UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

# DETECÇÃO DE FRAUDES NO CONSUMO DE ENERGIA ELÉTRICA USANDO ÁRVORES DE DECISÃO

YASMIN CHRISTINE CORREA MATOS

DM 28/2017

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2017

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

YASMIN CHRISTINE CORREA MATOS

# DETECÇÃO DE FRAUDES NO CONSUMO DE ENERGIA ELÉTRICA USANDO ÁRVORE DE DECISÃO

DM 28/2017

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2017

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

YASMIN CHRISTINE CORREA MATOS

# DETECÇÃO DE FRAUDES NO CONSUMO DE ENERGIA ELÉTRICA USANDO ÁRVORE DE DECISÃO

Dissertação submetida à Banca Examinadora do Programa de Pós-Graduação em Engenharia Elétrica da UFPA para a obtenção do Grau de Mestre em Engenharia Elétrica.

Área de Concentração: Sistemas de Energia Elétrica.

Orientador: João Paulo Abreu Vieira

UFPA / ITEC / PPGEE
Campus Universitário do Guamá
Belém-Pará-Brasil
2017

Dados Internacionais de Catalogação - na – Publicação (CIP) Sistema de Bibliotecas da UFPA

---

Matos, Yasmin Christine Correa, 1991 -

Detecção de fraudes em unidades consumidoras de energia elétrica usando árvores de decisão / Yasmin Christine Corrêa Matos .-2017.

Orientador : João Paulo Abreu Vieira

Dissertação (Mestrado) - Universidade Federal do Pará, Instituto de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica, Belém, 2016.

1. Energia elétrica – consumo - custos. 2. Energia elétrica – crime fiscal. 3. Energia elétrica - distribuição. 4. Mineração de dados. I. Título.

CDD 23. ed. 333.7932

---

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

# "DETECÇÃO DE FRAUDES EM UNIDADES CONSUMIDORAS DE ENERGIA ELÉTRICA USANDO ÁRVORES DE DECISÃO"

AUTOR: **VASMIN CHRISTINE CORREA MATOS**

DISSERTAÇÃO DE MESTRADO SUBMETIDA À BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA, SENDO JULGADA ADEQUADA PARA A OBTENÇÃO DO GRAU DE MESTRE EM ENGENHARIA ELÉTRICA NA ÁREA DE SISTEMAS DE ENERGIA ELÉTRICA.

APROVADA EM: 11/07/2017

**BANCA EXAMINADORA:**

**Prof. Dr. João Paulo Abreu Vieira**
(Orientador – PPGEE/UFPA)

**Prof. Dr. Marcus Vinicius Alves Nunes**
(Avaliador Interno – PPGEE/UFPA)

**Prof.ª Dr.ª Maria Emília de Lima Tostes**
(Avaliadora Interna – PPGCC/UFPA)

**Prof. Dr. Filipe de Oliveira Saraiva**
(Avaliador Externo ao Programa – PPGCC/UFPA)

**VISTO:**

**Prof. Dr. Evaldo Gonçalves Pelaes**
(Coordenador do PPGEE/ITEC/UFPA)

# DEDICATÓRIA

A minha família.

# AGRADECIMENTOS

A Deus, pela capacitação e pela presença constante em minha vida.

A minha família, pelo amor e apoio para a realização deste trabalho.

Ao orientador João Paulo e a equipe do projeto, pela oportunidade de participar deste projeto P&D.

Ao Rodrigo e a Flávia, pelas incontáveis ajudas e imensa disposição para ajudar na realização deste trabalho.

# SUMÁRIO

# LISTA DE FIGURAS

# LISTA DE TABELAS

# LISTA DE ABREVIAÇÕES

| | |
|---|---|
| **ANEEL:** | Agência Nacional de Energia Elétrica |
| **ABRADEE:** | Associação Brasileira de Distribuidoras de Energia Elétrica |
| **CELPA:** | Centrais Elétricas do Estado do Pará |
| **DT:** | *Decision Tree* |
| **EBITDA:** | *Earnings Before Interest, Taxes, Depreciation and Amortization* |
| **IEEE:** | *Institute of Electric and Electronic Engineers* |
| **KDD:** | *Knowledge Discovery in Database* |
| **PRODIST:** | Procedimentos de Distribuição de Energia Elétrica no Sistema Elétrico Nacional |
| **SISGPQ:** | Sistema de Gestão de Perdas e Qualidade de Energia Elétrica |

# RESUMO

Os prejuízos causados nos últimos anos pelas perdas comerciais às concessionárias de distribuição de energia elétrica no Brasil têm sido estimados aproximadamente em R$ 7 bilhões. Essa realidade representa, um desafio para algumas das distribuidoras do país, as quais necessitam de medidas eficazes no combate às perdas comerciais. Neste cenário, a presente dissertação de mestrado, apresenta uma metodologia capaz de detectar fraudes no consumo de energia elétrica, usando uma técnica de mineração de dados, conhecida como árvore de decisão. Testes de desempenho do método foram realizados usando dados reais do histórico de consumo de energia elétrica e de fiscalização de irregularidades em unidades consumidoras (UC's) da região metropolitana de Belém. Os resultados mostraram que o método proposto baseado em árvore de decisão possui bom desempenho na detecção de fraudes no consumo de energia elétrica.

**PALAVRAS-CHAVES:** Perdas comerciais de energia elétrica, Distribuição de energia elétrica, mineração de dados, árvore de decisão.

# ABSTRACT

In recent years, the injury caused by the nontechnical losses to power distribution utilities, in Brazil have been estimated at R$ 7 billion per year. This reality represents a challenge for some of country's utilities, who need effective measures to combat commercial losses. In this scenario, this dissertation presents a methodology able of detecting fraud in the consumption of electric energy, using a technique of data mining, known as decision tree. Performance tests of the method were done using real data from the history of electricity consumption and the inspection of consumer units (CU's) suspected of being irregular in the metropolitan region of Belém. The results showed that the proposed decision-tree based method performs well in the detection of fraud in the electric power consumption.

**KEY-WORDS:** Nontechnical loss, power distribution, data mining, decision tree.

# 1 INTRODUÇÃO

## 1.1 CONSIDERAÇÕES INICIAIS

No Brasil, o faturamento das concessionárias de distribuição de energia elétrica ainda é fortemente afetado com a prática de fraudes no consumo da energia elétrica. O nível de perdas comerciais em várias distribuidoras do país chega a atingir patamares alarmantes, causando prejuízos não somente as distribuidoras, mas também ao governo, que por sua vez deixa de arrecadar impostos e aos consumidores adimplentes, que acabam pagando pela energia furtada, como forma de compensar o montante desviado.

O problema do furto de energia é que ele se tornou um ciclo vicioso, pois gera maior custo na compra da energia, menor número de UC's pagantes, menos tributos recolhidos, logo não há racionalidade no consumo, exigindo a construção de mais usinas que são cada vez mais caras. Enquanto as perdas comerciais atingem percentuais próximos a 9% da energia consumida no mundo, segundo Simão (2012), no Brasil, o nível de perdas está em torno de 13% que em cifras, corresponde a 7 bilhões de reais, sendo que mais de 60% desse valor estratosférico não é faturado pelas distribuidoras da região norte do país (LEAL, 2012).

Para as distribuidoras, a redução das perdas comerciais implica, de maneira geral, em aumento de receita, porém, há casos em que a complexidade socioeconômica da área de concessão afeta de forma relevante a eficácia no combate às perdas, normalmente em virtude da falta de ação do poder público.

A identificação das unidades consumidoras (UC's) com comportamento fraudulento ou problemas em medição é uma tarefa complexa e atualmente, é realizada por uma pré-análise ineficiente sobre o comportamento dos clientes, uma vez que envolve inspeções *in loco*, as quais são feitas aleatoriamente, ou a partir da experiência do responsável. A razão entre o número de fraudes detectadas e o número de inspeções realizadas é inferior a 10% (FILHO, 2006), resultando em baixas taxas de acertos e alto custo, quando tal abordagem de identificação é adotada.

A mineração de dados ou (do inglês, *data mining*), surge como uma ferramenta promissora para melhorar a confiabilidade das inspeções *in loco* de UC's suspeitas de fraudes no consumo de energia elétrica. O processo de *data mining* tem como objetivo a descoberta

de conhecimentos valiosos em grandes bases de dados, como por exemplo a detecção de fraudes em cartões de crédito. A árvore de decisão é uma das técnicas de mineração de dados, que realiza o aprendizado através da tarefa de classificação. É uma representação simples do conhecimento que constrói classificadores que predizem classes baseadas nos valores de atributos de um conjunto de dados (CORREA, 2015).

O grande volume de dados que é gerado diariamente nas empresas do setor elétrico é pouco ou quase não explorado para extrair informações úteis ao processo de tomada de decisões ou auxílio na solução de problemas. As concessionárias de energia elétrica possuem um grande volume de dados com informações sobre os clientes, como por exemplo, o histórico do consumo, que podem ser transformados em conhecimentos relevantes no combate às perdas comerciais de energia elétrica.

## 1.2 OBJETIVO GERAL

O objetivo geral deste trabalho é abordar o problema das perdas comerciais de energia elétrica e apresentar uma metodologia capaz de detectar fraudes no consumo de energia elétrica em redes de distribuição.

## 1.3 OBJETIVO ESPECÍFICO

O objetivo específico deste trabalho consiste no uso de uma metodologia de mineração de dados, mais especificamente a árvore de decisão, para detectar fraudes no consumo de energia elétrica, através da base de dados da região metropolitana de Belém (RMB), fornecida pela CELPA.

## 1.4 REVISÃO BIBLIOGRÁFICA

Fraudes e desvios no consumo de energia representam um problema grave para as distribuidoras de energia elétrica. Avanços nas diferentes técnicas e metodologias relacionadas com a Inteligência Artificial permitiram a detecção, classificação, redução e previsão desses problemas.

Nesta seção são apresentados os principais artigos relativos a aplicação da mineração de dados na detecção de fraudes do consumo de energia elétrica em concessionárias de distribuição de energia elétrica.

Em 2010, Nagi, J. et. all., publicaram um artigo sobre a aplicação de uma técnica de classificação de padrões, a fim de detectar e identificar padrões de consumo de energia elétrica em clientes fraudadores. Os autores utilizaram a técnica *Support Vector Machine (SVM)* para a tarefa de classificação usando dados do histórico de consumo dos clientes. A SVM utiliza como informação o perfil de carga do cliente e atributos adicionais para expor um comportamento anormal que é conhecido por ser altamente correlacionado com atividades anormais de perdas não técnicas. Os autores mostram que a técnica utilizada é viável e muito promissora.

Em 2011, em um artigo publicado por Ramos, C. C. O. et all., é apresentado o uso da técnica *Optimum-Path Forest Classifier (OPF)*, que tem o objetivo de caracterizar e detectar possíveis consumidores irregulares por meio de uma metodologia rápida que combina o classificador OPF com a técnica de busca *Harmony Search (HS)*, a fim de selecionar os recursos mais representativos. Foi proposto um novo algoritmo híbrido e rápido para realizar a tarefa de seleção de características usando Harmony Search e Optimum-Path Forest (HS-OPF). A ideia principal é usar a precisão OPF sobre um conjunto e avaliar o valor da aptidão para orientar o algoritmo Harmony Search a encontrar a melhor combinação de recursos. A proposta foi validada em dois conjuntos de dados rotulados, obtidos de uma companhia de energia elétrica do Brasil, sendo um composto por consumidores comerciais e outro por industriais.

Ainda em 2011, Nagi, J. et all. publicaram um artigo que apresenta um modelo de detecção baseado em SVM com a introdução de um sistema de interferência difusa (FIS), sob a forma de regras Fuzzy If-Then. Com a implementação desta metodologia, a taxa de acerto aumentou de 60% (no trabalho anterior) para 72%.

Em 2014, Benítez et all., utilizaram classificação em séries temporais com algoritmos de clusterização dinâmica em um histórico de consumo de energia de perfis de carga. Um algoritmo de agrupamento dinâmico é aplicado em um banco de dados de perfis de carga diários de centenas de clientes durante dois anos. Segundo os autores, as técnicas usadas e a análise realizada provam que a ferramenta é adequada para a classificação de clientes de

acordo com seus padrões de consumo de energia, bem como para avaliar as tendências globais.

Em 2015, Spirić, J. V. et all., é utilizado séries temporais. O objetivo desse artigo é promover a estratégia estatística de controle de processo para a detecção de clientes suspeitos. A verificação deste método foi testada em séries temporais através de um conjunto de clientes que foram pegos furtando energia durante um certo tempo.

## 1.5 ESTRUTURA DA DISSERTAÇÃO

Esta dissertação está organizada em 6 capítulos, distribuídos da seguinte forma:

O capítulo 1 é a introdução, onde nela é feita uma breve apresentação do cenário nacional das perdas comerciais de energia elétrica, seus impactos para as concessionárias de energia, para a economia e consumidores.

No capítulo 2, as perdas comerciais são explanadas, com foco no estudo de caso, a CELPA. O capítulo 3 é dedicado a apresentar a mineração de dados, com foco na técnica utilizada nesta dissertação: a Árvore de Decisão.

O capítulo 4 foi dedicado à metodologia utilizada e seus respectivos passos: recebimento e pré-processamento dos dados. O capítulo 5 apresenta os resultados e suas análises, a partir de uma base de dados real, fornecida pela CELPA.

O 6º e último capítulo trata da conclusão, as considerações finais e sugestões sobre possíveis trabalhos futuros, a fim de se aprimorar a técnica proposta.

## 1.6 PUBLICAÇÃO REALIZADA

Artigo: Y. C. S. Correa, R. L. S. Carvalho, F. P. Monteiro, A. S. Tobias, J. P. A. Vieira, M. V. A. Nunes, U. H. Bezerra, M. E. L. Tostes, R. C. de Oliveira. Irregularities Detection in Electricity Consumption Using Decision Tree – ISGT – LA 2015, Montevideo, Uruguay. Neste trabalho foi apresentada a metodologia utilizada na detecção de irregularidades no consumo de energia elétrica em redes de distribuição, usando árvore de decisão.

# 2 PERDAS COMERCIAIS

## 2.1 INTRODUÇÃO

As perdas de energia elétrica são a diferença entre a energia recebida dos agentes fornecedores, subtraindo a energia medida pelas distribuidoras nas UC's. Essas perdas são classificadas entre técnicas e não técnicas.

As perdas técnicas correspondem à perda inerente ao sistema de energia elétrica, ou seja, é a perda sofrida pelo transporte de energia desde a condução, transformação e distribuição, já que a passagem de energia ocasiona efeito joule e as perdas não técnicas, ou comerciais, provém em geral, de ligações clandestinas, fraudes e irregularidades encontradas nas redes de distribuição de energia elétrica.

Perdas técnicas ocorrem entre a geração de energia elétrica até os sistemas de distribuição. São intrínsecas ao sistema elétrico e inerentes ao transporte da energia elétrica na rede, ligadas à transformação da energia elétrica em térmica por efeito joule, por efeito corona, por correntes de Foucault, e também ao estado de conservação de medidores de energia, as perdas nos núcleos dos transformadores, perdas ligadas às correntes de fuga no ar, de isoladores e de outros equipamentos que compõe a rede de energia elétrica. Podem ser calculadas a partir da revisão tarifária periódica, de acordo com as regras definidas no módulo 7 do PRODIST. [1]

A redução dessas perdas depende, principalmente, da tecnologia utilizada, da qualidade dos serviços de manutenção, da ampliação do sistema elétrico em acordo com a evolução do mercado consumidor e modo de operação dos sistemas, em que alguns pontos chave do controle são: níveis de carregamento dos condutores, demandas de energia reativa, perfis de tensão, etc.

---

1. São procedimentos de distribuição, elaborados pela ANEEL, que normatizam e padronizam as atividades técnicas relacionadas ao funcionamento e desempenho dos sistemas de distribuição de energia elétrica. Contêm 9 módulos, onde, o 7º, refere-se ao cálculo de perdas na distribuição.

Já as perdas não técnicas ou comerciais estão associadas à distribuição de energia elétrica, causada na maioria das vezes, pelo uso ilícito e ou/ inadequado da energia, através da ação de terceiros. É muito comum no Brasil, ocorrerem perdas de energia por ligações clandestinas, popularmente chamadas de “gato”, e fraudes. Essas perdas somam prejuízos superiores à 7 bilhões de reais por ano ao país, sendo correspondente à 13% da energia consumida (SIMÃO, 2012).

Segundo a ABRADEE (Associação Brasileira de Distribuidoras de Energia Elétrica), no Brasil, dependendo da área de concessão, as perdas não técnicas respondem por boa parte do custo da energia elétrica. Isso significa que os consumidores regulares pagam parte do consumo irregular de consumidores que utilizam de práticas ilegais em sua conexão com a distribuidora.

## 2.2 CLASSIFICAÇÃO DAS PERDAS COMERCIAIS

Pode ser classificada como:

- Furto (desvios ou ligações clandestinas);
- Fraude no medidor de energia;
- Erro de leitura, erro no faturamento;
- Irregularidade técnica no equipamento.

Atualmente, as inspeções nas unidades consumidoras são realizadas por técnicos da empresa quando são detectadas significativas variações no consumo de energia, e/ou quando o leiturista da concessionária ao realizar a leitura do medidor, observa algo anormal, como por exemplo uma derivação antes da medição. Contudo, nesse critério de seleção ocorrem falhas na identificação de irregularidades, que ocasionam prejuízos à imagem da concessionária perante a população.

Quando encontradas irregularidades, as UC’s são notificadas, regularizadas e rotuladas em uma das três categorias: normal (nada irregular foi detectado), fraudador (comprovação de fraude no medidor ou furto), ou irregularidade técnica (problema técnico no medidor de energia). Esses dois últimos, quando detectados, rotulam a UC como “Cliente Irregular”.

## 2.3 PERDAS POR INTERFERÊNCIA DA AÇÃO DO CONSUMIDOR

As perdas não técnicas por ação do consumidor podem ser divididas em furtos e fraudes. As duas ações são tidas como roubo de energia elétrica e segundo o código penal brasileiro, o furto de energia está tipificado no Art. 155:

> "É a subtração, para si ou para outrem, coisa alheia móvel:
>
> § *3º Equipara-se à coisa móvel a energia elétrica ou qualquer outra que tenha valor econômico."*

Furtos e fraudes caracterizam-se por conexões diretas ao sistema de distribuição, onde os cabos de ligação das UC's são conectados aos cabos de fornecimento da concessionária diretamente sem passar pelo equipamento de medição no caso de furtos e com adulteração do medidor quando é caso de fraude.

Em geral, a energia furtada é utilizada para alimentar equipamentos elétricos de alto consumo, como por exemplo geladeiras, condicionadores de ar, chuveiros elétricos, e são ligados à parte, estando o resto dos equipamentos da residência conectados ao sistema de medição. Esse tipo de ato pode ser observado, geralmente, em áreas de invasão, periferias e bairros com população de baixa renda, estando normalmente associado ao fato dessas pessoas quererem usufruir de serviços de conforto sem poderem arcar com o custo dos mesmos. Mas também é comumente encontrado em bairros e áreas nobres, pois o alto custo de vida pode estar associado a um gasto elevado de energia elétrica e a fatura da mesma pode representar uma parcela considerável das contas do fim do mês.

As perdas comerciais por furto de energia são associadas em razão de:

- Ligação de condutores diretamente à rede de distribuição, não havendo qualquer vínculo do cliente com a concessionária (ligação clandestina);
- Desvio de corrente não medida, caracterizado pelo rompimento ou desconexão do condutor neutro e utilização de neutro artificial por outra fonte de aterramento ou outra instalação (ramal);

- Desvio de corrente não medida, geralmente do ramal de entrada, sem o emprego da chave inversora;

- Desvio de corrente não medida, geralmente no ramal de entrada, com o emprego de chave inversora.

Para Reis (2005), as principais irregularidades na rede de distribuição são as com desvio em 1, 2 ou 3 fases e no ramal de ligação, podendo ser vistos com: desvio em 1, 2 ou 3 fases; com saída aérea, *bypass* em 1, 2 ou 3 fases; desvio através de fenda no eletroduto de entrada ou no cabo concêntrico (tubo sangrado).

O ato consciente de uma pessoa para eliminar ou reduzir a energia faturada, quando o medidor é adulterado ou quando é feito um desvio no ramal de entrada, antes do medidor é a fraude de energia. O consumidor faz um aumento de carga à revelia da concessionária, num circuito clandestino e, em muitos casos, de modo até sofisticado.

Segundo Calili (2005), por muitas vezes, o medidor é o caminho para a fraude, o que faz deste aparelho não só o causador da perda técnica, em razão das avarias nele, mas também pela perda comercial.

De acordo com os processos analisados pela ANEEL, os tipos de irregularidades mais encontradas com perfil de perdas por furtos e fraudes são:

- Desvio no ramal de entrada (antes do medidor);
- Ponteiros do medidor deslocados;
- Ligações do medidor invertidas;
- Terminal de prova aberto;
- Ligação direta à rede secundária;
- Bobina de potencial interrompida;
- Engrenagem do medidor substituída;
- Chave de aferição aberta;
- Curto circuito na entrada ou saída do medidor;
- Injeção de corrente no medidor com queima de uma ou mais bobinas do mesmo;
- Alimentação do motor de temporização de demanda interrompida;

## 2.4 FATORES DE INFLUÊNCIA

A questão das perdas comerciais é de alta complexidade, uma vez que estão envolvidas questões socioeconômicas da região consumidora, o desenvolvimento social, econômico e educacional da população. Vieiralves (2005) propõe que dentre outros fatores, as perdas comerciais sofrem expressiva influência de fatores exógenos: questões sociais, nível de emprego e renda, composição da balança comercial regional e nacional.

Para Smith (2004), as perdas comerciais estão associadas a problemas de governança, como aspectos de processos políticos, liberdade civil, burocracia, independência dos serviços públicos às pressões políticas, instabilidade política e violência, além da efetividade do judiciário na aplicação da lei, mostrando a correlação entre estes aspectos e o nível de perdas de energia elétrica.

Em Araújo (2007), a inadimplência é definida como a relação, em termos percentuais, entre o montante das contas não pagas até o último dia do mês de referência, incluindo tributos, e o total de contas faturadas no mesmo mês. Esse fator é amplamente discutido por muitos autores como Calili (2005), Araújo (2007), Ortega (2008) e Penin (2008), e para eles existe uma grande relação.

Em geral, furto e inadimplência estão relacionados às restrições orçamentárias dos consumidores. No contexto atual de tarifas altas e o crescente número de equipamentos eletroeletrônicos alimentados por energia elétrica, o aumento de consumo é considerável, logo o consumidor depara-se com uma situação em que pagar a conta de energia é um fardo crescente. O poder concedente e o agente regulador procuram minimizar esse efeito, através de uma tarifa social, que consiste num subsídio a um grupo de consumidores que não é limitado aos de baixa renda. Existem fatores sociais e culturais, além das desvantagens financeiras, que pressionam o consumidor ao furto de energia e à inadimplência. Um nível considerável de consumidores, que têm seu fornecimento de energia suspenso, recorre ao furto de energia e, de maneira análoga, parte do mercado, que é recuperado através de ações de combate às perdas, torna-se inadimplente.

Além da inadimplência, existe a questão cultural que envolve furtos e fraudes. Afim de justificar esses atos ilícitos nos deparamos com opiniões de que a tarifa de energia é cara, que não há mal algum em se fazer um “gato”. É tão forte a questão cultural que existe uma “indústria de irregularidades”, estando presente tanto consumidores como eletricistas que fomentam esses crimes. Por maior que seja a fiscalização para combater o “gato”, observa-se

que a cultura popular não considera essa prática ilícita, logo as concessionárias detectam as irregularidades, regularizam a medição e mais tarde são surpreendidas ao constatar que os mesmos consumidores tornaram a cometê-lo.

Vários estudos comprovam o aumento de ligações irregulares, quando da deterioração econômica da região ou do país em que se encontram (PENIN, 2008). As altas taxas de perdas comerciais são altamente influenciadas por esses fatores, mesmo existindo leis que asseguram a prática como crime inafiançável e pena prevista de um a quatro anos de prisão. Quando detectado furto ou fraude na unidade consumidora, é permitido às distribuidoras a cobrança do retroativo em até cinco anos; porém, numa ação conjunta com as agências reguladoras estaduais, esse período de cobrança acaba reduzindo, diminuindo o efeito da punição.

Segundo as concessionárias com maior índice de perdas não técnicas, os bairros conhecidos como “áreas vermelhas” - nas estatísticas policiais são a razão dos altos índices de delitos como roubos - são os que ocorrem uma maior incidência de furto e fraude e quando uma equipe é enviada para vistoria e faz o desligamento dos furtos, logo após sua saída, os moradores refazem as ligações clandestinas.

Há que se considerar que os problemas socioeconômicos, enfrentados por parte da população, na maioria das vezes, levam-na a práticas que buscam adequar seu perfil de despesas à sua renda reduzida. Assim, por questão cultural, dentro dessas comunidades, tal ato não é considerado delito e, sim, uma forma de administração de suas necessidades.

Dentro deste contexto, torna-se evidente a necessidade de programas sociais e educativos nessas áreas sujeitas a elevados índices de perdas comerciais, pois poderão auxiliar na redução daqueles, visto que serão uma potente maneira de combate e prevenção a fraudes e furtos.

## 2.5 O COMBATE ÀS PERDAS NO BRASIL

Em uma experiência da LIGHT em 2011 no município de Nilópolis, na baixada fluminense, a empresa usou de estratégia a aproximação com os clientes com equipes dedicadas que trabalharam intensamente para quebrar a relação: “ou o cliente furtava a energia ou não pagava a conta”. As perdas reduziram de 35% para 7% e o nível de adimplência subiu de 94% para 98%. O pensamento é que somente a utilização de medidores eletrônicos e blindagem da rede como soluções no combate às perdas comerciais podem

impor uma distância entre a empresa e os clientes. Para a LIGHT, a solução é combinar a aproximação com a instalação dessas novas tecnologias de medição.

Já com as cinco distribuidoras administradas pelo do Grupo Energisa, o combate às perdas mudou de foco, de regularização passou para inspeção. Segundo o grupo, 37% dos 2,7 milhões de medidores que fazem parte do parque das distribuidoras são eletrônicos e ainda contam com um centro unificado que abrange as cinco distribuidoras, o Centro de Inteligência no Combate às Perdas (CICOP), composto por 12 especialistas em análise e direcionamento das perdas, engenheiros, analistas de sistemas e estatísticos que identificam potenciais de irregularidades.

O grupo aponta que é necessário um planejamento adequado, treinamento das equipes que vão a campo e um forte trabalho de mídia junto à população para auxiliar na redução dos índices de perdas não técnicas. Em 2011 a Energisa Paraíba continuou reduzindo gradativamente as perdas, alcançando um recorde histórico de 11,05%, com ganho de 1,43 p.p.[2] em relação ao ano anterior e em cinco anos a distribuidora conseguiu uma redução de 20,43% em 2006 para 13,68% em 2011.

No estado do Ceará, a Coelce (Companhia de Energia Elétrica do Ceará) em parceria com o Instituto Federal do Ceará (IFCE), no âmbito de um programa de Pesquisa & Desenvolvimento (P&D) da Aneel, criaram um equipamento gerador de ruídos (interferências) na corrente elétrica, instalado no transformador de distribuição, impedindo que a energia furtada possa ser utilizada sem antes passar pelo medidor do consumidor que conta com um removedor de ruídos (filtro) para que a energia volte a ser utilizável. Essa energia “suja” quando é usada, gera por exemplo distorções na luminosidade de luz de uma lâmpada, ligamento e desligamento de televisores, a geladeira vibra e pode pular, ou seja, consequências incômodas ao cliente fraudador.

A redução do índice de perdas foi evidente no Ceará. Foram escolhidos dois locais para a instalação do “provocador de ruídos”, um circuito de baixa tensão na cidade de Fortaleza, com 74 consumidores, que havia muitas ligações clandestinas e furtos na rede elétrica, com índice de 49% de perdas, sendo reduzido a 2,7%.

---

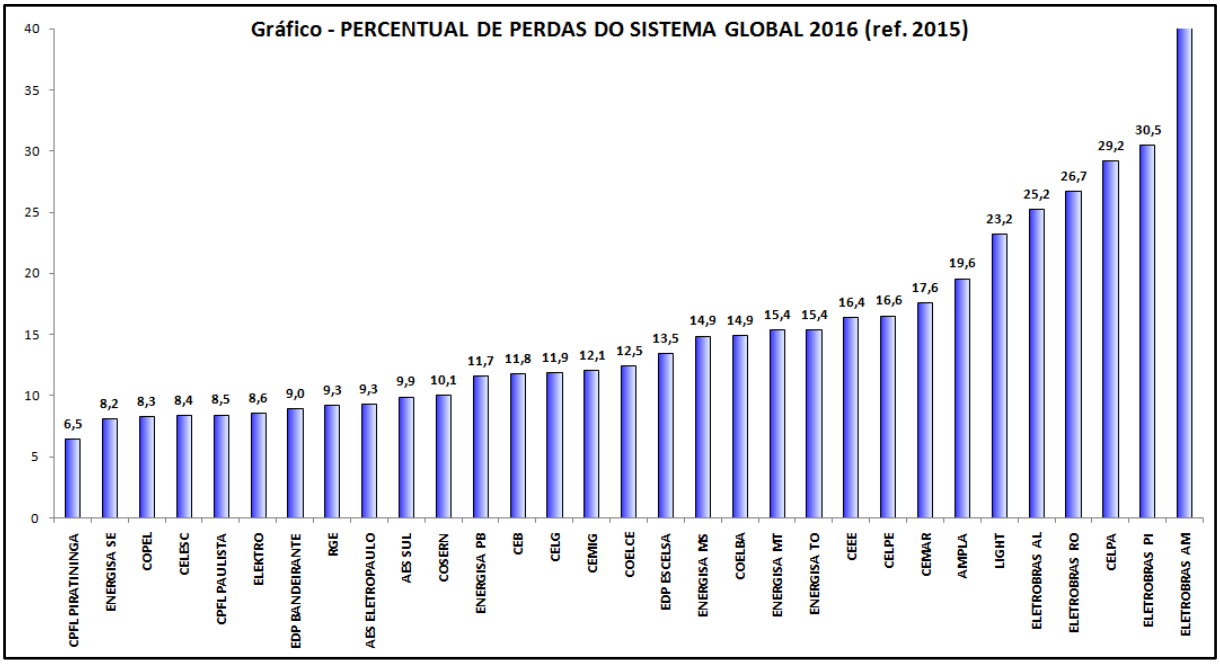
2. p.p: ponto percentual, é uma unidade que indica o valor absoluto da diferença entre percentagens.

E uma comunidade agrícola, em que o sistema de baixa tensão atravessa as propriedades dos consumidores, tornado fácil o uso de “garras” ou “ganchos” para uso no bombeamento dos sistemas de irrigação de cultura agrícola. Nessa área a redução foi de 39%.

## 2.6 PERDAS COMERCIAIS NO ESTADO DO PARÁ – CASO CELPA

O último resultado divulgado pela ABRADEE, mostra o percentual de perdas do sistema global nas distribuidoras de energia elétrica do país, estando duas delas com percentual superior à 30%, como exibe a figura 1.

Figura 1 – Gráfico de perdas nas 31 distribuidoras em 2015.




Fonte: Associação Brasileira de Distribuidores de Energia Elétrica - ABRADE, 2016.

A CELPA, concessionária do estudo de caso desta dissertação, atende os 144 municípios do estado do Pará, correspondendo a mais de 2 milhões de clientes. O Pará é o terceiro no ranking de perdas comerciais no Brasil e de acordo com a empresa, 1/3 da energia elétrica distribuída é subtraída. São cerca de 300 mil ligações clandestinas e aproximadamente 180 mil delas ocorrem na Região Metropolitana de Belém.

Em 2015, os comentários de desempenho da CELPA mostram que para o primeiro trimestre (1T15), a demanda de energia cresceu 4,6%, chegando a atingir 1.967 GWh[3], enquanto que as perdas totais de energia dos últimos 12 meses encerrados no primeiro

trimestre de 2015, representaram 30,8% da energia requerida — o que corresponde a 0,4 p.p. de queda, em relação aos 31,2%, verificados no findar do ano anterior, 2014.

As perdas totais, no segundo trimestre de 2015, representaram para a concessionária, 31,8% da energia requerida, representando um aumento de 1,0 p.p, quando comparadas aos 30,8%, verificados no primeiro trimestre de 2015; ao passo que, as perdas não técnicas, sobre o mercado de baixa tensão, atingiram 45,2%.

Já no terceiro trimestre de 2015, as perdas totais de energia tiveram uma queda de 0,5 p.p., ou seja, 31,3% da energia requerida, enquanto que as perdas não técnicas, sobre o mercado de baixa tensão, atingiram o percentual de 44,2%.

As perdas totais encerraram o último trimestre de 2015 em 29,2% da energia requerida, representando queda de 2,1 p.p. em relação aos 31,3% verificados no trimestre anterior, enquanto que as perdas não técnicas, sobre o mercado de baixa tensão, atingiram o percentual de 38,6%.

A tabela 1 mostra o balanço energético da empresa, em comparação com os trimestres anteriores e as ilustrações seguintes, os gráficos das figuras 2 e 3, correspondem as perdas totais sobre a energia requerida e às perdas não técnicas sobre o mercado de baixa tensão nos últimos meses do ano de 2015, respectivamente.

Tabela 1 - Balanço energético da CELPA

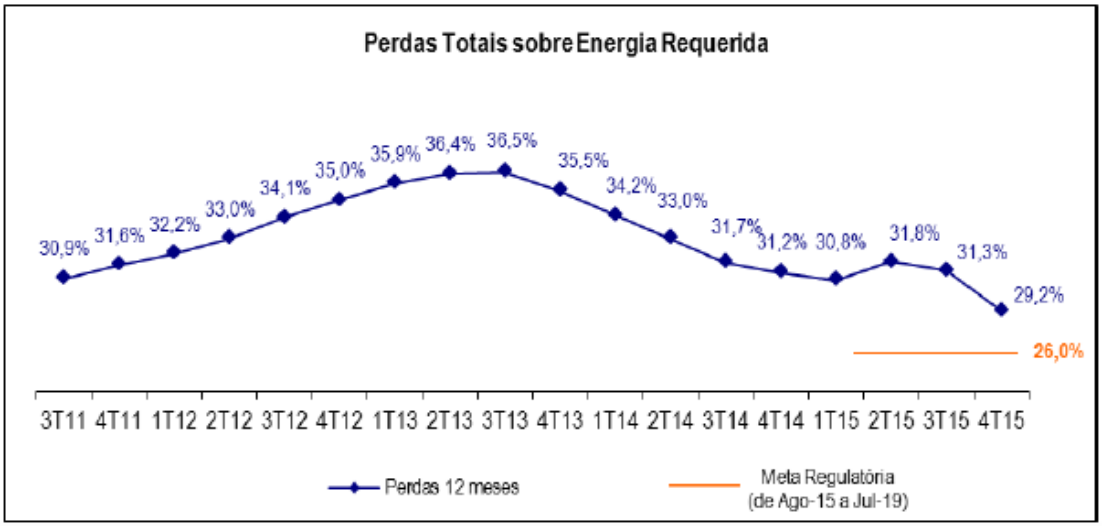
| BALANÇO ENERGÉTICO (GWh) | 4T14 | 3T15 | 4T15 | Var. | 2014 | 2015 | Var. |
|---|---|---|---|---|---|---|---|
| Energia Vendida (Cativo + Cons. Próprio) | 2.062 | 2.096 | 2.354 | *14,1%* | 7.755 | 8.138 | *4,9%* |
| Mercado Livre | 98 | 78 | 68 | *-29,9%* | 376 | 317 | *-15,8%* |
| Perdas Totais | 979 | 898 | 734 | *-25,0%* | 3.693 | 3.488 | *-5,6%* |
| **Energia Requerida** | **3.139** | **3.073** | **3.157** | ***0,6%*** | **11.824** | **11.943** | ***1,0%*** |
| Geração Própria | 127 | 109 | 116 | *-9,2%* | 469 | 441 | *-5,8%* |
| Compra de Energia (Contratos) | 3.011 | 2.963 | 3.041 | *1,0%* | 11.355 | 11.502 | *1,3%* |

(*) Inclui venda às classes, consumo próprio e merc. livre.

Fonte: Comentários de Desempenho 4T15, CELPA

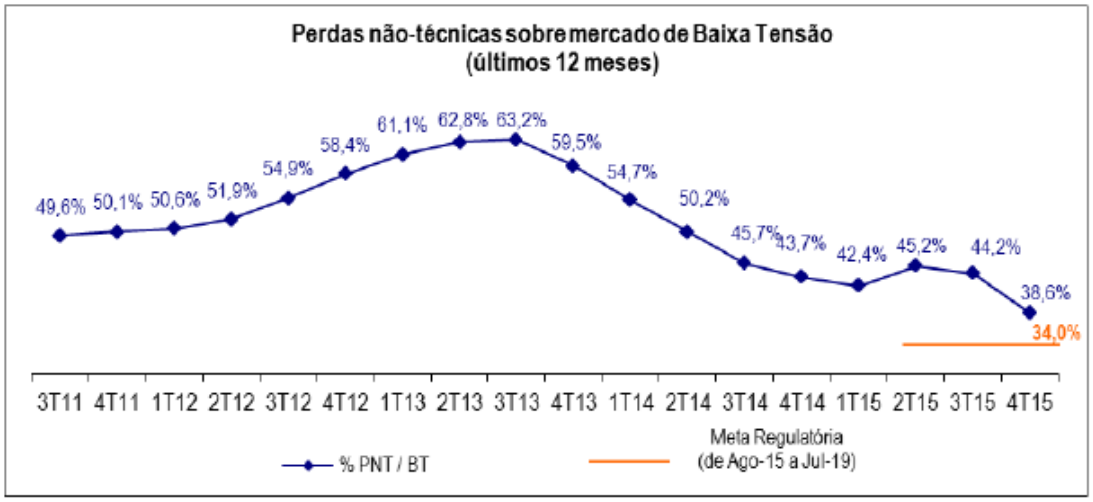
3 GWh: Giga-Watt-hora, em que watt-hora é a unidade de medida de potência elétrica pelo tempo no qual ela é consumida. Giga significa 1 bilhão.

Figura 2 - Perdas totais sobre a energia requerida




Fonte: Comentários de Desempenho 4T15, CELPA

Figura 3 - Perdas não técnicas sobre o mercado de baixa tensão nos últimos 12 meses




Fonte: Comentários de Desempenho 4T15, CELPA

As quedas nos percentuais de perdas de energia são frutos da intensificação e maior sucesso nas ações de combate, com consequente regularização e faturamento do consumo não registrado - quando o consumidor é regularizado após uma inspeção e, verificada a fraude, é faturado pelo seu consumo em meses anteriores. Tal redução ocorre após reformularmos o processos e ações de combate às perdas após percebermos uma necessidade de intensificação do programa, cujo resultado estamos observando neste trimestre, apesar da maior resistência às ações de combate em vista do atual momento econômico e da complexidade da área de concessão.

No primeiro trimestre de 2016, as perdas totais não tiveram queda, mas mantiveram o índice de 29,9% da energia requerida. Segundo a distribuidora, após a implantação de um novo sistema SAP/CCS no faturamento, o percentual de perdas poderia chegar a 29,2%.

Esse novo sistema comercial foi implantado no primeiro trimestre de 2016 e para a empresa ele representa (i) economia de aproximadamente 7 milhões de reais em custos anuais nas duas companhias combinadas (CELPA e CEMAR); (ii) unificação dos sistemas e processos em ambas as companhias; (iii) revisão de todos os procedimentos comerciais e suas interfaces, e; (iv) adoção de tecnologia de ponta, possibilitando maior agilidade e eficiência nas operações cotidianas.

Algumas atividades de faturamento e combate às perdas foram afetadas no mês de março. Estima-se que houve impacto entre 50 e 80 GWh de faturamento, o que ensejaria um crescimento de volume de até 3,4%, em linha com o crescimento de energia requerida apresentada no período. As perdas estariam num patamar de 29,2% e, consequentemente, o EBITDA[4] estaria R$ 21 milhões maior, atingindo R$ 127 milhões no trimestre. Mesmo após a estimativa de impactos da implantação do novo sistema, no cálculo do EBITDA, houve impacto de R$ 31 milhões negativos de renda não faturada neste trimestre (R$ 1 milhão negativo no 1T15), que sazonalmente deve ser revertido ao longo deste ano.

Com a sua implantação tendo ocorrido em março, os efeitos do período de estabilização no trimestre para a CELPA são mais significativos. A expectativa é que essa transição ainda possa afetar parte dos resultados do 2T16, com recuperação prevista do referido impacto até o final desse ano.

O segundo trimestre obteve redução nos níveis de perdas, representando 28,6% da energia requerida e queda de 1,3 p.p., ao passo que as perdas não técnicas sobre o mercado de baixa tensão atingiram 38,6%, uma diminuição de 2,6 p.p. em relação ao trimestre anterior.

---

4 EBITDA: *Earnings Before Interest, Taxes, Depreciation and Amortization.* Que significa lucros antes de juros, impostos, depreciação e amortização, em português. É um indicador financeiro que representa o quanto uma empresa gera de recursos através de suas atividades operacionais, sem contar impostos e outros efeitos financeiros. É capaz de medir a produtividade e a eficiência de uma empresa.

Após o impacto no 1T16 da migração para o novo sistema comercial da companhia, o percentual de perdas retoma uma trajetória de queda atingindo o seu menor nível desde que a Celpa foi adquirida no final de 2012.

No terceiro trimestre os níveis de perdas representaram 27,7% da energia requerida, ficando a 0,9 pontos da meta regulatória, ao passo que as perdas não técnicas sobre o mercado de baixa tensão atingiram 36,8%, uma diminuição de 2,3 p.p. em relação ao fechamento do trimestre anterior.

As perdas totais encerraram o último trimestre de 2016 em 28,3% da energia injetada, um aumento de 0,6 p.p. em relação aos 3T16. As perdas não técnicas, sobre o mercado de baixa tensão, atingiram o percentual de 37,8%, um crescimento de 1,6 p.p. comparado ao trimestre anterior.

Segundo a companhia, a eficácia das ações de combate às perdas reflete-se na queda dos índices nos últimos 12 meses encerrados em dezembro de 2016, relativamente ao mesmo período do ano anterior. Os percentuais de perdas totais e não técnicas sobre mercado de baixa tensão apresentaram redução significativa, mesmo apesar do cenário recessivo e dos desafios enfrentados na implantação do sistema comercial durante o primeiro semestre de 2016.

A tabela 2 mostra o balanço energético da empresa, em comparação com os trimestres anteriores e as ilustrações seguintes, os gráficos das figuras 4 e 5, correspondem as perdas totais sobre a energia requerida e às perdas não técnicas sobre o mercado de baixa tensão nos dois primeiros trimestres de 2016, respectivamente.

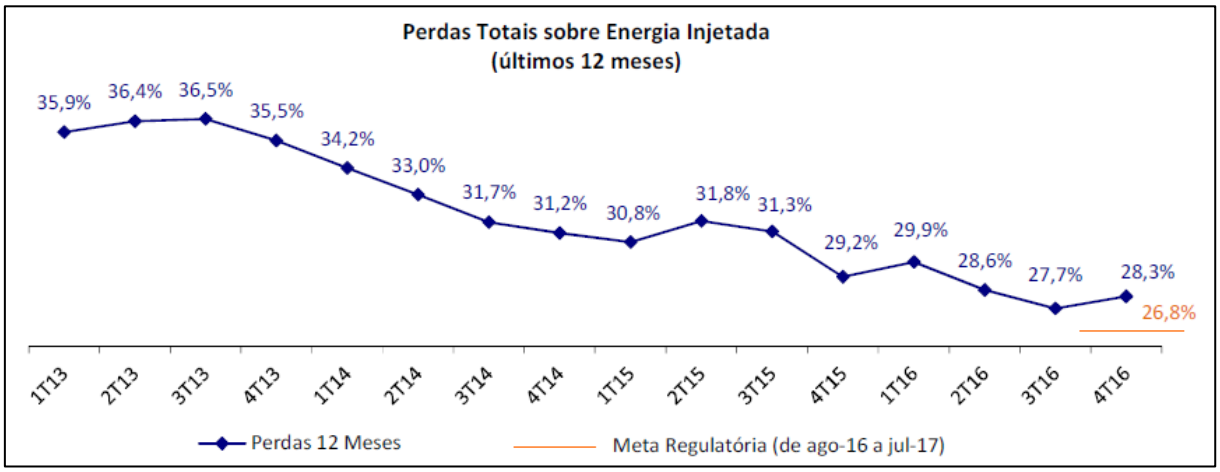
Tabela 2 - Balanço energético da CELPA

| Bal. Energético (MWh) - CEMAR | 4T15 | 4T16 | *Var.* | 2015 | 2016 | *Var.* |
|---|---|---|---|---|---|---|
| Sistema Interligado | 1.948.278 | 1.981.507 | *1,7%* | 7.235.690 | 7.530.671 | *4,1%* |
| **Energia Injetada** | **1.948.278** | **1.981.507** | ***1,7%*** | **7.235.690** | **7.530.671** | ***4,1%*** |
| Energia Distribuída* | 1.600.308 | 1.618.595 | *1,1%* | 5.960.762 | 6.175.379 | *3,6%* |
| Perdas Totais | 347.969 | 362.913 | *4,3%* | 1.274.927 | 1.355.291 | *6,3%* |

(*) Inclui mercados cativo e livre, consumo próprio e fornecimento a Estados vizinhos.

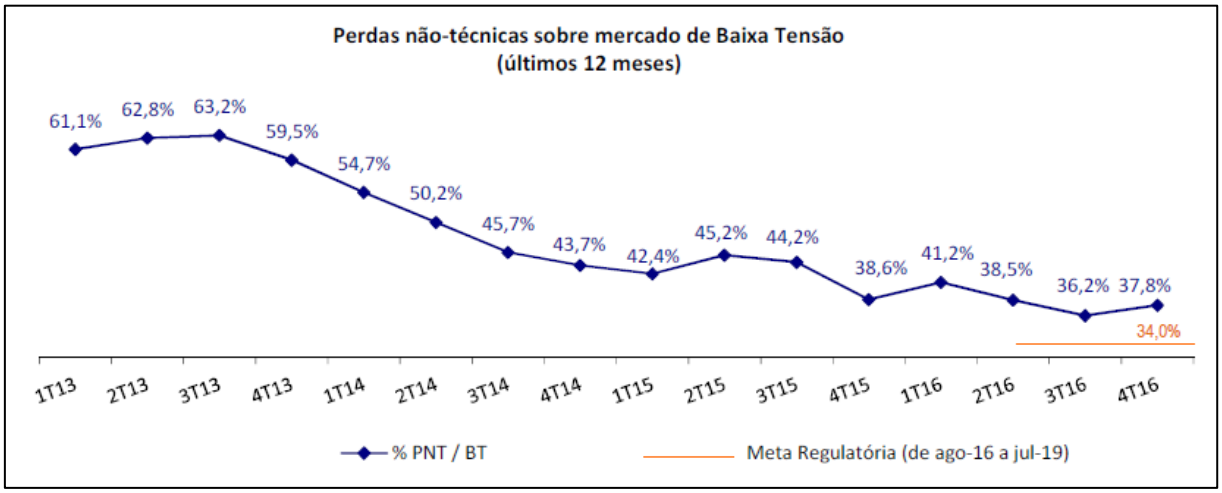
Fonte: Comentários de Desempenho 4T16, CELPA

Figura 4- Perdas totais sobre a energia requerida




Fonte: Comentários de Desempenho 4T16, CELPA

Figura 5 - Perdas não técnicas sobre o mercado de baixa tensão nos últimos 12 meses




Fonte: Comentários de Desempenho 4T16, CELPA

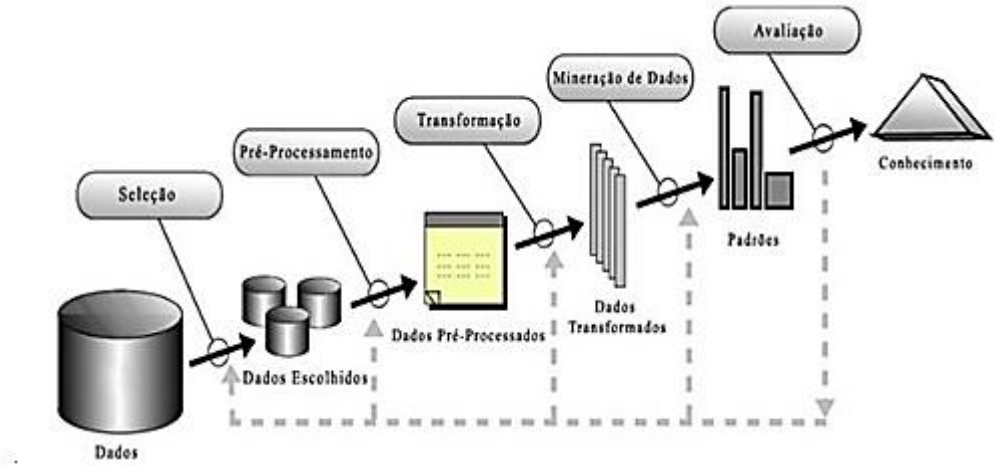
# 3 MINERAÇÃO DE DADOS E ÁRVORE DE DECISÃO

## 3.1 INTRODUÇÃO

Capturar, organizar e armazenar grandes quantidades de dados, obtidas de operações diárias ou pesquisas científicas, são tarefas árduas que as empresas adotam, a fim de detectar padrões, tomada de decisões, estatísticas, entre outras.

Um processo amplo, de maior pesquisa, denominado Busca de Conhecimento em Banco de Dados (*Knowledge Discovery in Database - KDD*), é uma metodologia própria para preparação e exploração dos dados, interpretação de seus resultados e assimilação dos conhecimentos minerados. Fayyad o define como “um processo não trivial de novos padrões válidos, úteis e compreensivos”. No entanto, a mineração de dados tornou-se mais conhecida do que o próprio processo de KDD, em função de ser a etapa onde são aplicadas as técnicas de busca de conhecimento. Na figura 6 tem-se a representação dos processos do KDD.

Figura 6 - Esquemático de representação do processo de KDD




Fonte: FAYYAD, 1996

São diversos os processos que definem e padronizam as fases e atividades da mineração de dados, que apesar de todos os casos serem particulares, no geral, possuem a mesma estrutura.

## 3.2 SELEÇÃO E PRÉ PROCESSAMENTO

A fase de seleção de dados é a primeira no processo de descobrimento de informação e possui impacto significativo sobre a qualidade do resultado final, uma vez que nesta fase é escolhido o conjunto de dados contendo todas as possíveis variáveis (também chamadas de características ou atributos) e registros (também chamados de casos ou observações) que farão parte da análise. O processo de seleção é bastante complexo, uma vez que os dados podem vir de uma série de fontes diferentes (data warehouses, planilhas, sistemas legados) e podem possuir os mais diversos formatos. É comum a necessidade do uso de um software específico para a carga dos dados, já que nem sempre as ferramentas de carga existentes conseguem dar contas das peculiaridades de cada aplicação.

O processo de preparação dos dados para a mineração é denominado pré-processamento e, serve para melhorar a qualidade dos dados. Essa etapa compreende funções relacionadas à captação, organização, tratamento e à preparação da base de dados para a etapa de *Data Mining*. É uma etapa fundamental no processo de Descoberta de Conhecimento em Base de Dados – *KDD*, pois a qualidade dos dados é quem determina a eficiência dos algoritmos de mineração, sendo relatada em trabalhos científicos como fase vital para buscar resultados satisfatórios e, em alguns casos, chega a tomar até 70% do tempo do projeto de mineração.

O efeito da aplicação da etapa de pré-processamento é observado na performance das aplicações de técnicas computacionais, tanto para a classificação quanto para o aprendizado de máquina. Isso torna a construção do modelo mais próximo da distribuição real dos dados, contribuindo também para a redução da complexidade computacional.

As técnicas de pré-processamento minimizam e podem eliminar problemas existentes em um conjunto de dados e ao depender dos algoritmos utilizados, podem deixar os dados fáceis para manipulação, além de não existir ordem fixa para aplicação das técnicas do pré-processamento. A seguir, algumas das técnicas de pré-processamento que foram utilizadas na metodologia desse trabalho:

- Integração de dados;
- Modificações para adequação dos tipos de atributos;
- Técnicas de amostragem;

- Redução de dimensionalidade;
- Balanceamento de dados;
- Limpeza de dados;
- Transformações de dados.

### 3.2.1 Eliminação manual de atributos

Em uma base de dados, alguns atributos não possuem relação como o problema a ser solucionado. Torna-se um caso de classificar esse(s) atributo(s) como irrelevante. Outra situação que deve ser verificada é a relevância dos atributos, ou seja, o atributo que possui o mesmo valor para todos os objetos.

### 3.2.2 Integração dos dados

Antes da aplicação das técnicas para classificação dos dados é preciso verificar se o conjunto de dados está distribuído entre outros conjuntos. Caso esteja, é necessário que seja realizada a integração dos dados. Outro problema é quando um atributo está identificado de modo diferente pelos conjuntos de dados ou quando os dados podem ter sido atualizados em momentos diferentes. Para esse caso, a solução consiste em utilizar metadados.

Os metadados são dados sobre dados que ao descrever as suas principais características são usados para evitar erros no processo de integração. O processo de integração origina um depósito ou repositório de dados (data warehouse) quer funciona como base de dados centralizada.

O número elevado de atributos de uma grande base de dados pode influenciar o desempenho do algoritmo pelo problema chamando de maldição de dimensionalidade. Já em caso de número elevado de objetos ocorre o problema de saturação de memória e aumento do tempo computacional.

### 3.2.3 Amostragem de dados

Alguns algoritmos de aprendizagem de máquina podem apresentar problemas na classificação de grandes números de objetos, como por exemplo algoritmos que são baseados em instâncias, ou que apresentam saturação de memória de grande número de dados. Quanto

mais dados são utilizados maior será a acurácia do modelo e menor a eficiência computacional. Para obtenção de ponto de equilíbrio entre acurácia e a eficiência computacional podem ser utilizados subconjuntos de dados, com o objetivo de diminuir o custo computacional.

### 3.2.4 Desbalanceamento de dados

Em vários conjuntos de dados reais podem ocorrer variação da quantidade de objeto para diferentes números de classes, cujos dados de subconjunto das classes aparecem com maior frequência que em outras classes.

O problema do desbalanceamento de dados afeta diversos algoritmos de classificação, como por exemplo, o algoritmo pode favorecer a classificação de novos dados para a classe majoritária. Para solução desses problemas, técnicas para balancear artificialmente o conjunto de dados podem ser utilizadas. Principais soluções:

- Redefinir o tamanho do conjunto de dados;
- Utilizar diferentes peso de classificação para as classes;
- Utilizar modelos para classificação de cada classe.

### 3.2.5 Redução de dimensionalidade

Uma forma de minimizar o efeito da dimensionalidade dos dados é por meio da eliminação de parte dos atributos ou pela combinação deles. A redução dos atributos pode melhorar o desempenho e reduzir o custo computacional do modelo, além de tornar mais fácil a compreensão dos resultados.

Quando cada atributo for visto como uma coordenada em um espaço d-dimensional em que “d” é número de atributos, o hipervolume que representa esse espaço cresce exponencialmente com a adição de novos atributos.

As técnicas para redução de dimensionalidade podem ser divididas em duas abordagens: a agregação e a seleção de atributos.

***Agregação*** **–** As técnicas de agregação utilizam novos atributos a partir da combinação de grupos de atributos originais da base de dados. Entre as principais técnicas de agregação está a Análise de componentes principais, a qual correlaciona estaticamente os exemplos e reduz o

efeito de dimensionalidade através de redundâncias.

Em aplicação nas áreas que há a importância de preservar os valores dos atributos, como áreas de biologia, finanças, medicina e monitoramento, são mais frequente utilização de técnicas de seleção de atributos para redução de dimensionalidade.

***Seleção de atributos*** **–** Além da maldição da dimensionalidade, parte dos atributos podem ser irrelevantes, redundante ou conter grande número de ruídos. Por meio da seleção de atributos pode-se:

- Identificar atributos importantes;
- Melhorar o desempenho das técnicas de aprendizado de máquina;
- Reduzir o custo computacional;
- Simplificar o modelo gerado;
- Facilitar a interpretação dos dados.

Uma forma de selecionar os atributos é através da eliminação manual, a qual pode se tornar complexa em casos com grandes números de exemplos ou atributos, e também pelas relações complexas entre atributos. Existem várias técnicas automáticas na literatura para seleção de atributos quer estão divididas em:

- Embutida;
- Baseada em filtro;
- Baseada em wrapper;

Nas técnicas embutidas, a seleção do subconjunto é feita pelo próprio algoritmo de aprendizado, assim como é feito pelos algoritmos de árvore de decisão, os quais realizam seleção interna de atributos; a técnica baseada em filtro é uma etapa de pré-processamento que utiliza o filtro sobre o conjunto de atributos originais, sem levar em consideração o algoritmo de aprendizado que utilizará esse conjunto de dados; a técnica baseada em wrapper utiliza o próprio algoritmo de aprendizado como caixa-preta para seleção. Essa técnica é normalmente combinada com uma técnica de amostragem. Para cada subconjunto, o algoritmo é consultado e o subconjunto que apresentar a melhor combinação entre relação da taxa de erro e redução do número de atributos será selecionado.

## 3.3 TRANSFORMAÇÃO DOS DADOS

A Transformação do Dados é a fase do KDD que antecede a fase de Data Mining. Após serem selecionados, limpos e pré-processados, os dados necessitam ser armazenados e formatados adequadamente para que os algoritmos possam ser aplicados. Em grandes corporações é comum encontrar computadores rodando diferentes sistemas operacionais e diferentes Sistemas Gerenciadores de Bancos de Dados (SGDB). Estes dados que estão dispersos devem ser agrupados em um repositório único.

A etapa de transformação dos dados merece destaque. Alguns algoritmos trabalham apenas com valores numéricos e outros apenas com valores categóricos. Nestes casos, é necessário transformar os valores numéricos em categóricos ou os categóricos em valores numéricos. Não existe um critério único para transformação dos dados e diversas técnicas podem ser usadas de acordo com os objetivos pretendidos. Algumas das técnicas empregadas nesta etapa são: suavização (remove valores errados dos dados), agrupamento (agrupa valores em faixas sumarizadas), generalização (converte valores muito específicos para valores mais genéricos), normalização (colocar as variáveis em uma mesma escala) e a criação de novos atributos (gerados a partir de outros já existentes).

Na metodologia adotada utiliza-se o método de sumarização, que tem como objetivo encontrar uma descrição simples e compacta dos dados. Para isso, podem ser utilizadas desde medidas estatísticas simples, como mínimo, média e desvio padrão, até técnicas sofisticadas de visualização e de determinação de relações funcionais entre atributos (HAN; PEI; KAMBER, 2011; MIRKIN, 2011).

Uma série temporal pode ser definida como um conjunto de observações feitas sequencialmente no tempo. Essa padronização é muito importante, pois no presente projeto de P&D são utilizadas informações de consumo mensais dos consumidores.

## 3.4 MINERAÇÃO DE DADOS

A mineração de dados é uma metodologia que tem por objetivo a obtenção de conhecimento, extraindo informações potencialmente úteis, e não triviais, de grandes conjuntos de dados. O conhecimento extraído, pode ser expresso, através de regras que descrevem as propriedades dos dados, os padrões mais frequentes, agrupamentos de objetos na base de dados, ou ainda, pela classificação dos dados.

Como mostrado na figura 6, a mineração de dados é apenas uma das etapas do KDD, nela envolve ainda a seleção, transformação, a mineração e por fim, a avaliação do resultado, com as informações úteis extraídas.

Para a etapa de mineração, foi escolhida uma técnica de aprendizado supervisionado, conhecida como Árvore de Decisão (Decision Tree). É uma técnica bastante utilizada na implementação de sistemas especialistas, que vem ganhando destaque entre trabalhos científicos, está em constante aperfeiçoamento, possui forma eficiente de construir classificadores, é robusta e tem bom desempenho, com grande quantidade de informação em pouco tempo, alcançando resultados satisfatórios, sendo também de fácil entendimento e interpretação. Todos esses são fatores que chamam a atenção, qualificando-a como a técnica que mais se adequa ao perfil da problemática em questão, devido à sua característica de classificação de padrões. Vale ressaltar a necessidade de se fazer uma análise detalhada dos dados que serão usados para garantir bons resultados.

### 3.4.1 TAREFA DE CLASSIFICAÇÃO

Pode-se classificar a mineração de dados pela sua capacidade de realizar determinadas tarefas (LAROSE, 2005). Essas tarefas podem ser divididas em duas categorias principais: Tarefas Descritivas e Tarefas Preditivas.

**Tarefas Descritivas (Não Supervisionadas) –** Seu objetivo é derivar padrões como: correlações, tendências, anomalias, grupos e trajetórias que representem relações comuns descobertas na base de dados. Na mineração de dados, as tarefas descritivas requerem técnicas de pós processamento para validar e explicar os resultados. Destacam-se Regras de Associação, Clustering e Sumarização dentre as tarefas descritivas.

**Tarefas Preditivas (Supervisionadas) –** Constroem uma hipótese, a partir da generalização de exemplos com classes previamente definidas, na tentativa de prever um tipo de comportamento para novos casos. O atributo cujo valor se deseja descobrir é conhecido como “variável dependente” ou “alvo”, enquanto que os atributos usados para fazer a predição são chamados de variáveis independentes ou explicativas. Os principais tipos de problemas são Classificação e Regressão.

A tarefa de classificação tem como função examinar um conjunto de registros rotulados e elaborar descrições das características dos registros determinando em que classe

se encontram. Um objeto é examinado e classificado, de acordo com uma classe definida. “A tarefa de classificação pode ser considerada uma tarefa mal definida, indeterminística, que é inevitável pelo fato de envolver predição”.

A mineração de modelos de classificação em base de dados é um processo composto de duas fases: aprendizado e teste. Na fase de aprendizado, um algoritmo classificador é aplicado sobre um conjunto de dados de treinamento. Como resultado, obtêm-se a construção do classificador, propriamente dito. Tipicamente, o conjunto de treinamento corresponde a um subconjunto de observações, selecionadas de maneira aleatória, a partir da base de dados que se deseja analisar. Cada observação do conjunto de treinamento é caracterizada por dois tipos de atributo: o atributo classe, que indica a classe à qual a observação pertence, e os atributos preditivos, cujos valores serão analisados para que seja descoberto o modo como eles se relacionam com o atributo classe.

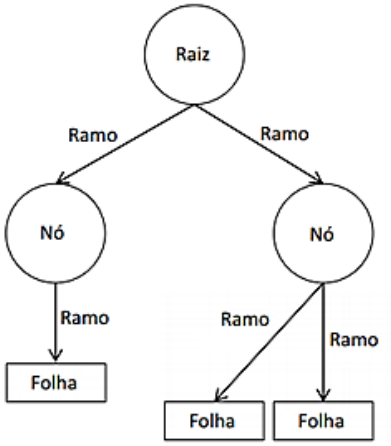
As principais técnicas de classificação atualmente usadas são: Árvore de Decisão, Redes Neurais Artificiais com Perceptron e Kohonen, Classificadores baseados em regras, Bayesianos, Máquina de Vetor de Suporte ou SVM – do inglês, Support Vector Machine – Sistemas Fuzzy, Máquina de Aprendizagem Extrema ou ELM (do inglês, Extreme Learning Machine) e uma nova técnica: Florestas de Caminhos Ótimos (OPF – Optimum-Path Forest). Quem ordena qual método será utilizado é o tipo de base de dados disponível.

### 3.4.2 ÁRVORE DE DECISÃO

A árvore de decisão (DT, do inglês *Decision Tree)* é um algoritmo classificador, possui habilidade de aprender através de exemplos com o objetivo de classificar registros em uma base de dados. Pode transformar ou decompor grandes e complexos problemas em subproblemas mais simples de uma forma recursiva, assim os subproblemas também podem ser decompostos quantas vezes forem necessárias para que uma melhor análise seja realizada. Após a construção da árvore de decisão, os resultados obtidos, são formados por dados organizados de maneira simples e de fácil entendimento e podem servir como ferramenta de apoio à tomada de decisão. Sua resposta possui visualização gráfica simbólica e compreensível.

Uma árvore de decisão é basicamente uma série de declarações *if-then*, que quando aplicados a um registro de uma base de dados, resultam na classificação daquele registro.

Figura 7 - Composição de uma árvore de decisão




Fonte: do autor

A figura 7 exemplifica uma árvore hipotética e é composta de:

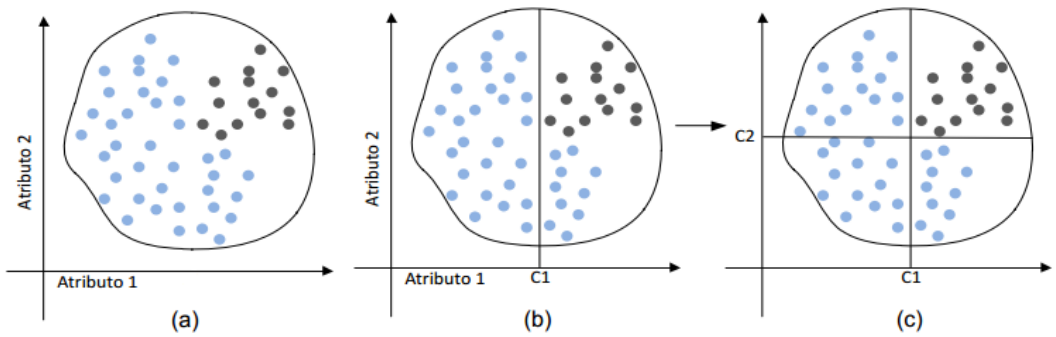
- Raiz: é o nó topo da árvore;
- Nós: são todos os elementos que estão conectados por ramos;
- Ramos: são ligações entre os nós;
- Folhas: representam as classes do conjunto de dados.

### 3.4.3 Top-Down Induction of Decision Tree - TDIDT

Árvores de decisão são construídas com base no modelo hierárquico Top-Down Induction of Decision Tree, ou TDIDT, isto é, do nó raiz em direção às folhas. Em geral, há diferenças na forma de realizar os passos, mas os algoritmos utilizam o processo de indução de uma Árvore de Decisão baseado na estratégia “dividir para conquistar”.

A técnica consiste em dividir o espaço definido pelos atributos em subespaço menores, podendo, cada subespaço, ser subdividido em novos subespaços ou associados a uma classe. A seguir, na figura 8, exemplificam-se os passos de como ocorre esse processo de “dividir para conquistar”, na classificação de duas classes. Na figura 8 (a), são mostrados os atributos; na 8 (b), o valor de um atributo, que deve ser comparado com a constante C1, dividindo o espaço em duas partes, para a separação em classes; porém, não foi suficiente para que houvesse completa separação, sendo necessária uma nova fronteira, como em 8 (c).

Figura 8 - Processo de classificação de duas classes utilizando árvore de decisão: espaço com os atributos (a), obtenção da primeira fronteira de decisão (b), e segunda fronteira de decisão (c).




Fonte: Adaptada de BORGES, 2013

### 3.4.4 Seleção de atributos

No processo de construção de uma árvore de decisão faz-se necessário a escolha correta do atributo preditivo, que definirá o sucesso do algoritmo de indução. São vários os critérios de seleção, sendo isso, uma das variações entre distintos algoritmos de indução, de árvores de decisão, que são definidos em termos da distribuição de classe dos exemplos, antes e depois da divisão.

A maioria dos algoritmos de indução de árvores de decisão procuram obter o melhor atributo a ser utilizado num nó, através da utilização de métodos que são responsáveis por verificar cada atributo candidato, selecionando aquele que melhor discrimine uma classe.

Entre os vários critérios de seleção de um atributo candidato a nó, será abordado neste trabalho: *Ganho de Informação*, *Razão de Ganho* e *Gini*. A maior parte dos algoritmos de indução busca dividir os dados de um nó, de forma a minimizar o grau de impureza dos nós filhos. Quanto menor o grau de impureza, mais desbalanceada é a distribuição de classe. Num determinado nó, a impureza é nula, se todos os seus exemplos pertencerem à mesma classe. Analogamente, o grau de impureza é máximo no nodo, se houver o mesmo número de exemplos para cada classe possível.

O critério de ganho de informação está fundamentado em uma medida conhecida como *Entropia*, desenvolvida por Quinlan em 1993 e pode ser definida como a medida de informação calculada pelas probabilidades de ocorrência de eventos individuais ou combinados.

A entropia caracteriza a (im)pureza de uma coleção despótica de exemplos. Para medir a diminuição da entropia, deve-se comparar o grau de entropia do nó-pai antes da divisão,

com o grau de entropia do nó-filho, após a divisão e, então, o atributo que gera uma diferença maior, é escolhido como condição teste.

O grau de entropia é definido por:

$$entropia(t) = -\sum_{t=1}^{k} p(i|t) \log_2 p\,(i|t) \qquad (3,1)$$

Em que:

p(i | t) é a fração dos exemplos pertencentes à classe i, no nó t;

k é o número de classes.

Quanto maior a entropia de um atributo, mais uniforme é a distribuição dos seus valores; entropia próximo de zero indica que as classes são pouco uniformes; entropia igual a zero significa que ocorreu apenas uma classe no conjunto de dados e será igual à um se o número de amostras de cada classe for igual.

***Ganho de Informação –*** O ganho de informação é uma medida baseada na impureza, e mede a redução da entropia, causada pela partição do conjunto. No processo de construção da árvore, o atributo que possuir o maior ganho de informação deve ser colocado no nó raiz da árvore, pois será este atributo que fornecerá a maior redução na entropia, possibilitando a classificação dos dados de maneira rápida.

O ganho é dado pela soma das entropias individuais menos a entropia conjunta, sendo uma medida de correlação entre duas variáveis. Segundo Castanheira, é uma propriedade estatística que mede como um determinado atributo separa as amostras de treinamento de acordo com sua classificação.

Deve ser calculada a entropia conjunta (para todo conjunto de dados) e a entropia individual (para cada atributo do conjunto de dados), para assim, poder determinar o valor do ganho de informação, que é dado por:

$$ganho = entropia(pai) - \sum_{j=1}^{n} \frac{N(v_j)}{N}\, entropia\left(v_j\right) \qquad (3,2)$$

Em que:

n é o número de valores do atributo, ou seja, o número de nós filho;

N é o número total de exemplos do nó pai;

$N(v_j)$ é o número de exemplos associados ao nó filho $v_j$.

O uso do ganho de informação pode gerar um problema: ele dá preferência a atributos com muitos valores possíveis, pois, ao utilizar um atributo totalmente irrelevante, seria criado um nó para cada valor possível, e o número de nós seria igual ao número de identificadores. Assim, cada um desses, teria apenas um exemplo, o qual pertence a uma única classe, e os exemplos seriam totalmente esclarecidos. O valor da entropia seria mínimo porque, em cada nó, todos os exemplos pertencem à mesma classe, e essa divisão geraria um ganho máximo.

***Razão de Ganho*** **–** Em QUINLAN (1993), foi proposta uma solução para o problema do ganho de informação: um ganho de informação relativo (ponderado) como critério de avaliação. Abaixo, a equação que define a razão de ganho:

$$Razão\ de\ Ganho(t) = \frac{ganho}{entropia(t)} \tag{3,3}$$

Em que:

t é um nó;

Para a entropia igual a zero, a razão de ganho não pode ser definida;

Quanto menor a entropia, maior a razão de ganho.

Quilan sugere que a razão de ganho seja realizada em duas etapas. É feito o cálculo do ganho de informação para todos os atributos e, após, se deve considerar apenas aqueles atributos, cujo ganho de informação esteve acima da média para escolher aquele que apresentar melhor razão de ganho.

***Gini*** **–** Desenvolvido por *Breiman*, em 1998, utiliza-se esse índice de dispersão estatística, para minimizar a impureza de cada nó. A impureza do nó é máxima quando todas as classes possuem igual distribuição e é mínima quando existe apenas uma classe.

O índice *Gini* $gini_{index}$ é definido pela equação a seguir:

$$gini_{index}(nó) = 1 - \sum_{t=1}^{k} p(i|t)^2 \tag{3,4}$$

Em que, k é o número de classes.

Semelhante ao ganho de informação, o Gini também precisa ser calculado por uma diferença, e é entre o $gini_{index}$ antes e após a divisão. É selecionado o atributo que gerar um maior valor para Gini. Segue a equação abaixo para representar essa diferença:

$$Gini = gini_{index}(pai) - \sum_{j=1}^{n} \frac{N(v_j)}{N} gini_{index}(v_j) \quad (3,5)$$

### 3.4.5 Poda

Na construção de árvores de decisão, os ramos podem refletir ruídos ou erros, ocasionando um problema chamado overfitting ou sobreajuste. O overfitting ocorre quando houve um aprendizado bem específico do conjunto de treinamento, não permitindo ao modelo generalizar.

Para evitar o overfitting, muitos algoritmos se valem de uma técnica conhecida como "Poda", do inglês pruning, que é a redução do tamanho da árvore de decisão. Consiste em eliminar alguns ramos da árvore, com base em medidas estatísticas, podendo ocorrer com a árvore já concluída (pós-poda), com a eliminação de alguns ramos considerados desnecessários ou durante a construção dela (pré-poda), com a introdução precoce de um nó em ramos com baixa importância estatística, por exemplo.

A poda melhora a taxa de acerto do modelo para novos exemplos, os quais não foram utilizados no conjunto de treinamento, tornando-se mais simples e de fácil interpretação para o usuário. Apesar de ser necessária a poda, deve-se tomar cuidado para que não ocorra o underfitting, que é quando a árvore é podada demais, e o modelo de classificação não aprende o suficiente sobre os dados de treinamento. Com a etapa de seleção, a poda varia de acordo com diferentes algoritmos de árvore de decisão.

### 3.4.6 Avaliação dos Classificadores

Um algoritmo de indução de árvores de decisão recebe como entrada um conjunto de treinamento para gerar as árvores de decisão, que se utilizam do conhecimento adquirido para prever os valores das classes de exemplos até então desconhecidos. Porém, na previsão dos valores das classes, podem ocorrer classificações equivocadas, tornando-se necessária a avaliação da qualidade da árvore.

Para avaliar o desempenho de uma árvore de decisão, podem ser utilizadas observações que mostram a qualidade da árvore, como a Matriz de Confusão, que é uma matriz que ilustra a qual classe cada exemplo pertence e também, a qual classe ele foi classificado pelo classificador. Nas matrizes com apenas duas classes, uma delas pode ser considerada positiva e a outra, negativa. A tabela 3 representa uma matriz de confusão de duas classes.

Tabela 3 - Matriz Confusão de duas classes

| | Classe Predita | |
|---|---|---|
| Classe Real | Positiva | Negativa |
| Positiva | VP | FN |
| Negativa | FP | VN |

Fonte: (BASGALUPP, 2010)

Em que:

- VP – Verdadeiros Positivos: são exemplos que pertencem à classe positiva e foram classificados corretamente;
- FP – Falsos Positivos: exemplos que pertencem à classe negativa e foram classificados positivos pelo classificador;
- FN - Falsos Negativos: exemplos que pertencem à classe positiva e foram classificados como negativos pelo classificador;
- VN - Verdadeiros Negativos: exemplos que pertencem à classe negativa e foram classificados negativos pelo classificador.

O que se espera nos resultados da matriz é que as taxas de sucesso para Verdadeiro Positivo e Verdadeiro Negativo sejam altas e as outras, baixas. Fazendo uma relação entre VP, FP, VN e FN, define-se uma métrica de desempenho para as taxas de acerto e erro, a acurácia.

A acurácia estima a probabilidade do classificador em acertar suas previsões. Análogo à acurácia, o desempenho de um modelo de classificação pode ser expresso em termos de sua taxa de erro, sendo que o complemento da acurácia estima a probabilidade do classificador em

errar suas previsões. A acurácia pode ser definida como o número total de amostras, especificadas corretamente pelo número total de classificações:

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \times 100\% \quad (3,6)$$

### 3.4.7 Algoritmos

***Random Forest –*** Random Forest, ou Floresta Aleatória, foi proposto por Breiman, em 2001, e, publicado no periódico Machine Learning. Consiste numa técnica de agregação de classificadores do tipo árvore, construídos de forma que a estrutura esteja composta de maneira aleatória, formados por um conjunto de árvores de decisão.

Para determinar a classe de uma instância, o método combina o resultado de várias árvores de decisão, por meio de um mecanismo de votação. Cada árvore dá um voto que indica sua decisão sobre a classe à qual pertencerá determinado objeto. O objeto pertencerá à classe que obtiver o maior número de votos, entre todas as árvores da floresta.

O modelo gerado elege a classe mais frequente entre as opções individuais de cada árvore. Desse modo, a seleção dos atributos é feita no instante de construção do modelo de classificação, identificando a seleção do tipo embutida (embedded). A vantagem desse classificador é que ele permite bases de dados com um número grande de atributos. Contudo, é suscetível ao overfitting em determinadas bases.

***CART –*** O Cart é um algoritmo proposto por Breiman et al. (1984), técnica que consiste em indução de árvores de classificação em atributos nominais e em árvores de regressão em algoritmos contínuos.

Possui a particularidade de ser uma árvore binária, que pode ser percorrida da raiz às folhas, respondendo apenas questões simples, do tipo “sim” ou “não”. Sua leitura e interpretação são de fácil trato e tem como principal característica a “capacidade de gerar árvores de reduzidas dimensões, de elevado desempenho e possuindo grande capacidade de generalização” (GARCIA, p.51, 2003).

# 4 METODOLOGIA

## 4.1 INTRODUÇÃO

No presente capítulo é apresentada uma metodologia para detectar perfis propensos à fraude com base na técnica árvore de decisão. O capítulo explora desde a aquisição dos dados, o pré-processamento com o uso de ferramentas computacionais e a construção do modelo classificador.

A finalidade da utilização da técnica de mineração de dados, árvore de decisão, é a obtenção de um modelo de classificação para detecção de irregularidades no sistema de distribuição da CELPA, classificando-o como fraudador e não fraudador, de acordo com a existência ou não de unidades consumidoras detectadas com fraude.

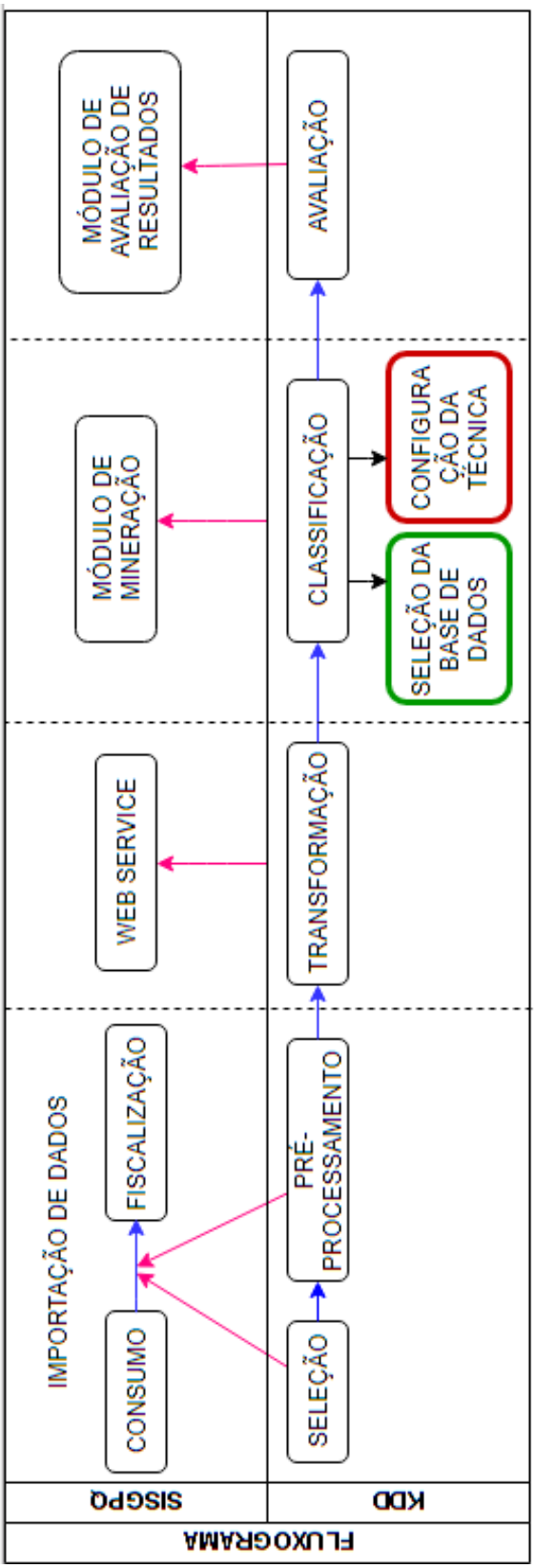
O modelo da árvore de decisão usado neste trabalho, apresenta duas estruturas, uma do tipo binária, o CART, e a outra é o Random Forest, em ambos a classificação final será dada por rótulos “FRAUDE” e “REGULAR”.

A metodologia foi desenvolvida baseada no conceito do KDD, criando um sistema computacional que utiliza técnicas de programação, aprendizagem de máquina e tecnologias de banco de dados.

O sistema computacional para detecção de fraudes utiliza o padrão de *software* MVC (Modelo Visão Controlador). Através do padrão de software adotado foram desenvolvidos módulos específicos para as etapas de importação e filtragem de dados, módulo de classificação e análise de resultados, por meio de gráficos com percentual das unidades consumidoras com suspeitas de fraudes e a localização geográfica das respectivas UC’s por meio do mapa da cidade.

A figura 9 representa um fluxograma mostrando a metodologia utilizada nesse trabalho. No fluxograma está representado as etapas da metodologia e feita uma associação entre elas. O KDD e suas etapas sendo direcionado como funciona no minerador e as respectivas telas exibidas.

Figura 9 – Fluxograma da metodologia utilizada




Fonte: do autor.

## 4.2 IMPORTAÇÃO DOS DADOS

Os dados foram fornecidos pela CELPA e correspondem à Região Metropolitana de Belém, RMB, Pará, entre os anos de 2013 a 2015. São dados comerciais - informações das UC's, como mostrado na tabela 4 - e os dados de fiscalização dos mesmos anos. O arquivo foi recebido nos formatos ".xls", ".xlsx", ".csv" e ".txt". A base de dados contém 31 atributos, são dados, cuja finalidade é organizá-los e extrair informações sobre os consumidores. Para o armazenamento dessas informações foi desenvolvido um módulo de importação tanto para consumidores quanto para fiscalização.

O arquivo da base de fiscalização disponibilizado pela Celpa contém informações das fiscalizações realizadas *in loco* nos anos de 2013, 2014 e 2015 contendo a identificação da UC, o rótulo de status da UC, se regular ou irregular, a data de fiscalização, informação do código de retorno da vistoria e se ocorreu a troca do medidor.

O fluxograma da figura 10 exibe o processo de classificação das UCs com perfil de fraude. O funcionamento do módulo de classificação, tem opção de utilização de modelos criados pela equipe do projeto ou de modelos desenvolvidos pelo próprio usuário através do SISGPQ, as outras opções disponíveis são a classificação de toda RMB ou a classificação por filtros mais específicos como Bairro, Tipo de Ligação, Subgrupo de Tensão Faturamento, Subgrupo de Tensão Origem, Classe de Consumo.

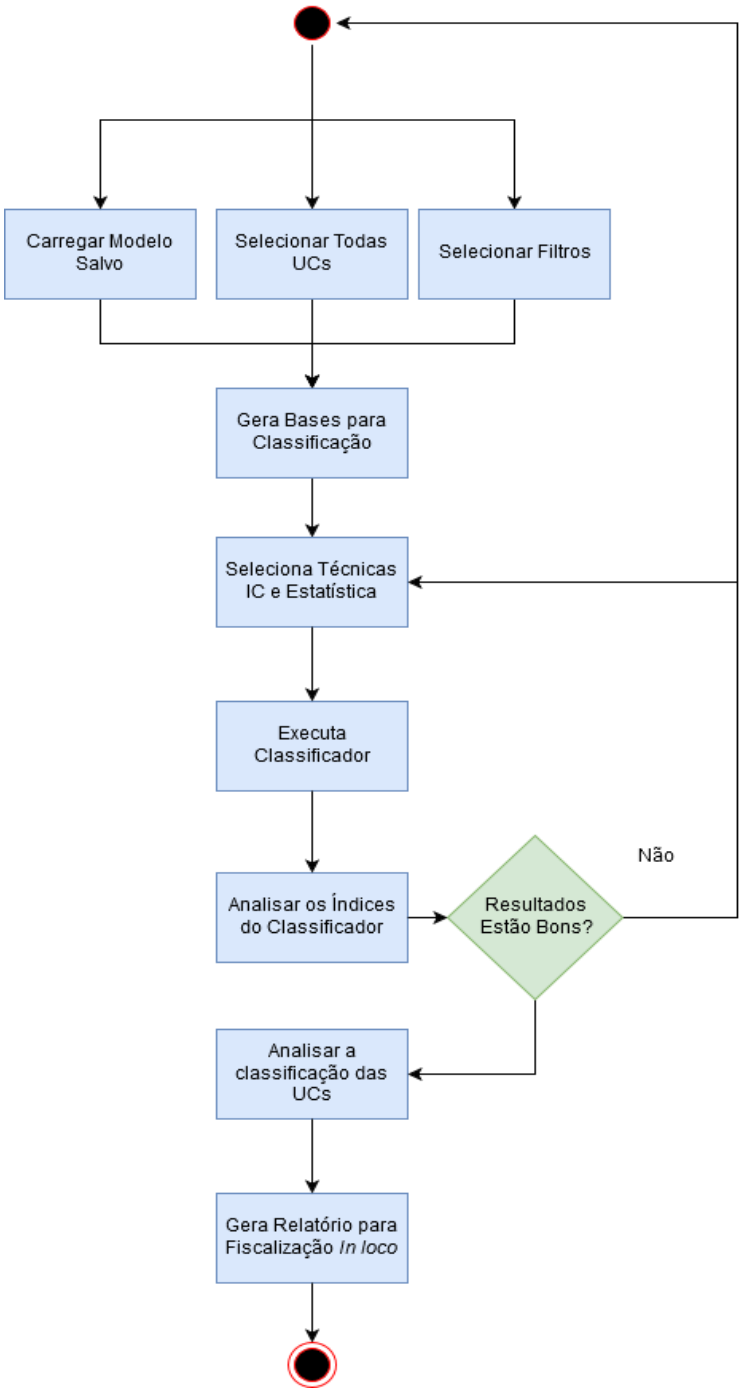
Tabela 4 - Dados Comerciais da Região Metropolitana de Belém fornecidos pela Celpa.

| Atributo | Descrição |
|---|---|
| Unidade Consumidora | Conjunto de instalações e equipamentos elétricos caracterizado pelo recebimento de energia elétrica em um só ponto de entrega, com medição individualizada e correspondente a um único consumidor, podendo estar na situação ligada ou desligada, com ou sem aparelho de medição. |
| Grupo de tensão origem | Fornecimento de tensão: A para tensões acima de 2,3 kV, e B quando é abaixo de 2,3 kV. |
| Subgrupo de tensão origem | Grupo A varia de A1 à A4, em diferentes níveis de tensão; Grupo B varia de B1 à B4, caracterizado se é residencial, rural, comercial ou iluminação pública. |
| Grupo de tensão faturamento | Fornecimento de tensão: A para tensões acima de 2,3 kV, caracterizados por tarifa binômia e B quando é abaixo de 2,3 kV, caracterizados por tarifa monômia. |
| Subgrupo de tensão faturamento | Grupo A varia de A1 à A4, em diferentes níveis de tensão; Grupo B varia de B1 à B4, caracterizado se é residencial, rural, comercial ou iluminação pública. |
| Código Tipo Tarifa | De 1 a 4. |
| Status da UC | Ligada ou Desligada. |
| Regional | A área de concessão é dividida em regiões. Tem-se, Norte, Nordeste, Sul e Oeste. |
| Município | Divisão administrativa de um estado, distrito ou região, com autonomia administrativa e constituído de órgãos político-administrativos próprios. |
| Localidade | Área de um município, distrito. |
| Etapa | É a divisão lógica dentro de uma localidade que funciona como facilitador para a tomada de leituras. |
| Livro | Conjunto de várias unidades consumidoras ordenados de composição lógica, chamado de Rotas, com mesmo calendário de faturamento. |
| Sequência | É a ordem em que o leiturista deve seguir ao fazer a leitura das unidades consumidoras de um livro. |
| Bairro | Cada uma das partes em que se divide uma cidade ou vila, para facilitar a orientação e possibilitar administração pública eficaz. |
| Fase de Ligação | Monofásico, Bifásico e Trifásico. |
| Classe de Consumo | Se Comercial, Industrial, Residencial, Rural, Poder Público ou outros. |
| Atividade | Atividade desenvolvida na unidade consumidora. |
| Tarifa | Convencional, Horária Azul, Horária Verde, Branca e Social. |
| EQUIP_PRIN | Medidor de consumo de energia elétrica. |
| Consumo Janeiro a Dezembro | Consumo nos 12 meses do ano. |

Fonte: do autor.

Após a seleção da base de dados escolhida para classificação o usuário escolhe entre as duas métricas estatísticas, média e desvio padrão, para ser utilizada pelo algoritmo de classificação. O algoritmo de classificação do SISGPQ para mineração de dados foi adaptado a partir dos algoritmos de árvore de decisão disponíveis na biblioteca Scikit-Learn do Python, sendo possível utilizar o Cart e Random Forest.

Figura 10 - Fluxograma do Módulo do Minerador




Fonte: do autor

## 4.3 PRÉ-PROCESSAMENTO E TRANSFORMAÇÃO DOS DADOS

Na etapa de pré-processamento é importante a formação da base de dados e para isso faz se necessário o uso de um bom sistema. Nesse trabalho foi escolhido o Sistema de Banco de Dados MongoDB, que mapeia o problema de grande volume de dados e armazenamento de dados não estruturados, tornando o banco de dados flexível à entrada de novos campos de informações.

O MongoDB armazena informações em documentos no formato BSON (Binary JSON), esse formato suporta estruturas como *arrays* e *embedded objects*. Os Documentos do MongoDB podem ser armazenados dentro de coleções (*collections*), onde são efetuadas operações de busca (*query*) e indexação (*indexing*). O BSON torna o MongoDB mais rápido fazendo um computador processar e efetuar pesquisas nos documentos com eficácia.

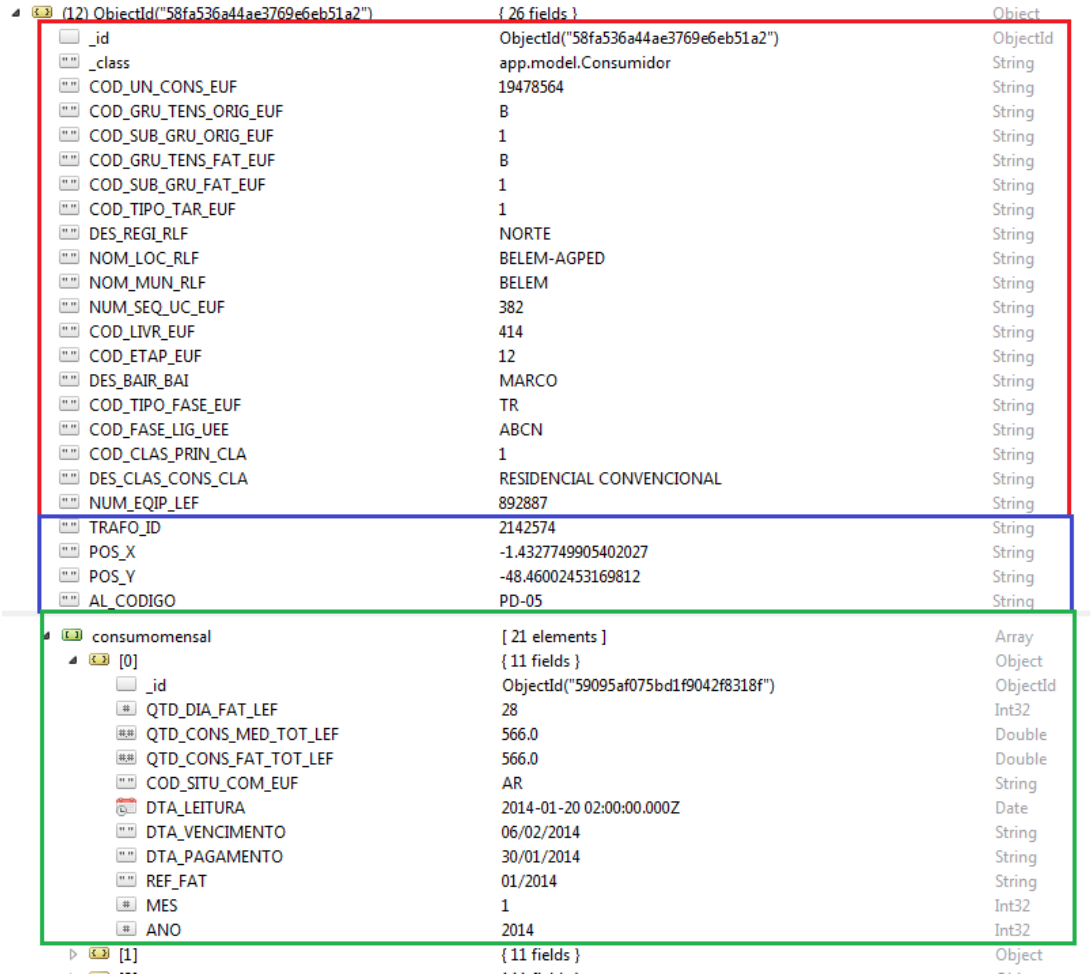
Todas as informações referentes às UC's estão armazenadas em um único documento, podendo ser atualizado individualmente ou alterado de modo independente de qualquer outro documento. A figura 11 mostra a estrutura do banco de dados MongoDB, utilizando a base comercial já armazenada.

No primeiro quadro em vermelho encontram-se as informações gerais das UC's, através desses dados são selecionados os atributos para entrada do minerador sendo eles, sub-grupo de tensão de origem, sub-grupo de tensão de faturamento, fase de ligação e código da classe de consumo.

No segundo quadro em azul encontram-se informação de localização das UC's esses dados permitem mostrar no mapa da cidade a localização das UC's.

No terceiro quadro em verde encontram-se informações sobre o consumo de energia, cada novo mês importado para base de dados cria um objeto dentro do *array* "consumomensal", que ficam as informações do mês, ano, data de pagamento, data de vencimento, data de leitura, quantidade de dias correspondentes a fatura, quantidade consumo medido e a quantidade de consumo faturado, sendo esse último utilizado para os cálculos estatísticos.

Figura 11 - Base de Dados no MongoDB




Fonte: do autor

É utilizado o WebService que mapeia todo o tráfego de informações entre o banco de dados e a camada de núcleo do software para a classificação de fraudes. No WebService foram desenvolvidas funções para a otimização do processo de envio e recebimento de dados do software. Entre as funções implementadas estão:

- Importação e tratamento de dados ausentes nos arquivos de entrada
- Criação das *collections* para base de treinamento e base de validação.
- Criação de *query* de busca por bairro.
- Criação de *query* de consulta do histórico de fiscalização.
- Criação de *query* de consulta do histórico de consumo.

Após a etapa de importação, o sistema computacional gera no banco de dados, MongoDB, através do software WebService, duas *colletions*, a primeira *colletion* é formada com a série temporal utilizando os cálculos de média e desvio padrão de cada mês para cada UC. A segunda *colletion* contém as UC's com fiscalização ocorrida no último mês junto com série temporal com cálculo de média e desvio padrão de cada UC.

Na metodologia adotada utiliza-se o método de sumarização, que tem como objetivo encontrar uma descrição simples e compacta dos dados. Para isso, podem ser utilizadas desde medidas estatísticas simples, como mínimo, média e desvio padrão, até técnicas sofisticadas de visualização e de determinação de relações funcionais entre atributos (HAN; PEI; KAMBER, 2011; MIRKIN, 2011).

O algoritmo de mineração de dados foi desenvolvido em *Python* utilizando os algoritmos de pré-processamento e árvore de decisão da biblioteca *Scikit-Learn* e para cálculos a biblioteca *Numpy*. O algoritmo desenvolvido foi criado para receber de entrada as variáveis provenientes da interface do software de mineração, sendo elas:

- Base de Treino: contém informações das UC's com rótulo de fiscalização junto com série temporal de mês a mês com cálculo das médias e desvio padrão e as informações dos demais atributos padrões. Essa base é utilizada para treinamento do modelo;
- Base de Teste: contém informações das UC's com rótulo de fiscalização junto com série temporal de mês a mês com cálculo das médias e desvio padrão e as informações dos demais atributos padrões. Essa base é utilizada para teste do modelo;
- Base de Validação: contém informações das UC's com série temporal de mês a mês com cálculo das médias e desvio padrão e as informações dos demais atributos padrões sem o rótulo de fiscalização. Essa base é utilizada para classificação das UC's como suspeitas de fraudes;
- Variável Estatística: o algoritmo recebe a variável que será utilizada como entrada do cálculo dos atributos mês a mês, entre as opções média e desvio padrão;
- Variável Árvore de Decisão: o algoritmo recebe a variável que será utilizada para escolha da técnica de árvore de decisão que será utilizada entre CART e Random Forest;

- Parâmetros da Árvore de Decisão: o algoritmo recebe os parâmetros selecionados na interface do classificador.

A partir desse algoritmo fonte foi criado pelo software o código.py com todas as variáveis selecionadas pelo usuário, estando prontas para execução do classificador. Os itens abaixo representam o funcionamento do algoritmo de mineração de dados.

1. Recebe a base de treino;
2. Faz o pré-processamento e conversão da base;
3. Lê a variável estática;
4. Lê e seta as colunas da base de entrada;
5. Divide a base entre treino e teste através do cross-validation;
6. Executa as técnicas de árvore de decisão e os parâmetros escolhidos;
7. Monta a matriz confusão;
8. Cria csv com a classificação da base de treino;
9. Cria csv com a classificação da base de teste;
10. Lê a base de validação;
11. Rotula a base de validação;
12. Cria csv com a classificação da base de validação.

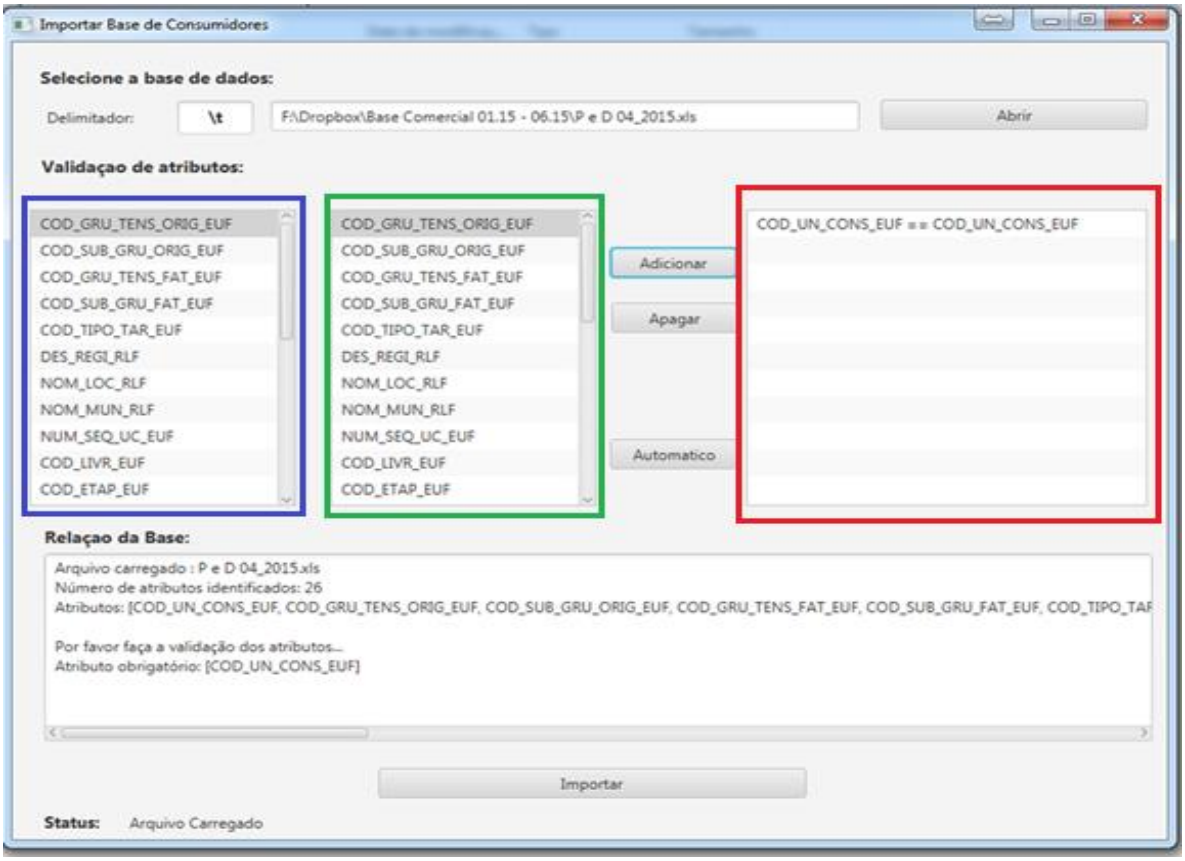
## 4.4 MINERAÇÃO DE DADOS

No software desenvolvido, a mineração de dados para detecção de fraudes funciona através da integração dos módulos de mineração e importação de dados. O módulo de importação de dados é subdividido entre os dados dos consumidores, do qual é importado da base comercial da Celpa – contendo informações sobre as unidades consumidoras – e os dados de fiscalização.

Ocorre formação das bases a partir do cruzamento das informações das UC’s, com histórico de consumo e o histórico de fiscalização que são armazenadas no banco, a partir dos dados que serão carregados pelo usuário, tendo a tarefa de alimentar a base a partir dos módulos de importação específicos.

Para a tarefa de importação de dados comerciais, deve-se “ selecionar a base de dados” como mostra a figura 12 exibindo a interface gráfica do software. Feita a seleção da base de dados, o próximo passo é validar os atributos, que são mostrados no quadro azul da área “Validação de atributos”, e no quadro verde estão os itens mapeados como padrões, normalmente similares em todas as bases apresentadas. Deve-se então interligar as duas colunas com os respectivos atributos que irão sendo exibidos no quadro vermelho, através do botão “automático” – interligação das colunas que possuem o mesmo nome - ou do botão “adicionar” – adiciona atributos específicos.

Figura 12 – Tela de importação Base de Consumidores



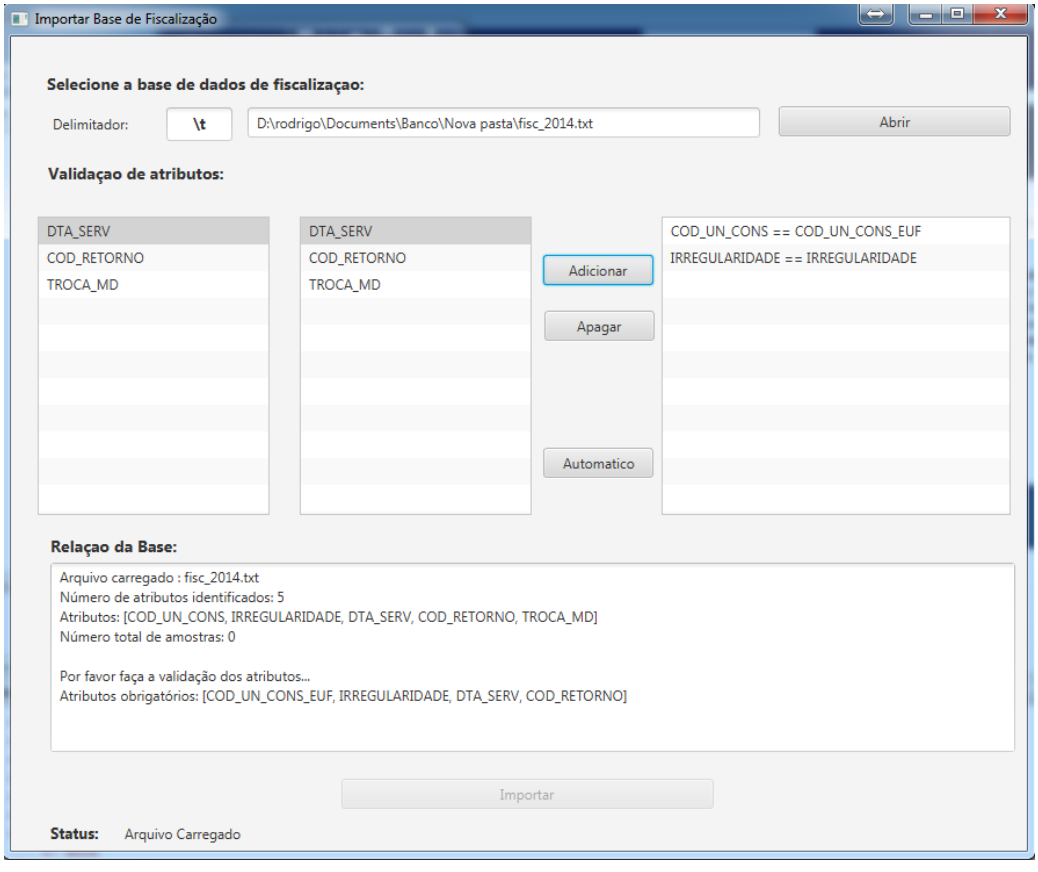

Fonte: do autor.

Na área “Relação da Base”, são exibidas as informações gerais da base selecionada, como o nome do arquivo, número de atributos, quais atributos a base contém e as observações para submeter a base. Em seguida, o botão “Importar” é habilitado e o usuário poderá submeter a base comercial para o banco de dados.

A figura 13, mostra a tela da importação dos dados da base de fiscalização. Na importação de dados da fiscalização, os dados contêm a seleção da base de dados, a parte da validação de atributos em que os atributos padrões são:

- Código da Unidade Consumidora: Número da conta contrato;
- Data do serviço de fiscalização: Data da inspeção in loco;
- Código do serviço: Código do serviço realizado;
- O *label* do estado da UC: “0” caso a UC esteja normal e “1” caso ela cometa algum tipo de irregularidade;
- Troca do medidor: “0” caso o medidor não tenha sido trocado e “1” caso o mesmo tenha sido substituído;
- Código de Retorno: são diversos códigos utilizados pela equipe de fiscalização sobre o status da UC; a tabela 5 mostra os códigos de retorno mais encontrados.

Figura 13 – Tela de importação da base de fiscalização




Fonte: do autor.

Tabela 5 – Códigos de Retorno mais frequentes com suas respectivas descrições.

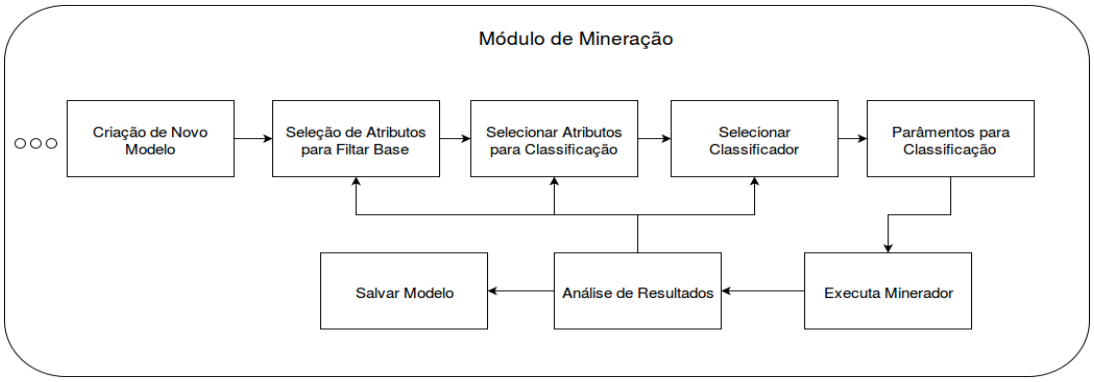
| CÓDIGO | DESCRIÇÃO |
|---|---|
| 565 | Desvio antes do Medidor |
| 300 | Medição Normal |
| 0 | Não há irregularidade |
| 371 | Manutenção em Rede |
| 515 | Medidor Avariado |
| 303 | Deficiência no Padrão de Entrada |
| 567 | UC desligada no sistema e ligada em campo à revelia da Celpa |
| 554 | Ligação Clandestina |
| 203 | Ligação Direta Equipe do Plantão |
| 568 | Interligação entre Linha e Carga |
| 201 | Medidor com Defeito |

Fonte: do autor.

Feita a validação dos atributos, importa-se as informações para o banco de dados. Logo após, é realizado o processo de mineração, podendo ser criado um novo modelo para a mineração, ou pode-se utilizar um modelo *default* com melhor classificação para a RMB.

A figura 14 abaixo, mostra um fluxograma do primeiro processo de mineração que parte do “zero” na criação de um novo modelo. Nesse processo, tem-se todas as etapas de seleção de atributos para filtrar a base, seleção de atributos para classificação, escolha do classificador, configuração dos parâmetros do classificador, o processo de análise dos resultados e a opção de salvar o modelo criado, bem como a consolidação dos dados classificados na base de dados.

Figura 14 – Fluxograma Modelo Novo

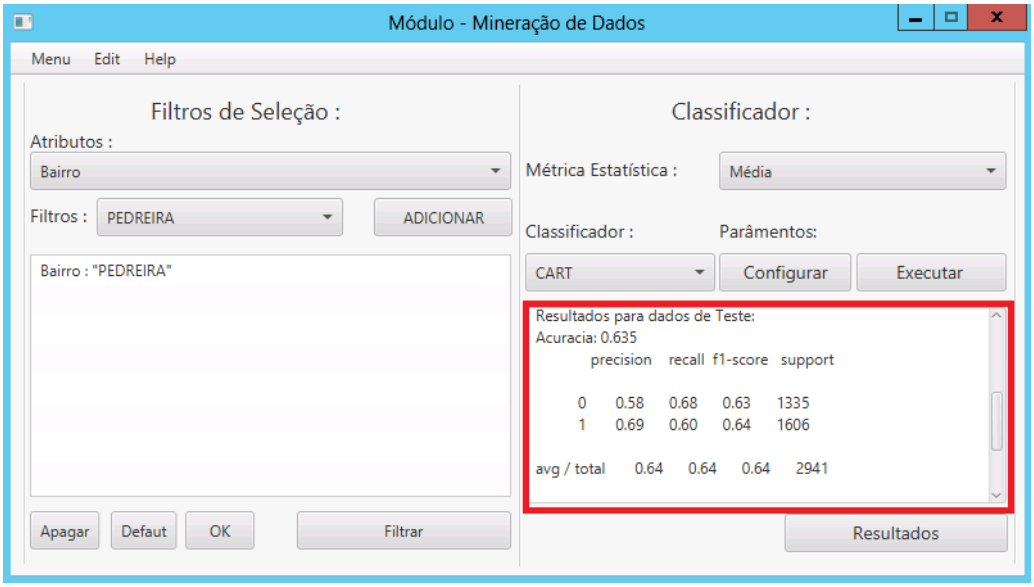



Fonte: do autor.

Em seguida, é exibida a tela da figura 15, “Filtros de Seleção e Mineração”, e para a aba de “Filtros de Seleção” são mostradas opções dos atributos que podem ser selecionados para a criação da base de dados mais específica, sendo as opções oferecidas para o usuário:

- Bairro;
- Sub grupo de tensão origem;
- Sub grupo de tensão faturamento;
- Tipo de fase de ligação;
- Código da classe de consumo.

Figura 15 – Tela módulo de mineração de dados: Filtros de Seleção e Mineração




Fonte: do autor.

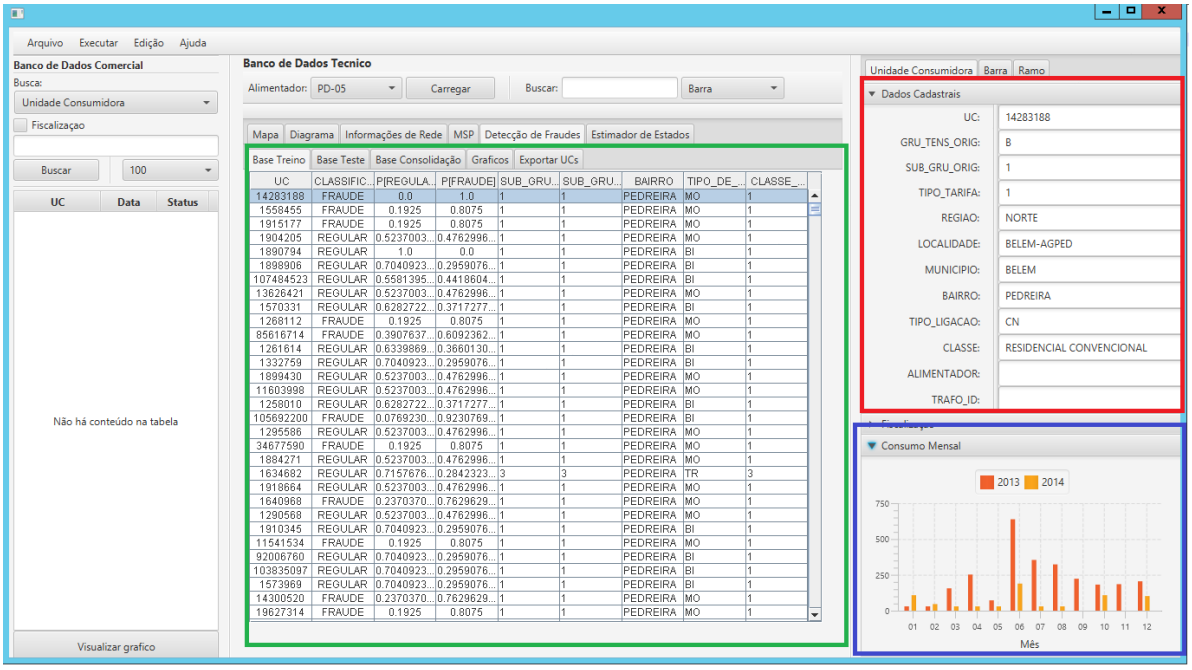
A tela exibe como atributo o bairro, como filtro “PEDREIRA”, fazendo com que todas as UC’s do Bairro da Pedreira sejam selecionadas para o processo de classificação. Em seguida, é feita a seleção do classificador, sendo disponível os classificadores *Random Forest* e *Cart*, após a configuração dos parâmetros e executado o processo de classificação, a tela exibe os “resultados para dados de treino” e “resultados para dados de teste” com as respectivas taxas de acurácia e matrizes confusão.

O minerador usa o “Holdout” para construir o classificador e consiste em dividir os exemplos em uma porcentagem fixa de exemplos $p$ para treinamento e (1-$p$) para teste, considerando normalmente $p$>1/2. Valores típicos são $p$ = 2/3 e (1-p) = 1/3, embora não existam fundamentos teóricos sobre esses valores. (REZENDE, 2005)

Ao clicar em “Resultados” o programa exibe uma tela com um módulo de análise de resultados, encontrada na tela principal do SISGPQ, mostrada na figura 16. É oferecido ao usuário a lista das UC’s classificadas pelo minerador com adição do rótulo de REGULAR ou FRAUDE após o processo de mineração de dados.

O quadro em verde exibido na figura 16, possui a lista com informações das UC’s classificadas, na coluna 1 o número correspondente à UC; na coluna 2 o rótulo de classificação; nas colunas 3 e 4 o grau de probabilidade de assertividade da UC estar regular e da UC estar fraudando respectivamente; as colunas 5 a 9 contém informações de sub-grupo de tensão de origem, sub-grupo de tensão de faturamento, fase de ligação e código da classe de consumo. O quadro vermelho contém todas as informações cadastradas das UC’s. O quadro azul mostra ao usuário o histórico de consumo de energia elétrica da UC selecionada.

Figura 16 – Tela de análise de resultados




Fonte: do autor.

# 5 RESULTADOS

## 5.1 INTRODUÇÃO

Neste capítulo, são apresentados os resultados das simulações realizadas com a árvore de decisão, com dados do ano de 2015 de três bairros da RMB.

São apresentados os resultados das validações das simulações, ou seja, a partir de uma base de dados não usada no treinamento do classificador baseado em árvore de decisão. A base de dados usada no treinamento da árvore de decisão corresponde ao período de 2013 e 2014, através do *holdout*.

Todas as classificações foram realizadas por bairro e os mesmos escolhidos devido à base de dados serem maiores e amostras com semelhanças na complexidade sócio-econômica; são eles: Guamá, Marambaia e Sacramenta.

## 5.2 SIMULAÇÕES

As simulações foram realizadas com parte da base de dados usada para treino e outra para teste, e os resultados obtidos pelo teste do minerador foram praticamente iguais, independente do algoritmo utilizado, Random Forest e CART e da métrica, AVG (média do consumo) e STD (desvio padrão do consumo) utilizadas.

O atributo bairro foi designado para ser o filtro nas classificações, devido a facilidade em validar os resultados de uma região simulada. Na importação de dados, na base de dados da fiscalização foram utilizados todos os atributos e dentre os 31 atributos da base comercial da CELPA, os escolhidos para as simulações foram:

- Unidade Consumidora;
- Grupo de tensão de origem;
- Grupo de tensão de faturamento;
- Fase de ligação;
- Classe de Consumo;
- Atividade;

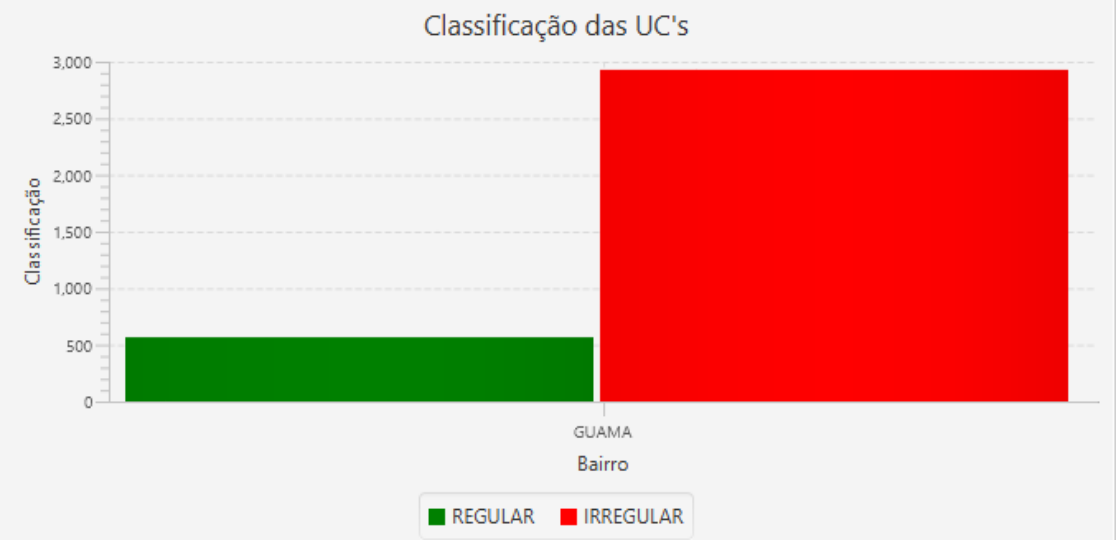
- Consumo.

Para serem selecionados apenas esses atributos, foi levado em consideração as taxas de acerto obtidas e a relevância deles na classificação.

### 5.2.1 Bairro Guamá

No bairro do Guamá, o universo de amostras da base de dados de validação é composto de 2900 UC's irregulares e 600 UC's regulares, totalizando 3500 amostras, como pode ser visto na figura 17.

Figura 17 - Gráfico com o número de amostras do bairro do Guamá.

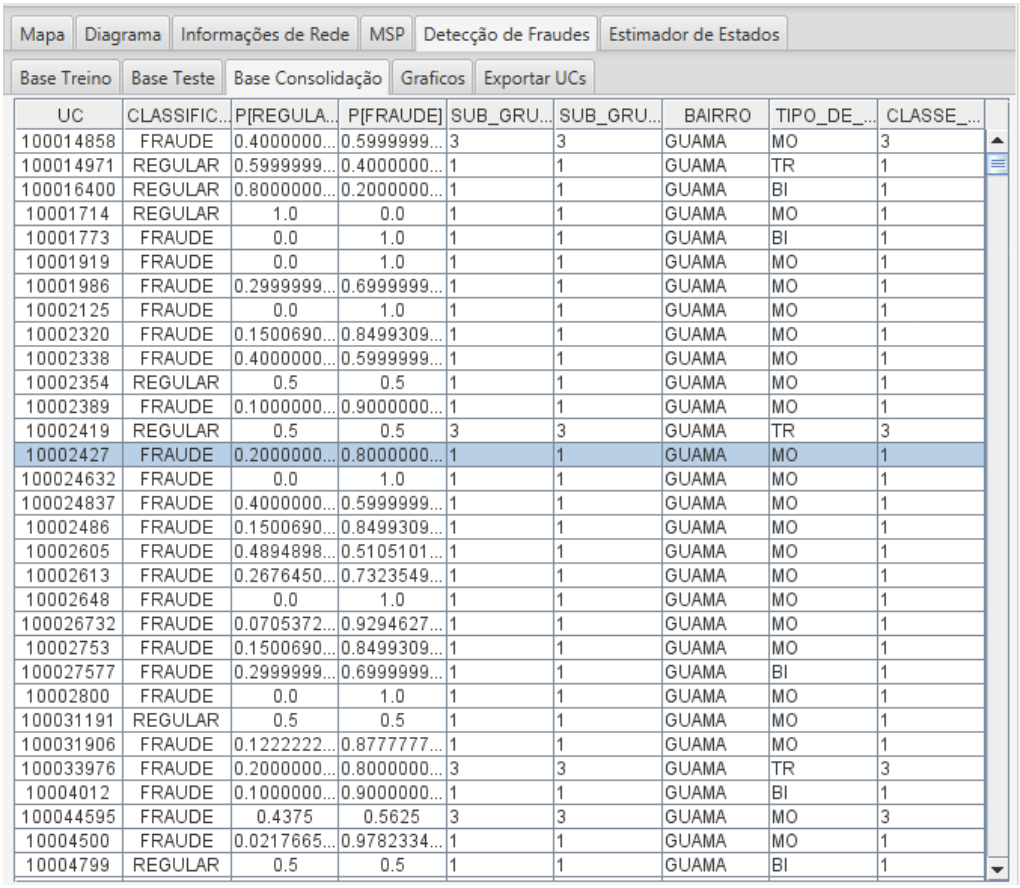



Fonte: do autor.

Para a validação, é necessário um cruzamento entre a informação obtida pelo SISGPQ e os dados de fiscalização do ano de 2015. A figura 18 exibe a tela do programa em que na aba de "Consolidação" estão as UC's rotuladas pelo minerador como "FRAUDE" ou "REGULAR", que pode ser confirmada através da aba "Fiscalização, onde encontram-se as informações das fiscalizações ocorridas nas UC's.

Na figura 18, a UC sublinhada de azul é um exemplo em análise, em que foi rotulada como "FRAUDE" pelo minerador e apresenta índice de probabilidade para fraude 0,8, ou seja, a chance de ser verdadeiro o rótulo é considerável.

Figura 18 - Tela do SISGPQ.

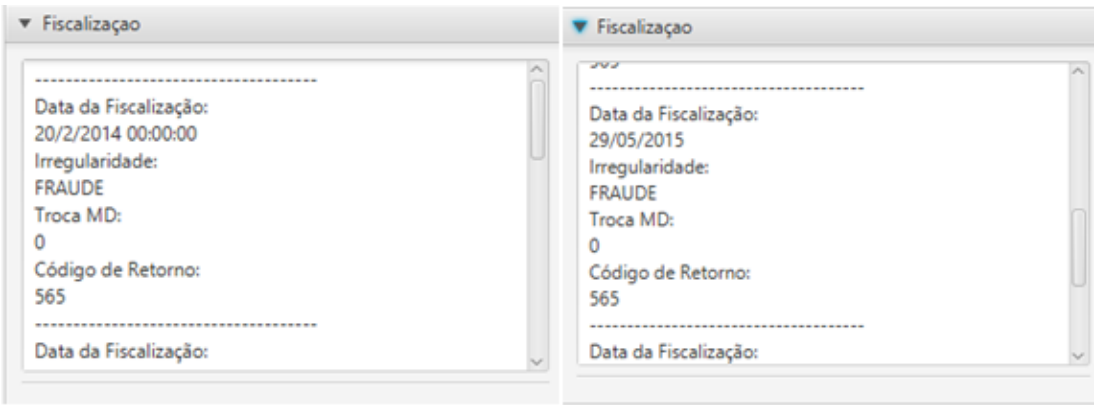



| UC | CLASSIFIC... | P[REGULA... | P[FRAUDE] | SUB_GRU... | SUB_GRU... | BAIRRO | TIPO_DE_... | CLASSE_... |
|---|---|---|---|---|---|---|---|---|
| 100014858 | FRAUDE | 0.4000000... | 0.5999999... | 3 | 3 | GUAMA | MO | 3 |
| 100014971 | REGULAR | 0.5999999... | 0.4000000... | 1 | 1 | GUAMA | TR | 1 |
| 100016400 | REGULAR | 0.8000000... | 0.2000000... | 1 | 1 | GUAMA | BI | 1 |
| 10001714 | REGULAR | 1.0 | 0.0 | 1 | 1 | GUAMA | MO | 1 |
| 10001773 | FRAUDE | 0.0 | 1.0 | 1 | 1 | GUAMA | BI | 1 |
| 10001919 | FRAUDE | 0.0 | 1.0 | 1 | 1 | GUAMA | MO | 1 |
| 10001986 | FRAUDE | 0.2999999... | 0.6999999... | 1 | 1 | GUAMA | MO | 1 |
| 10002125 | FRAUDE | 0.0 | 1.0 | 1 | 1 | GUAMA | MO | 1 |
| 10002320 | FRAUDE | 0.1500690... | 0.8499309... | 1 | 1 | GUAMA | MO | 1 |
| 10002338 | FRAUDE | 0.4000000... | 0.5999999... | 1 | 1 | GUAMA | MO | 1 |
| 10002354 | REGULAR | 0.5 | 0.5 | 1 | 1 | GUAMA | MO | 1 |
| 10002389 | FRAUDE | 0.1000000... | 0.9000000... | 1 | 1 | GUAMA | MO | 1 |
| 10002419 | REGULAR | 0.5 | 0.5 | 3 | 3 | GUAMA | TR | 3 |
| 10002427 | FRAUDE | 0.2000000... | 0.8000000... | 1 | 1 | GUAMA | MO | 1 |
| 100024632 | FRAUDE | 0.0 | 1.0 | 1 | 1 | GUAMA | MO | 1 |
| 100024837 | FRAUDE | 0.4000000... | 0.5999999... | 1 | 1 | GUAMA | MO | 1 |
| 10002486 | FRAUDE | 0.1500690... | 0.8499309... | 1 | 1 | GUAMA | MO | 1 |
| 10002605 | FRAUDE | 0.4894898... | 0.5105101... | 1 | 1 | GUAMA | MO | 1 |
| 10002613 | FRAUDE | 0.2676450... | 0.7323549... | 1 | 1 | GUAMA | MO | 1 |
| 10002648 | FRAUDE | 0.0 | 1.0 | 1 | 1 | GUAMA | MO | 1 |
| 100026732 | FRAUDE | 0.0705372... | 0.9294627... | 1 | 1 | GUAMA | MO | 1 |
| 10002753 | FRAUDE | 0.1500690... | 0.8499309... | 1 | 1 | GUAMA | MO | 1 |
| 100027577 | FRAUDE | 0.2999999... | 0.6999999... | 1 | 1 | GUAMA | BI | 1 |
| 10002800 | FRAUDE | 0.0 | 1.0 | 1 | 1 | GUAMA | MO | 1 |
| 100031191 | REGULAR | 0.5 | 0.5 | 1 | 1 | GUAMA | MO | 1 |
| 100031906 | FRAUDE | 0.1222222... | 0.8777777... | 1 | 1 | GUAMA | MO | 1 |
| 100033976 | FRAUDE | 0.2000000... | 0.8000000... | 3 | 3 | GUAMA | TR | 3 |
| 10004012 | FRAUDE | 0.1000000... | 0.9000000... | 1 | 1 | GUAMA | BI | 1 |
| 100044595 | FRAUDE | 0.4375 | 0.5625 | 3 | 3 | GUAMA | MO | 3 |
| 10004500 | FRAUDE | 0.0217665... | 0.9782334... | 1 | 1 | GUAMA | MO | 1 |
| 10004799 | REGULAR | 0.5 | 0.5 | 1 | 1 | GUAMA | BI | 1 |

Fonte: do autor.

Em seguida, pode-se observar na figura 19 as fiscalizações ocorridas na UC, sendo uma delas em maio de 2015, identificado o código de retorno 565, que segundo a tabela 5 encontrada no capítulo anterior, é um desvio antes do medidor. Pode-se comprovar então que a validação para essa UC obteve sucesso.

Figura 19 –Fiscalizações ocorridas na UC.




Fonte: do autor.

A tabela 6, exibe os resultados com a taxa de acurácia de 0,793 e o total de amostras da base de validação utilizada.

Tabela 6 – Simulação do bairro Guamá.

| Tipo de Base | Acurácia | Total de amostras |
|---|---|---|
| **VALIDAÇÃO** | 0,793 | 3500 |

Fonte: do autor.

Já a matriz confusão pode ser visualizada na tabela 7, em que os perfis regulares que foram classificados como regulares foram de 162 UC's e 2577 UC's com perfil irregular classificadas corretamente. As amostras classificadas incorretamente foram num total de 761.

Tabela 7 - Matriz Confusão do bairro Guamá.

| | Classificados como Regulares | Classificados como Irregulares |
|---|---|---|
| **UC's Regulares** | 201 | 399 |
| **UC's Irregulares** | 123 | 2577 |

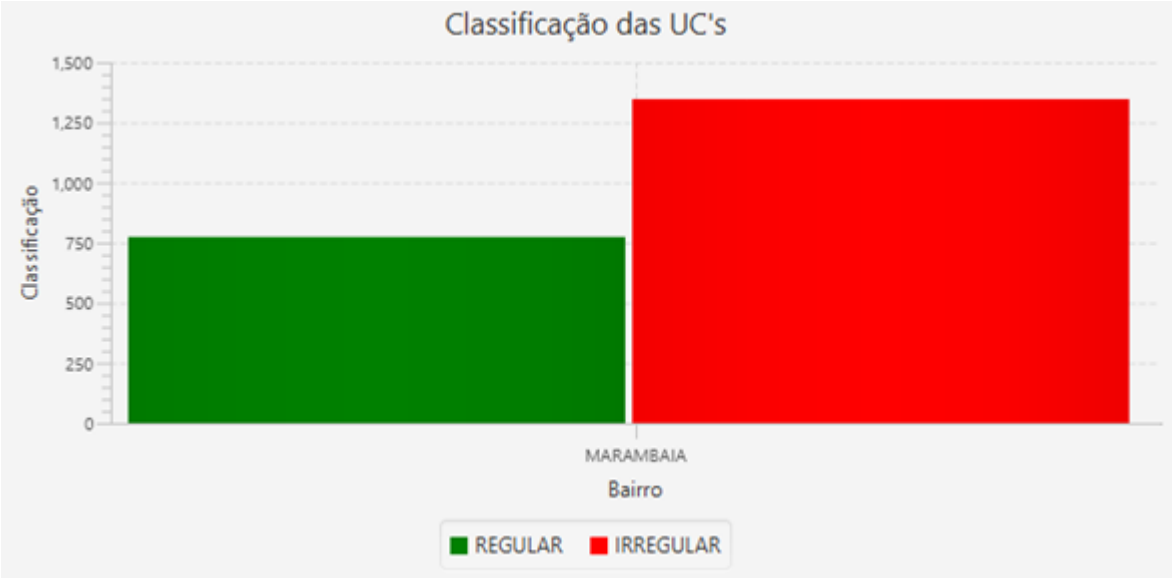
Fonte: do autor.

Para uma rápida avaliação da matriz confusão, pode-se olhar para a diagonal principal da matriz, em que os valores ali depositados correspondem aos acertos do minerador, devendo assim apresentar os maiores valores comparados aos da diagonal secundária, pois ela demonstra os valores dos erros do minerador.

### 5.2.2 Bairro Marambaia

A figura 20 exibe o gráfico com o universo de amostras contidas na simulação da base de dados de validação do bairro da Marambaia no ano de 2015 e apresenta um total de 2150 UC's, sendo 800 UC's regulares e 1350 irregulares.

Figura 20 - Gráfico com o número de amostras do bairro da Marambaia.




Fonte: do autor.

A tabela 8 exibe uma acurácia de 0,65 para a validação desse bairro.

Tabela 8 – Simulação do bairro da Marambaia.

| **Tipo de Base** | **Acurácia** | **Total de amostras** |
|---|---|---|
| **VALIDAÇÃO** | 0,657 | 2150 |

Fonte: do autor.

A seguir, na tabela 9, os resultados da matriz confusão. Foram 255 UC's regulares e 1159 irregulares classificadas corretamente. UC's classificadas incorretamente são 191 irregulares como regulares e 545 regulares classificadas como regulares.

Tabela 9 - Matriz Confusão do bairro Marambaia.

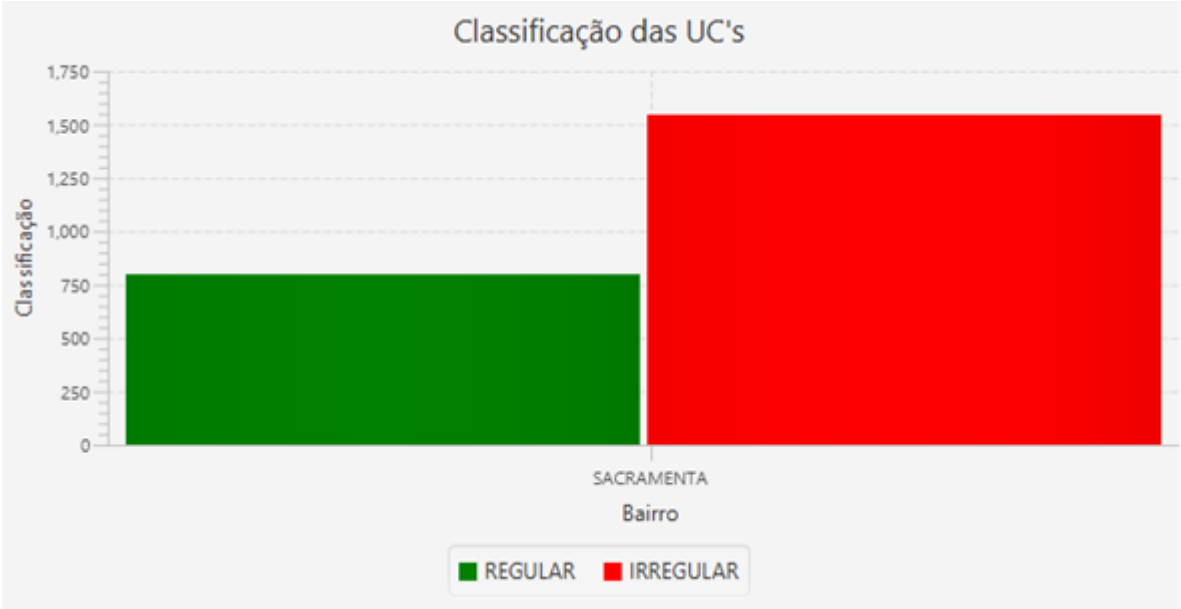
| | Classificados como Regulares | Classificados como Irregulares |
|---|---|---|
| Perfis Regulares | 255 | 545 |
| Perfis Irregulares | 191 | 1159 |

Fonte: do autor.

### 5.2.3 Bairro Sacramenta

As simulações realizadas com a base de dados do bairro da Sacramenta totalizam um universo de 2400 amostras, onde 1600 UC’s são irregulares e 800 UC’s regulares, como mostra o gráfico da figura 21.

Figura 21 – Gráfico com o número de amostras do bairro da Sacramenta.




Fonte: do autor.

A tabela 10 exibe a taxa de acerto obtida nessa simulação, de 0,71.

Tabela 10 – Simulação do bairro da Sacramenta.

| Tipo de Base | Acurácia | Total de amostras |
|---|---|---|
| **VALIDAÇÃO** | 0,71 | 2400 |

Fonte: do autor.

A matriz confusão para essa validação é exibida na tabela 11. Foram 458 UC's regulares e 1246 irregulares classificadas de maneira correta. As classificações incorretas foram de 342 UC's regulares classificadas como irregulares e 354 irregulares como regulares.

Tabela 11 - Matriz Confusão do bairro Sacramenta.

| | Classificados como Regulares | Classificados como Irregulares |
|---|---|---|
| Perfis Regulares | 458 | 342 |
| Perfis Irregulares | 354 | 1246 |

Fonte: do autor.

## 5.3 CONSIDERAÇÕES FINAIS

Analisando os resultados obtidos nas simulações, verifica-se que a taxa de acerto do minerador, varia entre 0,657 a 0,793. Considerando a complexidade do problema de furto e fraude de energia elétrica, dos dados envolvidos e da literatura encontrada, em que a taxa de acerto encontra-se nessa faixa, como em Nagi et all (2010) que utiliza *Support Vector Machines* (*SVM*), obteve 60% de taxa de acerto ao usar dados reais de uma concessionária de energia elétrica na Malásia; e em 2011 ao combinar essa técnica com Fuzzy, obteve 72% de acurácia.

# CAPÍTULO 6 – CONCLUSÕES

## 6.1 CONSIDERAÇÕES FINAIS

No desenvolvimento deste trabalho foi apresentado o problema das perdas comerciais de energia elétrica, a importância do assunto, que é de abrangência mundial, com sua complexidade socioeconômica, a falta de investimento em educação por parte do governo, a cultura popular e a inadimplência são fatores de extrema relevância neste contexto de altas taxas de perdas comerciais. É revelado o problema na concessionária CELPA, (Centrais Elétricas do Pará), uma das empresas com as taxas mais altas no Brasil, revelando a urgência em combater o problema com técnicas mais eficazes rápidas.

A metodologia aqui apresentada para detecção de fraudes no consumo da energia elétrica, utiliza uma técnica de mineração de dados: a árvore de decisão, baseada no KD, sendo seguidas e descritas as etapas, como a seleção da base de dados, o pré-processamento realizado e a mineração de dados, paralelamente com ferramentas necessárias, como WebService e MongoDB.

Neste trabalho foi desenvolvida uma metodologia baseada na árvore de decisão para a detecção de unidades consumidoras com comportamento fraudulento, como uma ferramenta computacional. Primeiramente são selecionadas as bases de dados de consumo e de fiscalização, feito um cruzamento entre elas, para treinar e testar o classificador, obtendo assim um modelo, pronto para classificar uma base de dados nunca antes utilizada pelo programa, podendo assim, rotular UC’s suspeitas de fraudes e auxiliar na tomada de decisão para realizar inspeções *in loco*.

A base de dados possui amostras de diversos perfis de consumidores, que vão desde os residenciais aos industriais, compostas de diversas características de consumo, sendo obtido um perfil para cada consumidor, a partir de uma sumarização do consumo, média e desvio padrão para cada mês do ano, obtendo um padrão, auxiliando o classificador na tomada de decisões.

As simulações realizadas foram com a maior base de dados possível afim de se obter melhores taxas de acerto e uma maior confiabilidade no desempenho da árvore de decisão. As taxas de acerto obtidas refletem a complexidade do problema, do minerador em todos os

bairros classificados e mostra que a metodologia proposta é de grande aplicabilidade e uma poderosa ferramenta.

# REFERÊNCIAS BIBLIOGRÁFICAS

ABRADEE, **“**Uso racional de energia”. Disponível em: <http://www.abradee.com.br/imprensa/noticias/21-setor-de-distribuicao.> Acessado em 26 de julho de 2016.

ANEEL, Agência Nacional de Energia Elétrica. Medição, Faturamento e Combate a Perdas Comerciais. Disponível em: <http://www.aneel.gov.br/visualizar_texto.cfm ?idtxt=1623>. Acesso em 7 de out. 2016.

_____. *Perdas de Energia.* Disponível em: <http://www.aneel.gov.br/area.cfm?idArea=801>. Acesso em 11 de out. 2016.

ARAÚJO, A. C. M. *Perdas e Inadimplência na atividade de distribuição de energia elétrica no Brasil*. 2007. 227 p. Tese (Doutorado em Engenharia Elétrica) – Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2007.

BASGALUPP, M. P. LEGAL. *Tree: um algoritmo genético multi-objetivo para indução de árvore de decisão. Tese (Doutorado em Ciências Matemáticas e de Computação)*. Universidade de São Paulo, São Carlos, 2010.

BORGES, F.A.S. Extração de Características Combinadas com Árvore de Decisão para Detecção e Classificação dos Distúrbios de Qualidade de Energia Elétrica. 2013. Dissertação (Mestrado em Engenharia Elétrica) – Universidade de São Paulo, São Carlos, 2013.

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

CELPA, Equatorial Energia. *Comentários de Desempenho 1T15. 2015.* Disponível em: http://www.equatorialenergia.com.br/. Acessado em 12 de setembro de 2016.

_____. *Comentários de Desempenho 2T15. 2015.* Disponível em: http://www.equatorialenergia.com.br/. Acessado em 12 de setembro de 2016.

_____. *Comentários de Desempenho 3T15. 2015.* Disponível em: http://www.equatorialenergia.com.br/. Acessado em 12 de setembro de 2016.

_____. *Comentários de Desempenho 4T15. 2015.* Disponível em: http://www.equatorialenergia.com.br/. Acessado em 12 de setembro de 2016.

_____. *Comentários de Desempenho 1T16. 2016.* Disponível em: http://www.equatorialenergia.com.br/. Acessado em 14 de setembro de 2016.

_____. *Comentários de Desempenho 2T16. 2016.* Disponível em: http://www.equatorialenergia.com.br/. Acessado em 14 de setembro de 2016.

_____. *Comentários de Desempenho 3T16. 2016.* Disponível em: http://www.equatorialenergia.com.br/. Acessado em 15 de março de 2017.

_____. *Comentários de Desempenho 4T16. 2016.* Disponível em: http://www.equatorialenergia.com.br/. Acessado em 27 de março de 2017.

CORREA, Y. C.S. *Detecção de Irregularidades no Consumo de Energia Elétrica Usando Árvore de Decisão.* 2015. 82 p. Trabalho de Conclusão de Curso (Engenharia Elétrica) – Universidade Federal do Pará, Belém, 2015.

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

FILHO, J. R. *Sistema Inteligente Baseado em Árvore de Decisão, para Apoio ao Combate às Perdas Comerciais na Distribuição de Energia Elétrica*. 2006. 174 p. Dissertação (Mestrado em Engenharia Elétrica) – Universidade Federal de Uberlândia, Uberlândia, 2006.

GARCIA, S. C. *O uso de árvores de decisão na descoberta de conhecimento na área da saúde*. Porto Alegre, 2003.

HAN, J; KAMBER, M. *Data Mining: Concepts and Techniques*. Elsevier, 2006.

HAN, J.; PEI, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. [s.l.] Elsevier Science, 2011.

LAROSE, D. T. *Discovering Knowledge in Data: An introduction to Data Mining*. John Wiley and Sons, Inc, 2005.

LEAL, M. A guerra fria das perdas comerciais. 2012. Disponível em: http://www.krj.com.br/. Acessado em 29 de setembro de 2016.

MIRKIN, B. Core concepts in data analysis: summarization, correlation and visualization. [s.l.] Springer London, 2011.

ORTEGA, G.V.C. *Redes neurais na identificação de perdas comerciais no setor elétrico.* 2008. 184 p. Dissertação (Mestrado em Engenharia Elétrica) - Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2008.

PENIN, C. A. S. *Combate, prevenção e otimização das perdas comerciais de energia elétrica*. 2008. 214 p. Tese (Doutorado em Engenharia Elétrica) – Escola Politécnica da Universidade de São Paulo, São Paulo, 2008.

REIS FILHO, J. *Sistema inteligente baseado em árvore de decisão, para apoio ao combate às perdas comerciais na distribuição de energia elétrica.* 2006. 174 f. Dissertação (Mestrado em Engenharia Elétrica) – Universidade Federal de Uberlândia, Uberlândia, 2006.

REZENDE, S. O. *Sistemas inteligentes: fundamentos e aplicações*. 2005.

SIMÃO, E. *'Gatos' causam prejuízos de R$ 7 bi e encarecem tarifas*. O Estado de São Paulo, Brasília, 12 fev. 2012. Disponível em: <http://www.estadao.com. br/noticias/impresso,gatos-causam-prejuizo-de-r-7-bi-e-encarecem-tarifas-,834 723,0.htm>. Acessado em: 28 de outubro de 2016.

SMITH, T. B. (2004). *"Electricity theft: a comparative analysis".* Energy Policy 32: 2067-2076.

VIEIRALVES, E. X. *Proposta de uma metodologia para avaliação das perdas comerciais dos sistemas elétricos: o caso Manaus.* 2005. 180 p. Dissertação (Mestrado em Planejamento de Sistemas Energéticos) – Universidade Estadual de Campinas, Campinas, 2005.