Instituto Federal da Bahia
Campus Salvador

Programa de Pós-Graduação em Engenharia de Sistemas e Produtos

WATER FRAUD ANALYTICS – UM MODELO DE MACHINE LEARNING PARA DETECÇÃO DE FRAUDES EM CONSUMO DE ÁGUA

Márcio Nunes de Souza

DISSERTAÇÃO DE MESTRADO

Salvador
15 de dezembro de 2021

MÁRCIO NUNES DE SOUZA

WATER FRAUD ANALYTICS – UM MODELO DE MACHINE LEARNING PARA DETECÇÃO DE FRAUDES EM CONSUMO DE ÁGUA

Esta Dissertação de Mestrado foi apresentada ao Programa de Pós-Graduação em Engenharia de Sistemas e Produtos da Instituto Federal da Bahia, como requisito parcial para obtenção do grau de Mestre em Engenharia de Sistemas e Produtos.

Orientador: Renato Lima Novais
Coorientador: Rodrigo Tripodi Calumby

Salvador
15 de dezembro de 2021

Biblioteca Raul V. Seixas – Instituto Federal de Educação, Ciência e Tecnologia da Bahia - IFBA – Campus Salvador/BA.
Responsável pela catalogação na fonte: Samuel dos Santos Araújo - CRB 5/1426.

S729w Souza, Márcio Nunes de.
Water fraud analytics – um modelo de machine learning para detecção de fraudes em consumo de água / Márcio Nunes de Souza. Salvador, 2021.
65 f. ; 30 cm.

Dissertação (Mestrado Profissional em Engenharia de Sistemas e Produtos) – Instituto Federal de Educação, Ciência e Tecnologia da Bahia.
Orientador: Prof. Dr. Renato Lima Novais.
Coorientador: Prof. Dr. Rodrigo Tripodi Calumby.

1. Detecção de fraudes. 2. Aprendizado de máquina. 3. Perdas de água. ´ 4. Perdas aparentes. I. Novais, Renato Lima. II. Calumby, Rodrigo Tripodi. III. IFBA. IV. Título.

CDU 2 ed. 628.1

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA BAHIA
R. Emídio dos Santos - Bairro Barbalho - CEP 40301-015 - Salvador - BA - www.portal.ifba.edu.br

INSTITUTO FEDERAL DA BAHIA

PRÓ-REITORIA DE PESQUISA, PÓS-GRADUAÇÃO E INOVAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE SISTEMAS E PRODUTOS - PPGESP

"WATER FRAUD ANALYTICS - UM MODELO DE *MACHINE LEARNING* PARA DETECÇÃO DE FRAUDES EM CONSUMO DE ÁGUA"

MÁRCIO NUNES DE SOUZA

Produto(s) Gerado(s): Dissertação;

Orientador: Prof. Dr. Renato Lima Novais
Coorientador: Prof. Dr. Rodrigo Tripodi Calumby

Banca examinadora:

Prof. Dr. Renato Lima Novais
Orientador – Instituto Federal da Bahia (IFBA)

Prof. Dr. Rodrigo Tripodi Calumby
Coorientador e Membro Externo - Universidade Estadual de Feira de Santana (UEFS)

Prof. Dr. Francisco José da Silva Borges de Santana
Membro Interno – Instituto Federal da Bahia (IFBA)

Prof. Dr. Alexandre da Costa e Silva Franco
Membro Externo – Instituto Federal da Bahia (IFBA)

Trabalho de Conclusão de Curso aprovado pela banca examinadora em 15/12//2021

Em 05 de dezembro de 2021.

Documento assinado eletronicamente por **RENATO LIMA NOVAIS**, **Docente Permanente**, em 15/12/2021, às 19:03, conforme decreto nº 8.539/2015.

Documento assinado eletronicamente por **FRANCISCO JOSE DA SILVA BORGES DE SANTANA**, **Docente Colaborador(a)**, em 15/12/2021, às 20:29, conforme decreto nº 8.539/2015.

Documento assinado eletronicamente por **Rodrigo Tripodi Calumby**, **Usuário Externo**, em 17/12/2021, às 08:25, conforme decreto nº 8.539/2015.

Documento assinado eletronicamente por **ALEXANDRE DA COSTA E SILVA FRANCO**, **Professor Efetivo**, em 18/12/2021, às 14:35, conforme decreto nº 8.539/2015.

A autenticidade do documento pode ser conferida no site http://sei.ifba.edu.br/sei/controlador_externo.php?acao=documento_conferir&acao_origem=documento_conferir&id_orgao_acesso_externo=0 informando o código verificador **2106682** e o código CRC **3568983B**.

Dedico este trabalho aos meus pais, Sr. Nivaldo Nunes $(in\ memory)$ e Sra. Ana Martins. Eles me ensinaram a ser resiliente, mesmo sem conhecer o significado desta palavra.

AGRADECIMENTOS

Agradeço a Deus pela oportunidade em realizar este trabalho. Agradeço à minha família por compreender a necessidade de ausência em muitos momentos importantes. Agradeço também aos meus orientadores por me guiarem em todo o processo doloroso da pesquisa científica.

A persistência é o caminho do êxito.

—AUTOR (Charles Chaplin)

RESUMO

As perdas de água na distribuição ocorrem com muita frequência no setor de saneamento. Dentre os tipos de perdas, as fraudes correspondem ao volume de água furtado pelos usuários e, tradicionalmente, são combatidas por meio da inspeção *in loco* da rede de abastecimento. A identificação das possíveis fraudes é uma atividade complexa e sua inspeção bastante custosa, acarretando na baixa taxa de mitigação e na manutenção de grande percentual das perdas. Neste contexto, o desenvolvimento de soluções tecnológicas que ajudem nessa tarefa é de bastante valia. Dentre elas, técnicas estatísticas e de aprendizado de máquina têm sido aplicadas para detecção de fraudes com resultados promissores. Contudo, estudos anteriores têm sido realizados com diversas limitações em termos do volume de dados utilizado e do rigor científico das análises. Este trabalho apresenta um estudo para a detecção de fraudes no consumo de água usando algoritmos de aprendizado de máquina supervisionada, considerando uma base de dados em larga escala e procedimentos experimentais com foco na adequada construção e avaliação dos modelos preditivos e do seu poder de generalização. Foram utilizadas técnicas de pré-processamento de dados, otimização de modelos baseada em validação cruzada e avaliação considerando dados independentes. Foram avaliados diversos algoritmos de aprendizagem de máquina, com o melhor modelo com acurácia geral de 79.62%, precisão de 81.70% e revocação de 76.34%. Assim, o modelo é capaz de identificar corretamente mais de 76% das fraudes e cerca de 83% dos registros idôneos.

Palavras-chave: Detecção de Fraudes, Aprendizado de Máquina, Perdas de Água, Perdas Aparentes

ABSTRACT

Water losses in distribution occur very frequently in the sanitation sector. Among the types of losses, frauds correspond to the volume of water stolen by users and, traditionally, are fought through inspection *in loco* of the supply network. The identification of possible frauds is a complex activity and costly, resulting in a low rate of mitigation and maintenance of a large percentage of losses. In this context, the development of technological solutions is of great value. Among they, statistical and machine learning techniques have been applied to detect fraud in various sectors with promising results. However, previous studies carried out have had several limitations in terms of the volume of data and the scientific rigor of the analyses. This work presents a study of fraud detection in water consumption using a set of supervised machine learning algorithms. We considered a large-scale database and experimental procedures focusing on the construction and evaluation of predictive models and their power of generalization. Among the evaluated machine learning algorithms, the best model with an overall accuracy of 79.62%, precision of 81.70% and recall of 76.34%. Thus, the model is able to correctly identify more than 76% of frauds and about 83% of reputable records.

Keywords: Fraud Detection, Machine Learning, Water Loss, Apparent Water Loss

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

Capítulo 1

INTRODUÇÃO

1.1 CONTEXTO

No setor do saneamento, a diferença entre o volume de água produzido e o volume consumido é caracterizado como perdas. No Brasil, de acordo com o Sistema Nacional de Informações sobre Saneamento (SNIS), a média do Índice de Perdas na Distribuição (IPD) nos últimos anos é de aproximadamente 38% (BRASIL, 2019)(BRASIL, 2020). De acordo com o Instituto Trata Brasil, essas perdas são equivalentes a R$ 10,5 bilhões em prejuízos financeiros (BRASIL, 2020). Este é um indicador que demonstra o alto índice de ineficiência das empresas de saneamento do país. De acordo com os padrões mundiais, cidades com excelência no abastecimento de água têm indicadores de perdas menores do que 20% (BRASIL, 2020). A recente aprovação do novo marco legal do saneamento básico — Lei federal nº 14.026/2020 (BRASIL, 2020) — inclui metas para a redução destas perdas pelas empresas de saneamento.

As perdas são classificadas em perdas reais e perdas aparentes. Essa classificação é importante, devido às diversas ferramentas para a gestão e o combate para as respectivas perdas (BRASIL, 2019).

O volume de perdas de um sistema de abastecimento de água é um fator chave na avaliação da eficiência de um operador de saneamento (BRASIL, 2020). O elevado índice de perdas implica na redução do faturamento, impactando na diminuição da capacidade das empresas em realizar investimentos para expansão e melhorias nos serviços, além dos danos ao meio ambiente pela exploração dos mananciais (IFC, 2013).

A maioria das empresas de saneamento realizam campanhas de combate às perdas, concentrando os maiores esforços no combate às perdas reais. Em relação às perdas aparentes, o combate se dá a partir da substituição dos hidrômetros, melhorias no cadastro comercial e com o combate às fraudes.

A detecção das fraudes enfrenta obstáculos de difícil resolução, como a variação no consumo, a sazonalidade em decorrência do clima e as mudanças de premissas que não são controladas pelas empresas. Com isso, tradicionalmente, é realizada a partir de denúncias de outros consumidores, através das inspeções em redes com altos índices de perdas e

evidentes alterações no padrão de consumo (GUMIER; LUVIZOTTO JUNIOR et al., 2007)(AESBE, 2015)(DE CASTRO FETTERMANN et al., 2015)(TARDELLI FILHO, 2016).

As perdas decorrentes de fraudes resultam em prejuízos financeiros para as empresas de saneamento. As fraudes causam a redução do faturamento e o aumento dos custos com insumos e mão-de-obra para a operação dos sistemas (GUMIER; LUVIZOTTO JUNIOR et al., 2007)(DE CASTRO FETTERMANN et al., 2015).

1.2 JUSTIFICATIVA

O desenvolvimento de sistemas de suporte à decisão que orientem as práticas de manutenção de sistemas e controle de perdas é um campo de investimento em tecnologias e inovação no saneamento (NASCIMENTO; HELLER, 2005).

Muitas soluções para detecção e prevenção de fraudes foram propostas e desenvolvidas ao longo do tempo. Tais soluções são largamente aplicadas aos mais variados ramos de negócios, como telecomunicações, transações bancárias, lavagem de dinheiro e consumo de energia elétrica (BOLTON; HAND, 2002)(FAWCETT; PROVOST, 1996)(PHUA et al., 2010). Porém, poucos trabalhos foram encontrados com propostas de soluções para as fraudes no saneamento, a exemplo de Passini e Toledo (2002), Humaid e Barhoum (2013), Fernandes (2014), De Castro Fettermann et al. (2015), Monedero et al. (2015), Detroz e Silva (2017), Morote e Hernández-Hernández (2018), Al-Radaideh e Al-Zoubi (2018), Uddin et al. (2019), SRIRAMULU et al. (2020), GOPAL e BALAJI (2020), SREEDEVI e SWATHI (2021), Espinosa, Gisselot e Arriagada (2020), Sreekanth e Thinakaran (2021).

Estes estudos demonstram a aplicação de diversas técnicas para a detecção de fraudes no saneamento, porém, os que utilizam técnicas de *machine learning* apresentam resultados insatisfatórios ou falhas de metodologia (e.g. baixo volume de dados, dados para avaliação usados durante o treinamento do modelo, configurações dos algoritmos ou método simples de avaliação baseado apenas em treino/teste). Estes problemas podem comprometer os resultados dos modelos na prática. A Seção 2.3 descreve com mais detalhes os trabalhos relacionados que foram encontrados durante a realização da pesquisa bibliográfica.

A maioria das soluções propostas para detecção de fraudes, nos mais variados ramos de negócios, são baseadas em métodos e técnicas de *machine learning*, como a classificação e clusterização dos clientes de acordo os perfis e características comuns, a previsão de consumo, a detecção de *outliers* e redes de relacionamentos (BAESENS; VLASSELAER; VERBEKE, 2015).

Neste contexto, a motivação deste trabalho é a necessidade de redução de perdas e a falta de soluções adequadas para identificar fraudes no consumo de água. Essa motivação baseia-se na hipótese de que há ainda muito espaço para investir esforços em soluções para detectar tais fraudes a partir de características comuns dos usuários, que possam ser identificadas e aprendidas por algoritmos de *machine learning*.

1.3 DEFINIÇÃO DO PROBLEMA

Detectar fraudes de consumo em sistemas de distribuição de água não é algo trivial e apresenta várias dificuldades. Entre os desafios encontrados, um dos principais é a identificação de anormalidades no consumo, devido aos diversos fatores que impactam na variação do consumo de água: o clima, a quantidade de pessoas, a classe social, o padrão dos imóveis, dentre outros (DE CASTRO FETTERMANN et al., 2015).

Além do mais, ainda existem muitas dificuldades em detectar fraudes pelos métodos tradicionais. Isso devido ao alto custo das inspeções nas redes de distribuição, da dificuldade em aumentar a quantidade de análises manuais, e da dificuldade em descobrir novos métodos de fraudes até então desconhecidos.

Por outro lado, diversas aplicações de detecção de fraudes usam *machine learning* como meio de solução para este problema. Entretanto, apesar de os trabalhos relacionados a fraudes no consumo de água, descritos na Seção 2.3, demonstrar a aplicabilidade de técnicas de análises de dados na resolução do problema, não foi possível encontrar uma solução desenvolvida com maior rigor científico e resultados satisfatórios.

Isto posto, o problema tratado neste trabalho é o de detectar fraudes no consumo de água através de técnicas de *machine learning*. A solução apresentada obteve resultados melhores do que os trabalhos anteriores (acurácia geral 79,62%) e fez contribuições com o estado da arte do uso de *machine learning* para a detecção de fraudes no consumo de água (e.g. alto volume de dados e rigorosa metodologia para desenvolvimento do modelo).

1.4 OBJETIVOS

Para satisfazer a necessidade de inovar nos métodos para detecção e prevenção de fraudes no consumo de água, o objetivo geral deste trabalho visa o desenvolvimento de um modelo analítico que usa *machine learning* para auxiliar na detecção de fraudes no consumo de água e, consequentemente, proporcionar a redução das perdas aparentes de água.

O modelo proposto é capaz de identificar consumidores com suspeita de fraudes, os quais devem ser alvo de fiscalização *in loco*.

Para alcançar este objetivo, as seguintes etapas foram desenvolvidas:

1. Estudo de trabalhos anteriores;
2. Análise exploratória dos dados existentes;
3. Análise comparativa com diferentes algoritmos;
4. Desenvolvimento do modelo;
5. Estudo experimental para avaliação dos resultados.

1.5 CONTRIBUIÇÕES REALIZADAS

As contribuições realizadas por este trabalho à comunidade científica são:

1. Desenvolvimento de modelo analítico para a detecção de fraudes no consumo de água;
2. Contribuição para o Estado da Arte no uso de *machine learning* para a detecção de fraudes no saneamento;
3. Registro do *software*;
4. Submissão de artigo científico.

1.6 LIMITAÇÕES DE ESCOPO

Para o desenvolvimento deste estudo, foram obtidos dados referentes a consumidores das cidades de Salvador e Feira de Santana (Bahia, Brasil) do ano 2018. Esta foi uma limitação no volume e diversidade de dados para desenvolver os modelos analíticos, pois num cenário ideal, seriam necessários dados de diversos anos para analisar as variações de consumo.

O modelo implementado no contexto deste trabalho foi feito sob a forma de um protótipo funcional para identificar consumidores que podem estar cometendo fraudes no consumo de água. Aspectos relacionados a usabilidade, segurança e desempenho não foram tratados devido às restrições de escopo. Em trabalhos futuros ou eventual desenvolvimento de sistema produtivo, esses aspectos serão tratados.

1.7 ORGANIZAÇÃO DO TRABALHO

O restante deste trabalho está organizado da seguinte forma: O Capítulo 2 apresenta os principais conceitos a respeito de perdas aparentes decorrentes de fraudes (Seção 2.1), conceitos de *machine learning* (Seção 2.2) e trabalhos relacionados (Seção 2.3). O estudo exploratório realizado para o desenvolvimento dos modelo analítico e os resultados obtidos são vistos no Capítulo 3. Por fim, o Capítulo 4 apresenta as considerações finais.

Capítulo 2

REFERENCIAL BIBLIOGRÁFICO

Este capítulo apresenta os temas mais importantes que estão relacionados com o desenvolvimento deste trabalho. A seção 2.1 contextualiza as perdas de água no saneamento e os respectivos subtipos. Aborda também sobre as fraudes no consumo de água e as formas de combate. A seção 2.2 apresenta sobre as técnicas de *machine learning* que são comumente usadas para detecção de fraudes e, por fim, a seção 2.3 apresenta os trabalhos relacionados que foram avaliados.

2.1 PERDAS DE ÁGUA

As discussões sobre a situação atual e o futuro das águas em todo o mundo foram incentivadas pela progressiva deterioração dos rios e mananciais, além do agravamento de conflitos entre os diversos setores usuários das águas (GUMIER; LUVIZOTTO JUNIOR et al., 2007). No Brasil, o cenário do setor de saneamento é bastante desafiador devido aos altos índices de perdas de água tratada (IFC, 2013).

O Sistema Nacional de Informações sobre Saneamento (SNIS) reporta que no ano de 2018 a média nacional do Índice de Perdas na Distribuição (IPD) foi de 38,5% (BRASIL, 2020), enquanto que em países mais avançados, os níveis de perdas estão abaixo de 20% (BRASIL, 2020). De acordo com o Instituto Trata Brasil, essas perdas são equivalentes a R$ 10,5 bilhões em prejuízos financeiros (BRASIL, 2020). A recente aprovação do novo marco legal do saneamento básico (lei 14.026/2020) (BRASIL, 2020), exige das empresas de saneamento, metas para a redução das perdas.

2.1.1 Tipos de perdas

As perdas são caracterizadas pela diferença do volume de água produzido e do volume micromedido nos pontos de consumo. Podem ser classificadas em perdas reais ou perdas aparentes (KUSTERKO et al., 2018). De acordo com o SNIS, a distinção entre perdas aparentes e reais é importante, pois as ferramentas para a gestão e para o combate são diferentes para cada um dos tipos (BRASIL, 2019).

As perdas reais referem-se aos vazamentos, enquanto que as perdas aparentes se dão por problemas de gestão (e.g. falta de manutenção preventiva), comerciais (e.g. falhas no cadastro), além de fraudes de usuários e erros de medição (KUSTERKO et al., 2018).

O volume de perdas de um sistema de abastecimento de água é um fator chave na avaliação da eficiência de um operador de saneamento (BRASIL, 2020). O elevado índice de perdas implica na redução do faturamento, impactando na diminuição da capacidade das empresas em realizar investimentos para expansão e melhorias nos serviços, além dos danos ao meio ambiente pela exploração dos mananciais (IFC, 2013).

Andrade Sobrinho e Borja (2016) afirmam que as perdas de água nos sistemas de abastecimento público geram desperdício dos recursos públicos, o que normalmente é repassado para os consumidores através de aumentos nas tarifas. Portanto, a redução dos índices de perdas deve ser o principal desafio das companhias de saneamento, uma vez que implicam, além da captação de um volume hídrico acima do previsto, no consequente aumento dos custos com insumos e mão de obra para operação do sistema (GUMIER; LUVIZOTTO JUNIOR et al., 2007).

Kusterko et al. (2018) afirmam também que a redução das perdas é estratégica para o processo de tomada de decisão, a fim de garantir a sustentabilidade e, até mesmo, a competitividade da companhia diante da concorrência.

2.1.2 Perdas Aparentes

De acordo com a AESBE (2015), as perdas aparentes são decorrentes de submedição dos hidrômetros, erros no tratamento de dados e consumo não-autorizado (fraudes).

As perdas aparentes têm impacto direto sobre a receita das empresas, tendo-se em vista que elas equivalem a volumes produzidos e consumidos, mas não faturados (BRASIL, 2020).

Tardelli Filho (2016) destaca as principais ações para o combate às perdas aparentes: Substituição periódica dos hidrômetros (preventiva) e imediata dos hidrômetros quebrados (corretiva); Combate às fraudes, a partir de denúncias, análises de variações atípicas de consumo ou quaisquer outros indícios ou evidências; Além do aprimoramento da gestão comercial das companhias (cadastros e sistemas comerciais).

2.1.3 Fraudes

Segundo De Castro Fettermann et al. (2015), as fraudes são decorrentes de intervenções improvisadas nos hidrômetros, para que sejam medidas apenas uma parcela da água consumida. De acordo com a AESBE (2015), o consumo não-autorizado, ou fraude, é o volume de água furtado pelo usuário de algum modo, por meio de ligações clandestinas, ligações diretas (*by-pass*) e violações no medidor. Queiroga (2005) afirma que as fraudes na utilização de redes de abastecimento de água e de energia elétrica são similares, pois acontecem na adulteração dos dispositivos de medição, ou da conexão direta na rede de distribuição.

As fraudes resultam em prejuízos financeiros para as empresas de saneamento em razão do volume de água não faturado (DE CASTRO FETTERMANN et al., 2015). No ano 2000, as fraudes ocorridas no município de Campinas (SP) contribuíram em 5% dos

26,6% de perdas na distribuição (PASSINI; TOLEDO, 2002).

As empresas de abastecimento de água enfrentam dificuldades para localizar os pontos onde ocorrem fraudes e vazamentos. A dificuldade de detecção de fraudes acontece em razão das diversas fontes de variabilidade existente no padrão de consumo, tais como alteração na quantidade de moradores, viagens, condições climáticas, vazamentos, entre outros (DE CASTRO FETTERMANN et al., 2015).

De acordo com Queiroga (2005), a natureza mutável das fraudes é um dos maiores desafios para o saneamento. Quando algum tipo de fraude é descoberto e combatido, novos tipos de fraudes são então desenvolvidos, de maneira que sempre existem casos ainda desconhecidos.

Gumier, Luvizotto Junior et al. (2007) afirmam que o principal método de detecção utilizado é a inspeção percorrendo toda a extensão das tubulações, utilizando equipamentos acústicos em busca dos pontos de vazamento. Mesmo quando existe a tentativa de direcionar a inspeção, a seleção das áreas é realizada empiricamente, dificultando a localização das fraudes.

Em muitas companhias, métodos subjetivos e com poucos fundamentos técnicos são usados para estimar os volumes fraudados, muitas vezes baseados apenas na opinião dos operadores e sem ter feito qualquer cálculo ou obedecido a algum critério minimamente razoável (AESBE, 2015).

A Figura 2.1 apresenta resumidamente o problema da perda de água decorrente de fraudes, os impactos, a classificação e as dificuldades que foram descritas nesta seção.

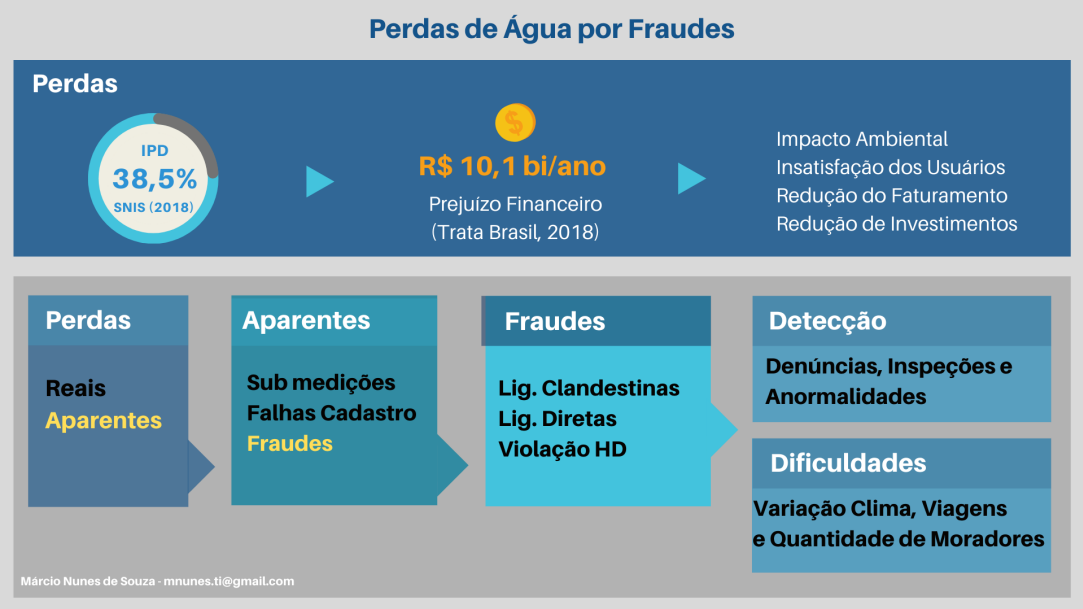

Figura 2.1 Problema: Perdas de Água por Fraudes.

2.1.4 Detecção de Fraudes

Assim como em outras áreas de negócios, muitos métodos para detecção de fraudes podem ser aplicados também no setor de saneamento. De Castro Fettermann et al. (2015) citam como possíveis métodos para o combate às fraudes no consumo de água: o Controle Estatístico de Processo (CEP), que procura identificar potenciais problemas (vazamento/fraude); Técnicas de 'previsão de demanda' a partir de procedimentos fundamentados em modelos estatísticos, matemáticos, econométricos ou subjetivos; e metodologias para a identificação de dados atípicos (*outliers*).

Fernandes (2014) sugere que a implementação de Distritos de Medição e Controle (DMC) é um importante recurso no combate aos consumos fraudulentos. Para o autor, em um determinado DMC com maior tendência de fraudes, será possível identificar as fraudes cometidas pelos usuários ao comparar os volumes faturados com os volumes verdadeiramente consumidos.

A estatística e *machine learning* fornecem tecnologias eficazes para detecção de fraude e foram aplicados com sucesso para detectar atividades como lavagem de dinheiro, fraude de cartão de crédito, fraudes em comércio eletrônico, fraude de telecomunicações e invasão de computadores (BOLTON; HAND, 2002) (BAESENS; VLASSELAER; VERBEKE, 2015).

Fawcett e Provost (1996) sugerem a verificação de alterações no comportamento dos usuários como um dos métodos para detectar fraudes. Os autores defendem também que um sistema tenha capacidade de aprender automaticamente as regras necessárias para a detecção das fraudes.

De acordo com Bolton e Hand (2002), os métodos estatísticos de detecção de fraudes podem ser supervisionados ou não supervisionados. Nos métodos supervisionados, amostras de registros fraudulentos e não fraudulentos são usadas para construir modelos que permitem classificar novas observações a uma das duas classes. Por outro lado, os métodos não supervisionados buscam simplesmente os registros que são muito diferentes do normal.

Segundo Phua et al. (2010), as abordagens mais comuns de mineração de dados para a detecção de fraudes são o uso único de algoritmos supervisionados ou, para melhores resultados, empregar modelos híbridos, composto pelo uso de múltiplos algoritmos supervisionados, ou algoritmos supervisionados e não supervisionados conjuntamente.

A detecção de fraudes por meio de algoritmos de *machine learning* é amplamente usada em diversas áreas de negócios, porém ainda é rara no setor de saneamento. Dentre os trabalhos relacionados que foram identificados durante a realização deste estudo, poucos utilizaram *machine learning* para tratar o problema. A seguir está descrito um resumo de tais trabalhos, enquanto que a seção 2.3 apresenta os trabalhos com mais detalhes.

Passini e Toledo (2002) realizaram um projeto piloto para detecção de fraudes no consumo de água no município de Campinas (São Paulo, Brasil). Foram desenvolvidos dois modelos baseados em agrupamento neural buscando conhecer os perfis de clientes fraudadores e um modelo de classificação por árvore de decisão para predizer o tipo de fraude cometido, que obteve taxa de acerto (recall) de 58% e taxa de erro de 42%.

Nos testes práticos, com investigação de 47 consumidores em campo, não foi possível constatar fraudes em nenhum dos casos analisados. Nota-se portanto, que o modelo apresentou baixa eficácia.

De Castro Fettermann et al. (2015) apresentam um processo para detecção de fraudes em consumo de água e desenvolveram dois modelos de *machine learning* com os algoritmos de detecção de *outliers* Z-Score e o Z-Score modificado. Foram utilizados dados referentes a 13 meses de consumo de 55 clientes residenciais, sendo 5 com fraudes, da cidade de Jequié (Bahia, Brasil). O modelo com o algoritmo Z-Score identificou quatro dos cinco consumidores com fraudes confirmadas, enquanto que com o algoritmo Z-Score modificado, todos os consumidores com fraudes confirmadas foram identificados. Apesar da alta taxa de acerto alcançada, destaca-se que o volume de dados utilizado foi extremamente limitado e não traz segurança em termos de generalização do modelo preditivo.

Por sua vez, Humaid e Barhoum (2013) desenvolveram um modelo de classificação com base na técnica de indução de regras para detectar fraudes em consumo de água de 67 áreas da cidade de Gaza (Palestina). O modelo apresentado apenas indica a zona ou área com maior probabilidade de ocorrências de fraudes de acordo com as estações do ano.

Monedero et al. (2015) desenvolveram três modelos para identificação de fraudes em consumo de água usando os dados (357.920 registros) de clientes da região de Sevilha (Espanha). Cada modelo foi criado propondo solução para um tipo de anormalidade de consumo: quedas progressivas de consumo, quedas repentinas de consumo e consumo anormalmente baixo. Os modelos foram avaliados com uma fração dos dados (35.147 registros). Nota-se que não houve uma base independente para a realização das avaliações dos modelos. Por fim, foram realizadas análises dos resultados e inpeções em campo, sendo identificados entre 2% a 10% dos clientes com suspeita de fraudes. Apesar da baixa eficácia, a metodologia trouxe ganhos para a empresa pois a mesma realizava as inspeções sem critérios definidos.

Al-Radaideh e Al-Zoubi (2018) propuseram o uso de classificação supervisionada com os algoritmos SVM e KNN para detectar clientes suspeitos de fraude no consumo de água, com dados de consumidores da cidade de Irbid (Jordânia). Os modelos foram desenvolvidos usando apenas as configurações padrões dos algoritmos. Os experimentos demonstraram que o modelo com o algoritmo SVM obteve acurácia de 72,4%, enquanto que o modelo com o algoritmo KNN obteve acurácia de 74,3%. Apesar de o algoritmo KNN obter resultados levemente superiores ao algoritmo SVM, observa-se que ambos os classificadores apresentaram desempenho próximo na detecção de fraudes. Segundo os autores, os resultados obtidos demonstram que os modelos podem aumentar a produtividade das equipes de inspeção, reduzindo os custos e as perdas aparentes de água.

Diante do exposto, verifica-se a existência de poucas estudos para detectar fraudes no saneamento com técnicas de *data mining* e *machine learning*, apesar de serem amplamente utilizados em muitos outros tipos de negócios. De acordo com Queiroga (2005), as técnicas de *data mining* aliadas à inteligência computacional, estão aptas a lidarem com grandes volumes de dados que seriam inviáveis se fossem avaliados por pessoas, por maior que fosse a equipe disponível.

Considerando os impactos causados pelas fraudes no sistema de fornecimento de água,

a dificuldade de identificação precisa das ocorrências, a baixa eficácia dos métodos existentes e as limitações de escala de dados e rigor científico dos estudos anteriores, é imprescindível que novas pequisas sejam realizadas para o desenvolvimento de soluções robustas e confiáveis. Assim, este trabalho realiza o desenvolvimento e validação experimental rigorosa de modelos de *machine learning* para a detecção de fraudes no consumo de água, visando alcançar poder de generalização em um cenário de larga escala e viabilizar sua aplicação prática.

Na seção a seguir, serão apresentados os principais conceitos relacionados a *machine learning* que foram utilizados no desenvolvimento deste trabalho.

2.2 MACHINE LEARNING (ML)

Machine Learning (Aprendizado de Máquina), pode ser definido como métodos computacionais que usam a experiência para melhorar a eficiência e fazer previsões precisas (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018). Batista et al. (2003) afirmam que *machine learning* (ML) é uma importante subárea de pesquisa da Inteligência Artificial.

Jordan e Mitchell (2015) destacam que *machine learning* é um dos campos técnicos que mais cresce atualmente, na intersecção da ciência da computação e estatística e no núcleo da inteligência artificial e da ciência de dados. Para os autores, *machine learning* aborda a questão de como construir programas de computadores que melhoram automaticamente com a experiência. Como o sucesso de um algoritmo de ML depende dos dados usados, o aprendizado de máquina está inerentemente relacionado à análise de dados e estatísticas (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018).

De acordo com Buczak e Guven (2015), pode haver uma confusão sobre o entendimento dos termos *Machine Learning* (ML), *Data Mining* (DM) e *Knowledge Discovery in Databases* (KDD) dentro da Ciência de Dados. KDD é um processo não trivial, interativo e iterativo, que busca identificar novos padrões de dados válidos, úteis e compreensíveis a partir de grandes conjuntos de dados (FAYYAD et al., 1996). Segundo Buczak e Guven (2015), o KDD é um processo completo que lida com a extração de informações úteis e previamente desconhecidas dos dados. *Data Mining* é a principal atividade do processo KDD e compreende a busca efetiva por conhecimentos úteis com a aplicação de algoritmos sobre os dados (GOLDSCHMIDT; PASSOS, 2005). Para Buczak e Guven (2015), o DM é usado para descrever uma etapa específica do KDD que lida com a aplicação de algoritmos específicos para extrair padrões de dados.

Existe uma sobreposição significativa entre *machine learning* e *data mining*, pois geralmente empregam os mesmos métodos (BUCZAK; GUVEN, 2015). Para os autores, o ML concentra-se na classificação e previsão, com base nas propriedades conhecidas previamente aprendidas com os dados de treinamento. Enquanto que *data mining* se concentra na descoberta de propriedades anteriormente desconhecidas nos dados.

2.2.1 Classificação

Classificação e Regressão estão entre as tarefas mais populares de *machine learning*. A classificação trata do problema de atribuir uma categoria a cada item, enquanto que

regressão trata de prever um valor real para cada item (MOHRI; ROSTAMIZADEH; TALWALKAR, 2018).

De acordo com Sun, Wong e Kamel (2009), um modelo de classificação é construído para prever os rótulos da classe para os dados de entrada desconhecidos, revelando o relacionamento entre o conjunto de atributos e o rótulo da classe. Camilo e Silva (2009) destacam que os modelos de classificação analisam o conjunto de registros fornecidos e, com o objetivo de identificar a qual classe um determinado registro pertence, os modelos aprendem a classificar os novos registros.

Tradicionalmente, os métodos de mineração de dados são classificados em aprendizado supervisionado (preditivo) e não-supervisionado (descritivo) (CAMILO; SILVA, 2009). Quando as instâncias são dadas com rótulos conhecidos, então o aprendizado é chamado de supervisionado, em contraste com o aprendizado não-supervisionado, onde as instâncias não são rotuladas (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007). Para Camilo e Silva (2009), a diferença entre os dois tipos está no fato de que os métodos supervisionados precisam de uma pré-categorização para os registros.

O objetivo da aprendizagem supervisionada é construir um modelo conciso de aplicação de rótulos de classe de acordo com características do preditor. O classificador resultante é então usado para atribuir os rótulos de classe às instâncias onde os valores das características do preditor são conhecidas, mas o valor do rótulo de classe é desconhecido (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007).

2.2.2 Pré-processamento

Dentre os fatores que impactam o sucesso do aprendizado de máquina, a representação e qualidade dos dados estão em primeiro lugar. Informações irrelevantes, redundantes, dados ruidosos e não confiáveis tornam a descoberta de conhecimento mais difícil (KOTSIANTIS; KANELLOPOULOS; PINTELAS, 2006). De acordo com os autores, o pré-processamento de dados pode ter um impacto significativo no desempenho de generalização de um algoritmo de aprendizagem supervisionado.

O pré-processamento pode ser aplicado para que a massa original de dados seja reduzida, mas mantendo a representatividade dos dados originais, permitindo que os algoritmos sejam executados com mais eficiência, mas mantendo a qualidade do resultado (CAMILO; SILVA, 2009). Também podem ser aplicadas com o objetivo de conhecer mais a respeito dos dados, solucionar problemas existentes e prepará-los para a fase seguinte no processo KDD (BATISTA et al., 2003).

A fase de pré-processamento pode incluir atividades de limpeza de dados, normalização, transformação, extração de recursos e seleção de dados (KOTSIANTIS; KANELLOPOULOS; PINTELAS, 2006).

2.2.3 Algoritmos de Aprendizagem Supervisionada

Uma grande variedade de algoritmos de *machine learning* foi aplicada com sucesso para classificação em muitos domínios de negócios. Árvores de decisão, redes bayesianas, redes neurais, *Support Vector Machine* (SVM), *naive bayes*, *logistic regression* e regras de associação são exemplos mais comuns destes algoritmos (SUN; WONG; KAMEL, 2009)(BA-

ESENS; VLASSELAER; VERBEKE, 2015)(KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007).

Frequentemente a adoção de algoritmos que combinam diversos classificadores tem obtido desempenho melhor do que o uso individual dos algoritmos tradicionais. O objetivo da integração do resultado dos algoritmos de classificação é gerar resultados mais precisos (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007). Devido a estes melhores resultados, ultimamente tem havido muito interesse na aprendizado por conjuntos (*ensemble learning*), que são os métodos que geram muitos classificadores e agregam seus resultados (LIAW; WIENER et al., 2002).

Bagging e *Boosting* são dois conhecidos métodos de *ensemble learning*. No *Bagging*, são usadas árvores sucessivas independentes, onde, no final, um simples voto da maioria é usada para previsão. Enquanto que no *Boosting*, árvores sucessivas e dinâmicas dão peso extra a pontos previstos incorretamente pelo preditor anterior. No final, um voto ponderado é levado para a previsão (LIAW; WIENER et al., 2002). *Random Forest* é um exemplo de algoritmo do tipo *Bagging*, enquanto que *Gradient Boosting* é um algoritmo do tipo *Boosting*.

Para a identificação de fraudes, geralmente os modelos são desenvolvidos usando algoritmos podem ser utilizados de forma única ou composto pelo uso de múltiplos algoritmos conjuntamente (PHUA et al., 2010).

2.2.4 Avaliação e Seleção de modelos

O uso correto de técnicas de avaliação e seleção de modelos é vital para o aprendizado de máquina (RASCHKA, 2018). De acordo com Kotsiantis, Zaharakis e Pintelas (2007), a avaliação do classificador é frequentemente baseada na precisão da previsão, ou seja, da porcentagem de previsão correta dividido pelo número total de previsões.

Existem algumas técnicas para se obter a precisão de um classificador. O *Cross Validation* é amplamente utilizado para evitar problemas como viés pessimista e a variância, geralmente ocasionados pela limitação do volume de dados rotulados para realizar todas as etapas do desenvolvimento do modelo (RASCHKA, 2018).

De acordo com Kotsiantis, Zaharakis e Pintelas (2007), no *Cross Validation*, o conjunto de treinamento é dividido em 'k' subconjuntos mutuamente exclusivos e de tamanhos iguais. A cada iteração da validação cruzada, um subconjunto é separado para avaliação do modelo e o classificador é treinado na união de todas os outros subconjuntos.

Este procedimento resultará em 'k' modelos diferentes. Com isso, esses modelos foram ajustados para conjuntos de treinamento distintos e parcialmente sobrepostos. E foram avaliados em conjuntos de validação não sobrepostos (RASCHKA, 2018). A média da taxa de erro de cada subconjunto é, portanto, uma estimativa da taxa de erro do classificador (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007).

Apesar de ser uma técnica computacionalmente mais cara, o *Cross Validation* proporciona uma maior precisão para um classificador (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007).

Durante o processo de treinamento, diversas técnicas devem ser testadas e combinadas a fim de escolher o melhor modelo gerado (CAMILO; SILVA, 2009). Para realizar a seleção

do modelo, pode-se utilizar a técnica de validação cruzada em conjunto com a técnica de otimização de hiperparâmetros (RASCHKA, 2018).

De acordo com Raschka (2018), os hiperparâmetros são as características de ajuste de um algoritmo de aprendizado de máquina. Eles são usados para controlar o comportamento dos algoritmos, encontrando o equilíbrio certo entre viés e variância. Para cada configuração de hiperparâmetro, pode ser aplicado o método de validação cruzada no conjunto de treinamento, resultando em múltiplas modelos e estimativas de desempenho.

Os resultados destes modelos então são avaliados, para que seja possível escolher o modelo com melhor ajuste. Um método comum para comparar modelos de *machine learning* é realizar comparações estatísticas da precisão dos classificadores treinados (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2007).

2.3 TRABALHOS RELACIONADOS

Nesta seção, são apresentados alguns trabalhos com análises e propostas de solução para as fraudes no consumo de água.

Um projeto piloto de mineração de dados de fraudes no consumo de água no município de Campinas (São Paulo, Brasil) foi realizado por Passini e Toledo (2002). Foram desenvolvidos três modelos, sendo dois deles baseados em agrupamento neural buscando conhecer os perfis de clientes fraudadores e um modelo de classificação por árvore de decisão para predizer o tipo de fraude cometido.

Nos modelos de agrupamento, os consumidores foram agrupados de acordo com as características comuns (Status da ligação, Categoria, Idade da ligação, Parcelamentos, Cortes, Retificações e Média de consumo) e os respectivos tipos de fraudes identificadas previamente. Não foram detalhados o total de registros utilizados nesta etapa do experimento. Para o desenvolvimento do modelo de classificação, foram utilizados 7.184 registros de consumidores com fraudes, sendo utilizado 80% (5.765) dos registros para treinamento. A partir do modelo treinado, os testes foram realizados com 20% (1.419) dos registros, acertando 58% (823 registros) e taxa de erro de 42% (596 registros).

Na fase final do experimento, a partir dos três modelos desenvolvidos, foram selecionados 47 consumidores sem histórico de fraudes para a realização da investigação em campo. Não foi possível constatar fraudes em nenhum dos casos analisados. Com base nestes resultados, constatou-se que o modelo não estava bom, pois apresentou alta taxa de erro e não foi possível identificar nenhum registro de fraude com a investigação em campo. Para trabalhos futuros, foi sugerido que modelos de padrões sequenciais ou associação possam ser testados, além de novos estudos com modelos baseados em classificação por árvore de decisão considerando o histórico de visitas em campo para classificação de fraudes.

Humaid e Barhoum (2013) desenvolveram um modelo de classificação para detectar fraudes em consumo de água com base na técnica de indução de regras. O modelo proposto permite que um interessado nos dados navegue pelas regras para obter *insights* no domínio dos dados. Foram utilizados dados de 34.000 consumidores de 67 áreas da cidade de Gaza (Palestina).

De acordo com os autores, com a utilização do modelo, foi possível identificar as áreas

da cidade com maior probabilidade de fraudes de acordo com as estações do ano. Com isso, ressaltam a importância do uso dos métodos de mineração de dados como a indução de regras em detrimento das regras heurísticas manuais, que não conseguem detectar todas as fraudes de consumo de água, especialmente ao lidar com grandes bancos de dados. Os autores sugerem a construção de um novo modelo considerando o perfil dos clientes individualmente para tornar a detecção de fraudes mais realista.

A implementação de Distritos de Medição e Controle (DMC) para melhor gerenciar as perdas de águas em determinadas regiões com maior incidência de fraudes na cidade de Gaia (Portugal) foi proposta por Fernandes (2014). Para a implantação de DMC é necessário um investimento alto em infra-estrutura para permitir o gerenciamento, muitas vezes em tempo real, dos volumes distribuídos e consumidos pelos clientes.

De Castro Fettermann et al. (2015) propuseram um processo para detecção de fraudes em consumo de água composto por cinco etapas: Análise descritiva dos dados, Tratamento dos dados, Seleção do método de detecção de fraudes, Aplicação do método selecionado e Análise dos resultados.

Para demonstrar a aplicação da proposta, os autores obtiveram os dados referentes a 13 meses de consumo de 55 clientes residenciais, sendo 5 com fraudes, da cidade de Jequié (Bahia, Brazil). A análise descritiva dos dados apresentou grande variabilidade do consumo. Para reduzir essa variabilidade, foi realizado um tratamento com o uso de análise de *clusters* hierárquicos, sendo o conjunto de dados subdividido em dois grupos por volume consumido. Em seguida, os algoritmos de detecção de *outliers* Z-Score e o Z-Score modificado foram selecionados e aplicados ao conjunto de dados. Foram considerados como fraude os casos com identificação de *outliers* por, pelo menos, três períodos consecutivos.

Analisando os resultados, o modelo com o algoritmo Z-Score identificou quatro dos cinco consumidores com fraudes confirmadas, enquanto que com o algoritmo Z-Score modificado, todos os consumidores com fraudes confirmadas foram identificados. Também foram identificados dois consumidores com possíveis fraudes em ambos os modelos. Os resultados indicam uma maior taxa de acerto do modelo com o algoritmo Z-score modificado. Apesar da alta taxa de acerto alcançada, destaca-se que o volume de dados utilizado foi extremamente limitado e não traz segurança em termos de generalização do processo preditivo. Para trabalhos futuros, os autores sugerem a utilização de algum critério de classificação dos consumidores para reduzir a variabilidade do consumo; a utilização de mais dados, permitindo a aplicação de séries temporais para previsão de demanda e identificação de fraudes; e a automatização do processo proposto através do uso de software.

Monedero et al. (2015) desenvolveram três modelos para identificação de fraudes em consumo de água usando um *dataset* com os dados de consumo trimestrais de 357.920 clientes da região de Sevilha (Espanha). Cada modelo foi criado propondo solução para um tipo de anormalidade de consumo: quedas progressivas de consumo, quedas repentinas de consumo e consumo anormalmente baixo.

O primeiro modelo foi desenvolvido usando o coeficiente de correlação de Pearson entre a quantidade de leituras e o consumo normalizado *per capita*. O segundo modelo foi desenvolvido com um conjunto de regras geradas pela análise do consumo normalizado,

comparando os valores em janelas anuais de tempo. O terceiro modelo foi desenvolvido com regras geradas pela análise de clientes com consumo anual abaixo de 10 m3 e a respectiva quantidade de pessoas residentes. Os modelos foram avaliados com uma fração dos dados (35.147 registros), sendo identificados entre 2% a 10% dos clientes com suspeita de fraudes. Nota-se que não houve uma base independente para a realização das avaliações dos modelos. A seleção dos clientes para inspeção será feita de acordo com os critérios dos modelos e da quantidade de clientes que a empresa deseja inspecionar.

Após o desenvolvimento dos modelos, foram escolhidos os dados de 859 clientes de uma região para a aplicação prática. Os algoritmos identificaram 85 (10%) clientes com suspeita de fraudes e desses, depois da realização das inspeções, foram identificados 6 (7%) casos de fraudes. Com o objetivo de identificar as áreas geográficas com maiores probabilidades de fraudes, um segundo teste com os dados de 1.164 clientes foi realizado, analisando as áreas geográficas com a ferramenta Google Earth. A utilização dos algoritmos neste *dataset* detectou 334 (28,7%) clientes que foram selecionados para serem inspecionados, sendo 30 (9%) clientes confirmados com fraudes, demonstrando que o uso de informações geográficas contribuem para a identificação de fraudes. Os autores concluem que, apesar da aparente baixa eficácia, a metodologia trouxe ganhos para a empresa pois a mesma realizava as inspeções sem critérios definidos.

Detroz e Silva (2017) propuseram um sistema de visão computacional para detecção automatizada de irregularidades em hidrômetros através do uso de técnicas de reconhecimento de padrões. De acordo com os autores, o sistema será capaz de identificar alguns tipos de irregularidades através da análise de imagens. O *framework* proposto atingiu uma acurácia média de 81,29%, concluindo que o uso de técnicas de visão computacional é uma estratégia promissora e tem potencial para beneficiar a análise de detecção de fraudes no consumo de água.

Morote e Hernández-Hernández (2018) realizaram um estudo com o objetivo de analisar a evolução e os fatores determinantes de fraudes no consumo de água em domicílios da cidade de Alicante (Espanha). Foram realizadas análises sobre os casos de fraudes identificados no período compreendido entre 2005-2017. De acordo com os autores, após as análises dos dados, foi possível identificar que a maior parte das fraudes ocorreram em imóveis compactos (83%) localizados no Distrito Norte da cidade (70%). Foi constatado também uma grande incidência de fraudes na zona rural. A inter-relação de uma série de fatores de natureza econômica (renda per capita, crise econômica e aumento do preço da água) explica a marcante concentração dos casos de fraudes na cidade de Alicante.

Al-Radaideh e Al-Zoubi (2018) propuseram o uso de classificação supervisionada com os algoritmos SVM e KNN para detectar clientes suspeitos de fraude no consumo de água. Para a realização do trabalho, foram obtidos dados de 90 mil consumidores da cidade de Irbid (Jordânia).

Após a análise exploratória dos dados, foram realizados filtros para eliminar redundâncias e dados irrelevantes, restando 16.761 registros, sendo 16.114 (96%) registros de consumidores sem fraudes e 647 (4%) registros de consumidores com fraudes. Por fim, foi aplicada a técnica de *undersampling* para balancear o conjunto de dados, permanecendo 1.294 registros, sendo 647 (50%) registros de fraudes e mais 647 (50%) registros aleatórios de consumidores sem fraudes.

Os experimentos foram executados apenas com as configurações padrões dos algoritmos. O primeiro experimento foi realizado com a validação cruzada com 10 *folds*. Os resultados demonstram que os algoritmos SVM e KNN obtiveram acurácia de 71% e 70% respectivamente. Os registros de fraudes classificadores corretamente correspondem a 61% com o SVM e 68% com o KNN. No segundo experimento, os dados foram particionados em treino (75%) e testes (25%). Os resultados demonstram que os algoritmos SVM e KNN obtiveram acurácia de 72,4% e 74,3% respectivamente. Os registros de fraudes classificados corretamente correspondem a 68% com o SVM e 73% com o KNN.

Apesar de o algoritmo KNN obter resultados levemente superiores ao algoritmo SVM, observa-se que ambos os classificadores apresentaram desempenho próximo na detecção de fraudes. Segundo os autores, os resultados obtidos demonstram que os modelos podem aumentar a produtividade das equipes de inspeção, reduzindo os custos e as perdas aparentes de água.

Uddin et al. (2019), SRIRAMULU et al. (2020), GOPAL e BALAJI (2020), SREEDEVI e SWATHI (2021) desenvolveram conjuntamente um estudo com o objetivo de apresentar um novo modelo para detecção de Perdas Não-Técnicas (NTL) no consumo de água usando técnicas de mineração de dados. Foram utilizadas as técnicas de mineração *Support Vector Machines* (SVM) e *K-Nearest Neighbor* (KNN).

Para identificar o perfil dos clientes fraudadores, foram realizadas tarefas de limpeza e preparação dos dados, restando 16.114 registros de clientes sem fraudes e 647 registros de clientes fraudulentos. Em seguida, os dados foram divididos em duas partes: 80% para o treinamento e 20% para os testes do modelo.

Os modelos foram treinados usando os parâmetros padrões dos classificadores, enquanto que a avaliação dos resultados foi realizada com base nas métricas obtidas por matriz de confusão. De acordo com os autores, a precisão dos modelos gerados atingiram uma taxa de mais de 74%, que é um resultado melhor do que a previsão manual realizada atualmente.

Apesar do referido estudo estar publicado em diversos artigos científicos, os autores não disponibilizaram detalhes do processo de construção dos modelos. Nota-se que foram utilizadas apenas as configurações padrões dos algoritmos e um método simples *(Hold-out)* para avaliação dos resultados. De acordo com (RASCHKA, 2018), o método *Hold-out* para avaliação de modelos pode ocasionar em problemas de viés pessimista e a variância nos resultados.

Espinosa, Gisselot e Arriagada (2020) realizaram um estudo com o objetivo de prever o consumo fraudulento de água potável através de técnicas de mineração de dados. De acordo com os autores, as perdas aparentes correspondem a um terço da água potável produzida no Chile. As ligações clandestinas, por sua vez, explicam entre 8% a 10% destas perdas.

Para a realização do estudo, foram utilizados 23.005 registros, onde 12.250 correspondem ao consumo regular e 10.755 ao consumo fraudulento. Os autores não relataram se utilizaram alguma técnica de balanceamento dos dados. O conjunto de dados foi particionado usando a técnica *Hold-Out* para treinamento e avaliação do preditor. 70% dos registros foram separados aleatoriamente para a realização da fase de treinamento, enquanto que 30% foram separados para a realização dos testes. Em seguida foram

realizadas tarefas de limpeza, preparação dos dados e seleção de *features*.

Foram utilizadas cinco técnicas de aprendizado de máquina usando os algoritmos incorporados na biblioteca *Scikit-Learn*: *Decision Tree, Naive Bayes, Neural Net, Support Vector Machine (SVM) e K-Nearest Neighbors (KNN)*. Para otimizar o desempenho preditivo de cada algoritmo, os autores realizaram ajustes em parâmetros de cada algoritmo.

O melhor desempenho geral foi obtido com o algoritmo *Decision Tree*. A previsão específica da classe de fraude mostra um bom desempenho com *recall* acima de 77% e precisão acima de 88%. O modelo obteve Acurácia geral de 88.16%. De acordo com os autores, isso implica que no geral, o modelo identifica bem o consumo fraudulento.

Recentemente, Sreekanth e Thinakaran (2021) realizaram um estudo com objetivo de prever com precisão a fraude de água metropolitana. Eles desenvolveram dois modelos usando o algoritmo de Rede Neural Convolucional (CNN) e o algoritmo Rede Neural Recorrente (RNN). Os modelos foram implementados e testados em um conjunto de dados que consiste em 8.002 registros e 8 colunas. Os autores não apresentaram detalhes do processo de avaliação e treinamento dos modelos.

De acordo o estudo, os modelos obtiveram uma precisão média de 94,52% usando o algoritmo RNN e precisão de 93,49% usando o algoritmo de CNN. Após realizarem análises estatísticas dos resultados, os autores chegaram à conclusão que a significância estatística da Rede Neural Recorrente é alta. Nota-se que o estudo apresenta resultados promissores, apesar da falta de detalhes sobre o processo de treinamento e avaliação dos modelos.

Apesar de resultados promissores em muitos dos trabalhos citados, não foi possível estabelecer uma comparação direta com o presente estudo devido à indisponibilidade dos mesmos dados e de detalhes dos processos construtivos empregados.

Dentre os trabalhos apesentados, é possível perceber diversas abordagens para tratar a temática. O projeto de Fernandes (2014) propôs solução com base na melhoria da infra-estrutura de distribuição de água através da implantação de DMC e equipamentos de medição em tempo real. Esta é uma solução de difícil implementação devido o alto custo dos equipamentos. O trabalho de Detroz e Silva (2017) propõe um *framework* desenvolvido com técnicas de visão computacional para identificar irregularidades em hidrômetros a partir da análise de imagens. Apesar de bons resultados, este tipo de solução propõe solução apenas as fraudes ocorridas diretamente nos hidrômetros. No estudo realizado por Morote e Hernández-Hernández (2018), os autores identificaram através da análise dos dados, os perfis de consumidores e localização da maioria das ocorrências de fraudes. Também identificaram que fatores sócio-econômicos explicam a marcante concentração dos casos de fraudes.

Os trabalhos avaliados com propostas de solução usando mineração de dados e *machine learning* demonstraram a potencialidade destas técnicas para solucionar o problema das fraudes no saneamento. Porém, alguns destes estudos apresentaram resultados insatisfatórios ou problemas metodológicos (e.g. baixo volume de dados, dados para avaliação usados durante o treinamento do modelo, configurações dos algoritmos ou método simples de avaliação baseado apenas em treino/teste). Estes problemas podem comprometer os resultados dos modelos na prática.

O trabalho aqui apresentado busca trazer a análise para um patamar superior em

termos da quantidade de registros, baseando-se em dados das duas maiores cidades do Estado da Bahia (Brasil), que representam uma população estimada de 3.506.307 habitantes (IBGE, 2020). Além disso, neste trabalho objetiva-se identificar as fraudes individualmente, ao invés da solução por zona ou área proposta por Humaid e Barhoum (2013). Por fim, otimizações de algoritmos e validação rigorosa foi realizada em busca de resultados mais confiáveis e generalizáveis do que os apresentados em Passini e Toledo (2002), Monedero et al. (2015), Al-Radaideh e Al-Zoubi (2018), Uddin et al. (2019), SRIRAMULU et al. (2020), GOPAL e BALAJI (2020), SREEDEVI e SWATHI (2021), Espinosa, Gisselot e Arriagada (2020), Sreekanth e Thinakaran (2021).

Portanto, este estudo objetivou o desenvolvimento e validação experimental de um modelo de *machine learning* para detecção de fraudes no consumo de água utilizando algoritmos de aprendizagem supervisionada.

Capítulo

3

MODELO DE DETECÇÃO DE FRAUDES EM CONSUMO DE ÁGUA

O combate às fraudes por meio de *machine learning* é uma inovação no saneamento, com o objetivo de proporcionar a redução de perdas aparentes em sistemas de abastecimento de água. Este capítulo apresenta o processo realizado com o objetivo de construir um modelo de detecção de fraudes em consumo de água. Para isso, foi conduzido um estudo experimental que avaliou diferentes modelos de aprendizado de máquina supervisionado para a detecção de fraudes cometidas por clientes de uma companhia pública de saneamento.

Dentre as diversas técnicas de aprendizado de máquina existentes, o aprendizado supervisionado foi escolhido para realizar este trabalho devido à grande quantidade de registros de fraudes existentes para treinar o modelo. Esta técnica também é utilizada em diversas soluções para identificar fraudes em outros tipos de negócios.

A Seção 3.1 apresenta, em detalhes, os materiais e método utilizados para o desenvolvimento deste estudo. Ao final, na Seção 3.2, são apresentados os resultados e discussão.

3.1 MATERIAIS E MÉTODO

Esta seção apresenta o processo usado para desenvolver o modelo de *machine learning* para detecção de fraudes no saneamento. A Figura 3.1 apresenta as etapas deste processo. Resumidamente temos que os dados foram obtidos e analisados, em seguida, foram realizadas as etapas de pré-processamento e transformação dos dados, otimização de parâmetros e validação cruzada, treinamento e testes do modelo final. Por fim, foram realizadas tarefas de análise dos resultados.

Figura 3.1 Big Figure - processo de desenvolvimento do modelo analítico.

Para a execução do processo, foi construído um *workflow*, apresentado na Figura 3.2, que tem como base o processo de descoberta de conhecimento em bancos de dados – do inglês, *Knowledge Discovery in Databases (KDD)* – para projetos de mineração de dados (FAYYAD et al., 1996). Todas as atividades foram realizadas usando a Knime Analytics Platform (Knime), que é uma ferramenta robusta de código aberto para mineração de dados e aprendizado de máquina, desenvolvido na Universidade de Konstanz na Alemanha e mantido pela KNIME AG (KNIME, 2020).

Em cada fase do *workflow* são realizadas tarefas para preparar os dados, treinar, otimizar e avaliar os modelos. Resumidamente, temos que na fase "Seleção dos dados" (Seção 3.1.1), os dados a serem utilizados no processo foram selecionados a partir do conjunto de dados original. Na fase "Pré-processamento e Transformação" (Seção 3.1.2), foram aplicadas técnicas para preparação dos dados, amostragem e particionamento. Em seguida, na fase "Modelagem" (Seção 3.1.3), foram realizadas as atividades para treinamento e otimização dos algoritmos de aprendizagem supervisionada. Por fim, na fase "Avaliação" (Seção 3.1.4), os resultados dos algoritmos foram avaliados para subsidiar a escolha daquele a ser considerado para uma possível implantação e utilização prática. Cada uma destas fases são descritas em detalhes nas seções a seguir.

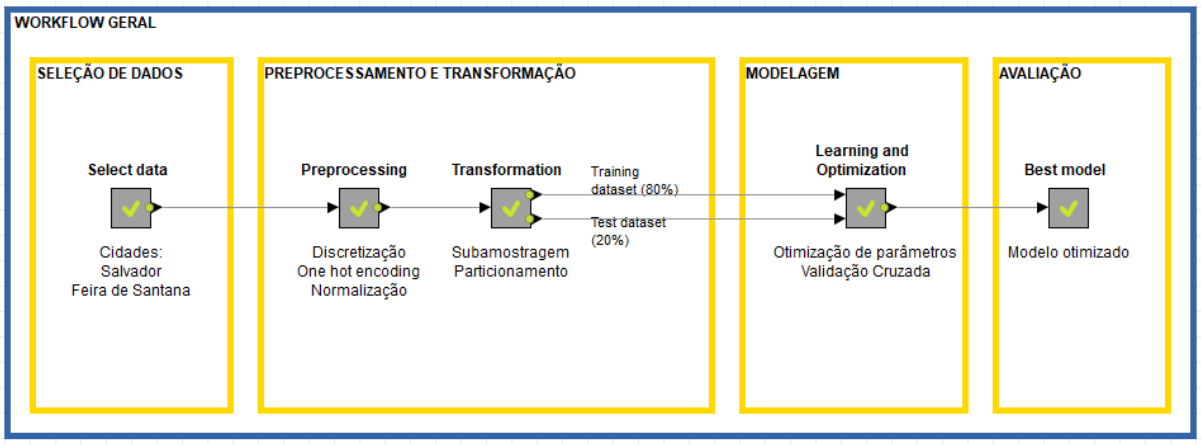

Figura 3.2 *Workflow* para construção e avaliação dos modelos preditivos.

3.1.1 Seleção dos dados

O conjunto de dados original usado para treinar os modelos de *machine learning* foi obtido por meio da Empresa Baiana de Águas e Saneamento S/A (Embasa) (EMBASA, 2020). Os dados dizem respeito ao cadastro de clientes (4.390.219 matrículas) e seus registros de consumo de água do ano de 2018.

A análise exploratória foi realizada com o objetivo de conhecer os dados dos clientes e auxiliar na escolha das variáveis que serão usadas para realizar o treinamento dos modelos na continuidade do trabalho. A Tabela 3.1 apresenta a descrição das variáveis selecionadas na base de dados.

A Figura 3.3 apresenta uma visão com as correlações existentes entre estas variáveis selecionadas. Observa-se que as fraudes possuem correlação positiva com as varáveis relacionadas a cortes por falta de pagamento (CorteFaltaPgto), cortes por infiltração (CorteInfiltracao), e ligação inativa (5-LigacaoInativa). Há também correlação negativa com ligações ativas (3-LigacaoLigada). Outros campos não possuem correlações relevantes.

Analisando os dados, foram identificados 56.265 clientes com fraudes, distribuídas nos municípios atendidos pela Embasa, conforme apresentada na Figura 3.4. O município de Salvador representa 39,87% dos registros de fraudes identificadas, seguida por Feira de Santana (13,45%), Camaçari (5,92%), Barreiras (4,58%), Candeias (4,26%), Dias D´ávila (3,20%), Lauro de Freitas (3,10%) e Simões Filho (3,02%). A soma dos demais municípios equivale a 22,60% dos casos de fraudes identificas.

Em seguida, foram selecionadas as matrículas das cidades de Salvador e Feira de Santana. Além de serem as duas maiores cidades da Bahia, com população estimada de 2.886.698 e 619.609 habitantes respectivamente (IBGE, 2020), elas são as cidades que possuem as maiores quantidades de fraudes identificadas, representando 53% do total de fraudes registradas em 2018. A Embasa também possui contratos com empresas terceirizadas para a inspeção de fraudes nestas duas cidades. Estes contratos são usados para contratar equipes para realizar as verificações em campo das matrículas com suspeitas de fraudes.

Variável	**Tipo**	**Descrição**
CortePedido	Numérico	Quantidade de cortes a pedido do cliente
CorteFaltaPgto	Numérico	Quantidade de cortes por inadimplência
CorteInfiltracao	Numérico	Quantidade de cortes por infiltração
SupressaoPedido	Numérico	Quantidade de supressão a pedido do cliente
SupressaoFaltaPgto	Numérico	Quantidade de supressão por inadimplência
SupressaoInfiltracao	Numérico	Quantidade de supressão por infiltração
PontosUtilizacao	Numérico	Quantidade de pontos de utilização no imóvel
AreaConstruida	Numérico	Tamanho da área construída
ConsumoAnual	Numérico	Volume de consumo por ano
ConsumoMinimo	Numérico	Volume mínimo de consumo
ConsumoMaximo	Numérico	Volume máximo de consumo
ConsumoMediana	Numérico	Mediana do volume consumido
ConsumoMedia	Numérico	Média do volume consumido
SituacaoImovel	Categórico	Status da situação do imóvel (Habitado, Desabitado, Em construção)
SituacaoAgua	Categórico	Status da ligação de água (Ligada, Cortada, Suprimida)
TipoPessoa	Categórico	Tipo de pessoa (Pessoa física, Pessoa jurídica, Não Definida)
SituacaoEsgoto	Categórico	Status da ligação de esgoto (Ligada, Suprimida, Potencial)
StatusFraude	Alvo	Status de fraude

Tabela 3.1 Tabela de detalhes das variáveis que compõem o *dataset*.

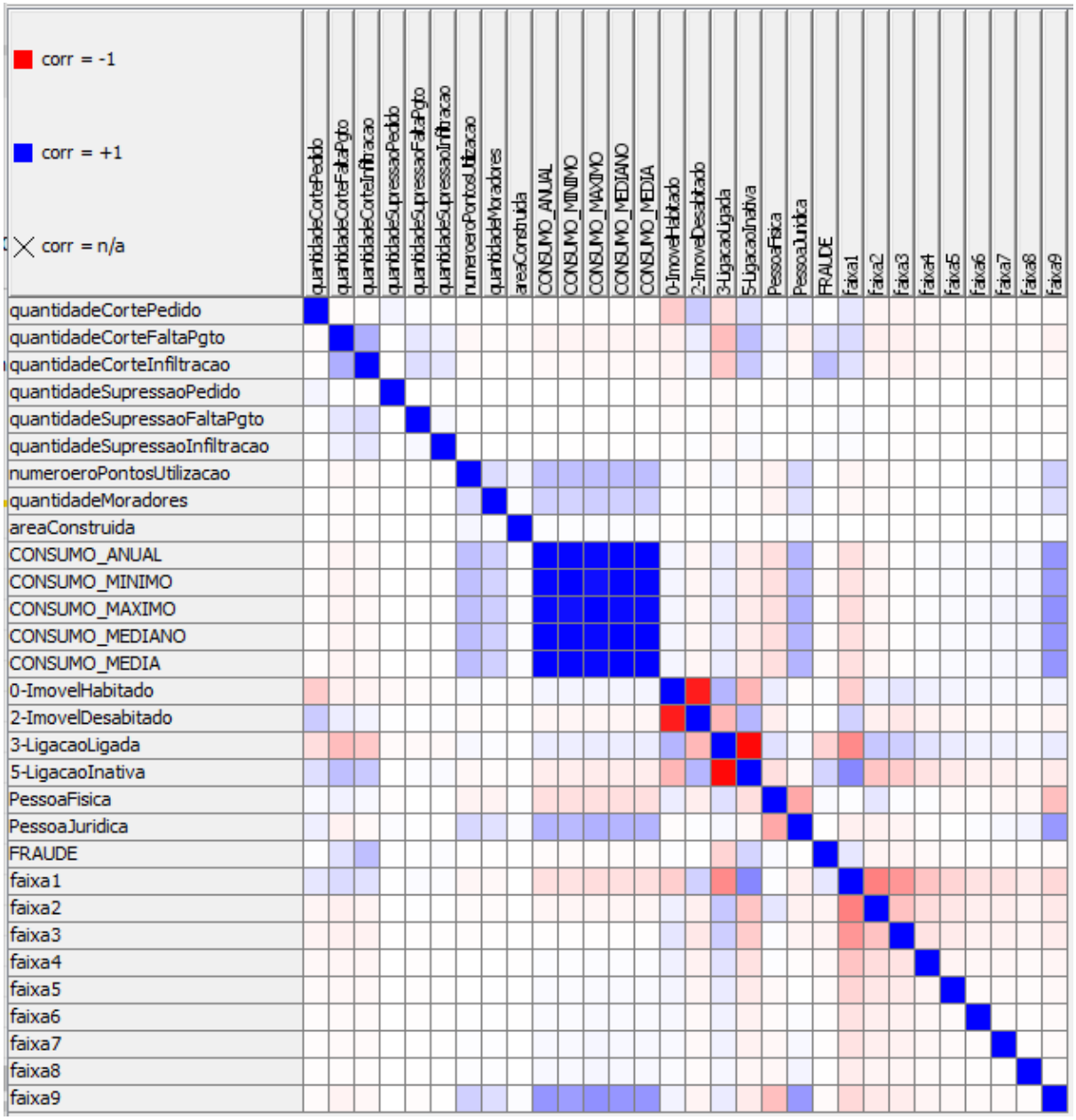

Figura 3.3 Gráfico de correlação das variáveis.

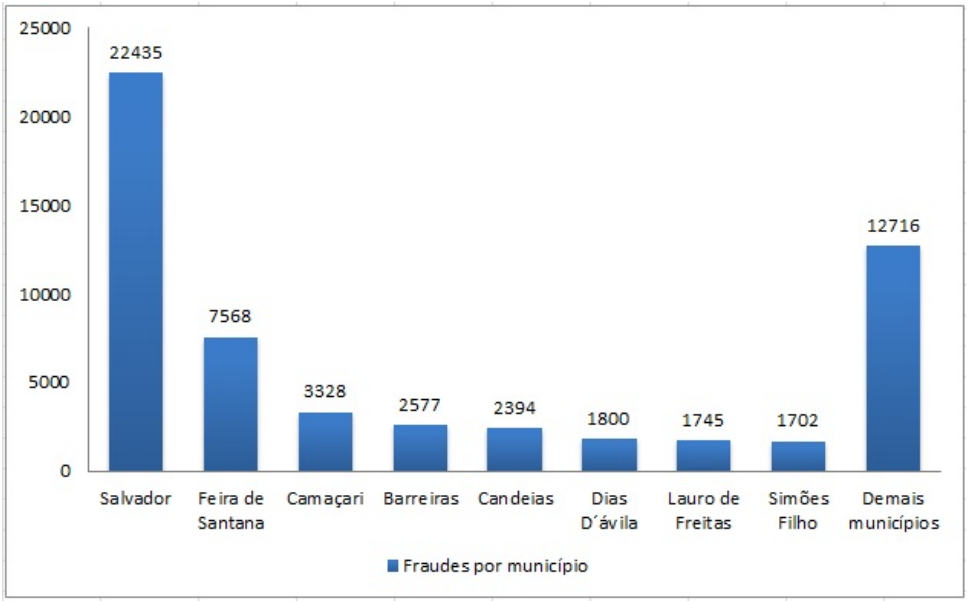

Figura 3.4 Quantidade de registros de fraudes por município (2018).

A Tabela 3.2 apresenta o detalhamento dos registros das cidades selecionadas. O *dataset* contém 823.954 registros, sendo 603.089 (73,19%) de clientes estabelecidos em Salvador e 220.865 (26,81%) de clientes em Feira de Santana. Do total, 30.003 (3,64%) são matrículas com fraude e 793.951 (96,36%) são matrículas sem fraude, o que configura-se um *dataset* desbalanceado.

Cidade	**Fraude**	**Não-Fraude**	**Total**
Salvador	22.435	580.654	603.089
Feira de Santana	7.568	213.297	220.865
Total	30.003	793.951	823.954

Tabela 3.2 Registros de fraude e não fraude por cidade.

Nesta etapa, foram realizadas as tarefas de seleção dos dados que serão usados no processo de desenvolvimento do modelo analítico. Em seguida, serão aplicadas técnicas de pré-processamento e transformação, preparando os dados para uso com os algoritmos de aprendizado de máquina.

3.1.2 Pré-processamento e Transformação

As técnicas de pré-processamento e transformação foram aplicadas com o objetivo de preparar os dados de modo a estarem adequados para processamentos com os algoritmos de classificação supervisionada. De acordo com Camilo e Silva (2009), esta é uma etapa fundamental para o sucesso do modelo, pois alguns algoritmos trabalham apenas com valores numéricos, enquanto que outros apenas com valores categóricos, sendo necessário a transformação dos tipos de dados.

Na etapa de pré-processamento, foram aplicadas as técnicas de discretização, *one hot encoding* e normalização, de acordo com o tipo das variáveis (Tabela 3.1). A Figura 3.5 apresenta o *workflow* das etapas do pré-processamento dos dados.

De acordo com Kotsiantis, Kanellopoulos e Pintelas (2006), a maioria dos algoritmos de *machine learning* são capazes de extrair conhecimento de *datasets* com dados discretos. Em caso de dados contínuos, são recomendadas a adoção de técnicas de discretização. A discretização transformam os dados contínuos em atributos discretos, reduzindo significativamente o número de valores possíveis e, consequentemente, contribuindo para a velocidade e eficiência do modelo resultante (KOTSIANTIS; KANELLOPOULOS; PINTELAS, 2006).

A discretização foi aplicada sobre a variável 'ConsumoMedia' (média do volume consumido), usando o componente *Numeric Binner* (ver *workflow* apresentado na Figura 3.5). Foram criadas categorias de acordo com as nove faixas de consumo adotadas pela Embasa, transformando faixas de valores contínuos em valores discretos. A Tabela 3.3 apresenta a discretização dos dados em faixas de consumo adotados pela Embasa.

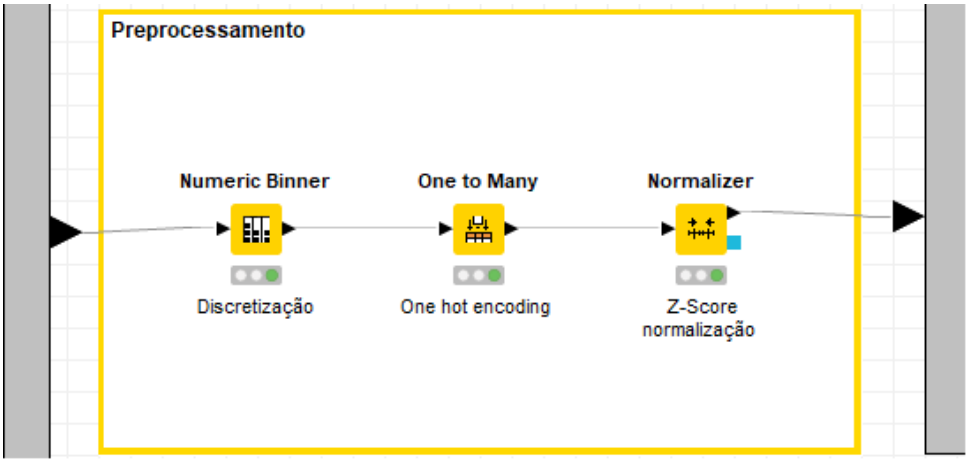

Figura 3.5 *Workflow* para o pré-processamento dos dados.

Faixas	**Faixa de Valores**
Faixa01	< 6m3
Faixa02	6 a 10m3
Faixa03	10 a 15m3
Faixa04	15 a 20m3
Faixa05	20 a 25m3
Faixa06	25 a 30m3
Faixa07	30 a 40m3
Faixa08	40 a 50m3
Faixa09	> 50m3

Tabela 3.3 Tabela de Discretização

Para as variáveis categóricas, usando o componente *One to Many* (segundo passo da Figura 3.5), aplicou-se a técnica *one hot encoding* para transformar cada registro das categorias em variáveis numéricas binárias (as quais são também conhecidas como variáveis *dummies*). Os novos recursos gerados podem levar à criação de classificadores mais concisos e precisos. Além disso, a descoberta de recursos significativos contribuem para uma melhor compreensão do classificador produzido, e melhor compreensão do conceito aprendido (KOTSIANTIS; KANELLOPOULOS; PINTELAS, 2006).

Em relação às variáveis numéricas, foi aplicada a técnica de normalização usando o componente *Normalizer* com o método Z-Score (terceiro passo da Figura 3.5), transformando os valores que estão em escalas diversas em uma escala comum. A normalização reduz a dimensionalidade do dados e permite que algoritmos de aprendizagem operem com mais rapidez e tenham mais eficiência (KOTSIANTIS; KANELLOPOULOS; PINTELAS, 2006).

A etapa de transformação dos dados foi realizada aplicando as técnicas de subamostragem e particionamento, conforme apresentada na Figura 3.6. Esta etapa realiza as atividades finais do processo de preparação dos dados para a devida utilização no processo construtivo dos modelos.

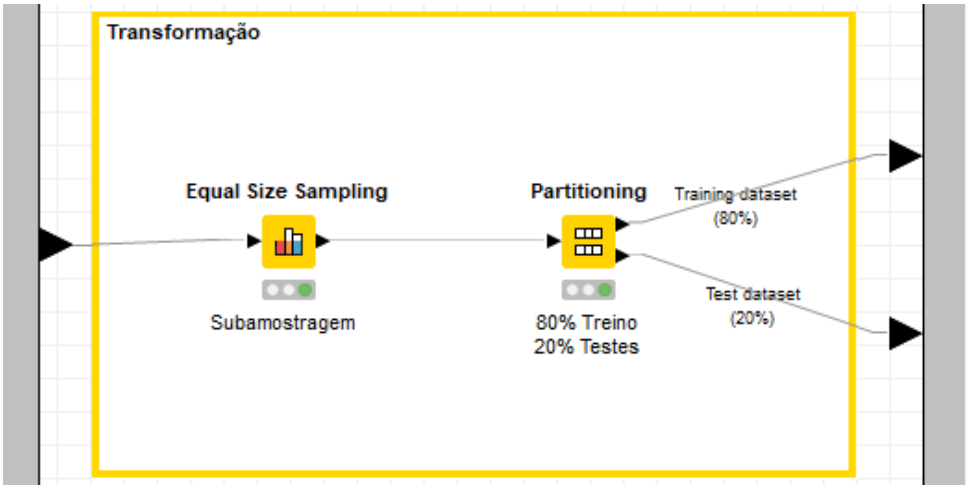

Figura 3.6 *Workflow* para a transformação dos dados.

Conforme descrito na Seção 3.1.1 “Data Selection”, os dados originalmente disponíveis estavam desbalanceados em relação ao número de registros com e sem fraude. Estudos realizados por Thabtah et al. (2020) demonstram que o desbalanceamento de classes tem um impacto negativo significativo no desempenho de um classificador. Dentre os métodos para tratar o desbalanceamento entre as classes, a subamostragem visa balancear o conjunto de dados por meio da eliminação de exemplos da classe majoritária (THABTAH et al., 2020).

Assim, no *workflow* da Figura 3.6, no componente *Equal Size Sampling*, foram selecionados todos os registros de fraudes (30.003 matrículas) e selecionados aleatoriamente 30.003 matrículas sem fraudes, totalizando 60.006 matrículas para o processo de construção dos modelos.

Para subsidiar as etapas de otimização, treinamento e teste dos modelos, os dados foram particionados usando o componente *Partitioning*, seguindo o padrão 80-20 onde: 48.004 registros (80%) foram destinados para otimização e treinamento, denominando "*Training dataset*"; enquanto que 12.002 registros (20%) foram separados para o teste final do modelo, denominando "*Test dataset*". Para isso, utilizou-se o método de amostragem estratificada dos registros com base na variável alvo "StatusFraude".

Nesta etapa, foram realizadas as tarefas de pré-processamento e transformação dos dados, preparando-os para uso com os algoritmos de aprendizagem supervisionada. Na próxima seção, aplicaremos as técnicas de otimização de parâmetros dos algoritmos de *machine learning* e validação cruzada. Estas são etapas fundamentais do processo de construção de modelos de *machine learning*.

3.1.3 Modelagem

No processo de construção e avaliação de preditores, foram considerados algoritmos de aprendizagem para classificação supervisionada tradicionais como: *Decision Tree*, *Support Vector Machine (SVM)*, *Logistic Regression* e *Naive Bayes*. Também foram usados algoritmos do tipo *ensemble learning* como *Random Forest* e *Gradient Boosting*, que geralmente apresentam resultados mais precisos do que os algoritmos tradicionais.

Dentre as diversas técnicas de avaliação de modelos de *machine learning*, foi aplicada a técnica de validação cruzada. Assim, o conjunto de treinamento foi subdivido em 10 subconjuntos (*folds*), sendo que a cada iteração, nove foram utilizados para construção do modelo e um utilizado para avaliação. Com isso, os modelos foram treinados e validados com amostras independentes por 10 vezes.

Os algoritmos de *machine learning* possuem diversos parâmetros que podem ser configurados com o objetivo de promover melhores resultados e desempenho. Para subsidiar a escolha da melhor configuração de cada algoritmo, foi aplicada a técnica de otimização dos hiperparâmetros via *Grid Search*. Assim, um conjunto de parâmetros foram previamente escolhidos. A Tabela 3.4 apresenta os hiperparâmetros e configurações avaliadas para cada algoritmo.

Algoritmo	**Hiperparâmetros**
Decision Tree	Number records to store for view: (5000, 10000, 15000, 20000 e 25000) Min number records per node: (1, 2, 3, 4 e 5) Pruning metod: No pruning
Random Forest	Number of Models: (100, 200, 300, 400, 500, 600, 700, 800, 900 e 1000) Split criterion: Gini Index
Gradient Boosting	Number of models: (100, 200, 300, 400, 500, 600, 700, 800, 900 e 1000) Limit number of levels: (1, 2, 3, 4 e 5) Learning rate: (0.1, 0.2, 0.3, 0.4 e 0.5)
Logistic Regression	Maximal number of epochs: (1000, 1100, 1200, 1300, 1400, 1500, 1600, 1700, 1800, 1900 e 2000) Epsilon: 1.0E-4 Learning rate strategy: LineSearch Regularization Prior: Laplace
Linear SVM	Number of iterations: (100, 150, 200, 250, 300, 350, 400, 450 e 500) Regularizer: L2 Loss Function: Logistic
Naive Bayes	Minimum standard desviation: (0.6, 0.7, 0.8, 0.9 e 1.0) Default probability: 0.001 Threshold standard desviation: 0.5

Tabela 3.4 Tabela de Hiperparâmetros utilizadas na otimização dos modelos.

Para cada algoritmo, foi criado no Knime um componente com um *workflow* para realizar a otimização de parâmetros e validação cruzada. O componente recebe os dados de treinamento e testes separadamente. Os resultados da otimização e validação cruzada será a saída de cada componente. Estes resultados serão agrupados para uso posterior no treinamento dos modelos finais de cada algoritmo.

A Figura 3.7 apresenta o *workflow* geral criado para a construção dos modelos, enquanto que a Figura 3.8 apresenta o *workflow* contido no componente do algoritmo, responsável por realizar a otimização dos parâmetros e validação cruzada.

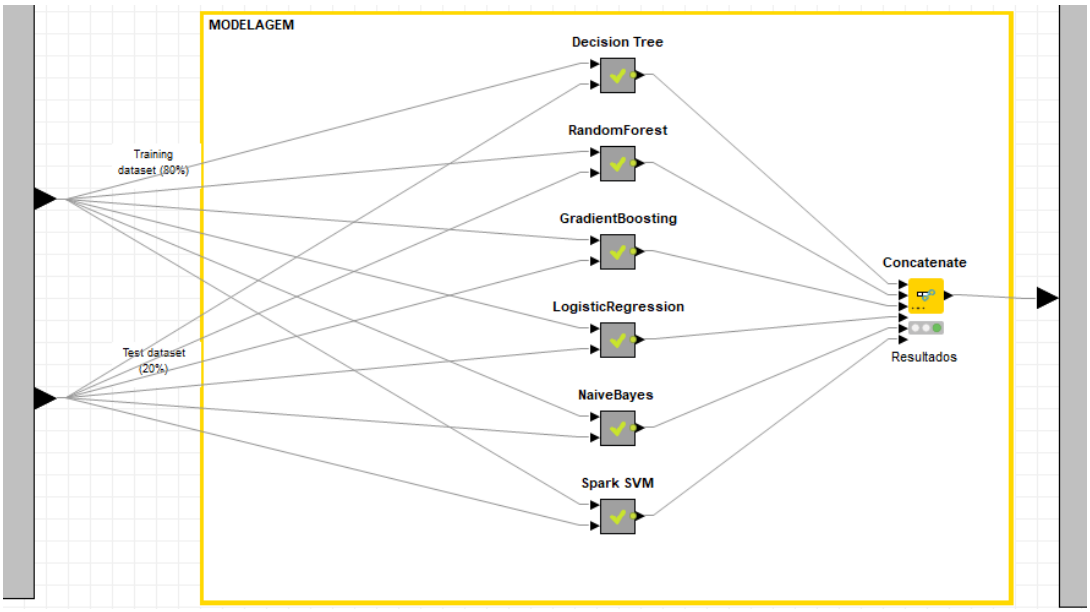

Figura 3.7 *Workflow para desenvolvimento do modelo (Modelagem).*

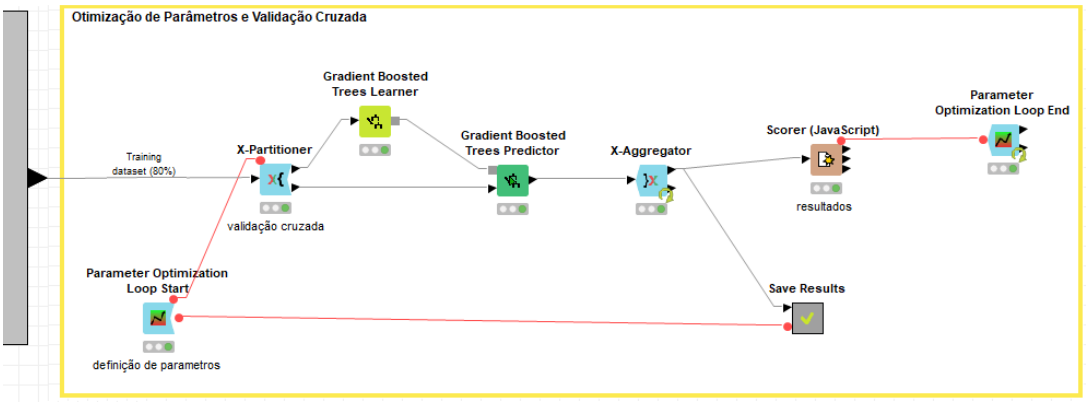

Figura 3.8 *Workflow* para otimização de parâmetros e validação cruzada.

O *workflow* de otimização e avaliação recebe apenas os dados de treinamento (*Training dataset*). Os componentes *Parameter Optimization Loop Start* e *Parameter Optimization Loop End* são responsáveis por realizar as iterações com as opções de parâmetros do *Grid*

Search. Os componentes *X-Partitioner* e *X-Aggregator* são responsáveis por realizar o *loop* da validação cruzada. Os componentes *Gradient Boosted Trees Learner* e *Gradient Boosted Trees Predictor* são responsáveis pela configuração do algoritmo e realizar o treinamento e avaliação do modelo. Estes componentes variam a depender do algoritmo em execução. Por sua vez, o componente *Scorer* é o responsável por consolidar os resultados e criar a matriz de confusão.

Por fim, os resultados foram salvos usando o componente *Save Results*. Para cada iteração no conjunto de hiperparâmetros do *Grid Search*, foram gerados 10 resultados da validação cruzada. Com base nesses dados, foram calculados a média e desvio padrão de medidas de avaliação da área (*Recall, Precision, F-measure, Accuracy e Error*). Esses resultados serão usados para realizar a escolha da melhor configuração do algoritmo, para o treinamento definitivo do modelo.

Nesta etapa, foram realizadas as tarefas de otimização de parâmetros e validação cruzada dos seis modelos com os algoritmos de aprendizagem supervisionada. Os resultados foram salvos para subsidiar o treinamento dos modelos finais e, finalmente, a escolha do melhor modelo. A próxima seção detalha o processo realizado para realizar tais tarefas.

3.1.4 Avaliação

A avaliação dos resultados foi realizada a partir da construção dos modelos usando as melhores configurações obtidas na otimização paramétrica. Foi construído um *workflow* para cada algoritmo, vide Figura 3.9. Todo o conjunto de dados de treino (*Training dataset*) foi utilizado para treinar o modelo, enquanto que os dados de testes (*Test dataset*) foram utilizados para realizar o teste independente do modelo. Como saída do *workflow*, os resultados foram usados para comparar com os modelos dos demais algoritmos.

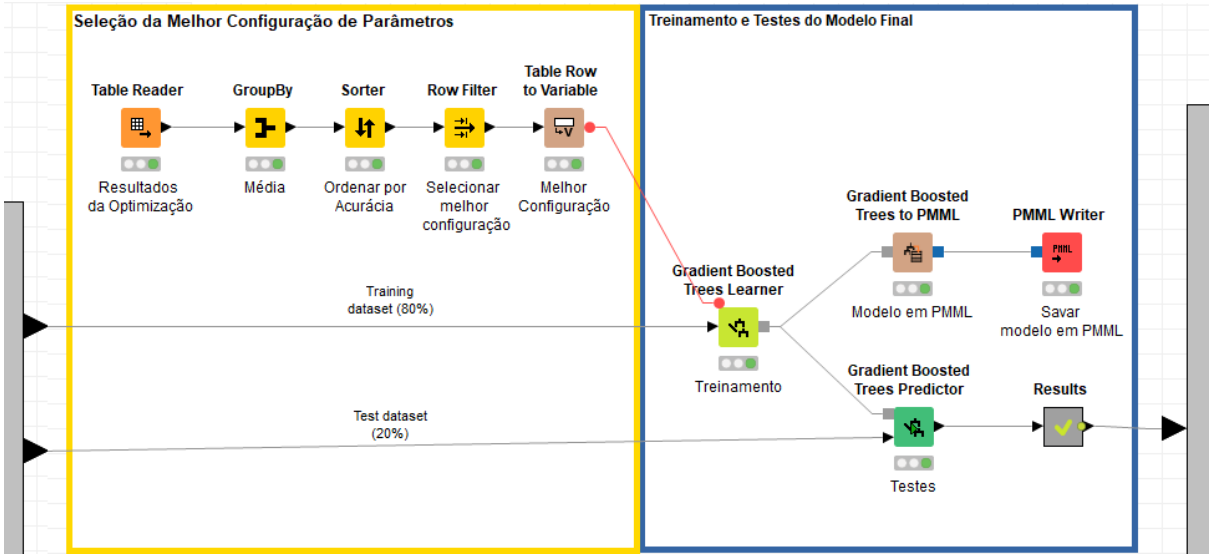

Figura 3.9 *Workflow* para treinamento e teste do modelo final.

Os resultados da otimização de hiperparâmetros e da validação cruzada foram obtidos com o componente *Table Reader*, e calculados as médias e desvio padrão das métricas

de avaliação. Em seguida, com o componente *Sorter*, os registros foram ordenados em ordem decrescente pela métrica Acurácia e realizado o filtro para selecionar o registro do melhor resultado com o componente *Row Filter*. Foram selecionados os registros com a maior acurácia de cada algoritmo. A Tabela 3.5 apresenta a melhor configuração de hiperparâmetros obtida para cada um dos algoritmos.

Os valores das melhores configurações de hiperparâmetros foram transformados em variáveis pelo componente *Table Row to Variable* e passados para o componente *Gradient Boosting Trees Learner*, onde foi realizado o treinamento do modelo com todo o conjunto de dados de treino (Figura 3.9). O modelo criado foi testado com o componente *Gradient Boosted Trees Predictor* e os dados de teste. Estes componentes mudam de acordo com o algoritmo utilizado.

Por fim, os resultados foram salvos no componente *Results* para comparação com os resultados dos demais modelos. O modelo gerado foi salvo em arquivo no formato PMML utilizando os componentes *Gradient Boosted Trees to PMML* e *PMML Writer*.

A Seção 3.2 apresenta os resultados obtidos e discussão sobre o uso do modelo no processo de combate a fraudes no consumo de água.

Algoritmo	**Hiperparâmetros**
Decision Tree	Number records to store for view: 25000 Min number records per node: 5 Pruning metod: No pruning
Random Forest	Number of models: 1000 Split criterion: Gini Index
Gradient Boosting	Number of models: 900 Limit number of levels: 4 Learning Rate: 0.2
Logistic Regression	Maximal number of epochs: 1500 Epsilon: 1.0E-4 Learning rate strategy: LineSearch Regularization Prior: Laplace
Linear SVM	Number of iterations: 150 Regularizer: L2 Loss Function: Logistic
Naive Bayes	Minimum standard desviation: 0.74 Default probability: 0.001 Threshold standard desviation: 0.5

Tabela 3.5 Tabela com os melhores hiperparâmetros para os algoritmos avaliados.

3.2 RESULTADOS E DISCUSSÕES

O processo de otimização paramétrica dos algoritmos de aprendizado foi realizado para construção e avaliação dos modelos de detecção de fraudes no consumo de água e seleção daquele mais eficaz. As principais medidas de avaliação dos modelos foram computadas para subsidiar a escolha da melhor configuração de cada um dos algoritmos.

A medida de acurácia geral dos modelos foi adotada para escolha das melhores configurações e também para definição do melhor modelo. A Tabela 3.6 apresenta os resultados obtidos por cada um dos modelos para o conjunto de teste final.

Algoritmo	**Verdadeiro Positivo**	**Falso Positivo**	**Verdadeiro Negativo**	**Falso Negativo**	**Acurácia**
Gradient Boosting	76.34%	17.10%	82.90%	23.66%	79.62%
Random Forest	75.15%	17.41%	82.59%	24.85%	78.87%
Logistic Regression	69.91%	16.81%	83.19%	30.09%	76.55%
Linear SVM	71.04%	20.00%	80.00%	28.96%	75.52%
Naive Bayes	73.82%	25.35%	74.65%	26.18%	74.24%
Decision Tree	70.97%	23.08%	76.92%	29.03%	73.95%

Tabela 3.6 Tabela de Resultados dos modelos.

Dentre os seis algoritmos avaliados, pode-se observar que os modelos com os algoritmos do tipo *ensemble learning* apresentaram os melhores resultados, sendo o *Gradient Boosting* com índice de acurácia geral de 79,62% e o *Random Forest* com (78,87%) para o mesmo índice. Os demais modelos, obtiveram os seguintes resultados: *Logistic Regression* (76,55%), *Linear SVM* (75,52%), *Naive Bayes* (74,24%) e *Decision Tree* (73,95%). Estes resultados demonstram que geralmente o uso de diversos preditores em conjunto produz modelos com maior índice de assertividade do que modelos com algoritmos tradicionais.

Apesar da pouca diferença do resultado final, o modelo com o algoritmo *Gradient Boosting* apresentou o melhor resultado e, por isso, foi escolhido para continuidade do estudo. O modelo salvo em formato PMML poderá ser implantado em ambiente produtivo.

A Figura 3.10 apresenta a matriz de confusão e estatísticas dos resultados obtidos no modelo treinado com o algoritmo *Gradient Boosting*.

De acordo com a matriz de confusão, o modelo classificou como fraude um total de 5.607 registros, sendo que destes, foi possível identificar corretamente 4.581 registros de fraudes (verdadeiros positivos), resultando em um índice *recall* de 76.34% e precisão de 81.70%. Em relação aos registros sem fraudes, o modelo classificou 6.395 registros, sendo que destes, 4.975 foram classificados corretamente (falsos positivos). Com isso, obteve-se o índice *recall* de 82,90% e precisão de 77.80%.

Por fim, o modelo classificou incorretamente 1.420 (23,66%) registros como não-fraudes (falsos negativos) e 1.026 (17.09%) registros como fraudes (falsos positivos).

Confusion Matrix

Rows Number : 12002	FRAUDE (Predicted)	NFRAUDE (Predicted)
FRAUDE (Actual)	4581	1420
NFRAUDE (Actual)	1026	4975

Class Statistics

Class	True Positives	False Positives	True Negatives	False Negatives	Recall	Precision	Sensitivity	Specificity	F-measure
FRAUDE	4581	1026	4975	1420	76.34%	81.70%	76.34%	82.90%	78.93%
NFRAUDE	4975	1420	4581	1026	82.90%	77.80%	82.90%	76.34%	80.27%

Overall Statistics

Overall Accuracy	Overall Error	Cohen's kappa (κ)	Correctly Classified	Incorrectly Classified
79.62%	20.38%	0.592	9556	2446

Figura 3.10 Matriz de confusão e estatísticas gerais do melhor modelo com o algorítmo *Gradient Boosting*.

Totalizando taxa média de erro de 20.38%. No geral, os números demonstram que o modelo possui uma alta capacidade de classificação correta dos registros de fraude e não-fraude (79,62%), podendo se tornar uma importante ferramenta no combate às ligações clandestinas e às perdas aparentes do consumo de água.

A partir da aplicação do modelo desenvolvido será possível a recomendação e direcionamento das inspeções *in loco* de modo mais assertivo, podendo proporcionar ganho de desempenho das equipes operacionais, redução de custos com as inspeções físicas e redução de perdas com a regularização dos serviços aos usuários do sistema de abastecimento de água.

Não é possível estabelecer um comparativo direto e exato entre os resultados aqui apresentados e os outros disponíveis na literatura, devido às limitações de indisponibilidade dos mesmos dados e dificuldade de replicação de estudos anteriormente propostos (Seção 2.3). Contudo, pode-se verificar que, de modo geral, as taxas de eficácia apresentadas aqui são superiores àquelas obtidas em estudos anteriores, como (PASSINI; TOLEDO, 2002)(MONEDERO et al., 2015). Além disso, destaca-se que este estudo baseou-se em uma base de dados em larga escala e considerou um processo de validação experimental rigoroso.

Dado o volume de dados aqui utilizado, os modelos foram expostos à uma grande diversidade de situações de fraudes, além do uso de técnicas de otimização baseada em validação cruzada e utilização de amostra de dados independente para avaliação final. Isso permitiu validar rigorosamente a capacidade de generalização dos modelos, superando limitações metodológicas de trabalhos anteriores (DE CASTRO FETTERMANN et al., 2015)(AL-RADAIDEH; AL-ZOUBI, 2018).

As fraudes no consumo de água representam uma parcela importante da perda de faturamento das empresas de saneamento e, portanto, devem ser combatidas para evitar a disseminação entre os consumidores e a respectiva queda de receitas das empresas. Por outro lado, a seleção equivocada de possíveis fraudes pode acabar gerando mais custos com equipes de inspeções em campo. Devido a isso, sistemas de detecção de fraudes devem buscar a menor taxa de erro possível.

Para reduzir as taxas de erros do classificador, deve-se avaliar os resultados em busca

de padrões entre os registros classificados equivocadamente, buscando perspectivas que possam auxiliar na evolução da solução e/ou propostas de novas soluções.

Diante disso, comparando os resultados da classificação com os dados originais do *dataset* usado para teste do modelo (*Test Dataset*), temos que o modelo acertou 60% dos registros de ligações ativas, 89% dos registros de ligações inativas e 67% dos demais tipos. A Figura 3.11 apresenta a comparação dos resultados por tipo de ligação.

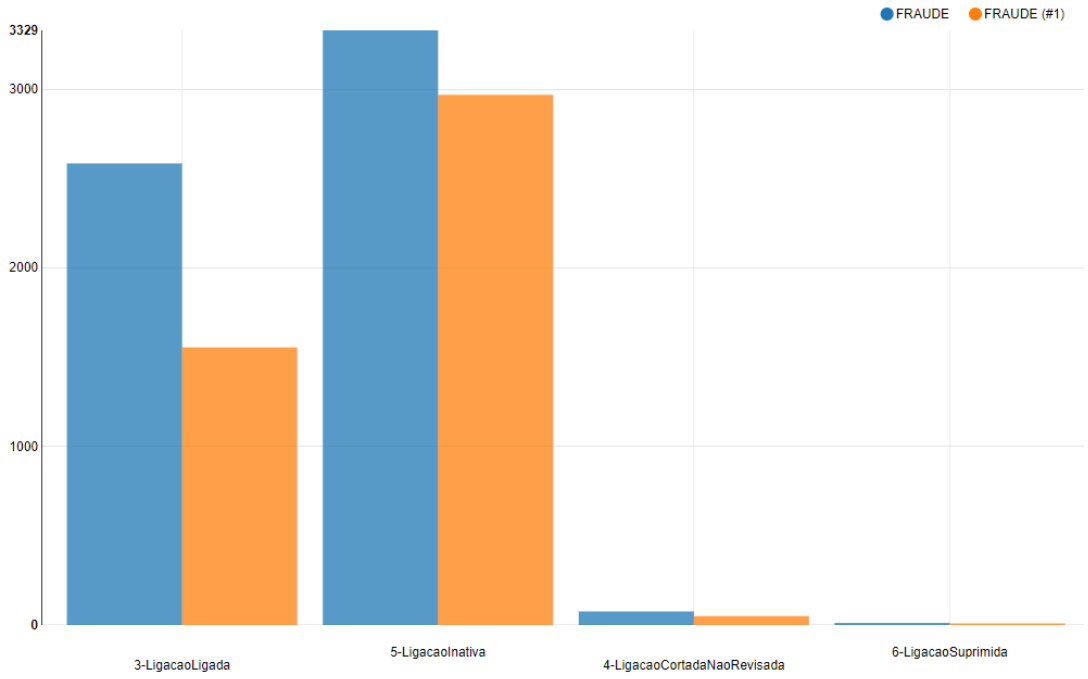

Figura 3.11 Comparação do Resultado por Tipo de Ligação.

Analisando os registros de falsos negativos (registros de fraudes classificados como não-fraudes), percebe-se que 73% (1.030) são consumidores com ligações ativas, 25% são consumidores com ligações inativas e 2% de outros tipos. A Figura 3.12 apresenta a distribuição por tipo de ligação dos registros de falsos negativos.

Portanto, apesar do modelo apresentar bons resultados no geral, observa-se que ainda encontra dificuldades em classificar corretamente as fraudes ocorridas em consumidores com ligações ativas.

A identificação de fraudes em ligações ativas é bastante complexa devido às diversas variáveis que influenciam no consumo de água. Dificilmente pode-se atribuir uma redução de consumo à existência de algum tipo de fraude ou mudança de hábitos dos consumidores, a redução na quantidade de moradores ou viagens temporárias.

Os algoritmos de *machine learning* de aprendizagem supervisionada funcionam bem a partir da aprendizagem com os registros disponibilizados para treinamento do modelo. Conforme demonstrado na Figura 3.3, os registros com Ligações Ativas (3-LigacaoLigada) possuem uma correlação negativa com a variável alvo (FRAUDE) e, portanto, pode ter impactado nos resultados para esta categoria.

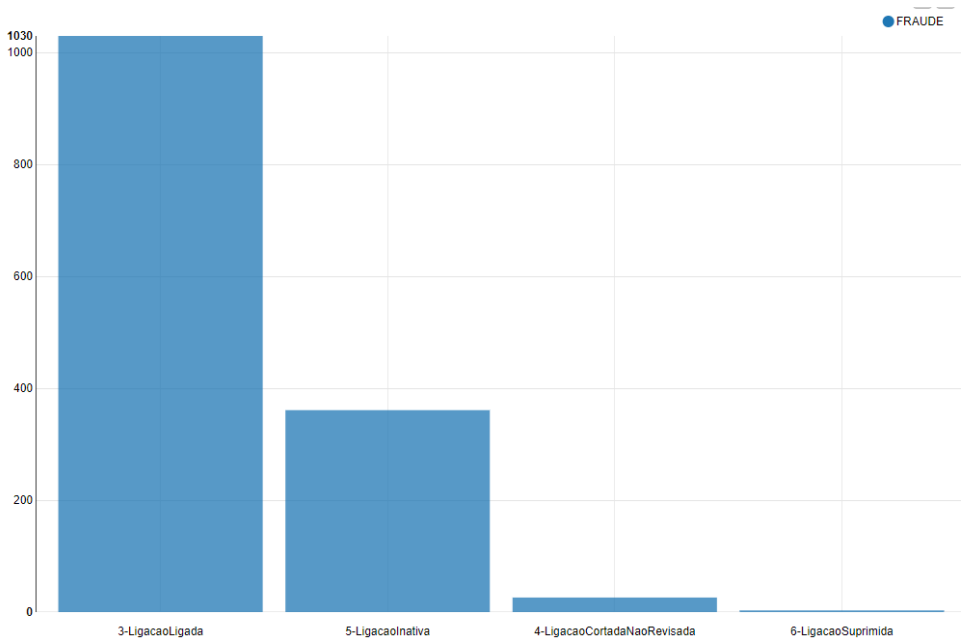

Figura 3.12 Resultado de Falsos Negativos por Tipo de Ligação.

Para melhorar o desempenho do modelo usando algoritmos de classificação, será preciso obter mais exemplos de registros de fraudes em ligações ativas e realizar uma minuciosa análise em busca de obter variáveis com melhor correlação com as fraudes.

Uma outra opção é o desenvolvimento de novos modelos utilizando outras técnicas de inteligência computacional como Aprendizagem Profunda (*Deep Learning*), Redes Neurais (*Neural Networks*) ou Séries Temporais (*Times Series Analytics*) para uso complementar com o modelo apresentado.

Métodos supervisionados e não-supervisionados são poderosas ferramentas complementares para detecção de fraudes (BAESENS; VLASSELAER; VERBEKE, 2015). De acordo com Phua et al. (2010), o uso de modelos híbridos podem proporcionar melhores resultados do que o uso individual de modelos de aprendizagem supervisionada. Assim, sugerimos que novos estudos sejam realizados buscando melhorar os resultados na identificação de fraudes em ligações ativas.

Reconhecidas as limitações, destaca-se aqui a relevante contribuição que este trabalho traz, do ponto de vista prático, para a área de detecção de fraudes em consumo de água.

Capítulo 4

CONCLUSÃO

No presente estudo, foram utilizadas técnicas de *machine learning* para detecção de fraudes no consumo de água em uma empresa pública de saneamento. Foram aplicadas técnicas de pré-processamento de dados, otimização de hiperparâmetros e validação cruzada para construção de modelos preditivos. Foram utilizados algoritmos de aprendizagem supervisionada tradicionais como: *Decision Tree*, *Support Vector Machine (SVM)*, *Logistic Regression* e *Naive Bayes*, além de algoritmos do tipo *ensemble learning* como *Random Forest* e *Gradient Boosting*.

Na etapa de seleção de dados, foram selecionadas 60.006 matrículas das cidades de Salvador e Feira de Santana (Bahia, Brasil), sendo que 30.003 matrículas com fraude e, após o processo de balanceamento do *dataset*, foram selecionadas aleatoriamente mais 30.003 matrículas sem fraudes.

Dentre os seis algoritmos avaliados, o algoritmo *Gradient Boosting* apresentou o melhor resultado, com índice de acurácia geral de 79,62%. O modelo classificou como fraude um total de 5.607 registros, sendo que destes, foi possível identificar corretamente 4.581 registros de fraudes (verdadeiros positivos), resultando em um índice *recall* de 76.34% e precisão de 81.70%. Em relação aos registros sem fraudes, o modelo classificou 6.395 registros, sendo que destes, 4.975 foram classificados corretamente (falsos positivos). Com isso, obteve-se o índice *recall* de 82,90% e precisão de 77.80%.

Os resultados obtidos denotam a eficácia do modelo construído, demonstrando alto poder de detecção de casos de fraudes e não fraudes. O modelo foi treinado com técnicas de alto rigor científico, contribuindo para o avanço do estado da arte em detecção de fraudes no saneamento usando *machine learning*.

4.1 RESULTADOS ALCANÇADOS

Como resultados alcançados, o principal produto foi o modelo desenvolvido para detecção de fraudes no saneamento. Este modelo, denominado *Water Fraud Analytics*, obteve o registro de software junto ao Instituto Nacional da Propriedade Industrial (INPI). Em anexo está o certificado de registro do software.

Foi escrito um artigo abordando o processo de desenvolvimento do modelo analítico para combate às fraudes. Este artigo está em processo de ajustes para publicação em revista científica conceituada.

O projeto também participou do Programa Centelha-Bahia de incentivo a criação de empreendimentos inovadores, sendo selecionado na primeira fase do edital (2019). Por fim, também foi selecionado no programa de inovação interna da Embasa, onde será executado um prova de conceito (POC) em algumas localidades.

4.2 LIMITAÇÕES

Devido ao período da pandemia mundial da Covid-19 (OPAS, 2021), iniciada em meados de março/2020, não foi possível a realização de experimentos com a efetiva inspeção das ligações pois estava em vigor medidas de distanciamento social e Termo de Ajuste de Conduta (TAC) (BAHIA, 2021) que impediam a Embasa de executar serviços que poderiam implicar na realização de cortes de fornecimento de água. Portanto, assim que possível, o modelo apresentado neste trabalho poderá ser alvo de uma avaliação comparativa com os resultados obtidos por equipes de especialistas na identificação de fraudes no saneamento.

4.3 TRABALHOS FUTUROS

Como trabalhos futuros, pode-se buscar outros modelos de aprendizado, como redes neurais profundas (*deep learning*) e outras técnicas de identificação de fraudes, como redes de relacionamento e identificação de anomalias de consumo com análise de *outliers*, principalmente para identificar fraudes em ligações ativas.

É sabido que o consumo de água pode sofrer variações temporais de acordo com o clima e também outros fatores como a quantidade de moradores, renda familiar e tamanho dos imóveis. Estudos para construir modelos de *machine learning* com séries temporais podem ser desenvolvidos para predizer o consumo de clientes e também auxiliar na identificação de fraudes.

Adicionalmente, as técnicas de *machine learning* também podem ser usadas para analisar e identificar outros tipos de perdas, como a análise de hidrômetros para evitar submedições ou para auxílio na identificação de vazamentos.

Todos esses modelos podem ser usados em um único software, tornando-se assim, uma plataforma analítica para gestão de perdas de água no saneamento.

Finalmente, o modelo construído poderá ser incorporado a um *software* de gestão operacional, auxiliando os técnicos na identificação de fraudes no consumo de água, contribuindo para a redução das perdas e aumento da eficiência operacional.

REFERÊNCIAS BIBLIOGRÁFICAS

AESBE, A. B. D. E. E. D. S. *Guia prático para estimação de consumo não autorizados e volumes não apropriados por falhas de cadastro*. 2015. [Online; accessed 30-September-2019]. Disponível em: ⟨http://www.aesbe.org.br/guias_praticos/⟩.

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

ANDRADE SOBRINHO, R.; BORJA, P. C. Gestão das perdas de água e energia em sistema de abastecimento de água da embasa: um estudo dos fatores intervenientes na rms. *Eng. sanit. ambient*, v. 21, n. 4, p. 783–795, 2016.

BAESENS, B.; VLASSELAER, V. V.; VERBEKE, W. *Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection*. [S.l.]: John Wiley & Sons, 2015.

BAHIA, D. P. da. *TAC - Embasa suspende corte de água*. 2021. [Online; accessed 08-11-2021]. Disponível em: ⟨https://www.defensoria.ba.def.br/noticias/coronavirus-embasa-firma-acordo-com-a-defensoria-e-suspende-corte-de-agua-de-usuarios-carentes/⟩.

BATISTA, G. E. d. A. P. et al. *Pré-processamento de dados em aprendizado de máquina supervisionado*. Tese (Doutorado) — Universidade de São Paulo, 2003.

BOLTON, R. J.; HAND, D. J. Statistical fraud detection: A review. *Statistical science*, JSTOR, p. 235–249, 2002.

BRASIL. Lei nº 12.026, de 15 de julho de 2020. marco legal do saneamento básico. *Diário Oficial da República Federativa do Brasil*, Brasília, DF, 2020. Disponível em: ⟨http://www.planalto.gov.br/ccivil_03/_Ato2019-2022/2020/Lei/L14026.htm⟩.

BRASIL, M. do Desenvolvimento Regional. Secretaria Nacional de S. S. *Sistema Nacional de Informações sobre Saneamento: Diagnóstico dos Serviços de Água e Esgotos -– 2017*. 2019. [Online; accessed 30-September-2019]. Disponível em: ⟨http://snis.gov.br/diagnostico-agua-e-esgotos/diagnostico-ae-2017⟩.

BRASIL, M. do Desenvolvimento Regional. Secretaria Nacional de S. S. *Sistema Nacional de Informações sobre Saneamento: Diagnóstico dos Serviços de Água e Esgotos -– 2018*. 2020. [Online; accessed 20-Julho-2020]. Disponível em: ⟨http://www.snis.gov.br/diagnostico-anual-agua-e-esgotos/diagnostico-dos-servicos-de-agua-e-esgotos-2018⟩.

BRASIL, T. Perdas de água 2020 (snis 2018): desafios para disponibilidade hídrica e avanço da eficiência do saneamento básico. *São Paulo*, 2020. [Online; accessed 30-September-2021]. Disponível em: ⟨https://www.tratabrasil.org.br/images/estudos/Relatorio_Final_-_Estudo_de_Perdas_2020_-_JUNHO_2020.pdf⟩.

BUCZAK, A. L.; GUVEN, E. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, IEEE, v. 18, n. 2, p. 1153–1176, 2015.

CAMILO, C. O.; SILVA, J. C. d. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. *Universidade Federal de Goiás (UFC)*, p. 1–29, 2009.

DE CASTRO FETTERMANN, D. et al. Uma sistemática para detecção de fraudes em empresas de abastecimento de água. *Interciencia*, Asociación Interciencia, v. 40, n. 2, p. 114–120, 2015.

DETROZ, J. P.; SILVA, A. T. da. Fraud detection in water meters using pattern recognition techniques. In: *Proceedings of the Symposium on Applied Computing*. [S.l.: s.n.], 2017. p. 77–82.

EMBASA. *EMBASA - Empresa baiana de águas e saneamento S/A*. 2020. [Online; accessed 30-08-2020]. Disponível em: ⟨https://www.embasa.ba.gov.br/index.php/institucional/a-embasa/apresentacao⟩.

ESPINOSA, F. H. T.; GISSELOT, F. F. P.; ARRIAGADA, I. R. B. Predicción de fraudes en el consumo de agua potable mediante el uso de minería de datos. *Universidad Ciencia y Tecnología*, v. 24, n. 104, p. 58–66, 2020.

FAWCETT, T.; PROVOST, F. J. Combining data mining and machine learning for effective user profiling. In: *KDD*. [S.l.: s.n.], 1996. p. 8–13.

FAYYAD, U. et al. From data mining to knowledge discovery in databases. *AI magazine*, v. 17, n. 3, p. 37–37, 1996.

FERNANDES, J. M. C. Redução de perdas aparentes em sistemas de abastecimento de água: Definição de critérios para identificação de consumos fraudulentos. 2014.

GOLDSCHMIDT, R.; PASSOS, E. *Data mining: um guia prático*. [S.l.]: Gulf Professional Publishing, 2005.

GOPAL, G. V.; BALAJI, V. Detection of fraudulent behaviour in water consumption using machine learning algorithms. 2020.

GUMIER, C. C.; LUVIZOTTO JUNIOR, E. et al. Aplicação de modelo de simulação-otimização na gestão de perda de água em sistemas de abastecimento. *Engenharia Sanitaria e Ambiental*, SciELO Brasil, 2007.

HUMAID, E. H.; BARHOUM, T. Water consumption financial fraud detection: A model based on rule induction. In: IEEE. *2013 Palestinian International Conference on Information and Communication Technology*. [S.l.], 2013. p. 115–120.

IBGE. *Instituto Brasileiro de Geografia e Estatística - IBGE*. 2020. [Online; accessed 30-08-2020]. Disponível em: ⟨https://cidades.ibge.gov.br/⟩.

IFC, I. F. C. *Manual sobre Contratos de Performance e Eficiência para Empresas de Saneamento em Brasil*. 2013.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. *Science*, American Association for the Advancement of Science, v. 349, n. 6245, p. 255–260, 2015.

KNIME. *Knime Analytics Platform*. 2020. [Online; accessed 15-march-2020]. Disponível em: ⟨https://www.knime.com/knime-open-source-story⟩.

KOTSIANTIS, S.; KANELLOPOULOS, D.; PINTELAS, P. Data preprocessing for supervised leaning. *International Journal of Computer Science*, Citeseer, v. 1, n. 2, p. 111–117, 2006.

KOTSIANTIS, S. B.; ZAHARAKIS, I.; PINTELAS, P. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, Amsterdam, v. 160, n. 1, p. 3–24, 2007.

KUSTERKO, S. et al. Gestão de perdas em sistemas de abastecimento de água: uma abordagem construtivista. *Engenharia Sanitária e Ambiental*, SciELO Brasil, v. 23, n. 3, 2018.

LIAW, A.; WIENER, M. et al. Classification and regression by randomforest. *R news*, v. 2, n. 3, p. 18–22, 2002.

MOHRI, M.; ROSTAMIZADEH, A.; TALWALKAR, A. *Foundations of machine learning*. [S.l.]: MIT press, 2018.

MONEDERO, I. et al. An approach to detection of tampering in water meters. *Procedia Computer Science*, Elsevier, v. 60, p. 413–421, 2015.

MOROTE, Á.-F.; HERNÁNDEZ-HERNÁNDEZ, M. Unauthorised domestic water consumption in the city of alicante (spain): A consideration of its causes and urban distribution (2005–2017). *Water*, Multidisciplinary Digital Publishing Institute, v. 10, n. 7, p. 851, 2018.

NASCIMENTO, N. d. O.; HELLER, L. Ciência, tecnologia e inovação na interface entre as áreas de recursos hídricos e saneamento. *Engenharia sanitária e ambiental*, SciELO Brasil, v. 10, n. 1, p. 36–48, 2005.

OPAS. *Histórico da Pandemia Covid-19*. 2021. [Online; accessed 08-11-2021]. Disponível em: ⟨https://www.paho.org/pt/covid19/historico-da-pandemia-covid-19⟩.

PASSINI, S. R. R.; TOLEDO, C. M. T. Mineração de dados para detecção de fraudes em ligações de água. *XI SEMINCO-Seminário de computação*, 2002.

PHUA, C. et al. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*, 2010.

QUEIROGA, R. M. Uso de técnicas de data mining para detecção de fraudes em energia elétrica. *Biblioteca Central da Universidade Federal do Espírito Santo*, 2005.

RASCHKA, S. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

SREEDEVI, E.; SWATHI, M. R. Data mining based model for detection of fraudulent behavior in water consumption. 2021.

SREEKANTH, D.; THINAKARAN, K. Metro water fraudulent prediction in houses using convolutional neural network and recurrent neural network. *REVISTA GEINTEC-GESTAO INOVACAO E TECNOLOGIAS*, v. 11, n. 4, p. 1177–1187, 2021.

SRIRAMULU, M. et al. Detection of fraudulent behaviour in water consumption. 2020.

SUN, Y.; WONG, A. K.; KAMEL, M. S. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, World Scientific, v. 23, n. 04, p. 687–719, 2009.

TARDELLI FILHO, J. Aspectos relevantes do controle de perdas em sistemas públicos de abastecimento de água. *Revista DAE*, v. 64, n. 201, p. 6–20, 2016.

THABTAH, F. et al. Data imbalance in classification: Experimental evaluation. *Information Sciences*, Elsevier, v. 513, p. 429–441, 2020.

UDDIN, S. A. et al. A data mining based model for detection of fraudulent behaviour in water consumption. 2019.

REPÚBLICA FEDERATIVA DO BRASIL

MINISTÉRIO DA ECONOMIA

INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL

DIRETORIA DE PATENTES, PROGRAMAS DE COMPUTADOR E TOPOGRAFIAS DE CIRCUITOS INTEGRADOS

Certificado de Registro de Programa de Computador

Processo Nº: **BR512021003119-9**

O Instituto Nacional da Propriedade Industrial expede o presente certificado de registro de programa de computador, válido por 50 anos a partir de 1° de janeiro subsequente à data de 21/05/2019, em conformidade com o §2°, art. 2° da Lei 9.609, de 19 de Fevereiro de 1998.

Título: Combate à Fraude

Data de publicação: 21/05/2019

Data de criação: 21/05/2019

Titular(es): COMBATEAFRAUDE TECNOLOGIA DA INFORMACAO S.A.

Autor(es): RAFAEL RODRIGUES VIANA

Linguagem: JAVA SCRIPT; PYTHON; OUTROS

Campo de aplicação: IN-02

Tipo de programa: AT-01; GI-01; GI-06; IA-01; PD-01; SO-07

Algoritmo hash: SHA-512

Resumo digital hash:
c12bc0c4eae9263acaffe423dc8c4875a512d641459ecb22af7a09b335d30ed0bb2e9b1436a84d703dd948d25bf622c0015745cccd74b0f0c63646ef3f9117d3

Expedido em: 21/12/2021

Aprovado por:
Carlos Alexandre Fernandes Silva
Chefe da DIPTO