INTELIGÊNCIA ARTIFICIAL BASEADA EM RANDOM FOREST PARA A CLASSIFICAÇÃO DE ANOMALIAS EM UM POÇO DE PETRÓLEO

NATÁSSIA RAFAELLE MEDEIROS SIQUEIRA[1]; RONEY MENEZES MEIRELLES JÚNIOR[2]; SOFIA PAGLIARINI[2]; JOÃO INÁCIO MOREIRA BEZERRA[2]; OSCAR MAURICIO HERNANDEZ RODRIGUEZ[3]; MARLON MAURÍCIO HERNANDEZ CELY[2]

[1]*Universidade Federal de Pelotas – natassiamsads@gmail.com*
[2]*Universidade Federal de Pelotas – rjmeirelles999@gmail.com*
[2]*Universidade Federal de Pelotas – sofiapagliarini@gmail.com*
[2]*Universidade Federal de Pelotas – jimbezerra@inf.ufpel.edu.br*
[3] *Escola de Engenharia de São Carlos – oscarmhr@sc.usp.br*
[2]*Universidade Federal de Pelotas – marlon.cely@ufpel.edu.br*

1. INTRODUÇÃO

Os campos de identificação e predição de falhas estão em rápida evolução, demandando metodologia e algoritmos computacionais, baseados em aprendizado de máquina precisas e eficientes para essas tarefas. A literatura mostra exemplos de aplicação em fluxos bifásicos (JIANG, 2024), sistemas mecânicos(SHCHERBAKOV, 2022) e em manufatura (YANG, 2020).

Tradicionalmente, a identificação de falhas é realizada de maneira subjetiva, com a identificação humana de acordo com observações de câmeras e sensores. Estes métodos são baratos, porém com baixa confiabilidade, de forma que técnicas de aprendizado de máquina vêm ganhando destaque nos últimos anos, em virtude de sua capacidade de extrair características dos dados e de sua escalabilidade em lidar com dados complexos (TANG, 2022).

Em escoamentos bifásicos, em especial aqueles que envolvem óleo e gás, um desafio considerável é a baixa disponibilidade de dados abertos, o que prejudica o desenvolvimento de métodos de inteligência artificial para prever e classificar anomalias. A principal razão é a confidencialidade de informações de alto custo, enquanto outra é a dificuldade em reconhecer e rotular todos os possíveis eventos improváveis a partir dos dados disponíveis. Este paradigma foi quebrado no trabalho de Vargas (2019), em que foi apresentada uma base de dados pública com múltiplos tipos de anomalias, denominada 3W, baseada em séries temporais, com o objetivo de fomentar o desenvolvimento de técnicas de aprendizado de máquina para identificar essas anomalias.

Neste contexto, neste trabalho aplica-se um algoritmo *Random Forest* para a detecção de anomalias em um poço de petróleo, de acordo com a base de dados 3W. Este algoritmo é composto por um conjunto de árvores de decisão, de acordo com as variáveis da base de dados, e cada uma das árvores faz uma classificação. Na saída, a classificação mais comum realizada pelas árvores é tratada como a classificação do *Random Forest.* O algoritmo demonstrou uma acurácia de praticamente 100% na detecção de anomalias no poço de petróleo específico.

2. METODOLOGIA

Em Vargas (2019), foi apresentada uma base de dados pública denominada 3W, que contém oito tipos de anomalias encontradas em poços de petróleo. Os arquivos que compõem a base de dados são formados por séries temporais, podendo ser extraídas de sensores em poços de petróleo ou ainda simuladas, e estão divididos em pastas da seguinte forma, de acordo com o comportamento (normal ou falha), e o tipo de falha.

- Pasta 0: Comportamento Normal.
- Pasta 1: Aumento abrupto de BSW.
- Pasta 2: Fechamento espúrio de DHSV.
- Pasta 3: Pistoneamento severo.
- Pasta 4: Instabilidade de fluxo.
- Pasta 5: Perda rápida de produtividade.
- Pasta 6: Restrição rápida em CKP.
- Pasta 7: Incrustação em CKP.
- Pasta 8: Hidrato em linha de produção.

Estes eventos estão divididos ao longo de 18 poços, e foi selecionado o poço 1 para este trabalho, por possuir vários tipos de anomalias, sendo composto pelas seguintes instâncias:

- 92 instâncias de comportamento normal, representado pela classe 0.
- 1 instância de aumento abrupto de BSW, representado pela classe 1.
- 1 instância de pistoneamento severo, representado pela classe 3.
- 32 instâncias de instabilidade de fluxo, representado pela classe 4.

Cada uma destas instâncias é composta por séries temporais, de tamanho variável, podendo representar um período de algumas horas até alguns dias. As observações são obtidas a cada segundo, e são compostas por oito variáveis, extraídas de sensores posicionados em locais variáveis dentro do poço de petróleo. Para reduzir a dimensão da base de dados neste trabalho trabalhou-se com observações a cada minuto, obtendo a média de cada uma das 60 observações de cada variável durante um minuto.

No poço 1, três variáveis apresentam valores nulos, dessa forma são desconsideradas neste experimento, e as variáveis que possuem valores não nulos são as seguintes:

- P-TPT: pressão no transdutor de temperatura e pressão.
- T-TPT: temperatura no transdutor de temperatura e pressão.
- P-MON-CKP: pressão de subida do estrangulador de produção.
- T-JUS-CKP: temperatura de descida do estrangulador de produção.
- P-JUS-CKGL: pressão de descida do elevador de gás.

Após a divisão das observações por minuto, considerando as observações com valor não nulo, foi realizada a normalização dos dados, de acordo com a distribuição normal. Esta normalização se dá de forma a evitar que as diferentes escalas das diferentes variáveis confundam o algoritmo de aprendizado de máquina na hora da classificação. Por fim, os dados normalizados foram divididos para treinamento e teste, da seguinte forma: 80% para treinamento e 20% para teste. O algoritmo Random Forest usou estes dados para realizar a classificação.

3. RESULTADOS E DISCUSSÃO

Na Figura 1, apresenta-se a matriz de confusão para a classificação das observações neste trabalho. A matriz mostra que o classificador proposto atinge uma acurácia de quase 100%, mesmo com o conjunto de dados sendo

desbalanceado, ou seja, certos comportamentos possuindo mais observações do que outros comportamentos, da seguinte forma:

- Acurácia de 100% na classificação de dados normais, com as 5753 observações na base de dados de teste sendo classificadas de maneira correta.
- Acurácia de 100% na classificação de dados que representam aumento abrupto de BSW, com as 69 observações na base de dados de teste sendo classificadas de maneira correta.
- Acurácia de 100% na classificação de dados que representam pistoneamento severo, com as 49 observações na base de dados de teste sendo classificadas de maneira correta.
- Acurácia de 99.89% na classificação de dados que representam instabilidade de fluxo, com 870 observações sendo classificadas corretamente e 1 das observações sendo classificada como comportamento normal.

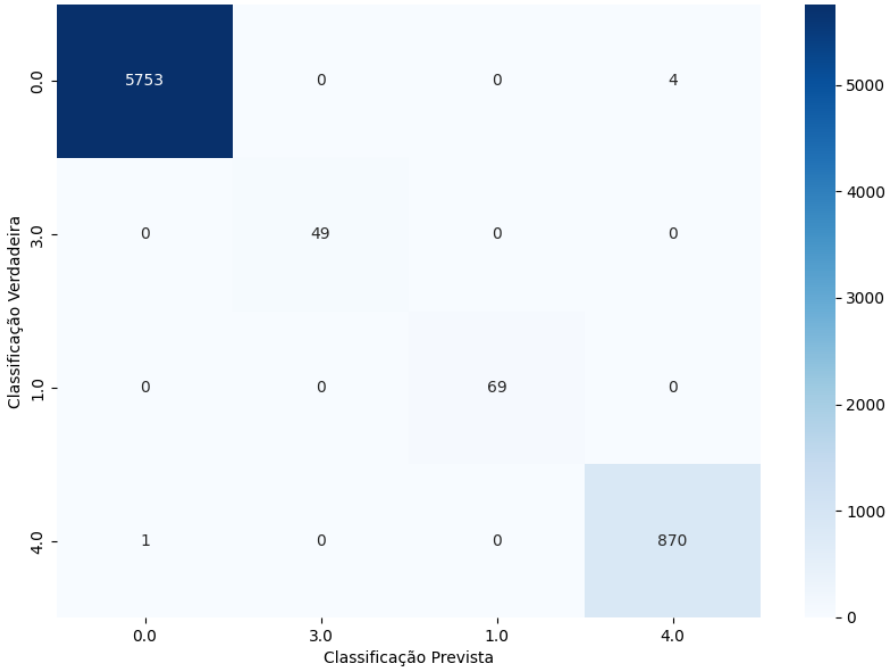

Figura 1: Matriz de confusão para a classificação dos dados que representam comportamentos normal ou falha.

4. CONCLUSÕES

Este trabalho mostrou que o algoritmo *Random Forest* é eficaz em classificar observações que representam comportamento normal ou comportamento anômalo em poços de petróleo, pela sua acurácia de praticamente 100%. Em trabalhos futuros, o objetivo é não apenas classificar a anomalia após ela ter acontecido, mas conseguir prever essa anomalia a partir de observações transientes, de forma que a anomalia possa ser controlada antes de que prejuízos sejam causados à

produção de petróleo. Isso se dará com técnicas que envolvem o processamento de linguagem natural, como por exemplo *transformers.*

AGRADECIMENTOS

Agradeço ao meu orientador, Marlon Maurício Hernandez Cely, pela orientação essencial e pelas contribuições valiosas durante o desenvolvimento deste trabalho. Sou grato aos colegas de pesquisa, Jairo, Letícia, Sofia, Roney, João Inácio, Eden, Carlos e Oscar pela colaboração e discussões produtivas que enriqueceram esta investigação. Agradeço também à Petrobras, à Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao Laboratório de Escoamentos Multifásicos Industriais (LEMI-USP) pelo suporte financeiro e recursos técnicos.

5. REFERÊNCIAS BIBLIOGRÁFICAS

JIANG, Yuxiao et al. A Flowrate Estimation Method for Gas-Water Two-Phase Flow Based on Multimodal Sensors and Hybrid LSTM-CNN Model. **IEEE Transactions on Instrumentation and Measurement**, 2024.

SHCHERBAKOV, Maxim; SAI, Cuong. A hybrid deep learning framework for intelligent predictive maintenance of cyber-physical systems. **ACM Transactions on Cyber-Physical Systems (TCPS)**, v. 6, n. 2, p. 1-22, 2022.

TANG, Shengnan; ZHU, Yong; YUAN, Shouqi. Intelligent fault identification of hydraulic pump using deep adaptive normalized CNN and synchrosqueezed wavelet transform. **Reliability Engineering & System Safety**, v. 224, p. 108560, 2022.

YANG, Jing et al. Using deep learning to detect defects in manufacturing: a comprehensive survey and current challenges. **Materials**, v. 13, n. 24, p. 5755, 2020.

VARGAS, Ricardo Emanuel Vaz et al. A realistic and public dataset with rare undesirable real events in oil wells. **Journal of Petroleum Science and Engineering**, v. 181, p. 106223, 2019.