Centro de Engenharia Elétrica e Informática

Departamento de Engenharia Elétrica

JOÃO PEDRO DA COSTA SOUZA

TRABALHO DE CONCLUSÃO DE CURSO

ANÁLISE DA APLICAÇÃO DE MÁQUINAS DE VETORES DE SUPORTE NA DETECÇÃO DE PERDAS NÃO-TÉCNICAS EM SISTEMAS DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA

Campina Grande, Paraíba.
Dezembro de 2018

JOÃO PEDRO DA COSTA SOUZA

ANÁLISE DA APLICAÇÃO DE MÁQUINAS DE VETORES DE SUPORTE NA DETECÇÃO DE PERDAS NÃO-TÉCNICAS EM SISTEMAS DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA

Trabalho de Conclusão de Curso submetido à Unidade Acadêmica de Engenharia Elétrica da Universidade Federal de Campina Grande como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciências no Domínio da Engenharia Elétrica.

Área de Concentração: Processamento de Energia

Orientador:
Jalberth Fernandes de Araujo, D. Sc.

Campina Grande
Dezembro de 2018

JOÃO PEDRO DA COSTA SOUZA

ANÁLISE DA APLICAÇÃO DE MÁQUINAS DE VETORES DE SUPORTE NA DETECÇÃO DE PERDAS NÃO-TÉCNICAS EM SISTEMAS DE DISTRIBUIÇÃO DE ENERGIA ELÉTRICA

Trabalho de Conclusão de Curso submetido à Unidade Acadêmica de Engenharia Elétrica da Universidade Federal de Campina Grande como parte dos requisitos necessários para a obtenção do grau de Bacharel em Ciências no Domínio da Engenharia Elétrica.

Área de Concentração: Processamento de Energia

Aprovado em 06 / 12 / 2018

Professor Avaliador
Universidade Federal de Campina Grande
Avaliador

Jalberth Fernandes de Araujo, D. Sc.
Universidade Federal de Campina Grande
Orientador, UFCG

Dedico este trabalho à minha avó, Alaíde (*in memorian*), como desculpa por nunca ter jogado aquela última mão de sueca.

Agradecimentos

A qualquer força de existência questionável que rege o universo, por desempenhar um papel que desconheço.

Aos meus pais, Francisca Neta e José Antônio, pelo simples fato de existirem.

Aos meus irmãos, Mariana Myrtes, Ana Cecília e Carlos Victor, por sempre me lembrarem que tenho um lugar pra voltar.

Ao meu amigo e orientador pancada, Jalberth Fernandes, por toda a tutela, puxões de orelha, e por ter acreditado em mim, talvez mais do que eu mesmo acreditei.

A Helem, Iago e Rafael, pelo apoio enorme ao projeto, correções e conversas, tão construtivas para realização deste trabalho.

A Arthur Francisco, pela amizade e suporte em tantos projetos.

A Natália Caroline, minha comparsa, e a Myria Maraço, grande parceira, por me manterem vivo durante todos esses anos.

Aos meus irmãos de escolha, Kaio Nikelisson, Ítalo Marcos, Luiz Felipe, Rômulo Deyvid, Luiz Fernandes e Lucas Oliveira, por terem sido tão presentes em minha vida, ainda que distantes.

Às minhas grandes parcerias que a universidade deu de presente, Robson, Ravi, Victor, José Adeilmo Jr., Ulisses, Giuseppe, Matheus, Paulo, Ítalo e Vandilson, por toda a amizade, auxílio técnico-científico (leites) e por terem tornado a graduação suportável.

Ao meu psicanalista, Felipe Pê, pela manutenção da minha sanidade mental.

Às irmãs Campos, Iris e Ísis, e a Rafaela Myrlis, por me resgatarem do fundo do poço, tantas e tantas vezes.

Aos meus professores, que me ajudaram a construir os conhecimentos que me fizeram ultrapassar as dificuldades do curso, especialmente a Leimar, Núbia, Joelson, Edgar, Luciana, Eustáquio e Edson.

A todos que conheci, com quem conversei ou simplesmente vi, pois a experiência da graduação foi construída com vocês.

"[...] E o universo
reconstruiu-se-me sem ideal nem esperança, e o dono da Tacaria sorriu."

Álvaro de Campos.

Resumo

Perdas não técnicas representam um grave problema no setor de distribuição de energia elétrica. Para combatê-las, concessionárias se utilizam de inspeções *in loco*, que, no entanto, representam alto custo financeiro às concessionárias e se mostram ineficientes. A seleção de consumidores suspeitos se torna, assim, de grande importância, de modo que novas tecnologias baseadas em mineração de dados e aprendizado de máquina têm surgido. Nesse aspecto, Máquinas de Vetores de Suporte têm se mostrado um método bastante eficiente, porém ainda não estudado de forma concreta. Assim, objetivou-se a análise da aplicação de Máquinas de Vetores de Suporte para detecção de perdas não técnicas em sistemas de distribuição de energia elétrica. Para tanto, foi utilizado um banco de dados contendo informações de consumo de 9177 consumidores, fornecido por uma distribuidora do estado da Paraíba, sob os quais foi realizada uma série de testes que identificaram a melhor estratégia de separação entre os conjuntos de treinamento e teste a ser utilizada, o melhor período de tempo e proporção de perdas não técnicas do banco de dados e o melhor tipo de entrada. Os melhores resultados foram obtidos com utilização da estratégia de separação de bases de validação cruzada com 10 *folders*, com 36 meses de consumo e proporção de perdas não técnicas no banco de dados de 50%. O melhor tipo de entrada foram as informações de consumo, sendo alcançadas taxas de sucesso em inspeção de até 76,6% e acurácia na detecção de irregularidades de 63,32%, que são resultados bastante superiores ao dos métodos utilizados atualmente pelas distribuidoras de energia elétrica. A aplicação de Máquinas de Vetores de Suporte se mostra, assim, uma alternativa viável para detecção de perdas não técnicas em sistemas de distribuição de energia elétrica.

Palavras-chave: Perdas Não Técnicas, Máquinas de Vetores de Suporte, Mineração de Dados, Distribuição de Energia Elétrica

Abstract

Non-technical losses represent a serious problem in electric power distribution sector. To combat them, concessionaires use on-site inspections which represent a high financial cost to utilities and are inefficient. Therefore, the selection of suspect consumers carries great importance, so that new technologies based on Data Mining and Machine Learning have emerged. In this regard, Support Vector Machines have proved to be a very efficient method, but has not been studied in a concrete way. The purpose of this study was to analyze the application of Support Vector Machines to detect non-technical losses in electric power distribution systems. A database containing consumer information from 9177 consumers, provided by a distributor from the state of Paraíba, Brazil, was used to carry out a series of tests that identified the best strategy for separating training and test sets to be used, the best time period, proportion of non-technical losses in database, and the best type of input. The best results were obtained with the use of 10 folders – cross validation, with 36 months of consumption and proportion of non-technical losses in the database of 50%. The best type of input was consumption information, with inspections success rates of up to 76.6% and accuracy of detecting irregularities of 63.32%, which are much better than the methods currently used by energy utilities. The application of Support Vector Machines is thus a viable alternative for the detection of non-technical losses in power distribution systems.

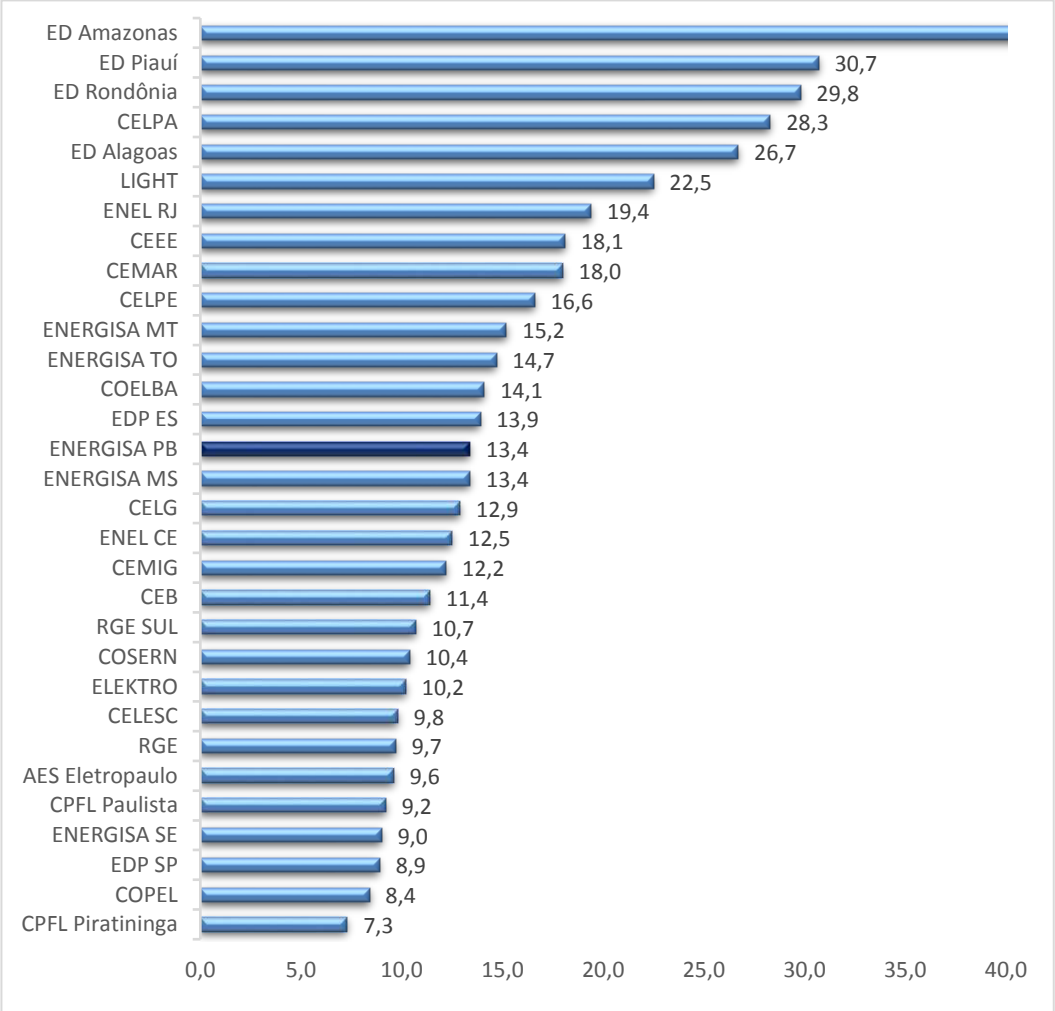
Keywords: Non-technical losses, Support Vector Machines, Data Mining, Power Distribution

SUMÁRIO

1 INTRODUÇÃO

Uma das maiores adversidades enfrentadas no setor de distribuição de energia elétrica são as perdas, as quais podem atingir níveis relativamente altos e são objeto de cuidadoso monitoramento por parte das empresas do setor. De acordo com dados da Associação Brasileira de Distribuidores de Energia Elétrica (ABRADEE), o percentual de perdas elétricas no Brasil foi de 13,9% em 2017 (ABRADEE, 2017). A título de ilustração, o percentual de perdas do sistema global no referido ano é apresentado na Figura 1.

Figura 1 – Perdas elétricas por distribuidora no Brasil em 2017.



Fonte: adaptado de ABRADEE (2017).

De acordo com Monedero *et al.* (2012), perdas elétricas podem ser divididas em perdas técnicas, que são inerentes ao sistema elétrico e originam-se de fenômenos físicos como o efeito Joule; e perdas não técnicas (PNT), que estão associadas à energia elétrica não contabilizada ou à energia elétrica contabilizada, porém não paga. A não contabilização da energia elétrica pode decorrer de erros no medidor ou de comportamentos indevidos de consumidores, como fraude e furto de energia elétrica.

O aumento das perdas não técnicas ocasiona não apenas redução de receita para as distribuidoras, como também aumento no faturamento dos consumidores não fraudadores, tendo em vista o aumento de taxas de contribuição, além de diminuição na arrecadação de impostos pelo estado. Desse modo, a identificação de fontes de perdas não técnicas torna-se de interesse geral.

O problema é comumente enfrentado pelas distribuidoras a partir de campanhas de prevenção e da realização de inspeções técnicas nos locais de consumo. Entretanto, além da inspeção de todos os consumidores atendidos ser inviável, os métodos de inspeção utilizados atualmente se mostram, em geral, ineficientes e onerosos, o que motiva o estudo de novas abordagens de combate ao problema (VIEGAS *et al*, 2017).

A detecção de PNT tem recebido interesse crescente da academia e da indústria, que fomentam o desenvolvimento de técnicas focadas em Mineração de Dados (capazes de tratar uma grande quantidade de dados) e Aprendizado de Máquina na busca pela determinação de um modelo que permita identificar possíveis unidades fraudadoras. Uma das grandes vantagens dessas técnicas é o baixo custo financeiro, enquanto uma de suas principais desvantagens é que os métodos são baseados em uma aprendizagem fundamentada em dados do passado, o que não garante a resolução de problemas do presente e futuro (ARAUJO, 2017; QUEIROGA, 2005).

A respeito de técnicas relacionadas ao aprendizado de máquina, ainda não há solução unificada para a detecção de perdas não técnicas. No entanto, o algoritmo de Máquinas de Vetores de Suporte ou, no inglês, *Support Vector Machine* (SVM), é um mais utilizados na bibliografia, e tem se destacado na classificação de PNT devido ao seu alto desempenho (superior a 60% de classificações corretas). Ademais, SVM possui menor complexidade em comparação a outras técnicas, como redes neurais artificiais (VIEGAS *et al.*, 2017).

Para buscar reduzir custos e fornecer um melhor atendimento aos consumidores, pode-se empregar a técnica SVM com o intuito de detectar perdas não técnicas nos sistemas de distribuição de energia elétrica. A aplicação da técnica, entretanto, ainda não

está completamente descrita na bibliografia. Informações como a quantidade de meses a ser utilizada como entrada, tratamento do desbalanceamento dos bancos de dados e das melhores entradas ainda não estão bem definidas. A determinação dessas características é essencial para implementação eficiente da técnica.

1.1 OBJETIVOS GERAIS E ESPECÍFICOS

O objetivo geral deste trabalho é analisar a aplicação da técnica de Máquinas de Vetores de Suporte para detecção de perdas não-técnicas em sistemas de distribuição de energia elétrica.

Além disso, este trabalho possui os seguintes objetivos específicos:

- Avaliar qual a melhor estratégia de separação de bases para aplicação da técnica, com relação ao banco de dados empregado.
- Determinar o período de tempo ótimo a ser utilizado para detecção de perdas não técnicas em sistemas de distribuição com utilização de Máquinas de Vetores de Suporte;
- Analisar as implicações do desbalanceamento do banco de dados com relação ao número de unidades consumidoras regulares na detecção de perdas não técnicas;
- Avaliar diferentes tipos de entrada para o método, como informações de consumo puras, variações entre os consumos mensais e variações referentes às variações de consumo mensais, além de dados estatísticos obtidos dessas informações, comparando-os.

1.2 ORGANIZAÇÃO DO EXTO

Este trabalho está organizado em seis seções, descritas a seguir:

Nesta seção, o problema de perdas elétricas foi apresentado, juntamente à proposta do trabalho e seus objetivos. Também é descrita a organização do documento.

Na seção 2, é realizado o embasamento teórico relativo a perdas não técnicas, mineração de dados, aprendizado de máquina, com foco no método de Máquinas de Vetores de Suporte e em suas principais características. Também são feitas algumas considerações sobre bancos de dados e sua utilização na detecção de perdas não técnicas.

Na seção 3, a revisão bibliográfica é apresentada, evidenciando-se os principais trabalhos que tratam da detecção de perdas não técnicas, especialmente a partir de Máquinas de Vetores de Suporte, e que inspiraram o presente estudo.

Na seção 4 é apresentada a metodologia empregada para análise da aplicação de Máquinas de Vetores de Suporte realizada.

Não seção 5, os resultados da pesquisa são descritos e discutidos.

A seção 6 compreende às conclusões e sugestões de trabalhos futuros, visando à continuidade da pesquisa apresentada.

Ao fim são apresentadas as referências utilizadas neste trabalho.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção, são apresentados os conceitos teóricos necessários para o entendimento dos procedimentos adotados e para a análise dos resultados obtidos. Os fundamentos relativos a perdas elétricas não técnicas, mineração de dados, aprendizado de máquina e, por fim, Máquinas de Vetores de Suporte são destacados a seguir.

2.1 PERDAS NÃO-TÉCNICAS

Perdas não-técnicas (PNT), também denominadas de perdas comerciais, são aquelas associadas à comercialização e pagamento de energia elétrica fornecida ao usuário e se referem à toda energia elétrica entregue, porém não faturada, provocando prejuízo às concessionárias (RAMOS *et al.,* 2016). As PNT englobam ocorrências como inadimplência, problemas de cobrança e as duas principais categorias, fraude e furto de energia elétrica.

O furto é caracterizado pelo desvio direto de energia da rede elétrica pelo consumidor clandestino (não cadastrado pela distribuidora), de modo que a energia utilizada não é contabilizada. Em geral, as ligações clandestinas são realizadas no alimentador de baixa tensão ou no transformador de serviço, o que torna a identificação visual possível. Entretanto, por estar normalmente associado a áreas de risco, a fiscalização e o combate a esse tipo de PNT tornam-se de difícil execução. Ademais, a identificação do furto de energia elétrica a partir de dados de consumo é inviável, tendo em vista que o consumidor não está cadastrado na concessionária (SAISSE, 2016).

Na fraude, em contraposição ao furto, ocorre registro do consumidor por parte da distribuidora. Porém, são realizadas intervenções que alteram as marcações de medição de consumo. Assim, apesar de consumida uma determinada quantidade de energia elétrica, o consumidor paga efetivamente apenas por uma fração do consumo total (ABRADEE, 2017).

As formas de intervenção na fraude são diversas, variando das mais rústicas até às mais sofisticadas, como isolação do neutro, inversão de fases e manipulações indevidas no medidor. Alterações no medidor, em especial, representam uma grande parcela dos tipos de fraude. Informações adicionais acerca dos tipos de irregularidades em medidores

de energia elétrica podem ser obtidas em Foiatto (2009), que destaca a importância do estudo para a redução de PNT, um problema enfrentado em todo o território nacional.

No Brasil, a quantidade de energia elétrica perdida por motivos de irregularidades, furtos e fraudes vem crescendo nos últimos anos, como é demonstrado na Figura 2. De acordo com dados da ABRADEE (2017), em 2017 as perdas comerciais no Brasil alcançaram 6,08% da energia total injetada no sistema global, composto por 63 distribuidoras (ABRADEE, 2017; ARAUJO, 2017).

Figura 2 - Percentual de perdas não técnicas em relação à energia injetada no sistema global.

Fonte: adaptado de ABRADEE (2018).

A partir da Figura 2, nota-se certa estabilidade das PNT na última década, de modo que elas representam um grande problema às distribuidoras de energia elétrica. De acordo com relatório da ANEEL, PNT correspondem a mais de R$ 12,3 bilhões em tarifas, o que equivale a 8% da receita do setor ou 29% da receita das distribuidoras (R$ 42 bilhões). Esse prejuízo é considerado como receita irrecuperável, sendo repassado à sociedade como um todo, prejudicando consumidores corretos e adimplentes (ANEEL, 2018).

Além dos prejuízos financeiros, furto e fraude de energia elétrica representam riscos à segurança da população, como o de curto-circuito, e ao fornecimento de energia elétrica. Ligações clandestinas podem ainda interferir negativamente na qualidade da energia elétrica (ROJAS & GALLEGO, 2015).

O combate às PNT é realizado pelas empresas em diversos âmbitos, tais como a regularização de focos de possíveis ligações clandestinas, como favelas, a implementação de políticas comerciais, como o atendimento à comunidade a partir de explicações, negociações e treinamento sobre consumo de energia elétrica. O principal meio de combate reside nos programas de inspeção, que têm por objetivo à verificação da integridade do sistema de medição, detectando falhas ou adulterações no medidor, fraudes e desvio de energia elétrica (RAMOS, 2014).

As inspeções são realizadas *in loco*, com atuação de técnicos especializados. A abordagem mais comum é a inspeção guiada pela seleção de unidades consumidoras (UC) consideras suspeitas ou pela técnica de varredura, em que uma área é escolhida e percorrida por uma equipe de fiscais no intento de identificar possíveis perdas comerciais (MONEDERO *et al.*, 2012).

As inspeções são normalmente realizadas com base em um conjunto de metodologias heurísticas para identificar os clientes suspeitos. Entretanto, a utilização desse método representa alto custo financeiro para as concessionárias, porém possui baixa eficiência, com índice de sucesso das inspeções variando entre 10% e 30% (CHAUHAN & RAJVANSHI, 2013; NIZAR & DONG, 2009; QUEIROGA, 2005). Em suma, a cada 10 inspeções, no máximo 3 resultam na identificação de irregularidades.

Outro fator agravante é a impossibilidade de se inspecionar todos os consumidores, de forma que a seleção correta de consumidores a serem inspecionados constitui o melhor mecanismo para otimização do combate às PNT (QUEIROGA, 2005; TREVIZAN *et al.,* 2015).

Diversas metodologias têm sido propostas visando à seleção mais eficaz de consumidores suspeitos. Dentre elas, metodologias baseadas no processamento de informações e características dos consumidores a partir de Mineração de Dados e Aprendizado de Máquina tem se destacado pela flexibilidade e taxas de sucesso acima de 60%. A fundamentação teórica necessária para o entendimento do processo de Mineração de Dados e da tecnologia de Aprendizado de Máquina aplicados à detecção de PNT são apresentados a seguir.

2.2 MINERAÇÃO DE DADOS

Mineração de Dados (*Data Mining)* é uma tecnologia em desenvolvimento, que pode ser descrita como a combinação das tecnologias de inteligência artificial e bancos

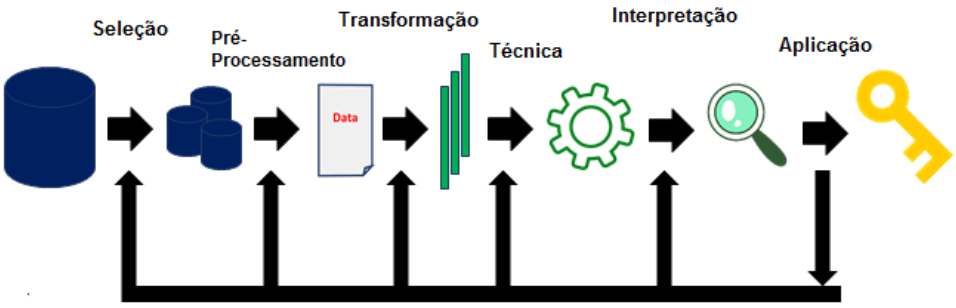
de dados, visando à obtenção de conhecimentos úteis às soluções dos mais diversos problemas. É um processo cujo objetivo principal é analisar, extrair e armazenar informações relevantes de grandes quantidades de informações dispersas (não refinadas) a partir de técnicas de reconhecimento e identificação de padrões (QUEIROGA, 2005; ROJAS & GALLEGO, 2015).

De acordo com Fayyad (1996), o processo de Mineração de Dados se divide em diversas etapas, que podem ser reiniciadas de forma cíclica, e que culminam na obtenção de um determinado conhecimento. As etapas estão a seguir:

- *Seleção de dados:* trata-se da correta seleção de dados que servirão de base para obtenção do conhecimento em foco, destacando-se as informações úteis para solução do problema;
- *Pré-processamento:* refere-se à preparação dos dados selecionados, tratando dados desconhecidos e/ou discrepantes que podem prejudicar a análise. Engloba procedimentos como a normalização de dados;
- *Transformação de dados:* refere-se à redução dos dados pré-processados, para aumentar a eficiência do processo e adaptação, representando as informações da melhor forma possível.
- *Aplicação de técnica e reconhecimento de padrões:* trata-se da aplicação de técnicas que extraiam padrões e modelem o problema proposto.

Ao fim das etapas, obtém-se o conhecimento útil, que pode então ser aplicado e trabalhado. O processo de Mineração de dados pode ser resumido pelo diagrama apresentado na Figura 03.

Figura 3 – Processo de Mineração de Dados.



Fonte: adaptado de Fayyad (1996).

A Mineração de Dados é utilizada com frequência para detecção de fraude em sistemas que contém grande quantidade de dados, como cartões de crédito (GOSH &

REILLY, 1994), sendo uma ferramenta importante para tomada de decisões. Outro problema que tem sido alvo da tecnologia é a detecção de PNT em sistemas de distribuição (QUEIROGA, 2005; RAMOS, 2014).

As metodologias de detecção de PNT se baseiam nas informações de consumo e em características de consumidores, visando a identificação de padrões nos quais possa ser identificado o perfil de uma unidade irregular a partir de modelos (ROJAS & GALLEGO, 2015). A etapa de identificação de padrões é realizada a partir de diversas técnicas de Aprendizado de Máquina, tais como Árvores de Decisão, Redes Neurais Artificiais (RNA) e Máquinas de Vetores de Suporte (*Support Vector Machines – SVM*). Os principais aspectos relativos ao Aprendizado de Máquinas são tratados a seguir.

2.3 APRENDIZADO DE MÁQUINA

O Aprendizado de Máquina se caracteriza pela capacidade de uma máquina de alterar sua estrutura, programação ou dados, de modo a melhorar seu desempenho futuro, e se relaciona diretamente à detecção automática de padrões significativos em dados. Nos últimos anos, essa ferramenta se tornou comum para qualquer tarefa que envolva extração de informações de grandes bancos de dados (SHALEV-SHWARTZ & BEN-DAVID, 2014).

O aprendizado de máquina pode ser divido em aprendizado supervisionado e não supervisionado. No primeiro, é necessário que exemplos sejam fornecidos com base em classes pré-estabelecidas, de modo que a classificação de um novo exemplo possa ser realizada. No aprendizado não supervisionado, por sua vez, ocorrências semelhantes são selecionadas e agrupadas, de modo que são configuradas novas classes (SMOLA & VISHWANATHAN, 2008).

Um dos principais problemas que o aprendizado de máquina busca solucionar é o de classificação binária que consiste em classificar um dado em duas classes distintas, como é o caso da classificação entre consumidores regulares e irregulares em sistemas de distribuição de energia elétrica. Para resolver problemas como o da classificação de consumidores regulares e irregulares, diversas técnicas foram propostas, como a de Máquinas de Vetores de Suporte (SVM).

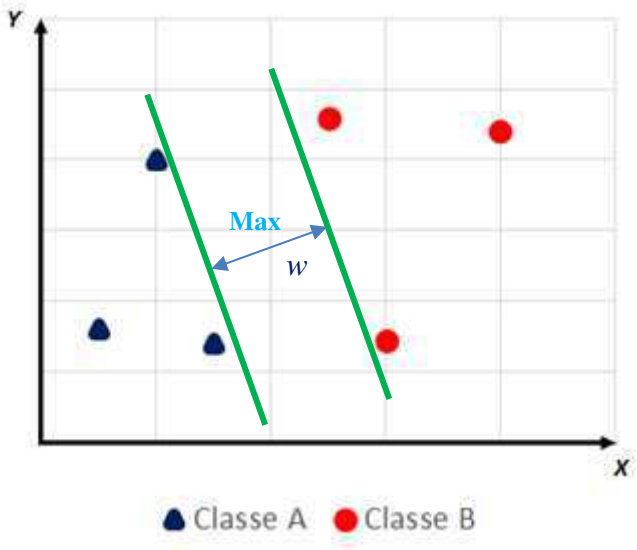
2.3.1 MÁQUINAS DE VETORES DE SUPORTE

Máquinas de Vetores de Suporte (SVM), do inglês *Support Vector Machines*, foram introduzidas por Vapnik e Lerner (1963), e são um grupo de métodos de aprendizado supervisionado, com os quais se que pode analisar dados e reconhecer padrões ou tendências com respeito à saída (DEPURU *et al.,* 2011).

SVM têm sido um dos grupos de métodos mais utilizados para classificação de consumidores suspeitos, para detecção de PNT (VIEGAS *et al.*, 2017). A técnica possui fundamentação mais sólida do que Redes Neurais Artificiais e pode substituí-las com desempenho melhor ou semelhante (RAMOS, 2014).

O funcionamento de SVM se baseia no desenvolvimento de um hiperplano (ou conjunto de hiperplanos) baseado nos dados a serem classificados, de modo a promover a maior distância de separação possível entre as classes a serem separadas. Quanto maior a distância que separa as classes, menor o erro do classificador (CHAUHAN & RAJVANSHI, 2013). O princípio básico de funcionamento pode ser ilustrado na Figura 4.

Figura 4 – Funcionamento de SVM na separação de classes.



Fonte: adaptado de Scholkopf & Smola (2002).

Na Figura 4, as classes A (triângulos azuis) e B (círculos vermelhos) são linearmente separáveis a partir de uma distância *w*, que deve ser a máxima possível. A base matemática do algoritmo de Máquinas de Vetores de Suporte está apresentada a

seguir. Como enunciado, os princípios iniciais foram desenvolvidos por Vapnik e Lerner (1963). A fundamentação a seguir é baseada na interpretação de Ramos (2014).

Considerando-se uma classe de hiperplanos $\boldsymbol{\mathcal{H}}$ com produto interno indicado pela Equação (1).

$$\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = 0\,, \tag{1}$$

em que $\boldsymbol{w} \in \boldsymbol{\mathcal{H}}$, $b \in \mathbb{R}$, correspondendo a funções de decisão da forma apresentada na Equação (2).

$$f(x) = sgn(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b). \tag{2}$$

Em que $sgn()$ representa a função sinal. De acordo com Vapnik (1963), dentre todos os hiperplanos que separam os dados, apenas um hiperplano ótimo é distinguido pela margem de máxima separação entre qualquer ponto de treinamento e este hiperplano. Essa solução é dada a partir da Equação (3):

$$\max_{\boldsymbol{w} \in \boldsymbol{\mathcal{H}}, b \in \mathbb{R}} \min\{ \|\boldsymbol{x} - \boldsymbol{x_i}\| \; \boldsymbol{x} \in \mathcal{H}, \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = 0\}. \tag{3}$$

Ademais, a capacidade da classe de hiperplanos de separação decresce com o crescimento da margem de separação. A construção de um hiperplano ótimo é obtida a partir da resolução da Equação (4).

$$\min_{\boldsymbol{w} \in \boldsymbol{\mathcal{H}}, b \in \mathbb{R}} \tau(\boldsymbol{w}) = \frac{1}{2}\|\boldsymbol{w}\|^2, \tag{4}$$

em que τ é denominada de função objetivo. O hiperplano está sujeito às restrições de desigualdade representadas pela Inequação (5),

$$y_i(\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b) \geq 1 \text{ para todo } i = 1, 2, \cdots, m, \tag{5}$$

garantindo que $f(x_i)$ será $+1$ para $y_i = +1$ e para -1 e para $y_i = -1$, e também fixando a escala de $\boldsymbol{w}$, sendo *m* a dimensão do vetor. A equação (4) e a inequação (5) originam o denominado problema de otimização restrita, resolvido a partir do Lagrangiano apresentado na Equação (6),

$$\mathrm{L}(\boldsymbol{w}, b, \alpha) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{m} \alpha_i (y_i(\langle \boldsymbol{x_i}, \boldsymbol{w} \rangle + b) - 1), \tag{6}$$

em que $\alpha_i \geq 0$ são os multiplicadores de *Lagrange* e *L* tem que ser minimizada com relação às variáveis *w* e *b* e maximizada com relação a α_i, de modo que deve ser encontrado um ponto de sela. Tem-se então a Equação (7).

$$\frac{\partial}{\partial b} L(\boldsymbol{w}, b, \alpha) = 0 \text{ e } \frac{\partial}{\partial \boldsymbol{w}} L(\boldsymbol{w}, b, \alpha) = 0, \quad (7)$$

Que, por sua vez, origina as Equações (8) e (9),

$$\sum_{i=1}^{m} \alpha_i y_i = 0\,, \quad (8)$$

$$\boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x_i}\,, \quad (9)$$

de forma que o vetor-solução se resume em uma expansão de um subconjunto dos padrões de treinamento, especificamente aqueles em $\alpha_i \neq 0$. Sendo esses os denominados vetores de suporte ou *Support Vectors.*

Substituindo as Equações (8) e (9) no Lagrangiano, eliminam-se as variáveis $\boldsymbol{w}$ e b, resultando no problema de otimização dual, alvo das Máquinas de Vetores de Suporte, que é representado pela Equação (10).

$$\max_{\alpha \epsilon \Re^m} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle, \quad (10)$$

e que está sujeito à Equação (11)

$$\alpha_i \geq 0 \text{ para todo } i = 1, \ldots, m \quad (11)$$

e à Equação (12)

$$\sum_{i=1}^{m} \alpha_i y_i = 0. \quad (12)$$

A função hiperplano de decisão (Equação (2)), pode então ser escrita conforme apresentado na Equação (13).

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i \langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle + b \right). \quad (13)$$

Para o cálculo de b, tem-se que todos os pontos definidos pela Equação (14),

$$\alpha_i [y_i (\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b) - 1] = 0 \text{ para todo } i = 1, \ldots, m, \quad (14)$$

são vetores de suporte na margem e, assim, $\alpha_i > 0$, e assim tem-se a Equação (15),

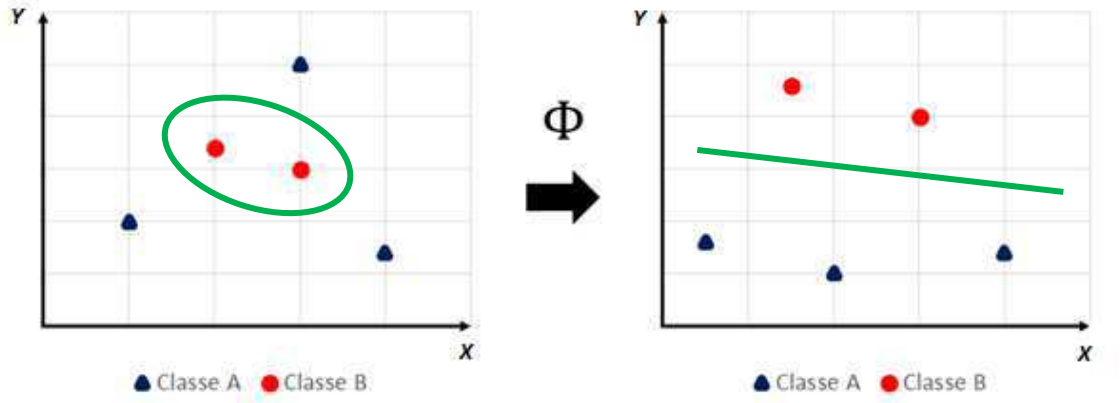
$$\langle \boldsymbol{x_i}, \boldsymbol{w} \rangle + b = y_i, \quad (15)$$

e a Equação (16),

$$b = y_i - \langle \boldsymbol{x_i}, \boldsymbol{w} \rangle. \quad (16)$$

Para o caso de classes não-separáveis, é necessário realizar um mapeamento das entradas para um espaço de maior dimensionalidade, acrescentando-se artifícios de tolerância às SVM. Em suma, o procedimento se configura como a transformação do espaço de entradas, que representa um problema não linearmente separável, em um espaço de características, normalmente, de maior dimensionalidade e linearmente separável, como indicado na Figura 5.

Figura 5 – Ilustração da transformação do espaço de entradas.



Fonte: adaptado de Scholkopf & Smola (2002).

Matematicamente, tem-se que a partir dos padrões de entrada $\mathcal{X}$, o produto interno dos vetores $\boldsymbol{x}$ e $\boldsymbol{x'}$ é empregado em termos do núcleo u estimado pelos elementos de entrada x e x', como é apresentado na Equação (17),

$$u(x, x') = \langle \boldsymbol{x}, \boldsymbol{x'} \rangle, \quad (17)$$

sendo essa substituição conhecida como truque de núcleo ("*kernel trick*"), utilizada para o mapeamento não linear, que objetiva tornar os dados linearmente separáveis.

O truque de núcleo pode ser usado desde que todos os vetores de característica ocorram apenas em produtos internos, o que pode ser observado nas equações (10) e (13). Dessa maneira, o vetor ponderado, determinado na Equação (9), não corresponderá mais ao seu respectivo vetor no espaço de entradas, tornando-se uma expansão no espaço de características. A função de decisão torna-se então a Equação 18

$$f(x) = sgn\left(\sum_{i=1}^{m} y_i \alpha_i \langle \boldsymbol{\phi}(x), \boldsymbol{\phi}(x_i) \rangle + b\right) = sgn\left(\sum_{i=1}^{m} y_i \alpha_i u(x, x_i) + b\right), \quad (18)$$

de modo que o problema de otimização assumirá a forma apresentada na Equação (19)

$$\max_{\alpha \epsilon \Re^m} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j u(x_i, x_j), \quad (19)$$

Também está sujeito às condições 1 e 2. Entretanto, é possível que o hiperplano de separação não exista, como é o caso de bancos de dados com ruído em que haja sobreposição de classes. Introduz-se então as denominadas variáveis de afrouxamento ρ_i, apresentadas na Equação (20):

$$\rho_i \geq 0 \text{ para todo } i = 1, \ldots, m, \quad (20)$$

o que leva às restrições a assumirem a forma apresentada na Equação (21).

$$y_i(\langle \boldsymbol{w}, \boldsymbol{x_i} \rangle + b) \geq 1 - \rho_i \text{ para todo } i = 1, \ldots, m. \quad (21)$$

A função-objetivo, definida na Equação (4), transforma-se na Equação (22):

$$\tau(\boldsymbol{w}, \rho) = \frac{1}{2} \|\boldsymbol{w}\|^2 + C \sum_{i=1}^{m} \rho_i, \quad (22)$$

em o equilíbrio entre maximização da margem de separação entre classes e a minimização de erros durante o treinamento é determinado a partir da condição que $C > 0$. Ademais, quanto maior o valor de C, maior o peso das variáveis de afrouxamento, o que reduz os erros de treinamento, mas afeta a capacidade de generalização da máquina.

O problema pode, então, ser reescrito em termos dos multiplicadores de Lagrange, o que resulta novamente na maximização apresentada na Equação (19), sujeita às restrições impostas pelas Equações (23) e (24),

$$0 \leq \alpha_i \leq C \text{ para todo } i = 1, \ldots, m. \quad (23)$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0. \quad (24)$$

Em que a primeira restrição (Equação (23)) ganha limite superior baseado em C, o que torna a influência dos padrões individuais limitada, enquanto a segunda restrição (Equação (24)) permanece idêntica à do problema linearmente separável (Equação (12)).

Por fim, para o caso não separável, b pode ser calculado tendo em vista que para os vetores de suporte x_i com $\alpha_i \leq C$, a variável de afrouxamento ρ é nula e, assim, tem-se a Equação (25),

$$\sum_{j=1}^{m} \alpha_j y_j u(x_i, x_j) + b = y_i . \tag{25}$$

A função de núcleo apresentada como produto interno na Equação (17) é chamada de função Núcleo Linear. Outra função de mapeamento bastante utilizada é a Gaussiana, chamada de Função de Base Radial ou RBF, do inglês *Radial Basis Function*. A função de núcleo é apresentada na Equação (26),

$$u(x, x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}}, \tag{26}$$

Em que $\sigma > 0$ é um parâmetro da função do núcleo.

Expostos os conceitos relativos à aplicação das Máquinas de Vetores de Suporte, torna-se necessário apresentar os critérios de avaliação de modelos de aprendizado supervisionado, iniciando com a separação entre os conjuntos de treinamento e teste.

2.3.2 Conjuntos de Treinamento e Teste

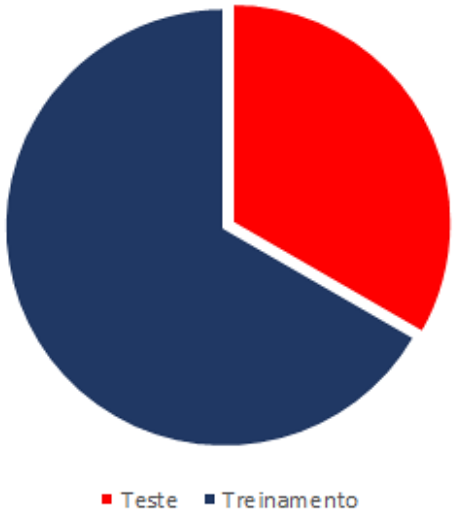
Antes de tratar da avaliação de modelos de Aprendizado de Máquina como as SVM, é necessário estabelecer os fundamentos necessários acerca da separação entre os conjuntos de treinamento, teste e validação.

Em técnicas de aprendizado supervisionado, o modelo proposto deve ser desenvolvido a partir dos dados de treinamento, com os resultados de seu aprendizado aplicados a dados reais. Nesse aspecto, pode-se dizer que a característica mais desejável de um modelo é a generalização da habilidade de classificação aprendida durante a fase de treinamento para novos conjuntos, ainda não vistos (SAISSE, 2016).

Assim, o conjunto de treinamento apenas treina o modelo. Porém, a avaliação dos resultados finais deve ser realizada sobre o conjunto de testes. Nenhum parâmetro deve ser modificado, nem treinamento realizado para aplicação do teste, para que os resultados não sejam enviesados. A separação do conjunto de dados nestes dois grupos de forma adequada é essencial e pode ser realizada a partir de diversas técnicas. As principais são a técnica de *Holdout* e a de Validação Cruzada (ARAUJO, 2017).

A separação entre conjuntos de treinamento e teste, também conhecida como *Holdout*, consiste na realização de um único experimento, que é treinar e testar, com divisão usual de 2/3 dos dados para treinamento e 1/3 para testes. Entretanto, uma única partição pode levar a resultados imprecisos, especialmente para pequenos e médios bancos de dados (menos de 100 mil instâncias), tendo em vista que pode enviesar o modelo, especializando-o. Essa especialização é conhecida como *overfitting* (QUEIROGA, 2005). A técnica de separação *Holdout* está ilustrada na Figura 6:

Figura 6 – Separação de bases a partir da técnica de *Holdout*.

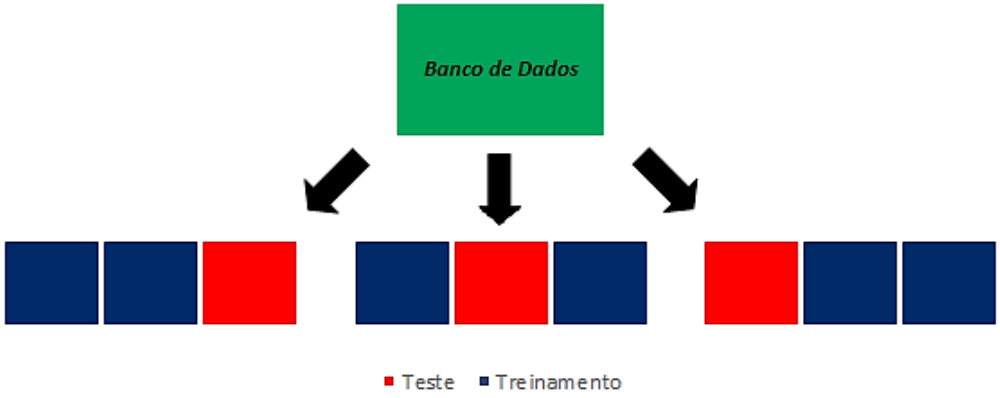


Fonte: adaptado de Queiroga (2005).

Para evitar especialização (*overfitting*) do modelo em relação ao conjunto de treinamento, utiliza-se em geral outro método de separação de conjuntos, denominado Validação Cruzada. Na Validação Cruzada, também conhecida como *k-fold Cross Validation*, o conjunto de dados é dividido em k subconjuntos (*folds*) mutuamente exclusivos de tamanho aproximadamente igual. O classificador é então treinado com *k-1 folds,* deixando o último como conjunto de testes. O processo se repete k vezes, com revezamento do *fold* de testes, sendo o desempenho do classificador tomado como média das classificações (QUEIROGA, 2005).

Devido ao número de etapas, essa técnica pode gerar alto custo computacional para grandes bancos de dados (mais de 100 mil instâncias), sendo mais utilizada para conjuntos menores. A utilização de $k = 10$ é bastante comum, sendo suficiente para representação eficiente da população (ARAUJO, 2017). A técnica está ser ilustrada na Figura 7.

Figura 7 – Separação de bases a partir da técnica de *Cross Validation.*



Fonte: adaptado de Araujo (2017).

Discutidas as estratégias de separação de bases, é importante observar as formas de como avaliar os modelos, de modo a compará-los e identificar os melhores.

2.3.3 AVALIAÇÃO DOS MODELOS

A avaliação dos modelos de Aprendizado de Máquina é realizada a partir da análise de diversas métricas de desempenho. No caso de problemas de classificação binária, uma ferramenta comumente utilizada para avaliação é a matriz de confusão.

Uma matriz de confusão é uma tabela 2x2 que contém os quatro possíveis resultados da aplicação de um classificador binário. Na Tabela 1, A, B, C e D representam a quantidade de elementos que se enquadram em cada uma das posições descritas a seguir (ARAUJO, 2017).

- Verdadeiros Positivos (VP) – quantidade de predições corretas da classe "Positivo".
- Falsos Positivos (FP) – quantidade de predições incorretas para a classe "Positivo".
- Falso Negativo (FN) – quantidade de predições incorretas para a classe "Negativo".
- Verdadeiro Negativo (VN) – quantidade de predições corretas para a classe "Negativo".

Tabela 1: Matriz de confusão.

		Referência	
		Positivo	**Negativo**
Predição	**Positivo**	**VP**	**FP**
	Negativo	**FN**	**VN**

Fonte: adaptado de Araujo (2017).

Diversas métricas de desempenho podem ser derivadas a partir da matriz de confusão. Algumas das principais métricas utilizadas são descritas a seguir:

- Acurácia (*Acc*): refere-se à quantidade de classificações corretas dentre todas as classificações realizadas. Em suma, representa a exatidão do modelo (ARAUJO, 2017). No que tange à detecção de PNT, trata-se da quantidade de classificações corretas de consumidores como regulares e irregulares, medindo o número total de acertos. Pode ser definidada pela Equação (27):

$$Acc = \frac{VP + VN}{VP + FP + VN + FN}. \tag{27}$$

- Sensibilidade (*Sen*): definida por (28), também conhecida como *recall,* é uma medida da proporção de verdadeiros positivos (VP), ou seja, de irregularidades, encontradas (GLAUNER *et al.*, 2016).

$$Sen = \frac{VP}{VP + FN}. \tag{28}$$

- Especificidade (*Esp*): a especificidade se refere ao total de classificações corretas relativas à classe que não é a de interesse, ou seja, de consumidores regulares. A partir da especificidade pode-se ainda obter o índice de falsos negativos, dado por *1 – Esp.* (GLAUNER *et al.,* 2016). Pode ser definida a partir da Equação (29).

$$Esp = \frac{VN}{VN + FP}. \tag{29}$$

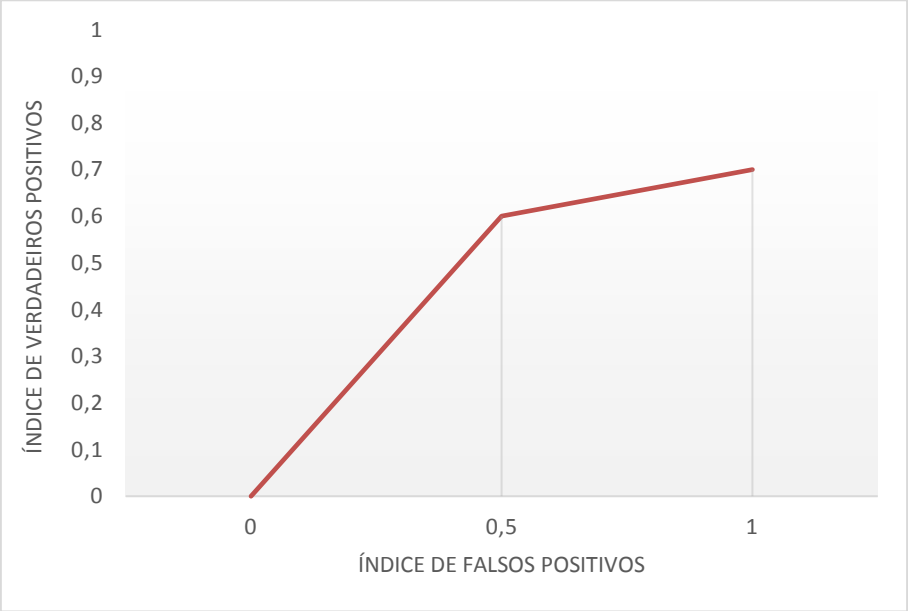
- Precisão (*Pre*): indica a quantidade de instâncias classificadas corretamente relativamente à classe de interesse dentre todas as classificações de classe positiva

realizadas (ARAUJO, 2017). No âmbito de detecção de PNT, refere-se ao número de irregularidades dentre todas as UC classificadas como irregulares. Pode definida a partir da Equação (30).

$$Pre = \frac{VP}{VP + FP}. \tag{30}$$

- Área abaixo da Característica de Operação do Receptor ou *Receiving Operator Characteristic* – ROC*:* a partir da ROC é possível relacionar a sensibilidade (*recall*) e o índice de falsos positivos *(1 – Esp*) , sendo particularmente útil na detecção de PNT, pois permite confrontar resultados positivos e negativos de inspeção, como é ilustrado na Figura 8. A área abaixo da ROC, AUC (do inglês *area under the curve*) é uma métrica de desempenho que varia entre 0 e 1. Para problemas de classificação binários, tem-se que uma $AUC > 0{,}5$ indica chance melhor que a sorte (GLAUNER *et al.*, 2016).

Figura 8 – Exemplo de Receiving Operator Curve.



Fonte: adaptado de Witten *et al.* (2016).

Para avaliar os modelos, diversas métricas de desempenho podem ser utilizadas, a depender do objetivo final. Pode-se estabelecer uma função de desempenho f_d, baseada nas métricas mais importantes para o problema proposto, de modo a tornar possível a comparação de modelos (ARAUJO, 2017).

Feitas as observações acerca de Máquinas de Vetores e avaliação de métodos de Aprendizado de Máquina, algumas considerações sobre bancos de dados para utilização na detecção de PNT devem ser feitas.

2.3.4 Considerações Sobre Bancos de Dados na Determinação de PNT

A utilização de Aprendizado de Máquina Supervisionado, como é o caso do método baseado em SVM, torna necessário o conhecimento prévio da classificação de cada elemento para realização da etapa de treinamento. Desse modo, cada UC deve ter sido inspecionada *in loco*, determinando-se a presença ou não de irregularidade. Outro ponto se relaciona ao desbalanceamento do banco de dados.

Em sistemas reais, o número de UC irregulares é muito inferior ao de regulares, de modo que Viegas *et al.* (2017) constataram em seu trabalho que o desbalanceamento de dados é um dos maiores problemas enfrentados em estudos de detecção de PNT.

Entretanto, no que se refere ao treinamento de classificadores, Glauner *et al.* (2016) observaram que a eficiência do modelo melhora de modo considerável para bancos de dados com 60% de consumidores irregulares.

Os dados de entrada mais comuns em métodos que utilizam mineração de dados e aprendizado de máquina para detecção de PNT são as informações de consumo dos consumidores (VIEGAS *et al.,* 2017). Alguns trabalhos, como o de Depuru *et al.* (2011), utilizam informações de consumo diárias, advindas de medidores inteligentes. Entretanto, essas informações podem afetar a privacidade do cliente, de modo que a utilização de consumos mensais, que é uma informação comum às empresas e consumidores, torna a aplicação dos métodos mais prática.

A normalização de dados é bastante comum, sendo um dos principais métodos utilizados o de normalização sigmoidal, descrito por (31)

$$x_0 = \frac{1}{1 + e^{-\frac{x-\bar{x}}{\sigma_x}}}. \tag{31}$$

Em que x é o valor original, $\bar{x}$ e σ_x representam a média e o desvio padrão do conjunto de dados, respectivamente, e x_0 é o valor normalizado. O método garante que todos os dados estarão presentes no intervalo [0,1].

Outro procedimento, o de janelamento de dados, por sua vez, permite que os dados sejam analisados sob um mesmo intervalo de tempo. A quantidade de meses a ser utilizada, porém, não está bem definida na bibliografia. Glauner *et al.* (2016) defendem

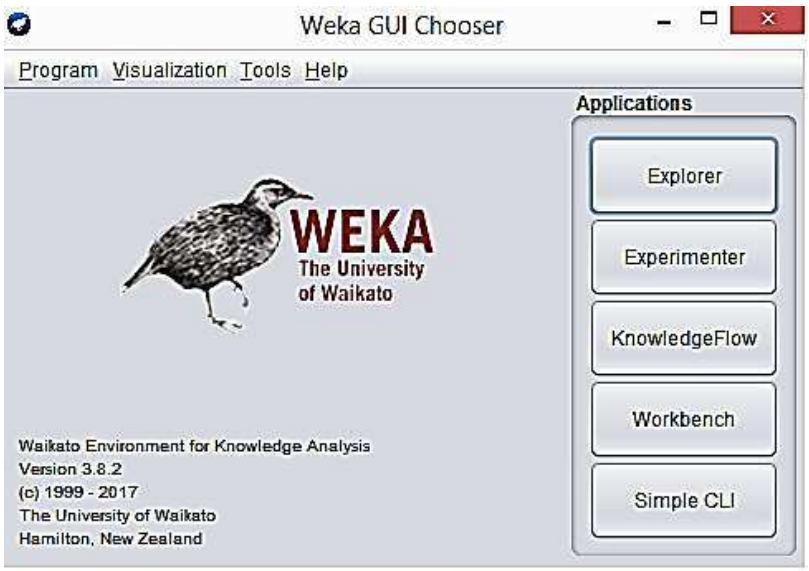
em seu trabalho que a utilização de uma quantidade maior do que 12 meses pode causar *overfitting*, enquanto autores como Nagi *et al.* (2011) informam que um número maior de meses é necessário para observação de padrões de consumo.

Organizado o banco de dados, a detecção de PNT em sistemas de distribuição de energia elétrica pode ser realizada a partir de diversas técnicas, como exposto previamente, porém algumas ferramentas tem surgido para facilitar tal aplicação, como é o caso do *software* WEKA.

2.4 WEKA

O *Waikaito Environment for Knowledge Analysis* (WEKA) é uma plataforma que compreende algoritmos de aprendizado de máquina e ferramentas de pré-processamento de dados. O *software* abrange o processo de mineração de dados desde a preparação dos dados de entrada, avaliação de esquemas de aprendizado, visualização dos dados de entrada e resultados. Na Figura 9, é apresentada tela de inicialização do WEKA.

Figura 9 – Janela inicial do WEKA.



Fonte: Witten *et al.* (2016).

O WEKA foi desenvolvido pela Universidade de Waikaito, na Nova Zelândia, sendo possível, a partir do *software,* a implementação de algoritmos de Aprendizado de Máquina em bancos de dados, incluindo uma grande variedade de ferramentas para transformação de dados, tais como algoritmos de discretização e amostragem. Inclui

também métodos relacionados aos principais problemas de mineração de dados, como regressão e classificação. Nesse aspecto, algoritmo baseado em SVM mais utilizado pertence à biblioteca LibSVM, criada por Chang e Lin (2013).

A partir da ferramenta também é possível obter as métricas usadas para avaliar a eficiência dos modelos, como a matriz de confusão, além da assertividade, acurácia, precisão, sensibilidade, *Receiving Operating Curve* (ROC), entre outras.

3 REVISÃO BIBLIOGRÁFICA

Nesta seção são comentados alguns dos principais estudos referentes à detecção de PNT com foco em Máquinas de Vetores de Suporte.

Nizar (2009) compara, em seu trabalho, os algoritmos de Máquina de Vetores de Suporte e Máquina de Extremo Aprendizado, observando resultados bastante similares entre os métodos, porém com superioridade da Máquina de Extremo Aprendizado.

Em seu trabalho, Xu *et al.* (2010) propuseram um método de cálculo baseado em Máquinas de Vetores de Suporte para perdas em linhas de transmissão, substituindo métodos tradicionais utilizando Redes Neurais Artificiais. Os autores mostraram que o emprego do SVM proporcionou maior acurácia com menor esforço computacional, quando comparado ao uso de Redes Neurais Artificiais.

Nagi *et al.* (2010) utilizaram Máquinas de Vetores de Suporte em sua pesquisa para melhorar a detecção de PNT em sistemas de distribuição da empresa malaia Tenaga Nasional Berhad, utilizando dados de consumo e outros atributos, como tipo de medidor e informações de fraude prévia. A partir do estudo, os autores sugerem ser possível aumentar a taxa de acerto da empresa na detecção de irregularidades de 3% para 60%.

Em um novo estudo, Nagi *et al.* (2011) utilizaram um sistema de inferência difusa e conhecimentos heurísticos para aprimorar o método de detecção de PNT proposto em seu último trabalho, de modo a indicar possíveis suspeitos de atividades fraudulentas. A partir das melhorias, a taxa de acerto subiu de 60% para 72%.

Depuru *et al.* (2011) simulam diversos padrões de consumo, baseados em critérios como localização geográfica, estação do ano, tipo de consumidor (rural ou comercial, grande, pequeno, etc.) e dados históricos. São utilizadas informações de medidores inteligentes, com dados coletados a cada 15 minutos. Os padrões de consumo simulados foram então utilizados para treinar um classificador baseado em Máquinas de Vetores de Suporte, que foi aplicado a dados de consumo reais. Os resultados encontrados foram considerados satisfatórios.

Glauner *et al.* (2016) investigaram em seu trabalho o comportamento de três classificadores para detecção de PNT atuando sobre bancos de dados com diferentes proporções entre consumidores regulares e irregulares. Foram analisados grandes bancos de dados (mais de 100 mil consumidores, com quatro anos de registro) a partir das seguintes técnicas: regras *booleanas*, lógica *fuzzy* e Máquinas de Vetores de Suporte. A

métrica utilizada para comparação foi a AUC, sendo que o classificador baseado em SVM apresentou os melhores resultados.

Meira (2017) também se utilizou de três classificadores em sua análise: *random forest,* Regressão Logística e Máquinas de Vetores de Suporte, usando como métrica de desempenho a AUC. O autor seleciona diversos atributos caracterizados pelos princípios de localidade, similaridade e infraestrutura e utiliza-os como entrada para os algoritmos supracitados. Concluiu-se que dados derivados do consumo são suficientes para proporcionar classificação satisfatória de PNT com relação a resultados obtidos de parâmetros dependentes das distribuidoras.

Nota-se, assim, uma tendência à utilização de Aprendizado de Máquina e Mineração de Dados para detecção de PNT em sistemas de distribuição de energia elétrica. Entretanto, alguns aspectos do problema não foram devidamente tratados. Embora a utilização de SVM tenha sido realizada com sucesso em alguns trabalhos, como o de Nagi *et al.* (2010), que alcançou acurácia de 72% com utilização de algoritmos evolutivos para determinação de parâmetros, a determinação da melhor estratégia de separação de bases, o estudo da melhor quantidade de meses a ser utilizada, para o caso de utilização de consumos mensais, e uma análise acerca do desbalanceamento do banco de dados não são devidamente tratados – ou o são, de forma separada. Apenas Glauner *et al.* (2016) trata dos estudos supracitados, com exceção da estratégia de separação de bases, porém a definição do melhor tipo de entrada não é discutida apropriadamente. Assim, esse trabalho supre essa lacuna na bibliografia, envolvendo todos os estudos supracitados, de modo a realizar-se uma ampla análise relativa à utilização de SVM na detecção de PNT em sistemas de distribuição de energia elétrica.

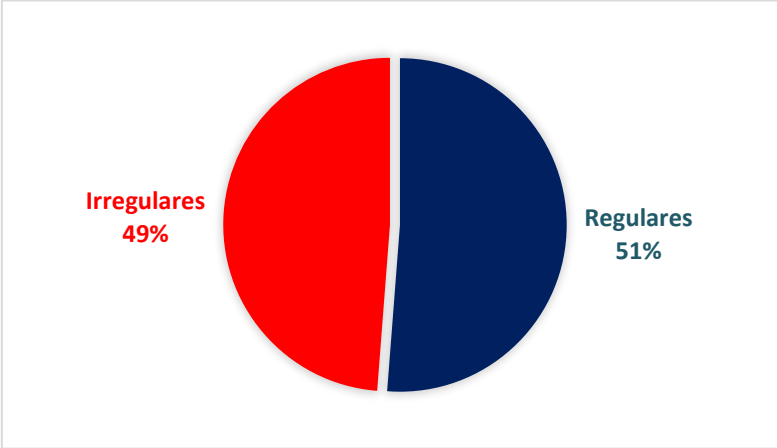
4 MATERIAL E MÉTODOS

Nesta seção são descritos o material e os métodos utilizados durante a elaboração do presente trabalho, de modo a garantir clareza e permitir a reprodutibilidade dos procedimentos adotados. Foram realizados estudos referentes às estratégias de separação de bases, à quantidade de meses a ser utilizada na análise com SVM, ao desbalanceamento do banco de dados e ao tipo de entrada mais relevante para detecção de PNT. A análise foi iniciada com a preparação do banco de dados.

4.1 BANCO DE DADOS UTILIZADO

O banco de dados utilizado neste trabalho possui dados de consumo em kWh/mês de 9177 UC, obtidos entre os anos de 2014 a 2017 e fornecidos por uma distribuidora de energia elétrica do estado da Paraíba. As UC em foco foram sujeitas a inspeções entre os anos de 2016 e 2017, sendo constatadas 4476 UC regulares e 4701 irregulares. A divisão entre as classes pode ser observada na Figura 10**:**

Figura 10 – Distribuição de UC em regulares e irregulares do banco de dados utilizado.



Fonte: próprio autor.

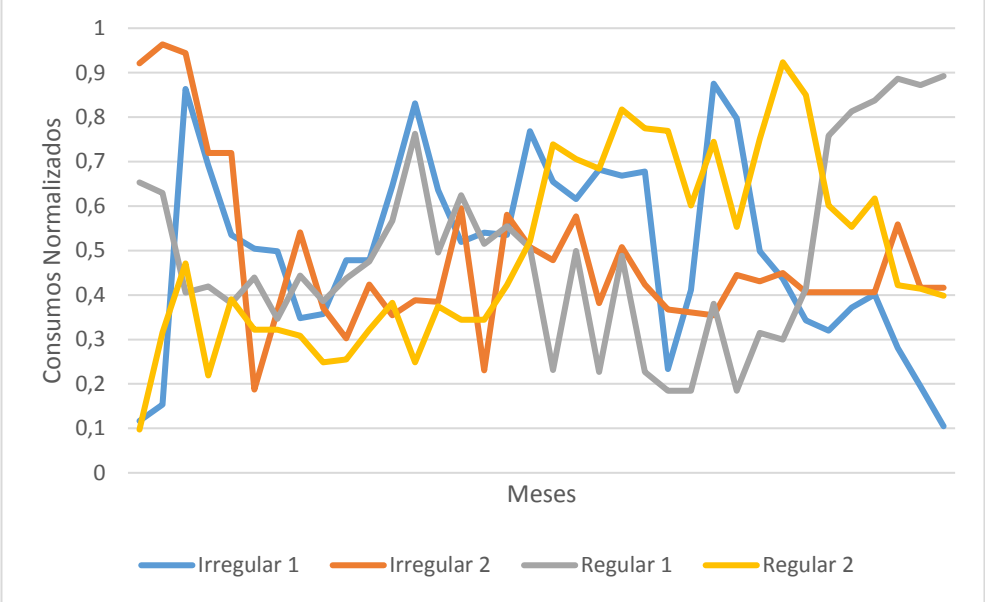
De posse do banco de dados, realizou-se a preparação dos dados para utilização posterior na determinação da melhor estratégia de separação de bases, na definição do período de tempo mais relevante e no melhor desbalanceamento, além da determinação do melhor tipo de entrada.

4.2 PREPARAÇÃO DE DADOS

Os dados foram pré-processados, sendo realizada a normalização e o janelamento dos dados de consumo. A normalização foi realizada utilizando o método sigmoidal, visando a uniformização dos dados, impedindo que a discrepância entre as ordens de grandeza de diferentes parâmetros dificulte o processo de classificação.

Para o janelamento, foram considerados os 36 meses anteriores à inspeção para todas as unidades consumidoras, abrangendo três anos de consumo de energia elétrica. Na Figura 11 são ilustrados comportamentos de consumo de UC regulares e irregulares.

Figura 11 – Exemplos de consumidores regulares e irregulares.



Fonte: próprio autor.

É possível observar a dificuldade de identificação de um padrão de consumo, tendo em vista que ele varia de UC para UC, sendo a utilização de Mineração de Dados e Aprendizado de Máquina fundamental. Após o tratamento de dados, foram realizados estudos referentes à definição da melhor estratégia de separação de bases, da melhor quantidade de meses a ser adotada e do desbalanceamento do banco de dados. A comparação entre modelos oriundos de cada entrada foi realizada a partir da função de desempenho f_d apresentada a seguir.

4.3 FUNÇÃO DE DESEMPENHO

A análise dos resultados foi realizada mediante avaliação das matrizes de confusão fornecidas pelo *software* WEKA, sendo os modelos comparados pela função de desempenho f_d apresentada em (32), cujos pesos foram definidos empiricamente.

$$f_d = 0{,}6 * AUC + 0{,}4 * Acc. \quad (32)$$

A AUC é o principal parâmetro de comparação utilizado na função de desempenho estipulada, tendo em vista que nela são correlacionados os principais resultados das inspeções em campo (inspeção em cliente irregular e inspeção em cliente regular), o que a torna essencial para o problema de detecção de PNT e, desse modo, recebeu maior peso. A acurácia, por sua vez, é essencial para determinação do desempenho de um modelo de Aprendizado de Máquina, tendo em vista que representa o número total de acertos. Entretanto, tendo em vista que nãos e refere apenas à classe de interesse, esta métrica recebeu menor peso.

4.4 ESTUDO DA MELHOR SEPARAÇÃO DE BASES

Para determinar a melhor separação de bases a ser utilizada, as opções de Validação Cruzada com 3, 4, 5, 6, 7, 8, 9 e 10 *folders* e de *Holdout* (66% treinamento e 33% teste) foram aplicadas ao banco de dados a partir do *software* WEKA. Após determinada a melhor forma de separação, seguiu-se o estudo relativo à quantidade de meses mais relevante.

4.5 ESTUDO DO NÚMERO DE MESES A SER ADOTADO

Para otimização do método, foi realizado um estudo a fim de definir a melhor quantidade de meses a ser utilizada para aplicação de Máquina de Vetores de Suporte. Foram realizados testes com 12, 18, 24, 30 e 36 meses a fim de definir o período de tempo mais relevante para detecção de PNT em sistemas de distribuição de energia elétrica, que possibilite representar os perfis de consumos irregulares de forma ótima, sem provocar *overfitting* e utilizar uma quantidade desnecessária de dados.

4.6 ESTUDO DO DESBALANCEAMENTO DO BANCO DE DADOS

Após realizada a determinação da melhor separação de bases e de quantidade de meses, o banco de dados foi modificado, de modo a se obter diferentes proporções de PNT, listando-se 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% e 90%. O processo foi realizado a partir de amostragem do banco original, visando a determinação da proporção de PNT que gerassem detecção ótima, sem gerar *overfitting* do modelo. Foram observados e comparados os resultados para cada proporção. O último estudo referiu-se à definição do tipo de entrada mais relevante à aplicação de SVM.

4.7 ESTUDO DAS MELHORES ENTRADAS

Após o pré-processamento e definição da melhor estratégia de separação de bases, do número de meses mais relevantes ao estudo de PNT e da proporção adequada de PNT para treinamento, os dados foram transformados, de modo a obter-se informações novas. Foram obtidas as variações entre os consumos mensais e as variações relativas às variações entre os consumos mensais. Também foram obtidos dados estatísticos relativos a cada tipo entrada, dos quais listam-se:

- Média;
- Primeiro quartil;
- Mediana;
- Terceiro quartil;
- Variância;
- Desvio Padrão;
- Curtose;
- Assimetria.
- Mínimo;
- Máximo;
- Amplitude.

A configuração de entradas pode ser resumida pela Tabela 2. Para os consumos puros, foram obtidas 36 entradas, uma para cada mês. As variações entre consumos mensais foram obtidas a partir da subtração simples entre o consumo relativo ao mês $n + 1$ e o consumo n, com n variando entre 1 e 35. Um processo idêntico foi utilizado para determinação das variações entre as variações de consumo mensais. Por fim, cada tipo de entrada resultou em 11 parâmetros estatísticos, que por sua vez também foram utilizados como entrada no *software* WEKA.

Tabela 2 – Atributos utilizados para treinamento.

Caso	Descrição	Número de Entradas
1	Consumos	36
2	Variações entre consumos	35
3	Variações entre variações	34
4	Estatísticas de Consumos	11
5	Estatísticas das Derivadas	11
6	Estatísticas das Variações entre Variações	11

Fonte: próprio autor.

Os resultados obtidos a partir da utilização do material e procedimentos apresentados nesta seção são expostos e discutidos a seguir.

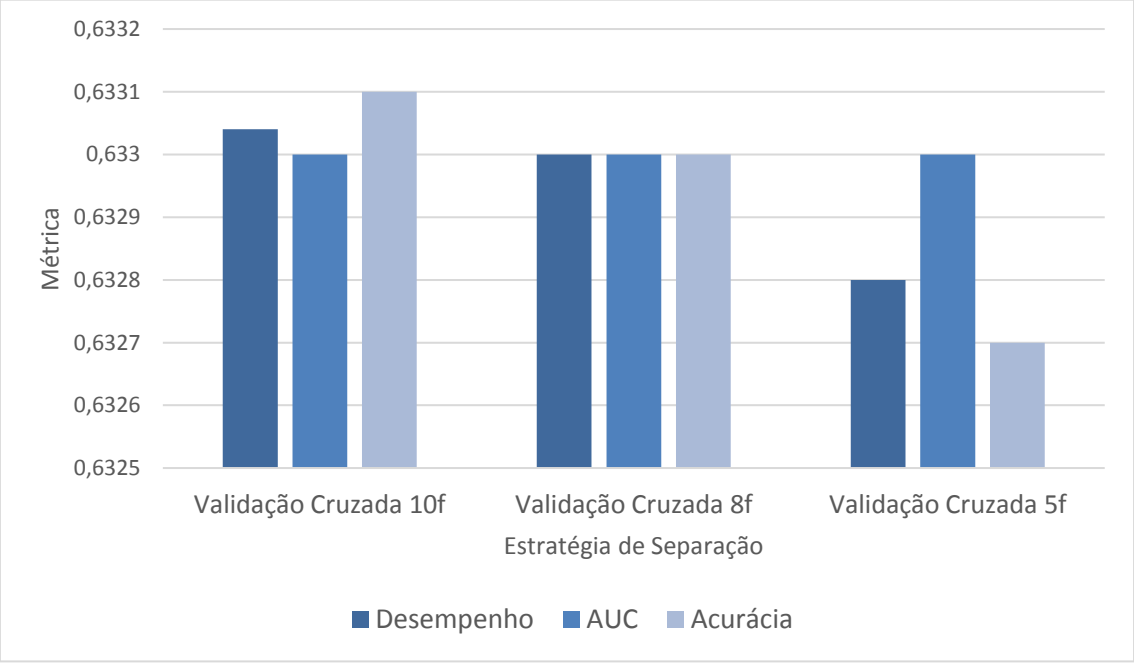
5 RESULTADOS E DISCUSSÕES

Nesta seção, os resultados obtidos a partir da metodologia adotada são apresentados e analisados. Inicialmente, são tratados os resultados dos estudos relativos à determinação da separação de bases. A seguir, as análises sobre o número de meses a serem utilizados e do desbalanceamento do banco de dados são expostas. Por fim, os resultados referentes ao estudo das entradas são discutidos.

5.1 ESTUDO DA MELHOR SEPARAÇÃO DE BASES

A partir dos testes relacionados à separação de bases realizados no *software* WEKA, definiu-se que a melhor forma de separação entre os conjuntos de treinamento e testes é o sistema de Validação Cruzada com 10 *folders.* As métricas relativas aos três melhores resultados obtidos podem ser visualizadas na Figura 12. Os outros resultados foram omitidos por não acrescentarem novas informações, de modo que os três resultados expostos são suficientes para determinação da melhor estratégia de separação de bases a ser utilizada.

Figura 12 – Melhores resultados para estudo de separação de bases.



Fonte: próprio autor.

É possível observar que os três melhores resultados obtidos são relativos às validações cruzadas com 10, 8 e 5 *folders*, respectivamente. Os resultados são bastante similares, com AUC de aproximadamente 0,63 e acurácia superior a 60%, porém com superioridade da validação cruzada com 10 *folders*, que foi a escolhida para realização dos descritos adiante.

A proximidade relativa às métricas usadas para comparação e falta de padrão referente à quantidade de *folders* usados implica na possibilidade de que a organização dos dados pode afetar o modelo a partir da divisão de conjuntos. Vale salientar os piores resultados foram obtidos a partir da utilização do método de *Holdout.* Tal fator comprova o que foi constatado por Queiroga (2005), que observou a inaplicabilidade do método *Holdout* para pequenos bancos de dados, tendo em vista a ocorrência de *overfitting.*

5.2 ESTUDO DO NÚMERO DE MESES A SER ADOTADO

Os resultados obtidos a partir do estudo da quantidade de meses usados para obtenção do modelo estão resumidos na Tabela 2.

Tabela 3 – Resultados obtidos para estudo do melhor período a ser utilizado.

Quantidade de meses	Acurácia	AUC	VP	FP	VN	FN	fd
12	0,6151	0,615	3315	2285	2191	1161	0,61504
18	0,6231	0,623	3261	2159	2317	1215	0,62304
24	0,6247	0,625	3245	2129	2347	1231	0,62488
30	0,6325	0,632	3316	2130	2346	1160	0,6322
36	**0,6331**	**0,633**	**3430**	**2238**	**2238**	**1046**	**0,63304**

Fonte: próprio autor.

Observa-se melhor desempenho com utilização de 36 meses de consumo, tendo em vista que a função de desempenho f_d apresentou melhores resultados. Para esse período, observa-se acurácia de 63,31%, com respectiva AUC de 0,633.

É interessante dizer que esse período de tempo também corresponde ao prazo máximo para que a distribuidora possa realizar a cobrança ao cliente, dada a identificação de irregularidade (ANEEL, 2010). Assim, utilizando-se o método baseado em SVM proposto neste trabalho, é possível identificar possíveis fraudadores no limite de cobrança pelas irregularidades realizadas.

Determinado o melhor período de tempo para detecção de PNT, realizou-se o estudo relativo ao desbalanceamento do banco de dados.

5.3 ESTUDO DO DESBALANCEAMENTO DO BANCO DE DADOS

Quanto ao desbalanceamento de dados, observou-se que para proporções de PNT inferiores ou iguais à 30% e superiores ou iguais à 70% ocorre *overfitting* nos modelos obtidos. Os dados são, assim, classificados de forma indistinta como regulares para baixas proporções de PNT e, igualmente, como irregulares para proporções elevadas. As matrizes de confusão obtidas para os modelos oriundos dos bancos de dados com proporções de 10% e 90% de PNT podem ser observadas nas Figuras 13 (a) e (b), respectivamente.

Figura 13 – Matrizes de confusão para modelos com *overfitting*.

		Referência	
		Irregulares	Regulares
Predição	Irregulares	0	0
	Regulares	4476	497

(a) Matriz de confusão para 10% de PNT

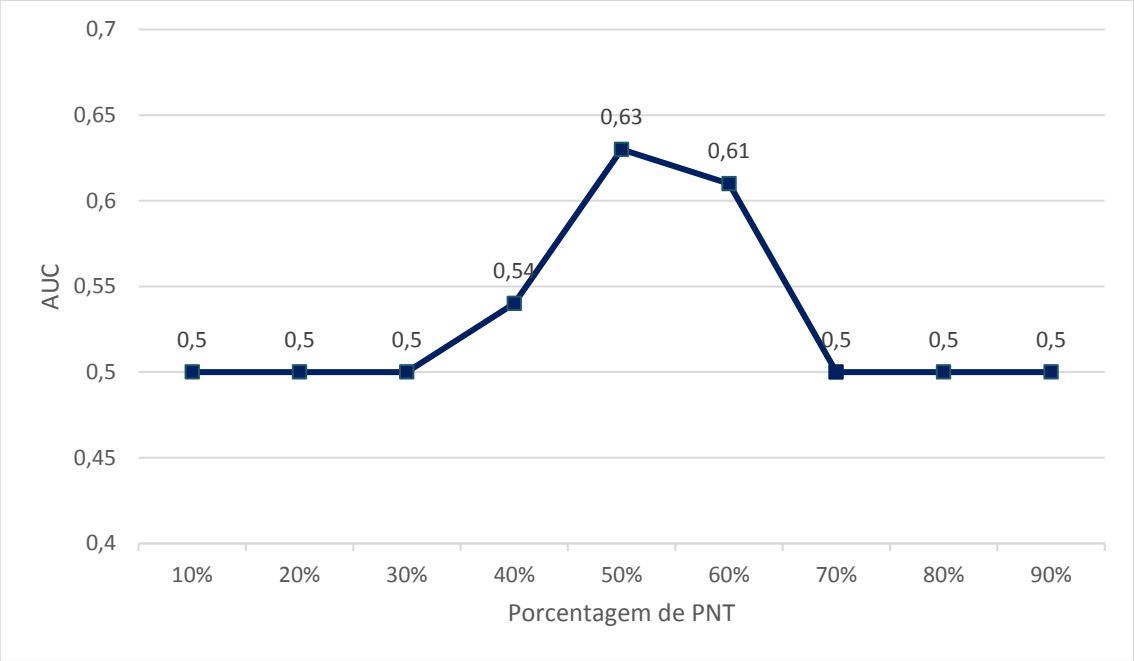
		Referência	
		Irregulares	Regulares
Predição	Irregulares	4476	497
	Regulares	0	0

(b) Matriz de confusão para 90% de PNT

Para os casos observados na Figura 12, embora a classificação tenha ocorrido de forma indistinta, ainda proporcionou acurácia de 90% para ambos os casos, tendo em vista que apenas os 10% relativos à classe menos significativa foram erroneamente classificados.

Dito isto, tem-se que a utilização da função de desempenho f_d apresentada na Equação (32) não pode ser utilizada de forma eficaz, pois proporcionaria erros de interpretação. Assim, para o presente estudo, apenas a AUC foi considerada. Os resultados obtidos encontram-se na Figura 14.

Figura 14 – *AUC* para os modelos oriundos das diversas proporções de PNT testadas.



Fonte: próprio autor.

Como pode ser constatado a partir da observação da Figura 14, a AUC é igual a 0,5 para proporções de PNT menores ou iguais a 30% e maiores ou iguais a 70%, de modo que os modelos apresentam o mesmo desempenho de uma escolha aleatória, ao acaso. Para proporções maiores que 30%, a AUC cresce até 0,63, em 50%, que representou o melhor resultado.

Os resultados obtidos são semelhantes aos obtidos por Glauner *et al.* (2016), que, no entanto, obteve os melhores resultados para um banco de dados com proporção de PNT de 60%.

A seguir são expostos e discutidos os resultados relativos ao estudo dos tipos de entrada mais relevantes.

5.4 Estudo das Entradas

Após definição da melhor estratégia de separação de bases, período de tempo a ser utilizado e desbalanceamento, foram analisados os tipos de entrada. Os estudos de caso são apresentados a seguir:

Os resultados obtidos estão resumidos na Tabela 4:

Tabela 4 – Resultados dos estudos relativos ao melhor período de tempo a ser utilizado.

Tipos de Modelo	**Acurácia**	**AUC**	**Precisão**	**Recall**	**VP**	**FP**	**VN**	**FN**	f_d
Caso 1	0,6332	0,633	0,695	0,766	3430	2238	2238	1046	0,63308
Caso 2	0,5856	0,586	0,559	0,706	3160	2394	2082	1316	0,58584
Caso 3	0,5326	0,533	0,523	0,742	3319	3027	1449	1157	0,53284
Caso 4	0,6157	0,616	0,626	0,576	2576	1540	2936	1900	0,61588
Caso 5	0,5987	0,599	0,613	0,536	2400	1516	2960	2076	0,59888
Caso 6	0,5960	0,590	0,618	0,504	2255	1396	3080	2221	0,5924

Fonte: próprio autor.

Para o Caso 1, foram classificados 5668 consumidores como irregulares (63,31% da população), dos quais 3430 corretamente. O número de falsos negativos, ou seja, de consumidores irregulares não identificados foi de 1046. Desse modo, tem-se uma precisão (taxa de acerto em inspeções) de 69,5%, com sensibilidade (*recall*) de 76,6%. A acurácia do modelo foi de 63,33% e a AUC foi de 0,633, resultando em um desempenho de 0,63308.

Para o Caso 2, relativo às variações entre os consumos mensais, 5564 consumidores foram classificados como irregulares (62,04% da população), dos quais 3160 corretamente. O número de consumidores irregulares não identificados foi de 1316, o que implica em sensibilidade do modelo de 70,6% e precisão de 55,90%. A acurácia do modelo foi de 58,56% e a AUC foi de 0,586, resultando em um desempenho de 0,58584.

Para as variações entre as variações de consumos mensais, relativas ao Caso 3, a classificação de UC irregulares alcançou 52,3% de precisão e 74,2% de sensibilidade, de modo que a partir desse modelo é possível abranger de forma mais completa os indivíduos irregulares, porém a se deu às custas da classificação de 6346 UC (70,9% do total) como fraudadoras. A acurácia do modelo foi de 53, 26%.

Para o Caso 4, que se refere aos dados estatísticos oriundos dos consumos mensais, tem-se precisão de 62,6% e sensibilidade (*recall*) de 56,7%. Foram identificados 2576 consumidores irregulares de forma correta, dos 4476 presentes no banco de dados. Entretanto, 1900 UC irregulares não foram detectadas. A acurácia do modelo foi de 61,57% e AUC de 0,616, com desempenho de 0,61588.

Para os dados estatísticos relativos às variações de consumo mensal (Caso 5), obteve-se precisão de 61,3% e sensibilidade (*recall*) de 53,6%. Para o modelo em questão, foram classificadas 3916 UC como irregulares (43,74% da população), das quais 2400 corretamente. A acurácia foi de 63,32% e a AUC de 0,599, com desempenho de 0,59888.

Por fim, tem-se, para o Caso 6, referente aos dados estatísticos oriundos das variações entre as variações de consumo mensal, obteve-se precisão de 61,8% e sensibilidade (*recall*) de 50,4%. Foram classificadas 3651 UC como irregulares (40,71% da população), das quais 2255 corretamente, porém 2221 UC irregulares não foram detectadas pelo modelo. A acurácia foi de 59,60% e AUC de 0,590, com desempenho de 0,5924.

É possível notar que os modelos construídos a partir das estatísticas dos dados apresentam, em geral, um índice relativamente pequeno de detecção de PNT, com valores de sensibilidade menores do que 60%, porém a classificação se torna, em geral, mais precisa, de forma que os casos estatísticos, com exceção dos consumos mensais, apresentaram resultados melhores do que relativamente à aplicação dos dados dos quais foram originados.

Os Casos 1, 2 e 3, embora apresentem alta taxa de *recall*, possuem uma precisão mais baixa, de modo que os modelos são sensíveis à detecção de PNT, porém a classificação carece de exatidão.

Os resultados encontrados estão de acordo com aqueles observados na bibliografia, a exemplo de Nagi *et al.* (2010), que obteve acurácia de 60%. A partir desses resultados, constata-se que é possível aumentar a taxa de acertos das inspeções para aproximadamente 70%, com até 76,6% de clientes irregulares identificados dentre todos os irregulares. Isso é possível a partir da utilização de validação cruzada com 10 *folders*, com utilização de um período 36 meses de dados de consumo e com banco de dados balanceado (50% de UC irregulares e 50% de UC regulares).

6 CONCLUSÕES

Neste trabalho foi apresentada uma análise da aplicação de Máquinas de Vetores de Suporte para detecção de perdas não técnicas em sistemas de distribuição. A análise foi fundamentada em aspectos como a determinação da melhor estratégia de separação de bases, avaliação do melhor período de tempo a ser utilizado e estudo da influência do desbalanceamento do banco de dados. Ademais, foram analisados diferentes tipos de entrada ao método.

Os resultados corroboram com a bibliografia consultada, relativamente à determinação de estratégias de separação de base, com utilização de validação cruzada com 10 *folders* proporcionando resultados superiores em relação às demais. Entretanto, a análise referente ao período ótimo a ser utilizado revela a necessidade de estudos adicionais, tendo em vista a divergência apresentada para com outros trabalhos.

A proporção de PNT a ser utilizada para treinamento de modelos visando a detecção de PNT também é um ponto a ser discutido, pois a desproporcionalidade entre as classes de consumidor (regulares e irregulares) se apresenta de forma natural, com um número muito maior de consumidores regulares, porém este estudo demonstra que a construção de bancos de dados com proporção controlada pode ser útil para a tarefa de detecção proposta.

Observou-se, ao fim, que o modelo alcançado referente ao Caso 1, dos estudos do tipo de entrada, que engloba todos os estudos anteriores, atende de forma satisfatória as necessidades das distribuidoras de energia elétrica, podendo ser utilizado para determinação de consumidores irregulares com taxas de sucesso na inspeção até 2,5 vezes superiores ao método utilizado atualmente. A aplicação de Máquinas de Vetores de Suporte se apresenta, assim, como um método eficaz para detecção de PNT em sistemas de distribuição, com acurácia elevada, de até 63,32%, podendo ser utilizado pelas concessionárias para aumentar seus índices de sucesso em inspeções.

6.1 TRABALHOS FUTUROS

Alguns aspectos que podem ser investigados a partir dos estudos apresentados neste trabalho são:

- Aplicação de métodos para determinação dos melhores parâmetros de refinamento do modelo, como o custo *C*;
- Investigação de combinações de entradas, bem como utilização de outros tipos de entrada, como coeficientes de *Fourier* e *Wavelets*;
- Associação do método de Máquinas de Vetores de Suporte a outros, como algoritmos evolutivos;
- Utilização de estatística multivariada para determinar os parâmetros com maior influência na identificação de consumidores irregulares.

REFERÊNCIAS

ABRADEE. Residencial. **Associação Brasileira de Distribuidores de Energia Elétrica**, 2015. Disponivel em: <http://www.abradee.com.br/financeiro/mapas-aliquotas-icms/residencial>.

ABRADEE. Furto e Fraude de Energia. **Associação Brasileira de Distribuidores de Energia Elétrica**, 2017. Disponivel em: <http://www.abradee.com.br/setor-de-distribuicao/perdas/furto-e-fraude-de-energia>.

AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA. **Perdas de Energia** - Distribuição, 2018. Disponível em < 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> Acesso em 20 de Novembro de 2018.

ARAUJO, B. S. **Métodos de Inteligência Computacional para Detecção de Fraudes de Energia Elétrica**. Universidade Federal do Rio de Janeiro, Rio de Janeiro – RJ, 2017. 57 p. (Monografia).

BOTEV, V; ALMGREN, M; GULISANO, V; LANDSIEDEL, O; PAPATRIANTAFILOU, M; VAN ROOIJ, J. ***Detecting Non-Technical Energy Losses through Structural Periodic Patterns in AMI data***. In: IEEE International Conference on Big Data (Big Data), 2016.

CHAUHAN, A.; RAJVANSHI, S. ***Non-Technical Losses in Power System: A Review***. *In:* International Conference of Power, Energy and Control, 2013.

DEPURU, S. S. S. R; WANG, L; DEVABHAKTUNI, V. Support Vector Machine Based Data Classification for Detection of Electricity Theft. IEEE, 2011.

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

FOIATTO, N. **Sistematização do reconhecimento de irregularidades que caracterizam fraude em medidores de energia elétrica**. Universidade Federal do Rio Grande do Sul, Porto Alegre - RS, 2009.

GLAUNER, P.; BOECHAT, A.; DOLBERG, L.; STATE, R.; BETTINGER, F.; RANGONI, Y.; DUARTE, D. ***Large-Scale Detection of Non-Technical Losses In Imbalanced Data Sets***, 2016.

GLAUNER, P.; MEIRA, A. J.; DOLBERG, L.; STATE, R.; BETTINGER, F.; RANGONI, Y. ***Neighborhood features help detecting non-technical losses in big data sets.*** In: 3rd International Conference on Big Data Computing, Applications and Technologies, 2016.

GOSH, S.; REILLY, D. L. ***Credit card fraud detection with a neural-network.*** In: Proceedings of the Twenty-Seventh Annual Hawaii International Conference on System Sciences, 1994.

MONEDERO, I; BISCARRI, F; LEÓN, C. GUERRERO, J.I; BISCARRI, J.MILLÁN, R. ***Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees.*** In: Electrical Power and Energy Systems. 34ª edition. p. 90-98, 2012

MEIRA, J. A.; GLAUNER, P.; STATE, R.; VALTCHEV, P.; DOLBERG, L.; BETTINGER, F.; DUARTE, D. **Distilling Provider-Independent Data for General Detection of Non-Technical Losses,** IEEE, 2017.

NAGI, J.; YAP, K. S.; TIONG, S. K.; AHMED, S. K.; MOHAMAD, M. **Non-technical Loss Detection for Metered Customers in Power Utility Using Support Vector Machines.** IEEE Transactions on Power Delivery, vol. 25, 2010.

NAGI, J.; YAP, K. S.; TIONG, S. K.; AHMED, S. K.; NAGI, F. ***Improving SVM-Based Nontechnical Loss Detection in Power Utility Using the Fuzzy Inference System.*** IEEE Transactions on Power Delivery, vol. 26, 2011.

NIZAR, A. H.; DONG, Z. Y. ***Identification and Detection of Electricity Customer Behaviour Irregularities***. IEEE, 2009.

QUEIROGA, R. M. **Uso de Técnicas de Data Mining para Detecção de Fraudes em Energia Elétrica.** Universidade Federal do Espírito Santo, Vitória - ES, 2005 (dissertação).

RAMOS, C. C. O.; RODRIGUES, D.; SOUZA, A. N. ; PAPA, J. P. ***On the Study of Commercial Losses in Brazil: A Binary Black Hole Algorithm for Theft Characterization.*** In: IEEE Transactions on Smart Grid, 2016.

RAMOS, C. C. O. **Caracterização de Perdas Comerciais em Sistemas de Energia Através de Técnicas Inteligentes**, Universidade de São Paulo, São Paulo - SP, 2014.

RODRIGUES, D; RAMOS, C. C. O; PAPA, J. P. ***Black Hole Algorithm for Non-technical Losses Characterization***. IEEE, 2015.

ROJAS, G. A. Q.; GALLEGO, R. A. ***Advanced Analytics for Non-Technical Losses of Energy***. *In:* IEEE Innovative Smart Grid Technologies Latin America, 2015.

SAISSE, R. W. **Detecção de Perdas Não Técnicas em Redes de Distribuição Radiais Utilizando Estimação de Estado**. Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2016. 103 p.

SCHOLKOPF, B; SMOLA, A. J. ***Learning with kernels***. MIT Press, 2002.

SHALEV-SHWARTZ, S.; BEN-DAVID, S. ***Understanding Machine Learning: from theory to algorithms***. Cambridge University Press, 2014.

SMOLA, A.; VISHWANATHAN. ***Introduction to Machine Learning***. Cambridge University Press, 2008.

TREVIZAN, R. D.; MARTIN, R. P.; BRETAS, N. G.; BETTIOL, A. L. ***Non-technical Losses Identification Using Optimum Path Forest and State Estimation***. *In:* PowerTech, At Eindhoven, Holanda, 2015.

VAPNIK, V. N. ***An overview of statistical learning theory***. In: IEEE Transactions on Neural Networks, vol. 10, p. 988-999, 1999.

VAPNIK, V. N; LERNER, A. (1963) ***A Pattern Recognition Using Generalized Portrait***. Automation and Remote Control, 24, 6.

VIEGAS, J. L.; ESTEVES, P. R.; MELÍCIO, R.; MENDES, V. M. F.; VIEIRA, S. M. ***Solutions for detection of non-technical losses in the electricity grid: A review***. In: Renewable and Sustainable Energy Reviews, v. 80, p. 1256–1268, 2017.

XU, R. Z.; WANG, Y. F.; LI, Y. K. ***A Novel Calculation Method for the Line Losses Based on Support Vector Machine,*** In: 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE), 2010.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. ***The Weka Workbench.*** 4ª Edição, 2016.