Apontamentos de Clientes de Distribuidoras de Energia Elétrica com Indícios de Fraude Utilizando Machine Learning

J. J. Borges Oliveira *, V. H. Ferreira. *

Universidade Federal Fluminense,
Programa de Pós-graduação em Engenharia Elétrica e de Telecomunicações,
Niterói, RJ
RJ, (e-mail: josuejbo@id.uff.com.br, vhferreira@id.uff.br)

Abstract: This article proposes the use of the supervised learning method to establish a scenario of classification and identification of possible customers of electric energy distributors with signs of irregularities. Using this method, the failures identified by energy utilities during the target generation process can be mapped, contributing to better assertiveness and decision making in the field. The model will use data belonging to a large electricity distributor in the state of Rio de Janeiro, to make a comparison between different processing and pre-processing approaches to point out evidence of fraud. For this, tests were carried out with different classifier algorithms and feature extraction methods, obtaining as a best result an average accuracy close to 70%.

Resumo: Este artigo propõe a utilização do método de aprendizagem supervisionada para estabelecer um cenário de classificação e identificação de possíveis clientes de distribuidoras de energia elétrica com indícios de irregularidades. Usando este método, as falhas identificadas pelas concessionárias de energia durante o processo de geração de alvos poderão ser mapeadas, contribuindo para melhor assertividade e tomada de decisão em campo. O modelo utilizará dados pertencentes à uma grande distribuidora de energia elétrica do estado do Rio de Janeiro, para realizar uma comparação entre diferentes abordagens de processamento e pré-processamento para o apontamento de indícios de fraude. Para isso, foram realizados testes com diferentes algoritmos classificadores e métodos de extração de características, obtendo como melhor resultado uma média de acerto próxima a 70%.

Keywords: Supervised Learning; Classification Algorithms; Electricity distributors; System Distribution; Feature Extraction Methods; Identification; Fraud Reporting.

Palavras-chaves: Aprendizagem Supervisionada; Algoritmos de Classificação; Distribuidora de Energia Elétrica; Sistema de Distribuição; Método de Extração de Características; Identificação; Indícios de Fraude.

1. INTRODUÇÃO

O Brasil tem enfrentado grande instabilidade econômica que vem se arrastando por anos. Além disso, o impacto causado pelo novo coronavírus agregado à essa crise, culminou na ampliação do cenário de desigualdade social, derivada do aumento do desemprego e da redução dos rendimentos reais.

Segundo os *White Papers* do instituto Ascende Brasil (2017), a taxa de desocupação no país chegou a 14,6% no terceiro trimestre do ano, o que representa alta de 1,3 ponto percentual na comparação com o trimestre anterior, quando ficou em 13,3%. Em consequência disso, o setor de distribuição de energia elétrica vivencia uma adversidade desafiadora, que é o aumento das perdas não técnicas de eletricidade (PNT).

Entende-se por perda não técnica, ou comercial, aquela oriunda de roubo de energia (desvio da rede de distribuição ou ligações clandestinas) ou de fraude de energia (modificações no medidor). Segundo a ANNEL em 2018 o sistema elétrico registrou cerca 6,6% de perda não técnica do total da energia injetada. Já a perda técnica, é inerente ao processo de distribuição de energia, sendo causadas pelo processo de dissipação durante o transporte e devido à excitação magnética quando em vazio.

As perdas comerciais de energia elétrica por irregularidades, por exemplo, representam para o país um prejuízo médio total de R$ 5 bilhões por ano em média (Dutra, 2016). Esse montante reflete cerca de 5% da energia total consumida no país (Ascende Brasil, 2017). Grande parte dessas perdas ocorre na rede da empresa Light S.A., que atende uma média de 7 milhões de unidades consumidoras em 31 municípios do Rio de Janeiro (Light, 2021).

A distribuidora de energia estudada. mapeou geograficamente as áreas onde há maior índice de perdas e concluiu que existe uma enorme relação entre o nível de perdas não técnicas da empresa e as peculiaridades do Rio de Janeiro (RJ). Com objetivo de combater esse grande ofensor, a instituição estuda meios para mitigar o alto valor de perda identificado.

Para tanto, nota-se que uma maneira objetiva de aferir esses resultados baseia-se nos apontamentos de clientes suspeitos de fraudes. O modelo de identificação de clientes com possíveis irregularidades atualmente praticado na empresa se baseia em técnicas computacionais para detecção de possíveis locais em fraude, tendo por base regras heurísticas construídas a partir da experiência de cada operador. Este modelo é pautado em um conjunto de metodologias e regras que identifica clientes de baixa tensão suspeitos de estarem cometendo algum tipo de irregularidade. Por meio deste processo, os indicadores da empresa mostraram que o resultado médio de acerto obtido na comprovação de clientes irregulares no ano de 2020 foi de 26%. Para tal resultado, foram consideradas algumas premissas, como clientes com fornecimento a baixa tensão do grupo convencional e resultado de notas geradas no programa de inspeção (UBI) no ano de 2020. Dessa forma, verifica-se que o procedimento praticado apresenta uma oportunidade de melhoria no que diz respeito à sua taxa de assertividade na identificação do cliente com fraude.

Neste contexto, este artigo busca apresentar um modelo de aprendizado de máquina capaz de interpretar corretamente os elementos presentes nos dados e indicar alvos suspeitos de fraudes, dando mais eficiência ao processo de combate e recuperação de energia. Para isso, foram testadas diferentes técnicas de extração de características e algoritmos de classificação, permitindo a delimitação da melhor estratégia a ser implementada. Essa estratégia foi aplicada a uma base de dados real de inspeções obtendo resultados promissores em termos de taxa de assertividade.

2. TRABALHOS RELACIONADOS

A utilização de novos modelos estatísticos tem a capacidade de proporcionar um impacto significativo na solução de problemas relacionados a geração, comércio, consumo de energia e principalmente na detecção de perdas não técnicas. Um exemplo disso são as Redes Neurais Artificiais, que podem ser especialmente úteis para questões desse tipo de problema (Megahed et al., 2019). Outro exemplo são as árvores de decisão, aplicadas em Filho et al. (2004) para selecionar clientes sujeitos a inspeções.

Na literatura, diversos trabalhos já registraram as dificuldades deste tipo de investigação, principalmente no que se refere à avaliação de registros classificados como normais, porém que podem estar contaminados por irregularidades que não foram detectadas durante a inspeção em campo (Rauber et al., 2005).

Um método de pré-processamento de dados foi empregado para melhorar o desempenho de detecção em (Ahamad, 2022). Vários algoritmos, incluindo aumento adaptativo, aumento categórico, aumento extremo, floresta aleatória e árvores extras, foram testados para encontrar suas taxas de falso positivo e detecção.

Cabral et al. (2004) usaram *Rough Sets* para o descobrimento de irregularidades de medição. Os resultados obtidos com esse modelo apresentaram um baixo valor de precisão (20%) devido à presença de ruídos existentes na base de dados da empresa cujos dados foram utilizados no estudo.

Rauber et al. (2005) e Rong et al. (2002) abordam de forma semelhante o problema de classificação de irregularidades. Nestes trabalhos, foram usados somente consumos históricos. Como os consumos apresentam natureza temporal, foi necessária a aplicação de métodos de análise de séries temporais nestes estudos (Pollock, 1999) na busca de extrair novas características invariantes. Além disso, Cabral et al. (2004) trazem uma revisão sobre os principais métodos dedicados à detecção de perdas não técnicas, presentes nos medidores inteligentes que podem contribuir para a solução do problema.

Nizar et al. (2007) e Angelos et al. (2011) utilizaram algoritmos de *clustering* para detecção de fraudadores. Depois da etapa de agrupamento, as análises dos perfis de consumo foram feitas pela comparação com os grupos selecionados. Já em Muniz et al. (2009), via método de classificação, uma rede neuro-fuzzy hierárquica foi utilizada buscando melhorar o desempenho.

Faria et al. (2014) propõe uma solução para identificar padrões de roubo de energia adicionando procedimentos forenses através da identificação de equipamentos eletrônicos adulterados. Por sua vez, Buzau et al (2018) emprega aprendizado de máquina para detecção de fraudes com utilização específica do classificador *XGBoost*.

Além do ótimo desempenho observado na fase de teste, a escolha do *Random Forest* como método de inteligência artificial foi pela capacidade do algoritmo em correlacionar e combinar as características conhecidas hoje, possibilitando a construção de um modelo capaz de se adaptar a mudanças e sazonalidades de dados permitindo uma adaptação aos eventos que levam a fraude e que sofrem evolução com o tempo.

2.2 Random Forest

O *Random forest* é um algoritmo que cria diversas árvores de decisão e faz combinações entre elas para se obter uma maior acurácia na predição. Sendo mais específico, ele é um conjunto de diversas árvores de decisão que possuem diferentes nós, sendo estes gerados aleatoriamente para a classificação desejada. Ao final, assim que todas as árvores tenham terminadas sua classificação individual, o algoritmo realiza um comitê para validar de fato qual a classificação mais indicada.

Para melhor entendimento do método, digamos que exista um problema bem simples para classificar o formato e a cor de um objeto, podendo este ser apenas um triângulo ou um quadrado, preto ou branco, e que além disso, cada pergunta feita diante do problema proposto, seja considerada um nó. Sendo assim, todo "nó" identificado irá dividir a árvore em dois caminhos diferentes que no decorrer do problema não irão se cruzar em nenhum momento.

Nesta situação, a criação do primeiro nó de decisão aconteceria mediante a pergunta de quantos lados possui a figura, como por exemplo, "o objeto tem 4 lados?". Se sim, a parte relevante da árvore se torna a da direita com a classificação de um "quadrado", caso a resposta seja negativa a relevância seria a da esquerda classificado como "triângulo". E assim segue a sequência de perguntas até que o resultado final indique a cor e a forma do objeto em questão.

Vale ressaltar que ao fim do modelo teremos 4 registros: triângulo preto, triângulo branco, quadrado preto e quadrado branco, apresentados na figura 1.

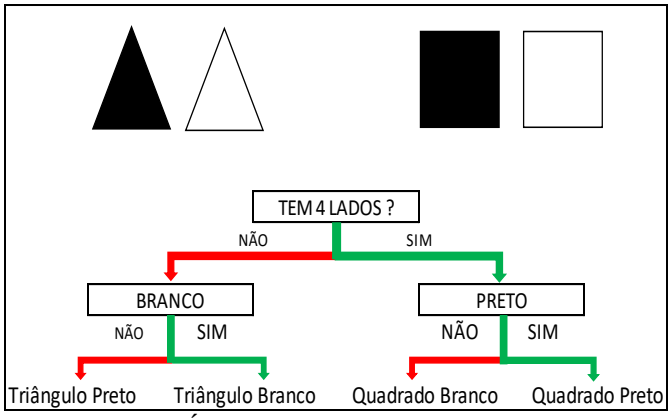


Fig. 1 Exemplo: Árvore de decisão - RF.

3. ANÁLISE EXPLORATÓRIA DOS DADOS

A área de concessão da Companhia abrange cerca de 26% (11.307 mil km²) do Estado do Rio de Janeiro e comporta uma população de 11 milhões de pessoas, representando 64% da população total do Estado. Dos 92 municípios do Estado com um total de 7 milhões de consumidores de energia elétrica, a Companhia atua em 31 municípios, representando 34% dos municípios totais, e possui uma base de cerca de 4,5 milhões de clientes (Light, 2021).

A abrangência das UCs consideradas no estudo corresponde a 20 municípios pertencentes ao Estado do Rio de Janeiro, que representam um volume de 80 mil instalações aleatórias pertencentes ao grupo B convencional (baixa tensão) e inspecionados no ano de 2020 (janeiro a dezembro), cuja base de dados corresponde a 52.608 registros da classe de clientes com resultados ditos regulares e 27.392 irregulares. Para análise dos casos de reincidência, foram considerados todos os apontamentos gerados pelo sistema da distribuidora no ano de 2019. O objetivo desta restrição é não comprometer o resultado final do teste com o fornecimento de informações anterior a análise.

Os dados utilizados foram fornecidos pela concessionária de energia elétrica e são originários de diversas bases de dados que contêm atributos gerais dos clientes residenciais da empresa. A Tabela 1 elenca uma lista com os atributos disponíveis na base de dados disponibilizada.

Tabela 1. Atributos Selecionados na Base de dados

Atributo	Descrição
Dados do Consumidor	Localização geográfica da Unidade Consumidora (UC).
Dados de Consumo	Histórico de consumo dos clientes, consumo medido e faturado.
Notas de leitura	Apontamento de irregularidade constatada pelo leiturista/campo
Notas de Serviço	Fechamento da nota de serviço executado pela Equipe

Os dados do consumidor são informações que podem ser obtidas através de formulários de cadastro e ajudam a caracterizar melhor a unidade consumidora (UC), tais como: pessoa física ou jurídica; localização geográfica; classe de consumo (residencial, comercial, industrial ou poder público).

Os dados de consumo são formados pela data de referência em que ocorreu a leitura, a energia consumida/medida e o consumo faturado. É também considerada uma informação muito valiosa no quesito tarifário e de grande importância na escolha do vetor de característica que corresponde a definição do modelo de aprendizado de máquina.

As notas de leitura estão relacionadas aos apontamentos que os leituristas realizam durante a verificação em campo. Em caso de suspeita de irregularidade, fraude ou qualquer outra anomalia que esteja impactando o real registro de consumo do cliente, deve-se realizar o registro desta situação. Apesar de não expressar categoricamente uma certeza de que o consumidor está irregular, esses direcionamentos servem de input para o sistema de seleção de alvos da distribuidora, e inserem no sistema a informação de cliente suspeito com apontamento de indício. Exemplo disso são as discriminações de irregularidades (IR) identificadas no registro do cliente.

As notas de serviço referem-se ao código lançado no sistema SAP pela equipe após inspeção que caracteriza a real condição do cliente vista pela equipe durante inspeção, tais como: irregularidade (IR); fraude (FR); inspecionado (INSP); não inspecionado (NI) e nada apurado (NA). As notas fechadas como NI podem ter vários significados, dentre eles: endereço não encontrado, área de risco, impedimento de acesso, entre outros que justifique o não inspecionado. As notas fechadas como NA, representam que durante a inspeção a equipe não verificou anomalias que pudesse comprometer o registro do consumo do cliente.

Dentre os 20 municípios pertencentes ao Estado do Rio de Janeiro, para melhor análise foram agrupados o número de ocorrências inspecionadas, bem como seus respectivos resultados de irregularidade, fraude encontrada ou nada apurado. Com base nessas informações, foi possível realizar um ranking com a lista de municípios que lideram a categoria dos maiores ofensores no quesito de anomalias identificadas, através da quantidade de notas de serviços com registro de irregularidade verificada, como segue na figura 2.

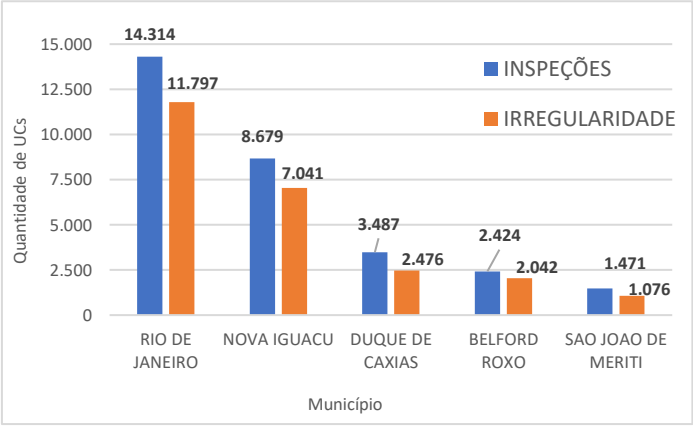


Fig. 2 Volume de Inspeções x Apontamentos de Irregularidade

Como observado, a figura 2 representa cinco municípios com maior registro de inspeções realizadas e identificadas na base do estudo, considerando o período de janeiro a dezembro de 2020. Além desta análise, a figura caracteriza também os maiores números de irregularidades encontradas. Nota-se que o volume de inspeções realizadas nos munícipios destacados, coincide com a quantidade de irregularidade identificada, mostrando assim que a seleção dos alvos verificados foi bem eficaz.

Hoje, algumas das grandes distribuidoras de energia elétrica, já utilizam algumas metodologias baseadas em determinados critérios de consumo para avaliar os clientes de baixa tensão na busca de identificar aqueles com irregularidades. Por meio desta análise, é possível detectar tais variações incomuns, como o degrau por exemplo. Uma técnica muito proveitosa que se baseia na variação do consumo medido mensal do cliente. Nos casos em que essa variação é muito brusca, entende-se que há grande possiblidade do cliente estar em fraude ou irregularidade.

Além da metodologia comentada, tal aplicação pode ser feita também realizando um comparativo com o consumo lido no mesmo mês do ano anterior. Para todas essas situações comentadas, se a variação de utilização for negativa e inferior a 40%, a instalação deste cliente entrará no fluxo de abertura de nota de serviço para inspeção em campo.

Essas metodologias e diferentes critérios são desenvolvidos por profissionais técnicos da área, com intuito de serem mais eficientes quanto aos resultados das ações geradas. Por este motivo, todas essas experiências são importadas por meio de um código na plataforma de inteligência utilizada onde serão tratadas e agrupadas por localidades. Além de obter a indicação do cliente fraudador, este sistema aponta consumidores notados com grande potencial de recuperação financeira, que significa dizer que a energia incorporada após a normalização será maior e mais assertiva quando comparada aos demais clientes com indício.

A aplicação das diferentes metodologias utilizadas na distribuidora estudada, tem conduzido a índices de valor positivo preditivo (VPP) médio de 26%. O VPP é dado pela proporção de clientes comprovadamente irregulares entre todos os clientes que foram classificados como suspeitos de estarem cometendo alguma irregularidade. Este indicador expressa o percentual de clientes irregulares no conjunto daqueles que são suspeitos (Rauber et al., 2005).

A Tabela 2 apresenta uma matriz de confusão típica para o caso de um problema de duas classes como o verificado, por exemplo, clientes normais e irregulares.

Tabela 2. Clientes Suspeitos x Clientes Inspecionados

	Normal	Irregular
Normal	a	b
Irregular	c	d

Em uma matriz de confusão (Tabela 2), os dados contidos nas células indicam o nº de exemplos que possuem a referente classificação, sendo que as linhas indicam a classificação dada pelo modelo proposto, enquanto as colunas indicam os clientes inspecionados/situação real (Cyro et al., 2008).

• 'a' e 'b': indicam exemplos classificados pelo modelo proposto como Normal';

• 'c' e 'd': indicam exemplos classificados pelo modelo proposto como "Fraudador/Irregular'.

Por outro lado, as colunas da matriz de confusão indicam a classificação realizada na inspeção:

• 'a' e 'c': indicam exemplos onde a classificação real é 'Normal';

• 'b' e 'd': indicam exemplos onde a classificação real é 'Fraudador/Irregular'.

4. METODOLOGIA PROPOSTA

A metodologia apresentada neste artigo é composta por três módulos, chamados de pré-processamento, treinamento e inferência. Na fase de pré-processamento, os dados são tratados visando melhor performa do modelo. Como explicitado no fluxograma da Fig.3, para o estudo são utilizadas duas estratégias: “janelamento” e “inspeção”.

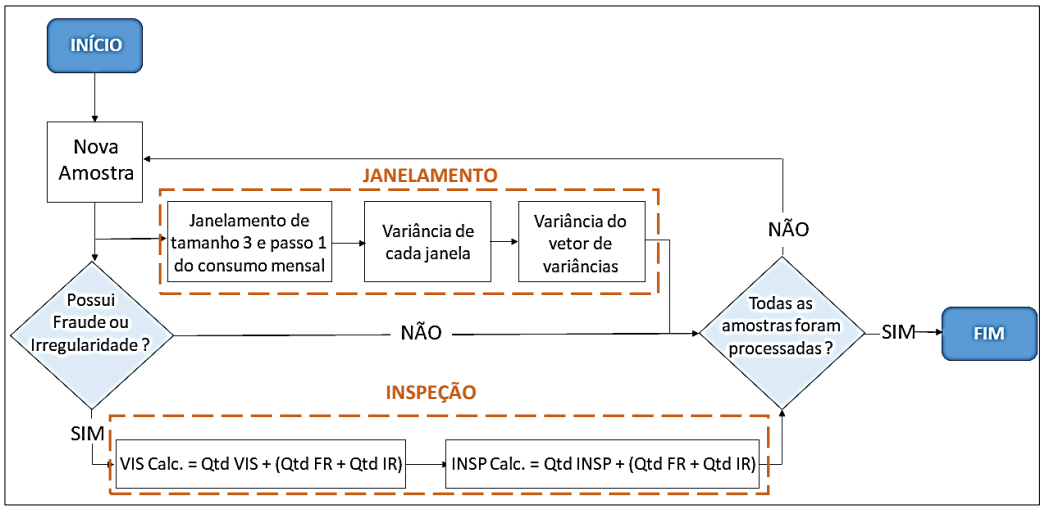


Fig. 3 Fluxograma: Pré-Processamento ou Extração de características.

Porém, antes de detalhar as estratégias abordadas nesta etapa de pré-processamento, destaco que as seis variáveis de entrada utilizadas no estudo correspondem aos resultados das notas de serviços executados pelas equipes de campo, descritas na figura abaixo.

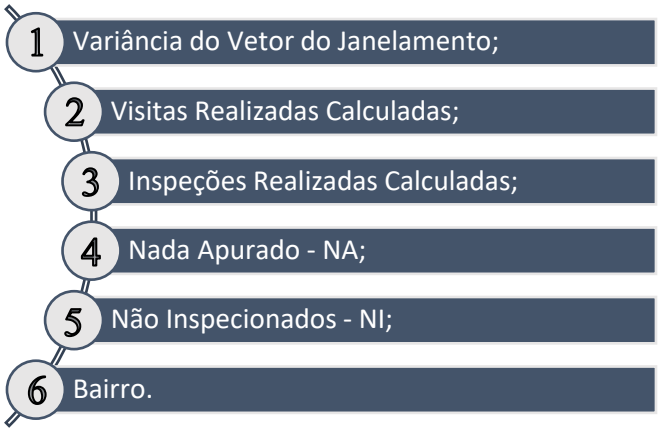
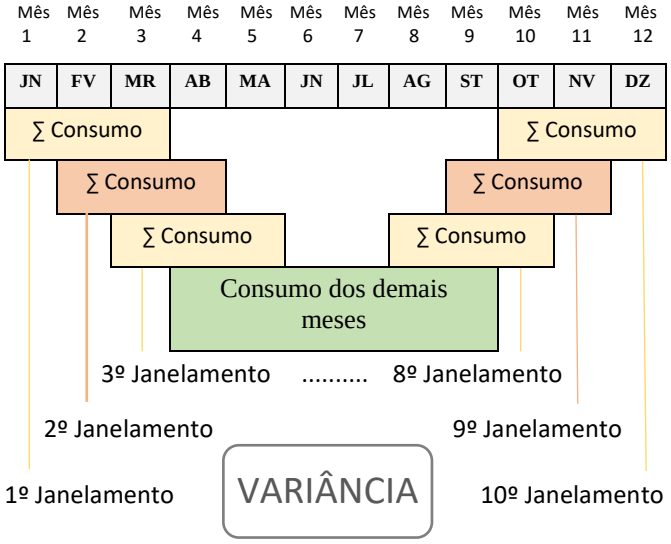


Fig. **4** Demonstração: Variáveis de entrada do modelo.

Quanto as duas estratégias apontadas na figura 3, chamadas de janelamento e inspeção, ambas se referem ao tratamento da base de dados na utilização das informações das UCs como parte das entradas do modelo.

Sendo assim, o modelo de janelamento é definido pelo agrupamento do consumo medido mensal de cada UC no intervalo de três meses consecutivos, como descrito na Tabela3:

Tabela 3. Representação: Janelamento – 12 meses



No exemplo ilustrado acima, a primeira janela será formada pelo somatório do consumo do mês 1, com o consumo do mês 2 mais o consumo do mês 3. Após formação desse janelamento, o modelo avançará para o mês subsequente considerando a mesma ordem de progresso, formando assim a segunda janela, que por sua vez será composta pelo consumo do mês 2, mais o consumo do mês 3, mais o consumo do mês 4. Essa metodologia se aplica para todos os meses do ano de 2020.

Desta forma, o modelo avançará criando grupos de janelamentos até o décimo segundo mês do ano, tendo por fim de processo a formação de 10 janelamentos. Após formação de todos as 10 janelas, vale ressaltar que para cada janelamento se terá também um valor de variância unitário, tendo ao fim dez resultados de variância.

Diante desta análise, o modelo utilizará como uma de suas entradas uma única variância desse conjunto de dez variâncias construídas, o que servirá de *imput* para o algoritmo de classificação.

O cálculo da variância é representado pela fórmula a seguir:

$$VAR = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + (x_3 - \bar{x})^2}{3} \quad (1)$$

x_1, x_2 *e* x_3 = representam o consumo mensal;

$\bar{x}$ = média aritmética;

Para o pré-processamento, foi considerado também uma aplicação chamada de método de inspeção, sendo esta classificada como duas das entradas do modelo: Visitas Realizadas Calculadas e Inspeções Realizadas Calculadas.

O objetivo é evitar que o modelo preveja uma informação de fraude ou irregularidade anteriormente verificada. Desta forma, aplica-se o seguinte cálculo:

$$VIS\ Calculadas = Tot.VIS - (FR + IR)$$
$$INSP\ Calculadas = Tot.INSP - (FR + IR) \quad (2)$$

Tot. VIS = total de visitas realizadas;

Tot. INSP = total de inspeções realizadas;

FR = fraude encontrada;

IR = irregularidade encontrada;

A aplicação deste método permite que o modelo não enxergue informações prévias, fazendo com que a veracidade dos dados seja determinística. Torno a ressaltar que as variáveis INSP e VIS calculadas servem de input para o modelo de classificação, assim como a variância do vetor de dez posições extraído pelo janelamento (Fig.3).

Em relação as demais entradas do modelo, como já comentadas anteriormente estão os resultados das notas de inspeções como: nada apurado (NA), que representa a não verificação de irregularidade que comprometa o real consumo da UC; o não inspecionado (NI) que representa a não execução de inspeção devido algum impeditivo qualquer e o indicativo de Bairro, o que possibilita entender melhor a representação geográfica de concentração das perdas.

Para aplicação do modelo foi utilizado um processador Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz 2.11 GHz, 64 bits e 8GB RAM, sistema operacional *Windows Feature Experience Pack* e uma linguagem de programação *Python*.

É importante citar que, para cada um dos três folds pré-estabelecidos haverá a divisão dos conjuntos de dados em treino e também em testes, como mostrado na figura 5. Através desta divisão, os conjuntos de treino passaram a compor 2/3 da base, enquanto o conjunto de treino será definido por 1/3 da mesma. Desta forma, para cada um fold, será gerado um modelo onde será testado e terá o seu valor registrado. Por fim, se terá três resultados de testes diferentes definidos, o que significa dizer que toda a base será verificada.

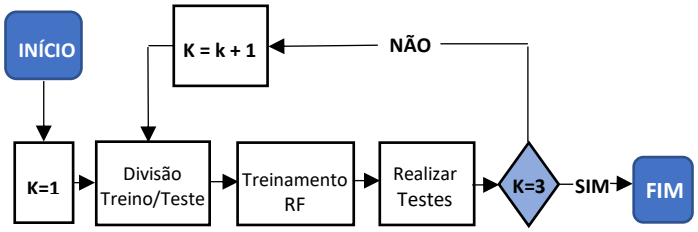


Fig. 5 Fluxograma: Treinamento

Durante este processo as informações dos consumidores são transmitidas ao modelo, e por sua vez tem seus parâmetros ajustados. A base de dados resultante é empregada no treinamento do módulo de classificação, o qual é responsável por indicar se o consumidor pertence à classe regular (normal) ou irregular, correspondendo a um problema de classificação binária.

Durante o processo de treinamento, foi verificado que os dados utilizados possuíam uma característica desbalanceada devido a diferença no quantitativo das classificações de clientes ditos como fraudadores e regulares, tornando complexo para o algoritmo realizar essa tarefa de separação das classes.

Consequência disso, é que o modelo não vai ter a generalidade necessária para trabalhar com o conjunto de dados de teste, sofrendo então o indesejado efeito de sobreajuste (*overfitting*). Em (Gunturi, 2021) o problema é solucionado com a introdução da técnica de sobreamostragem minoritária sintética (SMOTE) e modelos *Near-miss*. Para o trabalho em questão foi feito o uso de subamostragem da classe com maior número de registros, o que solucionou o problema e regularizou o sistema.

Após finalização dos ajustes, o desempenho do modelo é avaliado quando submetido a um conjunto de teste através de diferentes algoritmos.

Em um primeiro momento, para efeito de análise comparativa foi utilizado nos testes um algoritmo de rede neural (Megahed et al., 2019), entretanto, o resultado obtido não foi satisfatório. Além deste, foi testado o *Support Vector Machine*, um algoritmo muito utilizado para classificação em conjunto de pontos onde se busca uma linha de separação entre duas classes distintas (Jang et al., 2016). Durante testes, foi utilizado também o *Xgboost*, um algoritmo baseado em árvore de decisão e que utiliza uma estrutura de *Gradient boosting* (Megahed et al., 2019). Outra verificação foi com a utilização do algoritmo *K Nearest Neighbor*, que determina o rótulo de classificação de uma amostra baseado nas amostras vizinhas advindas de um conjunto de treinamento (Pinto et al., 2019); e por fim de análise, o *Random Forest*, um método de classificação ou regressão que funciona por meio da construção de várias árvores de decisão durante o treinamento (Reddy and Sodhi, 2018).

Dentre os métodos de *Machine Learning* testados nesta etapa, o *Random Forest* obteve uma assertividade maior quando comparado aos demais modelos de classificação. Esta característica foi comprovada no momento em que o modelo propôs resultados significativamente melhores e ajustados.

5. RESULTADOS

Com objetivo de verificar a eficiência do modelo, foi realizado um comparativo de três bases distintas: o resultado do modelo construído, o resultado do modelo utilizado pela distribuidora de energia elétrica e o resultado das inspeções verificadas pelas equipes de Campo.

Em relação aos resultados obtidos pelos diferentes modelos, foi considerado para análise de resultado as saídas "AF" (Acerto Fraude) para os clientes com apontamentos de indícios de irregularidade, saída "AR" (Acerto Regular) para clientes com apontamentos de regulares e "ERRO" para os resultados divergentes do verificado pelas equipes de Campo.

Como o modelo de inteligência utilizado pela distribuidora é acumulativo e tem a capacidade de quantificar diversos apontamentos de indícios para um único cliente de acordo com as suas regras e critérios de seleção de alvos, o trabalho utilizou como parâmetro apenas os clientes que tiveram um saldo de três ou mais indícios, com objetivo de se aproximar do processo de inspeção já realizado pela empresa. Já em relação aos resultados das equipes de campo, foram considerados dois critérios cruciais: apontamentos de cliente com fraude e irregularidade.

Para maior clareza dos resultados observados, as tabelas 4 e 5 apresentam um panorama geral do resultado individual das bases verificadas comparado ao observado em campo:

Tabela 4. Distribuidora vs Campo

	Regular	Fraude	Erro Regular	Erro Fraude
Distribuidora	40.512	8.302	12.096	19.090
Campo	52.608	27.392	-	-
Acerto	77%	30%	**RECALL**	

Tabela 5. Modelo vs Campo

	Regular	Fraude	Erro Regular	Erro Fraude
Modelo	43.687	18.874	8.921	8.518
Campo	52.608	27.392	-	-
Acerto	83%	69%	**RECALL**	

A Tabela 4 elenca o resultado final da ferramenta utilizada atualmente pela distribuidora de energia elétrica, ao passo que a Tabela 5 apresenta o resultado final do modelo construído, comparado com o retorno de campo, que por sua vez foi usado como referência de estudo. É possível observar que o modelo apresenta uma taxa de erro total menor que os demais, assim como uma taxa de erro nas predições ditas como fraudes.

Contudo, nota-se que a taxa de acerto nas predições ditas como fraude obteve uma média de 70%, o que representa uma ótima evolução do modelo no quesito averiguação de fraudes. Além disso, o tempo gasto durante treinamento foi de 2.55min. Sendo, 0.81min no 1/3 folds, 1.62min para o 2/3 folds e 2.41min no 3/3 folds.

A interpretação do resultado final para análise de assertividade é demonstrada na tabela 6 com a utilização do algoritmo de classificação *Random Forest* e por meio de dados estatísticos que definem um melhor desempenho do processo:

Tabela 6. Resultado Final do Modelo

Rede	Matriz de Confusão			Métricas
1		N	F	**Classificador: Random Forest** Accuracy: 70,181% F1-Score: 68.35% Precision: 67.904% Recall: 68.812%
	N	43696	8912	
	F	8543	18849	

Nesta Tabela o código F indica os clientes com irregularidade técnica e não técnica e o código N os clientes normais. Através dela, observa-se também resultados de quatro métricas distintas: acurácia média dos folds, a precisão, F1-*Score e recall*.

A acurácia representa a quantidade de acertos do modelo construído em relação ao total da amostra, ou seja, a proporção de clientes regulares ou fraudadores que foram corretamente classificados. Sejam eles, em clientes ditos como regulares ou fraudulentos, essa métrica leva em consideração todos os acertos. Já a precisão, leva em consideração todos os clientes ditos como fraudadores que de fato estão fraudando, assim como, clientes ditos como regulares que realmente são regulares. O *recall* representa o percentual de dados classificados como fraudadores comparado com a quantidade de fraudadores existente na amostra e o F1-*score*, une precisão e recall afim de trazer um número único que determine a qualidade geral do modelo.

O resultado da acurácia obtida destaca a eficácia do modelo, principalmente quando se compara a taxa de acerto do sistema desenvolvido com o praticado pela distribuidora durante os anos de 2020 e 2021, cujos indicadores apresentaram resultados inferiores a 40%.

Tabela 7. Média de acerto do sistema utilizado na Distribuidora

Acerto (%)	Ano (2020)	Ano (2021)	Média (%)
Regional Leste	21%	31%	24%
Regional Centro Sul	18%	12%	10%
Regional Vale	17%	19%	19%
Regional Baixada	47%	48%	39%
Regional Oeste	38%	42%	37%
Média Regional	26%	38%	26%

A primeira coluna “Acerto (%)” da Tabela 7 representa a subdivisão das cinco regionais pertencentes a distribuidora de energia, com abrangência de atuação por área geográfica, como: Regional Leste, Regional Centro-Sul, Regional Vale, Regional Baixada, e Regional Oeste. Além disso, apresenta também o percentual de acerto da ferramenta atual utilizada durante o ano de 2020 e 2021. A última coluna demonstra a média de acerto obtido por cada regional em ambos períodos.

Quanto a análise de eficiência do modelo obtidos na fase de teste, de forma resumida, nota-se que o F1-*Score* consiste no balanceamento do valor da acurácia em relação ao total de dados de falta do algoritmo, enquanto a precisão contrabalanceia o acerto na classe de interesse em relação as demais amostras, e por fim, para a métrica *recall* os valores estão acima de 60%, o que caracterizam o bom desempenho do modelo.

A interpretação gráfica do resultado final para comparação entre as variâncias médias é demonstrada na Fig. 6

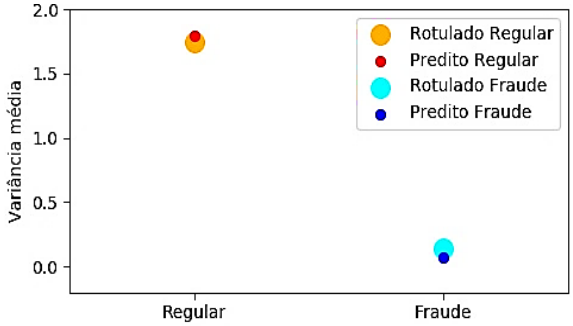


Fig. 6 Interpretação gráfica: Resultado final para comparação entre as variâncias médias

Através desta figura, pode-se notar que as predições do modelo tiveram grande desempenho. Percebe-se que dentre o conjunto de predições rotulados como regulares listados na base do estudo, o modelo foi capaz de apontar origens bem pontuais. Este mesmo comportamento pode ser notado por meio da representação gráfica do conjunto de dados rotulados como fraudadores, com uma leve discrepância. Em termos de resultados, o desejado era que ambos os pontos estivessem coincidindo com os apontamentos observados na base, porém é perceptível a evolução do modelo criado.

6. CONCLUSÃO

Este artigo apresentou a construção de uma ferramenta baseada em *Machine Learning*, cuja finalidade é permitir aos analistas de uma determinada distribuidora de energia elétrica o desenvolvimento de modelos de inteligência para o combate as perdas comerciais. Desta forma, espera-se que o modelo proposto seja capaz de interpretar corretamente os elementos presentes na base de dados da distribuidora e realize indicações de alvos suspeitos de fraudes para atuação, melhorando assim a eficiência do processo de combate e recuperação de energia atualmente praticado.

Mostrou-se e que a capacidade de classificação do algoritmo *Random Forest* foi superior a alcançada quando comparado aos demais algoritmos testados. Vale ressaltar também que

esta comparação foi a mais justa possível levando em consideração o mesmo banco de dados para todos os casos. Além disso, o tempo total gasto durante execução do modelo foi de 2.55min, apresentando assim uma acurácia de 70,181 %, podendo assim concluir que os resultados apresentados demonstram que o modelo proposto é bastante promissor no problema de identificação de irregularidades em baixa tensão.

AGRADECIMENTOS

Agradeço a Deus pela oportunidade concedida e a colaboração do meu orientador Vitor Hugo Ferreira durante o estudo.

7. REFERÊNCIAS

Ascende brasil (2017). *Perdas Comerciais E Inadimplência No Setor Elétrico*. White Paper, Edição nº 18, fev/ 2017. Acesso em 21 abr. 2022.

Angelos, E. W. S., Saavedra, O. R., Cortés, O. A. C., Souza, A. N. (2011). *Detection and Identification of Abnormalities in Customer Consumptions in Power Distribution Systems*. IEEE Transactions on Power Delivery, 26(4): 2436-2442.

Araujo, B., Almeida, H., and Mello, Filho (2019). *Computational Intelligence Methods Applied to the Fraud Detection of Electric Energy Consumers* IEEE Latin America Transactions, vol. 17, no. 1, january 2019.

Ahamad, T., Hongyu, Z., (2022). *Energetics Systems and artificial intelligence: Applications of industry 4.0*. Sciencedirect. Volume 8, November 2022, Pages 334-361.

Buzau, M., Tejedor-Aguilera, J., Cruz-Romero, C., and *Gómez-Expósito, A., Detection of Non-Technical Losses Using Smart Meter Data and Supervised Learning, IEEE Transactions on Smart Grid* (Early Accept), 2018. doi: 10.1109/TSG.2018.2807925.

Cristina, I. (2020). Repórter da Agência Brasil - Rio de Janeiro. Agência Brasil. *Taxa de desemprego passa de 13,3% para 14,6% no terceiro trimestre*.

Cabral, J. E., Gontijo, E. M., Pinto, J. O. P., and Filho, J. R. (2004). *Fraud detection in electrical energy consumers using rough* sets. In: 2004 IEEE International Conference on Systems, Man and Cybernetics, 4: 3625–3629.

Cyro, M., Karla, F., Marley, V., Marco, P. e Gustavo, C. (2008). *Indicações de Suspeitos de Irregularidade em Instalações Elétricas de Baixa Tensão*. Learning and Nonlinear Models - Revista da Sociedade Brasileira de Redes Neurais (SBRN), 6(1): 16-28

Dutra, B (2016). *Furto de energia eleva conta de luz de quem paga em 17%*. Revista Extra: Globo Comunicações. Acesso em 21 abr. 2022.

Faria, R. A. D.; Fonseca, K. V. O.; Schneider, B.; Nguang, S. K. Collusion and Fraud Detection on Electronic Energy Meters - *A Use Case of Forensics Investigation Procedures, 2014 IEEE Security and Privacy Workshops*, San Jose, CA, pp.65-68, 2014. doi: 10.1109/SPW.2014.19

Filho, J. R.; Gontijo, E. M.; Delaiba, A. C.; Mazina, E.; Cabral, J. E.; Pinto, J. O. P.Fraud *identification in electricity company customers using decision tree, 2004 IEEE International Conference on Systems, Man and Cybernetics* (IEEE Cat. No.04CH37583), pp.3730-3734, v.4, 2004. doi: 10.1109/ICSMC.2004.1400924.

Gunturi, S. K., Sarcar, D., (2021). *Ensemble machine learning models for the detection of energy theft*.Electric Power Systems Research. Volume 192, March 2021, 106904. Sciencedirect.

Jang, H.S., Bae, K.Y., Park, H.S., Sung, D.K., 2016. *Solar power prediction based on satellite images and support vector machine.* IEEE Trans. Sustain. Energy 7, 1255–1263. http://dx.doi.org/10.1109/TSTE.2016.2535466.

Jokar, P.; Arianpoo, N.; Leung, V.C.M. *A survey on security issues in smart grids, Secur*. Commun. Netw., v.9, pp.262-273, 2012. doi: 10.1002/sec.559

Light, Histórico e Perfil corporativo, Home. A companhia. Perfil corporativo. 2021. Disponível em http://ri.light.com.br/a-companhia/historico-e-perfil-corporativo/

Muniz, C., M. Vellasco, M., Tanscheit, R., Figueiredo, K. *A Neuro-fuzzy System for Fraud Detection in Electricity Distribution*. IFSAEUSFLAT 2009, 6(3): 1096-1101.

Megahed, T.F., Abdelkader, S.M., Zakaria, A., 2019. *Energy management in zero- energy building using neural network predictive control*. IEEE Internet of Things J. 6, 5336–5344. http://dx.doi.org/10.1109/JIOT.2019.2900558.

Nizar, A. H., Dong, Z. H., Zhao, J. H., e Zhang, P. (2007). *A Data Mining Based NTL Analysis Method*. IEEE Power Engineering Society (PES) General Meeting, 1(4): 1-8

Pollock, D. S. G. (1999). *A Handbook of TimeSeries Analysis, Signal Processing and Dynamics*. Academic Press, New York, San Diego Edition.

Pinto, T., Faia, R., Navarro-Caceres, M., Santos, G., Corchado, J.M., Vale, Z., 2019. *Multi-agent-based CBR recommender system for intelligent energy management in buildings.* IEEE Syst. J. 13, 1084–1095. http://dx.doi.org/10.1109/JSYST.2018.2876933.

Rauber, T.; Drago, I., Varejão, F. e Queiroga, R. (2005) *Extração e Seleção de Características na Identificação de Perdas Comerciais na Distribuição de Energia Elétrica*. XXV Cong. Soc. Bras. Comp.

Rong, J., Tagaris, H., Lachsz, A., and Jeffrey, M. (2002). *Wavelet based feature extraction and multiple classifiers for electricity fraud detection.* In 2002 Trans. and Distribution Conf. and Exhibition 2002: Asia Pacific. IEEE/PES, 3: 2251–2256.