

Laura Pellizari Pereira

AutoDrill: *Framework* AutoML para Previsão da Taxa de Perfuração em Poços de Petróleo

Brasil

Junho de 2026

Laura Pellizari Pereira

AutoDrill: *Framework* AutoML para Previsão da Taxa de Perfuração em Poços de Petróleo

Dissertação submetida ao Programa de Pós-Graduação em Ciência e Tecnologia da Computação da UNIFEI (área de concentração: Matemática da Computação), como parte dos requisitos necessários para a obtenção do Título de Mestre em Ciência e Tecnologia da Computação.

Universidade Federal de Itajubá – UNIFEI

Programa de Pós-Graduação em Ciência e Tecnologia Computação

Orientador: Prof. Dr. João Paulo R. R. Leite

Coorientador: Prof. Dr. Andreas Nascimento

Brasil

Junho de 2026

*Este trabalho é dedicado àqueles que
enxergam no conhecimento uma forma de liberdade.*

Agradecimentos

Agradeço, primeiramente, ao meu orientador, Prof. Dr. João Paulo Ribeiro Rodrigues Leite, pelo constante apoio, orientação e incentivo ao longo do desenvolvimento deste trabalho. Sua experiência, dedicação e disponibilidade foram fundamentais para a realização desta pesquisa.

Agradeço também ao meu coorientador, Prof. Dr. Andreas Nascimento, pelas valiosas contribuições e discussões, que enriqueceram significativamente este estudo.

À Universidade Federal de Itajubá e ao Programa de Pós-Graduação em Ciência e Tecnologia da Computação, agradeço pela oportunidade e pelos recursos disponibilizados para a realização deste mestrado.

Aos meus colegas e amigos, que compartilharam comigo momentos de aprendizado, estudo e reflexão, tornando essa jornada mais leve e enriquecedora.

Por fim, agradeço à minha família, pelo amor, apoio, paciência e incentivo constantes, sem os quais este trabalho não teria sido possível.

*“Whatever we do, we have to make our life vectors.
Lines with force and direction.”
Liam Neeson*

Resumo

A predição da Taxa de Perfuração (ROP, do inglês *Rate of Penetration*) desempenha papel central na otimização de operações na indústria de óleo e gás, influenciando diretamente a eficiência operacional e os custos associados à perfuração. Entretanto, modelos de aprendizado de máquina apresentam limitações nesse contexto em razão da forte não-linearidade das variáveis envolvidas e, sobretudo, da elevada heterogeneidade litológica e da baixa confiabilidade de sensores de fundo de poço, que resultam em séries temporais ruidosas e com lacunas. Neste trabalho, propõe-se o AutoDrill, um *framework* de *Automated Machine Learning* (AutoML) especificamente projetado para o domínio de perfuração. Em contraste com abordagens generalistas, o AutoDrill incorpora um *pipeline* de pré-processamento especializado em sinais de perfuração (incluindo suavização adaptativa, correção de transientes e interpolação de dados) acoplado a um mecanismo automatizado de busca e ajuste de hiperparâmetros. A avaliação experimental foi conduzida em diversos conjuntos de dados e cenários geológicos, utilizando registros provenientes de três repositórios *offshore* (Equinor Volve, USROP e NEATS NOPIMS). O desempenho do AutoDrill foi comparado aos *frameworks* AutoGluon e H2O AutoML, considerados representativos do estado da arte em AutoML. Os resultados indicam que o AutoDrill apresenta desempenho consistente e competitivo, alcançando coeficiente de determinação (R^2) médio de 0,91 ao longo de todos os cenários analisados, com destaque para ambientes caracterizados por elevada heterogeneidade e ruído, como no caso do NEATS NOPIMS. Nesse cenário específico, o AutoDrill superou de forma significativa os *frameworks* de referência, reduzindo o erro absoluto médio (MAE) em até 80% e elevando o R^2 de níveis inferiores a 0,20 para 0,94. Conclui-se que o conhecimento de domínio é determinante para a melhoria da predição em tempo real. Nessa perspectiva, o AutoDrill configura-se como uma abordagem prática, escalável e orientada ao domínio para a predição de ROP, contribuindo para o avanço da aplicação de técnicas de aprendizado de máquina na engenharia de petróleo.

Palavras-chave: AutoML, Previsão, ROP, Aprendizado de Máquina, Perfuração de Poços, Petróleo, Otimização.

Abstract

Predicting the Rate of Penetration (ROP) plays a central role in optimizing operations in the oil and gas industry, directly influencing operational efficiency and drilling-related costs. However, conventional machine learning models have limitations in this context due to the strong non-linearity of the variables involved and, above all, the high lithological heterogeneity and low reliability of downhole sensors, resulting in noisy and gapped time series. This work proposes AutoDrill, an Automated Machine Learning (AutoML) framework specifically designed for the drilling domain. In contrast to generalist approaches, AutoDrill incorporates a preprocessing pipeline specialized in drilling signals (including adaptive smoothing, transient correction, and data interpolation) coupled with an automated hyperparameter search and tuning mechanism. The experimental evaluation was conducted on various datasets and geological scenarios, using records from three offshore repositories (Equinor Volve, USROP, and NEATS NOPIMS). The performance of AutoDrill was compared to the AutoGluon and H2O AutoML frameworks, considered representative of the state of the art in AutoML. The results indicate that AutoDrill exhibits consistent and competitive performance, achieving an average coefficient of determination (R^2) of 0.91 across all analyzed scenarios, particularly in environments characterized by high heterogeneity and noise, such as in the case of NEATS NOPIMS. In this specific scenario, AutoDrill significantly outperformed the reference frameworks, reducing the mean absolute error (MAE) by up to 80% and raising the R^2 from levels below 0.20 to 0.94. In conclusion, the systematic incorporation of domain knowledge into the machine learning automation cycle is crucial for the robustness of real-time prediction. From this perspective, AutoDrill presents itself as a practical, scalable, and domain-oriented approach for ROP prediction, contributing to the advancement of machine learning techniques in petroleum engineering.

Keywords: AutoML, Forecasting, Rate of Penetration (ROP), Machine Learning, Well Drilling, Petroleum Engineering, Optimization.

Lista de ilustrações

Figura 1 – Fluxo Ilustrativo de uma Random Forest. Figura do autor.	26
Figura 2 – Fluxo de Treinamento GBR. Figura do autor.	27
Figura 3 – Componentes de um sistema AutoML. Figura do autor.	32
Figura 4 – Exemplo de sinal ruidoso e suavizado com filtro de <i>Savitzky-Golay</i> . Figura do autor.	36
Figura 5 – Arquitetura geral do sistema <i>AutoDrill</i> . Figura do autor.	42
Figura 6 – Sinal original e suavizado (<i>smooth</i>) de WOB. Figura do autor.	47
Figura 7 – Sinal original e após correção de (<i>spikes</i>) de RPM. Figura do autor. . .	49

Lista de tabelas

Tabela 1	–	Desempenho de técnicas de aprendizado de máquina na previsão de ROP.	19
Tabela 2	–	Estatísticas das principais variáveis do conjunto de dados <i>Equinor Volve</i> .	44
Tabela 3	–	Estatística das principais variáveis do conjunto de dados <i>USROP</i> . . .	44
Tabela 4	–	Estatística das principais variáveis do conjunto de dados <i>NEATS NOPIMS</i> .	45
Tabela 5	–	Comparação do desempenho de <i>frameworks</i> de AutoML em diferentes conjuntos de dados e litologias, agrupados pelo método AutoML. Valores em negrito indicam a melhor métrica (mais baixa) em geral, exceto para R^2 , em que é destacado o valor mais alto.	57
Tabela 6	–	Comparação de modelos de predição de ROP presentes na literatura. .	58

Lista de abreviaturas e siglas

AutoML	Automated Machine Learning
BDIA	Bit Diameter
BYM	Bourgoyne & Young
DEPT	Depth
ET	Extra Trees Regressor
GBR	Gradient Boosting Regressor
KNN	K-Nearest Neighbors
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MFIA	Mud Flow In Area
ML	Machine Learning
MSE	Mean Squared Error
R^2	Coefficiente de Determinação
RF	Random Forest
RMSE	Root Mean Squared Error
RNA	Redes Neurais Artificiais
ROP	Rate of Penetration
RPM	Rotations Per Minute
SVM	Support Vector Machine
SVR	Support Vector Regressor
TTR	Transformed Target Regressor
WOB	Weight on Bit
XGB	XGBoost Regressor

Sumário

1	INTRODUÇÃO	12
1.1	Contextualização	12
1.2	Objetivos	13
1.2.1	Objetivo Geral	13
1.2.2	Objetivos Específicos	14
1.3	Justificativa	14
1.4	Estrutura do Trabalho	15
2	REVISÃO BIBLIOGRÁFICA	16
2.1	Predição da Taxa de Perfuração com <i>Machine Learning</i>	16
2.2	Técnicas de AutoML	19
2.3	Desafios e Oportunidades	20
2.4	Considerações Finais	22
3	FUNDAMENTAÇÃO TEÓRICA	23
3.1	Aprendizado de Máquina	23
3.1.1	K-Nearest Neighbors (KNN)	24
3.1.2	Random Forest (RF)	25
3.1.3	Gradient Boosting Regressor (GBR)	26
3.1.4	XGBoost Regressor (XGB)	28
3.1.5	Extra Trees Regressor (ET)	28
3.1.6	Support Vector Regressor (SVR)	29
3.1.7	Transformed Target Regressor com ExtraTrees	30
3.2	AutoML	31
3.2.1	AutoGluon	32
3.2.2	H2O	33
3.3	Filtro de sinais	35
3.3.1	Filtro de Savitzky-Golay	35
3.4	Método de Tukey para Detecção de Outliers	37
3.5	Avaliação do Desempenho dos Modelos	38
3.5.1	Coeficiente de determinação (R^2)	39
3.5.2	Erro Absoluto Médio (MAE)	39
3.5.3	Erro Quadrático Médio (MSE)	39
3.5.4	Raiz do Erro Quadrático Médio (RMSE)	40
3.5.5	Erro Percentual Absoluto Médio (MAPE)	40

4	MATERIAIS E MÉTODOS	41
4.1	Arquitetura do AutoDrill	41
4.1.1	Conjunto de Dados	41
4.1.1.1	Conjunto de Dados <i>Equinor Volve</i>	43
4.1.1.2	Conjunto de Dados <i>USROP</i>	43
4.1.1.3	Conjunto de Dados <i>NEATS NOPIMS</i>	44
4.1.2	Pré-Processamento de Dados	45
4.1.3	Filtragem Digital e Suavização de Sinais	46
4.1.4	Detecção e Correção de Descontinuidades (<i>Spikes</i>)	48
4.1.5	Otimização e Seleção de Modelos	49
4.2	Estratégia de Testes e Validação Experimental	51
4.2.1	Cenários de Teste	51
4.2.2	Análise Comparativa com Soluções AutoML Existentes	52
4.2.3	Métricas de Avaliação	52
4.3	Aspectos de Implementação e Reprodutibilidade	53
4.3.1	Ferramentas e Ambiente	53
4.3.2	Limpeza e Pré-processamento de Dados	53
4.3.3	Configuração do AutoML e Busca de Hiperparâmetros	54
4.3.4	Reprodutibilidade	55
5	RESULTADOS E DISCUSSÃO	56
5.1	Avaliação Comparativa com AutoML Existentes	56
5.2	Comparação com a Literatura	58
5.3	Análise e Discussão	59
5.4	Principais Contribuições	60
6	CONCLUSÕES	62
6.1	Trabalhos Futuros	63
	REFERÊNCIAS	64

1 Introdução

1.1 Contextualização

As operações de perfuração na indústria de petróleo e gás representam uma parcela significativa dos custos de exploração e produção, especialmente em ambientes *offshore*. Fatores como a localização do poço, profundidade e complexidade geológica podem aumentar consideravelmente os gastos, sendo que poços *offshore* frequentemente apresentam custos de três a cinco vezes superiores aos de poços equivalentes em terra firme (BRET-ROUZAUT, 2022). Os custos diários de perfuração podem ultrapassar os US\$ 500.000, a depender das especificações e requisitos de cada projeto (TUNKIEL; SUI; WIKTORSKI, 2022). Esses elevados custos operacionais evidenciam a necessidade premente otimização da eficiência do processo de perfuração visando a redução do tempo não produtivo (NPT) e, conseqüentemente, o aumento da lucratividade geral dos projetos (NASCIMENTO; NOGUEIRA, 2010).

Entre os indicadores de desempenho disponíveis em operações de perfuração, a Taxa de Penetração (ROP) é particularmente importante, pois quantifica a velocidade com que a broca avança através das formações rochosas. Geralmente, valores mais elevados de ROP tendem a se traduzir em menor tempo de perfuração e, portanto, em menores custos operacionais. Por este motivo, a previsão ROP torna-se importante tanto para o planejamento quanto para a tomada de decisões em tempo real durante as atividades de perfuração. Entretanto, a estimativa da ROP com precisão continua sendo uma tarefa desafiadora, devido à sua dependência de múltiplos fatores inter-relacionados, incluindo o peso sobre a broca (WOB), a taxa de rotação (RPM), o torque, a vazão de fluido (ou lama) de perfuração e as propriedades do reservatório. Essas variáveis exibem interações complexas e não lineares que variam significativamente entre diferentes formações geológicas e condições operacionais. Em geral, modelagens tradicionais baseadas na Física têm dificuldade em capturar essas dinâmicas, frequentemente falhando em fornecer previsões confiáveis em cenários de tempo real ou em reservatórios não convencionais (HEGDE et al., 2017).

Nos últimos anos, técnicas de aprendizado de máquina (ML, do inglês *Machine Learning*) têm demonstrado grande potencial na modelagem e previsão da ROP com maior precisão do que os métodos tradicionais baseados na Física. Diversos estudos comprovaram a eficácia de abordagens orientadas por dados na melhoria do desempenho da perfuração (ELKATATNY; AL-ABDULJABBAR; ABDELGAWAD, 2020; SINGH et al., 2019; BRENJKAR; DELIJANI, 2022). Apesar desses avanços, os modelos de ML existentes geralmente requerem engenharia de atributos específica do domínio, pré-processamento

manual e ajustes conduzidos por especialistas para alcançar um desempenho satisfatório. Essa dependência de conhecimento especializado e de processos trabalhosos dificulta sua escalabilidade e limita sua adoção em operações de perfuração em tempo real.

Para enfrentar esses desafios, o Aprendizado de Máquina Automatizado (*Automated Machine Learning*, ou *AutoML*) surgiu como uma abordagem promissora para preencher a lacuna entre o alto poder preditivo dos modelos de ML e sua aplicação prática em contextos reais (HE; ZHAO; CHU, 2021). Estruturas AutoML automatizam etapas críticas do *pipeline* de ML, incluindo o pré-processamento de dados, a engenharia de atributos, a seleção de modelos e a otimização de hiperparâmetros (ZÖLLER; HUBER, 2021). Ao simplificar essas tarefas tradicionalmente complexas e demoradas, o AutoML reduz a dependência de conhecimento técnico especializado, diminuindo assim a barreira de adoção por profissionais da área, mantendo ao mesmo tempo um desempenho preditivo competitivo (HUTTER; KOTTHOFF; VANSCHOREN, 2019).

Entretanto, ainda que tornem a tecnologia mais acessível, as ferramentas AutoML de propósito geral foram concebidas majoritariamente para dados tabulares estáticos e genéricos. No contexto da perfuração, essas soluções geralmente falham ao ignorar a natureza física e sequencial das variáveis, além de serem vulneráveis às severas descontinuidades e ruídos inerentes aos sensores de campo. Essa limitação impõe a necessidade de arquiteturas de AutoML que incorporem conhecimentos de domínio diretamente em seus *pipelines*.

Por este motivo, este estudo apresenta o AutoDrill, o primeiro *framework* de AutoML totalmente especializado em dados complexos de perfuração. O AutoDrill foi desenvolvido com o objetivo de processar dados operacionais de perfuração de forma eficiente, automatizando etapas críticas do processo de aprendizado de máquina, incluindo o pré-processamento com métodos específicos ao contexto da perfuração e a seleção e otimização de modelos preditivos. A estrutura busca adaptar-se às particularidades de cada conjunto de dados, proporcionando robustez, precisão preditiva e aplicabilidade em diferentes cenários operacionais.

1.2 Objetivos

1.2.1 Objetivo Geral

Desenvolver e avaliar uma estrutura baseada em técnicas de AutoML, denominada AutoDrill, voltada à previsão de ROP em operações de perfuração, com o objetivo de aumentar a eficiência operacional por meio da automatização de etapas críticas do *pipeline* de modelagem, como a limpeza dos dados, o pré-processamento especializado para o contexto de perfuração e a seleção do modelo com base na distribuição estatística dos dados.

1.2.2 Objetivos Específicos

- **Curadoria e Consolidação de Dados:** Consolidar repositórios de dados de perfuração provenientes de diferentes contextos geográficos e operacionais, garantindo uma base de dados heterogênea para validar a capacidade de generalização do sistema proposto.
- **Pipeline de Pré-processamento Automatizado:** Implementar e integrar ao *framework* técnicas de tratamento de sinais especializadas no domínio de perfuração, incluindo a suavização adaptativa via filtro de *Savitzky-Golay* e a correção sistemática de discontinuidades (*spikes*) e anomalias de sensores.
- **Otimização e Seleção de Modelos:** Empregar estratégias de busca para a seleção automática de algoritmos de regressão e otimização de hiperparâmetros, maximizando o desempenho preditivo sobre a variável alvo ROP.
- **Validação e Benchmarking de Domínio:** Validar o desempenho do AutoDrill por meio de uma análise comparativa com *frameworks* de AutoML de propósito geral amplamente adotados na literatura (AutoGluon e H2O), utilizando métricas de erro residual (MAE, RMSE) e aderência (R^2).
- **Avaliação de Robustez em Dados Reais:** Analisar a eficácia do sistema em cenários de baixa fidelidade de aquisição e alta complexidade litológica, verificando a estabilidade das previsões em comparação a abordagens que não consideram o conhecimento de domínio.

1.3 Justificativa

A relevância deste trabalho fundamenta-se na lacuna tecnológica existente entre o avanço dos algoritmos de aprendizado de máquina e a sua aplicação prática e confiável no setor de óleo e gás. Embora a otimização de ROP seja um fator crítico para a viabilidade econômica de projetos *offshore*, a modelagem tradicional baseada em física frequentemente não acomoda a estocasticidade e a alta dimensionalidade dos dados reais de sensores. Adicionalmente, observa-se que, em campo, o processo decisório pode incluir ainda elementos subjetivos, como a experiência e a intuição do operador, o que evidencia o desafio de traduzir integralmente a prática operacional para modelos puramente determinísticos.

No campo teórico, a justificativa reside na necessidade de transcender o uso de *frameworks* de AutoML de propósito geral. Tais ferramentas, embora robustas para dados tabulares estáticos, mostram-se insuficientes diante da natureza ruidosa e das transições litológicas abruptas típicas da perfuração. O desenvolvimento do AutoDrill fundamenta-se, portanto, na integração inédita de técnicas de processamento de sinais de domínio ao

ciclo de automação do aprendizado de máquina, permitindo que a inteligência artificial se adapte à física do poço em vez de apenas ajustar parâmetros estatísticos superficiais.

Sob a ótica operacional, a criação de uma estrutura automatizada e especializada reduz a barreira de entrada para a adoção de inteligência artificial por profissionais do setor. Ao automatizar o tratamento de anomalias instrumentais e a seleção de modelos, o AutoDrill mitiga a dependência de especialistas em ciência de dados para o retreino de modelos em campo, promovendo uma solução escalável para o suporte à tomada de decisão em tempo real.

1.4 Estrutura do Trabalho

Este trabalho está organizado em seis capítulos que detalham o desenvolvimento e a validação do *framework* proposto. Após esta Introdução, o Capítulo 2 apresenta a Revisão Bibliográfica, na qual são discutidos os modelos clássicos de predição de ROP e o estado da arte das técnicas de Aprendizado de Máquina e AutoML aplicadas ao setor de energia, evidenciando as lacunas tecnológicas que motivam esta pesquisa. O Capítulo 3 estabelece a Fundamentação Teórica, abordando os conceitos de aprendizado supervisionado, o filtro de Savitzky-Golay para o tratamento de sinais e as métricas estatísticas utilizadas para a avaliação de desempenho dos modelos. No Capítulo 4, descrevem-se os Materiais e Métodos, detalhando a arquitetura do AutoDrill, os procedimentos de pré-processamento automatizado, como a detecção de *spikes*, e a estratégia de validação experimental baseada em múltiplos repositórios. O Capítulo 5 é dedicado à apresentação e discussão dos Resultados, estabelecendo um *benchmarking* entre a solução proposta e os *frameworks* generalistas AutoGluon e H2O AutoML. Por fim, o Capítulo 6 sintetiza as conclusões do estudo, discute suas contribuições para a engenharia de petróleo e aponta diretrizes para trabalhos futuros voltados à otimização em tempo real.

2 Revisão Bibliográfica

Neste capítulo, são apresentados os principais trabalhos voltados à predição de ROP em operações de perfuração, com ênfase na transição dos modelos físicos para abordagens baseadas em Aprendizado de Máquina e técnicas de AutoML.

Primeiramente, são apresentados estudos que exploram diversos algoritmos de aprendizado de máquina aplicados à modelagem da ROP, delineando seus avanços e limitações diante da natureza multivariada e não linear dos dados de perfuração. Na sequência, discute-se o AutoML como ferramenta de automação no ciclo de desenvolvimento de modelos preditivos, destacando seu potencial em reduzir a dependência de especialistas e ampliar a aplicabilidade dos modelos a cenários reais. Por fim, são examinados os desafios e oportunidades descritos na literatura, incluindo aspectos como heterogeneidade dos dados, ruídos de sensores, transições litológicas e a necessidade de robustez nos modelos de predição. Essa análise evidencia oportunidades de pesquisa provenientes de limitações nas abordagens atuais, as quais motivam a proposição do *framework* AutoDrill.

2.1 Predição da Taxa de Perfuração com *Machine Learning*

A predição da ROP tem se consolidado como um dos temas centrais em pesquisas voltadas para a otimização de operações de perfuração na indústria de óleo e gás. Dada a correlação direta entre a taxa de penetração e a viabilidade econômica das operações de perfuração, a literatura técnica tem evoluído na busca por modelos que superem as limitações das abordagens empíricas tradicionais.

Historicamente, a previsão do ROP baseava-se em modelos empíricos ou físicos, desenvolvidos a partir da relação entre parâmetros operacionais e propriedades da formação. Entre os modelos mais citados na literatura destaca-se o modelo de [Bourgoyne e Young \(1974\)](#), proposto em 1974, que utiliza uma abordagem de regressão múltipla para representar o efeito combinado de variáveis como peso sobre a broca, rotação da coluna, propriedades da formação e profundidade sobre a taxa de perfuração. Esse modelo tornou-se uma das formulações clássicas mais utilizadas na indústria devido à sua capacidade de integrar diferentes fatores operacionais em uma única expressão matemática, mas outros modelos empíricos relevantes foram propostos.

O modelo de [Bingham \(1964\)](#), por exemplo, descreve o processo de perfuração com base na relação entre o esforço aplicado na broca e a resistência mecânica da formação, sendo amplamente utilizado em estudos iniciais de mecânica da perfuração. Já o modelo de [Teale \(1965\)](#) introduziu o conceito de Energia Mecânica Específica (MSE), que relaciona

a energia necessária para remover um volume de rocha com parâmetros operacionais da perfuração. Esse conceito tornou-se particularmente importante para análise de eficiência da perfuração e otimização de parâmetros operacionais.

Na literatura, há avaliações quantitativas do desempenho de modelos físicos e empíricos clássicos utilizados na previsão da taxa de perfuração. Em análises comparativas conduzidas por [Hegde et al. \(2019\)](#), diferentes formulações do modelo de BYM apresentaram valores médios de erro absoluto entre 28% e 30% e valores de RMSE normalizado próximos de 30% em dezenove formações geológicas distintas, evidenciando limitações na modelagem de cenários heterogêneos e altamente não lineares.

De forma semelhante, [Al-AbdulJabbar, Elkatatny e Mahmoud \(2021\)](#) compararam modelos empíricos tradicionais com versões modificadas do BYM em poços carbonáticos com perda de circulação. Os resultados mostraram que o modelo clássico de BYM obteve coeficiente de determinação de aproximadamente $R^2=0,72$ e RMSE superior a 30 em alguns cenários operacionais, evidenciando as limitações das formulações empíricas convencionais em ambientes complexos. Além disso, [Kor et al. \(2021\)](#) destacam que o modelo de BYM apresenta desempenho satisfatório principalmente em formações homogêneas, mas sua acurácia tende a degradar significativamente em ambientes litologicamente heterogêneos, devido à dificuldade de representar adequadamente as interações geomecânicas e operacionais presentes no processo de perfuração.

Dessa forma, apesar de sua relevância histórica e contribuição conceitual, em geral, essas abordagens assumem relações funcionais relativamente simples entre as variáveis de entrada e o ROP, o que pode não refletir adequadamente a natureza altamente não linear e multifatorial do processo de perfuração ([HOSSAIN; AL-MAJED, 2015](#)).

Com o propósito de superar as limitações dos modelos físicos, a literatura tem adotado, de forma crescente, técnicas de aprendizado de máquina, em razão de sua capacidade de capturar relações complexas entre as variáveis de perfuração, adaptando-se a condições operacionais e geológicas variáveis ([HORNIK; STINCHCOMBE; WHITE, 1989](#); [GANDELMAN, 2012](#)). Essa característica torna tais abordagens relevantes no contexto das operações modernas de perfuração, cuja complexidade tem aumentado significativamente, especialmente em ambientes como águas profundas ou formações geológicas heterogêneas. Nesses cenários, múltiplos fatores (incluindo propriedades da rocha, parâmetros hidráulicos, características da broca e condições operacionais) interagem de maneira dinâmica e frequentemente não linear, dificultando sua representação por meio de modelos empíricos tradicionais. Dessa forma, métodos baseados em aprendizado de máquina têm se destacado como alternativas promissoras, uma vez que permitem modelar essas interações complexas e explorar grandes volumes de dados operacionais para otimizar o processo de perfuração.

De acordo com [Gandelman \(2012\)](#), as abordagens orientadas a dados demonstram a capacidade de gerar previsões em tempo real de ROP, permitindo ajustes dinâmicos nos

parâmetros operacionais durante a perfuração e favorecendo decisões mais ágeis diante de variações litológicas, condições de pressão e desgaste da broca. Essa adaptabilidade não apenas reduz custos associados ao tempo não-produtivo (NPT, de *Non-Productive Time*), mas também mitiga riscos operacionais relacionados a vibrações excessivas e falhas prematuras de equipamentos, promovendo maior segurança e eficiência.

Barbosa et al. (2019) reforçam essa perspectiva ao destacar o crescimento expressivo da literatura na última década, evidenciando a consolidação do aprendizado de máquina como ferramenta na engenharia de perfuração. Segundo os autores, os modelos preditivos aplicados à ROP não se restringem apenas à estimativa da taxa de penetração, mas também viabilizam a otimização dinâmica e contínua de variáveis críticas como peso sobre a broca, rotação e vazão de fluido de perfuração, possibilitando alcançar o equilíbrio entre maior desempenho e preservação da integridade mecânica dos equipamentos.

Diversos algoritmos de aprendizado de máquina vêm sendo aplicados à previsão de ROP, incluindo Redes Neurais Artificiais (RNA), *Support Vector Machines* (SVM) e métodos de conjuntos (*ensembles*) baseados em árvores, como *Random Forests* e *Gradient Boosting*. A literatura recente demonstra que esses modelos frequentemente superam os métodos tradicionais baseados em física, alcançando reduções nos erros de previsão e maior estabilidade preditiva. Embora não haja consenso na literatura quanto às métricas de avaliação, os estudos evidenciam os ganhos de desempenho proporcionados por essa abordagem:

Conclui-se, portanto, que as técnicas de aprendizado de máquina superam abordagens tradicionais de modelagem da taxa de perfuração de forma consistente, especialmente em cenários caracterizados por elevada variabilidade operacional e litológica, dados multivariados e presença de ruído, devido à sua capacidade de capturar relações não lineares complexas entre os parâmetros de perfuração (HU et al., 2020).

Apesar dos avanços, no entanto, a aplicação prática de modelos de aprendizado de máquina em ambientes reais de perfuração ainda enfrenta desafios consideráveis. A heterogeneidade dos dados é um dos principais obstáculos, visto que formações rochosas apresentam propriedades mecânicas e petrofísicas bastante distintas — xisto, arenito, carbonato e basalto, por exemplo, respondem de maneira diferenciada aos mesmos parâmetros operacionais (BARBOSA et al., 2019; SOARES; DAIGLE; GRAY, 2016). Além disso, questões relacionadas à qualidade dos dados, como ruídos em sensores, valores ausentes e taxas de amostragem inconsistentes, podem comprometer significativamente o desempenho dos modelos. Técnicas de pré-processamento e engenharia de *features* são fundamentais para mitigar esses problemas, mas demandam elevado nível de expertise e intervenção manual, o que limita a escalabilidade das soluções.

Essas limitações ressaltam a necessidade do desenvolvimento de abordagens de ML mais automatizadas, capazes de reduzir a dependência de ajustes manuais e, ao mesmo

Técnica	Desempenho Reportado	Referência
Redes Neurais Artificiais	MAPE de 8,6% e coeficiente de correlação de 0,94, superando modelos empíricos	(ELKATATNY; AL-ABDULJABBAR; ABDEL-GAWAD, 2020)
LSTM	$R^2 > 0,95$ com baixos erros relativos em ambientes litologicamente variáveis	(LIU et al., 2022)
SVM	Erro percentual de 6,15% com elevada generalização em dados ruidosos	(GAN et al., 2019; RIAZI et al., 2022)
Híbridos (PSO-MLP, ACO, DE)	$R^2 \approx 0,95$ no treinamento; RMSE entre 2,22 e 7,18 no teste	(FARHAD et al., 2022)
Baseados em árvores (RF)	MAPE $\approx 13\%$ em dezenas de poços horizontais	(CHERAGHI; KORD; MASHAYEKHI-ZADEH, 2021)

Tabela 1 – Desempenho de técnicas de aprendizado de máquina na previsão de ROP.

tempo, garantir robustez frente a imperfeições nos dados, abrindo espaço para a utilização de técnicas de AutoML em contextos de perfuração, tema discutido na seção seguinte.

2.2 Técnicas de AutoML

Embora o aprendizado de máquina tenha vasto potencial para a previsão de ROP, seu uso em operações reais de perfuração é frequentemente prejudicado pela necessidade de tarefas especiais, como limpeza de dados, engenharia de recursos, seleção de modelos e ajuste de hiperparâmetros, que exigem experiência em ciência de dados e operações de perfuração. Nesse contexto, as estruturas AutoML surgem como uma solução para acelerar o desenvolvimento de modelos, automatizando etapas fundamentais como pré-processamento, seleção de recursos, seleção de modelos e otimização de hiperparâmetros (HE; ZHAO; CHU, 2021; ZÖLLER; HUBER, 2021). Isso permite que o AutoML reduza a necessidade de conhecimento especializado, o que pode facilitar a adoção de modelos preditivos no ambiente industrial.

Entre os *frameworks* mais consolidados na literatura de AutoML estão o AutoGluon (SHIRKOV et al., 2020) e o H2O AutoML (LEDELL; POIRIER, 2020), selecionados neste trabalho como referências comparativas por representarem abordagens distintas e

amplamente validadas em *benchmarks* gerais de aprendizado de máquina.

O AutoGluon estrutura seus modelos por meio da integração de algoritmos em múltiplas camadas de empilhamento (*multi-layer stacking*), permitindo combinar diferentes modelos base e meta-modelos de forma hierárquica. Essa estratégia reduz a necessidade de ajustes extensivos de hiperparâmetros e tem se mostrado eficaz na obtenção de desempenho competitivo em diferentes tipos de dados tabulares (SHIRKOV et al., 2020).

O H2O AutoML, por sua vez, opera sobre uma plataforma distribuída que testa automaticamente múltiplos algoritmos, aloca recursos computacionais de forma adaptativa e combina os resultados por meio de modelos de *ensemble*. Essa abordagem permite explorar um amplo espaço de modelos e hiperparâmetros, sendo reconhecida por sua eficácia em ambientes que demandam respostas rápidas e previsões confiáveis (LEDELL; POIRIER, 2020).

De modo geral, esses *frameworks* demonstram desempenho sólido em cenários de dados estruturados de propósito geral, conforme evidenciado por avaliações comparativas em diferentes *benchmarks* de AutoML, nas quais são analisadas métricas como erro quadrático médio (RMSE), erro absoluto médio (MAE) e coeficiente de determinação (R^2) (HE; ZHAO; CHU, 2021; TRUONG et al., 2019). Esses resultados reforçam sua relevância como bases de comparação para avaliar abordagens alternativas de modelagem.

No entanto, *frameworks* de propósito geral podem apresentar limitações quando aplicados a domínios altamente especializados, como o da perfuração de petróleo e gás. Pesquisas de Karmaker et al. (2021) e Sun et al. (2023) indicam que tais estruturas geralmente não incorporam especificidades do domínio, como condições geológicas variáveis, transições litológicas e restrições operacionais dinâmicas. Além disso, problemas de qualidade de dados inerentes ao ambiente de perfuração podem comprometer a confiabilidade das previsões quando tratados de forma inadequada ao contexto operacional. Essas limitações reforçam a necessidade de *frameworks* AutoML orientados ao domínio, capazes de incorporar conhecimento especializado no *pipeline* de modelagem, tratar as particularidades dos dados de operação e fornecer níveis adequados de interpretabilidade para apoiar a tomada de decisão em campo.

2.3 Desafios e Oportunidades

Apesar do avanço de *frameworks* AutoML de propósito geral, sua aplicação em operações de perfuração ainda enfrenta limitações significativas devido a desafios específicos do domínio. Estudos recentes, como o de Elahifar e Hosseini (2024), evidenciam que os dados de perfuração apresentam altos níveis de ruído, valores ausentes e *outliers*, decorrentes de degradação de sensores, impactos mecânicos e inconsistências na aquisição. Além disso, a variabilidade geológica introduz dinâmicas complexas e pouco intuitivas entre os

parâmetros operacionais e a taxa de perfuração. Essas características tornam o processo de modelagem peculiar e laborioso, exigindo abordagens capazes de lidar com dados não ideais e relações complexas entre variáveis.

Nesse contexto, observa-se uma lacuna na literatura no que se refere à capacidade de generalização de *frameworks* AutoML aplicados à perfuração. A maioria dos estudos concentra-se no desenvolvimento de modelos ajustados para conjuntos de dados específicos ou para condições operacionais particulares. Embora esses modelos possam apresentar bom desempenho em cenários restritos, sua aplicação em novos conjuntos de dados geralmente resulta em perda de precisão, exigindo a reinicialização de grande parte do processo de ciência de dados, incluindo etapas de preparação de dados, seleção de atributos e ajuste de hiperparâmetros.

Frameworks AutoML de propósito geral frequentemente apresentam limitações quanto à disponibilidade de *pipelines* de pré-processamento para o tratamento contextualizado dessas irregularidades. De acordo com [Karmaker et al. \(2021\)](#), apesar dos avanços das ferramentas AutoML, muitas ainda carecem de suporte completo para tarefas essenciais, como compreensão das características específicas dos dados de um domínio, definição adequada do problema de previsão e construção apropriada de conjuntos de treinamento. Essas etapas frequentemente exigem a intervenção de um cientista de dados, limitando a autonomia das ferramentas e dificultando que especialistas do domínio utilizem essas tecnologias. Além disso, a ausência de automação em etapas críticas do processo de modelagem de aprendizado de máquina restringe a efetiva implementação de AutoML em ambientes industriais que demandam respostas rápidas e previsões com alta taxa de assertividade.

Diante desses desafios, destaca-se a necessidade de desenvolver *frameworks* AutoML orientados ao domínio da perfuração que conciliem duas características fundamentais: capacidade de generalização e adaptação automática. Em vez de construir modelos específicos para cada conjunto de dados, uma abordagem eficaz consiste em desenvolver um *framework* generalista capaz de adaptar automaticamente seu *pipeline* de modelagem às características de diferentes conjuntos de dados de perfuração. Nesse contexto, o sistema atua como uma estrutura flexível que automatiza etapas como análise dos dados, pré-processamento, seleção de atributos, treinamento e avaliação de modelos, permitindo que o próprio *pipeline* se especialize para cada cenário operacional e geológico analisado.

A implementação de *frameworks* AutoML com essas características pode contribuir para a eficácia do processo de perfuração, ao permitir que os modelos capturem as particularidades de diferentes condições geológicas e operacionais sem a necessidade de reconstrução manual do pipeline analítico. Dessa forma, tais abordagens têm potencial para oferecer suporte mais confiável, escalável e eficiente à tomada de decisão operacional em ambientes industriais complexos.

2.4 Considerações Finais

A revisão bibliográfica evidencia que, embora o uso de aprendizado de máquina e AutoML para a predição da ROP seja uma tendência consolidada, as soluções atuais carecem de robustez diante da heterogeneidade e do ruído inerentes aos dados de sensores. A lacuna identificada reside na falta de *frameworks* que integrem o conhecimento de domínio ao ciclo de desenvolvimento de modelos de aprendizado de máquina. Este trabalho propõe o *AutoDrill* para preencher tal lacuna, oferecendo uma abordagem especializada na perfuração *offshore*. Para fundamentar a construção desta solução, o próximo capítulo detalha os conceitos matemáticos e computacionais que sustentam o *framework*.

3 Fundamentação Teórica

Este capítulo estabelece o suporte teórico necessário para o desenvolvimento do sistema proposto, detalhando as arquiteturas de aprendizado de máquina, as técnicas de processamento de sinais e as métricas de avaliação de desempenho.

Inicialmente, apresentam-se os algoritmos de aprendizado supervisionado selecionados para a predição de ROP, incluindo o *K-Nearest Neighbors* (KNN), *Gradient Boosting Regressor* (GBR), *Random Forest* (RF), *Extra Trees Regressor* (ET), *XGBoost Regressor* (XGB), *Support Vector Regressor* (SVR) e o *Transformed Target Regressor* (TTR). Tais métodos foram escolhidos por sua eficácia em modelar as complexas interações não lineares intrínsecas ao processo de perfuração. Na sequência, detalha-se o filtro de *Savitzky-Golay*, fundamental para a atenuação de ruídos em séries temporais de sensores sem a perda de transientes operacionais relevantes.

Complementarmente, abordam-se os *frameworks* de AutoML, que possibilitam a orquestração e automação do *pipeline* preditivo. O capítulo encerra-se com a definição das métricas estatísticas (MAE, MSE, RMSE, MAPE e R²), que fundamentam a análise comparativa e a validação do rigor científico dos resultados obtidos.

3.1 Aprendizado de Máquina

O aprendizado de máquina é o campo da inteligência artificial dedicado ao desenvolvimento de algoritmos capazes de extrair padrões de dados para realizar previsões ou subsidiar processos decisórios. O crescimento do poder computacional e da disponibilidade de bases de dados têm impulsionado a expansão de suas aplicações em vários domínios (RATTAN; PENRICE; SIMONETTO, 2022; SIMON; ALIFERIS, 2024; YING, 2019). Epistemologicamente, pode ser compreendido como uma disciplina fundamentada na convergência entre a ciência da computação e conceitos de estatística clássica.

De acordo com Rattan, Penrice e Simonetto (2022) e Simon e Aliferis (2024), o aprendizado de máquina pode ser classificado em três paradigmas principais:

- **Aprendizado Supervisionado:** os modelos são treinados a partir de pares de entrada e saída rotulados, sendo empregado em tarefas de regressão e classificação.
- **Aprendizado Não Supervisionado:** voltado para a exploração de estruturas latentes e identificação de padrões ocultos em dados não rotulados.
- **Aprendizado por Reforço:** O algoritmo é treinado a partir de estratégias de decisão por meio da interação com o ambiente e do recebimento de recompensas.

Em essência, os algoritmos de aprendizado de máquina ajustam seus parâmetros internos com base na experiência extraída dos dados, permitindo a generalização do conhecimento adquirido para novos cenários (JORDAN; MITCHELL, 2015; GOODFELLOW; BENGIO; COURVILLE, 2016). Essa capacidade de adaptação é particularmente relevante em domínios que envolvem dados de grande volume, alta dimensionalidade e relações não lineares complexas, como ocorre na perfuração de poços de petróleo, em que a variabilidade litológica e operacional dificulta a construção de modelos puramente determinísticos. Além disso, o uso de algoritmos capazes de lidar com fluxos contínuos de dados favorece abordagens mais dinâmicas, adaptativas e orientadas a dados, que superam as limitações de modelos estatísticos tradicionais ao capturar a complexidade intrínseca dos processos de perfuração (KUBAT, 2017; LECUN; BENGIO; HINTON, 2015).

Nas seções subsequentes, são detalhados os algoritmos de aprendizado de máquina empregados na construção da biblioteca AutoML desenvolvida neste trabalho, com ênfase em suas principais características, vantagens e limitações.

3.1.1 K-Nearest Neighbors (KNN)

O *K-Nearest Neighbors* (KNN) é um algoritmo de aprendizagem supervisionada que realiza previsões com base na similaridade entre dados de entrada e observações anteriores (COVER; HART, 1967). Por ser não paramétrico, o KNN não faz nenhuma suposição sobre a distribuição dos dados e pode capturar relações não lineares e complexas, o que o torna apropriado para ambientes em que as relações entre as variáveis são locais ou com heterogeneidade acentuada (REGRESSION, 1992).

O funcionamento do algoritmo fundamenta-se em três etapas principais:

- **Determinação do valor de k :** etapa responsável por definir a quantidade de vizinhos a serem considerados na previsão, exigindo atenção especial à sensibilidade associada a valores muito baixos ou muito altos. Valores reduzidos de k podem tornar o modelo sensível a ruídos, enquanto valores excessivamente elevados podem suavizar padrões locais relevantes.
- **Cálculo das distâncias entre os pontos de dados:** quantifica a proximidade entre pontos em um espaço multidimensional. Utiliza métricas como *Euclidean*, *Manhattan* ou *Minkowski*, fundamentais para identificar quais amostras influenciarão a previsão. A métrica mais comum é a Distância Euclidiana, definida entre dois pontos p e q como:

$$d(p, q) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2} \quad (3.1)$$

- **Definição da saída:** para problemas de regressão, utiliza-se a média ou a mediana dos k vizinhos; para problemas de classificação, considera-se a classe majoritária.

Embora adaptativo, o KNN apresenta limitações. O custo computacional cresce com o volume de dados, uma vez que o algoritmo necessita calcular a distância entre o ponto de consulta e todos os exemplos do conjunto de treinamento a cada predição, o que pode torná-lo inviável em cenários de larga escala. Adicionalmente, a presença de ruído ou variáveis em escalas distintas pode reduzir sua precisão, sendo necessária a aplicação de técnicas de pré-processamento, normalização e redução de dimensionalidade (REGRESSION, 1992).

No domínio industrial, especialmente na previsão de parâmetros operacionais em perfuração de poços, o KNN pode ser aplicado para identificar padrões locais em dados multivariados e oferecer suporte à decisão em cenários de alta variabilidade operacional e geológica. O algoritmo já foi empregado em domínios distintos, incluindo diagnóstico médico, detecção de anomalias e previsão de séries temporais, demonstrando flexibilidade diante de estruturas de dados complexas (ALTMAN, 1992). Sua capacidade de generalizar localmente o comportamento dos dados o torna uma opção complementar em *pipelines* de aprendizado de máquina, especialmente quando utilizado em conjunto com métodos de *ensemble* ou integrado em *frameworks* AutoML

3.1.2 Random Forest (RF)

O *Random Forest Regressor* (RF), proposto por Breiman (2001), é um método de *ensemble learning* que combina múltiplas árvores de regressão, proporcionando redução da variância e aumento de robustez em comparação a árvores individuais. Cada árvore é construída a partir de uma amostragem com reposição (*bootstrap*) do conjunto de dados original, utilizando o método CART (*Classification and Regression Trees*) e critérios como a Diminuição da Impureza de Gini (*Decrease Gini Impurity*, DGI) ou o erro quadrático médio para determinar os pontos de divisão.

O CART consiste em um método de construção de árvores de decisão aplicável a problemas de regressão e classificação. O algoritmo realiza particionamento recursivo do conjunto de dados, dividindo-o em subconjuntos mais homogêneos com base nas variáveis preditoras. Para regressão, as divisões são definidas de modo a minimizar a variância ou o erro quadrático médio das amostras nos nós terminais. Cada nó terminal representa uma predição baseada nas observações que chegam até ele. Devido à sua estrutura arbórescente, o CART favorece a interpretação do modelo por meio de regras lógicas do tipo “se-então” (COULSTON et al., 2016).

A Figura 1 ilustra o funcionamento do *Random Forest Regressor*. Em cada nó das árvores, apenas um subconjunto aleatório de variáveis é considerado como candidato à divisão, o que promove diversidade entre os modelos e reduz o risco de sobreajuste, mesmo em dados ruidosos ou de alta dimensionalidade. As árvores são construídas até o atendimento dos critérios de parada, sem aplicação de poda, possibilitando a captura

de padrões complexos. As previsões individuais são então agregadas por meio da média, resultando na predição final do modelo (COURONNÉ; PROBST; BOULESTEIX, 2018).

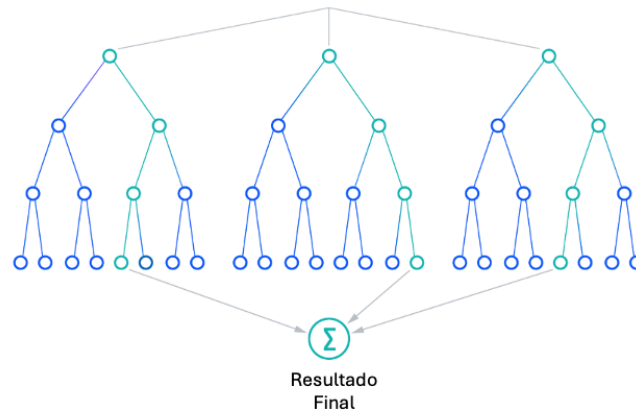


Figura 1 – Fluxo Ilustrativo de uma Random Forest. Figura do autor.

Entre os hiperparâmetros críticos do RF destacam-se o número de árvores que compõem a floresta (`n_estimators`), a profundidade máxima das árvores (`max_depth`), o número mínimo de amostras necessárias para realizar uma divisão (`min_samples_split`), o número mínimo de amostras em cada nó terminal (`min_samples_leaf`), o número de variáveis preditoras consideradas em cada divisão (`max_features`) e a estratégia de reamostragem (`bootstrap`). O parâmetro `replace` define se a reamostragem é com reposição (`bootstrap`) ou sem reposição (subamostras). Ajustes adequados desses parâmetros equilibram a capacidade do modelo de capturar padrões complexos e evitar sobreajuste (*overfitting*), embora o *Random Forest Regressor* seja reconhecido por sua robustez a diferentes configurações de hiperparâmetros. Devido à sua estabilidade e baixo custo de parametrização, o RF é frequentemente utilizado como *baseline* em cenários industriais com dados heterogêneos e ruidosos (NATEKIN; KNOLL, 2013).

3.1.3 Gradient Boosting Regressor (GBR)

O *Gradient Boosting Regressor* (GBR) é um método de aprendizado supervisionado que combina múltiplos modelos fracos (*weak learners*), geralmente árvores de decisão rasas — isto é, árvores com poucos níveis de profundidade, capazes de capturar apenas padrões simples — para construir um preditor mais robusto e preciso (FRIEDMAN, 2001). A técnica se baseia em aprendizado sequencial, no qual cada nova árvore é ajustada para corrigir os resíduos do modelo anterior, reduzindo progressivamente o erro de previsão. Esse processo iterativo torna o GBR altamente eficaz na modelagem de relações não lineares e complexas, mesmo em dados ruidosos ou heterogêneos (CLARK; LEE, 2025).

O algoritmo pode ser descrito nas seguintes etapas, conforme ilustrado na Figura 2:

- **Inicialização:** define-se uma predição inicial constante, geralmente a média dos valores da variável alvo (y) no conjunto de treinamento;
- **Cálculo dos Gradientes (Resíduos):** para cada iteração, calcula-se o gradiente negativo da função de perda em relação às predições atuais, identificando a direção de maior erro;
- **Ajuste do Modelo Base:** uma nova árvore de decisão é ajustada aos gradientes calculados (pseudorespostas), buscando capturar o padrão do erro remanescente;
- **Atualização com Encolhimento:** a nova árvore é adicionada ao modelo existente, ponderada por uma taxa de aprendizado (*learning rate*), atualizando as predições de forma incremental.

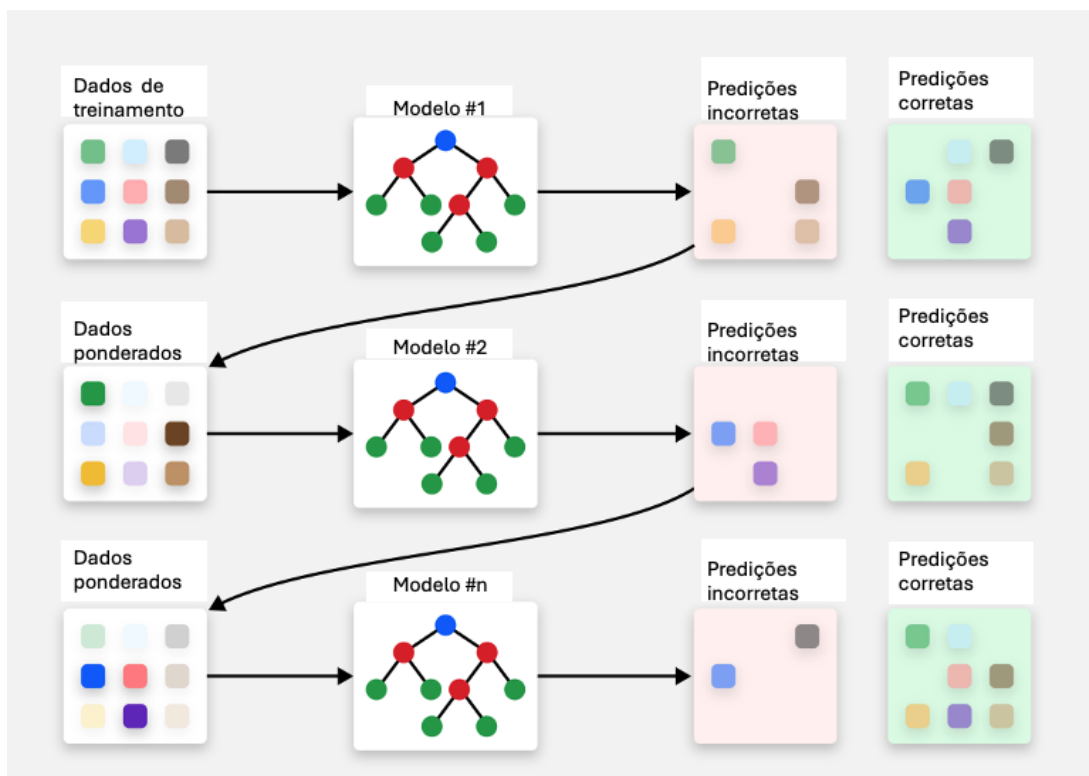


Figura 2 – Fluxo de Treinamento GBR. Figura do autor.

O parâmetro de taxa de aprendizado (*learning rate*) é crítico, controlando a contribuição de cada árvore e equilibrando o risco de *overfitting* e a velocidade de convergência. Outros hiperparâmetros importantes incluem o número de estimadores e a profundidade das árvores, que influenciam diretamente a capacidade do modelo de capturar padrões complexos. Técnicas como regularização, poda de árvores e *early stopping* podem ser aplicadas para prevenir sobreajuste (CLARK; LEE, 2025).

3.1.4 XGBoost Regressor (XGB)

O *XGBoost Regressor* é uma implementação otimizada do método de *gradient boosting* baseada em árvores de decisão, proposta por (CHEN; GUESTRIN, 2016). Esse algoritmo introduz diversas melhorias em relação aos métodos tradicionais de *gradient boosting*, tornando-o mais eficiente e escalável.

Entre seus diferenciais técnicos, destacam-se:

- **Otimização de Segunda Ordem:** Diferente do *gradient boosting* convencional, o *XGBoost* utiliza a expansão de Taylor de segunda ordem para aproximar a função de perda, incorporando informações tanto do gradiente quanto da Hessiana. Isso resulta em uma convergência mais precisa e eficiente para o valor alvo de ROP.
- **Regularização L1 e L2:** o *XGBoost* incorpora penalizações *L1* (*Lasso*) e *L2* (*Ridge*) sobre os pesos das árvores, permitindo controlar a complexidade do modelo e reduzir o risco de *overfitting*, especialmente em conjuntos de dados ruidosos.
- **Treinamento paralelizado:** diferentemente de implementações clássicas de *gradient boosting*, o *XGBoost* realiza operações paralelas durante a construção das árvores, reduzindo significativamente o tempo de treinamento e possibilitando o processamento de grandes volumes de dados.
- **Busca de divisão sensível à esparsidade:** o algoritmo lida de forma eficiente com valores ausentes ou dados esparsos, identificando automaticamente as melhores divisões mesmo em conjuntos de dados com numerosas lacunas.
- **Escalabilidade e robustez:** o *XGBoost* é altamente escalável, capaz de lidar com conjuntos de dados de alta dimensionalidade, variáveis heterogêneas e relações não lineares complexas, mantendo desempenho consistente mesmo em cenários com grande variabilidade operacional.

Devido à robustez e à capacidade de lidar com incertezas e heterogeneidade nos dados, o *XGBoost* é particularmente adequado para a previsão da taxa de penetração (ROP) e de outros parâmetros operacionais, modelando relações complexas entre variáveis operacionais, geológicas e de equipamento.

3.1.5 Extra Trees Regressor (ET)

O *Extra Trees Regressor* (Árvores Extremamente Aleatórias), introduzido por Geurts, Ernst e Wehenkel (2006), é um algoritmo de aprendizado de máquina baseado em árvores de decisão, caracterizado por sua alta aleatoriedade na construção das árvores, o que resulta em modelos com menor variância e maior velocidade de treinamento. Enquanto

os *Random Forests* buscam o melhor ponto de divisão em cada nó por meio de otimização local, *Extra Trees* seleciona os pontos de divisão de forma totalmente aleatória (MAME et al., 2025). Essa abordagem, embora possa gerar divisões menos otimizadas localmente, promove diversidade entre as árvores do *ensemble*, o que tende a reduzir a correlação entre elas e, conseqüentemente, diminuir a variância do modelo final (DEVELOPERS, 2024).

Outra diferença importante é que, por padrão, *Extra Trees* não depende do *bootstrap* para criar subconjuntos do treinamento; em vez disso, utiliza todo o conjunto de dados disponível para treinar cada árvore (GEURTS; ERNST; WEHENKEL, 2006). Isso contribui para um aprendizado estável, especificamente em conjuntos de dados grandes (MAME et al., 2025). O aumento da aleatoriedade traz ainda benefícios computacionais: a seleção aleatória de pontos de divisão é consideravelmente mais rápida do que a busca exaustiva realizada pelas *Random Forests* (DEVELOPERS, 2024). Como resultado, o tempo de treinamento do *Extra Trees* costuma ser menor, permitindo a construção de modelos complexos de forma mais eficiente (GEURTS; ERNST; WEHENKEL, 2006).

3.1.6 Support Vector Regressor (SVR)

O *Support Vector Regression* (SVR) é uma técnica derivada das *Support Vector Machines* (SVMs) aplicada a problemas de regressão. Diferente da SVM para classificação, o SVR busca determinar uma função $f(x)$ que se aproxime dos dados de treinamento de forma que os erros de predição fiquem dentro de uma tolerância pré-definida ϵ (SMOLA; SCHÖLKOPF, 2004; LORENA; CARVALHO, 2007). A função de regressão no SVR é determinada a partir de um subconjunto dos pontos de treinamento, denominados vetores de suporte, elementos usados para definir o modelo final. A função pode ser expressa como:

$$f(x) = \langle w, \phi(x) \rangle + b \quad (3.2)$$

onde w representa o vetor de pesos, b é o termo de *bias*, e $\phi(x)$ é uma função de mapeamento que projeta os dados de entrada para um espaço de características de dimensão superior. Esse mapeamento permite que o SVR capture relações complexas e não lineares entre variáveis. (SMOLA; SCHÖLKOPF, 2004).

Para lidar explicitamente com não linearidades, o SVR pode empregar funções *kernel*, como o *kernel* de base radial (RBF). Nesse caso, o *kernel* calcula diretamente a similaridade entre pares de pontos no espaço transformado, evitando a necessidade de computar explicitamente as coordenadas no espaço de alta dimensão. Essa abordagem torna possível modelar relações não lineares complexas entre as variáveis de entrada e saída, mantendo a eficiência computacional (LORENA; CARVALHO, 2007).

O treinamento do SVR consiste em resolver um problema de otimização convexa, equilibrando a complexidade do modelo e a precisão da predição. Esse equilíbrio garante robustez contra ruído e alta dimensionalidade do conjunto de características. Entre suas

principais vantagens estão a capacidade de generalização, a robustez a dados ruidosos e a flexibilidade para modelar relações não lineares. Entretanto, o SVR apresenta limitações, como a alta demanda computacional em grandes conjuntos de dados e a sensibilidade à escolha de hiperparâmetros, como C , ϵ e parâmetros do *kernel* (SMOLA; SCHÖLKOPF, 2004; LORENA; CARVALHO, 2007).

3.1.7 Transformed Target Regressor com ExtraTrees

O *Transformed Target Regressor* (TTR) é uma técnica de meta-estimativa que aplica uma transformação matemática à variável alvo (y) antes do ajuste do modelo, realizando a transformação inversa após a predição. Essa abordagem visa aproximar a distribuição da variável alvo de uma distribuição normal (Gaussiana) e estabilizar sua variância, o que mitiga o impacto de assimetrias e heterocedasticidade comuns em dados de perfuração (YEO; JOHNSON, 2000; RONALD; CONOVER, 1983).

Neste trabalho, utiliza-se a transformação de *Yeo-Johnson*, que expande a lógica da transformação *Box-Cox* ao permitir o tratamento de valores nulos e negativos, garantindo maior flexibilidade estatística. A transformação é definida formalmente pela Equação 3.3:

$$y^{(\lambda)} = \begin{cases} \frac{(y+1)^\lambda - 1}{\lambda}, & y \geq 0, \lambda \neq 0 \\ \ln(y+1), & y \geq 0, \lambda = 0 \\ -\frac{(-y+1)^{2-\lambda} - 1}{2-\lambda}, & y < 0, \lambda \neq 2 \\ -\ln(-y+1), & y < 0, \lambda = 2 \end{cases} \quad (3.3)$$

Após a predição, aplica-se a transformação inversa para retornar os valores à escala original, garantindo interpretabilidade (YEO; JOHNSON, 2000), conforme a Equação 3.4:

$$y = \begin{cases} (\lambda y^{(\lambda)} + 1)^{1/\lambda} - 1, & y^{(\lambda)} \geq 0, \lambda \neq 0 \\ \exp(y^{(\lambda)}) - 1, & y^{(\lambda)} \geq 0, \lambda = 0 \\ 1 - [-(2-\lambda)y^{(\lambda)} + 1]^{1/(2-\lambda)}, & y^{(\lambda)} < 0, \lambda \neq 2 \\ 1 - \exp(-y^{(\lambda)}), & y^{(\lambda)} < 0, \lambda = 2 \end{cases} \quad (3.4)$$

A integração do *Extra Trees* com o TTR permite que o algoritmo de *ensemble* opere sobre uma variável alvo estatisticamente mais estável, reduzindo a influência de heteroscedasticidade ou assimetrias em sua distribuição. Embora árvores de decisão sejam robustas a *outliers*, a normalização da distribuição alvo facilita a redução do erro residual em regiões de dados esparsos ou com variabilidade extrema (YEO; JOHNSON, 2000).

As principais vantagens técnicas dessa abordagem incluem:

- Linearização parcial da variável alvo, facilitando o aprendizado do modelo;
- Estabilização da variância, reduzindo o impacto de *outliers* ou distribuições assimétricas;
- Flexibilidade não linear, graças ao *Extra Trees*, que consegue capturar interações complexas entre características;
- Robustez a distribuições não Gaussianas, devido à transformação adaptativa da variável alvo.

3.2 AutoML

O AutoML consiste em um conjunto de métodos que visam automatizar etapas do desenvolvimento de modelos de aprendizado de máquina, desde o pré-processamento de dados, seleção de algoritmos, ajuste de hiperparâmetros e avaliação de desempenho (HE; ZHAO; CHU, 2021). A automação dessas etapas reduz a necessidade de intervenção manual especializada e permite a construção sistemática de modelos preditivos por profissionais com diferentes níveis de conhecimento em ciência de dados, além de acelerar o processo de construção e retreino do modelo.

Um sistema AutoML geralmente é estruturado em três componentes principais, ilustradas na Figura 3: *search space*, *search algorithm* e *evaluation strategy* (HE; ZHAO; CHU, 2021). O *search space* define o conjunto de configurações possíveis para o modelo, abrangendo algoritmos, hiperparâmetros e técnicas de pré-processamento. O *search algorithm* explora esse espaço para identificar a configuração mais adequada, empregando métodos como *grid search*, *random search* ou algoritmos evolutivos. Por fim, a *evaluation strategy* avalia o desempenho das configurações, geralmente por meio de validação cruzada, estimando a capacidade de generalização do modelo.

A comparação objetiva entre diferentes sistemas AutoML é fundamental para avaliar seu desempenho e identificar melhores práticas. No contexto de tarefas de regressão em dados tabulares, *frameworks* recentes têm incorporado estratégias de *ensemble* e otimização automática de modelos. Nesse sentido, o AutoGluon, proposto por Erickson et al. (2020), foi desenvolvido para problemas de aprendizado supervisionado em dados tabulares, incluindo regressão, sendo avaliado em *benchmarks* com conjuntos de dados do OpenML e outras fontes públicas, como em Michael (2025). Seu funcionamento baseia-se na combinação de modelos por meio de técnicas de *multi-layer stacking* e *bagging*, o que permite explorar de forma sistemática o espaço de modelos e favorecer a capacidade de generalização.

Por outro lado, o H2O AutoML, conforme apresentado por LeDell e Poirier (2020), caracteriza-se pela capacidade de integração de diferentes algoritmos amplamente utilizados

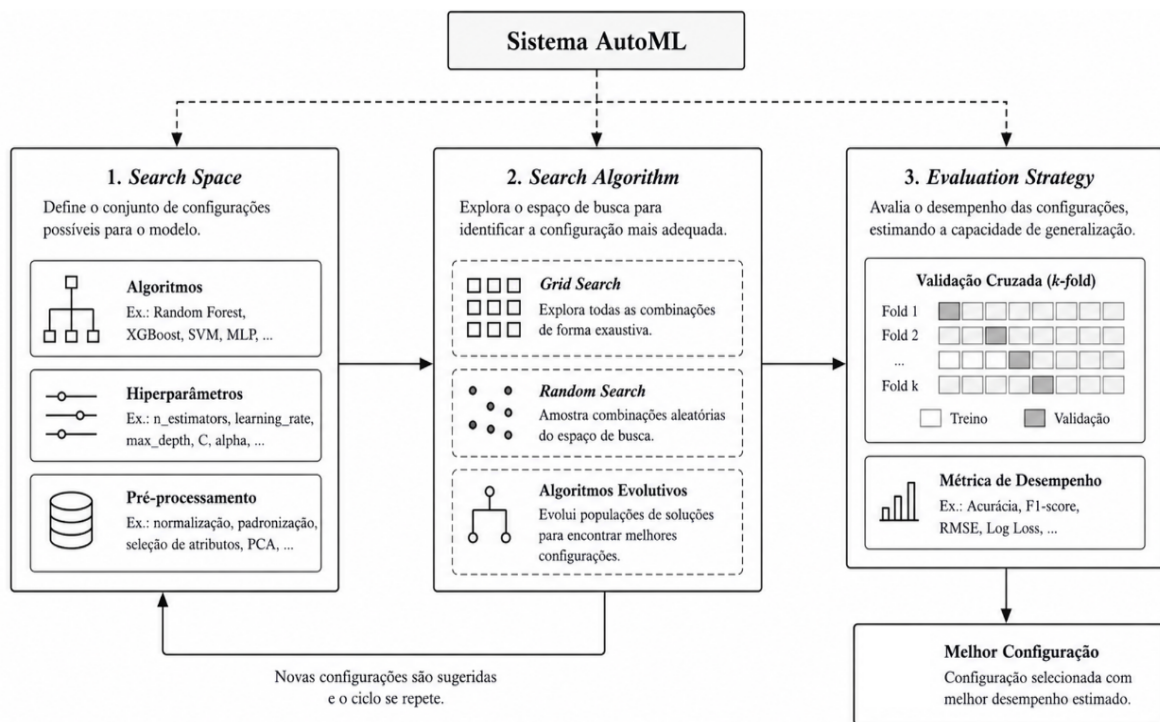


Figura 3 – Componentes de um sistema AutoML. Figura do autor.

em tarefas de regressão, como *Generalized Linear Models*, *Gradient Boosting Machines* e *Random Forests*, além de arquiteturas de redes neurais profundas. O *benchmark* conduzido por [Gijssbers et al. \(2019\)](#) indica que não há um método universalmente superior entre as ferramentas AutoML, uma vez que o desempenho varia de acordo com as características dos dados e o tipo de tarefa. No entanto, o H2O AutoML apresenta comportamento consistente em diferentes cenários, associado ao uso de estratégias de *ensemble*, o que contribui para a estabilidade dos resultados em problemas de regressão.

Dessa forma, a escolha conjunta de AutoGluon e H2O AutoML neste trabalho está fundamentada na complementaridade entre as abordagens. Enquanto o AutoGluon prioriza a combinação hierárquica de modelos para exploração do espaço de soluções, o H2O AutoML oferece uma estrutura consolidada com múltiplos algoritmos e suporte a diferentes configurações de treinamento. Essa combinação mostra-se adequada para problemas de regressão com elevada complexidade, nos quais é necessário capturar relações não lineares e manter consistência no desempenho dos modelos.

3.2.1 AutoGluon

O AutoGluon-Tabular ([ERICKSON et al., 2020](#)) é um *framework* de AutoML de código aberto desenvolvido no âmbito da *Amazon Web Services* (AWS), com o objetivo de facilitar o treinamento de modelos de aprendizado de máquina em dados tabulares com mínima intervenção do usuário. Diferentemente de abordagens tradicionais que priorizam

a seleção de um único modelo e a otimização de seus hiperparâmetros, o AutoGluon adota uma estratégia baseada na combinação de múltiplos modelos por meio de técnicas de *ensemble*, organizadas em diferentes níveis hierárquicos.

O *framework* realiza automaticamente etapas de pré-processamento, incluindo a identificação dos tipos de variáveis (numéricas, categóricas e textuais) e a aplicação de transformações apropriadas. Entre essas transformações estão a codificação de variáveis categóricas, o tratamento de valores ausentes e a vetorização de atributos textuais. Essas etapas são integradas ao *pipeline* de modelagem, reduzindo a necessidade de intervenção manual e permitindo a utilização direta de dados tabulares heterogêneos.

Durante o treinamento, o AutoGluon emprega um conjunto diversificado de modelos de base (*base learners*), o que permite capturar diferentes tipos de relações entre as variáveis, incluindo padrões lineares e não lineares. Os modelos incluem:

- Métodos baseados em árvores, como *Random Forest* e *Extremely Randomized Trees*;
- Algoritmos de *gradient boosting*, como LightGBM e CatBoost;
- Modelos lineares regularizados;
- Redes neurais profundas adaptadas para dados tabulares.

Os modelos de base são organizados em um esquema de empilhamento em múltiplas camadas (*multi-layer stack ensembling*), em que cada camada recebe as previsões da camada anterior como entradas adicionais. Modelos de nível superior (*meta-learners*) aprendem a ponderar e combinar as previsões das camadas inferiores, aumentando a eficácia do modelo final e reduzindo tanto o viés quanto a variância. O AutoGluon também utiliza técnicas de *bagging* e *blending*, criando múltiplas instâncias de cada modelo de base com diferentes subconjuntos de dados ou parâmetros, combinando suas previsões para melhorar a estabilidade e a generalização do modelo (ERICKSON et al., 2020).

3.2.2 H2O

O H2O AutoML é um *framework* de aprendizado de máquina automatizado, desenvolvido para acelerar a construção de modelos preditivos em dados tabulares, com ênfase em escalabilidade e eficiência computacional (LEDELL; POIRIER, 2020). O sistema combina a exploração do espaço de algoritmos e hiperparâmetros com a seleção automatizada de modelos, permitindo a identificação de configurações competitivas com mínima intervenção do usuário.

A execução do H2O AutoML ocorre em um ambiente baseado em *cluster*, operando sobre a *Java Virtual Machine* (JVM), no qual são realizados o processamento dos dados,

o treinamento e a avaliação dos modelos. O *framework* oferece interfaces em linguagens como Python, R, Java e Scala, além de uma interface gráfica intuitiva, ampliando sua acessibilidade em diferentes contextos de uso (JÄÄSKELÄINEN, 2022).

No que se refere ao tratamento dos dados, o H2O AutoML suporta diretamente variáveis numéricas e categóricas, realizando automaticamente o tratamento de valores ausentes e a codificação de atributos categóricos conforme necessário. Essas etapas são integradas ao *pipeline* de modelagem, reduzindo a necessidade de pré-processamento manual e garantindo consistência na entrada dos modelos.

O *framework* emprega uma variedade de modelos de base, treinados em paralelo, entre os quais se destacam:

- *Gradient Boosting Machines*, que refinam iterativamente as previsões minimizando os erros residuais;
- *Random Forests*, construídas a partir de múltiplas árvores independentes, promovendo diversidade e resistência a ruído;
- Redes Neurais, capazes de aprender representações não lineares complexas e interações entre características;
- Modelos Lineares Generalizados (GLMs), para capturar relações lineares fundamentais;
- Modelos de *kernel* e *ensembles* simples, utilizados quando apropriado para otimização de desempenho.

Após o treinamento dos modelos de base, o H2O AutoML constrói modelos do tipo *stacked ensemble*, nos quais um *meta-learner* é responsável por combinar as previsões individuais. Tipicamente, são gerados diferentes *ensembles*, incluindo um que utiliza todos os modelos disponíveis e outro baseado nos melhores modelos de cada família, visando equilibrar desempenho e custo computacional. Durante todo o processo, os modelos são avaliados em um *leaderboard* (um quadro de classificação, ou ranking), sendo o modelo final selecionado com base na métrica de desempenho definida (JÄÄSKELÄINEN, 2022).

O tratamento de *outliers* e ruído no H2O AutoML ocorre tanto pela natureza intrínseca dos modelos de base, como GBMs e *Random Forests*, quanto pelo empilhamento, que suaviza o efeito de observações extremas. Adicionalmente, o H2O AutoML incorpora mecanismos de controle de recursos computacionais, como limitação de tempo de execução e priorização de modelos mais promissores, permitindo maior eficiência no uso de infraestrutura. Sua arquitetura distribuída possibilita execução em ambientes escaláveis, incluindo integração com plataformas como Hadoop e Spark, o que favorece sua aplicação em problemas de maior porte (LEDELL; POIRIER, 2020).

3.3 Filtro de sinais

O pré-processamento de sinais desempenha papel fundamental na melhoria da qualidade dos dados, reduzindo ruídos que podem comprometer a estabilidade numérica e a capacidade preditiva de modelos subsequentes. Entre as diversas técnicas de suavização, destaca-se o uso de filtros digitais que preservam as características estruturais do sinal original, ao mesmo tempo em que eliminam flutuações indesejadas (CHEN et al., 2004; JOHN; SADASIVAN; SEELAMANTULA, 2021). A escolha do filtro é determinante em aplicações nas quais a integridade do sinal deve ser preservada, como no monitoramento de variáveis de sensores em tempo real ou em séries temporais de instrumentação.

3.3.1 Filtro de Savitzky-Golay

Neste trabalho, o filtro digital adotado foi o *Savitzky-Golay* (SAVITZKY; GOLAY, 1964), cuja seleção se justifica por suas características de suavização com preservação da forma do sinal, ajuste polinomial local, baixa distorção de fase e manutenção da amplitude relativa. Suas características o tornam particularmente adequado para dados de perfuração, apresentando propriedades essenciais para capturar mudanças operacionais e litológicas sem deformar características relevantes do comportamento físico dos parâmetros medidos.

O filtro consiste em uma técnica de suavização digital que realiza ajuste polinomial local via mínimos quadrados sobre janelas deslizantes dos dados. Para cada ponto central de uma janela de tamanho $w = 2m + 1$ (em que m representa o número de pontos considerados em cada lado do ponto central da janela), ajusta-se um polinômio de ordem p aos w vizinhos e se avalia esse polinômio no ponto central, conferindo máxima fidelidade a formas do sinal original — tais como picos, vales, curvaturas e pontos de inflexão — que muitas outras técnicas de suavização tendem a atenuar ou distorcer (JOHN; SADASIVAN; SEELAMANTULA, 2021).

Os parâmetros críticos do método são, portanto, o tamanho da janela w e a ordem p do polinômio ajustado. Enquanto as janelas largas reduzem a variância do ruído, introduzem viés inserindo ruído local; polinômios de ordem mais alta capturam mais complexidade do sinal, mas podem amplificar ruídos globais se não forem bem regulados. Além disso, o filtro pressupõe dados equiespaçados; se os pontos não estiverem uniformemente distribuídos, a interpolação ou pré-processamento para uniformizar os espaçamentos podem ser necessários (KONSTANTINOVSKY, 2024).

O ajuste polinomial é formalizado como um problema de mínimos quadrados. Para uma janela centrada em x_0 , o polinômio de ordem p é definido por:

$$y(x) \approx \sum_{k=0}^p a_k (x - x_0)^k, \quad (3.5)$$

onde a_k são os coeficientes do polinômio, determinados minimizando a soma dos quadrados dos erros:

$$\min_{a_0, \dots, a_p} \sum_{i=-m}^m \left[y(x_0 + i) - \sum_{k=0}^p a_k i^k \right]^2. \quad (3.6)$$

O valor filtrado no ponto central x_0 é então dado por $y(x_0) = a_0$. Equivalentemente, a suavização pode ser implementada como uma convolução discreta:

$$y_{\text{smooth}}(x_0) = \sum_{i=-m}^m c_i y(x_0 + i), \quad (3.7)$$

onde os coeficientes de convolução c_i são pré-calculados a partir da matriz de Vandermonde associada à janela:

$$\mathbf{c} = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{e}_1, \quad (3.8)$$

com $\mathbf{A}$ sendo a matriz de Vandermonde formada pelos deslocamentos $i = -m, \dots, m$ elevados a potências de 0 a p , e $\mathbf{e}_1$ um vetor unitário selecionando o coeficiente constante a_0 (KONSTANTINOVSKY, 2024).

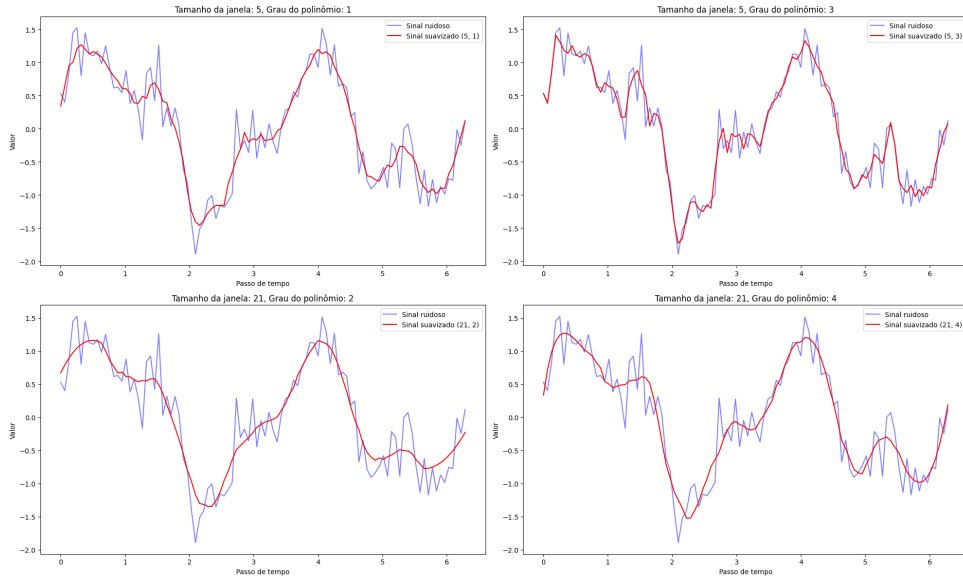


Figura 4 – Exemplo de sinal ruidoso e suavizado com filtro de *Savitzky-Golay*. Figura do autor.

A Figura 4 ilustra a aplicação do filtro de *Savitzky-Golay* em um sinal sintético composto por múltiplas frequências e ruído aditivo. O sinal resultante está em vermelho. Cada gráfico demonstra diferentes combinações de tamanho de janela e grau do polinômio, evidenciando o efeito da parametrização na suavização.

- **Subfigura superior esquerda** ($w = 5, p = 1$): utiliza uma janela reduzida e um polinômio de primeira ordem, resultando em uma suavização aproximadamente linear em cada intervalo local. Embora o ruído seja parcialmente reduzido, observa-se menor capacidade de representar curvaturas mais complexas do sinal.
- **Subfigura superior direita** ($w = 5, p = 3$): mantém o mesmo tamanho de janela, porém utiliza um polinômio de terceira ordem. Essa configuração proporciona maior flexibilidade para representar variações locais do sinal, preservando melhor picos, vales e mudanças abruptas, ainda que com menor atenuação do ruído.
- **Subfigura inferior esquerda** ($w = 21, p = 2$): emprega uma janela mais ampla e um polinômio de segunda ordem. O aumento da janela promove suavização mais intensa, reduzindo significativamente as flutuações de alta frequência. Em contrapartida, parte das variações locais do sinal é suavizada.
- **Subfigura inferior direita** ($w = 21, p = 4$): combina uma janela ampla com um polinômio de quarta ordem. Essa parametrização mantém o elevado nível de suavização proporcionado pela janela extensa, mas aumenta a capacidade de representar curvaturas mais complexas do sinal em comparação ao caso anterior.

De forma geral, a Figura 4 evidencia o compromisso existente entre suavização e preservação das características do sinal. Janelas maiores tendem a reduzir o ruído de maneira mais efetiva, enquanto polinômios de ordem superior aumentam a capacidade de adaptação às variações locais dos dados.

Do ponto de vista computacional, apresenta complexidade linear $\mathcal{O}(N)$, e sua formulação permite calcular o filtro eficientemente, mesmo em séries longas e aplicações em tempo real. A implementação é realizada a partir de convoluções com coeficientes pré-computados para cada combinação de w e p . Operações de fronteira exigem tratamento especial, como espelhamento (*padding*) ou ajuste da janela, para evitar artefatos causados pela ausência de vizinhos suficientes. Essa eficiência, aliada à preservação da forma do sinal original, justifica sua ampla adoção em contextos como instrumentação, análise espectral e pré-processamento de dados experimentais (SAVITZKY; GOLAY, 1964).

3.4 Método de Tukey para Detecção de Outliers

A identificação e o tratamento de valores discrepantes (*outliers*) representam uma etapa importante no pré-processamento de dados, especialmente em aplicações de aprendizado de máquina envolvendo dados operacionais reais. Em operações de perfuração, ruídos de sensores, falhas de aquisição e variações abruptas de medição podem introduzir valores extremos capazes de comprometer o treinamento e a capacidade de generalização dos modelos preditivos.

Entre os métodos estatísticos mais utilizados para detecção de *outliers* destaca-se o método de Tukey, proposto por [Tukey \(1977\)](#), baseado na análise da amplitude interquartil (*Interquartile Range* – IQR). O método utiliza os quartis da distribuição dos dados para estabelecer limites inferiores e superiores de aceitabilidade estatística, sem assumir distribuição normal dos dados.

A amplitude interquartil é definida como a diferença entre o terceiro quartil (Q_3) e o primeiro quartil (Q_1):

$$IQR = Q_3 - Q_1 \quad (3.9)$$

Com base no *IQR*, são definidos limites para identificação de valores discrepantes. Valores localizados abaixo de:

$$Q_1 - 1.5 \times IQR \quad (3.10)$$

ou acima de:

$$Q_3 + 1.5 \times IQR \quad (3.11)$$

são considerados potenciais *outliers*. Já valores além dos limites de:

$$Q_1 - 3 \times IQR \quad (3.12)$$

e

$$Q_3 + 3 \times IQR \quad (3.13)$$

são classificados como *outliers* extremos.

A principal vantagem do método de Tukey está em sua robustez estatística e simplicidade computacional, tornando-o adequado para conjuntos de dados com distribuições assimétricas ou presença de ruídos, características frequentemente observadas em dados de perfuração de poços de petróleo. Dessa forma, o método foi empregado neste trabalho para identificar e remover valores discrepantes antes do treinamento dos modelos preditivos.

3.5 Avaliação do Desempenho dos Modelos

A avaliação quantitativa do desempenho dos modelos de regressão é essencial para comparar abordagens e identificar quais apresentam melhor capacidade de generalização. Neste trabalho, utiliza-se um conjunto de métricas que permite analisar o erro sob diferentes

perspectivas estatísticas e operacionais, incluindo o coeficiente de determinação R^2 , Erro Absoluto Médio (MAE), Erro Quadrático Médio (MSE), Raiz do Erro Quadrático Médio (RMSE) e Erro Percentual Absoluto Médio (MAPE). Essas métricas são amplamente adotadas em aplicações industriais e de pesquisa por fornecerem diferentes ângulos sobre a qualidade das previsões (CHAI; DRAXLER, 2014; HYNDMAN; KOEHLER, 2006).

3.5.1 Coeficiente de determinação (R^2)

O coeficiente de determinação (R^2) (Equação 3.14) mede a proporção da variabilidade da variável alvo que é explicada pelo modelo de regressão, sendo definido como:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3.14)$$

em que y_i representa os valores observados, $\hat{y}_i$ os valores preditos pelo modelo e $\bar{y}$ a média dos valores observados. Valores de R^2 próximos de 1 indicam maior capacidade explicativa do modelo, enquanto valores próximos de 0 sugerem baixo poder de ajuste. Em alguns casos, o coeficiente pode assumir valores negativos, indicando desempenho inferior ao de um modelo baseado apenas na média dos dados observados.

3.5.2 Erro Absoluto Médio (MAE)

O Erro Absoluto Médio (MAE) é definido como a média dos valores absolutos dos erros de previsão, conforme a Equação 3.15:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3.15)$$

onde y_i representa o valor real, $\hat{y}_i$ o valor predito, e n o número de observações. Essa métrica mede o erro médio em unidades da variável alvo, sendo facilmente interpretável. Por não elevar os erros ao quadrado, o MAE apresenta menor sensibilidade a outliers em comparação ao MSE (WILLMOTT; MATSUURA, 2005).

3.5.3 Erro Quadrático Médio (MSE)

O Erro Quadrático Médio (MSE) é uma das métricas mais tradicionais em regressão, definida pela Equação 3.16:

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (3.16)$$

Essa métrica penaliza de forma mais severa erros maiores devido à operação de elevação ao quadrado. Essa propriedade torna o MSE útil em aplicações nas quais grandes

desvios não são tolerados, embora também o torne mais sensível à presença de *outliers* (BISHOP, 2006). Além disso, o MSE possui uma interpretação estatística relevante, pois é diretamente relacionado ao estimador de máxima verossimilhança sob a suposição de erros Gaussianos (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

3.5.4 Raiz do Erro Quadrático Médio (RMSE)

A Raiz do Erro Quadrático Médio (RMSE), dada pela Equação 3.17, é obtida pela raiz quadrada do MSE:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (3.17)$$

Diferentemente do MSE, o RMSE retorna o erro médio na mesma unidade da variável alvo, o que facilita a interpretação prática (CHAI; DRAXLER, 2014). Trata-se de uma métrica bastante utilizada em séries temporais e aplicações de previsão em engenharia, por oferecer uma medida clara da magnitude do erro. Entretanto, assim como o MSE, o RMSE é sensível a valores discrepantes.

3.5.5 Erro Percentual Absoluto Médio (MAPE)

O Erro Percentual Absoluto Médio (MAPE) é definido pela Equação 3.18:

$$MAPE = \frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (3.18)$$

Essa métrica expressa o erro como uma porcentagem média em relação aos valores reais, sendo frequentemente utilizada em contextos aplicados em que a interpretabilidade é essencial (HYNDMAN; KOEHLER, 2006). No entanto, apresenta limitações em situações em que y_i assume valores próximos de zero, podendo inflar artificialmente o erro percentual (ARMSTRONG; COLLOPY, 1985). Apesar dessas limitações, o MAPE continua sendo amplamente empregado em estudos de previsão pela sua facilidade de comunicação dos resultados em termos percentuais.

4 Materiais e Métodos

Este capítulo detalha o percurso metodológico e os recursos computacionais empregados no desenvolvimento e nos testes do *AutoDrill*. A estrutura aqui apresentada fundamenta as escolhas técnicas que orientaram a concepção do sistema, garantindo a reprodutibilidade dos experimentos e o rigor na análise dos resultados. Inicialmente, descreve-se a arquitetura modular da solução, com ênfase no *pipeline* de pré-processamento automatizado, que abrange etapas críticas de limpeza, tratamento de *outliers*, filtragem adaptativa de sinais e normalização de atributos. Na sequência, detalha-se a estratégia para a previsão da ROP, fundamentada no uso de *frameworks* de AutoML para a seleção de algoritmos e otimização de hiperparâmetros. Por fim, apresenta-se o protocolo experimental, descrevendo a origem e as características dos conjuntos de dados utilizados, bem como os cenários de teste projetados para avaliar a robustez, a estabilidade e a capacidade de generalização do sistema frente a diferentes contextos geológicos e operacionais.

4.1 Arquitetura do AutoDrill

A Figura 5 apresenta o diagrama da arquitetura do sistema, e sintetiza as etapas do fluxo adotado, desde o pré-processamento dos dados até a geração das previsões da taxa de penetração. O diagrama oferece uma visão geral dos módulos que compõem a solução, permitindo compreender a sequência lógica das operações aplicadas. Cada um dos elementos representados é detalhado nas subseções seguintes.

4.1.1 Conjunto de Dados

Para avaliar o desempenho do *framework* proposto em condições operacionais reais, empregou-se diferentes conjuntos de dados provenientes de operações de perfuração *offshore* disponibilizados publicamente. A adoção de bases originadas de diferentes regiões geográficas, teve como objetivo validar a capacidade de generalização do modelo, assegurando que seu desempenho não se limitasse a um contexto ou ambiente operacional, mas que se estendesse a cenários industriais distintos e representativos da realidade de campo.

Foram selecionados três conjuntos de dados amplamente utilizados na literatura técnica e científica recente: o repositório *Volve Data Sharing* da Equinor ([Equinor, n.d.](#)), o conjunto USROP (*University of Stavanger Rate of Penetration*) ([TUNKIEL, 2020](#)), e a base NEATS NOPIMS (*National Offshore Petroleum Information Management System*) ([National Offshore Petroleum Titles Administrator \(NOPTA\), n.d.](#)). A escolha por esses repositórios fundamentou-se tanto na qualidade e abrangência dos registros quanto na sua

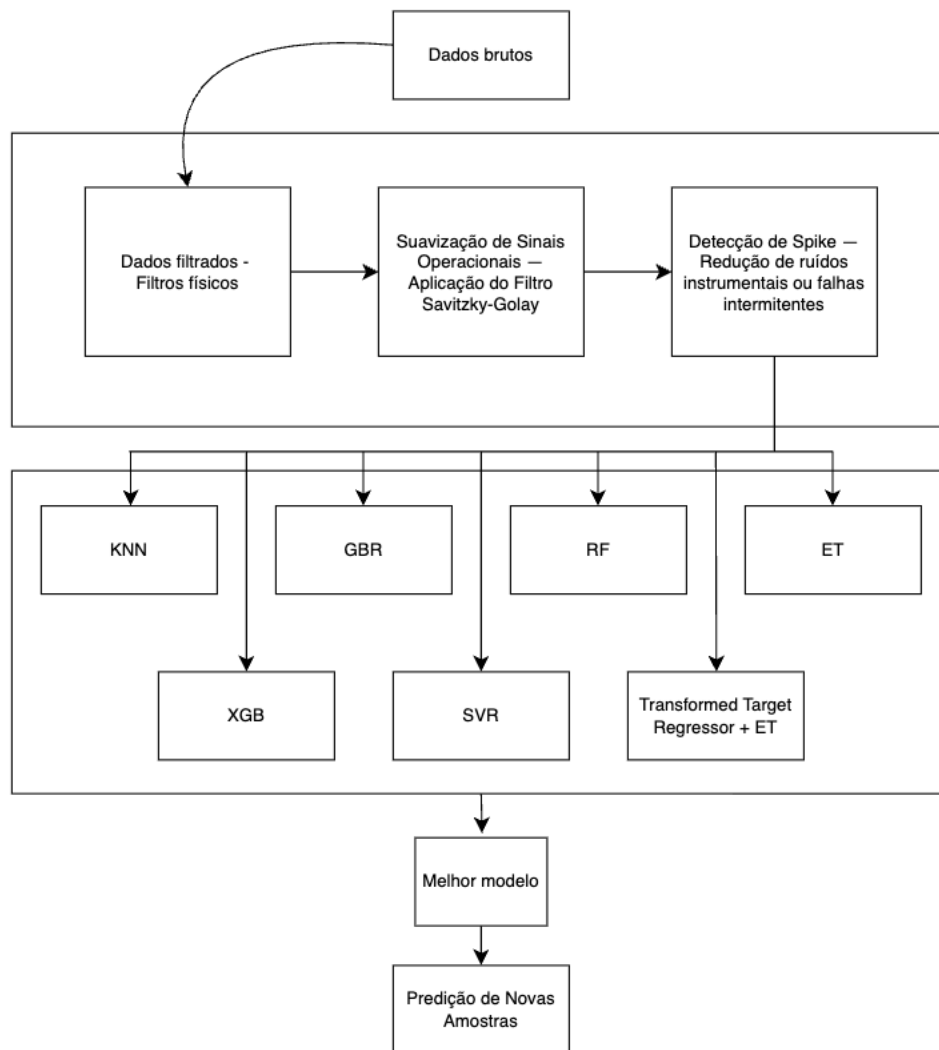


Figura 5 – Arquitetura geral do sistema *AutoDrill*. Figura do autor.

relevância para estudos de modelagem preditiva aplicados à perfuração de poços.

A escolha desse conjunto de dados permitiu contemplar diferentes perfis litológicos, variações significativas nas propriedades mecânicas das formações, mudanças nos regimes operacionais e diferenças nos equipamentos empregados, incluindo diâmetros de broca, características de vazão de fluido, estratégias de controle de peso sobre a broca e velocidade de rotação. Além disso, a heterogeneidade temporal dos repositórios, incluindo dados provenientes de perfurações realizadas em períodos distintos, reflete variações nos sistemas de monitoramento e automação ao longo do tempo. Isso permitiu uma avaliação experimental mais ampla, contemplando possíveis sensibilidades do modelo a mudanças operacionais e aprimoramentos de instrumentação.

4.1.1.1 Conjunto de Dados *Equinor Volve*

O conjunto de dados Volve foi disponibilizado pela Equinor por meio da iniciativa Volve Data Sharing ([Equinor, n.d.](#)). Trata-se de um repositório amplamente utilizado em pesquisas relacionadas à análise de dados de perfuração, em função de sua abrangência, documentação consistente e caráter público, o que favorece a reprodutibilidade e comparação de resultados reportados na literatura. Os dados referem-se a operações conduzidas no campo Volve, localizado no setor norueguês do Mar do Norte, onde foram perfurados poços produtores, injetores e exploratórios entre 2008 e 2016, em ambiente *offshore* com condições operacionais variadas.

Do ponto de vista geológico, o campo insere-se majoritariamente na Formação Hugin, composta por intercalações de arenitos, siltitos, folhelhos e camadas de carvão. Essa heterogeneidade litológica resulta em diferenças relevantes de resistência mecânica, abrasividade e porosidade, influenciando diretamente ROP e os demais parâmetros operacionais. Como consequência, a resposta da broca e o desempenho da perfuração podem apresentar variações expressivas ao longo da sequência de camadas penetradas.

O conjunto inclui registros detalhados de variáveis operacionais, como profundidade medida (DEPT), peso sobre a broca (WOB), rotação (RPM) e taxa de penetração (ROP), torque, vazão do fluido (MFIA), além de informações referentes à lama de perfuração, tipos e diâmetros de broca e litologia interpretada. Os dados são disponibilizados em alta resolução temporal, frequentemente associados a telemetria contínua, cobrindo milhares de horas de operação. Essa granularidade possibilita aplicações em modelagem preditiva, avaliação de desempenho da broca, detecção de eventos e otimização operacional. Por se tratar de um conjunto proveniente de operações reais, o repositório permite ainda analisar efeitos de mudanças de procedimentos, equipamentos e estratégias ao longo do tempo.

As variáveis selecionadas neste trabalho obedeceram a um critério prático importante: foram escolhidas por serem parâmetros que permitem intervenção operacional direta durante o processo de perfuração. Diferentemente de variáveis puramente descritivas ou derivadas da formação geológica, esses parâmetros podem ser ajustados em tempo real, possibilitando ações de otimização da taxa de perfuração.

A combinação de variabilidade litológica e diversidade operacional torna o conjunto Volve particularmente adequado para estudos de predição de ROP, calibração de modelos em diferentes contextos e avaliação de generalização entre poços.

A Tabela 2 apresenta estatísticas descritivas das variáveis selecionadas neste estudo.

4.1.1.2 Conjunto de Dados *USROP*

O conjunto de dados USROP foi desenvolvido pela Universidade de Stavanger com o propósito de fornecer uma base padronizada para estudos de modelagem e otimização

Variável	Média	Desvio Padrão	Mínimo	Máximo	Unidade
DEPT	2092.44	1181.84	144.00	5331.00	m
WOB	6.16	3.93	0.00	110.57	kN
RPM	155.69	42.74	0.00	311.26	rpm
MFIA	3339.68	887.74	-455.02	5345.09	L/min
ROP	32.62	28.78	0.26	373.39	m/h

Tabela 2 – Estatísticas das principais variáveis do conjunto de dados *Equinor Volve*.

da taxa de penetração em operações de perfuração (TUNKIEL, 2020). O repositório reúne registros provenientes de diferentes campanhas de perfuração *offshore* realizadas em plataformas ao longo da costa norueguesa entre o início dos anos 2000 e o início da década de 2010. Os dados são organizados em arquivos segmentados por seções de poço e apresentados de forma normalizada em função da profundidade, o que facilita comparações entre poços distintos e reduz variações decorrentes de diferenças de aquisição e procedimentos operacionais.

O conjunto contempla parâmetros operacionais essenciais, incluindo DEPT, WOB, RPM, MFIA, torque e ROP, e diâmetro de broca. A seleção desses parâmetros é particularmente relevante para estudos de otimização, pois eles representam diretamente o comportamento mecânico do processo de perfuração e sua interação com as propriedades da formação.

As operações registradas no USROP foram realizadas predominantemente em intervalos sedimentares, compostos por arenitos, folhelhos, siltitos e carbonatos. Essa heterogeneidade litológica introduz variação significativa nas respostas operacionais, tornando o conjunto adequado para a avaliação da capacidade de generalização de modelos preditivos em cenários geológicos com variabilidade marcante. A Tabela 3 apresenta estatísticas descritivas das principais variáveis empregadas neste estudo.

Variável	Média	Desvio Padrão	Mínimo	Máximo	Unidade
DEPT	2411.78	1066.44	225.17	4090.00	m
WOB	6.09	3.96	0.00	31.41	kN
RPM	152.37	49.99	0.00	311.23	rpm
MFIA	2873.58	1055.41	185.42	4538.45	L/min
ROP	24.98	15.09	0.33	99.21	m/h

Tabela 3 – Estatística das principais variáveis do conjunto de dados *USROP*

4.1.1.3 Conjunto de Dados *NEATS NOPIMS*

O conjunto de dados NEATS NOPIMS (*National Offshore Petroleum Information Management System*) é mantido pela *National Offshore Petroleum Titles Administrator* (NOPTA) e disponibilizado como parte de um esforço governamental australiano voltado à transparência e ao incentivo à pesquisa na área de exploração e produção *offshore* (National

[Offshore Petroleum Titles Administrator \(NOPTA\)](#), n.d.). A base reúne registros de perfuração provenientes de diferentes campos localizados, em sua maioria, nas bacias de Browse e Carnarvon, situadas na plataforma continental da Austrália Ocidental. Essas bacias estão inseridas em um contexto geológico marcado por unidades sedimentares compostas predominantemente por arenitos, siltitos, folhelhos e carbonatos, que exibem variações significativas em termos de propriedades geomecânicas e condições de fraturamento.

Os dados disponibilizados incluem parâmetros operacionais essenciais ao monitoramento da eficiência de perfuração, tais como profundidade medida, WOB, RPM, MFIA, ROP e diâmetro da broca. A frequência e precisão dos registros são adequadas para análises temporais e para a modelagem preditiva baseada em aprendizado de máquina, uma vez que os valores são capturados em tempo real durante as operações.

Além dos parâmetros de perfuração, o NEATS NOPIMS contém informações relacionadas à caracterização de reservatório e histórico de produção, o que permite análises integradas quando necessário. No entanto, para o escopo desta pesquisa, foram selecionadas apenas as variáveis diretamente relacionadas ao desempenho mecânico da perfuração, devido ao foco na modelagem e otimização da taxa de penetração.

As operações representadas nesse conjunto abrangem um intervalo temporal extenso, incluindo poços perfurados ao longo de diferentes décadas, em um contexto de evolução tecnológica gradual. Isso introduz variabilidade operacional adicional ao banco de dados, relacionada tanto à adoção de novas técnicas de perfuração quanto à diversificação de sistemas de circulação de fluido, configurações de coluna e estratégias de controle de pressão. Essa heterogeneidade é particularmente relevante, pois contribui para testar a capacidade de generalização do *framework* proposto quando aplicado a cenários operacionais distintos.

A Tabela 4 sintetiza as estatísticas descritivas das principais variáveis utilizadas na análise, apresentando seus valores médios, desvios padrão e valores extremos observados ao longo dos poços perfurados.

Variável	Média	Desvio Padrão	Mínimo	Máximo	Unidade
DEPT	2375.91	893.42	168.00	3939.00	m
WOB	15.56	7.19	-8.25	37.00	kN
RPM	168.11	29.50	76.00	221.00	rpm
MFIA	1094.71	155.28	611.00	1330.00	L/min
ROP	54.07	40.58	-999.25	388.29	m/h

Tabela 4 – Estatística das principais variáveis do conjunto de dados *NEATS NOPIMS*.

4.1.2 Pré-Processamento de Dados

A etapa inicial do pré-processamento consiste na aplicação de procedimentos de limpeza destinados a assegurar a consistência física dos dados. Visto que os dados

analisados foram adquiridos em ambiente operacional sujeito a flutuações instrumentais, ruídos de transmissão e interrupções de leitura, é essencial a remoção sistemática de registros inconsistentes ou não representativos do processo real de perfuração.

O fluxo se inicia com a extração de todos os registros com valor nulo na coluna de profundidade (DEPT), variável que atua como referência espacial ao longo do poço e que, adicionalmente, orienta o ordenamento temporal das observações. A ausência dessa variável inviabiliza a reconstituição sequencial da operação, de modo que sua exclusão constitui um requisito fundamental para assegurar a integridade das análises subsequentes.

O próximo passo caracteriza-se pelo estabelecimento dos limites inferior e superior de profundidade, de modo a restringir o conjunto de dados a intervalos específicos. Essa segmentação permitiu isolar trechos de interesse, reduzindo as heterogeneidades associadas a mudanças litológicas, transições de regime operacional ou substituição de ferramentas.

Em seguida, é realizada a padronização do diâmetro da broca (BDIA), mantendo-se apenas as observações correspondentes aos valores de diâmetro mais empregados nas operações consideradas. Tal padronização visa reduzir variações atribuídas à geometria da ferramenta, uma vez que o diâmetro influencia diretamente a relação entre parâmetros mecânicos, como WOB, RPM e ROP.

Em seguida, removeram-se valores fisicamente inviáveis, tais como leituras nulas ou negativas de WOB, RPM, ROP e MFIA. Esse tipo de ocorrência é normalmente associado a falhas transitórias de sensores, reinicialização de sistemas de aquisição ou inconsistências momentâneas na transmissão de dados. A exclusão desses registros evitou a incorporação de padrões artificiais que poderiam induzir viés nas análises estatísticas ou comprometer a capacidade de generalização dos modelos preditivos.

Adicionalmente, eliminaram-se registros duplicados, garantindo que cada linha represente um único estado operacional. Em continuidade, reindexou-se o conjunto de dados de modo a preservar coerência sequencial e rastreabilidade das observações.

4.1.3 Filtragem Digital e Suavização de Sinais

Para reduzir o impacto de ruídos de alta frequência — isto é, flutuações aleatórias ou variações irregulares que podem corromper a tendência real dos parâmetros monitorados — aplicou-se um processo de suavização aos principais parâmetros operacionais contínuos, tais como WOB, MFIA, RPM e ROP, uma vez que esses sinais apresentam variações contínuas e sua suavização deve preservar a forma da curva e a dinâmica operacional.

Nesta etapa, estabeleceu-se a distinção entre ruído e valores anômalos. Considerou-se ruído como pequenas variações aleatórias em torno do valor real do sinal, enquanto os valores anômalos (*spikes*) foram definidos como desvios pontuais de grande magnitude, potencialmente associados a falhas de sensores, eventos operacionais abruptos ou saturação

momentânea de instrumentação. Essa distinção é fundamental, pois ruídos são tratados por suavização, ao passo que valores anômalos requerem mecanismos de detecção e correção específicos.

Para reduzir o ruído preservando a forma do sinal, empregou-se o filtro de Savitzky–Golay, discutido na Seção 3.3.1, que realiza uma suavização local por meio do ajuste de polinômios em janelas móveis ao longo da série temporal. Diferentemente de técnicas mais simples, como médias móveis ou suavização exponencial, o filtro de Savitzky–Golay é reconhecido por preservar características essenciais do sinal, como pontos de inflexão, amplitudes relativas e curvaturas locais, o que o torna particularmente adequado ao processamento de dados instrumentais. Os parâmetros do filtro, especialmente o tamanho da janela e a ordem do polinômio, foram definidos de maneira ajustável, permitindo controlar o grau de suavização de acordo com o comportamento esperado de cada variável. Janelas maiores e polinômios de ordem superior produzem suavização mais intensa, porém podem reduzir a sensibilidade a variações operacionais sutis; parâmetros mais conservadores preservam microtransições relevantes para a interpretação operacional. A Figura 6 exemplifica a aplicação do filtro digital para o sinal de WOB.

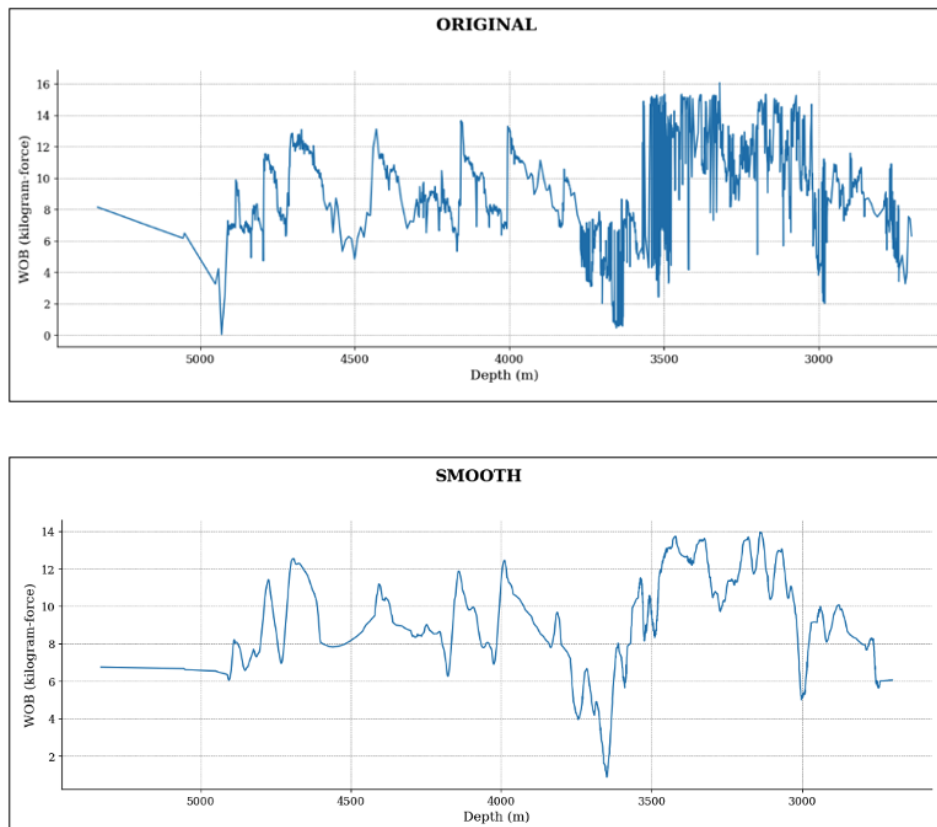


Figura 6 – Sinal original e suavizado (*smooth*) de WOB. Figura do autor.

4.1.4 Detecção e Correção de Descontinuidades (*Spikes*)

Determinados parâmetros operacionais, como RPM e MFIA, apresentam padrões de sinal característicos, porém suscetíveis à ocorrência de descontinuidades abruptas ou *spikes*. Esses eventos correspondem a desvios pontuais e acentuados em relação ao comportamento esperado da série temporal e podem ter origem em falhas de sensores, eventos transitórios não representativos das condições reais de operação, perturbações mecânicas ou ruídos introduzidos pelo sistema de aquisição de dados. A permanência desses valores espúrios nos dados compromete a consistência estatística das variáveis, introduzindo vieses significativos na modelagem, especialmente em algoritmos sensíveis a outliers.

Para mitigar esse problema, é realizada a implementação de um método sistemático de detecção e correção de spikes, composto pelas seguintes etapas principais:

1. **Detecção dos *Spikes*:** Inicialmente, extraiu-se a diferença absoluta entre o sinal original (não suavizado) e o sinal previamente suavizado por meio do filtro *Savitzky–Golay*. Considerou-se como ponto discrepante qualquer valor cuja diferença excedesse um limiar (δ) determinado empiricamente, ajustado conforme amplitudes de variação admissíveis para cada parâmetro operacional. Ressalta-se que a escolha inadequada de δ pode resultar na remoção de variações legítimas (quando muito reduzido) ou na não detecção de anomalias relevantes (quando demasiadamente elevado).
2. **Remoção dos Valores Anômalos:** Após a identificação das descontinuidades, os valores considerados *spikes* foram substituídos por marcações de dados faltantes (NaN). Esse procedimento visa evitar que anomalias pontuais distorçam análises posteriores, mantendo a coerência da estrutura temporal sem introduzir artefatos.
3. **Redução de Granularidade:** Os valores remanescentes foram arredondados a múltiplos específicos, por exemplo, 10, 25 ou 50 unidades, definidos empiricamente. Essa etapa reduz flutuações de alta frequência relacionadas à resolução dos sensores e alinha as séries aos intervalos operacionais usualmente observados, favorecendo maior estabilidade estatística e capacidade de generalização dos modelos.
4. **Interpolação para Reconstrução da Continuidade:** Aplicou-se interpolação linear para preencher os pontos marcados como ausentes, assegurando continuidade temporal e evitando a formação de rupturas artificiais. A interpolação linear é adequada porque essas variáveis apresentam evolução progressiva, sem mudanças abruptas ao longo do tempo ou da profundidade.

A Figura 7 exemplifica a dinâmica descrita acima, representando o sinal original de RPM e o sinal após a remoção de *spikes*.

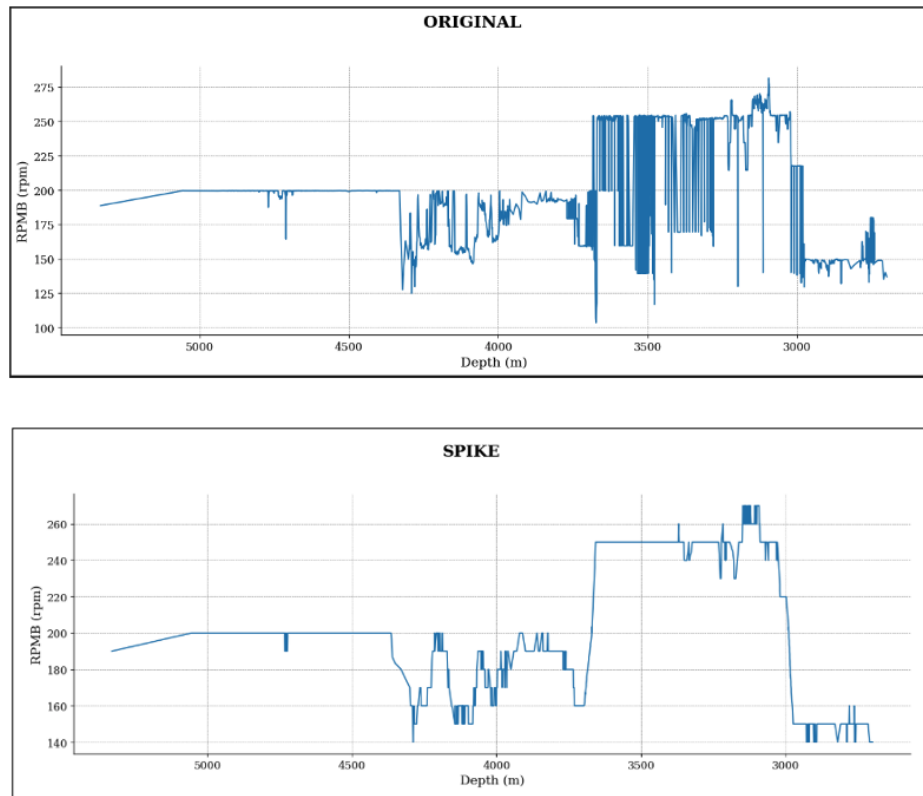


Figura 7 – Sinal original e após correção de (*spikes*) de RPM. Figura do autor.

O procedimento descrito visa preservar a integridade física e operacional das variáveis medidas, reduzindo interferências transitórias e distorções que possam comprometer análises subsequentes ou modelos de aprendizado de máquina.

4.1.5 Otimização e Seleção de Modelos

A etapa de otimização e seleção de modelos teve como finalidade identificar o algoritmo de regressão mais apropriado para a estimativa de ROP, considerando a variabilidade e a complexidade inerentes aos dados operacionais de perfuração. Para esse propósito, selecionou-se um conjunto diversificado de modelos, discutidos na Seção 3.1, representando diferentes paradigmas de aprendizagem, de modo a abranger tanto relações lineares quanto não lineares entre as variáveis.

Inicialmente, os modelos foram organizados em um mapeamento estruturado associado aos respectivos espaços de hiperparâmetros. Para cada algoritmo, aplicaram-se procedimentos de busca estocástica, utilizando o método *RandomizedSearchCV*, o qual seleciona combinações de hiperparâmetros a partir de amostragens aleatórias dentro de um intervalo previamente definido. Diferentemente da busca exaustiva (*GridSearchCV*), que avalia sistematicamente todas as combinações possíveis, o *RandomizedSearchCV* realiza apenas um subconjunto dessas avaliações, o que reduz substancialmente o custo

computacional enquanto mantém boas chances de identificar regiões de desempenho ótimo no espaço de busca. Assim, torna-se possível explorar eficientemente modelos com grande número de hiperparâmetros, equilibrando custo computacional e qualidade do ajuste.

Além disso, foi empregada uma estratégia de validação cruzada do tipo *k-fold cross-validation*, com o objetivo de aumentar a estatística da avaliação dos modelos. Nesse procedimento, o conjunto de treinamento é particionado em k subconjuntos (*folds*), nos quais, iterativamente, um subconjunto é utilizado para validação enquanto os demais são empregados no treinamento. Esse processo permite que diferentes porções dos dados sejam avaliadas ao longo das iterações, reduzindo a dependência de uma única divisão entre treino e validação e proporcionando estimativas mais estáveis da capacidade de generalização do modelo.

A utilização da validação cruzada também contribui para reduzir riscos de *data leakage*, uma vez que os dados utilizados para validação permanecem isolados durante cada etapa de treinamento e ajuste de hiperparâmetros. Dessa forma, operações de aprendizado e otimização são realizadas exclusivamente sobre os subconjuntos de treinamento de cada iteração, preservando a independência entre os dados de treino e validação. Esse procedimento evita que informações do conjunto de validação influenciem indiretamente o treinamento do modelo, reduzindo vieses otimistas na estimativa de desempenho e promovendo maior confiabilidade nos resultados experimentais.

Por fim, essa abordagem permitiu ajustar parâmetros relevantes, como profundidade de árvores, número de estimadores, taxa de aprendizado e parâmetros de regularização, reduzindo riscos de sobreajuste (*overfitting*) e favorecendo maior capacidade de generalização dos modelos avaliados.

Os modelos incluídos nesta etapa contemplaram diferentes categorias:

- **Modelos baseados em instâncias (KNN):** utilizados para capturar padrões locais de comportamento operacional. Por se tratar de um método sensível ao espaço de vizinhança, sua inclusão permite avaliar cenários nos quais pequenas variações dos parâmetros resultam em flutuações significativas no valor da ROP.
- **Modelos de árvores em conjunto (Random Forest, Extra Trees, Gradient Boosting e XGBoost):** selecionou-se esta classe pela capacidade comprovada de representar interações complexas entre variáveis e pela robustez diante de ruídos e valores anômalos. Esses modelos realizam agregação estatística de múltiplos estimadores, contribuindo para a estabilidade preditiva e para a redução da variância do estimador final.
- **Máquina de Vetores de Suporte (SVR):** incluiu-se este método com o intuito de avaliar situações em que o conjunto de dados apresenta distribuição heterogênea

ou regiões esparsas. A maximização da margem, característica do SVR, possibilita maior estabilidade e generalização em cenários de dados irregulares.

- **Transformed Target Regressor com PowerTransformer:** aplicou-se transformação nos valores-alvo para reduzir assimetria e heterocedasticidade, frequentemente observadas em variáveis operacionais de perfuração. Essa transformação melhora a estabilidade numérica do treinamento e favorece o desempenho de modelos sensíveis à distribuição dos resíduos.

Para cada modelo, realizou-se um processo de busca randômica com o objetivo de identificar a melhor configuração de hiperparâmetros dentro do espaço previamente definido e discutido na Seção 4.3. A avaliação das combinações seguiu a métrica apropriada a cada algoritmo, conforme sua função de desempenho padrão ou aquela explicitamente especificada durante a etapa de otimização. Após a seleção da configuração considerada mais adequada, procedeu-se ao ajuste final (*fit*) do modelo sobre o conjunto de dados, disponibilizando ao usuário tanto o modelo treinado quanto as métricas resultantes de seu desempenho para a base analisada.

As métricas de desempenho disponíveis para análise e utilização final do usuário foram: coeficiente de determinação (R^2), erro absoluto médio (MAE), erro quadrático médio (RMSE) e erro percentual absoluto médio (MAPE).

Dessa forma, estabeleceu-se um conjunto de modelagem capaz de acomodar diferentes perfis de dados e condições operacionais. Esse arranjo favorece a obtenção de estimativas de ROP consistentes, em conformidade com requisitos técnicos observados em contextos industriais.

4.2 Estratégia de Testes e Validação Experimental

4.2.1 Cenários de Teste

Os experimentos foram conduzidos a partir dos três conjuntos de dados apresentados na Seção 4.1.1, abrangendo operações de perfuração *offshore* caracterizadas por diferentes condições litológicas, parâmetros operacionais e configurações de sensores.

No caso específico do conjunto de dados Equinor Volve, foram consideradas formações geológicas heterogêneas, incluindo intervalos compostos por arenitos, carbonatos, folhelhos e lamias consolidadas, refletindo condições estratigraficamente complexas e susceptíveis a variações significativas nas propriedades mecânicas da rocha. Adicionalmente, diferentes diâmetros de broca foram empregados, principalmente nas medidas de 8,5" e 12,25", proporcionando diversidade nos regimes de interação broca-formação e permitindo examinar o impacto desse parâmetro na taxa de penetração alcançada.

Cada cenário de teste foi estruturado de maneira a representar combinações operacionais plausíveis, considerando tanto variações na profundidade quanto alterações nos parâmetros de perfuração ao longo do avanço do poço. Os intervalos de profundidade analisados variaram entre 1.660 e 5.330 metros, abrangendo desde seções iniciais de perfuração até intervalos mais profundos e geomecanicamente desafiadores. Em todos os cenários, o conjunto mínimo de variáveis de entrada incluiu profundidade medida, WOB, RPM e MFIA, por constituírem parâmetros fundamentais para a caracterização do desempenho mecânico da perfuração e por serem variáveis que podem ser alteradas durante o processo de perfuração, constituindo um sistema de retroalimentação com o sistema proposto.

Assim, os cenários de teste foram concebidos de forma a aproximar as condições experimentais das situações encontradas em ambientes industriais, permitindo examinar o desempenho do sistema em contextos variados, desde formações homogêneas e previsíveis até ambientes litológicos complexos e altamente heterogêneos.

4.2.2 Análise Comparativa com Soluções AutoML Existentes

Com o objetivo de contextualizar o desempenho do sistema proposto, realizou-se uma comparação com dois *frameworks* AutoML consolidados na literatura: o AutoGluon e o H2O AutoML. Embora tais ferramentas não sejam originalmente direcionadas a aplicações em perfuração de poços de petróleo, representam soluções de AutoML maduras, abrangentes e reconhecidas por sua capacidade de explorar automaticamente diferentes arquiteturas de modelos, estratégias de ajuste de hiperparâmetros e mecanismos de combinação de modelos. Dessa forma, constituem referenciais adequados para avaliação comparativa.

A comparação do AutoDrill com soluções de AutoML existentes foi conduzida com o objetivo de situá-lo no contexto de ferramentas de aprendizado automático. Diferentemente de *frameworks* generalistas, o AutoDrill integra etapas de pré-processamento orientadas por conhecimento físico, tratamento sistemático de discontinuidades e seleção de modelos compatíveis com a natureza não linear de ROP. Essa abordagem permite assegurar que os procedimentos de modelagem sejam adequados à heterogeneidade dos dados operacionais e às particularidades do domínio de perfuração.

4.2.3 Métricas de Avaliação

A avaliação do desempenho dos modelos foi conduzida utilizando o conjunto de métricas previamente apresentado na Seção 3.5, a saber: MAE, MSE, RMSE, R^2 e MAPE. A escolha conjunta dessas métricas justifica-se pelo caráter complementar entre elas, permitindo analisar não apenas o erro absoluto e quadrático, mas também a capacidade explicativa do modelo e a proporcionalidade dos desvios em relação ao valor real.

No contexto da predição de ROP, tais métricas possibilitam avaliar diferentes

aspectos relevantes para a tomada de decisão operacional. O MAE e o MAPE fornecem indicação direta da magnitude média do erro, auxiliando na interpretação prática dos desvios entre valores previstos e observados. O MSE e o RMSE, por sua vez, permitem identificar a dispersão dos erros e penalizam discrepâncias maiores, o que é particularmente útil em cenários de transições abruptas de litologia. Já o coeficiente de determinação (R^2) fornece uma medida global da capacidade do modelo em explicar a variabilidade dos dados ao longo da profundidade analisada.

Assim, o uso simultâneo dessas métricas assegura uma avaliação adequada e equilibrada do desempenho dos modelos, em conformidade com recomendações estabelecidas em estudos de aprendizado de máquina aplicados a problemas de regressão em dados industriais. Dessa forma, torna-se possível comparar de maneira consistente diferentes arquiteturas e configurações de modelos, bem como quantificar melhorias decorrentes das estratégias automáticas de seleção e ajuste implementadas no sistema proposto.

4.3 Aspectos de Implementação e Reprodutibilidade

Esta seção descreve o ambiente computacional, os procedimentos específicos de pré-processamento e as configurações de modelagem empregados na avaliação experimental do sistema. O objetivo é assegurar transparência metodológica e permitir a reprodução integral dos resultados.

4.3.1 Ferramentas e Ambiente

O código foi desenvolvido em *Python*, com uso extensivo das bibliotecas *NumPy* e *Pandas* para manipulação e pré-processamento de dados, *SciPy* para aplicação do filtro de *Savitzky-Golay* na suavização de sinais, *scikit-learn* para seleção de modelos, construção de *pipelines* e cálculo de métricas, e *XGBoost* para modelos de *gradient boosting*. Todos os experimentos foram conduzidos em um ambiente reprodutível, utilizando sementes aleatórias fixas e divisões de validação cruzada sempre que possível.

4.3.2 Limpeza e Pré-processamento de Dados

Antes do treinamento dos modelos, os dados brutos de poços passaram por diversas etapas de pré-processamento, descritas a seguir:

- **Filtragem por profundidade:** seleção dos registros dentro de um intervalo de profundidade de 1.660 e 5.330 metros;
- **Filtragem por diâmetro da broca:** consideração apenas de registros com diâmetros de broca de 8,5" e 12,25";

- **Remoção de valores inválidos:** eliminação de linhas contendo valores não físicos ou incoerentes, como ROP ou WOB negativos ou iguais a zero;
- **Deteção de *outliers*:** identificação de valores discrepantes utilizando o método de Tukey, baseado na amplitude interquartil (IQR). Foram considerados como outliers os valores localizados fora dos limites definidos pelos quartis da distribuição, utilizando limiares de $1.5 \times \text{IQR}$ e $3 \times \text{IQR}$ para identificação de possíveis outliers e outliers extremos, respectivamente. Os valores identificados foram removidos do conjunto de dados antes do treinamento dos modelos;
- **Suavização:** aplicação do filtro de *Savitzky-Golay* em WOB, ROP, RPM e MFIA, com janela de 51 amostras e polinômio de ordem 2, definidos empiricamente;
- **Redução de picos (*Spike Removal*):** arredondamento dos sinais para o múltiplo mais próximo de 10 ou 50, visando melhorar a consistência do sinal.

4.3.3 Configuração do AutoML e Busca de Hiperparâmetros

Foi construído um *pipeline* de aprendizado de máquina flexível utilizando *scikit-learn*, permitindo integração de múltiplos modelos de regressão, transformação opcional do alvo com *PowerTransformer* (método Yeo-Johnson) e ajuste de hiperparâmetros por meio de *RandomizedSearchCV*, com validação cruzada de 5 *folds* e processamento paralelo (`n_jobs=-1`).

Os seguintes modelos e espaços de hiperparâmetros foram explorados:

- **Random Forest:**
`n_estimators = [100, 200, 300, 500],`
`max_depth = [10, 20, 30, 40, None],`
`min_samples_split = [2, 5, 10, 15],`
`min_samples_leaf = [1, 2, 4, 6],`
`max_features = ['auto', 'sqrt', 'log2'],`
`bootstrap = [True, False].`
- **Extra Trees:** mesma grade de hiperparâmetros do Random Forest, com redução de variância aprimorada devido a divisões aleatórias.
- **Gradient Boosting:**
`n_estimators = [50, 100, 200],`
`learning_rate = [0.01, 0.05, 0.1, 0.2],`
`max_depth = [3, 5, 7],`

```
subsample = [0.8, 1.0].
```

- **XGBoost:**

```
learning_rate = [0.01, 0.05, 0.1, 0.3],
```

```
max_depth = [3, 5, 7],
```

```
gamma = [0, 0.1, 0.5],
```

```
min_child_weight = [1, 3, 5],
```

```
n_estimators = [50, 100, 200].
```

- **SVR (Support Vector Regression):**

```
C = [0.1, 1, 10, 100],
```

```
epsilon = [0.01, 0.1, 0.5],
```

```
gamma = ['scale', 'auto', 0.01, 0.001].
```

- **K-Nearest Neighbors (KNN):**

```
n_neighbors = [3, 5, 10, 15],
```

```
weights = ['uniform', 'distance'],
```

```
p = [1, 2] (distância Manhattan ou Euclidiana).
```

4.3.4 Reprodutibilidade

Todos os modelos foram inicializados com *random_state* = 42. As partições de validação cruzada foram mantidas constantes entre todos os métodos avaliados. Além disso, registros de execução, parâmetros de configuração e versões das bibliotecas foram documentados, assegurando que o *pipeline* experimental possa ser integralmente reproduzido em futuros estudos.

5 Resultados e Discussão

Este capítulo apresenta os resultados experimentais obtidos com o AutoDrill na predição de ROP nos cenários geológicos e conjuntos de dados definidos, comparando-os com dois *frameworks* de AutoML existentes, AutoGluon e H2O. A avaliação destaca tanto o desempenho preditivo quanto o custo computacional, enfatizando o desempenho sob condições adversas com ruído e irregularidades típicas de operações de perfuração reais. Além disso, ao final do capítulo, os resultados são contextualizados através da comparação com resultados de estudos previamente reportados na literatura.

5.1 Avaliação Comparativa com AutoML Existentes

A Tabela 5 resume o desempenho dos *frameworks* (AutoDrill, AutoGluon e H2O) em diferentes conjuntos de dados. Foram utilizadas métricas padrão de regressão, incluindo MAE, MSE, RMSE, R^2 e MAPE. Os conjuntos de dados contemplam perfis litológicos diversos, como calcário, folhelho, carbonato e formações mistas, proporcionando um conjunto abrangente de cenários de teste. Os melhores resultados em cada cenário estão destacados em negrito.

Os resultados indicam que o AutoDrill apresenta desempenho competitivo e frequentemente superior em diferentes cenários geológicos. Nos conjuntos de dados EQUINOR, o AutoDrill alcançou RMSEs de 2,65 (calcário), 3,13 (folhelho) e 3,57 (carbonato), comparáveis ou melhores que AutoGluon e significativamente superiores ao H2O. Apesar do AutoGluon apresentar MAE ligeiramente menor em folhelho (1,47 versus 1,65), o AutoDrill obteve melhor RMSE. No conjunto USROP, todos os *frameworks* apresentaram desempenho competitivo, com o AutoDrill alcançando o melhor MAE (0,006) e RMSE (0,12). Esse comportamento pode estar associado à maior padronização e consistência operacional do conjunto USROP, cujos dados apresentam menor variabilidade e ruído quando comparados aos cenários geológicos mais heterogêneos dos demais conjuntos analisados.

A discrepância de desempenho observada no conjunto NEATS NOPIMS é o indicativo mais forte da eficácia do AutoDrill. Enquanto *frameworks* genéricos como H2O e AutoGluon apresentaram coeficientes de determinação muito baixos ($R^2 \leq 0,19$), o AutoDrill manteve a estabilidade preditiva ($R^2 = 0,94$, $RMSE = 1,70$). Esse fenômeno decorre do fato de o NEATS NOPIMS possuir alta densidade de *outliers* e sinais transientes que confundem os modelos de *stacking* convencionais, mas que são tratados pelo *pipeline* de filtragem e transformação do sistema proposto. O desempenho inferior de ferramentas genéricas de AutoML nessas condições desafiadoras deve-se a três fatores principais: (i) presença de seções litologicamente heterogêneas (calcilutito e siltito), resultando em

Data	Litologia	AutoML	MAE	MSE	RMSE	MAPE	R^2
EQUINOR	Calcário	AutoDrill	1,54	7,03	2,65	11,08%	0,85
		AutoGluon	1,58	6,85	2,61	12,73%	0,86
		H2O	2,25	14,22	3,45	21,10%	0,72
EQUINOR	Argilito/Folhelho	AutoDrill	1,65	9,80	3,13	11,27%	0,90
		AutoGluon	1,47	9,60	3,09	8,95%	0,89
		H2O	2,16	18,17	3,46	11,49%	0,82
EQUINOR	Carbonato	AutoDrill	1,98	12,74	3,57	14,34%	0,88
		AutoGluon	2,38	18,17	4,26	16,49%	0,83
		H2O	2,77	18,22	5,11	19,13%	0,82
USROP	-	AutoDrill	0,006	0,016	0,12	0,04%	0,99
		AutoGluon	0,032	0,44	0,66	0,15%	0,99
		H2O	0,95	3,22	1,75	5,15%	0,99
NEATS	Calclutito/Siltito	AutoDrill	0,94	2,90	1,70	4,36%	0,94
		AutoGluon	3,52	1205,26	34,17	13,07%	0,19
		H2O	4,93	1691,16	41,12	18,00%	0,01

Tabela 5 – Comparação do desempenho de *frameworks* de AutoML em diferentes conjuntos de dados e litologias, agrupados pelo método AutoML. Valores em negrito indicam a melhor métrica (mais baixa) em geral, exceto para R^2 , em que é destacado o valor mais alto.

dinâmicas não estacionárias; (ii) dados de sensores incompletos e ruidosos, não tratados adequadamente pelos pré-processamentos padrão; e (iii) alta variabilidade e assimetria na distribuição da ROP devido a transições operacionais e mudanças geológicas. O *pipeline* de pré-processamento do AutoDrill, com suavização, correção de picos e interpolação, demonstra adaptabilidade e se sobressai em cenários irregulares e heterogêneos.

Esses achados reforçam a relevância de estratégias de pré-processamento específicas ao domínio. Dados reais de perfuração frequentemente apresentam ruído, falhas de sensores e inconsistências operacionais que podem comprometer o aprendizado dos modelos. Nesse contexto, os mecanismos integrados do AutoDrill de limpeza, suavização e correção de picos contribuem para reduzir os efeitos dessas irregularidades, permitindo uma representação mais consistente da dinâmica de perfuração.

Além do pré-processamento, o desempenho observado também está associado à etapa de modelagem adotada pelo sistema. O AutoDrill combina transformações especializadas dos sinais com a seleção de modelos mais adequados a dados heterogêneos e parcialmente ruidosos, como abordagens baseadas em *ensembles* de árvores. Em particular, a utilização de estratégias como o *Transformed Target Regressor* (TTR) em conjunto com modelos Extra Trees contribui para maior adaptação frente à variabilidade dos dados e distribuições assimétricas da ROP, o que favorece a estabilidade preditiva mesmo em

cenários operacionais ruidosos.

As métricas de desempenho oferecem perspectivas complementares sobre a qualidade do modelo. Por exemplo, embora o AutoGluon tenha superado levemente o AutoDrill em MAE no conjunto folhelho, o AutoDrill apresentou R^2 superior, indicando menor presença de *outliers* extremos e maior captura da variância total. Isso evidencia a relevância de múltiplos critérios de avaliação, especialmente em dados operacionais reais, nos quais grandes erros podem gerar impactos operacionais significativos.

Observou-se redução significativa de desempenho em cenários de maior variabilidade e presença de ruído, como no conjunto NEATS. Nessas condições, os baixos valores de R^2 e os elevados erros de predição indicam dificuldades na generalização dos modelos gerados automaticamente, o que pode limitar sua aplicação prática sem intervenção manual pesada.

5.2 Comparação com a Literatura

Visando situar o desempenho do AutoDrill no panorama do estado da arte, a Tabela 6 apresenta comparações com estudos recentes sobre predição de ROP utilizando técnicas de aprendizado de máquina. Os valores representam as métricas centrais reportadas, visto que o desvio padrão não é uniformemente informado na literatura consultada. Considerou-se, para referência, o valor médio de desempenho do AutoDrill nos cinco cenários de teste descritos na Tabela 5, que inclui diferentes contextos geológicos e condições operacionais.

Estudo	Modelo	MAE	R^2
(HEGDE et al., 2017)	Random Forest	8,3	0,84
(SABAH et al., 2019)	MLP – PSO	1,12	0,93
(MEHRAD et al., 2020)	LSSVM-COA	0,41	0,80
(ALALI et al., 2021)	ANN	9,22	0,94
(XI; GENG; WANG, 2024)	CNN-LSTM-Attention	0,97	0,93
AutoDrill (Média)	AutoML	1,22	0,91

Tabela 6 – Comparação de modelos de predição de ROP presentes na literatura.

Os modelos da literatura diferem em arquitetura e características dos conjuntos de dados, o que limita comparações diretas. Ainda assim, o AutoDrill alcança desempenho comparável ou superior em termos de MAE e R^2 em múltiplos conjuntos de dados e litologias. A média do MAE (1,22) e do R^2 (0,91) obtida pelo AutoDrill demonstra não apenas poder preditivo competitivo, mas também adaptabilidade frente a variações na qualidade dos dados e litologia. Talvez o ponto mais importante de destaque seja que, embora os outros trabalhos sejam otimizados para conjuntos específicos de dados, e alguns deles obtenham resultados levemente superiores, o AutoDrill manteve um desempenho

competitivo em cenários mais amplos e diversos, utilizando um *pipeline* adaptável e generalizado com mínima intervenção manual.

A análise evidencia que o AutoDrill oferece abordagem escalável e transferível para predição de ROP, com benefícios práticos em operações nas quais as condições dos dados variam significativamente.

5.3 Análise e Discussão

Os resultados obtidos evidenciam o impacto da integração do pré-processamento orientado ao domínio e automação de modelagem no desempenho de *frameworks* AutoML aplicados à perfuração. Em particular, o tratamento adequado de dados provenientes de sensores de campo — incluindo etapas de suavização, correção de picos e filtragem de ruído — mostrou-se fundamental para reduzir artefatos presentes nas séries operacionais, que frequentemente comprometem o aprendizado de modelos quando *pipelines* genéricos são utilizados. Esse comportamento é consistente com a natureza dos dados de perfuração, que costumam apresentar irregularidades decorrentes de falhas de sensores, vibrações mecânicas e mudanças abruptas nas condições operacionais.

Outro aspecto relevante observado nos experimentos é a capacidade do AutoDrill de manter desempenho estável mesmo em conjuntos de dados heterogêneos e com diferentes características litológicas. Essa habilidade decorre da combinação entre automação do processo de modelagem e mecanismos de pré-processamento adaptados ao domínio, que permitem reduzir a sensibilidade do *pipeline* a ruído, valores ausentes e distribuições não estacionárias. Em contraste, *frameworks* genéricos de AutoML tendem a assumir características estatísticas mais regulares nos dados de entrada, o que pode resultar em degradação de desempenho quando aplicados diretamente a dados operacionais de perfuração.

Essas propriedades tornam o AutoDrill adequado para cenários operacionais dinâmicos, como operações de perfuração *offshore*, nas quais as condições geológicas e operacionais podem variar significativamente ao longo do tempo. Nesses contextos, a capacidade de adaptação automática do *pipeline* de modelagem e a necessidade de intervenção manual mínima tornam-se fatores essenciais para viabilizar a adoção de modelos de aprendizado de máquina em ambientes industriais. Além disso, os tempos de treinamento relativamente reduzidos observados nos experimentos indicam que o *framework* pode ser aplicado em cenários que demandam atualização frequente dos modelos, como em estratégias de re-treinamento contínuo ou em aplicações que requerem previsões de baixa latência.

Comparado à literatura, o AutoDrill posiciona-se como uma alternativa equilibrada, considerando não somente a capacidade de generalização, mas também custo computacional e facilidade de adoção e adaptação. Embora alguns modelos especializados possam alcançar

desempenho elevado em conjuntos de dados específicos, tais abordagens frequentemente dependem de extensiva engenharia de atributos e ajustes manuais de hiperparâmetros. O AutoDrill, por sua vez, mantém desempenho competitivo em diferentes cenários geológicos e conjuntos de dados, o que sugere que a combinação entre automação e adaptação ao domínio pode representar uma estratégia eficaz para aplicações de aprendizado de máquina em engenharia de perfuração. Dessa forma, o *framework* se apresenta como uma abordagem escalável e prática para apoiar processos de tomada de decisão em fluxos de trabalho de otimização de perfuração baseados em dados.

5.4 Principais Contribuições

Este trabalho apresenta contribuições relevantes para a aplicação de técnicas de aprendizado de máquina na predição de ROP em operações de petróleo e gás. As principais contribuições podem ser sintetizadas nos seguintes pontos:

- **Proposição do framework AutoDrill:** desenvolvimento de um *framework* de AutoML orientado ao domínio da perfuração, projetado para automatizar etapas críticas do *pipeline* de aprendizado de máquina, incluindo pré-processamento, treinamento, seleção de modelos e avaliação de desempenho.
- **Integração de pré-processamento específico ao domínio:** implementação de mecanismos de tratamento de dados operacionais de perfuração, como suavização de séries temporais, correção de picos e interpolação de valores ausentes, visando mitigar ruídos e irregularidades frequentemente presentes em dados provenientes de sensores de campo.
- **Capacidade de adaptação a diferentes contextos geológicos:** desenvolvimento de um *pipeline* capaz de ajustar automaticamente o processo de modelagem às características de diferentes conjuntos de dados e litologias, permitindo maior robustez e generalização em cenários operacionais variados.
- **Avaliação comparativa com frameworks AutoML consolidados:** realização de experimentos comparativos entre o AutoDrill e *frameworks* amplamente utilizados na literatura, como AutoGluon e H2O AutoML, demonstrando desempenho competitivo ou superior em diferentes conjuntos de dados de perfuração.
- **Demonstração de aplicabilidade em dados reais de perfuração:** validação experimental do *framework* utilizando múltiplos conjuntos de dados com características geológicas distintas, evidenciando a capacidade do AutoDrill de lidar com dados heterogêneos e condições operacionais complexas.

- **Contribuição para a automação de *pipelines* de análise de dados em perfuração:** proposição de uma abordagem escalável que reduz a necessidade de intervenção manual no processo de modelagem, facilitando a integração de técnicas de aprendizado de máquina em ambientes industriais de engenharia de perfuração.

De forma geral, as contribuições deste trabalho demonstram que a integração entre automação de modelagem e conhecimento específico do domínio pode ampliar a eficácia de aplicações de aprendizado de máquina em problemas complexos da indústria de óleo e gás.

6 Conclusões

O *framework* AutoDrill, proposto neste trabalho, tem como objetivo fornecer uma abordagem estruturada e automatizada para a predição da Taxa de Perfuração em operações de óleo e gás, incorporando conhecimento específico do domínio ao longo de todo o *pipeline* de aprendizado de máquina. Diferentemente de *frameworks* AutoML de propósito geral, o AutoDrill foi concebido para lidar com as particularidades dos dados de perfuração, incluindo ruído de sensores, valores ausentes, presença de *outliers* e variações decorrentes de mudanças geológicas e operacionais.

Ao longo deste estudo, foi demonstrado que a integração de técnicas de pré-processamento orientadas ao domínio — como suavização de séries temporais, correção de picos e interpolação — exerce papel fundamental na melhoria da qualidade dos dados de entrada e, conseqüentemente, no desempenho dos modelos preditivos. Os resultados experimentais evidenciam que o AutoDrill apresenta desempenho competitivo e, em diversos cenários, superior a *frameworks* consolidados, como AutoGluon e H2O AutoML, especialmente em conjuntos de dados caracterizados por maior heterogeneidade e irregularidade.

Outro aspecto relevante do AutoDrill é sua capacidade de generalização aliada à adaptação automática. O *framework* foi projetado como uma solução generalista que se especializa dinamicamente para diferentes conjuntos de dados, ajustando seu *pipeline* de modelagem conforme as características específicas de cada contexto geológico e operacional. Essa propriedade permite reduzir significativamente a necessidade de intervenção manual, tornando o processo de modelagem mais eficiente, reproduzível e escalável.

Além disso, os tempos de treinamento observados dos modelos indicam que o AutoDrill possui potencial para aplicação em ambientes industriais, incluindo cenários que demandam atualização frequente de modelos ou suporte à tomada de decisão em tempo quase real. Dessa forma, o *framework* contribui não apenas para o avanço acadêmico na área de AutoML aplicado à engenharia de perfuração, mas também para a viabilização prática de soluções baseadas em dados no contexto operacional.

Em síntese, este trabalho demonstra que a combinação entre automação de modelagem e incorporação de conhecimento específico do domínio constitui uma estratégia eficaz para lidar com a complexidade inerente à predição de ROP. O AutoDrill se apresenta, portanto, como uma abordagem promissora para ampliar a utilização de técnicas de aprendizado de máquina na indústria de petróleo e gás, oferecendo suporte mais confiável e eficiente à tomada de decisão.

6.1 Trabalhos Futuros

O AutoDrill tem como objetivo fornecer suporte para futuras equipes de engenharia, cientistas de dados e operadores implementarem soluções de perfuração de poços de forma mais fácil e sistemática. Portanto, um dos principais pontos de evolução consiste na validação do AutoDrill em um conjunto mais amplo de dados operacionais, incluindo diferentes bacias, formações geológicas e condições de perfuração, a fim de avaliar sua robustez em cenários ainda mais diversos.

Outra direção relevante refere-se à incorporação de informações temporais e sequenciais mais complexas, por meio da integração com modelos baseados em séries temporais ou arquiteturas profundas, como redes neurais recorrentes ou modelos baseados em atenção. Essa abordagem pode permitir uma melhor captura de dependências temporais presentes nos dados de perfuração.

Adicionalmente, a inclusão de mecanismos de interpretabilidade representa uma extensão importante, possibilitando compreender com maior clareza a influência das variáveis operacionais nas previsões de ROP. Isso é particularmente relevante em ambientes industriais, onde a transparência dos modelos pode impactar diretamente sua adoção.

Por fim, a integração do AutoDrill com sistemas de monitoramento em tempo real e plataformas de suporte à decisão pode ampliar significativamente seu impacto prático, contribuindo para a otimização de operações de perfuração e redução de custos operacionais. Essas perspectivas reforçam o potencial do AutoDrill como uma base para o desenvolvimento de soluções mais avançadas e adaptativas no contexto da engenharia de perfuração.

Referências

- AL-ABDULJABBAR, A.; ELKATATNY, S.; MAHMOUD, M. Rop prediction in carbonate formations with lost circulation using modified bourgoyne and young models. *Journal of Petroleum Science and Engineering*, Elsevier, v. 208, p. 109632, 2021. Citado na página 17.
- ALALI, A. M. et al. Hybrid data driven drilling and rate of penetration optimization. *Journal of Petroleum Science and Engineering*, Elsevier, v. 200, p. 108075, 2021. Citado na página 58.
- ALTMAN, N. S. *An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression*. [S.l.]: The American Statistician, 1992. v. 46. 175–185 p. Citado na página 25.
- ARMSTRONG, J. S.; COLLOPY, F. Long-range forecasting: From crystal ball to computer. *John Wiley & Sons*, 1985. Citado na página 40.
- BARBOSA, L. F. F. et al. Machine learning methods applied to drilling rate of penetration prediction and optimization-a review. *Journal of Petroleum Science and Engineering*, Elsevier, v. 183, p. 106332, 2019. Citado na página 18.
- BINGHAM, M. G. A new approach to interpreting rock drillability. *Oil and Gas Journal*, v. 62, p. 93–96, 1964. Citado na página 16.
- BISHOP, C. M. *Pattern Recognition and Machine Learning*. [S.l.]: Springer, 2006. Citado na página 40.
- BOURGOYNE, A. T.; YOUNG, F. S. A multiple regression approach to optimal drilling and abnormal pressure detection. *Society of Petroleum Engineers Journal*, v. 14, n. 04, p. 371–384, 1974. Citado na página 16.
- BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, p. 5–32, 2001. Citado na página 25.
- BRENJKAR, E.; DELIJANI, E. B. Computational prediction of the drilling rate of penetration (rop): A comparison of various machine learning approaches and traditional models. *Journal of Petroleum Science and Engineering*, Elsevier, v. 210, p. 110033, 2022. Citado na página 12.
- BRET-ROUZAUT, N. Economics of oil and gas production. In: _____. *The Palgrave Handbook of International Energy Economics*. Cham: Springer International Publishing, 2022. p. 3–22. ISBN 978-3-030-86884-0. Disponível em: <https://doi.org/10.1007/978-3-030-86884-0_1>. Citado na página 12.
- CHAI, T.; DRAXLER, R. R. Root mean square error (rmse) or mean absolute error (mae)?—arguments against avoiding rmse in the literature. *Geoscientific Model Development*, Copernicus GmbH, v. 7, n. 3, p. 1247–1250, 2014. Citado 2 vezes nas páginas 39 e 40.

- CHEN, J. et al. A simple method for reconstructing a high-quality ndvi time-series data set based on the savitzky–golay filter. *Remote sensing of Environment*, Elsevier, v. 91, n. 3-4, p. 332–344, 2004. Citado na página 35.
- CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. [S.l.: s.n.], 2016. p. 785–794. Citado na página 28.
- CHERAGHI, Y.; KORD, S.; MASHAYEKHIZADEH, V. Application of machine learning techniques for selecting the most suitable enhanced oil recovery method; challenges and opportunities. *Journal of Petroleum Science and Engineering*, Elsevier, v. 205, p. 108761, 2021. Citado na página 19.
- CLARK, B.; LEE, F. *What is Gradient Boosting?* 2025. Acesso em: 30 ago. 2025. Disponível em: <<https://www.ibm.com/think/topics/gradient-boosting>>. Citado 2 vezes nas páginas 26 e 27.
- COULSTON, J. W. et al. Approximating prediction uncertainty for random forest regression models. *Photogrammetric Engineering & Remote Sensing*, American Society for Photogrammetry and Remote Sensing, v. 82, n. 3, p. 189–197, 2016. Citado na página 25.
- COURONNÉ, R.; PROBST, P.; BOULESTEIX, A.-L. Random forest versus logistic regression: a large-scale benchmark experiment. *BMC bioinformatics*, Springer, v. 19, n. 1, p. 270, 2018. Citado na página 26.
- COVER, T.; HART, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, IEEE, v. 13, n. 1, p. 21–27, 1967. Citado na página 24.
- DEVELOPERS, S. learn. *ExtraTreesRegressor*. 2024. <<https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html>>. Acesso em: 30 ago. 2025. Citado na página 29.
- ELAHIFAR, B.; HOSSEINI, E. Automated real-time prediction of geological formation tops during drilling operations: an applied machine learning solution for the norwegian continental shelf. *Journal of Petroleum Exploration and Production Technology*, Springer, v. 14, n. 6, p. 1661–1703, 2024. Citado na página 20.
- ELKATATNY, S.; AL-ABDULJABBAR, A.; ABDELGAWAD, K. A new model for predicting rate of penetration using an artificial neural network. *Sensors*, MDPI, v. 20, n. 7, p. 2058, 2020. Citado 2 vezes nas páginas 12 e 19.
- Equinor. *Volve Data Sharing*. n.d. <<https://www.equinor.com/energy/volve-data-sharing>>. Accessed: 2025-04-21. Citado 2 vezes nas páginas 41 e 43.
- ERICKSON, M. et al. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020. Citado 3 vezes nas páginas 31, 32 e 33.
- FARHAD, M. et al. Hybrid machine learning models for drilling rate of penetration prediction using metaheuristic optimization algorithms. *Journal of Petroleum Science and Engineering*, Elsevier, v. 213, p. 110459, 2022. Citado na página 19.
- FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, JSTOR, p. 1189–1232, 2001. Citado na página 26.

- GAN, C. et al. Prediction of drilling rate of penetration (rop) using hybrid support vector regression: A case study on the shennongjia area, central china. *Journal of Petroleum Science and Engineering*, Elsevier, v. 181, p. 106200, 2019. Citado na página 19.
- GANDELMAN, R. A. *Predição da ROP e otimização em tempo real de parâmetros operacionais na perfuração de poços de petróleo offshore*. Tese (Doutorado) — Ph. D. thesis, UFRJ, 2012. Citado na página 17.
- GEURTS, P.; ERNST, D.; WEHENKEL, L. Extremely randomized trees. *Machine learning*, Springer, v. 63, p. 3–42, 2006. Citado 2 vezes nas páginas 28 e 29.
- GIJSBERS, P. et al. An open source automl benchmark. *arXiv preprint arXiv:1907.00909*, 2019. Citado na página 32.
- GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. Citado na página 24.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. [S.l.]: Springer, 2009. Citado na página 40.
- HE, X.; ZHAO, K.; CHU, X. Automl: A survey of the state-of-the-art. *Knowledge-Based Systems*, Elsevier, v. 212, p. 106622, 2021. Citado 4 vezes nas páginas 13, 19, 20 e 31.
- HEGDE, C. et al. Analysis of rate of penetration (rop) prediction in drilling using physics-based and data-driven models. *Journal of petroleum science and Engineering*, Elsevier, v. 159, p. 295–306, 2017. Citado 2 vezes nas páginas 12 e 58.
- HEGDE, C. et al. Real-time predictive capabilities of analytical and machine learning rate of penetration (rop) models. *Journal of Petroleum Science and Engineering*, Elsevier, v. 178, p. 834–852, 2019. Citado na página 17.
- HORNIK, K.; STINCHCOMBE, M.; WHITE, H. Multilayer feedforward networks are universal approximators. *Neural networks*, Elsevier, v. 2, n. 5, p. 359–366, 1989. Citado na página 17.
- HOSSAIN, M. E.; AL-MAJED, A. A. *Drilling Engineering*. [S.l.]: Wiley, 2015. Citado na página 17.
- HU, H. et al. Dynamic response and strength failure analysis of bottomhole under balanced drilling condition. *Journal of Petroleum Science and Engineering*, Elsevier, v. 194, p. 107561, 2020. Citado na página 18.
- HUTTER, F.; KOTTHOFF, L.; VANSCHOREN, J. Automated machine learning: Methods, systems, challenges. *Automated Machine Learning*, Springer, p. 1–33, 2019. Citado na página 13.
- HYNDMAN, R. J.; KOEHLER, A. B. Another look at measures of forecast accuracy. *International Journal of Forecasting*, Elsevier, v. 22, n. 4, p. 679–688, 2006. Citado 2 vezes nas páginas 39 e 40.
- JÄÄSKELÄINEN, J. A. Automl performance in model fitting: a comparative study of selected machine learning competitions in 2012-2019. 2022. Citado na página 34.

JOHN, A.; SADASIVAN, J.; SEELAMANTULA, C. S. Adaptive savitzky–golay filtering in non-gaussian noise. *IEEE Transactions on Signal Processing*, IEEE, v. 69, n. 5021–5036, 2021. Citado na página 35.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. *Science*, American Association for the Advancement of Science, v. 349, n. 6245, p. 255–260, 2015. Citado na página 24.

KARMAKER, S. K. et al. Automl to date and beyond: Challenges and opportunities. *Acm computing surveys (csur)*, ACM New York, NY, v. 54, n. 8, p. 1–36, 2021. Citado 2 vezes nas páginas 20 e 21.

KONSTANTINOVSKY, T. *Introduction to the Savitzky-Golay Filter: A Comprehensive Guide (Using Python)*. 2024. Medium article. Disponível em: <<https://medium.com/pythoneers/introduction-to-the-savitzky-golay-filter-a-comprehensive-guide-using-python-b2dd07a8e2ce>>. Citado 2 vezes nas páginas 35 e 36.

KOR, K. et al. Penetration rate prediction in heterogeneous formations: A geomechanical approach through machine learning. *Journal of Petroleum Science and Engineering*, Elsevier, v. 207, p. 109138, 2021. Citado na página 17.

KUBAT, M. *An Introduction to Machine Learning*. [S.l.]: Springer, 2017. Citado na página 24.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, Nature Publishing Group, v. 521, n. 7553, p. 436–444, 2015. Citado na página 24.

LEDELL, E.; POIRIER, S. H2o automl: Scalable automatic machine learning. In: *Proceedings of the AutoML Workshop at ICML*. [S.l.: s.n.], 2020. v. 2020, p. 24. Citado 5 vezes nas páginas 19, 20, 31, 33 e 34.

LIU, H. et al. Rate of penetration prediction method for ultra-deep wells based on lstm–fnn. *Applied Sciences*, MDPI, v. 12, n. 15, p. 7731, 2022. Citado na página 19.

LORENA, A. C.; CARVALHO, A. C. P. L. F. de. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, v. 14, n. 2, p. 43–67, Dec. 2007. Disponível em: <https://seer.ufrgs.br/index.php/rita/article/view/rita_v14_n2_p43-67>. Citado 2 vezes nas páginas 29 e 30.

MAME, M. et al. Application of extra-trees regression and tree-structured parzen estimators optimization algorithm to predict blast-induced mean fragmentation size in open-pit mines. *Applied Sciences*, MDPI, v. 15, n. 15, p. 8363, 2025. Citado na página 29.

MEHRAD, M. et al. Developing a new rigorous drilling rate prediction model using a machine learning technique. *Journal of Petroleum Science and Engineering*, Elsevier, v. 192, p. 107338, 2020. Citado na página 58.

MICHAEL, I. Benchmarking automl frameworks for buyer intent prediction: Autoglun vs. h2o. ai vs. google automl. 2025. Citado na página 31.

NASCIMENTO, A.; NOGUEIRA, L. *Exploração de petróleo em camadas do pré-sal no Brasil: um estudo de caso no poço 1-sps-50*. Tese (Doutorado) — Dissertação de mestrado, UFItajubá, 2010. Citado na página 12.

- NATEKIN, A.; KNOLL, A. Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, Frontiers Media SA, v. 7, p. 21, 2013. Citado na página 26.
- National Offshore Petroleum Titles Administrator (NOPTA). *National Offshore Petroleum Information Management System (NOPIMS)*. n.d. Acesso em: 21 abr. 2025. Disponível em: <<https://public.neats.nopta.gov.au/nopims>>. Citado 2 vezes nas páginas 41 e 45.
- RATTAN, P.; PENRICE, D. D.; SIMONETTO, D. A. Artificial intelligence and machine learning: what you always wanted to know but were afraid to ask. *Gastro Hep Advances*, Elsevier, v. 1, n. 1, p. 70–78, 2022. Citado na página 23.
- REGRESSION, N. An introduction to kernel and nearest-neighbor. *The American Statistician*, v. 46, n. 3, p. 175–185, 1992. Citado 2 vezes nas páginas 24 e 25.
- RIAZI, M. et al. Modelling rate of penetration in drilling operations using rbf, mlp, lssvm, and dt models. *Scientific Reports*, Nature Publishing Group, v. 12, n. 1, p. 11650, 2022. Citado na página 19.
- RONALD, I.; CONOVER, W. *A modern approach to statistics*. [S.l.]: John Wiley and Sons, New York, 1983. Citado na página 30.
- SABAH, M. et al. A machine learning approach to predict drilling rate using petrophysical and mud logging data. *Earth Science Informatics*, Springer, v. 12, p. 319–339, 2019. Citado na página 58.
- SAVITZKY, A.; GOLAY, M. J. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, ACS Publications, v. 36, n. 8, p. 1627–1639, 1964. Citado 2 vezes nas páginas 35 e 37.
- SHIRKOV, A. et al. Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505*, 2020. Citado 2 vezes nas páginas 19 e 20.
- SIMON, G. J.; ALIFERIS, C. Artificial intelligence and machine learning in health care and medical sciences: best practices and pitfalls. Springer nature, 2024. Citado na página 23.
- SINGH, K. et al. Cloud-based rop prediction and optimization in real time using supervised machine learning. In: URTEC. *SPE/AAPG/SEG Unconventional Resources Technology Conference*. [S.l.], 2019. p. D023S041R002. Citado na página 12.
- SMOLA, A. J.; SCHÖLKOPF, B. A tutorial on support vector regression. *Statistics and computing*, Springer, v. 14, p. 199–222, 2004. Citado 2 vezes nas páginas 29 e 30.
- SOARES, C.; DAIGLE, H.; GRAY, K. Evaluation of pdc bit rop models and the effect of rock strength on model coefficients. *Journal of Natural Gas Science and Engineering*, Elsevier, v. 34, p. 1225–1236, 2016. Citado na página 18.
- SUN, Y. et al. Automl in the wild: Obstacles, workarounds, and expectations. In: *Proceedings of the 2023 CHI conference on human factors in computing systems*. [S.l.: s.n.], 2023. p. 1–15. Citado na página 20.
- TEALE, R. The concept of specific energy in rock drilling. *International Journal of Rock Mechanics and Mining Sciences*, v. 2, n. 1, p. 57–73, 1965. Citado na página 16.

- TRUONG, A. et al. Towards automated machine learning: Evaluation and comparison of automl approaches and tools. *2019 IEEE International Conference on Big Data*, p. 1471–1479, 2019. Citado na página 20.
- TUKEY, J. W. *Exploratory Data Analysis*. [S.l.]: Addison-Wesley, 1977. Citado na página 38.
- TUNKIEL, A. *USROP: University of Stavanger Rate of Penetration Data*. 2020. Accessed: 2025-04-21. Disponível em: <<https://github.com/AndrzejTunkiel/USROP>>. Citado 2 vezes nas páginas 41 e 44.
- TUNKIEL, A. T.; SUI, D.; WIKTORSKI, T. Impact of data pre-processing techniques on recurrent neural network performance in context of real-time drilling logs in an automated prediction framework. *Journal of Petroleum Science and Engineering*, Elsevier, v. 208, p. 109760, 2022. Citado na página 12.
- WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate Research*, v. 30, n. 1, p. 79–82, 2005. Citado na página 39.
- XI, C.; GENG, Y.; WANG, H. A new rop prediction method based on cnn-lstm-attention. In: IEEE. *2024 International Conference on New Trends in Computational Intelligence (NTCI)*. [S.l.], 2024. p. 379–383. Citado na página 58.
- YEO, I.-K.; JOHNSON, R. A. A new family of power transformations to improve normality or symmetry. *Biometrika*, Oxford University Press, v. 87, n. 4, p. 954–959, 2000. Citado na página 30.
- YING, X. An overview of overfitting and its solutions. In: IOP PUBLISHING. *Journal of physics: Conference series*. [S.l.], 2019. v. 1168, p. 022022. Citado na página 23.
- ZÖLLER, M. L.; HUBER, M. F. Benchmark and survey of automated machine learning frameworks. *Journal of Artificial Intelligence Research*, v. 70, p. 409–472, 2021. Citado 2 vezes nas páginas 13 e 19.