GRUPO DE ESTUDO DE SUBESTAÇÕES E EQUIPAMENTOS DE ALTA TENSÃO - GSE

ENSIGHTS: MONITORAMENTO INTELIGENTE DE ATIVOS DE TRANSMISSÃO

MARCELO DE CARVALHO(1);ANA CRISTINA DE FREITAS MAROTTI(1);ALEX DE VASCONCELLOS GARCIA(2);ANA CLÁUDIA RODRIGUES(1);LÁZARO MENEZES BRITO(1);CARLA CHRYSTINA DE CASTRO PACHECO FERREIRA(2);GABRIEL RESENDE MACHADO(2);EDMILSON VAREJAO(2);PEDRO HENRIQUE SCHNEIDER(2);JEFFERSON DE BARROS SANTOS(2);MAURICIO RAPHAEL WAISBLUM BARG(3);CLAUDIO GÜTTLER(1);EDWARD HERMANN HAEUSLER(2)
FURNAS(1); PUC-RIO(2);RADIX ENGENHARIA E DESENVOLVIMENTO DE SOFTWARE S/A(3)

RESUMO

Através do emprego de técnicas de Inteligência Artificial (Machine Learning) em bases de dados de ativos de transmissão de uma Subestação, é possível construir modelos preditivos que permitam aprimorar os planos de manutenção desses equipamentos, reduzindo o tempo e a quantidade de desligamentos e, consequentemente, a indisponibilidade das Funções de Transmissão correspondentes. Como resultado, almeja-se aprimorar a manutenção preditiva realizada atualmente em FURNAS, empregando técnicas de Data Science aplicadas aos dados disponíveis sobre os equipamentos, incluindo os dados de sensoriamento já existentes.

PALAVRAS-CHAVE
Data Science, Subestações AT, Modelos Analíticos, Monitoramento Inteligente, Ativos de Transmissão, Agnostic PAC Learning.

1.0 INTRODUÇÃO

Transformação Digital é o processo em que empresas usam tecnologias digitais para solucionar problemas tradicionais, como: quedas no desempenho, produtividade, agilidade e eficácia.
Dentre as diretrizes estratégicas de Furnas, destacam-se o estabelecimento de uma cultura focada em dados (data driven), e desenvolvimento da cultura do uso estratégico da informação, por meio do estímulo a habilidades analíticas no negócio e Data & Analytics. Tal diretriz estratégica é responsável por fomentar o uso estratégico dos dados na resolução de problemas de negócio.
Nesse aspecto, a Inteligência Artificial (IA) destaca-se como a principal tecnologia capaz de lidar com grandes quantidades de dados e extrair informação relevante para o negócio.
Dentro deste contexto, o presente trabalho tem por objetivo apresentar uma das iniciativas de Transformação Digital com uso de IA, através do desenvolvimento de um Projeto de P&D+I, intitulado “Aplicabilidade e Implementação de nova tecnologia voltada para o desenvolvimento de um modelo de monitoramento inteligente de ativos de transmissão”.
Essa iniciativa permitirá aplicar técnicas de Inteligência Artificial nos planos de manutenção de ativos de transmissão de Furnas. O projeto terá como produtos: (i) data lake de dados de operação e manutenção de ativos de transmissão, (ii) modelos analíticos de ativos de transmissão e (iii) plataforma para aprimoramento dos planos de manutenção.

2.0 RESULTADOS ESPERADOS

Neste projeto, espera-se obter uma plataforma computacional capaz de reduzir o nível de interrupções em equipamentos pertencentes aos ativos de transmissão. É sabido que a parada de funcionamento não programada de máquinas e equipamentos é uma das principais dores de negócios de empresas de energia. Não obstante os interesses da empresa, essas paradas, normalmente desencadeadas por falhas, afetam diretamente a operação do Sistema Elétrico. Nesse contexto, a previsão da ocorrência de falhas, e a consequente manutenção preventiva pode reduzir a incidência de falhas e suas consequências para o sistema e a empresa.
A manutenção preditiva determina o agendamento das ações de manutenção de forma adaptativa, ao invés da forma fixa como é o caso da manutenção preventiva; ou seja, em vez de depender de estatísticas de vida média industrial ou interna (ou seja, tempo médio até a falha) para programar atividades de manutenção, a manutenção preditiva usa monitoramento direto da condição mecânica, eficiência do sistema e outros indicadores para determinar a tomada de decisão de manutenção.

marcarv@furnas.com.br

Dentre as metodologias de monitoramento da condição mecânicas, as técnicas mais modernas se utilizam de modelos analíticos com foco em ações preditivas, via aplicação de algoritmos que são capazes de identificar eventos de falhas a partir do monitoramento de grande número de variáveis em alta frequência - empregando solução computacional para testes e produção de Big Data, Analytics e IA. As técnicas de Machine Learning (ML) permitem identificar padrões significativos em grandes quantidades de dados e gerar novos insights para melhorar a disponibilidade de ativos. Isso se dá porque esses algoritmos não são construídos como um conjunto predefinido de regras, como na programação de software tradicional. Em vez disso, esses algoritmos são de autoaprendizagem, isto é, eles inferem regras, realizando uma série de tentativas em um conjunto de dados de treinamento e, assim, constroem seu próprio modelo. Cada quantidade subsequente de dados é, então, usada para refinar esse modelo e melhorar seus poderes preditivos.
Como resultado, é possível utilizar algoritmos de Machine Learning sobre dados de subestações, para obter uma melhor confiabilidade em termos de identificação de eventos de falhas futuras. Com isso, espera-se:

- Menor incidência de falhas, o que, consequentemente, gera redução do tempo de inatividade não planejado, menos necessidade de inspeções redundantes e medidas de manutenção preventiva ineficazes.
- Custos reduzidos em consequência do aumento do ciclo de vida do equipamento por meio de melhor desempenho e vida útil prolongada do equipamento.
- Benefícios indiretos, incluindo qualidade aprimorada, retrabalho reduzido, defeitos reduzidos, segurança aprimorada e maior eficiência energética.

De acordo com dados da McKinsey, as ferramentas de manutenção preditiva podem reduzir o tempo de inatividade da máquina de manufatura de 30% a 50% e aumentar a vida útil da máquina em 20% a 40%.

3.0 CONSIDERAÇÕES TEÓRICAS

Nesta seção, traçamos algumas considerações teóricas sobre a complexidade amostral para construção do modelo de predição de falhas. Na construção de modelos de classificação supervisionada, ou aprendizagem supervisionada, na área de aprendizagem de máquina, a qualidade dos dados usados na fase de aprendizado, ou treinamento, é de crucial importância. De igual importância é o tamanho do conjunto de dados usado no treinamento, sendo determinante para a qualidade do aprendizado. A intuição é que quanto mais exemplos do que queremos que a máquina aprenda fornecermos, mais apurada será a aprendizagem. Uma boa aprendizagem deve levar em conta também uma boa distribuição de exemplos em relação ao universo de amostras possíveis.
Chamamos de complexidade amostral o limitante inferior para a quantidade de exemplos fornecidos a um algoritmo de aprendizado necessários para que, com confiança alta, este classifique com alta acurácia. A complexidade amostral depende do algoritmo, da distribuição de probabilidade dos dados, da acurácia desejada, bem como da confiança com a qual desejamos atingir esta acurácia.
Em (1), é apresentada uma abordagem geral de teoria do Aprendizado de Máquina (AM). Em (2), é apresentado um teorema que consegue estimar esta quantidade mínima de eventos, independentemente do algoritmo de ML a ser utilizado e da distribuição de probabilidade associada às amostras na base. O trabalho de Valiant (1) define matematicamente quando um algoritmo de aprendizado é adequado para fornecer um modelo de classificação/predição com acurácia e confiança fornecidas *a priori*. Haussler (2) fornece uma estimativa inferior na complexidade amostral para que o modelo forneça classificações com acurácia e confiança dadas a priori.

3.1 Um limitante inferior para o tamanho do conjunto de treinamento

No aprendizado indutivo simbólico supervisionado, ou Aprendizado de Máquina, um algoritmo (de aprendizado) recebe como entrada um conjunto de exemplos de treinamento. Cada exemplo está classificado como pertencente a uma determinada classe. O objetivo do algoritmo é produzir uma função de classificação, que pode ser descrita como um conjunto de regras, uma rede neural ou outro artefato computacional adequado. Essa função é, então, utilizada para predizer, com alguma precisão, a classe dos novos exemplos. Nesta seção, denominamos estas classes como *conceitos*.
Vamos considerar os exemplos de um conceito a ser aprendido como um vetor de atributos $\langle e_1, \dots, e_n \rangle$. Por exemplo, um algortimo que deseja classificar o tipo físico de pessoas em *magro* ou *gordo*, sendo cada indivíduo descrito por um vetor de duas componentes: $\langle altura, peso \rangle$. O objetivo do algoritmo de aprendizado é aprender uma função $f(altura, peso) = 1$, se o indivíduo é gordo, e 0, caso contrário. Tal função é dita ser um classificador binário.
O conjunto de treinamento T é constituído por pares $(\vec{e}, f(\vec{e}))$. Um algoritmo de aprendizado, então, recebe o conjunto T e produz uma função h (*hipótese*) que aproxima f. É desejável determinar qual a menor cardinalidade do conjunto de treinamento necessária para um aprendizado com boa aproximação. A teoria do aprendizado computacional vem responder à questão que indaga como saber que h se aproxima bem de f se não sabermos quem é f.
Valiant (1) apresenta a noção de sistema de *aprendizado provavelmente correto*. Ele usa conceitos da Teoria da Computação e Complexidade computacional para análise mais precisa das ferramentas e algoritmos existentes em aprendizado indutivo computacional. A premissa do sistema de aprendizado provavelmente correto diz que: qualquer h que não classifica bem um conceito c, tem uma alta probabilidade de ser reconhecida como ruim após um determinado número de exemplos, quando realizará uma predição incorreta. Desta forma uma h consistente com um

número grande de exemplos de treinamento é improvável de estar incorreta. Isto é, f é *provavelmente aproximadamente correta*.
A terminologia a seguir ajuda a expressar o teorema que exibe um limitante inferior para o tamanho do conjunto de treinamento apropriado para a obtenção de uma boa hipótese h dentro de um conjunto de possíveis hipóteses para um conceito binário arbitrário. Vamos considerar:

- Uma população X e uma distribuição de probabilidade D sobre X;
- Um conceito $f_c(x)$ e um conjunto de treinamento $T \subseteq X$ sob a mesma distribuição de probabilidade D;
- Um conjunto H de todas as possíveis hipóteses, e uma hipótese $h(x)$ consistente com T obtida a partir de algum algoritmo de aprendizado consistente;
- $m = card(T)$

Dada uma acurácia $\varepsilon, 0 < \varepsilon < 1$, e, uma confiança $\delta, 0 < \delta < 1$, consideramos as seguintes definições:

$$erro(h) = Prob[f_c(x) \neq h(x), x \text{ escolhido em } X \text{ de acordo com } D]$$

Um algoritmo de aprendizado A é (ε, δ)-bom sempre que h é consistente com T, temos que:

$$Prob[erro(h) > \varepsilon] < \delta$$

De certa forma, a noção de acurácia usada aqui é, na realidade, o limite do erro de f em relação a f_c. A acurácia usualmente usada em Aprendizado de Máquina é $1 - \varepsilon$. O mesmo pode ser dito sobre a confiança δ. Temos, então, o teorema de Haussler que fornece um limitante inferior para o tamanho de um conjunto de treinamento relativo a um algoritmo de aprendizado (ε, δ)-bom. Isto também é conhecido como a *complexidade da amostragem*, ou *sample complexity*.

Teorema [Haussler]. Sejam $\varepsilon, 0 < \varepsilon < 1$, $\delta, 0 < \delta < 1$, $T \subseteq X$, $n = card(H)$ e $m = card(T)$. Tem-se que se um algoritmo é (ε, δ)-bom, então:

- Se X é finito então $m \geq \frac{1}{\varepsilon} \times (\ln n + \ln(\frac{1}{\delta}))$
- Se X é infinito então $m \geq \frac{1}{\varepsilon} \times (VCDim(H) + \ln(\frac{1}{\delta}))$

$VCDim(H)$ é a dimensão de *Vapnik-Chervonenkis* de H. Trata-se de um conceito que mede a complexidade do conjunto de classificadores. Para os propósitos deste trabalho, basta sabermos que $\dim(X) \leq VCDim(H)$, onde $\dim(X)$ é a dimensão do espaço vetorial X.

3.2 Análise da complexidade amostral do modelo desenvolvido

Para o modelo preliminar desenvolvido para o projeto e apresentado na Seção 5, a dimensão do espaço vetorial dos dados de transformadores de potência disponíveis para *transfer learning* (transferência de aprendizado) das características de óleos e outras propriedades dos transformadores de potência é 9. Originalmente, a base de *transfer learning* classificava em 6 tipos de falha, além da classe *sem falha*, totalizando 7 classes. Para nos aproximarmos das necessidades previstas para classificadores binários, buscamos agrupar as classes, ficando, ao final, com 2 tipos de falha, além da classe *sem falha*.
Um fato teórico importante é como o tipo de classificador ou algoritmo de aprendizado associado está estritamente ligado à dimensão *VCdim*. Somente para ilustrar e justificar a nossa escolha pelo uso de *Random Forests* no nosso modelo, mostramos na Tabela 1, contendo a relação entre alguns algoritmos de classificação supervisionada e a *VCdim* dos classificadores resultantes. Outros classificadores como NN, CNN e RNN (tipos de redes neurais) possuem milhares de pesos associados às sinapses e, consequentemente, VC muito mais alto que o tamanho dos conjuntos de treinamento existentes. Claramente, não estamos no escopo de classificadores lineares. Dentre os classificadores, os que temos mais condições de trabalhar para o uso mais adequado dos conjuntos de treinamento existentes, que não são grandes, são as árvores de decisão. Este foi, de fato, o classificador que foi desenvolvido. Em se tratando de *Random Forests*, não existe uma expressão analítica que sintetize a VC destas. Na prática estima-se a *VCdim a posteriori*, após a remoção das árvores que têm por raiz *features* de baixa relevância.
Finalmente, é importante relatar que a escolha de um modelo não é somente determinada pelo critério de superação da complexidade amostral pelo conjunto de treinamento. Outros critérios são Cross-Validation (CV), Aikake Information Criteria (AIC), Bayesian Information Criteria (BIC) e Minimização de Risco Estrutural com dimensão *VCdim*.

TABELA 1 - Tipos de Classificadores e sua dimensão VCDim.

	Classificador	Dimensão VC
1	Linear	$VC = dim(E) + 1$
2	Gaussiano	$VC = \infty$
3	SVM	$VC = min(dim(E), \frac{D^2}{M^2})$
4	Redes Neurais	$VC = dim(E) \times \|N\|$
5	Árvores de decisão T	$VC \approx \#nos(T)$
6	Redes Bayesianas B	$VC \approx \#params(T)$

4.0 ARQUITETURA DA SOLUÇÃO

Apesar de, a princípio, lidarmos basicamente com fontes de dados estruturados (provenientes de bancos de dados relacionais) nesse projeto, decidimos utilizar uma arquitetura típica de projetos *Big Data*. Arquiteturas para *Big Data* precisam lidar com fontes de dados heterogêneas e em diferentes formatos (estruturados e não estruturados), que trafegam em grandes volumes e, muitas vezes, precisam ser processados à medida que são transmitidos (que muitas vezes acontece em grande velocidade – como dados de sensores e monitoramento de tempo real). Ao longo dos anos, várias arquiteturas de referência foram propostas para projetos desse tipo, como as arquiteturas Lambda (3) e Kappa (4). Princípios e fundamentos de plataformas de dados com essas características podem ser vistas em (5).
A Figura 1 apresenta a arquitetura de referência que utilizamos nesse projeto. Uma arquitetura de referência apresenta os componentes de software existentes em um sistema computacional e define o papel de cada um desses componentes e as relações entre eles. Essa descrição é feita em alto nível, sem considerar os detalhes técnicos e tecnologias específicas. Essa abordagem foi essencialmente importante dada a característica do nosso projeto de que a contratação da tecnologia seria estabelecida ao longo do projeto em processo formal de licitação de fornecedores.

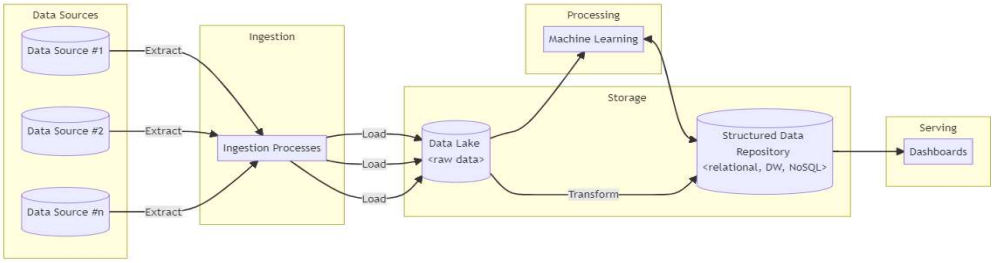

FIGURA 1 - Arquitetura de Referência utilizada no Projeto

Usamos uma arquitetura de 4 camadas, bastante difundida na indústria (6). Embora essas camadas lógicas sejam adequadas para descrever arquiteturas de diferentes níveis de complexidade, elas estão populadas e adaptadas com os componentes mínimos necessários para os objetivos do nosso projeto. A camada de Ingestão (*Ingestion*) é responsável pela obtenção dos dados das fontes originais e seu carregamento na plataforma de dados em nuvem. Uma vez carregados na plataforma, os dados são carregados em um *Data Lake*, um repositório de dados em sua forma bruta. Processos automatizados de limpeza e tratamento dos dados realizam as transformações necessárias para alimentar um Repositório de Dados Estruturados (que aqui pode ser implementado como um *Data Warehouse*, banco de dados relacional ou NoSQL). Esses dois componentes formam a camada de Armazenamento (*Storage*). Na camada de Processamento (*Processing*), os dados são utilizados para análise exploratória e para treinamento e testes de modelos de Aprendizado de Máquina. Por fim, na camada Serviço (*Serving*), os dados alimentam *dashboards* e aplicativos de visualização.
Essa arquitetura foi instanciada para cada provedor de Nuvem considerado para o projeto: cada componente foi mapeado em um serviço particular do fornecedor. Essa abordagem nos permitiu comparar de forma mais fácil e precisa as soluções disponíveis para implementação. Por fim, optamos pela implementação da solução usando a plataforma Azure da Microsoft.

5.0 METODOLOGIA E RESULTADOS PRELIMINARES

A parte mais desafiadora da arquitetura apresentada na Seção 3 é a criação e avaliação de modelos que contribuam para os objetivos apresentados na Seção 2. Trata-se de um trabalho que envolve muita experimentação e que melhor caracteriza a componente de pesquisa deste projeto de P&D.
Nesta fase inicial do projeto, concentramos os esforços em transformadores de potência, devido à sua importância, uma vez que sua indisponibilidade acarreta grande custo à empresa. A presente seção descreve os modelos criados a partir da análise cromatográfica dos gases dissolvidos em óleo isolante utilizado nos transformadores. Outros dados, como aqueles disponíveis no SAGE SCADA/EMS, também estão sendo processados e farão parte da solução.

Uma vez que o os transformadores são equipamentos confiáveis e que apresentam poucas falhas, há uma carência de dados de ensaios cromatográficos em casos de falha. Portanto, foi necessário buscar o transfer learning como ponto de partida para a criação dos modelos que implementam estes diagnósticos. Para tanto, foi conduzida uma pesquisa bibliográfica, de modo que fossem reunidos em um único dataset, dados rotulados de natureza semelhante aos dados in situ FURNAS, i.e., os atributos da mesma natureza, porém contendo os diagnósticos de falha.

5.1 Coleta dos dados

A coleta dos dados para a elaboração do *dataset* unificado utilizado para o procedimento de concepção do modelo de aprendizado reuniu seis conjuntos de dados disponibilizados por quatro trabalhos disponíveis na literatura (7, 8, 9, 10). A seguir, é mencionada a origem e atribuído um identificador para cada um desses seis *datasets* coletados a partir desses trabalhos.

- ***IEEE1:*** *dataset* extraído diretamente da Tabela 9 presente em (7);
- ***IEEE2:*** *dataset* extraído do arquivo PDF e disponível online por (7);
- ***IBRAHIM:*** dados disponibilizados *online* por (8);
- ***IECTC:*** *dataset* que contém apenas registros de transformadores diagnosticados como "*Normal*", retirados da Tabelas 1 e 2 do Anexo 2 de (9);
- ***UFSC1***: dados extraídos do artigo (10), das Tabelas *IEC TC10 - Normal* (Anexo A) e *CEPEL - Normal* (Anexo B);
- ***UFSC2:*** dados classificados como "*Normal*" extraídos das Tabelas que compõem os dados históricos do Anexo C de (10).

As classes de diagnóstico (ou rótulos) identificados a partir desses *datasets* são: ***NF - (No Faults)***; ***PD - (Partial Discharge)***; ***D1 - (Low Energy/Spark Discharge)***; ***D2 - (High Energy/Arc Discharge)***; ***T1 - Low Temperature Fault (t < 300ºC)***; ***T2 - Middle Temperature Fault (300ºC ≤ t ≤ 700ºC)***; ***T3 - High Temperature Fault (t > 700ºC)***.

5.2 Pré-processamento dos dados

Com a elaboração do dataset unificado, iniciou-se a etapa de pré-processamento dos dados, que consiste em duas fases: (i) limpeza e (ii) normalização/padronização dos dados.

O processo de limpeza dos dados consistiu em duas etapas principais: (1) substituição de valores nulos e inválidos e (2) remoção de amostras duplicadas ou com valores em branco. Durante a coleta e unificação dos *datasets* descritos pela Seção 5.1, a Etapa 1 iniciou-se de modo que alguns dados nulos ou inválidos foram tratados. Os *datasets* que necessitaram algum tratamento do tipo foram: ***IBRAHIM***, ***IECTC***, ***UFSC1*** e ***UFSC2***.

Após a limpeza e a unificação dos *datasets*, iniciou-se a Etapa 2, na qual foi executado um algoritmo de análise e remoção de amostras duplicadas. Após a remoção de amostras duplicadas, o *dataset* unificado apresentou a distribuição dos dados descrita pela Tabela 2, de acordo com as classes de diagnósticos.

TABELA 2 - Distribuição dos dados no *dataset* unificado após a remoção de amostras duplicadas

Dataset	Gases Medidos							Quantidade de Registros por Classe							Total de Registros
	H2	CH4	C2H2	C2H4	C2H6	CO	CO2	NF	PD	D1	D2	T1	T2	T3	
IEEE1	✓	✓	✓	✓	✓	×	×	0	10	10	10	10	10	10	60
IEEE2	✓	✓	✓	✓	✓	×	×	0	16	49	54	19	9	38	185
IBRAHIM	✓	✓	✓	✓	✓	×	×	0	50	84	143	109	68	110	564
IECTC	✓	✓	✓	✓	✓	✓	✓	27	0	0	0	0	0	0	27
UFSC1	✓	✓	✓	✓	✓	×	×	119	0	0	0	0	0	0	119
UFSC2	✓	✓	✓	✓	✓	×	×	191	0	0	0	0	0	0	191
Total por Classe								**337**	**76**	**143**	**207**	**138**	**87**	**158**	**1146**
Total por Classe (remoção de linhas duplicadas)								**333**	**70**	**133**	**183**	**119**	**62**	**122**	**1022**

Após a unificação e limpeza dos dados ex situ FURNAS, descrito pela Tabela 2, foi realizado o cálculo de cinco razões (frações) entre os gases. Esses quocientes foram calculados por dois motivos principais: (i) para a realização de um estudo comparativo do modelo de aprendizado com os principais métodos clássicos de diagnóstico em óleo isolante (Figura 2), que fazem uso desses quocientes durante o processo que determina o diagnóstico final do equipamento, e (ii) enriquecer o dataset com novos atributos. Os cinco quocientes de gases calculados e inseridos como atributos (colunas) no dataset unificado usaram a correção de valores não detectáveis sugerida em (4), portanto: R1 = (CH4 / H2) + 0,4; R2 = (C2H2 / C2H4) + 0,4; R4 = (C2H6 / CH4) + 0,4 e R5 = (C2H4 / C2H6) + 0,4.

A normalização/padronização é um processo de escalonamento de valores de atributos que possuem intervalos numéricos muito discrepantes entre si, o que para a maioria dos algoritmos de aprendizado pode afetar a importância de alguns atributos com maiores ordens de grandeza em detrimento de outros com menores ordens de grandeza, criando, desta forma, vieses indesejáveis.

Há diversas premissas para a padronização de variáveis em conjunto de dados. Entretanto, para a padronização do *dataset* unificado, foi escolhida a medida baseada em IQR (*Interquartile Range*), por ser robusta aos *outliers,* O cálculo feito pela padronização IQR, para um determinado valor x, é descrito pela Equação 1. A Equação 1 foi aplicada para os seguintes atributos do *dataset* unificado: (i) H2, (ii) CH4, (iii) C2H4, (iv) C2H6, (v) R1, (vi) R4 e (vii) R5. Já os atributos correspondentes ao gás acetileno (C2H2) e à razão R2 precisaram de um tipo de padronização específica, devido à distribuição dos seus valores possuir muitos *outliers* e valores iguais a zero. Para a padronização do acetileno, foi utilizada a Equação 2, onde $x \in C2H2$ e, para a padronização da razão R2, foi utilizada a Equação 3, sendo $y \in R2$.

$$Padr_{IQR} = \frac{x - Q_{2/4}}{Q_{3/4} - Q_{1/4}} \quad (1)$$

$$Padr_{C2H2} = \frac{log(x + 1)}{max(log(C2H2 + 1))} \quad (2)$$

$$Padr_{R2} = \frac{y}{max[R2]} \quad (3)$$

Por fim, vale ressaltar que as mesmas etapas de normalização também foram aplicadas no *dataset in situ* FURNAS, que, por sua vez, foi utilizado após a concepção e avaliação do modelo para fins de apresentação em um *dashboard* (vide Figura 2).

5.3 Concepção do Modelo de Aprendizado

Com as etapas de unificação e pré-processamento do *dataset ex situ* FURNAS concluídas, foi necessário conduzir um estudo comparativo entre alguns algoritmos de AM (a partir do *software Weka* (15)) para definir, dentre eles, o melhor. Para o estudo comparativo, foram escolhidos os seguintes algoritmos de AM presentes no *Weka*:

- *Support Vector Machines* (LibSVM) (11):
- *Random Forest* (12);
- *Fuzzy Unordered Rule Induction Algorithm* (FURIA) (13);
- *Random Trees* (14).

Para a comparação dos algoritmos, foram usados os resultados de acurácia média obtidos pelos modelos correspondentes em procedimentos de validação cruzada com 5 e 10 partições nos seguintes cenários de avaliação:

- ***A1:*** agregação das classes *PD, D1* e *D2* em (i) *Falha Elétrica*; *T1, T2* e *T3* em (ii) *Falha Térmica* e (iii) *Normal*;
- ***A2:*** agregação das classes *D1, D2*, *T1, T2* e *T3* em (i) *Descarga Elétrica e/ou Falha Térmica; PD* em (ii) *Arco Elétrico* e (iii) *Normal*;
- ***I:*** sem agregação, *i.e.* as sete classes individuais.

Vale ressaltar que os três cenários de avaliação foram combinados com mais três configurações, no tocante à presença de determinados quocientes de gases:

- ***R1, R2 e R5:*** *dataset* unificado com as razões R1, R2 e R5, sendo R4 descartada;
- ***sem R5:*** *dataset* unificado com as razões R1, R2 e R4, sendo R5 descartada;
- ***com R5:*** *dataset* unificado com a presença das razões R1, R2, R4 e R5.

A Tabela 3 apresenta os resultados obtidos a partir dos experimentos, onde os melhores resultados em cada cenário de avaliação estão destacados em amarelo. A partir da Tabela 3, é possível observar que a *Random Forest* apresentou os melhores resultados em todos os cenários de avaliação com 10 partições de validação cruzada, sendo o seu melhor resultado obtido no cenário de avaliação A1 com R5 em 10 validações cruzadas (92,66%).

TABELA 3 - Resultados dos experimentos conduzidos em cada cenário de avaliação.

ALGORITMOS (*Weka*)	A1 sem R5	A1 com R5	A1 R1, R2, R5	A2 sem R5	A2 com R5	A2 R1, R2, R5	I sem R5	I com R5	I R1, R2, R5
LibSVM	74,76%	74,17%	70,55%	72,11%	73,09%	71,43%	61,35%	62,82%	62,72%
Random Forest	92,37%	92,66%	91,39%	90,22%	90,02%	90,31%	83,27%	85,42%	84,64%
FURIA	89,53%	89,24%	88,65%	86,01%	85,42%	86,20%	77,20%	79,84%	79,55%
Random Tree	80,82%	81,12%	80,14%	75,83%	72,50%	74,76%	63,70%	64,38%	63,70%
# Regras *fuzzy* (FURIA)	25	25	23	26	23	32	38	35	42

Com a unificação do *dataset ex situ* FURNAS e a definição do melhor algoritmo de aprendizado, deu-se início à etapa de treinamento e avaliação do modelo de aprendizado. A partir dessa fase, foi utilizada a linguagem de programação

Python, com a biblioteca *Scikit-learn*. Primeiramente, o *dataset* unificado *ex situ* FURNAS foi dividido em dois conjuntos de dados: (i) o conjunto de treinamento, contendo 80% das amostras e (ii) o conjunto de teste, contendo os 20% restantes. Após, foi utilizado o método *RandomizedSearchCV* no conjunto de treinamento com 10 validações cruzadas para a calibração dos hiperparâmetros da *Random Forest.* Foi possível concluir que os valores de hiperparâmetros padrão fornecidos pelo *Scikit-learn* para o *Random Forest* produziram melhores resultados de acurácia e métrica F1, que por sua vez ficaram bem próximos ao melhor resultado de acurácia da *Random Forest* no *Weka* (92,66%).

Com a definição dos melhores valores de hiperparâmetros para a *Random Forest,* o modelo produzido após o ajuste no conjunto de treinamento foi, enfim, validado no conjunto de teste, juntamente com os métodos clássicos para diagnóstico em óleo isolante: (i) *Rogers*, (ii) *Doernenburg*, (iii) NBR 7274, (iv) IEC 599 e (v) Triângulo de Duval. Foram também avaliados os métodos refinados de (vi) *Rogers-R* (Rogers refinado) e (vii) IEC-R (IEC refinado) (16), além de duas abordagens híbridas: (viii) *Doernenburg + Duval (Doerneval)* e (ix) *Doernenburg + IEC Ibrahim (DIEC-R)*. Os resultados dos experimentos encontram-se na Tabela 4, onde os piores desempenhos estão destacados em vermelho e os melhores, em azul.

A partir da Tabela 4, observa-se que o método clássico de *Doernenburg* apresentou, dentre os demais métodos clássicos, o melhor resultado de acurácia. Por esse motivo, ele foi escolhido para formar, juntamente com os melhores métodos de diagnósticos de falhas, os métodos híbridos *Doerneval* e *DIEC-R*.

O método híbrido DIEC-R apresentou o melhor resultado de 0,732 de acurácia e 0,713 na métrica F1, enquanto a Random Forest apresentou uma acurácia de 0,922 e métrica F1 de 0,921, uma diferença de, aproximadamente, 19 pontos percentuais, mostrando-se o melhor indicador para o *dashboard* para a predição dos diagnósticos de ensaios cromatográficos no *dataset in situ* de FURNAS.

TABELA 4 - Resultados dos experimentos dos nove métodos de diagnóstico e o modelo no conjunto de teste.

Método	Acurácia (somente classe 0)	Acurácia (somente falhas)	Acurácia (todas as classes)	Acurácia (conjunto de teste)	F1-score (conjunto de teste)
Rogers	0.081	0.476	0.347	0.351	0.242
Rogers (refinado)	0.102	0.680	0.491	0.468	0.278
Doernenburg	0.399	**0.013**	**0.139**	**0.136**	**0.145**
NBR 7274	0.192	0.665	0.511	0.517	0.433
IEC Ratio	0.192	0.620	0.481	0.512	0.410
IEC (refinado)	0.144	**0.908**	0.660	0.644	0.566
Triângulo de Duval	**0.000**	0.901	0.608	0.605	0.388
Doernenburg + Duval	0.399	0.871	0.717	0.697	0.523
Doernenburg + IEC (Ibrahim)	**0.450**	0.882	**0.742**	0.732	0.713
Random Forest	-	-	-	**0.922**	**0.921**
	Número de amostras				
	333	689	1022	205	

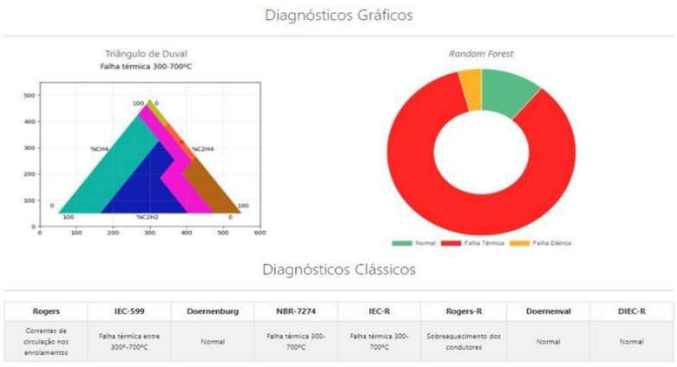

Figura 2 - Exemplo de tela com informações sobre o último ensaio cromatográfico realizado no equipamento, como também indicadores gráficos e resultados dos principais diagnósticos clássicos e híbridos

6.0 CONCLUSÕES

Planejamos uma arquitetura de referência com o objetivo de facilitar a comunicação entre os diferentes participantes, criando um vocabulário comum para as diferentes etapas de manipulação de dados no projeto. Instanciando essa arquitetura com os serviços e tecnologias de cada fornecedor de Nuvem considerado no projeto, conseguimos uma comparação mais adequada entre os mesmos e, consequentemente, podemos decidir de forma mais clara por qual contratar.

Elencamos algumas especificações funcionais referentes à uma primeira versão de *dashboard* em formato de aplicativo *Web*, voltada para a análise cromatográfica de óleo isolante em ativos, mais especificamente, em transformadores. Para isso, foi descrito todos os processos relacionados à coleta, ao pré-processamento dos dados e à elaboração e avaliação do modelo de aprendizado *Random Forest* utilizado como indicador em um protótipo de *dashboard,* que será evoluído no decorrer do cronograma do Projeto *EnSights* de modo a, no final, atender às necessidades operacionais de FURNAS.

Os produtos em desenvolvimento possibilitarão o aprimoramento dos planos de manutenção dos ativos de transmissão, utilizando técnicas de Inteligência Artificial (Machine Learning), a partir da correlação dos dados de operação e manutenção de equipamentos em subestações de Furnas; o apoio do processo de Gestão de ativos de Transmissão de Furnas, e a redução da incidência de Parcela Variável (PV).

O projeto de P&D+I, objeto deste trabalho, está em consonância com o processo de Inovação Tecnológica e Transformação Digital, elencado como tema estratégico através da Resolução no 2 do Ministério de Minas e Energia de 10/2/2021, e com o Decreto 10.332 de 28/4/2020 do Governo Federal, que institui a Estratégia de Governo Digital e estabelece, entre outras iniciativas, como a seguinte: Iniciativa 8.2. Implementar recursos de inteligência artificial em, no mínimo, doze serviços públicos federais, até 2022.

7.0 REFERÊNCIAS BIBLIOGRÁFICAS

(1) VALIANT, Leslie. A theory of the learnable. Commun. ACM, 27(11):1134–1142, 1984.
(2) EHRENFEUCHT, Andrzej, HAUSSLER, David, KEARNS, Michael, and VALIANT, Leslie. A general lower bound on the number of examples needed for learning. In Proceedings of the First Annual Workshop on Computational Learning Theory, COLT ’88, Cambridge, MA, USA, August 3-5, 1988, pages 139–154. ACM/MIT, 1988.
(3) MARZ., & WARREN. (2015). Big data: Principles and best practices of scalable real-time data systems. Manning.
(4) KREPS, J. (2014). Questioning the lambda architecture. Online Article, July, 205. https://www.oreilly.com/radar/questioning-the-lambda-architecture/
(5) KLEPPMANN, M. (2017). Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems (1a edição). O’Reilly Media.
(6) ZBURIVSKY, D., & PARTNER, L. (2021). Designing Cloud Data Platforms (1a edição). Manning Publications.
(7) LI, Enwen; WANG, Linong; SONG, Bin. Fault diagnosis of power transformers with membership degree. IEEE Access, v. 7, p. 28791-28798, 2019.
(8) IBRAHIM, Saleh I.; GHONEIM, Sherif SM; TAHA, Ibrahim BM. DGALab: an extensible software implementation for DGA. IET Generation, Transmission & Distribution, v. 12, n. 18, p. 4117-4124, 2018.
(9) DUVAL, Michel; DEPABLA, A. Interpretation of gas-in-oil analysis using new IEC publication 60599 and IEC TC 10 databases. IEEE Electrical Insulation Magazine, v. 17, n. 2, p. 31-41, 2001.
(10) MORAIS, Diego Roberto *et al*. Ferramenta inteligente para detecção de falhas incipientes em transformadores baseada na análise de gases dissolvidos no óleo isolante. 2004.
(11) CHANG, Chih-Chung; LIN, Chih-Jen. LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), v. 2, n. 3, p. 1-27, 2011.
(12) BREIMAN, Leo. Random forests. Machine learning, v. 45, n. 1, p. 5-32, 2001.
(13) HÜHN, Jens; HÜLLERMEIER, Eyke. FURIA: an algorithm for unordered fuzzy rule induction. Data Mining and Knowledge Discovery, v. 19, n. 3, p. 293-319, 2009.
(14) PFAHRINGER Bernhard. Random model trees: an effective and scalable regression method. University of Waikato, New Zealand, http://www.cs.waikato.ac.nz/~bernhard
(15) HALL, Mark et al. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, v. 11, n. 1, p. 10-18, 2009.
(16) TAHA, Ibrahim BM; GHONEIM, Sherif SM; DUAYWAH, Abdulaziz SA. Refining DGA methods of IEC Code and Rogers four ratios for transformer fault diagnosis. In: 2016 IEEE Power and Energy Society General Meeting (PESGM). IEEE, 2016. p. 1-5.

DADOS BIOGRÁFICOS

Bacharel em Ciência da Computação (2010), especialista em Arquitetura Empresarial e Sistemas Corporativos (2017) e Mestrando em Ciência de Dados pela PUC-RIO. Atualmente é o gerente substituto do Departamento de Transformação Digital de Furnas e Head da tribo de monetização desse Departamento. Coordenador Técnico do Projeto de Pesquisa e Desenvolvimento intitulado "Aplicabilidade de nova tecnologia voltada para o desenvolvimento de um modelo de monitoramento inteligente de ativos de transmissão", codinome ENSIGHTS.

(2) ANA CRISTINA DE FREITAS MAROTTI. Engenheira Eletricista pela UFRJ; Mestre e Doutora em Engenhaira Elétrica pela COPPE/RJ e Pós-graduação em Transformação Digital pelo MIT. Possui experiência em planejamento e estudos elétricos; equipamentos de alta tensão e ensaios elétricos. Atua na coordenação técnica, pesquisadora e gestora técnica de projetos, prioritariamente relacionados à Inovação Tecnológica com ênfase em Transformação Digital aplicada a otimização e melhoria de processos de O&M e de Gestão de Ativos da empresa. Coordenada e

lidera grupos de trabalho e comitês de estudos na área de Energia Elétrica e na área de Inovação Tecnológica.

(3) ALEX DE VASCONCELLOS GARCIA. Eng. de Computação pelo IME (1989). Mestre em Sistemas e Computação pelo IME (1992) e Doutor em Informática pela PUC-RIO (2000). Possui 33 anos de experiência em desenvolvimento de Software nas áreas de IA, TEF e Sistemas Embarcados. É professor do IME e pós-doutorando na PUC-RIO.

(4) EDWARD HERMANN HAEUSLER. BSc Matemática UnB (1983). MSc (1986) e DSc (1990) em informática PUC-RIO. Foi professor do DCC da UFF, 1986-1991. É professor do DI, desde 1991. Pós-Docs Univ-Aarhus, Dinamarca (1994), EKUT-Tübingen, Alemanha (2002) e INRIA-França (2012).Foi coordenador de pós-graduação em Informática da PUC-Rio e graduação em engenharia e ciência da computação. Tem mais de 70 artigos em periódicos, 119 em conferências, 14 livros-coletâneas de artigos. Coordena projetos com a indústria, assim como projetos de pesquisa patrocinados por CNPq, CAPES, FAPERJ, DAAD (Alemanha), NSF (EUA), STICAmSuD (França). Pesquisa complexidade computacional , Lógica, Formalização de raciocínio em IA e Eng. de Software.

(5) ANA CLÁUDIA RODRIGUES
Bacharel em Matemática com Pós graduação em Análise de Sistemas pela PUC-RJ, MBA em Gestão de Projetos pela FGV, MBA em Gestão de Negócios com Ênfase no Setor Elétrico pelo IBMEC e Formação em Tranformação Digital pelo MIT. Possui mais de 20 anos de experiência em TI com certificações ITIL, COBIT, BMPM. Atualmente em Furnas, é Head de Transformação Digital. Apoia negócios inovadores como mentora pela ABMEN(Associação Brasileira de Mentores de Negócio) e no InovAtiva Brasil. Também é instrutora do BBI of Chicago.

(6) LÁZARO MENEZES BRITO. Gestor de Projetos de P&D em FURNAS com MBA em Gestão Financeira, Controladoria e Auditoria pela FGV, com graduação em Processamento de Dados e Administração de Empresas.

(7) CARLA CHRYSTINA DE CASTRO PACHECO FERREIRA. Doutora em Engenharia de Defesa pelo IME (2018). Mestre em Engenharia da Computação pelo IME (2012). Bacharel em Informática pela UNESA (2008). Pós-Doutorado em Matemática Aplicada, em projeto de IA em óleo & gás, na PUC-RIO (2019-2020). Sua tese sobre Detecção de Robôs nas Redes Sociais foi tema de palestras na FGV-EMAp, Rede Globo, CEFET-RJ e CEP/FDC - Exército. Possui experiência em grandes empresas. Prêmio de Melhor Projeto de Startup FGV/EBAPE & IME: Pitch para investidores (2017). Pesquisadora do DI/PUC-Rio, onde participa do projeto de P&D I FURNAS/EnSights.

(8) GABRIEL RESENDE MACHADO. Bacharel em Ciência da Computação pelo UNIFESO (2016), mestre em Sistemas e Computação pelo IME (2019). Participa do projeto de pesquisa sobre Fake News sobre COVID-19 (CNPq proc. 401662/2020-9) e pesquisador do DI/PUC-Rio, onde participa do projeto de P&D I FURNAS/Ensights.

(9) EDMILSON VAREJAO. Bacharel (2005), Mestre (2008) e Doutor (2021) em Economia pela FGV/RJ. Foi Visiting Graduate Researcher na University of California, Los Angeles (UCLA) (2018). Foi Economista chefe na Kyros Investimentos, Economista Sênior na Tendências Consultoria e Pesquisador na FGV/RJ. Sócio e CEO da AI Consult.

(10) PEDRO HENRIQUE SCHNEIDER. Engenheiro Elétrico pela PUC-RIO (2000), Mestre em Matemática Aplicada pela FGV/RJ(2016) e doutorando em informática pela PUC-RIO desde 2019, na linha de pesquisa em ciência de dados. Teve mais de 15 anos de experiência no mercado de telecomunicações, trabalhando em grandes empresas multinacionais (2014), pesquisador da FGV-EMAP/RJ (2017) e atuou como cientista de dados na área de antifraude da fintech Zoop (2019). Atualmente, além do doutorado, atua como cientista de dados na AI Consult.

(11) JEFFERSON DE BARROS SANTOS. Doutor em Informática pela PUC-Rio, com período sanduíche no INRIA, na França. Membro colaborador do Laboratório de Tecnologias em Métodos Formais (TecMF) da PUC-Rio, onde realiza pesquisas na área de Teoria da Computação e Lógica, especificamente no estudo de provadores de teoremas, raciocínio automático e inteligência artificial. Atualmente é professor de disciplinas relacionadas com Computação e Tecnologia nos cursos de graduação e mestrado profissional da EBAPE/FGV. Na EBAPE, também coordena o Núcleo de Computação, que desenvolve projetos de desenvolvimento tecnológico para a Escola.

(12) MAURICIO RAPHAEL WAISBLUM BARG. Engenheiro Eletricista graduado pela UFF (2018) e Mestrado em Informática pela PUC-Rio (2021). Possui experiência na criação de algoritmos de machine learning e data science com aplicações voltadas para o setor elétrico. Atua executando projetos de P&D ANEEL voltados para inovação no setor elétrico. Interesse são arquitetura de software, machine learning, data science e otimização.

(13) CLAUDIO GÜTTLER. Engenheiro Eletricista pela PUC-Rio e Mestre em Sistemas e Computação na área de Pesquisa Operacional, linha Simulação de Sistemas pelo IME. Desde 2002 em Furnas – Centrais Elétricas S.A. passando por áreas de TI, Qualidade, Governança Corporativa e Novos Negócios em Energias Renováveis. Atualmente trabalha como Pesquisador e Gestor Técnico de Projetos de P&D I relacionados à Inovação Tecnológica. Integra grupos de trabalho em Transformação Digital.