Revista ARACÊ
ISSN: 2358-2472

MODELANDO O CUSTO DE CONSTRUÇÃO DE UMA SONDA DE PERFURAÇÃO MARÍTIMA ATRAVÉS DE FLORESTA ALEATÓRIA

MODELING THE COST OF CONSTRUCTING AN OFFSHORE DRILLING RIG USING RANDOM FOREST

MODELADO DEL COSTO DE CONSTRUCCIÓN DE UNA PLATAFORMA DE PERFORACIÓN MARINA UTILIZANDO RANDOM FOREST

https://doi.org/10.56238/arev7n11-001

Data de submissão: 03/10/2025 **Data de publicação:** 03/11/2025

Ricardo de Melo e Silva Accioly
D.Sc.
Instituição: Universidade do Estado do Rio de Janeiro (UERJ)
E-mail: raccioly@ime.uerj.br
Orcid: https://orcid.org/0000-0001-6513-3443

Fernanda da Serra Costa
D.Sc.
Instituição: Universidade do Estado do Rio de Janeiro (UERJ)
E-mail: fcosta@ime.uerj.br
Orcid: https://orcid.org/0000-0002-4414-7755

RESUMO

As sondas de perfuração marítimas são equipamentos vitais para a exploração e o desenvolvimento de campos de petróleo, consequentemente uma boa estimativa de seus custos de construção é fundamental para o planejamento de projetos de construção de sondas. Este artigo explora o desenvolvimento de um modelo de floresta aleatória (random forest) para prever os custos de construção de sondas de perfuração marítimas. O modelo busca fornecer estimativas precisas e confiáveis dos direcionadores de custos, com base em um conjunto de dados robusto que inclui custos históricos de construção e características de projeto das sondas. A abordagem de aprendizado por combinação de previsões, utilizada no algoritmo de floresta aleatória, permite capturar de forma eficaz relacionamentos e interações complexas nos dados, melhorando a precisão da previsão em comparação com os métodos de regressão tradicionais. Serão detalhadas a construção, a validação e as descobertas significativas do modelo, destacando sua capacidade de minimizar erros de estimativa e de apoiar a tomada de decisões na elaboração do orçamento do projeto. Os objetivos desta pesquisa abrangem a gestão de recursos, o controle de custos e o planejamento estratégico em projetos de construção de sondas.

Palavras-chave: Random Forest. Sonda de Perfuração Marítima. Direcionadores de Custos. Floresta Aleatória.

ABSTRACT

Offshore drilling rigs are vital equipment for oilfield exploration and development; therefore, accurate estimates of their construction costs are crucial for planning rig construction projects. This paper explores the development of a random forest model to forecast offshore drilling rig construction costs. The model aims to provide accurate and reliable estimates of cost drivers based on a robust dataset that includes historical construction costs and rig design characteristics. The prediction-based learning

Revista ARACÊ
ISSN: 2358-2472

approach used in the random forest algorithm effectively captures complex relationships and interactions in the data, improving forecast accuracy compared to traditional regression methods. The model's construction, validation, and significant findings will be detailed, highlighting its ability to minimize estimation errors and support decision-making in project budgeting. The objectives of this research encompass resource management, cost control, and strategic planning in rig construction projects.

Keywords: Random Forest. Offshore Drilling Rig. Cost Drivers.

RESUMEN

Las plataformas de perforación offshore son equipos vitales para la exploración y el desarrollo de yacimientos petrolíferos; por lo tanto, la estimación precisa de sus costos de construcción es crucial para la planificación de proyectos de construcción de plataformas. Este artículo explora el desarrollo de un modelo de bosque aleatorio para pronosticar los costos de construcción de plataformas de perforación offshore. El modelo busca proporcionar estimaciones precisas y confiables de los factores de costo basándose en un conjunto de datos robusto que incluye costos históricos de construcción y características de diseño de las plataformas. El enfoque de aprendizaje basado en predicciones utilizado en el algoritmo de bosque aleatorio captura eficazmente relaciones e interacciones complejas en los datos, mejorando la precisión del pronóstico en comparación con los métodos de regresión tradicionales. Se detallarán la construcción, validación y hallazgos significativos del modelo, destacando su capacidad para minimizar errores de estimación y respaldar la toma de decisiones en la presupuestación de proyectos. Los objetivos de esta investigación abarcan la gestión de recursos, el control de costos y la planificación estratégica en proyectos de construcción de plataformas.

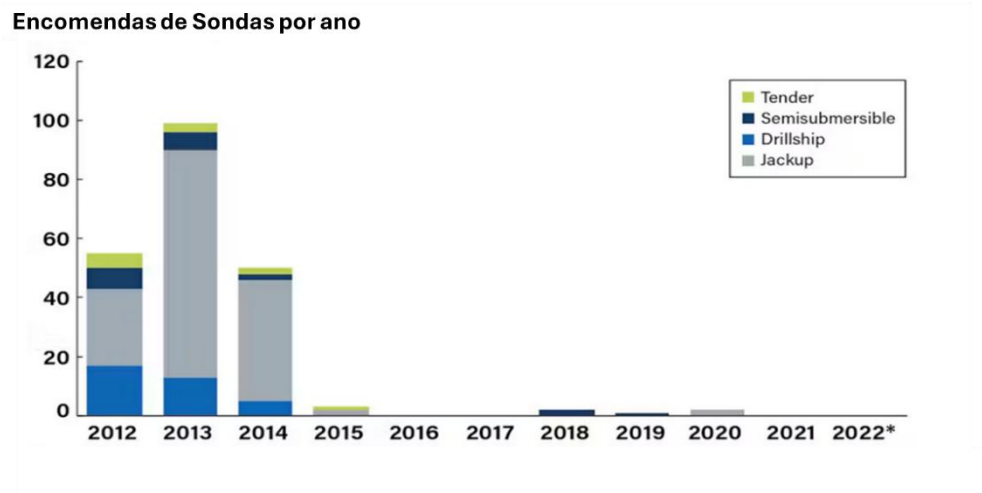
Palabras clave: Bosque Aleatorio. Plataforma de Perforación Offshore. Factores de Costo.

Revista ARACÊ
ISSN: 2358-2472

1 INTRODUÇÃO

No início da década de 2010, o mercado de construção de plataformas estava extremamente aquecido, com muitas novas construções, algumas delas especulativas, o que indicava que não havia contratos firmados para financiá-las. Conforme apresentado por Boman (2008), naquela época, 140 unidades móveis de perfuração offshore (MODU – MOBILE OFFSHORE DRILLING UNIT) estavam em construção, com outras encomendadas e em planejamento. Na Figura 1, vemos os pedidos de construção de sondas de 2012 a 2022 (SMITH, 2022), onde se pode observar que quando os preços do petróleo despencaram em 2014, o mercado de construção de plataformas foi fortemente afetado. Este efeito persiste nos anos seguintes.

FIGURA 1 – Encomendas de sondas entre 2012 e 2022



Fonte: Smith (2022)

De acordo com a Offshore Magazine (2023), a construção de um novo navio-sonda exige um investimento de quase 1 bilhão de dólares. Embora atualmente não haja perspectivas de um novo ciclo de construção, é essencial avaliar os fatores de custo que influenciam o CAPEX de uma nova MODU. Kaiser e Snyder (2012) e Kaiser et al. (2013) estudaram extensivamente a indústria de perfuração offshore e a construção de plataformas no Golfo do México e apontaram diversos fatores que influenciam os custos de construção das plataformas em seus estudos de 2012 e 2013, respectivamente. O modelo desenvolvido por eles (modelo linear multifatorial) para estimar os custos de construção, devido às suas características, limitou o tipo de relacionamento entre a variável resposta e as variáveis explicativas a um padrão linear.

Breiman et al. (1984) criaram modelos de árvores, que são poderosos para tarefas de classificação e regressão. O método divide os dados em subconjuntos com base em valores de variáveis explicativas, criando uma estrutura semelhante a uma árvore que representa decisões que conduzem a

Revista ARACÊ
ISSN: 2358-2472

resultados. No entanto, os modelos baseados em uma única árvore geralmente não são competitivos quanto à precisão de previsão e à estabilidade do modelo. Além disso, o número de previsões de uma árvore é finito e é determinado pelo número de nós terminais. Finalmente, essas árvores sofrem de um viés de seleção: preditores com valores mais distintos são favorecidos em relação a preditores mais granulares (HASTIE et al., 2009).

Posteriormente, Breiman (1996) aprimorou o desenvolvimento inicial por meio de técnicas de combinação de previsões (ensemble), utilizando o que denominou "bagging". "Bagging" é uma abreviação de "agregação por bootstrap". Em 2001, ele melhorou sua metodologia por meio de previsões com florestas aleatórias, que corrigiram o problema de correlação entre as árvores geradas pelas amostras de bootstrap, aprimorando ainda mais o método.

Esse tipo de modelagem apresenta várias características interessantes, tais como, ela aceita diferentes tipos de variáveis (esparsas, assimétricas, contínuas, categóricas etc.) sem a necessidade de pré-processamento; não exigem que o usuário especifique como as variáveis explicativas se relacionam com a resposta, como faz um modelo de regressão linear; podem lidar efetivamente com dados ausentes e conduzir implicitamente à seleção de variáveis, aspectos desejáveis para muitos problemas de modelagem da vida real.

2 REFERENCIAL TEÓRICO

2.1 MODELOS BASEADOS EM ÁRVORES

Na década de 1990, técnicas de combinação, ou "ensemble" (métodos que combinam previsões de vários modelos), começaram a surgir. *Bagging*, uma abreviação de "bootstrap aggregation", foi originalmente proposto por Leo Breiman e foi uma das primeiras técnicas de combinação desenvolvidas (BREIMAN, 1996). O *bagging* é uma abordagem geral que usa o bootstrapping em conjunto com qualquer modelo de regressão (ou de classificação) para gerar um conjunto de previsões.

A agregação por *bootstrap*, ou *bagging*, é um procedimento de uso geral para reduzir a variância de um método de aprendizagem estatístico. Ele é particularmente útil e frequentemente utilizado no contexto das árvores. Lembre-se de que, dado um conjunto de n observações independentes Z1, Z2, …, Zn, cada uma com variância σ^2, a variância da média das n observações é dada por σ^2/n. Em outras palavras, a média de um conjunto de observações reduz a variância. Em geral, isso não é prático porque não temos acesso a vários conjuntos de treinamento. Por outro lado, podemos usar o *bootstrap*, obtendo novas "amostras" por meio da reamostragem com reposição do conjunto de dados de treinamento (único). Nesta abordagem, geramos B conjuntos de dados de

Revista ARACÊ
ISSN: 2358-2472

treinamento distintos. Em seguida, treinamos o nosso método no b-ésimo conjunto de treinamento *bootstrap*, a fim de obter $\hat{f}^{*b}(x)$, uma previsão no ponto x.

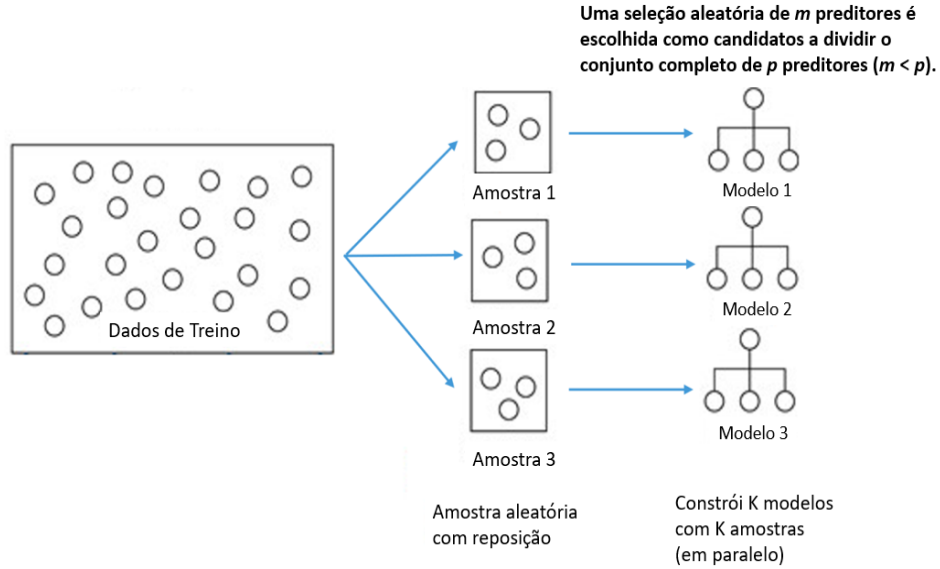
A média de todas as previsões é:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} \hat{f}^{*b}(x) \qquad (1)$$

A floresta aleatória (random forest) proporciona uma melhoria sobre o *bagging* por meio de um pequeno ajuste que descorrelaciona as árvores. Isso reduz a variância ao calcular a média das árvores. Neste método, ao construir as árvores, cada vez que uma divisão em uma árvore é considerada, uma seleção aleatória de m preditores é escolhida como candidatos a dividir o conjunto completo de p preditores ($m < p$). A divisão somente é permitida para um desses m preditores. Uma nova seleção de m preditores é realizada em cada modelo.

No contexto das regressões, Breiman (2001) recomenda que m seja igual a um terço do número de preditores. As Figuras 2 e 3 ilustram os conceitos mencionados acima.

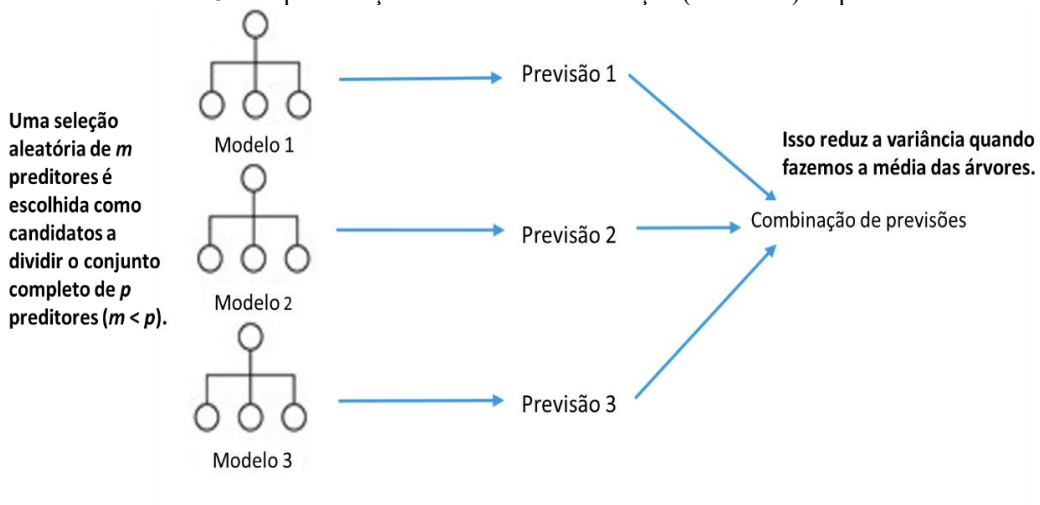
FIGURA 2 – Representação da metodologia de floresta aleatória



Fonte: Autoria própria (2025)

Revista ARACÊ
ISSN: 2358-2472

FIGURA 3 – Representação do conceito combinação (ensemble) de previsões



Fonte: Autoria própria (2025)

2.2 INTERPRETAÇÃO DE MODELOS COMPLEXOS

Uma questão importante quando se trabalha com previsões é interpretá-las, de modo a compreendê-las e justificá-las. As técnicas de combinação, ou "ensemble" (métodos que combinam previsões de muitos modelos), aumentaram a capacidade preditiva dos modelos ao reduzir os erros de previsão. No entanto, no caso das árvores, ao combinarem as respostas de diversas árvores, perdeu-se a facilidade de interpretação de uma única árvore. Este problema ocorre em todos os modelos que utilizam a combinação de previsão, mas também em outros modelos que costumam ser denominados caixas pretas ("black-box"), tais como redes neurais e SVM. Para melhorar esta questão, foram introduzidas algumas técnicas, tais como LIME (Local Interpretable Model-agnostic Explanations) e o SHAP (SHapley Additive exPlanations), segundo Molnar (2019). Aqui, usaremos o LIME, desenvolvido por Ribeiro et al. (2016).

O LIME é uma abordagem baseada na ideia de que modelos complexos podem ser aproximados localmente por modelos simples. Para interpretar a predição de um dado ponto:

- São geradas várias versões perturbadas dessa observação (por exemplo, alterando ligeiramente os valores das variáveis).
- O modelo complexo é usado para prever essas versões.
- Um modelo linear (ou outro simples) é ajustado localmente com base nessas previsões.
- Os coeficientes desse modelo explicativo indicam a importância das variáveis na predição local.

Revista ARACÊ
ISSN: 2358-2472

3 METODOLOGIA

A metodologia de floresta aleatória requer ajustar um conjunto de parâmetros para definir o comportamento do modelo de previsão. Os principais parâmetros deste modelo são: o número de árvores, o número de variáveis consideradas em cada divisão, o tamanho mínimo de observações em cada nó terminal (folha), a regra de divisão e o método de cálculo da importância das variáveis. No problema em questão, a regra de divisão é a variância, método usual em problemas de regressão. Já para o cálculo da importância será utilizada a permutação, pois o custo computacional neste problema é pequeno, mas permite avaliações mais robustas e confiáveis (BREIMAN, 2001 e STROBL et al., 2007).

Os demais parâmetros serão definidos por meio de validação cruzada, utilizando um grid de valores para cada um deles.

A base de dados contém 101 sondas de perfuração marítima, entregues e em construção entre 2001 e 2010, e foi coletada em 2011 por meio de consultas a operadores de sonda, publicações da IADC (International Association of Drilling Contractors), da Offshore Magazine e de outras fontes. A partir da base de dados com 40 variáveis explicativas (features) relacionadas a essas sondas, foi construído um modelo de floresta aleatória para prever e avaliar a importância de cada característica. Essas variáveis explicativas atuam como direcionadores de custo e são essenciais para prever os custos reais de construção. A Tabela 1 apresenta as variáveis explicativas, sendo a variável dependente o CAPEX em MMUS$ de 2011.

TABELA 1 – Variáveis (features) da base de dados

Nome da Variável	Descrição
Rig Type	Tipo de Sonda (Semi-submersível ou Navio Sonda)
Rig Water Depth	Lâmina D'água (pés)
DP	Sonda de Posicionamento Dinâmico (S/N)
DP Class	Classe de Posicionamento Dinâmico
Design Water Depth	Lâmina d'água de projeto (pés)
Drilling Depth	Profundidade de perfuração (pés)
BOP WP Max	Pressão máxima suportada pelo BOP (psi)
Draft	Comprimento submerso (pés)
Dual Activity	Dupla atividade (S/N)
EWT Capable	Capacidade de perfuração EWT (S/N)
Harsh Environment	Capacidade de operação em ambiente severo (S/N)
North Sea Capable	Capacidade de operação no mar do Norte (S/N)
Zero Discharge	Descarga zero (S/N)
Thruster Assist	Com presença de thrusters (S/N)
Top Drive	Com presença de topdrive (S/N)
Quarters Capacity	Capacidade de acomodação
Variable Load	Carga variável (mt)
Hull Newbuild	Casco Novo (S/N)
AnoTerm	Ano de término da Construção
AnoPed	Ano do pedido da sonda

Revista ARACÊ
ISSN: 2358-2472

CAPEX 2011	Capex em MMUS$ de 2011
Spec Build	Construção especulativa (S/N)
Year Hull Built	Ano de construção do casco
Year In Service	Anos em operação
Norwegian SUT	Declaração de conformidade da Noruega (S/N)
Hull Length	Comprimento do casco (pés)
Hull Breadth	Largura do casco (pés)
Hull Depth	Profundidade do casco (pés)
Bulk Cement	Capacidade de cimento (pés cúbicos)
Bulk Mud	Capacidade de lama (pés cúbicos)
Liquid Mud	Capacidade de lama (bbl)
Drill Water	Capacidade água industrial (bbl)
Potable Water	Capacidade água potável (bbl)
BOP Qty	Quantidade de BOPs
Engine Qty	Quantidade de máquinas
Generator Qty	Quantidade de Geradores
Derrick Or Mast	Torres ou Mastro (T/M)
Derrick Capacity	Capacidade de carga da Torre (lbs)
Mudpump Qty	Quantidade de bombas de lama
Tensioner Qty	Quantidade de tensionadores de riser
Tensioner Capacity	Capacidade do Tensionador Riser (lbs)

Fonte: Autoria própria (2025)

Para facilitar a interpretação dos resultados de previsão, utilizou-se a metodologia LIME (MOLNAR, 2019).

4 APLICAÇÃO

Para a aplicação do modelo de floresta aleatória foi desenvolvido um programa na linguagem R, versão 4.5.1, utilizando o pacote ranger (WRIGHT e ZIEGLER, 2017), versão 0.17.0. Para a interpretação dos resultados, utilizou-se a biblioteca lime (HVITFELDT et al., 2022), na versão 0.5.3. O ambiente de programação foi o RStudio 2025.09.1.

A base de dados continha dados de 101 sondas, sondas semi-submersíveis e navios-sonda, com 40 especificações técnicas (features), que foram separados em dois conjuntos: 75% usados para treino e os demais para teste do modelo. Para a definição do melhor conjunto de parâmetros, foi definido um grid de valores para o número de árvores (num.trees no ranger), o número de variáveis consideradas em cada divisão (mtry), o tamanho mínimo de observações em cada nó terminal (min.node.size) e a regra de divisão (*splitrule*: somente variância). O grid continha 70 configurações diferentes. A Tabela 2 apresenta os valores selecionados.

Revista ARACÊ
ISSN: 2358-2472

TABELA 2 – Grid de parâmetros do modelo de floresta aleatória

Parâmetro	Valores
num.trees (# árvores)	100, 200, 300, 400, 500
mtry (# máx. var. explicativas na divisão)	2, 4, 6, 8, 10, 12, 14
min.node.size (# mín. de observações no nó)	1, 5

Fonte: Autoria própria (2025)

A escolha dos melhores parâmetros foi realizada por meio de validação cruzada no conjunto de treino, com 5 envelopes de dados. Na Tabela 3 vemos os resultados.

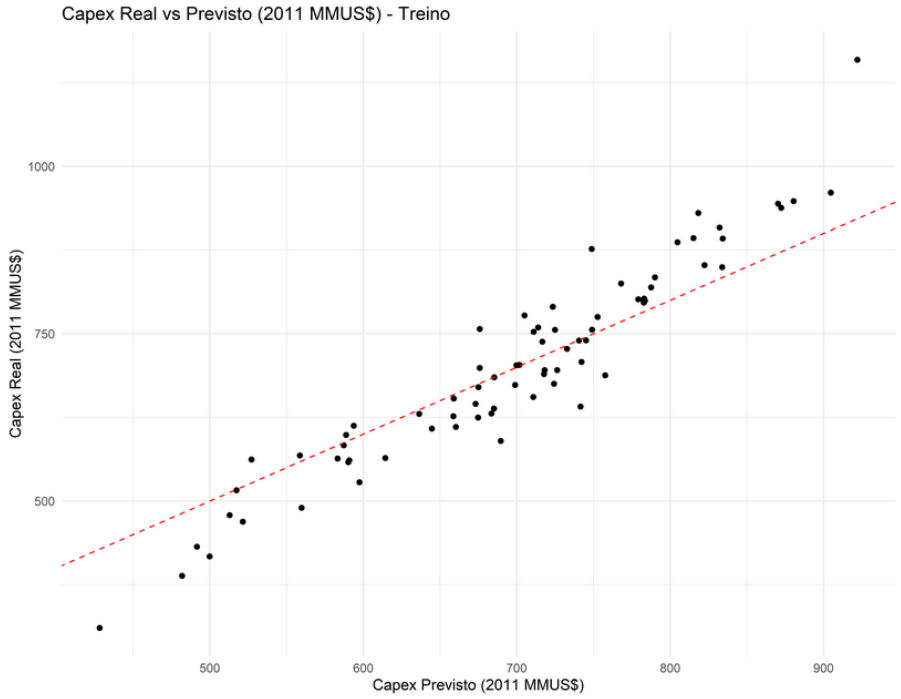
TABELA 3 – Melhores (cinco) resultados da validação cruzada

mtry	min.node.size	splitrule	num.trees	RMSE médio
2	1	Variância	100	105,2259
2	1	Variância	200	106,0781
2	1	Variância	300	106,3491
2	1	Variância	400	106,3901
2	1	Variância	500	106,4608

Fonte: Autoria própria (2025)

A Figura 4 apresenta a comparação entre os valores previstos e os valores reais (conjunto de treino), evidenciando um bom desempenho do modelo.

FIGURA 4 – Comparação entre valor real e previsão com floresta aleatória - Treino

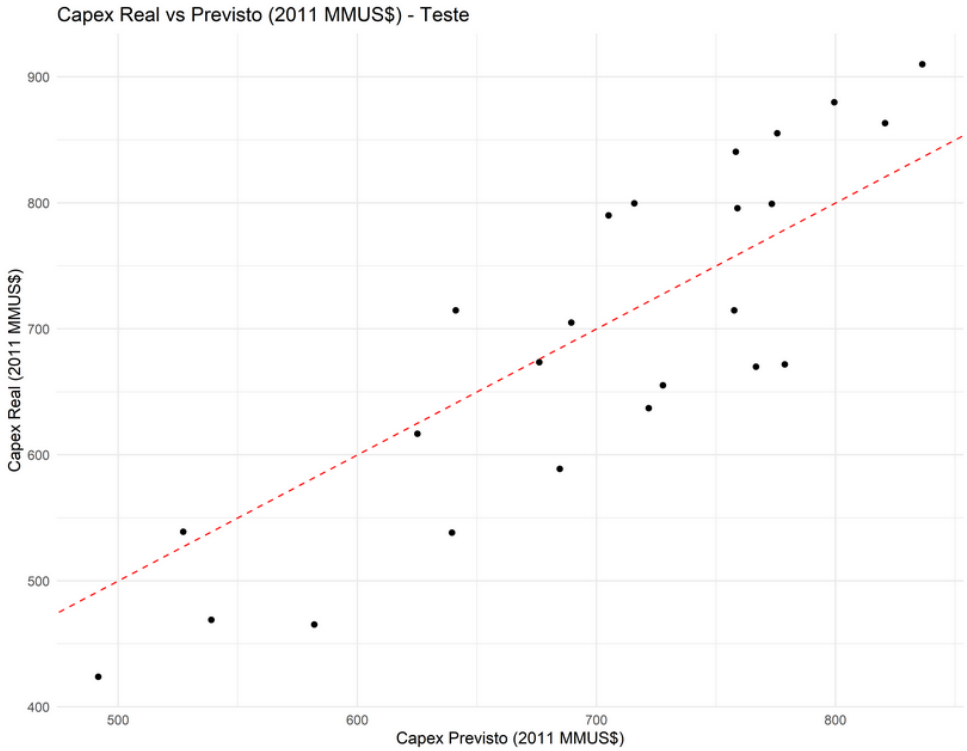


Fonte: Autoria própria (2025)

A Figura 5 apresenta a comparação entre os valores previstos e os valores reais (conjunto de teste), evidenciando um bom desempenho do modelo.

Revista ARACÊ
ISSN: 2358-2472

FIGURA 5 – Comparação entre valor real e previsão com floresta aleatória – Teste



Fonte: Autoria própria (2025).

Na Tabela 4 vemos as estatísticas comparativas dos resultados dos ajustes dos conjuntos de treino e de teste.

TABELA 4 – Métrica por conjunto de dados (treino e teste)

Conjunto	RMSE	MAE	MAPE	R2
Treino	58,14	44.44	6.70	0,91
Teste	72,63	64,89	9,94	0,76

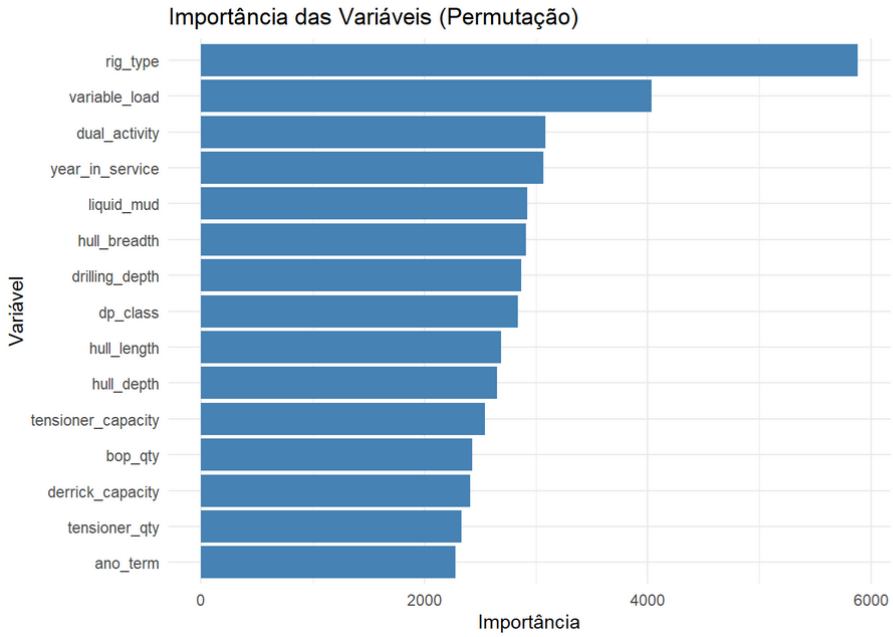
Fonte: Autoria própria (2025)

As estatísticas indicam um bom desempenho do modelo tanto na fase de treinamento quanto na de teste. O resultado mais importante é o de teste, em que, observando o MAPE, temos um erro previsto em torno de 10%, o que pode ser considerado adequado neste contexto.

A importância das variáveis pode ser observada na Figura 6. Só foram apresentadas as 15 variáveis mais importantes. Podemos observar que o tipo de sonda foi a variável mais importante. Os navios sondas têm custo de construção maior do que o das semi-submersíveis, como podemos ver em Kaiser et al. (2013). O navio sonda tem maior capacidade de carga variável (variable_load) do que as sondas semi-submersíveis.

Revista ARACÊ
ISSN: 2358-2472

FIGURA 6 – Importância das variáveis do modelo



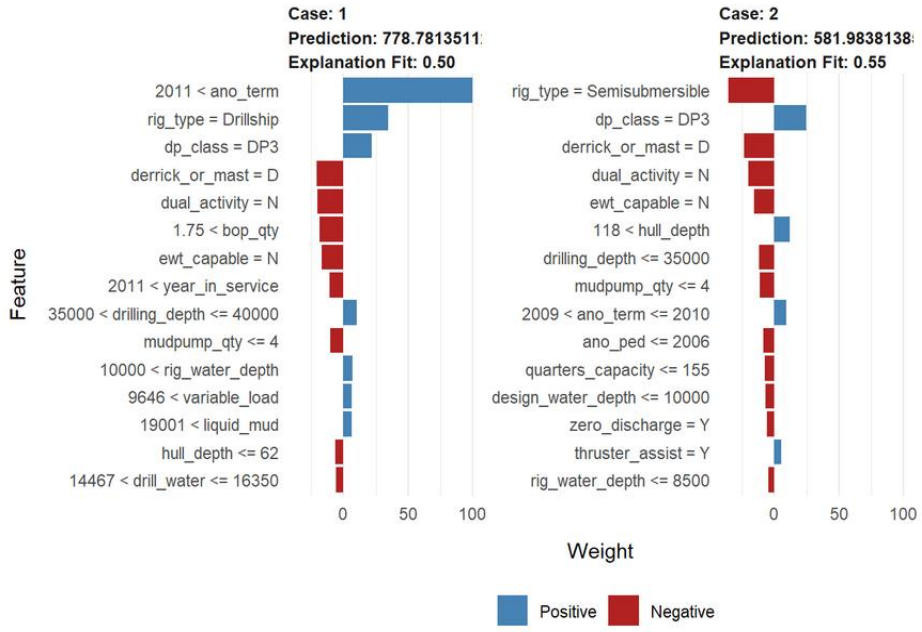
Fonte: Autoria própria (2025)

Na Figura 7 vemos o resultado do LIME nas 1ª e 2ª observações do conjunto de teste. Nela, podemos perceber o impacto do tipo de sonda: aumento de valor quando é um navio-sonda e perda de valor quando é uma sonda semi-submersível. O tipo de classe de posicionamento dinâmico foi positivo em ambas as observações. Outro ponto interessante é que as duas sondas perderam valor por não terem dupla atividade (dual_activity). Sondas com dupla atividade possuem duas torres, o que permite acelerar certas operações, ou seja, ganhar tempo na construção dos poços.

Nestes dois exemplos, foi possível explorar os fatores que afetaram as previsões, o que é muito importante para compreendê-las e justificá-las. Esta análise também permite identificarmos as fontes de desvio em relação aos valores reais.

Revista ARACÊ
ISSN: 2358-2472

FIGURA 7 – Avaliação dos resultados das 1ª e 2ª observações do conjunto de teste



Fonte: Autoria própria (2025)

5 CONCLUSÕES

A floresta aleatória representa uma metodologia poderosa na construção de modelos de previsão, pois trabalha com diversos tipos de variáveis e permite relacionamentos complexos (não lineares) entre elas. A identificação da importância das variáveis é fundamental para selecionar as mais relevantes, o que facilitaria a obtenção de um modelo mais parcimonioso.

A existência de técnicas de interpretação de modelos complexos sana a dificuldade de compreender os resultados desses modelos, permitindo identificar o que levou à obtenção de uma determinada previsão. Neste trabalho, utilizamos o LIME, mas o SHAP (MOLNAR, 2019) vem tendo seu uso ampliado.

Quanto aos resultados, o modelo permitiu identificar que o tipo de sonda é o fator de maior impacto no custo de construção. Em segundo lugar, a capacidade de carga variável da sonda. Em terceiro lugar, a dupla atividade, ou seja, a capacidade de realizar operações simultâneas. Este tipo de informação é importante no processo de seleção das características técnicas de um projeto.

Revista ARACÊ
ISSN: 2358-2472

REFERÊNCIAS

BOMAN, K. Rig construction boom continues. Offshore Magazine, 2008. Disponível em: https://www.offshore-mag.com/business-briefs/equipment-engineering/article/16761659/rig-construction-boom-continues. Acesso em: 13 jun. 2025.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A., & STONE, C. J. Classification and Regression Trees. Wadsworth and Brooks/Cole Advanced Books & Software, 1986.

BREIMAN, L. Random Forests. Machine Learning, 45(1), 5-32, 2001.

HASTIE, T., TIBISHIRANI, R., & FRIEDMAN, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., Springer, 2009.

HVITFELDT, E., PEDERSEN, T.L., BENESTY, M. lime: Local Interpretable Model-Agnostic Explanations, 2022. Disponível em: https://doi.org/10.32614/CRAN.package.lime. Acesso em: 15 jun. 2025.

KAISER, M. J., SNYDER, B. F. Reviewing rig construction cost factors, 2012. Disponível em: https://www.offshore-mag.com/business-briefs/equipment-engineering/article/16760123/reviewing-rig-construction-cost-factors. Acesso em: 13 jun. 2025.

KAISER, M. J., SNYDER, B., PULSIPHER, A.G. Offshore Drilling Industry and Rig Construction Market in the Gulf of Mexico, Louisiana State University Center for Energy Studies Coastal Marine Institute, OCS Study BOEM 2013-0112, 2013.

OFFSHORE MAGAZINE Report: No resurgence seen for newbuild rig construction, 2023. Disponível em: https://www.offshore-mag.com/rigs/article/14298424/westwood-global-energy-group-report-no-resurgence-seen-for-newbuild-rig-construction. Acesso em: 13 jun. 2025.

MOLNAR, C. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, Lulu.com, 2019.

RIBEIRO, M.T., SINGH, S., GUESTRIN, C Why Should I Trust You? Explaining the Predictions of Any Classifier, arXiv:1602.04938, 2016. Disponível em: https://doi.org/10.48550/arXiv.1602.04938. Acesso em: 13 jun. 2025.

SMITH, J. Rig construction market remains quiet but with room for long-term possibilities. Offshore Magazine, 82(6), 2022.

WRIGHT, M. N., ZIEGLER, A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R, Journal of Statistical Software, 77(1), 1–17, 2017.