UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE ENERGIAS ALTERNATIVAS E RENOVÁVEIS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

FERNANDA RODRIGUES PAULO

DETECÇÃO DE FRAUDE EM UNIDADES CONSUMIDORAS NÃO TELEMEDIDAS COM USO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA

JOÃO PESSOA
2020

FERNANDA RODRIGUES PAULO

DETECÇÃO DE FRAUDE EM UNIDADES CONSUMIDORAS NÃO TELEMEDIDAS COM USO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA

Dissertação de mestrado a ser apresentado ao programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal da Paraíba para qualificação como parte dos requisitos para obtenção do título de Mestre em Engenharia Elétrica.

ORIENTADOR: PROF. DR. JUAN MOISES MAURICIO VILLANUEVA
CO-ORIENTADO-R: PROF. DR. HELON DAVID DE MACÊDO BRAZ

JOÃO PESSOA
2020

Catalogação na publicação
Seção de Catalogação e Classificação

P331d Paulo, Fernanda Rodrigues.
Detecção de fraude em unidades consumidoras não telemedidas com uso de técnicas de aprendizado de máquina / Fernanda Rodrigues Paulo. - João Pessoa, 2020.
104 f. : il.

Orientação: Juan Moises Mauricio Villanueva.
Coorientação: Helon David de Macêdo Braz.
Dissertação (Mestrado) - UFPB/CEAR.

1. Energia elétrica. 2. Fraude de energia elétrica. 3. Inteligência artificial - Detecção de fraude. 4. Leitura e faturamento - Energia elétrica. I. Villanueva, Juan Moises Mauricio. II. Braz, Helon David de Macêdo. III. Título.

UFPB/BC CDU 620.91(043)

Elaborado por Larissa Silva Oliveira de Mesquita - CRB-15/746

UNIVERSIDADE FEDERAL DA PARAÍBA – UFPB
CENTRO DE ENERGIAS ALTERNATIVAS E RENOVÁVEIS – CEAR
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA - PPGEE

A Comissão Examinadora, abaixo assinada, aprova a Dissertação

DETECÇÃO DE FRAUDE EM UNIDADES CONSUMIDORAS NÃO TELEMEDIDAS COM USO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA

Elaborada por

FERNANDA RODRIGUES PAULO

como requisito parcial para obtenção do grau de
Mestre em Engenharia Elétrica.

COMISSÃO EXAMINADORA

PROF. DR. JUAN MOISES MAURICIO VILLANUEVA (Presidente)

PROF. DR. HELON DAVID DE MACEDO BRAZ

PROF. DR. YURI PERCY MOLINA RODRIGUEZ

PROF. DR. IVANOVITCH MEDEIROS DANTAS DA SILVA

João Pessoa/PB, 31 de julho de 2020

AGRADECIMENTOS

Agradeço aos meus pais, Democlitus e Ivaneide, por todo o apoio e incentivo ao longo dos anos, por sempre tentarem ao máximo proporcionar tudo de melhor. São meus espelhos de dedicação e do valor da educação. Sou grata também ao meu irmão, Gustavo, por me encorajar e torcer por mim, pela paciência nos dias ruins.

Agradeço também aos meus orientadores, Juan e Helon, por todos os ensinamentos, pelas orientações valiosas, pela gentileza e, principalmente, por toda a compreensão. Com uma rotina tão dinâmica, o apoio e a confiança de vocês no meu trabalho foi essencial. O meu muito obrigada também a Ivanovitch pelas importantes contribuições tanto nesse trabalho, como para estudos futuros; parabenizo pelo profissionalismo e pela atenção. Agradeço também a Yuri por aceitar participar da banca e poder contribuir para esse trabalho.

Sou muito grata a minha amada amiga Elayne por ter sido porta de tantas coisas boas na minha vida. O tema desse trabalho começou com uma indicação de estágio e, hoje, cresceu além do que eu esperava. Agradeço por todos os aprendizados, os incentivos, as lágrimas e alegrias que compartilhamos nesses dois anos de mestrado.

Agradeço ao meu amigo Danillo por todas as essenciais contribuições a esse trabalho, por sanar minhas dúvidas com tanta brandura e compartilhar ideias. Sou grata especialmente pela força, pelo respeito mútuo e por acreditar em mim.

Agradeço a Samuel por todo o apoio e compreensão, por dividir comigo as minhas lutas e celebrar as minhas conquistas. Sou grata também a Jordan por comemorar tão genuinamente as minhas vitórias, sempre desejando o meu crescimento e pronto para ajudar em tudo que é possível. Agradeço a Alexandre, por sempre se preocupar comigo, pelos aconselhamentos e por nunca me deixar desistir.

Sou muito grata ao meu coordenador Luciano, por confiar no meu trabalho, pelas flexibilidades em meio ao mestrado e pela inspiração para desenvolver o caminho da minha pesquisa desde o início do estágio. Agradeço também ao meu gerente Manoel e a cada um do CICOP, em especial, a Juliano pelos ensinamentos e ajustes nesse trabalho e pelo grande incentivo e, também, a Elisa e Gabriel por terem me dado suporte para cumprir as entregas do setor. Grata também aos amigos do DECP/DESC e COM.

Agradeço ainda aos professores Nady, Darlan e Protásio das disciplinas, cursadas no mestrado, pela oportunidade, flexibilização e compreensão diante a minha rotina de trabalho.

Por fim, agradeço a todos aqueles que, mesmo de maneira indireta, contribuíram na minha vida durante esse tempo e torceram pelo meu sucesso.

RESUMO

Estima-se que em 2018 cerca de 310 TWh foram destinados a alimentação de ligações e medições irregulares no Brasil, aproximadamente R$ 9 bilhões de prejuízo para as distribuidoras. Para a concessionária de estudo, são observadas dificuldades para a detecção de fraudes, devido, principalmente, ao volume de dados e a limitação de encontrar padrões sem uma ferramenta estruturada. Considerando esse cenário, propõe-se o desenvolvimento de uma metodologia automatizada para detecção de fraude em clientes da baixa tensão, não telemedidos, com a utilização de ferramentas de inteligência artificial. Foram extraídas informações do banco de dados da empresa, gerados atributos, selecionadas as principais variáveis e, então, avaliados os modelos. A principal variável proposta compara a média de consumo da unidade com os vizinhos geográficos mais próximos com características de porte semelhantes. Também são propostas variáveis que detectam o momento que houve uma redução de consumo, bem como o percentual, através de cálculos estatísticos. As técnicas de aprendizado de máquina mais utilizadas na literatura foram testadas e, no fim, quatro modelos foram propostos: *Support Vector Machine* para unidades com indicação de suspeita de fraude; Gradiente *Boosting* para unidades residenciais sem suspeita de fraude; *Random Forest* para unidades rurais; Rede Neural Perceptron Multicamadas para as demais classes de consumo. Os modelos foram qualificados e as técnicas selecionadas a partir de um novo indicador, proposto como alternativa as métricas usuais de avaliação, que computa o percentual do benefício de energia teoricamente recuperada pelo modelo em relação a toda a energia que poderia ter sido recuperada. Em testes teóricos, foi possível obter uma efetividade de 39,4%, ultrapassando 19,5% a metodologia atual da empresa, com uma cobertura 69,8% maior. O indicador de benefício evidencia também que o método apresentado foi capaz de recuperar 59,5% de todo montante de energia disponível, 153,2% superior ao modelo da empresa. Novas pesquisas envolvem a aplicação da metodologia proposta a base da empresa para classificação das unidades e envio de inspeções para verificar o resultado do trabalho em campo.

Palavras-chave: Perdas comerciais. Perdas não técnicas. Fraude de energia. Classificação de padrões. Aprendizado de máquina. Inteligência Artificial. Detecção de fraude. Recuperação de consumo.

ABSTRACT

In 2018, estimates that about 310 TWh were destined to supply irregular connections and measurements in Brazil, approximately R$ 9 billion losses for distributors. The concessionaire of this study faces challenges to detect fraud, mainly due to the volume of data and the limitation on finding patterns without a structured tool. Considering this scenario, the development of an automated methodology is proposed to detect fraud in low voltage customers, without telemetry, using artificial intelligence tools. Information was extracted from the company's database, attributes were implemented, the main variables were selected and then the models were evaluated. The main variable proposed compares the average consumption of the unit with the closest geographic neighbors with similar size characteristics. Variables are also proposed aiming to detect the moment of a reduction in the energy consumption, as well as its value. The most common Machine Learning techniques were tested and four models were proposed: Support Vector Machine was used for consumers with an indication of possible fraud; for residential units without this indication, Gradient Boosting was used; for rural units, Random Forest was used; for the other classes, a Multilayer Perceptron Neural Network was used. The models were qualified based on a new metric, proposed as an alternative to the usual evaluation metrics, which computes the percentage of the energy benefit theoretically recovered by the model in relation to all the energy that could have been recovered. In theoretical tests, it was possible to obtain an accuracy of 39.4%, surpassing 19.5% the current methodology of the company, with 69.8% greater recall. The energy benefit metric also shows that the proposed methodology was able to recover 59.5% of the total amount of energy available, 153.2% higher than the company's current model. New research involves the application of the proposed methodology to the company's base for the classification of the consumers and inspections will be sent to verify the results.

Keywords: Commercial losses. Non-technical losses. Energy fraud. Pattern classification. Machine learning. Artificial intelligence. Fraud detection. Consumption recovery.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Neste capítulo, inicialmente será apresentada a motivação do trabalho a ser desenvolvido. Em seguida, será exposto o mapeamento realizado do estado da arte para entendimento do histórico do tema e os últimos feitos na área. Finalmente, são mostradas as contribuições do trabalho, bem como os objetivos geral e específicos traçados.

1.1 Motivação

Uma distribuidora de energia elétrica deve fornecer conexão, atendimento e entrega efetiva de energia aos consumidores (ABRAADE, 2018a). A distribuição é a etapa final do fornecimento de energia, ligada ao subsistema de transmissão através de subestações. As redes podem ser compostas por linhas de: alta, em que a tensão entre fases é superior ou igual a 69 kV; média, em que a tensão entre fases é superior a 1 kV e inferior a 69 kV; e baixa tensão, em que a tensão entre fases é igual ou inferior a 1 kV (ANEEL, 2018).

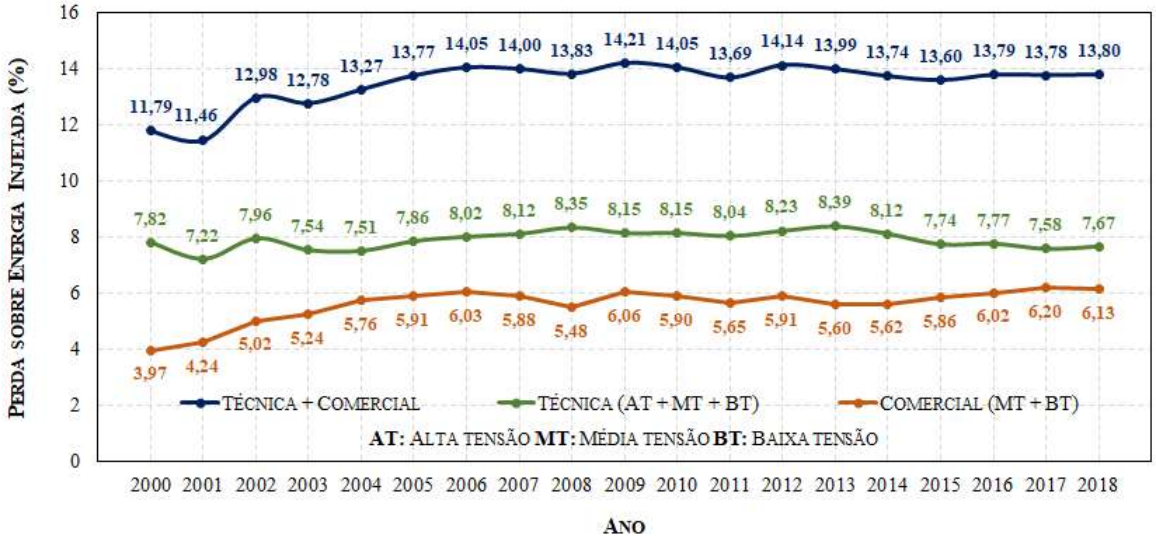
A Agência Nacional de Energia Elétrica (ANEEL) normatiza e padroniza as atividades das distribuidoras através dos Procedimentos de Distribuição de Energia Elétrica no Sistema Elétrico Nacional (PRODIST) e das resoluções normativas que têm por objeto o estabelecimento de diretrizes, obrigações, condições, regras, procedimentos ou quaisquer direitos e deveres dos agentes e consumidores. Os preços de uma distribuidora também são regulados pela ANEEL, que estabelece valores máximos permitidos de serem aplicados pelas empresas através das revisões tarifárias (ABRAADE, 2018a). Os reajustes das tarifas levam em conta os investimentos em infraestrutura, eficiência na gestão dos custos, níveis mínimos de qualidade, ganhos de escala e a variação inflacionária. Dessa maneira, existe um incentivo para as distribuidoras serem mais eficientes (ABRAADE, 2018b). A tarifa de energia elétrica dos consumidores exclusivos da concessionária é constituída basicamente pelos custos com a aquisição de energia, custos relativos ao sistema de distribuição e transmissão, encargos, impostos e perdas. Dessa composição tarifária, os custos relativos à perda de energia possuem um diferencial: observada ineficiência da gestão da distribuidora, o repasse das perdas devido a fraude e furto na conta de energia é limitada (ANEEL, 2015).

A fraude e o furto de energia são algumas das temáticas em destaque nas empresas de distribuição de energia. Elas compõem uma parcela da chamada perda global ou perda na distribuição, definida como a energia que é comprada ou gerada pela distribuidora, mas que não

chega a ser comercializada (ANEEL, 2015). A perda global pode ser dividida em técnica e não-técnica. A perda técnica refere-se ao montante dissipado no sistema decorrente das leis físicas relativas ao processo de transporte, transformação e medição de energia (ANEEL, 2018). Já a perda não técnica, também denominada por perda comercial, refere-se a todas as demais perdas associadas à distribuição, e decorre, principalmente, da fraude, do furto e de erros de medição (ANEEL, 2018).

Na Figura 1, é possível verificar o percentual de perda em relação a energia injetada das distribuidoras do Brasil. Nota-se que houve um crescimento da perda comercial nos últimos 18 anos, enquanto a técnica possui menos variações no mesmo período. Em 2018, o consumo de energia elétrica das distribuidoras foi de aproximadamente 310 TWh, enquanto a tarifa média de fornecimento, desconsiderando os tributos, foi de R$ 474,99/MWh (ANEEL, 2019). Para uma perda comercial de 6,13%, a perda de receita anual no Brasil foi superior a R$ 9 bilhões.

Figura 1 - Percentual de Perdas em relação à energia injetada no sistema das distribuidoras do Brasil.



Fonte: ABRADEE (2018c). Adaptado pelo autor.

Como citado anteriormente, as perdas técnicas são inerentes ao sistema e repassados em sua totalidade para a tarifa de energia. Já para a perda comercial, os níveis utilizados na tarifa são determinados através de modelos estatísticos que correlacionam essas perdas às características socioeconômicas de cada área de concessão (INSTITUTO ACENDE BRASIL, 2017). De fato, existe um reconhecimento do órgão fiscalizador que parte dessa perda não depende exclusivamente da distribuidora. De maneira geral, a metodologia utilizada para determinar o nível máximo de perda não técnica busca estabelecer valores que sejam compatíveis com a região, mas que, ao mesmo tempo, motivem as empresas a se empenharem no combate às perdas comerciais (INSTITUTO ACENDE BRASIL, 2017). Vale salientar que, além da redução do valor da tarifa de energia, combater essas perdas traz benefícios que incluem

o rateio dos custos com a energia suprida pelas geradoras, com o serviço de transmissão, com os encargos, com os tributos, reduz o consumo inconsciente e melhora a qualidade do fornecimento (ANEEL, 2015).

As perdas comerciais podem decorrer de furtos, fraudes, impedimentos de leitura e falhas de medição ou de faturamento. As falhas são consideradas responsabilidade da distribuidora e podem decorrer de um defeito nos equipamentos de medição, ausência ou equívoco da leitura de energia e erros sistêmicos. Os procedimentos a serem tomados pela empresa para esses casos são definidos pelos artigos 113 e 115 da resolução normativa nº 414. Já os procedimentos irregulares, que incluem a fraude e o furto, são definidos pelos artigos de 129 a 133 que citam as providências necessárias para caracterização e apuração do consumo não faturado. O termo furto é utilizado quando uma unidade consumidora é ligada de maneira irregular diretamente à rede da distribuidora, enquanto fraude refere-se à eliminação ou redução do consumo faturado através da adulteração do medidor ou de desvios de energia.

A perda de receita devido aos procedimentos irregulares é um dos principais focos das distribuidoras devido aos prejuízos que ela acarreta. Para identificar uma fraude ou um furto, é necessário enviar uma equipe técnica ao local para que seja feita uma inspeção. Entretanto, cada deslocamento e fiscalização gera custos a empresa que podem chegar a ter até 8 milhões de consumidores (ANEEL, 2019). Dessa forma, essas companhias buscam aprimorar as técnicas de identificação de fraude a fim de otimizar as inspeções, recuperando mais energia para cada visita realizada às unidades consumidoras. Normalmente, alguns indicadores estatísticos são acompanhados para determinar a produtividade das inspeções, como a efetividade – percentual de irregularidades em relação a quantidade de inspeções – e a energia recuperada por visita.

O Decreto-Lei 2.848 de 1940 do Código Penal classifica os crimes que resultam em perdas comerciais em dois tipos principais: furto ou estelionato (fraude) (INSTITUTO ACENDE BRASIL, 2017). O furto, quando pode ser identificado sem a presença de peritos criminais, possui como a pena reclusão de um a quatro anos e multa; quando não, a pena é dobrada. A fraude é classificada como estelionato, com pena prevista de reclusão de um a cinco anos e multa.

1.2 Estado da Arte

Em Viegas et al. (2017) foi elaborada uma revisão bibliográfica dos trabalhos científicos na área de detecção de perdas não técnicas. O artigo teve como objetivo analisar quais linhas de pesquisa foram adotadas e quais as limitações ainda existentes. Dos 103 estudos selecionados, 6 foram teóricos, 25 propuseram soluções de *hardware* e 72 propuseram

aplicações sem *hardware*. Para os estudos teóricos e de *hardware*, os autores citam as limitações de uma baixa precisão de detecção de perdas não técnicas ou de um capital significativo de despesas para a distribuidora. Entretanto, essas soluções podem se tornar viáveis para identificar e combater perda em locais considerados críticos. As soluções sem *hardware* compreenderam a maior parte dos estudos, por serem soluções mais acessíveis e por aproveitarem e transformarem as informações dos consumidores e medidores em dados para a detecção da probabilidade de um comportamento ilegal.

As técnicas mais comuns utilizadas para detecção de perda comercial em soluções sem *hardware* envolvem métodos de classificação como *Support Vector Machine* (SVM), Perfil de Carga, Redes Neurais Artificiais (RNA) e Árvores de Decisão, nessa ordem. Esses classificadores são capazes de inferir um indicador binário ou uma probabilidade da presença de perdas a partir de um conjunto de entradas. O uso dessas técnicas geralmente consiste no processamento de dados de entrada, ajuste do modelo de classificação aos dados, avaliação do desempenho e implantação do modelo. Como fonte de dados, a maior parte dos artigos utilizou informações de consumo de energia, do perfil do consumidor, da carga, tensão e correntes medidas e dos resultados de inspeção. Grande parte também fez combinações das variáveis, utilizando o consumo junto com a informação do consumidor ou com os resultados de inspeção.

Algumas limitações foram encontradas por Viegas et al. (2017) para as soluções de *software*. Parte delas pressupõe que a presença de perdas não técnicas resulta em uma mudança nas informações de consumo coletadas de um cliente. Entretanto, se a solução considera apenas a evolução do consumo, ela não será adequada para detectar irregularidades presentes desde o primeiro dia da ligação elétrica, ou que foram inseridas para desviar uma nova carga. Além disso, a maior parte das soluções dependem de dados de consumo de alta resolução, com medições diárias ou até mesmo horárias de consumo, demanda, tensão e corrente, que utilizam equipamentos mais avançados de medição e terão altos custos associados se a infraestrutura não estiver implementada. Viegas et al. (2017) também cita como as técnicas são muito dependentes de dados, mas não há uma análise sobre o efeito no desempenho do modelo ao utilizar diferentes tipos de amostras, variáveis, atrasos na coleta e diferentes soluções.

Outro problema associado aos trabalhos estudados por Viegas et al. (2017) é a falta de padrão na maneira de avaliar as técnicas utilizadas nas pesquisas. Isso dificulta a comparação dos métodos propostos e a real eficácia das soluções em um cenário real.

Neste trabalho, selecionaram-se artigos para revisão de literatura por suas contribuições para o tema de perdas comerciais e para a metodologia desta pesquisa. Através do estudo do estado da arte, foi possível observar a carência de trabalhos que incluíssem a comprovação da

aplicabilidade e da efetividade dos métodos em cenários reais, validando-os através de inspeções em campo. A maior parte deles limitou-se a uma base de teste teórica que não necessariamente refletiria os resultados obtidos em um cenário real. Os artigos de Nagi et al. (2010) e Guerrero et al. (2014) são exceções para esse caso.

Em Nagi et al. (2010), os autores objetivavam a utilização de técnicas de mineração de dados e classificação de padrões para detectar e identificar padrões de consumo em unidades com fraude. O método utilizado foi o classificador SVM juntamente com um algoritmo para otimizar seus parâmetros. Através de técnicas de extração de variáveis, o conjunto de informações selecionado para compor o sistema de identificação de irregularidades foram dados de consumo mensal normalizados pela quantidade de dias de faturamento, representando a média diária de consumo no mês, para 24 meses e um parâmetro do sistema da empresa que identifica clientes que, intencionalmente, evitam o pagamento das contas. Dos consumidores considerados no banco de dados, clientes sem consumo (0 kWh), que cessaram o contrato com a distribuidora ou que solicitaram ligação nova foram descartados. Além disso, os valores de consumo mensais foram tratados e filtrados retirando estimativas feitas via sistema e inconsistências. No final, o conjunto de dados considerado por Nagi et al. (2010) possuía cerca de 33 mil consumidores com fraude, com uma média de reincidências de 3,2 vezes.

Para identificar perfis de consumo suspeitos, foi feito um estudo a partir da construção de um classificador SVM binário para separar a curva de carga dos consumidores em dois tipos. O primeiro – classe 1 – consistia em unidades com fraude que apresentavam degrau de consumo. Já o segundo, unidades sem fraude que não apresentavam degrau de consumo – classe 2. Para treinar esse algoritmo, foram solicitadas inspeções em unidades com e sem histórico de fraude para segregação nas classes 1 e 2. No total, 383 consumidores foram utilizados para construir esse classificador. Como a razão entre as duas classes é desequilibrada, o classificador SVM foi ponderado para equilibrar a proporção da amostra. Assim, os pesos foram ajustados dividindo o número total de amostras do classificador pelas amostras individuais da classe.

A precisão do método utilizado foi estimada otimizando os parâmetros do SVM a partir de um *grid search*, que é simplesmente uma busca exaustiva em um subconjunto especificado. A acurácia, ou seja, o índice total de acerto entre o indicado e o real foi de 86,43% e a efetividade teórica, ou seja, a taxa de acerto de fraude dentre as indicadas foi de 77,41%. Um pós-processamento também foi realizado, em que se integrou a classificação do SVM com as informações de histórico de contas e reincidências de fraude já citadas. Em campo, obteve-se uma efetividade de 26% separados em 7% de anormalidade e 19% de atividades fraudulentas.

As anormalidades incluíam medidores trocados, casas abandonadas, mudanças de titularidade e defeitos na fiação do medidor.

Para melhorar a efetividade do método, Nagi et al. (2010) propôs selecionar apenas os clientes com as maiores probabilidades de fraude a partir de tomadas de decisões com base em valores de parâmetros das variáveis que compõem o sistema. Os valores desses parâmetros foram determinados pela inspeção de perfis de carga de clientes com fraude, já identificados anteriormente, analisando as características comuns que diferenciam os casos normais dos casos com atividades de fraude. A efetividade após implementado o sistema de tomada de decisão aumentou de 26% para 64%. Não foi citada a quantidade de unidades disparadas para inspeções, o que limita o conhecimento sobre o alcance da metodologia proposta.

Além da ausência de informação sobre a cobertura do método, uma limitação citada pelo próprio autor é que o classificador não identifica irregularidades com mais de 2 anos. Essa pesquisa também inclui algumas das deficiências citadas por Viegas et al. (2017), já que ela necessita que haja uma alteração no comportamento do consumo para identificar uma fraude. Há também uma alta dependência com uma variável que aponta riscos de fraude devido a reincidência, outro indicativo de que a metodologia possui um baixo percentual de cobertura.

Outro trabalho que incluiu resultados de inspeção foi o apresentado por Guerrero et al. (2014). Nele foi utilizado um sistema baseado em conhecimentos e em *text mining* para identificar perdas não técnicas. O conjunto de regras foi montado com base em entrevistas realizadas com os melhores inspetores de uma distribuidora de energia da Espanha. O objetivo principal era desenvolver um sistema para automatizar o processo manual de inspeção.

Para construir a base de regras, foram implementados os seguintes procedimentos para aquisição de conhecimento: entrevista pessoal, entrevista estruturada em objetivos, observação da técnica de trabalho dos especialistas e observação do protocolo de trabalho. Após as entrevistas, um resumo do conhecimento adquirido foi escrito e enviado para aprovação dos inspetores. Além da extração de regras a partir dessa base de conhecimento, no processo de análise manual de uma inspeção é feita uma avaliação para determinar se o consumo do cliente está coerente. Dessa maneira, dados de demanda contratada, localização geográfica, atividade econômica e estação do ano também foram incluídas no sistema proposto. Por fim, os autores utilizaram técnicas de processamento de linguagem natural para extrair informações não estruturadas dos comentários dos inspetores e estruturá-las em quatro categorias: correto, incorreto, consumo baixo ou fechado.

Todas as informações citadas foram convertidas em regras do sistema baseado em conhecimentos. Ao todo, foram geradas 177 regras. Elas foram aplicadas a um conjunto de

50.014 consumidores, em que 5.136 foram indicados como tendo problemas, mas apenas 2.403 exigiam inspeção para confirmar. Para teste em campo, um conjunto de 116 consumidores foi selecionado com base na quantidade de regras que foram apontadas pelo sistema. O resultado da inspeção retornou 10 casos de fraude, 7 defeitos e 20 anomalias sem perda, o que computa 32% de efetividade total, mas apenas 15% para casos com perda não técnica.

Os trabalhos de Nagi et al. (2010) e de Guerrero et al. (2014) demonstram a dificuldade de se obter altos índices de efetividade quando as metodologias propostas são aplicadas em campo. O tema de perdas comerciais se enquadra em problemas de classificação com desbalanço de dados, já que a proporção de fraude no conjunto de total é muito baixa. Angelos et al. (2011) analisou como o percentual de unidades fraudadoras na base de teste pode afetar o resultado de um modelo. Nessa pesquisa, foi utilizado um algoritmo de clusterização do C-Means baseado em lógica Fuzzy para encontrar consumidores com perfis de consumo semelhante. As variáveis utilizadas incluíram o consumo médio, máximo, mínimo e o desvio padrão da curva, além da quantidade de inspeções e a média de consumo na área residencial do cliente. Após separadas as unidades em grupos, ou clusters, uma classificação Fuzzy foi feita para identificar possíveis fraudadores ou padrões irregulares de consumo. Para avaliar o algoritmo, Angelos et al. (2011) realizou algumas análises variando, por exemplo, a classe de consumo, a distribuição de casos irregulares e a influência da sazonalidade na base de teste. Quando utilizadas porcentagens maiores de amostras irregulares, os autores perceberam que o método apresentava uma maior assertividade. Para um percentual de 90% de casos anormais no banco de teste, o modelo obteve 97,7% de efetividade e 2,5% de cobertura, enquanto para um percentual de 10%, ele obteve 20% de efetividade e 5,2% de cobertura. Esse resultado ilustra como os testes teóricos podem divergir muito dos práticos, já que, usualmente, utiliza-se uma base balanceada para verificar a eficácia do método.

O tema de combate às perdas não técnicas continua sendo intensamente pesquisado nos dias atuais. Em Ramos et al. (2018) foi proposta a detecção de clientes irregulares a partir de uma técnica de otimização meta-heurística chamada algoritmo do buraco negro, ou *Black Hole Algorithm*, utilizando dados de consumo, demanda e contrato como variáveis do modelo. Em Zheng et al. (2018), uma rede neural convolucional foi proposta com dois componentes para identificar roubo de energia através da curva de consumo semanal em duas dimensões: uma referente a periodicidade do consumo e outra referente aos aspectos globais da curva. Nesses trabalhos, foram utilizadas informações normalmente associadas a clientes da média e alta tensão, como por exemplo, demanda contratada, consumo horário e fator de carga. Na prática, a maior parte dos consumidores ainda não possui modelos de medição com monitoramento em

bases menores que mensal, o que torna a identificação de perdas mais desafiadora. Araujo et al. (2019), por sua vez, utilizou uma Rede Neural Artificial para determinar a probabilidade de existir uma irregularidade em uma unidade consumidora através de variáveis estatísticas de consumo mensal e observações de leituristas apontadas durante a leitura de energia. Nota-se que, mesmo na literatura mais recente, os desafios apontados por Viegas et al (2017) ainda são válidos. A alteração de consumo continua sendo a variável mais frequente para detecção de perda comercial, bem como a falta de padrão para avaliar os modelos dificulta a verificação da real eficácia dos métodos.

Massaferro et al. (2020), mais recentemente, propôs uma solução de aprendizado de máquina para identificação de irregularidades otimizado de modo que o retorno econômico para a empresa seja maximizado. Esse retorno foi calculado considerando tanto a receita recuperada, quanto o custo da inspeção, em que a perda por unidade consumidora é calculada com base, ou na demanda contratada, ou através de um algoritmo de regressão aplicado ao histórico de consumos recuperados, e a despesa é calculada através da curva de custo projetada para diferentes capacidades operacionais assumindo que ele seja proporcional a quantidade de inspeções. Para estimar o volume de energia perdida através de regressão, os autores utilizaram o *Random-Forest-Regressor* e uma Rede Neural. O algoritmo de melhor performance para identificação de fraude considerando a maximização do retorno econômico foi o classificador *Random Forest*, obtendo 15,6% de efetividade e 78,8% de cobertura.

1.3 Contribuições

Considerando o cenário atual de combate as perdas de energia na concessionária em estudo, as contribuições deste trabalho envolvem: a apresentação das etapas de modelagem dos dados que resultaram na construção e identificação das principais variáveis para detecção de fraude; implementação de variáveis que auxiliam no reconhecimento de uma irregularidade sem presumir que há variação de consumo através de comparativos entre a unidade consumidora e seus vizinhos com características semelhantes; proposta de uma nova métrica de avaliação de modelos com foco em representar o retorno econômico, com base na recuperação de energia e custo de inspeção por unidade consumidora, com objetivo de aprimorar a forma de avaliação dos algoritmos; avaliar diferentes técnicas de aprendizado de máquina através das principais métricas da literatura, como a matriz de confusão, a efetividade, a cobertura e o F-score, e do novo indicador proposto, selecionando os melhores modelos que produzam uma lista de clientes a serem inspecionados priorizando o maior retorno econômico para a distribuidora; fornecer

uma metodologia automatizada de detecção de fraude que aumentou o percentual de efetividade e cobertura em relação à metodologia usual da distribuidora.

Enfatiza-se que são descritos todos os passos desde a obtenção das informações no banco de dados da empresa, passando pelo tratamento e geração da base, extração de atributos, seleção das variáveis, até a aplicação dos métodos.

1.4 Objetivos

Este trabalho visa desenvolver uma metodologia com base em aprendizado de máquina para detecção de fraude em clientes da baixa tensão, não telemedidos, em uma empresa de distribuição de energia elétrica do Brasil, com a finalidade de automatizar o processo, melhorar a assertividade e identificar padrões não observados pelo método de análise atual da empresa. Para isso, como objetivos específicos, pretende-se:

a) extrair as variáveis mais relevantes para identificar uma fraude a partir de um banco de dados histórico;
b) propor uma nova métrica de avaliação de modelos priorizando o retorno econômico;
c) fazer um comparativo entre diferentes técnicas de aprendizado de máquina para o reconhecimento de padrões e escolher a combinação que oferece a maior recuperação de energia;
d) aumentar a efetividade das inspeções em campo em relação a metodologia atual da distribuidora.

O trabalho está dividido conforme descrito a seguir: no capítulo 2 estão especificados os principais conceitos que fundamentam e facilitam a compreensão da metodologia proposta; o capítulo 3 detalha a metodologia utilizada, que envolve a preparação e modelagem dos dados, bem como a aplicação dos modelos; o capítulo 4 apresenta e discute os resultados obtidos; o capítulo 5 trata das conclusões e trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Neste tópico, serão apresentados os principais conceitos utilizados neste trabalho, sendo possível classificá-los em definições sobre perda comercial, estatística e aprendizado de máquina. São apresentadas definições fundamentais sobre perdas tendo como referência principal as resoluções normativas da agência reguladora e os conhecimentos de especialistas da distribuidora de estudo deste trabalho.

2.1 A Unidade Consumidora Não Telemedida do Grupo B

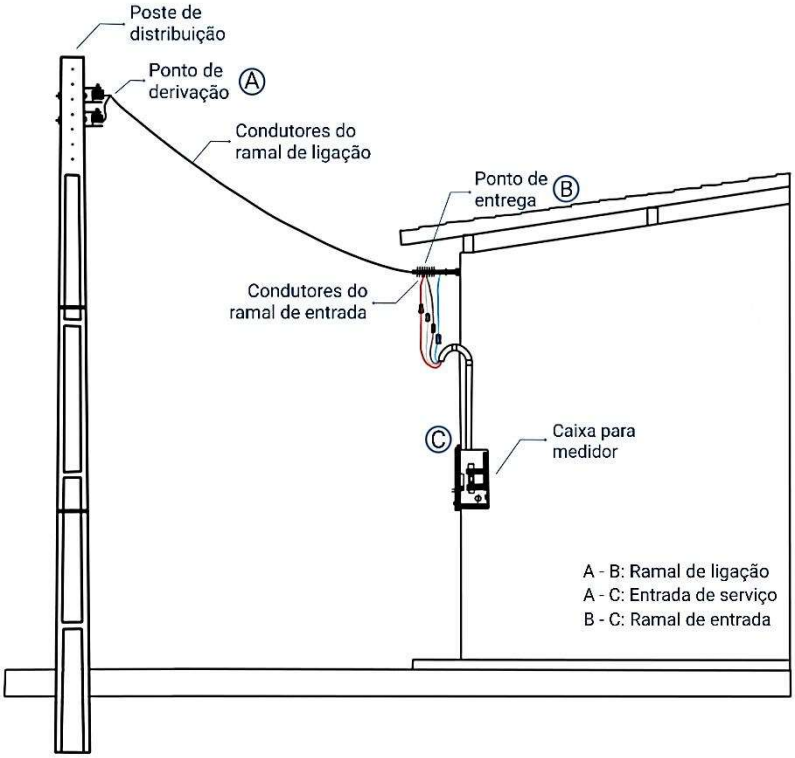
Uma Unidade Consumidora (UC) compreende o conjunto de instalações e equipamentos destinados ao recebimento de energia em um só ponto de entrega com medição individual correspondente a um único consumidor. Os direitos e deveres dos consumidores e das distribuidoras são regidos pela resolução normativa nº 414, que descreve os procedimentos e condições gerais de fornecimento de energia elétrica (ANEEL, 2010).

Cada consumidor com interesse em requisitar uma ligação e se conectar ao sistema elétrico deve fazer uma solicitação de fornecimento junto a concessionária da região. Para efetivar a ligação, é necessário que o padrão de entrada de energia esteja de acordo com as normas da distribuidora e seja fornecida a documentação pertinente. Uma UC é composta por instalações, ramal de entrada, equipamentos elétricos, condutores e acessórios, incluindo a subestação para fornecimento em tensão primária. O consumidor é responsável pelo zelo do ramal de entrada, caixa de medição, poste, dispositivos de proteção e de equipamentos mantidos sob lacre. O ponto de entrega define o ponto de conexão que a concessionária deve fornecer energia elétrica, caracterizando o limite de responsabilidade do fornecimento. Com exceção de casos excepcionados em norma, a tensão de fornecimento para a unidade será determinada de acordo com a carga instalada. Para cargas instaladas iguais ou inferiores a 75 kW, a tensão de fornecimento será secundária com valores padronizados inferiores a 2,3 kV (ANEEL, 2010). O grupamento das unidades pertencentes a esse nível de tensão é denominado de Grupo B (GB). Para cargas superiores a 75 kW, a tensão de fornecimento será primária e o grupamento é denominado Grupo A (GA). Normas técnicas da distribuidora definem se o fornecimento será por meio de ligação monofásica, bifásica ou trifásica.

O padrão de medição e fornecimento de um cliente do Grupo B com ligação direta, ou seja, que não utiliza transformadores de instrumentos e, portanto, o medidor é ligado

diretamente no circuito entre a fonte e a carga, pode ser observado na Figura 2. A entrada de serviço pode ser instalada em poste auxiliar, muro, mureta, pontalete, parede, ou mesmo ser subterrânea. O ramal de entrada pode ser aéreo ou subterrâneo com a caixa de medição interna a propriedade do cliente, no limite do terreno com o visor voltado para a via pública ou externa fixada em mureta, muro ou poste auxiliar na divisa do lote. Em geral, as concessionárias solicitam que a proteção do ramal de saída seja instalada após os equipamentos de medição. O medidor pode ser do tipo eletromecânico, considerado obsoleto, ou eletrônico.

Figura 2 - Exemplo de padrão de medição para grupo B direto.



Fonte: Elaborado pelo autor.

O padrão de medição pode ser alterado por solicitações do cliente ou a partir de exigências da própria distribuidora. Essas exigências, em geral, visam combater os impedimentos de leitura e os procedimentos irregulares através da externalização do padrão da medição, utilização de Caixas Padrão Rede (CPRede), blindagem dos clientes e/ou instalação do Dispositivo de Lacre do Compartilhamento de Borne (DLCB) nos medidores. As CPRedes, Figura 3, são caixas de medição que possuem uma lente de aumento instaladas em poste junto a via pública. Possuem uma tampa ou porta que dispõe de dispositivo de selagem e segurança para fechamento, além de vedação para evitar a penetração de água. As lentes de aumento facilitam a leitura do consumo, reduzindo os impedimentos devido à ausência do cliente no imóvel (QUEIROZ JR. et al, 2000). A utilização desses padrões de caixa tende a facilitar as

fiscalizações das unidades consumidoras, entretanto, nos dias atuais, evita-se sua prática devido a outros problemas que seu uso acarreta, como lentes embaçadas e fora de foco. O DLCB, Figura 4, refere-se ao dispositivo instalado nos bornes dos medidores que funciona como um lacre de segurança dificultando desvios de energia através do by-pass, em que se conecta a carga direto à fonte (BRITO, 2002). A blindagem de circuitos e clientes, Figura 5, consiste em instalar uma proteção mecânica na rede multiplex através de malha metálica, borracha vulcanizada e isolante conectando o cliente através de conector específico a aproximadamente 2 m de distância do poste e utilizar uma caixa blindada no padrão de medição do cliente.

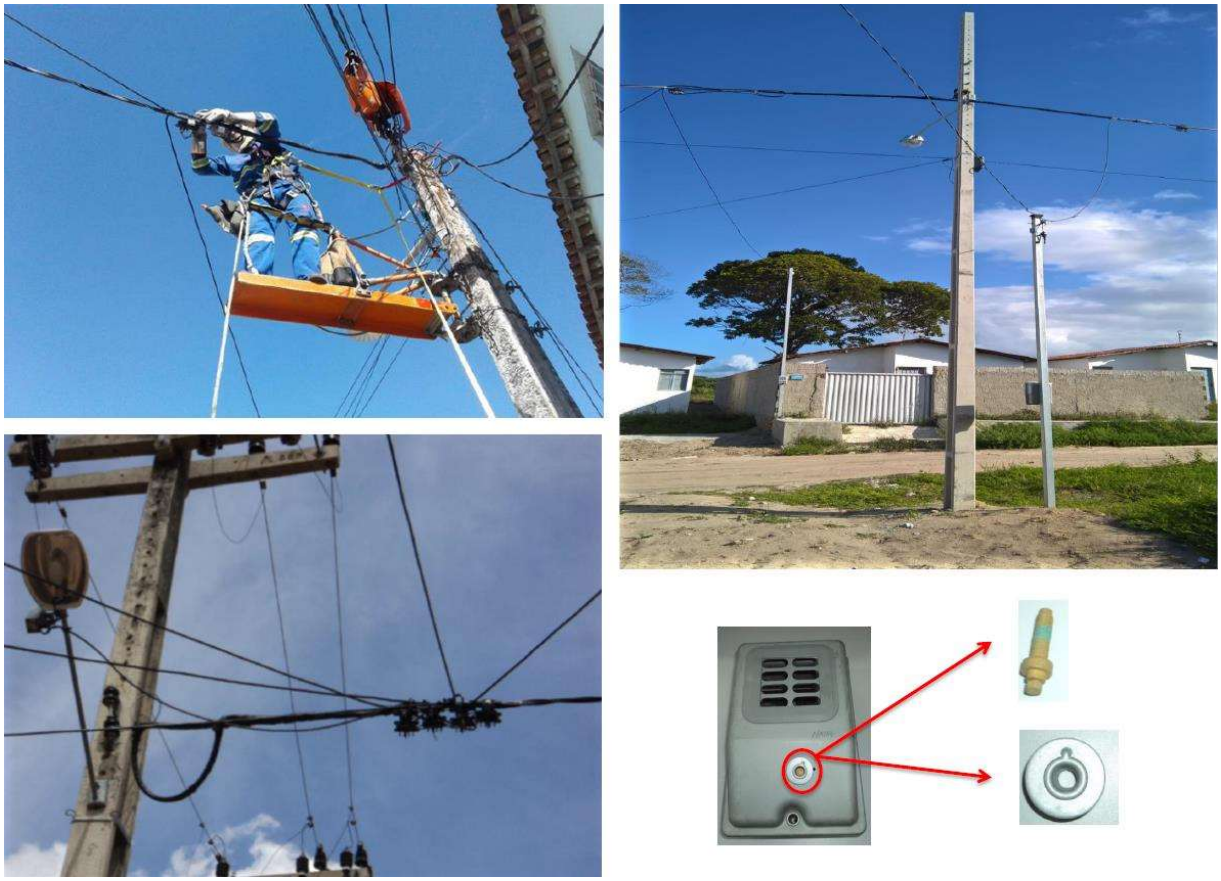
Figura 3 - Padrão externo com Caixas Padrão Rede (CPRede).

Fonte: Elaborado pelo autor.

Figura 4 - Medidor com Dispositivo de Lacre do Compartilhamento de Borne (DLCB).

Fonte: Elaborado pelo autor.

Figura 5 - Blindagem de rede através da proteção mecânica e caixa de medição blindada.



Fonte: Elaborado pelo autor.

Para fins de aplicação tarifária, as UCs podem ser classificadas de acordo com a atividade e a finalidade de utilização da energia elétrica. Conforme a resolução normativa nº 414, as classes de consumo estão indicadas no Quadro 1.

Desde 2012, os consumidores podem gerar a própria energia elétrica através de fontes renováveis ou cogeração qualificada para potências instaladas de até 5 MW (ANEEL, 2012). Os beneficiários da energia gerada podem incluir a própria UC com micro ou minigeração, integrantes de múltiplas unidades consumidoras, geração compartilhada e unidades do mesmo titular situadas em outro local dentro da mesma área de concessão. O total da energia ativa consumida é descontada da energia injetada do mês podendo existir um excedente que se mantém como crédito para abater o consumo da unidade consumidora em até 60 meses. Para essas unidades é necessário que seja instalado um medidor do tipo bidirecional que será capaz de registrar tanto a energia ativa consumida, quanto a energia ativa gerada.

Quadro 1 - Classe de consumo das unidades consumidoras.

Classe	Descrição
Residencial	UCs com fins residenciais. Dividida em seis subclasses: residencial, residencial baixa renda, residencial baixa renda indígena, residencial baixa renda quilombola, residencial baixa renda benefício de prestação continuada da assistência social e residencial baixa renda multifamiliar.
Industrial	UCs que desenvolvem atividades industriais ou realizem o transporte de matéria-prima, insumo ou produto resultante do seu processamento desde que realizado de forma integrada fisicamente à unidade indústria enquadrados de acordo com a Classificação Nacional de Atividades Econômicas (CNAE).
Comercial	UCs onde sejam desenvolvidas atividades de prestação de serviços não previstas nas outras classes. Subdividida em: comercial; serviços de transporte; serviços de comunicação e telecomunicação; associações e entidades filantrópicas; templos religiosos; administração condominial; iluminação em vias; semáforos, radares e câmeras de monitoramento de trânsito; outros.
Rural	Unidades que desenvolvam as atividades dispostas nas seguintes subclasses: agropecuária rural, agropecuária urbana, residência rural, cooperativa de eletrificação rural, agroindústria, serviço público de irrigação rural, escola agrotécnica, aquicultura.
Poder Público	UCs de pessoas jurídicas de direito público, subdividindo-se em: poder público federal, estadual e municipal.
Iluminação Pública	Unidades destinadas exclusivamente a prestação do serviço público de iluminação.
Serviço Público	UCs destinadas ao fornecimento para motores, máquinas e cargas essenciais à operação de serviços públicos de água, esgoto, saneamento e tração elétrica urbana ou ferroviária, explorados diretamente pelo poder público
Consumo Próprio	UCs de titularidade da própria distribuidora.

2.2 Leitura e Faturamento

A cobrança do consumo de energia elétrica é feita através das leituras feitas mensalmente ao medidor de energia elétrica através do leiturista ou de telemetria. O leiturista é o funcionário da distribuidora responsável por visitar as unidades não telemedidas e registrar as leituras exibidas no medidor. As unidades com telemetria são lidas diretamente através de um equipamento que fornece acesso remoto a esses valores.

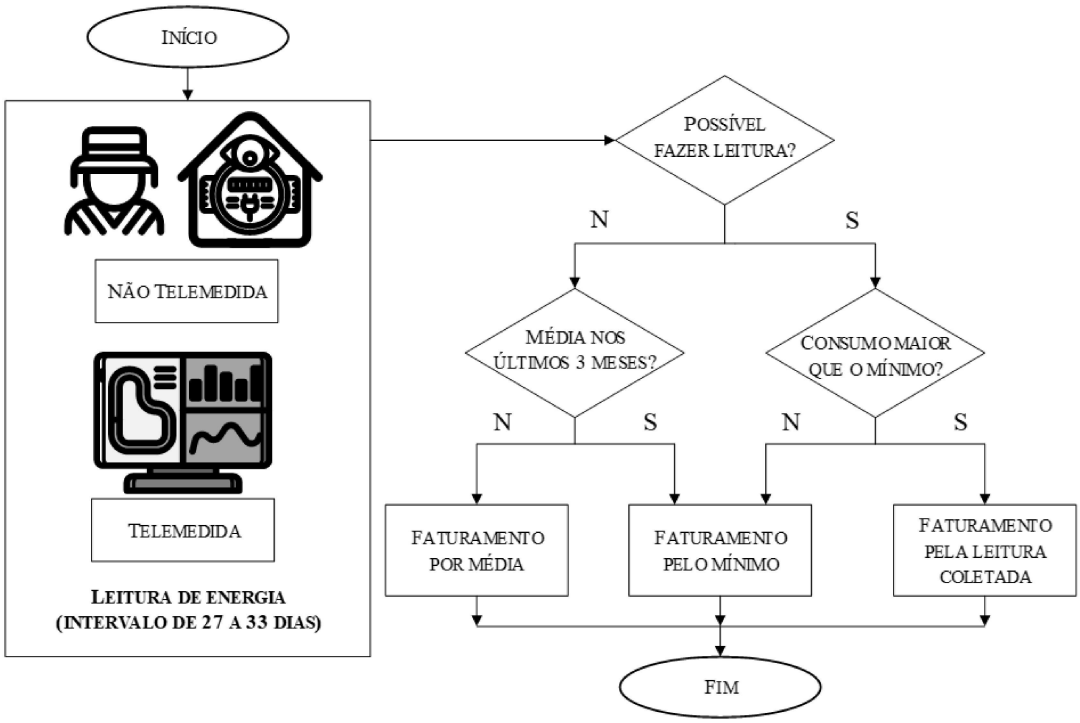
A modalidade tarifária de uma unidade do grupo B pode ser do tipo convencional, caracterizada por ser monômia, isto é, aplicável apenas ao consumo de energia, ou branca, em que são aplicadas tarifas diferenciadas de acordo com as horas de utilização do dia e segmentada em três postos horários: ponta, intermediário e fora ponta. As UCs do grupo B são faturados pelos valores de energia ativa, não sendo permitido cobrar reativos.

As leituras devem ser efetuadas em intervalos de 27 a 33 dias de acordo com o calendário de leitura. As unidades consumidoras são agrupadas em conjuntos com datas de leitura em comum, divididas de modo que seja possível visitar todas as unidades necessárias e obedecer a quantidade de dias máxima para o faturamento do mês. Caso o consumo medido ou

estimado da unidade aplicado ao faturamento do mês seja inferior ao custo de disponibilidade do sistema elétrico, será faturado os valores equivalentes ao próprio custo de disponibilidade. Ele é de 30 kWh para ligações monofásicas, 50 kWh para ligações bifásicas ou 100 kWh para ligações trifásicas. Vale ressaltar que, para beneficiários de micro ou minigeração na baixa tensão, ainda que a energia injetada seja superior ao consumo, essa regra também será aplicada.

Em caso de impedimento de acesso ao imóvel para fins de leitura, será faturada a média aritmética dos últimos 12 meses disponíveis, anteriores a constatação do impedimento. Esse procedimento pode ser aplicado por até 3 ciclos consecutivos de faturamento; após isso deve ser faturado o custo de disponibilidade, também conhecido por faturamento pelo mínimo da fase. Ao normalizar o acesso a unidade é possível realizar um acerto de faturamento. Na Figura 6 é possível observar o resumo dessa rotina de leitura e faturamento.

Figura 6 - Rotina simplificada da leitura e faturamento de uma unidade consumidora.



Fonte: Elaborado pelo autor.

As regras para o faturamento incorreto são semelhantes as aplicadas ao impedimento de acesso. Para acerto dos valores faturados, após a constatação da falha, caso o faturamento tenha sido a maior, devem ser devolvidas ao consumidor as quantias recebidas indevidamente limitadas até 36 ciclos de faturamento; caso o faturamento tenha sido a menor, ou houve ausência de faturamento, é possível cobrar até três ciclos das quantias não recebidas quando o motivo para o faturamento incorreto é de responsabilidade da distribuidora. Se houver comprovação de motivo atribuível ao consumidor, o prazo de cobrança se estende a 36 meses.

Para defeito na medição, os períodos máximos de cobrança e recuperação de consumo se enquadram na responsabilidade da distribuidora, ou seja, limitados a 3 meses. As regras utilizadas para compensação do faturamento, para esse caso, devem seguir os seguintes critérios em ordem de disponibilidade: aplicação do fator de correção do erro de medição; média aritmética dos últimos 12 meses proporcionalizados em 30 dias; faturamento imediatamente posterior à regularização da medição.

Em geral, as distribuidoras determinam códigos em seus sistemas para acompanhar os faturamentos pela média ou pelo mínimo, bem como seus motivos. Denomina-se irregularidade de leitura apontamentos de leituristas na coleta de leitura mensal das unidades. Esses apontamentos podem indicar impedimentos, defeitos na medição ou mesmo suspeitas de manipulação na medição. Denomina-se irregularidade de faturamento códigos gerados automaticamente pelo sistema que indicam que o consumo lido pode ter sido divergente do consumo faturado, ou seja, utilização da média, do mínimo da fase ou acerto de faturamento.

2.3 Procedimentos Irregulares

Os procedimentos irregulares referem-se ao ato ilícito de eliminar ou reduzir o consumo faturado. Para caracterizar a irregularidade e apurar o consumo não faturado, a distribuidora deve levantar o conjunto de evidências através de perícia técnica, avaliação do histórico de consumo e grandezas elétricas, medição fiscalizadora e/ou recursos visuais. Em caso de violação do medidor, é necessário também elaborar um relatório de avaliação técnica. O Termo de Ocorrência e Inspeção (TOI) deve ser emitido e uma cópia deve ser entregue ao consumidor ou a quem acompanhar a inspeção.

Comprovada a irregularidade, para apurar as diferenças não faturadas, a concessionária deve utilizar um dos cinco critérios descritos pela norma 414 aplicados de forma sucessiva: utilizar o consumo de medição fiscalizadora proporcionalizado em 30 dias; aplicar fator de correção obtido por meio de aferição do erro de medição; utilizar a média dos 3 maiores consumos proporcionalizados em 30 dias disponíveis em 12 ciclos de faturamento imediatamente anteriores ao início da irregularidade; determinar o consumo a partir da carga desviada ou carga instalada no momento da identificação da irregularidade; utilizar o maior consumo proporcionalizado em 30 dias nos 3 ciclos imediatamente posteriores a regularização da medição. Caso exista grande variação sazonal do consumo da unidade consumidora, essa condição deve ser considerada.

O período de cobrança deve ser determinado através de análise do histórico de consumo ou outros estudos técnicos que comprovem o início da irregularidade. O prazo máximo de

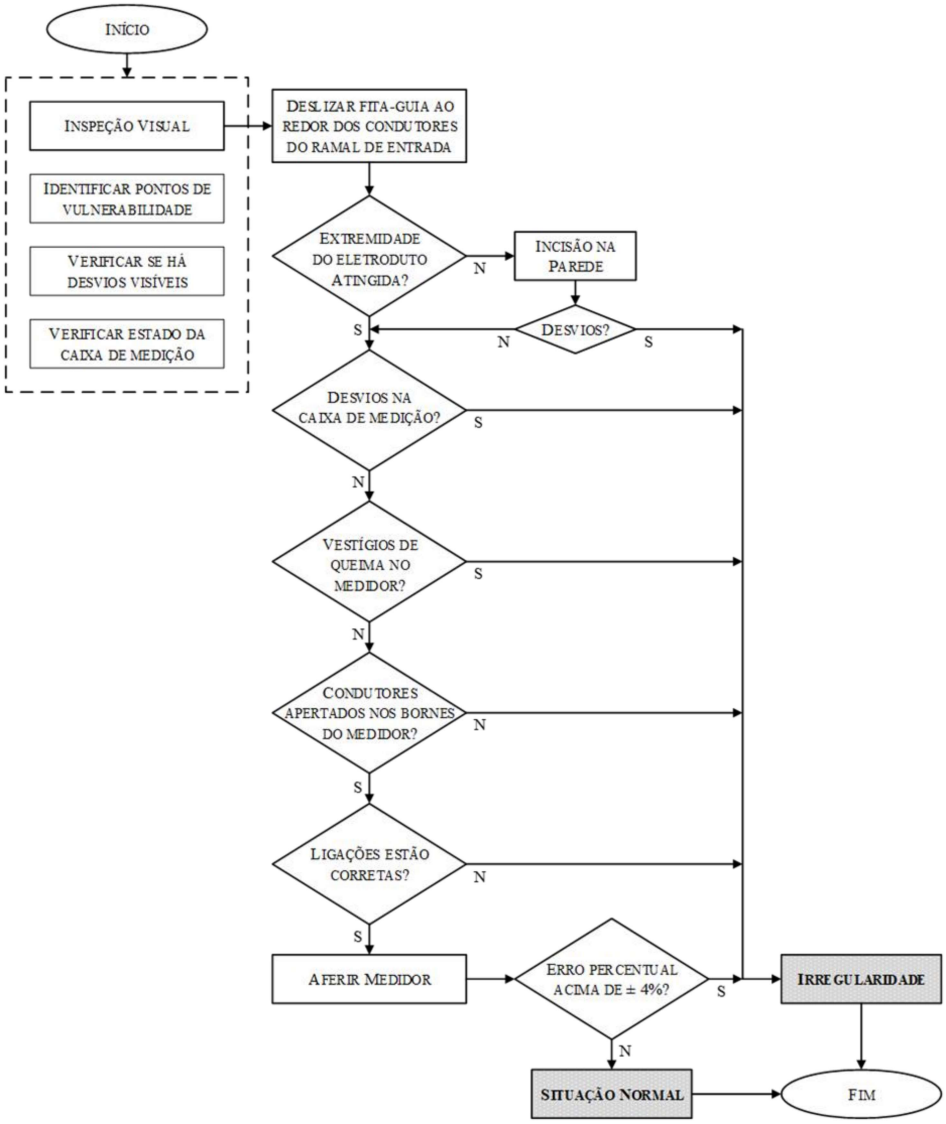
retroativo é de 36 meses. Caso não seja possível determinar o período de duração do procedimento irregular, esse prazo é limitado a 6 ciclos de faturamento. Além disso, restringe-se a retroatividade da recuperação de receita à última inspeção nos equipamentos de medição da unidade consumidora em questão, bem como ao período sob responsabilidade do atual titular.

Os procedimentos irregulares são comumente classificados em fraude e furto de energia. O furto ocorre através da ligação de condutores diretamente à rede de distribuição, caracterizando uma ligação clandestina. Já a fraude pode ser associada a diversas irregularidades e pode depender do tipo de medidor e do tipo de ligação utilizado na unidade consumidora. Alguns casos de fraude citados por Ortega (2008) e Curado (2015) incluem: desvio de cargas para outro potencial em vez de conectadas ao neutro, fazendo com que parte da corrente não circule pelo medidor; inversão dos cabos de fase e neutro; desconexão do neutro do medidor e utilização de outra fonte de aterramento; *by-pass* do medidor ao ligar a carga diretamente à rede da distribuidora sem passar pelos equipamentos de medição; ligação de carga em paralelo ao medidor de maneira que seu consumo não é registrado; elemento móvel do medidor bloqueado por meio de perfuração da caixa e introdução de objeto estranho.

O procedimento para inspecionar uma unidade consumidora a fim de identificar uma irregularidade está resumido na Figura 7.

Ele inicia-se com uma inspeção visual em que se busca observar pontos de vulnerabilidade, presença de desvios no ramal de serviço ou em caixas de passagens e caixa de medição amassada, furada ou com os lacres violados. Em seguida, fiscaliza-se o ramal de entrada, através do uso de fita-guia em que se envolve uma linha de nylon ao redor dos condutores e passa pelo eletroduto até a outra extremidade. Se houver impossibilidade de deslocar a fita, quebra-se a parede a fim de detectar um desvio. Para inspeção da caixa de medição e do medidor, verifica-se a existência de desvios dentro da caixa, o estado dos bornes de entrada e saída, se as fases e o neutro estão corretamente ligados, se há sinais de violação ou adulteração do medidor, como furos ou marcas, além de realizar testes no medidor através da utilização de uma carga e comparando os valores informados pelo medidor com um multímetro ou outros equipamentos de medição de grandezas elétricas. A ausência dos lacres nos medidores, caixas e cubículos corroboram com o indício de irregularidade, visto que estes só podem ser rompidos por representante credenciado da distribuidora (ANEEL, 2010). Quando é necessário retirar o medidor, ele é lacrado em invólucro específico e encaminhado em transporte adequado para realização da avaliação técnica.

Figura 7 - Procedimentos de inspeção de uma UC do GB com medição direta.



Fonte: Elaborado pelo autor.

2.4 Combate às Perdas Comerciais

A perda de energia associada ao consumo não faturado é denominada de perda comercial ou não-técnica. Geralmente as estratégias de combate a essa perda de receita e de contabilização de energia são divididas de acordo com sua origem e podem ser classificados em quatro tipos: defeito ou erro na medição, ausência de medição, fraude e furto.

Para os defeitos e erros de medição, utilizam-se os dados fornecidos pelo leiturista ou pela telemetria para identificar problemas através dos códigos de irregularidade e faturamento citados na seção 2.2. Nesses casos, a urgência da regularização é determinada pelo tempo de recuperação que se encontra limitado a três meses. O procedimento para confirmar o defeito é semelhante ao utilizado na inspeção de fraude, sendo que devem ser utilizados os mesmos critérios para que seja descartada a possibilidade de violação do medidor.

Os casos com ausência de medição podem ser devido a falha da concessionária, normalmente por erros de cadastro, ou devido a impedimentos de leitura. Para esses casos, são utilizadas medidas para eliminar ou evitar esses problemas, como normalizar o cadastro interno, externalizar o padrão da unidade consumidora, fazer limpeza e ajuste de foco das lentes em CPRedes.

Para os furtos de energia, busca-se regularizar diretamente a medição e o cadastro dos clientes fornecendo o padrão aos que não possuem e aplicando blindagem na rede de distribuição para evitar roubo na rede BT. A identificação de clandestinos é feita, normalmente, através de inspeções visuais nas redes da distribuidora. Estudos de balanço de energia podem ser feitos para tentar identificar áreas críticas com maior probabilidade de encontrar essas ligações diretas.

Nos casos de suspeita de fraude, procura-se levantar a maior quantidade de dados que indiquem um comportamento irregular na unidade consumidora. Dentre eles, incluem-se a indicação de leiturista, reduções de consumo e denúncias da população. A indicação do leiturista é dada quando, no momento da leitura, foi percebido algum tipo de irregularidade na UC que indique uma suspeita de fraude. Esse dado é considerado confiável nas distribuidoras, entretanto é limitado a fraudes visíveis que não foram retiradas no momento da leitura, o que limita os casos detectáveis. A redução de consumo pode ser rastreada através de visualizações gráficas ou através do cálculo do degrau conforme equação (1).

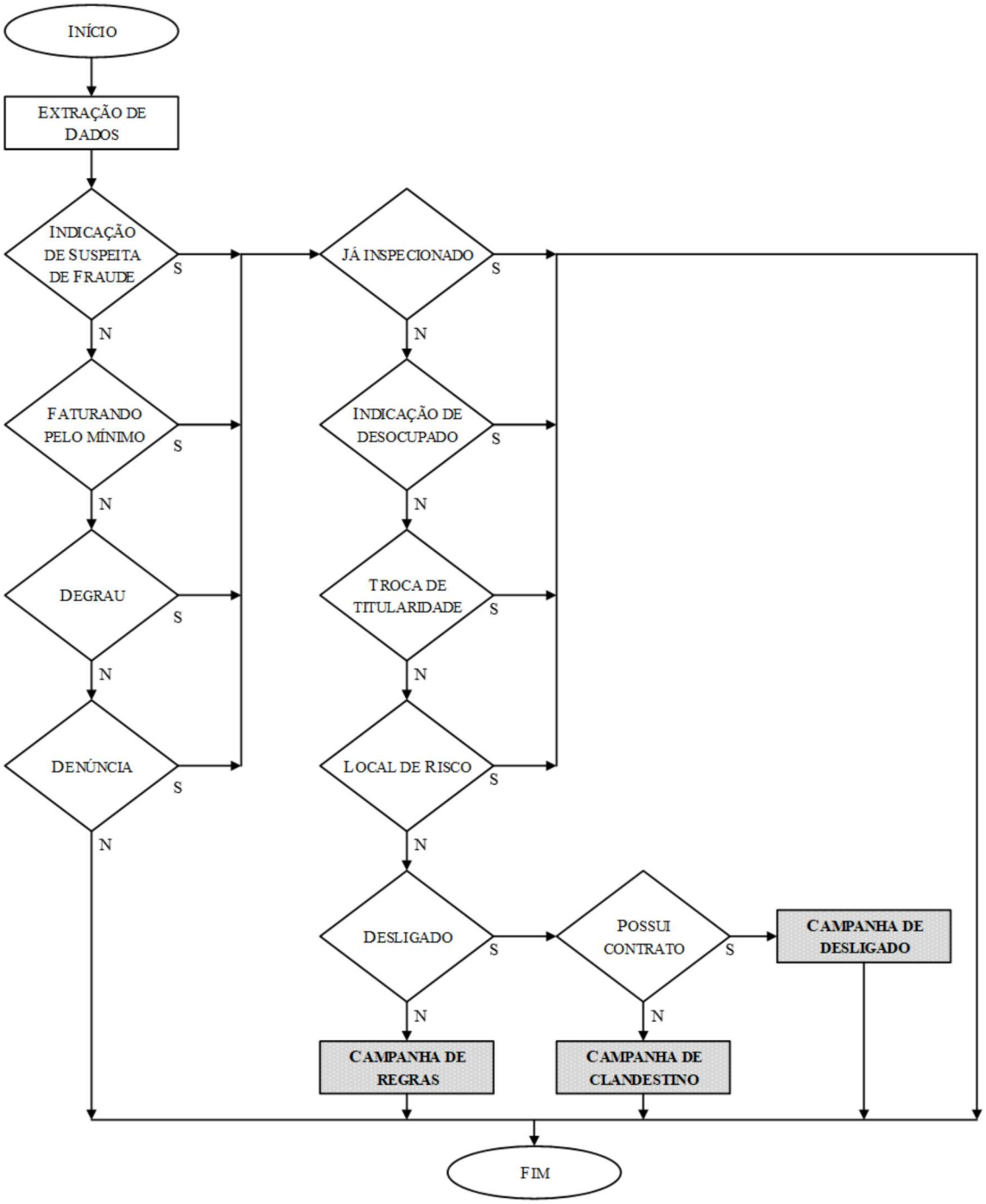
$$\text{degrau} = \begin{cases} \dfrac{\text{cons}_{\text{atual}} - \text{cons}_{\text{ant}}}{\text{cons}_{\text{ant}}}, & \text{cons}_{\text{ant}} \neq 0 \\ 0, & \text{cons}_{\text{ant}} = 0 \text{ e } \text{cons}_{\text{atual}} = 0 \\ 1, & \text{cons}_{\text{ant}} = 0 \text{ e } \text{cons}_{\text{atual}} \neq 0 \end{cases} \quad (1)$$

Em que: $\text{cons}_{\text{atual}}$ é o consumo atual, podendo referir-se a energia do mês ou a média de um intervalo arbitrário; e cons_{ant} é o consumo anterior de referência para comparativo com o $\text{cons}_{\text{atual}}$ devendo ambos estarem na mesma unidade, em geral kWh ou MWh.

Vale salientar que uma redução no patamar de consumo de uma unidade nem sempre é devido a procedimentos irregulares. Residências de veraneio e imóveis de aluguel são exemplos de UCs que podem reduzir drasticamente seu consumo. Outros casos são unidades comerciais e industriais que variam de acordo com a demanda de mercado ou de acordo com manutenções ou reformas. Até mesmo a substituição de equipamentos obsoletos pode resultar em um degrau negativo de consumo. A variável degrau definida em (1) objetiva verificar variações do consumo comparando uma média de consumo atual a uma média de consumo anterior. Os intervalos a serem comparados são arbitrários.

As unidades a serem inspecionadas ou regularizadas são normalmente indicadas por departamentos estratégicos no combate às perdas que levantam os dados, fazem análises e inserem as UCs em campanhas. Campanha é a denominação utilizada para o conjunto de unidades consumidoras selecionadas para: inspeção, quando existe suspeita de procedimentos irregulares; regularização, quando o caso é defeito, impedimento ou ausência de medição, normalmente com base em regras bem definidas. O fluxo descrito para geração de campanha pode ser resumido conforme Figura 8.

Figura 8 - Fluxo para geração de campanhas.



Fonte: Elaborado pelo autor.

A eficácia de uma campanha pode ser mesurada de diversas formas. Os principais indicadores são: a efetividade, que determina a taxa de acerto; o recuperado por TOI (Termo de Ocorrência e Inspeção), que determina a média de energia recuperada por irregularidade encontrada; e o recuperado por inspeção, que determina a média de energia recuperada por visita realizada na campanha.

2.5 Dados e Estatística

O estudo de dados reais ou experimentais deve envolver a área de estatística como forma de coletar, analisar e tirar conclusões sobre um conjunto de informações. Os dados resultam da observação de uma ou mais variáveis simultâneas a um processo, podendo ser constituído por todos os membros do grupo (população) ou por uma parcela desse grupo (amostra). Normalmente, a coleta de dados exige a seleção de parte da população e, dessa forma, é necessário considerar a incerteza associada à extração da amostra antes realizar inferências estatísticas.

Define-se variável como uma característica que descreve um membro de uma amostra. Ela pode ser do tipo: discreta ou categórica, quando seus valores possíveis forem finitos ou seguirem uma sequência de contagem; ou do tipo contínua, quando seus valores constituírem um intervalo completo. Variáveis contínuas são usualmente associados a medições. Uma estatística é qualquer quantidade cujo valor possa ser calculado com base nos dados amostrais (DEVORE, 2018).

Um conjunto de dados pode ser observado para análise através de tabelas e gráficos. As tabelas são de difícil interpretação, entretanto são utilizadas diretamente para tratamento, mineração e aplicação de técnicas de aprendizado de máquina. Os gráficos são melhores para interpretação, mas podem induzir conclusões errôneas, sendo necessário uma avaliação crítica. Eles incluem os gráficos de pontos, de barras, de linhas, de setor, histogramas e diagramas de caixas, ou *boxplots*.

Resumos visuais são indicados para impressões iniciais, mas uma análise completa exige o cálculo e interpretação de medidas numéricas que servem para caracterizar o conjunto de dados. Dentre elas estão as medidas de tendência central (média, mediana e quartis) e as medidas de dispersão (variância e desvio padrão).

A média aritmética, $\bar{x}$, de uma amostra $\{x_1, x_2, \ldots, x_n\}$ é definida pela equação (2). É a medida mais conhecida e usada, entretanto pode ser inadequada em algumas aplicações, pois seu valor pode ser bastante afetado pela presença de outliers, observações atípicas com valores muito maiores ou muito menores que os demais da série.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2}$$

Como alternativa, a mediana não é afetada por outliers e pode ser obtida a partir da ordenação crescente dos n valores seguindo os critérios da equação (3).

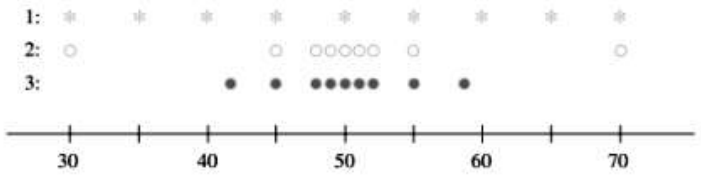
$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}}, & \text{para n ímpar} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n+2}{2}}}{2}, & \text{para n par} \end{cases} \tag{3}$$

Tanto a média quanto a mediana representarão onde os dados estão centralizados. Entretanto, a menos que a distribuição seja simétrica, essas medidas não serão iguais.

Diferente da mediana que divide o conjunto de dados em duas partes, os quartis são valores que dividem o conjunto em quatro partes. O chamado primeiro quartil, Q_1, separa o quarto inferior da série, o segundo quartil é igual a mediana e o terceiro quartil, Q_3, separa as observações do quarto superior do conjunto de dados (DEVORE, 2018). Para obter os três valores deve-se: após encontrar a mediana da série, dividir cada subsérie ao meio; esses valores serão os quartis.

As medidas de tendência central fornecem apenas informações parciais dos dados. Amostras com mesmos valores de média e mediana, mas diferentes entre si, podem ser observados na Figura 9 a seguir.

Figura 9 - Amostras com valores iguais de medidas centrais, mas com medidas de dispersão diferentes.



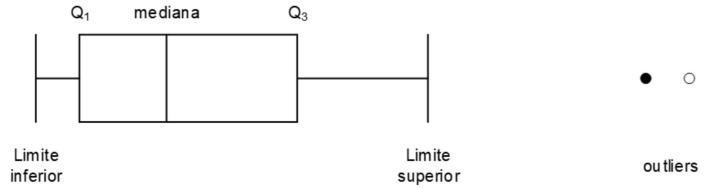
Fonte: Devore (2018).

As medidas de dispersão fornecem em que aspectos esses casos se diferenciam. Uma delas é a amplitude, dado pela diferença entre o maior e o menor valor da amostra. Essa medida, no entanto, depende apenas das observações extremas, ignorando as variações dos demais valores. Outra medida de dispersão é o desvio padrão, dado pela equação (4). Ele representa o tamanho de um desvio típico da média para a amostra selecionada.

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \tag{4}$$

Os *boxplots* são normalmente utilizados para apresentar as características mais proeminentes do conjunto de dados a partir das medidas citadas. Um esquemático de sua construção pode ser observado na Figura 10.

Figura 10 - Esquemático de um *boxplot*.



Fonte: Elaborado pelo autor.

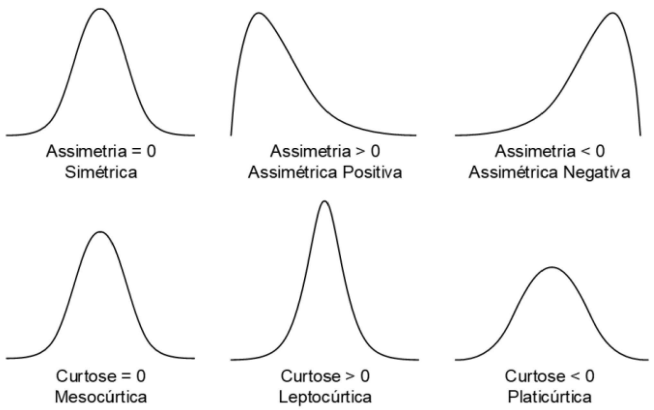
A partir do *boxplot* e das definições de quartis, é possível definir a amplitude interquartil, ou *interquartile range* (IQR), definidor por (5), que representa 50% de todos os valores observados no conjunto.

$$\mathrm{IQR} = \mathrm{Q_3} - \mathrm{Q_1} \tag{5}$$

É possível representar a presença de outliers através de marcações por fora dos limites superior e inferior. Com base no pressuposto de que a distribuição da população é do tipo normal, o que acontece para muitos exemplos reais, considera-se que qualquer observação mais distante do que 1,5·IQR de Q_1 ou Q_3 é um outlier. Um outlier é extremo se essa distância ultrapassar 3·IQR. Esses valores decorrem das propriedades da distribuição normal que afirmam que 99% dos valores estão a 1,5 desvios padrões da média.

A distribuição normal, ou gaussiana, é a mais importante distribuição de probabilidade, por representar muitas populações numéricas. Se a distribuição de uma população for normal, cerca de 68% dos seus valores estão a 1 desvio padrão da média, 95% dos seus valores estão a 2 desvios padrões da média e 99,7% dos seus valores estão a 3 desvios padrões da média. A assimetria mede a falta de simetria de uma distribuição em relação ao seu ponto central. A curtose mede o grau de achatamento da curva de distribuição de probabilidade em relação a curva normal. Essas medidas estão representadas na Figura 11. Alguns testes de normalidade, como o de Shapiro-Wilk, podem ser realizados nos dados amostrais para determinar se eles vieram de uma população com distribuição normal.

Figura 11 - Assimetria e curtose em uma distribuição de probabilidade.

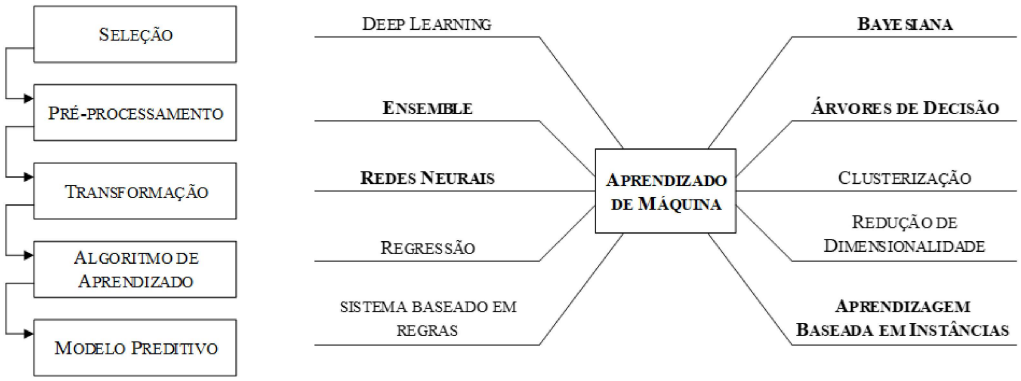


Fonte: Elaborado pelo autor.

2.6 Aprendizado de Máquina

Com o aumento na complexidade da solução de problemas reais a partir de um volume cada vez maior de dados, torna-se clara a necessidade de ferramentas computacionais capazes de resolver problemas de difícil visualização direta. O processo de induzir uma hipótese ou função a partir da experiência passada por meio de algoritmos para solução de problemas denomina-se Aprendizado de Máquina (AM), ou *Machine Learning*. As etapas usuais do desenvolvimento de modelo de AM, bem como as classificações dos tipos de algoritmos, podem ser observadas na Figura 12.

Figura 12 – Etapas de desenvolvimento de um modelo de Aprendizado de Máquina e os principais tipos de algoritmos.



Fonte: Elaborado pelo autor.

No aprendizado de máquina, os algoritmos aprendem a partir de um princípio de inferência, denominado de treinamento, no qual se obtêm conclusões genéricas a partir de um subconjunto de dados. O modelo deve ser capaz de relacionar os valores dos atributos de entrada ao seu respectivo atributo de saída, também chamado alvo ou *target*, mesmo quando aplicado a novos dados nunca antes apresentados ao algoritmo. Essa propriedade de manter-se válido para novos objetos é conhecida por generalização de um modelo. Se o algoritmo estiver com baixa capacidade de generalização, diz-se que o modelo está superajustado aos dados de treinamento, ou em *overfitting*, e não será capaz de apresentar resultados consistentes para dados inéditos. No caso inverso, o modelo está subajustado, ou em *underfitting*, não sendo capaz de produzir uma alta taxa de acerto, mesmo no conjunto de treinamento, normalmente porque os exemplos disponíveis são pouco representativos ou o modelo utilizado não foi capaz de capturar os padrões existentes nos dados (FACELI, et al, 2011).

De acordo com a forma que se dá o sistema de aprendizado, pode-se classificar os algoritmos de AM em aprendizado supervisionado e aprendizado não supervisionado. O termo supervisionado é utilizado devido à presença de um supervisor externo, ou professor, que conhece a saída desejada para cada exemplo. Aplicações do aprendizado supervisionado

incluem problemas de classificação, em que o objetivo é atribuir cada entrada a um número finito de categorias discretas, ou regressão, em que as saídas consistem de uma ou mais variáveis contínuas. Em um aprendizado não supervisionado, a figura do professor não existe e o atributo de saída não é diretamente utilizado. O objetivo é encontrar grupos semelhantes nos dados (clusterização), determinar sua distribuição (estimação de densidade) ou projetar os dados de alta dimensão em duas ou três dimensões.

2.6.1 Preparação dos Dados

Para utilizar um conjunto de dados em um algoritmo de aprendizado de máquina é necessário adequá-lo a partir de técnicas de pré-processamento. Essas técnicas envolvem principalmente a limpeza dos dados, a seleção de atributos e a transformação das variáveis através de normalização e redistribuição.

O banco de dados de organizações e empresas podem apresentar dados ruidosos, inconsistentes, redundantes ou incompletos. Substituir valores faltantes é de extrema importância, pois estes podem causar problemas nos modelos de AM. No entanto, deve-se considerar também os padrões dos valores que estão faltando e as informações que eles contêm. O uso de valores inapropriados para substituí-los pode perturbar o padrão dos dados e danificar informações que pudessem ser relevantes ao modelo. A captura da variabilidade presente em um conjunto de dados pode ser usada para inferir esses valores de maneira a causar menos danos ao conteúdo dos atributos. Conhecimento sobre o problema também pode auxiliar na escolha desses valores. Algumas técnicas incluem o emprego de estatísticas de tendência central, estimativa com base em outras variáveis ou valores fixos. Em casos extremos, em que a quantidade de dados reais é considerada insuficiente, elimina-se o objeto. A mensuração de insuficiência pode variar de acordo com a aplicação e a quantidade de dados disponíveis sendo normalmente de escolha do especialista. Dados inconsistentes também devem ser tratados e referem-se aos casos em que regras ou relações conhecidas são violadas, como, por exemplo, datas negativas e ultrapassagem de valores máximos. As técnicas para substituição desses dados são os mesmos para valores faltantes. Para casos de dados redundantes, é necessário eliminar as redundâncias encontradas como parte do processo de seleção de variáveis (PYLE, 1999).

A seleção de variáveis possui uma grande importância na área de aprendizado de máquina e análise de dados. O objetivo é selecionar o melhor subconjunto dos atributos originais preservando toda ou a maior parte da informação dos dados, eliminando aqueles que são irrelevantes ou redundantes. Outras abordagens incluem combinar as variáveis através de métodos lineares ou não lineares. Selecionar variáveis torna-se necessário pois, em algoritmos

de aprendizado, menos entradas significa menos parâmetros adaptativos a serem determinados com maior generalização do modelo (BISHOP, 1995). Mesmo que, em geral, uma redução na dimensão do vetor de entrada signifique uma redução de informação, em aplicações reais, como a quantidade de dados é limitada, a maldição da dimensionalidade leva a dados esparsos e pode reduzir a performance de sistemas de classificação (BISHOP, 1995). O termo maldição da dimensionalidade refere-se ao aumento exponencial da quantidade de dado necessária para determinar o mapeamento dos dados à medida que o número de dimensões ou variáveis aumenta. Mesmo para as ferramentas mais complexas existe um nível de dimensionalidade que derrotará qualquer esforço de construir um modelo adequado. Atributos redundantes trazem ainda outro problema que, por participarem mais de uma vez do processo de ajuste dos parâmetros do modelo, contribuem mais que outros objetos para a definição do resultado final e pode dar a falsa impressão de que esse perfil de objeto é mais importante que os demais.

A redundância de um atributo está relacionada a sua correlação com os demais atributos do conjunto de dados. Uma forma de verificar se duas variáveis estão relacionadas é através de coeficientes de correlação como os de Pearson e Spearman. Correlações positivas implicam que à medida que uma variável aumenta a outra também aumenta, enquanto correlações negativas implicam que enquanto uma aumenta a outra diminui. O coeficiente de correlação de Pearson, r, mede o grau da correlação linear entre dois conjuntos e é dado por (6).

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \cdot \sqrt{\sum(y_i - \bar{y})^2}} \tag{6}$$

Em que x_1, x_2, ..., x_n e y_1, y_2, ..., y_n são os valores que compõem os dois conjuntos.

Já o coeficiente de correlação de postos de Spearman, ρ, dado por (7), é uma medida que avalia relações monótonas, sejam elas lineares ou não, em que uma correlação perfeita indica que uma das variáveis é uma função monótona perfeita da outra. Na equação (7) representa a diferença entre os dois postos, ou *ranking*, de cada observação.

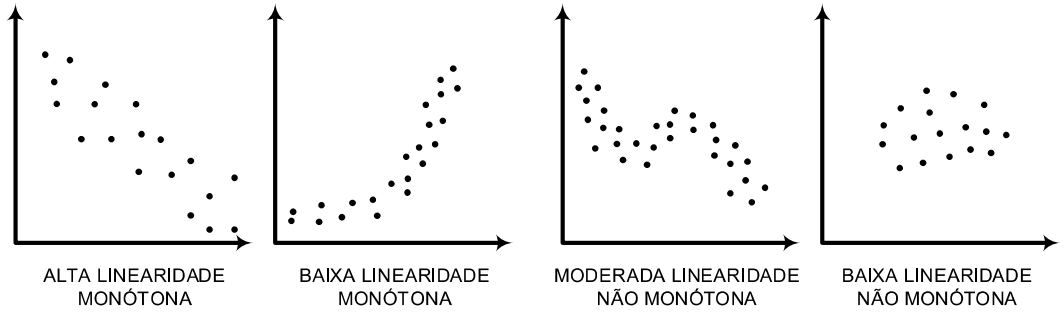
$$\rho = 1 - \frac{6 \cdot \sum {d_i}^2}{n \cdot (n^2 - 1)} \tag{7}$$

Em que d_i é a diferença entre os valores de postos das variáveis x: x_1, x_2, ..., x_n e y: y_1, y_2, ..., y_n e n é o número de observações.

Tanto r, quanto ρ variam entre -1 e +1 e uma correlação é considerada forte quando esses coeficientes são maiores que 0,8 ou menores que -0,8 (DEVORE, 2018). Um valor de r ou ρ próximo de 0 não significa que não existe uma forte relação entre as variáveis, mas apenas

que não há uma relação linear no caso de r, ou que não há uma relação monótona no caso de ρ. Casos com diferentes coeficientes podem ser observados na Figura 13 a seguir.

Figura 13 - Relações de linearidade e monotonicidade entre duas variáveis.



Fonte: Elaborado pelo autor.

Além da limpeza dos dados mencionado anteriormente, a transformação das variáveis através de normalização e redistribuição é de extrema importância para a maior parte das ferramentas de modelagens. Mesmo as que não exigem essas manipulações, podem se beneficiar delas.

O método a ser utilizado para normalizar um intervalo de uma variável deve possuir a menor distorção possível e ser tolerante a valores fora do intervalo (PYLE, 1999). Como os dados utilizados no treinamento de uma ferramenta de AM são apenas uma amostra da população, valores inéditos poderão ser apresentados ao modelo e devem ser levados em consideração. Há dois problemas principais com valores fora do intervalo. O primeiro é determinar os valores máximos, mínimos e a frequência da série para ajustá-los corretamente a normalização e a redistribuição. O segundo é que esses valores representam uma parte da informação do padrão da população que o modelo não será exposto durante o treinamento.

Se os valores máximos e mínimos de uma variável contínua são conhecidos, pode-se aplicar uma transformação por re-escala, conhecido por normalização mín-máx, em que novos limites de máximos e mínimos são definidos para o atributo. Para esse método, os valores são transformados a partir da equação (8). A maior vantagem desse método é que ele não introduz distorções a distribuição da variável.

$$x_{novo} = min_{novo} + \frac{x - min}{max - min}(max_{novo} - min_{novo}) \quad (8)$$

Para lidar com valores fora do intervalo, em que se desconhece o valor máximo e mínimo da série, é necessário considerar outras metodologias. Uma delas é o *clip*, em que se limita os valores fora do intervalo estabelecido ao grampeá-los. O problema desse método é que se assume que os números que estão fora do intervalo são equivalentes aos números dentro do intervalo (PYLE, 1999). Em algumas aplicações, especialmente as relacionadas a fraude,

isso pode não ser verdade já que, atividades fraudulentas, por exemplo, podem ser mais prováveis de sair do intervalo determinado, já que novos padrões de fraude estão em constante desenvolvimento. A menos que toda a população esteja disponível para modelagem, este é um problema que não pode ser evitado.

Para as variáveis categóricas simbólicas, como as técnicas de AM só lidam com dados numéricos, é necessário convertê-las preservando sua informação. As que assumem apenas dois valores, um dígito binário é suficiente para sua representação. As que assumem mais de dois valores, a conversão depende de o atributo ser ordinal, em que existe uma relação de ordem para os valores numéricos, ou nominal. Para os atributos nominais, codifica-se as diferentes entradas por uma sequência de c bits. Esse método é conhecido por codificação 1-de-c, ou ainda variáveis *dummy*, e cada sequência possui apenas um bit com valor 1 e os demais são 0. São necessárias c-1 variáveis para representar um atributo com c classes.

Dependendo do número de valores nominais, a sequência binária para representar cada valor pode ficar muito longa e pode-se optar por utilizar uma representação binária, em que a quantidade de atributos será determinado por $\log_2 c$. As duas representações podem ser observadas no Quadro 2.

Quadro 2 – Codificação 1-de-c para atributos nominais.

Atributo	**Código 1-de-c**	**Código binário**
Preto	000	00
Amarelo	100	01
Vermelho	010	10
Branco	001	11

Quando existe uma relação de ordem, a transformação é dada após a ordenação parar, em seguida, utilizar um código de acordo com a posição de ordem do valor. Caso seja necessário converter em códigos binários, pode-se utilizar o código cinza, em que apenas um bit varia de um número para o outro como um contador; ou o código termômetro, em que o aumento dos valores se assemelha ao aumento de temperatura em um termômetro analógico, conforme Quadro 3.

Quadro 3 – Codificação cinza ou termômetro para atributos ordinais.

Atributo	**Código cinza**	**Código termômetro**
Primeiro	000	0000
Segundo	001	0001
Terceiro	011	0011
Quarto	010	0111

Como parte da transformação das variáveis, além da normalização, também é necessário redistribuir objetos que são muito esparsos. Esses objetos podem não carregar muita informação, mas podem ser necessários para o modelo no caso de algumas aplicações. Uma solução utilizada é reunir esses atributos em conjuntos, através da técnica de *binning* ou quantização. Essa técnica envolve agrupar intervalos dos valores e usar rótulos para esses intervalos considerando-os como substitutos dos valores reais. Quando não há uma demarcação clara para determinar os limites dos intervalos, deve-se buscá-los de maneira que cada novo rótulo contenha uma mesma quantidade aproximada de casos. Esse procedimento também pode ser utilizado quando se deseja eliminar a ordem natural dos números e, após aplicar o remapeamento, a variável deve ser tratada como nominal.

2.6.2 Avaliação de Modelos Preditivos

Não é possível estabelecer a priori qual a melhor técnica de AM para um determinado problema, visto que a performance de um algoritmo pode depender da aplicação, das variáveis utilizadas e até mesmo dos parâmetros escolhidos no modelo. Dessa forma, torna-se necessário aplicar uma experimentação estruturada.

De início, é necessário fazer uma amostragem do conjunto de dados, em que devem ser separados um conjunto para treinamento e outro para teste com dados não observados pelo preditor durante a fase de treino. Calcular o desempenho preditivo do modelo utilizando os mesmos objetos do treino produz estimativas otimistas já que os algoritmos tendem a melhorar o seu desempenho preditivo nesses objetos (FACELI et al, 2011). As medidas de desempenho utilizadas devem ser reportadas junto com seus valores de desvio padrão associado. Um alto desvio padrão indica uma alta variabilidade nos resultados e uma instabilidade do modelo perante mudanças nos objetos (FACELI et al, 2011).

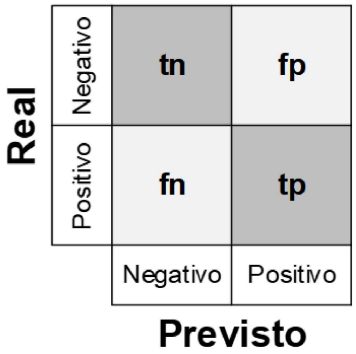
Para subdividir o conjunto de dados, existe a opção de empregar o *holdout*, em que se divide o conjunto de dados em uma proporção de p% para treinamento e (100% - p%) para teste. A amostragem desse conjunto pode ser aleatória ou aleatória estratificada, em que se garante que grupos da população são representados em uma proporção adequada em relação ao todo. Para tornar o resultado menos dependente da escolha da partição, é possível aplicar o *holdout* diversas vezes com diferentes partições aleatórias e obter a média de desempenho.

O método mais utilizado para testar o desempenho de um algoritmo de AM é a validação cruzada. Na validação cruzada do tipo k-Fold, o conjunto de dados é dividido em k partes, em que, para cada partição, utiliza-se k - 1 partes para treino e a parte remanescente para teste. Isso

é repetido para as k partições. Esse método produz uma estimativa mais fiel do desempenho preditivo do modelo.

Para avaliar a performance de um modelo, existe uma variedade de métricas que podem ser adequadas para uma aplicação em particular. As métricas de erro mais comuns em problemas de classificação incluem a acurácia, a precisão e a sensibilidade. Essas medidas são extraídas da matriz de confusão que ilustra o número de predições corretas e incorretas em cada classe e pode ser montada conforme mostrado na Figura 14 para duas classes.

Figura 14 - Matriz de confusão.



Fonte: Elaborado pelo autor.

As nomenclaturas tn, fp, fn e tp vêm do inglês true negative (verdadeiro negativo), false positive (falso positivo), false negative (falso negativo) e true positive (verdadeiro positivo). Os falsos positivos também são conhecidos como erro tipo 1 e os falsos negativos como erro tipo 2.

A acurácia, ac, pode ser definida conforme a equação (9) a seguir.

$$\mathrm{ac} = \frac{\mathrm{tp} + \mathrm{tn}}{\mathrm{n}} \tag{9}$$

Em que n é a quantidade de elementos da matriz de confusão. Essa métrica também é denominada de taxa de acerto e determinará a proporção de exemplos corretamente classificados. Como apontado por Kubat e Matwin (1997), para aplicações em que o conjunto de dados é desbalanceado, ou seja, uma classe é bem menos representada que as demais, o desempenho de um algoritmo de AM não pode ser expresso em termos da acurácia. De fato, se o modelo ignorar a existência da classe menos representada, ainda assim é possível obter altos índices de acurácia. Para uma aplicação em que apenas 10% da população é da classe positiva, por exemplo, é possível obter uma taxa de acerto de 90% ao classificar todos os itens como negativos. Como alternativa, são utilizadas outras métricas para melhor representar a performance do modelo.

A efetividade, ou precisão, é definida pela equação (10) e indica a proporção de positivos classificados corretamente dentre todos os classificados como positivos pelo modelo.

$$\text{efet} = \frac{\text{tp}}{\text{tp} + \text{fp}} \tag{10}$$

A cobertura, ou sensibilidade, é definida pela equação (11) e corresponde a taxa de acerto da classe positiva, ou seja, quantas unidades da classe positiva foram corretamente classificados dentre os que realmente eram positivos.

$$\text{cob} = \frac{\text{tp}}{\text{tp} + \text{fn}} \tag{11}$$

Para agregar as duas métricas, é possível utilizar a função F-score, ou F-measure, dada pela equação (12). Ela se baseia na medida E-measure, definido por Rijsbergen (1979), ao considerar um mesmo peso de ponderação entre a precisão e a sensibilidade.

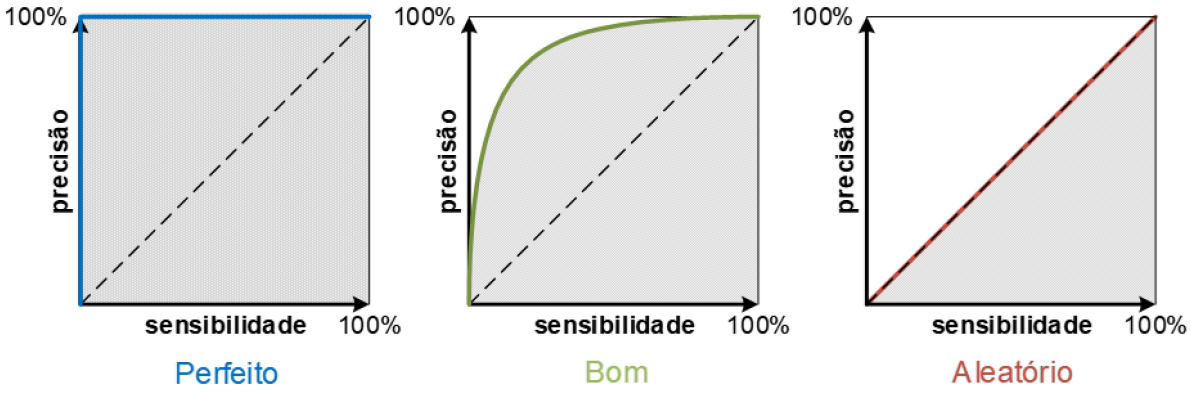
$$\text{F-score} = 2 \cdot \frac{\text{efet} \cdot \text{cob}}{\text{efet} + \text{cob}} = \frac{2 \cdot \text{tp}}{2 \cdot \text{tp} + \text{fn} + \text{fp}} \tag{12}$$

Outra função utilizada para agregar a efetividade e a cobertura é o G-measure, equação (13), que representa a média geométrica entre as duas métricas. O F-score, por sua vez, caracteriza a média harmônica.

$$\text{G-measure} = \sqrt{\text{efet} \cdot \text{cob}} = \text{tp} \cdot \sqrt{\frac{1}{(\text{tp} + \text{fp}) \cdot (\text{tp} + \text{fn})}} \tag{13}$$

A relação entre efetividade e cobertura pode ser visualizada graficamente através da Curva Característica de Operação do Receptor, ou curva ROC. Ela também é utilizada como uma forma de avaliar classificadores de duas classes. A curva ROC é um gráfico bidimensional plotado com o eixo x representando a precisão e o eixo y, a sensibilidade, conforme Figura 15. Ela é construída para diversos limiares de discriminação entre as classes. Quanto mais próximo à curva estiver das bordas superior e esquerda, mais preciso é o modelo. Outra forma de visualizar é a partir da área abaixo da curva ou AUC, calculada através de sua integral. Quanto maior a área sob a curva, mais preciso é o modelo. A diagonal do gráfico representa um modelo aleatório e qualquer curva abaixo dela representa um classificador pior que a utilização da aleatoriedade para classificar os padrões.

Figura 15 - Curva ROC e sua interpretação para diferentes modelos.

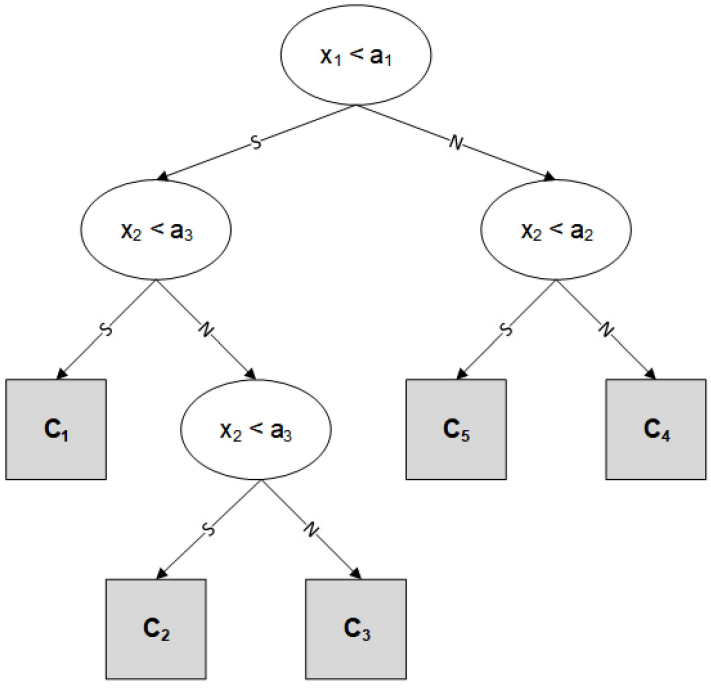


Fonte: Elaborado pelo autor.

2.6.3 Random Forest

Dentre os modelos preditivos existentes na área de aprendizado de máquina, encontram-se os métodos baseados em árvores de decisão, em que a solução do problema é produzida a partir de divisões e mapeamentos de subproblemas. Uma árvore de decisão clássica é um grafo acíclico direcionado, conforme Figura 16.

Figura 16 - Árvore de decisão.



Fonte: Elaborado pelo autor.

Os métodos baseados em árvore são simples de interpretar, entretanto geralmente não apresentam bons resultados de precisão quando comparados a outras técnicas de AM. Isso ocorre pois elas são instáveis, ou seja, uma pequena alteração nos dados pode gerar uma grande

alteração na árvore final já que, a cada nó, o critério para dividir a árvore tem como base o melhor atributo, sendo que dois ou mais atributos podem ser classificados similarmente e pequenas variações podem levar a árvores completamente diferentes (FACELI et al., 2011). No entanto, a performance pode ser aprimorada substancialmente ao agregar várias árvores de decisão ao mesmo modelo como ocorre para as florestas aleatórias, ou como são mais conhecidas, *Random Forests*.

Proposto por Breiman (2001), as *Random Forests* são combinações de árvores de decisão treinadas a partir do método de *bagging* em que cada árvore possui um conjunto de atributos escolhidos aleatoriamente para cada divisão com o objetivo de mantê-las não correlacionadas. *Bagging*, ou *bootstrap aggregation*, é um algoritmo do tipo *ensemble* que foi desenvolvido para reduzir a variância de modelos ruidosos e aumentar sua precisão. Ele consiste em gerar diversos conjuntos de treinamento a partir de uma amostragem uniforme e com reposição, de modo que observações podem ocorrer mais de uma vez, treinar os modelos e, então, combinar as saídas através de média ou de eleição.

2.6.4 Gradient Tree Boosting

Outro algoritmo baseado em árvores de decisão é o *Gradient Tree Boosting*. O *Gradient Boosting*, assim como o *bagging*, é um classificador *ensemble* em que um conjunto de modelos são treinados para reduzir enviesamento e variância de modelos instáveis e, consequentemente, aumentar estabilidade e precisão da técnica. Construídos os modelos, o *Gradient Boosting* generaliza-os a partir da otimização de uma função de perda diferenciável arbitrária. Normalmente esse algoritmo é utilizado com árvores de decisão.

Enquanto no *bagging* o conjunto de treinamento era gerado de modo que todas as observações possuem a mesma probabilidade de serem selecionadas para treinar os modelos, no *boosting* elas são ponderadas e algumas farão parte do conjunto com mais frequência. Cada classificador do *boosting* é treinado considerando o sucesso do classificador anterior. A cada etapa de treinamento, os pesos das ponderações são redistribuídos de modo a enfatizar os casos mais difíceis de se classificar, para que os próximos classificadores se concentrem nele.

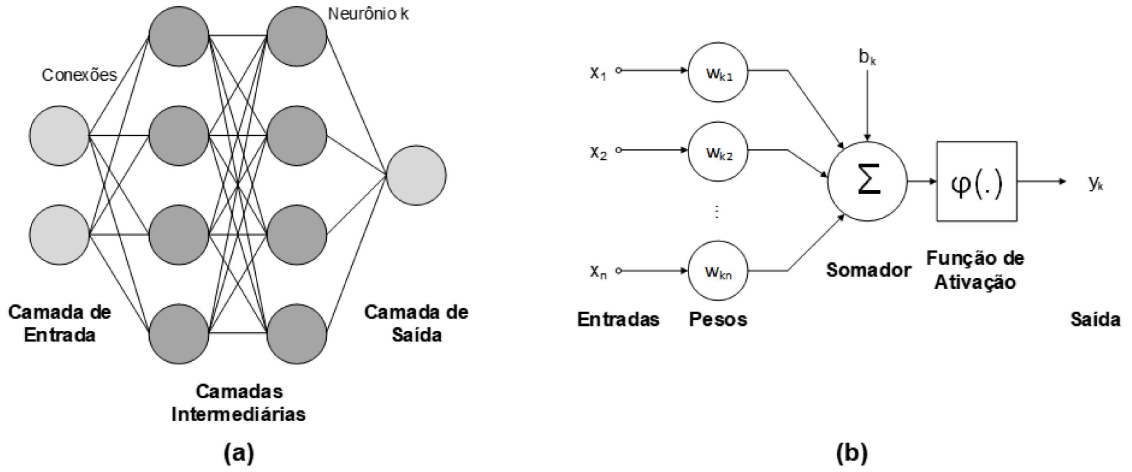
A saída é obtida também de maneira ponderada, diferente do *bagging*. Um modelo com um bom resultado de classificação recebe um peso maior ao fim de cada treinamento, e a saída será determinada pela média ponderada das estimativas de cada modelo.

2.6.5 Redes Neurais Artificiais

Outra técnica amplamente utilizada de aprendizado de máquina é a Rede Neural Artificial (RNA). As RNAs foram projetadas tomando como inspiração a maneira como o cérebro humano adquire conhecimento e a sua capacidade de aprendizado. Seus sistemas são compostos de unidades de processamentos denominadas neurônicos que computam funções matemáticas e estão dispostas em camadas, onde são interligadas por conexões normalmente unidirecionais. Essas conexões possuem pesos associados que ponderam a entrada recebida por cada neurônio e o processo de treinamento, também conhecido aqui por processo de aprendizado, consiste em ajustar os valores desses pesos codificando o conhecimento adquirido.

Na Figura 17(a) é possível observar uma rede *feedforward* multicamadas composta por: camada de entrada, em que os padrões são apresentados ao modelo; camadas intermediárias, onde é feita a maior parte do processamento; camada de saída, em que são entregues as saídas do sistema. Em uma rede *feedforward* a informação flui sem realimentação de valores de saídas. Cada camada pode ser composta por diversos neurônios. Como pode ser observado pela Figura 17(b), as entradas de um neurônio recebem os valores e estes são ponderados e combinados através de um somador para então serem convertidos na saída após aplicar uma função de ativação. Essa função é normalmente do tipo sigmoidal, ao menos para os neurônios da camada de saída, já que permite que o resultado final seja interpretado de maneira probabilística (BISHOP, 1995). Para as camadas intermediárias, a função linear retificada ou *Rectified Linear Unit* (ReLu) têm sido amplamente utilizada graças a sua simplicidade e eficácia (RAMACHANDRAN; ZOPH; LE, 2017).

Figura 17 - Representação de uma rede neural. (a) Rede neural multicamadas. (b) Modelo de um neurônio.



Fonte: Elaborado pelo autor.

Além de sua arquitetura, que inclui a quantidade de camadas e neurônios, função de ativação e a forma das conexões, as redes neurais também são caracterizadas pelo seu algoritmo

de aprendizado, que consiste nas regras utilizadas para ajuste dos pesos. Vários algoritmos foram propostos pela literatura, sendo os mais utilizados para o aprendizado supervisionado os algoritmos de otimização, como o gradiente descendente, o RMSProp e o Adam, que buscam minimizar uma função objetivo, normalmente uma função de erro. Em geral, o treinamento é realizado em conjunto com a técnica de retropropagação (*backpropagation*) para computar o gradiente da função objetivo. O termo *backpropagation* refere-se apenas ao método para calcular o gradiente, enquanto o algoritmo de otimização é usado para realizar o aprendizado, ou seja, atualizar os pesos, usando esse gradiente (GOODFELLOW, et al., 2016).

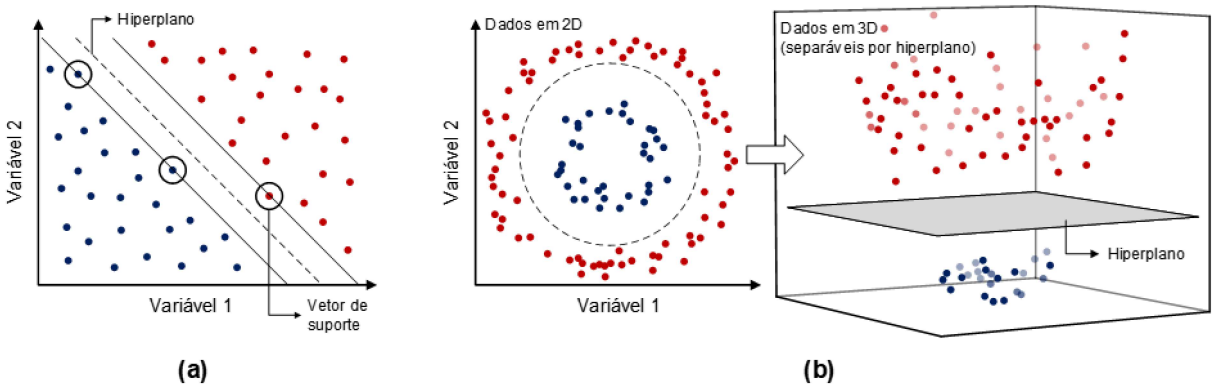
2.6.6 Máquinas de Vetores de Suporte

As máquinas vetores de suporte, *support vector machines* (SMVs) são modelos de aprendizado supervisionado com algoritmos de regras de associação que analisa dados e reconhece padrões. Elas têm sido amplamente utilizadas nos últimos anos e seus resultados comparados a técnicas populares como as RNAs. Sua teoria é embasada pela teoria de aprendizado estatístico que estabelece condições matemáticas que auxiliam na escolha de um classificador a partir de um conjunto de dados de treinamento. Essas condições consideram o desempenho do classificador no treino e sua complexidade, objetivando um bom desempenho para dados não apresentados ao modelo.

Considerando um conjunto de dados para treinamento, com os exemplos separados em categorias, o algoritmo do SVM busca construir um modelo que determinará a qual classe pertencem novos padrões. Nesse modelo, os exemplos são representados como pontos e mapeados de modo que as categorias sejam separadas por um espaço bem definido. Esse espaço é conhecido como hiperplano e ele é determinado de modo que a margem entre as classes seja maximizada. Os pontos que tocam as margens são conhecidos por vetores de suporte. Na Figura 18(a) é possível observar um exemplo para uma classificação binária em duas dimensões, com o hiperplano sendo representado por uma reta.

SVMs lineares são eficazes em conjuntos de dados aproximadamente lineares mesmo com a presença de ruídos e outliers. Entretanto, quando não é possível separar as classes diretamente, SVMs não lineares mapeiam o conjunto de dados de seu espaço original para um novo de maior dimensão, conhecido por espaço de características. A escolha apropriada desse novo espaço faz com que o conjunto de treinamento mapeado possa ser separado por um hiperplano, conforme a SVM linear. Esse caso é ilustrado pela Figura 18(b).

Figura 18 - Representação da lógica de Máquinas Vetores de Suporte. (a) SVM linear. (b) SVM não-linear.

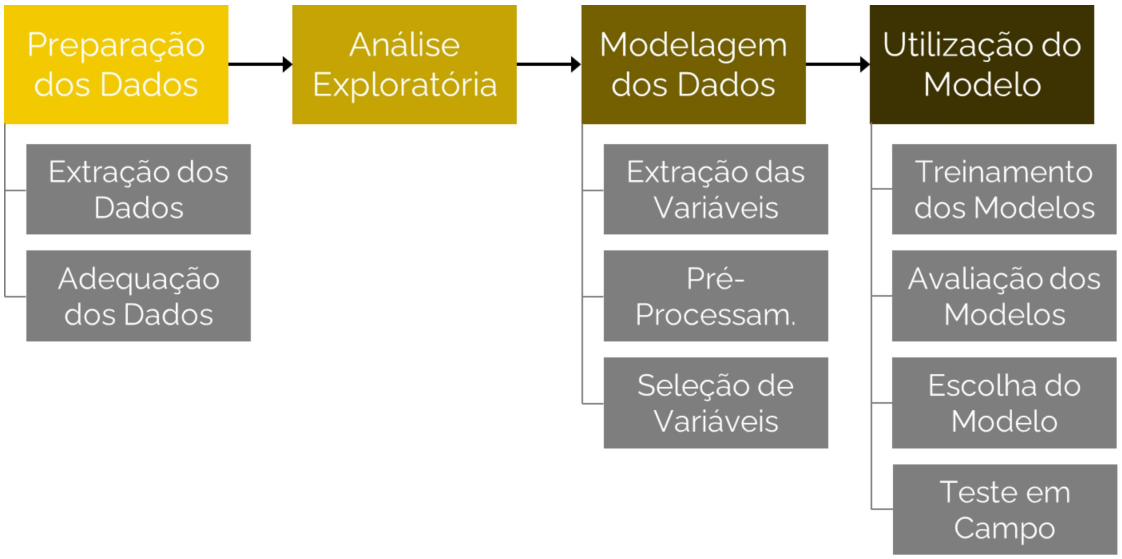


Fonte: Elaborado pelo autor.

3 PROPOSTA DO TRABALHO

Neste capítulo é apresentada a proposta do trabalho da dissertação, objetivando-se desenvolver uma metodologia para detecção de fraude em clientes da baixa tensão, com a finalidade de aumentar os índices de efetividade. Para tanto, o fluxo de processos empregado teve como base o proposto por Pyle (1999) e encontra-se resumido na Figura 19 a seguir.

Figura 19 - Fluxo da metodologia empregada.



Fonte: Elaborado pelo autor.

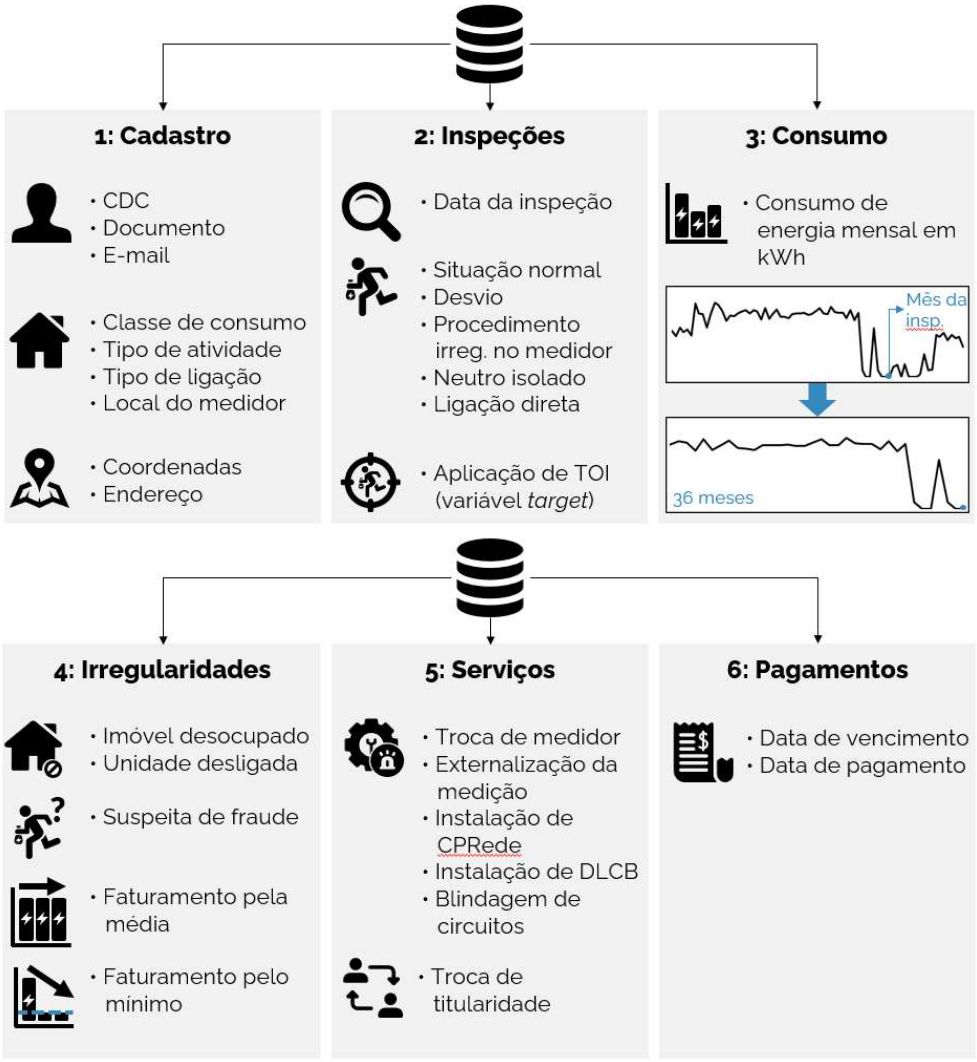
Cada etapa da metodologia será explanada e exemplificada. Assim, também serão apresentadas as considerações para a implementação utilizando a linguagem de programação Python e seu conjunto de bibliotecas para toda a extração, manipulação e classificação dos dados. Os principais pacotes utilizados foram:

a) NumPy: para operações com *arrays* e cálculos matemáticos;
b) Pandas: para importação, manipulação e análise de dados através do uso de *DataFrames*;
c) Seaborn e Matplotlib: para análise exploratória e geração de gráficos estatísticos;
d) Geopy: para localização de coordenadas a partir de endereços usando a ferramenta de geolocalização Nominatim;
e) OSMnx: para traçar rotas de carro em espaços geográficos;
f) NLTK: para processamento de linguagem natural;
g) Sklearn: para implementação, aplicação e avaliação dos modelos de aprendizado de máquina.

3.1 Construção do Banco de Dados

Os sistemas comerciais da empresa do estudo de caso possuem informações sobre os clientes relacionadas tanto ao serviço prestado, quanto a unidade consumidora. As informações utilizadas neste trabalho foram classificadas em seis categorias: cadastro, inspeções, consumo, irregularidade de leitura, irregularidade de faturamento, serviços e pagamentos. Essas categorias e suas principais informações estão presentes na Figura 20.

Figura 20 – Tabelas presentes nos sistemas comerciais da distribuidora utilizadas para construção do banco de dados.



Fonte: Elaborado pelo autor.

A tabela de cadastro contém dados relacionados ao consumidor responsável pela unidade, à localização geográfica e ao imóvel. Cada unidade consumidora (UC) possui um Código de Consumidor (CDC) que a caracteriza e que será utilizado no banco de dados como a chave primária. Os principais atributos considerados estão presentes no Quadro 4.

Quadro 4 - Atributos de cadastro obtidos dos sistemas.

Atributo	Categoria	Descrição
CDC	Consumidor	Código do Consumidor
Nome Consumidor	Consumidor	Nome do responsável ou da unidade
CPF Responsável	Consumidor	Número do CPF do responsável
Documento	Consumidor	Número do documento do responsável
Tipo do Documento	Consumidor	Tipo do documento (CNPJ, carteira de identidade, etc.)
E-mail	Consumidor	Endereço eletrônico do responsável
Telefone	Consumidor	Telefone do responsável ou da unidade
Classe Consumo	Imóvel	Classe especificado para tarifas pela ANEEL
Tipo Imóvel	Imóvel	Indica porte do imóvel
Divisão Atividade	Imóvel	Subclasse da unidade conforme ANEEL
Situação	Imóvel	Indica se a unidade está ligada ou desligada no sistema
Tipo Ligação	Imóvel	Tipo de ligação que atende à unidade
Local Medidor	Imóvel	Local de instalação do medidor (poste, CPRede, interno)
CEP	Localização	Código de Endereçamento Postal
Complemento	Localização	Complemento do endereço
Longitude	Localização	Longitude da coordenada
Latitude	Localização	Latitude da coordenada
Endereço	Localização	Endereço (rua e número, quando disponível)
Regional	Localização	Divisão do estado em Leste, Centro e Oeste
Município	Localização	Município
Livro	Localização	Número que caracteriza a data de leitura do consumo
Local	Localização	Cidade
Rota	Localização	Conjunto de ruas em um livro que determinam o trajeto de leitura do leiturista
Bairro	Localização	Bairro da unidade
Rota Rural	Localização	Indica se a rota da unidade está em zona rural

Os dados de inspeções referem-se as comprovações feitas em campo da existência ou não de uma irregularidade na medição do consumo de energia. No banco de dados, foram trazidas a data da inspeção, a indicação de aplicação do TOI e a ocorrência apontada. Para as irregularidades, considerou-se apenas as ocorrências que indicam a manipulação do faturamento através de fraude ou furto. As ocorrências extraídas para o banco estão presentes no Quadro 5. O atributo aplicação de TOI (Termo de Ocorrência e Inspeção) será a variável *target* para classificação.

A tabela de consumo do banco de dados refere-se ao consumo lido mensal para cada unidade. Esse valor de consumo pode ser diferente do valor faturado em casos de refaturamento, ajuste manual de leitura ou leitura pela média ou pelo mínimo nos meses considerados. Essa métrica foi obtida para os meses de janeiro de 2013 a julho de 2019.

Quadro 5 - Ocorrências consideradas para montar o banco de dados.

Atributo	Aplicação de TOI
Desvio de energia no ramal de entrada	Sim
Desvio de energia no ramal de ligação	Sim
Desvios nos bornes do medidor	Sim
Ligação direta – sem medidor	Sim
Ligação direta – intervenção de terceiros	Sim
Medidor inclinado/deitado – intervenção de terceiros	Sim
Neutro isolado	Sim
Procedimento irregular no medidor	Sim
Situação normal	Não

As informações de irregularidade de leitura ou faturamento são representadas por códigos presentes nos sistemas e podem indicar alguma característica da unidade para o mês de referência. As mais relevantes para este trabalho podem ser encontradas no Quadro 6. Como definido na Seção 2.2, as irregularidades de faturamento indicam quando o consumo lido pode ter sido divergente do faturado, enquanto as irregularidades de leitura são apontamentos da coleta de leitura mensal e podem justificar a irregularidade de faturamento. A indicação de suspeita de fraude é uma das mais importantes na escolha das unidades para campanha, como já citado.

Quadro 6 – Irregularidades de leitura e faturamento consideradas no banco de dados.

Descrição da Irregularidade	Tipo
Imóvel desocupado	Leitura
Suspeita de fraude	Leitura
Faturamento pelo mínimo	Faturamento
Faturamento pela média	Faturamento
Unidade desligada	Faturamento
Acerto de faturamento	Faturamento

A tabela de serviços refere-se a intervenções feitas na unidade por equipes da empresa. Elas foram utilizadas neste trabalho para extrair características – como por exemplo, a instalação de DLCB na medição - ou para observar mudanças de comportamento na unidade – exemplo, mudança de titularidade. As Ordens de Serviço (OSs) consideradas foram: suspensão de energia por falta de pagamento, transferência de titularidade, externalização da medição, instalação de CPRede, instalação de DLCB e blindagem de circuitos.

Finalmente, a tabela de pagamentos traz informações acerca das datas de pagamento e vencimentos das contas de energia dos clientes.

Originalmente, foram selecionadas 95.597 unidades consumidoras para composição do banco de dados. Para essa triagem, levou-se em consideração o sistema de armazenamento e consolidação de dados da empresa que disponibiliza apenas uma janela de 5 anos devido ao grande volume de informações. Dessa forma, selecionaram-se apenas as UCs com inspeção nos últimos 2 anos com uma das ocorrências do Quadro 5. Esse intervalo de tempo foi escolhido para que houvesse disponibilidade de pelo menos 3 anos de consumo anteriores a aplicação do TOI para cada cliente.

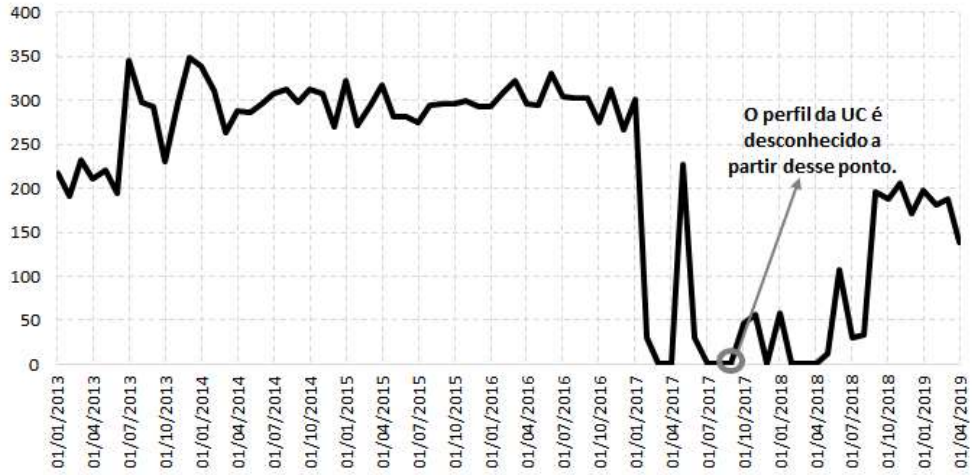
3.1.1 Adequação da Base de Dados

Como mencionado, a tabela de inspeções possui a variável *target*, ou alvo, da classificação. Como cada unidade consumidora foi inspecionada em uma data diferente, é necessário fazer uma adequação das tabelas considerando uma data de inspeção como a data referência. Sabendo que uma unidade consumidora pode ter mais de uma inspeção, priorizou-se aquela que houvesse evidenciado fraude, já que esses casos são menos frequentes. Para situações em que ocorreram mais de uma inspeção, optou-se pela mais recente delas.

O perfil de irregularidade antes ou após a data de referência é desconhecido, tendo em vista que a única maneira de comprovar uma fraude é através de aferição em campo. No banco de dados, apenas o consumo anterior à data de referência, data em que se conhece o perfil da unidade, deve ser considerado, afinal de contas é ele quem representa o perfil do imóvel para classificar em fraude ou situação normal e o objetivo do sistema é identificar a irregularidade enquanto ela está ocorrendo. Na Figura 21 é possível observar o consumo completo de uma UC inspecionada em 15/09/2017, tomada para exemplificação.

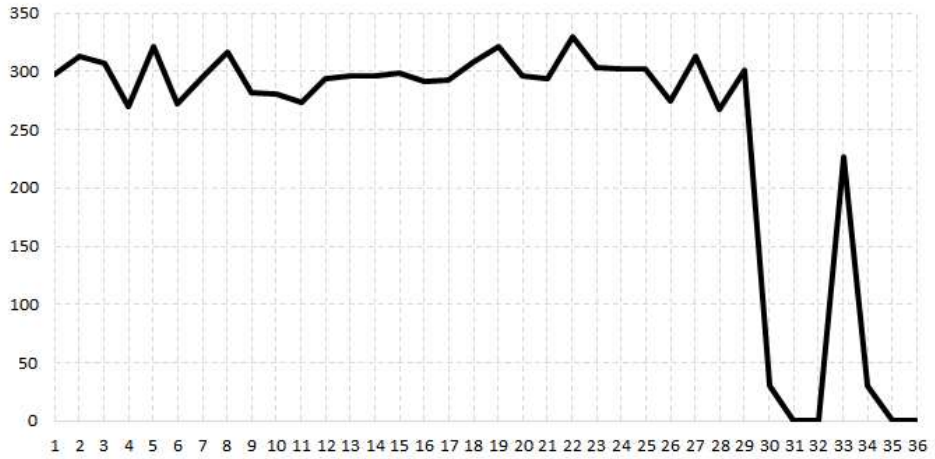
Para considerar a mesma quantidade de pontos para todos os exemplos do banco, assumiu-se uma janela de 36 meses anteriores a data de referência. Essa janela foi aplicada as tabelas de consumo, irregularidades e serviços. Além disso, foram atualizados os dados de cadastro com as informações válidas naquela data específica. Para implementar essa alteração, foi necessário extrair do banco da empresa um cadastro versionado, ou seja, um cadastro contendo todas as versões já existentes para aquela UC, em que se considerou apenas a versão válida para o intervalo da data de referência. O consumo da unidade da Figura 21, no banco de dados, será como na Figura 22. As unidades com menos de 24 pontos de consumo foram expurgadas da base.

Figura 21 - Consumo disponível nos sistemas da empresa para uma unidade.



Fonte: Elaborado pelo autor.

Figura 22 – Alteração da curva de consumo da Figura 22 para 36 meses com base na data de inspeção.



Fonte: Elaborado pelo autor.

3.2 Extração de Variáveis

A partir da construção do banco de dados descrita na seção anterior, algumas variáveis puderam ser extraídas como possíveis candidatas para compor o conjunto final dos dados de treinamento. Algumas características foram convertidas diretamente em variáveis categóricas, enquanto outras foram transformadas. O mapeamento para eliminar a ordem de variáveis contínuas foi feito quando julgado necessário.

As variáveis categóricas extraídas de forma direta foram a classe de consumo, o tipo de ligação e a variável *target* denominada fraude. Para o tipo de ligação, embora sejam possíveis encontrar ligações monofásicas, bifásicas e trifásicas nas medições, apenas a primeira e a última existem na empresa em questão. Conforme norma, as classes de consumo para cada tipo de consumidor são: residencial, industrial, comercial, rural, poder público, serviço público,

iluminação pública e consumo próprio. As duas últimas classes são tratadas de maneira diferenciada para análise e intervenção dentro da distribuidora e não foram consideradas no banco de dados final.

Outras variáveis obtidas para compor as categóricas foram extraídas a partir das ordens de serviço e do próprio cadastro. São elas: DLCB, blindagem, CPRede e externalização. Os atributos indicam a existência dessas características no padrão de medição do imóvel, levando em consideração apenas os serviços que foram executados na unidade antes do mês de referência.

As variáveis contínuas do banco de dados estão associadas principalmente ao consumo da unidade, mas outros parâmetros foram criados considerando a quantidade de irregularidades de leitura ou faturamento e o pagamento de contas.

A série temporal do consumo de energia elétrica considera até 36 meses de dados. O primeiro conjunto de variáveis foi construído a partir dos parâmetros estatísticos da curva subdividindo-a em 3. Esses parâmetros foram: máximo, mínimo, média, desvio padrão, assimetria e curtose. Nesse passo, foram geradas 18 métricas. Optou-se por séries de 12 meses para evitar sazonalidades. Unidades irrigantes, por exemplo, costumam reduzir o consumo de energia nos meses chuvosos. Esse comportamento estaria incluído nas 3 séries de 12 meses do consumidor.

O segundo conjunto de variáveis relacionadas ao consumo envolve a comparação das séries subdivididas anteriormente. Para isso, utilizou-se os coeficientes de correlação de Pearson e Spearman e a variável de desvio definida em (1) denominada degrau de consumo.

Por fim, o terceiro conjunto de variáveis envolveu toda a série de 36 meses de consumo. Uma das métricas considerou o p-valor do teste de Shapiro-Wilk para identificar se os dados possuem uma distribuição normal. As demais métricas referiram-se ao valor máximo e mínimo da primeira derivada da série de consumo. A utilização da primeira derivada, que consiste basicamente em realizar a diferença entre os meses, objetivou verificar alterações bruscas de consumo entre meses consecutivos.

3.2.1 Mapeamento de Variáveis Contínuas

Quando desejava-se eliminar o ordenamento ou reduzir a dispersão das variáveis contínuas, foi utilizado um mapeamento para transformá-las em categóricas através do método de quantização ou *binning*.

A quantidade de irregularidades de leitura ou faturamento do Quadro 6 são exemplos de variáveis dispersas, heterogêneos, com alta dispersão. Suas distribuições, desconsiderando as

unidades que não possuem essas características, foram utilizadas para determinar os pontos limites para construção das categorias. A irregularidade de acerto de faturamento foi utilizada apenas na etapa de pré-processamento dos dados, enquanto a de suspeita de fraude foi empregada em modelos específicos, a fim de evitar enviesamento dos algoritmos de aprendizado devido à precisão dela em relação às demais na detecção de fraudes.

Para as irregularidades de faturamento por média e UC desligada, as categorias foram divididas em subtipos denominados “ausente”, “pouco”, “médio” e “muito”. Eles foram delimitados da seguinte forma: se não possui o atributo, é denominado ausente; se possui ao menos uma vez, é denominado pouco; se possui mais de um, mas menos de 5 do atributo, então é denominado médio; se possui 5 ou mais, é denominado muito. As irregularidades de leitura de imóvel desocupado e suspeita de fraude também foram divididas nos 4 subtipos mencionados, entretanto com outras delimitações. Para desocupado, considerou-se que: inferior a 2, é denominado pouco; se possui mais de 2, mas menos de 5 indicações, médio; se possui 5 ou mais, muito. Para a suspeita, considerou-se que: se possui 1 indicação, é denominado pouco; se possui entre 2 e 4, médio; se possui 5 ou mais, muito.

Outro caso de variável dispersa considerada no banco foi a quantidade de inspeções já feitas na unidade consumidora. Como o percentual de UCs que já foram inspecionadas anteriormente é limitado, já que se utiliza o resultado de uma das inspeções como o target, a subdivisão foi binária para cada uns dos resultados da inspeção. Dessa forma, considerou-se uma variável indicando presença de inspeção com situação normal e outra indicando presença de inspeção com fraude. Além disso, dentro desses atributos, contabilizou-se também inspeções em outras UCs que pertencessem ao mesmo proprietário.

Para as características de pagamento de contas, consideraram-se duas variáveis, uma para indicar a média de dias que o cliente leva para realizar um pagamento e outra para indicar quantas contas foram atrasadas em 36 meses. Ambas as métricas foram categorizadas de modo que a quantidade de atrasos foi separada em 3 classes e a média de dias de pagamento, em 4 classes. Os limites foram definidos conforme Quadro 7 e Quadro 8 a seguir

Quadro 7 - Limites de categorização para a variável quantidade de atrasos.

Quantidade de Atrasos	Categoria
(0, 3)	Baixa
[3, 6)	Média
[6, 36]	Alta

Quadro 8 - Limites de categorização para a variável média dias de pagamento.

Média Dias de Pagamento	Categoria
(0, 8)	Baixa
[8, 15)	Média
[15, 31)	Acima da Média
[31, ∞)	Alta

Por fim, as últimas variáveis contínuas implementadas que necessitavam mapeamento foram relacionadas a quantidade de pontos na curva de consumo fora do intervalo entre o 1º e o 3º quartil. Quatro atributos foram criados: dois para a série completa e dois para a série referente aos últimos 12 meses. Os dois atributos de cada foram separados em: quantidade de pontos abaixo do 1º quartil e quantidade de pontos acima do 3º quartil. Os limites determinados para categorização das variáveis foram: para a série completa, menor que 3, menor que 5 e maior que 5; para a série de 12 meses, igual a 1, igual a 2 e maior ou igual a 3. Foram gerados também outros dois atributos semelhantes aos citados para a série da primeira derivada do consumo. Os limites considerados para essas variáveis foram idênticos aos utilizados para a série de 12 meses.

3.2.2 Variáveis Geradas por Atividade Econômica

Uma importante característica das unidades consumidoras pertencentes à classe de consumo comercial, industrial, poder público ou serviço público é o tipo de atividade. No Brasil, a classificação oficial adotada é a Classificação Nacional de Atividades Econômicas (CNAE) gerida pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Com base no CNAE Subclasses 2.3 disponível em IBGE (2020), as atividades econômicas são divididas em 20 seções, 86 divisões, 282 grupos, 670 classes e 1331 subclasses.

No sistema da empresa de estudo, está disponível a divisão de atividade por unidade consumidora. Entretanto, o que se observa é a existência de informações obsoletas ou imprecisas para grande parte dos dados. Dessa maneira, propõe-se uma correção desse cadastro de atividades com base nos dados públicos do Cadastro Nacional da Pessoa Jurídica (CNPJ) fornecido pela Receita Federal do Brasil.

Para a correção fez-se:

a) O cruzamento do CNPJ da base pública com o CNPJ do cadastro da empresa, quando disponível, para identificação da atividade econômica;
b) Para os casos que não foram identificados, o cruzamento das informações de e-mail, telefone e/ou nome da unidade consumidora para especificação do CNPJ;

c) Verificação da localidade, incluindo o bairro, dos dois cruzamentos, a fim de evitar novas inconsistências.

A atualização da atividade econômica a partir do CNPJ incluiu também as informações referentes ao porte da empresa, indicação de optante do Microempreendedor individual (MEI) e indicação de optante pela tributação SIMPLES. As três foram convertidas em variáveis para o modelo.

Em relação a atividade econômica, optou-se por considerar a Seção presente no CNAE como variável para o modelo. A quantidade de UCs em cada seção pode ser observado no Quadro 9 a seguir. Para as ocorrências zeradas, decidiu-se expurgar tais variáveis do modelo, agregando-as em três outras com seções genéricas: outros comércios, outras indústrias e outros poderes públicos. Essas seções genéricas foram utilizadas também para incluir as UCs em nome de pessoas físicas, que não possuem dados em bases públicas, quando não disponível a informação na referência de Divisão de Atividade da tabela de cadastro.

Quadro 9 - Quantidade de UCs na base de treinamento por Seção.

Seção - Descrição	Qtd na Base
Administração pública, defesa e seguridade social	1601
Agricultura, pecuária, produção florestal, pesca e aquicultura	31
Água, esgoto, atividades de gestão de resíduos e descontaminação	107
Alojamento e alimentação	1078
Artes, cultura, esporte e recreação	406
Atividades administrativas e serviços complementares	501
Atividades imobiliárias	135
Comércio; reparação de veículos automotores e motocicletas	1735
Educação	240
Indústrias de transformação	641
Informação e comunicação	414
Outras atividades de serviços	735
Saúde humana e serviços sociais	167

Para os comércios praticados por pessoas físicas, como não existe dados em bases públicas, foi utilizada a referência de Divisão de Atividade da tabela de cadastro.

3.2.3 Variáveis Geradas por Processamento de Linguagem Natural

Algumas características das unidades consumidoras nem sempre estão presentes diretamente no cadastro da empresa ou não foram informadas pelo leiturista ou estão desatualizadas nos sistemas comerciais. Uma maneira de obtê-las é utilizando o processamento de linguagem natural em mensagens e observações que foram digitadas pela central de

atendimento ou pelos inspetores técnicos ao executar um serviço. O processamento de linguagem natural foi utilizado neste trabalho para obter informações significantes a partir de textos não estruturados identificando tópicos, padrões e palavras-chave consideradas relevantes. Para implementação dos códigos, foram usadas as bibliotecas do Natural Language Toolkit (NLTK) para Python 3.

O primeiro passo da técnica utilizada consistiu em retirar as pontuações e as denominadas *stopwords*, expressões frequentes de uma determinada língua sem significado relevante para a informação final do texto. Exemplos de *stopwords* são artigos, pronomes e advérbios. Após retirados os termos menos significantes, aplicou-se o processo de *stemming* que identifica variações morfológicas nas palavras, mantendo apenas uma das formas ou o seu radical; exemplo, externo, externalizado, externalização são transformados em "extern".

Por fim, observou-se o espectro de frequência das palavras mais comuns a fim de extrair informações desconhecidas ou complementares para a identificação de fraude. Essa metodologia foi aplicada nas observações de serviço, ou OSs, e no endereço dos imóveis. O conteúdo obtido foi classificado em uma das variáveis a seguir: presença de CPRede, medição externa, unidade desligada, unidade desocupada, medição faz parte de um conjunto habitacional. A maior parte dessas variáveis serviu apenas de complemento para as implementadas anteriormente, com exceção da última citada. Para essa, gerou-se um novo atributo que indica se o imóvel faz parte de um conjunto habitacional, como é o caso de um apartamento ou uma casa em condomínio fechado.

3.2.4 Variáveis Geradas por Georreferenciamento

As métricas de consumo geradas até então não solucionam o problema exposto por Viegas et al. (2017) relativo à identificação de irregularidades em unidades que não possuem variação no histórico de consumo. Dessa maneira, propõe-se uma nova variável em que o consumo de uma unidade é comparado com a de seus vizinhos semelhantes. O objetivo é encontrar imóveis com padrões de consumo abaixo do esperado para a mesma classe de consumo de uma mesma região.

Para isso, fez-se o uso das coordenadas disponíveis na tabela de cadastro presente no Quadro 4. Como toda visita a uma UC atualiza os dados de latitude e longitude nos sistemas, essa informação é considerada confiável e, em geral, completa na base de dados da empresa. Para os poucos casos sem essa referência, utilizou-se a API Nominatim para buscar dados na base pública e cooperativa do *OpenStreetMap* (OSM) através do endereço e complemento também disponíveis na tabela de cadastro. Se mesmo assim não forem encontrados valores de

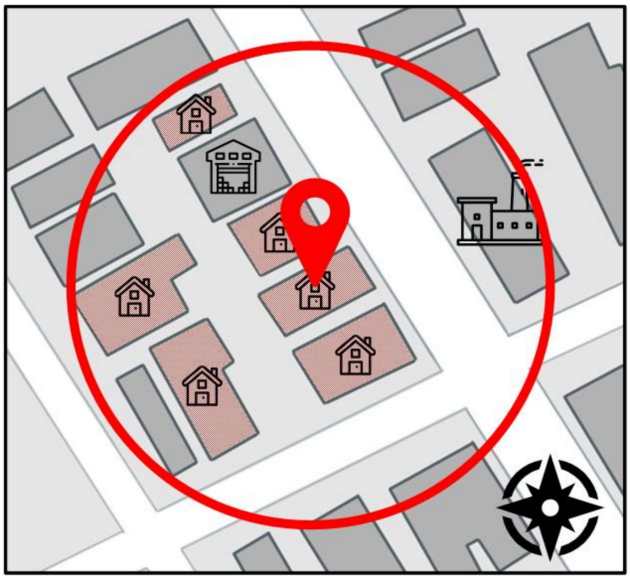
latitude e longitude para o imóvel, considerou-se a mesma coordenada de outra unidade no mesmo bairro ou no mesmo município, respectivamente.

A base dos vizinhos foi montada a partir de uma nova busca no banco de dados com todas as unidades consumidoras da empresa que atendessem os seguintes critérios:

a) Sem histórico de fraude ou irregularidades na medição que gerassem perda;
b) Não ficaram desocupadas ou desligadas no período de análise;
c) Não faturaram pelo mínimo no período de análise;
d) Sem indicação de leiturista de suspeita de fraude ou medidor com defeito;
e) UC ligada e faturando há pelo menos 12 meses;
f) Possui média de consumo mensal maior que 30 kWh.

A concepção da métrica encontra-se ilustrada na Figura 23, em que o símbolo de local, no centro da circunferência, indica a UC de referência, e as cinco casas indicadas ao redor, dentro do raio vermelho, são os vizinhos semelhantes mais próximos a ela.

Figura 23 - Concepção das variáveis associadas ao consumo que compara vizinhos geograficamente próximos e com características semelhantes a uma UC



Fonte: Elaborado pelo autor.

Para determinar quais vizinhos serão utilizados para compor a média de consumo comparativa de um cliente, primeiro um conjunto de consumidores da mesma classe de consumo e com o mesmo tipo de ligação foi selecionado. A partir desse conjunto, utilizou-se a fórmula de Haversine para calcular as distâncias entre o cliente de referência e as demais unidades a partir de suas latitudes e longitudes. Em seguida, tomou-se os 5 pontos mais próximos geograficamente e descartou-se os de maior e menor consumo, a fim de evitar grandes desvios no valor da média. Quando possível, foram priorizadas as UCs do mesmo bairro. Por fim, a variável denominada “Média Vizinhos” foi calculada através da média dos 3

consumidores finais. O número de unidades para composição da média foi determinado de forma empírica, em que o valor ótimo obtido para uma maior efetividade foi de 3.

Outra métrica gerada foi a chamada “Degrau Vizinhos”. Essa variável foi obtida calculando o desvio percentual da média de consumo dos últimos 12 meses da UC em relação à média dos vizinhos semelhantes, análogo ao degrau obtido através da equação (1).

Para as UCs associadas a algum tipo de atividade econômica, gerou-se também as variáveis “Média Vizinhos Atividade” e “Degrau Vizinhos Atividade” que, além de considerar as unidades mais próximas geograficamente, considerou também a Seção de CNAE com base no Quadro 9 e, quando possível, o porte da empresa.

3.2.5 Variáveis Geradas pelo Teste de Chow

Detecção de quebras estruturais é uma temática da área de econometria e estatística que busca observar mudanças nos parâmetros dos modelos de regressão que fazem com que predições e análises de impactos sejam prejudicadas (Berry et al., 1995). O parâmetro clássico para detecção dessas quebras é o teste de Chow e suas variações.

Esse teste foi empregado nas curvas de consumo de cada unidade, mais especificamente, para detecção de degraus negativos, em que houve uma redução brusca no montante de energia. A aplicação do teste de Chow seguiu a metodologia conhecida por Teste Sup F, em que se observa a curva mês a mês a fim de avaliar se houve mudanças ao longo do tempo.

O resumo em pseudo-código pode ser observado na Figura 24, em que o f_critico foi calculado com base no valor crítico da distribuição F. Nota-se que uma verificação é feita para identificar se a quebra estrutural encontrada corresponde a um aumento de consumo. Se sim, a variável degrau é zerada, bem como o mês de detecção do degrau. A menos que haja outra quebra com redução de consumo, não será considerado existência de degrau para esse caso.

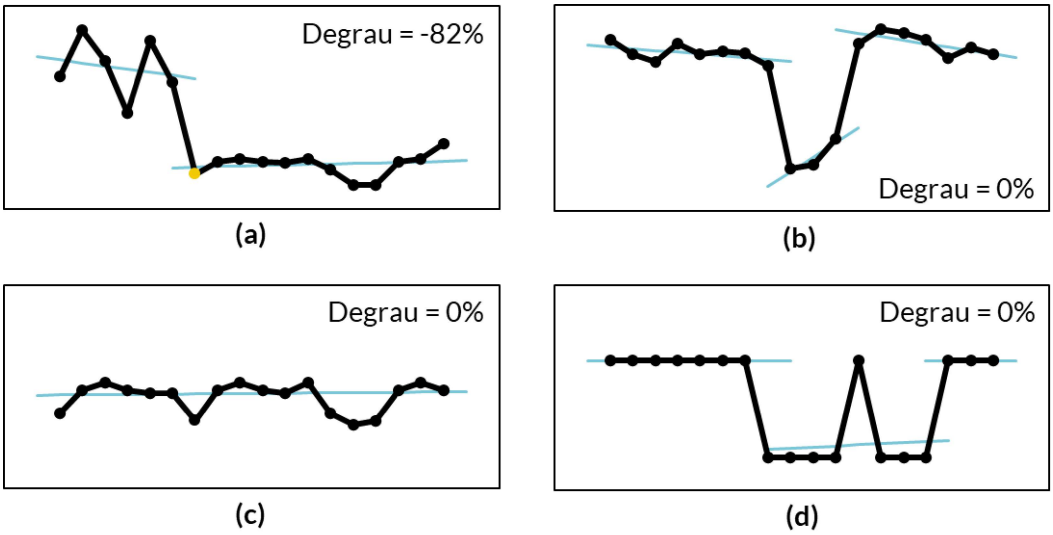
Figura 24 – Pseudo-código para detecção de degrau pelo Teste de Chow.

```
Para cada uc:
    degrau_min = 0
    mês_degrau = 0

    Para mês em curva:
        Calcula f_chow

        Se f_chow > f_critico:
            degrau = (curva[mês:] - curva[:mês])/curva[:mês]

            Se degrau < degrau_min:
                degrau_min = degrau
                mês_degrau = mês
            Se não se degrau > 20%:
                degrau_min = 0
                mês_degrau = 0
```

Fonte: Elaborado pelo autor.

De posse do mês em que houve uma quebra estrutural, calcula-se o degrau a partir de (1), em que $cons_{atual}$ é a média de consumo após o mês da quebra e $cons_{ant}$ é a média de consumo mês anterior a ele. Essa variável foi denominada Degrau Chow. Para uma quebra estrutural com redução de consumo, conforme a Figura 25(a), o degrau de consumo é identificado pela variável Degrau Chow, bem como seu mês de ocorrência. Para quebras estruturais conforme a Figura 25(b) e (d), em que o consumo retomou os padrões anteriores, não há identificação de degrau. Na Figura 25(c), não há quebra estrutural.

Figura 25 – Exemplos do comportamento da variável degrau Chow.



Fonte: Elaborado pelo autor.

O mês de quebra foi utilizado também como referência para geração de outras variáveis que tinham como objetivo justificar ou mitigar a redução de consumo. Dessa maneira, implementou-se uma variável de indicação de manutenções ou inspeções na unidade consumidora após o degrau Chow, indicação de desocupado, desligado ou faturamento pelo mínimo após o degrau Chow e indicação de suspeita de fraude após o degrau Chow.

3.3 Pré-Processamento

Após extração das características dos clientes, faz-se necessário limpar a base de unidades que possam prejudicar o desempenho do treinamento dos modelos, agrupar, redistribuir e normalizar as variáveis obtidas. Essas técnicas foram consideradas visando a redução da taxa de erro e reduzir o tempo de construção de um modelo.

O filtro aplicado ao banco, conforme citado anteriormente, consistiu em retirar as UCs com menos de 24 pontos de consumo para cálculo das métricas. O objetivo desse filtro foi desconsiderar unidades desligadas durante muito tempo ou aquelas que foram ligadas em um

tempo recente. Em ambos os casos, a maioria das variáveis de consumo se tornam irrelevantes, já que não é possível calculá-las. Além disso, observa-se que, comumente, os clientes não atingem os patamares de consumo padrão nos primeiros 12 meses da data de ligação.

A partir da análise exploratória dos dados, verificou-se que algumas variáveis possuem uma baixa frequência em relação as demais. Dessa maneira, fez-se uma varredura nos atributos categóricos e optou-se por agrupar: as seis classes de consumo em três (residencial, rural e outros); as características blindagem e DLCB; as características CPRede e externalização. As classes de consumo foram assim classificadas devido a diferença tanto na quantidade de UCs quanto no padrão de consumo que existe entre a classe residencial e as demais. Já as demais características foram devido às semelhanças entre seus objetivos: tanto a CPRede quanto a externalização objetivam manter o padrão de medição fora do imóvel, facilitando a observação de um procedimento irregular; enquanto o DLCB e a blindagem visam dificultar fisicamente a manipulação ou o furto de energia.

Aplicados os filtros e finalizada a organização da base de dados, foram utilizadas as técnicas de normalização das variáveis contínuas. Levando em consideração as vantagens e desvantagens dos diferentes métodos existentes, as técnicas de transformação linear (mín-máx) em conjunto com o *clip* para valores fora de uma faixa determinada foram utilizadas. O escalonamento foi feito para manter os valores sempre entre 0 e 1.

O resumo dos métodos utilizados, bem como os valores limites considerados, está presente no Quadro 10.

Quadro 10 - Tipos de normalização por variável e os limites utilizados

Variáveis		**Técnica**	**Limites**
Máximo, Mínimo, Média e Consumo Atual	Residencial	Min-Max + Clip	0 – 12.300
	Rural	Min-Max + Clip	0 – 69.200
	Outros	Min-Max + Clip	0 – 46.500
Desvio Padrão	Residencial	Min-Max + Clip	0 – 6.150
	Rural	Min-Max + Clip	0 – 34.600
	Outros	Min-Max + Clip	0 – 23.250
Curtose		Min-Max + Clip	-3 – 7
Degrau		Min-Max + Clip	-1 – 5
Assimetria		Min-Max	-3 – 3
Pearson		Min-Max	-1 – 1
Shapiro-Wilk Test		Min-Max	0 – 1

Para duas variáveis com a mesma característica, como degrau de consumo e degrau vizinhos, foi utilizada a mesma técnica e o mesmo limite. Na maioria dos casos, esses limites foram determinados com base no observado para a base de treinamento. Para as variáveis de

consumo, no entanto, levou-se em consideração o maior consumo por classe no histórico da empresa. Não foi utilizado o máximo do banco de treinamento, pois o maior consumo possível de ser aplicado no modelo pode não estar presente nesse conjunto de dados e utilizá-lo pode ser uma forma de enviesar a base e manter um problema com números fora do intervalo. Os problemas de números fora do intervalo, conforme visto, envolvem o desconhecimento do alcance de uma variável e a ausência de exemplos para as ferramentas de modelagem preverem seus comportamentos.

3.4 Seleção de Variáveis

Com todas as variáveis extraídas e normalizadas, o próximo passo envolve a seleção dos atributos que funcionarão como entrada do sistema. *Features* redundantes podem ser eliminados sem que haja perda de informação. Em aplicações reais, como existe uma quantidade de dados limitada, deseja-se evitar a maldição da dimensionalidade definida no tópico 2.6.1 para que não haja uma redução na performance dos classificadores.

Dessa forma, o primeiro passo na seleção de variáveis é aplicar técnicas para reconhecer correlações. Inicialmente, utilizou-se o coeficiente de correlação de Pearson para identificar todos os atributos com r maior que 0,8 ou menor que -0,8. De posse desses atributos, os conjuntos correlacionados foram analisados e apenas uma das variáveis foi escolhida para compor a entrada do sistema. Teoricamente, essa escolha pode ser arbitrária, já que as variáveis carregam a mesma informação. Entretanto buscou-se selecionar aquelas que, para um especialista da área de combate às perdas, pareciam as mais relevantes para identificação de uma fraude.

3.5 Aplicação dos Modelos

Com as variáveis normalizadas e selecionadas, partiu-se para a divisão da base em treinamento e teste a fim de aplicá-las aos modelos. Para separar os exemplos, selecionou-se UCs aleatoriamente e buscou-se manter a mesma proporção de fraude da base original na base de teste. Isso foi feito para que o resultado obtido na simulação do modelo fosse o mais próximo possível do que deve ser encontrado em campo. Como mencionado anteriormente, o percentual de unidades fraudadoras no conjunto de teste pode afetar o resultado, conforme analisado por Angelos et al. (2011). Para a base de treinamento, ora utilizaram-se as UCs restantes, ora utilizou-se um conjunto com a mesma proporção de fraude e situação normal. Essa escolha foi determinada de acordo com a técnica de aprendizado de máquina empregada: algumas só conseguem classificar se for utilizada uma base de treino proporcional, enquanto outras podem

obter um melhor desempenho com uma base desproporcional. As técnicas avaliadas incluíram: Árvore de Decisão, Naive Bayes, K-Vizinhos Mais Próximos (KNN), *Random Forest*, *Support Vector Machine*, RNA e *Gradient Boosting*.

Com a base separada, o processo de avaliação foi semelhante para todos os modelos aplicados e seguiu-se conforme as alíneas apresentadas a seguir:

a) Escolha dos principais parâmetros: para cada técnica determinou-se empiricamente valores para os principais parâmetros e diferentes modelos foram avaliados para uma mesma técnica;
b) Treinamento com *holdout* aleatório estratificado, aplicada 10 vezes para exportação da matriz de confusão, f-score e intervalo de confiança obtidos;
c) Treinamento com base completa;
d) Classificação da base teste: exportação da matriz de confusão, tomado como uma simulação de um caso real.

O segundo passo auxiliou na determinação das técnicas que seriam utilizadas para os passos c e d. No final, 4 modelos distintos foram montados a partir dos resultados do treinamento e teste: residencial sem indicação de suspeita de fraude; residencial com indicação de suspeita de fraude; rural; outros. As técnicas que apresentaram os melhores desempenhos para cada base foram escolhidas na etapa b. O *holdout* estratificado foi aplicado para determinar o modelo que possuía a melhor performance com uma base de teste tão desbalanceada quanto se espera encontrar em campo.

Escolhida as técnicas a serem empregados nos modelos, partiu-se para o treinamento com a base completa de treinamento e fez-se a simulação da efetividade e cobertura utilizando um novo conjunto de dados com uma proporção próxima a esperada na base real. Os resultados obtidos foram descritos no Capítulo 4.

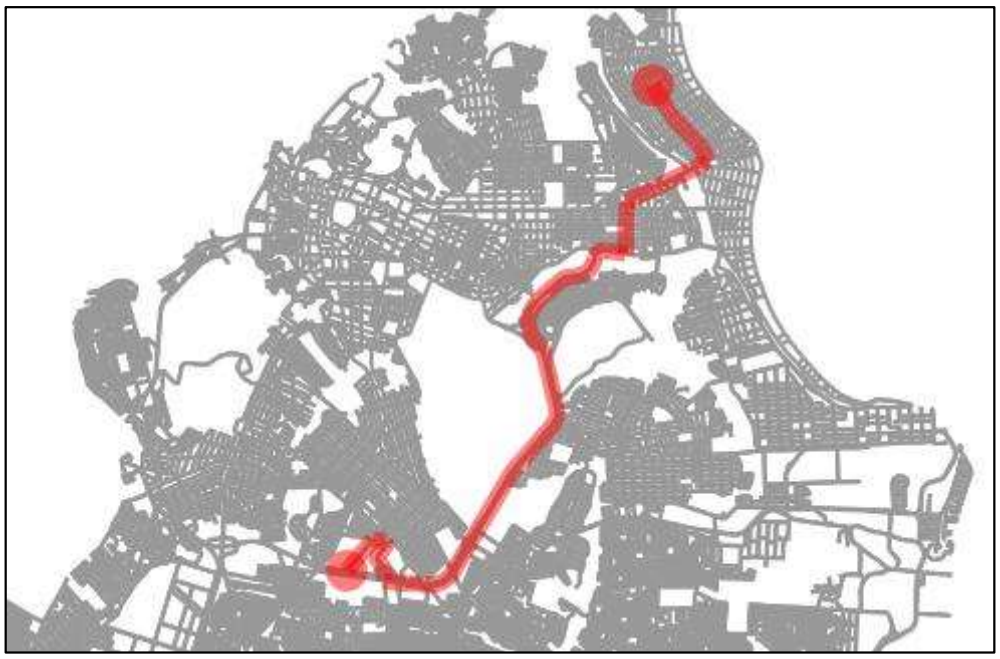
3.5.1 Indicador Benefício do Modelo

As principais métricas de avaliação da literatura, como o f-score e a AUC ROC haviam sido, inicialmente, escolhidas para determinar os modelos. Entretanto, notou-se uma necessidade de um novo indicador devido a semelhança entre os f-scores de algumas técnicas, que possuíam efetividades e coberturas distintas. Em termos de benefício para a empresa, não era possível afirmar, através dessas métricas, qual seria o modelo mais eficaz. Dessa maneira, o Indicador de Benefício foi proposto. A concepção do Indicador Benefício do Modelo inclui o conceito de custo de inspeção e recuperação de energia no caso de identificação de um procedimento irregular.

O custo da inspeção é composto basicamente pelos custos de deslocamento, da mão de obra e do material utilizado. Para esse cálculo, foi considerada a distância da rota entre a base da empresa e a unidade consumidora através da biblioteca OSMnx proposto por Boeing (2017). A rota é calculada pelo caminho mais curto considerando uma rede de rotas por viagens com carro, como pode ser observado na Figura 26. A partir das coordenadas da base da empresa e as coordenadas da UC aplicado ao OSMnx, o tempo de deslocamento foi obtido dividindo a distância da rota por uma velocidade média de 55 km/h. Essa velocidade foi assumida a partir de um estudo empírico interno para as viaturas com limite máximo de 80 km/h. A duração da inspeção também deve ser considerada e a média da empresa de estudo é 20 min para padrões do grupo B com ligação direta.

Considerando a soma do tempo de deslocamento e de inspeção, o custo total, em reais, foi calculado a partir do valor da hora de uma equipe. Para a transformação desse montante em energia (kWh), foi utilizada a tarifa média de venda sem imposto em R$/kWh.

Figura 26 – Exemplo de rota para inspeção de uma UC.



Fonte: Elaborado pelo autor.

A recuperação de energia, por sua vez, foi estimada de duas maneiras, a depender da existência de um degrau de consumo ou não. Caso exista o degrau, detectado a partir do teste de Chow mencionado no tópico 3.2.5, a recuperação será conforme regras dos procedimentos irregulares da resolução normativa nº 414 discutidas na Seção 2.3: serão recuperados os meses desde o degrau, com limite de 36 meses, considerando a média de consumo dos 3 maiores valores nos últimos 12 meses como referência para cálculo do consumo não faturado. Caso não exista o degrau, e como não há uma maneira de aplicar os demais critérios da resolução para o

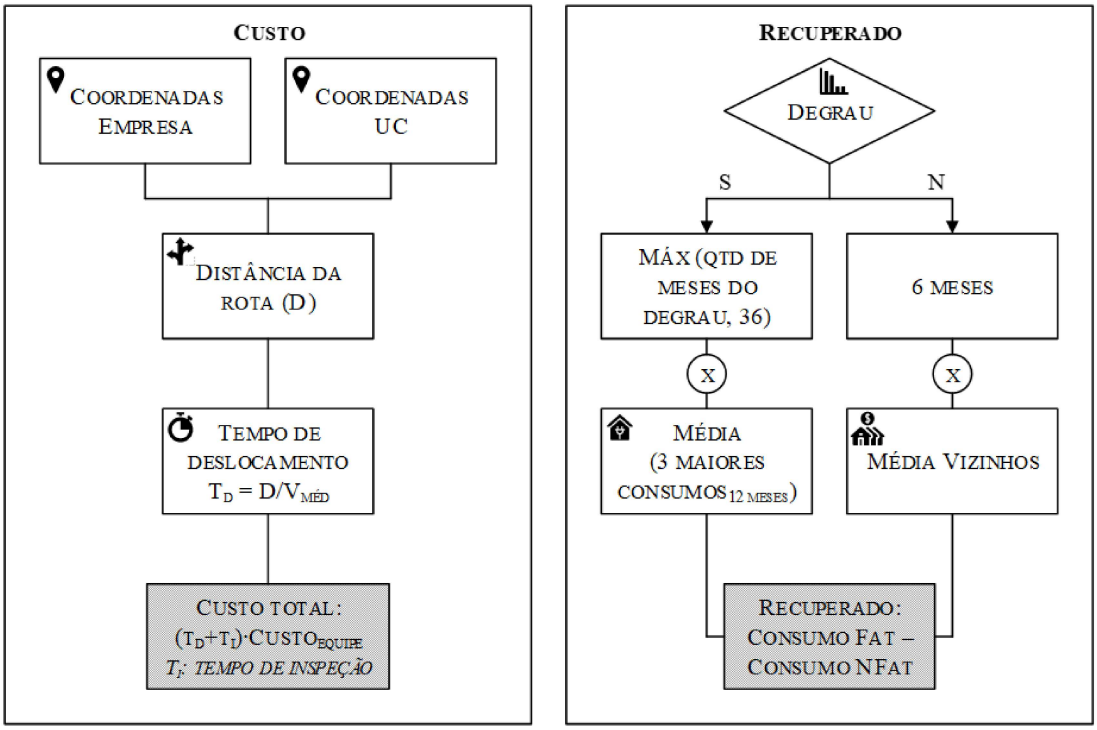
cálculo do consumo de referência, propõe-se uma simulação do consumo real esperado tendo como base unidades semelhantes a ela. Naturalmente, a variável Média Vizinhos apresentada na seção 3.2.4 mostra-se como propícia para tal atribuição, já que considera UCs de mesma classe de consumo, mesmo tipo de ligação e com maior proximidade geográfica. Dessa forma, o consumo não faturado é calculado pela subtração da Média Vizinhos com cada consumo dos últimos 6 meses, limitação pela norma devido a impossibilidade de identificar o período de duração da irregularidade.

Finalmente, o Indicador Benefício do Modelo proposto por este trabalho é calculado a partir da equação (14).

$$\text{Indicador Benefício do Modelo} = \frac{\text{tp} \cdot (\text{rec} - \text{cust}) - \text{fp} \cdot (\text{cust})}{(\text{tp} + \text{fn}) \cdot (\text{rec} - \text{cust})} \qquad (14)$$

Em que tp, fp e fn são os verdadeiros positivos, falsos positivos e falsos negativos, respectivamente, da matriz de confusão da Figura 14, e rec e cust referem-se ao recuperado e ao custo total, nessa ordem, calculados em kWh. O resumo das etapas de cálculo do indicador encontra-se no fluxograma da Figura 27.

Figura 27 - Fluxograma simplificado do cálculo do custo e do recuperado que compõem o Indicador Benefício do Modelo.



Fonte: Elaborado pelo autor.

Como métrica de avaliação, o indicador representa o percentual do benefício líquido do modelo em relação ao benefício total que poderia ter sido recuperado se todas as UCs fossem corretamente classificadas. Se o custo para inspecionar as unidades indicadas superar a energia

recuperada, o indicador será negativo e o modelo não compensa. Ao mesmo tempo, quanto mais assertivo em unidades consumidoras com potencial de alta recuperação, melhor será o indicador.

Um exemplo do uso do indicador pode ser observado no Quadro 11. Existem 4 unidades consumidoras nessa base fictícia, em que três possuem fraude e uma não. O Modelo 1 classificou a UC A corretamente com uma recuperação líquida de 2.555 kWh, o Modelo 2, a UC B, com 2.322 kWh e o Modelo 3, a UC C, com 172 kWh. Nota-se que a maior recuperação bruta é a do Modelo 1; entretanto, como o custo para inspeção dessa unidade é mais elevado, a recuperação líquida é menor que a do Modelo 2. O Modelo 3, por sua vez, apesar de classificar a UC que detém a menor distância até a base, possui um recuperado bem inferior aos outros e, portanto, possui um alto custo associado. Para os três casos, não há diferenciação entre os valores de efetividade, cobertura e, consequentemente, de f-score, o que ratifica a necessidade de uma nova maneira de avaliar os modelos.

Quadro 11 – Comparativo do Indicador Benefício para modelos diversos.

UC	Fraude	Rota (km)	Recup. (kWh)	Custo (kW)	Modelo 1	Modelo2	Modelo 3
A	S	191	3.200	845	S	N	N
B	S	20	3.000	155	N	S	N
C	S	1	250	78	N	N	S
D	N	1	6.200	78	N	N	N
Indicador Benefício do Modelo					**44%**	**53%**	**3%**
Efetividade					**100%**	**100%**	**100%**
Cobertura					**33%**	**33%**	**33%**

Se o perfil da UC D for de fraude, no entanto, o indicador é consideravelmente modificado, conforme pode ser visto no Quadro 12. Como a recuperação de consumo desse cliente representa um montante significativo em relação as demais UCs, a performance dos modelos que não foram capazes de classificá-la corretamente caem.

Quadro 12 – Alteração do Quadro 11 para unidade D com perfil de fraude.

UC	Fraude	Rota (km)	Recup. (kWh)	Custo (kW)	Modelo 1	Modelo2	Modelo 3
A	S	191	3.200	845	S	N	N
B	S	20	3.000	155	N	S	N
C	S	1	250	78	N	N	S
D	S	1	6.200	78	N	N	N
Indicador Benefício do Modelo					**20%**	**25%**	**1%**
Efetividade					**100%**	**100%**	**100%**
Cobertura					**25%**	**25%**	**25%**

4 RESULTADOS

O banco de dados para treinamento foi construído com 95.597 unidades consumidoras, em que 18.760, ou seja, 20% do total, foram classificadas como fraudadoras, constituindo um problema de classes desbalanceadas. Originalmente foram gerados 130 atributos, em que 91 eram categóricos e 39 contínuos. A análise exploratória e a seleção de variáveis contribuíram para a redução desse número, já que foram identificas possibilidades de agregar ou retirar objetos e, no final, foram considerados 105 atributos, sendo 81 categóricos e 24 contínuos. Os testes realizados em seguida levaram em conta apenas esse último conjunto.

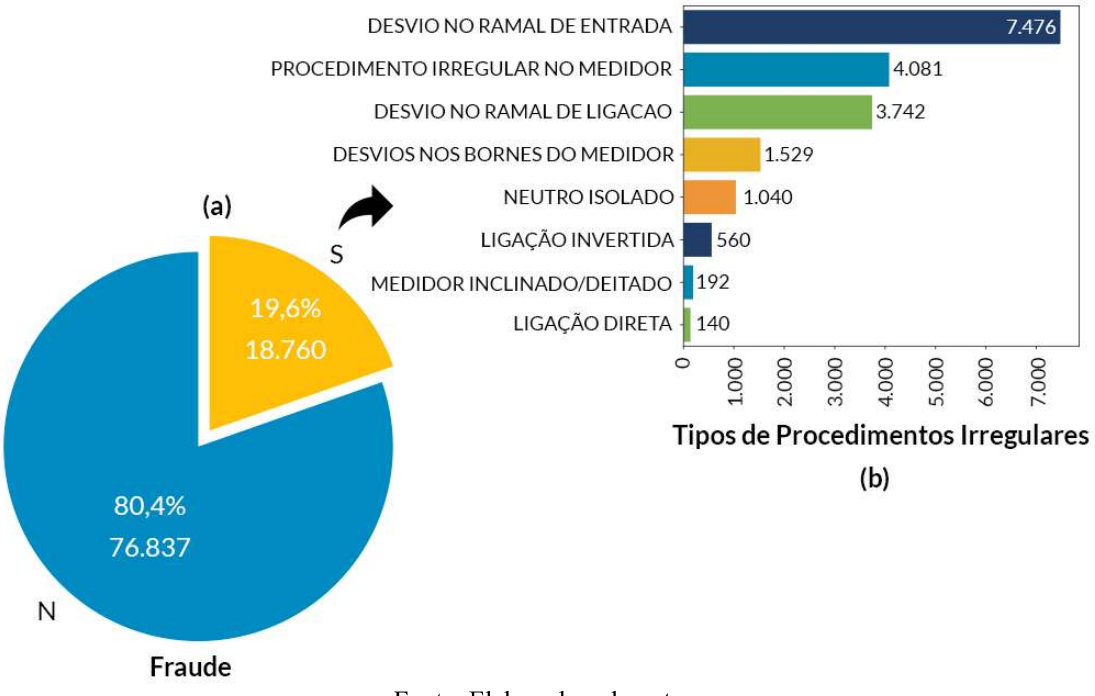
4.1 Análise Exploratória e Seleção de Variáveis

A análise exploratória consiste em verificar o espaço dos dados para descobrir e avaliar problemas apropriados, definir soluções e estratégias de implementação e produzir resultados mensuráveis (PYLE, 1999). A maior parte dessa análise foi feita com o auxílio de ferramentas gráficas que ajudam na compreensão das variáveis utilizadas.

4.1.1 Variáveis Categóricas

A distribuição da variável *target* pode ser vista na Figura 28(a) a seguir.

Figura 28 - Distribuição das unidades no banco de dados. (a) Variável *target* fraude. (b) Tipos de irregularidades.

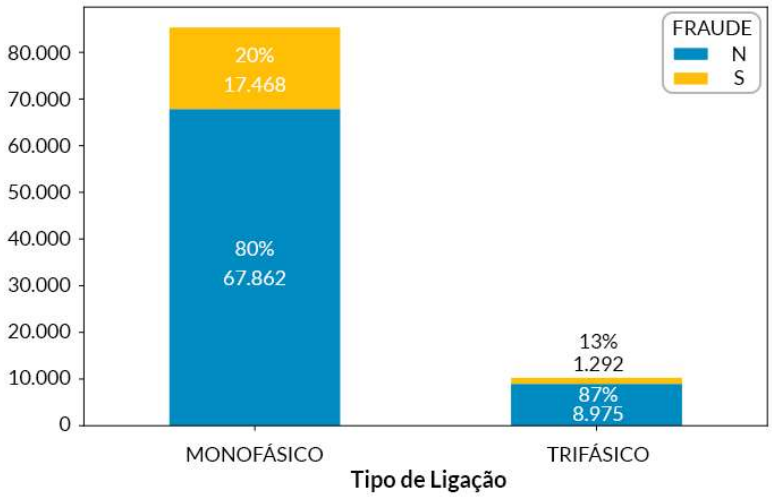


Fonte: Elaborado pelo autor.

Verifica-se, inicialmente, que existe um desbalanço no conjunto de treinamento, em que a classe de não fraudadores é muito superior à de fraudadores. É esperado, ainda, que essa desproporção seja maior quando observada toda a população. As ocorrências mais comuns, como pode ser observado na Figura 28(b), são os desvios e os procedimentos irregulares nos medidores, respectivamente.

Ao segregar esses atributos por tipo de ligação, como se observa na Figura 29, é possível verificar uma maior presença de ligações monofásicas, associadas ao volume de unidades residenciais. Nota-se também, que a frequência de fraude em UCs trifásicas é menor que nas monofásicas.

Figura 29 – Ocorrências de fraude por tipo de ligação.



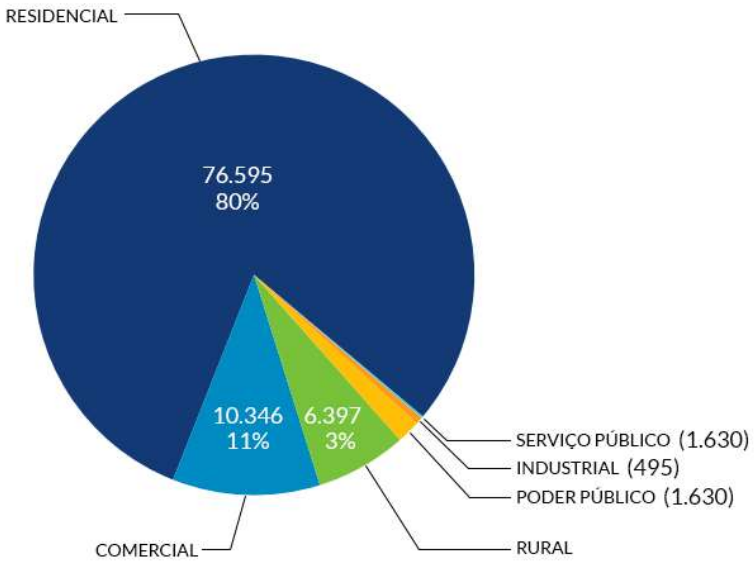
Fonte: Elaborado pelo autor.

A classe de consumo está distribuída conforme a Figura 30, em que se verifica que 80% de toda a amostra dos clientes do Grupo B são da classe residencial. Apesar dessa classe corresponder a grande maioria da quantidade de unidades, em termos de consumo mensal ela pode se tornar menos significante.

Na Figura 31(a) e (b), apresenta-se o cenário de consumo por classe. Na Figura 31(a), tem-se a média de consumo em kWh por classe, em que a classe residencial apresentou a menor das médias. Isso significa que encontrar uma única irregularidade em uma UC industrial pode corresponder ao montante de energia de 27 unidades residenciais. Para redução de custos de deslocamento e material, deseja-se encontrar irregularidades em unidades de maior porte, especialmente por significar uma maior redução na perda. O acumulado da média mensal dos consumos, em MWh, por classe está mostrado na Figura 31(b). Um destaque pode ser dado para a classe comercial, que apesar de corresponder a apenas 11% das unidades, o consumo acumulado dessas unidades chega a se equiparar ao de toda a classe residencial. Em contra

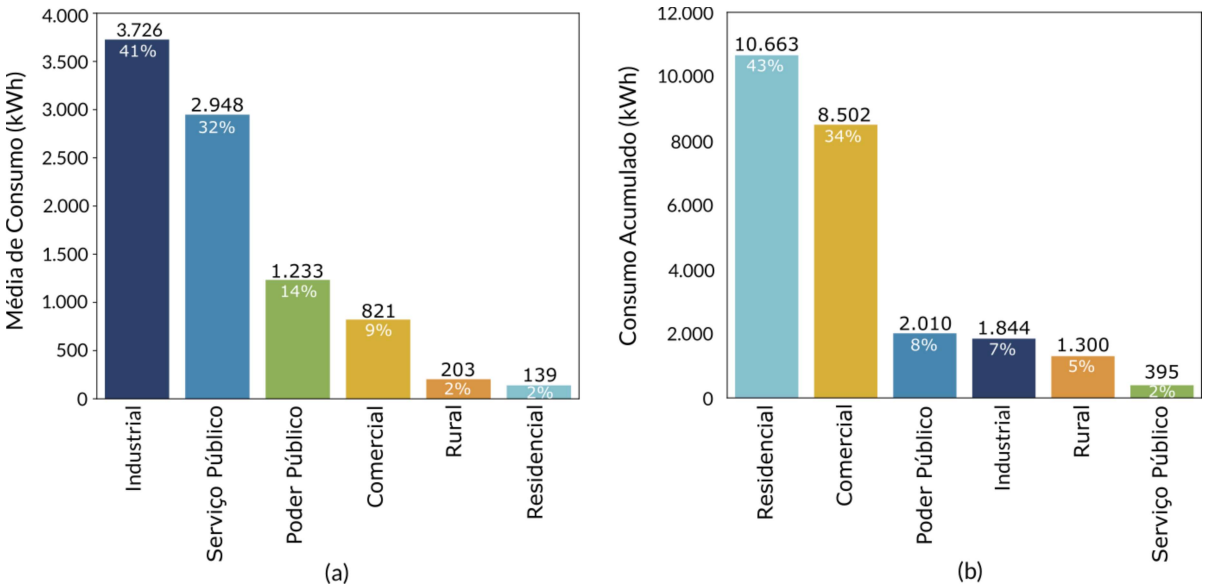
partida, identificar fraudes em grupos de clientes que possuem uma média de consumo maior pode ser mais desafiante. No banco de dados, o percentual de fraude tanto dos clientes industriais como comerciais é de 13%, enquanto dos clientes residenciais é de 20%.

Figura 30 - Distribuição da classe de consumo.



Fonte: Elaborado pelo autor.

Figura 31 – Comparativo da energia consumida por classe. (a) Média dos consumos mensais em kWh por classe de consumo. (b) Acumulado da média dos consumos mensais em MWh por classe de consumo.

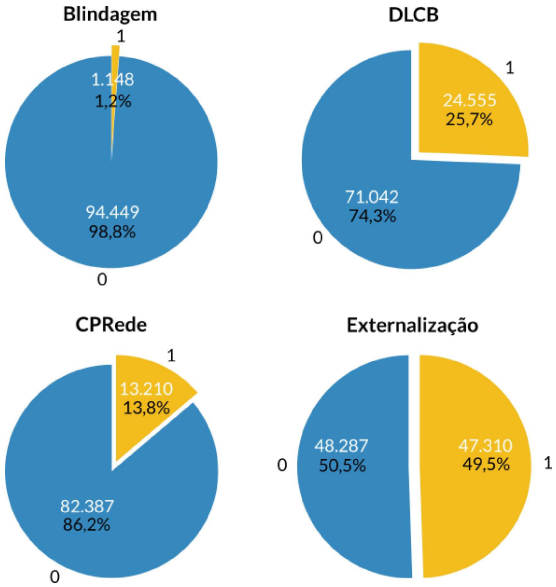


Fonte: Elaborado pelo autor.

Como a distribuição das classes de consumo é muito esparsa, levando em conta a quantidade de unidades por categoria, a semelhança da média de consumo e a curva de carga esperada, as classes foram agrupadas em três: residencial, rural e outros. Além disso, em vez de considerá-las como variáveis de entrada, serão contemplados diferentes modelos para cada uma dessas classes, já que os perfis de consumo podem ser muito divergentes entre si.

Outras variáveis que foram combinadas após observar as suas distribuições foram as características da medição (blindagem, DLCB, CPRede e externalização), conforme mencionado na seção 3.3. A quantidade de casos existentes no banco de dados pode ser verificada pela Figura 32, em que 1 indica a presença da característica e 0 indica a ausência. As variáveis blindagem e CPRede foram agregadas as de DLCB e externalização, respectivamente.

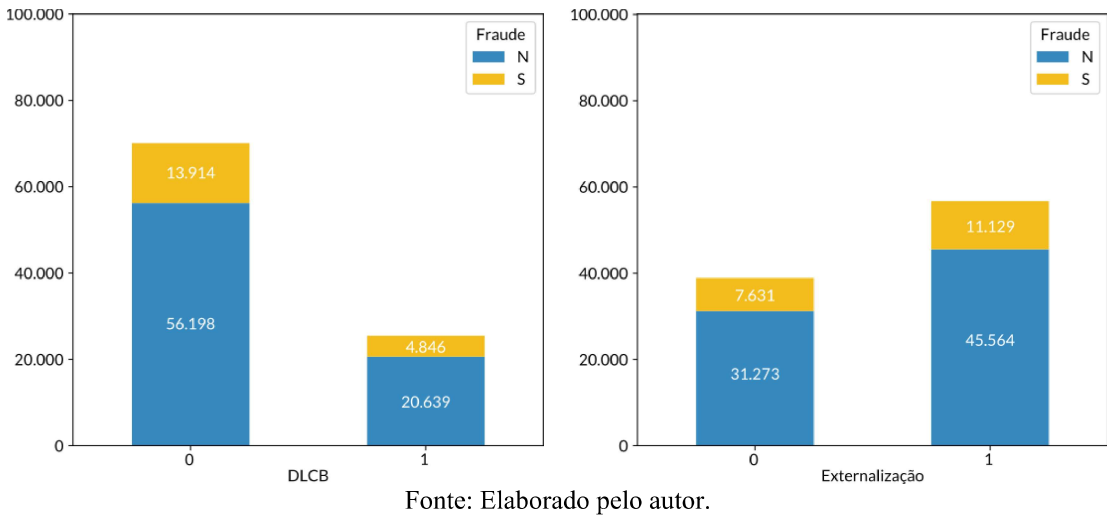
Figura 32 – Distribuição das variáveis de padrão de medição, em que 0 indica ausência e 1, presença.



Fonte: Elaborado pelo autor.

Considerando as variáveis após a integração, é possível observá-las em relação às ocorrências de fraude na Figura 33. Para todos os casos, a proporção de fraude manteve-se em 20%, indicando que essas variáveis isoladamente não apresentam melhora em relação a aleatoriedade. Vale ressaltar que isso não implica ineficácia da variável, já que agregada a outras informações, esta pode fornecer importantes referências para a classificação adequada das unidades.

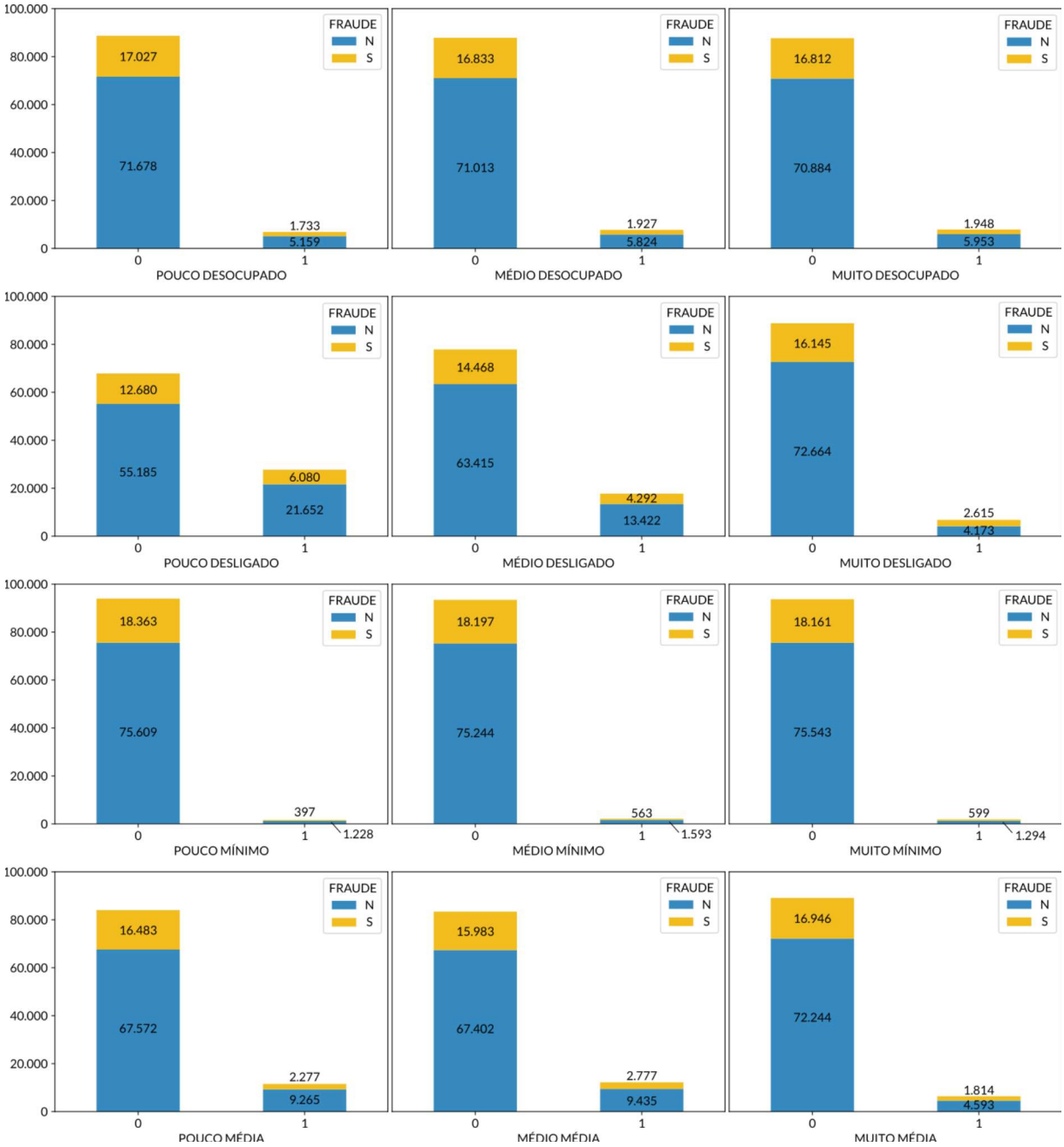
Figura 33 – Ocorrências de fraude nas variáveis de padrão de medição.



Fonte: Elaborado pelo autor.

Ainda acerca dos atributos categóricos, para as variáveis esparsas ou que necessitavam de mapeamento para eliminação de ordem, a distribuição foi fundamental para determinação dos limites para o *binning*. A conversão para atributos categóricos com a transformação em variáveis *dummy* está presente na Figura 34. Cada gráfico representa uma variável do modelo. O "0" indica a ausência da característica, enquanto o "1", indica a presença. Pode ser observado também a frequência de fraude para cada valor das métricas.

Figura 34 – Variáveis de irregularidade de leitura após o *binning* e conversão para *dummy*.

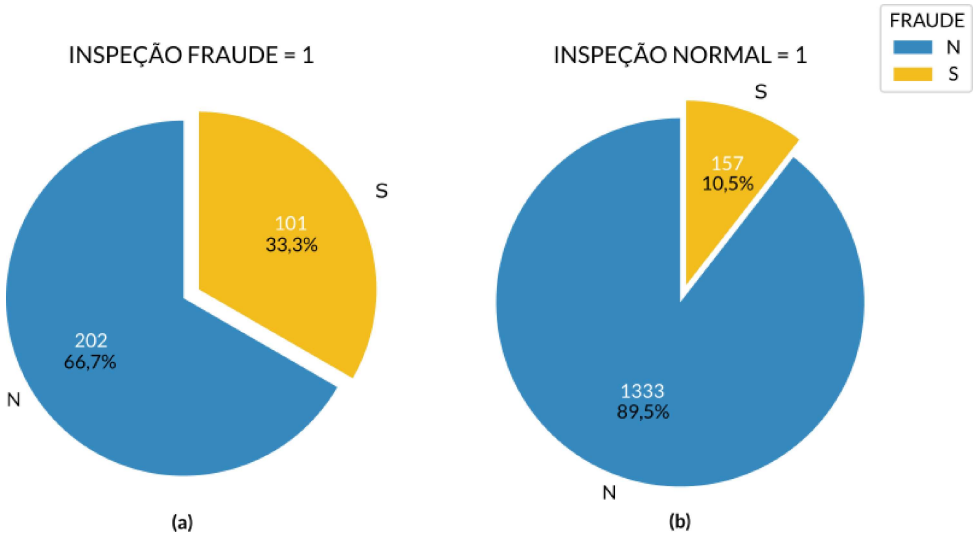


Fonte: Elaborado pelo autor.

Outro caso de variáveis esparsas considerado foi a quantidade de inspeções anteriores na unidade consumidora ou em outras unidades do mesmo dono. Como é possível observar na

Figura 35, existem pouquíssimos casos com a presença dessa variável, entretanto, observa-se que, isoladamente, ela pode trazer conclusões importantes. A variável indica que uma unidade em que já foi identificada com fraude anteriormente pode estar mais propícia a fraudar novamente. Para o banco utilizado, 33% das UCs que já foram identificadas com fraude voltaram a praticar irregularidades. Existe também indicativo de que as UCs que já foram inspecionadas e não houve identificação de anormalidade, mantém esse mesmo perfil. No banco esse percentual é de 89%.

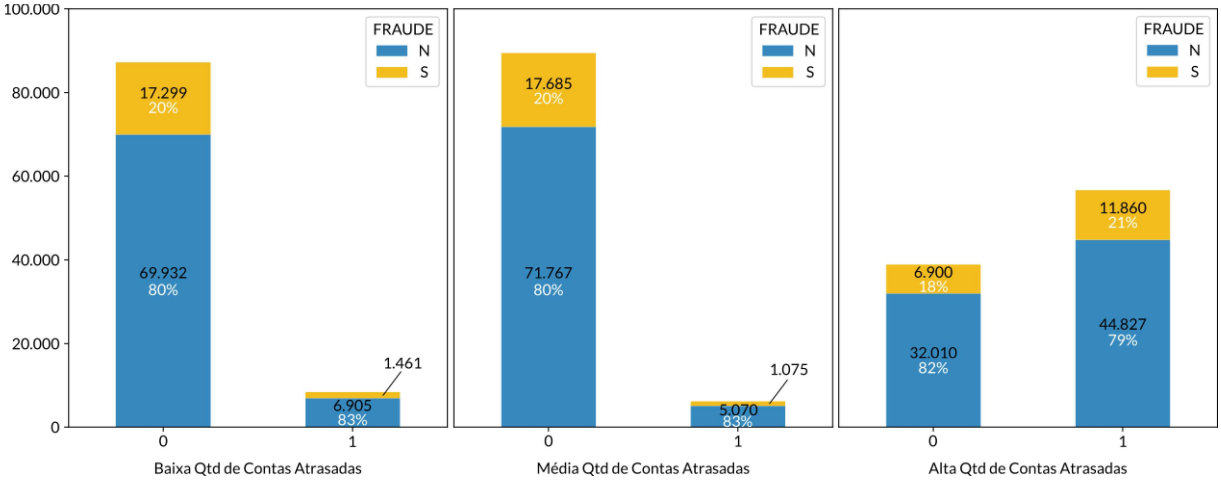
Figura 35 - Variáveis sobre inspeções anteriores separadas por ocorrências de fraude da variável *target*. (a) Inspeções anteriores em que se detectou fraude. (b) Inspeções anteriores em que não se detectou irregularidade.



Fonte: Elaborado pelo autor.

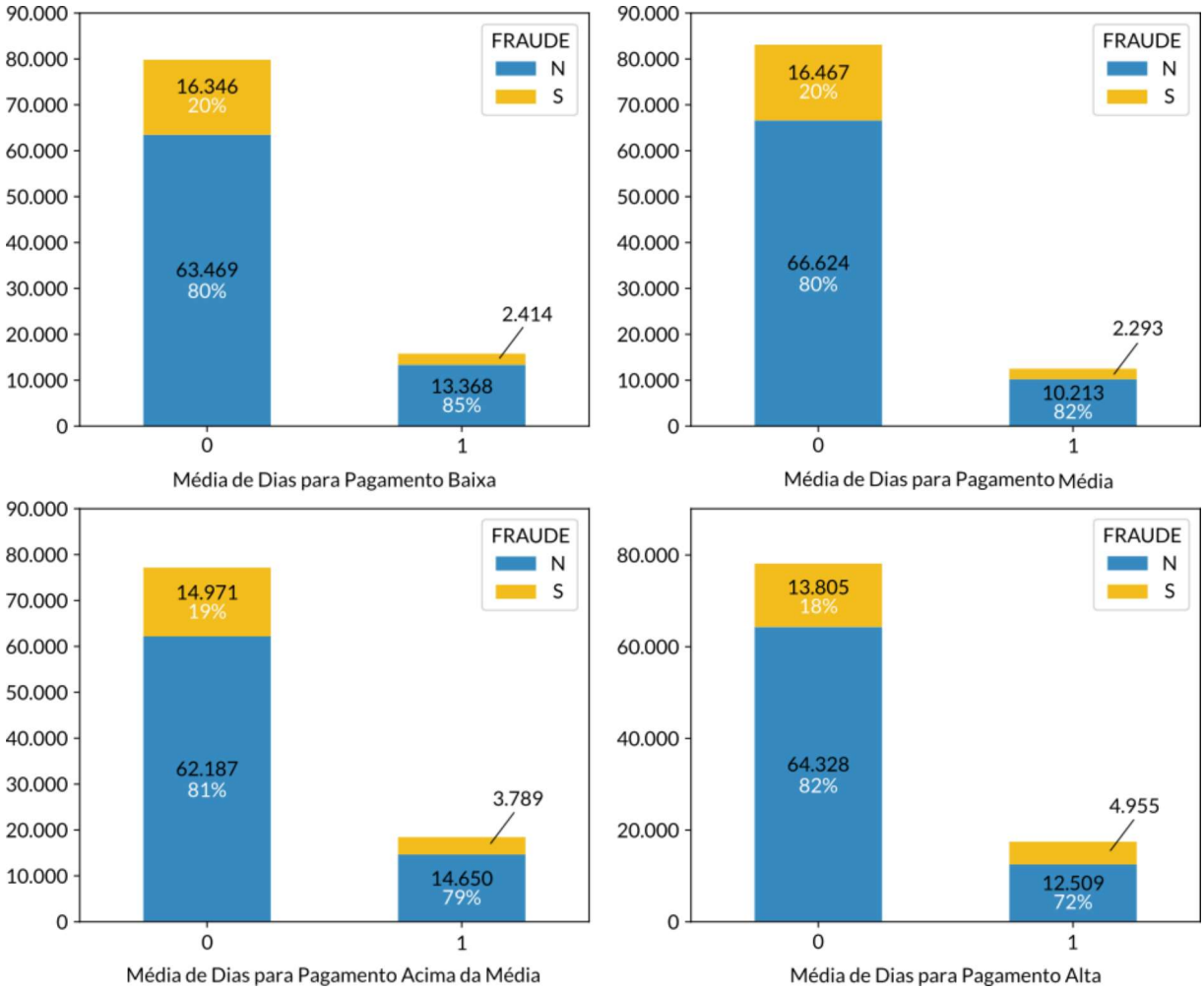
As variáveis que caracterizam o pagamento das contas de energia foram implementadas visando evidenciar as unidades inadimplentes. Como já mencionado, elas foram categorizadas em 3 para a quantidade de contas atrasadas, Figura 36, e em 4 para a média de dias para pagamento das contas, Figura 37.

Figura 36 - Variável referente a média de dias de pagamento após o *binning* e conversão para *dummy*.



Fonte: Elaborado pelo autor.

Figura 37 - Variável referente a média de dias de pagamento após o *binning* e conversão para *dummy*.

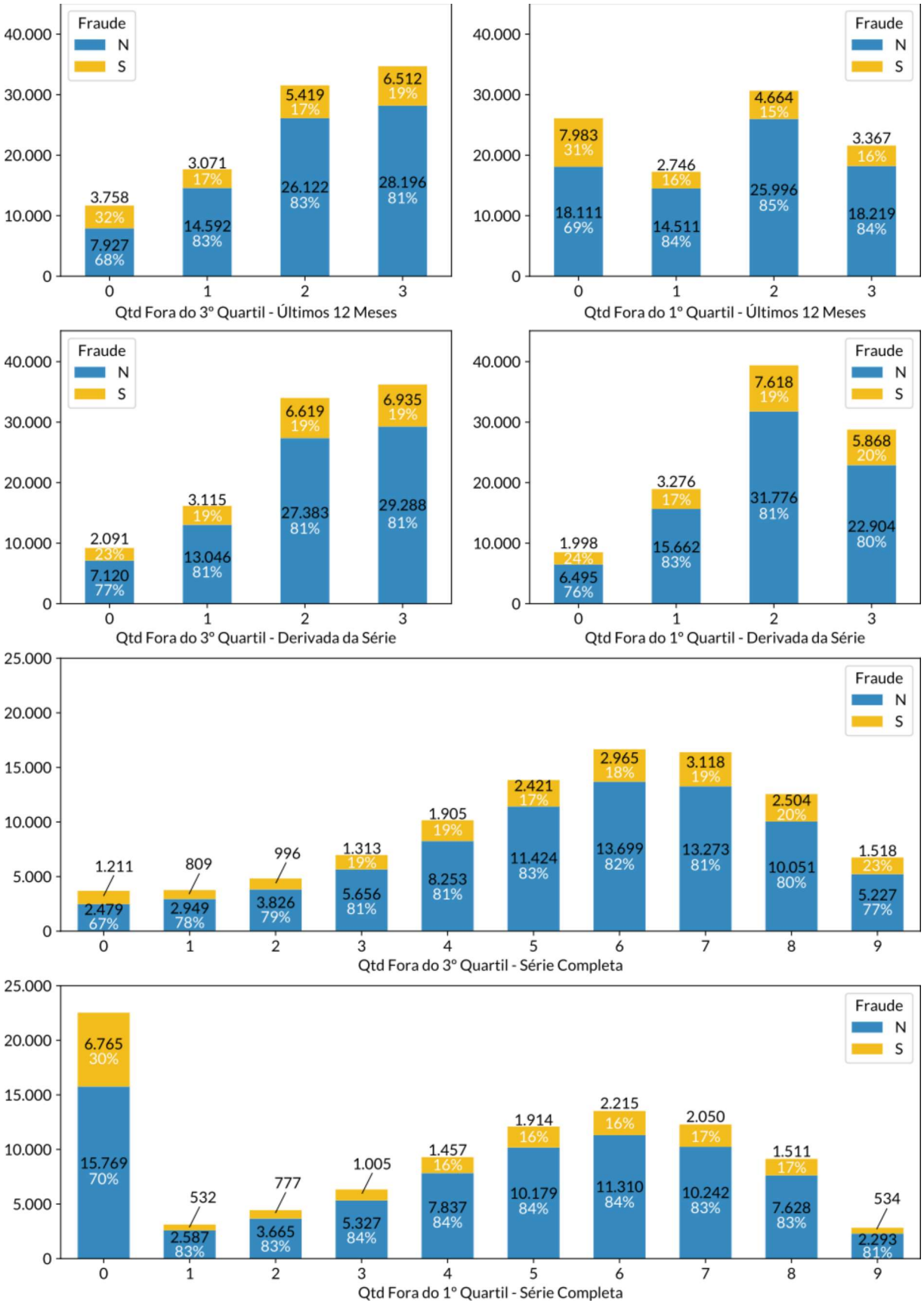


Fonte: Elaborado pelo autor.

Observa-se uma certa correlação entre inadimplência e fraude, já que as variáveis com alto índice de contas atrasadas e com elevada média de dias para pagamento refletem em um maior percentual de casos irregulares.

Finalmente, o último conjunto de variáveis categóricas são os seis referentes a quantidade de outliers na série e podem ser visualizadas na Figura 38. Como mencionado na Seção 3.3, esse conjunto foi categorizado em 4 classes, em que zero representa a ausência de pontos fora do 3º ou 1º quartil. Nota-se que algumas classes dessas variáveis apresentam um percentual de fraude maior do que o valor de referência aleatório de 20%. Isso indica variáveis que possivelmente serão promissoras para o modelo.

Figura 38 - Variáveis referentes as quantidades de outliers.



Fonte: Elaborado pelo autor.

Todos as variáveis consideradas que foram utilizadas no sistema após a análise, o agrupamento e transformação estão presentes no Quadro 13.

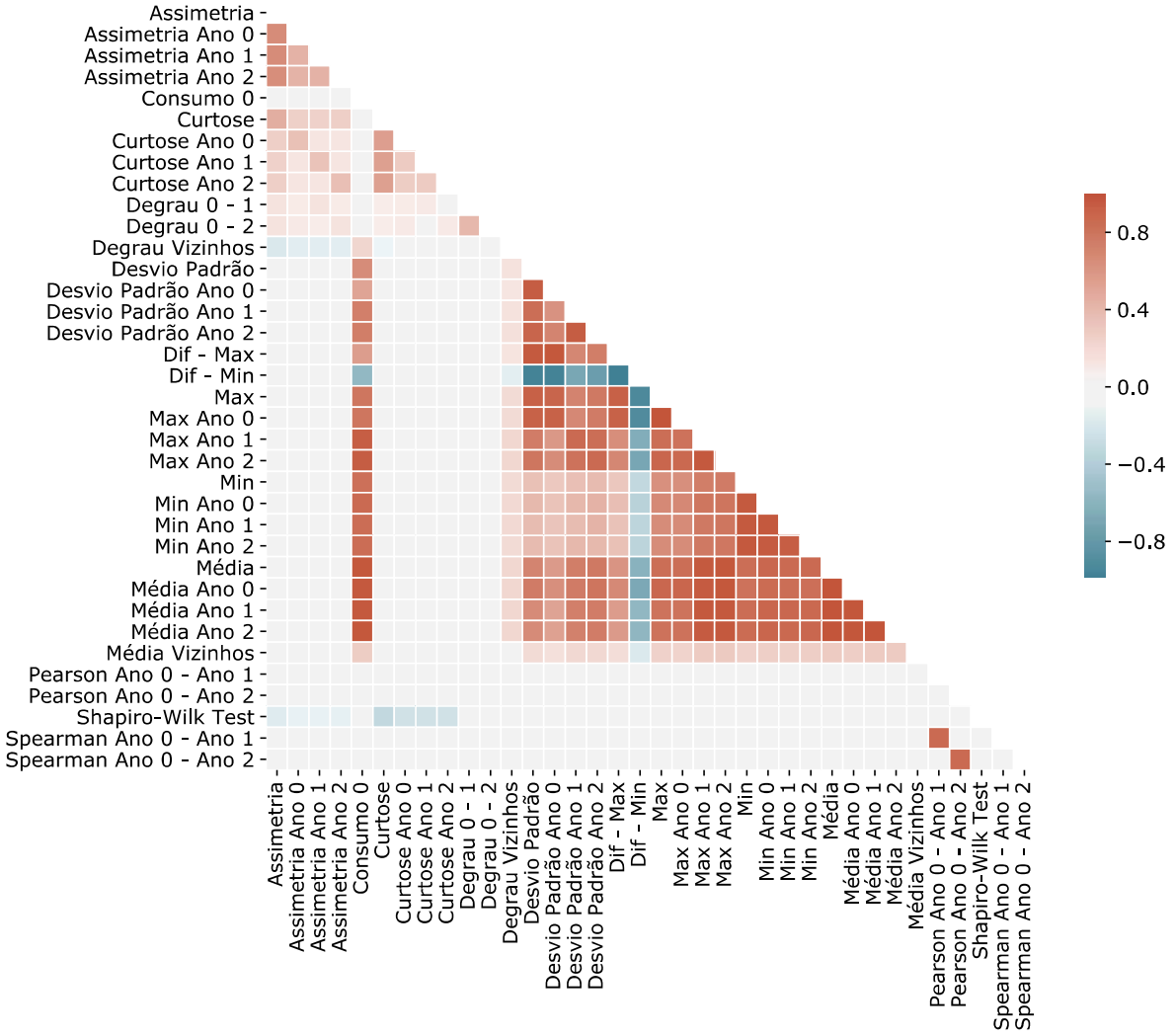
Quadro 13 - Variáveis categóricas consideradas no modelo.

Variável	Qtd	Descrição
Fraude (Saída)	1	Perfil da UC após inspeção.
Trifásico	1	Tipo de ligação trifásica na UC.
Conjunto	1	UC faz parte de um conjunto habitacional.
Desligado	3	Vezes que a UC permaneceu desligada em 36 meses.
Desocupado	3	Vezes que a UC permaneceu desocupada em 36 m.
Suspeita de Fraude	3	Vezes que a UC possui indicação de suspeita de fraude em 36 meses. Não incluído no modelo de residencial.
DLCB	1	UC possui DLCB ou blindagem de rede.
Externalização	1	UC possui medição externa.
Faturamento pela média	3	Vezes que a UC faturou pela média em 36 m.
Faturamento pelo mínimo	3	Vezes que a UC faturou pela mínimo em 36 m.
Inspeção Fraude	1	Vezes que a UC foi inspecionada com fraude.
Inspeção Normal	1	Vezes que a UC foi inspecionada e retornou normal.
Média Dias Pagamento	4	Média de dias que a UC leva para pagar uma conta.
Outlier Abaixo Q1 – Últimos 12 M	3	Quantidade de pontos abaixo do 1º quartil em 12 m.
Outlier Abaixo Q1 – Série Derivada	3	Quantidade de pontos abaixo do 1º quartil na série derivada.
Outlier Abaixo Q1 – Série Completa	3	Quantidade de pontos abaixo do 1º quartil em 36 m.
Outlier Acima Q3 – Últimos 12 M	3	Quantidade de pontos acima do 3º quartil em 12 m.
Outlier Acima Q3 – Série Derivada	3	Quantidade de pontos acima do 3º quartil na série derivada.
Outlier Acima Q3 – Série Completa	3	Quantidade de pontos acima do 3º quartil em 36 m.
Quantidade de Contas Atrasadas	3	Vezes que a UC atrasou uma conta.
Justificativa Degrau	1	Indica se a unidade esteve desocupada, desligada ou faturando pelo mínimo como justificativa para presença de um degrau.
Suspeita de Fraude após Degrau	1	Indica se a unidade possui uma indicação de leiturista de suspeita de fraude após uma redução de consumo. Não incluído no modelo de residencial.
Visita após Degrau	1	Indica se a unidade foi visitada para inspeção, regularização ou serviços semelhantes após uma redução de consumo.
Pequeno Comércio	1	Indica se a unidade é um pequeno comércio, identificada a partir da localização ou dados cadastrais da UC. Não incluído no modelo outros.
Seção do Tipo de Atividade	21	Agrupamento do CNAE do tipo de atividade. Incluído apenas no modelo outros.
MEI	1	UC optante pelo MEI. Incluído apenas no modelo outros.
SIMPLES	2	UC optante pelo SIMPLES. Incluído apenas no modelo outros.
Porte Empresa	4	Porte da empresa: micro, pequeno, outros, não informado. Incluído apenas no modelo outros.

4.1.2 Variáveis Contínuas

O primeiro passo para selecionar as variáveis contínuas para o modelo foi verificar a correlação entre elas através da utilização do coeficiente de Pearson conforme Figura 39, em que a cor vermelha indica correlação positiva e a cor azul indica correlação negativa. O Quadro 14 traz o resumo das variáveis que foram descartadas devido a correlação positiva ou negativa superior a 0,8. A variável Consumo 0, apesar de possuir alta correlação com as variáveis Mínimo Ano 0 e Média Ano 0, foi mantida por entender-se que, em alguns casos específicos, ela pode ser útil para a classificação da fraude. O mesmo foi feito com as variáveis Média Ano 0 (alta correlação com Máximo e Mínimo Ano 0) e Desvio Padrão Ano 0 (alta correlação com Máximo).

Figura 39 - Grau de correlação entre as variáveis contínuas medido através do coeficiente de Pearson.



Fonte: Elaborado pelo autor.

Quadro 14 – Variáveis excluídas devido à alta correlação com outras.

Variáveis Excluídas	Correlações
Spearman Ano - Ano 1, Spearman Ano 0 - Ano 2	Pearson Ano 0 - Ano 1, Pearson Ano 0 - Ano 2
Min, Min Ano 1, Min Ano 2, Média, Média Ano 1, Média Ano 2	Consumo – Último Mês, Min Ano 0, Média Ano 0
Dif – Max, Dif – Min	Desvio Padrão, Desvio Padrão Ano 0, Max, Max Ano 0
Desvio Padrão	Desvio Padrão Ano 0, Desvio Padrão Ano 1, Desvio Padrão Ano 2
Max Ano 2, Máx Ano 1	Desvio Padrão Ano 1, Desvio Padrão Ano 2, Max, Max Ano 0, Min Ano 0, Média Ano 0
Max Ano 0	Max, Desvio Padrão Ano 0
Desvio Padrão Ano 2	Desvio Padrão Ano 1

As variáveis consideradas no modelo estão presentes no Quadro 15.

Quadro 15 - Variáveis contínuas consideradas no modelo.

Variável	Descrição
Assimetria	Assimetria de toda a série de consumo.
Assimetria Ano 0	Assimetria da série correspondente ao último ano, mais recente a data de referência.
Assimetria Ano 1	Assimetria da série correspondente aos 12 meses intermediários.
Assimetria Ano 2	Assimetria da série correspondente aos 12 meses primeiros anos.
Consumo 0	Consumo do último mês, mais recente a data de referência.
Curtose	Curtose de toda a série de consumo.
Curtose Ano 0	Curtose da série correspondente ao último ano, mais recente a data de referência.
Curtose Ano 1	Curtose da série correspondente aos 12 meses intermediários.
Curtose Ano 2	Curtose da série correspondente aos 12 primeiros meses.
Degrau 0 - 1	Degrau calculado conforme equação (1) para atual = média de consumo do último ano e ant = média de consumo dos 12 meses intermediários.
Degrau 0 - 2	Degrau calculado conforme equação (1) para atual = média de consumo do último ano e ant = média de consumo dos 12 meses intermediários.
Degrau Vizinhos	Degrau calculado conforme equação (1) para atual = média de consumo do último ano e ant = média de consumo do último ano dos vizinhos mais próximos.
Degrau Vizinhos Atividade	Degrau calculado conforme equação (1) para atual = média de consumo do último ano e ant = média de consumo do último ano dos vizinhos mais próximos considerando a mesma seção e porte da empresa.
Degrau Chow	Degrau calculado conforme equação (1) para atual = média de consumo após índice de Chow e ant = média de consumo antes do índice de Chow.
Desvio Padrão Ano 0	Desvio padrão da série correspondente ao último ano, mais recente a data de referência.

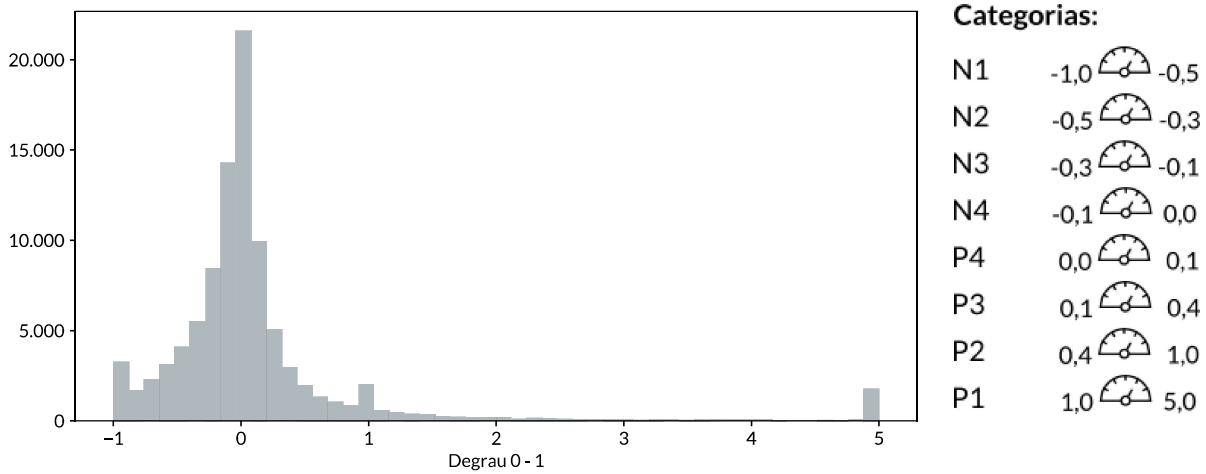
Continua

Conclusão

Variável	Descrição
Assimetria	Assimetria de toda a série de consumo.
Desvio Padrão Ano 1	Desvio padrão da série correspondente aos 12 meses intermediários.
Max	Máximo de toda a série de consumo.
Min Ano 0	Mínimo da série correspondente ao último ano, mais recente a data de referência.
Média Ano 0	Média da série correspondente ao último ano, mais recente a data de referência.
Média Vizinhos	Média de consumo do último ano dos vizinhos mais próximos.
Média Vizinhos Atividade	Média de consumo do último ano dos vizinhos mais próximos considerando a mesma seção e porte da empresa.
Pearson Ano 0 - Ano 1	Coeficiente de correlação de Pearson entre as séries de consumo do último ano e dos 12 meses intermediários.
Pearson Ano 0 - Ano 2	Coeficiente de correlação de Pearson entre as séries de consumo do último e primeiro ano.
Shapiro-Wilk Test	Teste de normalidade da distribuição da série de consumo.

Uma das variáveis mais utilizadas na metodologia para geração de campanha na empresa de estudo deste trabalho foi o degrau. Considerando a distribuição dessa variável a partir da Figura 40, e das regras utilizadas para gerar as listas de inspeção, os valores foram categorizados para facilitar a análise das ocorrências de fraude. A subdivisão foi feita a partir das 8 categorias da figura, determinada empiricamente, em que degraus negativos, ou seja, que houve redução de consumo, possuem índice N e degraus positivos, possuem índice P. Pelo formato da distribuição nota-se que a maior parte dos casos se encontra próximos a zero.

Figura 40 - Distribuição da variável degrau.

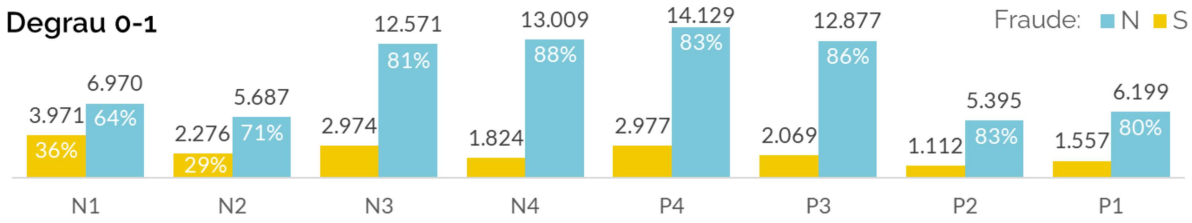


Fonte: Elaborado pelo autor.

A partir das categorias traçadas, monta-se a Figura 41. É possível observar que, nesta variável, apesar de serem obtidas efetividades superiores aos 20% da base completa para as categorias N1, N2 e N3, esse efeito também se repete para as categorias positivas, não

apresentando discrepância entre queda e evolução de consumo para servir como indício de fraude em uma UC. Além disso, como já observado pela distribuição do atributo, a cobertura é baixa, já que a maior parte das unidades possuem degraus entre -10% e 10%.

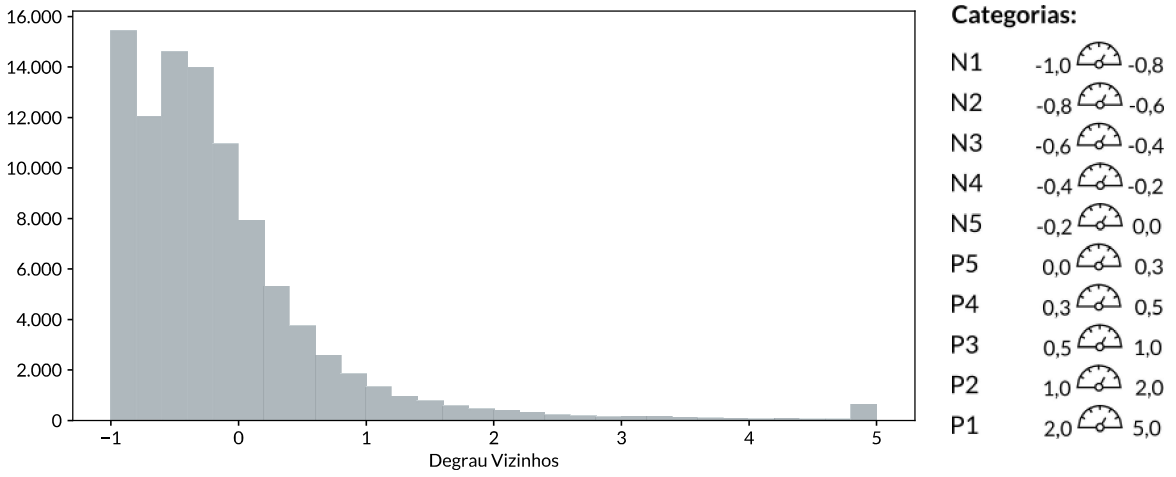
Figura 41 - Variável de degrau separado por classes para análise de ocorrências de fraude.



Fonte: Elaborado pelo autor.

A variável de degrau alternativa proposta foi o comparativo da média de consumo da unidade com UCs vizinhas mais próximas com a mesma classe de consumo e o mesmo tipo de ligação. Sua distribuição pode ser observada na Figura 42, bem como as subdivisões consideradas para análise da variável. Nota-se como o formato dessa distribuição diverge da vista anteriormente. Para a variável Degrau Vizinhos, existe uma disposição mais heterogênea dos casos encontrados, não havendo concentração da maior parte das unidades com o mesmo valor de degrau da forma como acontecia para a Figura 40.

Figura 42 - Distribuição da variável degrau vizinhos.

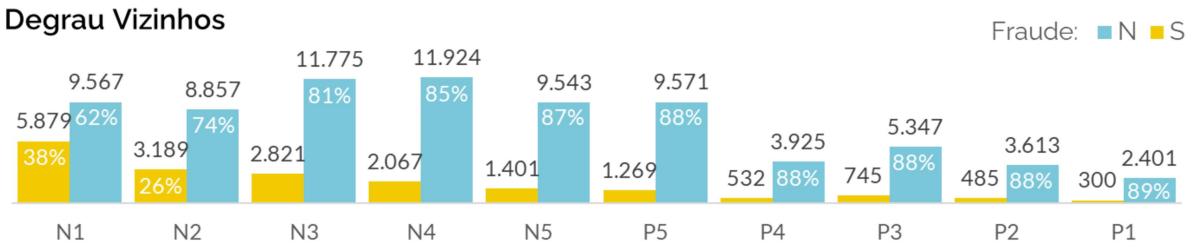


Fonte: Elaborado pelo autor.

Aplicando a segregação por meio das categorias delimitadas da Figura 42, é possível construir o gráfico da Figura 43. Diferente da variável de degrau utilizada para geração de campanhas, a métrica Degrau Vizinhos segue a lógica de que degraus positivos possuem, necessariamente, menos ocorrências de fraude que degraus negativos. Além disso, o percentual de acerto para as classes N1 a N3 são superiores ou equivalentes à do degrau usual com uma cobertura superior, atingindo uma maior quantidade de unidades. Essa variável, portanto,

apresenta-se como promissora para o modelo, podendo ser testada isoladamente em campo a fim de comparar sua performance em relação a outras regras de campanhas nos principais indicadores de uma campanha.

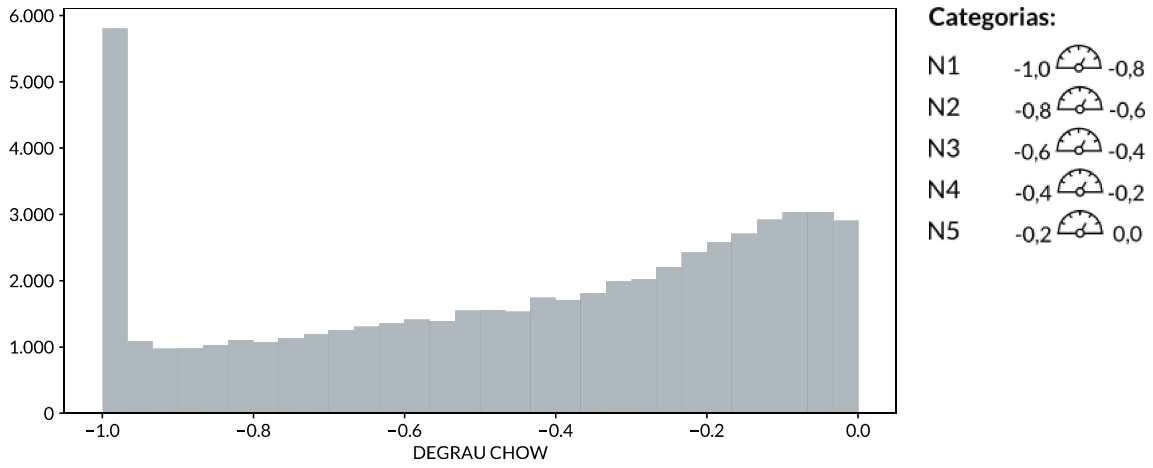
Figura 43 - Variável Degrau Vizinhos separado por classes para análise de ocorrências de fraude.



Fonte: Elaborado pelo autor.

A variável degrau Chow, por sua vez, possui a distribuição conforme a Figura 44, quando desconsiderados os valores zerados. Como não é possível identificar uma quebra de estrutura na curva de consumo de todas as unidades, existe um pico dessa variável em zero. Entretanto, nas as unidades que possuem essa quebra, ela pode ser bastante promissora.

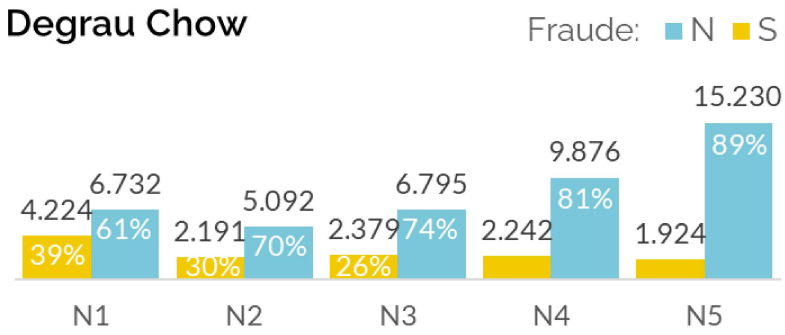
Figura 44 - Distribuição da variável degrau Chow.



Fonte: Elaborado pelo autor.

Aplicando a mesma segregação das categorias negativas da variável Degrau Vizinhos, é possível construir o gráfico da Figura 45. Apesar de restrita a menos unidades, o percentual de acerto das classes N1 a N3 da Degrau Chow são superiores ao próprio Degrau Vizinhos e também pode ser testada isoladamente em campo ou agregada às outras variáveis de degrau propostas devido a sua viabilidade.

Figura 45 - Variável Degrau Chow separado por classes para análise de ocorrências de fraude.

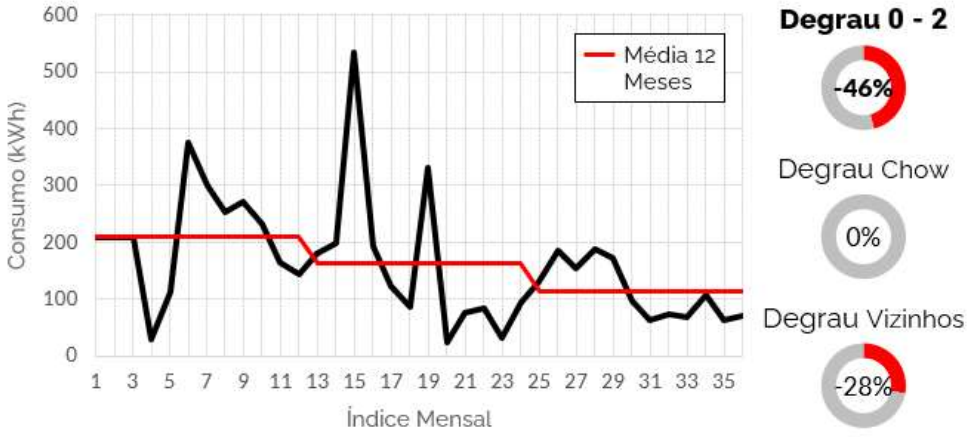


Fonte: Elaborado pelo autor.

Cada variável de indicação de degrau possui sua relevância. Existem casos de irregularidades na medição que podem não ser identificados através de uma, mas pode ser pelas outras, conforme ilustrado pelos 3 casos a seguir.

Na Figura 46, observa-se uma unidade consumidora do banco de dados que teve uma redução de consumo gradual ao longo dos meses. Por não possuir quebras estruturais na curva estatisticamente significantes, não há detecção de degrau por meio do teste de Chow. A variável degrau vizinhos, apesar de evidenciar um consumo 28% abaixo da média dos seus vizinhos, está classificado na categoria N4 da Figura 43, que possui uma frequência de fraude de apenas 15%, podendo ser julgada como não relevante pelo modelo. A variável Degrau 0 - 1 e, especialmente, a Degrau 0 - 2, no entanto, refletem como o consumo da unidade reduziu no decorrer dos anos.

Figura 46 – Exemplo de redução de consumo gradual em uma UC com desvio de energia. Em preto o consumo mensal, em vermelho a média de cada 12 meses de referência para as variáveis Degrau 0 – 1 e Degrau 0 – 2.

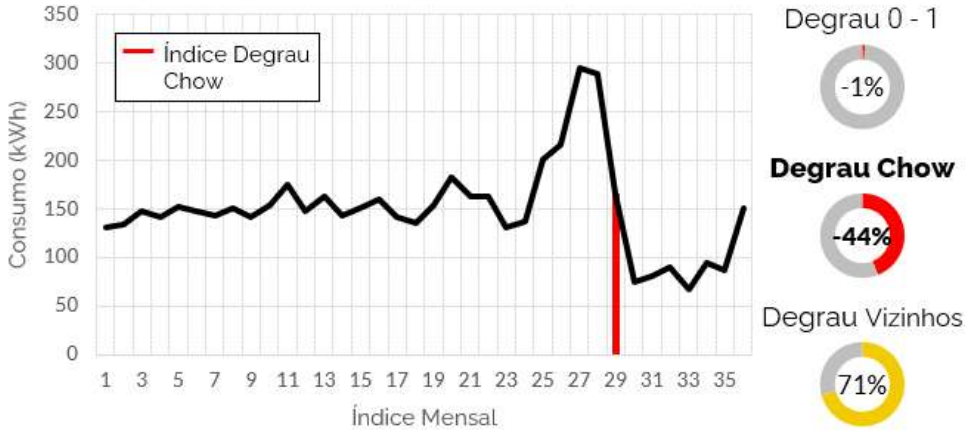


Fonte: Elaborado pelo autor.

Na Figura 47, observa-se uma unidade consumidora com uma redução de consumo súbita. O Degrau Chow identificou o seu início no mês de índice 29, como ilustrado em vermelho, entretanto, como houve um aumento no consumo logo antes do degrau ocorrer, a

variável Degrau 0 – 1 e Degrau 0 - 2 não identificaram essa redução. Além disso, a UC aparenta possuir um patamar de consumo superior aos seus vizinhos semelhantes, e, portanto, não foi identificada através da variável Degrau Vizinhos.

Figura 47 – Exemplo de redução súbita em UC com desvio de energia. Em preto o consumo mensal, em vermelho o início do degrau.



Fonte: Elaborado pelo autor.

No consumo da Figura 48 a seguir, diferente dos outros casos, não é possível observar reduções de consumo significativas. Nem o Degrau Chow, nem os Degraus 0 – 1 e 0 – 2 possuem qualquer tipo de indicação de irregularidade na UC, pelo contrário, existe uma indicação de crescimento da média de consumo. Entretanto, quando se compara o consumo da UC com sua variável Média Vizinhos, ou seja, a média de consumo das unidades semelhantes e próximas geograficamente, encontra-se um desvio de -65%.

Figura 48 – Exemplo de UC sem redução de consumo com desvio de energia. Em preto, o consumo mensal, em vermelho a média de consumo dos vizinhos semelhantes, em cinza o consumo após a regularização.

Fonte: Elaborado pelo autor.

4.2 Testes Teóricos

Visando comparar a metodologia que está sendo proposta com o que é utilizado atualmente na empresa de estudo, apresenta-se, na Tabela 1, os principais indicadores verificados para uma campanha, separados pela regra utilizada para as inspeções feitas entre 2017 e 2019.

Tabela 1 – Principais indicadores de campanhas de inspeção descriminado pela regra utilizada.

Regra	Inspeções	TOIs Aplicados	Efetividade	Recuperado (kWh)	Recuperado por TOI	Recuperado por Inspeção
Suspeita	14.537	5.268	36%	9.719.708	1.845	669
Diversos	17.118	4.786	28%	8.359.854	1.747	488
Avulso	116.506	14.374	12%	33.384.176	2.323	287
Degrau	8.326	889	11%	4.710.217	5.298	566
Varredura	1.525	107	7%	490.373	4.583	322

As regras de suspeita são aquelas que utilizam a indicação de leiturista de suspeita de fraude. Como pode ser verificado pela Tabela 1, essas campanhas são as que possuem maior percentual de efetividade, sendo aplicados 36% de TOIs por inspeção. Vale ressaltar que esse número não reflete, necessariamente, a quantidade de fraudes que são encontradas, visto que podem existir defeitos e outros tipos de perda comercial associados a esse percentual. A segunda regra com maior índice de acerto é a Diversos, que, em geral, une outros tipos de regras como o degrau e a própria suspeita. Com percentual menor de efetividade, uma campanha de avulso refere-se a unidades que foram visitadas pelos inspetores sem a utilização de uma regra baseado em dados estruturados. Sua efetividade é muito próxima a do próprio degrau, entretanto possui um recuperado por TOI ou por inspeção muito inferior a essa última regra. O degrau, apesar de possuir uma efetividade menor, possui o maior recuperado por TOI dentre todas as outras campanhas. Isso ocorre principalmente pela facilidade de determinar o momento em que a irregularidade se iniciou, uma vantagem da regra de degrau. Finalmente, as campanhas de varredura são voltadas para inspeções da maior parte das unidades de um local pré-determinado, com o objetivo de “varrer” a área.

Com a menor das efetividades, por ser uma regra que considera uma amostra mais aleatória dentre as citadas, o percentual de acerto da varredura será tomado como parâmetro para determinar a quantidade de casos de fraude a serem incluídas no banco de teste de unidades residenciais sem indicação de suspeita de fraude.

4.2.1 Comparativo das Principais Técnicas de Aprendizado de Máquina

Para comparar as técnicas de AM, utilizou-se a validação cruzada do tipo k-Fold para k = 5 como parte do treinamento e geração da base de teste através de *holdout* aleatório estratificado aplicado 10 vezes a própria base de treinamento. Em alguns algoritmos de aprendizado, o balanceamento do banco para treino é imprescindível, como é o caso da RNA. Para aqueles que não possuem tal característica, esse critério foi determinado de acordo com aquele que obtivesse a melhor performance do modelo. Para os testes, buscou-se utilizar uma proporção adequada para o que se espera encontrar em campo e, dessa forma, o banco de dados foi separado conforme a Tabela 2.

Como mencionado na Seção 3.5, quatro modelos foram montados:

a) Modelo 1: Residenciais sem suspeita de fraude;
b) Modelo 2: Unidades com suspeita de fraude;
c) Modelo 3: Rurais;
d) Modelo 4: Outras classes de consumo, que inclui comerciais, industriais, poderes

Tabela 2 - Quantidade de unidades consumidoras consideradas no banco de dados.

Banco de Dados		**Modelo 1**	**Modelo 2**	**Modelo 3**	**Modelo 4**
Treinamento	**Fraude**	14.647	9.459	1.467	1.446
	Situação Normal	47.663	9.219	4.216	7.409
Teste	**Fraude**	1.000	200	50	150
	Situação Normal	13.285	355	664	3.600
	Proporção Fraude	7%	36%	7%	4%

A avaliação das técnicas para o Modelo 1 está presente na Tabela 3 a seguir, em que são apresentadas as principais métricas de avaliação dos modelos. O intervalo de confiança foi calculado com base no Indicador Benefício proposto na seção 3.5.1.

Tabela 3 - Avaliação das técnicas aplicadas a unidades residenciais sem indicação de suspeita de fraude.

Técnica	**Efetividade**	**Cobertura**	**F-Score**	**Indicador Benefício**	**Intervalo de Confiança**
Árvore de Decisão	16,9%	47,6%	25,0%	16,3%	3,8%
Random Forest	40,4%	38,5%	39,4%	26,3%	3,7%
Gradient Boosting	**32,3%**	**45,4%%**	**37,7%**	**28,8%**	**3,9%**
K-Vizinhos Mais Próximos (KNN)	26,9%	23,8%	25,2%	21,8%	10,1%
Complement Naive Bayes	13,4%	57,2%	21,7%	19,7%	10,0%

Continua

Conclusão

Técnica	Efetividade	Cobertura	F-Score	Indicador Benefício	Intervalo de Confiança
Rede Neural MLP	16,7%	74,7%	27,2%	5,6%	15,9%
Support Vector Machine	23,3%	52,3%	32,2%	26,0%	14,9%

Para esse caso, houve um destaque para o *Gradient Boosting* e o *Random Forest*, ambos com valores de F-Score próximos entre si, em que o primeiro se destaca por possuir maior cobertura e o segundo, maior efetividade. Como mencionado, o Indicador Benefício do Modelo expressa, em termos numéricos, se uma maior cobertura compensa uma queda na efetividade, para f-scores parecidos. Nesse caso, observa-se que sim e o *Gradient Boosting* foi escolhido como técnica para a classificação das unidades residenciais sem indicação de leiturista de suspeita de fraude.

Para o UCs com essa indicação, o resultado da avaliação das técnicas encontra-se na Tabela 4 (Modelo 2). Assim como nas próprias campanhas da empresa, a efetividade, quando se utiliza essa variável, é superior as demais. A técnica do *Support Vector Machine* possuiu um melhor desempenho em termos de Indicador Benefício: mesmo com a menor das efetividades, a recuperação de consumo compensa ao se identificar 91% de todas as fraudes.

Tabela 4 - Avaliação das técnicas aplicadas a unidades com indicação de suspeita de fraude.

Técnica	Efetividade	Cobertura	F-Score	Indicador Benefício	Intervalo de Confiança
Árvore de Decisão	47,4%	61,5%	53,5%	54,1%	17,9%
Random Forest	52,0%	73,4%	60,8%	65,0%	17,1%
Gradient Boosting	55,1%	74,9%	63,5%	63,8%	11,1%
K-Vizinhos Mais Próximos (KNN)	47,6%	58,2%	52,3%	50,1%	12,6%
Complement Naive Bayes	53,6%	62,5%	57,7%	56,6%	6,2%
Rede Neural MLP	53,9%	72,6%	61,8%	61,8%	11,4%
Support Vector Machine	**44,4%**	**91,1%**	**59,7%**	**73,9%**	**10,4%**

Para as UCs rurais do Modelo 3, obteve-se o resultado da Tabela 5, com destaque para a técnica do *Random Forest* que, além de possuir o melhor indicador dentre as demais, obteve, também, uma efetividade e f-score superior. Observa-se que os modelos montados com a

técnica de Árvore de Decisão e Redes Neurais possuíram o Indicador Benefício negativa, o que representa que o custo associado as inspeções das unidades indicadas seriam superioras à energia recuperada pelas fraudes encontradas.

Tabela 5 - Avaliação das técnicas aplicadas a unidades da classe de consumo rural.

Técnica	Efetividade	Cobertura	F-Score	Indicador Benefício	Intervalo de Confiança
Árvore de Decisão	16,8%	51,8%	25,4%	-18,9%	74,4%
Random Forest	**37,7%**	**41,6%**	**39,5%**	**24,2%**	**47,9%**
Gradient Boosting	22,0%	44,6%	29,5%	10,8%	38,4%
K-Vizinhos Mais Próximos (KNN)	35,1%	19,4%	25,0%	16,5%	29,9%
Complement Naive Bayes	16,5%	62,2%	26,1%	23,8%	73,9%
Rede Neural MLP	17,5%	65,2%	27,1%	-4,3%	87,1%
Support Vector Machine	24,7%	48,6%	32,8%	17,4%	29,7%

Por fim, para as unidades da classe de consumo comercial, industrial, serviço público ou poder público, os resultados das avaliações das técnicas podem ser vistos na Tabela 6.

Tabela 6 - Avaliação das técnicas aplicadas a unidades da classe de consumo rural.

Técnica	Efetividade	Cobertura	F-Score	Indicador Benefício	Intervalo de Confiança
Árvore de Decisão	7,10%	28,10%	11,40%	21,10%	17,50%
Random Forest	25,60%	7,60%	11,60%	11,20%	18,30%
Gradient Boosting	13,90%	26,10%	18,10%	27,90%	21,50%
K-Vizinhos Mais Próximos (KNN)	19,10%	2,90%	5,10%	2,00%	5,50%
Complement Naive Bayes	8,30%	59,90%	14,50%	39,80%	19,40%
Rede Neural MLP	**7,60%**	**72,70%**	**13,80%**	**53,10%**	**20,10%**
Support Vector Machine	16,20%	26,70%	20,20%	18,20%	14,00%

A importância da utilização de um indicador que retrate o benefício líquido esperado é evidenciada para o Modelo 4. Observa-se que as técnicas apresentam uma diversidade de

resultados em termos de efetividade e cobertura muito maior que nos outros modelos. Os f-scores, por suas vezes, nos melhores casos, ficam em torno de 10% e 20%. Sem a presença do Indicador Benefício, compreender qual técnica deve ser utilizada pode ser desafiador ou, ainda, não trazer a melhor recuperação possível. Dessa forma, mesmo com a efetividade abaixo da média em relação as demais técnicas, a Rede Neural MLP foi capaz de recuperar 53,1% de toda a energia líquida disponível.

4.2.2 Teste Teórico dos Modelos

Considerando as técnicas escolhidas para cada um dos modelos com base no Indicador Benefício proposto, após treinamento com a base completa, aplicou-se o teste com o banco de dados das unidades consumidoras inspecionadas entre agosto de 2019 e janeiro de 2020 a fim de simular a efetividade em campo. Esse banco era composto por 14.003 unidades, em que 2.794 foram identificadas com procedimentos irregulares no conjunto de medição, em que 2.120 pertenciam a classe Residencial, 408 a classe Rural e 266 as demais classes.

Para comparação da metodologia proposta com a utilizada atualmente na empresa, foi aplicada, a essa base, o procedimento ilustrado a partir do fluxograma da Figura 8 para campanha de regras. A matriz de confusão obtida para a metodologia-empresa está apresenta na Tabela 7 a seguir.

Tabela 7 - Matriz de confusão do teste com a metodologia da empresa.

Matriz de Confusão		**Previsto**	
		N	S
Real	N	9.062	2.147
	S	1.738	1.056

Para a metodologia deste trabalho, considerando a união dos quatro modelos propostos, a matriz de confusão obtida pode ser observada na Tabela 8.

Tabela 8 - Matriz de confusão do teste com a metodologia proposta.

Matriz de Confusão		**Previsto**	
		N	S
Real	N	8.450	2.759
	S	1.001	1.793

O resumo das técnicas e parâmetros utilizados encontra-se no Quadro 16. Tais configurações foram determinadas de maneira empírica através dos testes com *holdout* aleatório da seção 4.2.1.

Quadro 16 - Principais parâmetros das técnicas utilizadas.

Modelo	Entradas	Técnica	Parâmetros
Residencial (1)	69	*Gradient Boosting*	Estimadores: 140; taxa de aprendizado: 1; profundidade máxima: 3
Ind. Suspeita de Fraude (2)	74	SVM	Gamma: scale; kernel: rbf; peso por classe: N - 1, S - 2
Rural (3)	72	*Random Forest*	Estimadores: 150; critério: Gini
Outros (4)	97	RNA	Algoritmo: Adam; função de ativação: tanh; alfa: 10^{-5}; arquitetura: (110, 130) neurônios

A Tabela 9 a seguir resume os principais indicadores dos quatro modelos utilizados e a união deles, que representa a proposta deste trabalho, denominada de Modelo Final. Foi incluído também o resultado da metodologia da empresa a título de comparação entre os métodos.

Tabela 9 – Principais indicados de avaliação dos modelos para o teste teórico.

Modelo	Efetividade	Cobertura	F-Score	Indicador Benefício
Modelo 1	52,3%	46,2%	49,0%	44,4%
Modelo 2	48,0%	86,9%	61,8%	73,1%
Modelo 3	64,7%	32,8%	43,5%	27,4%
Modelo 4	19,3%	79,1%	31,1%	74,1%
Modelo Final	**39,4%**	**64,2%**	**48,8%**	**59,5%**
Metodologia Empresa	33,0%	37,8%	35,2%	23,5%

Observa-se que foi possível obter efetividades superiores a metodologia da empresa por 19,5%, com uma cobertura 69,8% maior. Isso indica uma maior assertividade atingindo uma maior quantidade de unidades consumidoras.

Em termos de benefício líquido para a empresa, o Indicador Benefício do Modelo aponta uma melhoria em 153,2% da possibilidade de recuperação. Ao mesmo tempo, essa métrica indica que seria possível recuperar quase 60% de toda a energia disponível para recuperação.

Os resultados de uma campanha gerada com base no conjunto de dados inspecionados entre agosto de 2019 e janeiro de 2020 pela indicação da metodologia proposta, podem ser vistos na Tabela 10. Comparando com o resultado de campanhas de regras (suspeita, diversos e degrau) da Tabela 1, foi possível obter uma efetividade 43,9% maior. O recuperado por TOI também foi superior – em 38,8% – se considerados todas as 3 regras; entretanto a regra de

degrau continua sendo a maior recuperação por TOI aplicado, principalmente devido a facilidade em encontrar o início da irregularidade, não se limitando a 6 meses e sim a 36, com possibilidade de recuperar toda a energia não fatura. O recuperado por inspeção da metodologia proposta, por sua vez, foi 99,6% maior, chegando a superar até mesmo a regra de suspeita de fraude que possui 669 kWh de recuperado por inspeção, a maior da empresa. Esse último resultado evidencia o foco do modelo em recuperação de energia com menor custo.

Tabela 10 – Indicadores simulados para uma campanha com a metodologia proposta.

Indicador	**Campanha**
Quantidade de Inspeções	4.552
Efetividade	39,4%
Energia Recuperada	5.181.601 kWh
Recuperado por TOI	2.890 kWh
Recuperado por Inspeção	1.138 kWh
Recuperação Financeira	R$ 2.476.961
Custo	R$ 377.702

5 CONCLUSÕES

Nesse trabalho, foram abordados temas relacionados as perdas não técnicas de energia em que seu combate surge como desafio para as distribuidoras de energia devido à dificuldade na detecção de fraudes. A revisão bibliográfica demonstrou que, apesar de terem sido obtidos avanços no âmbito da utilização de técnicas de inteligência artificial para detecção de unidades consumidoras com irregularidades de medição, os resultados ainda se apresentam pouco aprofundados, sem padrões de avalição das técnicas e com pouco foco para cenários reais.

A primeira contribuição deste trabalho foi referente a construção de variáveis para detecção de fraude. Foram propostos atributos diversos com base nos dados da empresa de estudo que buscaram avaliar o comportamento do cliente de diversos ângulos. Como o volume de informação é muito elevado, a utilização de aprendizado de máquina contribui para a melhoria dos processos de combate à perda de energia, automatizando e encontrando padrões antes não observados pelo analista. A principal variável proposta buscou comparar a média de consumo da unidade com os vizinhos geográficos mais próximos que possuíam características de porte semelhantes. Através de análises exploratórias, foi verificada que a nova variável é capaz de englobar uma maior quantidade de clientes e possui melhor taxa de precisão quanto mais negativo for seu valor.

As demais contribuições envolveram diretamente as técnicas de aprendizado de máquina que tiveram suas performances avaliadas de maneira estruturada e de acordo com as principais métricas sugeridas pela literatura. Também foi proposto um novo indicador para avaliação de técnicas e modelos para avaliação do percentual do benefício de energia líquida recuperada em relação ao montante total disponível. Quatro modelos foram construídos a partir da separação da base em três classes de consumo mais todas as unidades previamente indicadas por leiturista com suspeita de fraude. Observou-se um destaque para as técnicas *Gradient Boosting, Support Vector Machine, Random Forest* e Rede Neural Artificial que foram selecionados para testes teóricos por possuírem os maiores valores do Indicador Benefício proposto em testes com *holdout* aleatório. No teste teórico, com o modelo formado pela união dos quatro, obteve-se uma efetividade de 39,4% e cobertura de 64,2%. Essa precisão ultrapassa 19,5% os obtidos utilizando a metodologia da empresa considerando a mesma base de referência e 44% os obtidos através de campanhas com regras que analisam consumo e

indicações de suspeita de fraude por leituristas. O modelo proposto também obteve uma cobertura 69,8% e 153,2% de benefício líquido superior quando utilizado o indicador proposto.

O uso de técnicas de inteligência artificial tem se mostrado promissor para auxiliar a distribuidora de energia na detecção de fraude, visto que esta tem obtido efetividades próximas a 11% quando não existe uma indicação clara de irregularidade na unidade consumidora. O volume de dados é o principal ofensor desse número; como existem muitas fontes e muitas variáveis diferentes, torna-se necessário a utilização de uma metodologia estruturada para o combate as perdas.

Como trabalhos futuros, as variáveis e os modelos serão aplicados na base da empresa e inspeções serão feitas em campo para verificar a real viabilidade do trabalho. Nessa etapa serão avaliados os ganhos em energia e no âmbito financeiro. Serão avaliados também os impactos da pandemia frente aos resultados obtidos até então. Outros passos envolvem aplicar técnicas avançadas de seleção de variáveis e realizar estudos específicos para resolver as limitações de se utilizar um banco de dados desbalanceado a fim de verificar as possibilidades de ganhos para os modelos propostos.

REFERÊNCIAS

ABRADEE. **A Distribuição de Energia**. In: ABRADEE: Segmento de Distribuição. Brasília, [2018a]. Disponível em: https://www.abradee.org.br/setor-de-distribuicao/a-distribuicao-de-energia/. Acesso em: 24 jan. 2020.

ABRADEE. **Tarifas de Energia**. In: ABRADEE: Segmento de Distribuição. Brasília, [2018b]. Disponível em: https://www.abradee.org.br/setor-de-distribuicao/tarifas-de-energia/. Acesso em: 24 jan. 2020.

ABRADEE. **Furto e Fraude de Energia**. In: ABRADEE: Segmento de Distribuição. Brasília, [2018c]. Disponível em: http://www.abradee.org.br/setor-de-distribuicao/furto-e-fraude-de-energia. Acesso em: 24 jan. 2020.

ANEEL. Agência Reguladora. **Resolução normativa nº 414, de 9 de setembro de 2010**. Estabelece as Condições Gerais de Fornecimento de Energia Elétrica de forma atualizada e consolidada. Brasília: Agência Reguladora, 2010.

ANEEL. Agência Reguladora. **Resolução normativa nº 482, de 17 de abril de 2012**. Estabelece as condições gerais para o acesso de microgeração e minigeração distribuída aos sistemas de distribuição de energia elétrica, o sistema de compensação de energia elétrica, e dá outras providências. Brasília: Agência Reguladora, 2012.

ANEEL. **Perdas de Energia**. In: ANEEL: Metodologia de Cálculo Tarifário da Distribuição. Brasília, 25 nov. 2015. Disponível em:http://www.aneel.gov.br/metodologia-distribuicao/-/asset_publisher/e2INtBH4EC4e/content/perdas/654800. Acesso em: 24 jan. 2020.

ANEEL. Agência Reguladora. **Resolução Normativa nº 842, de 26 de dezembro de 2018**. Procedimentos de Distribuição de Energia Elétrica no Sistema Elétrico Nacional (PRODIST), Módulo 1. Brasília: Agência Reguladora, 2018.

ANEEL. **Relatórios de Consumo e Receita de Distribuição: Consumidores, Consumo, Receita e Tarifa Média – Região**. In: ANEEL: Tarifas consumidores. Brasília, [2019]. Disponível em: http://www.aneel.gov.br/relatorios-de-consumo-e-receita. Acesso em: 24 jan. 2020.

ANGELOS, E. W. S.; SAAVEDRA, O. R.; CORTÉS, O. A. C.; SOUZA, A. N. Detection and Identification of Abnormalities in Customer Consumptions in Power Distribution Systems. **IEEE Transactions on Power Delivery**, Edmonton, v. 26, n. 4, p. 2436-2442, 2011.

ARAUJO, B. S. d.; ALMEIDA, H. L. S. d.; MELLO, F. L. d. Computational Intelligence Methods Applied to the Fraud Detection of Electric Energy Consumers. **IEEE Latin America Transactions**, [S.I.], v. 17, n. 01, p. 71-77, 2019.

BERRY, D. A.; CHALONER, K. M.; GEWEKE, J. K. **Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner**. New York: John Wiley & Sons, Inc., 1995.

BISHOP, C. M. **Neural Networks for Pattern Recognition**. Oxford: Clarendon Press, 1995.

BREIMAN, L. Random Forests. **Machine Learning**, [S.I.], v. 45, n. 1, p. 5 – 32, 2001.

BRITO, N. B. Experiências e Ações no Combate a Perdas Comerciais. In: XV Seminário Nacional de Distribuição de Energia Elétrica, 2002, [S.I.]. **Anais** [...]. [S.I.]: ABRADEE e EDP, 2002.

BOEING, G. OSMnx: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks. **Computers, Environment and Urban Systems**, [S.I.], v. 65, p. 126-139, 2017.

CURADO, M. I. C. **Localização de Perdas Não Técnicas de Energia em Sistemas de Distribuição Utilizando o Método PQ**. 2015. Monografia (Graduação em Engenharia Elétrica) – Universidade de São Paulo, São Carlos, 2015.
2015

DEVORE, J. L. **Probabilidade e estatística para engenharia e ciências**. 9. ed. São Paulo: Cengage, 2018.

FACELI, K.; LORENA, A. C.; GAMA, J.; CARVALHO, A. C. P. L. F. **Inteligência Artificial**: Uma Abordagem de Aprendizagem de Máquina. Rio de Janeiro: LTC, 2011.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. London: The MIT Press, 2016.

GUERRERO, J. I.; LEÓN, C.; MONEDERO, I.; BISCARRI, F.; BISCARRI, J. Improving Knowledge-Based Systems with statistical techniques, text mining, and neural networks for non-technical loss detection. **Knowledge-Based Systems**, [S.I.], v. 71, p. 376-338, 2014.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**: Data Mining, Inference, and Prediction. 2 ed. Stanford: Springer, 2008.

IBGE. Concla: Comissão Nacional de Classificação. **CNAE_Subclasses_2_3_Estrutura_Detalhada.xlsx**: Classificação de Atividade Econômicas por Tema, CNAE-Subclasses 2.3. [S.I.], 2020. Disponível em: https://concla.ibge.gov.br/classificacoes/por-tema/atividades-economicas. Acesso em: 12 jul. 2020.

INSTITUTO ACENDE BRASIL. Perdas Comerciais e Inadimplência no Setor Elétrico. **White Paper**, São Paulo, n. 18, 2017.

KUBAT, M.; MATWIN, S. Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In: Fourteenth International Conference on Machine Learning, 1997, San Francisco. **Proceedings** […]: Morgan Kaufmann Publishers, San Francisco, 1997.

MASSAFERRO, P; MARTINO, J, M; FERNÁNDEZ, A. Fraud Detection in Electric Power Distribution: An Approach That Maximizes the Economic Return. **IEEE Transactions on Power Systems**, [S.I.], v. 35, n. 1, p. 703-710, 2020.

NAGI, J.; YAP, K. S.; TIONG, S. K.; AHMED, S. K.; MOHAMAD, M. Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines. **IEEE Transactions on Power Delivery**, Edmonton, v. 25, n. 2, p. 1162-1171, 2010.

ORTEGA, G. V. C. **Redes Neurais na Identificação de Perdas Comerciais do Setor Elétrico**. 2008. Dissertação (Mestrado em Engenharia Elétrica) - Pontifícia Universidade Católica Do Rio De Janeiro - PUC-Rio, Rio de Janeiro, 2008.

PYLE, D. **Data Preparation for Data Mining**. San Francisco: Morgan Kaufmann Publishers, 1999.

QUEIROZ JR, J.; BRITO, N. B.; VALÉRIO, D. P.; SAMPAIO, S.; PARENTE, D.; FERNANDES, G. Caixa Padrão Rede-CPREDE: Projeto Medições as Claras. In: XIV Seminário Nacional de Distribuição de Energia Elétrica, 2000, Foz do Iguaçu. **Anais** [...]. Foz do Iguaçu: ABRADEE e EDP, 2000.

RAMACHANDRAN, P.; ZOPH, B., LE, Q. V. **Searching for Activation Functions**. v. 2. [S.I.]: arXiv, 2017. Disponível em: https://arxiv.org/abs/1710.05941. Acesso em: 30 jan. 2020.

RAMOS, C. C. O.; RODRIGUES, D.; SOUZA, A. N.; PAPA, J. P. On the Study of Commercial Losses in Brazil: A Binary Black Hole Algorithm for Theft Characterization. **IEEE Transactions on Smart Grid**, [S.I.], v. 9, n. 2, p. 676-683, 2018.

RIJSBERGEN, C. J. **Information Retrieval**. 2 ed. London: Butterworths, 1979.

VIEGAS, J. L.; ESTEVES, P. R.; MELÍCIO, R.; MENDES, V.M.F.; VIEIRA, S. M. Solutions for detection of non-technical losses in the electricity grid: A review. **Renewable and Sustainable Energy Reviews**, [S.I.], v. 80, p. 1256-1268, 2017.

ZHENG, Z.; YANG, Y.; NIU, X.; DAI, H.; ZHOU, Y. Wide and Deep Convolutional Neural Networks for Electricity-Theft Detection to Secure Smart Grids. **IEEE Transactions on Industrial Informatics**, Taipei, v. 14, n. 4, p. 1606-1615, 2018.