologia F no Cenário B.

Tabela 25 - Sistema Inteligente x Clientes Inspecionados – Cenário B ao B-5 da Metodologia E.

				Clientes Inspecionados Pela Empresa A	
				Não Fraudador (0)	*Fraudador (1)*
Melhor Modelo	**CB**	**G**	*N. Fraud. (0)*	3494	488
			Fraud. (1)	1021	340
		P	*N. Fraud. (0)*	59	49
			Fraud. (1)	181	382
		T	*N. Fraud. (0)*	2414	88
			Fraud. (1)	327	106
	CB-1	**G**	*N. Fraud. (0)*	431	59
			Fraud. (1)	185	48
		P	*N. Fraud. (0)*	5	4
			Fraud. (1)	30	70
		T	*N. Fraud. (0)*	454	4
			Fraud. (1)	19	3
	CB-2	**G**	*N. Fraud. (0)*	600	46
			Fraud. (1)	185	22
		P	*N. Fraud. (0)*	8	6
			Fraud. (1)	43	53
		T	*N. Fraud. (0)*	330	11
			Fraud. (1)	16	9
	CB-3	**G**	*N. Fraud. (0)*	394	85
			Fraud. (1)	144	87
		P	*N. Fraud. (0)*	7	9
			Fraud. (1)	34	53
		T	*N. Fraud. (0)*	397	14
			Fraud. (1)	79	39
	CB-4	**G**	*N. Fraud. (0)*	426	111
			Fraud. (1)	120	62
		P	*N. Fraud. (0)*	8	10
			Fraud. (1)	23	67
		T	*N. Fraud. (0)*	394	28
			Fraud. (1)	109	41
	CB-5	**G**	*N. Fraud. (0)*	678	78
			Fraud. (1)	162	57
		P	*N. Fraud. (0)*	22	25
			Fraud. (1)	17	47
		T	*N. Fraud. (0)*	368	57
			Fraud. (1)	10	17

Tabela 26 - Sistema Inteligente x Clientes Inspecionados – Cenário B da Metodologia F.

				Clientes Inspecionados Pela Empresa A	
				Não Fraudador (0)	*Fraudador (1)*
Melhor Modelo	**CB**	**G**	*N. Fraud. (0)*	1207	171
			Fraud. (1)	848	296
		P	*N. Fraud. (0)*	71	56
			Fraud. (1)	56	143
		T	*N. Fraud. (0)*	311	21
			Fraud. (1)	294	54

5.6. Interpretação dos Resultados e Indicação de Possíveis Fraudes

Como dito anteriormente na seção 4.3, a métrica de taxa de acerto utilizada para medir o desempenho do sistema será a precisão, indicada pela Equação (21). Enquanto que a taxa de acerto da empresa – Equação (22) – será utilizada para comparar com a alcançada pelo sistema.

5.6.1. Metodologia Evolutiva

A seguir, da Tabela 27 até a Tabela 29, são apresentados os resultados obtidos para cada análise utilizando a Metodologia E.

i. **Análise I:**

Tabela 27 - Taxa de acerto Análise I – Metodologia E.

	CA		Empresa A	
	Acert.	Ind./ Tot.	Acert. Empresa	Ind./ Tot.
G	**17,07%**	126 / 738	9,19%	681 / 7407
P	**27,81%**	252 / 906	17,20%	662 / 3848
T	**9,57%**	9 / 94	3,01%	34 / 1130
Legenda	*Acert: Acertividade; Ind.: Indicações; Tot.: Total.*			

ii. Análise II:

Tabela 28 - Taxa de acerto Análise II – Metodologia E.

	C1		C2		C3		C4		C5		C6		Empresa A	
	Acert.	Ind./ Tot.	Acert.	Ind./ Tot.	Acert.	Ind./ Tot.	Acert.	Ind./ Tot.	Acert.	Ind./ Tot.	Acert.	Ind./ Tot.	Acert. Empresa	Ind./ Tot.
G	19,28%	43	18,24%	58	19,94%	66	19,37%	43	20,09%	44	**23,73%**	28	12,69%	173
		223		318		331		222		219		118		1363
P	70,73%	58	74,51%	38	**79,63%**	43	74,71%	65	75,00%	69	78,26%	54	66,92%	87
		82		51		54		87		92		69		130
T	7,51%	13	11,11%	8	11,76%	6	11,50%	13	9,86%	7	**12,22%**	11	3,36%	18
		173		72		51		113		71		90		536
Legenda	*Acert: Acertividade; Ind.: Indicações; Tot.: Total.*													

iii. Análise III:

Tabela 29 - Taxa de acerto do Sistema Análise III - Metodologia E.

	G		Empresa A		P		Empresa A		T		Empresa A	
	Acert.	Ind./ Tot.	Acert. Empresa	Ind./ Tot.	Acert.	Ind./ Tot.	Acert. Empresa	Ind./ Tot.	Acert.	Ind./ Tot.	Acert. Empresa	Ind./ Tot.
CB-1	**21,13%**	30	14,80%	107	**70,00%**	70	67,89%	74	**13,64%**	3	1,46%	7
		142		723		100		109		22		480
CB-2	**10,63%**	22	7,97%	68	**55,21%**	53	53,64%	59	**36,00%**	9	5,46%	20
		207		853		96		110		25		366
CB-3	**37,66%**	87	24,23%	172	**61,62%**	61	60,19%	62	**33,05%**	39	10,02%	53
		231		710		99		103		118		529
CB-4	**34,07%**	62	24,06%	173	**74,44%**	67	71,30%	77	**27,33%**	41	12,06%	69
		182		719		90		108		150		572
CB-5	**26,03%**	57	13,85%	135	**73,44%**	47	64,86%	72	**22,97%**	17	5,97%	27
		219		975		64		111		74		452
CB	**24,98%**	340	15,50%	828	**67,85%**	382	64,23%	431	**24,48%**	106	6,61%	194
		1361		5343		563		671		433		2935
Legenda	*Acert: Acertividade; Ind.: Indicações; Tot.: Total.*											

A partir dos resultados obtidos conclui-se que o sistema superou o desempenho da empresa em todos os cenários. Baseado na Análise I e na Análise II, pode-se dizer que a diferença percentual do sistema na primeira análise se manteve próximo dos 10 pontos percentuais, assim como na segunda. Talvez fosse esperado uma maior melhora no que se diz respeito ao Cenário 6 pois está englobando todos os registros da base de dados. Isto pode indicar que há de fato uma mudança no perfil de consumo dos clientes.

A partir da Análise II, com o decorrer dos cenários, nota-se um leve crescimento de desempenho no mesmo período de teste, isso se dá pela adição de novos registros a cada cenário de treinamento, fazendo com que o sistema possa generalizar melhor. Outro ponto importante é a mudança de proporção das classes da Região P. Esse fato mostra como o modelo conseguiu fazer bem a distinção das classes, mesmo em uma situação oposta (maior número de Fraudadores do que Não Fraudadores) ao que vinha normalmente acontecendo antes da pandemia da COVID-19.

Já na Análise III, pode-se concluir que o sistema se comporta muito bem mesmo que o treinamento tenha sido feito até 5 meses antes. Em outras palavras, a empresa não precisaria treinar o sistema a cada mês. Além disso, poderia ser definido um limiar de erro acima do qual seria indicada a necessidade de o sistema ser treinado novamente incluindo os dados mais recentes.

5.6.2. Metodologia por Filtragem

A abordagem por filtragem é restritiva por utilizar um filtro para melhor definição da separação das classes, por isso, como dito anteriormente, será avaliada apenas nos cenários chaves. A seguir da Tabela 30 até a Tabela 32 estão representadas as taxas de acertos alcançadas pela Metodologia F e comparado com a segunda abordagem em cada uma das análises.

i. Análise I:

Tabela 30 - Metodologia E x Metodologia F – Análise I.

CA	Metodologia E		Metodologia F		Empresa A	
	Acert.	Ind./ Tot.	Acert.	Ind./ Tot.	Acert. Empresa	Ind./ Tot.
G	17,07%	126 / 738	19,40%	218 / 1124	9,19%	681 / 7407
P	27,81%	252 / 906	44,10%	187 / 424	17,20%	662 / 3848
T	9,57%	9 / 94	15,89%	17 / 107	3,01%	34 / 1130
Legenda	*Acert: Acertividade; Ind.: Indicações; Tot.: Total.*					

ii. Análise II:

Tabela 31 - Metodologia E x Metodologia F - Análise II.

C6	Metodologia E		Metodologia F		Empresa A	
	Acert.	Ind./ Tot.	Acert.	Ind./ Tot.	Acert. Empresa	Ind./ Tot.
G	23.73%	28	17.60%	63	12.69%	173
		118		358		1363
P	78.26%	54	62.79%	27	66.92%	87
		69		43		130
T	12.22%	11	12.28%	7	3.36%	18
		90		57		536
Legenda	*Acert: Acertividade; Ind.: Indicações; Tot.: Total.*					

iii. Análise III:

Tabela 32 - Metodologia E x Metodologia F - Análise III.

CB	Metodologia E		Metodologia F		Empresa A	
	Acert.	Ind./ Tot.	Acert.	Ind./ Tot.	Acert. Empresa	Ind./ Tot.
G	25,20%	343	25,87%	296	15,50%	828
		1361		1144		5343
P	68,32%	371	71,86%	143	64,23%	431
		543		199		671
T	24,48%	106	15,52%	54	6,61%	194
		433		348		2935
Legenda	*Acert: Acertividade; Ind.: Indicações; Tot.: Total.*					

A Metodologia F, de acordo com os resultados das análises e pela natureza do algoritmo, tende a ser uma alternativa, em casos onde se possua dados em abundância. Isso fica evidente quando observadas as análises separadamente. Na análise I a janela de teste possui 6 meses, diferentemente da segunda análise, que teve o pior desempenho quando comparado com a segunda metodologia ou mesmo em relação à empresa. Porém, pode ser observado um potencial da abordagem, visto que alcançou quase 20 pontos percentuais a mais do que a Metodologia E na Região P do Cenário A.

5.6.3. Metodologia Completa

A fim de ter uma comparação mais direta foi abordado a unificação das duas metodologias em uma terceira, chamada de completa, como dito anteriormente. Assim como a Metodologia F tem restrições sobre os cenários que se desejava analisar, a Metodologia C também as possui, por ter a filtragem de registros.

i. Análise I:

Tabela 33 - Metodologia E/F x Metodologia C – Análise I.

CA	Metodologia E		Metodologia F		Metodologia C		Empresa A	
	Acert.	Ind./ Tot.	Acert.	Ind./ Tot.	Acert.	Ind./ Tot.	Acert. Empresa	Ind./ Tot.
G	17,07%	126 738	19,40%	218 1124	16,28%	423 2598	9,19%	681 7407
P	27,81%	252 906	44,10%	187 424	32,18%	232 721	17,20%	662 3848
T	9,57%	9 94	15,89%	17 107	8,81%	14 159	3,01%	34 1130
Legenda	*Acert: Acertividade; Ind.: Indicações; Tot.: Total.*							

ii. Análise II:

Tabela 34 - Metodologia E/F x Metodologia C – Análise II.

C6	Metodologia E		Metodologia F		Metodologia C		Empresa A	
	Acert.	Ind./ Tot.	Acert.	Ind./ Tot.	Acert.	Ind./ Tot.	Acert. Empresa	Ind./ Tot.
G	23.73%	28 118	17.60%	63 358	23.29%	34 146	12.69%	173 1363
P	78.26%	54 69	62.79%	27 43	73.40%	69 94	66.92%	87 130
T	12.22%	11 90	12.28%	7 57	14.06%	9 64	3.36%	18 536
Legenda	*Acert: Acertividade; Ind.: Indicações; Tot.: Total.*							

iii. Análise III:

Tabela 35 - Metodologia E/F x Metodologia C – Análise III.

CB	Metodologia E		Metodologia F		Metodologia C		Empresa A	
	Acert.	Ind./ Tot.	Acert.	Ind./ Tot.	Acert.	Ind./ Tot.	Acert. Empresa	Ind./ Tot.
G	25,20%	343	25,87%	296	24,11%	224	15,50%	828
		1361		1144		929		5343
P	68,32%	371	71,86%	143	66,67%	376	64,23%	431
		543		199		564		671
T	24,48%	106	15,52%	54	23,30%	123	6,61%	194
		433		348		528		2935
Legenda	*Acert: Acertividade; Ind.: Indicações; Tot.: Total.*							

A junção das metodologias trouxe uma melhora significativa para a Metodologia F principalmente no Cenário 6 e Cenário B. No Cenário A, não conseguiu superar a performance alcançada pela Metodologia F. Os outros resultados foram bastante similares àqueles alcançados pelas duas metodologias apresentadas.

5.7. Limiar de Decisão

Outra característica do sistema é sua perspectiva de modificação do limiar de decisão. Este limiar está diretamente ligado com o tamanho a lista de suspeitos de fraude indicados pelos modelos: quanto maior o limiar, maior a confiança e menor o número de ligações indicadas para inspeção; e quanto menor o limiar, menor a confiança e maior é o número de ligações a serem vistoriadas.

As Tabela 36Tabela 37 e Tabela 38, e as Figura 32Figura 33Figura 34 mostram, respectivamente, as variações do limiar de decisão para as Regiões G, P e T do Cenário B.

Tabela 36 - Variação do Limiar de Decisão para a Região G.

Região G			
Limiar de Decisão	**Acertividade**	**Fraudador**	**Inspeções**
0.50	24,98%	340	1361
0.55	30,14%	220	730
0.60	34,48%	100	290
0.65	46,67%	28	60
0.70	83,33%	5	6

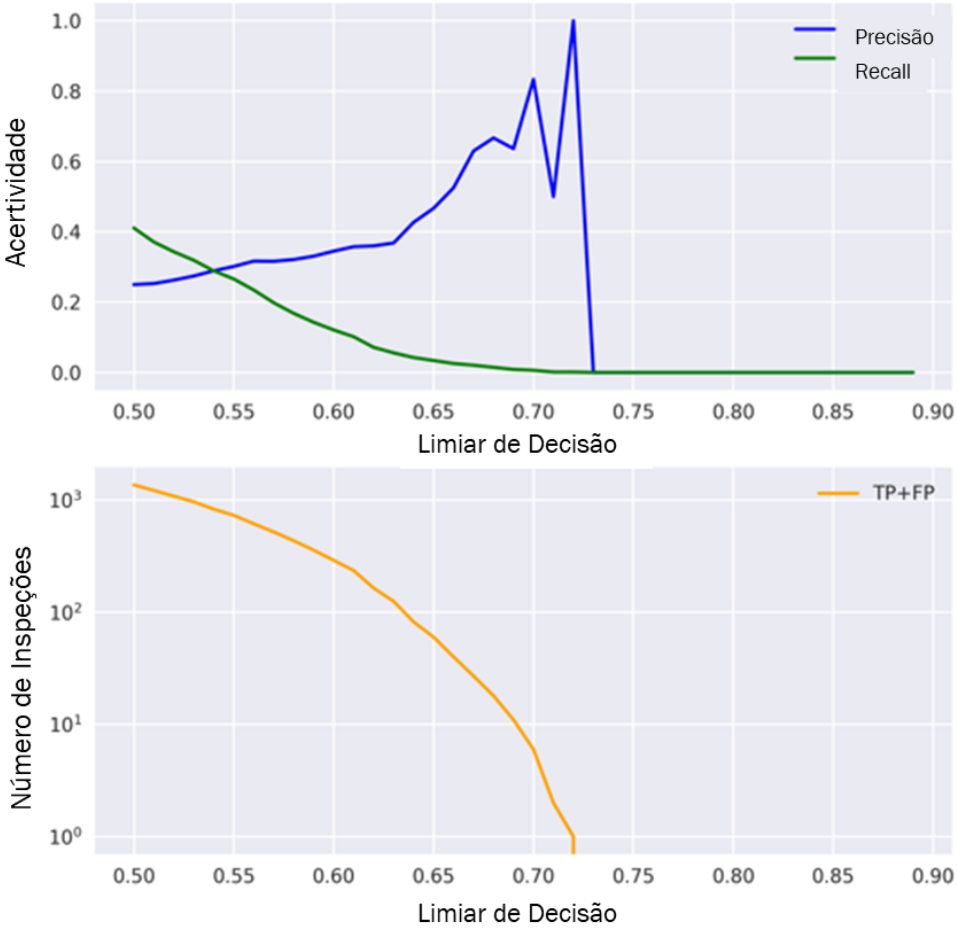


Figura 32 - Variação do Limiar de Decisão parar a Região G.

Tabela 37 - Variação do Limiar de Decisão para a Região P.

Região P			
Limiar de Decisão	**Acertividade**	**Fraudador**	**Inspeções**
0.50	67,85%	382	563
0.52	68,07%	356	523
0.60	65,64%	235	358
0.65	62,20%	153	246
0.70	55,47%	76	137

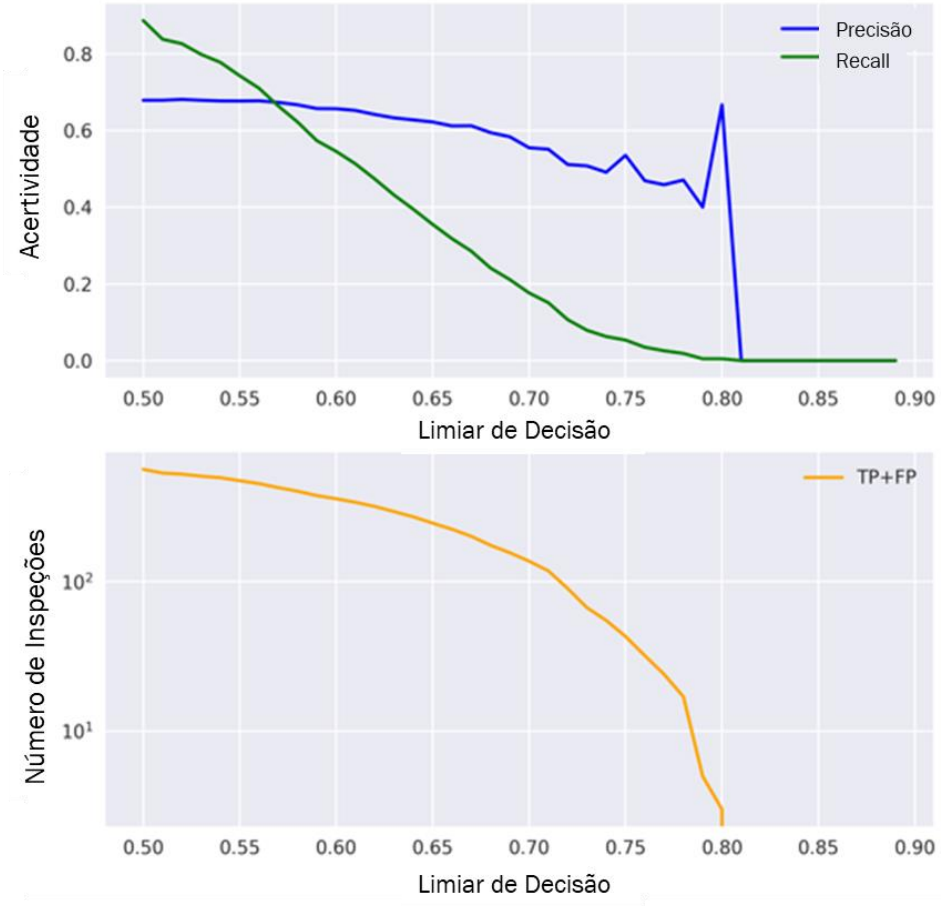


Figura 33 - Variação do Limiar de Decisão parar a Região P.

Essa técnica depende da quantidade de dados que estão sendo testados, ou seja, se houver abundância de registros, como pode ser visto na Região G, o resultado pode ser mais beneficiado pela mudança do limiar. Em outros casos, como na Região P e Região T, se houver menos registros, o impacto pode ser negativo ou nulo, respectivamente.

Tabela 38 – Variação do Limiar de Decisão para a Região T.

Região T			
Limiar de Decisão	**Acertividade**	**Fraudador**	**Inspeções**
0.50	20,65%	139	673
0.55	22,12%	119	538
0.60	24,29%	111	457
0.65	26,65%	101	379
0.70	27,05%	79	292

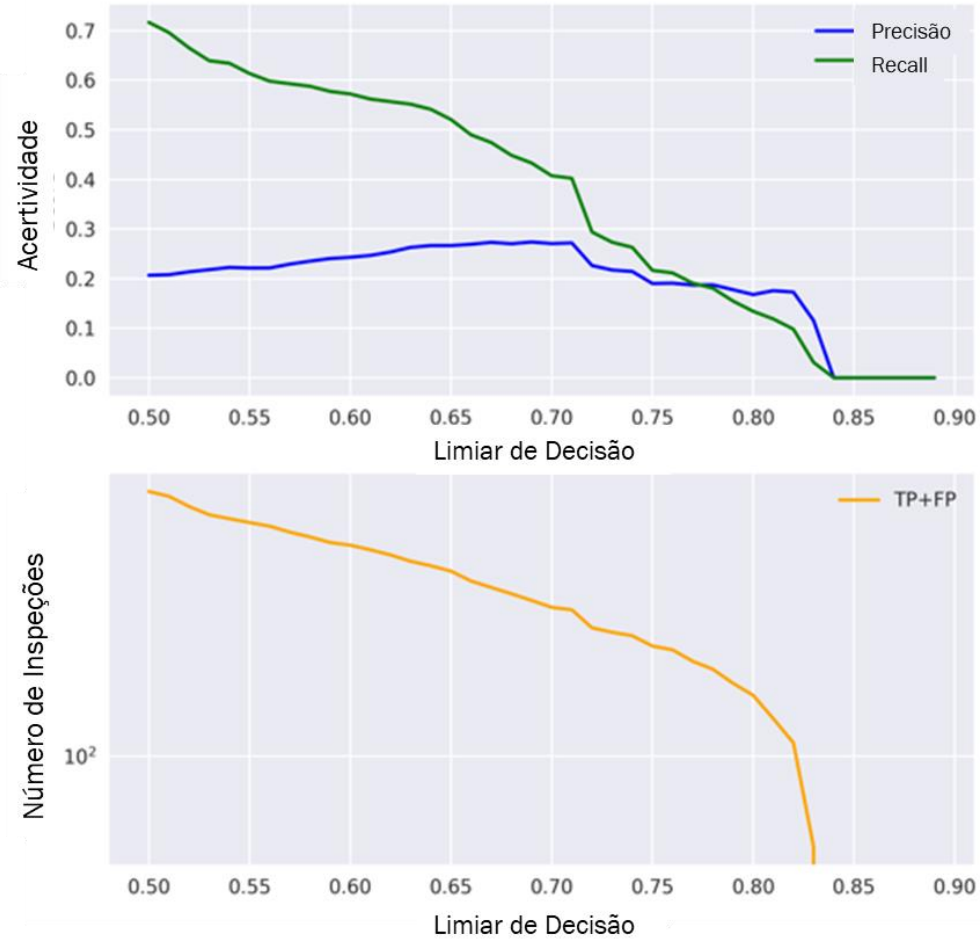


Figura 34 - Variação do Limiar de Decisão parar a Região T.

5.8. Testes Estatísticos

Para garantir que as amostras dos resultados apresentam diferenças significativas foi feito um teste estatístico como especificado na seção 4.3.1. O *Mann-Whitney U test* foi utilizado nas Análises II e III, pois são análises que comparam mais de um cenário com amostras diferentes e independentes. Os valores dos testes de significância das comparações do Cenário 6 na segunda análise e do Cenário B na terceira análise, podem ser vistos, respectivamente, nas Tabela 39 e Tabela 40. Foi utilizado o valor referência de p igual a 0,05 para rejeitar a hipótese nula.

Tabela 39 - *Mann-Whitney U Test* para Análise II.

p-valor		Cenários				
Regiões	X	C1	C2	C3	C4	C5
G	C6	7,65E-43	4,73E-14	3,14E-12	1,28E-57	4,48E-42
P		0,015759	6,80E-07	2,07E-05	0,018632	0,015759
T		1,25E-07	8,34E-38	0,004140	3,09E-34	8,72E-34

Tabela 40 - *Mann-Whitney U Test* para Análise III.

p-valor		Cenários				
Regiões	X	CB-1	CB-2	CB-3	CB-4	CB-5
G	CB	0,000326	1,91E-21	2,96E-05	0,051983	0,022947
P		0,004146	0,007979	0,00019	0,004501	6,77E-11
T		6,11E-10	1,79E-05	0,000006	7,91E-12	0,00956

A partir dos resultados obtidos pelos testes estatísticos, a hipótese nula de identidade pode ser rejeitada em todos os casos. Com isso pode-se indicar que um resultado é de fato superior a outro.

Já para comparar as Metodologias foi utilizado o *Wilcoxon Sign Rank Test*, pois se tratam de dois algoritmos diferentes em amostras semelhantes independentes. Os resultados dos valores de significância entre as duas metodologias para cada cenário estão mostrados na Tabela 41.

Tabela 41 – *Wilcoxon Sign Rank Test* entre as Metodologias F e E.

X p-valor		Metodologia II		
	Cenários	CA	C6	CB
	Regiões			
Metodologia I	G	*1,53E-09*	*0,003662*	*0,000163*
	P	*1,64E-16*	*0,004351*	*0,012142*
	T	*0,048503*	*0,004655*	*0,027402*

Novamente, através dos valores obtidos, pode se rejeitar a hipótese nula de identidade em todos os casos. Com isso pode-se indicar que um resultado é de fato superior a outro.

6. Conclusões e Trabalhos Futuros

6.1. Conclusões

O cenário nacional de distribuição de água tem tido bastante prejuízo no que se respeito às perdas. A maior parte dessas perdas tem origem comercial, ou se denomina aparente, que representam as ligações clandestinas, fraude ou qualquer tipo de irregularidade de natureza técnica.

Esse tipo de irregularidade tem origem quando o cliente quer reduzir custo ou até mesmo zerar a conta mensal de água levando a um prejuízo enorme para a prestadora. Esse prejuízo tem impacto direto tanto em âmbito social, econômico e até mesmo ambiental. A redução dessas perdas faz com que as empresas de saneamento aumentem o faturamento substancialmente. Além de trazer benefícios aos setores responsáveis por coletas de impostos e até mesmo para os clientes que, com uma melhora na distribuição de água, iriam pagar taxas menores.

Para diminuir essas perdas e irregularidades, as empresas prestadoras possuem heurísticas que, em um primeiro momento, levam à uma suspeita da fraude no consumo de água e, posteriormente, a empresa realize inspeções em busca dessas irregularidades. Porém, inspecionar cada cliente suspeito de fraude presente em uma região representa um serviço extremamente custoso, tanto em questão de tempo quanto em termos econômicos. Além disso, nem toda inspeção consegue detectar o cliente fraudador, mantendo o prejuízo da empresa.

O foco dessa dissertação foi a criação de um Sistema Inteligente para Identificação do Suspeito de Fraude no Consumo de Água, com o propósito de aumentar a taxa de acerto na detecção desses clientes irregulares, oferecendo maior suporte nas decisões.

A partir dos resultados, pode-se concluir que a inserção de novos atributos (Metodologia E) oriundos da série de consumo histórica e das variáveis exógenas, melhoraram substancialmente o resultado obtido quando se tratou de problemas com menos registros (Análise II) em comparação com a Metodologia F. Além disso, a composição dos resultados através de comitês de classificadores se mostrou benéfico em alguns cenários (Cenários 4 e 6).

A filtragem executada na Metodologia F se mostrou uma técnica poderosa quando há abundância de dados (Cenário A), porém, após o começo da pandemia, menos dados foram sendo registrados mensalmente e a diferença de proporção de fraudadores, no caso da Região P, prejudicou a etapa de aprendizado da metodologia fazendo com que fosse pior do que as heurísticas da própria empresa.

A Metodologia Completa apresentou ligeira melhora em relação a Metodologia por Filtragem, mostrando que a inserção de atributos e a junção das duas técnicas principais pode ser benéfico para o sistema. Quando comparado diretamente com a Metodologia Evolutiva, não demonstrou melhora significativa, por outro lado, manteve o bom desempenho alcançado. Com isso, conclui-se que o sistema pode alcançar precisões ainda maiores se a configuração das duas técnicas principais for otimizada.

Com os resultados obtidos no Capítulo 5, pode-se concluir também, que o sistema melhorou as métricas alcançadas pela Empresa A. Além disso, os diferentes Cenários comprovam como o sistema pôde se adaptar às diversas condições do problema, evidenciado no processo de generalização do que foi aprendido. Para a primeira análise os ganhos percentuais variam de 10 a 27% dependendo da região, enquanto que na segunda e na terceira análise os ganhos variam, respectivamente, de 6 a 12% e 3 a 9%. Esses ganhos representam um impacto direto na redução do furto de água e também do desperdício, resultando em melhoria financeira nas empresas de fornecimento de água, e, de forma geral, para a sociedade como um todo.

6.2. Trabalhos Futuros

Com objetivo de melhorar o sistema proposto, aumentando a robustez e a confiabilidade, alguns pontos têm espaço para aperfeiçoamentos e podem ser realizados futuramente:

- Investigar modelos aplicados a períodos menores do que 12 meses, beneficiando clientes que tenham históricos menores;
- Segmentar os clientes por patamares de consumo com o objetivo de verificar melhor aderência dos modelos, unificando ligações de alto consumo, como condomínios e clientes comerciais, por exemplo;

- Agregar o conjunto de dados de várias regiões, visando contornar o problema de volume de dados insuficiente para treinamento do modelo, no caso de pequenas quantidades de determinados tipos de clientes por região (clientes industriais, comerciais, etc.);
- Uso de outras informações exógenas, tais como localização geográfica do fraudador (bairro, CEP, etc.), número de fraudadores identificados em uma janela espaço-temporal na região avaliada, etc.;
- Outros algoritmos de aprendizado de máquina mais robustos, por exemplos os baseados em Redes Neurais Recorrentes, que demandam maior capacidade computacional, porém com potencial melhor de aprendizado e, consequentemente, possibilidade de alcançar maiores taxa de acertos.

Referências Bibliográficas

ABS DA CRUZ, A., VELLASCO, M., PACHECO, M.: **Quantum-inspired evolutionary algorithm for numerical optimization**. In: Proc. CEC, pp. 2630–2637 (2006).

AL-RADAIDEH, Q. A.; AL-ZOUBI, M. M. **A data mining based model for detection of fraudulent behaviour in water consumption**. In: 2018 9th International Conference on Information and Communication Systems (ICICS). IEEE, 2018. p. 48-54.

BACK, T.; FOGEL, D. B. ; MICHALEWICZ, Z., editors. **Handbook of Evolutionary Computation**. Institute of Physics Publishing, 1997. 1.1;

BLOCKEEL, H; DE RAEDT, L. **Top-down induction of first-order logical decision trees**. Artificial intelligence, v. 101, n. 1-2, p. 285-297, 1998.

BREIMAN, L. **Random forests**. Machine learning, v. 45, n. 1, p. 5-32, 2001.

BURIAN, R.: **Underappreciated pathways toward molecular genetics as illustrated by Jean Brachet's cytochemical embryology**. In: Sarkar, S. (ed.) The Philosophy and History of Molecular Biology: New Perspectives, pp. 67–85. Kluwer, Dordrecht (1996);

DARWIN, C.: **On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life**. Murray, London (1859);

CARVALHO, F. S. de; PEPLAU, G. R.; CARVALHO, G. S. de e PEDROSA, V. A. **Estudos Sobre Perdas No Sistema De Abastecimento De Água Da Cidade De Maceió**. In: VII Simpósio de Recursos Hidricos do Nordeste, 2004, São Luís. Anais do VII Simpósio de Recursos Hidricos do Nordeste, 2004.

CHEN, T; GUESTRIN, C. **Xgboost: A scalable tree boosting system**. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016. p. 785-794.

CORDER, Gregory W.; FOREMAN, Dale I. **Nonparametric statistics for non-statisticians**. 2011.

FERREIRA, Rita Cavaleiro de et al. **Caderno temático: Perdas de água e eficiência energética**. 2019.

FOGEL, D.; FOGEL, L.; ATMAR, J. Wirt. **Meta-evolutionary programming. In: Conference record of the twenty-fifth asilomar conference on signals, systems & computers**. IEEE computer Society, 1991. p. 540,541,542,543,544,545-540,541.

FOGEL, D.; FOGEL, L. **Evolutionary computation**. IEEE Transactions on neural networks, v. 5, n. 1, p. 1-2, 1994.

FOGEL, D.; FOGEL, L. **An introduction to evolutionary programming. In: European conference on artificial evolution**. Springer, Berlin, Heidelberg, 1995. p. 21-33.

FOGEL, L. **Intelligence through simulated evolution: forty years of evolutionary programming**. John Wiley & Sons, Inc., 1999.

GLASSNER, A.: **Quantum computing**, part 2. IEEE Comput. Graph. Appl. 86–95 (2001a);

GOPAL, G. V.; BALAJI, V. **Detection of fraudulent behaviour in water consumption using machine learning algorithms**. Proceedings of Journal of Engineering Sciences, v. 11, n. 7, p. 399-406, 2020.

HAN, K., KIM, J.: **Genetic quantum algorithm and its application to combinatorial optimization problem**. In: Proc. CEC, vol. 2, pp. 1354–1360 (2000);

HAN, K., KIM, J., **Quantum-inspired evolutionary algorithm for a class of combinatorial optimization,** IEEE Trans. Evol. Comput. 6(6), 580–593 (2002);

HAN, K., KIM, J.: **Quantum-inspired evolutionary algorithms with a new termination criterion, h-epsilon gate, and two-phase scheme**. IEEE Trans. Evol. Comput. 8(2), 156–169 (2004).

HEY, T.: **Quantum computing: an introduction**. Comput. Control Eng. J. 10(3), 105–112 (1999) Hinterding, R. :Representation, constraint satisfaction and the knap sack problem .In:Proc.CEC,pp.1286– 1292 (1999);

INSTITUTO TRATA BRASIL. **Perdas De Água 2021 (Snis 2019): Desafios Para Disponibilidade Hídrica E Avanço Da Eficiência Do Saneamento Básico**. Disponível em: http://www.tratabrasil.org.br/images/estudos/Perdas_dágua/Estudo_de_Perdas_2021.pdf. Acesso em: 22 jan. 2022.

KITTLER, J. and ALKOOT, F. M., **Sum versus vote fusion in multiple classifier systems**, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 1, pp. 110-115, jan. 2003, doi: 10.1109/TPAMI.2003.1159950.

KUNCHEVA, L.,

LIEMBERGER, Roland et al. **The Challenge of Reducing Non-Revenue Water in Developing Countries--How the Private Sector Can Help: A Look at Performance-Based Service Contracting**. 2006. Disponível em: http://documents1.worldbank.org/curated/en/385761468330326484/pdf/394050Reducing1e0water0WSS81PUBLIC1.pdf. Acesso em: 04 fev. 2022.

LIU, H., MOTODA, H., SETIONO, R. and ZHAO, Z., **Feature Selection: An Ever Evolving Frontier in Data Mining**, J. Mach. Learn. Res. Work. Conf. Proc. 10 Fourth Work. Featur. Sel. Data Min., pp. 4–13, 2010.

MANN, H.B. and WHITNEY, D.R. (1947) On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other. Annals of Mathematical Statistics, 18, 50-60. Disponível em: http://dx.doi.org/10.1214/aoms/1177730491 .

MINISTÉRIO DO DESENVOLVIMENTO REGIONAL. SECRETARIA NACIONAL DE SANEAMENTO - SNS. Sistema Nacional de Informações sobre Saneamento (SNIS). Diagnóstico Temático Serviços de Água e Esgoto. Brasília. 2021.

MOORE, M., NARAYANAN, A.: **Quantum-inspired computing. Technical Report**, Department of Computer Science, University Exeter, Exeter, UK (1995);

NARAYANAN, A., MOORE, M.: **Quantum-inspired genetic algorithms**. In: Proc. CEC, pp. 61–66 (1996);

NIELSEN, A.M., CHUANG, I.L.: **Quantum Computation and Quantum Information**. Cambridge University Press, Cambridge (2000);

ORTEGA, G.V.C.: Redes Neurais na. Identificação de Perdas Comerciais do Setor Elétrico. Rio de Janeiro, 2008. 184p. Dissertação de Mestrado - Departamento de Engenharia Elétrica.

PINHO, A. G., VELLASCO, M. and DA CRUZ, A. V. A., **A new model for credit approval problems: A quantum-inspired neuro-evolutionary algorithm with binary-real representation,"** 2009 World Congress on Nature & Biologically Inspired Computing (NaBIC), Coimbatore, 2009, pp. 445-450, doi: 10.1109/NABIC.2009.5393327.

RAMOS, A. C., GONZALES, R. and VELLASCO M., **Feature Selection methods applied to Motor Imagery task classification,** in 2016 IEEE Latin American Conference on Computational Intelligence (LA-CCI), 2016, pp. 1–6.

ROSENBLATT, F. **Principles of neurodynamics. perceptrons and the theory of brain mechanisms**. Cornell Aeronautical Lab Inc Buffalo NY, 1961.

SCHAPIRE, R. E. **Explaining adaboost**. In: Empirical inference. Springer, Berlin, Heidelberg, 2013. p. 37-52.

VAPNIK, V. **SVM method of estimating density, conditional probability, and conditional density**. In: 2000 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE, 2000. p. 749-752.

Apêndices

Apêndice A.1 – Perdas Reais (Físicas)

Como foi dito anteriormente, as perdas reais ou físicas, demonstram a quantidade ou o volume de água que é desperdiçado no processo natural de manuseio da água até o cliente – captação, tratamento, armazenamento e distribuição. A seguir, são apresentados os principais motivadores, assim como o impacto dessas perdas para a prestadora.

I. **Adução de Água Bruta:**
   a. **Origem:** Vazamento nas tubulações e limpeza do poço de sucção;
   b. **Impacto:** Variável, em função do estado das tubulações e da eficiência operacional.

II. **Tratamento:**
   a. **Origem:** Vazamentos estruturais, lavagem de filtros e descarga de lodo.
   b. **Impacto:** Significativo, em função do estado das tubulações e da eficiência operacional.

III. **Reserva:**
   a. **Origem:** Vazamentos estruturais, extravasamentos e limpeza;
   b. **Impacto:** Variável, em função do estado das tubulações e da eficiência operacional.

IV. **Adução de Água Tratada:**
   a. **Origem:** Vazamento nas tubulações e limpeza do poço de sucção e descargas;
   b. **Impacto:** Variável, em função do estado das tubulações e da eficiência operacional.

V. **Distribuição:**
   a. **Origem:** Vazamentos na rede, vazamentos em ramais e descargas;
   b. **Impacto:** Significativo, em função do estado das tubulações e principalmente das pressões.

Os tipos de vazamentos estão apresentados na Figura 35 e descritos logo a seguir.

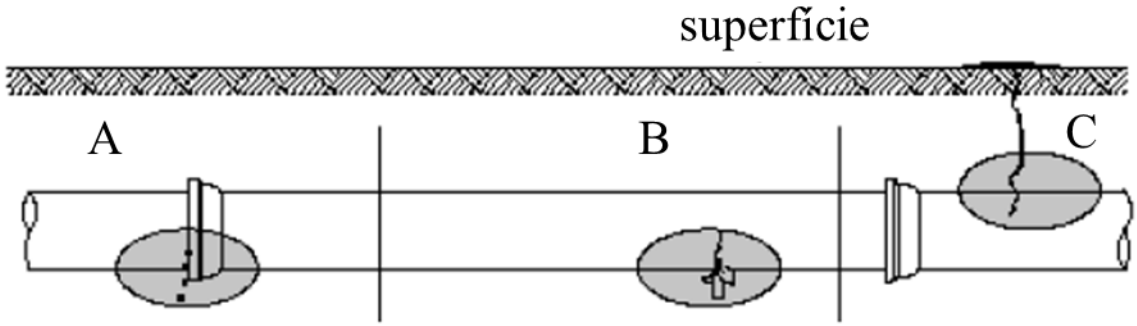


Figura 35 - Tipos de vazamentos em uma rede de distribuição.

Fonte: Modificado de Carvalho et al., 2004.

A. Vazamentos de baixa vazão, não visíveis, não aflorantes e não detectáveis por métodos acústicos. Equivale a 25% do volume perdido. Ações corretivas: Redução de pressão, uso de materiais e mão de obra de melhor qualidade.
B. Vazamentos não visíveis, não aflorantes e detectáveis por métodos acústicos. Representa 30% do volume perdido. Ações de correção: Redução de pressão e pesquisa de vazamentos.
C. Vazamentos visíveis aflorantes ou correntes nos cavaletes. Representam 45% do volume perdido. Ações corretivas: Redução de pressão.

Quando há um alto nível de perdas reais, acarreta uma obtenção e produção de água maior do que o necessário, afetando diretamente o custo de produção e a demanda hídrica. Além disso, gera ineficiência em:

- Produção:
    - Maior custo dos insumos químicos, energia para bombeamento, entre outros fatores de produção;
    - Maior manutenção da rede e equipamentos;
    - Uso excessivo da capacidade de produção e de distribuição; e
    - Maior custo oriundo da possível utilização de fontes de abastecimento alternativas de menor qualidade ou de difícil acesso.

- Ambiental:
    - Pressão excessiva sobre as fontes de abastecimento do recurso hídrico; e
    - Maior custo de mitigação dos impactos negativos dessa atividade (externalidade).

Apêndice A.2 – Caracterização e Nível de Eficiência das Perdas

Na Tabela 42, estão descritas as principais causas e consequências das perdas reais e aparentes em um sistema de abastecimento de água potável.

Tabela 42 - Caracterização de Perdas reais e aparantes.

Itens	**Características Principais**	
	Perdas Reais	**Perdas Aparentes**
Tipo de ocorrência mais comum	Vazamento	Erro de medição
Custos associados ao volume de água perdido	Custo de produção	- Tarifa - Receita Operacional
Efeitos no Meio Ambiente	- Desperdício do Recurso Hídrico - Necessidades de ampliações de mananciais	-
Efeitos na Saúde Pública	Risco de contaminação	-
Empresarial	Perda do produto	Perda de receita
Consumidor	- Imagem negativa (ineficiência e desperdício)	-
Efeitos no Consumidor	- Repasse para tarifa - Desincentivo ao uso racional	- Repasse para tarifa

		- Incitamento a roubos e fraudes

Fonte: ITB, 2021.

Sabe-se que não há a possibilidade de uma rede de distribuição de água não ter vazamentos, ou seja, é inviável eliminar completamente as perdas de água. Por isso a IWA, estabeleceu limites eficientes para a redução de perdas, são eles: 1) limites econômicos, que representam o volume cujos custos para reduzir as perdas são maiores do que o valor dos volumes recuperados; 2) limite técnico – ou "perdas inevitáveis" – é o volume mínimo definido pelo alcance das tecnologias atuais dos materiais, das ferramentas, dos equipamentos e logística. Na Figura 36 é possível observar um gráfico conceitual para o nível econômico ótimo de vazamentos assim como o nível mínimo de vazamentos. Observa-se que o custo total se dá pela soma do custo da água – que é diretamente proporcional ao tempo decorrido entre o início do vazamento e a conclusão do reparo – mais o custo de detecção e reparo – que varia conforme os ciclos de identificação. Portanto, o nível ótimo se dá pelo ponto da curva onde o custo total atinge o seu valor mínimo, ou em outras palavras, o nível econômico de vazamento. Já o nível mínimo de vazamento se dá pelas perdas que não podem ser evitadas. Consequentemente, haverá um volume mínimo de água perdida.

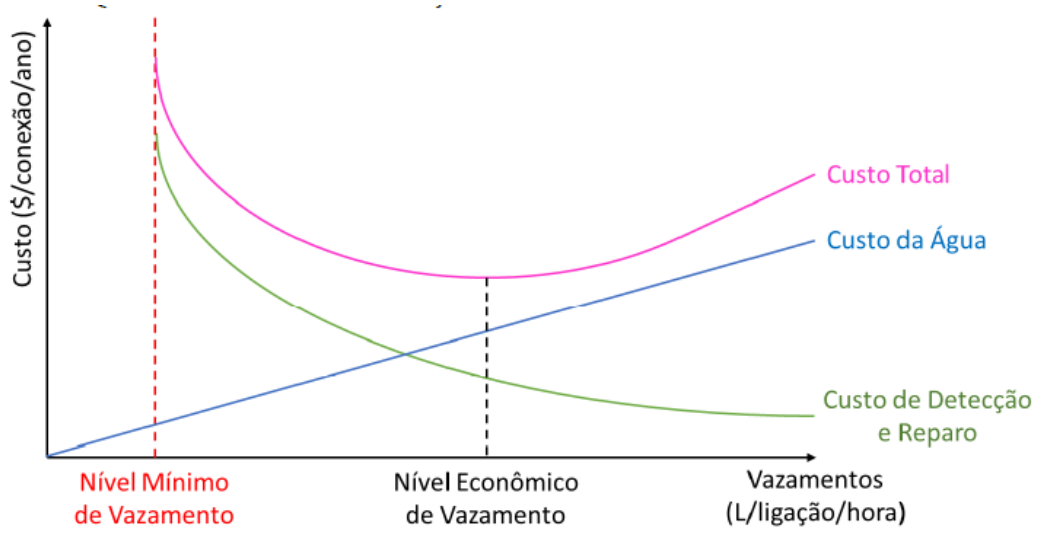


Figura 36 - Determinação do nível eficiente de perdas.

Fonte: United States Environmental Protection Agency (USEPA). Elaboração: ITB, 2021.

Apêndice B

Tabela 43 - Precisão do Conjunto de Validação dos Algoritmos para o Cenário A – Metodologia E.

Cenário	CA		
Região / Algoritmos	Região G	Região P	Região T
MLP	0,1649	0,2389	0,0806
SVM	0,1662	0,1763	0,0623
DT	0,1889	0,2505	0,0782
RF	0,1801	0,2592	0,0715
XGB	0,1822	0,2397	0,0706
Votação Maioria	0,1785	0,2494	0,0823
Soma Ponderada	0,1723	0,2134	0,0812
Fusão Prob.	0,1794	0,2265	0,078

Tabela 44 - Precisão do Conjunto de Validação dos Algoritmos para o Cenário 1 e 2 – Metodologia E.

Cenário	C1			C2		
Região	G	P	T	G	P	T
Algoritmos						
MLP	0,1766	0,8214	0,1940	0,2461	0,6634	0,2297
SVM	0,2571	0,5633	0,1928	0,2666	0,7446	0,1704
DT	0,4078	0,8148	0,2222	0,2501	0,8095	0,2400
RF	0,3867	0,8802	0,3132	0,3489	0,8245	0,2116
XGB	0,3733	0,8181	0,2878	0,3083	0,7966	0,2982
Votação Maioria	0,3736	0,8204	0,2650	0,2734	0,8291	0,2264
Soma Ponderada	0,3691	0,8726	0,2472	0,2668	0,8076	0,2160
Fusão Prob.	0,3728	0,8342	0,2507	0,2734	0,8351	0,2244

Tabela 45 - Precisão do Conjunto de Validação dos Algoritmos para o Cenário 3 e 4 – Metodologia E.

Cenário	C3			C4		
Região	G	P	T	G	P	T
Algoritmos						
MLP	0,1948	0,6987	0,1654	0,1913	0,6595	0,1587
SVM	0,2232	0,6833	0,1093	0,2346	0,7042	0,1593
DT	0,2218	0,6407	0,1717	0,2311	0,7492	0,1474
RF	0,2241	0,7311	0,1479	0,2162	0,7528	0,1692
XGB	0,2561	0,7590	0,1567	0,2542	0,7404	0,1666
Votação Maioria	0,2272	0,7327	0,1540	0,2646	0,7671	0,1722
Soma Ponderada	0,2246	0,7319	0,1664	0,2436	0,7490	0,1702
Fusão Prob.	0,2367	0,6713	0,1876	0,2269	0,7554	0,1608

Tabela 46 - Precisão do Conjunto de Validação dos Algoritmos para o Cenário 5 e 6 – Metodologia E.

Cenário	C5			C6		
Região	G	P	T	G	P	T
Algoritmos						
MLP	0,1975	0,6386	0,1191	0,1556	0,2927	0,0890
SVM	0,1795	0,6639	0,1518	0,1651	0,2338	0,0799
DT	0,1883	0,6562	0,1827	0,1984	0,3166	0,1100
RF	0,1934	0,7090	0,1486	0,1986	0,2949	0,1183
XGB	0,2002	0,6666	0,1339	0,2105	0,3319	0,1131
Votação Maioria	0,1866	0,6867	0,1605	0,2173	0,3298	0,1201
Soma Ponderada	0,1788	0,6692	0,1605	0,2058	0,3296	0,1215
Fusão Prob.	0,1907	0,6867	0,1579	0,2021	0,3466	0,1187

Tabela 47 - Precisão do Conjunto de Validação dos Algoritmos para o Cenário B e B-1 – Metodologia E.

Cenário	CB			CB-1		
Região	G	P	T	G	P	T
Algoritmos						
MLP	0,1601	0,2859	0,0957	0,1561	0,2833	0,0454
SVM	0,1498	0,2519	0,0779	0,1198	0,2023	0,0753
DT	0,1897	0,2845	0,0989	0,1984	0,2808	0,0748
RF	0,1901	0,2847	0,0701	0,1817	0,2770	0,0744
XGB	0,1900	0,2827	0,0708	0,1902	0,2840	0,0709
Votação Maioria	0,1742	0,2713	0,0913	0,1957	0,2938	0,0730
Soma Ponderada	0,1750	0,2713	0,0796	0,1661	0,2958	0,0722
Fusão Prob.	0,1503	0,2817	0,0821	0,1890	0,2732	0,0746

Tabela 48 - Precisão do Conjunto de Validação dos Algoritmos para o Cenário B-2 e B-3 – Metodologia E.

Cenário	CB-2			CB-3		
Região / Algoritmos	G	P	T	G	P	T
MLP	0,1550	0,2689	0,0809	0,1547	0,2791	0,0896
SVM	0,1232	0,2439	0,0892	0,1573	0,2461	0,0849
DT	0,1897	0,2918	0,1075	0,1931	0,2731	0,0998
RF	0,1831	0,2825	0,0904	0,1866	0,2755	0,0963
XGB	0,1878	0,2983	0,0867	0,1891	0,2708	0,0910
Votação Maioria	0,1720	0,2992	0,1045	0,1889	0,2862	0,0952
Soma Ponderada	0,1673	0,2902	0,1017	0,1838	0,2895	0,0973
Fusão Prob.	0,1773	0,2883	0,0956	0,1768	0,2732	0,0833

Tabela 49 - Precisão do Conjunto de Validação dos Algoritmos para o Cenário B-4 e B-5 – Metodologia E.

Cenário	CB-4			CB-5		
Região / Algoritmos	G	P	T	G	P	T
MLP	0,1493	0,2442	0,0759	0,1584	0,2760	0,0714
SVM	0,1577	0,2519	0,0830	0,1538	0,2136	0,0780
DT	0,2024	0,3087	0,0931	0,1951	0,2842	0,0732
RF	0,2058	0,3158	0,0857	0,1856	0,2828	0,0794
XGB	0,1907	0,2945	0,0771	0,1889	0,2827	0,0781
Votação Maioria	0,1908	0,2869	0,0821	0,1821	0,2765	0,0721
Soma Ponderada	0,1892	0,2963	0,0786	0,1662	0,2765	0,0742
Fusão Prob.	0,1823	0,2861	0,0856	0,1523	0,2732	0,0753

Tabela 50 - Precisão do Conjunto de Validação dos Algoritmos para o Cenário A – Metodologia F.

Cenário	CA		
Região / Algoritmos	Região G	Região P	Região T
RF	0,2241	0,2331	0,0286
ADAB	0,2110	0,2390	0,0337
MLP	0,2014	0,2031	0,0208
SVM	0,2138	0,2372	0,0393
XGB	0,1979	0,2371	0,0370
Votação Maioria	0,2177	0,2423	0,0352
Soma Ponderada	0,2133	0,2401	0,0312
Fusão Prob.	0,2023	0,2226	0,0277

Tabela 51 - Precisão do Conjunto de Validação dos Algoritmos para o Cenário 6 – Metodologia F.

Cenário	C6		
Região / Algoritmos	Região G	Região P	Região T
RF	0,1863	0,2902	0,1073
ADAB	0,1811	0,2823	0,1071
MLP	0,1869	0,2777	0,1135
SVM	0,1967	0,3116	0,0861
XGB	0,1904	0,3135	0,1011
Votação Maioria	0,1846	0,3032	0,1156
Soma Ponderada	0,1878	0,2978	0,1096
Fusão Prob.	0,1789	0,3086	0,1012

Tabela 52 - Precisão do Conjunto de Validação dos Algoritmos para o Cenário B – Metodologia F.

Cenário	CB		
Região / Algoritmos	Região G	Região P	Região T
RF	0,1960	0,2834	0,1082
ADAB	0,1897	0,2808	0,1020
MLP	0,1952	0,2585	0,0930
SVM	0,2074	0,2738	0,0932
XGB	0,2040	0,2701	0,0820
Votação Maioria	0,1922	0,2725	0,1033
Soma Ponderada	0,1879	0,2802	0,0903
Fusão Prob.	0,1898	0,2709	0,0988