Classificação de Unidades Consumidoras Irrigantes de Arroz Para Análise de Perdas Não Técnicas Utilizando o Método de *Random Forest*

[1]Henrique S. Eichkoff, [1]Daniel P. Bernardon, [1]Natália B. de Souza, [1]Pedro Marcolin, [1]Júlia Madaloz,
[2]Luciana M. Kopp, [3]Lucas M. Chiara, [3]Juliano A. Silva
[1]Universidade Federal de Santa Maria, [2]Universidade Federal de Pelotas, [3]CPFL Energia
henriquekoff@gmail.com, dpbernardon@ufsm.br, rhyo.natalia@gmail.com, pedro_marcolin@hotmail.com,
juliacmadaloz@gmail.com, lucianakopp@gmail.com, lucaschiara@cpfl.com.br, julianoandrade@cpfl.com.br

***Resumo*—A detecção de irregularidades no consumo de energia elétrica, para as empresas do setor elétrico, ainda é um grande desafio. As perdas comerciais, ou perdas não-técnicas, têm impacto negativo na receita das concessionárias de energia elétrica, além de impactar no bolso do consumidor. Neste intuito, este trabalho tem como objetivo utilizar técnicas de aprendizado de máquina para fins de classificação, utilizando o método de *Random Forest*, e do tipo não supervisionado (*Clustering*) para criar grupos de consumidores que serão utilizados para a predição do modelo classificador. A representação e clusterização do *dataset* foi desenvolvida com os valores da média e desvio padrão dos consumos dos anos 2017, 2018 e 2019 onde, para a obtenção dos resultados da metodologia proposta pelo trabalho, foram utilizados como dados de entrada os consumos reais do ano de 2020 das unidades consumidoras localizados no município de Uruguaiana, no estado do Rio Grande do Sul, do banco de dados da CPFL Energia.**

***Palavras-chave* – Clusterização; Aprendizado de Máquina; Perdas Não Técnicas; Random Forest**

I. Introdução

As perdas não técnicas, também denominadas de perdas comercias, correspondem à diferença entre as perdas globais ou totais de energia elétrica e as perdas técnicas. As mesmas estão associadas a problemas de faturamento das concessionárias e aspectos socioeconômicos das áreas de concessão das empresas de energia elétrica. De acordo com [1], as ocorrências de perdas não técnicas são mais frequentes no sistemas de distribuição em comparação aos sistemas de geração e transmissão, pois estes estão mais expostos a ações ilegais por consumidores, principalmente na redes secundárias de distribuição (baixa tensão). As perdas não técnicas são causadas principalmente por [2][3]:

- Furtos de energia elétrica (Conexões clandestinas na rede secundária de distribuição);
- Falhas ou fraudes nos medidores (Equipamento defeituoso ou ação intencional de violação para registrar menor consumo de energia);
- Irregularidades no faturamento da distribuidora (estimativa equivocada do consumo e instalações sem medidores).

As perdas não técnicas proporcionam enormes prejuízos financeiros as concessionárias de energia elétrica. No Brasil, segundo a Agência Nacional de Energia Elétrica (ANEEL), no ano de 2018 o custo relativos as perdas não técnicas ultrapassaram o valor de R$ 5,0 bilhões [4]. Nos Estados Unidos e Canadá, os prejuízos anuais estimados são de US$ 1,6 bilhões e US$ 100 milhões, respectivamente. A Malásia durante o ano de 2004, registrou um recorde no deficit de custos de perdas não técnicas, aproximadamente US$ 229 milhões. Na Índia, o prejuízo anual ocasionado pelas perdas comercias gira em torno de US$ 4,5 bilhões [5][6].

A identificação da existência de perdas não técnicas em unidades consumidoras (UC) é considerado um processo de avaliação complexa pela distribuidora. O procedimento mais usual para a detecção desse tipo de perdas no sistemas de distribuição ainda é a inspeção local. No entanto, existem alguns aspectos que entravam essa prática, como despesas com equipes de manutenção e tempo de inspeção para a investigação de UC suspeitas. Dessa forma, metodologias para detecção de perdas não técnicas empregando métodos computacionais de Inteligência Artificial (IA) vem sendo utilizadas pelas distribuidoras para auxiliar equipes de inspeção e indicar consumidores suspeitos de irregularidades. Segundo [7], as técnicas de IA mais utilizadas são Redes Neurais Artificiais (RNA), Lógica Fuzzy e Algoritmos Genéticos (AG). Entretanto, outros métodos *Expert Systems* (ES), *Multi-Agent Systems* (MAS), *Decision Tree* (DT), *K-Nearest Neighbor* (KNN), *Random Forest* e Algoritmos de *Clustering*, estão sendo aplicados em metodologias mais avançadas.

O processo de identificação de perdas não técnicas é mais complexo em UC localizadas nas área rurais em relação as instalações presentes no âmbito urbano. Isso se deve, principalmente, a presença de consumidores em locais com acesso remoto e grandes extensões das redes de distribuição rurais [8]. Esses fatores acabam dificultando a inspeção pela distribuidora. Assim, identificar e compreender os perfis de consumos das UC correspondentes a essas regiões é uma avaliação necessária no processo de identificação de possíveis consumidores irregulares.

O consumo de energia elétrica em áreas rurais é direcionado

https://doi.org/10.53316/sepoc2021.056

em grande parte para a produção agroindustrial. No Rio Grande do Sul, destaca-se o cultivo e a produtividade de arroz irrigado, sendo o maior estado produtor do Brasil, onde se utiliza energia elétrica em sistemas de bombeamento de água nas lavouras. Esses sistemas são compostos por bombas hidráulicas, motores de acionamentos, tubulação e peças especiais, e, sua funcionalidade é distribuir a água captada de uma fonte primária (rios, lagos, riachos ou barragens) para os canais de distribuição de água na lavoura. A atividade de irrigação representa uma parcela significativa no consumo de energia elétrica, pois os sistemas de bombeamento normalmente estão localizados no fim de uma rede alimentadora rural, concentrada e de uso praticamente contínuo durante o período da safra [9].

Nesse contexto, esse trabalho tem como objetivo apresentar uma metodologia para a identificação de possíveis ocorrências de perdas não técnicas levando em consideração, perfis de consumo de energia elétrica de UC irrigantes no município de Uruguaiana, na Região da Fronteira Oeste do Rio Grande do Sul, através de dados de consumo de energia elétrica dos últimos 4 anos.

II. Metodologia

Nesta seção é apresentada a metodologia implementada para classificação do conjunto de dados que será apresentado na Seção III. Serão brevemente contextualizados, o método utilizado para o agrupamento de dados, bem como os preparativos para a implementação do aprendizado de máquina e as técnicas empregadas para esse estudo.

A. Agrupamento de Dados Utilizando k-Means

O *k-Means* (KM) realiza o agrupamento através do cálculo de centroides dos dados, onde um determinado dado pertence à um grupo se a distância euclidiana de seus preditores ao centroide deste grupo se encaixar em uma tolerância. Sendo assim, o algoritmo agrupa dados de mesma categoria pelas menores distâncias destes aos clusters [10].

Neste trabalho, o método de KM realiza as predições das classes do consumo de energia elétrica de unidades consumidoras. Posteriormente, com as alterações no consumo, serão verificadas possíveis ocorrências de perdas não-técnicas.

B. Random Forest

O método de *Random Forest* (RF) foi desenvolvido por [11] e é baseado em árvores de decisão para classificação e regressão, partindo do princípio de que uma combinação de classificadores agregados possui melhor desempenho que um único classificador [12].

Seu princípio de funcionamento consiste em criar diversas árvores de decisão, onde o atributo de decisão é escolhido de forma aleatória. Em cada iteração, o algoritmo ignora a parte aleatória selecionada no conjunto de dados de treino, que é utilizada somente para criar uma árvore. O processo se repete diversas vezes, construindo um conjunto de árvores onde, para classificar uma instância, a mesma é passada em todas as árvores do conjunto, que a analisam separadamente. O resultado final é obtido pela combinação dos resultados de cada árvore [13].

C. Métricas de Avaliação

A matriz de confusão, Tabela I, foi usada para calcular a precisão baseada na classificação correta e incorreta. A matriz de confusão é capaz de representar duas ou múltiplas classes problemas. No entanto, seu uso na literatura em pesquisas relacionadas ao conjunto de dados desequilibrados é mais concentrado nas duas classes problemas, também conhecidos como problemas binários ou binomiais [14] o qual a classe menos frequente é nomeada como positiva e as demais classes são mescladas e nomeadas como negativas. Algumas das medidas mais conhecidas, derivadas dessa matriz, são as taxas de erro e precisão.

Tabela I
Matriz de Confusão

	Previsão Positiva	Previsão Negativa
Classe Positiva	Verdadeiro Positivo (VP)	Falso Negativo (FN)
Classe Negativa	Falso Positivo (FP)	Verdadeiro Negativo (VN)

A taxa de classificação incorreta ou de erro é dada por (1). A acurácia por (2), outras métricas associadas são precisão, revocação (*Recall*) e *F-1 score*.

$$\text{Erro} = \frac{FP + FN}{VP + FN + FP + VN} \tag{1}$$

$$\text{Acurácia} = \frac{VP + VN}{VP + FN + FP + VN} \tag{2}$$

A precisão (3) fala sobre quão preciso/exato nosso modelo está fora dos previstos positivos, quantos deles são positivos realmente.

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{3}$$

Revocação(4) calcula quantos dos atuais positivos nosso modelo captura, rotulando-os como positivos (Positivos Verdadeiro).

$$\text{Revocação} = \frac{VP}{VP + FN} \tag{4}$$

F-1 score (5) é em função da precisão e da revocação e é necessária quando se deseja buscar um equilíbrio entre eles. A precisão pode ser amplamente contribuída por um grande número de verdadeiros negativos, quando o *dataset* está desequilibrado, a precisão pode não ser suficiente, porque ao prever todas as amostrar como sendo da classe principal, ainda pode apresentar alta precisão. Então, *F-1 Score* pode ser uma medida melhor para se usar caso seja necessário procurar o equilíbrio entre precisão e revocação, obtendo-se uma desigual distribuição de classe. Para um conjunto de dados desbalanceados como o nosso caso, as métricas de

precisão e revocação são as que permitem melhor avaliação de desempenho do modelo preditivo, sumarizados pela pontuação do f1-score.

$$F\text{-}1\ score = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \quad (5)$$

III. Conjunto de Dados e Pré-Processamento

Nessa seção serão apresentados o *dataset* empregado nesse estudo e os procedimentos de pré-processamento dos dados utilizados como parte no desenvolvimento da metodologia desse trabalho. O estágio de pré-processamento foi uma atividade desenvolvida com o objetivo de identificar UC que apresentassem dados de cadastro nulos, ausentes ou incorretos, identificar correlações de atributos e melhor compreensão das informações presentes no *dataset*.

A. Conjunto de Dados: Dados de Consumo de Clientes Rurais do Município de Uruguaiana - RS

O conjunto de dados inicial utilizado como objeto de estudo neste artigo foi fornecido pela CPFL Energia, contendo o histórico de consumo mensal de energia elétrica de clientes rurais do município de Uruguaiana no estado do Rio Grande do Sul dos anos de 2017, 2018, 2019 e 2020. Originalmente o *dataset* fornecido possuía as informações de consumo em linhas, totalizando 100534, e, dentre as 9 colunas originais, selecionamos três colunas de interesse, sendo elas: código de instalação, data de referência (no formato 4 dígitos para ano e 2 dígitos para o mês), energia faturada.

Na primeira etapa do pré-processamento foi realizada a "transposição"da coluna data de referência, transformando as linhas em colunas com o valor de energia faturada associado à cada data, o *dataset* no fim dessa etapa possui 1714 linhas (UC) e 48 colunas.

Por conseguinte, na segunda etapa são inseridos no *dataframe* 2 novas colunas: Desvio padrão e Média, com os dados de desvio padrão e média do consumo de cada UC. As colunas são apresentadas na Tabela II.

Na terceira etapa de pré-processamento foram eliminados os *outliers* do conjunto de dados de maneira que tenhamos apenas os pontos dentro dos limites de 25% e 75% dos valores observados de média e desvio padrão, resultando em 1258 linhas e 51 colunas. A Figura 1 apresenta o gráfico do Desvio Padrão (kWh) x Média (kWh) do conjunto de dados pós processamento.

Tabela II
Dataset: Atributos Selecionados.

Atributo	Descrição
COD_NSTALACAO	código de identificação único para cada UC
201701	Consumo de Energia: aaaamm
...	...
202012	Consumo de Energia: aaaamm
DESVIO_PADRAO	Desvio Padrão do Consumo da UC
MEDIA	Média de Consumo de cada UC

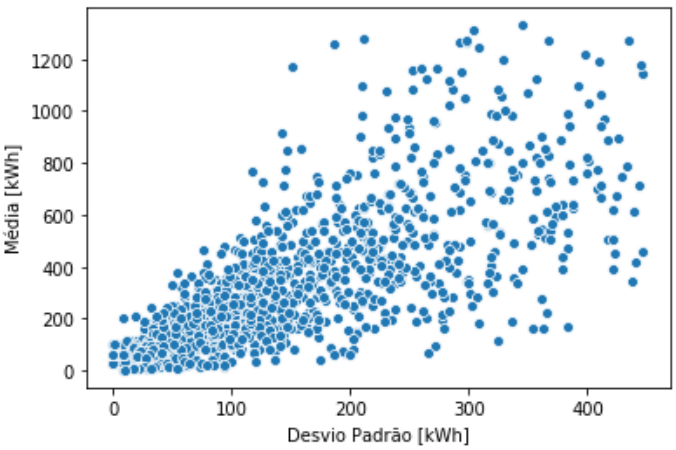

Figura 1. Desvio Padrão vs Consumo Médio Para Cada Cliente.

B. Cluesterização e Classificação de Clientes

O objetivo de aplicar *Clustering* é agrupar os pontos de maneira que os pontos em um grupo específico sejam semelhantes entre si e menos semelhantes aos pontos em outros grupos, utilizando algum critério como a similaridade dos dados para formar os grupos, para tanto deixamos a cargo do algoritmo encontrar padrões nos dados e agrupá-los [15]. É importante ressaltar que algoritmos de clusterização feitos para resolver problemas numa área específica geralmente fazem deduções baseadas em suposições em favor da área especifica, sendo uma limitação que afeta inevitavelmente a performance quando se aplica estes algoritmos em áreas cujos problemas não satisfazem essas premissas [16].

Como o objetivo do trabalho apresentado neste artigo é o desenvolvimento de um algoritmo para detecção de possíveis ocorrências de perdas não técnicas por perfil de consumo, foi utilizada a técnica de aprendizado de máquina não-supervisionado, *Agrupamento* (do inglês *Clustering*), para geração de classes, utilizando como entrada os atributos de Desvio Padrão e Média.

O algoritmo de clusterização utilizado foi o de método de partição baseado no erro quadrado: *K-means*, esse método associa um conjunto de indivíduos a *K* grupos sem criar uma estrutura hierárquica. O KM é bem simples e pode ser facilmente implementado para resolver muitos problemas práticos [16]. Para esta etapa foi utilizado o valor de K (n° de *clusters*) igual a 5. O resultado da etapa de *Agrupamento* é apresentado na figura 2, onde os grupos gerados são os Perfis de Consumo 0, 1, 2, 3 e 4. Para melhor entendimento das amostras, os valores de mínimo e máximo das classes são apresentados na Tabela III.

O gráfico em barras da Figura 3 apresenta o número de clientes em cada grupo gerado pela etapa de clusterização. Nota-se que o grupo 0 possui a maior parte de UC, 430 unidades consumidoras, o grupo 4 apresenta 371 UC, o grupo 2 possui 245 UC, o grupo 3 é o segundo menor com 151 UC e o grupo 1 é o de minoria, com 61 UC. Esse desbalanceamento

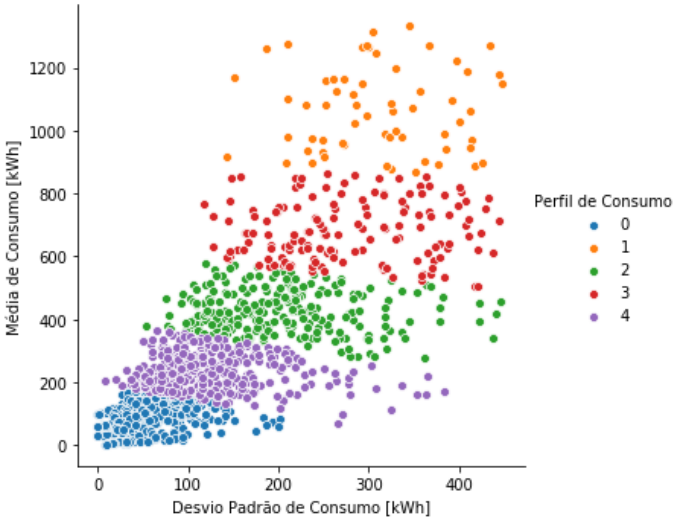

Figura 2. Resultado da Etapa de Clusterização.

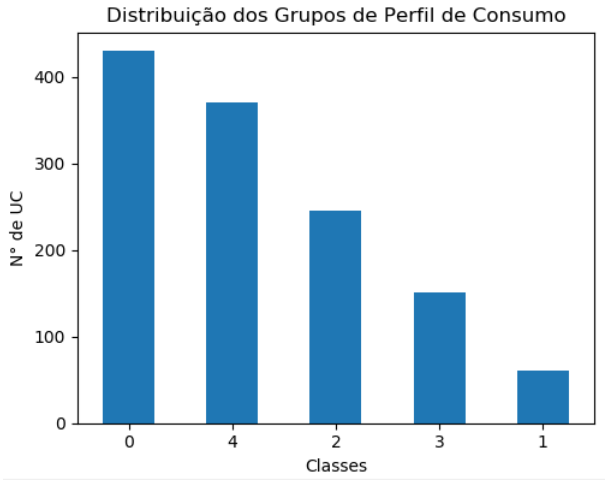

Figura 3. Distribuição do Grupos de Perfil de Consumo: Resultado da Clusterização.

Tabela III
PERFIL DE CONSUMO: VALORES MÍNIMOS E MÁXIMOS (kWh).

Classes		MEDIA	DESVIO_PADRAO
0	Max	181,886	202,215
	Min	1,486	0
4	Max	364,657	383,657
	Min	71,429	8,7777
2	Max	578,914	446,122
	Min	275,4	54,138
3	Max	863,057	443,754
	Min	504,4	117,34
1	Max	1330,95	446,904
	Min	869,714	142,388

entre grupos pode gerar problemas posteriores de classificação, uma vez que ao construir o modelo preditivo ele terá baixo desempenho de classificação correta para o grupo minoritário (Perfil de Consumo 1).

Existem várias abordagens para lidar com dados desbalanceados, cada uma com seus prós e contras. Para mitigar possíveis problemas preditivos do modelo causados pelos desbalanceados, foi implementado no algorítimo a técnica de *Subamostragem* das classes majoritárias (Perfil de Consumo: 0, 4, 2, 3). A *Subamostragem* envolve reduzir a classe (grupo) que contém a maioria dos registros de caso e reduzi-la para corresponder ao número de registros de caso na classe secundária. Apos realizada esta etapa deu-se inicio ao algoritmo de modelo preditivo, com dados de treino provenientes do novo conjunto de dados resultante da *Subamostragem*, Figura 4, contendo 61 amostras de todos os grupos, totalizando 305 amostras. O modelo de classificação foi montado com 70% das 305 UCs (213), e os outros 30% de dados de teste (92 amostras), para validação do modelo foi utilizado o conjunto de dados com as 1271 UCs. A distribuição do número de UCs dos conjuntos de dados são apresentados na Tabela IV.

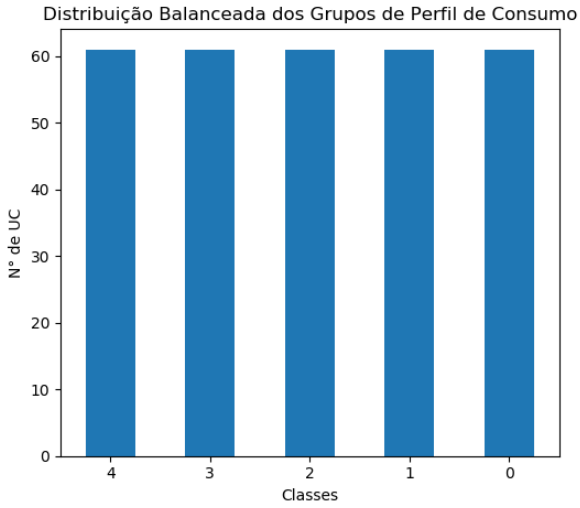

Figura 4. Distribuição Balanceada do Grupos de Perfil de Consumo

Tabela IV
N° DE UCS POR CONJUNTO DE DADOS DE TREINO, TESTE, VALIDAÇÃO

	N° Dados		
Classes	Treino	Teste	Validação
0	42	19	430
1	43	18	61
2	43	18	245
3	42	19	151
4	43	18	371

IV. Resultados e Discussões

Nesta seção, são apresentados os resultados das simulações e os comentários referentes ao modelo implementado.

A. Modelo Preditivo Aplicando Random Forest

Os resultados preditivos dos dados de consumo em relação aos perfis de consumo do conjunto de dados e subconjuntos de treino e teste são apresentado no relatório impresso ao fim da simulação do algoritmo, Figura 5. Os dados de entrada para o modelo foram: X = 12 colunas do histórico de consumo de energia elétrica das UC do ano de 2019; y = grupos gerados após o processo de clusterização (Perfil de Consumo: 0,1,2,3 e 4).

Também é apresentado na Figura 5 a matriz de confusão para a predição das 1258 UC. Nota-se que a pontuação de acurácia do modelo é de 82%, um bom resultado, contudo, os grupos dos Perfis de Consumo 0 e 4, foram os que apresentaram pior performance, observado pelos valores baixos de revocação (recall) e f1-score.

```
Acurácia do classificador RF nos dados de treino: 1.0000
Acurácia do classificador RF nos dados de teste: 0.7826
Acurácia do classificador RF nos dados completos: 0.8251


Matriz de Confusão da predição do conjunto de dados completo:
[[371   0   0   1  58]
 [  0  58   0   3   0]
 [  1   0 179  19  46]
 [  0   5  10 135   1]
 [ 50   0  22   4 295]]


              precision    recall  f1-score   support

           0       0.88      0.86      0.87       430
           1       0.92      0.95      0.94        61
           2       0.85      0.73      0.79       245
           3       0.83      0.89      0.86       151
           4       0.74      0.80      0.77       371

    accuracy                           0.83      1258
   macro avg       0.84      0.85      0.84      1258
weighted avg       0.83      0.83      0.83      1258

Pontuação de Acurácia: 0.8251192368839427
```

Figura 5. Resultado de Simulação do RF

B. Modelo Preditivo Utilizando Random Forest: Inserção de Dados com Consumo Alterado

Para exemplificação de detecção de clientes fora do seu Perfil de Consumo, na situação de possível ocorrência de perdas não técnicas, após a implementação do modelo classificatório de *Random Forest* foram inseridos como entrada no algoritmo uma lista de 13 elementos contendo: o código do cliente sobre análise (atributo Código de Instalação) e 12 valores de consumo do ano de 2020. Os valores de entrada são apresentados na Tabela V, onde os consumos dos meses 01, 02, 03, 09, 10, 11 e 12 foram decrescidos 40% do valor real (faturado pela CPFL), simulando o cliente irrigante consumindo menos que o comumente nos períodos de safra do arroz. As duas situações (Consumo Real e Consumo Alterado) foram simuladas.

Tabela V
Dados de Entrada Para Simulação de Possível Ocorrência de Perda Não Técnica

Código de Instalação : 3091534181		
Atributos	Consumo Mensal Atual [kWh]	Consumo Mensal Atual (alterado) [kWh]
CONSUMO_02001	180,4	108,24
CONSUMO_02002	168,1	100,86
CONSUMO_02003	100	60
CONSUMO_02004	159	159
CONSUMO_02005	100	100
CONSUMO_02006	134	134
CONSUMO_02007	125,05	125,05
CONSUMO_02008	100	100
CONSUMO_02009	116	69
CONSUMO_02010	106.6	64
CONSUMO_02011	100	60
CONSUMO_02012	100	60
Média Anual	124,1	95,06

As saídas para ambas entradas são apresentadas nas Figuras 6 e 7, respectivamente. Os relatórios de saída contém: o Cod_Instalação; apresentação das médias de consumo de energia elétrica do cliente para os anos anteriores presentes base de dados da CPFL e do ano vigente; a classe de Perfil de Consumo prevista para os consumos de entrada; a classe pré-definida pelo modelo na sua construção com os dados de consumo do ano de 2019; análise de possível ocorrência de perdas não técnicas ou não, dada pela mudança de Perfil ou não.

Nota-se que, para a entrada dos dados de consumo atual real, a classe prevista é a mesma da classe original, Figura 6. Para a entrada de consumo atual alterado, há uma mudança de classificação em que o cliente de origem: Perfil de Consumo 4, passou a ser classificado com Perfil de Consumo 0, figura 7, alertando uma situação de possível ocorrência de perda não técnica.

```
COD_INSTALACAO: 3091534181 | Consumo Mensal Atual


Ano    Média de Consumo Anual (kWh)
==================================
2017   338.091
2018   340.988
2019   205.742
2020   124.096


Classe de Perfil de Consumo Prevista: 4
Classe de Perfil de Consumo Original: 4
Situação: Não Houve Mudança de Perfil de Consumo
```

Figura 6. Saídas para Teste de Classificação: Consumo Real

V. Conclusão

Neste trabalho, foi apresentado um algoritmo classificador em *Random Forest* com o objetivo de classificar consumidores em *clusters*, chamados classes de Perfil de Consumo, utilizando como dados de entrada o consumo mensal de energia elétrica do ano de 2020 da UC, identificando se houve ou

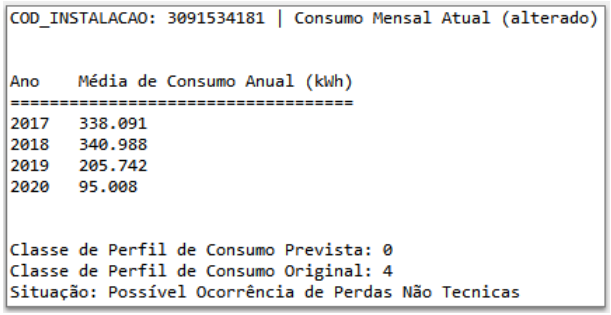

Figura 7. Saídas para Teste de Classificação: Consumo Alterado

não mudança de perfil e comparando a classe prevista com a cadastrada para a UC na fase de Clusterização. Desta maneira poderemos utilizar as saídas do algoritmo para identificar possíveis divergências no perfil de consumo de unidades consumidoras irrigantes do município de Uruguaiana - RS. O modelo apresentado foi implementado utilizando técnica de subamostragem com acurácia de 82% nas classificações, firmando importância para a correta classificação das classes com menor número de amostras. O contrário poderia ter sido implementado: inserir amostras nas classes de minorias de maneira a termos um conjunto de dados balanceado com base na classe majoritária, contudo essa técnica não fez-se presente neste trabalho. Algumas perspectivas para projetos futuros seriam: melhorias na precisão de classificação; teste do equilíbrio entre o número de consumidores das classes por sobre-amostragem; melhor delimitação no processo de agrupamento definindo faixas de valores para cada *cluster*. Em suma, para avaliar se um cliente da classe 0 divergiu de seu consumo para menos (Possível Perda Não Técnica) não haveria troca de classes, neste caso foi utilizado o comparativo com as médias de consumo dos anos anteriores, como auxilio na divergência.

AGRADECIMENTOS

Os autores gostariam de agradecer o apoio técnico e financeiro da CPFL Energia ao projeto "Sistema de Detecção de Perdas não Técnicas em Áreas de Irrigação Empregando Técnicas de Inteligência Artificial" (desenvolvido no âmbito do Programa de PD da ANEEL PD-00063-3065 / 2020). Este estudo também foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código Financeiro 001 e do National Instituto de Ciência e Tecnologia em Geração Distribuída (INCT-GD) da Universidade Federal de Santa Maria - UFSM, Brasil (processo CNPq 465640 / 2014-1, processo CAPES 23038.000776 / 2017-54 e FAPERGS 17 / 2551-0000517-1).

REFERÊNCIAS

[1] M. Madrigal, J. J. Rico, e L. Uzcategui, "Estimation of non-technical energy losses in electrical distribution systems," *IEEE Latin America Transactions*, vol. 15, no. 8, pp. 1447–1452, 2017, doi: 10.1109/TLA.2017.7994791.

[2] M. M. Buzau, J. T. Aguilera, P. C. Romero, e A. G. Expósito, "Detection of non-technical losses using smart meter data and supervised learning," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 2661–2670, 2019, doi: 10.1109/TSG.2018.2807925.

[3] J. R. Agüero, "Improving the efficiency of power distribution systems through technical and non-technical losses reduction," em *PES TD 2012*, Montevideo, Uruguai, 2012, doi: 10.1109/TDC.2012.6281652.

[4] ANEEL, *Relatório de Perdas de Energia Elétrica na Distribuição*, 1ª ed. Brasília, Brasil: Agência Nacional de Energia Elétrica, 2019.

[5] S. S. S. R. Depuru, "Modeling, detection, and prevention of electricity theft for enhanced performance and security of power grid," Tese de Doutorado, University of Toledo, Toledo, Estados Unidos, 2012.

[6] S. Chatterjee, V. Archana, K. Suresh, R. Saha, R. Gupta, e F. Doshi, "Detection of non-technical losses using advanced metering infrastructure and deep recurrent neural networks," em *2017 IEEE International Conference on Environment and Electrical Engineering and 2017 IEEE Industrial and Commercial Power Systems Europe (EEEIC / ICPS Europe)*, Milão, Itália, Jul. 2017, doi: 10.1109/EEEIC.2017.7977665.

[7] K. Warwick, A. Ekwue, e R. Aggarwal, *Artificial Intelligence Techniques in Power Systems*, 1ª ed. Reino Unido: Institution of Electrical Engineers, 1997, doi: 10.5555/261069.

[8] M. C. Evaldt, "Sistema neural artificial para identificação de perdas não técnicas em consumidores rurais," Dissertação de Mestrado, Universidade Federal de Santa Maria, Santa Maria, Brasil, 2018.

[9] L. L. Pfitscher, D. P. Bernardon, L. M. Kopp, M. V. T. Heckler, J. Behrens, P. B. Montani, e B. Thomé, "Automatic control of irrigation systems aiming at high energy efficiency in rice crops," em *2012 8th International Caribbean Conference on Devices, Circuits and Systems (ICCDCS)*, Playa del Carmen, México, 2012, doi: 10.1109/ICCDCS.2012.6188944.

[10] V. Marotta, *Aprendizado não supervisionado com o kMeans*. Universidade Federal de Viçosa, 2019.

[11] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[12] D. H. Costa, "Uso de séries temporais sentinel 1 na identificação de culturas agrícolas utilizando modelos de machine learning," Dissertação de Mestrado, Universidade de Brasília, Brasília, Brasil, Mar. 2020.

[13] M. C. Schiaffino, "Desenvolvimento de um método para classificação de comportamentos de ratos wistar utilizando o algoritmo de aprendizado supervisionado florestas aleatórias (random forests)," Dissertação de Mestrado, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brasil, 2020.

[14] M. Beckmann, N. Ebecken, e B. S. L. P. Lima, "A knn undersampling approach for data balancing," *Journal of Intelligent Learning Systems and Applications*, vol. 7, pp. 104–116, 2015, doi: 10.4236/JILSA.2015.74010.

[15] J. Bezdek e N. Pal, "Some new indexes of cluster validity," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 28, no. 3, pp. 301–315, 1998, doi: 10.1109/3477.678624.

[16] N. L. Cavalcanti, "Clusterização baseada em algoritmos fuzzy," Dissertação de Mestrado, Universidade Federal de Pernambuco, Recife, Brasil, Mar. 2006.