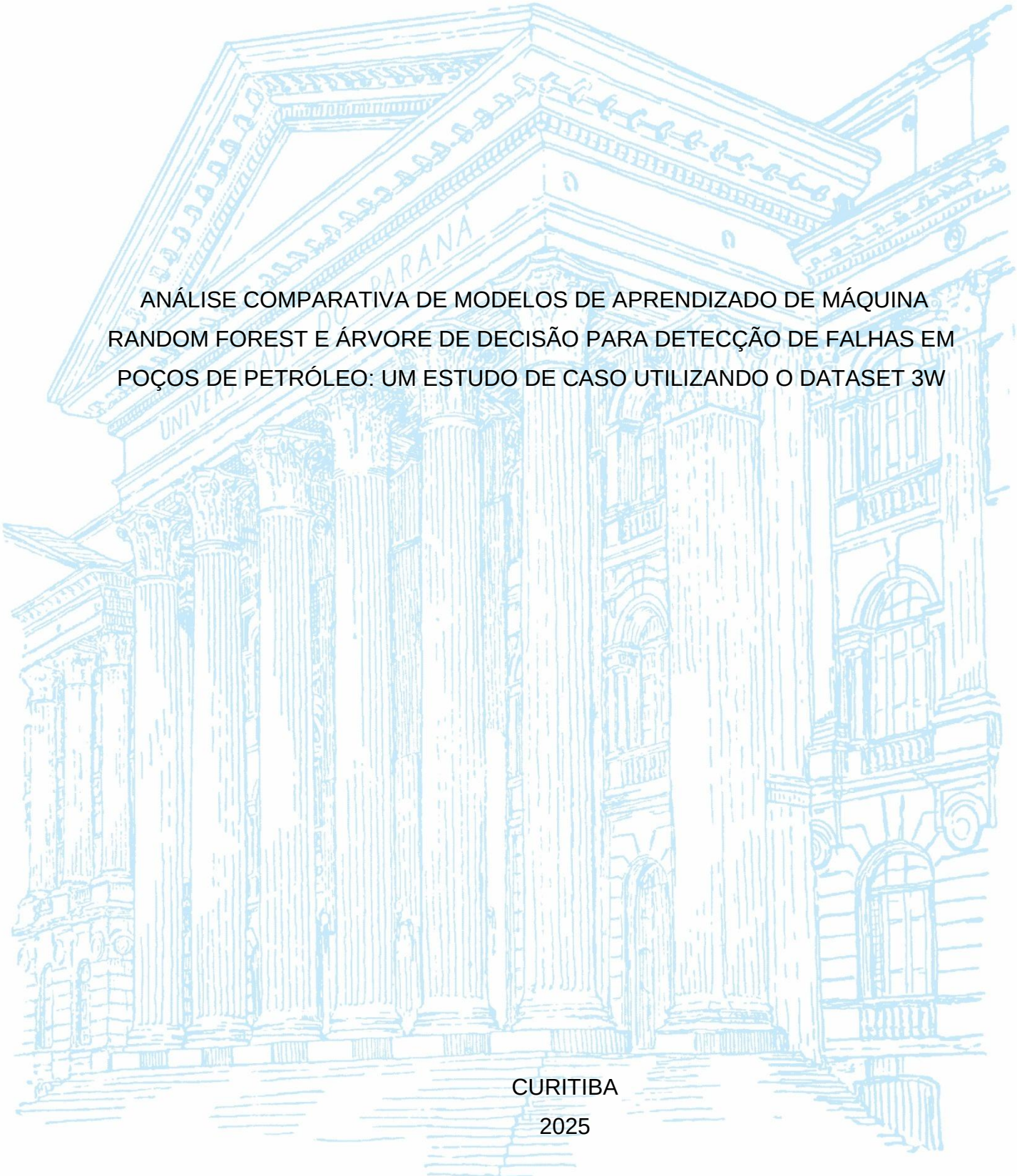
UNIVERSIDADE FEDERAL DO PARANÁ

JUAN FERNANDO COSTA DE MATOS



ANÁLISE COMPARATIVA DE MODELOS DE APRENDIZADO DE MÁQUINA RANDOM FOREST E ÁRVORE DE DECISÃO PARA DETECÇÃO DE FALHAS EM POÇOS DE PETRÓLEO: UM ESTUDO DE CASO UTILIZANDO O DATASET 3W

CURITIBA

2025

Juan Fernando Costa de Matos

ANÁLISE COMPARATIVA DE MODELOS DE APRENDIZADO DE MÁQUINA RANDOM FOREST E ÁRVORE DE DECISÃO PARA DETECÇÃO DE FALHAS EM POÇOS DE PETRÓLEO: UM ESTUDO DE CASO UTILIZANDO O DATASET 3W

TCC apresentado ao curso de Graduação em Engenharia Mecânica, Setor de Tecnologia, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Bacharel em Engenharia Mecânica.

Orientado: Prof. Dr. Pablo Valle

CURITIBA

2025

UNIVERSIDADE FEDERAL DO PARANÁ

PARECER Nº **7/2025/UFPR/R/TC/DEMEC**
PROCESSO Nº 23075.041715/2025-19
INTERESSADO: JUAN FERNANDO COSTA DE MATOS

TERMO DE APROVAÇÃO

Título: ANÁLISE COMPARATIVA DE MODELOS DE APRENDIZADO DE MÁQUINA RANDOM FOREST E ÁRVORE DE DECISÃO PARA DETECÇÃO DE FALHAS EM POÇOS DE PETRÓLEO: UM ESTUDO DE CASO UTILIZANDO O DATASET 3W

Autor: JUAN FERNANDO COSTA DE MATOS

Trabalho de Conclusão de Curso apresentado como requisito parcial para a obtenção do título de bacharel em Engenharia Mecânica. Aprovado pela seguinte banca examinadora:

Prof. Pablo Deivid Valle (UFPR/DEMEC) – Orientador

Prof. João Morais da Silva Neto (UFPR/DEMEC)

Eng. Antonio Celso Wasconcellos de Luca (Mestrando/PPGEM)

Curitiba, 14 de julho de 2025

Documento assinado eletronicamente por **PABLO DEIVID VALLE**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 18/07/2025, às 09:54, conforme art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **JOAO MORAIS DA SILVA NETO**, **CHEFE DO DEPARTAMENTO DE ENGENHARIA MECANICA - TC**, em 06/08/2025, às 15:55, conforme art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **ANTONIO CELSO WASCONCELLOS DE LUCA**, **Usuário Externo**, em 26/08/2025, às 11:21, conforme art. 1º, III, "b", da Lei 11.419/2006.

A autenticidade do documento pode ser conferida aqui informando o código verificador **7953954** e o código CRC **4E6FB835**.

A Deus, por ser a minha fortaleza. À minha esposa, pelo amor, paciência e apoio incondicional em cada etapa desta jornada. À minha mãe, por todos os sacrifícios e por sempre acreditar nos meus sonhos.

AGRADECIMENTOS

Agradeço, ao meu orientador, Professor Pablo Valle pela confiança e pela oportunidade de desenvolver esta pesquisa.

Estendo minha imensa gratidão à minha família, em especial à minha esposa, Thayssa Vilhena, e à minha mãe, Shirlene Costa, pelo amor, apoio incondicional e pela compreensão durante toda a jornada. Sem o incentivo e a paciência de vocês, especialmente nos momentos mais desafiadores, este trabalho não seria possível.

Um agradecimento especial à minha amiga, Kelly Castro. Sua amizade foi um pilar, e seus ensinamentos, uma verdadeira aula. Por toda a ajuda, pelas horas de estudo compartilhadas e por ter sido uma professora nesta caminhada, minha sincera gratidão.

O propósito da computação é o insight, não os números. (EDUARD HAMMING, 1962)

RESUMO

A identificação precoce de falhas em poços de petróleo é crucial para a otimização da produção e a mitigação de riscos operacionais e ambientais na indústria de óleo e gás. Este trabalho tem como objetivo desenvolver e comparar um pipeline de aprendizado de máquina, utilizando os algoritmos Árvore de Decisão e Random Forest, para a classificação da anomalia de "perda rápida de produtividade". A metodologia emprega dados reais do dataset 3W, disponibilizado pela Petrobras, utilizando um poço para treinamento (Poço 16) e outro distinto para validação externa (Poço 20). O pipeline de processamento inclui a segmentação dos dados em janelas temporais, a extração de atributos estatísticos e a redução de dimensionalidade com a Análise de Componentes Principais (PCA), configurada para reter 95% da variância dos dados. Adicionalmente, a técnica de oversampling foi aplicada para balancear as classes no conjunto de treinamento. Os resultados da validação interna no Poço 16 mostraram um desempenho perfeito para ambos os modelos. Contudo, na validação externa no Poço 20, o modelo Random Forest demonstrou uma superioridade expressiva, com uma ROC AUC de 0.9432 contra 0.5287 da Árvore de Decisão, indicando uma capacidade de generalização muito maior. Conclui-se que a arquitetura de ensemble do Random Forest é mais robusta e eficaz para a tarefa, mitigando o superajuste observado no modelo de árvore única e representando uma solução mais confiável para a detecção de anomalias em cenários operacionais reais.

Palavras-chave: Aprendizado de Máquina; Random Forest; Árvore de Decisão; Detecção de Anomalias; Indústria de Petróleo e Gás; Dataset 3W.

ABSTRACT

Early fault detection in oil wells is crucial for optimizing production and mitigating operational and environmental risks in the oil and gas industry. This study aims to develop and compare a machine learning pipeline, using the Decision Tree and Random Forest algorithms, for the classification of the "rapid productivity loss" anomaly. The methodology uses real data from the 3W dataset, provided by Petrobras, utilizing one well for training (Well 16) and a distinct one for external validation (Well 20). The processing pipeline includes segmenting data into temporal windows, extracting statistical features, and reducing dimensionality with Principal Component Analysis (PCA), configured to retain 95% of the data's variance. Additionally, an oversampling technique was applied to balance the classes in the training set. The internal validation results on Well 16 showed perfect performance for both models. However, in the external validation on Well 20, the Random Forest model demonstrated expressive superiority, with a ROC AUC of 0.9432 compared to the Decision Tree's 0.5287, indicating a much greater generalization capability. It is concluded that the ensemble architecture of the Random Forest is more robust and effective for the task, mitigating the overfitting observed in the single tree model and representing a more reliable solution for anomaly detection in real operational scenarios.

Keywords: Machine Learning; Random Forest; Decision Tree; Anomaly Detection; Oil and Gas Industry; 3W Dataset.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS OU SIGLAS

AUC	- Área sob a Curva ROC
3W	- Nome do dataset da Petrobras utilizado
FAR	- Taxa de Falsos Alarmes
KPI	- Indicador-Chave de Desempenho (Key Performance Indicator)
LSTM	- Long Short-Term Memory
PCA	- Principal Component Analysis
ROC	- Receiver Operating Characteristic
SPE	- Especificidade

SUMÁRIO

1 INTRODUÇÃO

A indústria de óleo e gás está sempre buscando por maneiras de tornar seu processo mais eficiente, minimizando os custos, aumentando a produção, trazendo melhorias em segurança e ambientais. Tendo isto em vista, Melo *et al*. (2024) discorre sobre como é essencial para atingir este objetivo que qualquer falha ou anomalia seja identificada precocemente a fim de mitigar os seus efeitos.

Como evidencia do Machine learning como ferramenta útil para a detecção e mitigação de problemas, tem-se o estudo de Abisoye *et al*. (2025), onde é possível observar o machine learning sendo utilizado para auxiliar na prevenção de vulnerabilidades relacionadas a cybersegurança.

Além disso, estratégias de manutenção preditiva utilizando machine learning, podem reduzir falha em até 70% e aumentar a vída útil de um equipamento entre 20% a 40% (Berta, 2021).

Além do uso da predição, o machine learning também tem outro uso, ele auxilia na classificação de falhas, identificando padrões e analisando uma grande quantidade de dados. Segundo Wu (2025), o uso do Machine learning voltado a classificação possibilitou relevar três categorias distintas de falhas que permitiram a identificação de microfraturas antes que resultassem em maiores instabilidades.

Tendo isto em mente, o objetivo deste trabalho é a comparação da eficácia entre dois modelos de machine learning para a classificação de falhas em poços de petróleo com base no dataset 3W disponibilizado pela Petrobras. Os modelos serão utilizados para identificar um tipo de falha. Um poço será utilizado para treinamento e outro, distinto para validação da métrica.

1.1 OBJETIVOS

1.1.1 Objetivo geral

Criar um pipeline de aprendizado de máquina fazendo uma comparação entre os métodos Árvore de Decisão e Random Forest para classificação de anomalias em poços de petróleo, fazendo uso do dataset 3W da Petrobras.

1.1.2 Objetivos específicos

a) Realizar uma revisão bibliográfica detalhada sobre técnicas de detecção de falhas em poços de petróleo;
b) Analisar os resultados de estudos anteriores aplicados ao dataset 3W, identificando lacunas e oportunidades de melhoria;
c) Desenvolver um pipeline de engenharia de atributos utilizando técnicas como PCA;
d) Implementar e avaliar diferentes métodos de aprendizado de máquina, Random Forest e Árvore de Decisão;
e) Validar a abordagem proposta utilizando dados reais de poços – treinando com dados de um poço e validando com dados de outro – e utilizando métricas como F1-Score e FAR;
f) Estimar o possível ganho econômico ao implementar o melhor método de aprendizado de máquina avaliado neste trabalho ao sistema da Petrobras;

1.2 JUSTIFICATIVA

A Petrobras (2022) em seus estudo sobre machine learning concorda sobre a necessidade de empregar esta tecnologia para melhoria da eficiência de seus serviços. A demora na percepção de uma falha pode levar a consequências severas de segurança, meio ambiente e custos.

Segundo Vargas *et al*. (2019), eventos indesejáveis como aumento abrupto de sedimentos básicos e fluxo intermitente de líquido ou gás, levaram a perdas de produção de 1.514.000 barris de petróleo correspondente à 75,7 milhões de dólares no ano de 2016 para a base da Petrobras localizada no Espírito Santo, fato este que corrobora sobre a importância da predição e classificação rápida e assertiva dos problemas na indústria de petróleo e gás.

Como mencionado anteriormente neste trabalho, a utilização de técnicas de machine learning permite análise de grandes volumes de dados, o que possibilita a percepção de padrões sutis que indicam o surgimento de anomalias. (Anzai *et al*., 2023).

De acordo com Melo (2024), é desafiador o estudo para detecção de anomalias, devido as características complexas dos dados. Dito isto, é de suma

importância a compreensão aprofundada das vantagens e limitações de cada modelo para se ter soluções robustas e confiáveis.

Logo, através do pipeline de treinamento dos modelos e uso de dados reais evidencia o poder de generalização e adaptação da metodologia para cenários reais. Desta forma, assim a evidencia-se que a pesquisa presente neste trabalho contribui de maneira significativa para a melhoria contínua na indústria petrolífera integrando Big Data e classificação para detecção de anomalias em ambiente de alta complexidade e elevado risco.

2 REVISÃO BIBLIOGRÁFICA

2.1 DATASET 3W

O conjunto de dados 3W foi criado pela Petrobras com o objetivo de fornecer uma base realista e acessível, destinada à pesquisa e ao desenvolvimento de algoritmos de Aprendizado de Máquina. Ele tem como foco a detecção precoce e a classificação de eventos indesejáveis em poços de petróleo offshore, que envolvem a exploração de petróleo em ambientes marinhos de grandes profundidades.

Estudos recentes têm destacado o uso do conjunto de dados 3W como uma referência para a avaliação de abordagens de Aprendizado de Máquina. Alguns exemplos disso são os trabalhos de Turan e Jaschke (2021), que investigaram a influência da seleção de características e de diferentes classificadores em conjuntos de dados multivariados. Gatta *et al*. (2024) mostraram como autoencoders podem ser eficazes na extração de características, enquanto Dias *et al*. (2024) propuseram uma arquitetura modular baseada em wavelets e Aprendizado de Máquina para aumentar a robustez nos processos de classificação.

O projeto 3W é composto pelo banco de dados 3W, um conjunto de dados realista e público que registra eventos raros e indesejáveis ocorridos em poços de petróleo, e pelo 3W Toolkit, um pacote de software desenvolvido para facilitar a experimentação com o 3W dataset. Sendo criado em 2019 e lançado oficialmente em em 2022, o projeto foi uma iniciativa estratégica da Petrobras, conduzida pelo departamento responsável pela Garantia de Fluxo e pelo CENPES (Centro de Pesquisas, Desenvolvimento e Inovação Leopoldo Américo Miguez de Mello). Ele marcou o lançamento do primeiro piloto do programa "Conexões para Inovação - Módulo Open Lab", que tem como objetivo encontrar soluções inovadoras para os desafios empresariais por meio da colaboração aberta. Na versão 2.0.0, disponibilizada em 25 de julho de 2024, o banco de dados passou a incluir informações detalhadas sobre eventos indesejáveis, categorizados por tipo e impacto operacional, possibilitando a aplicação de técnicas de Aprendizado de Máquina para identificar padrões complexos e prever falhas.

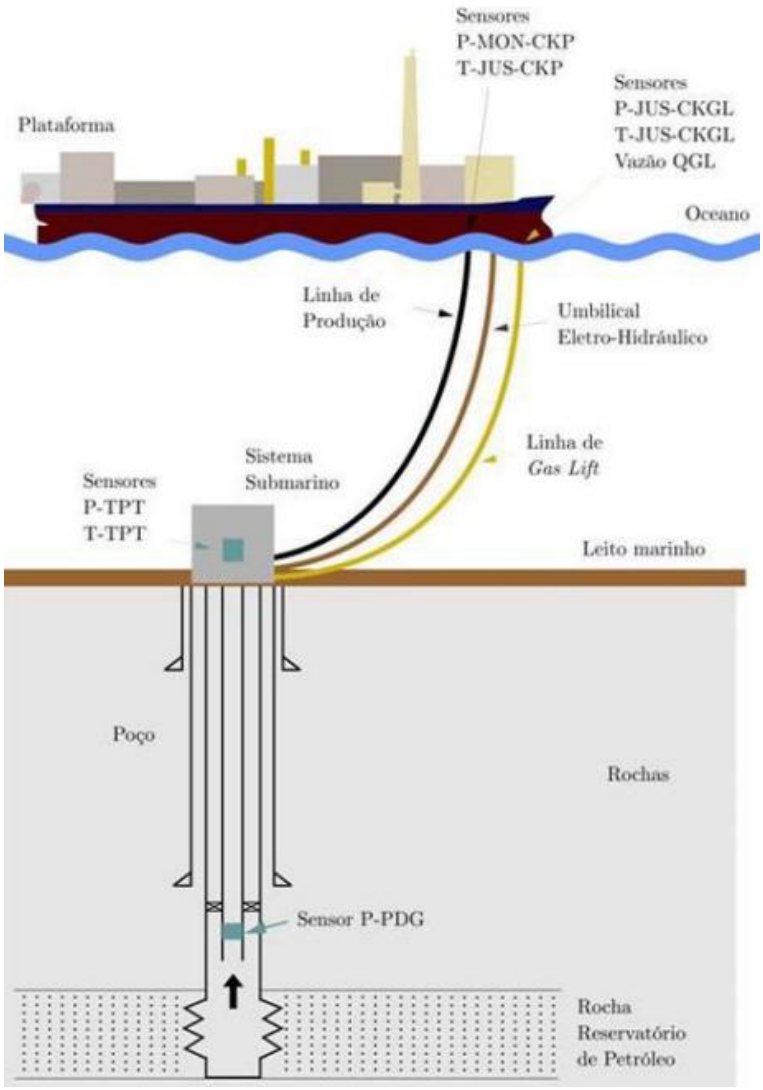
2.2 CONTEXTUALIZAÇÃO DOS POÇOS DE PETRÓLEO OFFSHORE E SEUS SENSORES

Na extração de petróleo em alto-mar, muitos poços utilizam a elevação natural para o transporte dos fluidos até a superfície. Esse processo depende de uma série de equipamentos e sensores distribuídos tanto no fundo do poço quanto na plataforma, permitindo o acompanhamento contínuo das operações. Os principais componentes desse sistema incluem as estruturas de suporte, os equipamentos de perfuração e o sistema responsável pelo fluxo dos fluidos extraídos.

Os sensores desempenham um papel essencial na aquisição de dados operacionais. Em poços offshore, essa coleta de informações é realizada por dispositivos como o sensor permanente de fundo (PDG), que mede a pressão no reservatório, e o transdutor de pressão e temperatura (TPT), responsável pelo monitoramento simultâneo dessas variáveis na Árvore de Natal Molhada (ANM). Além disso, sensores localizados na válvula de controle de produção (CKP) e ao longo da linha de Gas Lift – incluindo aqueles que registram a pressão e a temperatura próximas ao choke de gás lift (P-JUS-CKGL e T-JUS-CKGL) – ampliam a capacidade de monitoramento do sistema. Esse conjunto de medições possibilita um controle mais preciso das operações, especialmente quando a elevação natural é complementada por mecanismos artificiais para otimizar a extração.

A precisão dos sensores pode sofrer variações, uma vez que, devido aos elevados custos de intervenção e à complexidade de acesso aos poços offshore, a calibração nem sempre ocorre de forma regular. Essa inconsistência dificulta ainda mais a identificação de anomalias, pois os dados coletados podem conter ruídos, lacunas ou valores discrepantes, demandando a aplicação de técnicas avançadas de pré-processamento e análise para garantir a confiabilidade das informações.

FIGURA 1 – REPRESENTAÇÃO DA POSIÇÃO DOS SENSORES DE UM POÇO MARÍTIMO SURGENTE DE PETRÓLEO



FONTE: Junior (2022)

2.2.1 O Desafio Operacional: O Poço como um Paciente Monitorado

Para contextualizar a complexidade do problema, uma analogia se mostra eficaz: um poço de petróleo offshore é como um paciente em uma unidade de terapia intensiva (UTI), conectado a diversos monitores que medem seus "sinais vitais" 24 horas por dia. Os sensores mencionados, como o PDG e o TPT, atuam como esses monitores, medindo a "pressão arterial" (pressão) e a "temperatura" do poço em pontos críticos.

A anomalia estudada neste trabalho, a "perda rápida de produtividade", seria o equivalente a uma "queda de pressão súbita" ou uma "arritmia cardíaca" neste

paciente. Se não for detectada e tratada a tempo, o poço pode sofrer uma "parada cardíaca", ou seja, parar de produzir completamente, resultando em prejuízos financeiros e operacionais significativos.

O objetivo deste trabalho, portanto, é construir e comparar dois "médicos digitais" (modelos de Machine Learning) para vigiar esses sinais vitais. Compara-se a eficácia de uma abordagem mais simples, a Árvore de Decisão, com uma mais robusta e complexa, o Random Forest, para determinar qual deles é mais confiável para diagnosticar a "condição cardíaca" do poço antes que ela se agrave.

2.2.2 A Física da Perda de Produtividade

Para que um modelo de aprendizado de máquina seja eficaz, é imperativo que ele aprenda a partir de padrões nos dados que são manifestações de fenômenos físicos reais. A perda de produtividade em poços de petróleo raramente é um evento sem causa; ela é, na maioria das vezes, o sintoma de uma condição adversa conhecida como dano à formação (formation damage) (Halim, Hamidi, Akisanya, 2021). O dano à formação é definido como a redução da permeabilidade natural da rocha do reservatório na região próxima ao poço, o que prejudica o fluxo de hidrocarbonetos e impacta diretamente a produtividade (Faergestad, 2019. Segundo Ezenweichu e Laditan (2015), o dano à formação é uma das principais causas de redução da produtividade e pode ser induzido por uma complexa interação de efeitos mecânicos, químicos, biológicos e térmicos. Os mecanismos mais relevantes para a detecção por sensores de pressão e temperatura são os mecânicos e químicos.

a) Danos Mecânicos:

- Esta categoria refere-se à obstrução física dos canais porosos da formação. O mecanismo mais comum é a migração de finos, que ocorre quando partículas de pequena granulometria (quartzo, argilas), naturalmente presentes na rocha, se desprendem devido às forças de cisalhamento do fluido e obstruem as gargantas dos poros, reduzindo a permeabilidade. A análise de You et al. (2024) demonstra que a força capilar durante o fluxo multifásico (óleo e água) é um fator de desprendimento muito mais significativo do que as forças de arrasto em fluxo monofásico. Outro dano mecânico é o bloqueio por sólidos

externos, onde fluidos de operações de perfuração ou completação carregam sólidos para dentro da formação, causando entupimento;

b) Danos Químicos:

- Esta categoria envolve reações adversas entre os fluidos do poço e os minerais da formação. O inchamento de argilas (clay swelling) ocorre quando argilas reativas entram em contato com fluidos de baixa salinidade, expandindo seu volume e estrangulando os canais de fluxo. Segundo Faergestad (2016), este é um mecanismo químico comum que pode ser mitigado com o uso de fluidos de alta salinidade. A precipitação de incrustações (scale deposition) acontece quando mudanças de pressão e temperatura levam à precipitação de minerais, como carbonatos de cálcio, que formam uma "pele" (skin) de baixa permeabilidade na face da rocha;

Além do dano à formação, a depleção do reservatório e as variações de temperatura e viscosidade também influenciam a produtividade. A extração contínua de fluidos reduz a pressão do reservatório, diminuindo a força motriz que impulsiona os hidrocarbonetos (Techtac, 2023). A queda de pressão também induz uma queda de temperatura, que pode aumentar a viscosidade do óleo e levar à precipitação de parafinas e asfaltenos, restringindo o fluxo (Ezenweichu, Laditan, 2015).

2.2.3 Coleta e Qualidade de dados em Ambientes Offshore

A eficácia de qualquer modelo de detecção de anomalias depende da qualidade dos dados de entrada. O ambiente submarino impõe desafios singulares à instrumentação e coleta de dados (Skalvik *et al*., 2023). A Figura 1 ilustra a disposição geral dos sensores em um poço marítimo.

A qualidade dos dados coletados é afetada por múltiplos fatores sendo eles, ambiente hostil e degradação do sensor, limitações de manutenção e calibração, desafios de comunicação e energia.

Com relação ao ambiente hostil e degradação do sensor, os sensores operam sob condições extremas de alta pressão, alta temperatura e em um meio corrosivo, o que leva à degradação física, desvios (*drifts*), ruído e falhas completas (Skalvik *et al*., 2023).

Ainda, devido aos custos e à complexidade logística, a manutenção e a calibração dos sensores são infrequentes. Conforme destacado por Fascista (2022), essa falta de calibração periódica pode gerar desvios sistemáticos que se acumulam ao longo do tempo.

Por fim, a transmissão de dados do fundo do mar para a superfície é um desafio. A comunicação acústica, método mais viável, sofre de baixa largura de banda, altos atrasos e alto consumo de energia, resultando em dados com baixa frequência de amostragem e lacunas (Jindal, Saxena, Singh, 2014).

A natureza ruidosa e incompleta dos dados de sensores offshore é a principal razão pela qual modelos de *ensemble* como o Random Forest são mais robustos do que uma Árvore de Decisão única, que é mais sensível ao ruído e propensa ao superajuste (overfitting).

2.3 CARACTERISTICAS DO DATASET E ANOMALIAS

O dataset 3W é uma base de dados pública e realista, desenvolvida pela Petrobras, que registra eventos em poços de petróleo, incluindo tanto condições normais de operação quanto falhas e anomalias detectadas por sensores offshore. Composto por séries temporais multivariadas, esse conjunto de dados armazena informações sobre variáveis como pressão, temperatura, vazão e o status operacional das válvulas, sendo cada instância registrada dentro de uma janela amostral específica.

O dataset 3W, apresenta um total de 2.228 instâncias de eventos, compostas por dados reais, simulações e representações manuais. Esses eventos são categorizados em nove tipos de falhas, além das operações normais. As instâncias contêm registros multivariados provenientes de sensores que monitoram variáveis como pressão, temperatura e vazão. Neste trabalho, serão utilizadas exclusivamente as instâncias que correspondem a dados reais de poços de petróleo.

Segundo Vargas *et al*. (2019) O dataset também é divido em 10 pastas de 1 a 9, sendo cada uma dessas pastas dedicadas a eventos anômalos específicos. O evento que será avaliado será a perda rápida de produção que corresponde a pasta 5

Além da categorização por tipo de anomalia, cada evento também é classificado com base no estado operacional do poço, que pode estar aberto ou

fechado, e na classe de observação, que indica o nível da anomalia. Essa classificação é dividida em três categorias: normal (0), transiente de anomalia (101 a 109) e estado permanentemente anômalo (1 a 9). O período normal indica a ausência de anomalias, enquanto o período transiente representa a fase de transição entre o funcionamento normal e a estabilização da anomalia, momento em que a dinâmica associada ao evento ainda está em evolução. Quando essa fase transitória se encerra, inicia-se o estado estável da anomalia, caracterizando um desvio permanente nas condições operacionais do poço (Vargas, 2019). Essa classificação está presente na coluna classe que será utilizada para definir se as janelas analisadas apresentam anomalia ou não.

2.4 PRINCIPAL COMPONENT ANALYSIS (PCA)

A Análise de Componentes Principais (PCA) é uma técnica estatística multivariada utilizada para a redução de dimensionalidade em conjuntos de dados complexos. Seu objetivo é transformar um conjunto de variáveis possivelmente correlacionadas em um novo conjunto de variáveis linearmente não correlacionadas, denominadas componentes principais. Esses componentes são ordenados de forma que o primeiro retenha a maior variância (informação) possível dos dados originais, e cada componente subsequente capture a maior parte da variância restante, sob a restrição de ser ortogonal aos anteriores. (Jolliffe, 2002).

2.4.1 Fundamentos Matemáticos do PCA

O PCA opera sobre a matriz de covariância dos dados, que descreve como as variáveis se movem em conjunto. Matematicamente, a técnica encontra os autovetores e autovalores dessa matriz. De forma intuitiva, os autovetores representam as direções de máxima variância nos dados (os eixos dos novos componentes principais), e os autovalores correspondentes indicam a magnitude dessa variância, ou seja, quanta "informação" cada componente principal captura dos dados originais (Jolliffe, Cadima, 2016).

Para reduzir a dimensionalidade, selecionam-se os *k* primeiros componentes que, juntos, explicam uma porcentagem desejada da variância total (neste trabalho, 95%), descartando os componentes restantes que representam principalmente ruído.

Essa abordagem é amplamente utilizada em estudos de caso para detecção de falhas em equipamentos industriais, como bombas submersíveis elétricas, que apresentam desafios de dados similares aos dos poços de petróleo (Peng *et al*., 2021).

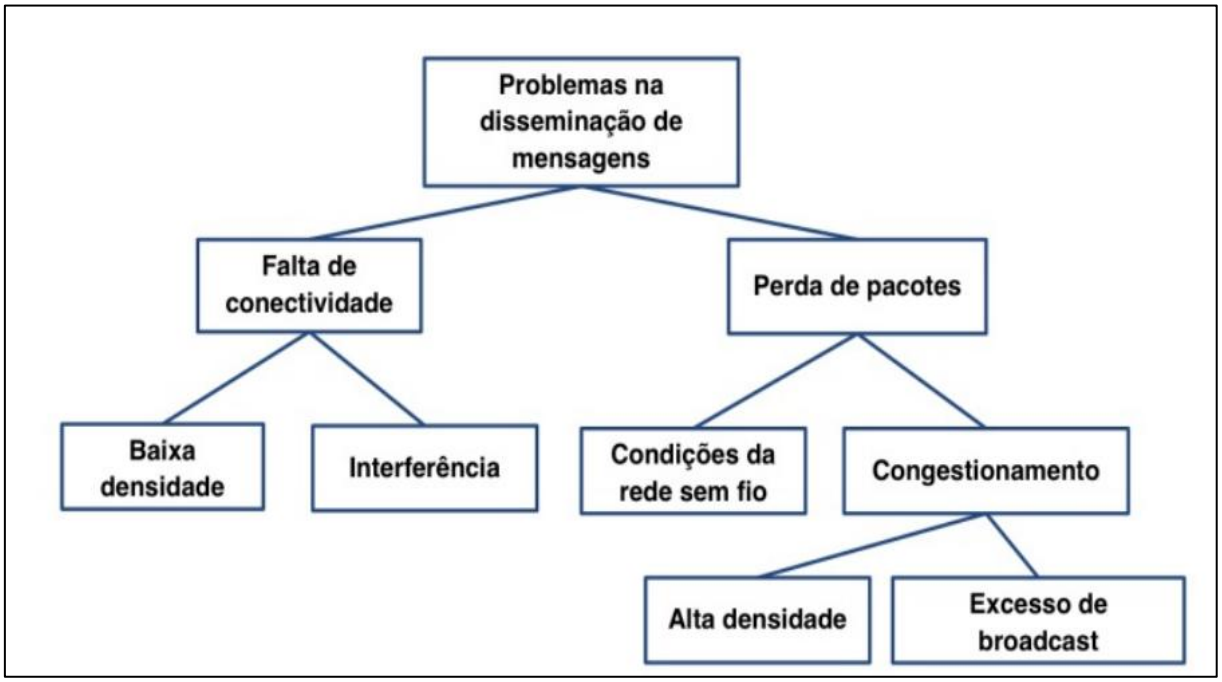
2.5 MODELOS RANDOM FOREST E ARVORE DE DECISÃO

Segundo Liu *et al*. (2012), os modelos baseados em árvores são ferramentas fundamentais no campo do Aprendizado de Máquina, notáveis por sua capacidade de modelar relações complexas e não lineares nos dados. Sua estrutura intuitiva os torna aplicáveis a problemas de classificação e regressão. Neste trabalho, foram selecionados dois algoritmos proeminentes desta classe: a Árvore de Decisão, como modelo base, e o Random Forest, como uma evolução baseada na técnica de ensemble.

2.5.1 A Lógica da Árvore de Decisão (Decision Tree)

Segundo Breiman *et al*. (1984), a Árvore de Decisão é um modelo de aprendizado supervisionado cuja estrutura se assemelha a um fluxograma. O modelo funciona dividindo recursivamente o conjunto de dados em subconjuntos mais puros, onde cada nó interno da árvore representa um teste condicional sobre um atributo, cada ramificação é o resultado desse teste, e cada nó folha corresponde a uma decisão final ou rótulo de classe. A FIGURA 2 ilustra essa estrutura hierárquica.

FIGURA 2 - EXEMPLIFICAÇÃO DE ESTRUTURA HIERARQUICA DE UMA ÁRVORE DE DECISÃO



FONTE: Albano, Nogueira e Silva (2015).

A construção da árvore é guiada por algoritmos que buscam otimizar as divisões em cada nó. Conforme Quinlan (1986), pioneiro no desenvolvimento de árvores de decisão, o processo de seleção de atributos é realizado com base em métricas estatísticas, como o Ganho de Informação (derivado da Entropia) ou o Índice Gini. O objetivo dessas métricas é encontrar o atributo que, ao ser usado para dividir os dados, resulta nos subconjuntos mais homogêneos possíveis em relação à classe de destino.

A fórmula da entropia para um conjunto S com c classes é:

$$H(S) = -\sum_{i=0}^{c} p_i \log_{2(}\, p_{i)}$$

Onde pi é a proporção de amostras da classe *i* em S.

O Índice Gini mede a probabilidade de um elemento ser classificado incorretamente e sua fórmula é:

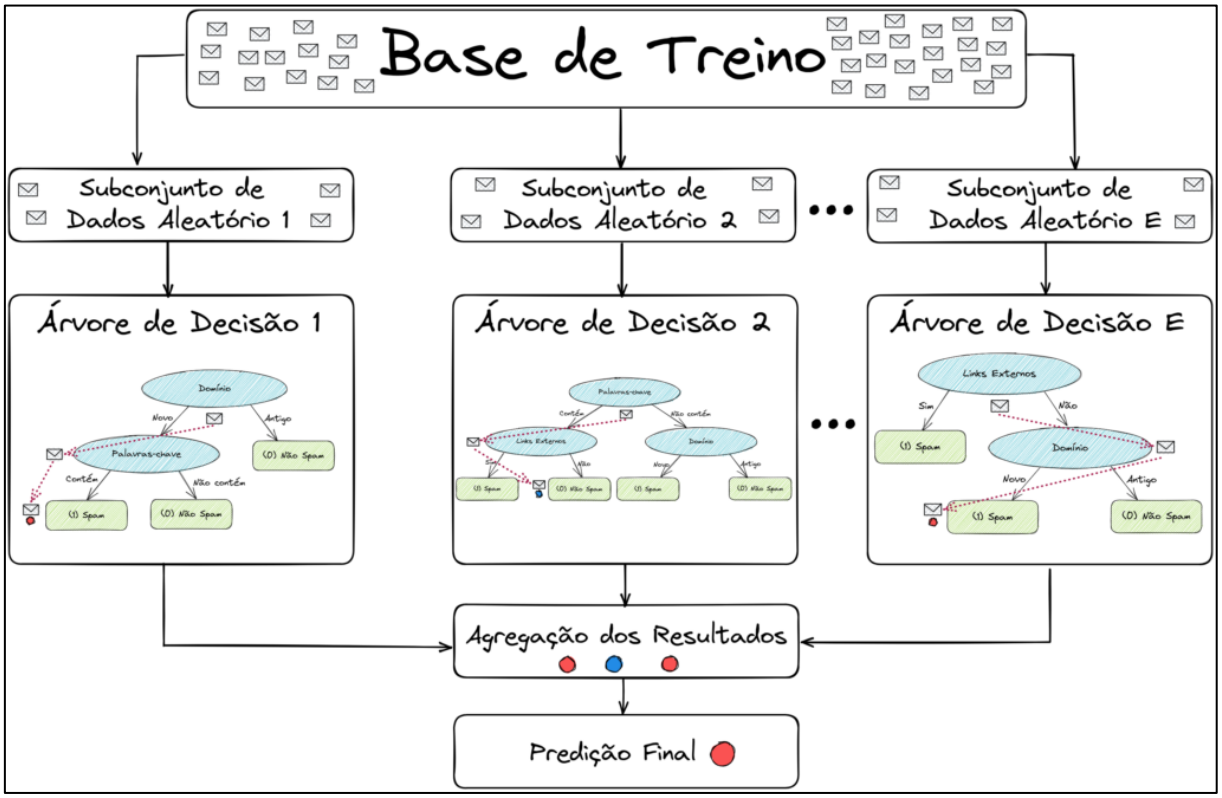
$$Gini(S) = 1 - \sum_{i=0}^{c} {p_i}^2$$

A principal vantagem deste modelo, amplamente discutida na literatura, é sua alta interpretabilidade (Breiman *et al*., 1984). As regras lógicas "se-então" podem ser extraídas diretamente da estrutura da árvore, permitindo que especialistas do domínio compreendam facilmente o processo de tomada de decisão do modelo. Contudo, conforme apontado por Breiman *et al*. (1984), a principal desvantagem da árvore de decisão é sua alta sensibilidade aos dados de treinamento, o que a torna propensa ao superajuste (overfitting). Uma árvore excessivamente complexa pode acabar "decorando" o ruído presente nos dados, em vez de aprender o padrão subjacente, resultando em um baixo desempenho em dados novos e não vistos. Técnicas como a poda (pruning) são empregadas para simplificar a árvore e melhorar sua capacidade de generalização.

2.5.2 Random Forest

Para solucionar as limitações da árvore única, especialmente o overfitting, Breiman (2001) propôs o algoritmo Random Forest. Trata-se de um método de *ensemble* que, como o nome sugere, constrói uma "floresta" de múltiplas árvores de decisão durante o treinamento e combina suas predições para obter um resultado mais acurado e estável. A robustez do método advém de dois mecanismos de aleatorização e a representação do modelo pode ser observado a partir da FIGURA 3.

FIGURA 3 - REPRESENTAÇÃO DO MODELO RANDOM FOREST



FONTE: Lopes (2023).

O primeiro pilar é a técnica de Bagging (Bootstrap Aggregating), introduzida por Breiman (1996), que consiste em criar diversas amostras de treinamento por meio de um processo de amostragem com reposição a partir do conjunto de dados original. Cada árvore da floresta é, então, treinada em uma dessas amostras, garantindo diversidade entre elas.

O segundo mecanismo, detalhado em Breiman (2001), é a Seleção Aleatória de Atributos. Diferente de uma árvore de decisão padrão, que avalia todos os atributos para encontrar a melhor divisão em um nó, o Random Forest restringe essa busca a um subconjunto aleatório de atributos. Este processo de dupla aleatorização (nos dados e nos atributos) resulta na criação de árvores altamente descorrelacionadas. Segundo Liu *et al*. (2012), a agregação das predições de múltiplos modelos descorrelacionados por meio de um sistema de votação majoritária faz com que os erros individuais de cada árvore tendam a ser compensados, resultando em um modelo com menor variância, menos suscetível ao overfitting e com um poder de generalização superior.

2.6 DETECÇÃO DE FALHAS NA INDÚSTRIA 4.0

O presente trabalho se insere no contexto da transformação digital da indústria de óleo e gás, um movimento alinhado aos princípios da Indústria 4.0. Esta transformação é caracterizada pela integração de tecnologias como Internet das Coisas (IoT), Big Data e Inteligência Artificial (IA) para otimizar operações (Ferreira, 2024). A disponibilidade de grandes volumes de dados de sensores em tempo real permitiu a transição para a manutenção preditiva, uma estratégia proativa que utiliza algoritmos de aprendizado de máquina para prever falhas antes que elas ocorram (Ohalete, 2023).

A detecção de anomalias em sistemas industriais é um campo de pesquisa ativo. Para contextualizar a escolha dos modelos deste trabalho, é relevante analisar outras técnicas aplicadas em cenários similares:

a) Support Vector Machines (SVM): Diversos estudos demonstram a eficácia do SVM para a detecção de vazamentos e anomalias em dutos, frequentemente alcançando altas taxas de acurácia em tarefas de classificação binária (Aljameel, 2022).
b) Redes Neurais e Deep Learning: Para a análise de séries temporais complexas, arquiteturas de deep learning como as Redes Neurais Recorrentes (RNNs) e, mais especificamente, as Long Short-Term Memory (LSTM), são amplamente utilizadas. O trabalho de Santos *et al*. (2020), por exemplo, emprega LSTMs para o monitoramento em tempo real e detecção de anomalias em poços de petróleo.
c) Outros Métodos de Ensemble: Além do Random Forest, outras técnicas de ensemble, como o Gradient Boosting, também são aplicadas com sucesso, construindo modelos de forma sequencial para corrigir os erros do modelo anterior (Aljameel, 2022).

2.7 FERRAMENTAS COMPUTACIONAIS PARA APRENDIZADO DE MÁQUINA

A implementação de modelos de aprendizado de máquina, como os analisados neste trabalho, depende de um ecossistema de ferramentas computacionais que inclui linguagens de programação, ambientes de desenvolvimento e bibliotecas especializadas.

2.7.1 A Linguagem Python na Ciência de Dados

A linguagem de programação Python, criada por Guido van Rossum, estabeleceu-se como a principal ferramenta para ciência de dados e aprendizado de máquina. Sua popularidade deriva de uma sintaxe simples e legível, que reduz a curva de aprendizado e permite que pesquisadores e engenheiros desenvolvam soluções complexas com menos linhas de código em comparação com outras linguagens (Jyothi, Yamaganti, 2019). Sendo uma linguagem de código aberto (open-source), Python se beneficia de uma vasta e ativa comunidade global, que contribui para um ecossistema robusto de bibliotecas e ferramentas. Esse ecossistema inclui bibliotecas poderosas para manipulação de dados, computação científica e visualização, que são fundamentais para o pipeline de aprendizado de máquina.

2.7.2 Google Colaboratory como Ambiente de Pesquisa

O Google Colaboratory, ou "Colab", é um ambiente de desenvolvimento interativo baseado em nuvem que permite escrever e executar código Python diretamente no navegador. Baseado nos Jupyter Notebooks, o Colab elimina a necessidade de configuração de hardware e software local, fornecendo um ambiente pré-configurado com as principais bibliotecas de aprendizado de máquina e acesso gratuito a recursos computacionais de alta performance, como Unidades de Processamento Gráfico (GPUs) (Google Colab, 2022). Essas características democratizam o acesso à pesquisa em aprendizado de máquina, permitindo que estudantes e pesquisadores desenvolvam e testem modelos complexos sem a necessidade de hardware caro. Além disso, suas funcionalidades de colaboração em tempo real e fácil compartilhamento o tornam uma ferramenta ideal para projetos de pesquisa e educação (Carneiro *et al.*, 2018).

2.7.3 Bibliotecas Essenciais

O poder do Python para ciência de dados é amplificado por suas bibliotecas especializadas. Para este trabalho, as mais relevantes são:

a) Scikit-learn: É a biblioteca de aprendizado de máquina mais utilizada em Python. Ela oferece uma vasta gama de algoritmos eficientes para tarefas de classificação, regressão, clusterização e redução de dimensionalidade, todos acessíveis por meio de uma interface de programação (API) consistente e unificada. Essa consistência facilita a experimentação e a comparação entre diferentes modelos.
b) Pandas: É a biblioteca padrão para manipulação e análise de dados estruturados em Python. Seu principal objeto, o DataFrame, é uma estrutura de dados bidimensional, similar a uma planilha, que permite carregar, limpar, transformar e analisar dados tabulares de forma intuitiva e eficiente
c) Matplotlib: É a biblioteca fundamental para a criação de visualizações de dados estáticas e de alta qualidade em Python. Sua flexibilidade permite a criação de uma ampla variedade de gráficos, como os gráficos de linha e as matrizes de confusão utilizadas neste trabalho para avaliar e apresentar os resultados dos modelos.

2.8 PRINCIPAIS DE AVALIAÇÃO DE MODELOS DE MACHINE LEARNING

2.8.1 Acurácia

A acurácia mede a proporção de todas as previsões corretas do modelo em relação ao total de casos. É definida por (TP+TN)/(TP+TN+FP+FN), onde TP e TN são os verdadeiros positivos e negativos, e FP e FN são falsos positivos e negativos. Em termos gerais, a acurácia indica o “grosso” desempenho do modelo. No entanto, essa métrica pode ser enganosa em conjuntos de dados desbalanceados: se a classe de interesse (por exemplo, falhas) for muito rara, um classificador que sempre prevê a classe majoritária pode apresentar acurácia alta sem realmente identificar a classe minoritária. Por exemplo, num cenário em que apenas 1% dos casos são falhas, um modelo que sempre prevê “não-falha” alcança 99% de acurácia apesar de não detectar nenhuma falha real. Em síntese, a acurácia é fácil de interpretar, mas tende a mascarar o desempenho quando há forte desbalanceamento de classes.

2.8.2 Precisão (Precision)

precisão avalia quantos dos exemplos que o modelo classificou como positivos são de fato positivos. Matematicamente, é dada por TP/(TP+FP). Em detecção de falhas, uma alta precisão significa que poucos dos alertas de falha emitidos são falsos – ou seja, quando o modelo sinaliza uma falha, essa predição é normalmente correta. Essa métrica é útil quando falsos positivos têm custo elevado (por exemplo, muitos alarmes falsos). Por outro lado, a precisão ignora os falsos negativos: um modelo que quase nunca sinaliza falhas pode obter precisão alta apesar de perder a maior parte das falhas reais. Em cenários muito desbalanceados, portanto, confiar apenas na precisão pode ser insuficiente, pois ela não penaliza a omissão de casos positivos (Google Developers, 2025).

2.8.3 Recall ou Sensibilidade

A Sensibilidade, true positive rate ou recall mede a proporção de exemplos positivos reais que foram corretamente identificados. Calcula-se por TP/(TP+FN). Em detecção de falhas, ela representa a taxa de falhas efetivamente detectadas pelo modelo. A sensibilidade é crítica quando falsos negativos são custosos (por exemplo, falhas não identificadas podem levar a danos maiores), pois reflete a capacidade do modelo de capturar todos os casos positivos. Em conjuntos desbalanceados, a sensibilidade costuma ser mais informativa que a acurácia, justamente por focar na classe minoritária. Contudo, a sensibilidade não leva em conta falsos positivos: um modelo pode obter alta sensibilidade sinalizando muitas falhas (incluindo muitos falsos alarmes), o que reduz a precisão. Há, portanto, um trade-off entre sensibilidade e precisão (Google Developers, 2025)

2.8.4 F1-Score

O F1-Score é a média harmônica entre precisão e sensibilidade, definido por 2 * (Precisão * Sensibilidade) / (Precisão + Sensibilidade) (Google Developers, 2025). Essa métrica equilibra os dois indicadores: obtém valores altos apenas quando ambos precisão e sensibilidade são altas. Em problemas com desbalanceamento, o F1 é preferível à acurácia (Google Developers, 2025), pois penaliza de forma equilibrada

falsos positivos e falsos negativos. Assim, o F1-Score é muito usado em detecção de falhas para avaliar o desempenho global na detecção da classe rara. Como limitação, o F1 ignora verdadeiros negativos, não refletindo diretamente o quão bem o modelo acerta a classe negativa. Além disso, pesquisas demonstram que F1 (assim como acurácia) pode dar valores otimistas em situações de forte desequilíbrio (Chicco, Jurman, 2023), de modo que deve ser analisado em conjunto com outras métricas.

2.8.5 Curva ROC e AUC

A curva ROC (Receiver Operating Characteristic) ilustra o desempenho do classificador plotando a taxa de verdadeiros positivos (revocação) em função da taxa de falsos positivos para diversos limiares de decisão. A área sob essa curva, o AUC (Area Under the ROC Curve), resume a qualidade do modelo independente de limiar. Intuitivamente, o AUC é a probabilidade de que o modelo atribua uma pontuação maior a um exemplo positivo do que a um negativo aleatórios (Google Developers, 2025). O valor do AUC varia de 0,5 (classificador aleatório) a 1,0 (classificador perfeito) (Google Developers, 2025). Essa métrica é útil para comparar modelos de forma geral, mas apresenta limitações em conjuntos desbalanceados: partes da curva ROC de pouca relevância prática podem inflar o AUC (Google Developers, 2025). Em especial na detecção de falhas, recomenda-se complementar o AUC com curvas de precisão-sensibilidade, pois elas avaliam melhor o desempenho na classe minoritária.

2.8.6 Especificidade e Taxa de Falsos Alarmes (FAR)

A especificidade (taxa de verdadeiros negativos) mede a proporção de exemplos negativos que foram corretamente classificados. É calculada por TN / (TN + FP) (Qlik Cloud, 2025). Em outras palavras, indica a fração de casos sem falha que o modelo identificou corretamente. Já a taxa de falsos alarmes (FAR, ou false positive rate) é FP / (FP + TN) (Qlik Cloud, 2025), complementando a especificidade. Um FAR baixo significa que poucos casos negativos foram sinalizados erroneamente como positivos. Especificidade e FAR são importantes quando falsos positivos têm alto custo. No entanto, se houver poucos exemplos da classe negativa, essas métricas podem ficar instáveis. Em geral, especificidade e sensibilidade devem ser avaliadas

conjuntamente (assim como precisão e FAR) para compreender completamente o desempenho em cada classe.

2.8.7 Coeficiente de Correlação de Matthews (MCC)

O coeficiente de correlação de Matthews (MCC) fornece uma única medida que considera TP, TN, FP e FN simultaneamente. Seu cálculo é:

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

O MCC varia de -1 (predições totalmente invertidas) a +1 (classificação perfeita), valendo 0 para desempenho equivalente a um palpite aleatório (Qlik Cloud, 2025). De acordo com Chicco e Jurman (2023) e A principal vantagem do MCC é ser equilibrado: ele só alcança valor alto se o classificador apresentar boa sensibilidade **e** boa especificidade ao mesmo tempo. Por isso, o MCC é considerado robusto em cenários desbalanceados, dando uma visão completa do desempenho. Em contrapartida, sua interpretação é menos intuitiva e é raramente usado por analistas não especializados.

3 MATERIAIS E MÉTODOS

Este capítulo detalha a metodologia empregada, desde a seleção e preparação dos dados até o treinamento e validação dos modelos de aprendizado de máquina.

3.1 DATASET

O dataset 3W, que serve de base para este estudo, apresenta um total de 2.228 instâncias de eventos, compostas por uma mistura de dados reais, simulações e representações manuais. Essa composição foi necessária para garantir uma base de dados abrangente, capaz de representar a diversidade de condições operacionais e anomalias. Os eventos são categorizados em nove tipos de falhas, além das operações normais. Para este trabalho, serão utilizadas exclusivamente as instâncias que correspondem a dados reais de poços de petróleo, a fim de garantir a máxima fidelidade ao cenário operacional.

Segundo Vargas et al. (2019), o dataset é dividido em pastas numeradas de 1 a 9, cada uma dedicada a um tipo de evento anômalo específico. Para esta análise, foram selecionados exclusivamente os dados reais referentes à anomalia do tipo 5, "perda rápida de produtividade". Além da categorização por tipo de anomalia, cada evento é classificado com base no estado operacional do poço (aberto ou fechado) e na classe de observação, que indica o nível da anomalia. Essa classificação é dividida em três categorias: normal (0), transiente de anomalia (101 a 109) e estado permanentemente anômalo (1 a 9). O período normal indica a ausência de anomalias, enquanto o período transiente representa a fase de transição entre o funcionamento normal e a estabilização da anomalia. Quando essa fase transitória se encerra, inicia-se o estado estável da anomalia, caracterizando um desvio permanente nas condições operacionais do poço. Essa classificação, presente na coluna "class", é a base para a rotulagem das janelas de dados como anômalas ou normais.

O Poço 16 (WELL-00016_20180426131710.parquet) foi utilizado para o treinamento dos modelos, enquanto o Poço 20 (WELL-00020_20140319110000.parquet) foi reservado como um conjunto de dados totalmente independente para a validação final.

3.2 MODELOS

Para a classificação das anomalias, foram utilizados dois modelos de aprendizado de máquina: a Árvore de Decisão e o Random Forest. A escolha desses modelos foi estratégica. Segundo Breiman *et al*. (1984), a Árvore de Decisão foi selecionada como modelo base por sua alta interpretabilidade, permitindo que as regras de classificação sejam facilmente compreendidas por especialistas do domínio. Em contrapartida, o Random Forest foi escolhido como um modelo de ensemble por sua conhecida robustez e capacidade de mitigar o superajuste (overfitting), um problema comum em árvores de decisão únicas. Conforme Soltan *et al*. (2022), a comparação entre um modelo simples e um de ensemble é uma prática comum para avaliar o ganho de performance em problemas de detecção de anomalias.

3.3 ENGENHARIA DE ATRIBUTOS

Os dados brutos coletados pelos sensores ao longo do tempo são séries temporais de alta frequência, o que representa um desafio para os modelos de aprendizado de máquina. Para que um modelo consiga identificar padrões relevantes, é impraticável analisar cada ponto de dado individualmente. Por isso, a engenharia de atributos foi um passo fundamental para transformar esse fluxo de dados em uma representação mais estruturada e informativa.

Os dados foram segmentados em janelas temporais não sobrepostas de 30 minutos (1800 linhas de dados). Essa janela foi escolhida como um equilíbrio para capturar dinâmicas operacionais significativas sem ser excessivamente influenciada por ruídos de curto prazo.

Para cada janela temporal e para cada sensor, foram calculados quatro atributos estatísticos: a média (mean), o desvio padrão (std), o valor mínimo (min) e o valor máximo (max). Esta abordagem condensa os 1800 pontos de dados de cada sensor em um resumo conciso do seu comportamento no período, gerando features que representam a tendência central, a dispersão e a amplitude do sinal.

A rotulagem de cada janela foi definida com base no valor máximo da coluna "class" dentro daquele intervalo. Se qualquer observação dentro da janela

apresentasse um código de anomalia (neste caso, valores 5 para anomalia estável ou 105 para transiente), a janela inteira era rotulada como anômala (is_anomaly = 1). Caso contrário, era considerada normal (is_anomaly = 0). Por fim, foram removidas as colunas de sensores que não apresentavam dados para a anomalia do tipo 5, garantindo que os modelos fossem treinados apenas com informações pertinentes.

Também foram retiradas as colunas que não eram aplicáveis ao problema 5, ou seja, as colunas que não apresentavam valores para o parquet (arquivo com os dados de base) utilizado.

3.4 PRÉ-PROCESSAMENTO E REDUÇÃO DE DIMENSIONALIDADE COM PCA

Foi realizada uma interseção entre os conjuntos de dados dos dois poços para garantir que apenas os sensores presentes e com dados completos em ambos fossem utilizados, resultando em um conjunto comum de atributos para treinamento e validação.

Em seguida, a técnica VarianceThreshold foi aplicada com um limiar de 0.0 para remover quaisquer atributos que não apresentassem variação ao longo do conjunto de dados de treinamento, pois estes não carregam informação discriminatória.

Após a engenharia de atributos, o desafio consistia em lidar com um conjunto de 44 características distintas. Com isso, os dados foram normalizados utilizando o StandardScaler, que padroniza as features removendo a média e escalonando para a variância unitária. Este passo é fundamental, pois a Análise de Componentes Principais (PCA) é sensível à escala das variáveis, e a normalização garante que atributos com maiores amplitudes numéricas não dominem a análise.

A Análise de Componentes Principais (PCA) foi aplicada para reduzir a dimensionalidade do conjunto de 44 features. O algoritmo foi configurado para reter o número mínimo de componentes principais que explicassem 95% da variância total dos dados (n_components=0.95). Seguindo as boas práticas para evitar vazamento de dados (data leakage), o modelo PCA foi ajustado (fit) exclusivamente nos dados de treino do Poço 16, e essa mesma transformação foi subsequentemente aplicada aos dados de validação do Poço 20.

Os componentes principais resultantes, que representam de forma compacta a informação dos sensores, foram então utilizados como a entrada final para os modelos de classificação Random Forest e Árvore de Decisão.

3.5 PIPELINE DE IMPLEMENTAÇÃO

Para garantir a reprodutibilidade e a integridade metodológica da análise, foi desenvolvido um pipeline de implementação estruturado, seguindo as melhores práticas para projetos de aprendizado de máquina. O fluxo de trabalho foi dividido nas seguintes etapas:

a) Definição dos Conjuntos de Dados;
   - Foram designados dois poços distintos para as fases de treinamento e validação. O Poço 16 foi utilizado para o desenvolvimento do modelo (treinamento e validação interna), enquanto o Poço 20 foi reservado como um conjunto de dados completamente independente para a validação externa, testando a capacidade de generalização do modelo em um cenário operacional diferente;

b) Pré-processamento e Engenharia de Atributos;
   - Cada arquivo de dados (referente a um poço) foi carregado e segmentado em janelas temporais de 30 minutos (1800 linhas);
   - Para cada janela, foram extraídos os atributos estatísticos (média, desvio padrão, mínimo e máximo) de cada sensor;
   - A rotulagem de cada janela como anômala (1) ou normal (0) foi realizada com base no valor máximo da coluna "class";
   - As colunas de sensores não aplicáveis à anomalia do tipo 5 foram removidas;

c) Preparação dos Dados para Modelagem (exclusivamente no Poço 16);
   - Divisão Treino-Teste: Os dados do Poço 16 foram divididos em 70% para treinamento e 30% para teste interno utilizando a função train_test_split;
   - Balanceamento de Classes: Devido à natureza rara dos eventos de falha, o conjunto de treinamento era altamente desbalanceado. Para mitigar o risco de o modelo se tornar enviesado para a classe majoritária

(operação normal), foi aplicada a técnica de oversampling com o RandomOverSampler no conjunto de treinamento. Essa técnica equaliza a distribuição de classes criando cópias aleatórias das amostras da classe minoritária.

- Normalização e Redução de Dimensionalidade: Os dados de treinamento foram normalizados com o StandardScaler e, em seguida, o PCA foi ajustado para reter 95% da variância. A mesma transformação (normalização e PCA) foi aplicada ao conjunto de teste do Poço 16 e, posteriormente, a todo o conjunto de dados do Poço 20;

d) Treinamento e Validação dos Modelos;

- Os modelos Random Forest e Árvore de Decisão foram treinados individualmente utilizando os dados de treinamento processados do Poço 16 (features X_train_bal e rótulos y_train_bal);
- O Modelo Random Forest foi configurado com 100 árvores (n_estimators=100) e ambos os modelos foram configurados com um random_state=42 para garantir a reprodutibilidade dos resultados;
- O desempenho foi avaliado em duas fases: primeiro, na validação interna com os dados de teste do Poço 16; segundo, na validação externa com os dados do Poço 20;

e) Métricas de Avaliação;

- Acurácia: Percentual geral de acertos;
- F1-Score: Média harmônica entre precisão e recall, ideal para dados desbalanceados;
- ROC AUC: Mede a capacidade do modelo de distinguir entre as classes;
- Especificidade (SPE): Capacidade de identificar corretamente os casos normais (verdadeiros negativos);
- Taxa de Falsos Alarmes (FAR): Proporção de casos normais classificados incorretamente como anomalias. A FAR é uma métrica de grande importância operacional, pois um valor alto pode inviabilizar a implementação do sistema;

3.6 TIPO DE ANOMALIA SELECIONADA

Os algoritmos serão treinados para identificar o tipo de problema cinco especificado por Vargas et.al. (2019). A anomalia em questão é referente a perda rápida de produtividade que pode ser causada por diferentes motivos como perda de pressão do poço, mudança de viscosidade do fluido de produção, entre outros fatores.

Dessa forma, quando essas mudanças de propriedades do poço mudam, o sistema pode não ter energia suficiente para manter a produção. Sendo assim, é importante classificar esse tipo de anomalia de maneira rápida para que o time de operações possa direcionar esforços para corrigir e analisar os problemas.

Dentro deste problema, os poços a serem analisados serão o 16 e o 20. Sendo eles os que contêm a menor quantidade de valores em branco entre todo o dataset.

4 DISCUSSÃO DOS RESULTADOS

A avaliação dos modelos de classificação foi realizada em duas etapas distintas: uma validação interna, utilizando o conjunto de teste do mesmo poço usado para treinamento (Poço 16), e uma validação externa, aplicando os modelos treinados aos dados de um poço distinto (Poço 20).

4.1 PRINCIPAL COMPONENTE ANALYSIS (PCA)

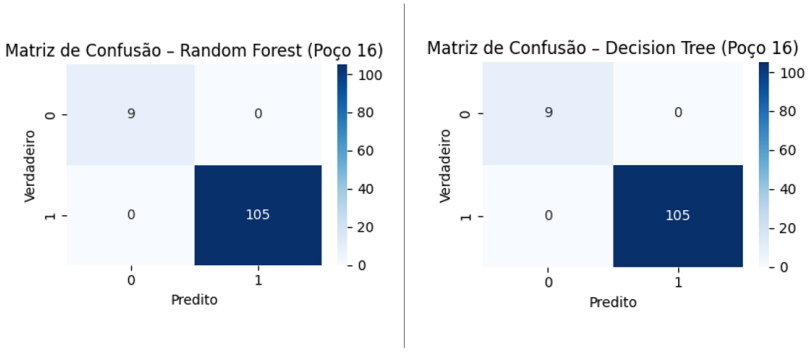
A aplicação do PCA resultou em uma redução dimensional expressiva. O conjunto de 44 *features* de entrada, obtido após a remoção de atributos com variância nula, foi condensado com sucesso em apenas 10 componentes principais.

A análise da variância explicada acumulada confirmou a eficácia do método. Conforme os resultados, o décimo componente foi o necessário para ultrapassar o limiar estabelecido, alcançando um total de 95,84% da variância contida nos dados originais. Dessa forma, o PCA permitiu a criação de um conjunto de features mais enxuto e computacionalmente eficiente para o treinamento dos modelos de classificação, minimizando a perda de informação relevante.

4.2 DESEMPENHO NA VALIDAÇÃO INTERNA (POÇO 16)

A partir das FIGURA 4 observa-se a matriz de confusão para os poços 16, para cada um dos algoritmos utilizados.

FIGURA 4 – MATRIZ DE CONFUSÃO PARA O POÇO 16



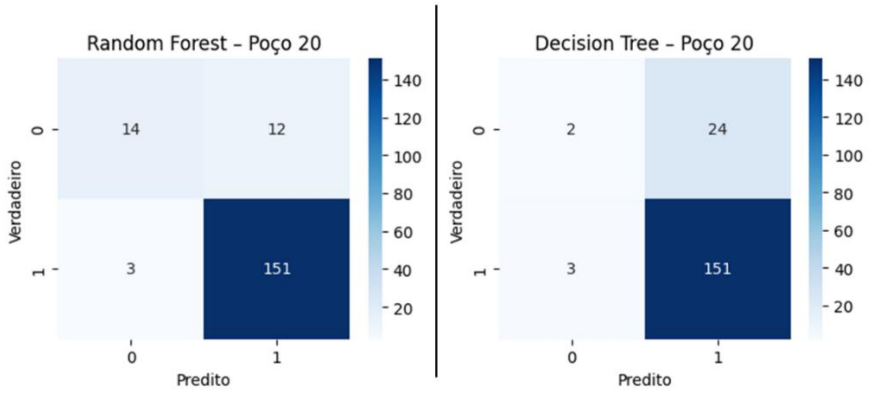
FONTE: O autor (2025)

Como é possível observar através da matriz de confusão, a taxa de acerto dos dois modelos foi de 100%. Embora ideal na teoria, esse desempenho perfeito em um ambiente de teste interno levanta uma suspeita de superajuste (overfitting). Este cenário pode ser comparado a um aluno que decorou as respostas exatas de uma prova antiga: ele gabaritou o simulado, mas isso não garante que ele tenha aprendido o conteúdo para se sair bem em uma prova nova e desconhecida. Por isso, a validação externa no Poço 20 é o teste definitivo da capacidade de generalização dos modelos.

4.3 DESEMPENHO NA VALIDAÇÃO EXTERNA (POÇO 20)

A validação externa, aplicando os modelos treinados no Poço 16 aos dados do Poço 20, revelou diferenças significativas de desempenho, conforme resumido na TABELA 1 e na FIGURAS FIGURA 5 FIGURA 6.

FIGURA 5 – MATRIZ DE CONFUSÃO PARA O POÇO 20



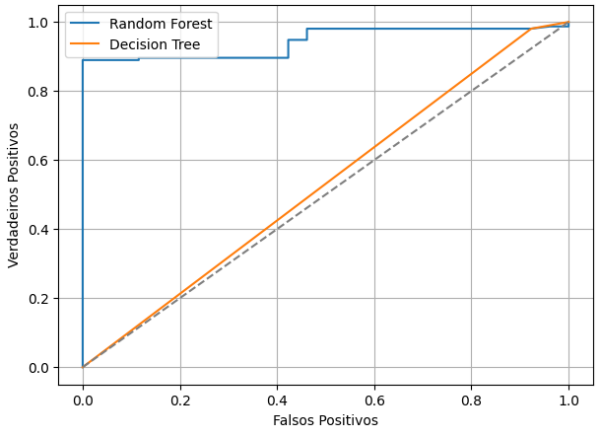
FONTE: O autor (2025)

TABELA 1 – RESULTADO DAS MÉTRICAS UTILIZADAS NO POÇO 20

Modelos	Acurácia	F1 Score	AUC	Especificidade	FAR
Random Forest	0.9167	0.9527	0.9432	0.5385	0.4615
Decision Tree	0.85	0.9179	0.5287	0.0769	0.9231

FONTE: O autor (2025)

FIGURA 6 – CURVA ROC PARA O POÇO 20



FONTE: O autor (2025).

De modo geral, o modelo Random Forest obteve resultado mais elevados em todas as métricas se comparadas ao modelo Árvore de Decisão (Decision Tree). A acurácia, representando a porcentagem total de previsões que o modelo acertou, foi superior para o Random Forest, bem como o F1 Score que para o caso de classes desbalanceadas como o caso deste trabalho, mostra a habilidade do modelo em não

criar alarmes falsos (Precisão) e encontrar todos os casos positivos para anomalia (Recall).

Já a curva ROC que plota a "Taxa de Verdadeiros Positivos" (Recall) contra a "Taxa de Falsos Positivos" (FAR) em vários limiares de classificação e a AUC que é a área sob essa curva, como mostrado na FIGURA 6, indica grande capacidade do modelo Random Forest em distinguir as classes positivas e negativas. A ROC AUC funciona como um "medidor de confiança" e o valor de 0.9432 do Random Forest indica que ele é extremamente confiável em sua tarefa de separar um poço saudável de um poço em falha. Em contrapartida, com o valor da AUC de 0.5287, o modelo de Árvore de Decisão é o equivalente a jogar uma moeda para adivinhar o resultado: um acerto por puro acaso.

Por fim, a especificidade, ou seja, o indicador que mede de todos os casos que eram realmente negativos (classe 0), qual a porcentagem que o modelo conseguiu identificar corretamente, mostrou que o Random Forest conseguiu identificar aproximadamente 54% dos casos negativos, ou seja, os casos que não apresentavam anomalia. A importância da Taxa de Falsos Alarmes (FAR) é imensa na prática: uma FAR alta, como os 92% da Árvore de Decisão, a torna inútil, pois seria como um alarme de incêndio que dispara a cada poucos minutos sem motivo – todos o ignorariam. O Random Forest, com uma FAR de 46%, ainda gera alarmes falsos, mas ao reduzir o "barulho" em quase metade, torna-se uma ferramenta de monitoramento muito mais viável e confiável.

5 CONSIDERAÇÕES FINAIS

Em suma, este trabalho demonstrou que é possível construir um "vigia digital" inteligente e confiável para poços de petróleo, que aprende com os dados de um poço para monitorar outro. A análise comparativa revelou que a escolha da arquitetura correta – neste caso, o modelo de ensemble Random Forest – é crucial para que este vigia seja eficaz e não gere um volume excessivo de alarmes falsos, representando uma solução robusta para a detecção de anomalias em cenários operacionais reais.

5.1 PERFORMANCE

Na validação interna, observou-se que tanto o modelo Random Forest quanto a Árvore de Decisão alcançaram um desempenho perfeito, com todas as métricas de avaliação atingindo o valor máximo de 1.0. A análise da matriz de confusão para o Poço 16 (Figura 2) corrobora este resultado, não exibindo nenhum erro de classificação. Embora este cenário pareça ideal, ele sugere que os modelos possuem alta capacidade de se ajustar aos padrões específicos dos dados de treinamento, levantando a hipótese de superajuste (*overfitting*). Este achado reforça a necessidade da validação externa como um teste definitivo para avaliar a capacidade de generalização dos modelos em um ambiente operacional diferente, o que foi feito utilizando o poço 20.

No cenário do Poço 20, o Random Forest foi superior ao modelo Decision Tree. Enquanto o Decision Tree se mostrou "viciado" em prever a classe majoritária (classe 1), o Random Forest conseguiu um desempenho muito mais equilibrado, sendo eficaz em ambas as classes.

O diferencial mais evidente foi seu poder de discriminação, medido pela métrica ROC AUC, que atingiu o valor de 0.9432. Este resultado mostra uma excelente capacidade do modelo em distinguir corretamente entre as classes, contrastando drasticamente com o desempenho da Árvore de Decisão (0.5287), que se mostrou marginalmente superior a uma classificação aleatória.

Ainda, o Random Forest exibiu maior robustez na identificação da classe negativa. Sua especificidade de 0.5385, embora moderada, foi muito superior à da Árvore de Decisão (0.0769), que se mostrou praticamente incapaz de generalizar para este grupo. Em conjunto, esses resultados demonstram o Random Forest como um

algoritmo mais estável e com maior poder de generalização, representando a escolha mais segura para implementação.

Além disso, A superioridade do modelo Random Forest em relação à Árvore de Decisão, conforme observado na análise, é uma consequência direta da sua arquitetura de ensemble. Árvores de Decisão individuais são suscetíveis à alta variância e ao superajuste (overfitting) (BRAIMAN, 2001), pois tendem a se adaptar excessivamente às particularidades do conjunto de treinamento. O Random Forest mitiga essa limitação ao construir múltiplas árvores sobre diferentes subamostras dos dados (técnica de bagging) (BRAIMAN, 1996) e ao restringir as divisões em cada nó a um subconjunto aleatório de features. Este processo de dupla aleatorização gera árvores descorrelacionadas, cujos erros individuais são compensados quando as previsões são agregadas por um sistema de votação. O resultado é um modelo final com menor variância, maior robustez e uma capacidade de generalização significativamente superior para dados não vistos, justificando seu desempenho mais consistente e confiável na presente análise.

5.2 ANÁLISE ECONÔMICA

Visando um possível ganho ao implementar o algoritmo Random Forest ao sistema da Petrobras, segue a análise econômica.

Para quantificar o ganho, parte-se do prejuízo de US$ 75,7 milhões anuais associado aos eventos indesejáveis (VARGAS, 2019). Embora os dados não detalhem a contribuição exata da anomalia "perda rápida de produtividade" para este total, é razoável assumir que ela represente uma fração significativa, visto que impacta diretamente o volume produzido.

Considerando um cenário hipotético conservador em que a "perda rápida de produtividade" seja responsável por 15% do total de perdas anuais, o impacto financeiro desta anomalia seria de:

$$0{,}15 \times US\$\ 75.700.000{,}00 = US\$\ 11.355.000{,}00$$

A implementação de um sistema de detecção eficaz como o modelo Random Forest proposto, que demonstrou alta performance em um cenário de validação realista, permite uma intervenção precoce que pode mitigar ou reverter a maior parte

dessas perdas. Se a atuação ágil, possibilitada pelo alerta do sistema, conseguir prevenir 80% das perdas associadas a esta anomalia específica, o ganho econômico anual seria de:

$$0{,}8 \times US\$\ 11.355.000{,}00 = US\$\ 9.084.000{,}00$$

Contudo, ao analisar a Taxa de Falsos Alarmes (FAR), que para o modelo Random Forest foi de 0.4615, observa-se um valor considerado alto para um ambiente operacional, embora represente uma melhoria drástica em relação à Árvore de Decisão (FAR de 0.9231). Ele implica que quase metade dos alertas gerados durante a operação normal do poço seriam, na verdade, falsos. Este fator introduz um novo custo operacional significativo, o custo de verificação: Cada alarme falso exige tempo e recursos da equipe de operações para investigar a sua causa, desviando a atenção de outras tarefas e gerando um custo de mão de obra.

Portanto, o desafio não é apenas detectar a falha corretamente, mas fazê-lo com um nível de ruído (alarmes falsos) operacionalmente gerenciável. A quantificação do ganho deve agora balancear o potencial de economia pela detecção correta de anomalias com os custos gerados pelos falsos alarmes.

Para estimar este custo, assume-se que um poço opera em estado normal durante 99% do tempo ao longo de um ano (Considerando o evento anômalo de perda rápida de produtividade como sendo raro, aparecendo em menos de 1% de todos os dados coletados, segundo Vargas et.al. (2019)). Com janelas de análise de 30 minutos, um ano possui 17.520 janelas. O número de falsos alarmes seria:

$$(Janelas\ por\ ano) \times (\%\ Operação\ Normal) \times (FAR)$$

$$(17.520\ ,00) \times (0.99) \times (0.4615) \approx 8.000\ falsos\ alarmes\ por\ ano$$

Assumindo um custo hipotético de US$ 255 por falso alarme para cobrir a análise e verificação pela equipe de engenharia, sendo esse custo de US$ 85/hora de salário de um engenheiro e 3 horas de serviço, o custo anual é de:

$$(8.000{,}00) \times (US\$\ 255{,}00) = US\$\ 2.040.000{,}00$$

Desta forma, o ganho líquido é a diferença entre a economia bruta e o custo dos falsos alarmes:

$$US\$9.084.000{,}00 \; - \; US\$\; 2.040.000{,}00 = \; US\$\; 6.964.000{,}00$$

Assim, mesmo considerando o impacto operacional negativo da alta taxa de falsos alarmes, a implementação do modelo Random Forest ainda apresenta um ganho econômico líquido potencial estimado em aproximadamente US$ 7 milhões anuais. Este resultado evidencia que, embora o modelo não seja uma solução perfeita, ele transforma um sistema de detecção praticamente inutilizável (Árvore de Decisão) em uma ferramenta de alto valor agregado, cujo benefício na prevenção de perdas de produção supera significativamente os custos operacionais associados às suas imperfeições.

5.3 SUGESTÃO PARA TRABALHOS FUTUROS

Embora os resultados sejam promissores, existem diversas oportunidades de aprimoramento e expansão da metodologia proposta. Nesse sentido, trabalhos futuros poderiam explorar técnicas de balanceamento de dados mais avançadas. Em substituição ao *RandomOverSampler* utilizado neste pipeline, a implementação de métodos como o SMOTE, que cria instâncias sintéticas, poderia refinar a capacidade do modelo de generalizar, visando elevar a taxa de verdadeiros negativos.

Além disso, outra frente de pesquisa consiste na exploração de algoritmos mais complexos como redes neurais LSTM (Long Short-Term Memory) e na otimização da engenharia de atributos. A engenharia de atributos, que neste trabalho utilizou a extração de features estatísticas e a redução via PCA, poderia ser enriquecida com a extração de features temporais mais sofisticadas, como tendências, e o uso de métodos de seleção que preservem a interpretabilidade dos resultados, como o RFE (Recursive Feature Elimination).

Por fim, é possível ampliar o escopo da validação para além da abordagem adotada neste trabalho, que treinou o modelo com dados de um poço e o validou com sucesso em outro levando em consideração o mesmo tipo de anomalia. Uma expansão dessa metodologia seria a implementação de um esquema de validação cruzada entre múltiplos poços, o que forneceria uma medida mais rigorosa da

performance. Seria igualmente crucial testar a eficácia do pipeline nos outros oito tipos de falhas catalogados no dataset 3W. Essa validação mais abrangente permitiria verificar a adaptabilidade e a escalabilidade da solução para diferentes cenários operacionais, contribuindo de maneira ainda mais significativa para a segurança e eficiência dos processos na indústria petrolífera.

Em suma, este trabalho confirma o imenso potencial do aprendizado de máquina como ferramenta para a otimização e segurança na indústria de óleo e gás. O modelo Random Forest demonstrou ser uma solução robusta e confiável, e as direções apontadas para trabalhos futuros abrem um leque de oportunidades para a criação de sistemas de monitoramento ainda mais inteligentes e eficazes.

REFERÊNCIAS

AJAYI, Abisoye et al. Using AI and machine learning to predict and mitigate cybersecurity risks in critical infrastructure. **International Journal of Engineering Research and Development**, [S. l.], v. 21, n. 2, p. 205-224, fev. 2025.

ALBANO, W. A.; NOGUEIRA, M.; SOUZA, J. N. de. A taxonomy for resilience in vehicular ad hoc networks. **IEEE Latin America Transactions**, Piscataway, v. 13, n. 1, p. 228-234, jan. 2015.

ALJAMEEL, S. S. et al. An anomaly detection model for oil and gas pipelines using machine learning. **Computation**, Basel, v. 10, n. 138, ago. 2022. DOI: 10.3390/computation10080138.

ANZAI, T. K. et al. Catching failures in 10 minutes: an approach to no code, fast track, AI-based real time process monitoring. *In*: OFFSHORE TECHNOLOGY CONFERENCE BRASIL (OTC), 2023. **Anais**[...] p. D011S004R004.

BERTA, Ramesh. AI in oil and gas: predicting equipment failures and maximizing uptime. **International Journal on Science and Technology (IJSAT)**, v. 12, n. 1, jan./mar. 2021.

BREIMAN, L. et al. **Classification and Regression Trees**. New York: Routledge, 1984.

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

BREIMAN, L. Random forests. **Machine Learning**, v. 45, n. 1, p. 5-32, 2001.

CARNEIRO, Tiago et al. Performance analysis of Google Colaboratory as a tool for accelerating deep learning applications. **IEEE Access**, Piscataway, v. 6, p. 61643-61655, 2018. DOI: 10.1109/ACCESS.2018.2874767.

CHICCO, D.; JURMAN, G. *The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification*. BioData Mining, v.16, art.4, 2023. DOI: 10.1186/s13040-023-00322-4.

DIAS, T. L. B. et al. Development of oilwell fault classifiers using a wavelet-based multivariable approach in a modular architecture. **SPE Journal**, v. 29, p. 4542-4556, 2024.

ESALQ. **Entropia, Ganho de informação e Decision trees**. [S. l.: s. n.], [s.d.]. Disponível em: https://www.esalq.usp.br/lepse/imgs/conteudo_thumb/Entropia--Ganho-de-informa--o-e-Decision-trees.pdf.

EZENWEICHU, C. P. The causes, effects and minimization of formation damage in horizontal wells. **PC**, v. 2, 2015. Disponível em: https://www.vurup.sk/wp-content/uploads/dlm_uploads/2017/07/pc_2_2015_ezenweichu_323_rev.pdf.

FAERGESTAD, I. (ed.). Formation damage. **Oilfield Review**, The Defining Series, [S. l.], 2016. Disponível em: https://www.slb.com/defining. Acesso em: 30 jun. 2025.
FERREIRA, Jonas. Transformação digital na indústria de óleo e gás. **Sydle**, 29 nov. 2024. Disponível em: https://www.sydle.com/br/blog/transformacao-digital-oleo-e-gas-6749b64578f5ea35da1fbc31. Acesso em: 30 jun. 2025.

GATTA, F. et al. Predictive maintenance for offshore oil wells by means of deep learning features extraction. **Expert Systems**, v. 41, n. 2, e13128, 2024.

GOOGLE. *Classification: accuracy, precision, recall, and related metrics*. Machine Learning Crash Course, Google Developers, 2025. Disponível em: https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall. Acesso em: 30 jun. 2025.

GOOGLE. *ROC and AUC*. Machine Learning Crash Course, Google Developers, 2025. Disponível em: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc. Acesso em: 30 jun. 2025.

GUSTINELI, Murilo. Random Forests: algoritmos baseados em árvores. **Brains**, 14 ago. 2023. Disponível em: https://brains.dev/2023/random-forests-algoritmos-baseados-em-arvores/. Acesso em: 30 jun. 2025.

HALIM, M. C.; HAMIDI, H.; AKISANYA, A. R. Minimizing formation damage in drilling operations: a critical point for optimizing productivity in sandstone reservoirs intercalated with clay. **Energies**, Basel, v. 15, n. 1, p. 162, dez. 2021. DOI: 10.3390/en15010162.

HAMMING, Richard W. **Numerical Methods for Scientists and Engineers**. New York: McGraw-Hill, 1962.

JINDAL, H.; SAXENA, S.; SINGH, S. Challenges and issues in underwater acoustics sensor networks: a review. *In*: INTERNATIONAL CONFERENCE ON PARALLEL, DISTRIBUTED AND GRID COMPUTING (PDGC), 9., 2014, Waknaghat. **Anais** [...]. Piscataway: IEEE, 2014. p. 251-255. DOI: 10.1109/PDGC.2014.7000751.

JOLLIFFE, I. T. **Principal Component Analysis**. 2. ed. New York: Springer, 2002.

JOLLIFFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, London, v. 374, n. 20150202, 2016. DOI: 10.1098/rsta.2015.0202.

JUNIOR, W. F. **Comparação de classificadores para detecção de anomalias em poços produtores de petróleo**. 2022. Dissertação (Mestrado em Computação Aplicada) – Instituto Federal do Espírito Santo, Serra, 2022.

JYOTHI, P. N. S.; YAMAGANTI, R. A review on python for data science, machine learning and IOT. **International Journal of Scientific & Engineering Research**, Houston, v. 10, n. 12, p. 851-858, dez. 2019. DOI: 10.13140/RG.2.2.10700.48000.

KHAN, A. et al. A Review on Underwater Data Collection Using Wireless Sensor Networks. **Wireless Communications and Mobile Computing**, 2021. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC8051310/.

KUHN, M.; JOHNSON, K. **Applied Predictive Modeling**. New York: Springer, 2013.
LAU, K. **Random Forest Ensemble Visualization**. 2014. Disponível em: https://www.cs.ubc.ca/~tmm/courses/547-14/projects/ken/report.pdf.

LIU, Y.; ZHOU, Z.; LI, X. **Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions**. New York: Springer, 2012.
LUCIDCHART. O que é uma árvore de decisão. **Lucidchart**, 2024. Disponível em: https://www.lucidchart.com/pages/pt/o-que-e-arvore-de-decisao.

MELO, A.; CÂMARA, M. M.; PINTO, J. C. Data-driven process monitoring and fault diagnosis: a comprehensive survey. **Processes**, v. 12, n. 2, p. 251, 2024.

OHALETE, N. C. et al. Advancements in predictive maintenance in the oil and gas industry: a review of AI and data science applications. **World Journal of Advanced Research and Reviews**, [S. l.], v. 20, n. 3, p. 167-181, 2023. DOI: 10.30574/wjarr.2023.20.3.2432.

PENG, L. et al. Predictive approach to perform fault detection in electrical submersible pump systems. **ACS Omega**, Washington, DC, v. 6, p. 8104-8111, mar. 2021. DOI: 10.1021/acsomega.0c05808.

PETROBRAS. **Relatório de monitoramento operacional dos poços**. 2022. Disponível em: https://agencia.petrobras.com.br/documents/d/agencia/relatorio_producao_e_vendas_-_4t22-pdf. Acesso em: 30 nov. 2024.
POPILIN NETO, M.; PAULOVICH, F. V. Explainable Matrix -- Visualization for Global and Local Interpretability of Random Forest Classification Ensembles. **arXiv:2005.04289**, 2020. Disponível em: https://arxiv.org/abs/2005.04289.

QLIKTECH (Qlik Cloud). *Pontuando modelos de classificação binária*. Qlik Cloud Help. Disponível em: https://help.qlik.com/pt-BR/cloud-services/Subsystems/Hub/Content/Sense_Hub/AutoML/scoring-binary-classification.htm. Acesso em: 20 jun. 2025.

QUINLAN, J. R. Induction of decision trees. **Machine Learning**, v. 1, n. 1, p. 81-106, 1986.

QUINLAN, J. R. **C4.5: Programs for Machine Learning**. San Francisco: Morgan Kaufmann, 1996.

SANTOS, A. C. et al. Ferramenta para Monitoramento em Tempo Real e Detecção de Anomalias em Poços de Petróleo Utilizando Aprendizagem Profunda. *In*: RIO OIL & GAS EXPO AND CONFERENCE, 2020. Disponível em: https://icongresso.ibp.itarget.com.br/arquivos/trabalhos_completos/ibp/3/final.IBP0938_20_27112020_085551.pdf.

SCIKIT-LEARN. **Post pruning decision trees with cost complexity pruning**. 2024. Disponível em: https://scikit-learn.org/stable/auto_examples/tree/plot_cost_complexity_pruning.html.

SKÅLVIK, A. M. et al. Challenges, limitations, and measurement strategies to ensure data quality in deep-sea sensors. **Frontiers in Marine Science**, [S. l.], v. 10, p. 1152236, abr. 2023. DOI: 10.3389/fmars.2023.1152236.

TAHA, M. H. N.; MANSOUR, A. A Review of Predictive Analytics Models in the Oil and Gas Industries. **Sensors**, v. 24, n. 12, p. 4013, 2024. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC11207882/.

TECHTAC. **Pressure Drops in Oil Wells: Understanding the Causes & Impacts**. 2023. Disponível em: https://www.techtac.com/understanding-the-causes-and-impacts-of-pressure-drops-in-oil-wells/. Acesso em: 30 jun. 2025.

TURAN, E. M.; JÄSCHKE, J. Classification of undesirable events in oil well operation. *In*: INTERNATIONAL CONFERENCE ON PROCESS CONTROL (PC), 23., 2021, Strbske Pleso. **Anais**[...]. New York: IEEE, 2021. p. 157-162.

TURNER, Shruti. *Specificity and sensitivity in data science*. Medium (Trusted Data Science), 17 jun. 2024. Disponível em: https://medium.com/trusted-data-science-haleon/specificity-and-sensitivity-in-data-science-9d0c19b102ab. Acesso em: 20 jun. 2025.

VARGAS, Ricardo Emanuel Vaz et al. A realistic and public dataset with rare undesirable real events in oil wells. **Journal of Petroleum Science and Engineering**, v. 181, p. 106223, 2019.

WELCOME to Colaboratory. [S. l.]: Google, [2022]. Disponível em: https://colab.research.google.com/. Acesso em: 30 jun. 2025.

WU, Shan et al. Combining acoustic emission and unsupervised machine learning to investigate microscopic fracturing in tight reservoir rock. **Engineering Geology**, v. 347, p. 107939, 2025.

YOU, Z. et al. Well Productivity Decline During Oil & Water Production due to Fines Migration. **SPE-221295-MS**, 2024. Disponível em: https://www.researchgate.net/publication/384840482_Well_Productivity_Decline_During_Oil_Water_Production_due_to_Fines_Migration

APÊNDICE 1 – CÓDIGO EM PYTHON

https://drive.google.com/file/d/1hkyIWiEDyvAe1UcwK_CWFE7GROrA1wU9/view?usp=sharing