JOSE EDILSON BRUNO FILHO	SIDARTA SILVA GALAS	MANOLIDIS EFSTRATIOS JUNIOR
Enel Distribuição Ceara	Enel Distribuição Ceara	Enel Distribuição Ceara
joseedilsonbrunofilho@gmail.com	sidartagalas@gmail.com	efstratios777@gmail.com

Ciência de Dados aplicada no combate ao furto de energia elétrica

Palavras-chave

Ciência de Dados
Data Mining
Furto
KDD
Machine Learning
Perdas de energia

Resumo

As perdas de energia elétrica nos últimos anos tem sido assunto prioritário para as distribuidoras do país devido a agressividade de mercado e as formas de como são realizados tais furtos nas redes de energia. Com fraudes cada vez mais sofisticadas, as distribuidoras de energia elétrica têm se deparado com muitas dificuldades para detectar tais problemas, pois o número de clientes só aumenta a cada dia e, inspecionar a todos seria uma tarefa árdua para a empresa. Diante do problema exposto, o atual trabalho propõe-se, em parceria com a Enel Distribuição Ceará, a aplicação de estratégias baseadas na Ciência de Dados e técnicas de Data Mining (DM) com o objetivo de prever clientes com algum tipo de irregularidade em sua unidade consumidora, seja fraude e/ou defeito no equipamento de medição. Foram utilizadas diversas técnicas de Machine Learning, desde as mais tradicionais, Árvore de Decisão e as Redes Neurais Artificiais, até as mais sofisticadas como Random Forest e Gradient Boosting. Os resultados alcançados validam a abordagem em Ciência de Dados adotada, contribuindo consideravelmente para previsão de clientes com algum tipo de problema, seja ele causado pela fraude ou por problemas intrínsecos da medição.

1. Introdução

A Enel Brasil integrante do Grupo Italiano Enel SPA, é uma das maiores empresas privadas do setor elétrico brasileiro e desempenha papel de liderança no desenvolvimento das fontes renováveis de energia no país. Atua em toda a cadeia

energética, com atividades nas áreas de geração, distribuição, conversão, transmissão e comercialização, além de soluções em energia. Por intermédio de três distribuidoras, nos estados do Rio de Janeiro, Goiás e Ceará, leva energia com qualidade a cerca de 10 milhões de clientes residenciais, comerciais, industriais, rurais e do setor público. A Enel Distribuição Ceará, é uma empresa no ramo de distribuição e geração de energia elétrica com atuação em todo o estado do Ceará com sede na capital Fortaleza. Sua área de concessão abrange atualmente os 184 municípios do estado do Ceará, que possuem uma população estimada em mais de 9 milhões de habitantes e com 149 km² em extensão territorial. Detém prémios de peso no setor elétrico como, Melhor Distribuidora de energia elétrica do Brasil, Melhor Distribuidora de energia elétrica do Nordeste, Prêmio Nacional de Qualidade – PNQ, Responsabilidade Social, Empresa-modelo em Sustentabilidade dentre outros. Em 2017 registrava 3,9 milhões de clientes segmentados nas classes Residencial (2,9 milhões), Rurais (742 mil), Comercial (225 mil), Institucionais (48 mil) e os Industriais (5 mil). Grande parcela dos seus clientes são residenciais (74%) seguidos dos Rurais (19%) e os Comerciais (6%).

O setor de distribuição de energia elétrica no Brasil anualmente registra elevadas taxas de perdas decorrentes de vários fatores técnicos e/ou de ordem comercial. Tais perdas, também conhecidas como Perdas Globais, remetem à energia elétrica que foi injetada no Sistema Interligado e nas redes das distribuidoras, mas que não chega a ser comercializada. Conforme estimativas da Associação Brasileira de Distribuidoras de Energia Elétrica (ABRADEE), o montante desta energia perdida chegou a 13,9% no ano de 2016, percentual um pouco maior quando comparado com o ano de 2015, que se chegou a registrar 13,6% (ABRADEE, 2016).

As perdas globais de energia podem ser categorizadas quanto à origem, em Perdas Técnicas e Perdas Não Técnicasou Perdas Comerciais, como são mais comumente conhecidas (LIMA, 2005; PENIN, 2008; QUEIROGA, 2005; VIEIRALVES, 2005). Tais perdas, técnicas e comerciais, correspondem respectivamente a 7,9% e 6,1% do total de energia perdida em 2016 mencionado anteriormente (ABRADEE, 2016). Conforme delineado a seguir, a Figura 1 mostra que ao longo dos anos de 2000 a 2015 ocorreu um crescimento de 30,48% no índice que mede os níveis de perdas comerciais no país.

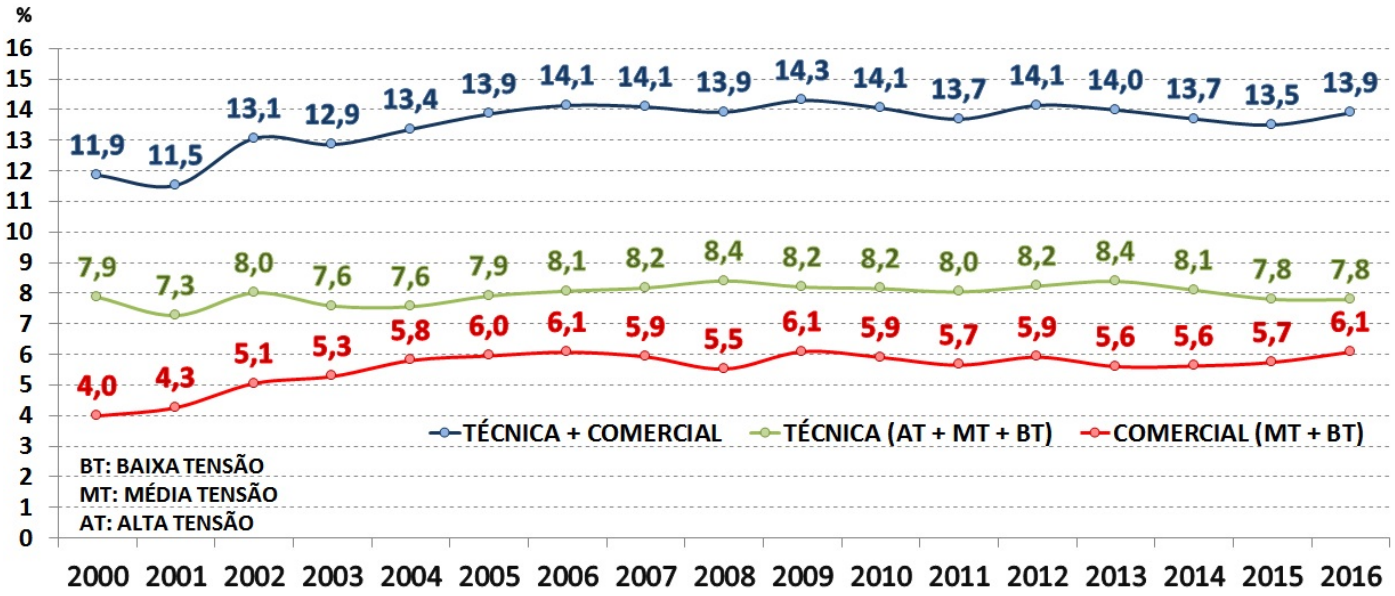

Figura 1 - Percentual de Perdas em Relação à Energia Injetada no SGD.

As diversas formas como são realizadas as fraudes e os furtos estão se tornando cada vez mais sofisticadas, a geografia extensa do estado acabam trazendo consigo dificuldades para as companhias de energia elétrica na detecção

desse tipo de ato. A principal ferramenta das concessionárias no combate as perdas comerciais ainda são as inspeções de campo realizadas por equipes treinadas, com o objetivo de detectar fraudes, furtos e outras irregularidades, como equipamentos adulterados ou defeituosos. Em meio ao universo total de seus clientes no estado do Ceará, que chega na casa de milhões, fica difícil separar aqueles que estão com consumos normais daqueles que estão causando perdas comerciais, sendo inviável para a Distribuidora inspecionar sua carteira completa de clientes sem um direcionamento inteligente de busca. Entretanto, diante da pluralidade e complexidade que cerca o contexto, a ciência de dados contribui para a mudança do cenário apresentado.

2. Desenvolvimento

2.1 Motivações

Os atuais Sistemas de Informação (SI) conseguem coletar e armazenar dados a uma taxa de velocidade quase inimaginável. Tais dados são provenientes de fontes heterogêneas, a saber: transações bancárias, cartões de crédito, medicina moderna, telecomunicações, fotografias compartilhadas em um determinado site da web, pesquisas espaciais e o crescente volume de informações disponível na internet (FAYYAD & PIATETSKY-SHAPIRO & SMYTH, 1996a); (BRAMER, 2013); (TAN & STEINBACH & KUMAR, 2009).

Junto com essa grande quantidade de dados e os avanços das tecnologias de banco de dados, as quais facilitaram o armazenamento e o gerenciamento de todo esse universo digital, com um custo relativamente baixo, surge a percepção de que nesses dados, estão escondidas “pepitas” de conhecimento que podem ser fundamentais para o avanço e competitividade em muitas áreas de negócio. No entanto, boa parte desses dados continua apenas armazenada, tornando o mundo mais rico em dados, porém pobre em conhecimento (BRAMER, 2013).

Conseguir gerenciar e, mais que isso, analisar e entender todo esse universo de dados estruturados e não estruturados, se tornou uma tarefa árdua, quase impossível, considerando apenas a capacidade analítica humana. Diante deste cenário, se fez necessário o surgimento de novas ferramentas e técnicas computacionais para subsidiar a extração de conhecimento útil e novo a partir desse conglomerado de dados. O campo da Ciência de Dados envolve muitas etapas, que vão desde a manipulação e recuperação dos dados, fundamentos matemáticos e estatísticos, pesquisa e raciocínio (FAYYAD & PIATETSKY-SHAPIRO & SMYTH, 1996a); (CHEN & HAN & YU, 1996).

2.2 Descoberta de Conhecimento em Banco de Dados (KDD)

Historicamente, o processo de descoberta de padrões úteis e novos em meio aos dados tem sido tratado com pluralidade de nomes por diversos pesquisadores, dentre eles, extração de conhecimento, descoberta de informação, arqueologia de dados, colheita de informação, processamento de padrões em dados e a própria Mineração de Dados (FAYYAD & PIATETSKY-SHAPIRO & SMYTH, 1996a); (CHEN & HAN & YU, 1996). O termo *Knowledge Discovery in Databases* (KDD) teve sua origem no final da década dos anos 1980 em Detroit, durante o Workshop de KDD, onde foram discutidas questões importantes sobre a ideia de descoberta de conhecimento em banco de dados enfatizando que o conhecimento é o produto final dessa descoberta, tendo se popularizado no campo de Inteligência Artificial (IA) e Aprendizado de Máquina (PIATETSKY-SHAPIRO, 1991).

De acordo com Fayyad, Shapiro e Smyth, o processo global do KDD (Ver Figura 2) inclui a avaliação e interpretaçãodos padrões encontrados na etapa de mineração para indicar quais os padrões podem ser considerados como conhecimento novo (FAYYAD & PIATETSKY-SHAPIRO & SMYTH, 1996b). Tal processo é considerado iterativo e interativo, pois envolve alguns passos que requerem decisões do usuário do processo (MAIMOM; ROKACH, 2010). As etapas básicas do processo KDD são delineadas a seguir:

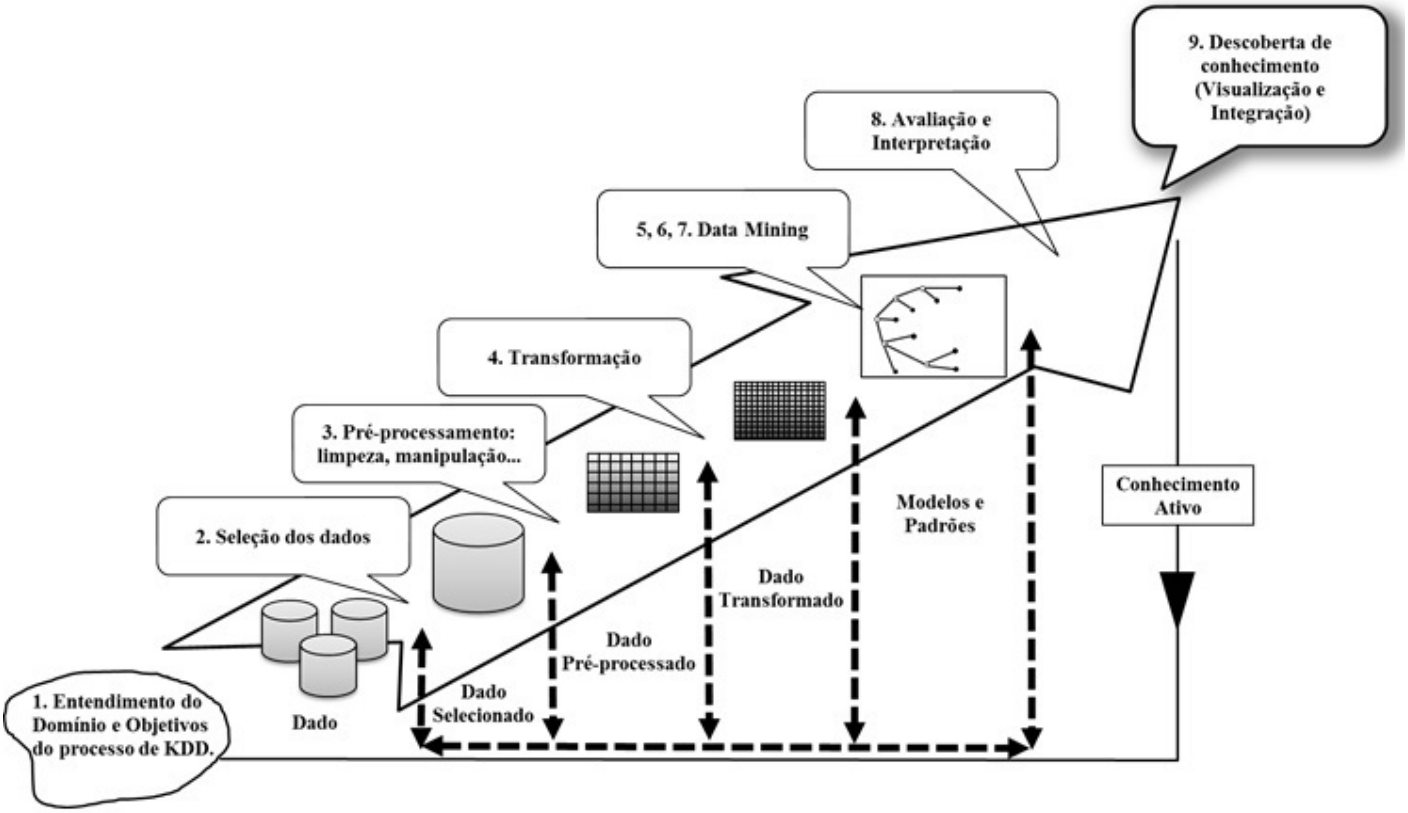

Figura 2 - Processo global do KDD.

I. **Entendimento do domínio:** Compreender o domínio dos dados onde será aplicado o processo é um pré-requisito essencial para se conseguir extrair algum conhecimento útil a partir de grandes volumes de dados. Neste momento se prepara o ambiente para entender o que deve ser feito com as muitas decisões.

II. **Seleção de dados:** Uma vez definidos os objetivos, os dados que serão utilizados para a descoberta de conhecimento devem ser estabelecidos. Isto inclui descobrir que dados estão disponíveis, a obtenção de dados adicionais necessários e depois integrar todos os atributos que serão considerados para o processo. Esta será a base de conhecimento (evidências) para a construção dos modelos.

III. **Pré-processamento dos dados:** Nesta etapa a confiabilidade dos dados é reforçada. Ela inclui a limpeza dos dados, como também a manipulação de dados em falta, remoção de ruídos e ou de outliers.

IV. **Transformação e redução dos dados:** Nesta etapa, a geração de melhores dados para a mineração de dados é preparado e desenvolvido. Métodos que incluem a redução das dimensões, por exemplo, a seleção de recurso, extração e amostragem de registros, e transformação de atributos, como a discretização de atributos numéricos e transformação funcional.

V. **Escolha da tarefa de mineração dos dados apropriada:** Nesta etapa o minerador está pronto para decidir que tipo de mineração de dados irá utilizar, por exemplo, classificação, regressão ou clustering. Isso

está diretamente relacionado com as metas do KDD, como também com as etapas precedentes. Existem dois objetivos principais na Mineração de Dados: predição e descrição. A Previsão comumente citada como um método de Aprendizagem Supervisionado, enquanto o método Descritivo inclui aspectos de Aprendizagem Não Supervisionada e de Visualização.

VI. **Escolha do(s) algoritmo(s) de mineração dos dados:** Estando com a estratégia definida, agora é o momento para decidir que técnicas serão utilizadas. É nesta fase que especificamos o(s) algoritmo(s) utilizado(s) na pesquisa por padrões. Por exemplo, considerando precisão versus clareza dos modelos, o primeiro é melhor utilizando Redes Neurais, ao passo que o segundo é melhor com Árvore de Decisão.

VII. **Mineração dos dados:** Esta etapa gera os padrões de uma forma de representação especial, tais como regras de classificação, árvore de decisão, modelos de regressão, dependências, etc. Nesta etapa poderá ser necessário aplicar o(s) algoritmo(s) várias vezes até que um resultado satisfatório seja obtido. Esta etapa será coberta com mais detalhes no próximo capítulo.

VIII. **Interpretação dos novos padrões:** Nesta etapa é feita a avaliação e interpretação dos padrões minerados (regras, confiabilidades, etc.), em referência aos objetivos definidos na primeira etapa. O objetivo desta etapa é a compreensão e utilidade do modelo induzido, visualizando os padrões extraídos, removendo os padrões redundantes ou irrelevantes, e traduzindo aqueles que são úteis e compreensíveis ao usuário. Aqui o conhecimento novo que foi descoberto está documentado para posterior utilização. Dependendo da necessidade do minerador, é possível retornar para qualquer uma das etapas passadas, com o intuito de melhorar o modelo.

IX. **Consolidação e utilização dos novos padrões:** Neste momento o minerador está apto para incorporar o novo conhecimento em algum sistema para ações futuras. O conhecimento torna-se ativo no sentido de que podemos fazer alterações no sistema e medir seus efeitos. Na verdade, o sucesso deste passo determina a eficácia do processo global do KDD.

2.3 Data Mining

Data Mining tem se tornado a “menina dos olhos” de muitos pesquisadores, cientistas e profissionais de diversos ramos da indústria, em virtude da alta disponibilidade da grande massa de dados do nosso atual Universo Digital e da necessidade iminente de transformar esses dados em informações e conhecimento útil, podendo ser usados em aplicações que vão desde a análise de mercado, detecção de fraudes, retenção de clientes, controle de produção, exploração científica, etc. (CHEN et al, 1996); (HAN, KAMBER, 2006).

Nossa capacidade humana de raciocínio, comprovadamente, consegue realizar até oito comparações em paralelo. O papel do Data Mining é precisamente ampliar essas comparações para “infinitas” tornando isso perceptível ao olho humano. Ainda existe muito conhecimento desconhecido na grande massa de dados disponíveis nas bases corporativas. E com as técnicas que o Data Mining oferece, é possível transformar esses dados em conhecimento valioso, que agregará valor para a organização. Diferentemente das técnicas estatísticas que verificam padrões hipotéticos, o Data Mining manuseia os próprios dados para descobrir tais padrões, mas de forma alguma substitui as técnicas tradicionais estatísticas, pelo contrário, o mesmo acabou se tornando uma extensão desses métodos, que em parte são o resultado de uma modificação maior na estatística. Por ser considerada interdisciplinar, Data Mining tem variações com o campo de atuação de alguns autores. As suas três áreas que são consideradas como tendo maior expressividade são: a Estatística, Aprendizado de Máquina e Banco de Dados. (LEMOS, 2003); (THEARLING, 2000).

2.4 Abordagens em Ciência de dados na Enel Distribuição Ceará

Atualmente a empresa conta com uma área especializada em inteligência computacional no combate as perdas de energia. A empresa investiu em tecnologia de ponta, contando agora com uma ferramenta de nome do mercado, o SAS. O sistema está baseado em quatro áreas de conhecimento principais (distintas, porém fortemente correlacionadas): Banco de Dados (TI), KDD (knowledge discovery), Data Mining e Estatística aplicada. A seleção de alvos para inspeção opera sobre uma arquitetura de banco de dados.

Inicialmente, os dados são extraídos do *Synergia* (sistema legado, proprietário) e carregados em banco de dados Oracle. Através do Sistema SAS a equipe acessa o Data Mart de Perdas, composto pelas tabelas Oracle (Fatos/Dimensões). No Sistema SAS, são implementadas as regras dos especialistas do negócio e as análises são direcionadas para a melhoria da rentabilidade e da energia recuperada. Investigam-se as bases de modo a determinar os perfis de clientes com maior propensão à fraude, o que envolve desde a construção de tabelas e consultas (SAS) quanto às atividades de mineração propriamente ditas.

Para isso, o Sistema SAS disponibiliza duas ferramentas que se integram e se completam: SAS Guide e SAS Miner. Com o SAS Guide, são realizadas: a) análise de consistência e validação dos dados; b) preparação e organização dos dados em tabelas que sejam estatisticamente adequadas para a posterior etapa de mineração e execução dos modelos estatísticos; c) análises estatísticas básicas e consultas diversas; d) importação dos Flat Files (tais como índices pluviométricos e renda per capita, IBGE); e etc. Através do SAS Miner, são executados três modelos que permitem aumentar a disponibilidade de clientes para uma campanha de inspeção SAS.

Ao final do processo, o SAS Miner gera a lista dos clientes eleitos para a inspeção. Diversas regras são aplicadas (baseadas nos especialistas das áreas), regras que formam variáveis de entrada e outras que refinam a seleção para campo potencializando a indicação por probabilidade (SAS Miner). Variáveis de entrada são formadas considerando reincidência de sucesso, ganho por cliente negativo e etc. Das regras que refinam a indicação para campo, podemos citar: contatos do cliente com a distribuidora, troca de titularidade, altas quedas percentuais e de energia no histórico de consumo do cliente, dentre outras. A partir de um fluxo de dados bem definido a área responsável pela geração de clientes candidatos a inspeção, executa semanalmente seus processos de acordo com o fluxograma mostrado na Figura 3.

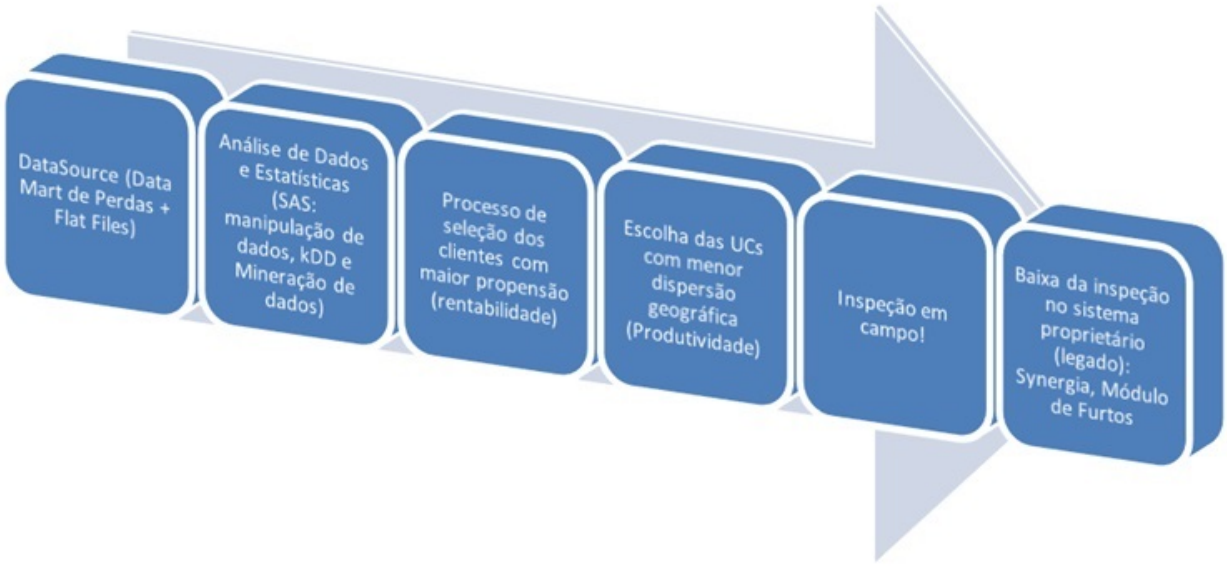

Figura 3 - Fluxo geral que consolida o conhecimento aplicado para a geração das inspeções: uma visão sistêmica.

Dentro do Fluxo Geral descrito anteriormente, o processo de análise de dados se dá pela execução do seguinte subprocesso mostrado na Figura 4.

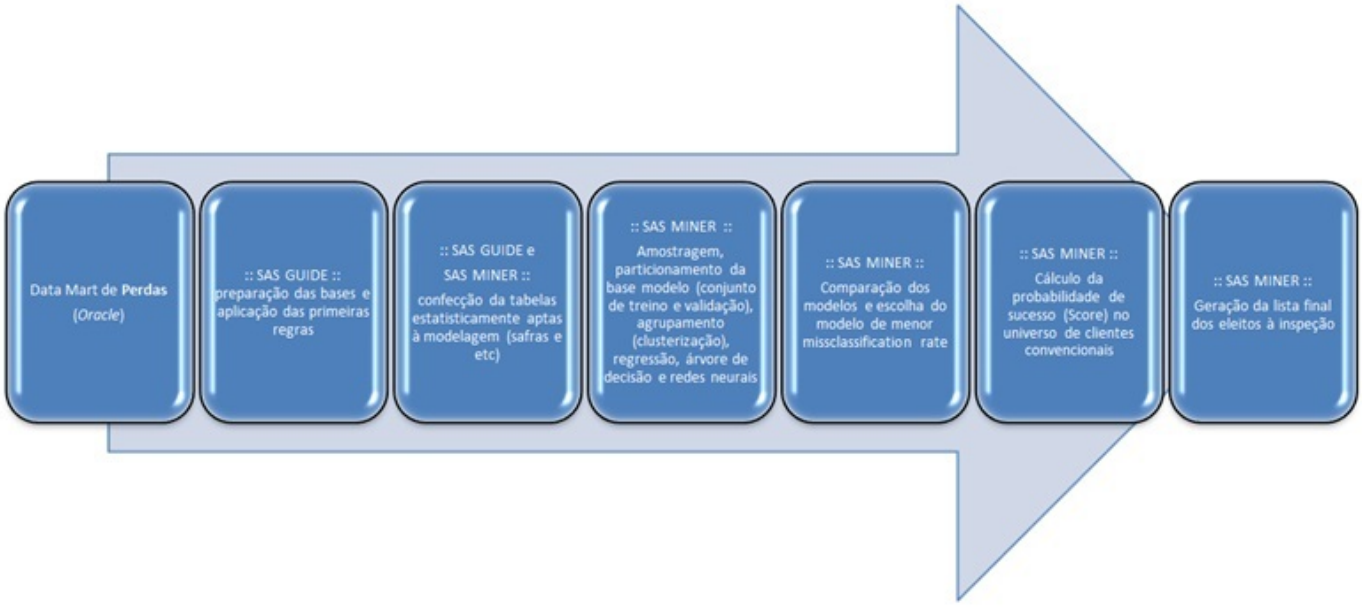

Figura 4 - Fluxo geral do sistema SAS GUIDE e SAS MINER

Desde que a abordagem em Data Mining começou a ser implementada na Enel Distribuição Ceará, em meados de 2009, conforme podemos observar na Figura 5, mesmo com uma redução de quase 50% dos serviços de inspeção de campo, a performance dos primeiros modelos utilizando as técnicas de Arvore de Decisão, Redes Neurais e Regressão já mostravam resultados promissores que confirmavam a eficácia das técnicas até então nunca utilizada pelo time de *Analytics*. Contudo, na mesma proporção em que se avançava na aplicação de abordagens inteligentes no combate ao furto de energia elétrica, a agressividade do mercado e a indústria da fraude também se sobressaíam levando a necessidade de inovar com técnicas em *Machine Learning.* Em razão ao aumento do furto de energia utilizando as mais diversas fraudes engenhosas, vislumbrou-se a necessidade de avançar ainda mais na implementação de novas técnicas de *Machine Learning*, tais como *Random Forest* e *Gradiente Boosting*.

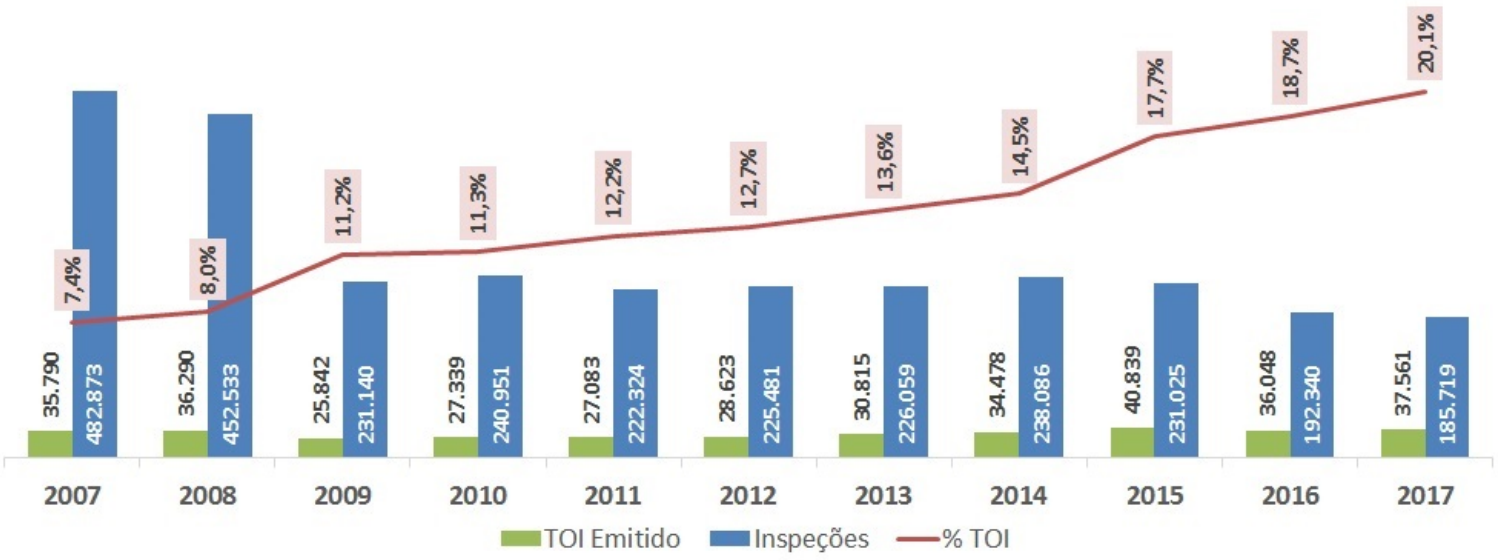

Figura 5 – Histórico de Operações de Inspeções Enel Distribuição Ceará

Random Forest é uma das novas técnica em *Machine Learning* utilizada principalmente na resolução de problemas de classificação. Utilizando uma abordagem diferenciada de um conjunto Árvores de Decisão, seu diferencial não está apenas no resultado de um modelo especificamente, mas da média dos resultados de várias Árvores de Decisão combinadas, reduzindo assim consideravelmente o problema do *overfitting* em seu conjunto de treinamento e teste (HO, 1995, p. 278–282). Assim como *Random Forest ,* a técnica de *Gradient Boosting* é amplamente utilizada para problemas de regressão e classificação, que produz um modelo de previsão na forma de um conjunto de modelos de previsão fracos, geralmente árvores de decisão. Esta técnica, produz o modelo em um modo de *stage*, como outros métodos de otimização, e os generaliza ao permitir a otimização de uma função de perda arbitrariamente diferenciável. (HASTIE & TIBSHIRANI & FRIEDMAN, 2008)

Após a implementação das novas técnicas citadas, dentro do processo de seleção de clientes, nos anos que seseguiam de 2015 até 2017 os resultados alcançados apresentaram uma melhora considerável em termos de acurácia dos modelos (*target*). De acordo com os dados apresentados na Figura 5, a quantidade de inspeções com TOI emitido no período, superaram em cerca de 17% do resultado alcançado nos anos de 2007 a 2009, onde foram inspecionados cerca de 1,7 milhões de clientes. Os resultados também atestam que a energia recuperada com os TOIs faturados superou em mais de 250%, cerca de 54 GWh em comparação ao período que mais tivemos inspeções executadas, em 2007-2009.

3. Conclusões

Diante do atual cenário no estado do Ceará, na qual a Enel Distribuição é detentora da concessão total, o presente trabalho traz mais uma vez a problemática recorrente que são as Perdas Comerciais para as distribuidoras de energia elétrica. Com isso, é exposto a importância do combate às fraudes e furtos de energia, através de técnicas computacionais inteligentes, viabilizando de forma otimizada, o combate a tais práticas ilícitas.

Técnicas bem conhecidas e utilizadas pela comunidade de Ciência de Dados foram expostas para trazer solução ao problema de combate as Perdas de energia elétrica, a saber, os classificadores por árvore de decisão e os classificadores por redes neurais. Conforme a agressividade do mercado e indústria da fraude avançavam em novas

formas de cometer o furto de energia, a equipe de *Analytics* implementou novas técnicas de *Machine Learning*, tais como *Random Forest* e *Gradiente Boosting*.

Os resultados alcançados pelos novos modelos gerados alcançaram resultados satisfatórios superando em mais de 250% a quantidade de TOIs emitidos com uma recuperação de energia de 87,3 GWh faturados. Desta forma, a abordagem de Ciência de Dados contribui consideravelmente para previsão de clientes com algum tipo de problema, seja ele causado pela fraude ou por problemas intrínsecos da medição.

4. Referências bibliográficas

ABRADEE, Associação Brasileira de Distribuidores de Energia Elétrica. Sistema de Informação para Gestão. Acesso em 30/01/2018, disponível em http://www.abradee.com.br/setor-de-distribuicao/perdas/furto-e-fraude-de-energia.

CHEN, Ming-Syan; HAN, Jiawei; YU, Philip S., Data Mining: an overview from database perspective. IEEE Transaction on Knowledge and Data Engineering, New York, Vol. 8, Nº. 6, 1996.

ENEL BRASIL. Conheça. Sobre a Enel. Acesso em 29/01/2018, disponível em <https://www.eneldistribuicao.com.br/ce/AEnelNoBrasil.aspx>.

ED CEARÁ. Sobre a Enel Distribuição Ceará. Acesso em 29/01/2018, disponível em < https://pt.wikipedia.org/wiki/Enel_Distribui%C3%A7%C3%A3o_Cear%C3%A1>.

FAYYAD, Usama; PIATETSKY-SHAPIRO, G.; SMYTH, Padhraic. 1996a. From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, 1996/ Vol. 17, Nº. 3.

FAYYAD, Usama; PIATETSKY-SHAPIRO, G.; SMYTH, Padhraic. 1996b. The KDD Process for Extracting Useful Knowledge from Volumes of Data, ACM, Nov. 1996/Vol. 39, Nº. 11.

HAN, Jiawei; KAMBER, Micheline.Data Mining: Concepts and Techniques. 2. ed. San Francisco. Morgan Kaufmann Publishers, 2006. 772 p.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome (2008). The Elements of Statistical Learning (2nd ed.). Springer. ISBN 0-387-95284-5.

HO, Tin Kam (1995). Random Decision Forests. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, 14–16 August 1995. pp. 278–282.

LIMA, Davi Antunes. Convergência tarifária: remédio regulatório para o livre acesso. Brasília: ANEEL, 2005. 16 p. (Textos para discussão, n. 2).

MAIMOM, Oded; ROKACH, Lior. Data Mining and Knowledge Discovery Handbook. 2. ed. New York, USA. 2010. 1285 p.

PENIN, Carlos Alexandre de Sousa. Combate, Prevenção e Otimização das Perdas Comerciais de Energia Elétrica. 2008, 227f.. Dissertação (Doutorado em Engenharia Elétrica), Escola Politécnica da Universidade de São Paulo, São Paulo.

PIATETSKY-SHAPIRO, G. 1991. Knowledge Discovery in Real Databases, AI Magazine, Winter 1991.

QUEIROGA, R. M. Uso de técnicas de data mining para detecção de fraudes em energia elétrica. 2005. 147f..

Dissertação (Mestrado em Informática) –Universidade Federal do Espírito Santo, Vitória.

VIEIRALVES, Eduardo de Xerez. Proposta de uma Metodologia para Avaliação das Perdas Comerciais dos Sistemas Elétricos. O Caso Manaus. 2005. Dissertação (Mestrado Planejamento de Sistemas Energéticos) – Unicamp, São Paulo.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. Introdução a Mineração de Dados. Rio de Janeiro: Editora Ciência Moderna. 2009. 900 p.