DOI: 10.48011/asba.v2i1.1337

Agrupamento e Classificação de Consumidores de Energia Rural Utilizando *Random Forest* e *K-Nearest Neighbors*

Natalia Bastos de Sousa * **Daniel Pinheiro Bernardon** *
Henrique Eichkoff * **Pedro Marcolin** * **Julia Madaloz** *
Lucas Melo de Chiara ** **Juliano Andrade Silva** **
Luciana Marini Kopp ***

*Universidade Federal de Santa Maria, RS, (e-mail: rhyo.natalia@gmail.com).

*Universidade Federal de Santa Maria, RS, (e-mail: dpbernardon@ufsm.br).

*Universidade Federal de Santa Maria, RS, (e-mail: henriquekoff@gmail.com).

*Universidade Federal de Santa Maria, RS, (e-mail:pedro_marcolin@hotmail.com).

*Universidade Federal de Santa Maria, RS, (e-mail:juliacmadaloz@gmail.com).

**CPFL Energia Power Utility, SP, (e-mail:lucaschiara@cpfl.com.br).

**CPFL Energia Power Utility, SP, (e-mail:julianoandrade@cpfl.com.br).

***Faculdade de Agronomia Eliseu Maciel, Departemento de Eng. Rural, RS, (e-mail:lucianakopp@gmail.com).

Abstract: Detecting irregularities in energy consumption is an important challenge for energy companies. Commercial losses due to irregularities cause losses in the concessionaire's revenues and losses for the consumer. To detect possible occurrences of non-technical losses, this work aims to use machine learning techniques, both of the supervised type, for classification purposes, such as K-nearest neighbors and Random Forest, and unsupervised type (Clustering), with the objective of creating consumer groups, which will be used for classification prediction. The objective is to organize the consumers of the data set into groups based on consumption, as the data set presented has twelve months of data, the representation was made according to the average consumption by the deviation. Real characteristic data from various consumer units located in a rural region of the state of São Paulo were used to obtain the results of the methodology proposed in this work.

Resumo: Detectar irregularidades no consumo de energia é um desafio importante para as empresas de energia. Perdas comerciais por irregularidades causam perdas nas receitas da concessionária e perdas para o consumidor. Para detectar possíveis ocorrências de perdas não técnicas, este trabalho tem como objetivo utilizar técnicas de aprendizado de máquina, tanto do tipo supervisionado, para fins de classificação, como *K-nearest neighbors* e *Random Forest*, quanto do tipo não supervisionado (*Clustering*), com o objetivo de criar grupos de consumidores, que serão usados para previsão de classificação. O objetivo é organizar os consumidores do conjunto de dados em grupos com base no consumo, como o conjunto de dados apresentado possui doze meses de dados, a representação foi feita de acordo com o consumo médio pelo desvio. Dados característicos reais de diversas unidades consumidoras localizados em uma região rural do estado de São Paulo foram utilizados para a obtenção dos resultados da metodologia proposta desse trabalho.

Keywords: K-Nearest Neighbors (KNN); Random Forest; Clustering; Supervised Classification; Unsupervised; Machine Learning.

Palavras-chaves: K-Nearest Neighbors (KNN); Random Forest; Classificação supervisionada; Não supervisionado; Machine Learning.

1. INTRODUÇÃO

Metodologias utilizando técnicas de Inteligência Artificial (IA), estão cada vez mais presentes em análises nos sistemas elétricos de potência (SEP). As aplicações dessas técnicas, consistem em apresentar respostas mais precisas e eficazes em diversas áreas de estudo de um sistema elétrico. A classificação das unidades consumidoras de determinadas regiões, é uma análise que pode ser realiza a partir de métodos de modelagem computacional de IA. Esse estudo, consiste em classificar os consumidores em um respectivo conjunto de unidades e realizar comparações entre eles, utilizando dados de entrada como consumo diário de energia elétrica, demanda contratada da instalação e localizações geográficas.

As técnicas mais elementares de IA aplicadas aos sistemas de energia são Redes Neurais Artificias (RNA), Lógica Fuzzy e Algoritmos Genéticos (AG) (Warwick et al. 1997). Entretanto, outros métodos como *Expert Systems* (ES), *Multi-Agent Systems* (MAS), *Decision Tree* (DT), *K-Nearest Neighbor* (KNN) e *Random Forest*, também podem ser utilizados no desenvolvimento de metodologias de análises nos sistemas elétricos.

O KNN é um método comum utilizado para classificação, devido à sua simplicidade e precisão. A ideia central do algoritmo, é sempre que um novo ponto é previsto em um determinado conjunto, as amostras vizinhas mais próximas são escolhidas a partir dos dados de treinamentos definidos (Zhang et al. 2015). O *Random Forest* (em português, Floresta Aleatória) é um método utilizado para classificação e regressão, baseado num conjunto de árvores de decisão (*Decision Tree*) combinadas para resolver problemas de classificação. Cada árvore de decisão é criada a partir de uma amostra aleatória inicial dos dados e posteriormente classificados em subconjuntos aleatórios para definir os atributos mais informativos (Breiman 2001). Os atributos mais relevantes na formação da floresta, são estabelecidos pela importância acumulada do atributo nas divisões dos nós de cada árvore da floresta (James et al. 2013).

A classificação de consumidores não é relevante apenas para determinar e classificar os dados característicos das unidades consumidoras de uma região. Para as concessionárias de energia, essa análise pode ser aplicada para detectar possíveis irregularidades em sua área de concessão. Essas irregularidades estão relacionadas a problemas de faturamento ou por ações ilegais dos consumidores, como fraude e roubo de energia. Essas ações levam a perdas comerciais ou perdas não técnicas nos sistemas de distribuição e ocasionam enormes prejuízos econômicos para as distribuidoras, além de prejudicar a segurança e a qualidade da energia da rede elétrica, causando problemas de sobrecarga e níveis de tensão abaixo dos limites operacionais adequados.

As perdas não técnicas correspondem à diferença entre perdas elétricas na distribuição e perdas técnicas e estão associadas à falta de receita das concessionárias de energia elétrica e as características socioeconômicas das áreas de concessão. Essas perdas são mais específicas para os sistemas de distribuição, pois os mesmos estão mais expostos a ações ilegais em relação aos sistemas de geração e transmissão (Madrigal et al. 2017). As perdas não técnicas são causadas principalmente por roubo de energia, adulteração de medidores e ineficiências no setor comercial da distribuidora. Isso inclui: (i) consumo de energia não faturado (conexões clandestinas na rede secundária, iluminação pública, etc.); (ii) alteração da precisão dos medidores (ação intencional de violação para registrar menor consumo, equipamento defeituoso, etc.) e (iii) erros de leitura (estimativa incorreta do consumo de energia elétrica em unidades consumidoras localizadas em locais remotos, instalações sem medidores, etc.) (Agüero 2012).

O procedimento mais comum para detectar as perdas não técnicas ainda é a inspeção local. Entretanto, existem vários fatores que dificultam essa prática, como custos com equipes de manutenção e tempo de inspeção para investigar unidades consumidoras suspeitas. Em áreas rurais, grandes extensões de redes e dificuldades de acesso tornam complexa a inspeção. Assim, as metodologias para detecção de perdas não técnicas utilizando métodos computacionais de IA, podem auxiliar as equipes de manutenção, limitando as regiões de busca e indicando os consumidores suspeitos de irregularidades (Evaldt 2014).

Esse trabalho tem como objetivo apresentar, uma metodologia de classificação de consumidores que possa auxiliar na detecção de possíveis ocorrências de perdas não técnicas em diversos conjuntos de unidades consumidoras. Os algoritmos de *K-Nearest Neighbors* e *Random Forest*, além de técnicas de agrupamento (*Clustering*) foram utilizados para classificar as unidades consumidoras com base no consumo faturado. Para a obtenção dos resultados, foram utilizados dados reais de consumidores de uma região rural do estado de São Paulo.

Esse trabalho está estruturado da seguinte forma. A Seção 2 descreve os dados utilizados paras simulações na seção de metodologia e os procedimentos de pré-tratamento desses dados para as simulações. A Seção 3 apresenta a metodologia aplicada para a classificação dos dados. A Seção 4 demonstra os resultados obtidos com a metodologia proposta e a Seção 5, as considerações finais sobre o trabalho.

2. CONJUNTO DE DADOS E PRÉ-PROCESSAMENTO

Nesta seção são apresentados o dataset utilizado no estudo e os procedimentos de pré-processamento dos dados. A etapa de pré-processamento foi realizada para identificação de dados nulos, ausentes ou com erro de cadastro, também neste etapa foi identificada possíveis correlações de atri-

butos e melhor compreensão das informações contidas no dataset.

2.1 Conjunto de dados: Consumo de Clientes Rurais do Estado de São Paulo

O conjunto de dados inicial, utilizado como objeto de estudo neste artigo, é composto por 82 atributos de natureza diversa (colunas) de 10.000 consumidores do estado de São Paulo. Dentre os 82 atributos, citamos: identificação do consumidor rural ou urbano, 12 dados de consumo mensal, nome da cidade do consumidor, região do cliente (nordeste ou noroeste), descrição do sistema setorial, latitude e longitude do consumidor (supõe-se: medidor de energia), e status da conexão do consumidor (ativo ou inativo). O atributo de status da conexão foi utilizado como filtro, para que somente os clientes ativos fossem utilizados para o estudo, também foram selecionados clientes classificados como rurais.

Por conseguinte, na primeira etapa do pré-processamento dos dados, somente conexões ativas, reduziu o número de valores para 8.349 clientes, já a segunda etapa de seleção de clientes da região nordeste, reduziu para 6.795 valores (linhas). A terceira etapa de pré-processamento selecionou clientes com o atributo Sistema Setorial classificados como agricultura rural, reduzindo o numero de clientes para 6.511. Por fim, foram selecionados os atributos com os quais iremos trabalhar, resultando em 15 atributos, em seguida as linhas com valores nulos foram eliminadas. Resultando em um conjunto de dados de dimensão 3172x15. A tabela 1 apresenta os atributos selecionados, em que, no atributo CONUSMO_X, X possui valores de 1 a 12 , representando os meses do ano.

Tabela 1. Dados: Atributos Selecionados.

Atributo	Descrição	Tpo
COD_INSTALACAO	código de identificação único para cada cliente	Numérico
DESC_MUNICIPIO	Nome do município do cliente	Nominal
IND_CALIZACAO	Indicador de cliente rural ou urbano	Caractere
CONSUMO_X	12 atributos de consumo (kWh)	Numérico

2.2 Conjunto de dados: segunda análise e tratamento de dados

Nesta etapa foram inseridas duas novas colunas contendo os valores do consumo médio do cliente (avg) ,utilizando os 12 valores do consumo de energia elétrica do cliente fornecido pelo conjunto de dados, e de desvio padrão (std). Ressalta-se que os valores de consumo de energia estão na notação 10^8 kWh. A figura 1 apresenta o gráfico desses dois atributos.

Pela figura 1 percebe-se pontos distantes que podem causar classificações errôneas, logo, esses pontos foram eliminados do conjunto de dados para que trabalhemos apenas com os pontos dentro dos limites de 25% e 75% dos valores observados. O novo conjunto de dados com os valores dentro deste limite é apresentado na figura 2.

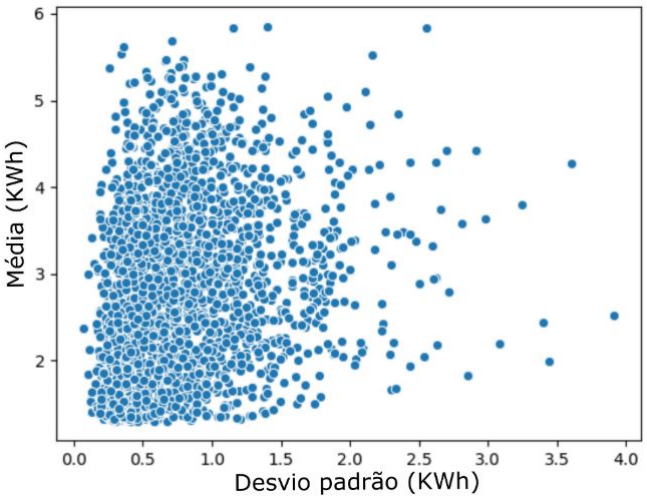

Figura 1. Consumo médio vs Desvio Padrão para cada cliente.

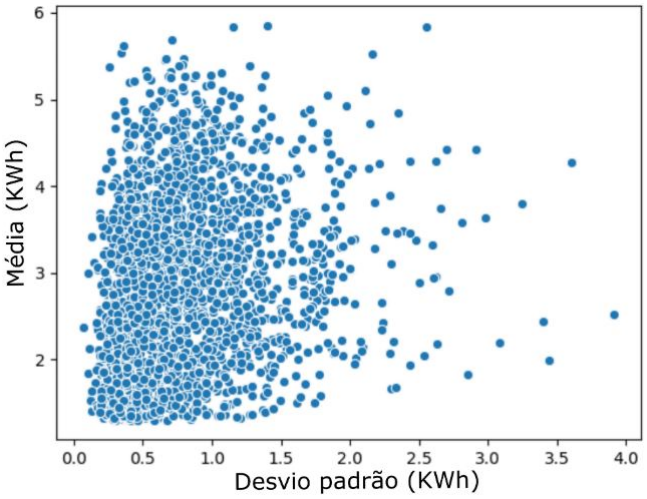

Figura 2. Novo Consumo médio vs Desvio Padrão para cada cliente.

2.3 Agrupamento e classificação dos clientes

Como estamos trabalhando com o objetivo de classificação dos clientes com base no consumo histórico de energia, foi calculado a entropia dos dados que servirão de entrada (12 atributos de consumo) para termos melhor ideia de seu comportamento e da medição de sua aleatoriedade, o resultado é apresentado na tabela 2.

Classificadores geralmente precisam gerenciar a entropia de maneira que ela não seja nem muito baixa e nem muito alta, normalmente dentro apresentando valores de 0 a 1. Como podemos verificar na tabela 2 os valores são altos, o que indica um alto grau de incerteza e aleatoriedade associados aos valores de consumo. Juntando o alto grau de entropia ao fato de não temos um atributo que sirva como classificador, de maneira que tenhamos clientes com perfil de consumo (avg x std) parecidos na mesma classe, os autores optaram por usar técnica de *clustering* para criação de classes de clientes. Dessa forma, teríamos classificadores para incorporar os algoritmos de aprendizado de máquina usando Random Forest e K-Nearest Neighbor (KNN). Segundo (Weiss et al. 2001), uma escolha importante para

minimizar o impacto do número de valores do conjunto de dados de treinamento (*training set*) ao desempenho do classificador, é a adequada distribuição de valores nas classes.

Tabela 2. Atributos de Consumo: Entropia.

Atributo	Entropia
CONSUMO_1	5.8429
CONSUMO_2	5.9343
CONSUMO_3	5.8496
CONSUMO_4	5.6616
CONSUMO_5	5.6411
CONSUMO_6	5.5625
CONSUMO_7	5.6460
CONSUMO_8	5.6017
CONSUMO_9	5.5811
CONSUMO_10	5.6143
CONSUMO_11	5.7228
CONSUMO_12	5.7362

O objetivo de aplicar *Clustering* é agrupar os pontos de maneira que os pontos em um cluster específico sejam semelhantes entre si e menos semelhantes aos pontos em outros *clusters*. Para tanto deixamos a cargo do algoritmo encontrar padrões nos dados e agrupá-los.

A abordagem usada foi a de Conectividade (*Connectivity*), nessa abordagem os pontos que são conectados ou imediatamente próximos um do outro são colocados na mesma classe (*cluster*).

O clustering espectral (*Spectral clustering*) é uma técnica que segue essa abordagem. Em *python*, o módulo *sklearn.cluster* reúne algoritmos populares de *cluster* não supervisionados, como o clustering espectral, usado neste artigo. Os parâmetros foram definidos da seguinte forma: o número de clusters foi determinado como 3; a afinidade definida como 'rbf', que constrói a matriz de afinidade usando um núcleo de função de base radial (RBF); atribuir_labels configurados para "discretizar"; e aleatório_state definido como 170.

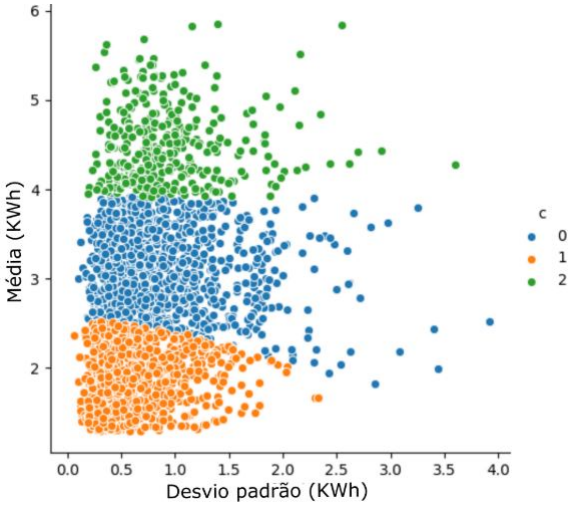

Figura 3. *Clusters* Gerados.

A figura 5 mostra o resultado da técnica de agrupamento utilizada. O resultado da aplicação do algoritmo é um *array* com valores de 0, 1 e 2 representando as classes criadas. O número de clientes por classe é apresentado na tabela 3. Percebe-se o baixo número de clientes na classe 2 em comparação com as classes 0 e 1, o que pode causar baixo índice de previsão em comparação com as outras classes. A tabela 4 apresenta os valores mínimos e máximos referentes a média de consumo e ao desvio padrão de cada classe.

Tabela 3. Número de Clientes por Classe.

Classes	N°de Clientes
0	966
1	881
2	336

Tabela 4. Valores Mínimos e Máximos (10^8 kWh).

Classes		CONSUMO_MEDIA	STD_VALUE
1	Min	1.3025	0.06598
	Max	2.5250	2.33325
0	Min	1.8233	0.10296
	Max	3.92	3.91468
2	Min	3.9208	0.18802
	Max	5.8483	3.6016

2.4 Análise dos Municípios Constituintes nas Classes Geradas

Como descrito anteriormente, no conjunto de dados, objeto de estudo deste artigo, um dos atributos (DESC_ MUNICIPIO) indica o município de cada cliente cadastrado na base de dados, no total tem-se 51 municípios, para representação gráfica os autores escolheram trabalhar com dez grandes regiões que englobam os 51 municípios, identificado pelo atributo DESC_COORDENADOR. A tabela 5 apresenta as dez regiões e o número de clientes cadastrados correspondente.

Para termos uma ideia do perfil de consumo por região (municípios), juntamos as informações do gráfico da figura 5 com as da tabela 5, gerando a figura 4.

Pela análise da figura 4 podemos concluir que as informações utilizadas até esta etapa não são suficientes para classificação do perfil de consumo dos clientes por região municipal: O perfil de consumo apresenta comportamento semelhante para as regiões. Um dos motivos para a geração das classes de classificação.

Tabela 5. Regiões e Número de Clientes.

Regiões	N° de Clientes
IBITINGA	496
MONTE ALTO	388
BARRETOS	269
BEBEDOURO	255
JABOTICABAL	178
BROTAS	172
SAO CARLOS	157
ARARAQUARA	127
MATAO	115
OLIMPIA	26

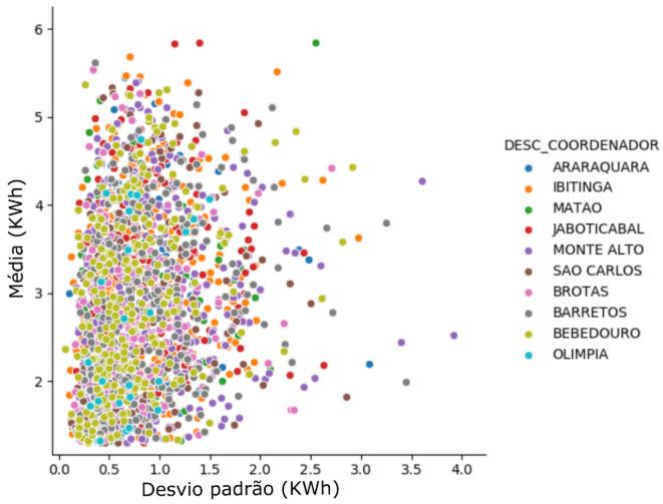

Figura 4. Consumo médio e desvio por grande Região.

3. METODOLOGIA

Nesta seção é apresentada a metodologia implementada para classificação do conjunto de dados apresentado na seção II.

Com dimensão final de 2183x15, são apresentados nesta seção os preparativos para a implementação do aprendizado de máquina e as técnicas utilizadas.

3.1 Random Forest

O método de classificação *Random Forest* foi desenvolvido por (Breiman 2001), onde o mesmo define o método como sendo uma combinação de árvores de decisões, onde cada árvore depende dos valores de um vetor amostrado aleatoriamente e com a mesma distribuição para todas as árvores.

Em outras palavras, dentro do vetor de características de cada árvore, são selecionados aleatoriamente alguns dos atributos que a mesma possui. Um vez feito isso, calcula-se a entropia apresentada por cada atributo e aquele que possuir a maior é escolhido para separar as classes naquela posição da árvore. A saída do classificador será aquela em que a classe foi retornada como a resposta pela maioria das árvores pertencentes à floresta (Guedes 2014).

Como mencionado, o método de *Random Forest* se utilizada de várias *Decision Trees*, combinando o resultado de classificação entre eles, conforme é possível ver na figura 5. Por isso, o método de *Random Forest* acaba se tornando muito mais poderoso comparado ao *Decision Tree*.

3.2 KNN

O método de *K-Nearest Neighbors* (KNN) armazena todas as amostras de treinamento (incluindo suas características) em um espaço de acordo com suas métricas sem processamento ou cálculo (Liu 2018). Quando o modelo recebe um objeto para ser previsto, coloca o novo objeto nesse espaço (também de acordo com as métricas). O modelo, então, faz previsões olhando para os vizinhos mais próximos do novo objeto. Geralmente, a previsão é o rótulo em que ocorre o máximo entre essas k amostras. O método KNN é um

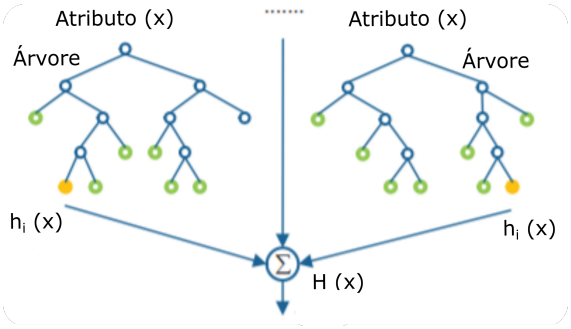

Figura 5. Método de *Random Forest* (Fonte: Zhao et al. 2015).

algoritmo que classifica todas as estatísticas disponíveis com base em uma medida de similaridade. Cada um dos dados de treinamento consiste em um conjunto de vetores e rótulos de classe associado a cada vetor. No caso mais simples, será + ou – (para classes positivas ou negativas). Porém, o KNN pode funcionar igualmente bem com um número arbitrário de classes (Thirumuruganathan 2010) que é o nosso caso. O algoritmo de classificação é realizado de acordo com as seguintes etapas (Beckmann et al. 2015):

1) Calcular a distância (geralmente euclidiana) entre um caso xi e todos os casos do conjunto de treinamento T;

2) Selecionar os k vizinhos mais próximos;

3) Os xi casos são classificados (rotulados) com a classe mais frequente entre os k vizinhos mais próximos. Também é possível usar a distância dos vizinhos para ponderar a decisão de classificação.

Ainda de acordo com Beckmann et al. (2015), o valor de k é dependente do *training-data*. Um pequeno valor de k significa que o ruído terá uma maior influência no resultado. Um grande valor torna computacionalmente caro e derrota a filosofia atrás do KNN: pontos próximos podem ter densidades ou classes semelhantes.

Para implementação do KNN, a seleção *sklearn.model* foi utilizada para dividir os dados de teste e treinamento, onde os valores de X e Y foram: CONSUMO 1,...,CONSUMO 12 e as matrizes de classificação, c (da etapa de clusterização), respectivamente. A divisão foi feita usando 0.8 dos dados para o treinamento e 0.2 para os testes, onde o estado aleatório foi definido pelo número 42, esse valor foi aleatoriamente definido. O código implementado encontra o melhor valor para k, considerando a variação de 30, calculando a acurácia e precisão de cada valor de k e o erro correspondente.

3.3 Métricas de Avaliação

A matriz de confusão, tabela 6, foi usada para calcular a precisão baseada nas classes corretas e incorretas. 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 (Beckmann et al. 2015) 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 6. 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.

$$\textit{F-1 score} = 2 \times \frac{\text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{5}$$

4. RESULTADOS E DISCUSSÕES

Nesta seção, são apresentados os resultados das simulações e os comentários referentes aos modelos *Random Forest* e KNN.

4.1 Random Forest

O resultado para a previsão das amostras de CONSUMO em relação às classes geradas no processo de *clustering* é apresentado no relatório impresso ao fim da simulação do algoritmo, figura 6.

```
Acurácia do classificador RF nos dados de treino: 1.00
Acurácia do classificador RF nos dados de teste: 0.97

              Precisão  Revocação  f-1 score  Suporte
0                 0.96       0.98       0.97      201
1                 0.98       0.98       0.98      173
2                 0.98       0.92       0.95       63

acurácia                               0.97      437
macro média       0.97       0.96       0.97      437
média ponderada   0.97       0.97       0.97      437

Pontuação de acurácia: 0.9702517162471396
Tempo: 2.503961299999901 seconds
```

Figura 6. Relatório de Simulação: modelo *Random Forest*.

A pontuação da previsão para a acurácia, considerando somente as amostras de teste (20% do conjunto de dados) foi de 0,97 (97%). Para os dados de treino (80% do conjunto de dados) a acurácia do modelo foi de 100%.

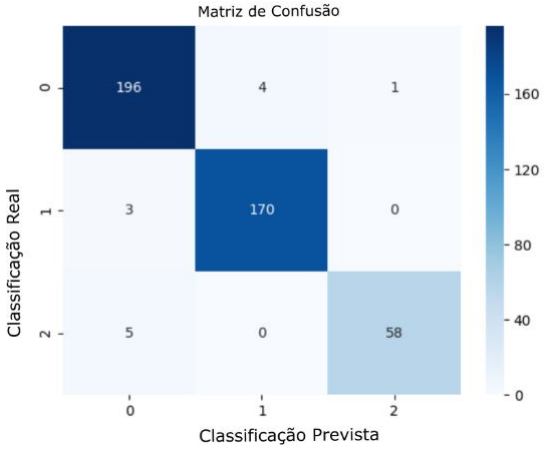

Figura 7. Matriz de Confusão do modelo *Random Forest*.

A figura 7 apresenta a matiz de confusão (*Confusion matrix*) onde podemos verificar pela diagonal principal as previsões corretas. Considerando a primeira linha da matriz verificamos que do total de 201 amostras da classe 0: 196 foram corretamente classificadas; 4 foram erroneamente classificadas como sendo da classe 1; 1 amostra foi erroneamente classificada como sendo da classe 2. Seguindo o mesmo raciocínio, do total de 173 amostras pertencentes à classe 1: 170 foram corretamente classificadas; 3 foram erroneamente classificadas como pertencentes à classe 0. Já para a classe 2, que apresenta menor número de amostras por ser uma classe com menor número de clientes, do total de 63 amostras: 58 foram classificadas corretamente; 5 foram classificadas como pertencentes à classe 0.

No geral, este modelo apresentou ótimos resultados de classificação com tempo de processamento de 2,5 segundos, aproximadamente. Avaliando os resultados das métricas de avaliação: precisão e revocação, percebe-se que estão próximas do ponto ideal de 1.00.

4.2 KNN

O resultado para as previsões usando o modelo KNN é apresentado na figura 8. Já a figura 9 apresenta a matriz de confusão do modelo KNN.

A melhor acurácia foi de 0.9748283752860412 com k= 12

Relatório de Classificação:

	Precisão	Revocação	f-1 score	Suporte
0	0.98	0.94	0.96	201
1	0.93	1.00	0.96	173
2	1.00	0.95	0.98	63
acurácia			0.96	437
macro média	0.97	0.96	0.97	437
média ponderada	0.97	0.96	0.96	437

Tempo: 3.0851744999999937 segundos

Figura 8. Relatório de Simulação: modelo *KNN*.

Conforme mostrado no relatório da figura 8, a melhor acurácia de classificação encontrada variando k de 1 a 30 é de aproximadamente 0,97 para k igual a 12. Os valores de precisão e revocação também apresentam equilíbrio e estão próximos da relação 1.00x1.00, *f1-score*, também está próximo de 1.00 para todas as classes (indicando o bom balanço de precisão e revocação).

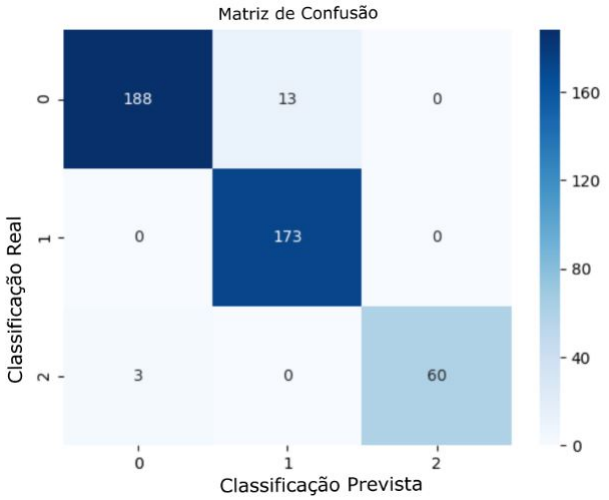

Figura 9. Matriz de Confusão para k=12 do modelo *KNN*.

Se verificarmos os valores da diagonal principal da matriz de confusão da figura 9, temos as amostras que foram corretamente classificadas, valores de verdadeiro positivo, do total de 437 amostras, 421 foram classificadas corretamente. Novamente, analisando a diagonal principal da matriz, para a classe 0:do total de 201 amostras, 188 foram corretamente classificadas e 13 foram erroneamente classificadas como sendo da classe 1; para a classe 1: todas as 173 amostras foram corretamente classificadas; para a classe 2: do total de 63 amostras, 60 foram corretamente classificadas e 3 classificadas como pertencentes à classe 0. O tempo de processamento do modelo KNN de classificação foi de 3.08 segundos, um tempo superior ao modelo de *Random Forest*.

4.3 Exemplo de Aplicação do Modelo Random Forest

Para exemplo de detecção de clientes fora da sua classe definida na seção 2.3 de *clustering*, após implementação do modelo de *Random Forest* entrou-se com uma lista de dados contendo o código do cliente (atributo COD_INSTALACAO) e os 12 valores de consumo (atributos CONSUMO_X), no formato apresentado na figura 10. O primeiro valor refere-se ao código do cliente e os outros 12 valores ao consumo de energia mensal (em 10^8 kWh). A saída é a classe prevista pelo algoritmo e a classe definida se seção 2.3 para a instalação de entrada, figura 11.

```
Entre com a lista de valores = COD_INSTALACAO, CONCUMO_1,...,
CONSUMO_12:82287 4.31 4.65 4.37 4.15 4.5 3.9 4.65 2.75 4.82 3.82 4.91
4.13
```

Figura 10. Dados de Entrada para classificação.

Pela figura 11 verifica-se que a classificação do cliente referente aos consumos de entrada é a mesma classe à que pertence, resultado positivo. Para verificarmos a detecção de cliente fora de sua classe original foram alterados os consumos dos três primeiros meses e implementado novamente o modelo.

```
Entre com a lista de valores  = COD_INSTALACAO, CONCUMO_1,...,
CONSUMO_12:82287 4.31 4.65 4.37 4.15 4.5 3.9 4.65 2.75 4.82 3.82 4.91
4.13
Classe Prevista  : [2]
COD_INSTLACAO: 82287
Classe Original  : 2
```

Figura 11. Dados de Saída para classificação.

A figura 12 apresenta o resultado para os novos dados de entrada, com alteração de consumo para o mesmo cliente. Verifica-se que com os três dados de consumo alterados o cliente pertencente a classe 2 foi classificado como pertencente a classe 0, resultado negativo.

```
Entre com a lista de valores  = COD_INSTALACAO, CONCUMO_1,..., CONSUMO_12:
82287 3.31 3.65 3.37 4.15 4.5 3.9 4.65 2.75 4.82 3.82 4.91 4.13
Classe Prevista  : [0]
COD_INSTLACAO: 82287
Classe Original  : 2
```

Figura 12. Novos Dados de Saída para classificação.

5. CONCLUSÃO

Neste artigo foram apresentados dois tipos de classificadores,*RF* e *KNN*, com o objetivo de agrupar consumidores em *clusters* que futuramente podem ser úteis na identificação do desvio de consumo do cliente de acordo com sua classe de identificação. Os dois modelos apresentaram acurácia de 97% nas classificações, contudo o modelo *KNN* foi o que apresentou melhores resultados para as classes 1 e 2, enquanto que o modelo *RF* apresentou os melhores resultados para a classe 0.

Perspectivas para projetos futuros seriam melhoras na precisão de classificação, equilíbrio entre o número de consumidores nas classes, melhor delimitação no processo de agrupamento, uso de mais dados históricos de consumo, neste trabalho foram utilizados apenas dados de um ano, e, o uso de *clusters* para classificar perfis de consumo dos consumidores como uma ferramenta de identificação de desvios do consumo, identificando possíveis perdas não técnicas.

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 financiado em parte pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código Financeiro 001 e pelo 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

Agüero, J. R. (2012). Improving the efficiency of power distribution systems through technical and non-technical losses reduction, Em *PES TD 2012*, Montevidéu, Uruguai.

Ammar, A. (2015) A Decision Tree Classifier for Intrusion Detection Priority Tagging, *Journal of Computer and Communications*, vol.3, pp. 52-58.

Beckmann, M., Ebecken, N., e Lima, B. P. (2015) A KNN Undersampling Approach for Data Balancing, *Journal of Intelligent Learning Systems and Applications*, vol.7, pp. 104-116.

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

Evaldt, M. C. (2014). Uma Metodologia para a identificação de Perdas Não Técnicas em Grandes Consumidores Rurais, Dissertação de Mestrado, Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Santa Maria, Santa Maria, Brasil.

Guedes, A. R. M; Guimaraes, V. L. (2014). *Sistemas de Reconhecimento baseado em Random Forest para caracteres de captchas.*, Disponível em: http://www.decom.ufop.br/menotti/rp142/trab/trab1-dp2-artigo.pdf, Ouro Preto, 2014.

James, G., Witten, D., Hastien, T., e Tibshirani, R. (2013). *An Introduction to Statistical Learning*, Capítulo 8. Springer, Nova York, Estados Unidos.

Liu, R. (2018). Machine Learning Approaches to Predict Default of Credit Card Clients, *Modern Economy*, vol. 9, pp. 1828-1838.

Madrigal, M., Rico, J. J., e Uzcategui, L. (2017). Estimation of Non-Technical Energy Losses in Electrical Distribution Systems, *IEEE Latin America Transactions*, vol. 15 (num. 8), pp. 1447-1452.

Thirumuruganathan, S. (2010). A Detailed Introduction to K-Nearest Neighbor (KNN) Algorithm.

Warwick, K., Ekwue, A., e Aggarwal, R. (1997). *Artificial Intelligence Techniques in Power Systems*, Capítulo 1. Institution of Electrical Engineers (IEE), Londres, Reino Unido.

Weiss, G. M., e Provost, F. (2001). The Effect of Class Distribution on Classifier Learning: An Empirical Study, Technical Report MLTR-43, Department of Computer Science - Rutgers University, New Brunswick, Estados Unidos.

Zhang, W. (2017). Machine Learning Approaches to Predict Default of Credit Card Clients, *Journal of Financial Risk Management*, vol. 6, pp. 364-374.

Zhang, Y., e Wang, J. (2015). GEFCom2014 probabilistic solar power forecasting based on k-nearest neighbor and kernel density estimator, Em *2015 IEEE Power Energy Society General Meeting*, Denver, Estados Unidos.

Zhao, Y., e Ma, X. (2017). *Study on credit evaluation of electricity users based on random forest.* 2017 Chinese Automation Congress (CAC), Qingdao, China.