DOI: 10.48011/asba.v2i1.1560

# Detecção de Fraude em Unidades Consumidoras Não Telemedidas com Uso de Técnicas de Aprendizado de Máquina

**Fernanda R. Paulo*. Juan M. M. Villanueva****
**Helon D. M. Braz*****

**Pós-Graduação Em Engenharia Elétrica, Universidade Federal da Paraíba, João Pessoa, PB ,Brasil (e-mail: fernanda.paulo@cear.ufpb.br)*
*** Departamento de Engenharia Elétrica, Universidade Federal da Paraíba, João Pessoa, PB ,Brasil (e-mail: jmauricio@cear.ufpb.br)*
**** Departamento de Engenharia Elétrica, Universidade Federal da Paraíba, João Pessoa, PB ,Brasil (e-mail: helon@cear.ufpb.br).*

**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, variables were selected and then six techniques were evaluated. The main variable proposed compares the average consumption of the unit with the closest geographic neighbors with similar size characteristics. In theoretical tests, it was possible to obtain an accuracy of 35% with the Random Forest model, which exceeds the company's indicators by 218% for inspection lists based on rules of consumption reduction.

**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 ao volume de dados e a limitação de encontrar padrões. Considerando esse cenário, propõe-se o desenvolvimento de uma metodologia 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 variáveis e, então, avaliadas seis técnicas. A principal variável proposta compara a média de consumo da unidade com os vizinhos geográficos mais próximos com características elétricas semelhantes. Em testes teóricos, foi possível obter uma efetividade de 35% com o modelo do *Random Forest*, o que ultrapassa os indicadores da empresa em 218% para listas de inspeções com base em regras de redução de consumo.

*Keywords*: Commercial losses; non-technical losses; energy fraud; pattern classification; machine learning; artificial intelligence; fraud detection; imbalanced classification.

*Palavras-chaves*: Perdas comerciais; perdas não técnicas; fraude de energia; classificação de padrões; aprendizado de máquina; inteligência artificial; detecção de fraude; dados desbalanceados.

## 1. INTRODUÇÃO

As distribuidoras de energia elétrica devem garantir o fornecimento efetivo de energia aos consumidores conectados ao seu sistema de distribuição. Em 2018, o consumo de energia das distribuidoras do Brasil foi de aproximadamente 310 TWh (ANEEL 2019). Parte desse montante transformado foi perdido devido às perdas técnicas, referente a dissipação de energia inerente ao sistema em virtude das leis físicas, e perdas não técnicas (ou perdas comerciais), referente as demais perdas e decorrente, principalmente, da fraude, do furto e de erros de medição.

A perda de receita devido a procedimentos irregulares é um dos principais focos das distribuidoras devido aos prejuízos que ela acarreta. Em 2018, a perda não técnica foi de 6,6% no Brasil, representando uma receita anual de mais de R$ 9,8 bilhões. Para caracterizar uma fraude ou um furto, é necessário enviar uma equipe especializada no local para que seja feita uma inspeção e cada deslocamento e fiscalização gera custos a empresa. Dessa forma, essas companhias buscam aprimorar as técnicas de identificação de fraude a fim de otimizar as inspeções, maximizando a energia recuperada por visita realizada. O número total de consumidores de uma única concessionária pode chegar até 8 milhões em algumas regiões do país.

A Agência Nacional de Energia Elétrica (ANEEL), órgão fiscalizador, incentiva as distribuidoras a serem mais eficientes a partir da composição tarifária. Enquanto toda a perda técnica é repassada na tarifa, a perda não técnica possui um repasse limitado de acordo com valores estabelecidos pelo órgão com base na região e empresas de mesmo porte, motivando as concessionárias a se empenharem no combate às perdas comerciais. Além da redução do valor da tarifa, combater essas perdas traz benefícios que incluem a melhora da qualidade do fornecimento e reduz o consumo inconsciente (ANEEL 2015).

Os trabalhos científicos na área de detecção de perdas não técnicas são, em sua maior parte, soluções de software que utilizam métodos de aprendizado de máquina (Viegas et al., 2013). Isso se dá 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. 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 por analistas. Entretanto, não há metodologia generalizada, já que os padrões podem diferir para cada estudo de caso, bem como as informações disponíveis para aplicação nos modelos.

Grande parte dos trabalhos utiliza dados de consumo de energia, do perfil do consumidor, da carga, tensões e correntes medidas e resultados de inspeções (Viegas et al., 2013). Além disso, a maior parte das soluções depende 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. Vale ressaltar que, apesar de haver um notável crescimento na utilização de *smart grids* e medidores inteligentes pelas distribuidoras no mundo, nos países em desenvolvimento e subdesenvolvidos, essa implementação ainda se encontra em fase inicial, com foco em consumidores de grande porte ou pequenos municípios devido a seus problemas socioeconômicos inerentes (Ponce-Jara et al., 2017). Dessa forma, se uma infraestrutura já não estiver implementada na empresa, soluções que sejam dependentes desses equipamentos terão altos custos de implantação associados.

Considerando o cenário atual de combate às perdas de energia na concessionária em estudo, este trabalho visa desenvolver uma metodologia baseada 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. A base de dados foi montada a partir das inspeções realizadas pela concessionária entre julho de 2017 e julho de 2019.

As principais contribuições envolvem aumentar o percentual de efetividade nas listas geradas para inspeção, construir e identificar as principais variáveis na detecção de fraude a partir de uma metodologia automatizada, implementar novas métricas que auxiliem na identificação da irregularidade sem presumir que há variação de consumo e avaliar diferentes modelos aplicados no combate às perdas. Todos os resultados serão avaliados através das métricas mais utilizadas na área de aprendizado de máquina e identificação de padrões, como efetividade e cobertura, derivadas da matriz de confusão, curva ROC e F-score. Uma alta efetividade ou precisão está relacionada a uma baixa taxa de falsos positivos, ou seja, a uma maior assertividade nas fraudes indicadas pelo modelo. Uma alta cobertura ou sensibilidade está relacionada a uma baixa taxa de falsos negativos, ou seja, a uma alta identificação das fraudes existentes pelo modelo.

O trabalho está dividido conforme descrito a seguir: na seção 2 está descrito o estado da arte, com os principais trabalhos relacionados a esta pesquisa; a seção 3 fornece informações referentes ao estudo de caso; a seção 4 detalha a metodologia utilizada, que envolve a preparação e modelagem dos dados, bem como a análise exploratória e a aplicação dos modelos; a seção 5 apresenta e discute os resultados obtidos; a seção 6 trata das conclusões e trabalhos futuros.

## 2. TRABALHOS RELACIONADOS

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. 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 *Support Vector Machine* (SVM) juntamente com um algoritmo para otimizar seus parâmetros. A acurácia, ou seja, o índice total de acerto entre o indicado pelo modelo 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%. Em campo, obteve-se uma efetividade de 26% divididos em 7% de anormalidade (casas abandonadas, mudança de titularidade e defeitos na fiação do medidor) e 19% de atividades fraudulentas. Não foi citada a quantidade de unidades enviadas para inspeções, o que limita o conhecimento sobre o alcance da metodologia proposta.

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. Os autores utilizaram técnicas de processamento de linguagem natural para extrair informações não estruturadas dos comentários dos inspetores e estruturá-las em categorias. 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 Guerrero et al. (2014) evidenciam 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 total é muito baixa. Em Angelos et al. (2011), os autores analisaram 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 C-Means baseado em lógica Fuzzy para encontrar consumidores com perfis de consumo semelhante. Quando utilizado percentuais 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 se utilizada uma base balanceada para verificar a eficácia do método.

Os trabalhos mencionados possuem limitações que foram evidenciadas por Viegas et al. (2017) em sua revisão do estado da arte. As metodologias utilizadas assumem 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.

Atualmente, o tema de combate às perdas não técnicas continua sendo intensamente pesquisado. Ramos et al. (2018) propôs 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. Zheng, et al. (2018) propôs uma rede neural convolucional com dois componentes para identificar roubo de energia através da curva de consumo semanal em duas dimensões. Já Araujo et al. (2019) utilizou uma Rede Neural Artificial para determinar a probabilidade de existir uma irregularidade em uma unidade consumidora (UC) 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, a alteração de consumo continua sendo a variável mais frequente para detecção de perda comercial, assim como a falta de padrão para avaliar os modelos, o que dificulta a verificação da real eficácia dos métodos.

## 3. ESTUDO DE CASO

A distribuidora de energia utilizada neste estudo é composta em sua maior parte por consumidores atendidos em baixa tensão, ou seja, com tensão de fornecimento inferior a 2,3 kV, representando 79% do consumo total da área de concessão. Esses consumidores possuem cargas instaladas de até 75 kW e são, em sua maior parte, clientes residenciais, que constituem 82% das unidades e 46% do consumo total.

Os principais indicadores das inspeções realizadas entre 2017 e 2019 estão apresentados na Tabela 1. A efetividade refere-se a taxa de acerto das inspeções realizadas, ou seja, a quantidade de irregularidade dividido pela quantidade de inspeções. O recuperado é a energia que foi ressarcida após a cobrança ao cliente. Outros indicadores também são acompanhados pela empresa, como o recuperado por irregularidade encontrada ou o recuperado pela quantidade de inspeções realizadas.

**Tabela 1. Principais indicadores das inspeções realizadas entre 2017 e 2019 descriminados pela regra utilizada**

| Regra | Inspeções | Irregulares | Efetividade | Recuperado (MWh) |
|---|---|---|---|---|
| Suspeita | 14.537 | 5.268 | 36% | 9.720 |
| Diversos | 17.118 | 4.786 | 28% | 8.360 |
| Avulso | 116.506 | 14.374 | 12% | 33.384 |
| Degrau | 8.326 | 889 | 11% | 4.710 |
| Varredura | 1.525 | 107 | 7% | 490 |

Na Tabela 1, os resultados foram separados pela regra utilizada:

1) Suspeita: inspeções realizadas com base em indicações de leituristas de suspeita de fraude. São as que possuem maior percentual de efetividade, tendo sido encontradas 36% de irregularidades nas unidades visitadas.

2) Diversos: em geral, une outros tipos de regras como a redução de consumo, denominado aqui de degrau, e a própria suspeita.

3) Avulso: refere-se a unidades que foram visitadas pelos inspetores sem a utilização de uma regra baseada em dados estruturados. Sua efetividade é muito próxima a do próprio degrau, entretanto é a que menos recupera energia para cada inspeção dentre as demais regras.

4) Degrau: baseado em regras de redução de consumo. Apesar de possuir uma efetividade menor que as regras anteriores, possui a maior recuperação por irregularidade encontrada dentre todas as outras.

5) Varredura: são campanhas 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.

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 classificação de dados desbalanceados. Esse banco foi montado a partir dos dados dos sistemas comerciais da empresa de estudo. As informações utilizadas neste trabalho foram classificadas em:

1) Cadastro: contém dados relacionados ao consumidor responsável pela unidade, à localização geográfica e ao imóvel.

2) Inspeções: referem-se as comprovações históricas feitas em campo da existência ou não de uma irregularidade na medição do consumo de energia.

3) Consumo: refere-se ao consumo lido mensal para cada unidade.

4) Irregularidade de leitura: apontamentos da coleta de leitura mensal. Exemplos: imóvel desocupado, suspeita de fraude.

5) Irregularidade de faturamento: indicam quando o consumo lido pode ter sido divergente do faturado. Exemplo: faturado pela média, faturado pelo mínimo da fase, unidade desligada.

6) Serviços: intervenções feitas na unidade por equipes da empresa. Elas foram utilizadas neste trabalho para extrair características ou para observar mudanças de comportamento na unidade.

7) Pagamentos: informações acerca das datas de pagamento e vencimentos das contas de energia dos clientes.

O banco de dados foi montado com base nos clientes inspecionados entre julho de 2017 e julho de 2019 com ocorrências de fraude ou que foram comprovadas não possuírem irregularidades na medição. Para essa triagem, foi levado em consideração o sistema de dados da empresa que disponibiliza apenas uma janela de 5 anos devido ao grande volume de informações. Dessa forma, selecionaram-se apenas as unidades com inspeção nesses 2 anos para que fosse garantido que houvesse a disponibilidade de pelo menos 3 anos de consumo para cada cliente.

Dessa maneira, o banco de dados foi separado conforme a Tabela 2. Alguns métodos necessitam que haja um equilíbrio entre os perfis a serem classificados na base de treinamento, como é o caso da Rede Neural Artificial (RNA). Dessa forma, foi montado também um banco que atendesse a esse critério. As unidades descartadas com o perfil "Normal" foram escolhidas de maneira aleatória.

**Tabela 2. Distribuição das unidades consumidoras consideradas no banco de dados**

| Banco de Dados | Fraude | Situação Normal | Total de UCs |
|---|---|---|---|
| Treinamento Desequilibrado | 17.735 | 63.528 | 81.263 |
| Treinamento Equilibrado | 17.735 | 17.735 | 35.470 |
| Teste / Validação dos Modelos | 1.000 | 13.285 | 14.285 |

## 4. METODOLOGIA

Objetivando-se desenvolver uma metodologia para detecção de fraude em clientes da baixa tensão, com a finalidade de aumentar os índices de efetividade, o fluxo de processos empregado encontra-se resumido na Fig. 1. Para implementação, foi utilizada a linguagem de programação Python e seu conjunto de bibliotecas para toda a extração, manipulação e classificação dos dados.

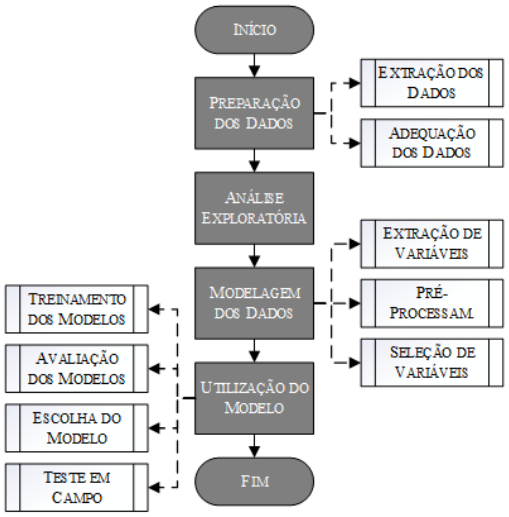


Fig. 1 Fluxo da metodologia empregada.

### *2.1 Preparação dos Dados*

Como mencionado, a extração dos dados foi feita a partir dos sistemas comerciais da empresa. Inicialmente, as tabelas exportadas precisaram ser adequadas para considerar uma das datas de inspeção da unidade como a data de referência em que se conhece a classificação da UC, seja ela fraude ou situação normal. O perfil antes ou após a data de referência é desconhecido, tendo em vista que a única maneira de comprovar uma fraude é através de inspeção em campo.

No banco de dados, apenas o consumo anterior à data de referência deve ser considerado, pois 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. Após essa data, não é possível determinar se há uma irregularidade de medição na unidade consumidora, a menos que haja uma nova inspeção.

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 às tabelas de consumo, irregularidades e serviços. Além disso, os dados de cadastro foram atualizados com as informações válidas naquela data específica. O consumo de uma unidade no banco de dados será conforme Fig. 2. As unidades com menos de 24 pontos de consumo foram expurgadas da base. Essas curvas de consumo são utilizadas posteriormente para extração de variáveis.

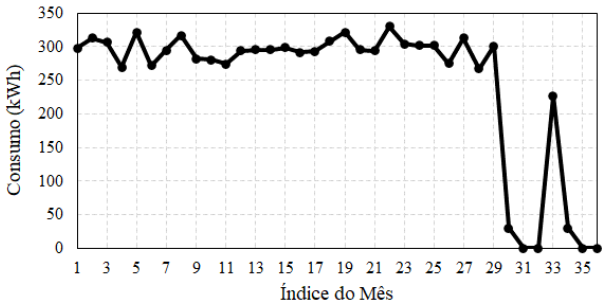


Fig. 2 Exemplificação de uma curva de consumo ajustada para 36 meses com base na data de inspeção.

### *2.2 Análise Exploratória*

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

Nessa etapa foi observada a presença de um desbalanço da variável *target* no conjunto de dados, em que a classe de não fraudadores é muito superior à de fraudadores, conforme observa-se na Fig. 3. A variável target, também denominada alvo, é aquela que se deseja classificar a partir dos valores de entrada. É a saída do modelo. Nesse caso, representa a presença, "S", ou não, "N", de fraude em uma UC.

É esperado, ainda, que essa desproporção seja maior quando observada toda a população. As ocorrências mais comuns são os desvios e os procedimentos irregulares nos medidores, respectivamente.

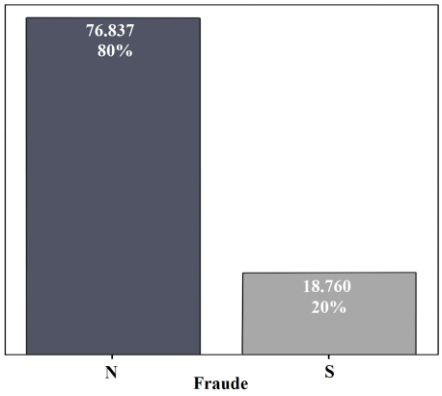


Fig. 3 Distribuição das unidades no banco de dados classificadas de acordo com a variável *target* fraude.

Nessa etapa de análise exploratória, os dados foram investigados com o objetivo de encontrar informações que pudessem ser convertidas em variáveis, possibilidades de agrupamentos, desbalanços e identificação de elementos individualmente promissores para caracterizar uma fraude. Apesar dessa etapa ter sido evidenciada como o segundo passo no fluxograma da Fig. 1, ela faz parte de todo o processo, já que auxilia a compreensão dos modelos e a tomada de decisão.

### *2.3 Metodologia Proposta para Detecção de Fraude*

Após a análise exploratória inicial e a partir das tabelas ajustadas, variáveis puderam ser extraídas como possíveis candidatas para compor o conjunto final dos dados de treinamento.

Características como o tipo de ligação e indicação de conjunto habitacional foram convertidas diretamente em variáveis categóricas, enquanto outras, como as relacionadas à quantidade, foram transformadas por meio da categorização. As informações relacionadas ao consumo da unidade foram utilizadas como variáveis contínuas.

Quando desejava-se eliminar o ordenamento ou reduzir a dispersão das variáveis contínuas, utiliza-se um mapeamento para transformá-las em categóricas através do método de quantização ou *binning*. A quantidade de irregularidades de leitura ou faturamento são exemplos de variáveis dispersas. Elas foram categorizadas em subtipos denominados “ausente”, “pouco”, “médio” e “muito”. Suas distribuições foram utilizadas para determinar os pontos limites para construção das categorias, como pode ser observado na Fig. 4 para a irregularidade de faturamento “Desligado” e “Média”. Dessa maneira, UCs que estiveram apenas um mês desligadas no sistema de faturamento foram categorizadas como “Pouco Desligado”; as que estiveram entre dois e quatro meses foram categorizadas como “Médio Desligado”; e as que passaram cinco ou mais meses desligados foram categorizadas como “Muito Desligado”.

Outros casos de variáveis dispersas convertidas em dados categóricos foram: quantidade de inspeções já feitas na unidade consumidora; média de dias que o cliente leva para realizar um pagamento; quantidade de contas atrasadas em 36 meses; e quantidade de pontos na curva de consumo fora do intervalo entre o 1º e o 3º quartil.

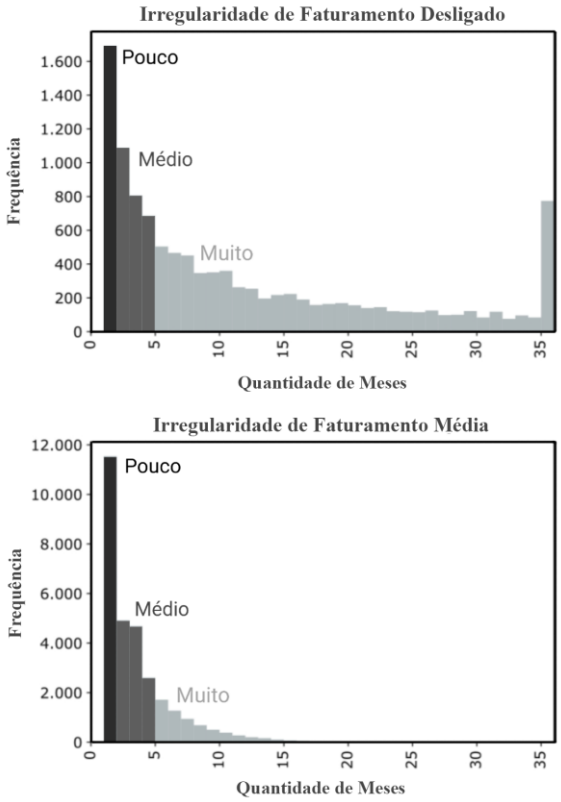


Fig. 4 Limites para categorização das variáveis Desligado e Média determinados através do histograma.

É possível que existam características das unidades consumidoras que não 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 compilando, identificando tópicos, padrões e palavras-chave relevantes. Baseada em Desai et al. (2015), a metodologia utilizada consistiu em:

1) Retirar as pontuações e *stopwords*.

2) Aplicar o processo de *stemming*.

3) Analisar o espectro de frequência das palavras a fim de extrair informações desconhecidas ou complementares para a identificação de fraude.

4) Extrair informação: classificar resultados em medição externa, unidade desligada, unidade desocupada, medição faz parte de um conjunto habitacional.

A maior parte das métricas de consumo foram construídas a partir das estatísticas da curva subdividindo-a em 3 séries de 12 meses. Essas estatísticas foram: máximo, mínimo, média, desvio padrão, assimetria e curtose. Optou-se por séries de 12 meses para evitar sazonalidades. Outras métricas incluíram a comparação entre essas séries. Para isso, os coeficientes de correlação de Pearson e Spearman, bem como a variável de redução de consumo definida em (1), denominada degrau, foram utilizadas:

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

Em que: $\text{cons}_{\text{atual}}$ é o consumo atual, podendo se referir a energia do mês ou a média de um intervalo arbitrário; e $\text{cons}_{\text{ant}}$ 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, imóveis de aluguel, unidades comerciais e industriais que variam de acordo com a demanda de mercado, manutenções ou reformas são exemplos de UCs que podem reduzir drasticamente seu consumo sem necessariamente ocorrer fraude.

As variáveis mencionadas até então não solucionam o exposto por Viegas et al. (2017) relativo à identificação de irregularidades em unidades que não possuíram variação no histórico de consumo. Propõe-se, então, uma nova variável em que o consumo de uma unidade é comparado com a de seus vizinhos semelhantes, denominada Degrau Vizinhos. O objetivo é encontrar imóveis com padrões de consumo abaixo do esperado para a mesma classe de consumo de uma mesma região. A concepção dessa métrica encontra-se ilustrada na Fig. 5, 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.

Fig. 5 Concepção das variáveis associadas ao consumo que compara vizinhos geograficamente próximos e com características semelhantes a uma UC.

Para determinar quais vizinhos seriam utilizados para compor a média de consumo comparativa de um cliente, fez-se seleção de um conjunto de consumidores da mesma classe de consumo, com o mesmo tipo de ligação (monofásico, bifásico ou trifásico) e sem histórico de fraude. Em seguida, 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. Por fim, uma variável denominada Média Vizinhos foi calculada através da média dos 3 consumidores restantes.

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.

A métrica "Degrau Vizinhos" 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 de (1). Em comparativo com a regra de "Degrau" da Tabela 1, que também será utilizada como uma das entradas do modelo, a variável "Degrau Vizinhos" possui vantagens evidentes. A Fig. 6 foi gerada a fim de compreender melhor o comparativo entre as duas métricas, onde se categorizou os valores para facilitar essa análise. Degraus negativos, ou seja, em que houve uma redução de consumo, possuem índice N e degraus positivos, possuem índice P. O percentual indicado corresponde a frequência da presença ou ausência de fraude em cada categoria criada, ou seja, na categoria N1, a variável "Degrau 0-1" possui 33% de casos de Fraude ("S"), enquanto a variável "Degrau Vizinhos" possui 35%.

É possível observar que, diferente da variável "Degrau" utilizada para geração de listas de inspeção atualmente (Tabela 1), a métrica "Degrau Vizinhos" segue a lógica de que degraus positivos possuem menos ocorrências de fraude que degraus negativos. Além disso, o percentual de acerto, isto é, a quantidade de fraudes pelo total de unidades indicadas pela classe, para N1, N2 e N3 são superiores ou equivalentes à do degrau usual com uma cobertura superior, atingindo uma maior quantidade de unidades.

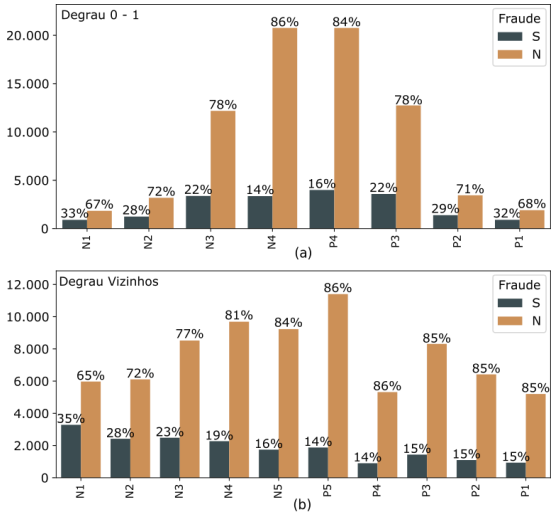


Fig. 6 Variável de degrau separado por classes para análise de ocorrências de fraude. (a) Degrau 0-1. (b) Degrau Vizinhos.

Após a extração de todas as variáveis, faz-se necessário limpar a base de unidades que possam atrapalhar o 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 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 às demais. Dessa maneira, fez-se uma varredura nos atributos categóricos e optou-se por agrupá-los.

Ademais, a partir das técnicas apresentadas em Pyle (2019) para normalização, foi utilizada a transformação linear (mín-máx) em conjunto com o clip para valores fora de uma faixa determinada para cada variável com base no observado para a base de treinamento. O escalonamento foi feito para manter os valores sempre entre 0 e 1.

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. Utilizou-se o coeficiente de correlação de Pearson para identificar todos os atributos com $r > 0,8$ ou $r < -0,8$. De posse desses atributos, foram analisados os conjuntos correlacionados e escolheu-se apenas uma das variáveis 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 de perdas, pareciam as mais relevantes para identificação de uma fraude.

As variáveis categóricas do modelo, além da saída Fraude, incluem: tipo de ligação; indicação de conjunto habitacional; quantidade de irregularidades do tipo desligado, desocupado, média e mínimo; indicação de presença de proteção nos bornes do medidor; indicação de medição externa; quantidade de inspeções já realizadas na UC; média de dias para realizar o pagamento das contas; quantidade de contas atrasadas; e quantidade de pontos fora do intervalo interquartil para o consumo do último ano e para o consumo dos 36 meses.

As variáveis contínuas do modelo incluem: máximo, média, mínimo, desvio padrão, assimetria, curtose da série de consumo; último consumo da UC; degrau calculado conforme equação (1) para atual = média de consumo do último ano e ant = média de consumo dos 12 meses intermediários; degrau semelhante a variável anterior, mas para ant = média de consumo dos 12 primeiros meses; degrau vizinhos; média dos vizinhos; coeficiente de correlação de Person entre as séries de 12 meses de consumo; e teste de normalidade de Shapiro-Wilk na série de consumo.

Ao final de todo o processo, foram geradas 65 variáveis, sendo 44 categóricos e 21 contínuos (Fig. 7).

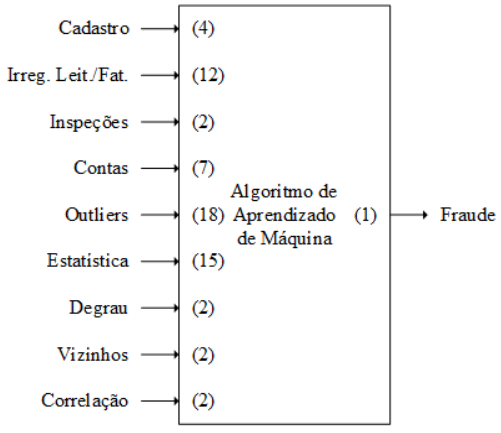


Fig. 7 Esquemático do modelo no formato entrada versus saída.

Não é possível estabelecer a priori qual a melhor técnica de Aprendizado de Máquina (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. As técnicas testadas foram escolhidas com base nas mais observadas na literatura. Para validação e escolha dos modelos, foi separada uma base de teste e aplicado o seguinte processo de avaliação:

1) Escolha dos principais parâmetros: para cada técnica determinou-se empiricamente valores para os principais parâmetros e diferentes modelos foram avaliados para uma mesma técnica.

2) Treinamento com validação cruzada: treinamento com validação cruzada k-fold para k = 5 e exportação da matriz de confusão, curva ROC, f-score e intervalo de confiança obtidos.

3) Treinamento com base completa: sem a presença da validação cruzada.

4) Classificação da base teste: exportação da matriz de confusão, tomado como uma simulação de um caso real.

A validação cruzada foi aplicada para determinar o modelo que possuía a melhor performance com base, principalmente, no f-score, na AUC ROC e no intervalo de confiança.

## 5. RESULTADOS

Para comparar as técnicas de Aprendizado de Máquina (AM), utilizou-se a validação cruzada do tipo k-Fold para k = 5 e a classe de consumo residencial como primeiro teste. Em alguns modelos, 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 as técnicas LDA, Árvore de Decisão, *Random Forest* e *Gradient Boosting* foi utilizado o banco desbalanceado. Para as demais, foi utilizado o balanceado. Na Tabela 3 são apresentadas as principais métricas de avaliação dos modelos após a utilização da validação cruzada. Em que: E é a efetividade, C é a cobertura, ROC AUC é a área abaixo da curva ROC, F1 é o f-score e I.C é o intervalo de confiança calculado com base no f-score.

**Tabela 3. Principais indicadores das inspeções realizadas entre 2017 e 2019 descriminado pela regra utilizada**

| Técnica | E. | C. | ROC AUC | F1 | I.C. |
|---|---|---|---|---|---|
| LDA | 59,1% | 19,6% | 74,0% | 29,5% | 1,1% |
| Árvore de Decisão | 14,0% | 41,0% | 60,6% | 21,0% | 1,1% |
| *Random Forest* | 68,4% | 23,8% | 78,9% | 35,3% | 1,0% |
| *Support Vector Machine* | 50,7% | 25,9% | 70,7% | 34,3% | 0,9% |
| Rede Neural MLP | 66,1% | 65,4% | 71,8% | 65,8% | 1,3% |
| *Gradient Boosting* | 62,1% | 34,8% | 78,5% | 44,6% | 1,6% |

Considerando os modelos que apresentaram melhores percentuais de f-score e AUC da curva ROC, aplicou-se o

teste com o banco de dados da Tabela 2 a fim de simular a efetividade em campo. Para aplicação desse teste utilizou-se a técnica de *holdout* aleatório estratificado aplicado 10 vezes a base para obter os resultados que serão apresentados.

Para a Rede Neural, 658 unidades com fraude foram corretamente classificadas com uma efetividade de 12% e uma cobertura de 66%, o que indica um alto custo associado a esse modelo por exigir muitas inspeções com pouca assertividade. A arquitetura da rede foi montada com o algoritmo L-BFGS, 100 neurônios na primeira camada, 80 na segunda camada e o ReLu como função de ativação.

Para o *Gradient Boosting*, 350 unidades com fraude foram corretamente classificadas com uma efetividade de 28% e uma cobertura de 35%. O algoritmo considerou 100 estimadores e profundidade máxima de 2.

Para o *Random Forest*, 240 unidades foram classificadas corretamente com uma efetividade de 35% e uma cobertura de 24%. Os parâmetros utilizados consideraram 150 estimadores e um balanceamento de classes de 1 para a não fraudadores e 2 para fraudadores.

Na Tabela 4, os principais indicadores dos três modelos utilizados são resumidos. A maior efetividade foi obtida através do *Random Forest*, com valores próximos ao que se obtém hoje pela regra de suspeita, considerada referência na empresa, e 218% superior à de degrau, sendo esta última a regra que deve ser considerada para comparativo com a metodologia. Vale ressaltar que os resultados obtidos nesse teste estão mais próximos ao que se deve obter em campo devido a proporção da base de dados utilizada. Na validação cruzada não foram utilizados os mesmos 7% devido a subdivisão em 5 partes iguais do banco de treinamento, seu objetivo foi verificar a performance geral do modelo.

**Tabela 4. Resultado do teste teórico para classificação de fraude**

| Técnica | E. | C. | ROC AUC | F1 |
|---|---|---|---|---|
| Random Forest | 35% | 24% | 60% | 28% |
| Rede Neural MLP | 12% | 66% | 65% | 21% |
| Gradient Boosting | 28% | 35% | 64% | 31% |

## 6. CONCLUSÕES

Neste trabalho, foram propostos diversos atributos com base nos dados da empresa de estudo que buscaram avaliar o comportamento do cliente de diversos ângulos. A principal variável proposta buscou comparar a média de consumo da unidade com os vizinhos geográficos mais próximos que possuíam características de porte semelhantes.

As técnicas de aprendizado de máquina tiveram suas performances avaliadas de maneira estruturada e de acordo com as principais métricas sugeridas pela literatura. Observou-se um destaque para os modelos do *Random Forest*, Rede Neural Artificial e *Gradient Boosting* que foram selecionados para testes teóricos por possuírem os maiores f-scores e área abaixo da curva ROC na validação cruzada. No teste teórico, a efetividade do *Random Forest* ultrapassa os obtidos pela empresa em campanhas com regras que analisam o consumo em 218%, atingindo 35% de efetividade.

Como trabalhos futuros, serão avaliadas as possibilidades de melhoria dos métodos a partir de novas variáveis, técnicas de seleção de variáveis e alteração de parâmetros. As variáveis e os modelos serão aplicados na base da empresa e inspeções serão feitas em campo para verificar os ganhos em energia e financeiros para a distribuidora.

## REFERÊNCIAS


ANEEL., (2015). *Perdas de Energia*. [Online]. Disponível em: http://www.aneel.gov.br/metodologia-distribuicao/-/asset_publisher/e2INtBH4EC4e/content/perdas/654800.

ANEEL., (2019). *Relatórios de Consumo e Receita de Distribuição: Consumidores, Consumo, Receita e Tarifa Média – Região*. [Online]. Disponível em: http://www.aneel.gov.br/relatorios-de-consumo-e-receita.

Angelos, E. W. S., Saavedra, O. R., Cortés, O. A. C. and Souza, A. N., (2011). Detection and Identification of Abnormalities in Customer Consumptions in Power Distribution Systems. *IEEE Transactions on Power Delivery*. Vol. 26, no. 4, 2436-2442.

Araujo, B., Almeida, H. and Mello, F., (2019). Computational Intelligence Methods Applied to the Fraud Detection of Electric Energy Consumers. *IEEE Latin America Transactions*. Vol. 17, no. 01, 71-77.

Desai, P. G., Sarojadevi, H. and Chiplunkar, N. N., (2015). A template Based Algorithm for Automatic Summarization and Dialogue Management for Text Documents. *International Journal of Research in Engineering and Technology*. Vol. 04, no. 11, 334-340.

Guerrero, J. I., León, C., Monedero, I., Biscarri, and F., Biscarri, J., (2014). Improving Knowledge-Based Systems with statistical techniques, text mining, and neural networks for non-technical loss detection, *Knowledge-Based Systems*. Volume 71, 376-388.

Nagi, Yap, J., K. S., Tiong, S. K., Ahmed, S. K. and Mohamad, M., (2010). Nontechnical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines. *IEEE Transactions on Power Delivery*. Vol. 25, no. 2, 1162-1171.

Ponce-Jara, M.A., Ruiz, E., Gil, R., Sancristóbal, E., Pérez-Molina, C., and Castro, M., (2017). Smart Grid: Assessment of the past and present in developed and developing countries. *Energy Strategy Reviews*. Volume 18, 38-52.

Pyle, D., (1999). Data Preparation for Data Mining. San Francisco: Morgan Kaufmann Publishers, 1999.

Ramos, C. C. O., Rodrigues, D., Souza, A. N. and Papa, J. P., (2018). On the Study of Commercial Losses in Brazil: A Binary Black Hole Algorithm for Theft Characterization. *IEEE Transactions on Smart Grid*. Vol. 9, no. 2, 676-683.

Viegas, J. L., Esteves, P. R., Melício, R., Mendes, V.M.F., and Vieira S. M., (2017). Solutions for detection of non-technical losses in the electricity grid: A review. *Renewable and Sustainable Energy Reviews*. Volume 80, 1256-1268.

Zheng, Z., Yang, Y., Niu, X., Dai, H. and Zhou, Y., (2018). Wide and Deep Convolutional Neural Networks for Electricity-Theft Detection to Secure Smart Grids. *IEEE Transactions on Industrial Informatics*. Vol. 14, no. 4, 1606-1615.