

UNIVERSIDADE FEDERAL DE ALAGOAS
CENTRO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA CIVIL

ANTONIO PAULO AMANCIO FERRO

**MODELOS PREDITIVOS PARA ROP COMO SUPORTE À OTIMIZAÇÃO EM
TEMPO REAL DE PARÂMETROS OPERACIONAIS NA PERFURAÇÃO DE POÇOS
DE PETRÓLEO**

Maceió-AL

2024

ANTONIO PAULO AMANCIO FERRO

**MODELOS PREDITIVOS PARA ROP COMO SUPORTE À OTIMIZAÇÃO EM
TEMPO REAL DE PARÂMETROS OPERACIONAIS NA PERFURAÇÃO DE POÇOS
DE PETRÓLEO**

Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Engenharia Civil da Universidade Federal de Alagoas, como parte dos requisitos para obtenção do título de Mestre em Engenharia Civil na área de concentração em Estruturas.

Orientadora: Prof.^a Dr.^a Aline da Silva Ramos Barboza

Maceió-AL

2024

Catálogo na Fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico

Bibliotecário: Marcelino de Carvalho Freitas Neto – CRB-4 – 1767

F395m Ferro, Antonio Paulo Amancio.

Modelos preditivos para ROP como suporte à otimização em tempo real de parâmetros operacionais na perfuração de poços de petróleo / Antonio Paulo Amancio Ferro. - 2024.

149 f. : il.

Orientadora: Aline da Silva Ramos Barboza.

Dissertação (Mestrado em Engenharia Civil) – Universidade Federal de Alagoas. Centro de Tecnologia. Maceió, 2024.

Bibliografia: f. 124-131.

Apêndices: f. 132-149.

1. Perfuração. 2. Taxa de penetração. 3. Aprendizagem de máquina. 4. Tempo real. I. Título.

CDU:621.951



Universidade Federal de Alagoas – UFAL
Unidade Acadêmica Centro de Tecnologia – CTEC
Programa de Pós-Graduação em Engenharia Civil – PPGEC



MODELOS PREDITIVOS PARA ROP COMO SUPORTE À OTIMIZAÇÃO EM TEMPO REAL DE PARÂMETROS OPERACIONAIS NA PERFURAÇÃO DE POÇOS DE PETRÓLEO

ANTONIO PAULO AMANCIO FERRO

Dissertação submetida à banca examinadora do Programa de Pós-Graduação em Engenharia Civil da Universidade Federal de Alagoas e aprovada no dia 21 do mês de março do ano de 2024.

Banca Examinadora:



Documento assinado digitalmente

ALINE DA SILVA RAMOS BARBOZA

Data: 21/03/2024 12:32:08-0300

Verifique em <https://validar.iti.gov.br>

Profa. Dra. Aline da Silva Ramos Barboza
(Orientadora – PPGEC/UFAL)



Documento assinado digitalmente

EDUARDO NOBRE LAGES

Data: 21/03/2024 12:53:26-0300

Verifique em <https://validar.iti.gov.br>

Prof. Dr. Eduardo Nobre Lages
(Avaliador Interno – PPGEC/UFAL)



Documento assinado digitalmente

CHARLTON OKAMA DE SOUZA

Data: 21/03/2024 16:54:27-0300

Verifique em <https://validar.iti.gov.br>

Dr. Charlton Okama de Souza
(Avaliador Externo ao Programa – PETROBRAS)

AGRADECIMENTOS

Aos meus pais, José Climério Ferro Neto e Maria Betânia Amancio Ferro, por todo apoio e incentivo ao longo da minha trajetória e por estarem presentes me ajudando em todos os momentos difíceis. Sem vocês, não seria possível.

À minha orientadora, Prof.^a Dr.^a Aline da Silva Ramos Barboza, pela confiança no desenvolvimento deste trabalho, pela paciência, pelos conselhos valiosos e também por todas as oportunidades oferecidas que contribuíram na minha formação acadêmica e que motivaram o desenvolvimento deste trabalho.

Ao Prof. Dr. Lucas Pereira de Gouveia, por também ter acompanhado de perto o desenvolvimento deste trabalho, pelas longas e frutíferas discussões, que contribuíram para a finalização desta dissertação, e em minha formação acadêmica, por sanar dúvidas e por estar sempre disposto a ajudar.

A todos os colegas do projeto CAESAR 2.0, em especial a Francisco Binas, pelos conselhos e discussões que contribuíram no desenvolvimento deste trabalho, direta e indiretamente, e a Erasmo Bezerra, pelo auxílio na indicação de correlações empíricas utilizadas neste trabalho.

Aos professores do Programa de Pós-graduação em Engenharia Civil da UFAL: Dr. Adeildo Ramos, Dr. Eduardo Nobre, Dr. Eduardo Toledo, Dr. William Lira, Dr. David Amorim, Dr. Francisco Patrick por todos os ensinamentos ao longo do curso.

Aos colegas e amigos do mestrado, Otávio Rodrigues, Gilberto Lucas e Kevin Tenório, pelos momentos de descontração, companheirismo e também por toda ajuda ao longo do curso.

Aos membros da banca examinadora pelos comentários, sugestões e contribuições, que ajudaram a melhorar a qualidade deste trabalho.

Ao Laboratório de Computação Científica e Visualização (LCCV/UFAL), pela infraestrutura disponibilizada, à Fundação Universitária de Desenvolvimento de Extensão e Pesquisa - FUNDEPES, pelo apoio administrativo, e à PETROBRAS, pelo suporte financeiro, por meio do projeto de pesquisa, desenvolvimento e inovação identificado pelo número ANP 22763-7.

À CAPES pelo apoio financeiro no início do curso.

A todos que de alguma forma contribuíram para a realização deste trabalho. Muito obrigado.

*Este trabalho é dedicado aos meus pais,
José Climério e Maria Betânia.*

RESUMO

FERRO, Antonio Paulo Amancio. **Modelos preditivos para ROP como suporte à otimização em tempo real de parâmetros operacionais na perfuração de poços de petróleo**. 2024.

Dissertação (Mestrado em Engenharia Civil) – Programa de Pós-Graduação em Engenharia Civil, Universidade Federal de Alagoas, Maceió - Alagoas.

A taxa de penetração (*Rate of Penetration* - ROP) é um parâmetro de grande interesse na otimização em tempo real da perfuração. Uma ROP maior diminui o tempo de perfuração e pode levar a reduções significativas de custo na construção do poço. Modelos preditivos são aplicados para prever a ROP resultante baseado em dados medidos durante a perfuração e viabilizam a determinação de parâmetros operacionais ótimos, como peso sobre a broca (*Weight on Bit* - WOB), rotação da broca (*Revolutions per Minute* - RPM) e a vazão do fluido, quando aliados a técnicas de otimização. Obter modelos de ROP precisos é uma tarefa difícil, dada a grande quantidade de fatores que afetam a ROP. O presente trabalho envolve o estudo de diferentes modelos de ROP. Modelos tradicionais: Bourgoyne & Young e Energia Mecânica Específica, são comparados com modelos de aprendizagem de máquina: Redes Neurais Artificiais (*Artificial Neural Networks* - ANN) e Florestas Aleatórias (*Random Forest* - RF). Dados públicos de 7 poços são estruturados em um dataset com dados relevantes para avaliação da performance de diferentes modelos preditivos, como parâmetros operacionais, dados de brocas, litologia, perfis geofísicos, gradiente de poro pressão e resistência não confinada. Nas análises comparativas, métricas de erro, como o erro médio absoluto (*Mean Absolute Error* - MAE) e a raiz do erro médio quadrático (*Root Mean Square Error* - RMSE), são comparadas entre os diferentes modelos considerando cada um dos 7 poços. Analisa-se a significância estatística com modelo de Bourgoyne & Young para compreender os efeitos mais significativos na ROP. A interpretabilidade dos modelos tradicionais em conjunto com o refino de hiperparâmetros são abordagens adotadas para empregar modelos de aprendizagem de máquina com *inputs* significativos e com maior capacidade preditiva. Em seguida, duas estratégias encontradas na literatura para utilização de modelos preditivos para ROP, aplicados na otimização em tempo real, são comparadas: modelos treinados com dados de poços de correlação, ou com os dados do próprio poço, simulando processo de aquisição gradual dos dados (aprendizagem contínua). Os resultados obtidos neste trabalho indicam melhor performance de modelos de aprendizagem de máquina em relação aos modelos tradicionais. O modelo RF apresenta melhor performance nas análises comparativas, tendo erros menores e menor custo computacional. É possível identificar a relevância do torque e da inclusão de dados da formação (Delta-T compressional) nos modelos de aprendizagem de máquina. Identifica-se que a estratégia de aprendizagem contínua é capaz de alcançar menores erros, ainda que se verifica que ambas estratégias são capazes de gerar previsões adequadas.

Palavras-chaves: Perfuração. Taxa de Penetração. Aprendizagem de Máquina. Tempo Real.

ABSTRACT

FERRO, Antonio Paulo Amancio. **Predictive models for ROP as support for real-time optimization of operational drilling parameters in oil well drilling**. 2024. Dissertation (Master in Civil Engineering) – Postgraduate Program in Civil Engineering, Federal University of Alagoas, Maceió - Alagoas.

The Rate of Penetration (ROP) is a parameter of great interest in real-time drilling optimization. A higher ROP reduces drilling time and can lead to significant cost reductions in well construction. Predictive models for ROP are applied to predict ROP based on measured data while drilling, enabling the determination of optimal operational parameters such as RPM, WOB, and fluid flow when combined with optimization techniques. Obtaining more accurate ROP models is a challenging task due to the large number of factors interacting nonlinearly. This study involves the examination of different ROP models, including traditional ones, like Bourgoyne & Young and Specific Mechanical Energy, and their adaptations, compared with machine learning models such as Artificial Neural Networks (ANN) and Random Forests (RF). Public data from 7 wells were structured into a dataset with relevant information for evaluating the performance of different models in estimating ROP, including operational parameters, drill-bit data, lithology, geophysical logging data, pore pressure gradient, and unconfined compressive strength. In comparative analyses, error metrics such as Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are compared among the different models for each of the 7 wells. Statistical significance analysis is performed with the Bourgoyne & Young model to understand more significant effects on ROP. The interpretability of traditional models, along with hyperparameter tuning, is adopted to employ machine learning models with more meaningful inputs and greater predictive capacity. Next, two strategies found in the literature for using predictive models for ROP in real-time optimization are compared: models trained with offset well data or with the well's own data, simulating a gradual data acquisition process (continuous learning). The results indicate better performance of machine learning models compared to traditional models. The RF model shows overall better performance in comparative analyses, with smaller errors and lower computational cost. The relevance of torque and the inclusion of formation data (Delta-T compressional) in machine learning models is identified. Also, continuous learning strategy can achieve lower errors, although both strategies are capable of generating appropriate predictions.

Keywords: Drilling. Rate of Penetration. Machine Learning. Real-Time.

LISTA DE FIGURAS

Figura 1 – Evolução da produção de petróleo e gás (<i>onshore</i> e <i>offshore</i>) de 2021 a 2023	15
Figura 2 – Aplicação da aprendizagem de máquina para estimativa da ROP em sistema de otimização	18
Figura 3 – Etapas gerais da metodologia	20
Figura 4 – Esquema de uma sonda de perfuração rotativa	22
Figura 5 – Exemplo de brocas com partes móveis	24
Figura 6 – Exemplo de brocas sem partes móveis	25
Figura 7 – Representação dos cones de uma broca RCB e seus parâmetros principais .	26
Figura 8 – Posicionamento dos cortadores em brocas PDC	27
Figura 9 – Disposição de equipamentos do LWD no BHA	30
Figura 10 – Resposta típica da ROP ao aumento do WOB	32
Figura 11 – Resposta típica da ROP ao aumento do RPM	33
Figura 12 – Relação entre a ROP e diferentes propriedades dos fluidos	35
Figura 13 – Efeito do aumento do diferencial de pressão na ROP	35
Figura 14 – Distribuição de métodos de aprendizagem de máquina aplicados ao problema de predição da ROP	45
Figura 15 – Duas abordagens de modelos preditivos que podem ser aplicadas na otimização em tempo real	48
Figura 16 – Os paradigmas principais da aprendizagem de máquina	49
Figura 17 – O problema geral do aprendizado supervisionado	50
Figura 18 – Representação gráfica de um <i>perceptron</i>	53
Figura 19 – Representação gráfica da arquitetura de MLP	55
Figura 20 – Funções de ativação e suas derivadas	56
Figura 21 – Representação do funcionamento do gradiente descendente	57
Figura 22 – Curvas de nível em domínios com e sem correção da escala	61
Figura 23 – Princípio dos segmentos proporcionais utilizado no escalonamento	61
Figura 24 – Esquema de agregação nas RF	62
Figura 25 – Processo executado no método de árvores de decisão	63
Figura 26 – Esquema do método de reamostragem <i>bootstrap</i>	65
Figura 27 – Exemplo de particionamento do espaço dos preditores no modelo RF	67
Figura 28 – Representação do sobre-ajuste e sub-ajuste	68
Figura 29 – Representação do processo de validação cruzada <i>k-fold</i>	70
Figura 30 – Esquema do método de parada antecipada	72
Figura 31 – Esquema de funcionamento da busca em <i>grid</i>	74
Figura 32 – Balanço precisão e interpretabilidade dos modelos citados	74
Figura 33 – Mapeamento dos poços utilizados no novo <i>dataset</i> por meio do <i>Google Earth</i>	79
Figura 34 – Comparação entre o Delta-T medido de um poço em diferentes tipo de arquivo	80

Figura 35 – Dados agrupados de todos os poços do <i>dataset</i>	82
Figura 36 – Estratégias de predição da ROP comparadas neste trabalho	84
Figura 37 – Detalhe da estratégia de predição da ROP por aprendizagem contínua	85
Figura 38 – Resultado da modelagem da ROP com 4 modelos tradicionais para o poço F05	87
Figura 39 – Resultado da modelagem da ROP com 4 modelos tradicionais para o poço F14	88
Figura 40 – Resultado da modelagem da ROP com 4 modelos tradicionais para o poço F15S	88
Figura 41 – Resultado da modelagem da ROP com 4 modelos tradicionais para o poço F04	89
Figura 42 – Resultado da modelagem da ROP com 4 modelos tradicionais para o poço F11_T2	89
Figura 43 – Resultado da modelagem da ROP com 4 modelos tradicionais para o poço F10	90
Figura 44 – Resultado da modelagem da ROP com 4 modelos tradicionais para o poço F12	90
Figura 45 – Significância estatística dos preditores do modelo BYM sem multicolinearidade	97
Figura 46 – Significância estatística dos preditores do modelo BYMmod sem multicoli- nearidade	98
Figura 47 – <i>Heatmap</i> com as métricas de erro em m/h dos modelos tradicionais para cada poço	98
Figura 48 – <i>Heatmap</i> com as métricas de erro em m/h dos modelos de aprendizagem de máquina com preditores diferentes e para cada poço	102
Figura 49 – Resposta dos modelos RF e ANN nos poços F05 e F14	103
Figura 50 – Resposta dos modelos RF e ANN nos poços F15S e F04	103
Figura 51 – Resposta dos modelos RF e ANN nos poços F11_T2 e F10	104
Figura 52 – Resposta dos modelos RF e ANN no poço F12	104
Figura 53 – <i>Heatmap</i> com as métricas de erro em m/h na predição para cada poço	107
Figura 54 – <i>Heatmap</i> com as métricas de erro em m/h dos modelos de aprendizagem de máquina com preditores diferentes e para cada poço na estratégia todos para um	108
Figura 55 – <i>Heatmap</i> com as métricas de erro em m/h com modelos tradicionais e de aprendizado de máquina na estratégia todos para um	109
Figura 56 – Resposta dos modelos RF e ANN nos poços F05 e F14	109
Figura 57 – Resposta dos modelos RF e ANN nos poços F15S e F04	110
Figura 58 – Resposta dos modelos RF e ANN nos poços F11_T2 e F10	110
Figura 59 – Resposta dos modelos RF e ANN no poço F12	111
Figura 60 – Resposta dos modelos RF e ANN no poços F12	111
Figura 61 – <i>Heatmap</i> com as métricas de erro em m/h dos modelos de aprendizagem de máquina com preditores diferentes e para cada poço na estratégia de aprendizagem contínua	113
Figura 62 – <i>Heatmap</i> com as métricas de erro em m/h com modelos tradicionais e de aprendizado de máquina na estratégia de aprendizado contínuo	114
Figura 63 – Comparação entre estratégia de aprendizagem contínua e todos para um para todos os modelos e em todos os poços	115

Figura 64 – Resultados da aprendizagem contínua (Preditores 4) nos poços F05 e F14 ao longo da profundidade em comparação com a litologia dos poços	116
Figura 65 – Resultados da aprendizagem contínua (Preditores 4) nos poços F15S e F04 ao longo da profundidade em comparação com a litologia dos poços	116
Figura 66 – Resultados da aprendizagem contínua (Preditores 4) nos poços F11_T2 e F10 ao longo da profundidade em comparação com a litologia dos poços	117
Figura 67 – Resultados da aprendizagem contínua (Preditores 4) nos poços F12 ao longo da profundidade em comparação com a litologia dos poços	117
Figura 68 – Comparação entre os tempos de treinamento na estratégia de aprendizagem contínua (Preditores 4) dos modelos de aprendizagem de máquina	118
Figura 69 – Comparação entre os hiperparâmetros dos modelos de aprendizagem de máquina (Preditores 4) na estratégia de aprendizagem contínua	119
Figura 70 – Comparação entre gradiente de pressão de poros de dois poços e medições pontuais pelo método MDT	133
Figura 71 – Detalhes sobre o código <i>Dull grading</i> segundo a IADC	136
Figura 72 – Histogramas agrupados das principais informações presentes no <i>dataset</i> para cada poço	138
Figura 73 – Plots das principais informações do poço F05 ao longo da profundidade medida	139
Figura 74 – Plots das principais informações do poço F14 ao longo da profundidade medida	140
Figura 75 – Plots das principais informações do poço F15S ao longo da profundidade medida	141
Figura 76 – Plots das principais informações do poço F04 ao longo da profundidade medida	142
Figura 77 – Plots das principais informações do poço F11 T2 ao longo da profundidade medida	143
Figura 78 – Plots das principais informações do poço F10 ao longo da profundidade medida	144
Figura 79 – Plots das principais informações do poço F12 ao longo da profundidade medida	145
Figura 80 – Plot da ROP, UCS e litologia para o Poço F05	146
Figura 81 – Plot da ROP, UCS e litologia para o Poço F14	146
Figura 82 – Plot da ROP, UCS e litologia para o Poço F15S	147
Figura 83 – Plot da ROP, UCS e litologia para o Poço F04	147
Figura 84 – Plot da ROP, UCS e litologia para o Poço F11_T2	148
Figura 85 – Plot da ROP, UCS e litologia para o Poço F10	148
Figura 86 – Plot da ROP, UCS e litologia para o Poço F12	149

LISTA DE TABELAS

Tabela 1 – Parâmetros importantes nas brocas RCB	26
Tabela 2 – Parâmetros de tamanho dos cortadores em brocas PDC	27
Tabela 3 – Parâmetros de orientação dos cortadores em brocas PDC	27
Tabela 4 – Fatores que afetam a ROP	32
Tabela 5 – Parâmetros intrínsecos do modelo	38
Tabela 6 – Valores dos parâmetros de normalização	38
Tabela 7 – Sugestões para alteração nas normalizações	41
Tabela 8 – Parâmetros de aprendizagem do modelo e hiperparâmetros principais da RF e ANN	71
Tabela 9 – Informações sobre poços analisados	78
Tabela 10 – Intervalos de disponíveis para os novos poços	79
Tabela 11 – Atributos no novo <i>dataset</i> presente em cada poço	80
Tabela 12 – Litologias encontradas nos poços analisados e suas descrições	81
Tabela 13 – Informações necessárias nos modelos tradicionais analisados	83
Tabela 14 – Informações sobre os intervalos da aprendizagem contínua para cada poço .	86
Tabela 15 – Coeficientes do modelo BYM para cada poço	91
Tabela 16 – Coeficientes do modelo BYMmod para cada poço	91
Tabela 17 – Coeficientes do modelo BYMSoares para cada poço	91
Tabela 18 – Coeficientes do modelo MSE para cada poço	92
Tabela 19 – FIV calculado para o modelo BYM em cada um 7 poços	94
Tabela 20 – FIV calculado para o modelo BYMmod em cada um 7 poços	95
Tabela 21 – FIV calculado para o modelo BYM sem problema de multicolinearidade em cada um 7 poços	95
Tabela 22 – FIV calculado para o modelo BYMmod sem problema de multicolinearidade em cada um 7 poços	96
Tabela 23 – Conjuntos de preditores utilizados nos modelos de aprendizagem de máquina	99
Tabela 24 – <i>Grid</i> de hiperparâmetros do modelo RF	99
Tabela 25 – <i>Grid</i> de hiperparâmetros do modelo ANN	100
Tabela 26 – Melhores hiperparâmetros e tempo de treinamento do modelo ANN (Predito- res 4) treinado com os dados de cada poço	105
Tabela 27 – Melhores hiperparâmetros e tempo de treinamento do modelo RF (Preditores 4) treinado com os dados de cada poço	105
Tabela 28 – RMSE médio das 5 partições na busca em <i>grid</i> do modelo ANN considerando pior e melhor combinação de hiperparâmetros	106
Tabela 29 – RMSE médio das 5 partições na busca em <i>grid</i> do modelo RF considerando pior e melhor combinação de hiperparâmetros	106

Tabela 30 – Melhores hiperparâmetros e tempo de treinamento do modelo ANN (Predito-
res 2) na estratégia todos para um 112

Tabela 31 – Melhores hiperparâmetros e tempo de treinamento do modelo RF (Preditores
2) na estratégia todos para um 113

Tabela 32 – Informações extraídas sobre as brocas que são utilizadas nos poços analisados 134

LISTA DE ABREVIATURAS E SIGLAS

ANN	<i>Artificial Neural Network</i>
BYM	<i>Bourgoyne & Young Model</i>
ECD	<i>Equivalent Circulating Density</i>
FIV	Fator de Inflação da Variância
HSI	<i>Hydraulic Horsepower per Square Inch</i>
MD	<i>Measured Depth</i>
MWD	<i>Measurement While Drilling</i>
MAE	<i>Mean Absolute Error</i>
MSE	<i>Mechanical Specific Energy</i>
PDC	<i>Polycrystalline Diamond Cristal</i>
PPG	<i>Pore Pressure Gradient</i>
RF	<i>Random Forest</i>
ROP	<i>Rate of Penetration</i>
RPM	<i>Revolutions per Minute</i>
RMSE	<i>Root Mean Absolute Error</i>
RCB	<i>Roller Cone Bit</i>
TVD	<i>True Vertical Depth</i>
WOB	<i>Weight on Bit</i>
LWD	<i>Logging While Drilling</i>
UCS	<i>Unconfined Compressive Rock Strength</i>
SVM	<i>Support Vector Machine</i>
TFA	<i>Total Flow Area</i>

LISTA DE SÍMBOLOS

ρ	Densidade do fluido de perfuração
q	Vazão do fluido de perfuração
T	Torque
F_j	Força hidráulica de impacto do jato aplicado pela broca
D_b	Diâmetro da broca
h	Fração desgastada do dente da broca
ΔP_b	Perda de pressão na broca
P_{of}	<i>Pump-Off Force</i>
γ_b	Inclinação do poço
μ_s	Coefficiente de atrito da coluna de perfuração
μ_b	Coefficiente de atrito da broca
E_m	Efficiência mecânica da broca
R^2	Coefficiente de determinação
Δt	Delta-T Compressional

SUMÁRIO

1	INTRODUÇÃO	15
1.1	Considerações iniciais	15
1.2	Motivação	17
1.3	Objetivos	19
1.4	Metodologia	20
1.5	Estrutura do trabalho	21
2	PERFURAÇÃO: CONSIDERAÇÕES BÁSICAS	22
2.1	Sonda de perfuração	22
2.2	Característica das brocas	23
2.3	Sistema de monitoramento da perfuração	28
3	MODELAGEM DA ROP: PREDIÇÃO E OTIMIZAÇÃO	31
3.1	Fatores que influenciam a ROP	31
3.2	Modelos tradicionais para ROP	36
3.2.1	Modelos baseados na energia mecânica específica	42
3.3	Modelos de aprendizado de máquina para ROP	44
3.4	Otimização aplicada a modelo baseado em dados	46
4	APRENDIZAGEM DE MÁQUINA	49
4.1	Aprendizado supervisionado	50
4.2	Modelos de aprendizagem aplicados à regressão	52
4.2.1	Redes Neurais Artificiais	53
4.2.1.1	Estratégia de treinamento da ANN	56
4.2.1.2	Método de retropropagação	58
4.2.1.3	Efeito de escala nos dados de entrada em ANNs	60
4.2.2	Florestas Aleatórias	62
4.2.2.1	Árvores de Decisão	62
4.2.2.2	A estratégia <i>bagging</i> e árvores de decisão descorrelacionadas	64
4.2.2.3	Processo de treinamento dos modelos RF	66
4.3	Seleção de modelos e sobre-ajuste	67
4.3.1	Partição dos dados	69
4.3.2	Refino de hiperparâmetros e o método de busca em <i>grid</i>	70
4.4	Interpretabilidade dos modelos	74
4.4.0.1	Importância de atributos em regressão linear	75
5	CONJUNTO DE DADOS E PLANEJAMENTO DAS ANÁLISES COM- PARATIVAS	77

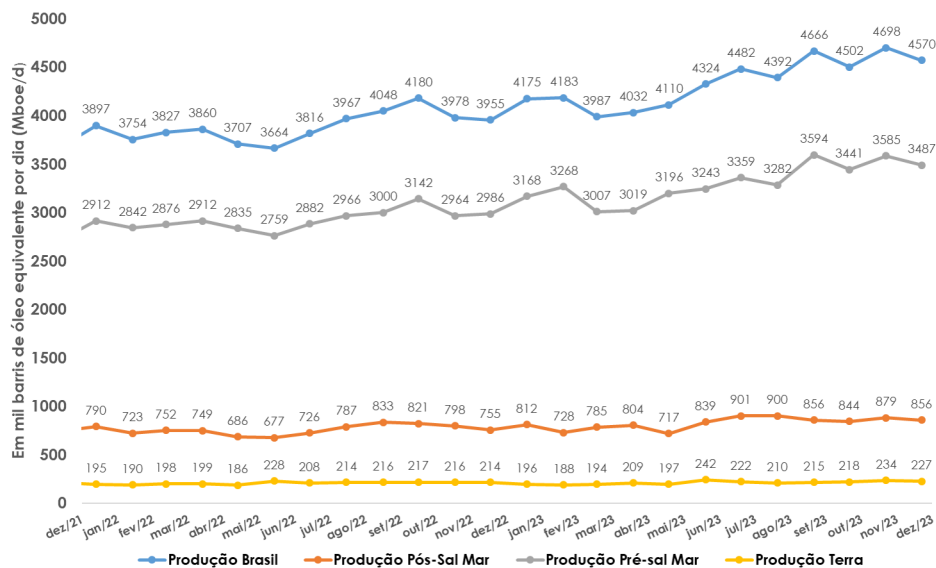
5.1	Novo <i>dataset</i> de referência para estudos sobre a ROP	78
5.1.1	Aspectos relevantes sobre os dados utilizados	80
5.2	Modelos utilizados neste trabalho	83
5.3	Estratégias para predição da ROP investigadas	84
5.4	Recursos empregados	86
6	RESULTADOS E DISCUSSÕES	87
6.1	Modelagem do ROP: Métodos tradicionais	87
6.2	Modelagem do ROP: Aprendizagem de máquina	99
6.3	Estratégia de predição da ROP: Todos para um	106
6.4	Estratégia de predição da ROP: Aprendizagem contínua	113
7	CONCLUSÕES	121
7.1	Considerações finais	121
7.2	Sugestões de trabalhos futuros	123
	REFERÊNCIAS	124
	APÊNDICE A – GRADIENTE DE PRESSÃO DE POROS	132
	APÊNDICE B – ATRIBUTOS DAS BROCAS	134
	APÊNDICE C – CORRELAÇÕES EMPÍRICAS PARA UCS	137
	APÊNDICE D – DADOS RELEVANTES NOS POÇOS <i>VOLVE</i>	138

1 INTRODUÇÃO

1.1 Considerações iniciais

A extração de petróleo é uma atividade que desempenha papel crucial na economia mundial, uma vez que, para além da produção de combustíveis, seus derivados são utilizados na fabricação de produtos importantes, como plásticos, detergentes, produtos farmacêuticos, solventes, tintas, além outros inúmeros produtos que impactam a vida moderna (THOMAS et al., 2004). Segundo dados do boletim mensal da produção de petróleo e gás natural (ANP, 2024a), a produção brasileira segue aumentando nos últimos anos e, como pode ser visto na Figura 1, a maior parte da produção vem da região do pré-sal, que envolve poços perfurados em águas profundas. Essa região impõe maior complexidade à execução da perfuração e, por sua vez, demanda custos elevados (NASCIMENTO et al., 2015; PETROBRAS, 2024). As projeções de investimentos para 2024, neste setor, correspondem a U\$ 1,96 bilhão, sendo que desses investimentos, 87% são direcionados à perfuração de novos poços (ANP, 2024b). Desse modo, qualquer avanço para melhorar a eficiência de processos na construção desses poços representa uma redução considerável de custos para o setor.

Figura 1 – Evolução da produção de petróleo e gás (*onshore* e *offshore*) de 2021 a 2023



Fonte: Adaptado de ANP (2024a)

De acordo com Elmgerbi et al. (2021), a etapa de perfuração efetiva (“broca no fundo”) é responsável, aproximadamente, por 20% a 30% do tempo e custo total da construção do poço. Desse modo, uma abordagem tradicionalmente considerada no controle dos custos do empreendimento é buscar a redução do tempo envolvido na atividade de perfuração. Portanto, a taxa de penetração, do inglês *Rate of Penetration* (ROP), é um dos principais indicadores considerados na análise da eficiência da perfuração. Sendo a ROP definida como a velocidade

com que a broca remove o material rochoso e penetra a formação, aumentar a ROP dentro de um limite seguro pode trazer economias significativas.

Mohamud (2020) menciona que, durante as últimas décadas, novas tecnologias de perfuração de poços foram desenvolvidas, o que possibilitou, de modo geral, elevar as taxas de penetração (ROP) nos poços, a exemplo do uso de brocas com material e design mais apropriados. No entanto, Mohamud (2020) observa que esses avanços, exclusivamente, não foram responsáveis por reduzir o tempo não produtivo significativamente durante a perfuração, o que motiva o emprego de estratégias adicionais para atingir o limite técnico de eficiência na perfuração. Essas estratégias geralmente envolvem a busca por maior controle operacional durante a perfuração, de modo que vibrações excessivas ou desgaste de brocas, por exemplo, sejam melhor controlados (GANDELMAN, 2012). Esse controle é atingido com a escolha de parâmetros operacionais ideais.

O principal aspecto de interesse no estudo da ROP é a otimização, pois de fato relaciona a ROP à redução de custo, ao minimizar o tempo de perfuração sem prejuízos de integridade dos equipamentos. Em campo, é possível melhorar a ROP por meio de testes para determinar os melhores parâmetros operacionais aplicados à coluna de perfuração, principalmente rotações por minuto, ou velocidade de rotação, e peso sobre a broca, respectivamente *Rotations per Minute* (RPM) e *Weight on bit* (WOB). Esses testes envolvem a verificação da resposta da ROP, conforme alterações sucessivas do par RPM e WOB, no chamado *drill-off test*. No entanto, o método envolve interrupções na perfuração e consistem apenas na modificação desses dois parâmetros. Muitos estudos de otimização utilizando dados históricos também foram explorados na literatura, nos quais técnicas tradicionais, por meio de expressões analíticas ou relações empíricas, são desenvolvidos para correlacionar os vários parâmetros de influência na perfuração e a ROP (NAJJARPOUR et al., 2021; FERRO et al., 2023; SILVA et al., 2023). A aplicabilidade desses modelos depende de uma série de restrições, como por exemplo o tipo de broca e as propriedades da formação.

Como mencionado em Payette et al. (2017), uma área que oferece oportunidades para melhoria da performance na perfuração é a melhor utilização de dados coletados em tempo real nos poços. Os sistemas de sensoriamento durante a perfuração causaram uma revolução, pois permitiram maior controle, identificação de problemas na perfuração em tempo real e também contribuíram com uma considerável massa de dados para estudos em poços futuros. Como mencionado por Xue (2020), atualmente o papel dos dados aumentou significativamente, e processar esses dados corretamente resulta diretamente em vantagens técnicas e tecnológicas para as empresas da indústria de óleo e gás. Como mencionado em Hossain e Al-Mejed (2015), enquanto a modelagem convencional é feita durante o planejamento do poço para identificar cenários indesejáveis, na modelagem em tempo real os modelos que representam o fenômeno recebem os *inputs* corretos durante a perfuração, permitindo, portanto, monitoramento contínuo e tomada de decisão em tempo real.

A grande quantidade de dados disponíveis recentemente possibilitou a utilização das técnicas de aprendizagem de máquina em vários âmbitos da indústria. Aprendizagem de máquina é, de modo geral, a teoria que envolve encontrar padrões complexos e desconhecidos utilizando dados, e faz parte da área de inteligência artificial. Como mencionado em Noshi e Schubert (2018), técnicas de análise de dados mais tradicionais não são suficientes para extrair informações de forma rápida, precisa e ainda possibilitando visualização clara em conjunto de dados complicados, o que abre espaço para técnicas mais sofisticadas. Um exemplo de aplicação relevante do uso da aprendizagem de máquina na indústria de petróleo e gás pode ser encontrada em Sobrinho et al. (2020) e Aranha et al. (2023), que buscam utilizar a aprendizagem de máquina para detecção de anomalias do poço por meio de dados adquiridos em tempo real (temperatura e pressão de válvulas no sistema de produção do poço). Noshi e Schubert (2018) mencionam que a demanda crescente da indústria motivou o interesse em aplicar aprendizagem de máquina nas áreas de otimização da perfuração e estimativa de parâmetros operacionais.

O controle da perfuração em tempo real demanda respostas imediatas à modificação do ambiente, como alteração de formações rochosas, por exemplo. Esse cenário complexo e dinâmico motivou pesquisas para avaliar a resposta de modelos preditivos para ROP em cenários realistas, que são alimentados com dados durante a perfuração e são capazes de gerar estimativas da ROP. Técnicas tradicionais, como mencionado, que envolvem toda a experiência pregressa com estudos acerca da modelagem da ROP, e a aprendizagem de máquina são estratégias empregadas para realizar essas previsões (BARBOSA et al., 2019; NAJJARPOUR et al., 2021).

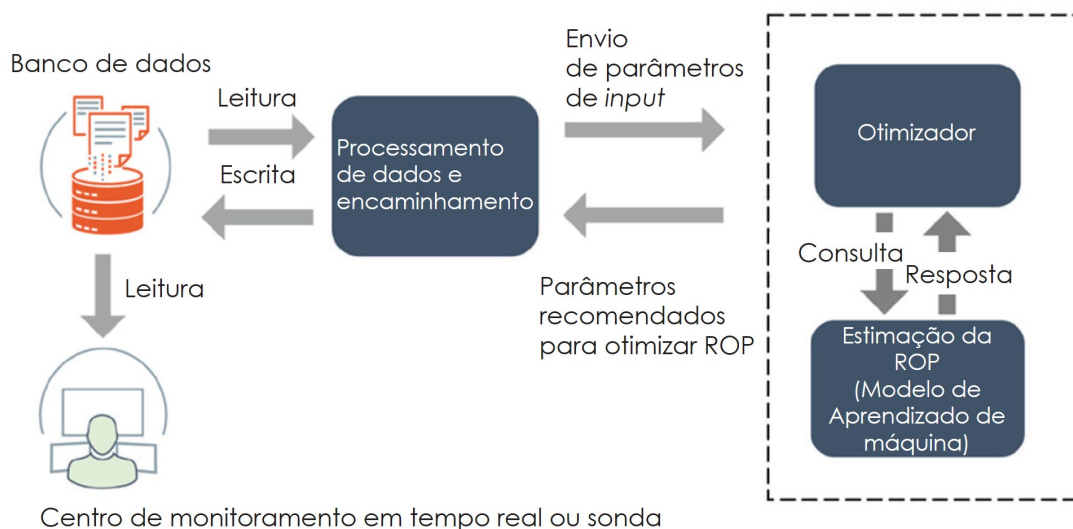
1.2 Motivação

Uma tendência de desenvolvimento para o futuro, na indústria, é cada vez mais a combinação efetiva de algoritmos em softwares de processamento de dados de fundo de poço, junto com análise teórica e interpretação (XUE, 2020). Nesse sentido, como também é mencionado em Contessi et al. (2022), diversos segmentos da indústria vêm abraçando os chamados gêmeos digitais, que são representações virtuais de um ativo e que são alimentados em tempo real com dados para simular e fazer previsões do comportamento desse ativo. Silva et al. (2020) mencionam que o uso de gêmeos digitais na indústria de óleo e gás representa ganhos reais do ponto de vista operacional.

Em Gandelman (2012), é indicado que avanços recentes viabilizaram a incorporação de *softwares* de sistemas de controle, e mesmo com milhares de variáveis e várias restrições, os problemas de otimização em tempo real podem ser executados nesses *softwares* e, consequentemente, empregados em campo. Usar modelos preditivos para ROP viabiliza essa tarefa pois contribui na automatização do processo quando aliado a algoritmos de otimização dentro de *frameworks* preparados para lidar com tratamento dos dados (HEGDE; GRAY, 2018; HÖHN et al., 2022). Como explicado em Hegde e Gray (2018), os modelos são funções objetivas dentro de um problema de otimização, que ao ser solucionado por meio de algoritmos apropriados

os parâmetros operacionais ideais são obtidos. Na Figura 2 é apresentado um esquema com exemplo de aplicação desses modelos.

Figura 2 – Aplicação da aprendizagem de máquina para estimativa da ROP em sistema de otimização



Fonte: Adaptado de Robinson et al. (2022)

Alguns exemplos de *softwares*, com aplicações em campo documentadas, são descritos na literatura, como é o caso de Payette et al. (2017), onde o *software* chamado *Drilling Advisory System* (DAS) é descrito. O DAS é uma aplicação desenvolvida para dar suporte às decisões de perfuração, ao executar *drill-off tests* de modo automatizado por meio de busca de pares WOB e RPM ótimos em superfícies de respostas ajustadas a dados em tempo real. Estratégias de otimização são aplicadas nessas superfícies de respostas que, por sua vez, envolvem a combinação da estimativa da ROP e da energia mecânica específica, incluída, segundo os autores, para evitar a recomendação de pares de WOB e RPM associados a processos de perda energética na perfuração, como é o caso da vibração excessiva. Como descrito em Xue (2020), um outro exemplo de solução é o *eDrilling*, descrito como um sistema que emprega modelagem em tempo real para monitorar e otimizar o processo de perfuração. Nesse sentido, são empregadas estratégias de aprendizagem de máquina para estimar padrões em dados históricos de perfuração, considerando vários parâmetros importantes na perfuração, como WOB, RPM, tipo de broca e vazão do fluido de perfuração. Por fim, um exemplo recente é o *software* GOLD (*Gas Oil Logging While Drilling*), vencedor do prêmio ANP 2023, com a seguinte descrição:

"Algoritmos de Inteligência Artificial (IA) foram desenvolvidos para otimizar parâmetros controláveis dentro de uma faixa operacional segura, reduzindo a Energia Específica Mecânica (MSE) e otimizando a Taxa de Penetração (ROP), recomendando modificações operacionais para ganho de eficiência."(ANP, 2023)

Barbosa et al. (2019) e Najjarpour et al. (2021) realizaram revisão de literatura de grande quantidade de modelos tradicionais e modelos baseados em dados, respectivamente. Na literatura é possível verificar que vários trabalhos recentes focam no uso de modelos de aprendizagem de máquina aplicado ao problema de predição da ROP (SOARES; GRAY, 2019; KOR; ALTUN, 2020; MOAZZENI; KHAMEHCHI, 2021; HÖHN et al., 2022). Nesses trabalhos, os modelos baseados em dados são citados como opções que retornam boa capacidade preditiva em relação aos modelos tradicionais. Em muitos trabalhos recentes os modelos tradicionais são usados como referência em benchmarks junto com os modelos de aprendizagem de máquina ou são adaptados para utilização em tempo real (SOARES; GRAY, 2019; MOAZZENI; KHAMEHCHI, 2021). Essas comparações são importantes para estabelecer possíveis limitações entre diferentes métodos, determinar melhores estratégias para desenvolvimento desses modelos preditivos e determinar informações relevantes para aumentar precisão dessas predições, uma vez que, como mencionado em Gandelman (2012), o processo de perfuração é um processo intrinsecamente transiente, com mudanças constantes das condições operacionais e, por isso, impõe maior complexidade na tarefa de atingir boas predições. Nesse sentido, um importante aspecto relacionado às comparações publicadas na literatura é, como mencionado em (TUNKIEL et al., 2021), que a maioria dos resultados publicados não são reproduzíveis, seja por falta de informações detalhadas dos modelos ou, principalmente, pelos dados utilizados, que muitas vezes possuem tamanho reduzido e são sigilosos.

Desse modo, tendo em vista as aplicações recentes que vêm sendo desenvolvidas para a indústria, é possível observar a relevância de aprimorar modelos preditivos para ROP. Estudos aprofundados acerca desse problema são contribuições relevantes, assim como a disponibilidade de dados de perfuração que sejam representativos, não sigilosos, com quantidade de informações adequadas e úteis para o desenvolvimento de modelos mais precisos.

1.3 Objetivos

O objetivo geral deste trabalho é a investigação, implementação e análise comparativa de técnicas tradicionais e de aprendizagem de máquina aplicadas à modelagem e à predição da ROP. Como objetivos específicos podem ser citados:

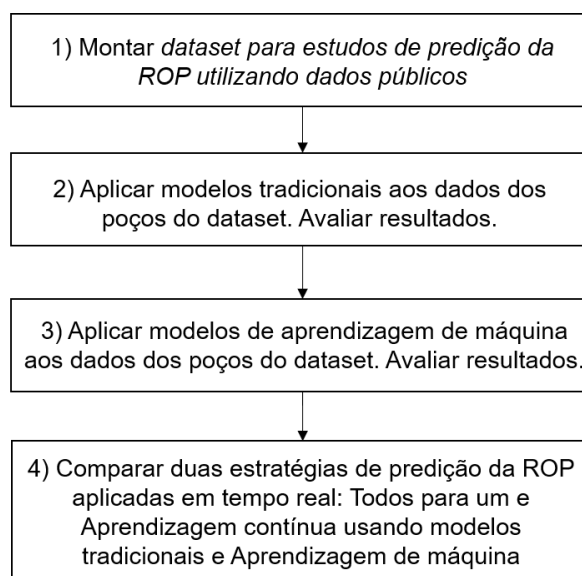
- Definir um *dataset* com dados públicos de perfuração de poços de petróleo para a comparação dos modelos de ROP.
- Realizar análise comparativa de modelos de ROP, com métodos tradicionais e de aprendizagem de máquina, com base nos dados de cada poço do *dataset*.
- Avaliar a influência de parâmetros de perfuração empregados nos modelos de ROP analisados.

- Realizar análise comparativa de estratégias de predição da ROP em tempo real com base na resposta de modelos tradicionais e de aprendizagem de máquina.

1.4 Metodologia

Inicialmente, faz-se uma revisão bibliográfica acerca dos métodos aplicados na modelagem da ROP, bem como uma revisão sobre a teoria de aprendizagem de máquina. Para alcançar todos os objetivos mencionados, as etapas descritas na Figura 3 são aplicadas neste trabalho.

Figura 3 – Etapas gerais da metodologia



Fonte: Autor (2024)

Na primeira etapa, define-se o conjunto de dados. Os dados públicos do campo petrolífero *Volve*, na região do Mar do Norte na Noruega, são utilizados. Esses dados foram tornados públicos pela Equinor (2018). Após a definição de informações relevantes e do tratamento necessário, dados de 7 poços são estruturados para compor o *dataset* empregado neste trabalho.

Na segunda etapa, os métodos tradicionais, definidos com base na revisão bibliográfica feita, são aplicados aos dados de cada um dos 7 poços e os resultados são analisados, onde se verifica a representatividade das informações coletadas e compara-se a capacidade dos modelos em capturar o padrão da ROP, bem como é feita a análise de interpretabilidade desses modelos tradicionais, que, por sua vez, é empregada na construção dos modelos de aprendizagem de máquina.

Da mesma forma, na terceira etapa são aplicados os métodos de aprendizagem de máquina relevantes, conforme revisão bibliográfica, aos dados de cada poço. Avaliando as configurações dos modelos e comparando diferentes parâmetros de entrada nesses modelos.

Por fim, na quarta etapa, empregam-se modelos para predição da ROP em duas estratégias referenciadas na literatura como aplicações à otimização de parâmetros operacionais por meio

de modelos baseado em dados: utilizando os dados disponíveis de poços de correlação (todos para um) e utilizando dados do próprio poço de modo incremental (aprendizado contínuo).

1.5 Estrutura do trabalho

Nesta seção, é apresentada a organização geral do trabalho e a síntese dos assuntos tratados em cada capítulo. Sendo o presente capítulo a introdução, foram apresentados a contextualização do problema tratado, a relevância do tema para indústria de petróleo, os objetivos a serem alcançados com o trabalho e a metodologia empregada.

No Capítulo 2, são apresentados conceitos básicos que envolvem o processo de perfuração, como equipamentos e informações relevantes sobre o sistema de monitoramento da perfuração. Essas informações são úteis para a compreensão das discussões levantadas ao longo do texto.

No Capítulo 3, são abordados os principais aspectos da modelagem da ROP, incluindo estratégias tradicionais, parâmetros de perfuração relevantes e a sua influência na ROP. A formulação de modelos tradicionais é apresentada, assim como revisão de literatura sobre modelos de aprendizado de máquina para a predição da ROP. Por fim, duas estratégias para otimização da perfuração em tempo real, com base em modelos preditivos, são apresentadas.

No Capítulo 4, são introduzidos os conceitos técnicos acerca de modelos de aprendizagem de máquina. É feita revisão dos métodos principais utilizados, com a descrição do processo de treinamento. Critérios para avaliação da performance na predição e a interpretação estatística de modelos simples de regressão também são discutidas.

No Capítulo 5, são apresentados mais detalhes acerca dos dados utilizados. Como Informações sobre os poços de origem, o tratamento empregado e detalhes sobre como esses dados são utilizados, em conjunto com os modelos, para a comparação das estratégias de predição investigadas.

No Capítulo 6, resultados e discussões são apresentados. Os resultados de modelos tradicionais aplicados em cada poço, assim como os modelos de aprendizagem de máquina, são reproduzidos e discutidos. Os resultados das duas estratégias de predição da ROP aplicáveis na otimização em tempo real são comparados.

No Capítulo 7 são apresentadas as conclusões do trabalho e as sugestões de trabalhos futuros.

Por fim, nos apêndices (Apêndice A, Apêndice B, Apêndice C e Apêndice D), são apresentados mais detalhes sobre o conjunto de dados utilizados, como métodos específicos empregados para calcular atributos presentes no conjunto de dados, além de gráficos que dão suporte à interpretação dos dados.

2 PERFURAÇÃO: CONSIDERAÇÕES BÁSICAS

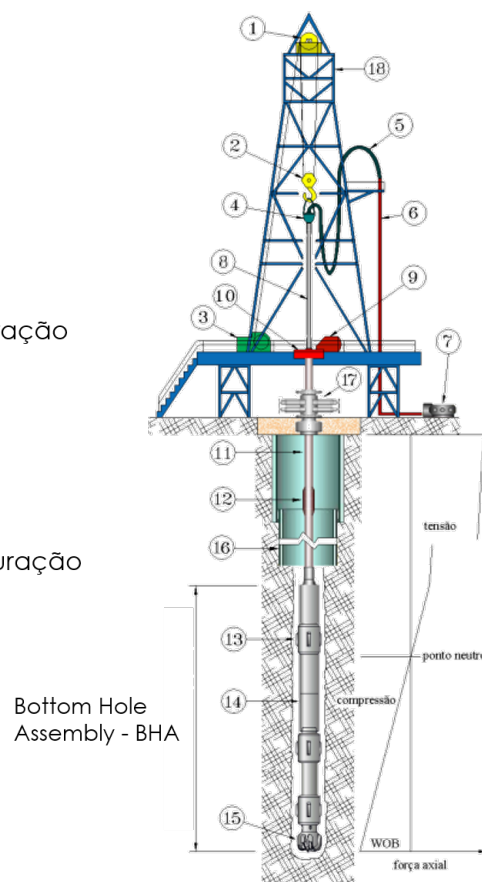
Este capítulo apresenta conceitos básicos acerca dos sistemas de perfuração rotativa de poços de petróleo. Como mencionado em Thomas et al. (2004), a perfuração de um poço de petróleo é realizada por meio de uma sonda. Uma série de sistemas interligados fazem parte dessa sonda, de maneira a possibilitar a penetração de uma broca, acoplada à extremidade de um conjunto de tubos conectados, em formações rochosas por meio de rotação e remoção contínua de fragmentos usando fluido de perfuração. Os sistemas de rotação e monitoramento que compõem a sonda são apresentados em mais detalhes ao longo deste capítulo.

2.1 Sonda de perfuração

As sondas de perfuração rotativa variam em tamanho, ambiente, capacidade de perfuração, além do nível de automação (MITCHELL et al., 2011). A Figura 4 apresenta esquema básico de uma sonda terrestre com indicativo de seus componentes. Esse esquema pode ser estendido para todos os tipos de sonda.

Figura 4 – Esquema de uma sonda de perfuração rotativa

1. Bloco de Coroamento
2. Catarina e Gancho
3. Guincho
4. Swivel (Junta Rotativa)
5. Mangueira do Kelly
6. Tubo
7. Bomba do Fluido de Perfuração
8. Kelly
9. Acionador Rotativo
10. Mesa Rotativa
11. Tubos de Perfuração
12. Conexão do Tubo de Perfuração
13. Estabilizadores
14. Comando de Perfuração
15. Broca
16. Revestimento
17. Preventor de Erupção
18. Torre de Perfuração



Fonte: Adaptado de Goulart (2019)

Na Figura 4, é indicado o chamado *Bottom Hole Assembly* (BHA), que é a região inferior da coluna de perfuração e que agrega os estabilizadores, a broca e os comandos de perfuração (SOUZA, 2011). A seguir, apresenta-se uma breve descrição dos principais equipamentos e suas contribuições ao processo de perfuração de poços:

- **Mesa rotativa:** É a parte da sonda responsável por aplicar a rotação no sistema. As mesas rotativas são geralmente utilizadas em sondas terrestres. No caso de sondas marítimas, utiliza-se o motor de topo, que é responsável pela mesma função.
- **Kelly:** É um elemento de transmissão conectado a mesa rotativa e que transmite a rotação da mesa para a coluna de perfuração, permitindo ainda levantar e descer a coluna enquanto rotaciona.
- **Catarina e Gancho:** São equipamentos de movimentação e elevação de cargas. A catarina é um conjunto de polias, e o gancho conecta a junta rotativa e a catarina (THOMAS et al., 2004; GOULAT, 2019). Como mencionado em Goulat (2019), um dos parâmetros de controle utilizados neste tipo de operação é a carga no gancho.
- **Bomba de fluido de perfuração:** Controla a vazão do fluido de perfuração. O fluido de perfuração é mencionado em Thomas et al. (2004) como sendo misturas complexas de sólidos, líquidos e produtos químicos que são especificados de modo a garantir perfuração rápida e segura.
- **Estabilizadores:** Como mencionado em Goulat (2019), fornece estabilidade para o BHA, auxilia no controle direcional, reduz vibração e centraliza a coluna de perfuração.
- **Tubos de perfuração:** Tubos conectados para formar um elemento estrutural responsável por transmitir a rotação, e também conduzir o fluido de perfuração, até a broca.
- **Comando de perfuração:** Elemento tubular de alta espessura e que fornece rigidez à coluna de perfuração.
- **Brocas:** Segundo Mitchell et al. (2011), a principal ferramenta do processo de perfuração é a broca acoplada ao fim da coluna de perfuração. Além disso, possui relação direta com a eficiência na penetração. Desse modo, a seguir são apresentados mais detalhes sobre este elemento da sonda.

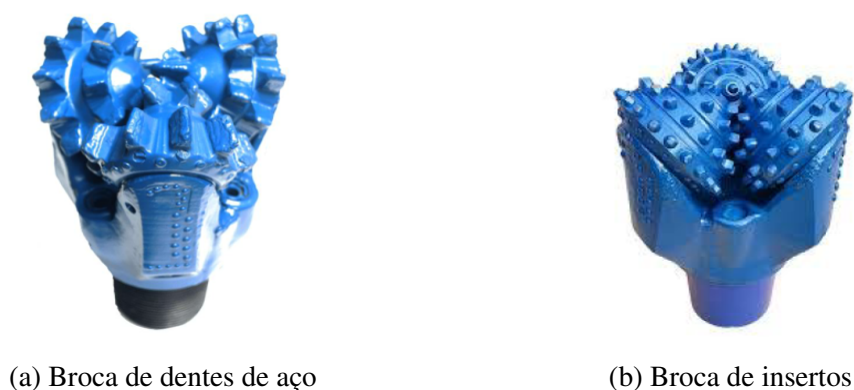
2.2 Característica das brocas

Os sistemas de perfuração rotativos utilizam basicamente dois tipos de brocas: as *Roller-Cone Bits* (RCB), que são as brocas de partes móveis, e as *Fixed Cutter Bits* (FCB), que são as brocas sem partes móveis. Nos dois tipos de brocas é permitido o fluxo de fluido internamente para resfriamento do material durante o processo de perfuração, e também para permitir a limpeza

do poço, uma vez que as brocas possuem orifícios por onde é expelido o jato para carrear o material sólido desagregado. A combinação da ação de indentação (causada pelo WOB aplicado) e a ação de corte (devido à rotação) causam a remoção do material rochoso. A depender do design da broca e de seu tipo, a ação de corte na rocha muda e torna-se mais apropriada para diferentes cenários (ALSENWAR, 2017). A seguir, apresentam-se os tipos principais de brocas:

- **Brocas com partes móveis:** As partes móveis são os cones acoplados ao corpo da broca. O mais comum é a broca três cones (tricônica). Podem ainda ser classificadas em brocas de dentes de aço (*Milled Tooth Bit*) e brocas de insertos (*Tungsten Carbide Insert*). Como mencionado em Thomas et al. (2004), as brocas de dentes de aço possuem a estrutura cortante fresada no próprio cone. As brocas de insertos possuem dentes manufacturados separadamente e unido ao cone, além que o material geralmente é de carbureto de tungstênio (FRANKIEWICZ, 2019). As ações de corte nessas brocas ocorrem com a combinação do efeito de raspagem, lascamento e esmagamento (THOMAS et al., 2004). A Figura 5 apresenta as duas configurações.

Figura 5 – Exemplo de brocas com partes móveis



(a) Broca de dentes de aço

(b) Broca de insertos

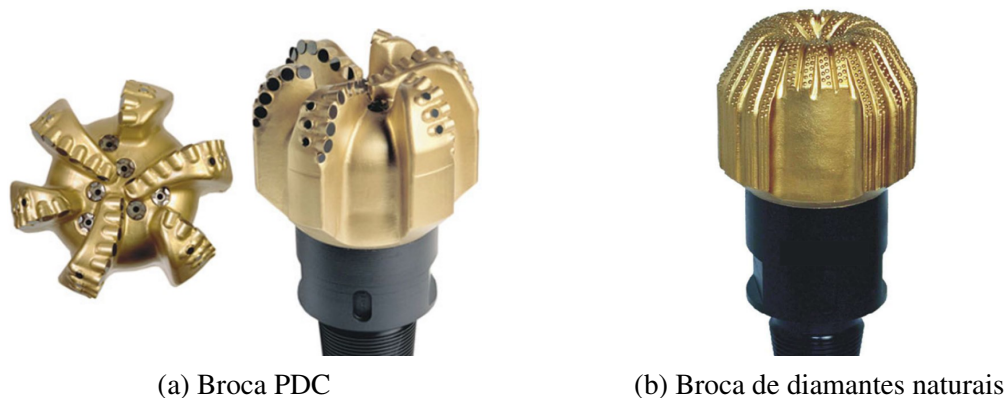
Fonte: Adaptado de Frankiewicz (2019)

São aplicadas a vários tipos de formações. Brocas de insertos são utilizadas em formações de média a alta dureza. Como mencionado em Frankiewicz (2019), para um dado WOB, maior espaço entre os dentes/inserções melhora a ROP, e de modo geral a alteração nessas configurações aumenta o desempenho para diferentes tipos de formações. É indicada para formações mais homogêneas, pois a alteração na capacidade de perfuração devido à formação, impacta a vida útil da broca.

- **Brocas sem partes móveis:** Essas brocas utilizam o efeito de cisalhamento para remoção do material rochoso. Em geral existem dois grupos: brocas PDC (*Polycrystalline Diamond Compact*) feitas com pastilhas de diamante sintético e as brocas impregnadas, com diamantes naturais ou com TSP (*Thermally Stable Polycrystalline*) que formam uma matriz envolvendo o corpo da broca. Como mencionado em Thomas et al. (2004), as brocas PDC

foram introduzidas para aplicação em formações moles com ROP alto, e para atingir maior vida útil. A Figura 6 apresenta os dois tipos de brocas.

Figura 6 – Exemplo de brocas sem partes móveis



Fonte: Adaptado de Ma et al. (2016)

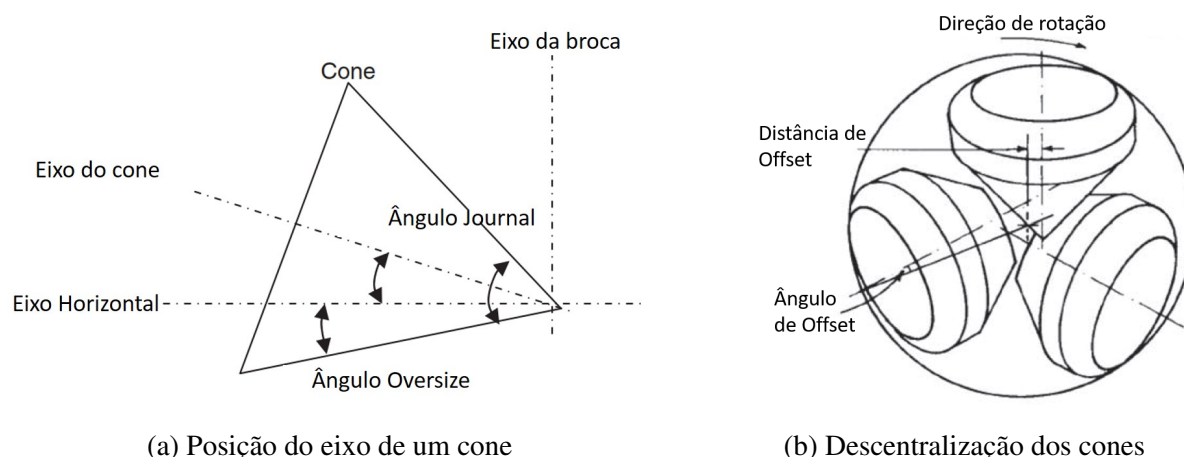
Segundo Mitchell et al. (2011), a ação de corte dessas brocas requer significativamente menor energia do que as brocas RCB, ou seja, menor WOB é requerido, no entanto, para atingir maiores ROP, são necessárias grandes rotações. As brocas sem partes móveis são menos suscetíveis a falhas e tem maior vida útil. Segundo Frankiewicz (2019), as brocas PDC garantem maior controle na perfuração direcional do que as brocas RCB. As brocas impregnadas com diamantes naturais são mais indicadas para formações mais duras do que as PDC, pois suportam temperaturas maiores. Como mencionado em Mitchell et al. (2011), brocas sem partes móveis possuem desempenho muito bom em formações heterogêneas.

Para cada tipo de broca, existem parâmetros relacionados à configuração do modelo que influenciam a performance durante a perfuração, pois possibilita maior eficiência de acordo com mecanismo de corte empregado na broca e a formação a ser perfurada. Em brocas RCB, o mecanismo de corte envolve a raspagem, lascamento e esmagamento (THOMAS et al., 2004; MITCHELL et al., 2011). Para controlar como cada mecanismo se sobrepõe aos demais, alteram-se as características dos dentes/insertos, o chamado ângulo *offset* dos cones e o ângulo *journal* (GUAN et al., 2021; THOMAS et al., 2004). A Figura 7 apresenta esquema de posicionamento dos cones de uma broca RCB, de onde é possível identificar os ângulos mencionados.

Como citado em Mitchell et al. (2011), o ângulo *journal* é o ângulo formado pela reta perpendicular ao eixo da broca e o eixo do cone. O ângulo *offset* é o ângulo que define a excentricidade dos cones na broca. Como mencionado em Mitchell et al. (2011), um ângulo *journal* menor é mais apropriado para formações mais moles e, segundo Frankiewicz (2019), um ângulo *offset* menor é mais apropriado para formações mais duras. Em rochas duras o efeito de esmagamento é mais predominante, já em rochas moles ocorre o efeito de raspagem, desse modo, a modificação dos dois ângulos alteram a geometria da broca favorecendo esses mecanismos. Em relação às propriedades dos dentes ou insertos, Mitchell et al. (2011) mencionam que quanto

mais dura a formação rochosa menor o espaçamento e tamanho dos dentes para evitar o dano desses elementos.

Figura 7 – Representação dos cones de uma broca RCB e seus parâmetros principais



Fonte: Adaptado de Mitchell et al. (2011)

A Tabela 1 apresenta os valores adotados para os ângulos mencionados conforme o uso em diferentes formações segundo Frankiewicz (2019). Além disso, as características dos dentes/insertos, conforme comentado em Mitchell et al. (2011), são apresentadas.

Tabela 1 – Parâmetros importantes nas brocas RCB

Parâmetro	Formação mole	Formação média	Formação dura
Ângulo <i>Offset</i>	4°	3°- 0°	0°
Ângulo <i>Journal</i>	33°	34°-36°	39°
Características dos dentes/insertos	Maior espaçamento e dentes maiores	Espaçamento mediano dentes atarracados	Espaçamento e dentes pequenos

Fonte: Autor (2024)

As brocas do tipo PDC são mais eficientes do que as brocas do tipo RCB para formações moles e medianas. A performance pode ser controlada por parâmetros que influenciam propriedades significativas para o processo de corte, como a eficiência na limpeza, prolongamento da vida útil das lâminas e agressividade do corte. Os parâmetros mais significativos são o número de cortadores nas brocas, e as suas dimensões, o ângulo de ataque (*Backrake angle*) e o ângulo de inclinação lateral (*Siderake angle*). Segundo Carrapatoso (2011), se a quantidade de cortadores na broca é pequena a profundidade de corte aumenta e a ROP é maior. Em relação às dimensões utilizadas em cada cortador, a Tabela 2 apresenta valores apropriados quando condicionado a uma formação específica.

A Figura 8 representa o posicionamento de um cortador de uma broca PDC. Têm-se o ângulo de ataque (α_c), que é o ângulo entre a vertical da formação e a face do cortador, e o

ângulo de inclinação lateral (θ_c), que é a medida da orientação do cortador da esquerda para direita.

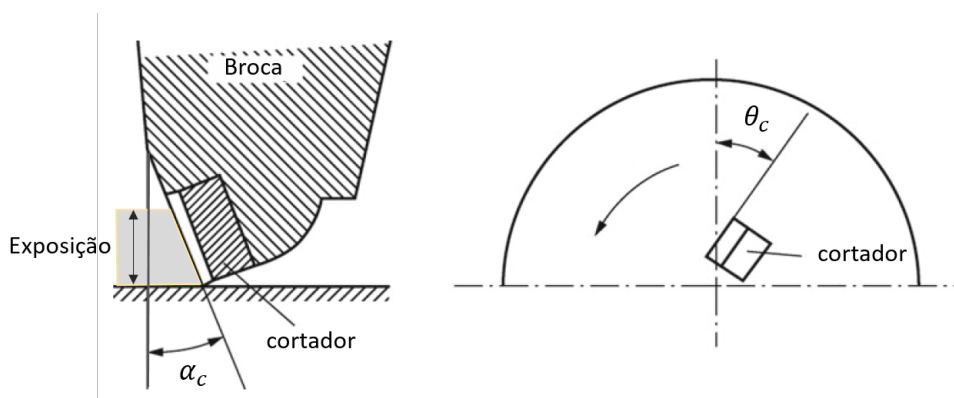
Tabela 2 – Parâmetros de tamanho dos cortadores em brocas PDC

Diâmetro (mm)	Altura (mm)	Formação apropriada
8,00	8,00	Dura
13,44	8,00	Média-Dura (Homogênea)
13,44	13,20	Média-Dura (Heterogênea)
16,00	8,00	Média (Homogênea)
16,00	13,20	Média (Heterogênea)
19,05	8,00	Mole (Homogênea)
19,05	13,20	Mole (Heterogênea)

Fonte: Autor (2024)

Como mencionado em Mitchell et al. (2011), ângulos de ataque (α_c) baixos tornam a broca mais agressiva (maior ROP dado um WOB), mas suscetível a danos por impacto. Carra-patoso (2011) cita que o cascalho possui menor tendência de ficar preso na broca quando α_c é menor, o que contribui na performance da perfuração. Segundo Mitchell et al. (2011), o ângulo de inclinação lateral tem função de produzir estabilidade no corte e também contribui na limpeza do processo de perfuração. A Tabela 3 apresenta o intervalo de valores dos ângulos utilizados nos cortadores da broca.

Figura 8 – Posicionamento dos cortadores em brocas PDC



Fonte: Adaptado de Guan et al. (2021)

Tabela 3 – Parâmetros de orientação dos cortadores em brocas PDC

Parâmetros	Intervalo
Ângulo de ataque	10° - 25°
Ângulo de inclinação lateral	0° - 10°

Fonte: Guan et al. (2021)

2.3 Sistema de monitoramento da perfuração

As sondas de perfuração modernas possuem sistemas de monitoramento que permitem a transmissão e registro simultâneo da maioria dos parâmetros importantes na operação de perfuração, permitindo a detecção e correção de problemas (MITCHELL et al., 2011; SOUZA, 2011). Segundo Mitchell et al. (2011), os parâmetros que não podem ser determinados automaticamente em tempo real são medidos e registrados constantemente (como propriedades do fluido de perfuração). Os principais parâmetros de importância que são obtidos no sistema de monitoramento são apresentados a seguir:

- **Profundidade:** Utiliza a altura do gancho, medidas conhecidas de tamanho do BHA, bem como dos demais tubos da coluna. Somando esses parâmetros e fazendo as devidas compensações, obtém-se a medida da profundidade (TAVARES, 2006).
- **Peso sobre a broca:** Considera-se a carga do gancho e o peso dos componentes da coluna que são sustentados para determinar a carga suspensa a ser transmitida ao fundo do poço. A medição é realizada quando o sistema encontra-se fora do fundo e quando a broca é apoiada no fundo do poço. O controle é feito transmitindo a carga suspensa cuidadosamente para o fundo do poço ou removendo essa carga (CARRAPATOSO, 2011; TAVARES, 2006).
- **RPM:** É um parâmetro fornecido pela própria sonda. Mas pode ser medido em tempo real por sensores instalados na mesa rotativa ou motor de topo para acompanhar o processo de rotação do sistema (TAVARES, 2006).
- **Torque:** O torque na superfície pode ser obtido diretamente na sonda. Alternativamente é possível determinar por meio de medida indireta de equipamentos de medição da corrente elétrica no motor de topo (TAVARES, 2006).
- **Vazão do fluido:** Calculado pelo número de *strokes* por minuto da bomba, medido por sensores de proximidade no pistão, e parâmetros conhecidos: volume da câmara e eficiência (TAVARES, 2006).
- **Vazão de retorno:** Sensores são colocados no conduto para retorno do fluido que circulou dentro do poço. É continuamente monitorada pois pode indicar problemas como *kick* ou perda de circulação (CARRAPATOSO, 2011).
- **Pressão de injeção:** É medida da pressão no tubo bengala (componente do sistema de circulação do fluido) usando sensor de pressão do tipo membrana (TAVARES, 2006). É também conhecido como *Standpipe Pressure* (SPP).
- **Taxa de penetração:** É calculada por meio da medida do avanço na perfuração (acompanhamento constante da profundidade do poço) pela unidade de tempo (geralmente horas) conforme mencionado em Carrapatoso (2011).

- **Peso do fluido:** Com a utilização de sensores, é medido o peso do fluido na entrada e na saída do poço. O peso do fluido é medido pela diferença entre leituras de um sensor do tipo membrana em dois níveis diferentes do fluido (TAVARES, 2006).
- **Carga no gancho:** É a carga sustentada pelo gancho. Pode ser medida por meio de extensômetros (TAVARES, 2006).
- **Propriedade dos fluidos:** As propriedades principais são a densidade, temperatura, viscosidade e conteúdo de sólidos (MITCHELL et al., 2011).

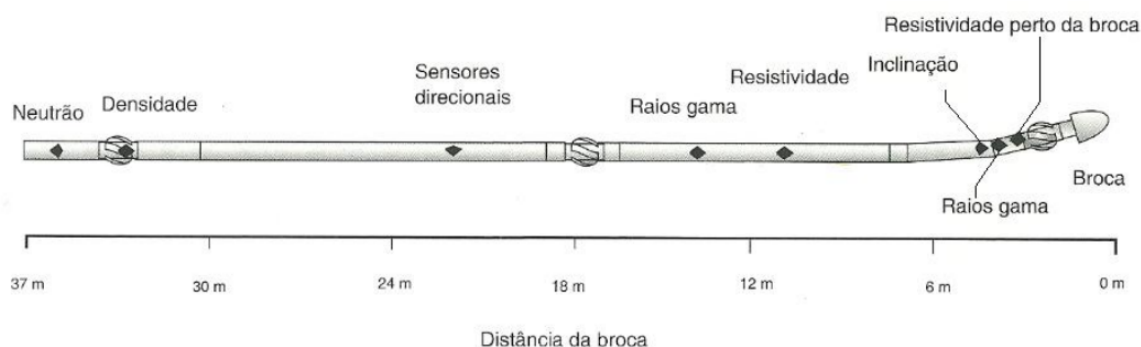
Os parâmetros operacionais mencionados são obtidos pelos sensores instalados na própria sonda, no chamado sistema *mud logging*. Vale ressaltar que o sistema de monitoramento é constituído de vários equipamentos localizados em diferentes posições do sistema de perfuração, de modo que existe aquisição de dados em equipamentos de superfície e também de fundo de poço, que empregam os métodos *Measure While Drilling* (MWD), *Logging While Drilling* (LWD) e *Pressure While Drilling* (PWD). Os dados são processados e agregados em relatórios para registro e análise futuras.

Como mencionado em Tavares (2006), a técnica do *mud logging* não utiliza nenhum sensor dentro do poço e o número de parâmetros medidos pode variar a depender das características da operação de perfuração. Vale ressaltar que o sistema *mud logging* permite caracterizar a formação perfurada (o tipo e determinar litologias) por meio da análise das amostras de material rochoso presente no fluido de perfuração. Além disso, alguns parâmetros importantes da formação, como a pressão de poros, podem ser estimados por correlações teóricas com WOB, ROP e vazão do fluido.

O sistema de monitoramento no fundo do poço consiste em empregar equipamentos acoplados no BHA. Segundo Tavares (2006), o MWD é responsável principalmente pelo fornecimento dos dados direcionais para controle da trajetória (a profundidade vertical verdadeira - TVD, pode ser determinada com essas informações). A transmissão de dados é feita por meio do fluido de perfuração utilizando perturbações na coluna de fluido que geram pulsos, e que, por sua vez, são decodificados em tempo real na superfície. Parâmetros de operação, como o torque, o WOB, RPM e ROP medidos no fundo do poço próximo à broca, são possíveis em certas configurações de BHA. Esses dados são obtidos em tempo real e são considerados mais precisos, no entanto não são informações obtidas comumente (BELASKIE et al., 1993).

O LWD é a técnica que emprega equipamentos de perfilagem acoplados ao BHA para fazer a aquisição de dados utilizando sistema de transmissão semelhante ao MWD. É considerado uma extensão do MWD e executa medições equivalentes à perfilagem a cabo tradicional. Com o LWD pode-se obter informações relacionadas às formações perfuradas. Um parametro importante determinado com os dados do LWD é o gradiente de pressão de poros (EPP). A Figura 9 apresenta um esquema com a posição dos equipamentos de LWD em um BHA.

Figura 9 – Disposição de equipamentos do LWD no BHA



Fonte: Adaptado de Rocha et al. (2011 apud SOUZA, 2011)

A seguir, são apresentados os diferentes tipos de perfilagem extraídas com o LWD e as informações obtidas de cada um:

- **Raios gama (*Gamma ray*):** Detecta a radioatividade da formação geológica, oferecendo portanto informações para identificação de litologia.
- **Sônico:** Utiliza equipamento para emissão de ondas mecânicas. A leitura da velocidade dessas ondas fornece os dados de tempo de trânsito (o Delta-T, que mede o tempo que uma onda leva para percorrer a distância de 1 *ft* na formação rochosa, dada em $\mu s/ft$). Esses dados permitem auxiliar na caracterização da litologia e efetivamente é utilizado para determinar a porosidade das formações. Utilizando esses dados é possível estimar a resistência das rochas utilizando correlações teóricas da literatura (CHEN et al., 2018; SOARES; GRAY, 2019).
- **Densidade e neutrão:** Segundo Thomas et al. (2004), são duas técnicas de perfilagem que medem raios gama após excitação com diferentes fontes. São úteis para medição de porosidade, identificação de zonas de gás.
- **Resistividade:** Mede a resistência à condução de corrente elétrica. Segundo Souza (2011), auxilia a identificar tipos de fluidos contido nos poros.

O PWD é, segundo Tavares (2006), o sistema que mede a pressão do fundo do poço. Emprega uma série de sensores alojados no BHA, e que utiliza o processo de transmissão de dados por pulsos do fluido como mencionado anteriormente. O dado principal fornecido por esse sistema é o chamado *Equivalent Circulation Density* (ECD), que é determinado por estimativa obtida da pressão anular e representa a pressão dinâmica de circulação no fundo do poço (TAVARES, 2006). Esse tipo de medição só é disponível em certas configurações de BHA, mas uma vez disponível é possível obter esse parâmetro em tempo real (SOARES; GRAY, 2019; TAVARES, 2006).

3 MODELAGEM DA ROP: PREDIÇÃO E OTIMIZAÇÃO

Muitos estudos foram desenvolvidos ao longo dos anos tendo como objetivo a formalização de um modelo preciso que representasse a relação de fatores que influenciam a ROP (BARBOSA et al., 2019; NAJJARPOUR et al., 2021). Para Hegde e Gray (2018), o esforço para desenvolver modelos da ROP consiste na busca para representar o objeto, ou a resposta do modelo (ROP), como uma função de parâmetros controláveis e incontroláveis, e, segundo Barbosa et al. (2019), o objetivo final dos modelos preditivos da ROP é a otimização da perfuração. Soares e Gray (2019) afirmam que a precisão desses modelos influencia diretamente no sucesso da otimização, uma vez que modelos preditivos precisos são essenciais como funções objetivo dentro do problema de otimização.

Este capítulo introduz os principais direcionamentos tomados na literatura acerca da modelagem da ROP com base nos parâmetros operacionais envolvidos na perfuração. São apresentados os fatores que influenciam a ROP, bem como os principais modelos analíticos tradicionais e adaptações dos modelos para tempo real, que envolve a relação entre parâmetros medidos em tempo real. Estratégias mais recentes aplicando os modelos baseados em dados são pontuadas. Em seguida é introduzido o tópico de otimização dos modelos da ROP.

3.1 Fatores que influenciam a ROP

Os primeiros modelos de previsão da ROP foram desenvolvidos durante a década de 1960. No entanto, em trabalhos anteriores como Sikes (1936) e Bielstein e Cannon (1950), discussões sobre quais variáveis afetam a ROP já são apresentadas. Essas discussões culminaram no desenvolvimento posterior dos modelos analíticos da ROP com capacidade preditiva. Hossain e Islam (2018) mencionam que os fatores que afetam a ROP ainda não são totalmente conhecidos. Porém, pesquisas recentes demonstraram a influência notável de vários parâmetros de perfuração na ROP. Em Fear (1999), esses parâmetros são investigados e, por sua vez, listados conforme a classificação da Tabela 4.

Os fatores controláveis na Tabela 4 são aqueles modificados na superfície de imediato durante a perfuração, ou que pelo menos possibilitam certo controle dos operadores. Em contrapartida, os fatores incontroláveis, também conhecidos como fatores ambientais, estão relacionados a limitações tecnológicas de acesso ou fatores que dependem das características da própria formação geológica. Na Tabela 4, informações dos fluidos e pressão no fundo do poço podem ser consideradas como fatores incontroláveis por conta que há requisitos necessários para atingir o objetivo (reservatório de petróleo) e que dependem das propriedades da formação perfurada. A densidade do fluido, por exemplo, tem papel fundamental para evitar entrada de fluidos da formação no poço e é definida em projeto. Do mesmo modo, o tamanho da broca deve obedecer requisitos do projeto do poço, como produção de petróleo e passagem de equipamentos,

e, portanto, não é modificado para atender a eficiência da perfuração. As relações entre as variáveis na Tabela 4 configuram o complexo sistema de interação broca/rocha durante a perfuração, sendo essa relação altamente não linear entre alguns desses fatores (SOARES; GRAY, 2019).

Tabela 4 – Fatores que afetam a ROP

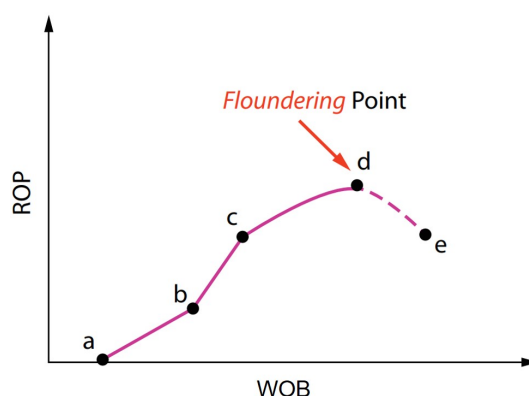
Fatores controláveis	Fatores incontroláveis
Design da broca	Tipo da formação
WOB	Propriedades da formação
RPM	Tipo do fluido
Vazão do fluido	Densidade do fluido
Potência hidráulica da broca	Outras propriedades do fluido
Arranjo dos ofícios da broca	<i>Overbalance</i> e/ou Pressão no fundo do poço
Geometria do motor/turbina	Tamanho da broca

Fonte: Adaptado de Fear (1999)

As influências dos principais fatores na ROP são apresentadas em mais detalhes a seguir:

- **WOB (Peso sobre a broca):** É considerado um dos fatores mais significativos, pois o processo de perfuração demanda que a compressão exercida pela coluna de perfuração sobre a rocha supere a resistência à compressão da formação, caso contrário não há penetração significativa, principalmente em brocas RCB. É o primeiro parâmetro a ser alterado para aumentar a ROP (GANDELMAN, 2012). Maurer (1962) investiga as relações entre os parâmetros no caso de brocas RCB, e observa que em condições apropriadas de limpeza do poço, ou seja, com remoção ideal de sólidos durante a perfuração, a ROP seria diretamente proporcional ao quadrado de WOB e que, em condições ideais de limpeza do poço, a relação entre esses parâmetros é linear até certo ponto, independentemente da formação. Em brocas PDC, o cisalhamento é o mecanismo de corte principal, desse modo, é produzido maior ROP com um WOB menor (GEOFFROY et al., 1998). Em Bourgoyne et al. (1986) a relação descrita por Maurer (1962) é atualizada conforme a Figura 10.

Figura 10 – Resposta típica da ROP ao aumento do WOB

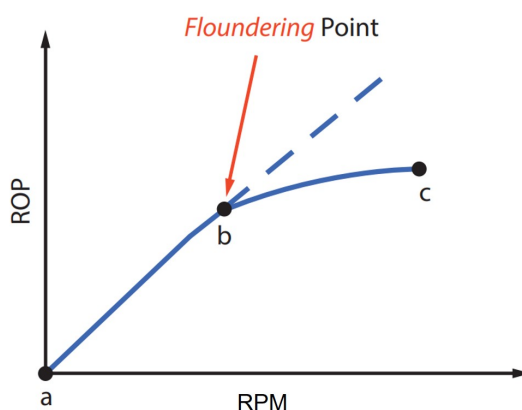


Fonte: Adaptado de Hossain e Al-Mejed (2015)

A Figura 10 é obtida ao manter constante os demais parâmetros e variar apenas o WOB. A partir do ponto *a* é vencida a resistência inicial da rocha e o processo de perfuração é iniciado. No segmento *ab* há relação linear, com um crescimento gradual da ROP com o aumento de WOB. Em seguida, no segmento *bc*, há uma região onde também existe relação linear entre a ROP e o WOB, porém com maior eficiência no processo de perfuração (maior inclinação da reta). Para além do ponto *c*, os aumentos no WOB ocasionam pouco aumento na ROP até atingir o *Floundering Point*, mencionado por Maurer (1962). Os valores de WOB excessivos para além do ponto *d* resultam em diminuição da ROP. Os motivos principais são, primeiramente, a grande quantidade de sólidos gerados, que dificulta a rápida limpeza por meio do fluido de perfuração, e outro motivo seria de caráter estrutural, uma vez que a coluna de perfuração começa a apresentar vibrações excessivas e flambagem. Problemas de limpeza durante a perfuração resultam em retrabalho da broca, porque as partículas se sedimentam e energia precisa ser gasta para perfurar em material rochoso removido previamente. Vibração excessiva e flambagem dissipam a energia que deveria ser empregada totalmente na perfuração da formação rochosa (GANDELMAN, 2012).

- **RPM:** Segundo Maurer (1962), a relação entre RPM e ROP é linear quando a condição de limpeza do poço é ideal, porém em caso de baixa remoção dos sólidos não há comportamento linear e a ROP diminui com o aumento do RPM. Esse fenômeno ocorre para valores de RPM muito elevados. A Figura 11 apresenta gráfico semelhante ao obtido na Figura 10, no qual a investigação feita por Bourgoyne et al. (1986) revela comportamento semelhante ao relatado em Maurer (1962).

Figura 11 – Resposta típica da ROP ao aumento do RPM



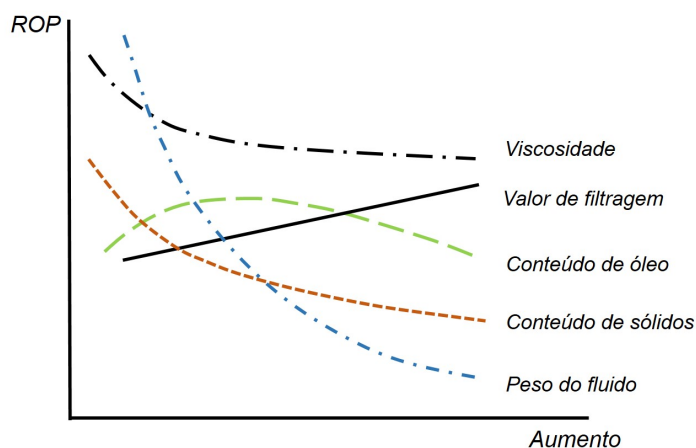
Fonte: Adaptado de Hossain e Al-Mejed (2015)

Na Figura 11, o ponto *b* também é conhecido como *Floundering Point* por conta de marcar a transição entre valores de RPM associados à maior eficiência da perfuração e valores elevados que contribuem para a perda de eficiência do processo. A diminuição da eficiência na perfuração está relacionada à baixa remoção dos sólidos e também por problemas de vibração excessiva, de forma análoga ao efeito causado por WOB elevado. Outro efeito

latente ocasionado por rotações elevadas é o desgaste da broca, que a longo prazo afeta a ROP.

- **Características da formação:** Esse fator engloba a resistência da rocha, o seu limite elástico, a porosidade e permeabilidade, a abrasividade e a sua dureza, dentre outros parâmetros relacionados (HOSSAIN; ISLAM, 2018). Dos fatores não controláveis é certamente o que mais afeta a ROP. Segundo Hossain e Islam (2018), as propriedades da formação mais importantes que afetam a ROP são o seu limite elástico e a sua resistência limite. Como mencionado em Gandelman (2012), diferentes litologias requerem valores diferentes de energia empregada no processo de perfuração, e, com isso, rochas mais resistentes induzem naturalmente uma ROP menor, pois demandam mais energia do sistema para ultrapassar a resistência da rocha perfurada. Segundo Bourgoyne et al. (1986), uma maior permeabilidade permite o fluido de perfuração circular mais facilmente na rocha, dessa forma atua modificando o diferencial de pressão agindo na broca. Além disso, rochas com abrasividade promovem maior desgaste da broca, que, por sua vez, afeta a ROP.
- **Propriedade do fluido de perfuração:** Parâmetros reológicos do fluido de perfuração também afetam a ROP de forma significativa, principalmente a densidade do fluido (NAJJARPOUR et al., 2021). Segundo Hossain e Islam (2018), as propriedades do fluido que afetam a ROP são: densidade, propriedades reológicas, características de filtração, conteúdo de resíduos sólidos e sua distribuição granulométrica e a composição química do fluido. Segundo Najjarpour et al. (2021), existe uma relação inversamente proporcional entre a ROP e alguns dos fatores mencionados como a viscosidade, a quantidade de sólidos no fluido e também a densidade (Figura 12). Um aumento da capacidade de filtração do fluido, no entanto, causa um aumento da ROP. Um dos fatores controláveis importantes é a vazão do fluido de perfuração, e esse parâmetro se relaciona de forma mais complexa com a ROP, do que o RPM e o WOB. Como mencionado em Soares e Gray (2019), o aumento da vazão leva ao aumento da perda de carga no anular e, conseqüentemente, levam a maior pressão de fundo do poço. Ao mesmo tempo, o aumento da vazão promove maior remoção de resíduos sólidos da perfuração, o que diminui a pressão de fundo do poço. Dessa forma, a vazão provoca efeitos concorrentes na ROP.
- **Overbalance:** É o diferencial entre a pressão exercida no poço e a pressão dos fluidos da formação (pressão de poros da formação), provocado por gases ou líquidos (GANDELMAN, 2012). O diferencial positivo é o *overbalance*, e consiste na pressão de fundo de poço aplicada ser maior que a pressão de poros. É de interesse na perfuração pois contribui na maior segurança em uma perfuração convencional, uma vez que evita influxo de fluidos da formação. Quando o diferencial positivo resulta em valores extremos, induz os resíduos sólidos a permanecerem nos poros e dessa forma dificulta o processo de perfuração (baixa ROP). A Figura 13 ilustra a relação entre o aumento do diferencial de pressão e o logaritmo

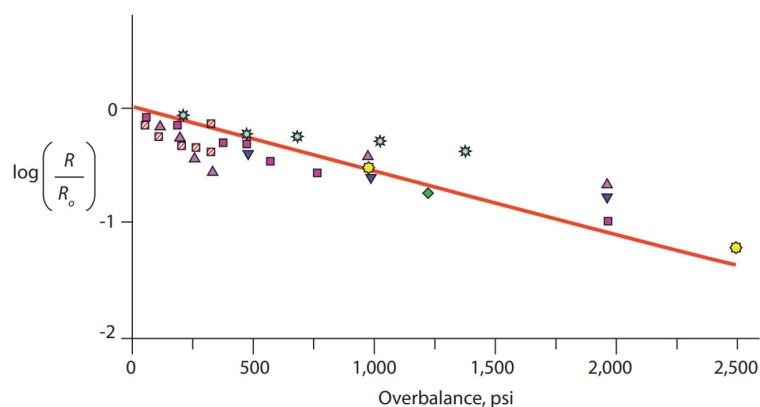
Figura 12 – Relação entre a ROP e diferentes propriedades dos fluidos



Fonte: Adaptado de Najjarpour et al. (2021)

da razão entre a ROP e o valor da ROP quando o *overbalance* é zero. Nessa imagem, são incluídos resultados de diferentes brocas.

Figura 13 – Efeito do aumento do diferencial de pressão na ROP



Fonte: Adaptado de Hossain e Al-Mejed (2015)

As propriedades do fluido também contribuem na alteração do diferencial de pressão. Como mencionado em Gamal et al. (2022), *overbalance* é um fator crítico na interação do fluido com a rocha e carece de estudos mais aprofundados em relação aos impactos na permeabilidade e porosidade da rocha. De forma geral, determina-se o *overbalance* a ser aplicado no poço durante o projeto, e isso é feito levando em consideração critérios de segurança, do tipo de fluido aplicado e também das propriedades da formação, além disso, a forma de ajuste durante a operação é alterando o peso do fluido (GAMAL et al., 2022; GANDELMAN, 2012; ALAWAD, 2016). Perfis de resistividade e sônico da região a ser perfurada são informações úteis nessa tomada de decisão no projeto, como mencionado

em Hossain e Al-Mejed (2015), com esses perfis é possível identificar zonas de pressão anormal nas formações perfuradas.

- **Design da broca e desgaste:** Primeiramente, a escolha do tipo da broca tem um impacto significativo na ROP. Brocas RCB não são aplicadas em formações de maior dureza, diferente das PDC, por exemplo, e danos prematuros gerados por escolha incorreta do tipo de broca afeta a ROP. O tamanho da broca é definido em projeto levando em conta parâmetros como a formação e sua espessura, por exemplo. Um peso aplicado em broca de diâmetro menor eleva a tensão aplicada na rocha, dessa forma, em fases que demandam uma broca menor, há tendência de manter um ROP maior se a formação for a mesma (GANDELMAN, 2012). Em relação ao fator hidráulico, como mencionado em Hossain e Islam (2018), a ação do jato expelido da broca é de promover melhor limpeza de fundo do poço e também da própria broca. A velocidade apropriada desse jato também influencia a ROP. O tamanho dos orifícios da broca influencia na velocidade do jato, assim como as propriedades do fluido. Em Bourgoyne et al. (1986) é mencionado que, naturalmente, as brocas tendem a perfurarem lentamente com o progresso da perfuração devido ao desgaste, e o desgaste, por sua vez, depende do tipo da broca (RCB ou PDC) e da formação. A redução da ROP não é significativa até ocorrer desgaste acentuado e quebra dos dentes da broca.

3.2 Modelos tradicionais para ROP

Em Graham e Muench (1959), relações rudimentares entre a ROP e variáveis como WOB e RPM já integravam um plano de otimização de custo envolvido na perfuração (FEAR, 1999; ISLAM; HOSSAIN, 2021). Os primeiros modelos da ROP relacionavam apenas RPM e WOB para obter a ROP por meio de relações simples e aplicadas em condições específicas de perfuração. No entanto, apenas em trabalhos posteriores como Cunningham (1960), Maurer (1962), Galle e Woods (1963), Bingham (1964) e Bourgoyne e Young (1974) modelos analíticos com maior capacidade preditiva começaram a ser propostos (SOARES; GRAY, 2019; ISLAM; HOSSAIN, 2021).

Inicialmente os modelos foram desenvolvidos considerando brocas RCB. A partir do final da década de 1970 surgiram os primeiros modelos adaptados para brocas do tipo PDC (NAJJARPOUR et al., 2021). O modelo apresentado em Bingham (1964) é um exemplo de modelo da ROP para broca RCB, e que serviu de referência em diversos estudos na literatura (NAJJARPOUR et al., 2021). Esse modelo é dado por

$$ROP = K \cdot \left[\frac{WOB}{D_b} \right]^a \cdot RPM \quad (3.1)$$

em que D_b é o diâmetro da broca, K e a são constantes relacionadas a formação rochosa e à

própria broca, respectivamente. O modelo de Bourgoyne e Young (1974) (BYM), no entanto, é conhecido como o mais compreensivo e minucioso dos modelos tradicionais da ROP. Esse modelo foi atualizado pelos autores em Bourgoyne et al. (1986) e foi originalmente desenvolvido para brocas RCB. O modelo foi proposto originalmente aliado a um processo de regressão linear múltipla, a fim de determinar os coeficientes (a_1 até a_8) com dados históricos de perfuração disponíveis. Dessa forma, a ROP do poço atual é determinada utilizando os coeficientes ajustados com esses dados históricos. A expressão do modelo é dada por

$$ROP = f_1 \cdot f_2 \cdot f_3 \cdot f_4 \cdot f_5 \cdot f_6 \cdot f_7 \cdot f_8 \quad (3.2)$$

em que a ROP é definida como o produto de 8 subfunções. As subfunções f são detalhadas nas seguintes expressões:

$$f_1 = e^{2.303 \cdot a_1} \quad (3.3)$$

$$f_2 = e^{2.303 \cdot a_2 \cdot ((TVD)_N - TVD)} \quad (3.4)$$

$$f_3 = e^{2.303 \cdot a_3 \cdot TVD^{0.69} \cdot (PPG - (PPG)_N)} \quad (3.5)$$

$$f_4 = e^{2.303 \cdot a_4 \cdot TVD \cdot (PPG - ECD)} \quad (3.6)$$

$$f_5 = \left[\frac{\left(\frac{WOB}{D_b} \right) - \left(\frac{WOB}{D_b} \right)_t}{\left(\frac{WOB}{D_b} \right)_N - \left(\frac{WOB}{D_b} \right)_t} \right]^{a_5} \quad (3.7)$$

$$f_6 = \left(\frac{RPM}{(RPM)_N} \right)^{a_6} \quad (3.8)$$

$$f_7 = e^{-a_7 \cdot h} \quad (3.9)$$

$$f_8 = \left(\frac{F_j}{(F_j)_N} \right)^{a_8} \quad (3.10)$$

Neste modelo, além dos coeficientes que são obtidos por regressão linear múltipla e os parâmetros intrínsecos, normalizações obtidas por meio de estudos com dados de campo, feito pelos autores, são incluídos no modelo. A Tabela 5 expõe os parâmetros intrínsecos do modelo e na Tabela 6 os valores de normalização propostos em Bourgoyne et al. (1986) são listados.

Tabela 5 – Parâmetros intrínsecos do modelo

Parâmetros	Descrição
RPM	Rotações por minuto medido na superfície
WOB	Peso sobre a broca medido na superfície
PPG	Valor atual do gradiente de pressão de poros
ECD	Valor atual da densidade equivalente dinâmica
TVD	Profundidade vertical verdadeira
D_b	Diâmetro externo da broca
h	Fração desgastada do dente da broca
F_j	Força hidráulica de impacto do jato aplicado pela broca

Fonte: Bourgoyne et al. (1986)

Tabela 6 – Valores dos parâmetros de normalização

	Valores	Unidade
$(TVD)_N$	10000,00	<i>ft</i>
$(PPG)_N$	9,00	<i>lb/gal</i>
$\left(\frac{WOB}{D_b}\right)_N$	4,00	<i>kpi</i>
$(F_j)_N$	1000	<i>lbf</i>
$(RPM)_N$	100	<i>rev/min</i>

Fonte: Bourgoyne et al. (1986)

Cada subfunção f na Equação 3.2 é uma função que define efeitos específicos gerados pelos parâmetros que afetam a ROP. A seguir, é apresentada cada função e as suas premissas.

f_1 - Representa o fator relacionado à resistência das formações, o tipo da broca utilizada e o efeito dos resíduos sólidos da perfuração (MITCHELL et al., 2011; NAJJARPOUR et al., 2021). O parâmetro a_1 é o coeficiente que captura essas relações e é conhecido como parâmetro de resistência da formação. Segundo Islam e Hossain (2021), não há explicação física para a forma exponencial, apenas sentido matemático de acentuação dos efeitos. De fato, esta subfunção agrega os demais efeitos relevantes na ROP e que não são explicitamente englobados nas demais subfunções, em que o principal efeito seria a resistência da formação.

f_2 - Engloba os efeitos de compactação da formação na ROP. A expressão exponencial decresce à medida que aumenta a profundidade e é normalizada para um valor de profundidade de até 10000 ft (Tabela 6). Como mencionado em Islam e Hossain (2021), essa suposição é uma simplificação visto que a compactação está relacionada ao modo de deposição, que, por sua vez, varia extensivamente. O parâmetro a_2 é o expoente da tendência de compactação normal.

- f_3 - Representa os efeitos relacionados à pressão de poros na ROP. Como mencionado em Eren e Ozbayoglu (2010), essa expressão também leva em conta, em certa medida, a característica de compactação da formação. Segundo Bourgoyne et al. (1986), f_3 está modelando o efeito de subcompactação presente em formações com pressão de poros excepcionais. O parâmetro a_3 é o expoente de sob compactação.
- f_4 - Inclui os efeitos do diferencial de pressão que é causado pela relação entre a pressão da formação e aquela causada pela densidade do fluido de perfuração (*overbalance*). O parâmetro a_4 é conhecido como expoente do diferencial de pressão. Assim como a f_3 inclui a medida do gradiente de pressão de poros, que, por sua vez, não é medido diretamente e necessita de modelo específico para a estimativa desta grandeza, conforme descrito em mais detalhes no Apêndice A.
- f_5 - Representa a influência do tamanho da broca e do WOB em conjunto. Como mencionado em Eren e Ozbayoglu (2010), na função f_5 o fator $\left(\frac{WOB}{D_b}\right)_t$ é o limite a ser ultrapassado para dar início ao processo de perfuração. Esse limite é assumido como sendo zero em formações pouco resistentes, e de modo geral pode ser diretamente obtido com os *drill-off tests*. a_5 é o expoente do peso sobre a broca (ISLAM; HOSSAIN, 2021).
- f_6 - É a função que inclui a influência do RPM no modelo. Como mencionado anteriormente, após um aumento de RPM que gera um aumento da ROP, existe um ponto (*Floundering Point*) onde esse comportamento deixa de existir e a ROP decai, essa função atribui ao modelo esse comportamento. O parâmetro a_6 é conhecido como expoente da velocidade de rotação. Como mencionado por Islam e Hossain (2021), o efeito de vibração desempenha um papel na interação entre o RPM e a ROP, mas não é diretamente considerada nessa equação.
- f_7 - Essa função inclui as considerações acerca do estado da broca, ou seja, o nível de desgaste dos dentes. Esse fenômeno em si depende também da abrasividade da formação perfurada, WOB, RPM, geometria dos dentes e a eficiência na remoção dos cascalhos no poço (ISLAM; HOSSAIN, 2021). O parâmetro a_7 é chamado de expoente de desgaste da broca. A subfunção f_7 relaciona a ROP com a variável h , que é a medida da altura remanecente dos dentes. Essa altura só é acessada no intervalo final de uso da broca. No entanto, trabalhos como Soares e Gray (2019) utilizam uma estratégia de interpolação linear para obter os valores de h nas demais profundidades (Apêndice B).
- f_8 - Representa os efeitos da força hidráulica de impacto do fluido de perfuração. Esse fator depende do tipo de fluido usado e de características da broca utilizada. A influência na ROP diz respeito principalmente às condições de limpeza do poço, ou seja, da remoção de sólidos da perfuração que notadamente tem efeitos indiretos na penetração da broca. O parâmetro a_8 é o expoente hidráulico. A expressão para F_j e demais parâmetros calculados com base nas brocas utilizadas são apresentados no Apêndice B.

Segundo Islam e Hossain (2021), esse modelo provou ser fundamental em trazer a tecnologia de perfuração para a era da informação, uma vez que foi utilizado em conjunto com dados de diferentes tecnologias de monitoramento em tempo real. Além disso, este modelo serviu como base para o desenvolvimento de outros modelos para ROP. Um exemplo é o modelo proposto por Reza e Alcocer (1986), que utiliza análise dimensional, aplicando o teorema de *Buckingham* π , para determinar os fatores independentes que influenciam a ROP em brocas RCB, resultando em um modelo de regressão não linear com 4 variáveis que relacionam WOB, RPM, viscosidade do fluido de perfuração, resistência à compressão da rocha, diferencial de pressão, vazão do fluido de perfuração e diâmetro da broca. No entanto, segundo Najjarpour et al. (2021), o modelo BYM é o mais comumente utilizado em estudos relacionados à ROP usando métodos tradicionais, e um dos motivos, certamente, é por correlacionar de forma explícita os principais parâmetros operacionais, e também efeitos notoriamente relevantes por meio de relações matemáticas simples, como é o caso do desgaste da broca.

Adaptações do BYM foram propostas por alguns autores com o objetivo de tornar o modelo mais representativo para outros tipos de brocas, principalmente fazendo alterações em subfunções ou normalizações (OSGOUEI, 2007; DARWESH et al., 2020; NAJJARPOUR et al., 2021). Devido ao fato de haver diferentes mecanismos de corte entre brocas do tipo PDC e RCB, a relação entre WOB, RPM e torque na broca provoca diferentes efeitos na ROP. Uma modificação que é considerada em Osgouei (2007) e Darwesh et al. (2020) é a adaptação da subfunção f_5 , que corresponde ao efeito do WOB na ROP, e que é descrita por

$$f_5 = \left[\frac{C_r \cdot \left(\frac{WOB}{D_b} \right) - 0,942 \cdot \Delta P_b \cdot \left(\frac{D_b - 1}{D_b} \right)}{800} \right]^{a_5} \quad (3.11)$$

que propõe a inclusão de uma relação empírica que corresponde ao efeito do chamado *Pump-off Force*, que é determinada por

$$P_{of} = 0,942 \cdot \Delta P_b \cdot \left(\frac{D_b - 1}{D_b} \right) \quad (3.12)$$

e essa expressão, por sua vez, depende diretamente da perda de pressão na broca (ΔP_b), que é apresentada no Apêndice B.

O *Pump-off Force* surge como uma força de reação que reduz a carga aplicada na broca. Ramsey (2019) menciona que esse efeito é predominante em brocas de diamantes naturais e também em brocas do tipo PDC, porém em menor grau. Esse efeito surge pela pequena distância entre a face da broca e a rocha perfurada. Desse modo, a expressão modificada tenta modelar esse efeito e, portanto, torna a subfunção f_5 possível de melhor adaptar o comportamento do WOB com a ROP para outros tipos de brocas que não somente as brocas de partes móveis. Darwesh et al. (2020) utilizam a Equação 3.11 com a sugestão de redução de 0,942 para 1/3. Na

Equação 3.11, ainda é incluído o fator C_r , que representa a fração da carga que de fato é aplicada na broca quando alargadores são utilizados. Algumas sugestões recentes também foram feitas em Darwesh et al. (2020) para alterar normalizações no modelo, como indicadas na Tabela 7.

Tabela 7 – Sugestões para alteração nas normalizações

Subfunções	Valor original	Valor modificado
f_2	10000	12000
f_4	9,0	12,0
f_6	100	100 para RCB e 160 para PDC

Fonte: Autor (2024)

Considerando a inclusão de informações de trajetória na formulação da subfunção f_5 , Eren e Ozbayoglu (2010) incluem a seguinte expressão simplificada dentro da subfunção f_5 do modelo Bourgoyne et al. (1986) original, que emprega a informação da inclinação medida (γ):

$$WOB_a = \frac{WOB}{\cos(\gamma)} \quad (3.13)$$

que pode ser incluída na Equação 3.11 relacionada ao efeito do WOB.

Outra modificação proposta é a adaptação da subfunção f_8 para a inclusão da chamada potência hidráulica por polegada quadrada (HSI), em vez do F_j , que é a força hidráulica de impacto (mais detalhes no Apêndice B). Desse modo, a subfunção f_8 dentro do modelo foi adaptada para considerar a seguinte expressão:

$$f_8 = \left(\frac{\Delta P_b / \left(1714 \cdot \frac{\pi \cdot D_b}{4} \right)}{2} \right) \quad (3.14)$$

Como indicado em Nasser (2012), considerações sobre o torque deveriam ser adicionadas ao modelo de Bourgoyne et al. (1986), além disso, relações diferentes existem entre torque e ROP para os diferentes tipos de brocas. Como mencionado em Hossain e Al-Mejed (2015), brocas do tipo PDC geralmente requerem torque elevado quando comparado com brocas do tipo RCB. A inclusão do torque pode ser relevante em certas situações onde seu efeito sobre a ROP seja significativo, como por exemplo em cenários com mudanças significativas na litologia (KAKOLI et al., 2018). Uma modificação feita em Mohamud (2020), para um cenário com brocas do tipo PDC, é a inclusão do torque como subfunção f_6 no modelo de Bourgoyne et al. (1986). Essa modificação deve ser avaliada mediante casos específicos, considerando também a qualidade e representatividade das medições adotadas.

Outra modificação do modelo BYM é feita por Soares e Gray (2019). Como mencionado por esses autores, o modelo BYM representa os principais aspectos da perfuração, no entanto, nem todos os parâmetros são facilmente obtidos em tempo real, sendo, portanto, estimados. Um exemplo é o desgaste e o gradiente de poro pressão. Segundo o autor, profundidade, WOB, RPM

e vazão do fluido são parâmetros sempre disponíveis em tempo real. Medidas como a ECD e a estimativa da resistência da rocha podem ser disponibilizadas em tempo real a depender dos sensores utilizados e possivelmente de correlações empíricas representativas. Nesse contexto, Soares e Gray (2019) propuseram modificações à formulação do modelo BYM a fim de torná-la mais apropriada para utilização em tempo real. O modelo (BYMSoares) corresponde à seguinte expressão:

$$ROP = a_1 \cdot TVD^{a_2} \cdot WOB^{a_5} \cdot RPM^{a_6} \cdot q^{a_8} \quad (3.15)$$

no qual TVD é a profundidade vertical verdadeira e q é a vazão do fluido de perfuração. A expressão é mais simples que o BYM e elimina a utilização de qualquer normalização.

3.2.1 Modelos baseados na energia mecânica específica

O modelo de energia mecânica específica, do inglês *Mechanical Specific Energy* (MSE), que é aplicado na análise da perfuração foi desenvolvido inicialmente por Teale (1965). O conceito consiste no trabalho realizado para escavar uma unidade de volume da rocha. Como argumentado em Teale (1965), o trabalho realizado na perfuração é a soma do trabalho induzido pelo WOB aplicado e o trabalho gerado pela torção T exercida na rotação da broca. O MSE é amplamente conhecido na indústria como medida de eficiência do processo de perfuração. Uma das aplicações dessa estratégia é na predição e otimização da ROP, em que o MSE é utilizado para garantir que o processo não resulte em vibrações excessivas e desgaste prematuro da broca (alterações no balanço energético). A razão entre o trabalho exercido na perfuração e o volume removido de rocha por unidade de tempo é definida por

$$\frac{Trabalho}{Volume} = \frac{WOB \cdot ROP + 60 \cdot 2\pi \cdot RPM \cdot T}{A_b \cdot ROP} \quad (3.16)$$

e, portanto, a expressão clássica do MSE, conforme apresentado em Teale (1965), corresponde à seguinte expressão:

$$MSE = \frac{WOB}{A_b} + \frac{120\pi \cdot RPM \cdot T}{A_b \cdot ROP} \quad (3.17)$$

na qual os parâmetros A_b e T correspondem à área da superfície da broca e ao torque aplicado na broca, respectivamente. Como mencionado em Chen et al. (2018), o torque na broca é o principal parâmetro no modelo, e apesar de ser diretamente medido em testes experimentais, as medições em campo não são obtidas facilmente, pois grande parte dos dados obtidos em campo são medidas dos parâmetros na superfície. Dessa forma, a Equação 3.17 foi atualizada por Pessier e Fear (1992) para a seguinte formulação:

$$MSE = WOB \cdot \left(\frac{1}{A_b} + \frac{13.33 \cdot \mu_b \cdot RPM}{D_b \cdot ROP} \right) \quad (3.18)$$

na qual é utilizada a relação estimada para o torque na broca em função do WOB, que é dada por

$$T_b = \frac{\mu_b \cdot D_b \cdot WOB}{36} \quad (3.19)$$

em que μ_b é o coeficiente de atrito da broca.

Chen et al. (2018) mencionam que os parâmetros deste modelo são fáceis de obter com medidas na superfície, e por isso este modelo tem bastante uso na indústria. No entanto, quando utiliza-se o WOB que é medido na superfície, e não o WOB_b (WOB real que atua na broca), os valores do torque na broca (Equação 3.19) não são precisos e, conseqüentemente, o MSE não é preciso. Na verdade, em perfuração direcional e horizontal, há maiores diferenças no WOB da superfície e o WOB_b , por conta das perdas por atrito causado pelo contato da coluna de perfuração ao longo do poço. Os valores de MSE calculados com o WOB apresentam erros consideráveis.

Contribuições feitas posteriormente levaram à formalização de um modelo para aplicação em perfuração direcional. Hammoutene (2012) e Amadi e Iyalla (2012) incluíram o conceito de eficiência mecânica da broca no modelo por meio do fator E_m de minoração, que é incluído para distinguir a eficiência na transmissão de energia para diferentes tipos de broca e diferentes formações. Em Chen et al. (2014) é proposta melhoria no modelo por meio da inserção da expressão que relaciona o WOB da superfície e o WOB na broca, que é dada por

$$WOB_b = WOB \cdot e^{-\mu_s \gamma_b} \quad (3.20)$$

e adiciona essa expressão para o torque na broca conforme a relação a seguir:

$$T_b = \frac{\mu_b \cdot WOB_b \cdot D_b}{36} = \frac{\mu_b \cdot WOB \cdot e^{-\mu_s \gamma_b} \cdot D_b}{36} \quad (3.21)$$

Dessa forma, a Equação 3.18 é atualizada para a expressão a seguir:

$$MSE = E_m \cdot WOB \cdot e^{-\mu_s \gamma_b} \cdot \left(\frac{1}{A_b} + \frac{13.33 \cdot \mu_b \cdot RPM}{D_b \cdot ROP} \right) \quad (3.22)$$

em que os coeficientes μ_s e γ_b nesta equação correspondem ao coeficiente de atrito da coluna e a inclinação da coluna, respectivamente.

A relação entre a ROP e a MSE, portanto, é dada pela seguinte expressão:

$$ROP = \frac{13.33 \cdot \mu_b \cdot RPM}{D_b \cdot \left(\frac{MSE}{E_m \cdot WOB \cdot e^{-\mu_s \cdot \gamma_b}} - \frac{1}{A_b} \right)} \quad (3.23)$$

Para utilização dessa expressão, a MSE é relacionada a resistência da formação na condição de eficiência na perfuração, quando a energia é empregada totalmente no processo de perfuração. Na literatura essa relação é feita com a resistência não confinada, do inglês *Unconfined Compressive Strength* (UCS), como indicado em Belayneh (2019), ou com a resistência confinada, do inglês *Confined Compressive Strength* (CCS), conforme mencionado em Chen et al. (2018). Desse modo, na Equação 3.23 a MSE é substituída pela medida de resistência da formação, que, por sua vez, pode ser determinada por correlações empíricas usando medições obtidas com perfilagem (como mencionado no Apêndice C).

3.3 Modelos de aprendizado de máquina para ROP

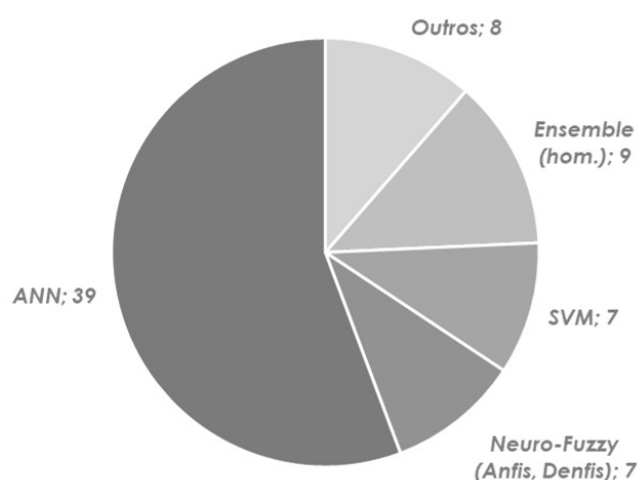
Como explicado em Abu-Mostafa et al. (2012), as técnicas de aprendizagem são utilizadas em casos em que a descrição analítica do fenômeno é difícil ou mesmo impossível, porém há disponibilidade de dados a partir dos quais os padrões existentes podem ser capturados utilizando algoritmos de aprendizagem. Dessa forma, o problema de encontrar modelos que representem a complexa relação entre os parâmetros operacionais da perfuração e a ROP enquadra-se nessa estratégia. Além da relação entre esses parâmetros ser desconhecida, a evolução tecnológica dos sistemas de perfuração possibilitou a coleta de grande quantidade de dados que podem ser analisadas usando as técnicas de aprendizagem de máquina.

Um conceito importante dentro da teoria dos modelos baseados em dados é o balanço viés e variância. A relação fortemente não linear entre os vários fatores que influenciam a ROP, quando capturada por meio dos modelos tradicionais, resultam em uma relação analítica de alto viés. O viés dos modelos é definido por James et al. (2014) como o erro que é introduzido por aproximar um problema real complexo por um modelo mais simples. Soares e Gray (2019) argumentam que modelos analíticos assumem uma relação matemática, como a de uma função potência, por exemplo, entre algumas variáveis mensuráveis, porém o processo de perfuração é muito mais complexo do que essas relações. A variância diz respeito à variabilidade do modelo aos dados utilizados para treino. Como indicado por Soares e Gray (2019), os modelos analíticos possuem baixa variância, no entanto, para gerar modelos com capacidade ideal de generalização (aprender os padrões nos dados usados para treino e gerar previsões), é necessário um balanço entre essas duas medidas. A teoria que envolve os métodos de aprendizagem de máquina oferecem técnicas que permitem garantir esse balanço e, conseqüentemente, gerar uma previsão mais precisas.

A capacidade preditiva dos modelos de aprendizagem de máquina para previsão da ROP superam as obtidas pelos métodos tradicionais segundo vários autores (SOARES; GRAY, 2019;

BARBOSA et al., 2019). Os primeiros a empregar técnica de aprendizagem de máquina para o problema da ROP foram Bilgesu et al. (1997), que empregaram redes neurais para predição da ROP. Como mencionado em Höhn et al. (2022), após mais de uma década sem desenvolvimentos expressivos, o uso de técnicas de aprendizagem de máquina em outras indústrias inspirou esforços de pesquisa para aplicar novas técnicas na predição e otimização da ROP. Em Barbosa et al. (2019), é feita a revisão de literatura com apresentação de trabalhos que aplicaram vários métodos de aprendizado de máquina, como pode ser identificado na Figura 14, que apresenta uma distribuição de métodos de aprendizagem de máquina conforme a literatura.

Figura 14 – Distribuição de métodos de aprendizagem de máquina aplicados ao problema de predição da ROP



Fonte: Adaptado de Barbosa et al. (2019)

Na Figura 14, é possível perceber que o modelo *Artificial Neural Networks* (ANN) foi o mais aplicado ao problema da ROP, no entanto, o *Support Vectors Machine* (SVM) e os métodos *Ensembles* são estratégias mais recentes que apresentam capacidade preditiva superior segundo alguns autores (SOARES; GRAY, 2019; BARBOSA et al., 2019; HÖHN et al., 2022). O *Ensemble*, como mostrado na Figura 14, foi bastante aplicado em vários trabalhos da literatura e consiste em uma família de métodos, onde o RF é o mais aplicado no problema da ROP. Segundo a literatura, o *Random Forest* (RF) permite gerar resultados precisos, e também com uma facilidade de interpretação do modelo maior que de outros métodos (HEGDE; GRAY, 2018; SOARES; GRAY, 2019; BARBOSA et al., 2019; HÖHN et al., 2022).

Como mencionado em Barbosa et al. (2019), existe a tendência dos modelos baseados em dados para predição da ROP serem construídos levando em conta menos preditores (*inputs* do modelo) quando se comparam trabalhos mais recentes com os primeiros estudos para predição da ROP com métodos de aprendizagem de máquina. Como abordado ainda em Barbosa et al. (2019), modelos com até 8 preditores compreendem a grande maioria dos estudos recentes, sendo os preditores mais comuns WOB, RPM, Profundidade e vazão do fluido.

Algumas das considerações que devem ser feitas na aplicação dos modelos de aprendi-

zagem de máquina para o problema de predição da ROP, segundo a literatura, são indicadas a seguir de forma objetiva:

- **Tratamento dos dados:** Como os modelos que usam aprendizagem de máquina precisam de dados, estes devem ser tratados, se necessário, pois dados poluídos podem inviabilizar a aplicação dos modelos. A utilização de dados representativos é crucial para o processo de aprendizagem dos modelos.
- **Capacidade preditiva geral do modelos:** O interesse principal é que o modelo seja capaz de gerar valores da ROP precisos (valor das métricas de erro menor possível). Existem considerações adicionais relacionadas à precisão da resposta quando há poucos dados disponíveis ou se há melhoria da capacidade preditiva quanto mais dados são fornecidos aos modelos.
- **Interpretabilidade dos modelos:** É importante por fornecer meios para compreender o fenômeno e a relação entre suas variáveis por meio da análise do modelo.
- **Inputs do modelo:** É importante determinar quais os preditores e também a quantidade necessária. Geralmente busca-se incluir no modelo apenas os preditores mais relevantes, tornando o modelo menos complexo. Nesse contexto, utilizar toda a experiência pregressa com o desenvolvimento dos modelos tradicionais, como os mencionados neste trabalho, é de grande auxílio na seleção das variáveis significativas como *input* desses modelos (AHMED et al., 2018). Considerações sobre a disponibilidade desses dados são de grande relevância para viabilizar, por exemplo, a aplicação do modelo em tempo real. Como mencionado em Soares e Gray (2019), a performance desses modelos pode ser melhorada considerando *inputs* que não são frequentemente analisadas como o *Gamma ray* e *Delta-T* Compressional (vide Capítulo 2), se disponíveis em tempo real.

3.4 Otimização aplicada a modelo baseado em dados

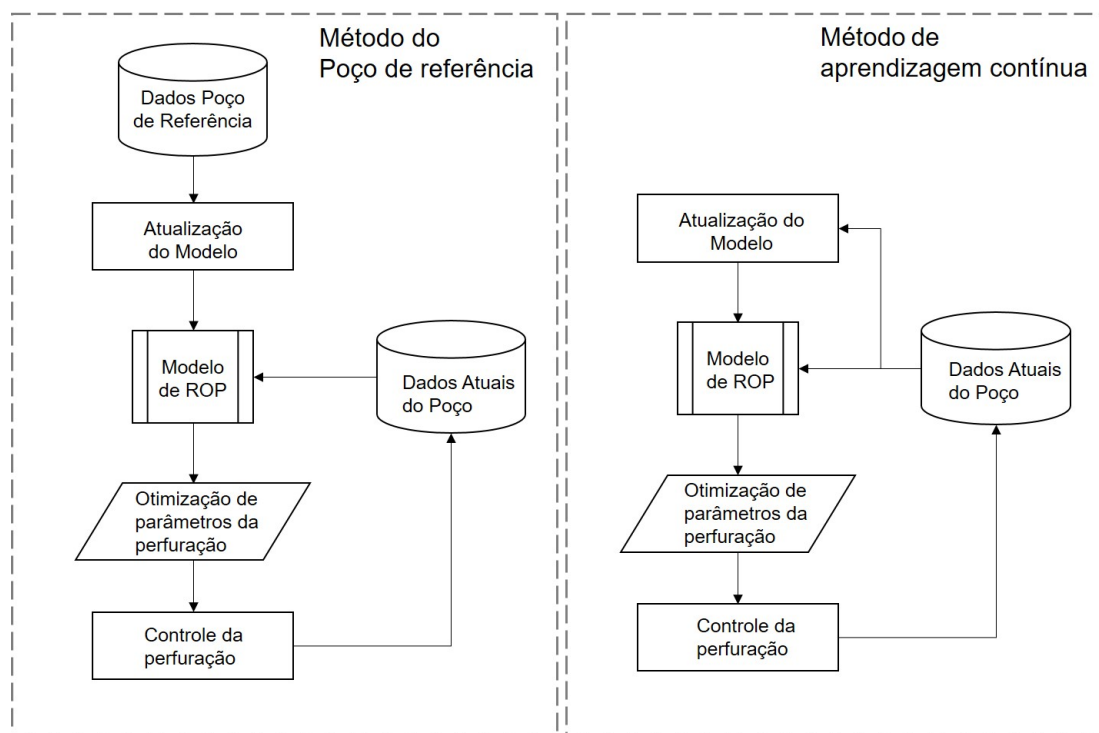
A otimização no processo de perfuração engloba vários aspectos. Durante as etapas de planejamento, por exemplo, novas configurações e equipamentos podem ser ajustados na composição da coluna de perfuração de forma a atingir um melhor comportamento dinâmico para maximizar a ROP. Estudos também podem ser desenvolvidos na seleção de brocas para otimizar a eficiência na perfuração em diferentes tipos de formações. No entanto, como citado em Soares et al. (2016), fatores como os mencionados não podem ser alterados em tempo real para permitir o ajuste e mitigação de problemas na perfuração que podem ocorrer. A necessidade de maior controle operacional para evitar comportamentos inesperados durante alterações na litologia e problemas de limpeza, por exemplo, aliado ao advento de instrumentos de aquisição de dados durante a perfuração, levou a indústria de óleo e gás a considerar a otimização de parâmetros controláveis em conjunto com os estudos realizados nas etapas de planejamento

e projeto (ISLAM; HOSSAIN, 2021; GANDELMAN, 2012). Como apresentado em Islam e Hossain (2021), essas linhas de estudo começaram a se desenvolver na década de 1980.

Uma abordagem utilizada na indústria para otimizar os parâmetros operacionais WOB e RPM durante a perfuração é o *drill-off test*. Esse método consiste em um teste de tentativa e erro para determinar valores de RPM e de WOB ótimos e que resultem em uma boa eficiência durante a perfuração. A ideia é testar combinações de RPM e WOB usando estratégias para variar apenas um desses dois fatores no processo deixando o outro constante. É realizado principalmente ao iniciar a perfuração em uma nova litologia e busca estimar o *Floudering Point*. Duas metodologias podem ser executadas: o teste ativo e o passivo. O teste ativo é para situação de perfuração rápida e o teste passivo é aplicado em condições de perfuração mais lenta, conforme são descrito em Mitchell et al. (2011) e Vegar et al. (2016). Apesar de ser uma prática que aumenta a eficiência, as decisões e controle são fortemente dependentes da experiência e intuição dos operadores (HEGDE et al., 2018). Outro aspecto é que não leva em conta a variação dos outros parâmetros no processo, como a vazão do fluido de perfuração. Como mencionado em Barbosa (2019), para implementar técnicas de controle em tempo real, modelos mais sofisticados são necessários, e os modelos preditivos da ROP usando aprendizagem de máquina são uma opção.

Tunkiel et al. (2021) mencionam que a otimização em tempo real dos modelos da ROP pode ser feita de múltiplas formas. Na Figura 15 duas estratégias principais são apresentadas: *a)* Método do poço de referência e *b)* Método de aprendizagem contínua. O método do poço de referência é uma estratégia que pode ser utilizada quando existem dados de um poço semelhante ao que será perfurado. O modelo é treinado com os dados desses poços de referência e aplicado ao novo poço. Dessa forma, desenvolvem-se modelos que aprendem os padrões presentes nos processos de perfurações anteriores da mesma região para serem aplicados no poço atual. É de fácil implementação em relação à outra opção, no entanto demanda a existência de dados de um poço perfurado na região, que podem ou não estar disponíveis. O modelo treinado com os dados de um poço de referência pode ser otimizado durante a perfuração para gerar os parâmetros ótimos de RPM e WOB, por exemplo, que devem ser mantidos para acelerar o processo de perfuração. Observa-se que, dentro dessa estratégia, modelos preditivos podem ser treinados para cada formação também (GANDELMAN, 2012).

Figura 15 – Duas abordagens de modelos preditivos que podem ser aplicadas na otimização em tempo real



Fonte: Adaptado de Tunkiel et al. (2021)

A segunda estratégia é a de aprendizagem contínua, na qual se aplicam métodos de separação em intervalos da perfuração e modelos são treinados continuamente ao longo desses intervalos. Em um próximo intervalo, esses modelos são otimizados para obter os parâmetros de perfuração desejados. O modelo pode ser atualizado com mais dados à medida que a perfuração ocorre (conforme representado na atualização do modelo ilustrada na Figura 15). Como citado anteriormente, Soares e Gray (2019) investigam essa estratégia e indicam que esse processo seria ideal para o controle em tempo real. As vantagens principais são que os modelos são treinados com dados do próprio poço, além disso, também não é necessário ter perfurado outros poços na região. No entanto, algumas limitações dessa abordagem podem ser mencionadas, como é o caso que a capacidade preditiva pode não ser adequada tendo em vista a menor quantidade de dados e também a capacidade de melhorar ou não as previsões quando mais dados são incluídos no modelo durante a perfuração. Essas possíveis limitações tornam essa estratégia mais difícil de implementar na prática de modo que seja possível assegurando boas previsões, garantindo o sucesso da otimização, ao longo da perfuração.

4 APRENDIZAGEM DE MÁQUINA

A área de aprendizado de máquina é um campo dentro da inteligência artificial cujo objetivo é encontrar padrões com base em dados (IZBICKI; SANTOS, 2020). Modelos de aprendizagem de máquina determinam esses padrões ocultos utilizando algoritmos sedimentados na teoria de aprendizagem estatística. Nesse contexto, aprender é descobrir o padrão em um conjunto de dados analisados a partir do qual é possível, em seguida, generalizar para dados não analisados. Encontrar esse padrão pode ser encarado como estimar uma função alvo que é desconhecida, e que não há um meio claro de determiná-la diretamente por meio de relações matemáticas. Além disso, na aprendizagem de máquina procura-se fazer o mínimo de suposições ou restrições possíveis sobre as relações que existem no conjunto de dados.

Essa área desenvolveu-se de forma interdisciplinar, sendo os principais pilares a estatística, a otimização e a ciência da computação. Como mencionado em Abu-Mostafa et al. (2012), a premissa de aprender com os dados é ampla o suficiente para motivar o surgimento de paradigmas que tratam de diferentes premissas e situações. A Figura 16 apresenta os paradigmas do aprendizado de máquina.

Figura 16 – Os paradigmas principais da aprendizagem de máquina



Fonte: Autor (2024)

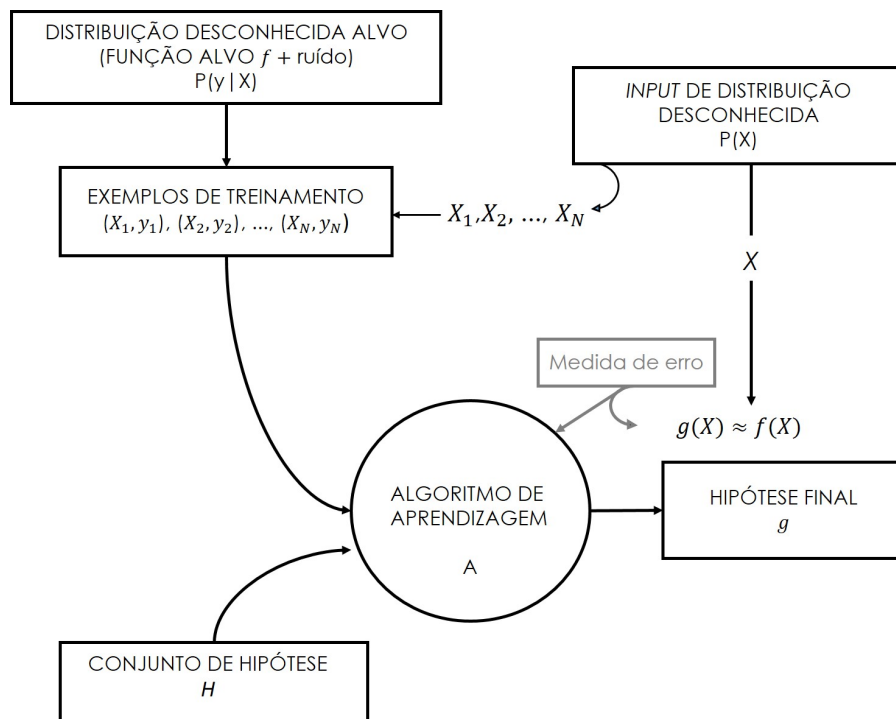
Pode-se definir o aprendizado supervisionado como aquele no qual são fornecidos os chamados dados rotulados, ou seja, os dados possuem exemplos explícitos de como o *output* do modelo de aprendizagem de máquina deveria ser, dado os *inputs* desse modelo (ABU-MOSTAFA et al., 2012). No aprendizado não supervisionado não há dados rotulados, e os algoritmos são utilizados para extrair um nível maior de representação e compreensão dos

padrões presentes nos dados, e assim entender como estão estruturados. No aprendizado por reforço, um sistema preparado para assimilar informações de um ambiente é definido, e o que é fornecido ao modelo é um *feedback* externo no qual o sistema desenvolvido reage de acordo com um sistema de recompensa e penalidades (GÉRON, 2021). Neste trabalho, o foco é o aprendizado supervisionado, que é discutido em mais detalhes a seguir.

4.1 Aprendizado supervisionado

O paradigma de aprendizado de máquina mais utilizado é o supervisionado (ABU-MOSTAFA et al., 2012; GÉRON, 2021; MURPHY, 2022). Formalmente, a tarefa executada é determinar uma função f que mapeia *inputs* $X = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$, conhecidos como atributos ou preditores, para um *output* Y , conhecido como variável resposta. Do problema supervisionado surgem duas abordagens: classificação e regressão. Quando a variável resposta é um valor quantitativo, o problema é de regressão, e se o valor é categórico, temos um problema de classificação (IZBICKI; SANTOS, 2020). A Figura 17 apresenta o problema abordado no aprendizado supervisionado de forma esquematizada.

Figura 17 – O problema geral do aprendizado supervisionado



Fonte: Adaptado de Abu-Mostafa et al. (2012)

Na Figura 17, procura-se determinar uma função alvo f desconhecida. Como nas aplicações práticas não se têm dados gerados por uma função determinística, mas sim dados gerados por uma distribuição probabilística, a relação $Y = f(X)$ é substituída por $P(Y|X)$. Nesse sentido, a variável resposta Y é uma variável aleatória. Uma forma alternativa de analisar é considerar a existência de uma função determinística f associada a um ruído com certa distribuição.

No aprendizado supervisionado, a estimativa da função alvo é feita utilizando os chamados dados de treinamento, que são efetivamente utilizados pelos algoritmos de aprendizagem para estimar a função alvo f e, desse modo, esses são os dados históricos que contêm o padrão a ser aprendido. O termo treinamento é o equivalente de processo de estimação dentro da teoria de aprendizagem de máquina (IZBICKI; SANTOS, 2020). A Figura 17 apresenta os dados de treinamento como parte de um conjunto de *input*, que, por sua vez, possui uma certa distribuição. Como mencionado em Abu-Mostafa et al. (2012), os dados de *input* precisam ter uma distribuição, do contrário não há o que ser aprendido. Além disso, parte desses dados deve ser utilizados para verificar se a função estimada g é próxima de f utilizando as medidas de erro. Esses dados são chamados de dados do conjunto de teste e as métricas de erro ajudam a determinar a capacidade preditiva do modelo. A função g deve aproximar f e possuir capacidade de generalizar essa estimativa para outras amostras dessa distribuição. Para viabilidade desse processo o conjunto de dados deve ser particionado, e dados distintos utilizados no conjunto de treinamento e de teste.

Na Figura 17, a parte principal é formada pelo conjunto de hipótese e o algoritmo de aprendizagem. Dentro de um conjunto de hipótese, é retirada a melhor função g que estima f . A busca dentro desse conjunto é feita pelo algoritmo de aprendizagem. Sempre que um modelo de aprendizagem de máquina é definido, tem-se um conjunto de hipóteses e o algoritmo de aprendizagem (ABU-MOSTAFA et al., 2012). Nesse contexto, o conjunto de hipótese pode ser todas as possibilidades de pesos em uma arquitetura de rede neural, e nesse caso o algoritmo de aprendizagem seria a retropropagação aplicada, que consiste em um processo de otimização desses pesos, como explicado em mais detalhes ainda neste capítulo. Nesse processo é necessário ter ainda uma métrica de desempenho (medida de erro) para possibilitar a escolha da melhor função g no conjunto de hipóteses. Isso pode ser feito utilizando a definição de funções de perda, que em problemas de regressão geralmente envolve a perda absoluta (Norma L_1) ou a quadrática (Norma L_2) (IZBICKI; SANTOS, 2020). Considerando essas duas normas, as métricas de desempenho mais comumente usadas no processo de treinamento são dadas por

$$C(\mathbf{X}, \mathbf{Y}) = \frac{1}{N} \cdot \sum_{i=1}^N L_1(g(X_i), y_i) = \frac{1}{N} \cdot \sum_{i=1}^N |g(X_i) - y_i| \quad (4.1)$$

em que $\mathbf{X}$ e $\mathbf{Y}$ definem o conjunto de dados de treinamento, com todos os *inputs* e *outputs* do modelo. L_1 é a função perda absoluta, N é o tamanho dos dados de treinamento utilizados no processo de aprendizagem, y_i corresponde à variável resposta de cada instância de treinamento. X_i são os *inputs* do modelo na forma $X_i = (x_{i,1}, x_{i,2}, \dots, x_{i,d})$ com d preditores e $g(X_i)$ é a estimativa da variável resposta. A função perda quadrática é dada por

$$C(\mathbf{X}, \mathbf{Y}) = \frac{1}{N} \cdot \sum_{i=1}^N L_2(g(X_i), y_i) = \frac{1}{N} \sum_{i=1}^N (g(X_i) - y_i)^2 \quad (4.2)$$

Ambas equações (Equação 4.1 e Equação 4.2), conforme apresentadas, também são

conhecidas como funções de custo. Essas expressões são a base de métricas para seleção de modelos conforme será discutido na seção 4.3.

4.2 Modelos de aprendizagem aplicados à regressão

Os modelos de aprendizagem de máquina, enquanto modelos de regressão, podem ser paramétricos ou não paramétricos. Os modelos paramétricos assumem uma forma para a função a ser estimada, desse modo f é parametrizada por um número finito de parâmetros. Com os métodos não paramétricos, não é feita suposições sobre a complexidade de f (JAMES et al., 2014; IZBICKI; SANTOS, 2020). Em ambos os casos, dois objetivos podem ser alcançados com esses modelos: o preditivo e o inferencial. No primeiro, valores de uma variável resposta podem ser previstos com base em novos atributos fornecidos ao modelo, e o segundo, determina-se a significância desses atributos na variável resposta.

Considerando inicialmente os modelos paramétricos, o modelo de regressão mais simples é o linear, que é expresso por

$$g(X_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_d x_{i,d} \quad i = 1, 2, \dots, N \quad (4.3)$$

Neste modelo, o coeficiente β_0 é chamado de intercepto (ou *bias*), e os demais coeficientes são os coeficientes de regressão. O número de preditores do modelo é d e N o número de instâncias no conjunto de treinamento. Em forma matricial essa relação é dada por

$$g(\mathbf{X}) = \boldsymbol{\beta}^T \mathbf{X} \quad (4.4)$$

no qual o estimador dos coeficientes de regressão é deduzido analiticamente com o método dos mínimos quadrados, conforme dado por

$$\boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \quad (4.5)$$

Izbicki e Santos (2020) demonstram que, mesmo quando a função regressão verdadeira, a função f (vide Figura 17), consiste em uma relação não linear, a estimativa via mínimos quadrados (Equação 4.5) converge para o melhor estimador linear com base em $\mathbf{X}$. O melhor estimador linear pode ser representativo ou não. Além disso, para casos de mais parâmetros do que dados, esse estimador não é bem definido.

Uma relação não linear pode ser expressa, em alguns casos, como uma relação linear conforme a Equação 4.3 quando se utilizam transformações adequadas (MONTGOMERY; RUNGER, 2016). Dessa forma, a expressão dada por

$$\bar{g}(X_i) = \beta_0 + \beta_1 \theta_{i,1} + \beta_2 \theta_{i,2} + \dots + \beta_d \theta_{i,d} \quad i = 1, 2, \dots, N \quad (4.6)$$

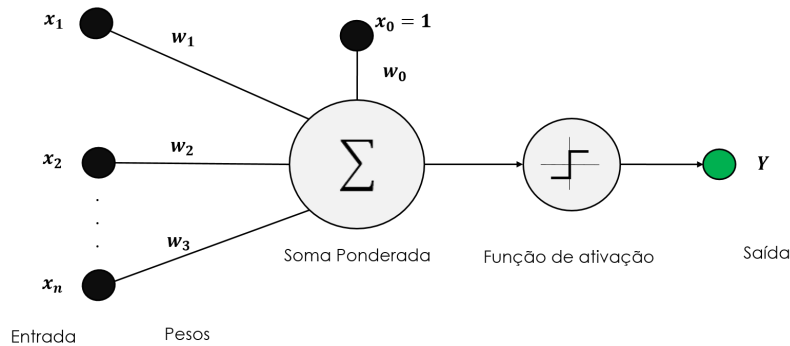
é um modelo de regressão linear equivalente a Equação 4.3, em que $\theta_{i,j}$, com $j \in \{1, 2 \dots d\}$, são variáveis transformadas. Desse modo, se $\theta_{i,j} = x_{i,j}^j$ temos a chamada regressão polinomial, e devido a esse artifício, a Equação 4.6 pode ser resolvida por meio da mesma Equação 4.5.

A seguir, dois modelos não paramétricos de regressão são explorados em mais detalhes. Os dois são modelos de aprendizado de máquina clássicos que permitem boas estimativas em problemas de regressão que envolvem estimar padrões complexos nos dados (IZBICKI; SANTOS, 2020).

4.2.1 Redes Neurais Artificiais

As Redes Neurais Artificiais, do inglês *Artificial Neural Network* (ANN), consistem em uma classe de algoritmos de aprendizagem que foram desenvolvidas tendo inicialmente a motivação de reproduzir o funcionamento biológico das redes neurais cerebrais. O desenvolvimento das ANNs foi iniciado por McCulloch e Pitts (1943), que formalizaram o primeiro modelo de neurônio artificial. Posteriormente, como mencionado em Goodfellow et al. (2016), o chamado *perceptron*, descrito formalmente em Rosenblatt (1958), definiu um modelo de neurônio artificial com um processo de aprendizagem bem definido e que é considerado, como mencionado em Géron (2021), a arquitetura mais simples de uma ANN. A Figura 18 apresenta esse modelo.

Figura 18 – Representação gráfica de um *perceptron*



Fonte: Adaptado de Mitchell (1997)

Neste modelo, é executada uma soma ponderada das suas entradas que são aplicadas em seguida na função de ativação. Esse processo corresponde ao limiar a ser excedido (por analogia com o limiar de impulsos em um neurônio biológico), que, por sua vez, gera o resultado final (MITCHELL, 1997; GÉRON, 2021). O *perceptron* foi aplicado inicialmente em problemas de classificação binária, e a função de ativação mais comum no contexto de classificação é a função sinal, que é dada por

$$\text{sgn}(z) = \begin{cases} -1, & \text{se } z < 0 \\ 0, & \text{se } z = 0 \\ 1, & \text{se } z > 0 \end{cases} \quad (4.7)$$

em que $z = w_1x_1 + w_2x_2 \dots w_nx_n$ e o limiar é definido como $z = 0$, w_i corresponde aos pesos do *perceptron* e x_i as entradas, conforme Figura 18. No caso de problemas de regressão, o *perceptron* não é associado a nenhuma função de ativação e, dessa forma, o problema se reduz a uma regressão linear análoga à mencionada anteriormente (vide Equação 4.3).

O processo de aprendizagem com o *perceptron* é efetivo em problemas de classificação com dados chamados linearmente separáveis, nos quais uma fronteira de decisão pode ser definida para classificar corretamente todas as instâncias de treinamento em suas classes. No entanto, como mencionado em Géron (2021), o *perceptron* não consegue apreender padrões complexos, o que motivou o desenvolvimento de novas estratégias com esses modelos. Identificou-se que a maioria das limitações pode ser resolvida entrelaçando vários *perceptrons* e definindo assim modelos mais complexos usando uma mesma unidade básica. Segundo Evsukoff (2020), as topologias mais importantes de ANN são: o modelo *perceptron* de múltiplas camadas, conhecido como *Multilayer Perceptron* (MLP); o modelo de rede convolucional, chamados de *Convolutional Neural Network* (CNN); o modelo de redes neurais concorrentes, conhecido como *Recurrent Neural Network* (RNN); os autocodificadores, conhecido como *Autoencoders* (AE) e o modelo generativo adversarial ou *Generative Adversarial Network* (GAN). O modelo MLP é o mais clássico e também o mais difundido em aplicações científicas, com utilização em diversas áreas.

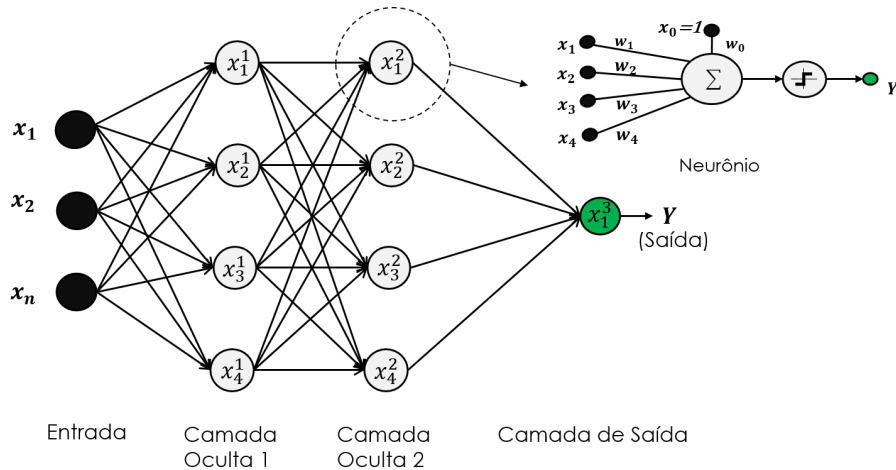
Esse modelo é um estimador não linear que pode ser aplicado para aprendizado de padrões complexos nos dados (IZBICKI; SANTOS, 2020; GÉRON, 2021). A estrutura MLP consiste no acoplamento de camadas ocultas que, por sua vez, contêm vários neurônios. As camadas são chamadas de ocultas porque os *outputs* de cada camada não são conhecidos ao longo do processo (GOODFELLOW et al., 2016). Os neurônios correspondem à estrutura de um *perceptron*. Ou seja, em cada um dos múltiplos neurônios das camadas, é feita a soma ponderada dos pesos e aplicada em uma função de ativação. De forma geral, o MLP pode ter várias camadas ocultas e um número diferente de neurônios em cada camada sendo que a escolha da quantidade de cada componente define a chamada arquitetura da rede neural, sendo que a escolha da melhor arquitetura depende do problema. No MLP, o sinal circula apenas na direção das entradas às saídas: as camadas de entrada recebem os sinais de *input*, as camadas intermediárias extraem e armazenam o conhecimento e a camada de saída emite o resultado do estímulo, desse modo, essa arquitetura também é conhecida como *feedforward* de múltiplas camadas. A Figura 19 apresenta o esquema do modelo MLP. Neste trabalho, será usado ANN para representar a arquitetura MLP.

Considerando a arquitetura de rede apresentada na Figura 19, o cálculo executado nos neurônios da primeira camada é dado por

$$z_j^1 = \phi \left(w_{0,j}^0 + \sum_{i=1}^n w_{i,j}^0 x_i^0 \right) \quad (4.8)$$

em que o sobrescrito indica a camada, sendo a camada de entrada com índice zero e, nesse contexto, $j \in \{1, 2, 3, 4\}$ corresponde ao número de neurônios nesta camada, sendo $i \in \{1, 2, \dots, n\}$

Figura 19 – Representação gráfica da arquitetura de MLP



Fonte: Autor (2024)

o número de neurônios na camada anterior ou, neste caso, o número de entradas na rede. ϕ é a função de ativação.

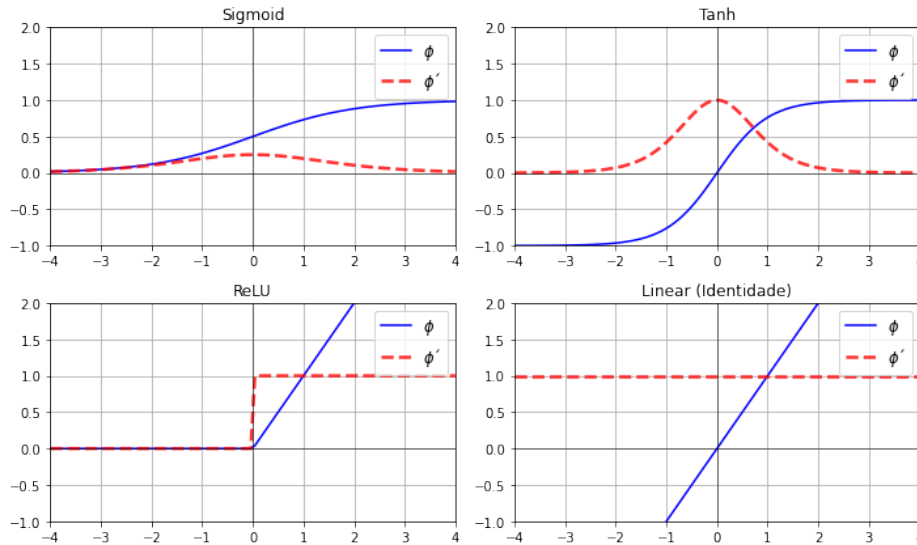
O processamento feito nos neurônios conforme a Equação 4.8 é propagado na rede, para demais neurônios e camadas, seguindo o mesmo princípio. As principais funções de ativação são apresentadas a seguir:

- Linear: $\phi(x) = x$
- Função sigmoideal: $\phi(x) = \frac{1}{1 + e^{-x}}$
- Tangente hiperbólica: $\phi(x) = \tanh(x) = \frac{e^{2x} - 1}{e^{2x} + 1}$
- ReLU: $\phi(x) = \max(0, x)$

Em modelos aplicados a problemas de regressão, o neurônio da camada de saída sempre possui uma função de ativação linear. Nas camadas ocultas, no entanto, são especificadas funções de ativação diferentes. Como mencionado em Géron (2021), funções de ativação precisam possuir não linearidade pois, caso contrário, o método como um todo fica limitado apenas a uma transformação linear. Além disso, funções de ativação como a apresentada na Equação 4.7 não são utilizadas nessa estrutura de rede mais complexa pois dificultam a estratégia que é comumente utilizada para treinar a ANN, o chamado *backpropagation* ou retropropagação. Essa estratégia demanda continuidade e, além disso, a função deve ser diferenciável. A Figura 20 ilustra as funções de ativação junto com as suas derivadas.

Desse modo, uma função de ativação linear nas camadas ocultas será útil apenas quando os dados possuírem padrão linear, uma vez que a rede será um aproximador linear. Como mencionado em Goodfellow et al. (2016), é demonstrado que uma ANN de apenas uma camada,

Figura 20 – Funções de ativação e suas derivadas



Fonte: Autor (2024)

com número suficiente de neurônios nessa camada, é capaz de representar qualquer função contínua. No entanto, não há garantia que o processo de treinamento utilizado de fato aprenda essa função. Desse modo, é necessário especificar funções de ativação adequadas e, como mencionado em (GOODFELLOW et al., 2016), em muitas circunstâncias, aumentar o número de camadas reduz o número de neurônios necessários para representar a função desejada. A escolha do tipo de função de ativação é uma especificação atrelada aos neurônios da rede e a arquitetura da ANN, e depende do problema analisado, como discutido em mais detalhes na subseção 4.3.2.

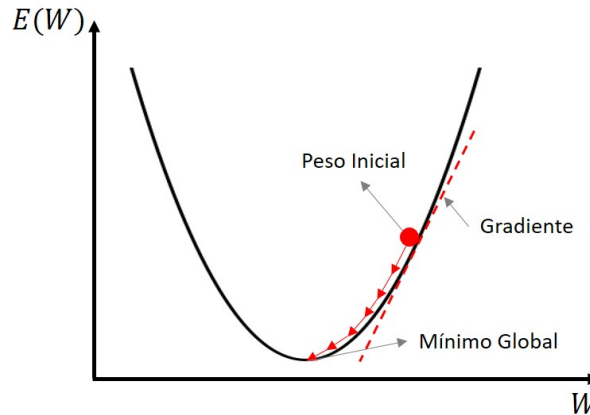
4.2.1.1 Estratégia de treinamento da ANN

O objetivo do treinamento da ANN é estimar os pesos de modo a reduzir a função custo (seja Equação 4.1 ou Equação 4.2). A alternativa mais empregada com esses modelos é a aplicação do tradicional método do gradiente descendente combinado com um processo chamado de retropropagação, que, por sua vez, é um método eficiente que foi desenvolvido para permitir a estimativa das componentes do vetor gradiente (IZBICKI; SANTOS, 2020; GÉRON, 2021).

O gradiente descendente é um método de otimização que utiliza apenas a informação da direção do gradiente para a busca do mínimo da função erro definida no processo de aprendizagem (a função de custo). Em Géron (2021), é mencionado que o método é utilizado para encontrar a solução ótima em uma grande variedade de problemas e, como mencionado também em Goodfellow et al. (2016), esse método funciona bem com vários tipos de ANNs. Como o gradiente informa a direção de maior crescimento de uma função, o gradiente descendente envolve caminhar na direção oposta ao vetor gradiente, portanto, em direção ao ponto de mínimo. A Figura 21 ilustra a estratégia de minimização.

Na Figura 21, a função erro $E(W)$ possui um mínimo global que é atingido em passos

Figura 21 – Representação do funcionamento do gradiente descendente



Fonte: Autor (2024)

tomados em direção ao mínimo quando é definido um valor de peso inicial na rede. O gradiente descendente, portanto, funciona iterativamente utilizando uma regra de atualização dos pesos que determina o passo ideal para caminhar na direção que produz a decida em direção ao mínimo. A expressão da regra de atualização empregada no método é dada por

$$W^{[i+1]} = W^{[i]} + \Delta W = W^{[i]} - \eta \nabla E(W^{[i]}) \quad (4.9)$$

em que os valores de peso da rede são atualizados com a informação do gradiente. O sobrescrito em colchetes indica as iterações. O $\nabla E(W)$ é o vetor gradiente e η é conhecido como taxa de aprendizado, que é responsável por controlar o tamanho do passo e que possui importância crucial para a convergência, pois um passo grande pode ser responsável por impossibilitar a convergência (GÉRON, 2021; GOODFELLOW et al., 2016). No método do gradiente descendente tradicional, utiliza-se em cada iteração toda as N instâncias do conjunto de treinamento para definir a métrica de erro. Considerando a métrica quadrática (norma L_2), a função de erro é dada por

$$E(W) = \frac{1}{N} \sum_{i=1}^N E_i(W) = \frac{1}{N} \sum_{i=1}^N \frac{1}{2} \sum_{k=1}^K (g_k(X_i, W) - y_{i,k})^2 \quad (4.10)$$

neste caso, o fator $1/2$ é adicionado por conveniência matemática e K representa o número de neurônios da camada de saída, uma situação mais geral que corresponde ao caso de redes com mais de uma saída.

Uma abordagem alternativa, conhecida como gradiente descendente estocástico é, segundo Goodfellow et al. (2016), a abordagem mais comum no treinamento de ANNs. Essa abordagem calcula o gradiente e atualiza os pesos para cada instância, selecionada aleatoriamente, dos dados de treinamento. Como mencionado em Géron (2021), devido a aleatoriedade do processo, a decida não é suave como no método tradicional e, desse modo, a decida para o mínimo é feita na média. O método apresenta vantagens como o menor custo computacional e,

ainda segundo Géron (2021), quando a função de custo possui irregularidades significativas, o gradiente descendente estocástico terá uma chance maior de evitar mínimos locais porque usa vários vetores gradientes em vez de apenas um para guiar a busca do mínimo.

4.2.1.2 Método de retropropagação

A estratégia empregada para encontrar o gradiente da Equação 4.9 de modo eficiente é conhecida como retropropagação e consiste em executar o gradiente descendente em uma ordem específica: Atualizando os pesos da última camada e, em seguida, atualizam-se os pesos da camada anterior e assim por diante até a camada de entrada (IZBICKI; SANTOS, 2020). O princípio da retropropagação é uma aplicação direta da regra da cadeia e é comumente utilizado com o gradiente descendente estocástico. Desse modo, é uma forma rápida e precisa para determinar todas as derivadas parciais do $\nabla E_n(W)$, para uma instância n qualquer dos dados, conforme apresentado por

$$\nabla E_n(W) = \begin{bmatrix} \frac{\partial E_n}{\partial w_1} \\ \frac{\partial E_n}{\partial w_2} \\ \vdots \\ \frac{\partial E_n}{\partial w_H} \end{bmatrix} \quad (4.11)$$

em que H é o número total de pesos da ANN. Considerando a saída da ANN como sendo definida, de maneira simplificada, por $\hat{y}$, que depende da instância de entrada e dos pesos, o erro é dado por

$$E_n = \frac{1}{2} \cdot \sum_{k=1}^K (\hat{y}_{n,k} - y_{n,k})^2 \quad (4.12)$$

Considerando $w_{j,i}$ a designação genérica do peso entre a entrada z_i (saída do neurônio anterior) e a saída v_j , tem-se a seguinte expressão para as componentes do vetor gradiente:

$$\frac{\partial E_n}{\partial w_{j,i}} = \frac{\partial E_n}{\partial v_j} \frac{\partial v_j}{\partial w_{j,i}} \quad (4.13)$$

na qual v_j é chamada de variável de ativação, ou seja, o resultado da ponderação das entradas em um neurônio. Este valor é calculado com os elementos da camada anterior, conforme:

$$v_j = \sum_{i=0}^P w_{j,i} z_i = w_{j,0} z_0 + w_{j,1} z_1 + w_{j,2} z_2 + \dots + w_{j,P} z_P \quad (4.14)$$

em que P é o número de elementos em uma camada anterior. Essa expressão é aplicada na função de ativação resultando em

$$z_j = \phi(v_j) \quad (4.15)$$

Na Equação 4.13, definindo-se a primeira derivada parcial convenientemente como δ_j e resolvendo a segunda derivada parcial usando a Equação 4.14, tem-se a seguinte expressão:

$$\frac{\partial E_n}{\partial w_{j,i}} = \delta_j \frac{\partial}{\partial w_{j,i}} \left(\sum_{i=0}^P w_{j,i} z_i \right) = \delta_j z_i \quad (4.16)$$

e, considerando o cenário da camada seguinte como sendo a camada de saída, tem-se a seguinte expressão:

$$\delta_k = \frac{\partial E_n}{\partial v_k} = \frac{\partial E_n}{\partial \hat{y}_{n,k}} = \frac{\partial}{\partial \hat{y}_{n,k}} \left(\frac{1}{2} \sum_{k=1}^K (\hat{y}_{n,k} - y_k)^2 \right) = \hat{y}_{n,k} - y_{n,k} \quad (4.17)$$

Considerando a Equação 4.14 e Equação 4.15, a variável de ativação calculada com os neurônios em uma camada anterior obtém-se a seguinte expressão:

$$v_k = \sum_{j=0}^P w_{k,j} z_j = \sum_{j=0}^P w_{k,j} \phi(v_j) \quad (4.18)$$

considerando um neurônio j qualquer de uma camada oculta, conectado a vários na camada seguinte (saída ou outra camada oculta), com K neurônios, e também considerando a regra da cadeia multivariável temos

$$\delta_j = \frac{\partial E_n}{\partial v_j} = \sum_{k=1}^K \frac{\partial E_n}{\partial v_k} \frac{\partial v_k}{\partial v_j} \quad (4.19)$$

que é equivalente a

$$\delta_j = \sum_{k=1}^K \frac{\partial E_n}{\partial v_k} \frac{\partial v_k}{\partial v_j} = \sum_{k=1}^K \delta_k w_{k,j} \phi'(v_j) \quad (4.20)$$

Retornando à Equação 4.16, e arranjando os termos considerando a Equação 4.20, pode-se expressar de uma forma geral as componentes do gradiente conforme a seguinte expressão:

$$\frac{\partial E_n}{\partial w_{j,i}} = \delta_j z_i = z_i \cdot \phi'(v_j) \left(\sum_{k=1}^K \delta_k \cdot w_{k,j} \right) \quad (4.21)$$

Essa expressão auxilia a determinação das componentes na Equação 4.11 que, por sua vez, é utilizada na Equação 4.9 para atualização dos pesos.

No processo de treinamento das ANNs, conforme descrito, é necessário definir valores iniciais de pesos para o começo do processo de treinamento. Essa definição geralmente é feita de forma aleatória e nem sempre resulta em uma configuração de peso perto do ideal. Como mencionado em Silva et al. (2016), é benéfico para convergência executar o treinamento mais de uma vez com diferentes valores de pesos. Passar os dados de entrada várias vezes na rede ajuda a atualizar os pesos para a direção ideal. Desse modo o treinamento das redes é feito em épocas. Uma época corresponde à passagem do conjunto de treinamento completamente pela rede. Os pesos encontrados em uma época são utilizados na época seguinte. O número de épocas do treinamento pode estabelecer um critério de parada no processo de aprendizagem. Encontrar um valor ideal de épocas depende do problema e do conjunto de dados conforme será discutido na subseção 4.3.2.

O procedimento pode ser resumido nas seguintes etapas:

Passo 1 - Define-se chute inicial para os pesos da ANN. Normalmente esses pesos iniciais são escolhidos de maneira aleatória.

Passo 2 - Nesta etapa, com os pesos iniciais definidos, executa-se a propagação para frente dos dados de entrada da rede. Determina-se os sinais transferidos camada por camada até o neurônio de saída. Desse modo, z_i e v_j são determinados inicialmente nesta etapa.

Passo 3 - Com o valor obtido no neurônio de saída, é possível determinar δ_k conforme Equação 4.17

Passo 4 - Partindo do valor de δ_k , calculam-se os valores de δ_j para todos os elementos das camadas ocultas.

Passo 5 - δ_j e z_i são utilizados para determinar as derivadas da função erro em relação aos parâmetros $w_{i,j}$, como é relacionado na Equação 4.16 e Equação 4.21

Passo 6 - Com os valores das derivadas, atualizam-se os pesos conforme Equação 4.9

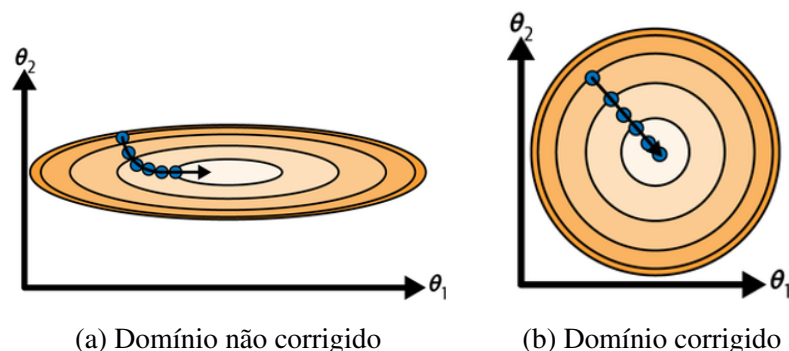
Passo 7 - Repetem-se os passos 2 até 6, apresentando nova instância dos dados de treinamento.

Passo 8 - Interrompe-se o processo de treinamento até que um determinado critério de convergência ser atingido.

4.2.1.3 Efeito de escala nos dados de entrada em ANNs

Certos métodos de aprendizagem de máquina são sensíveis ao efeito de escala presente nos dados de entrada. Para ANNs, esse efeito é importante pois afeta o processo de treinamento. Em cenário onde as entradas de dados possuem ordens de grandeza muito diferentes, por exemplo, é possível que o algoritmo de aprendizagem das ANNs esteja suscetível a uma maior dificuldade para atingir o mínimo. Esse efeito pode ser ilustrado na Figura 22.

Figura 22 – Curvas de nível em domínios com e sem correção da escala

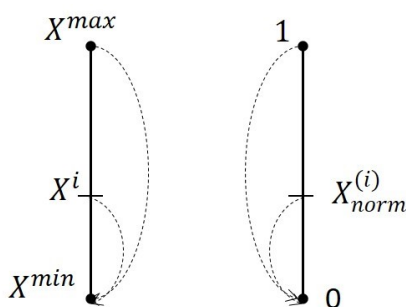


Fonte: Adaptado de Géron (2021)

No caso da Figura 22, um dos preditores do modelo possui escala muito maior e, portanto, o domínio é distorcido na Figura 22a. A convergência demanda um maior número de iterações para atingir o mínimo, caminhando em uma direção e em seguida para a outra. Na Figura 22b essa caminhada é direta para o mínimo. Como mencionado em Silva et al. (2016), não considerar esse efeito de escala nos dados pode também afetar as transformações que ocorrem nas funções de ativação, e, portanto, pode inviabilizar o processo de aprendizagem.

Duas abordagens principais são utilizadas em relação ao processo de correção da escala dos preditores: subtrair cada entrada de dado pela média do conjunto e dividir pelo desvio padrão dos dados de treinamento, que é um caso apropriado para quando as informações de entrada da rede seguem distribuição normal, processo que é conhecido como padronização (AOUN; MADARÁSZ, 2022). Outra opção, é transformar os dados de modo que o valor resultante esteja entre 0 e 1 ou entre -1 e 1 , esta transformação é baseada no princípio dos segmentos proporcionais de Thales e é uma das formas mais comuns de correção do efeito de escala nas ANNs. A Figura 23 ilustra o princípio utilizado para o caso em que o domínio esteja no intervalo fechado de 0 a 1. Esse processo é conhecido como normalização min-max.

Figura 23 – Princípio dos segmentos proporcionais utilizado no escalonamento



Fonte: Adaptado de Silva et al. (2016)

A expressão deduzida tomando por base a Figura 23 é dada por

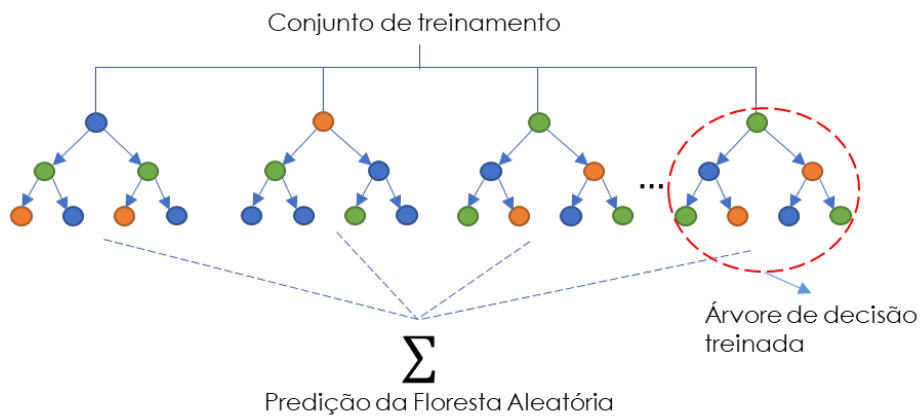
$$x_{norm}^{(i)} = \frac{x^{(i)} - x_{min}}{x_{max} - x_{min}} \quad (4.22)$$

em que, $x_{norm}^{(i)}$ é o novo valor de cada instância i . $x^{(i)}$ é o valor antigo, x_{max} e x_{min} correspondem ao maior e menor valor absoluto dentro do conjunto de dados.

4.2.2 Florestas Aleatórias

O método chamado Florestas Aleatórias, do inglês *Random Forest* (RF), faz parte de um tipo de método de aprendizagem de máquina chamado de aprendizagem por agrupamento ou *Ensembles*, que consiste na estratégia de agrupar múltiplos modelos que juntos produzem uma resposta melhor que a alternativa de usá-los isoladamente (MURPHY, 2022). Quando o agrupamento é feito com vários modelos do mesmo tipo, o *Ensemble* é homogêneo, caso contrário é chamado de heterogêneo. As RF recebem esse nome pois agrupam múltiplas árvores de decisão (*Ensemble* homogêneo). Essa estratégia foi desenvolvida no trabalho de Breiman (2001). A Figura 24 ilustra o esquema deste tipo de modelo.

Figura 24 – Esquema de agregação nas RF



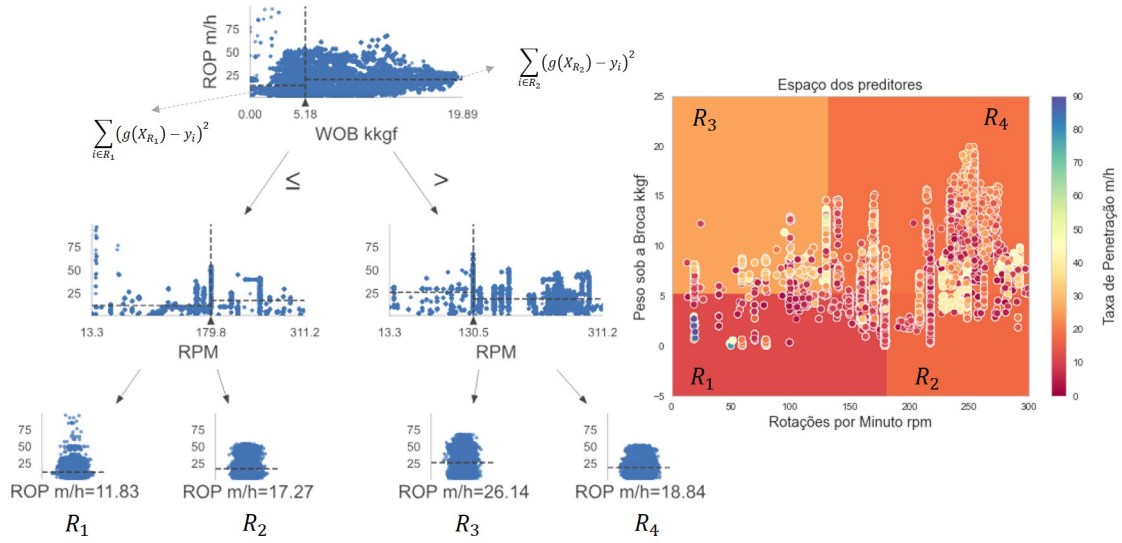
Fonte: Autor (2024)

Dessa forma, para compreender os modelos RF, é necessário estabelecer os conceitos principais da estimativa por árvores de decisão quando aplicada em problemas de regressão.

4.2.2.1 Árvores de Decisão

Segundo Izbicki e Santos (2020), uma árvore de decisão é construída inicialmente por particionamento recursivos no espaço dos preditores (os *inputs* do modelo). Cada particionamento é feito de modo que essas regiões não sejam sobrepostas. Neste processo, é definida uma condição de decisão (chamado de nó) a partir do qual será feito esse particionamento. Os resultados finais são chamados de folhas. A Figura 25 apresenta a estrutura de uma árvore de decisão e a partição correspondente no espaço dos preditores.

Figura 25 – Processo executado no método de árvores de decisão



Fonte: Autor (2024)

As regiões R_1 , R_2 , R_3 e R_4 da Figura 25 são utilizadas para obter a predição por meio do cálculo de médias. Ou seja, para um novo dado que se deseja obter a predição com o modelo, é determinado na divisão dos nós qual região R_j essa nova instância analisada pertence e, dessa forma, o preditor é simplesmente a média atrelada a essa região, que foi determinada com os dados de treinamento e com o particionamento encontrado no processo de aprendizagem.

Para o particionamento dessas regiões são executadas múltiplas divisões binárias, sendo necessário definir nessas divisões a variável de predição (X_j) junto com um valor s , que é o valor de corte. Desse modo, em cada divisão binária criada é definido um critério de decisão que separa duas regiões, como dado pelos seguintes conjuntos:

$$R_1(j, s) = \{X | X_j \leq s\} \quad (4.23)$$

$$R_2(j, s) = \{X | X_j > s\} \quad (4.24)$$

Segundo James et al. (2014) e Izibicki e Santos (2020), X_j e s são escolhidos em cada divisão binária como a melhor combinação que leva à redução da expressão:

$$\sum_{i: x_i \in R_1(j, s)} (g(X_{R_1}) - y_i)^2 + \sum_{i: x_i \in R_2(j, s)} (g(X_{R_2}) - y_i)^2 \quad (4.25)$$

Ou seja, em cada divisão determina-se a variável preditora e o valor de s que resulta na menor perda quadrática (Equação 1.2) nas duas regiões. $g(X_{R_j})$ é a predição com os dados da região R_j , que corresponde ao valor médio da variável resposta nas instâncias de treino da região R_j e y_i o valor da variável resposta para cada instância i na região R_j .

Na Figura 25, em cada divisão são indicados a reta vertical (s) e o valor $g(X_{R_1})$ (reta horizontal de cada região). O processo de busca executado é conhecido como busca gulosa, que é sempre feito de cima para baixo e sempre consiste em escolher o melhor valor para o passo analisado, sem escolher, por exemplo, uma divisão que resulte em uma melhor árvore em um passo futuro (JAMES et al., 2014). Prosseguindo com essas escolhas de divisão do espaço, novos espaços são criados e o processo é repetido nessas novas regiões até um critério de parada ser atingido. Um critério, por exemplo, é quando a quantidade de instâncias dentro de uma região é atingida.

Uma característica expressiva deste método é a sua alta interpretabilidade. Como mencionado em Izbicki e Santos (2020), os modelos de árvore de decisão podem, por exemplo, indicar quais as variáveis são mais importantes, utilizando, por exemplo, o nível de redução do erro quadrático em cada nova partição. No entanto, os métodos de árvores de decisão também possuem limitações bem conhecidas (GÉRON, 2021). Os modelos geralmente são muito dependentes dos dados utilizados no treinamento, o que leva à tendência de sobre-ajuste (JAMES et al., 2014; IZBICKI; SANTOS, 2020). Como mencionado em Géron (2021), a grande limitação é relacionada à instabilidade dos modelos de árvore de decisão, pois são sensíveis a pequenas variações nos dados de treinamento (variância do modelo é alta). Tais limitações motivaram o desenvolvimento de métodos mais sofisticados.

É importante mencionar que todos os modelos baseados em árvores de decisão não são sensíveis a escala dos preditores. Tendo em vista que uma árvore de decisão realiza o particionamento baseado em apenas um único preditor por vez, sendo esse processo independente dos demais, não é influenciado pela unidade de cada preditor.

4.2.2.2 A estratégia *bagging* e árvores de decisão descorrelacionadas

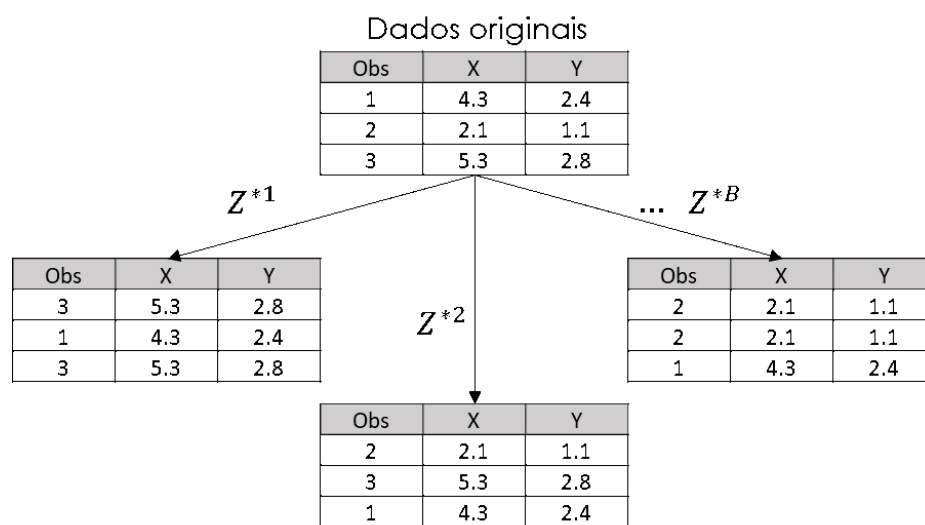
As RF utilizam uma técnica de reamostragem no processo de treinamento. Como mencionado anteriormente, várias árvores de decisão são utilizadas no treinamento de modelos RF e, por causa disso, o cenário ideal seria obter vários conjuntos de dados independentes para treinar essas árvores de decisão. A estratégia utilizada aproveita ao máximo o único conjunto de dados de treinamento e utiliza o chamado *bootstrap*, que consiste em retirar amostras aleatórias com reposição em um conjunto sempre com o mesmo tamanho (IZBICKI; SANTOS, 2020). A Figura 26 ilustra esse processo de reamostragem.

O modo que as RF agregam essas árvores de decisão é por meio de uma adaptação no processo conhecido como *bagging*. O *bagging* consiste simplesmente na média das predições de cada árvore de decisão ($g_b(X)$) conforme dado por

$$g(X) = \frac{1}{B} \sum_{b=1}^B g_b(X) \quad (4.26)$$

Dado um conjunto de observações independentes, cada uma com variância σ^2 , a média da variância das observações é dada por σ^2/n , em outras palavras, a estimativa pelo *bagging* possui menor variância do que a obtida por meio das árvores de decisão individuais (IZBICKI; SANTOS, 2020; AOUN; MADARÁSZ, 2022). Na construção do modelo preditivo por meio do *bagging*, em média dois terços das amostras do conjunto de treinamento são utilizados no processo de treinamento de cada preditor, o terço que não é utilizado recebe o nome de conjunto “*out of bag*”. Além disso, a predição geralmente melhora com a quantidade de árvores de decisão utilizadas até atingir um certo limite.

Figura 26 – Esquema do método de reamostragem *bootstrap*



Fonte: Autor (2024)

A estratégia do *bagging* é alterada nos modelos RF para tornar as árvores de decisão, desenvolvidas com cada amostra, menos correlacionadas. Essa modificação é feita para que não seja fornecida previsões muito semelhantes nas árvores de decisão. Dessa forma, é estipulado um limite de escolha das variáveis de predição X_j dentro de cada nó dessas árvores, de modo que apenas seja escolhido, de forma aleatória, sempre uma quantidade $p < P$, sendo P o número total de preditores considerados no treinamento, e que é menor que a quantidade de variáveis predictoras. Em cada nó um novo subconjunto com m variáveis é escolhido (JAMES et al., 2014; IZBICKI; SANTOS, 2020).

4.2.2.3 Processo de treinamento dos modelos RF

O processo de treinamento das RF pode ser descrito como uma sucessão de passos conforme apresentado a seguir:

Passo 1 - Executar o método de reamostragem com reposição no conjunto de treinamento. Cada amostra será a base de entrada de uma árvore da floresta.

Passo 2 - Para a primeira divisão do espaço dos preditores é amostrado aleatoriamente, com reposição, um número $p < P$ de preditores utilizado no modelo.

Passo 3 - Para cada um dos preditores amostrados $X_{j(1)}, X_{j(2)} \dots X_{j(p)}$ aplica-se o algoritmo de divisão do espaço:

Determina o valor $s_{j(k)}$ que divide a região em duas, com $R_1 = \{X | X_{j(k)} \leq s_{j(k)}\}$ e $R_2 = \{X | X_{j(k)} > s_{j(k)}\}$ medindo o erro quadrático em cada partição.

Passo 4 - Seleciona a variável $X_{j(k)}$ e o valor de corte $s_{j(k)}$ que produz o valor mínimo conforme a Equação 4.25.

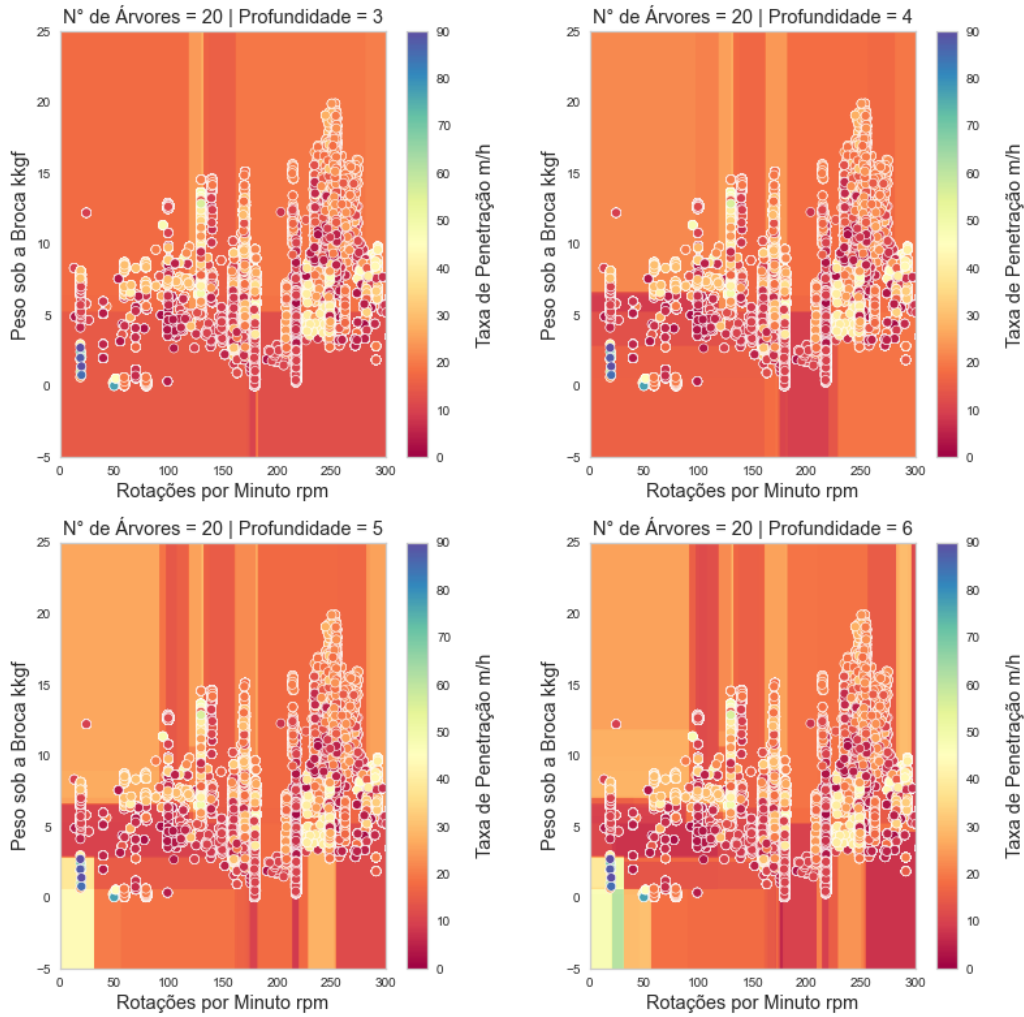
Passo 5 Verifica-se em R_1 se o critério de parada foi atingido. Se o critério for atingido temos uma folha, caso contrário temos um nó que poderá ainda ser particionado (retorna ao passo 2 até atingir o critério de parada).

Passo 6 Verifica-se em R_2 se o critério de parada foi atingido. Se o critério for atingido temos uma folha, caso contrário temos um nó que poderá ainda ser particionado (retorna ao passo 2 até atingir o critério de parada)

Passo 7 Retorna ao passo 1, utilizando outra amostra do método *bootstrap*. O processo é executado para todas as árvores de decisão.

A Figura 27 ilustra o particionamento do espaço de preditores utilizando várias árvores de decisão com diferentes profundidades, de onde é possível verificar o aumento do particionamento das regiões. Parâmetros como a profundidade das árvores e a quantidade de árvores utilizadas no modelo impactam na performance final do modelo. A predição final com o modelo é feita conforme a Equação 4.26.

Figura 27 – Exemplo de particionamento do espaço dos preditores no modelo RF



Fonte: Autor (2024)

4.3 Seleção de modelos e sobre-ajuste

Como mencionado anteriormente, a divisão dos dados em dois conjuntos na Figura 17 é uma etapa importante do processo. Ao separar os dados, é possível estimar o erro de generalização apropriadamente e, desse modo, é possível avaliar o modelo preditivo e também comparar o desempenho preditivo de diferentes modelos. A avaliação do modelo é feita com a estimativa do erro usando os dados de teste (dados não utilizados no treinamento) que corresponde ao erro de generalização do modelo. As principais métricas de erro utilizadas com modelos de regressão nos dados de teste são o erro médio absoluto, do inglês *Mean Absolute Error* (MAE) e a raiz do erro médio quadrático, do inglês *Root Mean Squared Error* (RMSE). O MAE é determinado pela seguinte expressão:

$$MAE = \frac{1}{M} \sum_{i=1}^M |y_i - g(X_i)| \quad (4.27)$$

e o RMSE pode ser determinado por

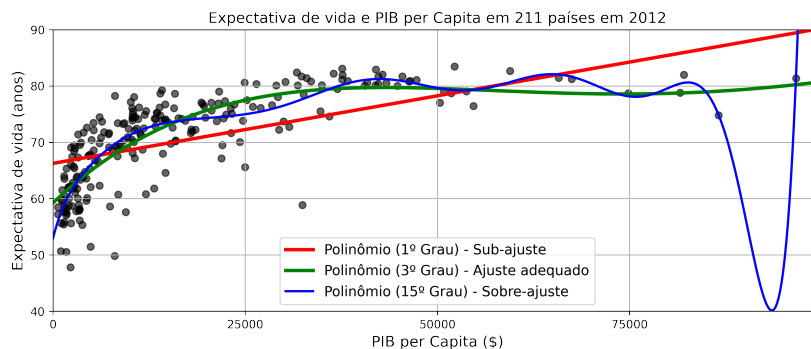
$$RMSE = \sqrt{\frac{1}{M} \sum_{i=1}^M (y_i - g(X_i))^2} \quad (4.28)$$

As duas expressões se relacionam diretamente com as funções de custo apresentadas na Equação 4.1 e na Equação 4.2, sendo aqui determinadas no conjunto de teste (de tamanho M) e utilizadas para estimar o erro de generalização e selecionar modelos, entre vários que seja necessário comparar (GÉRON, 2021).

O MAE indica o quanto que o modelo erra, em média, para cada nova observação. Esse tipo de métrica possui a mesma unidade da variável resposta e é menos sensível à presença de *outliers* nos dados (SOARES; GRAY, 2019; GÉRON, 2021). A RMSE, por considerar inicialmente o quadrado dos erros, é mais influenciada por *outliers*. Opta-se pelo RMSE para que as medidas de erro possuam mesma unidade do MAE. As duas métricas podem servir como critério de seleção para a performance dos modelos, sendo que geralmente o RMSE é maior que o MAE e, na presença de muitos *outliers*, as diferenças entre o MAE e a RMSE aumentam. (BOTCHKAREV, 2018; SOARES; GRAY, 2019; GÉRON, 2021).

Estimar o erro de generalização de um modelo de aprendizagem de máquina com dados utilizados no treinamento pode ser uma estimativa muito enviesada, uma vez que o processo de aprendizagem já envolve a redução do erro do modelo em relação a esses dados e, comparar diferentes modelos usando os erros do treinamento não resulta em selecionar um modelo que apresentará melhor predição para novos dados. O motivo para não haver garantia de desempenho na generalização do modelo preditivo quando se utilizam os dados de treinamento se deve a possibilidade de sobre-ajuste. O sobre-ajuste é quando o modelo captura não apenas o padrão presente nos dados mas também o ruído dos dados de treinamento, sendo que o modelo resultante torna-se dependente desses dados de treinamento e não apresenta generalização aceitável. Alternativamente, o sub-ajuste ocorre quando o modelo tem desempenho de ajuste ruim nos dados de treinamento por não ter complexidade adequada para capturar a função f . A Figura 28 ilustra o fenômeno de sobre-ajuste e sub-ajuste.

Figura 28 – Representação do sobre-ajuste e sub-ajuste



Fonte: Adaptado de Izbicki e Santos (2020)

Na Figura 28, a curva azul é um caso de sobre-ajuste. Se essa curva for utilizada com novos dados, erros elevados podem ser obtidos, pois a curva polinomial de 15º grau começa a interpolar instâncias de treinamento, resultando em um modelo mais complexo que o padrão seguido pelos dados. A curva vermelha também não retornaria valores adequados (caso de sub-ajuste), pois, nesse caso, o ajuste é feito com um modelo muito simples, que não é ideal para capturar o real padrão dos dados. Esse tipo de modelo mais simples é, na linguagem do aprendizado de máquina, um modelo com alto viés. A curva verde da Figura 28 seria a que proporcionaria melhor generalização (melhor predição para novos dados). De modo geral, a escolha do modelo intermediário entre sub-ajuste e super-ajuste é o ideal.

Modelos mais complexos possuem menor viés e, por isso, estão associados a maior risco de sobre-ajuste. Além disso, como mencionado em James et al. (2014), modelos mais complexos possuem variância maior devido a maior flexibilidade e, por isso, são sensíveis a pequenas alterações nos dados de treinamento. Para obter modelos de regressão com boa capacidade preditiva em novas instâncias de dados é preciso controlar tanto o viés quanto a variância. O controle da complexidade pode ser feito com uma boa avaliação e seleção de modelos e com a definição dos hiperparâmetros ótimos, ou seja, aqueles parâmetros que controlam a arquitetura do modelo (como mencionado na subseção 4.3.2).

4.3.1 Partição dos dados

Para o treinamento, avaliação de performance e seleção de modelos, é necessário aplicar uma estratégia de partição dos dados. A opção mais simples de particionamento é separar os dados em dois conjuntos, um para treinamento e outro para teste, como mencionado anteriormente. A literatura indica porcentagens para divisão ideal dos dados (por exemplo 70% e 30%, respectivamente) e recomenda utilizar seleção randômica neste particionamento (IZBICKI; SANTOS, 2020). No entanto, essa opção possui desvantagens quando se pretende realizar um refinamento dos hiperparâmetros, por exemplo. Como indicado em (GÉRON, 2021), ao selecionar hiperparâmetros é necessário realizar grande quantidade de testes para, de fato, encontrar o nível de complexidade necessário e aumentar a capacidade de generalização, no entanto, se esse processo for executado sempre nos mesmos dados os parâmetros encontrados não serão aqueles que oferecerão maior desempenho. Esse fato decorre da tendência do modelo de sofrer sobre-ajuste neste conjunto de dados (IZBICKI; SANTOS, 2020; GÉRON, 2021). A alternativa que é utilizada envolve criar um terceiro conjunto chamado de conjunto de validação para executar esse refinamento do modelo. Uma estratégia recomendada nesse processo é aplicar o método de validação cruzada (IZBICKI; SANTOS, 2020; GÉRON, 2021). A Figura 29 a seguir apresenta o esquema do método chamado validação cruzada *k-fold*.

Figura 29 – Representação do processo de validação cruzada *k-fold*



Fonte: Adaptado de Murphy (2022)

Na Figura 29 os dados do conjunto são separados em k lotes, nesse caso $k = 5$, e o modelo é treinado na primeira iteração com $k - 1$ lotes. O erro de predição do modelo é determinado com o lote deixado de fora no treinamento (lotes em destaque na Figura 29). O processo é repetido iterativamente de modo que o lote deixado de fora é sempre trocado em cada iteração, até que todos os lotes sejam utilizados no cálculo do erro de generalização. Esse método possibilita aproveitar melhor a quantidade de dados em análise, sendo ideal para conjuntos com poucas instâncias. O erro final determinado na validação cruzada é a média dos erros em cada iteração. Esse valor pode ser utilizado para comparar diferentes modelos ou selecionar os seus hiperparâmetros.

4.3.2 Refino de hiperparâmetros e o método de busca em *grid*

Modelos complexos de aprendizagem de máquina possuem muitos hiperparâmetros, como é o caso de modelos de redes neurais profundas convolucionais (GOODFELLOW et al., 2016). Os chamados hiperparâmetros são os parâmetros que configuram os modelos de aprendizagem de máquina e que geralmente são definidos antes do processo de treinamento em si e, portanto, não fazem parte do processo de ajuste de parâmetros durante o treinamento. Esses parâmetros controlam a arquitetura do modelo e flexibilizam o grau de complexidade.

Ao aplicar modelos de aprendizagem de máquina é necessário incluir um processo de refinamento desses hiperparâmetros de modo que seja possível melhorar a performance preditiva em um problema específico e também controlar o grau de complexidade do modelo para evitar o sobre-ajuste. Esse refino envolve obter hiperparâmetros ideais, com métodos específicos de otimização, quando se procura desenvolver soluções robustas. Segundo Géron (2021), a alternativa mais comum para encontrar esses hiperparâmetros ideais é realizar a busca em *grid* que, por sua vez, consiste em um método de força bruta que é utilizado em conjunto com a validação cruzada para selecionar modelos com os melhores hiperparâmetros. A Tabela 8 apresenta os principais hiperparâmetros dos dois modelos de aprendizagem de máquina discutidos neste trabalho, incluindo também os parâmetros ajustáveis durante o processo de

treinamento (parâmetros ordinários) que são comentados neste capítulo.

Tabela 8 – Parâmetros de aprendizagem do modelo e hiperparâmetros principais da RF e ANN

Modelo	Parâmetros de aprendizagem	Hiperparâmetros
RF	- Preditor utilizado em cada nó de cada árvore de decisão	- N° de estimadores (N° de árvores)
	- Valor de corte em cada nó de cada árvore de decisão	- Profundidade das árvores
ANN	Pesos	- Quantidade de preditores selecionados aleatoriamente na partição das regiões
		- Número mínimo de dados na folha de cada árvore de decisão
		- Número de camadas ocultas e neurônios em cada camada
		- Função de ativação
		- Número de épocas
		- Taxa de aprendizado

Fonte: Autor (2024)

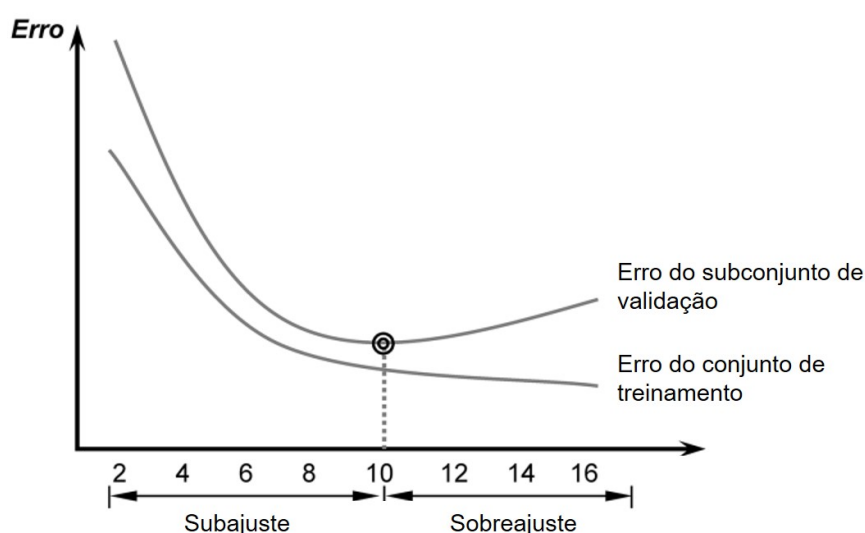
No caso do modelo RF, os hiperparâmetros principais controlam a estrutura das árvores de decisão individuais, a aleatoriedade presente no treinamento dessas árvores de decisão, assim como o tamanho da floresta. Aumentar o número de árvores de decisão no modelo geralmente induz melhoria na predição até determinado ponto, onde é atingido um valor assintótico. Aumentar a partir desse ponto apenas eleva o custo computacional do treinamento do modelo. Como mencionado em Probst (2019), a maior parte do aumento de performance pode ser atingida enquanto cresce as 100 primeiras árvores. O número de árvores de decisão ideal pode ser afetado pelo tamanho do conjunto de treinamento e também quando temos árvores menos correlacionadas, cenário onde geralmente é necessário um maior número de árvores de decisão.

Como mencionado em Géron (2021), reduzir o valor máximo permitido de profundidade atua como uma medida de regularização e, desse modo, diminui o risco de sobre-ajuste. Como mencionado em Bruce et al. (2020), pode-se variar a profundidade máxima geralmente no intervalo de 5 até 30. O hiperparâmetro que controla o número mínimo de dados na folha de cada árvore é um dos mais importantes para a performance do modelo, geralmente com valores entre 1 e 20 (MANTOVANI et al., 2018). Este hiperparâmetro também ajuda a controlar a profundidade das árvores, onde valores maiores reduzem a profundidade das árvores. A quantidade de preditores selecionados aleatoriamente exerce um papel significativo para descorrelacionar as árvores de decisão e, segundo Probst (2019), se há muitos preditores relevantes no modelo, selecionar um conjunto menor desses preditores em cada nó explora melhor os preditores com menor efeito na variável resposta. Desse modo, não são selecionados apenas os preditores influentes nas partições.

No caso do modelo ANN, os principais hiperparâmetros, conforme apresentados na Tabela 8, possuem influência significativa na performance do modelo. Géron (2021) informa que o hiperparâmetro mais importante é a taxa de aprendizagem, que é um valor entre 0 e 1. Uma taxa

de aprendizagem maior proporciona um treinamento mais rápido e pode provocar divergência no método do gradiente descendente ou uma convergência para mínimos locais. Um valor muito pequeno, implica em pequenas mudanças nos pesos em cada iteração e torna o processo de treinamento muito lento. O número de épocas não deve ser muito grande pois torna o custo computacional do processo de treinamento elevado e, principalmente, um número elevado de épocas pode induzir ao sobre-ajuste. Se o número de épocas for pequeno, pode ocorrer sub-ajuste. O número ideal é um hiperparâmetro importante, que poderia ser incluído em estratégias de otimização de hiperparâmetros. No entanto, uma técnica alternativa é a parada antecipada (*Early Stopping*), que contorna a necessidade do refino específico desse hiperparâmetro. A Figura 30 ilustra o funcionamento dessa técnica.

Figura 30 – Esquema do método de parada antecipada



Fonte: Adaptado de Silva et al. (2016)

Na parada antecipada, é separada uma parcela dos dados durante o treinamento como um conjunto de validação, e periodicamente o conjunto de validação é utilizado para monitorar o erro de predição da rede treinada em cada época, e dessa forma, caso o erro comece a aumentar o treinamento é interrompido e os parâmetros do modelo com melhor desempenho em uma época é salvo. Essa técnica atua como uma medida de regularização da rede e, como mencionado em Goodfellow et al. (2016), é provavelmente a estratégia mais comum de regularização. Essa técnica pode ser utilizada sozinha ou aliada a outras estratégias de regularização, como por exemplo a chamada regularização L_2 , que penaliza a função de perda adicionando um novo termo (produto da soma dos quadrados de todos os pesos da rede por um parâmetro α) conforme adicionado em mais detalhes em (RASCHKA; MIRJALILI, 2017).

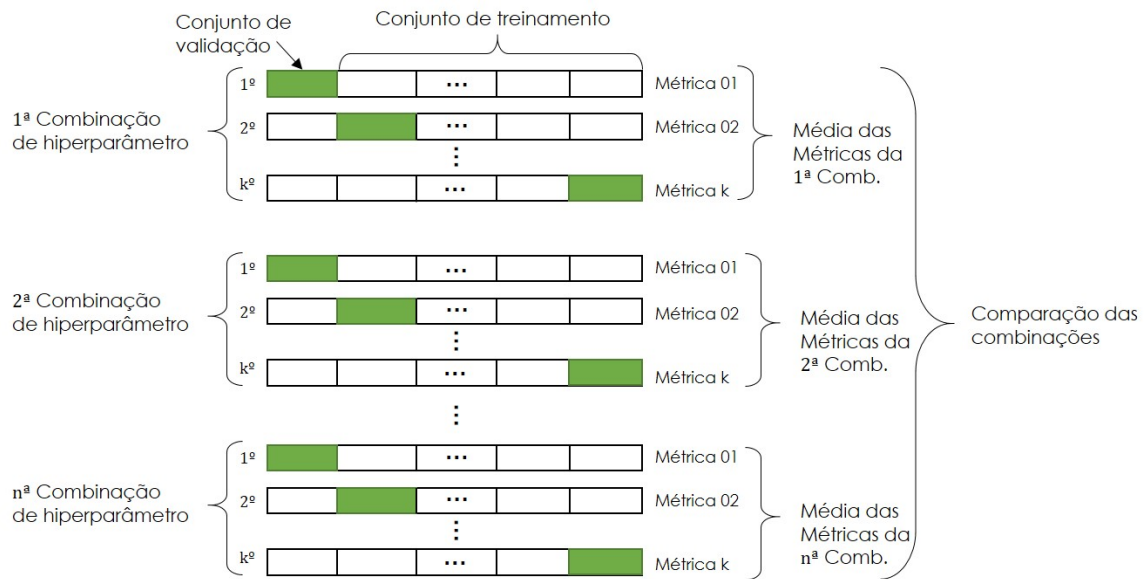
O número de neurônios e a quantidade de camadas ocultas são os parâmetros que definem a arquitetura do modelo ANN e são comumente objetos de análise em técnicas de otimização de hiperparâmetros. Os valores ideais dependem em grande parte do problema analisado. Para a definição do número de neurônios existem relações empíricas que auxiliam na definição desse

valor, no entanto, é importante testar diferentes valores que podem ser especificados com base em experiência com problemas similares. Em relação ao número de camadas ocultas, como mencionado por Géron (2021), é mais simples e eficiente escolher modelos com mais camadas e neurônios do que seria necessário e, em seguida, usar técnicas de regularização como a parada antecipada. Além disso, existe normalmente uma relação de custo benefício maior ao aumentar o número de camadas em vez do número de neurônios por camada. Geralmente, considera-se o mesmo número de neurônios para todas as camadas sendo que a medida que existe mais camadas ocultas a modelagem de funções complexas pode ser feita com cada vez menos neurônios em cada camada.

A função de ativação conhecida como ReLU geralmente é um bom padrão para todas as camadas ocultas. É mencionado a ampla utilização dessa função de ativação nos modelos ANN mais complexos, sendo responsável por evitar problemas inerentes no treinamento desses modelos e também por aumentar a velocidade desse processo (GOODFELLOW et al., 2016; GÉRON, 2021). Em camadas ocultas, funções tangente hiperbólica são preferíveis à sigmoide, ainda que não tão amplamente explorada quanto a ReLU e suas adaptações (GÉRON, 2021). Funções de ativação do tipo sigmoide e tangente hiperbólica ainda possuem aplicabilidade em certos problemas. A Função tangente hiperbólica pode ser adequada em problemas onde as variáveis resposta podem assumir valores negativos e a função sigmoide tem aplicações em modelos probabilísticos por retornar valores entre 0 e 1. Por isso que é um hiperparâmetro que precisa ser levado em conta nos modelos ANN (FONTANA, 2020).

Uma das estratégias mais simples para encontrar hiperparâmetros ótimos é fazer a chamada brusca em *grid*. Esse método é uma comparação por força bruta de modelos com diferentes hiperparâmetros. A estratégia consiste em utilizar o método de validação cruzada onde em cada partição dos dados é testado uma combinação diferente de hiperparâmetros. Desse modo, é possível comparar os desempenhos das diferentes versões do modelo e escolher o que apresenta menor erro. Com o modelo selecionado, o erro de generalização pode ser determinado no conjunto de teste. A Figura 31 ilustra esse processo.

Figura 31 – Esquema de funcionamento da busca em *grid*

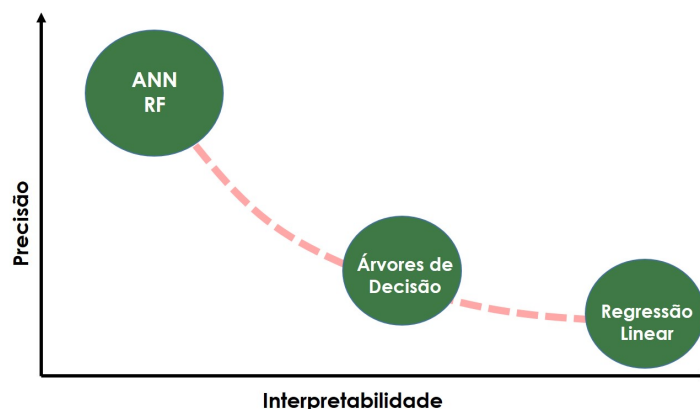


Fonte: Autor (2024)

4.4 Interpretabilidade dos modelos

A capacidade preditiva de modelos de regressão é o objetivo principal na maioria das aplicações. No entanto, a interpretabilidade de modelos de regressão também é um fator importante, uma vez que possibilita uma compreensão mais aprofundada do problema, de onde novos *insights* acerca das relações entre as variáveis de predição podem ser obtidos (IZBICKI; SANTOS, 2020). Modelos interpretáveis são aqueles de onde as relações entre as variáveis preditoras para gerar a resposta de predição é compreendida pelo usuário. A utilização de modelos com boa capacidade preditiva e boa interpretabilidade seria ideal, no entanto, nem sempre um modelo mais interpretável tem uma boa performance no problema analisado. A Figura 32 ilustra a relação entre diferentes modelos preditivos quanto à precisão e interpretabilidade.

Figura 32 – Balanço precisão e interpretabilidade dos modelos citados



Fonte: Autor (2024)

Na Figura 32 é possível identificar que o modelo de regressão linear é facilmente interpretável, no entanto, geralmente não atinge a mesma precisão de modelos mais complexos como ANN e RF. A menor interpretabilidade de certos modelos, chamados de modelos caixa preta, motivaram o desenvolvimento de métodos de interpretabilidade mais sofisticados, ainda que muitos possuam limitações e utilizem aproximações locais usando regressão linear (IZBICKI; SANTOS, 2020; MITA, 2021; MOLNAR, 2022). Segundo Mita (2021), uma das maneiras de compreender as decisões de um algoritmo de aprendizagem de máquina é medir a influência de cada variável preditora na variável resposta. Essa abordagem é chamada de importância de atributos.

4.4.0.1 Importância de atributos em regressão linear

A importância dos regressores de um modelo linear de regressão pode ser mensurada pelo valor absoluto do valor-t (t_β) (MOLNAR, 2019). O t_β pode ser estimado pela seguinte expressão:

$$t_\beta = \frac{\beta_j}{\sqrt{\hat{\sigma}^2 \cdot C_{jj}}} \quad (4.29)$$

em que, β_j corresponde ao coeficiente da regressão e o denominador corresponde ao chamado de erro padrão do coeficiente, e depende da variância do erro $\hat{\sigma}^2$ (estimada pela soma dos quadrados dos resíduos conforme mencionado em Montgomery e Runger (2016)) e o elemento da diagonal principal da matriz $C = (X^T X)^{-1}$, em que X é a matriz formada pelas instâncias de dados, conforme Equação 4.5. O t_β está associado a cada preditor de regressão específica, e estabelece uma métrica que pode ser comparada entre diferentes variáveis, uma vez que é adimensional. Além disso, o t_β pode ser comparado com valores críticos e métricas estatísticas, como o valor-p, que indicam a significância estatística dos preditores para o modelo. Modelos lineares revelam explicações confiáveis da relação entre a variável modelada e os preditores escolhidos desde que o modelo linear seja apropriado para a modelagem do fenômeno estudado.

Montgomery e Runger (2016) mencionam que a existência de multicolinearidade entre as variáveis pode afetar significativamente a interpretabilidade do modelo, pois a estimativa dos coeficientes da regressão linear é afetada. Desse modo, é necessário verificar se há problemas de multicolinearidade ou não. Uma forma de fazer essa verificação é calculando o chamado Fator de Inflação da Variância (FIV) das variáveis, que não deve ser superior a 5. A expressão do FIV é dada por

$$FIV(\beta_j) = \frac{1}{1 - R_j^2} \quad (4.30)$$

em que R_j é o popular coeficiente de determinação múltipla associado a cada variável regressora, que consiste em uma métrica frequentemente usada para avaliação de modelos de regressão

linear, conforme apresentado em mais detalhes em Montgomery e Runger (2016). Para solucionar o problema da multicolinearidade, uma vez identificado, uma possibilidade é a remoção do preditor que causa o problema de multicolinearidade, desde que mantenha a qualidade do ajuste e os FIVs determinados com os preditores remanecentes fiquem abaixo de 5.

5 CONJUNTO DE DADOS E PLANEJAMENTO DAS ANÁLISES COMPARATIVAS

Considerando as estratégias descritas no Capítulo 3, é perceptível que o problema de predição da ROP é relevante e de difícil modelagem considerando a grande quantidade de variáveis que interagem ao mesmo tempo durante a perfuração. Muitas estratégias tradicionais foram desenvolvidas, seja aplicando técnicas de regressão linear envolvendo relações entre parâmetros relevantes da perfuração, ou então utilizando o conceito da energia mecânica específica. Uma abordagem mais recente envolve empregar modelos baseados em dados usando aprendizagem de máquina, para captar as relações entre os parâmetros envolvidos na perfuração e a ROP utilizando apenas dados, sem especificar, de antemão, a complexidade matemática do padrão a ser aprendido nesses dados. Estratégias para predição da ROP, que possibilitam otimizar os parâmetros operacionais em tempo real, são viáveis por meio de modelos para predição da ROP precisos, portanto, para construir esses modelos e testar a sua acurácia e adequabilidade na predição da ROP é necessário empregar análises em grande quantidade de dados e de preferência em vários poços (SOARES; GRAY, 2019; TUNKIEL et al., 2021; HÖHN et al., 2022).

Como mencionado por Nagy (2019), a exposição e aplicação crescente dos modelos baseados em dados demandam o desenvolvimento de *datasets* confiáveis e consistentes para treinar modelos de aprendizagem de máquina. Além disso, o autor menciona que o processo de compartilhamento e transferência de dados de experiência de campo, entre a indústria de petróleo e os institutos de pesquisa, ganhou relevância nos últimos anos. Nesse contexto, uma importante iniciativa foi a liberação de dados da Equinor (2018) com licença de uso permissiva para o público. Os dados do campo petrolífero *Volve*, na região do mar do norte na Noruega, foram tornados públicos pela Equinor (2018). Os dados consistem em cerca de 5 *Tb* de informação divididos em 40 mil arquivos com dados de perfuração, de produção, interpretação geofísica, perfilagens, sísmica e relatórios diários. A posição oficial da empresa em relação aos dados liberados é melhor descrita por meio da seguinte citação:

"Nós acreditamos que o potencial de aprendizado para estudantes é enorme quando eles podem treinar com dados reais...Nós compartilhamos esses dados também para encorajar maior produtividade e inovação na indústria. Nós esperamos que não ajudará apenas futuros inovadores na área de energia em seus trabalhos, mas também contribuirá em operações mais eficientes e possivelmente uma melhor interação entre os *players* de nossa indústria."(EQUINOR, 2018 apud NAGY, 2019, tradução nossa)

Os dados do campo *Volve* proporcionam uma oportunidade para aplicar diferentes estratégias de predição e otimização da ROP. Em Tunkiel et al. (2021), os autores discutem esse cenário e estruturam um *dataset* de referência para esses estudos, o *dataset USROP*, que agrega dados de 7 poços. No entanto, como mencionado pelos autores, os dados disponíveis do campo *Volve* contêm ainda muitas informações que não foram processadas e incluídas nesse *dataset*, sendo que informações de litologia, desgaste da broca e outros parâmetros necessários para a

implementação de modelos, como o BYM, são algumas sugestões relevantes dada pelos autores para a melhoria, ou desenvolvimento de outros *datasets*, que contribuam para ampliar as análises com esses modelos.

5.1 Novo *dataset* de referência para estudos sobre a ROP

Para aplicar a maioria dos modelos mencionados no Capítulo 3, por exemplo, seria necessário adicionar novas informações ao *dataset USROP*. No entanto, quando se analisam os demais dados disponíveis para cada um dos poços desse *dataset*, percebe-se, por exemplo, que não é possível aplicar a maioria dos modelos do Capítulo 3 de forma adequada. Os poços considerados em Tunkiel et al. (2021) são: F09A, F07, F14, F15, F15S, F05 e F09 (seguindo nomenclatura adotada para os poços do campo *Volve*). Portanto, neste trabalho, opta-se por analisar outros poços e montar um *dataset* com todas as informações necessárias.

A Tabela 9 apresenta as coordenadas dos poços escolhidos e o ano de início de cada um. Além disso, na Figura 33 é possível entender melhor a localização desses poços no mar do norte e a distância relativa entre eles.

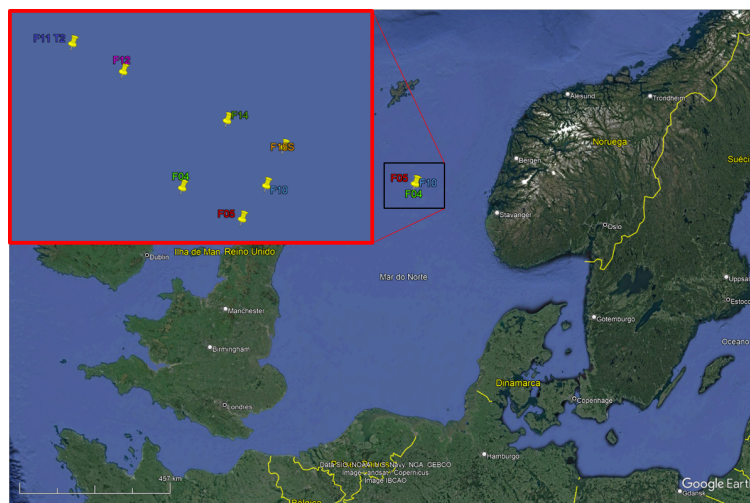
Tabela 9 – Informações sobre poços analisados

Poços	N-S	E-W	Início do poço
F05	58° 26' 29.66" N	1° 53' 14.99" E	2007
F14	58° 26' 29.77" N	1° 53' 15.08" E	2007
F15S	58° 26' 29.71" N	1° 53' 15.15" E	2007
F04	58° 26' 29.72" N	1° 53' 14.92" E	2007
F11_T2	58° 26' 29.96" N	1° 53' 14.87" E	2013
F10	58° 26' 29.68" N	1° 53' 15.07" E	2009
F12	58° 26' 29.89" N	1° 53' 14.94" E	2007

Fonte: Autor (2024)

Como pode ser observado, 4 novos poços são incluídos no conjunto de poços analisados por Tunkiel et al. (2021), que possuem informações suficientes (F05, F14 e F15S). Além disso, é possível observar a proximidade entre todos os poços apresentados. O critério de seleção desses novos poços é a quantidade de dados disponíveis, em que é possível agregar as informações consideradas por Tunkiel et al. (2021) no *dataset USROP* e novas informações que são necessárias para empregar os modelos mencionados no Capítulo 3. Essas informações adicionadas auxiliam também na melhoria de modelos de aprendizagem de máquina, uma vez que é possível avaliar melhores estratégias de treinamento dos modelos considerando todas essas novas informações e também interpretar melhor os resultados da predição da ROP com os diferentes modelos.

Na Tabela 10, são apresentadas algumas informações relevantes sobre o conjunto de dados utilizados neste trabalho, como o intervalo disponível para análise, a quantidade de pontos (linhas do *dataset*) e também a ROP média de cada poço.

Figura 33 – Mapeamento dos poços utilizados no novo *dataset* por meio do *Google Earth*

Fonte: Autor (2024)

Tabela 10 – Intervalos de disponíveis para os novos poços

Poços	MD Inicial (m)	MD Final (m)	Intervalo disponível (m)	ROP média (m/h)	Nº de Pontos
F05	2931,17	3765,91	834,74	28,28	5459
F14	2780,28	3465,85	685,57	18,02	5457
F15S	2554,83	4062,98	1508,15	14,27	16212
F04	2750,06	3483,86	733,81	19,95	4474
F11_T2	3600	4522,9	922,9	18,88	8307
F10	3800,13	4299,725	499,6	21,25	4818
F12	2450,14	3442,75	992,62	11,28	9474

Fonte: Autor (2024)

Dentro do intervalo de profundidade apresentado na Tabela 10, as informações que são coletadas para cada poço estão apresentadas na Tabela 11. Os dados principais referenciados na Tabela 11 podem ser visualizados para cada poço, ao longo da profundidade, no Apêndice B e Apêndice D. É possível perceber que as informações extraídas para cada poço correspondem a medições de parâmetros operacionais, dados de brocas e dados da formação perfurada. Parâmetros importantes para os modelos discutidos neste trabalho, como o gradiente de poro pressão e o UCS, que são determinados por meio dos atributos apresentados na Tabela 11, também são apresentados no Apêndice B e Apêndice D. Os métodos utilizados para determinar o gradiente de poro pressão e o UCS, para cada poço, são apresentados em mais detalhes no Apêndice A e Apêndice C. Todos os dados que compõem o novo *dataset* possibilitam executar as análises comparativas entre as diferentes estratégias de predição da ROP em tempo real investigadas neste trabalho.

Tabela 11 – Atributos no novo *dataset* presente em cada poço

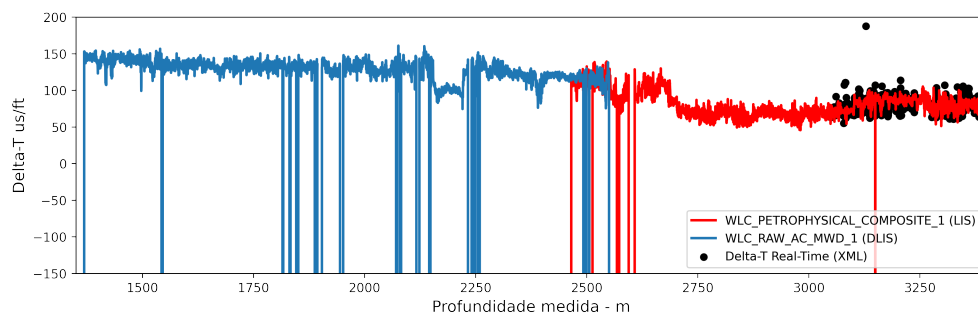
Atributos	Unidade
Profundidade medida	<i>m</i>
Peso sobre a broca	<i>kgf</i>
Pressão de injeção média	<i>kPa</i>
Torque médio na superfície	<i>kN · m</i>
Taxa de penetração (ROP)	<i>m/h</i>
Velocidade de rotação média	<i>rev/min</i>
Vazão do fluido	<i>L/min</i>
Densidade do fluido	<i>g/cm</i>
Diâmetro da broca	<i>mm</i>
Peso médio no gancho	<i>kgf</i>
Profundidade verdadeira (TVD)	<i>m</i>
Densidade equivalente dinâmica (ECD)	<i>g/cm³</i>
Delta-T compressional	<i>us/ft</i>
<i>Bulk density</i>	<i>g/cm³</i>
<i>Gamma ray</i>	<i>gAPI</i>
Inclinação	<i>deg</i>
TFA	<i>in²</i>
<i>Dull grading</i>	-
Tipo de broca	-
Litologia	-

Fonte: Autor (2024)

5.1.1 Aspectos relevantes sobre os dados utilizados

Como explicado em Tunkiel et al. (2020) e Tunkiel et al. (2021), o trabalho de desenvolvimento do *dataset USROP* baseou-se na extração de dados presente em arquivos *.XML* estruturados conforme um padrão muito usado na indústria para transmissão de dados em tempo real, o WITSML (do inglês *Wellsite Information Transfer Standard Markup Language*). Os arquivos *.XML* agregam dados que vêm de diferentes fontes, como do *mud logging*, e também dados coletados dos equipamentos LWD/MWD. No entanto, para um mesmo poço, dados adicionais podem ser encontrados em outros arquivos, como indicado na Figura 34.

Figura 34 – Comparação entre o Delta-T medido de um poço em diferentes tipo de arquivo



Fonte: Autor (2024)

Na Figura 34, os dados encontrados em um arquivo *.XML* pertencem a um trecho do poço e os demais dados podem ser obtidos em outros arquivos, como, por exemplo, em formato *.dlis*, *.lis* e *.las*, que também são formatos de arquivos comuns na indústria e disponíveis para muitos poços presente nos dados da Equinor (2018). Desse modo, é possível comparar dados entre diferentes arquivos e compatibilizar essas informações para um mesmo intervalo. Para obter grande parte dos atributos conforme a Tabela 11, é preciso comparar as informações entre diferentes arquivos e se certificar de que correspondem ao mesmo poço. A maioria das informações relacionadas a medições de parâmetros operacionais e de dados dos perfis geofísicos são obtidos dessa forma.

A identificação da litologia pode ser encontrada para alguns poços em arquivos *mud logging* de extensão *.dlis* ou *.lis*, no entanto, esses dados correspondem a informações codificadas e não a litologia interpretada. A descrição da litologia de cada poço pode ser obtida em arquivos *.pdf* chamados de perfis de avaliação de formação, em que é possível extrair os intervalos de cada litologia ao longo da profundidade. Esses dados são extraídos manualmente desses arquivos e, em seguida, inseridos no *dataset* utilizando a informação da profundidade para relacionar o tipo correto de litologia. A Tabela 12 apresenta as litologias que ocorrem em todos os poços presentes no novo *dataset* e, no Apêndice D, é possível visualizar os perfis de litologia.

Tabela 12 – Litologias encontradas nos poços analisados e suas descrições

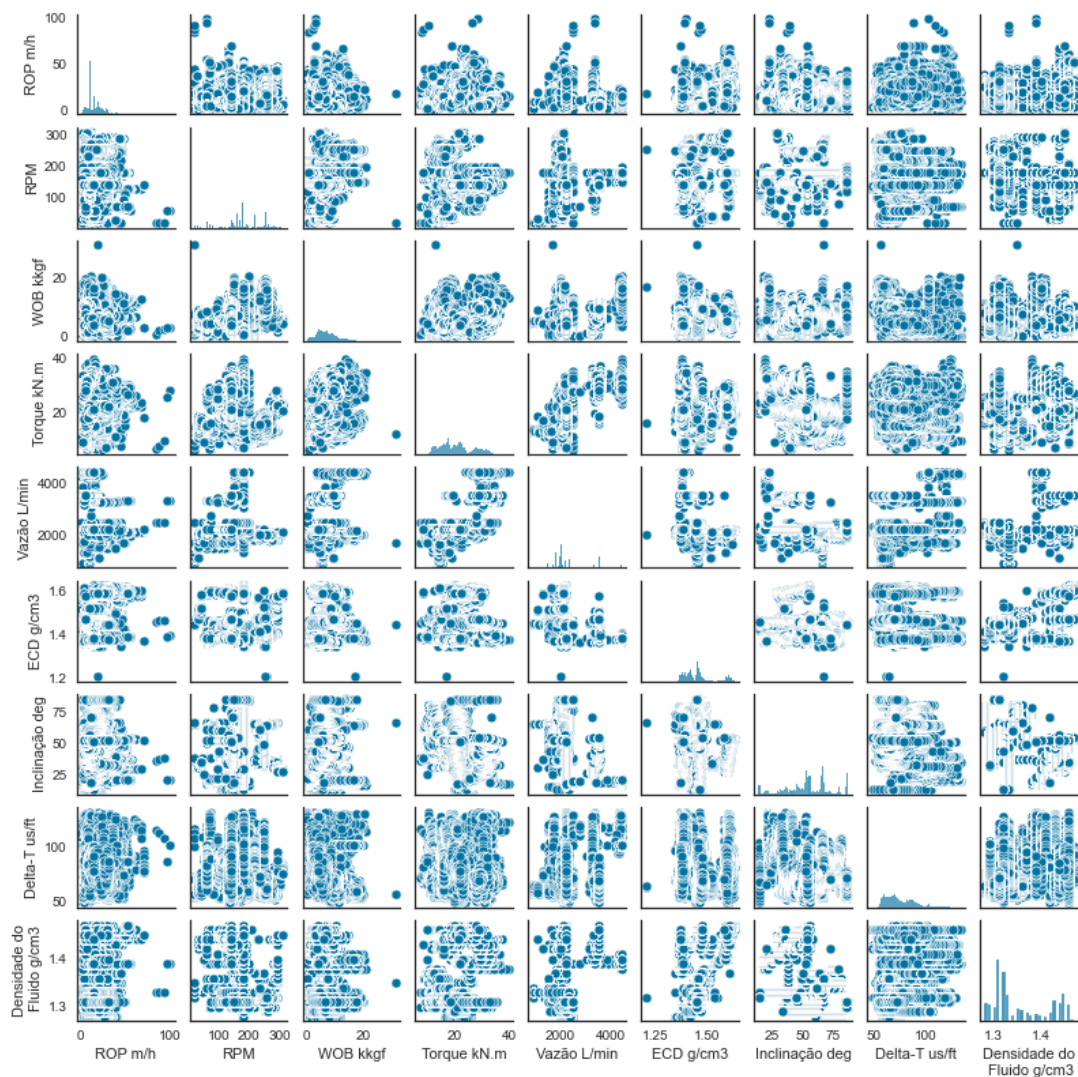
Litologia	Características
Argilito (<i>Claystone</i>)	Uma granulação fina de rocha sedimentar consistindo principalmente de argila compactada e endurecida e semelhante ao xisto, mas sem laminações.
Cálcario (<i>Limestone</i>)	Rocha sedimentar consistindo principalmente de carbonato de cálcio, depositado com os restos calcários de animais ou quimicamente precipitado a partir do mar. Rocha sedimentar consistindo principalmente de carbonato de cálcio, depositado com os restos calcários de animais marinhos ou quimicamente precipitado a partir do mar.
Marga (<i>Marl</i>)	Uma mistura de carbonatos de cálcio, argilas, magnésio e restos de conchas.
Arenito (<i>Sandstone</i>)	Quaisquer grupos de rochas sedimentares comuns constituídas por grãos de areia consolidadas com materiais como quartzo, hematite e minerais de argila.
Siltito (<i>Siltstone</i>)	Material sedimentar consistindo de partículas muito finas de tamanho intermediário entre areia e argila.
Carvão (<i>Coal</i>)	Rocha sedimentar contendo grande quantidade de material orgânico.

Fonte: Adaptado de Leite (2012)

Os dados relacionados às brocas utilizadas em cada poço são obtidos dos relatórios de brocas disponíveis para cada poço. As informações sobre a broca que compõe a Tabela 11 (Tipo de broca, *Dull grading* e TFA) são disponibilizadas em mais detalhes no Apêndice B.

Como é possível perceber na Figura 34, certos dados carecem de tratamento, uma vez que é possível identificar valores fisicamente impossíveis e que sinalizam erros claros na medição. É o caso da presença de valores -999.99 em certas medições (como na Figura 34). Esses valores são removidos. Além disso, valores espúrios, como ROP, WOB e RPM negativos, que correspondem a possíveis erros de medição, também são removidos. No trabalho de Tunkiel et al. (2021), para desenvolvimento do *dataset USROP*, alguns valores de referência foram utilizados para remover valores discrepantes nos dados, como, por exemplo: pressão de injeção acima de 25000 kPa, ROP acima de 100 m/h e WOB acima de 35 kkgf. Esses valores também são adotados. Métodos de identificação de *outliers* também são aplicados, como o popular método *z-score* (mais detalhes em Bruce et al. (2020)). No entanto, ocasionalmente pontos considerados como *outliers* para um determinado poço são mantidos, quando considerado valores razoáveis em relação aos dados dos demais poços. A Figura 35 apresenta visualização dos dados agrupados de todos os poços da Tabela 9, considerando os principais atributos utilizados neste trabalho.

Figura 35 – Dados agrupados de todos os poços do *dataset*



Fonte: Autor (2024)

Na Figura 35 é apresentado o histograma de frequências de cada atributo na diagonal e, em cada linha, o gráfico de dispersão entre o atributo da linha e cada um dos demais atributos de cada coluna. É possível perceber que, de modo geral, não há valores extremos e os dados encontram-se dentro de um intervalo realista de valores. Nenhum atributo do *dataset* possui mais que 4,5% de ocorrências identificadas como *outliers* usando o método *z-score* no intervalo entre -3 e 3. Além de valores extremos e valores numéricos fisicamente impossíveis, existem situações de valores ausentes de algum atributo específico, mesmo que os demais estejam presentes, caso em que surge a presença de *NaN* (valores não numéricos) no *dataset*. Como comentado em Tunkiel et al. (2020), uma alternativa simples é utilizar a propagação adiante, em que o valor imediatamente anterior é repetido para o seguinte, e a propagação reversa, em que o valor imediatamente posterior é repetido para o anterior. Ambas técnicas são utilizadas.

5.2 Modelos utilizados neste trabalho

Como mencionado no Capítulo 3, o BYM (Equação 3.2) é um dos modelos tradicionais mais relevantes, pois serviu de base para o desenvolvimento de outros modelos para ROP e, como também mencionado em Najjarpour et al. (2021), é o modelo mais utilizado para comparação de performance com demais estratégias, como, por exemplo, métodos de aprendizagem de máquina. Considerando modificações encontradas na literatura, conforme apresentadas na Equação 3.11, 3.13, 3.14 e, incluindo também, a modificação da subfunção f_6 (substituindo o RPM pelo torque) é possível definir uma versão modificada do BYM, o BYMmod. O torque utilizado é aquele medido na superfície, conforme mencionado na Tabela 11. As novas normalizações incluídas na Tabela 7 também são consideradas neste modelo modificado. Essas modificações tornam o modelo mais representativo para cenário de perfuração com outros tipos de broca e, portanto, tornam o modelo relevante como referência para outras estratégias. Os modelos BYMSoares (Equação 3.15) e MSE (Equação 3.23) são modelos que agregam informações mais facilmente obtidas em tempo real sendo, portanto, aplicáveis nas estratégias comparadas neste trabalho, conforme detalhado mais adiante. A Tabela 13 lista os modelos tradicionais aplicados neste trabalho com o indicativo dos parâmetros considerados em cada modelo.

Tabela 13 – Informações necessárias nos modelos tradicionais analisados

Modelo	Parâmetros medidos	Parâmetros calculados (Apêndice A, B, C, D)
BYM	TVD, ECD, D_b , WOB, RPM	PPG, h , F_j
BYMmod	TVD, ECD, D_b , WOB, Inclinação, Torque	PPG, h , ΔP_b , HSI
BYMSoares	TVD, WOB, RPM, Vazão do fluido	-
MSE	WOB, RPM, D_b , Inclinação	UCS

Fonte: Autor (2024)

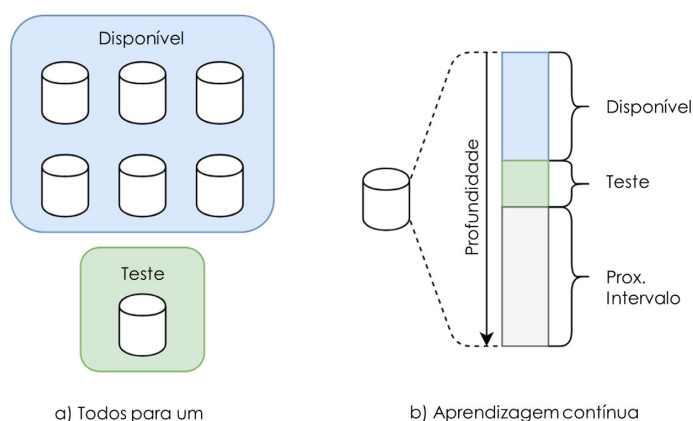
Em relação aos modelos de aprendizagem de máquina, conforme Capítulo 3, os métodos mais aplicados no problema de predição da ROP são os modelos baseados em redes neurais

e os modelos *Ensembles*. Desse modo, os modelos ANN e RF, apresentados em detalhes no Capítulo 4, são empregados neste trabalho. Adota-se a função perda quadrática no treinamento desses modelos (vide Equação 4.2). Além disso, como mencionado no Capítulo 4, modelos de regressão linear são também considerados modelos simples de aprendizagem de máquina que, apesar de possuírem menor capacidade preditiva, possuem maior capacidade de interpretabilidade e, portanto, os modelos BYM e BYMmod, que possuem maior quantidade de informação, e são considerados modelos de regressão linear, são analisados desse modo. A análise mais aprofundada desses dois modelos, BYM e BYMmod, com os dados de cada poço, permite servir como referência aos modelos de aprendizagem de máquina. Como comentado na Capítulo 3, um dos aspectos mais importantes considerado no desenvolvimento de modelos de aprendizagem de máquina aplicados ao problema de predição da ROP é a seleção dos *inputs* do modelo, desse modo, a análise da interpretabilidade do BYM e BYMmod é considerada na construção dos modelos de aprendizagem de máquina aplicados neste trabalho.

5.3 Estratégias para predição da ROP investigadas

Considerando os dois métodos para otimização de parâmetros operacionais em tempo real, indicados na Figura 15, duas estratégias propostas por Tunkiel et al. (2021) para comparação da performance de modelos na predição da ROP são apresentadas na Figura 36. Neste trabalho, essas estratégias são adotadas utilizando os dados do novo *dataset* e possibilitam comparar diferentes métodos tradicionais e de aprendizagem de máquina afim de definir modelos com maior performance na predição e melhor estratégia para aplicação nos processos de otimização.

Figura 36 – Estratégias de predição da ROP comparadas neste trabalho



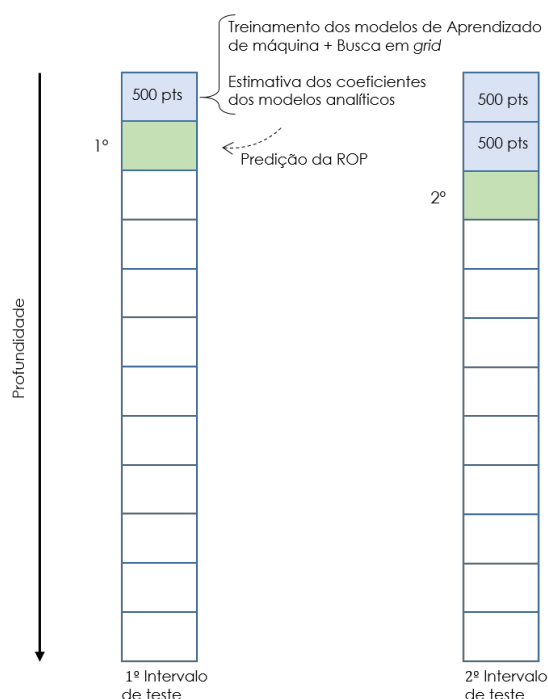
Fonte: Adaptado de Tunkiel et al. (2021)

Na Figura 36, a primeira estratégia consiste em considerar os poços de correlação no treinamento. Desse modo, a estratégia todos para um utiliza dados de 6 poços (todos os demais poços disponíveis) para treinamento dos modelos de aprendizagem de máquina, ou na estimativa dos coeficientes de modelos tradicionais, para prever a ROP em outro poço. Nessa abordagem, considera-se que os poços são do mesmo campo, que há semelhança entre eles e que a maior

quantidade de dados contribui no processo de aprendizagem dos modelos. Além disso, o fato de utilizar todos os poços é uma medida que facilita a aplicação do método, uma vez que não requer critérios de seleção de dados desses poços. Na segunda estratégia, são utilizados os dados do próprio poço, onde o modelo é treinado inicialmente com poucos dados (no início do poço) e as previsões são feitas em um intervalo seguinte. Esse processo é continuado ao longo da profundidade usando cada vez mais dados e fazendo previsões em vários intervalos. Nessa abordagem, apesar do processo de aprendizagem ser feito com os dados do próprio poço, a menor quantidade de dados aliado a possíveis alterações significativas ao longo da profundidade, como no caso de diferentes litologias, podem reduzir a capacidade preditiva nos intervalos analisados.

Todos os poços da Tabela 10 são avaliados nas duas estratégias da Figura 36. Desse modo, com uma maior quantidade de poços é possível compreender mais detalhadamente as limitações de cada estratégia. As métricas de erros, dadas na Equação 4.27 e Equação 4.28, são utilizadas para avaliar os erros na predição dentro dos dados de teste indicados na Figura 36. Além disso, as duas estratégias são comparadas em termos de precisão alcançada pelos modelos. Nos modelos de aprendizado de máquina, o método de busca em *grid* é utilizado no treinamento dos modelos, conforme descrito no Capítulo 4, para obter hiperparâmetros adequados no treinamento. O pré-processamento adotado para os modelos ANNs é o escalonamento min-max com intervalo de 0 a 1, que é considerado apropriado tendo em vista a distribuição dos dados do novo *dataset*, que não seguem padrão normal (Apêndice D) e os principais atributos são valores positivos (Figura 35). Para a estratégia de aprendizado contínuo, é adotada a divisão de intervalos iguais de 500 pontos. A Figura 37 ilustra o processo em mais detalhes.

Figura 37 – Detalhe da estratégia de predição da ROP por aprendizagem contínua



Fonte: Autor (2024)

O valor de 500 pontos é considerado suficiente para treinar os modelos de aprendizado de máquina e executar a busca em *grid*. Considerando ainda a quantidade de pontos em cada poço, conforme apresentado na Tabela 10, utilizar intervalos de 500 pontos resulta em quantidade similar de pontos no último intervalo para a maioria dos poços. Conforme pode ser conferido na Tabela 14.

Tabela 14 – Informações sobre os intervalos da aprendizagem contínua para cada poço

Poços	Intervalo médio (m)	Nº de intervalos	Pontos no último intervalo
F05	78,68	10	459
F14	64,86	10	457
F15S	43,82	32	212
F04	83,35	8	474
F11_T2	54,10	16	307
F10	50,76	9	318
F12	53,23	18	474

Fonte: Autor (2024)

5.4 Recursos empregados

Neste trabalho é utilizada a linguagem de programação *Python* para implementação das análises, manipulação dos dados e em todos aspectos relacionados à construção dos modelos baseados em dados. Algumas bibliotecas em *Python* para extração e manipulação de dados são utilizadas, como é o caso do *pandas*, e bibliotecas específicas para tipos de arquivo, como o *dlisio* e *lasio*, para extrair informações dos arquivos em formato *.dlis*, *.lis* e *.las*. Além disso, a solução de modelos estatísticos de regressão tradicionais, com os testes e métricas estatísticas relacionadas, são obtidos com suporte de outra biblioteca: *statsmodels*. Para determinação de coeficientes do modelo MSE, utiliza-se o algoritmo *Nelder-Mead* da biblioteca *Scipy*. Recursos dentro da biblioteca *scikit-learn* são empregados na construção dos modelos de aprendizagem de máquina citados neste trabalho. Os resultados deste trabalho são alcançados utilizando Notebook com processador Core i7 – 7500U de 2.7GHz com Turbo Boost de até 3.5GHz e 8GB DDR4.

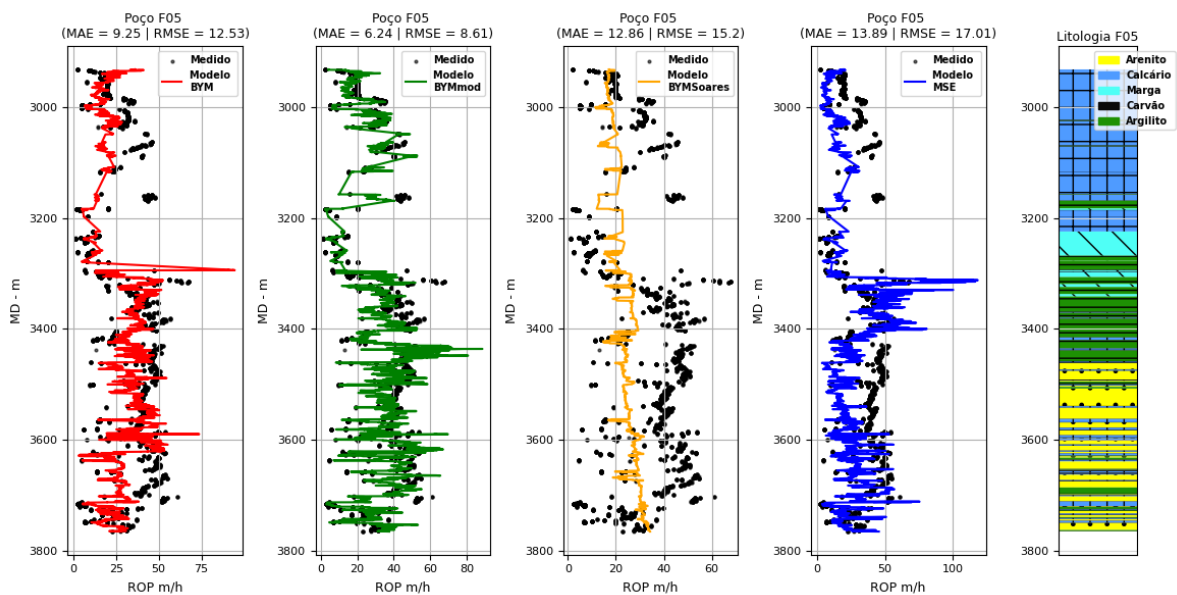
6 RESULTADOS E DISCUSSÕES

Neste capítulo são apresentados os resultados e discussões acerca das análises propostas neste trabalho. Inicialmente, resultados acerca da modelagem da ROP com as duas técnicas são apresentados: Com métodos tradicionais e com métodos de aprendizagem de máquina. Os principais aspectos relacionados a cada estratégia são discutidos, e a análise dos resultados de cada modelo para todos os 7 poços investigados é apresentada. Além disso, analisam-se em detalhes as particularidades de cada poço. Em seguida, os resultados da predição da ROP usando as duas abordagens que viabilizam a otimização em tempo real são apresentados e comparados: todos por um e aprendizagem contínua, conforme descritas no Capítulo 5.

6.1 Modelagem do ROP: Métodos tradicionais

A seguir são apresentados os resultados dos modelos tradicionais mencionados no Capítulo 3 e também referenciados no Capítulo 5. Os coeficientes do modelo são estimados com os dados do próprio poço, de onde é possível avaliar, inicialmente, nas Figuras 38, 39, 40, 41, 42, 43 e 44, a qualidade do ajuste e, portanto, a representatividade de cada modelo em capturar o padrão da ROP de cada poço (junto com sua litologia). Como pode ser visto no Apêndice B, os dados de alguns poços analisados pertencem a fases diferentes, em certos casos empregando brocas de tipos e tamanhos diferentes. É possível perceber que, de modo geral, a resposta dos modelos tradicionais apresentados indicam que o conjunto de coeficientes estimados para o poço todo são suficientes para representar adequadamente o padrão da ROP e que os dados utilizados são representativos dos fenômenos que levaram a ROP resultante em cada poço.

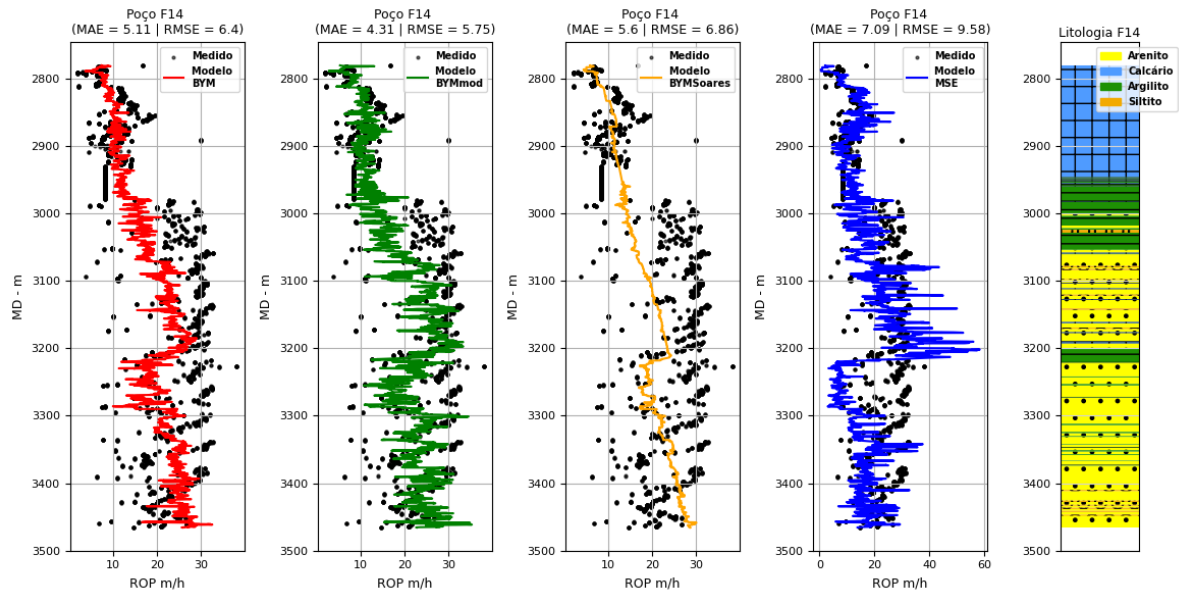
Figura 38 – Resultado da modelagem da ROP com 4 modelos tradicionais para o poço F05



Fonte: Autor (2024)

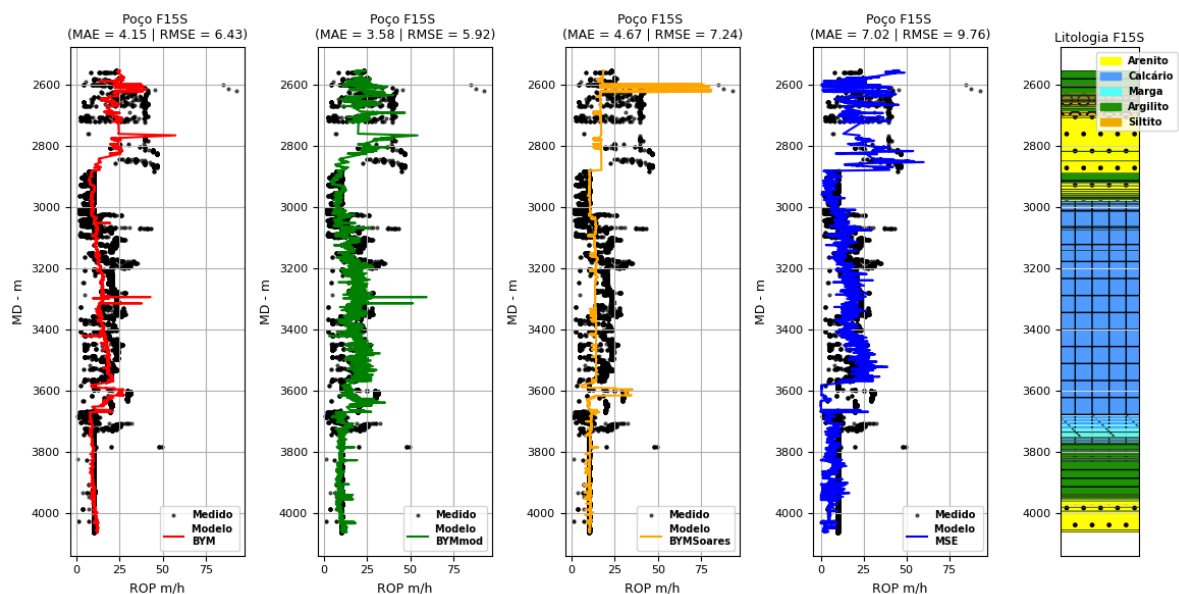
Os coeficientes de cada modelos são apresentados nas Tabelas 15, 16, 17 e 18 para os modelos BYM, BYMmod, BYMSoares e MSE, respectivamente. Para os modelos de regressão BYM, BYMmod e BYMSoares a relação não linear entre a ROP e os parâmetros de perfuração é transformada em uma relação linear, assim como o modelo linear apresentado na Equação 4.6, quando se utiliza transformação logarítmica na variável resposta e nos *inputs* do modelo.

Figura 39 – Resultado da modelagem da ROP com 4 modelos tradicionais para o poço F14



Fonte: Autor (2024)

Figura 40 – Resultado da modelagem da ROP com 4 modelos tradicionais para o poço F15S

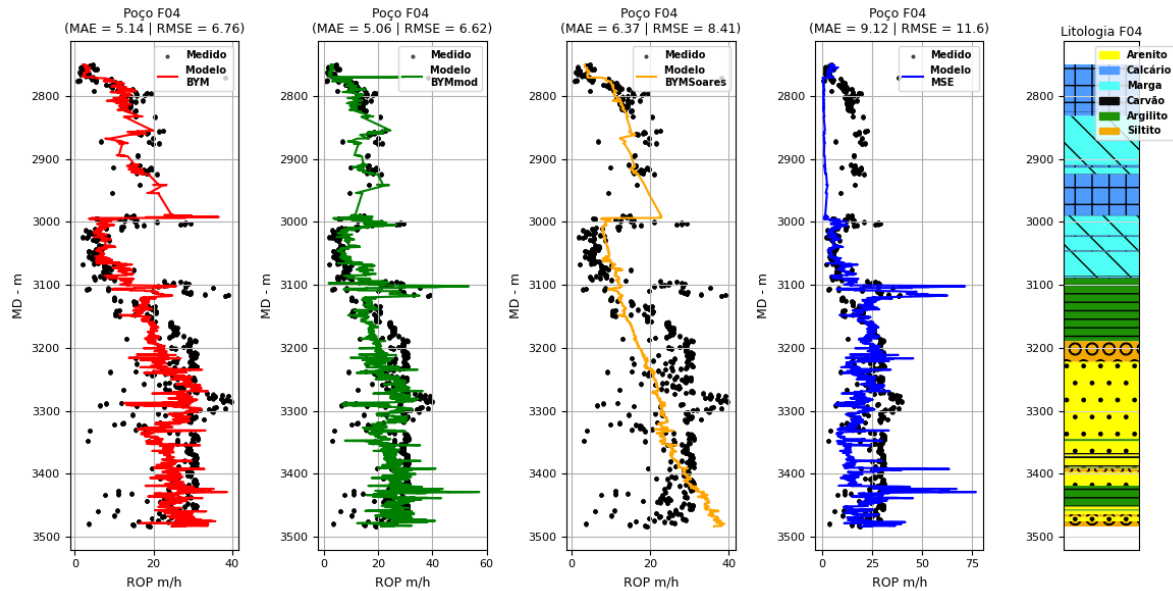


Fonte: Autor (2024)

Com a transformação em modelos de regressão linear múltipla, os coeficientes podem

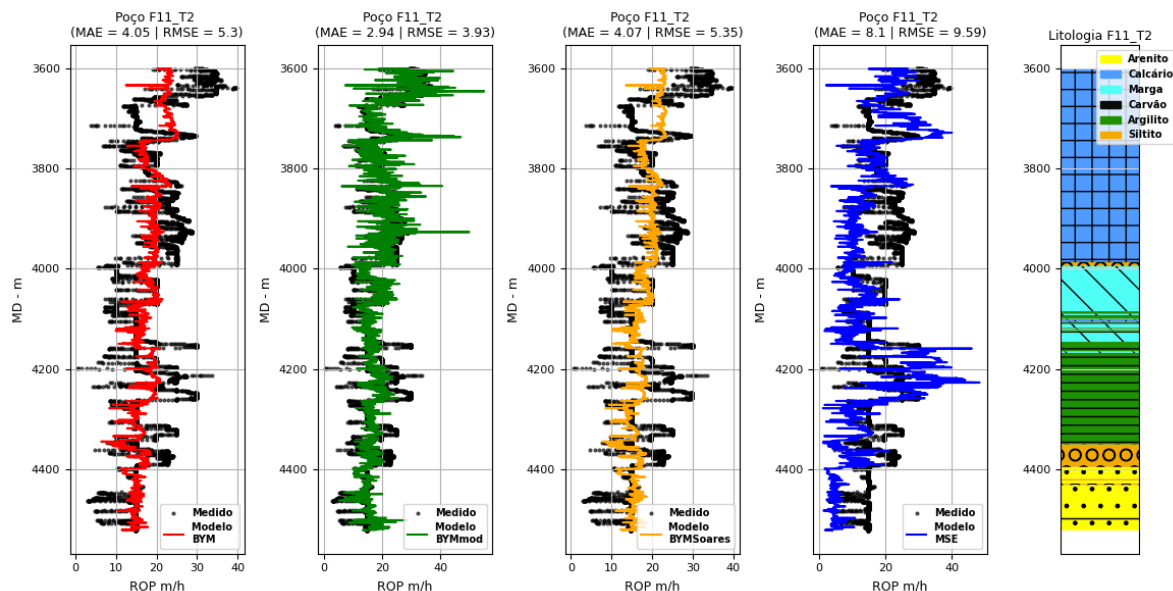
ser obtidos por meio da técnica de mínimos quadrados ordinários conforme apresentado na Equação 4.5 e, portanto, a interpretação do modelo pode ser feita conforme mencionado no Capítulo 4.

Figura 41 – Resultado da modelagem da ROP com 4 modelos tradicionais para o poço F04



Fonte: Autor (2024)

Figura 42 – Resultado da modelagem da ROP com 4 modelos tradicionais para o poço F11_T2

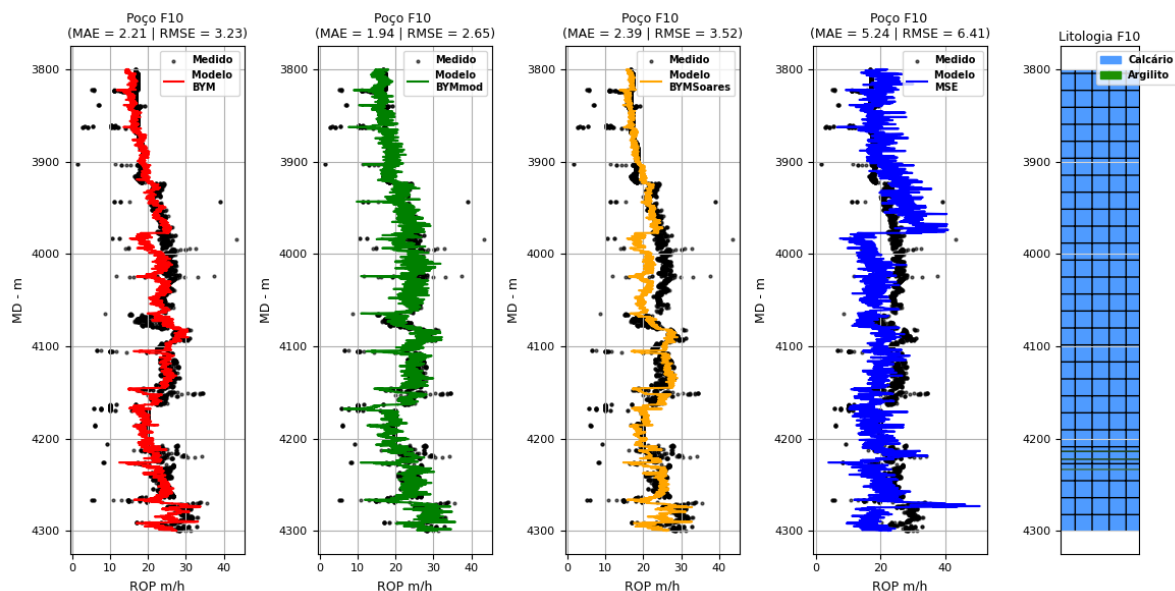


Fonte: Autor (2024)

Para o caso do modelo MSE, os coeficientes obtidos na Tabela 18 são determinados utilizando o método *Nelder-Mead* para resolver o problema de otimização que envolve a minimização do erro quadrático entre a medida da ROP de cada poço e a Equação 3.23. A minimização

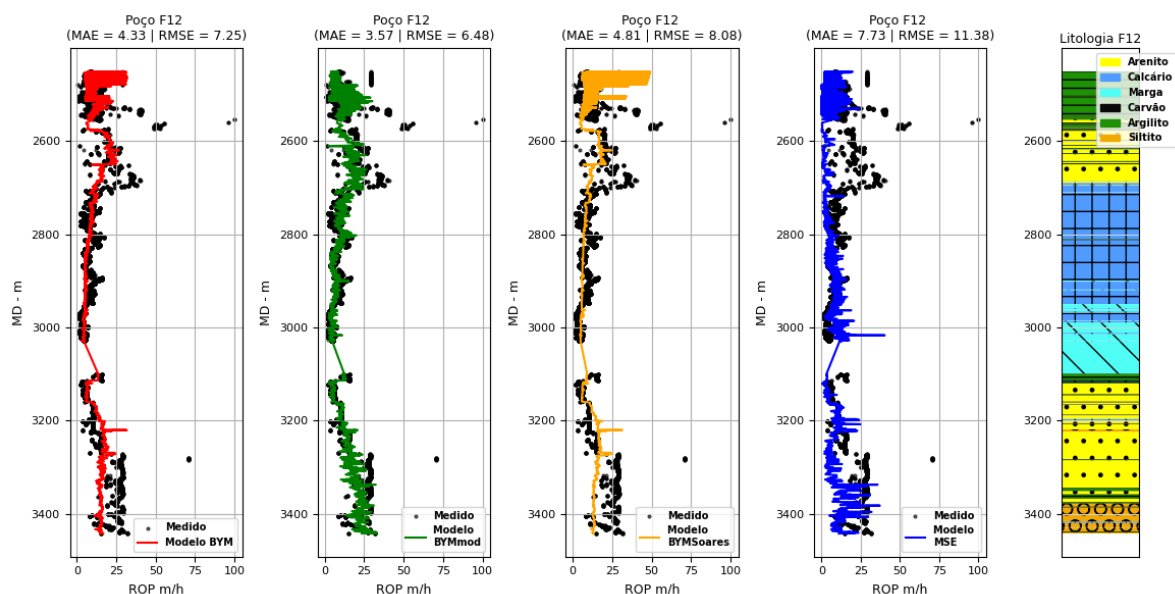
dessa função perda quadrática é feita utilizando os recursos mencionados no Capítulo 5 e com chute inicial: $\mu_b = 0,5$ e $\mu = 0,5$. Para todos os poços o processo de minimização é satisfatório uma vez que o método de otimização converge para todos os casos. Para utilização do modelo MSE, é necessária a determinação da resistência da formação (conforme a Tabela 13).

Figura 43 – Resultado da modelagem da ROP com 4 modelos tradicionais para o poço F10



Fonte: Autor (2024)

Figura 44 – Resultado da modelagem da ROP com 4 modelos tradicionais para o poço F12



Fonte: Autor (2024)

As informações das litologias que são extraídas para cada poço são utilizadas em conjunto com as correlações empíricas descritas no Apêndice C. As correlações são específicas para o tipo

de litologia e são utilizadas para determinar a resistência não confinada da rocha (UCS) usando a medida de Delta-T da formação.

Tabela 15 – Coeficientes do modelo BYM para cada poço

Poços	BYM							
	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
F05	2,79	$-8,85 \cdot 10^{-4}$	$-1,67 \cdot 10^{-4}$	$1,70 \cdot 10^{-5}$	0,75	2,44	9,02	0,21
F14	2,21	$-4,36 \cdot 10^{-4}$	$-8,70 \cdot 10^{-5}$	$1,00 \cdot 10^{-5}$	0,21	0,66	-2,35	$2,96 \cdot 10^{-3}$
F15S	0,29	$2,31 \cdot 10^{-4}$	$7,43 \cdot 10^{-4}$	$-4,30 \cdot 10^{-5}$	0,03	-0,10	-5,20	0,75
F04	1,10	$-4,55 \cdot 10^{-4}$	$6,05 \cdot 10^{-4}$	$-2,90 \cdot 10^{-5}$	0,48	-0,85	0,22	0,44
F11 T2	1,83	$7,63 \cdot 10^{-5}$	$1,98 \cdot 10^{-5}$	$5,86 \cdot 10^{-7}$	0,27	0,48	0,00	0,49
F10	56,95	$-4,27 \cdot 10^{-2}$	$-8,50 \cdot 10^{-4}$	$5,30 \cdot 10^{-5}$	0,37	-0,63	226,00	0,37
F12	2,04	$2,20 \cdot 10^{-5}$	$-3,48 \cdot 10^{-4}$	$2,30 \cdot 10^{-5}$	-0,13	1,15	1,74	-1,27

Fonte: Autor (2024)

Tabela 16 – Coeficientes do modelo BYMmod para cada poço

Poços	BYMmod							
	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8
F05	1,69	$2,53 \cdot 10^{-4}$	$3,09 \cdot 10^{-4}$	$-1,30 \cdot 10^{-5}$	0,24	5,96	3,41	0,81
F14	2,32	$3,31 \cdot 10^{-4}$	$1,16 \cdot 10^{-3}$	$-6,30 \cdot 10^{-5}$	0,09	1,70	-2,35	-0,22
F15S	1,22	$3,69 \cdot 10^{-4}$	$8,45 \cdot 10^{-4}$	$-4,80 \cdot 10^{-5}$	-0,09	2,73	-4,89	0,07
F04	3,86	$-1,62 \cdot 10^{-4}$	$3,20 \cdot 10^{-5}$	$7,00 \cdot 10^{-6}$	0,61	2,55	4,54	0,10
F11 T2	1,83	$2,59 \cdot 10^{-4}$	$3,38 \cdot 10^{-4}$	$-1,60 \cdot 10^{-5}$	0,05	4,39	0,00	0,17
F10	20,37	$-1,48 \cdot 10^{-2}$	$-2,90 \cdot 10^{-4}$	$1,80 \cdot 10^{-5}$	0,16	2,29	78,87	0,14
F12	3,02	$-3,98 \cdot 10^{-4}$	$-8,52 \cdot 10^{-4}$	$5,70 \cdot 10^{-5}$	-0,06	3,08	1,86	-1,01

Fonte: Autor (2024)

Tabela 17 – Coeficientes do modelo BYMSoares para cada poço

Poços	BYMSoares				
	a_1	a_2	a_5	a_6	a_8
F05	-52,89	3,36	0,24	$7,68 \cdot 10^{-2}$	4,02
F14	-114,70	13,10	0,19	0,26	-0,50
F15S	-12,03	0,23	$2,95 \cdot 10^{-2}$	-0,50	2,61
F04	-122,70	13,84	0,16	-0,68	0,44
F11 T2	5,64	-0,95	0,24	0,58	0,56
F10	-223,10	25,45	0,31	-0,90	$-3,36 \cdot 10^{-2}$
F12	58,55	-4,75	-0,16	1,26	-2,65

Fonte: Autor (2024)

Para o modelo BYM, conforme descrito no Capítulo 3, informações como o gradiente de poro pressão e o desgaste da broca são estimadas seguindo métodos próprios conforme apresentado em mais detalhes no Apêndice A e Apêndice B. O parâmetro $\left(\frac{WOB}{D_b}\right)_t$ é adotado

no modelo como igual a zero, uma vez que se verificou melhor aderência da resposta do modelo com a medida da ROP para todos os poços.

Tabela 18 – Coeficientes do modelo MSE para cada poço

Poços	MSE	
	a_1	a_2
F05	$3,86 \cdot 10^{-2}$	-2,79
F14	$3,63 \cdot 10^{-1}$	$3,16 \cdot 10^{-1}$
F15S	$1,24 \cdot 10^{-1}$	$-4,10 \cdot 10^{-1}$
F04	$5,44 \cdot 10^{-2}$	-1,94
F11_T2	$1,33 \cdot 10^{-1}$	$-8,50 \cdot 10^{-1}$
F10	$5,70 \cdot 10^{-1}$	$2,85 \cdot 10^{-2}$
F12	$1,68 \cdot 10^{-2}$	-3,00

Fonte: Autor (2024)

Em relação ao modelo BYMmod, adota-se $C_b = 1$, pois não há utilização de alargamento nos poços analisados. Assim como recomendado por Darwesh et al. (2020), a expressão empírica do *Pump-Off Force* na Equação 3.12 é alterada de modo que o coeficiente 0,942 é reduzido. Neste caso, adota-se um valor pequeno de $2 \cdot 10^{-5}$, uma vez que se identifica melhoria na resposta dos modelos para todos os poços. Na substituição da f_6 , conforme mencionado no Capítulo 3, é incluso o torque na superfície substituindo o RPM. A normalização adicionada na f_6 é o torque médio das medições em cada poço.

Nas Figuras 38, 39, 40, 41, 42, 43 e 44, é também apresentadas as métricas de erro calculadas com base na resposta de cada modelo seguindo as Equação 4.27 e Equação 4.28. Essas métricas são comparadas mais detalhadamente na Figura 47 e servem de referência de qualidade do ajuste para cada modelo. É possível observar que modelo BYMmod aparenta capturar melhor as medidas da ROP para todos os poços (menores erros). Este modelo adiciona novas informações, como trajetória e o torque, ao modelo BYM, além de modificar normalizações propostas inicialmente para se adequar a outros tipos de brocas. Como pode ser analisado na Figura 38, para o poço F05, essas novas informações permitiram capturar melhor a resposta em certos trechos quando comparado com o modelo BYM, como é o caso do trecho de 3000 m até 3200 m, ou do trecho de 2800 m até 3000 m para o caso do poço F04 (Figura 41).

Em relação ao modelo BYMSoares, quando comparado com a resposta dos demais modelos, aparenta ser simplificado ao não capturar o padrão da ROP em certos poços, apenas a tendência de aumento e diminuição da ROP. Essa observação pode ser feita novamente para o caso do poço F04 (Figura 41). No entanto, em poços como o F10 (Figura 43), o modelo é capaz de se ajustar adequadamente os dados de ROP, de modo que as métricas de erro entre os modelos BYM e BYMSoares são próximas. Esse caso é um indicativo que apenas as quatro variáveis não são suficientes para capturar o padrão da ROP na maioria dos poços, como pode ser visto mais claramente também na Figura 38. No entanto, para o poço F10, os efeitos principais que resultam

na ROP são explicáveis, em grande parte, com as variáveis incluídas no modelo BYMSoares, que são WOB, RPM, vazão do fluido e o TVD.

É possível perceber que, no caso do poço F05, as diferenças maiores identificadas no modelo BYM, e que são reduzidas com o modelo BYMmod, são também identificadas na resposta do modelo MSE. Essas diferenças são perceptíveis no trecho de 3400 m até 3600 m na resposta do modelo MSE. Para o poço F14, também é possível identificar trecho onde a resposta do MSE apresenta maior diferença com a ROP medida, sendo ainda mais perceptível no caso da Figura 41 para o caso do trecho de 2800 m até 3000 m. Em ambos os casos, essas diferenças são perceptíveis em maior e menor grau na mesma região para o modelo BYM, e são reduzidas na resposta do modelo BYMmod, o que indica que as informações adicionadas ao modelo contribuem para capturar melhor o padrão da ROP nesses trechos.

A diferença observada no poço F04 para o caso do modelo MSE é a maior dentre todos os casos mencionados. Como é possível perceber no Apêndice D, o trecho é marcado por mudanças bruscas no RPM acompanhado por redução significativa do WOB no trecho próximo de 3000 m. Essas alterações bruscas também podem ser identificadas nas outras regiões comentadas anteriormente onde o modelo MSE apresenta maior diferença, notadamente nos poços F05, F04 e F14. Pode-se relacionar essas diferenças com quedas significativas do WOB mantendo valores elevados de RPM. Esse comportamento observado nos parâmetros operacionais em conjunto com as suposições feitas no modelo MSE, que envolve uma relação entre a energia normalmente empregada no processo de perfuração e a ROP. Pode-se inferir que esses trechos estão relacionados a processos ineficientes na perfuração. A análise dos dados e a resposta dos modelos são corroborados por informações retiradas dos relatórios de cada poço sobre esses trechos, conforme apresentado a seguir:

"Após 3000 m MD os valores de *stick-slip* medidos pela ferramenta *TeleScope* aumentaram, acima das especificações da ferramenta. Tentamos mitigar o *stick-slip* aumentando o RPM da superfície, e aliviando o peso sobre a broca. Também foi tentado parar a rotação e retirar a coluna do fundo, liberando o torque na coluna, mas sem sucesso." [tradução nossa do documento: 15-9-F-5_nr_002_MWD_DD_EOWR, p.09 do poço F05] (EQUINOR, 2018)

"Uma vez que a seção de ganho de ângulo iniciou com 50% de steering ratio para 70% para alcançar +4% DLS, a resposta melhora uma vez que entra no grupo geológico Cromer Knoll. Como esperado, os choques e o *stick-slip* também aumentaram e tivemos que ajustar os parâmetros de perfuração frequentemente para mitigá-los. O ECD aumentou gradualmente e foi necessário reduzir a vazão do fluido, já que o RPM foi aumentado para 180 devido ao *stick-slip*." [tradução nossa do documento: 15-9-f-14_nr_002_MWD_DD_EOWR. p.20 do poço F14] (EQUINOR, 2018)

"Retomando a perfuração em 2993 m. A ROP foi muito baixa comparada com fases anteriores e ocorreu *stick-slip* severo. Mitigar os choques e *stick-slip* fez a ROP cair ainda mais, de 14m/h para 4-6 m/h." [tradução nossa do documento: 15-9-f-4_nr_003_MWD_DD_EOWR. p. 23 do poço F04] (EQUINOR, 2018)

Como é possível perceber, nesses trechos há identificação de problemas na perfuração relacionados ao *stick-slip*, que é o fenômeno de vibração torcional. Como mencionado em Andås (2017), esse fenômeno ocorre por conta da fricção entre a broca e a formação. Nesse caso, o coeficiente de atrito estático da broca é maior que o coeficiente de atrito cinético inicial, de modo que a broca adere mais a formação. Neste caso, fatores como as propriedades da formação perfurada e características da broca são influentes. Esse fenômeno é predominante em brocas do tipo PDC (tipo de broca predominante nos poços investigados) e, como mencionado ainda em Andås (2017), flutuações nas medidas de torque na superfície podem indicar esse fenômeno.

Portanto, nos casos dos trechos com diferenças maiores mencionadas, uma das informações relevantes para a redução dessas diferenças, quando é comparado o modelo BYM e o BYMmod, é o torque, que nesse caso pode ter contribuído para melhor estimar a ROP devido ao fenômeno de vibração torcional nesses trechos. Outra diferença relevante nos poços analisados pode ser identificada para o caso do poço F12 (Figura 44). No trecho entre 2500 m e 2600 m, picos da ROP não são precisamente capturados por nenhum modelo, e maiores diferenças podem ser observadas no caso do modelo MSE. Como pode ser analisado no Apêndice B, esse é o único trecho com outro tipo de broca, RCB de dentes de aço em vez de PDC. Portanto, o conjunto único de coeficientes apresentado nas Tabela 15, 16, 17 e 18 não são representativos nesse trecho em particular, tendo em vista que os demais dados usados na determinação desses coeficientes pertencem à brocas PDC, sendo o ideal obter coeficientes para trechos de brocas distintas. No entanto, nota-se que o BYMmod ainda captura bem o padrão da ROP de modo geral.

Como indicado na Capítulo 4, um fator importante para melhor interpretar modelos de regressão linear diz respeito ao problema de multicolinearidade em que os preditores do modelo são correlacionados, de modo que adicionam informações redundantes. Quando é identificada, a multicolinearidade afeta a interpretabilidade do modelo. Os valores de FIV (Equação 4.30), para os modelos BYM e BYMmod são apresentados na Tabela 19 e Tabela 20, respectivamente.

Tabela 19 – FIV calculado para o modelo BYM em cada um 7 poços

Variáveis BYM	F05	F14	F15S	F04	F11_T2	F10	F12
X1	-	-	-	-	-	-	-
X2	22,97	16,04	2,22	14,10	6,90	3499,08	4,12
X3	85,80	291,82	482,43	241,15	225,94	1655,41	158,99
X4	121,75	259,75	470,36	271,58	187,74	1665,71	160,95
X5	1,99	1,28	2,04	1,400	1,85	1,780	2,07
X6	8,30	6,64	3,24	9,77	1,15	1,77	1,87
X7	2,40	7,57	1,79	3,18	-	3484,37	2,07
X8	1,48	5,28	2,86	1,53	1,47	5,87	3,44

Fonte: Autor (2024)

Tabela 20 – FIV calculado para o modelo BYMmod em cada um 7 poços

Variáveis BYMmod	F05	F14	F15S	F04	F11_T2	F10	F12
X1	-	-	-	-	-	-	-
X2	13,17	18,60	3,70	11,16	4,92	3686,21	3,90
X3	91,26	264,21	274,28	142,79	276,02	1644,98	155,14
X4	124,24	236,90	281,92	155,63	245,23	1665,24	161,67
X5	1,20	3,24	2,78	1,68	6,01	1,68	2,03
X6	1,84	3,34	3,40	3,17	1,64	1,73	3,77
X7	1,30	10,47	2,08	1,23	-	3659,17	2,25
X8	1,28	6,42	2,19	1,36	1,95	5,66	2,79

Fonte: Autor (2024)

Os modelos BYM e BYMmod, como observado nos exemplos discutidos, estimam adequadamente o padrão da ROP nos 7 poços quando comparados com os demais modelos. Além disso, como relacionam a ROP com uma série de informações, são modelos de interesse para uma análise de interpretabilidade estatística, possível com modelos de regressão linear. No entanto, levando em conta a recomendação de que o FIV deve ficar abaixo de 5, a Tabela 19 e Tabela 20 indentificam algumas variáveis como correlacionadas. Uma solução possível é a remoção das variáveis que causam essa multicolinearidade (MONTGOMERY; RUNGER, 2016). A Tabela 21 e Tabela 22 apresentam os valores do FIV para o modelo sem as variáveis que causam a multicolinearidade.

Tabela 21 – FIV calculado para o modelo BYM sem problema de multicolinearidade em cada um 7 poços

Variáveis BYM	F05	F14	F15S	F04	F11_T2	F10	F12
X1	-	-	-	-	-	-	-
X2	2,07	2,73	1,56	3,70	1,88	-	2,54
X3	-	-	-	-	-	-	-
X4	1,72	1,080	2,34	1,59	2,05	1,27	1,42
X5	1,02	1,15	2,00	1,39	1,85	1,74	2,02
X6	-	3,90	1,96	4,76	1,10	1,54	1,85
X7	1,18	-	1,72	2,15	-	-	1,81
X8	1,47	2,01	2,57	1,52	1,44	1,97	3,07

Fonte: Autor (2024)

Tabela 22 – FIV calculado para o modelo BYMmod sem problema de multicolinearidade em cada um 7 poços

Variáveis BYMmod	F05	F14	F15S	F04	F11_T2	F10	F12
X1	-	-	-	-	-	-	-
X2	2,882	1,715	3,696	2,638	4,579	-	3,530
X3	-	-	-	-	-	-	-
X4	1,709	1,080	2,709	1,689	2,283	1,273	2,179
X5	1,175	2,554	2,772	1,620	4,487	1,641	2,020
X6	1,75	1,890	3,406	2,999	1,632	1,376	3,478
X7	1,222	-	2,045	1,213	-	-	2,113
X8	1,244	1,176	2,167	1,212	1,635	1,545	2,782

Fonte: Autor (2024)

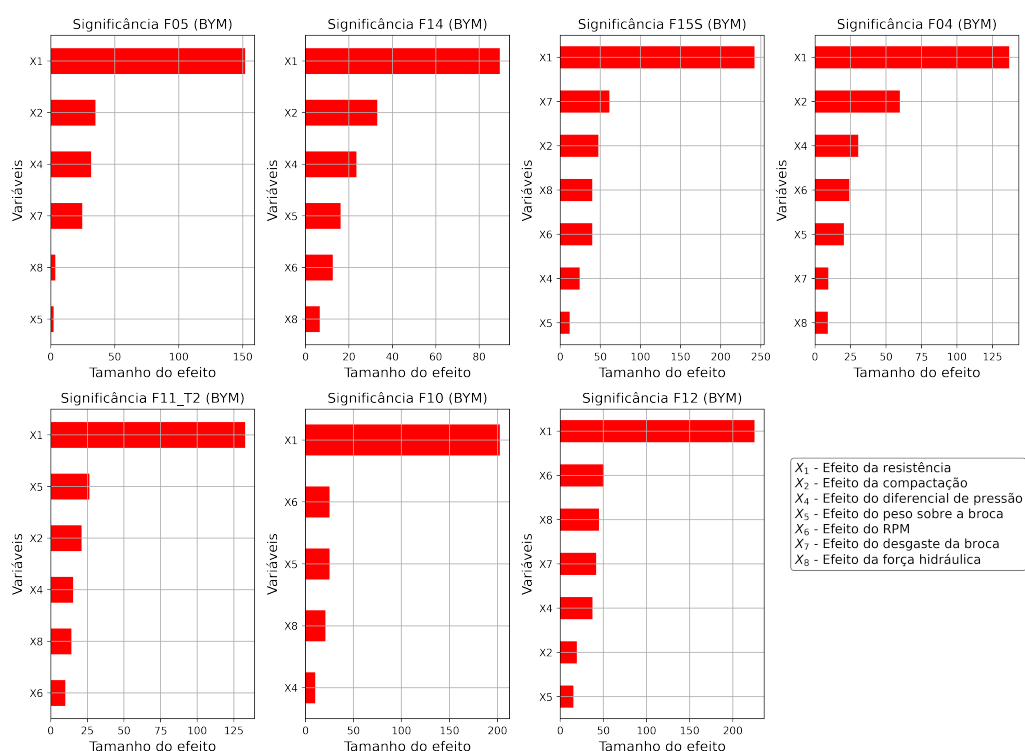
A abordagem empregada é remover incrementalmente as variáveis com maior FIV e recalculá-lo novamente com base em um modelo de regressão sem as variáveis removidas. Como é possível perceber na Tabela 19 e Tabela 20, os maiores valores estão associados às características de formação, conforme detalhado no Capítulo 3, sendo que todas essas variáveis explicativas, em conjunto, são informações redundantes. Quando os preditores X3 e X4 possuem valores de FIV muito elevados, opta-se por remover primeiro o preditor X3, uma vez que o X4 agrega mais informações da perfuração (três medições). Ao final desse processo, é determinado se ainda há multicolinearidade ou não. Caso não exista, o processo é interrompido. Como pode ser analisado, na Tabela 21 e Tabela 22, as variáveis remanescentes possuem FIV menor que 5.

Com as variáveis indicadas nos modelos sem multicolinearidade, pode-se utilizar a Equação 4.29 para determinar a importância dos efeitos e, desse modo, analisar a significância das variáveis explicativas, conforme é apresentado na Figura 45 e Figura 46. É possível perceber que, na Figura 45, a importância maior é atribuída ao X1 em todos os modelos. Como comentado no Capítulo 3, a função f_1 é conhecida como função de resistência da formação, apesar de que a medida de resistência não é diretamente considerada no modelo, como é no modelo MSE. Essa definição pode ser interpretada pela consideração de que o maior efeito na ROP, não considerado diretamente no modelo, seria a resistência. Desse modo, os resultados da Figura 45 parecem indicar que existem mais efeitos que não são considerados no modelo BYM tradicional. É necessário ressaltar que o modelo não foi originalmente desenvolvido para representar o comportamento de brocas do tipo PDC, que são mais suscetíveis a certos fenômenos, como o próprio *stick-slip*, do que brocas RCB.

Em relação ao BYMmod, na Figura 46, é possível perceber que há maior equivalência entre a significância das variáveis explicativas do modelo para cada poço, quando comparado com os resultados do BYM. É necessário ressaltar também que os efeitos significativos consideram a

regressão do modelo para os dados do poço todo. Em trechos menores, outras variáveis seriam significativas. Esse motivo pode explicar, por exemplo, a menor relevância da variável relacionada ao efeito do WOB na ROP, por exemplo. Como comentado no Capítulo 4, é possível comparar o tamanho dos efeitos na Figura 45 e Figura 46 com valores críticos e, portanto, verificar se todos os efeitos são significativos de fato. Utilizando os recursos do *statsmodels*, é possível obter o valor-p, que ao ser comparado com nível de significância de 5% conclui-se que há evidência estatística para considerar a variável X8 no poço F15S como não significativa. Desse modo, todas os efeitos considerados nos demais poços na Figura 45 e Figura 46 são relevantes e não correlacionados.

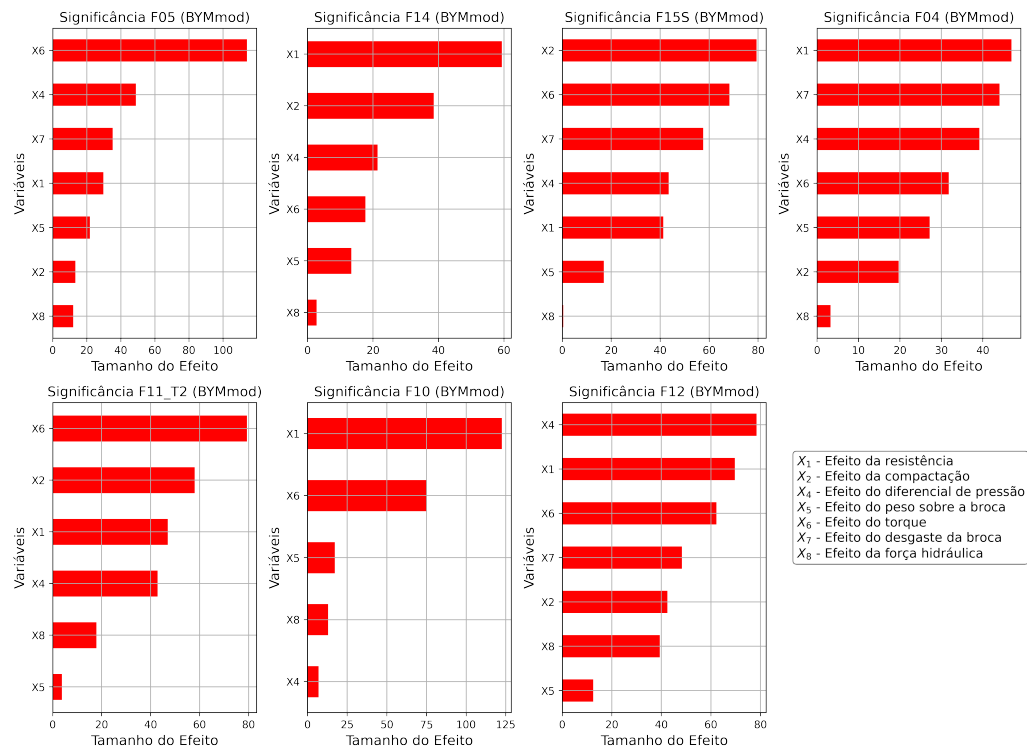
Figura 45 – Significância estatística dos preditores do modelo BYM sem multicolinearidade



Fonte: Autor (2024)

Pode-se perceber que, de modo geral, para o poço F05 (Figura 46) o efeito mais significativo é relacionado ao torque, assim como para o poço F11_T2. Este resultado corrobora com as observações anteriores que, devido ao fenômeno de *stick-slip*, a adição da informação do torque contribuiu para a redução das diferenças entre a ROP medida e a resposta do modelo nos trechos comentados para o poço F05. É possível perceber também que em poços como F11_T2 e F15S e F14 a variável X2, relacionada ao efeito de compactação da formação na ROP, é uma das mais significativas. Observa-se também que o poço F10 é o que atribui menor importância para os efeitos da formação (único modelo sem a variável X2). Tendo em vista a análise do resultado do modelo BYMSoares, o padrão da ROP é estimado razoavelmente com um modelo mais simples. O poço F12 e F04 são os que possuem maior quantidade de efeitos em nível de relevância equivalente, como, por exemplo, percebe-se uma maior relevância do desgaste.

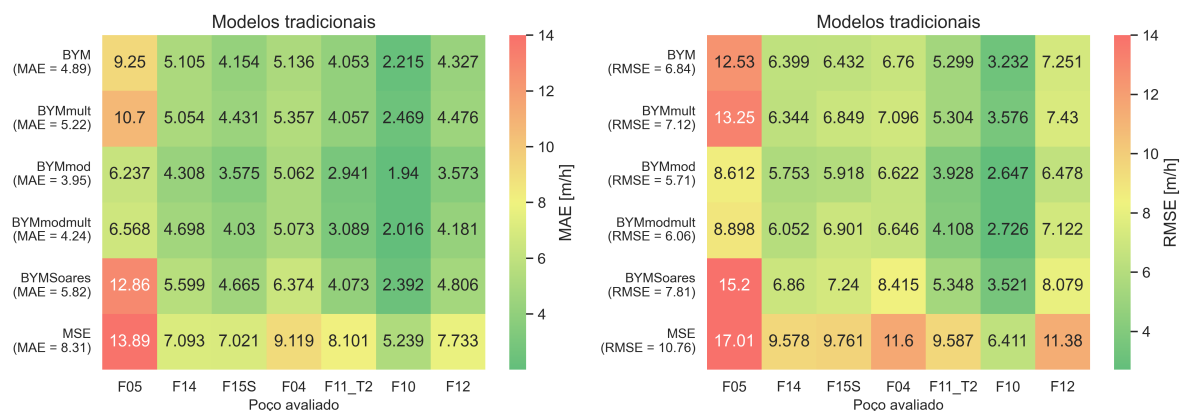
Figura 46 – Significância estatística dos preditores do modelo BYMmod sem multicolinearidade



Fonte: Autor (2024)

Na Figura 47, é apresentado o cálculo das métricas de erro para cada poço e modelo utilizado. Percebe-se que os modelos sem multicolinearidade (BYMmult e BYMmodmult) possuem erros muito próximos dos modelos originais, isso indica que a resposta do modelo não é significativamente alterada com as remoções das variáveis que causavam a multicolinearidade nos modelos. No geral, o poço F05 é o que apresenta erros maiores e o F10 os menores.

Figura 47 – Heatmap com as métricas de erro em m/h dos modelos tradicionais para cada poço



Fonte: Autor (2024)

6.2 Modelagem do ROP: Aprendizagem de máquina

Tendo em vista a análise dos modelos tradicionais, percebe-se que há variáveis relacionadas a características da formação com grande relevância (X2 e X4, conforme Figura 46). Além disso, certos parâmetros operacionais, como o torque, possuem relevância significativa em certos poços. Ao mesmo tempo, observa-se nos Apêndices A e C que informações como gradiente de poro pressão, presente na variável X4, e o UCS, presente no modelo MSE, são informações possíveis de estimar por meio da medição do Delta-T (Capítulo 2). Para a aplicação dos modelos de aprendizagem de máquina neste trabalho, são extraídos os resultados considerando diferentes preditores, tendo em vista determinar um conjunto de preditores mais adequados e suficientes para aprender o padrão da ROP e melhorar a performance dos modelos de aprendizagem de máquina. Os preditores investigados são apresentados na Tabela 23.

Tabela 23 – Conjuntos de preditores utilizados nos modelos de aprendizagem de máquina

Preditores	Conjunto de parâmetros
1	WOB, RPM, Vazão do fluido
2	WOB, RPM, Vazão do fluido, Delta-T
3	WOB, Torque, Vazão do fluido, Delta-T
4	WOB, RPM, Torque, Vazão do fluido, Delta-T

Fonte: Autor (2024)

Os parâmetros do conjunto Preditores 1, são os mais comuns em modelos de aprendizagem de máquina para predição da ROP, conforme mencionado em Barbosa et al. (2019), empregado, por exemplo, no trabalho de Elmergi et al. (2021). Analisando os demais conjuntos de preditores, pode-se avaliar, nos resultados, a influência do Delta-T e a influência na mudança do RPM pelo torque, conforme a discussão feita com os resultados dos modelos tradicionais.

Nos modelos de aprendizagem de máquina construídos neste trabalho, utiliza-se a perda quadrática (Equação 4.2) no treinamento, que também foi aplicada na obtenção dos coeficientes do modelo MSE. Como comentado no Capítulo 5, emprega-se a busca em *grid* para estimar os hiperparâmetros adequados. O método utilizado é *k-fold* com 5 partições. A Tabela 24 apresenta os hiperparâmetros comparados no processo de treinamento do modelo RF. O nome dos hiperparâmetros são aqueles designados no *scikit-learn*.

Tabela 24 – *Grid* de hiperparâmetros do modelo RF

Hiperparâmetros	Conjunto de valores
<i>n_estimators</i>	10, 50, 100, 200
<i>min_samples_leaf</i>	8, 10, 12, 15, 20
<i>max_features</i>	3, "None"
<i>max_depth</i>	8

Fonte: Autor (2024)

Na Tabela 24, $n_estimators$ corresponde à quantidade de árvores de decisão consideradas no modelo RF. Como este *grid* é utilizado no processo de treinamento para todos os cenários investigados, a quantidade de dados disponível neste processo pode ser significativamente diferente. Considerando a teoria comentada no Capítulo 4, é comparado um número de 10 até 200 de árvores de decisão, para considerar a possibilidade do modelo necessitar de baixa quantidade de árvores para atingir boa performance na busca em *grid*, caso em que também diminui-se o custo computacional. Além disso, considera-se o caso em que o valor comumente referenciado (100 árvores de decisão) não seja suficiente. Em relação ao número mínimo de dados na folha (*min_samples_leaf*), opta-se pelo valor mínimo de 8, para garantir que uma quantidade razoável de dados seja utilizada no cálculo da média da predição de cada árvore de decisão. Esse valor é relativamente alto, considerando o intervalo usual de 1 a 20. Esse aumento é uma medida de regularização para minimizar a possibilidade de sobre-ajuste e melhorar a generalização do modelo. Medida semelhante é utilizada em Boukredera et al. (2023) para um caso de predição da ROP com árvores de decisão. Outra medida neste sentido é estipular um número máximo permitido para a profundidade das árvores (*max_depth*), em que diminui-se o tamanho das árvores de decisão e, portanto, a complexidade das árvores é melhor controlada. Adota-se o limite de 8, que também utilizado em Boukredera et al. (2023), por exemplo. Adotam-se duas opções para o número máximo de preditores utilizado em cada nó (*max_features*): subconjunto com 3 preditores e "None", que significa que todos os preditores são considerados. O valor 3 é definido por considerar que ao menos 3, de todos os preditores considerados na Tabela 23, devem ser suficientes para gerar predições razoáveis. A Tabela 25 apresenta os hiperparâmetros do modelo ANN.

Tabela 25 – *Grid* de hiperparâmetros do modelo ANN

Hiperparâmetros	Conjunto de valores
<i>hidden_layer_sizes</i>	5, 8, 10, [2, 2], [3, 3], [4, 4] [5, 5], [2, 2, 2], [3, 3, 3], [4, 4, 4]
<i>learning_rate</i>	10^{-4} , 10^{-3} , 10^{-2} , 10^{-1}
<i>activation</i>	ReLU
<i>max_iter</i>	200
<i>alpha</i>	0,01
<i>n_iter_no_change</i>	10
<i>validation_fraction</i>	10%
<i>solver</i>	SGD

Fonte: Autor (2024)

No caso do modelo ANN, para todos os cenários comparados neste trabalho, utiliza-se o método de parada antecipada. Os termos *max_iter*, *n_iter_no_change* e *validation_fraction* configuram esse método. Desse modo, um valor alto de épocas é definido (200 épocas), e o processo iterativo é interrompido caso não seja detectado melhoria após uma determinada quantidade de épocas (10 épocas). A detecção da melhoria é feita usando uma fração do conjunto

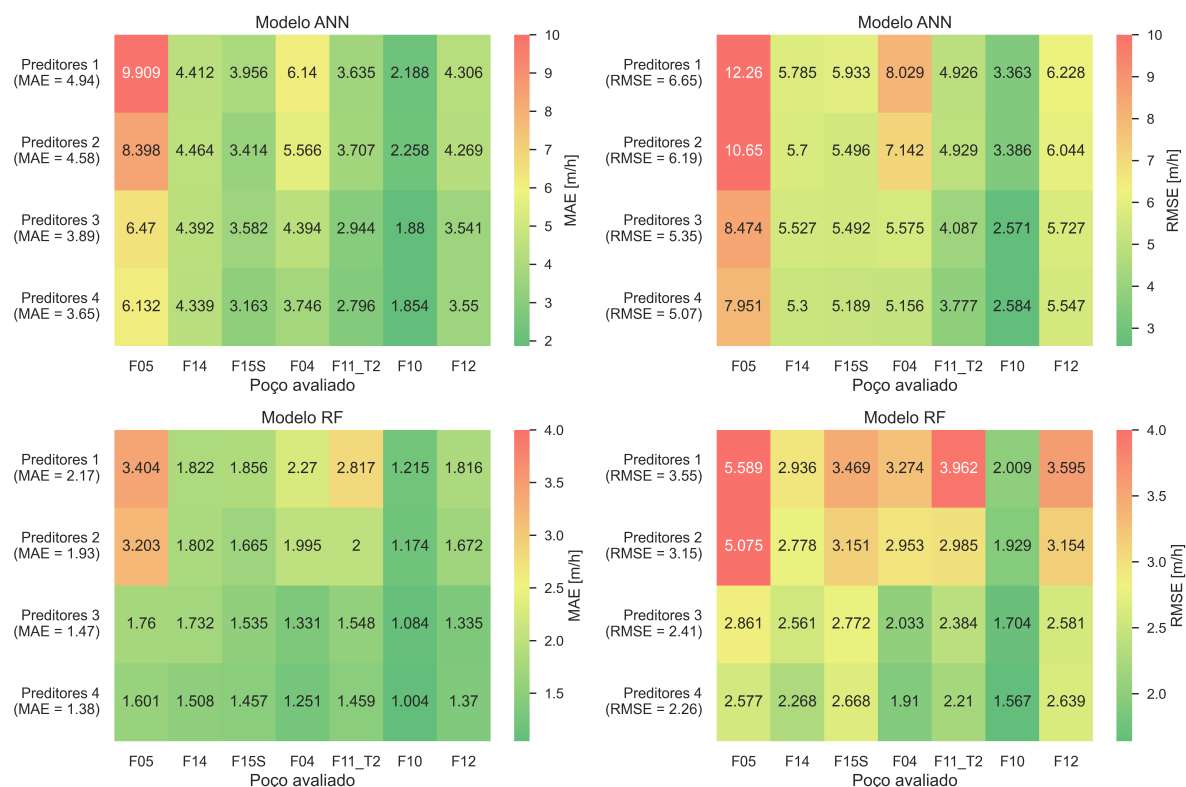
de dados de treinamento (10%). Além disso, todos os modelos ANN deste trabalho são resolvidos por meio do gradiente descendente estocástico utilizando a retropropagação (SGD na Tabela 25). A função de ativação escolhida é a ReLU, que além de ser uma das funções mais comuns para modelos ANNs ajuda a reduzir o custo computacional pela sua simplicidade. Essa função de ativação foi utilizada, por exemplo, no trabalho de Aoun e Madarász (2022) para predição da ROP com um modelo ANN com três camadas ocultas. O α na Tabela 25 corresponde a mais uma medida de regularização para reduzir o risco de sobre-ajuste. Em Soares e Gray (2019) esse termo é comparado na busca em *grid* com valores entre 0,1 e 0,00001. É atribuído o valor 0,01 por ser observado melhores resultados nos cenários analisados neste trabalho.

A taxa de aprendizado (*learning_rate*) é um dos hiperparâmetros mais importante para os modelos ANNs, portanto, é comparado diferentes ordens de grandeza para a taxa de aprendizado. O hiperparâmetro *hidden_layer_sizes* configura o número de camadas e de neurônios, em que [3, 3], por exemplo, representa duas camadas com 3 neurônios cada. Adota-se a prática comum de utilizar a mesma quantidade de neurônios em todas camadas. Tendo em vista o número de preditores na Tabela 23, utilizam-se regras empíricas como referência na escolha da quantidade de neurônios, como é o caso da regra que indica um número de neurônios em uma camada oculta como sendo o dobro do número de *inputs* da rede mais 1 (SILVA et al., 2016). Considerando o número de preditores na Tabela 23, adota-se o valor máximo de 10 neurônios. Para os casos em que o número de neurônios é muito maior do que a referência empírica (Preditores 1) as medidas de regularização implementadas (parada antecipada) atuam, em conjunto com a busca em *grid*, para impedir o sobre-ajuste nesses casos. Como mencionado em Silva et al. (2016), regras empíricas servem apenas como referência, e o valor ideal é determinado pelo tipo de problema. Em Elmgerbi et al. (2021), são especificados 20 neurônios para uma camada oculta em um modelo com 3 preditores, e em Aoun e Madarász (2022) são especificados 12 neurônios em cada uma das 3 camadas para um modelo com 5 preditores. Portanto, também foi incluído modelos com até 3 camadas.

Nas Tabelas 24 e 25, são comparados 40 candidatos de cada modelo. Por exemplo, no caso da Tabela 24 o número de combinações é $4 \cdot 5 \cdot 2 \cdot 1 = 40$. Como apresentado na Figura 31, cada modelo é treinado 5 vezes, e a seleção dos 40 modelos é feita com o melhor resultado nas partições não utilizadas no treinamento. Desse modo, cada vez que a busca em *grid* é utilizada nas comparações neste trabalho, têm-se $40 \cdot 5 = 200$ ajustes realizados. Portanto, o custo computacional para aplicar os modelos de aprendizagem de máquina é muito maior do que nos modelos tradicionais.

Assim como foi feita a comparação do resultado dos modelos tradicionais com os dados medidos da ROP em cada poço (Figura 47), a Figura 48 apresenta o *heatmap* com os resultados das métricas de erros para os dois modelos, RF e ANN, usando os preditores da Tabela 23 e para cada um dos 7 poços.

Figura 48 – *Heatmap* com as métricas de erro em m/h dos modelos de aprendizagem de máquina com preditores diferentes e para cada poço

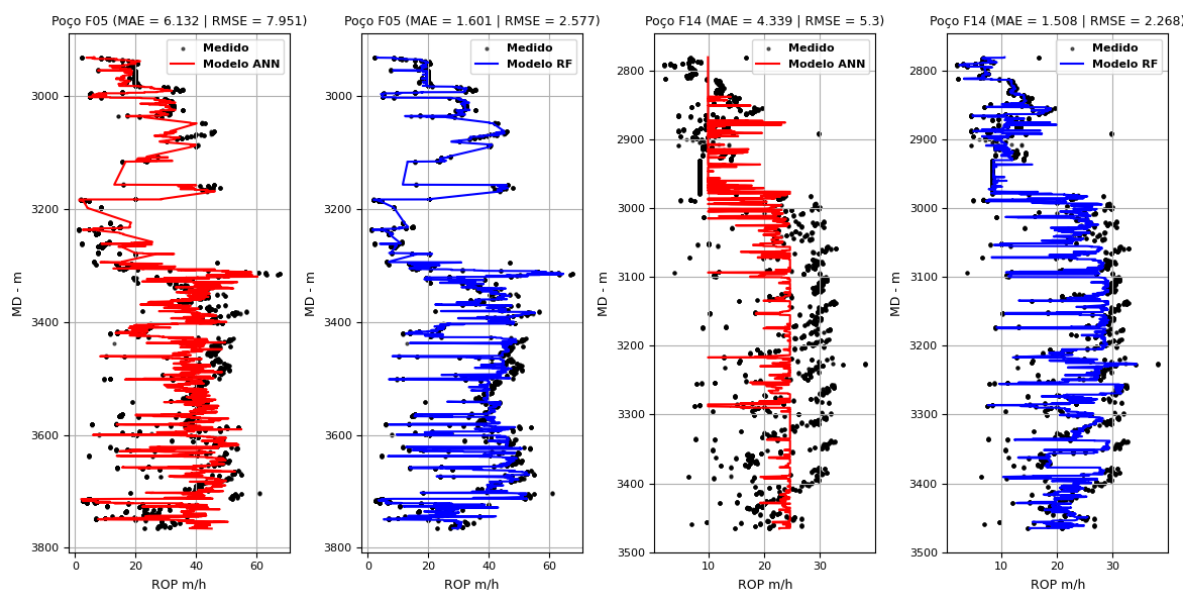


Fonte: Autor (2024)

Percebe-se, inicialmente, que os erros são menores em relação aos resultados dos modelos tradicionais apresentados anteriormente. Além disso, como pode ser visualizado, melhor estimativa é alcançada com o conjunto Preditores 4 (Tabela 23), para o qual a média de cada métrica é menor. As menores médias para todos os poços no modelo ANN são MAE = 3,65 m/h e RMSE = 5,07 m/h, para o modelo RF são MAE = 1,38 m/h e RMSE = 2,28 m/h. Portanto, observam-se erros menores para o modelo RF. É possível perceber também que para o poço F05, os modelos sem o parâmetro de torque (Preditores 1 e Preditores 2) são os que apresentam menor desempenho, o que está de acordo com a análise de significância do modelo BYMmod. Além disso, é possível perceber que para o poço F10, mesmo para o conjunto Preditores 1, os erros são significativamente menores. Essa observação está de acordo com a análise dos modelos tradicionais, nos quais parâmetros operacionais são suficientes para capturar o padrão da ROP neste poço. É possível perceber, comparando as duas métricas, que para o conjunto Preditores 4 existe maior proximidade entre o MAE e RMSE indicando a maior uniformidade dos erros. As Figuras 49, 50, 51 e 52 apresentam a comparação da resposta de cada modelo com as medidas da ROP em cada poço considerando o conjunto Preditores 4.

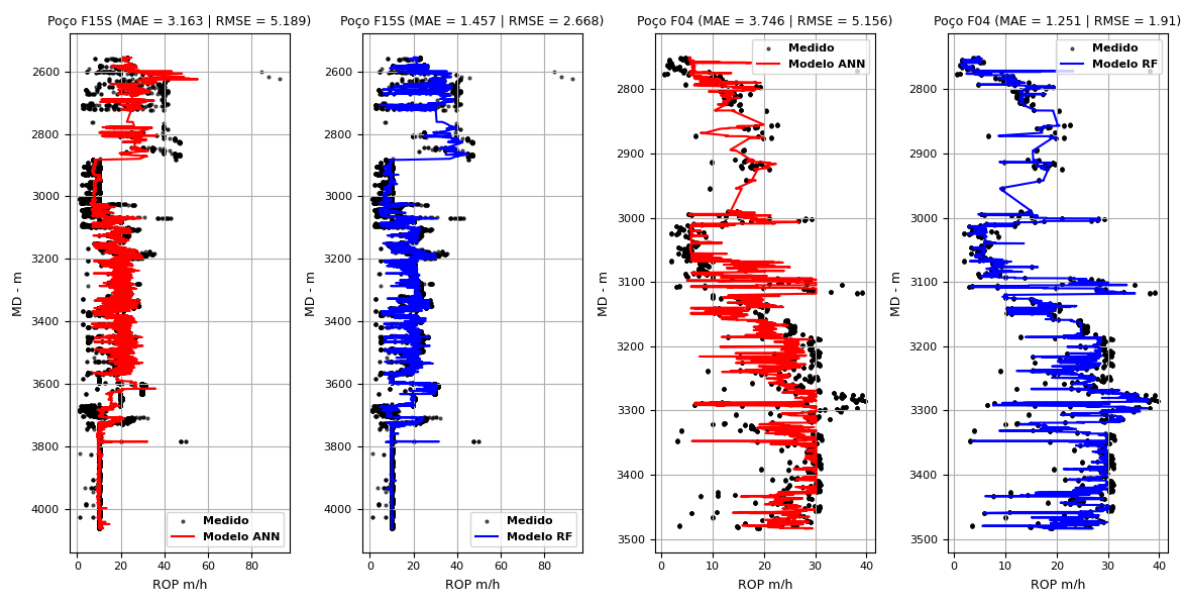
Novamente, percebe-se que os dados utilizados são representativos e que com os pre-

Figura 49 – Resposta dos modelos RF e ANN nos poços F05 e F14



Fonte: Autor (2024)

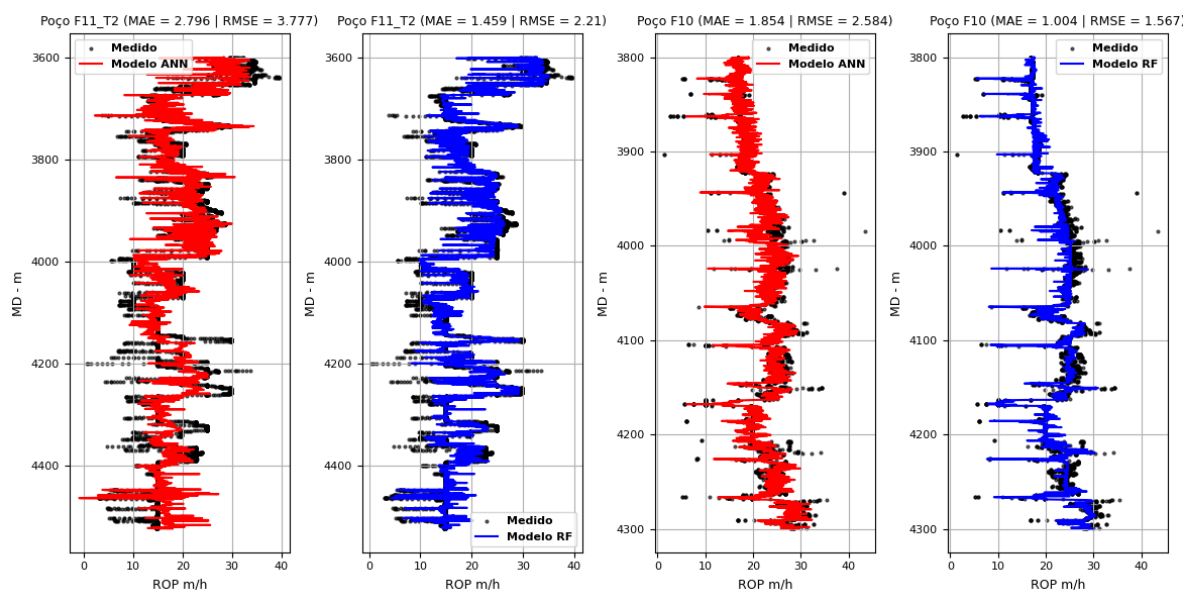
Figura 50 – Resposta dos modelos RF e ANN nos poços F15S e F04



Fonte: Autor (2024)

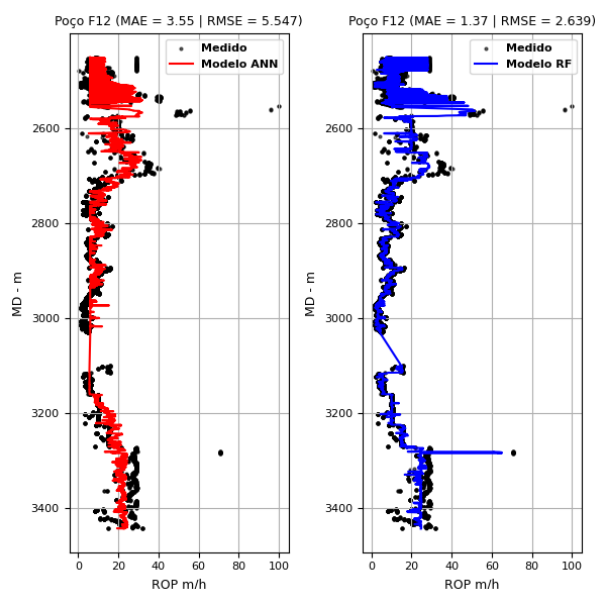
ditores utilizados é possível capturar adequadamente o padrão da ROP em todos os poços. A aderência das informações do modelo e as medições são melhores do que as alcançadas pelos modelos tradicionais considerando todas aquelas informações incluídas nos modelos, conforme discutido anteriormente. Como a busca em *grid* é utilizada, as diferentes versões dos modelos RF e ANN são comparadas e, portanto, os resultados apresentados nas Figura 49, 50, 51, 52 correspondem a modelos com diferentes hiperparâmetros. Na Tabela 26 e 27 são apresentados os melhores hiperparâmetros de cada método e também o tempo de execução no treinamento.

Figura 51 – Resposta dos modelos RF e ANN nos poços F11_T2 e F10



Fonte: Autor (2024)

Figura 52 – Resposta dos modelos RF e ANN no poço F12



Fonte: Autor (2024)

Inicialmente, ao comparar os resultados na Tabela 26 e 27, nota-se que os tempos de treinamento dos modelos RF foram menores para 5 dos 7 poços. O tempo de treinamento é maior para o poço F15S do que os demais, uma vez que esse poço é o que possui mais dados (Tabela 10). Além disso, percebe-se que os hiperparâmetros do modelo ANN variam muito mais nos diferentes poços do que os melhores hiperparâmetros do modelo RF. Na Tabela 26, apenas os poços F05 e F15S tiveram a mesma configuração. No caso do modelo RF, os poços F14, F15S, F11_T2 e F12 apresentaram o mesmo conjunto de hiperparâmetros, sendo que todos

possuem a mesma quantidade de dados na folha. Desse modo, nota-se a maior dificuldade de especificar uma arquitetura de modelo ANN única para todos os poços, uma vez que se observam níveis diferentes de complexidade de rede em cada poço. Para o poço F10, poucos neurônios são necessários em apenas uma camada, para o poço F12, é necessário maior quantidade de neurônios e duas camadas na rede (maior quantidade de parâmetros). O nível de complexidade desses modelos estão de acordo com as análises dos modelos analíticos (vide Figura 46). No modelo RF, para a maioria dos poços, é observada uma menor quantidade de preditores considerado em cada partição das árvores. Essa observação pode indicar a necessidade de ressaltar a influência de preditores menos significativos, no conjunto Preditores 4, e a ROP, conforme comentado no Capítulo 4.

Tabela 26 – Melhores hiperparâmetros e tempo de treinamento do modelo ANN (Preditores 4) treinado com os dados de cada poço

Poços de teste	Nº de camadas ocultas e neurônios	Taxa de aprendizagem	Tempo de treinamento (s)
F05	10	0,01	42,00
F14	[3,3]	0,01	68,25
F15S	10	0,01	90,11
F04	[3,3]	0,001	54,82
F11_T2	10	0,01	71,50
F10	5	0,01	56,88
F12	[5,5]	0,001	80,82

Fonte: Autor (2024)

Tabela 27 – Melhores hiperparâmetros e tempo de treinamento do modelo RF (Preditores 4) treinado com os dados de cada poço

Poços de teste	Max preditores	Nº de estimadores	Nº min. de dados na folha	Tempo de treinamento (s)
F05	None	200	8	40,07
F14	3	200	8	44,36
F15S	3	200	8	145,12
F04	None	100	8	37,91
F11_T2	3	200	8	99,80
F10	3	100	8	51,66
F12	3	200	8	78,68

Fonte: Autor (2024)

As Tabelas 28 e 29 apresentam os resultados da busca em *grid*. É apresentado o RMSE médio das 5 partições não utilizadas no treinamento, onde para cada conjunto de preditor é

apresentado o que possui pior e melhor performance (maior e menor erro respectivamente). Em cada conjunto de preditor, os resultados são semelhantes aos apresentados na Figura 48, já que é o mesmo modelo. É possível perceber que no modelo ANN a diferença entre o pior e o melhor é maior quando comparado com os resultados do modelo RF. Para o conjunto Preditores 4, a maior diferença é 7,61 m/h e, para os resultados do modelo RF, é 1,37 m/h. Portanto, esses resultados mostram a relevância de comparar modelos com diferentes hiperparâmetros e também mostram o impacto, no resultado final, de diferentes hiperparâmetros para o modelo ANN.

Tabela 28 – RMSE médio das 5 partições na busca em *grid* do modelo ANN considerando pior e melhor combinação de hiperparâmetros

Poços	Preditores 01		Preditores 02		Preditores 03		Preditores 04	
	Pior	Melhor	Pior	Melhor	Pior	Melhor	Pior	Melhor
F05	15,30	12,26	15,28	10,7	15,28	8,56	15,28	7,67
F14	8,76	5,77	8,76	5,71	8,76	5,62	8,77	5,12
F15S	8,12	6,03	8,12	5,59	8,12	5,54	8,12	5,28
F04	9,76	8,12	9,76	7,20	9,76	5,66	9,77	5,15
F11_T2	6,25	4,93	6,24	4,91	6,24	4,11	6,24	3,71
F10	4,75	3,43	4,74	3,41	4,75	2,58	4,75	2,55
F12	8,86	6,34	8,86	6,07	8,86	5,79	8,86	5,56

Fonte: Autor (2024)

Tabela 29 – RMSE médio das 5 partições na busca em *grid* do modelo RF considerando pior e melhor combinação de hiperparâmetros

Poços	Preditores 01		Preditores 02		Preditores 03		Preditores 04	
	Pior	Melhor	Pior	Melhor	Pior	Melhor	Pior	Melhor
F05	6,67	5,83	6,61	5,58	4,49	3,35	4,46	3,10
F14	3,68	3,15	3,57	3,10	3,43	2,88	3,00	2,53
F15S	3,91	3,72	3,68	3,44	3,37	3,06	3,28	2,96
F04	4,38	3,55	4,15	3,34	3,20	2,43	3,15	2,29
F11_T2	4,28	4,24	3,44	3,27	2,86	2,66	2,69	2,47
F10	2,51	2,32	2,48	2,24	2,12	1,95	2,01	1,81
F12	4,39	3,84	4,29	3,50	3,76	3,00	3,71	3,07

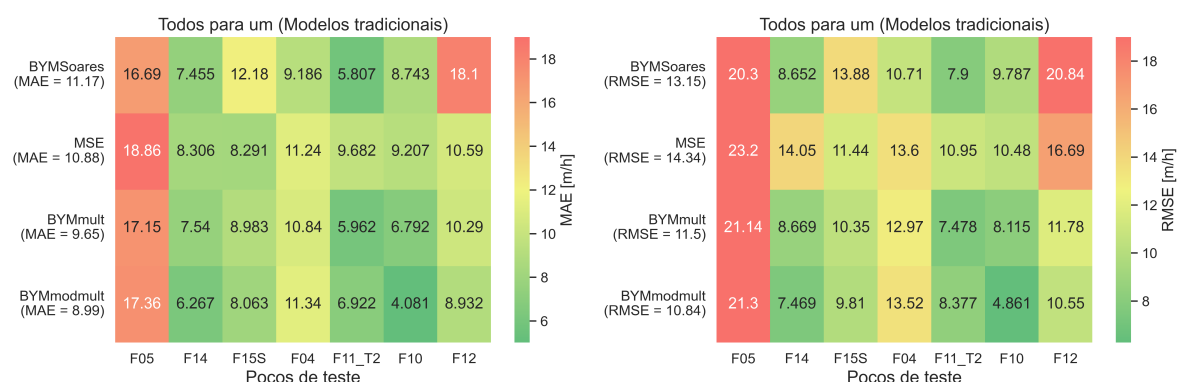
Fonte: Autor (2024)

6.3 Estratégia de predição da ROP: Todos para um

Na Figura 53, os resultados da predição da ROP com modelos tradicionais e estratégia todos para um são apresentados. Desse modo, é feita a predição da ROP para um poço de teste utilizando os dados de todos os demais poços para determinar os coeficientes dos modelos tradicionais, utilizando os mesmos métodos comentados anteriormente. Inicialmente, como é possível perceber, as métricas de erro são maiores do que as obtidas na Figura 47, caso em que os coeficientes do modelo foram estimados com os dados do próprio poço. Considerando os

resultados dos modelos tradicionais e de aprendizagem de máquina já apresentados, nota-se, por exemplo, que um MAE em torno de 6,0 m/h geralmente está atrelado a uma boa aderência entre a ROP medida e prevista, conforme as curvas apresentadas anteriormente. Na Figura 53, são apresentados também os resultados dos modelos BYMmult e BYMmodmult como valores de referência, uma vez que apenas os modelos BYMSoares e MSE possuem aplicabilidade para predições em tempo real, pois não dependem de informações que não são possíveis de estimar com precisão nesta situação, como o desgaste e o gradiente de poro pressão.

Figura 53 – *Heatmap* com as métricas de erro em m/h na predição para cada poço



Fonte: Autor (2024)

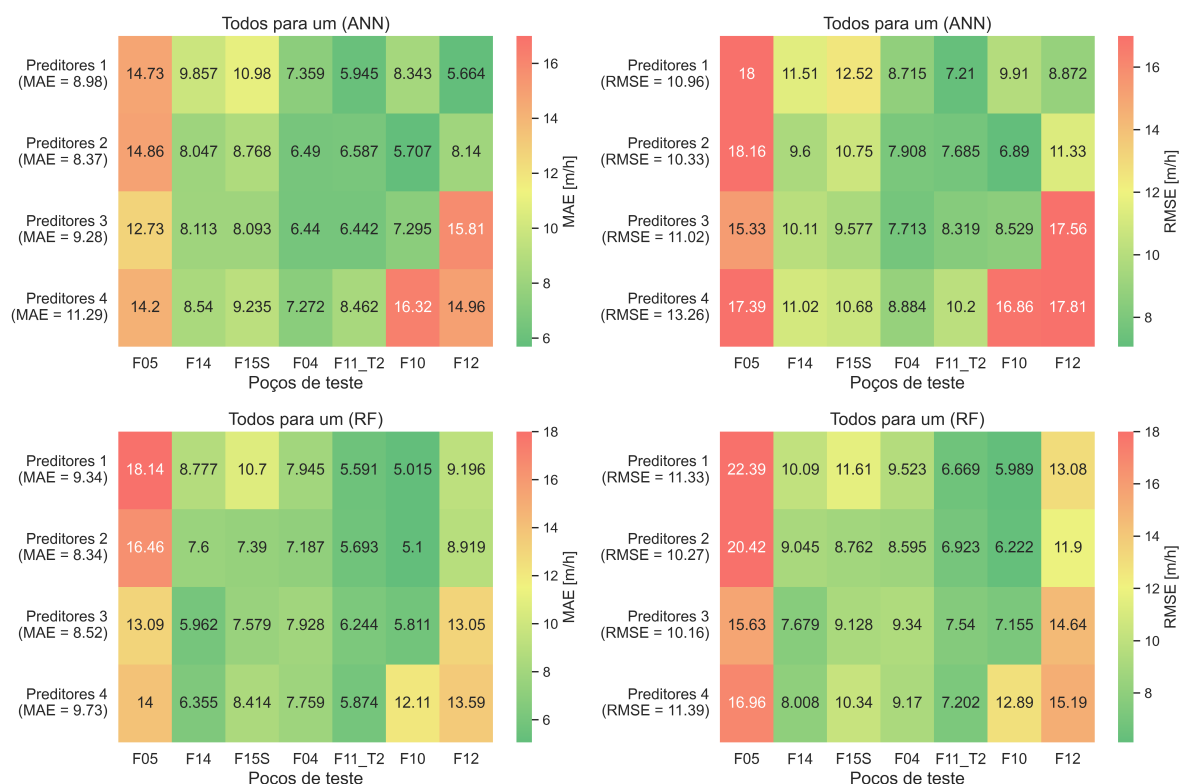
É possível perceber que, para todos os casos, a predição do poço F05 utilizando as informações de outros poços resultou em erros muito maiores. No modelo BYMSoares, é possível perceber que em ao menos três poços (F05, F15S e F12) os erros são elevados em relação aos resultados de todos os modelos para os demais poços. Além disso, é possível perceber maior diferença entre as métricas de erro MAE e RMSE para o modelo MSE, o que sinaliza a presença de diferenças elevadas (maior nos poços F14 e F12). O modelo que apresenta menores erros de modo geral é o BYMmodmult, quando é comparado a média das duas métricas (MAE = 8,99 m/h e RMSE = 10,84 m/h) com a média dos demais modelos. No entanto, como mencionado, esse modelo não é viável em tempo real.

Na Figura 54, são apresentados os resultados para os modelos de aprendizagem de máquina. Novamente, é possível perceber o aumento esperado dos erros quando comparado com os valores apresentados na Figura 48. As métricas de erro apresentadas correspondem ao erro de generalização dos modelos de predição, uma vez que são dados que não foram utilizados em nenhuma etapa do processo de treinamento ou de busca dos hiperparâmetros.

É possível perceber que, de modo geral, utilizar o conjunto de Preditores 4 não induziu menores erros quando é comparado a média das métricas. Esse resultado é observado para os dois modelos. Percebe-se que esse resultado é devido, principalmente, a erros maiores na predição do poço F12, em que também é possível notar que a informação do torque não contribuiu na melhoria da predição da ROP para esse poço, visto que o conjunto de Preditores 3 e Preditores 4

possuem erros muito maiores do que nos demais conjuntos de preditores. Para o poço F05, a informação do torque é relevante, no entanto, assim como foi identificado na Figura 53, erros elevados são observados para esse poço. Nota-se, portanto, que com os modelos utilizados, não é possível estimar um padrão claro entre a ROP e as informações consideradas no conjunto Preditores 4 para o caso desses dois poços. Relações específicas entre a ROP e esses preditores, que não são generalizadas para todos os poços de treinamento, é uma possível causa. Essa observação é entendida quando são considerados os problemas de vibração torcional que são identificados no poço F05 e, por exemplo, a presença de outro tipo de broca no poço F12. Desse modo, nessa estratégia todos para um, o conjunto de preditores que apresentam melhor performance é aquele em que as relações entre os preditores e a ROP estão presentes nos poços utilizados no treinamento.

Figura 54 – *Heatmap* com as métricas de erro em m/h dos modelos de aprendizagem de máquina com preditores diferentes e para cada poço na estratégia todos para um

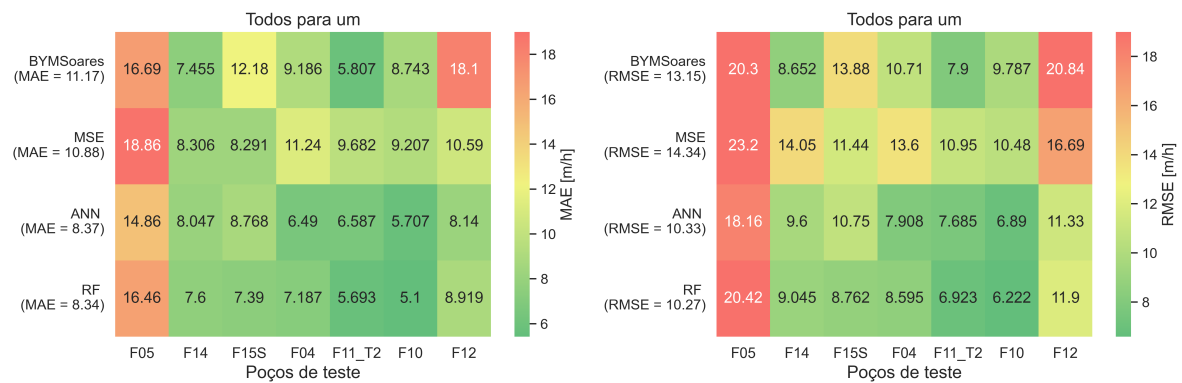


Fonte: Autor (2024)

Os modelos de aprendizagem de máquina apresentam melhor performance do que os modelos tradicionais, uma vez que os erros são menores e, considerando o conjunto Preditores 2, percebe-se que, de modo geral, os modelos de aprendizagem de máquina geram erros menores para a maioria dos poços. Essa conclusão é feita incluindo os modelos BYMmult e BYMmodmult, o que mostra a boa capacidade preditiva, uma vez que poucas informações são incluídas no modelo (poucos preditores). O preditor Delta-T mostra ser uma inclusão que contribui na

performance do modelo, pois, na Figura 54, o conjunto Preditores 2 possibilita diminuir os erros na maioria dos poços em relação ao conjunto Preditores 1. A Figura 55 mostra a comparação entre os modelos tradicionais e os modelos de aprendizagem de máquina. Comparando os dois modelos de aprendizagem de máquina, percebe-se que o modelo RF apresenta melhor performance, uma vez que na maior parte das combinações, as métricas de erros são menores para o modelo RF, mesmo caso ocorre na Figura 54.

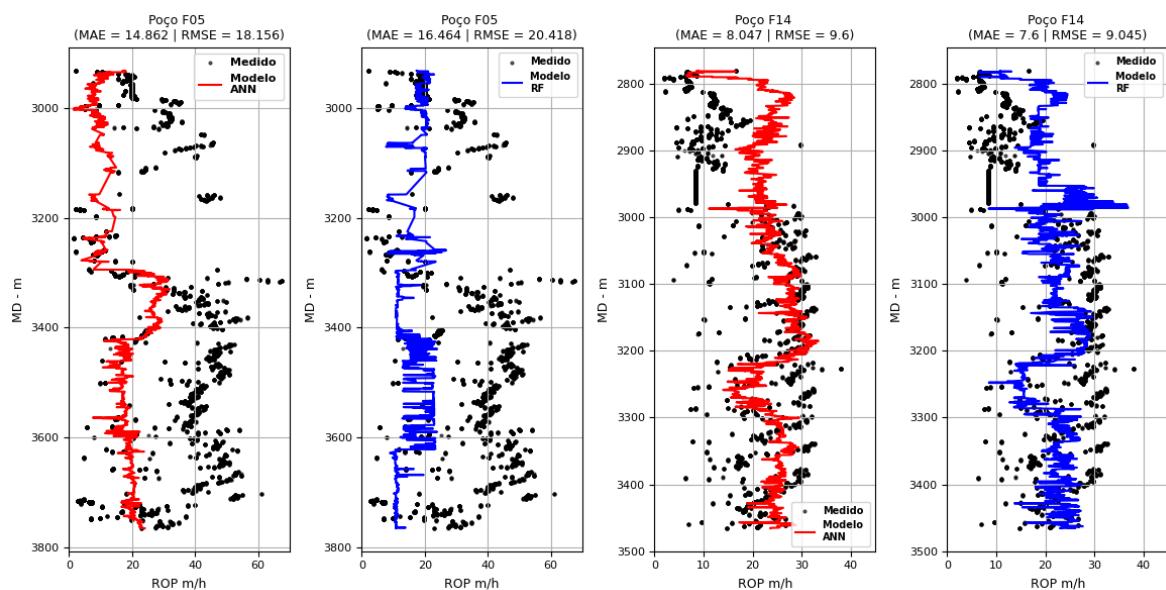
Figura 55 – *Heatmap* com as métricas de erro em m/h com modelos tradicionais e de aprendizado de máquina na estratégia todos para um



Fonte: Autor (2024)

As Figuras 56, 57, 58 e 59 apresentam os perfis que indicam a resposta do modelos de aprendizagem de máquina (utilizando o conjunto Preditores 2) ao longo da profundidade.

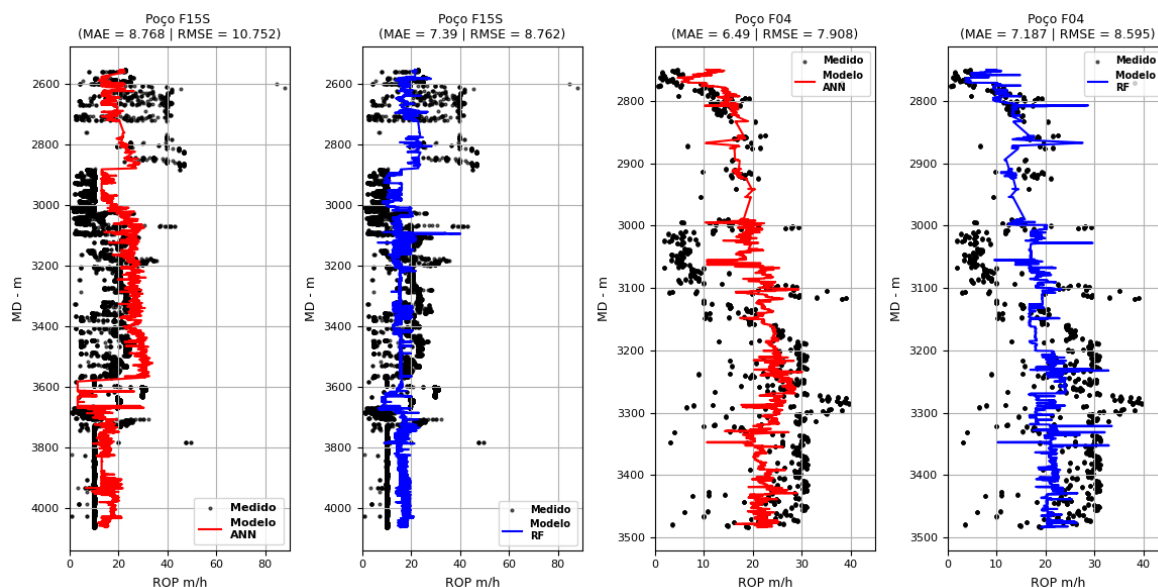
Figura 56 – Resposta dos modelos RF e ANN nos poços F05 e F14



Fonte: Autor (2024)

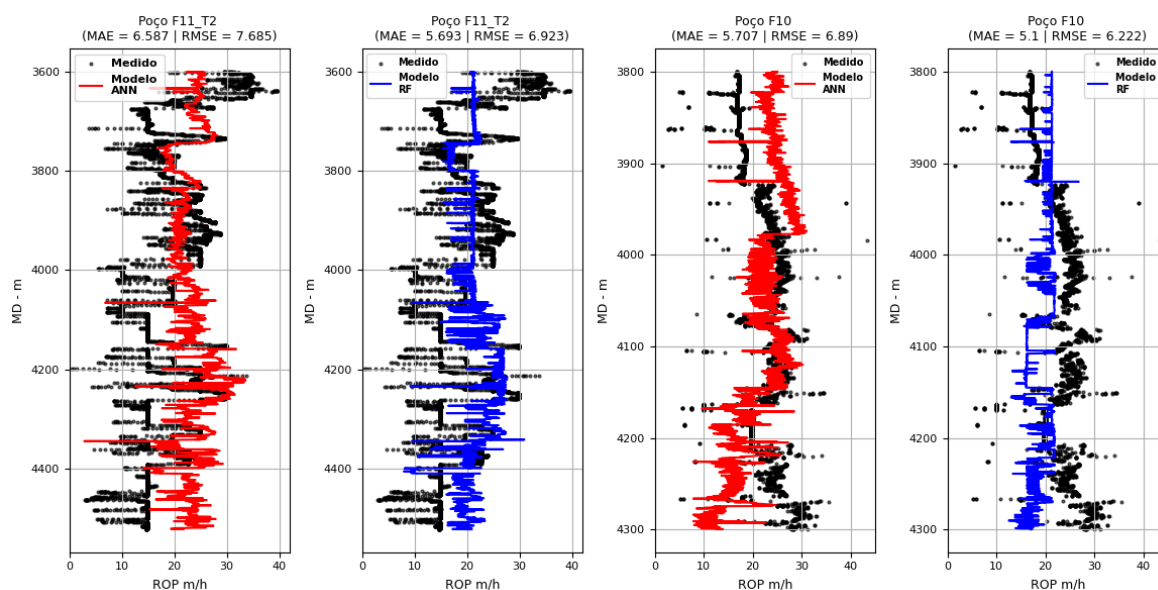
Percebe-se que os modelos treinados com dados de poços de correlação capturam o padrão da ROP razoavelmente bem para alguns poços, como observado, por exemplo, na Figura 57. Esse fato demonstra mais um vez a capacidade preditiva dos modelos de aprendizagem de máquina, pois capturam padrões em grande quantidade de dados de poços distintos. No entanto, em todos os poços é possível verificar trechos em que existem diferenças significativas, principalmente para o poço F05 na Figura 56.

Figura 57 – Resposta dos modelos RF e ANN nos poços F15S e F04



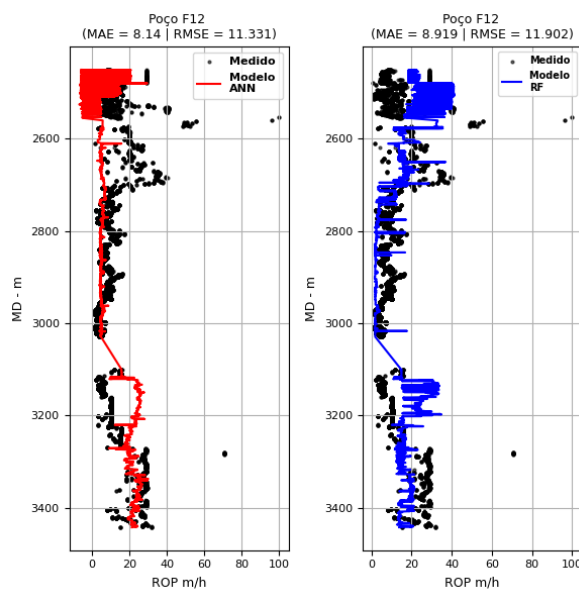
Fonte: Autor (2024)

Figura 58 – Resposta dos modelos RF e ANN nos poços F11_T2 e F10



Fonte: Autor (2024)

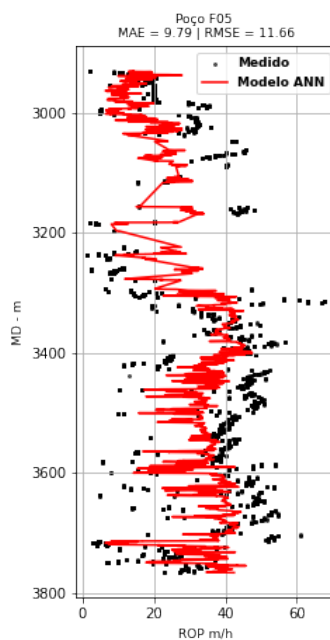
Figura 59 – Resposta dos modelos RF e ANN no poço F12



Fonte: Autor (2024)

Considerando o poço F05, e as informações discutidas anteriormente sobre a detecção de problemas de vibração *stick-slip* nos poços F05, F14 e F04, a Figura 60 apresenta o resultado da predição da ROP no poço F05 com apenas dados dos poços F14 e F04.

Figura 60 – Resposta dos modelos RF e ANN no poços F12



Fonte: Autor (2024)

No exemplo da Figura 60, é utilizado o conjunto Preditores 3 e o modelo ANN, que obtiveram melhor performance para esse poço, porém ainda com as métricas de erro muito elevadas. Observa-se a redução significativa dos erros e também que a resposta do modelo captura bem o padrão da ROP ao longo da profundidade, inclusive nos intervalos com registro de vibração excessiva, com resultados próximos aqueles obtidos com os modelos tradicionais que utilizam os próprios dados do poço F05 (Figura 38). A arquitetura da rede empregada utiliza 1 camada oculta com 10 neurônios e taxa de aprendizado de 0,001. Esse resultado mostra que é possível gerar modelos que capturam bem o padrão da ROP com poços de correlação, sendo, portanto, uma estratégia possível para fazer previsões e utilizar esses modelos na otimização em tempo real de parâmetros da perfuração. No entanto, é necessário selecionar adequadamente os dados desses poços e empregar preditores representativos, pois poços de treinamento com características diversas diminuem a capacidade preditiva do modelo. Interpretabilidade dos dados de perfuração, utilizar dados de uma mesma broca e de poços em região semelhante auxiliam neste processo.

As Tabelas 30 e 31 apresentam os hiperparâmetros selecionados na busca em *grid* dos modelos da Figura 55. Novamente, os tempos de treinamento para o modelo RF são menores. Em relação à arquitetura dos modelos de aprendizagem de máquina, percebe-se menor variação entre os hiperparâmetros apresentados para os dois modelos. Devido às limitações identificadas neste método de predição, os modelos de aprendizagem de máquina, com o conjunto Preditores 2 e hiperparâmetros apresentados nas Tabelas 30 e 31, não representam os modelos ideais para cada poço, pois o padrão da ROP não foi aprendido em todos os poços, como no poço F05. Fica claro nessa estratégia que é necessária uma metodologia para selecionar dados de treinamento representativos e preditores relevantes, como feito na Figura 60.

Tabela 30 – Melhores hiperparâmetros e tempo de treinamento do modelo ANN (Preditores 2) na estratégia todos para um

Poços de teste	Nº de camadas ocultas e neurônios	Taxa de aprendizagem	Tempo de treinamento (s)
F05	5	0,01	472
F14	[4,4]	0,001	387
F15S	[3,3]	0,0001	306
F04	[4,4]	0,001	402
F11_T2	[4,4]	0,001	333
F10	[4,4]	0,001	389
F12	[4,4]	0,001	352

Fonte: Autor (2024)

Tabela 31 – Melhores hiperparâmetros e tempo de treinamento do modelo RF (Preditores 2) na estratégia todos para um

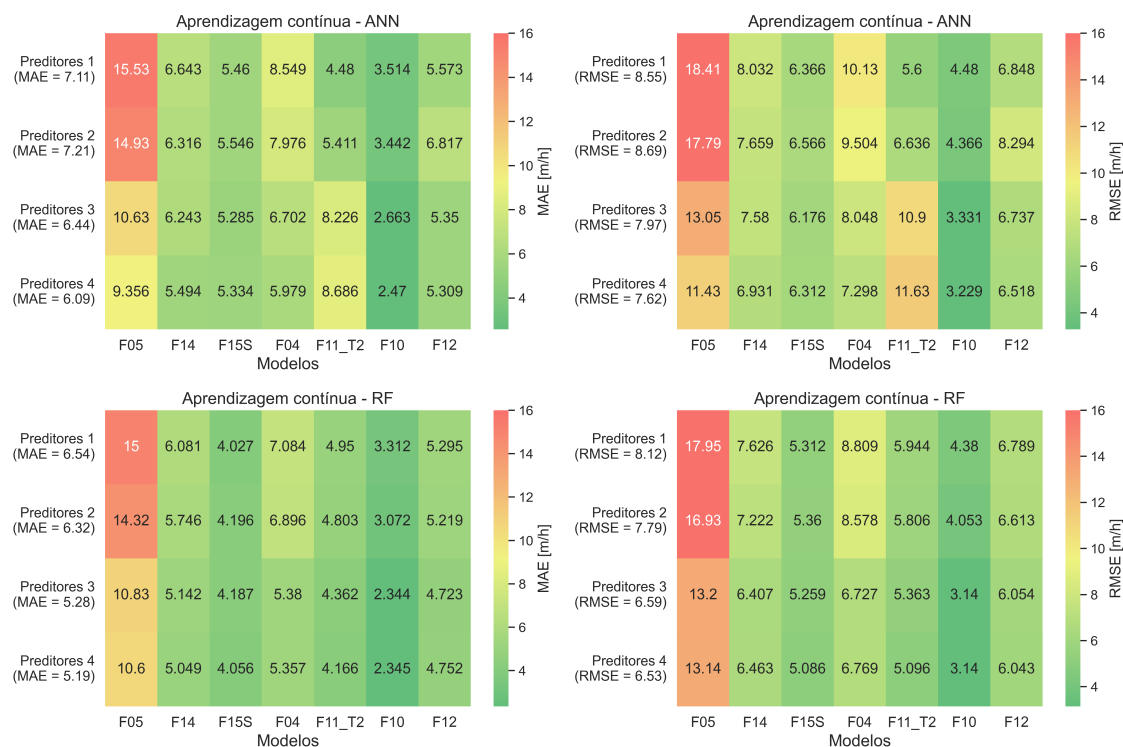
Poços de teste	Max preditores	Nº de estimadores	Nº min. de dados na folha	Tempo de treinamento (s)
<i>F05</i>	3	100	8	224
<i>F14</i>	3	100	8	216
<i>F15S</i>	3	100	8	176
<i>F04</i>	3	200	8	200
<i>F11_T2</i>	3	100	8	205
<i>F10</i>	3	200	8	227
<i>F12</i>	3	100	8	172

Fonte: Autor (2024)

6.4 Estratégia de predição da ROP: Aprendizagem contínua

A Figura 61 apresenta os resultados dos modelos de aprendizagem de máquina na estratégia de aprendizagem contínua para cada poço e considerando os diferentes preditores da Tabela 23.

Figura 61 – *Heatmap* com as métricas de erro em m/h dos modelos de aprendizagem de máquina com preditores diferentes e para cada poço na estratégia de aprendizagem contínua

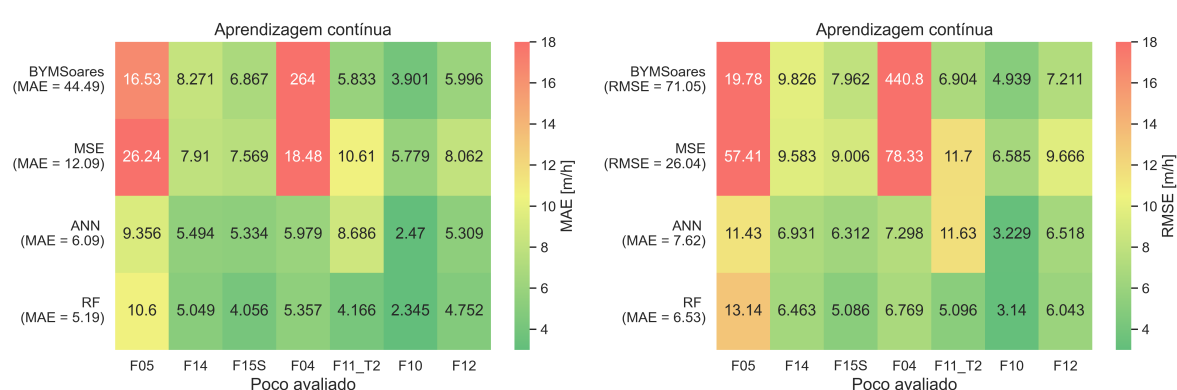


Fonte: Autor (2024)

Como explicado no Capítulo 5, a aprendizagem contínua envolve a simulação de um cenário em tempo real na qual o modelo é continuamente treinado com cada vez mais dados do próprio poço, e as previsões são feitas em intervalos à frente da broca (intervalos de teste). Os resultados apresentados na Figura 61 correspondem às médias dos intervalos de teste ao longo do poço que, por sua vez, possuem a mesma quantidade de pontos (adotam-se 500 pontos). Desse modo, as métricas de erro apresentadas são os erros de generalização, assim como os obtidos na estratégia todos para um, em que as métricas de erro são calculadas em dados não utilizados no treinamento e na seleção dos hiperparâmetros. Portanto, é possível perceber que, de modo geral, os valores encontrados são menores do que os obtidos na estratégia todos para um, conforme foi apresentado na Figura 54.

Como pode ser observado, o conjunto Preditores 4 é o que apresenta menor média das métricas para todos os poços, estando de acordo com o que foi apresentado na Figura 48, em que foi verificada a capacidade dos modelos de aprendizagem de máquina, com esses preditores, de capturar o padrão da ROP em cada poço. Percebe-se, também, que os dois modelos apresentam erros com tendência semelhante, com erros menores no poço F10 e maiores no poço F05. Verifica-se, ainda, a maior importância do torque neste poço (erros muito maiores para os conjuntos de preditores sem torque). Para os dois modelos, na maioria dos poços, há redução das métricas de erro quando é considerado o Delta-T. É possível perceber que novamente o modelo RF apresenta erros menores considerando todas as métricas. A Figura 62 apresenta a comparação dos modelos de aprendizagem de máquina (Preditores 4) com os modelos BYMSoarees e MSE, conforme feito na Figura 55.

Figura 62 – *Heatmap* com as métricas de erro em m/h com modelos tradicionais e de aprendizado de máquina na estratégia de aprendizado contínuo

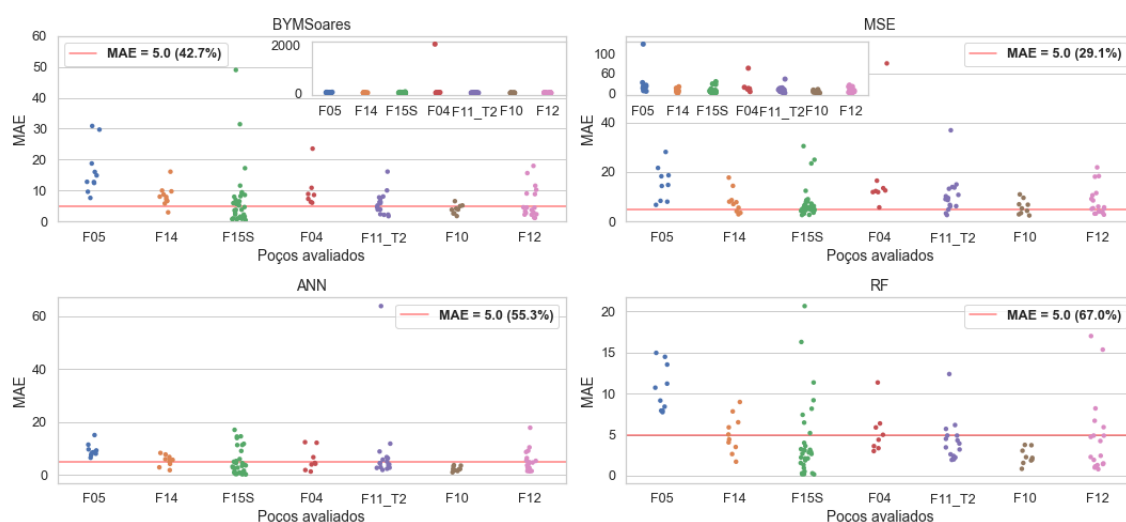


Fonte: Autor (2024)

Observam-se erros muito significativos para os dois modelos tradicionais em ao menos dois poços: F05 e F04. Além disso, considerando as menores médias de erro geral obtidas na estratégia todos para um: MAE = 8,34 m/h e RMSE = 10,27 m/h, no modelo RF, percebe-se que os dois modelos tradicionais também apresentam erros elevados em outros poços, como

no F14 e F11_T2, para o BYMSoares e o MSE, respectivamente. Além disso, considerando essa referência, nota-se um valor elevado também no modelo ANN (poço F11_T2), caso em que é possível perceber também uma maior diferença entre MAE e RMSE, indicando valores discrepantes. Na Figura 63, é apresentada, em mais detalhes, a comparação entre os erros da aprendizagem contínua e a estratégia todos para um, para as quais os valores das métricas MAE nos intervalos de teste de cada poço (cada ponto corresponde a média dos 500 pontos em cada intervalo) são comparados com o menor valor da estratégia todos para um (conforme Figura 55).

Figura 63 – Comparação entre estratégia de aprendizagem contínua e todos para um para todos os modelos e em todos os poços



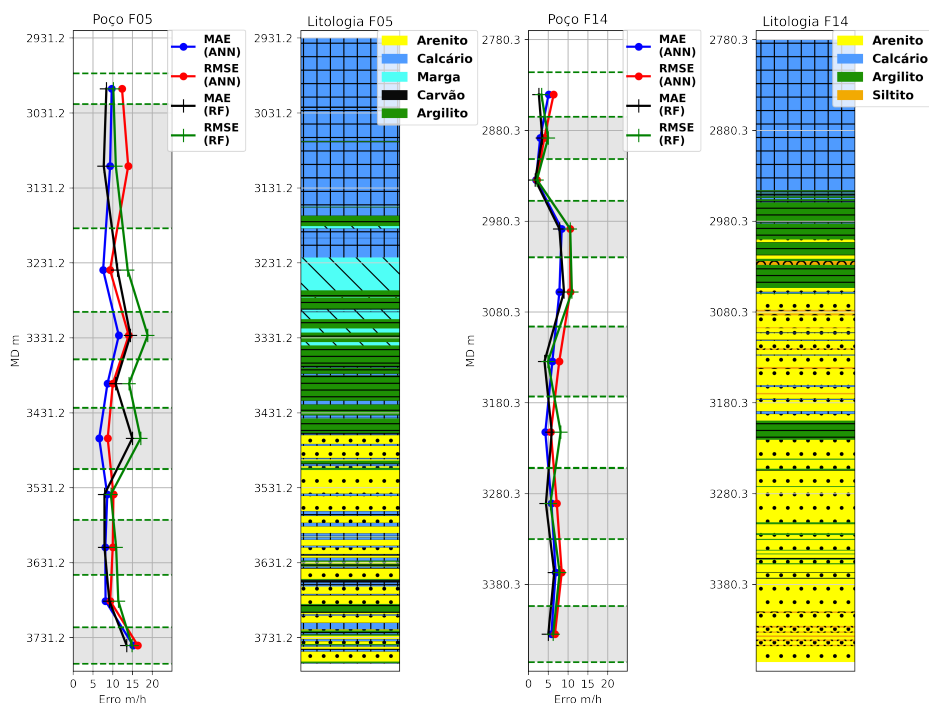
Fonte: Autor (2024)

Na Figura 63, é possível visualizar os valores discrepantes dos erros nos intervalos de teste para cada poço. Ao comparar os pontos com o menor valor da estratégia todos para um (reta horizontal), percebe-se que, para os modelos de aprendizagem de máquina, a maioria dos valores encontra-se abaixo da reta. Para o modelo RF, 67,0% dos erros calculados nos intervalos são menores que este valor, conforme Figura 63. Portanto, percebe-se que a estratégia de aprendizagem contínua possibilita fazer previsões com menores erros mesmo considerando os poucos dados utilizados no treinamento. Além disso, novamente é identificado melhor performance do modelo RF e melhor performance dos modelos de aprendizagem de máquina.

Nas Figuras 64, 65, 66 e 67 são apresentados os valores das métricas de erro calculadas nos intervalos de teste para cada poço ao longo da profundidade. Além disso, a informação de litologia é apresentada como referência. Os intervalos destacados, no plot à esquerda em cada figura, mostram as divisões entre intervalos de teste da estratégia de aprendizagem contínua. Essas retas delimitam o intervalo de profundidade definido pelos 500 pontos utilizados. Como pode ser observado, os intervalos não possuem o mesmo tamanho, isso ocorre por conta de *gaps* entre medições. No entanto, as litologias representadas entre essas retas correspondem às litologias registradas para os 500 pontos analisados. Com a informação da litologia é possível

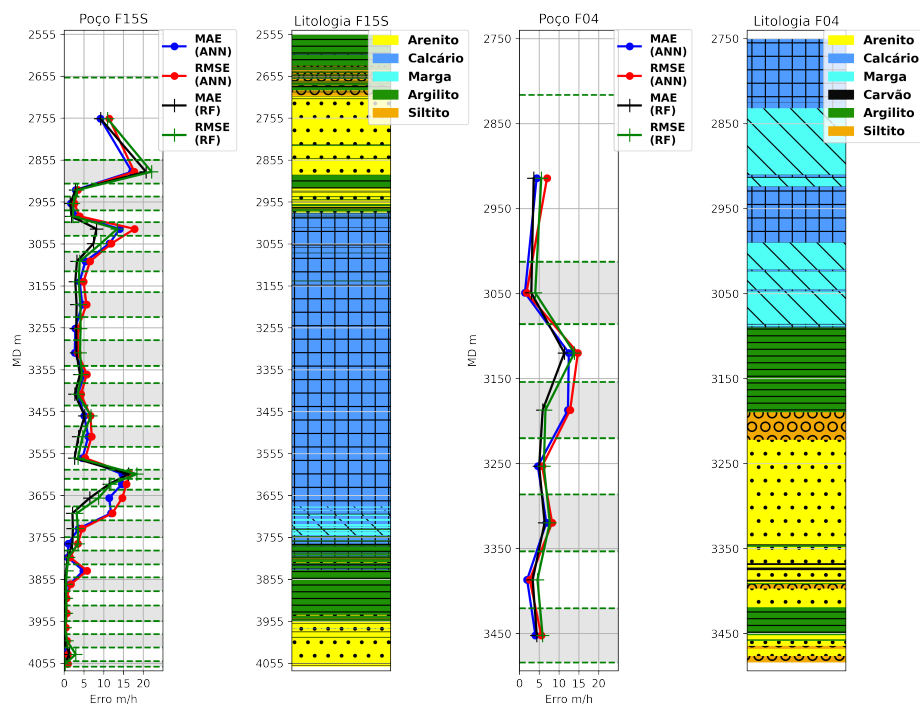
analisar, em mais detalhes, os resultados obtidos pelos modelos de aprendizagem de máquina.

Figura 64 – Resultados da aprendizagem contínua (Preditores 4) nos poços F05 e F14 ao longo da profundidade em comparação com a litologia dos poços



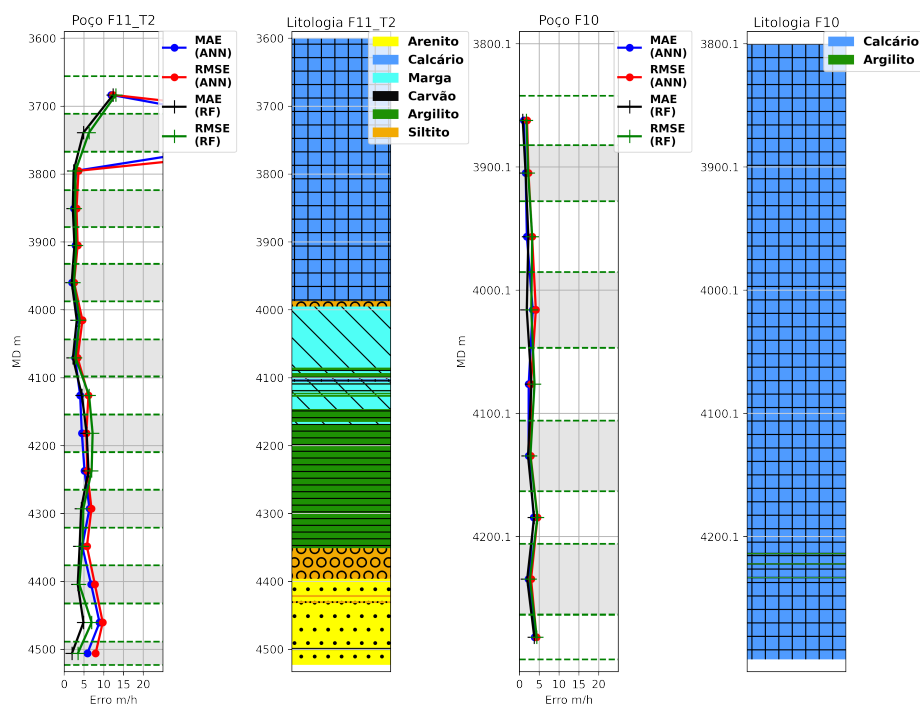
Fonte: Autor (2024)

Figura 65 – Resultados da aprendizagem contínua (Preditores 4) nos poços F15S e F04 ao longo da profundidade em comparação com a litologia dos poços



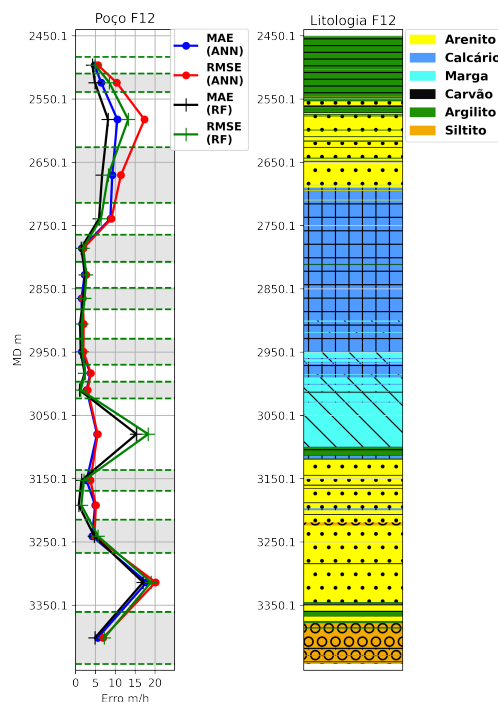
Fonte: Autor (2024)

Figura 66 – Resultados da aprendizagem contínua (Preditores 4) nos poços F11_T2 e F10 ao longo da profundidade em comparação com a litologia dos poços



Fonte: Autor (2024)

Figura 67 – Resultados da aprendizagem contínua (Preditores 4) nos poços F12 ao longo da profundidade em comparação com a litologia dos poços

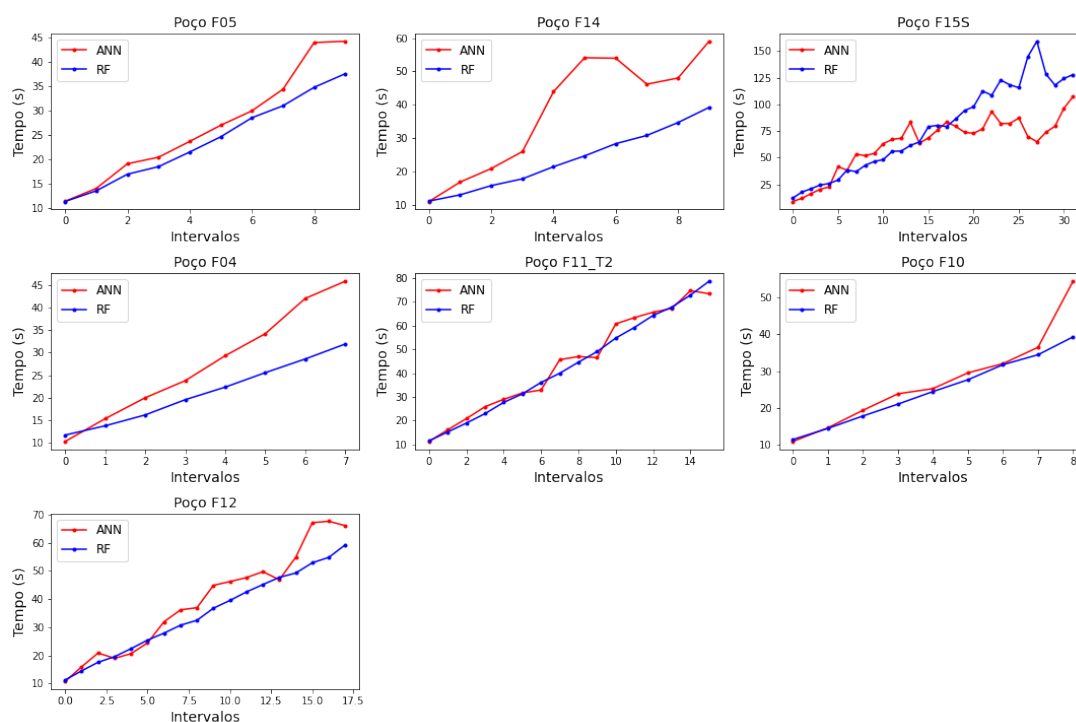


Fonte: Autor (2024)

É possível perceber que a tendência dos resultados dos dois modelos, de modo geral, é

semelhante. Esse fato mostra que ambos estimam o mesmo padrão da ROP. Nos poços F11_T2 e F12 é possível identificar diferenças maiores entre os dois modelos. Para o poço F05, percebe-se que este é o poço que possui maior intercalações entre diferentes litologias. Esse fato ajuda a explicar a variabilidade alta da ROP após 3300 m e, além disso, a maior influência do torque, pois a intercalação entre as formações, com diferentes resistências, favorecem o surgimento de vibrações torcionais (ANDÁS, 2017). No Apêndice D são apresentadas comparações entre a ROP, UCS e a litologia correspondente para todos os poços, onde é possível notar picos, no intervalo mencionado, para o poço F05 (comportamento semelhante no poço F04 e F14). É possível perceber que em regiões de rocha calcária a ROP é menor, e em rochas do tipo argilito a ROP é maior. Esse fato é relacionado com a resistência dessas rochas, em que as rochas calcárias apresentam valores maiores de UCS do que o argilito. Alterações bruscas na ROP devido a mudanças na litologia e à resistência impactam nas previsões de um intervalo para o outro, como é possível verificar no poço F15S (intervalo 2955 m-3055 m e 3555 m - 3655 m). Para o poço F11_T2, a modificação brusca e a pouca quantidade de dados no início do poço resulta em erro elevado para o modelo ANN. A influência das modificações na litologia ajuda a explicar a relevância do Delta-T para a melhoria da capacidade preditiva dos modelos, visto que auxilia a capturar essas variações. Na Figura 66, é possível perceber que o poço F10 possui apenas uma litologia dominante, fato este que também ajuda a explicar a boa resposta dos modelos com apenas informações das variáveis operacionais. A Figura 68 apresenta os tempos de treinamento para os modelos de aprendizagem de máquina.

Figura 68 – Comparação entre os tempos de treinamento na estratégia de aprendizagem contínua (Preditores 4) dos modelos de aprendizagem de máquina



Fonte: Autor (2024)

nos erros podem ser encontradas comparando redes com diferentes hiperparâmetros. No modelo RF, o número mínimo de dados na folha é o mesmo da Tabela 27, no entanto, percebe-se, na Figura 69, que existe variabilidade em relação ao número de árvores e à quantidade de preditores ao longo da profundidade nos poços. No caso do poço F10, onde erros pequenos são obtidos nos dois modelos, diferentes hiperparâmetros são selecionados ao longo do poço. Em geral, observa-se que a relação entre tamanho dos dados e possíveis mudanças sutis na distribuição dos preditores no intervalo posterior, que passam a ser adicionados no conjunto de treinamento, podem dificultar o processo de aprendizagem do padrão presente nesses dados, alterando o nível de complexidade necessário nos modelos para estimar esse padrão.

Por isso, apesar dessa estratégia resultar, na maioria dos casos, em menores erros do que a estratégia todos para um, é necessário definir melhor critérios para a inclusão de novos dados, seleção de hiperparâmetros, assim como a definição do tamanho dos intervalos, para que não resulte em erros significativos, como, por exemplo, no caso do modelo ANN no poço F11_T2.

7 CONCLUSÕES

7.1 Considerações finais

Os estudos desenvolvidos neste trabalho necessitaram de dados de perfuração de poços. Esses dados foram coletados e um *dataset* foi estruturado com informações relevantes que possibilitaram comparar estratégias de predição da ROP que podem, por sua vez, serem utilizadas na otimização da ROP com modelos baseados em dados. Os resultados apresentados com modelos tradicionais e com modelos de aprendizagem de máquina mostram que as informações coletadas são representativas, uma vez que foi possível fazer boas predições da ROP e interpretação da perfuração com os modelos tradicionais, além de informações que foram corroboradas com as experiências relatadas durante a perfuração. Desse modo, os dados estruturados para o presente trabalho podem ser utilizados em estudos futuros para melhorar o desenvolvimento de modelos baseados em dados.

Em relação às técnicas tradicionais para modelagem da ROP, percebe-se que modelos tradicionais podem ser utilizados para interpretar a perfuração e ajudar a selecionar dados para a predição da ROP, quando é feita a análise do ajuste levando em conta as premissas de cada modelo. Nesse sentido, os modelos de regressão linear, como o BYM e o BYMmod, possibilitam maior interpretabilidade, sendo possível analisar a significância dos efeitos relacionados no modelo, que por sua vez adiciona mais informações na análise dos dados históricos.

O BYM é um modelo que consegue capturar bem as interações entre os parâmetros de perfuração e a ROP. No entanto, o modelo não considera diretamente informações relevantes como é o caso da resistência da rocha e o torque. Ambas informações são consideradas em outros modelos, como o de Reza e Alcocer (1986) e o MSE original, e como foi possível perceber nos estudos com os dados dos 7 poços analisados, efeitos importantes na ROP podem ser percebidos com a inclusão de dados sobre a formação geológica e o torque medido. Esses efeitos foram notados na análise de interpretabilidade do modelo BYMmod. As modificações incluídas no BYMmod, para tornar o modelo de BYM original mais representativo para o cenário com brocas PDC, permitiram capturar melhor o padrão da ROP em todos os poços e possibilitaram uma maior compressão de efeitos relevantes na ROP. Foi possível identificar que o modelo MSE pode servir para identificar regiões com aparente problemas de ineficiência na análise dos dados históricos, o que também auxilia na interpretação dessas informações e, possivelmente, auxilia a selecionar dados para o treinamento dos modelos de predição.

Nas estratégias de predição investigadas neste trabalho, e que seriam duas possíveis implementações para otimização de modelos baseados em dados em tempo real, foi identificado que a predição da ROP é possível com modelos de aprendizagem de máquina e que esses modelos, em geral, possuem a vantagem de gerar erros menores para uma maior quantidade de cenários quando comparado com modelos tradicionais. Foi possível identificar que preditores

que carregam informações da formação geológica são relevantes e, nesse caso, a medida de Delta-T possibilita adicionar informações ao modelo que permitiram reduzir erros de forma geral nos poços. Essa medida, em geral, é utilizada para determinar o UCS e também pode ser utilizada na estimativa do gradiente de pressão de poros, possibilitando usar os modelos tradicionais. Além disso, a informação do torque foi necessária para melhorar a predição em certos poços que possuem características como grande alteração litológica e suscetibilidade a vibrações torcionais (*stick-slip*), o que mostrou que essa medida também é relevante para os modelos de aprendizagem de máquina nos poços investigados. No entanto, também foi possível verificar que, em casos sem grandes alterações litológicas, é possível fazer predições do ROP adequadas considerando apenas os dados operacionais principais.

Nota-se que os métodos de aprendizagem de máquina são capazes de capturar padrões em dados de uma grande quantidade de poços. No entanto, foi possível identificar que a grande quantidade de dados que é utilizada na estratégia de aprendizagem com poços de correlação pode resultar em diferenças maiores em certos trechos do poço de teste. Além disso, certos poços possuem características particulares que tornaram a predição da ROP problemática ao utilizar essa estratégia de treinamento. Nesse sentido, foi possível concluir que utilizar uma grande quantidade de dados para o treinamento dos modelos de aprendizagem de máquina, nessa estratégia, não necessariamente melhora a performance desses modelos. No entanto, constatou-se que a aprendizagem com poços de correlação é possível, e capaz de gerar predições adequadas, se houver dados representativos, como ficou constatado ao utilizar uma parcela dos dados para melhorar a predição no poço problemático.

A estratégia de aprendizagem contínua resulta em erros menores quando comparada com a estratégia de predição com dados de poços de correlação, levando em conta os dados e casos empregados neste trabalho. No entanto, foi possível perceber que ao analisar as predições, ao longo de cada poço, existem trechos onde os erros aumentam significativamente. Esse aumento dos erros ocorreu em maior parte nas regiões de alteração litológica, por conta de diferenças significativas na resistência, junto com casos onde há poucos dados utilizados no treinamento. Essa diminuição da performance preditiva pode estar relacionada com a inclusão de novos dados, ao conjunto de treinamento, que possuem uma distribuição estatística distinta, de modo que a estimativa do padrão presente nos dados agregados é dificultada. Portanto, é necessário investigar abordagens para melhorar ainda mais as predições, evitando erros significativos que podem ocorrer, seja definindo melhor critérios para o treinamento desses modelos ou definindo melhor os dados utilizados nesse processo.

Em relação à performance dos modelos de aprendizagem de máquina analisados, foi possível detectar que o RF se destacou, com erros menores e, no geral, com custo computacional menor. Essa conclusão foi perceptível nas duas metodologias de predição da ROP analisadas. O refino dos hiperparâmetros se mostrou necessário uma vez que a performance é melhorada com a inserção dessa metodologia. Observou-se também que o modelo ANN apresentou maior

modificação de hiperparâmetros ótimos nos casos investigados sendo, portanto, um modelo mais sensível as modificações nos seus hiperparâmetros.

7.2 Sugestões de trabalhos futuros

Como sugestão de trabalhos futuros pode-se sugerir os seguintes tópicos:

- Utilizar o *dataset* para a avaliação de outros modelos preditivos para ROP;
- Realizar estudo comparando os parâmetros operacionais ótimos obtidos com os modelos de aprendizagem de máquina e com estratégias tradicionais;
- Continuar o aperfeiçoamento das estratégias preditivas inserindo novos preditores, definindo novas combinações de preditores que permitam melhorar ainda mais a performance dos modelos de aprendizagem de máquina;
- Avaliar performance de outras arquiteturas de ANNs, definindo melhores hiperparâmetros para melhorar a predição com as estratégias todos para um e aprendizagem contínua;
- Definir critérios para selecionar dados de poços de correlação para melhorar a performance dos modelos preditivos para ROP;
- Realização de estudos para definir critérios de treinamento durante a aprendizagem contínua que possibilitem reduzir ainda mais os erros;
- Usar o *dataset* para desenvolver pesquisa com modelos baseados em dados com aprendizagem de máquina para predição de outras variáveis de interesse em tempo real, como litologia e gradiente de poro pressão.

REFERÊNCIAS

- ABU-MOSTAFA, Y. S.; MAGDON-ISMAIL, M.; LIN, H.-T. *Learning from Data: A Short Course*. 1. ed. Pasadena: AMLBook, 2012.
- AHMED, O.; ADENIRAN, A.; SAMSURI, A. Rate of penetration prediction utilizing hydromechanical specific energy. In: . [S.l.]: InTech, 2018.
- ALAWAD, M. N. J. A rock mechanical model for overbalanced, managed pressure, and underbalanced drilling applications wellbore integrity and fracture seal materials view project. *International Journal of Rock Mechanics and Geomechanics for Gas, Petroleum and Water*, v. 1, p. 103, 2016. Disponível em: <<http://ijrmg.upg-ploiesti.ro>>.
- ALSENWAR, M. *NCS Drilling Data Based ROP Modelling and its Application*. Dissertação (Mestrado) — University of Stavanger, Noruega, 2017.
- AMADI, K.; IYALLA, I. Application of mechanical specific energy techniques in reducing drilling cost in deepwater development. In: SPE. *Deepwater Drilling and Completions Conference 2012*. Galveston, Texas, 2012. p. 626–635. 10.2118/156370-MS.
- ANDÁS, E. *Mitigating Stick-Slip Vibrations Using Surface Control Software: A Validation of SoftSpeed II Using High-Speed Along-String Dynamics Measurements in a Norwegian Offshore Field*. Dissertação (Mestrado) — University of Stavanger, Noruega, 2017.
- ANP. *Agência Nacional do Petróleo Gás Natural e Biocombustíveis (Prêmio ANP de Inovação Tecnológica 2023)*. 2023. Disponível em: <<https://www.gov.br/anp/pt-br/assuntos/pesquisa-desenvolvimento-e-inovacao/premio-anp-de-inovacao-tecnologica/premio-anp-inovacao-tecnologica-2023>>. Acesso em: 30 Jan. 2024.
- ANP. *Agência Nacional do Petróleo Gás Natural e Biocombustíveis (Boletim Mensal da Produção de Petróleo e Gás Natural)*. 2024. Disponível em: <<https://www.gov.br/anp/pt-br/centrais-de-conteudo/publicacoes/boletins-anp/boletins-boletim-mensal-da-producao-de-petroleo-e-gas-natural>>. Acesso em: 30 Jan. 2024.
- ANP. *Agência Nacional do Petróleo Gás Natural e Biocombustíveis (Painel Dinâmico de Produção de Petróleo e Gás Natural)*. 2024. Disponível em: <https://www.gov.br/anp/pt-br/canais_atendimento/imprensa/noticias-comunicados/investimentos-em-exploracao-podem-chegar-a-us-1-96-bilhao-em-2024>. Acesso em: 30 Jan. 2024.
- AOUN, M. A. B.; MADARÁSZ, T. Applying machine learning to predict the rate of penetration for geothermal drilling located in the utah forge site. *Energies*, v. 15, n. 12, 2022. ISSN 1996-1073. Disponível em: <<https://www.mdpi.com/1996-1073/15/12/4288>>.
- ARANHA, P. E.; LOPES, L. G. O.; SOBRINHO, E. S. P.; OLIVEIRA, I. M. N.; ARAÚJO, J. P. N. de; SANTOS, B. B.; JÚNIOR, E. T. L.; SILVA, T. B. da; VIEIRA, T. M. A.; LIRA, W. W. M.; POLICARPO, N. A.; SAMPAIO, M. A. A system to detect oilwell anomalies using deep learning and decision diagram dual approach. *SPE Journal*, Society of Petroleum Engineers (SPE), p. 1–14, nov. 2023. ISSN 1930-0220. Disponível em: <<http://dx.doi.org/10.2118/218017-PA>>.

BARBOSA, L. F. F. M. *Drilling optimization of petroleum and natural gas wells: application of artificial intelligence*. Dissertação (Mestrado) — Universidade Estadual Paulista, São Paulo, 2019.

BARBOSA, L. F. F. M.; NASCIMENTO, A.; MATHIAS, M. H.; CARVALHO, J. A. de. Machine learning methods applied to drilling rate of penetration prediction and optimization - a review. *Journal of Petroleum Science and Engineering*, Elsevier B.V., v. 183, 12 2019. ISSN 09204105.

BELASKIE, J. P.; DUNN, M. D.; CHOO, D. K. Distinct applications of MWD, weight on bit, and torque. *SPE Drilling & Completion*, v. 8, p. 111–117, 6 1993. ISSN 1064-6671.

BELAYNEH, M. New alternative MSE based ROP modelling and analysis with north sea field data. *International Journal of Engineering and Technology*, v. 12, n. 12, 2019. ISSN 0974-3154.

BIELSTEIN, W. J.; CANNON, G. E. Factors affecting the rate of penetration of rock bits. *Spring Meeting South-Western District Division of Production*, Dallas, Texas, 1950.

BILGESU, H.; TETRICK, L. T.; ALTMIS, U.; MOHAGHEGH, S.; AMERI, S. A new approach for the prediction of rate of penetration (ROP) values. In: *SPE Eastern Regional Meeting*. Lexington, Kentucky: [s.n.], 1997. 10.2118/39231-MS.

BINGHAM, M. G. Rock drillability: How rock properties are related to drilling. *The Oil and Gas Journal*, p. 94–101, 1964.

BOTCHKAREV, A. Evaluating performance of regression machine learning models using multiple error metrics in azure machine learning studio. *SSRN Electronic Journal*, 2018. ISSN 1556-5068.

BOUKREDERA, F. S.; YUCEFI, M. R.; HADJADJ, A.; EZENKWU, C. P.; VAZIRI, V.; APHALE, S. S. Enhancing the drilling efficiency through the application of machine learning and optimization algorithm. *Engineering Applications of Artificial Intelligence*, Elsevier BV, v. 126, p. 107035, nov. 2023. ISSN 0952-1976. Disponível em: <<http://dx.doi.org/10.1016/j.engappai.2023.107035>>.

BOURGOYNE, A.; YOUNG, F. A multiple regression approach to optimal drilling and abnormal pressure detection. *Society of Petroleum Engineers Journal*, v. 14, p. 371–384, 8 1974. ISSN 0197-7520.

BOURGOYNE, A. T.; MILLHEIM, K. K.; CHENEVERT, M. E.; JR, F. S. Y. *Applied Drilling Engineering*. United States of America: SPE Textbook Series, 1986.

BREIMAN, L. Random forest. *Machine Learning*, v. 45, p. 5–32, 2001. ISSN 08856125.

BRUCE, P.; BRUCE, A.; GEDECK, P. *Practical Statistics for Data Scientists*. Toronto: Wiley, 2020.

CARRAPATOSO, C. M. *Análise dos modelos analíticos de otimização de perfuração baseados em energia específica para formações evaporíticas*. Dissertação (Mestrado) — Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2011.

CHANG, C.; ZOBACK, M. D.; KHAKSAR, A. Empirical relations between rock strength and physical properties in sedimentary rocks. *Journal of Petroleum Science and Engineering*, v. 51, n. 3, p. 223–237, 2006. ISSN 0920-4105. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S0920410506000167>>.

CHEN, X.; FAN, H.; GUO, B.; GAO, D.; WEI, H.; YE, Z. Real-time prediction and optimization of drilling performance based on a new mechanical specific energy model. *Arabian Journal for Science and Engineering*, Springer Verlag, v. 39, p. 8221–8231, 10 2014. ISSN 21914281.

CHEN, X.; YANG, J.; GAO, D. Drilling performance optimization based on mechanical specific energy technologies. In: _____. IntechOpen, 2018. Disponível em: <<https://doi.org/10.5772/intechopen.75827>>.

CONTESSI, B. A.; FONTES, C. E.; SILVA, F. R. G. D.; FERRAZ, R.; CAMPAGNAC, L. A.; DAOUD, I. L. A. Simulation-based digital twin for predictive maintenance. *Rio Oil and Gas Expo and Conference*, Instituto Brasileiro de Petróleo e Gas, v. 22, n. 2022, p. 315–316, set. 2022. ISSN 2525-7579. Disponível em: <<http://dx.doi.org/10.48072/2525-7579.rog.2022.315>>.

CUNNINGHAM, R. A. *Laboratory Studies of the Effect of Rotary Speed on Rock-bit Performance and Drilling Cost*. American Petroleum Institute. New York, 1960.

DARWESH, A. K.; RASMUSSEN, T. M.; AL-ANSARI, N. Controllable drilling parameter optimization for roller cone and polycrystalline diamond bits. *Journal of Petroleum Exploration and Production Technology*, v. 10, 2020.

ELMGERBI, A. M.; ETTINGER, C. P.; TEKUM, P. M.; THONHAUSER, G. Application of machine learning techniques for real time rate of penetration optimization. In: . [s.n.], 2021. v. 17, p. 25–27. Disponível em: <<http://onepetro.org/SPEMEDT/proceedings-pdf/20MEDT/2-20MEDT/D021S007R003/2444379/spe-202184-ms.pdf/1>>.

EQUINOR. *Disclosing all Volve data*. 2018. Disponível em: <<https://www.equinor.com/news/archive/14jun2018-disclosing-volve-data>>. Acesso em: 09 de Jul. 2022.

EREN, T.; OZBAYOGLU, M. E. Real time optimization of drilling parameters during drilling operations. *SPE Oil and Gas Índia Conference and Exhibition*, 2010.

EVSUKOFF, A. G. *Inteligência Computacional: Fundamentos e aplicações*. Rio de Janeiro: E-papers, 2020. ISBN 978-65-8706-502-1.

FEAR, M. J. How to improve rate of penetration in field operations. *SPE Drill & Completion*, 1999.

FERRO, A. P. A.; BARBOZA, A. da S. R.; GOUVEIA, L. P. de. Performance analysis of random forest and bourgoyne and young models in rate of penetration prediction. In: *Encontro Nacional de Construção de Poços de Petróleo e Gás - ENAHPE*. Matinhos, PR: [s.n.], 2023.

FONTANA, E. *Introdução aos Algoritmos de Aprendizagem Supervisionada*. Curitiba, PR, 2020. Acesso em: 14 out. 2023. Disponível em: <https://fontana.paginas.ufsc.br/files/2018/03/apostila_ML_pt2.pdf>.

FRANKIEWICZ, J. *The Application of Data Analytics and Machine Learning for Formation Classification and Bit Dull Grading Prediction*. Dissertação (Mestrado) — University of Stavanger, Noruega, 2019.

GALLE, E.; WOODS, H. Best constant weight and rotary speed for rotary rock bits. *Drilling and Production Practice*, New York, USA, 1 1963.

GAMAL, H.; ELKATATNY, S.; ADEBAYO, A. The role of overbalance pressure on mud induced alteration of sandstone rock pore system. *Scientific Reports*, Springer Science and Business Media LLC, v. 12, 12 2022.

GANDELMAN, R. A. *Predição da ROP e Otimização em Tempo Real de Parâmetros Operacionais na Perfuração de Poços*. Dissertação (Mestrado) — Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2012.

GEOFFROY, H.; MINH, D. N.; MAITOURNAM, H.; BERGUES, J.; PUTOT, C. Evaluation of drilling parameters of a PDC bit. In: _____. *Advances in Rock Mechanics*. WORLD SCIENTIFIC, 1998. p. 301–314. ISBN 9789812839640. Disponível em: <http://dx.doi.org/10.1142/9789812839640_0028>.

GÉRON, A. *Mãos à Obra: Aprendizado de Máquina com Scikit-Learn, Keras & TensorFlow*. 2. ed. [S.l.]: Alta Books, 2021. ISBN 9788550815480.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016.

GOULAT, J. D. V. *Dinâmica de uma coluna de perfuração utilizando a teoria de Cosserat*. Dissertação (Mestrado) — Pontifícia Universidade Católica, Rio de Janeiro, 2019.

GRAHAM, J.; MUENCH, N. Analytical determination of optimum bit weight and rotary speed combinations. In: . [S.l.]: SPE, 1959.

GUAN, Z.; CHEN, T.; LIAO, H. *Theory and Technology of Drilling Engineering*. [S.l.]: Springer Singapore, 2021. ISBN 978-981-15-9326-0.

HAMMOUTENE, C. Fea modelled mse/ucs values optimise pdc design for entire hole section. *Society of Petroleum Engineers - North Africa Technical Conference and Exhibition 2012, NATC 2012: Managing Hydrocarbon Resources in a Changing Environment*, v. 1, p. 66–76, 2012.

HEGDE, C.; DAIGLE, H.; GRAY, K. E. Performance comparison of algorithms for real-time rate-of-penetration optimization in drilling using data-driven models. *SPE Journal*, Society of Petroleum Engineers (SPE), v. 23, p. 1706–1722, 10 2018. ISSN 1086055X.

HEGDE, C.; GRAY, K. Evaluation of coupled machine learning models for drilling optimization. *Journal of Natural Gas Science and Engineering*, Elsevier B.V., v. 56, p. 397–407, 8 2018. ISSN 18755100.

HÖHN, P.; ODEBRETT, F.; SHAHID, K.; PAZ, C.; OPPELT, J. Framework for automated generation of real-time rate of penetration models. *Journal of Petroleum Science and Engineering*, Elsevier BV, v. 213, p. 110369, 6 2022. ISSN 09204105.

HORSRUD, P. Estimating mechanical properties of shale from empirical correlations. *SPE Drilling & Completion*, v. 16, n. 02, p. 68–73, 2001. Disponível em: <<https://doi.org/10.2118/56017-PA>>.

HOSSAIN, M. E.; AL-MEJED, A. A. *Fundamentals of Sustainable Drilling Engineering*. [S.l.]: Scrivener Publishing, 2015.

HOSSAIN, M. E.; ISLAM, M. R. *Drilling Engineering Problems and Solutions: A Field Guide for Engineers and Students*. [S.l.]: Wiley, 2018. ISBN 9781118998342.

ISLAM, M. R.; HOSSAIN, M. E. *Drilling Engineering: Towards Achieving Total Sustainability*. [S.l.]: Elsevier, 2021. ISBN 9780128201930.

IZBICKI, R.; SANTOS, T. M. dos. *Aprendizado de máquina: uma abordagem estatística*. [S.l.: s.n.], 2020. ISBN 9786500024104.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. *An Introduction to Statistical Learning: with Applications in R*. [S.l.]: Springer New York, 2014.

KAKOLI, M.; EBRAHIMABADI, A.; MIRSHEKARI, B.; SHEIKHZAKARIAEE, J. A model to predict the performance of drilling operation using svm approach in one of the iranian oilfields. *International Journal of Petroleum and Petrochemical Engineering*, v. 4, 2018.

KOR, K.; ALTUN, G. Is support vector regression method suitable for predicting rate of penetration? *Journal of Petroleum Science and Engineering*, Elsevier B.V., v. 194, 11 2020. ISSN 09204105.

KUNG, S. de C. *Previsão de taxas de penetração na perfuração de poços de petróleo com brocas de cortadores fixos*. Dissertação (Mestrado) — Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2016.

LEITE, V. R. C. *Uma análise da classificação de litologias utilizando SVM, MLP e métodos Ensemble*. Dissertação (Mestrado) — Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2012.

MA, T.; CHEN, P.; ZHAO, J. Overview on vertical and directional drilling technologies for the exploration and exploitation of deep petroleum resources. *Geomechanics and Geophysics for Geo-Energy and Geo-Resources*, v. 2, p. 365–395, 12 2016. ISSN 2363-8419.

MANTOVANI, R. G.; HORVÁTH, T.; CERRI, R.; JUNIOR, S. B.; VANSCHOREN, J.; CARVALHO, A. C. P. d. L. F. de. An empirical study on hyperparameter tuning of decision trees. 2018.

MARCHESI, V. R. *Técnicas De Modelagem Aplicadas à Previsão De Pressão De Poros Em Ambientes Geologicamente Complexos*. Tese (Doutorado) — Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2015.

MAURER, W. The "perfect-cleaning" theory of rotary drilling. *Journal of Petroleum Technology*, v. 14, p. 1270–1274, 1962.

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, v. 5, p. 115–133, 12 1943. ISSN 0007-4985.

MCNALLY, G. H. Estimation of coal measures rock strength using sonic and neutron logs. *Geoexploration*, v. 24, n. 4, p. 381–395, 1987. Disponível em: <<https://www.sciencedirect.com/science/article/pii/0016714287900081>>.

MITA, G. *Toward Interpretable Machine Learning, with Applications to Large-scale Industrial Systems Data*. Tese (Doutorado) — Sorbonne University, França, 2021.

MITCHELL, R. F.; MISKA, S. Z. et al. *Fundamentals of drilling engineering*. [S.l.]: Society of Petroleum Engineers, 2011.

MITCHELL, T. M. *Machine Learning*. [S.l.]: McGraw-Hill Science, 1997. ISBN 0070428077.

- MOAZZENI, A. R.; KHAMEHCHI, E. A new hybrid analytical-machine learning method for real-time rop modeling. *Biointerface Research in Applied Chemistry*, AMG Transcend Association, v. 11, p. 7589–7604, 1 2021. ISSN 20695837.
- MOHAMUD, Y. I. H. *ROP Modelling, Analysis and Optimization Study Based on Morvin Field 6506/11 Drilling Data*. Dissertação (Mestrado) — University of Stavanger, Noruega, 7 2020.
- MOLNAR, C. *Practical Wellbore Hydraulics and Hole Cleaning*. [S.l.: s.n.], 2019.
- MOLNAR, C. *Interpretable Machine Learning: A guide for making black box models explainable*. 2. ed. [s.n.], 2022. Acessado em 30 de Agosto de 2022. Disponível em: <<https://christophm.github.io/interpretable-ml-book/>>.
- MONTGOMERY, D. C.; RUNGER, G. C. *Estatística Aplicada e Probabilidade para Engenharia*. 6. ed. [S.l.]: LTC, 2016.
- MURPHY, K. *Probabilistic Machine Learning: An Introduction*. [S.l.]: MIT Press, 2022. ISBN 9780262046824.
- NAGY, A. *Availability and Quality of Drilling Data in the Volve Dataset*. Master thesis — Faculty of Science and Technology, 2019.
- NAJJARPOUR, M.; JALALIFAR, H.; NOROUZI-APOURVARI, S. A review on half a century of experience in rate of penetration management: Application of analytical, semi-analytical and empirical models. *Advances in Geo-Energy Research*, Yandy Scientific Press, v. 5, p. 252–273, 2021. ISSN 2208598X.
- NASCIMENTO, A.; KUTAS, D. T.; ELMGERBI, A.; THONHAUSER, G.; MATHIAS, M. H. Mathematical modeling applied to drilling engineering: An application of bourgoyne and young ROP model to a presalt case study. *Mathematical Problems in Engineering*, Hindawi Publishing Corporation, v. 2015, 2015. ISSN 15635147.
- NASSER, P. A. S. *Real Time Rate of Penetration, Prediction and Optimization During Drilling Operations*. Dissertação (Mestrado) — Universiti Teknologi Petronas, Malásia, 2012.
- NOSHI, C. I.; SCHUBERT, J. J. The role of machine learning in drilling operations; a review introduction to data analytics. In: . [S.l.: s.n.], 2018.
- OSGOUEI, R. E. *Rate of Penetration Estimation Model for Directional and Horizontal Wells*. Tese (Doutorado) — Middle East Technical University, Turquia, 2007.
- PAYETTE, G. S.; SPIVEY, B. J.; WANG, L.; BAILEY, J. R.; SANDERSON, D.; KONG, R.; PAWSON, M.; EDDY, A. A real-time well-site based surveillance and optimization platform for drilling: Technology, basic workflows and field results. In: . [S.l.]: SPE, 2017.
- PESSIER, R. C.; FEAR, M. J. Quantifying common drilling problems with mechanical specific energy and a bit-specific coefficient of sliding friction. *Proceedings - SPE Annual Technical Conference and Exhibition*, OnePetro, Delta, p. 373–388, 10 1992.
- PETROBRAS. *Tecnologias pioneiras do pré-sal*. 2024. Disponível em: <<https://nossaenergia.petrobras.com.br/energia/tecnologias-pioneiras-do-pre-sal/>>. Acesso em: 30 Jan. 2024.

PROBST, P. *Hyperparameters, Tuning and Meta-Learning for Random Forest and Other Machine Learning Algorithms*. Tese (Doutorado) — Ludwig Maximilian University of Munich, Alemanha, 2019.

RAMSEY, P. E. M. S. *Practical Wellbore Hydraulics and Hole Cleaning*. [S.l.]: Gulf Professional, 2019. ISBN 9780128170885.

RASCHKA, S.; MIRJALILI, V. *Python Machine Learning, 2nd Ed.* 2. ed. Birmingham, UK: Packt Publishing, 2017. ISBN 978-1787125933.

REZA, M. R.; ALCOCER, C. F. A unique computer simulation model well drilling: Part i—the reza drilling model. In: . [s.n.], 1986. (SPE Western Regional Meeting, All Days), p. SPE-15108-MS. Disponível em: <<https://doi.org/10.2118/15108-MS>>.

ROBINSON, T. S.; BATRUNY, P.; GOMES, D.; HASHIM, M. M. H. M.; YUSOFF, M. H.; ARRIFFIN, M. F.; MOHAMAD, A. Successful development and deployment of a global rop optimization machine learning model. In: *Day 2 Wed, March 23, 2022*. OTC, 2022. (22OTCA). Disponível em: <<http://dx.doi.org/10.4043/31680-MS>>.

ROCHA, L. A. S.; AZEVEDO, C. T. d. Projetos de poços de petróleo: geopressões e assentamento de colunas de revestimentos. *Interciência, Rio de Janeiro*, v. 511, 2009.

ROCHA, L. A. S.; AZUAGA, D.; ANDRADE, R.; VIEIRA, J. L.; SANTOS, O. L. A. *Perfuração direcional*. 3ª edição. ed. [S.l.]: Editora Interciência, RJ, 2011.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, v. 65, 1958.

SIKES, S. T. Drilling speeds. *API Drilling and Production Practice*, p. 54-59, 1936.

SILVA, E. A. B.; FERRO, A. P. A.; JÚNIOR, F. de A. V. B.; GOUVEIA, L. P. de; BARBOZA, A. da S. R. Estudo de caso da estimativa da resistência à compressão confinada de rochas com base em dados geofísicos e parâmetros operacionais. In: *Encontro Nacional de Construção de Poços de Petróleo e Gás - ENAHPE*. Matinhos, PR: [s.n.], 2023.

SILVA, F. R. G. da; CRUZ, M. de S.; BARDUCHI, B.; BELLUMAT, E.; SANTOS, M. Vieira dos; MATOS, V. Barroso de; GANDELMAN, R. A.; MARTINS, A. L. Six years operating a real-time drilling problem detection software in deepwater environments: Results and challenges. In: *Day 4 Thu, July 30, 2020*. SPE, 2020. (19LACP). Disponível em: <<http://dx.doi.org/10.2118/199077-MS>>.

SILVA, I. N. da; SPATTI, D. H.; FLAUZINO, R. A. *Redes Neurais Artificiais Para Engenharia e Ciências Aplicadas*. 2. ed. São Paulo: Artliber, 2016.

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 B.V., v. 34, p. 1225–1236, 8 2016. ISSN 18755100.

SOARES, C.; GRAY, K. Real-time predictive capabilities of analytical and machine learning rate of penetration (ROP) models. *Journal of Petroleum Science and Engineering*, Elsevier B.V., v. 172, p. 934–959, 1 2019. ISSN 09204105.

SOBRINHO, E. da S. P.; OLIVEIRA, F. L. D.; ANJOS, J. L. R. D.; SILVA, T. B. D.; LOPES, L. G. O.; ARAÚJO, J. P. N. D.; GONÇALVES, C.; FERREIRA, M. V. D.; GOUVEIA, L. P. D.; LIRA, W. W. M. A framework to detect production anomalies using deep learning and decision tree. *Rio Oil and Gas Expo and Conference*, v. 20, p. 437–438, 12 2020. ISSN 25257579.

SOUZA, L. Z. *Estudo do Estado da Arte da Perfuração Direcional*. Monografia (Graduação em Engenharia Mecânica) — Universidade Federal do Espírito Santo, Vitória, 2011.

TAVARES, R. M. *Interpretação e Análise de Dados de Perfuração em Poços de Petróleo*. Dissertação (Mestrado) — Universidade Estadual de Campinas, São Paulo, 2006.

TEALE, R. The concept of specific energy in rock drilling. *International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts*, Pergamon, v. 2, p. 57–73, 3 1965. ISSN 0148-9062.

THOMAS, J. E. et al. *Fundamentos de engenharia de petróleo*. 2ª edição. ed. [S.l.]: Editora Interciência, RJ, 2004.

TUNKIEL, A. T.; SUI, D.; WIKTORSKI, T. Reference dataset for rate of penetration benchmarking. *Journal of Petroleum Science and Engineering*, Elsevier B.V., v. 196, 1 2021. ISSN 09204105.

TUNKIEL, A. T.; WIKTORSKI, T.; SUI, D. Drilling dataset exploration, processing and interpretation using volve field data. In: *Volume 11: Petroleum Technology*. American Society of Mechanical Engineers, 2020. (OMAE2020). Disponível em: <<http://dx.doi.org/10.1115/OMAE2020-18151>>.

VEGAR, P.; ØYVIND, B.; TVEIT, S. *Model for evaluating drilling efficiency based on the concept of Mechanical Specific Energy*. Dissertação (Mestrado) — Norwegian University of Science and Technology, Noruega, 2016.

XUE, Q. *Data Analytics for Drilling Engineering*. [S.l.]: Springer International Publishing, 2020. ISBN 978-3-030-34034-6.

APÊNDICE A – GRADIENTE DE PRESSÃO DE POROS

A pressão de poros é a pressão do fluido contido nos poros da rocha. Segundo Rocha e Azevedo (2009), o gradiente de pressão de poros, do inglês *Pore Pressure Gradient* (PPG), é a razão entre a pressão de poros e a respectiva profundidade. Sua importância está na determinação do peso do fluido ideal a ser utilizado na perfuração que, por sua vez, exercerá pressão hidrostática adequada para manter a estabilidade do poço.

Como mencionado em Marchesi (2015), a medição direta da pressão de poros pode ser feita com equipamentos que são empregados nos métodos *Repeated Formation Test* (RFT) e *Measured Direct Test* (MDT). No entanto, como mencionado em Rocha e Azevedo (2009), as medições diretas são registros pontuais, não sendo viável a determinação de curvas ao longo do poço por meio de medições diretas. O método de Eaton é o método indireto mais utilizado para estimativa da pressão de poros, sendo possível utilizar para determinar os perfis do poço (MARCHESI, 2015). Esse método relaciona o gradiente de pressão de poros com o gradiente de sobrecarga e dados de perfilagem geofísica, geralmente resistividade ou sônico. A expressão é dada por

$$G_{pp} = G_s - \left[(G_s - G_N) \cdot \left(\frac{\Delta t_n}{\Delta t_0} \right)^m \right] \quad (\text{A.1})$$

na qual G_N é uma constante chamada de gradiente de pressão de poros normal, cujo valor fica entre 8,5 lb/gal e 9,0 lb/gal, mas geralmente é considerada igual a 8,5 lb/gal. Para a constante m são utilizados os valores: 1, 2; 2, 0 ou 3, 0. A escolha desses valores pode ser feita com medidas diretas e informações de estudos sobre a região analisada. Δt_o é a medição do perfil sônico, podendo também ser utilizado a medida de resistividade. Δt_n é o valor da reta de tendência normal, que representa a tendência de compactação da formação e é determinada aproximando uma reta nas medições Δt_o . Por fim, G_s é o gradiente de sobrecarga, que consiste na relação entre a pressão de sobrecarga, devido ao peso de todas as camadas de rocha sobrepostas, e a profundidade vertical verdadeira. G_s é determinado pela seguinte expressão:

$$G_s = \frac{\sigma_{ov}}{0,1704 \cdot D} \quad (\text{A.2})$$

na qual D é a profundidade vertical verdadeira e σ_{ov} é a pressão de sobrecarga expressa em psi, que é convertido em lb/gal pela Equação A.2. A pressão σ_{ov} é dada por

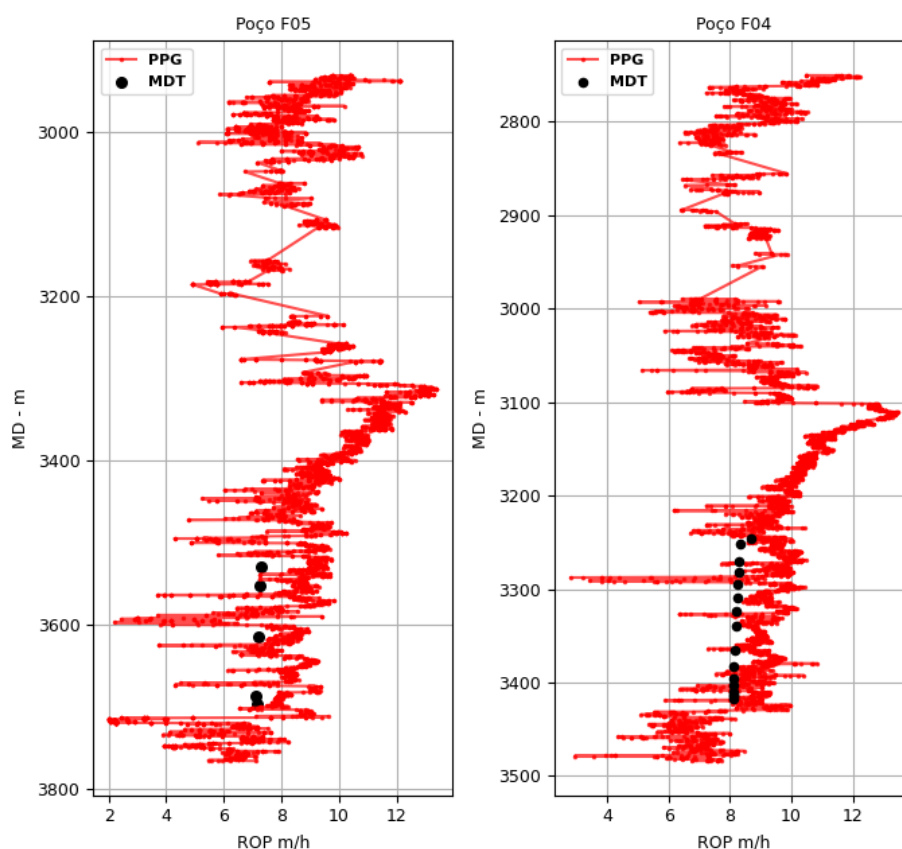
$$\sigma_{ov} = 1.422 \cdot \left(\rho_w \cdot D_w + \sum_{bi}^n \cdot \rho_{bi} \cdot \Delta D_i \right) \quad (\text{A.3})$$

Nessa expressão, D_w é a lâmina d'água em m, os valores de ΔD_i são os intervalos de profundidade para n camadas. ρ_w é a densidade da água do mar e ρ_{bi} é a densidade de

cada camada da formação, ambos em g/cm^3 . Utilizando a Equação A.3, obtêm-se σ_{ov} em psi . Os valores de densidade das rochas podem ser determinados por correlações empíricas ou diretamente utilizando medição direta de densidade (perfil *Bulk Density*).

Para a determinação do gradiente de poro pressão nos poços do campo *Volve* utilizou-se $D_w = 91\text{ m}$, tendo como base informações de relatórios técnicos desses poços. A densidade da água do mar utilizada é $\rho_w = 1,02\text{ g/cm}^3$. O perfil de densidade medida foi utilizado para todos os poços. Para as profundidades iniciais do poço, em que as rochas são mais porosas, Rocha e Azevedo (2009) mencionam que os valores de densidade variam de 1,5 a 1,95 g/cm^3 . Desse modo, considerou-se para os primeiros 50 m o valor de $\rho_{bi} = 1,5\text{ g/cm}^3$ e para o intervalo após os primeiros 50 m e o início das medições de densidade foi utilizado $\rho_{bi} = 1,95\text{ g/cm}^3$. Esses valores foram definidos considerando análise do comportamento do perfil de densidade. Por fim, foi utilizado $m = 1,2$ na Equação A.1, tendo como base a comparação com medições diretas pelo método MDT disponível em alguns poços. A Figura 70 ilustra essa comparação.

Figura 70 – Comparação entre gradiente de pressão de poros de dois poços e medições pontuais pelo método MDT



Fonte: Autor (2024)

APÊNDICE B – ATRIBUTOS DAS BROCAS

A Tabela 32 apresenta os dados das brocas utilizadas em cada poço analisado neste trabalho. Essas informações incluem o *Dull grading*, que é a avaliação da broca após o fim da perfuração, detalhado na Figura 71, e alguns atributos importantes como a *Total Flow Area* (TFA), o seu tipo e diâmetro. As fases de cada poço são delimitadas pela profundidade medida inicial e final (MD Inicial e MD Final respectivamente) correspondente ao uso de cada broca.

Tabela 32 – Informações extraídas sobre as brocas que são utilizadas nos poços analisados

Poço	<i>Dull grading</i>	Tipo	TFA (in ²)	Diâmetro (in)	MD Inicial (m)	MD Final (m)
F05	1-3-WT-G-X-IN-BF-BHA	PDC	0,70	8,50	2931,17	3279,83
	1-1-WT-N-X-IN-CT-TD	PDC	0,70	8,50	3293,08	3765,91
F14	1-2-CT-S-X-I-WT-TD	PDC	1,05	12,25	2780,28	2787,92
	1-1-WT-A-X-IN-NO-TD	PDC	0,70	8,50	2788,16	3465,85
F15S	1-2-WT-S-IN-X-CT-BHA	PDC	0,72	8,50	2554,83	3669,85
	1-1-WT-A-IN-X-NO-TD	PDC	0,51	8,50	3670,08	4062,98
F04	4-3-CT-N-X-IN-WT-TD	PDC	1,11	12,25	2750,06	2770,00
	1-1-NO-A-X-IN-NO-BHA	PDC	0,58	8,50	2770,02	2993,00
	1-1-WT-A-X-I-CT-TD	PDC	0,58	8,50	2993,14	3483,86
F11 T2	0-0-NO-A-X-I-NO-TD	PDC	0,95	8,50	3600,00	4522,90
F10	1-1-NO-A-X-IN-NO-BHA	PDC	1,04	8,50	3442,00	4299,73
	4-4-BT-A-X-IN-CT-PR	PDC	1,45	17,50	2450,14	2512,96
F12	1-1-WT-A-E-IN-WT-TD	RCB (Dentes de aço)	1,42	17,50	2513,07	2573,00
	1-1-CT-A-X-IN-WT-TD	PDC	1,11	12,25	2573,73	3113,96
	0-1-WT-S-X-IN-NO-TD	PDC	0,66	8,50	3114,41	3442,75

Fonte: Autor (2024)

Alguns atributos relevantes que são relacionados às brocas da Tabela 32, e que são necessárias nos modelos tradicionais discutidos neste trabalho, são: a perda de pressão na broca (ΔP_b); a potência hidráulica por polegada quadrada, do inglês *Hydraulic Horsepower per Square Inch* (HSI) e a Força hidráulica de impacto do jato (F_j).

A perda de pressão na broca é devido a mudança da velocidade do fluido nos orifícios da broca. ΔP_b é dado por

$$\Delta P_b = \frac{q^2 \cdot \rho}{12031 \cdot TFA^2} \quad (\text{B.1})$$

em que q é vazão do fluido de perfuração em *gal/min*, ρ a densidade do fluido em *lb/gal* e a TFA em *in²*, obtida na Tabela 32, que representa a área total dos orifícios da broca. A HSI, que representa uma medida de energia hidráulica, e é dada por

$$HSI = \frac{\Delta P_b}{1714 \cdot \frac{\pi \cdot D_b}{4}} \quad (B.2)$$

em que ΔP_b é dado em *psi* e, portanto, HSI é obtido em Hp/in^2 . A expressão que indica a força exercida na rocha pelo fluido de perfuração F_j é estimada por

$$F_j = 0,000518 \cdot \rho \cdot q \cdot v_n \quad (B.3)$$

em que F_j é expresso em *lbf*, ρ e q são as variáveis comentadas conforme Equação B.1, v_n é a velocidade do fluido nos orifícios da broca, dada em *ft/s*, e expressa por

$$v_n = 0,3208 \cdot \frac{q}{TFA} \quad (B.4)$$

Utilizando o *Dull grading* (Figura 71), é possível determinar a fração desgastada do dente da broca que é utilizada no modelo BYM. Determina-se inicialmente o DG, que é calculado com os valores dos dois primeiros campos do código e que correspondem aos danos nos dentes da broca para região interna e externa em uma escala de 1 a 8. Esses valores são interpretados após uma avaliação ao fim do uso da broca. Desse modo, determina-se o h , no intervalo de uso da broca, por meio da expressão dada por

$$h = \frac{D - D_{in}}{D_{out} - D_{in}} \cdot \frac{DG}{8} \quad (B.5)$$

que corresponde a uma interpolação linear para determinar o desgaste da broca em uma profundidade intermediária qualquer (D), e a profundidade final (D_{out}) e inicial (D_{in}) que são valores conhecidos.

Figura 71 – Detalhes sobre o código *Dull grading* segundo a IADC

CODIFICAÇÃO DOS DESGASTES DE BROCAS							
D				R	O	Observações	
Estrutura Cortante		Desgaste Principal		Rolamento ou Selos	Calibre 1/16"	Outro Desgaste	Motivo da Retirada
Interna	Externa	Característica	Localização				
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
5	8	IQ	M	E	2	IL	TP
(1) ESTRUTURA CORTANTE INTERNA				(2) ESTRUTURA CORTANTE EXTERNA			
2/3 Internos da Estrutura Cortante				1/3 Externo da Estrutura Cortante			
BROCAS DENTE DE AÇO - Uma medida linear do desgaste dos dentes devido à abrasão ou dano. Ex.: 0 - Nenhuma perda, desgaste e/ou quebra de insertos. 8 - Todos os insertos perdidos, desgastados e/ou quebrados.				BROCAS DE INSERTO - Uma medida linear de perda, desgaste ou quebra dos insertos. Ex.: 0 - Nenhuma perda, desgaste e/ou quebra de insertos. 8 - Todos os insertos perdidos, desgastados e/ou quebrados.		BROCAS DE DIAMANTE, PDC e TSP - Uma medida linear da perda, desgaste ou quebra dos cortadores. Ex.: 0 - Nenhuma perda, desgaste e/ou quebra dos cortadores. 8 - Todos os cortadores perdidos, desgastados e/ou quebrados.	
(3) DESGASTE PRINCIPAL				(4) LOCALIZAÇÃO		(8) MOTIVO DE RETIRADA DA BROCA	
Estes códigos também são usados na coluna (7).				BROCAS DE CONES		English - Português - Por extenso em português	
English - Português - Por extenso em português				N	I	WC	AM
SS	AA	Desgaste por Auto-Afiação		M	M	BU	EN
SD	AD	Aba da Perna Desgastada		G	E	HR	BF
-	AE	Aleta erodida		A	T	CM	CL
SP	AQ	Aba da Perna Quebrada		BROCAS DE DIAMANTE		CM	CM
PN	BE	Bocal do Jato Erodido		C	C	CT	DP
RG	CA	Cone com Calibre Arredondado		N	N	HR	DU
QC	CD	Cone c/ Desgaste Descentralizado		T	P	DSF	FC
ER	CE	Cone Erodido		S	O	DTF	FF
CI	CI	Cone com Interferência		G	G	DMF	FM
LC	CP	Cone Perdido		A	T	FP	FO
BC	CQ	Cone Quebrado		CONE # OU # CONES #s		PN	JO
CC	CT	Cone Trincado		1		BHA	MB
-	DC	Descolamento do cortador		2		FM	MF
WT	DD	Dentes Desgastados		3		CP	MP
CT	DL	Dentes Lascados		(5) ROLAMENTOS / SELOS		LOG	PE
FC	DP	Desgaste Plano nos Dentes		ROLAMENTOS NÃO SELADOS - Uma escala		TD	PF
BT	DQ	Dentes Quebrados		linear estimando a vida útil do rolamento já utilizado.		HP	PP
CR	DT	Cone c/ Desgaste p/ Testemunho		Ex.: 0 = Nenhuma vida utilizada		LIH	QJ
TR	EG	Desgaste por Engrenagem		8 = Toda vida útil utilizada		PP	QP
BU	EN	Desgaste por Enceramento		ROLAMENTOS SELADOS		TW	RD
-	FL	Falha de Liga do Cortador		E	E	RIG	RS
WT	ID	Insertos/Cortadores Desgastados		F	F	CP	TE
OG	IG	Insertos/Cortadores Girados		X	D	DST	TF
CT	IL	Insertos/Cortadores Lascados		(6) CALIBRE		TQ	TO
LT	IP	Insertos/Cortadores Perdidos		Ex.: 0 = Nenhuma perda de calibre		PR	TP
BT	IQ	Insertos/Cortadores Quebrados		1 = 1/16" de perda de calibre		WO	WO
LN	JP	Jato Perdido		2 = 2/16" de perda de calibre...			
RO	MA	Marca em Forma de Anel		(7) OUTRO DESGASTE			
JD	MF	Marca de Ferro		Códigos usados a partir da coluna (3).			
JD	MI	Marca de Inserto					
NO	NC	Nenhuma Característica de Desgaste					
PB	PE	Perna Empenada					
BL	PQ	Perna Quebrada					
RR**	RR	Broca Reutilizável					
CD*	TR	Cone Travado					
HC	TT	Trinca no Inserto ou Desintegração do Cortador p/ Temperatura					
OBSERVAÇÕES GERAIS							
Seguindo o método Petrobrás, temos:							
1) Iniciar a medida do desgaste dos cortadores do centro para a periferia.							
2) Considerar desgaste igual a 8 quando o PDC estiver tangente à superfície da aleta da broca.							
3) O nariz é composto pelo cortador posicionado na parte mais alta da aleta e seus vizinhos. A escolha da quantidade de cortadores vizinhos vai depender da conicidade da broca e do bom senso do técnico.							

Fonte: Kung (2016)

APÊNDICE C – CORRELAÇÕES EMPÍRICAS PARA UCS

As correlações empíricas utilizadas neste trabalho são obtidas por meio de calibração com dados experimentais obtidos de ensaios e publicados na literatura. As principais correlações encontradas para o UCS correspondente às litologias identificadas nos poços da *Volve* são estabelecidas entre a porosidade e o chamado tempo de trânsito compressional (Delta-T Compressional), que corresponde ao chamado perfil sônico.

Para as litologias do tipo: Argilito (*Claystone*) e Siltito (*Siltstone*), que são folhelhos, a correlação empírica é dada por

$$UCS = 0,77 \cdot \left(\frac{304,8}{\Delta t} \right)^{2,93} \quad (C.1)$$

Essa expressão é documentada em Horsrud (2001). Para a litologia do tipo Calcário (*Limestone*) e Marga (*Marl*), a correlação utilizada é expressa por

$$UCS = \frac{10 \left(2,44 + \frac{109,14}{\Delta t} \right)}{145} \quad (C.2)$$

Essa equação foi obtida em Chang et al. (2006). Essa expressão pode ser utilizada para rochas carbonáticas, como o calcário. Para a marga, que é uma rocha carbonática com alto teor de argila, opta-se por aplicar a Equação C.2 devido às intercalações observadas entre essas duas rochas nos dados de litologia (vide os gráficos de litologia no Apêndice D). Para o Arenito (*Sandstone*) e o Carvão (*Coal*), a correlação dada por

$$UCS = 1200 \cdot e^{-0,036 \cdot \Delta t} \quad (C.3)$$

é aplicada. Essa expressão, por sua vez, é extraída do estudo de McNally (1987).

APÊNDICE D – DADOS RELEVANTES NOS POÇOS VOLVE

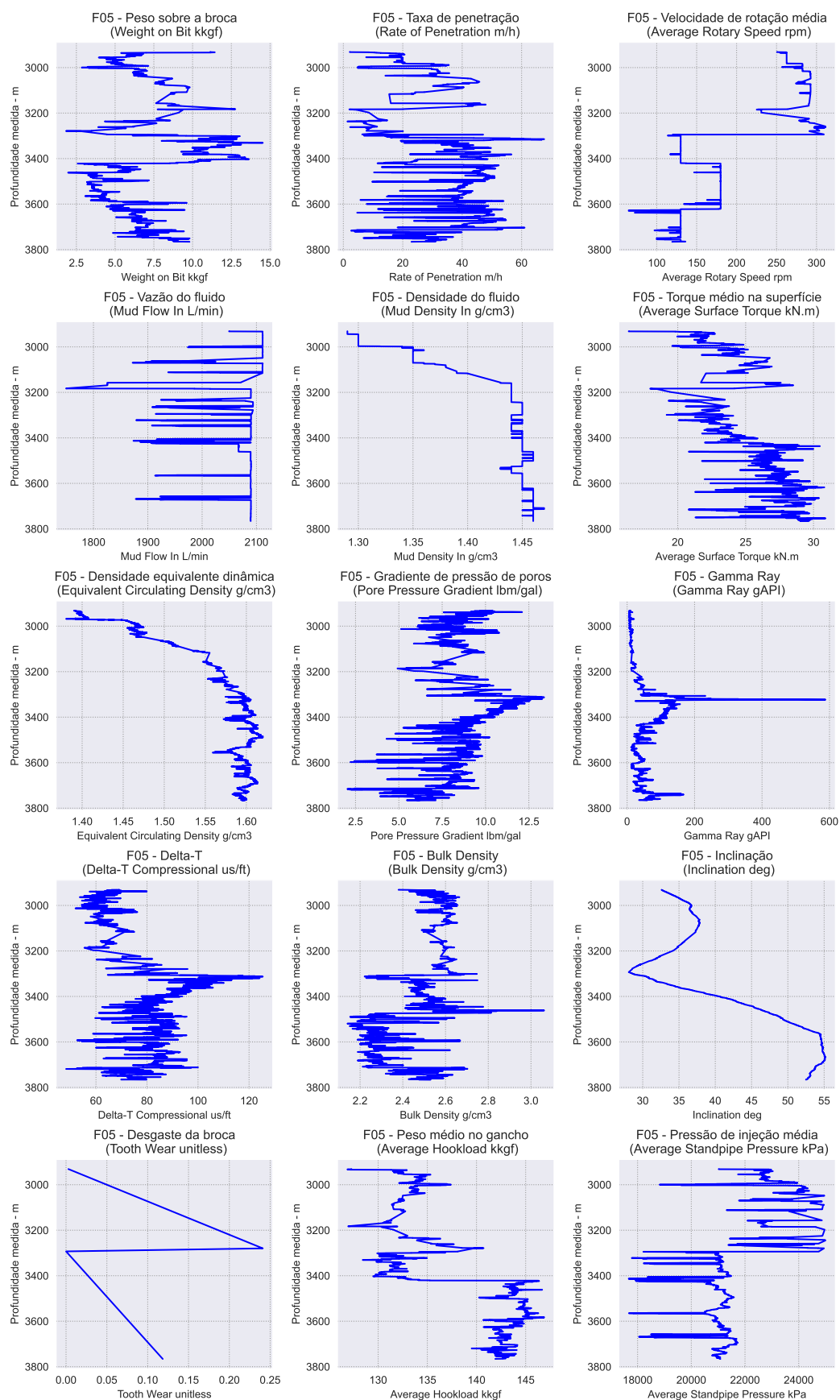
Na Figura 72 são apresentados histogramas com os dados de todos os poços e para algumas das variáveis de perfuração extraídas dos poços analisados. Nas Figuras 73, 74, 75, 76, 77, 78 e 79 são apresentados cada variável ao longo da profundidade medida. Nas Figuras 80, 81, 82, 83, 84, 85 e 86 são apresentadas as informações de UCS e litologia em comparação com a ROP para cada poço.

Figura 72 – Histogramas agrupados das principais informações presentes no *dataset* para cada poço



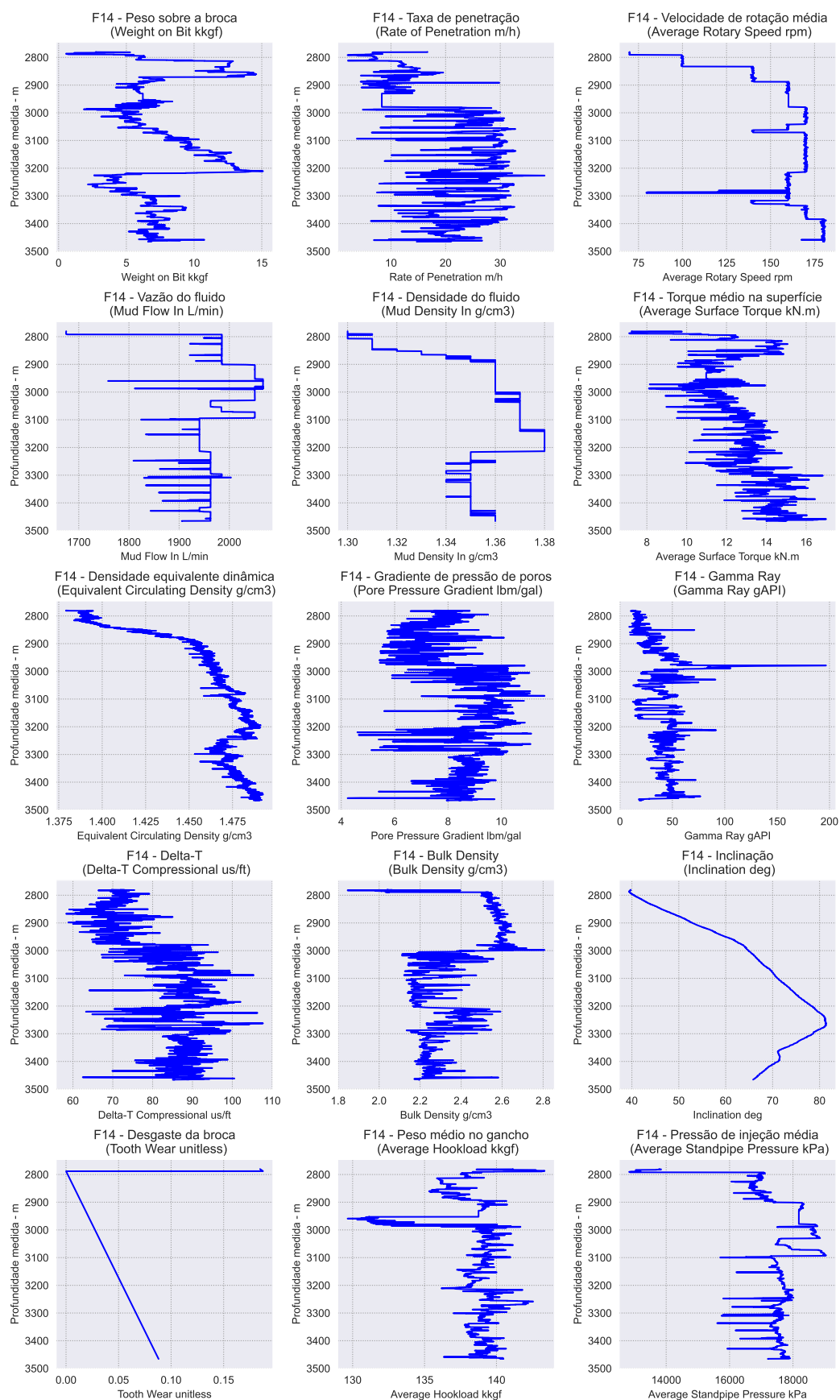
Fonte: Autor (2024)

Figura 73 – Plots das principais informações do poço F05 ao longo da profundidade medida



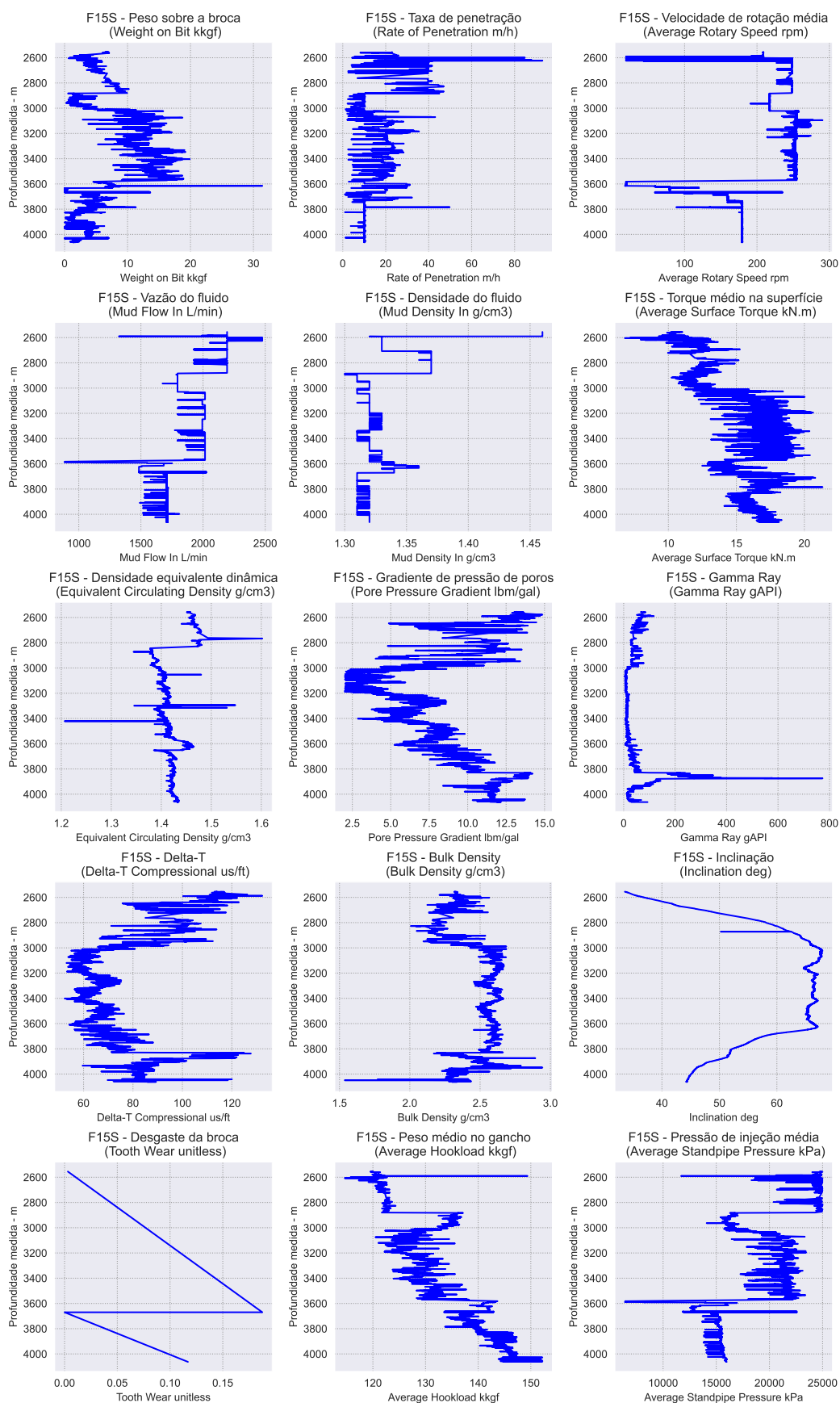
Fonte: Autor (2024)

Figura 74 – Plots das principais informações do poço F14 ao longo da profundidade medida



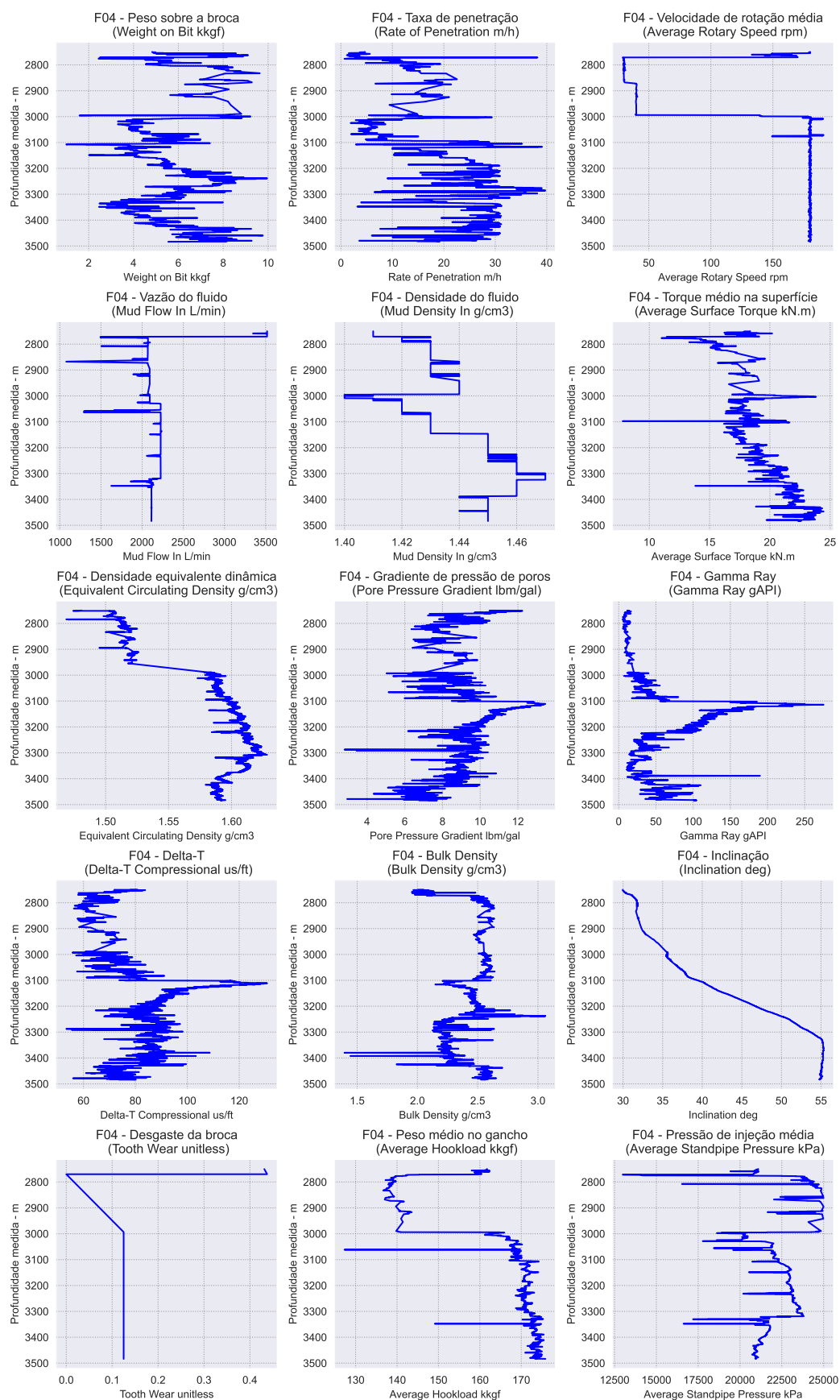
Fonte: Autor (2024)

Figura 75 – Plots das principais informações do poço F15S ao longo da profundidade medida



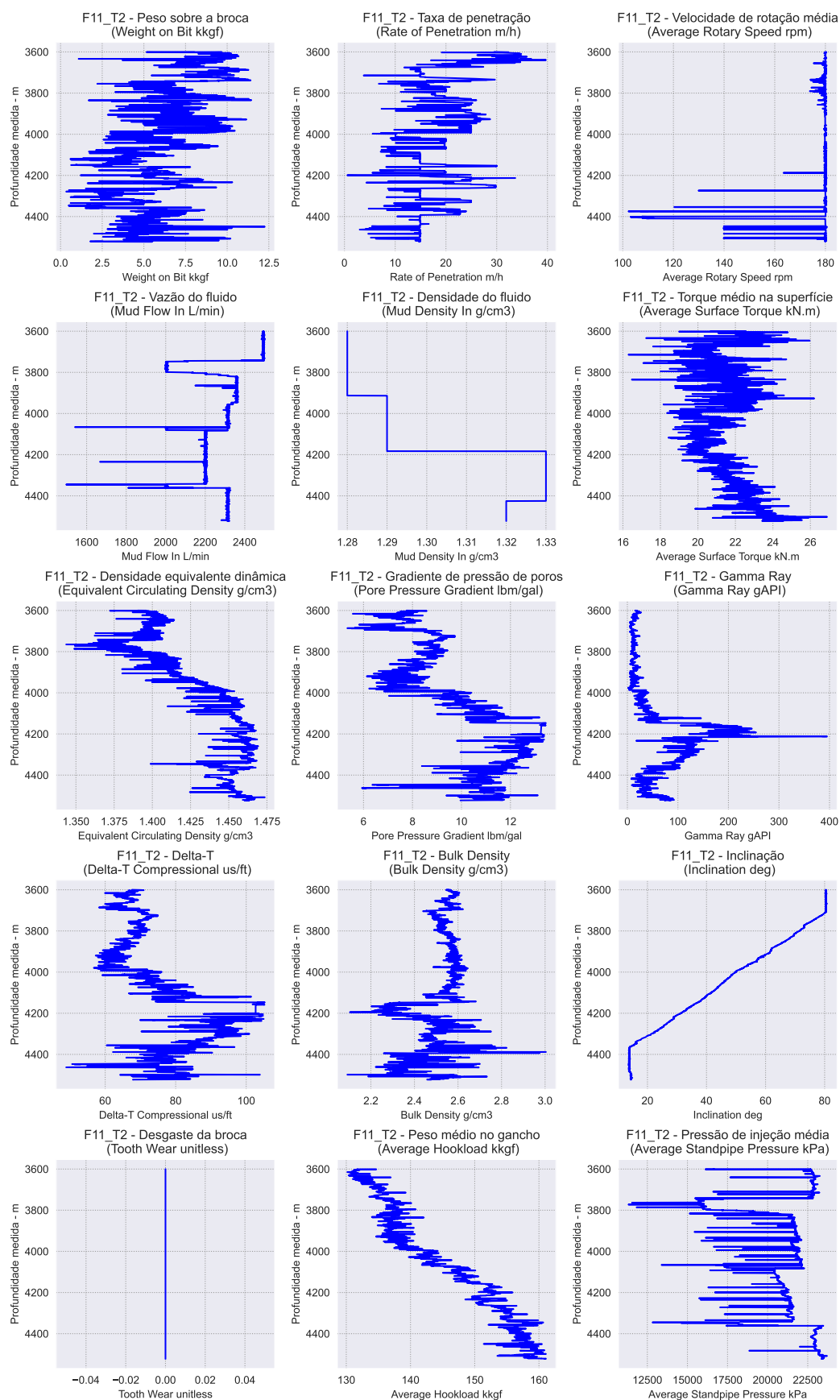
Fonte: Autor (2024)

Figura 76 – Plots das principais informações do poço F04 ao longo da profundidade medida



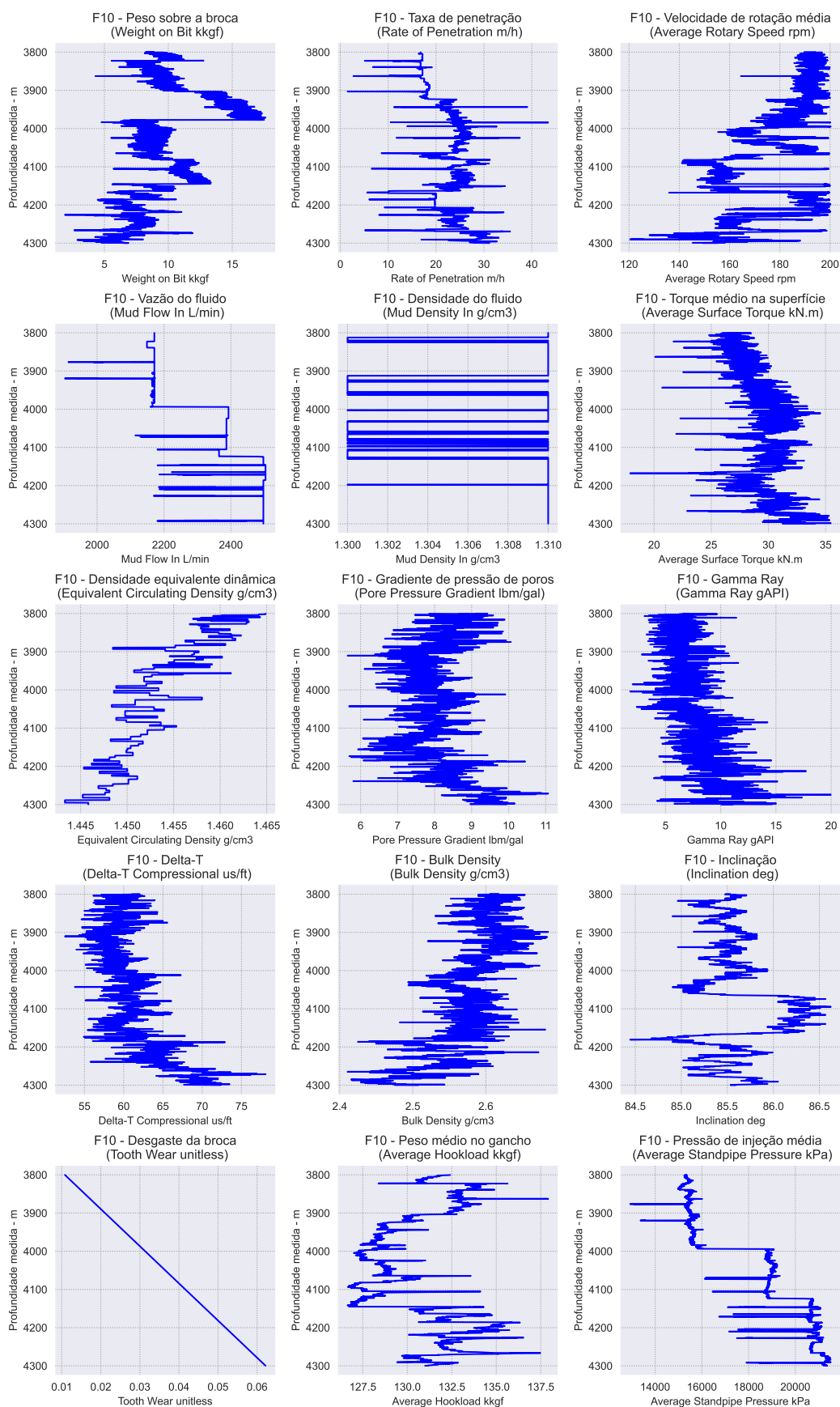
Fonte: Autor (2024)

Figura 77 – Plots das principais informações do poço F11 T2 ao longo da profundidade medida



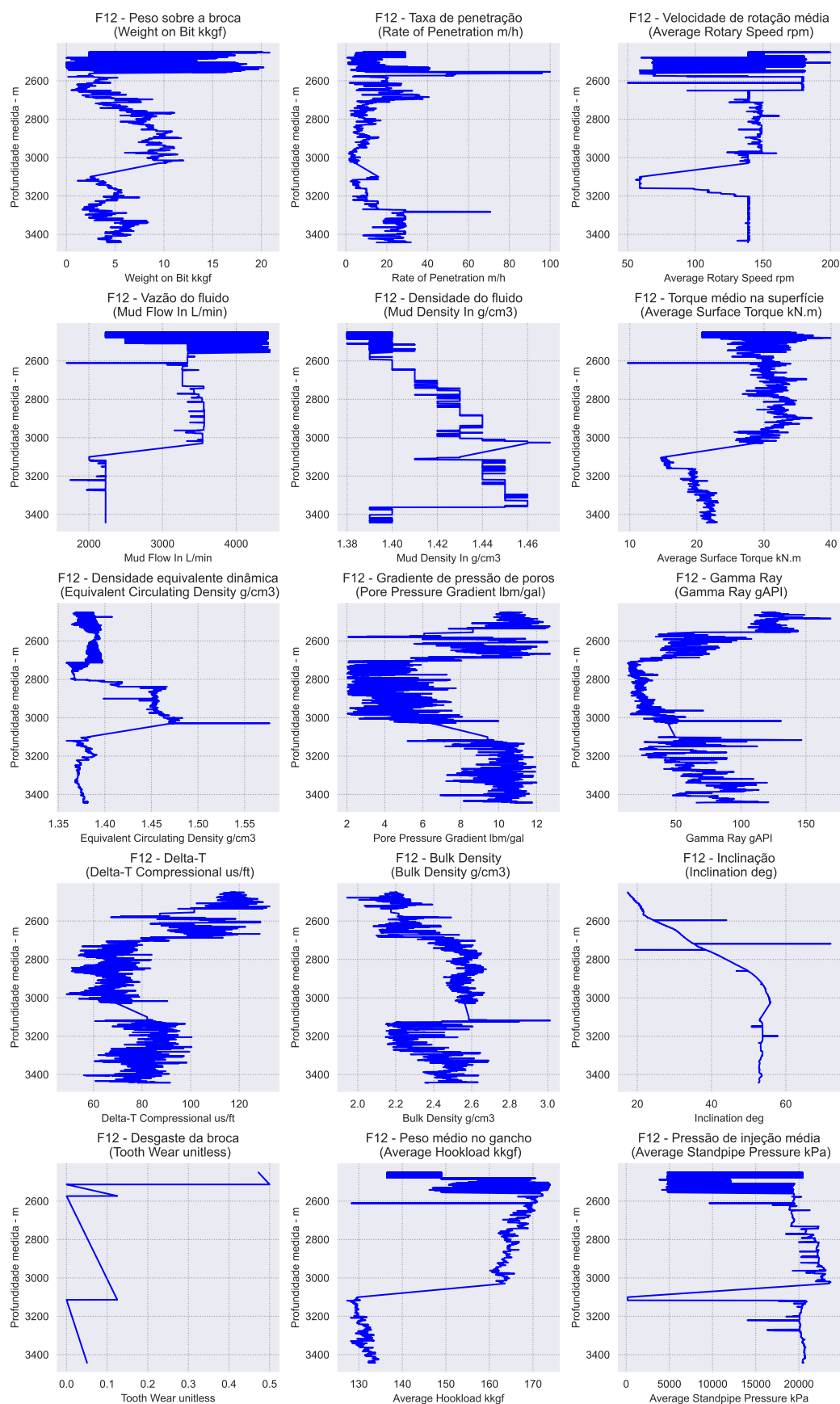
Fonte: Autor (2024)

Figura 78 – Plots das principais informações do poço F10 ao longo da profundidade medida



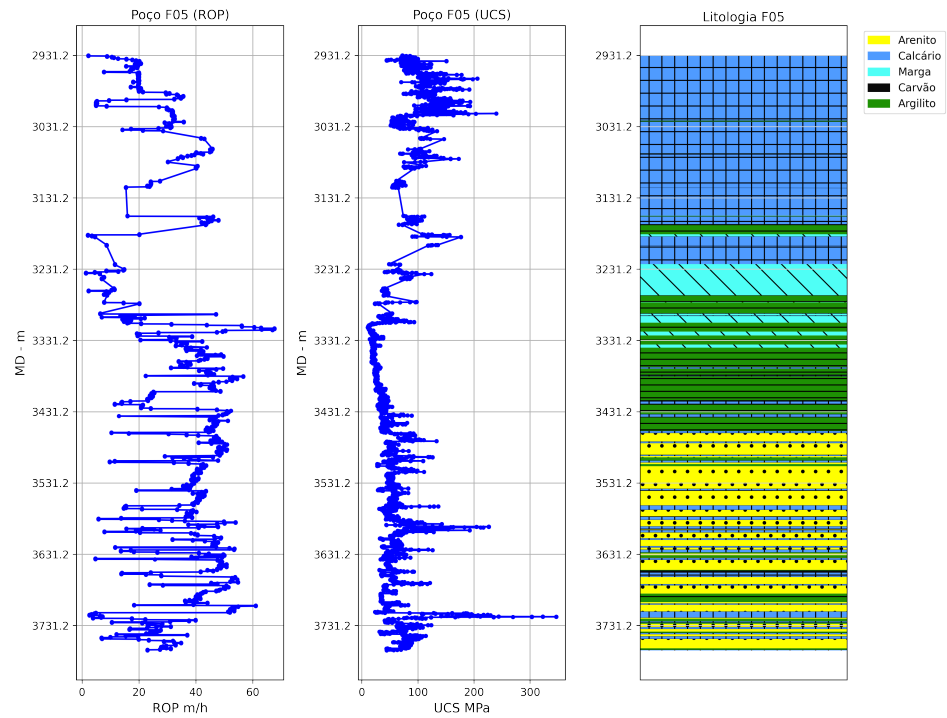
Fonte: Autor (2024)

Figura 79 – Plots das principais informações do poço F12 ao longo da profundidade medida



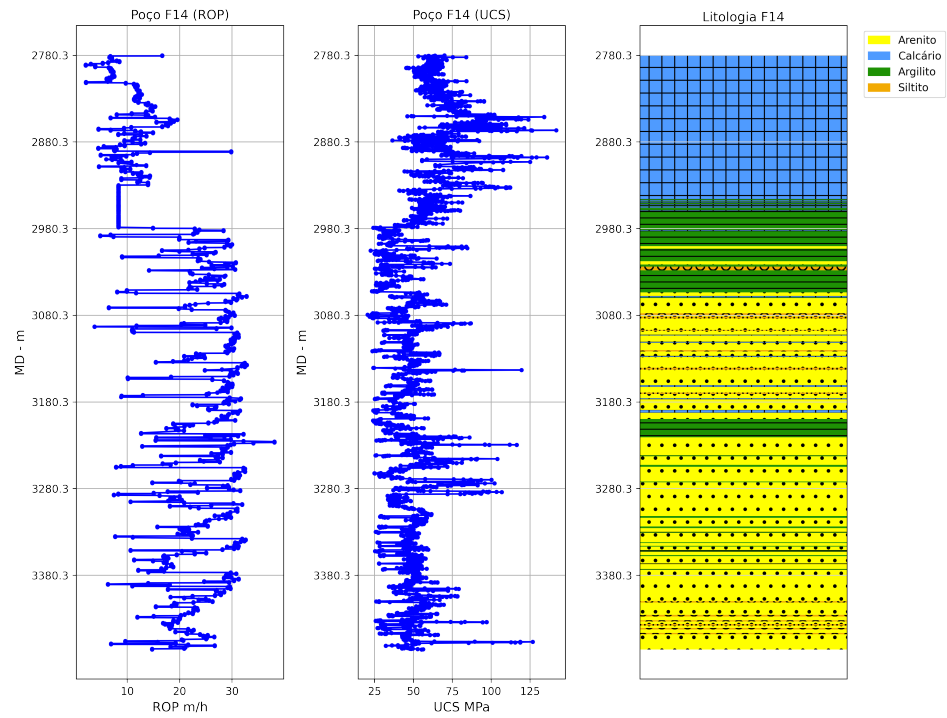
Fonte: Autor (2024)

Figura 80 – Plot da ROP, UCS e litologia para o Poço F05



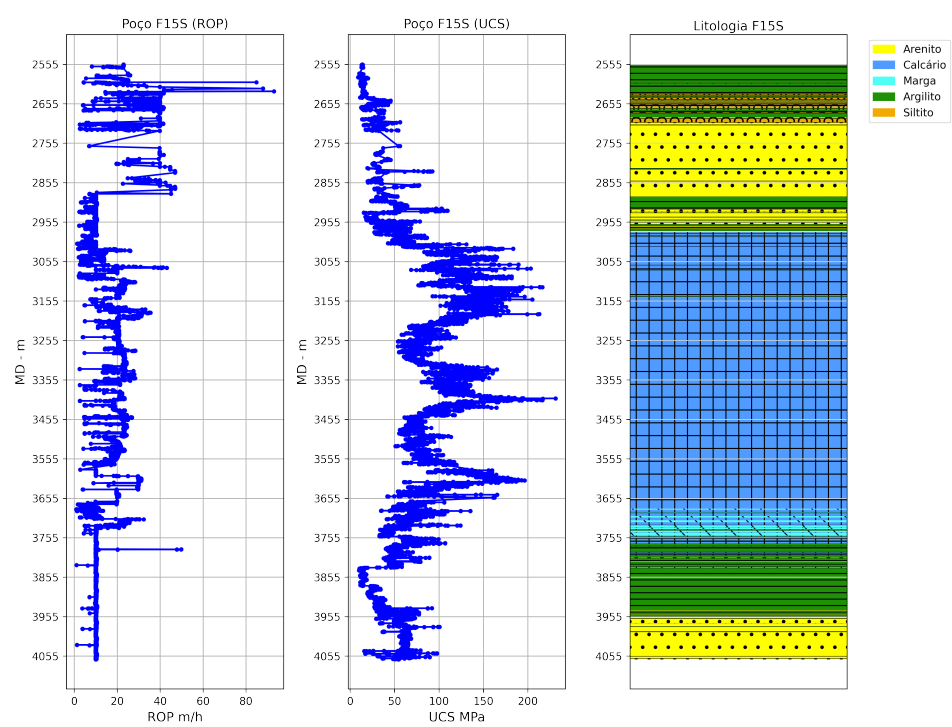
Fonte: Autor (2024)

Figura 81 – Plot da ROP, UCS e litologia para o Poço F14



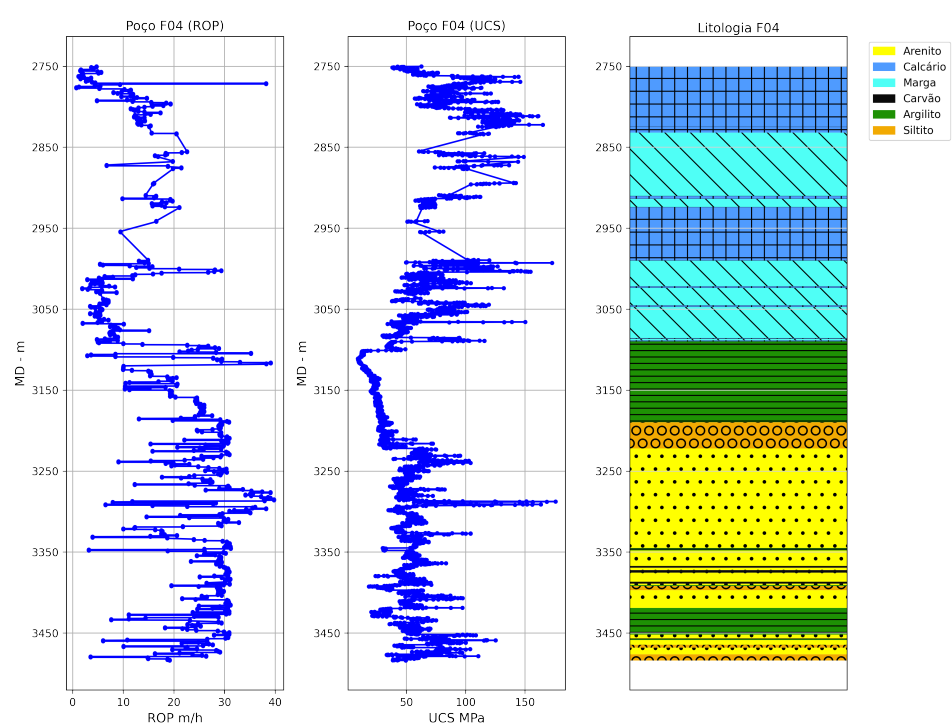
Fonte: Autor (2024)

Figura 82 – Plot da ROP, UCS e litologia para o Poço F15S



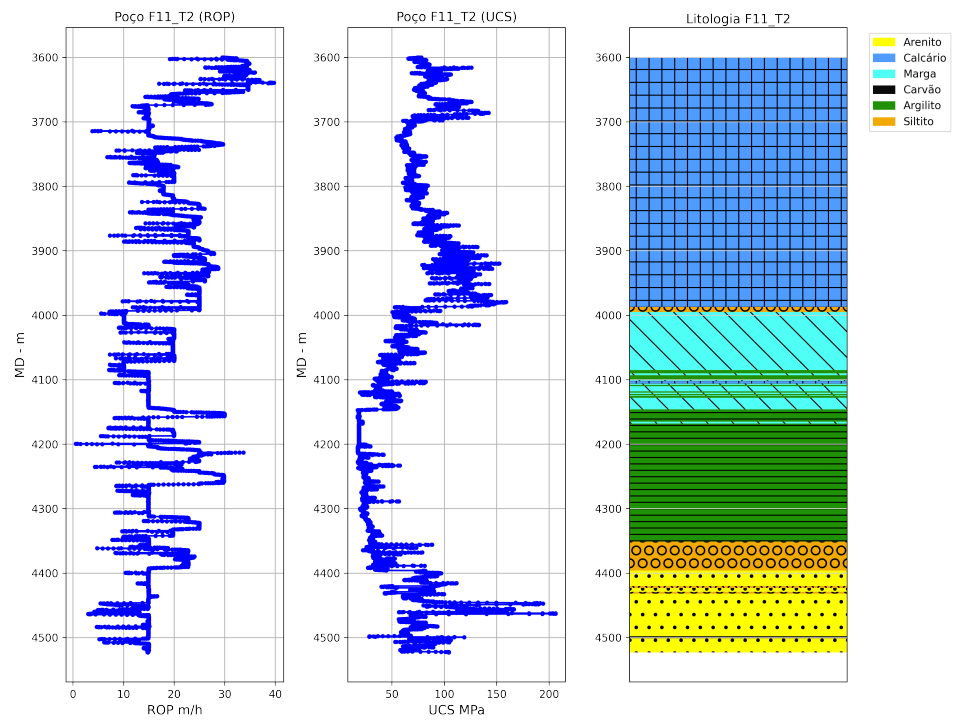
Fonte: Autor (2024)

Figura 83 – Plot da ROP, UCS e litologia para o Poço F04



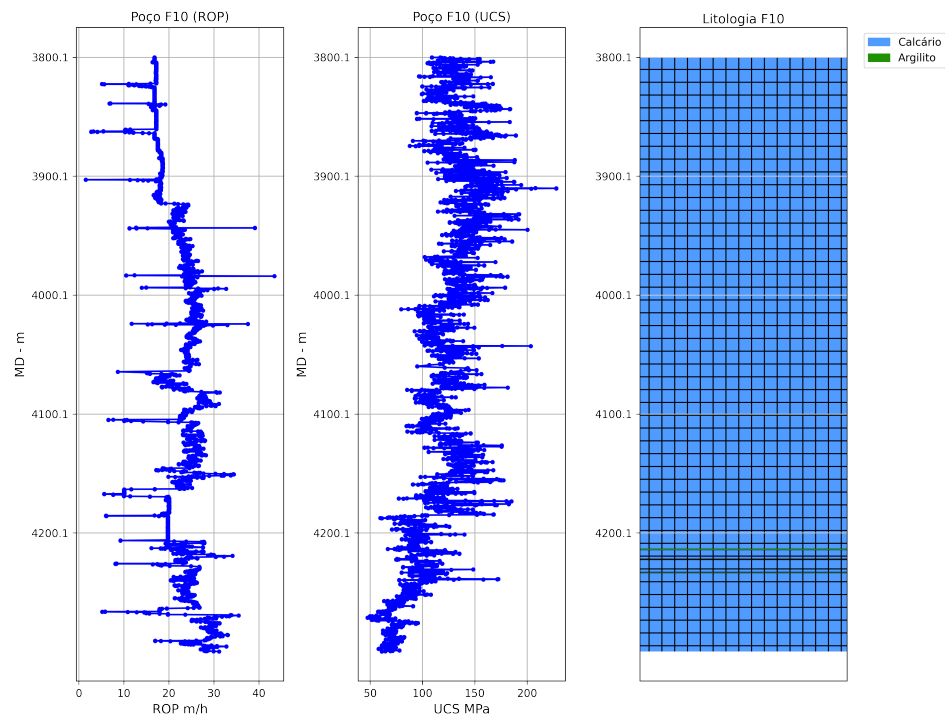
Fonte: Autor (2024)

Figura 84 – Plot da ROP, UCS e litologia para o Poço F11_T2



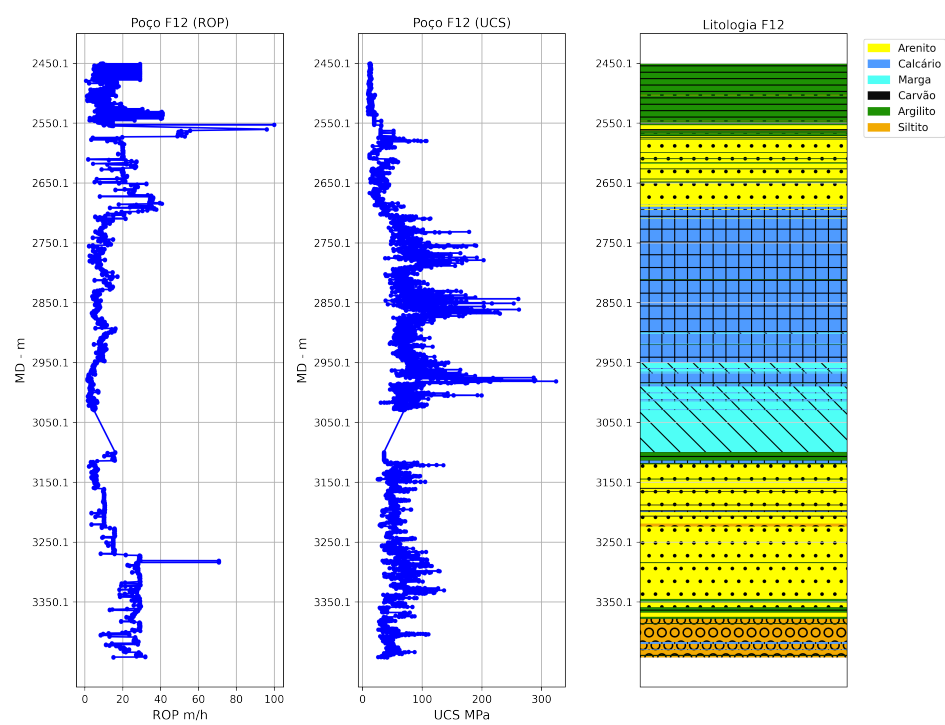
Fonte: Autor (2024)

Figura 85 – Plot da ROP, UCS e litologia para o Poço F10



Fonte: Autor (2024)

Figura 86 – Plot da ROP, UCS e litologia para o Poço F12



Fonte: Autor (2024)