Centro de Engenharia Elétrica e Informática

Curso de Pós-Graduação em Engenharia Elétrica

HELEM MONYELLE DE MÉLO ALVES

ANÁLISE DA CONTRIBUIÇÃO DE ATRIBUTOS DERIVADOS DO HISTÓRICO DE CONSUMO PARA A DETECÇÃO DE PERDAS NÃO TÉCNICAS

Campina Grande - PB.

Julho - 2019

UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
Centro de Engenharia Elétrica e Informática
Programa de Pós-Graduação em Engenharia Elétrica

ANÁLISE DA CONTRIBUIÇÃO DE ATRIBUTOS DERIVADOS DO HISTÓRICO DE CONSUMO PARA A DETECÇÃO DE PERDAS NÃO TÉCNICAS

HELEM MONYELLE DE MÉLO ALVES

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Campina Grande como parte dos requisitos necessários para obtenção do grau de Mestre em Engenharia Elétrica.

Área de Concentração: Processamento de Energia

Prof. Edson Guedes da Costa
Prof. Jalberth Fernandes de Araújo
Orientador(es)

Campina Grande - PB.

Julho - 2019

A474a Alves, Helem Monyelle de Mélo.

Análise da contribuição de atributos derivados do histórico de consumo para a detecção de perdas não técnicas / Helem Monyelle de Mélo Alves. – Campina Grande, 2019.

64 f. : il. color.

Dissertação (Mestrado em Engenharia Elétrica) – Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2019.

"Orientação: Prof. Dr. Edson Guedes da Costa, Prof. Dr. Jalberth Fernandes de Araújo".

Referências.

1. Processamento de Energia. *2. Correlation Based Feature Selection.* 3. Mineração de Dados. 4. Perdas Não Técnicas. 5. Redes Neurais Artificias. 6. *Relief.* 7. Seleção de Atributos. I. Costa, Edson Guedes da. II. Araújo, Jalberth Fernandes de. III. Título.

CDU 621.31(043)

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECÁRIA SEVERINA SUELI DA SILVA OLIVEIRA CRB-15/225

"ANÁLISE DA CONTRIBUIÇÃO DE ATRIBUTOS DERIVADOS DO HISTÓRICO DE CONSUMO PARA A DETECÇÃO DE PERDAS NÃO TÉCNICAS"

HELEM MONYELLE DE MÉLO ALVES

DISSERTAÇÃO APROVADA EM 19/07/2019

EDSON GUEDES DA COSTA, D.Sc., UFCG
Orientador(a)

JALBERTH FERNANDES DE ARAÚJO, Dr., UFCG
Orientador(a)

BENEMAR ALENCAR DE SOUZA, D.Sc., UFCG
Examinador(a)

GEORGE ROSSANY SOARES DE LIRA, D.Sc., UFCG
Examinador(a)

CAMPINA GRANDE - PB

Dedico este trabalho a Deus e ao meu avô Pedro José de Mélo (in memoriam).

AGRADECIMENTOS

Agradeço a Deus que em todos os momentos da minha vida esteve ao meu lado e é a razão de eu estar onde estou. Agradeço a meus pais, minha irmã e meus avós que sempre me apoiaram emocionalmente e financeiramente. Agradeço a meus amigos por todas as palavras de incentivo que me foram tão necessárias durante o período em que estive envolvida com esta dissertação. Agradeço aos professores Edson e Jalberth por terem aceitado a missão de me orientar e por todos os conselhos e paciência que tão gentilmente me concederam. Por fim, agradeço o apoio acadêmico de Iago Batista e João Pedro.

“Soli Deo Gloria”

RESUMO

As perdas não técnicas são resultantes majoritariamente do consumo irregular de energia elétrica, por meio de fraudes ou furtos. A redução delas é um dos principais objetivos das concessionárias de distribuição de energia elétrica. Atualmente, as concessionárias têm utilizado sobretudo inspeções *in loco* para identificação de clientes irregulares. Entretanto, as inspeções frequentemente estão associadas a um alto custo e uma baixa eficácia. Neste sentido, as concessionárias têm recorrido a técnicas de mineração de dados com o intuito de aumentar a assertividade na seleção de clientes irregulares para inspeções. Neste trabalho, é analisada a contribuição de atributos derivados do histórico de consumo de energia elétrica na detecção de perdas não técnicas, utilizando técnicas de mineração de dados. Para isto, são criados novos atributos a partir dos dados de consumo, considerando características de sazonalidade, informações estatísticas, variações mensais, taxas de queda e informações do consumo no domínio da frequência. Para definir quais os melhores atributos (considerando-se os atributos originais e os atributos criados posteriormente) são utilizados os métodos para seleção de atributos *Correlation Based Feature Selection* e *Relief*. Na sequência, o algoritmo de Redes Neurais Artificias do tipo *mutilayer perceptron* é aplicado para classificar os clientes da base de dados entre regulares e irregulares a partir dos atributos selecionados. A partir dos resultados obtidos, verificou-se que a adição de novos atributos contribuiu para o aumento da assertividade do algoritmo de redes neurais artificias, proporcionando um ganho de aproximadamente 10 pontos percentuais, o que pode representar uma economia significativa no dinheiro gasto pelas concessionárias com inspeções improcedentes. Com isso, pode-se destacar que a análise de atributos pode contribuir para a redução de custos associados a detecção de perdas não técnicas ao melhorar a assertividade na identificação de potenciais clientes irregulares.

Palavras-chave: *Correlation Based Feature Selection*, mineração de dados, perdas não técnicas, Redes Neurais Artificias, *Relief*, seleção de atributos.

ABSTRACT

Non-technical losses are mainly caused by the electricity irregular consumption due to fraud or theft. Their reduction is one of the main objectives of electricity distribution companies. Currently, companies have mainly used in loco inspections to identify irregular customers. However, these inspections are often associated with high costs and low effectiveness. Then, many companies have resorted to data mining techniques in order to increase assertiveness in the selection of irregular customers for inspections, using cadastral information such as class, supply voltage, type of connection and, mainly, historical consumption data. In this work, the contribution of attributes derived from the consumption electric energy history in non-technical losses detection using data mining techniques is analyzed. Therefore, new attributes are created from the consumption data, using seasonality characteristics, statistical information, monthly consumption variations, fall rates and consumption information in the frequency domain. In order to define the best attributes considering the original attributes and the attributes created subsequently, the attribute selection methods Correlation Based Feature Selection and Relief are used. Afterwards, the multilayer perceptron artificial neural networks algorithm is applied to classify the database clients between regular and irregular using the selected attributes. From the results, it was verified that the new attributes addition contributed to increase the artificial neural networks assertiveness, providing approximately a 10 percentage point gain, which can represent significant savings on the money spent by concessionaires with not assertive inspections. Therefore, it can be emphasized that the attributes analysis presented in this work can be used to reduce costs associated with non-technical losses detection by improving assertiveness in the potential irregular client's identification.

Key-words: Correlation Based Feature Selection, data mining, non-technical losses, Artificial Neural Networks, Relief, attribute selection.

.

SUMÁRIO

1 INTRODUÇÃO

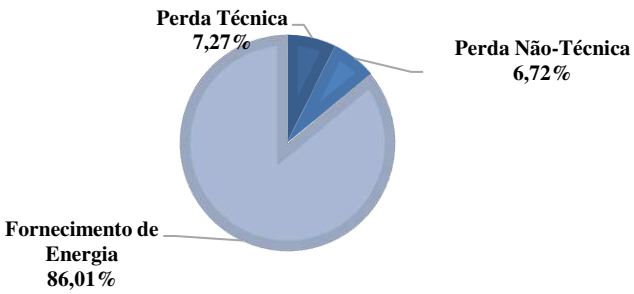
As perdas no sistema elétrico são equivalentes à energia elétrica comprada pelas concessionárias de energia que não é faturada a seus consumidores. Elas podem ser classificadas em dois tipos: perdas técnicas e não técnicas.

Segundo definição dada pela Agência Nacional de Energia Elétrica (ANEEL, 2018), as perdas técnicas estão relacionadas principalmente à: transformação de energia elétrica em energia térmica nos condutores (efeito Joule), perdas nos núcleos dos transformadores, e perdas dielétricas, além disso, também podem ser inseridas no sistema elétrico as perdas decorrentes da corrente de fuga volumétrica e superficial, como também as perdas por ionização (BOGORODITSKY, 1981). Elas podem ser reduzidas por meio de: investimentos na construção de novas redes de distribuição e transmissão de energia elétrica, repotencialização e elevação dos níveis de tensão das redes, correta manutenção e melhoria dos equipamentos instalados nas redes de distribuição e transmissão de energia elétrica (COMETTI, 2004).

Já as perdas não técnicas são calculadas como a diferença entre as perdas totais e as perdas técnicas. Portanto, elas são equivalentes a todas as demais perdas associadas à distribuição de energia elétrica (ANEEL, 2018). Furtos (ligações clandestinas, desvios diretos das redes de distribuição de energia elétrica) ou fraudes de energia (adulterações no medidor) representam as principais fontes de perdas não técnicas. Outras possíveis fontes são erros de leitura, medição e faturamento ou defeito nos equipamentos de medição (ANEEL, 2018).

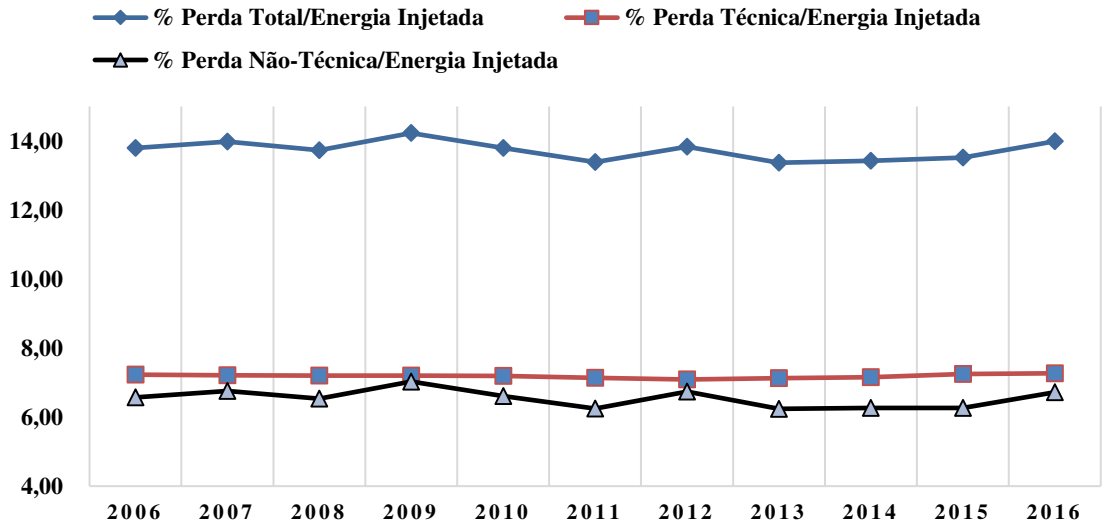
De acordo com relatório divulgado pela ANEEL em 2018, as perdas elétricas nas redes de distribuição no Brasil no ano de 2016 corresponderam à aproximadamente 14% da energia elétrica nelas injetada. Isto significa dizer que, para cada 1 kWh de energia faturada, as concessionárias tiveram que adquirir, em média, 1,16 kWh. As perdas não técnicas representam cerca de R$ 12,3 bilhões nas tarifas, o que equivale a 8% da receita do setor elétrico (R$ 156 bilhões) ou 29% da receita das distribuidoras (R$ 42 bilhões). No gráfico da Figura 1 é apresentada a quantidade percentual de perdas técnicas e não técnicas no montante total de energia injetada no ano de 2016 e, no gráfico da Figura 2, a evolução das perdas entre os anos de 2006 e 2016.

Figura 1: Participação percentual das perdas no total da energia elétrica injetada nos sistemas de distribuição no ano de 2016.



Fonte: adaptado de ANEEL (2018).

Figura 2: Evolução das perdas no setor de distribuição no Brasil com relação à energia elétrica injetada.



Fonte: adaptado de ANEEL (2018).

Como pode ser constatado a partir das informações do relatório da ANEEL e dos gráficos apresentados, as perdas não técnicas são um problema grave e recorrente para as concessionárias de energia elétrica no Brasil. De fato, desde 2013, a participação percentual das perdas não técnicas na energia injetada está em crescimento, alcançando o elevado valor de 6,72% no ano de 2016 (o maior valor registrado desde 2012), o que corrobora a necessidade da sua detecção, mitigação e prevenção.

Diversas medidas têm sido e vem sendo empregadas pelas concessionárias de distribuição de energia elétrica para a detecção e mitigação das perdas não técnicas. Algumas concessionárias utilizam técnicas para blindagem da rede de distribuição e medição remota do consumo de energia elétrica de seus clientes. É também comum a realização de campanhas de conscientização da sociedade, com o objetivo de divulgar os

riscos e as consequências criminais do furto de energia elétrica e o incentivo a realização de denúncias.

Além das medidas supracitadas, tarifas de energia elétrica com valores mais baixos são estabelecidas para “clientes de baixa renda”, que são consumidores em situação de vulnerabilidade social. Segundo Huback (2018), as altas tarifas implicam em altos índices de inadimplência e furto de energia em diversas regiões do país, especialmente em áreas de baixa renda. Com a redução do valor da tarifa, espera-se que os consumidores intitulados “clientes de baixa renda” sejam incentivados à prática de um consumo regular de energia elétrica, e não recorram a fraudes e/ou a furtos.

Embora as medidas descritas sejam relevantes, atualmente a principal ação adotada pelas concessionárias de distribuição de energia elétrica para a detecção das perdas não técnicas consiste na atividade de inspeção a partir da análise de dados de consumo dos clientes. Geralmente, a presença de queda de consumo ou consumo no mínimo da tarifa são indícios de irregularidade.

Contudo, a seleção de alvos para inspeção não é trivial. Ela envolve a extração de informações das bases de dados das concessionárias, que possuem dados de milhões de clientes, o que contribui para tornar o seu uso menos frequente do que o necessário. Outro problema com o uso de informações cadastrais consiste na definição de regras para seu tratamento e aplicação na detecção e mitigação das perdas não técnicas. As regras geralmente tendem a saturar, ou perder a sua efetividade. À medida que são aplicadas, o número de possíveis alvos enquadrados por elas diminui (COMETTI, 2004).

Para melhorar, ou aumentar, a efetividade das regras de tratamento, o uso de técnicas de mineração de dados e aprendizado de máquina para auxílio na detecção de perdas não técnicas no setor elétrico é importante por adicionar inteligência artificial aos processos de tratamento e análise de dados cadastrais. Com algoritmos de classificação como Árvores de Decisão, *Support Vector Machine*, Redes Neurais Artificiais e Redes *Bayesianas*, é possível criar modelos para identificação de clientes com perdas não técnicas baseando-se em informações cadastrais e no histórico de consumo (MESSINIS, 2018; BASTOS, 2011). Os modelos tendem a ter maior durabilidade em comparação às regras arbitrárias e com possibilidade para introdução de uma variedade e quantidade maior de clientes.

De fato, diversas técnicas de mineração de dados têm sido utilizadas para solução de problemas envolvendo bancos de dados. Elas podem ser utilizadas para identificação de pacientes com câncer, estudos sobre o comportamento de ações da bolsa de valores, e

até mesmo em campanhas eleitorais. Uma característica importante presente em alguns estudos que utilizam técnicas de mineração de dados é a adição de novos atributos aos dados originais e a determinação dos principais atributos do conjunto de dados resultante.

A identificação dos melhores atributos pode ser realizada por meio da aplicação de métodos para seleção de atributos. Os métodos mais empregados envolvem a utilização de medidas estatísticas como a correlação, para ordenar os atributos, e o emprego de procedimentos para comparar e avaliar os resultados obtidos a partir de diferentes combinações de atributos (até que se identifique o conjunto de atributos com a maior contribuição para a solução do problema em estudo). De acordo com Karegowda (2010), a seleção dos melhores atributos pode contribuir significativamente para o aumento da assertividade de estudos de mineração de dados envolvendo a classificação de um conjunto de dados.

No que diz respeito ao problema da detecção de perdas não técnicas, é possível que informações importantes do perfil dos clientes possam ser obtidas com a criação de novos atributos a partir do histórico de consumo desses clientes. Entretanto, há que se ter maior cuidado com relação à adição de atributos. Quanto maior o número de atributos, mais demorada pode ser a aplicação de técnicas de mineração de dados e há chances de que a capacidade de classificação dos algoritmos utilizados seja prejudicada. É preciso certificar-se de que os melhores atributos estejam sendo utilizados.

Diante do exposto, este trabalho tem como objetivo analisar a contribuição de atributos derivados do histórico de consumo para detecção das perdas não técnicas.

Para isto, em uma primeira etapa, são criados atributos a partir dos dados de consumo de energia elétrica de clientes de uma concessionária de distribuição de energia elétrica com atuação no Brasil. Na sequência, os melhores atributos do banco de dados resultante são identificados a partir da aplicação de métodos de seleção de atributos.

Após a identificação dos melhores atributos disponíveis, um algoritmo de Redes Neurais Artificias é utilizado para determinar quais clientes dentre o número total de clientes analisados estão contribuindo para o aumento das perdas não técnicas.

1.1 OBJETIVOS

O objetivo geral deste trabalho é a análise da contribuição de atributos derivados do histórico de consumo de energia elétrica de clientes. O histórico de consumo adveio de uma concessionária de distribuição de energia elétrica, com atuação no Brasil, para a detecção das perdas não técnicas.

Além disso, este trabalho possui os seguintes objetivos específicos:

- Criar atributos a partir do histórico de consumo de energia elétrica de clientes de uma concessionária de distribuição de energia elétrica;
- Analisar e identificar, a partir do uso de seleção de atributos, os melhores atributos dentre o conjunto de atributos original e os atributos criados, conforme descrito no objetivo específico anterior;
- Classificar os clientes de uma concessionária de distribuição de energia elétrica entre irregulares (clientes que praticam furto de energia elétrica, ou com medição de energia elétrica incorreta) ou regulares;
- Identificar a combinação entre algoritmo de classificação (Redes Neurais Artificiais) e dados de entrada (os dados de consumo originais, ou os melhores atributos identificados a partir dos métodos de seleção de atributos) que resulta no melhor índice de assertividade na detecção de perdas não técnicas.

1.2 ORGANIZAÇÃO DO TEXTO

Este trabalho está organizado em sete capítulos, descritos a seguir.

No Capítulo 2 é realizado o embasamento teórico sobre técnicas de mineração de dados. Em seguida, os parâmetros estatísticos que são utilizados para criação de novos atributos são descritos, bem como os métodos para seleção de atributos que são utilizados no trabalho, o comportamento de algoritmos de classificação em estudos de mineração de dados, as definições do algoritmo de Redes Neurais Artificiais e a definição das métricas que são utilizadas para avaliação dos resultados.

No Capítulo 3 uma revisão bibliográfica sobre o uso de técnicas de mineração de dados para a detecção de perdas não técnicas é apresentada. Trabalhos que propuseram a

criação de novos atributos derivados dos dados de consumo e a utilização da representação no domínio da frequência dos dados de consumo são discutidos. Além disso, são destacadas as contribuições deste trabalho em relação a trabalhos anteriores.

No Capítulo 4 é apresentada a metodologia empregada para a avaliação da contribuição dos atributos derivados do histórico de consumo na detecção de perdas não técnicas, destacando as características do banco de dados usado. Neste capítulo, também são descritos os procedimentos adotados para a criação dos atributos, para a aplicação dos métodos de seleção de atributos e para a classificação dos clientes.

No Capítulo 5 são apresentados os resultados obtidos neste trabalho.

No Capítulo 6 são apresentadas as conclusões e as sugestões de trabalhos futuros, com o intuito de continuar a linha de pesquisa apresentada neste trabalho.

No Capítulo 7 são apresentados os artigos que foram publicados, aceitos para publicação e submetidos.

Por fim, são apresentadas as referências utilizadas.

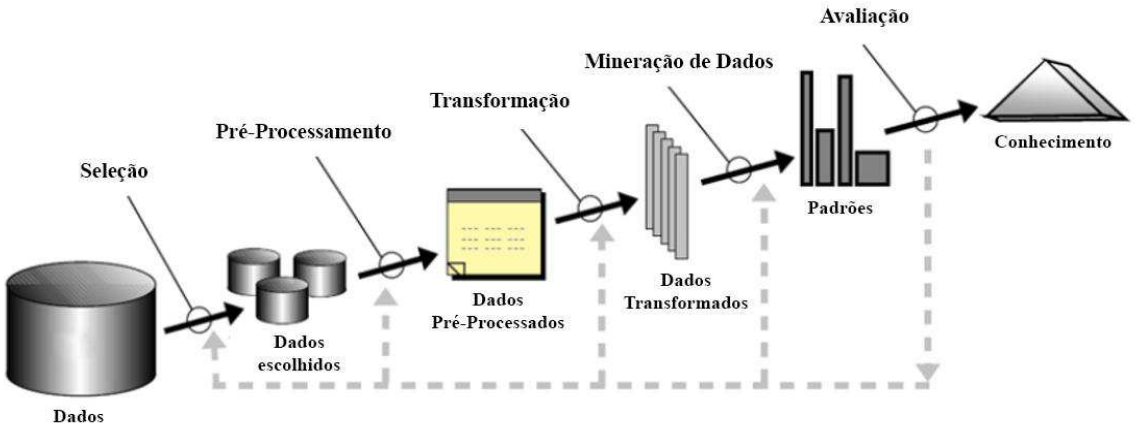
2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo é apresentada a fundamentação teórica necessária ao entendimento do tema proposto. Assim, são apresentadas informações sobre o processo de mineração de dados, os métodos de classificação, o algoritmo de redes neurais artificiais, algumas métricas costumeiramente utilizadas para a avaliação de estudos de classificação utilizando mineração de dados, alguns parâmetros que são utilizados para obtenção de atributos derivados do histórico de consumo e os métodos *Correlation Based Feature Selection* e *Relief* aplicados para a seleção de atributos.

2.1 MINERAÇÃO DE DADOS

Muitos estudiosos compreendem a mineração de dados como uma etapa de um processo complexo e iterativo chamado de *Knowledge Discovery in Databases* ou, em tradução livre, Descoberta de Conhecimento nas Bases de Dados (KDD). Segundo Faayad (1996), o KDD é um processo necessário para solucionar o problema da sobrecarga de dados causado pela chamada "Era da Informação". Na Figura 3 pode-se observar uma representação das etapas do processo de KDD.

Figura 3: Fluxograma com as etapas do processo de KDD.



Fonte: Adaptado de Faayad (1996).

A primeira etapa do processo de KDD, segundo Faayad (1996), é chamada de seleção de dados, e requer o conhecimento do domínio do problema para que o melhor

conjunto de dados seja escolhido. Por exemplo, para detectar a presença de câncer em pacientes, é necessário, no mínimo, ter dados com informações clínicas dos pacientes. Enquanto que informações sobre as contas bancárias desses pacientes são desnecessárias. Assim, a seleção adequada de dados significa a utilização mais objetiva das informações disponíveis e a eliminação de informações que não irão ajudar na resolução do problema estudado.

Na sequência, a etapa de pré-processamento dos dados deve ser realizada visando o tratamento de valores desconhecidos e discrepantes que podem prejudicar a análise. Nessa etapa, também é recomendável a integração dos dados, utilizando, por exemplo, técnicas de normalização. Uma técnica de normalização que pode ser aplicada é a Sigmoidal (HANN, 2000), a qual é apresentada na Equação (1). Com a utilização dessa técnica, pode ser garantido que todos os dados estejam no intervalo entre 0 e 1:

$$y_0 = \frac{1}{1 + e^{-\frac{y-\bar{y}}{\sigma_y}}}, \qquad (1)$$

em que y é o valor original, $\bar{y}$ e σ_y são a média e o desvio padrão do conjunto de dados respectivamente, e y_0 é o valor normalizado.

Na etapa de transformação de dados, os dados pré-processados necessitam ser submetidos ao processo de redução, permitindo, assim, o aumento da eficiência do processo com uma base menor e mais consistente. A transformação de dados pode ser feita utilizando-se mecanismos para representação eficiente dos dados. Alguns mecanismos são: a redução da quantidade de atributos (restando apenas os realmente necessários), a redução do conjunto de dados usado para treinamento, ou a utilização da técnica de *under-sampling*, que consiste na retirada aleatória de algumas amostras da classe dominante da base de dados para evitar a ocorrência de enviesamento amostral (QUEIROGA, 2005).

Finalmente, tem-se a etapa de mineração de dados, que consiste na aplicação de algoritmos para identificação e reconhecimento de padrões em um banco de dados com o objetivo de extrair informações deste banco de dados. Nesta etapa, os algoritmos que mais se adéquem ao problema devem ser selecionados, mesmo que o processo de escolha exija um longo período de testes. Além disto, a utilização de dois ou mais algoritmos de forma integrada não é incomum, pois proporciona um ganho na confiabilidade do método. Exemplos de algoritmos utilizados, de forma isolada ou em conjunto são: Redes Neurais

Artificiais, Regras de Indução, Árvores de Decisão, Redes Probabilísticas, Redes *Bayesianas*, *Support Vector Machine* e *Random Forests*.

O número de áreas do conhecimento que utilizam técnicas de mineração de dados em seu dia-a-dia tem crescido gradativamente ao longo dos anos, sobretudo devido à popularização de banco de dados e informações cadastrais. As tecnologias computacionais empregadas na mineração de dados permitem identificar e extrair informações úteis e, assim, classificar em níveis diferentes ou aglutinar dados com características similares.

Após a etapa de mineração de dados, tem-se a etapa de avaliação, na qual os resultados obtidos são avaliados. É importante destacar que em técnicas para mineração de dados os resultados podem ser utilizados para retroalimentação do processo com o intuito de obter resultados mais consistentes. Além disso, deve-se ressaltar que a mineração de dados pode ser utilizada para várias aplicações, como predição, classificação, clusterização e associação. Em estudos para detecção de fraudes é mais comum utilizar a classificação (MESSINIS, 2018).

No próximo tópico são apresentados mais detalhes sobre técnicas de classificação em mineração de dados.

2.2 TÉCNICAS DE CLASSIFICAÇÃO EM MINERAÇÃO DE DADOS

As técnicas de classificação em mineração de dados podem ser de dois tipos: supervisionadas e não-supervisionadas. Elas são comumente usadas para prever valores de variáveis do tipo categóricas (variáveis que contém um número finito de categorias ou grupos). A partir delas pode-se, por exemplo, criar um modelo que classifica os clientes de um banco como especiais ou de risco; ou um laboratório pode usar sua base histórica de voluntários e verificar em quais indivíduos uma nova droga pode ser melhor aplicada. Em ambos os cenários, um modelo é criado para classificar a qual categoria um certo registro pertence, por exemplo: especial ou de risco e voluntários A, B ou C.

Na técnica de classificação por aprendizado supervisionado, algoritmos são treinados utilizando exemplos rotulados, como uma entrada em que a saída desejada é conhecida. Por exemplo, um equipamento pode ter pontos de dados rotulados como “F” (falha) ou “E” (exatidão). O algoritmo de aprendizagem recebe um conjunto de entradas com as saídas corretas correspondentes e “aprende” ao comparar a saída real com as

saídas corretas para encontrar padrões. O algoritmo, então, modifica o modelo, utilizando padrões para prever os valores de rótulos em dados adicionais não-rotulados. O desempenho do algoritmo é verificado na etapa de teste, quando é verificada a quantidade de dados não rotulados classificada corretamente. A aprendizagem supervisionada é comumente utilizada em aplicações nas quais dados históricos preveem eventos futuros prováveis. Por exemplo, ela pode antecipar quando transações via cartão de crédito são passíveis de fraude ou qual segurado tende a reivindicar sua apólice (MACHADO, 2018).

Já a técnica de classificação por aprendizado não supervisionado é utilizada majoritariamente em dados que não possuem rótulos históricos. A resposta correta não é fornecida durante a análise. O algoritmo deve descobrir o que está sendo mostrado a partir da identificação de padrões. O objetivo é explorar os dados e encontrar alguma estrutura dentro deles.

A aprendizagem não supervisionada funciona bem com dados transacionais. Por exemplo, ela pode identificar segmentos de clientes com atributos similares que são tratados de modo igualmente similar em campanhas de *marketing*; ou ela pode encontrar os atributos principais que separam segmentos distintos de clientes (MACHADO, 2018).

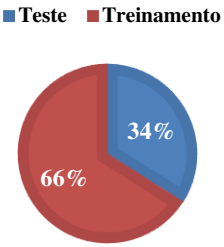
Em estudos para detecção de perdas não técnicas é possível a utilização tanto de técnicas de classificação por aprendizado supervisionado quanto de técnicas de classificação por aprendizado não supervisionado, embora a primeira seja mais comum na bibliografia. Assim, na próxima subseção são explicados conceitos importantes necessários à etapa de separação dos dados para criação das bases de treinamento e teste em técnicas de classificação por aprendizado supervisionado.

2.2.1 Estratégias para Separação de Bases

Como já discutido, na técnica de classificação por aprendizado supervisionado, é necessário que informações sobre o comportamento dos dados sejam previamente fornecidas ao algoritmo utilizado na classificação. Sendo assim, deve-se separar uma parte dos dados para treinamento do algoritmo e outra parte para teste. As estratégias de separação mais utilizadas em técnicas de classificação para mineração de dados são:

- *Percentage split*: nesta técnica, o banco de dados é separado aleatoriamente em duas partes, uma para treinamento e outra para teste. Na bibliografia, é usual utilizar 66% dos dados para treinamento e 34% dos dados para teste, conforme ilustrado da Figura 4.

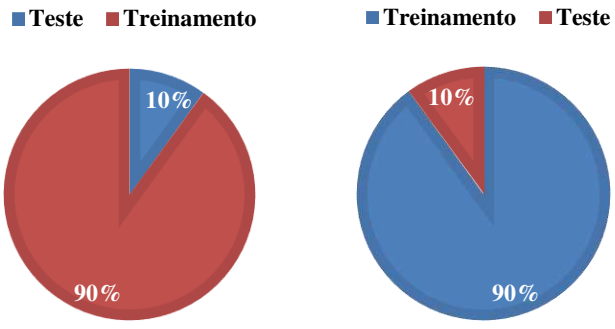
Figura 4: Ilustração da técnica de *percentage split.*



Fonte: Autor (2019).

- *Cross validation*: nesta técnica, o banco de dados é separado em subconjuntos de treinamento e teste, e o resultado final é obtido fazendo-se uma média de todos os resultados de teste. Em mineração de dados é comum a utilização de *cross validation with 10 folders*. Neste caso, como mostrado na Figura 5, os dados são divididos em 10 subconjuntos sendo 9 subconjuntos utilizados para treinamento e 1 para teste. Na sequência, há uma nova divisão, de modo que o subconjunto utilizado para teste será utilizado no treinamento e 1 dos subconjuntos de treinamento será utilizado como teste, e assim sucessivamente, até que cada subconjunto seja utilizado como teste. Ao final, restarão 10 resultados de teste, e o resultado final será a média dos valores.

Figura 5: Ilustração da técnica de *cross validation with* 10 *folders*.



Fonte: Autor (2019).

A escolha da estratégia pode depender do tipo do estudo, da característica do banco de dados e da capacidade de processamento à disposição. De acordo com um levantamento histórico por Viegas (2017), o algoritmo de Redes Neurais Artificiais é comumente utilizado em estudos para detecção de perdas não técnicas em mineração de dados. Assim, na próxima seção mais detalhes são fornecidos sobre o algoritmo.

2.3 REDES NEURAIS ARTIFICIAIS

As Redes Neurais Artificiais (RNA) têm origem na psicologia e na neurobiologia, e foram desenvolvidas com base no funcionamento do sistema nervoso biológico. Seu objetivo é simular o comportamento dos neurônios de um cérebro e, a partir disso, definir um modelo capaz de aprender a reconhecer padrões.

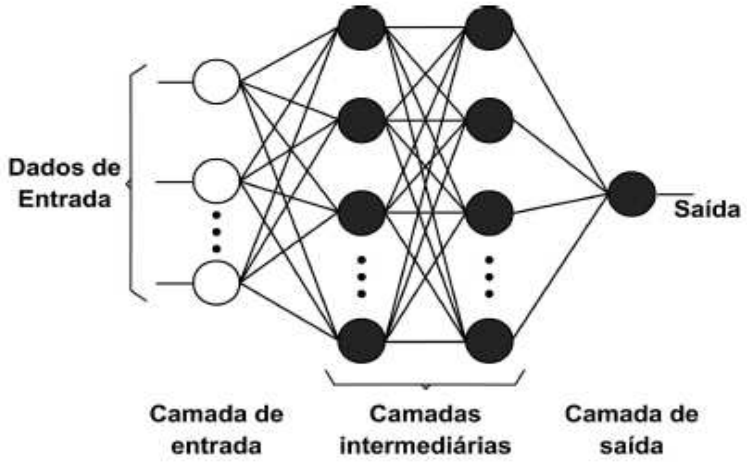
O neurônio biológico é formado pelo soma, que é o corpo do neurônio, pelo axônio, que conduz o impulso nervoso de um neurônio para outro, por dendritos, que são ramificações que fazem a comunicação entre neurônios, e pelas sinapses, que propagam o sinal neural ativando a comunicação com o próximo neurônio (LOESCH, 1996). Os neurônios artificiais, por sua vez, são comumente definidos por meio de uma soma ponderada das entradas. No neurônio artificial, os dendritos são representados pelos valores de entrada, os pesos são as sinapses, a função de soma faz o papel do corpo do neurônio e a função de transferência representa o axônio (MCCULLOCH, 1943). É por meio de um conjunto de neurônios artificiais que as RNA se tornam capazes de aprender a realizar uma determinada tarefa. Essa capacidade advém do treinamento da rede com relação ao problema de interesse.

Um tipo de RNA bastante utilizado é a *multilayer perceptron* (MLP), cuja estrutura básica consiste em 3 camadas: a camada de entrada, a camada intermediária e a camada de saída, conforme ilustrado na Figura 6.

Para treinamento de uma Rede Neural Artificial, é geralmente aplicada a técnica de *backpropagation*, na qual são usados pares de entradas e saídas com valores conhecidos para ajustar os pesos sinápticos da rede por meio de um mecanismo de correção de erros. Ou seja, durante o processo de treinamento, o valor de saída é gerado e comparado com o valor conhecido. O erro obtido é usado para ajustar os pesos sinápticos, a fim de reduzir gradualmente o erro (RUMELHART, 1986).

O erro é propagado da camada de saída para a camada de entrada. Portanto, os pesos sinápticos das camadas internas são corrigidos de acordo com o erro quando é feita a retropropagação. Na fase de treinamento, os pesos são ajustados até a rede ser capaz de identificar uma entrada padrão e processar uma resposta correta com relação ao padrão. O tempo de treinamento é influenciado por vários fatores, mas deve-se usar um critério de parada, como a taxa de erro, o número máximo de ciclos ou períodos de treinamento (RUMELHART, 1986).

Figura 6: Representação gráfica de uma RNA do tipo MLP com duas camadas intermediárias.



Fonte: Fiorin (2011).

Por ter uma elevada capacidade de aprendizado, as RNA são comumente utilizadas em problemas de classificação, nos quais se pretende identificar um padrão ou comportamento. Na próxima seção são feitas considerações sobre algumas métricas utilizadas para a avaliação do desempenho de técnicas de classificação em estudos envolvendo mineração de dados.

2.4 MÉTRICAS DE DESEMPENHO

Existem algumas métricas que são normalmente aplicadas em estudos de mineração de dados para avaliação dos resultados. Uma das mais utilizadas é a matriz de confusão. Para explicar o seu conceito, é utilizado o exemplo de um classificador desenvolvido para agrupar clientes de uma concessionária de energia elétrica em duas classes: regulares ou irregulares. Como pode ser visto na Tabela 1, tem-se uma matriz 2×2, em que *I* representa os clientes irregulares e *R* os clientes regulares.

Tabela 1: Exemplo de matriz de confusão.

	I	***R***
I	*VP*	*FP*
R	*FN*	*VN*

Fonte: Autor (2019).

Na Tabela 1, as colunas representam as reais classes dos dados, e as linhas representam a classificação estimada pelo método. Para entender a matriz de confusão, é

necessário compreender o que cada célula significa. A célula representada por *VN* contém a quantidade de exemplos regulares classificados corretamente como regulares (verdadeiros negativos). A célula representada por *FP* contém a quantidade de exemplos regulares classificados incorretamente como irregulares (falsos positivos). A célula representada por *FN* contém a quantidade de exemplos irregulares classificados incorretamente como regulares (falsos negativos). Já a célula representada por *VP* contém a quantidade de exemplos irregulares classificados corretamente como irregulares (os verdadeiros positivos). A partir dos valores de cada uma destas células, pode-se definir as métricas tais como a taxa de erros, a acurácia, a sensibilidade e a precisão. As métricas são essenciais à avaliação de desempenho do método classificação.

A taxa de erros é obtida a partir da razão entre o número de casos classificados incorretamente e o número total de casos, como mostrado na Equação (2), utilizando os dados da Tabela 1. Já a acurácia é a métrica na qual é mostrada a taxa de acertos, isto é, a razão entre o número de classificações corretas e o número total de exemplos utilizados, como mostrado na Equação (3) (SALARI, 2014):

$$taxa\ de\ erros\ = \frac{FN + FP}{FN + FP + VN + VP}, \tag{2}$$

$$acurácia\ = \frac{VN + VP}{FN + FP + VN + VP} = 1 - taxa\ de\ erros. \tag{3}$$

A sensibilidade é a razão entre o número de casos irregulares corretamente classificados e o número total de casos irregulares existentes, como mostrado na Equação (4). Já a precisão é a razão entre o número de casos irregulares corretamente classificados e o número total de exemplos classificados como irregulares, como mostrado em (5) (SALARI, 2014):

$$sensibilidade\ = \frac{VP}{VP + FN}, \tag{4}$$

$$precisão = \frac{VP}{VP + FP}. \quad (5)$$

Para usar a acurácia e a taxa de erros para avaliar o desempenho de um classificador em casos de detecção de fraudes, é necessário realizar uma análise cautelosa, pois elas podem ser fortemente influenciadas pelo número de casos regulares corretamente classificados.

No contexto de perdas não técnicas, a sensibilidade proporciona uma noção da cobertura do algoritmo classificador, isto é, o percentual do conjunto de clientes irregulares identificados pelo algoritmo. A precisão fornece uma noção da exatidão do algoritmo, isto é, o percentual de sucessos na identificação de clientes irregulares dentre o total de classificados como irregulares (COMETTI, 2004).

Muitos trabalhos utilizam a sensibilidade para avaliar o resultado obtido por um classificador. Entretanto, para se ter uma análise completa do classificador, é importante avaliar também a métrica de precisão.

Na próxima seção são dadas as definições dos parâmetros utilizados para a criação dos atributos derivados do histórico de consumo.

2.5 DEFINIÇÕES DE PARÂMETROS UTILIZADOS PARA A CRIAÇÃO DE NOVOS ATRIBUTOS

Nesta seção são fornecidas as definições de parâmetros estatísticos que são utilizados para a criação de novos atributos a partir dos dados do histórico de consumo. Os parâmetros são: média, moda, mediana, desvio padrão, variância, desvio médio, valor mínimo, valor máximo, amplitude, assimetria, 1º quartil, 3º quartil e curtose. De acordo com Meyer (1983), as definições são:

- **Média:** a média é uma das medidas de tendência central, e pode ser interpretada como o valor que indica a concentração de dados de uma distribuição. A definição matemática da média $\bar{x}$ de um conjunto de valores *x* é apresentada na Equação (6):

$$\bar{x} = \frac{x_1 + x_{2+} x_{3+\cdots+} x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad (6)$$

em que *n* representa o número de amostras e x_i os valores do conjunto.

- **Moda:** a moda é também uma das medidas de tendência central, sendo definida como o valor mais frequente em um conjunto de dados. Caso o conjunto de dados *x* seja uma matriz, então tem-se a moda dos elementos ao longo de cada dimensão.
- **Mediana:** a mediana é o valor que separa a amostra em populações de igual probabilidade de ocorrência. Ela é o elemento que divide a distribuição em 50% para cada lado, ou seja, o valor do meio. Caso o número de amostras seja par, a mediana refere-se ao valor médio dos dois elementos centrais.
- **Variância:** a variância é uma medida da dispersão de um conjunto de dados. A variância de uma população é definida como a média dos quadrados dos desvios dos valores em relação à média. A variância amostral é definida pela Equação (7):

$$Var = \frac{\sum(x_i - \bar{x})^2}{n-1}. \qquad (7)$$

- **Desvio Médio:** o desvio médio é uma medida da dispersão dos dados em relação à média de uma sequência, o "afastamento" em relação a essa média. Ele representa a média das distâncias entre cada elemento da amostra e seu valor médio, calculada conforme mostrado na Equação (8):

$$DM = \frac{\sum(x_i - \bar{x})}{n}, \qquad (8)$$

em que *n* representa o número de amostras, x_i os valores do conjunto, e $\bar{x}$ a média.

- **Desvio Padrão:** o desvio padrão é uma medida de dispersão em torno da média de uma variável, sendo comumente representado pela letra grega σ. Matematicamente, é obtido por meio da raiz quadrada da variância.
- **Amplitude:** a amplitude refere-se à diferença entre o maior (valor máximo) e o menor (valor mínimo) valor presente no conjunto de dados.
- **Quartis:** os quartis são os valores que dividem a distribuição em quatro partes. O primeiro quartil, Q_1, separa o menor quarto de valores do restante do conjunto; o segundo quartil, Q_2, divide o conjunto em parcelas iguais, de 50%, coincidindo então com a mediana; o terceiro quartil (Q_3), por fim, divide o conjunto em 75 % / 25 % dos dados.
- **Coeficiente de assimetria:** o coeficiente de assimetria permite distinguir distribuições assimétricas, de modo que um valor negativo indica que a cauda do lado esquerdo da função de densidade de probabilidade é maior do que no lado direito. Um valor positivo, por sua vez, implica em uma cauda maior no lado direito. Para um coeficiente nulo, tem-se dados distribuídos de forma relativamente igual em ambos os lados, não implicando, necessariamente, numa distribuição simétrica.
- **Curtose:** a curtose é uma medida de achatamento em comparação à distribuição normal. A curtose é definida pela expressão apresentada na Equação (9):

$$Curtose = \frac{1}{n}\sum\left[\frac{x_i - \bar{x}}{\sigma}\right]^4, \qquad (9)$$

em que n representa o número de amostras, x_i os valores do conjunto, $\bar{x}$ a média e σ o desvio padrão. Se a curtose assume valor 3, a distribuição tem o mesmo achatamento que a distribuição normal. Caso a curtose seja menor que 3, a distribuição é mais achatada que a distribuição normal. Caso contrário, a distribuição tem um pico mais abrupto que a normal.

Na seção textual subsequente é feita a fundamentação teórica acerca de métodos para seleção de atributos.

2.6 MÉTODOS DE SELEÇÃO DE ATRIBUTOS

Os métodos para seleção de atributos podem ser classificados principalmente em dois tipos: *Wrapper* e *Filter*. Em métodos do tipo *Wrapper* são utilizados algoritmos de classificação para medir a importância de um conjunto de atributos. O algoritmo de classificação deve avaliar todas as combinações possíveis entre os conjuntos de atributos para selecionar os melhores atributos. Dessa forma, os atributos selecionados irão depender das características do método de classificação aplicado. Além disso, para a aplicação de métodos do tipo *Wrapper* em grandes bases de dados, é necessária elevada capacidade de processamento computacional.

Já em métodos *Filter* é utilizado um parâmetro ou medida para identificar os melhores atributos. Neste caso não é aplicado nenhum algoritmo para a seleção dos atributos. Alguns exemplos de medidas utilizadas por métodos *Filter* são a auto correlação, e o coeficiente de Pearson. Esses métodos são computacionalmente mais rápidos e simples, e uma vez que a análise dos atributos é concluída, eles podem ser utilizados como entrada para diversos algoritmos de classificação.

Alguns exemplos de métodos aplicados para seleção de atributos são: *Correlation-based Feature Selection* (CFS) (HALL, 1999), *Principal Component Analysis* (PCA) (WOLD, 1987), *Gain Ratio* (GR) (QUINLAN, 1986), A*ttribute Evaluation* (MARCELIS, 1990), *Chi-square Feature Evaluation* (JIN, 2006), *Fast Correlation-based Feature selection* (FCBF) (YU, 2003), *Information Gain*, (QUINLAN, 1986), *Euclidean Distance* (HOWARD, 1984), *Markov blanket filter* (FU, 2010) e *Relief*. Nas duas subseções a seguir são dados mais detalhes sobre os métodos de CFS e *Relief* (SANTORO, 2005), ambos do tipo *Filter*.

2.6.1 *CORRELATION-BASED FEATURE SELECTION*

No método de CFS a relevância de um conjunto de atributos é determinada considerando o poder preditivo de cada atributo juntamente com o grau de redundância entre eles. O CFS é utilizado para estimar a correlação entre o conjunto de atributos e a classe (ou variável alvo) do problema em estudo, assim como a correlação entre cada um dos atributos.

A relevância de um conjunto de atributos cresce com o aumento da correlação entre os atributos e a classe e diminui com o aumento da correlação entre eles. O CFS é

usado para determinar o melhor subconjunto de atributos e geralmente é combinado com estratégias de busca como *Forward Selection*, *Backward Elimination*, *Bi-directional Search*, *Best-first Search* e *Genetic Search*.

O equacionamento do método de CFS é descrito pela Equação (10). (KAREGOWDA, 2010).

$$r_{zc} = \frac{k\overline{r_{zi}}}{\sqrt{k + k(k-1)r_{ii}}}, \tag{10}$$

em que r_{zc} é a correlação entre a soma dos conjuntos de atributos e a classe, k é o número de conjuntos de atributos, r_{zi} é a média do valor de correlação entre os conjuntos de atributos e a classe, e r_{ii} é a média do valor de correlação entre os atributos.

2.6.2 RELIEF

O método de *Relief* para seleção de atributos foi desenvolvido por Kira e Rendell, em 1992, a partir da análise da capacidade de cada atributo distinguir a qual classe pertence às instâncias mais próximas (SANTORO, 2005). Em um conjunto de dados, as instâncias são as linhas e os atributos são as colunas.

Por ser um método para avaliação individual de atributos, na aplicação do *Relief* uma variável estatística é atribuída a cada atributo para estimar a importância do atributo com relação à classe do problema em estudo. Por exemplo, em problemas para detecção de fraudes é possível estimar a importância de cada atributo para a classe "cliente irregular". O peso de cada atributo pode variar entre -1 (pior) e +1 (melhor), ou seja, quanto maior a importância do atributo para o problema em estudo mais próximo de 1 será o valor do peso calculado pelo método de *Relief*.

De acordo com Santoro (2015), na aplicação do método *Relief* é utilizada distância euclidiana para identificar as instâncias vizinhas mais próximas de cada instância com as seguintes características:

- A instância mais próxima com a mesma classe da instância (*NearHit*);
- A instância mais próxima com classe diferente da classe da instância (*NearMiss*).

A partir disto, é empregada a técnica de ponderação, na qual pesos diferentes são atribuídos aos atributos de acordo com a capacidade deles para discriminar as instâncias.

Aos atributos que conseguem discriminar instâncias de classes diferentes e não discriminar instâncias da mesma classe são atribuídos pesos maiores.

O peso de um atributo para cada instância será calculado a partir da diferença dos valores encontrados entre uma instância e seu *NearMiss* e *NearHit* respectivamente, como mostrado na Equação (11):

$$\begin{aligned} W[Z] = \text{diferença (valor de Z, instância, } \textit{NearMiss}) - \\ \text{diferença (valor de Z, instância, } \textit{NearHit}), \end{aligned} \tag{11}$$

em que W[Z] é o peso do atributo Z, e a função diferença (Z, instância1, instância2) retorna a diferença entre a instância1 e a instância2 relativa ao atributo Z. Ao final do estudo, todos os pesos dos atributos são somados e eles são então ranqueados.

Neste capítulo foram apresentados os conceitos teóricos acerca dos temas necessários para fundamentar este trabalho. Discutiu-se como a mineração de dados pode auxiliar na detecção de fraudes no sistema elétrico, como os algoritmos de classificação funcionam, e as principais estratégias de separação de bases para treinamento e teste. Ademais, foram apresentados os conceitos do algoritmo de Redes Neurais Artificiais, as principais métricas de avaliação utilizadas em problemas de classificação, as definições de alguns parâmetros estatísticos que auxiliarão na obtenção dos atributos derivados do histórico de consumo, e por fim, as definições acerca dos algoritmos de seleção de atributos CFS e *Relief*.

3 REVISÃO BIBLIOGRÁFICA

Neste capítulo são apresentados, em ordem cronológica, trabalhos correlatos ao tema desta pesquisa e que foram considerados contribuições relevantes ao delineamento do presente trabalho. São discutidos trabalhos relacionados à detecção de fraudes utilizando mineração de dados, à utilização da técnica de classificação com Redes Neurais Artificiais para identificação de clientes fraudulentos no setor de energia elétrica, à utilização de atributos derivados do histórico de consumo de clientes e à aplicação de algoritmos para seleção de atributos, com o intuito de se extrair o maior número de informações e garantir maior eficácia do estudo.

Em 1984, Gosh utilizou técnicas de mineração de dados para identificar a ocorrência de fraudes em cartões de crédito. No trabalho, foi utilizado um algoritmo baseado em uma rede neural artificial, que foi treinada com os casos de fraude resultantes de perda de cartão de crédito, roubo, fraudes deliberadas, e fraudes via sistema de correios. Ao final do estudo, o autor constatou que a habilidade de detecção de perfis de fraude pela rede neural possibilitou uma redução de 40% para 20% no percentual de fraudes. Foi desenvolvido um *software*, que foi instalado e agregado ao ambiente de produção de um banco.

O próximo trabalho relevante no tema de perdas não técnicas com a criação de atributos e utilização de redes neurais artificiais é de Eller (2003), no qual RNAs e mineração de dados são aplicadas para descobrir comportamentos suspeitos entre clientes da Centrais Elétricas de Santa Catarina S.A. (Celesc). Duas tarefas das redes neurais foram exploradas: classificação e segmentação. A classificação foi utilizada para se trabalhar com consumidores residenciais e comerciais, e a segmentação voltou-se a consumidores industriais. Foram calculados também atributos derivados das curvas de consumo. Entretanto, o autor destaca que muitas são as variáveis que determinam o comportamento de consumidores e o conjunto de variáveis que foi possível reunir para efetuar a análise foi muito limitado, o que tornou a busca de padrões comportamentais típicos de consumidores irregulares um tanto quanto prejudicada.

Em 2004, foi apresentada no trabalho de Cometti uma metodologia baseada no uso de técnicas de inteligência computacional para detectar possíveis ocorrências de uso ilícito de energia elétrica e instalações irregulares. O objetivo do trabalho foi aumentar as chances de sucesso nas inspeções em campo realizadas pelas empresas concessionárias.

Foram utilizadas técnicas de mineração de dados e sistemas baseados em conhecimento. Informações do perfil dos consumidores e seus históricos de consumos mensais nos anos anteriores foram usados durante o processo de mineração de dados. Resultados de inspeções anteriores foram utilizados como exemplos para o aprendizado supervisionado e testes dos sistemas classificadores. Da curva de consumo, foram extraídas várias medidas, como média, desvio padrão, coeficientes de amplitude e fase da série de Fourier, e coeficientes de polinômios aproximados pelo Método de Quadrados Mínimos. Entretanto, o autor cita que não foram abordadas no estudo a influência do comportamento sazonal dos clientes. Os resultados utilizando essas técnicas foram aplicados em testes de campo, os quais tiveram uma assertividade de aproximadamente 23%.

Rauber (2005) focou na determinação de quais atributos relacionados aos dados de clientes são mais importantes para a identificação de perdas não técnicas. Foram utilizados dois métodos de seleção de características: *Best Features*, que analisa cada característica independente das demais; e SFS, que dado o melhor conjunto atual, examina todos os candidatos restantes juntamente com o conjunto atual e agrega o melhor dos candidatos ao conjunto. Os atributos analisados foram: o coeficiente de Fourier, o coeficiente de *wavelet*, a série de consumos original e o polinômio ortogonal derivado desta série. Após a obtenção dos resultados, o autor concluiu que a extração de novas características de dados disponíveis melhora o desempenho de um classificador.

Em 2007, Todesco construiu um sistema para identificação de possíveis irregulares de energia elétrica, empregando o processo *Knowledge Discovery in Databases* (KDD). Para a etapa de mineração dos dados, foram utilizadas somente informações sobre o consumo e definiu-se uma medida chamada *score* acumulado. Trata-se de uma medida que calcula a diferença entre o consumo do mês atual e o consumo do mesmo mês no ano anterior para 12 meses. Consumidores com *score* acumulado acima de determinado valor limiar são candidatos à inspeção. Após ajuste do valor limiar, a taxa de acerto para o grupo de consumidores residenciais foi de 64% e a taxa média de acerto para o grupo de consumidores comerciais (padarias, lanchonetes e postos de gasolina) foi de 80%.

Penin (2008) utilizou um algoritmo de redes neurais artificiais na identificação de perdas não técnicas. Os atributos estatísticos coeficiente de variação, média e desvio padrão, derivados dos dados de consumo, foram as únicas entradas para o algoritmo de redes neurais artificiais. O índice de casos corretamente classificados foi de

aproximadamente 32%. Destaca-se que o trabalho possuiu um foco voltado para a análise econômica da perda não-técnica.

Em 2012, Mondero estudou a detecção de perdas não técnicas utilizando dados de clientes da *Endesa Company*. No estudo, são levados em consideração o coeficiente de Pearson e árvores de decisão. Alguns atributos estatísticos como o valor máximo e mínimo da curva de consumo mensal e bimestral e a média semestral de consumo de cada cliente são também considerados para a análise. Segundo o autor, os resultados obtidos foram melhores do que os até então registrados na bibliografia.

Ramos (2012) utilizou o *software* WEKA para aplicar os algoritmos de *K nearest neighbor*, *support vector machine* e Redes Neurais Artificiais em um banco de dados e comparou os resultados obtidos entre os três classificadores. Alguns atributos como fator de potência, potência reativa, demanda contratada e demanda consumida foram utilizados em conjunto com os dados de consumo (que não sofreram nenhuma adaptação ou modificação). Como conclusão, o autor destaca que a utilização de *features* (atributos criados de dados já existentes) melhora a assertividade do classificador.

Costa (2013) utilizou redes neurais para detectar o comportamento irregular em clientes de uma distribuidora. Uma análise de autocorrelação foi realizada para a determinação dos atributos que seriam utilizados no estudo. Com relação aos dados de consumo, foram usados apenas os valores mensais. O autor destaca que a metodologia utilizada possibilitou a obtenção de resultados melhores, mas ressalta a necessidade de testar o método com mais algoritmos de classificação.

A utilização de mineração de texto, redes neurais e técnicas estatísticas foi a estratégia adotada por Guerrero (2014) para extrair informações de dados de clientes da empresa Endesa, que seriam convertidas em regras de detecção de fraudes. No que diz respeito às técnicas estatísticas, foram derivados dos dados de consumo a média, os valores mínimos e máximos e o desvio típico de consumo. O foco do trabalho foi determinar, a partir dessas métricas, o perfil do cliente regular e assim identificar o perfil irregular. Foi alcançada uma taxa de acerto de cerca de 33,6%.

Já Kosut (2015) utilizou a análise de consumo de clientes de uma concessionária como ferramenta para a detecção de fraudes em sistemas de distribuição de energia elétrica. Para auxiliar no estudo foram criados 30 novos atributos, baseados em características dos clientes como tipo de tarifa, tensão de alimentação, quantidade de irregularidades, dias desde a última inspeção, potência consumida, entre outros. Na sequência foi feito um estudo para selecionar os melhores atributos utilizando métodos

para seleção de atributos do tipo *filte*r e *wrapper* e foi aplicado o algoritmo de árvores de decisão para classificação dos clientes. Ao final do estudo foi comprovado que com o acréscimo e a posterior seleção de novos atributos foi possível aumentar o acerto na classificação dos clientes.

Como pode ser compreendido pela exposição realizada, a maioria dos trabalhos utiliza a criação de atributos juntamente com técnicas de mineração de dados para detecção de perdas não técnicas. Entretanto, poucos são os trabalhos em que é utilizado um método para seleção dos melhores atributos e são exploradas diversas características dos clientes.

Portanto, neste trabalho é proposta a análise da contribuição de atributos derivados do histórico de consumo para a detecção de perdas não técnicas. São criados atributos a partir dos dados de consumo utilizando parâmetros estatísticos, representações no domínio da frequência, e métricas que consideram parâmetros como sazonalidade e variações e quedas de consumo. Na sequência, os melhores atributos são determinados utilizando-se métodos para seleção de atributos e o algoritmo de redes neurais artificiais é utilizado para classificar com base nos atributos resultantes os clientes de uma concessionária de energia entre regulares e irregulares.

A fim de sintetizar as contribuições dos trabalhos supracitados, na Tabela 2 são apresentadas as principais referências e contribuições abordadas nesta revisão bibliográfica. Adicionalmente, na Tabela 2 também é destacada a contribuição desta pesquisa.

Neste capítulo foi discorrido acerca de trabalhos que utilizaram técnicas de mineração de dados, criação de atributos e seleção de atributos para a detecção de perdas não técnicas, enfatizando-se as contribuições e melhorias propostas por este trabalho ao estudo. No próximo capítulo, é descrita a metodologia aplicada neste trabalho para avaliação da contribuição de atributos derivados do histórico de consumo na detecção de perdas não técnicas.

Tabela 2 - Resumo das principais referências e contribuições dos pesquisadores citados na revisão bibliográfica.

Pesquisadores	Contribuições				
	Mineração de dados para detecção de fraudes	**Mineração de dados para detecção de perdas não técnicas**	**Criação de novos atributos a partir dos dados de consumo**	**Algoritmos para seleção de atributos**	**Exploração de características temporais, sazonais e estatísticas dos dados de consumo**
Gosh (1984)	X				
Eller (2003)	X	X	X		
Cometti (2004)	X	X	X		
Rauber (2005)	X	X	X	X	
Todesco (2007)	X	X	X		
Penin (2008)	X	X	X		
Mondero (2012)	X	X	X		
Ramos (2012)	X	X	X	X	
Costa (2013)	X	X			
Guerrero (2014)	X	X	X		
Kosut (2015)	X	X		X	
Este trabalho (2019)	X	X	X	X	X

Fonte: Autor (2019).

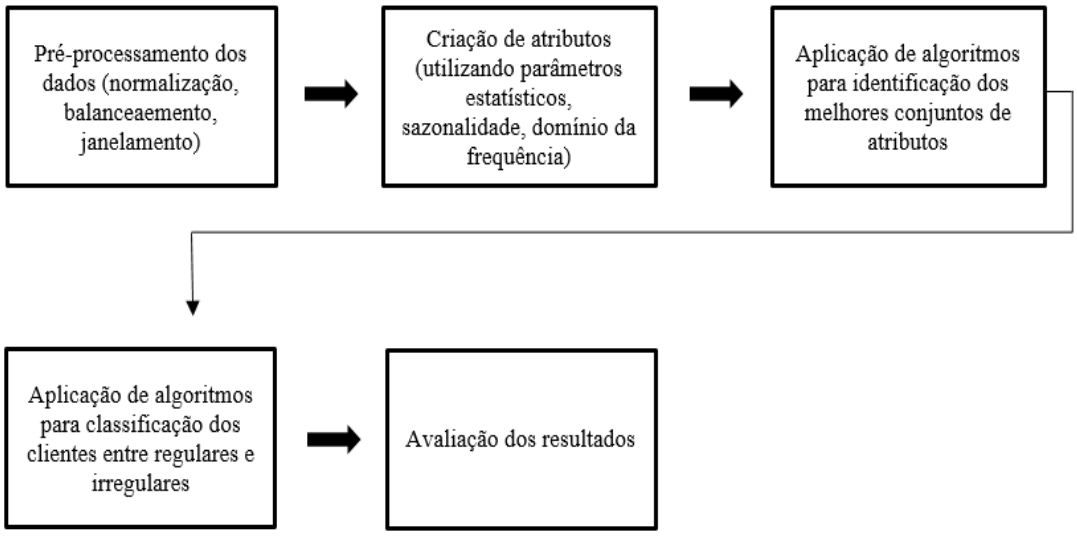
4 METODOLOGIA

Neste capítulo é apresentada a metodologia proposta para a análise das contribuições dos atributos derivados do histórico de consumo de energia elétrica para a detecção de perdas não técnicas, enfatizando-se os processos empregados nas etapas de pré-processamento dos dados, seleção dos atributos, e avaliação dos resultados. A metodologia está baseada na utilização de técnicas de mineração de dados e aprendizado de máquina para classificação dos clientes de uma concessionária de energia elétrica com atuação no Brasil entre regulares e irregulares.

Os procedimentos adotados são: tratamento do banco de dados com as informações sobre o histórico de consumo de energia elétrica dos clientes da concessionária; criação de atributos a partir do histórico de consumo; emprego de métodos de seleção de atributos para identificação dos melhores atributos; utilização de algoritmo para classificação dos clientes entre regulares e irregulares a partir dos melhores atributos identificados; e avaliação dos resultados por meio da matriz de confusão e das métricas: assertividade, acurácia e precisão e sensibilidade.

Para explicar, de forma mais detalhada, as etapas da metodologia aplicada para a análise da contribuição dos atributos derivados do histórico de consumo na detecção de perdas não técnicas, foi construído o diagrama de blocos apresentado na Figura 7, no qual é indicada a ordem dos procedimentos supracitados.

Figura 7 - Diagrama de blocos representativo da metodologia proposta.



Fonte: Autor (2019).

Como pode ser constatado, a primeira etapa consiste no pré-processamento dos dados, em que ocorre a normalização, o balanceamento e o janelamento dos dados. Na sequência, atributos são criados a partir das informações do histórico de consumo. Algoritmos para seleção de atributos como: *Correlation Based Feature Selection* e *Relief* são aplicados para identificação dos melhores atributos. Por fim, o algoritmo de redes neurais artificiais é empregado para a classificação dos clientes utilizando os atributos selecionadas na etapa anterior como entrada. Convém ressaltar que o objetivo principal é analisar qual a contribuição dos atributos derivados do histórico de consumo na detecção das perdas não técnicas.

4.1 MATERIAL

Nesta seção, o banco de dados utilizado durante o estudo é descrito. Para analisar a contribuição de atributos derivados do histórico de consumo de energia elétrica na detecção de perdas não técnicas, foram utilizadas informações reais de consumo mensal de clientes da base cadastral de uma concessionária de distribuição de energia elétrica com atuação em um dos estados do Brasil. Por motivos de confidencialidade, a concessionária será chamada de concessionária A.

Como tratado na fundamentação teórica, na seção 2.2, para a aplicação da técnica de aprendizagem supervisionada, durante a etapa de treinamento do algoritmo, é necessário que se conheça previamente a classificação de cada elemento da base de dados que será utilizada.

Dessa forma, foram selecionados dados de consumo mensal referentes a 9.177 unidades consumidoras entre outubro de 2014 e setembro de 2017. As unidades consumidoras foram inspecionadas pela concessionária A no mês de setembro de 2017.

Para a criação de parte dos atributos derivados do histórico de consumo é utilizado o *software* MATLAB®. Já para realização da seleção de atributos e classificação de clientes é utilizado o *software* de mineração de dados WEKA. Este *software open source* foi desenvolvido em linguagem Java pela Universidade de Waikato, na Nova Zelândia. No ambiente do *software*, além de ser possível aplicar algoritmos para seleção de atributos, é permitido o uso de várias técnicas de mineração de dados (classificação, clusterização, predição, etc.) e algoritmos para avaliação dos resultados (Redes Neurais

Artificiais, Redes Bayesianas, Árvores de Decisão, Support Vector Machine, etc). Além disso, os parâmetros para cada algoritmo utilizado podem ser editados pelo usuário.

4.2 MÉTODOS

Nesta seção, os procedimentos adotados para a realização do trabalho, os quais foram apresentados na Figura 7, são descritos. São explicadas as etapas de pré-processamento, as premissas adotadas para a criação dos atributos derivados do histórico de consumo, a utilização dos algoritmos de seleção de atributos e dos algoritmos de classificação e, a estratégia para avaliação dos resultados.

4.2.1 PRÉ-PROCESSAMENTO DOS DADOS

A primeira etapa do estudo consiste no tratamento dos dados selecionados. Os procedimentos utilizados para realizar o tratamento são: normalização, balanceamento e janelamento dos dados.

A técnica de normalização é aplicada ao banco de dados com o intuito de evitar discrepâncias na ordem de grandeza dos dados de diferentes naturezas, agrupando-os em uma mesma faixa de valores. Isto é importante, tendo em vista que valores de consumo muito distintos uns dos outros podem fazer com que um atributo se sobreponha aos demais, prejudicando a qualidade da informação. Para realizar o tratamento dos dados, é necessário aplicar o método de normalização explicado na seção 2.1 da fundamentação teórica. A técnica de normalização deve ser também empregada após a criação dos novos atributos, o que garante que todos os atributos estejam contemplados no mesmo intervalo de grandeza. Neste caso, entre 0 e 1.

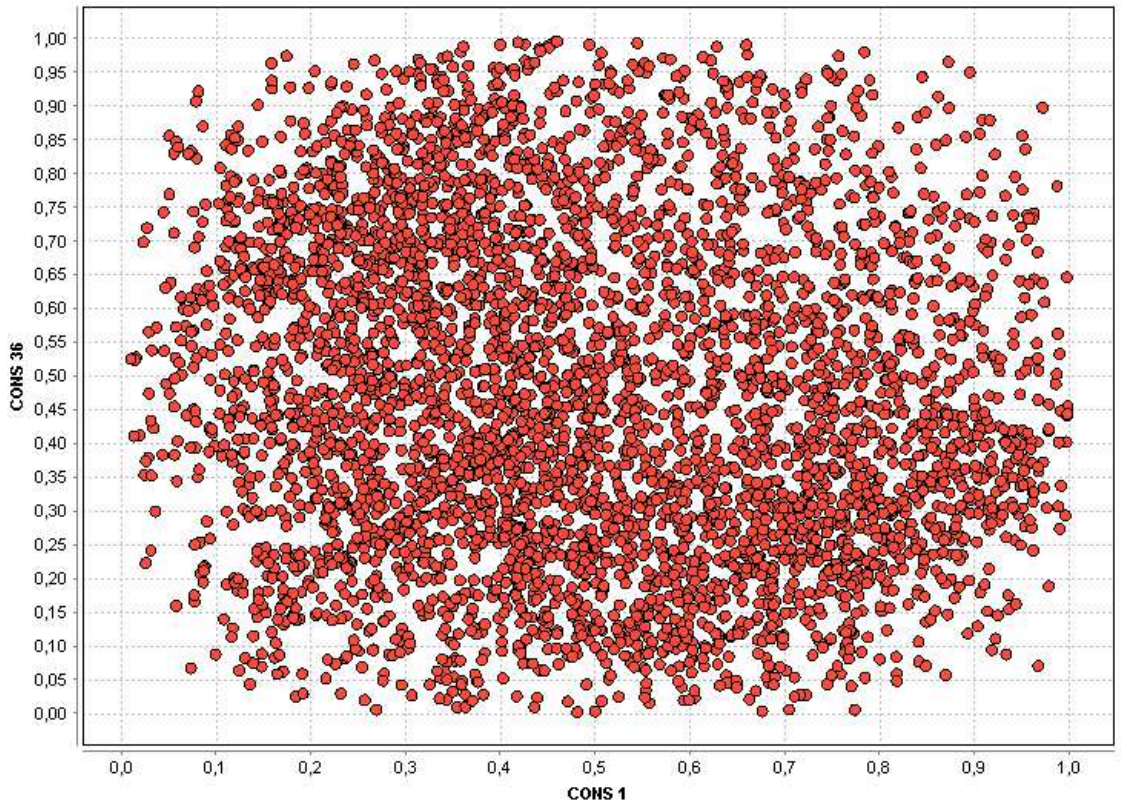
O banco de dados obtido possui uma característica desbalanceada com relação à variável alvo, ou variável *target* (cliente irregular), tendo em vista que, dos 9.177 clientes inspecionados, foi encontrada fraude em 4.476 casos (49% da base de dados). Sendo assim, como a redução no volume da classe dominante não provoca redução considerável no volume total do banco de dados, é recomendável realizar o balanceamento dos dados para evitar o enviesamento amostral (quando uma das amostras está presente em maior quantidade nos dados de análise). Assim, é necessário realizar o procedimento conhecido

com *undersampling* (discutido na seção 2.1 da fundamentação teórica), resultando em um banco de dados com 8.952 casos, 4.476 com fraude (50%) e 4.476 sem fraude (50%).

Com relação ao janelamento dos dados, o procedimento se faz necessário para tratar todos os casos dentro de um mesmo intervalo de valores e tempo. No banco de dados havia consumidores com períodos de consumo distintos. É recomendável que a análise seja feita com o mesmo número de dados de consumo mensal para todas as unidades consumidoras. Dessa forma, para fins de padronização, foi considerado o período de 36 meses de informação antes da realização da inspeção em todas as unidades. O período de 36 meses é conveniente por ser o intervalo máximo permitido por lei para que a concessionária de energia elétrica possa cobrar ao cliente, uma vez que a fraude é detectada (ANEEL, 2010).

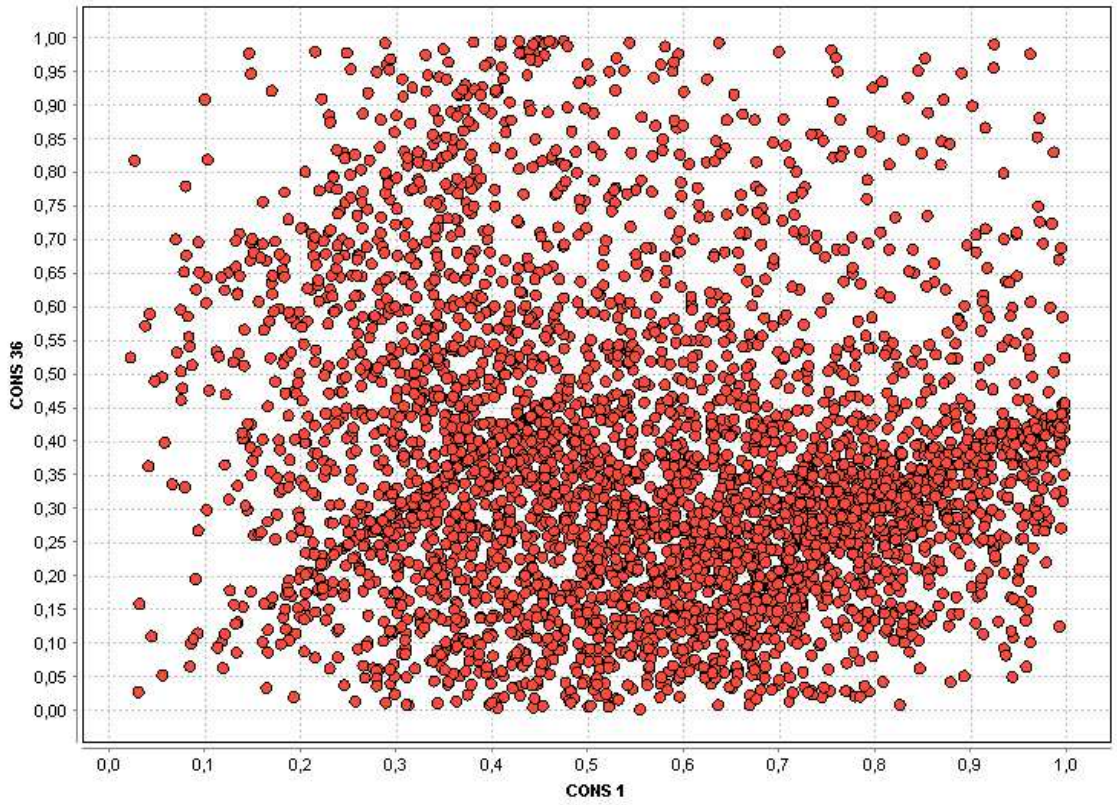
Para ilustrar o comportamento de consumo dos clientes dessa base de dados, são mostrados nas Figuras 8 e 9, em forma de gráfico, os valores de consumo normalizados dos clientes para o mês 1 (outubro de 2014), eixo horizontal, e para o mês 36 (setembro de 2017), eixo vertical. Os dados dos clientes regulares são mostrados na Figura 8 e os dados dos clientes irregulares na Figura 9.

Figura 8: Disposição dos valores de consumo dos clientes regulares para o mês 1 (eixo horizontal) e para o mês 36 (eixo vertical).



Fonte: Autor (2019).

Figura 9: Disposição dos valores de consumo dos clientes regulares para o mês 1 (eixo horizontal) e para o mês 36 (eixo vertical).



Fonte: Autor (2019).

Há uma mudança de perfil de consumo dos clientes irregulares ocorrida durante o período de 36 meses considerado, em que a queda de consumo fica evidente em função da concentração da maior parte dos clientes irregulares na região inferior do gráfico. Como dito anteriormente, no eixo Y são representados os valores de consumo registrados no mês 36, e no gráfico a maior parte dos clientes possui valores de consumo entre 0 e 0,5 (considerando uma escala normalizada), enquanto que no mês 1 esse consumo se distribui entre 0,3 e 0,9 (também considerando uma escala normalizada). Quanto aos clientes regulares, o perfil de consumo é mais diversificado, com os valores de consumo sendo distribuídos por toda a região do gráfico.

Na próxima seção são concedidos mais detalhes acerca da estratégia adotada para criação de novos atributos.

4.2.2 CÁLCULO DOS ATRIBUTOS DERIVADOS DOS DADOS DE CONSUMO

Após a etapa de pré-processamento, é necessário criar atributos derivados do histórico de consumo. A criação dos atributos é realizada com o intuito de explorar informações adicionais dos dados de consumo que podem estar sendo inutilizadas durante a etapa de classificação. É analisada a contribuição dos novos atributos quando empregados como fonte de dados para a detecção de perdas não técnicas. Com a criação desses atributos são consideradas as seguintes características:

- Variação mensal do perfil de consumo de energia elétrica dos clientes em análise;
- Informações estatísticas do perfil de consumo de energia elétrica dos clientes em análise;
- Informações no domínio da frequência do perfil de consumo dos clientes em análise;
- Características sazonais do perfil de consumo dos clientes em análise;
- Contribuição da queda de consumo na caracterização do perfil de consumo de clientes irregulares.

A informação da variação mensal de consumo por unidade consumidora permite a extração de informações adicionais sobre o comportamento do cliente, as quais podem contribuir para a detecção de fraudes. Por exemplo, alguns clientes irregulares costumam controlar seu consumo com o intuito de ludibriar a concessionária de energia elétrica (em um mês consomem normalmente, no mês seguinte fazem uso da fraude para reduzir o consumo e assim sucessivamente).

Já os atributos obtidos a partir da aplicação de parâmetros estatísticos aos dados de consumo e as variações ou degraus de consumo, podem ser utilizados para investigar se há algum padrão de comportamento no perfil dos clientes irregulares. Espera-se que esse tipo de informação possa contribuir para melhorar a assertividade dos algoritmos aplicados para a classificação dos clientes. Os atributos são calculados considerando todo o intervalo de consumo de 36 meses para os dados de consumo, e de 35 meses para os degraus de consumo. Os parâmetros estatísticos utilizados estão listados na Tabela 3.

Tabela 3 - Parâmetros estatísticos escolhidos.

Parâmetros Estatísticos	
Média	Mediana
Moda	Desvio padrão
Desvio médio	Valor mínimo
Valor máximo	Amplitude
Curtose	Assimetria
1º quartil	3º quartil
Variância	

Fonte: Autor (2019).

Como destacado na revisão bibliográfica, em alguns estudos para detecção de perdas não técnicas utilizando mineração de dados, resultados satisfatórios foram obtidos com a utilização da informação de consumo dos clientes no domínio da frequência. Portanto, atributos são criados a partir da aplicação da transformada de Fourier aos dados de consumo dos clientes.

Para contribuir com o estudo de detecção de perdas não técnicas, a informação de sazonalidade é também utilizada. São criados atributos a partir do cálculo da média de consumo dos meses de verão e inverno, da diferença entre a média de consumo dos meses do verão (e inverno) atual e a média de consumo dos meses do verão (e inverno) do ano anterior, e da diferença entre o consumo registrado no mês atual e o consumo registrado no mesmo período do ano passado.

A última característica utilizada como premissa para a criação dos atributos é a medida da participação da queda de consumo no perfil de clientes irregulares. Para avaliar a importância da queda de consumo foi criado um atributo ao qual é atribuído o valor unitário a cada queda de consumo superior a 30% com relação ao consumo medido no mês anterior.

Ao final são criados 124 novos atributos a partir dos 36 atributos com a informação do consumo de energia elétrica medido. Por questões de organização da estrutura do texto, os nomes e as definições de cada um dos 160 atributos utilizados neste trabalho são apresentados no Apêndice A. Cada vez que um atributo for citado no texto, a sua definição é fornecida na sequência.

4.2.3 SELEÇÃO DOS MELHORES ATRIBUTOS DERIVADOS DOS DADOS DE CONSUMO

Após a etapa de criação dos novos atributos devem ser utilizados os métodos de *Correlation Based Feature Selection* e *Relief* para determinação dos atributos mais importantes na identificação de clientes irregulares utilizando técnicas de classificação e mineração de dados.

Os algoritmos são aplicados aos 160 atributos (os 36 atributos originais e os 124 criados posteriormente). Para o método de *Correlation Based Feature Selection* são escolhidos todos os atributos indicados após a aplicação do método e é utilizado o algoritmo de busca de *Best First.* Já para o método *Relief* são escolhidos os 24 atributos melhor ranqueados. Dessa forma, são determinados 3 conjuntos diferentes de atributos, considerando os atributos selecionados pelos algoritmos supracitados e os atributos originais. Os atributos selecionados por cada algoritmo são mostrados na seção de Resultados.

4.2.4 CLASSIFICAÇÃO DOS CLIENTES

Após a determinação dos conjuntos com os melhores atributos, cada conjunto é utilizado como entrada para o algoritmo de Redes Neurais Artificiais. Neste trabalho, é utilizada uma rede neural artificial com estrutura *Multilayer Perceptron*, com técnica de treinamento do tipo *backpropagation*, taxa de treinamento de 0,2 e de aprendizado de 0,3. Os valores das taxas de treinamento e aprendizado foram definidos após avaliações dos primeiros resultados obtidos.

Para a análise de classificação, o banco de dados deve ser separado em duas bases, uma para treinamento e outra para teste. Para separação das bases é utilizada a estratégia de *percentage split,* em função do tempo para processamento dos resultados e do tamanho do banco de dados. Os resultados são avaliados por meio da matriz de confusão fornecida pelo WEKA após a classificação. As métricas de desempenho utilizadas são a sensibilidade e a precisão. Combinando estas duas métricas, conforme é mostrado na Equação (12), é possível determinar a assertividade do classificador.

$$assertividade = \frac{s \cdot pr}{0{,}3 \cdot pr + 0{,}7 \cdot s} \qquad (12)$$

em que s é o valor da sensibilidade, e pr é a precisão. Dado que o intuito é obter um classificador que indique o maior número possível de inspeções assertivas, à métrica precisão foi concedido um peso maior no cálculo da assertividade dos classificadores. Neste trabalho, o resultado da classificação dos clientes será avaliado conforme o valor de assertividade, calculado de acordo com a Equação (12).

Neste capítulo foram apresentados a metodologia proposta para a análise das contribuições dos atributos derivados do histórico de consumo de energia elétrica para a detecção de perdas não técnicas, enfatizando-se os processos empregados nas etapas de pré-processamento dos dados, seleção dos atributos, e avaliação dos resultados. No capítulo seguinte é feita uma análise crítica dos resultados obtidos, bem como um ordenamento mostrando os resultados gerados por cada classificador (do melhor para a pior).

5 RESULTADOS

Neste capítulo são apresentadas as análises dos resultados obtidos utilizando-se a metodologia descrita no capítulo anterior. A apresentação dos resultados está dividida em duas seções. A primeira trata da sumarização e análise das características dos atributos selecionados pelos algoritmos de *Correlation Based Feature Selection* e *Relief* e a segunda corresponde à avaliação dos resultados obtidos pelo algoritmo de redes neurais artificiais quando da utilização como entrada dos dados de consumo originais e dos atributos selecionados na primeira etapa.

5.1 SELEÇÃO DE ATRIBUTOS

Os atributos apresentados na Tabela 4 foram selecionados a partir da aplicação do método de *Correlation Based Feature Selection*. Da análise da tabela, é possível constatar que apenas 7 atributos faziam parte da lista de atributos originais (é importante notar que os valores dos 36 meses de consumo são os atributos originais), e 17 (ou 70,83% dos atributos selecionados) foram criados posteriormente. Além disso, os atributos selecionados estão correlacionados às características de variação de consumo mensal com mais ocorrências nos meses próximos à data da inspeção (Der22, Der26, Der28, Der29, Der30, Der31, Der32, Der33, Der34, Der35), às informações estatísticas dos dados de consumo originais (mínimo, máximo, amplitude e desvio médio), e à sazonalidade (Var20,Var24, DifVerão).

Tabela 4 – Atributos selecionados pelo método de *Correlation Based Feature Selection*.

Rótulo do atributo	**Descrição**
Cons1	Consumo Mês 1
Cons2	Consumo Mês 2
Cons5	Consumo Mês 5
Cons7	Consumo Mês 7
Cons34	Consumo Mês 34
Cons35	Consumo Mês 35
Cons36	Consumo Mês 36
Minimo Cons	Valor mínimo dos 36 meses de consumo
Maximo Cons	Valor máximo dos 36 meses de consumo
Amplitude Cons	Amplitude dos 36 meses de consumo
Der22	Diferença de consumo entre o mês 23 e o mês 22

Der26	Diferença de consumo entre o mês 27 e o mês 26
Der28	Diferença de consumo entre o mês 29 e o mês 28
Der29	Diferença de consumo entre o mês 30 e o mês 29
Der30	Diferença de consumo entre o mês 31 e o mês 30
Der31	Diferença de consumo entre o mês 32 e o mês 31
Der32	Diferença de consumo entre o mês 33 e o mês 32
Der33	Diferença de consumo entre o mês 34 e o mês 33
Der34	Diferença de consumo entre o mês 35 e o mês 34
Der35	Diferença de consumo entre o mês 36 e o mês 35
DesvioM Der	Desvio médio das derivadas dos 36 meses de consumo
Var20	Diferença entre o consumo do mês 32 e a média de consumo dos últimos 12 meses
Var24	Diferença entre o consumo do mês 36 e a média de consumo dos 12 meses anteriores
DifVerão	Diferença entre a média dos consumos dos meses de verão do ano da inspeção com relação a média de consumo dos meses de verão do ano anterior

Fonte: Autor (2019).

Já na Tabela 5 é possível observar os atributos selecionados a partir da aplicação do método de *Relief*. Da análise da Tabela, constata-se que apenas 3 atributos faziam parte da lista de atributos originais, e 21 (ou 87,5% dos atributos selecionados) foram criados posteriormente, a partir dos dados originais. Neste caso, todas as características citadas (variações de consumo, sazonalidade, informações estatísticas) estão presentes nos atributos selecionados. Entretanto, há predominância de atributos relacionados a informações estatísticas (mínimo, amplitude, desvio padrão, variância, moda, média, mediana, quartil) tanto dos dados de consumo originas quanto das variações mensais de consumo.

Tabela 5 – Atributos selecionados pelo método de *Relief*.

Var24	Diferença entre o consumo do mês 36 e a média de consumo dos 12 meses anteriores
Minimo Cons	Valor mínimo dos 36 meses de consumo
Cons36	Consumo Mês 36
Amplitude Cons	Amplitude dos 36 meses de consumo
Desvio 01 Cons	Desvio padrão dos 36 meses de consumo
Der35	Diferença de consumo entre o mês 36 e o mês 35
Var23	Diferença entre o consumo do mês 35 e a média de consumo dos 12 meses anteriores

Variancia Cons	Variância dos 36 meses de consumo
Var22	Diferença entre o consumo do mês 34 e a média de consumo dos 12 meses anteriores
Cons35	Consumo Mês 35
Moda Cons	Moda dos 36 meses de consumo
Media Cons	Média dos 36 meses de consumo
Mediana Cons	Mediana dos 36 meses de consumo
Desvio 01 Der	Desvio padrão das derivadas dos 36 meses de consumo
Mediana Der	Mediana das derivadas dos 36 meses de consumo
Der34	Diferença de consumo entre o mês 35 e o mês 34
Variancia Der	Variância das derivadas dos 36 meses de consumo
DesvioM Cons	Desvio médio dos 36 meses de consumo
Cons34	Consumo Mês 34
Der33	Diferença de consumo entre o mês 34 e o mês 33
Media Der	Média das derivadas dos 36 meses de consumo
Inverno	Média de consumo dos meses de inverno
Quartil 3 Cons	Valor do 3º quartil dos 36 meses de consumo
Media 6 Cons	Média dos últimos 6 meses de consumo antes da data da inspeção

Fonte: Autor (2019).

Por fim, os atributos que foram selecionados concomitantemente pelo método de *Correlation Based Feature Selection* e pelo método de *Relief* são mostrados em ordem alfabética na Tabela 6. Neste caso, o número de atributos pertencentes a lista de atributos originais é de 3, e o número de atributos que foram criados a partir dos dados originais corresponde a 6 ou 66,6% do conjunto total.

Na próxima seção são mostrados e discutidos os resultados obtidos pelo algoritmo de classificação usando redes neurais artificiais quando utilizado como entrada os atributos originais (dados de consumo dos 36 meses) e os atributos das Tabelas 4, 5 e 6.

Tabela 6 – Atributos selecionados pelo método de *Correlation Based Feature Selection* e *Relief*.

Amplitude Cons	Amplitude dos 36 meses de consumo
Cons34	Consumo Mês 34
Cons35	Consumo Mês 35
Cons36	Consumo Mês 36
Der33	Diferença de consumo entre o mês 34 e o mês 33
Der34	Diferença de consumo entre o mês 35 e o mês 34
Der35	Diferença de consumo entre o mês 36 e o mês 35
Minimo Cons	Valor mínimo dos 36 meses de consumo

Var24	Diferença entre o consumo do mês 36 e a média de consumo dos 12 meses anteriores

Fonte: Autor (2019).

5.2 CLASSIFICAÇÃO DE CLIENTES

A seguir são mostrados os resultados obtidos para cada um dos conjuntos de atributos, cujos parâmetros foram utilizados como entrada para o algoritmo de classificação usando Redes Neurais Artificiais. Para avaliação dos resultados, são utilizadas matrizes de confusão, em que *I* denota o número de consumidores irregulares e *R* representa o número de consumidores regulares. Posteriormente, uma discussão acerca da avaliação e comparação dos resultados é realizada.

No Quadro 1, são mostrados os resultados obtidos para o primeiro conjunto com os atributos originais (dados de consumo dos 36 meses). Como pode ser constatado, a partir de uma análise do Quadro 1, com o uso do algoritmo baseado na utilização de Redes Neurais Artificiais foi possível identificar corretamente apenas metade dos clientes irregulares (738 em um universo de 1.484 casos). Já no que se refere aos clientes apontados como irregulares pelo classificador, cerca de 60% realmente eram irregulares (738 em um universo de 1.236 casos). A assertividade calculada foi de aproximadamente 56%.

Quadro 1 - Matriz de confusão – atributos originais

Matriz de Confusão			Métricas (%)	
	I	R	Acurácia	59,00%
I	738	498	Sensibilidade	50,00%
R	746	1062	Precisão	60,00%
			Assertividade	56,00%

Fonte: Autor (2019).

No Quadro 2, são mostrados os resultados para o segundo conjunto com os atributos selecionados pelo método de *Correlation Based Feature Selection*. Como pode ser constatado, com o uso do algoritmo baseado na utilização de Redes Neurais Artificiais foi possível identificar corretamente apenas 54% dos clientes irregulares (802 em um universo de 1.484 casos). Já, no que se refere aos clientes apontados como irregulares pelo classificador, cerca de 69% foram corretamente identificados (802 em um universo de 1.167 casos). Levando em consideração as duas métricas, a assertividade calculada foi de aproximadamente 64%. Houve um aumento em todas as métricas consideradas se comparado com os valores calculados utilizando apenas os dados de consumo originais.

Isto significa que a utilização da metodologia de criação e seleção de atributos trouxe ganhos na detecção de clientes irregulares.

Quadro 2 - Matriz de confusão – *Correlation based feature selection.*

Matriz de Confusão			Métricas	
	I	R	Acurácia	66,00%
I	802	365	Sensibilidade	54,00%
R	682	1195	Precisão	69,00%
			Assertividade	64,00%

Fonte: Autor (2019).

Os resultados obtidos com os atributos selecionados pelo método *Relief* são apresentados no Quadro 3. Como pode ser constatado, com o uso do algoritmo baseado na utilização de Redes Neurais Artificiais foi possível identificar corretamente 57% dos clientes irregulares (841 em um universo de 1.484 casos). Quanto aos clientes apontados como irregulares pelo classificador, cerca de 70% foram corretamente apontados (841 em um universo de 1.216 casos). Levando em consideração as duas métricas, a assertividade calculada foi de aproximadamente 65%. Nesse caso, com a utilização dos atributos selecionados pelo método de *Relief*, os valores de todas as métricas consideradas foram superiores às métricas calculadas utilizando apenas os atributos originais e além disso, o valor de assertividade foi superior ao valor registrado utilizando os atributos selecionados pelo método de CFS.

Quadro 3 - Matriz de confusão – *Relief.*

Matriz de Confusão			Métricas	
	I	R	Acurácia	67,00%
I	841	375	Sensibilidade	57,00%
R	643	1185	Precisão	70,00%
Relief			Assertividade	65,00%

Fonte: Autor (2019).

Por fim, os resultados obtidos com a junção dos atributos selecionados pelo método de CFS e pelo método de *Relief* são mostrados no Quadro 4. Como pode ser constatado, com o uso do algoritmo baseado na utilização de Redes Neurais Artificiais foi possível identificar corretamente 61,00% dos clientes irregulares (903 em um universo de 1.484 casos). Já, no que se refere aos clientes apontados como irregulares pelo classificador, cerca de 68,00% foram corretamente identificados (903 em um universo de 1.330 casos). Levando em consideração as duas métricas, a assertividade calculada foi de aproximadamente 66,00%. Nesse caso, os valores das métricas foram superiores aos valores registrados utilizando os atributos originais, os atributos selecionados apenas pelo método de CFS e os atributos selecionados pelo método de *Relief*.

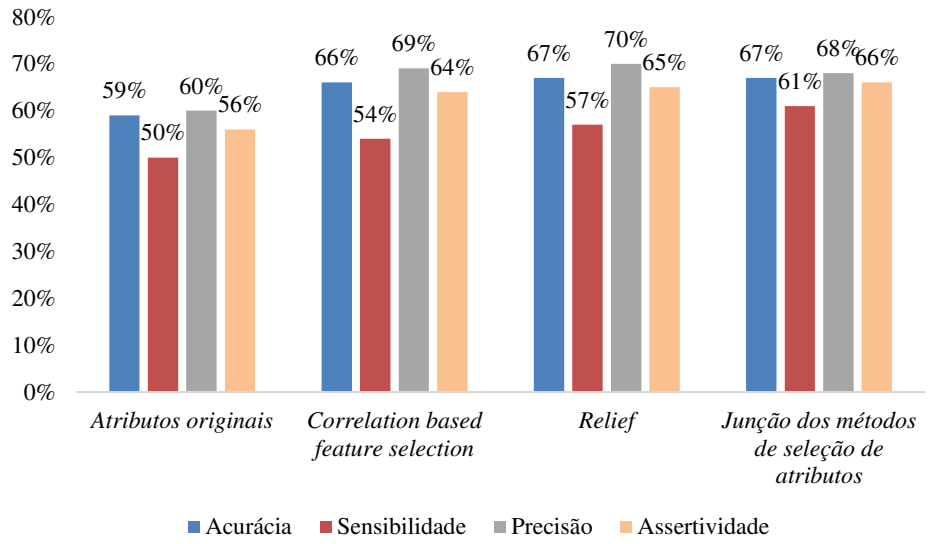
Quadro 4 - Matriz de confusão – Junção dos métodos de seleção de atributos.

Matriz de Confusão			Métricas	
	I	R	Acurácia	67,00%
I	903	427	Sensibilidade	61,00%
R	581	1113	Precisão	68,00%
			Assertividade	66,00%

Fonte: Autor (2019).

Na Figura 10, os resultados para todos os conjuntos de atributos são mostrados de forma gráfica. É possível constatar que o melhor resultado foi obtido com a utilização dos atributos selecionados tanto pelo método de *Correlation Based Feature Selection* quanto pelo método de *Relief*, com uma assertividade de 66,00%, e, o pior resultado foi obtido com os atributos originais, com uma assertividade de 56,00%, totalizando um ganho de 10 pontos percentuais após a aplicação da metodologia proposta neste trabalho.

Figura 10 – Síntese dos Resultados.



Fonte: Autor (2019).

A princípio, o ganho obtido pode aparentar ser baixo. Entretanto, muitas concessionárias de energia elétrica no Brasil registram, em média, um índice de assertividade de cerca de 10% a 12% na seleção de clientes irregulares utilizando modelos e regras convencionais como direcionadores. Considerando que sejam realizadas 100 inspeções com o custo de R$ 100,00 por inspeção e com uma assertividade de 10%, têm-se um gasto não recuperado de R$ 9.000,00. Considerando-se o mesmo cenário, mas com uma assertividade de 20%, o gasto não recuperado cai para R$ 8.000,00. Tendo em vista que, em alguns estados do Brasil, são realizadas em média 150.000 inspeções por ano,

um ganho de 10 pontos percentuais na assertividade na seleção de alvos proporcionaria (no cenário considerado de R$ 100,00 por inspeção) uma economia anual de R$ 1.500.000,00.

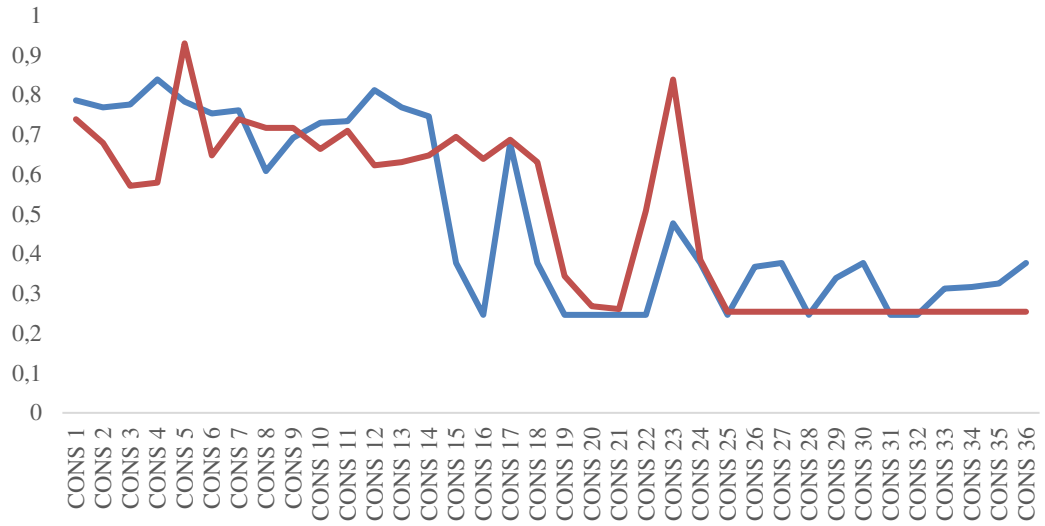
Além da diminuição do custo total das inspeções, com o aumento de assertividade obtido por meio da metodologia proposta neste trabalho, é possível aumentar o montante de energia a ser recuperado ou faturado pela concessionária a partir da identificação da irregularidade (fraude ou defeito). Dado que a ANEEL permite que seja feita a cobrança da energia não faturada durante o período em que o cliente estava em situação de irregularidade (até 36 meses retroativos em casos de fraudes e até 3 meses retroativos em casos de irregularidade por defeito).

Com relação aos atributos selecionados pelos dois métodos de seleção aplicados neste trabalho, é possível entender as principais características do perfil de consumo dos clientes que contribuíram para uma maior assertividade na detecção de clientes irregulares. Dos 9 atributos selecionados concomitantemente pelos métodos de CFS e *Relief*, 6 possuíam informações acerca do consumo e da variação de consumo em meses próximos a data da inspeção, 2 atributos continham informações sobre a diferença entre o valor máximo e o valor mínimo do conjunto de dados e a informação do valor mínimo do conjunto de dados, e no atributo restante havia a exploração da característica de variação anual de consumo.

Dessa forma, pode-se concluir que nas variações de consumo há informações importantes para algoritmos de classificação baseados no uso de Redes Neurais Artificiais obterem melhores resultados em estudos de detecção de perdas não técnicas. Isto ocorre porque grande parte dos clientes irregulares possui um perfil de consumo com variações abruptas e com acentuadas quedas em comparação a períodos anteriores. A título de ilustração, no gráfico da Figura 11, é mostrado o perfil de consumo de dois clientes irregulares durante os 36 meses avaliados.

Da análise do gráfico da Figura 11, é possível constatar que existem diferenças entre o perfil de clientes irregulares. Porém, a característica de queda de consumo e/ou maior variação do perfil de consumo é típica. No caso do perfil de consumo do cliente 1 (linha azul do gráfico da Figura 11), as elevações de consumo registradas após a queda, não são equivalentes ao consumo medido no início da análise. Já no caso do perfil de consumo do cliente 2 (linha vermelha do gráfico da Figura 11), provavelmente têm-se a descrição de um problema de defeito no medidor, tendo em vista que o consumo permanece inalterado no mínimo da tarifa.

Figura 11 – Perfil de consumo de dois clientes irregulares.



Fonte: Autor (2019).

Ademais, vale ressaltar que utilizando da metodologia aplicada é possível avaliar a contribuição de cada atributo criado a partir dos dados do histórico de consumo de clientes de uma concessionária de distribuição de energia elétrica.

Por fim, é importante destacar a necessidade de aplicar esta metodologia sempre que for feito um estudo em um banco de dados diferente, tendo em vista que os melhores atributos a serem selecionados podem variar conforme o banco de dados utilizado devido às características do conjunto de clientes. Na Tabela 7 os principais resultados obtidos a partir deste trabalho são descritos.

Neste capítulo, os resultados deste trabalho foram apresentados e discutidos. No próximo capítulo são apresentadas as conclusões e as propostas para futuros trabalhos.

Tabela 7 – Síntese dos Resultados.

Criação de novos atributos a partir dos dados de consumo de energia elétrica de clientes de uma concessionária de energia elétrica com área de concessão no Brasil, explorando sobretudo características sazonais e queda de consumo
Identificação dos melhores atributos do conjunto de dados resultante (utilizando métodos de seleção de atributos) para a detecção de perdas não técnicas
Obtenção de ganho de 10 pontos percentuais de assertividade na classificação de clientes contribuindo para perdas não técnicas ao aplicar o algoritmo de redes neurais artificiais no banco de dados com os melhores atributos

Fonte: Autor (2019).

6 CONCLUSÕES

Nesta dissertação foi apresentada a análise da contribuição de atributos derivados do histórico de consumo para a detecção de perdas não técnicas. Para tanto, foram criados atributos a partir dos dados de consumo originais utilizando parâmetros estatísticos, a representação dos dados de consumo no domínio da frequência e regras para consideração de características temporais e de sazonalidade do perfil de consumo. Na sequência, foram utilizados os algoritmos *Correlation Feature Based Selection* e *Relief* para selecionar os melhores atributos dentre o total de atributos resultante (originais e derivados).

Os atributos originais, os atributos selecionados pelo método de CFS, os atributos selecionados pelo método de *Relief* e os atributos selecionados concomitantemente pelos métodos de CFS e *Relief* foram utilizados como entrada para o algoritmo baseado no uso de Redes Neurais Artificiais, para a classificação dos clientes de uma concessionaria de distribuição de energia entre regulares e irregulares.

Os resultados da classificação realizada pelo algoritmo baseado no uso de Redes Neurais Artificiais, considerando as 4 entradas descritas anteriormente foram avaliados a partir das métricas de acurácia, sensibilidade e precisão fornecidas por uma matriz de confusão. O melhor valor de assertividade foi obtido quando utilizado como entrada os atributos selecionados concomitantemente pelos dois métodos de seleção de atributos aplicados (CFS e *Relief*).

Dos 9 atributos selecionados concomitantemente pelos métodos de CFS e *Relief*, 6 possuíam informações acerca do consumo e da variação de consumo em meses próximos a data da inspeção, 2 atributos continham informações sobre a diferença entre o valor máximo e o valor mínimo do conjunto de dados e a informação do valor mínimo do conjunto de dados, e no atributo restante havia a exploração da característica de variação anual de consumo.

Com a criação de novos atributos e a avaliação da contribuição deles como proposto pela metodologia apresentada neste trabalho, foi possível obter um ganho de 10 pontos percentuais na assertividade na identificação de clientes irregulares utilizando o algoritmo de Redes Neurais Artificiais e técnicas de mineração de dados. Esse ganho representa uma diminuição do custo empregado pelas concessionárias em inspeções improcedentes e um aumento no volume de recuperação de energia obtido pelas concessionárias em ações de inspeções em campo.

Assim, a partir dos resultados obtidos neste trabalho, é possível concluir que:

- Conforme exposto no primeiro objetivo específico deste trabalho, foram criados atributos a partir dos dados de consumo de clientes de uma concessionária de distribuição de energia elétrica. Todos os atributos criados estão compilados na Tabela I.A do Apêndice A. Com a criação dos atributos, foi possível explorar algumas características dos clientes, como o comportamento do perfil de consumo em períodos sazonais e as variações do consumo ao longo do período de análise;
- A partir do segundo objetivo específico deste trabalho, foi possível identificar, por meio do uso de seleção de atributos, os melhores atributos dentre o conjunto de atributos original e os atributos criados. Os atributos selecionados pelos métodos de seleção adotados (CFS e *Relief*), estão apresentados nas Tabelas 4 e 5. É importante notar que a maioria dos atributos selecionados pelos métodos, ou seja, a maioria dos melhores atributos, pertence à lista de atributos criados para cumprimento do primeiro objetivo específico;
- Para cumprimento dos terceiro e quarto objetivos específicos deste trabalho, foi utilizado o algoritmo baseado no uso de redes neurais artificiais utilizando como entrada os atributos originais, os atributos selecionados pelo método de CFS, os atributos selecionados pelo método de *Relief* e os atributos selecionados concomitantemente pelos métodos de CFS e *Relief*. Os valores de assertividade estão apresentados nos Quadros 1, 2, 3 e 4 para todas as entradas que foram descritas. O melhor valor de assertividade foi obtido com a utilização dos atributos selecionados concomitantemente pelos métodos de CFS e *Relief* e houve um ganho de 10 pontos percentuais com relação ao pior valor de assertividade (obtido utilizando apenas os atributos originais como entrada).

Dessa forma, a partir das conclusões apresentadas, é possível constatar que a metodologia utilizada neste trabalho permite:

- Explorar características e informações do perfil de consumo que podem contribuir significativamente para o aumento da assertividade na seleção de clientes irregulares;

- Identificar os melhores atributos em um conjunto de dados, contribuindo para o aumento da assertividade dos resultados e o tempo de processamento computacional das análises com a diminuição do número total de atributos;
- Reduzir os custos de operação ao aumentar a assertividade na seleção de alvos para inspeções em campo.

Com a apresentação das conclusões obtidas por meio desta dissertação, podem ser destacadas algumas perspectivas de trabalhos futuros. As perspectivas estão apresentadas na subseção a seguir.

6.1 TRABALHOS FUTUROS

Algumas linhas de investigação que podem ser seguidas a partir dos estudos apresentados neste trabalho, bem como as questões não aprofundadas nele, estão listadas abaixo:

- Empregar a metodologia apresentada considerando a utilização de outros algoritmos de seleção de atributos (*Gain Ratio* e *InfoGain*) e algoritmos de classificação (árvores de decisão e redes bayesianas);
- Avaliar a contribuição de técnicas de clusterização na detecção de perdas não técnicas, utilizando segmentos de clientes com características semelhantes (porte de consumo, classe de tensão, atividade empregada);
- Comparar a contribuição de regras criadas a partir de combinações de atributos para a detecção de perdas não técnicas. Por exemplo, determinar qual seria a assertividade encontrada a partir da seleção de clientes por meio de uma regra que considere clientes com queda nos últimos 3 meses maior que 30%, e classe de tensão trifásica e média de consumo do verão deste ano abaixo da média de consumo do verão do ano anterior.
- Determinar a probabilidade da classificação de cada cliente como irregular, com o intuito de identificar os casos mais suscetíveis e assim aumentar a assertividade em inspeções feitas a partir da seleção obtida por algoritmos de classificação de clientes;

- Empregar a metodologia apresentada neste trabalho em campo e realizar análises dos resultados obtidos.

7 PUBLICAÇÕES

Até o presente momento e durante o desenvolvimento da pesquisa, alguns artigos foram publicados, submetidos e aceitos para publicação. Os artigos estão apresentados na Tabela 8.

Tabela 8 - Artigos publicados, submetidos e aceitos para publicação.

Artigos	Autores	Título	Congresso ou Revista	Ano
PUBLICADOS	**ALVES, H. M. M.** OLIVERIA, I. B. BARROS, R. M R. ARAUJO, J. F. COSTA, E. G.	**Detecção de Perdas Não Técnicas Utilizando Mineração de Dados**	IWADA	2019
	DANTAS, F. B. SILVA, W. P. **ALVES, H. M. M.** COSTA, E. G.	Geração e Medição de Alta Tensão Alternada e Contínua com Aplicações em Laboratório	IWADA	2019
	DANTAS, F. B. SILVA, W. P. **ALVES, H. M. M.** COSTA, E. G.	Distribuição de Tensão Impulsiva nos Enrolamentos de um Transformador	IWADA	2019
	ANDRADE, A. F. COSTA, E. G. **ALVES, H. M. M.** ANDRADE, F. L. M.	Influence of harmonics on the electromechanical stresses in a power transformer	SBSE	2018
	ANDRADE, A. F. **ALVES, H. M. M.** DINIZ, L LUCIANO, B. A.	Analytical and computational study of the inductance in a power reactor	SBSE	2018
	ANDRADE, A. F. **ALVES, H. M. M.** FERNANDES, J. M. B COSTA, E. G.	Computational modelling of heat transfer in a porcelain-housed surge arrester	SBSE	2018
	LUCENA, M. D. **ALVES, H. M. M.** DINIZ, L COSTA, E. G.	Electrical Breakdown Analysis in a Sphere-Plane Utilizing the Finite Element Method	SBSE	2018
	LUCENA, M. D. DINIZ, L **ALVES, H. M. M.** COSTA, E. G.	Influence of the Electronegativity in the Breakdown Voltage of N2 and SF6 Mixtures of Circuit Breakers	SBSE	2018
	ANDRADE, A. F. FERNANDES, J. M. B **ALVES, H. M. M.** COSTA, E. G.	Thermal Behavior Analysis in a Porcelain-Housed ZnO Surge Arrester by Computer Simulations and Thermography	ICHVE	2018
	SILVA, W. P. **ALVES, H. M. M.** DANTAS, F. B. NASCIMENTO, M. L. COSTA, E. G.	Utilização de Metodologia Prática para o Entendimento do Circuito de Greinacher	COBENGE	2018
	ANDRADE, A. F. **ALVES, H. M. M.** GERMANO, A. D MATIAS, P. S	Solar Radiation and Ambient Temperature Influence on Electrothermal Behavior of a Polymeric Surge Arrester	ISH	2017
	ANDRADE, A. F.		ISH	2017

	COSTA, E. G. **ALVES, H. M. M.** GERMANO, A. D	Influence of Impurities Movement on the Dielectric Strength of Insulating Oil		
ACEITO PARA PUBLICAÇÃO	ANDRADE, A. F. COSTA, E. G. FERNANDES, J. M. B **ALVES, H. M. M.** AMORIM, C. R.	Thermal Behaviour Analysis in a Porcelain-Housed ZnO Surge Arrester by Computer Simulations and Thermography	Revista High Voltage	2019

Fonte: Autor (2019).

REFERÊNCIAS

COMETTI, E. S., VAREJÃO, F. M. **Melhoramento da Identificação de Perdas Comerciais Através da Análise Computacional Inteligente do Perfil de Consumo e dos Dados Cadastrais de Consumidores**. UFES - Universidade Federal do Espírito Santo. Espírito Santo, 2004.

ANEEL–Agência Nacional de Energia Elétrica. <http://www2.camara.leg.br/atividade-legislativa/comissoes/comissoes-permanentes/cme/audiencias-publicas/2018/audiencia-publica-16-05-2018/ANEEL%20-%20%20Perdas%20Eletricas%20-%20Davi%20Lima.pdf>.Audiência Pública, maio de 2018. Acesso em 05 de novembro de 2018.

HUBACK, Vanessa Barroso da Silva. **Medidas ao Combate a Perdas Elétricas Não Técnicas em Áreas com Severas Restrições a Operação de Sistemas de Distribuição de Energia Elétrica**. Dissertação de mestrado. Universidade Federal do Rio de Janeiro, 2018.

MESSINIS, George M., HATZIARGYRIOU, Nikos D. Å. ***Review of non-technical loss detection methods.*** *Electric Power Systems Research*, Elsevier, janeiro de 2018.

BASTOS, Paulo Roberto F. de Moura. **Diagnóstico de Perdas Comerciais de Energia Elétrica na Distribuição Usando Redes Bayesianas**. Tese de doutorado. Universidade Federal de Campina Grande, 2011.

KAREGOWDA, Asha Gowda; MANJUNATH A. S.; JAYARAM, M. A. ***Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection.*** *International Journal of Information Technology and Knowledge Management,* dezembro de 2010, Volume 2, No. 2, pgs. 271-277.

FAYYAD, U; PIATETSKY-SHAPIRO, G; SMYTH, P. ***From Data Mining to Knowledge Discovery in Databases.*** *American Association for Artificial Intelligence*, 1996.

QUEIROGA, Rodrigo Mendonça.– Operador Nacional do Sistema Elétrico. **Uso de Técnicas de Data Mining Para Detecção de Fraudes em Energia Elétrica**. Dissertação de mestrado. Universidade Federal do Espírito Santo, 2005.

MACHADO, Felipe Nery Rodrigues. ***Big Data O Futuro dos Dados e Aplicações***. Saraiva educação. 1 ed, 2018.

VIEGAS, Joaquim L., ESTEVES, Paulo R., MELÍCIO R., MENDES, V. M. F., VIEIRA Susana M. ***Solutions for detection of non-technical losses in the electricity grid: A review.*** *Renewable and Sustainable Energy Reviews,* 2017.

LOESCH, Cláudio. **Redes neurais artificiais, fundamentos e modelos**. Blumenau: Editora da FURB, 1996, p. 166.

MCCULLOCH, W. S.; PITTS, W. ***A logical calculus of the ideas immanent in nervous activity***. *Bulletin of Mathematical Biophysics*, 1943.

RUMELHART, David E. HINTON, Geoffrey E. WILLIAMS, Ronald J**.** ***Learning representations by back-propagating errors.*** Nature, 1986.

FIORIN, Daniel V. et al. **Aplicações de redes neurais e previsões de disponibilidade de recursos energéticos solares**. Rev. Bras. Ensino Fís. São Paulo, v. 33, n. 1, p. 01-20, 2011.

HONG, Yao-Ming. ***Feasibility of using artificial neural networks to forecast groundwater levels in real time.*** *Landslides.* 1-12. 10.1007/s10346-017-0844-5. 2017.

SALARI, Nader & Shohaimi, Shamarina &Najafi, Farid & Nallappan, Meenakshii & Karishnarajah, Isthrinayagy. ***A Novel Hybrid Classification Model of Genetic Algorithms, Modified k-Nearest Neighbor***

and Developed Backpropagation Neural Network. PloS one. 9. e112987. 10.1371/journal.pone.0112987, 2014.

MEYER, Paul. **Probabilidade: Aplicações a Estatística**. LTC Editora. 2 ed, 1983

WOLD, Svante. ESBENSEN, Kim. GELADI, Paul. ***Principal Component Analysis***. Chemometrics and Intelligent Laboratory Systems, 2 pp 37-52, 1987.

QUINLAN, J.R. Mach Learn (1986) 1: 81. https://doi.org/10.1007/BF00116251.

MARCELIS, A.J.J.M. ***On the classification of attribute evaluation algorithms***. Science of Computer Programming, Volume 14, Issue 1, June 1990, Pages 1-24.

JIN, Xin. XU, Anbang. BIE, Rongfang. GUO, Ping. ***Machine Learning Techniques and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profiles***. BioDM 2006, LNBI 3916, pp. 106–115, 2006.

YU, Lei. LIU Huan. ***Feature Selection for High-Dimensional Data:A Fast Correlation-Based Filter Solution***. Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.

HOWARD, Anton. ***Elementary Linear Algebra.*** (7th ed.), John Wiley & Sons, pp. 170–171, ISBN 978-0-471-58742-2, 1984.

FU, Shunkai. DESMORAIS, Michel C. ***Markov Blanket based Feature Selection: A Review of Past Decade***. Proceedings of the World Congress on Engineering 2010 Vol I WCE 2010, June 30 - July 2, 2010.

SANTORO, Daniel Monegatto. **Sobre Processos de Seleção de Subconjuntos de Atributos - As abordagens Filtro e *Wrapper***. Dissertação de mestrado. Universidade Federal de São Carlos, 2005.

GHOSH, Sushmito. REILLY, Douglas L. ***Credit Card Fraud Detection With a Neural Network.*** *Proceedings of the Twenty-Seventh Annual Hawaii International Conference on System Sciences*. 1994.

ELLER, Nery Artur. **Arquitetura de Informação para o Gerenciamento de Perdas Comerciais de Energia Elétrica**. Tese de doutorado. Universidade Federal de Santa Catarina, 2003.

RAUBER, Thomas W. DRAGO, Idilio. VAREJÃO, Flávio M. QUEIROGA, Rodrigo M. **Extração e Seleção de Características na Identificação de Perdas Comerciais na Distribuição de Energia Elétrica.** XXV Congresso da Sociedade Brasileira de Computação, 2005.

TODESCO, J. L. MORALES, A. B. T. RAUTENBERG, S. GARBELOTTO, L. A. ATHAYDE, E. D. **Aplicação de Técnicas de Mineração de Dados para detecção de Fraudes de Energia.** Universidade Federal de Santa Catarina, 2007.

PENIN, Carlos Alexandre de Souza. **Combate, Prevenção e Otimização das Perdas Comercias de Energia.** Tese de doutorado. Universidade de São Paulo, 2008.

MONEDERO, Iñigo. BISCARRI, Félix. LEON, Carlos. GUERRERO, Juan I. BISCARRI, Jesus. MILLAN, Rocio. ***Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees.*** Electrical Power and Energy Systems. 2012.

RAMOS, Caio C. O. SOUZA, André N. GASTALDELLO, Danilo Sinkiti. ***Identification and feature selection of non-technical losses for industrial consumers using the software WEKA***. *2012 10th IEEE/IAS International Conference on Industry Applications*. Fortaleza, 2012.

COSTA Breno C., ALBERTO Bruno. L. A., PORTELA André M., MADURO W., ELER EsdrasO. ***Fraud Detection In Electric Power Distribution Networks Using An Ann-Based Knowledge-Discovery Process***. *International Journal of Artificial Intelligence & Applications (IJAIA),* Vol. 4, No. 6, Novembro de 2013.

GUERRERO, Juan I. LEÓN, Carlos. MONEDERO, Iñigo. BISCARRI, Félix. BISCARRI, Jesus. ***Improving Knowledge-Based Systems with statistical techniques, text mining, and neural networks for non-technical loss detection***. *Knowledge-Based Systems Journal*. 2014.

KOSUT, Juan Pablo. SANTOMAURO, Fernando. JORYSZ, Andrés. FERNANDEZ, Alicia. LECUMBERRY, Federico. RODRIGUEZ, Fernanda. ***Abnormal Consumption Analysis For Fraud Detection UTE-UdelaR Joint Efforts***. *IEEE PES Innovative Smart Gride Technologies Latin America. 2015*.

HALL, Mark A. ***Correlation-based Feature Selection for Machine Learning***. Tese de doutorado. Universidade de Waikato, 1999.

BOGORODITSKY, N. P.; PASYNKOV, V. V.; TAREEV, B. M. ***Electrical Engineering Materials***. MIR Publishers Moscow. 1979.

RAMOS, Caio C. O., RODRIGUES Douglas, SOUZA, André N**. *On the Study of Commercial Losses in Brazil: A Binary Black Hole Algorithm for The ft Characterization***. *IEEE Transactions on Smart Grid*. 2016.

APÊNDICE A - LISTA DE ATRIBUTOS

Na Tabela I.A estão descritos todos os atributos utilizados neste trabalho.

Tabela I.A – Descrição dos atributos.

Cons1	Consumo do mês 1
Cons2	Consumo do mês 2
Cons3	Consumo do mês 3
Cons4	Consumo do mês 4
Cons5	Consumo do mês 5
Cons6	Consumo do mês 6
Cons7	Consumo do mês 7
Cons8	Consumo do mês 8
Cons9	Consumo do mês 9
Cons10	Consumo do mês 10
Cons11	Consumo do mês 11
Cons12	Consumo do mês 12
Cons13	Consumo do mês 13
Cons14	Consumo do mês 14
Cons15	Consumo do mês 15
Cons16	Consumo do mês 16
Cons17	Consumo do mês 17
Cons18	Consumo do mês 18
Cons19	Consumo do mês 19
Cons20	Consumo do mês 20
Cons21	Consumo do mês 21
Cons22	Consumo do mês 22
Cons23	Consumo do mês 23
Cons24	Consumo do mês 24
Cons25	Consumo do mês 25
Cons26	Consumo do mês 26
Cons27	Consumo do mês 27
Cons28	Consumo do mês 28
Cons29	Consumo do mês 29
Cons30	Consumo do mês 30
Cons31	Consumo do mês 31
Cons32	Consumo do mês 32
Cons33	Consumo do mês 33

Cons34	Consumo do mês 34
Cons35	Consumo do mês 35
Cons36	Consumo do mês 36
Média Cons	Média dos 36 meses de consumo
Mediana Cons	Mediana dos 36 meses de consumo
Moda Cons	Moda dos 36 meses de consumo
Desvio 01 Cons	Desvio padrão dos 36 meses de consumo
DesvioM Cons	Desvio médio dos 36 meses de consumo
Variancia Cons	Variância dos 36 meses de consumo
Minimo Cons	Valor mínimo dos 36 meses de consumo
Máximo Cons	Valor máximo dos 36 meses de consumo
Amplitude Cons	Amplitude dos 36 meses de consumo
Quartil 1 Cons	Valor do 1º quartil dos 36 meses de consumo
Quartil 3 Cons	Valor do 3º quartil dos 36 meses de consumo
Der1	Diferença de consumo entre o mês 2 e o mês 1
Der2	Diferença de consumo entre o mês 3 e o mês 2
Der3	Diferença de consumo entre o mês 4 e o mês 3
Der4	Diferença de consumo entre o mês 5 e o mês 4
Der5	Diferença de consumo entre o mês 6 e o mês 5
Der6	Diferença de consumo entre o mês 7 e o mês 6
Der7	Diferença de consumo entre o mês 8 e o mês 7
Der8	Diferença de consumo entre o mês 9 e o mês 8
Der9	Diferença de consumo entre o mês 10 e o mês 9
Der10	Diferença de consumo entre o mês 11 e o mês 10
Der11	Diferença de consumo entre o mês 12 e o mês 11
Der12	Diferença de consumo entre o mês 13 e o mês 12
Der13	Diferença de consumo entre o mês 14 e o mês 13
Der14	Diferença de consumo entre o mês 15 e o mês 14
Der15	Diferença de consumo entre o mês 16 e o mês 15
Der16	Diferença de consumo entre o mês 17 e o mês 16
Der17	Diferença de consumo entre o mês 18 e o mês 17
Der18	Diferença de consumo entre o mês 19 e o mês 18
Der19	Diferença de consumo entre o mês 20 e o mês 19
Der20	Diferença de consumo entre o mês 21 e o mês 20
Der21	Diferença de consumo entre o mês 22 e o mês 21
Der22	Diferença de consumo entre o mês 23 e o mês 22
Der23	Diferença de consumo entre o mês 24 e o mês 23
Der24	Diferença de consumo entre o mês 25 e o mês 24
Der25	Diferença de consumo entre o mês 26 e o mês 25

Der26	Diferença de consumo entre o mês 27 e o mês 26
Der27	Diferença de consumo entre o mês 28 e o mês 27
Der28	Diferença de consumo entre o mês 29 e o mês 28
Der29	Diferença de consumo entre o mês 30 e o mês 29
Der30	Diferença de consumo entre o mês 31 e o mês 30
Der31	Diferença de consumo entre o mês 32 e o mês 31
Der32	Diferença de consumo entre o mês 33 e o mês 32
Der33	Diferença de consumo entre o mês 34 e o mês 33
Der34	Diferença de consumo entre o mês 35 e o mês 34
Der35	Diferença de consumo entre o mês 36 e o mês 35
Media Der	Média das derivadas dos 36 meses de consumo
Mediana Der	Mediana das derivadas dos 36 meses de consumo
Moda Der	Moda das derivadas dos 36 meses de consumo
Desvio 01 Der	Desvio padrão das derivadas dos 36 meses de consumo
DesvioM Der	Desvio médio das derivadas dos 36 meses de consumo
Variancia Der	Variância das derivadas dos 36 meses de consumo
Minimo Der	Valor mínimo das derivadas dos 36 meses de consumo
Máximo Der	Valor máximo das derivadas dos 36 meses de consumo
Amplitude Der	Amplitude das derivadas dos 36 meses de consumo
Quartil 1 Der	Valor do 1º quartil das derivadas dos 36 meses de consumo
Quartil 3 Der	Valor do 3º quartil das derivadas dos 36 meses de consumo
Media 6 Cons	Média dos últimos 6 meses de consumo antes da data da inspeção
Media 12 Cons	Média dos últimos 12 meses de consumo antes da data da inspeção
Var1	Diferença entre o consumo do mês 13 e a média de consumo dos 12 meses anteriores
Var2	Diferença entre o consumo do mês 14 e a média de consumo dos 12 meses anteriores
Var3	Diferença entre o consumo do mês 15 e a média de consumo dos 12 meses anteriores
Var4	Diferença entre o consumo do mês 16 e a média de consumo dos 12 meses anteriores
Var5	Diferença entre o consumo do mês 17 e a média de consumo dos 12 meses anteriores
Var6	Diferença entre o consumo do mês 18 e a média de consumo dos 12 meses anteriores
Var7	Diferença entre o consumo do mês 19 e a média de consumo dos 12 meses anteriores
Var8	Diferença entre o consumo do mês 20 e a média de consumo dos 12 meses anteriores

Var9	Diferença entre o consumo do mês 21 e a média de consumo dos 12 meses anteriores
Var10	Diferença entre o consumo do mês 22 e a média de consumo dos 12 meses anteriores
Var11	Diferença entre o consumo do mês 23 e a média de consumo dos 12 meses anteriores
Var12	Diferença entre o consumo do mês 24 e a média de consumo dos 12 meses anteriores
Var13	Diferença entre o consumo do mês 25 e a média de consumo dos 12 meses anteriores
Var14	Diferença entre o consumo do mês 26 e a média de consumo dos 12 meses anteriores
Var15	Diferença entre o consumo do mês 27 e a média de consumo dos 12 meses anteriores
Var16	Diferença entre o consumo do mês 28 e a média de consumo dos 12 meses anteriores
Var17	Diferença entre o consumo do mês 29 e a média de consumo dos 12 meses anteriores
Var18	Diferença entre o consumo do mês 30 e a média de consumo dos 12 meses anteriores
Var19	Diferença entre o consumo do mês 31 e a média de consumo dos 12 meses anteriores
Var20	Diferença entre o consumo do mês 32 e a média de consumo dos 12 meses anteriores
Var21	Diferença entre o consumo do mês 33 e a média de consumo dos 12 meses anteriores
Var22	Diferença entre o consumo do mês 34 e a média de consumo dos 12 meses anteriores
Var23	Diferença entre o consumo do mês 35 e a média de consumo dos 12 meses anteriores
Var24	Diferença entre o consumo do mês 36 e a média de consumo dos 12 meses anteriores
Verão	Média de consumo dos meses de verão
Inverno	Média de consumo dos meses de inverno
DifVerão	Diferença entre a média dos consumos dos meses de verão do ano da inspeção com relação a média de consumo dos meses de verão do ano anterior

DifInverno	Diferença entre a média dos consumos dos meses de inverno do ano da inspeção com relação a média de consumo dos meses de inverno do ano anterior
Queda30%	Atribui um peso cada vez em que há uma queda de consumo maior ou igual a 30% entre o mês atual e o mês anterior
F1	Transformada de Fourier do mês 1
F2	Transformada de Fourier do mês 2
F3	Transformada de Fourier do mês 3
F4	Transformada de Fourier do mês 4
F5	Transformada de Fourier do mês 5
F6	Transformada de Fourier do mês 6
F7	Transformada de Fourier do mês 7
F8	Transformada de Fourier do mês 8
F9	Transformada de Fourier do mês 9
F10	Transformada de Fourier do mês 10
F11	Transformada de Fourier do mês 11
F12	Transformada de Fourier do mês 12
F13	Transformada de Fourier do mês 13
F14	Transformada de Fourier do mês 14
F15	Transformada de Fourier do mês 15
F16	Transformada de Fourier do mês 16
F17	Transformada de Fourier do mês 17
F18	Transformada de Fourier do mês 18
F19	Transformada de Fourier do mês 19
F20	Transformada de Fourier do mês 20
F21	Transformada de Fourier do mês 21
F22	Transformada de Fourier do mês 22
F23	Transformada de Fourier do mês 23
F24	Transformada de Fourier do mês 24
F25	Transformada de Fourier do mês 25
F26	Transformada de Fourier do mês 26
F27	Transformada de Fourier do mês 27
F28	Transformada de Fourier do mês 28
F29	Transformada de Fourier do mês 29
F30	Transformada de Fourier do mês 30
F31	Transformada de Fourier do mês 31
F32	Transformada de Fourier do mês 32
F33	Transformada de Fourier do mês 33
F34	Transformada de Fourier do mês 34

F35	Transformada de Fourier do mês 35