UNIVERSIDADE FEDERAL DA BAIHIA

ESCOLA POLITÉCNICA

CURSO DE GRADUAÇÃO EM ENGENHARIA QUÍMICA

VINICIUS JOSÉ MACEDO DE FREITAS

APLICAÇÃO DE SENSOREAMENTO VIRTUAL DATA-DRIVEN PARA A PREDIÇÃO DA VAZÃO DE SISTEMA DE BOMBEIO

SALVADOR - BAHIA

2023

VINICIUS JOSÉ MACEDO DE FREITAS

APLICAÇÃO DE SENSOREAMENTO VIRTUAL DATA-DRIVEN PARA A PREDIÇÃO DA VAZÃO DE SISTEMA DE BOMBEIO

Trabalho de Conclusão apresentado ao Curso de Graduação em Engenharia Química da Universidade Federal da Bahia, como requisito parcial à obtenção do título de Bacharel em Engenheira Química.

Orientador: Prof. Dr. Leizer Schnitman

Coorientador: Dr. Galdir Damasceno Reges Junior

SALVADOR - BAHIA

2023

FOLHA DE APROVAÇÃO

VINICIUS JOSÉ MACEDO DE FREITAS

APLICAÇÃO DE SENSOREAMENTO VIRTUAL DATA-DRIVEN PARA A PREDIÇÃO DA VAZÃO DE SISTEMA DE BOMBEIO

Trabalho de Conclusão de apresentado ao Curso de Graduação em Engenharia Química da Universidade Federal da Bahia, como requisito parcial à obtenção do título de Bacharel em Engenheira Química, foi avaliado pela seguinte banca examinadora:

BANCA EXAMINADORA

_____

Orientador

Prof. Dr. Leizer Schnitman (UFBA)

_____

1º Membro Avaliador

Prof. Dr. Delano Mendes de Santana (UFBA)

_____

2º Membro Avaliador

Prof. Me. Erbet Almeida Costa (UFBA)

Salvador, BA, 15 de dezembro de 2023.

AGRADECIMENTOS

Aos meus pais, Rodolfo e Gislane, que me deram todo o apoio necessário para seguir meus estudos.

À minha namorada Lais, que sempre esteve ao meu lado para me incentivar, desde o início até o final da graduação.

Aos integrantes do LEA, principalmente para Leizer, Galdir, Erbet e Odilon, que investiram um bom tempo para me auxiliar, e fizeram ser possível a realização deste trabalho.

Aos meus colegas de curso, que tornaram a jornada acadêmica muito mais leve e sem vocês não seria a mesma coisa.

À empresa iFood, que permitiu a flexibilização dos meus horários de trabalho durante o período de estudo.

Por fim, gostaria de agradecer à Universidade Federal da Bahia e todo seu corpo docente.

RESUMO

Este trabalho aborda a aplicação de sensoreamento virtual data-driven para a predição da vazão de um sistema de bombeio. O objetivo principal é medir o erro da predição da vazão e comparar o desempenho entre diferentes modelos de aprendizado de máquina. A metodologia empregada envolveu o uso das bibliotecas sklearn e Tensorflow em Python para a construção e análise de modelos, bem como o tratamento dos dados. Os passos utilizados na metodologia foram: definição dos regressores, definição dos modelos utilizados e preparação dos dados (normalização e separação dos dados em treino e teste). Os modelos utilizados foram: regressão linear, floresta aleatória, vetores de suporte e redes neurais. Os resultados revelaram que a regressão linear apresentou os maiores erros, enquanto as redes neurais tiveram o melhor resultado. Os resultados evidenciaram a flexibilidade e robustez das redes neurais para tratamento de dados, em comparação com a simplicidade da regressão linear.

Palavras-chave: Sensor Virtual, Aprendizado de Máquina e Bombeamento Centrífugo Submerso.

ABSTRACT

This work addresses the application of virtual data-driven sensing for predicting the flow rate of a pumping system. The main objective is to measure the prediction error of the flow rate and compare the performance among different machine learning models. The employed methodology involved the use of the sklearn and Tensorflow libraries in Python for building and analyzing models, as well as data preprocessing. The steps used in the methodology were: defining the regressors, specifying the models used, and preparing the data (normalization and splitting into training and testing sets). The models used included linear regression, random forest, support vector machines, and neural networks. The results revealed that linear regression exhibited the highest errors, while neural networks yielded the best performance. The findings underscored the flexibility and robustness of neural networks in data handling, in contrast to the simplicity of linear regression.

Keywords: Soft-sensor, Machine Learning and Electrical Submersible Pump.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

AID	*Automatic Interaction Detection*
BCS	Bombeamento Centrífugo Submerso
BCS-LEA	Planta de Bombeamento Centrífugo Submerso do Laboratório de Elevação Artificial da Universidade Federal da Bahia
BCP	Bombeamento por Cavidades Progressivas
BM	Bombeamento Mecânico com Hastes
CART	*Classification and Regression Tree*
CTAI	Centro de Capacitação Tecnológica em Automação Industrial
ELU	Unidade Linear Exponencial
GRU	*Gated Recurrent Units*
LEA	Laboratório de Elevação Artificial
LSTM	*Long Short-Term Memory*
MLP	*Multilayer Perceptron*
RBF	*Radial Basis Function*
RNA	Redes Neurais Artificiais
SVM	*Support Vector Machine*
SVR	*Support Vector Regression*
UFBA	Universidade Federal da Bahia

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO E JUSTIFICATIVA

Em qualquer aplicação industrial, o monitoramento das variáveis de processo é uma parte crucial da supervisão da saúde do processo, garantindo que ele funcione dentro dos limites desejados (Abeykoon, 2018). Tendo isso em vista, a necessidade de monitoramento das variáveis em caso de falha ou indisponibilidade de sensores físicos, atrelado ao avanço no ramo de aprendizado de máquina, faz com que a medição por sensores virtuais venha crescendo cada vez mais.

Graças ao avanço computacional, a utilização de sensores virtuais baseados em dados pode trazer mais confiabilidade para um sistema, visto que não possuem considerações como variáveis constantes (Kadlec *et al.*, 2009). Os sensores virtuais são baseados nos dados que refletem o comportamento da planta obtidos previamente por sensores físicos, conseguindo ser genéricos o suficiente para modelar com sucesso diversos estados de um sistema, como por exemplo o estado transiente e o estado estacionário.

Além disso, o sensoreamento virtual pode ser utilizado em diversas áreas da engenharia, sendo uma delas a engenharia de petróleo. De acordo com Thomas (2001), apesar da expansão de outras formas de geração de energia, o consumo de petróleo ainda é chave para o desenvolvimento de grande parte dos países, sendo um pilar da economia brasileira.

Com isso, este trabalho se propõe a executar um estudo de caso, da aplicação de sensores virtuais *data-driven* de um sistema de bombeio centrífugo submerso utilizando aprendizado de máquina, a fim de verificar a viabilidade da implementação de sensores virtuais em sistemas de produção de petróleo.

1.2 OBJETIVO DA PESQUISA

1.2.1 OBJETIVO GERAL

O objetivo geral deste trabalho é aplicar técnicas de sensoreamento virtual *data-driven*, usando aprendizado de máquina, a fim de prever a vazão na válvula choke do sistema de bombeio centrífugo submerso. Para validações práticas, este trabalho se baseia em dados gerados pela unidade piloto de Bombeamento Centrífugo

Submerso do Laboratório de Elevação Artificial da Universidade Federal da Bahia (BCS-LEA).

1.2.2 OBJETIVOS ESPECÍFICOS

a) Medir o erro médio quadrático da predição da vazão do sistema BCS-LEA, a partir de um sensor virtual, considerando dois cenários:
    a. Sensor virtual utilizando como dados de entrada: pressão e temperatura no manifold, abertura da válvula pneumática e frequência de operação.
    b. Sensor virtual utilizando como dados de entrada: abertura da válvula pneumática e frequência de operação.
b) Comparar o erro médio quadrático ao utilizar diferentes tipos de modelos, como: regressão linear múltipla, SVR, floresta aleatória e rede neural artificial.

2 LABORATÓRIO DE ELEVAÇÃO ARTIFICIAL

O Laboratório de Elevação Artificial (LEA) teve a sua planta BCS inaugurada em 2011 e é um dos laboratórios do Centro de Capacitação Tecnológica em Automação Industrial (CTAI), localizado na Universidade Federal da Bahia. O LEA conta com uma planta de bombeamento centrífugo submerso (BCS), em uma coluna com 32 metros de profundidade (Figura 1).

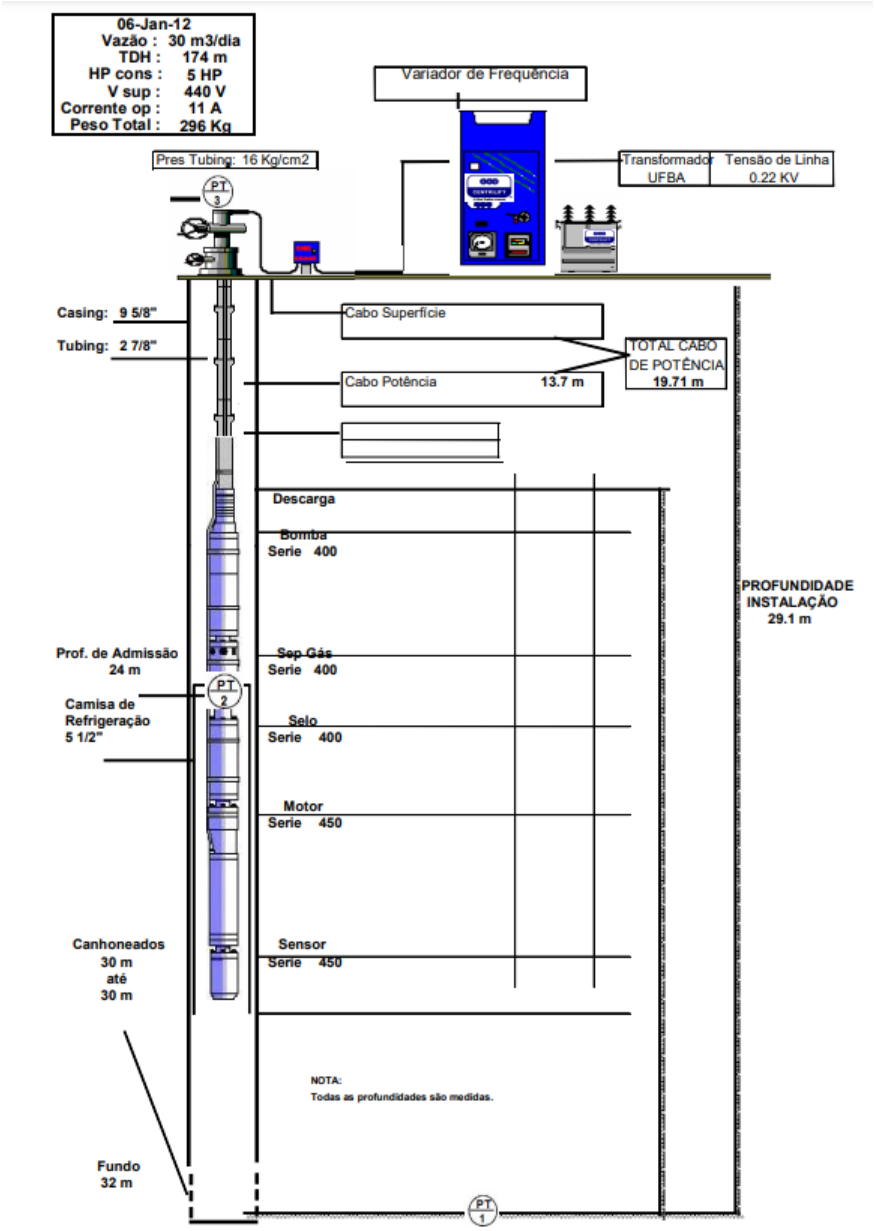

Figura 1: Planta BCS do Laboratório de Elevação Artificial da UFBA.

Fonte: Souza (2014).

Além disso, o poço possui instrumentos em diversos pontos, os quais permitem o

acompanhamento das variáveis de processo e simulações integradas com o MATLAB/Simulink.

O laboratório foi projetado para dar suporte às atividades de ensino e pesquisa, particularmente em engenharia de petróleo, controle e automação, química, mecânica e mecatrônica.

A infraestrutura do LEA proporcionou a realização de diversas pesquisas. Uma das pesquisas foi feita por Reges *et al.* (2021), na qual realizaram uma análise de vibração no sistema BCS submersa em óleo e em diferentes condições de operação. Outro trabalho foi realizado por Costa *et al.* (2021), no qual foram estimados parâmetros do sistema BCS a partir das variáveis de processo.

3 REVISÃO DE LITERATURA

3.1 BOMBEAMENTO CENTRÍFUGO SUBMERSO

3.1.1 INTRODUÇÃO

Caso a pressão de um reservatório seja suficientemente elevada, seus fluidos chegam naturalmente à superfície, caracterizando um poço surgente. Porém, nem todos os poços possuem pressão elevada o suficiente ou têm sua pressão reduzida devido à retirada de fluido (Thomas, 2001). Tendo isso em vista, é necessário um complemento de energia, para que a elevação do óleo seja bem sucedida, e isso caracteriza a elevação artificial.

Existem diversas técnicas de elevação artificial, tais como: bombeio centrífugo submerso (BCS), bombeamento por cavidades progressivas (BCP), bombeamento mecânico com hastes (BM) e gas-lift (Thomas, 2001). De acordo com Souza (2014), a escolha do método de elevação depende de vários critérios, dentre eles: razão gás-óleo, viscosidade dos fluidos, profundidade do reservatório e temperatura.

Neste contexto, este trabalho explorará em detalhes a elevação artificial de petróleo, com foco no BCS. O sistema BCS foi inventado e desenvolvido por Armais Arutunoff, no fim da década de 1910 (Takács, 2009).

A primeira instalação de um sistema BCS foi realizada em 1926, no campo “El Dorado” em Cansas. As primeiras unidades de BCS eram compostas de motores com até 6 metros de comprimento, com potências em torno de 105 HP. Essas unidades continham uma vedação logo acima do motor para evitar a infiltração de fluidos do poço, e eram alimentados por um cabo de três condutores vindos da superfície. Até hoje, esse conjunto de peças continuam sendo os principais componentes de um sistema BCS (Takács, 2009).

3.1.2 EQUIPAMENTOS DE SUBSUPERFÍCIE EM UM SISTEMA BCS

De acordo com Thomas (2001), os principais equipamentos de subsuperfície em um sistema BCS são: bomba, admissão da bomba, protetor, motor elétrico e cabo elétrico.

A bomba é do tipo centrífuga de múltiplos estágios, com cada estágio possuindo um impulsor e um difusor. A função do impulsor é fornecer energia cinética para o fluido, que é redirecionado pelo difusor, resultando em um aumento de pressão. Cada

estágio da bomba incrementa a pressão do fluido, e a quantidade de estágios da bomba deve ser suficiente para que os fluidos cheguem à superfície.

Na entrada da bomba há uma seção de admissão, que é responsável por separar o gás presente no fluido, minimizando a entrada de gás no primeiro estágio da bomba. Essa seção é extremamente importante, visto que evita que a bomba sofra com a interferência do gás, e que pode levar a condição de bloqueio (*gás lock)*.

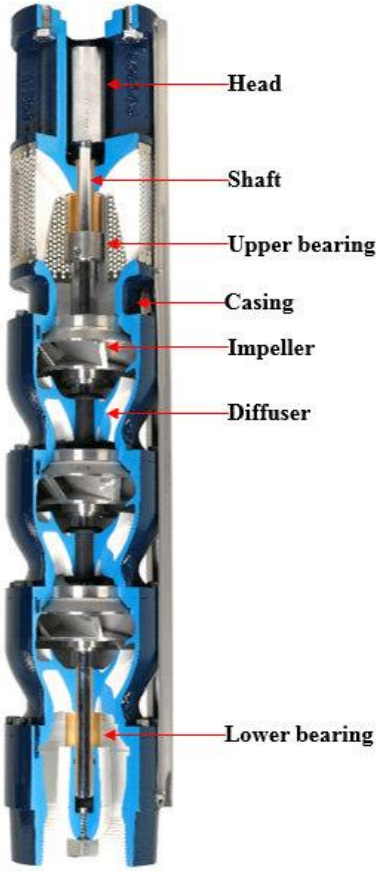

Figura 2: Bomba centrífuga de múltiplos estágios.

Fonte: Zhu (2018).

Os motores elétricos são responsáveis por fornecer energia cinética para o impulsor. Esses motores são do tipo trifásico, e devem ser preenchidos com um óleo mineral que garanta seu isolamento elétrico e o resfriamento do motor, tendo em vista que tais motores trabalham em condições extremamente severas.

Figura 3: Motor elétrico trifásico.

Fonte: Zhu (2018).

Os cabos elétricos, trifásicos de cobre ou alumínio, são responsáveis por transmitir energia da superfície ao motor trifásico.

De acordo com Kunnel (2000), 67% das falhas em BCS são de natureza elétrica, com a maioria com a motivação no cabo elétrico, principalmente na partida e parada do equipamento, visto que nesses momentos a corrente no cabo pode atingir valores até 7 vezes maiores que o valor de operação.

A figura a seguir ilustra a junção de todos os componentes supracitados.

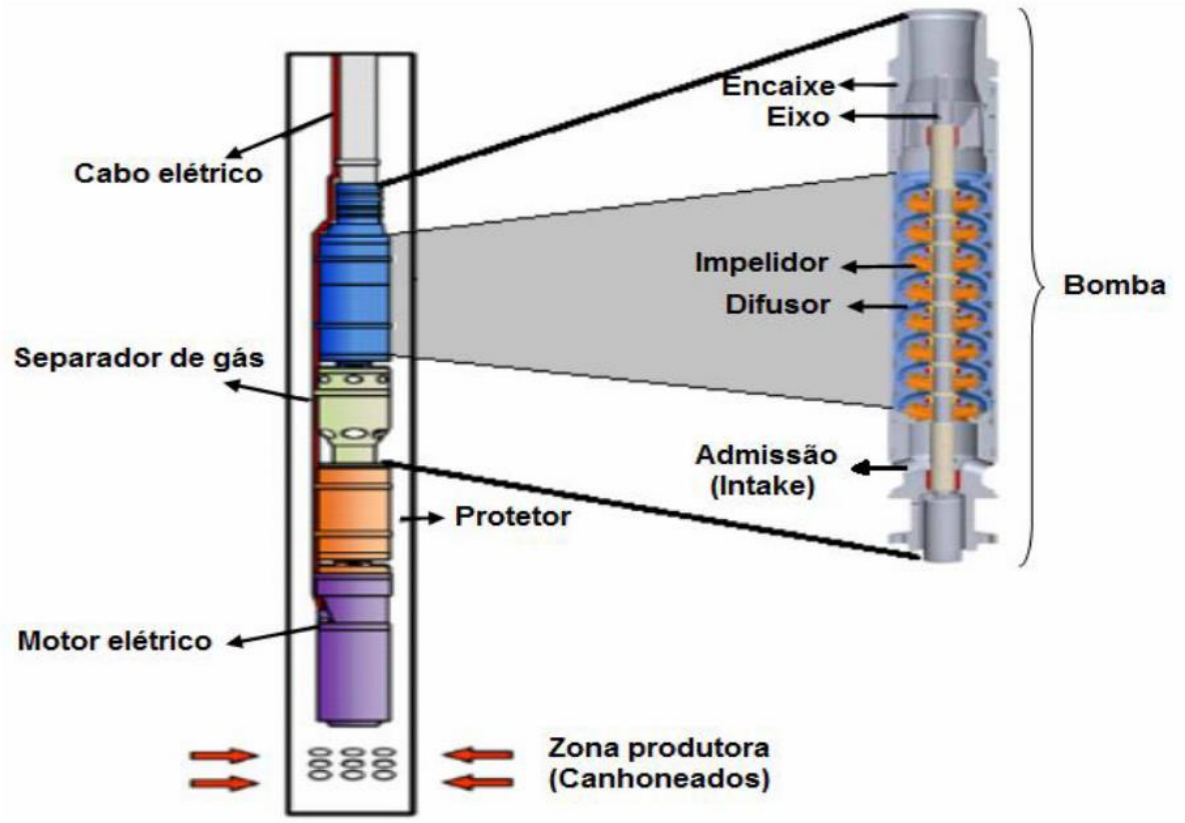

Figura 4: Conjunto de fundo do sistema de BCS.

Fonte: Souza (2014).

3.2 SENSOREAMENTO VIRTUAL

3.2.1 INTRODUÇÃO

Usualmente, o monitoramento das variáveis de processo é feito a partir da instalação de sensores físicos em locais desejados no processo (Kadlec *et al.*, 2009). No entanto, sensores físicos podem não ser adequados para certas aplicações devido a diversos fatores, tais como: ambientes de trabalho adversos que resultam na necessidade de manutenção frequente do equipamento, perturbações no fluxo do processo e na qualidade do produto, atrasos nas medições, altos custos, entre outros (Perera *et al.*, 2023).

A fim de resolver esses problemas, foi investigada a possibilidade de estimar essas variáveis de difícil medição através de modelos matemáticos, e isso originou o conceito de sensores virtuais.

Os sensores virtuais são amplamente utilizados em sistemas de BCS, devido a dificuldade de medição física das variáveis de processo. Uma modelagem de sensor

virtual foi feita por Aceros (2018), a partir de equações fenomenológicas.

3.2.2 TIPOS DE SENSORES VIRTUAIS

De acordo com Kadlec *et al.* (2009), existem dois principais tipos de sensores virtuais: *model-driven* e *data-driven*.

Os sensores virtuais *model-driven* são uma classe de sensores que se baseiam nas leis químicas e físicas que regem o processo, e foram os primeiros a serem utilizados em aplicações de monitoramento e controle de processos. A eficácia dos sensores virtuais *model-driven,* está diretamente ligada à qualidade da modelagem do processo em questão, bem como a todas as suposições feitas durante essa modelagem.

Tais suposições podem incluir: foco maior no estado estacionário do processo, ausência de perturbações significativas no processo, consideração de variáveis constantes, entre outros fatores. Qualquer desvio das condições previamente assumidas na modelagem do processo irá afetar a precisão e confiabilidade dos sensores virtuais *model-driven* (Kadlec *et al.*, 2009).

A fim de solucionar os problemas associados aos sensores virtuais *model-driven*, os sensores *data-driven* passaram a ser amplamente utilizados, visto que, são baseados em dados medidos, e representam as reais condições do processo. Consequentemente, há quem proponha que os sensores *data-driven* descrevem as variáveis de processo de uma forma mais real quando comparados aos sensores *model-driven* (Kadlec *et al.*, 2009).

Diversos tipos de algoritmos podem ser utilizados nos sensores *data-driven*. Chou *et al*. (2019), através do algoritmo *Gated Recurrent Units* (GRU) estimou a impureza do destilado e do produto de fundo de uma coluna de destilação. Hu *et al.* (2021) utilizou diversas abordagens baseadas em redes neurais, como *Long Short-Term Memory* (LSTM) e *Multilayer Perceptron* (MLP), para estimar o índice de fluxo de fusão do propileno em um processo de polimerização.

3.3 APRENDIZADO DE MÁQUINA

De acordo com Géron (2022), aprendizado de máquina é a ciência de programar computadores para que eles aprendam a partir de dados. Uma das principais vertentes do aprendizado de máquina é o aprendizado supervisionado, que consiste

em descrever a relação entre uma variável de interesse com uma ou mais variáveis explicativas, a partir de dados rotulados.

O aprendizado supervisionado é frequentemente utilizado em problemas de classificação e regressão. A distinção desses dois tipos de problemas está no tipo de dado do rótulo de cada um. A classificação apresenta dados rotulados com variáveis categóricas, como por exemplo a raça de um animal. Já a regressão, apresenta dados rotulados com variáveis numéricas e contínuas, como por exemplo o preço de um imóvel ou a pressão de um tanque. Como este trabalho requer técnicas de regressão, este será o tipo de problema em foco.

Para o algoritmo de aprendizado de máquina aprender a partir de dados, é necessária uma etapa de treino. O método de treino de cada algoritmo pode variar, porém, em geral, para o aprendizado supervisionado, as etapas a seguir são comuns a todos os algoritmos:

1. Separação dos dados em dados de treino e teste.
2. Treino do algoritmo com dados de treino.
3. Predição dos dados de teste, e avaliação do erro comparado com os dados de teste reais.

A etapa de separação dos dados em treino e teste é crucial para qualquer modelagem de aprendizado de máquina. Essa separação faz com que o algoritmo aprenda com os dados de treino, e receba os dados de teste como dados nunca vistos antes, simulando a aplicação do modelo em produção.

Para a avaliação do erro das predições em relação aos dados reais, em problemas de regressão, geralmente são utilizados o erro médio quadrático ou o erro médio absoluto.

A seguir, serão detalhados os algoritmos de aprendizado de máquina utilizados neste trabalho.

3.3.1 REGRESSÃO LINEAR MÚLTIPLA

De acordo com Daniya *et al.* (2020), a regressão linear múltipla explica a relação entre uma variável contínua Y, e duas ou mais variáveis independentes (x_1,x_2,x_3,...,x_k).

Esse modelo pode ser representado pela equação 1, a seguir.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad (1)$$

Em que ε é o resíduo, β_k são constantes, Y é a variável dependente, X_k são as

variáveis dependentes.

A inferência dos valores de β é realizada a partir do método de mínimos quadrados, que consiste em minimizar a soma dos quadrados dos resíduos, definida pela equação 2, a seguir.

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij})^2 \quad (2)$$

3.3.2 VETORES DE SUPORTE

O algoritmo baseado em vetores de suporte foi inicialmente proposto por Vapnik *et al.* (1963). Inicialmente, a ideia foi utilizar vetores de suporte para resolução de problemas de classificação, dando origem ao algoritmo conhecido como *Support Vector Machine* (SVM), que é capaz de dividir e classificar os dados a partir de hiperplanos (Zhang e O'Donnell, 2020).

Mais de 30 anos depois, este algoritmo foi adaptado para problemas de regressão, dando origem ao algoritmo chamado *Support Vector Regression* (SVR). O SVR, ao invés de separar os dados por um hiperplano, introduz uma nova função de perda denominada ε-*insensitive*, que define ε como o desvio máximo que a previsão da variável estimada deve ter em relação ao valor real de treino (Figura 5) (Zhang e O'Donnell, 2020).

Além do parâmetro ε, também foram propostas as variáveis ζ e ζ*, que representam pontos fora do desvio máximo aceito (Figura 5).

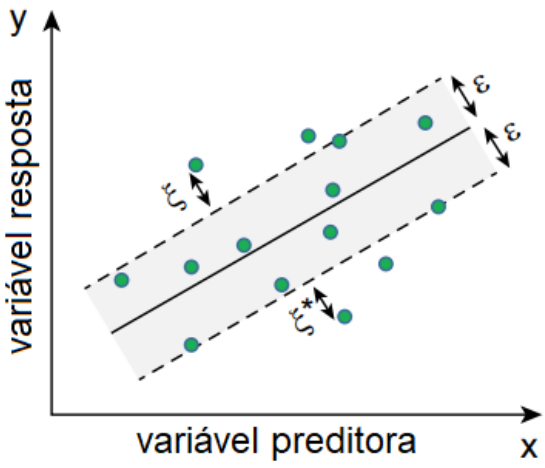

Figura 5: Representação gráfica de ε e ζ em uma função linear.

Fonte: Adaptado de Zhang e O'Donnell (2020).

Portanto, em caso de uma função linear do tipo:

$$y = f(x) = w^T x + b \tag{3}$$

em que w representa um vetor de pesos, x as variáveis independentes e b uma constante, a aproximação da função f pode ser feita com as equações representadas a seguir:

$$min_w \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \zeta_i + \zeta_i^* \tag{4}$$

$$em\ que \begin{cases} y_i - w^T x_i - b \leq \varepsilon + \zeta_i \\ w^T x_i + b - y_i \leq \varepsilon + \zeta_i^* \\ \zeta_i, \zeta_i^* \geq 0 \end{cases}$$

Com isso, a regressão se torna um problema de minimização do vetor de pesos w, ao mesmo tempo em que são minimizadas as variáveis ζ, e C é um parâmetro de regulação, que regula o peso das variáveis ζ.

Para a aproximação de sistemas não-lineares, foi introduzida uma "função kernel", que é responsável por transformar os dados de entrada não-lineares em dados lineares, a partir da adição de dimensões nos dados de entrada. A figura a seguir ilustra uma representação gráfica dessa transformação.

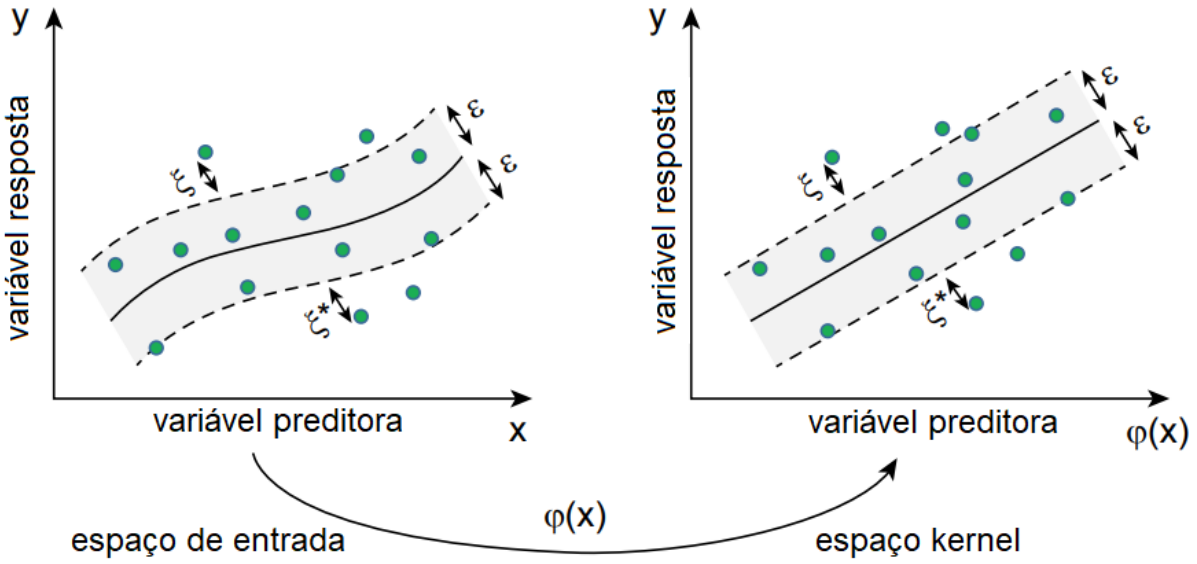

Figura 6: Representação gráfica da aplicação de uma função kernel.

Fonte: Adaptado de Zhang e O'Donnell (2020).

De acordo com Zhang e O'Donnell (2020), podem ser utilizadas diversas

funções kernel, como: *radial basis function* (RBF), polinomial, ANOVA RB, entre outros. A utilização de cada kernel depende da distribuição dos dados de entrada, sendo o kernel RBF utilizado de forma mais geral.

Tendo em vista a transformação dos dados de entrada pela função kernel, a regressão de dados não-lineares pode ser representada pela equação a seguir:

$$min_w \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l} \zeta_i + \zeta_i^* \tag{5}$$

$$em\ que \begin{cases} y_i - w^T\varphi(x_i) - b \leq \varepsilon + \zeta_i \\ w^T\varphi(x_i) + b - y_i \leq \varepsilon + \zeta_i^* \\ \zeta_i, \zeta_i^* \geq 0 \end{cases}$$

3.3.3 FLORESTA ALEATÓRIA

Para definir um algoritmo de floresta aleatória, é necessário primeiramente introduzir o conceito de árvores de decisão. Uma árvore de decisão é uma árvore onde cada nó interno (não terminal) representa um teste ou decisão sobre o item de dado considerado (Goebel e Gruenwald, 1999). De acordo Tan *et al*. (2005), sucesso pode ser explicado por diversos fatores, como:

• interpretabilidade do resultado;

• robustez para dados com ruídos;

• capacidade de lidar com atributos redundantes e irrelevantes, os quais, se não tratados adequadamente, podem prejudicar o desempenho do modelo;

As árvores de decisão podem ser separadas em dois grupos: árvores de classificação e árvores de regressão. As árvores de classificação têm o objetivo final de identificar uma classe de um conjunto de dados, ou seja, sua resposta é uma variável categórica. Já as árvores de regressão, a resposta final é uma variável contínua.

De acordo com Loh (2011), dois dos principais algoritmos de árvores de regressão são: *Automatic Interaction Detection* (AID) e *Classification and Regression Tree* (CART). Os dois algoritmos seguem a mesma lógica, definida a seguir.

1. Iniciar no primeiro nó da árvore.
2. Para cada entrada X, o algoritmo decide as variáveis de separação e seus valores, a fim de dividir o nó em outros dois nós.
3. Para cada nó filho, é calculada a média dos valores de y_i reais que pertencem

ao nó. Essa média é utilizada como o valor previsto de y_i, para todos os dados que pertencerem ao grupo.

4. O algoritmo verifica se alguma condição de parada foi atingida. Algumas condições que podem ser utilizadas são: tamanho máximo da árvore, limite de erro, entre outros.
5. Caso nenhuma condição de parada seja atingida, são repetidos os passos 2,3 e 4.

A Figura 7 a seguir representa visualmente o resultado do algoritmo, em que X_i são as variáveis de separação, t_i são os valores das variáveis de separação e R_i são os valores da previsão de y_i para os dados pertencentes aquele grupo, que é dado pela média dos valores reais de y_i.

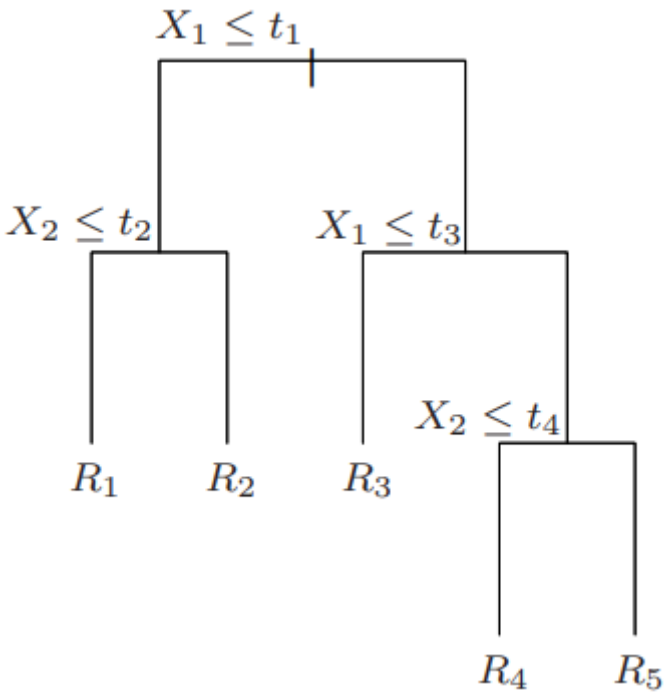

Figura 7: Representação gráfica de uma árvore de regressão.
Fonte: Hastie *et al*. (2001)

Tendo em vista a definição das árvores de regressão, podemos definir um regressor baseado em floresta aleatória como sendo uma média do resultado de múltiplas árvores de regressão. O processo de juntar diversos preditores do mesmo tipo é denominado *bagging*. Ao fazer isso, é gerado um modelo muito mais robusto, reduzindo a variância quando comparado a árvores de regressão separadas (Hastie *et al*., 2001).

3.3.4 REDES NEURAIS ARTIFICIAIS

O conceito de Redes Neurais Artificiais (RNA) foi introduzido pelos psiquiatras Warren McCulloch e Walter Pits (1943) no estudo "*A Logical Calculus of the Ideas Immament in Nervous Activity*". A proposta inicial foi replicar o comportamento dos neurônios biológicos, porém, a modelagem proposta pelos autores não era capaz de ser treinada. Com os avanços dos estudos, foram adicionados pesos nas conexões dos neurônios e foi introduzido o algoritmo de treino denominado *backpropagation*.

Hoje, a forma mais utilizada das RNAs é na forma de *Multilayer Perceptron* (MLP). Uma rede MLP é composta por unidades de processamento chamados de *perceptrons*, também conhecidos por neurônios. Estes neurônios podem ser alocados paralelamente em uma única camada, e em série com diversas camadas. Os neurônios em série são interconectados por pesos, que são adequados conforme o treinamento da rede, e visam minimizar o erro do modelo (Ludermir *et al*., 2000) (Figura 9).

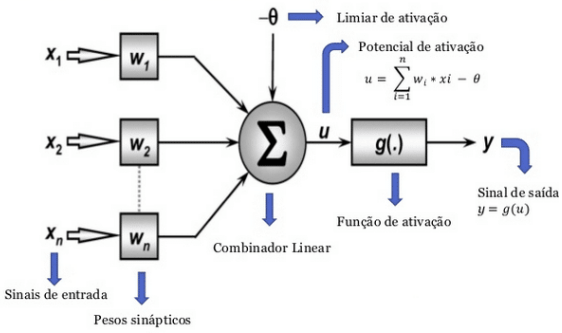

Figura 8: Rede Perceptron de uma única camada.

Fonte: Palmiere, S. E. 2022. Disponível em: https://embarcados.com.br/rede-perceptron-de-uma-unica-camada/. Acesso em: 30 de outubro de 2023.

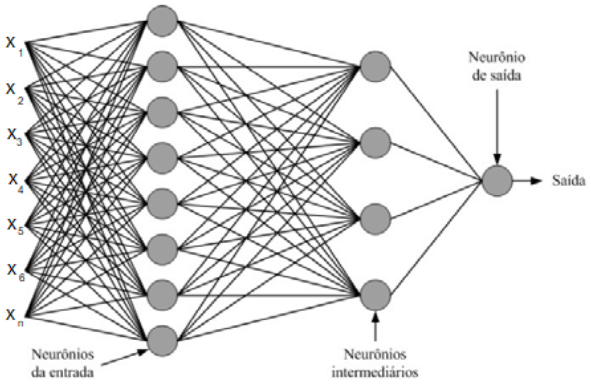

Figura 9: Representação gráfica de uma rede MLP.

Fonte: Adaptado de Rigo *et al*. (2016)

Como ilustrado na Figura 8, um neurônio é responsável por receber um vetor de pesos com seus respectivos sinais de entrada, um limiar de ativação, combinar linearmente esses valores, e, por fim, aplicar uma função de ativação g (Equação 6).

$$y = g\left(\sum_{i=1}^{n} w_i x_i - \theta\right) \quad (6)$$

Uma das funções de ativação mais utilizadas para problemas de regressão é a função Unidade Linear Exponencial (ELU) (Equação 7).

$$\begin{cases} x & , para\ x > 0 \\ \alpha(e^x - 1), & para\ x \leq 0 \end{cases} \quad (7)$$

O resultado da equação 6 é o sinal de saída, que é utilizado como sinal de entrada nas camadas posteriores da rede, ilustrado pela Figura 9.

O algoritmo de aprendizado da rede, *backpropagation*, pode ser dividido em duas partes: *forward pass* e *backward pass*. No *forward pass* os pesos da rede são fixados, e são computadas previsões através da aplicação da equação 6 em cada neurônio da rede, resultando em um valor final no neurônio de saída. Após isso, é aplicado o *backward pass*, que computa o erro da previsão atual com relação ao valor real, e com base nos erros, atualiza os pesos dos neurônios (Hastie *et al*., 2001).

4 ESTUDO DE CASO

Neste capítulo, será abordado um estudo de caso que se concentra na aplicabilidade do sensoreamento virtual na planta BCS-LEA. O objetivo principal deste estudo é avaliar como a implementação de um sensor virtual pode contribuir para a predição da vazão do óleo dessa planta. Este processo envolverá a aplicação das técnicas de modelagem de dados citadas no capítulo anterior, e a comparação dos resultados entre cada tipo de modelagem.

O estudo de caso na planta BCS-LEA oferece uma oportunidade valiosa para explorar os benefícios e desafios do sensoreamento virtual em um contexto prático. Através da análise das informações coletadas e dos resultados obtidos, será possível compreender como essa abordagem pode ser aplicada com sucesso na indústria.

A seguir, apresentaremos em detalhes a metodologia utilizada, os modelos empregados e os resultados obtidos durante o estudo de caso, permitindo uma visão abrangente do potencial do sensoreamento virtual na otimização de processos industriais.

4.1 METODOLOGIA

4.1.1 COLETA DE DADOS

A coleta de dados foi realizada a partir de um ensaio contínuo, durante período de 17/09/2021 a 20/09/2021, resultando em um conjunto de dados com 236454 pontos de dados distintos. O objetivo do ensaio foi observar o comportamento do sistema a partir da variação da frequência de operação e da abertura da válvula pneumática.

O autor não participou da coleta dos dados, os dados foram fornecidos pelos integrantes do LEA.

4.1.2 DEFINIÇÃO DOS REGRESSORES

Para a previsão da vazão do sistema, foram selecionados os seguintes regressores: pressão e temperatura no manifold, abertura da válvula pneumática e frequência de operação da bomba BCS. Todas essas variáveis apresentam alta correlação com a vazão do sistema (Figura 10), com destaque para a pressão no manifold, que apresenta correlação praticamente linear com a vazão do sistema. A abertura da válvula pneumática e a frequência de operação são as variáveis de

entrada do sistema, portanto, são alteradas pelos operadores da planta.

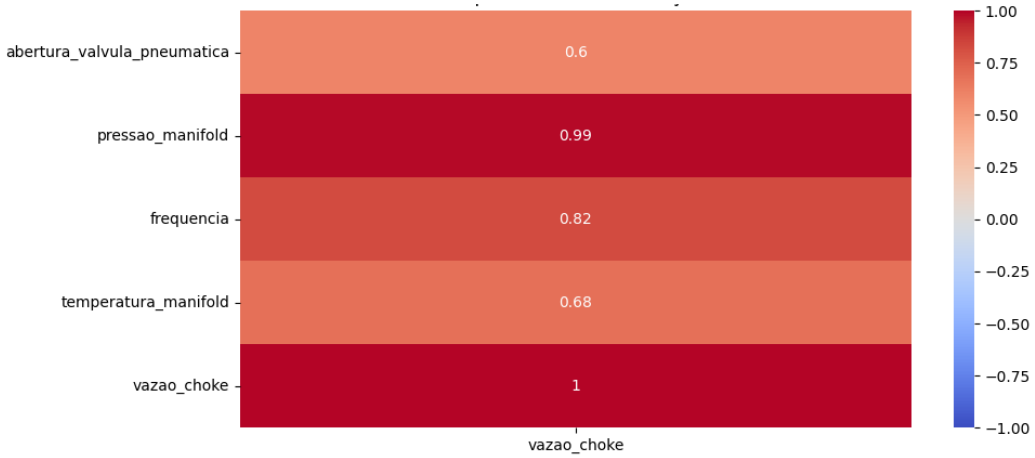

Figura 10: Correlação de Pearson da pressão e temperatura no manifold, abertura da válvula pneumática e frequência de operação, em relação a vazão na válvula choke.

Além disso, também foram utilizados os valores prévios dessas variáveis, visto que, sendo um sistema dinâmico, o valor atual da vazão (y) depende dos valores atuais e passados das variáveis de entrada (x) (Equação 8).

$$y(t) = f(x_t, x_{t-1}, x_{t-2}, x_{t-n}) \quad (8)$$

A definição do atraso n utilizado na modelagem foi feita a partir da aplicação de um único modelo com parâmetros iguais, apenas variando o valor de n, ou seja, variando a quantidade de amostras no passado utilizadas como variáveis de entrada. Baseado na variação do erro em relação a n, pode ser definido o valor ótimo.

Foi definido o algoritmo Random Forest como padrão, e n variando de 0 a 8. Após isso, foram treinados e testados os modelos, e o valor de n definido foi 5, visto que para valores de n maiores que 5, não há decréscimo significativo no erro do modelo (Figura 11).

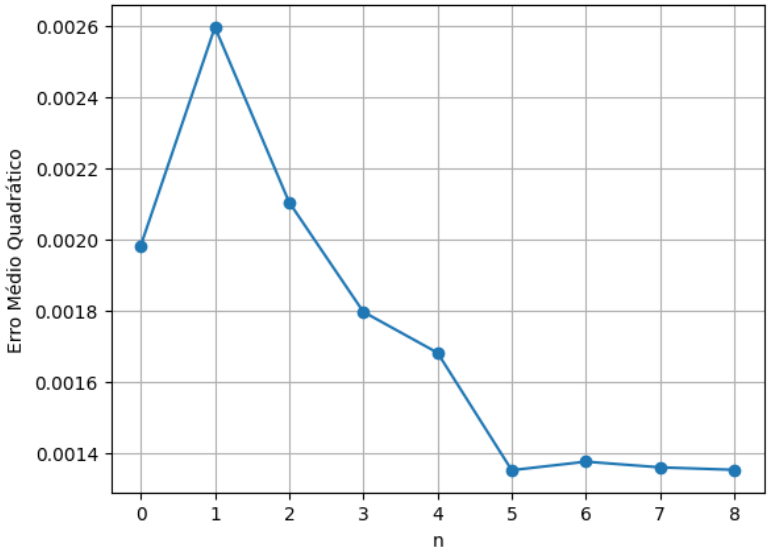

Figura 11: Erro médio quadrático X Atraso.

4.1.3 MODELOS UTILIZADOS

Os modelos utilizados neste trabalho foram: regressão linear múltipla, vetores de suporte, floresta aleatória e RNA.

Para a utilização dos três primeiros modelos, foi utilizada a biblioteca *sklearn*, que conta com diversos tipos de modelos, abstraídos para a linguagem *Python*. Os módulos utilizados foram, respectivamente, *sklearn.linear_model.LinearRegression*, *sklearn.svm.SVR, sklearn.ensemble.RandomForestRegressor*.

Já para a modelagem da RNA, foi utilizada a biblioteca *Tensorflow 2*. As bibliotecas citadas efetuam a modelagem a partir dos métodos explicados na revisão da literatura.

Para a definição dos parâmetros de cada modelo, foi utilizado uma busca aleatória, a partir de um *range* de parâmetros. Essa busca consiste em definir uma combinação de valores para cada parâmetro, e testar aleatoriamente uma combinação entre eles, especificando um número máximo de combinações distintas a serem testadas. O número máximo de combinações testado para cada modelo neste trabalho foi 50.

A seguir, estão os valores utilizados em cada parâmetro no modelo final, adicionado dos valores testados na busca aleatória, entre colchetes:

SVR:

- C = 1 [0,1; 1,0; 10,0]
- ε = 0,1 [0,01; 0.1; 0.2; 0.5]

- kernel = rbf [linear; poly; rbf; sigmoid]

Floresta Aleatória:

- n_estimators = 150 [25; 50; 75; 100; 125; 150; 175; 200]
- O restante dos parâmetros foram os parâmetros padrão do módulo.

Rede neural MLP:

- 3 camadas densas de 100 neurônios [1, 2, 3, 4, 5]
- Função de ativação elu nas camadas internas [elu; relu; linear]
- Função de ativação linear na camada externa [linear]
- EarlyStopping, com patience = 3. [3]
- Otimizador ADAM [ADAM]
- 100 épocas de treino [100]

A busca aleatória dos parâmetros para o SVR e a Floresta Aleatória foram feitos utilizando o módulo *sklearn.model_selection.RandomizedSearchCV*. Já a busca aleatória para a rede neural MLP foi feita utilizando um loop, feito manualmente, testando as combinações dos parâmetros citados acima.

4.1.4 PREPARAÇÃO DOS DADOS

A preparação adequada dos dados é uma das etapas cruciais para o sucesso de uma modelagem orientada a dados (Hastie *et al*., 2001). Para preparar os dados, foram realizadas duas etapas:

1. Separação dos dados em treino e teste
2. Normalização dos dados

De acordo Hastie *et al*. (2001), a proporção de dados de treino e teste varia conforme os dados utilizados e o problema a ser resolvido, porém, é comum utilizar de 70 a 80% dos dados para treino, e o restante para teste. Tendo isso em vista, para este trabalho, visto que se trata de uma série temporal, foram utilizados os primeiros 80% dos dados para treino, e os 20% restantes para teste (Figura 12).

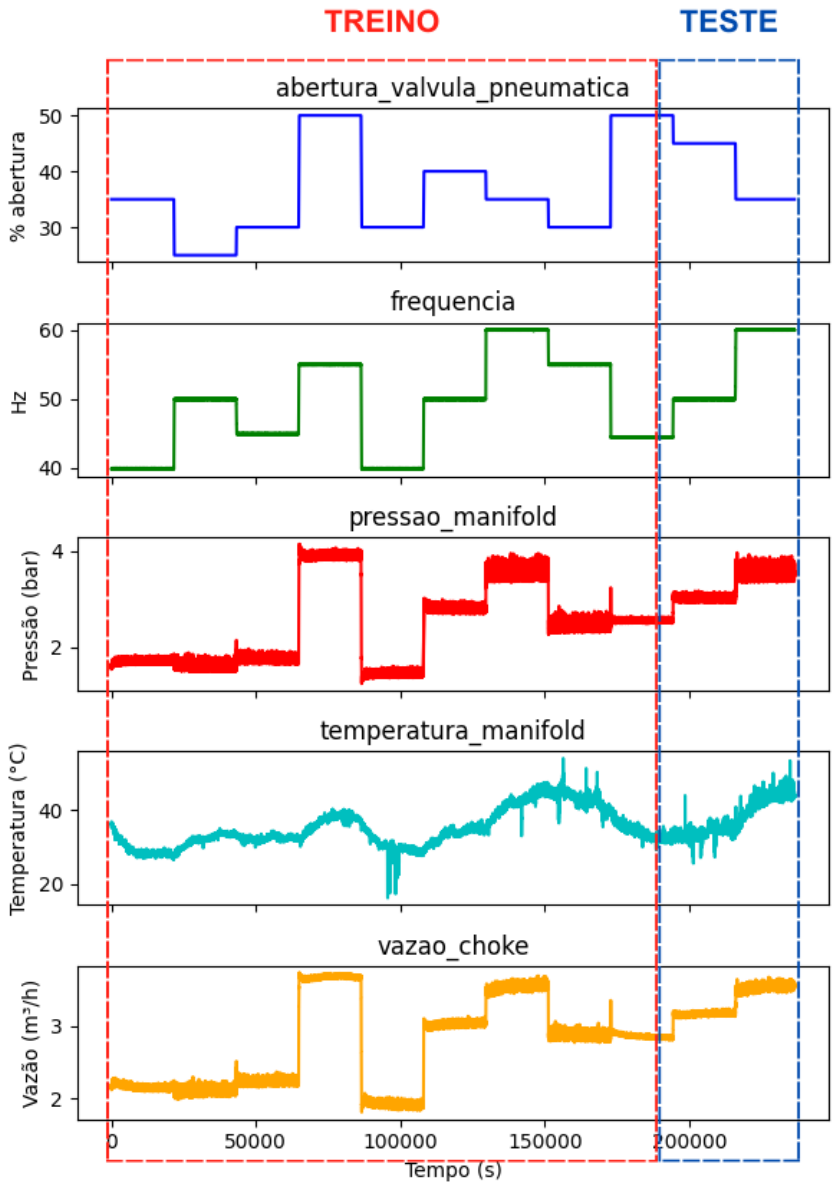

Figura 12: Ilustração dos dados de treino e teste.

A normalização dos dados é necessária na maioria dos problemas de aprendizado de máquina, devido a diferença de escala entre as variáveis de entrada. Ao normalizar os dados, colocamos todas as variáveis na mesma escala, sem perda de informação, gerando uma modelagem mais precisa (Hastie *et al*., 2001).

A técnica de normalização utilizada foi o escalonamento mínimo-máximo (Equação 9).

$$x_{esc} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{9}$$

Para aplicar corretamente a normalização dos dados de teste e treino, foram utilizados os valores de x_{min} e x_{max} apenas do subconjunto de treino, para que não haja nenhuma influência dos dados de teste na etapa de treino dos modelos.

4.2 RESULTADOS E DISCUSSÕES

Para a simulação do sensoreamento virtual foram propostos dois cenários:

- CENÁRIO 1 - Predição da vazão na choke com base na temperatura e pressão no manifold, abertura da válvula pneumática e frequência de operação.
- CENÁRIO 2 - Predição da vazão na choke apenas com base na abertura da válvula pneumática e frequência de operação, simulando uma perda dos sensores no manifold.

4.2.1 CENÁRIO 1

Para o primeiro cenário, utilizando os modelos descritos anteriormente, os resultados são ilustrados na Figura 13, em que a linha azul representa os dados reais de vazão, e a linha laranja os dados previstos por cada modelo.

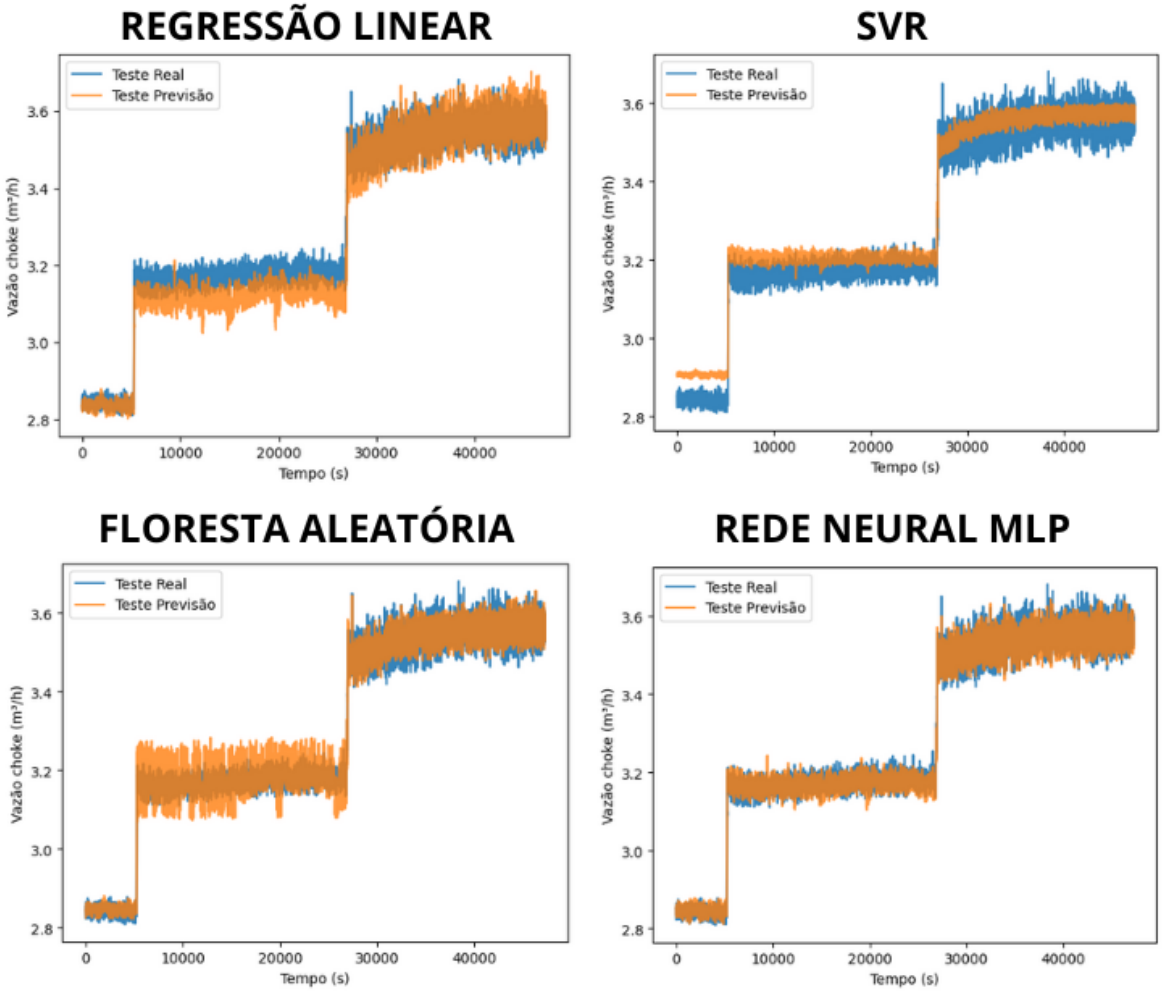

Figura 13: Resultados gráficos da modelagem no cenário 1.

Além da análise gráfica, foi medido o erro médio quadrático dos dados de teste reais em relação aos valores previstos (Tabela 1).

Tabela 1: Erro médio quadrático dos dados de teste reais em relação aos valores previstos, por modelo, no cenário 1.

Modelo	**Erro médio quadrático ($(m^3/h)^2$)**
Regressão Linear	*0,00185*
SVR	*0,00163*
Floresta Aleatória	*0,00083*
Rede Neural MLP	*0,00038*

De acordo com os resultados, é possível concluir que os quatro modelos apresentaram resultados satisfatórios, sendo a rede neural MLP o modelo com o menor erro.

Tendo em vista a alta correlação linear entre o valor da pressão no manifold e a vazão na choke, é esperado que a regressão linear consiga modelar o sistema com certa precisão, mesmo sendo o modelo mais simples entre os quatro testados. Apesar da precisão satisfatória, a análise gráfica (Figura 13) indica que a maior parte do erro está presente na segunda etapa estacionária do teste, o que provavelmente é reflexo dos outliers da temperatura no manifold nessa região dos dados de teste, ilustrados na Figura 14, a seguir.

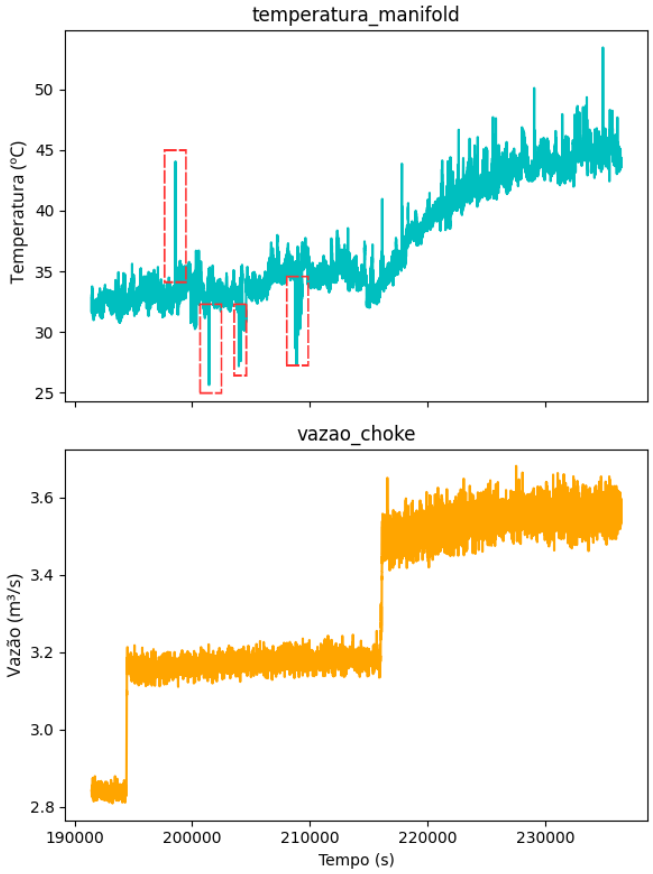

Figura 14: Dados da temperatura do manifold e vazão real da válvula choke, nos dados de teste. Outliers da temperatura destacados.

O modelo SVR apresentou menor variação em estado quase-estacionário nas previsões quando comparado ao restante das previsões e aos dados reais, o que pode ser atribuído ao parâmetro C, que, de acordo com a equação 5, controla o peso das variáveis ζ e ζ^*. Caso o valor de C fosse menor, veríamos uma maior variação nos valores de previsão.

O algoritmo de floresta aleatória, em termos de erro médio quadrático, apresentou o segundo menor erro, apenas perdendo para a rede neural. Apesar do baixo erro, podemos ver que, assim como a regressão linear, os resultados foram afetados pelos outliers de temperatura no manifold.

A rede neural MLP obteve os resultados com maior precisão, sendo cerca de 2x mais precisa quando comparada com a Floresta Aleatória. Tal resultado evidencia a flexibilidade das redes neurais, e sua grande capacidade de se adaptar aos dados de treino, e conseguir aprender e generalizar padrões, sem “decorá-los”.

4.2.2 CENÁRIO 2

Para o segundo cenário, não foram utilizados os dados dos sensores no manifold, apenas a abertura da válvula pneumática e a frequência de operação, que são parâmetros alterados pelos operadores. Os resultados são ilustrados na Figura 14, em que a linha azul representa os dados reais de vazão, e a linha laranja os dados previstos por cada modelo.

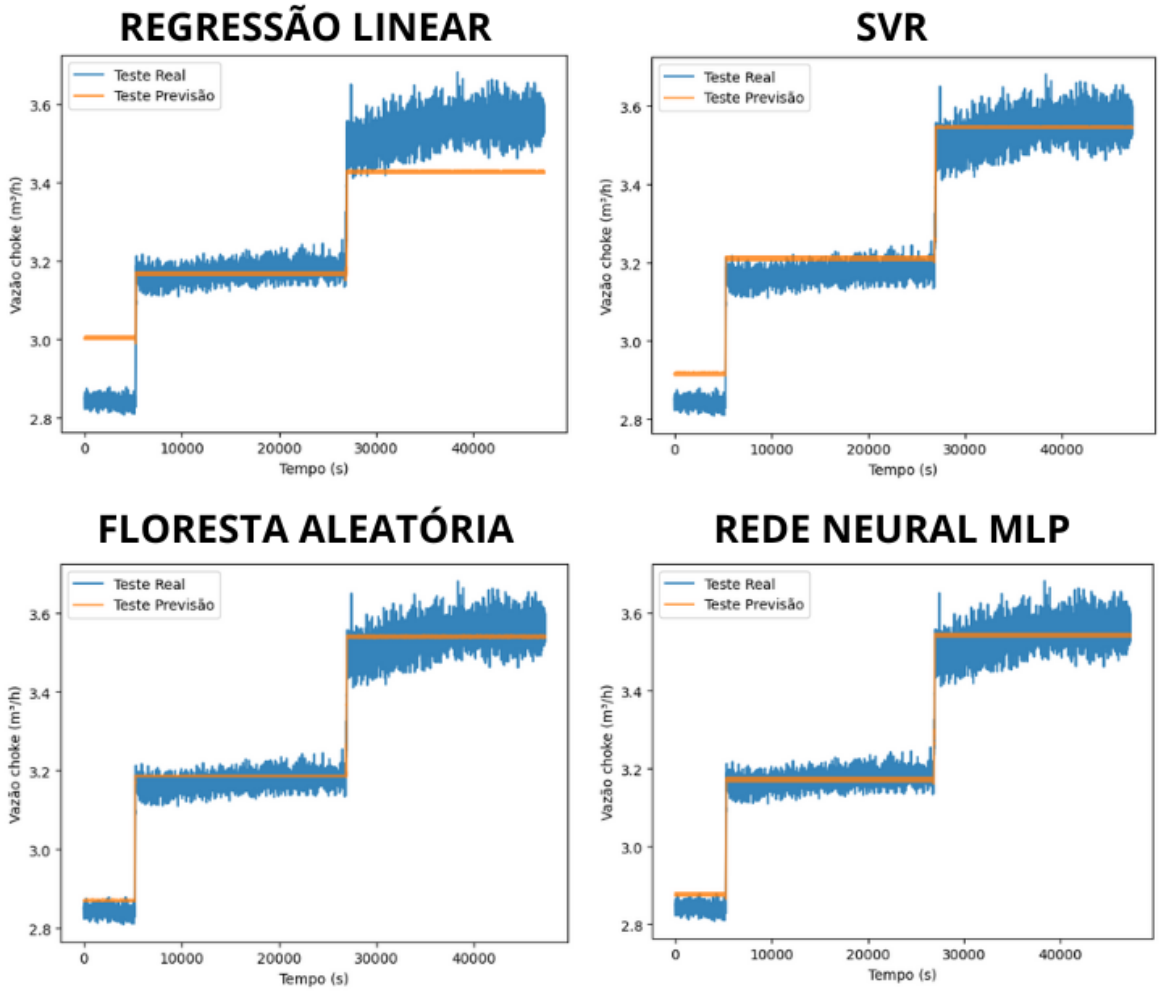

Figura 15: Resultados gráficos da modelagem no cenário 2.

Tabela 2: Erro médio quadrático dos dados de teste reais em relação aos valores previstos, por modelo, no cenário 2.

Modelo	**Erro médio quadrático ($(m^3/h)^2$)**
Regressão Linear	*0,00867*
SVR	*0,00210*
Floresta Aleatória	*0,00095*
Rede Neural MLP	*0,00089*

Observando os resultados do cenário 2, o principal ponto que é possível observar é a falta de variação nos dados previstos. Esse comportamento se dá devido à falta de variação dos dados de abertura da válvula pneumática e frequência de operação no estado quase-estacionário, visto que tais variáveis são definidas pela operação. Portanto, como não há ruídos nas variáveis preditoras, também não existe ruído nos valores de vazão previstos.

Além disso, é notável o aumento no erro quadrático médio em todos os modelos, sendo este um comportamento esperado, visto que não estão sendo utilizados os valores dos sensores do manifold, que são fortes preditores da vazão.

Apesar de todos os modelos perderem precisão com relação ao cenário 1, o modelo da regressão linear foi o que apresentou o maior aumento de erro (4,68x). Esse comportamento ocorre devido a não-linearidade da relação entre vazão na choke e abertura da válvula pneumática/frequência de operação da bomba.

Outro ponto de atenção é que, no cenário 2, os modelos não conseguem prever o aumento gradual da vazão no último degrau de teste, de 3,4 a 3,6 m^3/h. Esse comportamento pode ser explicado pela falta da temperatura no manifold, que é responsável por essa mudança gradual, visto que a temperatura continua mudando, enquanto a pressão no manifold se mantém igual.

Apesar do aumento do erro da predição da vazão, os resultados se mostraram satisfatórios.

4.2.3 COMPARAÇÃO ENTRE CENÁRIOS

Em suma, os resultados de ambos os cenários foram satisfatórios, e o cenário 1 se mostrou como sendo mais adequado para ser aplicado em um sensor virtual para

a planta BCS-LEA. Apesar do cenário 2 demonstrar um erro maior, tais modelos podem fornecer um sensor virtual com certa precisão em caso de perda dos sensores do manifold.

5 CONCLUSÃO

Esse trabalho pretendeu entender diferentes técnicas de sensoreamento virtual *data-driven*, a fim de utilizá-las para a predição da vazão do sistema BCS-LEA.

Para atingir tal objetivo, definiu-se dois objetivos específicos. O primeiro objetivo foi predizer a vazão do sistema BCS-LEA utilizando sensores virtuais em dois cenários. No primeiro cenário, os preditores foram: temperatura e pressão no manifold, abertura da válvula pneumática e frequência de operação da bomba. No segundo cenário, os preditores foram: abertura da válvula pneumática e frequência de operação da bomba.

Para esse primeiro objetivo, foi possível concluir que o primeiro cenário demonstrou uma maior acuracidade, como esperado. Apesar disso, em caso de perda dos sensores do manifold, seria possível estimar a vazão com certa precisão a partir do sensor proposto no segundo cenário.

O segundo objetivo foi comparar o erro dos seguintes modelos: regressão linear múltipla, SVR, floresta aleatória e rede neural MLP. A partir dos resultados foi possível concluir que a rede neural apresentou os melhores resultados, e a regressão linear múltipla, os piores, em ambos os cenários.

Com isso, foi possível confirmar a eficácia da aplicação de sensoreamento virtual *data-driven*, a partir de modelos de aprendizado de máquina.

5.1 LIMITAÇÕES E SUGESTÕES PARA TRABALHOS FUTUROS

Na modelagem de ambos os cenários, não foi realizado um teste exaustivo dos diferentes parâmetros possíveis para cada modelo devido a limitação de *hardware*, visto que os testes foram todos realizados no computador pessoal do autor. Portanto, os resultados aqui obtidos podem não refletir necessariamente a forma mais otimizada de cada modelo.

Para trabalhos futuros, é encorajado que seja testada uma gama maior de parâmetros, a fim de se aproximar do resultado mais otimizado possível. Além disso, é sugerido que a modelagem baseada em dados seja comparada com a modelagem fenomenológica do sistema.

6 REFERÊNCIAS

[1]. ABEYKOON, Chamil. Design and applications of soft sensors in polymer processing: A review. **IEEE Sensors Journal**, v. 19, n. 8, p. 2801-2813, 2018.

[2]. ACEROS, Egner et al. A First Principles Model for Virtually Sensing Operational Parameters in an ESP Well. **SPE Artificial Lift Conference and Exhibition-Americas**. OnePetro, 2018.

[3]. CHOU, Cheng-Hung et al. Physically consistent soft-sensor development using sequence-to-sequence neural networks. **IEEE Transactions on Industrial Informatics**, v. 16, n. 4, p. 2829-2838, 2019.

[4]. COSTA, E. A. et al. A Bayesian approach to the dynamic modeling of ESP-lifted oil well systems: An experimental validation on an ESP prototype. **Journal of Petroleum Science and Engineering**, v. 205, p. 108880, 2021.

[5]. DANIYA, T. et al. Least square estimation of parameters for linear regression. **International Journal of Control and Automation**, v. 13, n. 2, p. 447-452, 2020.

[6]. GÉRON, Aurélien. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow**. " O'Reilly Media, Inc.", 2022.

[7]. GOEBEL, Michael; GRUENWALD, Le. A survey of data mining and knowledge discovery software tools. **ACM SIGKDD explorations newsletter**, v. 1, n. 1, p. 20-33, 1999.

[8]. HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The elements of statistical learning: data mining, inference, and prediction**. Springer, 2001.

[9]. HU, Xuan et al. Novel soft sensor model based on spatio-temporal attention. In: **2021 International Joint Conference on Neural Networks (IJCNN)**. IEEE, 2021. p. 1-7.

[10]. KADLEC, Petr; GABRYS, Bogdan; STRANDT, Sibylle. Data-driven soft sensors in the process industry. **Computers & chemical engineering**, v. 33, n. 4, p. 795-814, 2009.

[11]. KUNNEL, B. Downhole pumps deliver broad gains. **Hart's E & P**, v. 73, n. 10, p. 71-80, 2000.

[12]. LOH, Wei-Yin. Classification and regression trees. **Wiley interdisciplinary reviews: data mining and knowledge discovery**, v. 1, n. 1, p. 14-23, 2011.Loh, W.-Y. (2011).

[13]. LUDERMIR, T. B.; BRAGA, A. P.; CARVALHO, A. Redes neurais artificiais: teoria e aplicações. Livros Técnicos e Científicos Editora, 2000.

[14]. MCCULLOCH, Warren S.; PITTS, Walter. A logical calculus of the ideas immanent in nervous activity. **The bulletin of mathematical biophysics**, v. 5, p. 115-133, 1943.

[15]. PERERA, Yasith S. et al. The role of artificial intelligence-driven soft sensors in advanced sustainable process industries: A critical review. **Engineering Applications of Artificial Intelligence**, v. 121, p. 105988, 2023.

[16]. RIGO JR, Luís Otávio et al. Aplicação de Multi-Layer Perceptron para Previsão de Emissão de Gases derivados de Veículos a Diesel. **Latin American Journal of Energy Research**, v. 3, n. 2, p. 1-11, 2016.

[17]. REGES, Galdir et al. Electric submersible pump vibration analysis under several operational conditions for vibration fault differential diagnosis. **Ocean Engineering**, v. 219, p. 108249, 2021.

[18]. SOUZA, L. **Controle avançado aplicado ao sistema BCS operando com escoamento monofásico.** Tese (Mestrado em Mecatrônica) – Escola Politécnica, Universidade Federal da Bahia. Bahia.

[19]. TAKACAS, G. Electrical Submersible Pumps-Manual. **British Library Cataloguing in**, 2009.

[20]. TAN, P.-N.; STEINBACH, M.; KUMAR, V. Introduction to Data Mining, (First Edition). **Addison-Wesley Longman Publishing Co., Inc.,** 2005.

[21]. THOMAS, José Eduardo. **Fundamentos de engenharia de petróleo**. Interciência, 2001.

[22]. VAPNIK, Vladimir N. Pattern recognition using generalized portrait method. **Automation and remote control**, v. 24, n. 6, p. 774-780, 1963.

[23]. ZHANG, Fan; O'DONNELL, Lauren J. Support vector regression. In: **Machine learning**. Academic Press, 2020. p. 123-140.

[24]. ZHU, Jianjun; ZHANG, Hong-Quan. A review of experiments and modeling of gas-liquid flow in electrical submersible pumps. **Energies**, v. 11, n. 1, p. 180, 2018.