Metodologia para Detecção de Possíveis Ocorrências de Perdas Não Técnicas em Consumidores Rurais Aplicando Métodos de *Machine Learning*

Natalia B. Sousa * **Daniel P. Bernardon** * **Henrique S. Eichkoff** *
Pedro Marcolin * **Daniel L. Lemes** * **Luciana M. Kopp** **
Juliano S. Andrade *** **Lucas M. Chiara** ***

* *Universidade Federal de Santa Maria (UFSM), RS, (e-mail: rhyo.gabaglia@hotmail.com; dpbernardon@ufsm.br; henriquekoff@gmail.com; pedro_macolin@hotmail.com; daniellemesee@gmail.com).*
** *Universidade Federal de Pelotas (UFPEL), RS, (e-mail: lucianakopp@gmail.com)*
*** *Companhia Paulista Força e Luz (CPFL Energia), SP, (e-mail: julianoandrade@cpfl.com.br; lucaschiara@cpfl.com.br)*

Abstract: The difficulty of detecting non-technical losses by electric energy concessionaires has been a great and constant challenge. Inspecting consumer units located in rural areas demands excessive time and expenses on the part of concessionaires, due to the distance from urban centers and the difficulty of access, without there being a previous technical indication of the occurrence of Non-Technical Losses. This work aims to present a methodology for estimating electricity consumption for rice crops that use flood irrigation, in the city of Uruguaiana, Rio Grande do Sul, implementing classification using artificial intelligence techniques (clustering, k-means and random forest), and with the help of indicators, report cases of possible non-technical losses.

Resumo: A dificuldade de detecção de perdas não técnicas por parte das concessionárias de energia elétrica, tem se mostrado um grande e constante desafio. Inspecionar unidades consumidoras localizadas em zonas rurais, demanda tempo e gastos excessivos por parte das concessionárias, em função da distância dos centros urbanos e da dificuldade de acesso, sem que haja um prévio indicativo técnico de ocorrência de PNT (Perdas Não Técnicas). Este trabalho tem o objetivo de apresentar uma metodologia de estimativa do consumo de energia elétrica para lavouras de arroz que utilizam irrigação por inundação, no município de Uruguaiana, Rio Grande do Sul, implementar classificação dos mesmos por meio de técnicas de inteligência artificial (*clustering*, *k-means* e *random forest*), e com auxilio de indicadores, informar casos de possível PNT.

Keywords: Non-technical losses; Electricity Consumption; Clustering; Random Forest; K-means; Rural Consumers.

Palavras-chaves: Perdas Não Técnicas; Consumo; Clutering; Random Forest; K-meas; Consumidores Rurais.

1. INTRODUÇÃO

O consumo de energia elétrica em unidades consumidoras rurais é direcionado em grande parte para a atividade agroindustrial, apresentando um elevado consumo sazonal durante a época de cultivo, sendo estes consumidores, extremamente relevantes para as distribuidoras. No Rio Grande do Sul, destaca-se o cultivo do arroz irrigado, sendo o estado com a maior produtividade do Brasil. A energia elétrica no âmbito rural, é utilizada para o funcionamento dos sistemas de irrigação, responsáveis por distribuir uma quantidade de água captada de uma fonte disponível (rios, barragens, lagos ou riachos) para todas as parcelas das lavouras de arroz. Esses sistemas, também conhecidos como estações de bombeamento, são compostos por tubulações, peças especiais, bombas hidráulicas e motores de indução com potências que variam entre 100 e 300 cv, e permanecem acionados continuamente durante o período da safra do arroz irrigado (Köpp et al. 2016) (Pfitscher et al. 2012).

O cultivo de arroz irrigado é tipicamente comum no Rio Grande do Sul, apresentando altos índices anuais de produtividade. Normalmente, as lavouras ocupam extensas áreas e são cultivadas em ecossistemas de várzeas, encontrados nas encostas de rios, lagoas e lagunas. A época da safra tem inicio em Setembro, com a preparação do solo, e o plantio ocorre no período entre os meses de Outubro a

Dezembro. Já a colheira, acontece entre o final de Janeiro e o início de Abril, variando de acordo com as características dos cultivares plantados e da região de plantio (Uberti 2017). A Fronteira Oeste do estado, através do município de Uruguaiana, destaca-se como a região que apresenta os maiores potenciais produtivos para a respectiva cultura.

As unidades consumidoras presentes nas áreas rurais da Região da Fronteira Oeste do Rio Grande do Sul, apresentam elevados consumos mensais de energia elétrica, devido a atividade de irrigação da lavoura, principalmente nos meses de Janeiro, Fevereiro, Novembro e Dezembro. Dessa forma, esses consumidores são significativos nos alimentadores de distribuição e obrigam as concessionárias a disporem de um conhecimento prévio da carga instalada de seus clientes rurais, a fim de evitar irregularidades em registros de medições de energia faturada, por exemplo, perdas não técnicas. Uma alternativa para identificar e compreender os perfis de consumo das unidades consumidoras correspondentes a essas regiões, é através de estudos de estimativas ou previsões de consumo de energia elétrica. As metodologias tracionais aplicadas a resolução desse tema, são desenvolvidas a partir de diversas técnicas, tais como: Regressão, Séries Temporais, Redes Neurais Artificias, Lógica Fuzzy, Sistemas Especializados, entre outros (Anwar et al. 2018). No entanto, para análises em consumidores rurais, pode-se empregar uma abordagem mais simples, utilizando apenas informações dos principais atributos associados aos sistemas de irrigação e as características da lavoura de arroz.

Dentre os principais atributos que compõem o processo do cultivo do arroz irrigado, destaca-se a área de plantio, variável que representa a parcela da lavoura cultivada com a cultura do arroz e que necessita de irrigação (Uberti 2017). No Rio Grande do Sul, é comum a rotatividade de áreas de plantio de uma mesma lavoura em um intervalo entre três ou mais safras, devido as características dos solos. No ponto de vista do consumo de energia elétrica da instalação rural, a área de plantio é fator imprescindível para o dimensionamento do sistemas de irrigação, pois a mesma determina a vazão total a ser aplicada na lavoura, e consequentemente, influencia no consumo de água e energia elétrica da unidade consumidora.

A irrigação de uma lavoura de arroz deve ser suficiente para a atender a necessidade hídrica da área de plantio, ou seja, saturar o solo, formar um nível de água e compensar a evapotranspiração e perdas nas tubulações (Uberti et al. 2017). No entanto, esse processo pode ocasionar um consumo de energia elétrica expressivo para o consumidor a fim de atender essa demanda hídrica, devido ao funcionamento praticamente contínuo das estações de bombeamento durante a época da safra. Dessa forma, são recomendados desligamentos dos sistema de irrigação em dias com níveis de chuvas significativos, pois a precipitação pluviométrica colabora com a reposição das perdas hidráulicas nas lavouras, além de ser uma forma natural de irrigação para a agricultura. Assim, quanto maiores os índices de precipitação no período da safra, menor será a necessidade de irrigação complementar na lavoura, e consequentemente, menor será o consumo de energia elétrica das estações de bombeamento (Marcolin et al. 2021).

Este trabalho tem por objetivo apresentar uma metodologia de detecção de possíveis situações de perdas não técnicas (PNT) em um conjunto de consumidores rurais do município de Uruguaiana/RS, empregando métodos de *Machine Learning* aplicados ao processo agrupamento e classificação de dados. Para isso, são propostos diferentes indicadores a partir de dados reais de consumo mensal de energia elétrica, para avaliar irregularidades para uma determinada unidade consumidora irrigante.

O presente trabalho está estruturado da seguinte forma: A Seção 2 apresenta o conjunto de dados dos clientes irrigantes e as etapas de pré-processamento dessas informações para aplicabilidade nos demais estágios. A Seção 3 descreve os algoritmos de aprendizagem de máquina utilizados na desenvolvimento da metodologia proposto, sendo esta, apresentada na Seção 4, contextualizando todos os estágios presentes na mesma. A Seção 5, contextualiza o estudo de caso, e seus resultados e discussões. Por mim, na Seção 6 serão apresentados as conclusões desse trabalhos e as propostas de continuidade do presente estudo.

2. DADOS E PRÉ-PROCESSAMENTO

Neste trabalho, foram utilizados conjunto de dados providos pela concessionária de energia elétrica e equipe de processamento de imagens. Os dados da distribuidora são de clientes rurais irrigantes do município de Uruguaiana, e foram utilizados para implementação do modelo classificador e para validação, enquanto que os dados da equipe de processamento de imagens foram utilizados na metodologia de associação de áreas de plantios para as UCs. Nesta seção são descritos os conjuntos de dados de ambas origens e a etapa de pré-processamento, para posterior uso dos mesmos na metodologia implementada.

2.1 Dados Cadastrais das UCs rurais de Uruguaiana

Um dos conjunto de dados utilizados contém registros de 440 Unidades Consumidoras (UCs) do município de Uruguaiana, e seus respectivos atributos cadastrados em 61 colunas. O conjunto de dados, além de conter cadastros das UCs de natureza comercial, contem os valores do consumo de energia elétrica mensal do ano vigente, no caso do estudo apresentado, o ano de 2021. Outro conjunto de dados contêm os valores históricos de consumo de energia elétrica das UCs para os anos: 2017, 2018, 2019 e 2020. Alguns desses atributos contém valores nulos, ausentes (NaN) ou erroneamente cadastrados.

Na etapa de pré-processamento os dois conjuntos de dados foram mesclados utilizando como chave de junção o código de identificação da UC, um atributos comum entre as tabelas. Foram retirados os dados Nan ou com possíveis erros de cadastro, resultando em 339 UCs disponíveis para serem utilizados na etapa de modelagem. Os atributos também foram filtrados de maneira que apenas fossem usadas colunas cujos dados fossem de interesse para a metodologia proposta, como por exemplo: dados de consumo de energia elétrica, código de identificação da UC, e, coordenadas de latitude e longitude do medidor da UC.

Os dados de consumo são 12 valores de consumo mensal para os anos de 2017, 2018, 2019 e 2020. Escolheu-se

utilizar os consumos de 2020 para gerar o modelo classificador proposto, enquanto que os anos de 2019 e 2021 foram utilizados para comparação de resultados e validação, descartando-se os anos de 2017 e 2018, que apresentavam maior número de dados ausentes. Dessa maneira, foram utilizados os consumos do ano de 2020 para gerar dois novos atributos: média e desvio de consumo. Esses dois novos atributos foram calculados para todas as 339 UCs com o objetivo de, posteriormente, utiliza-los na geração de *clusters* que indicarão Perfis de Consumo da UC. A Figura 1 apresenta o gráfico da média x desvio (σ) do resultado desta etapa de pré-processamento em kWh.

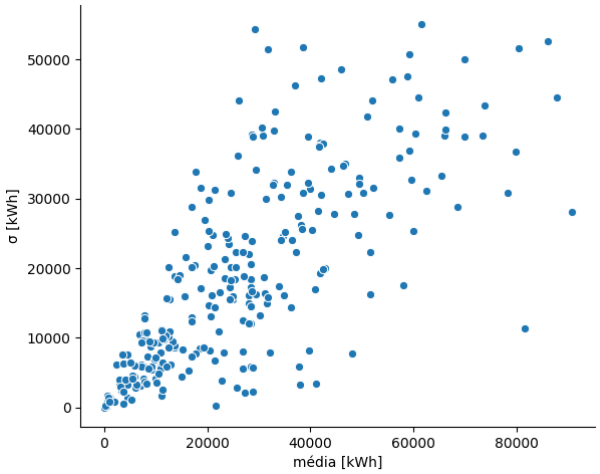

Figura 1. Resultado da etapa de pré-processamento.

2.2 Dados do Processamento de Imagens da Zona Rural de Uruguaiana

Os dados provenientes do estudo da zona rural de Uruguaiana, conhecida pela plantio de arroz, fornecem as seguintes informações e valores: coordenadas geográficas de latitude e longitude dos centroides de áreas de plantio, e tamanho da área em metros quadrados da área de plantio. Para o estudo de como associar as áreas das lavouras às UCs os autores levaram em conta os seguintes fatores na etapa de pré-processamento:

- Não é possível simplesmente associar cada área de plantio à uma UC, pois observou-se que alguns medidores de clientes rurais estão mais distantes das suas áreas de plantio, enquanto que estão mais próximos de áreas que não lhe pertencem. Logo, proximidades entre UC e área não é um fator decisivo para associação.
- Uma área associada a uma centroide não significa que ela é uma lavoura única, pois devido a diferença de relevos, características da lavoura, etc., na análise das imagens ocorre divisão de uma lavoura maior em mais de uma área de plantio, ou seja diferentes centroides (com valores menores de área) podem pertencer a uma única lavoura maior.
- Para contornar os problemas descritos nos itens acima, decidiu-se criar *clusters* que englobam UCs e centroides. Dessa maneira, o posterior estudo de estimativa de consumo se baseia no grupo de cada UC (*cluster*)
- Usar na geração de *clusters* apenas as UCs que estão próximas de fato de alguma área de plantio, para isso foram filtradas as UC distantes em mais de 30% da UC com menor distância de uma área de plantio.
- Alguns dados provenientes do processamento de imagens ainda estão em processo de aprimoramento, então, alguns centroides detectados se encontram nos limites das fronteiras do município.

A Figura 2 apresenta geograficamente as coordenadas dos medidores das UCs irrigantes (em azul) e as coordenadas dos centroides das áreas de plantio (em vermelho), resultantes da etapa de pré-processamento. Ressalta-se o número maior de áreas de plantio, nesta etapa temos 1183 pontos de centroides, em relação ao número de UCs, pois como comentado anteriormente, algumas lavouras são particionadas em áreas menores.

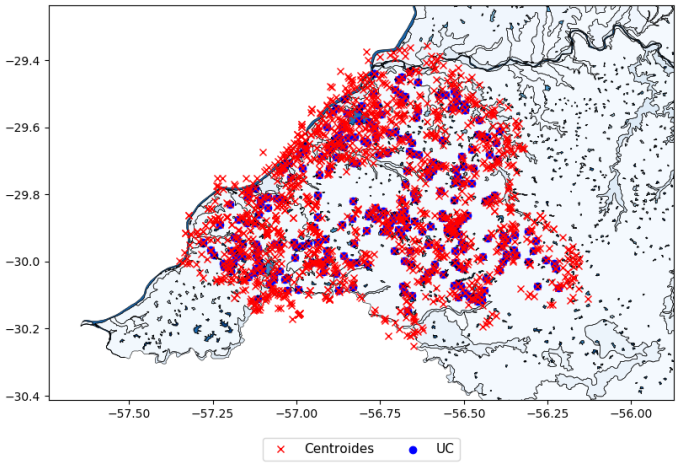

Figura 2. Mapa geográfico dos centroides de áreas de plantio e UCs irrigantes de Uruguaiana.

3. ALGORITMOS UTILIZADOS

Com as etapas de estudo e pré-processamento dos dados terminadas, temos um novo conjunto de dados para serem utilizados na implementação do modelo proposto. Na Seção 3 serão brevemente descritos os algoritmos de aprendizado de máquina utilizados no desenvolvimento da metodologia implementada pelos autores.

3.1 Kmeans

De acordo com (Fontana et al. 2009), K-means utiliza o conceito de centroides como protótipos representativos dos grupos, onde o centroide representa o centro de um grupo, sendo calculado pela média de todos os objetos do grupo. É uma técnica que usa o algoritmo de agrupamento de dados por K-médias (*K-means clustering*). O objetivo deste algoritmo é encontrar a melhor divisão de P dados em K grupos C_i, i = 1, ... K, de maneira que a distância total entre os dados de um grupo e o seu respectivo centro, somada por todos os grupos, seja minimizada (Pimentel et al. 2003). Em Pimentel et al. (2003), são descritos os passos do funcionamento do K-means:

(1) Atribuem-se valores iniciais para os protótipos seguindo algum critério, por exemplo, sorteio aleatório

desses valores dentro dos limites de domínio de cada atributo;
(2) Atribui-se cada objeto ao grupo cujo protótipo possua maior similaridade com o objeto;
(3) Recalcula-se o valor do centroide (protótipo) de cada grupo, como sendo a média dos objetos atuais do grupo;
(4) Repete-se os passos 2 e 3 até que os grupos se estabilizem.

K-Means possui uma complexidade computacional equivalente a O (N ∗ k ∗ T), pois, a cada iteração (T), é calculada a distância entre os N objetos até cada um dos k centroides. Se considerarmos nessa medida de complexidade o número de atributos, passaríamos para O (N ∗ k ∗ T ∗ n), já que cada comparação entre dois objetos tem complexidade computacional O (n).

3.2 Kmeans Constrained

O *Constrained K-means* tem na inicialização dos centroides a utilização de sementes e é um algoritmo semi supervisionado. A modificação principal feita por (Wagstaff et al. 2001) foi no momento da atualização dos centroides, o algoritmo garante que nenhuma das restrições antes especificadas sejam violadas, ou seja, o elemento que faz parte do conjunto do cálculo dos centroides iniciais não pode ter a classe que lhe foi dada alterada, garantindo assim que tais elementos não sejam rotulados erroneamente (Wagstaff et al. 2001). De forma análoga, os passos seguidos pelo *Constrained K-Means* são:

(1) Escolhe aleatoriamente k centros para os *clusters*;
(2) Atribui cada objeto para o *cluster* de centro mais próximo sem violar as restrições;
(3) Atualiza cada centro para a média dos objetos do *cluster* correspondente;

3.3 Random Forest - RF

Em 2001, L. Breiman introduziu pela primeira vez o termo *Random Forest* (RF), que é uma forma conglomerada de várias árvores de decisão autodeterminantes (Breiman 2001). Cada árvore neste método é atribuído com um voto unitário e, após a votação, o número máximo de votos diagnosticará o problema e calculará a precisão da classificação para resolver toda a técnica de classificação (Chakraborty et al. 2019).

O resultado final de uma amostra das entradas é determinado pela votação da classificação majoritária. No processo de treinamento do modelo, cada arvore de decisão é construída em uma amostra dos dados de treinamento, utilizando um subconjunto de variáveis selecionadas aleatoriamente. Logo, algumas amostras que não são utilizadas no processo de treinamento são chamadas de "*Out-of-Bag Samples* (OBB)", e são utilizadas para avaliar o desempenho geral da classificação. Além disso, as estimativas de erro da OBB fornecem uma avaliação imparcial da precisão, assim como da validação cruzada (Zhang et al. 2016).

4. METODOLOGIA E MODELO PROPOSTO

O modelo de detecção de possível ocorrência de PNT baseia-se na metodologia apresentada nesta seção, em que são usadas as técnicas de aprendizado de máquina, descritas anteriormente, no conjunto de dados resultantes da etapa de pré-processamento.

A Figura 3 resume a metodologia implementada, onde: primeiramente, realiza-se o agrupamento das áreas de plantio e UCs, inserindo no conjunto de dados um novo atributo referente ao resultado desse agrupamento; em seguida, são geradas classes de Perfil de Consumo (PF), utilizando os dados de média e desvio de consumo de 2020; se o agrupamento é satisfatório, é realizado a estimativa de consumo, onde são inseridos os valores calculados da estimativa de cada UC em um novo atributo no conjunto de dados; subsequentemente é dado inicio ao treino e teste do modelo RF. Os procedimentos e resultados dessas etapas são descritos a seguir.

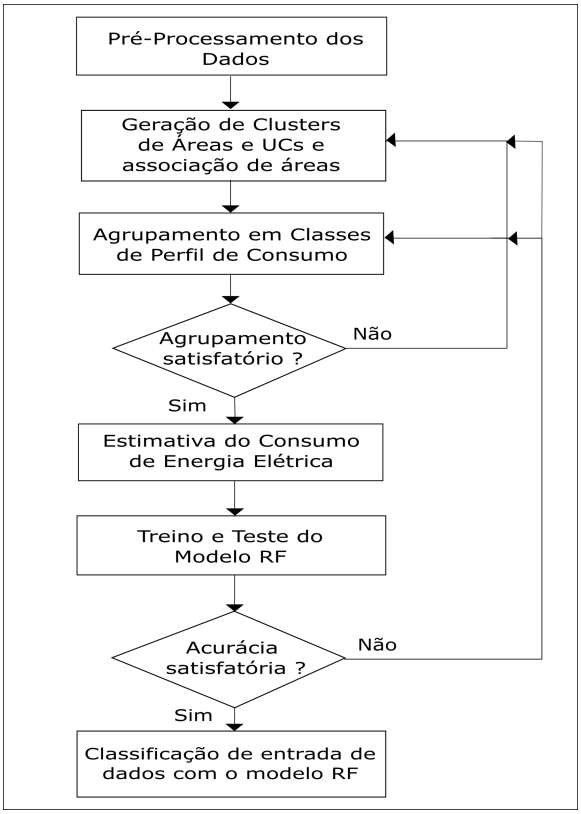

Figura 3. Fluxograma da Metodologia Proposta

4.1 Geração de Clusters *de Áreas de plantio e UC*

Como comentado anteriormente, o código de identificação da UC é único para cada uma das 339 amostras do conjunto de dados. Cada código apresenta valores cadastrados das coordenadas geográficas de latitude e longitude do seu ponto de medição, essas coordenadas juntamente com as coordenadas dos centroides das áreas de plantio apresentados, apresentado na Seção 3, são utilizadas como dados de entrada para a criação dos *clusters*. Com os 1183 pontos de centroides e mais os 339 pontos de UCs, temos no total 1522 pontos para serem agrupados utilizando o método *Kmeans*.

O número de *clusters* a serem gerados foi definido em 56. O resultado desse agrupamento, Figura 4, é inserido como um novo atributo (coluna) no conjunto de dados,

contendo o número de classificação (de 0 a 55) do *cluster* ao qual cada UC pertence. Paralelamente, também temos uma tabela contendo as centroides e a qual *cluster* (0 a 55) ela pertence. Por conseguinte, é gerada uma tabela auxiliar contendo os somatórios das áreas, em hectares, para cada *cluster*, que será utilizado posteriormente.

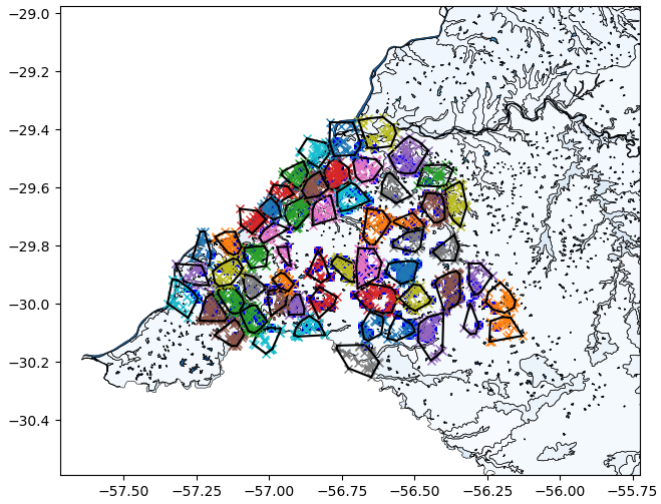

Figura 4. Resultado da Etapa de Geração de *Clusters* de áreas de plantio e UCs.

4.2 Associação de Áreas de Plantio às UCs

Após a definição das classes de áreas e UCs, podemos somar a os hectares totais de cada *cluster* e implementar uma abordagem inicial de associar cada UC a uma "parcela"da área de plantio total do *cluster* que ela pertence. Essa é uma metodologia inicial abordada pelos autores, ainda estuda-se outras abordagens para essa etapa a serem implementadas seguindo a obtenção de dados aprimorados.

Nessa etapa foram levados em consideração que: as imagens de satélite utilizadas pela equipe de processamento de imagens são do ano de 2020, e áreas maiores de lavouras resulta em consumo de energia maior esperado para o sistema irrigante da UC. Com isso, a metodologia utilizada nesta etapa segue os seguintes passos:

(1) É normalizado (de 0 a 1) a média de consumo do ano de 2020 das UCs, por *cluster*;
(2) O valor da normalização é utilizado como peso de cada UC no consumo de energia elétrica de cada *cluster*;
(3) A área total do *cluster* é multiplicada pelo valor de peso de cada UC;
(4) O resultado é atribuído como área de plantio para a UC.

No final do algoritmo dessa etapa temos uma área associada para cada umas das 339 UCs.

4.3 Geração de Classes de Perfil de Consumo (PF)

Para a geração das classes PF, decidiu-se retirar do conjunto de dados as amostras *outliers*, mantendo as amostras dentro dos limites de 25% e 75%. Logo, o resultado após a remoção dos *outliers* são 272 amostras. Escolheu-se remover os *outliers* após a associação das áreas para que a remoção dessas UCs não influencia-se na distribuição dos valores, já que, um número menor de UC resultaria em mais hectares para outras UCs.

Para esta etapa da metodologia 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 de consumo do ano de 2020. O algoritmo de clusterização utilizado foi *Kmeans Constrained*, com as restrições definidas como sendo o número máximo (*smax*) e mínimo (*smin*) de amostras (UCs) em cada classe gerada. Dessa maneira, o número de classes a serem geradas foi definido em 5, *smin* foi definido como o n° de amostras total (272) dividido pelo n° definido de classes (5), e *smax* definido como $smin + 1$. O resultado gráfico dessa clusterização é representado pela Figura 5. O n° de amostras por classe gerada, definidos nas restrições como smax = 55 e smin = 54, é apresentado na Tabela 1.

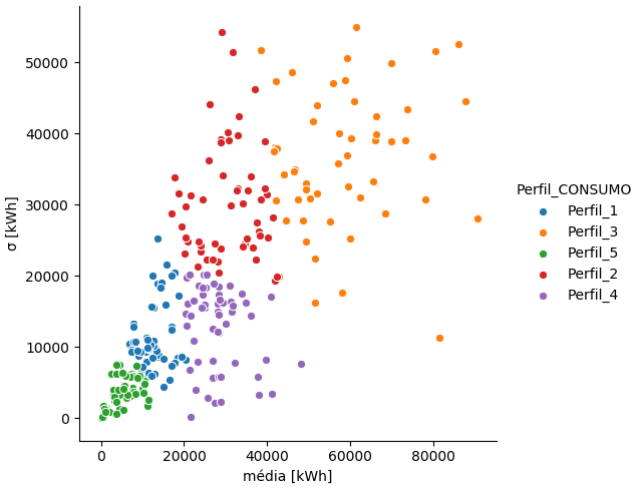

Figura 5. Resultado da Etapa de Geração de *Clusters* de Perfis de Consumo.

Tabela 1. Conjunto de dados: Atributos Selecionados.

Classes: Perfil de Consumo	N° Total de Amostras
Perfil_1	55
Perfil_2	54
Perfil_3	54
Perfil_4	54
Perfil_5	55

4.4 Estimativa de Consumo de Energia Elétrica das UCs Irrigantes de Arroz de Uruguaiana

O consumo de energia elétrica nas lavouras de arroz irrigado é direcionado, principalmente, para a atividade de irrigação. Esse atividade implica no acionamento de motores e bombas hidráulicas a fim de distribuir água para todas as parcelas da lavoura. Os sistemas de irrigação funcionam de maneira contínua, sendo desligados apenas em algumas situações especiais, como em faltas de energia ou em épocas com altos índices pluviométricos.

O principal método de irrigação empregado nas lavouras de arroz é a técnica de inundação contínua. No entanto, em algumas localidades vem sendo utilizado o método de irrigação por aspersão através de um pivô central, pois o mesmo pode apresentar uma economia de mais de 50da água (Stone et al. 2005). Para esse trabalho, serão estudas e consideradas apenas as UCs irrigantes que utilizem seus respectivos sistemas de irrigação através do método de inundação contínua. Os atributos associados ao consumo de energia elétrica nas lavouras de arroz irrigado e o método de estimativa de consumo de energia dos sistemas de irrigação por inundação aplicados a cultura do arroz, aplicados neste trabalho, são:

- Área Irrigada: A área irrigada representa a área total da lavoura cultivada com a cultura do arroz e que necessita de irrigação.
- Altura Manométrica Total: A altura manométrica (AMT) representa a energia que a bomba deverá transmitir a água para transportar uma determina vazão entre as tubulações de sucção e recalque. Para esse trabalho, a estimativa da altura manométrica total será realiza com base no estudo realizado em (Köpp et al. 2016).De acordo com o estudo, constatou-se que 65% das estações de bombeamento da das lavouras de arroz irrigado do município de Uruguaiana apresentaram alturas manométricas entre 5 e 15 m. Dessa forma, será atribuído para essa variável, o valor intermediário desse estudo, ou seja, 10 metros.
- Vazão: A vazão deriva da quantidade de água requerida pela lavoura de arroz e é considerada uma importante variável para o processo de dimensionamento dos sistema irrigantes. O dimensionamento dos conjuntos elevatórios é realizado a partir da vazão unitária, que por sua vez é dependente de fatores como lâmina adotada, tempo de bombeamento diário, aspectos físicos do solo e da topografia da área irrigada (Köpp et al. 2016). Para a manutenção de irrigação na lavoura é recomentada uma vazão unitária de 1,5 litros por segundo por hectare de área plantada (q = 1,5 L/s/ha). Para determinar a vazão total (Q) da área de plantio (A), utilizou-se a equação (1):

$$\mathrm{Q} = q * A \tag{1}$$

- Consumo de Energia Elétrica dos Sistemas de Irrigação: A estimativa do consumo de energia elétrica em UCs irrigantes, é inicia-se a partir do cálculo da estimativa da Potência Instalada do Sistema de Irrigação (P_{SI}) (Köpp et al. 2016). Dada pela formula (2), onde η é o rendimento do sistema. Em média, o rendimento global encontrado no estudo (Köpp et al. 2016) foi de 57 ± 15%. Neste trabalho utilizou-se o rendimento de 65%.

$$P_{SI} = \frac{Q * AMT}{\eta} * \frac{0.736}{75} \tag{2}$$

- Estimativa da Energia Elétrica Consumida pelo Sistema de Irrigação: É dada por (3). Onde E é a energia consumida pelo sistema de irrigação (kWh); P_{SI} é a potência do sistema irrigante (kW) e t é o tempo de funcionamento da estação de bombeamento (h). Para o valor de t considerou-se que as estações de bombeamento permanecem ligadas por um intervalo de 21 horas diárias durante um período que varia entre 80 e 100 dias (Köpp et al. 2016). Logo, em valores mensais, o tempo de bombeamento (t) utilizado foi de 450 horas.

$$\mathrm{E} = P_{SI} * t \tag{3}$$

4.5 Métricas de Avaliação do Modelo

Quando um modelo classificatório é elaborado, o mesmo precisa ser testado e avaliado para que se tenha uma dimensão do quão assertivo é o modelo para predição. Para tanto, métricas de avaliação são utilizadas para avaliação dos modelos de classificação de dados como acurácia, precisão e F-score (Mariano 2021). Uma das maneiras de representar os resultados de um método é através da matriz de confusão, mostrada na Tabela 2.

Tabela 2. Matriz de Confusão

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

Para calcular a taxa de acertos (Acurácia) e de erros, as classes são associadas conforme (4) e (5), respectivamente.

$$\text{Acurácia} = \frac{VP + VN}{VP + FN + FP + VN} \tag{4}$$

$$\text{Erro} = \frac{FP + FN}{VP + FN + FP + VN} \tag{5}$$

Outra métrica importante é a sensibilidade, que avalia a capacidade do método em detectar resultados classificados como positivos (Mariano 2021), dado por (6).

$$\text{Sensibilidade} = \frac{VP}{VP + FN} \tag{6}$$

Já a precisão avalia a quantidade de verdadeiros positivos sobre a soma de todos os valores positivos (Mariano 2021), dado por (7).

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{7}$$

Por fim, F-score é uma média harmônica calculada com base na precisão e sensibilidade (Mariano 2021), dado por (8).

$$\mathrm{F1} = 2 * \frac{preciso * sensibilidade}{preciso + sensibilidade} \tag{8}$$

4.6 Modelo RF

Na implementação do modelo RF, utilizado na predição das UCs em classes de Perfil de Consumo, a entrada são os valores de consumo do ano de 2020. A saída do modelo RF é a classificação PF (Perfil_1, Perfil_2, Perfil_3, Perfil_4 ou Perfil_5) da UC. Utilizando as métricas de avaliação descritas, obtivemos o melhor resultado de 76%

de acurácia, figura 6. Foi utilizado 80% do conjunto de dados para o treino do modelo, e 20% para os dados de teste.

Na Figura 6 também é apresentada a matriz de confusão, a precisão na classificação das classes e a pontuação F1, a coluna *suport* informa quantas amostras foram utilizadas na etapa de teste por classe. As classes (0 a 4) correspondem aos perfis de consumo (1 a 5), respectivamente.

```
Acuracia do Classificador RF nos dados de Treino: 0.93
Acuracia do Classificador RF nos dados de Teste: 0.76


Matriz de Confusão 2020:
[[ 7  0  0  2  1]
 [ 0  3  2  2  0]
 [ 0  3  6  0  0]
 [ 2  0  0 13  0]
 [ 1  0  0  0 13]]
```

	precision	recall	f1-score	support
0	0.70	0.70	0.70	10
1	0.50	0.43	0.46	7
2	0.75	0.67	0.71	9
3	0.76	0.87	0.81	15
4	0.93	0.93	0.93	14

Figura 6. Relatório Final de Simulação - Implementação do Modelo RF

5. ESTUDO DE CASO - RESULTADOS E DISCUSSÕES

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

O modelo RF foi utilizado tendo como entrada os dados atuais de consumo, ano de 2021, sua saída são: o resultado preditivo do PF atual; comparativo com o PF anterior (ano de 2020) (indicador 1); e três indicadores auxiliares.

Os indicadores auxiliares foram utilizados para considerarmos uma faixa de valor aceitável de variação entre o consumo estimado e a média de consumo. Visto que, durante o período de safra, condições adversas (chuvas, pandemia, PNT, etc.) podem resultar em consumos menores do que o estimado, decidiu-se considerar a média de consumo da UC como indicador de divergência ou não de consumo como um dos indicadores. Sendo assim, se a média de consumo da UC for menor do que o estimado, ou menor do que até 40% do valor estimado, a bandeira do indicador 2 é acionada para possível ocorrência de PNT. Os indicadores 3 e 4 referem-se as médias de consumos dos anos anteriores (2020 e 2019), respectivamente, fazendo o comparativo entre as curvas de consumo. As curvas de consumo, para este trabalho, utilizam apenas os valores médios observados.

O estudo de caso são apresentado é de uma UC especifica, escolhida aleatoriamente. Como comentado, seus consumos de 2021, presentes na base de dados da distribuidora de energia que atende a região, são utilizados como entrada no modelo RF. Para este caso, ocorre divergência entre as classes PF e entre os consumos, ou seja, um caso de possível ocorrência de PNT.

5.1 Estudo de Caso: Possível Ocorrência de PNT

No caso 1 exemplificamos uma UC em que os consumos do ano de 2021 resultaram em uma mudança na classe PF, e seus indicadores de consumo estimado e média de consumo também foram acionados. O relatório final de simulação deste caso é apresentado na Figura 7. Nota-se que ocorre mudança de PF 1 para PF 5, a energia consumida foi menor do que a estimada, e todos os indicadores foram acionados.

```
      Relatório Final de Simulação: Caso 1
Classe Prevista: Perfil_5
Classe Original: Perfil_1
Indicador 1: Mudança de PF
Situação: Possível Ocorrência de Perdas Não Técnicas


      Comparativo: Consumo Máximo 2021 & Consumo Estimado
Indicador 2: Consumo Médio Menor do que o Limite Estimado
Situação: Possível Ocorrência de Perdas Não Técnicas


      Comparativo: Média de Consumo 2021 & Consumo anterior (2020, 2019)
Indicador 3: Consumo Médio do ano vigente Menor do que o Consumo Médio de
2020
Situação: Possível Ocorrência de Perdas Não Técnicas
Indicador 4: Consumo Médio do ano vigente Menor do que o Consumo Médio de
2019
Situação: Possível Ocorrência de Perdas Não Técnicas
```

Figura 7. Relatório Final de Simulação - Caso 1

Na Figura 8 nota-se que o consumo do ano de 2021 para a UC testada foi muito menor do que os anos anteriores (gráfico superior), e pouco menor do que o estimado (gráfico inferior). Um ponto a ser observado é que o valor estimado aproximou-se do real nestes caso.

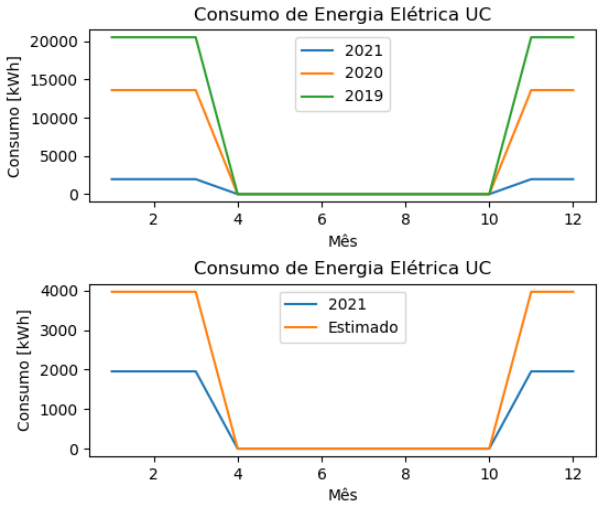

Figura 8. Gráfico de Consumos - Caso 1

6. CONCLUSÃO

O presente trabalho trouxe um estudo inicial de um modelo de detecção de perdas não técnicas em áreas de irrigação de arroz, especificadamente, neste trabalho, na fronteira oeste do Rio Grande do Sul. Foram associados diversos atributos para predizer o consumo das unidades consumidoras, além de utilizar algoritmos para geração de *clusters* e associação das áreas de plantação de arroz a sua respectiva unidade consumidora, além de ferramentas para avaliação do modelo proposto. A associação de áreas

de plantio às UCs é um ponto de grande importância para a estimativa de consumo de energia mais próxima possível do real, e, consequentemente, para classificação ou não de PNT. Utilizando dados iniciais das áreas, o total de hectares de cada *cluster* foi distribuído utilizando a média de consumo e pesos atribuídos de cada UC, logo, para o estudo da estimativa e indicadores for, os consumos foram avaliados pela média anual calculada. Por fim, foi apresentado um estudo de caso em que a unidade consumidora aciona os indicadores da metodologia, informando possível ocorrência de PNT.

AGRADECIMENTOS

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES/PROEX) – Código de Financiamento 001. 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 P&D da ANEEL PD-00063-3065 / 2020). Este estudo também foi financiado em parte pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código Financeiro 001 e pelo Instituto Nacional 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

Anwar, T., Sharma, B., Chakraborty , K., e Sirohia, H. (2018). Introduction to Load Forecasting. *International Journal of Pure and Applied Mathematics*. volume 119. n. 15. pp. 1527-1538.

Breiman, L. (2001). *Random Forests*. Machine Learning. volume 45. pp. 5-32.

Chakraborty, D., Sur, U., e Banerjee, P. K. (2019). Random Forest Based Fault Classification Technique for Active Power System Networks. Em *WIECON-ECE 2019*. Bangalore, Índia.

Fontana, A., e Naldi, M. C. (2009). Estudo e Comparação de Métodos para Estimação de Números de Grupos em Problemas de Agrupamento de Dados. Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação (ICMC). Universidade de São Paulo. São Carlos, Brasil.

Kopp, L. M., Peiter, M., Robaina, A. D., e Toescher, A. D. (2016). Caracterização de Estações de Bombeamento em Lavouras de Arroz do Rio Grade do Sul. *Journal of the Brazilian Association of Agricultural Engineering*. volume 36. n. 2. pp. 342-351.

Marcolin, P., Bernardon, D. P., Sousa, N. B., Eichkoff, H. S., Madaloz, J., Köpp, L. M., Chiara, L. M., e Silva, J. A. (2021). Metodologia para Detecção de Perdas Não Técnicas de Unidades Consumidoras Irrigantes de Arroz. Em *CBQEE 2021*. Foz do Iguaçu, Brasil.

Mariano, D. (2021). Métricas de Avaliação em Machine Learning. *Revista Brasileira de Bioinformática (BIOINFO)*. volume 1. pp. 1-9.

Pfitscher, L. L., Bernardon, D. P., Kopp, L. M., Heckler, M. V. T., Behrens, J., Montani, P. B., e Thomé, B. (2003). Automatic Control of Irrigation Systems Aiming at High Energy Efficiency in Rice Crops. Em *ICCDCS 2012*. Playa del Carmén, México.

Pimentel, E. P., França, V. F., e Omar, N. (2003). A Identificação de Grupos de Aprendizes no Ensino Presencial Utilizando Técnicas de Clusterização. Em *SBIE 2003*. Rio de Janeiro, Brasil.

Stone, L. F., Silveira, P. M., e Moreira, J. A. A. (2005). *Métodos de irrigação*. [Online]. Disponível em: https://www.agencia.cnptia.embrapa.br/gestor/arroz/arvore/CONT000foh49q3602wyiv8065610d5y5f5im.html. Acesso em 15 Jan. 2022. Brasília, Brasil.

Wagstaff, K., Cardie, C. S. R., e Schroedl, S. (2001). Constrained K-means Clustering with Background Knowledge. *Eighteenth Internacional Conference on Machine Learning*. volume 18. pp. 577–584.

Uberti, V. A. (2017). Lógica Fuzzy para Avaliação de Eficiência Energética em Sistemas de Irrigação de Lavouras de Arroz. Dissertação de Mestrado. Programa de Pós-Graduação em Engenharia Elétrica. Universidade Federal de Santa Maria. Santa Maria, Brasil.

Uberti, V. A., Abaide, A. R., Pfitscher, L. L., Evaldt, M. C., Prade, L. R., e Figueiredo, R. M. (2017). Fuzzy-based Methodology for Evaluation of Energy Efficiency in Rice Irrigation Systems. Em *UPEC 2017*. Heraklion, Grécia.

Zhang, C., Li, Y., Yu, Z., e Tian, F. (2016). Feature Selection of Power System Transient Stability Assessment based on Random Forest and Recursive Feature Elimination. Em *APPEEC 2016*. Xi'an, China.