Análise Comparativa de Técnicas de Aumento de Dados na Classificação de Escoamento Multifásico

Natássia Siqueira[1], **João Bezerra**[2], **Sofia Pagliarini**[2], **Marlon Cely**[2], **Daniel Palomino**[2]

[1] Programa de Pós-Graduação em Computação – Universidade Federal de Pelotas (UFPel)
Caixa Postal 15.064 – 91.501-970 – Pelotas – RS – Brazil

nrmsiqueira@inf.ufpel.br , Universidade Federal de Pelotas – UFPel
Praça Domingos Rodrigues, 96010-450 – Pelotas – RS – Brazil

jimbezerra@inf.ufpel.edu.br, sofiapagliarini@gmail.com

marlon.cely@ufpel.edu.br, dpalomino@inf.ufpel.edu.br

Resumo. *A aplicação de aprendizado de máquina em sistemas industriais de óleo e gás é frequentemente limitada pela escassez de dados rotulados e pelo desbalanceamento entre classes. Este trabalho investiga o impacto de técnicas de aumento de dados na classificação de regimes de escoamento multifásico utilizando dois conjuntos de dados experimentais. Foram avaliadas técnicas tradicionais de oversampling (SMOTE e ADASYN) e métodos generativos, incluindo Gaussian Copula, VAE e TVAE. Os experimentos foram conduzidos com Random Forest, XGBoost e LightGBM, utilizando validação cruzada estratificada e análise estatística baseada nos testes de Friedman e Nemenyi. Os resultados mostram que métodos generativos apresentaram melhor desempenho médio, com destaque para o VAE, que obteve o menor rank médio (1.17) e melhoria de até 0.03 no F1-score em relação ao cenário sem aumento de dados. Os resultados indicam que estratégias generativas podem melhorar a identificação de classes minoritárias em dados industriais desbalanceados.*
Palavras-chave*: aprendizado de máquina; data augmentation; escoamento multifásico; desbalanceamento de classes.*

Abstract. *Machine learning applications in oil and gas industrial systems are often limited by scarce labeled data and class imbalance. This work investigates the impact of data augmentation strategies on multiphase flow pattern classification using two experimental datasets. Traditional oversampling techniques (SMOTE and ADASYN) and generative methods, including Gaussian Copula, VAE and TVAE, were evaluated. Experiments were conducted using Random Forest, XGBoost and LightGBM, with stratified cross-validation and statistical analysis based on Friedman and Nemenyi tests. Results show that generative methods achieved the best average performance, with VAE obtaining the lowest average rank (1.17) and improvements of up to 0.03 in F1-score compared to the baseline without augmentation. These findings indicate that generative augmentation strategies can improve minority class recognition in imbalanced industrial datasets.*

1. INTRODUÇÃO

A aplicação de Inteligência Artificial (IA) na indústria de óleo e gás é frequentemente limitada pela escassez de dados rotulados e pelo desbalanceamento entre classes. Em sistemas de escoamento multifásico, regimes importantes como o regime de golfadas (*slug flow*) são frequentemente sub-representados devido ao elevado custo experimental e às restrições operacionais envolvidas na coleta de dados [Santos et al. 2024]. Como consequência, modelos de aprendizado de máquina aplicados ao monitoramento industrial podem apresentar baixa robustez e dificuldades de generalização [Kumar et al. 2024].

Técnicas de aumento de dados (AU) têm sido amplamente utilizadas para mitigar problemas de desbalanceamento em tarefas de classificação, permitindo ampliar a representação de classes minoritárias por meio da geração de amostras sintéticas [Dablain and Chawla 2024]. Entretanto, em dados de escoamento multifásico, as amostras geradas devem preservar relações físicas fundamentais do sistema [Borges et al. 2024]. Além disso, ainda existe incerteza sobre quais estratégias realmente proporcionam melhorias consistentes de desempenho, especialmente em dados tabulares industriais [Khan et al. 2025]. Neste trabalho investigamos o impacto de diferentes estratégias de aumento de dados na classificação de regimes de escoamento multifásico utilizando conjuntos de dados tabulares industriais. Foram avaliadas técnicas tradicionais de oversampling e modelos generativos, incluindo *Tabular Variational Autoencoders* (TVAE) [Zhang and Li 2023] e *Gaussian Copula* [Patel et al. 2024], utilizando métricas tradicionais de classificação e testes estatísticos de desempenho.

As principais contribuições deste trabalho incluem: (i) avaliação comparativa entre técnicas tradicionais e generativas de aumento de dados; (ii) proposição de um protocolo experimental reprodutível para cenários desbalanceados; e (iii) análise quantitativa do impacto das técnicas de AU no desempenho dos classificadores. Os resultados demonstram que estratégias generativas de aumento de dados podem contribuir para melhorar a robustez de modelos aplicados ao monitoramento de escoamento multifásico.

2. VISÃO GERAL DE DATA AUGMENTATION

Técnicas de *data augmentation* têm sido amplamente utilizadas para mitigar problemas de desbalanceamento e escassez de dados em tarefas de aprendizado de máquina, permitindo aumentar artificialmente a representatividade de classes minoritárias e melhorar a capacidade de generalização dos modelos [Mumuni and Mumuni 2022].

No contexto de dados tabulares industriais, abordagens tradicionais de oversampling, como SMOTE e ADASYN, geram novas amostras por interpolação entre instâncias existentes. Entretanto, esses métodos podem apresentar limitações em cenários com relações não lineares complexas e distribuições multimodais [A. Fernández and Herrera 2018]. Mais recentemente, modelos generativos baseados em aprendizado profundo, como Variational Autoencoders (VAE) e Tabular Variational Autoencoders (TVAE), passaram a ser explorados para geração de dados sintéticos mais consistentes do ponto de vista estatístico [L. Xu and Veeramachaneni 2019]. Além disso, métodos probabilísticos baseados em cópulas gaussianas têm sido utilizados para modelar dependências entre variáveis e gerar amostras sintéticas preservando características da distribuição original. Neste trabalho, foram avaliadas técnicas tradicionais e generativas

de aumento de dados em conjuntos tabulares de escoamento multifásico, considerando seus impactos no desempenho e na robustez de classificadores supervisionados.

3. METODOLOGIA

Esta seção descreve a metodologia empregada para investigar o impacto de técnicas de aumento de dados na classificação de regimes de escoamento multifásico utilizando modelos de aprendizado de máquina. O procedimento experimental compreende: (i) caracterização dos conjuntos de dados, (ii) geração de dados sintéticos, (iii) treinamento dos modelos de classificação e (iv) avaliação estatística dos resultados.

3.1. Conjuntos de Dados

O estudo utiliza dois conjuntos de dados experimentais amplamente utilizados na literatura para análise de escoamento multifásico.

3.1.1. Dataset I - Bifásico Horizontal (Al-Dogail)

O primeiro conjunto de dados foi desenvolvido por [Al-Dogail and Gajbhiye 2021] para investigar o comportamento hidrodinâmico de escoamentos gás–líquido em dutos horizontais. Os experimentos consideram diferentes propriedades físicas do fluido, incluindo tensão superficial, viscosidade e densidade do líquido, além de diferentes velocidades superficiais das fases gás e líquido. As principais variáveis medidas incluem o regime de escoamento observado e o gradiente de queda de pressão ao longo da tubulação.

3.1.2. Dataset II – Escoamento Óleo–Água

O segundo conjunto de dados foi compilado por [A. Al-Sarkhi and Hasan 2017] a partir de diferentes estudos experimentais sobre escoamento bifásico óleo–água, contendo aproximadamente 2560 pontos experimentais. Os dados abrangem diferentes misturas óleo–água, diâmetros de tubulação, ângulos de inclinação e combinações de velocidades superficiais das fases. Esse conjunto de dados é relevante para análise da transição entre regimes estratificados e não estratificados. Os conjuntos de dados relacionam propriedades físicas dos fluidos, condições operacionais e respostas hidrodinâmicas observadas. A Tabela 1 apresenta as principais variáveis utilizadas nos experimentos.

Tabela 1. Estrutura das variáveis utilizadas nos conjuntos de dados experimentais

Grupo	Variáveis	Tipo
Propriedades físicas	ρ, μ, σ	Entrada
Condições operacionais	$V_{SL}, V_{SG}, V_{SO}, V_{SW}, \theta$	Entrada
Variáveis hidrodinâmicas	$\Delta P/L$	Saída
Regime de escoamento	Classes observadas	Saída

3.2. Geração de Dados Sintéticos

Para lidar com o desbalanceamento entre classes, foram avaliadas técnicas tradicionais de oversampling e modelos generativos baseados em aprendizado profundo. As técnicas analisadas incluem SMOTE, ADASYN, Variational Autoencoder (VAE) e Tabular Variational Autoencoder (TVAE). Esses métodos foram utilizados para gerar amostras sintéticas das classes minoritárias, ampliando a diversidade do conjunto de treinamento.

3.3. Treinamento e Avaliação

Os conjuntos de dados originais e aumentados foram utilizados no treinamento de três modelos baseados em ensemble, sendo eles: Random Forest, XGBoost e LightGBM. Os modelos foram avaliados utilizando validação cruzada estratificada com 10 partições. Para os métodos de boosting, aplicou-se *early stopping* para reduzir sobreajuste. O desempenho foi avaliado utilizando acurácia, precisão, recall e F1-score.

Além disso, foi realizada análise estatística utilizando o teste não paramétrico de Friedman e o teste post-hoc de Nemenyi ($p < 0.05$). Todos os experimentos foram avaliados em um conjunto de teste independente composto por 2904 amostras, mantido isolado das etapas de treinamento e geração de dados sintéticos.

4. RESULTADOS E DISCUSSÃO

4.1. Configuração Experimental e Métricas de Avaliação

Os experimentos foram conduzidos utilizando três algoritmos amplamente empregados em tarefas de classificação com dados tabulares: *Random Forest*, *XGBoost* e *LightGBM*. Os conjuntos de dados apresentam desbalanceamento entre classes, especialmente em estados operacionais minoritários. Para mitigar esse problema, foram avaliadas técnicas tradicionais de oversampling e modelos generativos probabilísticos, incluindo SMOTE, ADASYN, Borderline-SMOTE, *Gaussian Copula*, *Variational Autoencoder* (VAE) e *Tabular Variational Autoencoder* (TVAE). O aumento de dados foi aplicado apenas nas classes minoritárias.

A avaliação foi realizada por validação cruzada estratificada em 10 *folds*. As métricas reportadas incluem *acurácia*, *precisão* e *F1-score*, com ênfase no comportamento das classes minoritárias.

4.2. Impacto do Data Augmentation no Desempenho de Classificação

A Figura 1 apresenta o desempenho dos classificadores sob diferentes técnicas de *data augmentation*. Na **Base de Dados I**, o XGBoost apresentou o melhor desempenho geral, mantendo valores de *F1-score* entre 0,93 e 0,96. Os melhores resultados foram obtidos com GMM e TVAE. O Random Forest apresentou comportamento estável, enquanto o LightGBM demonstrou maior dependência do aumento de dados, saindo de *F1-score* igual a 0,70 sem *augmentation* para aproximadamente 0,94 com TVAE. Na **Base de Dados II**, observou-se maior homogeneidade entre os resultados. O XGBoost novamente apresentou os melhores desempenhos, atingindo aproximadamente 0,93 com Borderline-SMOTE e VAE. LightGBM e Random Forest mantiveram desempenho relativamente estável entre os métodos avaliados.

De forma geral, os resultados indicam que técnicas de *data augmentation* contribuem para a melhoria do desempenho dos classificadores, especialmente em cenários

desbalanceados. Além disso, métodos generativos, como VAE e TVAE, apresentaram desempenho competitivo em ambos os conjuntos de dados.

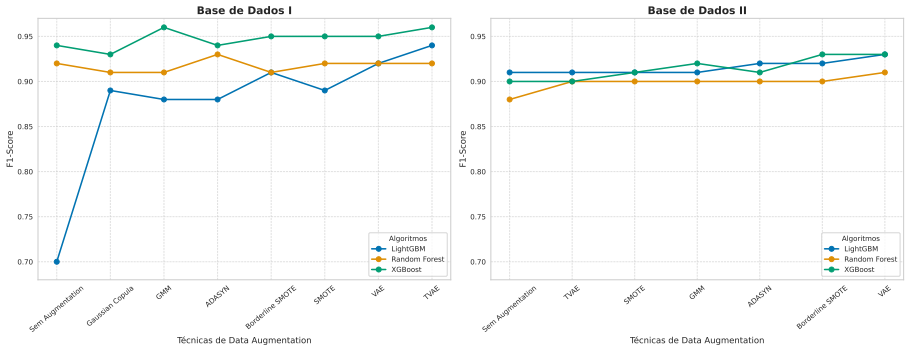

Figura 1. Comparação do desempenho dos classificadores sob diferentes estratégias de data augmentation.

4.3. Resultados Estatísticos e Análise Comparativa

A análise estatística demonstra que modelos generativos, particularmente VAE e TVAE, apresentaram tendência de desempenho superior em comparação com métodos tradicionais de sobremostragem. No Dataset I, embora o TVAE tenha apresentado o melhor ranking médio ($\bar{R} = 2.00$), o teste de Friedman ($p = 0.3216$) indicou ausência de diferenças estatisticamente significativas entre os métodos avaliados.

Por outro lado, o Dataset II apresentou diferenças significativas entre os métodos ($p = 0.0275$). Nesse cenário, o VAE apresentou o melhor desempenho, alcançando ranking médio de $\bar{R} = 1.17$. O teste pós-hoc de Nemenyi indicou desempenho superior em relação ao cenário sem aumento de dados. Os resultados evidenciam que a eficácia das técnicas generativas depende das características estatísticas do domínio analisado. Em cenários mais complexos, modelos generativos mostraram maior capacidade de capturar a estrutura latente dos dados.

A Figura 2 apresenta o diagrama de Diferença Crítica obtido a partir dos testes de Friedman e Nemenyi. Observa-se que o VAE apresentou o melhor desempenho global ($\bar{R} = 1.17$), enquanto o cenário sem *data augmentation* apresentou o pior ranking médio ($\bar{R} = 6.33$). Esses resultados reforçam que estratégias de geração sintética de dados podem contribuir para melhorar o desempenho de classificadores em cenários com desbalanceamento de classes.

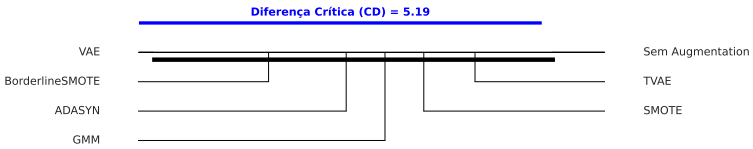

Figura 2. Diagrama de Diferença Crítica

5. CONCLUSÕES

Este trabalho investigou o impacto de estratégias de *data augmentation* na classificação de estados operacionais em dados industriais desbalanceados. Os resultados demonstram que técnicas de geração sintética contribuem para melhorar o desempenho dos modelos, especialmente nas classes minoritárias. Entre as abordagens avaliadas, modelos generativos como VAE e TVAE apresentaram desempenho mais estável em comparação com métodos clássicos de sobremostragem, sugerindo maior capacidade de capturar a estrutura estatística dos dados tabulares. Como trabalhos futuros, pretende-se investigar estratégias de geração sintética orientadas por restrições físicas e avaliar essas abordagens em cenários multimodais.

Referências

A. Al-Sarkhi, A. Sarica, S. A.-K. and Hasan, H. (2017). Dimensionless oil–water stratified to non-stratified flow pattern transition. *Journal of Petroleum Science and Engineering*, 151:284–291.

A. Fernández, S. G. and Herrera, F. (2018). Addressing the classification with imbalanced data: Smote and beyond. *Artificial Intelligence Review*, 52(3):665–691.

Al-Dogail, A. S. and Gajbhiye, R. N. (2021). Effects of density, viscosity and surface tension on flow regimes and pressure drop of two-phase flow in horizontal pipes. *J. Pet. Sci. Eng.*, 205:108719.

Borges, C. F. et al. (2024). Limitations of conventional data augmentation in engineering tabular datasets: A review. *Comput. Ind. Eng.*, 189:Article 109960.

Dablain, D. A. and Chawla, N. V. (2024). Data augmentation's effect on machine learning models when learning with imbalanced data. In *2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, San Diego, CA, USA.

Khan, S. T. et al. (2025). Establishing operational boundaries for tabular data augmentation in petroleum engineering. *Energy Rep.*, 11:3120–3129.

Kumar, A. V. et al. (2024). Safety constraints and optimization in offshore production systems: A data-driven approach. *IEEE Access*, 12:5500–5510.

L. Xu, A. Sklar, C. S. X. and Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. In *Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS)*.

Mumuni, A. and Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array*, 16:100258.

Patel, N. S. et al. (2024). Improving model generalization with gaussian copula-based data augmentation. *Expert Syst. Appl.*, 250:Article 124100.

Santos, M. A. et al. (2024). Operational cost analysis and data scarcity in multiphase flow facilities. *IEEE Trans. Ind. Appl.*, 60(1):100–109.

Zhang, H. M. and Li, Q. (2023). Enhancing tabular data generation via tabular variational autoencoders (tvae) for small datasets. *IEEE Access*, 11:99000–99010.