ANÁLISE EXPLORATÓRIA DO CONJUNTO DE DADOS 3W PARA DETECÇÃO DE FALHAS DE OPERAÇÃO DE POÇOS DE PETRÓLEO, USANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA

Antonio Alberto Moreira de Azevêdo

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientadores: Sergio Lima Netto
Ricardo Emanuel Vaz Vargas

Rio de Janeiro
Dezembro de 2024

ANÁLISE EXPLORATÓRIA DO CONJUNTO DE DADOS 3W PARA DETECÇÃO DE FALHAS DE OPERAÇÃO DE POÇOS DE PETRÓLEO, USANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA

Antonio Alberto Moreira de Azevêdo

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Orientadores: Sergio Lima Netto
Ricardo Emanuel Vaz Vargas

Aprovada por: Prof. Sergio Lima Netto, Ph.D.
Eng. Ricardo Emanuel Vaz Vargas, D.Sc.
Prof. Eduardo Antônio Barros da Silva, Ph.D.
Eng. Afranio José de Melo Junior, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
DEZEMBRO DE 2024

Moreira de Azevêdo, Antonio Alberto

Análise exploratória do conjunto de dados 3W para detecção de falhas de operação de poços de petróleo, usando técnicas de aprendizado de máquina/Antonio Alberto Moreira de Azevêdo. – Rio de Janeiro: UFRJ/COPPE, 2024.

XVI, 86 p.: il.; 29, 7cm.

Orientadores: Sergio Lima Netto

Ricardo Emanuel Vaz Vargas

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2024.

Referências Bibliográficas: p. 81 – 86.

1. Aprendizagem de máquina. 2. Multiclasse. 3. Manutenção preditiva. I. Netto, Sergio Lima *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

Agradecimentos

Especialmente a meu amor, Andréa, que me incentivou a reiniciar o Mestrado após a pausa, forçada pela nossa mudança para o Rio de Janeiro. Houve muita paciência e compreensão nestes longos meses. E aos meus filhos Hanna, Leonardo e Catarina, pelas palavras de incentivo e compreensão, nas horas de minhas ausências e perda de humor.

À Petrobras, por ter liberado 40% da minha carga horária de trabalho para a realização do Mestrado. Aos colegas Daniel Graziani e Fernando Fontanezzi, e no final desta jornada, à minha gerente, Erika Almirão pelo apoio.

Ao amigo, prof. Jugurta Filho, pelas orientações iniciais na UFS, e ao prof. Luiz Pereira Calôba, que me incentivou em um momento de grande dúvida na COPPE, a seguir para a qualificação, sugerindo caminhos.

Ao meu orientador, prof. Sergio Lima Netto, agradeço pela simplicidade e cordialidade ao apresentar as opções de pesquisa, sempre muito paciente com um estudante retornando à academia de engenharia trinta anos depois. Tive a certeza de estar seguindo o melhor caminho. Obrigado por tudo.

Ao novo orientador, Ricardo Emanuel Vaz Vargas, agradeço pela luz na busca incessante pelo *fio condutor* da dissertação. Apesar de pouco tempo, suas orientações foram importantes para aperfeiçoar a narrativa.

Aos colegas Matheus Marins e Thadeu Dias pelo suporte no *embarque* da plataforma MAIS.

Ao servidores da COPPE, Maurício Machado, pelo constante apoio na Secretaria do DEE, e Gabriella Trigueiro, pela prontidão no suporte nos acessos ao *cluster* do SMT.

Finamente, agradeço aos membros da banca, prof. Eduardo Antônio Barros da Silva e Afranio José de Melo Junior pela disponibilidade em participar do processo.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

ANÁLISE EXPLORATÓRIA DO CONJUNTO DE DADOS 3W PARA DETECÇÃO DE FALHAS DE OPERAÇÃO DE POÇOS DE PETRÓLEO, USANDO TÉCNICAS DE APRENDIZADO DE MÁQUINA

Antonio Alberto Moreira de Azevêdo

Dezembro/2024

Orientadores: Sergio Lima Netto
Ricardo Emanuel Vaz Vargas

Programa: Engenharia Elétrica

Apresenta-se, nesta dissertação, a aplicação de técnicas de aprendizado de máquina para a análise exploratória de detecção de falhas operacionais em poços de petróleo, utilizando o conjunto de dados 3W. O trabalho foca no monitoramento preditivo e aborda desafios como a dinâmica temporal das falhas, o desbalanceamento de classes e a inclusão de eventos simulados. Utilizando algoritmos de aprendizado supervisionado e técnicas de redução de dimensionalidade, o estudo demonstra a eficácia desses métodos na predição de falhas, contribuindo para a melhoria da manutenção de poços *offshore*.

Além disso, este trabalho sugere futuras pesquisas para aprimorar o tratamento de classes desbalanceadas e investigar melhor a influência de eventos simulados sobre os dados reais, com o objetivo de melhorar a robustez e a precisão dos modelos. Outro ponto relevante levantado é a importância de desenvolver soluções que possam ser aplicadas em sistemas de monitoramento em tempo real, com o intuito de criar sistemas preditivos que sejam efetivos para a detecção precoce de falhas. Dessa forma, o estudo oferece uma base para o desenvolvimento de modelos e abordagens para detecção antecipada de eventos indesejados em ambientes complexos, como os poços de petróleo *offshore*.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

EXPLORATORY ANALYSIS OF THE 3W DATASET FOR DETECTING OIL WELL OPERATION FAILURES, USING MACHINE LEARNING TECHNIQUES.

Antonio Alberto Moreira de Azevêdo

December/2024

Advisors: Sergio Lima Netto
Ricardo Emanuel Vaz Vargas

Department: Electrical Engineering

This dissertation presents the application of machine learning techniques for exploratory analysis of operational failure detection in oil wells, using the 3W dataset. The work focuses on predictive maintenance, addressing challenges such as the temporal dynamics of failures, class imbalance, and the inclusion of simulated events. By using supervised learning algorithms and dimensionality reduction techniques, the study demonstrates the effectiveness of these methods in failure prediction, improving offshore well maintenance.

In addition, this work suggests future research to improve the treatment of imbalanced classes and to better investigate the influence of simulated events on real data, with the aim of improving the robustness and accuracy of the models. Another relevant point raised is the importance of developing solutions that can be applied to real-time monitoring systems, in order to create predictive systems that are effective for early failure detection. Thus, the study provides a basis for the development of models and approaches for early detection of undesirable events in complex environments, such as offshore oil wells.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

3W	Conjunto de dados, p. 3
AEM	*Abnormal Event Management*, p. 3
ANP	Agência Nacional do Petróleo, Gás Natural e Biocombustíveis, p. 1
ARMA	Média Móvel Autorregressivo, p. 8
AR	Autorregressivo, p. 8
BSW	*Base Sediments and Water*, p. 4
BalAcc	Acurácia Balanceada, p. 32
Bpd	Barris por dia, p. 1
CBM	Condition-Based Monitoring, p. 2
CKGL	*Choke* de *Gas Lift*, p. 3
CKP	*Choke* de Produção, p. 3
COPPE	Instituto Alberto Luiz Coimbra de Pós-graduação e Pesquisa de Engenharia, p. 4
CSV	*Comma Separated Values*, p. 11
DD	Diagrama de Decisão, p. 19
DHSV	*Downhole Safety Valves*, p. 4
DS	*Dataset*, p. 10
DTW	*Dynamic Time Warping*, p. 19
EFB	*Exclusive Feature Bundling*, p. 24
EUA	Estados Unidos da América, p. 2

GBDT	*Gradient Boost Decision Tree*, p. 24
GNB	*Gaussian Naive Bayes*, p. 17
GOSS	*Gradient-based One-Side Sampling*, p. 24
Hz	Hertz, p. 11
ICA	*Independent Component Analysis*, p. 8
KNN	*K-nearest Neighbors*, p. 17
LSTM	*Long Short-term Memory*, p. 18
MAIS	*Modular Artificial Intelligence System*, p. 4
MSPC	Monitoramento Estatístico de Processos, p. 7
MTS	*Multivariate Time Series*, p. 10
OCC	*One Class Classifier*, p. 18
O&M	Operação e Manutenção, p. 2
P-CKGL	Pressão à Jusante do CKGL, p. 3
P-MON-CKP	Pressão à Montante do CKP, p. 3
P-PDG	Pressão no PDG, p. 3
P-TPT	Pressão no TPT, p. 3
PCA	*Principal Component Analysis*, p. 8
PDG	*Permanent Downhole Gauge*, p. 3
PEE	Programa de Engenharia Elétrica, p. 4
PETROBRAS	Petróleo Brasileiro S.A., p. 2
PSE	Engenharia de Sistemas de Processos, p. 7
QDA	*Quadratic Discriminant Analysis*, p. 17
QGL	Taxa de Fluxo de *Gas Lift*, p. 3
RUL	*Remaining Useful Life*, p. 9
SMT	Laboratório de Processamento de Sinais, Multimídia e Telecomunicações, p. 4

STFT	*Short-time Fourier Transform*, p. 8
SVM	, p. 17
T-CKGL	Temperatura à Jusante do CKGL, p. 3
T-JUS-CKP	Temperatura à Jusante do CKP, p. 3
T-PTP	Temperatura no TPT, p. 3
TPE	*Tree-structured Parzen Estimators*, p. 21
TPT	Transdutor de Pressão e Temperatura, p. 3
UFRJ	Universidade Federal do Rio de Janeiro, p. 4
UO-ES	Unidade de Operação Espírito Santo, p. 2

Capítulo 1

Introdução

Este capítulo descreve os principais tópicos introdutórios desta dissertação: Seção 1.1, contextualização e motivação, Seção 1.2, objetivos e contribuições, e Seção 1.3, a organização deste documento.

1.1 Contextualização e Motivação

Na indústria de petróleo e gás, as operações estão sujeitas a exigências regulatórias ambientais, especialmente na exploração e produção *offshore*. Essas exigências visam minimizar os impactos ambientais e garantir a segurança operacional, particularmente na gestão de poços. O controle de vazão e a segurança do sistema de extração são monitorados, devido ao risco associado a qualquer tipo de falha.

Atualmente, existem poços que operam com vazões superiores a 40.000 barris por dia (bpd), conforme relatado pela Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) [5]. A elevada vazão desses poços, essencial para atender à demanda por energia, também amplia as consequências de uma falha crítica, que pode interromper a produção e acarretar prejuízos financeiros. A interrupção de um poço, além de impactar a rentabilidade da operação, pode causar danos ao meio ambiente e à reputação da empresa.

Diante da complexidade e dos riscos inerentes às operações de extração em poços marítimos, a indústria petrolífera busca métodos e tecnologias que permitam prever e antecipar falhas operacionais. Esta antecipação é essencial para a continuidade das operações e para a mitigação de riscos que possam comprometer a integridade do sistema e a segurança dos trabalhadores. O desenvolvimento de técnicas avançadas de monitoramento e previsão de falhas visa reduzir a ocorrência de incidentes e acidentes, garantindo a continuidade da produção e minimizando as perdas associadas a paradas não planejadas.

O Departamento de Energia dos EUA [6] publicou um guia de melhores práticas sobre operação e manutenção (O&M), focado na eficiência energética. Estima-se que

essas práticas possam economizar entre 5% e 20% das contas de energia sem investimentos significativos de capital. O guia abrange diferentes tipos de programas de manutenção e suas definições, explorando tecnologias de manutenção, especialmente as tecnologias preditivas.

Segundo o Departamento de Energia, a manutenção preventiva pode ser definida como "Ações executadas em um cronograma baseado em tempo ou execução de máquina que detectam, impedem ou mitigam a degradação de um componente ou sistema com o objetivo de sustentar ou estender sua vida útil por meio do controle da degradação a um nível aceitável". A manutenção preditiva, por sua vez, pode ser definida como "Medições que detectam o início de um mecanismo de degradação, permitindo que os estressores causais sejam eliminados ou controlados antes de qualquer deterioração significativa no estado físico do componente". Estudos independentes apontam as seguintes economias médias industriais resultantes do início de um programa de manutenção preditiva funcional em relação a um programa que utiliza apenas manutenção preventiva [6]:

- Retorno do investimento: 10 vezes maior;
- Redução nos custos de manutenção: 25% a 30%;
- Eliminação de avarias: 70% a 75%;
- Redução no tempo de inatividade: 35% a 45%;
- Aumento da produção: 20% a 25%.

No entanto, a implementação da manutenção preditiva é dispendiosa, com um custo médio estimado superior a US$ 50.000, relacionado à qualificação do pessoal para utilizar eficazmente as tecnologias de manutenção preditiva.

Segundo MELO *et al.* [7], o monitoramento de processos é frequentemente usado de forma intercambiável com *detecção e diagnóstico de falhas*. No artigo dos autores é apresentada uma revisão da literatura sobre monitoramento de processo baseado em dados.

No trabalho de JARDINE *et al.* [8] é discutido o conceito de monitoramento baseado na condição (*Condition-Based Monitoring*, CBM), um programa de manutenção que recomenda decisões de manutenção com base nas informações coletadas. Trata-se de um sistema que suporta as decisões de manutenção preditiva e pode ser usado para fazer diagnósticos, prognósticos ou ambos.

Segundo VARGAS *et al.* [9], alguns poucos tipos de eventos indesejados foram responsáveis pela maioria da perda de produção da Unidade de Operação Espírito Santo (UO-ES), uma das Unidades da Petróleo Brasileiro S.A. (PETROBRAS). Em 2016 esta perda estimada foi equivalente a 1.514.000 barris, correspondendo a US$75,7 milhões, considerando um valor médio de US$50,00/barril naquele período.

A PETROBRAS decidiu publicar um *dataset* para ser empregado no desenvolvimento de Gerenciamento de Eventos Anormais ou *Abnormal Event Management*

(AEM) [10]. O conjunto de dados denominado 3W foi inicialmente publicado por VARGAS *et al.* [9] em 2019 e republicado em 2022 pela PETROBRAS, como oportunidade no Programa de Tecnologia e Inovação, no Módulo *Open Lab* [11] e no portal GitHub PETROBRAS [12]. O Módulo *Open Lab* é voltado para a publicação de desafios e oportunidades de desenvolvimento de software em uma abordagem colaborativa. O 3W serve como referência para o desenvolvimento de técnicas de aprendizado de máquina voltadas à análise de anomalias em poços de petróleo.

As variáveis monitoradas mais comuns em poços *offshore* da PETROBRAS estão presentes no conjunto de dados 3W. Excluindo as variáveis relativas a gás lift (P-CKGL, T-CKGL E QGL), as demais são as mais comuns em poços de fluxo natural. Poços de fluxo natural podem ser equipados para ser operados com métodos artificiais de elevação sob certas circunstâncias. Na Tabela 1.1 são listadas as variáveis [9].

Tabela 1.1: Variáveis monitoradas mais comuns em poços *offshore* da PETROBRAS.

Nome	**Descrição**	**Unidade**
P-PDG	Pressão no medidor de fundo de poço (*Permanent Downhole Gauge* - PDG)	Pa
P-TPT	Pressão no transdutor de pressão e temperatura (TPT)	Pa
T-TPT	Temperatura no TPT	ºC
P-MON-CKP	Pressão à montante do choke de produção (CKP)	Pa
T-JUS-CKP	Temperatura à jusante do CKP	ºC
P-CKGL	Pressão à jusante do *choke* de *gas lift* (CKGL)	Pa
T-CKGL	Temperatura à jusante do CKGL	ºC
QGL	Taxa de fluxo de gás lift (QGL)	m^3/s

Os eventos indesejados, que neste documento são chamados de falhas, são descritos detalhadamente em VARGAS *et al.* [9] e incluem:

1. Aumento Abrupto de Sedimentos e Água (*Base Sediments and Water*, BSW);
2. Fechamento Espúrio da Válvula de Segurança de Subsuperfície (*Downhole Safety Valves*, DHSV);
3. Golfada Severa;
4. Instabilidade de Fluxo;
5. Perda Rápida de Produtividade;
6. Restrição Rápida no *Choke* de Produção (CKP);
7. Incrustação (*Scaling*, depósito de sais minerais que pode ocorrer) no CKP;
8. Hidrato na Linha de Produção.

O 3W apresenta **desafios** para o desenvolvimento de um sistema de monitoração, que o tornam um *dataset* com nuances para pesquisa [1]: possível existência de dois tipos de amostras normais; variação da dinâmica temporal; e desbalanceamento entre as classes.

A UFRJ/COPPE/PEE, em seu Laboratório de Processamento de Sinais, Multimídia e Telecomunicações (SMT), desenvolveu a plataforma MAIS (*Modular Artificial Intelligence System*), dedicada ao estudo da base de dados 3W [1].

1.2 Objetivos e Contribuições

Com este estudo é pretendido realizar análise exploratória de dados, com técnicas de aprendizagem de máquina, no 3W, para detecção de eventos indesejados em poços de petróleo.

Para enfrentar algumas dificuldades do 3W, foram formuladas algumas perguntas de pesquisa e apresentadas na Seção 3.2. Utilizando a plataforma MAIS para experimentos multiclasse, estas e outras questões foram investigadas. Foram pesquisadas questões relativas ao tempo de treinamento, quantidade de experimentos, medidas de avaliação de desempenho, influência de dados simulados no desenvolvimento do classificador e possível existência de *clusters* na Classe 1 (Aumento Abrupto de BSW).

Este estudo resultou nas seguintes contribuições:

(i) Análise sobre a dinâmica temporal dos eventos normais e das classes de falha.

(ii) Estudo sobre a relevância dos dados simulados no *dataset*.

(iii) Proposta de um modelo de detecção de falha que demande pouco tempo de treinamento, para ser posto em produção como uma ferramenta de manutenção preditiva.

(iv) A última contribuição é uma análise sobre possíveis causas de existência de *clusters* na Classe de falha referente a Aumento Abrupto de BSW.

1.3 Organização da Dissertação

O Capítulo 2 procura situar o leitor no problema de detecção de falhas na operação de poços de petróleo. Para tal, é abordado sobre monitoramento de processo e sobre o programa de manutenção baseada na condição. Em seguida, é descrita a base de dados 3W. Depois é apresentada uma revisão bibliográfica de trabalhos utilizando 3W, é feita uma descrição da plataforma MAIS, e é realizada uma revisão bibliográfica de trabalhos sobre a ocorrência de BSW.

No Capítulo 3, foca-se na análise exploratória dos desafios específicos do conjunto de dados 3W. O objetivo é compreender as complexidades envolvidas na aplicação de

técnicas de aprendizado de máquina para o monitoramento de poços de petróleo. Na pesquisa busca-se propor soluções para melhorar a detecção de falhas. É apresentado um Experimento de Referência com seus resultados e questões de pesquisa.

No Capítulo 4, são apresentados os resultados da pesquisa, e analisada a melhor configuração para geração de alarmes de falha no sistema CBM, com foco na menor taxa de alarmes perdidos.

Finalmente, no Capítulo 5 são apresentadas as conclusões de cada questão apresentada, e aponta o leitor para possíveis direções de extensão da pesquisa no tema desta dissertação.

Capítulo 2

Monitoramento de Poços de Petróleo

Este capítulo procura situar o leitor no problema de detecção de falhas na operação de poços de petróleo, tema central desta dissertação. Para tal, na Seção 2.1 é abordado sobre monitoramento de processo. Na Seção 2.2 é abordado o programa de manutenção baseada na condição. Na Seção 2.3 é descrita a base de dados 3W, contendo uma série de medições associadas à operação (normal ou defeituosa) de poços de petróleo. Esses dados são a base de todos os estudos desenvolvidos nesta dissertação. Na Seção 2.4 são apresentas as medidas de desempenho empregadas neste trabalho. Na Seção 2.5 é feita uma revisão bibliográfica de trabalhos utilizando a base 3W, concluindo com uma discussão das contribuições em relação aos estudos anteriores. Na Seção 2.6 é feita uma descrição da plataforma MAIS (Modular Artificial Intelligence System), descrevendo todas as suas funcionalidades para o desenvolvimento eficiente de modelos de detecção e classificação de falhas na operação de poços. Na Seção 2.7 é feita uma Revisão bibliográfica de trabalhos sobre a ocorrência de Sedimentos e Água (*Base Sediments and Water*, BSW). O aumento abrupto de BSW é uma das falhas investigadas, sob a ótica de possível existência de *clusters* na base de dados. Por fim, na Seção 2.8 conclui-se o capítulo sumarizando as principais contribuições.

2.1 Monitoramento de Processo

O monitoramento de processos abrange práticas utilizadas na supervisão da produção e na identificação de falhas em sistemas industriais. Segundo MELO *et al.* [7], este campo evoluiu com a introdução dos gráficos de controle por SHEWHART [13], que marcaram o início do monitoramento estatístico de processos (MSPC). Esses gráficos distinguem variações comuns das variações especiais. Posteriormente, HOTELLING [14] introduziu a estatística T^2, permitindo a análise simultânea de múltiplas variáveis e suas interações, estabelecendo o monitoramento estatístico multivariado de processos (MSPC). A partir da década de 1980, a acessibilidade a computadores e

o desenvolvimento de modelos de variáveis latentes, como a análise de componentes principais (PCA), impulsionaram o MSPC. Desde os anos 2000, o avanço computacional e o surgimento de técnicas de aprendizado de máquina permitiram a análise de grandes volumes de dados e a identificação de padrões complexos. Apesar de discussões recentes classificarem modelos multivariados como aprendizado de máquina, na literatura de Engenharia de Sistemas de Processos (PSE), essas abordagens ainda são tratadas separadamente devido às suas origens históricas. Segundo os autores MELO *et al.* [7], no campo da PSE, o termo *monitoramento de processos* é frequentemente usado de forma intercambiável com *detecção e diagnóstico de falhas* [7]. Contudo, em outras áreas, como processamento de sinais e análise de dados médicos, conceitos semelhantes são denominados *detecção de anomalias* [15], *detecção de novidades* [16] e *detecção de pontos de mudança* [17]. Embora essas nomenclaturas abordem variações do mesmo problema fundamental, há uma falta de troca de informações entre as comunidades que investigam esses tópicos e a de PSE. Essa segregação resulta, em parte, das diferenças na natureza dos dados e das metodologias empregadas. Por exemplo, dados médicos geralmente são univariados e analisados em lotes, enquanto dados de processos industriais frequentemente apresentam alta redundância e colinearidade.

A confiabilidade das técnicas de detecção depende da qualidade dos dados coletados, bem como da implementação de sistemas robustos capazes de lidar com incertezas e ruídos inerentes aos processos industriais. O uso de indicadores como a taxa de falsos positivos auxilia na avaliação do desempenho das metodologias aplicadas [7]. O avanço na detecção de anomalias de processo representa um componente no contexto da Indústria, fornecendo ferramentas que otimizam a operação e manutenção de sistemas complexos, resultando em maior segurança, produtividade e eficiência energética.

2.2 Monitoramento Baseado na Condição

Os autores JARDINE *et al.* [8] discutem o conceito de monitoramento baseado na condição (*condition-based monitoring*, CBM), um programa de manutenção que toma decisões com base nas informações coletadas por meio do monitoramento contínuo da operação de um equipamento, sistema ou processo. Segundo JARDINE *et al.* [8], o processo de CBM completo trata da detecção, isolamento e identificação de falhas quando elas ocorrem. A detecção de falhas é uma tarefa para indicar se algo está errado no sistema monitorado; o isolamento de falhas é uma tarefa para localizar o componente que está com defeito; e a identificação da falha é uma tarefa para determinar a natureza da falha quando ela é detectada. Um processo de CBM pode ser usado para fazer diagnósticos, prognósticos ou ambos. O diagnóstico é a análise

posterior do evento e o prognóstico é a análise anterior do evento. Os prognósticos são muito mais eficientes do que os diagnósticos para alcançar um desempenho sem tempo de inatividade. O diagnóstico, entretanto, é necessário quando o prognóstico é equivocado e ocorre uma falha. O CBM consiste em três etapas principais: aquisição de dados, processamento de dados e suporte à decisão de manutenção. Neste trabalho, são exploradas as etapas de processamento dos dados e de suporte à decisão de manutenção.

2.2.1 Processamento de Dados

JARDINE *et al.* [8], a respeito do processamento de dados, destacam várias técnicas de análise de forma de onda dos dados para diagnóstico e prognóstico dos sistemas mecânicos, e comenta que na literatura existem três categorias de análise:

- Análise no domínio do tempo, baseada diretamente na própria forma de onda, onde a principal ideia é ajustar os dados da forma de onda para um modelo de série temporal paramétrico e extrair características baseadas neste modelo. Os modelos mais populares são o Autorregressivo (AR) e o de Média Móvel Autorregressivo (ARMA), cuja maior dificuldade é a determinação da ordem do modelo, que controla a complexidade do processo de modelagem e ainda a complexidade e a qualidade do modelo final;

- Análise no domínio da frequência, seja de forma não paramedida, que estima diretamente a densidade espectral de potência do sinal, ou de forma paramedida, que constrói um modelo para o sinal e estima a potência espectral baseada no modelo ajustado. Algumas ferramentas auxiliares de análise no domínio da frequência são a representação gráfica do espectro, filtros de frequência, análise de envelope e análise de estrutura de banda lateral. A transformada de Hilbert [18], que é uma ferramenta útil na análise de envelope, também tem sido usada para detecção e diagnóstico de falha de máquina;

- Análise de tempo-frequência, que representa a energia ou potência do sinal por meio de funções bidimensionais de tempo e frequência, para melhor revelar os padrões de falta para um diagnóstico mais acurado. As ferramentas mais comuns de análise de tempo- frequência são a transformada de Fourier de curta duração (*short-time Fourier transform*, STFT), o espectograma (a potência do STFT), a distribuição de Wigner-Ville, ou mesmo a transformada *wavelet*[18].

Alguns desses aspectos foram explorados com maior profundidade no decorrer desta dissertação.

2.2.2 Suporte à Decisão de Manutenção

As técnicas de suporte à decisão de manutenção num programa CBM podem ser divididas em duas categorias principais: diagnósticos e prognósticos. O diagnóstico de falhas concentra-se na detecção, isolamento e identificação de falhas quando elas ocorrem. Os prognósticos, entretanto, tentam prever faltas ou falhas antes que elas ocorram. Os prognósticos são superiores aos diagnósticos no sentido de que os prognósticos podem prevenir faltas ou falhas e, se impossível a prevenção, viabiliza a prontidão (com peças sobressalentes preparadas e recursos humanos planeados) para os problemas, poupando custos adicionais de manutenção não planejada [8].

Existem dois tipos principais de predição em prognósticos de máquinas. O prognóstico mais utilizado é prever quanto tempo resta antes que uma ou mais falha(s) ocorra(m), dada a condição atual da máquina ou sistema e o perfil de operação anterior. O tempo restante antes de observar uma falha é geralmente chamado de vida útil restante (*Remaining Useful Life*, RUL). Em algumas situações, especialmente quando uma falta ou falha é catastrófica, é mais desejável prever a chance de uma máquina operar sem falta ou falha até algum momento futuro, na próxima inspeção. Dada a dificuldade dessa segunda abordagem, a maioria dos artigos na literatura sobre prognósticos de máquinas discute apenas a estimativa RUL [8].

De modo geral, um monitoramento periódico é utilizado por ser mais econômico e fornecer diagnósticos mais precisos usando dados filtrados e/ou processados. Para GATTA *et al.* [4], as abordagens propostas no contexto da manutenção preditiva podem ser divididas em três metodologias: baseadas no conhecimento, baseadas em modelos e orientadas por dados. A primeira abordagem utiliza experiências anteriores para inferir regras de detecção de falhas; a segunda utiliza conhecimentos físicos ou métodos de estimativa estatística para fornecer uma representação do modelo que descreve o fenômeno observado; a terceira utiliza dados históricos e em tempo real para entender quando uma falha está prestes a ocorrer.

Durante muito tempo, as abordagens basearam-se apenas na base de conhecimento e técnicas baseadas em modelos. No entanto, esses tipos de abordagem apresentam vários inconvenientes, dado que não é fácil encontrar um modelo teórico capaz de explicar mecanismos tão complexos como os relacionados com a maquinaria industrial e, ao mesmo tempo, um modelo simplificado pode ser superficial para processar a informação fornecida pelos dados. Além disso, as abordagens baseadas no conhecimento e em modelos geralmente não são adequadas para atualização em tempo real, o que é uma situação comum quando se trabalha em estruturas de CBM. Por outro lado, o principal problema com o emprego de abordagens baseadas em dados é a necessidade de uma grande quantidade de dados e de uma forma eficiente de processá-los. Com o crescimento da internet das coisas, o barateamento dos

sistemas de aquisição, armazenamento e processamento de dados e os recentes avanços na inteligência artificial, uma abordagem orientada por dados, no contexto da manutenção preditiva, se tornou possível, conforme será investigado ao longo deste trabalho.

2.3 O Conjunto de Dados 3W

O estudo de VARGAS *et al.* [9] adotou a definição de séries temporais multivariadas (*Multivariate Time Series*, MTS), semelhante às utilizadas por ZHOU e CHAN [19].

Um *dataset* DS é um conjunto de m MTS ($S^i \,|\, i = 1, 2, \ldots, m;\ m \in \mathbb{N},\ m > 1$), e é definido como $DS = \{S^1, S^2, \ldots, S^m\}$. Cada MTS i é uma instância (também referenciada neste documento como objeto ou ocorrência), composta por um vetor de n *séries temporais univariáveis*, representado por $\mathbf{x}^i = (x_1^i, x_2^i, \ldots, x_n^i)$, onde x_j^i para $j = 1, 2, \ldots, n;\ n \in \mathbb{N},\ n > 1$, são as variáveis do processo (ou apenas variáveis). Assim, cada MTS é definida como $S^i = \{\mathbf{x}^i\}$. Cada variável j que compõe uma MTS i é, por sua vez, uma sequência temporal ordenada de p_i observações, extraídas no tempo t, denotada por $\mathbf{x}_j^i = (x_{j,1}^i, x_{j,2}^i, \ldots, x_{j,p_i}^i)$, onde $t = 1, 2, \ldots, p_i;\ p_i \in \mathbb{N},\ p_i > 1$. Portanto, cada MTS pode ser representada como uma matriz organizada de dimensões $n \times p_i$, definida como:

$$S^i = \begin{bmatrix} x_{1,1}^i & x_{1,2}^i & \cdots & x_{1,p_i}^i \\ x_{2,1}^i & x_{2,2}^i & \cdots & x_{2,p_i}^i \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1}^i & x_{n,2}^i & \cdots & x_{n,p_i}^i \end{bmatrix}.$$

Todas as instâncias têm um número fixo de variáveis n, mas cada instância pode ser composta por qualquer quantidade de observações p_i. Todas as variáveis de uma instância i têm número fixo de observações p_i.

O conjunto de dados 3W é composto por três tipos de instâncias, determinadas por suas fontes: reais, simuladas e desenhadas à mão. As instâncias reais foram extraídas do sistema de informações da planta utilizado para rastrear os processos industriais da UO-ES. Essa extração foi feita sem pré-processamento para manter seus aspectos realistas, como valores NaN, variáveis congeladas (devido a problemas de comunicação de sensores ou rede), instâncias com tamanhos diferentes e *outliers*. As instâncias geradas por dados simulados e desenhados à mão deram origem a séries temporais sem esses problemas [9].

Todas as instâncias simuladas foram geradas com o sistema OLGA, desenvolvido pela empresa Schlumberger, que é um simulador dinâmico de escoamento multifásico adotado por diversas empresas petrolíferas ao redor do mundo [20]. O OLGA é uma

ferramenta utilizada pela PETROBRAS para simulação de cenários em poços de petróleo e é um sistema que simula fenômenos dinâmicos (transientes defeituosos) [21].

Foi desenvolvida também uma ferramenta específica para o conjunto de dados 3W ser enriquecido com instâncias desenhadas à mão, a partir do conhecimento tácito dos especialistas sobre os tipos de eventos indesejáveis que foram considerados.

A versão 1.0.0 do 3W está disponível com a seguinte estrutura e características gerais: cada instância, seja real, simulada ou desenhada à mão, foi salva em um arquivo de valores separados por vírgula (*Comma Separated Values*, CSV) padronizado e dedicado. Todos os arquivos CSV foram agrupados em diretórios com base no rótulo da instância. Todas as instâncias foram geradas com observações obtidas com taxa de amostragem fixa (1 Hz). A fonte de cada instância foi incorporada ao nome do seu arquivo CSV. Todos os nomes reais dos poços foram substituídos por nomes genéricos.

Ao construir o banco de dados 3W, além dos períodos de falha, os autores indicaram para cada instância de evento um período em que as amostras não estão com falha (referido como normal) e um período de falha transitória antes da consolidação real da falha em estado estacionário, como representado na Figura 2.1. Tal procedimento de anotação permite detectar uma determinada falha durante seu estágio transitório inicial. Este é o comportamento pretendido no sistema proposto, de forma a minimizar custos de manutenção e perdas de produção.

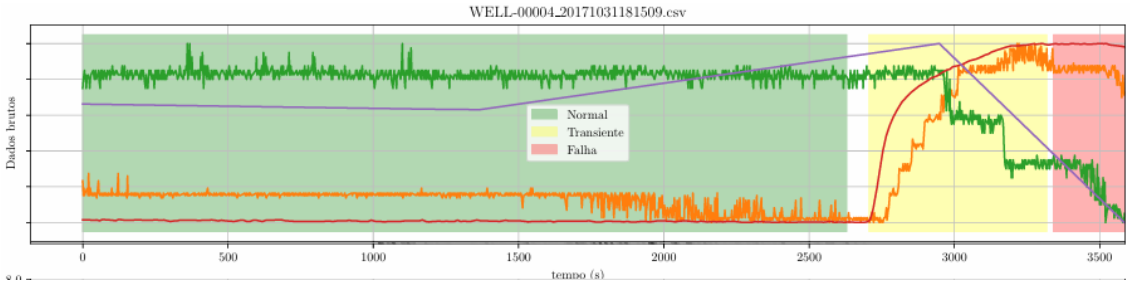

Figura 2.1: Exemplo de valores de tags (observações) normalizados de um evento (instância), onde a cor do fundo indica os estágios normal (verde), transiente (amarelo), e falha (vermelho).

Na Tabela 2.1 são apresentadas as quantidades de instâncias que compõem o conjunto de dados 3W na versão 1.0.0, divulgada em 01 de julho de 2019, por tipo de evento e por fonte de conhecimento (reais, simuladas e desenhadas à mão). Ao considerar apenas as instâncias reais, quatro classes de eventos indesejáveis se tornam raras (com menos de 1% dos eventos totais): 1, 6, 7 e 8. Se também forem consideradas as instâncias simuladas e as desenhadas à mão, apenas a Classe 7 pode ser considerada rara.

Tabela 2.1: Quantidade de instâncias que compõem o *dataset* 3W.

Tipo de evento	**Real**	**Simulada**	**Desenhada**	**Total**
0. Normal	597	-	-	597
1. Aumento Abrupto de BSW	5	114	10	129
2. Fechamento Espúrio de DHSV	22	16	-	38
3. Golfada Severa	32	74	-	106
4. Instabilidade de Fluxo	344	-	-	344
5. Perda Rápida de Produtividade	12	439	-	451
6. Restrição Rápida no CKP	6	215	-	221
7. Incrustação no CKP	4	-	10	14
8. Hidrato na Linha de Produção	3	81	-	84
TOTAL	1025	939	20	1984

Dasafios do 3W. No 3W estão contidos alguns desafios para o desenvolvimento de um sistema de monitoração, que o tornam um *dataset* interessante para muitas pesquisas [1]:

(i) Identificação da Classe 0 (atividade normal), onde amostras normais podem ser de dois tipos: número de amostras da Classe 0 (n_0), ou número de amostras do período normal inicial nas instâncias com falha (n_N). A base de dados original apresenta uma estrutura de tal modo que n_0 largamente se sobrepõe a n_N. Isto induz um viés, que leva a uma série de falsos alarmes, devido à classificação incorreta das amostras n_N;

(ii) Alguns sensores estão sempre faltando em algumas classes, tornando trivial identificar aquelas classes apenas por uma simples verificação da atividade do sensor;

(iii) A dinâmica temporal das falhas consideradas varia amplamente, com uma fase transitória variando de somente alguns minutos (Classe 2) a algumas horas (Classe 8);

(iv) O conjunto de dados é amplamente desbalanceado entre estas classes de falha, conforme Tabela 2.1.

2.4 Medidas de desempenho

2.4.1 Acurácia

Mede a acurácia média por classe, ajustando a influência do desbalanceamento entre classes ao calcular a acurácia separadamente para cada classe e depois tirar a média, onde:

$$\text{Acurácia} = \frac{\text{Número de predições corretas}}{\text{Número total de amostras}} \tag{2.1}$$

Ou, em termos de variáveis:

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.2}$$

Onde VP (verdadeiro positivo), VN (verdadeiro negativo), FP (falso positivo) e FN (falso negativo);

2.4.2 Acurácia balanceada

A medida de *acurácia balanceada* [22] é utilizada para lidar com problemas de desbalanceamento de classes em tarefas de classificação. Ela é definida como a média aritmética das taxas de acerto (*sensibilidade, S*) calculadas individualmente para cada classe. A fórmula geral é:

$$\text{Acurácia Balanceada} = \frac{1}{C}\sum_{c=1}^{C} \text{S}_c, \tag{2.3}$$

onde C é o número total de classes, e S_c para cada classe c é definido como:

$$\text{S}_c = \frac{\text{VP}_c}{\text{VP}_c + \text{FN}_c}. \tag{2.4}$$

Nesta fórmula:

- VP_c: Verdadeiros positivos da classe c,
- FN_c: Falsos negativos da classe c.

2.4.3 F1

A medida de avaliação de desempenho F balanceada (pontuação F1) é a média harmônica de precisão e sensibilidade, combinando essas duas medidas de maneira equilibrada. A precisão (P) mede a proporção de verdadeiros positivos em relação ao total de previsões positivas, enquanto a sensibilidade (S) mede a proporção de verdadeiros positivos em relação ao total de casos reais positivos.

$$\text{P} = \frac{VP}{VP + FP} \tag{2.5}$$

$$\text{S} = \frac{VP}{VP + FN} \tag{2.6}$$

$$\text{F1} = \frac{2.P.S}{P + S} \tag{2.7}$$

A pontuação F1, ao combinar essas medidas, fornece uma única medida de avaliação de desempenho que considera tanto a precisão quanto sensibilidade. Porém a F1 não considera desbalanceamento das classes.

2.4.4 Macro-F1

A medida *macro-F1* é uma média aritmética dos valores da medida $F1$-score calculados individualmente para cada classe em um problema de classificação. É definida como:

$$\text{Macro-F1} = \frac{1}{C}\sum_{c=1}^{C} F1_c, \tag{2.8}$$

onde:

$$F1_c = \frac{2 \cdot \text{P}_c \cdot \text{S}_c}{\text{P}_c + \text{S}_c}. \tag{2.9}$$

O termo P_c para cada classe c é definido como:

$$\text{P}_c = \frac{\text{VP}_c}{\text{VP}_c + \text{FP}_c} \tag{2.10}$$

E o termo S_c conforme equação 2.4
Nestas fórmulas:

- VP_c: Verdadeiros positivos para a classe c,
- FP_c: Falsos positivos para a classe c,
- FN_c: Falsos negativos para a classe c.

A *macro-F1* dá igual peso a cada classe, independentemente de seu tamanho ou proporção no conjunto de dados.

2.4.5 F1 ponderada

A medida F1-ponderada (*weighted*) [23][24] calcula medidas para cada rótulo e encontra sua média, ponderada pelo número de instâncias verdadeiras para cada rótulo. Isso altera a F1 para levar em conta o desequilíbrio do rótulo. Resulta em uma pontuação F que não está entre precisão e sensibilidade. Após calcular a F1 para cada classe, é feita uma média ponderada, onde o peso de cada classe corresponde à sua proporção no conjunto de dados.

A F1 ponderada é calculada como:

$$\text{F1 ponderada} = \frac{\sum_{c=1}^{n} w_c \cdot F1_c}{\sum_{c=1}^{n} w_c}, \tag{2.11}$$

Onde w_c é o peso de cada classe c, proporcional ao número de amostras da classe c:

$$w_c = \frac{\text{Número de amostras da classe } c}{\text{Número total de amostras}} \tag{2.12}$$

2.4.6 Coeficiente de Silhueta

O Coeficiente de Silhueta [25][26] é uma medida usada para avaliar a qualidade de *clusters* em algoritmos de agrupamento. Ele mede o quão similar um ponto é ao seu próprio *cluster* (coesão) em comparação com outros *clusters* (separação), variando entre -1 e 1. Valores próximos de 1 indicam bons agrupamentos, enquanto valores próximos de -1 sugerem que os pontos estão mal alocados.

O Coeficiente de Silhueta para uma amostra i é definido como:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{2.13}$$

onde:

- $a(i)$: É a distância média entre a amostra i e todas as outras amostras do mesmo *cluster*.
- $b(i)$: É a distância média entre a amostra i e todas as amostras do *cluster* mais próximo (ao qual i não pertence).

O coeficiente total para o conjunto de dados é obtido como a média de $s(i)$ para todas as amostras i:

$$S = \frac{1}{n} \sum_{i=1}^{n} s(i), \tag{2.14}$$

onde n é o número total de amostras.

O Coeficiente de Silhueta é útil para determinar o número ideal de *clusters* em uma análise de agrupamento. Ele ajuda a identificar estruturas naturais nos dados, fornecendo uma indicação da coesão interna dos *clusters* e da separação entre eles.

2.5 Revisão Bibliográfica Acerca da Base 3W

2.5.1 MARINS et al.

Visando a detecção e classificação das falhas presentes na base 3W, versão 1.0.0, MARINS *et al.* [2] propuseram um sistema CBM utilizando características estatísticas extraídas de um segmento de comprimento fixo dos sinais dos sensores e um classificador de *Random Forest* [27]. Foram feitos três tipos de experimentos para os detectores e classificadores de falha: de uma classe (normal x falha), binários para cada classe de falha e multiclasse.

Os autores consideraram alguns parâmetros para avaliar seu modelo. Dentre estes, foi definido um parâmetro para impor equilíbrio entre diferentes tipos de amostras normais, incluindo o número de amostras extraídas das instâncias da Classe

0, n_0, e o número de amostras da fase normal inicial nas instâncias com falha, n_N. Denotando este parâmetro por b, tem-se $b = n_0/n_N$. Portanto, o número total de amostras normais e com falha usadas durante o treinamento torna-se $n_{normal} = n_{defeituoso} = (b+1)n_N$. O uso do parâmetro b foi utilizado para poder controlar o desempenho do classificador nas amostras da fase normal inicial nas instâncias com falha.

Foram utilizadas apenas as fases normal e transitória do evento para treinar o classificador. As instâncias desenhadas foram ignoradas e a Classe 7 (Incrustação no CKP) foi desconsiderada por estar sub-representada.

A base foi dividida em conjuntos de treinamento, validação e teste, seguindo uma proporção de 70% (treinamento e validação) / 30% (teste). Aos dados foi aplicada a normalização *z-score* e após obter as características estatísticas foi aplicado PCA [28], com um número mínimo de componentes que garantisse 99% da variância explicada.

Foi realizado um treinamento de um modelo do tipo *Random Forest*, com uma validação cruzada [29] usando $k = 5$ grupos. Para evitar contaminação durante este procedimento, todas as amostras de um determinado evento pertencem ao mesmo grupo, mantendo as proporções de classe. O modelo proposto foi avaliado com 50 rodadas utilizando um otimizador bayesiano para busca dos melhores hiperparâmetros, testando um total de 2500 conjuntos. O resultado final foi um sistema de classificação de falhas alcançando uma acurácia geral de 94% no caso multiclasse.

2.5.2 TURAN e JASCHKE

O trabalho de TURAN e JASCHKE [3] seguiu caminho similar ao de MARINS *et al.* [2], com algumas diferenças:

- Foram excluídos os sinais provenientes dos sensores P-CKGL e T- CKGL, pois estes apresentavam alto número de amostras nulas. Além disso, para reduzir o tamanho e ruído dos dados, optaram por fazer uma amostragem das entradas na série temporal de 1 segundo para 10 segundos, utilizando uma média sobre o período;

- Na etapa de extração de características, foi utilizado o pacote TSFRESH [30] para calcular os seguintes atributos de uma janela: média, variância, assimetria, curtose, máximo, mínimo, mediana, quartil, coeficientes de variação da média móvel e segunda derivada da média, os coeficientes de um modelo linear e de um modelo polinomial de terceiro grau, e, por fim, a transformada de Fourier do segmento;

- Na seleção de características, foi apontado que preservando 99% de variância do conjunto de treinamento usando PCA [28], a dimensionalidade não é reduzida e mesmo assim o desempenho é inferior quando comparado com o conjunto sem redução;

- Foram considerados seis modelos de classificação: regressão logística, máquinas de vetores de suporte (*support vector machine*, SVM) [31], análise discriminante linear e quadrática [29], árvore de decisão, *Random Forest* e ADABOOST [32].

No treinamento, seguiram a mesma metodologia adotada por MARINS *et al.* [2], com respeito à separação de eventos de treinamento e teste (70% / 30%), e validação cruzada com cinco grupos. Usaram a estratégia de um-versus-resto para os modelos que não são multiclasse.

Segundo os autores, o classificador de árvores de decisão foi o que obteve melhor resultado na validação cruzada, sem utilizar um método de seleção de características, considerando uma menor janela deslizante do classificador (600 s). Neste modelo foram obtidos os valores de 0,94 de acurácia balanceada [22], e 0,91 de pontuação macro-F1.

2.5.3 GATTA et al.

Neste estudo, no pré-processamento do 3W, fase de extração de características, GATTA *et al.* [4] comparam o uso de características estatísticas, com o uso de um *Autoencoder* (aprendizagem profunda), que emprega um modelo conforme LI *et al.* [33]. No método de extração de características, foi adotada a mesma metodologia de 9 indicadores estatísticos explorada por MARINS *et al.* [2] e o *Autencoder* foi composto por duas camadas convolucionais 1D de codificação e duas camadas deconvolucionais 1D de decodificação.

Nesta abordagem, foi dividida cada série temporal em janelas deslizantes. O comprimento de cada janela é fixo. Foram realizados experimentos com observações de 301, 451 e 601 segundos. Foram ainda utilizadas as 3 fases de cada instância (normal, transiente e falha) para a composição da janela deslizante, não fazendo distinção entre transiente e falha. Também foi descartada uma variável do conjunto de dados, T-CKGL, por conter muitos dados nulos.

Para a comparação são usadas as fases de extração de característica, e a fase de treinamento, com aplicação de 4 classificadores: florestas aleatórias, *K-nearest neighbors* (KNN) [29], Classificação linear (*Gaussian Naive Bayes*, GNB) e discriminante quadrático (*Quadratic Discriminant Analysis*, QDA).

As características do *Autoencoder* mostram que as medidas parecem ser tão melhores quanto maiores são as janelas de tempo. Diferente dos trabalhos anteriores,

os autores optaram por separar os dados em treino/validação/teste seguindo a proporção 60%/20%/20%.

Os melhores resultados para o experimento multiclasse, foram alcançados com os atributos do *Autoencoder* combinados com o classificador de *Random Forest*, com a janela deslizante de 601 s, resultando em 0,795 de acurácia e 0,898 de F1.

2.5.4 MACHADO-2022 et al.

Neste artigo de MACHADO *et al.* [34] é apresentada uma metodologia para aperfeiçoar o desempenho de um classificador de uma classe (*one class classifier*, OCC) [29] usando dois métodos de aprendizagem não supervisionados, o autoencoder de memória de longo e curto prazos (*long short-term memory*, LSTM) [35] e o SVM de uma classe. Somente dois tipos de falhas com diferentes dinâmicas são analisados: Fechamento Espúrio de DHSV (Classe 2) e Hidrato na Linha de Produção (Classe 8).

No pré-processamento, foram eliminadas variáveis que apresentavam mais do que 18% dos dados perdidos (nulos ou NaN), resultando na seleção de somente 3 sensores para as Classe 2 e 5 variáveis para a Classe 8. Além disso, separaram o conjunto de dados para treinamento/teste em proporções diferentes nas duas classes: os dados da Classe 2 foram separados em 80%/20% e os da Classe 8 em 50%/50%. Os dados de transitório e de regime permanente de falha combinados em um único rótulo.

Um critério de deslocamento no tempo (antecipação) de mudança entre os rótulos de normal e falha foi proposto para melhorar a discriminação de dados normais e defeituosos durante a fase de treinamento, com o objetivo de reduzir o erro de reconstrução do LSTM. Foi selecionada uma das variáveis de entrada como referência para ser efetuado o deslocamento, onde o melhor momento foi calculado por uma equação estabelecida pelo autores. Para a Classe 2, foi selecionada a Variável "Temperatura no TPT" e para a Classe 8, "Pressão a Montante do CKP". Uma vez selecionada a variável, foi usada a variação da percentagem da variável em relação ao seu estado com o rótulo normal, onde os autores chamaram este limiar de y, para escolher o deslocamento ótimo do rótulo. No procedimento experimental, o desempenho dos dois classificadores foi avaliado para diferentes deslocamentos normalizados $y \in]0, 1]$.

Para se obter o modelo do autoencoder LSTM, os dados foram convertidos em um matriz 3-D no formato de amostras, intervalo de tempo e variáveis. O tamanho da janela (intervalo de tempo) foi de 10 amostras no tempo.

O melhor resultado para as Classes 2 e 8 foi obtido com o LSTM. Para a Classe 2, usando $y = 0{,}6$, foram obtidos os valores de 0,999 de acurácia e de 0,936 de F1. Para a Classe 8, usando $y = 0{,}9$, foram obtidos os valores de 0,966 de acurácia e 0,953 de F1.

2.5.5 ARANHA et al.

Os autores [36] apresentam uma proposta de CBM, composto por um OCC, que usa uma combinação de *autoencoder* LSTM com uma abordagem analítica baseada em regras sobre os dados dos sensores, denominada pelos autores diagrama de decisão (DD) e relatam que foi testada em 3 poços do pré-sal que operam na Bacia de Santos (SP, Brasil).

No caso do treinamento do DD, apresentam como exemplo: se as válvulas M1/DHSV deveriam estar abertas de acordo com o sistema de monitoramento, mas a pressão no PDG aumenta enquanto a pressão no TPT diminui, o DD alerta para uma anomalia. Outro cenário apresentado de indicação de anomalia é quando há um aumento repentino e simultâneo nas leituras de pressão dos sensores TPT e PDG. O módulo entra no modo de treinamento após adquirir aproximadamente 500 pontos de dados normais, o que cobre cerca de 5.000 segundos de produção em tempo real, quantidade determinada empregando testes com dados do setor.

No *autoencoder* LSTM, usando uma abordagem semelhante, o treinamento da rede profunda utiliza 500 instâncias de dados rotulados. Os autores observam que considerando que o monitoramento é realizado em tempo real, e as condições de operação de poço estão constantemente sendo alteradas, novo treinamento é realizado, no sistema em operação testado por eles, cada vez que 500 novas instâncias são salvas. Este número de instâncias requeridas para novo treinamento foi determinado, via testes de hiperparâmetros sobre dados reais, focando em alcançar uma quantidade suficiente de dados que minimize a hipersensibilidade de resultados, sem penalizar os recursos computacionais.

Também comparam o resultado do modelo proposto com os resultados dos trabalhos de MARINS *et al.* [2] (*Random Forest*), TURAN e JASCHKE [3](árvore de decisão) e MACHADO *et al.* [34] (LSTM), exclusivamente sobre a Classe 2 (Fechamento Espúrio de DHSV), testando o conjugado de LSTM com o DD, onde os dados dos sensores utilizados foram as variáveis do 3W. Nesta análise, obtiveram com o modelo conjugado os valores de 0,989 de acurácia e 0,992 de F1.

2.5.6 MACHADO-2024 et al.

Os autores [37] adotaram como estratégia de classificação a OCC, empregando como medida de similaridade a *Dynamic Time Warping* (DTW) [38] combinada com o agrupamento *k-means*, para em seguida, as variáveis e instâncias agrupadas serem aplicadas ao classificador LSTM. No trabalho, somente as Classes 1, 2 e 8 foram avaliadas em detalhe. MACHADO *et al.* [37] citam TAVENARD *et al.* [39] como referência da combinação de DTW com *K-means* e KEOGH e LIN [40] como referência do agrupamento de séries temporais.

Foram utilizadas as 3 fases de cada instância (normal, transiente e falha) mas não houve distinção entre transiente e falha, e somente 5 variáveis foram usadas para desenvolver o modelo: P-PDG, P-TPT, T-TPT, P-MON-CKP E T-JUS-CKP.

O agrupamento *k-means* foi empregado para cada uma das 5 variáveis, onde foi calculado o baricentro de cada variável da classe, na série temporal, usando o DTW como medida (*DTW Barycenter Average*, DBA). O algoritmo do *k-means* foi configurado com 2 *clusters* (grupos), onde a medida DTW foi aplicada primeiramente para separar a série temporal em instâncias de treinamento dentro dos dois grupos para cada variável, baseado na similaridade de cada um. Cada *cluster* obtido no final do processo, foi composto pelas instâncias que continham intersecção de variáveis dos *subclusters*.

Neste trabalho não foi estudado sobre a melhor quantidade de grupos para agrupamento, empregando uma medida de avaliação de agrupamento, isto é, não foi testado se com mais de 2 grupos, seria gerado resultado numericamente superior ao obtido com a escolha.

Segundo MACHADO *et al.* [37], "Deve-se ter cuidado ao comparar o desempenho de classificadores de classe única e multiclasse, uma vez que os dados usados para treinamento e teste nesses métodos são bastante diferentes (Hempstalk & Eibe, 2008)".

Considerando o parágrafo anterior, os resultados alcançados por este artigo não podem ser comparados aos anteriores. Além disso, os autores também não avaliaram o desempenho com medida de acurácia. Entretanto, são apresentados alguns dos resultados encontrados, com os melhores resultados de F1 nos testes:

- Para a Classe 1, no *cluster* M_3, composto por fusão das instâncias de dois *clusters* (M_1 e M_2), totalizando 90% das instâncias selecionadas para a validação e 89,17% das instâncias selecionadas para o teste, foi obtido 0,69;

- Para a Classe 2, no *cluster* M_1, totalizando 76,44% das instâncias selecionadas para a validação e 89,17% das instâncias selecionadas para o teste, foi obtido 0,76;

- Para a Classe 8, no *cluster* M_1, totalizando 77,83% das instâncias selecionadas para a validação e 74,86% das instâncias selecionadas para o teste, foi obtido 0,80.

No Capítulo 3, é aprofundada a questão da melhor quantidade de grupos na Classe 1 e as possíveis causas de haver grupos com características distintas na referida classe.

2.5.7 DIAS et al.

Neste trabalho foi apresentada a plataforma MAIS [41], e foram utilizados os mesmos critérios preconizados por MARINS *et al.* [2], com objetivos de padronização quanto à [1]:

- Seleção das instâncias e classes a serem pesquisadas, onde as instâncias desenhadas foram ignoradas e a Classe 7 foi desconsiderada por estar subrepresentada;
- Uso de parâmetro para pesquisar o balanceamento entre as amostras extraídas das instâncias da Classe 0 e as amostras da fase normal inicial nas instâncias com falha, visando pesquisar os desafios (i) e (iv) apresentados na Seção 2.4;
- Uso da normalização *z-score* e seleção de características com PCA [28];
- Separação dos dados para treinamento, validação e testes (50% / 20% / 30%); e a estrutura adotada de validação cruzada no treinamento, com 5 grupos.

Além de utilizar as fases normal e transitória do evento, foi utilizada a fase de falha para treinar o classificador.

Além do método de extração de características estatísticas, conforme apresentado por MARINS *et al.* [2], os autores apresentaram o método baseado em *wavelets* [18], que fornece análise tempo-frequência multiescala. Os desafios impostos pelo conjunto de dados 3W, apontados na Seção 2.2, são abordados combinando os recursos *wavelet* e estatísticos, resultando, na avaliação do desempenho do classificador no problema multiclasse, em acurácia balanceada (macro-*recall*) [22] de 0,986 e pontuação macro-F1 de 0,987.

Neste trabalho foi utilizado o LGBM [42] como técnica de classificação. O modelo proposto foi avaliado com 100 rodadas utilizando o *tree-structured Parzen estimators* (TPE)[43] para busca dos melhores hiperparâmetros, testando um total de 5.000 conjuntos.

Os autores apresentam matriz de confusão dos modelos testados e a matriz de confusão das anomalias agrupadas, visando a identificação de falso negativo de anomalias.

Na próxima seção são apresentados alguns detalhes do classificador, além de outros da plataforma MAIS.

2.6 MAIS

Conforme introduzido, nesta são descritas as características do MAIS (*Modular Artificial Intelligence System*) [41].

O MAIS é um pacote de software, escrito em Python 3. É uma estrutura de código aberto, modular, flexível e escalável, para prototipagem de classificador de falhas usando técnicas de aprendizagem de máquina.

Para encontrar os hiperparâmetros ideais, é empregada uma pesquisa baseada em *tree-structured Parzen estimators* (TPE) [43], onde o alvo de otimização é a média de medida de avaliação de desempenho encontrada na validação cruzada em 5 grupos, estratificados do subconjunto de treinamento.

Internamente, para cada pasta, a partição de treinamento é configurada com Parada Antecipada[1] (*Early Stopping*) com uma taxa de 0,2:0,8, onde a menor partição é reservada para a parada antecipada. Todas as etapas de particionamento são estratificadas por grupo de eventos, de modo que a distribuição das fontes de eventos seja tão consistente quanto possível em todos os subconjuntos.

Depois que o melhor conjunto de parâmetros é encontrado, o classificador é mais uma vez treinado com todo o conjunto de dados de treinamento (70%), e testado com os 30% restantes para avaliação dos resultados, conforme indicado por ALPAYDIN [29].

Na Figura 2.2 é apresentado o arcabouço do sistema.

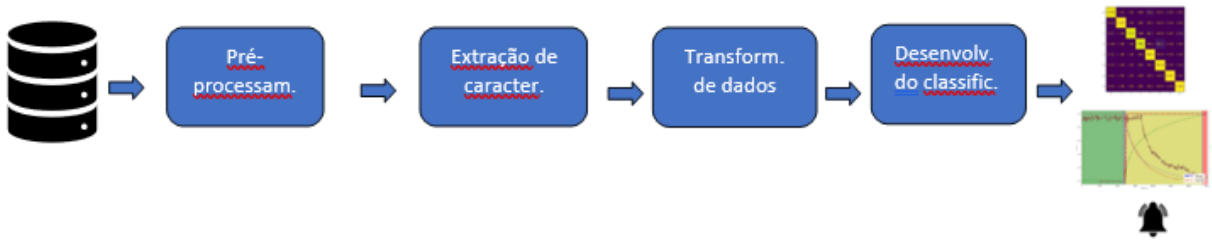

Figura 2.2: Arcabouço do MAIS.

2.6.1 Módulos do Pacote

O sistema está disponível em um arcabouço, com módulos independentes desempenhando o papel das seguintes técnicas e processos:

- Balanceamento das amostras normais da Classe 0 e das amostras normais que antecedem as amostras de falha. Considerando a mesma abordagem realizada

[1] A parada antecipada [28], é uma técnica utilizada em modelos de aprendizado de máquina e redes neurais para prevenir o superajuste (*overfitting*) durante o treinamento. Durante o treinamento de modelos em problemas de otimização de aprendizado supervisionado, o erro no conjunto de treinamento tende a diminuir à medida que o número de épocas (*epochs*) aumenta. No entanto, o erro no conjunto de validação pode inicialmente diminuir, mas eventualmente começa a aumentar devido ao superajuste. A parada antecipada monitora o erro no conjunto de validação e interrompe o treinamento quando o desempenho no conjunto deixa de melhorar, indicando que o modelo está começando a se ajustar excessivamente aos dados de treinamento.

por MARINS *et al.* [2], visto que o conjunto de dados original apresenta $n_0 \gg n_N$ (desafios *i* e *iv* da Seção 2.2), a população de segmentos n_0 é subamostrada tal que $n_0 = b$ x n_N, e o valor de b é encontrado em por meio da busca de hiperparâmetros, dentro de um intervalo de 1 a 10, números inteiros;

- Extração de janelas deslizantes. O tamanho da janela é um parâmetro de busca, cuja técnica depende do tipo de característica a ser extraída. Se forem estatísticas, é selecionada uma janela com tamanho entre 100 e 1.000 amostras; Se forem características baseadas em wavelets (atributos multiescala de frequência de tempo) [18], o número de escalas *wavelet*, K, determina o tamanho da janela, onde $N = 2^K$. O valor de K é encontrado por meio da busca de dentro de um intervalo de 4 a 10, números inteiros, em ordem para lidar com a dinâmica diversa do 3W, conforme mencionado no desafio *iii* da Seção 2.2;

- Transformação dos dados. É realizada a padronização dos dados de treinamento, removendo a média e dimensionando para a variância unitária;

- Pré-Processamento dos dados. Substituição de dados faltantes ou NaN com valores médios do sensor;

- Extração de características. Características estatísticas, conforme utilizadas por MARINS *et al.* [2], onde o montante de componentes principais encontradas, s, é alguma fração da variância explicada do sinal, onde $0 \leq s \leq 1$, e o valor de s é encontrado por meio da busca de parâmetro, dentro de um intervalo [0,9 , 1.0), números reais; e podem ser acrescentadas características baseadas em wavelets;

- Redução de dimensionalidade. É ofertada a possibilidade de usar PCA [28], onde o montante de componentes principais encontradas, s, é alguma fração da variância explicada do sinal, onde $0 < s \leq 1$, e o valor de s é encontrado por meio da busca de parâmetro, dentro de um intervalo [0,9 , 1.0), números reais; Ou um método de seleção de características com base no critério de impureza de Gini (avaliação de importância da característica), empregando um estimador baseado em uma *Random Forest* com 100 árvores, onde o valor da fração de subamostragem de recurso, também é encontrado por meio da busca de parâmetro, dentro de um intervalo de [0,1, 1,0];

- Desenvolvimento do classificador. É utilizado LGBM [42], com experimentos do tipo multiclasse, que considera um único sistema para discriminar todas as diferentes classes do 3W individualmente, configurado com 500 árvores, onde outro conjunto de parâmetros pode ser ajustado: fração de subamostragem de

um conjunto de observações, pesquisada dentro do intervalo [0,1, 1,0]; fração de subamostragem de recurso, entre 0,1 e 1,0; pesos de regularização *L1* ($\lambda 1$) e *L2* ($\lambda 2$), ambos os parâmetros entre 10^{-5} e 10 e, finalmente, o número máximo de folhas por árvore, M_l, tal que M_l em $\{4, 8, ..., 128\}$;

- Avaliação de desempenho. Na plataforma, o usuário pode escolher uma medida de avaliação de desempenho como alvo de otimização, dentre 13 medidas [1]: acurácia, acurácia balanceada, precisão, *recall*, pontuação F1, macro-precisão, macro-*recall*, macro-F1, micro-precisão, micro- *recall*, micro-F1, precisão ponderada, *recall* ponderada, F1 ponderada.

A adoção do LGBM propicia uma aceleração do treinamento em relação ao *Gradient Boost Decision Tree* (GBDT) [44] convencional [2]. No LGBM são usadas duas técnicas para otimização do tempo [42]:

- Amostragem *One-Side* baseada em Gradiente (*Gradient-based One-Side Sampling*, GOSS). As amostras de dados com pequenos gradientes são excluídas e é usado o restante para estimar o ganho de informação;

- Agrupamento de Recursos Exclusivos *(Exclusive Feature Bundling* - EFB). São agrupadas as características que são mutuamente exclusivas, com objetivo de redução de dimensionalidade, usando um algoritmo ganancioso [3] [45].

A arquitetura em módulos permite o pesquisador escolher se será considerado o período transitório do sinal, ou se são consideradas as amostras do início de uma instância (que normalmente não são com falha).

A plataforma dispõe de recursos de armazenamento dos hiperparâmetros, dos resultados conforme figuras de mérito escolhidas para análise, de matrizes de confusão de treinamento, validação e teste do sistema multiclasse, no *cluster* de reprodução.

2.7 Revisão Bibliográfica Acerca de BSW

O BSW (*Base Sediments and Water*) é um indicador, utilizado na engenharia de petróleo, de característica do poço e de estágio da vida produtiva. Ele é o quociente

[2] Ambos são modelos *ensemble* de árvores de decisão, porém no GBDT elas são treinadas em sequência, onde em cada iteração treina as árvores ajustando o erro residual. Implementações convencionais de GBDT precisam, para cada característica, escanear todas as instâncias de dados para estimar o ganho de informação de todos os possíveis pontos de divisão. Portanto, suas complexidades computacionais serão proporcionais ao número de características e ao número de instâncias. Isso torna essas implementações muito demoradas ao lidar com *big data*. KE *et al.* [42] publicaram que os experimentos em vários conjuntos de dados públicos mostram que o LightGBM pode acelerar o processo de treinamento em até 20 vezes, alcançando quase a mesma precisão que no GBDT convencional.

[3] Os algoritmos gananciosos (*greedy algorithms*) são estratégias utilizadas para resolver problemas de otimização, tomando decisões localmente ótimas em cada etapa na esperança de encontrar uma solução globalmente ótima [45].

entre a vazão de água mais os sedimentos que estão sendo produzidos e a vazão total de líquidos e sedimentos de um poço, ambos medidos sob condições normais de temperatura e pressão [46]. Esses componentes indesejados (água e sedimentos) podem se acumular no fundo dos tanques de armazenamento e nos equipamentos de processamento, impactando a eficiência das operações e a qualidade do petróleo produzido.

Analisando principais causas de de aumento de BSW em poços de petróleo, assim está consolidado na literatura:

- Condições do Reservatório: BSW pode ser causado por variações na permeabilidade do reservatório, pressões do reservatório, conificações de água, migração de fluidos de injeção, e a natureza das camadas de petróleo e água;
- Problemas Operacionais: Problemas como fissuras no *casing*, falhas no controle de água, e degradação de materiais;
- Erosão e Corrosão: Essas são outras causas que podem resultar em intrusão de água no poço, aumentando o BSW.

Durante o ciclo de vida de um poço, há expectativa de aumento do BSW, devido ao aumento de água produzida de reservatórios aquíferos naturais ou injeção para reduzir a queda de produção [9].

No trabalho de YOSHIOKA *et al.* [47] os autores afirmam que registros de temperatura têm sido usados para localizar entradas de água em poços de petróleo, e sua identificação é frequentemente realizada por intuição. Em poços horizontais de produção a temperatura dos fluidos não é afetada por mudanças de temperatura geotérmicas e a diferença primária de cada fase é causada por efeitos friccionais. Em outras palavras, entrada de água resulta em aquecimento ou resfriamento do poço. Estas inferências são qualitativas, um vez que não há meios de determinar o fluxo de água. A entrada de água aquecida é um resultado de fluxo de água de um aquífero localizado abaixo da zona de produção (conificações de água). Em compensação, água produzida pode ser mais fria do que o óleo produzido, devido a diferenças nas propriedades termais dos referidos fluidos.

Em contrapartida, em poços verticais a temperatura não é constante ao longo do poço, pois é afetada por mudanças geotérmicas [47].

2.8 Conclusões

Neste capítulo foram apresentados o programa CBM, algumas características do conjunto de dados 3W, seguidos pelos trabalhos publicados em periódicos, com pesquisas dedicadas ao conjunto de dados 3W, de MARINS *et al.* [2], TURAN e

JASCHKE [3], GATTA *et al.* [4], MACHADO *et al.* [34], ARANHA *et al.* [36], MACHADO *et al.* [37], e DIAS *et al.* [1]. Estes trabalhos exibiram resultados com diferentes modelos de extração de características e classificadores.

Observa-se que somente os trabalhos de MARINS *et al.* [2] e DIAS *et al.* [1] utilizaram classificador multiclasse, sobre todos os dados, exceto Classe 7 (Incrustação no CKP). TURAN e JASCHKE [3] adotaram multiclasse com este conjunto, porém excluíram duas variáveis (sensores P-CKGL e T- CKGL). GATTA *et al.* [4] empregaram multiclasse, porém excluíram a variável T-CKGL.

Os demais trabalhos usaram a abordagem de classificador OCC, porém restringindo a base de pesquisa:

- MACHADO *et al.* [34] classificaram sobre somente dois tipos de falhas: Fechamento Espúrio de DHSV (Classe 2) e Hidrato na Linha de Produção (Classe 8). Contudo eliminaram mais variáveis, resultando na seleção de somente 3 sensores para as Classe 2 e 5 para Classe 8;
- ARANHA *et al.* [36], pesquisaram sobre exclusivamente a Classe 2;
- MACHADO *et al.* [37], somente as Classes 1, 2 e 8 foram avaliadas em detalhe, descartando 3 variáveis dos dados brutos.

Destaca-se que esta falta de similaridade na seleção de classes e nos métodos de classificação dificulta muito uma comparação de modelos e resultados. Segundo ALPAYDIN [29], quando comparando dois algoritmos, ambos devem ser investigados com igual cuidado na execução. Tendo em vista estas peculiaridades, resume-se na Tabela 2.2 os resultados dos trabalhos multiclasse de MARINS *et al.* [2], GATTA *et al.* [4] e DIAS *et al.* [1], quanto às figuras de mérito de acurácia e F1.

Tabela 2.2: Medidas obtidas nos trabalhos de MARINS *et al.* [2], TURAN e JASCHKE [3], GATTA *et al.* [4] e DIAS *et al.* [1].

-	Acurácia (ou acurácia balanceada*)	F1 (ou macro-F1**)
MARINS *et al.* [2]	0,940	-
TURAN e JASCHKE [3]	0,940(*)	0,910 (**)
GATTA *et al.* [4]	0,795	0,898
DIAS *et al.* [1]	0,986(*)	0,987(**)

Outro aspecto importante é que poucos trabalhos lidam com os problemas do 3W, conforme mencionado nos desafios *i* a *iv* na Seção 2.3. Foram tratadas algumas destas lacunas neste trabalho.

Em seguida, neste capítulo, foi apresentada a descrição da plataforma MAIS, descrevendo todas as funcionalidades desde o carregamento do 3W até o classificador adotado. Esta é uma ferramenta que pode auxiliar na padronização de experimentos, dada a sua construção em *arcabouço*, além dos recursos de armazenamento de hiperparâmetros e resultados, permitindo comparar experimentos com eficiência.

Por último, realizada revisão bibliográfica sobre BSW, com o objetivo de fundamentar análise de causas de possível existência de *clusters* na base de dados.

No próximo capítulo, considerando a exploração das pesquisas aqui apresentadas, e os desafios do *dataset*, são investigadas hipóteses sobre os tipos de amostras, o desbalanceamento do *dataset* e a dinâmica temporal do sinais. Estas hipóteses são formuladas na Seção 3.2, considerando a utilização do MAIS como ferramenta e tendo como referência um experimento padrão do 3W.

Capítulo 3

Análise Crítica da Base de Dados 3W

O Capítulo 2 contextualiza o leitor no problema específico de detecção e classificação de falhas na operação de poços de petróleo, tema central abordado nesta dissertação. O objetivo do Capítulo 3 é explorar questões fundamentais detalhadas na Seção 2.3, referentes ao conjunto de dados 3W. Nesse sentido, são explorados aspectos como os dois tipos de amostras da Classe 0, ampla variação da dinâmica temporal das falhas e o desbalanceamento entre as classes de falha. A análise desses itens visa proporcionar uma compreensão das complexidades envolvidas na aplicação de técnicas de aprendizado de máquina para o monitoramento de poços de petróleo, contribuindo para a fundamentação teórica e prática necessária ao desenvolvimento da dissertação.

A análise critica foi iniciada com a reprodução de experimento, onde identificou-se a necessidade de investigar os métodos de extração e seleção de características, conforme é detalhado na Seção 3.1. Finalmente, após a definição do Experimento de Referência, são detalhadas as análises exploratórias da base 3W, para detecção de falhas de operação de poços de petróleo, na Seção 3.2, cujos resultados são apresentados no Capitulo 4.

3.1 Organização de Experimento de Referência

Configuração do MAIS. Conforme citado anteriormente, neste trabalho foi empregado o MAIS [41] como sistema CBM, visando investigar os desafios encontrados no *dataset* 3W.

Foram mantidos os requisitos de seleção de classes e sensores, e a mesma separação de instâncias de treinamento e teste estabelecidas por MARINS *et al.* [2] e DIAS *et al.* [1] no conjunto de dados.

Configuração adotada:

- Seleção das instâncias e classes a serem pesquisadas, onde as instâncias de-

senhadas foram ignoradas e a Classe 7 foi desconsiderada por estar sub-representada;

- Separação dos dados para treinamento, validação e testes (50% / 20% / 30%); e a estrutura adotada de validação cruzada no treinamento, com 5 grupos, conforme Tabela 3.2.

- Uso da normalização *z-score* nos dados de treinamento;

- Uso de parâmetro para pesquisar o balanceamento entre as amostras extraídas das instâncias da Classe 0 e as amostras da fase normal inicial nas instâncias com falha, visando pesquisar os desafios *i* e *iv* apresentados na Seção 2.4.

Levou-se em conta tanto o período transiente quanto o período permanente de falha do sinal. Para cada janela de análise, o rótulo (atributo Classe) da última amostra foi extraído e utilizado como representativo da classe correspondente. Essa abordagem garante que a classificação reflita o estado final da janela, proporcionando uma representação mais precisa da condição de falha do sistema.

Na Tabela 3.1 são exibidos os tipos de instâncias e períodos de amostras normais e de falha utilizados no treinamento e teste do modelo.

Tabela 3.1: Configuração do Modelo de Referência quanto aos tipos de instâncias e períodos.

Instâncias		Amostras		
Reais	Simuladas	Normais	Regime trans. de falha	Regime perm. de falha
Sim	Sim	Sim	Sim	Sim

Neste trabalho, foram construídas funções adicionais no módulo do MAIS [41], responsável por rotular as amostras das instâncias, com o objetivo de pesquisar questões relativas à dinâmica temporal [1]. No módulo principal, responsável pelo classificador multiclasse, foram criadas funções específicas para publicar os resultados dos experimentos e cenários associados, de treinamento e teste, inclusive os modelos treinados, no repositório MLFLOW [49] [2]. Foi criado, adicionalmente ao pacote de módulos do MAIS, um *notebook* para emissão dos desenhos das matrizes de confusão, matrizes de confusão das anomalias agrupadas, e para os gráficos representativos de evolução temporal dos sensores e das inferências por instância, usadas nos experimentos (alarme de evento), com legendas no idioma português, para uso nesta dissertação.

[1] Os dois módulos citados e o notebook foram publicados no GitHub do autor. [48]

[2] O MLFLOW é uma plataforma de gestão de experimentos para aprendizado de máquina. É uma ferramenta de código aberto que facilita a gestão do ciclo de vida de projetos de aprendizado de máquina. Ele oferece funcionalidades para o rastreamento e organização de experimentos, gestão de modelos e implantação de modelos em ambientes de produção.

O código foi mantido com as mesmas configurações de sementes aleatórias estabelecidas no repositório de UFRJ/COPPE/PEE [41]. Essa abordagem assegura a reprodutibilidade dos resultados e permite uma comparação direta e justa com estudos anteriores que utilizaram o mesmo conjunto de dados e parâmetros experimentais.

Tabela 3.2: Número de instâncias nos conjuntos de treinamento e teste, empregadas no desenvolvimento do sistema CBM proposto.

	Treinamento		Teste		Total
Tipo de evento	Real	Simulada	Real	Simulada	
0. Normal	468	0	129	0	597
1. Aumento abrupto de BSW	3	78	2	36	119
2. Fechamento Espúrio de DHSV	15	11	7	5	38
3. Golfada Severa	22	55	10	19	135
4. Instabilidade de Fluxo	260	0	84	0	344
5. Perda rápida de produtividade	8	340	4	99	451
6. Restrição Rápida no CKP	4	170	2	45	221
8. Hidrato na Linha de Produção	0	56	3	25	84
TOTAL	780	710	241	229	1960

Experimento de Referência. Para garantir a consistência e a eficácia das análises realizadas, um processo padronizado foi adotado, envolvendo a definição cuidadosa de diversos aspectos metodológicos fundamentais.

Primeiramente, foram estabelecidos critérios para a seleção e a separação dos dados destinados ao treinamento e teste dos modelos, usando a versão 1.0.0 do 3W, divulgada em 01 de julho de 2019 [12], primeira versão publicada, descrita no artigo de VARGAS *et al.* [9]. Este trabalho já estava em curso quando do lançamento da versão 2.0.0.

A escolha de um Experimento de Referência foi guiada pela necessidade de validar as hipóteses formuladas ao longo da pesquisa. Esse experimento foi selecionado com base em critérios de relevância, desempenho prévio e adequação aos objetivos da pesquisa.

Iniciando a pesquisa, foi reproduzido o Experimento 1 de DIAS *et al.* [1]. Neste processo, foi identificado que as 3 instâncias reais da Classe 8 não foram consideradas no teste. Com esta correção, conforme ilustrado na Tabela 3.3 e na Figura 3.1, foi observada uma perda no desempenho de teste.

Tabela 3.3: Resultados da reprodução do Experimento 1 de DIAS *et al.* [1].

	Validação		Teste	
	BalAcc	macro-F1	BalAcc	macro-F1
sem 3 instâncias	0,953± 0,018	0,956± 0,017	0,989	0,988
com 3 instâncias	0,955± 0,020	0,958± 0,015	0,976	0,975

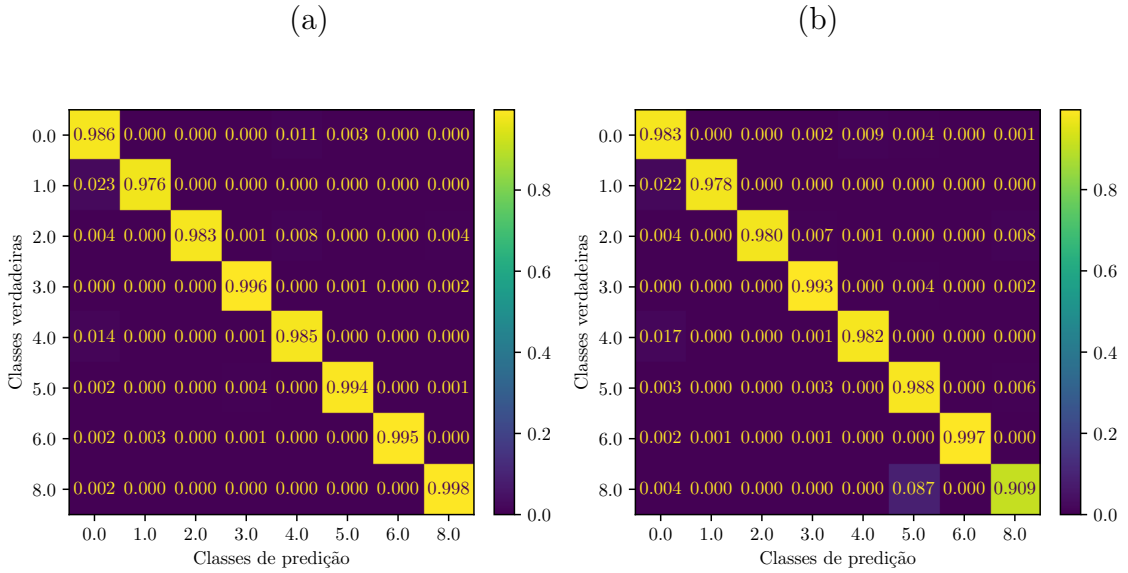

Figura 3.1: Matrizes de confusão de teste na reprodução do Experimento 1 de DIAS *et al.* [1], antes (a) e após inclusão de três instâncias reais (b).

A inclusão das três instâncias reais da Classe 8 ao conjunto de testes revelou dificuldade para classificar amostras das Classes 5 e 8, e a necessidade de refazer a comparação na classificação do *dataset* entre os três tipos de extração de características oferecidos pelo MAIS: estatísticas, baseadas em wavelets e combinadas.

O Experimento foi reproduzido, utilizando um serviço do Jupyter, hospedado em um sistema computacional comercial [3], onde o experimento demandou 23,15 h para completar 44 experiências. Diante disto, o tempo de processamento foi monitorado e utilizado como medida de comparação doravante nos experimentos.

Comparação de classificadores baseados nas características. Ao revisitar os trabalhos de TURAN e JASCHKE [3] e GATTA *et al.* [4], ficou evidente que os melhores resultados em seus modelos foram obtidos quando as características extraídas foram utilizadas sem redução de dimensionalidade. Além disso, conforme mencionado na Seção 2.6.1, o LGBM [42], empregando as técnicas EFB e GOSS, realiza a redução de dimensionalidade, visando diminuir a base de dados e acelerar a classificação.

Para assegurar a padronização no método de busca de hiperparâmetros, conforme adotado por DIAS *et al.* [1], o número de experiências [4] do experimento foi fixado em 100 para cada cenário avaliado. O ponto no espaço de hiperparâmetros que

[3]Google Colab Pro+ (Execução em segundo plano, que permite a execução contínua de código por até 24h e acesso a GPUs, como as NVIDIA Tesla V100 e A100).

[4]O termo experimento é utilizado para referir-se ao conjunto de experiências, culminando com o teste do classificador. E experiência é a implementação do arcabouço desde o pré-processamento até o desenvolvimento do classificador.

obteve o melhor resultado alvo foi selecionado como a configuração ideal, utilizando a técnica baseada em TPE [43], conforme mencionado na Seção 2.5.7. A medida de avaliação de desempenho utilizada para encontrar o melhor conjunto de parâmetros foi a acurácia balanceada (BalAcc) [22].

Em conformidade com os resultados publicados por MARINS *et al.* [2], GATTA *et al.* [4] e DIAS *et al.* [1], neste trabalho foi adotado, além da acurácia balanceada (BalAcc) [22], a medida de avaliação de desempenho macro-F1 para avaliação final. Foram realizados testes com o MAIS [5], comparando três tipos de extração de características: estatísticas, baseadas em wavelets e combinadas, conforme visto na Tabela 3.4. Esses testes foram conduzidos tanto com a aplicação da técnica de redução de dimensionalidade [28] no *arcabouço*, quanto sem sua aplicação.

Tabela 3.4: Resultados dos modelos por método de extração e redução de características, com busca de hiperparâmetros em 100 experiências, considerando cinco grupos de validação cruzada durante busca de hiperparâmetros.

		Validação		Teste	
Com redução de dimensionalidade (PCA) [28]					
Método	Tempo(h)	BalAcc	macro-F1	BalAcc	macro-F1
Estatística	14,3	0,962±0,014	0,970±0,009	0,964	0,965
Baseada em wavelets	14,4	0,983±0,005	0,984±0,004	0,979	0,981
Combinada	17,5	0,955±0,020	0,958±0,015	0,976	0,975
Sem redução de dimensionalidade					
Estatística	8,3	0,974±0,005	0,974±0,006	0,972	0,972
Baseada em wavelets	13,2	0,982±0,007	0,980±0,008	0,975	0,954
Combinada	15,1	0,954±0,034	0,952±0,040	0,975	0,964

[5] Experimentos realizados no *cluster* do SMT, resultando em tempo de processamento inferior ao obtido com o Colab Pro+.

As matrizes de confusão dos modelos com características estatísticas, baseadas em wavelets, e combinadas, todas com PCA, são apresentadas na Figura 3.2.

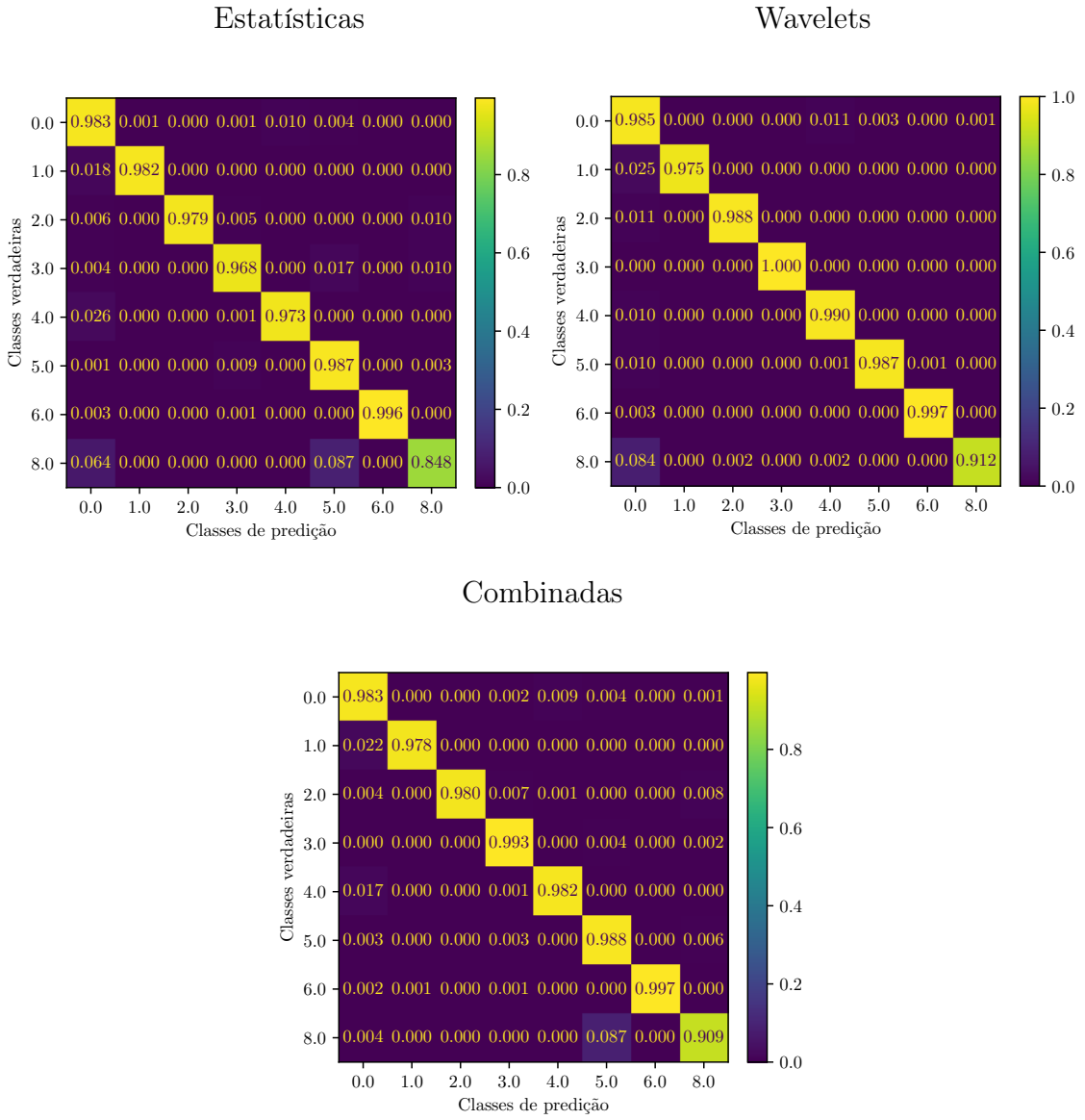

Figura 3.2: Matrizes de confusão dos testes para as configurações avaliadas, com aplicação de PCA, estatística, *wavelet* e combinada.

O modelo com características baseadas em wavelets com PCA apresentou os melhores resultados, especificamente influenciado pela qualidade da classificação da falha 8. Porém este modelo apresenta elevados valores de falsos negativos nas classes 1 (0,025), 2 (0,011) e 5(0,010) e 8 (0,084).

Na Figura 3.3 são apresentadas as matrizes de confusão das anomalias agrupadas, demonstrando os resultados obtidos. Ao analisar os resultados individualizados por classe de falha, constatou-se que o modelo com a menor taxa média de alarmes perdidos (0,007), ou seja, a menor taxa de falsos negativos das anomalias, foi o treinado com características combinadas (estatísticas e baseadas em wavelets) com

PCA. Esses resultados são consistentes com os obtidos por DIAS *et al.* [1]. Portanto, esse modelo foi selecionado como referência para comparações futuras ao longo dessa dissertação.

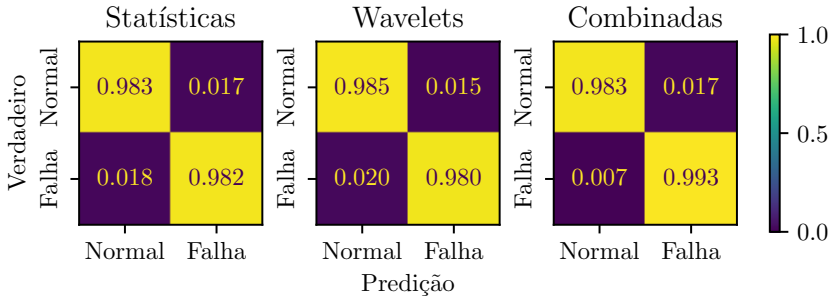

Figura 3.3: Avaliação do conjunto de testes por configuração de extração de característica, tendo as anomalias agrupadas, viabilizando a identificação de falso negativo de anomalias.

Os pontos selecionados no espaço de hiperparâmetros explorado durante o treinamento de cada classificador avaliado estão apresentados na Tabela 3.5. Os parâmetros tamanho da janela deslizante (*window_size*), balanceamento intra-classe 0 (*normal_balance*), *b*, e número de componentes (*n_components*) são específicos da metodologia proposta, e os demais parâmetros são inerentes ao classificador LGBM.

Tabela 3.5: Melhor conjunto de parâmetros encontrado no treinamento de cada classificador conforme tipo de característica.

Parâmetro	Estatístico	Wavelets	Combinado
Window_size	600	1024	1024
Normal_balance (b)	3	1	2
N_components	1	0,992	0,999
Subsample	0.1	0,55	0,1
Feature_fraction	0,25	1	0,95
Num_leaves	114	5	58
Lambda_l1	$7{,}494\,.\,10^{-04}$	0,153	$1{,}854.\,10^{-05}$
Lambda_l2	3,372	7,3	1,313

Observa-se que valores menores do parâmetro de balanceamento intra-classe *(b)* resultam em melhores desempenhos, indicando que amostras normais que precedem eventos de falha têm uma influência positiva. Esta hipótese é investigada mais detalhadamente na Seção 3.2.1.

Nesta configuração, apesar de ser utilizado o PCA [28], o melhor conjunto foi obtido com variância explicada próxima da unidade. O LGBM [42], com o objetivo

de reduzir a base de dados e acelerar a classificação, efetua redução de dimensionalidade nas características, usando EFB (parâmetro *feature_fraction*) e uma redução de amostras, empregando o GOSS (parâmetro *subsample*).

Modelo de referência. Considerando tempo de processamento e resultados de teste final, foi escolhido o modelo para fins de comparação de resultado com as próximas experiências executadas neste trabalho.

Podemos resumir as etapas realizadas no *arcabouço*, complementando as citadas acima, conforme citadas em 2.6.1:

- Balanceamento das amostras normais da Classe 0 e das amostras normais que antecedem as amostras de falha;
- Pré-processamento de dados - substituição de valores perdidos de algumas variáveis por valores médios;
- Extração de características - características combinadas (estatísticas e baseadas em wavelets);
- Transformação de dados - *z-score* [50] e PCA [28];
- Desenvolvimento do classificador LGBM com experimentos do tipo multiclasse.

3.2 Questões Orientadoras da Pesquisa

Objetivos e Perguntas de Pesquisa. Tendo obtido os resultados do Experimento de Referência, foi dado prosseguimento à realização dos testes de hipóteses.

Foram formuladas perguntas e realizados experimentos [5], que levam à exploração de hipóteses, constituindo a principal contribuição deste trabalho:

1. Experimento 1 - Como as observações classificadas como normais, que antecedem as observações de falha em eventos de falha, influenciam na eficiência da classificação quando comparadas com observações de eventos normais?
2. Experimento 2 - Como o classificador se comporta quando as amostras coletadas no regime permanente de falha são excluídas do conjunto de dados?
3. Experimento 3 - Como o sistema classificador se comporta quando é escolhida outra medida de avaliação de desempenho que não seja a acurácia balanceada, tendo em vista o desbalanceamento das classes?
4. Experimento 4 - Qual foi a influência das instâncias simuladas no desempenho do classificador?

5. Experimento 5 - Qual foi a quantidade mínima necessária de experiências que viabilizou uma aceitável detecção de evento não desejável, considerando a padronização de 100 experiências?

6. Experimento 6 - Investigada a existência de agrupamentos indesejados que possam reduzir a eficácia do modelo na identificação e classificação de falhas e as causas para a formação de mais de um grupo em uma dada classe de falha.

Esses seis pontos representam áreas críticas nas quais a pesquisa se concentrou, formulando hipóteses, buscando não apenas abordar os desafios identificados na literatura, mas também contribuir com novas percepções e soluções para o monitoramento eficaz e a gestão de poços de petróleo, empregando técnicas de aprendizado de máquina.

Os dados utilizados nos Experimentos (versão e separação de treinamento/teste) foram os mesmos utilizados no Experimento de Referência, com exceção das exclusões mencionadas em cada experimento.

3.2.1 Experimento 1

Como as observações classificadas como normais, que antecedem as observações de falha em eventos de falha, influenciam na eficiência da classificação quando comparadas com observações de eventos normais?

O conjunto de dados 3W possui instâncias que contêm exclusivamente amostras normais, conforme ilustrado na Figura 3.4 [6], bem como instâncias representativas de anomalias, onde amostras normais antecedem o período de falha transitória, conforme ilustrado na Figura 3.5 [7].

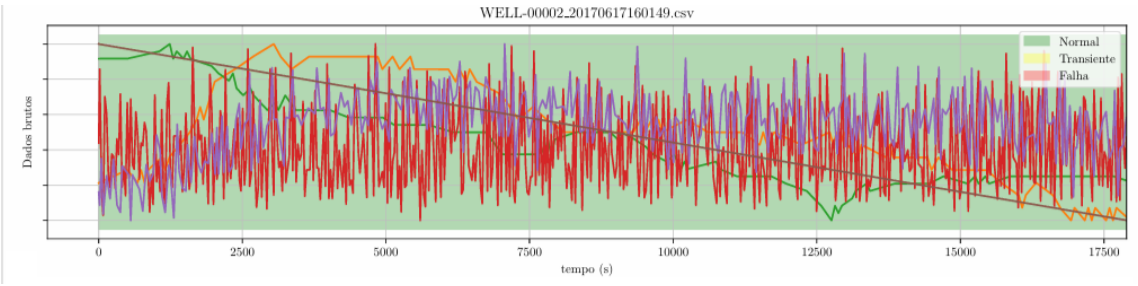

Figura 3.4: Exemplo de instância da Classe 0.

[6] A cor do fundo indica o estágio normal (verde).
[7] A cor do fundo indica os estágios normal (verde), transiente (amarelo), e falha (vermelho).

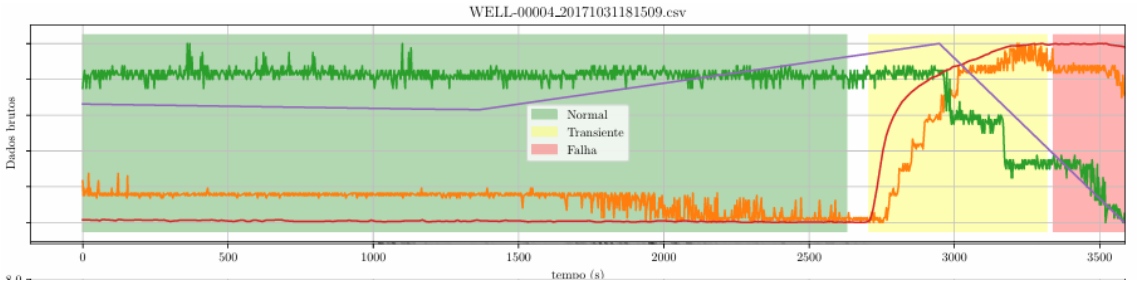

Figura 3.5: Exemplo de instância de falha da Classe 6.

Neste estudo, duas classes foram utilizadas para representar as amostras sem anomalia: a Classe 0, para observações de eventos normais, e a Classe 9, para os inícios normais (das classes de falha), conforme Figura 3.6.

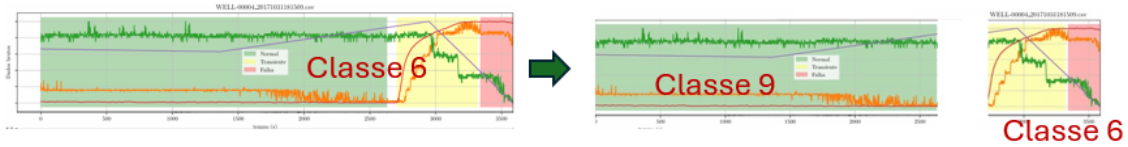

Figura 3.6: Ilustração de separação das amostras normais da instância da Classe 6 – WELL-0004_20171031181509.csv, com substituição do rótulo de 0 para 9.

Esse procedimento visa investigar o impacto dessa separação na eficiência da classificação. Na Figura 3.7 são exibidas as atividades realizadas no experimento.

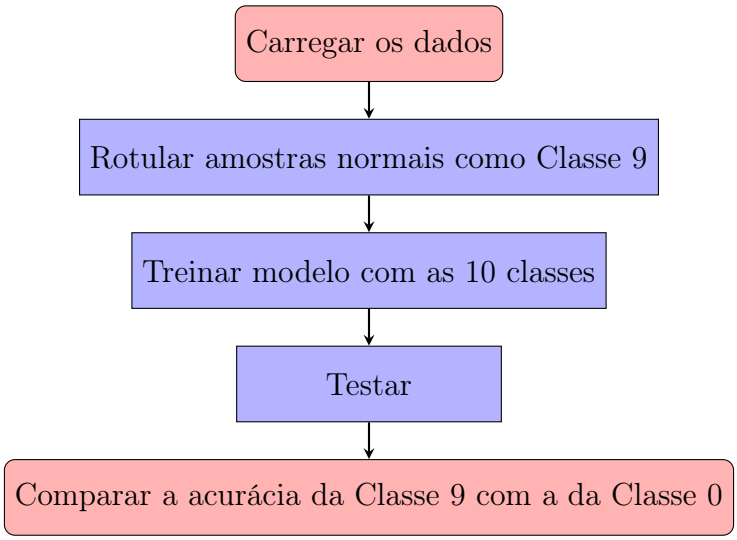

Figura 3.7: Diagrama das atividades realizadas no Experimento.

Na Tabela 3.6 são exibidos os tipos de instâncias e períodos de amostras normais e de falha utilizados no treinamento, validação e teste do modelo do Experimento.

Uma vez que as amostras foram separadas, experimentos com balanceamento e sem balanceamento das amostras normais foram realizados, empregando 100 experiências..

Tabela 3.6: Configuração do Experimento 1 quanto aos tipos de instâncias e períodos.

Fase	Instâncias		Amostras		
	Reais	Simuladas	Normais	Regime trans. de falha	Regime perm. de falha
Trein., valid. e teste	Sim	Sim	Separadas e rotuladas como Classe 9	Sim	Sim

3.2.2 Experimento 2

Como o classificador se comporta quando as amostras coletadas no regime permanente de falha são excluídas do conjunto de dados?

A análise da dinâmica temporal das falhas nos poços de petróleo é importante para a precisão e confiabilidade dos sistemas de monitoramento. Um aspecto particular dessa análise envolve a exclusão das amostras coletadas no regime permanente de falha do conjunto de dados. Em um sistema CBM em operação, cujo objetivo principal é a detecção de falha, não há amostras deste tipo para empregar no detector.

Seleção e exclusão das amostras no treinamento, conforme ilustrado na Figura 3.8. Este estudo visa entender como a exclusão afeta o desempenho do classificador.

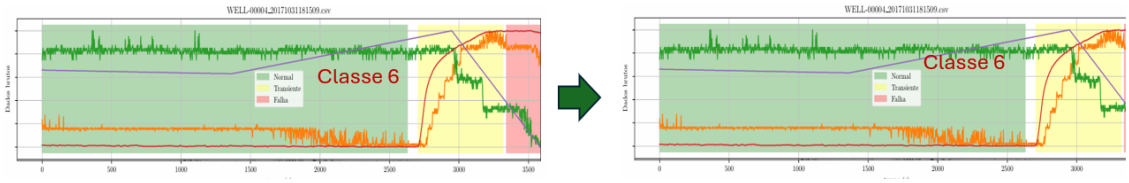

Figura 3.8: Ilustração de cenário de treino e teste sem amostras de falha em regime permanente.

Na Figura 3.9 são exibidas as atividades realizadas no experimento.

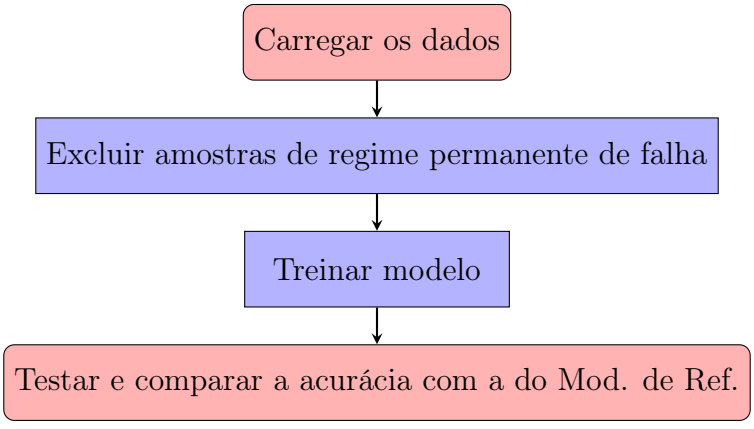

Figura 3.9: Diagrama das atividades realizadas no Experimento.

Na Tabela 3.7 são exibidos os tipos de instâncias e períodos de amostras normais e de falha utilizados no treinamento, validação e teste do modelo do Experimento.

Tabela 3.7: Configuração do Experimento 2 quanto aos tipos de instâncias e períodos.

	Instâncias		Amostras		
Fase	Reais	Simuladas	Normais	Regime trans. de falha	Regime perm. de falha
Trein., valid. e teste	Sim	Sim	Sim	Sim	Excluídas

Dado que os eventos das Classes 3 e 4 contêm exclusivamente amostras em regime permanente de falha, eles foram excluídos deste estudo.

Esse procedimento visa investigar o impacto dessa separação na eficiência da classificação, realizando 100 experiências.

3.2.3 Experimento 3

Como o sistema classificador se comporta quando é escolhida outra medida de avaliação de desempenho que não seja a acurácia balanceada, tendo em vista o desbalanceamento das classes?

Conforme exposto na Seção 2.6, o MAIS [41] permite a seleção da medida alvo de otimização. Este alvo é a média da medida de avaliação de desempenho encontrada na validação cruzada em 5 pastas (*folds*), estratificadas do subconjunto de treinamento. No Experimento de Referência o MAIS foi configurado com a acurácia balanceada [22] como alvo. Após a seleção do melhor conjunto de parâmetros baseado na medida, o *dataset* é novamente treinado com todos os dados do treinamento, e testado com o restante, visando a obtenção do índice final.

Esta experiência visou testar se outra medida de avaliação de desempenho poderia trazer um melhor resultado final nos testes, considerando o grande desbalanceamento entre as classes. Portanto, foi empregada a medida F1-ponderada (*weighted*) [23][24], realizando 100 experiências.

Na Figura 3.10 é ilustrado o processo.

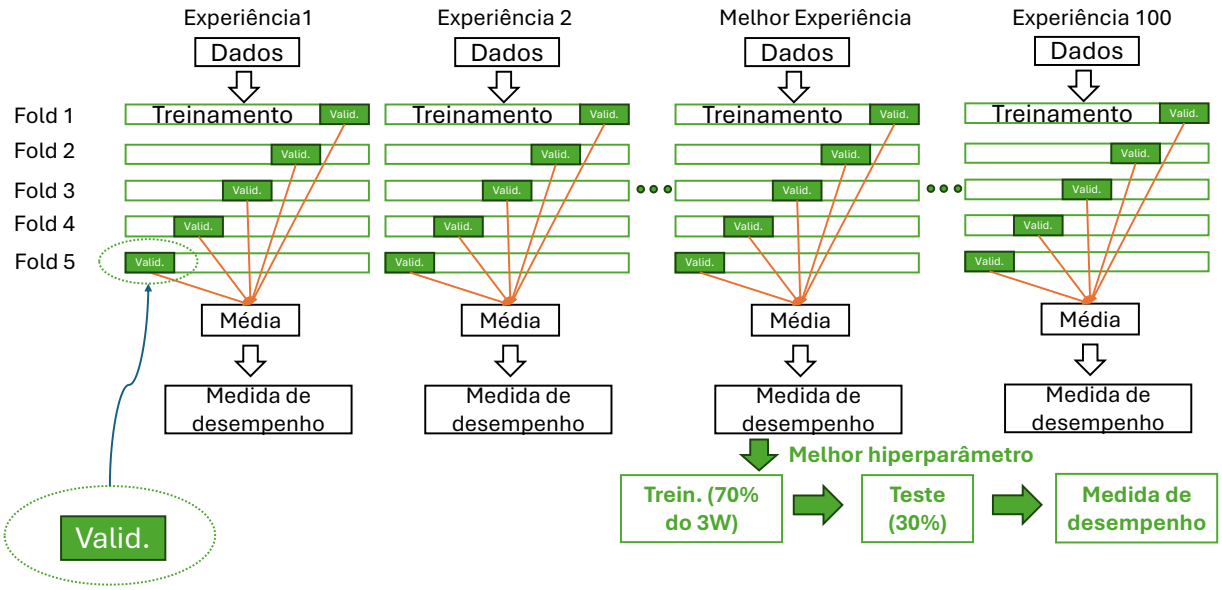

Figura 3.10: Diagrama ilustrando a busca de hiperparâmetros, seleção da experiência com melhor medida de desempenho na validação e teste final do modelo.

Na Figura 3.11 são exibidas as atividades realizadas no experimento.

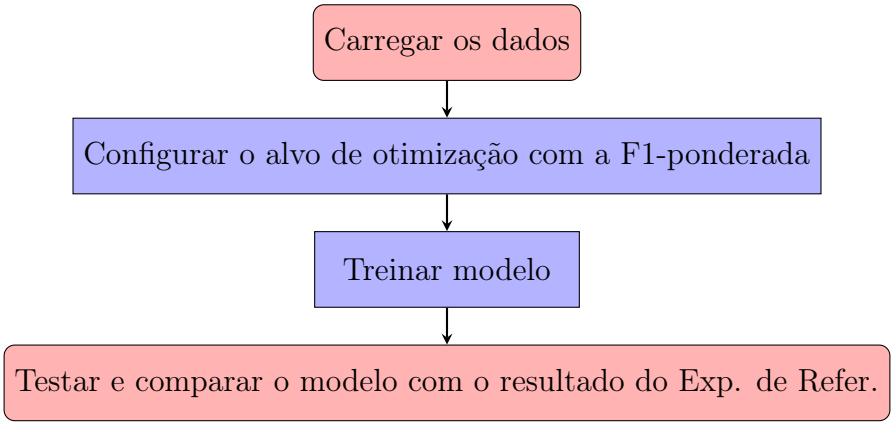

Figura 3.11: Diagrama das atividades realizadas no Experimento.

Na Tabela 3.8 são exibidos os tipos de instâncias e períodos de amostras normais e de falha utilizados no treinamento, validação e teste do modelo do Experimento.

Tabela 3.8: Configuração do Experimento 3 quanto aos tipos de instâncias e períodos.

Instâncias		Amostras		
Reais	Simuladas	Normais	Regime trans. de falha	Regime perm. de falha
Sim	Sim	Sim	Sim	Sim

3.2.4 Experimento 4

Qual foi a influência das instâncias simuladas no desempenho do classificador?

A utilização de eventos simulados é uma prática comum em projetos de aprendizado de máquina, especialmente em cenários onde os dados reais são limitados ou desbalanceados. No contexto da detecção e classificação de falhas em poços de petróleo, a avaliação da influência das instâncias simuladas no desempenho do classificador foi considerada importante para determinar a eficácia e a confiabilidade desses dados.

No 3W, os eventos simulados apresentam uma distribuição, que varia conforme classe. Na Tabela 3.9 é ilustrada essa relação. Observa-se que não há uma regularidade na distribuição.

Tabela 3.9: Porcentagem de instâncias simuladas no *dataset*.

Tipo de evento	% Simuladas
1. Aumento Abrupto de BSW	88
2. Fechamento Espúrio de DHSV	42
3. Golfada Severa	70
4. Instabilidade de Fluxo	0
5. Perda Rápida de Produtividade	97
6. Restrição Rápida no CKP	97
8. Hidrato na Linha de Produção	96

Na Figura 3.12 [8] é ilustrado o comportamento de uma instância real e uma instância simulada da Classe 2, e os respectivos resultados de detecção de falha usando o classificador do Experimento de Referência.

[8]Evolução temporal dos sinais de sensor, o correspondente estado de operação do poço monitorado, e o resultado do modelo treinado com o classificador de referência. No gráfico que ilustra a evolução temporal dos sinais de sensor o fundo verde indica operação regular, o fundo amarelo indica estado transitório de falha, e o vermelho indica estado permanente de falha.

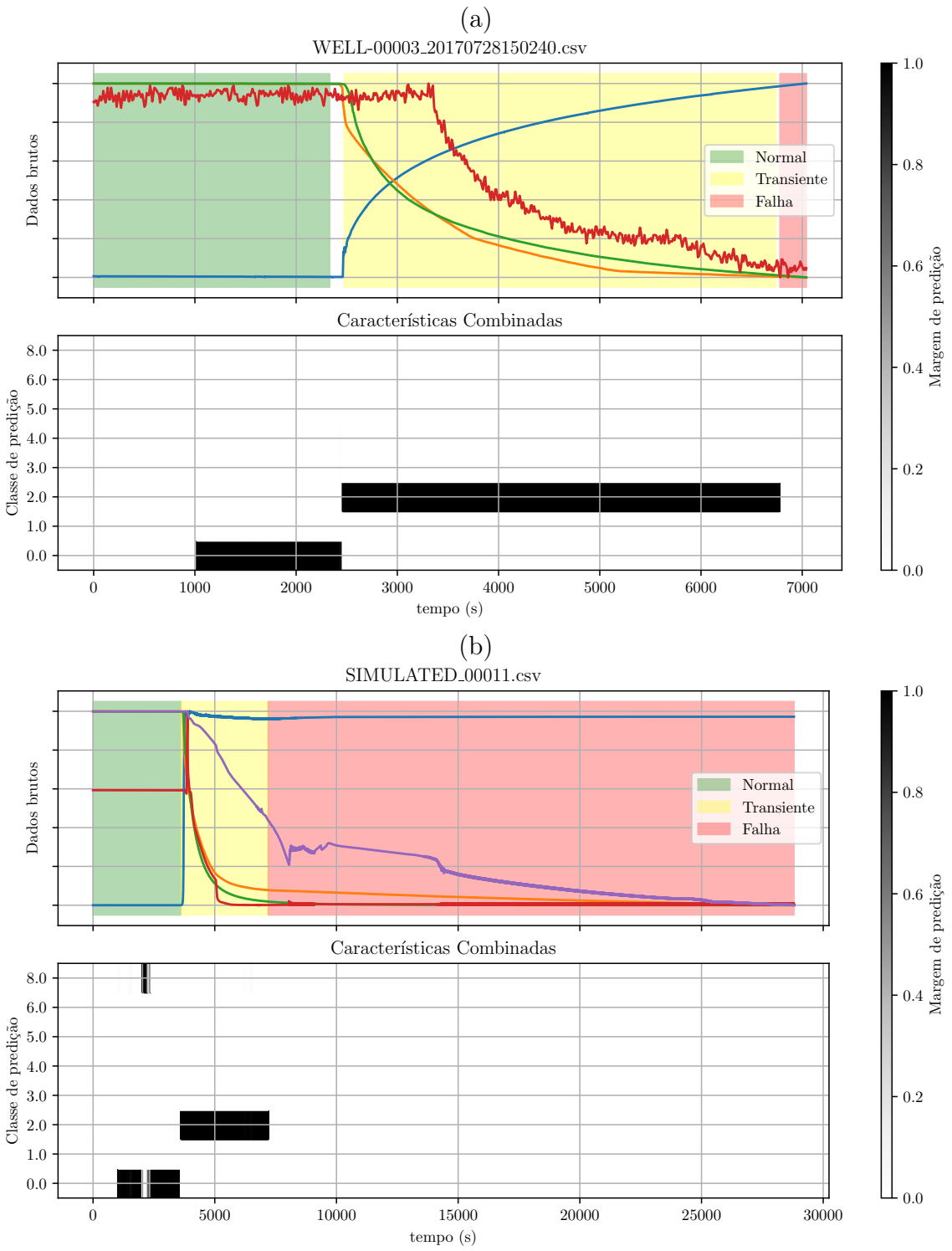

Figura 3.12: Gráficos ilustrando o comportamento de uma instância real (a) e uma instância simulada (b), da Classe 2, e os respectivos resultados de detecção de falha.

Para entender melhor a influência das instâncias simuladas, o desempenho do classificador foi avaliado em dois cenários distintos:

- Exclusão de instâncias simuladas: O treinamento foi realizado apenas com instâncias de eventos reais, conforme Tabela 3.10, onde são exibidos os tipos de

instâncias e períodos de amostras normais e de falha utilizados no treinamento, validação e teste do modelo do Experimento.

Tabela 3.10: Configuração do Cenário 1 quanto aos tipos de instâncias e períodos.

	Instâncias		Amostras		
Fase	Reais	Simuladas	Normais	Regime trans. de falha	Regime perm. de falha
Trein., valid. e teste	Sim	Excluídas	Sim	Sim	Sim

- Inclusão de instâncias simuladas no treinamento, como no Experimento de Referência, conforme Tabela 3.11, onde são exibidos os tipos de instâncias e períodos de amostras normais e de falha utilizados no treinamento, validação e teste do modelo do Experimento.

Tabela 3.11: Configuração do Cenário 2 quanto aos tipos de instâncias e períodos.

	Instâncias		Amostras		
Fase	Reais	Simuladas	Normais	Regime trans. de falha	Regime perm. de falha
Trein. e valid.	Sim	Sim	Sim	Sim	Sim
Teste	Sim	Excluídas	Sim	Sim	Sim

Esse procedimento foi executado em ambos os cenários, utilizando como padrão a realização de 100 experiências.

3.2.5 Experimento 5

Qual foi a quantidade mínima necessária de experiências que viabilizou uma aceitável detecção de evento não desejável, considerando a padronização de 100 experiências?

O processo de busca pelo melhor conjunto de parâmetros e treinamento para o Experimento de Referência, envolvendo 100 experiências, foi longo e exigente, consumindo aproximadamente 17,5 horas de tempo de processamento. Em um estudo analisado por BERGSTRA *et al.* [51], que investigou métodos para otimização de hiperparâmetros, foi observado que um número significativamente menor de experiências pode frequentemente produzir resultados próximos ao desempenho do melhor conjunto de hiperparâmetros identificado. Na Figura 3.13 é ilustrada a pouca evolução do BalAcc na validação, conforme experiências realizadas. Observa-se sutil linha de tendência da função alvo.

O objetivo desta análise é propor um sistema CBM que possa ser treinado com rapidez e apresente confiabilidade para alarmar possíveis falhas em poços em operação.

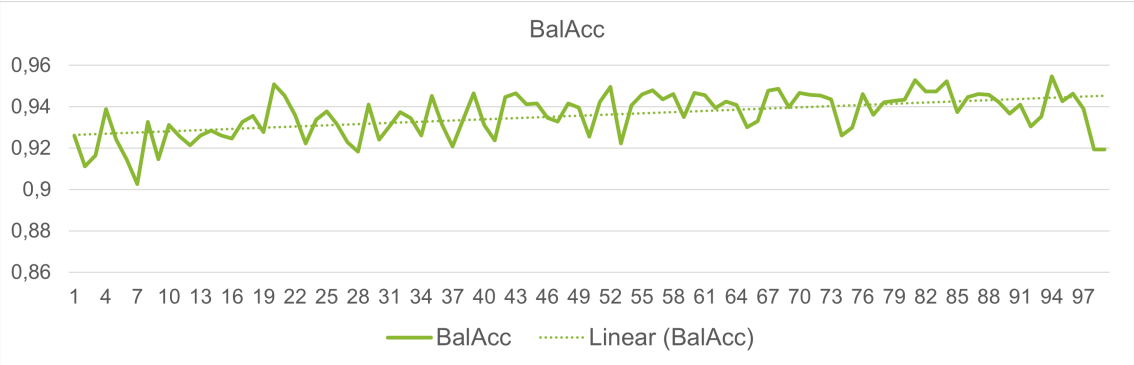

Figura 3.13: Gráfico de evolução do BalAcc em função da experiência de busca de hiperparâmetro.

Segundo as descobertas desses pesquisadores, resultados comparáveis ao desempenho do Experimento de Referência foram alcançados com uma quantidade mínima de experiências variando entre 40 e 60.

Assim, treinamentos adicionais foram conduzidos, incluindo a busca por hiperparâmetros, utilizando conjuntos crescentes de experiências, variando de 10 a 60.

Cada conjunto de experiências foi avaliado não apenas quanto à acurácia balanceada e macro-F1, mas também em termos de tempo de processamento. Essa abordagem sistemática visou não apenas validar a robustez do modelo inicial, mas também otimizar seu desempenho para cenários práticos.

Na Tabela 3.12 são exibidos os tipos de instâncias e períodos de amostras normais e de falha utilizados no treinamento do modelo do Experimento.

Tabela 3.12: Configuração do Experimento 5 quanto aos tipos de instâncias e períodos.

	Instâncias		Amostras		
Fase	Reais	Simuladas	Normais	Regime trans. de falha	Regime perm. de falha
Trein., valid. e teste	Sim	Sim	Sim	Sim	Sim

3.2.6 Experimento 6

O objetivo deste experimento foi investigar a existência de agrupamentos indesejados que possam reduzir a eficácia do modelo na identificação e classificação de falhas e as causas para a formação de mais de um grupo em uma dada classe de falha. No artigo de MACHADO *et al.* [37] foi apresentado que a identificação de *clusters* pode propiciar melhores resultados de classificação. Neste experimento foram avaliados apenas a melhor quantidade de *clusters* e as possíveis causas.

Não foi utilizada a plataforma MAIS [41] como instrumento, tendo em vista que esta análise prescinde de métodos não supervisionados [32] que extrapolam os

recursos do MAIS e não foi realizada a classificação após a análise de agrupamento.

O código publicado por MACHADO *et al.* [37][52] foi reproduzido, com o objetivo de melhor compreender as técnicas empregadas de agrupamento. Algumas classes e funções foram adicionadas ao código mencionado, com o objetivo de avaliar o desempenho dos agrupamentos e gerar os gráficos dos resultados[9].

A Classe 1, Aumento Abrupto de BSW, foi especificamente selecionada para uma análise aprofundada de agrupamento. Nesta abordagem, para implementar os objetivos do experimento, o *k-means* [39][40] foi configurado para encontrar possíveis *clusters*, usando DTW [38] como medida de similaridade.

O desempenho do agrupamento foi avaliado empregando o Coeficiente de Silhueta [25][26]. Após a melhor quantidade de *clusters* ter sido identificada, a variável que apresentou o maior coeficiente foi selecionada e os respectivos resultados obtidos com o agrupamento foram analisados.

A Figura 3.14 ilustra o arcabouço utilizado para empreender a análise.

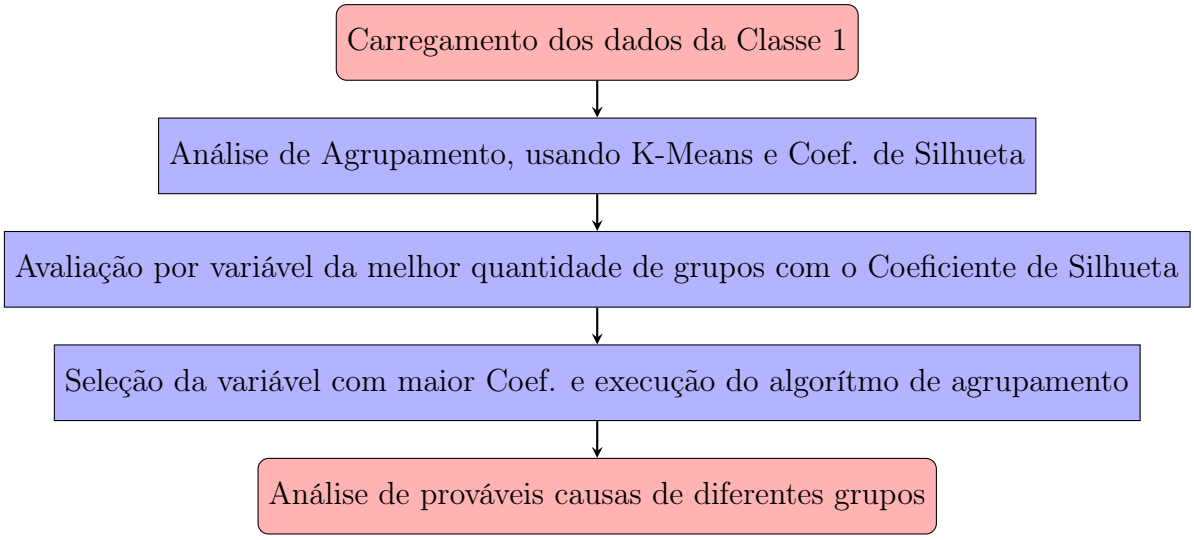

Figura 3.14: Atividades realizadas no experimento.

Inicialmente foi analisada a correlação dos agrupamentos com o tipo de instância de falha, se real ou simulado. Em seguida foram analisadas as possíveis causas de Aumento Abrupto de BSW com os grupos distintos de dados.

3.3 Conclusões

No capítulo anterior foi introduzida a questão da detecção e classificação de falhas na operação de poços de petróleo. Neste capítulo foca-se na exploração dos desafios específicos do conjunto de dados 3W, conforme detalhado na Seção 2.3, abordando

[9] O *notebook* com as inclusões foi publicado no GitHub do autor com o nome C0_Evaluation_-3W_Class1.ipynb [48]

questões como os dois tipos de amostras da Classe 0, a variação da dinâmica temporal e o desbalanceamento das classes de falha.

Na pesquisa é procurado esclarecer esses desafios e propor soluções para melhorar a detecção e classificação de falhas, usando o 3W como estudo de caso.

No próximo capítulo são apresentados os resultados da pesquisa, segregados por classe de falha e tipo de instância, e analisa a melhor configuração para geração de alarmes de falha no sistema CBM, com foco na menor taxa de alarmes perdidos.

Capítulo 4

Resultados e Discussões

Neste capítulo são apresentados os resultados da pesquisa conforme questões apresentadas na Seção 3.2, segregados por classe de falha e tipo de instância, e apresenta a análise da configuração para geração de alarmes de falha no sistema CBM, com foco na taxa de alarmes perdidos.

No MLFLOW [49], são registrados os parâmetros, medidas, artefatos (*artifacts*)[1] e versões de código associados a cada experimento realizado com o MAIS [41] e apresentados neste capítulo, exceto os apresentados na Seção 4.6, facilitando a análise comparativa de diferentes abordagens.

Na Seção 4.1, Experimento 1, são apresentados e discutidos resultados sobre separação das amostras normais que antecedem as amostras de falha. Na Seção 4.2, Experimento 2, são tratados os resultados sobre a exclusão das amostras de falha em regime estacionário. Na Seção 4.3, Experimento 3, são contemplados os resultados do classificador, utilizando a medida F1-ponderada como alvo na busca de hiperparâmetros. Na Seção 4.4, Experimento 4, são expostos os resultados de experimentos que contemplam a influência de eventos simulados na detecção. Na sequência, na Seção 4.5, Experimento 5, são apresentados os resultados sobre a quantidade mínima de experiências necessárias para treinar o classificador, tendo em vista a implementação prática de um sistema CBM. Finalmente, na Seção 4.6, Experimento 6, são analisados o resultado da análise de possível agrupamento dos dados da Classe 1 (Aumento Abrupto de BSW) e suas prováveis causas, e na Seção 4.7 são apresentadas as conclusões do capítulo.

4.1 Experimento 1

O conjunto de dados 3W possui instâncias que contêm exclusivamente amostras normais, bem como instâncias representativas de anomalias, onde amostras normais

[1]Nos artefatos estão armazenados os gráficos de matriz de confusão e os arquivos de dados treinados, com objetos serializados (com a extensão .pkl), empregados na geração de alarmes.

antecedem o período de falha transitória. No Experimento de Referência (Seção 3.1) as amostras normais antecedem o período de falha transitória e foram mantidas nas classes originais de falha.

Neste experimento foram usadas duas classes para representar as amostras sem anomalia, sendo a Classe 0, para observações de eventos normais e a Classe 9 para os inícios normais das falhas.

Foram realizados dois experimentos: considerando balanceamento de amostras normais e não o considerando. O experimento sem balanceamento contempla todas as amostras normais da Classe 0.

Os pontos selecionados no espaço de hiperparâmetros explorado durante o treinamento de cada classificador avaliado estão apresentados na Tabela 4.1:

Tabela 4.1: Melhor conjunto de parâmetros encontrado nos treinamentos.

Parâmetro	Exper. de Referência	Experimento 1	
		com balanceamento	sem balanceamento
Window_size	1024	1024	1024
Normal_balance (b)	2	5	sem balanceamento
N_components	0,999	0,998	0,995
Subsample	0.1	0,15	0.55
Feature_fraction	0,95	0,90	0,25
Num_leaves	58	10	49
Lambda_l1	$1{,}854.10^{-05}$	$2{,}855.10^{-05}$	0,264
Lambda_l2	1,313	0,368	7,553

Observa-se menor utilização de componentes do PCA (parâmetro *N_components*) e maior redução de dimensionalidade no EFB [42](parâmetro *feature_fraction*).

Foram obtidos resultados conforme apontado na Tabela 4.2 e na Figura 4.1.

Tabela 4.2: Resultados do classificador, do Experimento de Referência e do Experimento 1.

		Validação		Teste	
Experimento	Tempo(h)	BalAcc	macro-F1	BalAcc	macro-F1
Referência	17,5	$0{,}955 \pm 0{,}020$	$0{,}958 \pm 0{,}015$	0,976	0,975
1 com balanc.	17,1	$0{,}954 \pm 0{,}022$	$0{,}942 \pm 0{,}028$	0,974	0,969
1 sem balanc.	22,6	$0{,}956 \pm 0{,}018$	$0{,}955 \pm 0{,}018$	0,975	0,975

Observa-se na matriz de confusão do Experimento 1 sem balanceamento das amostras normais (b), na Figura 4.1, uma redução na taxa de alarmes perdidos na Classe 5 de 0,003, e na taxa de falsos alarmes da Classe 0, de 0,003. No entanto, há uma elevação na taxa de alarmes perdidos da Classe 4, de 0,017 para 0,029, e que as amostras normais da Classe 9 apresentam uma taxa de falso alarme sobre a Classe 5 de 0,023. O erro de detecção na Classe 9 evidencia uma possível necessidade de ajuste na anotação na transição entre as amostras normais e as amostras em regime transiente de falha da Classe 5 (Perda rápida de produtividade).

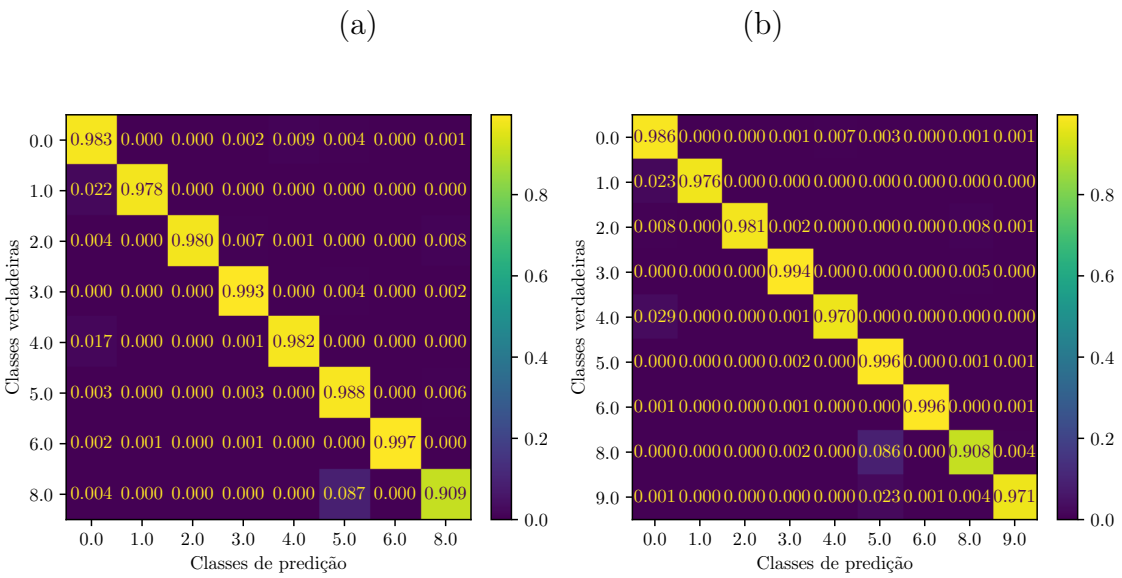

Figura 4.1: Matrizes de confusão de Referência (a) e Experimento 1 sem balanceamento das amostras normais (b)

Na Figura 4.2 (b) é apresentada a matriz de confusão com as amostras normais agrupadas (das Classes 0 e 9) e amostras das falhas (das Classes 1, 2, 3, 4, 5, 6 e 8) agrupadas do Experimento 1 sem balanceamento das amostras normais. Ao comparar com a matriz de referência (a), observa-se um aumento na taxa de alarmes perdidos de 0,007 para 0,011 e na taxa de falsos alarmes de 0,017 para 0,018.

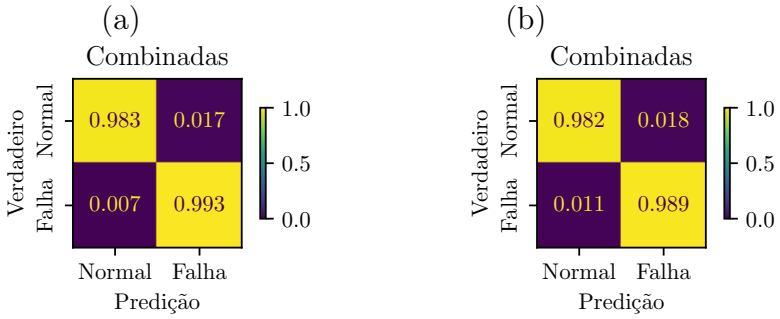

Figura 4.2: Matrizes de confusão das anomalias agrupadas do Experimento de Referência (a) e do Experimento 1, sem balanceamento das amostras normais (b).

A seguir são exibidos gráficos de inferência ou teste em produção sobre instância real da Classe 8, onde pode-se comparar os alarmes gerados. A inferência é realizada nos dois classificadores sem alteração da classe das amostras normais que antecedem as amostras de falha. Na Figura 4.3[2] são exibidos os gráficos do evento WELL-00004_20171031181509.csv, da Classe 8. Observa-se no resultado do Experimento 1, sem balanceamento das amostras normais (b), que há uma redução na margem de predição, sinalizada em vermelho.

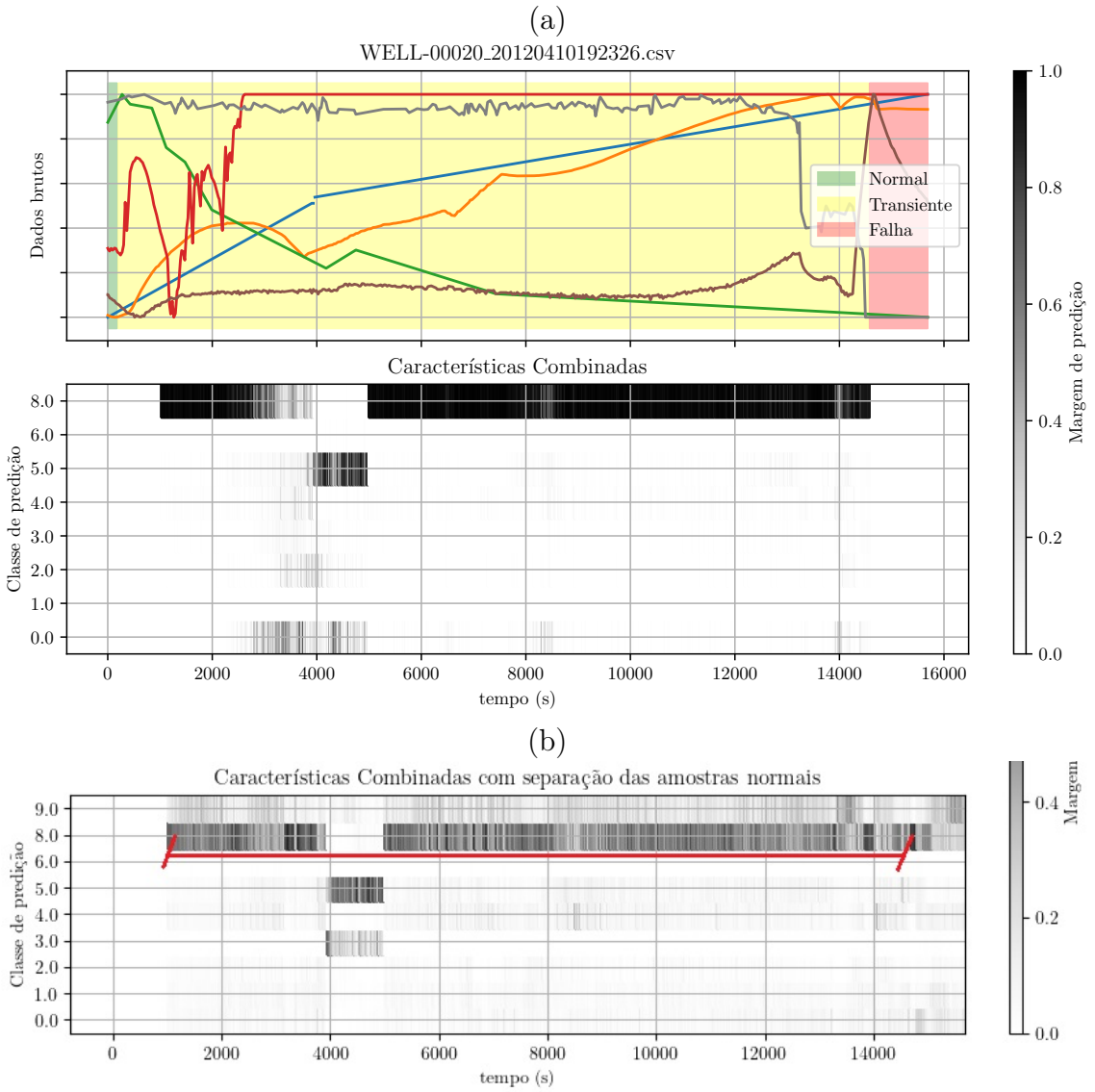

Figura 4.3: Gráficos de alarme de evento da Classe 8, onde pode-se comparar os alarmes gerados, empregando os classificadores treinados, de referência e e do Experimento 1, sem balanceamento das amostras normais (b).

[2]Evolução temporal dos sinais de sensor, o correspondente estado de operação do poço monitorado, e o resultado dos modelos treinados com os classificadores (*unpickling* dos dados serializados) [49][53]. No gráfico que ilustra a evolução temporal dos sinais de sensor, o fundo verde indica operação regular, o fundo amarelo indica estado transitório de falha e o vermelho indica estado permanente de falha. A margem de predição sinaliza a detecção de classe de determinada janela.

Análise dos Resultados. A separação das amostras normais em duas classes proporcionou uma visão mais clara sobre o comportamento dos dados. A acurácia balanceada (BalAcc) e a pontuação macro-F1 mantiveram-se elevadas tanto na validação quanto no teste, indicando que o classificador consegue lidar bem com a nova categorização das amostras normais.

No entanto, separar as amostras normais em outra classe não apresenta vantagem. Na Figura 4.1, observa-se que as amostras normais da Classe 9 têm uma acurácia de classificação menor, refletindo a complexidade em distinguir inícios de falha dos eventos normais, principalmente na Classe 5.

Apesar das instâncias da Classe 4 (Instabilidade de Fluxo) não conterem amostras normais, a taxa de alarmes perdidos aumentou com a introdução da Classe 9. Isso evidencia a dificuldade dos dois classificadores (do Experimento de Referência e do Experimento 1) separarem as amostras da Classe 0 das amostras da Classe 4.

A separação das amostras normais permitiu uma melhor identificação das fases transitórias, importante para a detecção precoce de falhas. A maior sensibilidade do modelo para identificar as transições entre os períodos normal e transitório de falha pode resultar em uma redução na detecção de falsos negativos, ou a taxa de alarmes perdidos, melhorando a capacidade preditiva do sistema. Este tema pode ser estudado em futuros trabalhos.

Na Seção 4.2 é analisada a separação das amostras de falha em estado permanente, trazendo uma outra visão sobre a fase transitória.

4.2 Experimento 2

Nesta seção é analisado como o classificador se comporta quando as amostras coletadas no regime permanente de falha são excluídas do conjunto de dados. O classificador foi configurado para operar sem as amostras no regime permanente de falha, ajustando seus parâmetros para maximizar a BalAcc na detecção precoce de falhas.

Os pontos selecionados no espaço de hiperparâmetros explorado durante o treinamento de cada classificador avaliado estão apresentados na Tabela 4.3.

Tabela 4.3: Melhor conjunto de parâmetros encontrado nos treinamentos.

Parâmetro	Exper. de Referência	Experimento 2
Window_size	1024	1024
Normal_balance (b)	2	9
N_components	0,999	0,980
Subsample	0.1	0.3
Feature_fraction	0,95	0,2
Num_leaves	58	94
Lambda_l1	1,854. 10^{-05}	0,015
Lambda_l2	1,313	2,403

Observa-se uma maior utilização das amostras normais da Classe 0, menor utilização de componentes do PCA (parâmetro *N_components*) e maior redução de dimensionalidade no EFB (parâmetro *feature_fraction*).

Foram obtidos resultados conforme a Tabela 4.4.

Tabela 4.4: Resultados dos classificadores, Experimento de Referência e Experimento 2.

		Validação		Teste	
Experimento	Tempo(h)	BalAcc	macro-F1	BalAcc	macro-F1
Referência	17,5	0,955±0,020	0,958±0,015	0,976	0,975
2	11,7	0,970±0,019	0,975±0,014	0,964	0,972

Nota-se que apesar de validação alcançar melhores resultados do que no Experimento de Referência, no teste a acurácia global (BalAcc) sofreu uma queda, o que sugere um maior erro de generalização na validação, em relação ao Experimento de Referência.

Na Figura 4.4 são apresentadas as matrizes do Experimento de Referência e do Experimento 2. No Experimento 2, os dados das Classes 3 e 4 não foram considerados, devido à ausência de amostras normais e em regime transiente, e nota-se a redução na taxa de falsos alarmes, porém uma queda na acurácia das Classes 2 e 8, acompanhada de aumento na taxa de alarmes perdidos nestas classes de falha.

(a) (b)

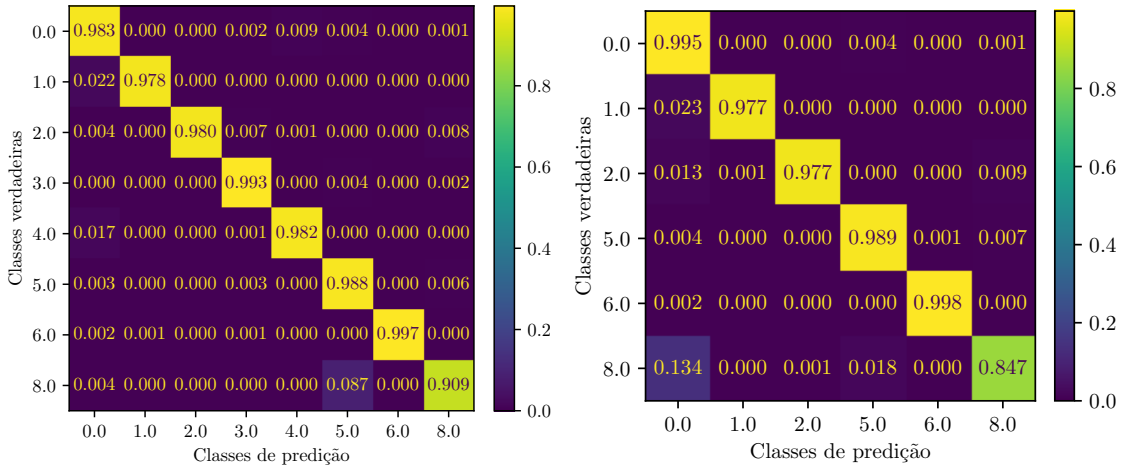

Figura 4.4: Matrizes de confusão dos testes do Experimento de Referência (a) e do experimento 2 (b).

Na Figura 4.5 é ilustrada a matriz de confusão das anomalias agrupadas do Experimento de Referência e a do Experimento 2. Ao comparar com a referência, observa-se o acréscimo na taxa de alarmes perdidos (de 0,007 para 0,035), causado pela classificação das instâncias das Classes 2 e 8, porém a redução na taxa de falsos alarmes (de 0,017 para 0,005), causado pela exclusão das amostras da Classe 4.

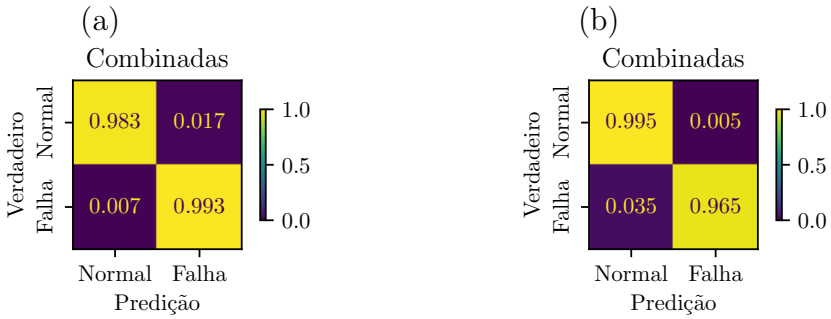

Figura 4.5: matrizes de confusão das anomalias agrupadas do Experimento de Referência (a) e do experimento 2 (b).

A seguir, são exibidos gráficos de inferência (segundo teste) sobre as instâncias reais das Classes 0, 6 e 8, onde pode-se comparar os alarmes gerados, empregando os classificadores treinados, o de referência e o Experimento 2. A inferência dos eventos é realizada nos dois modelos, com a leitura das instâncias de falha contendo as amostras em regime permanente de falha.

A redução na taxa de falsos alarmes pode ser ilustrada analisando a instância WELL-00001 20170421100251.csv da Classe 0, onde observa-se melhor desempenho do Experimento 2, conforme Figura 4.6 [3], e que o classificador detecta que o evento é normal em todas as amostras, enquanto o classificador do Experimento de Referência classifica as amostras nas duas Classes (0 e 4).

Há um indício de semelhança entre as amostras de Classe 0 e 4. Usando o classificador do Experimento de Referência, em testes de inferência, todos os eventos de teste da Classe 0 (129) foram analisados com mais profundidade, sendo gerados os gráficos, e foi detectado que as amostras classificadas na Classe 4 estão concentradas em somente 5 eventos (3,87%), incluindo a instância WELL-00001_-20170421100251.csv. O mesmo procedimento foi empregado com todos os eventos de teste da Classe 4 (84), e foi detectado que as amostras classificadas na Classe 0 estão concentradas em somente 10 eventos (11,90%). Foi observado que os demais eventos das Classes 0 e 4 apresentam classificação correta, gerando alarmes, sem apresentar outros comportamentos relevantes.

[3]Evolução temporal dos sinais de sensor, o correspondente estado de operação do poço monitorado, e o resultado dos modelos treinados com os classificadores (*unpickling dos dados serializados*) [49][53]. No gráfico que ilustra o estado de operação do poço monitorado, o fundo verde indica operação regular.

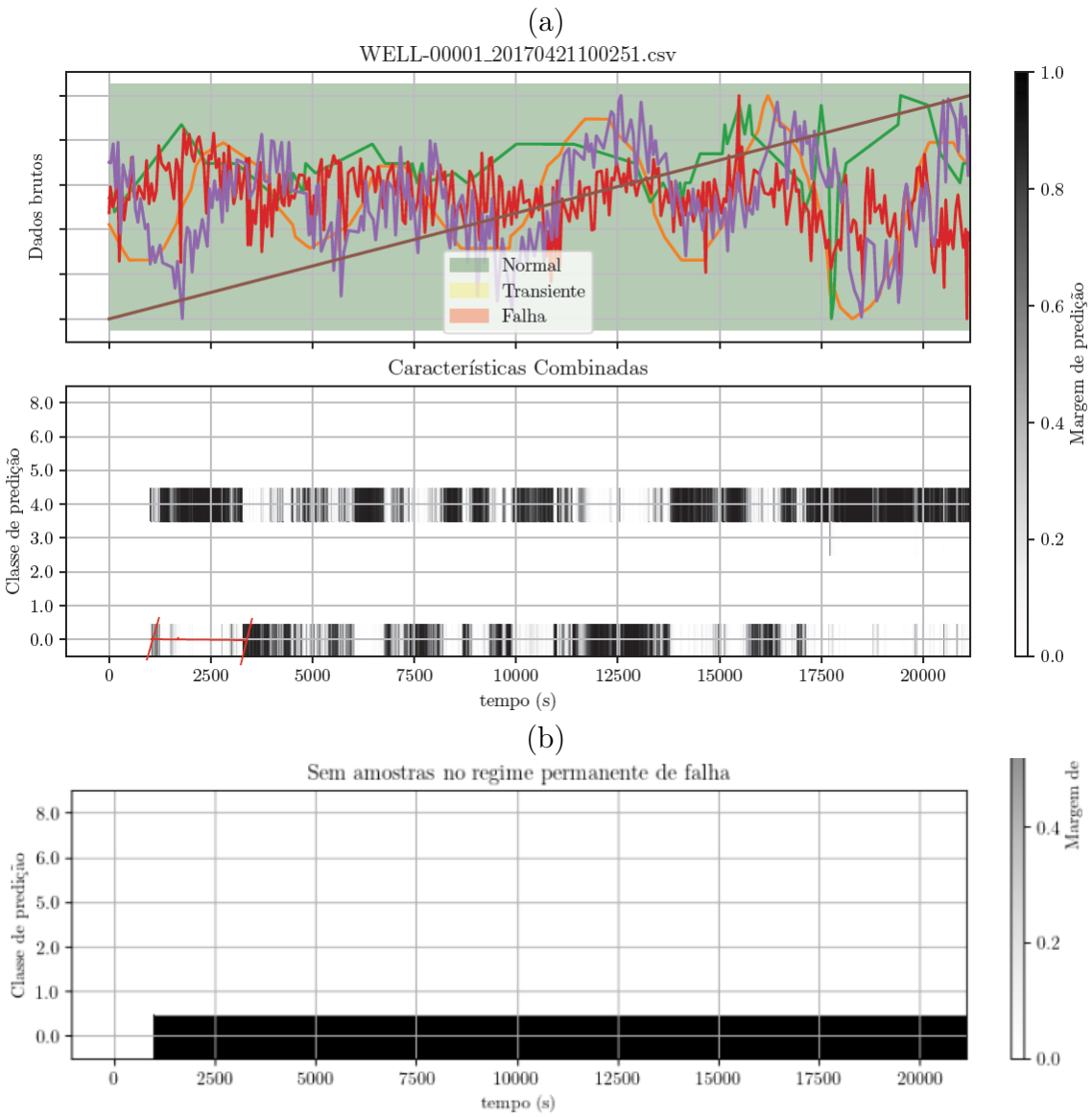

Figura 4.6: Gráficos de alarme de evento da Classe 0, do modelo treinado no Experimento de Referência (a) e do Experimento 2 (b).

Na detecção do evento da Classe 8, WELL-00020 20120410192326.csv, a ausência de amostras em regime permanente de falha inviabiliza a identificação da classe no modelo treinado com o classificador do Experimento 2, conforme ilustrado no gráfico (b) da Figura 4.7 [2] e sinalizado em vermelho.

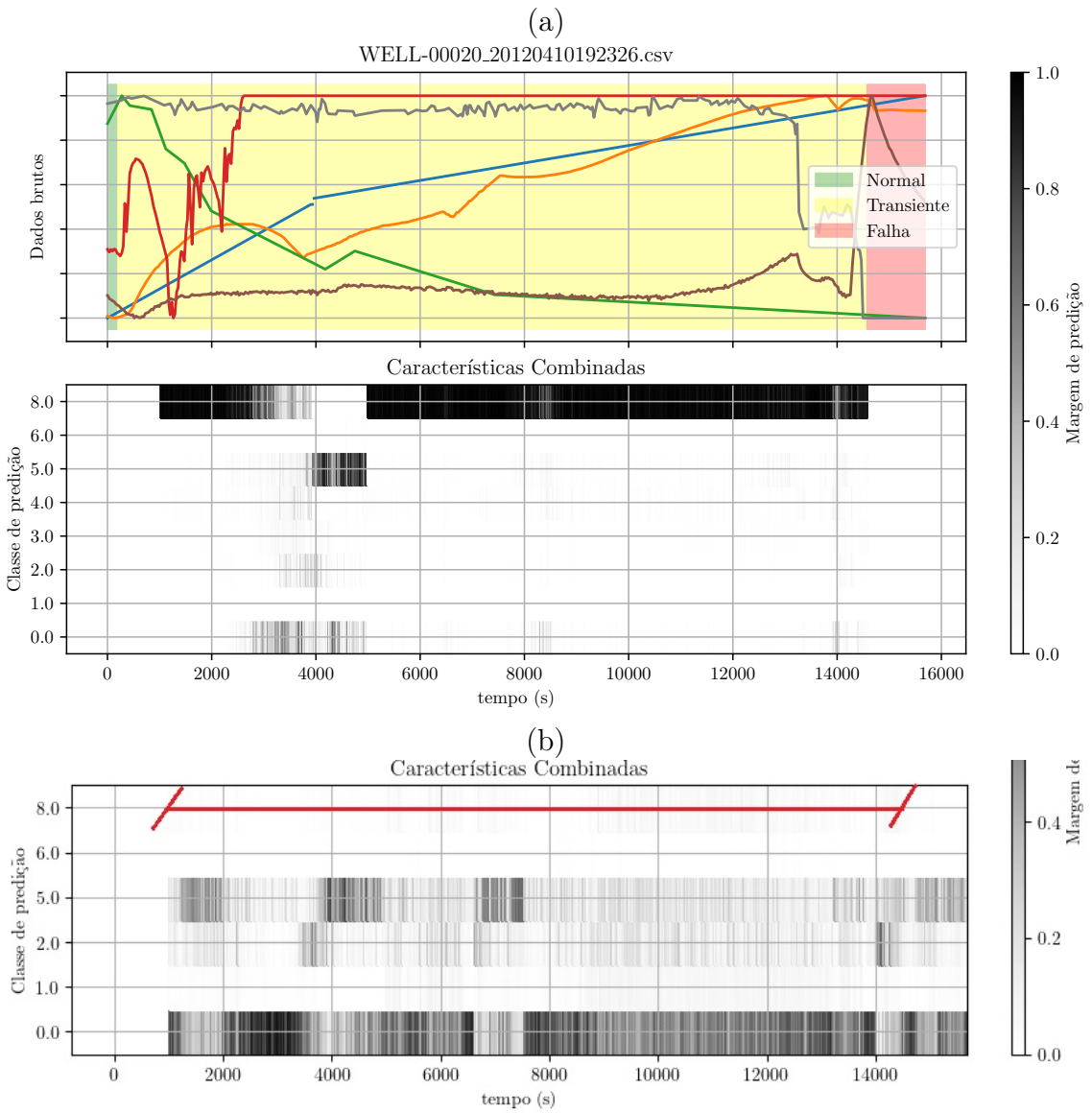

Figura 4.7: Gráficos de alarme de evento da Classe 8, do modelo treinado com o classificador de referência e sem as amostras em regime permanente de falha.

Discussão dos Resultados. A exclusão das amostras em regime permanente de falha desafiou o classificador a focar na detecção precoce de falhas durante o período transiente. A capacidade de detecção precoce não se mostrou aprimorada em relação ao Experimento de Referência, com respeito à detecção de falhas, onde houve aumento da taxa de alarmes perdidos.

A exclusão das Classes 3 e 4 do experimento apresentou melhor resultado de classificação da Classe 0, alcançando redução na taxa de falsos alarmes de 0,017 para 0,005, conforme Figura 4.5 (b). Assim, foi investigada a similaridade em algumas instâncias das Classes 0 e 4, que são responsáveis por redução na acurácia da classificação.

Devido a restrições de arquitetura do sistema MAIS [41], o classificador não foi

experimentado considerando as amostras em regime permanente de falha no treinamento e excluindo do teste. Este é um cenário que reflete a prática de monitoramento de sinais na operação de poços, onde a predição da falha com antecedência do evento é o alvo, e pode ser objeto de futura pesquisa.

4.3 Experimento 3

Nesta seção, é estudado como o classificador se comporta quando é escolhida outra medida de avaliação de desempenho, tendo em vista o desbalanceamento das classes.

Para isso, o MAIS foi configurado no processo de busca pelo melhor hiperparâmetro, utilizando a medida F1-ponderada. Em seguida, os resultados obtidos foram comparados com os resultados de referência, que utilizaram a acurácia balanceada (BalAcc).

Nesta abordagem, na Tabela 4.5 são apresentados os pontos selecionados no espaço de hiperparâmetros explorado durante o treinamento de cada classificador avaliado.

Tabela 4.5: Melhor conjunto de parâmetros encontrado no treinamento do Experimento de referência e do Experimento 3.

Parâmetro	Exper. de Refer.	Exper. 3
Window_size	1024	1024
Normal_balance (b)	2	9
N_components	0,999	0,968
Subsample	0.1	0.1
Feature_fraction	0,95	0,25
Num_leaves	58	56
Lambda_l1	1,854. 10^{-5}	5,181. 10^{-4}
Lambda_l2	1,313	0,379

Observa-se no Experimento 3 que há uma maior influência das amostras de instâncias normais (parâmetro b) e uma redução na taxa de subamostragem de características em cada árvore, conforme parâmetro *feature_fraction*.

Na Tabela 4.6 é apresentado um resumo dos resultados das diferentes abordagens, comparando a acurácia balanceada (BalAcc) e a pontuação macro-F1 em cenários de validação e teste.

Tabela 4.6: Resultados dos classificadores, Experimento de Referência e Experimento 3.

		Validação		Teste	
Cenário	Tempo(h)	BalAcc	macro-F1	BalAcc	macro-F1
Experim. de Refer.	17,5	0,955±0,020	0,958±0,015	0,976	0,975
Experimento 3	18	0,943±0,032	0,967±0,024	0,980	0,978

Na Figura 4.8 é exibida a matriz de confusão dos experimentos. No Experimento 3 nota-se uma melhoria na acurácia das Classes 0, 1, 5 e 8, entretanto uma queda na acurácia das Classes 2, 3, e 4, contribuindo para um aumento na taxa de alarmes perdidos em relação ao Experimento de Referência.

(a) (b)

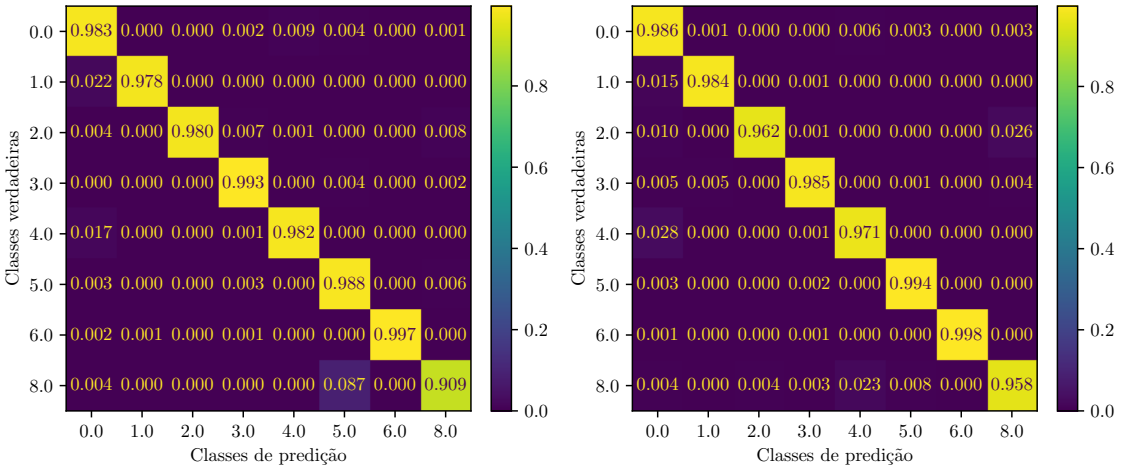

Figura 4.8: Matrizes de confusão do teste do Experimento de Referência (a) e Experimento 3 (b).

Na Figura 4.9 é ilustrada a matriz de confusão das anomalias agrupadas do Experimento de Referência e do Experimento 3. Ao comparar com a referência, observa-se que há um acréscimo na taxa de alarmes perdidos (de 0,07 para 0,010), causado pela menor acurácia na classificação das Classes 2, 3 e 4, porém uma redução na taxa de falsos alarmes (de 0,017 para 0,014).

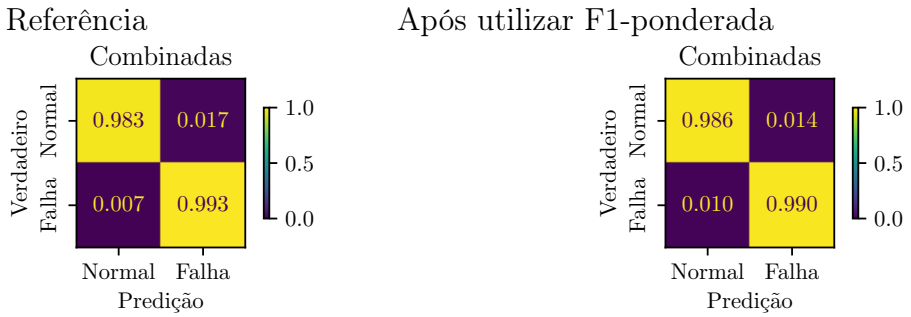

Figura 4.9: Matrizes de confusão das anomalias agrupadas do Experimento de Referência e do experimento após utilizar F1-ponderada como medida de alvo na busca de hiperparâmetros.

Analisando os gráficos de alarmes de falha de alguns eventos, pode-se observar o comportamento distinto do modelo treinado no Experimento 3 em relação ao modelo treinado do Experimento de Referência, conforme classe escolhida.

Selecionando a instância com dados reais da Classe 1, WELL-00006_-20170801063614.csv, conforme visto na Figura 4.10 [2], nota-se que o classificador do Experimento 3 (b) apresenta uma evolução sobre o classificador de referência (a), emitindo alarme com margem de predição (sucessivas janelas com detecção na classe 1) aproximadamente aos 1.600 s, no início da fase transiente de falha, ao tempo em que no de referência a predição com margem acontece aproximadamente aos 3.200 s.

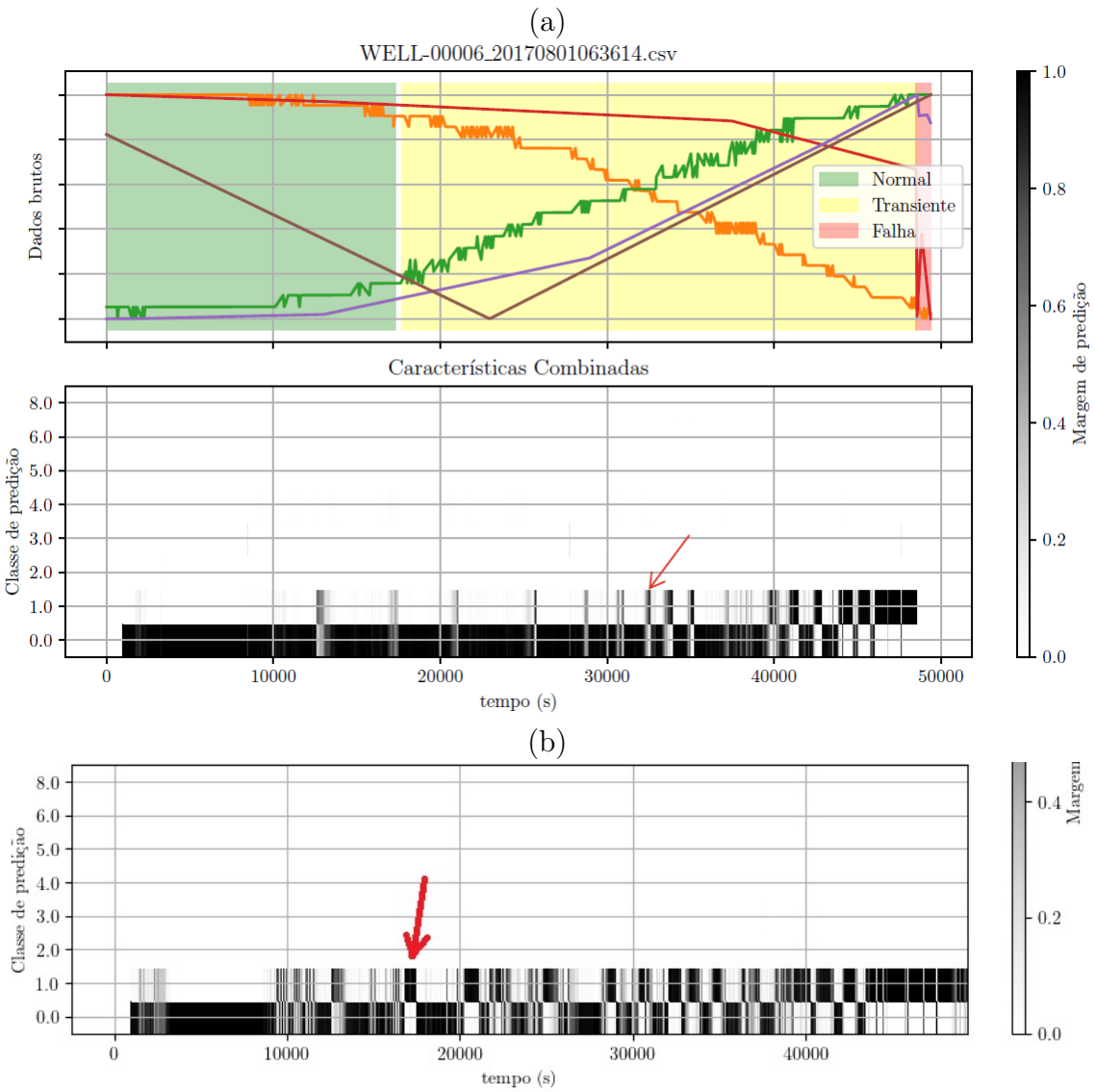

Figura 4.10: Gráficos de alarme de evento da Classe 1, com o classificador de referência (a) e Experimento 3 (b).

Nas inferências com as instâncias reais de falha das demais classes não foi observado melhor resultado do que com o modelo treinado com o classificador de referência.

Análise dos Resultados. Este experimento revela que apesar de as medidas de desempenho com o classificador modificado serem melhores, há um aumento na taxa de alarmes perdidos em relação ao Experimento de Referência de 0,07 para 0,010, causada pela redução na acurácia das Classes 2, 3 e 4.

4.4 Experimento 4

Nesta seção, é estudado qual é a influência das instâncias simuladas no desempenho do classificador.

Foram definidos dois cenários para avaliar a influência das instâncias simuladas:

- Exclusão de instâncias simuladas: O treinamento foi realizado apenas com instâncias de eventos reais;
- Inclusão de instâncias simuladas no treinamento, como no Experimento de Referência.

Os resultados dos dois cenários são apresentados a seguir.

Exclusão de Instâncias Simuladas. Neste cenário, o treinamento foi realizado apenas com eventos reais, excluindo-se os dados das Classes 1, 6 e 8 devido à quantidade de instâncias (não ha quantidade suficiente para treinar o classificador adotado). Na Tabela 4.7, é apresentado o número de instâncias utilizadas nos conjuntos de treinamento e teste.

Tabela 4.7: Número de instâncias reais nos conjuntos de treinamento e teste no cenário de exclusão de instâncias simuladas.

	Treinamento	Teste
0. Normal	468	129
2. Fechamento Espúrio de DHSV	15	7
3. Golfada Severa	22	10
4. Instabilidade de Fluxo	260	84
5. Perda rápida de produtividade	8	4
TOTAL	773	234

Os resultados, conforme ilustrado na Figura 4.11 (b), mostraram uma redução na acurácia das Classes 2, 4 e 5, ao comparar com o modelo de referência (a).

A reduzida quantidade de instâncias nas Classes 2, 3 e 5 influenciou negativamente também na classificação da Classe 4, cujo conjunto não contém instâncias simuladas.

(a) (b)

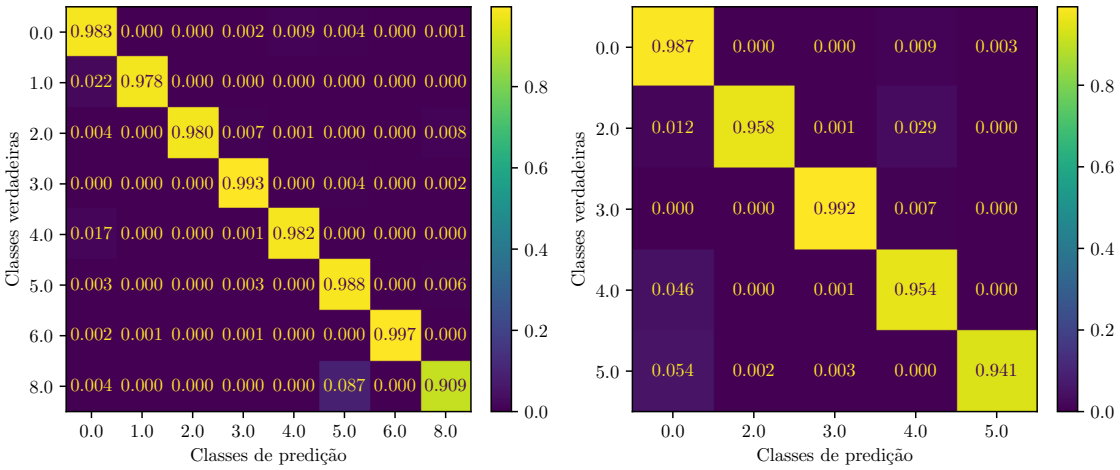

Figura 4.11: Matrizes de confusão do teste do Experimento de Referência (a) e Experimento 4, cenário de exclusão de instâncias simuladas (b).

Na detecção do evento da Classe 2, instância WELL-00003_20170728150240.csv, Figura 4.12 [2], observa-se no gráfico (b) que a ausência de instâncias simuladas no *dataset* resulta em uma detecção com menor acurácia do que no classificador de referência (gráfico a), classificando algumas amostras como Classe 4, conforme sinalizado em vermelho.

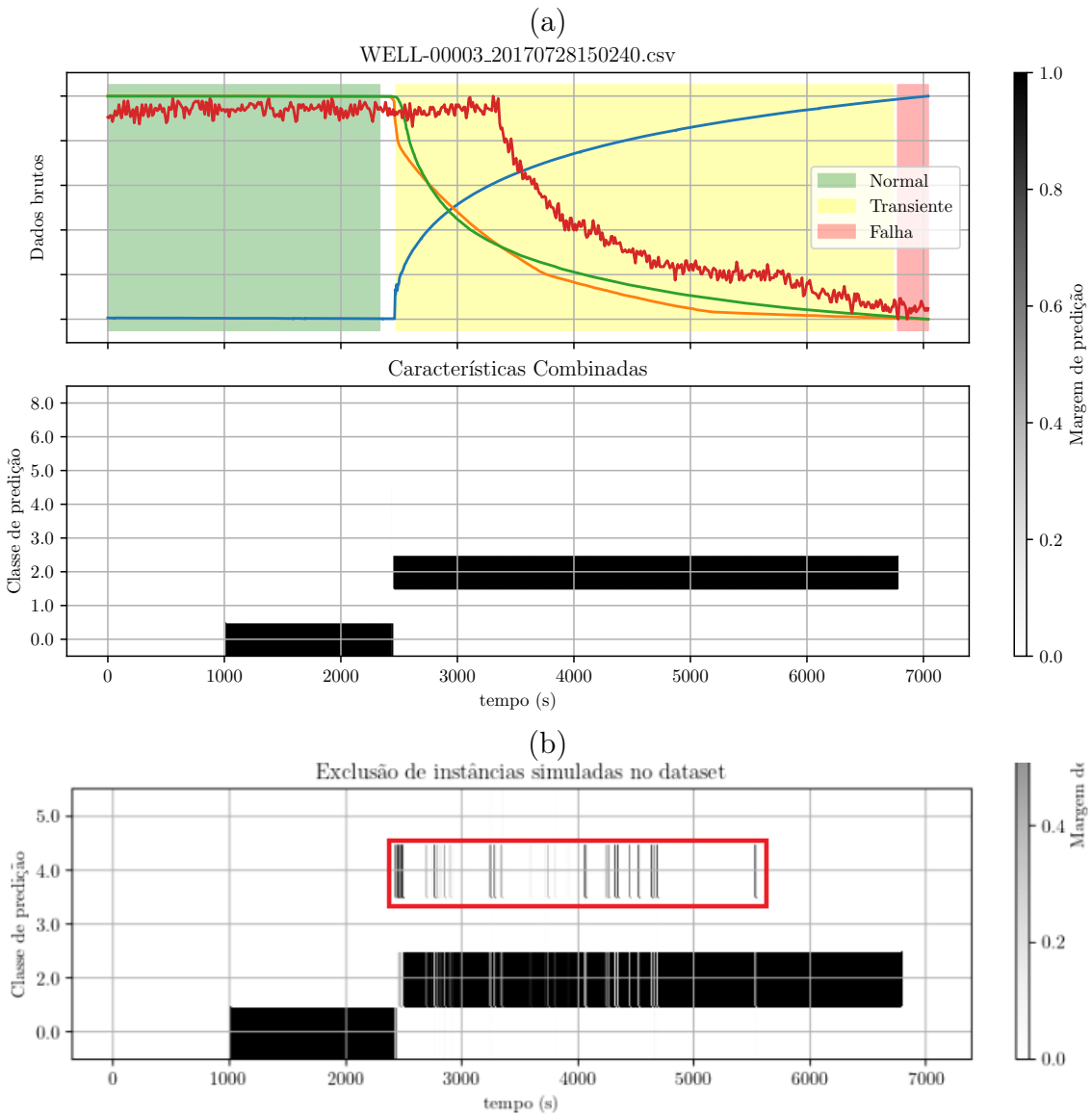

Figura 4.12: Gráficos de alarme de evento da Classe 2, do modelo treinado no Experimento de Referência (a) e no Experimento 4, cenário de exclusão de instâncias simuladas (b).

Na detecção do evento da Classe 4, instância WELL-00003_20170728150240.csv, Figura 4.13 [4], observa-se no gráfico inferior que a ausência de instâncias simuladas no *dataset*, resulta em uma detecção com menor acurácia, classificando algumas amostras como Classe 0, conforme sinalizado em vermelho.

[4]Evolução temporal dos sinais de sensor, o estado de operação do poço monitorado, e o resultado dos modelos treinados com os classificadores (*unpickling* dos dados serializados) [49][53]. O gráfico do topo ilustra a evolução temporal dos sinais de sensor e o correspondente estado de operação do poço monitorado. O fundo vermelho indica estado permanente de falha.

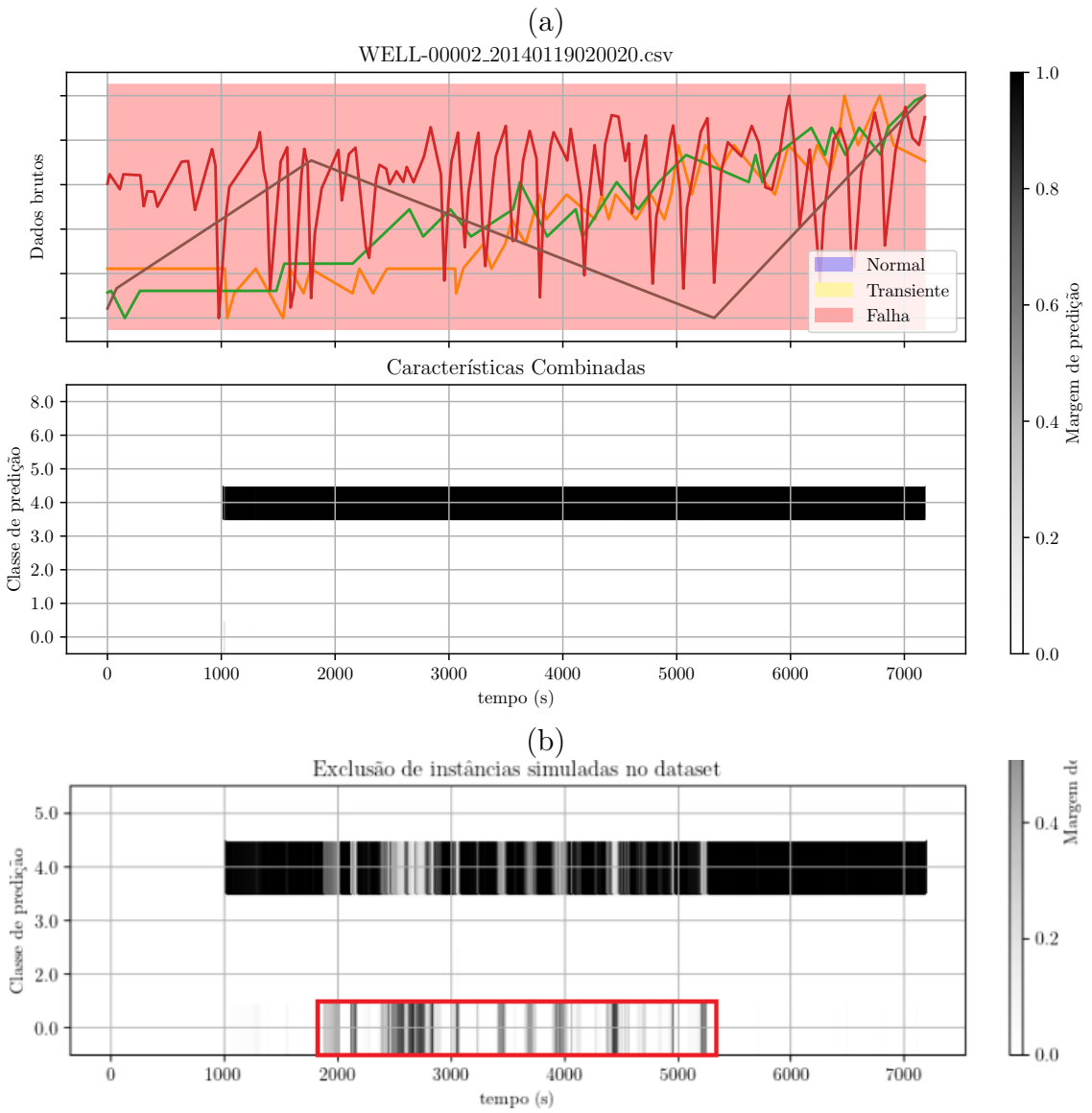

Figura 4.13: Gráfico de alarme de evento da Classe 4, classificador de referência (a) e no Experimento 4, cenário de exclusão de instâncias simuladas (b).

Inclusão de Instâncias Simuladas. A classificação de eventos reais foi avaliada utilizando o modelo treinado no Experimento de Referência.

Na Tabela 4.8, o número de instâncias empregado nos conjuntos de treinamento e teste é apresentado.

Tabela 4.8: Número de instâncias nos conjuntos de treinamento e teste, cenário de Inclusão de Instâncias Simuladas.

	Treinamento		Teste
	Reais	Simuladas	Reais
0. Normal	468	0	129
1. Aumento abrupto de BSW	3	78	2
2. Fechamento Espúrio de DHSV	15	11	7
3. Golfada Severa	22	55	10
4. Instabilidade de Fluxo	260	0	84
5. Perda rápida de produtividade	8	340	4
6. Restrição Rápida no CKP	4	170	2
8. Hidrato na Linha de Produção	0	56	3
TOTAL	780	710	241

Na matriz de confusão resultante do Experimento 4, cenário de Inclusão de Instâncias Simuladas, ilustrada na Figura 4.14 (b), nota-se que a menor população de eventos reais nas Classes 1, 6 e 8, quando comparadas às demais, resultou em uma queda na acurácia de classificação dessas anomalias, respectivamente de 0,978 para 0,117, de 0,997 para 0,778 e de 0,909 para 0,449.

(a) (b)

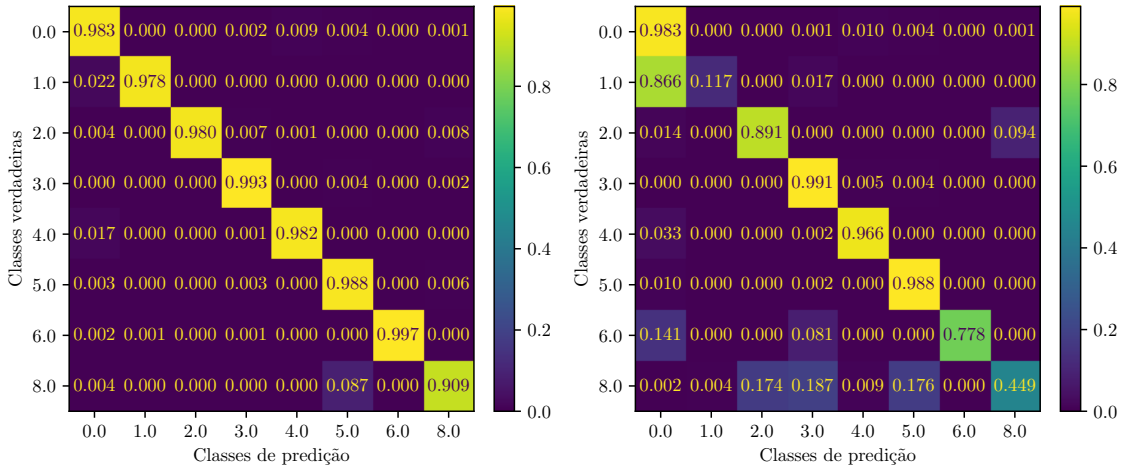

Figura 4.14: Matrizes de confusão do teste do Experimento de Referência (a) e do Experimento 4, cenário de Inclusão de Instâncias Simuladas (b).

Na Tabela 4.9 são apresentados os resultados dos Experimentos, comparando a acurácia balanceada (BalAcc) e a pontuação macro-F1 para cada cenário.

Tabela 4.9: Comparação de Desempenho: Instâncias Reais vs. Simuladas.

	Validação		Teste	
Cenário	BalAcc	macro-F1	BalAcc	macro-F1
Exper. 4, Exclusão de Simuladas	0,828±0,100	0,825±0,073	0,966	0,968
Exper. 4, Inclusão de Simuladas, teste com reais	0,955±0,020	0,958±0,015	0,770	0,770
Reais+Simuladas (Experimento de Referência)	0,955±0,020	0,958±0,015	0,976	0,975

Conforme exibido na Figura 4.15 [2], analisando a inferência do evento da Classe 6, WELL-00004_20171031181509.csv, no Experimento 4, cenário de Inclusão de Instâncias Simuladas, pode-se observar que há geração de alarme aproximadamente aos 1.000 s, porém com alguma perda de alarme em algumas amostras e outras com geração de alarmes na Classe 3 .

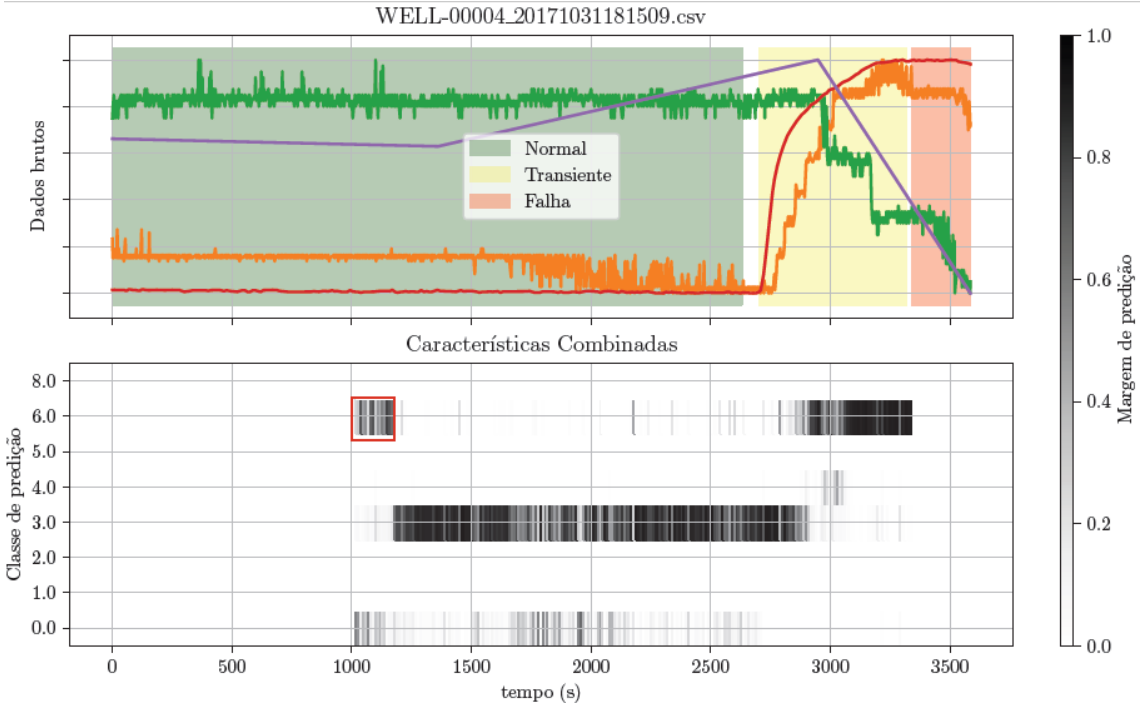

Figura 4.15: Gráfico de alarme de evento da Classe 6 no Experimento 4, cenário de Inclusão de Instâncias Simuladas.

Análise dos Resultados. Os resultados indicam que a inclusão de eventos simulados pode melhorar a cobertura geral e fornecer um conjunto de dados mais robusto para treinamento. No entanto, a precisão pode ser comprometida se os dados simulados não forem representativos das condições reais de operação, con-

forme ilustrado no exemplo de inferência no Experimento 4, cenário de Inclusão de Instâncias Simuladas, Figura 4.15, sobre a instância real da Classe 6.

4.5 Experimento 5

Para determinar a quantidade mínima necessária de experiências que viabiliza uma aceitável detecção de eventos indesejáveis, foi adotada a padronização de 100 experiências do MAIS. Este estudo analisou o treinamento do classificador durante a pesquisa de melhor conjunto de hiperparâmetros [43], verificando a quantidade mínima de experimentos que produz resultados próximos da acurácia balanceada (BalAcc) do Experimento de Referência. Em cada cenário de 10 a 60 experiências, foram executados de 3 a 4 treinamentos (busca de hiperparâmetros) e testes, e registrados os resultados dos melhores modelos, obtidos por validação cruzada, nos testes.

Resultados satisfatórios foram obtidos inicialmente no cenário com 20 experiências. Na Tabela 4.10 é apresentada uma comparação dos hiperparâmetros dos experimentos em relação aos resultados do Experimento de Referência. Comparando os parâmetros dos experimentos com 20 experiências com o Experimento de referência, nota-se que o alvo foi encontrado com maior redução de dimensionalidade nas amostras (parâmetro *subsample*), no entanto empregando uma maior taxa de amostragem de características (parâmetro feature_fraction) e maior número de componentes do PCA (parâmetro n_components).

Tabela 4.10: Melhor conjunto de parâmetros encontrado no Experimento de Referência e no Experimento 5, nos cenários de treinamento com 10, 20, 30, 40, 50 e 60 experiências.

Parâmetro	Refer.	10	20	30	40	50	60
Window_size	1024	1024	1024	1024	1024	1024	1024
Normal_balance (b)	3	2	2	4	2	6	7
N_components	0,971	0,951	0,999	0,986	0,988	0,994	0,994
Subsample	0,35	0,4	0.1	0.5	0,9	0,15	0,55
Feature_fraction	0,7	0,9	0,95	0,55	0,75	0,15	0,35
Num_leaves	92	99	58	95	16	69	16
Lambda_l1	0,042	$1{,}909.10^{-4}$	$1{,}854.10^{-5}$	3,742	0,010	$6{,}768.10^{-4}$	$9{,}413.10^{-4}$
Lambda_l2	0,094	0,374	1,313	3,691	6,989	8,773	0,224

Na Tabela 4.11 são apresentados os melhores resultados do Experimento 5, nos cenários de treinamento com 10, 20, 30, 40, 50 e 60 experiências e do Experimento de Referência.

Tabela 4.11: Melhores resultados dos experimentos, utilizando os conjuntos crescentes de experiências, variando de 10 a 60.

		Validação		Teste	
Experimento	Tempo(h)	BalAcc	macro-F1	BalAcc	macro-F1
10 experiências	1,6	0,931±0,024	0,928±0,016	0,969	0,966
20 experiências	3,7	0,946±0,021	0,950±0,019	0,977	0,977
30 experiências	5,1	0,948±0,015	0,946±0,020	0,976	0,973
40 experiências	8,1	0,952±0,021	0,944±0,018	0,976	0,977
50 experiências	10,1	0,964±0,016	0,968±0,012	0,970	0,976
60 experiências	11,5	0,953±0,021	0,955±0,022	0,974	0,975
100 experiências(Refer.)	17,5	0,955±0,020	0,958±0,015	0,976	0,975

Na Figura 4.16 são apresentadas as matrizes de confusão dos melhores resultados nos cenários dos modelos testados com 20 e 30 experiências, respectivamente.

(a) (b)

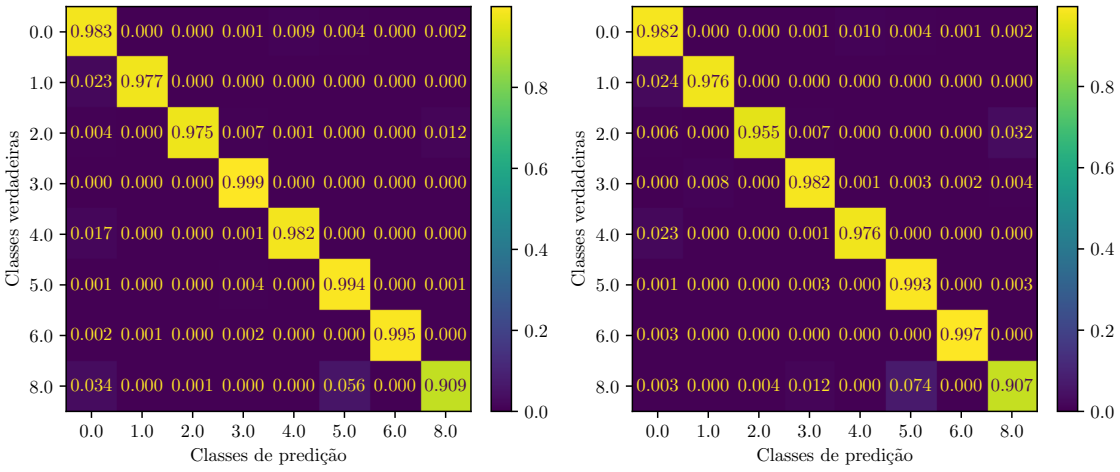

Figura 4.16: Matriz de confusão do teste após treinamento com busca de hiperparâmetros com 20 experiências (a) e 30 experiências (b).

Na Figura 4.17 são apresentadas as matrizes de confusão dos melhores resultados nos cenários dos modelos testados com 40 e 50 experiências, respectivamente.

(a) (b)

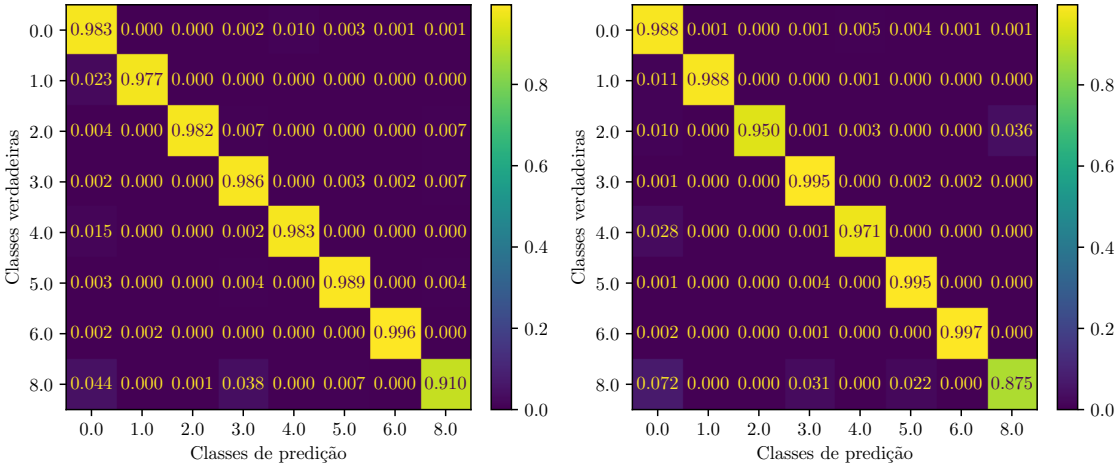

Figura 4.17: Matriz de confusão do teste após treinamento com busca de hiperparâmetros com 40 experiências (a) e 50 experiências (b).

Na Figura 4.18 são apresentadas as matrizes de confusão dos melhores resultados no cenário do modelo testado com 60 experiências (a) e do Experimento de Referência (b).

(a) (b)

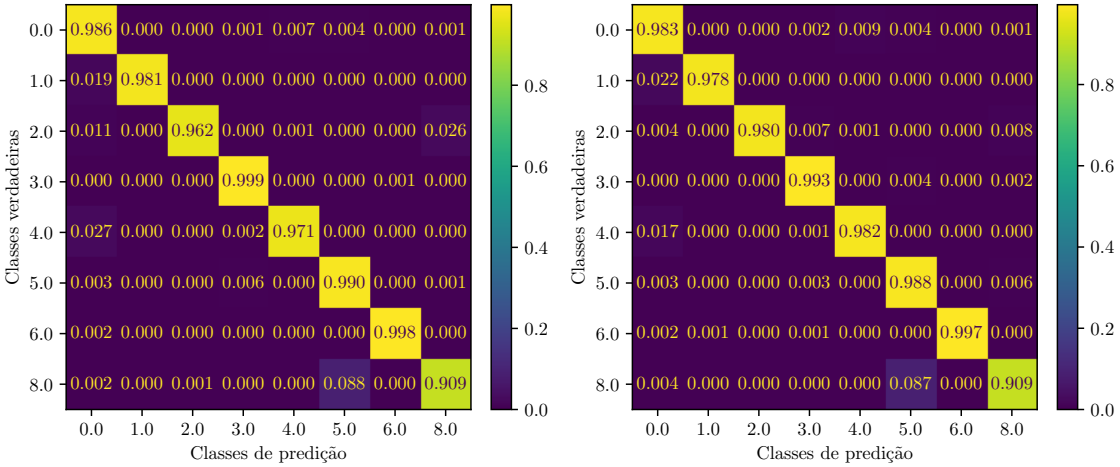

Figura 4.18: Matriz de confusão do teste após treinamento com busca de hiperparâmetros com 60 experiências (a) e 100 experiências (b).

Na Figura 4.19 [5] é ilustrada a evolução da acurácia balanceada em função do número de experiências.

Figura 4.19: Gráfico de evolução da acurácia balanceada por experimento.

Na Figura 4.20 [5] é mostrada a evolução do macro-F1 em função do número de experiências.

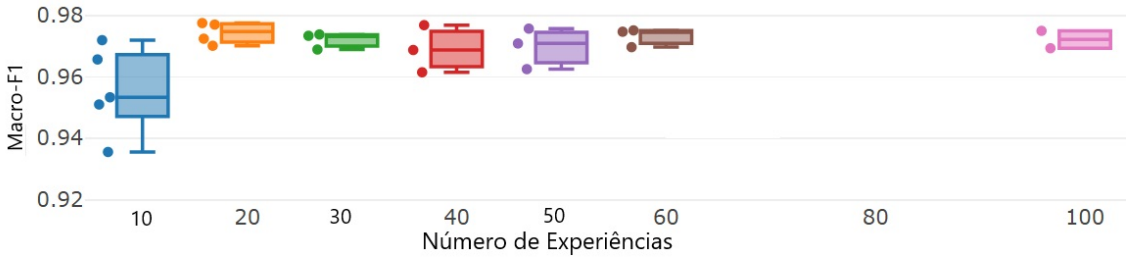

Figura 4.20: Gráfico de evolução do por experimento.

Análise dos Resultados. Na Seção 3.1, dentro do processo de comparação de classificadores baseados nas características, foi observado que o classificador que obteve o menor tempo de processamento foi o estatístico sem PCA, quando demandou 8,3 h, alcançando valores de BalAcc e macro-F1 de 0,972 e 0,972 respectivamente.

Neste Experimento, no cenário com experimentos de 20 experiências, observa-se uma redução de tempo quando comparado com o Experimento de Referência de 17,5h para 3,7h, e inferior até ao citado no parágrafo anterior, resultando em valores de BalAcc e macro-F1 superiores ao resultado obtido com o classificador com características estatísticas sem PCA.

Comparando o resultado por evento da matriz de confusão do classificador, do melhor resultado de experimento com 20 experiências (Figura 4.16 (a)) à matriz do classificador de melhor resultado com 100 experiências (Figura 4.18 (b)), pode ser constatada a queda no desempenho de classificação nas classes 1, 2 e 6, sem prejuízo

[5]O eixo horizontal representa o tamanho do experimento (quantidade de experiências por experimento). Os pontos marcam os valores exatos dos resultados encontrados nos experimentos. O retângulo central da haste possui três linhas que estão na horizontal: a linha de baixo, representada pelo contorno externo inferior do retângulo, indica o primeiro quartil; A linha de cima, que é o contorno externo superior do retângulo, indica o terceiro quartil; A linha interna indica a mediana. Gráfico gerado, empregando os dados armazenados no MLFLOW ([49]) e utilizando recursos de comparação da plataforma.

de detecção de falha, conforme exemplos a seguir, porém uma elevação na acurácia das classes 3 (de 0,993 para 0,999) e 5 (de 0,988 para 0,994). Comparando as medianas dos quatro experimentos do cenário com 20 experiências, com a mediana dos dois experimentos do cenário de referência, obtém-se os valores, respectivamente, de 0,974 e 0,972 de BalAcc e de 0,974 e 0,972 de macro-F1

Selecionando para inferência a instância com dados reais da Classe 1, WELL-00006_20170801063614.csv, nota-se que no modelo treinado com o classificador de referência, a predição com margem acontece aproximadamente aos 3.200 s, conforme Figura 4.21 [2], gráfico (a), sinalizado com a seta vermelha. O modelo treinado no Experimento 5, no cenário com 20 , gráfico (b), apresenta uma sutil desvantagem em relação ao modelo treinado com o classificador de referência, emitindo alarme com menor consistência de predição até aproximadamente 4.200 s, porém ainda na fase transiente de falha.

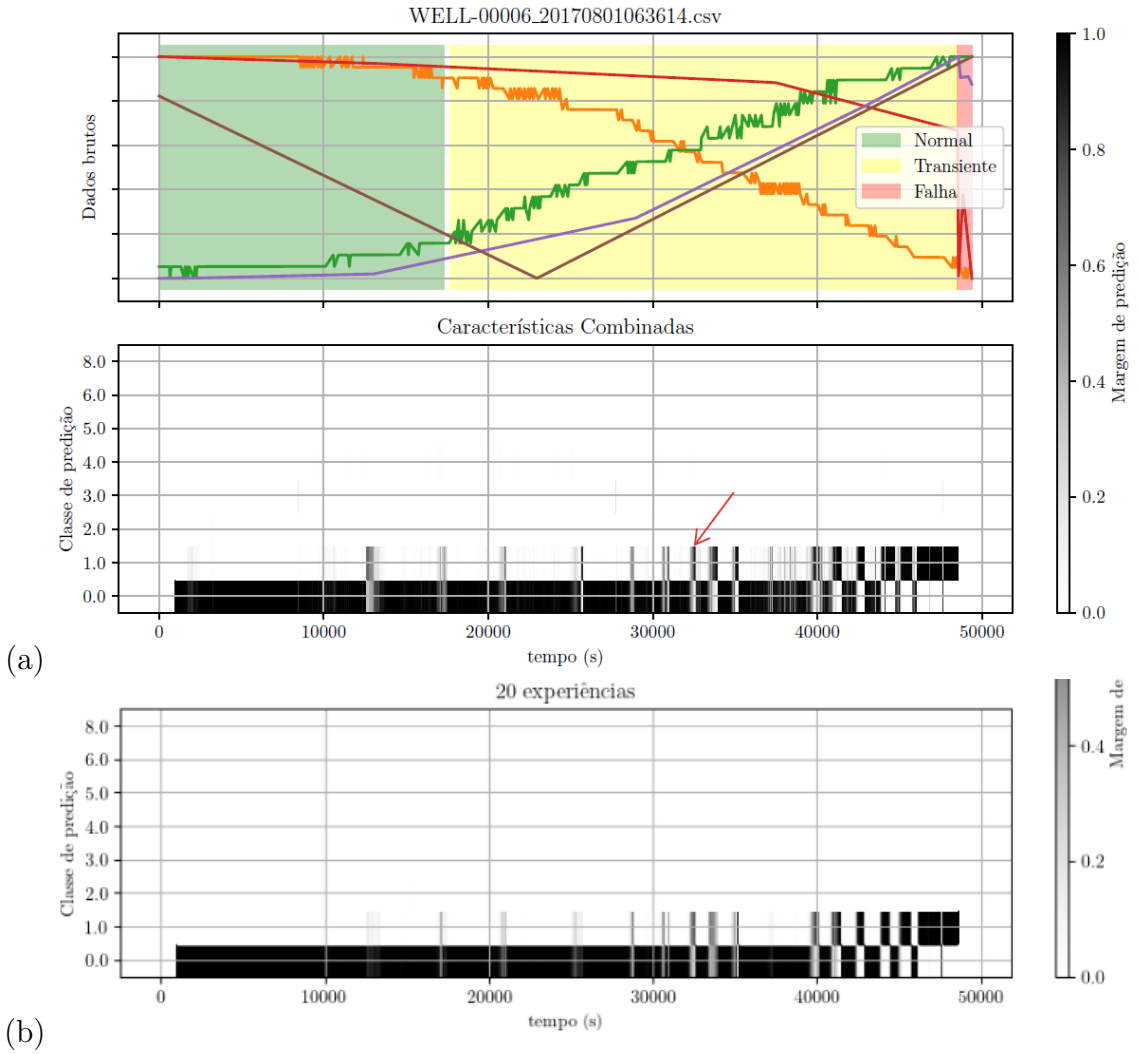

Figura 4.21: Gráficos de alarme de evento da Classe 1, do modelo treinado com o classificador de referência (a) e no Experimento 5, no cenário com 20 experiências (b).

Na detecção do evento da Classe 2, instância WELL-00003_20170728150240.csv, Figura 4.22 [2], observa-se no gráfico (b) que há detecção ainda no início da fase transiente, classificando algumas amostras a partir do centro da fase transiente como classe 4 ou 8, conforme sinalizado em vermelho, mas cumprindo a função de alarme no mesmo instante que o classificador de referência.

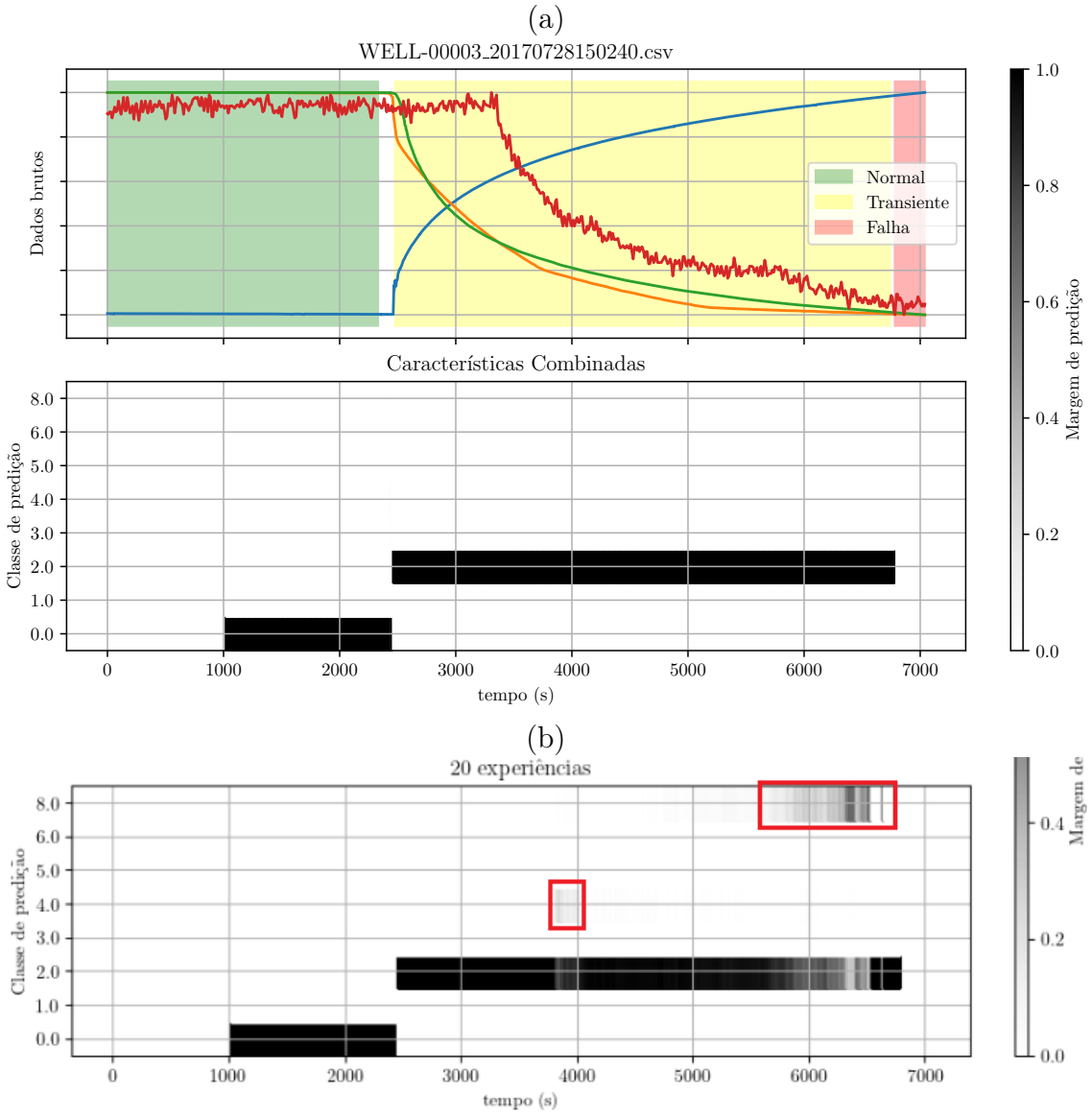

Figura 4.22: Gráficos de alarme de evento da Classe 2, do modelo treinado com o classificador de referência (a) e no Experimento 5, no cenário com 20 experiências (b).

4.6 Experimento 6

Com o objetivo de investigar a existência de agrupamentos indesejados que possam reduzir a eficácia do modelo na identificação e classificação de falhas, e as causas para a formação de mais de um grupo em uma dada classe de falha, foi selecionada a Classe 1 (Aumento Abrupto de BSW) de falha, para uma análise aprofundada de agrupamento. Na Tabela 4.12 é apresentada a quantidade de instâncias por tipo.

Tabela 4.12: Número de instâncias nos conjuntos de treinamento e teste (entre parenteses) da Classe 1.

Tipo de evento	Real	Simulada	Total
1. Aumento Abrupto de BSW	3(2)	78(36)	81(38)

Inicialmente, foi avaliada a eficácia do algoritmo em agrupar instâncias similares e separar as dissimilares, investigando qual a quantidade de *clusters* que proporciona a melhor agrupamento para cada variável.

Na Figura 4.23[6] são exibidos os resultados do coeficiente de silhueta de 81 instâncias de treinamento, para as variáveis P-PDG (Melhor Coeficiente de Silhueta: 0,53, para 3 *clusters*) e P-TPT (Melhor Coeficiente de Silhueta: 0,70, para 2 *clusters*), conforme a quantidade de *clusters* parametrizada.

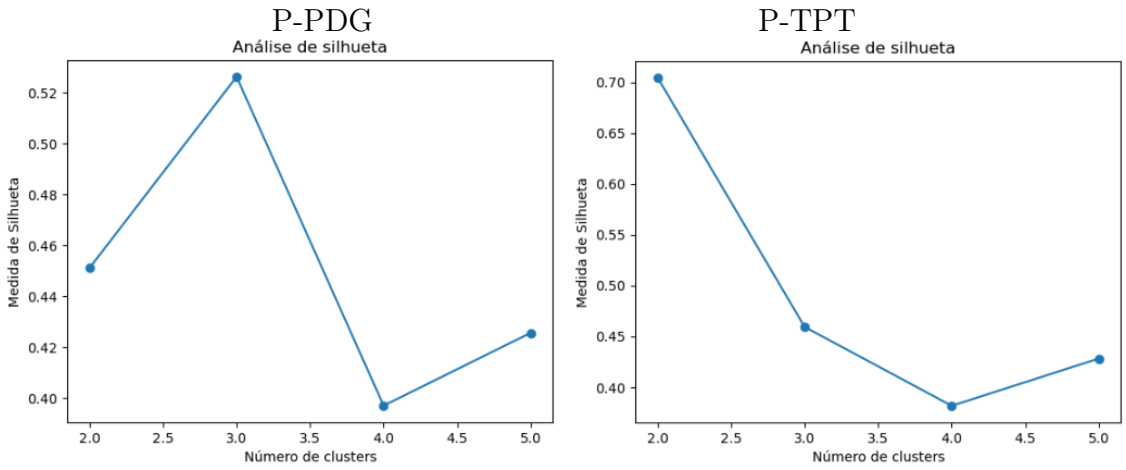

Figura 4.23: Resultados do coeficiente de silhueta para as variáveis P-PDG e P - TPT, conforme a quantidade de *clusters* parametrizada.

Na Figura 4.24[6] são exibidos os resultados do coeficiente de silhueta de 81 instâncias de treinamento, para as variáveis Variável T-TPT (Melhor Coeficiente de Silhueta: 0,90, para 2 *clusters*) e P-MON-CKP (Melhor Coeficiente de Silhueta: 0,69, para 2 *clusters*), conforme a quantidade de *clusters* parametrizada.

[6]O código e respectivos gráficos se encontram no *notebook* publicado no GitHub do autor. [48]

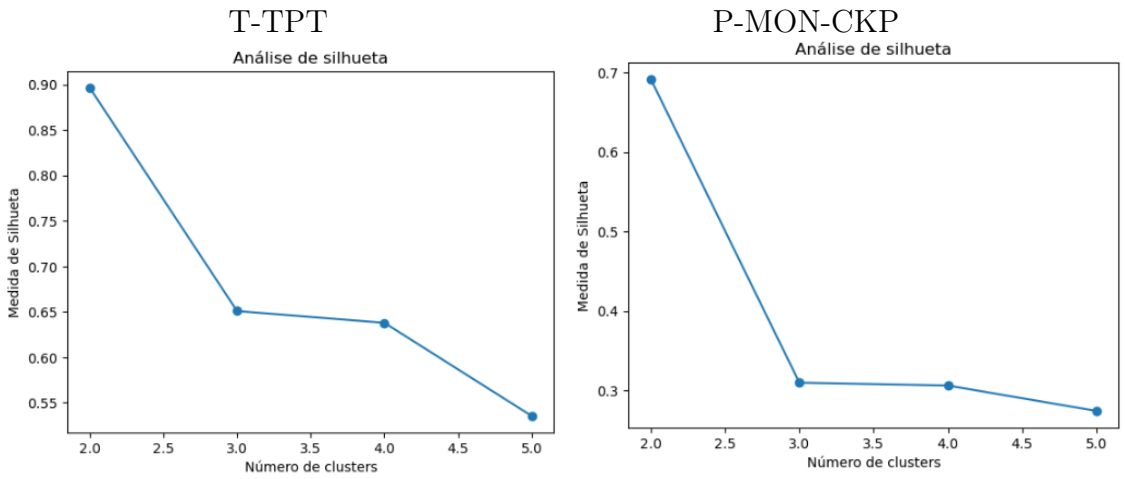

Figura 4.24: Resultados do coeficiente de silhueta para as variáveis T-TPT e P-MON-CKP, conforme a quantidade de *clusters* parametrizada.

Na Figura 4.25[6] são exibidos os resultados do coeficiente de silhueta de 81 instâncias de treinamento, para a Variável T-JUS-CKP (Melhor Coeficiente de Silhueta: 0,84, para 2 *clusters*), conforme a quantidade de *clusters* parametrizada.

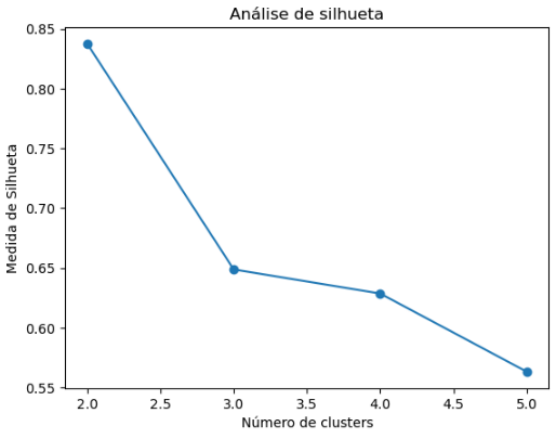

Figura 4.25: Resultados do coeficiente de silhueta para a variável T-JUS-CKP, conforme a quantidade de *clusters* parametrizada.

Nas análises de silhueta na Classe 1, somente na Variável P-PDG foi alcançado melhor resultado para 3 *clusters*. No entanto foi identificado que a Variável T-TPT apresentou o melhor coeficiente de silhueta para 2 *clusters* (0,90), corroborando os resultados encontrados por MACHADO *et al.* [37] em sua análise de agrupamento por similaridade com um grupo menor de instâncias.

Análise de agrupamento. As 81 instâncias de treinamento foram selecionadas para análise de agrupamento, empregando *k-means* [39][40] e usando DTW [38] como medida de similaridade, considerando a Variável T-TPT com 2 *clusters*. Na Figura 4.26[6] são exibidos os resultados dos agrupamentos obtidos.

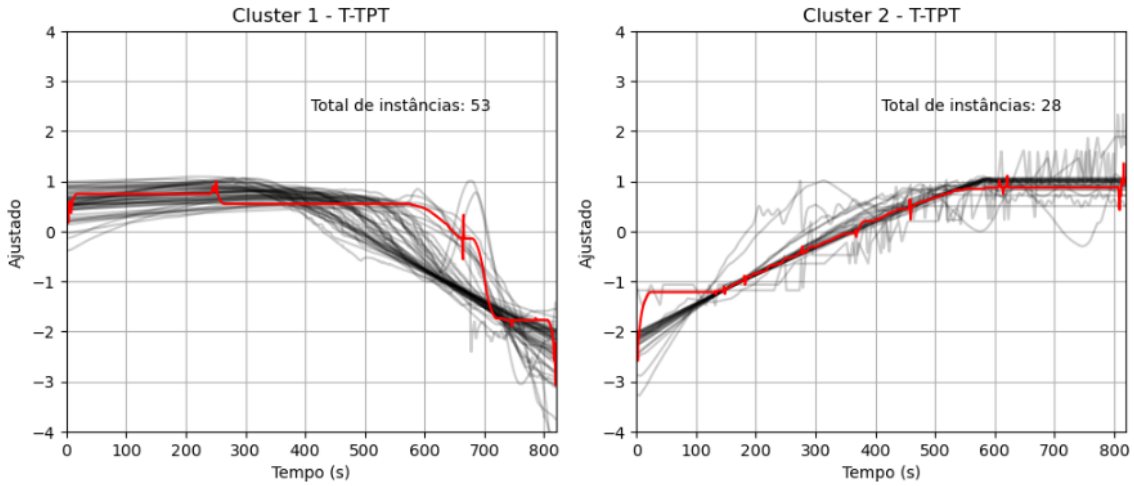

Figura 4.26: Resultado de agrupamento das instâncias da Variável T-TPT na Classe 1.

Pode-se visualmente constatar que nos dois *clusters* há um comportamento bem distinto de temperatura, onde no primeiro há resfriamento e no segundo o aquecimento dos fluidos do poço.

Análise de causas da existência de 2 *clusters* na Classe 1. Comparando primeiramente os dois *clusters* nos dados de treinamento da Classe 1 com o tipo de instância deles (3 reais e 78 simuladas), constata-se numericamente que não há correlação com a forma como os dados foram coletados (real ou simulado). Conforme mostrado na Seção 4.4, no cenário de inclusão de instâncias simuladas no treinamento, a influência das instâncias simuladas sobre a classificação de instâncias reais da Classe 1 é muito pequena, resultando em uma acurácia de 0,117 (Figura 4.14).

Na Seção 2.7 foram abordadas as principais causas de aumento de BSW em poços, citando condições do reservatório, problemas operacionais e erosão. Algumas peculiaridades a respeito da topologia do poço (vertical ou horizontal) e respectivas características de influência a partir de mudanças geotermais, características friccionais dos fluidos ou conificação de água, resultam em aquecimento ou resfriamento da água.

Devido ao sigilo das locações dos poços, não há condições de cruzar o resultado do agrupamento com as características geo-mecânicas dos poços ou os problemas operacionais enfrentados. Contudo, há a suspeita de que o *cluster* com eventos onde há elevação de temperatura está relacionado com poços horizontais, caracterizados por entrada de água aquecida. Esta entrada é resultado de fluxo de água de um aquífero, localizado abaixo da zona de produção (conificações de água). Analisando as curvas de evolução da variável, de três instâncias reais utilizadas no treinamento, conforme Figura 4.27 [2], pode-se constatar este comportamento.

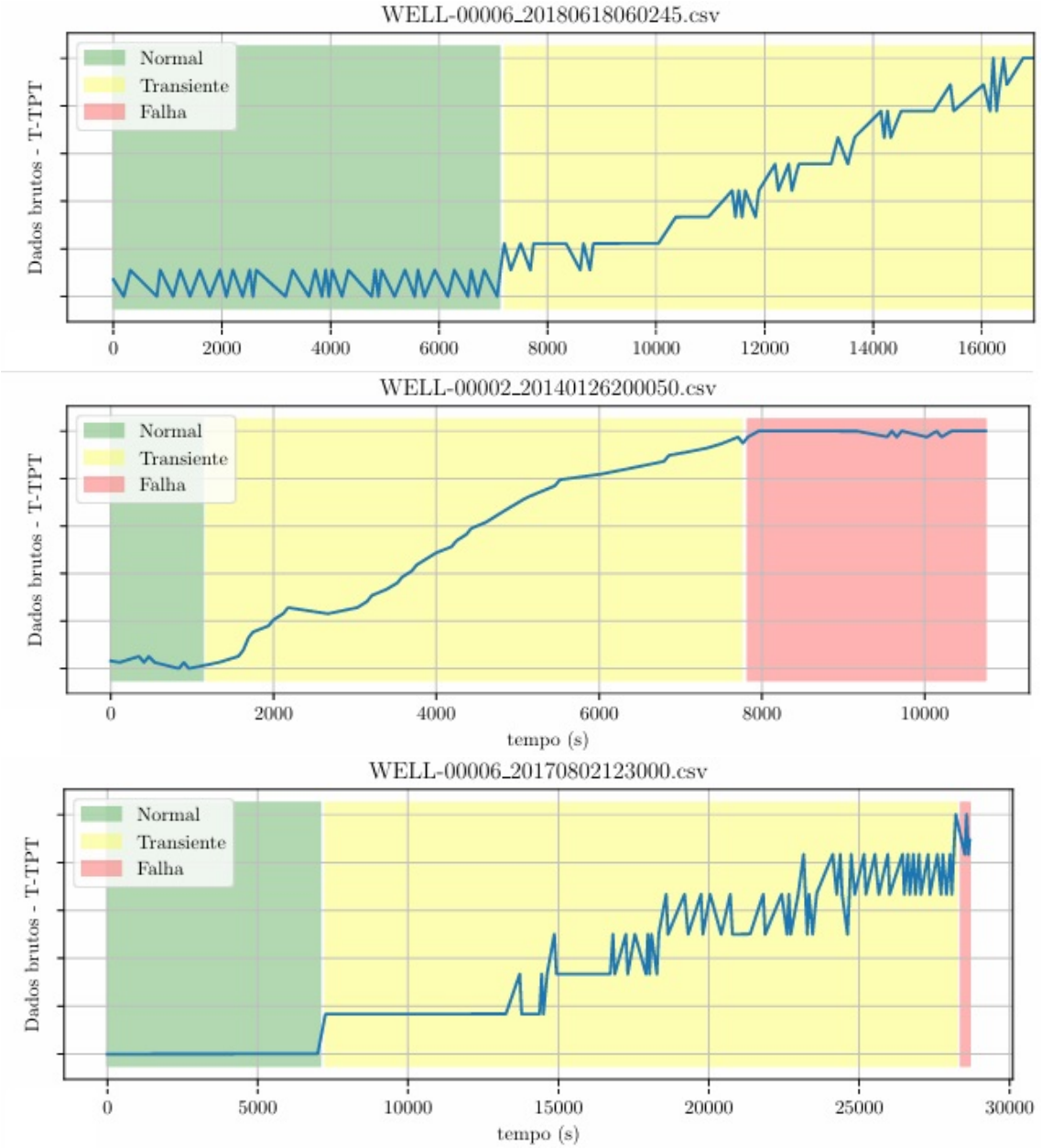

Figura 4.27: Gráficos de evolução da Variável T-TPT das três instâncias reais integrantes da base de dados de treinamento, WELL-00006201806180602 45.csv, WELL-00002201401262000 50.cs e WELL-00006201708021230 00.csv.

Analisando as curvas de evolução da Variável T-TPT, de duas instâncias reais utilizadas no teste, conforme Figura 4.28 [2], pode-se constatar o mesmo comportamento encontrado nas instâncias de treinamento.

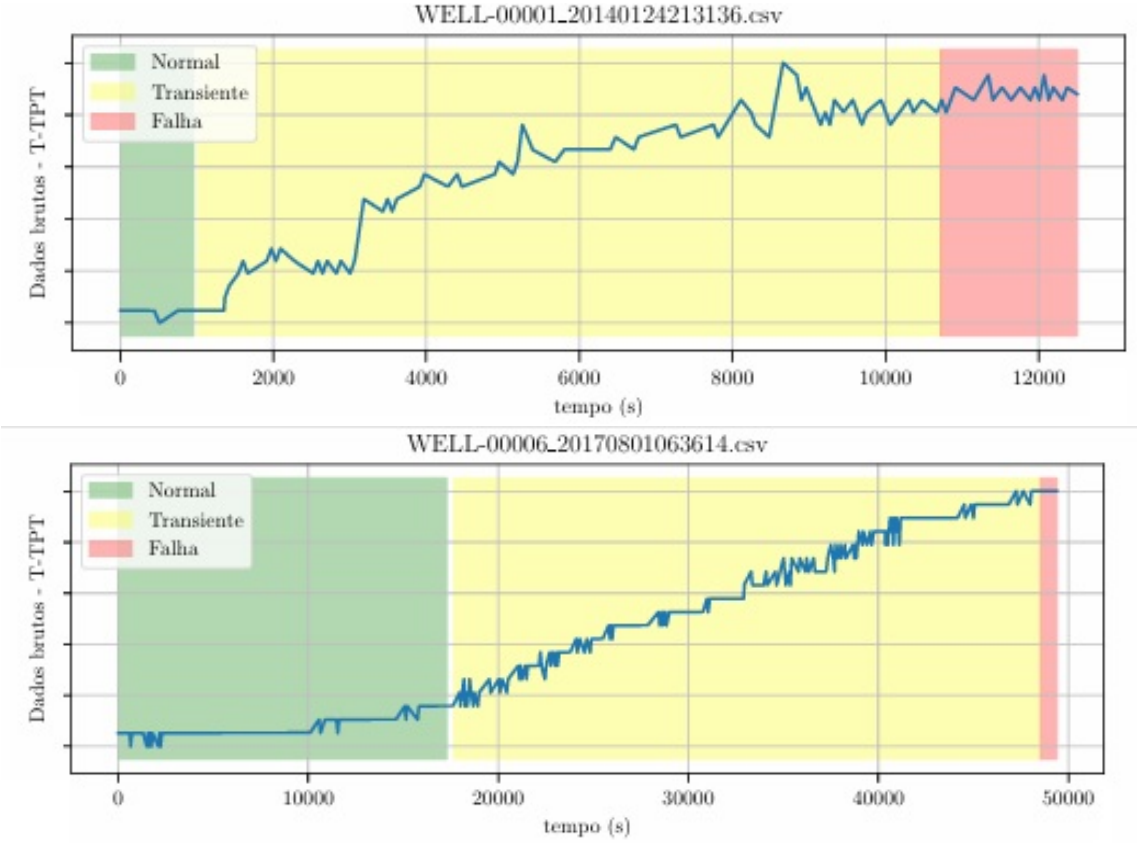

Figura 4.28: Gráficos de evolução da Variável T-TPT das duas instâncias reais integrantes da base de dados de teste, WELL-00001201401242131 36.csv e WELL-00006201708010636 14.csv.

4.7 Conclusões dos Resultados

Neste estudo, foram abordadas questões orientadoras com o objetivo de melhorar a eficiência e a precisão do sistema de classificação de eventos em poços de petróleo. As principais conclusões obtidas a partir das análises realizadas são as seguintes:

A análise da separação das amostras normais demonstrou que há um impacto na taxa de alarmes perdidos para certas classes de falha. Ajustes adicionais são necessários para minimizar o impacto negativo na detecção de falhas específicas e melhorar a acurácia de classificação das amostras normais que antecedem os eventos da falha.

A análise da dinâmica temporal, excluindo amostras em regime permanente de falha, não revela que o classificador pode se tornar sensível e eficaz na detecção precoce de falhas, contudo é observado que alguns eventos da Classe 0 tem similaridade com eventos da Classe 4, tornando difícil a separação.

A utilização de eventos simulados desempenha um papel na melhoria do modelo de classificação. No entanto, é essencial garantir que os eventos simulados sejam representativos e capturem as nuances dos dados reais. A combinação de instâncias

reais e simuladas mostrou-se a abordagem mais eficaz.

Ao utilizar a medida F1-ponderada como critério de avaliação em vez da acurácia balanceada, resulta em uma elevação na taxa de alarmes perdidos, no entanto em redução na taxa de falsos alarmes. Em algumas classes, o desempenho do classificador foi numericamente superior ao do classificador de referência.

A análise da quantidade mínima de experiências necessárias revelou que 20 experiências são suficientes para a detecção de eventos indesejáveis, onde pode ser constatada a queda no desempenho de classificação nas Classes 1, 2 e 6, sem prejuízo de detecção de falha, porém uma elevação na acurácia da Classes 3 (de 0,993 para 0,999 com 20 experiências) e 5 (de 0,988 para 0,994 com 20 experiências).

A análise de agrupamento para a Classe 1 mostrou que a Variável T-TPT indicou uma distinção clara entre os grupos. Isto sugere que diferentes características operacionais ou condições físicas dos poços influenciam a dinâmica de falha. Embora o sigilo da localização dos poços tenha limitado uma análise mais detalhada, o agrupamento revelou a possibilidade de associar determinados comportamentos de temperatura com tipos específicos de poços.

Em geral, este estudo apresentou uma melhor compreensão dos problemas apresentados, nas minúcias das dinâmicas temporais das amostras normais e das amostras de falha, da representatividade das instâncias simuladas e sobre a formação de *clusters* na Classe 1.

Capítulo 5

Conclusões

Este trabalho teve como objetivo a análise exploratória do 3W para a detecção de falhas em operações de poços de petróleo de reservatórios, usando a base de dados 3W, com foco em manutenção preditiva. Foram apresentados desafios do 3W e propostos temas de estudo, usando como ferramenta principal o MAIS.

No Capítulo 3 foi apresentada uma configuração de classificador multiclasse, e seu resultado empregado como referência para comparação com os resultados de experimentos apresentados neste estudo. A pesquisa se concentrou na dinâmica temporal das amostras normais e amostras de falha em regime estacionário; na busca de hiperparâmetros usando outra medida de desempenho, considerando o desbalanceamento das classes de falha; na influência dos dados simulados sobre a classificação dos dados reais; na busca de um modelo de treinamento rápido e com elevada acurácia para ser posto em produção; e finalmente, na análise de possível existência de *clusters* e suas causas na Classe 1.

Os resultados desta pesquisa foram apresentados no Capítulo 4. Quando as amostras normais, que antecedem as amostras de falha, foram separadas em outra classe (9), ficou evidente a necessidade de refinar os parâmetros do modelo para melhorar a distinção entre amostras normais e amostras de falha, principalmente da Classe 4.

A exclusão das amostras de falha em regime estacionário desafiou o classificador a focar na detecção precoce de falhas durante o período transiente, apresentando resultados de desempenho inferiores aos obtidos com o classificador de referência. Este procedimento exigiu a exclusão das instâncias das Classes 3 e 4, e revelou que 3,87% das instâncias da Classe 0 e 11,90% das instâncias da Classe 4 apresentam similaridade.

A busca de hiperparâmetros usando F1-ponderada como medida de desempenho resultou em um aumento na taxa de alarmes perdidos em relação ao Experimento de Referência, causado pela redução na acurácia das Classes 2, 3 e 4.

A análise a respeito da influência dos dados simulados sobre a classificação dos

dados reais revelou que as instâncias simuladas das Classes 1, 6 e 8 trazem pouca representatividade para o conjunto.

Nos experimentos de busca de um modelo de treinamento rápido e com elevada acurácia foi demonstrado que um experimento com 20 experiências é suficiente para realizar o treinamento do *dataset*, com desempenho muito próximo ao obtido com 100 experiências.

No que diz respeito ao agrupamento das falhas, foi identificado que a presença de *clusters* na Classe 1 (Aumento Abrupto de BSW) pode fornecer informações adicionais para a melhoria do processo de detecção. Há a suspeita de que o *cluster* com eventos, onde há elevação de temperatura, está relacionado com poços horizontais, caracterizados por entrada de água aquecida.

Recomendação para trabalhos futuros. Lacunas ainda precisam ser investigadas e melhorias podem ser sugeridas para pesquisas futuras.

Primeiramente, a influência dos dados simulados versus os dados reais no processo de treinamento dos modelos. Embora este estudo tenha mostrado que a inclusão de instâncias simuladas pode melhorar o desempenho do classificador, seria interessante desenvolver experimentos que avaliem detalhadamente o impacto de diferentes proporções de dados simulados versus reais. Nestes experimentos podem ser gerados gráficos de alarme de instâncias reais e de simuladas visando explicitar as diferenças na detecção.

Além disso, a questão da segmentação dos eventos de falha com base em características temporais, como a identificação de *clusters* nas falhas, apresenta grande potencial para pesquisas futuras. Estudos adicionais poderiam investigar em maior profundidade as razões para a formação desses *clusters* e como eles podem ser usados para criar modelos de predição mais precisos e específicos para cada tipo de falha.

Finalmente, o uso de técnicas de manutenção preditiva em tempo real, com integração a sistemas de monitoramento industrial, representa uma direção relevante a ser explorada. A criação de *pipelines* automatizados que integrem a coleta, o processamento e a análise dos dados dos sensores em tempo real, com a implementação de modelos preditivos, poderia levar a soluções mais ágeis e eficientes para a mitigação de falhas operacionais em ambientes de petróleo e gás.

Por fim, este trabalho demonstrou que, embora os algoritmos de aprendizado de máquina aplicados ao monitoramento de poços de petróleo apresentem grande potencial para a detecção precoce de falhas, a implementação prática de um sistema de manutenção preditiva depende de uma série de fatores, como a qualidade e a quantidade de dados disponíveis, o balanceamento entre classes e a capacidade de generalização dos modelos para novas instâncias. O desenvolvimento de técnicas que levem em consideração esses desafios é essencial para garantir a segurança e a continuidade das operações em ambientes de alta complexidade, como os poços de

petróleo *offshore*.

Referências Bibliográficas

[1] DIAS, T. L. B., MARINS, M. A., PAGLIARI, C. L., et al. "Development of Oil-well Fault Classifiers Using a Wavelet-based Multivariable Approach in a Modular Architecture", *SPE Journal*, 2024.

[2] MARINS, M. A., BARROS, B. D., I.H.SANTOS, et al. "Fault detection and classification in oil wells and production/service lines using random forest", *Journal of Petroleum Science and Engineering*, 2021.

[3] TURAN, E. M., JASCHKE, J. "Classification of undesirable events in oil well operation", *Proceedings of 23rd International Conference on Process Control*, p. 157–162, 2021.

[4] GATTA, F., GIAMPAOLO, F., CHIARO, D., et al. "Predictive maintenance for offshore oil wells by means of deep learning features extraction", *Expert Systems e13128*, 2022.

[5] AGÊNCIA NACIONAL DE PETRÓLEO, G. N. E. B. *Dados estatíticos / Produção de petróleo e gás natural / Produção por poço 2023*. Relatório técnico, AGÊNCIA NACIONAL DE PETRÓLEO, GÁS NATURAL E BIOCOMBUSTÍVEIS, 2023. Disponível em: <`https://www.gov.br/anp/pt-br/centrais-de-conteudo/dados-estatisticos`>. Acesso em: 15/02/2024.

[6] DEPARTMENT OF ENERGY, U. S. A. *Operations and maintenance best practices: a guide to achieving operational efficiency.* Relatório técnico, Federal Energy Management Program, 2002. Disponível em: <`https://www.pnnl.gov/main/publications/external/technical_reports/PNNL-13890.pdf`>.

[7] MELO, A., CÂMARA, M. M., PINTO, J. C. "Data-Driven Process Monitoring and Fault Diagnosis: A Comprehensive Survey", *Processes*, v. 12, n. 2, pp. 251, 2024. doi: 10.3390/pr12020251.

[8] JARDINE, A., LIN, D., BANJEVIC, D. "A review on machinery diagnostics and prognostics implementing condition-based maintenance", *Mechanical Systems and Signal Processing*, pp. 1483–1524, 2006.

[9] VARGAS, R. E. V., MUNARO, C. J., CIARELLI, P. M., et al. "A realistic and public dataset with rare undesirable real events in oil wells." *Journal of Petroleum Science and Engineering 181*, p. 62–77, 2019.

[10] RENGASWAMYB, R., YINC, K., KAVURID, S. N., et al. "A review of process fault detection and diagnosis Part I: Quantitative model-based methods", *Computers and Chemical Engineering 27 293 /311*, 2003.

[11] PETROBRAS. *Open Lab*. Relatório técnico, PETROBRAS, 2022. Disponível em: <`https://conexoes-inovacao.petrobras.com.br/modulo-open-lab`>. Acesso em: 13/11/2024.

[12] PETROBRAS. *3W*. Relatório técnico, PETROBRAS, abr 2019. Disponível em: <`https://github.com/petrobras/3W`>.

[13] SHEWHART, W. A. *Economic Control of Quality of Manufactured Product*. New York, Van Nostrand Reinhold, 1931.

[14] HOTELLING, H. "Multivariate Quality Control", *Techniques of Statistical Analysis*, pp. 111–184, 1947.

[15] XIA, X., PAN, X., LI, N., et al. "GAN-based anomaly detection: A review", *Neurocomputing*, v. 493, pp. 497–535, 2022.

[16] TARASSENKO, L., CLIFTON, D., BANNISTER, P., et al. "Novelty Detection". In: *Encyclopedia of Structural Health Monitoring*, cap. 35, John Wiley & Sons, Ltd, 2009.

[17] XUAN, X., MURPHY, K. "Modeling changing dependency structure in multivariate time series". In: *Proceedings of the Twenty-Fourth International Conference, ICML 2007*, pp. 1055–1062, Corvallis, Oregon, USA, 2007.

[18] DINIZ, P. S. R., NETTO, S. L., SILVA, E. A. B. "Filter banks". In: *Digital Signal Processing: System Analysis and Design*, 2ª ed., cap. 9, pp. 556–604, Porto Alegre, Cambridge University Press, 2010. ISBN: 9780521887755.

[19] ZHOU, P.-Y., CHAN, K. C. "A feature extraction method for multivariate time Series Classification Using Temporal Patterns". In: *Advances in Knowledge Discovery and Data Mining*, 2015.

[20] ANDREOLLI, I. *Introdução à Elevação e Escoamento Monofásico e Multifásico de Petróleo*. Brasil, Interciência, 2016.

[21] GRØDAHL, S. I. *Small scale multiphase flow experiments on surge waves in horizontal pipes*. Tese de Mestrado, Norwegian University of Science and Technology, 2014.

[22] BRODERSEN, K. H., ONG, C. S., ANDJOACHIM M. BUHMANN, K. E. S. "The balanced accuracy and its posterior distribution", *International Conference on Pattern Recognition*, 2010.

[23] SIBLINI, W., FRRY, J., HE-GUELTON, L., et al. "Master your Metrics with Calibration", *Worldline*, 2020. Disponível em: <`https://arxiv.org/pdf/1909.02827`>.

[24] SCIKIT_LEARN. *sklearn.metrics.f1_score*. Relatório técnico, Scikit_learn, 2024. Disponível em: <`https://scikit-learn.org/0.16/modules/generated/sklearn.metrics.f1_score.html`>.

[25] ROUSSEEUW, P. J. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis", *Journal of Computational and Applied Mathematics*, v. 20, pp. 53–65, 1987. doi: 10.1016/0377-0427(87)90125-7.

[26] GUILLEMOT, T., FOUCHET, A., LAYTON, R. *Unsupervised evaluation metrics*. Relatório técnico, scikit-learn, 2024. Disponível em: <`https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html`>.

[27] HO, T. K. "Random decision forests", *IEEE Proceedings of the Third International Conference on Document Analysis and Recognition*, 1995.

[28] S.HAYKIN. *Neural Networks and Learning Machines*. Hamilton, Ontario, Canada, Pearson – Prentice Hall, 2009.

[29] ALPAYDIN, E. *Introduction to Machine Learning*. Third. ed. New York, The MIT Press., 2014. ISBN: 978-0-262-02818-9.

[30] BRAUN, N., NEUFFER, J., KEMPA-LIERH, A. W., et al. "Time Series Feature Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package)", *Neurocomputing*, pp. 72–77, 2018.

[31] BISHOP, C. M. *Pattern Recognition and Machine Learning*. Cambridge CB3 0FB, U.K., Springer-Verlag, 2006. ISBN: ISBN-10: 0-387-31073-8.

[32] MURPHY, K. P. "Latent linear models; Adaptive basis function models". In: *Machine Learning: A Probabilistic Perspective*, cap. 12 e 16, Cambridge, Massachusetts, The MIT Press, 2012.

[33] LI, X., MA, H., LI, X. "Deep representation clustering-based fault diagnosis method with unsupervised data applied to rotating machinery", *Mechanical Systems and Signal Processing, 143, 106825*, 2020.

[34] MACHADO, A. P. F., VARGAS, R. E. V., CIARELLI, P. M., et al. "Improving performance of one-class classifiers applied to anomaly detection in oil wells", *Journal of Petroleum Science and Engineering*, 2022.

[35] NGUYEN, H. D., TRAN, K. P., THOMASSEY, S., et al. "Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management", *International Journal of Information Management*, 2021.

[36] ARANHA, P., LOPES, L., SOBRINHO, E., et al. "A System to Detect Oilwell Anomalies Using Deep Learning and Decision Diagram Dual Approach", *SPE Journal*, 2024.

[37] MACHADO, A. P. F., MUNARO, C., CIARELLI, P. M., et al. "Time series clustering to improve one-class classifier performance", *Expert Systems with Applications*, 2024.

[38] MÜLLER, M. "Dynamic Time Warping". In: *Information Retrieval for Music and Motion*, cap. 4, Institut für Informatik III, Universität Bonn, Römerstr. 164, 53117 Bonn, Germany, Springer-Verlag Berlin Heidelberg, 2007.

[39] TAVENARD, R., J.FAOUZI, G.VANDEWIELE, et al. "Tslearn, A machine learning toolkit for time series data", *Journal of Machine Learning Research*, 2020. Disponível em: <`http://jmlr.org/papers/v21/20-091.html`>.

[40] KEOGH, E., LIN, J. "Clustering of time-series subsequences is meaningless: implications for previous and future research", *Knowledge and Information Systems*, 2005. Disponível em: <`https://www.scopus.com/record/display.uri?eid=2-s2.0-21844471761&origin=inward`>.

[41] UFRJ/COPPE/PEE. *MODULAR ARTIFICIAL INTELLIGENCE SYSTEM (MAIS)*. Relatório técnico, UFRJ/COPPE/PEE, 2022. Disponível em: <`https://github.com/petrobras/3W`>. Acesso em: 15/02/2024.

[42] KE, G., MENG, Q., FINLEY, T., et al. "LightGBM: A highly efficient gradient boosting decision tree", *31st International Conference on Neural Information Processing Systems, NIPS'2017*, p. 3149–3157, 2017.

[43] R.BARDENET, Y.BENGIO, B.KÉGL, et al. "Algorithms for hyperparameter optimization", *Proceedings of the 24th International Conference on Neural Information Processing Systems, NIPS'11, Curran Associates Inc.,*, p. 2546–2, 2011.

[44] FRIEDMAN, J. H. "Greedy function approximation: a gradient boosting machine", *Annals of statistics,*, p. 1189–1232, 2001.

[45] CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L., et al. *Introduction to Algorithms*. 3rd ed. Cambridge, MA, MIT Press, 2009.

[46] THOMAZ, J. E., TRIGGIA, A. A., CORREIA, C. A., et al. *Fundamentos de Engenharia de Petróleo*. Rio de Janeiro, Editora Interciência, 2001.

[47] YOSHIOKA, K., ZHU, D., HILL, A. D., et al. "Prediction of Temperature Changes Caused by Water or GasEntry Into a Horizontal Well", *SPE Production and Operations*, 2007.

[48] DE AZEVEDO, A. A. M. *Módulos complementares do MAIS*. Relatório técnico, UFRJ/COPPE/PEE, 2024. Disponível em: <https://github.com/betomazevedo/Mestrado-em-Engenharia-Eletrica/blob/main/README.md>.

[49] DE AZEVEDO, A. A. M. *Resultados dos Experimentos*. Relatório técnico, mlflow, 2024. Disponível em: <https://dagshub.com/betomazevedo39/3W.mlflow/#/experiments/154?searchFilter=&orderByKey=attributes.start_time&orderByAsc=false&startTime=ALL&lifecycleFilter=Active&modelVersionFilter=All+Runs&datasetsFilter=W10%3D>. Acesso em: 13/11/2024.

[50] FENG, C., WANG, H., LU, N., et al. "Log-transformation and its implications for data analysis", *Shanghai Arch Psychiatry 26 (2): 105–109*, 2014.

[51] BERGSTRA, J., R.BARDENET, Y.BENGIO, et al. "Random search for hyper-parameter optimization", *Journal of Machine Learning Research*, p. 281–305, 2012.

[52] MACHADO, A. P. F., MUNARO, C., CIARELLI, P. M., et al. *Time series clustering to improve one-class classifier performance [Source Code]*. Relatório técnico, Code ocean, 2024.

[53] PYTHON. *pickle — Python object serialization.* Relatório técnico, Python, 2024. Disponível em: <`https://docs.python.org/3/library/pickle.html`>. Acesso em: 13/11/2024.