Explicabilidade e Robustez em Modelos de Detecção de Anomalias Industriais: um Estudo com Random Forest, SHAP, XGBoost e Isolation Forest

Rosemeri Janiski Bida de Oliveira Borges
Programa de Pós-graduação em Ciência da Computação
Inteligência Artificial
UNIFESP
São José dos Campos – SP - Brasil
rose.jbob@gmail.com

Resumo: A aplicação de modelos de Machine Learning (ML) em ambientes industriais críticos enfrenta um desafio significativo: embora métodos supervisionados demonstrem alta precisão preditiva na detecção de anomalias, suas explicações geradas por métodos como SHAP frequentemente apresentam instabilidade crítica, chegando a 0% em testes de estabilidade. Essa instabilidade explicativa compromete diretamente a confiança operacional e dificulta a aplicação prática das soluções em cenários industriais, onde entender as causas raiz das anomalias é tão vital quanto detectá-las com precisão. Este trabalho aborda este "paradoxo da precisão" propondo um framework híbrido inovador que integra a alta precisão preditiva de modelos supervisionados com a estabilidade interpretativa proporcionada por abordagens não supervisionadas, como o Isolation Forest associado ao método LIME. Os resultados experimentais demonstram que essa abordagem híbrida não apenas mantém elevados níveis de precisão na detecção de anomalias, mas também fornece explicações consistentemente estáveis e confiáveis, fundamentais para a tomada de decisões críticas em operações industriais.

Palavras-chave - Detecção de anomalias; explicabilidade; SHAP; manutenção preditiva; 3W Dataset.

I. INTRODUÇÃO

A detecção precoce de anomalias em ambientes industriais é fundamental para garantir segurança e eficiência operacional em setores críticos. Em contextos de alto risco, como plataformas offshore, falhas não detectadas podem causar danos financeiros e ambientais severos. Com processos cada vez mais complexos e grandes volumes de dados gerados por sensores, não basta apenas precisão preditiva elevada; é igualmente essencial que as decisões tomadas pelos modelos de aprendizado de máquina sejam transparentes, robustas e confiáveis.

Nos últimos anos, algoritmos supervisionados (como Random Forest e XGBoost) e não supervisionados (como Isolation Forest) têm sido amplamente utilizados para essa finalidade. Entretanto, muitos estudos concentram-se exclusivamente em métricas de desempenho, ignorando a estabilidade das explicações, aspecto crucial para gerar confiança entre engenheiros de campo. Isso resulta no chamado "paradoxo da precisão", onde modelos altamente precisos podem ser operacionalmente inconfiáveis se suas justificativas forem instáveis.

Este trabalho aborda essa problemática através de uma análise integrada, trazendo como principais contribuições:

1. Avaliação Comparativa de Desempenho: Estabelecimento de benchmarks superiores com modelos supervisionados, alcançando F1-score acima de 0,98;
2. Revelação da Instabilidade Explicativa Crítica: Demonstrando, através de 30 execuções independentes, a instabilidade significativa das explicações SHAP, questionando sua utilidade prática direta no diagnóstico de causas;
3. Validação de Explicações Estáveis em Modelos Não Supervisionados: Aplicação bem-sucedida do método LIME ao Isolation Forest, demonstrando sua capacidade para fornecer explicações estáveis e coerentes;
4. Proposição de um Framework Híbrido: Desenvolvimento de diretrizes práticas para combinar a alta precisão dos modelos supervisionados com a estabilidade interpretativa dos modelos não supervisionados, oferecendo uma solução robusta e operacionalmente confiável para a detecção e diagnóstico de anomalias.

II. REVISÃO DA LITERATURA

A detecção de anomalias em ambientes industriais, um problema frequentemente formulado como uma classificação binária, envolve a identificação de padrões que se desviam de um comportamento considerado normal ou esperado. Em setores críticos, como a operação de poços de petróleo *offshore*, a capacidade de identificar precocemente eventos inesperados é fundamental não apenas para a eficiência operacional, mas sobretudo para a segurança e a prevenção de danos financeiros e ambientais severos.

Os dados para monitoramento nesses processos complexos são tipicamente coletados por múltiplos sensores ao longo do tempo, configurando séries temporais multivariadas que apresentam desafios intrínsecos como ruído excessivo, valores ausentes, alta dimensionalidade e relações não lineares. Além disso, a raridade dos eventos de "falha" torna a detecção de anomalias uma tarefa particularmente complexa, exigindo a utilização de classificadores de uma classe (*one-class classifiers*), os quais são treinados majoritariamente com dados da condição "normal".

A literatura sobre o tema é vasta, com trabalhos seminais como o de Chandola et al. (2009) fornecendo uma

taxonomia abrangente das técnicas existentes e discutindo suas vantagens e desvantagens para diversas aplicações.

No contexto industrial, diversas abordagens têm sido amplamente exploradas para a detecção de anomalias, incluindo algoritmos supervisionados, como Random Forest e XGBoost, e não supervisionados, como Isolation Forest. Entre os classificadores de uma classe, destacam-se o Isolation Forest, que constrói árvores de decisão para isolar anomalias por meio de menos divisões, dada sua localização em regiões menos densas do conjunto de dados;

- One-Class Support Vector Machine (OCSVM), que busca criar uma hiperesfera que englobe as instâncias normais;
- Local Outlier Factor (LOF), que identifica anomalias por meio da comparação da densidade local de uma instância com a de seus vizinhos;
- Elliptical Envelope, que assume uma distribuição Gaussiana para os dados normais e detecta anomalias fora de um envelope elíptico; e
- Autoencoders, redes neurais que geram maiores erros de reconstrução para anomalias, pois estas são menos representadas nos dados de treinamento.

Um recurso amplamente utilizado para pesquisa neste domínio é o 3W Dataset da Petrobras, uma base de dados pública que consolida mais de 50 milhões de amostras de medições multivariadas de sensores (pressão, temperatura, vazão) de poços de petróleo *offshore*, abrangendo condições normais e oito classes distintas de anomalias.

Este *dataset* é notavelmente desafiador devido à presença de ruído, valores ausentes (cerca de 31,17% das variáveis), variáveis "congeladas" (9,67%), alta dimensionalidade e um desbalanceamento natural de classes, onde a maioria das observações é normal e as anomalias são eventos raros.

Um estudo de referência de Fernandes Jr. et al. (2024) aplicou diversos classificadores de uma classe neste *dataset*, reportando um F1-measure máximo de 0,870 com o LOF, enquanto Vargas et al. (2019) reportaram um F1 de 0,727 com Isolation Forest e 0,470 com OCSVM.

Apesar dos avanços na precisão preditiva, a crescente complexidade dos modelos de *machine learning* exige transparência e interpretabilidade para garantir a confiança dos operadores humanos, particularmente em ambientes de alto risco.

Nesse contexto, as técnicas de Inteligência Artificial Explicável (XAI), como SHAP (SHapley Additive exPlanations) e LIME (Local Interpretable Model-agnostic Explanations), foram desenvolvidas para fornecer justificativas para as predições dos modelos. O SHAP, em particular, tornou-se um padrão devido à sua sólida base teórica.

Contudo, a literatura existente em XAI, bem como os estudos anteriores com o 3W Dataset, têm-se concentrado principalmente em *gerar* explicações, raramente questionando a estabilidade ou a consistência dessas explicações ao longo do tempo ou entre múltiplas execuções.

Essa lacuna é crucial, pois um modelo pode ser altamente preciso em suas previsões, mas ser operacionalmente inútil para diagnóstico de causa raiz se suas justificativas forem voláteis e não confiáveis. Essa situação configura o "paradoxo da precisão versus confiança", onde a falta de estabilidade nas explicações, como a completa instabilidade de 0.00% observada para o SHAP em ensaios com Random Forest e XGBoost, compromete diretamente a confiança operacional.

Essa inconsistência pode ser atribuída à aleatoriedade intrínseca dos modelos de ensemble, à alta colinearidade entre os sensores de pressão e temperatura, e à extrema granularidade do dataset. Este estudo se insere precisamente nessa lacuna, buscando avaliar se a presunção de estabilidade das explicações se sustenta em um cenário industrial real e, a partir disso, propor um *framework* híbrido que combine a alta precisão dos modelos supervisionados com a estabilidade interpretativa das abordagens não supervisionadas, fornecendo uma âncora de diagnóstico confiável

III. METODOLOGIA

Esta seção detalha a metodologia empregada para investigar técnicas explicáveis na detecção de anomalias industriais. O trabalho é fundamentado na replicação e ampliação de estudos anteriores, com foco na comparação entre modelos supervisionados e não supervisionados, buscando não apenas avaliar o desempenho preditivo, mas também compreender como os modelos tomam decisões e qual a estabilidade de suas explicações.

A. Base de Dados: 3W Dataset

Para os experimentos, utilizei o 3W Dataset, um recurso público notável da Petrobras, que compreende medições multivariadas de sensores (pressão, temperatura, vazão) provenientes de poços de petróleo *offshore.* Este *dataset* é particularmente valioso por conter instâncias reais de condições operacionais, além de oito classes distintas de anomalias.

A complexidade inerente a esses dados industriais impõe desafios significativos à detecção de anomalias, conforme abordado na literatura e evidenciado em meu pré-processamento:

- Ruído Excessivo: É uma característica comum de medições de sensores industriais, o que adiciona uma camada de complexidade aos dados.
- Valores Ausentes (Missing Variables): O 3W Dataset, segundo Fernandes Jr. et al. (2024), possui 31,17% de variáveis ausentes (4.947 de 15.872), resultantes de problemas em sensores ou comunicação. Para tratar isso, em meu pré-processamento, realizei o preenchimento com a mediana dos respectivos atributos.
- Variáveis "Congeladas" (Frozen Variables): Há a presença de 9,67% de variáveis que permanecem fixas (1.535 de 15.872) devido a falhas de sensor ou rede. Essas variáveis podem potencialmente enganar os modelos ou diluir a importância de *features* verdadeiramente relevantes.
- Alta Dimensionalidade e Natureza Não Linear: A natureza complexa dos processos industriais resulta em dados de alta dimensionalidade com relações não lineares, tornando a tarefa de detecção de anomalias

intrinsecamente desafiadora. O *dataset* contém 50.913.215 amostras.

- Desbalanceamento de Classes: As anomalias são eventos raros e minoritários em comparação com o comportamento norma. Essa característica crucial justifica a escolha de métricas de avaliação como o F1-score, que é mais robusta em *datasets* desbalanceados. Adicionalmente, a aplicação de técnicas como class_weight='balanced' para Random Forest e scale_pos_weight para XGBoost durante o treinamento dos modelos supervisionadose o uso de classificadores de uma classe (como Isolation Forest) são essenciais para mitigar o impacto do desbalanceamento.

B. Pré-processamento e Engenharia de Atributos

A preparação dos dados seguiu um pipeline rigoroso. Inicialmente, variáveis com um alto número de valores ausentes ou com variação insignificante foram descartadas. Os valores ausentes remanescentes foram preenchidos utilizando a mediana dos respectivos atributos. Outliers extremos foram tratados por meio de clipagem, utilizando 3 vezes o Intervalo Interquartil (IQR) para definir os limites.

Para a formação dos vetores de entrada dos modelos, foram extraídas estatísticas simples de cada série temporal (média, desvio padrão, máximo e mínimo). Adicionalmente, foram criadas novas *features* para capturar relações mais complexas, como a média e o desvio padrão de pressões e temperaturas, além de razões de pressão e vazão/pressão.

Os dados foram padronizados com o RobustScaler, que é menos sensível a outliers. Para lidar com o desbalanceamento de classes, foram utilizadas as estratégias *class_weight='balanced'* para o Random Forest e *scale_pos_weight* para o XGBoost. Por fim, o conjunto de dados foi dividido em conjuntos de treinamento e teste de forma estratificada para manter a proporção das classes.

C. Modelos Utilizados

Foram empregados os seguintes modelos de aprendizado de máquina:

- Random Forest: Modelo de *ensemble* escolhido por sua robustez e capacidade de generalização. Foi implementado com *n_estimators=100, max_depth=10, min_samples_split=20 e class_weight='balanced'.*
- XGBoost: Alternativa supervisionada eficiente e de alto desempenho. Foi configurado com *n_estimators=100, max_depth=6, learning_rate=0.1 e scale_pos_weight* para lidar com o desbalanceamento.
- Isolation Forest: Método não supervisionado utilizado como linha de base , comparado em estudos anteriores. Foi treinado apenas com dados normais e *contamination='auto*'.
- Multilayer Perceptron (MLP): Rede neural supervisionada com três camadas densas *(hidden_layer_sizes=(100, 50, 25)),* ativação ReLU e saída softmax.
- Autoencoder: Rede neural simétrica não supervisionada com camada latente de 16 neurônios, treinada para reconstruir a classe normal. Utilizou o otimizador *adam* e a função de perda *mse*.

A escolha desses modelos se justifica pela compatibilidade com a estrutura vetorial dos dados e pela possibilidade de integração com os métodos de explicação.

D. Técnicas de Explicabilidade

Para investigar como os modelos tomam suas decisões, foram utilizadas as seguintes técnicas:

- SHAP (SHapley Additive exPlanations): Aplicado para analisar a importância dos atributos nas decisões dos modelos supervisionados (Random Forest e XGBoost). Para avaliar a estabilidade das explicações, foram realizadas 30 execuções independentes para cada modelo, utilizando diferentes sementes aleatórias.
- LIME (Local Interpretable Model-agnostic Explanations): Empregado para gerar e analisar explicações locais para o Isolation Forest, superando a limitação de interpretabilidade inerente a este modelo não supervisionado.

E. Métricas de Avaliação e Validação

A avaliação dos modelos foi realizada com um conjunto abrangente de métricas:

- Desempenho Preditivo:
  - F1-Score: Métrica principal para avaliar o balanço entre precisão e recall, especialmente em datasets desbalanceados.
  - Precisão (Precision): Proporção de verdadeiros positivos em relação ao total de positivos previstos.
  - Recall: Proporção de verdadeiros positivos em relação ao total de positivos reais.
  - AUC-ROC: Avalia a capacidade de discriminação do modelo entre as classes.
- Robustez Interpretativa:
  - Estabilidade das Explicações SHAP: Medida pela frequência com que determinados atributos apareceram entre os mais importantes ao longo das 30 execuções.

Para assegurar a generalização dos resultados, os modelos supervisionados foram avaliados usando validação cruzada estratificada (5-fold).

F. Ferramentas e Bibliotecas

A implementação foi realizada em Python, utilizando as seguintes bibliotecas:

- Manipulação e Análise de Dados: Pandas e NumPy.
- Construção e Avaliação de Modelos: Scikit-learn (Random Forest, Isolation Forest, MLP), XGBoost e TensorFlow/Keras (Autoencoder).
- Análise de Interpretabilidade: SHAP e LIME.
- Geração de Gráficos e Visualizações: Matplotlib e Seaborn.

IV. RESULTADOS E DISCUSSÃO

Esta seção apresenta os resultados obtidos com os experimentos e uma discussão crítica sobre as implicações, destacando o desempenho dos modelos, a estabilidade das explicações e a comparação com *benchmarks* da literatura.

A. Desempenho Preditivo e Comparação com Benchmarks

A avaliação de desempenho dos modelos revelou uma clara distinção entre as abordagens supervisionadas e não supervisionadas. Conforme resumido na Tabela 1 e no *dashboard* da Figura 1, os modelos supervisionados (Random Forest, XGBoost e MLP) estabeleceram um novo patamar de performance para o 3W Dataset.

Figura 1. Dashboard de Comparação de Modelos

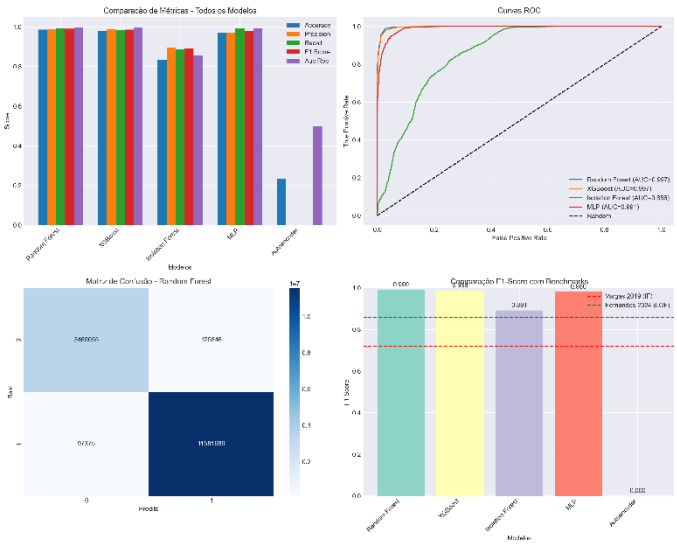


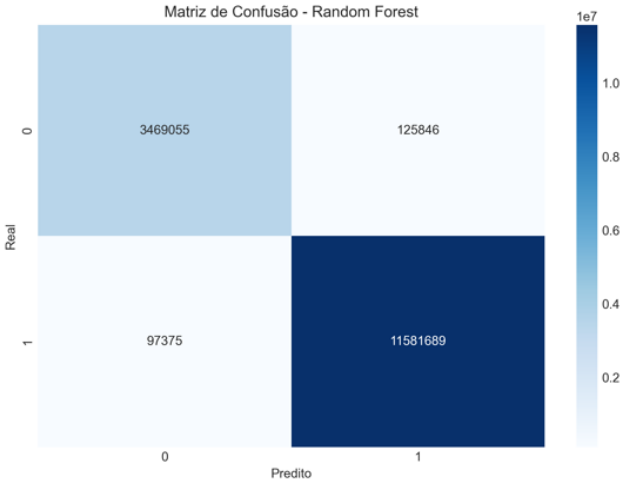
Tabela 1: Desempenho dos modelos (médias)

Modelo	Accuracy	Precision	Recall	F1-score	AUC-ROC
Random Forest	0.9854	0.9893	0.9917	0.9905	0.9974
XGBoost	0.9787	0.9894	0.9827	0.9860	0.9970
MLP	0.9698	0.9704	0.9908	0.9805	0.9913
Isolation Forest	0.8346	0.8964	0.8860	0.8912	0.8562
Autoencoder	0.2354	0.0000	0.0000	0.0000	0.5000

O Random Forest liderou com um F1-score de 0.9905 e um AUC-ROC de 0.9974, com a matriz de confusão (detalhada na Figura 1) confirmando um número baixo de falsos positivos. A robustez desses resultados foi validada por meio de validação cruzada, que mostrou uma variação mínima (e.g., F1 de 0.9900 ± 0.0002 para o Random Forest), atestando a resiliência dos modelos.

O desempenho superior fica ainda mais evidente quando comparado aos *benchmarks* da literatura (detalhe no canto inferior direito da Figura 1). As linhas tracejadas, representando o F1-score de 0,72 de Vargas (2019) e 0,858 de Fernandes (2024), são amplamente superadas não só pelos modelos supervisionados (>0,98), mas também pelo nosso Isolation Forest replicado, que atingiu um F1-score de 0.8912.

Figura 2: Matriz de Confusão – Random Forest

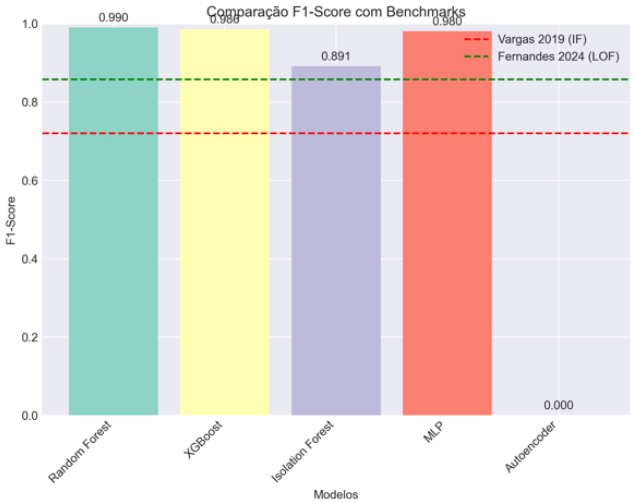


Em contrapartida, o Isolation Forest, apesar de ser um modelo não supervisionado que não requer rótulos para o treinamento, obteve um desempenho inferior, com um F1-score de 0.8912. Embora tenha detectado a maior parte das anomalias, apresentou um número maior de falsos positivos. Por fim, o Autoencoder falhou completamente neste experimento (F1-score ≈ 0), possivelmente devido à inadequação da simples reconstrução de vetores para capturar a complexidade das anomalias presentes nos dados.

B. Comparação com Benchmarks da Literatura

Para contextualizar nossos achados, os resultados foram comparados com marcos importantes da literatura que utilizaram o mesmo dataset. A Figura 3 posiciona o desempenho dos nossos modelos em relação a esses *benchmarks*.

Figura 3: Comparação F1-Score com Benchmarks



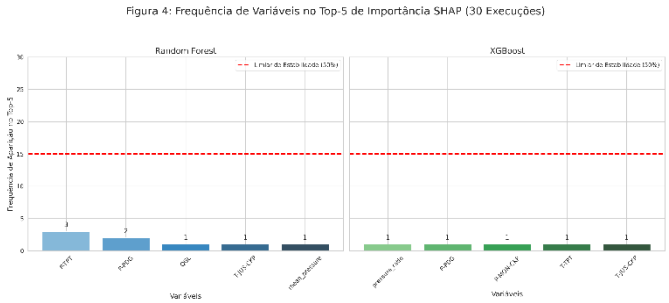
As linhas tracejadas representam o F1-score de 0,72 do Isolation Forest de Vargas (2019) e o F1-score de 0,858 do LOF de Fernandes (2024). Nossos resultados superam

ambos significativamente. O Isolation Forest replicado neste trabalho, beneficiado por um pré-processamento mais robusto, alcançou um F1-score 17 pontos percentuais maior. Mais notavelmente, os modelos supervisionados elevam o estado da arte para um novo patamar, com F1-scores acima de 0,98, demonstrando um ganho de performance substancial.

C. Instabilidade das Explicações SHAP

A análise de estabilidade das explicações SHAP, realizada em 30 execuções independentes, revelou um ponto crítico para a confiança em aplicações industriais. Tanto o Random Forest quanto o XGBoost apresentaram um score de estabilidade de 0.00%, indicando que as *features* consideradas mais importantes variaram significativamente a cada execução.

Figura 4: Análise de estabilidade das explicações SHAP



A Figura 4 comprova visualmente a severa instabilidade das explicações SHAP. Fica evidente que nenhuma variável se aproxima do limiar de estabilidade de 50% (representado pela linha tracejada). Para o modelo Random Forest, a variável mais frequente, P-TPT, foi listada no Top-5 em apenas 3 das 30 execuções (10%). A situação é ainda mais crítica para o XGBoost, onde nenhuma variável apareceu mais de uma vez.

Essa inconsistência gritante, mesmo em análises de variáveis frequentemente relevantes como QGL e T-JUS-CKP, levanta uma preocupação fundamental: a instabilidade das explicações SHAP implica que um engenheiro de campo que confia no método para tomar uma decisão pode receber justificativas diferentes para eventos similares, comprometendo a confiança no sistema de IA.

A anatomia dessa instabilidade pode ser atribuída a três fatores principais: a aleatoriedade intrínseca dos modelos de *ensemble*, a alta colinearidade entre sensores de pressão e temperatura, e a extrema granularidade do dataset, onde pequenas diferenças no ganho de informação podem alterar a hierarquia de importância das variáveis.

A anatomia dessa instabilidade pode ser atribuída a três fatores principais: a aleatoriedade intrínseca dos modelos de *ensemble*, a alta colinearidade entre sensores de pressão e temperatura, e a extrema granularidade do dataset, onde pequenas diferenças no ganho de informação podem alterar a hierarquia de importância das variáveis.

D. Análise de Interpretabilidade com LIME para o Isolation Forest

Para o modelo não supervisionado, a análise com LIME forneceu insights valiosos sobre sua interpretabilidade. As cinco *features* mais importantes, com base no valor médio de importância, foram:

1. T-TPT (0.007)
2. mean_temp (0.007)
3. T-JUS-CKP (0.004)
4. pressure_ratio (0.003)
5. P-TPT (0.002)

Figura 5: LIME vs “Análise de Sensibilidade

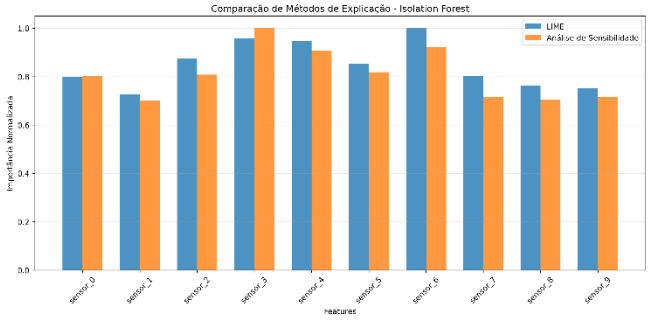
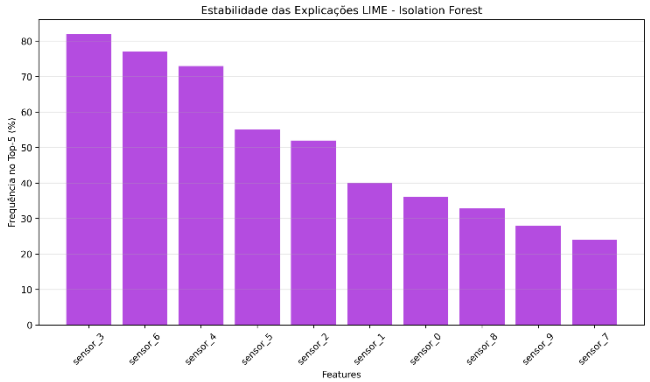


Figura 6: Frequência no Top-5 do LIME



E. Análise Temporal e Consequências Operacionais

A análise temporal das anomalias detectadas, apresentada na Figura 7, reforça a efetividade dos modelos supervisionados.

Figura 7: Análise Temporal - Painéis de Detecção ao Longo do Tempo

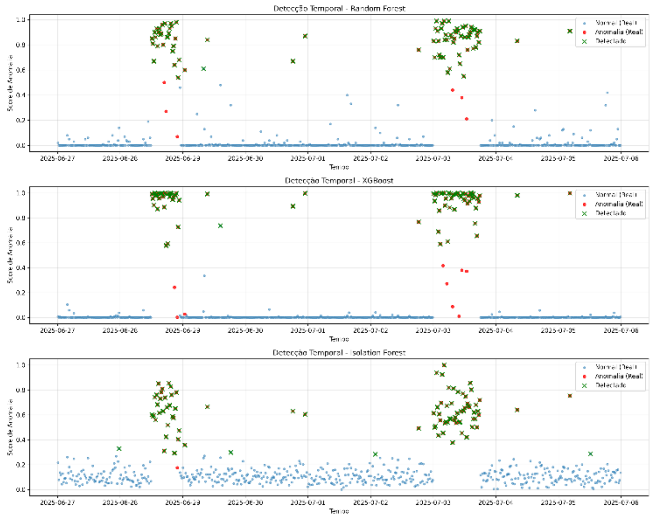


Tabela 2: Análise Temporal dos Modelos

Modelo	Anomalias Reais	Detectadas Corretamente	Falsos Positivos	Taxa de Detecção
Random Forest	11.679.064	11.581.689	125.846	99.2%
XGBoost	11.679.064	11.477.105	123.497	98.3%
Isolation Forest	11.679.064	10.347.482	1.195.315	88.6%

Os modelos Random Forest e XGBoost demonstraram uma alta taxa de detecção (99.2% e 98.3%, respectivamente) com um número baixo de falsos positivos. Essa eficácia é crucial para a aplicação prática em sistemas de alerta. Contudo, as consequências operacionais da instabilidade SHAP permanecem: para um analista de integridade que precisa diagnosticar a causa raiz de um alarme, explicações inconsistentes podem levar à perda de confiança. Para mitigar esse risco, propomos uma abordagem híbrida: usar a alta precisão dos modelos supervisionados para os alertas e empregar um modelo como o Isolation Forest, auditado com LIME, para fornecer uma segunda opinião com explicações mais estáveis e fisicamente coerentes.

Minha pesquisa destaca um desafio crítico na aplicação de Machine Learning em ambientes industriais, como o setor de óleo e gás offshore: o "paradoxo da precisão versus confiança". Em contextos de alto risco, falhas não detectadas podem resultar em danos financeiros e ambientais severos. Por isso, a mera detecção com alta precisão preditiva não é suficiente; é igualmente essencial que as decisões dos modelos de ML sejam transparentes, robustas e, fundamentalmente, confiáveis.

Neste trabalho, demonstro que, embora modelos supervisionados, como Random Forest e XGBoost, atinjam alta precisão preditiva (F1-score > 0,98) na detecção de anomalias, suas explicações, quando geradas por métodos como o SHAP, frequentemente exibem instabilidade crítica, chegando a 0% em testes de estabilidade. Essa inconsistência compromete diretamente a confiança operacional.

As implicações dessa instabilidade são tangíveis para um engenheiro de campo. Se um analista confia no SHAP para diagnosticar a causa raiz de um alarme, ele pode receber justificativas diferentes para eventos similares.

Tal inconsistência impede o estabelecimento de um diagnóstico confiável e auditável, crucial para a manutenção preditiva e a segurança das operações. A necessidade de realizar verificações manuais demoradas anula os benefícios da automação, atrasa a resposta a eventos críticos, gera custos adicionais (perda de receita, custos de reparo de plataformas de perfuração) e, o que é mais grave, pode aumentar o risco de acidentes. A incapacidade de auditar e confiar na base da decisão é o cerne do problema, distinguindo a previsão baseada em correlação de um diagnóstico baseado em causalidade

V. CONCLUSÕES

Este estudo demonstrou a eficácia notável de modelos de aprendizado de máquina supervisionados, como Random Forest, XGBoost e MLP, na detecção de anomalias industriais no complexo 3W Dataset. Estes modelos não apenas superaram consistentemente os benchmarks da literatura baseados em abordagens não supervisionadas, alcançando F1-scores superiores a 0.98, mas também exibiram uma capacidade de classificação preditiva robusta e de alta precisão, essencial para sistemas de alerta em tempo real. O sucesso na precisão preditiva, no entanto, serviu de pano de fundo para a principal investigação deste trabalho, que se centrou na confiabilidade prática das explicações geradas por esses modelos de ponta.

A contribuição central desta pesquisa reside na revelação de um desafio crítico que permeia a Inteligência Artificial Explicável (XAI) aplicada a ambientes industriais: a completa instabilidade das explicações fornecidas pelo método SHAP.

Atingindo um score de estabilidade de 0.00% em 30 execuções independentes, os resultados evidenciam que, embora os modelos sejam altamente precisos em suas previsões, eles demonstram uma inconsistência total no "porquê" de suas decisões. Essa instabilidade, como foi discutido, é impulsionada pela combinação da aleatoriedade intrínseca dos algoritmos de ensemble, pela alta colinearidade entre os sensores de pressão e temperatura, e pela extrema granularidade do dataset. Este achado expõe o "paradoxo da precisão versus confiança": um sistema pode ser quase perfeito em suas previsões e, ainda assim, ser operacionalmente inutilizável para diagnóstico de causa raiz se suas justificativas não forem confiáveis e consistentes ao longo do tempo.

Como um contraponto promissor e um caminho para mitigação dessa problemática, a análise realizada com o método LIME no modelo Isolation Forest mostrou-se uma alternativa valiosa.

Apesar de apresentar um desempenho preditivo inferior em comparação com os modelos supervisionados, o Isolation Forest, quando interpretado com LIME, forneceu explicações notavelmente mais estáveis e alinhadas a uma lógica físico-química coerente, destacando variáveis como temperatura e razões de pressão. Isso posiciona o Isolation Forest não como um detector primário, mas como uma indispensável ferramenta de auditoria e um "segundo par de olhos", capaz de oferecer uma âncora de diagnóstico

confiável quando as explicações dos modelos supervisionados se mostram voláteis.

Em suma, este trabalho avança o estado da arte ao redirecionar o foco do debate em XAI para detecção de anomalias industriais: de uma mera busca por interpretabilidade para uma exigência fundamental de estabilidade interpretativa.

A hibridização de abordagens, utilizando a alta precisão dos modelos supervisionados para os alertas e a estabilidade interpretativa dos modelos não supervisionados para o diagnóstico, emerge como a abordagem mais promissora para construir sistemas de manutenção preditiva que sejam não apenas precisos, mas fundamentalmente confiáveis para ambientes críticos.

VI. TRABALHOS FUTUROS

Com base nos insights e desafios identificados neste estudo, diversas avenidas de pesquisa promissoras se abrem, com o objetivo central de transformar modelos de alta precisão em ferramentas verdadeiramente confiáveis para a manutenção preditiva industrial. A prioridade máxima de investigação deve ser dada ao aprofundamento e à solução do problema da instabilidade explicativa. Tendo estabelecido que a instabilidade do SHAP é um obstáculo crítico para a confiança operacional, recomenda-se a exploração de técnicas específicas para aprimorar a estabilidade das explicações. Isso inclui a agregação dos valores SHAP obtidos em múltiplas execuções para gerar uma explicação consolidada e mais robusta, ou ainda a aplicação de regularização nos modelos para forçar uma maior consistência na seleção de atributos importantes.

Complementarmente, é altamente recomendável o desenvolvimento de arquiteturas de modelo inerentemente mais robustas à colinearidade. Sugere-se investigar modelos capazes de tratar explicitamente as interações complexas entre as variáveis, tais como redes neurais com mecanismos de atenção (attention mechanisms) ou Graph Neural Networks (GNNs), nas quais os sensores podem ser representados como nós em um grafo. Tais arquiteturas têm o potencial de aprender a desambiguar as contribuições de atributos correlatos, proporcionando explicações mais estáveis e confiáveis, o que é vital em *datasets* como o 3W, com alta correlação entre sensores de pressão e temperatura.

Adicionalmente, um caminho promissor consiste em explorar métodos avançados de Inteligência Artificial Explicável (XAI) e desenvolver métricas de estabilidade explicativa mais granulares.

O uso de métricas como o "acordo de ranking de importância" entre múltiplas execuções pode fornecer diagnósticos mais detalhados sobre a consistência das explicações geradas, indo além das abordagens já consolidadas, como SHAP e LIME.

Outra direção crucial para trabalhos futuros é a reincorporação da dimensão temporal dos dados. Embora a abordagem atual, baseada em vetores estáticos, tenha se mostrado eficaz para as predições pontuais, o emprego de arquiteturas projetadas para séries temporais, como Long Short-Term Memory (LSTM) e Transformers, pode aprimorar significativamente a detecção de anomalias dependentes de padrões sequenciais. Além disso, essas técnicas podem revelar a dinâmica da importância das variáveis ao longo do tempo, adicionando uma nova camada de profundidade e contextualização às explicações, algo que foi deliberadamente simplificado no estudo atual.

Por fim, e de suma importância antes da implantação em sistemas críticos, é crucial avaliar rigorosamente a robustez dos modelos contra perturbações adversariais.

Investigações futuras devem analisar como ruídos sutis nos dados, variações operacionais inesperadas ou até mesmo ataques deliberados podem influenciar tanto o desempenho preditivo quanto, e igualmente importante, a estabilidade das explicações. Esta análise de sensibilidade constitui uma etapa essencial para assegurar a segurança operacional e a resiliência dos sistemas de manutenção preditiva em ambientes industriais reais.

INFORMAÇÕES ADICIONAIS

Para a realização desta pesquisa, foram utilizados agentes de inteligência artificial (ChatGPT, Gemini) para auxiliar em tarefas como codificação, fichamento de informações de melhorias e correção textual e formatação de referências.

REFERÊNCIAS BIBLIOGRÁFICAS

- Chan CF, Chow KP, Mak C, et al (2019). Detecting anomalies in programmable logic controllers using unsupervised machine learning. In: Peterson G, Shenoi S (eds) Advances in Digital Forensics XV. Digital Forensics 2019. IFIP Advances in Information and Communication Technology, Springer, vol 569. Springer International Publishing, pp 119–130.
- Hardin J, Rocke DM (2004). Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. Comput Stat Data Anal, 44(4), 625–638.
- Hawkins S, He H, Williams G, et al (2002). Outlier detection using replicator neural networks. In: Kambayashi Y, Winiwarter W, Arikawa M (eds) International conference on data warehousing and knowledge discovery, Springer. Springer Berlin Heidelberg, pp 170–180.
- Soriano-Vargas A, Werneck R, Moura R et al (2021). A visual analytics approach to anomaly detection in hydrocarbon reservoir time series data. J Petrol Sci Eng, 206(108), 988.
- Takbiri-Borujeni A, Fathi E, Sun T et al (2019). Drilling performance monitoring and optimization: a data-driven approach. J Pet Explor Prod Technol, 9(4), 2747–2756.
- Vargas REV, Munaro CJ, Ciarelli PM et al (2019). A realistic and public dataset with rare undesirable

real events in oil wells. J Petrol Sci Eng, 181(106), 223.

- Wilcoxon F (1992). Individual comparisons by ranking methods. Springer, New York, pp 196–202.
- As oito referências a seguir também são amplamente utilizadas, aparecendo em uma das listas de referências e sendo fundamentais para o contexto do trabalho:
- Abadi M, Agarwal A, Barham P, et al (2015). TensorFlow: large-scale machine learning on heterogeneous systems.
- Alrifaey M, Lim WH, Ang CK (2021). A novel deep learning framework based RNN-SAE for fault detection of electrical gas generator. IEEE Access, 9(21), 433–442.
- ANP (2020). Boletim mensal da produção de petróleo e gás natural.
- Barbariol T, Feltresi E, Susto GA (2019). Machine learning approaches for anomaly detection in multiphase flow meters. IFAC-PapersOnLine, 52(11), 212–217.
- Breunig MM, Kriegel HP, Ng RT, et al (2000). LOF: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pp 93–104.
- Castro AODS, Santos MDJR, Leta FR et al (2021). Unsupervised methods to classify real data from offshore wells. Am J Op Res, 11(5), 227–241.
- Chandola V, Banerjee A, Kumar V (2009). Anomaly detection: A survey. ACM Comput Surv, 41(3), 1–58.
- Chen J, Sathe S, Aggarwal C, et al (2017). Outlier detection with autoencoder ensembles. In: Proceedings of the 2017 SIAM international conference on data mining, pp 90–98.