Identificação de Perdas Não Técnicas Através de Método de Classificação e Otimização de Hiperparâmetros, Baseado em Dados Endógenos e Exógenos

Bruno K. Hammerschmitt*. Alzenira R. Abaide*. Marcelo Bruno Capelleti*. Renato G. Negri. Fernando G. K. Guarda***. Lucio R. Prade****. Rafael G. Milbradt*****. Laura L. C. dos Santos******. Nelson Knak Neto******.**

**Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Santa Maria, Santa Maria – RS, Brasil (e-mail: brunokhammer@hotmail.com).*
***Curso de Graduação em Engenharia Elétrica, Universidade Federal de Santa Maria, Santa Maria – RS, Brasil.*
****Colégio Técnico Industrial, Universidade Federal de Santa Maria, Santa Maria – RS, Brasil.*
*****Escola Politécnica, Universidade do Vale dos Sinos, São Leopoldo – RS, Brasil.*
******Colégio Politécnico, Universidade Federal de Santa Maria, Santa Maria – RS, Brasil.*
*******Coordenação Acadêmica, Universidade Federal de Santa Maria, Cachoeira do Sul – RS, Brasil.*

Abstract: Non-Technical Losses (NTLs) are problems commonly found in electric power systems distribution, caused by theft or fraud of energy. These problems result in financial losses for distribution utilities as well as consumers, who partially bear the costs involved in the NTLs. In view of this, practices for the identification of consumer units that are committing some type of irregularity in their facilities must be applied. In this scenario data classification models emerge, which based on supervised learning based on endogenous and exogenous historical data, are able to interpret information and label them. One of these models is the Decision Tree, which associated with Machine Learning (ML) techniques for the optimization of hyperparameters can obtain results with high precision in the outliers identification. Thus, this study aims to implement the Decision Tree model to identify consumers with NTLs, and to propose the optimization of Decision Tree hyperparameters following three ML techniques, Bayes Search, Grid Search and Randomized Search. Finally, the results are discussed and analyzed, and considerations are performed on the models.

Resumo: As Perdas Não Técnicas (PNTs) são problemas comumente encontrados nos sistemas de distribuição de energia elétrica, ocasionados pelo furto ou fraude de energia. Estes problemas acarretam em prejuízos financeiros as concessionárias de distribuição, assim como aos consumidores, que arcam parcialmente com os custos envolvidos às PNTs. Diante disto, medidas para a identificação de unidades consumidoras que estejam cometendo algum tipo de irregularidade nas suas instalações devem ser aplicadas. Neste panorama surgem os modelos de classificação de dados, que a partir da aprendizagem supervisionada a partir de dados históricos de origem endógena e exógena consegue interpretar as informações e rotulá-las. Um desses modelos é a Árvore de Decisão, que associada a técnicas de *Machine Learning* (ML) para a otimização de hiperparâmetros consegue obter resultados com alta precisão na identificação de *outliers*. Assim, este estudo tem por objetivo a implementação do modelo Árvore de Decisão para a identificação de consumidores com PNTs, e a proposição da otimização de hiperparâmetros da Árvore de Decisão seguindo por três técnicas de ML, *Bayes Search*, *Grid Search* e *Randomized Search*. Por fim, são discutidos os resultados e realizadas análises e ponderações sobre os modelos.

Keywords: Non-Technical Losses; Outliers Identification; Decision Tree Method; Hyperparameter Optimization; Exogenous Data; Machine Learning.

Palavras-chaves: Perdas Não Técnicas; Identificação de *Outliers*; Método de Árvore de Decisão; Otimização de Hiperparâmetros; Dados Exógenos; *Machine Learning.*

1. INTRODUÇÃO

As perdas em Sistemas Elétrico de Potência são divididas em duas categorias, Perdas Técnicas (PTs) e Perdas Não Técnicas (PNTs) (ou perdas comerciais). As PTs acontecem devido ao efeito térmico causado pela corrente elétrica que atravessa condutores e equipamentos do sistema. Já as PNTs, são ocasionadas na apuração de energia elétrica consumida pelas Unidades Consumidoras (UCs) mas não contabilizadas pelos medidores de energia.

Segundo a Agência Nacional de Energia Elétrica (ANEEL), PNTs podem ser ocasionadas por erros de instalação, erros de medições, furtos de energia, que são relacionados a ligações clandestinas ou desvios da rede, e fraudes de energia

vinculadas a adulteração nos medidores ou desvios (ANEEL, 2021). As PNTs representam um sério problema para o sistema elétrico, ocasionando prejuízos financeiros para as concessionárias de distribuição de energia elétrica e para os consumidores. Essas perdas dificultam a operação do sistema provocando diversas consequências como: alterações na estabilidade da rede, redução de confiabilidade, variação de tensão em regime, desequilíbrio de tensão do sistema, podendo resultar em carga desequilibrada, assim como perdas devido à sobrecarga e distorções de harmônicas (P. J. Navani, 2009; Glauner *et al.*, 2017).

PNTs são problemas intrínsecos da condução de energia na transmissão e distribuição, traduzida pelas adversidades enfrentadas pelo setor eletroenergético. Essas perdas são mais evidentes em países com economia emergente, onde os maiores índices de PNTs são identificados, como por exemplo o Brasil e a Índia. Em todo o mundo, é estimado que as PNTs sejam responsáveis por prejuízos na ordem de 100 bilhões de dólares (Glauner *et al.*, 2018).

No Brasil, as perdas no sistema de distribuição de energia para o ano de 2020, que são obtidas pelas diferenças entre a energia adquirida pelas concessionárias de distribuição e a energia elétrica efetivamente faturada aos consumidores, foram de 15,1% do total de energia injetada na rede de distribuição de energia. Destes, 7,5% foram relacionados a PNTs, sendo que equivalem a cerca de 37,9 TWh. Realizando os cálculos relativos às PNTs, utilizando o preço médio de energia nos processos tarifários e sem considerar tributos, representaram um prejuízo por volta de R$ 8,6 bilhões, valor este que é repassado parcial ou totalmente aos consumidores finais, e que representa cerca de 2,9% do valor da tarifa de energia elétrica (ANEEL, 2021).

Visto que os orçamentos para investimentos do setor são bilionários e que as PNTs acompanham o crescimento da carga do sistema, constata-se a importância de projetos de mitigação de perdas no sistema elétrico. Assim, com objetivo de otimizar o uso da energia e diminuir os investimentos futuros, os quais são necessários para suprir o crescimento da demanda de energia elétrica no Brasil, o emprego de modelos computacionais para a identificação das PNTs são necessários.

Neste cenário, os sistemas de distribuição de energia elétrica no desenvolvimento de redes inteligentes são considerados como uma das importantes áreas de aplicação de Big Data (Lv *et al.*, 2019). A utilização de dados endógenos (dados históricos das unidades consumidores e de posse das distribuidoras de energia elétrica) e dados exógenos (dados de outras fontes, mas que possuem estreita relação com as PNTs) são vistas com potencial e grande valor agregado para identificação de PNTs (Capeletti *et al.*, 2021). Tais dados podem servir de insumos de entrada para modelos de *Machine Learning* (ML) ou aprendizado de máquinas, uma vez que, esses dados utilizados em conjunto podem fornecer padrões importantes para identificação de PNTs. Manualmente esses padrões são de difícil identificação, sendo necessário aplicações de métodos inteligentes para tal função, o que permite indicar consumidores com prováveis irregularidades de energia elétrica.

Diante disto, esse trabalho propõe um método de identificação de PNTs a partir do uso do método de classificação Árvore de Decisão e processos de otimização. O método foi empregado seguindo a técnica orientada a dados e com aprendizagem supervisionada de classificação, utilizando técnicas Inteligência Artificial (IA), especificamente no que compreende a ML para a otimização de hiperparâmetros do modelo. Para isso, contou com um grande banco de dados de uma concessionária de distribuição de energia elétrica do Brasil, Companhia Estadual de Energia Elétrica (CEEE) e Grupo Equatorial Energia, e de dados exógenos ao sistema, dados estes de origem em dados meteorológicos. Deste modo, com a otimização de hiperparâmetros por técnicas de ML, há melhoria de desempenho na identificação UCs que estejam cometendo fraude ou furto de energia.

2. CARACTERIZAÇÃO E PREMISSAS METODOLÓGICAS

A identificação de PNTs pode ser classificada como uma detecção de *outliers* (Hammerschmitt *et al.*, 2020). Esta denominação é representada pelo banco de dados ser extremamente desbalanceado. Dado o problema de identificação de PNTs em sistemas de distribuição de energia, existe um grande desafio no treinamento dos modelos de ML, devidos aos dados serem desbalanceados, fato que pode ocasionar má generalização dos algoritmos na aprendizagem. Este problema deve ser ponderado, sendo necessário o aperfeiçoamento do algoritmo de aprendizado, bem como sua otimização para obtenção de resultados assertivos sem distorções, não entrando nas áreas ineficazes anteriormente citadas (Li, Yan e Xu, 2021).

Na literatura existe uma grande quantidade de artigos que apresentam métodos para atenuar os efeitos do banco de dados desbalanceado. Na fase de aprendizagem do algoritmo de classificação é indispensável e necessário ter representatividade de ambos os casos de consumidores. Um dos exemplos de métodos é a sobreamostragem ou a subamostragem de dados, esses métodos de amostragem são revisados em Ghori et al. (2021). Neste estudo é realizada a comparação entre diversas técnicas de sobreamostragem e subamostragem, considerando o *recall* como uma métrica de desempenho, concluindo que subamostragem dos dados é a melhor técnica de amostragem (Ghori *et al.*, 2021).

Além da amostragem dos dados, a otimização de hiperparâmetros e/ou parâmetros dos modelos, a exemplo da Árvore de Decisão, desempenha um papel importante na classificação de UCs fraudulentas. Levando em consideração que a detecção de PNTs em modelos de ML é um exemplo de detecção em banco de dados desbalanceado, com a otimização de hiperparâmetros é possível atingir resultados significativamente melhores do que modelos com hiperparâmetros padrões (Hancock e Khoshgoftaar, 2021).

Em Saeed et al. (2020), os autores aplicaram o algoritmo *Boosted C5.0 Decision Tree* para classificar consumidores honestos e consumidores fraudulentos, alcançando resultados

superiores a outros modelos como *Random Forest*, *XGBoost Tree* e Redes Neurais (Saeed *et al.*, 2020). Assim, a otimização de hiperparâmetros do modelo, com objetivo de aprimorar a precisão de classificação, é uma das mais importantes tarefas na redução do erro na predição com isso aprimorando a assertividade (Oo e Thein, 2019).

Os hiperparâmetros são variáveis do algoritmo definidas antes do treinamento que representam características construtivas, o que difere dos parâmetros, que são ajustados durante o treinamento dos modelos. As técnicas de otimização de hiperparâmetros mencionadas abaixo foram empregadas neste estudo no método de Árvore de Decisão.

- *Bayes Search* (busca bayesiana): algoritimo de busca que tenta estimar qual é a combinação de hiperparâmetros que resultará na maior performance, com base numa distribuição criada a partir das combinações testadas anteriormente (Kumar *et al.*, 2017).
- *Grid Search* (busca em grade): algoritmo de busca que recebe um conjunto de valores de um ou mais hiperparâmetros e testa todas as combinações dentro dessa vizinhança (Pedregosa *et al.*, 2011a).
- *Randomized Search* (busca aleatória): algoritimo de busca que testa combinações aleatórias de hiperparâmetros (Pedregosa *et al.*, 2011b).

3. METODOLOGIA

O modelo de detecção de PNTs associado a dados endógenos e exógenos foi desenvolvida em linguagem de programação Python e implementado no software *Spyder* 3.7 do pacote ANACONDA®, em um notebook com processador Intel Core i5 de 2.5 GHz, 8 GB de memória RAM, DDR 4 de 2133 MHz, placa de vídeo GeForce MX 930, e um sistema operacional Microsoft Windows 10, por meio do uso das bibliotecas *Pandas*, *Sckit-learn*, *Sckit-optimize*, *Numpy* e *SciPy*. Para isso, foi necessária uma base dados endógenos fornecida pela concessionária de distribuição e dados exógenos, que são informações externas a concessionária de distribuição, mas com determinado grau de relação com o consumo de eletricidade e atividade suspeitas.

Para implementação do modelo foi necessário o pré-tratamento do grande volume de informações, e posterior utilização nos modelos de classificação. Deste modo, foi utilizada a base de dados de consumo e informações da UC, como: a carga instalada, localização, código da fase de atendimento da UC, e inspeções de campo. O intervalo de tempo utilizado foi de janeiro de 2015 a dezembro de 2019, que conta com 40.000 UCs da região em estudo. Além disso, também foi utilizado a base dados de temperatura ambiente média mensal da região em análise.

O método de classificação utilizado foi a Árvore de Decisão, onde foram aplicados três técnicas de otimização de hiperparâmetros, sendo elas: *Bayes Search*, *Grid Search,* e *Randomized Search*. A mesma base de dados foi utilizada para treinamento e teste dos modelos, onde foram utilizados dados de 2.000 UCs para treinamento e 38.000 UCs para teste. Os dados de treinamento contêm 50% das UCs com resultado de inspeção com irregularidades (apresentaram fraude ou furto de energia), e 50% sem irregularidades nas suas instalações. Os dados de teste contêm 1.000 UCs com irregularidades, e o restante dos dados sem irregularidades, constatadas durante a inspeções realizadas no intervalo de tempo mencionado anteriormente. A base de dados foi selecionada empiricamente, procedendo por amostras balanceadas para treinamento e amostras desbalanceadas para teste. Desta forma, será apresentado uma breve descrição do tratamento dos dados assim como os resultados obtidos pelos modelos.

3.1 Pré-Tratamento dos dados

Os parâmetros que necessitam de pré-tratamento serão descritos a seguir, com suas respectivas definições, ajustes e equacionamentos para possível reprodução.

3.1.1 Organização dos dados de consumo e temperatura

A compatibilização temporal dos dados de consumo e temperatura foi realizada para adequá-los para posterior análise, seguindo primeiramente pelos dados de consumo obtidos do histórico de dados da concessionária de distribuição, e em seguida pelos dados de temperatura ambiente que foram obtidos da estação meteorológica localizada na região em estudo, informações estas que podem ser encontradas no site do INMET (INMET, 2021).

- *Dados de consumo*: Os dados de consumo foram obtidos a partir da coleta mensal de dados de medidores, sendo a unidade de medida o kWh. Foram considerados para este estudo apenas consumidores com consumo médio mensal abaixo de 500 kWh. A série histórica destes dados correspondem a 60 leituras de consumo, com data inicial em janeiro de 2015 e final em dezembro de 2019.
- *Dados de temperatura*: os dados de temperatura ambiente são obtidos na unidade de graus Celsius (ºC), com resolução temporal de hora em hora, com data e hora inicial as 00 hora de 1º de janeiro de 2015 e final as 23 horas de 31 de dezembro de 2019. Para adequar os dados de temperatura, foi realizada a média mensal da série histórica, que segue por (1).

$$T = \frac{1}{n}\sum_{k=1}^{n} y_k \quad (1)$$

Onde: T é a média mensal de temperatura em ºC, y_k é a temperatura horária medida pela estação meteorológica, e n o número total de medidas de temperatura em cada mês.

3.1.2 Correlação de Pearson

A correlação de Pearson (r) corresponde a um valor absoluto, situado entre +1 e -1, que reflete a intensidade de dois conjuntos de dados, que podem ser classificados em correlação positiva (com intervalo de $0 < r \leq 1$) e correlação

negativa (com intervalo de -1 ≥ r > 0). A correlação positiva sinaliza que os dados são diretamente proporcionais, e a correlação negativa denota que as variáveis são correlacionadas, mas inversamente proporcionais. Se a correlação de Pearson for 0, não há relação entre os conjuntos de dados (Jawad *et al.*, 2020). A correlação de Pearson é descrito por (2), a qual possui como resultado o grau de relação entre as variáveis de consumo e temperatura.

$$r = \frac{\sum_{k=1}^{m}(x_{ik} - \bar{x})(y_{ik} - \bar{y})}{\sqrt{\sum_{k=1}^{m}(x_{jk} - \bar{x})^2} \cdot \sqrt{\sum_{k=1}^{m}(y_{jk} - \bar{y})^2}} \quad (2)$$

Onde: r é o valor absoluto da correlação de Pearson, x_{ik} os dados de consumo da UC em análise, $\bar{x}$ a média aritmética dos dados de consumo da UC, y_{ik} a temperatura média mensal obtida por (1), e $\bar{y}$ a média aritmética dos dados de temperatura.

3.1.3 Média, mediana e desvio padrão

Os resultados obtidos para a média, mediana e desvio padrão seguem como um balizador de como a unidade consumidora vem se portando, visto que são os índices mais comuns empregados na análise de banco de dados e que vem a auxiliar na detecção PNTs nas UCs. O cálculo da média aritmética de consumo ($\bar{x}$) segue por (3), com x_i sendo as amostras de consumo identificadas para a UC, e n o número total de amostras da UC.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad (3)$$

A mediana (*Md*) é identificada pelo valor central do conjunto de amostras, no caso em específico, ela denota o ponto médio das amostras de consumo da UC. Para o caso de o total de amostras ser um número ímpar, haverá apenas um valor central na série amostral. Caso o número total de amostras for um número par, (4) deverá ser utilizada para obter a *Md*, que consiste basicamente em realizar a média aritmética dos dois valores centrais do conjunto amostral, denominados por x_{Md1} e x_{Md2}.

$$Md = \frac{x_{Md1} + x_{Md2}}{2} \quad (4)$$

Já o desvio padrão (*Dp*) tem a função de informar o grau de dispersão do conjunto amostral da série histórica, pois quanto menor o resultado de *Dp* mais homogênea é a série histórica em análise, e que é obtida por (5):

$$Dp = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}} \quad (5)$$

Onde: x_i são as informações de consumo da UC, $\bar{x}$ é a média aritmética de consumo da UC, e n o número de total de amostras desta mesma UC.

3.1.4 Processo para obter os resultados de inspeções com fraude ou furto de energia

Com objetivo de empregar a aprendizagem supervisionada nos modelos de classificação, é necessário categorizar todas as UCs. Com esse objetivo, foi realizada uma varredura em todas as inspeções realizada nos anos que compreendem o período dos dados deste estudo. Diante disto, as UCs em que foi encontrada a fraude ou furto de energia pela inspeção em campo foi categorizada como consumidor irregular e rotulada com valor "1". Já para as UCs em que a inspeção retornou uma conclusão negativa, ou seja, consumidores em que a fraude ou furto de energia não foi detectada, assim como aos consumidores que durante o período não tiveram inspeções, foram categorizadas como UCs regulares e rotuladas com o valor "0". Essa informação é importante para a aprendizagem dos modelos, pois servirá de base para que os modelos consigam incorporar as informações que são disponibilizadas durante a fase de treinamento, interpretando melhor os padrões de dados para cada rótulo, e assim, obter uma maior assertividade durante a fase de testes.

3.2 Matriz de Confusão

Para avaliar a assertividade dos modelos de classificação binária, pode-se ter quatro avaliações de saída conforme em (Castro e Ferrari, 2016), sendo retratadas pela Matriz de Confusão ilustrada pela Tabela 1, onde, "1" é retratado como positivo e "0" como negativo, e que recebem as seguintes classificações: Verdadeiros Negativos (VN), originalmente negativas com saídas negativas (rotulado por "0" e classificado como "0"); Falsos Positivos (FP), originalmente negativas com saídas positivas (rotulado por "0" e classificado como "1"); Falsos Negativos (FN), originalmente positivas com saídas negativas (rotulado por "1" e classificado como "0"); e Verdadeiros Positivos (VP), originalmente positivas com saídas positivas (rotulado por "1" e classificado como "1").

Tabela 1. Matriz de Confusão

		Valor Predito	
		0	**1**
Valor Real	**0**	VN	FP
	1	FN	VP

Na literatura a matriz de confusão é utilizada para analisar estatisticamente o resultado dos classificadores (Dian-Gang *et al.*, 2018). A partir da matriz de confusão, pode-se definir métricas estatísticas para avaliação da assertividade dos modelos, sendo uma dessas métricas a precisão global, também denominada acurácia, definida em (6).

$$Precisão\ global = \frac{VN + VP}{VP + VN + FP + FN} \quad (6)$$

No caso em análise, as UCs identificadas anteriormente com irregularidades, aqui denominadas por UCs com fraude ou furto de energia, receberão o valor "1", e as UCs classificadas sem fraude ou furto receberão o valor "0".

Outros parâmetros quantificados pela matriz de confusão são a taxa de acerto sobre os dados reais obtida por (7), também conhecida por *recall*, e a precisão do modelo de classificação na predição de "0" e "1", expressa por (8).

$$Taxa\ de\ acerto\ (recall) = \frac{VN}{VN + FP}\ ou\ \frac{VP}{VP + FN} \quad (7)$$

$$Precisão = \frac{VN}{VN + FN}\ ou\ \frac{VP}{VP + FP} \quad (8)$$

3.3 Árvore de Decisão e Otimização dos Hiperparâmetros

A Árvore de Decisão é um método que pode ser utilizado para classificação, caracterizado como um algoritmo não paramétrico com objetivo de determinar um atributo alvo a partir dos atributos previsores. Trata-se de um método baseado em procura a partir de um espaço de possíveis hipóteses, criando estruturas simbólicas que sejam compreensíveis por humanos. Tem como características os chamados nodos, a árvore terá um nodo raiz como ponto de partida, o nodo raiz terá seus filhos, e esses também gerarão filhos, e assim sucessivamente, e por fim chegará ao nodo folha. Com isso, a Árvore de Decisão é capaz de armazenar as regras de classificação através de seus nodos, e o nodo final representa a decisão tomada (Lenz *et al.*, 2020).

Para este estudo foi utilizado o cálculo da entropia, conforme exposto por Lenz et al. (2020). Os ajustes de hiperparâmetros dos modelos seguiram pela definição de critérios pré-estabelecidos de maneira empírica pelos autores, mas coincidentes entre todos, onde, os critérios foram: a estratégia de divisão de validação cruzada, igual a 10; o número de configurações de parâmetros que são amostradas, igual a 10; e o espaço de pesquisa sobre os parâmetros do estimador fornecido, que variam de 1 a 100.

4. RESULTADOS

Os resultados deste estudo serão apresentados seguindo a Árvore de Decisão padrão (sem otimização), Árvore de Decisão com a otimização de hiperparâmetros seguindo respectivamente pelas técnicas de *Bayes Search*, *Grid Search* e *Randomized Search*. Para cada uma das aplicações serão apresentados os resultados da matriz de confusão, a precisão global, o tempo de processamento do modelo em análise, e o número de UCs que seriam listadas para inspeção em situação real. Por fim, serão discutidos os resultados já mencionados e sinalizados os resultados de maior granularidade, como a taxa de acerto e a precisão da predição para identificação das UCs fraudadoras ou não, onde será ponderado sobre os modelos que obtiveram melhores resultados.

4.1 Árvore de Decisão

Os resultados para a Árvore de Decisão padrão, sem otimização de seus parâmetros, podem ser vistos na matriz de confusão ilustrada pela Tabela 2. A precisão global do modelo foi de 60,59% com um tempo de processamento de 4,28 segundos.

Tabela 2. Matriz de confusão dos resultados da Árvore de Decisão.

		Valor Predito	
		0	**1**
Valor Real	**0**	22.370	14.630
	1	344	656

Conforme a Tabela 2, foram classificadas 22.370 UCs como VN, 14.630 UCs como FP, 344 UCs como FN, e 656 UCs como VP. Portanto, a lista de UCs a serem inspecionadas seria composta por FP e VP, totalizando 15.286 UCs.

4.2 Árvore de Decisão com Otimização de Hiperparâmetros por Bayes Search

Os resultados para a Árvore de Decisão com a otimização de hiperparâmetros pela técnica de *Bayes Search*, podem ser observados pela matriz de confusão demostrada pela Tabela 3. A precisão global do modelo foi de 80,83% com um tempo de processamento de 6,30 segundos.

Tabela 3. Matriz de confusão da Árvore de Decisão com otimização de hiperparâmetros por *Bayes Search*.

		Valor Predito	
		0	**1**
Valor Real	**0**	30.101	6.899
	1	385	615

Como demostrado na Tabela 3, foram classificadas 30.101 UCs como VN, 6.899 UCs como FP, 385 UCs como FN, e 615 UCs como VP. Assim, a lista de UCs a serem inspecionadas seria composta por FP e VP, totalizando 7.514 UCs.

4.3 Árvore de Decisão com Otimização de Hiperparâmetros por Grid Search

Já os resultados para a Árvore de Decisão com a otimização de hiperparâmetros pela técnica de *Grid Search*, são observadas pela matriz de confusão expressa pela Tabela 4. A precisão global do modelo foi de 80,82% com um tempo de processamento de 20,87 segundo.

Tabela 4. Matriz de confusão da Árvore de Decisão com otimização de hiperparâmetros por *Grid Search*.

		Valor Predito	
		0	**1**
Valor Real	**0**	30.098	6.902
	1	385	615

De acordo com a Tabela 4, foram classificadas 30.098 UCs como VN, 6.902 UCs como FP, 385 UCs como FN, e 615 UCs como VP. Deste modo, a lista de UCs a serem inspecionadas seria composta por FP e VP, totalizando 7.517 UCs.

4.4 Árvore de Decisão com Otimização de Hiperparâmetros por Randomized Search

Por fim, os resultados para a Árvore de Decisão com a otimização de hiperparâmetros pela técnica de *Randomized Search*, são exibidas pela matriz de confusão da Tabela 5. A precisão global do modelo foi de 80,82% com um tempo de processamento de 6,13 segundos.

Tabela 5. Matriz de confusão da Árvore de Decisão com otimização de hiperparâmetros por *Randomized Search*.

		Valor Predito	
		0	**1**
Valor Real	**0**	30.098	6.902
	1	385	615

Com base na Tabela 5, foram classificadas 30.098 UCs como VN, 6.902 UCs como FP, 385 UCs como FN, e 615 UCs como VP. Mediante a isto, a lista de UCs a serem inspecionadas seria composta por FP e VP, totalizando 7.517 UCs.

5. DICUSSÕES

Como pode ser observado pela Tabela 6, todas as aplicações que envolveram técnicas de otimização de hiperparâmetros dentro da Árvore de Decisão obtiveram melhores resultados que o modelo na sua forma padrão, e com redução em torno de 50% do número UCs a serem inspecionadas.

A Árvore de Decisão com a otimização de hiperparâmetros pela técnica de *Bayes Search* obteve melhor resultado no quesito precisão global, contudo, os resultados das outras duas técnicas de otimização, que tiveram os mesmos resultados de classificação, não tiveram significativa diferença sob o melhor resultado, que foi a diferença de 80,83% para 80,82%, respectivamente. Em se tratando do tempo de processamento, a Árvore de Decisão sem otimização aparece com o menor tempo de classificação dos dados de UCs que cometeram ou não fraude ou furto de energia, com o tempo de 4,28 segundos, seguido pelo *Randomized Search* com um tempo de 6,13 segundos, *Bayes Search* com tempo de 6,30 segundos, e por último o *Grid Search* com tempo de processamento de 20,87 segundos.

Quanto aos resultados para a taxa de acerto, relativa à identificação de UCs sem fraude ou furto (“0”), e UCs com fraude ou furto de energia (“1”), foram os mesmos resultados para todas as três técnicas, que foram 81,35% para a identificação de “0”, e “61,50%” para a identificação de “1”. Uma ressalva a ser feita é que a Árvore de Decisão sem otimização obteve uma taxa de acerto para identificação de UCs que cometeram fraude ou furto, sinalizada “1”, superior aos demais modelos, que foi de 65,60%.

Finalmente, a precisão dos três modelos de otimização na predição de UCs com e sem fraude ou furto de energia foi a mesma para todos, na qual os resultados foram de 98,74% para “0”, e 8,18% para “1”, sendo estes resultados superiores aos resultados obtidos pela Árvore de Decisão sem otimização. Todavia, os resultados aparentemente se mostram baixos em se tratando da precisão na identificação de UCs com PNTs, ocasionado pela discrepância entre o número de UCs sem fraude ou furto, onde, para os testes deste estudo foram de 37.000 UCs contra 1.000 UCs que cometeram fraude ou furto de energia comprovados pelas inspeções de campo, o que é normal para um banco de dados real em se tratando de UCs com e sem PNTs.

Tabela 6. Resultados da Árvore de Decisão e das técnicas de otimização de hiperparâmetros.

	Precisão Global (%)	**Tempo de processamento (segundos)**	**Taxa de acerto (%)**		**Precisão (%)**	
			0	1	0	1
Árvore de Decisão	60,59%	4,28	60,46%	65,60%	98,49%	4,29%
Árvore de Decisão - *Bayes Search*	80,83%	6,30	81,35%	61,50%	98,74%	8,18%
Árvore de Decisão - *Grid Search*	80,82%	20,87	81,35%	61,50%	98,74%	8,18%
Árvore de Decisão - *Randomized Search*	80,82%	6,13	81,35%	61,50%	98,74%	8,18%

6. CONCLUSÕES

Mediante aos resultados obtidos pelo modelo de classificação da Árvore de Decisão padrão, e com a aplicação das técnicas de ML para a otimização de hiperparâmetros, foi identificado que todas as aplicações que envolveram o processo de otimização obtiveram melhores resultados para precisão global, porém, com tempo de processamento mais elevado que a aplicação simples. Além disso, houve redução do número de UCs listadas para inspeção, este que resultaria na redução de custos de deslocamento de equipe para inspeção de campo. Ademais, a Árvore de Decisão com otimização de hiperparâmetros pela técnica *Bayes Search* e a pela técnica *Randomized Search*, foram os modelos que obtiveram melhores resultados em geral, considerando todos os parâmetros analisados.

Vale destacar que a aplicação da técnica de *Grid Search* teve os mesmos resultados da técnica *Randomized Search*, para a precisão global, taxa de acerto e precisão na predição de classificação das UCs em "0" e "1". Entretanto, o tempo de processamento da técnica *Grid Search* foi superior a 3 vezes aos outros modelos com otimização de hiperparâmetros, e cerca de 5 vezes sobre a Árvore de Decisão padrão. Embora os tempos de processamento estejam na faixa de poucos segundos devido a pequena quantidade de amostras de dados, em um banco de dados com maior número de UCs os tempos de processamento destes modelos devem apresentar maiores diferenças ao serem comparados.

Ressalta-se ainda que os resultados da precisão de predição dos modelos na identificação de UCs com PNTs foram baixos, mas justificados pelo banco de dados ser desbalanceado, e quando analisada a taxa de acerto dos modelos, fica evidente que os modelos obtiveram resultados satisfatórios para a classificação de UCs com e sem fraude ou furto de energia. Por fim, a utilização de dados endógenos associados a dados exógenos, no caso da temperatura, são poucos explorados na literatura, fato que motiva o desenvolvimento de estudos envolvendo a associações deste e outros dados exógenos com relação as PNTs.

AGRADECIMENTOS

Os autores agradecem ao apoio técnico e financeiro da Companhia Estadual de Energia Elétrica e do Grupo Equatorial Energia (Programa de P&D da ANEEL através do projeto CEEE/EQUATORIAL/UFSM nº 5000003849), Instituto Nacional de Ciência e Tecnologia em Sistemas de Geração Distribuída (INCTGD), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq - nº 465640/2014-1), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES - nº 23038.000776/2017-54), Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS - nº 17/2551-0000517-1) e Universidade Federal de Santa Maria (UFSM), Instituições Brasileiras.

REFERÊNCIAS

ANEEL (2021) *Perdas de Energia Elétrica na Distribuição*, *Agência Nacional de Energia Elétrica, Brasil*. Available at: https://www.aneel.gov.br/documents/654800/18766993/Relatório+Perdas+de+Energia_+Edição+1-2021.pdf/143904c4-3e1d-a4d6-c6f0-94af77bac02a#:~:text=Essas perdas%2C inevitáveis em qualquer,R%24 1%2C8 bilhão.&text=de 2021. (Accessed: 21 January 2022).

Capeletti, M. B. *et al.* (2021) 'Descriptive Data Analysis of Weather Inputs for Non-Technical Losses Detection System', in *2021 9th International Conference on Modern Power Systems (MPS)*. IEEE, pp. 1–5. doi: 10.1109/MPS52805.2021.9492593.

Castro, L. N. de and Ferrari, D. G. (2016) *Introdução à Mineração de Dados: Conceitos Básicos, Algoritmos e Aplicações*. 1ª. Brasil: Editora Saraiva. Available at: https://integrada.minhabiblioteca.com.br/reader/books/978-85-472-0100-5.

Dian-Gang, W. *et al.* (2018) 'Anomaly Behavior Detection Based on Ensemble Decision Tree in Power Distribution Network', in *2018 4th Annual International Conference on Network and Information Systems for Computers (ICNISC)*. IEEE, pp. 312–316. doi: 10.1109/ICNISC.2018.00069.

Ghori, K. M. *et al.* (2021) 'Treating Class Imbalance in Non-Technical Loss Detection: An Exploratory Analysis of a Real Dataset', *IEEE Access*, 9, pp. 98928–98938. doi: 10.1109/ACCESS.2021.3095145.

Glauner, P. *et al.* (2017) 'The Challenge of Non-Technical Loss Detection Using Artificial Intelligence: A Survey', *International Journal of Computational Intelligence Systems*, 10(1), pp. 760–775. doi: 10.2991/ijcis.2017.10.1.51.

Glauner, P. *et al.* (2018) *Non-Technical Losses in the 21st Century: Causes, Economic Effects, Detection and Perspectives*. Available at: https://www.researchgate.net/publication/325297875 (Accessed: 15 January 2022).

Hammerschmitt, B. K. *et al.* (2020) 'Non-Technical Losses Review and Possible Methodology Solutions', in *2020 6th International Conference on Electric Power and Energy Conversion Systems (EPECS)*. IEEE, pp. 64–68. doi: 10.1109/EPECS48981.2020.9304525.

Hancock, J. and Khoshgoftaar, T. M. (2021) 'Impact of Hyperparameter Tuning in Classifying Highly Imbalanced Big Data', in *2021 IEEE 22nd International Conference on Information Reuse and Integration for Data Science (IRI)*. IEEE, pp. 348–354. doi: 10.1109/IRI51335.2021.00054.

INMET (2021) *Instituto Nacional de Meteorologia*, *Brasil*. Available at: https://portal.inmet.gov.br/ (Accessed: 22 July 2021).

Jawad, M. *et al.* (2020) 'Machine Learning Based Cost Effective Electricity Load Forecasting Model Using Correlated Meteorological Parameters', *IEEE Access*, 8, pp. 146847–146864. doi: 10.1109/ACCESS.2020.3014086.

Kumar, M. *et al.* (2017) *Scikit-Optimize: Machine Learning*

in Python - Bayes Search. Available at: https://scikit-optimize.github.io/stable/modules/generated/skopt.BayesSearchCV.html#skopt.BayesSearchCV (Accessed: 21 January 2022).

Lenz, M. L. *et al.* (2020) *Fundamentos de Aprendizagem de Máquina*. Porto Alegre: SAGAH. Available at: https://integrada.minhabiblioteca.com.br/reader/books/9786556900902 (Accessed: 21 January 2022).

Li, Q., Yan, M. and Xu, J. (2021) 'Optimizing Convolutional Neural Network Performance by Mitigating Underfitting and Overfitting', in *2021 IEEE/ACIS 19th International Conference on Computer and Information Science (ICIS)*. IEEE, pp. 126–131. doi: 10.1109/ICIS51600.2021.9516868.

Lv, Z. *et al.* (2019) 'A Data Fusion and Data Cleaning System for Smart Grids Big Data', in *2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*. IEEE, pp. 802–807. doi: 10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00119.

Oo, M. C. M. and Thein, T. (2019) 'Hyperparameters optimization in scalable random forest for big data analytics', *2019 IEEE 4th International Conference on Computer and Communication Systems, ICCCS 2019*, pp. 125–129. doi: 10.1109/CCOMS.2019.8821752.

P. J. Navani (2009) 'Technical and Non-Technical Losses in Power System and its Economic Consequences in Indian Economy', *International Journal of Electronics and Computer Science Engineering (IJESCE)*, Volume 1(ISSN: 2277-1956),. No.2), pp. 757–761.

Pedregosa, F. *et al.* (2011a) *Scikit-learn: Machine Learning in Python - Grid Search*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html (Accessed: 21 January 2022).

Pedregosa, F. *et al.* (2011b) *Scikit-learn: Machine Learning in Python - Randomized Search*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html (Accessed: 21 January 2022).

Saeed, M. S. *et al.* (2020) 'An efficient boosted C5.0 decision-tree-based classification approach for detecting non-technical losses in power utilities', *Energies*, 13(12), pp. 1–19. doi: 10.3390/en13123242.