Pontifícia Universidade Católica
do Rio de Janeiro

Norma Alice da Silva Carvalho

Sistema Inteligente Híbrido para Classificação do Perfil de Pagamento dos Consumidores Não-Residenciais de Energia Elétrica

Tese de Doutorado

Tese apresentada como requisito parcial para obtenção do título de Doutor pelo Programa de Pós-Graduação em Engenharia de Produção da PUC-Rio.

Orientador: Prof. Eugenio Kahn Epprecht
Co-Orientador: Prof. Reinaldo Castro Souza

Rio de Janeiro
Dezembro de 2016

Norma Alice da Silva Carvalho

Sistema Inteligente Híbrido para Classificação do Perfil de Pagamento dos Consumidores Não-Residenciais de Energia Elétrica

Tese apresentada como requisito parcial para obtenção do título de Doutor pelo Programa de Pós-Graduação em Engenharia de Produção da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Prof. Eugenio Kahn Epprecht
Orientador
Departamento de Engenharia Industrial – PUC-Rio

Prof. Reinaldo Castro Souza
Co-Orientador
Departamento de Engenharia Elétrica – PUC-Rio

Prof. André Luís Marques Marcato
Departamento de Energia – UFJF

Prof. José Francisco Moreira Pessanha
Departamento de Otimização e Energética e Meio Ambiente – CEPEL

Prof. Nival Nunes de Almeida
Coordenação das Engenharias – PUC-Rio

Prof. Rodrigo Flora Calili
Departamento de Metrologia – PUC-Rio

Prof. Mônica Barros
Escola Nacional de Ciências Estatísticas – ENCE

Prof. Márcio da Silveira Carvalho
Coordenador Setorial do Centro Técnico Científico – PUC-Rio

Rio de Janeiro, 20 de dezembro de 2016

Norma Alice da Silva Carvalho

Graduada em Matemática pela Universidade Federal Fluminense (2009). Mestra em Metrologia pela Pontifícia Universidade Católica do Rio de Janeiro (2011).

Ficha Catalográfica

Carvalho, Norma Alice da Silva

Sistema inteligente híbrido para classificação do perfil de pagamento dos consumidores não-residenciais de energia elétrica/ Norma Alice da Silva Carvalho; orientador: Eugenio Kahn Epprecht; co-orientador: Reinaldo Castro Souza. – 2016.

109 f.: il. color.; 30 cm

Tese (doutorado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Industrial, 2016.

Inclui bibliografia

1. Engenharia Industrial – Teses. 2. Sistemas de distribuição de energia elétrica. 3. Inadimplência; 4. Busca de conhecimento em base de dados. 5. Mapas auto-organizáveis de Kohonen 6. Classificador Bayesiano Simples. I. Epprecht, Eugenio Kahn. II. Souza, Reinaldo Castro. III. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Engenharia Industrial. IV. Título.

CDD: 658.5

Aos que contribuíram para a conclusão desse grande desafio.

Agradecimentos

A Deus, por me conceder os dons da fortaleza e da sabedoria.

Aos familiares e amigos, pelas palavras de perseverança, apoio e incentivo.

Aos amigos e colegas da PUC-Rio, pelos momentos de aprendizagem e superação que juntos vivenciamos.

Ao orientador, Eugenio Kahn Epprecht, pelo apoio, aconselhamentos e acolhimento desse projeto em sua linha de pesquisa.

Ao co-orientador, Reinaldo Castro Souza, por acreditar na realização desse estudo, apoiar-me e contribuir com muitos ensinamentos que levarei para a vida.

Ao Professor José Francisco Moreira Pessanha, pelas valiosas contribuições a esse estudo e exemplo de amor ao ensino.

À Professora Maria Fátima Ludovico de Almeida, pelos valiosos ensinamentos de metodologia da pesquisa e pelas palavras de incentivo.

Aos Professores da PUC-Rio com os quais tive a oportunidade de aprender muitos conceitos.

À Ana Paiva, pelas palavras de perseverança e pelo apoio administrativo.

À equipe técnica e administrativa do Departamento de Engenharia Industrial, pelas informações administrativas e apoio técnico. Em especial, a Cláudia Guimarães Teti, sempre solícita a auxiliar em qualquer situação.

À Coordenação Central de Pós-graduação e à Coordenação Pós-Graduação do Departamento de Engenharia Industrial, por conceder-me isenção da mensalidade nos períodos de prorrogação para conclusão desse estudo.

À Coordenação de Aperfeiçoamento de Pessoal do Ensino Superior (CAPES), pelo apoio financeiro na realização desse estudo.

Resumo

Carvalho, Norma Alice da Silva; Epprecht, Eugenio Kahn; Souza, Reinaldo Castro. **Sistema inteligente híbrido para classificação do perfil de pagamento dos consumidores não-residenciais de energia elétrica.** Rio de Janeiro, 2016. 109p. Tese de Doutorado - Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

O **objetivo** desta pesquisa é classificar o perfil de pagamento dos consumidores não-residenciais de energia elétrica, considerando conhecimento armazenado em base de dados de distribuidoras de energia elétrica. A **motivação** para desenvolvê-la surgiu da necessidade das distribuidoras por um modelo de suporte a formulação de estratégias capazes de reduzir o grau inadimplência. A **metodologia** proposta consiste em um sistema inteligente híbrido composto por módulos intercomunicativos que usam conhecimentos armazenados em base de dados para segmentar consumidores e, então, atingir o objetivo proposto. O sistema inicia-se com o módulo neural, que aloca as unidades consumidoras em grupos conforme similaridades (valor fatura, consumo, demanda medida/demanda contratada, intensidade energética e peso da conta no orçamento), em sequência, o módulo bayesiano, estabelece um escore entre 0 e 1 que permite predizer o perfil de pagamento das unidades considerando os grupos gerados e os atributos categóricos (atividade econômica, estrutura tarifária, mesorregião, natureza jurídica e porte empresarial) que caracterizam essas unidades. Os **resultados** revelaram que o sistema proposto estabelece razoável taxa de acerto na classificação do perfil de consumidores e, portanto, constitui uma importante ferramenta de suporte a formulação de estratégias para combate à inadimplência. **Conclui-se** que, o sistema híbrido proposto apresenta caráter generalista podendo ser adaptado e implementado em outros mercados.

Palavras-chave

Sistemas de distribuição de energia elétrica; inadimplência; busca de conhecimento em base de dados; mapas auto-organizáveis de Kohonen; classificador Bayesiano Simples.

Abstract

Carvalho, Norma Alice da Silva; Epprecht, Eugenio Kahn; Souza, Reinaldo Castro. **Hybrid intelligent system for classification of non-residential electricity customers payment profiles.** Rio de Janeiro, 2016. 109p. DSc. Thesis – Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

The **objective** of this research is to classify the non-residential electricity customer payment profiles regarding the knowledge stored in electricity distribution utilities' databases. The **motivation** for development of the work from the need of electricity distribution by a support model to formulate strategies for tackling non-payment and late payment. The proposed **methodology** consists of a hybrid intelligent system constituted by intercommunicating modules that use knowledge stored in database to customer segmentation and then achieve the proposed objective. The system begins with the neural module, which allocates the consuming units in groups according to similarities (bill amount, consumption, measured demand/contracted demand, energy intensity and share of the electricity bill in the customer's income), in sequence, the Bayesian module establishes a score between 0 and 1 that allows to predict what payment profile of the units considering the generated groups and categorical attributes (business activity, tariff type, business size, mesoregion and company's legal form) that characterize these units. The **results** showed that the proposed system provides a reasonable success rate when classifying customer profiles and thus constitutes an important tool in the formulation of strategies for tackling non-payment and late payment. In **conclusion**, the hybrid system proposed here is a generalist one and could usefully be adapted and implemented in other markets.

Keywords

Electricity distribution systems; non-payment/late payment; knowledge discovery in databases; Kohonen's self-organizing maps; naive Bayes.

Sumário

Lista de figuras

Lista de quadros

Lista de tabelas

Lista de siglas

ANEEL:	Agência Nacional de Energia Elétrica
BMU:	*Best-Matching Unit*
CCEE:	Câmara de Comercialização de Energia Elétrica
CMSE:	Comitê de Monitoramento do Setor Elétrico
CNPE:	Câmara Nacional de Política Energética
CNPJ:	Cadastro Nacional de Pessoas Jurídicas
CRM:	*Customer Relationship Management*
DC:	Demanda Contratada
DM:	Demanda Medida
EPE:	Empresa de Pesquisa Energética
EQ:	Erro de Quantização
ET:	Erro de Topográfico
KDD:	*Knowledge Discovery in Databases*
MME:	Ministério de Minas e Energia
ONS:	Operador Nacional do Sistema
PCHs:	Pequenas Centrais Hidrelétricas
PIB:	Produto Interno Bruto
SC:	*Soft Computing*
SIN:	Sistema Interligado Nacional
SOM:	*Self-Organizing Maps*
SVM:	*Support Vector Machine*
TC:	Tarifa de Consumo
TD:	Tarifa de Demanda
TE:	Tarifa de Energia
TI:	Tecnologia da Informação
TUSD:	Tarifa do Uso do Sistema de Distribuição

"Os números governam o mundo"

Pitágoras

1
Introdução

O sistema de distribuição de energia elétrica é constituído por empresas que conectam a sociedade ao Setor Elétrico, conforme diretrizes instituídas pelo agente regulador deste Setor –no Brasil, a Agência Nacional de Energia Elétrica (ANEEL)– e contrato de concessão ou ato de permissão celebrado com o Poder Concedente para exploração do serviço, por determinado período de tempo, no território geográfico do qual cada uma dessas empresas detém o monopólio do fornecimento de energia elétrica (Agência Nacional de Energia Elétrica, 2008; Campos, 2010)

Atualmente, o Brasil possui 101 empresas distribuidoras de energia elétrica, sendo 63 concessionárias e 38 permissionárias (Agência Nacional de Energia Elétrica, 2016), que operam na redução da tensão da energia, produzida pelo sistema de geração e repassada ao sistema de transmissão, para atender consumidores em níveis de alta tensão (superior a 69 kV e inferior a 230 kV), média tensão (superior a 1 kV e inferior a 69 kV) e baixa tensão (igual ou inferior a 1 kV) conectados ao Setor. Em decorrência da função que exercem, a tarifa de fornecimento de energia elétrica –fonte de receita do Setor– é recolhida, essencialmente, por essas empresas em uma periodicidade mensal e engloba custos gerenciáveis (Parcela A= geração de energia + transmissão de energia + encargos setoriais + tributos) e custos não-gerenciáveis (Parcela B= custos operacionais + custos de capital + depreciação) pelo sistema de distribuição (Agência Nacional de Energia Elétrica, 2008).

Conforme Araújo (2007), Medeiros (2013) e Smith (2004), as empresas distribuidora de energia elétrica sofrem com perdas que impactam na sua gestão financeira. Tais perdas podem ser classificadas em: **técnicas –**provenientes à dissipação de energia nos componentes do sistema elétrico– e **não-técnicas (ou comerciais) –** provenientes à inadimplência e ao uso irregular da energia elétrica, sendo este último causado principalmente pela ação de terceiros (furto de energia e fraude no consumo de eletricidade) ou por equipamentos defeituosos (erros de leitura, falha no faturamento).

No Brasil, o principal mecanismo de recuperação dessas perdas é estabelecido pela ANEEL no processo de revisão tarifária periódica, realizado a cada quatro anos em média. Tal mecanismo consiste na definição de um valor teto para a tarifa de fornecimento de energia elétrica de cada distribuidora, considerando a quantidade de energia necessária para atender ao mercado consumidor e uma estimativa de perdas na distribuição. Os custos regulatórios definidos pela ANEEL podem ser maiores ou menores do que os custos reais praticados pelas empresas distribuidoras, no entanto, todas essas empresas são incentivadas a reduzirem seus custos e se tornarem mais eficientes para que na próxima revisão tarifária periódica esse ganho de produtividade seja repassado ao consumidor em prol da modicidade tarifária (Agência Nacional de Energia Elétrica, 2016). Assim, a cada ciclo tarifário, as distribuidoras de energia elétrica possuem um novo desafio em reduzir ainda mais as perdas ou arcar com a compra desta energia não coberta pela ANEEL.

Conforme Fonseca & Reis (2012, p. 93), "não são poucas as distribuidoras com nível de perdas acima dos limites permitidos e passíveis de serem reconhecidos e admitidos na composição das tarifas". A maioria das distribuidoras de energia elétrica que atuam no Brasil apresentam um prejuízo anual de milhões de reais com as perdas (Agência Nacional de Energia Elétrica, 2016). Por isso, cada vez mais, para minimizar os efeitos das perdas, as distribuidoras apostam na modernização tecnológica das redes e em ações de combate à inadimplência, à fraude[1] e ao furto[2] de energia.

A modernização tecnológica das redes reduz a ocorrência das perdas técnicas e inibe as perdas não-técnicas devido a possibilidade de monitoramento do consumo de energia em tempo real, porém requer investimento de longo prazo. Por outro lado, conforme Instituto Acende Brasil[3] (2007a), ao intensificar o combate à inadimplência, há aumento do furto e da fraude no sistema de distribuição. Desta forma, a gestão da inadimplência se caracteriza como um dos processos mais críticos para o negócio das distribuidoras (Fonseca & Reis, 2012).

[1]Fraude de energia= violação ou adulteração de medidores.

[2]Furto de energia= derivações no circuito elétrico de modo que a carga atendida por esses circuitos não seja medida.

[3]Instituto Acende Brasil= associação privada, criada em 2001, que busca oferecer aos diversos atores da sociedade informações qualificadas sobre os principais agentes econômicos e institucionais que moldam o Setor Elétrico Brasileiro.

1.1
Caracterização do Problema de Pesquisa

A inadimplência caracterizada pelo montante da receita faturada e não recebida é gerada por consumidores de diferentes setores. Em algumas empresas distribuidoras de energia elétrica, grande parte das faturas não pagas são provenientes dos segmentos residencial e comercial, enquanto em outras, são do segmento industrial (Fonseca & Reis, 2012).

Usualmente, a gestão comercial dessas empresas em conformidade com as regras estabelecidas ANEEL (2010) propõe ações e iniciativas para combater a inadimplência, veja Quadro 1. A ação de maior recorrência é a suspensão do fornecimento de energia elétrica. Embora prevista na legislação vigente, trata-se de uma ação complexa de ser realizada em muitas empresas distribuidoras de energia elétrica devido a questões operacionais ou por ser a unidade consumidora inadimplente uma prestadora de serviço essencial (por exemplo: hospital) ou ainda um particular que depende do fornecimento de energia para manter aparelhos médicos essenciais a vida (Campos, 2010; Nilsson, 2012).

Quadro 1- Ações de Combate a Inadimplência em Distribuidoras de Energia Elétrica

Ação	Enfoque
Campanhas de atualização cadastral	Atualizar cadastro de consumidores para facilitar cobrança. Medida complexa que exige atuação integrada da gestão comercial com a área responsável pelos Sistemas de Informação. O fator custo de uma campanha desse tipo somado às suas especificidades operacionais dificultam e desestimulam sua aplicação.
Sorteio de prêmios para pagamento em dia	Incentivar consumidores a serem adimplentes. Uma medida pontual, normalmente aplicada em campanhas específicas. Nessas campanhas também se busca um resultado positivo para a imagem da empresa para a população atendida em sua área de concessão. Depende, naturalmente, de um aporte de gastos com mídias de grande alcance (televisão, rádio, jornais, etc.).
Campanha de eficiência energética e reforma de instalações elétricas internas	Reduzir o valor da conta de energia em troca de lâmpadas e de eletrodoméstico antigos. É uma ação de caráter mais preventivo, aplicada em áreas de difícil acesso, de consumidores com menor nível de renda. Incentivar a reforma de instalações elétricas internas.
Substituição dos medidores analógicos pelos medidores digitais inteligentes (*smart meters)*	Monitorar consumo e condições de fornecimento de energia. Realizar leitura, suspensão e religação de modo remoto.
Teleaviso de inadimplência (7 a 23 dias em atraso)	Lembrar ao consumidor a falta de pagamento e saber data que será saldada a pendência, todavia depende do nível de atualização da base cadastral de gestão de consumidores.
Notificação da suspensão do fornecimento de energia (24 a 70 dias em atraso)	Comunicar ao consumidor, por escrito e com entrega comprovada de antecedência mínima de 15 dias, que o serviço de fornecimento de energia elétrica será suspenso por falta de pagamento.
Suspensão do fornecimento de energia por falta de pagamento (39 a 90 dias em atraso)	Realizar a suspensão do fornecimento de energia após notificação por atraso.

Quadro 1- Ações de Combate a Inadimplência em Distribuidoras de Energia Elétrica (continuação)

Ação	Enfoque
Ações no Serviço de Proteção ao Crédito e na justiça (acima de 90 dias em atraso)	Inscrever CNPJ do devedor no cadastro de inadimplentes dos Serviços de Proteção ao Crédito: uma ação de forte impacto capaz de gerar bons resultados. Usualmente, essa ação é realizada após findos os prazos legais e esgotadas todas as iniciativas. Porém, quando realizada antes do prazo de suspensão previsto, contribui para reduzir o número de unidades consumidoras suspensas.
Campanhas de parcelamento de débitos	Parcelar débitos de significativo valor, principalmente nos segmentos industrial e comercial. Também bastante eficiente quando aplicada em áreas de consumidores de baixa renda e com elevados níveis de inadimplência.
Terceirização da cobrança	Contratar empresas especializadas em recuperação de dívidas para trabalhar determinada carteira de débitos vencidos da distribuidora. É uma medida que pode ser aplicada para recuperar os débitos mais antigos, aqueles em relação aos quais a distribuidora já não teria muito que fazer.

Fonte: Adaptado de Fonseca & Reis (2012)

Souza et al. (2013) mencionam que esses instrumentos não têm sido suficientes e que novas estratégias devem ser desenvolvidas tanto para recuperar o faturamento perdido quanto intervir preventivamente na ocorrência das perdas não-técnicas.

Nos últimos anos, estudos têm sido desenvolvidos propondo sistemas que detectem e minimizem perdas não-técnicas (Araujo, 2007; Bastos, 2011; Cabral et al., 2006; Calili, 2005; Carvalho et al., 2014; Costa et al., 2013; Dias, 2006; Eller, 2003; Faria, 2012; Huang et al., 2013; Jiang et al. 2002; León et al., 2011; Medeiros, 2013; Monedero et al., 2012; Nagi et al., 2008; Nizar et al., 2006; Nizar et al., 2008; Ortega, 2008; Penin, 2008; Petkovic & Balaban, 2008; Queiroga, 2005; Ramos, 2014; Ramos et al., 2011; Reis Filho, 2006; Ribeiro et al., 2012; Souza et al., 2013; Trevisan, 2014). Todavia, há uma lacuna na literatura no que se refere a detecção do risco de inadimplência dos consumidores em empresas distribuidoras de energia elétrica. Sabe-se que, o tema risco de crédito é bem consolidado em estudos que envolvem instituições financeiras, mas em empresas distribuidoras de energia elétrica é incipiente devido a não disponibilidade de informações sobre os rendimentos domiciliares dos seus clientes.

O desenvolvimento de um sistema capaz de prever o comportamento de pagamento dos consumidores é um instrumento valioso na determinação do índice de cobrabilidade em distribuidoras de energia elétrica. Tal índice, definido como razão entre arrecadação do mês corrente e valor faturado anterior, deve ser estabelecido antes do início do mês de referência visando uma melhor

administração do lucro que será obtido pelas empresas distribuidoras (Carvalho, 2011).

Souza et al. (2013) desenvolveram um sistema *fuzzy* para avaliar, antecipadamente, a capacidade de pagamento de consumidores residenciais baixa tensão em uma empresa distribuidora de energia elétrica, considerando a média do histórico anual de atrasos desses consumidores e a participação estimada da conta no seu orçamento.

Assim, busca-se com esta pesquisa responder: como desenvolver um sistema que seja capaz de compreender o fator inadimplência em unidades consumidoras não-residenciais atendidas na média e alta tensão e, então, contribuir para que a gestão da inadimplência sane as insuficiências dos instrumentos vigentes?

1.2 Objetivo

A presente pesquisa tem por objetivo geral propor um sistema que classifica o perfil de pagamento de unidades consumidoras não-residenciais de distribuidora de energia elétrica considerando conhecimento armazenado em base de dados. Em termos específicos, a pesquisa busca:

- Descrever brevemente a cadeia de valor do setor elétrico brasileiro a fim de apontar a inadimplência como um dos principais fatores que ameaçam o equilíbrio econômico-financeiro das empresas distribuidoras de energia elétrica e, consequentemente, do setor;
- Definir e obter fatores (atributos) que influenciam a inadimplência no setor elétrico;
- Apresentar a mineração de dados (*Data mining*) como uma etapa do processo de Busca de Conhecimento em Base de Dados (*Knowledge Discovery in Databases* - KDD), na qual estão disponíveis ferramentas multidisciplinares diversas;
- Identificar na Gestão de Relacionamento com o Cliente (*Customer Relationship Management* –CRM) um sistema de apoio à estratégia de negócios através da segmentação do mercado consumidor;
- Apresentar as tecnologias *soft computing* como ferramentas recentes da mineração de dados para segmentação do mercado consumidor;

- Apresentar a Rede Neural *Self-Organizing Maps* e o Classificador Bayesiano Simples (*Naive* Bayes) como técnicas *soft computing* que podem ser aplicadas na segmentação do mercado consumidor.

1.3
Delimitação e Relevância do Estudo

A pesquisa fundamenta-se no tema perdas não-técnicas no sistema de distribuição de energia elétrica e, encontra-se orientada nos aspectos teóricos da busca de conhecimento em banco de dados, da segmentação do mercado consumidor e das técnicas *soft computing* para criar um sistema de suporte a gestão da inadimplência em empresas de distribuição de energia elétrica. Apresenta contribuições acadêmicas de modo que propõe um sistema para solucionar problema real enfrentado no âmbito econômico-político de uma empresa.

1.4
Metodologia

Conforme Vergara (2013), quanto aos fins essa pesquisa é classificada como: descritiva (pois expõe características de determinada população ou fenômeno), metodológica (visto que utiliza instrumentos de captação ou manipulação da realidade) e aplicada (por apresentar uma finalidade prática motivada pela necessidade de resolver problemas concretos). E, quanto aos meios de investigação, classifica-se em: bibliográfica (pois faz uso de material publicado em livros, redes eletrônicas, jornais e revistas científicas), documental (pois utiliza registros provenientes da base de dados de determinada empresa distribuidora de energia elétrica) e *ex post facto* (observa-se o comportamento das variáveis que compõe a base de dados da distribuidora para extração de padrões).

A sequência de desenvolvimento da pesquisa e o encadeamento das atividades que a compõem são apresentados na Figura 1.

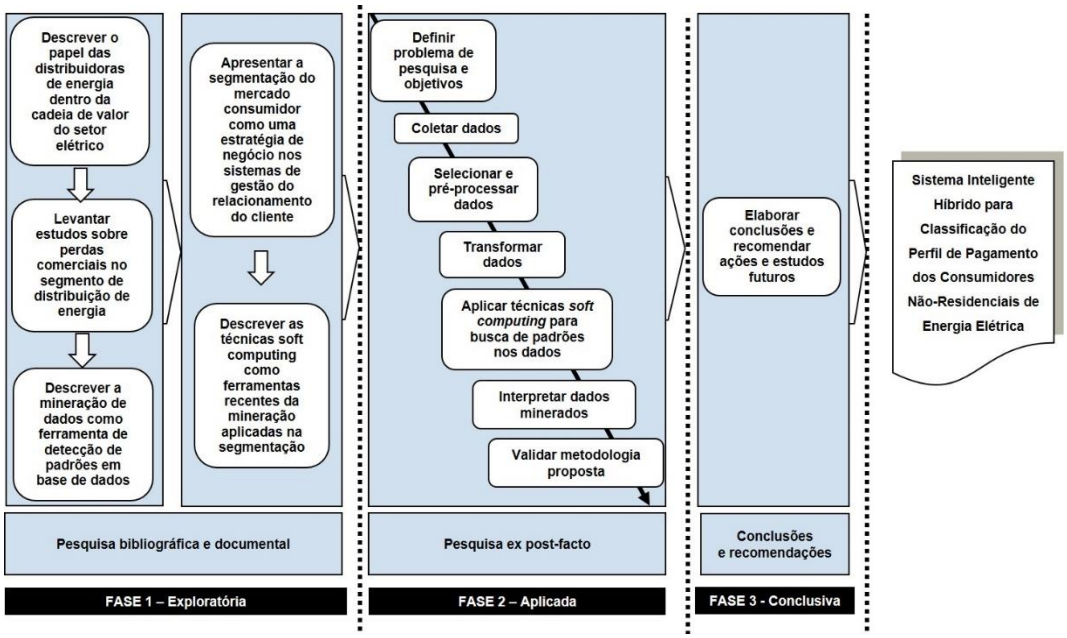

Figura 1- Desenho da Pesquisa
Fonte: Elaboração própria

1.5
Estrutura da Pesquisa

A presente pesquisa é composta por cinco capítulos. O primeiro pretende apresentar o problema de pesquisa, bem como, descrever a estrutura desta. O segundo capítulo apresenta a base conceitual para seu desenvolvimento. No terceiro capítulo, um sistema é proposto para solucionar o problema de pesquisa é apresentada. O quarto capítulo se reserva a aplicação e a discussão dos resultados obtidos com o sistema proposto. As conclusões e recomendações para futuros desdobramentos do estudo são apresentados no capítulo cinco. Por fim, encontram-se as referências bibliográficas utilizadas na elaboração desse estudo e os materiais complementares à pesquisa agrupados no apêndice.

2
Fundamentação Teórica

Este capítulo é dedicado à revisão da literatura das áreas de conhecimento que fundamentam a pesquisa proposta.

2.1
Setor Elétrico

O setor elétrico, nas últimas três décadas, passou por um conjunto de mudanças legais, normativas e tecnológicas que promoveram alterações na estrutura do setor, nos processos de comercialização de energia e na operação e expansão do sistema no Brasil e em muitos países (Agência Nacional de Energia Elétrica, 2008; Tovar et al., 2011).

No setor elétrico brasileiro atuam (Carvalho, 2011): (i) instituições vinculadas ao poder executivo, que planejam, regulam, fiscalizam, controlam e monitoram seu funcionamento; (ii) agentes da indústria (G=geradores, T=transmissores, D=distribuidores e C=comercializadores), que efetuam as transações econômicas necessárias para produção e distribuição de energia aos mercados consumidores e (iii) outros agentes que buscam interesses próprios ou de um grupo (consumidores, investidores, sindicatos, etc.). A Figura 2 apresenta uma visão geral do sistema de valor[4] do atual setor elétrico brasileiro.

[4]Sistema de valor= conjunto de atividades interdependentes geradoras de valor para um determinado segmento.

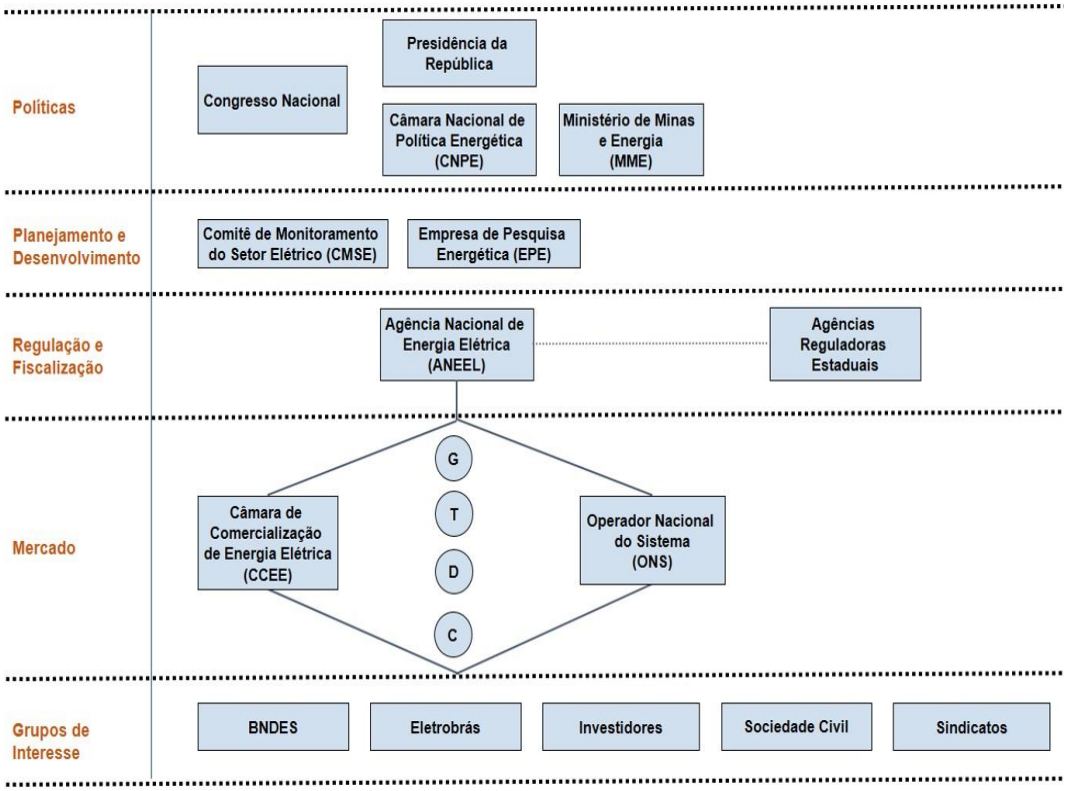

Figura 2- Visão Geral do Sistema de Valor do Setor Elétrico
Fonte: Adaptado ANEEL (2008)

2.1.1
O Mercado Consumidor de Energia Elétrica

O mercado consumidor de energia elétrica é composto por pessoas físicas ou jurídicas que solicitam o fornecimento de energia e assumem a responsabilidade pelo pagamento das faturas e outras obrigações fixadas pela ANEEL (Campos, 2010). Nesse mercado, tem-se (Câmara de Comercialização de Energia Elétrica, 2010):

- **Consumidor cativo**: aquele cujo fornecedor de energia é obrigatoriamente o agente distribuidor que detém o serviço de concessão para localidade na qual se encontra instalado. O consumidor cativo pode ser classificado como consumidor grupo A (alta e média tensão) ou consumidor grupo B (baixa tensão).
- **Consumidor livre**: aquele cujo fornecedor de energia lhe é facultado escolher através da livre negociação entre agentes de geração ou de comercialização.

- **Consumidor especial**: consumidor livre com demanda entre 500 kW e 3MW que opte em adquirir parte ou a totalidade da energia através da fonte incentivada[5].

Desta forma, a comercialização de energia pode ocorrer nos Ambientes de Contratação Regulada ou Livre, sendo suas operações de contabilização e liquidação dadas pela Câmara de Comercialização de Energia Elétrica (CCEE). A Figura 3 ilustra esses ambientes.

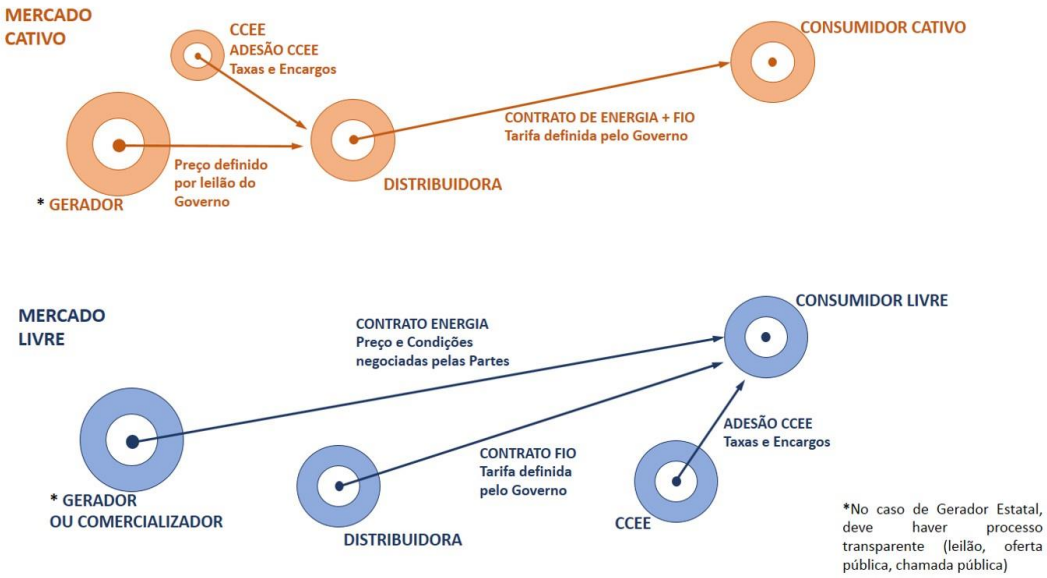

Figura 3- Mercados de Energia Elétrica
Fonte: Elaboração Própria

2.1.2
Política Tarifária, Modelo Regulatório e Composição da Tarifa

As regras que definem a responsabilidade do governo pela garantia da oferta, do serviço e o equilíbrio entre a tarifa justa para o consumidor e a remuneração adequada aos agentes da indústria são estabelecidas pela política tarifária.

Conforme Instituto Acende Brasil (2007b), no Brasil, a política tarifária adotada segue a regulação por incentivos (*Price-cap*). Tal modelo estabelece um valor teto para a tarifa, de modo que, as empresas de distribuição de energia sejam estimuladas a buscar eficiência na conexão e atendimento aos mercados consumidores para superação de parâmetros, pré-determinados pelo regulador, que lhes garantam ganhos de produtividade que cubram seus custos operacionais e retorne os investimentos realizados (Koliou et al., 2015).

[5]Fonte incentivada= energia proveniente de Pequenas Centrais Hidrelétricas (PCHs), eólica, biomassa ou solar

2.1.3
Estrutura Tarifária Consumidor Cativo

A tarifa de fornecimento de energia é segregada em tarifa do uso do sistema de distribuição (TUSD) e tarifa de energia (TE). A TUSD corresponde aos custos operacionais para distribuição de energia enquanto, a TE, aos custos associados a compra de energia. Assim, consumidor cativo paga TUSD e TE às empresas distribuidoras de energia elétrica enquanto, o consumidor livre, apenas TUSD (Carvalho, 2011).

A tarifa de fornecimento de energia do consumidor cativo (objeto de estudo desta pesquisa) é cobrada considerando a TUSD e a TE, sendo essa última definida com base na demanda de potência (kW) e consumo de energia (kMh). No caso do consumidor cativo baixa tensão (grupo B), a TE é estabelecida somente em função do consumo de energia (kMh) –tarifa monômia. E, para o consumidor cativo média e alta tensão (grupo A), considera-se tanto o consumo de energia (kMh) quanto a demanda de potência (kW) –tarifa binômia.

Os consumidores cativos baixa tensão podem ser enquadrados nas modalidades tarifárias convencional ou branca, conforme os seguintes critérios (Agência Nacional de Energia Elétrica, 2010, 2012):

- **Convencional**: aplicável compulsoriamente e automaticamente a todos consumidores do grupo B. Caracteriza-se por tarifas de consumo de energia, independentemente das horas de utilização do dia.
- **Branca**: aplicável aos consumidores do grupo B, exceto para o subgrupo B4 (iluminação pública) e subclasses baixa renda do subgrupo B1 (residencial), que optarem após publicação de resolução específica com definição dos procedimentos e critérios a serem observados. Caracteriza-se por tarifas diferenciadas de consumo de energia, de acordo com as horas de utilização do dia.

Enquanto, os consumidores cativos média e alta tensão podem ser enquadrados nas modalidades tarifárias convencional, horária verde ou horária azul, conforme os seguintes critérios (Agência Nacional de Energia Elétrica, 2010, 2012):

- **Convencional**: aplicável aos consumidores atendidos em tensão de fornecimento de energia inferior a 69 kV e com demanda de potência contratada inferior a 300 kW. Caracteriza-se por tarifas de consumo de energia e de demanda de potência, independentemente das horas de utilização do dia. Esta modalidade será extinta a partir do quarto ciclo de revisão tarifária periódica de cada distribuidora, que

- **Horária verde**: aplicável compulsoriamente aos consumidores atendidos em tensão de fornecimento de energia inferior a 69 kV e com demanda de potência contratada igual ou superior a 300 kW. Caracteriza-se por tarifas diferenciadas de consumo de energia, de acordo com as horas de utilização do dia, assim como de uma única tarifa de demanda de potência.
- **Horária azul**: aplicável compulsoriamente aos consumidores atendidos em tensão de fornecimento de energia igual ou superior a 69 kV. Caracteriza-se por tarifas diferenciadas de consumo de energia e de demanda de potência, de acordo com as horas de utilização do dia.

O Quadro 2 apresenta uma síntese das possíveis modalidades e estruturas tarifárias aplicáveis aos consumidores grupo A.

Quadro 2- Estrutura Tarifária por Modalidade Tarifária Consumidor Cativo Grupo A

Modalidade tarifária	**Compulsório para consumidores**		**Estrutura Tarifária**			
	Tensão (kV)	**Demanda contratada** (kW)	**Demanda de potência** (R$/kW)		**Consumo de energia** (R$/MWh)	
			Ponta (P)	**Fora ponta** (Fp)	**Ponta** (P)	**Fora ponta** (Fp)
Horária azul	>= 69	qualquer	$TD_{(P)}$	$TD_{(Fp)}$	$TC_{(P)}$	$TC_{(Fp)}$
Horária verde	< 69	>= 300	TD		$TC_{(P)}$	$TC_{(Fp)}$
Convencional*	< 69	< 300	TD		TC	

*Em atendimento ao disposto nos incisos I e II, § 6º do Art. 57 da Resolução Normativa ANEEL nº 414 de 09 de setembro de 2010, as unidades consumidoras cadastradas na modalidade tarifária Convencional deverão migrar para modalidade tarifária horária azul ou verde.

Ponta= período composto por três horas diárias consecutivas definidas pela distribuidora considerando a curva de carga de seu sistema elétrico, aprovado pela ANEEL para toda a área de concessão ou permissão, com exceção dos sábados, domingos e feriados nacionais; **Fora Ponta=** período composto pelo conjunto das horas diárias consecutivas e complementares àquelas definidas nos postos de ponta; **TD=** tarifa de demanda e **TC=** tarifa de consumo.
Fonte: Elaboração própria

É importante ressaltar que, a partir de 2015, os consumidores cativos, de todas as concessionárias e permissionárias conectadas pelo Sistema Interligado Nacional[6] (SIN), passaram a pagar pelos custos reais da geração de energia elétrica por meio do sistema de bandeiras tarifárias.

O custo real da geração de energia elétrica depende da fonte de energia usada para abastecer o mercado consumidor. No setor elétrico brasileiro, a fonte predominante é a hidráulica e, por isso, a geração de energia depende da chuva e do nível de água armazenado nos reservatórios. Assim, quando o nível dos reservatórios está baixo, as usinas térmicas são acionadas para garantir a segurança

[6]Sistema Interligado Nacional= sistema principal composto por usinas, linhas de transmissão e ativos de distribuição que abrange a maior parte do território brasileiro.

do sistema, porém o custo da geração sobe. A condição de operação do sistema hidrotérmico brasileiro é verificada mensalmente pelo ONS, que decide se há ou não operação de usinas termelétricas e o custo associado a essa geração. Então, com base nessas informações, a ANEEL aciona a bandeira tarifária vigente no mês seguinte (Agência Nacional de Energia Elétrica, 2016).

O sistema de bandeiras tarifárias funciona como um semáforo de trânsito que reflete custo associado a geração de energia (Agência Nacional de Energia Elétrica, 2016):

- **Bandeira verde:** indica condições favoráveis de geração de energia. Nesse caso, a tarifa de fornecimento de energia não sofre nenhum acréscimo.
- **Bandeira amarela:** indica condições de geração menos favoráveis. A tarifa de fornecimento de energia sofre acréscimo de R$ 0,015 para cada quilowatt-hora consumidos.
- **Bandeira vermelha – Patamar 1:** indica condições de geração mais custosa. A tarifa de fornecimento de energia sofre acréscimo de R$ 0,030 para cada quilowatt-hora consumidos.
- **Bandeira vermelha – Patamar 2:** indica condições de geração ainda mais custosa. A tarifa de fornecimento de energia sofre acréscimo de R$ 0,045 para cada quilowatt-hora consumidos.

2.1.4 Perdas no Setor Elétrico

Entende-se por perdas a quantidade de energia elétrica injetada nas redes de transmissão e distribuição que não serão pagas pelos usuários do sistema (Antmann, 2009). As perdas têm crescido consideravelmente em todo mundo, veja Tabela 1, sendo a maior parcela delas correspondente às perdas técnicas e não-técnicas do sistema de distribuição (Oliveira, 2009).

Tabela 1- Percentual de Perdas na Distribuição e Transmissão de Energia Elétrica

Região	% de Perdas na Transmissão e Distribuição			
	Ano 1980	Ano 1990	Ano 2000	Ano 2013
América Latina e Caribe	12,50	14,53	16,04	14,76
América do Norte	8,76	8,98	5,97	6,31
Europa e Ásia Central	7,09	7,40	9,20	8,18
Sul da Ásia	19,19	19,76	26,55	18,10
Leste da Ásia e Pacífico	6,43	6,27	6,32	5,66
Oriente Médio e Norte da África	8,87	9,38	11,71	11,99
Sul da África	7,63	6,03	8,21	8,49

Fonte: World Bank (2016)

A redução de perdas técnicas pode ser alcançada pela modernização das redes e otimização da alocação de equipamentos no sistema de distribuição, enquanto a redução de perdas não-técnicas depende de inspeções periódicas nas unidades consumidoras e de ações de combate à inadimplência (Trevisan, 2014).

Recentemente, um novo conceito de medição do consumo surgiu com os medidores inteligentes (*smart metering*), ou ainda medidores eletrônicos, –medidor de energia elétrica de alta funcionalidade que coleta dados sobre o uso da energia através de sensores e os informam automaticamente às empresas distribuidoras de energia elétrica. Com a implementação dos medidores eletrônicos, o conceito de redes inteligentes (*smart grids*) tem ganhado atenção dos agentes do setor elétrico. As redes inteligentes referem-se à adoção recursos de Tecnologia da Informação (TI) e da comunicação nos sistemas de distribuição e de transmissão, garantindo automação, maior segurança e eficiência operacional destes (Fang et al., 2012). No Brasil, as redes inteligentes estão em um estágio preliminar de definição e padronização de metodologias e tecnologias para migração das redes analógicas em inteligentes e de definição de políticas públicas para incentivar o desenvolvimento de equipamentos, por meio de projetos de Pesquisa e Desenvolvimento e projetos de demonstração em cidades inteligentes conduzidos por algumas empresas distribuidoras de energia elétrica (Lamin, 2013; Ramos, 2014). A implementação das redes inteligentes contribuirá para uso eficiente de energia pelos consumidores e permitirá às distribuidoras monitorar a rede continuamente para caracterizar melhor o perfil dos seus consumidores, além disso, um sistema flexível para consumo de energia por meio de um plano de pagamento pré-pago poderá ser adotado, minimizando perdas (Ramos, 2014).

As perdas técnicas podem ser precisamente detectadas pelo sistema através de informações sobre a energia total injetada na rede de distribuição e a energia total fornecida, enquanto as perdas não-técnicas necessitam ser estimadas (Depuru et al., 2011; Oliveira, 2009; Smith, 2004).

As perdas não-técnicas tendem a ser elevadas em países com alto nível de corrupção, baixa estabilidade política e baixa eficácia do governo (Smith, 2004). Estas perdas representam para as empresas distribuidoras de energia elétrica um obstáculo para a sustentabilidade e desenvolvimento do setor uma vez que interferem negativamente na qualidade de fornecimento de energia e impedem

melhorias e ampliação da rede (Depuru et al., 2011; Ramos, 2014). Desta forma, a redução de perdas não-técnicas é um dos temas de interesse contemplados nas chamadas públicas para projetos de Pesquisa e Desenvolvimento promovidos pela ANEEL que são executados pelos agentes distribuidores em conjunto com instituições de pesquisa (Ramos, 2014).

Inúmeros trabalhos têm proposto o uso de técnicas inteligentes e estatísticas tradicionais para classificar, identificar e caracterizar as perdas não-técnicas em distribuidoras de energia elétrica, considerando conhecimento armazenado em bases de dados. Assim, seguindo essa perspectiva, tem-se os estudos:

Araújo (2007) identificou um conjunto de fatores setoriais, sociais e econômicos que explicam a intensidade da ocorrência de perdas e inadimplência em distribuidoras de energia que atuam no Brasil, por meio de técnicas estatísticas tradicionais.

Bastos (2011) propôs uma metodologia para diagnosticar perdas não-técnicas usando redes Bayesianas.

Cabral et al. (2006) identificaram padrões de fraude, considerando dados históricos dos consumidores de distribuidoras de energia brasileiras, por meio de uma metodologia fundamentada na teoria dos *rough sets*.

Calili (2005) identificou comportamento de consumidor em um sistema de distribuição por meio da lógica *fuzzy*, considerando Pesquisa de Posses e Hábitos de Consumo realizada e grupos distintos de consumidores (adimplente, inadimplente e fraudulento) definidos pela rede auto-organizável de Kohonen.

Carvalho et al. (2014) propuseram o uso de um modelo neural não-supervisionado para segmentar consumidores de energia elétrica conforme perfil de atrasos e, com isso, contribuir para a formulação de estratégias de combate a inadimplência.

Costa et al. (2013) utilizaram algoritmos genéticos na formulação do processo de estratificação de consumidores de energia em subgrupos homogêneos e, então, selecionaram consumidores a serem inspecionados.

Dias (2006) propôs um sistema baseado em regras para identificar suspeitos de fraude.

Eller (2003) propôs um método para verificar comportamentos de fraude no consumo de energia elétrica através das redes neurais artificiais.

Faria (2012) propôs uma metodologia baseada em redes neurais e lógica *fuzzy*, para detectar unidades consumidoras com comportamento de furto e fraude, visando minimizar os procedimentos técnicos com inspeções.

Huang et al. (2013) detectaram defeitos e fraudes nos medidores por meio de uma estimativa de carga do transformador de distribuição e usaram a análise de variância (ANOVA) para criar uma lista de possíveis consumidores com problemas de medição.

Jiang et al. (2002) propuseram um método para detecção de fraude usando técnicas *wavelet* combinada com modelo classificador neural e bayesiano.

León et al. (2011) utilizaram a base de dados de uma distribuidora de energia elétrica espanhola no desenvolvimento de um sistema especialista integrado composto por módulos de mineração de dados e de texto para identificar a presença de algum tipo de perda não-técnica.

Medeiros (2013) desenvolveu um sistema de apoio a decisão baseado na experiência de uma distribuidora e, em procedimentos estatísticos, que auxiliasse a alocação de equipes de campo para viabilizar, técnica e economicamente, a inspeção de consumidores suspeitos de fraudar o consumo de energia.

Monedero et al. (2012) detectaram quedas anômalas no consumo de energia utilizando o coeficiente de Pearson e por meio das teorias de redes bayesianas e de árvore de decisão verificaram a ocorrência de fraude no sistema.

Nagi et al. (2008) propuseram o uso de máquinas de vetor de suporte (*support vector machine –SVM*) para detectar fraude no sistema, considerando perfil de carga dos consumidores.

Nizar et al. (2006) usaram métodos de caracterização de cargas e técnicas de mineração de dados para classificar, detectar e prever perdas não-técnicas devido a erros de medição e de faturamento.

Nizar et al. (2008) propuseram um sistema composto por três módulos dotados de técnicas inteligentes e de estatísticas tradicionais para detectar perdas não-técnicas no consumo de energia, a saber: **(i) 1º módulo**, agrupar consumidores conforme perfis de carga diária individual; **(ii) 2º módulo**, usar regras de aprendizagem para especificar se há perdas técnicas, não-técnicas ou comportamento suspeito e **(iii) 3º modulo**, predizer comportamento futuro do consumidor.

Ortega (2008) propôs um sistema inteligente para identificação do perfil do consumidor (normal ou irregular) por meio de técnicas de redes neurais.

Penin (2008) discutiu práticas para minimização de perdas não-técnicas, propôs melhorias nos procedimentos legais para recuperação de receitas e nos processos de combate e de prevenção dessas perdas, bem como testou algumas técnicas inteligentes (algoritmos genéticos e redes neurais) e estatísticas tradicionais (regressão logística, análise discriminante, análise de cluster) na detecção de fraude ou consumo irregular.

Petkovic e Balaban (2008) detectaram inadimplentes no sistema de distribuição de uma cidade através de uma metodologia neural que usou a base de dados desta.

Queiroga (2005) usou técnicas distintas de mineração de dados para detectar fraude e instalações irregulares.

Ramos et al. (2011) propuseram uma metodologia fundamentada no algoritmo de Floresta de Caminhos Ótimos para detectar fraude em sistema de distribuição.

Ramos (2014) desenvolveu uma metodologia para classificar e selecionar características dos padrões de comportamento de potenciais consumidores com irregularidades na medição de energia elétrica por meio de algoritmos híbridos baseados em técnicas evolutivas.

Reis Filho (2006) propôs uma metodologia para identificar consumidores potencialmente fraudadores ou com problemas em medidores usando a teoria de árvore de decisão.

Ribeiro et al. (2012) propuseram um modelo estatístico para analisar o histórico de ações tomadas na redução de perdas não-técnicas, alocando otimamente os recursos para um determinado período de tempo.

Souza et al. (2013) desenvolveram uma metodologia *fuzzy* para avaliar, antecipadamente, a capacidade de pagamento das unidades consumidoras de energia elétrica e, com isso, contribuíram para identificar possíveis inadimplentes na rede.

Trevisan (2014) identificou fraude e furto no sistema de distribuição por meio de uma metodologia fundamentada no algoritmo de Floresta de Caminhos Ótimos.

2.2
Busca de Conhecimento em Base de Dados

Nas últimas décadas, o crescimento explosivo da geração de dados pelas organizações tornou a capacidade humana em interpretá-los uma tarefa inviável. Tal fenômeno impulsionou a criação de sistemas automatizados e inteligentes de gerenciamento de base de dados. Nesse contexto, a Busca de Conhecimento em Base de Dados (*Knowledge Discovery in Databases* –KDD) surgiu como "um processo não-trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis em dados" (Fayyad et al., 1996, p. 6). Para melhor compreensão dessa definição, é necessário conceituar cada componente da mesma (Rezende et al., 2005):

Conhecimento: comparação e combinação de *n* elementos que são analisados e contextualizados dentro de determinados eventos.

Dados: conjunto de elementos em um repositório de dados.

Processo: conjunto de ações necessárias para extração de conhecimento de base de dados.

Não-trivial: complexidade para executar o processo devido a diversos fatores operacionais e de controle, tais como: necessidade de manipulação de grandes volumes de dados, dificuldade em definir tarefa (isto é, funcionalidade) e algoritmo para buscar padrões em dados.

Padrões válidos, novos e úteis: subconjunto de dados descrito em alguma linguagem que forneça informações adequadas e novas ao contexto de aplicação de Busca de Conhecimento em Base de Dados sendo possível incorporá-los no processo.

Compreensíveis: padrões descritos em alguma linguagem devem ser compreendidos pelos usuários permitindo análise profunda dos dados.

O processo Busca de Conhecimento em Base de Dados, representado pela Figura 4, reúne ferramentas, técnicas e meios computacionais junto a ação do homem em uma busca constante de conhecimentos em base de dados. Em síntese, trata-se de um sistema integrado, interativo e iterativo que envolve várias etapas, a saber (Fayyad et al., 1996; Goldschmidt & Passos, 2005):

1. **Seleção de dados:** refere-se a escolha de atributos ou registros de dados que serão considerados no processo de Busca de Conhecimento em Base de Dados;
2. **Pré-processamento:** refere-se à remoção de ruídos, escolha de decisões sobre estratégias de dados faltantes e codificação dos dados;
3. **Transformação de dados:** refere-se a redução, projeção e normalização de dados;
4. **Mineração de dados:** refere-se a escolha da técnica e algoritmos que serão usados para busca de padrões de interesses nos dados.
5. **Interpretação de padrões:** consiste em analisar e interpretar o modelo de conhecimento gerado. Conforme (Rezende et al., 2005, p. 321), "diversas medidas para avaliação de conhecimento têm sido pesquisadas com a finalidade de auxiliar o usuário no entendimento e na utilização do conhecimento adquirido". Tais medidas podem ser classificadas em medidas de desempenho (por exemplo: precisão, erro, tempo de aprendizado, etc.) e medidas de qualidade (medidas de discrepância).

6. **Ação sobre o conhecimento:** trata-se de incorporar o conhecimento dentro de outro sistema para ações futuras ou simplesmente documenta-lo e reporta-lo às partes interessadas.

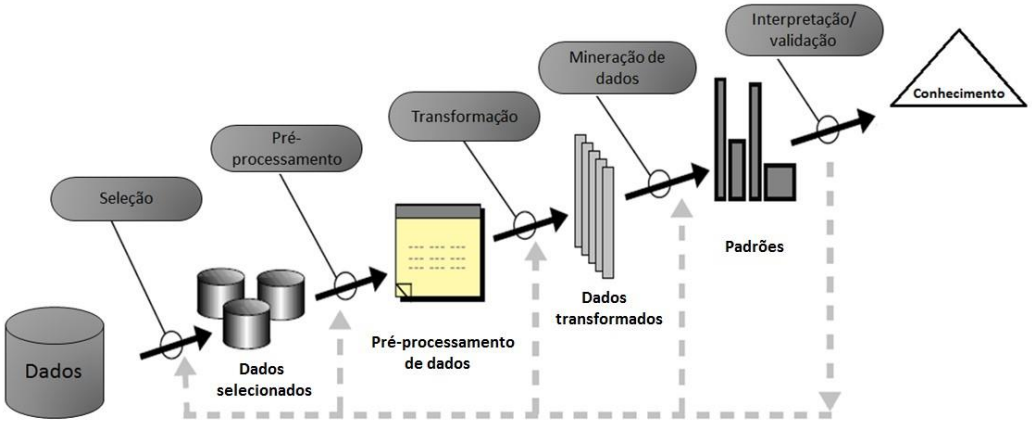

Figura 4- Visão Geral do Processo de Busca de Conhecimento em Base de Dados
Fonte: Adaptado de Fayyad et al. (1996)

2.2.1
Seleção, Pré-processamento e Transformação de Dados

Conforme Jain et al. (1999, p. 270), "não há diretrizes teóricas que sugerem quais conjuntos de exemplos[7] e atributos[8] são adequados para usar em uma situação específica". Desta forma, Rezende et al. (2005) acreditam que o conhecimento sobre o domínio de aplicação é que fornece ao analista subsídio para escolher conjunto de exemplos e definir atributos que serão usados para extração de padrões.

Normalmente, os dados não estão no formato adequado para extração de padrões, devendo o analista aplicar métodos de limpeza, transformação e redução para adequá-los ao formato. A seguir, a finalidade de cada método é descrita.

Limpeza: preencher, suavizar e corrigir dados incompletos, ruidosos e inconsistentes que podem estar presentes no conjunto de dados selecionados. Neste processo (Han & Kamber, 2006):

- **Dados incompletos** podem ser ignorados ou preenchidos com uma constante global (por exemplo: com rótulo "desconhecido", com valor tendendo a menos infinito, com valor médio do atributo da ocorrência ou com provável valor determinado por modelo de regressão). O preenchimento gera viés aos dados, porém, o preenchimento com valor provável determinado por modelo de regressão é uma estratégia muito usada. Conforme Hair Jr et al.

[7]Conjunto de exemplos= coleção de exemplos contendo valores de atributos bem como classe associada.

[8]Atributos= conforme Rezende et al. (2005, p. 95), "um atributo descreve uma característica ou um aspecto de um exemplo".

(2009), o dado incompleto pode ser ignorado quando acontecem de maneira não-aleatória e o número de dados incompletos para um caso ou observação for abaixo de 10%.

- **Dados ruidosos** podem ser detectados e "suavizados" pelo método *binning*[9] ou pelo método de regressão[10].
- **Dados inconsistentes** podem ser detectados utilizando alguma relação de dependência conhecida entre atributos e corrigidos manualmente utilizando referências externas.

Transformação: converter dados dentro de uma base de armazenamento de modo apropriado para extração de padrões. Esta conversão pode ser realizada para (Han & Kamber, 2006):

- **Integrar** dados de várias fontes (base de dados, cubo de dados e arquivos simples) em uma única base de armazenamento. Tal procedimento requer atenção do analista ao integrar as fontes, pois um dado atributo pode estar nomeado de modo distinto em cada fonte; além disso, os dados referentes a determinado atributo podem estar relacionados a elementos distintos, ou ainda, podem estar em escala diferente em cada fonte.
- **Consolidar** base de dados a fim de suavizar ruídos, adicionar atributos, sintetizar e normalizar dados. A normalização coloca os valores dos dados da base em um valor escalar entre 0 e 1. Esse procedimento é particularmente útil para algoritmos de classificação e agrupamento que utilizam redes neurais ou medidas de distância pois permite que os valores dos atributos tenham igual influência na obtenção do resultado. Veja algumas formas de normalização:

  Normalização *range*: transformação linear simples sobre os dados originais, veja Eq. (1).

$$x' = \frac{x - min_A}{max_A - min_A} \tag{1}$$

  Na Eq. (1): x representa o valor de determinado atributo A que desejamos transformar, min e max são, respectivamente, o valor mínimo e máximo desse atributo. Essa forma de normalização preserva o relacionamento entre os valores do dado original.

  Normalização *z-score*: consiste em transformar os valores de um atributo A com base na média e desvio padrão desse atributo, veja Eq. (2).

$$x' = \frac{x - \bar{x}}{\sigma_x} \tag{2}$$

[9]Os dados pertencentes a uma amostra são ordenados considerando os valores em torno dele. Os valores ordenados são distribuídos dentro de uma frequência de ocorrências a definir.

[10]Em geral, usa-se a regressão linear que envolve encontrar a equação da reta que melhor ajusta a duas variáveis e, assim, determinar quais são os dados ruidosos.

Na Eq. (2): x representa o valor de determinado atributo A que desejamos transformar, $\bar{x}$ e σ_x são, respectivamente, a média e o desvio padrão do atributo A. Esse método é útil quando o mínimo e máximo de um atributo são desconhecidos ou quando existem muitos outliers na normalização *range*.

Normalização por escala decimal: consiste em transformar os valores do atributo A em um número de ponto decimal, veja Eq. (3).

$$x' = \frac{x}{10^j} \tag{3}$$

Na Eq. (3): x representa o valor de determinado atributo A que desejamos transformar, j é o menor número inteiro tal que $Max(|x'|) < 1$.

Redução de dados: reduzir número de exemplos e de atributos da base de dados selecionada. Esta redução pode ser realizada de três maneiras (Rezende et al., 2005):

- **Redução do número de exemplos:** consiste em gerar amostras representativas do conjunto de dados original fazendo uso, geralmente, de métodos de amostragem.
- **Redução do número de valores de um atributo (discretização):** consiste na substituição de um atributo contínuo por atributos discretos. Essencialmente, um algoritmo de discretização aceita como entrada valores de um atributo contínuo e gera como saída uma pequena lista de intervalos ordenados. Técnicas como *binning* e análise de histograma são usados para esse propósito.
- **Redução do número de atributos:** consiste na detecção e remoção de atributos fracamente relevantes ou redundantes. Em geral, a opinião de especialista do domínio de aplicação ou métodos heurísticos básicos de seleção de subconjunto, como por exemplo: stepwise forward, stepwise backforward e indução por árvore de decisão são usados para essa finalidade.

2.2.2 Mineração de Dados

A mineração de dados é a etapa do processo de Busca de Conhecimento em Base de Dados na qual efetivamente ocorre a extração de conhecimento. A extração de conhecimento em base de dados requer que o usuário defina a princípio a tarefa (isto é, funcionalidade) e, então, escolha técnica mais apropriada para extração de conhecimento.

A tarefa é definida pelo usuário conforme meta desejada em atividades de **descrição** e **predição** (Côrtes et al., 2002; Fayyad et al., 1996; Goldschmidt &

Passos, 2005; Han & Kamber, 2006; Rezende et al., 2005). A Figura 5 apresenta uma síntese das tarefas relacionadas a cada atividade.

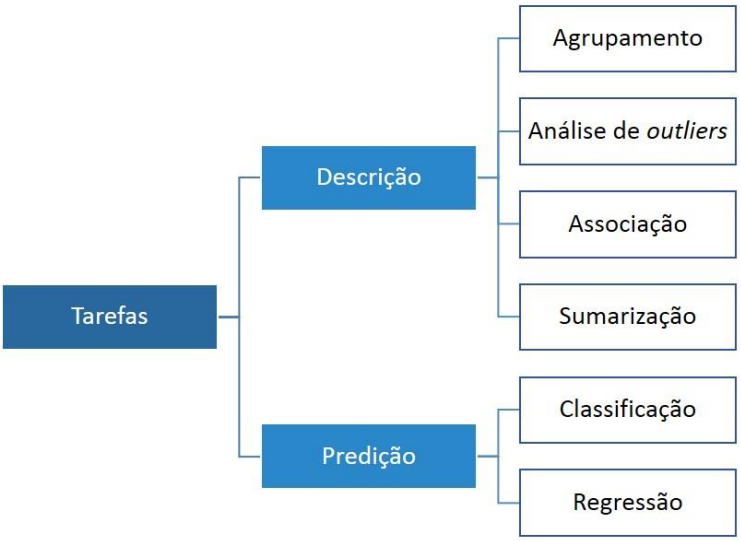

Figura 5-Tarefas de Busca de Conhecimento em Base de Dados
Fonte: Elaboração própria

As atividades de descrição consistem em identificar comportamentos intrínsecos do conjunto de dados. Enquanto as atividades de predição, em generalizar exemplos passados com respostas conhecidas em linguagem capaz de reconhecer a classe de um novo exemplo. É válido mencionar que, a classificação das tarefas de Busca de Conhecimento em Base de Dados não é um consenso dentro da literatura (Chen et al., 1996; Groth, 1998). A tarefas descritas pela Figura 5 podem se unir originando outras tarefas mais complexas (Goldschmidt & Passos, 2005).

Ao definir tarefa a ser empregada, o usuário deve especificar técnica para executá-la. Conforme Rezende et al. (2005, p. 327), "as técnicas descrevem um paradigma de extração de conhecimento e vários algoritmos podem seguir esse paradigma". Na literatura, existe uma variedade de técnicas pertencentes às áreas de estatística, tecnologia de base de dados, aprendizado em máquina, reconhecimento de padrão e inteligência artificial disponível para mineração de dados (Han & Kamber, 2006; Hand et al., 2001)

A seguir, as tarefas usadas no desenvolvimento desta pesquisa serão descritas, para detalhamento das demais, veja Han & Kamber (2006).

2.2.2.1
Tarefa de Sumarização

A tarefa de sumarização consiste em encontrar uma descrição compacta para um subconjunto de dados por meio de medidas numéricas ou gráficas (Rezende et al., 2005), permitindo ao pesquisador ter uma visão crítica das características dos dados (Hair Jr et al., 2009, p. 52). A análise descritiva dos dados pode seguir abordagem univariada, bivariada ou multivariada, dependendo da quantidade de variáveis em questão. As variáveis podem ser classificadas em:

Variáveis quantitativas: variáveis que podem ser medidas em uma escala numérica. Classificam-se em (Triola, 2011):

- ***Variáveis Discretas***: aquelas que possuem características mensuráveis em que somente os valores inteiros fazem sentido, normalmente proveniente de contagem. Exemplo: número de empresas.
- ***Variáveis Contínuas***: aquelas que possuem características mensuráveis em que somente os valores fracionários fazem sentido, consequentemente assumem valores em uma escala contínua. Exemplo: tempo de produção.

Variáveis qualitativas (ou categóricas ou de atributos): variáveis que não podem ser medidas em uma escala numérica, sendo definidas por categorias ou classificações. Classificam-se em (Triola, 2011):

- ***Variáveis Nominais***: aquelas que não apresentam ordenação entre as categorias. Exemplos: sexo, estado civil.
- ***Variáveis Ordinais***: aquelas que apresentam ordenação entre as categorias. Exemplo: mês de observação (janeiro, fevereiro, ..., dezembro).

As medidas numéricas são informações resumidas que caracterizam um conjunto de dados.

Medidas de tendência central são medidas numéricas que produzem um valor médio representativo de um conjunto de dados (Spiegel & Stephens, 2008). A média aritmética, mediana e moda são as medidas mais comumente usadas para esse fim. Cabe mencionar que, a moda é pouco usada com dados numéricos e o valor da média leva em consideração todos os valores e, por isso, pode ser dramaticamente afetado pela presença de valores extremos, enquanto a mediana não é tão sensível a essa presença (Triola, 2011). Desta forma, veja que não há uma única melhor medida de centro, cada uma apresenta vantagens e desvantagens.

- ***Média aritmética:*** encontrada pela adição dos valores e divisão do total pelo número de valores.
- ***Mediana***: valor do meio quando os dados originais estão arranjados em ordem crescente (ou decrescente) de magnitude. Caso o número

de valores for ímpar, a mediana será o número localizado no meio exato da lista. Caso o número de valores for par, a mediana será encontrada pelo cálculo da média aritmética dos dois números do meio.

- ***Moda:*** valor mais frequente em um conjunto de dados.

Medidas de dispersão são medidas numéricas que representam o grau de variabilidade entre dados (Spiegel & Stephens, 2008). Nesse caso, as mais utilizadas são: amplitude, desvio padrão e a variância.

- ***Amplitude:*** diferença entre o maior valor e o menor valor de um conjunto de dados.
- ***Desvio padrão***: medida de variação de todos os valores em torno da média. Para obtê-lo, deve-se: calcular os desvios entre cada valor individual e a média; elevar cada uma das diferenças ao quadrado e, então, dividir o somatório das diferenças ao quadrado pelo número de observações.
- ***Variância:*** medida da variação igual ao quadrado do desvio padrão.

As medidas gráficas complementam a informação dada pelas medidas numéricas, fornecendo uma representação visual das relações básicas. A seguir, detalham-se algumas das técnicas mais amplamente usadas para o exame de características da distribuição, das relações bivariadas e das diferenças de grupos (Hair Jr et al., 2009).

Histograma é uma representação gráfica de uma única variável que representa a frequência de ocorrências dentro de categorias de dados. As frequências são graficamente representadas para examinar a forma da distribuição de valores. Cabe mencionar que a forma de qualquer distribuição pode ser descrita por duas medidas (Hair Jr et al., 2009; Triola, 2011):

- **Assimetria:** usada para descrever o equilíbrio da distribuição. A medida de assimetria é baseada nas relações entre média, mediana e moda. Quando estas medidas são idênticas em valor, diz-se que a distribuição dos dados é simétrica. Quando a média se distancia da moda, situando-se a mediana em uma posição intermediária, diz-se que a distribuição dos dados é assimétrica. Assim, quanto maior for a distância (seja positiva ou negativa), maior é a assimetria da distribuição. Uma assimetria positiva denota uma distribuição deslocada para a esquerda, enquanto uma assimetria negativa reflete um desvio para a direita. As distribuições assimétricas podem ser transformadas calculando-se a raiz quadrada, logaritmos, quadrados ou cubos.
- **Curtose:** usada para descrever a altura de uma distribuição. A medida de altura (“alongamento” ou o “achatamento”) de uma distribuição é comparada com a Distribuição Normal. Quanto mais achatada for uma distribuição, maior a dispersão. Quanto mais

alongada, menor a dispersão. Para a distribuição achatada, a transformação mais comum é a inversa.

Diagrama de dispersão é o método mais usualmente empregado para examinar relações bivariadas (Triola, 2011). Trata-se de um gráfico de pontos que representa os valores conjuntos correspondentes de duas variáveis para qualquer caso dado. Três tipos de relação entre variáveis podem ser visualizados através do padrão de pontos: (i) linear, caracterizada pela forte organização dos pontos ao longo de uma linha reta; (ii) não-linear, caracterizada por um conjunto curvilíneo de pontos; (iii) indeterminada, caracterizada por um padrão aparentemente aleatório de pontos que indica relação alguma.

Gráfico de caixas é uma representação ilustrativa da distribuição de dados de uma variável métrica para cada grupo de uma variável não-métrica (Triola, 2011). O gráfico de caixas (Figura 6) é composto por quartis superior e inferior da distribuição de dados que formam os limites superior e inferior da caixa, com comprimento da caixa sendo a distância entre o 25º percentil e o 75º percentil, tendo a caixa 50% dos dados centrais. A mediana é representada por uma linha sólida dentro da caixa. O exame desse gráfico, permite-nos dizer que (Hair Jr et al., 2009): (i) quanto maior a caixa, maior a dispersão das observações; (ii) mediana próxima de um extremo da caixa, assimetria na direção oposta é indicada; (iii) linhas que se estendem a partir de cada caixa (chamadas de *whiskers*) representam a distância da menor e da maior das observações que estão a menos de um quartil da caixa; (iv) observações atípicas (variam entre 1,0 e 1,5 quartis de distância da caixa) e valores extremos (observações a mais de 1,5 quartis do extremo da caixa) são representados por símbolos fora dos *whiskers*.

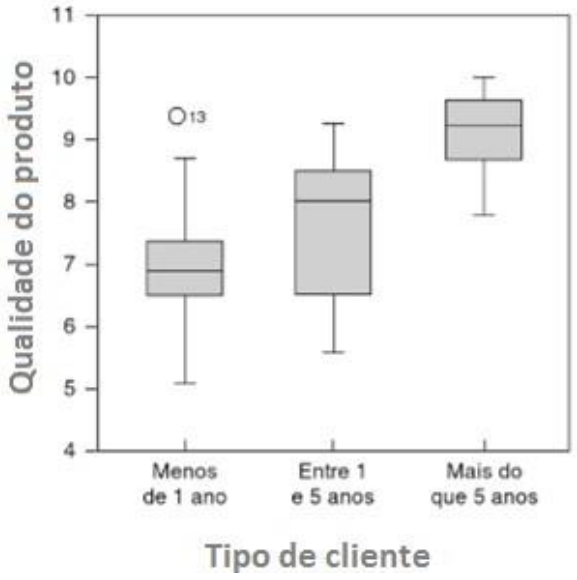

Figura 6- Gráfico de Caixas
Fonte: Hair Jr et al. (2009)

Gráfico dos perfis (Figura 7) é um retrato direto das diferenças e similaridades entre grupos (Hair Jr et al., 2009). Neste gráfico, cada grupo é

descrito por um conjunto de barras que representa a média das observações para cada variável em análise.

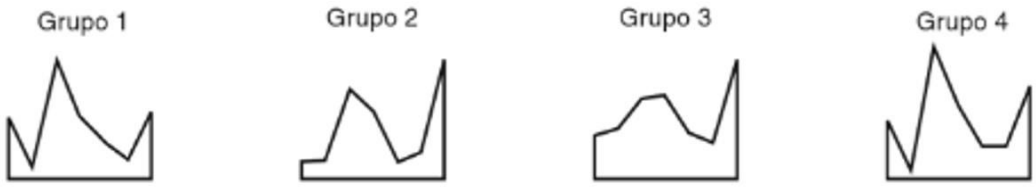

Figura 7- Gráfico dos perfis
Fonte: Hair Jr et al. (2009)

2.2.2.2 Tarefa de Agrupamento

O termo agrupamento é usado por diversas comunidades de pesquisa para se referir a partição de um conjunto de dados em *n* subconjuntos homogêneos tendo como critério a similaridade entre elementos. Em geral, o agrupamento é realizado para reduzir número de registros da base de dados ou, ainda, como etapa de pré-processamento para realização de outras tarefas– por exemplo, a classificação (Halkidi et al., 2001).

A tarefa de agrupamento consiste em um método de aprendizagem não supervisionado e envolve as seguintes etapas (Jain et al., 1999):

Seleção de dados (escolha do número de exemplos[11] e do número, tipo e escala de atributos): consiste em selecionar dados de interesse e colocá-los no formato adequado para extração de conhecimento pelo método escolhido.

Definir padrão de proximidade entre elementos: fator de central importância na identificação de grupos em um conjunto de exemplos, podendo ser medido diretamente, através da análise subjetiva do homem em relação aos objetos, ou indiretamente, através de vetores de medições ou características que descrevem cada objeto (Hand et al., 2001). Em muitas aplicações, a proximidade é medida indiretamente e, frequentemente, referida em pesquisas de agrupamento por dissimilaridade, distância ou similaridade (Everitt et al., 2011).

Escolha do método, técnica e algoritmo de agrupamento: na literatura, existem vários métodos de agrupamento (Jain et al., 1999). Tais métodos são classificados conforme modelo estatístico ou princípio de indução que se aplica ao algoritmo de agrupamento (Estivill-Castro, 2002). Para detalhamento, veja Rokach & Maimon (2005).

[11]Exemplo= conforme Rezende et al. (2005, p. 95), "um exemplo, também denominado caso, registro ou dado na literatura, [consiste em] uma tupla de valores de atributos (ou um vetor de valores de atributos)" sendo, a reunião dos *n* vetores de valores de atributos associados às suas respectivas classes (ou rótulos) denominada conjunto de exemplos.

Interpretação dos resultados: os mecanismos para interpretação das saídas do agrupamento dependerão do método escolhido para desempenhar tal tarefa.

Validação dos resultados: trata-se do procedimento de avaliação dos resultados do algoritmo. Na literatura, existem dois principais critérios de avaliação de desempenho de modelos de agrupamento (Chaimontree et al., 2010):

- **Critério interno:** refere-se a medidas que usam apenas informações contidas no próprio grupo de dados. Nesse caso, tem-se: ***medidas de coesão*** que validam a solidez dentro de um grupo e ***medidas de separação*** que validam o isolamento entre grupos, sendo essas validações feitas, por exemplo, através de medidas de proximidade de objetos (Liu et al., 2010).
- **Critério externo:** refere-se a medidas externas ao conjunto de dados. Nesse caso, os resultados do algoritmo de agrupamentos são avaliados com base em uma estrutura pré-definida, a qual é aplicada a um conjunto de dados e reflete nossa intuição sobre o agrupamento (Halkidi et al., 2001). Conforme Liu et al. (2010), a entropia é uma medida de validação externa pois avalia a "pureza" dos grupos com base nos rótulos dados.

2.2.2.3 Tarefa de Classificação

O termo classificação é usado para se referir à atribuição de um rótulo de classe a cada objeto dentro de um conjunto de dados. Conforme Jackson (2002, p. 275), "o rótulo de classe é um identificador qualitativo discreto" atribuído a cada objeto.

A tarefa de classificação consiste em "encontrar um conjunto de modelos (ou funções) que descrevem e distinguem classes [dentro de um conjunto de dados]" (Han & Kamber, 2006, p. 24). Trata-se de um método de aprendizagem supervisionado no qual é "fornecido ao algoritmo de aprendizado, ou indutor, um conjunto de exemplos de treinamento para os quais o rótulo da classe associada é conhecido [e serve como base para a classificação de novo conjunto de exemplos]" (Rezende et al., 2005, p. 91). A tarefa de classificação envolve as seguintes etapas:

Seleção de dados: existem duas abordagens gerais, a saber (Hand et al., 2001): **(i)** um algoritmo de classificação executa automaticamente a seleção de variáveis como parte da definição do modelo e **(ii)** o classificador é usado como uma "caixa preta" e tem um *loop* externo que sistematicamente acrescenta e subtrai variáveis do conjunto de dados. O conjunto de dados selecionados é representado por um conjunto de tupla de dados rotulados.

Escolha do método, técnica e algoritmo de classificação: a escolha do método depende de recursos do problema, dos dados e do objetivo em executar a tarefa de classificação (Hand et al., 2001). Han & Kamber (2006) acrescentam outros fatores para escolha do método de classificação: custo computacional envolvido na execução da tarefa, facilidade na interpretação dos resultados e capacidade de predição do modelo.

Validação dos resultados: trata-se do procedimento de avaliação dos resultados do algoritmo. Existem diversas maneiras de avaliar o desempenho de um modelo de classificação (Bramer, 2007; Hand et al., 2001). O critério de avaliação comumente utilizado é a acurácia – também denominada, precisão do classificador (Bramer, 2007; Goldschmidt & Passos, 2005). Conforme Bramer (2007) e Han & Kamber (2006), existem três principais estratégias usadas para medir a acurácia, a saber:

- **Conjunto *hold-out*:** consiste em dividir a base de dados em conjuntos treino e teste, veja Figura 8. Primeiramente, uma parte da base de dados (tipicamente entre 50% a 90% da base) é usada para construir um classificador. Posteriormente, o modelo classificador é usado para prever a classificação do conjunto de dados restante (conjunto teste).

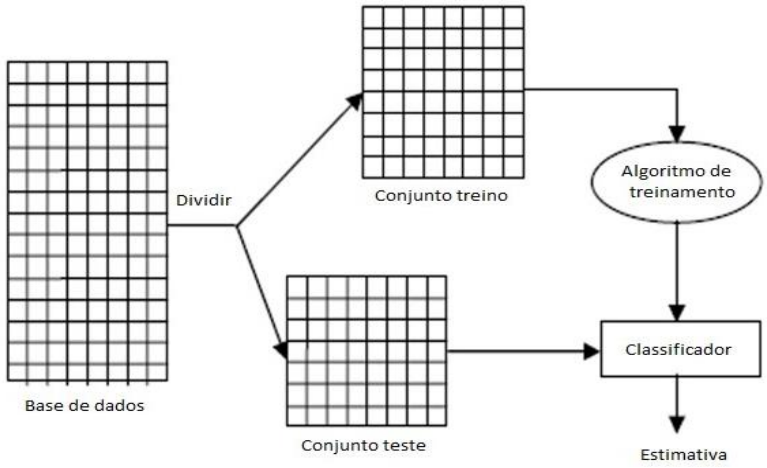

Figura 8- Validação Cruzada *hold-out*
Fonte: Adaptado de Bramer (2007)

- **Validação cruzada *k-fold*:** consiste em dividir o conjunto de dados composto por N elementos em k partes iguais, em geral, $k = 5$ ou $k = 10$. Assim, $k - 1$ partes são usados como conjunto de treino e k partes como conjunto teste. Este método é ilustrado pela Figura 9.

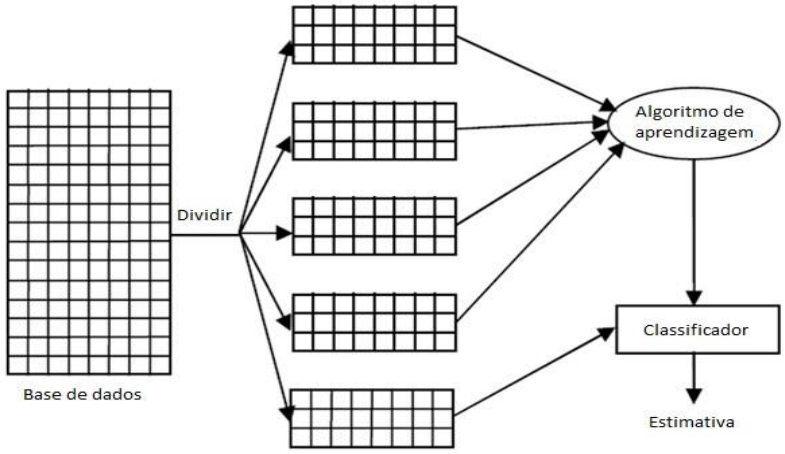

Figura 9- Validação Cruzada k-fold
Fonte: Adaptado de Bramer (2007)

- **Validação cruzada *N-fold* (*leaving-one-out*):** refere-se a um caso extremo de k-fold. O conjunto de dados é dividido em várias partes de modo que forme vários conjuntos com um elemento cada. Neste caso, N conjunto de dado são formados. N-1 conjuntos são usados para classificar um conjunto de um único elemento.

A acurácia é obtida pela Eq. (4):

$$Acc = 1 - Err(h) \tag{4}$$

Na Eq. (4), $Err(h)$ refere-se a taxa de classificação incorreta definida conforme Eq. (5).

$$Err(h) = \frac{\sum_{i=1}^{N} \|y_i \neq h(i)\|}{N} \tag{5}$$

Na Eq. (5), tem-se:

$\|E\|$: operador que retorna 1 se a operação E for verdadeira e 0, caso contrário;
y_i: classe real associada ao i-ésimo elemento
$h(i)$: classe indicada pelo classificador para o i-ésimo elemento

N: número de elementos da base de dados

2.3
Segmentação do Mercado Consumidor

Entende-se, por segmentação, a partição de um grupo heterogêneo em vários grupos homogêneos menores considerando características de demografia, comportamentos, valores, etc., sendo tal partição realizada através da tarefa de agrupamento ou classificação.

Conforme Hiziroglu (2013) e Tkaczynski & Rundle-Thiele (2011), a segmentação do mercado foi introduzida na literatura de marketing por Smith (1956), o qual a considera uma estratégia de gestão e marketing. Nesse mesmo sentido, conforme Ngai et al. (2009) e Kracklauer et al. (2004), técnicas de segmentação são utilizadas na Gestão de Relacionamento com o Cliente[12] visando identificar e maximizar o valor do cliente para a organização.

Tkaczynski & Rundle-Thiele (2011) mencionam que existem diferentes abordagens de segmentação do mercado consumidor que são utilizadas para satisfazer às necessidades dos pesquisadores. Goyat (2011), Sun (2009) e Yankelovich & Meer (2006) são exemplos de estudos revisionais sobre segmentação.

Na literatura, não existe uma orientação genérica sobre critérios para uma segmentação eficaz, visto que pesquisadores os estabelecem conforme área que atuam (Biggadike, 1981; Dibb, 1995, 1999; Goller et al., 2002; Kotler, 2003; Raaij & Verhallen, 1994; Wendel & Kamakura, 2000). Conforme Hiziroglu (2013), os critérios mais comuns para uma segmentação eficaz são os estabelecidos por Kotler (2003), a saber:

MENSURABILIDADE, capacidade de medir as variáveis dos segmentos;

ACESSIBILIDADE, capacidade de verificar a existência de determinado segmento como representativo de uma categoria de clientes;

CAPACIDADE DE DIFERENCIAÇÃO, os segmentos devem ser distinguíveis;

SUBSTANCIALIDADE, a dimensionalidade de segmentos deve ser proporcional ao estudo realizado;

CAPACIDADE DE AÇÃO, possibilidade de implementação de ações para alocar ou realocar clientes nos segmentos.

O estudo sobre segmentação, embora não seja um assunto de pesquisa novo, ainda tem muito a evoluir teórico e metodologicamente, acreditam Blocker & Flint (2007).

Recentemente, uma família de técnicas de mineração de dados denominada *Soft Computing* (SC) tem sido usada na segmentação do mercado consumidor (Hiziroglu, 2013). Tal termo foi criado por Zadeh no início dos anos 90 e se refere,

[12]Gestão de Relacionamento com o Cliente= sistema de apoio à estratégia de negócios composta por quatro elementos estruturais, a saber: identificação, atração, retenção e desenvolvimento do cliente.

basicamente, a "um corpo não homogêneo de conceitos e técnicas [formado por] uma parceria de métodos distintos que de uma forma ou de outra estão de acordo com o seu princípio orientador" que é explorar a imprecisão e incerteza para alcançar robustez, baixo custo da solução e melhor proximidade com a realidade (Zadeh, 1994, p. 49). Conforme Magdalena (2010, p. 150), SC é um conceito em constante evolução e tem como componentes tradicionais:

- **Lógica *fuzzy*:** extensão da teoria clássica de conjuntos criada para tratar graus de pertinência intermediários de elementos de um universo.

- **Redes neurais artificiais:** modelo matemático que se assemelha a estrutura dos neurônios biológicos com capacidade computacional de aprendizado e generalização.

- **Computação evolucionária:** família de técnicas de busca e otimização que utilizam modelos computacionais baseados na teoria da evolução natural. Também podem ser encontrados na literatura como algoritmos evolutivos. Enquadram-se nesse contexto: algoritmos genéticos, estratégias de evolução e programação genética.

- **Raciocínio probabilístico:** família de metodologias para raciocinar sobre o conhecimento incerto por meio da teoria da probabilidade. Enquadram-se nesse contexto: redes Bayesianas e redes de Markov.

Conforme Hiziroglu (2013), o uso de técnicas *soft computing* para mineração de dados tem se mostrado uma tendência em pesquisas recentes que abordam esse tema. Um estudo bibliométrico realizado por esse autor, revelou que:

- Técnicas *soft computing* têm crescido como área potencial em pesquisas de segmentação;
- Embora diversos setores da economia aplicam técnicas de *soft computing* em segmentação de clientes, ainda há pouca aplicação dessas técnicas em problemas de segmentação relacionados às áreas de negócios e gestão;
- Na maioria dos estudos, aproximadamente 80%, o uso das técnicas *soft computing* são exclusivamente voltadas para realização ou contribuição de melhorias na tarefa de agrupamento, sendo a aplicação dessas técnicas em tarefas de classificação ou em ambas tarefas ainda pouco usado.

2.4
Rede Neural Artificial *Self-Organizing Maps*

A rede neural artificial é um modelo matemático que se assemelha às estruturas neurais biológicas e possui capacidade computacional adquirida por meio de aprendizado e generalização (Rezende et al., 2005). McCulloch & Pitts (1943)

foram os primeiros a descreverem um modelo matemático que representasse um neurônio biológico. Posteriormente, diversas arquiteturas de redes foram desenvolvidas para resolver problemas de predição, de categorização e de otimização (Carvalho et al., 1998; Rezende et al., 2005).

A arquitetura de uma rede neural artificial é composta por unidades de processamento simples (nodos, neurônios) que são interligadas por um grande número de conexões– conhecidas como pesos sinápticos, os quais armazenam o conhecimento e o passa para a rede através de um processo de aprendizagem (Carvalho et al., 1998; Haykin, 2001). Tal processo consiste em uma forma iterativa de adaptação dos parâmetros neurais por um algoritmo de aprendizado. Na literatura, existem três paradigmas de aprendizagem que determinam como os parâmetros neurais são ajustados para alcançar uma tarefa desejada (Rezende et al., 2005):

- **Aprendizagem supervisionada:** caracteriza-se pela existência de um professor, ou supervisor, que fornecem os pares de entrada e saída desejados para a rede. Nesse caso, os parâmetros da rede são ajustados de forma a encontrar uma ligação entre os pares de entrada e saída fornecidos;
- **Aprendizagem não-supervisionada:** caracteriza-se pela não existência de saídas desejadas para as entradas. Nesse caso, os parâmetros da rede são ajustados considerando apenas os valores dos vetores de entrada;
- **Aprendizagem por reforço:** caracteriza-se por ser uma abordagem intermediária entre os paradigmas apresentados acima. Nesse caso, o conjunto de treinamento é formado pelos vetores de entrada e um crítico externo que retorna um sinal de reforço ou penalidade indicando se uma determinada saída está correta ou não, tendo como base a última ação da rede.

Conforme Haykin (2001), a escolha do paradigma de aprendizado depende do tipo de tarefa que a rede neural deve executar. No caso da tarefa de agrupamento –proposta dessa pesquisa–, aplica-se a aprendizagem não-supervisionada.

A Rede Neural Artificial *Self-Organizing Maps* (Rede SOM), introduzida na literatura por Teuvo Kohonen na década de 80, é uma rede neural artificial que segue paradigma de aprendizado não-supervisionado e têm sido amplamente aplicadas em tarefas de agrupamento em diversas áreas, tais como: indústria, finanças, ciências naturais e linguística (Kohonen, 2013; Vesanto, 1999; Yang et al., 2012). A seguir, apresenta-se detalhamento dessas redes.

A Rede *Self-Organizing Maps* é inspirada no funcionamento neurofisiológico do cérebro pois, baseia-se no mapa topológico presente no córtex cerebral, o qual

se subdivide em áreas e subáreas ricas em neurônios e responsáveis por funções específicas– tais como: fala, visão, etc.– que são desempenhadas através de um mapeamento interno de resposta ao órgão sensorial de mesma natureza (Carvalho et al., 1998; Kohonen, 1990).

A arquitetura da Rede SOM, ilustrada pela Figura 10, é formada por uma camada de entrada com $\boldsymbol{x}$ observações m-dimensionais e uma camada de saída representada por uma grade que consiste em k neurônios associados a w pesos m-dimensional (Everitt et al., 2011).

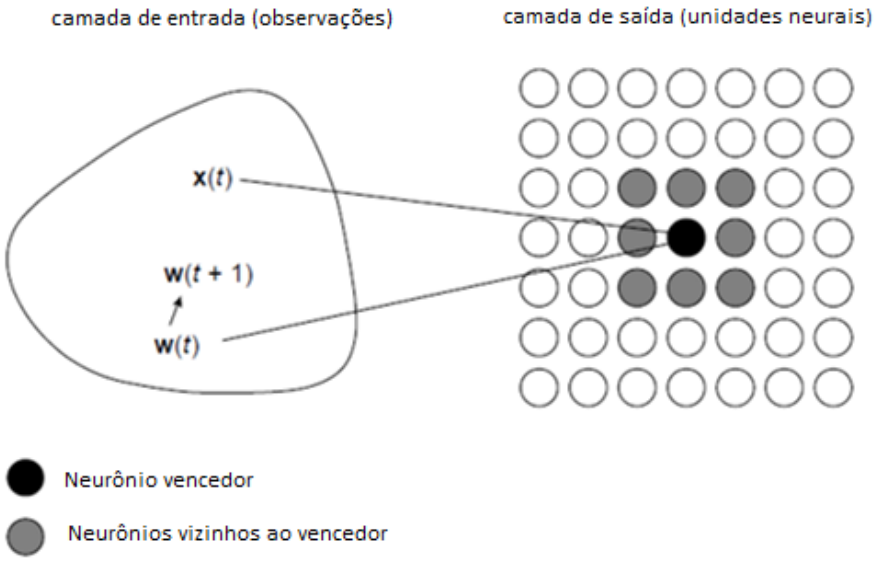

Figura 10- Rede SOM
Fonte: Everitt et al. (2011)

Nessa estrutura, cada vetor de entrada deve selecionar a unidade que melhor o represente, sendo tal unidade (neurônio vencedor) e suas unidades vizinhas (neurônio vizinho) modificadas a cada instante de tempo através de um processo iterativo até que uma melhor adequação dos dados de entrada seja alcançada (Kohonen, 2013).

Em uma Rede SOM, as unidades neurais são conectadas entre si por uma relação de vizinhança determinada pela topologia ou estrutura do mapa, que pode assumir vários formatos diferentes, tais como: hexagonal e quadrado, veja Figura 11 (Faria et al., 2010; Vesanto et al., 1999, 2000). Conforme Carvalho et al. (1998, p. 114), "a definição do formato mais adequado depende do problema atacado e da distribuição de seus dados. Geralmente, o formato da vizinhança é definido por tentativa e erro". Kohonen (1997a) afirma que a adaptação da rede é diretamente

influenciada pela formato da vizinhança e, que tradicionalmente, o formato hexagonal apresenta melhores resultados que o retangular.

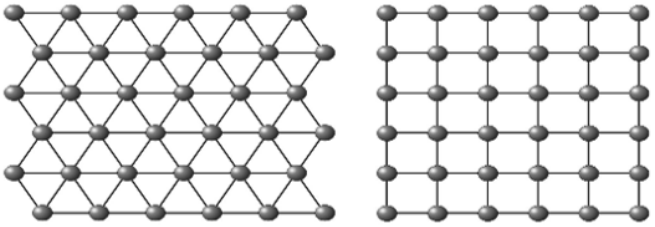

Figura 11- Formato da Vizinhança
Fonte: Elaboração própria

Em relação a quantidade de unidades neurais, Carvalho et al.(1998) mencionam que deve ser determinada, a priori, pelo usuário através de um processo de "tentativa e erro" na busca por uma melhor configuração da rede. De modo complementar, tem-se o estudo de Wilppu (1997) concluindo que essa escolha depende da grandeza do desvio que se deseja detectar entre as unidades da rede, de modo que, grandes desvios são notificados em redes menores e pequenos desvios são notificados em redes maiores. Ainda conforme esse autor, precisa-se de uma rede maior quando se tem um conjunto de dados de entrada grande pois, essa rede apresenta mais desiguais em uma unidade neural.

O processo de aprendizagem, ou simplesmente treinamento, da rede SOM pode ser realizado de duas maneiras distintas (Faria et al., 2010; Vesanto et al., 2000): **(i)** em batelada, a atualização dos pesos sinápticos é realizada somente após a apresentação de todos os elementos do conjunto de dados utilizados ou **(ii)** sequencial (incremental), a atualização dos pesos sinápticos é realizada toda vez que um exemplo é apresentado a rede. A seguir, os passos básicos que envolvem a aplicação do algoritmo são descritos.

1ª etapa: Inicialização dos parâmetros. Nessa etapa, **x** observações, m-dimensionais, são apresentadas a rede e, então, o algoritmo responsável pela formação do mapa auto organizável inicializa os pesos sinápticos aleatoriamente, atribuindo-lhes valores pequenos tomados de um gerador de números aleatórios, ou linearmente, considerando uma sequência de vetores bidimensionais gerados pelos dois maiores componentes principais de **x** (Haykin, 2001; Kohonen, 2013). Conforme Kohonen (1997b), a inicialização linear previne torções indesejáveis no mapeamento ao longo do treinamento. Outros parâmetros também são inicializados,

tais como (Haykin, 2001): taxa inicial de aprendizagem ($\eta_0 \cong 0{,}1$) e raio inicial entre neurônios na vizinhança topológica ($\sigma_0 =$ valor igual ao "raio da grade").

2ª etapa: competição entre neurônios. Nesta etapa, o "espaço contínuo de vetores de entrada [são mapeados] para um espaço discreto de saída de neurônios por um processo de competição entre os neurônios da grade" (Haykin, 2001, p. 488). Conforme Carvalho et al. (1998), a ativação de um neurônio é determinada pela distância entre seu peso e o vetor de entrada, veja Eq. (6). Na Rede SOM, a função de ativação é baseada na medida de distância Euclidiana (Carvalho et al., 1998; Kohonen, 1990, 2013).

$$i(\boldsymbol{x}) = arg\, \min_j \left\| \boldsymbol{x} - \boldsymbol{w}_j \right\|, \quad j = 1, 2, \cdots, l \tag{6}$$

Na Eq. (6), tem-se:

$i(\boldsymbol{x})$: neurônio que melhor casa com o vetor de entrada **x**– isto é, neurônio vencedor;
$\boldsymbol{x}$: vetor de entrada selecionado aleatoriamente do espaço de entrada com m-dimensão – representado, matematicamente, por $\boldsymbol{x} = [x_1, x_2, \cdots, x_m]^T$;
$\boldsymbol{w}_j$: vetor peso sináptico do neurônio j na grade com mesma dimensão do espaço de entrada– representado, matematicamente, por $\boldsymbol{w}_j = \left[w_{j1}, w_{j2}, \cdots, w_{jm}\right]^T$;
l: número total de neurônios na grade.

3ª etapa: cooperação entre neurônios. Nesta etapa, o "neurônio vencedor determina a localização espacial de uma vizinhança topológica de neurônios excitados [que com ele interagem]" (Haykin, 2001, p. 487). Conforme esse autor, a vizinhança topológica (h_{ij}, onde: i representa neurônio vencedor e j, neurônio excitado) é uma função unimodal da distância lateral entre neurônio vencedor e neurônio excitado (d_{ij}), sendo representada, tipicamente, pela função gaussiana ilustrada pela Figura 12 e definida conforme Eq. (7).

$$h_{j,i(\boldsymbol{x})}(t) = exp\left(-\frac{d_{j,i}^2}{2\sigma^2(t)}\right) \tag{7}$$

Na Eq. (7), tem-se:

$\sigma\,(t)$: representa a "largura efetiva" da vizinhança topológica no instante t. Esse parâmetro deve diminuir com o tempo a fim de satisfazer uma das condições da função de vizinhança topológica;
$d_{j,i}^2$: distância entre neurônio excitado j e neurônio vencedor i ao quadrado;
t: instante de tempo.

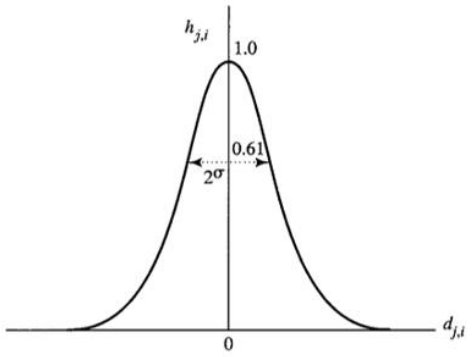

Figura 12- Função de Vizinhança Gaussiana
Fonte: Haykin (2011)

A "largura efetiva" da vizinhança topológica pode ser descrita por uma função de decaimento exponencial como descrito pela Eq. (8).

$$\sigma(t) = \sigma_0 \, exp\left(-\frac{t}{\tau_1}\right), \; t = 0,1,2,\cdots \tag{8}$$

Na Eq. (8), tem-se: σ_0 o valor de σ na inicialização do algoritmo SOM e τ_1 uma constante de tempo.

4ª etapa: adaptação entre neurônios. Nesta etapa, os "neurônios excitados [aumentam] seus valores individuais [da função de ativação] em relação ao padrão de entrada através de ajustes adequados aplicados a seus pesos, [conforme Eq. (9)]" (Haykin, 2001, p. 487–488). Tal etapa consiste em duas fases, a saber (Carvalho et al., 1998; Haykin, 2001; Kohonen, 1997): (i) Ordenação –consiste na ordenação topológica dos vetores peso, visando agrupar os padrões de entrada nos neurônios do mapa topológico, através de um processo que pode exigir 1000, ou mais, iterações do algoritmo SOM; (ii) Convergência –consiste no aprimoramento do agrupamento realizado, através de um processo que pode exigir um número de iterações de no mínimo 500 vezes o número de neurônios na rede com a taxa de aprendizagem igual ou próxima a 0,01 e raio da vizinhança do neurônio vencedor igual a 1 ou zero neurônios vizinhos.

$$\boldsymbol{w}_i(t+1) = \boldsymbol{w}_i(t) + \eta\,(t)\, h_{j,i(\boldsymbol{x})}(t)[\boldsymbol{x}(t) - \boldsymbol{w}_i(t)] \tag{9}$$

Na Eq. (9): $\eta\,(t)$ a taxa de aprendizagem do algoritmo.

A taxa de aprendizagem decresce gradualmente com o aumento do tempo. Em geral, opta-se pela função de decaimento exponencial dada pela Eq. (10).

$$\eta\,(t) = \eta_0 \, exp\left(-\frac{t}{\tau_2}\right), t = 0,1,2,\cdots \tag{10}$$

Na Eq. (10): τ_2 uma constante de tempo do algoritmo de aprendizagem do SOM.

O processo de treinamento é finalizado somente após o algoritmo conseguir convergir –isto é, até que não sejam observadas modificações significativas do mapa de características.

Ao finalizar o treinamento da rede SOM, tem-se uma grade de baixa dimensionalidade –conforme Carvalho et al. (1998, p. 112), "geralmente bidimensional, podendo algumas vezes ser unidimensional [e, muito raramente, tridimensional]"– constituída por neurônios. Conforme Vesanto et al. (2000), cada neurônio k é representado por um vetor peso m-dimensional (também denominado vetor protótipo ou vetor código) $w_k = [w_{k1}, \cdots, w_{km}]$, onde m é a dimensão do espaço de entrada. O vetor protótipo representa o neurônio mais próximo (neurônio vencedor, *the best-matching unit* –BMU) de cada vetor de entrada **x**. Os vetores protótipo BMU e seus vizinhos são atualizados através da Eq. (9) até os neurônios na grade se tornarem ordenados, isso ocorre quando neurônios vizinhos têm vetores protótipo similares.

A essência do agrupamento SOM é dividir o espaço de entrada em regiões compostas por unidades neurais cujos vetores protótipos representam todos os possíveis vetores que estiverem na sua vizinhança. A Figura 13 ilustra uma rede SOM gerada, na qual os neurônios da grade estão interligados por uma relação de vizinhança (topologia) hexagonal e rotulados com nome do vetor de entrada cuja distância do vetor protótipo foi a menor presente em cada unidade neural.

Figura 13- Representação de uma Rede SOM
Fonte: Vesanto et al.(2000)

É importante ressaltar, antes de mencionar sobre métodos de visualização, que embora o SOM combine as funções de um vetor de quantização– ao reduzir o espaço dimensional do vetor de entrada– e de um vetor de projeção– ao projetar os

dados do espaço de entrada de alta dimensão para o espaço de saída– é necessário adotar, adicionalmente, um método de projeção (por exemplo: Análise de Componentes Principais) para tornar a interpretação da estrutura global da grade mais fácil, visto que a projeção realizada pelo próprio SOM fornece somente a quantidade e a identificação de vetor de entrada por neurônio (Vesanto, 1999).

A interpretação gráfica da Rede SOM pode ser realizada através de diferentes tipos de métodos, os quais podem ser classificados em três categorias (Vesanto, 1999):

1ª categoria– métodos que fornecem uma ideia global e possível da estrutura de grupo. Nesse caso, tem-se: A Rede SOM, propriamente dita, representada por uma grade que é constituída por unidades neurais, cujo tamanho e estrutura são definidos a priori pelo usuário, nas quais as **x** observações m-dimensionais de entrada se encontram dispostas, conforme características de similaridade e dissimilaridade. Desta forma, a Rede SOM nos fornece a quantidade e a identificação do vetor de entrada por unidade neural, então, pode-se verificar que, regiões mais densas representam grupos de unidades neurais cujo vetor de atributos é similar para com seus vizinhos e, área menos densas, regiões de separação entre grupos.

2ª categoria– métodos de detecção de grupos. Nesse caso, tem-se:

- **U-matriz unificada:** matriz de distância unificada entre unidades vizinhas. A U-matriz tem dimensão $(2\mathrm{L}-1)\times(2\mathrm{C}-1)$, sendo L e C definidos pelo tamanho do mapa gerado. Esse método permite visualizar relações topológicas do mapa por meio de tonalidades de color, atribuída a cada unidade neural, conforme distância existente entre unidades vizinhas, de modo que, tons mais claros representem menor distância entre vetores de entrada e respectivo vetor protótipo e, tons mais escuros, maior distância entre esses vetores. Assim, grupos são detectados pelas delimitações dessas regiões.
- **Similaridade por cor:** matriz de distância unificada com atribuição de cor semelhante às unidades do mapa que possuem proximidade no espaço de entrada. Para isso, verifica-se a distância entre a projeção das **x** observações m-dimensionais de entrada sobre cada vetor protótipo do mapa gerado por meio da técnica de análise de componentes principais[13].

3ª categoria– métodos de análise das características dos grupos. Nesse caso, tem-se a **U-matriz por atributos**, matriz de distância por atributos entre unidades vizinhas. Esse método permite verificar a existência de correlação

[13] Técnica de sumarização de dados multidimensionais que transforma um conjunto $\mathbf{V} = \{\mathbf{v}_1, \dots, \mathbf{v}_N\}$ de vetores $\mathbf{v}_n = [\mathbf{v}_{n1}, \dots, \mathbf{v}_{nD}]^T \in \mathcal{R}^D, n = 1, \dots, N$, em uma dada base ortonormal e encontra uma nova base ortonormal $\{\boldsymbol{u}_1, \dots, \boldsymbol{u}_D\}$ capaz de gerar o espaço original, de modo que, ao optar por uma projeção de dados utilizando os P primeiros componentes principais, $P < D$, obtém-se uma representação do conjunto original em um espaço de menor dimensão.

entre os m-dimensionais atributos dos vetores de entrada **x,** de modo que, U-matrizes com padrões de coloração semelhantes indicam correlação positiva enquanto, padrões de coloração invertidos indicam correlação negativa.

Em relação à avaliação dos resultados do algoritmo SOM, tradicionalmente, a literatura utiliza duas métricas, a saber (Faria et al., 2010; Sassi, 2006; Vesanto et al., 2000):

- **Erro de quantização (EQ):** verifica o quão bem ajustados estão os vetores de entrada na rede. O erro de quantização é estimado pelas médias das distâncias entre cada vetor de dados **x** e o correspondente vetor protótipo do neurônio vencedor $\mathbf{w}_i$.
- **Erro topográfico (ET):** mede a capacidade do mapa em representar a topologia dos dados de entrada. O erro topográfico é calculado considerando para cada vetor de entrada **x** se seus primeiros e segundos neurônios vencedores são sempre vizinhos ou não, de modo que para cada vetor de entrada **x** seja atribuído valor 1 se não forem adjacentes e 0, caso contrário. Tal métrica é descrita em Kiviluoto (1996 apud; Vesanto et al., 2000).

A partir dos resultados obtidos pelas métricas, verifica-se que (Sassi, 2006):

- A medida EQ corresponde a resolução do mapa, que é inversamente proporcional ao número de neurônios na grade, isto é, erro diminui com aumento do número de neurônios;
- Se o número de neurônios for muito grande e/ou o raio da vizinhança tornar-se menor ou igual a 1 durante muito tempo, pode ocorrer de os neurônios se posicionarem sobre os objetos a serem representados. Nesse caso, EQ≅0 mas, a capacidade de representar a topologia dos dados é pedida e ET aumenta. O comportamento de ET nesta situação dependerá também do número de neurônios disponíveis no formato da vizinhança; ET aumenta se há poucos neurônios e diminui se há muitos neurônios.
- Caso EQ e ET sejam muitos baixos, a rede SOM, na tentativa de representar fidedignamente os dados, acaba perdendo sua capacidade de generalização e representando exatamente os dados, caracterizando o fenômeno de sobreajuste (*overfitting*).
- Em geral, valores muito baixos de ET associados a valores mais altos de EQ podem sugerir subajuste (*underfitting*). Este fenômeno ocorre quando um mapa é "rígido" demais, isso acontece quando há poucos neurônios para representar um número proporcionalmente grande de dados ou o raio de vizinhança final da função gaussiana for maior que 1 durante o treinamento.

2.5
Classificador Bayesiano Simples

O classificador Bayesiano simples (também conhecido como *naive* Bayes) é a estrutura mais simples de rede de crenças Bayesianas[14] usada na tarefa de classificação (Viaene et al., 2004).

O *naive* Bayes fundamenta-se no raciocínio probabilístico[15] de Thomas Bayes para quantificar e calcular a ocorrência de algum evento dado um conjunto de evidências (Han & Kamber, 2006). Thomas Bayes foi um matemático e teólogo inglês do século XVIII que formulou um teorema capaz de atualizar o grau de crença relacionado à ocorrência de determinado evento considerando um conjunto de informações já disponibilizadas (Coppin, 2004). Para compreender o pensamento Bayesiano, considere:

- A e B eventos definidos em um espaço amostral E, com chance de ocorrência determinada conforme Eq. (11) e Eq. (12) de modo que para cada evento seja atribuído um número real no intervalo [0, 1], tal que a probabilidade do espaço amostral inteiro seja 1.

$$\mathrm{P(A)} = \frac{\mathrm{n(A)}}{\mathrm{n(E)}} \tag{11}$$

$$\mathrm{P(B)} = \frac{\mathrm{n(B)}}{\mathrm{n(E)}} \tag{12}$$

  Onde:
  $\mathrm{n(A)}$= número de casos favoráveis ao evento A
  $\mathrm{n(B)}$= número de casos favoráveis ao evento B
  $\mathrm{n(E)}$= número de elementos do espaço amostral E.

- A probabilidade de dois eventos independentes A e B ocorrem simultaneamente é definida pela Eq. (13). No entanto, quando a ocorrência de um evento é modifica pelo outro, a probabilidade desses eventos ocorrerem simultaneamente é expressa conforme Eq. (14).

$$\mathrm{P(A \cap B)} = \mathrm{P(A)} \times \mathrm{P(B)} \tag{13}$$

$$\mathrm{P(A \cap B)} = \mathrm{P(A|B)} \times \mathrm{P(B)} = \mathrm{P(A)} \times \mathrm{P(B|A)} \tag{14}$$

  Onde:
  $\mathrm{P(A \cap B)} = \frac{\mathrm{n(A \cap B)}}{\mathrm{n(E)}}$, sendo $\mathrm{n(A \cap B)}$= número de casos $\mathrm{A \cap B}$
  $\mathrm{P(A|B)}$= probabilidade condicional de um evento A dado um evento B
  $\mathrm{P(B|A)}$= probabilidade condicional de um evento B dado um evento A

[14] Rede de crenças Bayesianas= grafo acíclico dirigido, no qual evidências (ou hipóteses) e relação de interdependência das hipóteses são representadas por nós e arcos, respectivamente.

[15] Raciocínio probabilístico= método de aquisição de conhecimento por indução que faz uso da teoria de probabilidades.

- Imagine que A seja um evento qualquer em E e $B_1, B_2, \dots, B_k$ uma partição do espaço amostral E. Para isso, considere a Eq. (15):

$$P(A) = P(A|B_1)\times P(B_1) + \cdots + P(A|B_k)\times P(B_k) \quad (15)$$

- Bayes relaciona a inferência racional à experiência empírica e à subjetividade das visões prévias rearranjando a Eq. (14) e considerando a Eq. (15), obtendo Eq. (16).

$$P(B|A) = \frac{P(A|B)\times P(B)}{P(A)} \quad (16)$$

Onde:
P(B|A)= probabilidade a posteriori
P(A|B)= probabilidade condicional obtida pela experiência empírica
$P(A)$ e $P(B)$ = probabilidades a priori

O *naive* Bayes determina em que classe[16] c_i um vetor $\mathbf{x}$ pertence, considerando a máxima probabilidade a posteriori obtida. Assim, para identificar a melhor classificação para um exemplo[17] particular $\mathbf{x} = (x_1, \dots, x_n)$, calcula-se a probabilidade a posteriori de cada possível classe c_i, conforme Eq. (17):

$$P(c_i|\mathbf{x}) = \frac{P(\mathbf{x}|c_i)\times P(c_i)}{P(\mathbf{x})} \quad (17)$$

No entanto, visando simplificar a Eq. (17), o *naive* Bayes considera $P(\mathbf{x})$ uma constante e assume que os valores dos atributos são condicionalmente independentes uns dos outros (Han & Kamber, 2006). Assim, a probabilidade a posteriori é obtida conforme Eq. (18):

$$P(c_i|\mathbf{x}) = P(x_1|c_i)\times P(x_2|c_i)\times \cdots \times P(x_n|c_i)\times P(c_i) \quad (18)$$

As probabilidades $P(x_k|c_i)$ são obtidas conforme Eq. (19):

$$P(x_k|c_i) = \frac{P(x_k \cap c_i)}{P(c_i)} \quad (19)$$

Onde: x_k é o valor do atributo A_k para vetor $\mathbf{x}$

Conforme Han & Kamber (2006), casos em que probabilidades $P(x_k|c_i)$ seja zero devem ser tratados com a correção de Laplace, definida por Eq. (20):

[16]Classe= rótulo atribuído aos exemplos a fim de caracterizá-los quanto a algum fenômeno de interesse.

[17]Exemplo= também denominado caso, registro ou dado na literatura, trata-se de uma tupla de valores de atributos (ou um vetor de valores de atributos) que descreve o objeto de interesse.

$$p(\mathrm{x_k}|\mathrm{c_i}) = \frac{a+1}{b+\mathrm{m}} \quad (20)$$

Onde:
a=número de ocorrência $\mathrm{x_k}$ na classe $\mathrm{c_i}$
b= número de exemplos da base treino na qual é dada a classificação $\mathrm{c_i}$
m=número de classes $\mathrm{c_i}$

A correção de Laplace leva em consideração que a base treino é tão grande que adicionar um a cada contagem de que precisamos será insignificante no valor da probabilidade estimada e convenientemente evita caso de valores de probabilidade zero (Han & Kamber, 2006).

Em relação a validação, existem diversas maneiras de fazê-la (Bramer, 2007; Han & Kamber, 2006) mas, comumente, usa-se a métrica acurácia –proporção de acertos do classificador. O *naive* Bayes tem apresentado bom desempenho em inúmeras aplicações, em especial naquelas em que os atributos não são fortemente correlacionados (Cheng & Greiner, 2001).

3
Sistema Inteligente Híbrido para Classificação do Perfil de Pagamento dos Consumidores Não-Residenciais de Energia Elétrica

Este capítulo é dedicado a formulação de um sistema que permita às empresas distribuidoras de energia elétrica classificar perfil de pagamento dos consumidores não-residenciais que pertencem a sua área de concessão. Para isso, considerou-se:

- Informações relacionadas ao cadastro, ao consumo, ao faturamento e à arrecadação dos consumidores que estão presentes na base de dados das empresas distribuidoras;
- Informações relativas à natureza jurídica desses consumidores disponibilizadas para consulta em *site* governamental;
- A intensidade energética que cada atividade econômica demanda para gerar um dólar na economia brasileira. Esta informação está disponível em *site* governamental do setor elétrico;
- Estimativa da participação da conta de energia elétrica no orçamento desses consumidores, considerando: dados de consumo e de faturamento dos consumidores, bem como, Produto Interno Bruto (PIB) do setor e consumo de energia elétrica final por setor a nível Brasil disponibilizados pela Empresa de Pesquisa Energética (2013);
- Técnicas *soft computing* de mineração de dados descritas detalhadamente nas seções 2.4 e 2.5 para segmentar mercado consumidor.

Um fluxograma com a formulação do sistema proposto é ilustrado na Figura 14. Tal sistema é composto por módulos independentes e intercomunicativos que usam técnicas *soft computing* distintas para obter uma classificação do perfil de pagamento dos consumidores, a saber:

Neural: aloca consumidores que apresentam similaridades em termos de valor da fatura, do consumo, da demanda, da intensidade energética e do peso da conta em uma mesma unidade neural ou em unidades neurais vizinhas.

Bayesiano: classifica perfil de pagamento dos consumidores, levando em consideração unidade neural na qual cada consumidor se encontra alocado e variáveis qualitativas (tais como: atividade econômica, estrutura tarifária, mesorregião, natureza jurídica e porte empresarial) que os caracterizam.

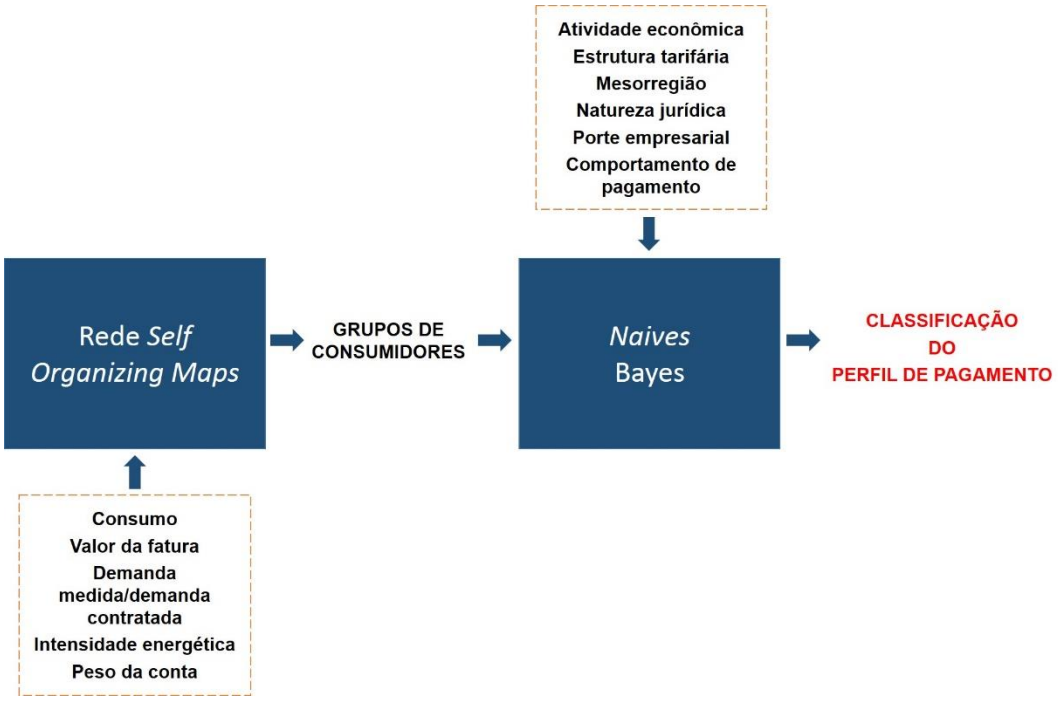

Figura 14- Sistema Inteligente Híbrido Proposto
Fonte: Elaboração Própria

A implementação do sistema inteligente híbrido deve ser inicializada somente após seleção, pré-processamento e transformação da base de dados da empresa distribuidora de energia elétrica selecionada para análise. Nesta etapa inicial:

- Seleciona-se universo e variáveis de pesquisa;
- Verifica-se ocorrência de dados faltantes no histórico dos consumidores;
- Corrige-se inconsistências relacionadas às informações cadastrais dos consumidores;
- Desconsidera consumidores com histórico de dados inferiores a 24 meses;
- Desconsidera consumidores que apresenta dados faltantes superiores a 10% nos históricos de consumo, de valor da fatura ou de demanda, dados faltantes;
- Cria-se a variável demanda medida/demanda contratada, considerando histórico de demanda fornecido;
- Adota-se como valor de entrada do módulo neural do sistema a mediana dos valores de consumo, de valor da fatura e de demanda medida/demanda contratada. A escolha da mediana se deve ao fato dela ser uma medida de tendência central não sensível a valores extremos que podem estar presentes no histórico dos consumidores;
- Verifica-se a frequência de atrasos de cada consumidor no histórico de arrecadação e cria-se a variável dias de atraso, considerando a mediana da frequência dos atrasos;

- Agrega informações de consumo, de valor da fatura e de demanda medida/demanda contratada dos consumidores que possuem várias instalações, de modo que: **(i)** registro do consumo corresponda a soma dos consumos das instalações; **(ii)** registro do valor da fatura, a soma dos valores das faturas; **(iii)** registro da demanda medida/demanda contratada corresponda a mediana desta razão e **(iv)** registro de atraso corresponda ao máximo de atraso verificado nas instalações. Em relação aos registros qualitativos, considera-se o mais frequente entre as instalações;
- Cria-se a variável renda estimada do consumidor (R$), considerando razão entre consumo anual (kWh) e intensidade energética (kWh/US$) do setor no qual o consumidor pertence. Após obter a razão, faz-se a conversão da moeda dólar para a moeda real, considerando valor médio da moeda dólar em relação à moeda real no período usado para obter a estimativa da renda;
- Cria-se a variável peso da conta, considerando a razão entre mediana do valor da fatura de energia (R$) e renda estimada do consumidor (R$);
- Cria-se a variável comportamento de pagamento, considerando a frequência de atrasos de cada consumidor no histórico de arrecadação. Consumidor que paga no vencimento ou antecipado é adimplente, caso contrário, inadimplente;
- Integraliza-se base de dados da pesquisa, considerando dados pré-processados advindos da base de dados da empresa distribuidora em análise e dados obtidos em fontes governamentais;
- Analisa-se comportamento de consumo, de valor da fatura, de demanda e de atraso dos consumidores por meio de técnicas de sumarização. Além disso, verifica-se a relação entre comportamento de pagamento e atividade econômica, bem como a relação entre peso da conta e intensidade energética. Aplicam-se transformações logarítmicas, inversas ou outras descritas na seção 2.2.2.1, caso os dados apresentem assimetria e curtose acentuadas.
- Normaliza-se as variáveis quantitativas da base de dados da pesquisa usando a normalização *range*, o que permitirá que os valores destas variáveis tenham igual influência na obtenção dos resultados;
- Separa-se a base de dados da pesquisa em dois conjuntos (conjunto treino e conjunto teste) por meio do método de amostragem aleatória simples.

O sistema inteligente híbrido deve ser implementado computacionalmente por meio de *software* livre ou comercial.

4
Implementação do Sistema Inteligente Híbrido

A implementação deste sistema constituído por módulos ocorreu em um *ultrabook* com sistema operacional de 64 *bits*, processador Intel(R) Core (TM) i5-3317U de 1,70 GHz e 4,00 GB de memória RAM. É importante ressaltar que, os procedimentos relacionados à seleção, ao pré-processamento e à caracterização da amostra de pesquisa foram realizados por meio de um sistema de gerenciamento de banco de dados e de uma planilha eletrônica. Os módulos neural e bayesiano foram implementados, respectivamente, no *Software* MatLab[18] usando o Toolbox SOM[19] e em uma planilha eletrônica.

4.1
Universo e Amostra de Pesquisa

A base de dados utilizada nesta pesquisa foi extraída de uma empresa de distribuição de energia elétrica brasileira que atende cerca de 2,8 milhões de consumidores residenciais, comerciais e industriais, em mais de 60 municípios, cobrindo uma área aproximada de 30.000 km^2. Tal distribuidora faz parte de uma holding controlada por um grupo italiano que atua nos demais segmentos do setor elétrico brasileiro. No Quadro 3, constam a relação das variáveis disponibilizadas pela diretoria comercial desta distribuidora de energia elétrica.

O estudo foi realizado considerando somente consumidores cativos não-residenciais pertencentes ao grupo A. Os comportamentos de consumo, de valor fatura, de demanda e de atraso desses consumidores foram observados no período jan./2009 a dez./2010. Antes da implementação da metodologia proposta, os procedimentos de limpeza, de redução e de transformação dos dados foram realizados na base de dados disponibilizada pela distribuidora.

[18]MatLab= *software* comercial desenvolvido pela MathWorks amplamente conhecido por acadêmicos e profissionais de diversas áreas para resolver problemas científicos e de engenharia

[19]Toolbox SOM= biblioteca de rotinas para implementação das redes SOM no MatLab. Este Toolbox foi desenvolvido pelo Laboratório de Computação e Ciências da Informação da Universidade de Tecnologia de Helsinki – Finlândia. É uma ferramenta de domínio público e pode ser encontrada para instalação no endereço: www.cis.hut.fi/projects/somtoolbox/. Existem outras bibliotecas de rotinas disponíveis para implementação das redes SOM, no entanto esta foi escolhida devido aos recursos gráficos que possui.

Quadro 3- Relação de Variáveis Disponibilizadas

Variável	Base de dados
Número unidade consumidora	Cadastro/ Faturamento e Arrecadação/ Histórico Consumo e Demanda
Ano Mês referência faturamento	Faturamento e Arrecadação/ Histórico consumo e demanda
Nome consumidor	Cadastro
Município unidade consumidora	
Porte unidade consumidora	
Data de compensação	Faturamento e Arrecadação
Data de pagamento	
Data de vencimento	
Valor fatura	
Ano Mês vencimento	Histórico consumo e demanda
Tensão	
Nome setor	
Demanda Contratada Fora Ponta	
Demanda Contratada Ponta	
Demanda Medida Fora Ponta	
Demanda Medida Ponta	
Consumo Ativo Fora Ponta	
Consumo Ativo Ponta	

Fonte: Elaboração própria

A base de dados da pesquisa foi estruturada conforme Figura 15.

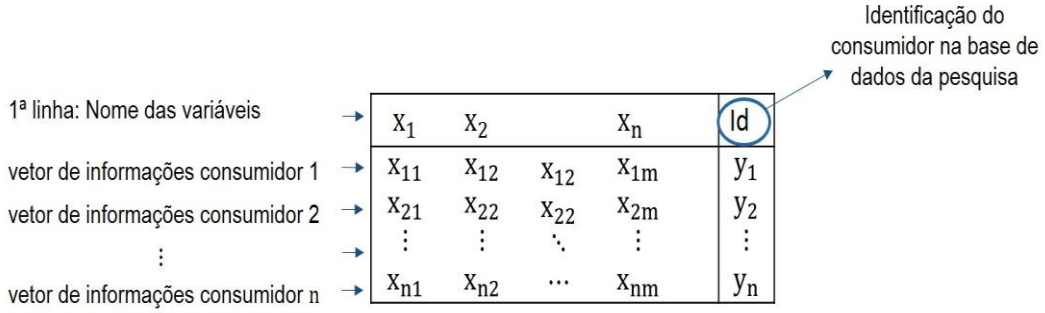

Figura 15- Estrutura da Base de Dados da Pesquisa
Fonte: Elaboração própria

As variáveis usadas na implementação do sistema proposto constam definidas no Apêndice A.

4.2
Análise Exploratória e Pré-processamento da Base de Dados da Pesquisa

O pré-processamento dos dados iniciou-se com o processo de limpeza que consistiu, primeiramente, em verificar e analisar a quantidade de dados perdidos por consumidores. Optou-se por não aplicar método de substituição de dados

perdidos e implementar base de dados com todos os dados válidos possíveis, desde que os dados perdidos fossem aleatórios e correspondessem em até 10% dos dados do histórico de consumo, de valor da fatura e da demanda de cada consumidor. Isso pois, a substituição poderia gerar viés aos dados; além disso, a mediana dessas variáveis é que serão usadas para analisar comportamento dos dados e implementar sistema proposto. Ainda nesse processo, corrigiu-se as inconsistências relacionadas às informações cadastrais dos consumidores e desconsiderou da base de dados da pesquisa os consumidores com histórico de dados inferiores a 24 meses e, ainda, aqueles que apresentavam dados faltantes superiores a 10% nos históricos de consumo, de valor da fatura ou de demanda.

No processo de transformação, o histórico de consumo, de valor da fatura e da demanda de cada consumidor foram sintetizados e representados pela mediana desses valores. Em seguida, criou-se novas variáveis considerando as variáveis disponibilizadas na base de dados da pesquisa e as informações relativas à natureza jurídica dos consumidores e à intensidade energética de cada atividade econômica que estão disponibilizadas em *site* governamental. Posteriormente, agregou-se consumidores cujo Cadastro Nacional de Pessoas Jurídicas possui mesma raiz, então integralizou-se a base de dados da pesquisa.

Então, após processo de limpeza e transformação da base de dados da pesquisa, analisou-se o comportamento de consumo, de valor da fatura, de demanda e de atraso dos consumidores por meio das técnicas de sumarização.

A Figura 16 apresenta como a variável consumo se encontra distribuída na base de dados da pesquisa. Por ela, pode-se verificar assimetria positiva da distribuição de dados do consumo, o que significa que a dispersão dos valores de consumo é desigual ao longo da base de pesquisa. Devido a esse motivo, aplicou-se uma transformação logarítmica nos dados do consumo, reduzindo consideravelmente a assimetria dessa distribuição (veja Figura 17).

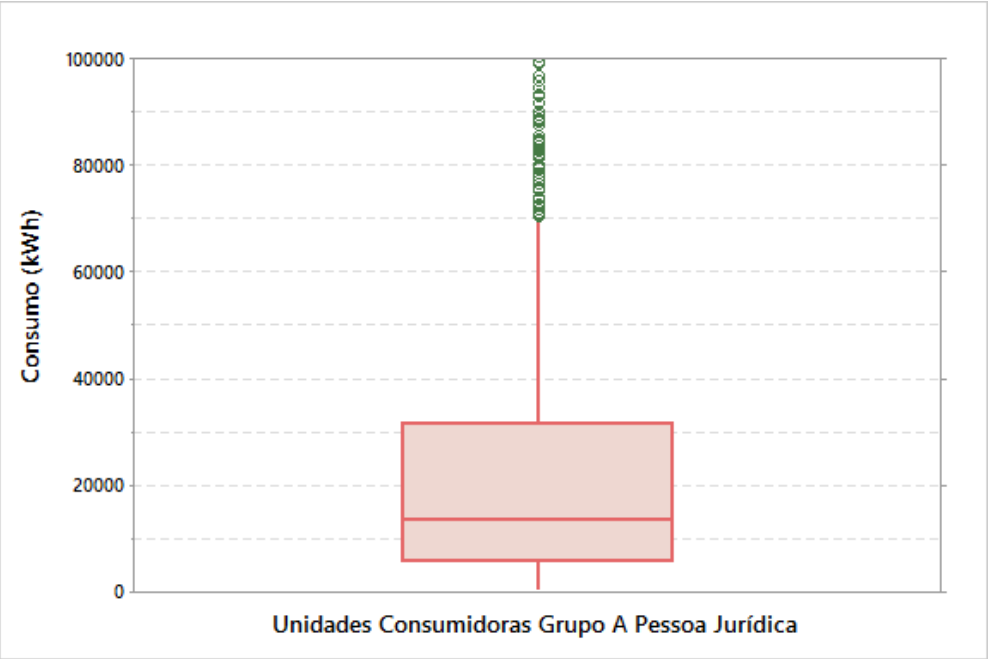

Figura 16- Diagrama de Caixa da Variável Consumo
Fonte: Elaboração própria

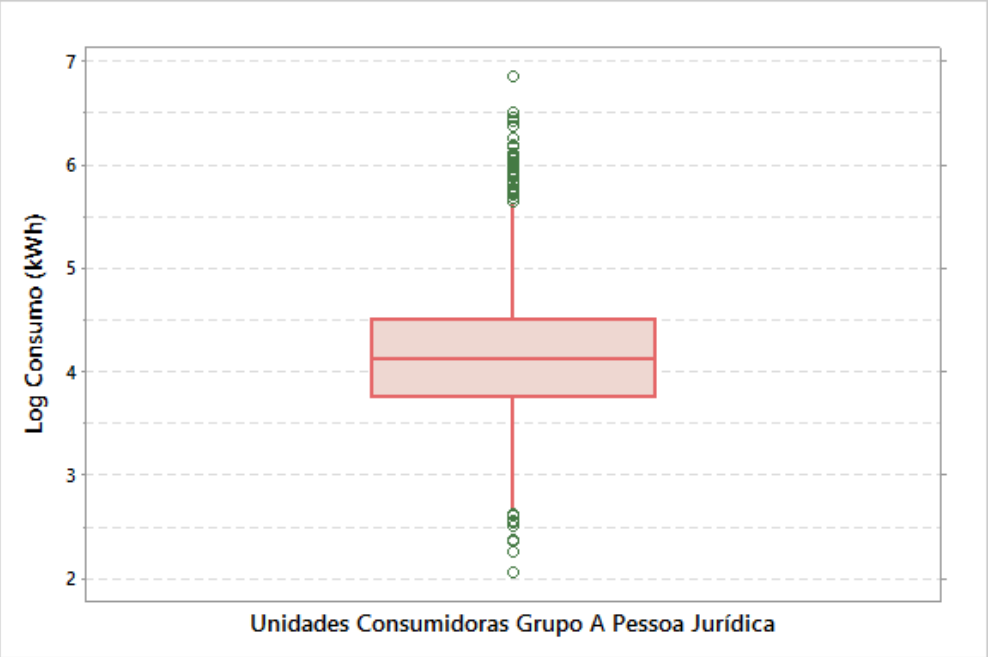

Figura 17- Diagrama de Caixa da Variável Log Consumo
Fonte: Elaboração própria

Na Figura 18, verificou-se a distribuição dos dados de valor da fatura na base de dados da pesquisa. Por ela, pode-se verificar assimetria positiva da distribuição de dados do valor fatura. Por isso, aplicou-se uma transformação logarítmica nos dados do consumo, o que tornou a distribuição consideravelmente simétrica (veja Figura 19).

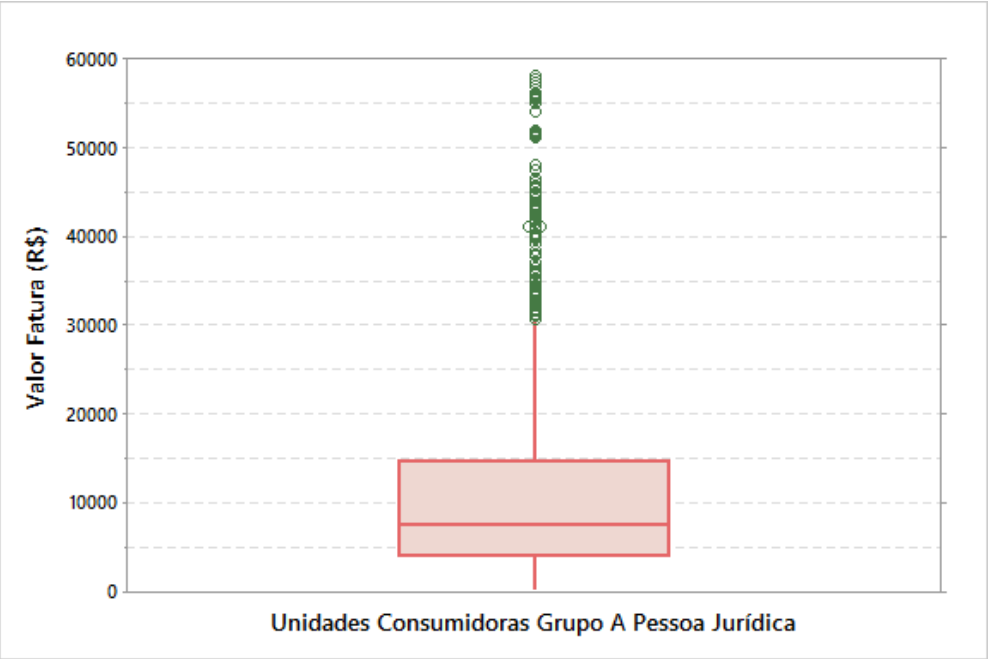

Figura 18- Diagrama de Caixa da Variável Valor Fatura
Fonte: Elaboração própria

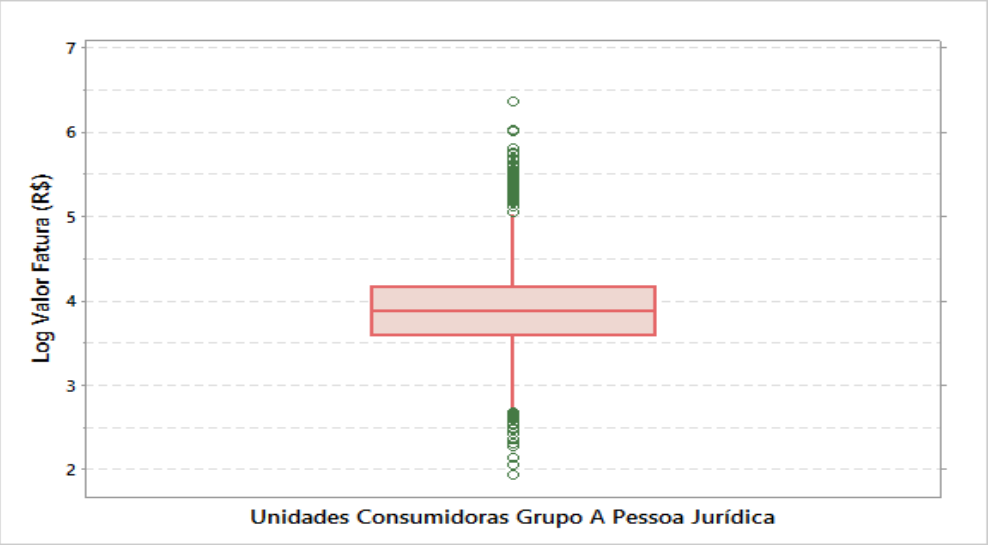

Figura 19- Diagrama de Caixa da Variável Log Valor Fatura
Fonte: Elaboração própria

A Figura 20 apresenta a distribuição de dados da razão Demanda Medida (DM) por Demanda Contratada (DC) na base de dados da pesquisa. Por ela, pode-se verificar assimetria negativa da distribuição de dados inerentes a essa razão. Devido a esse motivo, algumas transformações (como: quadrado, cubo e raiz quadrada) foram aplicadas. No entanto, todas essas transformações geraram um grande número de valores extremos e ainda contribuíram para o aumento da

assimetria na distribuição dos dados. Assim, optou-se por deixar a distribuição original desses dados.

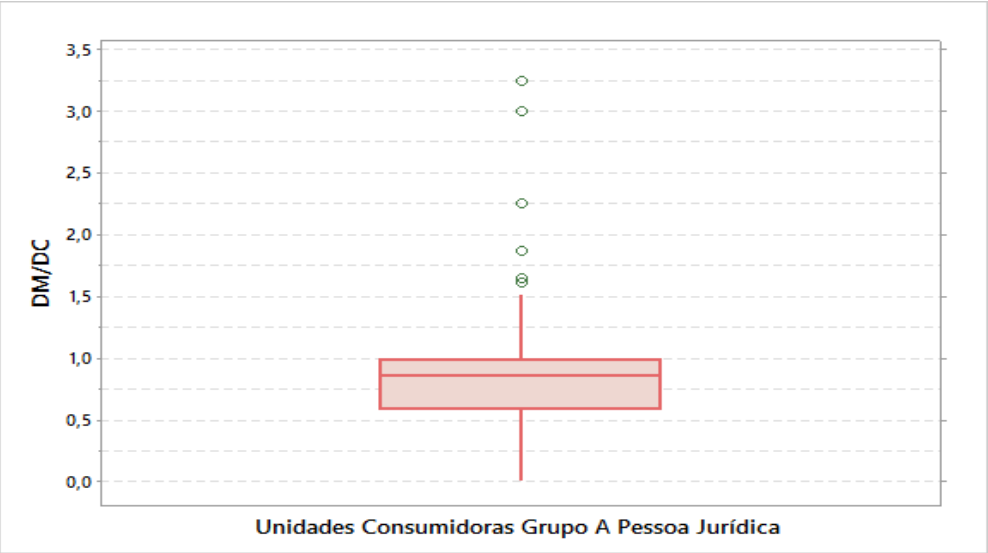

Figura 20- Diagrama de Caixa da Variável Demanda Medida/Demanda Contratada
Fonte: Elaboração própria

Em relação a distribuição dos dados das variáveis peso da conta e intensidade energética, verificou-se a existência de uma assimetria positiva nessas distribuições, assim, aplicou-se uma transformação logarítmica nesses dados, o que minimizou essa característica, embora não tenha sido eliminada.

A Tabela 2 apresenta as principais estatísticas da base de dados da pesquisa.

Tabela 2-Estatísticas Descritivas da Base de Dados da Pesquisa

Variável / Estatísticas	Log Consumo	Log Valor Fatura	DM/DC	Log IE	Log Peso Conta
Mínimo	2,0645	1,9345	0,0000	-1,6990	-3,5992
1º quartil	3,7501	3,5861	0,5825	-1,0970	-2,7441
Mediana	4,1250	3,8723	0,8600	-1,0970	-2,6360
3º quartil	4,4965	4,1634	0,9800	-0,7959	-2,3837
Máximo	6,8529	6,3720	3,25	0,5198	-0,8506
Média	4,1445	3,8890	0,72783	-0,9125	-2,5088
Desvio Padrão	0,5978	0,5251	0,37896	0,4127	0,4252
Assimetria	0,43	0,46	-0,59	1,39	1,10

Fonte: Elaboração própria

A Figura 21 ilustra a predominância de consumidores inadimplentes na base de dados da pesquisa, no entanto, muitos deles apresentam comportamento de pagamento variável (isto é, antecipam, pagam no vencimento e atrasam). Observou-se, pela Figura 22, que cerca de 50% das unidades consumidoras inadimplentes

tiveram entre 1 a 4 eventos de atraso no histórico, em torno de 20%, entre 5 a 10 eventos de atraso e, aproximadamente 30%, entre 11 e 24 eventos de atraso. Assim, não é recomendável descrever o comportamento de atraso dos consumidores por uma medida de locação.

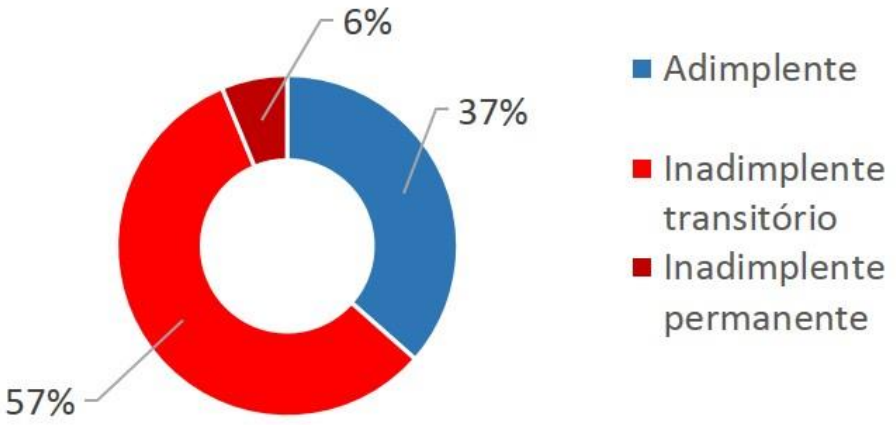

Figura 21- Comportamento de Pagamento Base de Dados da Pesquisa
Fonte: Elaboração própria

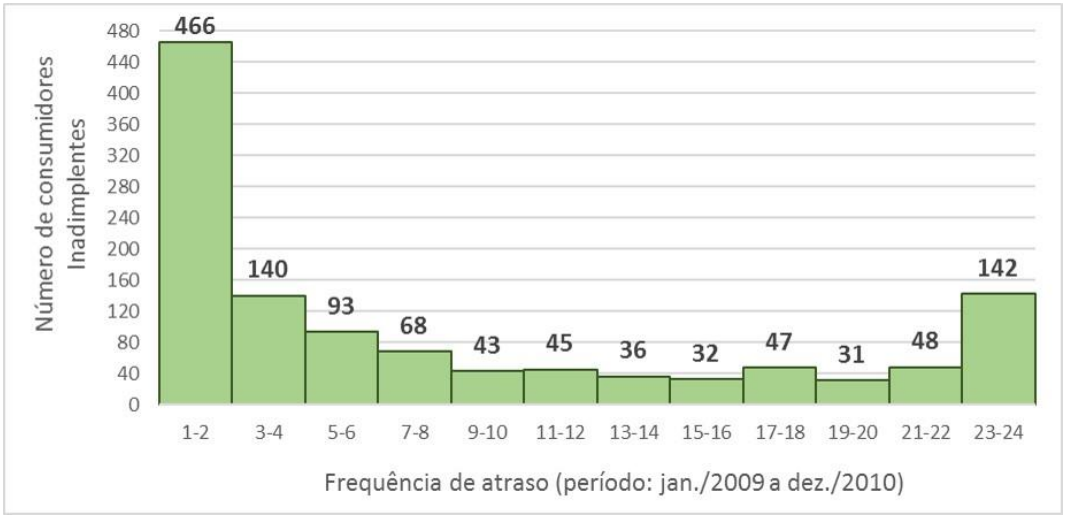

Figura 22- Frequência de Atrasos Base Pesquisa
Fonte: Elaboração própria

A Figura 23 informa que consumidores com maiores frequências de atraso tendem a pagar suas faturas muitos dias após a data de vencimento.

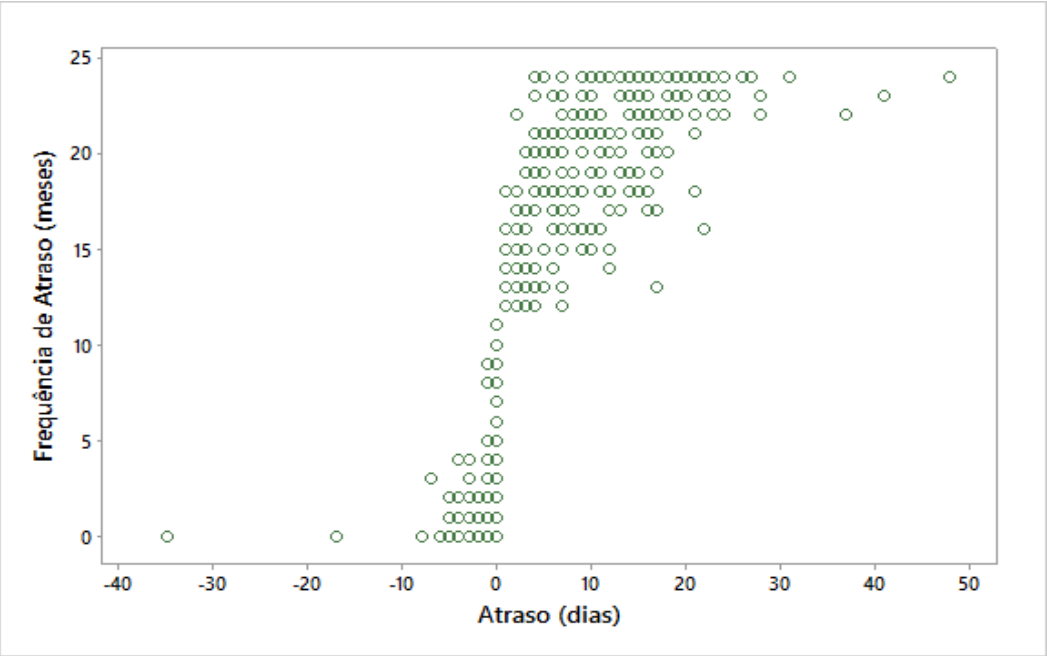

Figura 23- Diagrama de Dispersão das Variáveis Atraso e Frequência de Atraso
Fonte: Elaboração própria

Uma matriz de dispersão foi construída para verificar a relação estabelecida entre as variáveis consumo, valor da fatura e demanda medida/demanda contratada, veja Figura 24. Para isso, foram utilizadas essas variáveis em seu formato original (isto é, sem a transformação logarítmica) pois, as transformações poderiam mudar a interpretação das variáveis.

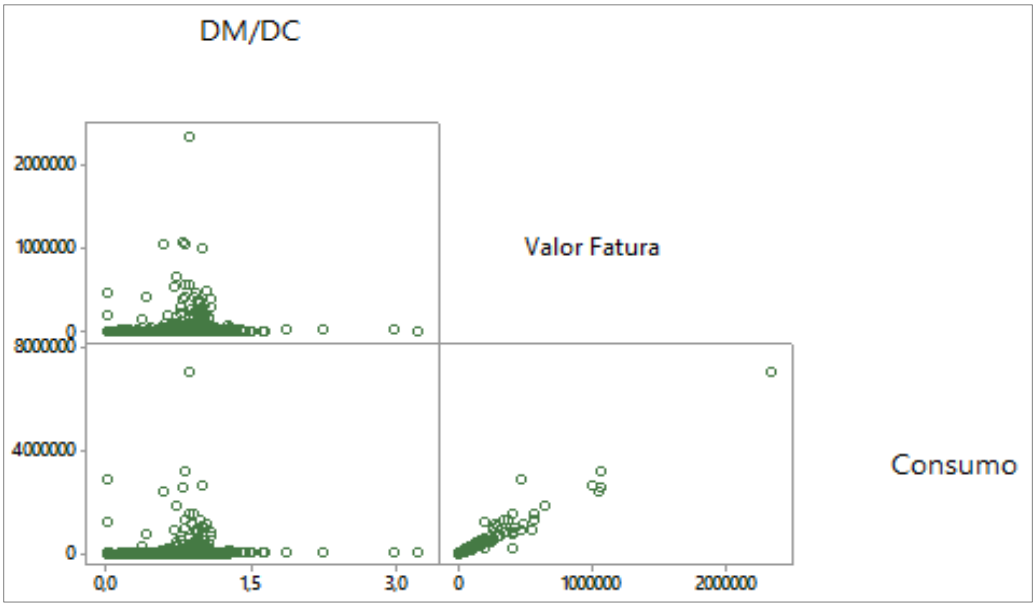

Figura 24- Matriz de Dispersão das Variáveis Consumo, Demanda Medida/Demanda Contratada e Valor Fatura
Fonte: Elaboração própria

Note que, o valor fatura e consumo são altamente correlacionados e um maior controle da razão Demanda Medida/Demanda Contratada foi observado à medida que o valor fatura aumentou, veja Figura 25.

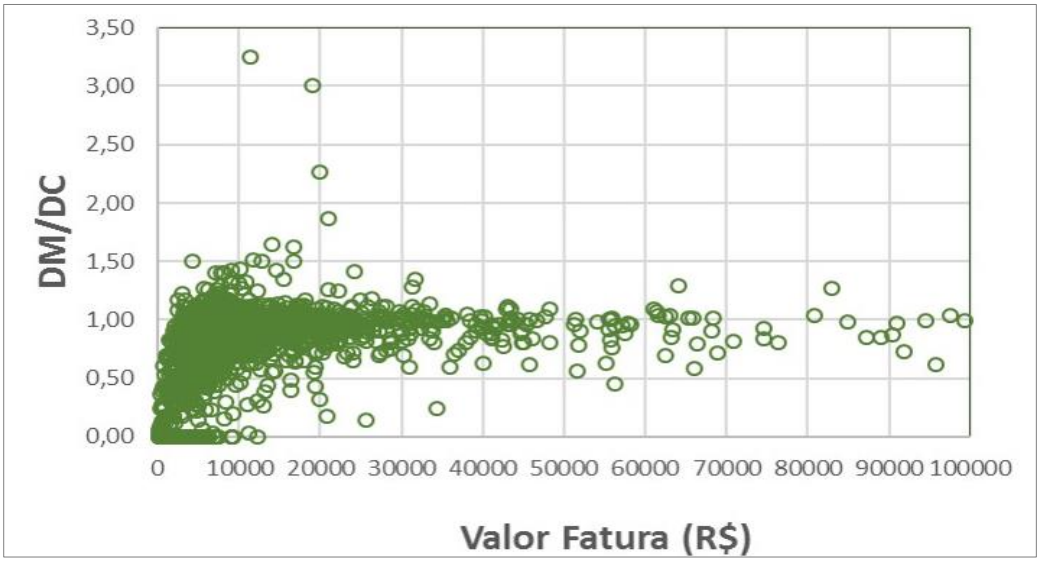

Figura 25- Diagrama de Dispersão das Variáveis Valor Fatura e Demanda Medidas/Demanda Contratada
Fonte: Elaboração própria

Outra relação verificada foi dias de atraso e valor da fatura, veja Figura 26. Note que, consumidores com valor fatura alto tem grandes possibilidades de serem adimplentes. A Figura 27 permite uma visualização melhor dessa relação, veja que consumidores com valor fatura abaixo de R$ 20.000,00 tem maiores possibilidades de serem inadimplentes.

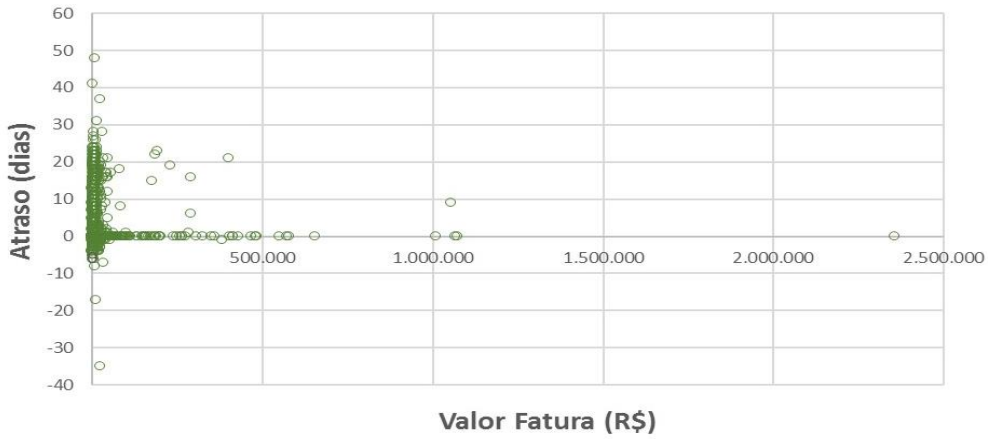

Figura 26- Diagrama de Dispersão entre Atraso e Valor Fatura
Fonte: Elaboração própria

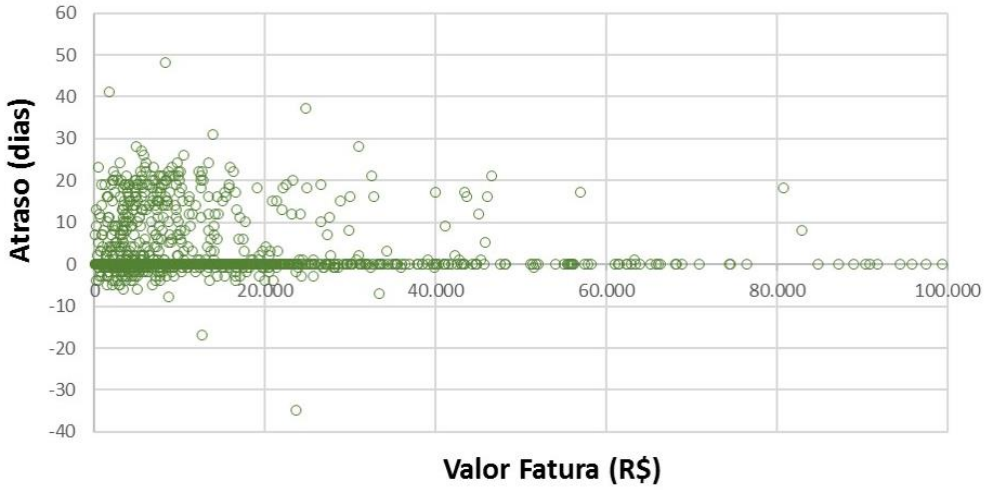

Figura 27- Diagrama de Dispersão entre Atraso e Valor Fatura (ampliado)
Fonte: Elaboração própria

Um fator interessante verificado pela análise de dados é que não há uma atividade econômica em si que detém maiores valores de faturas (acima de R$ 40.000,00). Todas as atividades econômicas têm pelo menos um consumidor que possui valor elevado de fatura, exceto a atividade econômica Indústria Metalúrgica (consumidores têm valor da fatura abaixo de sete mil reais). Adotando essas configurações, verificou-se que cento e quarenta e um consumidores constam na relação de maior valor da fatura. Dentre as atividades econômicas que predominam maior valor da fatura estão: Comércio e Público (com 70 consumidores), Outras Indústrias (com 28 consumidores) e Indústria de alimentos e bebidas (com 16 consumidores).

Por fim, antes de implementar o sistema inteligente híbrido, as variáveis quantitativas (demanda medida/demanda contratada, log consumo, log valor da fatura, log peso da conta e log intensidade energética) foram normalizadas aplicando a técnica de normalização *range* e a base de dados da pesquisa foi dividida em conjunto treino e teste pelo método de amostragem aleatória simples. O estudo foi conduzido com 1.876 consumidores, dos quais 1.600 formaram o conjunto treino.

4.3
Resultados da Implementação do Sistema Inteligente Híbrido

O módulo neural foi inicializado com o conjunto treino, composto pelas variáveis quantitativas supracitadas de 1.600 consumidores. Então, diferentes configurações para geração da rede SOM foram testadas:

- **Formato da vizinhança:** hexagonal e retangular;
- **Função de vizinhança:** gaussiana (ver seção 2.4);
- **Algoritmo de treinamento:** sequencial (nessa opção, o toolbox permite escolher o valor da taxa de aprendizagem inicial e função de aprendizagem);
- **Algoritmo de inicialização:** aleatório e linear;
- **Função taxa de aprendizagem:** exponencial (ver seção 2.4);
- **Taxa de aprendizagem inicial:** 0,1 (ver seção 2.4);
- **Raio final de vizinhança:** 1 (ver seção 2.4)
- **Tamanho de rede:** considerou-se uma fórmula heurística implementada pelo Toolbox SOM como sendo a rede de maior tamanho e, a partir dela, determinou-se outros tamanhos de rede. Conforme Faria et al. (Faria et al., 2010), a fórmula heurística para determinação das dimensões da rede é dada pela raiz quadrada da razão dos dois maiores autovalores do conjunto de dados, de modo que, o seu produto seja próximo ao número desejado de neurônios da rede, também definido por heurística, que corresponde a cinco vezes a raiz quadrada do tamanho da amostra de dados.

No Apêndice B, encontram-se as rotinas do MatLab para implementação da rede SOM.

A Tabela 3 apresenta os principais resultados de cada configuração de rede testada. Através dela, verificou-se o melhor ajuste –isto é, melhor resultado em relação a: tempo de treinamento, erro de quantização e erro topológico– em cada resolução de rede:

- Na rede SOM 8x2, formato de vizinhança hexagonal, ocorreu o fenômeno de subajuste, que pode ser justificado pela presença de poucos neurônios para representar um número proporcionalmente grande de dados. No entanto, nesse tamanho de rede, o formato de

vizinhança retangular apresenta um ajuste relativamente bom, sendo a melhor configuração desse tamanho de mapa obtida com o algoritmo de inicialização linear;

- Nas redes SOM de maiores resoluções, os melhores ajustes ocorreram com formato de vizinhança hexagonal e algoritmo de inicialização linear. Em mapas com resoluções maiores, a capacidade de representar a topologia dos dados é perdida com o formato de vizinhança retangular, pois o erro topológico aumentou;
- O erro de quantização decresceu, consideravelmente, da rede 8x2 para 10x4 para 12x8. Nas redes 14x10 e 16x12, esse decréscimo continuou, mas de forma sutil;
- O erro médio das redes 12x8, 14x10 e 16x12 apresenta diferenças mínimas;
- O tempo de treinamento aumenta, consideravelmente, à medida que o tamanho da rede aumenta.

Tabela 3- Configurações de Redes Testadas

Parâmetros / **Tamanho de rede**	Formato de vizinhança	Algoritmo de inicialização	Raio inicial	Tempo de treinamento	Erro Quantização (QE)	Erro Topológico (TE)
8x2	Hexagonal	Aleatório	1	797 s	0,1218	0,0456
	Retangular	Aleatório		803 s	0,1236	0,0831
	Hexagonal	Linear		796 s	0,1240	0,0525
	Retangular	Linear		811 s	0,1222	0,0713
10x4	Hexagonal	Aleatório	1	1.947 s	0,0935	0,0463
	Retangular	Aleatório		1.956 s	0,0911	0,0931
	Hexagonal	Linear		1.935 s	0,0949	0,0256
	Retangular	Linear		1.933 s	0,0916	0,0925
12x8	Hexagonal	Aleatório	3	4.923 s	0,0672	0,0306
	Retangular	Aleatório		4.934 s	0,0646	0,1094
	Hexagonal	Linear		4.762 s	0,0665	0,0206
	Retangular	Linear		4.648 s	0,0658	0,1181
14x10	Hexagonal	Aleatório	4	7.634 s	0,0599	0,0194
	Retangular	Aleatório		8.258 s	0,0565	0,1006
	Hexagonal	Linear		7.870 s	0,0582	0,0306
	Retangular	Linear		7.799 s	0,0565	0,0850

Tabela 3- Configurações de Redes Testadas (continuação)

Parâmetros / Tamanho de rede	Formato de vizinhança	Algoritmo de inicialização	Raio inicial	Tempo de treinamento	Erro Quantização (QE)	Erro Topológico (TE)
16x12	Hexagonal	Aleatório	5	12.278 s	0,0534	0,0188
	Retangular	Aleatório		11.164 s	0,0503	0,1069
	Hexagonal	Linear		12.102 s	0,0533	0,0263
	Retangular	Linear		12.170 s	0,0504	0,1031

Fonte: Elaboração própria

Assim, considerando o supracitado e que para o problema de pesquisa o interessante é detectar grandes desvios entre as unidades da rede, concluiu-se que a rede 12x8 com formato hexagonal e inicialização linear é a que melhor representa os dados do espaço de entrada nas unidades neurais, de modo que, dados similares no espaço de entrada sejam mapeados na mesma unidade neural ou em unidades vizinhas.

A Figura 28 apresenta a matriz de distância unificada (U-matriz) e as matrizes de distância por atributos para a configuração de rede adotada. Considerando o espectro de cores que atribui distância entre vetores de entrada e vetor protótipo, verificou-se:

- Na matriz de distância unificada, a presença de cinco possíveis grupos distintos;
- A forte correlação positiva entre as variáveis valor da fatura e consumo, assim como ocorre com os atributos intensidade energética e peso da conta;
- Características comuns entre consumidores de cada grupo:

  Grupo 1- consumidores apresentam valor da fatura e consumo mediano, intensidade energética e peso da conta baixo. Em relação a demanda, apresentam comportamento variável;

  Grupo 2- consumidores apresentam valor fatura e consumo baixo, com controle de demanda baixo a moderado baixo, intensidade energética e peso da conta moderado baixo;

  Grupo 3- consumidores apresentam intensidade energética e peso da conta moderado baixo e comportamento bastante heterogêneo em relação ao valor da fatura, consumo e demanda;

  Grupo 4- consumidores apresentam valor da fatura e consumo moderado alto a alto, intensidade energética e peso da conta mediano a moderado alto. Em relação a demanda, apresentam comportamento variável;

Grupo 5- consumidores apresentam valor fatura e consumo mediano a moderado alto, intensidade energética e peso da conta alto. Em relação a demanda, apresentam comportamento variável.

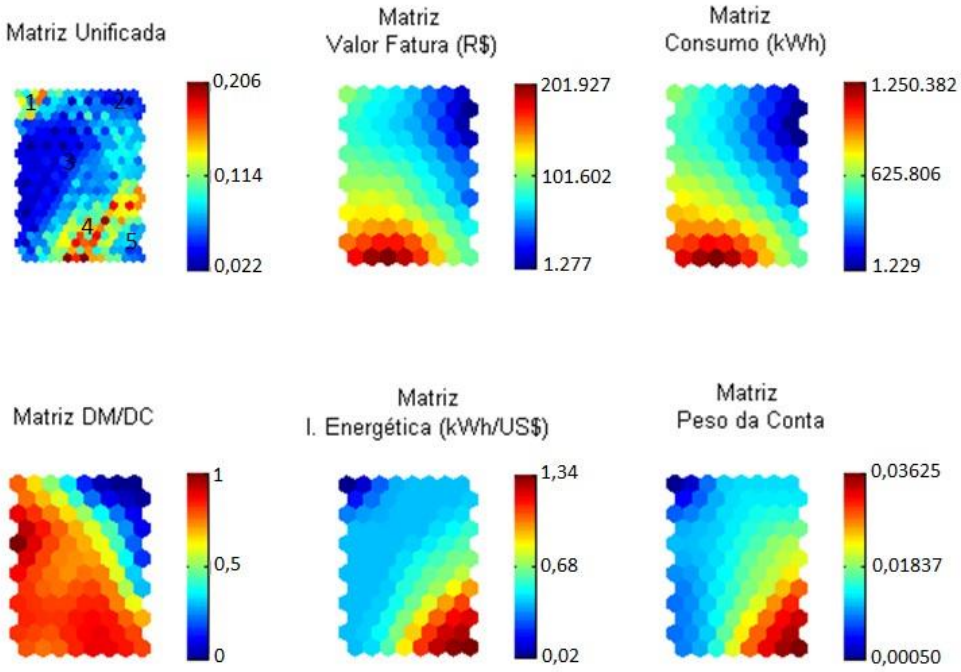

Figura 28- Matriz de Distância Unificada e Matrizes de Distância por Atributos
Fonte: Elaboração própria

A Figura 29 apresenta duas técnicas (b) e (c) que em conjunto evidenciam os grupos na rede SOM (d). Em (a), tem-se cada unidade neural rotulada com o primeiro vetor de entrada que a integrou; (b), a quantidade de vetores de entrada por unidade neural (*hits*) representada pela densidade de preenchimento das unidades. Ambas imagens representam a rede SOM propriamente dita. Pela imagem (b) verifica-se que, regiões mais densas apresentam similaridades para com seus vizinhos, constituindo assim, possíveis grupos separados por regiões menos densas. Em (c), uma classificação por cores (similaridade por cor) sobre cada unidade neural é apresentada distinguindo os grupos. Em (d), os grupos são evidenciados considerando a visualização conjunta entre (b) e (c).

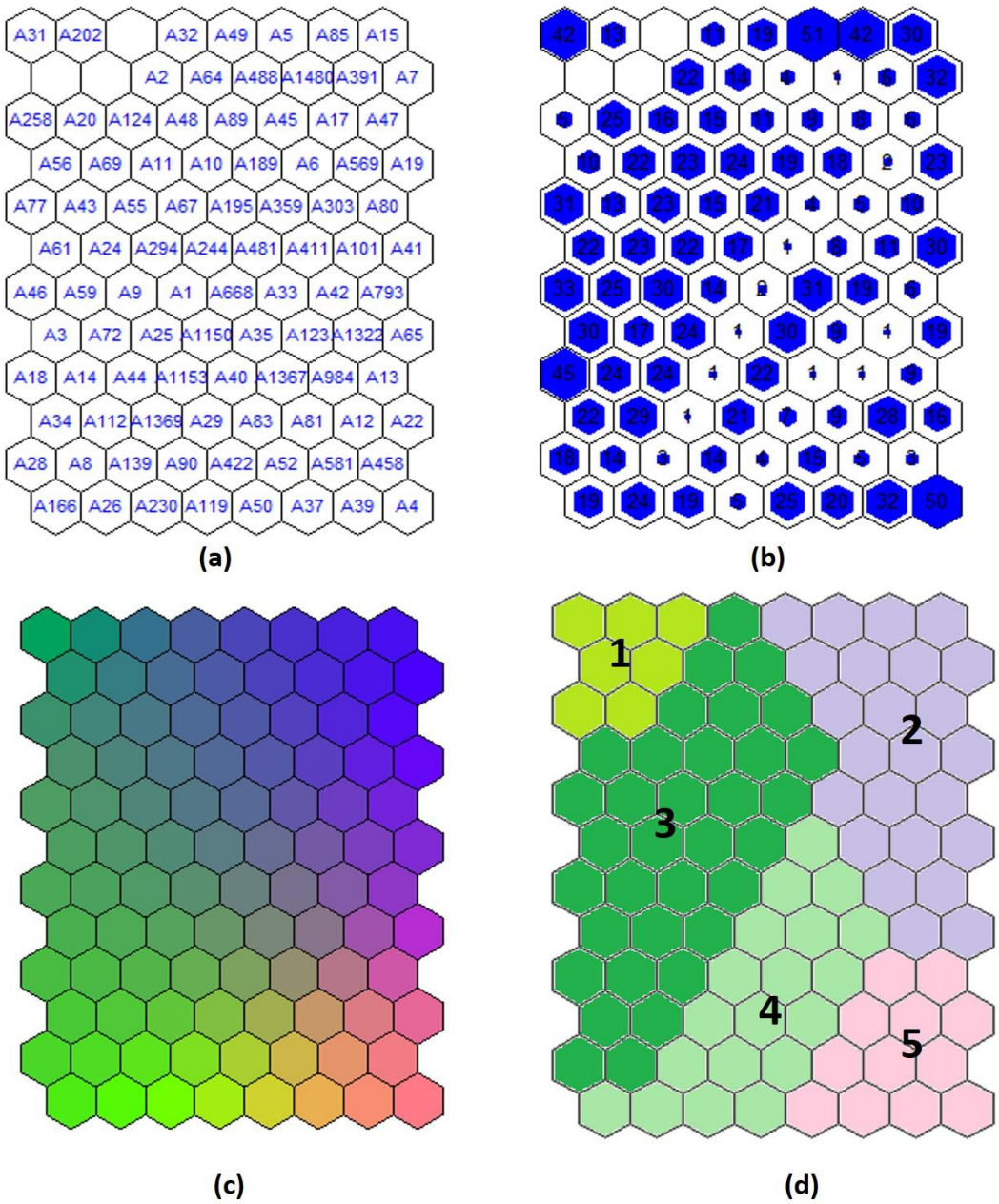

Figura 29- Grupos de Consumidores
Fonte: Elaboração própria

Assim, as unidades neurais que compõem cada grupo são descritas a seguir:

Grupo 1= {1, 2, 3, 13, 14, 15, 25}

Grupo 2= {49, 50, 61, 62, 63, 64, 65, 66, 73, 74, 75, 76, 77, 78, 79, 80, 85, 86, 87, 88, 89, 90, 91, 92}

Grupo 3= {4, 5, 6, 7, 8, 9, 10, 11, 16, 17, 18, 19, 20, 21, 22, 23, 26, 27, 28, 29, 30, 31, 32, 33, 37, 38, 39, 40, 41, 42, 43, 51, 52, 53}

Grupo 4= {12, 24, 34, 35, 36, 44, 45, 46, 47, 48, 54, 55, 56, 57, 58, 59, 67, 68, 69}

Grupo 5= {60, 70, 71, 72, 81, 82, 83, 84, 93, 94, 95, 96}

Em seguida, construiu-se um gráfico dos perfis para esses grupos (Figura 30). Na Figura 30, cada grupo é descrito por um conjunto de barras que representa a média das observações para cada variável em análise (os pontos da esquerda para direita, em cada perfil, se referem aos atributos: valor da fatura, consumo, demanda medida/demanda contratada, intensidade energética e peso da conta).

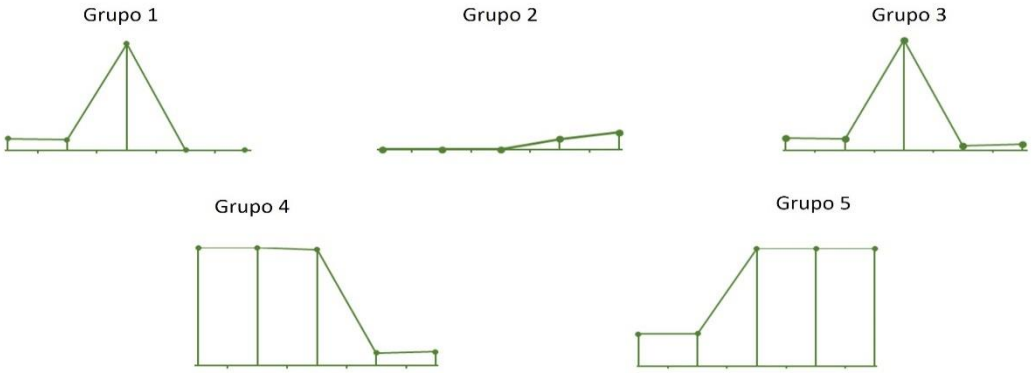

Figura 30- Perfil dos Grupos
Fonte: Elaboração própria

As informações contidas na Figura 28 expressam fidedignamente as informações da Figura 30. Na Tabela 4, constam informações que permitem caracterizar os grupos gerados.

Tabela 4- Dados Quantitativos dos Grupos

Grupo	Estatística	Valor da Fatura	Consumo (kWh)	DM/DC	IE (kWh/US$)	Peso da Conta
Grupo 1	Máximo	R$ 64.017,00	117.331	2,26	0,08	0,00171
	Média	R$ 11.263,96	23.297	0,86	0,04	0,00082
	Mínimo	R$ 1.650,00	2.846	0,00	0,02	0,00029
Grupo 2	Máximo	R$ 12.190,00	14.947	1,00	1,34	0,04561
	Média	R$ 2.576,54	3.514	0,16	0,14	0,00458
	Mínimo	R$ 86,00	116	0,00	0,02	0,00054
Grupo 3	Máximo	R$ 48.169,00	108.696	3,25	0,08	0,00468
	Média	R$ 11.696,18	24.267	0,88	0,08	0,00203
	Mínimo	R$ 1.768,00	2.500	0,14	0,08	0,00094
Grupo 4	Máximo	R$ 2.354.998,00	7.127.395	1,62	0,77	0,01333
	Média	R$ 89.890,75	224.809	0,92	0,16	0,00383
	Mínimo	R$ 2.439,00	3.383	0,00	0,08	0,00061
Grupo 5	Máximo	R$ 577.230,00	1.312.517	3,00	3,31	0,14104
	Média	R$ 26.356,28	64.181	0,93	1,11	0,02678
	Mínimo	R$ 1.473,00	1.780	0,00	0,49	0,00958

Fonte: Elaboração própria

Os passos seguintes foram: (i) caracterizar os grupos de consumidores com as variáveis qualitativas disponibilizadas para desenvolvimento desta pesquisa; (ii) implementar o conjunto teste no módulo neural e, então, (iii) implementar o módulo bayesiano.

A quantidade de consumidores alocados por grupos conforme variáveis qualitativas é apresentada na Tabela 5.

A implementação do conjunto teste na rede SOM ocorreu de modo a procurar a menor distância entre o vetor de entrada de cada consumidor com o vetor protótipo de cada unidade neural. Esse procedimento foi avaliado pelo erro de quantização que, em média, foi de 0,0656.

Tabela 5- Número de Consumidores Alocados por Unidade Neural Conforme Variáveis Qualitativas em Estudo

Conjunto	Subconjunto	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
Atividade econômica	Alojamento, Alimentação e Transportes	62	63	126	3	0
	Atividades Financeiras, Imobiliária e Administrativa	0	44	86	8	0
	Comércio Atacadista	2	22	68	3	0
	Comércio Varejista	19	24	256	15	0
	Construção	0	29	0	20	0
	Educação e Saúde	1	44	121	11	0
	Indústria de Alimentos, Bebidas e Agropecuários	0	19	0	6	72
	Indústria de Não-Metálicos	0	8	0	8	97
	Indústria Metalúrgica, de Máquinas e Equipamentos Diversos	0	30	0	50	2
	Indústria Diversas	0	38	0	76	42
	Serviços Diversos	1	43	66	15	0
Estrutura Tarifária	Convencional	34	321	404	52	28
	Horo-sazonal	51	43	319	163	185
Natureza Jurídica	Administração pública e Entidades Sem Fins Lucrativos	1	79	132	12	0
	Sociedade Empresária Limitada	25	49	76	51	34
	Outras Entidades Empresariais	59	236	515	152	179
Mesorregião	Baixadas	6	54	89	13	13
	Centro-Sul	8	65	71	20	14
	Metropolitana	46	152	429	126	89
	Noroeste	2	36	23	9	21
	Norte	23	57	111	47	76

Tabela 5- Número de Consumidores Alocados por Unidade Neural Conforme Variáveis Qualitativas em Estudo (continuação)

Conjunto	Subconjunto	Grupo 1	Grupo 2	Grupo 3	Grupo 4	Grupo 5
Porte	Grande	4	9	16	67	25
	Médio	56	182	446	90	40
	Pequeno	25	173	261	58	148

Fonte: Elaboração própria

Assim, a partir do conjunto (treino e teste) amostrado, obteve-se as $P(\mathbf{x}|c_i)$, onde c representa a variável classe (Adimplente e Inadimplente) e **x**, a ocorrência de cada variável característica dentre de um contexto de possibilidades para ela disponíveis. Posteriormente, obtém-se as probabilidades conjuntas $P(\mathbf{x}|c_i)$ e, então, a partir do conjunto teste (variáveis categóricas dos consumidores conjunto teste constam em Apêndice C), obteve-se as $P(c_i|\mathbf{x})$. Nas Tabelas 6 e 7, constam a base de cálculo para as probabilidades conjuntas $P(\mathbf{x}|c_i)$.

Tabela 6- Probabilidades Marginais

Conjuntos	Subconjuntos	Cardinalidade	Prob. Marginal
Comportamento de pagamento	Adimplente	573	0,36
	Inadimplente	1027	0,64
Atividade econômica	Alojamento, Alimentação e Transportes	254	0,16
	Atividades Financeiras, Imobiliárias e Administrativas	138	0,09
	Comércio Atacadista	95	0,06
	Comércio Varejista	314	0,20
	Construção	49	0,03
	Educação e Saúde	177	0,11
	Indústria de Alimentos, Bebidas e Agropecuários	97	0,06
	Indústria de Não-Metálicos	113	0,07
	Indústria Metalúrgica, de Máquinas e Equipamentos Diversos	82	0,05
	Indústrias Diversas	156	0,10
	Serviços Diversos	125	0,08
Estrutura tarifária	Convencional	839	0,52
	Horo-sazonal	761	0,48
Natureza jurídica	Administração pública e Entidades Sem Fins Lucrativos	224	0,14
	Sociedade Empresária Limitada	1141	0,71
	Outras Entidades Empresariais	235	0,15

Tabela 6- Probabilidades Marginais (continuação)

Conjuntos	Subconjuntos	Cardinalidade	Prob. Marginal
Porte	Grande	121	0,08
	Médio	814	0,51
	Pequeno	665	0,42
Mesorregião	Baixadas	175	0,11
	Centro-Sul	178	0,11
	Metropolitana	842	0,53
	Noroeste	91	0,06
	Norte	314	0,20
Rede SOM	Grupo 1	85	0,05
	Grupo 2	364	0,23
	Grupo 3	723	0,45
	Grupo 4	215	0,13
	Grupo 5	213	0,13

Fonte: Elaboração própria

Tabela 7- Probabilidades Condicionais

Variável Categórica	Adimplente	Inadimplente
Alojamento, Alimentação e Transportes	0,1658	0,1548
Atividades Financeiras, Imobiliárias e Administrativas	0,0942	0,0818
Comércio Atacadista	0,0750	0,0506
Comércio Varejista	0,2304	0,1772
Construção	0,0192	0,0370
Educação e Saúde	0,0838	0,1256
Indústria de Alimentos, Bebidas e Agropecuários	0,0576	0,0623
Indústria de Não-Metálicos	0,0803	0,0652
Indústria Metalúrgica, de Máquinas e Equipamentos Diversos	0,0401	0,0574
Indústrias Diversas	0,0785	0,1081
Serviços Diversos	0,0750	0,0798
Convencional	0,5305	0,5209
Horo-sazonal	0,4695	0,4791
Administração pública e Entidades Sem Fins Lucrativos	0,1449	0,1373
Sociedade Empresária Limitada	0,7382	0,6991
Outras Entidades Empresariais	0,1169	0,1636
Grande	0,0663	0,0808
Médio	0,5497	0,4859
Pequeno	0,3839	0,4333
Baixadas	0,1047	0,1120
Centro-Sul	0,1204	0,1061
Noroeste	0,0524	0,0594
Norte	0,1693	0,2113
Metropolitana	0,5532	0,5112

Tabela 7- Probabilidades Condicionais (continuação)

Variável Categórica	Adimplente	Inadimplente
Grupo 1	0,0646	0,0467
Grupo 2	0,2112	0,2366
Grupo 3	0,4834	0,4343
Grupo 4	0,1187	0,1431
Grupo 5	0,1222	0,1392

Fonte: Elaboração própria

A taxa de acerto do classificador bayesiano simples pode ser verificada pela Tabela 8 (isto é, matriz de confusão, na qual constam o número de acertos e erros do modelo por classe). Nesta tabela, incluiu-se também a quantia monetária relacionada a soma do valor mediano da fatura desses consumidores.

Tabela 8- Matriz de Confusão para os Perfis de Pagamento dos Consumidores

Perfil de Pagamento		Previsto	
		Adimplente	Inadimplente
Real	Adimplente	20 consumidores (R$ 261.159,00)	92 consumidores (R$ 2.412.325,00)
	Inadimplente	11 consumidores (R$ 205.184,00)	153 consumidores (R$ 3.157.910,00)

Fonte: Elaboração própria

O modelo obteve uma acurácia regular (63% dos casos foram detectados corretamente) pois foi afetado pelo baixíssimo valor preditivo do perfil adimplente (somente 18% dos casos foram detectados corretamente) mas, apresentou bom valor preditivo do perfil inadimplente (93% dos casos foram detectados). Esta ocorrência pode ser justificada pela prevalência da variável classe Inadimplência sobre a classe Adimplência, tendendo a classificar novos consumidores como pertencente à classe majoritária. Isto é indesejável quando as classes minoritárias (neste caso, Adimplência) são aquelas que possuem uma informação muito importante. Para os sistemas de distribuição de energia elétrica, o importante é verificar as chances de um consumidor ser inadimplente.

Todavia, a melhoria da acurácia do sistema proposto é interessante para fins de gestão do fluxo de caixa das distribuidoras de energia elétrica e pode ser alcançada balanceando as variáveis classes, de modo que, a classe inadimplência seja redistribuída em outras classes.

É válido mencionar que, a inserção de novos consumidores que não possuam dados quantitativos está condicionada a uma estimativa de alocação ao grupo, que pode ser realizada construindo um classificador bayesiano simples através dos dados da Tabela 5.

5
Conclusão e Perspectiva da Pesquisa

A gestão da inadimplência nas distribuidoras de energia elétrica é uma atividade vital para todo o Setor Elétrico, pois é a tarifa de fornecimento de energia, recolhida mensalmente por aquelas, a fonte de remuneração do Setor. A implementação de metodologias para combate à inadimplência nessas empresas tem sido, essencialmente, focada na recuperação do faturamento perdido. Nessa área, poucos estudos têm explorado o desenvolvimento de um sistema preventivo.

Assim, buscou-se responder: Como desenvolver um sistema que seja capaz de compreender o fator inadimplência em unidade consumidoras não-residenciais atendidas na média e alta tensão e, então, contribuir para que a gestão da inadimplência sane as insuficiências dos instrumentos vigentes?

Com relação a este questionamento, constatou-se que, o uso de técnicas *soft computing* aplicadas na mineração de dados, os quais são armazenados pelas operações diárias das distribuidoras, poderiam contribuir para compreender o comportamento de pagamento dos consumidores, visto que essas tecnologias têm mostrado ser, nos últimos anos, campo potencial para muitas pesquisas aplicadas à área gerencial.

Em decorrência disto, propôs-se a aplicação de um sistema inteligente híbrido intercomunicativo composto por módulos –neural, que alocasse consumidores com similaridades relacionadas às variáveis quantitativas e bayesiano, que estabelecesse um escore da propensão adimplência/inadimplência de determinado consumidor considerando a unidade neural no qual se encontrasse alocado e atributos categóricos que o caracterizassem– que foram alimentados por informações geradas pelas operações diárias de uma distribuidora de energia elétrica.

Os resultados revelaram que o sistema proposto para detecção do comportamento de pagamento apresentou razoável acurácia, permitindo assim que o objetivo geral "propor um sistema que classifica o perfil de pagamento de unidades consumidoras não-residenciais de distribuidora de energia elétrica considerando conhecimento armazenado em base de dados" fosse alcançado.

Pelos aspectos descritos e resultados gerados no desenvolvimento desta pesquisa, pode-se concluir que a presente pesquisa contribuiu para o desenvolvimento de uma metodologia complementar aos estudos que exploram meios preventivos de combate à inadimplência em distribuidoras de energia elétrica. É importante ressaltar que a proposta não está restrita ao uso específico das variáveis e configurações adotadas, porém, deve ser modificada e adaptada à realidade de cada empresa, a fim de melhor exprimir o perfil de consumidores delas.

Para trabalhos futuros de desdobramento da pesquisa e aprofundamento dos resultados, propõe-se a implementação deste sistema em um *software* livre e a investigação de outros fatores que possam compor a variável comportamento de pagamento e, assim, obter um balanceamento entre as variáveis classes do sistema e, consequentemente, melhorar a acurácia deste. Duas linhas de pesquisa poderiam ser investigadas:

> 1ª linha –a classe inadimplência da variável comportamento de pagamento ser definida em função da duração dos atrasos do pagamento da fatura mensal e dos procedimentos de cobrança geralmente adotados pelas distribuidoras, de modo que:
>
> - Inadimplente 1: 1 a 6 dias em atraso, aguardar pelo pagamento;
> - Inadimplente 2: 7 a 23 dias em atraso, realizar telecobrança;
> - Inadimplente 3: 24 a 37 dias em atraso, notificar corte;
> - Inadimplente 4: 38 a 44 dias, realizar corte;
> - Inadimplente 5: 45 dias ou mais de atraso, acionar a Justiça e o Serviço de Proteção ao Crédito.
>
> 2ª linha –criar um módulo *fuzzy* que estratificasse, por meio de regras fornecidas por especialistas ou extraídas dos dados, o comportamento de pagamento conforme variáveis: frequência de atraso das unidades consumidoras no período jan./2009 a dez/2010, dias de atrasos mensais e peso da conta.

Além disso, por meio de um projeto de pesquisa e desenvolvimento que reúna agentes do governo (ANEEL), da indústria de distribuição (uma distribuidora de energia elétrica) e da universidade (estudantes de Pós-graduação na área de interesse), os custos reais de operação dos consumidores e o impacto de variáveis macroeconômicas (taxa de inflação, taxa de desemprego, instabilidade política, etc.) sobre o comportamento de pagamentos dos consumidores poderiam ser investigados e adaptados ao sistema nesse estudo proposto.

Referências bibliográficas

AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA (Brasil). **Atlas de energia elétrica do** Brasil. 3. ed. Brasília: 2008. 236p.

AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA (Brasil). Resolução normativa nº 414 de 9 de setembro de 2010. **Diário Oficial**, 15 set. 2010, seção 1, v. 147, n. 177, p. 115. Disponível em: <www.aneel.gov.br/biblioteca>. Acesso em: dez.2016.

AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA (Brasil). Resolução normativa nº 479 de 3 de abril de 2012. **Diário Oficial**, 12 abr. 2012, seção 1, v. 149, n. 71, p. 48. Disponível em: <www.aneel.gov.br/biblioteca>. Acesso em: dez.2016.

AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA. Site com informações sobre o setor elétrico. Disponível em: <www.aneel.gov.br>. Acesso em: dez.2016.

ANTMANN, P. Reducing technical and non-technical losses in the power sector. **World Bank**, Washington, 2009.

ARAUJO, A. C. M. de. **Perdas e inadimplência na atividade de distribuição de energia elétrica no Brasil**. Rio de Janeiro, 2007. 116p. Tese de Doutorado – COPPE, Universidade Federal do Rio de Janeiro.

BASTOS, P. R. F. de M. **Diagnóstico de perdas comerciais de energia elétrica na distribuição usando redes Bayesianas**. Campina Grande, 2011. 125p. Tese de Doutorado – Programa de Pós-graduação em Engenharia Elétrica, Universidade Federal de Campina Grande.

BIGGADIKE, E. R. The contributions of marketing to strategic management. **Academy of Management Review**, v. 6, n. 4, p. 621–632, 1981.

BLOCKER, C. P.; FLINT, D. J. Customer segments as moving targets: integrating customer value dynamism into segment instability logic. **Industrial Marketing Management**, v. 36, n. 6, p. 810–822, 2007.

BRAMER, M. **Principles of Data Mining**. London: Springer Science & Business Media, 2007. 343p.

CABRAL, J. E. C.; PINTO, J. O. P.; LINARES, K. S. C.; PINTO, A. M. A. C. (2006). Methodology for fraud detection using rough sets. In: **IEEE International Conference on Granular Computing**. 2006, p. 244–249.

CALILI, R. F. **Desenvolvimento de sistema para detecção de perdas comerciais em redes de distribuição de energia elétrica**. Rio de Janeiro, 2005, 157p. Dissertação de Mestrado – Departamento de Engenharia Elétrica, Pontifícia Universidade Católica do Rio de Janeiro.

CÂMARA DE COMERCIALIZAÇÃO DE ENERGIA ELÉTRICA (Brasil). **Visão geral das operações na CCEE**: versão 2010. Brasília: 2010. 94p.

CAMPOS, C. **Curso básico de direito de energia elétrica**. Rio de Janeiro: Synergia, 2010. 168p.

CARVALHO, A. F.; BRAGA, A. P.; LUDERMIR, T. B. **Fundamentos de redes neurais artificiais**. Rio de Janeiro: DCC/IM, COPPE/SISTEMAS, NCE/UFRJ, 1998.

CARVALHO, N. A. da S. **Aplicação de modelos estatísticos para previsão e monitoramento da cobrabilidade de uma empresa distribuidora de energia elétrica no Brasil**. Rio de Janeiro, 2011. 154p. Dissertação de Mestrado – Programa de Pós-graduação em Metrologia para Qualidade e Inovação, Pontifícia Universidade Católica do Rio de Janeiro.

CARVALHO, N. A. S.; SOUZA, R. C.; EPPRECHT, E. K. Topologia do perfil de pagamento dos consumidores de alta tensão de um distribuidor de energia elétrica. In: **XLVI Simpósio Brasileiro de Pesquisa Operacional**. Bahia: 2014, p. 1194–1205.

CHAIMONTREE, S.; ATKINSON, K.; COENEN, F. Best clustering configuration metrics: towards multiagent based clustering. In: **International Conference on Advanced Data Mining and Applications**. Chongqing: Springer, 2010, p. 48–59.

CHEN, M. S.; HAN, J.; YU, P. S. Data mining: an overview from a database perspective. **IEEE Transactions on Knowledge and Data Engineering**, v. 8, n. 6, p. 866–883, 1996.

CHENG, J.; GREINER, R. Learning bayesian belief network classifiers: algorithms and system. In: **Proceeding of 14th Biennial Conference of the Canadian Society for Computational Studies of Intelligence**. p. 141–151, 2001.

CHIAVENATO, I. **Administração Geral e Pública**. 4. ed. Baurueri: Manole, 2016.

COPPIN, B. **Artificial Intelligence Illuminated**. Sudbury: Jones and Bartlett Plublishers, 2004. 698p.

CÔRTES, S. D. C.; PORCARO, R. M.; LIFSCHITZ, S. **Mineração de Dados – Funcionalidades, Técnicas e Abordagens**. Rio de Janeiro, 2002. Monografia – Informática, Pontifícia Universidade Católica do Rio de Janeiro.

COSTA, E. O. da; FABRIS, F.; LOUREIROS, A. R.; AHONEN, H.; VAREJAO, F. M. Using GA for the stratified sampling of electricity consumers. In: **IEEE Congress on Evolutionary Computation**. Cancún: 2013, p. 261–268.

DEPURU, S. S. S. R.; WANG, L.; DEVABHAKTUNI, V. Electricity theft: overview, issues, prevention and a smart meter based approach to control theft. **Energy Policy**, n. 39, p. 1007–1015, 2011.

DIAS, H. B. P. **Uma abordagem baseada em conhecimento para apoio ao combate às perdas comerciais na distribuição de energia elétrica**. Vitória, 2006. 96p. Dissertação de Mestrado – Programa de Pós-graduação em Informática – Universidade Federal do Espírito Santo.

DIBB, S. Developing a decision tool for identifying operational and attractive segments. **Journal of Strategic Marketing**, n. 3, p. 189–203, 1995.

DIBB, S. Criteria guiding segmentation implementation: reviewing the evidence. **Journal of Strategic Marketing**, n. 7, p. 107–109, 1999.

ELLER, N. A. **Arquitetura de informação para o gerenciamento de perdas comerciais de energia elétrica**. Florianópolis, 2003, 115p. Tese de Doutorado – Programa de Pós-graduação em Engenharia de Produção – Universidade Federal de Santa Catarina.

EMPRESA DE PESQUISA ENERGÉTICA (Brasil). **Balanço Energético Nacional 2013**: ano base 2012. Rio de Janeiro: 2013.

ESTIVILL-CASTRO, V. Why so many clustering algorithms: a position paper. **SIGKDD Explorations**, v. 4, n. 1, p. 65–75, 2002.

EVERITT, B.; LANDAU, S.; LEESE, M.; STAHL, D. **Cluster analysis**. 5. ed. Nova Jersey: Wiley, 2011.

FANG, X.; MISRA, S.; XUE, G.; YANG, D. Smart grid -the new and improved power grid: a survey. **IEEE Communications Surveys and Tutorial**, v. 14, n. 4, p. 944–980, 2012.

FARIA, E. L.; ALBUQUERQUE, M. P.; ALFONSO, J. L. G.; ALBUQUERQUE, M. P.; CAVALCANTE, J. T. P. **Introdução ao toolbox de redes neurais de Kohonen**. Rio de Janeiro: Centro Brasileiro de Pesquisas Físicas, 2010.

FARIA, L. T. de. **Sistema inteligente híbrido intercomunicativo para detecção de perdas comerciais**. Ilha Solteira, 2012. 112p. Dissertação de Mestrado – Universidade Estadual Paulista.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37–54, 1996.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. **Advances in knowledge discovery and data mining**. MIT Press, 1996.

FERRARI, E. L. **Contabilidade Geral:** Teoria e 950 questões. 4. ed. Rio de Janeiro: Impetus, 2003.

FONSECA, J. N.; REIS, L. B. **Empresas de distribuição de energia elétrica no Brasil: temas relevantes para gestão**. Rio de Janeiro: Synergia, 2012.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining:** um guia prático. Rio de Janeiro: Elsevier, 2005.

GOLLER, S.; HOGG, A.; KALAFATIS, S. P. A new research agenda for business segmentation. **European Journal of Marketing**, n. 36, p. 252–271, 2002

GOYAT, S. The basis of market segmentation: a critical review of literature. **European Journal of Business and Management**, n. 3, p. 45–54, 2011.

GROTH, R. **Data mining:** a hands-on approach for business professionals. Prentice-Hall, 1998.

HAIR JR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise Multivariada de Dados**. 6. ed. Porto Alegre: Bookman, 2009, 688p.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. **Journal of Intelligent Information Systems**, v. 17, n. 2–3, p. 107–145, 2001.

HAN, J.; KAMBER, M. **Data Mining:** concepts and techiques. 2. ed. Academic Press, 2006.

HAND, D.; MANNILA, H.; SMYTH, P. **Principles of data mining**. Massachusetts Institute of Technology, 2001.

HAYKIN, S. **Redes Neurais**. 2. ed. Porto Alegre: Bookman, 2001.

HIZIROGLU, A. Soft computing applications in customer segmentation: state-of-art review and critique. **Expert Systems with Application**, n. 40, p. 6491–6507, 2013.

HUANG, S.-C.; LO, Y.-L.; LU, C.-N. Non-technical loss detection using state estimation and analysis of variance. **IEEE Transactions on Power Systems**, v. 28, n. 3, p. 2959–2966, 2013

INSTITUTO ACENDE BRASIL. Perdas e inadimplência no setor elétrico. **Cadernos de política tarifária**, p. 1–9, 2007a.

INSTITUTO ACENDE BRASIL. Política tarifária e regulação por incentivos. **Cadernos de política tarifária**, p. 1–10, 2007b.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Divisão Regional do Brasil em Mesorregiões e Microrregiões Geográficas**. Rio de Janeiro, 1990.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Classificação Nacional de Atividades Econômicas**. Rio de Janeiro, 2007.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. Site com estudos e informações estatísticas nacionais de diversos setores. Disponível: <www.ibge.gov.br>. Acesso em: dez.2016.

JACKSON, J. Data mining a conceptual overview. **Comunications of the Association for Information Systems**, n. 8, p. 267–296, 2002.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM Computing Surveys**, v. 31, n. 3, p. 264–323, 1999.

JIANG, R.; TAGARIS, H.; LACHSZ, A.; JEFFREY, M. Wavelet based feature extraction and multiple classifiers for electricity fraud detection. In: **Transmission and Distribution Conference and Exhibition**. Asia Pacific: 2002, p. 2251–2256.

KIVILUOTO, K. Topology preservation in self-organizing maps. In **International Conference on Neural Networks**. 1996, p. 294–299.

KOHONEN, T. The self-organizing map. In: **Proceeding of the IEEE**. 1990, p. 1464–1480.

KOHONEN, T. Exploration of very large databases by self-organizing maps. In: **International Conference on Neural Networks**. Houston: IEEE, 1997a.

KOHONEN, T. **Self organizing maps**. 2. ed. Heidelberg: Springer, 1997b.

KOHONEN, T. Essentials of the self-organizing map. **Neural Networks**, n. 37, p.52–65, 2013.

KOLIOU, E.; BARTUSCH, C.; PICCIARIELLO, A.; EKLUND, T.; SÖDER, L.; HAKVOORT, R. A. Quantifying distribution-system operators' economic incentives to promote residential demand response. **Utilities Policy**, n. 35, p. 28–40, 2015.

KOTLER, P. **Marketing management**. 13. ed. New Jersey: Prentice-Hall, 2003.

KRACKLAUER, A. H.; MILLS, D. Q.; SEIFERT, D. **Collaborative customer relationship management:** talking CRM to the next level. Boston: Springer, 2004.

LAMIN, H. **Análise de impacto regulatório da implantação de redes inteligentes no Brasil**. Brasília: 2013. 300p. Tese de Doutorado – Departamento de Engenharia Elétrica – Universidade de Brasília.

LEÓN, C.; BISCARRI, F.; MONEDERO, I.; GUERRERO, J. I.; BICARRI, J.; MILLÁN, R. Integrated expert system applied to the analysis of non-technical losses in power utilities. **Expert Systems with Application**, n. 38, p. 10274–10285, 2011.

LIU, Y.; LI, Z.; XIONG, H.; GAO, X.; WU, J. Understanding of intenal clustering validation measures. In: **IEEE Internatinal Conference on Data mining**. Sidney: 2010, p. 911–916.

MAGDALENA, L. What is soft computing? Revisiting Possible Answer. **International Journal of Computational Intelligence Systems**, v. 3, n. 2, p. 148–159, 2010.

MARTINS, J. G. F. **Proposta de método para classificação do porte das empresas**. Natal, 2014. 78p. Dissertação de Mestrado – Programa de Pós-graduação em Administração – Universidade de Potiguar.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **Bulletin of Mathematical Biophysics**, n. 5, p. 115–133, 1943.

MEDEIROS, A. L. **Alocação de equipes de campo para avaliação de perdas não-técnicas de energia elétrica**: desenvolvimento de um sistema de apoio a decisão. Lavras, 2013. 186p. Tese de Doutorado – Programa de Pós-graduação em Administração – Universidade Federal de Lavras.

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. **Electrical Power and Energy Systems**, n. 34, p. 90–98, 2012.

NAGI, J.; MOHAMMAD, A. M.; YAP, K. S.; TIONG, S. K.; AHMED, S. K. Non-technical loss analysis for detection of electricity theft using support

vector machines. In: **2nd IEEE International Conference on Power and Energy**. Malásia: 2008, p. 907–912.

NGAI, E. W.; XIU, L.; CHAU, D. C. K. Application of data mining techniques in customer relationship management: a literature review and classification. **Expert Systems with Application**, n. 36, p. 2592–2602, 2009.

NILSSON, G. A suspensão do fornecimento do serviço essencial de energia elétrica ao usuário inadimplente - análise sob a perspectiva constitucional e consumerista. **Direito & Justiça**, v. 38, n. 2, p. 141–155, 2012.

NIZAR, A. H.; DONG, Z. Y.; JALALUDDIN, M.; RAFFLES, M. J. Load profiling method in detecting non-technical loss activities in a power utility. In: **First International Power and Energy Conference**. Malásia: 2006, p. 82–87.

NIZAR, A. H.; DONG, Z. Y.; ZHANG, P. Detection rules for non technical losses analysis in power utilities. In: **2008 IEEE Power and Energy Society General Meeting** - **Conversion and Delivery of Electrical Energy in the 21st Century**. Pensilvânia: 2008.

OLIVEIRA, M. E. de. **Avaliação de metodologias de cálculo de perdas técnicas em sistemas de distribuição de energia elétrica**. Ilha Solteira, 2009. 135p. Tese de Doutorado – Programa de Pós-graduação em Engenharia Elétrica – Universidade Estadual Paulista.

ORTEGA, G. V. C. **Redes neurais na identificação de perdas comerciais do setor elétrico**. Rio de Janeiro, 2008. 184p. Dissertação de Mestrado – Programa de Pós-graduação em Engenharia Elétrica – Pontifícia Universidade Católica do Rio de Janeiro.

PENIN, C. A. de S. **Combate, prevenção e otimização de perdas comerciais de energia elétrica**. São Paulo, 2008. 214p. Tese de Doutorado – Escola Politécnica – Universidade de São Paulo.

PETKOVIC, I.; BALABAN, N. Detecting defaulters for consumed electric energy with neural clustering. In: **6th International Symposium on Intelligent Systems and Informatics**. Subotica: 2008, p. 1-4.

QUEIROGA, R. M. **Uso de técnicas de data mining para detecção de fraudes em energia elétrica**. Vitória, 2005. 146p. Dissertação de Mestrado – Programa de Pós-graduação em Informática – Universidade Federal do Espírito Santo.

RAAIJ, W. F.; VERHALLEN, T. M. M. Domain-specific market segmentation. **European Journal of Marketing**, n. 28, p. 49–66, 1994.

RAMOS, C. C. O. **Caracterização de perdas comerciais em sistemas de energia através de técnicas inteligentes**. São Paulo, 2014. 128p. Tese de Doutorado – Escola Politécnica – Universidade de São Paulo.

RAMOS, C. C. O.; SOUSA, A. N. de; PAPA, J. P.; FALCÃO, A. X. A new approach for nontechnical losses detection based on optimum-path forest. **IEEE Transactions on Power Systems**, v. 26, n. 1, p. 181–189, 2011.

REIS FILHO, J. (2006). **Sistema inteligente baseado em árvore de decisão, para apoio ao combate às perdas comerciais na distribuição**

de energia elétrica. Uberlândia, 2006. 174p. Dissertação de Mestrado – Programa de Pós-graduação em Engenharia Elétrica – Universidade Federal de Uberlândia.

REZENDE, S. O.; EVSUKOFF, A. G.; GARCIA, A. C. B.; CARVALHO, A. C. P. L. F.; BRAGA, A. P.; MONARD, M. C.; LUDERMIR, T. B. **Sistemas inteligentes: fundamentos e aplicações**. Baurueri: Manole, 2005. 525p.

RIBEIRO, S. S.; CAZES, T.; MANO, R. F. Investment optimization methodology applied to investments on non-technical losses reduction actions. In: **Seventeenth IEEE Symposium on Computers and Communications**. Capadócia: 2012, p. 354–360.

ROKACH, L.; MAIMON, O. **Data mining and knowledge discovery handbook**. Springer, 2005.

SASSI, R. J. **Uma arquitetura híbrida para descoberta de conhecimento em base de dados**: teoria dos rough sets e redes neurais artificiais mapas auto organizáveis. São Paulo, 2006. 169p. Tese de Doutorado – Escola Politécnica – Universidade de São Paulo.

SMITH, T. B. Electricity theft: a comparative analysis. **Energy Policy**, n. 32, p. 2067–2076, 2004.

SMITH, W. R. Product differentiation and market segmentation as an alternative marketing strategy. **Journal of Marketing**, n. 21, p. 3–8, 1956.

SOUZA, R. C.; PESSANHA, J. F. M.; OLIVEIRA, F. L. C. A residential consumer payment capability index based on fuzzy logic inference. **Journal of Intelligent & Fuzzy Systems**, v. 25, n. 3, p. 649–657, 2013.

SPIEGEL, M. R.; STEPHENS, L. J. **Schaum's outline of theory and problems of statistics**. 4. ed. McGraw-Hill Inc, 2008. 577p.

SUN, S. An analysis on the conditions and methods of market segmentation. **International Journal of Business and Management**, v. 4, n. 2, p. 63–70, 2009.

TKACZYNSKI, A.; RUNDLE-THIELE, S. R. Event segmentation: a review and research agenda. **Tourism Management**, n. 32, p. 426–434, 2011.

TOVAR, B.; RAMOS-REAL, F. J.; ALMEIDA, E. F. de. Firm size and productivity. Evidence from the electricity distribution industry in Brasil. **Energy Policy**, n. 39, p. 826–833, 2011.

TREVISAN, R. D. **Detecção e identificação de perdas comerciais em sistemas de distribuição**: metodologia baseada em floresta de caminhos ótimos. Porto Alegre, 2014. 89p. Dissertação de Mestrado – Programa de Pós-graduação em Engenharia Elétrica – Universidade Federal do Rio Grande do Sul.

TRIOLA, M. F. **Introdução à estatística**. 10. ed. Rio de Janeiro: LTC, 2011. 695p.

VERGARA, S. C. **Projetos e relatório de pesquisa em administração**. 14. ed. São Paulo: Atlas, 2013. 94p.

VESANTO, J. SOM-based data visualization methods. **Intelligent Data Analysis**, n. 3, p. 111–126, 1999.

VESANTO, J.; HIMBERG, J.; ALHONIEMI, E.; PARHANKANGAS, J. Self organizing map in Matlab: the SOM toolbox. In: **Proceeding of the Matlab DSP Conference**. Espoo: 2011, p. 35-40.

VESANTO, J.; HIMBERG, J.; ALHONIEMI, E.; PARHANKANGAS, J. **SOM toolbox for Matlab 5**. Espoo: 2000. 59p.

VIAENE, S.; DERRIG, R. A.; DEDENE, G. A case study of applying boosting naive bayes to claim fraud diagnosis. **IEEE Transactions on Knowledge and Data Engineering**, v. 16, n. 5, p. 612–620, 2004.

WENDEL, M.; KAMAKURA, W. **Market Segmentation:** Conceptual and Methodological Foundations. 2. ed. Boston: Kluwer Academic Publishers, 2000.

WILPPU, E. **The visualisation capability of self organizing maps to detect deviations in distribution control**. Turku: TUCS Research Group, 1997.

WORLD BANK. Site com informações mundiais de diversos setores da economia. Disponível em: <www.worldbank.org>. Acesso em: dez.2016.

YANG, M.; HUNG, W. L.; CHEN, D. H. Self-organizing map for symbolic data. **Fuzzy set and systems**, n. 203, p. 49–73, 2012.

YANKELOVICH, D.; MEER, D. Rediscovering market segmentation. **Harvard Business Review**, n. 84, p. 121–132, 2006.

ZADEH, L. A. Soft computing and fuzzy logic. **IEEE Software**, v. 11, n. 6, p. 48–56, 1994.

Apêndice A: Definição das Variáveis de Pesquisa

Atividade Econômica: combinação de ação que resulta em certos tipos de produtos ou, ainda, uma combinação de recursos que gera bens e serviços específicos, gerando valor adicionado (Instituto Brasileiro de Geografia e Estatística, 2007). De acordo com informação sobre atividade econômica principal cadastrada em *site* governamental, que realiza registro de Pessoas Jurídicas, os consumidores da base de dados da pesquisa foram enquadrados em uma das seguintes atividades econômica:

- Alojamento, Alimentação e Transportes (AAT)
- Atividades Financeiras, Imobiliárias e Administrativas (AFIA)
- Comércio Atacadista (CA)
- Comércio Varejista (CV)
- Construção (CO)
- Educação e Saúde (ES)
- Indústria de Alimentos, Bebidas e Agropecuários (IABA)
- Indústria de Não-Metálicos (INM)
- Indústria Metalúrgica, de Máquinas e Equipamentos Diversos (IMMED)
- Indústria Diversas (ID) → compreende atividades relacionadas a indústrias distintas: Extração mineral; Geração de energia; Produção de têxteis e artigos do vestuário, calçados e acessórios; Impressão e reprodução de gravações; Fabricação de produtos de madeira, celulose, papel e produtos de papel; Fabricação de produtos de borracha e material plástico; Fabricação de produtos químicos, farmoquímicos e farmacêuticos; Fabricação de artigos de joalheria, bijuteria e semelhantes; Fabricação de instrumentos musicais; Fabricação de artefatos para pesca e esporte; Fabricação de brinquedos e jogos recreativos; Fabricação de instrumentos e materiais médicos e odontológicos; Fabricação de escovas, pincéis e vassouras; Fabricação de equipamentos e acessórios para segurança e proteção pessoal e profissional; Fabricação de produtos diversos não especificados anteriormente.
- Serviços Diversos (SD) → compreende atividades relacionadas a serviços distintos: Serviços Industriais de Utilidade Pública (nesse grupo estão incluídos eletricidade, gás e água), exceto geração de energia; Serviços de informação e comunicação; Serviços profissionais, científicos e técnicos; Serviços da administração pública, defesa e seguridade social; Serviços de organizações associativas; Serviços de lavanderia, tinturaria e toalheiros; Serviços de beleza; Serviços funerários e outros relacionados; Serviços pessoais não especificados anteriormente.

Cadastro Nacional de Pessoa Jurídica (CNPJ): número que identifica e reúne informações cadastrais das entidades de interesse das administrações tributárias da União, dos Estados, do Distrito Federal e dos Municípios junto à Receita Federal do Brasil. O CNPJ é composto por quatorze dígitos cujos oitos primeiros (raiz)

identificam a empresa e são comuns a todas as unidades locais/estabelecimentos, os quatro seguintes (sufixo) identificam os endereços de atuação da empresa e dois últimos são dígitos verificadores, que é resultado de uma equação com os doze números anteriores (Instituto Brasileiro de Geografia e Estatística, 2007).

Comportamento de pagamento: modo de extinguir obrigações através do cumprimento efetivo de um pagamento. Consumidor que paga no vencimento ou antecipado é adimplente, caso contrário, inadimplente.

Consumo: quantidade de energia elétrica gasta pelo consumidor, atribuída ao uso de equipamentos elétricos em um intervalo de tempo, sendo expresso em kWh (Agência Nacional de Energia Elétrica, 2008, 2010).

Demanda contratada (DC): demanda de potência ativa, expressa em quilowatts (kW), a ser obrigatória e continuamente disponibilizada pela distribuidora, no ponto de entrega, conforme valor e período de vigência fixados em contrato, e que deve ser integralmente paga, seja ou não utilizada durante o período de faturamento (Agência Nacional de Energia Elétrica, 2008, 2010).

Demanda medida (DM): maior demanda de potência ativa, expressa em quilowatts (kW) e verificada por medição, integralizada em intervalos de 15 (quinze) minutos durante o período de faturamento (Agência Nacional de Energia Elétrica, 2008, 2010).

Estrutura Tarifária conjunto de tarifas aplicáveis às componentes de energia elétrica e/ou demanda de potência ativa de acordo com a modalidade de fornecimento (Agência Nacional de Energia Elétrica, 2008). De acordo com legislação vigente e informações de demanda, de tensão e de setor disponibilizadas pela empresa distribuidora de energia elétrica, os consumidores foram enquadrados como Convencional (Con) ou Horário (Hor).

Intensidade Energética (IE): incidência do consumo de energia elétrica final sobre o PIB (Produto Interno Bruto) de uma classe econômica a nível Brasil. O Quadro A1 apresenta a intensidade energética por classe econômica.

Quadro A1- Consumo de Energia Elétrica por Valor agregado em Diversos Setores da Economia Brasileira

Classe Econômica	Atividades Relacionadas	Dados Ano 2010		
		Bilhões de kWh/ano	Bilhões de US$/ ano	Intensidade Energética (kWh/US$)
Agropecuário	Atividades de agricultura, pecuária, produção florestal, pesca e aquicultura	18,9	87,8	0,22
Comercial e Público	Atividades de comércio, de comunicações, das instituições financeiras, das administrações públicas, dos aluguéis e outros serviços e SIUP (Serviços Industriais de Utilidade Pública – nesse grupo estão incluídos eletricidade, gás e água), exceto geração de energia	106,7	1.298,4	0,08
Energético	Atividades de extração de petróleo, de gás natural e de carvão mineral, refino de petróleo, destilação de álcool, geração de eletricidade e produção de coque	26,8	87,1	0,31
Indústria Extrativa Mineral	Atividades de extração mineral, exceto às relacionadas com extração de petróleo, gás natural e carvão mineral	11,3	23,1	0,49
Indústria de Papel e Celulose	Atividades de fabricação de polpa moldada (produtos obtidos a partir de pastas químicas ou mecânicas e/ou aparas, na forma desejada e para uso específico), polpa, papel, papel-cartão e papelão e de produtos fabricados com papel, papel-cartão ou papelão ondulado, mesmo impressos, desde que a impressão de informação não seja a finalidade principal do produto	19,0	6,7	2,84
Indústria Química	Atividades de fabricação de produtos químicos inorgânicos e orgânicos, resinas e fibras, defensivos agrícolas e desinfetantes domissanitários, tintas e produtos afins, produtos e preparados químicos diversos, exceto atividades de fabricação de sabões, detergentes sintéticos, produtos de limpeza e polimento, produtos cosméticos e de perfumaria e de higiene pessoal, refino de petróleo, destilação de álcool e produção de coque	23,9	30,9	0,77
Indústria Têxtil	Atividades de preparação das fibras têxteis, a fiação e a tecelagem, bem como, a fabricação de tecidos de malha, exceto atividades do vestuário, calçados e artefatos de tecido	8,3	8,7	0,95

Quadro A1- Consumo de Energia Elétrica por Valor agregado em Diversos Setores da Economia (continuação)

Classe Econômica	Atividades Relacionadas	Dados Ano 2010		
		Bilhões de kWh/ano	Bilhões de US$/ ano	Intensidade Energética (kWh/US$)
Outras Indústrias	Atividades relacionadas à mecânica, material elétrico e comunicação, material de transporte, madeira, mobiliário, borracha, farmacêutica, perfumaria, sabões e velas, produção de matérias plásticas, fumo, construções e diversos.	40,4	246,9	0,16
Transportes	Atividades de transporte de passageiros ou mercadorias nas modalidades terrestre, aquaviário e aéreo, bem como, atividades de armazenamento, de correios e outras auxiliares dos transportes e de entrega.	1,7	82,7	0,02

Fonte: Adaptado de Empresa de Pesquisa Energética (2013)

Mesorregião: área composta por unidades político-administrativas que apresentam formas de organização do espaço geográfico distintas construídas ao longo do tempo pela sociedade que nela se formou (Instituto Brasileiro de Geografia e Estatística, 1990). Nesta pesquisa, os municípios atendidos pela empresa distribuidora em estudo foram agregados em regiões cujas características são descritas a seguir:

- **Baixadas** (BA)**:** área de 3.644 km^2, potencialmente turística e composta por 700.842 habitantes distribuídos em dez municípios cuja soma do PIB corresponde a R$ 26.880.963.000.
- **Centro-Sul** (CS)**:** área de 14.764 km^2 com expressiva industrialização e potencialidades turísticas. Possui 1.543.594 habitantes distribuídos em trinta municípios cuja soma do PIB corresponde a R$ 45.278.056.000.
- **Metropolitana** (ME)**:** com 12.578.485 habitantes, 10.256 km^2 e PIB de R$ 346.253.938.000 é uma área muito desenvolvida que compreende a capital dessa Unidade de Federação, na qual municípios do entorno crescem e são dependentes. Região, composta por trinta municípios, que se distingue em muitos aspectos das demais.
- **Noroeste** (NO)**:** área de 5.373 km^2 tradicionalmente agrícola e composta por 317.493 habitantes distribuídos em treze municípios cuja soma do PIB corresponde a R$ 3.968.067.000.
- **Norte** (N)**:** área de 9.746 km^2, tradicionalmente agroindustrial e com expressiva produção de petróleo e gás, composta por 849.515 habitantes distribuídos em nove municípios cuja soma do PIB corresponde a R$ 63.958.290.000.

Natureza jurídica: identificação da constituição jurídico-institucional das entidades públicas e privadas nos cadastros da administração pública do País, gerando um conjunto de direitos e obrigações vinculados ao CNPJ (Instituto Brasileiro de Geografia e Estatística, 2016). Nesta pesquisa, considerou-se

(Chiavenato, 2016; Ferrari, 2003; Instituto Brasileiro de Geografia e Estatística, 2016):

- **Administração Pública e Entidades Sem Fins Lucrativos** (AP&ESFL)**:** enquadram-se nesta categoria administração pública e entidades sem fins lucrativos.
  - **Administração Pública:** órgãos públicos, as autarquias e as fundações públicas da União, dos Estados, do Distrito Federal e dos Municípios.
  - **Entidades sem fins lucrativos:** instituição privada que não distribui, entre os seus sócios ou associados, conselheiros, diretores, empregados ou doadores, eventuais excedentes operacionais, dividendos, bonificações, participações ou parcelas do seu patrimônio, auferidos mediante o exercício de suas atividades, e que os aplica integralmente na consecução do respectivo objeto social.
- **Sociedade empresária limitada** (SEL)**:** sociedade constituída por, no mínimo, dois proprietários que respondem solidariamente pela integração do capital social, em que o capital social é dividido em quotas iguais ou desiguais, cabendo a cada sócio responder ao valor de suas quotas de forma restrita.
- **Outras entidades empresariais** (OOE)**:** enquadram-se nesta categoria cooperativas, consórcios de sociedades, empresas individuais de responsabilidade limitada (de natureza empresarial), empresários (individuais), empresas públicas, sociedades de economia mista, sociedades anônimas e sociedades simples.
  - **Cooperativa:** organização constituída por membros de determinado grupo econômico ou social que objetiva desempenhar, em benefício comum, determinada atividade;
  - **Consórcio de sociedades:** associação de companhias ou qualquer outra sociedade para obter finalidade comum ou determinado empreendimento de custo muito elevado. O consórcio é formalmente instituído por um contrato entre as empresas que o constituem, por isso, não tem personalidade jurídica própria. É válido mencionar que, as consorciadas somente se obrigam nas condições previstas no contrato, de modo que, cada uma responde por suas obrigações, sem presunção de solidariedade.
  - **Empresas Individuais de Responsabilidade Limitada (de natureza empresarial):** empresa constituída por apenas um sócio, o próprio empresário. Para abertura, é necessário, um capital mínimo de 100 vezes o valor do salário-mínimo no momento do registro da empresa, que constituirá o patrimônio empresarial, permitindo a separação entre patrimônio empresarial e privado.
  - **Empresários (individuais):** pessoa física que exerce em nome próprio uma atividade empresarial. O patrimônio da pessoa física e do empresário individual são os mesmos, logo o titular responde de forma ilimitada pelas dívidas.
  - **Empresas Públicas:** entidade de personalidade jurídica de direito privado, com patrimônio próprio e capital exclusivo da União, criado por lei para exploração de atividade econômica que o Governo exerça.

- **Sociedades de Economia Mista:** entidade de personalidade jurídica de direito privado, criada por lei para exploração de atividade econômica, sob a forma de sociedade anônima, cujas ações com direito a voto pertençam em sua maioria à União.
- **Sociedade anônima:** sociedade constituída por, no mínimo, dois proprietários, de natureza eminentemente empresarial, em que o capital social é dividido em ações de igual valor nominal, que são de livre negociabilidade, limitando-se a responsabilidade do acionista ao preço de emissão das ações subscritas ou adquiridas. A sociedade anônima (também denominada companhias) pode ser aberta ou fechada conforme os valores mobiliários de sua emissão estejam ou não adquiridos à negociação no mercado de valores mobiliários (ações, debêntures, partes beneficiárias).
- **Sociedade simples:** sociedade constituída por, no mínimo, dois proprietários que reciprocamente se obrigam a contribuir com bens ou serviços para o exercício de atividade econômica e partilha dos resultados entre si, não tendo por objeto o exercício de atividade própria de empresário. São sociedades formadas por proprietários que exercem profissão intelectual, de natureza científica, literária ou artística, mesmo se contar com auxiliares ou colaboradores, salvo se o exercício da profissão constituir elemento de empresa.

Porte empresarial: tamanho do negócio definido por meio de critérios quantitativos, qualitativos ou ambos, geralmente envolvendo as seguintes variáveis: número de empregados, faturamento, setor de atividade, capital social, lucro, patrimônio líquido, ativo fixo e etc. No Brasil, não existe um único critério adotado, uma vez que governo federal, empresas, institutos, pesquisadores e agências de fomento usam modelos classificatórios distintos para atenderem aos seus objetivos de investigação (Martins, 2014). Nesta pesquisa, a variável porte foi fornecida pela empresa distribuidora de energia elétrica. Os consumidores estão classificados quanto ao porte em: pequeno porte (P), médio porte (M) e grande porte (G).

Peso da Conta no Orçamento: participação da conta de energia no orçamento empresarial. Obtido pela razão entre valor da fatura de energia (R$) e a renda das unidades consumidoras (R$).

Valor da fatura: quantia monetária total (R$) que deve ser paga pelo consumidor à distribuidora, em função do fornecimento de energia elétrica, da conexão e uso do sistema ou da prestação de serviços.

Apêndice B: Rotinas do MatLab para Implementação da Rede Neural Artificial *Self-Organizing Maps*

clear all; close all; clc;

%Normalização base pesquisa

D= som_read_data('basepesquisa.txt'); **%cria estrutura de dados**

sD= som_normalize(D, 'range'); **%normaliza estrutura de dados**

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%Busca tamanho do mapa

sD2= som_read_data('basetreino.txt'); **%cria estrutura de dados**

sM = som_make(sD2); **%cria, inicializa e treina o mapa**

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%Treina mapa 8x2 formato hexagonal com algoritmo de inicialização aleatória

sD2= som_read_data('basetreino.txt'); **%cria estrutura de dados**

sT= som_topol_struct(sD2,'msize',[8 2], 'hexa', 'sheet'); **%cria estrutura do mapa**

sM = som_randinit(sD2, sT); **%cria e inicializa o mapa**

sM = som_seqtrain(sM, sD2, 'gaussian', 'radius', [1 1], 'alpha_ini', 0.1, 'power', 'trainlen', 9000, 'epochs'); **%treina o mapa**

[qe te]=som_quality(sM,sD2); **%qualidade do mapa**

%Treina mapa 8x2 formato retangular com algoritmo de inicialização aleatória

sD2= som_read_data('basetreino.txt'); **%cria estrutura de dados**

sT= som_topol_struct(sD2,'msize',[8 2], 'rect', 'sheet'); **%cria estrutura do mapa**

sM = som_randinit(sD2, sT); **%cria e inicializa o mapa**

sM = som_seqtrain(sM, sD2, 'gaussian', 'radius', [1 1], 'alpha_ini', 0.1, 'power', 'trainlen', 9000, 'epochs'); **%treina o mapa**

[qe te]=som_quality(sM,sD2); **%qualidade do mapa**

%Treina mapa 8x2 formato hexagonal com algoritmo de inicialização linear

sD2= som_read_data('basetreino.txt'); **%cria estrutura de dados**

sT= som_topol_struct(sD2,'msize',[8 2], 'hexa', 'sheet'); **%cria estrutura do mapa**

sM = som_lininit(sD2, sT); **%cria e inicializa o mapa**

sM = som_seqtrain(sM, sD2, 'gaussian', 'radius', [1 1], 'alpha_ini', 0.1, 'power', 'trainlen', 9000, 'epochs'); **%treina o mapa**

```
[qe te]=som_quality(sM,sD2); %qualidade do mapa

%Treina mapa 8x2 formato retangular com algoritmo de inicialização linear
sD2= som_read_data('basetreino.txt'); %cria estrutura de dados
sT= som_topol_struct(sD2,'msize',[8 2], 'rect', 'sheet'); %cria estrutura do mapa
sM = som_lininit(sD2, sT); %cria e inicializa o mapa
sM = som_seqtrain(sM, sD2, 'gaussian', 'radius', [1 1], 'alpha_ini', 0.1, 'power', 'trainlen',
9000, 'epochs'); %treina o mapa
[qe te]=som_quality(sM,sD2); %qualidade do mapa
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%Treina mapa 10x4 formato hexagonal com algoritmo de inicialização aleatória
sD2= som_read_data('basetreino.txt'); %cria estrutura de dados
sT= som_topol_struct(sD2,'msize',[10 4], 'hexa', 'sheet'); %cria estrutura do mapa
sM = som_randinit(sD2, sT); %cria e inicializa o mapa
sM = som_seqtrain(sM, sD2, 'gaussian', 'radius', [1 1], 'alpha_ini', 0.1, 'power', 'trainlen',
21000, 'epochs'); %treina o mapa
[qe te]=som_quality(sM,sD2); %qualidade do mapa

%Treina mapa 10x4 formato retangular com algoritmo de inicialização aleatória
sD2= som_read_data('basetreino.txt'); %cria estrutura de dados
sT= som_topol_struct(sD2,'msize',[10 4], 'rect', 'sheet'); %cria estrutura do mapa
sM = som_randinit(sD2, sT); %cria e inicializa o mapa
sM = som_seqtrain(sM, sD2, 'gaussian', 'radius', [1 1], 'alpha_ini', 0.1, 'power', 'trainlen',
21000, 'epochs'); %treina o mapa
[qe te]=som_quality(sM,sD2); %qualidade do mapa

%Treina mapa 10x4 formato hexagonal com algoritmo de inicialização linear
sD2= som_read_data('basetreino.txt'); %cria estrutura de dados
sT= som_topol_struct(sD2,'msize',[10 4], 'hexa', 'sheet'); %cria estrutura do mapa
sM = som_lininit(sD2, sT); %cria e inicializa o mapa
sM = som_seqtrain(sM, sD2, 'gaussian', 'radius', [1 1], 'alpha_ini', 0.1, 'power', 'trainlen',
21000, 'epochs'); %treina o mapa
[qe te]=som_quality(sM,sD2); %qualidade do mapa

%Treina mapa 10x4 formato retangular com algoritmo de inicialização linear
sD2= som_read_data('basetreino.txt'); %cria estrutura de dados
```

sT= som_topol_struct(sD2,'msize',[10 4], 'rect', 'sheet'); **%cria estrutura do mapa**
sM = som_lininit(sD2, sT); **%cria e inicializa o mapa**
sM = som_seqtrain(sM, sD2, 'gaussian', 'radius', [1 1], 'alpha_ini', 0.1, 'power', 'trainlen', 21000, 'epochs'); **%treina o mapa**
[qe te]=som_quality(sM,sD2); **%qualidade do mapa**
%%

%Treina mapa 12x8 formato hexagonal com algoritmo de inicialização aleatória
sD2= som_read_data('basetreino.txt'); **%cria estrutura de dados**
sT= som_topol_struct(sD2,'msize',[12 8], 'hexa', 'sheet'); **%cria estrutura do mapa**
sM = som_randinit(sD2, sT); **%cria e inicializa o mapa**
sM = som_seqtrain(sM, sD2, 'gaussian', 'radius', [3 1], 'alpha_ini', 0.1, 'power', 'trainlen', 49000, 'epochs'); **%treina o mapa**
[qe te]=som_quality(sM,sD2); **%qualidade do mapa**

%Treina mapa 12x8 formato retangular com algoritmo de inicialização aleatória
sD2= som_read_data('basetreino.txt'); **%cria estrutura de dados**
sT= som_topol_struct(sD2,'msize',[12 8], 'rect', 'sheet'); **%cria estrutura do mapa**
sM = som_randinit(sD2, sT); **%cria e inicializa o mapa**
sM = som_seqtrain(sM, sD2, 'gaussian', 'radius', [3 1], 'alpha_ini', 0.1, 'power', 'trainlen', 49000, 'epochs'); **%treina o mapa**
[qe te]=som_quality(sM,sD2); **%qualidade do mapa**

%Treina mapa 12x8 formato hexagonal com algoritmo de inicialização linear
sD2= som_read_data('basetreino.txt'); **%cria estrutura de dados**
sT= som_topol_struct(sD2,'msize',[12 8], 'hexa', 'sheet'); **%cria estrutura do mapa**
sM = som_lininit(sD2, sT); **%cria e inicializa o mapa**
sM = som_seqtrain(sM, sD2, 'gaussian', 'radius', [3 1], 'alpha_ini', 0.1, 'power', 'trainlen', 49000, 'epochs'); **%treina o mapa**
[qe te]=som_quality(sM,sD2); **%qualidade do mapa**

%Treina mapa 12x8 formato retangular com algoritmo de inicialização linear
sD2= som_read_data('basetreino.txt'); **%cria estrutura de dados**
sT= som_topol_struct(sD2,'msize',[12 8], 'rect', 'sheet'); **%cria estrutura do mapa**
sM = som_lininit(sD2, sT); **%cria e inicializa o mapa**
sM = som_seqtrain(sM, sD2, 'gaussian', 'radius', [3 1], 'alpha_ini', 0.1, 'power', 'trainlen', 49000, 'epochs'); **%treina o mapa**

[qe te]=som_quality(sM,sD2); **%qualidade do mapa**

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%Treina mapa 14x10 formato hexagonal com algoritmo de inicialização aleatória

sD2= som_read_data('basetreino.txt'); **%cria estrutura de dados**

sT= som_topol_struct(sD2,'msize',[14 10], 'hexa', 'sheet'); **%cria estrutura do mapa**

sM = som_randinit(sD2, sT); **%cria e inicializa o mapa**

sM = som_seqtrain(sM, sD2, 'gaussian', 'radius', [4 1], 'alpha_ini', 0.1, 'power', 'trainlen', 71000, 'epochs'); **%treina o mapa**

[qe te]=som_quality(sM,sD2); **%qualidade do mapa**

%Treina mapa 14x10 formato retangular com algoritmo de inicialização aleatória

sD2= som_read_data('basetreino.txt'); **%cria estrutura de dados**

sT= som_topol_struct(sD2,'msize',[14 10], 'rect', 'sheet'); **%cria estrutura do mapa**

sM = som_randinit(sD2, sT); **%cria e inicializa o mapa**

sM = som_seqtrain(sM, sD2, 'gaussian', 'radius', [4 1], 'alpha_ini', 0.1, 'power', 'trainlen', 71000, 'epochs'); **%treina o mapa**

[qe te]=som_quality(sM,sD2); **%qualidade do mapa**

%Treina mapa 14x10 formato hexagonal com algoritmo de inicialização linear

sD2= som_read_data('basetreino.txt'); **%cria estrutura de dados**

sT= som_topol_struct(sD2,'msize',[14 10], 'hexa', 'sheet'); **%cria estrutura do mapa**

sM = som_lininit(sD2, sT); **%cria e inicializa o mapa**

sM = som_seqtrain(sM, sD2, 'gaussian', 'radius', [4 1], 'alpha_ini', 0.1, 'power', 'trainlen', 71000, 'epochs'); **%treina o mapa**

[qe te]=som_quality(sM,sD2); **%qualidade do mapa**

%Treina mapa 14x10 formato retangular com algoritmo de inicialização linear

sD2= som_read_data('basetreino.txt'); **%cria estrutura de dados**

sT= som_topol_struct(sD2,'msize',[14 10], 'rect', 'sheet'); **%cria estrutura do mapa**

sM = som_lininit(sD2, sT); **%cria e inicializa o mapa**

sM = som_seqtrain(sM, sD2, 'gaussian', 'radius', [4 1], 'alpha_ini', 0.1, 'power', 'trainlen', 71000, 'epochs'); **%treina o mapa**

[qe te]=som_quality(sM,sD2); **%qualidade do mapa**

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%Treina mapa 16x12 formato hexagonal com algoritmo de inicialização aleatória

sD2= som_read_data('basetreino.txt'); **%cria estrutura de dados**

sT= som_topol_struct(sD2,'msize',[16 12], 'hexa', 'sheet'); **%cria estrutura do mapa**

sM = som_randinit(sD2, sT); **%cria e inicializa o mapa**

sM = som_seqtrain(sM, sD2, 'gaussian', 'radius', [5 1], 'alpha_ini', 0.1, 'power', 'trainlen', 97000, 'epochs'); **%treina o mapa**

[qe te]=som_quality(sM,sD2); **%qualidade do mapa**

%Treina mapa 16x12 formato retangular com algoritmo de inicialização aleatória

sD2= som_read_data('basetreino.txt'); **%cria estrutura de dados**

sT= som_topol_struct(sD2,'msize',[16 12], 'rect', 'sheet'); **%cria estrutura do mapa**

sM = som_randinit(sD2, sT); **%cria e inicializa o mapa**

sM = som_seqtrain(sM, sD2, 'gaussian', 'radius', [5 1], 'alpha_ini', 0.1, 'power', 'trainlen', 97000, 'epochs'); **%treina o mapa**

[qe te]=som_quality(sM,sD2); **%qualidade do mapa**

%Treina mapa 16x12 formato hexagonal com algoritmo de inicialização linear

sD2= som_read_data('basetreino.txt'); **%cria estrutura de dados**

sT= som_topol_struct(sD2,'msize',[16 12], 'hexa', 'sheet'); **%cria estrutura do mapa**

sM = som_lininit(sD2, sT); **%cria e inicializa o mapa**

sM = som_seqtrain(sM, sD2, 'gaussian', 'radius', [5 1], 'alpha_ini', 0.1, 'power', 'trainlen', 97000, 'epochs'); **%treina o mapa**

[qe te]=som_quality(sM,sD2); **%qualidade do mapa**

%Treina mapa 16x12 formato retangular com algoritmo de inicialização linear

sD2= som_read_data('basetreino.txt'); **%cria estrutura de dados**

sT= som_topol_struct(sD2,'msize',[16 12], 'rect', 'sheet'); **%cria estrutura do mapa**

sM = som_lininit(sD2, sT); **%cria e inicializa o mapa**

sM = som_seqtrain(sM, sD2, 'gaussian', 'radius', [5 1], 'alpha_ini', 0.1, 'power', 'trainlen', 97000, 'epochs'); **%treina o mapa**

[qe te]=som_quality(sM,sD2); **%qualidade do mapa**

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%Saída SOM

h=som_hits(sM,sD2); **%densidade dos neurônios**

U=som_umat(sM); **%Matriz de distância (U-matriz)**

Um=U(1:2:size(U,1),1:2:size(U,2)); **%U-matriz média**

Bmus= som_bmus(sM,sD2); **%encontra neurônio vencedor para cada vetor em sD**

%projeção dos dados sobre o espaço gerado por componentes principais

[Pd,V,me]=pcaproj(sD2.data,2); **%projeta dados**

Pm=pcaproj(sM.codebook,V,me); **%projeta protótipos**

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%Visualização SOM

som_show(sM,'umat','all';'comp',[1:5]); **%matriz de distância unificada e por atributos**

sM=som_autolabel(sM,sD2,'vote'); **%define forma de atribuição do rótulo**

som_show(sM, 'empty', 'Rótulo','footnote','')**%estrutura SOM**

som_show_add('label',sM,'Textsize', 8, 'Textcolor', 'b') **%rótulo do mapa**

som_show(sM, 'empty', 'Densidade','footnote','')**%estrutura SOM**

som_show_add('hit',h,'MarkerColor','b', 'Subplot', 1)**%densidade neurônios**

som_show_add('hit',h,'Text', 'on', 'Textcolor','k')**%densidade neurônios**

som_show_clear

C1=som_colorcode(Pm);

som_cplane(sM,C1);**%similaridade por cor**

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%%Inclusão de vetor após treino

sD3= som_read_data('baseteste.txt'); **%cria estrutura de dados**

[bmus2, qerrs] =som_bmus(sM, sD3); **% encontra BMUs (primeiro, segundo e terceiro neurônio vencedor) para um determinado conjunto de dados no mapa gerado (sM), bem como o correspondente erro de quantização**

%which, quais tipos de BMUs são retornados. Por padrão, é o vencedor.

%mask, vetor peso para ser usado na pesquisa do BMU. Por padrão, é sMap.mask ou, no caso, de uma matriz ser dada, algum (dim, 1). A máscara pode ser utilizada para ponderar o processo de pesquisa, isto é, para ponderar a influência dos componentes do cálculo da distância.

Apêndice C: Variáveis Categóricas dos Consumidores do Conjunto Teste

Quadro C1- Variáveis Categóricas dos Consumidores da Base Teste

Consumidor	A. econômica	N. jurídica	Mesorregião	Porte	E. Tarifaria	Grupo	Perfil Real
1601	CA	OEE	N	P	Hor	G1	Inadimplente
1602	ES	SEL	ME	M	Hor	G3	Inadimplente
1603	AFIA	OEE	ME	M	Hor	G3	Adimplente
1604	AFIA	SEL	N	P	Con	G2	Inadimplente
1605	IMMED	SEL	ME	G	Hor	G4	Inadimplente
1606	CV	SEL	BA	M	Con	G3	Adimplente
1607	CV	SEL	ME	M	Hor	G3	Adimplente
1608	IABA	OEE	ME	P	Con	G2	Adimplente
1609	AAT	SEL	ME	P	Con	G3	Adimplente
1610	AAT	SEL	ME	P	Con	G3	Adimplente
1611	AAT	SEL	N	P	Con	G2	Adimplente
1612	AAT	SEL	CS	P	Con	G2	Inadimplente
1613	CA	SEL	ME	M	Hor	G3	Adimplente
1614	CO	SEL	ME	M	Hor	G4	Inadimplente
1615	AAT	SEL	ME	M	Hor	G1	Inadimplente
1616	INM	SEL	BA	P	Hor	G5	Inadimplente
1617	IMMED	SEL	ME	P	Hor	G4	Inadimplente
1618	INM	SEL	ME	P	Hor	G5	Adimplente
1619	AFIA	OEE	ME	M	Hor	G3	Adimplente
1620	INM	SEL	BA	P	Hor	G5	Inadimplente
1621	SD	SEL	BA	M	Con	G2	Inadimplente
1622	SD	SEL	ME	P	Hor	G3	Inadimplente
1623	ID	SEL	ME	M	Hor	G4	Inadimplente
1624	AAT	SEL	ME	P	Con	G3	Inadimplente
1625	AFIA	SEL	ME	M	Hor	G4	Inadimplente
1626	AAT	OEE	ME	M	Hor	G3	Inadimplente
1627	AFIA	AP&ESFL	BA	M	Con	G3	Adimplente
1628	CV	SEL	ME	P	Hor	G3	Adimplente
1629	ID	SEL	ME	P	Hor	G4	Inadimplente
1630	CO	OEE	NO	M	Con	G4	Inadimplente
1631	SD	OEE	CS	M	Con	G2	Inadimplente
1632	CV	SEL	BA	M	Con	G3	Adimplente
1633	AFIA	AP&ESFL	ME	M	Con	G3	Inadimplente
1634	IMMED	SEL	N	G	Hor	G4	Adimplente
1635	ES	OEE	ME	M	Con	G3	Inadimplente
1636	AFIA	AP&ESFL	CS	M	Con	G2	Adimplente
1637	AAT	SEL	CS	P	Con	G2	Adimplente
1638	IABA	SEL	ME	M	Hor	G5	Inadimplente
1639	AFIA	AP&ESFL	BA	M	Con	G3	Inadimplente
1640	AAT	SEL	CS	P	Con	G2	Inadimplente
1641	IABA	SEL	ME	M	Hor	G5	Adimplente
1642	SD	SEL	ME	M	Hor	G3	Adimplente
1643	CV	SEL	NO	P	Con	G3	Inadimplente
1644	CV	SEL	ME	P	Con	G3	Adimplente
1645	ES	OEE	ME	P	Con	G3	Inadimplente
1646	AFIA	OEE	CS	M	Con	G3	Inadimplente
1647	SD	AP&ESFL	ME	M	Con	G2	Inadimplente
1648	SD	OEE	ME	G	Hor	G4	Adimplente
1649	ES	AP&ESFL	NO	M	Con	G2	Adimplente
1650	AAT	SEL	ME	P	Con	G2	Inadimplente
1651	ES	AP&ESFL	ME	M	Con	G2	Inadimplente
1652	CV	SEL	NO	P	Con	G3	Adimplente

Quadro C1- Variáveis Categóricas dos Consumidores da Base Teste (continuação)

Consumidor	A. econômica	N. jurídica	Mesorregião	Porte	E. Tarifaria	Grupo	Perfil Real
1653	CV	SEL	ME	M	Hor	G3	Adimplente
1654	AAT	SEL	ME	P	Hor	G3	Adimplente
1655	AAT	OEE	BA	M	Con	G1	Inadimplente
1656	ID	SEL	ME	P	Hor	G4	Inadimplente
1657	AAT	SEL	N	P	Con	G2	Inadimplente
1658	CV	SEL	ME	P	Con	G3	Inadimplente
1659	ES	AP&ESFL	CS	M	Con	G3	Inadimplente
1660	ES	AP&ESFL	ME	M	Con	G3	Adimplente
1661	CA	SEL	BA	P	Con	G3	Adimplente
1662	AFIA	SEL	BA	P	Con	G2	Adimplente
1663	ID	SEL	N	M	Hor	G4	Inadimplente
1664	CV	SEL	ME	P	Hor	G3	Adimplente
1665	ID	SEL	ME	P	Hor	G4	Adimplente
1666	CV	SEL	ME	M	Hor	G4	Inadimplente
1667	ES	AP&ESFL	ME	M	Con	G2	Adimplente
1668	AAT	SEL	BA	P	Con	G3	Inadimplente
1669	ID	SEL	ME	P	Hor	G2	Inadimplente
1670	AAT	SEL	ME	P	Hor	G3	Inadimplente
1671	CV	SEL	N	P	Hor	G3	Adimplente
1672	AFIA	AP&ESFL	ME	G	Hor	G4	Inadimplente
1673	ID	OEE	BA	G	Hor	G5	Adimplente
1674	INM	SEL	ME	P	Hor	G5	Inadimplente
1675	ES	SEL	N	M	Con	G2	Inadimplente
1676	CV	SEL	BA	M	Hor	G4	Inadimplente
1677	CV	SEL	ME	M	Hor	G3	Adimplente
1678	IABA	SEL	ME	P	Hor	G5	Inadimplente
1679	AAT	SEL	CS	P	Con	G3	Adimplente
1680	AFIA	OEE	ME	M	Hor	G4	Inadimplente
1681	INM	SEL	N	P	Hor	G5	Inadimplente
1682	AAT	SEL	BA	P	Hor	G3	Inadimplente
1683	ES	OEE	ME	M	Hor	G3	Adimplente
1684	AFIA	SEL	N	M	Hor	G3	Inadimplente
1685	CV	SEL	CS	P	Con	G3	Inadimplente
1686	AAT	SEL	N	M	Con	G3	Inadimplente
1687	AFIA	SEL	ME	M	Con	G2	Inadimplente
1688	IABA	OEE	ME	M	Hor	G5	Inadimplente
1689	ES	OEE	ME	P	Con	G3	Inadimplente
1690	CV	SEL	CS	M	Hor	G3	Inadimplente
1691	CO	SEL	BA	M	Con	G4	Inadimplente
1692	CA	OEE	ME	M	Hor	G3	Adimplente
1693	AAT	SEL	ME	P	Hor	G3	Inadimplente
1694	CV	SEL	ME	M	Hor	G3	Adimplente
1695	IMMED	SEL	ME	M	Con	G4	Adimplente
1696	CV	SEL	ME	M	Hor	G3	Inadimplente
1697	AAT	SEL	NO	P	Hor	G3	Inadimplente
1698	CV	SEL	N	P	Hor	G3	Adimplente
1699	CA	SEL	ME	M	Hor	G3	Adimplente
1700	CV	SEL	ME	M	Hor	G3	Adimplente
1701	CA	SEL	N	M	Hor	G4	Inadimplente
1702	CV	SEL	BA	M	Hor	G3	Inadimplente
1703	ES	SEL	ME	P	Con	G2	Adimplente
1704	CA	SEL	N	M	Con	G2	Inadimplente
1705	CV	SEL	ME	P	Con	G3	Inadimplente
1706	CV	SEL	N	M	Con	G2	Adimplente
1707	INM	SEL	ME	P	Hor	G5	Adimplente
1708	CO	OEE	ME	M	Con	G2	Inadimplente
1709	ID	SEL	N	M	Con	G4	Adimplente
1710	CA	OEE	ME	M	Hor	G3	Adimplente
1711	AAT	OEE	ME	M	Con	G1	Adimplente
1712	CV	SEL	ME	P	Hor	G3	Inadimplente
1713	AAT	SEL	ME	P	Hor	G3	Inadimplente

Quadro C1- Variáveis Categóricas das Unidades Consumidoras da Base Teste (continuação)

Consumidor	A. econômica	N. jurídica	Mesorregião	Porte	E. Tarifaria	Grupo	Perfil Real
1714	ID	SEL	ME	P	Hor	G4	Inadimplente
1715	ID	SEL	NO	G	Hor	G4	Adimplente
1716	CV	SEL	N	M	Hor	G3	Adimplente
1717	ES	AP&ESFL	ME	M	Hor	G3	Inadimplente
1718	ES	AP&ESFL	ME	M	Con	G3	Inadimplente
1719	ES	SEL	ME	P	Con	G2	Inadimplente
1720	ID	SEL	ME	G	Hor	G5	Inadimplente
1721	CA	OEE	ME	P	Con	G2	Inadimplente
1722	AAT	OEE	N	M	Con	G1	Inadimplente
1723	CV	OEE	ME	M	Con	G3	Inadimplente
1724	ES	OEE	ME	P	Hor	G3	Inadimplente
1725	INM	SEL	N	P	Hor	G5	Adimplente
1726	ES	SEL	N	M	Con	G2	Adimplente
1727	ID	OEE	CS	G	Hor	G4	Inadimplente
1728	IMMED	SEL	N	P	Con	G2	Inadimplente
1729	CA	OEE	NO	M	Con	G2	Inadimplente
1730	CV	SEL	ME	P	Hor	G3	Adimplente
1731	SD	SEL	N	M	Con	G2	Inadimplente
1732	AFIA	AP&ESFL	ME	M	Con	G3	Adimplente
1733	AFIA	AP&ESFL	N	G	Hor	G3	Inadimplente
1734	SD	OEE	ME	M	Hor	G3	Inadimplente
1735	CV	SEL	ME	M	Con	G3	Adimplente
1736	AFIA	AP&ESFL	ME	M	Con	G3	Adimplente
1737	AAT	SEL	ME	P	Hor	G3	Inadimplente
1738	SD	AP&ESFL	ME	M	Hor	G3	Inadimplente
1739	AAT	SEL	ME	P	Con	G1	Adimplente
1740	IMMED	SEL	ME	M	Con	G4	Inadimplente
1741	IABA	SEL	ME	P	Hor	G5	Inadimplente
1742	ES	AP&ESFL	ME	M	Con	G4	Adimplente
1743	AAT	SEL	ME	P	Hor	G3	Adimplente
1744	CV	SEL	ME	P	Hor	G3	Adimplente
1745	CV	SEL	BA	M	Con	G3	Adimplente
1746	INM	SEL	ME	P	Hor	G5	Inadimplente
1747	SD	AP&ESFL	ME	M	Con	G2	Inadimplente
1748	ES	OEE	ME	M	Con	G3	Inadimplente
1749	ES	AP&ESFL	ME	M	Con	G2	Inadimplente
1750	CV	SEL	ME	P	Hor	G1	Inadimplente
1751	AAT	OEE	ME	M	Con	G1	Adimplente
1752	ES	OEE	N	M	Hor	G3	Inadimplente
1753	IABA	SEL	ME	G	Hor	G5	Inadimplente
1754	AFIA	SEL	ME	M	Con	G2	Inadimplente
1755	IABA	OEE	N	P	Hor	G5	Inadimplente
1756	AFIA	AP&ESFL	BA	M	Con	G3	Inadimplente
1757	CA	SEL	ME	P	Con	G3	Inadimplente
1758	CV	SEL	ME	M	Hor	G3	Adimplente
1759	ES	SEL	ME	P	Con	G2	Inadimplente
1760	CA	SEL	ME	M	Con	G3	Adimplente
1761	ID	SEL	ME	P	Hor	G2	Adimplente
1762	CO	SEL	ME	P	Con	G2	Inadimplente
1763	AFIA	AP&ESFL	ME	M	Con	G3	Inadimplente
1764	IMMED	SEL	N	M	Hor	G4	Adimplente
1765	AFIA	AP&ESFL	CS	M	Con	G3	Adimplente
1766	CV	SEL	ME	M	Hor	G3	Inadimplente
1767	AFIA	AP&ESFL	ME	M	Con	G3	Inadimplente
1768	ES	SEL	CS	P	Hor	G2	Inadimplente
1769	ID	SEL	ME	P	Hor	G4	Inadimplente
1770	AFIA	OEE	ME	M	Con	G3	Inadimplente
1771	AAT	SEL	CS	P	Hor	G3	Inadimplente
1772	SD	OEE	CS	M	Con	G2	Adimplente
1773	ES	SEL	ME	M	Hor	G2	Adimplente
1774	IMMED	SEL	ME	M	Hor	G4	Adimplente

Quadro C1- Variáveis Categóricas das Unidades Consumidoras da Base Teste (continuação)

Consumidor	A. econômica	N. jurídica	Mesorregião	Porte	E. Tarifaria	Grupo	Perfil Real
1775	ES	AP&ESFL	ME	M	Con	G2	Inadimplente
1776	AAT	SEL	CS	P	Con	G2	Inadimplente
1777	INM	SEL	CS	G	Hor	G4	Adimplente
1778	ES	OEE	CS	P	Con	G2	Inadimplente
1779	CV	SEL	BA	P	Hor	G1	Inadimplente
1780	AAT	SEL	BA	P	Con	G2	Inadimplente
1781	AAT	SEL	ME	P	Con	G2	Inadimplente
1782	ES	SEL	NO	P	Con	G3	Inadimplente
1783	IMMED	SEL	ME	P	Hor	G5	Adimplente
1784	SD	AP&ESFL	ME	M	Con	G2	Inadimplente
1785	ES	SEL	ME	P	Hor	G3	Inadimplente
1786	ES	AP&ESFL	N	M	Con	G2	Adimplente
1787	AFIA	AP&ESFL	BA	M	Con	G2	Adimplente
1788	IMMED	SEL	ME	P	Hor	G5	Adimplente
1789	AAT	SEL	BA	P	Hor	G1	Inadimplente
1790	CV	SEL	ME	M	Hor	G1	Adimplente
1791	CV	SEL	ME	P	Con	G2	Adimplente
1792	AAT	SEL	CS	M	Con	G3	Inadimplente
1793	CV	SEL	BA	P	Hor	G3	Inadimplente
1794	CV	SEL	BA	P	Hor	G3	Adimplente
1795	SD	AP&ESFL	ME	M	Con	G2	Inadimplente
1796	CV	SEL	ME	M	Hor	G3	Adimplente
1797	CV	SEL	ME	M	Con	G3	Adimplente
1798	CV	SEL	NO	P	Hor	G3	Adimplente
1799	ES	OEE	ME	P	Con	G2	Adimplente
1800	CV	SEL	BA	M	Con	G3	Inadimplente
1801	AAT	SEL	CS	P	Con	G2	Adimplente
1802	CO	SEL	BA	M	Hor	G4	Inadimplente
1803	CA	SEL	NO	M	Hor	G3	Adimplente
1804	SD	AP&ESFL	ME	M	Con	G3	Inadimplente
1805	AAT	SEL	BA	P	Hor	G3	Adimplente
1806	AFIA	AP&ESFL	CS	M	Hor	G3	Inadimplente
1807	CV	SEL	CS	P	Hor	G3	Inadimplente
1808	INM	OEE	N	P	Hor	G5	Inadimplente
1809	AAT	SEL	N	P	Hor	G3	Inadimplente
1810	IABA	SEL	N	M	Hor	G5	Inadimplente
1811	ES	AP&ESFL	ME	M	Con	G3	Adimplente
1812	SD	SEL	N	M	Hor	G3	Inadimplente
1813	IABA	SEL	N	P	Hor	G4	Inadimplente
1814	CV	SEL	ME	P	Con	G3	Inadimplente
1815	AAT	SEL	BA	P	Con	G3	Adimplente
1816	AAT	SEL	ME	P	Hor	G3	Adimplente
1817	SD	AP&ESFL	ME	M	Hor	G3	Adimplente
1818	AAT	SEL	ME	P	Hor	G3	Inadimplente
1819	AFIA	OEE	ME	G	Hor	G3	Inadimplente
1820	CV	SEL	ME	M	Con	G3	Adimplente
1821	CV	SEL	CS	M	Con	G3	Inadimplente
1822	AAT	SEL	ME	M	Hor	G1	Adimplente
1823	SD	SEL	ME	P	Hor	G3	Inadimplente
1824	CA	SEL	ME	M	Hor	G3	Adimplente
1825	IMMED	SEL	ME	P	Con	G2	Inadimplente
1826	IABA	SEL	ME	P	Con	G4	Inadimplente
1827	INM	SEL	N	P	Hor	G5	Adimplente
1828	ID	SEL	ME	P	Hor	G4	Inadimplente
1829	ES	OEE	ME	G	Con	G4	Inadimplente
1830	INM	SEL	ME	M	Hor	G5	Inadimplente
1831	CV	SEL	BA	P	Con	G3	Adimplente
1832	CV	SEL	BA	M	Con	G3	Inadimplente
1833	CV	SEL	ME	M	Hor	G3	Adimplente
1834	SD	SEL	N	M	Hor	G3	Inadimplente
1835	CV	SEL	ME	M	Con	G3	Inadimplente

Quadro C1- Variáveis Categóricas das Unidades Consumidoras da Base Teste (continuação)

Consumidor	A. econômica	N. jurídica	Mesorregião	Porte	E. Tarifaria	Grupo	Perfil Real
1836	AAT	SEL	N	P	Con	G3	Adimplente
1837	ES	AP&ESFL	NO	M	Con	G3	Adimplente
1838	ES	SEL	ME	M	Hor	G3	Inadimplente
1839	SD	AP&ESFL	ME	M	Hor	G3	Inadimplente
1840	AAT	SEL	NO	P	Con	G1	Adimplente
1841	IMMED	SEL	ME	M	Hor	G4	Adimplente
1842	CA	SEL	CS	P	Hor	G3	Inadimplente
1843	CV	SEL	ME	M	Hor	G3	Inadimplente
1844	CV	SEL	ME	P	Con	G3	Inadimplente
1845	SD	OEE	BA	G	Hor	G4	Inadimplente
1846	ID	SEL	N	M	Hor	G4	Inadimplente
1847	CV	SEL	ME	M	Hor	G3	Inadimplente
1848	CV	SEL	ME	M	Hor	G3	Adimplente
1849	IABA	SEL	NO	M	Hor	G5	Inadimplente
1850	AAT	SEL	NO	P	Con	G3	Inadimplente
1851	IABA	SEL	ME	P	Hor	G5	Adimplente
1852	AAT	SEL	ME	P	Con	G2	Adimplente
1853	CV	SEL	ME	P	Hor	G3	Inadimplente
1854	AAT	SEL	BA	P	Con	G2	Adimplente
1855	ID	OEE	CS	G	Con	G2	Adimplente
1856	CV	SEL	ME	M	Hor	G4	Inadimplente
1857	INM	SEL	N	M	Hor	G5	Inadimplente
1858	ID	SEL	ME	M	Hor	G4	Inadimplente
1859	IMMED	SEL	ME	P	Con	G2	Adimplente
1860	CA	SEL	ME	P	Hor	G2	Inadimplente
1861	AAT	OEE	N	M	Con	G1	Inadimplente
1862	CV	SEL	BA	P	Con	G3	Inadimplente
1863	INM	SEL	N	P	Hor	G5	Adimplente
1864	INM	SEL	N	P	Hor	G5	Inadimplente
1865	ES	AP&ESFL	ME	M	Con	G3	Adimplente
1866	CV	SEL	ME	M	Hor	G1	Inadimplente
1867	CV	SEL	NO	P	Hor	G1	Adimplente
1868	AAT	SEL	ME	M	Hor	G1	Inadimplente
1869	CV	SEL	ME	P	Hor	G1	Adimplente
1870	SD	SEL	ME	M	Con	G2	Adimplente
1871	AFIA	AP&ESFL	ME	M	Hor	G3	Inadimplente
1872	AAT	SEL	ME	M	Con	G1	Inadimplente
1873	ES	SEL	ME	P	Con	G2	Adimplente
1874	SD	AP&ESFL	ME	M	Con	G2	Adimplente
1875	AAT	SEL	BA	P	Hor	G3	Adimplente
1876	AAT	SEL	BA	P	Con	G2	Inadimplente

Veja o que significa cada abreviação no Apêndice A.

Fonte: Elaboração própria