Machine Learning Aplicado à Predição da Obrigação de Investimento em P,D&I

Flávia Bravo, Luciana Sousa, Tatiana Escovedo, Helio Lopes, Marcos Kalinowski

Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro, RJ, Brasil

flaviabravo@gmail.com, luandrea.sousa@gmail.com,
{tatiana,lopes,kalinowski}@inf.puc-rio.br

Abstract. *Investments in Research, Development, and Innovation (R&D&I) from Brazil's oil and gas sector are substantial due to the obligation established by the National Agency of Petroleum, Natural Gas and Biofuels (ANP). Identifying the expectation of funding in an agile and simple way enables better planning, increasing the effectiveness of expenditures. This article proposes elaborating a machine learning model to predict the potential of mandatory investments that companies in the oil and gas sector must make in R&D&I, allowing better planning of the application of financial resources for universities and science and technology institutes.*

Resumo. *Os investimentos em Pesquisa, Desenvolvimento e Inovação (P,D&I) do setor de petróleo e gás do Brasil são substanciais devido à obrigatoriedade estabelecida pela Agência Nacional de Petróleo, Gás Natural e Biocombustíveis (ANP). Identificar a expectativa de financiamento de forma ágil e simples possibilita um melhor planejamento, aumentando a eficácia dos gastos. Este artigo propõe a elaboração de um modelo de machine learning para estimar o potencial de investimentos obrigatórios que as empresas do setor de óleo e gás devem realizar em P,D&I, permitindo um melhor planejamento da aplicação de recursos financeiros para universidades e institutos de ciência e tecnologia.*

1. Introdução

No Brasil, a Agência Nacional de Petróleo, Gás Natural e Biocombustíveis (ANP) estipula a obrigação pelas empresas petrolíferas de realização de atividades de Pesquisa, Desenvolvimento e Inovação (P,D&I) através de cláusulas nos contratos para exploração, desenvolvimento e produção de petróleo e gás natural, seguindo as regras estabelecidas no Regulamento Técnico ANP Nº 3/20 05 [Regulamento, 2015].

Assim, os investimentos em P,D&I oriundos do setor de óleo e gás no Brasil são vultosos devido à obrigação estabelecida pela ANP. Identificar de forma ágil e simples a expectativa de captação financeira viabiliza melhor planejamento, aumentando a eficácia dos dispêndios. Entretanto, antes de reivindicar o aumento deste investimento para se aproximar da média dos países da Organização para a Cooperação e Desenvolvimento Econômico (OCDE), é mais sensato priorizar o aumento da eficácia dos dispêndios existentes [Leal & Figueiredo, 2021].

Neste sentido, este artigo tem como objetivo principal auxiliar na melhora do planejamento para direcionamento de recursos e políticas públicas para Universidades e Institutos de Ciência e Tecnologia através da elaboração de um modelo para previsão do

potencial de investimentos obrigatórios que as empresas do setor de óleo e gás devem realizar em P,D&I.

A ANP estabelece fórmulas para cálculo do valor a ser investido obrigatoriamente em P,D&I. No entanto, tais fórmulas são complexas, pois envolvem diversos parâmetros difíceis de estimar, como o tipo de contrato de produção; as características de qualidade das correntes de óleo e gás, como grau API, teor de enxofre, nitrogênio, acidez naftênica; o tipo de rendimento das frações (leves, médias e pesadas); além dos diferenciais de preços dadas estas variadas características.

Desta forma, este trabalho visa avaliar a utilização de algoritmos de *Machine Learning* (ML) para simplificar o cálculo do valor da obrigação legal de investimento P,D&I, tornando-o mais rápido e simples. A posse desta informação de forma antecipada pode auxiliar na mudança de regras de distribuição de recursos de forma fácil e ágil. Esta análise pode ser feita em função do valor a ser destinado por ano, dos mercados em potencial do país e das instituições de Ciência e Tecnologia que devem ser envolvidas. Assim, os principais benefícios desta pesquisa são:

- Simplificação do cálculo de obrigação legal de P,D&I, possibilitando maior previsibilidade do potencial de investimentos obrigatórios;
- Visão de longo prazo, viabilizando que a ANP possa adequar políticas públicas conforme a previsão da receita, adequando a regulamentação técnica de P,D&I;
- Visibilidade da expectativa de investimentos obrigatórios em P,D&I para as demais instituições interessadas, como Universidades e instituições de ciência e tecnologia.

Todas estas ações contribuem para o melhor uso do dinheiro público, uma vez que o investimento eficaz em P,D&I é um direito social das gerações futuras do Brasil e um dever da geração presente [Leal & Figueiredo, 2021].

Este artigo está organizado em quatro seções adicionais. A Seção 2 apresenta a fundamentação teórica dos assuntos relacionados ao tema deste trabalho. Ressaltamos que não foram incluídos trabalhos relacionados, pois não foram encontrados na literatura trabalhos tratando especificamente o cálculo da obrigação de investimento em P,D&I. A Seção 3 descreve o problema que pretendemos resolver, descrevendo os motivadores e objetivos pretendidos. A Seção 4 detalha a solução proposta com os experimentos e resultados atingidos. Finalmente, na Seção 5 são relatadas as conclusões identificadas.

2. Fundamentação Teórica

Esta seção aborda conceitos e informações relevantes sobre a ANP, Obrigação Legal em P,D&I e *Machine Learning*, visando facilitar o entendimento global do trabalho.

2.1. Agência Nacional de Petróleo, Gás Natural e Biocombustível (ANP)

A Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP) foi estabelecida pelo Decreto nº 2.455, de 14 de janeiro 1998, e já estava prevista na Lei nº 9.478, de 6 de agosto de 1997, sendo o órgão encarregado de regular a indústria do petróleo e seus derivados, do gás natural e dos biocombustíveis, atuando em diversos segmentos.

Neste artigo, abordaremos o segmento de pesquisa, desenvolvimento e inovação. Em relação a este segmento, os contratos de exploração e produção de petróleo e gás incluem uma cláusula de P,D&I que exige das empresas signatárias o compromisso de investirem em projetos de pesquisa, desenvolvimento e inovação voltados para o setor regulado (obrigação legal em P,D&I). Estes recursos são aplicados mediante autorização da ANP [Acesso à Informação, 2020].

2.2. Obrigação Legal em P,D&I

Segundo o Regulamento Técnico ANP Nº 3/2005 [Regulamento, 2015], o fato gerador e valor da obrigação está vinculado à modalidade dos contratos originais e respectivos termos aditivos. Os recursos da obrigação de investimento em P,D&I devem ser aplicados em projetos ou programas de P,D&I executados no País.

A obrigação em P,D&I é um percentual da Receita Bruta Anual. Atualmente, para identificar a Receita Bruta, utiliza-se a combinação de diversas fórmulas:

Cálculo da Receita Bruta Mensal (RB mensal):

$$RBmensal = (VPFpetróleo\ X\ Ppetróleo) + (VPFgás\ X\ Pgás) \quad (1)$$

onde *VPFpetróleo* é volume de produção fiscalizada de petróleo, em m^3; *Ppetróleo* o preço de referência de petróleo, em R\$/m^3; *VPFgás* o volume de produção fiscalizada de gás natural, em m^3 e *Pgás* o preço de referência do gás natural, em R\$/m^3 [Manual, 2017]. Para identificação do preço de referência de petróleo utiliza-se a Equação 2 [Preços, 2020]:

$$Ppetróleo = TC\ X\ 6{,}2898\ X\ (PPref + Dq) \quad (2)$$

onde *TC* é a média mensal das taxas de câmbio diárias para compra de dólar americano, segundo o Banco Central; *6,2898* é uma constante utilizada para conversão volumétrica de metros cúbicos para barris de petróleo; *PPref* o valor médio mensal dos preços diários do petróleo utilizado como referência internacional, definido no art. 20, inciso XI, Resolução ANP nº 703/17, em dólares americanos por barril, para o mês cujo preço se calcula e *Dq* o diferencial de qualidade entre o petróleo nacional e o petróleo de referência. O diferencial de qualidade (*Dq*) obtém-se através da Equação 3 [Preços, 2020]:

$$Dq = VBPnac - VBPref - S - A - N \quad (3)$$

onde *VBPnac* é o valor bruto dos produtos derivados do petróleo nacional, em dólares americanos por barril, sendo o valor das frações (rendimentos) leves, médias e pesadas, decorrentes da destilação do petróleo nacional avaliado, calculado com base nos preços no mercado internacional de cada derivado; *VBPref* é o valor bruto dos produtos derivados do petróleo de referência, em dólares americanos por barril, sendo o valor das frações (rendimentos) leves, médias e pesadas, decorrentes da destilação do petróleo de referência, calculado com base nos preços do mercado internacional de cada derivado constante; *S* é o deságio dado aos petróleos com teor de enxofre superior a 0,60% m/m, em dólares americanos por barril; *A* é o deságio dado aos petróleos com TAN (Índice de acidez naftênica) superior a 0,50 mgKOH/g, em dólares americanos por barril e *N* é o deságio dado aos petróleos com teor de nitrogênio superior a 0,25% m/m, em dólares

americanos por barril. Para encontrar o Valor Bruto do Petróleo (VBP), tanto Nacional quanto o de referência, recorre-se à Equação 4 [Preços, 2020]:

$$VBP = (Fl\ X\ Pl) + (Fm\ X\ Pm) + (Fp\ X\ Pp) \qquad (4)$$

onde *Fl* é a fração dos destilados leves; *Fm* a fração dos destilados médios; *Fp* a fração dos destilados pesados; *Pl* o preço da fração dos destilados leves; *Pm* o preço da fração dos destilados médios e *Pp* o preço da fração dos destilados pesados.

É possível observar que tais fórmulas são complexas, envolvendo diversos parâmetros, o que torna complicado identificar, de forma simples e direta, a obrigação legal em P,D&I para fins de planejamento.

2.3. *Machine Learning*

Machine Learning, ou Aprendizado de Máquina, é o campo de estudo que fornece ao sistema computacional a capacidade de aprender coisas que não foram explicitamente programadas. É uma subárea de Inteligência Artificial que se concentra na descoberta de padrões ou de fórmulas matemáticas que expliquem o relacionamento entre os dados [Escovedo & Koshiyama, 2020]. Neste artigo, será abordado um problema de regressão, ou seja, quando se deseja fazer a predição de um valor numérico, contínuo ou discreto. Para melhorar o desempenho do modelo final, é possível a utilização de modelos do tipo *Ensemble*, no qual vários modelos são estrategicamente gerados e combinados para resolver um problema específico de inteligência computacional [Mahesh, 2020]. Esta combinação de dados e modelos constitui uma agregação de informações, uma vez que diferentes modelos podem extrair diferentes informações embutidas nos dados [Winkler, 1989].

3. Descrição do Problema

A indústria de óleo e gás movimenta um volume vultoso de recursos, ocupando o terceiro lugar no ranking das principais atividades econômicas no Brasil [Pesquisa, 2015]. Assim, impacta em diferentes setores e é um dos principais vetores do desenvolvimento socioeconômico do Brasil, alavancando a economia através da geração de emprego e renda, e do investimento em Pesquisa, Desenvolvimento e Inovação [Firmo, 2019].

Através deste grande volume de recursos, a ANP fomenta o desenvolvimento da pesquisa tecnológica e inovação no Brasil desde que estabeleceu regulamento que obriga empresas do setor a destinar um percentual de suas receitas a projetos de pesquisa, que podem ser realizados em parcerias entre centros de pesquisa, universidades e fornecedores, estimulando e viabilizando inovação na cadeia produtiva.

Cabe à ANP, através dos regulamentos, estabelecer como tais recursos devem ser investidos, como bolsas de estudos para formação de mão de obra especializada para inserção no mercado de trabalho e no desenvolvimento de novas pesquisas, fomento de inovações tecnológicas para ampliar participação dos biocombustíveis na matriz energética nacional, dentre outros [Polybio et al., 2012]. Sendo assim, é importante identificar qual será o volume disponível destes recursos para que se possa planejar os investimentos e revisitar o regulamento para direcioná-los aos projetos que mais contribuirão com o desenvolvimento tecnológico desejado, e gerenciá-los da forma eficiente e eficaz.

No entanto, conforme já mencionado, o cálculo da estimativa destes recursos é complexo, uma vez que depende do tipo de contrato de exploração de cada campo em produção e da estimativa de receita a ser obtida. A receita é estimada através de fórmulas que envolvem inúmeras questões, apresentadas na Seção 2, como o cálculo das variações de preços do óleo de cada campo dadas suas caraterísticas de qualidade.

Desta forma, o artigo avaliará um modelo de *Machine Learning* para possibilitar uma predição mais simples da obrigação de investimento em P,D&I no setor de petróleo, gás natural, biocombustíveis e outras fontes de energia renováveis correspondentes a este setor.

4. Solução Proposta

Esta Seção descreve as fontes de dados utilizadas e as operações de pré-processamento realizadas e detalha a construção dos modelos e experimentos realizados, apresentando as suas principais características e os resultados alcançados por cada um deles.

4.1. Fontes de Dados e Pré-Processamento

A base de dados para construção do modelo foi elaborada utilizando dados abertos do Governo, da ANP, do Banco Central do Brasil e sites de referência para séries históricas de preços de óleo e gás. As quatro fontes de dados encontram-se descritas a seguir:

1. Agência Nacional de Petróleo, Gás Natural e Biocombustíveis:

- Volume de Produção Nacional de Petróleo e Gás Natural [Dados, 2021].
- Volume de obrigações geradas por ano (de 2000 até de 2020). Esta base foi a fonte utilizada para identificação do Investimento em P,D&I por mês (R$) (mil) [Recursos, 2021] .

2. *Investing.com*:

- Valor médio mensal do petróleo *Brent* [Brent, 2021].
- Valor médio mensal do Gás Natural [Gás Natural, 2021].

3. Banco Central do Brasil:

- Taxa de câmbio (US$) diária [Estatísticas, 2021].

4. Instituto Brasileiro de Geografia e Estatística:

- IPCA – Número índice [IPCA, 2021].

As oito variáveis utilizadas destas fontes foram:

1. Ano: ano de referência dos atributos.
2. Mês: mês de referência dos atributos.
3. Produção de Petróleo (b) (mil): volume de produção de petróleo no Brasil em barris.
4. PP ref - *Brent* (último) (USD): preço médio do petróleo do tipo *brent* em USD por barril.

5. TC: Taxa de câmbio entre o Real e o Dólar (USD).
6. Produção nacional de gás natural (b): volume de produção de gás natural no Brasil em barris.
7. Preço Gás Natural (1 Mmbtu): preço médio do gás natural por Mmbtu.
8. Investimento P,D&I por mês (R$) (mil): valor destinado obrigatoriamente a investimentos em pesquisa e desenvolvimento no Brasil.

Como as bases foram obtidas de fontes diferentes, com unidades distintas, foi necessário consolidar o *dataset* gerado e realizar as seguintes operações de pré-processamento:

- Cálculo do investimento em P,D&I mensal obtido pela multiplicação do percentual do volume de produção de óleo e gás de cada mês pelo volume de obrigações geradas por ano, uma vez que o cálculo da obrigação é função do volume de produção;
- Cálculo da média mensal das cotações diárias de dólar, segundo o Banco Central, para identificação da cotação mensal do dólar;
- Conversão do volume de produção de gás natural e o volume de produção de petróleo de milhões para milhares;
- Ajustes nas bases para padronização da formatação, como transposição de linhas para colunas e ajustes nos títulos das colunas.

Após conclusão das transformações, a base de dados foi consolidada contendo 7 colunas e 252 linhas, representadas pelos meses dos anos de 2000 a 2020, conforme extrato do conjunto de dados constante na Figura 1.

Figura 1. Extrato do conjunto de dados

	Ano	Mês	Produção de Petróleo (b) (mil)	PP ref - Brent (último) (US$)	TC - taxa de Câmbio	Produção nacional de gás natural (b)	Preço Gás Natural (1 Mmbtu)	Investimento P&D por mês (R$) (mil)
0	2000	Janeiro	35.7940	25.97	1.8037	1.0039	2.662	7.4589
1	2000	Fevereiro	32.5014	28.09	1.7753	0.9030	2.761	6.7711
2	2000	Março	36.8677	24.77	1.7420	1.1339	2.945	7.7029
3	2000	Abril	34.8374	23.89	1.7682	1.1393	3.141	7.2925
4	2000	Maio	35.9348	28.31	1.8279	1.2352	4.356	7.5343

4.2. Construção dos Modelos e Experimentos Realizados

Considerando que o objetivo do artigo é desenvolver um modelo de aprendizagem supervisionada a partir de dados históricos, cujo resultado deverá ser numérico, utilizaremos algoritmos de regressão para que, dado um novo padrão de dados, seja possível estimar o valor esperado para a variável resposta.

Para avaliar cada modelo, serão utilizadas as métricas de erro quadrático médio, ou *mean square error* (MSE) e o coeficiente de determinação (R2). O MSE fornece uma perspectiva de magnitude do erro, porém sem indicar direção do erro. Já o R2, quanto mais próximo de 1, indica um melhor ajuste do modelo, ou seja, o quanto a variável resposta é explicada pelas variáveis independentes.

Os algoritmos de regressão testados foram: *Ridge*, *LASSO*, *ElasticNet*, *K-Nearest Neighbors* (KNN), Árvores de Decisão e *Support Vector Machines* (SVR). Os algoritmos do tipo *ensemble* testados foram: *AdaBoost* (AB), *Gradient Boosting* (GBM), *Random Forests* (RF) e *Extra Trees* (ET). Para avaliar melhorias no desempenho dos modelos construídos, foram testadas variações nos seus hiperparâmetros, estando em destaque na Tabela 2 os hiperparâmetros finais de cada experimento.

Os experimentos foram realizados no ambiente *Google Collaboratory*, na nuvem, com linguagem de programação *Python* e as bibliotecas *Pandas*, para a análise exploratória e manipulação de dados, e *Scikit-Learn*, para *Machine Learning*.

Foram desenvolvidos 14 experimentos alterando as seguintes variáveis: moeda; valor nominal e real; taxa de câmbio. Além disso, foram utilizadas diferentes partições de treino e teste. As decisões sobre os experimentos e os parâmetros utilizados foram norteadas em consequência dos resultados apresentados. Desta forma, o roteiro dos experimentos obedeceu a sequência relatada nos parágrafos subsequentes.

O primeiro experimento (E1) foi realizado utilizando todas as variáveis com seus valores nominais, com as moedas originais de cada base de dados, e incluindo a taxa de câmbio como variável, uma vez que os preços de óleo e gás são definidos em dólares. Além disso, o conjunto de treino e teste foi particionado em 80% e 20%, respectivamente. Foram testados todos os algoritmos de regressão mencionados, sendo o melhor MSE de teste obtido de 113,43 e o R2 de 0,9445 com o KNN.

Para identificar possibilidade de obter um melhor modelo, foi realizado o segundo experimento (E2) no qual todas as variáveis foram convertidas para Reais (R$), sendo mantida a taxa de câmbio como variável independente e mantida a partição de treino e teste de 80% e 20%. O melhor modelo testado permaneceu o KNN, com melhora no erro de teste, que passou para 62,1523, quase metade do erro anterior. Já o R2 teve um acréscimo pouco significativo, passando para 0,9696.

No terceiro experimento (E3), optou-se por manter os critérios do estudo (E2) exceto pela taxa de câmbio, que foi retirada da base de variáveis independentes. O menor MSE de teste encontrado foi de 106,1118, também com o KNN, retornando a um patamar próximo ao estudo (E1), enquanto o R2 melhorou sutilmente em relação ao (E2), passando para 0,9481.

Neste primeiro bloco de experimentos, observou-se que o melhor modelo construído foi com o algoritmo KNN do (E2), que utilizou a taxa de câmbio como variável independente, com as moedas das variáveis convertidas para Reais (R$).

No entanto, como a base de dados contém informações de períodos distantes, avaliou-se que seria interessante corrigir os preços e o valor do investimento em P,D&I, considerando a taxa de inflação histórica. A hipótese a ser testada é que valores atualizados são mais fidedignos e podem conduzir a um modelo mais aderente.

Para atualização destes valores, foi utilizado o Índice Nacional de Preços ao Consumidor Amplo (IPCA), que é o índice oficial de inflação do governo federal. A correção é calculada multiplicando o valor inicial por um fator, que é calculado dividindo o número-índice do mês final pelo número-índice do mês anterior ao mês inicial. Assim, após obtenção da tabela com a série histórica dos números-índices do IPCA no Sistema IBGE de Recuperação Automática (SIDRA), foram calculados os fatores de correção

mensais, que foram multiplicados pelos preços do Petróleo, preço do Gás e pelo valor do investimento em P,D&I.

Desta forma, as variáveis dos próximos experimentos passaram a ser:

1. Ano: ano de referência dos atributos.
2. Mês: mês de referência dos atributos.
3. Produção de Petróleo (b) (mil): volume de produção de petróleo no Brasil em barris.
4. VP PP ref - *Brent* (último) (R$): preço médio do petróleo do tipo *brent* em Reais (R$) por barril, corrigido pelo IPCA.
5. TC: Taxa de câmbio entre o Real (R$) e o Dólar (USD).
6. Produção nacional de gás natural (b): volume de produção de gás natural no Brasil em barris.
7. VP Preço Gás Natural (1 Mmbtu) (R$): preço médio do gás natural em Reais (R$) por Mmbtu, corrigido pelo IPCA.
8. VP Investimento P,D&I por mês (R$) (mil): valor em Reais (R$) destinado obrigatoriamente a investimentos em pesquisa e desenvolvimento no Brasil, corrigido pelo IPCA.

Assim, foi realizado um experimento (E4) sem a taxa de câmbio na lista de variáveis independentes e outro experimento (E5) incluindo a taxa de câmbio na lista de variáveis independentes, que resultaram respectivamente em MSE de teste de 154,4243 e 153,2624, e de R2 de 0,9174 e 0,9180, ambos novamente obtidos com o algoritmo KNN. Os resultados foram muito próximos entre si, mas inferiores aos estudos que não consideraram a correção monetária.

Após analisar todos os resultados obtidos nos experimentos anteriores, avaliou-se que, ao utilizar variáveis em Reais (R$), o modelo poderia estar carregando variações de preços que não são devido à produção de óleo e gás, mas sim devido a variações cambiais.

Para testar esta nova hipótese, foi realizado novo experimento (E6) convertendo todas as variáveis para Dólar (USD), mantendo a taxa de câmbio como variável independente, e todas as variáveis com seus valores nominais. Além disso, foi mantido o particionamento da base de treino e teste em 80% e 20%, respectivamente. O resultado obtido melhorou substancialmente, sendo o menor erro de teste encontrado de 9,9032 para o algoritmo KNN. O R2 do modelo se manteve no patamar dos estudos anteriores, sendo de 0,96068. Assim, optou-se por prosseguir nos próximos experimentos utilizando sempre as variáveis convertidas em Dólar.

Com o objetivo de obter um modelo ainda melhor, foi iniciado um novo experimento (E7), que manteve os parâmetros do experimento anterior (E6), mas retirou a taxa de câmbio da base de variáveis. O melhor modelo continuou sendo o KNN, com melhora nos resultados, já que o MSE de teste agora obtido foi de 9,2833 e o R2 de 0,9631, ainda em nível compatível com os anteriores.

A partir daí, uma nova análise para identificar se o modelo poderia ser melhorado ainda mais foi alterar a partição do conjunto de treino e teste. Ao invés de fixar uma proporção, foram realizados novos experimentos estabelecendo como base de teste

apenas o ano de 2020 e variando as bases de treino conforme a seguir, mantendo todos os demais parâmetros equivalentes ao experimento (E7) que obteve os melhores resultados desde o início, a saber:

- Experimento 8 (E8): base de treino com todos os dados entre 2000 e 2019;
- Experimento 9 (E9): base de treino com todos os dados entre 2015 e 2019;
- Experimento 10 (E10): base de treino com todos os dados entre 2018 e 2019.

O objetivo destes experimentos foi identificar se ao concentrar o período da base de treino o modelo poderia melhorar, considerando a hipótese de que o contexto da exploração e produção de óleo e gás mundial tende a ser mais homogêneo em períodos mais curtos. No entanto, os resultados obtidos foram piores, sendo o MSE de teste do (E8) de 41,1762, do (E9) de 40,8476 e do (E10) de 237,8758, para os algoritmos SVR, Lasso e LR, respectivamente.

Observando tais resultados, notou-se que o ano de 2020 poderia ter tido volume de produção e preços atípicos devido à pandemia de Covid-19. Para testar esta hipótese, foram realizadas novas partições nos experimentos subsequentes, ampliando a base de teste para o período de 2018 e 2019. As bases de treino foram estabelecidas da seguinte forma:

- Experimento 11 (E11): base de treino com todos os dados entre 2000 e 2017;
- Experimento 12 (E12): base de treino com todos os dados entre 2015 e 2017;
- Experimento 13 (E13): base de treino com todos os dados entre 2016 e 2017.

Os resultados também não foram satisfatórios, sendo o MSE de teste do (E11) de 56,7035, do (E12) de 77,6841 e do (E13) de 95,0420, para os algoritmos *ElasticNet*, LR e LR, respectivamente. A conclusão foi que possivelmente o menor volume de dados resultou em modelos piores. Assim, optou-se por avançar em nova análise, e não em desenvolver novos experimentos que utilizem períodos reduzidos como dados de treino.

Desta forma, optou-se por testar algoritmos *Ensemble* que, por serem meta-algoritmos que combinam vários modelos, podem gerar um melhor desempenho preditivo. No experimento seguinte (E14), foram mantidos os parâmetros do melhor modelo obtido até então (E7), ou seja, as variáveis foram mantidas em dólares, a taxa de câmbio não foi utilizada como variável independente, e os percentuais de particionamento da base de treino e teste foram mantidos em 80% e 20%, respectivamente. O resultado foi ainda superior que o (E7), pois o MSE de teste foi de 8,0734. O R2 novamente apresentou baixa variação comparando com os experimentos anteriores, pois foi de 0,9679.

No (E14), o *Random Forest* apresentou os melhores resultados. Como foi o melhor resultado obtido em comparação com todos os demais modelos, este foi o algoritmo selecionado. Em seguida, foi examinado o ajuste do número de estimadores para checar se variando este hiperparâmetro poderia ser obtida melhoria no modelo. Foram definidos valores entre 50 e 400, em incrementos de 50, para realização do experimento. No entanto, o modelo obteve melhor desempenho com o número padrão de estimadores (100). Assim, o modelo foi treinado com todo o conjunto de dados de treinamento e foram realizadas predições para o conjunto de dados de validação separado anteriormente para confirmar as descobertas.

4.3. Resultados

A Tabela 1 resume as características de cada experimento, incluindo algoritmo utilizado, valor, moeda, utilização da taxa de câmbio e particionamento das bases de treino e teste. A Tabela 2 apresenta de forma consolidada os resultados de cada experimento realizado, considerando o melhor modelo obtido e seus respectivos hiperparâmetros, erro de treino, erro de teste e R2.

É possível observar que os experimentos de 1 a 5 (E1 a E5), que utilizaram as variáveis em Reais, obtiveram erros de teste elevados. A partir do experimento 6 (E6), com utilização da moeda em Dólar, os resultados melhoraram, apresentando erros de teste reduzidos. A retirada da taxa de câmbio no experimento 7 (E7), acarretou em melhor resultado. As tentativas de fracionamento dos períodos das bases de treino e teste nos experimentos 8 a 13 (E8 a E13) pioraram os resultados. Desta forma, utilizando as conclusões dos experimentos 1 a 13 (E1 a E13), chegou-se no experimento 14 (E14), que foi desenvolvido utilizando como base o modelo 7, que foi o melhor obtido até então. Foram mantidos os parâmetros do experimento 7 (E7), aplicando *Ensembles* para avaliar a melhoria no modelo resultante. Assim, foi obtido o melhor resultado com o modelo *RandomForest*, com MSE de teste de 8,0734784 e R2 de 0,9679475.

Tabela 1. Características dos Experimentos

Experimento	Valor	Moeda	Taxa de Câmbio	Base de Treino	Base de Teste
1	Nominal	R$ e USD	SIM	80%	20%
2	Nominal	R$	SIM	80%	20%
3	Nominal	R$	NÃO	80%	20%
4	Presente	R$	NÃO	80%	20%
5	Presente	R$	SIM	80%	20%
6	Nominal	USD	SIM	80%	20%
7	Nominal	USD	NÃO	80%	20%
8	Nominal	USD	NÃO	2000 a 2019	2020
9	Nominal	USD	NÃO	2015 a 2019	2020
10	Nominal	USD	NÃO	2018 a 2019	2020
11	Nominal	USD	NÃO	2000 a 2017	2018 a 2019
12	Nominal	USD	NÃO	2013 a 2017	2018 a 2019
13	Nominal	USD	NÃO	2016 a 2017	2018 a 2019
14	Nominal	USD	NÃO	80%	20%

Tabela 2. Resultados dos Experimentos

Experimento	Melhor Modelo	Hiperparâmetros	Erro Treino	Erro Teste	R2
1	KNN	metric = manhattan, n_neighbors = 1	136,70337	113,43866	0,9445827
2	KNN	metric = euclidean, n_neighbors = 3	100,45565	62,152336	0,9696372
3	KNN	metric = euclidean, n_neighbors = 3	147,51742	106,1119	0,948162
4	KNN	metric = euclidean, n_neighbors = 9	283,05665	154,42433	0,9174243
5	KNN	metric = manhattan, n_neighbors = 3	223,96957	153,26245	0,9180456

Experimento	Melhor Modelo	Hiperparâmetros	Erro Treino	Erro Teste	R2
6	KNN	metric = manhattan, n_neighbors = 3	15,587722	9,903252	0,9606831
7	KNN	metric = manhattan, n_neighbors = 5	19,142283	9,283356	0,9631441
8	SVR	C = 1.0, kernel = rbf	25,186961	41,176288	-3,151437
9	Lasso	alpha = 1.0	28,612032	40,847687	-3,118307
10	LR	N/A	7,8424631	237,87583	-22,982893
11	ElasticNet	alpha = 1.0	37,179876	56,703529	-1,8583335
12	LR	N/A	23,41838	77,684134	-2,9159319
13	LR	N/A	19,025343	95,04208	-3,790918
14	Random Forest	n_estimators = 100	14,347836	8,0734784	0,9679475

5. Conclusão

O processo de inovação gera impactos no aumento da produtividade, agregando valor à economia do país. Por isso, o aumento da taxa de inovação é uma das principais condições para o Brasil acelerar o seu crescimento econômico e o seu desenvolvimento social, o que demanda investimentos e esforços deliberados [Leal & Figueiredo, 2021]. Identificar com antecedência recursos disponíveis para investimento em inovação é uma das formas de viabilizar melhor planejamento e aplicação destes recursos valiosos para o país.

Considerando que o investimento obrigatório em P,D&I pela indústria de óleo e gás é significativo, e que a forma de identificá-lo atualmente é através de fórmulas complexas envolvendo múltiplos parâmetros, o objetivo deste artigo foi a partir da utilização de *Machine Learning* obter um modelo que possa facilitar, simplificar e dar celeridade ao cálculo do valor nacional de recursos a serem aplicados em P,D&I, viabilizando melhor planejamento para direcionamento de recursos e políticas públicas para Universidades e Institutos de Ciência e Tecnologia, aumentando a eficácia dos gastos.

Ao final dos 14 experimentos realizados, observou-se que é possível avançar no caminho de utilizar Machine Learning para obter modelos simplificados e que alcancem resultados satisfatórios para estimar o investimento em P,D&I nacional obrigatório para as empresas de óleo e gás. Tais modelos viabilizam estimar o investimento de P,D&I de forma ágil e simplificada, sem utilização das fórmulas complexas necessárias atualmente. Assim, eles podem ser utilizados como ferramentas auxiliares para que a ANP melhore o planejamento de políticas públicas e direcionamento de recursos para Universidades, Empresas e Institutos de Ciência e Tecnologia.

Ao beneficiar a ANP, contribuindo para uma gestão mais eficiente e eficaz dos recursos, o benefício se estende à sociedade, que passa a contar com uma melhor organização e alocação de recursos de um setor cujo volume de receitas é significativo, e com uma aplicação que tem relação direta com a construção de um país mais dinâmico, competitivo e socialmente mais justo [Leal & Figueiredo, 2021].

É importante ressaltar que este estudo não buscou ser exaustivo, embora o modelo construído tenha apresentado resultados que indicam sua viabilidade de utilização. Assim, para trabalhos futuros, recomenda-se investigar o desenvolvimento de modelos adicionais, utilizando outros algoritmos de *Machine Learning* (como redes neurais) para

avaliar a possibilidade de alcançar resultados ainda melhores. Adicionalmente, é interessante buscar validar o modelo junto à agência ANP e às empresas do setor, captando sugestões de melhoria e adequando o modelo desenvolvido.

Referências

[Acesso à Informação, 2020] "Acesso à Informação - Institucional Agência Nacional do Petróleo, Gás Natural e Biocombustíveis" (2020) Agência Nacional do Petróleo, Gás Natural e Biocombustíveis. Disponível em https://www.gov.br/anp/pt-br/acesso-a-informacao/institucional (último acesso em outubro de 2021).

[Brent, 2021] "Dados Históricos do Brent" (2021) Investing.com. Disponível em: https://www.investing.com/commodities/brent-oil-historical-data (último acesso em outubro de 2021).

[Dados, 2021] "Dados Estatísticos - Produção Nacional de Petróleo e Gás Natural" (2021) Agência Nacional do Petróleo, Gás Natural e Biocombustíveis. Disponível em: https://www.gov.br/anp/pt-br/centrais-de-conteudo/dados-estatisticos (último acesso em outubro de 2021).

[Escovedo & Koshiyama, 2020] T. Escovedo & A. Koshiyama (2020) "Introdução a Data Science: Algoritmos de Machine Learning e Métodos de Análise". São Paulo: Casa do Código.

[Estatísticas, 2021] "Estatísticas econômico-financeiras (tipo de consulta: dólar)" (2021) Banco Central do Brasil.

[Firmo, 2019] "A Relevância do Petróleo & Gás para o Brasil" (2019) IBP. Disponível em: https://www.ibp.org.br/personalizado/uploads/2019/08/erelevancia-do-petroleo brasil.pdf (último acesso em outubro de 2021).

[Gás Natural, 2021] "Dados Históricos do Gás Natural" (2021) Disponível em: Investing.com. https://www.investing.com/commodities/natural-gas-historical-data (último acesso em outubro de 2021).

[IPCA, 2021] "IPCA - Número Índice," Instituto Brasileiro de Geografia e Estatística" (2021) Disponível em: https://sidra.ibge.gov.br/tabela/1737 (último acesso em outubro de 2021).

[Leal & Figueiredo, 2021] C. I. S. Leal & P. N. Figueiredo (2021) "Inovação Tecnológica no Brasil: Desafios e Insumos para Políticas Públicas", Revista de Administração Pública, vol. 55, no. 3, doi: 10.1590/0034-761220200583.

[Mahesh, 2020] B. Mahesh (2020) "Machine Learning Algorithms - A Review", International Journal of Science and Research (IJSR), vol. 9, no. 1, pp. 381–386, doi: 10.21275/ART20203995.

[Manual, 2017] "Manual de Procedimentos Cálculos, distribuição e Auditoria da Participação Especial" (2017) Agência Nacional do Petróleo, Gás Natural e Biocombustíveis, vol. 5. Brasil, pp. 1–25.

[Pesquisa, 2015] "Pesquisa Industrial Anual" (2015) Instituto Brasileiro de Geografia e Estatística. Instituto Brasileiro de Geografia e Estatística, Rio de Janeiro.

[Polybio et al., 2012] L. Polybio, B. Teixeira e M. N. Sales (2012) "Necessidade de P&D em Biocombustíveis: o Papel da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP)", doi:10.7447/dc.2012.007.

[Preços, 2020] "Memória de Cálculo - Preços de Referência do Petróleo" (2020) Agência Nacional do Petróleo, Gás Natural e Biocombustíveis. Brasil, pp. 1–17. Disponível em: https://www.gov.br/anp/pt-br/assuntos/royalties-e-outras-participacoes/arq-royalties/prp/mc/2020/memoria-calculo-fev-2020.pdf (último acesso em outubro de 2021).

[Recursos, 2021] "Recursos Financeiros das Cláusulas de Investimentos em PD&I" (2021) Agência Nacional do Petróleo, Gás Natural e Biocombustíveis. Disponível em: https://www.gov.br/anp/pt-br/assuntos/pesquisa-desenvolvimento-e-inovacao/investimentos-em-pd-i/recursos-financeiros-das-clausulas-de-investimentos-em-pd-i (último acesso em outubro de 2021).

[Regulamento, 2015] Regulamento Técnico ANP (2015) Brasil, pp. 1–49.

[Winkler, 1989] R. L. Winkler (1989) "Combining forecasts: A philosophical basis and some current issues", International Journal of Forecasting, vol. 5, no. 4, doi: 10.1016/0169-2070(89)90018-6.