

UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE TECNOLOGIA
CURSO DE GRADUAÇÃO EM ENGENHARIA QUÍMICA

VICTOR HUGO LOPES BENEDITO

**DESENVOLVIMENTO DE UM SISTEMA ENSEMBLE PARA A DETECÇÃO E
CLASSIFICAÇÃO DE ANOMALIAS EM POÇOS DE PETRÓLEO OFF-SHORE**

Maringá

2025

UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE TECNOLOGIA
CURSO DE GRADUAÇÃO EM ENGENHARIA QUÍMICA

VICTOR HUGO LOPES BENEDITO

**DESENVOLVIMENTO DE UM SISTEMA ENSEMBLE PARA A DETECÇÃO E
CLASSIFICAÇÃO DE ANOMALIAS EM POÇOS DE PETRÓLEO OFF-SHORE**

Trabalho de Conclusão de Curso de Graduação apresentado ao
Curso de Graduação em Engenharia Química da Universidade
Estadual de Maringá como parte dos requisitos para a obtenção do
título de Engenheiro Químico.

Orientador(a): Dr. Leandro Vitor Pavão

Maringá

2025

VICTOR HUGO LOPES BENEDITO

**DESENVOLVIMENTO DE UM SISTEMA ENSEMBLE PARA A DETECÇÃO E
CLASSIFICAÇÃO DE ANOMALIAS EM POÇOS DE PETRÓLEO OFF-SHORE**

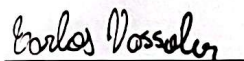
Trabalho de Conclusão de Curso de Graduação apresentado ao Curso de Graduação em Engenharia Química da Universidade Estadual de Maringá como parte dos requisitos para a obtenção do título de Engenheiro Químico.

APROVADO EM: 30 DE Janeiro DE 2025

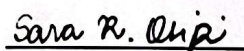
BANCA EXAMINADORA



Dr. Leandro Vitor Pavão - (DEQ/UEM)



Carlos Henrique Vassoler - (DEQ/UEM)



Dra. Sara Regina Osipi - (DEQ/UEM)

AGRADECIMENTOS

Primeiramente, agradeço a minha família, especialmente meus pais, pelo amor, apoio e dedicação em me proporcionar as condições necessárias para alcançar meus objetivos. Vocês são minha maior inspiração e motivação.

Aos meus professores do departamento de Engenharia Química, em especial o prof. Leandro, pela orientação durante toda a graduação em diversos projetos e pela liberdade no desenvolvimento desse projeto.

Aos meus amigos e colegas, cujo apoio foi essencial em diversos momentos durante essa trajetória.

Por fim, agradeço a todas as pessoas que, direta ou indiretamente, contribuíram para a realização deste TCC. Cada palavra de incentivo, cada ajuda oferecida, foi essencial para esta conquista.

RESUMO

A detecção e classificação de anomalias em equipamentos industriais são fundamentais para garantir a eficiência operacional e reduzir riscos em sistemas críticos, como poços de petróleo offshore. Utilizando os dados fornecidos pelo projeto 3W da Petrobras, este trabalho apresenta o desenvolvimento de um modelo baseado em técnicas de aprendizado de máquina, implementando um ensemble de algoritmos supervisionados para identificar e categorizar falhas nesses equipamentos. Abordando sete dos nove eventos de falhas disponíveis no projeto, foram desenvolvidos sete modelos para detecção de anomalias, cada um adaptado às características específicas de cada evento, e um modelo para a classificação dessas anomalias. Os resultados demonstraram a eficácia dos modelos propostos, tanto na identificação quanto na classificação precisa de anomalias, oferecendo uma abordagem prática e escalável para a gestão de falhas em poços de petróleo offshore. Este trabalho contribui para a aplicação de soluções baseadas em aprendizado de máquina, melhorando a confiabilidade e segurança das operações industriais.

Palavras-chave: Detecção de falhas, aprendizado de máquina, poços de petróleo offshore, ensemble de modelos, tecnologia.

ABSTRACT

Anomaly detection and classification in industrial equipment are essential to ensure operational efficiency and reduce risks in critical systems, such as offshore oil wells. Using data provided by Petrobras' 3W project, this work presents the development of a model based on machine learning techniques, implementing an ensemble of supervised algorithms to identify and categorize faults in this equipment. Addressing seven out of the nine fault events available in the project, seven models were developed for anomaly detection, each adapted to the specific characteristics of each event, and one model for anomaly classification. The results demonstrated the effectiveness of the proposed models in both accurately identifying and classifying anomalies, offering a practical and scalable approach to fault management in offshore oil wells. This work contributes to the application of machine learning-based solutions, enhancing the reliability and safety of industrial operations.

Keywords: Fault detection, machine learning, offshore oil wells, ensemble models, technology.

LISTA DE FIGURAS

Figura 1 - Formação do campo da Ciência de dados.....	11
Figura 2 - Representação esquemática do funcionamento de poços off-shore	14
Figura 3 - Esquema de tubulação e sensores da “árvore de natal”.....	16
Figura 4 - Distribuição dos Eventos Indesejados no banco de dados da 3W versão 2.0.0.....	18
Figura 5 - Fluxograma de tipos de aprendizado de máquina.....	21
Figura 6 - Comportamento da função logística em um plano 2D	22
Figura 7 - Árvore de decisão construída com LightGBM e XGBoost.....	24
Figura 8 - Impacto de differencing nos dados.	26
Figura 9 - Impacto da transformação logarítmica nos dados	26
Figura 10 - Fluxograma de funcionamento do SHAP	28
Figura 11 - Retorno do desempenho global das características para o modelo.....	29
Figura 12 - Validação cruzada tradicional.....	30
Figura 13 - Validação cruzada estratificada	31
Figura 14 – Esquemática de funcionamento do Grid Search	32
Figura 15 - Esquemática de funcionamento do Bayesian Search.....	33
Figura 16 - Estrutura do sistema de reconhecimento de anomalias	34
Figura 17 - Sistema de Reconhecimento e Classificação de Anomalias.....	34
Figura 18 - Distribuição das anomalias para cada modelo desenvolvido	41
Figura 19 - Matrizes de confusão dos modelos otimizados	43
Figura 20 - Modelo 1 reconhecendo a ausência de aumento repentino de sedimentos básicos e água durante teste de operação normal.....	44
Figura 21 - Modelo 1 reconhecendo a anomalia de aumento repentino de sedimentos básicos e água com pouco atraso durante teste de detecção de anomalia.....	44
Figura 22 - Modelo 1 apresentando leve oscilação durante teste de detecção de anomalia....	45
Figura 23 - Modelo 8 apresentando oscilações não contínuas por um período razoável durante teste de operação normal	45
Figura 24 - Modelo 8 apresentando oscilações contínuas por um período razoável durante teste de operação normal.....	46
Figura 25 - Modelo 8 apresentando muitas oscilações durante todo período de teste de detecção de anomalia.....	46
Figura 26 - Avaliação SHAP para o modelo 1 - beeswarm plot	48
Figura 27 - Avaliação SHAP para o modelo 1 - bar plot	48

Figura 28 - Teste de monitoramento do modelo 1 sobre a operação normal	49
Figura 29 - Teste de monitoramento do modelo 1 sobre ocorrência de evento	49
Figura 30 - Avaliação SHAP para o modelo 2 - beeswarm plot	50
Figura 31 - Avaliação SHAP para o modelo 2 - bar plot	50
Figura 32 - Teste de monitoramento do modelo 2 sobre a operação normal	51
Figura 33 - Teste de monitoramento do modelo 2 sobre ocorrência de evento	51
Figura 34 - Avaliação SHAP para o modelo 5 - <i>beeswarm plot</i>	52
Figura 35 - Avaliação SHAP para o modelo 5 - bar plot	52
Figura 36 - Teste de monitoramento do modelo 5 variáveis montante CKP	53
Figura 37 - Teste de monitoramento do modelo 5 variáveis jusante CKP	53
Figura 38 - Avaliação SHAP para o modelo 6 - beeswarm plot	54
Figura 39 - Avaliação SHAP para o modelo 6 - bar plot	54
Figura 40 - Teste de monitoramento de operação normal	55
Figura 41 - Teste de monitoramento de evento no modelo 6	55
Figura 42 - Avaliação SHAP para o modelo 9 - beeswarm plot	56
Figura 43 - Avaliação SHAP para o modelo 9 - bar plot	56
Figura 44 - Teste de monitoramento de operação normal	57
Figura 45 - Teste de monitoramento de evento no modelo 9	57
Figura 46 - Matriz de confusão do modelo otimizado de classificação	58

LISTA DE TABELAS

Tabela 1 - Variáveis da versão 1.0.0 do banco de dados 3W	14
Tabela 2 - Variáveis da versão 2.0.0 do banco de dados 3W	15
Tabela 3 - Distribuição de Anomalias da base de dados 3W versão 2.0.0	17
Tabela 4 - Separação dos arquivos de treino e teste	35
Tabela 5 - Análise de dados a serem tratados.....	36
Tabela 6 - Tempo médio de regime transiente para cada evento	42
Tabela 7 - Resultados dos modelos de detecção de anomalias otimizados	42
Tabela 8 - Estatísticas de monitoramento por modelo.....	47
Tabela 9 - Resultados da otimização do modelo de classificação	58
Tabela 10 - Hiperparâmetros usados nesse trabalho para os modelos LGBM.....	59
Tabela 11 - Hiperparâmetros usados nesse trabalho para os modelos Regressão Logística ...	59

SUMÁRIO

1	INTRODUÇÃO.....	10
1.1	CONTEXTUALIZAÇÃO	10
1.1.1	Ciência de dados	10
1.1.2	Machine Learning	11
1.1.3	Iniciativa 3W.....	12
1.2	OBJETIVOS	12
1.3	JUSTIFICATIVA.....	13
2	FUNDAMENTAÇÃO TEÓRICA	13
2.1	POÇOS DE PETROLEO <i>OFF-SHORE</i>	13
2.2	PROJETO 3W	14
2.2.1	Descrição Dos Eventos	18
2.3	APRENDIZADO DE MÁQUINA	20
2.3.1	Aprendizado Supervisionado e Problemas de classificação	21
2.3.2	Engenharia de Features	25
2.3.3	Validação Cruzada	30
2.3.4	Otimização de Hiperparâmetros.....	31
3	METODOLOGIA	33
3.1	Estrutura do Sistema de monitoramento	33
3.2	Pré-processamento	34
3.2.1	Extração de Características.....	37
3.3	Escolha de Modelo	38
3.3.1	Light Gradient Boosting Machine (LGBM)	38
3.3.2	Regressão Logística	39
3.4	Extração de dados da detecção de Anomalias	39
4	RESULTADOS	40
4.1	Otimização e Construção dos modelos de Detecção de anomalias	40
4.2	Análise de Impacto de características nos modelos propostos	47
4.3	Modelo de classificação de anomalias	57
4.4	Reprodutibilidade e Hiperparâmetros dos modelos	59
5	CONCLUSÃO	60
6	REFERENCIAS.....	62

1 INTRODUÇÃO

Conforme apontam Milani Júnior, Bomtempo e Pinto Júnior (2007), a indústria do petróleo desempenha um papel vital na economia global e é essencial para o progresso socioeconômico do Brasil. Esse setor abrange seis processos principais: exploração, desenvolvimento, produção, refino, transporte e distribuição. Esses processos podem ser organizados de diferentes maneiras, sendo a mais comum a divisão em *upstream* (exploração, desenvolvimento e produção) e *downstream* (refino, transporte e distribuição).

Dada a importância e competitividade do setor, as empresas do ramo petrolífero estão adotando cada vez mais tecnologias digitais. Esses sistemas aumentam a eficiência, promovem a integração de sistemas, e facilitam o processamento e a análise de dados, visando objetivos como tomada de decisões, aumento da produtividade e maior flexibilidade operacional.

A principal impulsionadora dessas inovações é a chamada Indústria 4.0, que está se tornando cada vez mais presente em nosso cotidiano. De acordo com Sony e Naik (2020), a Indústria 4.0 representa a convergência de tecnologias que promovem a interação entre os domínios físicos, digitais e biológicos por meio de sistemas físicos embutidos. Essa revolução faz com que fábricas atuem de maneira mais independente, tentando assim reduzir ao máximo o contato humano, principalmente em áreas de maior risco.

As indústrias de todos os ramos têm investido cada vez mais em tecnologias avançadas com o objetivo de aumentar sua produtividade e competitividade. Entre as áreas de destaque estão a automação industrial, que é um pilar fundamental para toda implementação de tecnologia, seguida da inteligência artificial (IA) e a Internet Industrial das Coisas (IIoT, do inglês *Industrial Internet of Things*), que têm impulsionado uma transformação significativa nos processos produtivos. Muitas inovações estão proporcionando melhorias expressivas na eficiência operacional, na qualidade e padronização dos produtos e na segurança do ambiente de trabalho.

1.1 CONTEXTUALIZAÇÃO

Para facilitar o entendimento da tese, esta seção abordará conceitos essenciais sobre ciência de dados, aprendizado de máquina, detecção de anomalias e a iniciativa 3W.

1.1.1 CIÊNCIA DE DADOS

A Ciência de Dados é uma área interdisciplinar que une princípios da estatística, ciência da computação e análise de dados, com o objetivo de extrair informações relevantes e gerar conhecimento a partir de grandes volumes de dados. Valendo-se de métodos científicos, processos sistemáticos, algoritmos avançados e sistemas tecnológicos, essa disciplina é capaz

de examinar dados estruturados e não estruturados, identificar padrões lineares e não lineares permitindo com que algoritmos cada vez mais complexos sejam capazes de existir.

Figura 1 - Formação do campo da Ciência de dados



Fonte: Ribeiro (2018).

O propósito dessa área é transformar dados em conhecimento aplicável, permitindo assim a resolução de problemas complexos, a otimização de processos e a criação de inovações em múltiplos contextos. Seu impacto é amplamente perceptível em diversas áreas do conhecimento e setores da economia.

1.1.2 MACHINE LEARNING

Machine Learning é um subcampo da inteligência artificial focado no desenvolvimento de algoritmos e modelos que permitem aos computadores aprenderem a partir de dados. Em vez de serem explicitamente programados para executar tarefas específicas, os sistemas de Machine Learning analisam grandes volumes de dados, identificam padrões e fazem previsões ou tomam decisões com base nesses padrões. Como destacado por Lurdemir (2021), "o aumento da capacidade dos computadores atuais é parcialmente em razão das técnicas de Aprendizado de Máquina. Entretanto, não é de hoje que se deseja fazer que o computador aprenda." Essa perspectiva reforça a longa trajetória de esforços na busca por tornar os computadores capazes de aprender e agir de maneira autônoma, com base em dados.

Uma das aplicações mais importantes do Machine Learning é na detecção de falhas e segurança. Em ambientes industriais, por exemplo, os algoritmos podem monitorar

continuamente dados de sensores para identificar sinais de desgaste ou mau funcionamento de equipamentos, prevenindo falhas catastróficas antes que elas ocorram. Da mesma forma, em sistemas de segurança, o Machine Learning é usado para detectar comportamentos anômalos que possam indicar uma possível violação ou ataque cibernético. Esses sistemas podem aprender a reconhecer padrões de atividades normais e, quando algo fora do comum é detectado, eles podem alertar os administradores para que ações preventivas sejam tomadas.

1.1.3 INICIATIVA 3W

O projeto 3W é uma iniciativa pioneira da Petrobras voltada para a experimentação e desenvolvimento de abordagens baseadas em Machine Learning aplicadas à detecção e classificação de eventos indesejáveis em poços de petróleo offshore. Este projeto destaca-se como um marco importante, sendo o primeiro repositório da Petrobras publicado no *GitHub*, refletindo um movimento em direção à colaboração aberta e ao avanço tecnológico na indústria de óleo e gás.

O principal objetivo desse projeto é explorar soluções de aprendizado de máquina para enfrentar desafios críticos relacionados a eventos indesejáveis em poços de petróleo, como falhas em sistemas de elevação artificial ou problemas de *flow assurance*. A detecção precoce desses eventos pode reduzir significativamente custos de manutenção, minimizar perdas de produção e mitigar riscos ambientais e humanos (VAZ VARGAS *et al.*, 2019).

A motivação por trás do projeto está ligada aos desafios enfrentados na operação de poços de petróleo offshore. Eventos indesejáveis, como falhas em sistemas de elevação artificial ou problemas relacionados à garantia de escoamento (*flow assurance*), podem gerar prejuízos expressivos tanto do ponto de vista financeiro quanto ambiental e humano.

De acordo com Vaz Vargas *et al.* (2019), estima-se que perdas de produção possam alcançar até 5% em cenários críticos, enquanto os custos de manutenção, especialmente aqueles envolvendo o uso de sondas marítimas, podem ultrapassar US\$ 500 mil por dia. Além disso, as consequências de falhas catastróficas podem incluir graves danos ao meio ambiente, riscos à segurança das pessoas e prejuízos irreparáveis à imagem da empresa. Diante disso, identificar esses eventos precocemente é essencial para prevenir impactos severos e otimizar operações.

1.2 OBJETIVOS

Este trabalho tem como objetivo validar a utilização de aprendizado de máquina dentro da indústria de óleo e gás, por meio da criação de um modelo robusto capaz de identificar anomalias e classificá-las. São objetivos específicos deste trabalho:

- Obter um bom desempenho na identificação de anomalias;

- Obter um bom desempenho na classificação de eventos;
- Realizar uma análise exploratória de dados nos modelos, anomalias, eventos e características;

1.3 JUSTIFICATIVA

Fundamentada no contexto apresentado, e a partir da relevância crescente do uso de tecnologias avançadas, como o aprendizado de máquina, para solucionar problemas críticos na indústria de óleo e gás, foi desenvolvida a solução presente neste trabalho. A operação de poços offshore apresenta desafios complexos relacionados à identificação e gestão de anomalias, cuja detecção tardia pode resultar em consequências significativas, como perdas de produção, aumento de custos operacionais e riscos ambientais.

A utilização de aprendizado de máquina oferece um potencial transformador, permitindo a automação e o aprimoramento na detecção e classificação dessas anomalias, o que, por sua vez, pode melhorar a eficiência e a segurança operacional. Esse tipo de tecnologia possibilita identificar padrões e comportamentos que muitas vezes passam despercebidos em análises tradicionais, contribuindo para uma tomada de decisão mais assertiva e ágil.

Além disso, o desenvolvimento de soluções baseadas em aprendizado de máquina para a indústria de óleo e gás está alinhado com a tendência global de digitalização e inovação tecnológica no setor. O investimento em modelos inteligentes não apenas fortalece a capacidade preditiva e preventiva, mas também promove a sustentabilidade, ao minimizar os riscos de acidentes ambientais e reduzir os custos associados às operações de manutenção corretiva.

Dessa forma, este trabalho justifica-se pela contribuição potencial para o avanço tecnológico do setor, ao validar a aplicação prática de aprendizado de máquina para identificar e classificar anomalias.

2 FUNDAMENTAÇÃO TEÓRICA

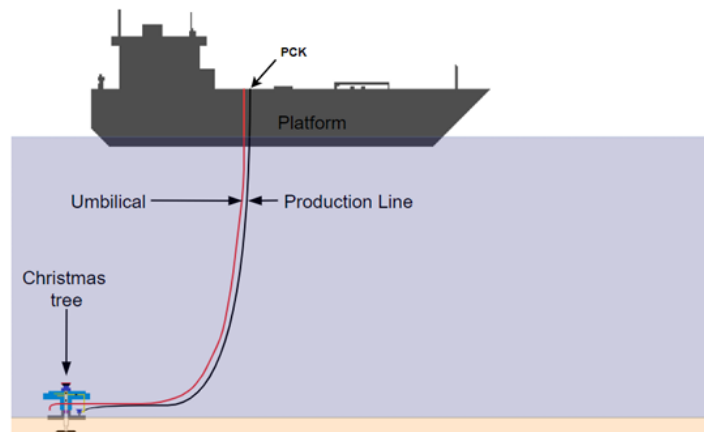
2.1 POÇOS DE PETRÓLEO *OFF-SHORE*

De acordo com Laik (2018), um poço de petróleo offshore é uma estrutura projetada para operar no fundo do mar, extraindo petróleo e gás de reservatórios, localizados abaixo do leito marinho. Este sistema complexo é composto por diferentes sistemas menores que desempenham funções específicas. Entre eles estão o tubo de produção, que transporta o fluido extraído até a superfície; a cabeça de poço, que garante a estabilidade e segurança estrutural durante as fases de perfuração e extração; e a "árvore de Natal", um equipamento instalado

sobre a cabeça de poço que regula o fluxo de produção e fornece acesso ao tubo de produção por meio de válvulas e sensores operados remotamente.

Laik (2018) ainda complementa que um componente fundamental desse sistema é o umbilical, um feixe de cabos e tubos submarinos que conecta a plataforma de produção à árvore de Natal no leito marinho. Ele desempenha um papel essencial na comunicação, integrando condutores elétricos para controle remoto de equipamentos, linhas hidráulicas para acionamento de dispositivos submarinos, e linhas de fluxo dedicadas à injeção de produtos químicos e recuperação de fluidos. A Figura 2, descreve um esquema de um poço *offshore*.

Figura 2 - Representação esquemática do funcionamento de poços *off-shore*



Fonte: Vaz Vargas *et al.* (2019).

2.2 PROJETO 3W

O conjunto de dados 3W é uma base de dados pública disponibilizada pela Petrobras no dia 30 de dezembro de 2022 (VAZ VARGAS *et al.*, 2019). Essa base é composta por dados reais, simulados e desenhados manualmente relacionados aos poços de petróleo em operação. Além disso, os dados incluem instâncias do poço de petróleo durante a operação normal e, mais importante, o registro de eventos indesejados no poço. A primeira versão do projeto contava apenas com apenas 24 poços de petróleo em seu banco, dos quais eram monitoradas 7 variáveis, como é mostrado na Tabela 1.

Tabela 1 - Variáveis da versão 1.0.0 do banco de dados 3W

Variável	Descrição
P-PDG	Pressão no PDG (permanent downhole gauge)
T- TPT	Temperatura no TPT (temperature and pressure transducer)
P-TPT	Pressão no TPT (temperature and pressure transducer)

T-JUS-PCK	Temperatura a jusante do PCK (production choke)
P-JUS-PCK	Pressão a jusante do PCK (production choke)
T-MON-PCK	Temperatura a montante do PCK (production choke)
P-MON-PCK	Pressão a montante do PCK (production choke)

Fonte: Vaz Vargas *et al.* (2019).

Com o constante apoio da Petrobras e parceiros no fomento do projeto, a versão 2.0.0 disponibilizada em 25 de julho de 2024 passou a ter informações de um total de 42 poços de petróleo e 20 novas variáveis monitoradas que são apresentadas na Tabela 2. Além disso, foi adicionado diversos novos registros para os eventos já existentes e a adição de um novo evento, apontado na Tabela 3.

Tabela 2 - Variáveis da versão 2.0.0 do banco de dados 3W

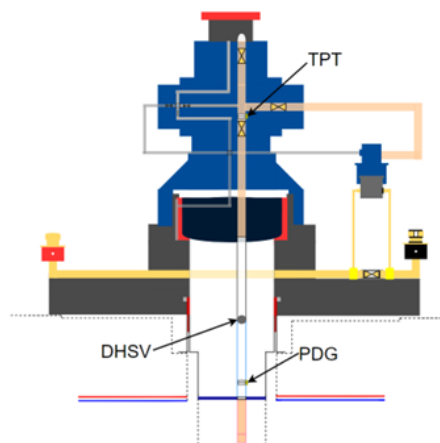
Variável	Descrição
ABER-CKGL	Abertura do GLCK (gas lift choke)
ABER-CKP	Abertura do PCK (production choke)
ESTADO-DHSV	Estado da DHSV (downhole safety valve)
ESTADO-M1	Estado da PMV (production master valve)
ESTADO-M2	Estado da AMV (annulus master valve)
ESTADO-PXO	Estado da PXO (pig-crossover) valve
ESTADO-SDV-GL	Estado da válvula de fechamento (shutdown valve) do gas lift
ESTADO-SDV-P	Estado da válvula de fechamento (shutdown valve) da produção
ESTADO-W1	Estado da PWV (production wing valve)
ESTADO-W2	Estado da AWV (annulus wing valve)
ESTADO-XO	Estado da válvula XO (crossover)
P-ANULAR	Pressão no anular do poço
P-JUS-BS	Pressão a jusante da SP (service pump)
P-JUS-CKGL	Pressão a jusante do GLCK (gas lift choke)
P-JUS-CKP	Pressão a jusante do PCK (production choke)
P-MON-CKGL	Pressão a montante do GLCK (gas lift choke)
P-MON-SDV-P	Pressão a montante da válvula de fechamento (shutdown valve) da produção

PT-P	Pressão a jusante da PWV (production wing valve) no tubo de produção
QBS	Vazão na SP (service pump)
QGL	Vazão de gas lift

Fonte: Vaz Vargas *et al.* (2019).

Sensores distribuídos no processo de extração esquematizado pelas Figuras 2 e 3, fazem medição de temperatura, pressão, vazão, status da válvula e a porcentagem de abertura das válvulas.

Figura 3 - Esquema de tubulação e sensores da “árvore de natal”.



Fonte: Vaz Vargas *et al.* (2019).

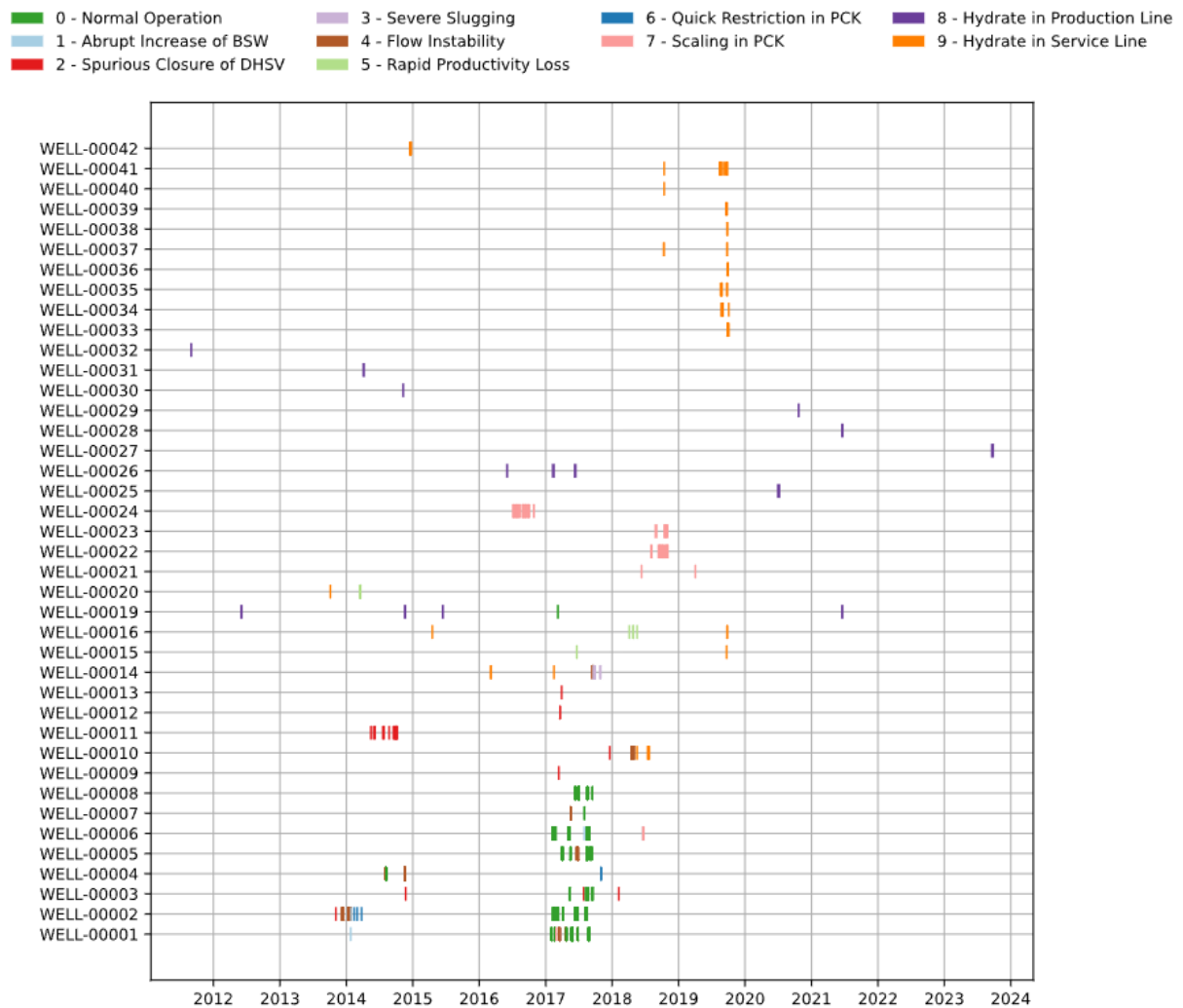
O sistema 3W armazena informações sobre nove categorias de eventos indesejáveis em poços de petróleo, dos quais sete são contemplados nesse trabalho. Fatores como presença de água, sedimentos, gás natural, além da razão e taxa de fluxo são essenciais para compreender esses eventos. O conjunto de dados incluem instâncias reais, simuladas e elaboradas manualmente. As instâncias reais foram obtidas diretamente do sistema de informações operacionais da Petrobras, enquanto as simuladas foram geradas por meio do OLGA, um simulador dinâmico amplamente utilizado na indústria petrolífera. Por outro lado, as instâncias criadas manualmente foram desenvolvidas por especialistas, com o objetivo de mitigar o desequilíbrio nos dados.

Tabela 3 - Distribuição de Anomalias da base de dados 3W versão 2.0.0

Evento	Fonte Real	Fonte Simulada	Fonte Desenhada a Mão	Total
0 - Normal	594	0	0	594
1 - Aumento Repentino de Sedimentos Básicos e Água	4	114	10	128
2 - Fechamento Irregular de Válvula de Segurança	22	16	0	38
3 - Intermittência Severa	32	74	0	106
4 – Instabilidade de Vazão	343	0	0	343
5 - Perda Rápida de Produtividade	11	439	0	450
6 - Restrição Rápida em CKP	6	215	0	221
7 - Presença de Incrustação em CKP	36	0	10	46
8 - Hidratos na Linha de Produção	14	81	0	95
9 - Hidratos na Linha de Serviço	57	150	0	207

Fonte: Autoria Própria.

Vale destacar que além das instâncias reais apresentarem uma distribuição desigual, como é destacado na Tabela 3, os eventos apresentam uma distribuição totalmente desigual entre si, ilustrado na Figura 4, com cerca de anomalias representando menos de 2% dos registros. Cada evento no conjunto de dados é representado como uma sequência contínua de observações, classificadas em três estados distintos: normal, transiente de falha e estacionário de falha. No estado normal, não há indícios de comportamento anômalo. No estado transiente de falha, as dinâmicas associadas ao evento indesejável estão em progresso. Após a cessação dessas dinâmicas, ocorre o término do estado estacionário de falha. Essa organização dos estados foi projetada para facilitar a detecção precoce de falhas.

Figura 4 - Distribuição dos Eventos Indesejados no banco de dados da 3W versão 2.0.0

Fonte: Autoria própria.

2.2.1 DESCRIÇÃO DOS EVENTOS

A subseções a seguir apresentam as descrições dos eventos aos quais vão ser estudadas neste presente trabalho.

2.2.1.1 AUMENTO REPENTINO DE SEDIMENTOS BÁSICOS E ÁGUA (BSW)

O Basic Sediment and Water (BSW) é definido como a razão entre a vazão de água e sedimentos produzidos e a vazão de líquido produzido, ambas medidas em condições padrão (ANDREOLLI, 2016; ABASS; BASS, 1988 apud VARGAS *et al.*, 2019).

Projetos de tubulações de poços consideram informações sobre o BSW, que são inicialmente obtidas por meio de modelagem de reservatório. Ao longo da vida útil de um poço, espera-se que o BSW aumente, devido ao maior volume de água produzido, proveniente tanto do aquífero natural do reservatório quanto da injeção artificial para evitar o declínio da produção. No entanto, um aumento abrupto do BSW pode gerar uma série de problemas, como

dificuldades na garantia de escoamento, redução da produção de óleo, elevação do petróleo, complicações no processamento na instalação industrial e impacto no fator de recuperação (ANDREOLLI, 2016 apud VARGAS *et al.*, 2019).

2.2.1.2 FECHAMENTO IRREGULAR DE VÁLVULA DE SEGURANÇA (DHSV)

A válvula de segurança de fundo de poço, também conhecida como Downhole Safety Valve (DSV), é instalada na coluna de produção de um poço com a finalidade de garantir o seu fechamento em emergências, como quando a unidade de produção é fisicamente desconectada do poço. O fechamento irregular dessa válvula pode ser causado por diversos fatores, como problemas de comunicação, problemas no atuador, condições operacionais fora da referência etc. Essa válvula é mantida aberta por um atuador hidráulico e se fecha automaticamente quando ocorre a desconexão entre o poço e a unidade de produção (SCHLUMBERGER, 2019; STANDARDS NORWAY, 2013 apud VARGAS *et al.*, 2019).

2.2.1.3 PERDA RÁPIDA DE PRODUTIVIDADE

A produtividade de um poço está diretamente relacionada a diversas características do reservatório, como a pressão estática, o percentual de sedimentos e água, o índice de produtividade, a razão gás/óleo, a viscosidade do fluido produzido, entre outras (HAUSLER; KRISHNAMURTHY; SHERAR, 2015 apud VARGAS *et al.*, 2019).

Conforme Vargas *et al.* (2019), à medida que o reservatório é explorado e sua capacidade diminui, essas propriedades sofrem alterações ao longo do tempo. Quando essas mudanças reduzem a energia disponível no sistema a ponto de não ser mais suficiente para superar as perdas, o fluxo de fluido deixa de alcançar a superfície, interrompendo a produção. Esse fenômeno é conhecido na indústria de exploração e produção (E&P) como "perda rápida de produtividade" e, em casos extremos, resulta na perda de surgência.

2.2.1.4 RESTRIÇÃO RÁPIDA EM CKP

De acordo com Vargas *et al.* (2019), a válvula CKP é instalada no início da unidade de produção, desempenhando um papel crucial no controle do poço na superfície. Embora a expressão "restrição rápida em CKP" não seja amplamente definida na literatura, ela é frequentemente utilizada na Petrobras. Esta restrição se refere a uma amplitude que excede um valor de referência durante um período curto.

Antecipar esse tipo de anomalia é desejável, pois, geralmente, essa válvula é manual e fechamentos não intencionais podem ser revertidos mais rapidamente, reduzindo, assim, as perdas de produção.

2.2.1.5 PRESENÇA DE INCRUSTAÇÃO EM CKP

O monitoramento da válvula CKP é essencial devido à sua susceptibilidade a depósitos inorgânicos, os quais podem reduzir drasticamente a produção de petróleo (SCHLUMBERGER, 2019 apud VARGAS *et al.*, 2019).

A identificação automática desta condição em tempo hábil é altamente desejável, pois permite que ações apropriadas, como a injeção de inibidor de incrustação, sejam adotadas para evitar perdas na produção.

2.2.1.6 HIDRATOS NA LINHA DE PRODUÇÃO E SERVIÇO

Os hidratos representam um dos principais desafios na indústria do petróleo. Eles são compostos cristalinos formados por água e gás natural, assemelhando-se ao gelo. A formação de hidratos exige a presença de água e gás natural, além de altas pressões e baixas temperaturas. Por isso, oleodutos que transportam óleo morto, sem gás natural, não sofrem com esse tipo de anomalia. No entanto, sua ocorrência é mais comum em gasodutos e poços produtores de gás. Hidratos também podem se formar em poços produtores de óleo, podendo até interromper completamente o fluxo de produção (ANDREOLLI, 2016; apud VARGAS *et al.*, 2019) (ELLISON; GALLAGHER; LORIMER, 2000; apud VARGAS *et al.*, 2019).

2.3 APRENDIZADO DE MÁQUINA

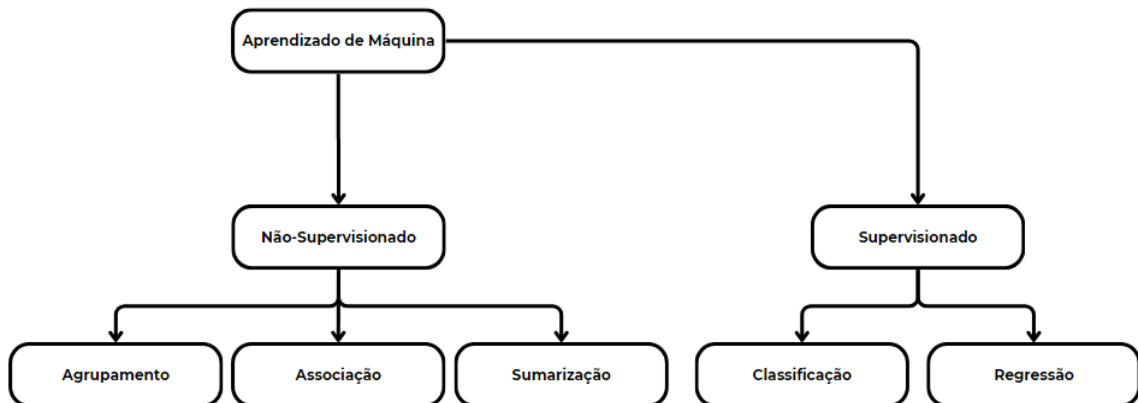
Como já contextualizado anteriormente, o aprendizado de máquina é uma subárea da inteligência artificial que busca construir sistemas capazes de melhorar automaticamente com base em dados e experiências passadas. Esses sistemas podem ser treinados para resolver uma vasta gama de problemas, como classificação de dados, reconhecimento de padrões, e detecção de anomalias. De acordo com Jordan e Mitchell (2015), o aprendizado de máquina se destaca não apenas pela sua aplicação prática em áreas como visão computacional e reconhecimento de fala, mas também por suas implicações teóricas, que envolvem entender as leis estatísticas, computacionais e teóricas da informação que regem todos os sistemas de aprendizado, desde computadores até seres humanos e organizações.

A crescente popularidade e sucesso de técnicas de aprendizado de máquina podem ser atribuídos à sua capacidade de extrair insights de grandes volumes de dados, o que tem sido essencial em diversas indústrias, como finanças, saúde e logística. Esse campo continua a evoluir rapidamente, impulsionado por avanços em algoritmos e a maior disponibilidade de dados e poder computacional.

2.3.1 APRENDIZADO SUPERVISIONADO E PROBLEMAS DE CLASSIFICAÇÃO

O aprendizado supervisionado é uma das abordagens utilizadas no campo de aprendizado de máquina. Nesse contexto, o aprendizado supervisionado é caracterizado por um processo em que um modelo é treinado com base em um conjunto de dados rotulados, ou seja, dados que já possuem a resposta ou o rótulo associado a cada instância de entrada.

Figura 5 - Fluxograma de tipos de aprendizado de máquina



Fonte: Autoria própria.

Nessa abordagem, conforme descrito na imagem 5, os problemas podem ser classificados principalmente em dois tipos: classificação e regressão. Na classificação, o objetivo é prever uma categoria ou classe de uma instância, como no caso de distinguir entre imagens de cães e gatos (Joulin *et al.*, 2017). Já na regressão, o objetivo é prever um valor contínuo, como a previsão do preço de uma casa com base em características como localização, número de quartos, entre outros (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

O principal objetivo de um problema de classificação é, portanto, categorizar uma instância em uma de várias classes possíveis. Dependendo da quantidade de classes, os problemas de classificação podem ser classificados em dois tipos: classificação binária e classificação multiclasse. A classificação binária envolve dois rótulos ou classes, como por exemplo, classificar e-mails como "spam" ou "não spam", enquanto a classificação multiclasse envolve mais de duas classes, como no caso da classificação de espécies de plantas em diferentes categorias baseadas em suas características.

Além disso, existem problemas de classificação mais complexos, como a classificação multirrótulo, em que uma instância pode pertencer a várias classes simultaneamente. Esse tipo de problema é comum em sistemas de recomendação, por exemplo, quando uma música pode ser associada a vários gêneros musicais, ou em sistemas de classificação de textos, onde um artigo pode tratar de múltiplos tópicos.

Alguns algoritmos populares para problemas de classificação incluem as Máquinas de Vetores de Suporte (SVM), Redes Neurais, Árvores de Decisão e K-Nearest Neighbors (K-NN), cada um com suas características e áreas de aplicação específicas. Neste trabalho serão abordados dois algoritmos de classificação, sendo eles a Regressão logística e o Light Gradient Boosting Machine (LGBM), ambos amplamente utilizados em tarefas de classificação devido às suas características particulares e eficácia em diferentes tipos de dados e problemas.

2.3.1.1 REGRESSÃO LOGÍSTICA

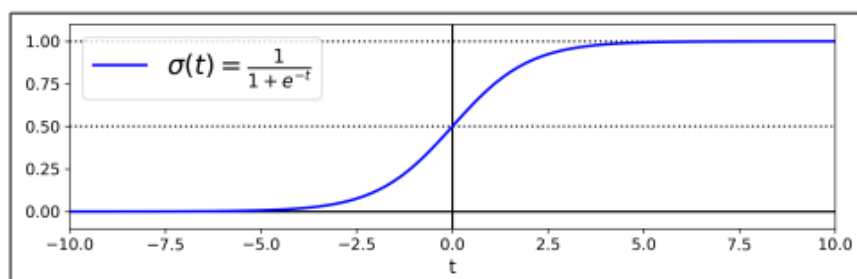
A Regressão Logística é uma técnica amplamente utilizada para resolver problemas de classificação, onde a variável dependente é categórica. Essa abordagem é particularmente útil quando o objetivo é prever a probabilidade de um evento ou resultado binário, como a probabilidade de um paciente sobreviver a uma doença ou se um cliente vai ou não cancelar um serviço. A Regressão Logística é frequentemente escolhida por sua simplicidade, capacidade de fornecer probabilidades de classificação e facilidade de interpretação dos resultados (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

Na Regressão Logística, o modelo é construído com base em uma função logística, também conhecida como função sigmoide, que transforma a saída do modelo em uma probabilidade, restrita ao intervalo de 0 a 1. A equação do modelo é dada por:

$$P(y = 1 | X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}}$$

Onde $P(y = 1 | X)$ é a probabilidade de uma instância pertencer à classe 1, x_1, x_2, \dots, x_n são as características de entrada (variáveis independentes), e $\beta_0, \beta_1, \dots, \beta_n$ são os coeficientes do modelo. O objetivo é estimar esses coeficientes de forma que a função logística forneça a melhor estimativa da probabilidade da instância pertencer à classe 1.

Figura 6 - Comportamento da função logística em um plano 2D



Fonte: Gerón (2019)

A principal vantagem da Regressão Logística sobre outros modelos de classificação é que ela fornece uma probabilidade associada a cada previsão, permitindo decisões mais informadas, como por exemplo, escolher um limiar de probabilidade para decidir a classe de uma instância. Em muitos casos, um valor de limiar de 0,5 é utilizado, classificando qualquer instância com probabilidade maior que 0,5 como pertencente à classe 1 e as demais como pertencentes à classe 0.

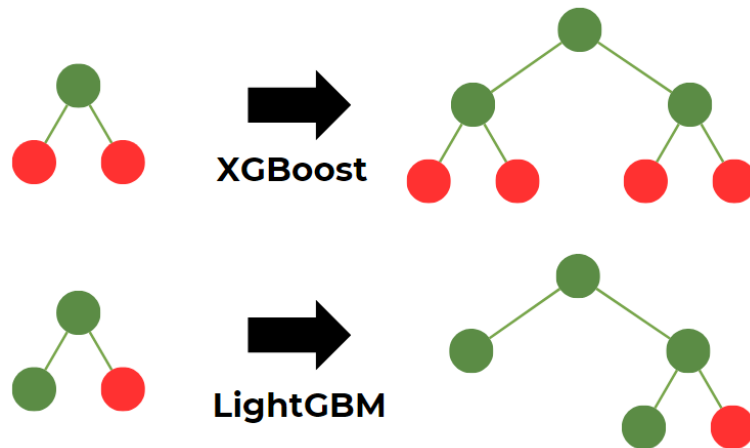
No entanto, a Regressão Logística não é isenta de limitações. Em situações onde as classes são separáveis de forma perfeita no espaço das variáveis independentes, a Regressão Logística pode falhar em fornecer uma solução única e ótima, uma vez que existem infinitas combinações de coeficientes que podem prever corretamente os dados de treinamento. Esse fenômeno ocorre porque a função de máxima verossimilhança se aproxima de zero quando as classes são perfeitamente separáveis, tornando a estimativa dos parâmetros difícil (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

2.3.1.2 *LIGHT GRADIENT BOOSTING MACHINE* (LGBM)

O *Light Gradient Boosting Machine* (LGBM) é uma técnica de aprendizado de máquina que se destaca no campo dos algoritmos baseados em *boosting*, sendo amplamente utilizada para problemas de classificação e regressão devido à sua eficiência e alto desempenho, especialmente em grandes conjuntos de dados. Desenvolvida pela *Microsoft Research*, a LGBM é uma versão otimizada do algoritmo *Gradient Boosting Decision Tree* (GBDT), com o objetivo de melhorar a velocidade de treinamento e reduzir o consumo de memória, mantendo a precisão dos modelos gerados.

O algoritmo combina vários modelos fracos (geralmente árvores de decisão) para criar um modelo forte. O processo de treinamento ocorre iterativamente, em que cada nova árvore tenta corrigir os erros residuais das árvores anteriores. Em vez de usar um método de busca exaustiva para encontrar os melhores pontos de divisão, como no GBDT tradicional, o LGBM introduz melhorias significativas na eficiência computacional por meio de estratégias de seleção e otimização. Essas melhorias incluem o uso de histogramas para armazenar e calcular os possíveis valores de divisão e a ordenação de valores em *buckets* discretos, o que reduz o custo de memória e processamento. Além disso, o LGBM suporta o crescimento de árvores de forma *leaf-wise* (orientada a folhas), em vez de *depth-wise* (orientada a profundidade), mostrado pela Figura 7, permitindo que o modelo foque em regiões de maior erro e atinja uma maior redução do *loss* a cada iteração.

Figura 7 - Árvore de decisão construída com *LightGBM* e *XGBoost*



Fonte: Autoria própria.

O algoritmo tradicional de GBDT enfrenta dificuldades de desempenho devido à necessidade de examinar todas as instâncias de dados para calcular o ganho de informação para cada possível ponto de divisão. Esse processo se torna muito custoso em termos de tempo computacional quando se trabalha com grandes conjuntos de dados. Para resolver essa limitação, o LGBM introduz duas técnicas inovadoras: o *Gradient-based One-Side Sampling* (GOSS) e o *Exclusive Feature Bundling* (EFB) (KE *et al.*, 2017).

A técnica GOSS busca otimizar a seleção das instâncias de dados a serem utilizadas no treinamento. O GOSS prioriza a retenção das instâncias com gradientes maiores, que são as mais influentes para o cálculo do ganho de informação. Isso permite uma amostragem mais eficiente, mantendo a precisão na estimativa do ganho, especialmente quando a variação do ganho de informação é significativa. Já a técnica EFB é aplicada para reduzir o número de características, aproveitando a natureza esparsa dos dados. Ao identificar características exclusivas, ou seja, aquelas que raramente assumem valores não nulos simultaneamente, o EFB permite agrupar essas características de maneira eficaz, diminuindo a dimensionalidade sem perda significativa de informação.

Essas inovações fazem do LGBM um dos algoritmos mais rápidos e eficientes para grandes volumes de dados. Os experimentos realizados por Ke *et al.* (2017) demonstraram que o LGBM pode acelerar o processo de treinamento em até 20 vezes, mantendo uma precisão comparável a dos métodos tradicionais de GBDT. A implementação dessas técnicas não só melhora o desempenho computacional, mas também torna o LGBM uma ferramenta extremamente eficaz para tarefas como classificação em grandes bases de dados.

2.3.2 ENGENHARIA DE FEATURES

A engenharia de features é uma etapa importante no processo de modelagem de dados, especialmente em projetos que envolvem aprendizado de máquina. Essa etapa consiste em transformar os dados brutos em um formato que possa ser melhor compreendido e explorado pelos algoritmos. A qualidade das features extraídas ou construídas pode impactar significativamente o desempenho do modelo, tornando a engenharia de features um dos pilares fundamentais de uma análise de dados bem-sucedida.

Existem diversas técnicas que podem ser aplicadas nos dados, dentre elas temos o uso de *differencing*, transformação logarítmica, média e desvio padrão.

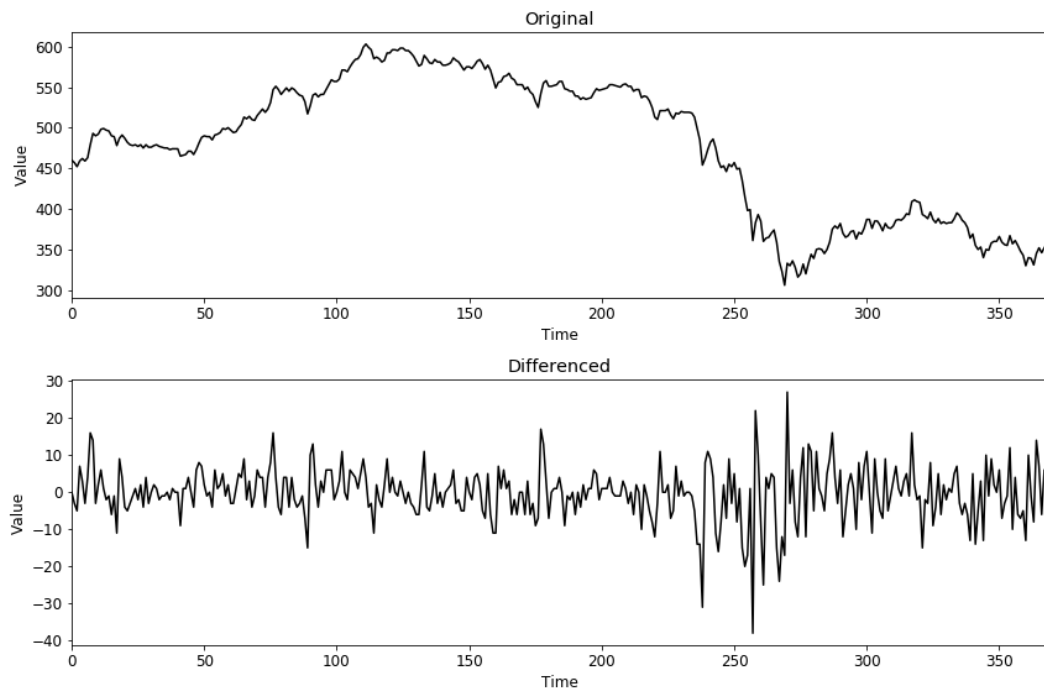
2.3.2.1 CRIAÇÃO DE FEATURES COM *DIFFERENCING*

O *differencing* é uma técnica comum em séries temporais para remover tendências ou padrões sazonais. Consiste em calcular as diferenças entre valores consecutivos da série para reduzir a autocorrelação e tornar a série mais estacionária, como é mostrado na imagem 8. Essa abordagem é útil, especialmente quando se deseja explorar os gradientes de mudança no modelo.

$$Y'_t = Y_t - Y_{t-1}$$

Sendo Y_t o valor da característica no tempo t , Y_{t-1} o valor da característica no tempo $t-1$ e Y'_t o valor do primeiro *differencing*, que representa a mudança entre os dois pontos. Essa técnica tem diversas aplicações práticas. Por exemplo, em séries financeiras, pode ser usada para calcular o retorno diário de ações ou ativos, eliminando a tendência de longo prazo dos preços. Já em sensores industriais, o *differencing* pode identificar mudanças rápidas em variáveis operacionais, auxiliando na detecção precoce de anomalias.

Figura 8 - Impacto de *differencing* nos dados.

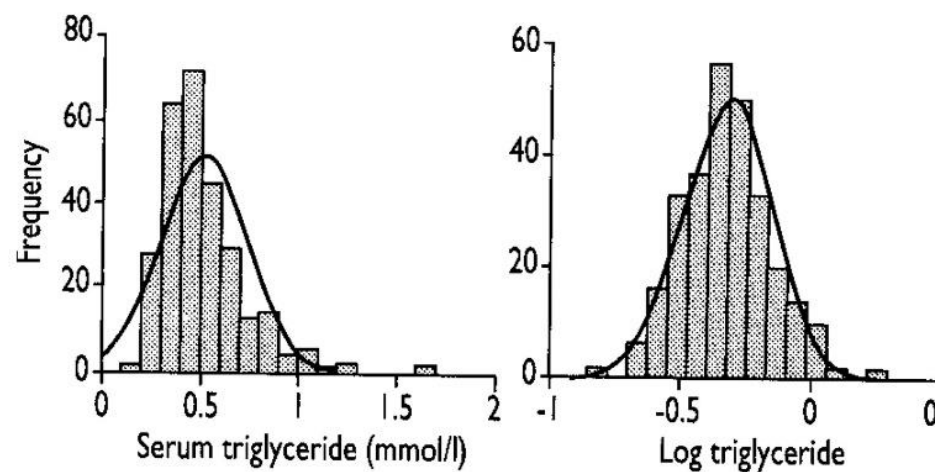


Fonte: Monigatti (2023)

2.3.2.2 CRIAÇÃO DE FEATURES COM TRANSFORMAÇÃO LOGARÍTMICA

A transformação logarítmica é uma técnica utilizada para reduzir a variabilidade nos dados, especialmente quando há uma grande diferença entre os valores de uma variável. Ao aplicar o logaritmo, é possível estabilizar a variância e melhorar a linearidade dos dados, o que é benéfico para a construção de modelos preditivos. (ALTMAN, 2000)

Figura 9 - Impacto da transformação logarítmica nos dados



Fonte: Bland *et al.* (1996).

Como mostrado na imagem 9, a transformação logarítmica é uma técnica capaz de comprimir a variabilidade dos dados, o que pode melhorar sua normalidade. Essa compressão reduz a influência de valores extremos, aproximando a distribuição dos dados de uma distribuição normal. Isso é particularmente útil em análises que assumem a normalidade dos dados, como muitos testes estatísticos e modelos preditivos. No entanto, essa técnica apresenta uma limitação quanto a valores negativos ou nulos, o que exige um tratamento prévio dos dados, como adicionar uma constante positiva para garantir que todos os valores sejam maiores que zero.

2.3.2.3 CRIAÇÃO DE FEATURES ESTATÍSTICAS

A análise estatística básica, como o cálculo de soma, média e desvio padrão, é essencial para compreender a distribuição e a variabilidade dos dados em um conjunto de observações. Essas estatísticas fornecem informações valiosas sobre o comportamento dos dados e podem ser utilizadas tanto para análise exploratória quanto como recursos em modelos de aprendizado de máquina.

A soma é a base para o cálculo de diversas estatísticas e representa o total acumulado de uma variável ao longo de todas as observações. Ao trabalhar com séries temporais, a soma permite identificar tendências ou padrões de comportamento global. Já a média μ é uma medida de tendência central que descreve o valor médio de uma variável. Ela é calculada dividindo a soma de todos os valores pelo número total de observações.

$$\mu_X = \frac{\sum_{i=1}^n X_i}{n}$$

O desvio padrão σ quantifica a dispersão dos dados em relação à média, fornecendo uma medida da consistência dos valores. Ele é calculado como a raiz quadrada da variância, que é a média dos quadrados das diferenças entre cada valor e a média.

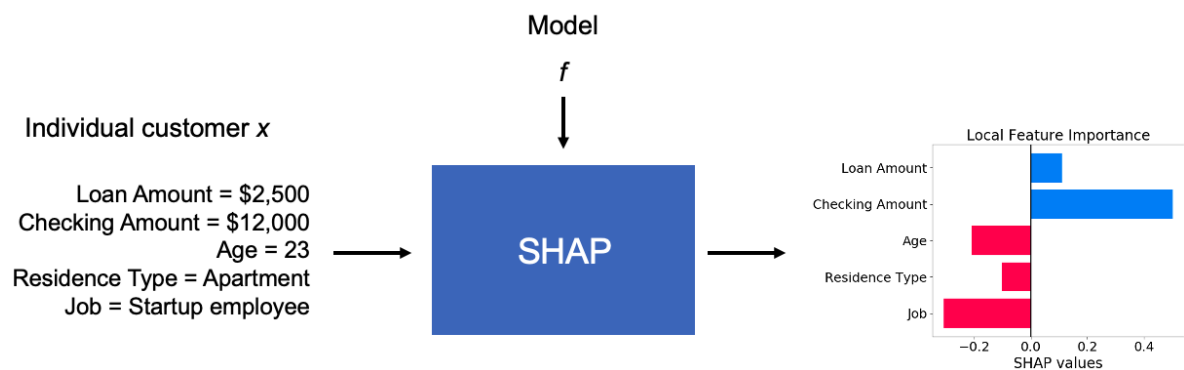
$$\sigma_X = \sqrt{\frac{\sum_{i=1}^n (X_i - \mu_X)^2}{n}}$$

Neste trabalho, a média e o desvio padrão são aplicados na soma das pressões e temperaturas. A primeira ajuda a identificar os valores típicos para essas variáveis, o que pode ser útil para definir condições operacionais normais, e o último ajuda a apontar a distância dessa média.

2.3.2.4 SHAPLEY ADDITIVE EXPLANATIONS (SHAP)

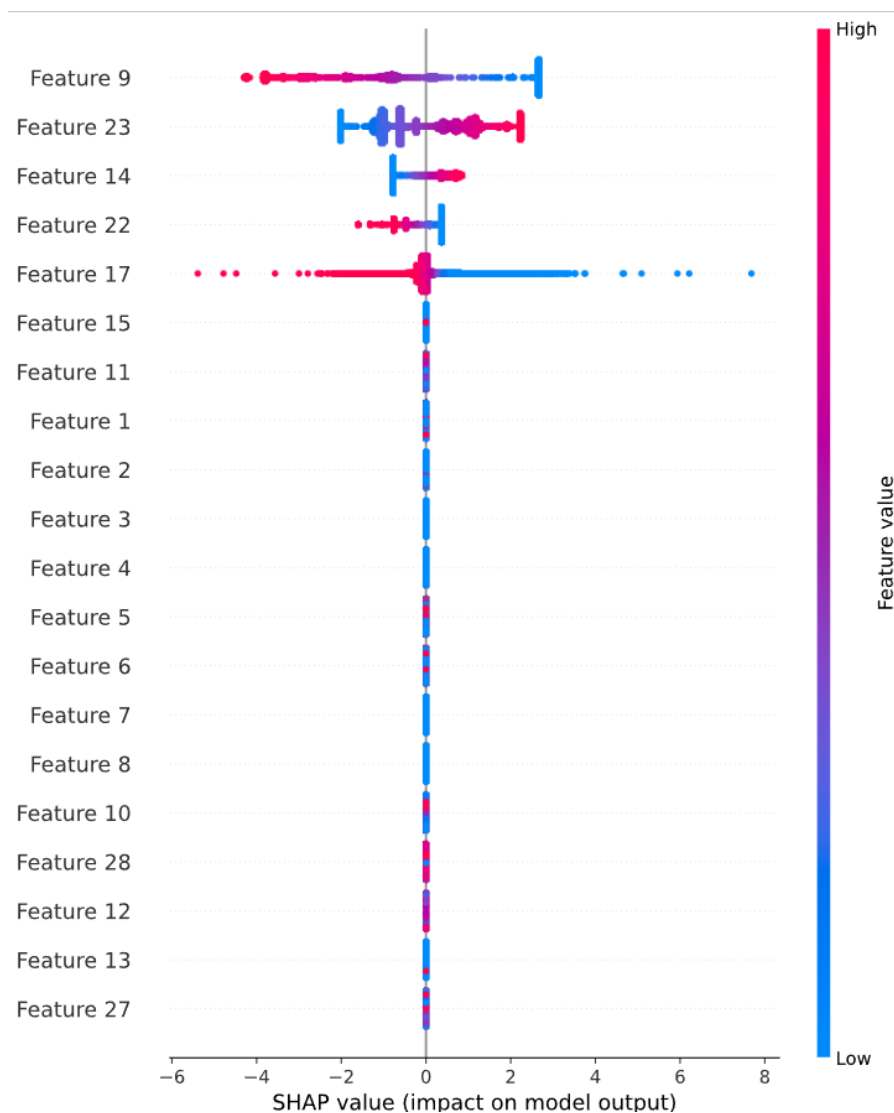
O Shapley Additive Explanations é uma técnica utilizada para explicar as previsões feitas por modelos de aprendizado de máquina, especialmente em cenários onde os modelos são complexos, como árvores de decisão, redes neurais e métodos de ensemble. Baseado na teoria dos valores de Shapley, originada na teoria dos jogos, o SHAP busca quantificar a contribuição individual de cada variável (feature) para a previsão realizada pelo modelo, oferecendo uma visão detalhada de como as entradas afetam os resultados (MOLNAR, 2020).

Figura 10 - Fluxograma de funcionamento do SHAP



Fonte: Covert (2020)

O objetivo principal do SHAP é atribuir uma "pontuação" a cada variável em cada instância analisada. Essas pontuações indicam se a variável contribuiu para aumentar ou diminuir a previsão, bem como a magnitude dessa influência. Por exemplo, em um modelo que prevê falhas em um sistema industrial, o SHAP pode identificar o quanto a pressão, temperatura ou outras variáveis influenciam o risco de falha para uma observação específica.

Figura 11 - Retorno do desempenho global das características para o modelo

Fonte: Autoria própria.

Uma das grandes vantagens do SHAP é que ele oferece explicações consistentes e matematicamente sólidas. Ele garante que a soma das contribuições das variáveis seja igual à diferença entre a previsão do modelo para uma instância específica e a média global das previsões. Essa propriedade aditiva torna as explicações intuitivas e fáceis de interpretar, mesmo para usuários não especializados (MOLNAR, 2020).

De acordo com Molnar (2020), o SHAP tem um papel essencial na promoção da transparência dos modelos preditivos, especialmente contextos em que decisões baseadas no modelo podem ter impactos significativos, como em setores regulados ou críticos, incluindo saúde, finanças e engenharia. Ele permite que os tomadores de decisão entendam quais fatores estão direcionando os resultados e em que medida.

O SHAP é amplamente reconhecido como uma das ferramentas mais robustas e interpretáveis disponíveis para análise de impacto das variáveis em modelos preditivos. Ele é

essencial não apenas para validar e explicar modelos, mas também para fornecer insights acionáveis que podem ser usados para otimizar processos e tomar decisões informadas.

2.3.3 VALIDAÇÃO CRUZADA

De acordo com Kohavi (1995), a validação cruzada é uma técnica estatística utilizada para avaliar o desempenho de modelos preditivos, especialmente em aprendizado de máquina. Seu objetivo principal é medir a capacidade de generalização do modelo, ou seja, sua habilidade de realizar previsões precisas em dados não utilizados durante o treinamento. A validação cruzada é útil para evitar problemas de sobreajuste (*overfitting*), nos quais o modelo se ajusta muito bem ao conjunto de treinamento, mas apresenta baixo desempenho em novos dados.

Figura 12 - Validação cruzada tradicional



Fonte: Leite (2020)

Na versão tradicional da validação cruzada, como mostrado na Figura 12, a *K-Fold*, o conjunto de dados é dividido em K subconjuntos aproximadamente iguais, chamados de *folds*. O modelo é treinado K vezes, utilizando K-1 desses subconjuntos para treinamento e reservando o subconjunto restante para validação. Em cada iteração, um *fold* diferente é usado como conjunto de validação, e os resultados são combinados para obter métricas de desempenho mais representativas. Esse método reduz a dependência dos dados escolhidos para validação, garantindo uma avaliação mais confiável do modelo (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

2.3.3.1 VALIDAÇÃO CRUZADA ESTRATIFICADA

O *K-Fold Stratified* é uma variação do método de validação cruzada *K-Fold*, que é utilizada particularmente quando se trabalha com conjuntos de dados desbalanceados. Esse método busca garantir que cada um dos K subconjuntos gerados na validação cruzada preserve a proporção original das classes presentes no conjunto de dados. Assim, ele contribui para que o modelo avaliado receba uma representação mais consistente dos dados durante o treinamento e a validação, minimizando o viés introduzido por distribuições não uniformes (KOHAVI, 1995).

Figura 13 - Validação cruzada estratificada



Fonte: Duan (2023).

O processo de *K-Fold Stratified* funciona, semelhante ao *K-fold*, dividindo o conjunto de dados em K subconjuntos (ou *folds*). Em cada iteração, um subconjunto é reservado para validação, enquanto os outros $K-1$ são usados para treinamento. O diferencial está no fato de que a divisão é feita de maneira estratificada, ou seja, a distribuição proporcional das classes é preservada em cada subconjunto. Essa característica é essencial em tarefas de classificação nas quais há desequilíbrio nas classes (BISHOP, 2006).

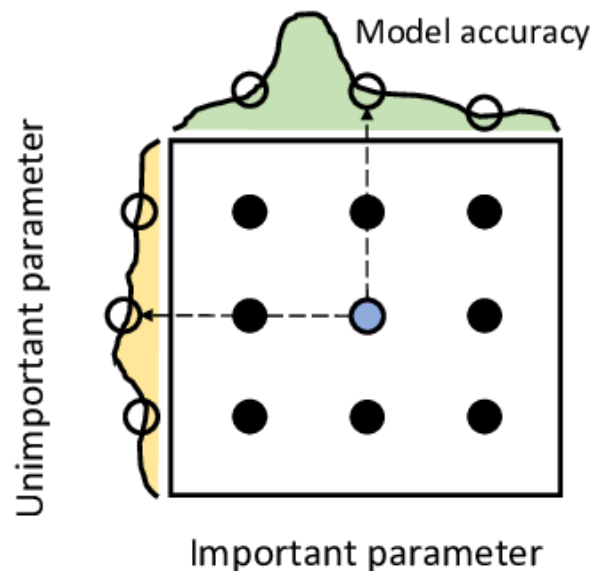
2.3.4 OTIMIZAÇÃO DE HIPERPARÂMETROS

A otimização de hiperparâmetros é um passo essencial no processo de desenvolvimento de modelos de aprendizado de máquina, pois visa encontrar as melhores combinações de parâmetros que maximizem o desempenho do modelo em um dado conjunto de validação. Os hiperparâmetros, que diferem dos parâmetros ajustados durante o treinamento, são definidos antes do processo de aprendizado e podem influenciar significativamente a capacidade preditiva do modelo (ERDEN *et al.*, 2023). Entre as diversas abordagens para a otimização de

hiperparâmetros, destacam-se os métodos de busca exaustiva, como o *Grid Search*, e os métodos probabilísticos, como a otimização Bayesiana.

O *Grid Search* é uma abordagem sistemática que explora todas as combinações possíveis de um conjunto pré-definido de valores de hiperparâmetros, como é apresentado na Figura 14. Esse método garante que todas as combinações sejam avaliadas, oferecendo a garantia de encontrar a configuração ótima dentro do espaço de busca especificado. No entanto, sua eficiência diminui em cenários com muitos hiperparâmetros ou valores possíveis, devido à alta demanda computacional.

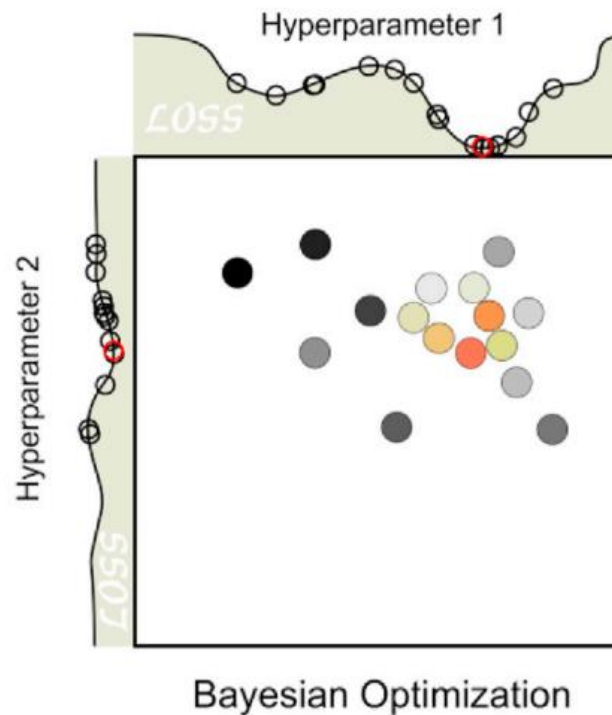
Figura 14 – Esquemática de funcionamento do Grid Search



Fonte: adaptado de Bergstra *et al.*, (2019).

Por outro lado, a otimização Bayesiana, representada esquematicamente na Figura 15, é uma técnica que visa superar as limitações do Grid Search, especialmente em termos de custo computacional. Baseando-se em princípios probabilísticos, esse método utiliza modelos probabilísticos, como o processo gaussiano, para criar uma função de aquisição que guia a busca pelos melhores hiperparâmetros (SHAHRIARI *et al.*, 2016). A principal ideia da otimização bayesiana é balancear a exploração de regiões ainda não investigadas do espaço de busca com a exploração local de áreas promissoras, levando a um processo mais eficiente de ajuste dos hiperparâmetros.

Figura 15 - Esquemática de funcionamento do *Bayesian Search*



Fonte: adaptado de Passos *et al*, (2022).

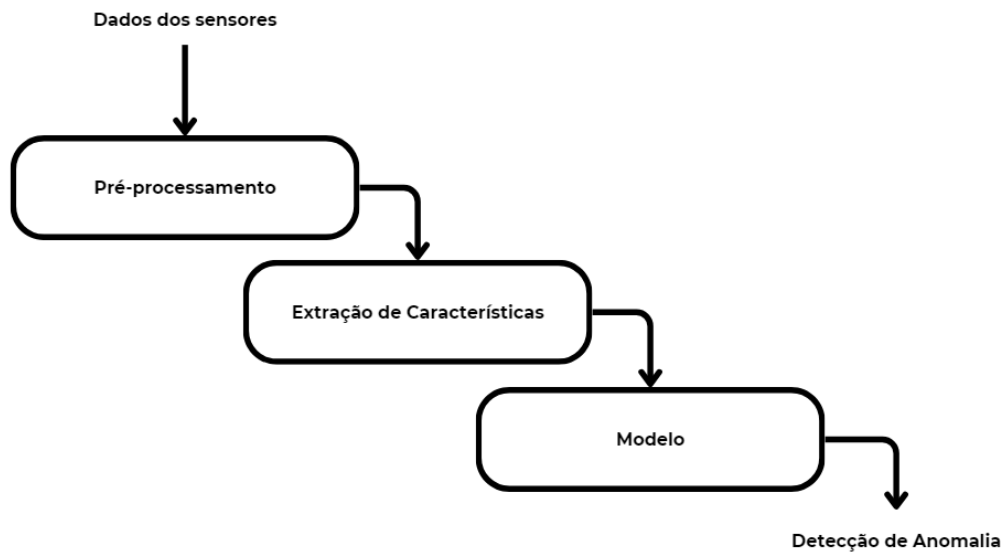
3 METODOLOGIA

Este capítulo aborda a metodologia e apresenta a estrutura do sistema de monitoramento para detectar e classificar anomalias. Primeiro, detalhamos como o sistema é treinado e em seguida apresentamos como os dados são preparados em três etapas distintas antes de serem alimentados no algoritmo de classificação escolhido, a fim de apontar a anomalia, seguindo para um modelo de votação que classificará o evento entre os sete disponíveis.

Este trabalho foi inteiramente realizado em um ambiente Python, utilizando pacotes disponibilizados pela Petrobras para garantir a padronização, além de alguns pacotes externos mencionados na seção 4.4, que trata da reprodutibilidade dos resultados.

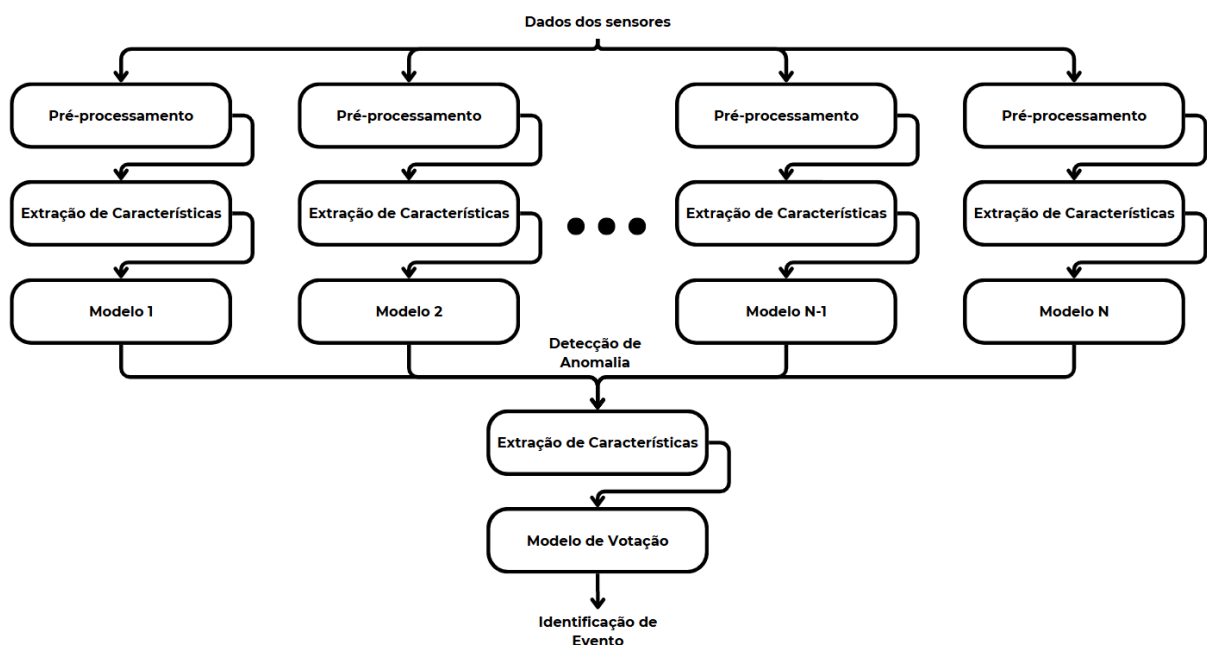
3.1 ESTRUTURA DO SISTEMA DE MONITORAMENTO

Esta subseção apresenta uma introdução ao *pipeline* proposto neste trabalho. O objetivo da primeira estrutura, referente a Figura 16, é atuar como um sistema de monitoramento de condições. Ou seja, o sistema deve ser capaz de identificar e diferenciar eventos indesejáveis (anomalias) de condições normais, utilizando sinais brutos de sensores como base.

Figura 16 - Estrutura do sistema de reconhecimento de anomalias

Fonte: Autoria Própria

Já o objetivo da segunda estrutura, mostrado na Figura 17, é atuar como um classificador de eventos, integrando a primeira estrutura de identificação de anomalias apresentada na figura anterior.

Figura 17 - Sistema de Reconhecimento e Classificação de Anomalias

Fonte: Autoria Própria

3.2 PRÉ-PROCESSAMENTO

O primeiro bloco do sistema corresponde à etapa de pré-processamento. Essa etapa é composta por duas ações principais: inicialmente, os dados são divididos em conjuntos de

treinamento e teste, permitindo que os algoritmos sejam treinados e avaliados em dados não vistos previamente. Em seguida, os dados brutos são tratados para garantir sua adequação às etapas posteriores, nas quais as features estatísticas serão extraídas. Na Tabela 4 é apresentado o número de instâncias para cada evento em cada conjunto de dados, com a divisão realizada de forma aleatória na proporção de 70/30. Adicionalmente, foram descartadas as instâncias que não possuem observações transientes. Por esse motivo, os eventos 3 e 4 foram excluídos, devido à sua falta observações em estado transientes. Além disso as Instâncias desenhadas a mão não são abordadas nesse trabalho, portanto são descartadas.

Tabela 4 - Separação dos arquivos de treino e teste

Evento	Instâncias de Treino	Instâncias de Teste	Instâncias Totais
0 - Normal	416	178	594
1 - Aumento Repentino de Sedimentos Básicos e Água	83	35	118
2 - Fechamento Irregular de Válvula de Segurança	27	11	38
5 - Perda Rápida de Produtividade	74	32	106
6 - Restrição Rápida em CKP	155	66	221
7 - Presença de Incrustação em CKP	25	11	36
8 - Hidratos na Linha de Produção	67	28	95
9 - Hidratos na Linha de Serviço	145	62	207

Fonte: Autoria própria.

Após a separação entre as instâncias de treino e teste, realiza-se o tratamento dos *outliers* físicos, uma vez que esses não correspondem a anomalias genuínas no sistema, mas sim a falhas de leitura decorrentes de interferências externas ou ruídos. Esses erros de leitura são bastante comuns em sistemas industriais e, embora não representem falhas operacionais reais, podem distorcer a análise dos dados. Além disso, a presença de outliers físicos afeta diretamente o desempenho do escalador utilizado na normalização, potencialmente prejudicando a qualidade do treinamento e da avaliação dos modelos de machine learning. Por esse motivo, essa etapa de pré-processamento é essencial para garantir uma representação consistente e confiável dos dados. São considerados outliers físicos nesse trabalho as temperaturas abaixo de $-273\text{ }^{\circ}\text{C}$, tal qual pressões e vazões negativas, visto que esses valores não só representam uma pequena

parcela de dados, mas também não apresentam sentido fisicamente, logo serão considerados ruídos.

Tabela 5 - Análise de dados a serem tratados

Evento	Contagem de Ruídos	Contagem de NaN	Contagem de Não-Rotulados
0 - Normal	2.365.222	203.091.772	2.138.400
1 - Aumento Repentino de Sedimentos Básicos e Água	176.916	185.520.026	14.400
2 - Fechamento Irregular de Válvula de Segurança	0	15.796.915	79.200
5 - Perda Rápida de Produtividade	0	261.187.181	39.600
6 - Restrição Rápida em CKP	0	123.481.257	21.600
7 - Presença de Incrustação em CKP	651.311	113.972.498	129.600
8 - Hidratos na Linha de Produção	1.760.248	96.072.564	50.400
9 - Hidratos na Linha de Serviço	274.148	172.509.344	205.200

Fonte: Autoria própria

Após a identificação dos dados que necessitavam de tratamento, foi elaborada uma estratégia cuidadosa e justificada para a limpeza dessas informações. O processo foi realizado instância por instância, a fim de evitar o vazamento e enviesamento de dados, adotando abordagens específicas para cada tipo de dado.

O pré-processamento se inicia descartando colunas que não possuem nenhum registro nas Instâncias de anomalia, as mesmas colunas descartadas para essa anomalia vão ser descartadas para a operação normal na construção do modelo referente a anomalia.

Em relação aos dados binários, os valores ausentes (NaN) foram substituídos por zero. Essa decisão foi fundamentada no fato de que, em variáveis binárias, os valores só podem assumir os valores 0 ou 1. A substituição por zero é uma abordagem mais conservadora, evitando a introdução de ambiguidades que poderiam afetar negativamente os resultados do modelo.

Para os dados não binários ruidosos optou-se por substituir pela média. A média é uma medida robusta que reflete o centro da distribuição dos dados, ajudando a suavizar os efeitos dos ruídos sem comprometer as informações essenciais.

Já para os dados não binários com valores ausentes, a mediana foi utilizada para preencher as lacunas. A escolha da mediana, em vez da média, foi feita para evitar a influência de outliers que poderiam distorcer a representação dos dados. A mediana, no geral, é menos sensível a valores extremos, assim assegurando uma melhor preservação da integridade da distribuição dos dados.

Por fim, os registros que apresentavam rótulos não identificados foram descartados. Rótulos ausentes ou errôneos podem introduzir ruídos significativos no processo de aprendizado supervisionado, prejudicando a performance do modelo. O descarte dessas instâncias foi uma medida necessária para garantir que apenas dados confiáveis fossem utilizados nas etapas subsequentes, além disso a quantidade de rótulos ausentes é ínfima comparado ao tamanho do banco de dados.

Essas ações de limpeza garantiram que os dados estivessem prontos para as próximas etapas.

3.2.1 EXTRAÇÃO DE CARACTERÍSTICAS

O segundo bloco da estrutura é dedicado à extração de características, uma etapa fundamental para transformar dados brutos em informações úteis que possam ser exploradas pelos modelos de aprendizado de máquina. Esse processo consiste em identificar e criar variáveis que capturem os padrões mais representativos dos dados, facilitando a detecção de falhas nos poços de petróleo offshore.

Abordadas na seção 2.3.2, diferentes técnicas de engenharia de features foram aplicadas para destacar características relevantes e reduzir a dimensionalidade dos dados. Entre elas, foi utilizado o método de *differencing*, que remove tendências sazonais ou de longo prazo e permite que o foco seja direcionado às variações locais mais significativas. Além disso, transformações logarítmicas foram realizadas para estabilizar a variância e lidar com distribuições assimétricas, tornando-as mais adequadas para os algoritmos. Estatísticas descritivas, como média e desvio padrão também foram extraídas, de modo a capturar informações importantes sobre a dispersão, centralidade e forma das distribuições.

Para complementar, a técnica SHAP (SHapley Additive exPlanations) foi empregada para explicar a importância das características utilizadas, permitindo identificar quais variáveis apresentam maior impacto nas previsões do modelo. Esse tipo de abordagem contribui para a

transparência e interpretabilidade do sistema, garantindo maior confiança nos resultados obtidos.

3.3 ESCOLHA DE MODELO

O último bloco da estrutura consiste na etapa de modelagem e classificação. Antes dessa fase, os dados brutos passaram por um processo de limpeza e as features estatísticas foram extraídas. O objetivo principal dessa etapa é mapear os dados de entrada para categorias específicas, analisando as características extraídas. Para isso, são utilizados algoritmos de classificação que fornecem uma saída probabilística, indicando a probabilidade de os dados de entrada pertencerem a uma determinada classe.

A escolha do algoritmo de classificação envolve a decisão entre abordagens de *machine learning* ou *deep learning*. Ambas possuem vantagens e limitações: enquanto os algoritmos de *deep learning* geralmente requerem grandes volumes de dados para treinamento e apresentam uma complexidade maior na configuração de hiperparâmetros e estrutura de redes, os algoritmos de machine learning se mostram mais adequados em cenários com conjuntos de dados menores e menor disponibilidade de tempo para ajustes.

Considerando o volume de dados disponível e a necessidade de eficiência no treinamento, optou-se por utilizar algoritmos de machine learning, com destaque para o LightGBM (LGBM) e a Regressão Logística (LR). A escolha do modelo para cada evento e seus respectivos hiperparâmetros serão definidos pelo algoritmo de otimização bayesiana, validadas pela validação cruzada estratificada e avaliada através de métricas no capítulo 4.

3.3.1 LIGHT GRADIENT BOOSTING MACHINE (LGBM)

Como já mencionado anteriormente na subseção 2.3.1.2, o LightGBM é um algoritmo de *gradient boosting* baseado em árvores de decisão, reconhecido por sua capacidade de lidar eficientemente com desafios como desbalanceamento de classes, presença de dados faltantes e alta dimensionalidade. Se escolhido, possui diversos hiperparâmetros para otimizar o desempenho do modelo.

Entre os principais hiperparâmetros que serão otimizados estão:

- **Tipo de *boosting*:** Determina a abordagem utilizada para construir o modelo. O LightGBM oferece três opções principais: *gbdt* (*Gradient Boosting Decision Tree*), que é o método padrão e amplamente utilizado; *goss* (*Gradient-based One-Side Sampling*), que foca em amostras com maiores gradientes para acelerar o treinamento sem perda significativa de desempenho; e *dart* (*Dropouts meet Multiple Additive Regression Trees*), que aplica *dropout* às árvores para reduzir o *overfitting*.

- **Número de Estimadores:** Controla o número total de árvores que serão construídas no modelo.
- **Taxa de Aprendizado:** Define o impacto de cada árvore no modelo final.
- **Profundidade Máxima das Árvores:** Limita a profundidade das árvores para controlar a complexidade do modelo.
- **Número de Folhas:** Representa o número máximo de folhas por árvore.
- **Frações de Características:** Controlam a proporção de features utilizadas em cada iteração.
- **Parâmetro de Regularização L1 e L2 (λ_1 e λ_2):** Adicionam penalizações aos coeficientes para evitar *overfitting* e melhorar a generalização do modelo.

3.3.2 REGRESSÃO LOGÍSTICA

Como discutido anteriormente na subseção 2.3.1.1, a Regressão Logística é um modelo de classificação amplamente utilizado, especialmente para problemas binários, devido à sua simplicidade, eficiência computacional e capacidade de fornecer probabilidades interpretáveis para cada classe. Caso seja escolhida, a Regressão Logística também possui alguns hiperparâmetros importantes que podem ser otimizados para melhorar o desempenho do modelo.

Entre os principais hiperparâmetros que serão otimizados estão:

- **Penalização:** Define o tipo de regularização aplicada ao modelo, como L1 (*Lasso*) ou L2 (*Ridge*), para evitar *overfitting* e melhorar a capacidade de generalização.
- **C:** Controla o grau de penalização dos coeficientes.
- **Número Máximo de Iterações:** Define o número máximo de iterações que o algoritmo de otimização pode executar antes de interromper.
- **Solver:** Especifica o método de otimização utilizado para ajustar os coeficientes do modelo. Os algoritmos disponíveis incluem métodos como liblinear, saga, e lbfgs, cada um com características específicas para diferentes tamanhos e naturezas de dados.

3.4 EXTRAÇÃO DE DADOS DA DETECÇÃO DE ANOMALIAS

A etapa de extração da detecção de anomalias foi realizada utilizando um método baseado na soma intermitente de reconhecimento de anomalias. Essa abordagem permitiu identificar e quantificar padrões discrepantes nos dados de forma eficiente, gerando um conjunto robusto de features que serviram de base para o modelo de votação.

Cada instância dos dados foi analisada, tratada, pré-processada e inserida dentro de todos os modelos de eventos, além disso vale ressaltar que o treinamento é baseado desde o começo da instancia até a metade do registro de anomalia, fazendo com que assim seja possível classificar o evento antes de chegar de fato na anomalia. Essa integração permitiu aproveitar as informações obtidas na etapa de detecção de anomalias, melhorando a aplicabilidade da solução proposta nessa tese.

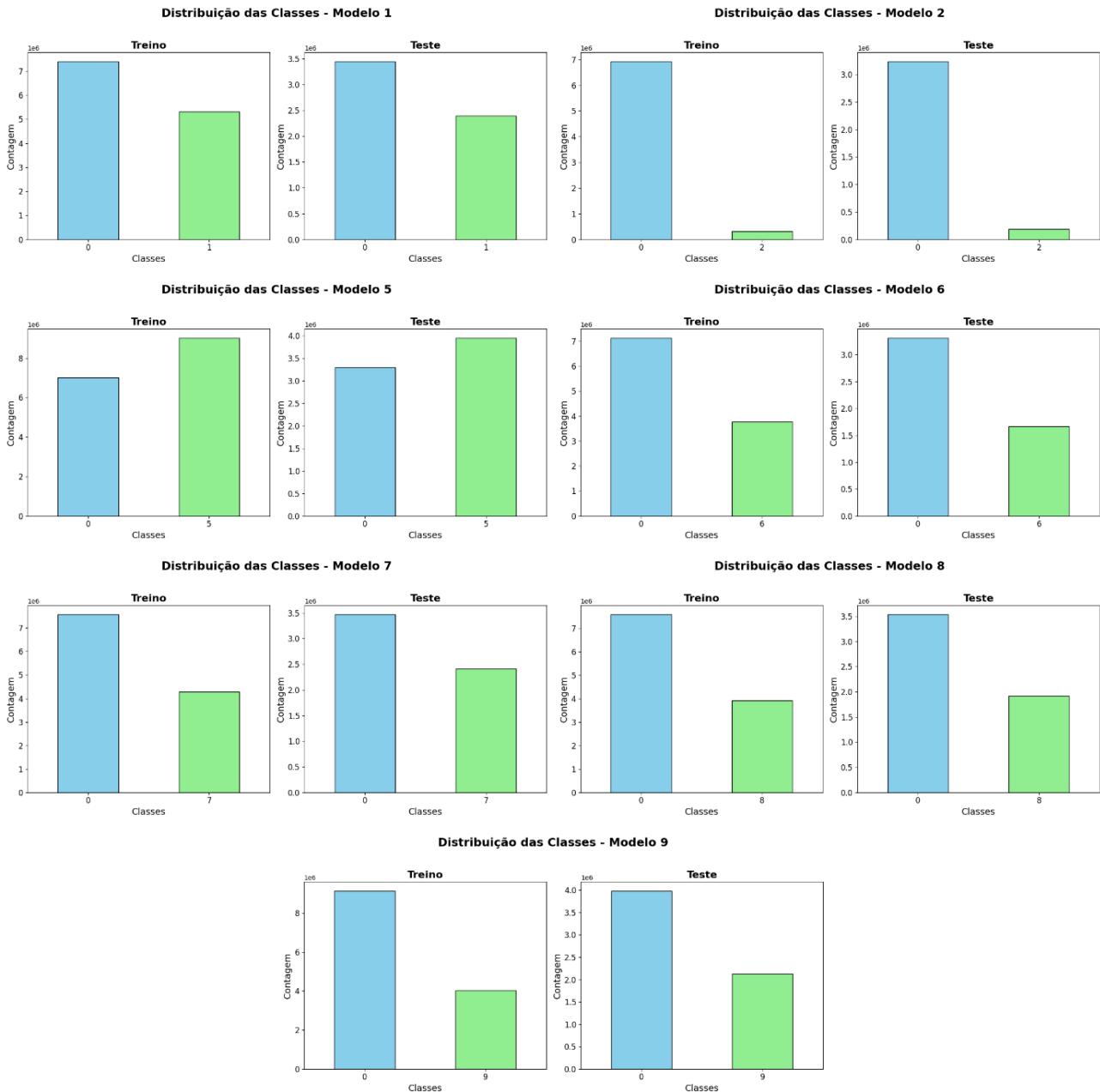
4 RESULTADOS

Este capítulo se dedica a expor os resultados obtidos através da metodologia previamente apresentada, separando-a em identificação de anomalias e classificação de eventos.

4.1 OTIMIZAÇÃO E CONSTRUÇÃO DOS MODELOS DE DETECÇÃO DE ANOMALIAS

Com o objetivo de desenvolver modelos robustos para a detecção de anomalias, as seguintes estratégias e configurações foram adotadas:

- Para cada modelo N, será utilizado dados do evento N e de operação normal 0;
- Escaladores disponíveis: *MinMax* e *Standard*;
- Modelos disponíveis: *LightGBM* e Regressão Logística;
- N° de *folds* dentro da validação cruzada estratificada: 5;
- Tempo de shift para a criação de variáveis *differencing* é de 3 minutos;
- Métrica objetivo: F1;
- Número de iterações dentro de cada cenário na otimização bayesiana: 100.

Figura 18 - Distribuição das anomalias para cada modelo desenvolvido

Fonte: Autoria própria.

Como identificado na análise da distribuição das classes, apresentado na Figura 18, os dados apresentavam um problema significativo de desbalanceamento, com a classe de anomalias substancialmente menor em relação à operação normal. Para corrigir isso, foi utilizada a seguinte equação para o cálculo dos pesos das classes:

$$\text{Peso}(c) = \frac{N^{\circ} \text{ total de amostras}}{N^{\circ} \text{ de classes} * N^{\circ} \text{ de amostras da classe } c}$$

O tempo escolhido para o *differencing*, o qual poderia ter sido ajustado como um hiperparâmetro, foi baseado na Tabela 6. Baseado nisso estipulou-se que um tempo adequado seria de 3 min, o que mais tarde pode ser avaliado através do SHAP.

Tabela 6 - Tempo médio de regime transiente para cada evento

Evento	Tempo médio de regime transiente (minutos)
1	682,44
2	64,33
5	89,75
6	117,5
7	3088,23
8	893,57
9	238,92

Fonte: Autoria própria.

Os resultados expostos na Tabela 7, indicam bons desempenhos tendo em média um F1 superior a 0,94 em todos os modelos. No entanto, esses resultados ainda precisam ser avaliados quanto ao risco de *overfitting*, que pode influenciar a generalização dos modelos para novos dados.

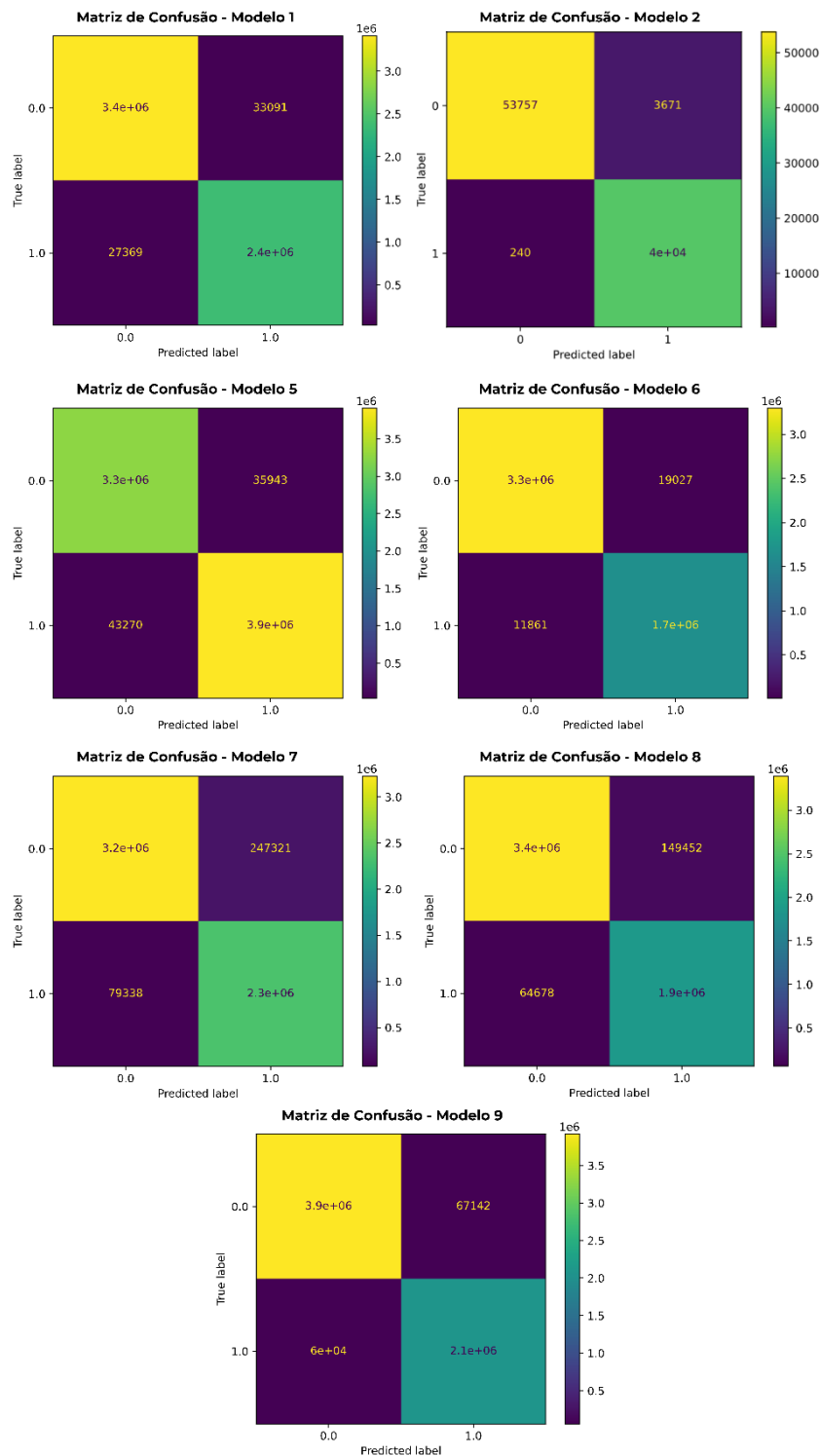
Tabela 7 - Resultados dos modelos de detecção de anomalias otimizados

Evento	Escalador	Modelo	Média F1	Média Recall	Média Precisão	Acurácia
1	<i>Standard</i>	<i>LightGBM</i>	0,987	0,99	0,99	0,99
2	<i>MinMax</i>	Regressão Logística	0,980	0,98	0,98	1,00
5	<i>MinMax</i>	Regressão Logística	0,990	0,99	0,99	0,99
6	<i>Standard</i>	Regressão Logística	0,990	0,99	0,99	0,99
7	<i>Standard</i>	<i>LightGBM</i>	0,940	0,95	0,94	0,94
8	<i>Standard</i>	<i>LightGBM</i>	0,943	0,96	0,95	0,96
9	<i>Standard</i>	<i>LightGBM</i>	0,980	0,98	0,98	0,98

Fonte: Autoria própria.

Ao analisar as matrizes de confusão exibidas na Figura 19, é possível observar a distribuição das previsões realizadas pelos modelos em comparação com as classes reais. Mesmo os modelos 7 e 8 apresentando um desempenho inferior comparados aos demais, eles ainda apresentam resultados satisfatórios, visto que os falsos positivos e falsos negativos apresentam 10 vezes menos ocorrências do que as predições corretas.

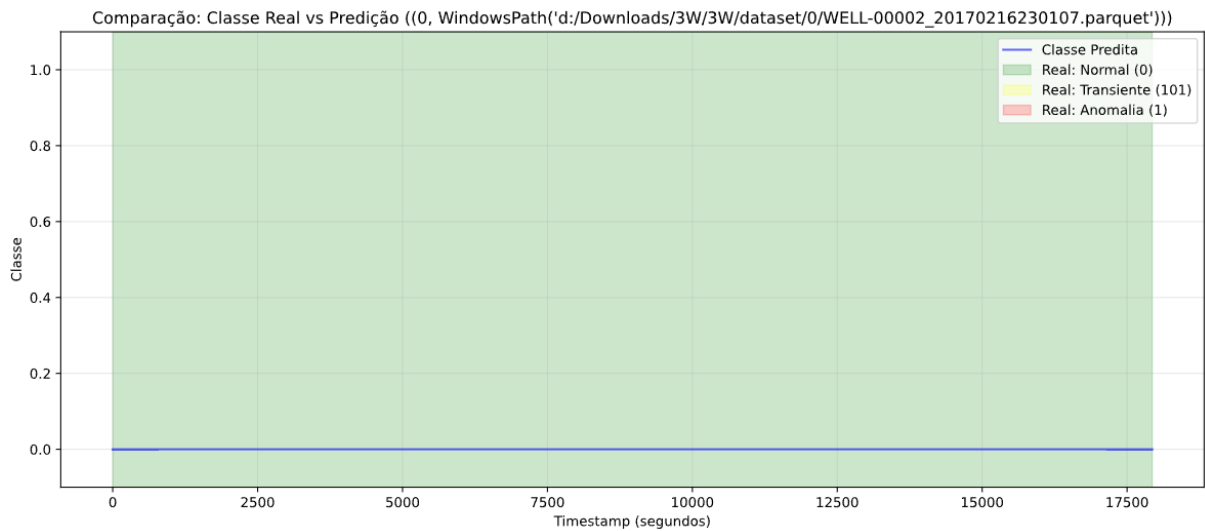
Figura 19 - Matrizes de confusão dos modelos otimizados



Fonte: Autoria própria.

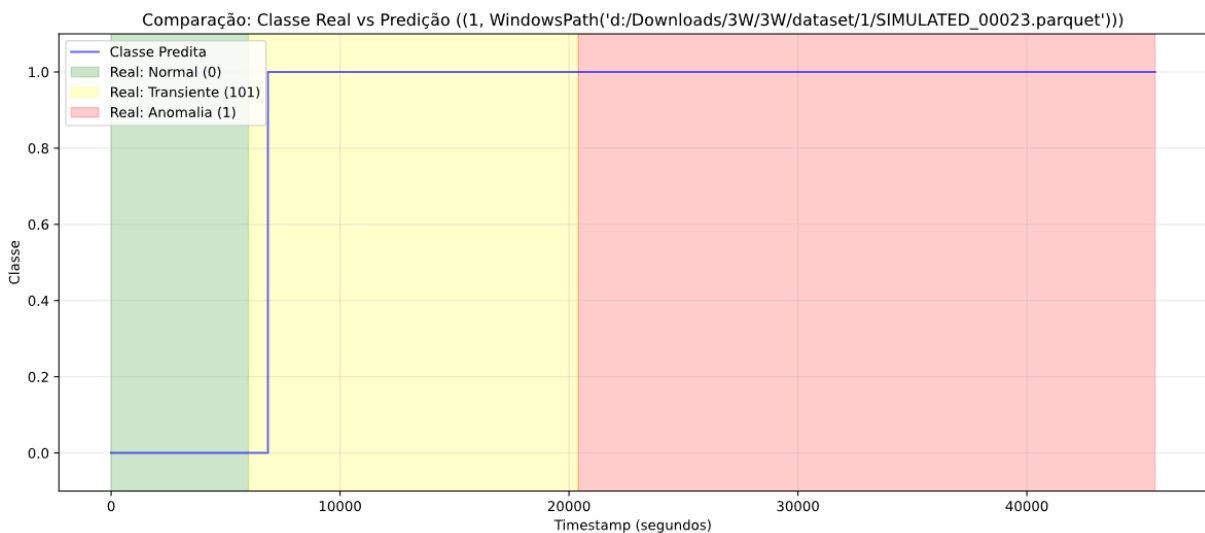
As Figuras 20, 21 e 22, exemplificam a operação normal dos modelos identificando as anomalias sem ou com pouca oscilação na decisão e sem muito atraso na detecção da anomalia, indicando um bom ajuste e confirmando um não *overfitting*, o que o torna um resultado ideal.

Figura 20 - Modelo 1 reconhecendo a ausência de aumento repentino de sedimentos básicos e água durante teste de operação normal



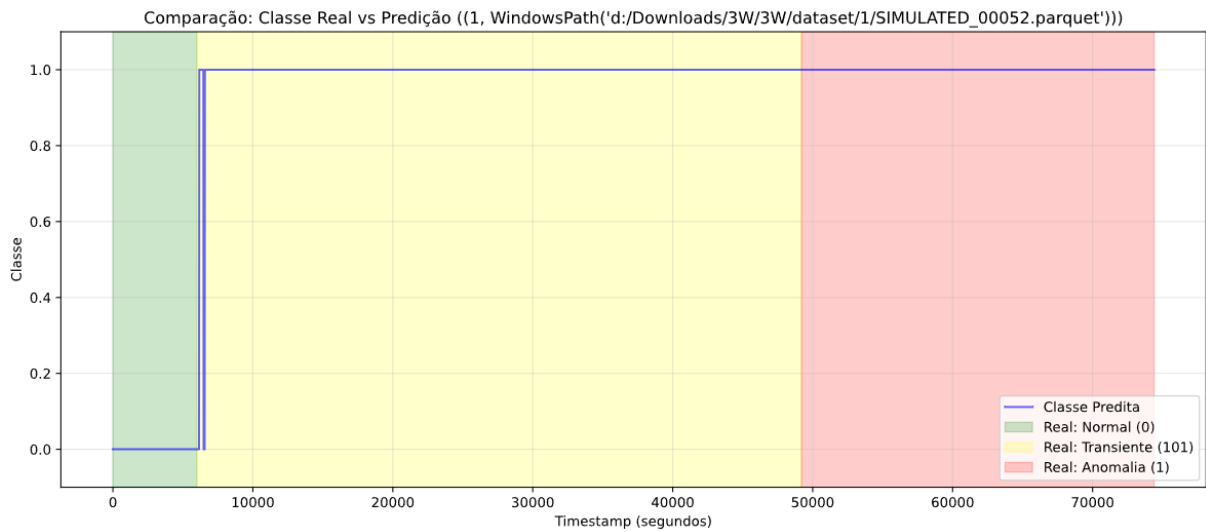
Fonte: Autoria própria.

Figura 21 - Modelo 1 reconhecendo a anomalia de aumento repentino de sedimentos básicos e água com pouco atraso durante teste de detecção de anomalia



Fonte: Autoria própria.

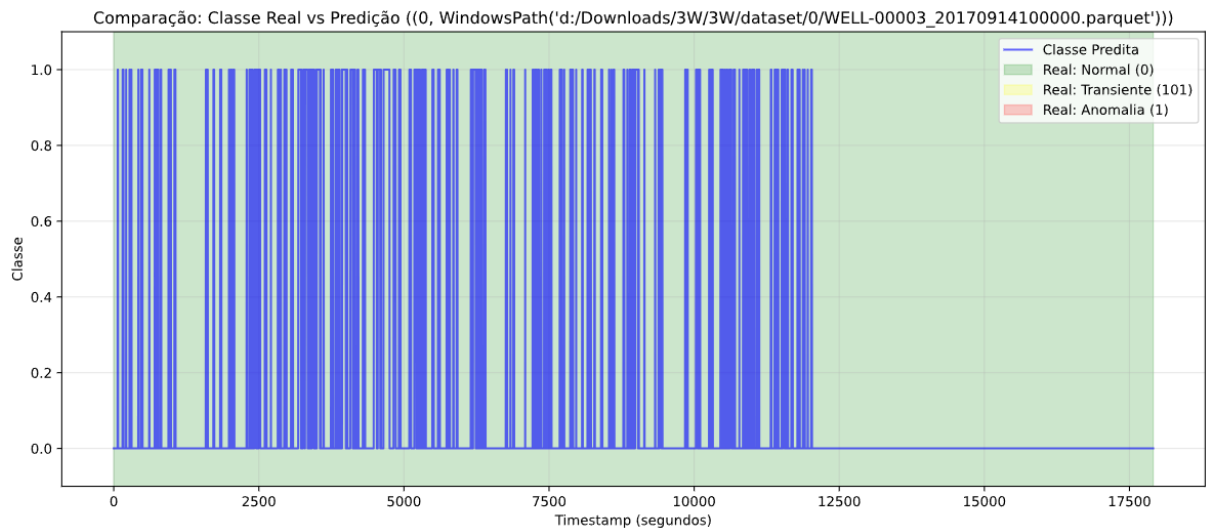
Figura 22 - Modelo 1 apresentando leve oscilação durante teste de detecção de anomalia



Fonte: Autoria própria.

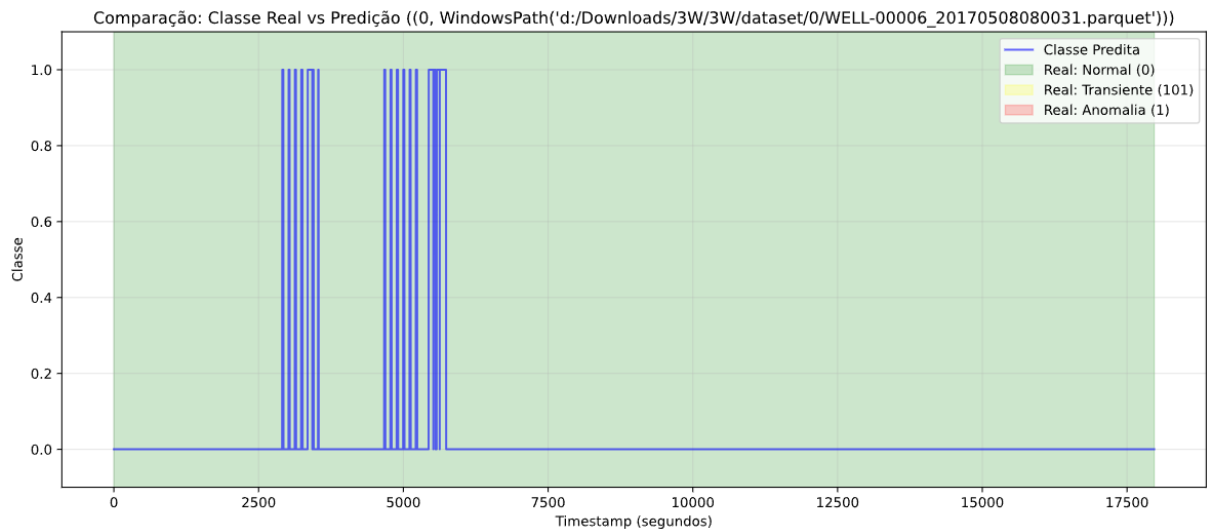
Nas Figuras 23, 24 e 25 são exemplificados gradualmente cenários onde o modelo não apresentou um bom ajuste aos dados, o que pode configurar um *underfitting* se provar ser um caso recorrente.

Figura 23 - Modelo 8 apresentando oscilações não contínuas por um período razoável durante teste de operação normal



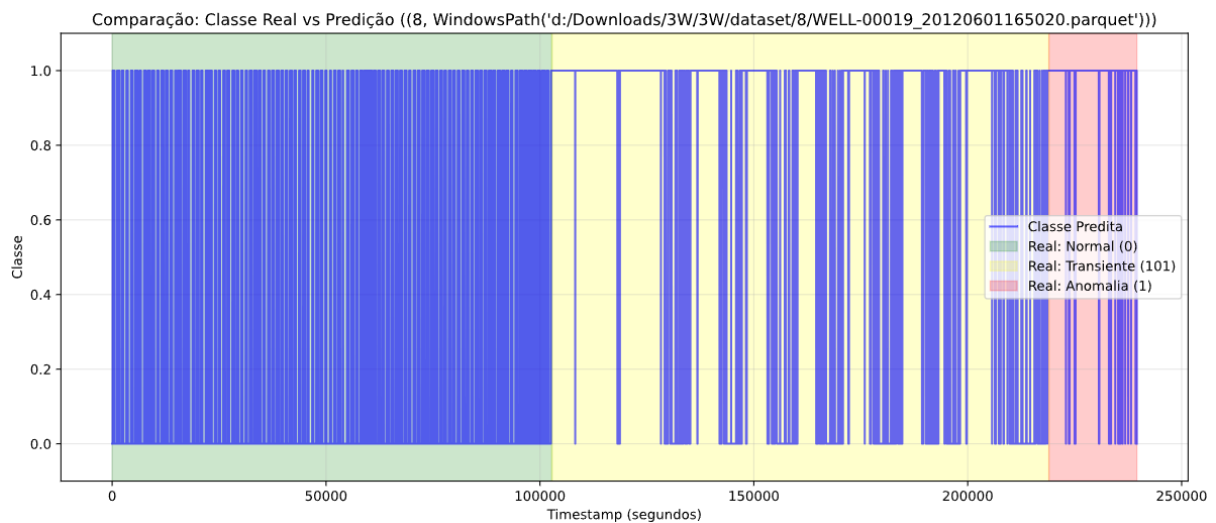
Fonte: Autoria própria.

Figura 24 - Modelo 8 apresentando oscilações contínuas por um período razoável durante teste de operação normal



Fonte: Autoria própria.

Figura 25 - Modelo 8 apresentando muitas oscilações durante todo período de teste de detecção de anomalia



Fonte: Autoria própria.

Os resultados apresentados na Tabela 8, mostram estatísticas do desempenho dos modelos quanto as suspeitas de *underfitting* e *overfitting*. Mesmo em modelos com menor desempenho, é possível ver que na maioria dos modelos o tempo de duração de falso negativo se manteve menor do que a média de duração do período transiente, exceto pelos modelos 5 e 9, o que poderia se criar um alerta de *underfitting*. Porém a baixa quantidade de testes com falsos negativos por mais de 30 minutos, indica que essas ocorrências são apenas ocasionais e que o modelo está respondendo bem as anomalias no geral.

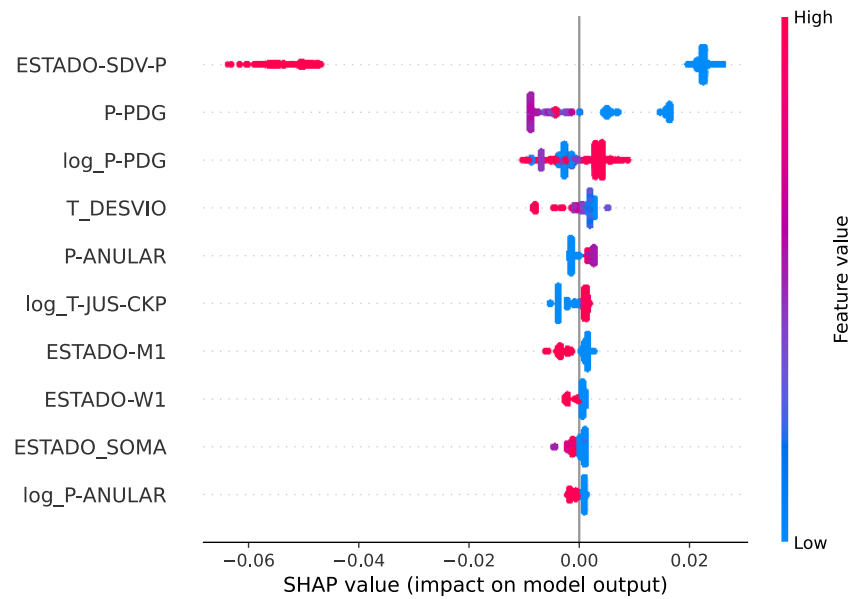
Tabela 8 - Estatísticas de monitoramento por modelo

Modelo	Porcentagem de testes com oscilações	Quantidade de teste com falsos positivos por mais de 30 minutos	Quantidade de teste com falsos negativos por mais de 30 minutos	Maior tempo de duração de falso positivo (minutos)	Maior tempo de duração de falso negativo (minutos)
1	1.65% (11)	2	1	325	133
2	1.65% (11)	2	2	54	54
5	6.45% (43)	2	2	125	104
6	9.15% (61)	1	0	70	20
7	12.74% (85)	8	1	417	121
8	2.55% (17)	1	1	348	89
9	3.15% (21)	10	3	100	380

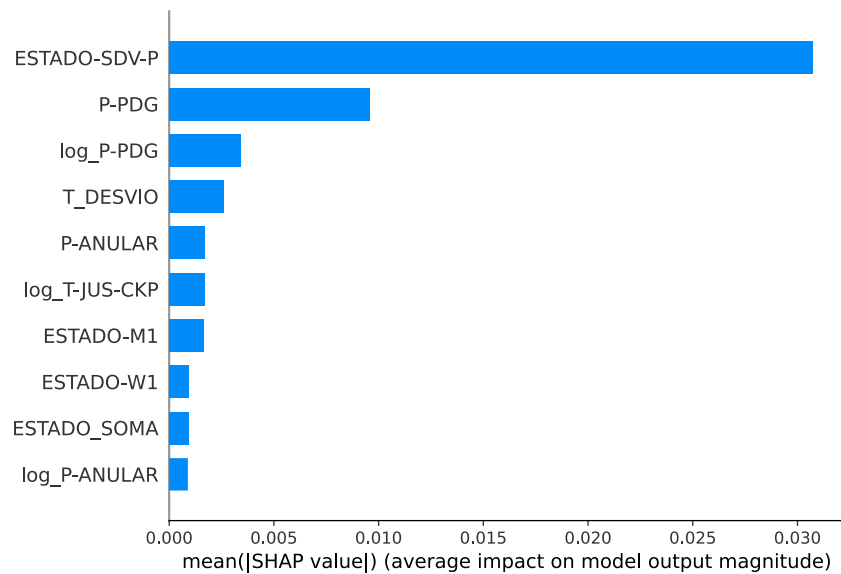
Fonte: Autoria própria.

4.2 ANÁLISE DE IMPACTO DE CARACTERÍSTICAS NOS MODELOS PROPOSTOS

Utilizando a ferramenta SHAP, já explicada anteriormente na subseção 2.3.2.4, foi possível extrair explicações sobre o comportamento do modelo em relação as features que o mesmo utilizou para treinar, serão exibidas apenas as 10 características com mais relevância. As Figuras 26 e 27 mostram 2 variáveis relevantes para o modelo 1, relacionado ao evento aumento repentino de sedimentos básicos e água, a característica binária ESTADO-SDV-P se apresenta bem separada, onde os valores maiores (nesse caso 1) contribuem para um resultado negativo para o modelo (nesse caso 0) sendo a operação normal. A característica P-PDG ou *permanent downhole gauge*, já não se apresenta tão segregado quanto a característica anterior, porém é possível notar que valores altos contribuem para a operação normal registrada pelo modelo.

Figura 26 - Avaliação SHAP para o modelo 1 - *beeswarm plot*

Fonte: Autoria própria.

Figura 27 - Avaliação SHAP para o modelo 1 - *bar plot*

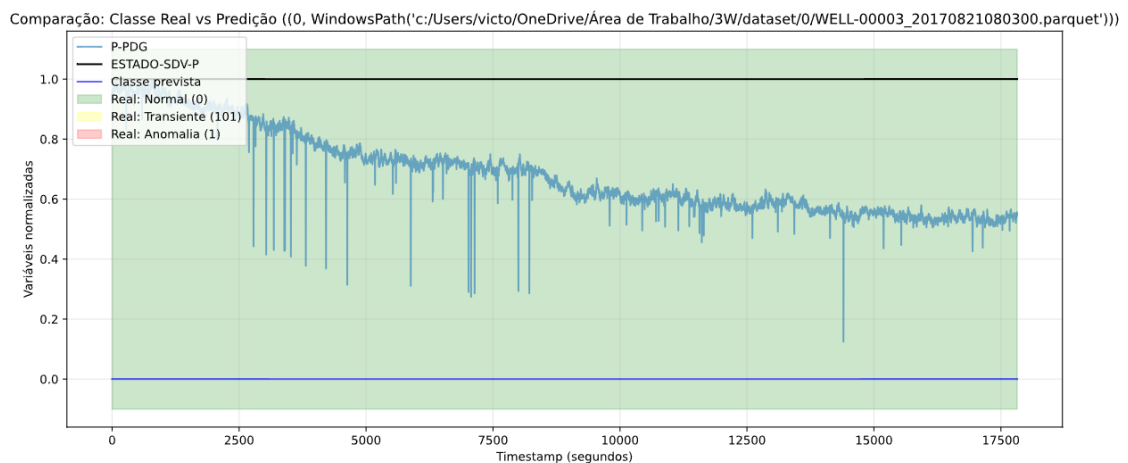
Fonte: Autoria própria.

As Figuras 28 e 29, **exemplificam** o comportamento sobre as 2 características citadas. Podemos ver o valor de 1 para ESTADO-SDV-P enquanto se mantém uma operação normal. A grande falta de registros sobre essa característica quanto à anomalia, a torna inconclusiva de atribuir ao fator de causadora ou efeito gerado. A característica P-PDG mostra uma situação inversa ao que foi mencionado anteriormente visto que a mesma mostra que um aumento de P-PDG aparenta estar correlacionada com a anomalia prevista no evento. Isso pode acontecer por conta de os valores serem escalados para entrar no SHAP, fazendo com que assim o mesmo o

atribua de forma inversa, reforçando a importância de realizar uma análise de monitoramento de teste em relação as características.

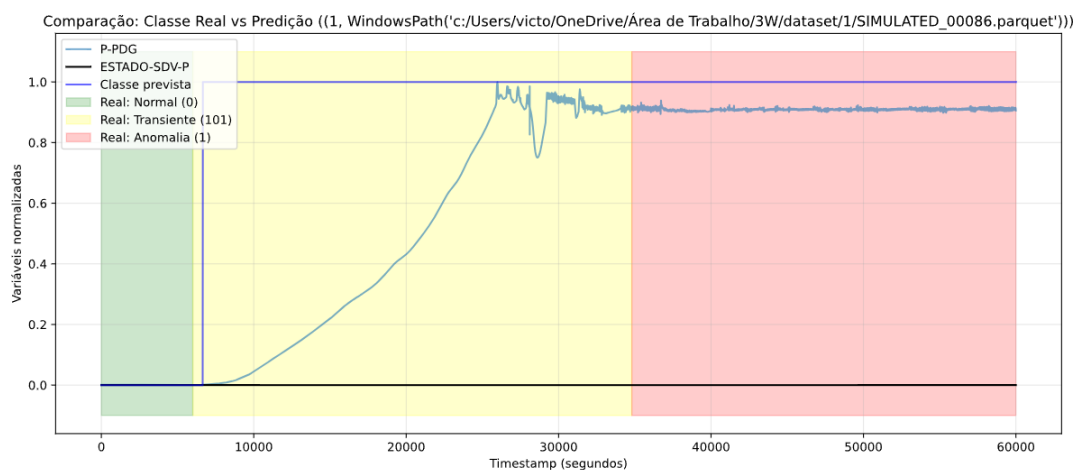
Visto que esse evento é descrito por um aumento de uma razão entre a vazão de água e sedimentos e a vazão de líquido produzido, isso significa que a quantidade de água e sedimentos produzidos está aumentando em relação à quantidade de óleo. A água extra e os sedimentos adicionam massa e volume ao fluxo que está sendo extraído do poço, podendo acarretar entupimentos ou depósitos na linha. Isso gera uma mudança nas propriedades de fluxo do sistema, consequentemente acarretando o aumento da pressão na linha. A vasta semelhança de diversos casos parecidos com a Figura 29, a grande quantidade de registros fornecidos para essa variável e a natureza física do evento, pode se atribuir uma possível relação de causalidade por parte do evento sobre essa variável.

Figura 28 - Teste de monitoramento do modelo 1 sobre a operação normal



Fonte: Autoria própria.

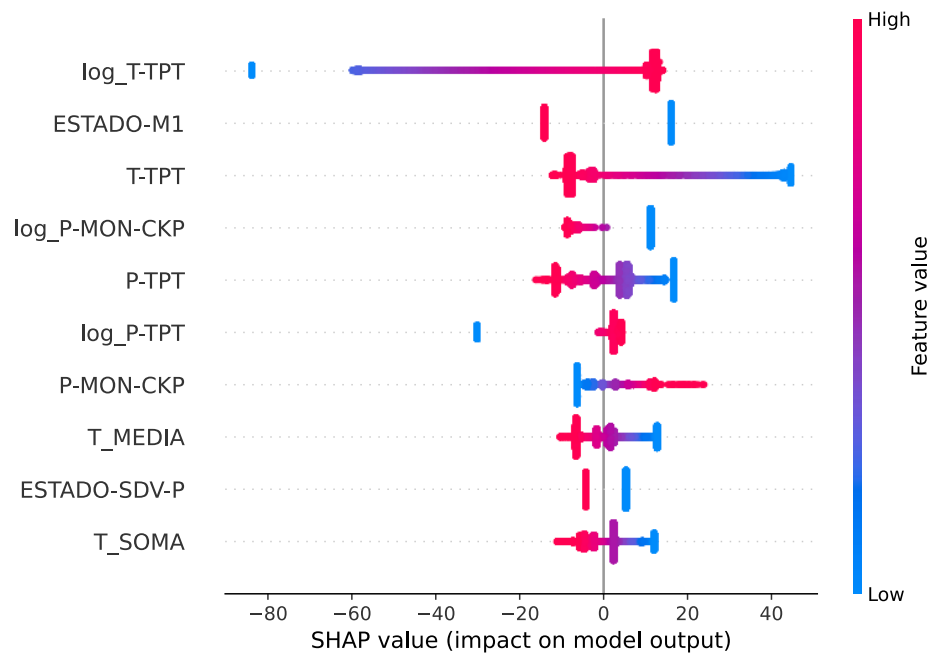
Figura 29 - Teste de monitoramento do modelo 1 sobre ocorrência de evento aumento repentino de sedimentos básicos e água



Fonte: Autoria própria.

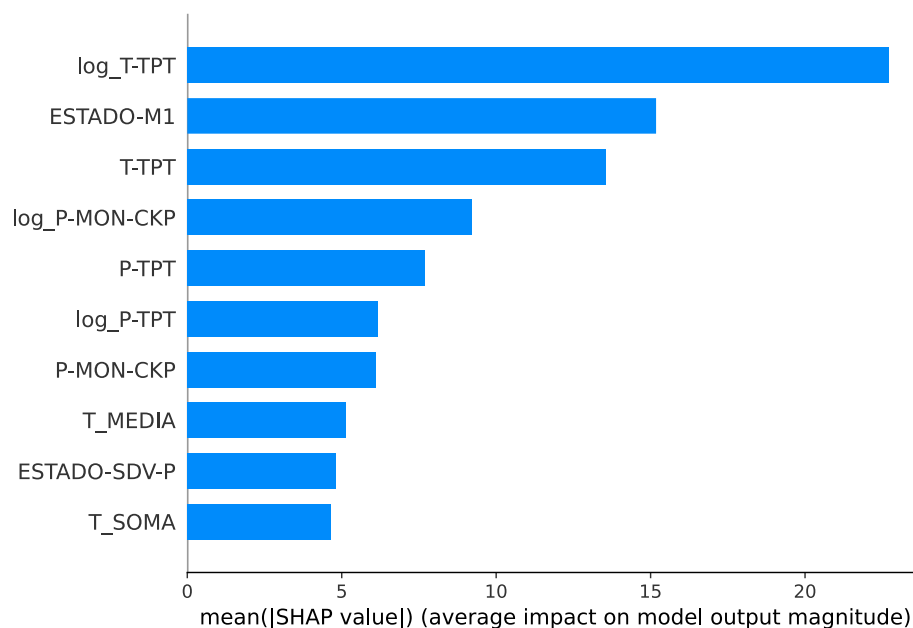
Repetiu-se o mesmo procedimento para o modelo 2, referente ao evento de fechamento irregular de válvula de segurança. A Figura 30 e 31 trazem grande relevância para a característica T-PTP, ESTADO-M1 e log_TPT. Devido a baixa quantidade de registros de ESTADO-M1, é inviável realizar uma análise de possível causa e correlação. As características de T-TPT e log_T-TPT apresentam claros comportamentos que precisam ser confirmados através de uma análise de monitoramento.

Figura 30 - Avaliação SHAP para o modelo 2 - *beeswarm plot*



Fonte: Autoria própria.

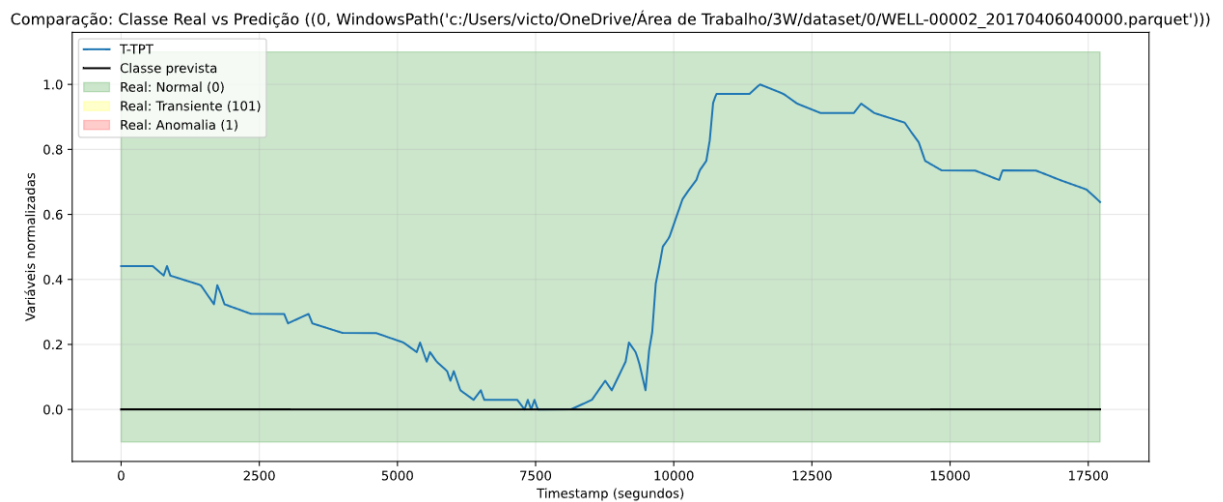
Figura 31 - Avaliação SHAP para o modelo 2 - *bar plot*



Fonte: Autoria própria.

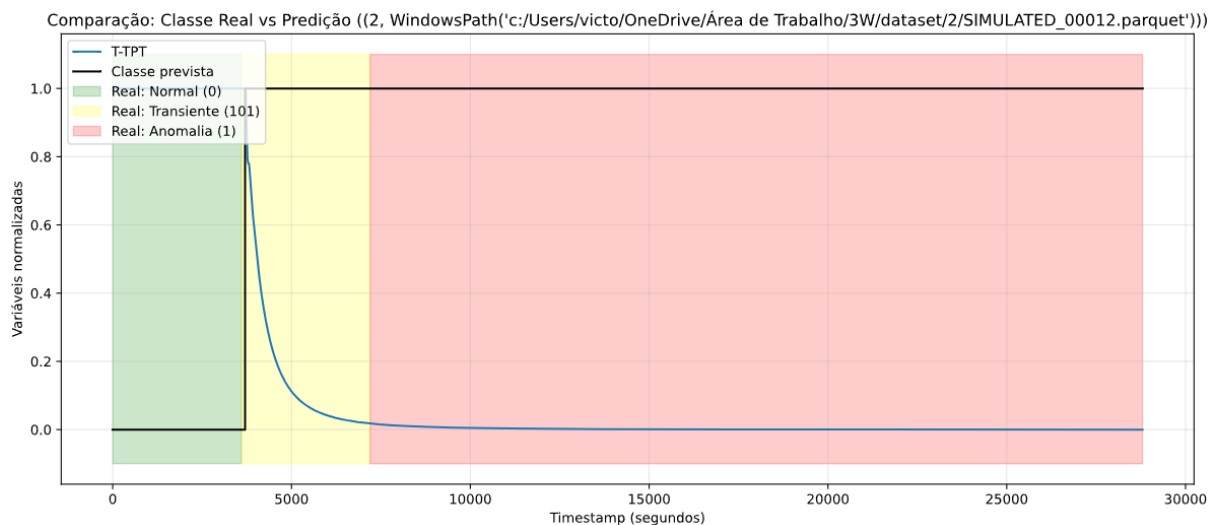
As Figuras 32 e 33, exemplificam o comportamento do evento 2. Durante testes de monitoramento de evento é possível notar o mesmo padrão da Figura 33 se repetir diversas vezes. Porém com uma menor quantidade de testes e com uma menor proporção de reincidência de padrão. Se torna inconclusivo associar unicamente a queda de T-TPT a uma possível causa ou correlação com o evento 2, o motivo disso pode estar associado ao valor dessa variável na Figura 31 que por mais que esteja no topo do ranking, não se encontra tão distante das outras variáveis.

Figura 32 - Teste de monitoramento do modelo 2 sobre a operação normal



Fonte: Autoria própria.

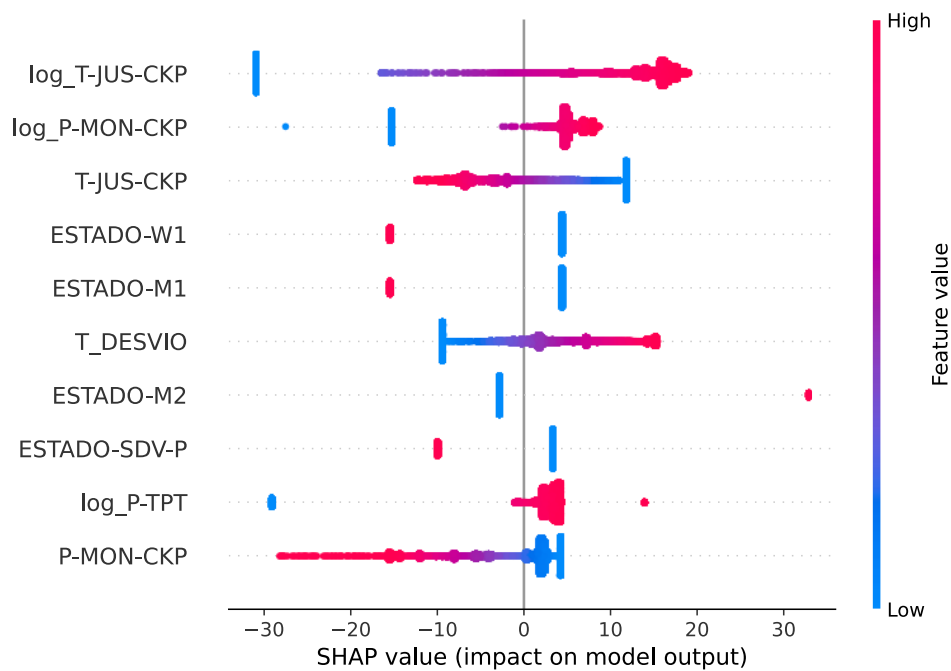
Figura 33 - Teste de monitoramento do modelo 2 sobre ocorrência do evento fechamento irregular de válvula de segurança



Fonte: Autoria própria.

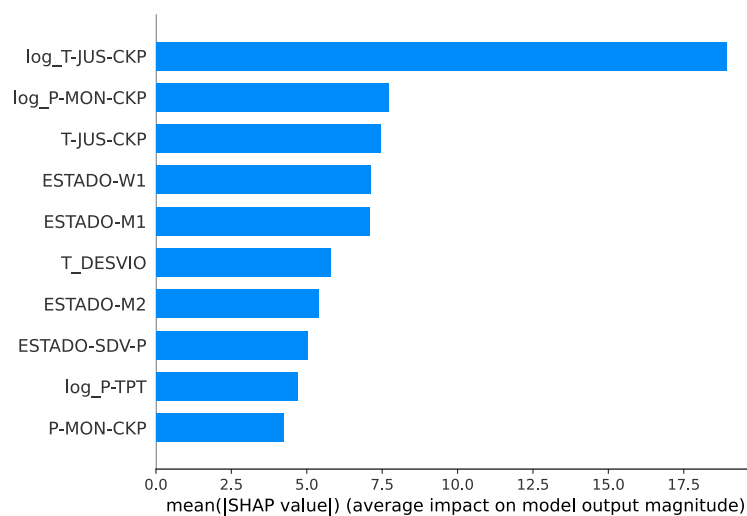
Para o modelo 5, referente ao evento de perda rápida de produtividade, a ferramenta indica, conforme mostrado nas Figuras 34 e 35, grande relevância as variáveis T-JUS-CKP, sua respectiva transformação logarítmica e para a transformação logarítmica de P-MON-CKP. A grande quantidade de registros dessas variáveis as torna viáveis de serem analisadas. As três variáveis apresentam um comportamento que precisa ser confirmado através de testes de monitoramento.

Figura 34 - Avaliação SHAP para o modelo 5 - *beeswarm plot*



Fonte: Autoria própria.

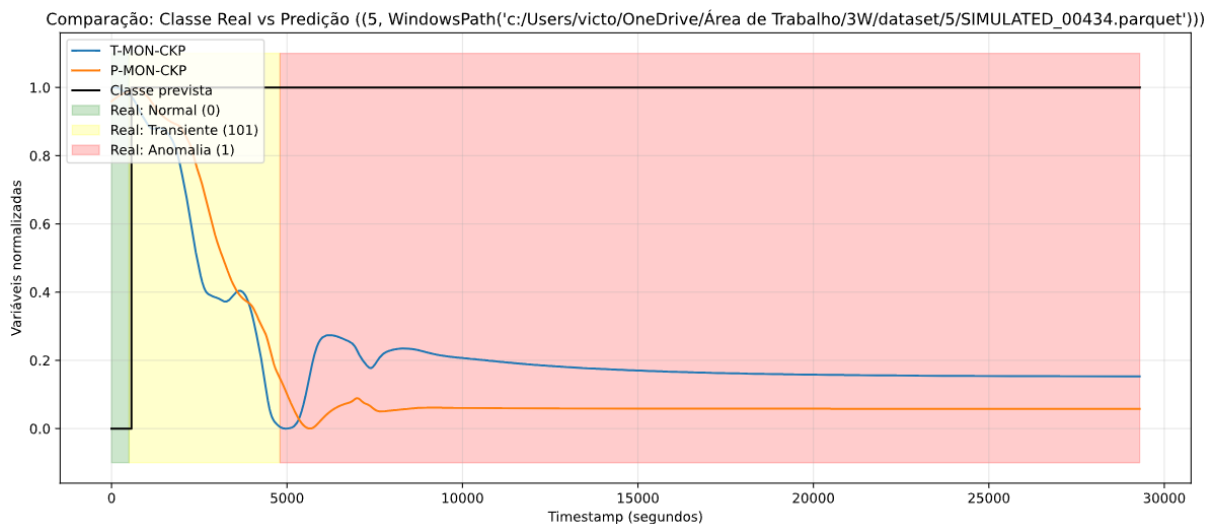
Figura 35 - Avaliação SHAP para o modelo 5 - *bar plot*



Fonte: Autoria própria.

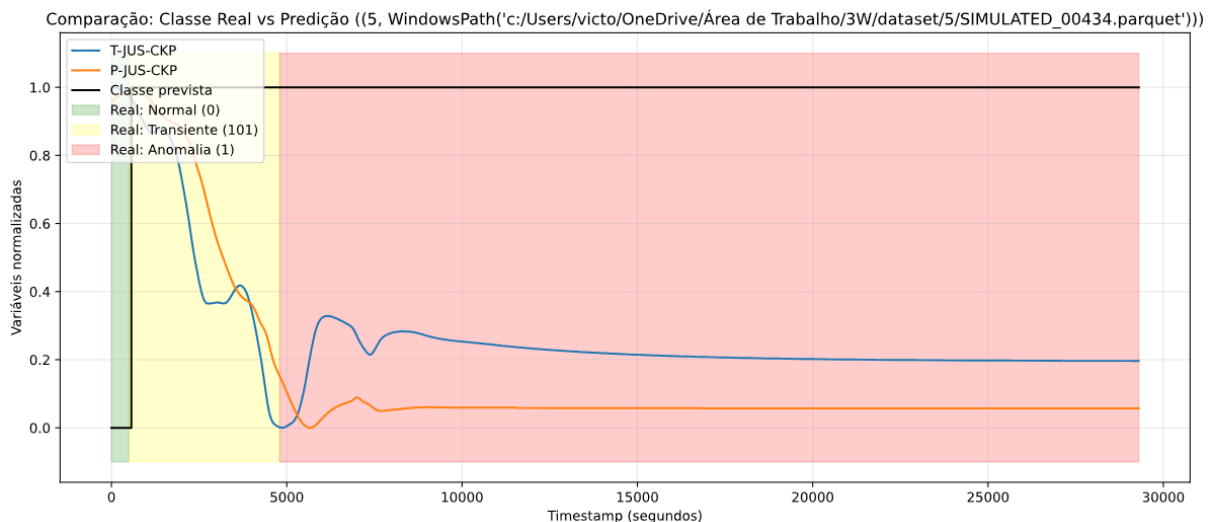
Visto que as variáveis da posição CKP obtiveram uma maior relevância, foi feito uma análise sobre as variáveis de pressão e temperatura da posição CKP, tanto na jusante quanto no montante, mostradas nas Figuras 36 e 37. A falta de registros de algumas variáveis durante a operação normal impede de afirmar com certeza, porém em muitas anomalias é notável que a queda de, principalmente temperatura, está correlacionada de alguma forma com a anomalia. Pode-se especular que conforme, o reservatório é esvaziado, a pressão na linha diminui, fazendo com que assim quando o petróleo passe por CKP tenha sua temperatura reduzida, por conta do efeito Joule-Thompson. Mostrando que possivelmente a baixa temperatura da jusante em CKP pode ser uma consequência do evento.

Figura 36 - Teste de monitoramento do modelo 5 sobre o evento perda rápida de produtividade com as variáveis montante CKP



Fonte: Autoria própria.

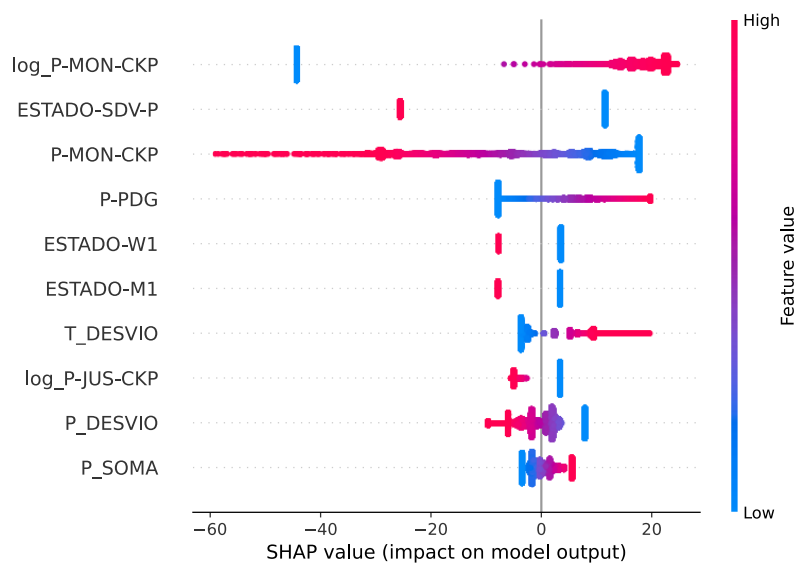
Figura 37 - Teste de monitoramento do modelo 5 sobre o evento perda rápida de produtividade com as variáveis jusante CKP



Fonte: Autoria própria.

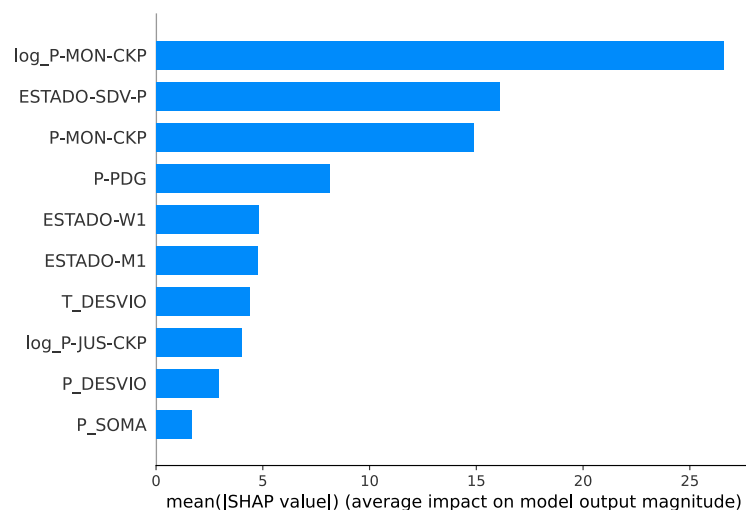
As Figuras 38 e 39, apresentam o modelo dando grande relevância as variáveis de log_P-MON-CKP e P-MON-CKP para o evento de restrição rápida em CKP. Visto que a literatura apresenta pouca referência a esse evento, é interessante realizar um teste de monitoramento para averiguar a resposta da anomalia quanto a variável. Porém dificulta-se em concluir qualquer dependência de causa ou correlação.

Figura 38 - Avaliação SHAP para o modelo 6 - *beeswarm plot*



Fonte: Autoria própria.

Figura 39 - Avaliação SHAP para o modelo 6 - *bar plot*

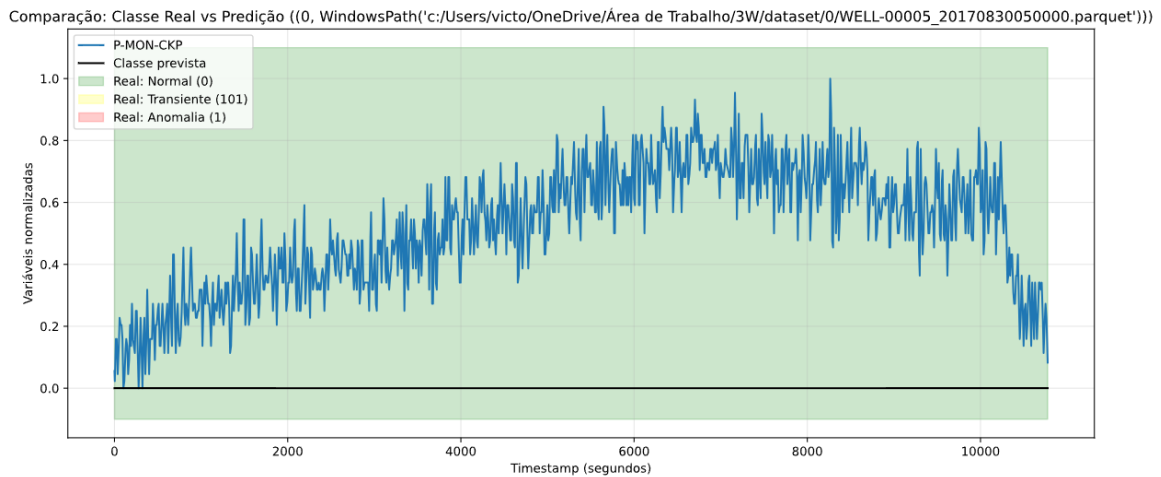


Fonte: Autoria própria.

Testes de monitoramento realizados, mostrados nas Figuras 40 e 41, apresentam resultados inconclusivos, pois existem diversos padrões de repetição na anomalia,

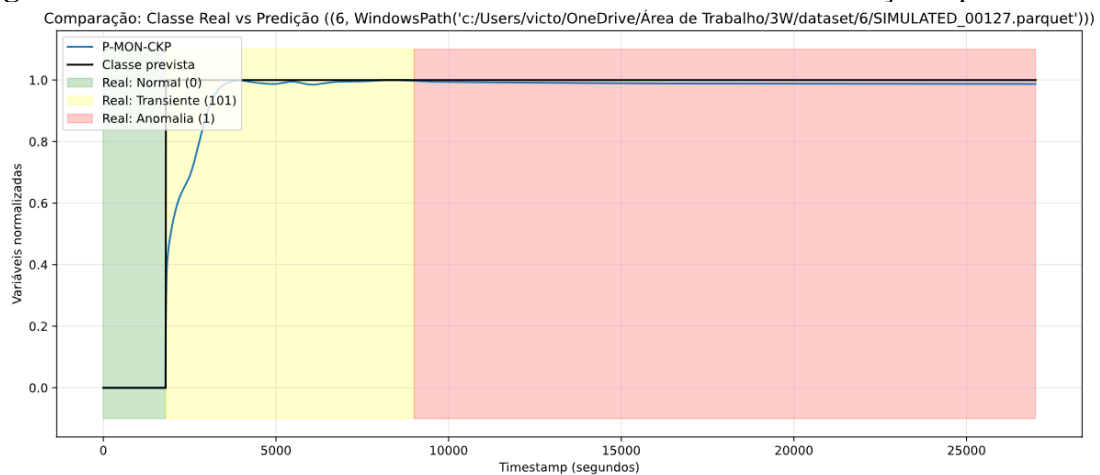
principalmente em relação ao abrupto aumento de pressão que não é visto em nenhuma operação normal, mas é inconclusivo devido à falta de explicações maiores sobre o evento.

Figura 40 - Teste de monitoramento no modelo 6 sobre operação normal



Fonte: Autoria própria.

Figura 41 - Teste de monitoramento do modelo 6 sobre o evento restrição rápida em CKP



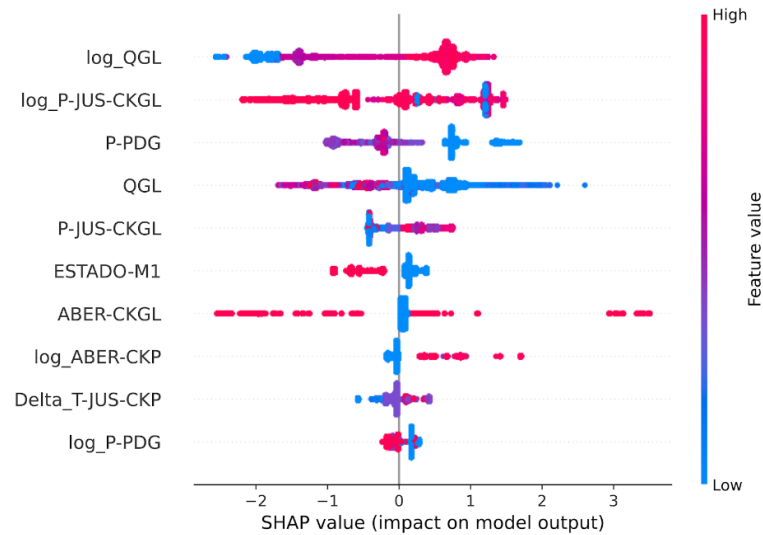
Fonte: Autoria própria.

Os modelos 7 e 8 precisam passar por uma análise mais profunda das características e ajustes de hiperparâmetros, visto que as métricas obtidas não são tão consistentes quanto os outros modelos. Por conta disso, não será abordada a análise SHAP desses modelos neste trabalho, pois seria necessário realizar diversos novos testes relacionados à criação de features, a experimentação com diferentes transformações de dados e provavelmente uma busca mais aprofundada de otimização de hiperparâmetros.

O modelo 9, referente ao evento de hidratos na linha de serviço, apresenta maior relevância para as variáveis de P-PDG, QGL, P-JUS-CKGL e suas respectivas formas logarítmicas, como representado nas Figuras 42 e 43. A quantidade de registros fornecidos por

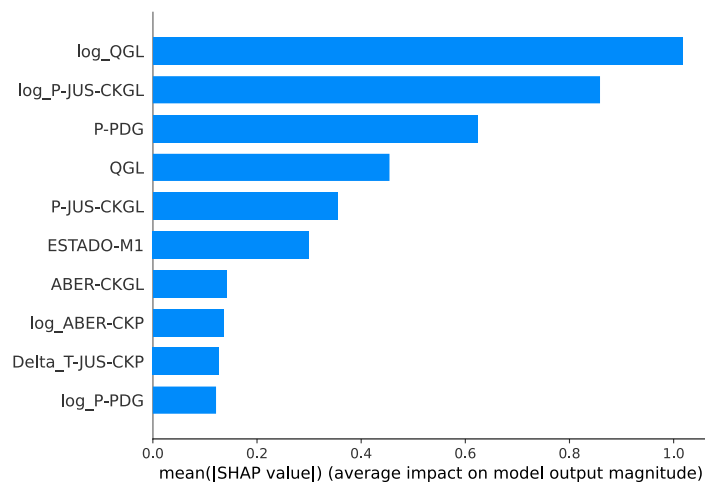
essas variáveis são suficientes para realizar testes de monitoramento e tentar constatar correlações de forma mais clara.

Figura 42 - Avaliação SHAP para o modelo 9 - *beeswarm plot*



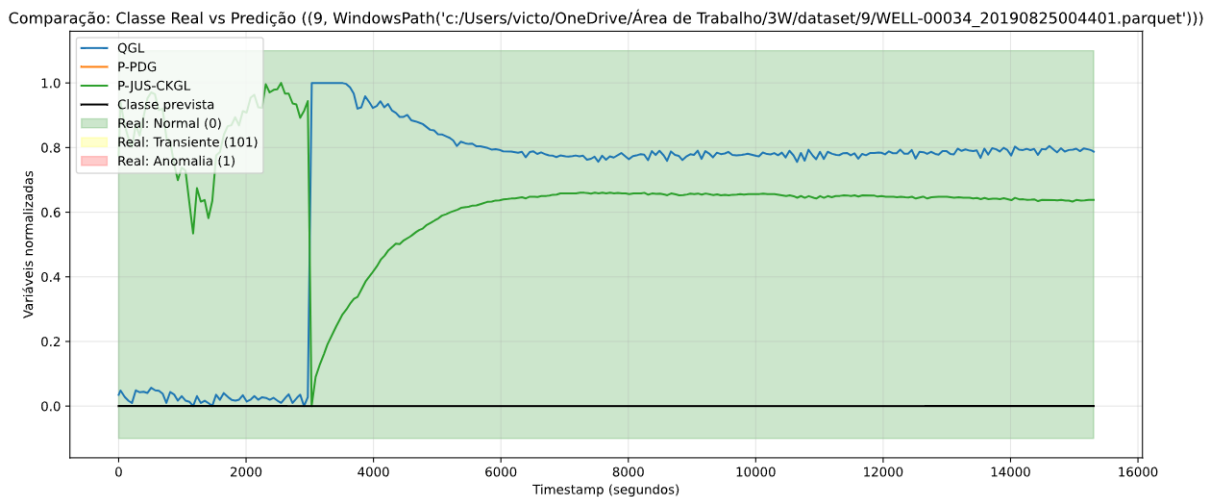
Fonte: Autoria própria.

Figura 43 - Avaliação SHAP para o modelo 9 - *bar plot*

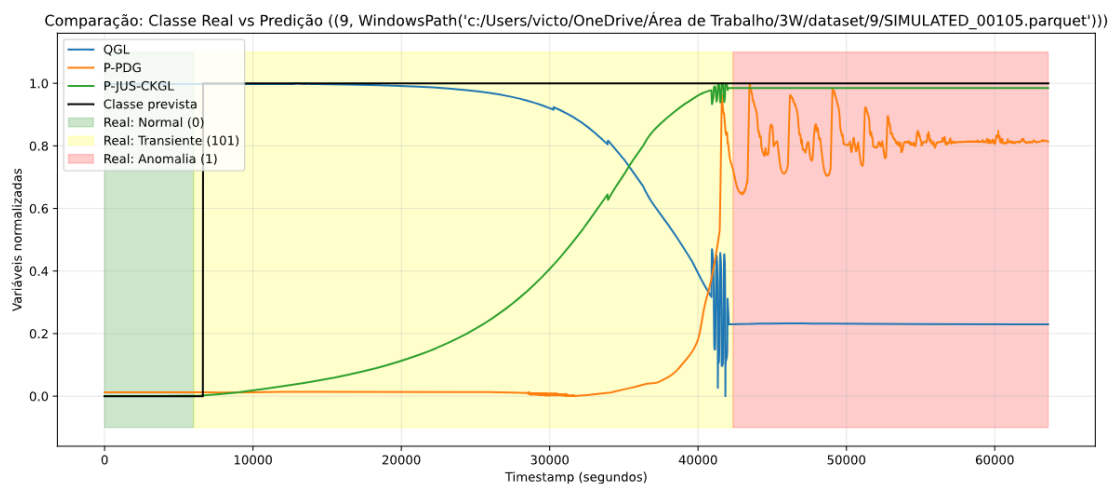


Fonte: Autoria própria.

As Figuras 44 e 45, exemplificam uma parte do comportamento recorrente dessa anomalia e da operação comum. Diversos outros cenários são perceptíveis, o que torna difícil concluir que nesse caso o aumento ou decrescimento de qualquer variável é exclusivamente causada pelo evento. Suposições baseadas na descrição do evento podem existir a fim de tentar explicar o comportamento anômalo, a mais plausível seria que o aumento da pressão seria causado pelos cristais ou por uma resposta de controle a fim de ajudar no escoamento. A baixa na QGL poderia ser uma resposta para evitar contribuir na formação desses cristais.

Figura 44 - Teste de monitoramento de operação normal

Fonte: Autoria própria.

Figura 45 - Teste de monitoramento do modelo 9 sobre o evento de hidratos na linha de serviço

Fonte: Autoria própria.

4.3 MODELO DE CLASSIFICAÇÃO DE ANOMALIAS

Por conta de cada modelo servir exclusivamente para cada evento, é comum que o modelo de outro evento detecte uma anomalia de outro evento, porém sem muita precisão e com resultado oscilatório. A fim de evitar isso, foi extraído o comportamento de todos os modelos para todos os eventos e computado o tamanho máximo de janela de reconhecimento das anomalias sem oscilar, esses dados foram extraídos sobre a condição de se utilizar apenas metade do tempo de anomalia registrada, ou seja, transiente + anomalia. Esses dados foram separados na proporção de 70/30 para treinamento e teste. Neste caso a seguinte estratégia foi proposta para otimizar o modelo:

- Escaladores disponíveis: MinMax e Standard;
- Modelos disponíveis: LightGBM;

- Métrica objetivo: Multi-Logloss;
- Número de iterações dentro de cada cenário na otimização bayesiana: 100.
- Sem peso de balanceamento nas classes.

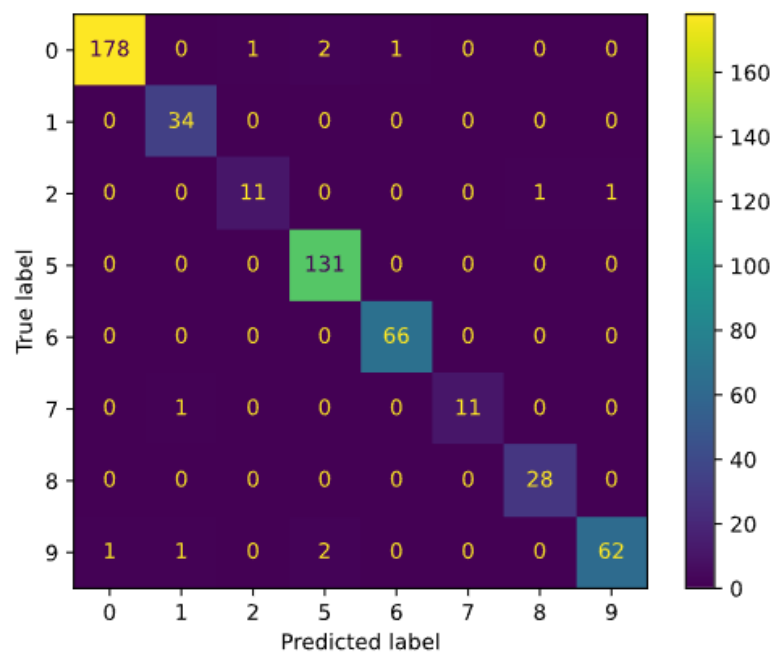
A Tabela 9 e a Figura 46 mostram os resultados obtidos a partir da otimização desse modelo. É possível ver uma baixíssima quantidade de falsos negativos e positivos, indicando um bom desempenho para o modelo, além das métricas de otimização se manterem com um resultado satisfatório.

Tabela 9 - Resultados da otimização do modelo de classificação

Hiperparâmetros	Escalador
Escalador	Standard
Modelo	LightGBM
LogLoss	0,979
Média Precisão	0,96
Média F1	0,98
Média Recall	0,97

Fonte: Autoria própria.

Figura 46 - Matriz de confusão do modelo otimizado de classificação



Fonte: Autoria própria.

4.4 REPRODUTIBILIDADE E HIPERPARÂMETROS DOS MODELOS

O projeto 3W oferece um ambiente com pacotes para que seja realizada a construção de modelos de forma padronizada afim de facilitar a reprodutibilidade dos resultados. Além dos pacotes atuais utilizados pelo projeto 3W, o pacote LightGBM na versão 4.5, o Optuna na versão 4.1.0 e a biblioteca SHAP na versão 0.46.0. Toda a parte randômica dos resultados se utiliza do estado 42. As Tabelas 10, 11 e 12 apresentam os resultados obtidos a partir da otimização dos hiper parâmetros e que foram utilizados neste trabalho.

Tabela 10 - Hiperparâmetros usados nesse trabalho para os modelos de LGBM na detecção de anomalias.

Hiperparamêtros	Modelo 1	Modelo 7	Modelo 8	Modelo 9
Tipo de <i>booster</i>	gbdt	gbdt	gbdt	gbdt
Número de Folhas	25	95	45	56
Profundidade Máxima	10	17	6	18
Taxa de aprendizado	0,00217	1,393	0,00097	0.0418
Número de estimadores	62	100	100	100
Frações de Características	1	0,92	1	1
Regularização L1	0,04471	3,924	0,00035	9,8258
Regularização L2	2,23285	0,206	2,0937	0,800
Peso classe 0	0,859	0,782	0,758	0,721
Peso classe 1	1,1956	1,3843	1,468	1,630

Fonte: Autoria própria.

Tabela 11 - Hiperparâmetros usados nesse trabalho para os modelos de regressão logística na detecção de anomalias.

Hiperparamêtros	Modelo 2	Modelo 5	Modelo 6
C	1,076	453,96	0,0086
Solver	Liblinear	Liblinear	Liblinear
Max. Iterações	133	186	156
Peso classe 0	0,523	1,143	0,764
Peso classe 1	11,29	0,888	1,444
Penalidade	L2	L2	L2

Fonte: Autoria própria.

Tabela 12 - Hiperparâmetros usados nesse trabalho para o modelo de LGBM na classificação de anomalias.

Hiperparâmetros	Modelo
Tipo de <i>booster</i>	gbdt
Número de Folhas	178
Profundidade Máxima	22
Taxa de aprendizado	0,01222
Número mínimo amostras por folha	63
Taxa de amostragem de dados no treinamento	0,6509
Taxa de amostragem de característica por árvore	0,8193
Número de estimadores	949
Frações de Características	1
Regularização L1	2,60187
Regularização L2	0,03128

Fonte: Autoria própria.

5 CONCLUSÃO

O presente trabalho teve como objetivo desenvolver e avaliar um sistema *ensemble* de aprendizado de máquina para a detecção e classificação de anomalias em poços de petróleo offshore. A abordagem adotada concentrou-se na identificação de anomalias para cada tipo de evento e posteriormente realizar a classificação do evento.

Os resultados obtidos demonstraram que esse sistema apresenta alta capacidade de identificar anomalias, mesmo em cenários complexos e desafiadores. A utilização de múltiplos algoritmos combinados permitiu uma maior robustez ao modelo, aproveitando os pontos fortes de cada técnica e minimizando suas limitações individuais. Esse desempenho foi reforçado pelo uso de dados devidamente escalados, contribuindo para uma melhor convergência e precisão dos modelos.

Além disso, os métodos supervisionados empregados mostraram-se eficazes para classificar e separar falhas em diferentes categorias, destacando o potencial prático do modelo para aplicações em sistemas de monitoramento em tempo real. A detecção precoce de anomalias em equipamentos e processos críticos pode não apenas evitar falhas catastróficas, mas também otimizar a eficiência operacional e reduzir custos associados a reparos corretivos e paralisações não planejadas.

Por fim, ressalta-se que, embora os resultados sejam promissores, há espaço para melhorias e futuras investigações. Principalmente na exploração de novas técnicas de *feature engineering* e algoritmos de aprendizado profundo, além da aplicação de modelos em dados mais amplos e diversificados. A implementação de um sistema em tempo real e a validação em cenários reais de operação offshore também seriam passos importantes para a consolidação e, principalmente, a expansão da aplicabilidade prática do modelo desenvolvido.

6 REFERÊNCIAS

- ABASS, H.; BASS, D. **The Critical Production Rate in Water-Coning System**. Permian Basin Oil and Gas Recovery Conference. Texas: Society of Petroleum Engineers, 1988. p. 351-360.
- ALTMAN, D. G. et al. **Statistics with confidence: confidence intervals and statistical guidelines**. 2. ed. London: BMJ Books, 2000.
- ANDREOLLI, I. **Introdução à Elevação e Escoamento Monofásico e Multifásico de Petróleo**. Rio de Janeiro: Interciência, 2016.
- BISHOP, Christopher M. **Pattern Recognition and Machine Learning**. 1. ed. New York: Springer, 2006. 738 p. (Information Science and Statistics). ISBN 9780387310732.
- BLAND, J. M.; ALTMAN, D. G. **Statistics notes: Transformations, means, and confidence intervals**. BMJ, London, v. 312, n. 7038, p. 1079, 1996. DOI: 10.1136/bmj.312.7038.1079.
- GERÓN, Aurélien. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. 2. ed. Sebastopol: O'Reilly Media, 2019.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. Springer, 2009.
- JORDAN, M. I.; MITCHELL, T. M. **Machine learning: Trends, perspectives, and prospects**. Science, v. 349, p. 255-260, 2015. DOI: <https://doi.org/10.1126/science.aaa8415>.
- JOULEIN, Armand; GRAVE, Edouard; BOJANOWSKI, Piotr; MIKOLOV, Tomas. **Bag of Tricks for Efficient Text Classification**. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. Valencia, Spain: Association for Computational Linguistics, 2017. p. 427–431.
- KOHAVI, R. **A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection**. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence. 1995. p. 1137-1143.
- LEITE, Rodrigo. **Introdução à Validação-Cruzada: K-Fold**. Medium, 6 out. 2020. Disponível em: <https://rodrigols89.medium.com/introdução-a-validação-cruzada-k-fold-2a6bcd32a90>.
- LUDERMIR, Teresa Bernarda. **Inteligência Artificial e Aprendizado de Máquina: Estado Atual e Tendências**. Estudos Avançados, v. 35, n. 101, p. 1-21, 2021. DOI: <https://doi.org/10.1590/s0103-4014.2021.35101.007>.
- MILANI JÚNIOR, A.; BOMTEMPO, J. V.; PINTO JÚNIOR, H. Q. **A Indústria do Petróleo como uma Organização Complexa: Modelagem de Negócios e Processo Decisório**. Production, v. 17, n. 1, p. 8-32, 2007.
- MOLNAR, Christoph. **Interpretable Machine Learning: A Guide for Making Black Box Models Explainable**. 2. Ed. 2022. Disponível em: <https://christophm.github.io/interpretable-ml-book/>.

MONIGATTI, Leonie. **Stationarity in Time Series — A Comprehensive Guide**. Towards Data Science, 11 abr. 2023. Disponível em: <https://towardsdatascience.com/stationarity-in-time-series-a-comprehensive-guide-8beabe20d68>.

PASSOS, Dário; MISHRA, Puneet. **A Tutorial on Automatic Hyperparameter Tuning of Deep Spectral Modelling for Regression and Classification Tasks**. Chemometrics and Intelligent Laboratory Systems, v. 223, p. 104520, 2022. DOI: <https://doi.org/10.1016/j.chemolab.2022.104520>.

RIBEIRO, Elieser. **O que é Ciência de Dados**. Medium, 20 dez. 2018. Disponível em: https://medium.com/@elieser_ribeiro/o-que-%C3%A9-ci%C3%Aancia-de-dados-5b2654b9fa08.

SCHLUMBERGER. **The Oilfield Glossary: Where the Oil Field Meets the Dictionary**. Schlumberger Web Site, 2019. Disponível em: <https://www.glossary.oilfield.slb.com/>.

STANDARDS NORWAY. **NORSOK Standard D-010**. Standards Norway, Lysaker, 2013.
SUZUMAR, Laik. **Offshore Petroleum Drilling and Production**. Boca Raton: CRC Press, 2018. ISBN 9781315157177. DOI: <https://doi.org/10.1201/9781315157177>.

VARGAS, R. E.; MUNARO, C. J.; MARQUES CIARELLI, P.; GONÇALVES MEDEIROS, A.; GUBERFAIN DO AMARAL, B.; CENTURION BARRIONUEVO, D.; DIAS DE ARAÚJO, J. C.; LINS RIBEIRO, J.; PIEREZAN MAGALHÃES, L. **A Realistic and Public Dataset with Rare Undesirable Real Events in Oil Wells**. Journal of Petroleum Science and Engineering, v. 181, 2019. DOI: <https://doi.org/10.1016/j.petrol.2019.106223>.