TRATAMENTO DE DADOS PARA CLASSIFICAÇÃO DE PADRÕES DE ESCOAMENTO USANDO APRENDIZADO DE MÁQUINA

EDEN TAYLOR DALA BARBA CORRÊA[1]; RONEY MENEZES MEIRELLES JÚNIOR[2]; SOFIA PAGLIARINI[3]; JOÃO INÁCIO MOREIRA BEZERRA[4]; MARLON MAURICIO HERNANDEZ CELY[5]; JAIRO VALÕES DE ALENCAR RAMALHO[6]

[1]*Universidade Federal de Pelotas – dalabarbacorrea@gmail.com*
[2]*Universidade Federal de Pelotas – rjmeirelles03@gmail.com*
[3]*Universidade Federal de Pelotas – sofiapagliarini@gmail.com*
[4]*Universidade Federal de Pelotas – jimbezerra@inf.ufpel.edu.br*
[5]*Universidade Federal de Pelotas – marlon.cely@ufpel.edu.br*
[6]*Universidade Federal de Pelotas – jairo.ramalho@ufpel.edu.br*

1. INTRODUÇÃO

O conhecimento dos padrões de escoamento provenientes do fluxo bifásico de água e óleo é de grande importância para a indústria óleo-gás, pois variáveis críticas para projetos nessa área têm forte correlação com o padrão de escoamento (Cely et al, 2024). Identificar com precisão esses padrões sob diferentes condições de desalinhamento do poço também é essencial para interpretar adequadamente o perfil de saída do registro de produção (Sun et al, 2023).

Frequentemente, em modelos de classificação, nos deparamos com bancos de dados que possuem classes muito desbalanceadas, como por exemplo: dados de diagnóstico de doenças raras, defeitos de fabricação, transações fraudulentas, etc. Treinar um modelo em um conjunto de dados com poucas observações de uma determinada classe resulta em um desempenho preditivo ruim do mesmo, especialmente para as observações pertencentes à classe minoritária (NORA, 2024).

A detecção de eventos indesejáveis em poços de petróleo e gás pode ajudar a prevenir perdas na produção, acidentes ambientais, vítimas humanas e reduzir os custos de manutenção (VARGAS, 2019).

Este trabalho aborda o problema da classificação de padrões de escoamento multifásico utilizando o algoritmo *Random Forests*[1]. Para melhorar o desempenho do modelo diante de dados desbalanceados, técnicas de balanceamento, como *SMOTE* (*Synthetic Minority Over-sampling Technique*[2]) e *Tomek Links*, foram empregadas, visando otimizar a distribuição das classes.

A pesquisa insere-se na área de aprendizado de máquina aplicado à engenharia, buscando contribuir para o aprimoramento das técnicas de análise de escoamento multifásico e fornecer ferramentas que possam ser aplicadas em diversos setores industriais. O presente estudo tem como objetivo investigar a eficácia dessas técnicas no contexto específico de classificação de padrões de escoamentos, propondo uma solução capaz de lidar com os desafios inerentes aos dados desequilibrados e de múltiplas classes.

---

[1] Florestas Aleatórias.
[2] Técnica de Superamostragem Sintética de Minorias.

2. METODOLOGIA

A criação dos modelos de classificação, validação desses mesmos modelos e o tratamento de dados foram feitos utilizando bibliotecas da linguagem de programação *Python* como *pandas*, *numpy*, *scikit-learn* e *imblearn*.

Neste trabalho foram utilizados dados que haviam sido coletados e analisados em (CELY et al, 2024) a partir de estudos disponíveis na literatura. Foram observados diferentes padrões de escoamento multifásico, incluindo: estratificado (ST), estratificado e mistura na interface (ST & MI), anular (A), intermitente (I), dispersão de óleo em água e água (D o/w & w), dispersão de óleo em água (D o/w), dispersão de água em óleo (D w/o), e dispersão de água em óleo e óleo em água (D w/o & D o/w).

Os padrões de escoamento compilados e os autores dos trabalhos de onde foram encontrados os dados são mostrados na Tabela 1. As referências da Tabela 1 podem ser encontradas em (CELY et al, 2024) e não foram incluídas ao final do trabalho devido à falta de espaço.

Autor	Classificação
(ABDUVAYT; et al. ,2004)	ST, ST & MI, D o/w & w , Do/w
(AL-SARKHI; et al.ed.2. ,2017)	ST, ST & MI,
(AL-SARKHI; et al.ed.3. ,2017)	ST, ST & MI
(AL-WAHAIBI; et al.ed.1. ,2014)	A, I, ST, D o/w, D w/o
(AL-WAHAIBI; et al.ed.2. ,2014)	A, I, ST, D o/w, D w/o
(AL-WAHAIBI; et al.ed.3. ,2014)	A, I, ST, D o/w, D w/o
(CAI; et al. ,2012)	ST, ST & MI, D w/o
(DASARI; et al. ,2013)	ST, I, ST & MI, D w/o, D o/w
(GRASSI; et al. ,2008)	ST, I, A, D o/w
(IBARRA; et al. ,2015)	ST, ST & MI, D o/w, D w/o
(MONTOYA; et al. ,2009)	ST, ST & MI, D o/w, D o/w & w
(NÄDLER; et al. 1997)	ST, D w/o & D o/w, ST & MI, D w/o, D o/w
(RODRIGUEZ; 2006)	ST, D w/o & D o/w, ST & MI, D w/o, D o/w
(SHI; et al. ed. 1. 2017)	A, I
(SHI; et al. ed. 2. 2017)	A, I
(WEGMANN; et al.ed. 1. ,2006)1	ST, I, A, D w/o
(WEGMANN; et al. ed. 2. ,2006)2	ST, I, A, D w/o

Tabela 1: Agrupamento por Autores dos Padrões de Escoamento

3. RESULTADOS E DISCUSSÃO

Antes de gerar os classificadores, foi realizada uma etapa de tratamento de dados, na qual 27 entradas com valores faltantes foram removidas, resultando em 2119 dados disponíveis. Desses, apenas 10 apresentavam o padrão D o/w & w, e apenas 20 tinham o padrão D w/o. Para contornar esse desbalanceamento, foram utilizadas ferramentas de balanceamento de dados da biblioteca *Imblearn*, aplicando subamostragem com *Tomek links* e sobreamostragem com *SMOTE*. Como resultado, obteve-se um conjunto de dados contendo 3592 amostras, sendo 449 pertencentes a cada um dos diferentes padrões de escoamento.

A avaliação do classificador foi realizada utilizando a curva ROC (*Receiver Operand Characteristic*[3]) e validação cruzada. Para avaliar os algoritmos classificadores, foi empregada a validação cruzada, que analisa a acurácia do

[3] Característica de Operação do Receptor.

classificador de forma geral e evita *overfitting*[4]. Os dados foram divididos em 10 conjuntos aleatórios, resultando em 10 validações, cada uma com um conjunto diferente para teste e os demais para treinamento. Cada um desses processos de validação cruzada foi executado 30 vezes para verificar se o modelo mantém sua precisão mesmo ao alterar os conjuntos de treinamento e teste.

Os melhores resultados foram obtidos quando o *tomek link* foi utilizado antes do *SMOTE*, ambos fazendo balanceamento de dados em todas as classes, a acurácia média do *random forest* foi de 97,28%, para comparação modelos de classificação utilizando KNN e SVM também foram avaliados, o KNN obteve acurácia média de 93,58% e o SVM obteve acurácia de 95,81%.

Além da validação cruzada, também foram utilizadas as curvas ROC para avaliar o classificador. A curva ROC (Figura 1) mostra a relação entre a taxa de verdadeiros positivos e a taxa de falsos positivos, sendo traçada ao plotar a taxa de verdadeiros positivos no eixo y e a taxa de falsos positivos no eixo x. Para comparar a eficácia de diferentes classificadores, pode-se utilizar as áreas abaixo das curvas (AUC).

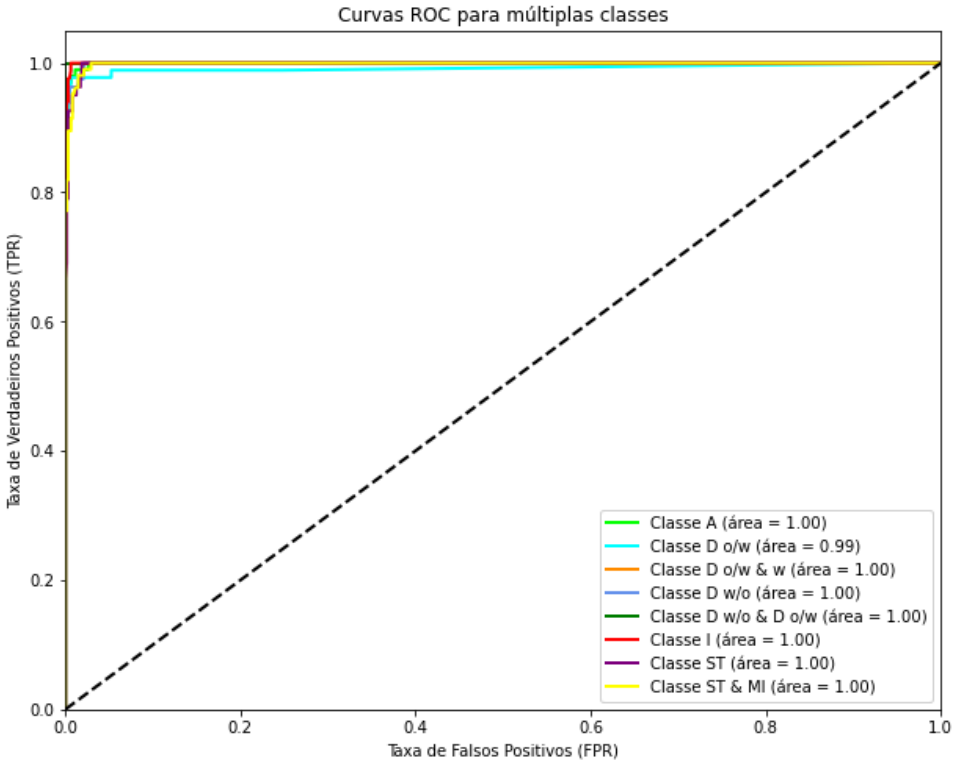

Figura 1: Curvas ROC do classificador *random forest*.

A seguir, a Tabela 2 apresenta as médias das áreas abaixo da curva (AUC) para cada padrão de escoamento.

Padrão de escoamento	AUC
Anular	99,93%
Dispersão de óleo em água	99,22%
Dispersão de óleo em água e água	100%
Dispersão de água em óleo	99,9%
Dispersão de água em óleo e dispersão de óleo em água	100%
Intermitente	99,98%
Estratificado	99,85%

[4] Sobreajuste.

Padrão de escoamento	AUC
Anular	99,93%
Dispersão de óleo em água	99,22%
Dispersão de óleo em água e água	100%
Estratificado e mistura na interface	99,84%

Tabela 2: AUC média para cada padrão de escoamento.

Como podemos observar as áreas abaixo da curva são bem próximas de 1, o que é um resultado ótimo. Para fins de comparação, a AUC média obtida com KNN foi de 98,8098%, e a AUC média do SVM foi de 99,62%.

4. CONCLUSÕES

Nesta pesquisa foi apresentado um algoritmo *Random Forest*, para classificação de padrões de escoamento bifásico óleo-água em tubulações horizontais.

Mostra-se que em determinados problemas de classificação, pode-se utilizar modelos relativamente simples de aprendizagem de máquina e conseguir resultados satisfatórios, ao invés de buscar uma arquitetura mais complexa para gerar o classificador. Neste trabalho, foi abordado um desbalanceamento de dados usando SMOTE para conseguir melhorar a precisão do modelo.

Com base nos resultados quantitativos, observa-se que os modelos têm ótima precisão e se ajustam bem a diferentes distribuições dos dados, o que é confirmado pelos resultados de validação cruzada, ROC e AUC.

5. REFERÊNCIAS BIBLIOGRÁFICAS

Cely, M., Diaz, C., Cavalheiro, H. de L., Camperos, J., Garcia, A. P., Pagliarini, S., Rodriguez, O., Rossi, M., Evald, P., & Peñaloza, E. (2024). **Identificação de padrões de escoamento bifásico água-óleo em tubulações horizontais da indústria óleo-gás através de técnicas de inteligência artificial**. Revista Científica Digital de São Paulo, 1(1), 92 páginas. https://doi.org/10.37885/978-65-5360-640-1

NORA, Andrielle Couto. *Modelo de classificação para dados desbalanceados: método SMOTE e variantes*. 2024. Trabalho de Conclusão de Curso (Graduação em Estatística) – Universidade Federal de São Carlos, São Carlos, 2024. Disponível em: https://repositorio.ufscar.br/handle/ufscar/19545. Acesso em: 09 out. 2024.

SUN, Y. et al. A comparative study of oil–water two-phase flow pattern prediction based on the GA-BP neural network and random forest algorithm. *Processes* (Basel, Switzerland), v. 11, n. 11, artígo 3155, 5 nov. 2023.

VARGAS, R. E. V. et al. A realistic and public dataset with rare undesirable real events in oil wells. *Journal of Petroleum Science and Engineering*, Netherlands, v. 181, p. 106223, 2019. Disponível em: https://www.sciencedirect.com/science/article/pii/S0920410519306357. Acesso em: 09 out. 2024.