Estimativa da porosidade usando ferramentas de aprendizagem de máquina não paramétrica

Isaac N. da S. Macedo (FAGEOF-UFPA), Jose F. V. Gonçalves (FAGEOF-UFPA), Celso R. L. Lima (CPGf-UFPA) , Jose J. S. de Figueiredo (CPGf e FAGEOF-UFPA),Pamela C. R. Bolsem (Fagoef-UFPA) e João L. M. da Silva (Fagoef-UFPA)

Resumo

É conhecido que a porosidade é um importante parâmetro petrofísico pra exploração de hidrocarbonetos. A porosidade pode ser obtida através da perfilagem e experimentalemnte de uma forma direta. Teoricamente, a porosidade pode ser obtida empiricamente. Neste trabalho, através do uso da Aprendizagem de Máquinas (AM) observamos que o uso de regressões não paramétricas são mais eficientes que a regressão linear para a predição desse parâmetro, pois tais métodos são mais robustos. A análise de performace de uso destes métodos foram verificados a partir de métricas de Regressão tais como: R^2, MSE, RMSE e MAE. Dos métodos utilizados o que mostrou melhor eficiência foi a Regressão Guassiana, seguido pelas Regressões Bayesiana e Linear.

Introdução

A perfilagem é uma técnica de obtenção de dados que busca avaliar o perfil geofísico encontrado com o intuito, de modo geral, obter dados capazes de serem associados a propriedades petrofísicas das rochas e que possam identificar locais de interesse de possíveis explorações de óleo e gás. Por definição, um perfil de poço é uma representação gráfica entre a profundidade e as propriedades petrofísicas da rocha através de poço. Ele é obtido através de uma sonda de perfilagem que desce dentro do poço em questão da exploração (Nery, 2013). As informações obtidas no perfil são compostas geralmente pela litologia, porosidade, espessura, profundidade do poço, temperaturas, pressão e propriedades das rochas como resistividade, densidade e taxa de raio gama. Com base na análise dos dados obtidos pelo perfil, são decidido os intervalos do poço que são de interesse a serem analisados, onde se dá a primeira etapa do processamento dos dados geofísicos.

Entretanto, ao passar dos anos surgiu a necessidade do desenvolvimento de tecnologias capazes de obterem respostas mais eficientes e robustas em relação a obtenção de propriedades petrofísicas no contexto teórico ou empírico. Em relação as pesquisas científicas, surgiu o Machine Learning (ML) ou aprendizagem de máquinas, conhecido comumente como Aprendizado de Máquina (AM) (Jiang et al., 2020; Nourani et al., 2022). Podemos definir o ML ou AM como a associação e análise de um conjunto de dados e métodos como classificação, agrupamento e/ou regressão, com o objetivo de entender o padrão e buscar uma análise de como o sistema se adapta em determinadas situações trazendo o intuito de melhorar a precisão dos dados (Raschka, 2015).

Neste trabalho usamos ferramentas de aprendizagem de máquina para regressão, assim como processamento de perfis de poço para obtenção do perfil de porosidade a partir de outros perfis. Estes estudos foram feitos em dados da perfilagem da Bacia de Campos. De forma direta, buscou-se predizer a porosidade por meio de regressões não paramétricas, Regressão Bayesiana e Gaussiana, e paramétrica tendo como hipótese que a regressão não-paramétrica supera a paramétrica (ex: Regressão Linear Multiparamétrica) por ser um método mais robusto para a predição.

Metodologia

A porosidade é definida como a relação entre o volume de espaços vazios de uma rocha e o volume total da mesma, em percentual, é uma propriedade estatística que depende das dimensões envolvidas (Schön, 1998). Existem alguns tipos de classificação para a porosidade, por exemplo: porosidade deposicional ou primária, pós-deposicional ou secundária. A porosidade primária são as rochas obtidas durante processos tectônicos deposicionais, e a porosidade pós-deposicional que é o resultado de um processo geológico após a transformação do sedimento em rocha.

Usam-se, também, os termos porosidade absoluta – que relaciona o volume total de vazios – e porosidade efetiva – que leva em conta apenas os espaços vazios interconectados. A porosidade efetiva é fundamental nos cálculos de interpretação do reservatório de hidrocarboneto, e devido sua importância comercial (Schön, 1998). Em perfilagem, através da ferramenta neutrônica, porosidade total é determinada. Para obtermos a porosidade efetiva é necessário uma análise laboratorial de uma amostra de testemunho (Nery, 2013).

Para uma dada situação no qual necessita-se determinar a porosidade, uma das formas que encontramos é de relacioná-la com outras propriedades físicas das rochas, sabendo que elas são parâmetros petrofísicos que podemos encontrar durante a perfilagem. Por exemplo, a partir de um modelo linear, podemos relacionar a porosidade com outros perfis, de acordo com a seguinte relação

$$\phi = a\Delta T + b\rho b + cV_{sh}, \quad (1)$$

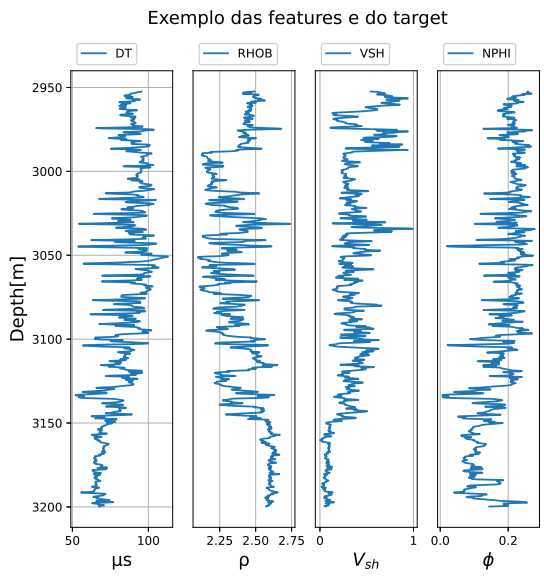

Figura 1 – Exemplo (poço 1 da Bacia de Campos) de perfis escolhidos como entrada e saída para regressão linear, Gaussiana e Bayesiana.

no qual ΔT é tempo de trânsito sônico, V_{sh} é volume de folhelho e (ρ) é a densidade da rocha. Esta equação nos mostra que a porosidade pode ser proposta por uma função linear de parâmetros petrofísicos e seus coeficientes por meio de uma regressão linear ou usando regressão por aprendizagem de máquinas. Neste caso, tendo como features -parâmetros de entrada- os perfis de densidade, tempo de trânsito e o volume de folhelho. Figura 1 mostra exemplo de perfis usado no nosso trabalho de regressão.

Na regressão, ao contrário da classificação, busca-se predizer quantidades- saídas (target's)- contínuas. Essa predição (regressão) pode ser paramétrica e não paramétrica. A primeira consiste nos métodos tradicionais, regressão linear (por exemplo), e as não paramétricas consistem em saídas definidas como uma distribuição sobre funções e inferência ocorrendo diretamente no espaço dessas (Casella et al., 2006). No caso do presente estudo, utilizamos o processo de regressão Gaussiana e regressão Bayesiana como métodos de inferência não paramétricos.

Em alguns caso, como veremos a seguir, método de regressão Linear foi menos eficiente que as predições não paramétricas. Neste ultimo caso, foi levantado um conjunto de hipóteses no qual as regressões não-paramétricas seriam foram as mais precisas para a predição da porosidade.

Como mencionado anteriormente, foi usado as regressões Bayesiana e a Gaussiana para a previsão de porosidade. A regressão Bayesiana pode descrita por:

$$p(\omega|\lambda) = \mathcal{N}(\omega|0,\lambda^{-1}\mathbf{I}_p), \tag{2}$$

tendo em vista que os valores a priori para ω e λ são escolhidas para podermos ter distribuições gama. Esses parâmetros,ω e λ, tem por objetivo ajustar o modelo de regressão (Bishop & Nasrabadi, 2006). Já a equação que descreve a regressão gaussiana é dada por:

$$f(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), \kappa(\mathbf{x},\mathbf{x}')\right), \tag{3}$$

no qual $m(\mathbf{x})$ é definida como a função média e $\kappa(\mathbf{x},\mathbf{x}')$ é definida como a função covariância (Williams & Rasmussen, 2006). Ademais, neste trabalho fora utilizado os kernels da regressão gaussiana pois esses determinam a forma a priori e a posteriori da regressão (Williams & Rasmussen, 2006). O primeiro kernel utilizado fora o Radial Basis Function (RBF). Esse kernel é do tipo estacionário, ou seja, depende apenas da distância de dois dados não dependendo de seus valores absolutos, além do mais são invariantes. Matematicamente esse Kernel é dado por:

$$\kappa\left(\mathbf{x}_i,\mathbf{x}_j\right) = exp\left(-\frac{d(\mathbf{x}_i,\mathbf{x}_j)^2}{2l^2}\right), \tag{4}$$

no qual $d(\cdot,\cdot)$ é a distância euclidiana entre os dados de teste (x_i) e treinamento (x_j) e l é o parâmetro de escala da hiperparametrização. Por fim, o outro kernel utilizado fora o rational quadratic que pode ser visto como uma mistura de kernel RBF, mas com diferentes escalas de comprimento. Matematicamente esse kernel é dado por:

$$\kappa\left(\mathbf{x}_i,\mathbf{x}_j\right) = \left(1+\frac{d(\mathbf{x}_i,\mathbf{x}_j)^2}{2\alpha l^2}\right)^{-\alpha}, \tag{5}$$

tendo que α e l tem que ser maior que 0, podendo ser visto como uma escala de mistura da raiz da função exponencial de covariância.

Na aplicação da nossa metodologia de trabalho - que pode ser observada na Figura 2- foram utilizados 3 poços (poço 1, 5 e 12) para realizar o teste/treinamento e um dado de validação (poço 3) que não foi utilizado no treinamento. Todos os poços foram oriundos da bacia de Campos.

Como pode ser observado no fluxograma da Figura 2, o trabalho começa com o carragamento dos dados, seguida pela análise de correlação por meio de gráfico de dipersão entre as variáveis de entrada e a variável de saída-predita, bem como pela correlação de Pearson, assim como pelas medidas de dispersão, ainda na etapa de visualização e análise buscou-se verificar se havia dados faltantes.

Ao término da análise estatística e visualização dos dados fora iniciado a etapa de pré-processamento sendo feito 'dropNaN' nos dados faltantes, após o corte com base nos gráficos de correlação e na correlação de pearson fora escolhido os melhores poços, julgou-se como melhor poço aqueles que continham grandes variações e correlações (ver Figura 3) pois esses permitem que a máquina aprenda sem enviesamentos e permite que ela generalize para poços não usados na etapa de processamento.

Na etapa de processamento aplicou-se as técnicas de ML para regressão, regressão linear, regressão Gaussiana e regressão Bayesiana com kernel e sem kernel. Essas regressões foram feitas em Python utilizando

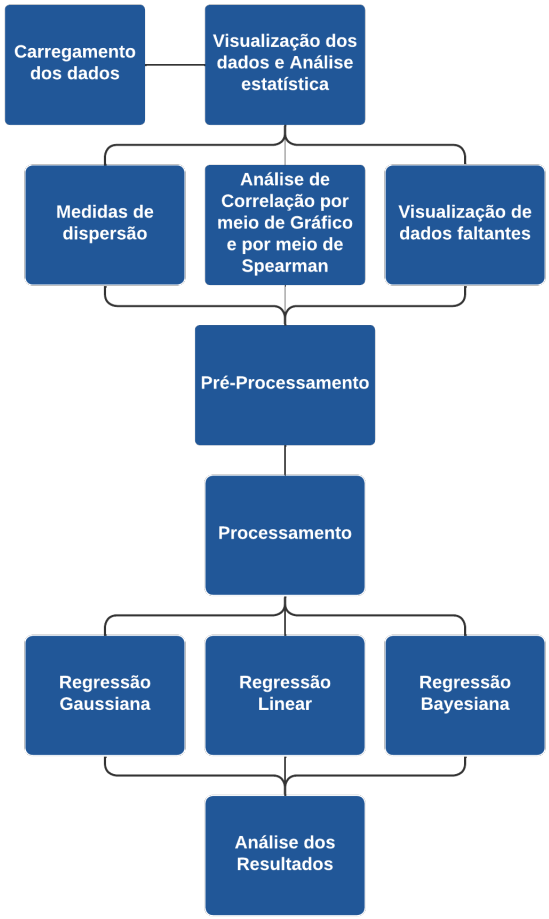

Figura 2 – Este fluxograma ilustra a metodologia geral deste trabalho.

as Bibliotecas do Framework de Machine Learning do scikit learn (https://scikit-learn.org/stable/). Tendo escolhido essas regressões, com o objetivo de encontrar os melhores hiperparâmetros fora aplicado GridSearch (também do scikit learn). Sabe-se que esta técnica de otimização consiste na crossvalidação e combinação de hiperparâmetros para encontrar o melhor ajuste de curva para a propriedade do target a ser previsto. Também foi feita a concatenação de poços com boas variações e correlações para que a máquina pudesse generalizar e assim obter mais precisão em um poço de validação com os hiperparâmetros encontrados, podendo ser visto estes na Tabela 1.

Tabela 1 – Valores dos hiperparâmetros utilizados na regressão linear, Bayesiana e Gaussiana. Esses valores foram obtidos usando o GridSearch do Scikit learn.

Hiperparâmetros da Regressão Bayesiana	Hiperparâmetros da Regressão Gaussiana RBF	Hiperparâmetros da Regressão Gaussiana Rational Quadratic
PolinomialFeatures grau = 2	alpha = 0.01	alpha = 1e-07
alpha_1 = 1e-10	length_scale = 0.1	alpha = 1
alpha_2 = 1e4	—	length_scale = 0.01
lambda_1 = 0.1	—	—
lambda_2 = 0.1	—	—
n_iter = 400	—	—
tol = 0.001	—	—

Por fim, na última etapa fora feito a análise dos resultados que consiste na análise das métricas do Mean Squared Error (MSE),

$$MSE(y,\hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2, \tag{6}$$

e a Mean absolute error (MAE),

$$MAE(y,\hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} |y_i - \hat{y}_i|, \tag{7}$$

no qual n é número de amostras, y_i é o valor real e $\hat{y}_i$ é o valor predito. As outras métricas analisadas foram o RMSE (Root-mean-square deviation),

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}, \tag{8}$$

e R-Squared é definido por

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{9}$$

no qual $\bar{y}$ é a média do valor real. E por ultimo, a métrica Mean absolute percentage error (MAPE) definido por

$$MAPE(y,\hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} \frac{|y_i - \hat{y}_i|}{\max(\varepsilon, |y_i|)}, \tag{10}$$

no qual ε é número arbitrário e positivo para evitar resultados indefinidos quando o y_i tende a zero.. Todas essas métricas são importantes para analisar a eficiência e performace dos algoritmos de regressão.

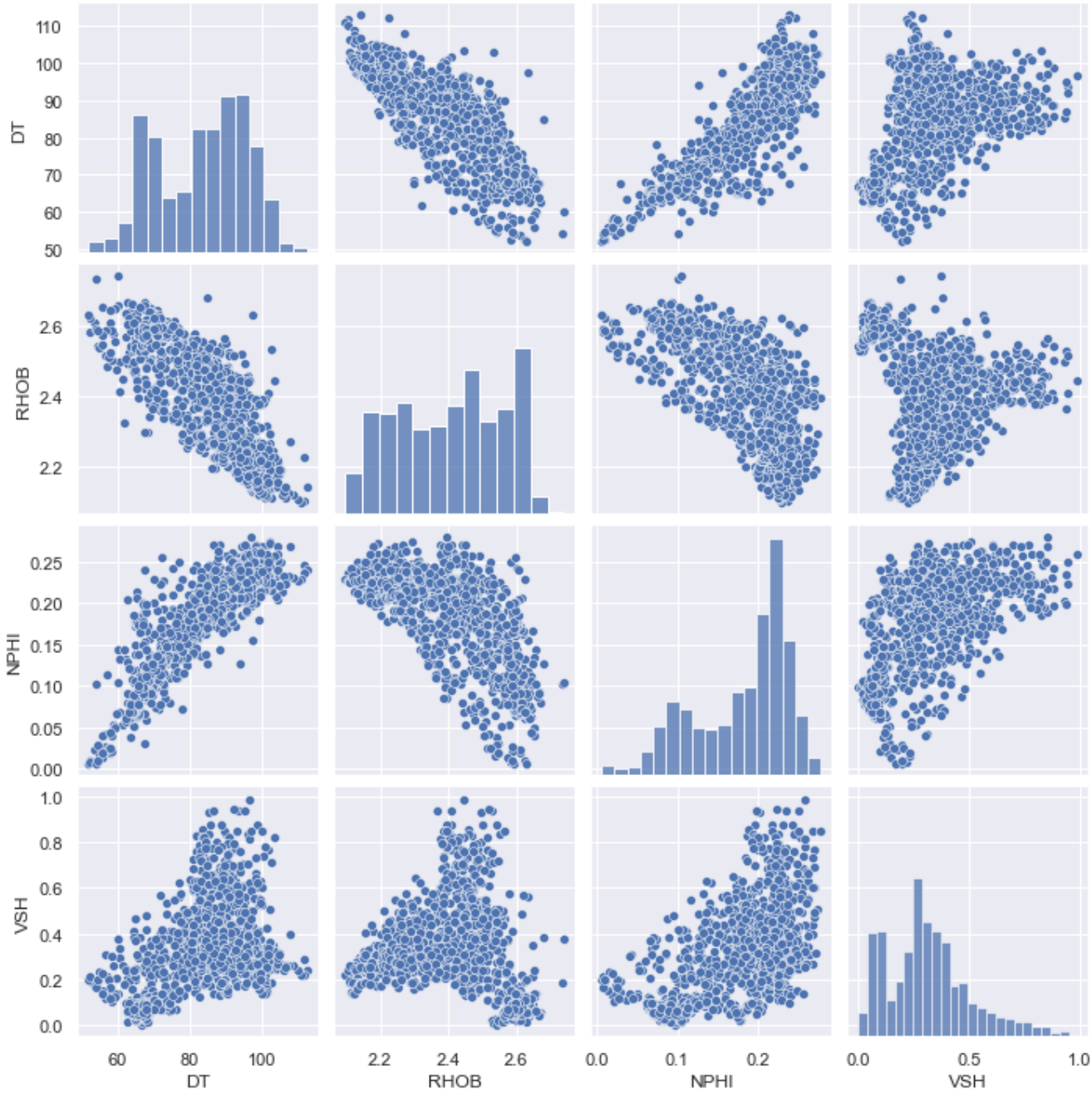

Figura 3 – Gráfico de correlação do poço 1 entre o target e as features.

Resultados

Como pode ser verificado na Tabela 2, a partir dos valores das métricas descritas na seção anterior, os métodos de regressão não paramétricas apresentaram uma melhor performace quando comparado ao método paramétrico baseado em regressão Linear. Destaca-se a métrica R^2 no qual apresentou o maior valor para dados de validação, quando comparado com valores da regressão Linear.

Pode ser notado a partir da Tabela 2, que entre os métodos de regressão não paramétrica, aquele que apresentou um melhor desempenho no modelo predito (segundo as métricas MSE e MAPE) foi o método de mistura Gaussiana. Entre os kernels analisados, destaca-se o kernel tipo "Rational Quadratics". Também pode ser observado, que as regressões não paramétricas no poço de validação- sabendo que o poço de validação é um poço que não fora usado em nenhuma parte do treino da máquina- tiveram menor erro em relação a regressão linear.

Após a análise das métricas fez-se a análise gráfica dos

Tabela 2 – Valores de métricas de Regressão para métodos lineares, regressão Bayesiana e Gaussiana, como se pode observar de todos os métodos destaca-se com melhor eficácia o método de Regressão Gaussiana.

Métricas	Regressão L. Poço Validação	Regressão L. Poço Teste	Regressão B. com Hiperparâmetros Poço Validação	Regressão B. com Hiperparâmetros Poço Teste	Regressão G. com Hiperparâmetros RBF Poço Validação	Regressão G. com Hiperparâmetros RBF Poço Teste	Regressão G. com Hiperparâmetros Rational Quadratic Poço Validação	Regressão G. com Hiperparâmetros Rational Quadratic Poço Teste
MAE	0.03117	0.02247	0.02903	0.02016	0.02794	0.01973	0.02842	0.01722
RMSE	0.03762	0.02928	0.03534	0.02732	0.03495	0.02710	0.03527	0.02456
MSE	0.00141	0.00086	0.00125	0.00075	0.00122	0.00073	0.00124	0.00060
R2	0.32595	0.73844	0.40512	0.73846	0.41810	0.77588	0.40744	0.81592
MAPE	0.14963	0.16308	0.14156	0.12876	0.13813	0.12166	0.13943	0.11065

dados, buscando-se verificar onde as regressões não paramétricas se ajustaram melhor ao dado real. Para isso fora marcado na Figura 4 as áreas (em vermelho) onde as curvas preditas pelos métodos não paramétricos se ajustaram melhor do que a regressão linear.

Na ultima coluna da Figura 4 mostra o erro relativo. Devida a presença de ”outliers”, o erro relativo superior a 10 % em todos os casos. No entanto, é possível observar na parte infeiror da ultima parte (detacado em azul) mostra que o desempenho do método Gaussiano é melhor que os demais.

Conclusões

Neste trabalho foi aplicado três metodologias de regressão para estimativa da porosidade a partir de perfis petrofísicos (sônico, Raio Gama e Densidade). A partir dos valores obtidos quantitativamente pelas métricas de regressão quanto, pela análise qualitativa através da visualização dos perfis real e predito, destaca-se:

1. Para todos casos - para dados de teste e validação- as regressões não paramétricas superam a regressão linear;

2. Das regressões não paramétricas destaca-se a regressão Gaussiana;

3. Em relação aos tipos de Kernel, no poço de validação destaca-se o RBF e o no teste destaca-se o Rational Quadratics;

4. O erro relativo variou entre $\pm 10\%$ (sem a presença de outliers) e o MAPE para os casos não paramétricos apresentaram o menor erro;

Para trabalhos futuros pretende-se realizar testes e validação destas metodologias com poços de outros campos Também usar outros modelos com diferentes feautures e comparações com outros modelos de ML e Deep Learning.

Agradecimentos

Os autores agradecem a Universidade Federal do Pará (UFPA) e ao Instituto de Geociências pelo apoio institucional. O discente Isaac N. da S. Macedo agradece ao CNPq pela concessão da bolsa de Iniciação Científica.

Referências

Bishop, C. M. & Nasrabadi, N. M., 2006. Pattern recognition and machine learning, vol. 4, Springer, New York, USA.

Casella, G., Fienberg, S. & Olkin, I., 2006. Nonparametric Regressio, Springer New York, New York, NY, doi: 10.1007/0-387-30623-4$_5$.

Jiang, L., Castagna, J. P. & Russell, B., 2020. Porosity prediction using machine learning, 3862–3866, doi: 10.1190/segam2020-w13-04.1.

Nery, G. G., 2013. Perfilagem geofísica em poço aberto, 1st ed., Sociedade Brasileira de Geofísica (SBGf), Rio de Janeiro, BR.

Nourani, M. et al., 2022. Comparison of machine learning techniques for predicting porosity of chalk, Journal of Petroleum Science and Engineering, vol. 209: 109853, doi: https://doi.org/10.1016/j.petrol.2021.109853.

Raschka, S., 2015. Python Machine Learning, 1st ed., Packt Publishing, Birmingham, UK.

Schön, J., 1998. Physical Properties of Rocks: Fundamentals and Principles of Petrophysics, Handbook of geophysical exploration: Seismic exploration, Elsevier.

Williams, C. K. & Rasmussen, C. E., 2006. Gaussian processes for machine learning, vol. 2, MIT Press, Massachusetts, USA.

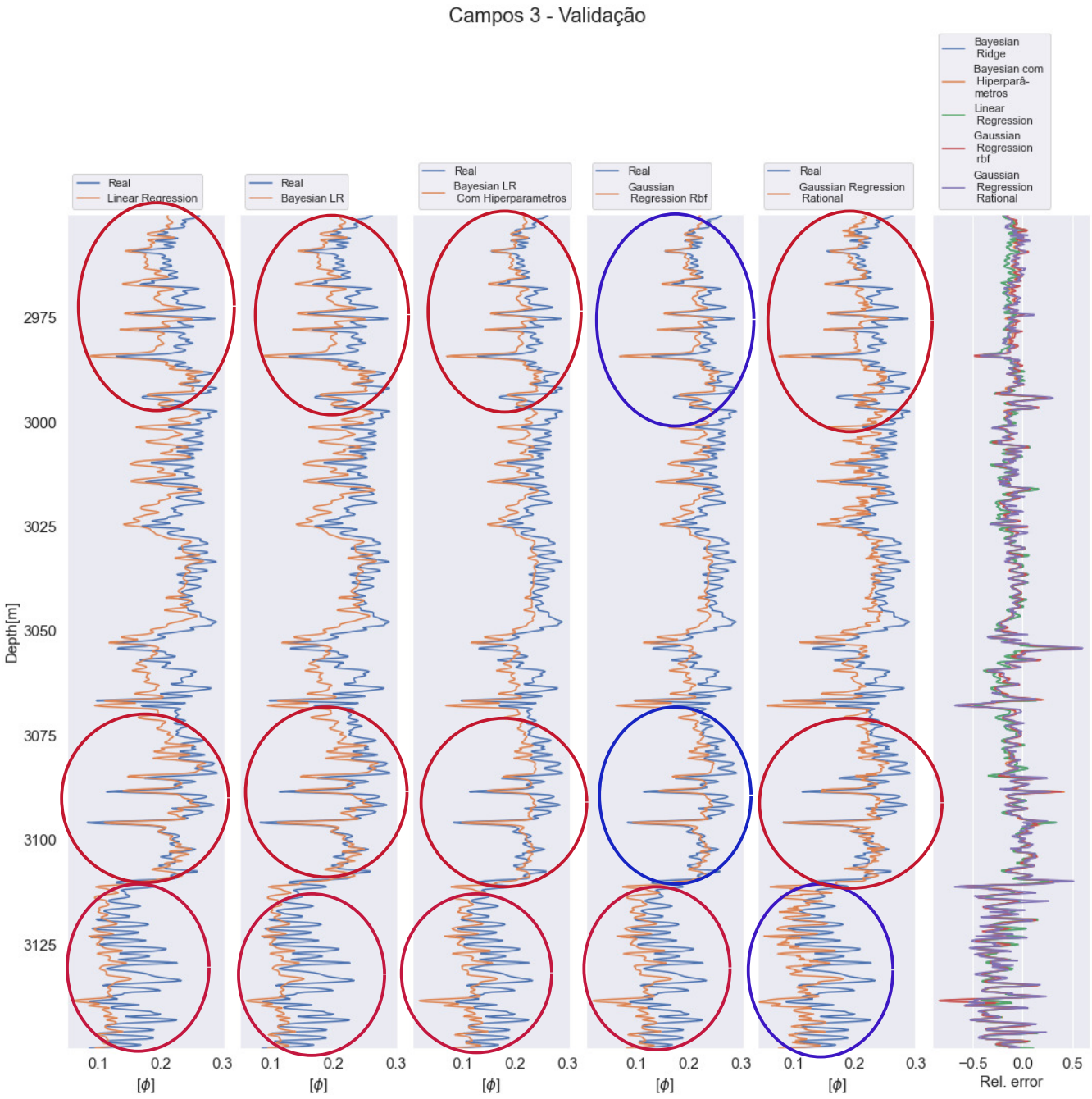

Figura 4 – Análise qualitativa entre as curvas reais (em azul) e preditas (em amarelo). A ultima coluna mostra o erro relativo na forma decimal.