IDENTIFICAÇÃO DE PERDAS NÃO TÉCNICAS DE ENERGIA UTILIZANDO TÉCNICA DE REGRESSÃO BASEADA EM BOOSTING

Amanda Vasconcellos Gueldini –amandagueldini@hotmail.com
Gabriel Antonio Mello Borges dos Santos – gabrielmello11@gmail.com
Cleber Roberto Guirelli (Orientador) – cguirelli@mackenzie.br

RESUMO

Ao identificar anomalias no consumo de energia, é possível detectar perdas comerciais através da utilização de algoritmos baseados em *boosting*. Utilizando a metodologia de árvores de decisão, via técnicas de regressão, em conjunto com um algoritmo denominado *Adaboost*, pode-se identificar possíveis anomalias no consumo. Ao recolher os dados dos consumidores, dividindo-os em grupos (*clusters*) de acordo com as curvas de consumo e aplicar o método estatístico RMSD nos dados permitiu classificá-los em normais e anormais. Isto possibilitou a inserção dos consumos no algoritmo para a realização da prova de sua competência relacionada à identificação de perdas não técnicas de energia com base nas informações disponibilizadas. Apesar do *Adaboost* apresentar-se um excelente classificador e estar entre os melhores, o resultado obtido ao comparado com outros algoritmos para a indicação das anomalias poderia ser aprimorado com a utilização de mais dados.
Palavras-chave: Classificador *Adaboost*. Perdas não técnicas.

IDENTIFICATION OF NON-TECHNICAL ENERGY LOSSES USING BOOSTING-BASED REGRESSION TECHNIQUES

ABSTRACT

By identifying anomalies in energy consumption, it is possible to detect commercial losses using algorithms based on boosting. Using the decision tree methodology, via regression techniques, combined with an algorithm called Adaboost, it is possible to identify possible anomalies in consumption. By collecting consumer data, dividing them into groups (clusters) according to consumption curves and applying the RMSD statistical method to the data, it was possible to classify them into normal and abnormal. This made it possible to insert the consumptions in the algorithm to perform the proof of its competence related to the identification of non-technical energy losses based on the information made available. Despite Adaboost being an excellent classifier and being among the best, the result obtained when compared to other algorithms for the indication of anomalies could be improved with the use of more data.

Keywords: Adaboost classifier. Non-technical losses.

1. INTRODUÇÃO

O fenômeno da perda de energia ocorre ao longo do sistema de transmissão e distribuição de energia elétrica. São as chamadas perdas técnicas que são as previstas pela concessionária, pois são efeitos das propriedades físicas dos componentes do sistema elétrico. Quanto maior for o comprimento da linha, maior será a perda.

O sistema é considerado de transmissão quando a tensão é superior a 230 kV, e de distribuição se a tensão é inferior a 69 kV. Porém, se a tensão permanece nesse intervalo entre 69 kV e 230 kV, o sistema é considerado de subtransmissão, ou seja, é a rede para casos particulares em que a distribuição não se conecta diretamente à transmissão, porém considerada parte da rede de distribuição.

Além das perdas técnicas, têm-se as perdas comerciais ou perdas não técnicas. São a diferença entre a geração, consumo e perdas técnicas. Podem ser consideradas perdas não técnicas aquelas relacionadas a problemas como faltas de medidores, furto de energia, erros de faturamento das unidades consumidoras, e qualquer tipo de perda que não seja considerada técnica.

No ano de 2018 o Brasil registrou 14% de perdas de energia, sendo 6,6% através das perdas comerciais (BRASÍLIA, 2019). A Região Norte do país está situada com o maior volume destas. Os níveis das perdas e como eles podem ser identificados dependem de uma série de fatores como a gestão das concessionárias de energia, dados socioeconômicos, e aspectos comportamentais de cada área analisada. De acordo com a Agência Nacional de Energia Elétrica (ANEEL) o aperfeiçoamento das metodologias para a identificação dessas anomalias evoluiu os dados do tema, permitindo melhor análise de identificação de divergências (ANEEL, 2019).

Os elevados números de consumidores fraudadores não prejudicam somente a concessionária, mas também a população, pois os impostos referentes a essas contas de energia dos inadimplentes não chegam aos serviços básicos; e os adimplentes, pois a tarifa de energia aumenta para que a distribuidora consiga de alguma maneira compensar uma parte da energia furtada. Dentre as classificações de perdas comerciais, a parte que inclui o roubo de energia devido a ligações clandestinas é perigosa e traz consequências não somente monetárias, mas também causa acidentes, pois existem casos de cidadãos sendo expostos a choques elétricos, curto circuitos e incêndios ao intervirem no sistema elétrico da rede.

Levando em conta a necessidade de detectar os diversos tipos de fraudes, o objetivo desse trabalho é recolher dados de consumidores e, a partir disso, utilizar a metodologia de árvores de decisão via técnicas de regressão em conjunto com o algoritmo baseado em *boosting*, denominado Adaboost,

para identificar possíveis anomalias no perfil de consumo. Após o término da simulação foi possível decidir sobre a eficácia do algoritmo utilizado.

2. REVISÃO DA LITERATURA

2.1. CLASSIFICAÇÃO DAS PERDAS DE ENERGIA

As perdas de energia elétrica geradas ocorrem nas linhas de transmissão e redes de distribuição. Perdas que ocorrem na distribuição podem ser calculadas como a subtração entre a energia elétrica fornecida pela distribuidora e a faturada pelos consumidores. Portanto, essas perdas podem ser técnicas ou não técnicas.

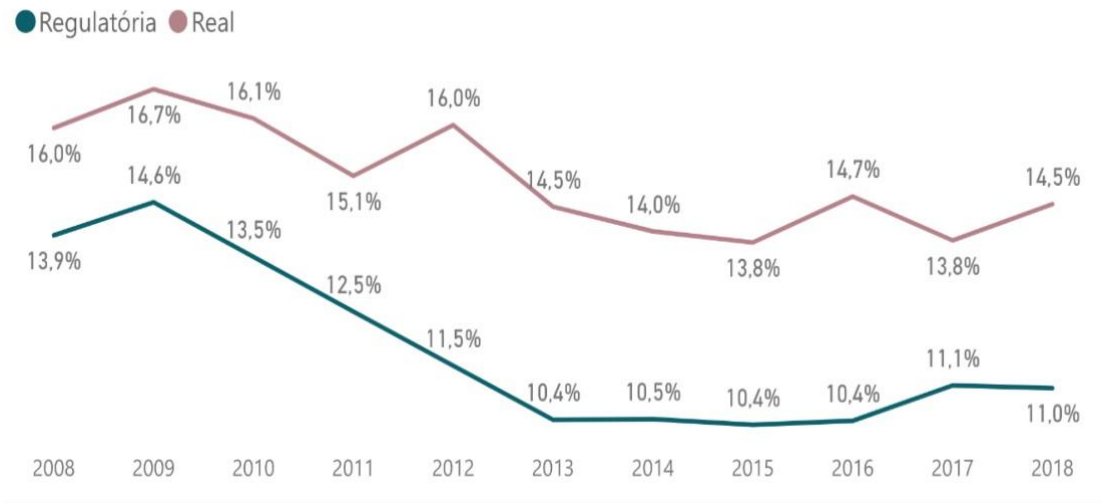
As perdas técnicas estão relacionadas as configurações e características da rede de distribuição das concessionárias. Podem ser relacionadas a distribuição da energia, pois as perdas ocorrem no processo de trânsito e transformação da tensão.

No entanto, as perdas não técnicas ou perdas comerciais são definidas como a diferença entre perdas totais e as perdas técnicas. São conhecidas como fraudes quando a origem se relacionar às ligações clandestinas. Mas também podem ocorrer por erros de leitura, faturamento e medição. A maioria ocorre em níveis de baixa tensão.

As perdas não técnicas podem ser divididas em reais e regulatórias. As regulatórias são reconhecidas na tarifa de energia, e são apuradas pela ANEEL. As reais são apuradas pela diferença entre as perdas totais, ou seja, os valores que realmente ocorrem (ANEEL, 2019).

O Gráfico 1 demonstra o comparativo de perdas não técnicas, em porcentagem sobre baixa tensão, do período de 2008 a 2018:

Gráfico 1- Comparativo Perdas Não Técnicas (% sobre BT)



Fonte: ANEEL (2019)

2.2. CONSUMIDORES FRAUDADORES

O processo de inspeção dos possíveis consumidores fraudadores é realizado após os técnicos da concessionária detectarem diferenças significativas no consumo de energia, ou quando a leitura do medidor não estiver de acordo com as normalidades. (MATOS, 2017)

A Energias de Portugal (EDP) registrou 3.794 fraudes de energia em residências, comércios e indústrias no primeiro semestre de 2019, somente nos arredores do Alto Tietê. A empresa realiza inspeções periódicas, e com tecnologia de última geração, chegaram a identificar uma fraude a cada três inspeções realizadas. Assim, foi possível recuperar 29.247 MWh devido a 13.004 inspeções.

Segundo as regras da ANEEL, o consumidor fraudador após ser flagrado, passa pelo processo de medição da energia furtada, e depois dos técnicos da empresa analisarem o valor total devido, é cobrado todo o período que não foi contabilizado anteriormente.

Assim, pode ser considerado em três categorias:

- Consumidor fraudador: a fraude foi confirmada no medidor ou o furto foi comprovado;
- Irregularidade técnica: medidor de energia com anomalias (problemas);
- Normal (sem fraudes ou anomalias).

Os consumidores que forem categorizados como fraudador ou irregularidade técnica são classificados pela concessionária como clientes irregulares, promovendo uma grande preocupação para as concessionárias de energia.

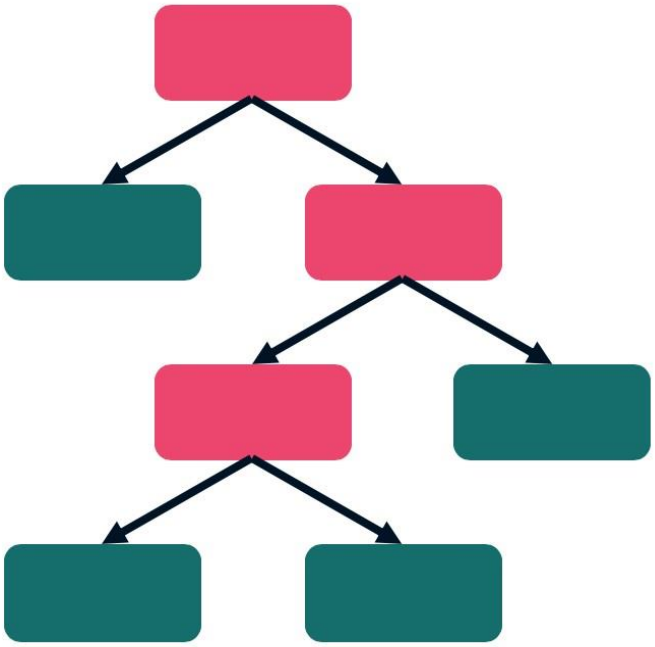
Ao longo do processo de aprimoramento das técnicas de detecção, diversas metodologias relacionadas a Inteligência Artificial foram criadas, melhoradas e modificadas anualmente com o intuito de auxiliar na diminuição das perdas, porém nenhuma se revelou muito eficaz. (MILES, 2019)

2.3. ÁRVORE DE DECISÃO

Árvore de decisão é um método que pode ser utilizado para a classificação de dados, através de diversos resultados de uma série de opções conjuntas, este método pode ser utilizado para criar modelos com aprendizado de máquina.

Geralmente a árvore de decisão se inicia a partir de um nó, que irá se dividir nos possíveis resultados, esses resultados geram outros nós, que vão se ramificando conforme surgem outras possibilidades, originando-se então uma árvore. O Diagrama 1 mostra o desenho de uma árvore de decisão com os nós e suas respectivas ramificações:

Diagrama 1 – Representação da árvore de decisão



Fonte: Do Autor

Há três modelos de nós (MILES, 2019):

- Probabilidade: demonstrado por um círculo, com as probabilidades dos resultados;
- Decisão: representado por um quadrado, demonstra a decisão que será tomada
- Término: representado por um triângulo demonstra o resultado de um caminho. Este método pode ser utilizado para ajustar modelos preditivos automatizados, que são aplicados em técnicas de *machine learning*, que leva em consideração ponderações de um item para verificar o valor dele.

No método de classificação, nós são dados e não decisões. As ramificações possuem diversos atributos e regras de classificação, que se associam a um rótulo de classe, geralmente nos extremos da ramificação. Por vezes, essas regras trabalham como um *if-else*, as decisões e valores de dados geram uma cláusula, onde caso as condições sejam cumpridas, resultam em algo concreto. Podendo também ser chamada de árvores de regressão.

2.4. BOOSTING

Atualmente existem algumas técnicas de *machine learning* para problemas de regressão e classificação, denominadas de *boosting*. Essa técnica produz um modelo de previsão, ou árvores de decisão. Os classificadores aprendem de forma espontanea, assim chegam a uma decisão mais forte. De acordo com Datalab Serasa Experian (2019), em sua definição matemática, *boosting* é uma forma de expansão, ajustando os dados em uma soma ponderada de funções elementares.

2.5. ADABOOST

Existem diversos algoritmos de classificação mas este trabalho aplica o algoritmo *Adaboost* para a classificação dos dados. O AdaBoost é o mais famoso algoritmo de *boosting* idealizado por Freund e Schapire (1997).

O AdaBoost utiliza diversos classificadores para elevar o acerto no processo de classificação, onde com um conjunto de dados de treinamento, o algoritmo de aprendizado de base produz um conjunto de classificadores base a fim de construir um classificador melhor, no qual seus resultados são combinados através do voto ponderado.

O funcionamento deste algoritmo é baseado na distribuição de pesos de acordo com o processo, na iteração é calculada a distribuição para normalizar os pesos distribuídos.

O modo mais comum de utilizar o Adaboost é em conjunto com as árvores de decisão. É possível projetar qualquer tarefa com essa estrutura (se um e-mail é spam ou não, se o tempo amanhã estará ensolarado ou não). Algumas árvores de decisão podem ser maiores do que outras, mas não há um tamanho pré-determinado para elas.

Existem três conceitos principais por trás do Adaboost:

- Nós: suas "árvores", diferentemente das árvores de decisão, são chamadas de Nós (ou *stumps*). Estes são categorizados por um nodo e duas folhas somente;
- Alguns nós terão maior poder de decisão do que outros;
- A ordem dos nós importa, pois o erro de uma influência na montagem do próximo. O primeiro passo é atribuir a cada linha um peso, o que indicará o quão importante é que a observação seja bem classificada. Cada nó utilizará apenas uma variável para fazer a sua verificação. No início esses pesos serão todos iguais e equivalentes a $1/N$, onde N é o número de linhas. Para melhor exemplificar o Adaboost, é suposta uma situação conforme a Tabela 1. Neste caso, o número de linhas é número de pacientes, e cada coluna (em verde) representa uma variável. No exemplo, baseando-se em cada variável, é possível detectar se o paciente apresenta problemas cardíacos.

Tabela 1 – Variáveis x Peso

Dores no Peito	Artérias obstruídas	Peso do Paciente (kg)	Problema no Cardíaco	Peso
Sim	Sim	92	Sim	1/8
Não	Sim	82	Sim	1/8
Sim	Não	95	Sim	1/8
Sim	Sim	76	Sim	1/8
Não	Sim	70	Não	1/8
Não	Sim	57	Não	1/8
Sim	Não	75	Não	1/8
Sim	Sim	78	Não	1/8

Fonte: Do autor

Após realizar as primeiras iterações com os nós os pesos serão alterados. Para construir o primeiro nó, os pesos iguais devem ser ignorados e as variáveis neste primeiro momento serão mais relevantes. Será verificado qual dessas variáveis melhor classificam o problema cardíaco.

- Dores no peito: Para as cinco amostras de pacientes que foram verificados com dores no peito, três foram constatados com problemas cardíacos, e dois pacientes foram classificados sem problemas. E das três amostras sem dores no peito, apenas um foi classificado como sem problemas cardíacos.
- Artérias obstruídas: Para as seis amostras de pacientes que foram verificados com as artérias obstruídas, três foram constatados com problemas cardíacos, e três pacientes foram classificados sem problemas. E a única amostra sem problemas na artéria foi classificada como sem problemas.
- Peso do paciente: Para as três amostras de pacientes que foram verificados com peso irregular (maior que 80 kg), três foram constatados com problemas cardíacos. E das cinco amostras sem peso irregular, apenas um foi classificado com problemas cardíacos.

Para calcular a variável que melhor classificou os problemas cardíacos será utilizada a taxa de Gini. Esta, conhecida como índice de Gini, tem a função de medir a probabilidade de uma variável específica ser classificada erroneamente ao ser escolhida aleatoriamente. O índice de Gini varia de 0 a 1, onde 1 indica que os dados são espalhados aleatoriamente entre diversas classes, e 0 indica que todos os dados são parte de uma única classe. Assim, um índice de Gini calculado como 0,5 indica dados igualmente espalhados em algumas classes. A fórmula (1) representa a equação de Gini, onde p_1 é a probabilidade de um dado ser classificado para uma classe escolhida:

$$Gini = 1 - \sum_{i=1}^{n} (p_1)^2 \quad (1)$$

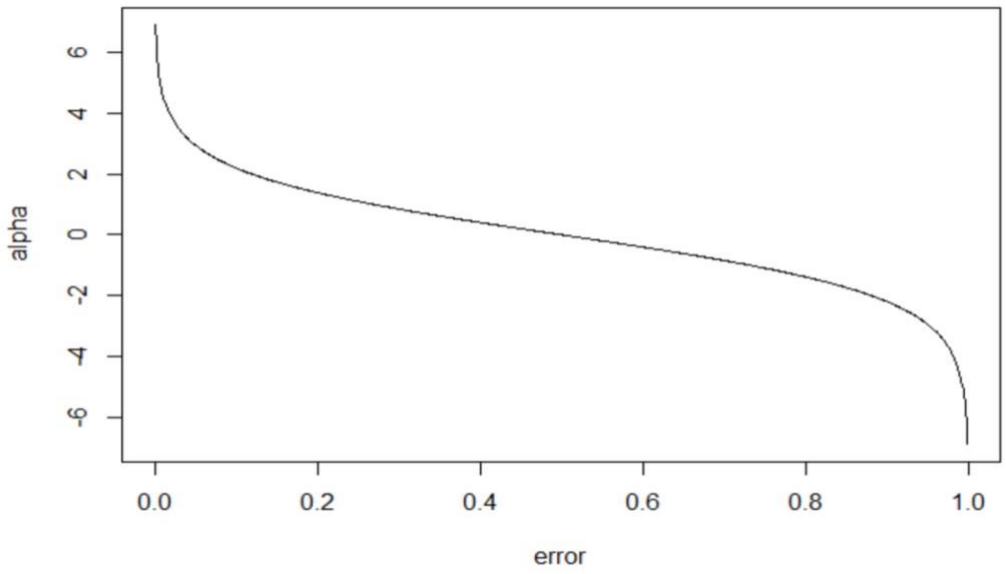
E assim:

- Dores no peito: 0,47;

- Artérias obstruídas: 0,5; ● Peso do paciente: 0.2.

A variável peso do paciente teve a menor taxa, portanto será definida como o nó principal neste momento. Portanto, os pesos das próximas iterações serão definidos de acordo com o erro, sendo que quanto menor o erro, menor o peso na próxima iteração, e assim sucessivamente. O Gráfico 2 apresenta os pesos dos erros por amostras. Quanto mais próximo de 1, pior e quanto mais próximo de zero, melhor.

Gráfico 2 - Peso do erro / Amostra



Fonte: ALTO, 2020

ln

É necessário calcular o quão significativo é o nó na classificação final, por isso utilizamos a fórmula (2) para encontrar os pontos (K) do Gráfico 2, ou seja, a quantidade de dizer que cada nó terá:

$$K = \frac{1}{2} \ln\left(\frac{1 - Erro\ Total}{Erro\ Total}\right) \qquad (2)$$

Onde o $Erro\ Total$ é dado pela soma dos pesos associados às classificações incorretas, conforme mostrado anteriormente. Então, o novo peso (np) é encontrado a partir da fórmula (3):

$$np = primeiro\ peso \times e^{K} \qquad (3)$$

Em seguida soma-se os novos pesos e cada peso é dividido pela somatória total para serem normalizados. Para o Adaboost classificar os nós e determinar o resultado ele utilizará a somatória dos nós, iterações e seus respectivos pesos serão fundamentais para a sua própria classificação.

3. METODOLOGIA

A metodologia do trabalho foi dividida em três etapas. A primeira foi a criação de uma base de dados a partir de uma pesquisa feita com diversos indivíduos, em que estes preencheram dados de suas contas de luz. São eles:

- consumo dos últimos 12 meses (set/2019 a set/2020);
- classificação da unidade consumidora;
- classe de tensão e tipo de medição;
- bairro/cidade;
- quantidade de pessoas que residem na residência;
- se após o começo da quarentena (março/2020) os indivíduos continuaram a residir na mesma residência;
- quantidade de indivíduos que começaram a trabalhar/estudar no modo home office;
- se o indivíduo já tinha conhecimento de algum episódio de furto de energia pelo bairro em que reside.

A segunda etapa refere-se à identificação de anomalias dos consumidores a partir do método de árvores de decisão baseado em regressão, na qual utilizou-se o algoritmo Adaboost, implementado em linguagem Python através do *Collab* da Google. Para isso, utilizamos a técnica de clusterização *k-means*, em que agrupamos os dados da quantidade de energia de cada conta de luz de acordo com curva de consumo de cada cliente. Normalizando os dados, colocando-os em um mesmo intervalo, para estarem em contexto com os outros, os inserimos como input para o algoritmo. Foi considerada a existência de quatro espécies com anomalia nessa base de dados, para assim ser possível a posterior comprovação da eficácia do algoritmo. Diversos testes foram feitos com o intuito de ajustar o código do algoritmo para termos um resultado mais preciso.

A terceira etapa refere-se à documentação dos testes e resultados obtidos, assim como a conclusão do trabalho e em conjunto com a revisão da metodologia proposta.

4. RESULTADOS E DISCUSSÃO

Para a escolha do algoritmo utilizado, considerou-se algoritmos baseados em árvores de decisão que possuem boa performance na classificação dos dados. Atualmente muitos trabalhos utilizam árvores de decisão para a classificação dos dados de consumo nas redes de energia elétrica, entretanto o Adaboost ainda é pouco utilizado.

A partir de uma base de dados recolhida através de uma pesquisa com moradores de São Paulo, com o auxílio da ferramenta *Google Forms*, filtrou-se os dados necessários para a conclusão da metodologia do trabalho. Estes são os dados de consumo de energia elétrica e a quantidade de indivíduos residentes na moradia, no período de um ano. No total foram recolhidas 42 amostras, e

após o tratamento dos dados removendo os valores inconsistentes e duplicados, obteve-se um total final de 32 amostras.

A Tabela 1 apresenta as primeiras 7 amostras da base de dados utilizada com os respectivos valores de cada mês, M1 até M13 (set/2019 - set/2020), em conjunto com a quantidade de indivíduos (NP).

Tabela 2 - Representação da base de dados

	ID	AN	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	NP
0	1	N	0.179	0.186	0.327	0.163	0.162	0.162	0.162	0.182	0.157	0.167	0.180	0.176	0.187	0.002
1	2	N	0.118	0.125	0.106	0.121	0.104	0.132	0.127	0.105	0.120	0.119	0.122	0.110	0.115	0.002
2	3	N	0.000	0.160	0.161	0.146	0.143	0.245	0.209	0.215	0.231	0.221	0.221	0.232	0.225	0.001
3	4	N	0.202	0.187	0.208	0.281	0.194	0.194	0.205	0.216	0.223	0.214	0.184	0.187	0.203	0.004
4	5	N	0.192	0.245	0.437	0.130	0.128	0.127	0.128	0.118	0.117	0.120	0.133	0.143	0.145	0.002
5	6	N	0.050	0.051	0.067	0.050	0.050	0.174	0.194	0.194	0.121	0.167	0.139	0.202	0.191	0.003
6	7	A	0.169	0.154	0.394	0.050	0.050	0.050	0.050	0.050	0.050	0.000	0.000	0.000	0.000	0.002

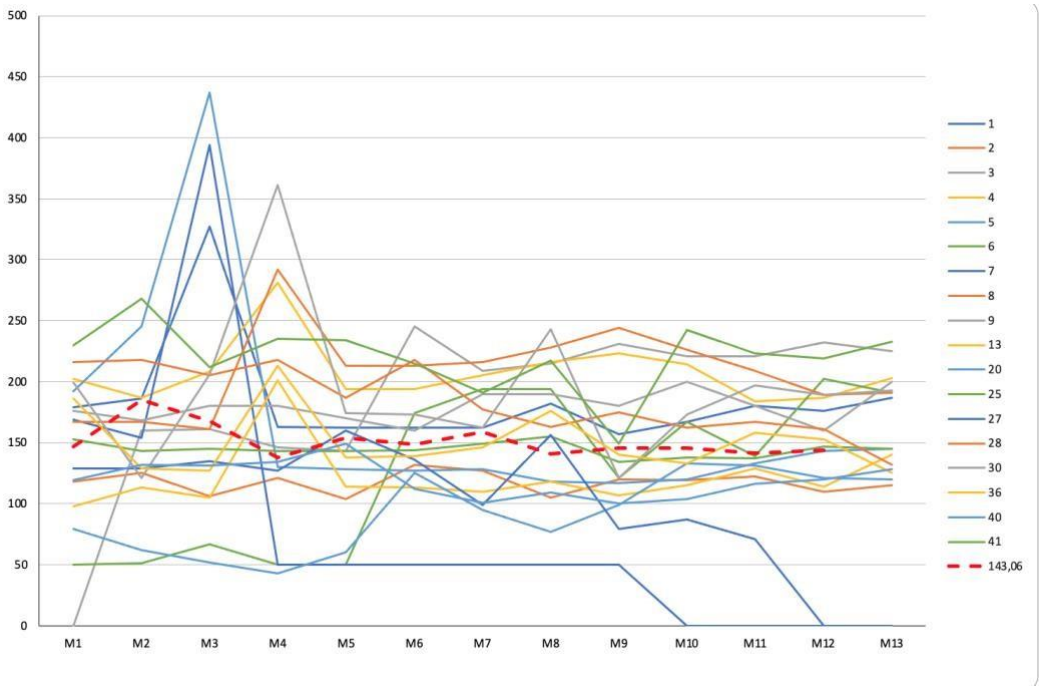
Fonte: Do autor

O método de clusterização foi utilizado para a divisão dos dados em dois grupos de acordo com as curvas de consumo, denominados Grupo 1 e Grupo 2. O método de clusterização (*k-means*) foi realizado utilizando o software Matlab. O primeiro grupo possui um total de 15 amostras, e é representado pelo Gráfico 3, enquanto o segundo grupo possui 17 amostras e é representado pelo Gráfico 4.

Gráfico 3 – Curvas de consumo do grupo 1

Fonte: Do Autor

Gráfico 4 – Curvas de consumo do grupo 2



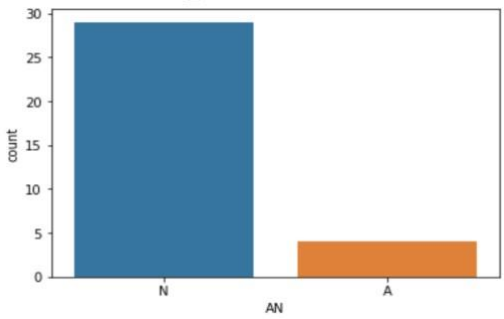
Fonte: Do Autor

Após o término desse processo aplicou-se o método estatístico RMSE (raiz quadrada do erro médio - *root-mean-square deviation*), encontrada a partir da fórmula (5) aplicado em todos os consumidores de ambos os grupos.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Predicted_i - Actual_i)^2}{N}} \qquad (5)$$

Desse modo foi possível determinar para cada grupo, dois valores de máximo e mínimo que foram considerados como anormais. Na Tabela 2 é possível observar na coluna “AN”, dois valores “A” (Anormal) ou “N” (Normal), resultados obtidos a partir da clusterização e do RMSE. Esses dados foram utilizados para auxiliar os testes e verificar a acuracidade do algoritmo em etapas posteriores. O Gráfico 5 demonstra a quantidade de amostras classificadas inicialmente como anormais e normais.

Gráfico 5 – Classificação das amostras



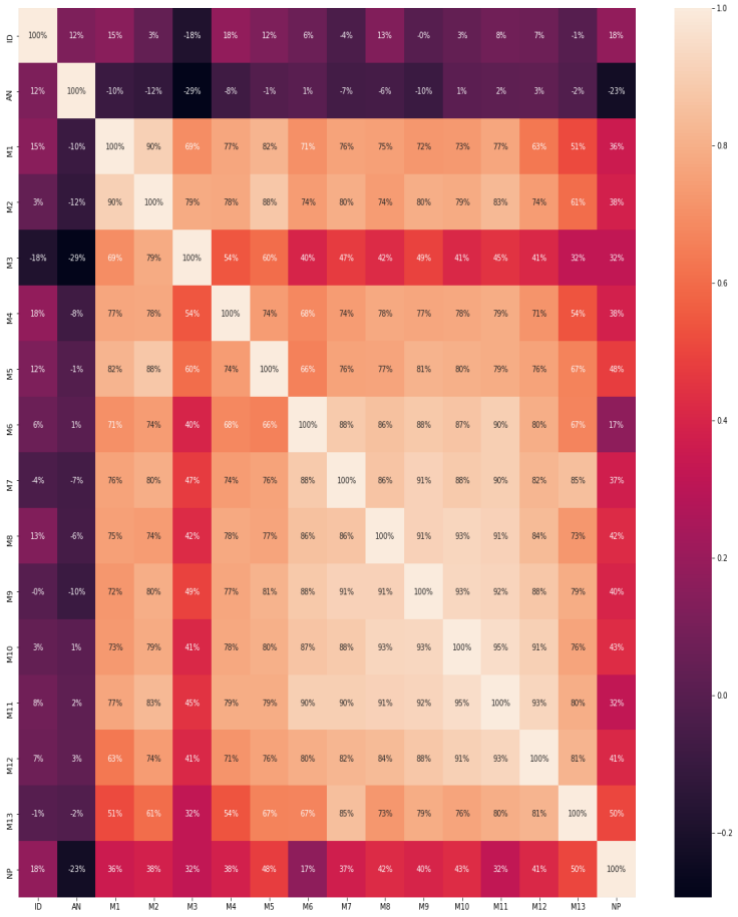
Fonte: Do Autor

O algoritmo tem uma melhor taxa de resultado ao trabalhar em uma mesma base (em um mesmo intervalo), sendo essa a razão para a mudança proposital dos dados para seguirem uma mesma escala. Como é possível notar também na Tabela 2, utilizou-se a escala de 0 a 1 para os dados de consumo de cada mês normalizados respeitarem o respectivo intervalo.

Utilizou-se a biblioteca Scikit-Learn, na qual possui o módulo "sklearn.ensemble". Neste está incluído o algoritmo baseado em boosting, denominado Adaboost (PEDREGOSA et al., 2011). Depois de tratados, classificados e normalizados os dados de consumo, inseriu-se os mesmos para realização dos testes com o algoritmo na plataforma Google Collab. Nesta plataforma além de utilizar o Adaboost, decidiu-se implementar outros 6 algoritmos (*Regressão Logística, K Nearest Neighbor, Support Vector Machine Linear Classifier, Support Vector Machine RBF Classifier, Gaussian Naive Bayes, Decision Tree Classifier e Random Forest Classifier*), que também atuam como classificadores, para efeitos de comparação de resultados.

Como todo classificador, este algoritmo correlaciona os dados que foram inseridos no programa, para que assim os treinos e testes sejam ainda mais precisos. Pode-se observar essa correlação a partir da Gráfico 6.

Gráfico 6 - Correlações



Fonte: Do Autor.

Os tons mais claros são correlações que o algoritmo entende como boas, e portanto possuem um peso maior na classificação dos dados. Os tons mais escuros têm um peso menor para essa classificação.

No código do algoritmo, alterou-se os percentuais de treino e teste para viabilizar o melhor resultado possível, utilizando-se 50% treino e 50% teste. Ou seja, das 32 amostras, 15 foram utilizadas para treino, e o restante para teste.

Após a finalização da simulação, obteve-se um resultado de 82,36 % de acuracidade. Para chegar a essa conclusão utilizamos a chamada Matriz de Confusão, representada pela Imagem 1. Essa matriz mostra o quanto de previsões foram corretas, ou seja, se os dados previstos como “positivo” realmente são “positivo” ou se na verdade são “falso”, e o mesmo procedimento ocorre com a previsão de “falso”. Em sua diagonal principal são demonstrados o volume de dados em que a previsão estava correta, e para determinar esse resultado foi necessário comparar o resultado previsto com o real, ou seja, a parte treinada com a prevista.

As colunas são as previsões da matriz, enquanto as linhas são os dados de fato correspondentes. Para a primeira linha considera-se apenas o “positivo”, assim como na primeira coluna, para a segunda linha é considerado o “falso”, como na segunda coluna.

Imagem 1 – Matriz de confusão

```
Model 7
[[ 0  2]
 [ 1 14]]
Testing Accuracy = "82.35294117647058"
```

Fonte: Do Autor

Para obter os resultados de acuracidade o algoritmo calcula a razão da soma dos dados da diagonal principal, pela soma das linhas (total), como exibido na fórmula (6) a seguir:

$$\frac{0 + 14}{2 + 15} \times 100\% = 82{,}35\% \quad (6)$$

Ao comparar os resultados com "R*andom Forest Classifier Training*", "Logistic Regression Training", "K Nearest Neighbor", "Support Vector Machine (Linear Classifier e RBF Classifier)" e "Gaussian Naive Bayes", estes obtiveram um resultado melhor com a mesma quantidade de amostras testadas. Nesse caso, a acuracidade foi de 88,23%, como ilustra a Tabela 3.

Tabela 3 – Acuracidade dos modelos

Modelo	Classificador	Acuracidade
Logistic Regression	94%	88%
K Nearest Neighbor	88%	88%
Support Vector Machine (Linear)	94%	88%
Support Vector Machine (RBF)	88%	88%
Gaussian Naïve Bayes	94%	88%
Decision Tree	100%	82%
Random Forest	94%	88%
Adaboost	100%	82%

Fonte: Do Autor

Para a exibição dos consumidores que possuem alguma anomalia nos consumos, no final do algoritmo Adaboost os resultados são exibidos ordenados primeiro pelos treinos e depois os testes realizados. Considerando "0" para consumidores anormais, e "1" para consumidores normais, como na Imagem 2.

Imagem 2 – Identificação de anormalidades

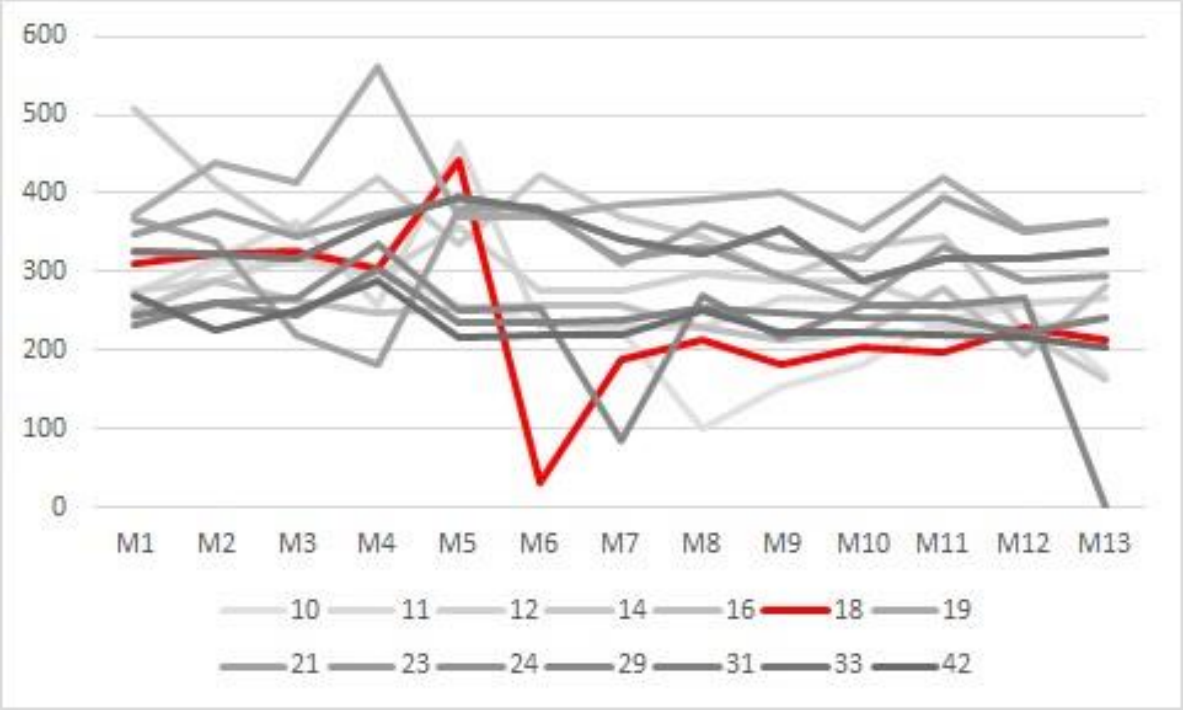
```
Train:
[1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1]

Test:
[0 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1]
```

Fonte: Do autor

Assim a identificação das anomalias fica ordenada de acordo com os dados inseridos no algoritmo, como a base de dados é pequena essa identificação fica mais ágil. Entretanto, para bases maiores é possível que essa exibição seja realizada de outra forma.

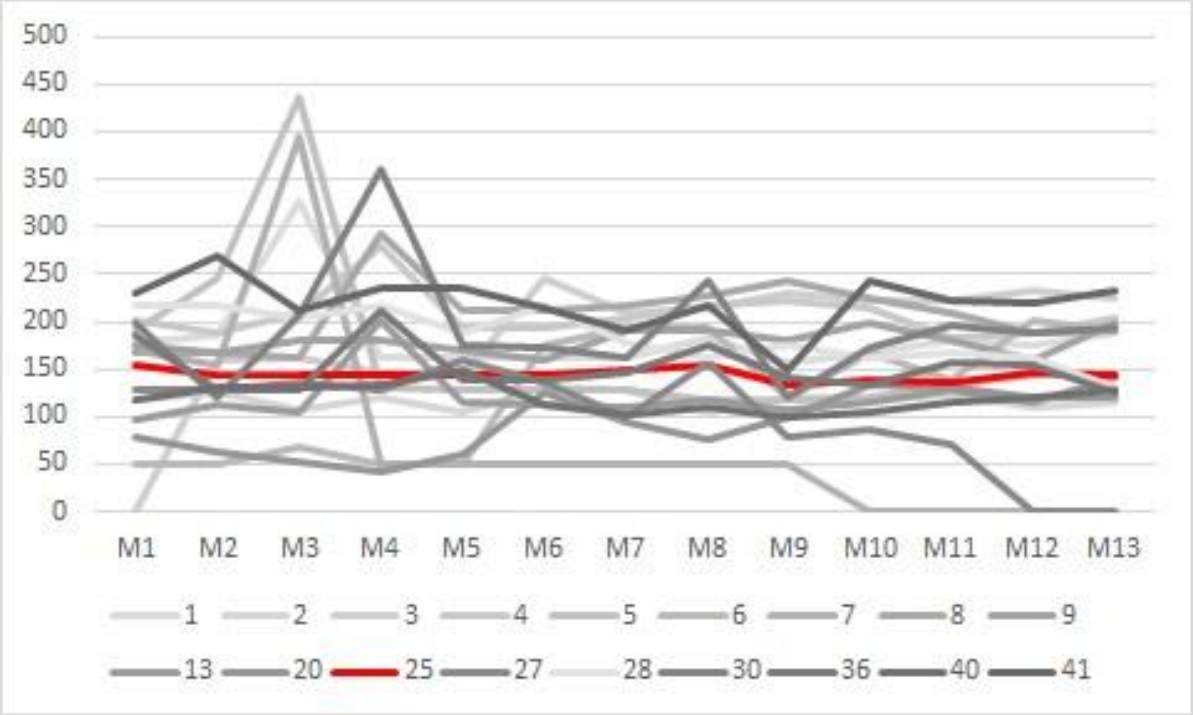
Os consumos identificados como anormais a partir do Adaboost são, respectivamente, os ID´s 18 e 25. A disposição das curvas de consumo, dos mesmos, é dada pelo Gráfico 7 e Gráfico 8. Nos quais as curvas destacadas são os ID`s mencionados.

Gráfico 7 – Curvas de consumo com identificação e anomalia- grupo 1



Fonte: Do autor

Gráfico 8 – Curvas de consumo com identificação e anomalia- grupo 2



Fonte: Do autor

5. CONSIDERAÇÕES FINAIS

A pesquisa demonstrou uma alternativa de identificação de anomalias e a forma com que ela pode ser endereçada. Dessa forma, procurando reduzir os prejuízos das concessionárias e dos consumidores afetados com maior assertividade. As concessionárias de energia têm o objetivo de identificar divergências no sistema elétrico, e com isso carrega muitos desafios. Para isso os estudos de novas tecnologias e investimentos nessa área tem crescido cada vez mais, com a intenção de identificar as perdas de forma efetiva e com baixo custo.

Inicialmente a proposta de obter uma base de dados a partir de uma liberação de dados de concessionárias de energia. Realizou-se reuniões com representantes de diversas concessionárias, porém não se obteve êxito nesse ramo. Assim, a solução de contorno baseou-se em criar essa base de dados a partir de uma pesquisa feita com residentes do Estado de São Paulo, resultando em uma quantidade de amostras que permitiram a continuação do trabalho.

Após a realização dos estudos do algoritmo, verificou-se que caso uma quantidade maior de amostras tivesse sido recolhida, os resultados seriam mais precisos. No entanto, mesmo com uma base de dados menos volumosa do que a desejada, obteve-se uma avaliação positiva do algoritmo Adaboost para a identificação de perdas não técnicas de energia. Conclui-se então que o Adaboost possui a capacidade de identificar as anomalias de consumo do circuito elétrico.

Apesar do algoritmo estudado ser um excelente classificador, ele não alcançou uma performance alta, como esperado.

Há atualmente muitas metodologias que permitem detectar e corrigir as fraudes que ocorrem no sistema. Porém, não é concreta a informação de qual método é o mais eficiente e nenhum deles previne e detecta em totalidade as ligações clandestinas e modificações dos medidores.

Para pesquisas futuras, indica-se a realização de testes com uma quantidade maior de amostras, para assim comparar novamente a eficácia do algoritmo Adaboost com outros algoritmos que tem a possibilidade de ser utilizados para a identificação de perdas não técnicas.

REFERÊNCIAS

ABRADEE (Vitória). **EDP alerta sobre segurança com a rede elétrica**. 2020. Disponível em: https://www.abradee.org.br/edp-registrou-mais-de-150-mil-clientes-sem-energia-por-conta-depipas-na-rede-eletrica-em-2019/. Acesso em: 08 jun. 2020.

ALTO, Valentina. **Understanding AdaBoost for Decision Tree**. 2020. Disponível em: https://towardsdatascience.com/understanding-adaboost-for-decision-tree-ff8f07d2851. Acesso em: 15 maio 2020

BRASÍLIA. Luís Carlos Carrazza. Agência Nacional de Energia Elétrica. **Perdas de Energia Elétrica na Distribuição**. Df: Agência Nacional de Energia Elétrica, 2019. 21 p. Disponível em: https://www.aneel.gov.br/documents/654800/18766993/Relat%C3%B3rio+Perdas+de+Energia_+Edi%C3%A7%C3%A3o+1-2019.pdf/b43e024e-5017-1921-0e66-024fa1bed575. Acesso em: 05 jun. 2020.

CURADO, Maria Isabel Coutinho. **Localização de Perdas Não Técnicas de Energia em Sistemas de Distribuição Utilizando o Método PQ**. 2015. 119 f. TCC (Graduação) - Curso de Engenharia Elétrica, Escola de Engenharia de São Carlos, Universidade de São Paulo, São Carlos, 2015.

DATALAB SERASA EXPERIAN. **Boosting (1): Árvores de Decisão e Gradient Boosting**. 2019. Disponível em: https://www.datalabserasaexperian.com.br/datalab/boosting-1-arvores-de-decisao-egradient-boosting/#:~:text=Em%20sua%20defini%C3%A7%C3%A3o%20matem%C3%A1tica%2C%20boosting,soma%20ponderada%20de%20fun%C3%A7%C3%B5es%20elementares%3A&text=onde%20%CE%B2m%20s%C3%A3o%20os,x%2C%20caracterizadas%20pelo%20par%C3%A2metro%20%CE%B3.. Acesso em: 15 maio 2020.

EDP. **EDP alerta sobre os riscos das ligações clandestinas.** 2018. Disponível em: https://www.edp.com.br/noticias/edp-alerta-sobre-os-riscos-das-ligacoesclandestinas#:~:text=Por%20esses%20motivos%2C%20a%20EDP,realizada%20com%20a%20m% C3%A1xima%20urg%C3%AAncia.. Acesso em: 08 jun. 2020.

FREUND, Yoav; SCHAPIRE, Robert e. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. **Journal of Computer and System Sciences**, [s.l.], v. 55, n. 1, p. 119-139, ago. 1997. Elsevier BV. http://dx.doi.org/10.1006/jcss.1997.1504. Disponível em: https://www.sciencedirect.com/science/article/pii/S002200009791504X?via%3Dihub. Acesso em: 20 jun. 2020.

LUCIDCHART. **O que é um diagrama de árvore de decisão?** 2020. Disponível em: https://www.lucidchart.com/pages/pt/o-que-e-arvore-de-decisao#section_0. Acesso em: 15 jun. 2020.

MATOS, Yasmin Christine Correa. **DETECÇÃO DE FRAUDES NO CONSUMO DE ENERGIA ELÉTRICA USANDO ÁRVORES DE DECISÃO**. 2017. 59 f. Dissertação (Mestrado) - Curso de Sistemas de Energia Elétrica, Engenharia Elétrica, Universidade Federal do Pará, Belém, 2017. Disponível em: http://ppgee.propesp.ufpa.br/ARQUIVOS/dissertacoes/DM%2028_2017%20Yasmin%20Christine%20Correa%20Matos.pdf. Acesso em: 8 jun. 2020.

MILES. **Leaky**: decision tree, random forests and adaboost. Decision Tree, Random Forests and AdaBoost. 2019. Disponível em: http://sonneblog.com/2019/03/14/Decision-Tree-Random-Forestsand-AdaBoost/. Acesso em: 20 set. 2020.

O DIÁRIO (São Paulo). **EDP descobre 3,74 mil fraudes de energia elétrica no Alto Tietê**. 2019. Disponível em: http://www.odiariodemogi.net.br/edp-descobre-374-mil-fraudes-de-energia-eletricano-alto-tiete/. Acesso em: 23 jun. 2020.

PACHECO JUNIOR, João Carlos. **MODELOS PARA DETECÇÃO DE FRAUDES UTILIZANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA**. 2019. 102 f. Dissertação (Mestrado) - Curso de Mestre em Economia, Escola de Economia de São Paulo, Fundação Getúlio Vargas, São Paulo, 2019. Disponível em: http://bibliotecadigital.fgv.br/dspace/bitstream/handle/10438/27166/Dissertacao_Joao_Carlos_Pacheco_VFinal_2.pdf?sequence=3&isAllowed=y. Acesso em: 08 jun. 2020.

PEDREGOSA, Fabian et al. Scikit-learn: Machine Learning in Python. Journal Of Machine Learning Research. Brookline, p. 2825-2830. out. 2011. Disponível em: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf. Acesso em: 08 ago. 2020.

AGRADECIMENTOS

Gostaríamos de agradecer em primeiro lugar aos nossos pais, e a nossa família, que desempenharam um papel fundamental com todo o suporte em nossas vidas.

E também, pelo ano, pelas reuniões, pelo suporte e apoio como orientador e professor, um imenso obrigado ao Cleber Roberto Guirelli, quem nos ajudou muito durante o percorrer do ano todo, e sem ele não poderíamos ter chegado até este ponto.

Aos nossos amigos que estiveram conosco durante esses 5 anos. Ao Fabrício e a Patricia, os quais nos deram um excelente suporte ao longo deste trabalho.

E, por fim, gostaríamos de agradecer o Engenheiro Victor Costa, responsável por compartilhar um pouco de seu conhecimento e estar sempre disposto a ajudar no que for necessário e em seu alcance.

Foi uma honra ter concluído este trabalho, e com essas pessoas o caminho ficou menos difícil.