UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
CAMPUS MACAÉ
CURSO DE BACHARELADO EM ENGENHARIA MECÂNICA

ANDRÉ BRANCO DE MENEZES RAFAEL DA SILVA

PREVISÃO DE TAXA DE PERFURAÇÃO EM POÇOS DE PETRÓLEO OFFSHORE UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

MACAÉ
2021

ANDRÉ BRANCO DE MENEZES RAFAEL DA SILVA

PREVISÃO DE TAXA DE PERFURAÇÃO EM POÇOS DE PETRÓLEO OFFSHORE UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

Trabalho de conclusão de curso de graduação apresentado ao Curso de Engenharia Mecânica do campus Macaé da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Engenharia Mecânica.

Orientadora: Profa. Janaína Sant'Anna Gomide Gomes

MACAÉ

2021

CIP - Catalogação na Publicação

S586p

Silva, André Branco de Menezes Rafael da
Previsão de taxa de perfuração em poços de petróleo offshore utilizando algoritmos de aprendizado de máquina / André Branco de Menezes Rafael da Silva. - Rio de Janeiro, 2021.
74 f.

Orientadora: Janaína Sant'Anna Gomide Gomes.
Trabalho de conclusão de curso (graduação) - Universidade Federal do Rio de Janeiro, Campus Macaé Professor Aloísio Teixeira, Bacharel em Engenharia Mecânica, 2021.

1. Aprendizado de Máquina. 2. Perfuração Offshore. 3. Engenharia de Petróleo. I. Gomes, Janaína Sant'Anna Gomide, orient. II. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

ANDRÉ BRANCO DE MENEZES RAFAEL DA SILVA

PREVISÃO DE TAXA DE PERFURAÇÃO EM POÇOS DE PETRÓLEO OFFSHORE UTILIZANDO ALGORITMOS DE APRENDIZADO DE MÁQUINA

Trabalho de conclusão de curso de graduação apresentado ao Curso de Engenharia Mecânica do campus Macaé da Universidade Federal do Rio de Janeiro como parte dos requisitos para obtenção do grau de Bacharel em Engenharia Mecânica.

Aprovado em ___ de _____ de _____

BANCA EXAMINADORA:

Janaína Sant'Anna Gomide Gomes
Doutora (Universidade Federal do Rio de Janeiro - Campus Macaé)

Lucas Lisbôa Vignoli
Doutor (Universidade Federal do Rio de Janeiro - Campus Macaé)

Raquel Jahara Lobosco
Doutora (Universidade Federal do Rio de Janeiro - Campus Macaé)

Leila Weitzel Coelho da Silva
Doutora (Universidade Federal Fluminense)

AGRADECIMENTOS

À minha professora orientadora, Janaína Gomide, pela orientação, auxílio e ensino feitos e pela confiança depositada.

À Ocyan, pelo incentivo e pela concessão de dados utilizados neste trabalho.

Aos meus colegas de trabalho, Bruno Chagas e Rodrigo Chamusca, que auxiliaram com opiniões e esclarecendo dúvidas.

Por fim, agradeço a meus amigos Adson Braga, Julian Santos e Gabriel Sant'Clair, pelo apoio e companhia durante o período de realização do trabalho.

RESUMO

A taxa de perfuração (ROP) de um poço de petróleo é uma métrica muito importante de se controlar, por influenciar diretamente na produtividade do poço, no desgaste da broca e na segurança do poço e operação. Este trabalho avalia a utilização de metodologias de Aprendizado de Máquina para previsão da ROP em poços de petróleo *offshore* como alternativa para modelos tradicionalmente utilizados pela indústria. Após uma revisão geral dos principais conceitos de engenharia de petróleo e aprendizado de máquina relevantes para o trabalho, apresentou-se um conjunto de trabalhos relacionados, que também aplicaram metodologias de aprendizado de máquina para a previsão de ROP, muitas vezes comparando o desempenho com os modelos tradicionais. Foram disponibilizados para realização dos experimentos numéricos deste trabalho os valores de profundidade do poço, profundidade da broca, peso sobre broca, torque do *top drive*, velocidade de rotação do *top drive*, vazão de entrada de fluido de perfuração, vazão de saída de fluido de perfuração, pressão de *standpipe* e inclinação do poço para porções de 4 poços *offshore* dos campos de Búzios e Sépia, no pré-sal da Bacia de Santos. Após a coleta e preparo dos dados, realizou-se diversos experimentos numéricos para cada poço, avaliando o desempenho de três modelos de aprendizado de máquina: regressão linear por gradiente descendente, árvore de decisão e floresta aleatória. Variou-se os hiper-parâmetros de número de amostras mínimo por folha na árvore de decisão, e de profundidade máxima e número de estimadores na floresta aleatória. A avaliação dos modelos foi feita utilizando as métricas de erro médio absoluto normalizado e R^2. Os melhores modelos foram, então, utilizados para avaliar a importância das características utilizadas e a possibilidade de aproveitar os dados de um poço para ajudar no treinamento de um modelo utilizado para prever a ROP em outro poço, utilizando aprendizado por transferência. De forma geral, os modelos de floresta aleatória apresentaram resultados superiores para cada poço, com exceção de um poço em que um modelo de árvore de decisão apresentou resultados levemente melhores. Houve diferenças também nos melhores hiper-parâmetros para cada poço, onde para alguns poços o resultado foi melhor para combinações de hiper-parâmetros que favorecessem a generalização do modelo, enquanto para outros os hiper-parâmetros que permitiam melhor ajuste mostraram-se mais adequados. De modo geral, os resultados foram inferiores àqueles obtidos na literatura, mas parâmetros de perfuração comumente considerados como extremamente importantes para a definição da ROP não estavam disponíveis para o trabalho, reduzindo o desempenho dos modelos.

Palavras-chave: aprendizado de máquina. perfuração offshore. engenharia de petróleo.

ABSTRACT

The rate of perforation (ROP) of an oil well is a very important metric to control, as it affects well productivity, bit wear, and well and operational security. This work evaluates the use of Machine Learning methodologies for predicting ROP in offshore oil wells as an alternative for other models traditionally used by the industry. After a general revision of the petroleum engineering and machine learning concepts most relevant for this work, related works which also applied machine learning methodologies for ROP prediction, often comparing the performance to traditional models, were presented. In this work the features available for numeric experiments were well depth, bit depth, weight on bit, top drive torque, top drive rotary speed, drilling fluid inlet flow rate, drilling fluid outlet flow rate, standpipe pressure, and well inclination, all available for portions of 4 offshore wells on the Búzios and Sépia fields, in Santos Basin's pre-salt. After collecting and preparing the data, many numeric experiments on each well were made, evaluating the performance of three machine learning models: gradient descent linear regression, decision tree, and random forest. For the decision tree, the minimum number of samples per leaf hyperparameter was varied, while the maximum depth and number of estimators hyperparameters were varied for the random forest. The models were evaluated using the normalized mean absolute error and the R^2. The best models were, then, used to evaluate the feature importance and the possibility of using the data of one well to improve the training of a model to predict the ROP in another well, using transfer learning. In general, random forest models performed the best for each well, except for one well, in which a decision tree model performed slightly better. The best hyperparameters were also different for each well; for some wells, the best result occurred for hyperparameter combination that favored model generalization, while for others hyperparameters that favored a better model fit performed better. In general, results were inferior to those obtained in the literature, but drilling parameters widely considered as extremely important for determining the ROP were not available for this work, reducing the models' performance.

Keywords: machine learning. offshore drilling. oil and gas engineering.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANP	Agência Nacional do Petróleo, Gás Natural e Biocombustíveis
BOP	*Blowout preventer*
CART	Árvore de Classificação e Regressão
MAE	Erro médio absoluto
MSE	Erro médio quadrático
MWD	*Measurement while drilling*
PV	Viscosidade plástica
R^2	Coeficiente de explicação
RMSE	Raiz do erro médio quadrático
ROP	Taxa de perfuração
RPM	Rotação por minuto
SPP	Pressão de *standpipe*
UCS	Resistência compressiva uniaxial
WOB	Peso sobre broca

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO DO TEMA

O petróleo é uma substância química amplamente utilizada pela sociedade moderna, com usos que vão de uso como combustível veicular, à produção de lubrificantes e asfalto, à produção de plásticos, e por isso é uma substância de enorme interesse econômico. Antes de o petróleo poder ser utilizado, deve-se primeiramente encontrar reservatórios economicamente viáveis, etapa conhecida como prospecção. Em seguida perfurar um poço para alcançar o reservatório, na etapa conhecida como perfuração, para que seja possível extraí-lo deste reservatório, na etapa da produção. Por fim, o petróleo é processado em diversos produtos e subprodutos, que podem ser utilizados como estão ou sofrerem processamentos adicionais para a produção de novos produtos. (THOMAS, 2001)

O petróleo é uma substância utilizada desde a antiguidade, quando era usado, por exemplo, na produção de tijolos, pavimentação de estradas e impermeabilização de potes, mas foi somente com a Segunda Revolução Industrial que descobriu-se o uso do petróleo como fonte de energia, dando destaque a ele e à sua extração. Foi assim que deu-se início à perfuração de poços de petróleo para fins comerciais, inicialmente por métodos percussivos, e posteriormente por métodos rotativos, alcançando profundidades cada vez maiores. De início os poços eram todos terrestres, ou *onshore*, mas logo buscou-se expandir a produção com reservatórios marítimos, ou *offshore*. (THOMAS, 2001)

Enquanto um grande número de descobertas de jazidas ocorreu ao redor do mundo a partir de 1860 após a perfuração do primeiro poço comercial, o Brasil permaneceu mais de 70 anos sem encontrar petróleo em quantidades significativas em seu território, até que foi encontrada uma acumulação de petróleo na Bahia em 1939, seguida de novas descobertas de campos com potencial comercial. Após estas descobertas, e outras que ocorreram na década seguinte, que confirmaram a existência de potencial petrolífero no Brasil, foi criada a PETROBRAS em 1953, com o objetivo de reduzir a dependência brasileira do petróleo importado. (MORAIS, 2013)

As primeiras descobertas de campos de petróleo *offshore* brasileiros deram-se entre 1968-1973, no Nordeste brasileiro, e a partir de 1974, na bacia de Campos, na qual a extração avançou ao longo do tempo de águas rasas (lâmina d'água de até 300 m) para águas profundas (300 m a 1500 m) e ultra-profundas (acima de 1500 m). Os avanços exploratórios e tecnológicos das décadas seguintes eventualmente levaram à descoberta de reservatórios gigantes e supergigantes de Pré-sal nas Bacias de Santos e de Campos a partir de 2006, dando início à Era do Pré-sal. Em 2010, o primeiro campo de pré-sal recebeu a Declaração de Comercialidade da Agência Nacional do Petróleo (ANP), confirmando presença de volumes comerciais na área, e foi seguida por Declarações de Comercialidade

para diversos outras regiões de pré-sal. (MORAIS, 2013)

Para que seja possível extrair petróleo de reservatórios, é necessário primeiramente realizar um programa de prospecção, que visa localizar áreas de uma bacia sedimentar em que possa ocorrer o acúmulo de petróleo e, posteriormente, determinar quais destas áreas têm maior probabilidade de encontrar petróleo em quantidades apreciáveis. (THOMAS, 2001)

Após o processo de prospecção, já definida a posição e a maneira como será perfurado o poço, é necessário perfurar o poço até a profundidade do reservatório de petróleo. A perfuração é realizada em seções progressivas, em que perfura-se até uma profundidade determinada, reveste-se o poço com tubos de aço e cimenta-se o espaço entre o revestimento e a parede do poço para aumentar a estabilidade do poço, e em seguida retoma-se a perfuração, com uma broca de diâmetro menor. Este processo é repetido até que a profundidade final seja alcançada. (THOMAS, 2001)

Concluída a perfuração, o poço precisa ser preparado para uma produção segura ao longo de sua vida, em um conjunto de processos chamado de completação. Após a completação, pode-se partir para a produção, em que o petróleo é extraído do reservatório, uma etapa tão complexa quanto as anteriores. (THOMAS, 2001)

A taxa de penetração, ou *rate of penetration* (ROP), representa a taxa de variação da profundidade do poço em relação ao tempo, calculada durante atividades de perfuração. O valor da ROP influencia diretamente no desgaste das brocas, na segurança do poço e da operação, bem como na eficiência geral do processo, e por isso é uma variável que deve ser controlada. Para controlar a ROP de forma efetiva, é necessário ser capaz de estimar seu valor para diferentes valores de entrada, e por isso o interesse em modelar seu comportamento (GANDELMAN, 2012).

Modelos experimentais para a ROP existem e são largamente utilizados pela industria, mas sua acurácia e capacidade de generalização para condições de perfuração diferentes são limitadas (GANDELMAN, 2012). Trabalhos acadêmicos recentes utilizam técnicas de aprendizado de máquina para obter modelos e reportam resultados significativamente melhores que os modelos tradicionais. (BARBOSA et al., 2019)

1.2 OBJETIVOS

Este trabalho tem como objetivo avaliar o uso de técnicas de aprendizado de máquina para a previsão da taxa de perfuração (ROP) durante a perfuração de poços de petróleo *offshore* de pré-sal da Bacia de Santos.

1.2.1 Objetivos Específicos

- Aplicar modelos de árvore de decisão e floresta aleatória na previsão da ROP durante a perfuração de poços de petróleo no pré-sal da Bacia de Santos, nos campos de

Búzios e de Sépia.

- Avaliar a performance de diferentes combinações de hiper-parâmetros para os modelos em cada poço avaliado.
- Avaliar a performance do uso de *transfer learning* para aproveitar conhecimentos adquiridos em um poço na previsão em outro poço.

1.3 MOTIVAÇÃO E JUSTIFICATIVA

A ROP é um parâmetro que comumente busca-se maximizar, o que permitiria perfurar uma mesma profundidade em menor tempo, aumentando a produtividade. Na realidade, entretanto, deve-se considerar outras variáveis, como o desgaste das brocas, amplificado por ROPs mais altas, especialmente quando perfura-se formações duras, e a limpeza poço (GANDELMAN, 2012; HEGDE et al., 2017). Desgastes exagerados na broca levam a trocas de broca mais frequentes, aumentando o tempo de manobra com trocas de broca e adicionando custos (HEGDE et al., 2017). A limpeza de poço também deve ser considerada, pois concentrações altas de sólidos no anular podem aumentar excessivamente a pressão no poço, que deve ser controlada de modo a evitar fratura de parede de poço e perda de fluido para formação. Seja para se maximizar ou se limitar, o controle do valor da ROP é uma das tarefas mais importantes durante a perfuração do poço, e para isso é importante conseguir estimar os valores da ROP para determinadas condições de perfuração (GANDELMAN, 2012).

Do ponto de vista acadêmico, este trabalho envolve fundamentos de aprendizado de máquina e sua aplicação, além de fundamentos em engenharia de petróleo, mais especificamente referentes ao processo de perfuração de poços *offshore*.

1.4 ESTRUTURA DO TRABALHO

O segundo capítulo desse trabalho refere-se à revisão de literatura, abordando de forma resumida os principais conceitos de aprendizado de máquina e engenharia de petróleo utilizados neste trabalho. Posteriormente, no mesmo capítulo, revisou-se trabalhos relacionados de aplicação de aprendizado de máquina na previsão da ROP.

O terceiro capítulo apresenta a metodologia utilizada no trabalho, delimitando as ferramentas e processos utilizados, e os experimentos numéricos que foram realizados.

O quarto capítulo apresenta os resultados obtidos e uma discussão analisando-os de forma comparativa.

O quinto capítulo contém a conclusão do trabalho, onde fecha-se as ideias do estudo e pesquisa realizados.

2 REVISÃO DE LITERATURA

2.1 PERFURAÇÃO DE POÇOS DE PETRÓLEO

Esta sessão foi, majoritariamente, baseada em (THOMAS, 2001).

O processo de perfuração de poços de petróleo antecede a extração do petróleo dos reservatórios, e é um processo complexo que, conforme Gabbay (2015), pode ser responsável por 40% a 60% dos custos totais de desenvolvimento dos campos de petróleo, e a perfuração de um único poço de petróleo *offshore* pode levar múltiplos meses.

A perfuração de um poço de petróleo é realizada através de uma sonda de perfuração, que no caso de poços *offshore* de águas profundas ou ultraprofundas são plataformas semissubmersíveis ou navios-sonda (MORAIS, 2013). As sondas possuem vários sistemas: conjuntos de equipamentos responsáveis pela correta operação das sondas para a perfuração, que podem estar direta ou indiretamente relacionados à atividade de perfuração (THOMAS, 2001).

2.1.1 Equipamentos da sonda de perfuração

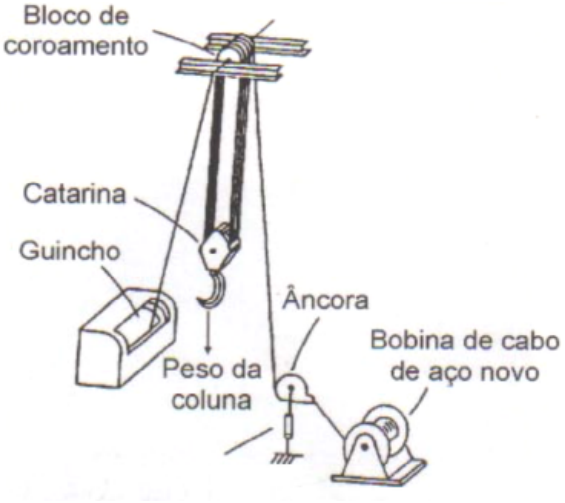
Os principais sistemas de uma sonda são os sistemas de: sustentação de cargas, de movimentação de cargas, de rotação, de circulação, de segurança de poço, de subsuperfície, de posicionamento dinâmico, de geração e transmissão de energia, e de monitoração. Dentre estes, os seis primeiros têm influência direta no processo da perfuração, enquanto os três últimos afetam-no indiretamente, mas são indispensáveis.

O sistema de sustentação de cargas é formado pela torre de perfuração, ou *derrick*, responsável por sustentar todas as cargas içadas durante a perfuração, e pela subestrutura, que apoia a torre e recebe suas cargas.

O sistema de movimentação de cargas realiza o içamento e movimento da coluna de perfuração, do revestimento, além de outras cargas gerais. O sistema é formado pelo guincho (que gera a energia mecânica necessária para o içamento das cargas), o cabo de perfuração (que transmite a energia mecânica entre os elementos), o bloco de coroamento no topo do mastro (formado por um conjunto de polias que sustenta as cargas do cabo de perfuração), a catarina (conjunto de polias suspenso pelo cabo de perfuração), o gancho (conectado à parte inferior da catarina e ao objeto a ser içado, sendo também responsável por amortecer os impactos causados pela movimentação de carga) e o elevador (equipamento em formato de anel utilizado para envolver os elementos tubulares da perfuração, permitindo seu movimento). Uma representação esquemática do sistema de movimentação de cargas pode ser vista na Figura 1.

O sistema de rotação é responsável por transmitir potência rotativa para a coluna de perfuração para fins de perfuração, bem como para a conexão de novos tubos à coluna de

Figura 1 – Sistema de Movimentação de Cargas



Fonte: (THOMAS, 2001)

perfuração. A rotação em uma sonda *offshore* é gerada pelos motores do *top drive*, um equipamento que é conectado ao gancho e que acopla-se aos tubos da coluna de perfuração para transmitir potência a ela, tanto para a operação de perfuração quanto para a conexão dos tubos para montagem da coluna de perfuração. No topo do top drive, há a cabeça de injeção, ou *swivel*, pelo qual o fluido de perfuração é injetado na coluna de perfuração. Uma fotografia de um *top drive* é mostrada na Figura 2.

Figura 2 – Top drive

Fonte: Concedida pela Ocyan

Adicionalmente ao *top drive*, há também o motor de fundo, posicionado no fundo da coluna de perfuração, sobre a broca, permitindo a rotação desta sem rotacionar toda a coluna. O motor de fundo é um motor hidráulico de tipo turbina ou de tipo deslocamento positivo movido pelo fluxo de fluido de perfuração, e é largamente utilizado para perfuração de poços direcionais e horizontais.

O sistema de circulação é formado pelo conjunto de equipamentos responsáveis pelo armazenamento, circulação e tratamento do fluido (lama) de perfuração: os tanques de lama, as bombas de lama, o tubo bengala (*standpipe*), o *swivel* e o subsistema de tratamento. O fluido de perfuração é muito importante para o processo de perfuração, tendo funções de limpar o poço do cascalho gerado durante a perfuração, levando-o à superfície, manter a estabilidade da parede do poço por meio de pressão hidrostática e resfriar e lubrificar a broca durante a perfuração.

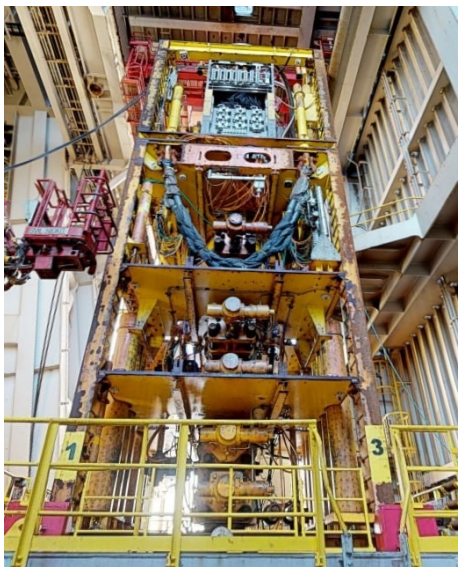
As bombas de lama são bombas de deslocamento positivo (ou volumétricas) alternativas de pistão, comumente de tipo triplex, *i.e.*, com 3 pistões. Elas retiram o fluido dos tanques de lama, que atravessa o tubo bengala (*standpipe*) em direção ao *swivel*, pelo qual é injetado na coluna de perfuração. O fluido atravessa a coluna até sair por orifícios na broca chamados de jatos da broca. O fluido, então, retorna à superfície pelo espaço anular entre a coluna de perfuração e as paredes do poço ou revestimento. Na superfície, o fluido passa pelo subsistema de tratamento, para eliminar os sólidos e gases que se incorporam a ele durante o processo e, quando necessário, adicionar substâncias para ajustar suas propriedades. Uma bomba de lama triplex é representada na imagem 3.

Figura 3 – Bomba de Lama Triplex

Fonte: Concedida pela Ocyan

O sistema de segurança do poço inclui, principalmente, o *blowout preventer*, ou BOP, e a cabeça de poço, e tem função de manter o controle sobre o poço. O equipamento mais importante desse sistema, o BOP, é um conjunto de válvulas que conecta-se ao topo da cabeça de poço e permite o fechamento do poço no caso de ocorrência de um *kick*, fluxo indesejável do fluido contido no poço, de modo a controlá-lo. Caso um *kick* não seja controlado com sucesso, ele pode levar a um *blowout*, acidente que pode causar enormes danos à sonda, ao ambiente e às pessoas. A Figura 4 mostra um BOP em superfície na sonda.

Figura 4 – *Blowout Preventer*



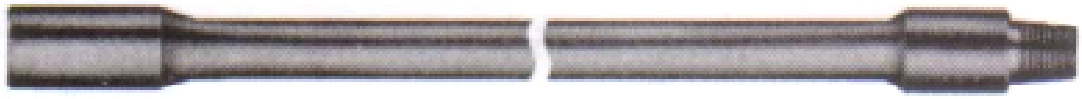
Fonte: Concedida pela Ocyan

O sistema de subsuperfície é composto pela broca, o motor de fundo e a coluna de perfuração, e é o sistema que atua de maneira mais direta na perfuração do poço. A coluna de perfuração conecta a sonda, na superfície, à broca, fornecendo peso e rotação à broca e transportando o fluido de perfuração para dentro do poço, enquanto a broca realiza a ruptura e desagregação das rochas, necessários para perfuração do poço.

A coluna de perfuração é formada, principalmente, por comandos, mais conhecidos como *drill collars*, e tubos de perfuração (*drill pipes*) normais e pesados. Todos estes são elementos tubulares, mas diferem-se principalmente por seu peso. Os *drill collars* são os com maior peso linear, e têm função de fornecer peso sobre a broca e rigidez para a coluna. Os *drill pipes* pesados são tubos de peso e rigidez intermediários aos *drill collars* e *drill pipes*, permitindo uma transição mais gradual entre eles. Os *drill pipes* são os tubos mais leves e formam o maior comprimento da coluna de perfuração. Uma imagem de um *drill pipe* pode ser encontrada na Figura 5. Os elementos tubulares são conectados entre si por conexões cônicas com roscas soldadas, conhecidas por *tool joints*.

Além destes elementos tubulares, há também os acessórios, como por exemplo os es-

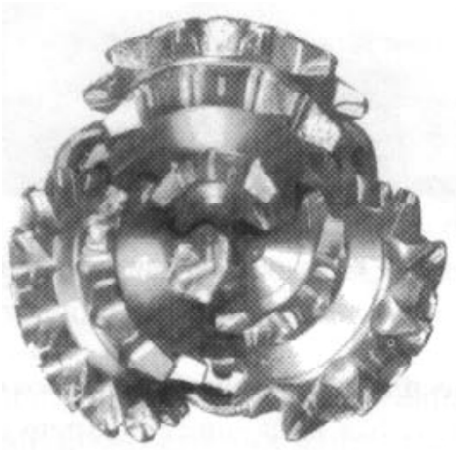
Figura 5 – *Drill Pipe*



Fonte: (THOMAS, 2001)

tabilizadores, que aumentam a rigidez da coluna e apoiam-se na parede do poço para manter a coluna centrada no poço, os amortecedores de vibração, que absorvem vibrações verticais da coluna, especialmente durante perfuração de formações mais duras, os alargadores, que permitem aumentar o diâmetro de trechos já perfurados, e o percussor de perfuração (*drilling jar*), que quando ativado aplica na coluna de perfuração um movimento brusco para cima ou para baixo, permitindo liberar a coluna de perfuração no evento de uma prisão de coluna.

Figura 6 – Broca Tricônica de dentes de aço



Fonte: (THOMAS, 2001)

As brocas podem ser de vários tipos diferentes, e precisam ser selecionadas de modo a ter melhor desempenho para a seção perfurada. Existem brocas sem partes móveis, como a broca de diamantes naturais, utilizada para perfuração de formações rochosas extremamente duras e abrasivas, e a broca de diamantes sintéticos, usada para formações moles. Há, também, as brocas com partes móveis, usadas para formações moles e duras, com um a quatro cones rotativos que cortam a rocha por uma combinação de ações de raspagem, lascamento, esmagamento e erosão por impacto do jato de lama. As brocas com partes móveis mais comuns são as tricônicas, como a mostrada na Figura 6, e podem ser de dentes de aço ou de insertos, comumente de carbureto de tungstênio. As brocas de partes móveis podem ter formatos de dente ou inserto diferentes para perfuração de diferentes formações. Uma última variável importante para a broca é seu diâmetro, que

deve reduzir entre a perfuração de uma seção superior e uma inferior.

O sistema de posicionamento dinâmico é utilizado em sondas de perfuração *offshore* dos tipos plataforma semi-submersível e navio-sonda, utilizadas na perfuração de poços profundos e ultraprofundos, e têm o objetivo de manter a sonda em uma posição aproximadamente vertical sobre a cabeça de poço. Para isso, são utilizados sensores de posicionamento e propulsores, mantendo a posição da sonda de maneira dinâmica (MORAIS, 2013).

O sistema de geração e transmissão de energia é formado, principalmente, pelas fontes de energia, comumente motores a diesel, pelos barramentos elétricos e pelas linhas de transmissão. Ele é responsável por alimentar a energia para toda sonda, incluindo sistemas de perfuração e auxiliares.

Por fim, o sistema de monitoramento é formado pelos equipamentos e *softwares* utilizados para medir, monitorar e registrar dados sobre os diferentes equipamentos e sistemas utilizados na sonda.

Enquanto os sistemas de posicionamento dinâmico, de geração e transmissão de energia e de monitoramento não atuam diretamente no processo de perfuração, o processo não pode ser realizado sem o funcionamento correto deles.

2.1.2 Operações da atividade de perfuração

Durante a atividade de perfuração, são realizadas uma série de operações necessárias para o processo.

A principal operação é a perfuração em si, caracterizada pela aplicação de peso e rotação sobre a broca enquanto esta encontra-se no fundo do poço e circula-se fluido de perfuração pela coluna de perfuração, causando o aumento da profundidade do poço.

Outras operações que ocorrem com peso sobre broca, rotação e circulação de fluido são o alargamento, em que perfura-se o poço novamente com uma broca de diâmetro maior (operação também que pode ser feita durante a perfuração com o uso de um alargador), e o repasse, em que passa a broca novamente em um trecho que se estreitou por algum motivo. Em nenhuma dessas duas operações a profundidade do poço é aumentada, e durante o repasse o peso sobre a broca e a rotação mantém-se baixos para evitar desgaste.

Como o movimento vertical do *top drive* é limitado pelo quanto o sistema de movimentação de cargas é capaz de levantá-lo, sempre que ele alcança a altura mais baixa durante o movimento da coluna de perfuração, é necessário realizar a operação de conexão. Nela, o *top drive* desconecta-se da coluna, eleva-se e conecta-se a um novo tubo, que é adicionado à coluna de perfuração pelo torque gerado pelo *top drive*. Um total de três novos tubos pode ser adicionado de uma vez à coluna, e depois retoma-se a operação de perfuração. A operação de conexão sempre ocorre sem circulação de fluido.

A operação de manobra consiste nos movimentos de retirada e de descida da coluna de perfuração. Durante a descida, é necessário realizar diversas conexões até alcançar

a profundidade desejada, e durante a retirada são realizadas operações de desconexão, inversas à conexão, até que toda a coluna de perfuração retorne à superfície. A manobra é comumente utilizada para a troca de brocas, seja para colocar uma broca mais adequada para a perfuração da formação específica, ou para trocar uma broca que tenha se desgastado em excesso.

A operação de circulação consiste em circular o fluido de perfuração com a broca próxima da profundidade máxima, e é realizada para limpar o poço, removendo os cascalhos do fundo do poço e levando-os para a superfície.

Como citado anteriormente, os poços de petróleo são perfurados em fases, cuja quantidade depende das características das formações que serão perfuradas e da profundidade final pretendida. Após a perfuração de cada fase, é necessário realizar o revestimento e a cimentação da fase.

Durante o revestimento, são descidos tubos de aço que cobrem a parede do poço, servindo para vários objetivos, principalmente: prevenir desmoronamento da parede do poço, permitir uso de fluidos e pressões de bombeio compatíveis com as formações mais profundas, mantendo maior estabilidade, e impedir a migração de fluidos de e para a formação, entre outros.

Na cimentação, preenche-se o espaço anular entre a parede do poço e o revestimento com cimento, fixando a tubulação e evitando circulação de fluido nesta região.

Após concluir a perfuração de uma fase do poço, também realiza-se a perfilagem, em que mede-se propriedades da formação para sua caracterização.

2.1.3 Perfuração Direcional

Os poços podem ser classificados entre poços verticais, poços direcionais e poços horizontais, diferenciados pelo ângulo perfurado em relação à vertical. Os poços verticais não são, na realidade, estritamente verticais, podendo desviar naturalmente da vertical. Quando um poço vertical desvia acima de limites de inclinação determinados, a inclinação deve ser corrigida utilizando técnicas de perfuração direcional, que também é utilizada na perfuração dos poços direcionais e horizontais.

Na perfuração direcional, desvia-se intencional e controladamente a trajetória de um poço. Isto pode ser realizado retirando a coluna do poço e descendo-a novamente com um *bent sub*, que tem a função de orientar o fundo da coluna de perfuração para a direção em que o poço deve ser perfurado, e ativando o motor de fundo, que perfura sem a rotação do resto da coluna. Também é possível perfurar utilizando um motor de fundo *steerable*, com uma deflexão no próprio corpo, permitindo que a direção de perfuração ajustada com a rotação da coluna e ativação do motor de fundo, e posteriormente continuar perfurando em linha reta usando o *top drive*, sem precisar retirar a coluna do poço.

Para controlar a direção do poço, é necessário também medir e registrar a direção e inclinação do poço. Uma ferramente muito utilizada é o MWD (*measurement while*

drilling), que envia as informações de inclinação e direção usando pulsos de pressão através do fluido de perfuração. As medições de direção e inclinação são registradas para diferentes profundidades no *directional survey*, para consulta posterior.

2.1.4 Taxa de Perfuração

A taxa de perfuração é a velocidade com que a broca perfura as rochas em um poço de petróleo, e é um valor muito importante de se controlar. GANDELMAN (2012) discute, em sua dissertação, os parâmetros que mais influenciam ROP. São eles: o peso sobre broca (WOB), o diâmetro e tipo da broca, o diferencial de pressão poço-formação (*overbalance*), a rotação da broca (RPM), a vazão do fluido de perfuração, a profundidade vertical do poço, a litologia da formação, o tempo de trânsito (de onda sonora) e a composição da coluna de perfuração. Dentre esses, o WOB e o RPM são os parâmetros mais facilmente ajustáveis para controlar a ROP, enquanto os outros parâmetros são escolhidos baseados em outros requisitos do projeto do poço, como no caso da vazão e diferencial de pressão (que são escolhidos baseado na limpeza e estabilidade de poço), ou estão completamente fora do controle dos operadores, como no caso da litologia.

Há, na literatura, modelos tradicionais utilizados para a previsão de ROP: equações matemáticas que buscam modelar o comportamento físico das variáveis de perfuração. Entretanto, como afirma GANDELMAN (2012), a eficiência destes modelos é consideravelmente limitada e não consegue generalizar bem para regiões e condições diferentes daquelas para que foram desenvolvidos. Dois dos modelos mais difundidos na indústria do petróleo são os modelos de Maurer e de Bourgoyne e Young.

O modelo de Maurer (MAURER et al., 1962) é baseado no diâmetro da broca, na perfurabilidade (dificuldade em se perfurar) da rocha, no WOB e no RPM. O modelo em questão calcula a ROP com a equação em 2.1, onde d_b é o diâmetro da broca e a e b são constantes experimentais características do tipo de rocha.

$$\mathrm{ROP} = a(\mathrm{RPM}) \left(\frac{\mathrm{WOB}}{d_b} - b \right)^2 \tag{2.1}$$

O modelo de Bourgoyne e Young (JR; JR et al., 1974) foi criado com o intuito de modelar o comportamento mais complexo da ROP em relação aos parâmetros de perfuração, que os modelos simples de Young e de Maurer são incapazes de modelar. O modelo de Bourgoyne e Young, representado na equação 2.2, modela a relação entre a ROP e os parâmetros de perfuração como exponencial, e baseia-se em 8 parâmetros: 1) Perfurabilidade da Rocha; 2) Profundidade; 3) Compactação; 4) *Overbalance*; 5) Peso sobre broca; 6) Velocidade de rotação da broca; 7) Desgaste da broca e; 8) Vazão.

$$\mathrm{ROP} = \exp\left(a_1 + \sum_{j=2}^{8} a_j x_j \right) \tag{2.2}$$

Barbosa et al. (2019) incluiu uma revisão rápida destes e outros métodos tradicionais de previsão da ROP. Os modelos citados incluem também uma versão adaptada do modelo de Bourgoyne, dois modelos baseados principalmente na interação broca-rocha, e um modelo recente baseado em análise de regressão (AL-ABDULJABBAR et al., 2019), mostrado na equação 2.3. Nesta equação T é o torque, SPP é a pressão de *standpipe*, Q é a vazão de fluido, ρ é a densidade do fluido de perfuração, PV é a viscosidade plástica do fluido e UCS é a resistência compressiva uniaxial. e e f são constantes obtidas por regressão não-linear.

$$\mathrm{ROP} = 16,96\frac{\mathrm{WOB}^{e}\times\mathrm{RPM}\times\mathrm{T}\times\mathrm{SPP}\times\mathrm{Q}}{d_b^2\times\rho\times\mathrm{PV}\times\mathrm{UCS}^{f}} \tag{2.3}$$

2.2 APRENDIZADO DE MÁQUINA

A criação do termo aprendizado de máquina (*machine learning*) é atribuída a Arthur Samuel (SAMUEL, 1959), onde ele utiliza o termo para referir-se ao processo de programar computadores a aprenderem por experiência como resolver tarefas, no lugar de programá-los com a metodologia específica e minuciosa de como resolvê-las, passo a passo. Uma definição, mais técnica, dada por Tom Mitchell (MITCHELL et al., 1997), diz que um programa de computador aprende pela experiência E, referente a uma classe de tarefas T, e uma métrica de desempenho P, se sua performance em tarefas T, medida pela métrica P, melhorar com a experiência E.

Diferentemente da programação tradicional, em que o programador define em um algoritmo a lógica que deve ser seguida para realizar a tarefa T, em uma abordagem de aprendizado de máquina, o algoritmo irá aprender por si só a lógica necessária para realizar a tarefa T. Assim, esta abordagem é comumente utilizada para: (i) problemas em que as soluções exigem a programação de um grande número de regras, (ii) problemas complexos em que soluções tradicionais não alcançam resultados adequados ou (iii) situações em que a resposta se altera com o tempo: um algoritmo de aprendizado de máquina pode se adaptar aos novos dados (GÉRON, 2019).

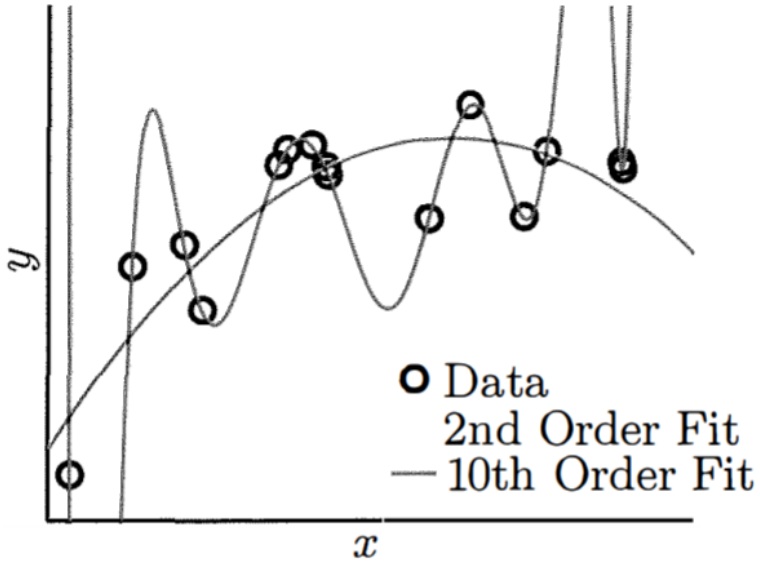
Ao aprender um conceito alvo, o algoritmo de aprendizado é dado um conjunto de exemplos de treino, um conjunto X de instâncias x, junto de seu resultado alvo y(x). O algoritmo, então, busca estimar a função y(x) usando funções de um conjunto considerado, chamado conjunto hipótese, assim aprendendo (MITCHELL et al., 1997).

Como durante o treino o algoritmo de aprendizado ajusta o modelo de forma que o erro cometido pela previsão seja mínimo para o conjunto de treino, este processo faz com que o modelo seja otimizado para realizar previsões no conjunto de testes. Entretanto, isto não significa que o modelo é capaz de generalizar suas previsões, *i.e.*, prever com sucesso os valores ou classes de exemplos que não estejam no conjunto de treino. Para poder avaliar o desempenho de um modelo de forma não enviesada, é necessário medir esse

desempenho com um conjunto de dados diferente, o conjunto de teste (ABU-MOSTAFA; MAGDON-ISMAIL; LIN, 2012).

Um problema comum durante a aplicação de algoritmos de aprendizado de máquina é o sobre-ajuste (*overfitting*) do modelo aos dados, ilustrado na Figura 7. O sobre-ajuste é o fenômeno que ocorre quando o modelo ajusta-se tão bem ao conjunto de dados de treino que perde sua capacidade de generalização. Isto ocorre com modelos que são mais complexos que o necessário. Estes modelos ajustam-se a aleatoriedades ou ruídos presentes na amostra de dados de treino, e portanto não representam corretamente o comportamento geral dos dados. Para evitar a ocorrência de sobre-ajuste, duas possíveis soluções são: utilizar um modelo menos complexo para as previsões, ou adicionar restrições ao modelo, um processo chamado de regularização (ABU-MOSTAFA; MAGDON-ISMAIL; LIN, 2012).

Figura 7 – Sobre-ajuste ilustrado com polinômios de 2ª e 10ª ordem



Fonte: (ABU-MOSTAFA; MAGDON-ISMAIL; LIN, 2012)

2.2.1 Tipos de Aprendizado

As abordagens de Aprendizado de Máquina são classicamente classificadas em aprendizado supervisionado e não-supervisionado. Se o conjunto de treino utilizado contiver os rótulos, *i.e.*, a resposta correta esperada, para cada dado de entrada, então trata-se de um caso de aprendizado supervisionado, e o algoritmo buscará gerar um modelo capaz de prever corretamente a resposta correta esperada. No caso do aprendizado não supervisionado, o conjunto de dados de treino não contém informações de resposta esperada pelo modelo; neste caso, o algoritmo buscará algum tipo de padrão no conjunto

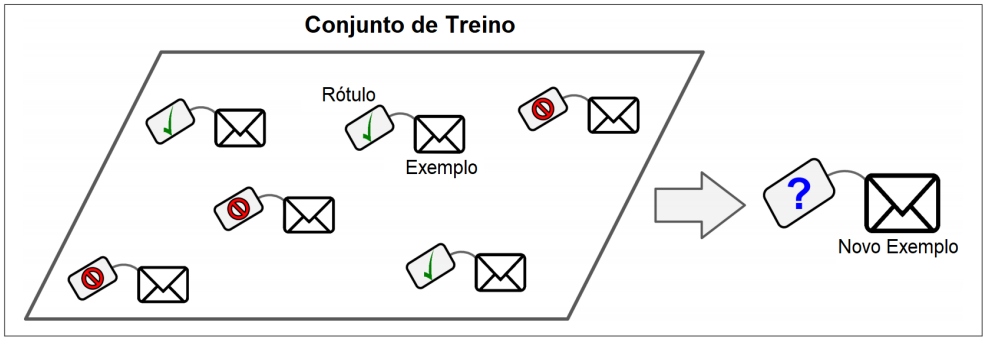
de dados, buscando agrupar os exemplos, reduzir suas dimensões ou encontrar anomalias (ABU-MOSTAFA; MAGDON-ISMAIL; LIN, 2012).

Há também o aprendizado semi-supervisionado, em que apenas parte dos exemplos do conjunto de dados utilizado para treino contém uma resposta esperada, e o algoritmo extrapola essa informação para os exemplos não-rotulados baseado na similaridade entre os dados, e com isso busca um modelo capaz de prever uma resposta esperada (GÉRON, 2019).

Um outro tipo de aprendizado é o aprendizado por reforço, que lida com a situação em que um agente está em um ambiente que ele pode observar e com que consegue interagir a partir de ações, e busca aprender como otimizar suas ações. Os objetivos do agente são representados por uma função de recompensa, que atribui um valor numérico a cada ação realizada pelo agente em cada estado distinto. Baseado em suas ações no ambiente e às recompensas obtidas por estas ações, o agente aprende uma estratégia de ação de forma a maximizar a recompensa (MITCHELL et al., 1997).

O aprendizado de máquina pode ser, adicionalmente, classificado em aprendizado incremental ou em pacotes. No aprendizado em pacote (*batch learning*), o algoritmo precisa treinar com todos os dados de treino de uma vez; assim, para que um modelo utilize novos dados, o treino deve ser realizado novamente para todos os dados presentes, o que pode ser custoso tanto em tempo quanto em processamente. Para o aprendizado incremental (*online learning*), o algoritmo consegue ajustar o modelo gerado utilizados novos dados introduzidos, sem precisar treinar novamente com os dados anteriores, o que é especialmente vantajoso para modelar eventos dinâmicos (GÉRON, 2019).

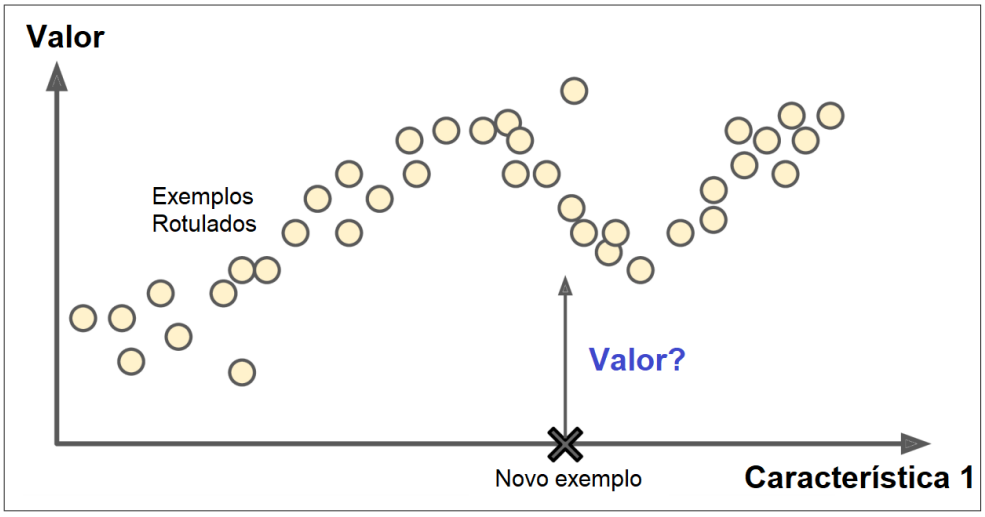
Figura 8 – Classificação



Fonte: Adaptada de (GÉRON, 2019)

O aprendizado supervisionado geralmente é utilizado para resolver um entre dois tipos de problema: Classificação ou Regressão (GÉRON, 2019). A classificação lida com o problema de atribuir, a novas entradas, uma entre um número de classes ou categorias discretas, como spam ou não-spam, no exemplo de classificação de e-mails. Já o problema de regressão lida com a previsão de variáveis contínuos, como por exemplo a previsão do

Figura 9 – Regressão



Fonte: Adaptada de (GÉRON, 2019)

preço de um carro.(BISHOP, 2006) Representações gráficas dos problemas de classificação e regressão podem ser vistos, respectivamente, nas figuras 8 e 9.

2.2.2 Métricas de Desempenho

As métricas de desempenho são parte fundamental do aprendizado de máquina. Elas são utilizadas tanto no treino dos modelos quanto em seu teste, de modo a avaliar quão bem o modelo se ajusta aos dados (GÉRON, 2019).

Durante o treino, a métrica de desempenho é utilizada para compor a função custo, que mede o erro cometido pelo modelo e deve ser minimizada. Para o teste, a métrica de desempenho é utilizada com o fim de comparar modelos e soluções diferentes. Esta comparação só pode ser feita utilizando métricas calculadas no teste porque só assim será possível avaliar a capacidade de generalização do modelo (GÉRON, 2019).

Para problemas de regressão, uma métrica de avaliação comumente utilizada é a Raiz do Erro Médio Quadrático (RMSE), que pode ser calculado pela equação 2.4. O quadrado do RMSE também pode ser usado, e recebe o nome de Erro Médio Quadrático (MSE) (GÉRON, 2019)

$$\text{RMSE}(\mathbf{y}, \hat{\mathbf{y}}) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2}, \tag{2.4}$$

onde: $\hat{y}^{(i)}$ é o valor previsto pelo modelo para o i-ésimo exemplo; $y^{(i)}$ é valor alvo do i-ésimo exemplo.

Uma outra métrica de desempenho comumente utilizada para problemas de regressão é o Erro Médio Absoluto (MAE), calculado pela equação 2.5. (GÉRON, 2019)

$$\text{MAE}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{m}\sum_{i=1}^{m}|\hat{y}^{(i)} - y^{(i)}|. \tag{2.5}$$

Também é possível realizar a normalização dos erros, que pode ser feito de algumas maneiras diferentes. Uma delas é a normalização pela média, em que divide-se o erro pelo valor médio da variável na amostra, $\bar{y}$:

$$\text{MAE Normalizado}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{\frac{1}{m}\sum_{i=1}^{m}|\hat{y}^{(i)} - y^{(i)}|}{\bar{y}} \tag{2.6}$$

Utilizar métricas diferentes para treino pode resultar em modelos diferentes: o MAE pode funcionar melhor que o RMSE para situações em que há muitos *outliers*, *i.e.*, exemplos cujo valor alvo é distante do comportamento comum do conjunto de dados. Isto ocorre porque o MAE coloca pesos proporcionalmente menores a valores mais distantes, quando comparado o RMSE; assim, ao minimizar o erro durante o treino do modelo, os *outliers*, que não são amostras representativas, terão menor influência no modelo final (GÉRON, 2019).

Uma outra métrica de desempenho muito utilizada para a avaliação de modelos de regressão, mas não como métrica de treino, é o coeficiente de determinação do modelo, R^2. O coeficiente de explicação indica a porção da variação explicada pelo modelo, e pode ser calculado pela equação 2.7:(BUSSAB; MORETTIN, 2010)

$$\text{R}^2(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \frac{\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{n}(y^{(i)} - \bar{y})^2}. \tag{2.7}$$

2.2.3 Algoritmos

Os algoritmos, em aprendizado de máquina, são os conjuntos de regras que definem como serão realizadas inferências sobre o problema a partir dos dados; a partir destas inferências, os algoritmos geram modelos, que contém as regras para a solução do problema especificado. Em contraste, em uma abordagem de *design*, o próprio algoritmo contem o conjunto de regras utilizado para resolver o problema especificado. (ABU-MOSTAFA; MAGDON-ISMAIL; LIN, 2012)

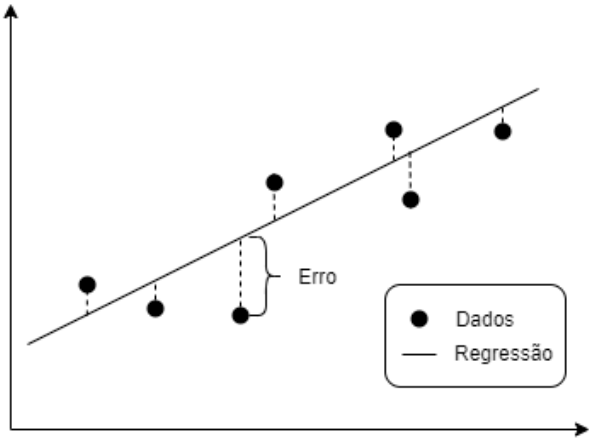
Quado um algoritmo treina um modelo, aquele ajusta os parâmetros deste de modo a otimizar a métrica de desempenho escolhida ao prever no conjunto de treino. A forma como estes parâmetros são ajustados dependem não só do algoritmo, mas também dos hiper-parâmetros definidos. Os parâmetros são definidos e ajustados de forma automática durante o treino, enquanto os hiper-parâmetros são valores definidos manualmente e que influenciam na forma como se busca por parâmetros ótimos.(GÉRON, 2019)

Ao estudar diferentes algoritmos de aprendizado de máquina, é importante estudar também os seus hiper-parâmetros e a maneira como eles afetam o processo de aprendizado, pois podem alterar completamente o modelo resultante.

2.2.3.1 Regressão Linear

A regressão linear, ilustrada na Figura 10 é um algoritmo de regressão baseado na aproximação da função alvo por um modelo linear: uma reta para duas dimensões, um plano para três, ou um hiperplano para mais dimensões.

Figura 10 – Regressão Linear



Um modelo linear de n dimensões é composto pela soma ponderada de n atributos, mais uma constante, chamada de termo de viés, como mostrado na equação 2.8 (GÉRON, 2019):

$$\hat{y} = h_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta} \cdot \mathbf{x}, \tag{2.8}$$

onde: $\mathbf{x}$ é o vetor $n+1$ de atributos, cujos elementos são os n atributos x_j, com $j > 0$ e x_0 é constante igual a 1; $\boldsymbol{\theta}$ é o vetor de parâmetros cujos elementos são os $n+1$ parâmetros θ_j; $\boldsymbol{\theta} \cdot \mathbf{x}$ é o produto escalar entre os vetores $\boldsymbol{\theta}$ e $\mathbf{x}$.

O treinamento de um modelo linear para um determinado conjunto de dados consiste em encontrar o valor de $\boldsymbol{\theta}$ de modo que o erro quadrático médio cometido seja mínimo. (GÉRON, 2019)

Tradicionalmente, utiliza-se o Método de Mínimos Quadrados para encontrar o valores dos parâmetros θ_j, uma fórmula fechada que minimiza o MSE entre a reta prevista e os exemplos:

$$\boldsymbol{\theta} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y}, \tag{2.9}$$

onde: $\boldsymbol{X}$ é a matriz $m \times (n+1)$ que contém todos os m exemplos $\boldsymbol{x}^{(i)}$ do conjunto de dados de treino, com todos os $n+1$ atributos por exemplo, inclusive o $x_0^{(0)}$ constante;

$\left(\boldsymbol{X}^T\boldsymbol{X}\right)$ é uma matriz inversível; $\boldsymbol{y}$ é o vetor contendo todos os valores alvo, de $y^{(1)}$ a $y^{(m)}$.

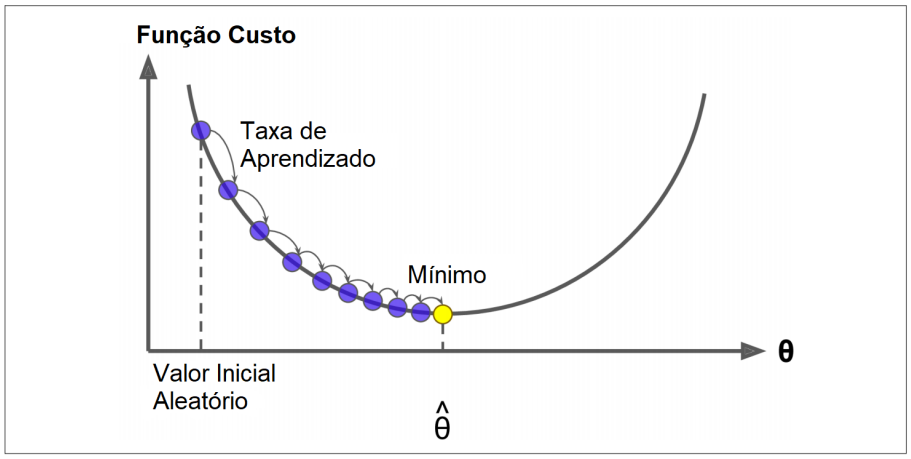
No caso de $\left(\boldsymbol{X}^T\boldsymbol{X}\right)$ ser uma matriz singular, *i.e.*, não inversível, ainda é possível calcular $\boldsymbol{\theta}$, bastando trocar a inversa da matriz da equação em 2.9 por uma pseudo-inversa, calculada pelo processo de Decomposição de Valores Singulares. (GÉRON, 2019)

A abordagem de calculo da regressão linear pelo Método de Mínimos Quadrados utiliza todos os dados de treino simultaneamente para calcular, exatamente, o valor dos parâmetros que minimize o erro quadrático médio entre os dados e a previsão pelo modelo linear. Neste processo, o cálculo da inversa ou da pseudo-inversa da matriz $\left(\boldsymbol{X}^T\boldsymbol{X}\right)$ é, geralmente, a parte mais computacionalmente custosa, e o custo computacional depende principalmente do número de atributos dos exemplos. (GÉRON, 2019)

Uma outra abordagem para o cálculo da regressão linear é o uso do método do Gradiente Descendente, uma técnica geral utilizada para minimizar funções duplamente diferenciáveis, como o erro quadrático médio da regressão linear. (ABU-MOSTAFA; MAGDON-ISMAIL; LIN, 2012)

O Gradiente Descendente inicia com uma inicialização randômica dos parâmetros do modelo, e segue com a melhoria progressiva do modelo, ajustando iterativamente os parâmetros no sentido de maior decrescimento da função de custo, *i.e.*, no sentido oposto ao gradiente da função. Este ajuste é feito a passos pequenos, definidos pela taxa de aprendizado, um importante hiper-parâmetro que afeta o quão rápido o algoritmo alcançará seu valor final. A Figura 11 exemplifica os passos realizados pelo algoritmo de gradiente descendente aplicados em uma função de custo convexa, como é o caso da função de custo da regressão linear. (GÉRON, 2019)

Figura 11 – Gradiente Descendente



Fonte: Adaptada de (GÉRON, 2019)

O Gradiente Descendente não irá, necessariamente, alcançar o mínimo global da função de erro para dada função de custo: algumas funções podem ter múltiplos mínimos locais, e

o gradiente descendente poderia parar em qualquer um destes, dependendo dos valores de inicialização e dos hiper-parâmetros definidos no algoritmo. No caso da regressão linear, entretanto, a função de custo é convexa e possuí somente um mínimo local, que é também o mínimo global, então é garantido que este será alcançado. (GÉRON, 2019)

Um passo do Gradiente Descendente é dado por:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^k - \eta \nabla_{\boldsymbol{\theta}} \mathrm{J}(\boldsymbol{\theta}), \tag{2.10}$$

onde: $\boldsymbol{\theta}^k$ é o vetor de parâmetros do modelo na iteração k; η é a taxa de aprendizado; $\nabla_{\boldsymbol{\theta}} \mathrm{J}(\boldsymbol{\theta})$ é o vetor gradiente da função de custo $\mathrm{J}(\boldsymbol{\theta})$, em relação a cada componente do vetor de parâmetros do modelo, dado pela equação 2.11 abaixo.

$$\nabla_{\boldsymbol{\theta}} \mathrm{J}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial}{\partial \theta_0} \mathrm{J}(\boldsymbol{\theta}) \\ \frac{\partial}{\partial \theta_1} \mathrm{J}(\boldsymbol{\theta}) \\ \vdots \\ \frac{\partial}{\partial \theta_n} \mathrm{J}(\boldsymbol{\theta}) \end{pmatrix} \tag{2.11}$$

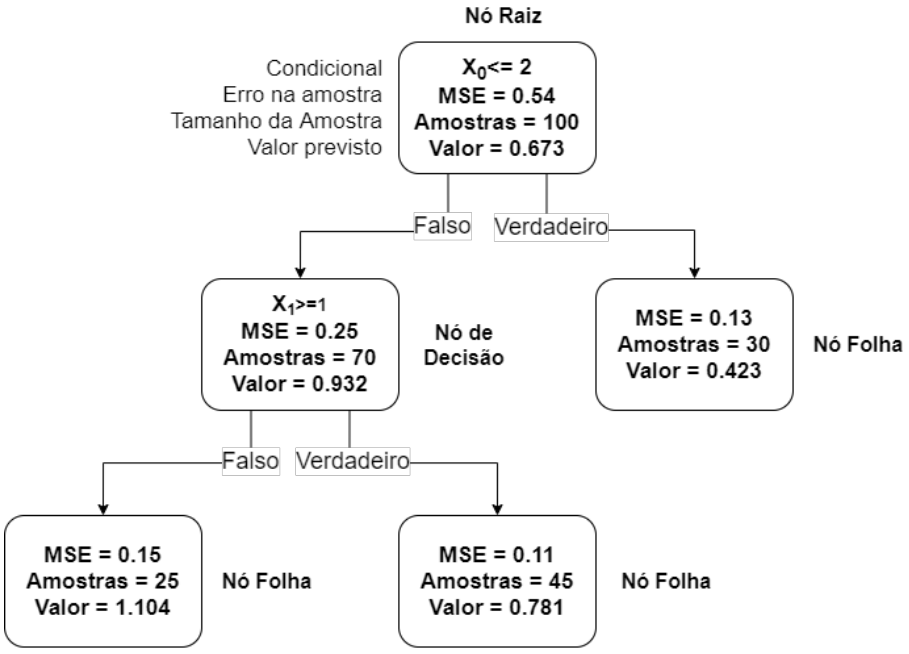
O método de gradiente descendente depende das escalas dos dados para alcançar os resultados de maneira mais rápida e com maior precisão. Caso as escalas dos dados sejam muito diferentes entre si, o algoritmo poderá demorar muito mais para alcançar um resultado ótimo, e muitas vezes não alcançá-lo. Para resolver este problema, é comum aplicar técnicas de normalização. Uma regularização comum é conhecida como a normalização *min-max*, em que os valores são subtraídos pelo mínimo e posteriormente divididos pela amplitude dos dados. Uma outra normalização, consiste na subtração da média dos dados e posterior divisão pelo desvio padrão. (GÉRON, 2019)

2.2.3.2 Árvore de Decisão

A árvore de decisão é um algoritmo de Aprendizado de Máquina simples e versátil, capaz de realizar tarefas de classificação e regressão não lineares, além de ser capaz de lidar com problemas com múltiplas saídas. A árvore de decisão é um algoritmo relativamente intuitivo e de fácil compreensão. (GÉRON, 2019)

As árvores de decisão podem ser representadas por conjuntos de decisões binárias, com todos os exemplos partindo do nó inicial, chamado de nó raiz, e seguindo por múltiplas decisões lógicas consecutivas pelos nós de decisão, que dividem progressivamente os dados até que eles alcancem um nó folha, que não se divide em mais nós. Os nós folhas são associados a um valor, no caso de problemas de regressão, ou a uma classe, no caso de problemas de classificação, e todos os exemplos recebem a classe ou valor do nó folha em que se encontram (MITCHELL et al., 1997). A Figura 12 exemplifica uma árvore de decisão de regressão usando duas características, X_0 e X_1.

Figura 12 – Árvore de Decisão



Um algoritmo muito utilizado para treino de uma árvore de decisão é o algoritmo Árvore de Classificação e Regressão (*Classification And Regression Tree* — CART). O CART inicia dividindo o conjunto de treino em dois usando uma característica k e um valor limite t_k, selecionando-os de forma que a função de custo $J(k, t_k)$ definida em 2.12 seja mínima (GÉRON, 2019) (BISHOP, 2006):

$$J(k, t_k) = \frac{m_{esq}}{m} E_{esq} + \frac{m_{dir}}{m} E_{dir}, \tag{2.12}$$

onde

- $m_{esq/dir}$ é o número de exemplos no nó da esquerda/direita;
- E_{esq} é a métrica de impureza dos subconjuntos: gini para problemas de classificação, e MAE ou MSE para regressão.

após encontrar com sucesso a regra que minimiza o custo da função em 2.12, dividindo o conjunto de testes inicial em dois subconjuntos, o processo é repetido para dividir os subconjuntos progressivamente, até que não seja mais possível fazer uma divisão que reduza o custo.

As árvores de decisão possuem uma grande quantidade de hiper-parâmetros que podem ser ajustados, adicionando critérios de parada do algoritmo CART ou de divisão das amostras, como por exemplo a profundidade máxima, o mínimo de amostras por divisão e o mínimo de amostras por folha. A profundidade máxima define o número máximo de nós que um subconjunto de dados pode passar antes de chegar em um nó de folha, o que pode forçar o algoritmo a parar mais cedo. O mínimo de amostras por divisão

define quantos exemplos são necessários para que um nó se divida, tornando-se um nó de decisão. O mínimo de amostras por folha define quantos exemplos são necessários para formar um nó folha. (GÉRON, 2019)

Estes e outros hiper-parâmetros das árvores de decisão atuam limitando a capacidade de ajuste do modelo, *i.e.*, são métodos de regularização, que atuam reduzindo a chance de ocorrer sobre-ajuste e consequentemente melhorando a generalização do modelo. Isso é especialmente importante para as árvores de decisão, que têm forte tendência de sobre-ajuste. (GÉRON, 2019)

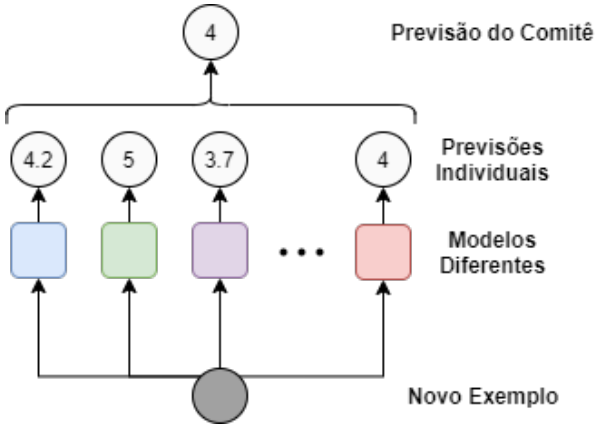
As árvores de decisão são capazes, também, de medir a importância de cada característica usada no treino, uma medida do quanto cada característica afeta na decisão final realizada pelo modelo. Para isso, calcula-se quanto cada nó que usa uma determinada característica reduz, em média, o erro cometido pela previsão, ponderado por quantos exemplos de treino passam por estes nós. Por fim, a importância das características é normalizada, de modo que a soma delas totalize 1. (GÉRON, 2019)

2.2.3.3 Floresta Aleatória

O algoritmo de floresta aleatória é um caso específico da implementação de um método de comitê (*ensemble*), uma técnica que utiliza a previsão de um conjunto de previsores, conhecido como comitê, para alcançar um resultado geralmente melhor que as previsões dos modelos individuais. Uma floresta aleatória é simplesmente um comitê de árvores de decisão treinados com porções diferentes do conjunto de dados. (GÉRON, 2019)

A previsão de um método de comitê pode ser dividido em duas partes: primeiro cada modelo do comitê realiza sua previsão para um exemplo específico, depois o método de agregação une as previsões de cada modelo. Para classificação, a agregação pode ser feita pela moda das classes previstas ou pela média das probabilidades previstas por classe, enquanto para a regressão a agregação é feita pela média entre os valores previstos. A Figura 13 ilustra o funcionamento de um comitê utilizado para a previsão em um problema de regressão. (GÉRON, 2019)

Figura 13 – Método Comitê



O método de comitê precisa de um conjunto de modelos diferentes; enquanto isto pode ser alcançado utilizando múltiplos algoritmos diferentes, uma outra forma possível é utilizar um mesmo algoritmo para treinar cada modelo, mas treiná-los com diferentes amostras do conjunto de dados. Nesta situação, a as amostras utilizadas para o treino de cada modelo podem ser construídas por uma reamostragem com substituição (agregação *bootstrap* (BISHOP, 2006), ou *bagging*) ou sem subsituição (*pasting*). (GÉRON, 2019)

O algoritmo de floresta aleatória é um comitê de árvores de decisão, geralmente treinadas utilizando *bagging*. O algoritmo da floresta aleatória adiciona, ainda, uma maior aleatoriedade ao processo, ao fazer com que, no treino de cada nó das árvores de decisão, o CART buscará a melhor característica entre um subconjunto aleatório das características dos exemplos. (GÉRON, 2019)

O algoritmo de floresta aleatória tem quase todos os mesmos hiper-parâmetros que as árvores de decisão voltados para controlar o crescimento das árvores individuais, além de hiper-parâmetros para controlar a agregação do comitê, como o número de árvores de decisão, o método de reamostragem, e o tamanho das amostras. (GÉRON, 2019)

Como a árvore de decisão, a floresta aleatória é capaz de calcular a importância das características utilizando o mesmo processo para cada árvore da floresta, e tomando a média dos valores entre todas as árvores, obtendo uma importância das características global para a floresta. (GÉRON, 2019)

2.2.4 Aprendizado por Transferência

O aprendizado por transferência (*transfer learning*) trata do uso de conhecimento adquirido ao treinar um modelo que realiza uma determinada tarefa, para auxiliar no aprendizado de uma novo modelo utilizado em tarefa semelhante. Uma definição mais rigorosa é: Dado um domínio origem D_S e uma tarefa de aprendizado T_S, um domínio alvo D_T e uma tarefa de aprendizado T_T, *transfer learning* objetiva melhorar o aprendizado de uma função preditiva alvo $f_T(\cdot)$ em D_T, usando a informação em D_S e T_S, onde $D_S \neq D_T$, ou $T_S \neq T_T$. (PAN; YANG, 2009)

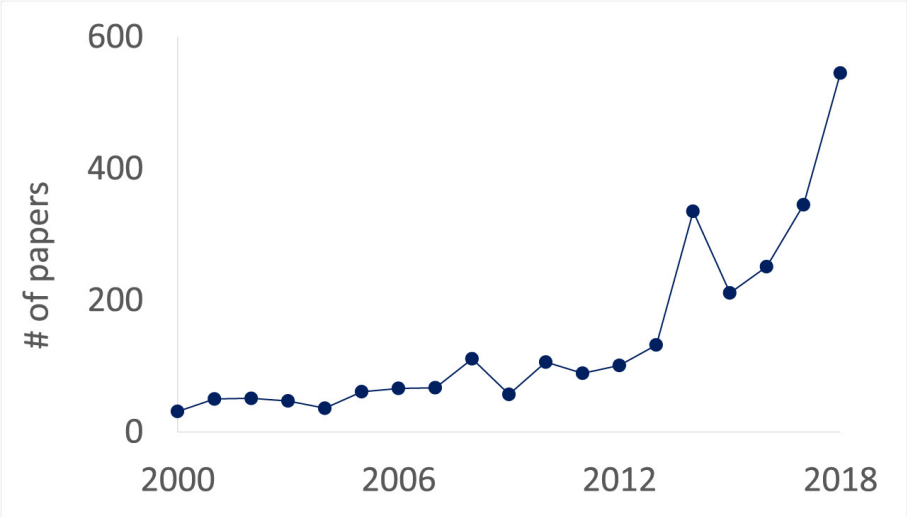
Transfer learning é comumente utilizado para casos em que deseja-se implementar um algoritmo em um contexto específico, mas não há dados suficientes para se obter um resultado satisfatório. Neste caso, utiliza-se um algoritmo genérico que possui bons resultados em um contexto diferente, mas semelhante, e aplica-se as técnicas de *transfer learning* para alcançar resultados melhores no contexto específico desejado. (SEGEV et al., 2016)

Em Segev et al. (2016), os autores desenvolveram dois algoritmos diferentes que permitem realizar uma transferência de modelo entre dois domínios utilizando florestas aleatórias, e os algoritmos desenvolvidos pelos autores podem ser implementados de forma conjunta.

2.3 TRABALHOS RELACIONADOS

O campo de aprendizado de máquina cresceu muito rapidamente nos últimos anos, tanto nas pesquisas realizadas no campo quanto nas aplicações. Na indústria de Óleo e Gás não é diferente, com um crescimento exponencial de publicações sobre aprendizado de máquina na indústria, como mostrado na Figura 14. Entretanto, a quantidade de projetos implementados é baixa quando comparado a outras indústrias, muitas vezes se limitando a publicações de pesquisa e projetos pilotos para prova de conceito. Apesar disso, a indústria possui grandes oportunidades para a implementação destes projetos, devido ao forte conhecimento dos especialistas e à grande quantidade de dados produzidos. (HAJIZADEH, 2019)

Figura 14 – Número anual de *papers* no jornal OnePetro com o tópico de Aprendizado de Máquina entre 2000 e 2018.



Fonte: (HAJIZADEH, 2019)

Em Alkinani et al. (2019), os autores realizaram uma revisão de literatura sobre as aplicações de redes neurais artificiais na indústria de petróleo. São citados diversos estudos para aplicação das redes neurais na prospecção, perfuração e produção de poços, bem como na engenharia de reservatórios.

Aplicações na perfuração de poços de petróleo citados no trabalho de Alkinani et al. (2019) incluem diagnóstico de desgaste em brocas durante a perfuração, seleção de melhor broca para perfuração de uma seção, previsão de problemas de perfuração e classificação de frases em relatórios de perfuração, entre outros.

Shadizadeh, Karimi e Zoveidavianpoor (2010) trabalham com a previsão, com redes neurais artificiais, da ocorrência de prisões de coluna durante a perfuração de poços de petróleo, enquanto o trabalho de Elmousalami e Elaskary (2020) realiza a classificação automática de casos de prisão de coluna de forma a acelerar a resolução do problema. O

trabalho de Gurina et al. (2020) lida com a detecção da ocorrência de prisões de coluna e outros acidentes de perfuração durante a perfuração direcional.

Quanto a trabalhos que realizam a previsão da ROP para poços de petróleo, Barbosa et al. (2019) realizaram uma revisão de literatura extensiva do tema, com foco em abordagens que utilizam técnicas de aprendizado de máquina e em como estes modelos podem ser usados para otimizar as atividades de perfuração. Entre os trabalhos revisados, houve a criação de um total de 70 modelos de aprendizado de máquina. Destes, a maioria consistiu de redes neurais artificiais, que apareceram 39 vezes, seguido de comitês homogêneos, como a floresta aleatória, utilizados 9 vezes.

Barbosa et al. (2019) também avaliou as entradas mais comumente utilizadas entre os 53 trabalhos. Entre estes, o peso sobre broca e o RPM foram os mais utilizados, aparecendo 41 vezes. Em seguida, há a profundidade do poço, o fluxo de fluido de perfuração, o peso do fluido de perfuração, o diâmetro da broca, a resistência compressiva uniaxial ou outra resistência de rocha, e o desgaste da broca.

O trabalho de Yuswandari, Prayoga e Purba (2019) utilizou inicialmente regressões lineares múltiplas para seleção de parâmetros, e posteriormente uma rede neural artificial não especificada para prever a ROP a cada 100 metros em um poço, após treinar em um outro poço presente no mesmo campo. Os dados, coletados de relatórios de perfuração, incluíam profundidade, peso sobre broca, velocidade de rotação, torque, pressão de *standpipe* e reologia dos fluidos de perfuração, mas foram selecionados somente o peso sobre broca e a velocidade de rotação.

Em Shi et al. (2016), os autores treinaram um modelo convencional de rede neural artificial para a previsão de ROP, e o utilizaram como base de comparação para duas outras arquiteturas de redes neurais da classe de máquinas de aprendizado extremo (ELM). Os dados de entrada foram escolhidos entre aqueles que poderiam ser coletados em tempo real, e incluíam diâmetro e tipo da broca, desgaste na broca, resistência compressiva uniaxial, abrasividade e perfurabilidade da formação, velocidade de rotação, peso sobre broca, pressão de bomba de lama, viscosidade e densidade do fluido de perfuração. Comparando R^2, MSE, RMSE e variância considerada pelo modelo, os autores reportaram que as máquinas de aprendizado extremo alcançaram resultados melhores que as redes neurais convencionais em um tempo menor de treino.

Bataee, Irawan e Kamyab (2014) realizaram previsão de taxa de perfuração e otimização de parâmetros de perfuração baseados em um modelo desenvolvido com redes neurais artificiais. Os dados utilizados estavam presentes a cada 100 metros, e continham o diâmetro da broca, a profundidade, o peso sobre a broca, a velocidade de rotação e a densidade do fluido de perfuração. Os resultados da previsão com redes neurais artificiais foi comparado com modelos tradicionais de Bingham, de Bourgoyne e Young, e de Warren pelo R^2, e o modelo de redes neurais mostrou-se mais eficiente.

O trabalho apresentado em Eskandarian, Bahrami e Kazemi (2017) consistiu na previ-

são da ROP utilizando modelos de floresta aleatória e uma rede neural do tipo perceptron multi-camadas monótono. O conjunto de dados inicial utilizado continha informações de peso sobre broca, velocidade de rotação, fluxo de fluido de perfuração, pressão de bomba, inclinação, azimute e profundidade do poço, e, para o fluido de perfuração, o peso, viscosidade de funil, viscosidade plástica, tensão limite de escoamento e forças géis de 10 segundos e 10 minutos. Antes de aplicar os modelos, utilizou o método Cubist para reduzir as entradas para 6 e 4 entradas, comparando os resultados destas duas abordagens. A rede neural apresentou o melhor resultado, mas não significativamente melhor que a floresta aleatória.

Na dissertação de GANDELMAN (2012), o autor utiliza redes neurais artificiais para predição de ROP e otimização dos parâmetros de perfuração. Os dados estavam em formato de série temporal, amostrados a cada 15 a 30 segundos, e continham velocidade de rotação, peso sobre broca, profundidade, fluxo de fluido de perfuração, diâmetro de broca e diferencial de pressão poço-formação (*overbalance*), além de uma variável categórica indicando a litologia perfurada. Os resultados obtidos para a rede neural foram comparados aos modelos tradicionais de Maurer e de Young, e observou-se melhores resultados para a rede neural.

Os autores de Hegde et al. (2017) compararam a performance dos modelos tradicionais de previsão de ROP de Bingham, de Motahhari e de Hareland com três modelos baseados em dados: Regressão Linear, Floresta Aleatória e um método de comitê não especificado. No trabalho, foram utilizados como entrada o peso sobre broca, o fluxo de entrada do fluido de perfuração, a resistência compressiva uniaxial da rocha, e a velocidade de rotação da broca, todos indexados na profundidade a cada $0,25$ pés, e coletados durante a perfuração. O melhor resultado obtido foi com o modelo de floresta aleatória, que resultou em menores erros normalizados e maiores R^2 que os modelos tradicionais até mesmo se treinado com somente 30% dos dados. Avaliou-se também a eficácia do modelo quando treinado 1) exclusivamente com dados da formação sendo testada, 2) somente com dados de outras formações ou 3) com dados misturados de formações, e observou-se que o primeiro caso é o mais eficaz, seguido do terceiro caso, e por fim pelo segundo.

Em outro trabalho, Hegde et al. (2020), os autores utilizaram os modelos de floresta aleatória desenvolvidos pela metodologia anterior para previsão de ROP, junto de modelos para previsão de outras variáveis, com o objetivo de realizar a otimização da perfuração.

Nesse trabalho, serão utilizados os algoritmos de regressão linear por gradiente descendente, árvore de decisão e floresta aleatória. As métricas utilizadas para avaliar serão o MAE normalizado pela média e o R^2. No final, o melhor modelo será utilizado para avaliar a possibilidade do uso de aprendizado por transferência, utilizando os dados de um poço para prever o resultado para outro poço no mesmo campo. As características disponíveis para treino são a velocidade de rotação do *top drive*, o torque do *top drive*, o peso sobre broca, a pressão de *standpipe*, a profundidade total perfurada, o fluxo de

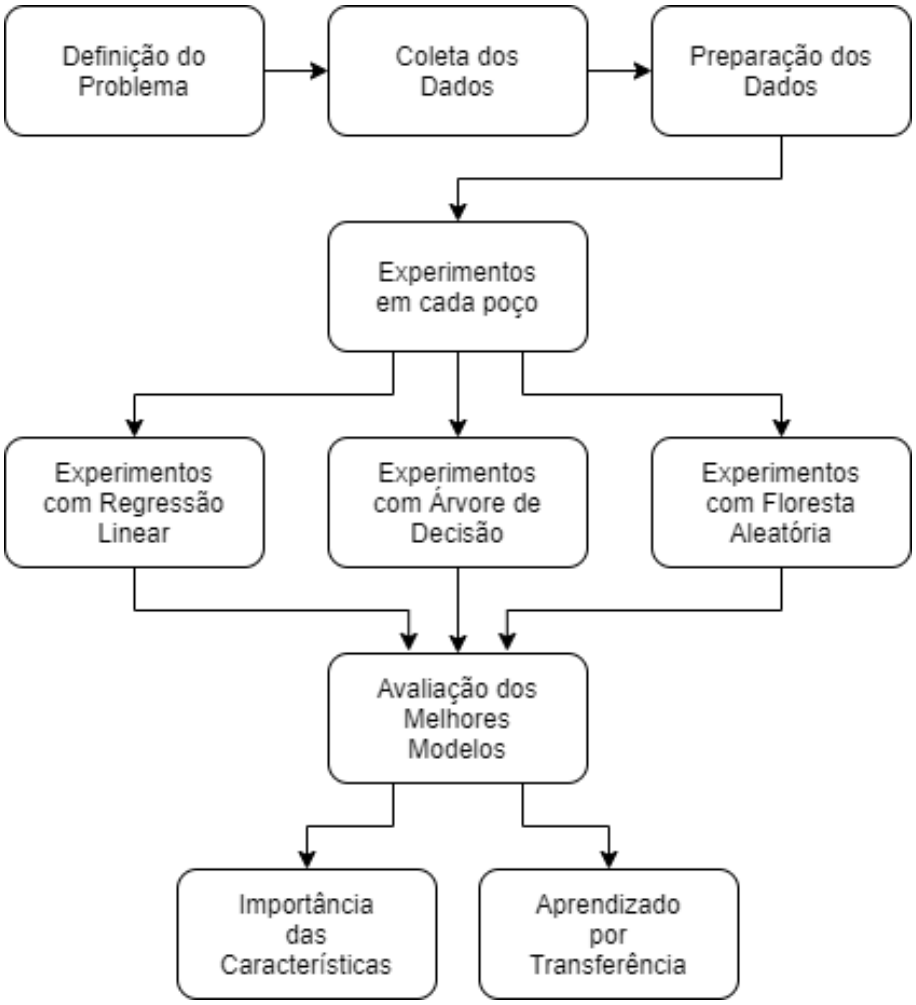
entrada do fluido de perfuração, o fluxo de saída do fluido de perfuração, e a inclinação do poço.

Inspirado pela maioria dos trabalhos, escolheu-se tratar os dados indexando-os no tempo. Baseado nos resultados positivos obtidos no trabalho de Hegde e Gray (2017), o algoritmo de floresta aleatória foi selecionado. Por falta de disponibilidade, parâmetros que informem propriedades da formação rochosa perfurada, geometria e tipo de broca, e propriedades físicas do fluido de perfuração, comumente utilizados nos trabalhos relacionados, não foram utilizados.

3 METODOLOGIA

As etapas desse trabalho estão delimitadas no fluxograma da Figura 15.

Figura 15 – Etapas do projeto de Aprendizado de Máquina



3.1 DEFINIÇÃO DO PROBLEMA

O a função a ser estimada pelo algoritmo de aprendizado, como definido na seção 1.2, é a taxa de perfuração de poços de petróleo *offshore* (ROP), o que será feito a partir de dados obtidos em tempo real.

Foram realizados experimentos computacionais para avaliar a eficácia de diferentes métodos de aprendizado de máquina para a previsão da taxa de perfuração.

3.2 COLETA DE DADOS

Os dados utilizados nos experimentos numéricos foram cedidos pela Ocyan, uma empresa do setor de óleo e gás *upstream offshore* atuante no Brasil e exterior. Os dados foram disponibilizados para 4 poços diferentes da Bacia de pré-sal de Santos, dois do campo de Sépia e dois do campo de Búzios. Informações sobre os poços podem ser encontrados na Tabela 1.

Tabela 1 – Descrição dos conjuntos de dados

	Campo	**Qntd Dados**	**Profundidade**
Poço 1	Búzios	761467	Lâmina d'água 1855 m e dados 3160–6016 m
Poço 2	Búzios	710753	Lâmina d'água 2001 m e dados 3040–6063 m
Poço 3	Sépia	365857	Lâmina d'água 2139 m e dados 5323–5600 m
Poço 4	Sépia	252955	Lâmina d'água 2120 m e dados 5376–5557 m

As entradas selecionadas para o estudo foram escolhidas baseado em características utilizadas em trabalhos relacionados, na opinião de especialistas da área para características que afetariam os dados, e por fim baseado na disponibilidade de cada uma delas. As características selecionadas foram:

- Profundidade do poço (m)
- Profundidade da broca (m)
- Peso sobre broca (klbf)
- Torque do *top drive* (klbf·ft)
- Velocidade de rotação do *top drive* (rpm)
- Vazão de entrada de fluido de perfuração (gal/min)
- Vazão de saída de fluido de perfuração (gal/min)
- Pressão de *standpipe* (psi)
- Inclinação do poço (^{o})

As oito primeiras características acima são dados em tempo real coletados por sensores em equipamentos na superfície da sonda e enviados para a base de operações da empresa por meio de protocolos de comunicação WITSML (*Wellsite information transfer standard markup language*) e armazenados em bancos de dados MongoDB, que são do tipo NoSQL, com uma frequência de 1 dado por segundo.

A inclinação do poço é obtida por meio de um *directional survey* e é disponibilizada para uma quantidade discreta e comparativamente baixa de valores de profundidade. Estes valores são medidos após a perfuração da determinada profundidade e disponibilizados

a intervalos de tempo variáveis. A informação é registrada por um operador na sonda de perfuração, onde é inicialmente armazenada e posteriormente transmitida para a base de operações da empresa.

Os dados considerados não incluem informações sobre as propriedades físicas das formações perfuradas, do fluido de perfuração utilizado ou de geometria e tipo de broca utilizados. Estes dados, apesar de recomendados por especialistas e serem utilizados em outros trabalhos por influenciarem consideravelmente a ROP, não estavam disponíveis.

Como a função alvo é a ROP, os dados desejados limitam-se àqueles durante a atividade de perfuração, desconsiderando, por exemplo, atividades de revestimento e repasse, que, embora vitais para o processo de perfuração, são desconsideradas para a estimativa da ROP. Para isso, durante a coleta de dados, utilizou-se os relatórios diários de perfuração, que indicam, entre outras informações, os períodos em que ocorreu perfuração, com precisão máxima de meia hora. Os dados disponibilizados foram coletados destes períodos identificados, com adicional de 15 minutos antes e depois, de modo a corrigir imprecisões devido à precisão limitada dos dados.

3.3 PREPARAÇÃO DOS DADOS

Para realizar a preparação dos dados, bem como para implementar as etapas seguintes, utilizou-se a linguagem de programação Python, versão 3.8.6 com as bibliotecas numpy[1] e pandas[2] para leitura, análise e tratamento dos dados, e as bibliotecas matplotlib[3] e plotly[4] para geração de gráficos de diferentes tipos.

3.3.1 Análise Exploratória dos Dados

A etapa de preparação dos dados inicia-se com uma análise exploratória, com o objetivo de conhecer melhor os dados, encontrar padrões e anomalias e verificar suposições sobre os dados a partir de gráficos e estatísticas. Os dados em tempo real e os dados obtidos pelo *directional survey* foram analisados separadamente.

Uma análise inicial de estatísticas dos dados em tempo real revelou o número de linhas de dados por poço, e também que, para cada uma das linhas, não havia colunas com dados faltantes. Outras estatísticas reveladas sobre cada características divida por poço estão resumidas na Tabela 2.

Observando inicialmente os dados medidos a cada segundo, notou-se ao analisá-los em formato de tabela que, para alguns poços, ocorria falhas ocasionais em que o tempo entre dois dados subsequentes era maior que 1 segundo, sem considerar os casos em que

[1] https://numpy.org/
[2] https://pandas.pydata.org/
[3] http://matplotlib.org/
[4] https://plotly.com/python/

Tabela 2 – Amplitude de dados em tempo real

	Amplitude			
Característica	Poço 1	Poço 2	Poço 3	Poço 4
Profundidade de poço (m)	3160,3–6016,6	3039,7–6063,4	5323,4–5600,2	5376,8–5556,8
Profundidade da broca (m)	3159,3–6016,6	3038,3–6063,4	5322,5–5600,2	5373,0 -5556,7
Peso sobre broca (klbf)	-79,7–930,7	-151,4–456,6	-55,9–519,3	-82,1–476-4
Torque (klbf · ft)	-22,1–55,4	-15,7–67,4	0,0–50,0	0,0–15603,4
Velocidade de rotação (rpm)	0,0–239,1	0,0–241,3	0,0–80,0	0,0–81,1
Vazão de entrada de fluido (gal/min)	0,0–5483,4	0,0–9298,6	0,0–558,7	0,0–675,4
Vazão de saída de fluido (gal/min)	-0,3–828,4	0,0–601,2	0,0–292,8	0,0–365,7
Pressão de *standpipe* (psi)	18,8–4438,2	0,0–5859,5	15,0–4127,2	15,0–5260,9

a perfuração foi interrompida para realizar outras atividades. Este é o primeiro problema identificado nos dados (**P1**).

Analisando estes dados em gráficos de formato de série temporal, notou-se alguns outros problemas nos dados:

P2 - Os dados continham porções em que não se estava perfurando: porções logo antes de iniciar ou logo após acabar de perfurar e porções em que realizava-se a conexão de novos tubos;

P3 - Nos poços 1 e 2, os valores de vazão de entrada de fluido de perfuração maiores que 1200 gal/min eram erros de medição ou transmissão, representando menos de $0,032\%$ dos dados e o menor valor destes era de acima de 4500 gal/min, valor muito superior às vazões comumente utilizadas na perfuração. O mesmo vale para os dados no poço 4 com torque maior que 50, em que o menor valor destes era de mais de 1000 klbf·m.

P4 - Em uma porção de 30 min de dados em um dos poços, o sinal de profundidade parou de crescer, indevidamente;

P5 - Havia algumas poucas instâncias de dados em que houve redução na profundidade do poço, que se consideradas poderiam ter grande efeito na ROP prevista pelos modelos, mesmo que não sejam exemplos representativos do comportamento dos dados.

P6 - Havia algumas poucas lacunas maiores nos dados de 2 dos poços, com porções de múltiplos metros perfurados sem dados;

A solução destes problemas será apresentada nas seções seguintes.

Quanto aos dados do *directional survey*, eles incluem a inclinação do poço medida graus em determinadas profundidades, com um total de linhas menor que 100 por poço. A exploração dos dados, neste caso, foi feita observando diretamente a tabela de valores. Não havia dados faltantes nas colunas, mas nem sempre os dados de inclinação estavam disponíveis para toda a amplitude de profundidade dos poços. Adicionalmente, no poço 3

foi identificado um caso em que a inclinação estava incorreta, porém ocorreu em profundidades abaixo de onde havia dados em tempo real para o poço, e portanto não afetaria o resultado.

As amplitudes dos dados de *directional survey* para as porções do poço com dados em tempo real estão na Tabela 3. Como mostrado na tabela, os poços 1 e 4 possuem dados de inclinação somente para parte dos dados, sendo ainda mais limitado no poço 1. Esta falta de dados para algumas profundidades será resolvida posteriormente, na subseção 3.3.4.

Tabela 3 – Amplitude de dados do *directional survey*

	Amplitude			
Características	Poço 1	Poço 2	Poço 3	Poço 4
Inclinação (^{o})	1,15–3,06	0,17–14,52	1,41–5,53	0,93–1,12
Profundidade de poço (m)	5703,6–5966,2	3967–5873	5296–5600	5403,7–5486,5

3.3.2 Correção de problemas nos dados

Antes da realização dos experimentos numéricos nos dados, é necessário primeiro limpá-los corrigindo os erros identificados anteriormente.

Primeiramente, eliminou-se os dados de operações diferentes da perfuração, o problema **P2**. Para isso, foi feito, inicialmente, uma limpeza manual dos dados, eliminando as porções identificadas visualmente como manobra e repasse; posteriormente, filtrou-se os dados, eliminando aqueles em que a diferença entre a profundidade de broca e a profundidade do poço fossem maiores que 1 metro e em que a vazão de fluido de perfuração fosse menor que 120 gal/min, assim eliminando a operação de conexão. O valor limite de vazão de fluido de perfuração foi escolhido como estimativa conservadora para englobar casos extremos considerando que, durante operações normais de perfuração, a vazão de fluido de perfuração se mantém entre 400 gal/min e 1200 gal/min. Já a distância de 1 metro foi escolhida para levar em consideração o efeito de *bit bounce*, em que a sonda perde contato com a formação momentaneamente.

Para resolver o problema **P3** aplicou-se um segundo filtro para os dados, excluindo aqueles em que o vazão de entrada do fluido de perfuração é maior que 2000 gal/min e o torque do *top drive* é maior que 100 klbf·ft.

Para resolver o problema **P4**, a porção em que o sinal de profundidade parou de ser atualizado, primeiro identificou-se o intervalo de tempo em que a profundidade não se atualizava, e posteriormente preencheu-se este intervalo com a regressão linear feito com os valores antes e depois do congelamento, como mostrado nas figuras 16 e 17.

Para resolver o problema **P5**, os dados em que a profundidade do poço reduziu foram eliminados após o cálculo da "ROP instantânea"de cada exemplo, calculada a partir da diferença entre as profundidades atual e do exemplo anterior: caso a "ROP instantânea"seja negativa, o exemplo é desconsiderado.

Figura 16 – Porção com profundidade congelada

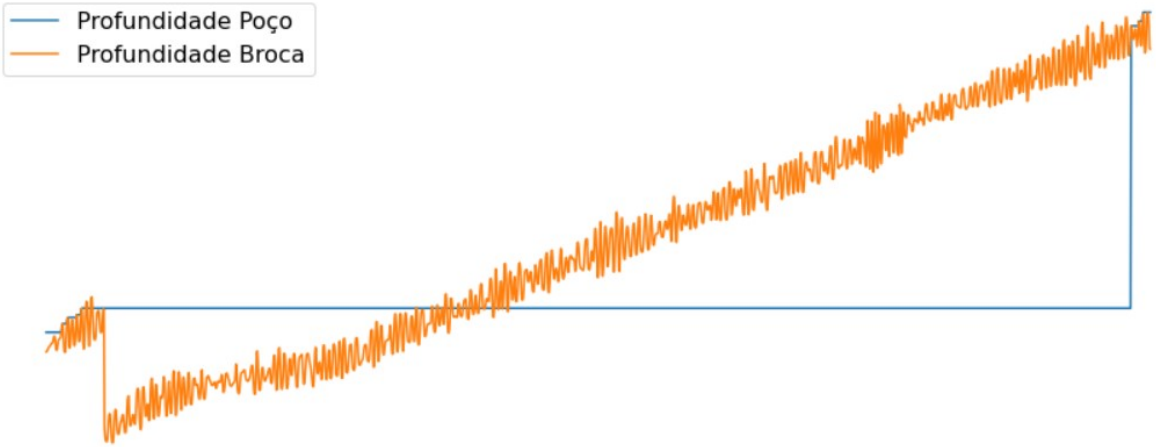
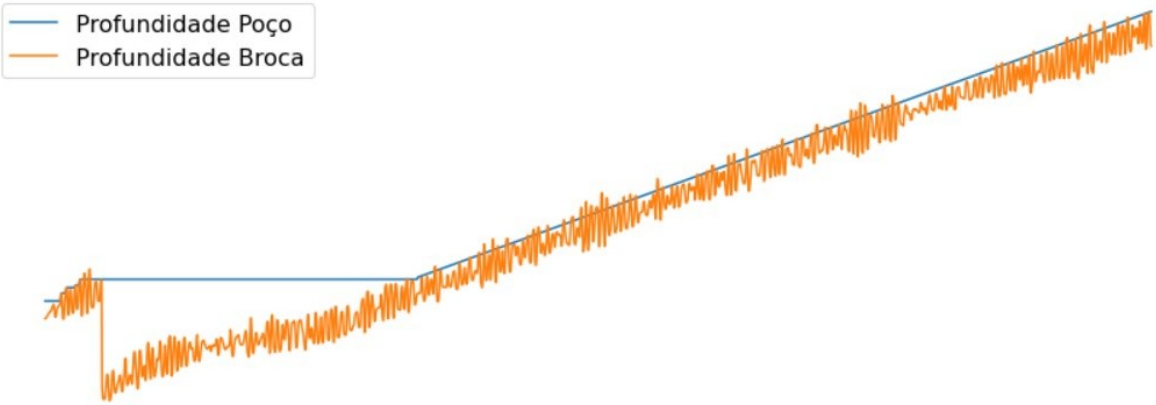


Figura 17 – Porção com profundidade corrigida por regressão linear



Por fim, os problemas das lacunas dos dados **P1 e P6** foram resolvidos de maneira automática em uma etapa posterior do processamento dos dados, definida na seção 3.3.4, em que agregou-se os dados na profundidade.

3.3.3 Engenharia de Características

O processo de engenharia de características consiste na seleção de características que serão utilizadas pelo algoritmo, bem como na criação de novas características a partir das já existentes, de modo a conferir melhores informações sobre a tarefa.

Neste processo, foram criados duas novas características: A proximidade da broca e a potência. A proximidade da broca foi obtida calculando a diferença entre a profundidade do poço e a profundidade da broca, e foi criada porque a relação entre a profundidade do poço e a profundidade da broca que afeta o processo da ROP, e não o valor isolado da profundidade da broca. Já a potência foi obtida pelo produto entre a velocidade de rotação e o torque; este valor foi criado na tentativa de modelar a energia utilizada para

perfurar a rocha, e foi baseado na equação 2.3, que contém o produto entre estas variáveis.

Após a criação das novas características, eliminou-se a profundidade da broca por considerar-se que seria um valor de pouca relevância para o modelo.

3.3.4 Transformação dos Dados

Neste passo, realizou-se a inclusão de uma coluna com a inclinação, retirada do *directional survey*, e a agregação dos dados pela profundidade. A motivação para essas alterações são explicadas a seguir.

Como a inclinação é medida para apenas um pequeno conjunto de profundidades específicas, é necessário estendê-la para as profundidades em que não há valor especificado. Essa extensão foi feita propagando a inclinação para as profundidades superiores, i.e., utilizando o valor da última inclinação medida para valores de profundidades maiores. Caso não haja inclinações anteriores, considerou-se que seu valor é 0.

Por fim, os dados foram agregados pela profundidade, retirando-os do domínio do tempo. Para isso, definiu-se faixas de $0,08$ metros de profundidade, agregando todos os valores medidos enquanto a profundidade encontrava-se em cada uma destas faixas utilizando a média. O valor de $0,08$ metros foi selecionado por ser aproximadamente igual a $0,25$ pés, valor utilizado no trabalho de (HEGDE et al., 2017).

Neste processo de agregação de dados, os problemas das lacunas de dados é resolvido: dados faltantes são desconsiderados ao calcular a média dos valores, utilizando somente os valores existentes que estejam na faixa de profundidade considerada. Caso a lacuna de dados seja tão grande que não exista dados para uma determinada profundidade, esta faixa de profundidade é excluída da análise.

Dessa forma, as características que são consideradas são a profundidade de poço, proximidade de broca, peso sobre broca, torque, velocidade de rotação, potência (torque vezes rotação), vazão de entrada de fluido, vazão de saída de fluido, pressão de *standpipe* e inclinação. Estatísticas das características após toda a limpeza e transformações podem ser encontradas na Tabela 4.

Tabela 4 – Amplitude dos dados após transformação

	Amplitude			
Característica	Poço 1	Poço 2	Poço 3	Poço 4
Profundidade de poço (m)	3160,3–6016,7	3039,7–6063,5	5323,4–5600,2	5376,9–5556,9
Proximidade de broca (m)	-0,004–0,839	0,000–0,628	0,000–0,688	0,000–0,708
Peso sobre broca (klbf)	-7,01–813,93	-14,71–54,85	-16,30–32,23	-4,35–39,31
Torque (klbf · ft)	-4,05–39,36	1,54–39,58	5,08–8,70	0,00–19,19
Velocidade de rotação (rpm)	0,00–231,34	0,00–234,60	29,51–79,69	15,81–79,71
Potência (rpm · klbf)	-462,55–5763,34	0,00–7638,00	0,02–596,84	0,00–843,86
Vazão de entrada de fluido (gal/min)	244,89–887,62	128,12–867,01	449,34–553,28	459,91–665,71
Vazão de saída de fluido (gal/min)	21,67–240,01	1,43–275,99	0,00–268,38	0,00–258,64
Pressão de *standpipe* (psi)	1515,6–4235,1	582,6–3456,3	2306,5–3896,4	2193,7–5183,8
Inclinação (^{o})	0,00–1,76	0,17–14,52	1,41–5,53	0,00–1,12

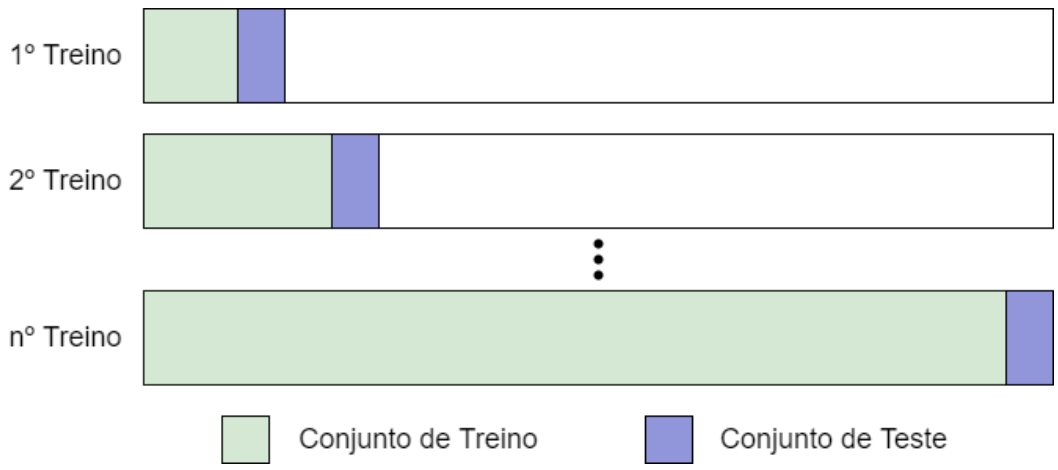
Nos experimentos numéricos, que serão definidos na seção 3.4, os experimentos com regressão linear utilizam o Gradiente Descendente para treino, e como apontado no final da subseção 2.2.3.1, este algoritmo depende que a escala dos dados seja semelhante entre si. Assim, nestes casos, implementou-se ainda a normalização *min-max*.

3.3.5 Divisão de Treino e Teste

Após o tratamento dos dados, é necessário organizar como serão divididos os conjuntos de treino e teste. Como a perfuração é um processo sequencial, em que não é possível perfurar uma determinada profundidade antes de perfurar as profundidades anteriores, e as propriedades das rochas perfuradas variam com a profundidade, é necessário tomar cuidado para não treinar com dados posteriores aos dados de teste, pois estes dados não estariam disponíveis para treino antes de serem utilizados para previsão.

Assim, utilizou-se o esquema de treino e teste representado na Figura 18, em que define-se uma porção dos dados para treino, e testa com os dados que os sucedem. Em seguida, aumenta-se o conjunto de treino, e testa com os dados seguintes, repetindo este processo até não haver mais dados. Considerou-se inicialmente o conjunto dos 360 primeiros dados do poço para treino, e os 180 seguintes para teste do modelo treinado. Para cada novo treino, o conjunto de treino foi aumentado em 360.

Figura 18 – Divisão de Treino e Teste



Os dados dos diferentes poços estão disponíveis para amplitudes de profundidade diferentes, conforme Tabela 1, então o número de treinos e testes realizados irá variar dependendo do poço. Esta informação está resumida na Tabela 5.

Tabela 5 – Número de treinos realizados por poço

	Poço 1	Poço 2	Poço 3	Poço 4
Quantidade de Treinos	89	95	9	5

Como o processo de perfuração é altamente não linear (BARBOSA et al., 2019) e o valor da ROP por profundidade é errático, realizou-se uma remoção de *outliers* do

conjunto de treino. De todos os dados de treino, foram removidos os $0,5\%$ de maior valor de ROP do conjunto de treinos. Esta remoção precisou ser feita após a divisão dos conjuntos de treino e teste pois ela é baseada em uma estatística calculada sobre o conjunto.

3.4 EXPERIMENTOS PARA CADA POÇO

Os experimentos numéricos a seguir foram feitos utilizando a biblioteca scikit-learn[5] versão 0.24.1 do Python para aplicação dos algoritmos de regressão linear, árvore de decisão para regressão e floresta aleatória para regressão. As estatísticas foram calculadas utilizando funções da biblioteca pandas versão 1.1.4, e os diagramas de caixa e outros gráficos foram criados com a biblioteca matplotlib versão 3.3.2.

Após a divisão dos conjuntos de treino e teste, definida na subseção 3.3.5, realizou-se os experimentos numéricos, em que treinou-se modelos de regressão linear, árvore de decisão e floresta aleatória com diferentes hiper-parâmetros em cada conjunto de treino de cada poço, e em seguida testou-se os modelos em cada conjunto de teste de cada poço.

Como foram realizados múltiplos testes para modelos treinados com o mesmo algoritmo, e portanto há múltiplos valores para as métricas, estas precisam ser analisadas de forma estatística. A quantidade de pares de conjunto de treino e teste foi mostrada na Tabela 5.

Para o treino da regressão linear, escolheu-se utilizar o método de Gradiente Descendente; assim, como falado no final da subseção 3.3.4, também foi necessário normalizar os dados, utilizando a normalização *min-max*. No treino da árvore de decisão, variou-se o hiper-parâmetro de tamanho de amostra mínimo por folha, definido a partir de uma porcentagem do tamanho do conjunto de treino, com os valores representados na Tabela 6. No treino da floresta aleatória, variou-se os hiper-parâmetros de profundidade máxima de árvore e de número de árvores (estimadores) do comitê, com os valores representados na Tabela 7.

Tabela 6 – Variação de Hiper-parâmetros — Árvore de Decisão

	Árvore de Decisão				
Tamanho de Amostra Mínimo por Folha	0,1%	0,2%	0,33%	0,5%	1%

Tabela 7 – Variação de Hiper-parâmetros — Floresta Aleatória

	Floresta Aleatória											
Profundidade Máxima de Árvore	3	6	8	10	3	6	8	10	3	6	8	10
Número de Árvores	100	100	100	100	200	200	200	200	400	400	400	400

[5] https://scikit-learn.org/

3.4.1 Importância das Características

Para medir quanto cada característica influencia na previsão final da taxa de perfuração para cada poço, utilizou-se a medida de importância das características para a árvore de decisão e floresta aleatória, como definido nas subseções 2.2.3.2 e 2.2.3.3.

As combinações de algoritmo e hiper-parâmetros utilizados para medir a importância das características foram aquelas que tiveram melhores resultados para cada poço. Estas combinações de algoritmo e hiper-parâmetros foram usadas para treinar um modelo usando todos os dados do poço, para cada poço, e este modelo foi usado na medida da importância das características desse poço.

3.4.2 Aprendizado por Transferência

Após os experimentos numéricos anteriores, o melhor modelo é escolhido para testar a possibilidade de aprendizado por transferência. Para este experimento numérico, foram escolhidos os poços 1 e 2 por ambos serem do campo de Búzios, e portanto possuírem características litológicas semelhantes, e por possuírem dados para uma maior extensão de poço que os poços 3 e 4.

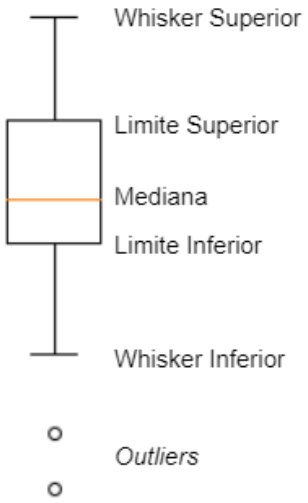
Os experimentos numéricos de aprendizado por transferência foram feitos tanto treinando no poço 1 e testando no poço 2 quanto treinando no poço 2 e testando no poço 1. Foram testadas duas situações: treinar um modelo todos os dados do primeiro poço e usar somente este modelo para prever no segundo poço, o que será chamado de modelo simples; e treinar um modelo com todos os dados do primeiro poço e um segundo modelo com os dados do segundo poço, utilizando o esquema de treino-teste da Figura 18, e usando a média de suas previsões como previsão final, o que será chamado de modelo composto. Para fins de comparação, mesmo que o experimento numérico com o modelo simples pudesse ser usado para prever para todos os exemplos do segundo poço, o teste será feito com os conjuntos de teste utilizado para os outros experimentos numéricos, de modo a coincidir com o experimento com o modelo composto.

3.4.3 Representação de Resultados

Os resultados serão apresentados em formato de diagrama de caixa, ou *box plot*. Os diagramas de caixa são uma maneira simplificada de representar a distribuição dos dados em uma amostra.

A Figura 19 apresenta um exemplo de diagrama de caixa. Nele, os limites inferior (LI) e superior (LS) representam, respectivamente, o 1^{o} e 3^{o} quartis dos dados (25% e 75%). A construção dos *whiskers* inferior (WI) e superior (WS) é feita a partir de LI e LS, de modo que os valores de WI e WS são calculados pelas fórmulas nas equações 3.1

Figura 19 – Diagrama de Caixa



e 3.2. Quaisquer dados que não se encontram entre os *whiskers* inferior e superior são considerados como *outliers* e são representados separadamente.

$$\mathrm{WI} = \mathrm{LI} - 1,5(\mathrm{LS} - \mathrm{LI}) \tag{3.1}$$

$$\mathrm{WS} = \mathrm{LS} + 1,5(\mathrm{LS} - \mathrm{LI}) \tag{3.2}$$

Em alguns casos nos resultados, os *outliers* serão suprimidos, não sendo representados. Isto foi feito para evitar que valores muito distantes dos outros atrapalhassem na análise dos dados.

Para conferir informações mais completas sobre sobre os resultados, serão apresentados, também, os mínimos, máximos e médias de cada modelo para cada poço.

4 RESULTADOS E DISCUSSÕES

Este capítulo apresenta os resultados e discussões sobre os experimentos numéricos delimitados no capítulo anterior, realizados com regressão linear, árvores de decisão e floresta aleatória, variando seus hiper-parâmetros nos 4 poços *offshore* considerados. Em seguida, os melhores modelos serão utilizados para calcular a importância das características por poço. Por fim, avaliou-se a possibilidade de aprendizado por transferência entre poços usando os poços 1 e 2.

4.1 EXPERIMENTOS NUMÉRICOS DO POÇO 1

As Figuras 20, 22 e 24 mostram os diagramas de caixa dos erros normalizados para, respectivamente, os modelos de regressão linear, árvore de decisão e floresta aleatória, enquanto as Figuras 21, 23, 25 mostram os diagramas de caixa dos valores de R^2 para os mesmos modelos. Nos rótulos dos gráficos de floresta aleatória, o primeiro número refere-se à profundidade máxima das árvores, enquanto o segundo número refere-se ao número de árvores utilizados na floresta.

Como os *outliers* dos diagramas foram suprimidos para ser possível comparar as caixas nos gráficos, as informações principais neles serão a mediana e o primeiro e terceiro quartis. Para uma visão completa dos dados, é necessário observar também as Tabelas 8, 9 e 10, que contêm a média, desvio padrão, máximo e mínimo para cada modelo.

Observando os gráficos de erro normalizado e R^2 da regressão linear (Figuras 20 e 21) e a Tabela 8, e comparando-os com os outros modelos, nota-se que a regressão linear apresenta resultados consideravelmente piores que todos os outros modelos tanto em erro normalizado quanto em R^2.

Figura 20 – Erro normalizado para Regressão Linear sem *outliers* — Poço 1

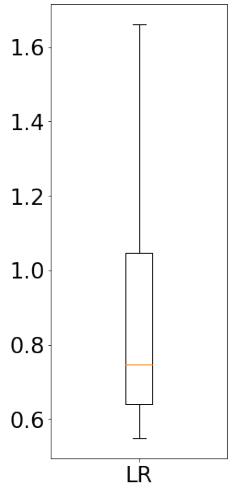


Figura 21 – R^2 para Regressão Linear sem *outliers* — Poço 1

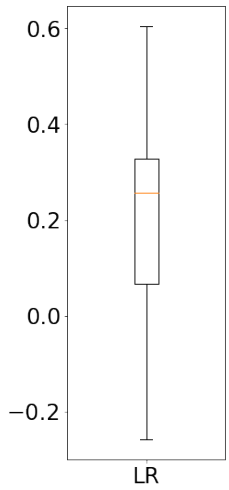


Tabela 8 – Resultados da Regressão Linear para diferentes hiper-parâmetros — poço 1

	Erro Normalizado				**R^2**			
	méd	dev	mín	máx	méd	dev	mín	máx
Regressão Linear	1,21	1,57	0,55	12,42	-0,33	2,18	-1,41	0,60

Analisando os diagramas de caixa das árvores de decisão para o erro normalizado na Figura 22 e a parte correspondente na Tabela 9, os menores erros ocorrem para a árvore de decisão com mínimo de amostras por folha igual a 1/200 ou 0,5% do tamanho do conjunto de treino. Os erros caem conforme aumenta-se este hiper-parâmetro, e depois voltam a crescer quando ele é igual a 1/100 ou 1%. Um processo inverso ocorre com o R^2 na Figura 23 e na Tabela 9, mas para esta métrica, quanto maior o valor, melhor. Assim, a melhor escolha do hiper-parâmetro mínimo de amostras por folha para a árvore de decisão no poço 1 é 0,5%.

Figura 22 – Erro normalizado para Árvores de Decisão sem *outliers* — Poço 1

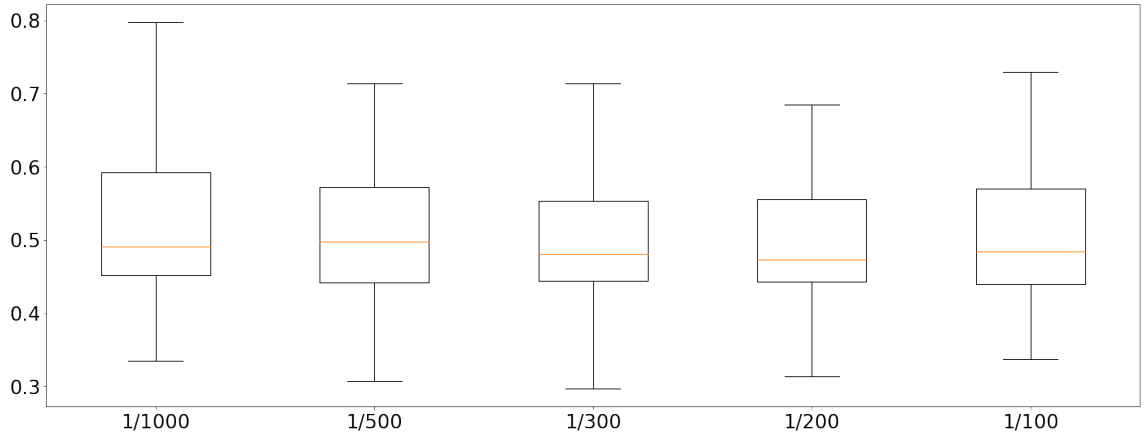


Figura 23 – R^2 para Árvores de Decisão sem *outliers* — Poço 1

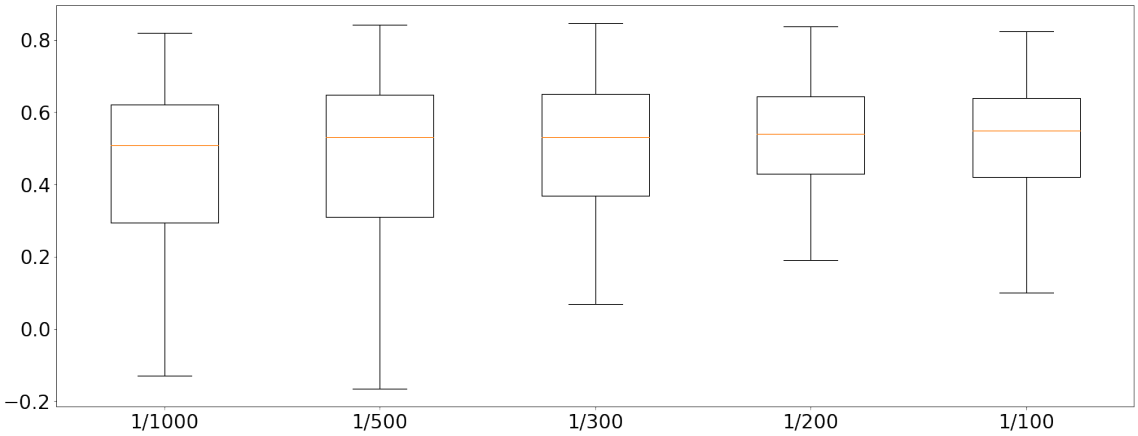
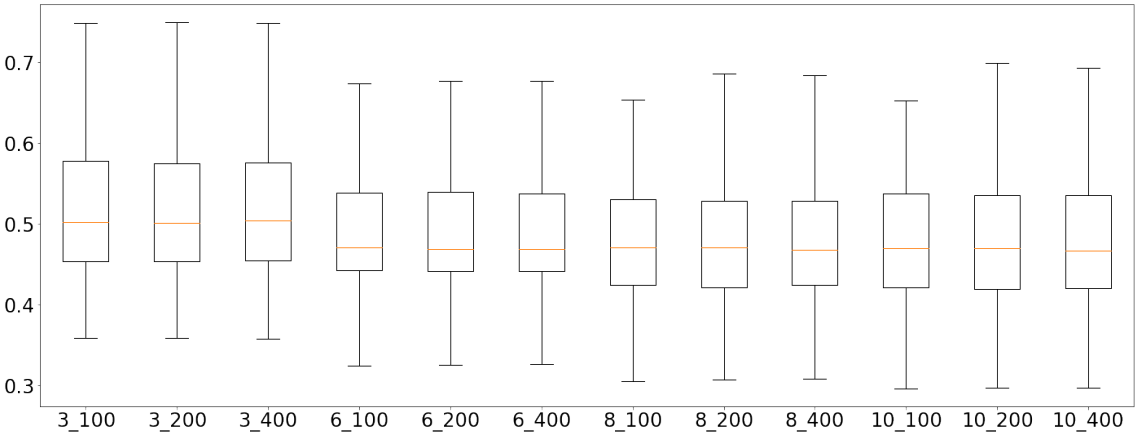


Tabela 9 – Resultados da árvore de decisão para diferentes hiper-parâmetros — poço 1

	Erro Normalizado				**R^2**			
Mín. Amostras/folha	méd	dev	mín	máx	méd	dev	mín	máx
0,1%	0,544	0,192	0,335	1,944	0,430	0,274	-0,555	0,819
0,2%	0,534	0,216	0,308	2,153	0,463	0,262	-0,766	0,843
0,33%	0,520	0,186	0,298	1,889	0,492	0,240	-0,733	0,846
0,5%	0,515	0,173	0,314	1,734	0,500	0,257	-1,180	0,837
1%	0,538	0,215	0,337	2,022	0,488	0,270	-1,319	0,823

Para as florestas aleatórias, a análise dos diagramas de caixa de erro normalizado da Figura 24 parece indicar que os modelos com profundidade máxima igual 8 ou 10 têm melhor resultado, sem diferenças perceptíveis entre eles ou para diferentes números de estimadores. A análise dos diagramas em 25 mostram o mesmo comportamento para o R^2. Analisando a Tabela 10, observa-se que há uma pequena melhora com o aumento de estimadores e da profundidade máxima, mas ela não é muito expressiva.

Figura 24 – Erro normalizado para Florestas Aleatórias sem *outliers* — Poço 1



Comparando os melhores modelos de árvore de decisão e de floresta aleatória, observa-

Figura 25 – R^2 para Florestas Aleatórias sem *outliers* — Poço 1

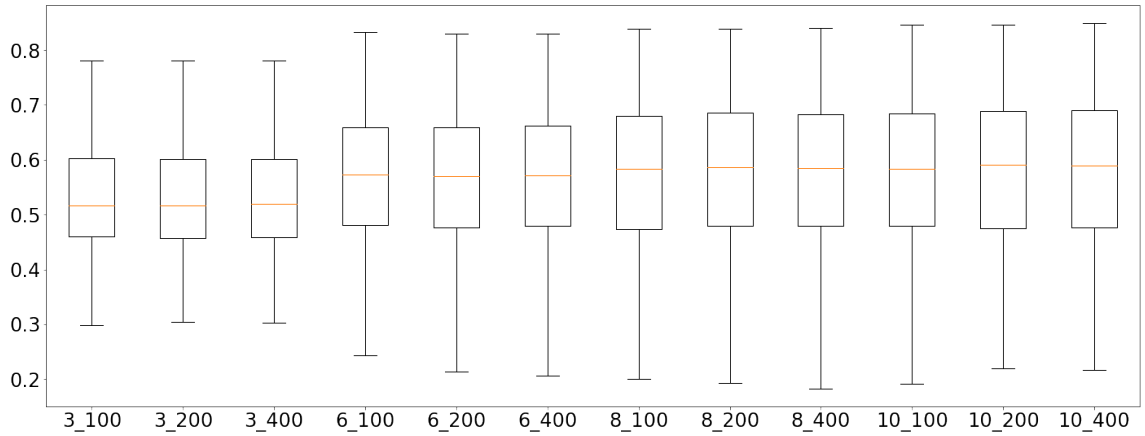


Tabela 10 – Resultados da floresta aleatória para diferentes hiper-parâmetros — poço 1

		Erro Normalizado				**R^2**			
Prof.	**# Arv.**	méd	dev	mín	máx	méd	dev	mín	máx
	100	0.587	0.342	0.359	3.142	0.494	0.244	-1.279	0.781
	200	0.587	0.342	0.358	3.138	0.493	0.248	-1.316	0.781
3	**400**	0.587	0.342	0.358	3.135	0.493	0.247	-1.295	0.781
	100	0.518	0.203	0.324	1.975	0.539	0.227	-1.066	0.833
	200	0.517	0.201	0.326	1.948	0.539	0.226	-1.042	0.830
6	**400**	0.517	0.201	0.326	1.953	0.539	0.224	-1.016	0.830
	100	0.506	0.190	0.305	1.906	0.551	0.203	-0.680	0.839
	200	0.505	0.189	0.308	1.882	0.551	0.204	-0.647	0.839
8	**400**	0.504	0.187	0.308	1.862	0.553	0.201	-0.601	0.840
	100	0.505	0.184	0.296	1.841	0.551	0.196	-0.351	0.846
	200	0.503	0.183	0.297	1.841	0.552	0.201	-0.340	0.846
10	**400**	0.502	0.182	0.297	1.830	0.553	0.200	-0.329	0.849

se que a floresta aleatória apresenta R^2 e erros normalizados melhores que a árvore de decisão.

4.2 EXPERIMENTOS NUMÉRICOS DO POÇO 2

Observando os resultados da regressão linear para o poço 2 nos gráficos das Figuras 26 e 27 e na Tabela 11, novamente nota-se que o modelo é o pior entre os considerados.

Tabela 11 – Resultados da Regressão Linear para diferentes hiper-parâmetros — poço 2

	Erro Normalizado				**R^2**			
	méd	dev	mín	máx	méd	dev	mín	máx
Regressão Linear	0,94	0,74	0,40	5,83	-1,83	18,4	-179	0,50

Para as árvores de decisão, a Figura 28 mostra uma tendência de decrescimento geral do erro normalizado com o aumento do mínimo de amostras por folha, atingindo seu

Figura 26 – Erro normalizado para Regressão Linear sem *outliers* — Poço 2

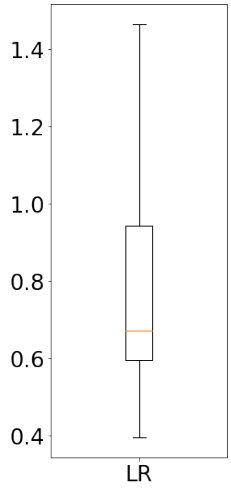
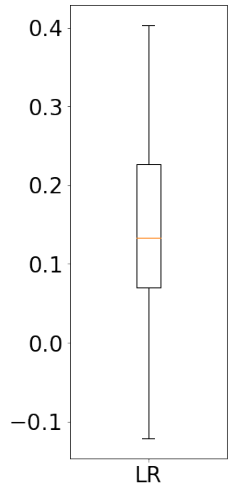


Figura 27 – R^2 para Regressão Linear sem *outliers* — Poço 2



melhor desempenho em 1/100. No gráfico da Figura 29, observa-se uma tendência de aumento do primeiro e terceiro quartis com o aumento do mínimo de amostras por folha, mas há uma queda na mediana de 1/200 e 1/100. Analisando a Tabela 12, confirma-se essa tendência, porém é possível notar pela análise do máximo do erro normalizado e mínimo do R^2, além dos desvios padrões para ambas métricas, há um aumento da dispersão dos valores dessas métricas quando aumenta-se o hiper-parâmetro de 1/200 para 1/100, mesmo que esta diferença não seja muito significativa para o R^2. Assim, apesar de o melhor resultado geral ser melhor para o valor de hiper-parâmetro igual a 1/100, pode-se justificar a escolha pelo 1/200.

No gráfico de erro normalizado das florestas aleatórias na Figura 30, os valores do primeiro quartil (representado pelo limite inferior da caixa) dos modelos é visualmente indistinguível, mas os valores da mediana e terceiro quartil (representado pelo limite su-

Figura 28 – Erro normalizado para Árvores de Decisão sem *outliers* — Poço 2

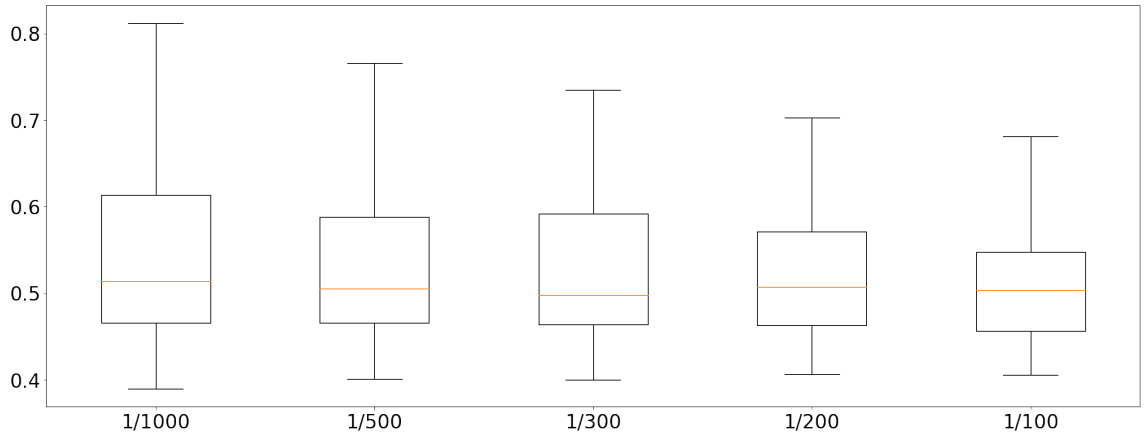


Figura 29 – R^2 para Árvores de Decisão sem *outliers* — Poço 2

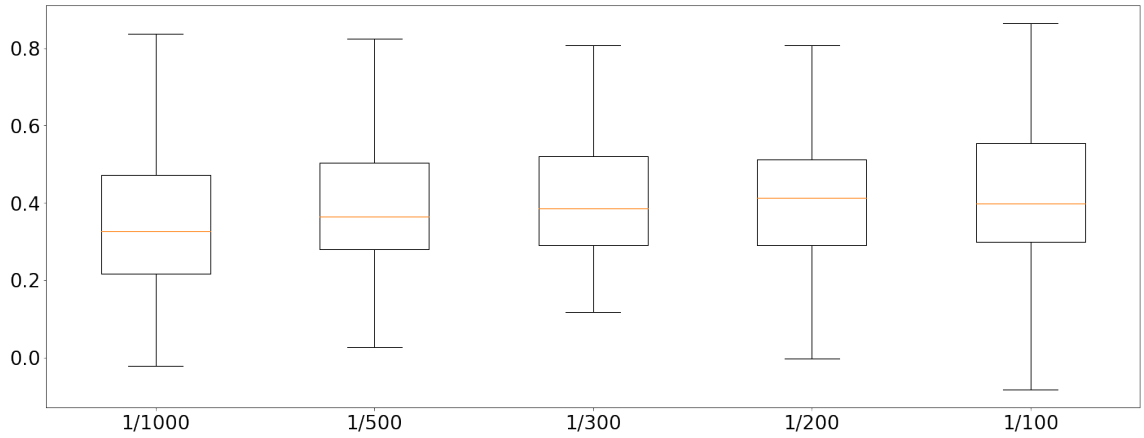


Tabela 12 – Resultados da árvore de decisão para diferentes hiper-parâmetros — poço 2

	Erro Normalizado				**R^2**			
Mín. Amostras/folha	méd	dev	mín	máx	méd	dev	mín	máx
0,1%	0,560	0,134	0,390	1,316	0,312	0,324	-1,528	0,838
0,2%	0,540	0,107	0,401	0,997	0,364	0,263	-0,954	0,900
0,33%	0,529	0,096	0,400	0,948	0,396	0,222	-0,609	0,866
0,5%	0,525	0,095	0,407	0,978	0,406	0,194	-0,229	0,856
1%	0,524	0,115	0,406	1,296	0,417	0,196	-0,243	0,864

perior da caixa) dos modelos com profundidade máxima igual a 6 são menores. Os valores para o R^2, na Figura 31, são visualmente indistinguíveis, não aparentando haver diferenças significativas entre os modelos. Em análise dos valores da Tabela 13, confirma-se que os modelos com profundidade máxima 6 têm menor erro normalizado, e enquanto apresentam menor R^2 que os modelos com profundidade máxima 3, a diferença não é significativa, o que não pode ser dito sobre a diferença entre os erros normalizados dos modelos. A variação no número de árvores utilizados para estimar não altera significativamente o desempenho do modelo.

Figura 30 – Erro normalizado para Florestas Aleatórias sem *outliers* — Poço 2

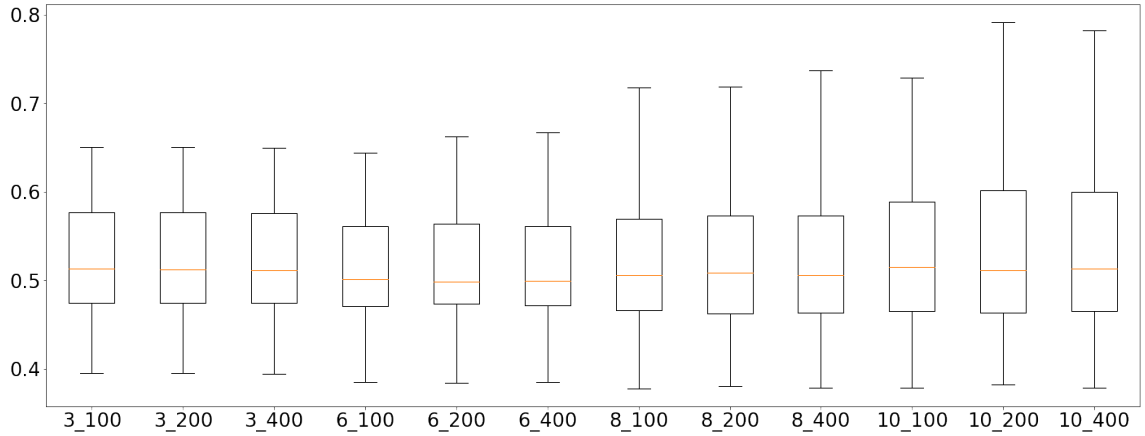
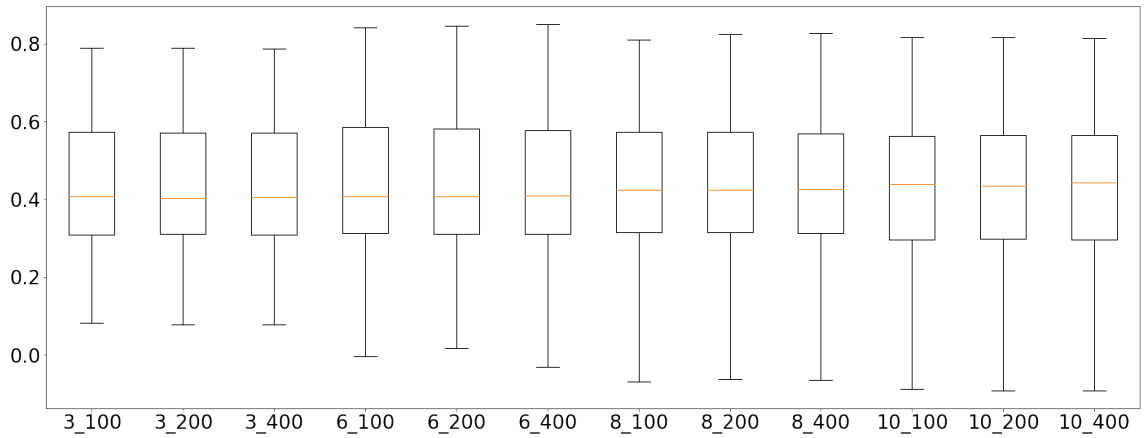


Figura 31 – R^2 para Florestas Aleatórias sem *outliers* — Poço 2



Comparando-se os resultados dos melhores modelos de árvore de decisão e floresta aleatória, o desempenho dos modelos é extremamente similar, com um resultado levemente superior para a árvore de decisão.

Tabela 13 – Resultados da floresta aleatória para diferentes hiper-parâmetros — poço 2

		Erro Normalizado				**R^2**			
Prof.	**# Arv.**	méd	dev	mín	máx	méd	dev	mín	máx
	100	0.546	0.140	0.395	1.452	0.426	0.181	-0.197	0.788
	200	0.546	0.141	0.395	1.452	0.426	0.181	-0.206	0.788
3	**400**	0.546	0.140	0.394	1.447	0.427	0.180	-0.187	0.787
	100	0.528	0.109	0.385	1.213	0.424	0.219	-0.315	0.840
	200	0.528	0.107	0.384	1.192	0.424	0.216	-0.249	0.845
6	**400**	0.529	0.110	0.385	1.219	0.423	0.218	-0.255	0.849
	100	0.538	0.128	0.378	1.355	0.409	0.248	-0.469	0.809
	200	0.538	0.129	0.380	1.367	0.408	0.250	-0.470	0.825
8	**400**	0.538	0.129	0.378	1.356	0.406	0.256	-0.466	0.826
	100	0.551	0.157	0.379	1.650	0.386	0.289	-0.616	0.816
	200	0.552	0.163	0.382	1.704	0.385	0.295	-0.645	0.816
10	**400**	0.552	0.161	0.378	1.672	0.383	0.298	-0.751	0.815

4.3 EXPERIMENTOS NUMÉRICOS DO POÇO 3

Como ocorreu nos outros poços, os resultados da regressão linear apresentados nas Figuras 32 e 33 e especialmente na Tabela 14 são consideravelmente piores que todos os outros modelos.

Figura 32 – Erro normalizado para Regressão Linear sem *outliers* — Poço 3

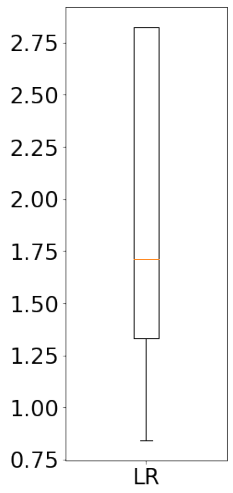
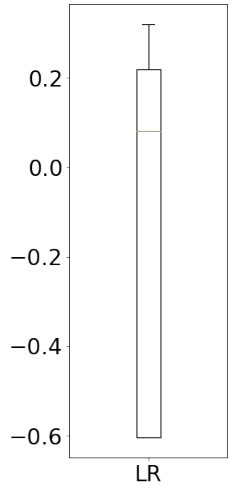


Tabela 14 – Resultados da Regressão Linear para diferentes hiper-parâmetros — poço 3

	Erro Normalizado				**R^2**			
	méd	dev	mín	máx	méd	dev	mín	máx
Regressão Linear	$2 \cdot 10^{12}$	$5 \cdot 10^{12}$	0,84	$2 \cdot 10^{12}$	$-4 \cdot 10^{24}$	$1 \cdot 10^{25}$	$-3 \cdot 10^{25}$	0,32

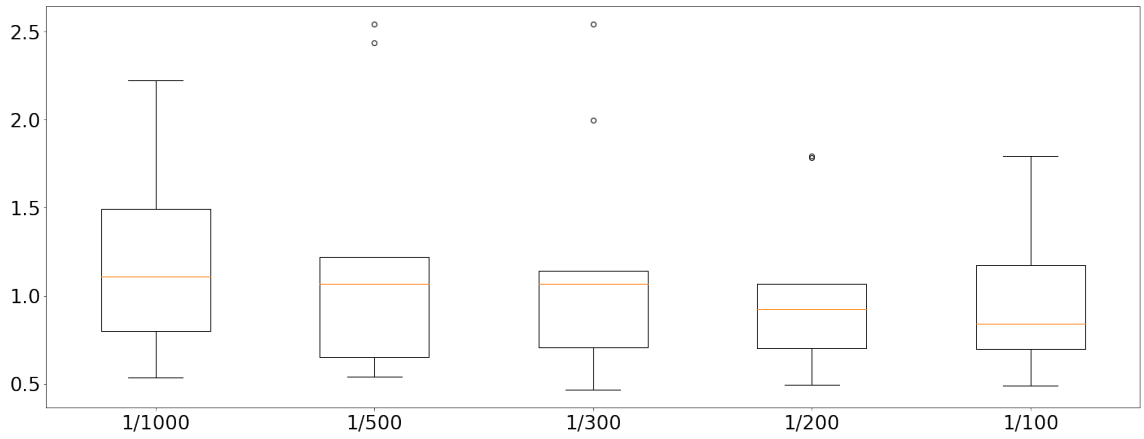
Para os erros normalizados das árvores de decisão, mostrados no gráfico da Figura 34, observa-se rapidamente que os modelos com 1/200 e 1/100 de mínimo de amostras por

Figura 33 – R^2 para Regressão Linear sem *outliers* — Poço 3



folha estão entre os melhores: têm valores máximos semelhantes entre si e menores que os outros modelos, o modelo de 1/200 amostras apresenta o menor valor de terceiro quartil, e segunda menor mediana, enquanto o modelo de 1/100 têm terceiro quartil maior que o de 1/200, sua mediana é significativamente mais baixa. Comparando em termos de R^2 com a Figura 35, o modelo com 1/100 aparenta ter melhor resultados, mas ao analisar as estatísticas na Tabela 15, observa-se que o valor mínimo para R^2 é consideravelmente mais baixo no modelo de 1/100 quando comparado ao 1/200 reduzindo o valor médio do R^2. Após uma análise completa, pode-se afirmar que o modelo de 1/100 apresenta um resultado geral melhor, com a média e desvio padrão dos erros normalizados menores, e com primeiro, segundo (mediana) e terceiro quartis maiores que 1/200.

Figura 34 – Erro normalizado para Árvores de Decisão — Poço 3



Para florestas aleatórias, o gráfico da Figura 36 revela um melhor desempenho em termos de erro normalizado para os modelos com profundidade máxima 3, enquanto o gráfico da Figura 37 revela comportamentos variados da R^2 entre os modelos, sem uma

Figura 35 – R^2 para Árvores de Decisão sem *outliers* — Poço 3

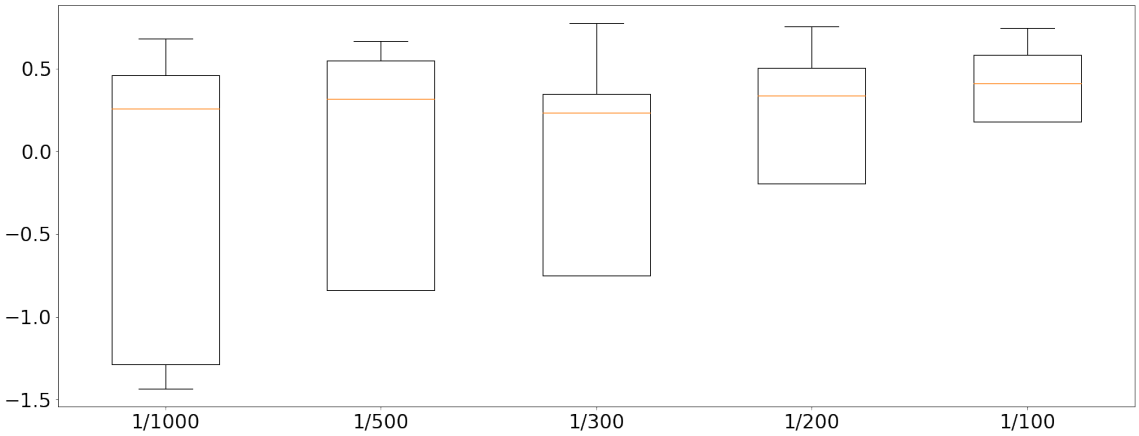
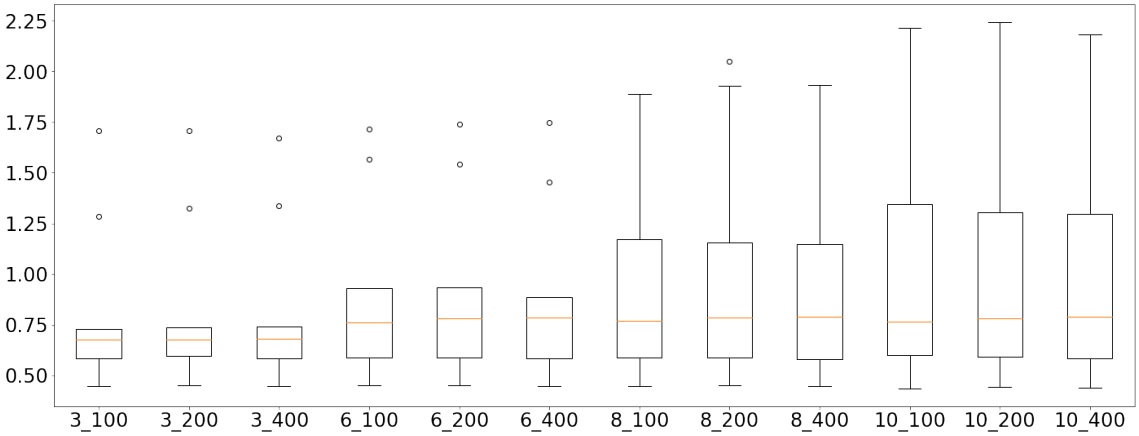


Tabela 15 – Resultados da árvore de decisão para diferentes hiper-parâmetros — poço 3

	Erro Normalizado				**R^2**			
Mín. Amostras/folha	méd	dev	mín	máx	méd	dev	mín	máx
0,1%	1,171	0,524	0,537	2,221	-0,793	2,249	-6,414	0,680
0,2%	1,246	0,746	0,541	2,541	-0,879	2,364	-6,365	0,667
0,33%	1,182	0,675	0,468	2,541	-1,353	3,565	-10,351	0,772
0,5%	1,014	0,484	0,494	1,793	-0,369	1,779	-4.830	0,753
1%	0,960	0,437	0,490	1,792	-0,450	2,333	-6,524	0,742

Figura 36 – Erro normalizado para Florestas Aleatórias — Poço 3



tendência ou melhores modelos claros. Pela Tabela 16, confirma-se que o erro normalizado é menor para a profundidade máxima 3, mesmo que os máximos e mínimos entre os modelos sejam semelhantes. Quanto ao R^2, observa-se que, em média, os modelos de profundidade máxima 3 são melhores. Em ambos os casos, não há diferenças significativas quando varia-se o número de árvores da floresta.

Comparando os melhores modelos de árvore de decisão e floresta aleatória, os modelos de floresta aleatória resultam em erro normalizado e R^2 significativamente melhores que da árvore de decisão.

Figura 37 – R^2 para Florestas Aleatórias sem *outliers* — Poço 3

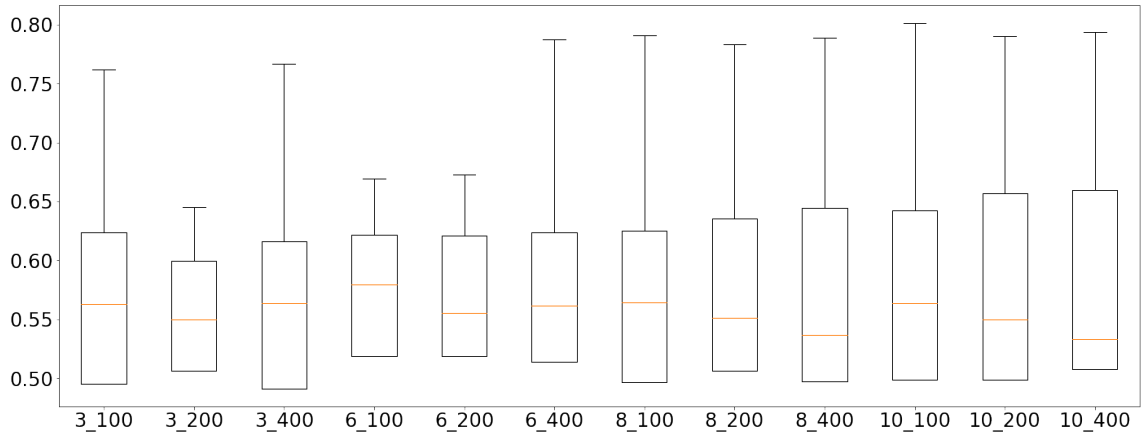


Tabela 16 – Resultados da floresta aleatória para diferentes hiper-parâmetros — poço 3

		Erro Normalizado				**R^2**			
Prof.	**# Arv.**	méd	dev	mín	máx	méd	dev	mín	máx
	100	0.810	0.412	0.449	1.706	-0.278	2.382	-6.592	0.762
	200	0.819	0.417	0.450	1.708	-0.291	2.396	-6.639	0.762
3	**400**	0.812	0.411	0.447	1.672	-0.269	2.327	-6.429	0.767
	100	0.892	0.452	0.452	1.714	-0.319	2.399	-6.633	0.785
	200	0.896	0.452	0.453	1.738	-0.342	2.445	-6.774	0.782
6	**400**	0.882	0.438	0.447	1.746	-0.375	2.537	-7.052	0.787
	100	0.982	0.551	0.448	1.887	-0.377	2.527	-7.026	0.791
	200	1.009	0.596	0.450	2.050	-0.434	2.672	-7.466	0.783
8	**400**	0.992	0.569	0.446	1.934	-0.467	2.770	-7.762	0.788
	100	1.051	0.648	0.437	2.215	-0.431	2.691	-7.511	0.801
	200	1.059	0.661	0.443	2.241	-0.504	2.874	-8.071	0.790
10	**400**	1.051	0.646	0.439	2.182	-0.527	2.934	-8.256	0.794

4.4 EXPERIMENTOS NUMÉRICOS DO POÇO 4

Novamente, os resultados da regressão linear apresentados nas Figuras 38 e 39 e especialmente na Tabela 17 são consideravelmente piores que todos os outros modelos.

Tabela 17 – Resultados da Regressão Linear para diferentes hiper-parâmetros — poço 4

	Erro Normalizado				**R^2**			
	méd	dev	mín	máx	méd	dev	mín	máx
Regressão Linear	23,2	48,3	1,00	110	-2393	5350	-11964	0,18

Entre as árvores de decisão, a árvore com tamanho de amostra mínimo por folha igual a 1/100 apresentou os melhores resultados de erro normalizado e R^2, como pode ser visto nas Figuras 40 e 41 e na Tabela 18, de maneira bastante clara.

Para as florestas aleatórias, os gráficos das Figuras 42 e 43 e a Tabela 19 apontam que os melhores modelos são aqueles com profundidade máxima igual a 3, tanto em R^2 quanto

Figura 38 – Erro normalizado para Regressão Linear sem *outliers* — Poço 4

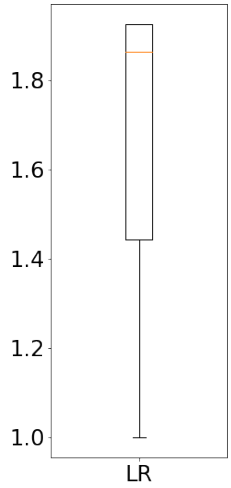
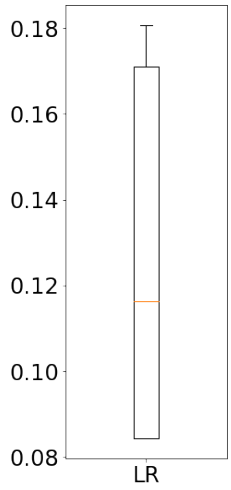


Figura 39 – R^2 para Regressão Linear sem *outliers* — Poço 4



em erro normalizado. Além disso, o número de árvores não influenciou significativamente no desempenho dos modelos.

Comparando os melhores modelos de árvore de decisão e floresta aleatória para o poço, os modelos de floresta aleatória resultam em erros normalizados e valores de R^2 melhores que os de árvore de decisão.

4.5 IMPORTÂNCIA DAS CARACTERÍSTICAS

A importância das características foi calculada para os melhores modelos para cada poço, avaliando o quanto cada uma delas é responsável por queda no erro cometido pelo modelo. Os resultados estão representados nas Figuras 44, 45, 46 e 47. Para todos os casos, a proximidade da broca é a característica que mais causa redução no erro cometido pelos modelos. Para os poços 1 e 2, a profundidade também teve importância significativa,

Figura 40 – Erro normalizado para Árvores de Decisão — Poço 4

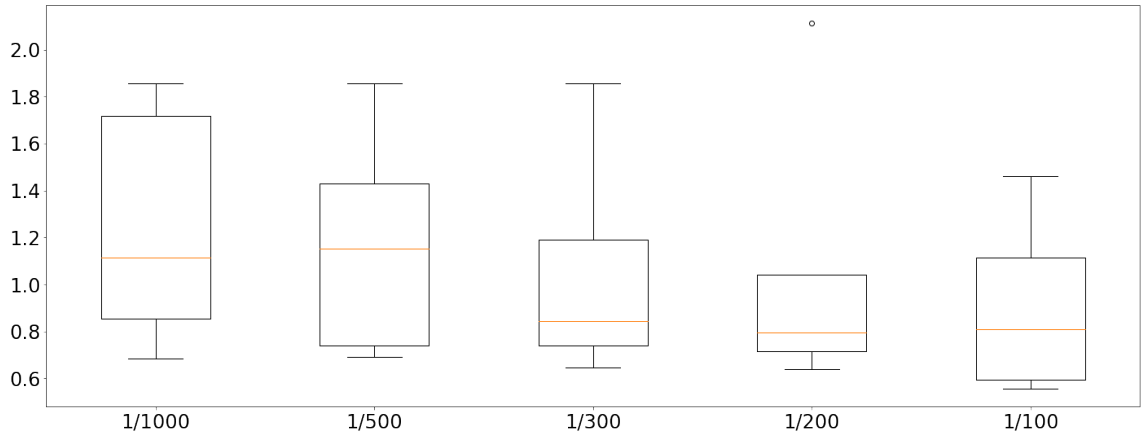


Figura 41 – R^2 para Árvores de Decisão — Poço 4

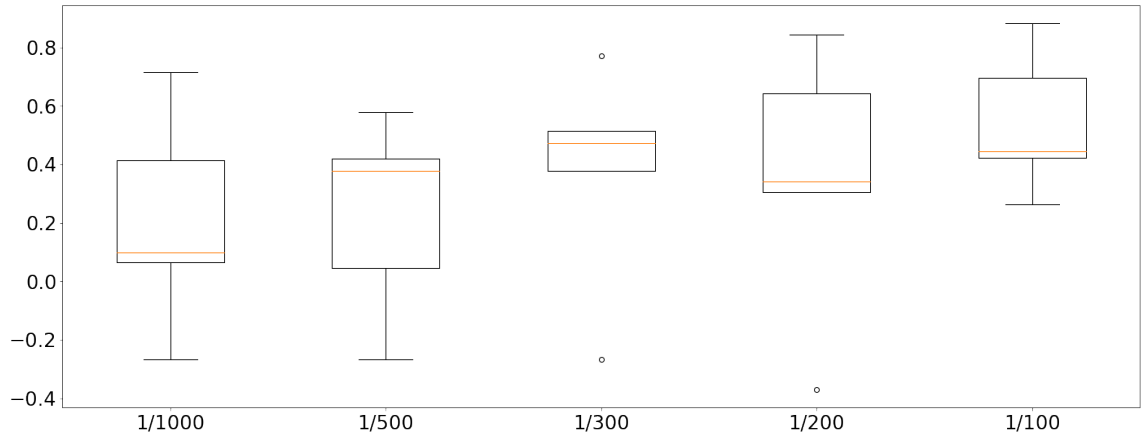


Tabela 18 – Resultados da árvore de decisão para diferentes hiper-parâmetros — poço 4

	Erro Normalizado				**R^2**			
Mín. Amostras/folha	méd	dev	mín	máx	méd	dev	mín	máx
0,1%	1,245	0,52	0,684	1,855	0,205	0,374	-0,268	0,715
0,2%	1,174	0,488	0,692	1,855	0,231	0,340	-0,268	0,580
0,33%	1,056	0,492	0,648	1,855	0,373	0,387	-0,268	0,770
0,5%	1,062	0,606	0,641	2,111	0,352	0,461	-0,370	0,843
1%	0,907	0,381	0,558	1,462	0,541	0,246	0,262	0,882

Figura 42 – Erro normalizado para Florestas Aleatórias — Poço 4

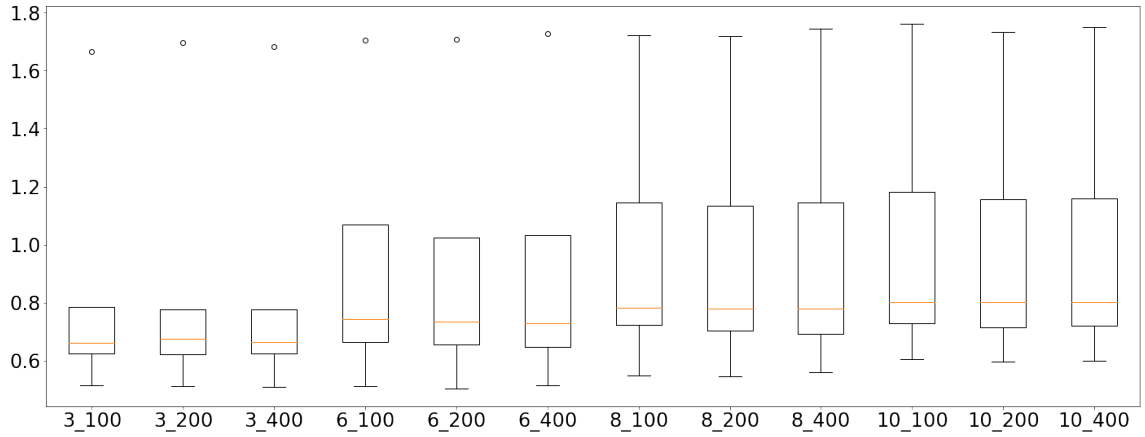


Figura 43 – R^2 para Florestas Aleatórias — Poço 4

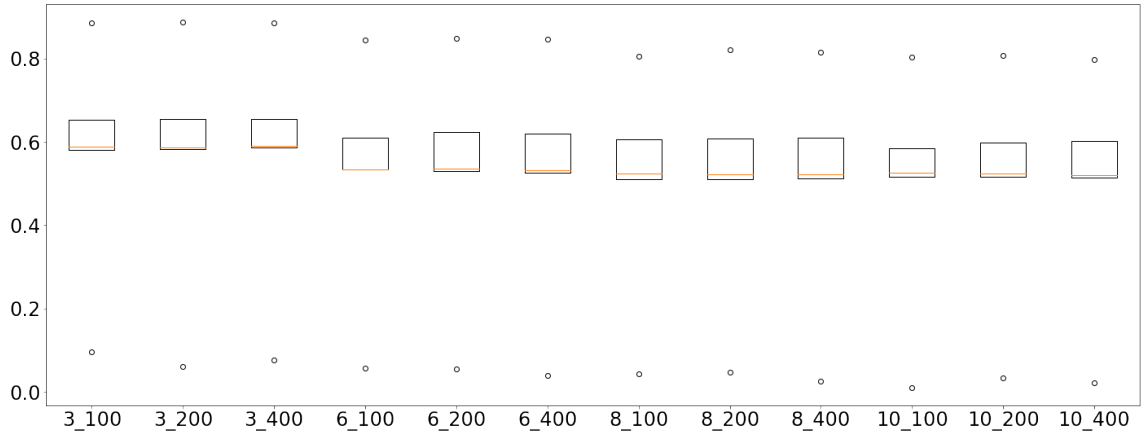


Tabela 19 – Resultados da floresta aleatória para diferentes hiper-parâmetros — poço 4

		Erro Normalizado				**R^2**			
Prof.	**# Arv.**	méd	dev	mín	máx	méd	dev	mín	máx
	100	0.850	0.466	0.514	1.665	0.561	0.288	0.095	0.886
	200	0.856	0.479	0.512	1.697	0.555	0.303	0.060	0.887
3	**400**	0.851	0.474	0.511	1.681	0.559	0.296	0.077	0.886
	100	0.939	0.475	0.512	1.707	0.516	0.286	0.058	0.844
	200	0.926	0.476	0.504	1.708	0.518	0.290	0.055	0.849
6	**400**	0.930	0.484	0.514	1.727	0.513	0.295	0.040	0.846
	100	0.985	0.466	0.549	1.722	0.497	0.280	0.042	0.805
	200	0.976	0.468	0.545	1.719	0.502	0.283	0.047	0.821
8	**400**	0.984	0.477	0.559	1.744	0.497	0.290	0.025	0.815
	100	1.015	0.469	0.606	1.760	0.488	0.291	0.011	0.804
	200	1.001	0.460	0.597	1.735	0.496	0.284	0.033	0.807
10	**400**	1.006	0.465	0.600	1.749	0.491	0.287	0.021	0.798

Figura 44 – Importância das características — Poço 1

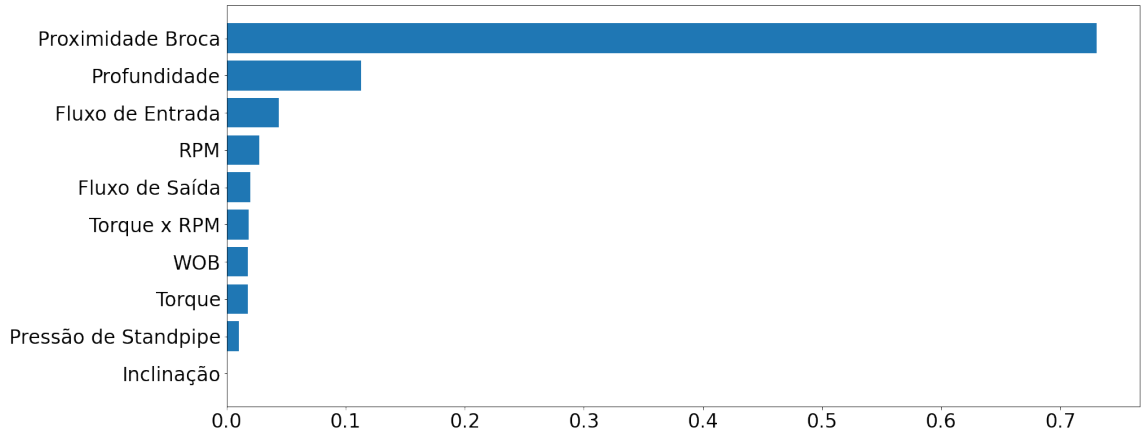
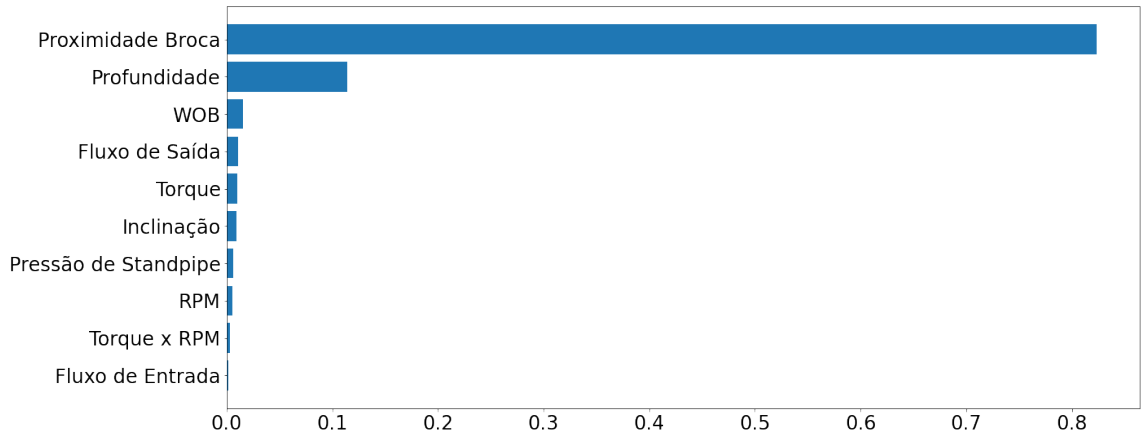


Figura 45 – Importância das características — Poço 2



mas as outras características apresentaram perto de nenhum efeito na previsão da ROP. No caso dos poços 3 e 4, nenhuma outra característica teve importância significativa.

A importância maior dada à profundidade para os poços 1 e 2 em comparação com os poços 3 e 4 pode ser explicada pela maior amplitude de profundidade deles, como mostrado na Tabela 1. O mesmo pode ser observado para a inclinação no poço 2, que varia até um valor de 14,52º, significativamente maior que as inclinações dois outros poços, como observado na Tabela 4.

4.6 APRENDIZADO POR TRANSFERÊNCIA

Como definido na subseção 3.4.2, para avaliar a possibilidade de aprendizado por transferência, serão utilizados os poços 1 e 2, com o melhor modelo. Como os melhores modelos encontrados para os poços 1 e 2 não foram comuns entre eles, escolheu-se o modelo de floresta aleatória para avaliar o aprendizado por transferência por ter tido o melhor desempenho geral, com a combinação de hiper-parâmetros de profundidade máxima da

Figura 46 – Importância das características — Poço 3

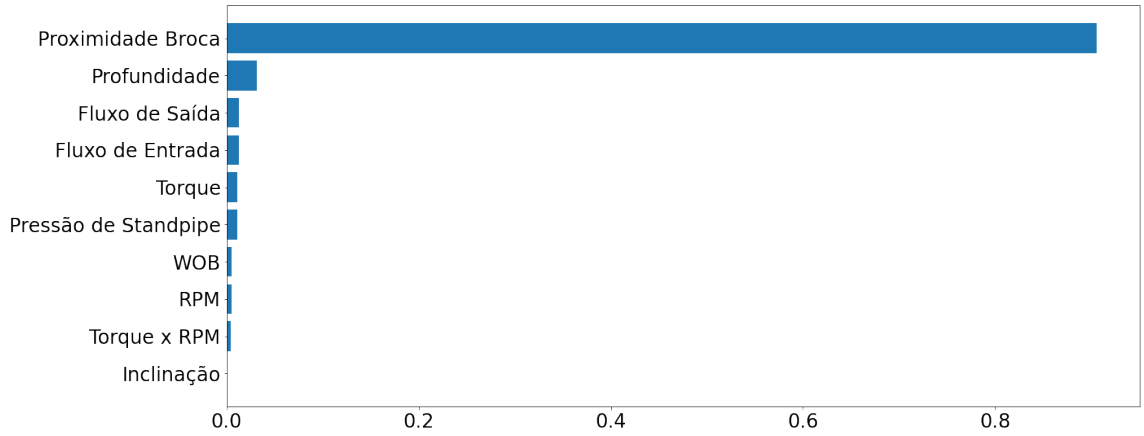
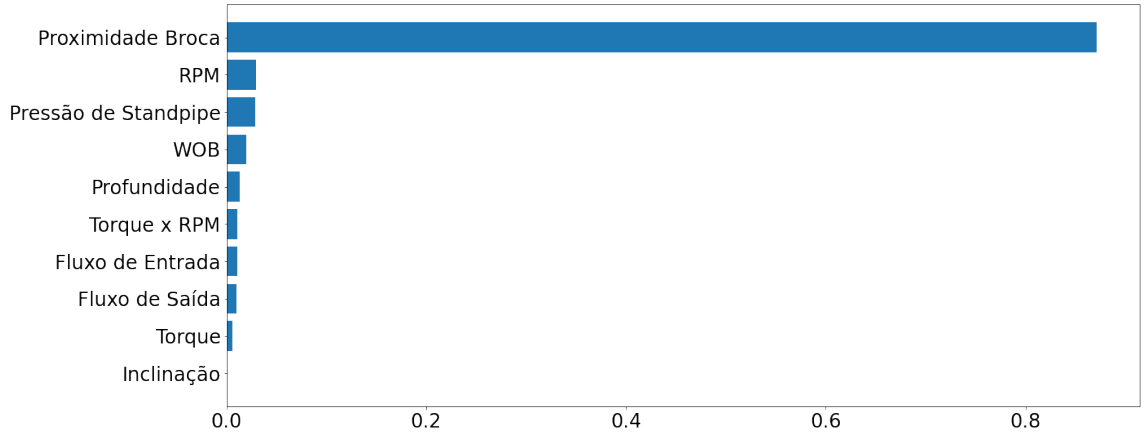


Figura 47 – Importância das características — Poço 4



árvore igual a 8 (intermediário entre os melhores resultados obtidos para cada poço) e 100 árvores.

Foram utilizados dois modelos para avaliar o aprendizado por transferência, o simples e o composto, definidos na subseção 3.4.2. O modelo simples corresponde à aplicação direta do modelo treinado em todo o primeiro poço para a previsão no segundo poço. Já o modelo composto é aquele que utiliza a média entre as previsões do modelo treinado em todo o primeiro poço e do modelo treinado com os dados de profundidades menores para o segundo poço para a previsão no segundo poço.

Tabela 20 – Resultados para o aprendizado por transferência

		Erro Normalizado				**R^2**			
Poço Teste	**Modelo**	méd	dev	mín	máx	méd	dev	mín	máx
	Simples	0.594	0.318	0.385	2.599	0.451	0.329	-1.330	0.804
2 ⟶ 1	**Composto**	0.606	0.363	0.410	3.484	0.358	0.204	-1.156	0.616
	Simples	0.693	0.247	0.421	2.334	0.232	0.391	-1.071	0.753
1 ⟶ 2	**Composto**	0.611	0.201	0.389	2.127	0.359	0.274	-0.556	0.781

Figura 48 – Erro normalizado na transferência para previsão no poço 1

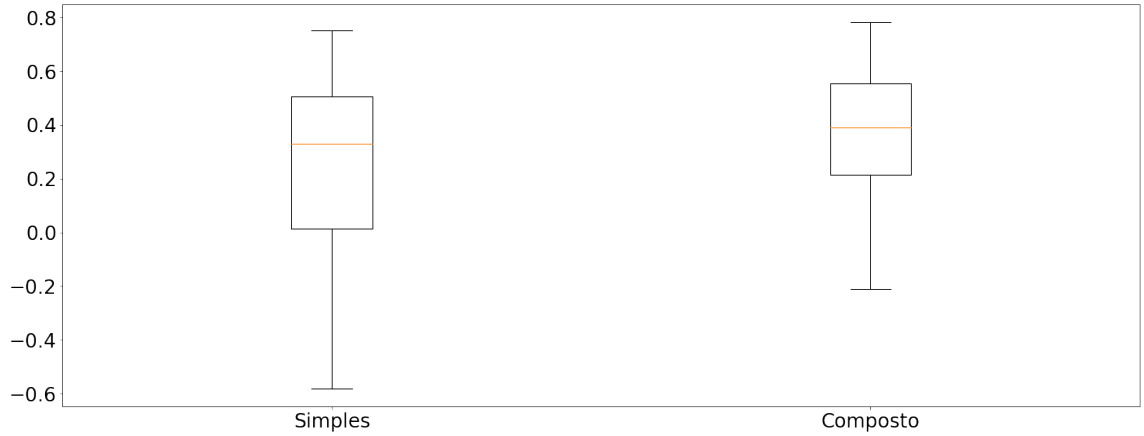
Figura 49 – R^2 na transferência para previsão no poço 1

Observando os gráficos de erro normalizado e R^2 para a previsão no poço 1 nas Figuras 48 e 49 e na Tabela 20, observa-se que o modelo composto tem pior performance que o modelo simples, especialmente em termos de R^2. Enquanto isto, para previsão do poço 2, observado na tabela anterior e nas Figuras 50 e 51, o desempenho foi significativamente melhor para o modelo composto.

Além disso, comparando-se os resultados obtidos nos experimentos numéricos de transferência de aprendizado com os resultados os experimentos anteriores, verifica-se que os resultados de transferência foram todos inferiores, mostrando uma transferência negativa de aprendizado. Para a previsão no poço 1, o melhor resultado médio entre os experimentos numéricos originais foram erro normalizado de 0,502 e R^2 de 0,553, comparados aos 0,594 de erro normalizado e 0,451 de R^2 do aprendizado por transferência. Já para a previsão no poço 2, o melhor resultado médio entre os experimentos numéricos originais foi erro normalizado de 0,524 e R^2 de 0,417, enquanto os resultados de aprendizado por transferência foram 0,611 de erro normalizado e 0,359 de R^2.

Figura 50 – Erro normalizado na transferência para previsão no poço 2

Figura 51 – R^2 na transferência para previsão no poço 2



4.7 DISCUSSÕES

Observa-se, pelos resultados gerais para os poços, que não foi encontrado um melhor conjunto de hiper-parâmetros válido para todos os poços, o que era esperado pelas diferenças entre as características dos poços. Entretanto, de modo geral, os modelos de floresta aleatória obtiveram os melhores resultados e, quando isso não ocorreu, como no caso do poço 2, a diferença entre o desempenho da árvore de decisão e da floresta aleatória não foi significativa.

Uma análise comparativa entre os resultados dos modelos nos diferentes poços revela que tanto os erros normalizados quanto o R^2 obtido foram melhores nos poços 1 e 2 que nos poços 3 e 4. Isto pode ser explicado pela amostra de dados disponível para cada um dos poços: os poços 1 e 2 possuíam dados para uma extensão de profundidade múltiplas vezes maior que a extensão de profundidade dos poços 3 e 4. Ao agregar os dados em profundidade como explicado na subseção 3.3.4, isso resultou em um menor número de exemplos disponível para o treino dos modelos, o que comumente resulta em desempenho

pior. Por fim, os dados disponíveis para poços 3 e 4 eram todos da parte final do poço, que comumente é mais imprevisível e difícil de se perfurar.

Os melhores modelos para os poços 3 e 4 tanto de árvore de decisão quanto de floresta aleatória foram aqueles gerados com hiper-parâmetros que mais favorecem a generalização do modelo, em troca de pior ajuste aos dados de treino. Isto leva a acreditar que as variáveis de entrada para os algoritmos, nestes casos, influenciavam menos no valor final da ROP, que provavelmente é determinada principalmente por outras variáveis que não foram consideradas.

Em contraste, os poços 1 e 2 tiveram melhores resultados em hiper-parâmetros intermediários, sem favorecer exageradamente a generalização ou o ajuste aos dados de treino. Isto indica que foi possível obter mais informação de cada exemplo de treino na previsão da ROP. Apesar disso, ainda é provável que outras variáveis não consideradas no modelo tenham grande influência na ROP, já que várias das variáveis tidas como determinantes no processo não puderam ser utilizadas.

De modo geral, observou-se também que, para as florestas aleatórias, a variação do número de árvores não surtiu muito efeito nas métricas de desempenho. É provável que a amplitude de valores considerada nos experimentos numéricos não foi adequada para uma boa comparação, e com 100 árvores as métricas já tenham alcançado um platô, não justificando o aumento da complexidade do modelo.

Observando individualmente os resultados de cada modelo para cada conjunto de testes em cada poço, valores negativos de R^2 ocorriam somente em um a três conjuntos de testes no poço, sempre na porção final dele. Isto tem um efeito negativo desproporcional nos valores médios, especialmente no poço 3, em que ocorria um único valor negativo, menor que -6, e todos os outros valores eram positivos. Como o valor de R^2 tem o valor 1 como limite superior, o único valor negativo, não representativo, fez com que a média do R^2 fosse negativa, enquanto o primeiro quartil (25%) já possuía valores maiores que $0,5$.

Para o teste de transferência de aprendizado, os resultados piores para os modelos simples e combinado mostram que a transferência de aprendizado utilizando estas estratégias é negativo, não sendo justificável utilizá-los. Isto, entretanto, não significa que a transferência não é possível, e outras abordagens como as mostradas em Segev et al. (2016) poderiam ter resultados melhores.

Quanto a importância de características, a maior importância dada à proximidade da broca pode ser visto como uma limitação do modelo, pois é uma característica mais relacionada ao comportamento do processo, não sendo um valor medido de antemão ou ajustado pelo operador.

Embora os resultados obtidos sejam inferiores a outros obtidos em literatura, alguns dos parâmetros de perfuração considerados como mais significativos para o processo, como perfurabilidade da formação, tipo e geometria da broca e propriedades físicas do fluido não puderam ser disponibilizadas, e resultados melhores provavelmente seriam obtidos

usando estes parâmetros como características para os modelos.

Adicionalmente, como mostrado em Hegde et al. (2017), o desempenho de um modelo treinado com dados pertencentes a formações diferentes daquela para que se está prevendo resulta em um desempenho menor da previsão, o que pode ter contribuído para um resultado pior.

5 CONCLUSÃO

Este trabalho teve como objetivo avaliar o uso de técnicas de aprendizado de máquina para a previsão da taxa de perfuração (ROP) durante a perfuração de poços de petróleo *offshore* de pré-sal da Bacia de Santos.

Iniciou-se com uma revisão geral de fundamentos da engenharia de petróleo focado em perfuração, especificamente *offshore*. Revisou-se primeiro os diferentes sistemas de equipamentos necessários para a perfuração do poço, e depois as diferentes operações contidas na atividade de perfuração e conceitos de perfuração direcional. Em seguida, revisou-se conceitos relacionados à taxa de perfuração, incluindo parâmetros que afetamna e diferentes modelos tradicionais comumente utilizados na previsão da ROP.

O próximo passo foi a apresentação dos fundamentos de aprendizado de máquina, incluindo os conceitos de o que é o aprendizado de máquina, como difere-se da programação adicional e sua metodologia básica. Os diferentes tipos de aprendizado foram explicados, especificando que tipos de problemas podem ser resolvidos com eles. Diferentes métricas de performance usadas para problemas supervisionados de regressão foram apresentados, seguido dos algoritmos que foram utilizados no trabalho. Apresentou-se o modelo da regressão linear e seu treino tanto utilizando o método de mínimos quadrados quanto o gradiente descendente. Os outros modelos expostos foram a árvore de decisão de regressão, treinada pelo algoritmo de CART, e seus principais hiper-parâmetros que podem ser ajustados, e a floresta aleatória, baseado na aplicação de métodos de comitê em múltiplos modelos de árvore de decisão. Por fim, apresentou-se os conceitos básicos do aprendizado por transferência, utilizado para transferir conhecimento entre domínios diferentes, mas semelhantes.

Apresentou-se também, trabalhos relacionados, iniciando com uma discussão rápida sobre trabalhos que utilizam metodologias de aprendizado de máquina em diferentes aplicações na indústria de óleo e gás, com foco na perfuração de poços. Em seguida, apresentou-se outros trabalhos que buscaram resolver o mesmo problema que este trabalho: a previsão da ROP a partir de parâmetros de perfuração de entrada. Algoritmos utilizados incluem redes neurais artificiais de diferentes arquiteturas, regressão linear, floresta aleatória e um método de comitê não especificado. A maior parte dos trabalhos utilizavam os dados indexados na profundidade, somente um utilizando-os indexado no tempo. Os principais parâmetros utilizados nos trabalhos foram o peso sobre broca, a rotação da broca, a profundidade, a densidade do fluido de perfuração, o diâmetro da broca e a perfurabilidade da formação. Aqueles trabalhos que compararam os resultados obtidos pelos modelos de aprendizado de máquina com aqueles obtidos por modelos tradicionais reportaram resultados consideravelmente superiores pela abordagem de aprendizado de máquina.

Em seguida, apresentou-se a metodologia utilizada no trabalho, incluindo a metodologia de coleta e preparação dos dados, e a definição dos experimentos numéricos que foram realizados. Os experimentos numéricos foram feitos em um total de 4 poços *offshore* da Bacia de Santos, nos campos de Sépia e Búzios, com dados disponíveis para diferentes extensões de profundidade do poço.

Identificou-se falhas nos conjuntos de dados e indicou-se como foram resolvidas para não interferirem negativamente nos experimentos numéricos. Em seguida, definiu-se novas características a serem utilizadas pelos modelos a partir de outras já existentes e uniu-se os dados em tempo real com os dados de *directional survey*. Por fim, agregou-se os dados na profundidade usando a média, para então dividí-los em conjuntos de treino progressivamente crescentes e conjuntos de teste de tamanho constante.

Os experimentos numéricos foram feitos com cada uma das duplas de conjuntos de treino e teste, utilizando algoritmos de regressão linear, árvore de decisão e floresta aleatória, variando hiper-parâmetros dos modelos. Para a árvore de decisão, variou-se o tamanho de amostra mínimo por folha, enquanto para a floresta aleatória variou-se o a profundidade máxima e o número de árvores utilizadas, hiper-parâmetros que influenciam nas capacidades de ajuste e generalização dos modelos.

Após os experimentos numéricos, avaliou-se também a capacidade de aprendizado por transferência usando informações adquiridas no treino de um poço para a previsão da ROP em outro. Esta avaliação foi feita utilizando duas estratégias simples: utilizar um modelo treinado com todos os dados do primeiro poço para prever no segundo, ou utilizando média entre o valor previsto por este modelo e o valor previsto por um modelo treinado no segundo poço para fazer a previsão.

Nos experimentos numéricos, não foi encontrado um conjunto modelo-hiper-parâmetros ótimo comum para todos os poços, mas identificou-se que, de forma geral, as florestas aleatórias obtiveram melhor desempenho. Nos poços para os quais os dados disponíveis eram apenas no final do poço, hiper-parâmetros que favorecessem a generalização do modelo tiveram resultados melhores, enquanto naqueles que havia dados para a maior parte da extensão do poço favoreceu-se modelos que se ajustavam de maneira mais fina aos dados. Os experimentos numéricos de aprendizado de transferência resultaram em uma transferência negativa, reduzindo o desempenho dos modelos. De forma geral, os resultados obtidos foram inferiores a outros trabalhos obtidos em literatura.

5.1 SUGESTÕES DE MELHORIA

Como afirmado anteriormente, diversos parâmetros considerados pela literatura como determinantes na previsão da ROP não estavam disponíveis durante a execução do trabalho, e por isso não puderam ser utilizados. Começar a utilizá-los provavelmente levará a uma melhoria considerável no desempenho dos modelos, especialmente os parâmetros

de formação litológica.

A coleta dos dados específico para perfuração, baseado nas informações dos relatórios diários de perfuração demandou um trabalho manual considerável, que poderia ser reduzido caso sejam utilizados métodos de identificação automática das operações de perfuração. Barbosa et al. (2019) apontaram alguns trabalhos que aplicam algoritmos para identificação dessas operações.

Enquanto os resultados de aprendizado por transferência obtidos foram negativos, isso não significa que a transferência não é possível neste modelo. Outras abordagens de aprendizado por transferência para florestas aleatórias e árvores de decisão podem ser encontrados em Segev et al. (2016).

REFERÊNCIAS

ABU-MOSTAFA, Y. S.; MAGDON-ISMAIL, M.; LIN, H.-T. **Learning From Data**. [S.l.]: AMLBook, 2012. ISBN 1600490069.

AL-ABDULJABBAR, A. et al. A robust rate of penetration model for carbonate formation. **Journal of Energy Resources Technology**, American Society of Mechanical Engineers Digital Collection, v. 141, n. 4, 2019.

ALKINANI, H. H. et al. Applications of artificial neural networks in the petroleum industry: a review. In: SOCIETY OF PETROLEUM ENGINEERS. **SPE Middle East Oil and Gas Show and Conference**. [S.l.], 2019.

BARBOSA, L. F. F. et al. Machine learning methods applied to drilling rate of penetration prediction and optimization-a review. **Journal of Petroleum Science and Engineering**, Elsevier, v. 183, p. 106332, 2019.

BATAEE, M.; IRAWAN, S.; KAMYAB, M. Artificial neural network model for prediction of drilling rate of penetration and optimization of parameters. **Journal of the Japan Petroleum Institute**, The Japan Petroleum Institute, v. 57, n. 2, p. 65–70, 2014.

BISHOP, C. M. **Pattern Recognition and Machine Learning**. [S.l.]: Springer, 2006.

BUSSAB, W. d. O.; MORETTIN, P. A. **Estatística básica**. [S.l.: s.n.], 2010.

ELMOUSALAMI, H. H.; ELASKARY, M. Drilling stuck pipe classification and mitigation in the gulf of suez oil fields using artificial intelligence. **Journal of Petroleum Exploration and Production Technology**, Springer, p. 1–14, 2020.

ESKANDARIAN, S.; BAHRAMI, P.; KAZEMI, P. A comprehensive data mining approach to estimate the rate of penetration: Application of neural network, rule based models and feature ranking. **Journal of Petroleum Science and Engineering**, Elsevier, v. 156, p. 605–615, 2017.

GABBAY, M. S. **Uma metodologia para estimar os custos de perfuração de poços de petróleo: um estudo de caso de dois campos onshore na região nordeste do Brasil**. Tese (Doutorado) — Universidade Federal do Rio Grande do Norte, 2015.

GANDELMAN, R. A. **Prediçao da ROP e otimizaçao em tempo real de parâmetros operacionais na perfuraçao de poços de petróleo offshore**. Dissertação (Mestrado) — Universidade Federal do Rio de Janeiro, 2012.

GÉRON, A. **Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. [S.l.]: O'Reilly Media, Incorporated, 2019. ISBN 9781492032649.

GURINA, E. et al. Application of machine learning to accidents detection at directional drilling. **Journal of Petroleum Science and Engineering**, Elsevier, v. 184, p. 106519, 2020.

HAJIZADEH, Y. Machine learning in oil and gas; a swot analysis approach. **Journal of Petroleum Science and Engineering**, Elsevier, v. 176, p. 661–663, 2019.

HEGDE, C. et al. Analysis of rate of penetration (ROP) prediction in drilling using physics-based and data-driven models. **Journal of Petroleum Science and Engineering**, Elsevier BV, v. 159, p. 295–306, nov. 2017. Disponível em: https://doi.org/10.1016/j.petrol.2017.09.020.

HEGDE, C.; GRAY, K. Use of machine learning and data analytics to increase drilling efficiency for nearby wells. **Journal of Natural Gas Science and Engineering**, Elsevier BV, v. 40, p. 327–335, abr. 2017. Disponível em: https://doi.org/10.1016/j.jngse.2017.02.019.

HEGDE, C. et al. Fully coupled end-to-end drilling optimization model using machine learning. **Journal of Petroleum Science and Engineering**, Elsevier BV, v. 186, p. 106681, mar. 2020. Disponível em: https://doi.org/10.1016/j.petrol.2019.106681.

JR, A. T. B.; JR, F. Y. et al. A multiple regression approach to optimal drilling and abnormal pressure detection. **Society of Petroleum Engineers Journal**, Society of Petroleum Engineers, v. 14, n. 04, p. 371–384, 1974.

MAURER, W. et al. The"perfect-cleaning"theory of rotary drilling. **Journal of Petroleum Technology**, Society of Petroleum Engineers, v. 14, n. 11, p. 1–270, 1962.

MITCHELL, T. M. et al. Machine learning. McGraw-hill New York, 1997.

MORAIS, J. M. d. **Petróleo em águas profundas: uma história tecnológica da Petrobras na exploração e produção offshore**. [S.l.]: Instituto de Pesquisa Econômica Aplicada (Ipea), 2013.

PAN, S. J.; YANG, Q. A survey on transfer learning. **IEEE Transactions on knowledge and data engineering**, IEEE, v. 22, n. 10, p. 1345–1359, 2009.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal of Research and Development**, v. 3, n. 3, p. 210–229, 1959.

SEGEV, N. et al. Learn on source, refine on target: A model transfer learning framework with random forests. **IEEE transactions on pattern analysis and machine intelligence**, IEEE, v. 39, n. 9, p. 1811–1824, 2016.

SHADIZADEH, S.; KARIMI, F.; ZOVEIDAVIANPOOR, M. Drilling stuck pipe prediction in iranian oil fields: An artificial neural network approach. **Iranian Journal of Chemical Engineering**, v. 7, n. 4, p. 29–41, 2010.

SHI, X. et al. An efficient approach for real-time prediction of rate of penetration in offshore drilling. **Mathematical Problems in Engineering**, Hindawi, v. 2016, 2016.

THOMAS, J. **Fundamentos de engenharia de petróleo**. [S.l.]: Interciência, 2001.

YUSWANDARI, A.; PRAYOGA, A.; PURBA, D. Rate of penetration (rop) prediction using artificial neural network to predict rop for nearby well in a geothermal field. **Proc. 44th Work. Geotherm. Reserv. Eng. Stanford Univ. Stanford, California, Febr. 11**, v. 13, n. 2019, 2019.