unesp UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
CAMPUS DE GUARATINGUETÁ

IAN RYUJI MATSUMOTO ROLIM

DETECÇÃO DE ANOMALIAS EM POÇOS PRODUTORES DE PETRÓLEO: UMA ABORDAGEM POR MODELOS DE APRENDIZADO DE MÁQUINA

Guaratinguetá - SP

2024

Ian Ryuji Matsumoto Rolim

Detecção de anomalias em poços produtores de petróleo: uma abordagem por modelos de aprendizado de máquina

Trabalho de Graduação apresentado ao Conselho de Curso de Graduação em Engenharia Elétrica da Faculdade de Engenharia e Ciências do Campus de Guaratinguetá, Universidade Estadual Paulista, como parte dos requisitos para obtenção do diploma de Graduação em Engenharia Elétrica.

Orientador: Prof. Dr. Daniel Julien Barros da Silva Sampaio

Guaratinguetá - SP

2024

Rolim, Ian Ryuji Matsumoto
R748d Detecção de anomalias em poços de petróleo: uma abordagem por modelos de aprendizado de máquina / Ian Ryuji Matsumoto Rolim - Guaratinguetá, 2024.
54 f : il.
Bibliografia: f. 51-53

Trabalho de Graduação em Engenharia Elétrica – Universidade Estadual Paulista, Faculdade de Engenharia e Ciências de Guaratinguetá, 2024.
Orientador: Prof. Dr. Daniel Julien Barros da Silva Sampaio

1. Aprendizado do computador. 2. Poços de petróleo. 3. Inteligência artificial - processamento de dados. 4.Banco de dados. I. Título.

CDU 622.323

Luciana Máximo
Bibliotecária/CRB-8 3595

unesp UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
CAMPUS DE GUARATINGUETÁ

IAN RYUJI MATSUMOTO ROLIM

ESTE TRABALHO DE GRADUAÇÃO FOI JULGADO ADEQUADO COMO PARTE DO REQUISITO PARA A OBTENÇÃO DO DIPLOMA DE
"GRADUADO EM ENGENHARIA ELÉTRICA"

APROVADO EM SUA FORMA FINAL PELO CONSELHO DE CURSO DE GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Prof. Dr. DANIEL JULIEN BARROS DA SILVA SAMPAIO
Coordenador

BANCA EXAMINADORA:

Prof. Dr. DANIEL JULIEN BARROS DA SILVA SAMPAIO
Orientador/UNESP-FEG

Prof. Dr. FRANCISCO ANTONIO LOTUFO
UNESP-FEG

Prof. Dr. MAURO HUGO MATHIAS
UNESP-FEG

Agosto de 2024

DADOS CURRICULARES

IAN RYUJI MATSUMOTO ROLIM

NASCIMENTO	22.10.1999 – São Paulo / SP
FILIAÇÃO	Sebastião Normando Rolim Audrea Mie Matsumoto Rolim
2018/2024	Curso de Graduação em Engenharia Elétrica – Universidade Estadual Paulista – UNESP – Câmpus de Guaratinguetá

AGRADECIMENTOS

Em primeiro lugar, agradeço à minha família, à minha mãe Audrea e ao meu pai Sebastião por terem investido tanto em minha formação, me fornecendo todo suporte emocional e financeiro que eu precisasse. Agradeço também à minha irmã Ane por sempre me dar o apoio e companhia nos momentos de lazer. E aos meus avós, Akira e Namie, que me apoiaram incondicionalmente e sempre estiveram presentes.

Em segundo lugar, agradeço a cada um dos meus amigos mais próximos que conheci na faculdade, em especial, Felipe Eleutério, Karen Gomes, Jaine Fernandes e Pedro Laurino, que de alguma maneira fizeram parte dessa caminhada e contribuíram de alguma forma. Também gostaria de agradecer à minha amiga de infância, Isabelle, que me acompanhou desde o início, e tornou a minha jornada mais leve e divertida.

Quero agradecer ao meu Orientador Daniel Julien da Silva Sampaio por todo suporte e por todas as oportunidades oferecidas durante a graduação, bem como pela paciência durante o desenvolvimento do projeto.

Por fim, agradeço ao apoio financeiro da ANP, FINEP, MCTI e FAPESP concedido por meio do programa PRH 34.1 FEG/UNESP.

RESUMO

A extração de petróleo é uma atividade vital para a economia global, fornecendo energia e uma vasta gama de produtos derivados. Contudo, o processo de extração está sujeito a anomalias que podem causar falhas operacionais e perdas financeiras significativas. Este trabalho tem como objetivo desenvolver um modelo de aprendizado de máquina para detectar anomalias em poços de petróleo offshore que operam por elevação natural. Para isso foi utilizado o banco de dados público 3W *dataset*, criado por Vargas (2019), e a metodologia de *benchmark* proposta pelo mesmo. Foram aplicados os algoritmos *Isolation Forest*, *One Class* SVM e o *Dummy*. A avaliação de desempenho foi realizada utilizando o F1 *Score* e o desvio padrão, comparando os resultados com os obtidos na literatura. Os experimentos indicaram que a maioria dos algoritmos testados superou os resultados obtidos na literatura, com exceção do *Isolation Forest*, que, apesar de um desempenho inferior ao reportado por VARGAS (2019) e JUNIOR (2022), ainda apresentou resultados significativos. A seleção de hiperparâmetros e as atualizações nas bibliotecas de Python foram fatores que influenciaram os resultados.

PALAVRAS-CHAVE: Aprendizado em máquina; Detecção de anomalias; Monitoramento de poços de petróleo.

ABSTRACT

Oil extraction is a vital activity for the global economy, providing energy and a wide range of derived products. However, the extraction process is subject to anomalies that can cause operational failures and significant financial losses. This work aims to develop a machine learning model to detect anomalies in offshore oil wells that operate by natural lift. For this purpose, the public 3W dataset created by VARGAS (2019) and the benchmark methodology proposed by him were used. The algorithms Isolation Forest, One Class SVM, and Dummy were applied. Performance evaluation was conducted using the F1 Score and standard deviation, comparing the results with those obtained in the literature. The experiments indicated that most of the tested algorithms outperformed the results obtained in the literature, with the exception of Isolation Forest, which, despite having a lower performance than reported by VARGAS (2019) and JUNIOR (2022), still showed significant results. Hyperparameter tuning and updates in Python libraries were factors that influenced the results.

KEYWORDS: Machine learning; Anomaly detection.; Oil well monitoring

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANM	Árvore de Natal Molhada
ANP	Agência Nacional do Petróleo, Gás Natural e Biocombustíveis
BSW	*Basic Sediment Water*
CKP	Válvula *Choke* de Produção
CSV	*Comma-Separated Values*
DHSV	*Downhole Safety Valve*
FN	Falso Negativo
FP	Falso Positivo
QGL	Sensor de Vazão de *Gas Lift*
PDG	*Permanent Downhole Gauge*
P-PDG	Pressão do Fluido no *Permanent Downhole Gauge*
P-JUS-CKGL	Pressão do Fluido Jusante à Válvula *Choke* de Produção de *Gas Lift*
P-MON-CKP	Pressão do Fluido Montante à Válvula *Choke* de Produção
SVM	*Support Vector Machine*
TN	Verdadeiro Negativo
TP	Verdadeiros Positivo
TPD	*Temperature and Pressure Transducer*
T-PDG	Temperatura do Fluido no *Permanent Downhole Gauge*
T-JUS-CKGL	Temperatura do Fluido Jusante à Válvula *Choke* de Produção de *Gas Lift*
T-JUS-CKP	Temperatura do Fluido Jusante à válvula *Choke* de Produção
UEH	Umbilical Eletro-hidráulico

SUMÁRIO

1 INTRODUÇÃO

No cenário econômico atual, a extração de petróleo representa uma atividade crucial para a economia global, visto que esta, além de fornecer uma fonte vital de energia para a população, com um barril de óleo bruto podendo gerar em média 1.680kWh de energia, fornece ao mercado global uma gama de produtos derivados das diferentes etapas de refino do petróleo. O gás de cozinha, lubrificantes, plásticos e gasolina são alguns exemplos desses produtos (GAUTO, 2016).

De acordo com a Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP, 2024), o consumo mundial de petróleo, no ano de 2022, totalizou 97,3 milhões de barris/dia, o que representa um crescimento de 3,1% (2,9 milhões de barris/dia) com relação ao ano de 2021. Assim, percebe-se que a demanda global de petróleo é alta e considerando suas inúmeras aplicações, garantir que o mesmo seja produzido sem impedimentos é de grande interesse global.

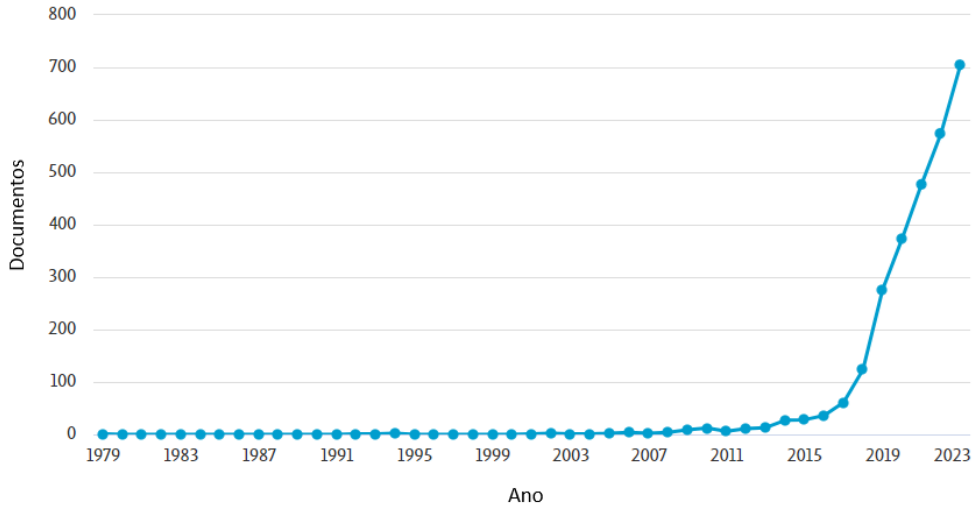
No entanto, durante o processo de extração, é comum ocorrerem anomalias que podem resultar em falhas operacionais e perdas financeiras significativas. O uso de técnicas de detecção de anomalias baseadas em inteligência artificial, como o machine learning, tem sido cada vez mais explorado para mitigar esses riscos e melhorar a eficiência das operações de extração de petróleo (VARGAS, 2019).

Nesse contexto, VARGAS (2019) desenvolveu, até o conhecimento do autor, o primeiro banco de dados público com instâncias reais, simuladas e desenhadas à mão das principais anomalias que ocorrem em poços surgentes de petróleo. Com esse banco de dados é possível treinar diversos algoritmos de aprendizado de máquina para diferentes finalidades, tais como a detecção de anomalias que ocorrem durante a extração de petróleo.

1.1 JUSTIFICATIVA

Para verificar a relevância do tema utilizou-se a plataforma Scopus para verificar a quantidade de publicações que continham as palavras chaves "*Oil and Gas*" e "*Machine Learning*", visto que são os principais conceitos envolvidos no trabalho.. Os resultados, apresentados na Figura 1.1 demonstram a relevância do uso de algoritmos de aprendizado de máquina na indústria do petróleo. Em 2023 houveram 395 documentos publicados e, além disso, pode-se perceber que a quantidade de documentos aumentou nos últimos anos, o que demonstra a crescente relevância do tema.

Figura 1.1 - Gráfico de quantidade de documentos publicados por ano com as palavras-chave "*Oil and Gas*" e "*Machine Learning*"



Fonte: Scopus (2024)

1.2 OBJETIVOS

O objetivo geral do presente trabalho é desenvolver um modelo de aprendizado de máquina para detecção de anomalias que ocorrem durante a extração de petróleo em poços surgentes de petróleo. Utilizando o banco de dados 3W dataset disponibilizado por VARGAS (2019) e tomado como base o benchmark proposto pelo mesmo para detecção de anomalias.

1.3 ORGANIZAÇÃO DO TRABALHO

O presente trabalho está dividido em cinco seções principais. O primeiro capítulo, a introdução, traz uma breve introdução ao tema abordado, bem como a relevância do tema, os objetivos do trabalho e a forma como o trabalho foi organizado.

No segundo capítulo são apresentados os elementos contextuais necessários para o entendimento do trabalho. Neste capítulo são apresentados os principais processos envolvidos na exploração de petróleo, bem como as principais anomalias que serão utilizadas no decorrer do trabalho. Após isso é contextualizado sobre o banco de dados utilizado, o 3W dataset e, por fim, sobre o aprendizado de máquina e os algoritmos utilizados.

Em seguida, apresenta-se a metodologia utilizada para obtenção dos resultados, que são apresentados e discutidos no capítulo seguinte. Por fim, encerra-se o trabalho com as conclusões e sugestões de trabalhos futuros.

2 ELEMENTOS CONTEXTUAIS

Neste capítulo serão descritos os principais conceitos envolvidos no presente trabalho. Primeiramente será contextualizado sobre a exploração de petróleo, bem como as principais anomalias que ocorrem durante esse processo. Após isso será descrito sobre o banco de dados utilizado e então sobre o aprendizado de máquina e os algoritmos utilizados.

2.1 EXPLORAÇÃO DE PETRÓLEO

A exploração de petróleo pode ser dividida conforme sua localização: onshore e offshore. A primeira se trata da exploração do petróleo em terra, enquanto a segunda se trata da exploração de petróleo no mar aberto e em altas profundidades. Em ambos os casos, as estruturas que realizam a extração do petróleo são denominadas poços de petróleo ou apenas poços (THOMAS, 2004).

No Brasil, a maior parte de suas reservas de petróleo se encontram localizadas no mar, ou seja, offshore. Atualmente, os campos marítimos produzem 97,7% do petróleo e 86,1% do gás natural, sendo que 74,7% do total produzido no Brasil é oriundo do Pré-sal (ANP, 2023).

O petróleo proveniente do Pré-sal, que representa mais da metade da produção marítima, é explorado em águas profundas. Essa exploração é complexa e demanda um alto nível de conhecimentos, tornando sua exploração muito mais desafiadora economicamente e tecnologicamente quando comparado à exploração onshore (MORAIS, 2013).

Entretanto, apesar de sua complexidade para exploração, o petróleo proveniente do pré-sal fez com que o Brasil reduzisse não só suas necessidades de importação de petróleo, mas também sua dependência energética dos países vizinhos (RICCOMINI C. et al 2012).

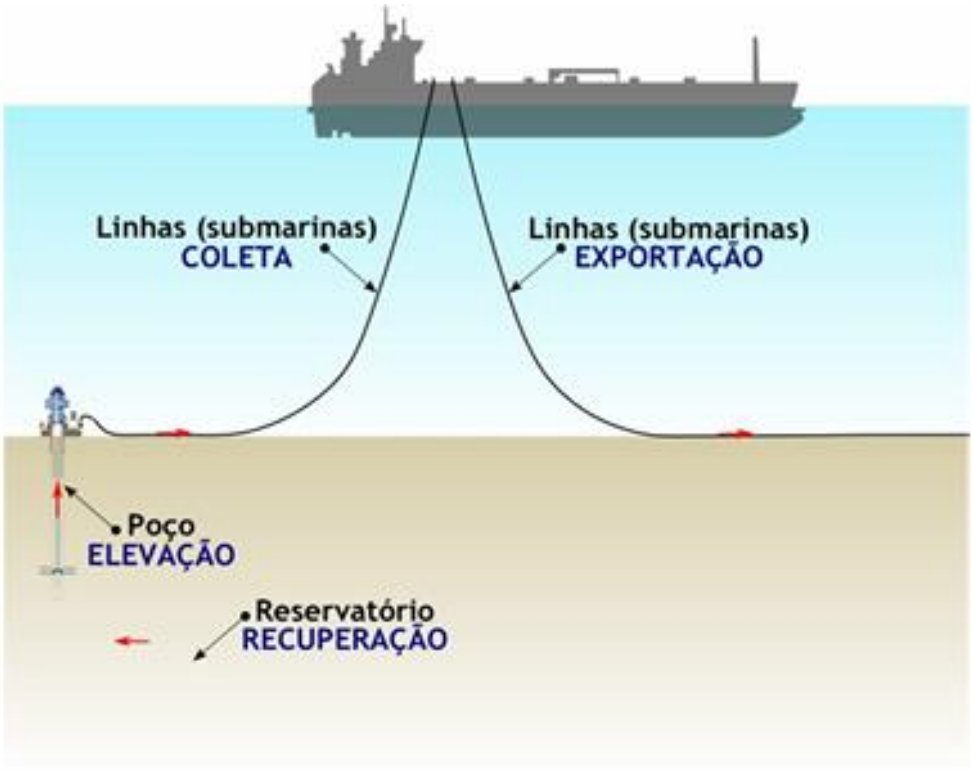
O banco de dados fornecido publicamente por (VARGAS, 2019), utilizado no decorrer desse trabalho, possui dados inerentes a poços produtores de petróleo offshore. Dessa forma, será contextualizado somente sobre o processo de produção de poços de petróleo desta natureza.

O processo de produção do petróleo em alto mar se inicia a partir do escoamento do petróleo do meio poroso até uma unidade estacionária de produção. Na plataforma, ocorre o tratamento, descarte de compostos indesejáveis e escoamento dos compostos de interesse econômico através de oleodutos e gasodutos (ANDREOLLI, 2016).

A Figura 2.1 representa uma linha de produção offshore, ilustrando o sistema empregado nesse tipo de exploração. Os processos envolvidos nessa cadeia de produção podem

ser divididos em quatro fases distintas: recuperação, elevação, coleta e exportação. Cada uma dessas etapas está sujeita à ocorrência de diversos tipos de eventos indesejáveis, uma vez que fenômenos físicos distintos ocorrem em cada uma delas (ANDREOLLI, 2016; VARGAS, 2019).

Figura 2.1 - Etapas do transporte de petróleo



Fonte: Vargas (2019)

A primeira etapa, chamada de recuperação, se trata do escoamento no meio poroso, ou reservatório. Na recuperação, apenas uma parcela do volume total é recuperável, sendo esta influenciada principalmente pelas características da rocha e fluido do reservatório. O primeiro se justifica pelo fato que existem algumas características físicas que o reservatório pode possuir que facilitam ou dificultam sua recuperação. De forma semelhante, a natureza do fluido também é de importância elevada, visto que fatores como a viscosidade influenciam na aderência nos poros e podem impedir a produção de parte do óleo. Além das características mencionadas, também existem outros fatores que impactam nessa etapa, sendo esses: a forma de produção, distribuição dos poços e método utilizado para a recuperação (ANDREOLLI, 2016).

A segunda etapa é chamada de elevação e representa o escoamento na coluna de produção, podendo ser natural ou artificial. Inicialmente, durante o processo inicial de exploração do poço, a pressão interna do reservatório é normalmente suficiente para que o

petróleo e gás cheguem até a superfície. Esse fenômeno é conhecido como elevação natural. Esta fase é mais econômica, pois dispensa o uso de equipamentos adicionais para extrair o petróleo, porém ela não se mantém por um período prolongado de tempo (ANDREOLLI, 2016).

À medida que a produção no reservatório avança e a sua pressão diminui, a elevação natural pode se tornar insuficiente para manter uma taxa de fluxo economicamente viável, tornando necessária a implementação de mecanismos externos para auxiliar na movimentação de petróleo até a superfície e garantir a viabilidade econômica da exploração. A partir do momento em que o poço de petróleo passa a produzir com auxílio desses mecanismos, diz-se que o poço está produzindo através de elevação artificial. Os métodos de elevação artificial são diversos, e cada um é escolhido de acordo com os fatores específicos do poço, como a viscosidade do óleo, comprimento das tubulações e geometria do poço. Dentre os métodos, os principais são: o *Gas lift* contínuo, o bombeio mecânico, o bombeio centrífugo submerso e o bombeio por cavidades progressivas (ANDREOLLI, 2016).

A etapa seguinte a elevação é chamada de coleta, e representa o escoamento que ocorre entre a tubulação, que está em contato com a água do mar, até a plataforma. A tubulação pode ser divida em dois trechos principais: flowline e riser. O primeiro, praticamente horizontal, transporta o petróleo dos poços até os risers e é projetado para suportar as condições ambientais do fundo do mar e possui a fricção como principal fator que influencia no fluxo. Já o riser é praticamente vertical e representa o trecho que faz a conexão entre as tubulações de fundo do mar à plataforma de produção. Nesse trecho, comparado às demais etapas, a componente gravitacional influencia menos no fluxo, já que devido a baixa pressão faz com que parte do gás se desprenda do líquido e se expanda, reduzindo a fração de líquido do fluído. (ANDREOLLI, 2016; VARGAS, 2019).

Por fim, temos a etapa de exportação, que consiste no escoamento dos fluidos já tratados através de um oleoduto ou gasoduto para uma base terrestre ou para outros navios coletores de óleo e gás. Essa etapa é crucial para garantir que o petróleo e gás extraído cheguem aos destinos finais de forma segura, eficiente e econômica. (ANDREOLLI, 2016; VARGAS, 2019).

Diante do exposto, é evidente que cada etapa possui características operacionais distintas e, consequentemente, está suscetível a diferentes tipos de anomalias. O não funcionamento correto de uma das etapas afeta toda a cadeia de produção, podendo ocasionar quedas na produção. Nesse contexto, a detecção dessas anomalias pode auxiliar na identificação precoce dos problemas operacionais, permitindo que medidas corretivas sejam tomadas (VARGAS, 2019).

2.1.1 Poços Marítimos Surgentes de Petróleo

O banco de dados utilizado no decorrer do presente trabalho utiliza dados inerentes a poços de petróleo offshore que utilizam operam através da elevação natural. Dessa forma, é necessário compreender quais os principais equipamentos utilizados e quais os principais sensores disponíveis no sistema.

Existem diversos itens que compõem esse sistema de exploração, sendo os três principais: as plataformas, o sistema de perfuração e o mecanismo de transmissão do petróleo para a plataforma (COSTA; NETO, 2007). Cada um desses componentes desempenha um papel crucial no processo de extração de petróleo, conforme introduzido no capítulo 2.1. A Figura 2.2 apresenta os principais componentes presentes nesse sistema.

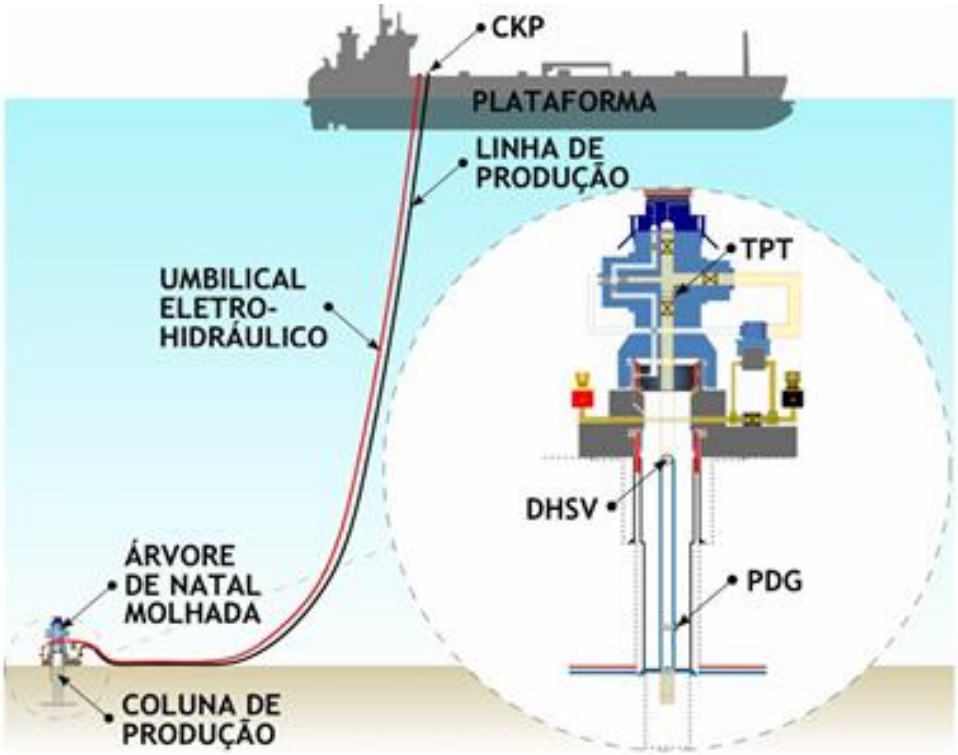
Diante do exposto no capítulo 2.1, referente às etapas envolvidas na cadeia de produção, temos que o petróleo flui inicialmente pela Coluna de Produção, passando pela Árvore de Natal Molhada (ANM), Linha de Produção e válvula Choke de Produção (CKP) até chegar na plataforma.

A ANM é um componente crucial no sistema de produção de petróleo offshore, instalado no leito marinho, sob a cabeça do poço. Sendo composta de um conjunto de válvulas, acessórios e conexões, a ANM tem como função controlar o fluxo de petróleo e gás extraído. Nela são medidos, através do Permanent Downhole Gauge (PDG) e o Temperature and Pressure Transducer (TPD), parâmetros como temperatura e pressão. Em casos de emergência, válvulas de segurança do tipo falha-fecha, como a Downhole Safety Valve (DHSV), podem fechar rapidamente para prevenir vazamentos descontrolados (ANDREOLLI, 2016; VARGAS, 2019).

O umbilical eletro-hidráulico (UEH) é quem faz as interligações entre os componentes da ANM aos da plataforma. O UEH possui três funcionalidades principais: operação dos poços de produção através do sistema hidráulico; fornecimento de energia elétrica para diversos equipamentos e transmissão de dados; e a injeção de produtos químicos, como inibidores de corrosão e hidratos (ANDREOLLI, 2016; VARGAS, 2019).

Por fim, a válvula CKP possui sensores de temperatura e pressão e é responsável pelo controle de abertura no poço. Essa válvula é instalada na plataforma. (VARGAS, 2019).

Figura 2.2 - Esquema simplificado de um poço marítimo surgente de petróleo



Fonte: Vargas (2019)

2.1.1.1 Sensores disponíveis

Foi mencionado anteriormente, no capítulo 2.1.1 alguns sensores disponíveis nos sistemas de elevação e escoamento de petróleo utilizados para medir parâmetros como temperatura e pressão. No geral, em poços de exploração de petróleo offshore possuem uma quantidade menor de sensores que poços onshore. Esse fato se dá por dois motivos principais: o custo elevado e a dificuldade de acesso (VARGAS, 2019).

Os sensores tipicamente disponíveis no sistema de elevação natural estão contidos em três equipamentos distintos, o PDG, TPT e CKP. Esses equipamentos, cujas posições foram apresentadas na Figura 2.2, tipicamente possuem até cinco sensores relevantes. É importante ressaltar que esses sensores podem ou não possuir uma boa confiabilidade, visto que devido aos altos custos de intervenções, a calibração nem sempre é garantida. Em alguns casos, os sensores podem ficar por períodos longos em falha sem que ocorra uma substituição ou calibração do sensor (VARGAS, 2019).

O primeiro sensor é o de pressão do fluido no PDG (P-PDG). O PDG se trata de um manômetro permanente no fundo de poço, cujas falhas estão tipicamente relacionadas ao nível de hostilidade do ambiente. Por se tratar de um equipamento que está na própria coluna de

produção, sua substituição em caso de falha demanda a substituição da própria coluna de produção, caracterizando um processo de custo elevado (VARGAS, 2019; JUNIOR, 2022).

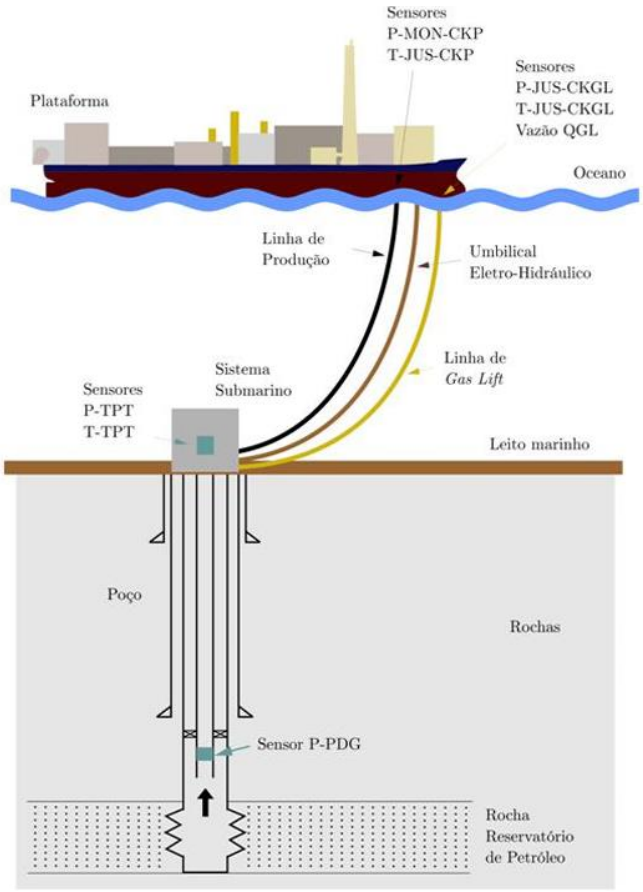
O TPT possui dois sensores, um transdutor de temperatura (T-TPT) e um de pressão (P-TPT). Localizado no interior da ANM, são sensores que possuem confiabilidade considerada boa pelos profissionais da área de Elevação e Escoamento de Petróleo (VARGAS, 2019; JUNIOR, 2022).

Na válvula CKP, localizada na plataforma, existem dois sensores. O primeiro sensor é o sensor de pressão do fluido montante à válvula CKP (P-MON-CKP), e, quando presente, é um sensor considerado de boa confiabilidade. Já o segundo é um sensor de temperatura do fluido jusante à válvula CKP chamado de T-JUS-CKP (VARGAS, 2019; JUNIOR, 2022).

Além dos sensores relacionados à elevação natural, também existem sensores relacionados à elevação artificial. Esses sensores estão localizados na linha de *Gas lift* na plataforma. São esses os sensores de pressão e temperatura do fluido jusante à válvula choke de *Gas lift* (P-JUS-CKGL e T-JUS-CKGL) e o Sensor de vazão de *Gas lift* (QGL).

A Figura 2.3 apresenta a posição aproximada dos cinco primeiros sensores apresentados. Os três últimos sensores, posicionados na linha de serviço, apesar de serem sensores referentes à elevação artificial, estão presentes no banco de dados utilizados pois um poço de elevação natural pode operar através de elevação artificial em alguns momentos. Além disso, suas leituras podem ser úteis no prognóstico de anomalias que ocorrem em ambos os métodos de elevação (VARGAS, 2019).

Figura 2.3 - Representação da posição dos sensores de um poço marítimo surgente de petróleo



Fonte: Junior (2022)

2.1.2 Anomalias em Poços de Petróleo

Uma anomalia pode ser definida como uma observação que se desvia significativamente das outras, sugerindo que a mesma foi gerada por um mecanismo diferente (HAWKINS, 1980). Anomalias também podem ser definidas como eventos que representam falhas nos processos ou sistemas, fazendo com que ocorram desvios das condições operacionais normais (VARGAS, 2019).

A ocorrência de anomalias durante a extração de petróleo pode resultar em impactos financeiros substanciais (VARGAS, 2019). No Brasil, a produção média de um poço de petróleo é de 17,48 mil barris por dia (ANP, 2024), e com o preço médio do barril de petróleo no ano de 2024 sendo de U$84,40 (MACROTRENDS, 2024), a interrupção da produção devido a uma anomalia pode causar uma perda de receita de aproximadamente U$ 1,475 milhões de dólares diários.

O banco de dados 3W dataset possui oito tipos distintos de anomalias, que serão explicadas nas subseções seguintes. Os mesmos estão numerados de acordo com o rótulo de instância utilizado no banco de dados:

1. Aumento Abrupto de BSW;
2. Fechamento Espúrio de DHSV;
3. Intermitência Severa;
4. Instabilidade de Fluxo;
5. Perda Rápida de Produtividade;
6. Restrição Rápida em CKP;
7. Incrustação em CKP;
8. Hidrato em Linha de Produção.

Essas anomalias ocorrem principalmente em poços produtores de petróleo marítimos operados via Elevação Natural e são anomalias que possuem potencial elevado para impactar na produção de petróleo (VARGAS, 2019).

Outro rótulo utilizado para classificar cada anomalia é referente ao nível de observação, podendo ser: normal, transiente de anomalia e estado normal de anomalia. No período normal, não existe anomalia. O período transiente representa o estado transitório entre o estado normal e o estado estável de anomalia. Nesse período, a dinâmica resultante da anomalia ainda está em curso. Por fim, quando essa dinâmica finaliza, inicia-se o período estável de anomalia (VARGAS, 2019).

Para confirmar a ocorrência real de uma anomalia, profissionais de monitoramento dos poços na Petrobras utilizam janelas temporais distintas. As estimativas dessas janelas de tempo estão presentes na Figura 2.4. O tamanho dessas janelas temporais podem ser utilizados para aprimorar o desempenho de algoritmos de aprendizado de máquina (VARGAS, 2019).

Figura 2.4 - Estimativas de tamanhos de janelas temporais para cada anomalia

TIPO DE ANOMALIA	TAMANHO DE JANELA
1 – Aumento Abrupto de BSW	12 h
2 – Fechamento Espúrio de DHSV	5 min – 20 min
3 – Intermitência Severa	5 h
4 – Instabilidade de Fluxo	15 min
5 – Perda Rápida de Produtividade	12 h
6 – Restrição Rápida em CKP	15 min
7 – Incrustação em CKP	72 h
8 – Hidrato em Linha de Produção	30 min – 5 h

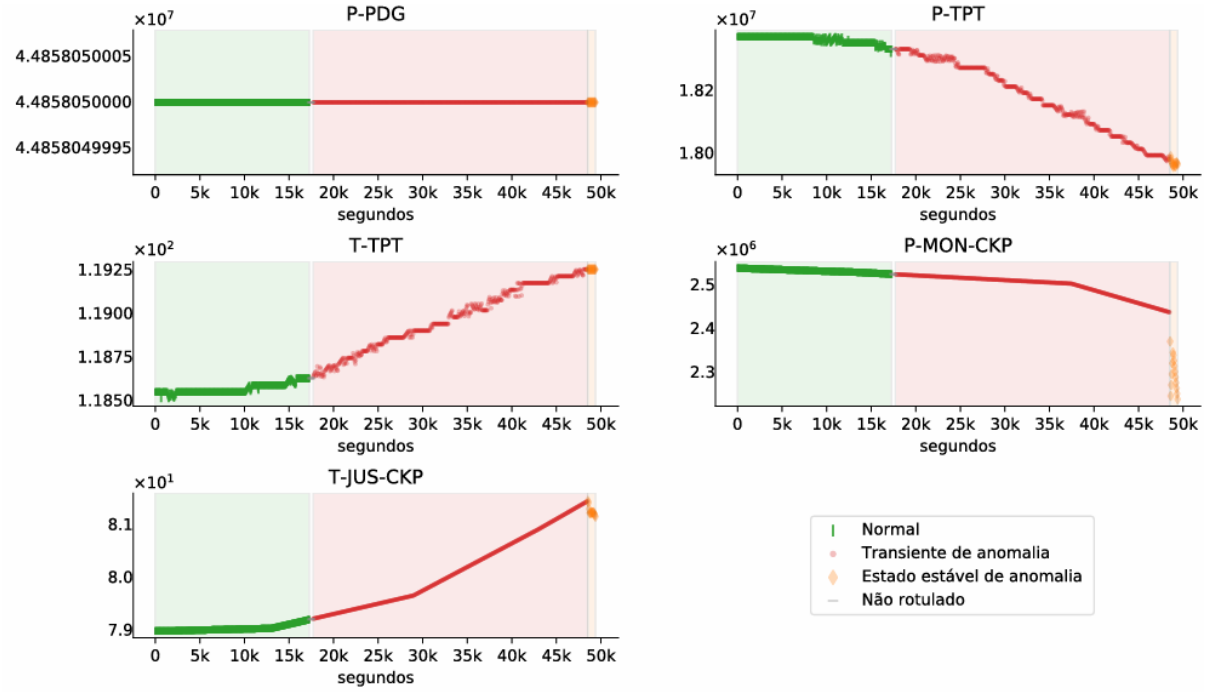
Fonte: Vargas (2019)

2.1.2.1 Aumento Abrupto de BSW

O *Basic Sediments and Water* (BSW) é definido por Andreoli (2016) como: “a razão entre a vazão ou volume de água e sedimentos produzidos e a vazão ou volume de líquido produzido, ambos medidos na condição padrão". No fundo do poço são instalados equipamentos para realizar a contenção de sedimentos, o que reduz a quantidade de sedimentos produzidos. Dessa forma, devido a quantidade reduzida de sedimentos, o BSW é utilizado, na prática, como a fração de água produzida do total de líquidos. Por exemplo, um BSW de 60% infere, na prática, que 60% da vazão volumétrica líquida produzida pelo poço se trata de água (ANDREOLLI, 2016).

Ao decorrer da vida útil do poço, é esperado que seu BSW aumente, uma vez que a água, proveniente do aquífero natural ou dos métodos de recuperação suplementares, atinge o poço. Entretanto, o aumento abrupto do BSW pode acarretar diversos problemas, por exemplo, à produção reduzida de óleo, à garantia de escoamento e à elevação do petróleo. Isso ocorre pois as mudanças no BSW podem alterar diversas variáveis em um sistema de elevação e escoamento de petróleo. No geral, é esperado que as pressões em pontos mais próximos à superfície diminuam e aumentem em pontos mais profundos. Em relação às temperaturas, no geral, ocorre um aumento em todos os equipamentos (ANDREOLLI, 2016; VARGAS, 2019). A Figura 2.5 ilustra um exemplo das leituras dos sensores quando ocorre um aumento abrupto de BSW.

Figura 2.5 - Variáveis de uma instância de aumento abrupto de BSW

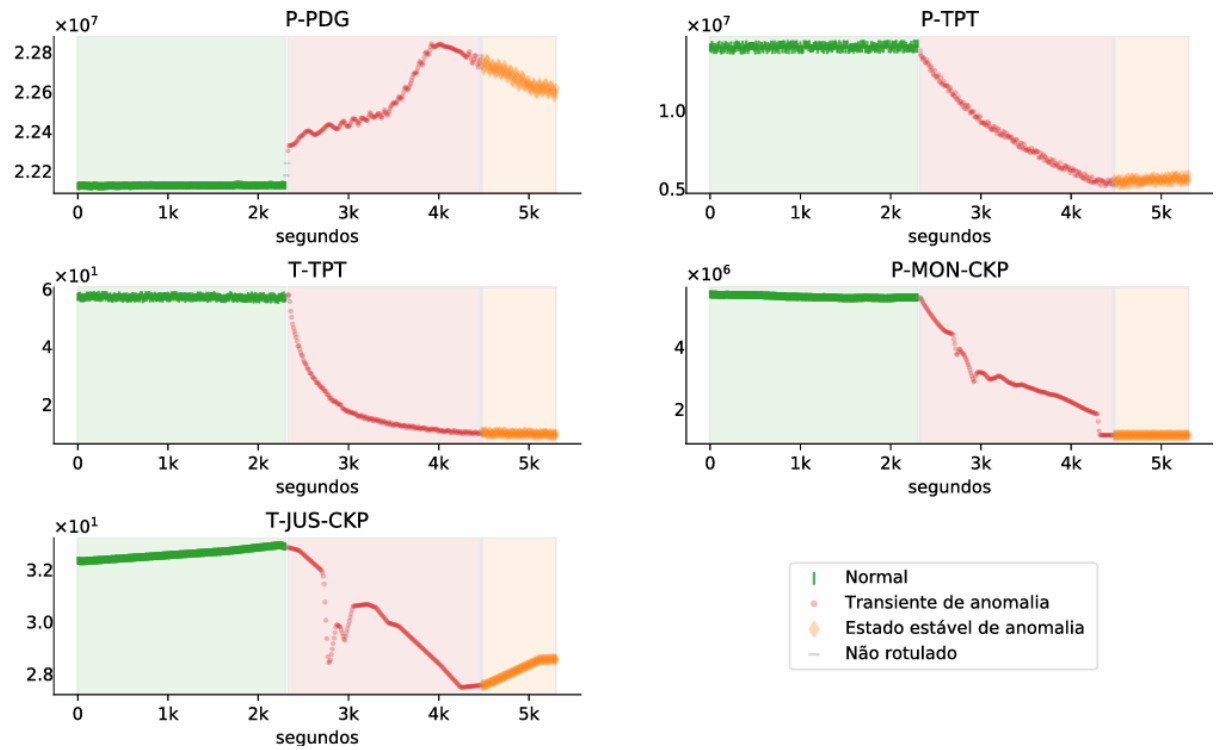


Fonte: Vargas (2019)

2.1.2.2 Fechamento Espúrio de DHSV

A *Downhole Safety Valve* (DHSV) se trata de uma válvula de segurança, mantida aberta por um atuador hidráulico, que assegura o fechamento do poço em caso de desconexão entre o poço e a unidade de produção. O fechamento espúrio dessa válvula impacta na produção de petróleo e pode levar a perda de produção e custos adicionais (VARGAS, 2019). A Figura 2.6 ilustra um gráfico com as leituras dos sensores para um exemplo de instância que passa pelo fechamento espúrio de DHSV .

Figura 2.6 - Variáveis de uma instância de fechamento espúrio de DHSV



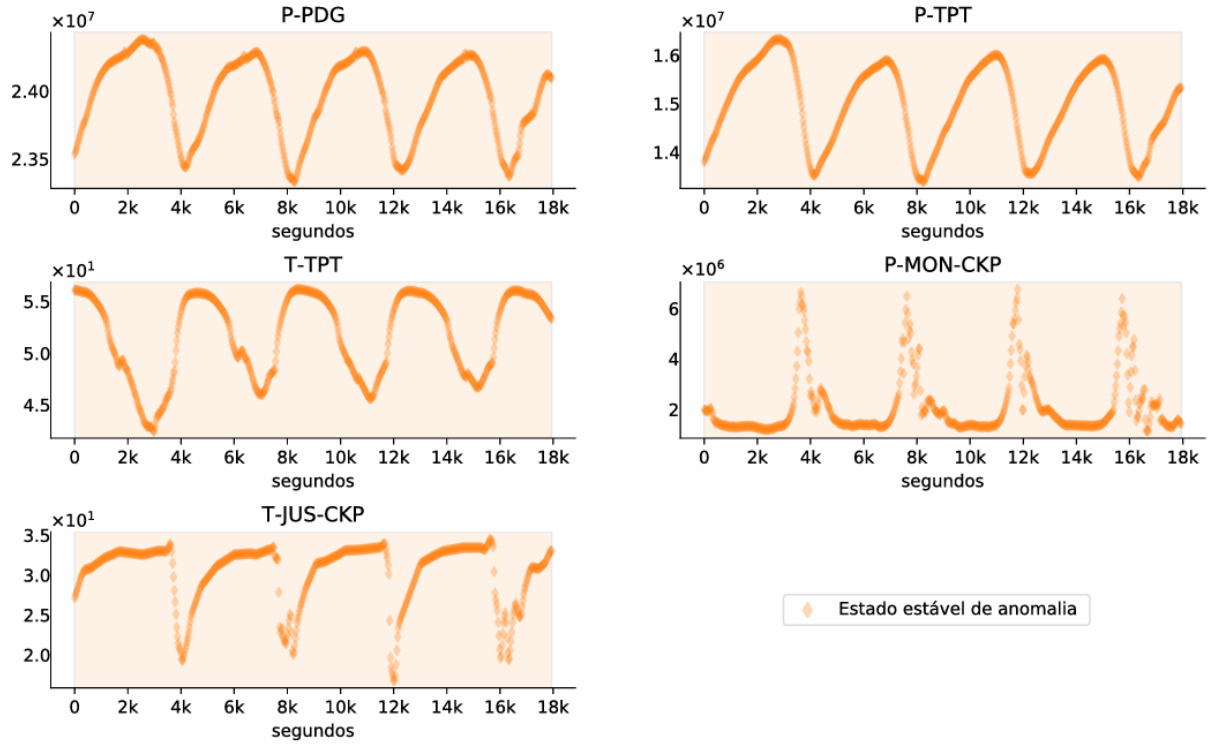
Fonte: Vargas (2019)

2.1.2.3 Intermitência Severa

A intermitência severa surge de uma distribuição não homogênea das fases de gás e líquido dentro das tubulações. Assim, fluem pelas tubulações bolhas de gás separadas por porções de líquido. Na saída, devido a presença dessas bolhas de gás, a produção alterna entre períodos de baixa e alta produção de líquido. Essa produção alterada gera problemas para o processo de separação, ocasionando uma queda geral na produção. Além disso, fatores como a compressibilidade das bolhas de gás e a variação do peso da coluna de líquido causam oscilações periódicas da pressão na tubulação, que por sua vez podem causar danos na instalação e levar a quedas na produção (MEGLIO et al., 2012).

Por se tratar de um fenômeno com periodicidade bem definida e intensidade suficiente, essa anomalia pode ser detectada por sensores ao longo do circuito de produção. Devido aos possíveis impactos na produção que essa anomalia pode causar, seu prognóstico permite que sejam tomadas ações operacionais para reverter o quadro e reduzir a severidade dos impactos causados (VARGAS, 2019). A Figura 2.7 ilustra um gráfico com as leituras dos sensores para um exemplo de instância que passa por intermitência severa.

Figura 2.7 - Variáveis de uma instância de intermitência severa



Fonte: Vargas (2019)

2.1.2.4 Instabilidade no Fluxo

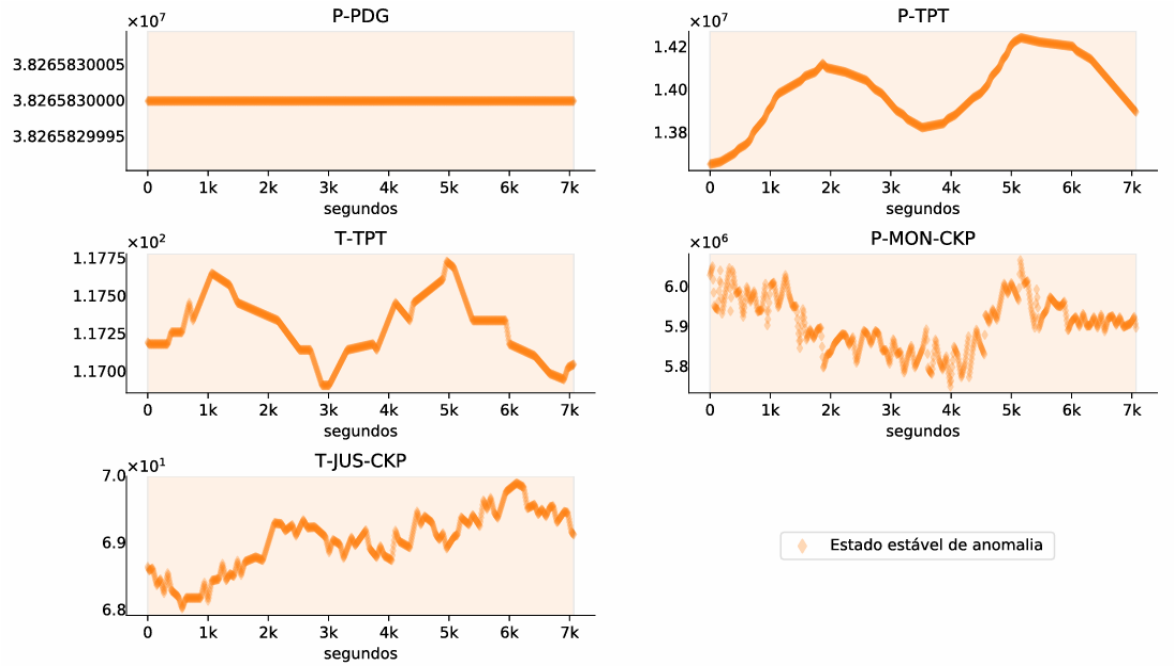
Apesar de muito similar a intermitência severa, a instabilidade no fluxo se diferencia por não possuir uma periodicidade definida e sua natureza é altamente complexa. A instabilidade no fluxo pode evoluir para uma intermitência severa, o que pode ocasionar consequências mais críticas (THEYAB, 2018; VARGAS, 2019). A figura 2.8 ilustra um exemplo das leituras dos sensores quando ocorre uma instabilidade no fluxo.

2.1.2.5 Perda Rápida de Produtividade

A perda rápida de produtividade ocorre quando as propriedades relativas ao reservatório, tais como a pressão estática, índice de produtividade e razão gás/óleo, se modificam à medida que o poço amadurece e fazem que com o fluido não consiga chegar mais até a superfície, levando ao cesar da produção. Esse fenômeno, que também pode ser entendido como a perda de surgência, pode ter suas consequências evitadas através do prognóstico da condição e tomada de ações operacionais (VARGAS, 2019). A Figura 2.9 ilustra um gráfico

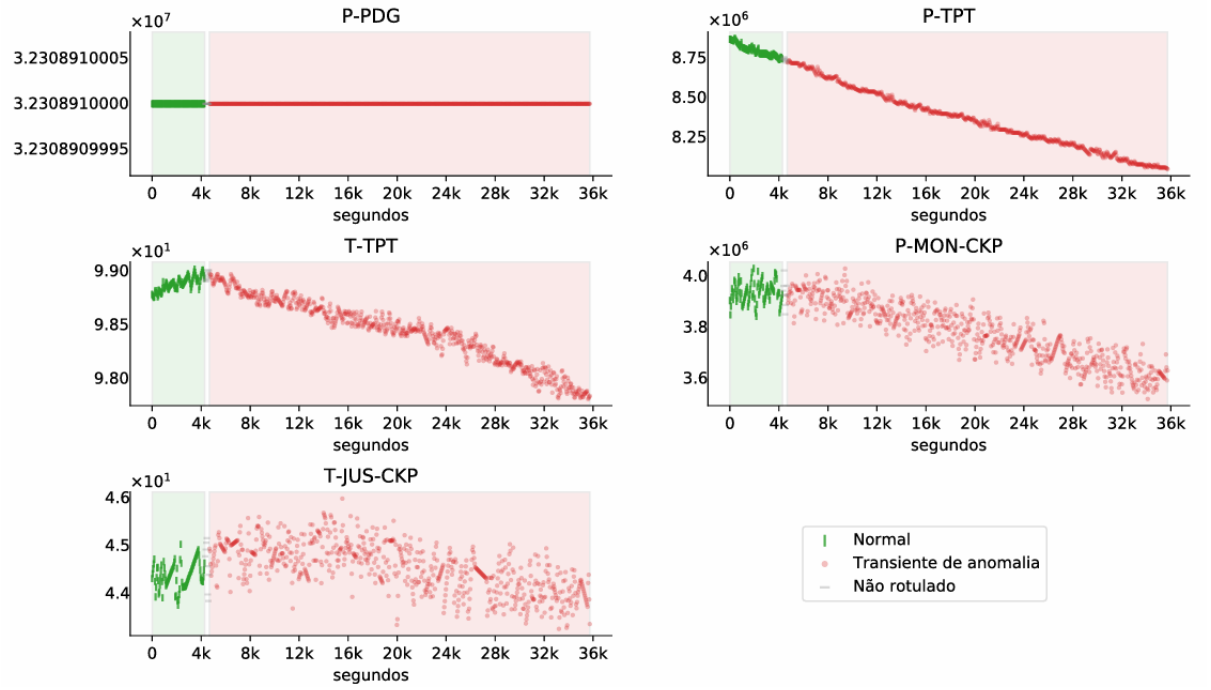
com as leituras dos sensores para um exemplo de instância que passa pela perda rápida de produtividade.

Figura 2.8 - Variáveis de uma instância de instabilidade no fluxo



Fonte: Vargas (2019)

Figura 2.9 - Variáveis de uma instância de perda rápida de produtividade

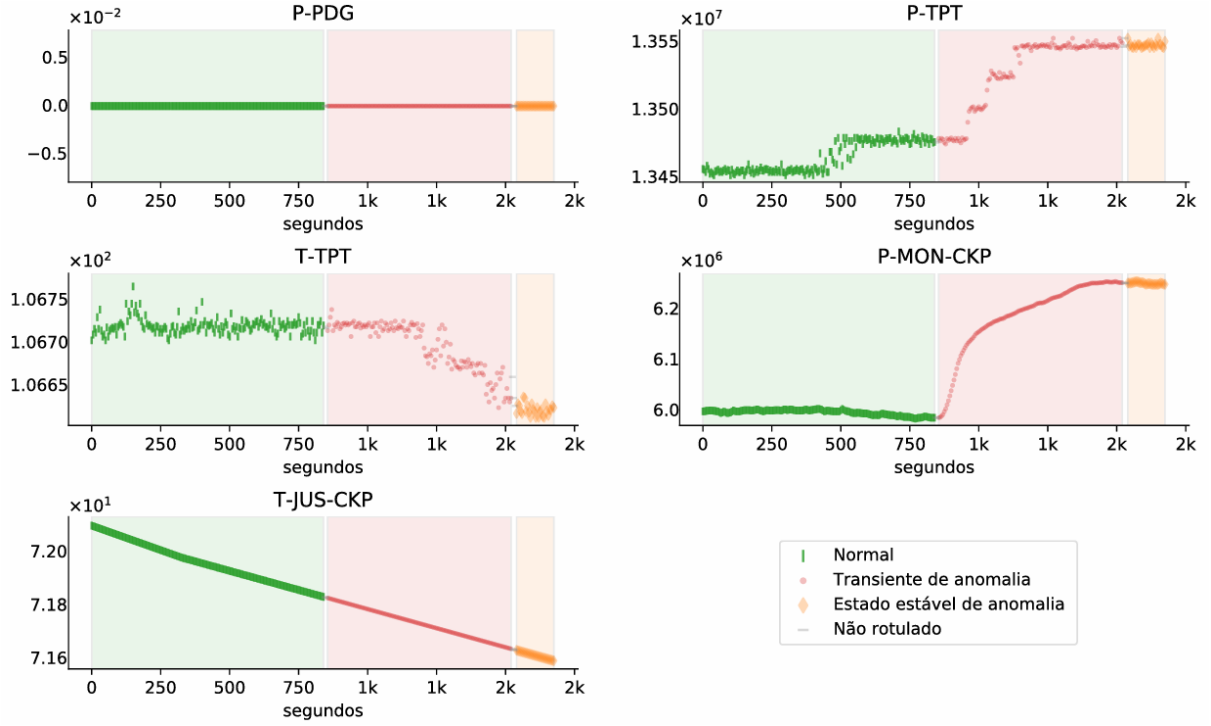


Fonte: Vargas (2019)

2.1.2.6 Restrição Rápida em CKP

Apesar de não ser uma expressão bem definida na literatura, a “restrição rápida em CKP”, é utilizada na Petrobras para definir problemas operacionais que ocorrem na válvula CKP, que instalada no início da unidade de produção, é responsável pelo controle do poço na superfície. Se tratam de restrições durante um período restrito e com amplitude acima de uma referência. No geral, essa válvula é manual, então fechamentos indesejados podem ser desfeitos com mais agilidade. (VARGAS, 2019). A Figura 2.10 ilustra um gráfico com as leituras dos sensores para um exemplo de instância que passa por restrição rápida em CKP.

Figura 2.10 - Variáveis de uma instância de restrição rápida em CKP

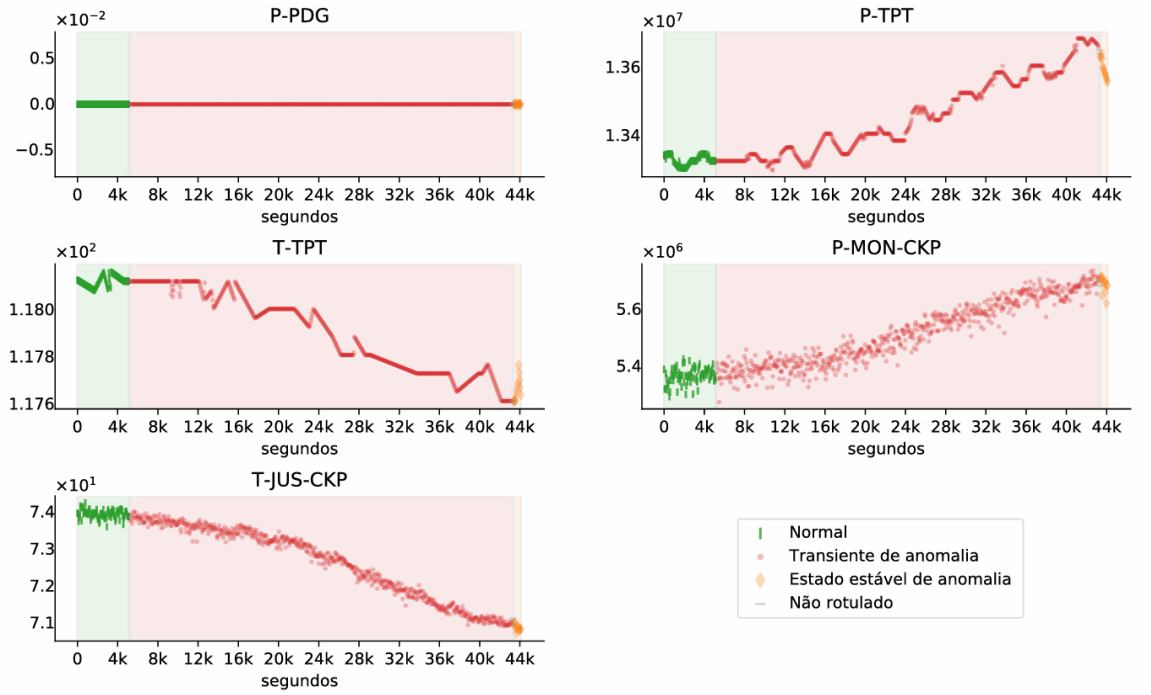


Fonte: Vargas (2019)

2.1.2.7 Incrustação em CKP

A incrustação pode ocorrer em diversas partes do sistema de escoamento de petróleo, e se trata de depósitos inorgânicos de dureza elevada que podem obstruir as tubulações. A válvula CKP tem susceptibilidade a esses depósitos, e a ocorrência de incrustação em CKP pode reduzir a produção de petróleo. A identificação dessa condição permite a utilização de recursos como a injeção de inibidor de incrustação para evitar uma perda elevada de produção (ANDREOLLI, 2016; VARGAS, 2019). A Figura 2.11 ilustra um gráfico com as leituras dos sensores para um exemplo de ocorrência de incrustação em CKP.

Figura 2.11 - Variáveis de uma instância de incrustação em CKP



Fonte: Vargas (2019)

2.1.2.8 Hidrato em Linha de Produção

Considerado um dos grandes problemas da indústria do petróleo, os hidratos são compostos cristalinos formados pela combinação de moléculas de água e gás, geralmente sob condições de alta pressão e baixas temperaturas. No cenário brasileiro de clima tropical, essas condições são encontradas em águas mais profundas. A formação dos hidratos ocorre com mais frequência em gasodutos e poços produtores de gás, porém também podem ocorrer em poços de petróleo (ANDREOLLI, 2016; VARGAS, 2019).

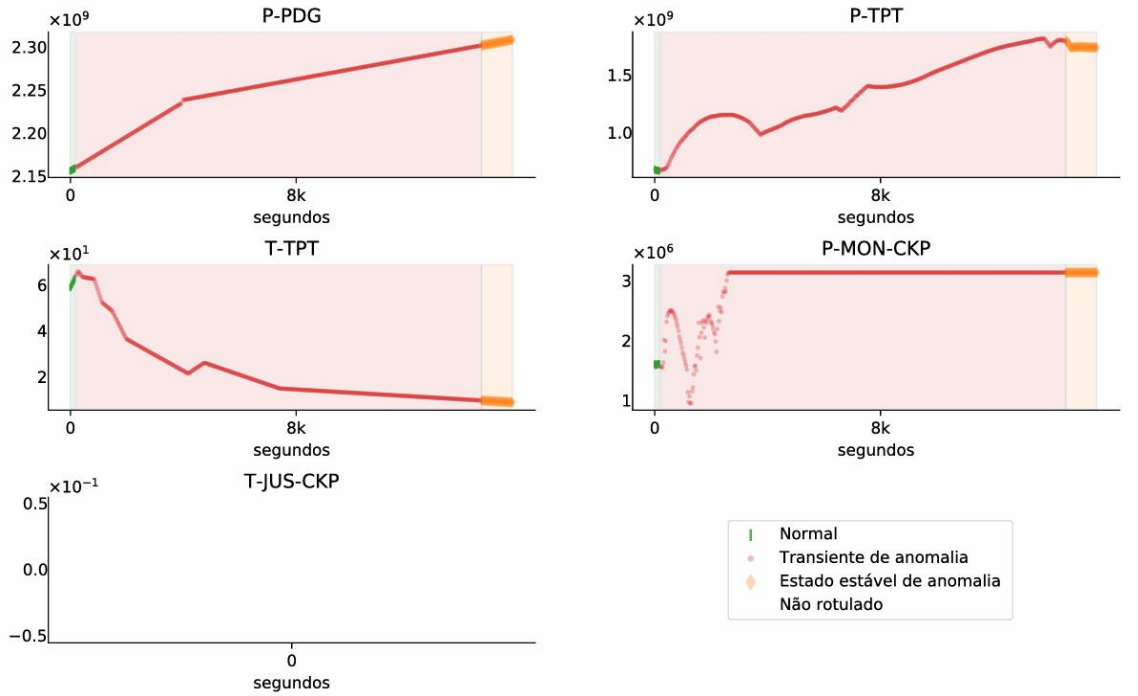
A ocorrência de hidratos na linha de produção pode ocasionar obstruções parciais ou totais de tubulações. Esses bloqueios nas tubulações não só possuem custos altos para desobstrução, mas também muitas vezes exigem a interrupção do sistema de produção, acarretando mais perdas de produção. Além disso, a formação de hidratos pode apresentar riscos associados ao aumento de pressão e temperatura, podendo acarretar rupturas nas tubulações (ANDREOLLI, 2016). A Figura 2.12 apresenta um exemplo de formação de hidrato em que ocorre obstrução total da tubulação em um poço da plataforma P-34 da Petrobras. A

Figura 2.13 ilustra um gráfico com as leituras dos sensores para um exemplo de ocorrência de hidrato em linha de produção.

Figura 2.12 - Hidrato em um poço da plataforma P-34 da Petrobras (banco de imagem da Petrobras

Fonte: Andreolli (2016)

Figura 2.13 - Variáveis de uma instância de incrustação em CKP



Fonte: Vargas (2019)

2.2 O BANCO DE DADOS

O 3W dataset, desenvolvido por Vargas et al. (2019) é um banco de dados que possui dados inerentes a poços de petróleo que operam offshore a partir da elevação natural. Ele é composto por séries temporais, capturadas por oito sensores diferentes. Essas séries podem ser classificadas de acordo com sua origem: reais, simuladas e desenhadas à mão. Os dados deste dataset podem representar o poço em diferentes estados de funcionamento, seja esse normal ou em algum dos oito eventos indesejáveis (VARGAS, 2019).

2.2.1 Estrutura Geral

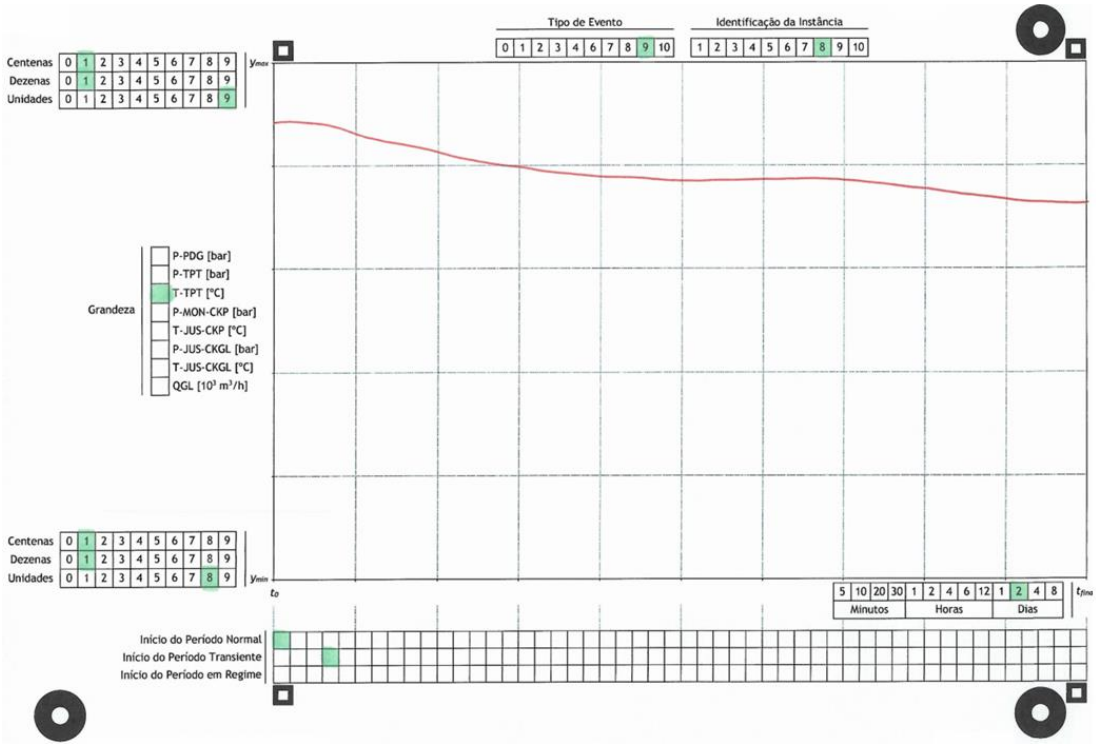
O dataset é composto por Séries Temporais Multivariadas (STM). Uma STM refere-se a uma coleção de séries temporais em que cada série temporal individual apresenta uma variável diferente, mas todas compartilham o mesmo índice de tempo. No caso do 3W dataset, temos que cada STM possui dez séries temporais univariadas compartilhando do mesmo índice de tempo, sendo essas: uma com o rótulo da observação; uma com o *timestamp* de captura do dado; e oito variáveis associadas aos sensores geralmente disponíveis. É importante ressaltar que a taxa de amostragem utilizada para gerar cada instância foi fixa e de 1 amostra por segundo (VARGAS, 2019).

As origens das instâncias do *dataset* podem ser três: reais, simuladas ou desenhadas à mão. As instâncias reais dizem respeito àquelas que realmente ocorreram no poços produtores de petróleo da Petrobras. Por se tratarem de dados extraídos sem pré-processamento, aspectos como dados faltantes ou congelados foram mantidos. Já as instâncias simuladas foram obtidas a partir da utilização de um simulador dinâmico monofásico, o OLGA (SCHLUMBERGER, 2019). Esse simulador é utilizado por diversas indústrias de petróleo ao redor do globo, incluindo a Petrobras (VARGAS, 2019).

Por fim, as instâncias desenhadas à mão são aquelas que foram construídas por especialistas da área, através de uma ferramenta desenvolvida para esse propósito, composta de um script para processamento de imagem e um modelo gráfico. Foram priorizados dois tipos raros de anomalias reais. A figura 2.14 apresenta um exemplo de utilização do modelo gráfico, com os diversos campos do modelo preenchidos por um especialista na área. (VARGAS, 2019).

O principal objetivo da utilização de instâncias simuladas e desenhadas à mão é reduzir o desbalanceamento do conjunto de dados causado quando se utilizam apenas instâncias reais (VARGAS, 2019).

Figura 2.14 - Exemplo de preenchimento do modelo gráfico



Fonte: Vargas (2019)

Cada instância é rotulada de duas formas, uma referente ao nível de instância e a outra ao nível de observação. A primeira se trata de um código numérico associado à operação normal ou a algum tipo de anomalia existente em algum instante dessa instância. Dessa forma, temos que nenhuma instância tem mais de uma anomalia (VARGAS, 2019). O código número utilizado para operação normal é o 0, enquanto que para as demais anomalias, segue-se a convenção apresentada no capítulo 2.1.2, que vai de 1 à 8, cada uma representando uma anomalia distinta.

Essa rotulação é importante pois é com base nela que ocorre a organização dos arquivos referentes ao banco de dados. Cada instância está salva em um arquivo dedicado com extensão *Comma-Separated Values* (CSV). Essas instâncias estão organizadas em nove pastas distintas, cujo nome representa os códigos numéricos de cada tipo de anomalia.

A segunda rotulagem, aplicada ao nível de observação, separa qualquer instância de qualquer tipo em três períodos distintos: normal, transiente de anomalia e estado estável de anomalia. Essa rotulagem foi realizada com o propósito de possibilitar o prognóstico, já que os períodos intermediários podem ser interpretados como um período de pré-anomalia e aprendidos para predizer períodos estáveis de anomalia (VARGAS, 2019). Os códigos numéricos desse rótulo utilizam o seguinte padrão:

- Período Normal: utiliza-se o código 0;
- Período Transiente de Anomalia: utiliza-se o código “10x”, sendo x um número de 1 à 8, referente à respectiva anomalia;
- Período Estável de Anomalia: utiliza-se um número de 1 à 8, de acordo com a respectiva anomalia.

Assim, podemos resumir as variáveis das séries temporais presentes do 3W *dataset* conforme apresentado na tabela 2.1. Além disso, também presentes nessa tabela, estão as unidades adotadas para cada variável.

Tabela 2.1 - Variáveis das séries temporais presentes no 3W *dataset*

Variável	**Descrição**	**Unidade**
timestamp	Carimbo de data e tempo	-
P-PDG	Pressão do fluido em PDG	Pa
P-TPT	Pressão do fluido em TPT	Pa
T-TPT	Temperatura do fluido em TPT	ºC
P-MON-CKP	Pressão do fluido montante à válvula *choke* de produção	Pa
T-JUS-CKP	Temperatura do fluido jusante à válvula *choke* de produção	ºC
P-JUS-CKGL	Pressão do fluido jusante à válvula *choke* de produção	Pa
T-JUS-CKGL	Temperatura do fluido jusante à válvula *choke* de produção	ºC
QGL	Vazão de *Gas Lift*	m^3/s
class	Valor numérico que representa o estado de cada anomalia ao longo da série temporal	-

Fonte: Adaptado de Vargas et al. (2019)

2.2.2 Quantitativo de Dados

O 3W *dataset* possui um total de 1978 instâncias, sendo essas 594 referentes à operação normal e 1384 anomalias. Essas instâncias são dividas em 1019 reais, 939 simuladas

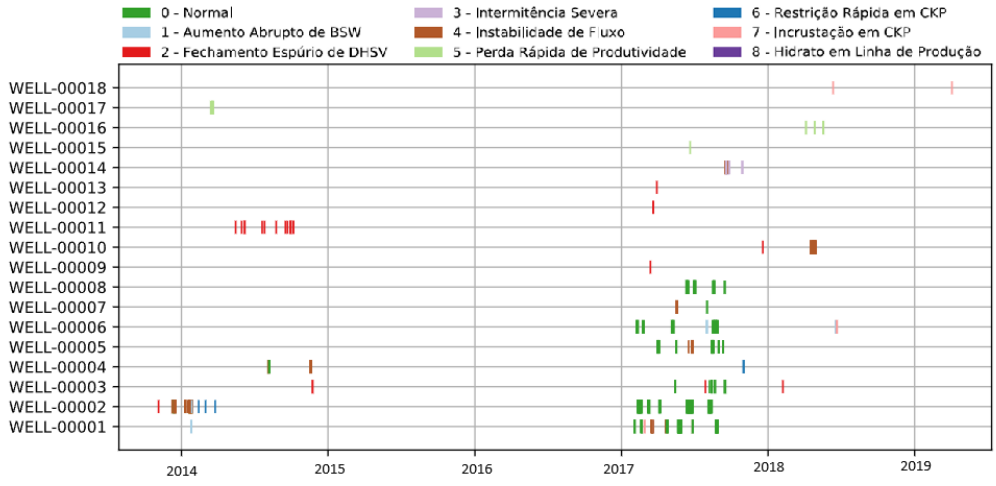
e 20 desenhadas. A tabela 2.2 apresenta uma visão resumida desses dados. Ao observarmos apenas as instâncias reais, temos que, considerando-se o limiar de 1%, temos quatro tipos de anomalias raras, as de código numérico 1, 6, 7 e 8. Porém, ao considerar também as instâncias simuladas, apenas a anomalia 7 pode ser considerada rara. Essa característica permanece mesmo se as instâncias desenhadas à mão forem consideradas (VARGAS, 2019).

Tabela 2.2 - Quantitativo de instâncias de cada tipo de evento e de cada fonte

Evento	Real	Simulado	Desenhado	Total
0 - Operação Normal	594	0	0	594
1 - Aumento Abrupto de BSW	5	114	10	129
2 - Fechamento Espúrio de DHSV	22	16	0	38
3 - Intermitência Severa	32	74	0	106
4 - Instabilidade de Fluxo	344	0	0	344
5 - Perda Rápida de Produtividade	11	439	0	450
6 - Restrição Rápida em CKP	6	215	0	221
7 - Incrustação em CKP	5	0	10	5
8 - Hidrato em Linha de Produção	0	81	0	81
Total	**1019**	**939**	**20**	**1978**

Fonte: Vargas et al. (2019)

Figura 2.15 - Mapa de dispersão das instâncias reais históricas do 3W dataset



Fonte: Adaptado de Vargas et al., 2019

A Figura 2.15 apresenta um mapa de dispersão das instâncias reais, no qual é possível ver o tipo de anomalia, de qual poço essa instância surgiu e em qual ano ela ocorreu. Já a tabela 2.3 apresenta dados referentes às dificuldades inerentes a dados reais. Pode-se perceber que 29,45% das variáveis estão faltando e 9,80% estão congeladas. À medida que se aumenta o

número de variáveis faltantes, mais espaço se torna o *dataset*, e isso pode impor dificuldade adicional nos algoritmos de aprendizado de máquina. As variáveis congeladas também impõem dificuldades adicionais, já que elas não manifestam os padrões associados às anomalias.

Tabela 2.3 - Quantitativo de variáveis ausentes, congeladas ou não rotuladas de dados reais

Tipo	**Quantidade**	**Percentual**
Variáveis Faltantes	4.660	29,45% de 158.324
Variáveis Congeladas	1.550	9,80% de 158.324
Observações Não Rotuladas	635.440	1,24% de 51.385.410

Fonte: Vargas et al. (2019)

2.3 APRENDIZADO DE MÁQUINA

A Inteligência Artificial é uma área da ciência que permite que os computadores pensem e tomem decisões de forma autônoma. Ao unir a inteligência humana e o poder computacional, a inteligência artificial permite a produção de soluções inteligentes e confiáveis para problemas extremamente não lineares e altamente complexos. O Aprendizado de Máquina é uma subárea da Inteligência Artificial que se concentra no desenvolvimento de algoritmos e técnicas que permitem que os computadores aprendam a partir de dados e façam previsões ou tomem decisões sem serem explicitamente programados para isso (TARIQ et al, 2021).

O aprendizado em máquina pode ser classificado de diversas maneiras, tomando como base diferentes critérios Eles podem ser separados dependendo de como são supervisionados durante o treinamento, se conseguem aprender incrementalmente ou não, e se são baseados em instâncias ou modelos (GÉRON, 2019).

Primeiramente, quanto à supervisão durante o treinamento, temos três categorias principais: aprendizado supervisionado, não supervisionado e por reforço (TARIQ et al, 2021). O aprendizado supervisionado utiliza de um conjunto de dados rotulados onde o algoritmo aprende a mapear entradas para saídas desejadas. Um exemplo de utilização desse método é de um desenvolvimento de um filtro que classifica *e-mails* como *spam*. O algoritmo é treinado com e-mails que estão rotulados de acordo com sua classe (*spam* ou não) e, com base nisso, o algoritmo deve aprender a classificar novos *e-mails* (GÉRON, 2019).

Já o aprendizado não supervisionado lida com dados não rotulados e é frequentemente utilizado para identificar padrões ou agrupamentos de dados (TARIQ et al, 2021). Outra utilização comum é a detecção de anomalias, na qual o algoritmo aprende a

diferenciar uma instância normal de uma anormal (GÉRON, 2019). Por fim, o aprendizado por reforço se baseia em um sistema de recompensas e punições para treinar um agente a tomar decisões sequenciais, visando maximizar uma recompensa cumulativa ao longo do tempo (SUTTON; BARTO, 2018). No presente trabalho, serão utilizados apenas algoritmos de aprendizado não supervisionado, visto que o intuito é realizar a detecção de anomalias.

Quanto à capacidade de aprender incrementalmente, os algoritmos podem ser divididos em aprendizado em lote (*batch*) e aprendizado em tempo real (*online)*. O aprendizado em lote envolve treinar o modelo utilizando o conjunto de dados completo de uma única vez. Após treinado, esse modelo não aprende mais, ele apenas aplica o que aprendeu. Em contraste, o aprendizado *online* permite que o modelo seja atualizado incrementalmente à medida que novos dados chegam (GÉRON, 2019). Nesse contexto, para o presente trabalho, a categoria utilizada será a de aprendizado *batch*, visto que está sendo treinado um modelo com a quantidade total do banco de dados, e o mesmo não aumenta dinamicamente.

Finalmente, com base na abordagem de aprendizado, os algoritmos podem ser categorizados em baseados em instâncias ou em modelos. No caso dos baseados em instâncias, o sistema aprende com base nos exemplos de treinamento e faz previsões com base na similaridade dos exemplos e das novas entradas. Já no caso do aprendizado baseado em modelos, o algoritmo constroi um modelo a partir dos dados de treinamento e utiliza esse modelo para realizar as predições (GÉRON, 2019).

2.3.1 Desenvolvimento de um modelo de aprendizado de máquina

Um projeto de aprendizado de máquina possui diversas etapas fundamentais que garantem que o modelo final seja aplicável aos dados em questão. A primeira etapa se trata do entendimento do problema, que envolve duas partes: entender qual é o problema que queremos solucionar e qual parte do problema pode ser resolvida pelo aprendizado de máquina (RUSSEL; NORVIG, 2020). No caso do presente trabalho, o problema que se deseja buscar uma solução são as perdas de produção ocasionadas pelas anomalias que ocorrem durante a extração de petróleo, e a parte que pode ser resolvida por aprendizado de máquina é o prognóstico e detecção dessas anomalias.

Com o problema definido, é necessário realizar a coleta de dados. Essa etapa envolve obter um conjunto representativo de dados que será utilizado para treinar e avaliar o modelo. Após isso, é necessário realizar a preparação dos dados, onde os dados são limpos e pré-processados (RUSSEL; NORVIG, 2020). A limpeza e pré-processamento de dados é

necessária para garantir que os dados estejam no formato adequado para o treinamento do modelo. Nessa etapa, são tratados os valores ausentes, outliers e é realizada a normalização dos dados (HAN; KAMBER; PEI, 2011).

A próxima etapa é a seleção de um modelo apropriado e treinamento do mesmo. Nessa etapa escolhe-se o tipo de algoritmo que será utilizado e esse é ajustado para o conjunto de dados de treinamento. Também é realizada nessa etapa o ajuste de alguns parâmetros chamados de hiperparâmetros. Eles controlam diretamente o comportamento do processo de aprendizado e influenciam a performance do modelo (RUSSEL; NORVIG, 2020). Os valores dos hiperparâmetros não são ajustados automaticamente pelo algoritmo e necessitam de ajuste manual (GOODFELLOW; BENGIO; COURVILLE, 2016).

Após o treinamento, é realizada a etapa de validação. Ao testar o desempenho do modelo com um conjunto de dados teste é possível verificar o quão confiável é o modelo e se o mesmo pode ser utilizado para a última etapa, que é a de implantação e monitoramento (RUSSEL; NORVIG, 2020).

2.3.2 *Ensemble Learning*

Antes de entender sobre os algoritmos utilizados é necessário entender uma técnica utilizada em alguns algoritmos de aprendizado de máquina. O *ensemble learning* refere-se a técnica de selecionar uma coleção ou conjunto de modelos base e combinar suas previsões através de métodos como votação ou médias. Ao realizar isso é possível produzir um resultado mais robusto e exato que os modelos individuais (RUSSEL; NORVIG, 2020).

O principal objetivo do *ensemble* é reduzir a variabilidade e melhorar a capacidade preditiva, explorando diferentes modelos que podem compensar as falhas uns dos outros. Existem várias abordagens para criação de *ensembles,* sendo algumas delas: *bagging, random forests, stacking* e *boosting* (RUSSEL; NORVIG, 2020). Desses, serão detalhados apenas os dois primeiros, visto que são os que serão utilizados neste trabalho.

O *bagging* funciona a partir da geração de subconjuntos de dados de treinamento a partir do conjunto original, treinamento de um modelo em cada subconjunto e, em seguida, a combinação das previsões por meio de votação ou média (BREIMAN, 1996).

O *random forests* utiliza o *bagging* e um conjunto de Árvores de Decisão aleatórios criados durante o treinamento. Uma Árvore de Decisão é a representação de uma função que transforma um conjunto de atributos em uma única decisão. A decisão é realizada através de uma série de testes começando na raiz e seguindo os ramos apropriados até chegar a uma folha.

Cada nó interno da árvore corresponde a um teste de atributo de entrada, os ramos são rotulados com os possíveis valores desse atributo, e as folhas indicam o valor de saída final da função (RUSSEL; NORVIG, 2020).

2.3.3 Algoritmos Utilizados

Para o desenvolvimento do presente trabalho, foram utilizados alguns dos algoritmos experimentados por Vargas (2019) em seu *benchmark* inicial. São esses o *Isolation Forest*, *One Class SVM* e o *Dummy*.

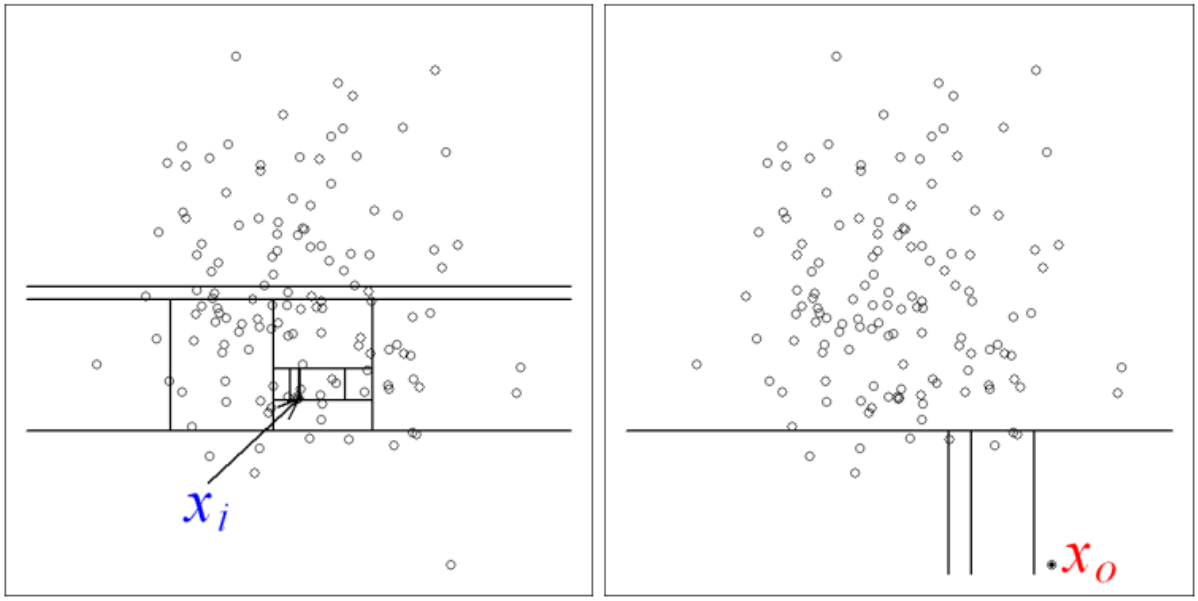
2.3.3.1 *Isolation Forest*

Desenvolvido por Liu, Ting e Zhou (2008), o *Isolation Forest* é um algoritmo não supervisionado criado para detecção de anomalias em um conjunto de dados. Esse algoritmo, baseado em modelo, utilizou de duas características das anomalias para seu desenvolvimento. A primeira é o fato de que as anomalias são a minoria e a segunda é a diferença significativa entre os padrões de dados de uma instância anômala e uma normal. Com essas duas características em mente, entende-se que as anomalias são poucas e diferentes, o que as torna mais suscetíveis ao isolamento que as instâncias normais (LIU; TING; ZHOU, 2008).

O *Isolation Forest* também é um modelo *ensemble* e funciona criando uma floresta de Árvores de Isolamento, onde cada árvore é construída através de partições aleatórias de dados. Os dados são repetidamente divididos de forma aleatória até que todas as instâncias estejam isoladas. As instâncias anômalas, por serem poucas e diferentes, necessitam de menos partições para serem isoladas que instâncias normais (LIU; TING; ZHOU, 2008).

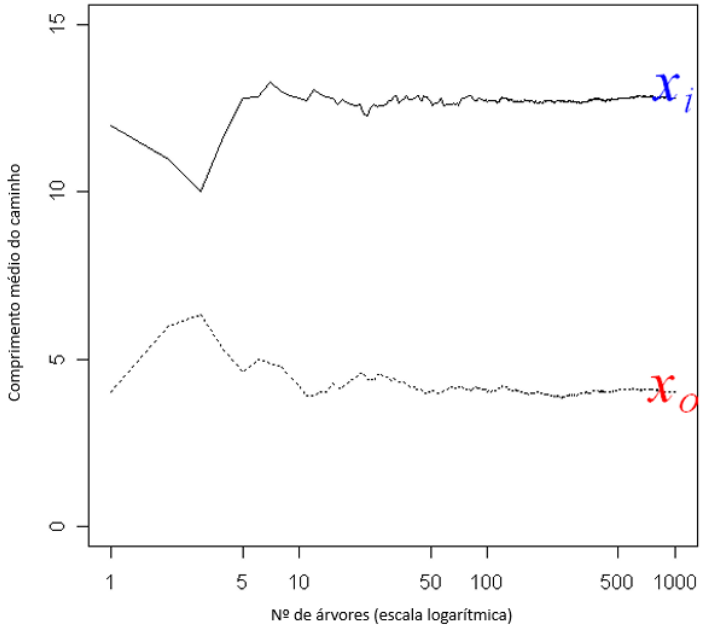
A Figura 2.16 apresenta um exemplo de aplicação do algoritmo para isolamento de uma instância normal (x_i) e uma anômala (x_o). Pode-se perceber, ao comparar a quantidade de partições aleatórias necessárias para isolar as instâncias, que o ponto normal necessitou de mais partições que o ponto anormal. No final, calcula-se o comprimento médio dos caminhos, conforme demonstrado na Figura 2.17. Nota-se, como esperado, que o comprimento médio de uma instância anômala é menor que a de uma normal.

Figura 2.16 - Exemplo de isolamento de uma instância normal (x_i) e uma anômala (x_o)



Fonte: Liu, Ting e Zhou (2008)

Figura 2.17 - Exemplo de comprimento médio uma instância normal (x_i) e uma anômala (x_o)



Fonte: Liu, Ting e Zhou (2008)

2.3.3.2 *One Class* SVM

One Class SVM é um algoritmo da família *Support Vector Machine* (SVM). O *One Class* SVM, ao contrário do SVM tradicional, é um algoritmo de treinamento não supervisionado que é utilizado para classificação binária, ou multiclasse. Ele é utilizado para

classificação de uma classe única (como por exemplo uma classe normal). Nesse caso, durante o treinamento o algoritmo busca aprender uma função e identificar se os novos exemplos pertencem a essa classe ou são anomalias, retornando valores positivos ou negativos, respectivamente (VARGAS, 2019), (SCHÖLKOPF et al., 2001).

2.3.3.3 *Dummy*

O algoritmo *Dummy* é um classificador simples utilizado como referência para comparar algoritmos mais complexos. Este algoritmo não tenta aprender padrões nos dados, mas sim utilizar estratégias triviais para realizar previsões. Existem diversas estratégias que podem ser aplicadas pelo algoritmo, tais como prever sempre a classe majoritária, selecionar classes ou valores aleatoriamente, ou utilizar médias e medianas (PEDREGOSA et al. 2011).

2.3.4 Métricas de Desempenho

Segundo (HAN; KAMBER; PEI, 2011), a fim de avaliar o desempenho e eficácia dos modelos de aprendizado de máquina são utilizadas as métricas de desempenho. Para entender as diferentes métricas de desempenho é necessário antes entender sobre a matriz de confusão. Essa matriz é uma ferramenta que visualiza o desempenho de um algoritmo de classificação, fornecendo uma visão detalhada sobre os tipos de erros e acertos que o modelo está cometendo. Ela é composta por quatro componentes principais:

- Verdadeiros Positivos (TP): Representa os casos em que a classe positiva foi corretamente identificada;
- Falsos Positivos (FP): Representa os casos em que a classe negativa foi classificada incorretamente como positiva;
- Verdadeiros Negativos (TN): Representa os casos em que a classe negativa foi corretamente identificada;
- Falsos Negativos (FN): Representa os casos em que a classe positiva foi classificada incorretamente como negativa.

A partir dos componentes da matriz de confusão é possível calcular algumas métricas para avaliar o modelo de aprendizado de máquina (HAN; KAMBER; PEI, 2011). Aqui serão descritos apenas sobre as métricas que foram utilizadas no trabalho: *Precision*, *Recall* e *F1 Score*.

O *precision* representa a proporção de verdadeiros positivos em relação ao total de exemplos classificados como positivos pelo modelo. Mede a exatidão das previsões positivas. Já o *recall*, representa a proporção de verdadeiros positivos em relação ao total de exemplos que realmente são positivos. Essa métrica mede a capacidade do modelo de identificar corretamente todos os exemplos positivos. Por fim, o *F1 Score* representa a média harmônica entre precisão e revocação, proporcionando uma única métrica que balanceia os dois aspectos. Quando esse indicador se aproxima de 1 pode-se concluir que o modelo tem um bom equilíbrio entre *precision* e *recall*. Essas métricas podem ser calculadas pelas equações 2.1, 2.2 e 2.3.

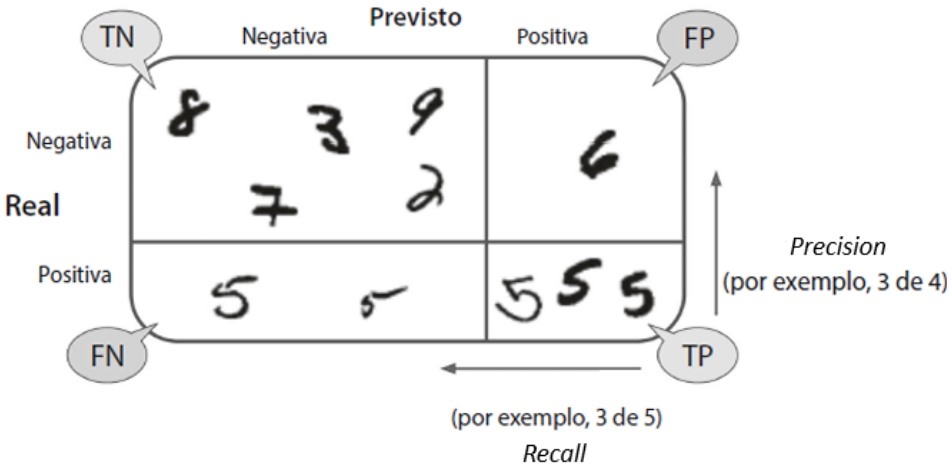
$$Precision = \frac{TP}{TP + FP} \tag{2.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.2}$$

$$F1\ Score = \frac{2 \times precision \times recall}{precision + recall} \tag{2.3}$$

A Figura 2.18 apresenta um exemplo de matriz de confusão para um algoritmo que busca classificar se um número manuscrito é o número 5 ou não.

Figura 2.18 - Exemplo de matriz de confusão para um algoritmo que classifica se o número manuscrito é o número 5 ou não



Fonte: Adaptado de Géron (2019)

3 METODOLOGIA

Para reprodução dos resultados obtidos inicialmente por VARGAS (2019), utilizou-se como base o código em Python disponibilizado pelo mesmo no *Github*. O código inicial foi desenvolvido em 2019 e, devido a atualizações de algumas bibliotecas que são utilizadas no

código, o mesmo precisou ser adaptado para que fosse compatível com as versões atuais das bibliotecas. Nesse código, são utilizados os algoritmos de *Isolation Forest*, *One Class* SVM com os quatros tipos de *kernel* e o *Dummy*.

3.1 TRATAMENTO DOS DADOS

A etapa de tratamento dos dados constituiu da seleção de alguns parâmetros que foram tomados como base para seleção dos dados que seriam utilizados para treinamento e validação do algoritmo de aprendizado de máquina. Além disso, foram consideradas algumas regras propostas por VARGAS (2019) em seu *benchmark* para detecção de anomalias.

A primeira regra considerada é o fato de que seriam utilizadas apenas instâncias reais com algum tipo de anomalia que possuem período normal maiores ou iguais a vinte minutos, ou seja, como a taxa de observação dos dados é de 1 amostra por segundo, as instâncias devem ter no mínimo 1200 observações no período normal. Assim, foram descartadas as anomalias com código numérico 3 e 4, visto que essas possuem apenas período estável de anomalia (VARGAS, 2019).

A segunda regra proposta por VARGAS (2019) é que devem ser realizadas múltiplas rodadas de treinamento e teste, sendo o número de rodadas igual ao número de instâncias. Em cada rodada, as amostras que foram utilizadas para o treinamento ou teste foram extraídas de um única instância.

A última regra que VARGAS (2019) propõe é que em cada rodada são calculadas as métricas *Precision*, *Recall* e F1 *Score*. Sendo que o valor médio do F1 *Score* deve ser considerado a principal métrica de desempenho.

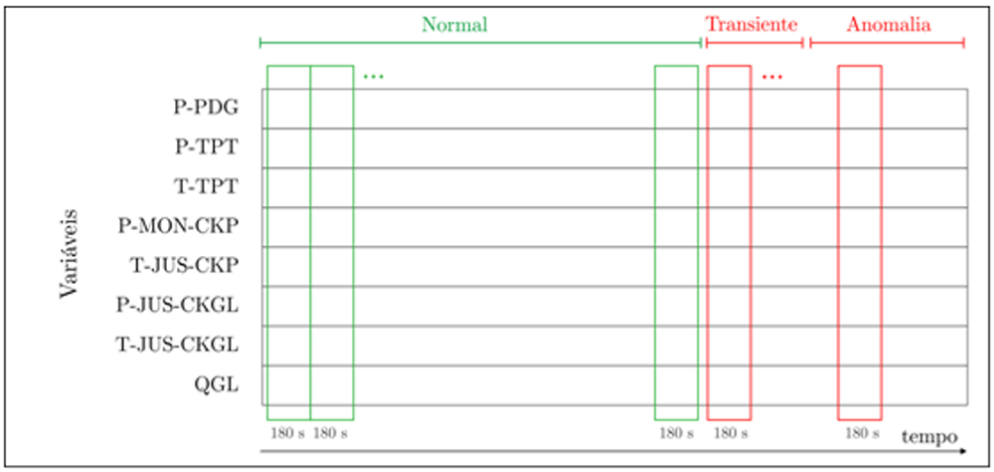
Além das regras consideradas, foram determinados alguns parâmetros a serem seguidos, também sugeridos por VARGAS (2019) para escolha dos dados que seriam utilizados. Foi utilizada uma estratégia específica de amostragem com janela deslizante com até 15 amostras com 180 observações cada. As observações de períodos normais foram divididas e as as primeiras observações foram utilizadas para treinamento (60%), enquanto as últimas para validação (40%). As observações provenientes de períodos anômalos, sejam esses de períodos transientes ou estáveis de anomalia, foram utilizadas apenas para validação.

Ademais, foram descartadas variáveis das amostras utilizadas que não possuíam desvios padrões mínimos de 1% ou possuíam um percentual de valores ausentes maior que 10%. Então, através da biblioteca *tsfresh*, foram calculadas as métricas: mediana, média, desvio padrão, variância, máximo e mínimo para cada variável. Além disso, antes de ser realizado o

treinamento, foi realizada a normalização das variáveis das amostras a partir da média e desvio padrão dos dados de treinamento. Os dados de teste também foram normalizados, tomando como base a média e desvio padrão obtidos dos dados de treinamento.

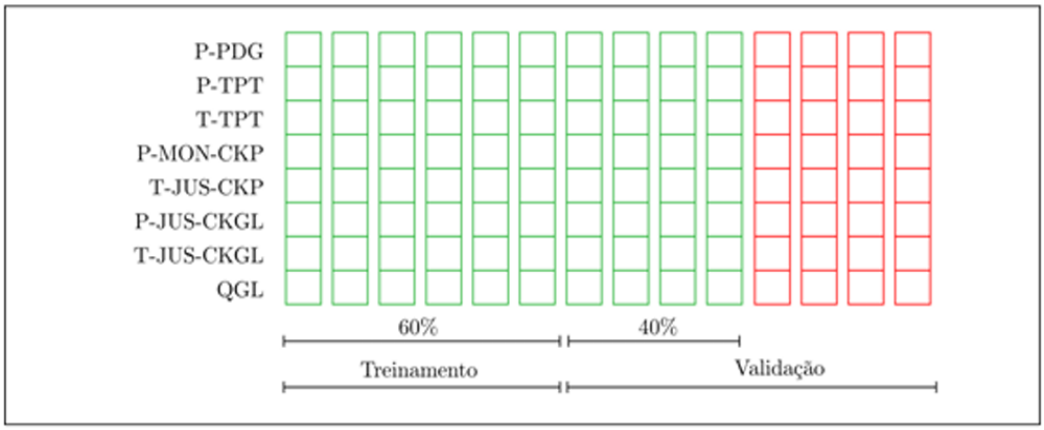
A Figura 3.1 ilustra a estratégia de amostragem de janela deslizante, enquanto a figura 3.2 ilustra como foram divididas as amostragens em treinamento e validação.

Figura 3.1 - Estratégia de amostragem com janelas deslizantes de 180 observações



Fonte: Adaptado de Junior (2022)

Figura 3.2 - Divisão das amostragens em dados de treinamento e dados de teste



Fonte: Adaptado de Junior (2022)

3.2 TREINAMENTO DOS MODELOS E AVALIAÇÃO DOS RESULTADOS

O processo de treinamento dos modelos segue um fluxo que se repete até que sejam obtidas as três métricas finais (*Precision*, *Recall* e F1 *Score*) utilizadas para avaliação do desempenho do modelo.

As etapas seguidas para cada uma das instâncias são: a extração das amostras; filtragem das variáveis; normalização dos dados; extração de características; treinamento do modelo; e o cálculo das métricas de performance. Assim, após obtidas as métricas de performance para cada instância, são calculadas as médias para obter um valor final. Esse processo é realizado para cada um dos algoritmos de aprendizado de máquina utilizados.

Conforme indicado por VARGAS (2019), a principal métrica utilizada para avaliar a performance dos modelos deve ser o F1 *Score*, então utilizou-se o mesmo como principal avaliação do desempenho do modelo treinado.

3.3 SELEÇÃO DOS HIPERPARÂMETROS

Após a obtenção dos resultados iniciais, foram testadas algumas variações de hiperparâmetros dos algoritmos, a fim de obter uma combinação de hiperparâmetros que resultasse em um desempenho (F1 *Score*) superior que os demais. Cada algoritmo utilizado apresenta diferentes hiperparâmetros que podem ser selecionados, conforme representado na Tabela 3.1.

Tabela 3.1 - Hiperparâmetros de cada algoritmo

Algoritmo	**Hiperparâmetros**
Isolation Forest	‘n_estimators’: [50, 100, 150, 200], ‘max_samples’: [‘auto’, 0.5, 0.75, 1.0], ‘contamination’:[‘auto’, 0, 0.05, 0.10], ‘bootstrap’: [True, False], ‘max_features’: [0.5, 0,75, 1.0]
One-Class SVM	‘kernel’: [‘linear’, ‘rbf’, ‘poly’, ‘sigmoid’], ‘gamma’: [‘auto’, ‘scale’, 1e-4, 1e-3, 1e-2, 0.1, 0.5, 1.0, 5.0, 10.0], ‘nu’: [1e-4, 1e-3, 1e-2, 0.10, 0.50, 1.0]

Fonte: Autoria própria

Tabela 3.2 - Combinação de hiperparâmetros utilizada por JUNIOR (2022)

Algoritmo	**Hiperparâmetros**
Isolation Forest	‘n_estimators’: [200], ‘max_samples’: [0.75], ‘contamination’:[0.10], ‘bootstrap’: [False], ‘max_features’: [1.0]
One-Class SVM	‘kernel’: [‘sigmoid’], ‘gamma’: [1e-2],

‘nu’: [1e-3]

Fonte: Adaptado de Junior (2022)

Segundo JUNIOR (2022), a combinação ideal ideal para cada um desses algoritmos é a combinação apresentada na Tabela 3.2. Foram realizados treinamentos sem a seleção de hiperparâmetros e posteriormente foram utilizados os mesmos hiperparâmetros definidos por JUNIOR (2022) para verificar se o desempenho (F1 *Score*) dos modelos se alterava positivamente ou não com relação aos resultados sem seleção de hiperparâmetros.

4 RESULTADOS E DISCUSSÃO

4.1 RESULTADOS OBTIDOS

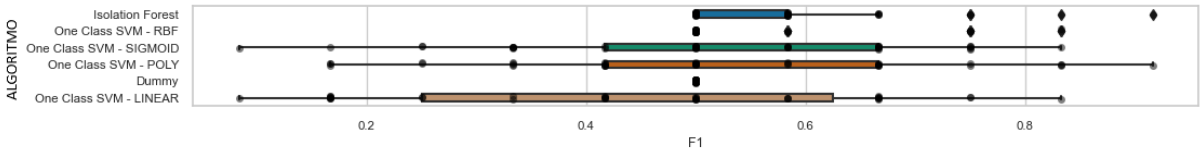
Os resultados obtidos estão compilados nas tabelas 4.1 e 4.2. Conforme estabelecido no capítulo 3.1, foram realizadas 38 rodadas de treinamento e validação, ou seja, o mesmo número de instâncias reais do tipo de anomalia considerado. Na tabela 4.1 estão apresentados os resultados obtidos sem definição de nenhum hiperparâmetro. Devido ao uso do parâmetro *micro* para o cálculo das métricas, foram obtidos valores iguais para as três métricas, F1 *Score*, *Precision* e *Recall* (SCIKIT-LEARN, 2019). Assim, optou-se por apresentar nas tabelas de resultados somente o F1 *Score* e o desvio padrão. A Figura 4.1 apresenta a dispersão dos valores do F1 *Score* para os valores da tabela 4.1.

Tabela 4.1 - F1 *Score* e Desvio Padrão para diferentes algoritmos

Algoritmo	**F1 Score**	**Desvio Padrão**
Isolation Forest	0,574	0,107
One Class SVM - RBF	0,543	0,098
One Class SVM - SIGMOID	0,521	0,179
One Class SVM - POLY	0,513	0,196
Dummy	0,500	0,000
One Class SVM - LINEAR	0,445	0,209

Fonte: Autoria própria

Figura 4.1 - Dispersão do F1 *Score* para diferentes algoritmos



Fonte: Autoria própria

A partir dos resultados apresentados na tabela 4.1 pode-se constatar que o algoritmo Isolation Forest apresentou F1 Score com melhor medida, com um valor de 0,574. Em seguida, o *One Class* SVM com o *kernel* RBF, com F1 Score de 0,543. Dos algoritmos testados, apenas um ficou abaixo do algoritmo de controle, o *Dummy*, obtendo um resultado de 0,445 para o F1 *Score*.

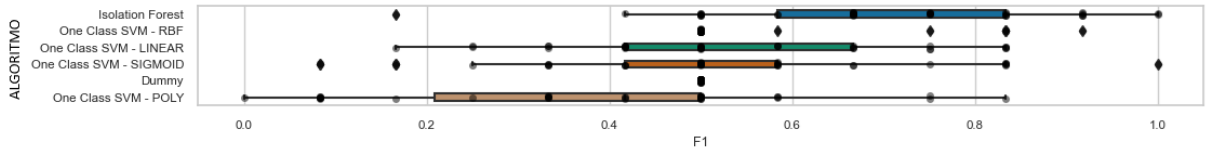
Os resultados apresentados na tabela 4.1 possuem os resultados obtidos para a seleção dos hiperparâmetros definidos na tabela 3.2. A Figura 4.2 ilustra a dispersão dos valores do F1 Score para os valores da tabela 4.1.

Tabela 4.2 - F1 *Score* e Desvio Padrão para diferentes algoritmos (com hiperparâmetros)

Algoritmo	**F1 Score**	**Desvio Padrão**
Isolation Forest	0,697	0,180
One Class SVM - RBF	0,546	0,115
One Class SVM - LINEAR	0,533	0,178
One Class SVM - SIGMOID	0,515	0,215
Dummy	0,500	0,000
One Class SVM - POLY	0,385	0,216

Fonte: Autoria própria

Figura 4.2 - Dispersão do F1 *Score* para diferentes algoritmos com seleção de hiperparâmetros



Fonte: Autoria própria

Ao verificar os resultados da tabela 4.2, pode-se constatar que os valores de F1 *Score* se alteraram com relação aos resultados da tabela 4.1. Alguns algoritmos, como o *Isolation Forest*, obtiveram uma melhora significativa, com um aumento no F1 *Score* de 21%. Porém outros, como o *One Class* SVM, com o *kernel* POLY, obteve uma redução de 25% no F1 *Score*. As diferenças percentuais entre os resultados da tabela 4.1 e 4.2 estão apresentados na tabela 4.3.

Tabela 4.3 - Comparação de F1 *Score* e Desvio Padrão

Algoritmo	**F1 Score (Sem seleção de hiperparâmetros)**	**F1 Score (Com seleção de hiperparâmetros)**	**Diferença (%)**
Isolation Forest	0,574	0,697	21%
One Class SVM - RBF	0,543	0,546	1%
One Class SVM - SIGMOID	0,521	0,515	-1%
One Class SVM - POLY	0,513	0,385	-25%
Dummy	0,500	0,500	0%
One Class SVM - LINEAR	0,445	0,533	20%

Fonte: Autoria própria

4.2 COMPARAÇÃO DOS RESULTADOS

Conforme explicado anteriormente, algumas bibliotecas utilizadas no código base foram atualizadas, e isso fez com que o código precisasse ser adaptado para que o mesmo fosse compatível com as versões atuais. Algumas bibliotecas de Python utilizadas também tiveram atualizações em seu código, como o Scikit-Learn. Além disso, o banco de dados utilizado, o 3W dataset foi atualizado e possui algumas diferenças quando comparado a versão utilizada por VARGAS (2019) em seu benchmark inicial. Nesse contexto, apesar de terem sido utilizados os mesmos procedimentos para o desenvolvimento do modelo de aprendizado de máquina, já se espera que os resultados finais se diferenciem um pouco dos resultados da literatura que utilizam o mesmo banco de dados de VARGAS (2019).

A principal diferença da versão do banco de dados utilizada por VARGAS (2019) e a utilizada neste trabalho está na quantidade de instâncias reais presentes. A tabela 4.4 apresenta o comparativo de quantitativo de dados entre a versão do banco de dados utilizada por VARGAS (2019) e a versão utilizada nesse trabalho.

Essa variação nas instâncias reais também gera uma alteração no quantitativo de variáveis faltantes, congeladas e de observações não rotuladas, como observado nas tabelas 4.5 e 4.6.

Tabela 4.4 - Comparativo de instâncias reais para as versões do banco de dados

Evento	Versão de Vargas (2019)	Versão Utilizada
0 - Operação Normal	597	594
1 - Aumento Abrupto de BSW	5	5
2 - Fechamento Espúrio de DHSV	22	22
3 - Intermitência Severa	32	32
4 - Instabilidade de Fluxo	344	344
5 - Perda Rápida de Produtividade	12	11
6 - Restrição Rápida em CKP	6	6
7 - Incrustação em CKP	4	5
8 - Hidrato em Linha de Produção	3	0
Total	**1025**	**1019**

Fonte: Autoria própria

Tabela 4.5 - Comparativo do quantitativo de dados faltantes, congelados e não rotulados para as diferentes versões do banco de dados

Tipo	Versão de Vargas (2019)	Versão Utilizada
Variáveis Faltantes	4.947	4.660
Variáveis Congeladas	1.535	1.550
Observações Não Rotuladas	5.130	635.440

Fonte: Autoria própria

Tabela 4.6 - Comparativo do percentual das variáveis faltantes, congeladas e observações não rotuladas entre as diferentes versões do banco de dados

Tipo	Versão de Vargas (2019)	Versão Utilizada
Variáveis Faltantes	31,17% de 15.872	29,45% de 15.824
Variáveis Congeladas	9,67% de 15.872	9,80% de 15.824
Observações Não Rotuladas	0,01% de 50.913.215	1,24% de 51.385.410

Fonte: Autoria própria

Nesse contexto, comparou-se os resultados obtidos aos do benchmark inicial de VARGAS (2019) e aos obtidos por JUNIOR (2022). A tabela 4.7 apresenta o comparativo dos valores do F1 Score bem como os do desvio padrão (em parênteses). Optou-se por comparar

nessa tabela os melhores valores de F1 Score obtidos, tomando como referência a tabela 4.3 para decisão dos valores.

Tabela 4.7 - Comparativo do F1 *Score* e Desvio Padrão (em parênteses) para diferentes algoritmos entre VARGAS (2019), JUNIOR (2022) e o autor do texto

Algoritmo	**F1 Score - VARGAS (2019)**	**F1 Score - JUNIOR (2022)**	**F1 Score - Autor do texto**
Isolation Forest	0,727 (0,182)	0,701 (0,176)	0,697 (0,180)
One Class SVM - RBF	0,532 (0,075)	-	0,546 (0,115)
One Class SVM - SIGMOID	0,470 (0,201)	0,477 (0,221)	0,521 (0,179)
One Class SVM - POLY	0,472 (0,188)	-	0,513 (0,196)
Dummy	0,500 (0,000)	-	0,500 (0,000)
One Class SVM - LINEAR	0,414 (0,195)	-	0,533 (0,178)

Fonte: Autoria própria

Ao comparar os valores obtidos para o algoritmo *Isolation Forest* pode-se perceber que o valor encontrado foi, apesar de próximo, inferior ao obtido por VARGAS (2019), porém está bem próximo do obtido por JUNIOR (2022). Em termos de desvio padrão, o valor obtido está próximo aos valores comparados.

Para o *One Class* SVM, os valores obtidos de F1 *Score* foram maiores que os obtidos por VARGAS (2019) em seu *benchmark* inicial. O desvio padrão sofreu aumento então para o *kernel* KBF quanto para o POLY, sendo que para o primeiro houve um aumento de 53% e para o segundo 4%. Para o *kernel* LINEAR e SIGMOID, o desvio padrão sofreu redução quando comparado ao benchmark de VARGAS (2019). Quando comparado o resultado de JUNIOR (2022) para o *kernel* SIGMOID, nota-se que tanto o F1 *Score* quanto o desvio padrão obtido foram melhores.

5 CONCLUSÃO

No presente trabalho, buscou-se aplicar algoritmos de aprendizado de máquina para desenvolver modelos capazes de realizar a detecção de anomalias que ocorrem durante a extração de petróleo em poços de petróleo *offshore* que operam por elevação natural. Para isso utilizou-se do banco de dados público 3W *Dataset*, bem como o código base elaborado por VARGAS (2019).

Para o desenvolvimento dos modelos, foi seguida a metodologia desenvolvida por VARGAS (2019) para *benchmark* dos algoritmos. Foram utilizados os algoritmos *Isolation Forest*, *One Class* SVM e o *Dummy*. Para avaliação de desempenho utilizou-se o F1 *Score* como métrica, bem como o desvio padrão. Essas métricas foram comparadas entre si e com os resultados obtidos na literatura por VARGAS (2019) e JUNIOR (2022).

Foram realizados treinamentos com e sem seleção de hiperparâmetros para os algoritmos. Os resultados obtidos através dos experimentos obtiveram, em sua maior parte, resultados superiores aos obtidos no *benchmark* inicial de VARGAS (2019) e obtidos por JUNIOR (2022). O único algoritmo utilizado que obteve resultado inferior aos obtidos por VARGAS (2019) e JUNIOR (2022) foi o *Isolation Forest*, que apresentou um F1 *Score* de 0,679 e desvio padrão de 0,180. No entanto, este algoritmo ainda apresentou desempenho significativamente maior que os demais testados, sendo que o segundo melhor desempenho foi o do *One Class* SVM - RBF, com F1 Score de 0,546.

Além disso, todos os algoritmos testados apresentaram um F1 *Score* superior ao do *Dummy*, que foi utilizado como referência para avaliar os desempenho. Isso não ocorreu no *benchmark* inicial de VARGAS (2019), e até mesmo o algoritmo que havia apresentado o menor valor de F1 *Score* para o autor, o *One Class* SVM - LINEAR, com F1 Score de 0,414, obteve, neste trabalho 0,533.

Tendo em vista a diferença dos resultados obtidos neste trabalho aos da literatura, é importante notar que essas diferenças ocorreram devido a diferenças nos bancos de dados utilizados e devido a atualizações nas bibliotecas de Python utilizadas. A versão do banco de dados utilizado apresenta diferenças significativas para o quantitativo de instâncias reais, dados faltantes, congelados e de observações não rotuladas.

Neste trabalho, foram utilizados apenas os mesmos algoritmos que os utilizados por VARGAS (2019) em seu *benchmark* inicial. Como sugestão de trabalhos futuros, sugere-se a exploração mais aprofundada do 3W *dataset*, que está em constante evolução, para avaliar o desempenho de outros algoritmos. Além disso, sugere-se comparar o desempenho de

algoritmos para as diferentes versões do 3W *dataset*, avaliando como as diferenças do quantitativo de instâncias reais, dados faltantes congelados e observações não rotuladas impactam o desempenho dos algoritmos.

REFERÊNCIAS

ANDREOLLI, I. **Introdução à elevação e escoamento monofásico e multifásico de petróleo.** Rio de Janeiro: Interciência, 2016. *E-book.* Disponível em: https://plataforma.bvirtual.com.br. Acesso em: 13 maio 2024.

AGÊNCIA NACIONAL DO PETRÓLEO, GÁS NATURAL E BIOCOMBUSTÍVEIS. **Anuário Estatístico Brasileiro do Petróleo, Gás Natural e Biocombustíveis 2023**. Brasília: ANP 2024. Disponível em: https://www.gov.br/anp/pt-br/centrais-de-conteudo/publicacoes/anuario-estatistico/anuario-estatistico-2023#Seção%201. Acesso em: 30 jun. 2024.

AGÊNCIA NACIONAL DO PETRÓLEO, GÁS NATURAL E BIOCOMBUSTÍVEIS. **Boletim mensal da produção de petróleo e gás natural**. Brasília: ANP, 2023. Disponível em: http://www.anp.gov.br/. Acesso em: 29 out. 2023.

AGÊNCIA NACIONAL DO PETRÓLEO, GÁS NATURAL E BIOCOMBUSTÍVEIS. **Boletim mensal da produção de petróleo e gás natural**. Brasília: ANP, 2024. Disponível em: http://www.anp.gov.br/. Acesso em: 30 maio 2024.

BREIMAN, L. Bagging predictors. **Machine learning**, Boston,v. 24, n. 2, p. 123-140, 1996.

CORTES, C.; VAPNIK, V. N. Support-vector networks. **Machine Learning**, Boston, v. 20, n. 3, p. 273-297, 1995.

FERNANDES JUNIOR, W. **Comparação de classificadores para detecção de anomalias em poços produtores de petróleo**. 2022. 64 f. Dissertação (Mestrado em Computação Aplicada) - Instituto Federal do Espírito Santo, Serra, 2022. Disponível em: https://github.com/ricardovvargas/3w_dataset/blob/master/docs/master_degree_dissertation_wander_junior.pdf. Acesso em: 29 out. 2023.

GAUTO, M. A. **Petróleo e gás**: princípios de exploração, produção e refino. Porto Alegre: Bookman, 2016.

GÉRON, A. **Hands-on machine learning with scikit-Learn, keras, and tensorflow**: Concepts, tools, and techniques to build intelligent systems. Sebastopol: O'Reilly Media, 2019.

GITHUB. **The first realistic and public dataset with rare undesirable real events in oil wells.** [California], 2019. Disponível em: https://github.com/ricardovvargas/3w_dataset. Acesso em: 29 out. 2023.

GOODFELLOW, I; BENGIO, Y; COURVILLE, A. **Deep learning**. Cambridge: MIT Press, 2016. Disponível em em: https://www.deeplearningbook.org/. Acesso em: 30 maio 2024.

HAN, J; KAMBER, M; PEI, J. **Data mining**: concepts and techniques. 3. ed. Burlington: Morgan Kaufmann, 2011.

HAWKINS, D. **Identification of outliers**. Netherlands: Springer, 1980. (Monographs on Statistics and Applied Probability).

LIU, F. T.; TING, K. M.; ZHOU, Z. H. Isolation forest. *In:* Proceedings- IEEE INTERNATIONAL CONFERENCE ON DATA MINING, 8, 2008, Pisa. **Proceedings** [...]. Pisa: IEEE, 2018. Disponível em: https://www.researchgate.net/publication/224384174_Isolation_Forest. Acesso em: 11 maio 2024.

MACROTRENDS. **Brent crude oil prices - 10 ýear daily chart**. 2024. Disponível em: https://www.macrotrends.net/. Acesso em: 30 maio 2024.

MEGLIO, F. *et al.* Stabilization of slugging in oil production facilities with or without upstream pressure sensors. **Journal of process control**, United Kingdom, v. 22, n. 4, p. 809–822, 2012. Disponível em: https://www.researchgate.net/publication/257407466_Stabilization_of_slugging_in_oil_production_facilities_with_or_without_upstream_pressure_sensors. Acesso em: 11 maio 2024.

MORAIS, J. M. **Petróleo em águas profundas**: uma história tecnológica da PETROBRAS na exploração e produção offshore. Brasília: Instituto de Pesquisa Econômica Aplicada, 2013. Disponível em: https://repositorio.ipea.gov.br/handle/11058/1147. Acesso em: 13 maio 2024.

OLGA Dynamic Multiphase Flow Simulator. **Schlumberger web site**, Houston, 2019. Disponível em: https://www.software.slb.com/products/olga. Acesso em: 09 jun. 2024.

ORTIZ NETO, J. B.; COSTA, A. J. D. A Petrobrás e a exploração de petróleo offshore no Brasil: um approach evolucionário. **Revista Brasileira de Economia**, Rio de Janeiro, v. 61, n. 1, 2007. Disponível em: https://www.scielo.br/j/rbe/a/bbJ3zjwJBFyhkthrtMQrvbF/. Acesso em 30 maio 2024.

PEDREGOSA, F. *et al*. Scikit-learn: machine learning in python**. Journal of machine learning research,** Cambridge, v. 12, p. 2825-2830, 2011. Disponível em: https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf?ref=https://githubhelp.com. Acesso em: 30 maio 2024.

RICCOMINI C.; SANT'ANNA, L. G.; TASSINARI, C. C. G. Pré-sal: geologia e exploração. **Revista USP**, São Paulo, 95:33–42, 2012. Disponível em: https://www.revistas.usp.br/revusp/article/view/52236 . Acesso em: 13 maio 2024.

RUSSELL, S.; NORVIG, P. **Artificial intelligence**: a modern approach**.** 4. ed. Hoboken: Pearson, 2020.

SCHÖLKOPF, B. *et al*. Estimating the support of a high-dimensional distribution. **Neural Computation,** Cambridge, v. 13, n. 7, p. 1443-1471, 2001. Disponível em: https://www.researchgate.net/publication/220499623_Estimating_Support_of_a_High-Dimensional_Distribution. Acesso em: 09 jun. 2024.

SCOPUS. **Analyze search results**. [Netherlands], 2024. Disponível em: https://www.scopus.com/term/analyzer.uri?sort=plf-f&src=s&sid=06b0fb7c8b1333428c02ab759233cb42&sot=a&sdt=a&sl=68&s=%28TITLE-ABS-KEY%28"oil+and+gas"%29+AND+TITLE-ABS-KEY%28"machine+learning"%29%29&origin=resultslist&count=10&analyzeResults=Analyze+results. Acesso em: 30 jun. 2024.

SOUZA, L.S.; SGARBI, G. N. C. Bacia de santos no brasil: geologia, exploração e produção de petróleo e gás natural. **Boletín de Geología**, Bucaramanga, 2019. Disponível em: http://www.scielo.org.co/scielo.php?pid=S0120-02832019000100175&script=sci_arttext. Acesso em: 10 jun 2024.

SUTTON, R. S.; BARTO, A. G. **Reinforcement learning**: an introduction. 2. ed. Cambridge: MIT Press, 2018. Disponível em: https://web.stanford.edu/class/psych209/Readings/SuttonBartoIPRLBook2ndEd.pdf. Acesso em 17 jun. 2024.

TARIQ, Z. *et al.* A systematic review of data science and machine learning applications to the oil and gas industry. **Journal of petroleum exploration production technology**, Germany, v. 11, 2021 Disponível em: https://link.springer.com/article/10.1007/s13202-021-01302-2 Acesso em: 16 jun. 2024.

THEYAB, M. Severe Slugging Control: Simulation of Real Case Study**. Journal of environmental research**, California, v. 2, n. 1, 2018. Disponível em: https://www.researchgate.net/publication/324248681_Severe_Slugging_Control_Simulation_of_Real_Case_Study Acesso em: 15 jun 2024.

THOMAS, J. E. **Fundamentos de engenharia de petróleo.** 2. ed. Rio de Janeiro: Editora Interciência, 2004.

VARGAS, R. E. V. **Base de dados e benchmarks para prognóstico de anomalias em sistemas de elevação de petróleo**. 2019. Tese (Doutorado em Engenharia Elétrica - Automação) - Universidade Federal do Espírito Santo, Vitória, 2019. Disponível em: https://github.com/ricardovvargas/3w_dataset/blob/master/docs/doctoral_thesis_ricardo_vargas.pdf. Acesso em: 29 out. 2023.

VARGAS, R. E. V. *et al.* A realistic and public dataset with rare undesirable real events in oil wells. **Journal of petroleum science and engineering**, Netherlands, v. 181, p. 106223, 2019. Disponível em: https://www.sciencedirect.com/science/article/pii/S0920410519306357 Acesso em 29 out. 2023.

APÊNDICE A – CÓDIGO UTILIZADO

Devido ao número elevado de linhas do código, optou-se por disponibilizar o código em: <https://github.com/ianryuji/TCC_Ian_RMR>