UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA QUÍMICA
PROGRAMA DE PÓS–GRADUAÇÃO EM ENGENHARIA QUÍMICA

DISSERTAÇÃO

DESENVOLVIMENTO DE MODELOS MATEMÁTICOS BASEADOS EM *MACHINE LEARNING* NA PERFURAÇÃO DE POÇOS DE PETRÓLEO

TATIANE SILVA SOUSA

Seropédica
Dezembro, 2024

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA QUÍMICA
PROGRAMA DE PÓS–GRADUAÇÃO EM ENGENHARIA QUÍMICA

DESENVOLVIMENTO DE MODELOS MATEMÁTICOS BASEADOS EM *MACHINE LEARNING* NA PERFURAÇÃO DE POÇOS DE PETRÓLEO

TATIANE SILVA SOUSA

Sob a Orientação da Professora
Márcia Peixoto Vega Domiciano, D.Sc.

Dissertação submetida como requisito parcial para a obtenção do grau de **Mestre em Engenharia Química**, no Programa de Pós-Graduação em Engenharia Química, Área de Concentração em Tecnologia Química.

Seropédica
Dezembro, 2024

Universidade Federal Rural do Rio de Janeiro
Biblioteca Central / Seção de Processamento Técnico

Ficha catalográfica elaborada
com os dados fornecidos pelo(a) autor(a)

S725d

Sousa, Tatiane Silva, 1984-
Desenvolvimento de modelos matemáticos baseados em machine learning na perfuração de poços de petróleo. / Tatiane Silva Sousa. - Nova Iguaçu, 2024.
174 f.: il.

Orientadora: Márcia Peixoto Vega Domiciano.
Dissertação(Mestrado). -- Universidade Federal Rural do Rio de Janeiro, Curso de Pós-Graduação em Engenharia Química, 2024.

1. Bullheading. 2. Kick. 3. Machine learning. I. Domiciano, Márcia Peixoto Vega, 1972-, orient. II Universidade Federal Rural do Rio de Janeiro. Curso de Pós-Graduação em Engenharia Química III. Título.

UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO
INSTITUTO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA QUÍMICA
PROGRAMA DE PÓS–GRADUAÇÃO EM ENGENHARIA QUÍMICA

TATIANE SILVA SOUSA

Dissertação submetida como requisito parcial para a obtenção do grau de **Mestre em Engenharia Química**, no Programa de Pós-Graduação em Engenharia Química, Área de Concentração em Tecnologia Química.

APROVADA em 18 de dezembro de 2024.

gov.br Documento assinado digitalmente
MARCIA PEIXOTO VEGA DOMICIANO
Data: 24/02/2025 11:40:14-0300
Verifique em https://validar.iti.gov.br

Márcia Peixoto Vega Domiciano, D.Sc. - DEQ/IT/UFRRJ
Orientador

gov.br Documento assinado digitalmente
MAURICIO BEZERRA DE SOUZA JUNIOR
Data: 20/02/2025 16:04:48-0300
Verifique em https://validar.iti.gov.br

Maurício Bezerra de Souza Júnior, D.Sc. - EQ/UFRJ

gov.br Documento assinado digitalmente
CLAUDIA OSSANAI
Data: 20/02/2025 10:35:06-0300
Verifique em https://validar.iti.gov.br

Cláudia Ossanai, D.Sc. - UFF

AGRADECIMENTOS

Agradeço a Deus, a família e aos amigos pelo apoio para concluir mais uma etapa em minha vida.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

RESUMO

SOUSA, Tatiane Silva. **Desenvolvimento de modelos matemáticos baseados em machine learning na perfuração de poços de petróleo.** 2024. 174p. Dissertação (Mestrado em Engenharia Química, Tecnologia Química). Instituto de Tecnologia, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2024.

A exploração de regiões complexas (pré-sal, formações rochosas depletadas, reservatórios carbonáticos) exige técnicas de perfuração não convencionais como *Pressurized Mud Cap Drilling* (PMCD). Nesses cenários ocorrem frequentemente distúrbios de *kick* e perda total de circulação. Dessa forma, é essencial o desenvolvimento de modelos matemáticos eficientes para prever pressões no poço, garantindo segurança no operacional, isto é, que a perfuração seja conduzida dentro da janela operacional: acima da pressão de poros e abaixo da pressão de fratura. Neste sentido, para descrever a técnica PMCD, o modelo matemático deve prever adequadamente as etapas de migração de gás e a operação de *bullheading* (bombeamento de fluido de sacrifício em contra corrente, sem retorno à superfície, forçando o gás e cascalho a retornarem para a formação). O presente trabalho de dissertação de mestrado desenvolveu modelos baseados em *machine learning* a partir de dados da unidade experimental do LEF/DEQ/IT/UFRRJ e de dados da literatura de poços reais. As métricas de avaliação estatísticas dos modelos (R^2, RMSE, MSE, SSE) e as simulações dinâmicas revelaram boa capacidade preditiva quando informação transiente é empregada.

Palavras-chave: *bullheading, kick, machine learning*

ABSTRACT

SOUSA, Tatiane Silva. **Development of mathematical models based on machine learning in oil well drilling.** 2024. 174p. Dissertation (Master in Chemical Engineering, Chemical Technology). Instituto de Tecnologia, Universidade Federal Rural do Rio de Janeiro, Seropédica, RJ, 2024.

The exploration of complex regions (pre-salt, depleted formations, carbonate reservoirs) requires non conventional drilling techniques such as Pressurized Mud Cap Drilling (PMCD). total loss and kick disturbances are often observed in these scenarios. As a result, it is developing of major importance to mathematical models in order to predict well pressure and assure operational safety, i.e, drilling inside the operational window: above porous pressure and below fracture pressure. As a result, for PMCD describing purposes, the mathematical model needs to properly predict the steps of gas migration and the bullheading operation (countercurrent pumping of sacrificial fluid, without surface return, forcing gas and cutting back to formation). This dissertation work developed machine learning based models using experimental data from LEF/DEQ/IT/UFRRJ unit and real well literature data. Statistical measures (R^2, RMSE, MSE, SSE) and dynamic simulations concerning the developed models showed good predictive performance when transient information is proved.

Keywords: bullheading, kick, machine learning

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SÍMBOLOS

x	dado original
x_{min}	valor mínimo
x_{max}	valor máximo
x'	dado normalizado
y	dado transformado
w	Shapiro-Wilk
g1	coeficiente de assimetria de Fisher-Pierson
k	coeficiente de curtose de Fisher-Pierson
a_1	coeficiente determinado pelos dados
n	número de observações da amostra ou exemplos
m_4	quarto momento central
s	desvio padrão
m	número de características
D	conjunto de dados
y_i	previsão ou alvo
k	número total de funções ou classes
F	espaço de árvores de regressão
q	valor da divisão dos exemplos em diferentes folhas
f	número total de folhas na árvore
f_k	uma árvore específica
w	peso das folhas
ϕ	expressa que y está em função de x
l	função de perda
$\hat{y}_i$	previsão
Ω	parâmetro de penalização da complexidade do modelo
γ	parâmetro de controle na divisão do nó
λ	parâmetro de regularização da penalização do modelo
N	número total de amostras
y_i	valor real da amostra
y	média de todos os valores reais
p_{ii}	probabilidade de se pertencer a classe
z_i	pontuação da classe
e	exponencial
Σ	somatório
%	percentagem

Hz	unidade de medida de frequência em hertz
psi	unidade de medida de pressão
g_i	estatística de gradiente de primeira ordem da função da perda
h_i	estatística de gradiente de segunda ordem da função da perda
I_i	conjunto de índices dos exemplos que terminam na folha j
I	conjunto de instâncias no nó atual
I_L	divisão do nó para o lado esquerdo
I_R	divisão do nó para o lado direito

LISTA DE ABREVIAÇÕES

ANN	*Artificial Neural Network*
API	*Application Programing Interface*
CNN	*Convolutional Neural Network*
COA	*Cuckoo Search Algorithm*
ECD	Densidade Equivalente de Circulação
GA	*Genetic Algorithm*
GRU	*Gated Recurrent Unit*
KNN	*K-Nearest Neighbor Network*
LAM	*Light Annular Mud*
LCM	*Lost Circulation Material*
LEF	Laboratório de Escoamento de Fluidos Giulio Massarani
LSTM	*Long-Short Term Memory*
LSSVM	*Least Squares Support Vector Machines*
MCD	*Mud Cap Drilling*
MPD	*Managed Pressure Drilling*
MLP	*Multilayer Perceptron*
MSE	*Mean Squared Error*
PCA	*Principal Component Analysis*
PI	Proporcional Integral
PID	Proporcional Integral Derivativo
PMCD	*Pressurized Mud Cap Drilling*
PSO	*Particle Swarm Optimization*
RBF	Função de Base Radial
RMSE	*Root Mean Squared Error*
ROP	Taxa de Penetração
SAC	*Sacrificial Fluid*
SSE	*Sun Squared Error*
SVM	*Support Vector Machine*
UFRRJ	Universidade Federal Rural do Rio de Janeiro
XGBoost	*Extreme Gradient Boosting*

SUMÁRIO

1 INTRODUÇÃO

Neste capítulo são apresentadas as motivações que conduziram ao desenvolvimento do presente estudo, envolvendo a operação de *bullheading* na perfuração de poços de petróleo através da técnica *Pressurized Mud Cap Drilling* – PMCD, aplicando *machine learning,* bem como os objetivos que norteiam o trabalho.

Segundo Helgeland (2014), devido ao aumento da demanda energética, novas áreas estão sendo exploradas, como regiões com reservatórios carbonáticos, que são áreas com uma grande quantidade de hidrocarbonetos, com isso, o processo de perfuração se torna mais complexo. Devido às propriedades dessas regiões serem altamente variáveis, a janela operacional fica mais estreita, com o risco de ocorrerem perdas de circulação, *kicks,* que são fluidos indesejáveis que migram da formação rochosa para dentro do poço, e *blowouts,* que são fluidos indesejáveis que migram da formação rochosa para a superfície. Nesse cenário, a perfuração convencional mostra-se incapaz de se manter dentro dos limites de perfuração estabelecidos, com isso, novas técnicas de perfuração foram desenvolvidas, como o *bullheading* que força o *kick* a retornar para a formação rochosa, sendo utilizado na técnica de perfuração PMCD, que tem se apresentado como uma solução satisfatória para estes cenários extremos (Thomas, 2001).

Segundo Liu (2023), durante o *bullheading,* controlam-se a migração do gás e a pressão do anular durante as operações de PMCD, que dependem da taxa de migração de gás e da velocidade de escoamento do gás no fluido de perfuração*,* e para que uma operação *bullheading* seja considerada bem-sucedida, é fundamental compreender as mudanças na pressão do poço e na distribuição de gás, de forma que novas fraturas não sejam criadas ou que não danifiquem a estrutura do poço.

Assim, a perfuração em cenários não convencionais (pré-sal, formações depletadas, reservatórios carbonáticos, com fraturas e formações cavernosas), apresentando características extremas, é realizada através da técnica PMCD. Portanto, quando ocorre um influxo de gás (muito comum em zonas fraturadas), as pressões anulares aumentam em decorrência da migração do gás em direção à superfície, o que pode comprometer a segurança operacional. Neste sentido, implementa-se a operação de *bullheading* que bombeia o fluido de sacrifício (água do mar), sem retorno para a superfície, em contra corrente, forçando o retorno do fluido invasor, bem como os cascalhos de volta para a formação. A partir do cenário apresentado, o objetivo primordial da presente dissertação é desenvolver modelos matemáticos para prever a operação de PMCD, empregando a abordagem de *Machine Learning.*

Este trabalho está organizado da seguinte forma: o Capítulo 1 aborda a introdução onde são apresentados: a motivação, o objetivo e a organização do trabalho, no Capítulo 2 apresenta-se a revisão bibliográfica, com informações sobre: a perfuração de poços de petróleo, os tópicos de relevância sobre o *bullheading,* os trabalhos desenvolvidos na unidade experimental e a técnica de *machine learning,* no Capítulo 3 serão introduzidos os fundamentos do *machine learning* e apresentando também as informações sobre a unidade experimental, com sua instrumentação e experimentos realizados. No Capítulo 4, Resultado e Discussões, são apresentados na sequência o Capítulo 5 com a conclusão e as sugestões para trabalhos futuros. Por fim, é apresentado no Capítulo 6 as Referências Bibliográficas, e os anexos, no Capítulo 7.

2 REVISÃO DA LITERATURA

Este capítulo apresenta de forma geral e sucinta os conceitos básicos da perfuração de poços, dos fluidos e influxos de perfuração, bem como os trabalhos desenvolvidos no Laboratório de Escoamento de Fluidos Giulio Massarani, na Universidade Federal Rural do Rio de Janeiro (LEF/UFRRJ). Em seguida, é apresentado o *machine learning,* os principais trabalhos que utilizam técnicas de *machine learning,* abordando os principais tratamentos e métricas de avaliação utilizados.

2.1 Perfuração de Poços de Petróleo

Segundo Thomas (2001), o petróleo é constituído, basicamente, por uma mistura de hidrocarbonetos que tem origem a partir da matéria orgânica depositada junto com os sedimentos. Informações sobre a utilização do petróleo pelo homem remonta a tempos antigos como, na antiga Babilônia: os tijolos eram assentados com asfalto e o betume era largamente utilizado pelos fenícios na calefação das embarcações, onde o petróleo era retirado de exsudações naturais encontrados em todos os continentes.

O início da exploração comercial na sociedade moderna data de 1859, nos Estados Unidos, logo após a descoberta de Drake de um poço de apenas 21 metros de profundidade, perfurado com um sistema de percussão movido a vapor, que produziu 2 m^3/dia. Descobriu-se que a destilação do petróleo resultava em produtos que substituem, com grande margem de lucro, o querosene obtido a partir do carvão e o óleo de baleia, que eram largamente utilizados para iluminação, com isso, se iniciava a era do petróleo.

Em 1900, no Texas, o americano Anthony Lucas, utilizando o processo rotativo, encontrou petróleo a uma profundidade de 354 metros, sendo considerado o marco importante na perfuração de poços de petróleo. Nos anos seguintes a perfuração rotativa se desenvolveu progressivamente com melhorias nos projetos, nas qualidades dos materiais utilizados e em novas técnicas de perfuração.

No Brasil, a história do petróleo se inicia em 1858, quando José Barros Pimentel extraiu mineral betuminoso para fabricação de querosene, nas margens do rio Marau, localizado na Bahia. Entretanto, as principais notícias sobre o petróleo ocorrem em Alagoas, em 1891, em função da existência de sedimentos argilosos betuminosos no litoral.

O primeiro poço brasileiro foi perfurado, em 1897, por Eugênio Ferreira Camargo, no município de Bofete, no estado de São Paulo, com profundidade de 488 metros, produzindo 0.5 m^3. Até o final de 1939, aproximadamente 80 poços tinham sidos perfurados, sendo o primeiro comercial sendo descoberto somente em 1941, em Candeias, BA. A partir de 1953, no governo Vargas, foi instituído o monopólio estatal do petróleo com a criação da Petrobras, que deu partida decisiva nas pesquisas do petróleo brasileiro.

A produção de petróleo no Brasil cresceu de 750 m^3/dia, na época da criação da Petrobras, para mais de 182.000 m^3/dia no final dos anos 90, graças aos contínuos avanços tecnológicos de perfuração e produção na plataforma continental (Thomas, 2001).

A camada pós-sal abriga a maioria das reservas do Brasil. O campo de Enchova tem a menor profundidade, 2.629 metros e o campo com maior profundidade a ser explorado até o pré-sal é Roncador, com 4.343 metros. A Figura 1 ilustra a evolução da exploração de petróleo pela Petrobras (Francisco, 2011).

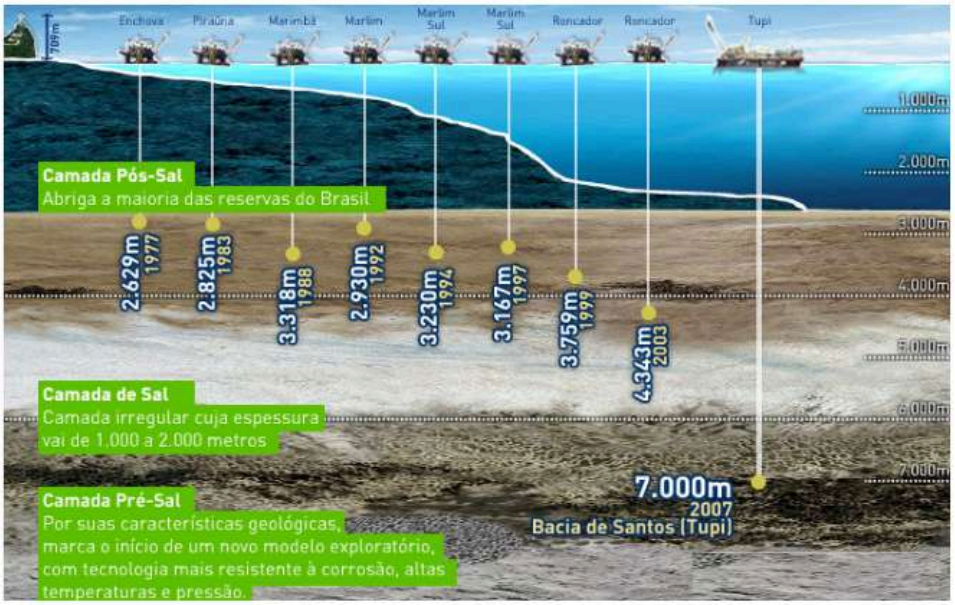

Figura *1* - Evolução da exploração de petróleo pela Petrobras. Fonte: Francisco, 2011.

A perfuração de poços é uma das etapas do processo de exploração do petróleo que ocorre após estudos geológicos e geofísicos indicarem uma possível região com reservas, e com base nesses estudos, são definidos vários fatores importantes: a profundidade da região a ser perfurada, quais os tipos de brocas serão necessárias e a composição do fluido de perfuração.

A perfuração rotativa ocorre através da sonda de perfuração, que é composta de uma torre e equipamentos que sustentam a coluna de perfuração conectada à broca. Através da rotação e do peso da coluna sobre a broca, as rochas são perfuradas, conforme mostrado na Figura 2.

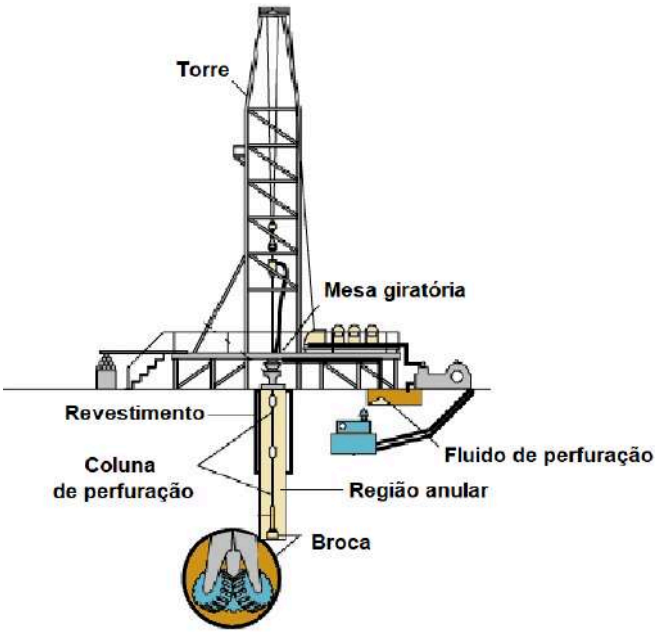

Figura *2* – Esquema de uma sonda de perfuração rotativa. Fonte: Adaptado de Chieza (2011).

Durante esse processo, o fluido de perfuração é bombeado para o interior da coluna, passando pela broca, atingindo o fundo do poço, se misturando aos fragmentos da rocha e retornando à superfície junto com esses fragmentos, através do espaço anular formado entre a coluna e as paredes do poço.

A perfuração de poços se inicia com um diâmetro, e após atingir uma determinada profundidade, a coluna de perfuração é retirada do poço e inserida uma coluna de aço que possui um diâmetro menor do que a broca, o espaço anular formado entre o tubo de aço e as paredes do poço é preenchido com cimento para evitar o desmoronamento das paredes do poço e a migração de fluidos entre as diversas zonas permeáveis atravessadas pelo poço, assegurando, dessa forma, a segurança do processo.

O número de fases e o comprimento das colunas de revestimento são determinados em função das pressões de poros e de fraturas previstas, que indicam o risco de prisão da coluna por diferencial de pressão, ocorrência de *kicks*, desmoronamento das paredes do poço ou perda de fluido de perfuração para as formações (Thomas, 2001).

A Figura 3 mostra as diversas fases caracterizadas pelas diferentes profundidades do poço e diâmetros das brocas.

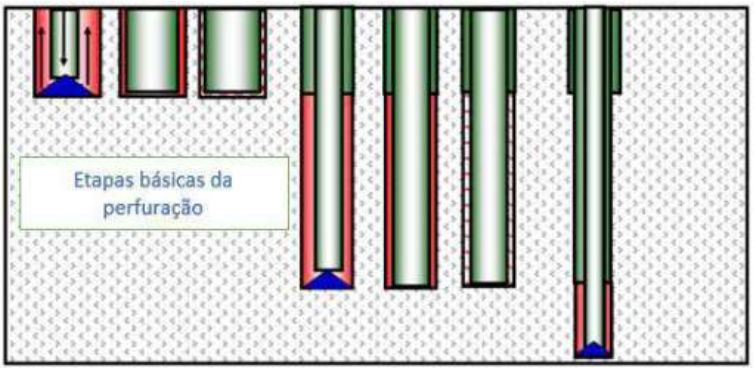

Figura *3* – Etapas básicas da perfuração. Fonte: Carvalho (2018).

Segundo Rocha (2007), com base em estudos de geopressões, que consiste no cálculo das pressões e tensões existentes nas formações rochosas, define-se a janela operacional, onde indicando-se o limite da pressão de fratura, bem como os intervalos mínimos e máximos para a densidade do fluido de perfuração em cada profundidade do poço.

A janela operacional refere-se ao intervalo entre os limites máximo e mínimo, dentro do qual o peso de fluido deve estar contido, para cada profundidade do poço. Este intervalo tem como limite superior a pressão de fratura, ou seja, o peso do fluido não pode ser alto o suficiente para fraturar a formação e induzir perdas severas de fluido para a formação. E como limite inferior, o peso do fluido não pode ficar abaixo da pressão de poros, o que permitiria um influxo do reservatório para dentro do poço, além disso, o peso do fluido não pode estar abaixo da pressão de colapso inferior, o que poderia levar à instabilidade das paredes do poço. Essas informações são demonstradas na Figura 4.

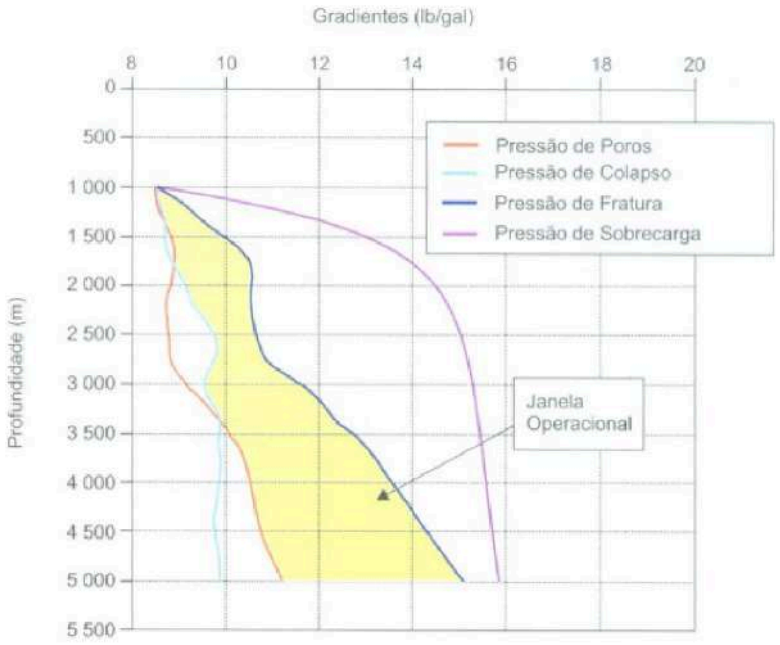

Figura *4* - Curva de geopressões Fonte: Rocha, (2007).

2.1.1 Técnicas de Perfuração de Poços de Petróleo

Segundo Ghauri (2014), a técnica de *Pressurized Mud Cap Drilling* (PMCD) é uma variante da tecnologia *Managed Pressure Drilling* (MPD) que reduz o risco do processo de perfuração em formações rochosas altamente fraturadas, sendo uma alternativa mais eficaz do que as operações convencionais, quando ocorrem altas perdas durante a perfuração.

2.1.1.1 *Managed Pressure Drilling* - MPD

Segundo Kaasa *et al.* (2011), o MPD é a perfuração sob pressão gerenciada que tem como objetivo controlar a pressão anular de fundo de poço, durante as operações de perfuração de poços de petróleo. O autor propôs um modelo hidráulico simples que estima a pressão de fundo de poço em tempo real e um algoritmo de controle de *feedback* que automatiza a abertura e fechamento da válvula *choke* para manter a contrapressão desejada, conforme mostrado na Figura 5.

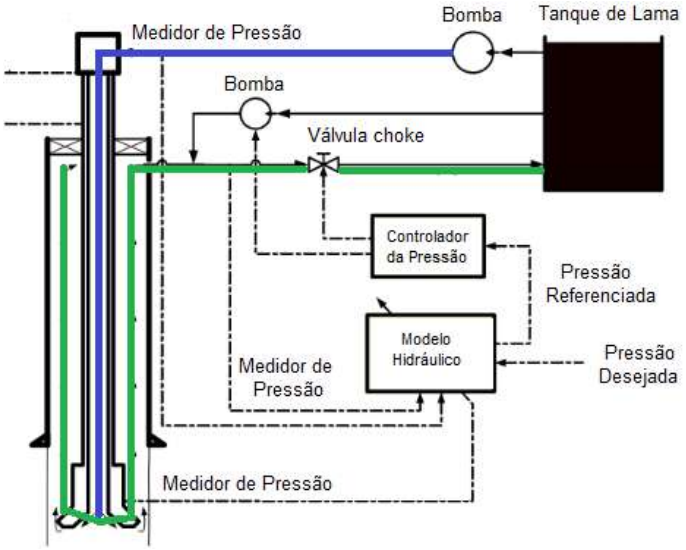

Figura *5* **–** Esquema simplificado de um sistema MPD. Fonte: Adaptado de Kaasa *et al.* (2011).

O fluido bombeado por dentro da coluna de perfuração está representado pela linha em azul, que ao atravessar a broca, retorna para a superfície pelo espaço anular, representado pela linha verde. O modelo hidráulico recebe as informações sobre a pressão de bombeamento do fluido pela coluna de perfuração, a pressão do fundo do poço e a pressão da saída do fluido pelo anular, e com base na pressão desejada o algoritmo de controle emprega como elementos finais de controle tanto a bomba quanto a válvula *choke*. Os fluxos dessas informações estão representados pelas linhas pontilhadas.

2.1.1.2 *Pressurized Mud Cap Drilling* - PMCD

Segundo Romualdo (2021), o PMCD é uma variação da operação de MPD que alimenta uma lama leve chamada de L*ight Annular Mud* (LAM) no topo da região anular, enquanto um fluido de sacrifício chamado de S*acrificial Fluid* (SAC) é bombeado através da coluna de perfuração, a fim de transportar os fragmentos da rocha perfurada e fluidos invasores de volta para a formação (Figura 6).

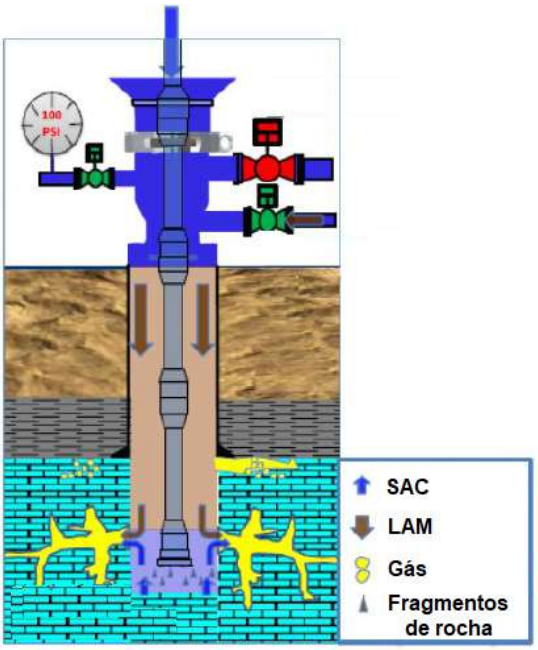

Figura *6* – Esquema de PMCD. Fonte: Adaptado de Ghauri (2014).

Segundo Ghauri (2014), as operações de perfuração são executadas normalmente até a ocorrência de zonas de perda ou fraturas. Pequenas perdas serão tratadas com *Lost Circulation Material* (LCM), mas quando ocorrerem perdas insustentáveis, a operação será convertida em *Pressurized Mud Cap Drilling* (PMCD).

Após a identificação do *kick*, é executada a operação *bullheading* para forçar os fluidos indesejáveis a retornarem para a formação rochosa. Na Figura 6, o fluido de perfuração SAC é inserido por dentro da coluna de perfuração, sendo representado pelas setas na cor azul, o fluido de perfuração LAM é inserido dentro do espaço anular do poço, representado pelas setas na cor marrom, e o gás contido na formação rochosa está representado pelas setas na cor amarela, que representa o *kick*.

2.2 Fluido de perfuração

Segundo Thomas (2001), os fluidos de perfuração são misturas complexas de sólidos, líquidos, gases e outros produtos químicos que devem ser especificados de forma a garantir uma perfuração rápida e segura, logo, é desejável que o fluido de perfuração apresente as seguintes características: ser quimicamente estável, manter a estabilidade das paredes do poço, ser facilmente separável dos cascalhos que são carreados para a superfície, permitir que as partículas sólidas fiquem em suspensão quando estiver em repouso, ser inerte em relação às rochas produtoras, aceitar qualquer tratamento físico ou químico, ser bombeável, apresentar baixo grau de corrosão e de abrasão em relação aos equipamentos utilizados na perfuração, facilitar interpretações geológicas do material retirado do poço e apresentar custo compatível com a operação de perfuração.

Além de transportar os fragmentos da rocha perfurada, o fluido de perfuração tem grande importância na lubrificação, resfriamento e limpeza de todos os equipamentos conectados por todo o caminho percorrido, além de exercer pressão hidrostática sobre as

formações rochosas, de modo a evitar o influxo de fluidos indesejáveis (*kicks*) e estabilizar as paredes do poço (Thomas, 2001).

2.3 Influxos de perfuração

Segundo Russano (2014), é importante que haja um controle da pressão anular, porque a operação fora de uma janela operacional causará problemas de segurança. Se a pressão anular de fundo encontra-se maior que a pressão de poros observa-se a perda do fluido de perfuração para a formação rochosa, que é chamada de perda de circulação.

A invasão do fluido de perfuração pode provocar danos irreversíveis ao reservatório, por reduzir sua produtividade. Em reservatórios de óleo pesado os problemas podem ser especialmente críticos devido à geração de emulsões estáveis entre o óleo e o filtrado. Enquanto a invasão em reservatórios de óleo leve é menos crítica devido a boas propriedades de mobilidade. (Calçada *et al.*, 2005).

Segundo Rocha (2007), caso a pressão de poros se torne maior que a pressão dentro do poço, poderá ocorrer um influxo do fluido da formação para o poço. Esta típica ocorrência indesejada chama-se *kick* e pode levar a grandes perdas de tempos não produtivos (NTP). Em casos mais severos e de total descontrole, o *kick* pode atingir a superfície, resultando no chamado *blowout,* que gera consequências desastrosas tais como a destruição total da plataforma, danos ao meio ambiente e aos seres humanos. Por outro lado, pressões dentro do poço muito maiores que a pressão de poros podem levar à prisão da coluna, fenômeno referido como prisão por pressão diferencial.

Diante de formações impermeáveis, o diferencial de pressão entre o poço e a formação não resulta em fluxo da formação para dentro do poço ou prisão por pressão diferencial. Entretanto, outros problemas podem ser acarretados, tais como instabilidade das formações que pode levar ao desmoronamento total ou parcial das paredes do poço, acarretando o aprisionamento da coluna de perfuração.

2.4 Trabalhos desenvolvidos no Laboratório de Escoamento de Fluidos Giulio Massarani (LEF), na Universidade Federal Rural do Rio de Janeiro (UFRRJ)

A unidade experimental, instalada no Laboratório de Escoamento de Fluidos Giulio Massarani, do Departamento de Engenharia Química, no Instituto de Tecnologia da Universidade Federal Rural do Rio de Janeiro (LEF/UFRRJ), vem contribuindo para os estudos relacionados ao processo de perfuração de poços de petróleo. Ao longo dos anos, foram desenvolvidos os seguintes estudos:

Vieira (2009) foi o pioneiro no laboratório, com o desenvolvimento de um modelo matemático não-linear para representar o sistema de perfuração e construiu uma unidade experimental capaz de representar as características mais relevantes da perfuração. Além disso, implementou uma estratégia de controle Proporcional Integral (PI) para regular a pressão anular de fundo, utilizando a vazão de alimentação como variável manipulada.

Freitas (2013) introduziu os estudos acerca do controle da pressão anular de fundo em cenários de influxos, realizando simulações e experimentos de controle em cenários de *kick* de líquido, através da manipulação da válvula *choke*.

Russano (2014) também utilizou a estratégia de controle para regular a pressão anular de fundo, através da manipulação da válvula *choke,* com foco na perda de circulação durante a perfuração, por meio de simulação e experimentos.

Oliveira *et al.* (2015) implementou a estratégia de compensação de tempo morto para o controle da pressão anular de fundo e comparou os resultados com os da estratégia de controle PI.

Patrício (2016), desenvolveu um modelo matemático que é uma estratégia de controle para regular o distúrbio de *kick* de gás, realizando simulações e testes experimentais utilizando a estratégia de controle PI e a estratégia de reconfiguração de controle (*feedback-feedforward*).

Costa (2016) estudou o controle da pressão anular de fundo no bombeio de diferentes fluidos durante a cimentação, utilizando a estratégia *feedback* PI para regular a pressão anular de fundo.

Carvalho (2018) estudou o controle de pressão nas operações de PMCD sob distúrbios de *kick* de gás, propondo um modelo matemático bifásico baseado em equações de conservação de massa e movimento e realizou testes na unidade experimental. Utilizou uma estratégia de reconfiguração da variável manipulada, em que o índice de abertura da *choke* era manipulado na operação MPD e a vazão de *bullheading* era manipulada na operação PMCD.

Ribeiro (2018) propôs a implementação de controladores baseados em redes neuronais para regular a pressão anular de fundo, onde diferentes distúrbios foram avaliados, como o *kick* de gás, perda de circulação e o procedimento de conexão de tubos. A partir dos 1566 dados experimentais, o algoritmo foi treinado, validado e testado.

Silva (2019) fez modificações na unidade experimental para mimetizar a solubilização do gás no fluido de perfuração, uma vez que o sistema utilizado foi água/ar comprimido.

Ramalho (2023) propôs a implementação de *machine learning* com o objetivo de estudar a perda de circulação, utilizando dados de poços reais e dados obtidos da unidade experimental, implementando estudos de regressão e classificação.

2.5 M*achine learning*

Segundo Géron (2021), o *machine learning* é a ciência e a arte de programar computadores para que sejam capazes de aprender com os dados. A técnica é indicada para: problemas que exigem muitos ajustes finos ou longas listas de regras, problemas complexos para os quais não há uma boa solução disponível e para ambientes e dados inconstantes ou flutuantes. Vale ressaltar que o *machine learning* simplifica as regras, têm um desempenho melhor do que a abordagem tradicional e é adaptável aos dados inconstantes.

Existem tantos tipos diferentes de sistemas de aprendizado de máquina que podem ser classificados com os seguintes critérios: serem ou não treinados com supervisão humana (supervisionado/não-supervisionado), se aprendem ou não gradativamente em tempo real (*batch* e online) e se funcionam simplesmente comparando novos pontos de dados com pontos de dados conhecidos, ou se detectam padrões em dados de treinamento e criam um modelo preditivo (baseado em instâncias ou baseados em modelos).

A classificação de aprendizado supervisionado/não-supervisionado tem relação com a supervisão humana ou não durante o treinamento, ou seja, está diretamente ligada aos dados de treino possuírem ou não rótulos definidos no modelo.

No caso do aprendizado supervisionado, o algoritmo deve possuir a capacidade de generalizar o conhecimento a partir de dados disponíveis com rótulos (possui variável alvo), de modo que possa ser usado para prever novos casos rotulados (dados inéditos). Já no caso do aprendizado não-supervisionado, o algoritmo é capaz de agrupar dados em categorias *clusters*

usando métodos automatizados em dados que não foram classificados ou rotulados anteriormente (Géron, 2021).

2.5.1 *Machine learning* na perfuração de poços de petróleo

Seguem os estudos na área de perfuração de poços de petróleo com aplicação de técnicas de *machine learning* mais relevantes para o desenvolvimento da presente dissertação de mestrado.

Kamyab *et al.* (2010) desenvolveram o estudo da detecção precoce de *kicks* usando análise de dados em tempo real, aplicando modelos baseados em rede neuronal.

Agostin *et al.* (2017) apresentaram uma metodologia para definir materiais ótimos para controle de perdas de fluidos, com o objetivo de minimizar esses eventos durante a perfuração, aplicando a técnica *Naive Bayes* para treinar o modelo.

Alouhali *et al.* (2018) desenvolveram um algoritmo de detecção automatizada de *kick* usando as técnicas: *decision tree*, *K-Nearest Neighbor* (KNN), *Artificial Neuronal Network* (ANN) e *Baysean Network*.

Xie et al. (2018) efetuaram a análise de *big data* para monitoramento de *kick* em projetos complexos de perfuração subaquática, utilizando as técnicas: ANN e *Generic Algorithm* (GA)

Al-Hameedi *et al.* (2018) previram volumes de perda de fluido, Densidade de Circulação Equivalente (ECD) e Taxa de Penetração (ROP), aplicando técnicas de estatísticas avançadas, como o *Principal Component Analysis* (PCA).

Abbas *et al.* (2019) desenvolveram um sistema especialista capaz de diagnosticar a perda de circulação e sugerir soluções eficazes com base em parâmetros de operação e características do fluido de perfuração, aplicando ANN e *Support Vector Machine* (SVM).

Fjetland et al. (2019) desenvolveram um algoritmo para detectar *kick* e estimar a taxa de influxo no poço empregando redes neuronais recorrentes.

Amed *et al.* (2020) fizeram a previsão de zonas de perda de circulação usando duas técnicas de inteligência artificial: função de base radial (RBF) e SVM.

Alkinani *et al.* (2020) aplicaram diversas técnicas de *Machine Learning*, incluindo SVM, árvores de decisão, regressão logística, ANN e árvores de conjunto, para desenvolver um modelo de previsão de perda de circulação.

Mardanirad et al. (2021) desenvolveram um algoritmo *deep learning* para distinguir com precisão as severidades de perdas de circulação em operações de perfuração de petróleo, usando *Convolutional Neural Network* (CNN), *Gated Recurrent Unit* (GRU) e *Long-Short Term Memory* (LSTM).

Sabah *et al.*, (2021) implementaram três algoritmos de busca heurística, sendo: GA, *Particle Swarm Optimization* (PSO) e algoritmo *Cuckoo Search* (COA) para prever a perda de circulação, acoplados as seguintes técnicas: redes neurais de *Multilayer Perceptron* (MLP), SVM e *Least Squares Support Vector Machines* (LSSVM)

2.5.2 Validação do modelo

Segundo Géron (2021), a única forma de mensurar até que ponto um modelo generalizará bem em casos novos, é testá-lo na prática, sendo uma das melhores opções a divisão dos dados em dois conjuntos: conjunto de treino e conjunto de teste, onde normalmente, 80% dos dados é destinado ao treino e 20% ao teste (inferência), mas esses valores variam, a depender da quantidade de dados disponível e da estratégia de validação escolhida.

O modelo deve ser treinado com base nos dados do conjunto de treino, ajustado no conjunto de validação e, por fim, avaliado com base nos dados do conjunto de teste, sendo a taxa de erro em novos casos denominada de erro de generalização. Ao analisar o modelo no conjunto de teste, é possível obter uma estimativa desse erro, indicando quão bem o modelo se comportará em instâncias não vistas anteriormente. Se o erro de treinamento for baixo, mas o erro de generalização for alto, isso sugere o sobreajuste (*overfitting*) dos dados de treinamento (Géron, 2021).

Segundo Xu & Goodacre (2018), a validação cruzada é um dos métodos mais comuns utilizados para dividir o conjunto de dados do treino para a seleção de modelos. Ela divide os dados em k partes (*k-folds*), usando uma parte como conjunto de validação, enquanto treina o modelo nas outras k-1 partes restantes. Esse processo é repetido k vezes, registrando o desempenho preditivo. O parâmetro ótimo é escolhido com base no melhor desempenho médio. A validação cruzada permite que se obtenha não somente uma estimativa de desempenho do modelo, como também um cálculo da precisão dessa estimativa, ou seja, o valor do desvio padrão (Géron, 2021).

2.5.3 Principais tratamentos

Segundo Géron (2021), se os dados de treinamento estiverem cheios de erros, *outliers* e ruídos, o sistema terá mais dificuldade para detectar os padrões básicos, logo é necessário tratar os dados antes de serem utilizados para treinar o modelo, com isso, se algumas instâncias são *outliers*, estes podem ser descartados ou serem tratados.

Caso falte algumas características para algumas instâncias, como campos nulos ou vazios, pode-se decidir ignorar completamente esta instância, caso tenha uma quantidade grande de campos nulos ou decidir preencher os campos nulos com algum valor como: média, moda, etc.

Pode-se aplicar a *feature engineering,* em que as características mais importantes são selecionadas para treinar o modelo, podendo ser utilizadas para combinar características existentes a fim de obter as características mais úteis, ou podem ser utilizadas para criar características, ao coletar dados novos.

Os algoritmos de *machine learning* não costumam lidar com dados textuais, por isso, os dados não numéricos precisam ser codificados para números.

Além disso, os dados são usualmente transformados aplicando-se a normalização e tratamentos de assimetria e curtose.

2.5.3.1 Normalização dos dados

Segundo Géron (2021), os algoritmos de *machine learning* não funcionam bem quando atributos numéricos de entrada possuem escalas muito diferentes, com isso, há duas técnicas comuns para que todos os atributos tenham a mesma escala: o *Min-Max Scaling,* onde o dado é transformado com base em seus valores mínimo e máximo, e o *Standard Normalization,* onde o dado é transformado com base nos valores da sua média e do desvio padrão.

2.5.3.2 Tratamentos de assimetria e curtose

Jonson e Witchern (2007), desaconselharam a utilização de dados não normais, principalmente para modelos estatísticos, já que a normalidade de resíduos é assumida como premissa, podendo gerar conclusões erradas, caso a distribuição dos dados seja muito fora da

normalidade. Desse modo, algumas transformações podem ser realizadas nos dados não-normais, como aplicação de raiz quadrada, logaritmo ou *box-cox*.

São calculados os valores da assimetria e curtose, para identificar se os dados possuem uma distribuição normal, ou se precisam de transformação. A assimetria indica o nível de deslocamento da distribuição, identificado através do cálculo do coeficiente de Fisher-Pierson, e a curtose representa o nível de achatamento da distribuição, identificado também através dos conceitos de Fisher. A assimetria e a curtose ideais devem estar mais próximas de 0.

Ademais, pode ser utilizado o teste de normalidade de Shapiro-Wilk em que a normalidade é considerada a hipótese nula. Através da análise do p-valor, quando o p-valor for maior que 0,05, então não há evidência suficiente para rejeitar a hipótese nula de normalidade, caso contrário, a hipótese nula pode ser rejeitada e assume-se que os dados não seguem uma distribuição normal. (Jonson e Witchern, 2007).

2.5.4 Métricas de avaliação

A tarefa de regressão (Huyen, 2022) emprega as seguintes métricas para avaliação da performance dos resultados: *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), *Sum of Squared Error* (SSE) e o coeficiente de determinação (R^2).

Cada métrica fornece informações diferentes e importantes a respeito da qualidade das previsões. O MSE atribui um peso maior aos erros altos, permitindo identificar se o modelo está errando muito, mesmo que seja para alguns casos; o RMSE traz a mesma informação, entretanto, na mesma escala do dado, o que permite compará-lo com a média, mediana e desvio padrão da variável alvo, a fim de quantificar com maior clareza o erro; o R^2 quantifica o quão bem o modelo se ajustou aos dados, apresentando como objetivo o mais próximo de 1. Já para as outras métricas, o objetivo sempre é alcançar valores próximos a 0.

Segundo Scikit-Learn (2019), a função de perda também pode ser uma forma de avaliar o modelo. As funções de perda mais utilizadas e comuns são *LogLoss* e *SoftMax.* A cada iteração busca-se reduzir a função de perda, objetivando chegar o mais próximo possível de 0, visto que a função de perda representa um erro do modelo (Huyen, 2022).

2.5.5 *Extreme Gradient Boosting* - XGBoost

O *Extreme Gradient Boosting* (XGBoost)*,* criado por Tianqi Chen, em 2014, tornou-se uma biblioteca altamente aplicada de código aberto para solução de problemas utilizando *machine learning*. Em 2015, a maioria das soluções vencedoras de desafios publicadas no blog do *Kaggle*, utilizaram o XGBoost. É nesta plataforma que muitas empresas, como por exemplo Google, Microsoft e NVIDIA, buscam soluções de *machine learning* e análise de dados. (Chen & Guestrin, 2016).

O XGBoost utiliza a técnica de *tree boosting* para previsões, principalmente em tarefas de classificação e regressão, que consiste em um conjunto de várias árvores de decisão construídas e treinadas de forma iterativa, buscando otimizar sempre os erros da árvore anterior.

Na Figura 7 ilustra a arquitetura do modelo XGBoost*,* para prever se uma determinada pessoa gostará ou não do jogo X, sendo que cada folha retorna uma pontuação para a predição baseada nas características inseridas, em seguida, unem-se várias árvores diferentes, cada uma com uma pontuação para a predição, sendo que a previsão se dá a partir da soma da pontuação de cada folha.

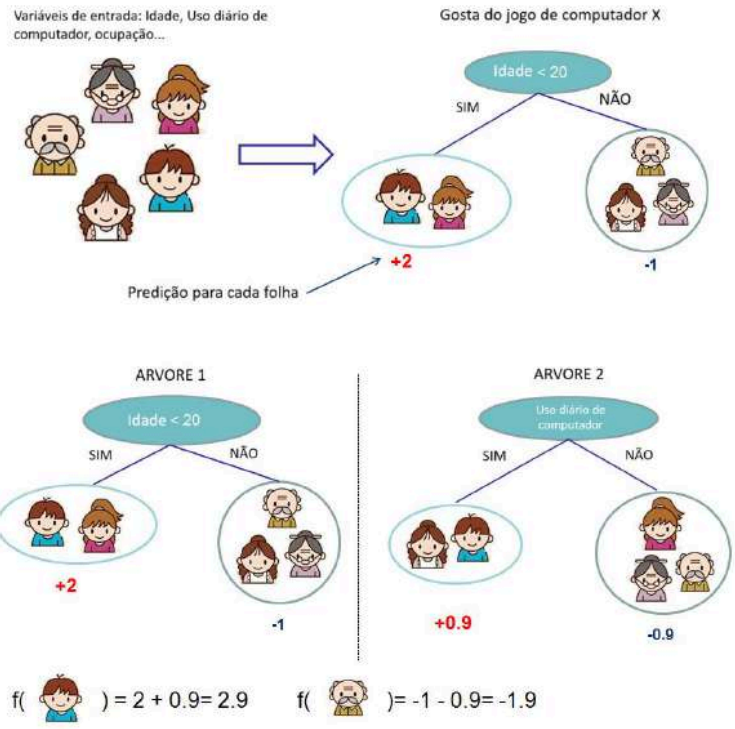

Figura *7* – Exemplo de XGBoost. Fonte: XGBoost *Documentation*.

Segundo Chen & Guestrin, (2016), normalmente, é impossível enumerar todas as possíveis estruturas de árvores. Portanto, é utilizado um algoritmo que começa com uma única folha e adiciona iterativamente ramos à árvore. Após cada divisão, a perda de informação é avaliada, onde é considerada como a divisão afeta a qualidade da árvore, levando em conta as informações das divisões das folhas da esquerda e da direita. Uma boa divisão é caracterizada por uma melhoria significativa na redução da função de perda, na obtenção de previsões mais precisas e na capacidade de generalização do modelo.

Esta busca iterativa das folhas direita e esquerda é chamada de *greedy algorithm*, sendo muito eficaz ao examinar todas as possíveis divisões, mas se torna lento quando há grande quantidade de dados ou configurações distribuídas. Para lidar com essas situações, existe a opção de utilizar um algoritmo aproximado. Nesse algoritmo, primeiro são propostas divisões candidatas com base em percentis das variáveis. Em seguida, as características são agrupadas em intervalos com base nessas candidatas, onde as estatísticas são calculadas e a melhor divisão nos nós é escolhida com base nessas estatísticas. Existem duas versões aproximadas: a versão global que efetua todas as divisões candidatas no início e as usa em todos os níveis da árvore, sendo a versão mais rápida, mas geralmente requerendo mais candidatas e a versão local que propõe novas candidatas após cada divisão. A versão local mais adequada para árvores profundas, tendo em vista que essas árvores possuem um grande número de camadas (ou níveis) e, portanto, muitos nós e divisões. Cada camada adicional na árvore agrega complexidade ao modelo, permitindo que a árvore capture padrões mais intrincados e específicos nos dados de treinamento.

2.5.6 Optuna

Apresentado por Akiba *et al.* (2019), o Optuna é um programa computacional de otimização de ajuste de hiperparâmetros, projetado com o princípio *define-by-run*, sendo o primeiro deste tipo, onde a definição dos espaços de busca e a avaliação das configurações de

hiperparâmetros são integradas ao próprio código do usuário, tornando o processo mais flexível e dinâmico.

O programa computacional está disponível sob a licença de uso livre, e possui como uma das propostas principais a *Application Programming Interface* (API) definida por execução, que permite aos usuários construir dinamicamente o espaço de busca de hiperparâmetros. O *software* formula a otimização de hiperparâmetros como um processo de minimização/maximização de uma função objetivo, recebendo um conjunto de hiperparâmetros como entrada e retornando sua pontuação de validação.

O Optuna possui uma implementação de estratégias eficientes para estimação de desempenho, obtido através da busca e consequente descarte dos parâmetros ruins (poda), que pode ser feita através da amostragem relacional, que é caracterizada pela escolha de configurações de hiperparâmetros com base nas configurações anteriores, usando informações acumuladas, direcionando a busca para regiões promissoras, ou através da amostragem independente, onde a escolha de cada configuração é realizada de forma isolada, sendo eficiente, mas menos direcionada, ignorando seu histórico (Akiba *et al,* 2019).

3 MATERIAL E MÉTODOS

Neste capítulo, apresenta-se o algoritmo XGBoost, a metodologia experimental bem como dados de poços reais em um cenário de perfuração de poços de petróleo empregando a técnica PMCD com o objetivo de treinar um modelo matemático baseado em *machine learning*.

3.1 Etapas no processo de *machine learning*

Para a construção do *machine learning* foram desenvolvidos programas com o uso da linguagem *Python* e utilizado a plataforma Colab para seu processamento, com o objetivo de treinar um modelo que possa aprender com os dados de *bullheading* e prever seu comportamento.

As etapas da construção do *machine learning*, são mostradas na Figura 8: coleta de dados, pré-processamento dos dados coletados, treinamento dos dados processados e, por fim, avaliação do modelo candidato, gerando o modelo escolhido.

Figura *8* – Etapas para a construção do modelo via *machine learning*. Fonte: Adaptado de Shade (2018).

3.1.1 Coleta dos dados

O desenvolvimento de modelos matemáticos baseados em *machine learning* para descrever a perfuração do tipo PMCD utilizou dados da unidade experimental do LEF/UFRRJ, obtidos por Carvalho (2018) e dados de poços reais disponíveis em Jayah *et al.* (2013), Zein *et al.* (2017) e Wattanasuwankorn *et al.* (2014).

3.1.1.1 Poços reais

Foram coletados dados de poços reais para a operação PMCD Jayah *et al.* (2013), apresentado na Figura 9, de Zein *et al.* (2017), na Figura 10 e de Wattanasuwankorn *et al.* (2014), na Figura 11.

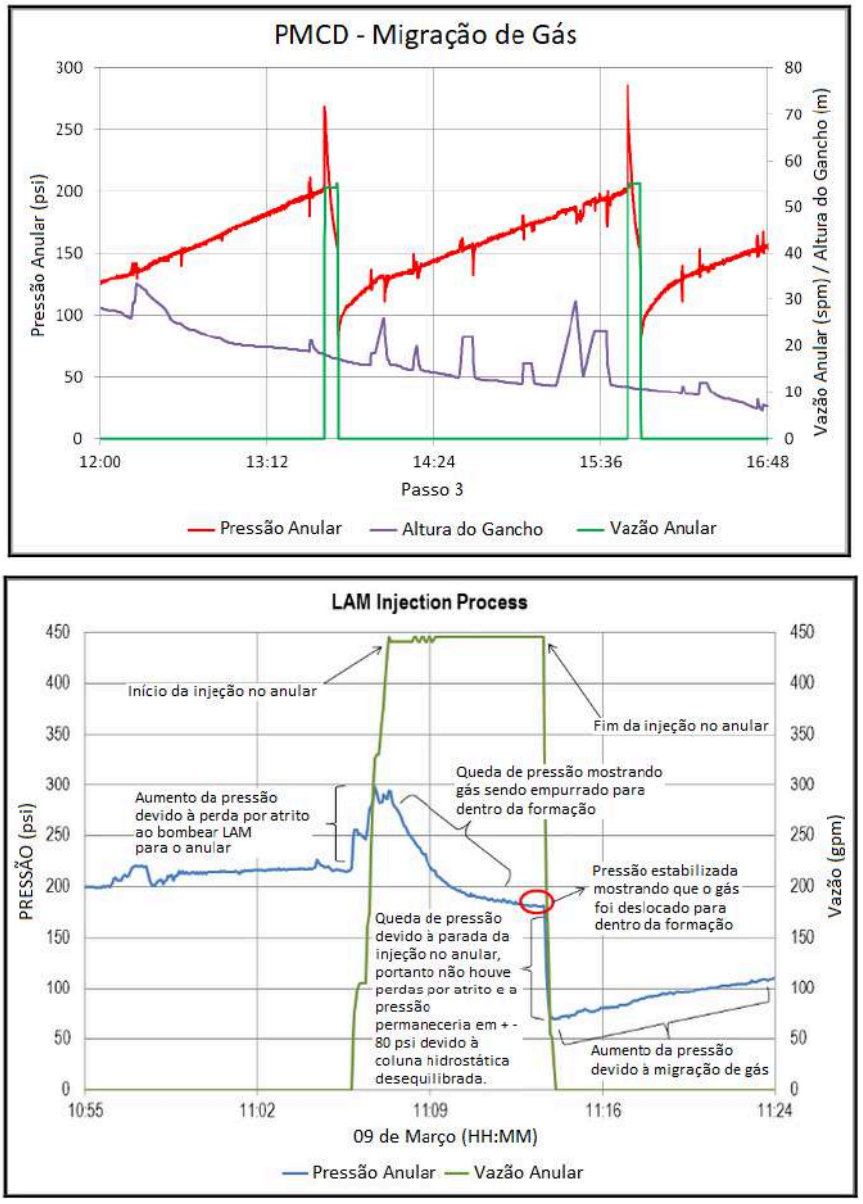

Figura *9* – Dois ciclos de *bullheading*. Fonte: Adaptado de Jayah *et al.* (2013).

Na Figura 9, o gráfico superior mostra o processo de perfuração PMCD com dois ciclos de *bullheading,* objetivando reduzir os picos de pressão no anular acima de 250 psi, representando o experimento 1, e no gráfico inferior é apresentado mais um procedimento de PMCD com a injeção de LAM para reduzir a pressão anular forçando o gás a migrar de volta para a formação. Na sequência, há novamente a migração de gás para o anular e a operação é retomada.

Figura 10, ilustra três ciclos de *bullheading,* com picos de pressão no anular acima de 150 psi, representando o experimento 2.

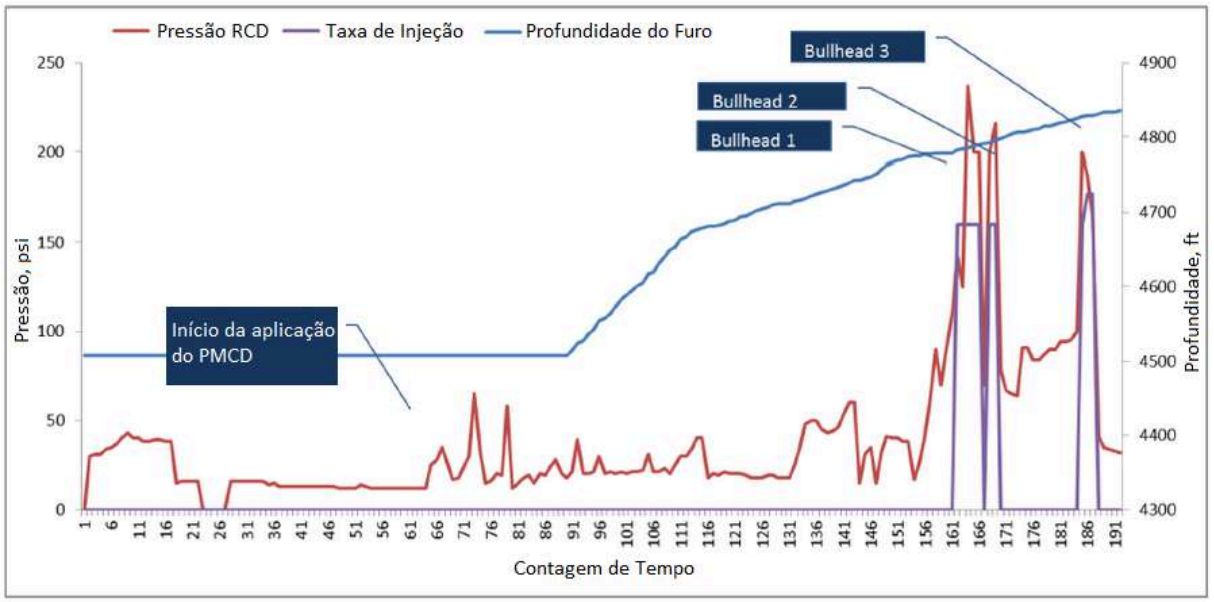

Figura *10* – Três ciclos de *bullheading*. Fonte: Adaptado de Zein *et al.* (2017).

Na Figura 11, o gráfico superior mostra ciclos de *bullheading* durante a operação de PMCD, com pressão superior a 100 psi, representando o experimento 3 e no gráfico inferior são mostrados o início e o fim do processo de *bullheading,* representando o experimento 4.

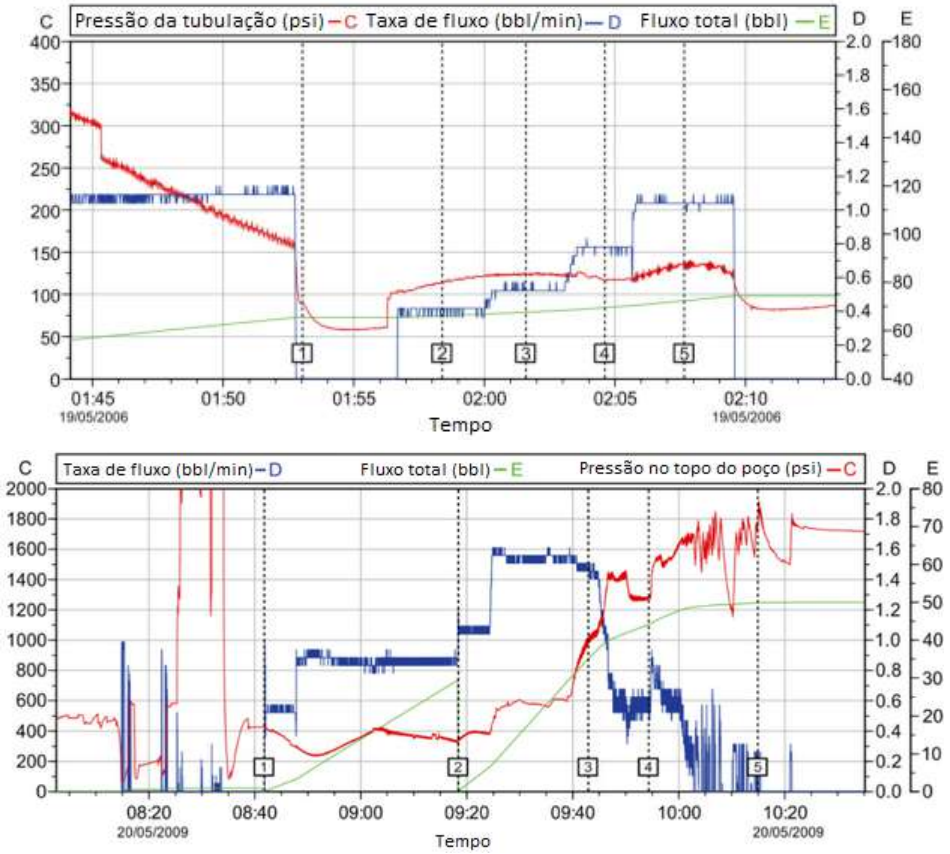

Figura *11* – Ciclos de *bullheading*. Fonte: Adaptado de Wattanasuwankorn *et al.* (2014).

Para coletar as informações dos trabalhos de Jayah *et al.* (2013), Zein *et al.* (2017) e Wattanasuwankorn *et al.* (2014), foi utilizado a ferramenta automeris.io, conforme exemplo mostrado na Figura 12, aplicado aos dados de Jayah *et al.* (2013).

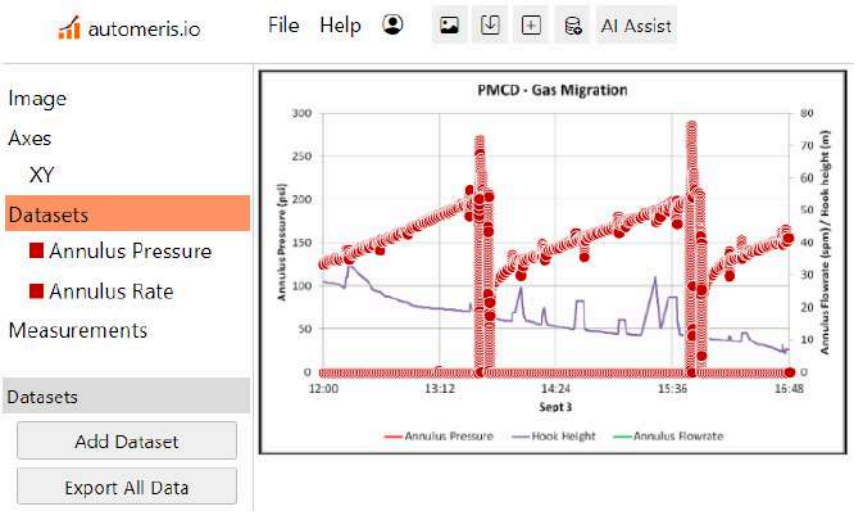

Figura *12* – Ferramenta automeris.io. Fonte: A autora.

3.1.1.2 A perfuração na unidade experimental

Os estudos do processo de perfuração de petróleo foram desenvolvidos na unidade experimental, localizada no Laboratório de Escoamento de Fluidos Giulio Massarani (LEF), do Departamento de Engenharia Química, no Instituto de Tecnologia da Universidade Federal Rural do Rio de Janeiro (UFRRJ), na Figura 13.

Figura *13* – Foto da unidade experimental. Fonte: A autora.

A estrutura hidráulica foi construída de forma a retratar os processos de perfuração MPD e PMCD*, s*endo utilizado a água como fluido de perfuração e o ar comprimido como *kick.* A unidade é formada pela região de injeção do fluido (coluna), a região do reservatório e a região do anular.

A região de injeção do fluido é composta por um conjunto de equipamentos compreendidos entre os tanques até a interseção das conexões entre a região do reservatório e a região do anular, esses equipamentos são: tanques que armazenam o fluido; sensor de nível; bomba helicoidal de deslocamento positivo, com pressão de recalque igual a 170 psi e o medidor de vazão mássica da marca Micro Motion.

A região anular é composta por uma mangueira flexível projetado helicoidalmente com 270 metros de comprimento, com uma polegada de diâmetro interno, sendo fabricada em borracha de acrilonitrilo butadieno, tendo a pressão recomendada de trabalho de até 300 psi, podendo ser utilizada com água, óleo ou gás. Sua inclinação é devido à necessidade do deslocamento ascendente da fase gasosa em um sistema fechado, simulando um poço fechado. Em cada volta em torno do suporte do anular há 5 metros de mangueira com uma inclinação de 88,8° em relação ao plano vertical. Transdutores de pressão estão instalados no início do anular e próximo à válvula *choke.*

A região anular também possui outros equipamentos como: o transdutor de pressão diferencial utilizado na identificação da presença de gás, que ocorre quando essa pressão diferencial diminui, tendo em vista que as perdas por atrito devem ser menores quanto maior for a presença de gás, o medidor de vazão mássica da marca Metroval, que monitora as perturbações como perda de circulação (redução de vazão) e *kick* (aumento da vazão) e a válvula de controle proporcional chamada de válvula *choke,* cuja manipulação do índice de abertura permite controlar a pressão no anular.

A região do reservatório é composta por um conjunto de equipamentos ligados ao tanque pressurizado que simulam: tanque de pressão, com volume máximo de 50 litros, suportando até 300 psi; transdutor de pressão, válvulas *on/off* para entrada e saída de ar, tornando possível controlar e ajustar a pressão; válvulas *on/off* instaladas na lateral do tanque para saída do líquido, que migrou para o tanque, devido a operação de *bullheading*; válvula de segurança, para que a pressão no tanque não exceda seu limite; válvula de controle proporcional, que faz o contato do tanque com a região anular permitindo testar diferentes permeabilidades e níveis de fratura; sensor de nível instalado na parte superior do tanque de pressão.

Na Figura 14 é apresentado o esquema do aparato experimental, onde as setas azuis representam a operação de perfuração MPD, as setas vermelhas representam as operações de PMCD com *bullheading* e as setas brancas representam o gás proveniente do reservatório.

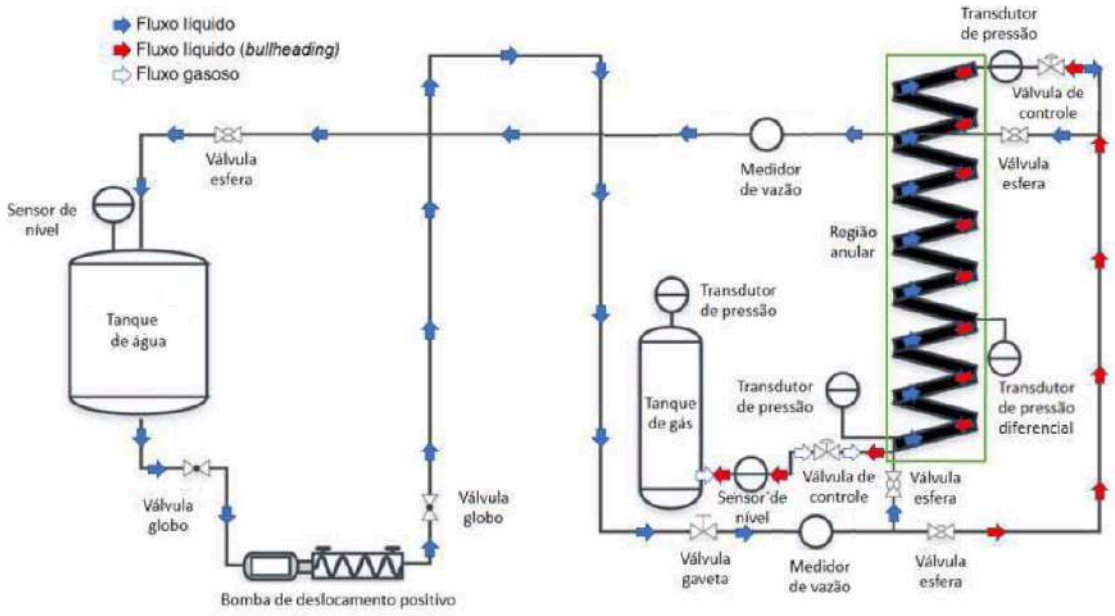

Figura *14* – Esquema dos fluxos da unidade experimental. Fonte: Adaptado de Carvalho (2018).

O monitoramento e controle da unidade experimental é realizado através do programa Petrocontrol T.M. (Martins et al. ,2019) desenvolvido em linguagem C++, que faz leitura em tempo real e envia os sinais de controle para os instrumentos instalados na unidade, a fim de atuarem nas etapas dos experimentos realizados (Figura 15).

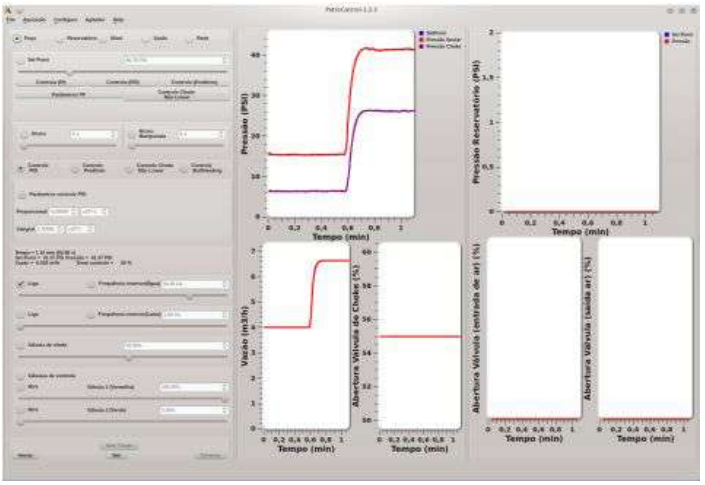

Figura *15* – Tela do sistema. Fonte: Adaptado de Ribeiro (2018).

Segundo Carvalho (2018), por meio do programa é possível estabelecer a forma de interação entre a região anular e a região do reservatório, podendo induzir o *kick* ou a perda de circulação, controlar as pressões da região anular, da região do reservatório a técnica de perfuração, especificar as aberturas das válvulas e a manipulação das variáveis. É possível escolher diferentes formas de controle: controle PID, controle preditivo e controle não linear, usando estratégias *feedback*, *feedforward* ou um esquema de controle por reconfiguração.

Carvalho (2018), apontou que durante a operação de perfuração MPD e o PMCD, é importante manipular a variável que possui maior influência em cada etapa, para que se possa controlar a pressão anular durante os experimentos. Para a etapa em que se opera com a técnica MPD utiliza-se o índice de abertura da válvula *choke* como variável manipulada. Durante a etapa de operação de PMCD, a ideia é manipular a frequência de rotação da bomba de forma a manipular a vazão do fluido de *bullheading*, causando assim um impacto mais relevante para se atingir o valor desejado na pressão anular. A Figura 16 mostra os fluxos de água representado pelas setas na cor vermelha e o fluxo de ar comprimido representado pelas setas na cor azul, durante um processo de perfuração com distúrbio de *kick* de gás

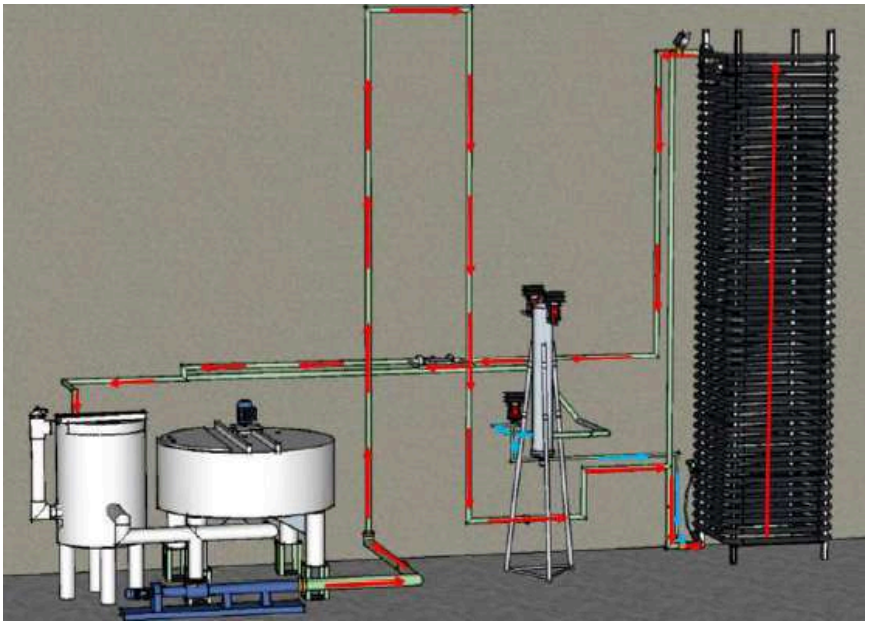

Figura *16* – Esquema da unidade experimental. Fonte: Ribeiro (2018).

A presente dissertação de mestrado foi desenvolvida com base nos experimentos realizados por Carvalho (2018), sendo eles: injeção/migração de gás, operação de *bullheading e* identificação de *kick* (Figura 17).

No experimento de injeção/migração de gás, o *kick* é simulado após o anular ser preenchido por água e a bomba ser mantida desligada, após esse distúrbio a região anular é fechada e busca-se analisar o comportamento da pressão durante a migração de gás. Seguem as etapas do experimento: primeiro o anular precisa ser preenchido com o fluido água e com bomba desligada, depois ocorre o distúrbio de *kick*, em seguida ocorre o fechamento do poço e por fim, ocorre a migração do gás (ar comprimido)

O experimento de operação de *bullheading* tem seu início de forma análoga ao experimento anterior*,* logo após, o *kick* é aplicado a operação de *bullheading,* com a injeção do fluido pelo anular em contra corrente, sendo a pressão controlada através da frequência de rotação da bomba. Entretanto, logo após, é realizada a análise do comportamento das pressões do poço. Seguem as etapas do experimento: primeiro o anular precisa ser preenchido com o fluido água e com bomba desligada, depois ocorre o distúrbio de *kick*, em seguida ocorre o fechamento do poço, logo depois, ocorre a migração do gás (ar comprimido), seguido da identificação de *kick* e por fim, ocorre a operação de *bullheading.*

O experimento de operação de identificação de *kick* inicia com o anular preenchido por água e a bomba é mantida desligada, logo após o *kick,* a migração ocorre com o poço fechado, sendo realizada a análise do comportamento das pressões do poço. Seguem as etapas do experimento: primeiro o anular precisa ser preenchido com água, com bomba desligada, implementa-se, ocorre o distúrbio de *kick*, em seguida, ocorre o fechamento do poço e a migração do gás (ar comprimido). Esses experimentos são análogos ao de injeção/migração de gás, mas foram realizados com a motivação de identificar o distúrbio de *kick* de gás.

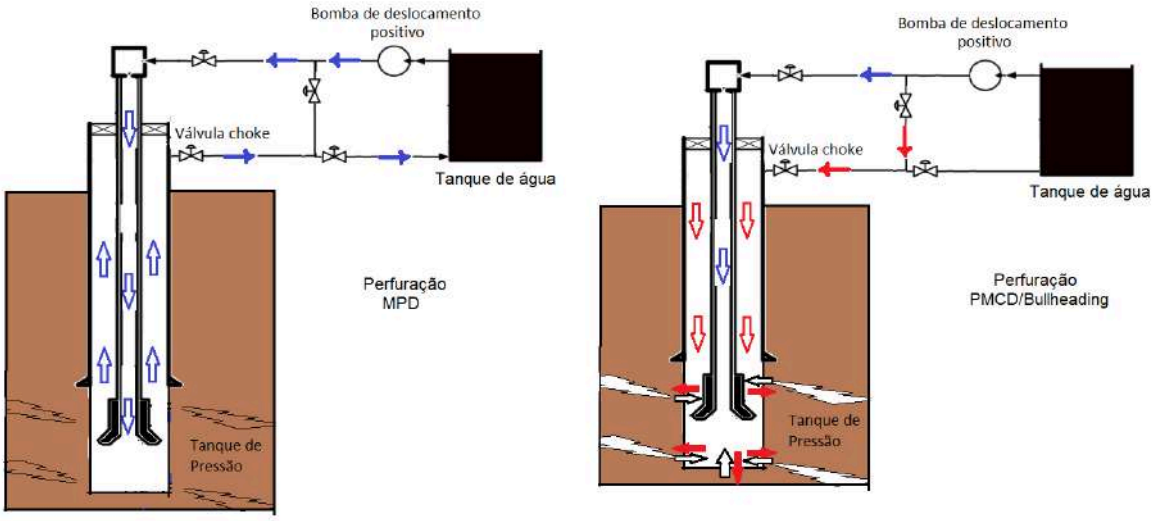

Figura *17* – Esquema da simulação na unidade experimental. Fonte: A autora.

Nas Figuras 18 a 26 apresentam-se todos os dados experimentais de Carvalho (2018) que foram utilizados para a geração do modelo matemático baseado em *machine learning.*

Inicialmente são apresentados os resultados dos dados experimentais coletados para este estudo. Os experimentos de *injeção/migração de gás* possuem os seguintes parâmetros: índice de abertura da válvula *choke* em 23%, frequência de rotação da bomba em 3 Hz, pressão do tanque no experimento 0 de 50 psi e, no experimento 1 de 60 psi (Figura 18).

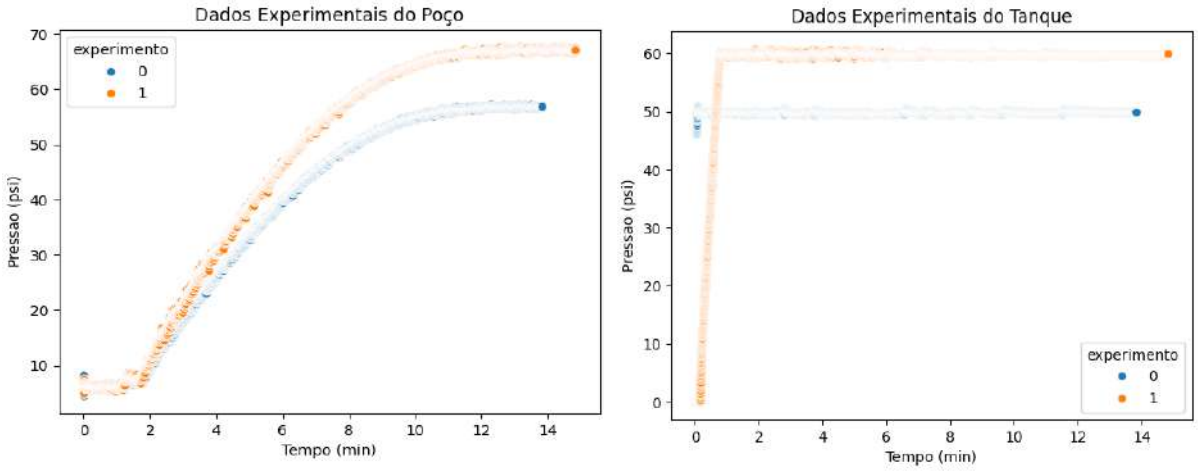

Figura *18* – Experimentos de injeção/migração de gás. Fonte: A autora.

Os experimentos de PMCD estão com *set point* da pressão no tanque de 60 psi e com os seguintes parâmetros: índice de abertura da válvula *choke* em 27%, frequência de rotação da bomba nos experimentos 2, 3, 4 e 5 em 13 Hz, 15 Hz, 17 Hz e 20 Hz, respectivamente, *set point* final com pressão do tanque em 20 psi, (Figura 19).

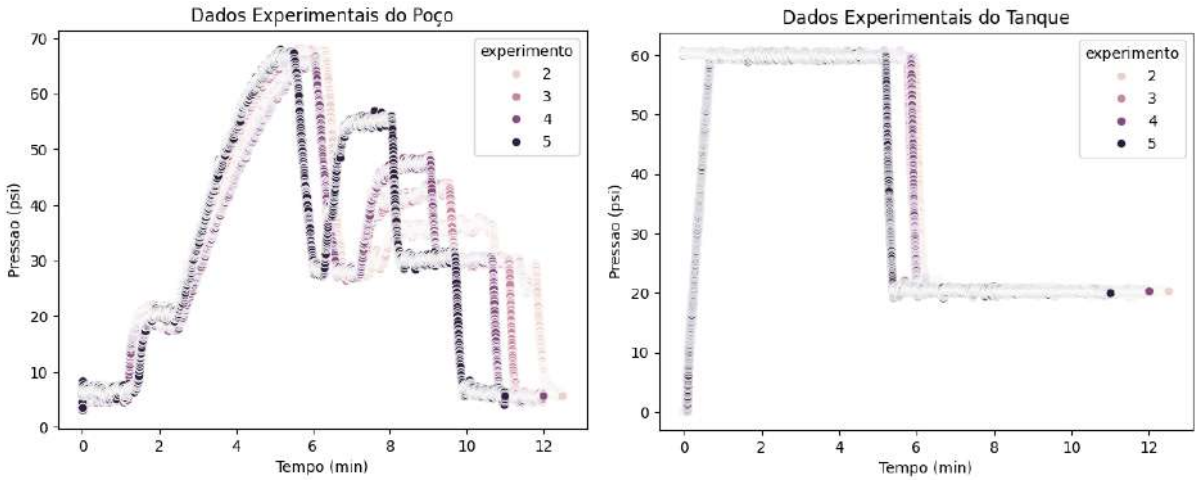

Figura *19* – Experimentos de 2 a 5 de PMCD. Fonte: A autora.

Os experimentos de PMCD, com *set point* da pressão no tanque de 60 psi, índice de abertura da válvula *choke* em 27%, frequência de rotação da bomba nos experimentos 6, 7, 8 e 9 em 13 Hz, 15 Hz, 17 Hz e 20 Hz, respectivamente, e com *set point* final com pressão do tanque em 60 psi, são ilustrados na Figura 20.

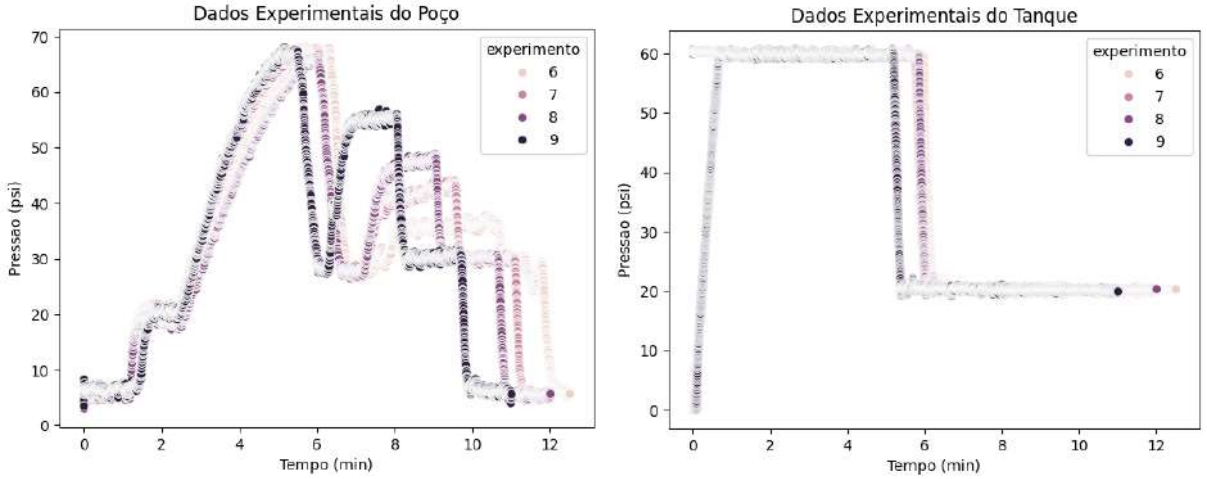

Figura *20* – Experimentos de 6 a 9 de PMCD. Fonte: A autora.

Os experimentos de PMCD, com *set point* da pressão no tanque de 70 psi, índice de abertura da válvula *choke* com 27%, frequência de rotação da bomba nos experimentos 10, 11, 12 e 13 em 13 Hz, 15 Hz, 17 Hz e 20 Hz, respectivamente e com *set point* final com pressão do tanque em 20 psi, são ilustrados na Figura 21.

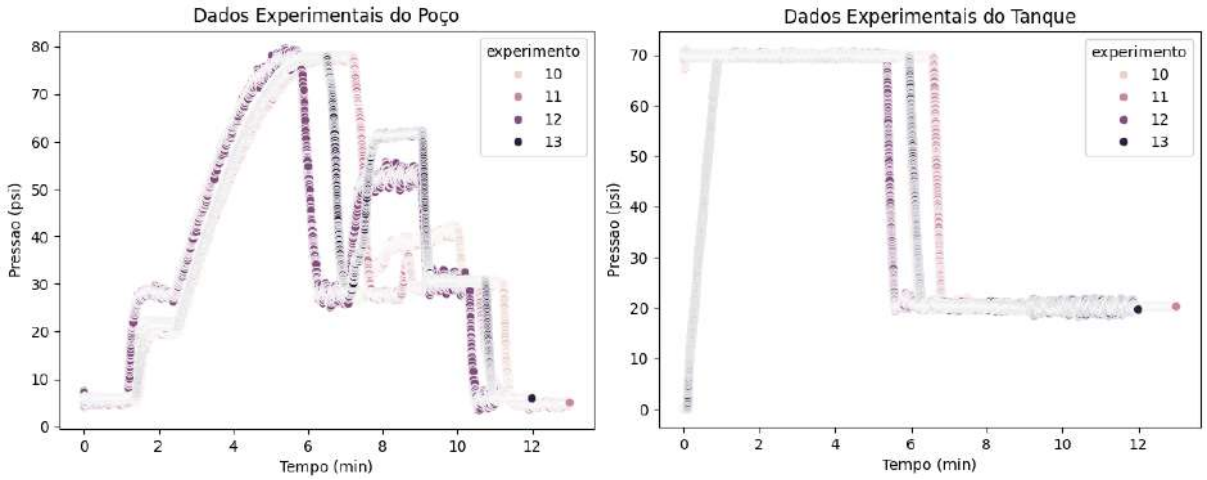

Figura *21* – Experimentos de 10 a 13 de PMCD. Fonte: A autora.

Os experimentos de PMCD, com *set point* da pressão no tanque de 70 psi, índice de abertura da válvula *choke* com 27%, frequência de rotação da bomba nos experimentos 14, 15, 16 e 17 em 13 Hz, 15 Hz, 17 Hz e 20 Hz, respectivamente, e com *set point* final com pressão do tanque em 70 psi, são ilustrados na Figura 22.

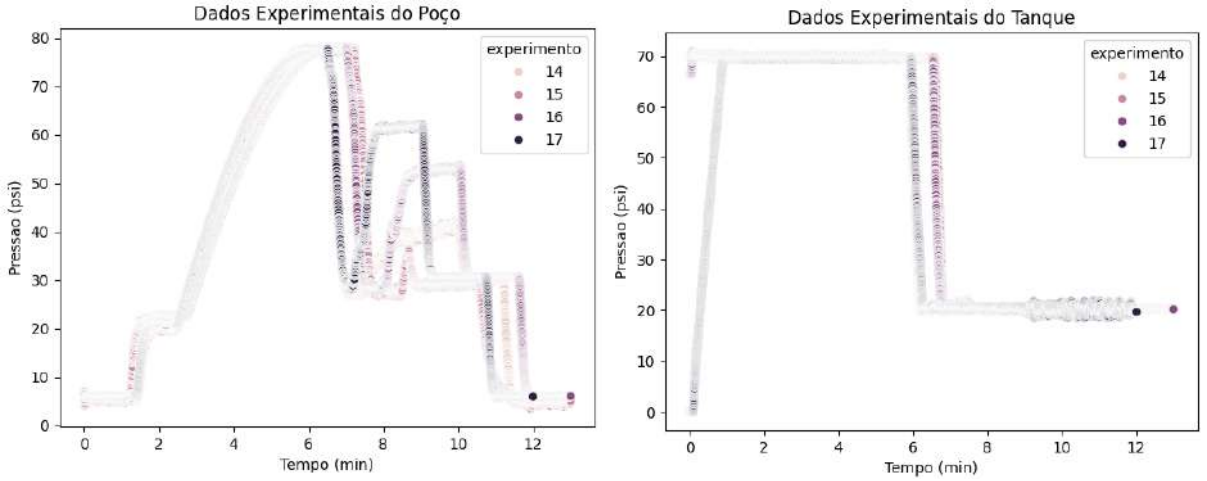

Figura *22* – Experimentos de 14 a 17 de PMCD. Fonte: A autora.

Os experimentos de PMCD, com *set point* da pressão no tanque de 80 psi, índice de abertura da válvula *choke* com 27%, frequência de rotação da bomba: nos experimentos 18, 19, 20 e 21 em 13 Hz, 15 Hz, 17 Hz e 20 Hz, respectivamente, e com *set point* final com pressão do tanque em 20 psi, são ilustrados na Figura 23.

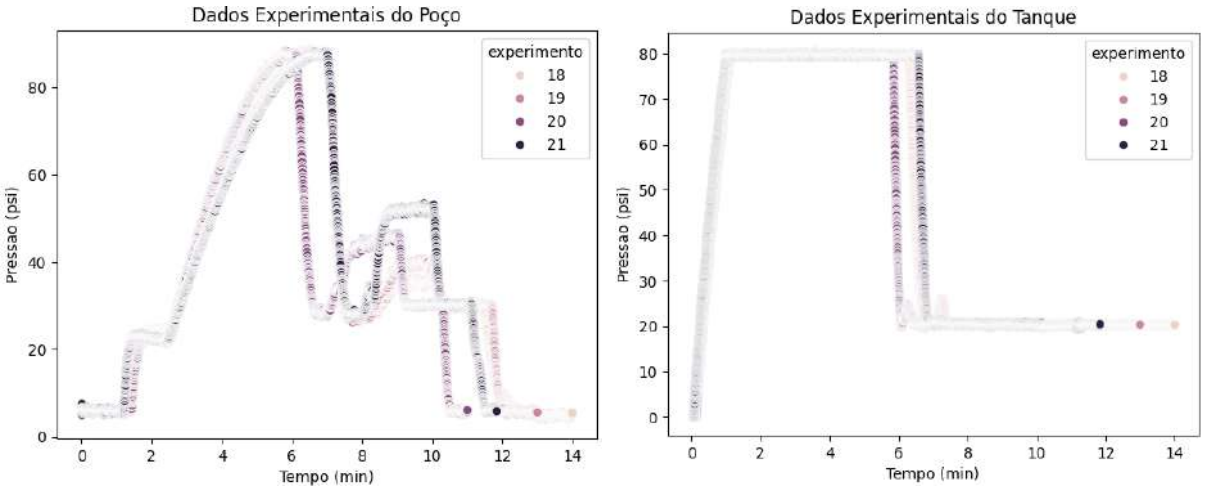

Figura *23* – Experimentos de 18 a 21 de PMCD. Fonte: A autora.

Os experimentos de PMCD, com *set point* da pressão no tanque de 80 psi, índice de abertura da válvula *choke* com 27%, frequência de rotação da bomba: no experimento 22 em 13 Hz e no experimento 23 em 15 Hz, e com *set point* final com pressão do tanque em 80 psi, são ilustrados na Figura 24.

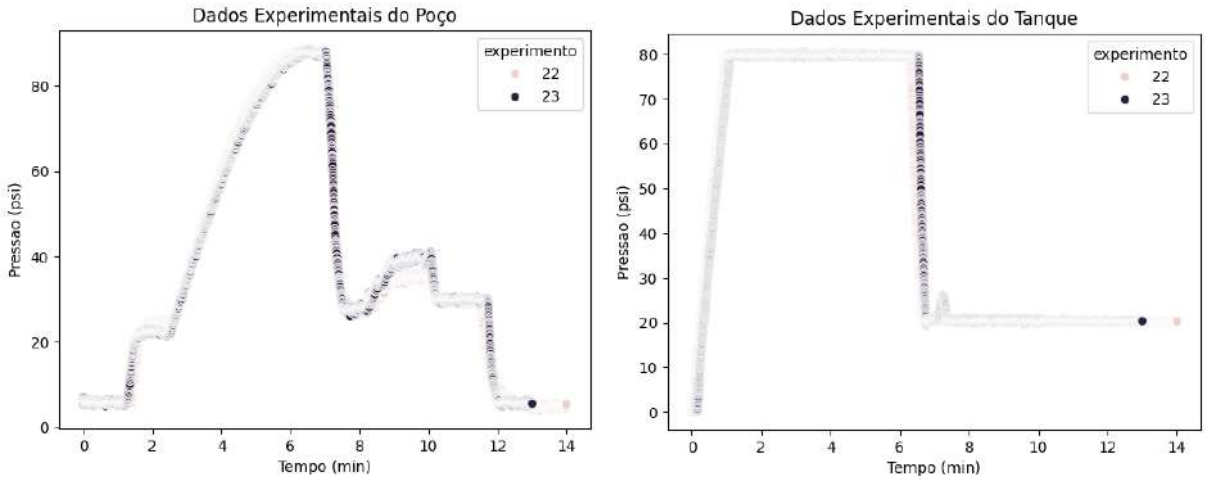

Figura *24* – Experimentos de 22 a 23 de PMCD. Fonte: A autora.

Os experimentos de identificação de *kick* foram executados variando-se a frequência de rotação da bomba e controlando-se a pressão no tanque em 20 psi, usando os seguintes parâmetros: índice de abertura da válvula *choke* com 100%, frequência de rotação da bomba nos experimentos 24, 25, 26 e 27 em 15 Hz, 17 Hz, 20 Hz e 25 Hz, respectivamente, e pressão do tanque inicialmente em 60 psi, (Figura 25).

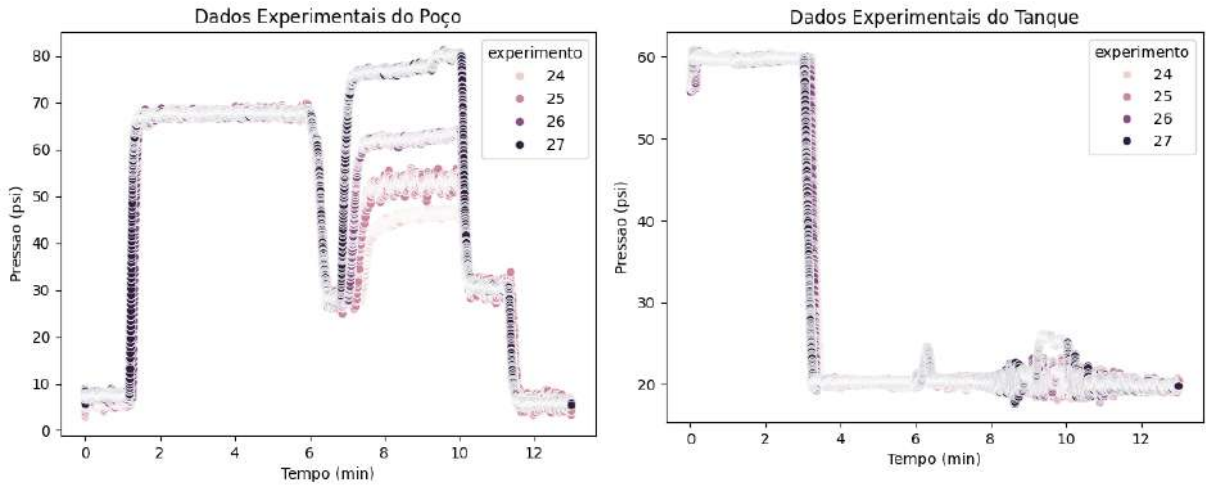

Figura *25* – Experimentos de 24 a 27 de opera*ção* de identificação de *kick*. Fonte: A autora.

Os experimentos de operação de identificação de *kick* foram executados variando-se a frequência de rotação da bomba, finalizados com pressão no tanque em 30 psi e com os seguintes parâmetros: índice de abertura da válvula *choke* com 100%, frequência de rotação da bomba nos experimentos 28, 29, 30 e 31 em 15 Hz, 17 Hz, 20 Hz e 25 Hz, respectivamente, e com pressão do tanque inicial em 60 psi (Figura 26).

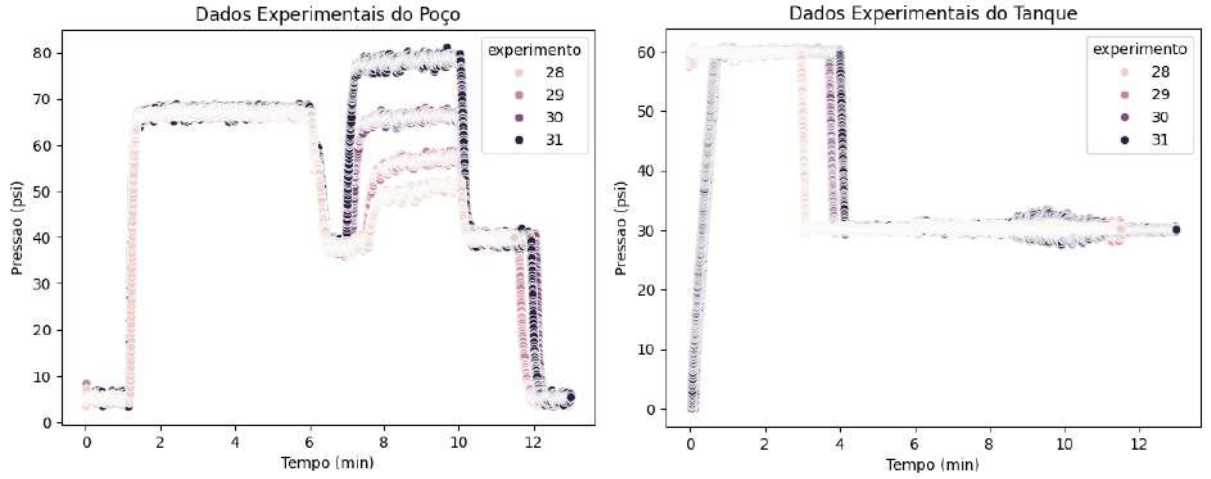

Figura *26* – Experimentos de 28 a 31 de operação de identificação de *kick*. Fonte: A autora.

3.1.2 Pré-processamento dos dados coletados

A partir dos dados coletados, em arquivo único para experimental e outro para dados de poços reais, é feito o pré-processamento, também chamado de tratamento dos dados, no qual foram realizadas as seguintes etapas: primeiro ocorre a seleção das variáveis, depois são tratados as linhas duplicadas, em seguida, é selecionada a quantidade de dados passados, por fim, ocorre a separação dos dados para o treino e para o teste.

Após essa separação dos dados, todas as etapas em seguida foram executadas para os dois conjuntos de dados (treino e teste): tratamento dos dados nulos, padronização da escala

numérica, a transformação dos dados, a normalização dos dados e por fim o tratamento de *outliers*.

Na seleção das variáveis de entrada foram selecionadas as variáveis mais importantes para descrever o processo de PMCD: tempo, a pressão da *choke*, a vazão, a frequência do inversor da bomba, a percentagem de abertura da válvula *choke*, a vazão, a percentagem da abertura da válvula do reservatório, sendo a variável alvo a pressão da *choke*.

Já para os dados de poços reais, foram selecionadas as seguintes variáveis: tempo, pressão na *choke* e vazão, sendo a variável alvo a pressão na *choke*.

Após a seleção das variáveis, são excluídas as linhas duplicadas encontradas nos dados coletados, tendo em vista que essas linhas duplicadas não trazem uma informação relevante. Logo em seguida, aplica-se a quantidade de dados passados, através da criação de novas variáveis, com informações relacionadas à quantidade de dados passados. Foi construído no código uma opção para selecionar a utilização ou não de dados passados para descrever adequadamente o comportamento dinâmico do processo de perfuração. Foram efetuados os seguintes testes: sem dados passados, com 2 dados passados, com 8 dados passados e 20 dados passados.

Por fim, ocorre a separação dos dados para treino e para teste, definido a partir do total de dados coletados, sendo 75% dos dados para o treino e o restante dos dados, 25% para o teste.

Após a separação, os dados nulos encontrados foram tratados através da sua eliminação, conforme recomenda Auyen (2022).

Logo em seguida, ocorre a padronização da escala numérica (Equação 1), com as seguintes escalas: de -4 a 4, de 0 a 1, de 0 a 4 e de 0 a 10.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Onde x é o dado original em cada amostra, x_{min} é o valor mínimo para a variável, x_{max} é o valor máximo para a variável e x' representa o dado normalizado.

Depois, foi aplicada a transformação dos dados, com o uso das funções sig e log, apresentadas nas Equações 2 e 3, respectivamente.

$$y = \frac{1}{1 - e^{-x}} \tag{2}$$

Enquanto que na função log, possui mais precisão, mesmo quando o valor do x seja próximo de zero.

$$y = ln(x + 1) \tag{3}$$

O teste de normalidade de Shapiro-Wilk, apresentado na Equação 4 (Gawali, 2021), informa se os dados possuem uma distribuição normal. A assimetria, apresentado na Equação 5, e a curtose, Equação 6, indicam o nível de deslocamento e o nível de achatamento da distribuição, respectivamente.

$$W = \frac{\frac{\sum_{i=1}^{n} ((a_i - x_{(i)})^2)}{n}}{\sum_{i=1}^{n} ((x_i - \bar{x})^2)} \tag{4}$$

$$g1 = \frac{\frac{\sum_{i=1}^{n} ((x_i - \bar{x})^3)}{n}}{\left(\frac{\sum_{i=1}^{n} ((x_i - \bar{x})^2)}{n}\right)^{\frac{3}{2}}} \tag{5}$$

$$K = \frac{m_4}{s^4} - 3 \tag{6}$$

Onde a_i representa os coeficientes determinados pelos dados, n é o número de observações na amostra, x_i representa cada observação na amostra, x é a média da amostra, m_4 é o quarto momento central em relação à média e s é o desvio padrão.

Por fim, ocorre o tratamento de *outliers*, onde os mesmos foram substituídos pela média.

3.1.3 Treinamento dos dados processados

Os dados de treino que já estão pré-processados são utilizados pelo Optuna para otimização dos hiperparâmetros, onde são definidos: a função objetivo, sendo aplicada como métrica a ser minimizada o RMSE. A quantidade de iterações foi de 5 e 30, gerando como resultado os melhores parâmetros que possuem o erro RMSE minimizado. Após o ajuste, implementa-se o treinamento, validação e teste com os melhores parâmetros. Vale ressaltar que foi realizada validação cruzada para evitar supertreinamento.

O treinamento dos dados, baseados nos conceitos matemáticos do algoritmo XGBoost, consideram um conjunto de dados com n exemplos e m características (variáveis), representado como $D = \{(x_i, y_i)\}$ sendo ($|D| = n$, $x_i \in \mathbb{R}^m$, $y_i \in \mathbb{R}$).

$$y_i = \varphi(x_i) = \sum_{k=1}^{K} (f_k(x_i)), f_k \in F \tag{7}$$

$$F = \{ f(x) = wq(x) \}, \text{ onde } (q: \mathbb{R}^m \rightarrow T, w \in \mathbb{R}^T) \tag{8}$$

Onde y_i representa a previsão, K representa o número total de funções aditivas que um modelo de árvore utiliza para realizar as previsões, F é o espaço de árvores de regressão, q representa como cada árvore divide os exemplos em diferentes folhas, T representa o número total de folhas em uma árvore, f_k representa uma árvore específica com sua própria maneira de dividir exemplos e pesos nas folhas, w representa os pesos de folha e ϕ é utilizado para expressar que y é uma função de x.

As regras nas árvores (representadas por q) são empregados para decidir em qual folha um exemplo se enquadra e, em seguida, somam-se as pontuações das folhas (representadas por w) para o cálculo da previsão final.

Para aprender o conjunto de funções usadas no modelo, minimiza-se o seguinte objetivo regularizado:

$$L(\varphi) = \sum l(\hat{y}_i, y_i) + \sum \Omega(f_k) \tag{9}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda(w)^2 \in F \tag{10}$$

Onde l representa uma função de perda convexa diferenciável que mede a diferença entre a previsão $\hat{y}_i$ e o alvo y_i, Ω penaliza a complexidade do modelo (ou seja, as funções da árvore de regressão), γ é um parâmetro que controla se um determinado nó da árvore será dividido com base em uma determinada métrica de ganho, λ é o parâmetro de regularização que controla a penalização sobre a complexidade do modelo. Ele adiciona uma penalidade proporcional à magnitude dos pesos w ao termo de perda da função objetiva. Um valor maior de λ leva a uma maior regularização, o que incentiva o modelo a preferir pesos mais próximos de zero, reduzindo a complexidade do modelo e as demais variáveis já foram apresentadas nas Equações 7 e 8.

A parte de regularização ajuda a suavizar os pesos que são aprendidos no final, de forma a evitar o *overfitting*, quando o modelo está pouco generalista, com sobreajuste para dados vistos.

O modelo de conjunto de árvores na Equação 9 é treinado de forma aditiva. Considera-se que $\hat{y}^t_i$ é a previsão da i-ésima instância na t-ésima iteração, dessa forma, é adicionado o f^t que melhora o modelo para minimizar o objetivo apresentado na Equação 11:

$$L(t) = \sum_{i=1}^{n} (l(y_i \hat{y}^{(t-1)} + f_t(x_i))) + \Omega(f_t) \tag{11}$$

A Equação 12 ilustra aproximação de segunda ordem usada para otimizar o objetivo no cenário geral:

$$L(t) = \sum_{i=1}^{n} [l(y_i \hat{y}^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{12}$$

$$g_i = \delta_{\hat{y}^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \tag{13}$$

$$h_i = \delta^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \tag{14}$$

Onde g_i e h_i são estatísticas de gradiente de primeira e segunda ordem na função de perda. Os termos constantes são removidos para obtenção do objetivo simplificado na etapa *t* (Equação 15):

$$\hat{L}(t) = \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{15}$$

Define-se $I_j = i \mid q(x_i) = j$ como o conjunto de instâncias da folha *j*, ou seja, I_j é o conjunto de índices dos exemplos no conjunto de treinamento que terminam na folha *j* após o processo de construção da árvore. Então a equação é simplificada expandindo Ω conforme as Equações 16 e 17:

$$\hat{L}(t) = \sum_{i=1}^{n} \left[g_i f_t\left(x_i\right) + \frac{1}{2} h_i f_t^2\left(x_i\right) \right] + \lambda T + \frac{1}{2} \lambda \sum_{j=1}^{T} (w_j^2) \tag{16}$$

$$\hat{L}(t) = \sum_{j=1}^{T} [(\sum_{i \in I_i}^{n} (g_i)) w_j + \frac{1}{2} (\sum_{i \in I_i}^{n} (h_i + \lambda)) w_j^2] + \gamma T \tag{17}$$

Para uma árvore fixa $q(x)$, calcula-se o peso ótimo w^*_i da folha *j* (Equação 18) e seu valor ótimo (Equação 19) é similar à função de pontuação de impureza, utilizada para avaliar árvores de decisão, medindo a qualidade de uma estrutura de árvore *q*.

$$w_j^* = - \frac{\sum_{i \in I_i}^{n} (g_i)}{\sum_{i \in I_i}^{n} (h_i + \lambda)} \tag{18}$$

$$\hat{L}(t) = - \frac{1}{2} \sum_{j=1}^{T} [\frac{(\sum_{i \in I_j}^{n} (g_i))^2}{\sum_{i \in I_j}^{n} (h_i + \lambda)}] + \gamma T \tag{19}$$

Como é impossível enumerar todas as possíveis estruturas de árvore *q,* utiliza-se um algoritmo que começa com uma única folha e adiciona-se iterativamente ramos à árvore. Após cada divisão, a perda de informação é avaliada, Equação 20, considerando como a

divisão afeta a qualidade da árvore, e levando em conta as informações das divisões nas folhas esquerda e direita.

$$L_{split} = \frac{1}{2}\sum_{j=1}^{T}\left[\frac{\left(\sum_{i\in I_L}^{n}(g_i)\right)^2}{\sum_{i\in I_L}^{n}(h_i+\lambda)} + \frac{\left(\sum_{i\in I_R}^{n}(g_i)\right)^2}{\sum_{i\in I_R}^{n}(h_i+\lambda)} + \frac{\left(\sum_{i\in I}^{n}(g_i)\right)^2}{\sum_{i\in I}^{n}(h_i+\lambda)}\right] - \gamma \quad (20)$$

Onde $I = I_L \cup I_R$, I se refere ao conjunto de índices das instâncias no nó atual, o índice L representa a divisão do nó para o lado esquerdo, o índice R representa a divisão do nó para o lado direito.

3.1.4 Análise do modelo gerado

Na avaliação da performance dos resultados, foram utilizadas as métricas: *Mean Squared Error* (MSE), *Root Mean Squared Error* (RMSE), *Sum of Squared Error* (SSE) e R^2, apresentados nas Equações 21, 22, 23 e 24, respectivamente.

$$MSE = \frac{1}{N}\sum_{j=1}^{N}\left((\hat{y}_j - y_j)^2\right) \quad (21)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left((\hat{y}_j - y_j)^2\right)} \quad (22)$$

$$SSE = \sum_{j=1}^{N}\left((\hat{y}_j - y_j)^2\right) \quad (23)$$

$$R^2 = 1 - \frac{\sum_{j=1}^{N}\left((\hat{y}_j - y_j)^2\right)}{\sum_{j=1}^{N}\left((y_j - \hat{y})^2\right)} \quad (24)$$

Onde $\hat{y}_j$ representa o valor previsto pelo modelo para cada amostra, y_j representa o valor real de cada amostra, N representa a quantidade total de amostras, $\hat{y}$ representa a média de todos os valores reais.

Para a análise das perdas, foram utilizadas as funções: *LogLoss* apresentada pela Equação 25 e *SoftMax* apresentada na Equação 26.

$$LogLoss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}(y_{ij}\log(p_{ij})) \quad (25)$$

$$(z)_i = \frac{e^{zi}}{\sum_{j=1}^{K}(e^{zj})} \quad (26)$$

Onde N representa o número total de amostras, K o número total de classes, y_{ij} é uma variável indicadora que é 1 se a i-ésima observação pertence à j-ésima classe e 0 caso o contrário. p_{ij} é a probabilidade predita da i-ésima observação pertencer à j-ésima classe, z_i representa a pontuação associada à classe i e $\Sigma^{k}_{j=1}(e^{zj}) = 1$ é a soma de todas as exponenciais das pontuações.

4 RESULTADOS E DISCUSSÕES

No presente capítulo, são apresentados os resultados obtidos com o uso do *machine learning* aplicado aos dados experimentais e também aos dados de poços reais utilizando o XGBoost, bem como as métricas de avaliação dos modelos gerados.

4.1 Resultado dos dados experimentais de Carvalho (2018)

Foram coletados dados dos experimentos realizados no LEF/UFRRJ por Carvalho (2018), lidos por um algoritmo desenvolvido em *Python,* para gerar, em um único arquivo, todos os experimentos, descrevendo: de injeção/migração de gás, que estão identificados como experimento 0 e 1; operação de *bullheading*, que estão numerados do 6 ao 23; *kick*, que estão numerados do 24 ao 31. Os dados coletados foram organizados em uma planilha (Figura 27), como o total de 603.300 linhas com 35 colunas (dados de entrada e saída).

Após a coleta dos experimentos, o código foi desenvolvido para efetuar o pré-processamento, treinamento e análise gerando gráficos, figuras, métricas e o modelo com o melhor aprendizado.

	index	Tempo_poco	SetPoint	Pressao	Vazao	Freq_Inversor	Freq_Inversor_Lama	Abertura_choke	Abertura_Valvula1	Abertura_Valvula2	...	Pressao_Choke
0	0	0.001767	8.206621	8.206621	0.502082	3.0	3.0	95.0	100.0	0.0	...	0.000000e+00
1	1	0.008800	6.222524	6.222524	0.540918	3.0	3.0	95.0	100.0	0.0	...	0.000000e+00
2	2	0.010567	7.026104	7.026104	0.532649	3.0	3.0	95.0	100.0	0.0	...	0.000000e+00
3	3	0.012183	4.504444	4.504444	0.533163	3.0	3.0	95.0	100.0	0.0	...	0.000000e+00
4	4	0.012183	4.504444	4.504444	0.533163	3.0	3.0	95.0	100.0	0.0	...	0.000000e+00
...	...	...	...	...	...	...	...	...	...	...	...	...
603295	19000	11.829117	5.858904	5.858904	0.003183	3.0	3.0	95.0	100.0	0.0	...	2.000000e-08
603296	19001	11.830650	6.009814	6.009814	0.003203	3.0	3.0	95.0	100.0	0.0	...	2.000000e-08
603297	19002	11.830650	6.009814	6.009814	0.003203	3.0	3.0	95.0	100.0	0.0	...	2.000000e-08
603298	19003	11.832283	5.790000	5.790000	0.003126	3.0	3.0	95.0	100.0	0.0	...	2.000000e-08
603299	19004	11.832283	5.790000	5.790000	0.003126	3.0	3.0	95.0	100.0	0.0	...	2.000000e-08

603300 rows × 35 columns

Figura *27* – Dados experimentais. Fonte: A autora.

Logo em seguida, após a seleção das variáveis para treinar o modelo, suas características (o total de valores únicos, o tipo de dado, suas médias, seus desvios padrão, valores mínimos, seus quartis e valores máximos) foram apresentados na Figura 28.

	Valores_Únicos	Tipo_de_Dado	mean	std	min	25%	50%	75%	max
Tempo_poco	141160	float64	5.91	3.64	0.00	2.78	5.63	8.92	14.83
Pressao	173611	float64	40.01	24.00	2.92	21.37	37.99	63.97	88.66
Vazao	135687	float64	0.18	0.35	-0.00	0.00	0.00	0.17	2.12
Freq_Inversor	6	float64	6.12	6.05	3.00	3.00	3.00	3.00	25.00
Abertura_choke	198	float64	63.65	43.61	0.00	3.00	95.00	95.00	100.00
Vazao2	135099	float64	0.18	0.35	-0.00	0.00	0.00	0.17	2.12
Abertura_Valvula_Reservatorio	4	float64	44.29	42.12	0.00	0.00	27.00	100.00	100.00
Pressao_Choke	155660	float64	35.59	26.45	0.00	4.40	36.30	60.21	94.68
Tempo_tanque	310483	float64	5.91	3.64	0.00	2.78	5.63	8.92	14.83
Pressao_Tanque	284968	float64	43.55	22.58	0.00	20.35	42.54	60.07	80.59
experimento	32	int64	15.61	9.35	0.00	7.00	16.00	24.00	31.00

Figura *28* – Resumo do *dataframe* dos experimentos. Fonte: A autora.

4.1.1 Dados experimentais tratados com a função sig e sem aplicar dados passados

Foram realizados testes para a geração do modelo matemático sem a introdução de dados passados, com 2 e 8 dados passados, sendo todos com aplicação da escala de -4 a 4.

Com a arquitetura inicial que não empregou informação dinâmica de dados passados, as variáveis vazão e frequência do inversor apresentam *outiliers,* logo foram tratadas com a substituição pelas suas médias, conforme mostrado na Figura 29.

```
Não há outliers para tratar em Tempo_poco
Não há outliers para tratar em Pressao
Tratando outliers em Vazao
Tratando outliers em Freq_Inversor
Não há outliers para tratar em Abertura_choke
Tratando outliers em Vazao2
Não há outliers para tratar em Abertura_Valvula_Reservatorio
Não há outliers para tratar em Tempo_tanque
Não há outliers para tratar em Pressao_Tanque
```

Figura *29* – Mostrando as variáveis com *outliers* com dados experimentais, tratados com a função sig e sem dados passados. Fonte: A autora.

Antes dos dados serem treinados pelo XGBoost, foi realizada a otimização dos hiperparâmetros com o Optuna gerando o melhor resultado na iteração 4 os melhores parâmetros são: 'n_estimators': 570, 'learning_rate': 0.2744, 'max_depth': 11, 'min_child_weight': 4, 'subsample': 0.8976657889958629, 'colsample_bytree': 0.6952, 'gamma': 0.7453, 'reg_alpha': 0.2214 e 'reg_lambda': 0.8697, em apenas 2 minutos e 2 segundos de execução, executado no *back-end* do Google Compute Engineer em Python 3, com 12.7 GB de memória e 107.7 GB disponíveis em disco.

Com os melhores parâmetros gerados pelo Optuna, os dados são treinados, validados e testados pelo XGBoost*,* o modelo com o melhor aprendizado e suas métricas (R^2, RMSE, SSE e MSE) são apresentados nas Figuras 30 e 31, apesar do R^2 elevado, as demais métricas de avaliação indicam que o modelo não foi treinado. RMSE tem métrica similar a MSE, entretanto a aplicação da raiz quadrada permite que seu valor tenha a mesma escala do dado original, indicando que o treinamento do modelo não foi eficaz.

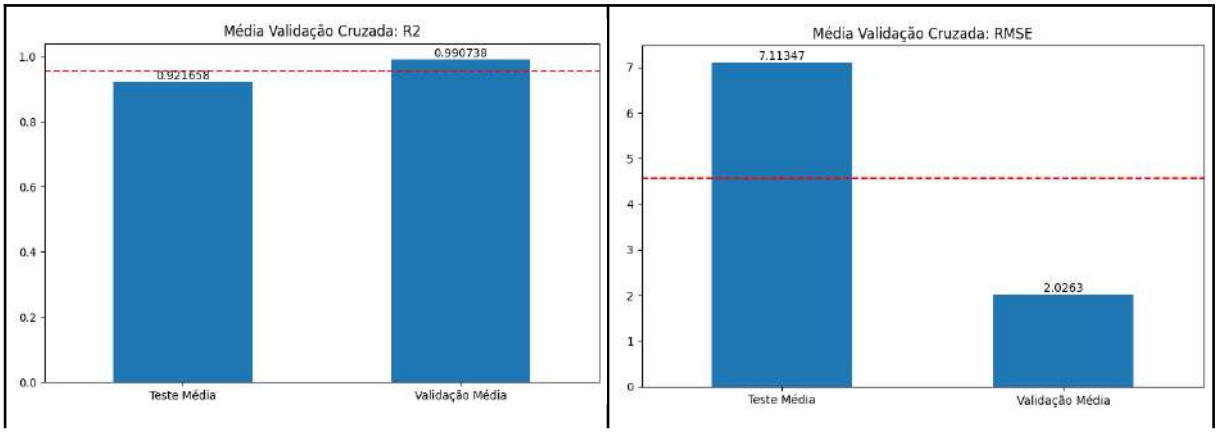

Figura *30* – Métricas de avaliação R^2 e RMSE com dados experimentais, tratados com a função sig e sem dados passados. Fonte: A autora.

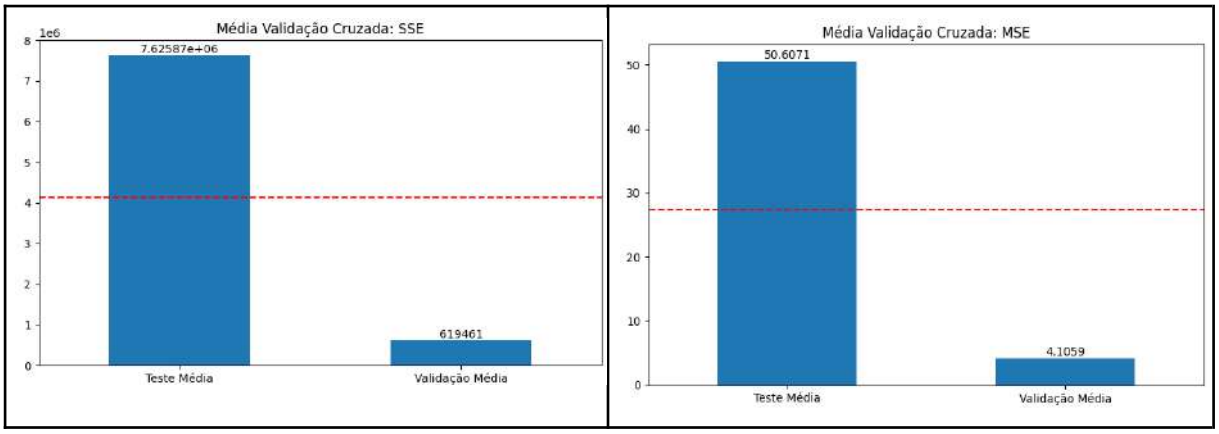

Figura *31* – Métricas de avaliação SSE e MSE com dados experimentais, tratados com a função sig e sem dados passados. Fonte: A autora.

São apresentados na Figura 32 os valores da densidade dos dados para a variável de saída no teste e na validação do modelo, sendo que quanto mais valores preditos de forma correta, ou seja, quando valor predito é igual ao valor real, mais será observada a presença da cor roxa nos gráficos. Verifica-se a ineficácia do modelo desenvolvido pela baixa concentração da cor roxa.

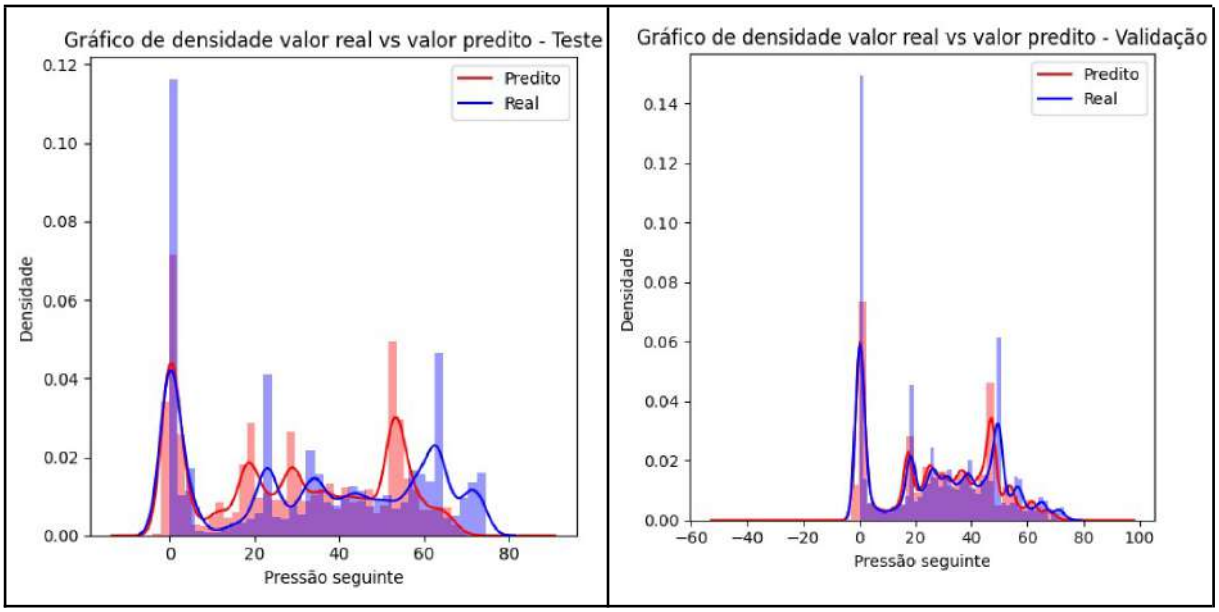

Figura *32* – Gráficos de densidade com dados experimentais, tratados com a função sig e sem dados passados. Fonte: A autora.

São apresentados na Figura 33 os valores da distribuição dos resíduos do modelo no teste e na validação resultando grande quantidade de valores residuais distantes de zero, indicando baixo potencial preditivo para o modelo.

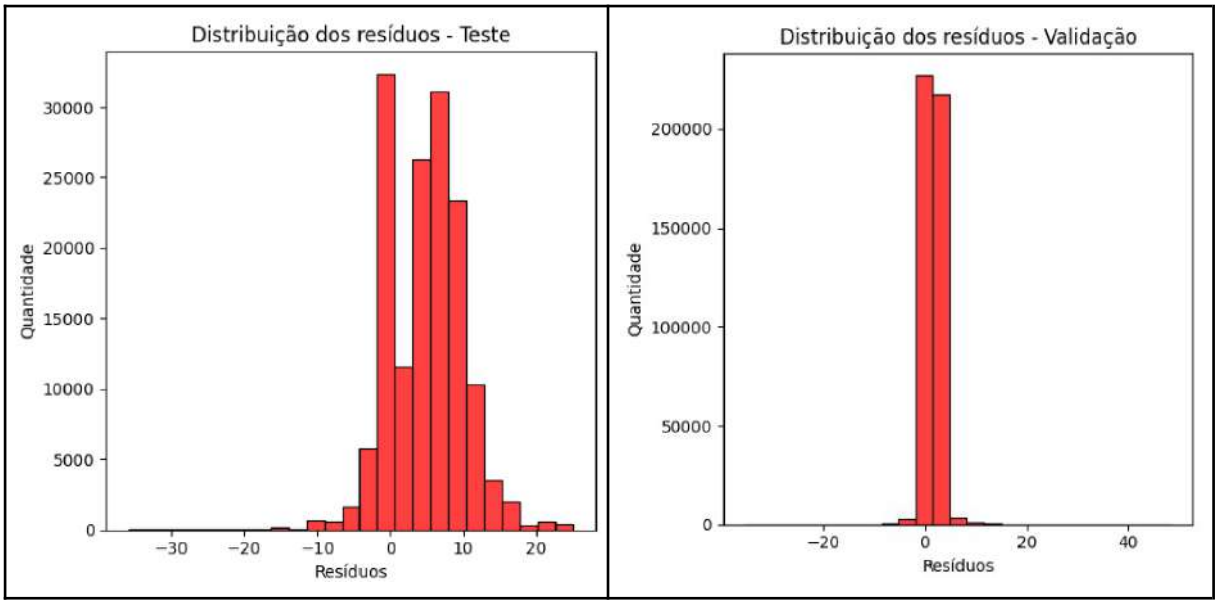

Figura *33* – Distribuição dos resíduos com dados experimentais, tratados com a função sig e sem dados passados. Fonte: A autora.

A Figura 34 apresenta os pontos na curva de comparação de valor real com o valor predito. Verifica-se que há uma grande distância entre os pontos e a reta em vermelho, o que indica a geração de um modelo inadequado.

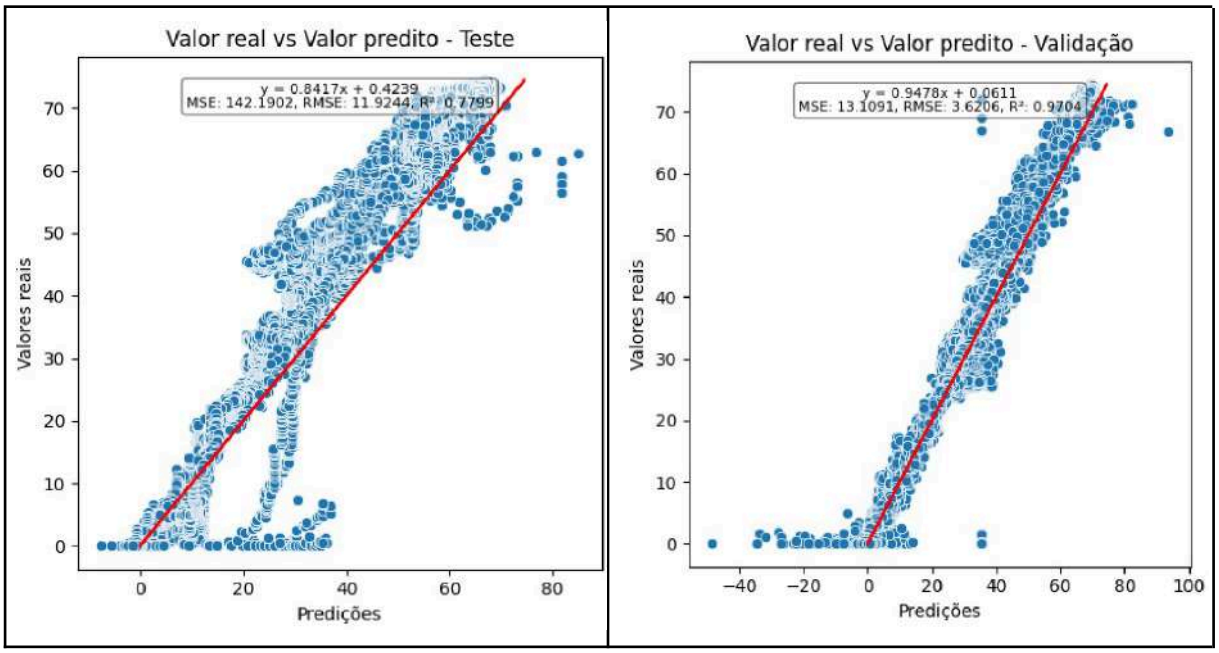

Figura *34* – Gráfico de evolução do modelo com dados experimentais, tratados com a função sig e sem dados passados. Fonte: A autora.

O gráfico da função de perda ao longo das épocas é apresentado na Figura 35. Os resultados revelam que a função da perda não se aproxima de zero, indicando a produção de um modelo ineficiente.

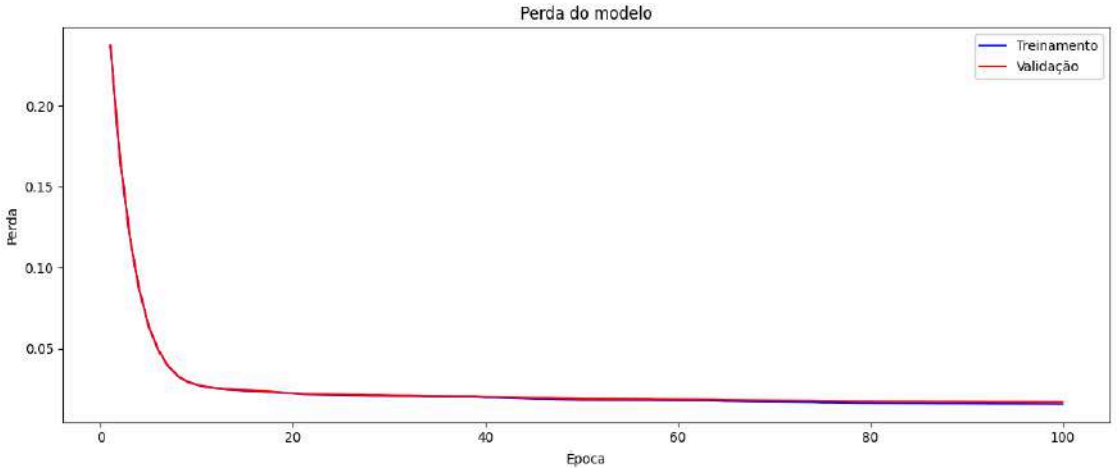

Figura *35* – Curva de perdas função sig e sem dados passados. Fonte: A autora.

O gráfico na Figura 36 representa a importância de cada variável para as previsões, sendo que a pressão, teve uma forte influência nas previsões, seguida da percentagem da abertura da válvula *choke*.

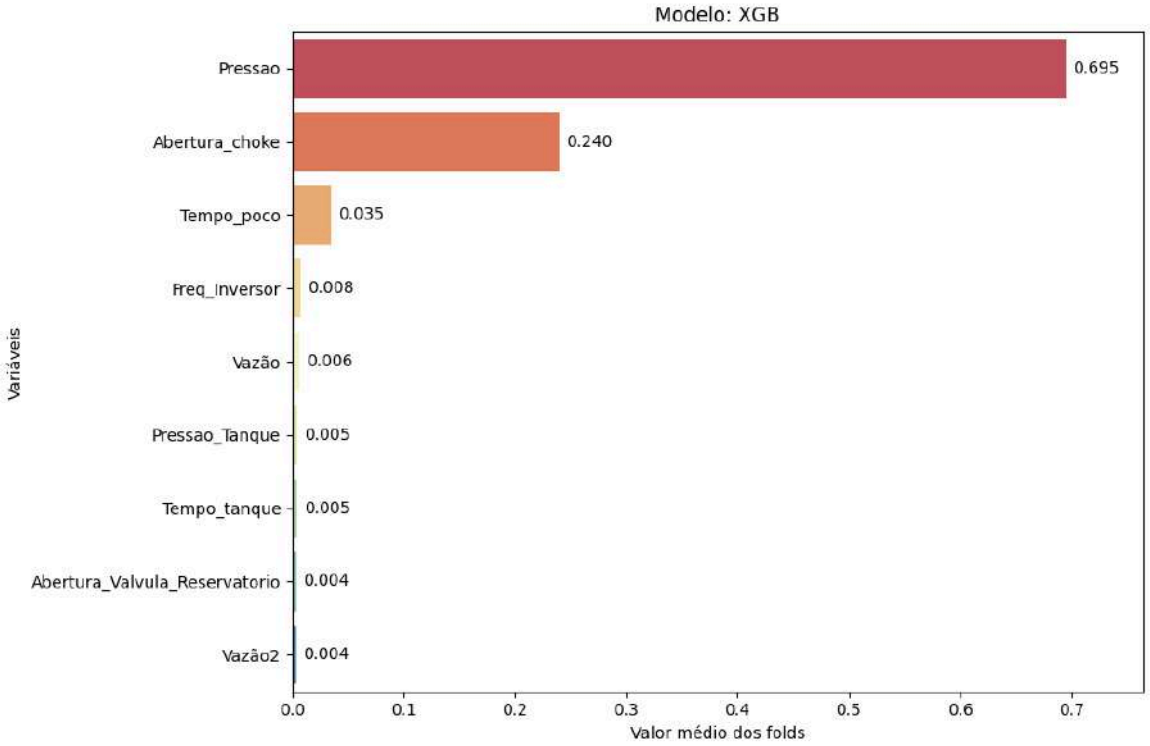

Figura *36* – Importância das variáveis com dados experimentais, tratados com a função sig e sem dados passados. Fonte: A autora.

As Figuras 37 a 40 representam o quanto o modelo consegue prever a operação de PMCD. Todos os resultados apontam que o modelo matemático não foi eficiente para

representar os dados experimentais. Os resultados para os demais experimentos estão no Anexo A.

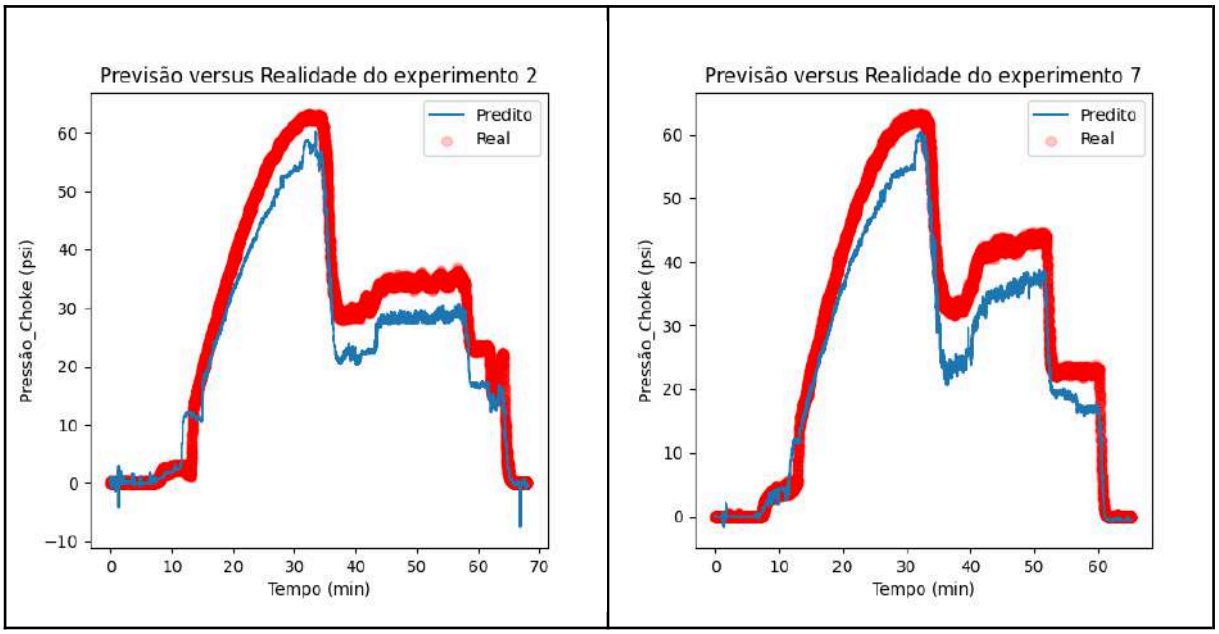

Figura *37* – Previsão versus realidade dos testes com dados experimentais, tratados com a função sig e sem dados passados, para os experimentos 2 e 7. Fonte: A autora.

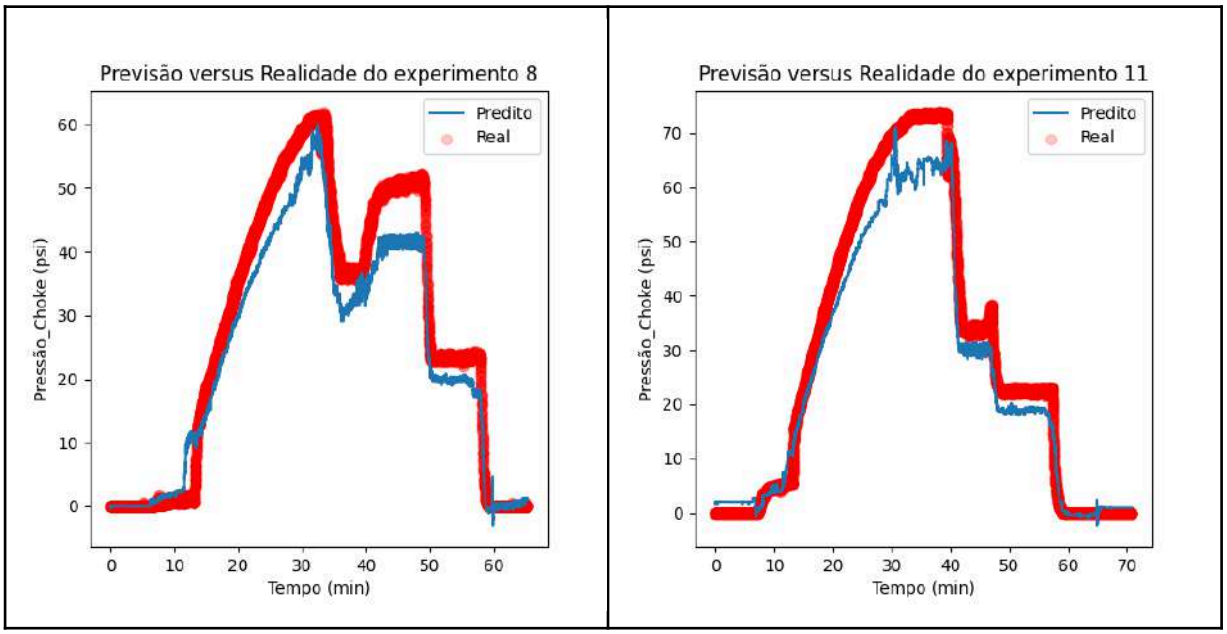

Figura *38* – Previsão versus realidade dos testes com dados experimentais, tratados com a função sig e sem dados passados, para os experimentos 8 e 11. Fonte: A autora.

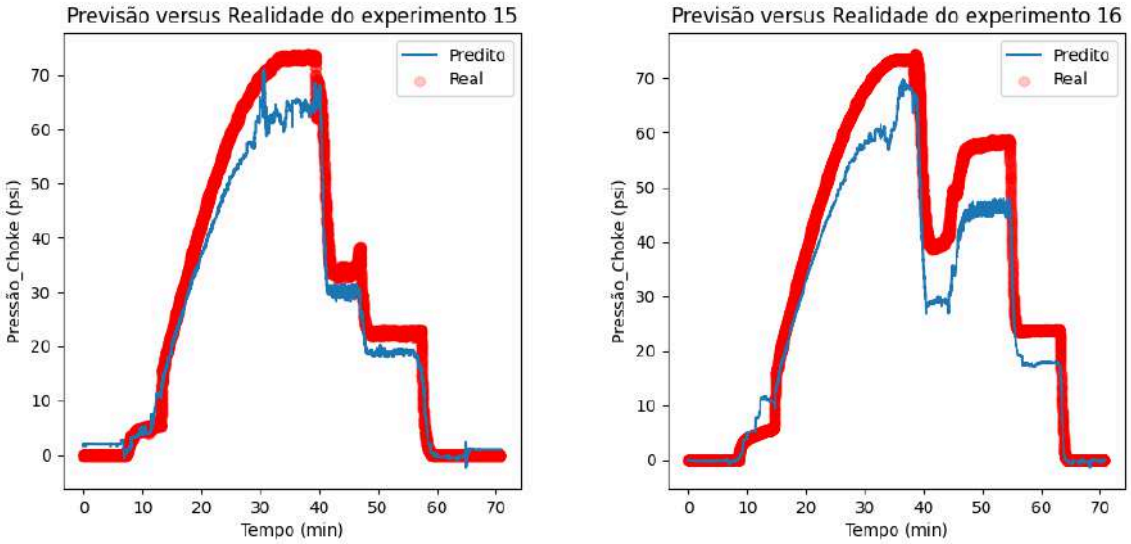

Figura *39* – Previsão versus realidade dos testes com dados experimentais, tratados com a função sig e sem dados passados, para os experimentos 15 e 16. Fonte: A autora.

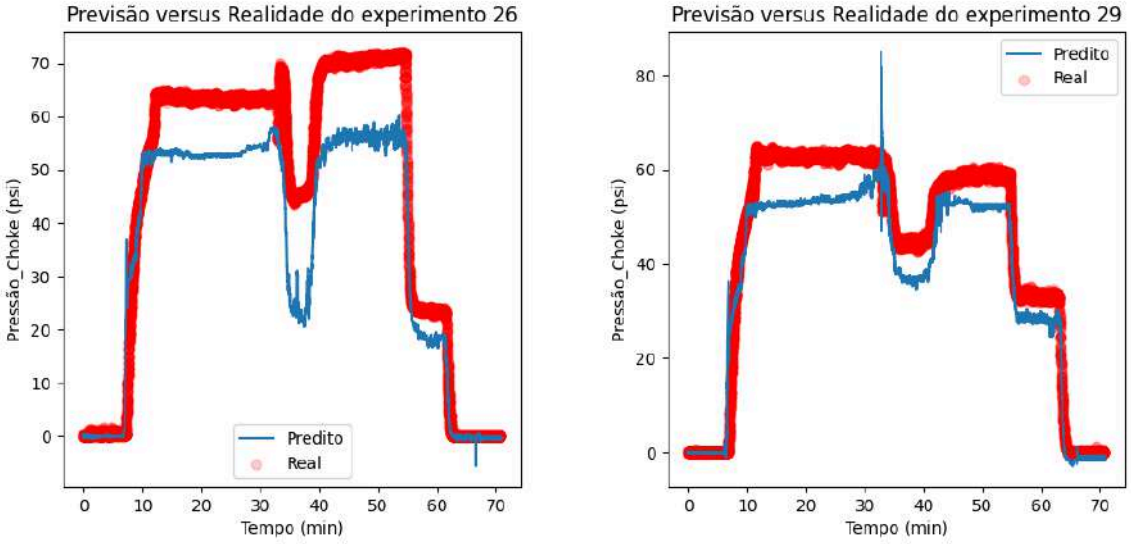

Figura *40* – Previsão versus realidade dos testes com dados experimentais, tratados com a função sig e sem dados passados, para os experimentos 26 e 29. Fonte: A autora.

Com relação à arquitetura do modelo, o modelo treinou 570 árvores, sendo demonstrado na Figura 41 uma parte de sua árvore. Cada nó contém a decisão de uma variável e as folhas representam a pontuação que será somada a cada amostra para gerar sua previsão ao final. Vale ressaltar que em cada nó os valores das variáveis variam de -4 a 4 por conta da normalização que é realizada antes do treinamento do modelo.

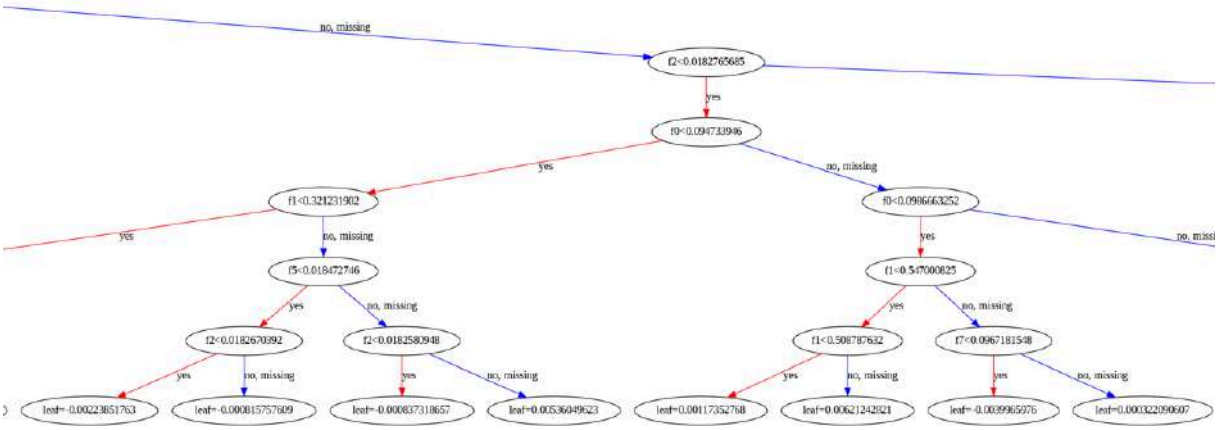

Figura *41* – Parte da árvore gerada pelo modelo. Fonte: A autora.

4.1.2 Dados experimentais tratados com a função sig e com 2 dados passados

Ao aplicar 2 dados passados, todas as variáveis são deslocadas, conforme mostrado na Figura 42, a vazão e a frequência do inversor por apresentarem *outiliers,* foram tratadas com a substituição pelas suas médias.

	Valores_Únicos	Tipo_de_Dado	mean	std	min	25%	50%	75%	max
Pressao_Choke_k-1 (psi)	176138	float64	0.41	0.35	0.02	0.03	0.33	0.78	0.98
Tempo_poco_k (min)	169124	float64	0.40	0.33	0.02	0.08	0.30	0.72	0.98
Pressao_k (psi)	195390	float64	0.44	0.36	0.02	0.10	0.35	0.87	0.98
Vazão_k (m³/h)	148591	float64	0.04	0.09	0.02	0.02	0.02	0.03	0.98
Freq_Inversor_k (Hz)	6	float64	0.09	0.20	0.02	0.02	0.02	0.02	0.98
Abertura_choke_k (%)	198	float64	0.64	0.44	0.02	0.02	0.97	0.97	0.98
Vazão2_k (m³/h)	147876	float64	0.04	0.09	0.02	0.02	0.02	0.03	0.98
Abertura_Valvula_Reservatorio_k (%)	4	float64	0.39	0.43	0.02	0.02	0.14	0.98	0.98
Tempo_tanque_k (min)	387451	float64	0.40	0.33	0.02	0.08	0.30	0.72	0.98
Pressao_Tanque_k (psi)	354017	float64	0.54	0.38	0.02	0.12	0.58	0.94	0.98
Pressao_Choke_k (psi)	176071	float64	0.41	0.35	0.02	0.03	0.33	0.78	0.98
Pressão_Choke_k+1 (psi)	155655	float64	35.60	26.45	0.00	4.40	36.31	60.21	94.68
experimento	32	int64	15.61	9.35	0.00	7.00	16.00	24.00	31.00

Figura *42* – Resumo do *dataframe* com dados experimentais, tratados com a função sig e com 2 dados passados. Fonte: A autora.

Antes dos dados serem treinados pelo XGBoost, é feito a otimização dos hiperparâmetros com o Optuna gerando os melhores parâmetros, quais sejam: 'n_estimators': 875, 'learning_rate': 0.0692, 'max_depth': 17, 'min_child_weight': 3, 'subsample': 0.5050, 'colsample_bytree': 0.7981, 'gamma': 0.0019, 'reg_alpha': 0.8570 e 'reg_lambda': 0.8459, em apenas 3 minutos de execução, utilizando o mesmo ambiente de execução do modelo anterior.

Com os melhores parâmetros gerados pelo Optuna, os dados são treinados, validados e testados pelo XGBoost. As Figuras 43 e 44 apresentam as métricas de avaliação.

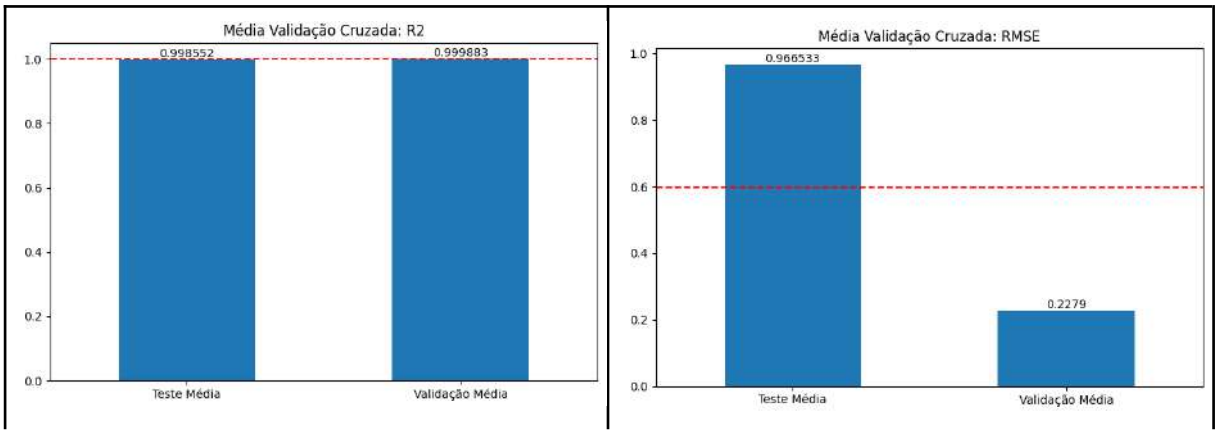

Figura *43* – Métricas de avaliação R² e RMSE com dados experimentais, tratados com a função sig e com 2 dados passados. Fonte: A autora.

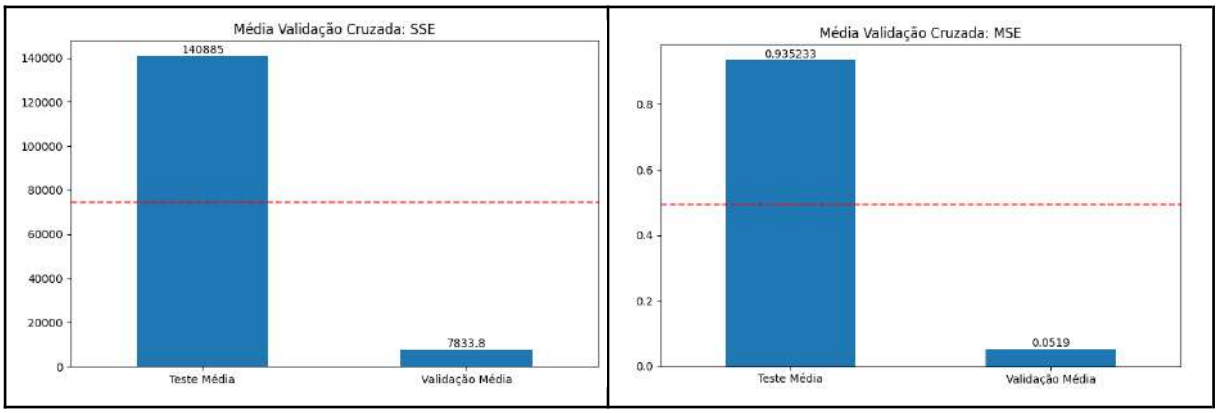

Figura *44* – Métricas de avaliação SSE e MSE com dados experimentais, tratados com a função sig e com 2 dados passados. Fonte: A autora.

A Figura 45 ilustra os valores da densidade dos dados para a variável de saída no teste e na validação, sendo que quanto mais valores preditos de forma correta, ou seja, quando valor predito é igual ao valor real, mais será observada a presença da cor roxa nos gráficos.

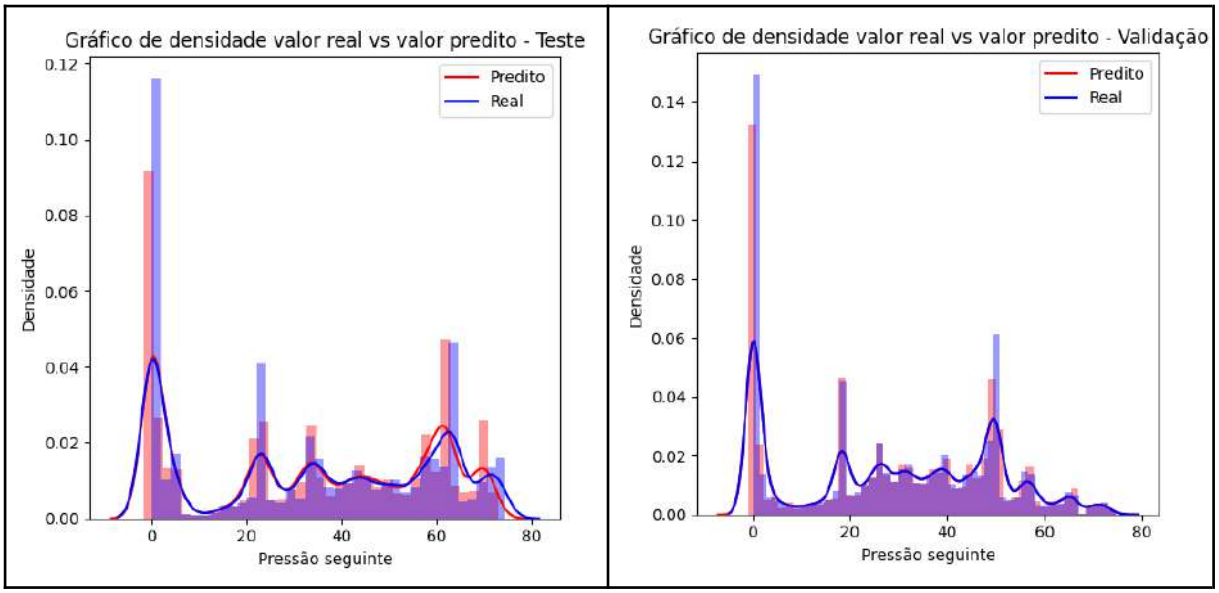

Figura *45* – Gráficos de densidade com dados experimentais, tratados com a função sig e com 2 dados passados. Fonte: A autora.

A Figura 46 contém os valores da distribuição dos resíduos no teste e na validação do modelo.

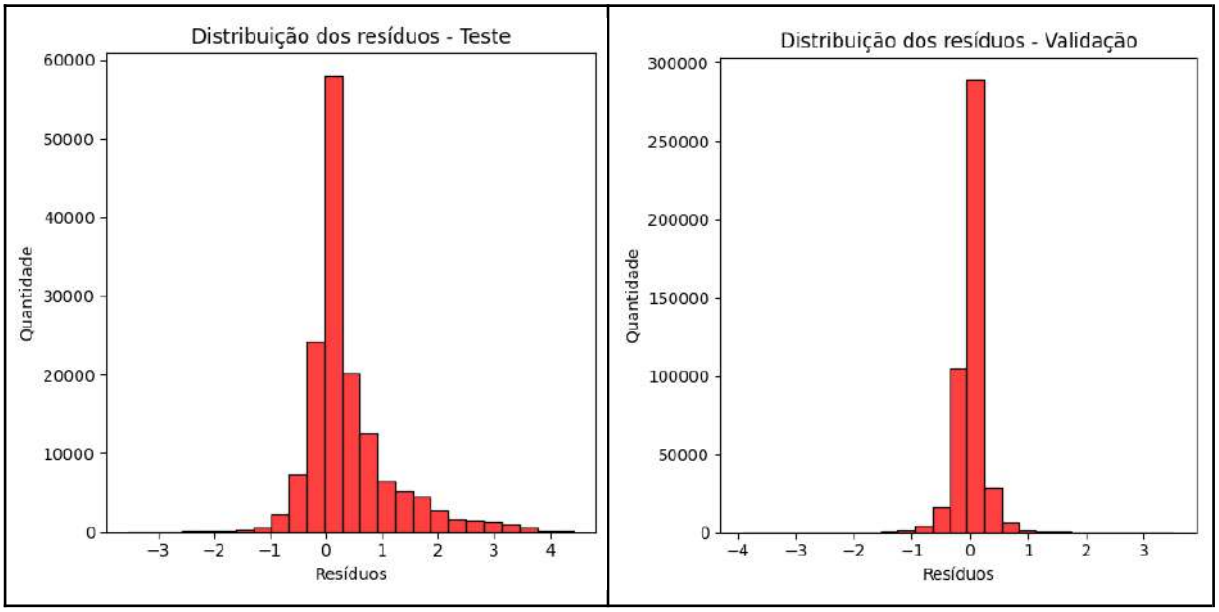

Figura *46* – Distribuição dos resíduos com dados experimentais, tratados com a função sig e com 2 dados passados. Fonte: A autora.

Na Figura 47, os pontos na curva de comparação de valor real e valor predito, ilustram quão distantes da reta (valor real) estão as previsões do modelo.

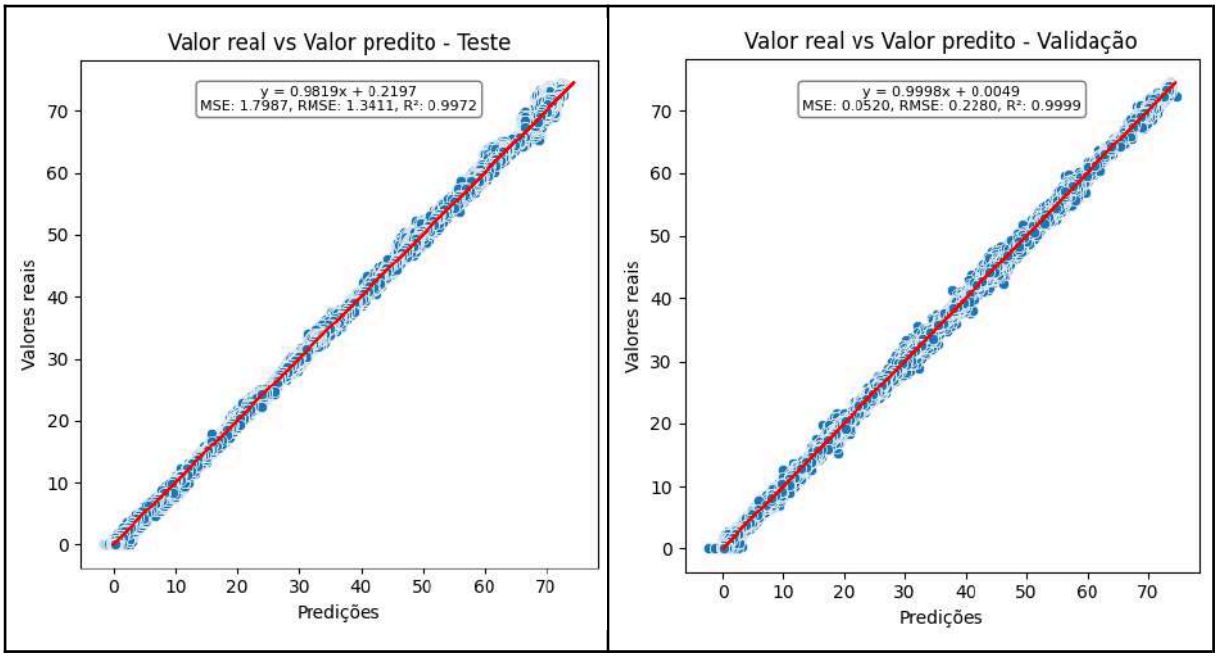

Figura *47* – Gráfico da evolução do modelo com dados experimentais, tratados com a função sig e com 2 dados passados. Fonte: A autora.

O gráfico da função de perda ao longo das épocas apresentado na Figura 48.

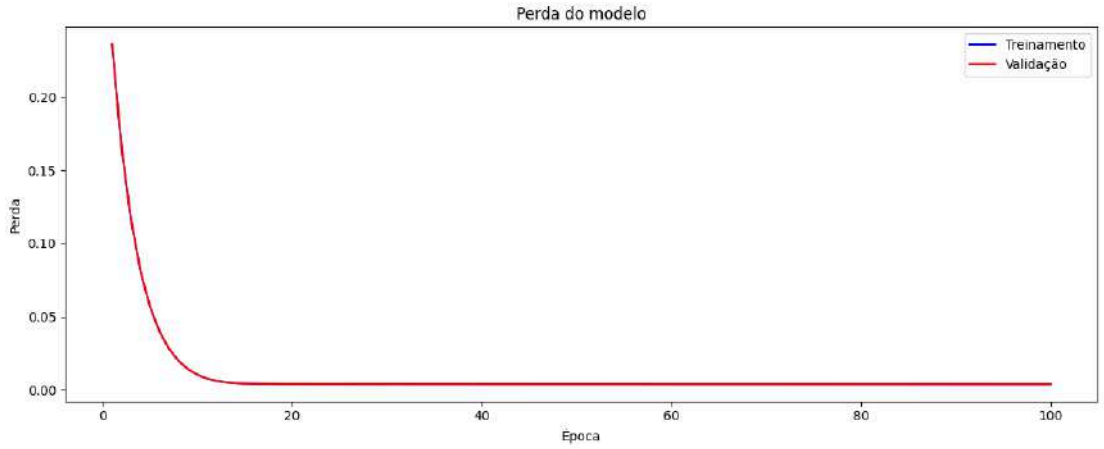

Figura *48* – Curva de perdas função sig e com 2 dados passados. Fonte: A autora.

O gráfico na Figura 49 representa a importância de cada variável para as previsões, sendo que os valores de pressão na *choke* defasadas no tempo, apresentam-se como mais importantes.

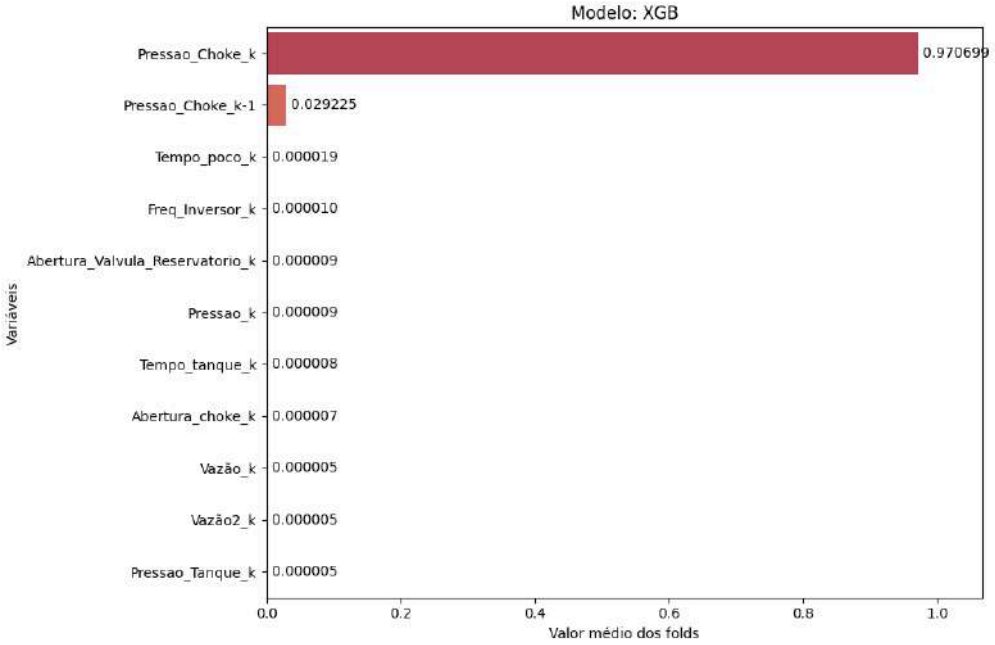

Figura *49* – Importância das variáveis com dados experimentais, tratados com a função sig e com 2 dados passados. Fonte: A autora.

As Figuras 50 a 53, que representam o quanto o modelo consegue prever a operação de PMCD, sendo que os resultados para os demais experimentos se encontram no Anexo B.

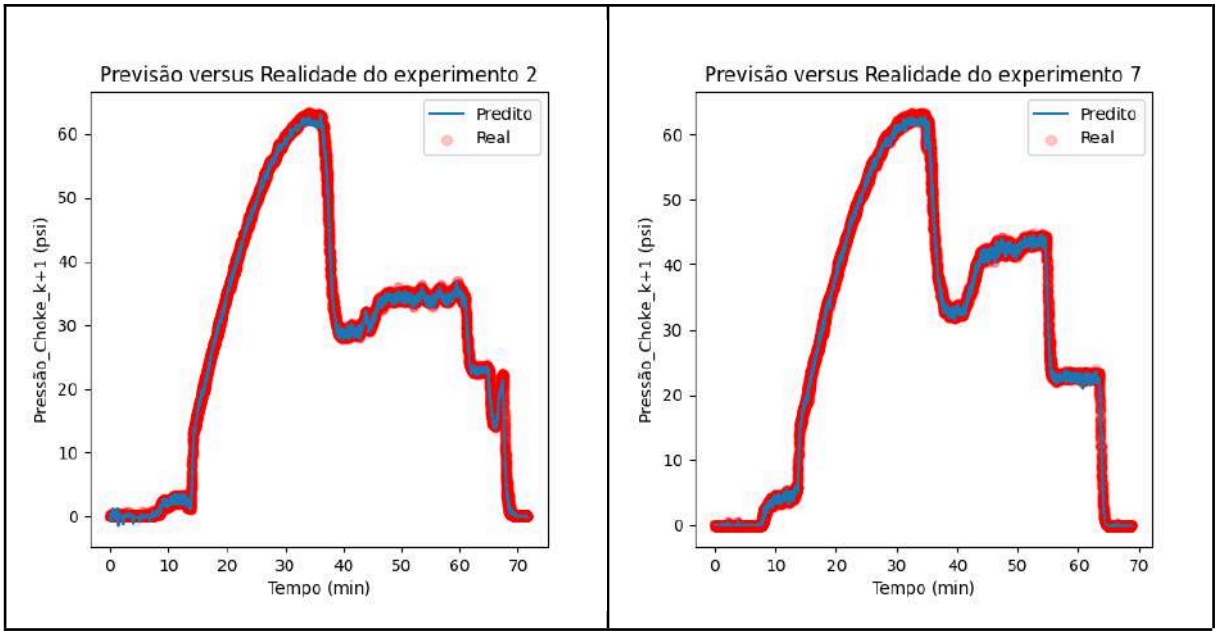

Figura *50* – Previsão versus realidade com dados experimentais, tratados com a função sig e com 2 dados passados, para os experimentos 2 e 7. Fonte: A autora.

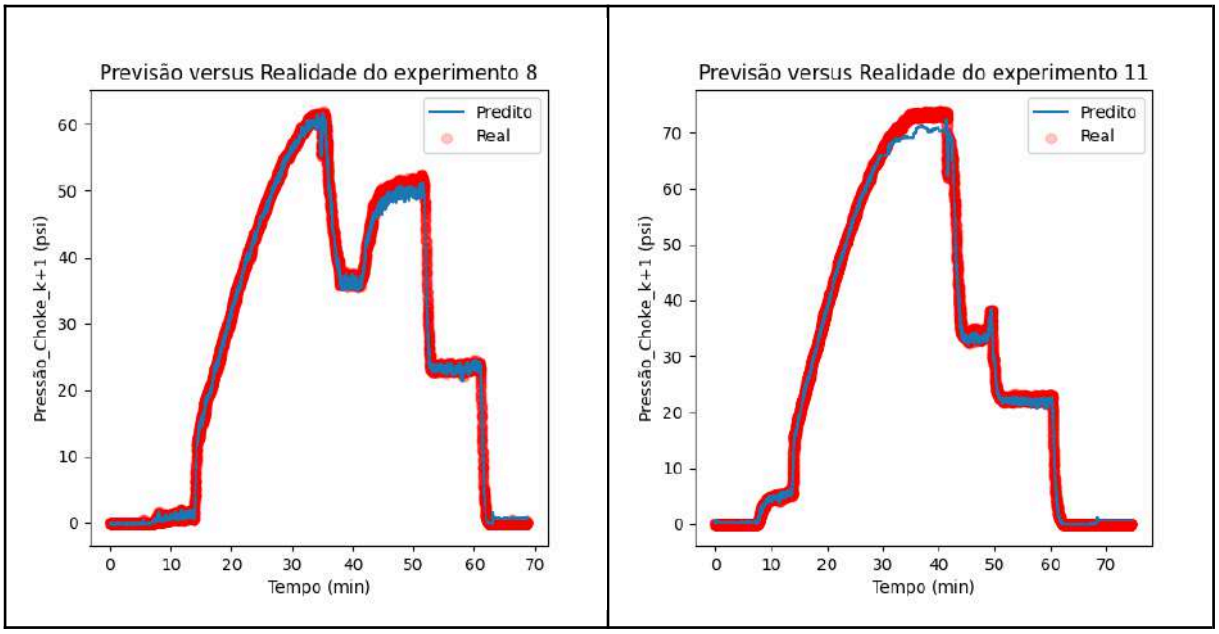

Figura *51* – Previsão versus realidade com dados experimentais, tratados com a função sig e com 2 dados passados, para os experimentos 8 e 11. Fonte: A autora.

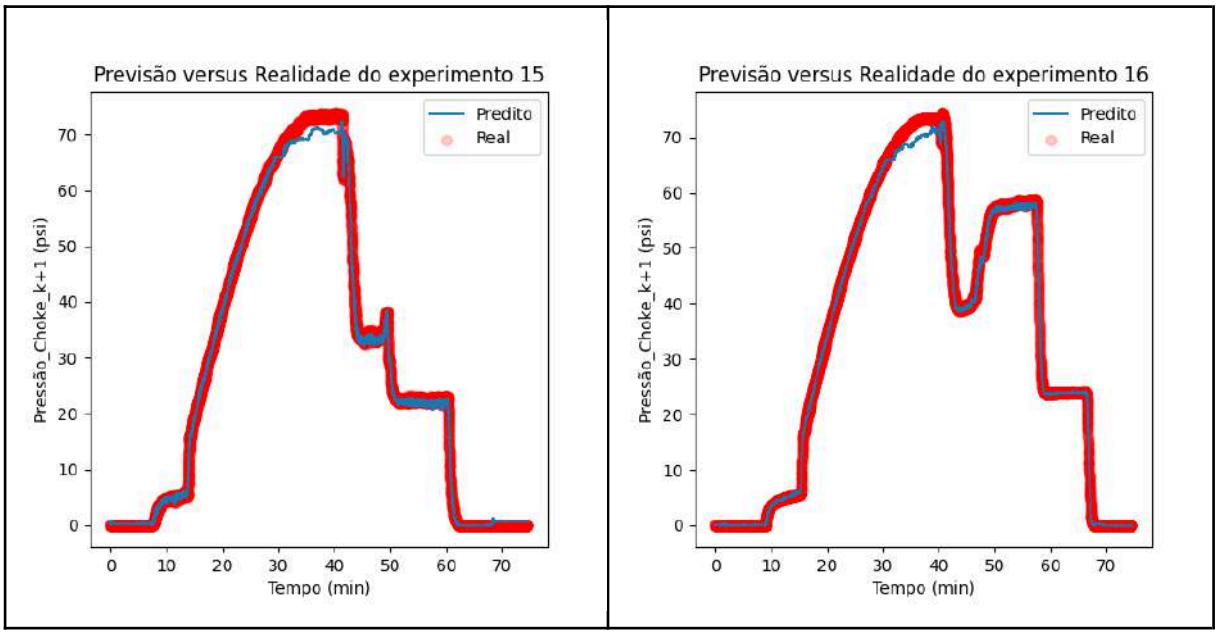

Figura *52* – Previsão versus realidade com dados experimentais, tratados com a função sig e com 2 dados passados, para os experimentos 15 e 16. Fonte: A autora.

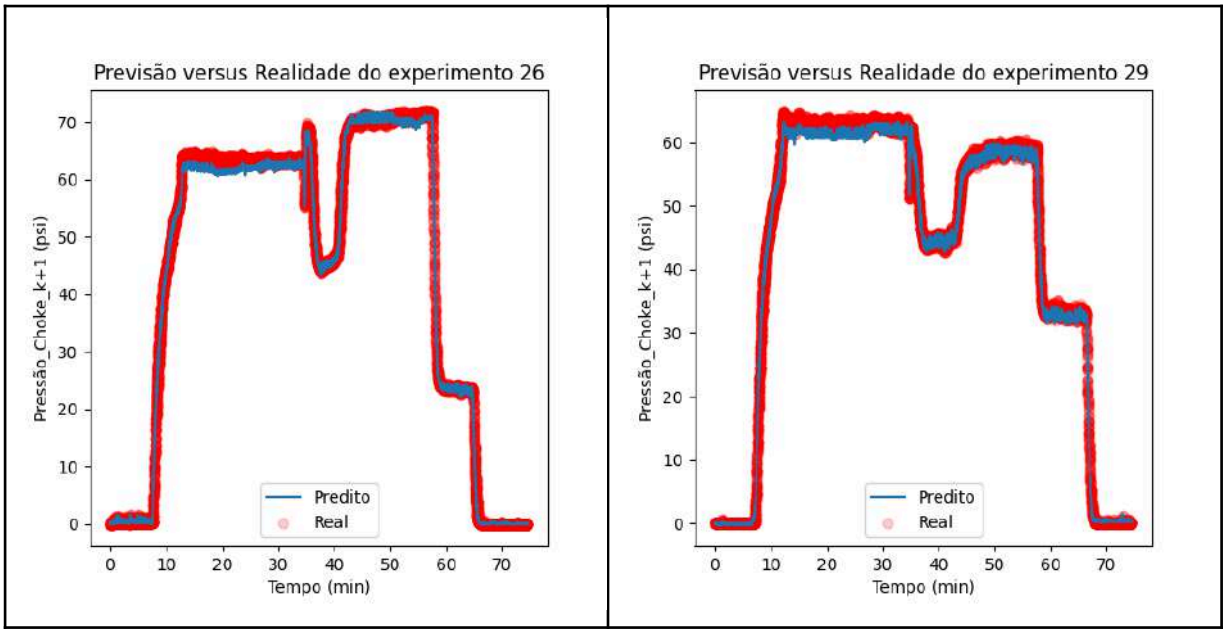

Figura *53* – Previsão versus realidade com dados experimentais, tratados com a função sig e com 2 dados passados, para os experimentos 26 e 29. Fonte: A autora.

Com relação à arquitetura do modelo, para a configuração utilizando a escala de -4 a 4, com a função sig e 2 dados passados, o modelo treinou 875 árvores em cada iteração. Conforme observado nas Figuras 43 a 53, todas as métricas de avaliação do modelo apresentam desempenho superior ao serem introduzidos dados passados, com a introdução de informação sobre a dinâmica do processo.

4.1.3 Dados experimentais tratados com a função sig e com 8 dados passados

Ao empregar 8 dados passados, todas variáveis são deslocadas, conforme mostrado na Figura 54. A vazão e a frequência do inversor, apresentando *outiliers,* foram tratadas com a substituição pelas suas médias.

	Valores_Únicos	Tipo_de_Dado	mean	std	min	25%	50%	75%	max
Pressao_Choke_k-1 (psi)	176138	float64	0.41	0.35	0.02	0.03	0.33	0.78	0.98
Pressao_Choke_k-2 (psi)	176142	float64	0.41	0.35	0.02	0.03	0.33	0.78	0.98
Pressao_Choke_k-3 (psi)	176136	float64	0.41	0.35	0.02	0.03	0.33	0.78	0.98
Pressao_Choke_k-4 (psi)	176132	float64	0.41	0.35	0.02	0.03	0.33	0.78	0.98
Pressao_Choke_k-5 (psi)	176116	float64	0.41	0.35	0.02	0.03	0.33	0.78	0.98
Pressao_Choke_k-6 (psi)	176115	float64	0.41	0.35	0.02	0.03	0.33	0.78	0.98
Pressao_Choke_k-7 (psi)	176113	float64	0.41	0.35	0.02	0.03	0.33	0.78	0.98
Tempo_poco_k (min)	169067	float64	0.40	0.33	0.02	0.08	0.30	0.72	0.98
Pressao_k (psi)	195323	float64	0.44	0.36	0.02	0.10	0.34	0.87	0.98
Vazão_k (m³/h)	148541	float64	0.04	0.09	0.02	0.02	0.02	0.03	0.98
Freq_Inversor_k (Hz)	6	float64	0.09	0.20	0.02	0.02	0.02	0.02	0.98
Abertura_choke_k (%)	198	float64	0.64	0.44	0.02	0.02	0.97	0.97	0.98
Vazão2_k (m³/h)	147826	float64	0.04	0.09	0.02	0.02	0.02	0.03	0.98
Abertura_Valvula_Reservatorio_k (%)	4	float64	0.39	0.43	0.02	0.02	0.14	0.98	0.98
Tempo_tanque_k (min)	387348	float64	0.40	0.33	0.02	0.08	0.30	0.72	0.98
Pressao_Tanque_k (psi)	353956	float64	0.54	0.38	0.02	0.12	0.58	0.94	0.98
Pressao_Choke_k (psi)	176071	float64	0.41	0.35	0.02	0.03	0.33	0.78	0.98
Pressão_Choke_k+1 (psi)	155655	float64	35.61	26.45	0.00	4.42	36.32	60.22	94.68

Figura *54* – Resumo do *dataframe* com dados experimentais, tratados com a função sig e com 8 dados passados. Fonte: A autora.

A otimização dos hiperparâmetros com o Optuna, para posterior utilização do XGBoost, fornece os seguintes parâmetros: 'n_estimators': 463, 'learning_rate': 0.1872, 'max_depth': 8, 'min_child_weight': 3, 'subsample': 0.7219, 'colsample_bytree': 0.9018, 'gamma': 0.3002, 'reg_alpha': 0.9437 e 'reg_lambda': 0.0624, em apenas 3 minutos e 42 segundos de execução, utilizando o mesmo ambiente de execução do primeiro modelo apresentado.

Com os melhores parâmetros gerados pelo Optuna, os dados são treinados, validados e testados pelo XGBoost, gerando as métricas apresentadas nas Figuras 55 e 56.

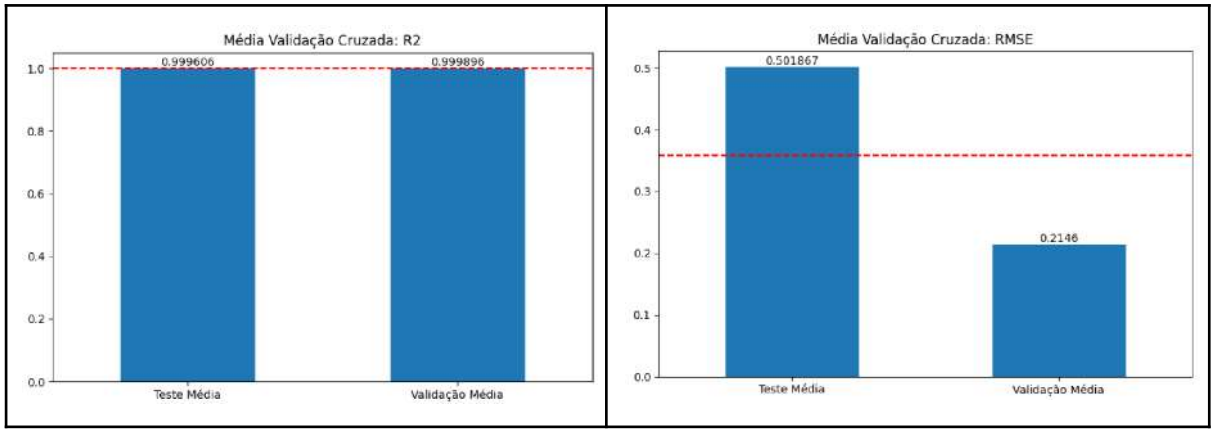

Figura *55* – Métricas de avaliação R² e RMSE com dados experimentais, tratados com a função sig e com 8 dados passados. Fonte: A autora.

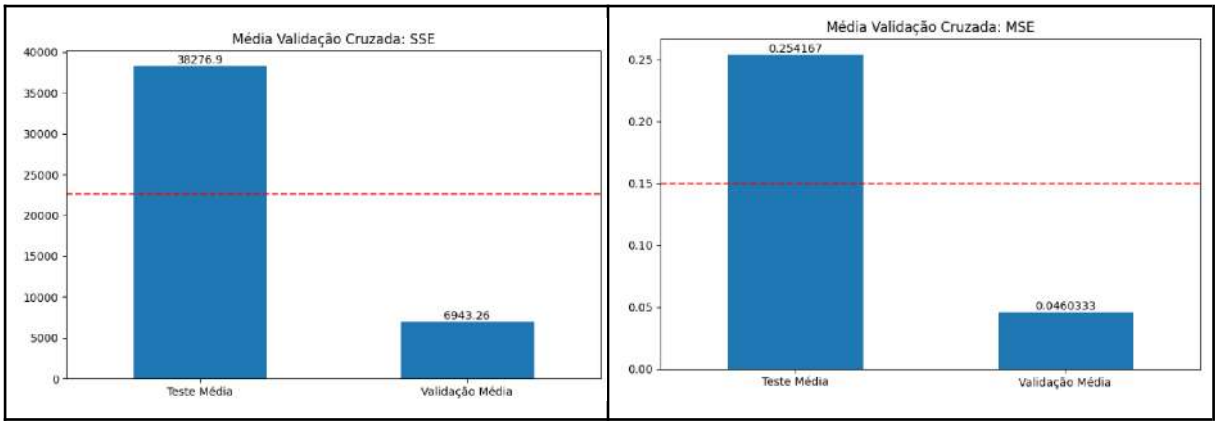

Figura *56* – Métricas de avaliação SSE e MSE com dados experimentais, tratados com a função sig e com 8 dados passados. Fonte: A autora.

São apresentados na Figura 57 os valores da densidade dos dados de pressão na *choke* para o teste e a validação do modelo, sendo que quanto mais valores preditos de forma correta, mais será observada a presença da cor roxa nos gráficos.

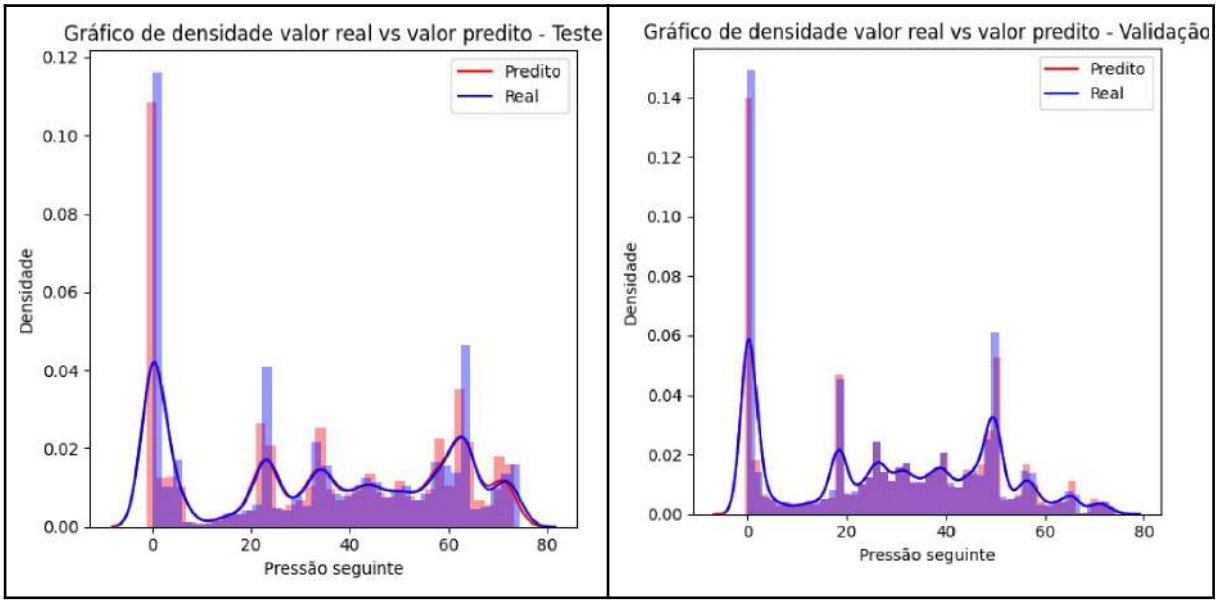

Figura *57* – Gráficos de densidade com dados experimentais, tratados com a função sig e com 8 dados passados. Fonte: A autora.

A Figura 58 apresenta os valores da distribuição dos resíduos no teste e na validação. Observa-se que a maior parte dos erros obtidos estão próximos de 0.

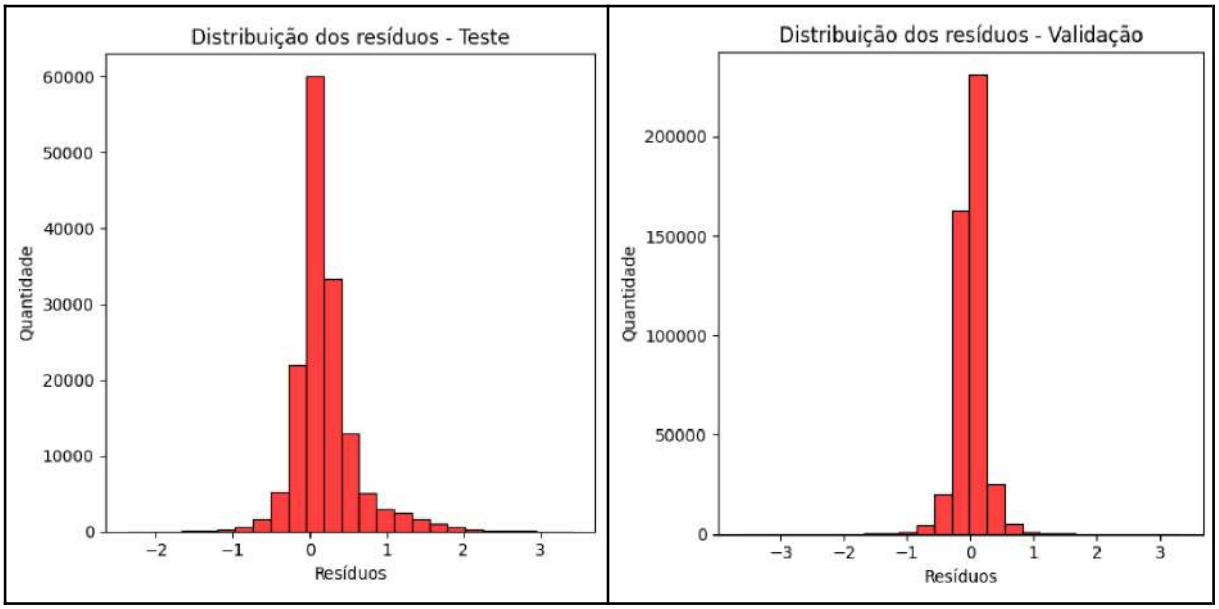

Figura *58* – Distribuição dos resíduos com dados experimentais, tratados com a função sig e com 8 dados passados. Fonte: A autora.

A Figura 59 apresenta os pontos na curva de comparação de valor real e valor predito, além de apresentar valores de R^2. Pode ser observado a proximidade dos pontos em relação à diagonal em vermelho, evidenciando a adequação do modelo gerado via *machine learning*.

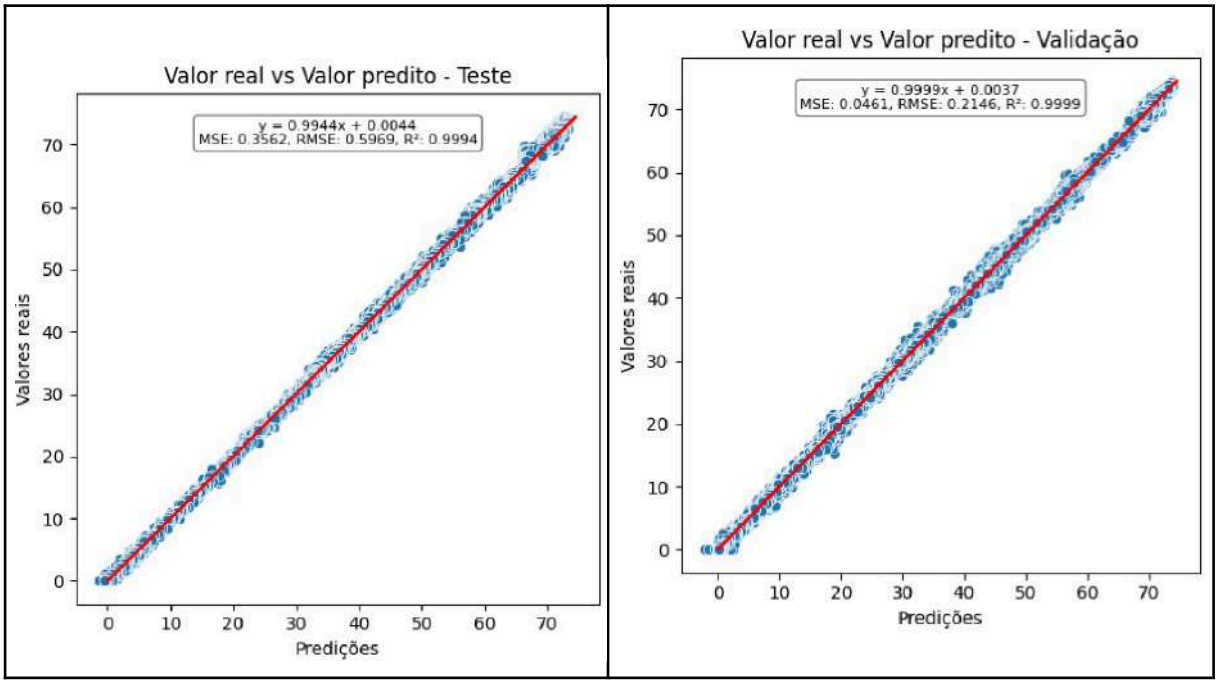

Figura *59* – Gráfico de evolução do modelo com dados experimentais, tratados com a função sig e com 8 dados passados. Fonte: A autora.

O gráfico da função de perda ao longo das iterações apresentado na Figura 60. Os resultados do modelo são satisfatórios por se aproximarem de 0 ao longo das épocas. O Anexo C contém todas as demais simulações referentes às métricas de avaliação.

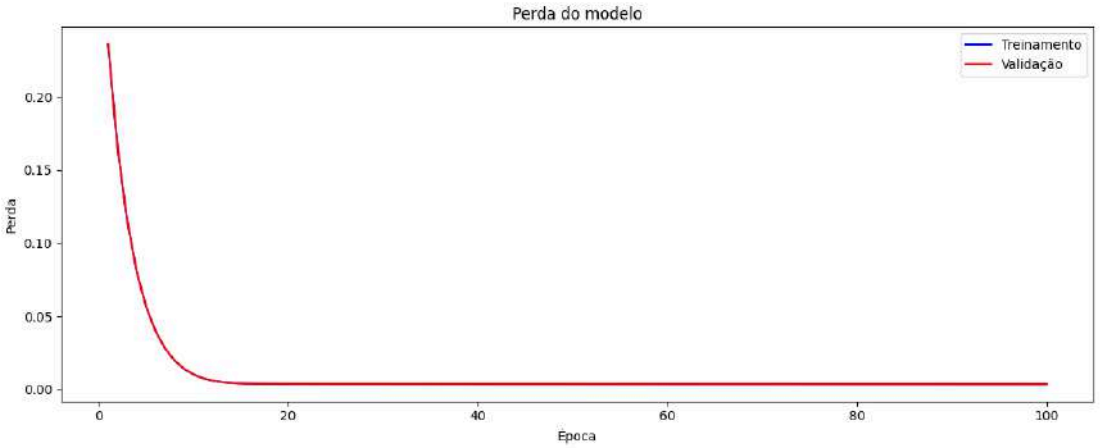

Figura *60* – Curva de perdas função sig e com 8 dados passados. Fonte: A autora.

O gráfico na Figura 61 representa a importância de cada variável para as previsões. Revelando que as variáveis primordiais são as pressões na *choke* defasadas no tempo.

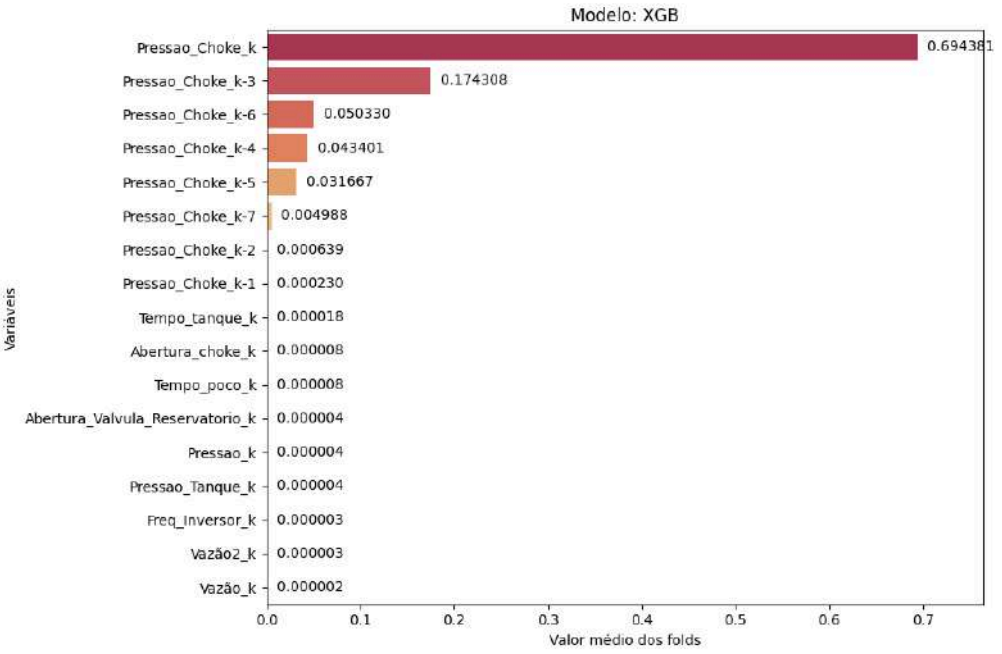

Figura *61* – Importância das variáveis com dados experimentais, tratados com a função sig e com 8 dados passados. Fonte: A autora.

Os resultados das previsões e realidade dos testes do modelo são apresentados nas Figuras 62 a 65, indicando quanto o modelo consegue prever a operação de PMCD. Os resultados para as validações estão registrados no Anexo C.

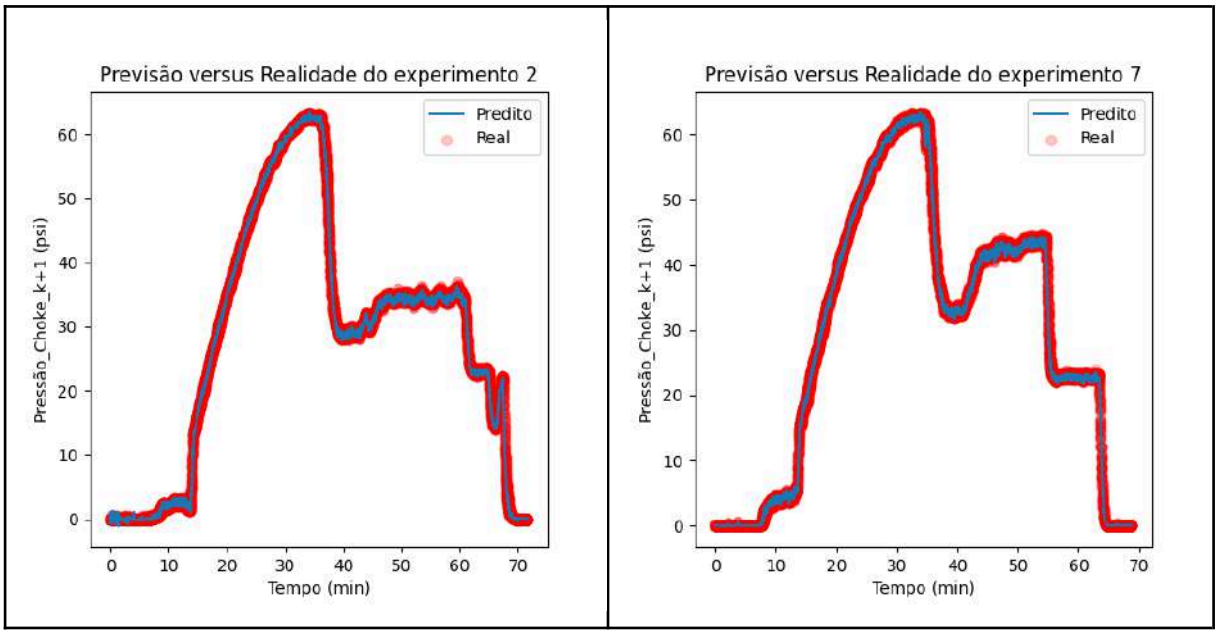

Figura *62* – Previsão versus realidade com dados experimentais, tratados com a função sig e com 8 dados passados, para os experimentos 2 e 7. Fonte: A autora.

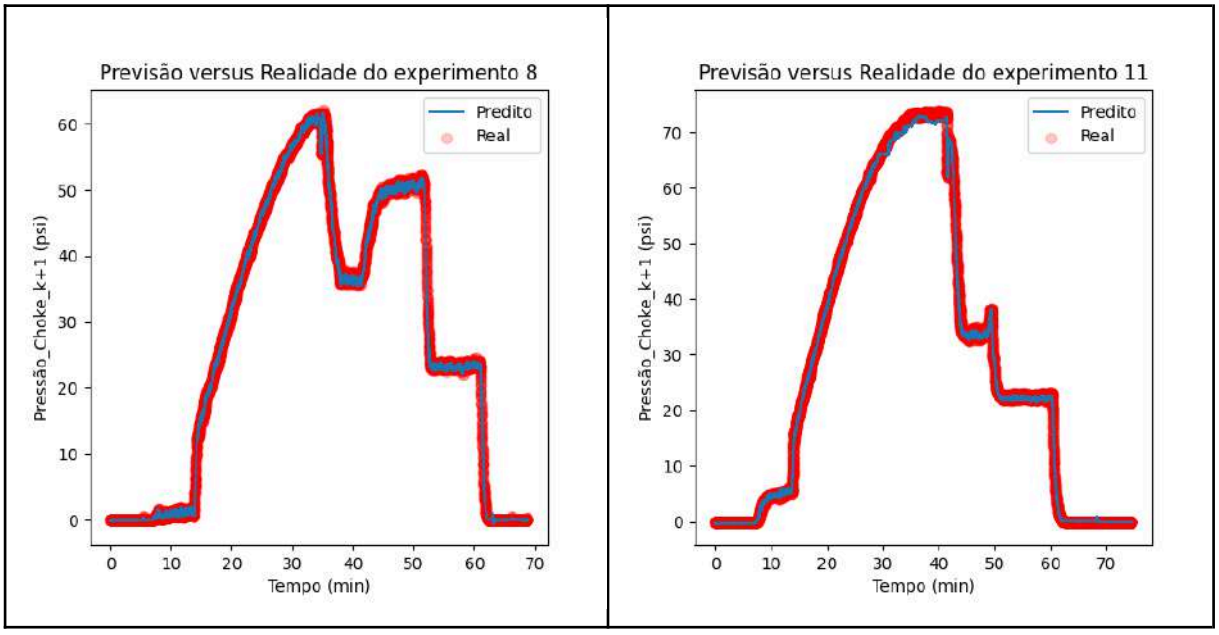

Figura *63* – Previsão versus realidade com dados experimentais, tratados com a função sig e com 8 dados passados, para os experimentos 8 e 11 Fonte: A autora.

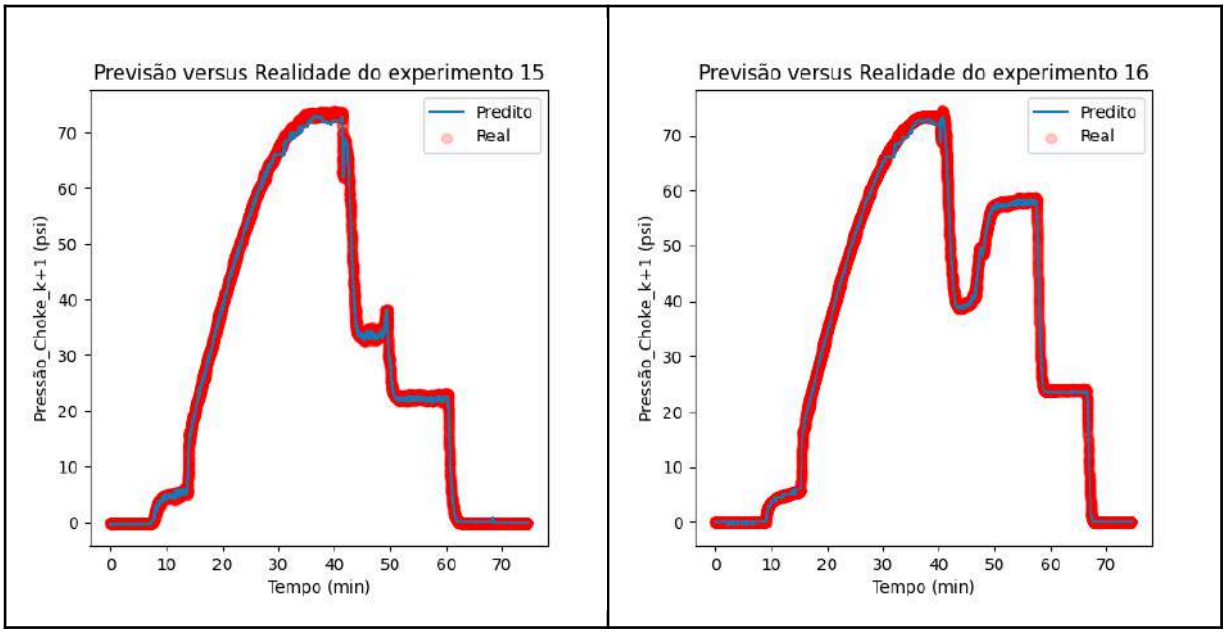

Figura *64* – Previsão versus realidade com dados experimentais, tratados com a função sig e com 8 dados passados, para os experimentos 15 e 16. Fonte: A autora.

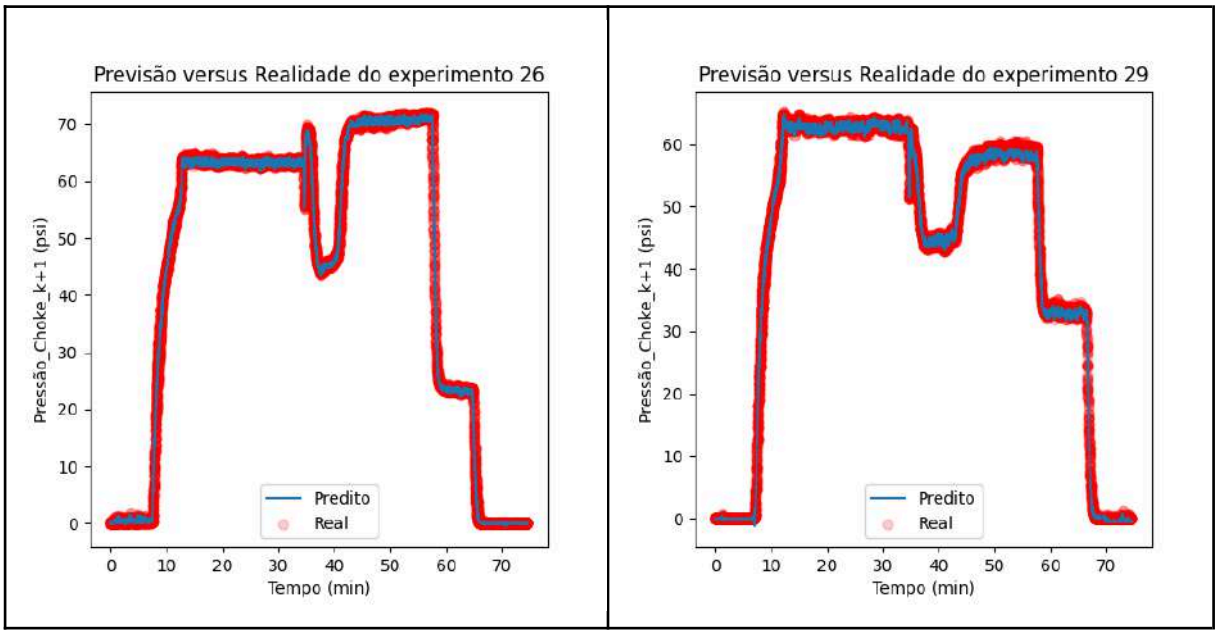

Figura *65* – Previsão versus realidade com dados experimentais, tratados com a função sig e com 8 dados passados, para os experimentos 26 e 29. Fonte: A autora.

Com relação à arquitetura do modelo, para a configuração utilizando a escala de -4 a 4, junto com a função sig e com 8 dados passados, o modelo treinou 463 árvores.

A análise comparativa dos resultados revela que a introdução de 8 dados passados, todas as métricas de avaliação do modelo foram satisfatórias.

4.1.4 Análise comparativa dos resultados dos dados experimentais com aplicação de 8 dados passados e tratados com a função sig com as escalas de 0 a 1, de -4 a 4 e de 0 a 4

Com base no resultado do modelo tratado com a função sig e com 8 dados passados, que apresentou o melhor aprendizado, foram desenvolvidos outros modelos, com o objetivo de avaliar se a mudança da padronização da escala interfere na aprendizagem. Os valores das suas métricas de avaliação da validação estão apresentados na Tabelas 1 e as métricas de avaliação do teste são apresentados na Tabela 2.

Tabela 1 – Métricas de avaliação da validação do modelo com 8 dados passados e tratados com a função sig em diferentes escalas. Fonte: A autora.

Métricas da Validação	Escala (0, 1)	Escala (-4, 4)	Escala (0, 4)
R^2	0.9999	0.9999	0.9999
RMSE	0.2117	0.2146	0.2126
MSE	0.0448	0.0461	0.0452

Tabela 2 – Métricas de avaliação do teste do modelo com 8 dados passados e tratados com a função sig em diferentes escalas. Fonte: A autora.

Métricas do Teste	Escala (0, 1)	Escala (-4, 4)	Escala (0, 4)
R^2	0.9994	0.9994	0.9994
RMSE	0.6422	0.5969	0.6349
MSE	0.4124	0.3562	0.4032

Por fim, são mostrados na Figura 66 os resultados da previsão versus realidade do experimento 11, sendo a imagem mais à esquerda com escala de 0 a 1, a do meio com escala de -4 a 4 e à direita com escala de 0 a 4. Todos os resultados apresentaram desempenho satisfatório. Dessa forma, o tratamento de dados com diferentes escalas não modifica significativamente a qualidade dos resultados.

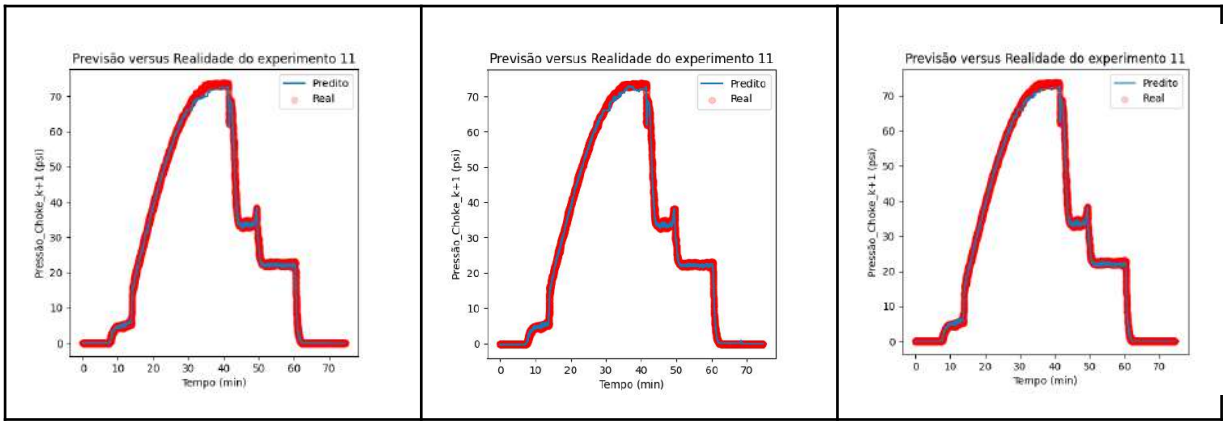

Figura *66* – Previsão versus realidade com dados experimentais, com 8 dados passados e tratados com a função sig em diferentes escalas. Fonte: A autora.

As Tabelas 3 a 5 ilustram os tempos computacionais requeridos para a construção dos modelos baseados em *machine learning* a partir dos dados experimentais de Carvalho (2018).

Tabela 3 – Tempos computacionais dos modelos com dados experimentais com sig para execução do Optuna. Fonte: A autora.

Modelos com dados experimentais com sig			
Tempo Total Optuna	**Escala (0, 1)**	**Escala (-4, 4)**	**Escala (0, 4)**
Sem dados passados	1 min 40 s	2 min 2 s	1 min 55 s
Com 2 dados passados	2 min 6 s	3 min	2 min 12 s
Com 8 dados passados	2 min 48 s	3 min 42 s	2 min 54 s

Tabela 4 – Tempos computacionais dos modelos com dados experimentais com sig para execução do XGBoost. Fonte: A autora.

Modelos com dados experimentais com sig			
Tempo Total XGBoost	**Escala (0, 1)**	**Escala (-4, 4)**	**Escala (0, 4)**
Sem dados passados	43 s	43.1 s	50.9 s
Com 2 dados passados	36.2 s	32 s	45.6 s
Com 8 dados passados	47.1 s	1 min	55.6 s

Tabela 5 – Tempos computacionais dos modelos com dados experimentais com sig para execução do Optuna e o XGBoost. Fonte: A autora.

Modelos com dados experimentais com sig			
Tempo Total Optuna e XGBoost	**Escala (0, 1)**	**Escala (-4, 4)**	**Escala (0, 4)**
Sem dados passados	143 s	165.1 s	165.9 s
Com 2 dados passados	162.2 s	212 s	177.6 s
Com 8 dados passados	215.1 s	282 s	229.6 s

Os Anexos D ao G apresentam os resultados para o tratamento dos dados com a função log.

4.2 Resultado dos dados de poços reais

Os dados de poços reais utilizados para a modelagem, são provenientes dos trabalhos de Jayah (2013) representados pelos experimentos 1 e 5, Zein (2017) representado como experimento 2 e Wattanasuwankorn (2014) representados pelos experimentos 3 e 4, gerando um total de 3.126 dados. Dados de tempo, pressão na *choke* e vazão foram utilizados para treinar, validar e testar o algoritmo de *machine learning* para regressão, com o uso do XGBoost e tendo a pressão *choke* no instante seguinte como variável alvo. O resumo dos dados é apresentado na Figura 60, onde são mostradas suas características com o total de valores únicos, o tipo de dado, suas médias, seus desvios padrão, valores mínimos, seus quartis e valores máximos. Observa-se que os dados reais apresentam maior desvio padrão que os dados do laboratório (Figura 22), indicando maior não uniformidade, conferindo maior complexidade para fins de síntese de modelos matemáticos.

	Valores_Únicos	Tipo_de_Dado	mean	std	min	25%	50%	75%	max
Tempo	3005	float64	4580.41	4068.73	0.0	1094.43	3480.61	7367.70	17221.00
Pressao_Choke	2314	float64	2886542.70	3760337.18	0.0	616845.46	1249794.79	2784078.96	13797026.86
Vazao	548	float64	0.01	0.11	0.0	0.00	0.00	0.00	4.50
experimento	5	int64	3.03	1.33	1.0	2.00	3.00	4.00	5.00

Figura *67* – Conjunto de dados de poços reais. Fonte: A autora.

Dessa etapa em diante foram feitos testes sem dados passados, com 2, 8 e 20 dados passados, sendo todos com a transformação dos dados utilizando as funções sig e log com aplicação da padronização da escala de 0 a 4 para a função sig e de 0 a 10 para a função log.

Ao se aplicar a quantidade de dados passados, são criadas novas variáveis. Foi construído no código uma opção para selecionar a quantidade de dados passados. A Figura 68, mostra o exemplo de dados sem a aplicação de dados passados.

	Tempo	Pressao_Choke	Vazao	experimento
0	0.00	8.607900e+05	0.0	1
1	65.00	8.695776e+05	0.0	1
2	114.00	8.695933e+05	0.0	1
3	135.00	8.812890e+05	0.0	1
4	179.00	8.766276e+05	0.0	1

Figura *68* – Exemplo de dados de poços reais sem dados passados. Fonte: A autora.

Com o uso de 2 dados passados, todas as variáveis são deslocadas, com isso, foram criadas as variáveis para a pressão *choke* para cada passo anterior, conforme mostrado na Figura 69.

	experimento	Pressao_Choke_k-1	Tempo_k	Vazao_k	Pressao_Choke_k	Pressao_Choke_k+1
0	1	860790.048971	65.0	0.0	869577.552128	869593.347520
1	1	869577.552128	114.0	0.0	869593.347520	881288.957873
2	1	869593.347520	135.0	0.0	881288.957873	876627.562150
3	1	881288.957873	179.0	0.0	876627.562150	888323.172502
4	1	876627.562150	200.0	0.0	888323.172502	894186.773071

Figura *69* – Resumo dos dados de poços reais com 2 dados passados. Fonte: A autora.

A próxima etapa ocorre a divisão dos dados sendo 75% para o treinamento e 25% dos dados para o teste. A Figura 70 apresenta as distribuições da variável alvo (pressão na *choke* no instante seguinte) para os dados de treinamento e teste. O lado esquerdo da imagem é referente ao treinamento e o lado direito ao teste. Verifica-se que essas distribuições não estão próximas, sinalizando a necessidade de mais dados para se obter uma divisão que facilite a previsão quanto aos dados inéditos.

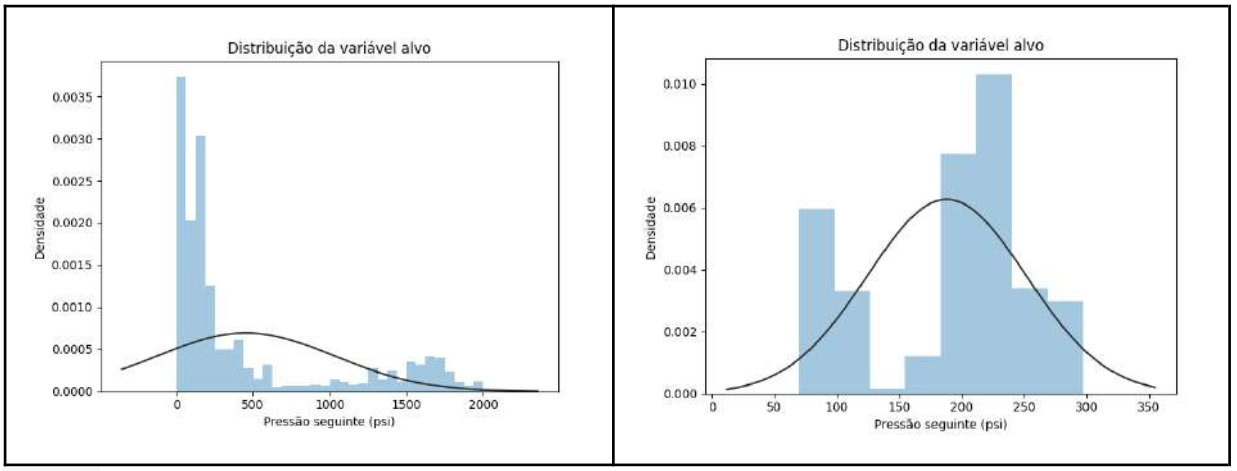

Figura *70* – Distribuição da pressão *choke* seguinte no treino e teste. Fonte: A autora.

Dessa etapa em diante foram feitos testes sem dados passados, com 2, 8 e 20 dados passados.

4.2.1 Dados de poços reais tratados com a função sig e sem dados passados

A análise estatística dos dados de entrada e saída é apresentada na Figura 71. A vazão apresentou *outiliers* tratados com a substituição pela média.

	Valores_Únicos	Tipo_de_Dado	mean	std	min	25%	50%	75%	max
Tempo (min)	3095	float64	0.74	0.15	0.5	0.59	0.77	0.87	0.98
Vazão (bbl/min)	466	float64	0.52	0.10	0.5	0.50	0.50	0.50	0.98
Pressao_Choke (psi)	2315	float64	418.66	545.39	0.0	89.47	181.27	403.80	2001.09
experimento	5	int64	3.03	1.33	1.0	2.00	3.00	4.00	5.00

Figura *71* – Resumo do *dataframe* com dados de poços reais, tratados com a função sig e sem dados passados. Fonte: A autora.

Antes dos dados serem treinados pelo XGBoost, é realizada a otimização dos hiperparâmetros com o Optuna gerando os seguintes parâmetros: 'n_estimators': 859, 'learning_rate': 0.0620, 'max_depth': 5, 'min_child_weight': 7, 'subsample': 0.9436, 'colsample_bytree': 0.9234, 'gamma': 0.5857, 'reg_alpha': 0.2353 e 'reg_lambda': 0.7179, em apenas 1,67 segundos de execução.

Com os melhores parâmetros gerados pelo Optuna, os dados são treinados, validados e testados pelo XGBoost. As métricas de avaliação do modelo são apresentadas na Figura 72, obtidas pela validação cruzada nos testes e na validação do modelo.

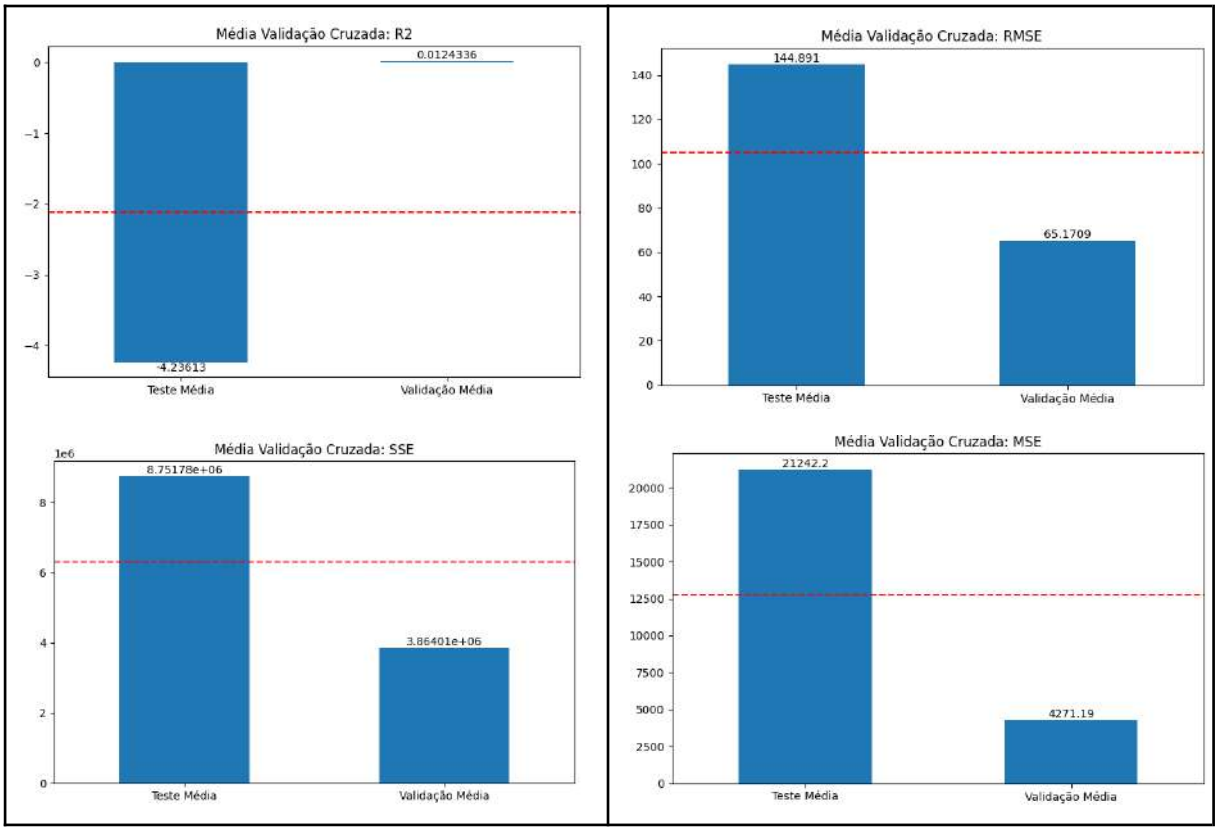

Figura *72* – Métricas de avaliação com dados de poços reais, tratados com a função sig e sem aplicação de dados passados. Fonte: A autora.

São apresentados na Figura 73 os valores da densidade dos dados para a pressão na *choke* no teste e na validação do modelo, sendo que quanto mais valores preditos de forma correta, ou seja, quando valor predito é igual ao valor real, mais será observada a presença da cor roxa nos gráficos.

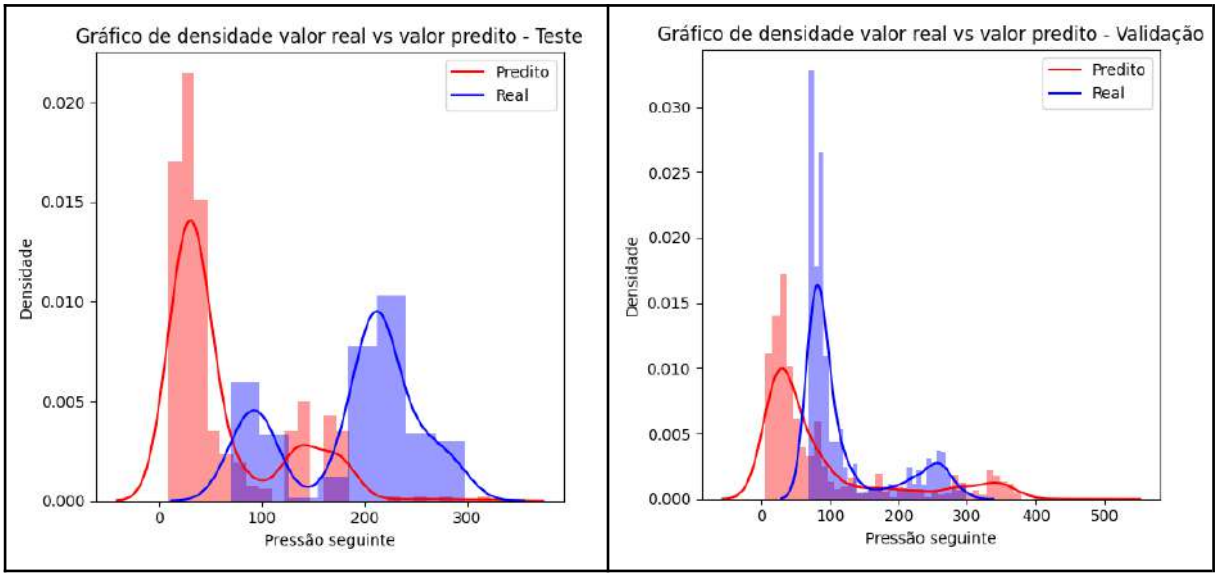

Figura *73* – Gráficos de densidade com dados de poços reais, tratados com a função sig e sem dados passados. Fonte: A autora.

São apresentados na Figura 74 os valores da distribuição dos resíduos no teste e na validação do modelo.

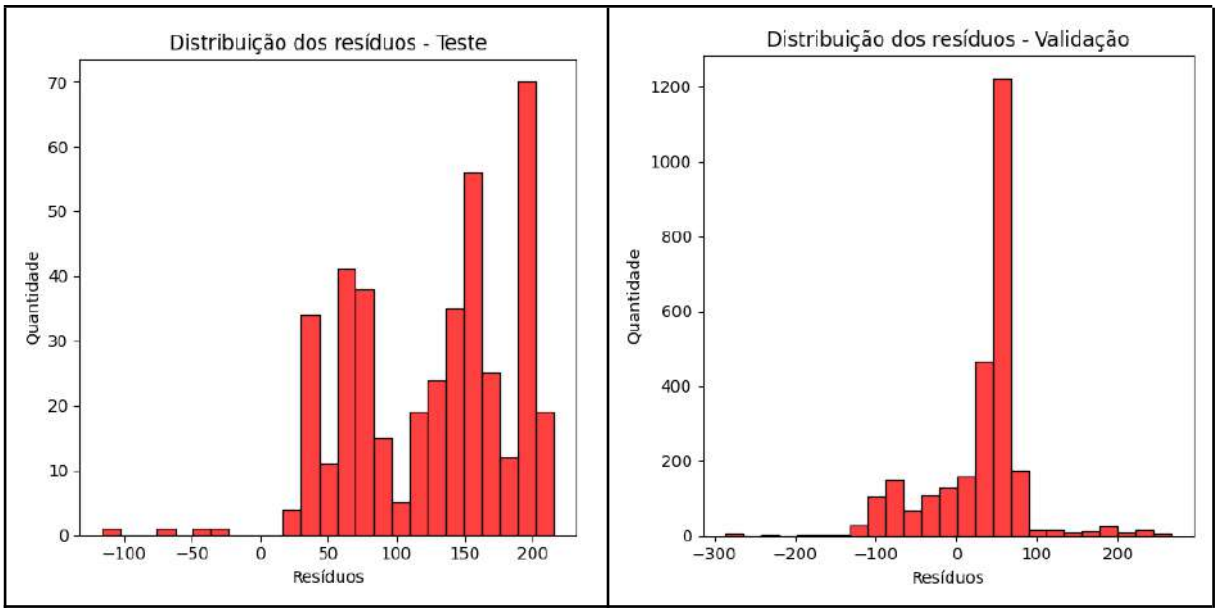

Figura *74* – Distribuição dos resíduos com dados de poços reais, tratados com a função sig e sem dados passados. Fonte: A autora.

A Figura 75 apresenta muitos pontos distantes da curva de comparação de valor real e valor predito apresentando valores de R^2 abaixo de 0,8 no teste e validação e os valores de MSE e de RMSE estão muito altos.

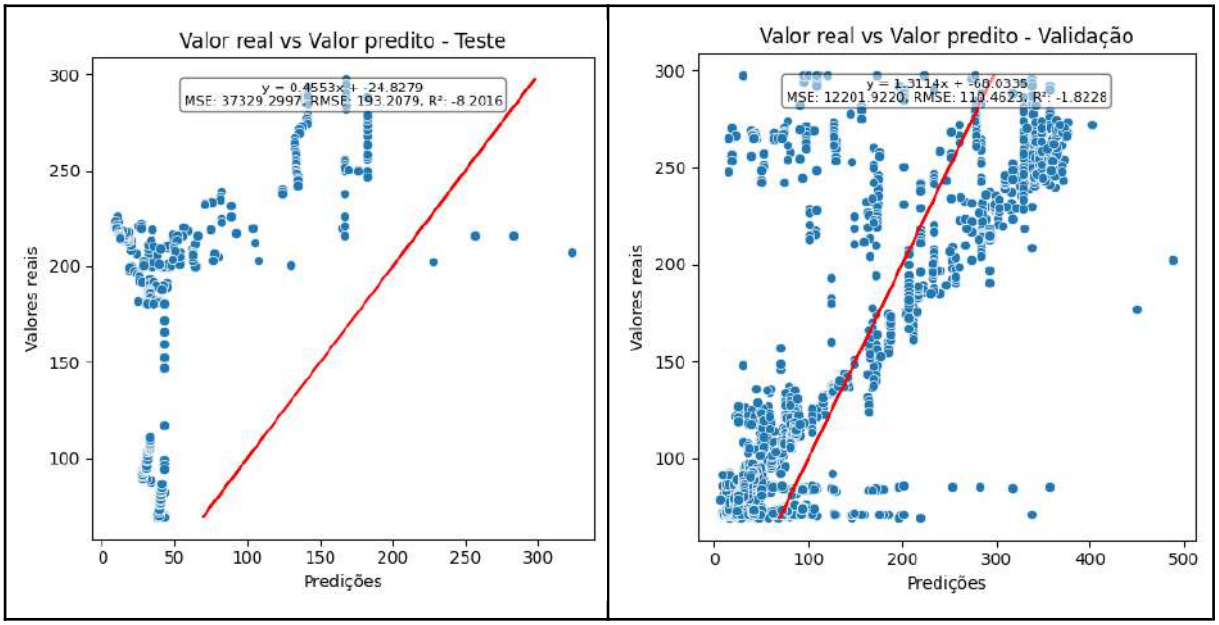

Figura *75* – Gráfico de evolução do modelo com dados de poços reais, tratados com a função sig e sem dados passados. Fonte: A autora.

O gráfico da função de perda é apresentado nas Figuras 76. Os resultados do modelo não são satisfatórios por apresentarem perdas significativas ao longo das épocas.

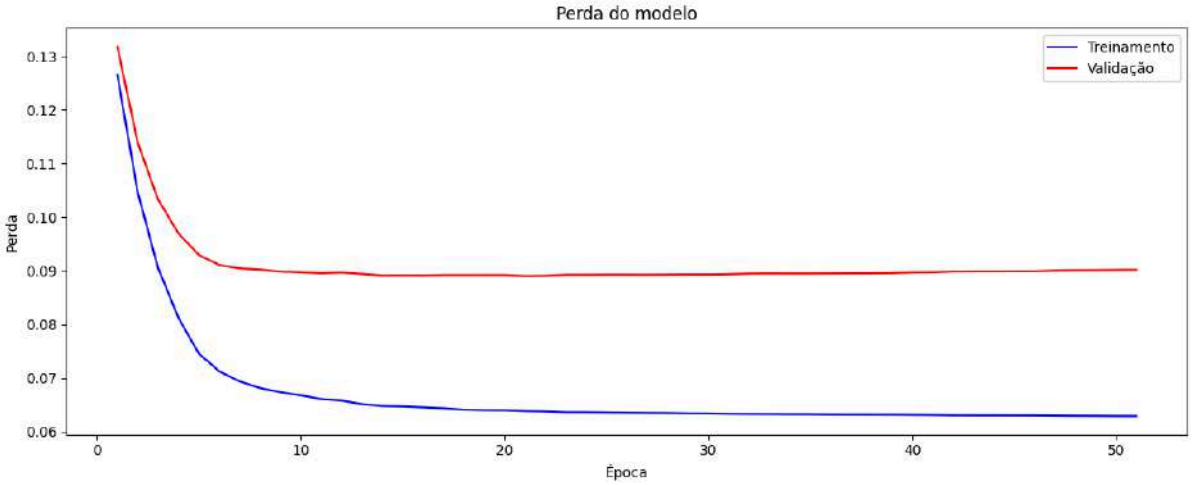

Figura *76* – Curva de perdas função sig e sem dados passados. Fonte: A autora.

O gráfico na Figura 77 representa a importância de cada variável para as previsões, sendo que a vazão, sendo que a vazão foi a variável mais relevante.

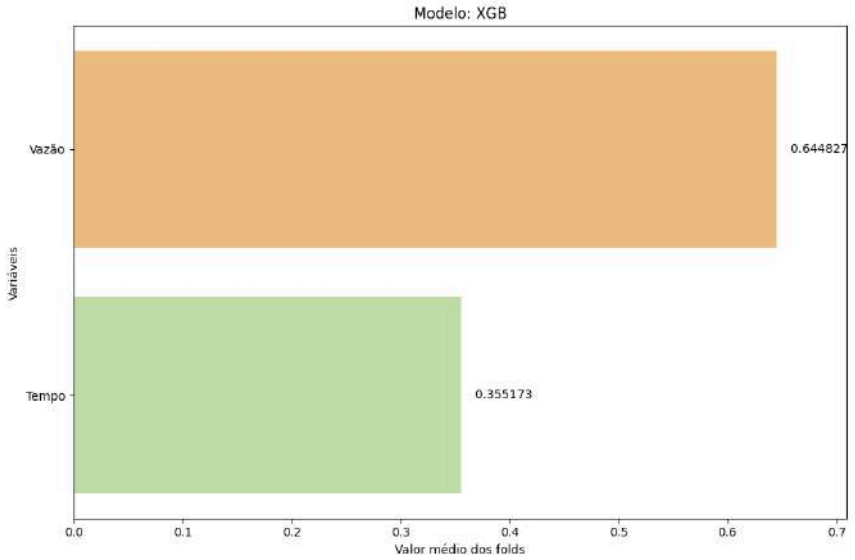

Figura *77* – Importância das variáveis com dados de poços reais, tratados com a função sig e sem dados passados. Fonte: A autora.

Os resultados das previsões e realidade do modelo são apresentados nas Figuras 78 a 80, indicando que o modelo não consegue prever a operação de PMCD adequadamente.

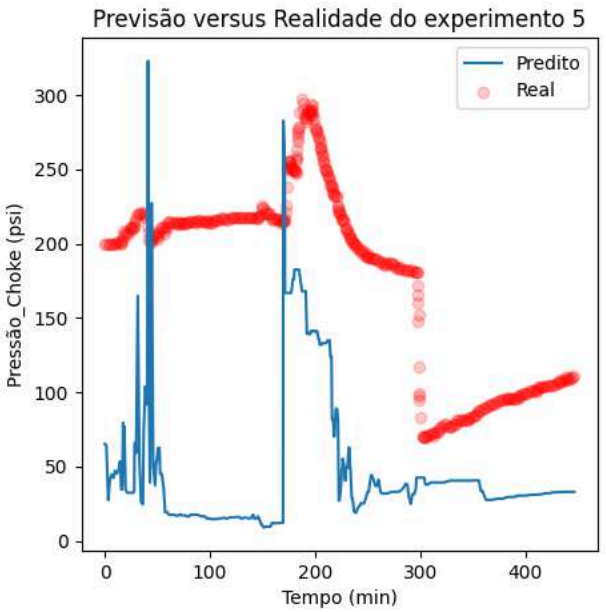

Figura *78* – Previsão versus realidade do teste com dados de poços reais, tratados com a função sig e sem dados passados, para o experimento 5. Fonte: A autora.

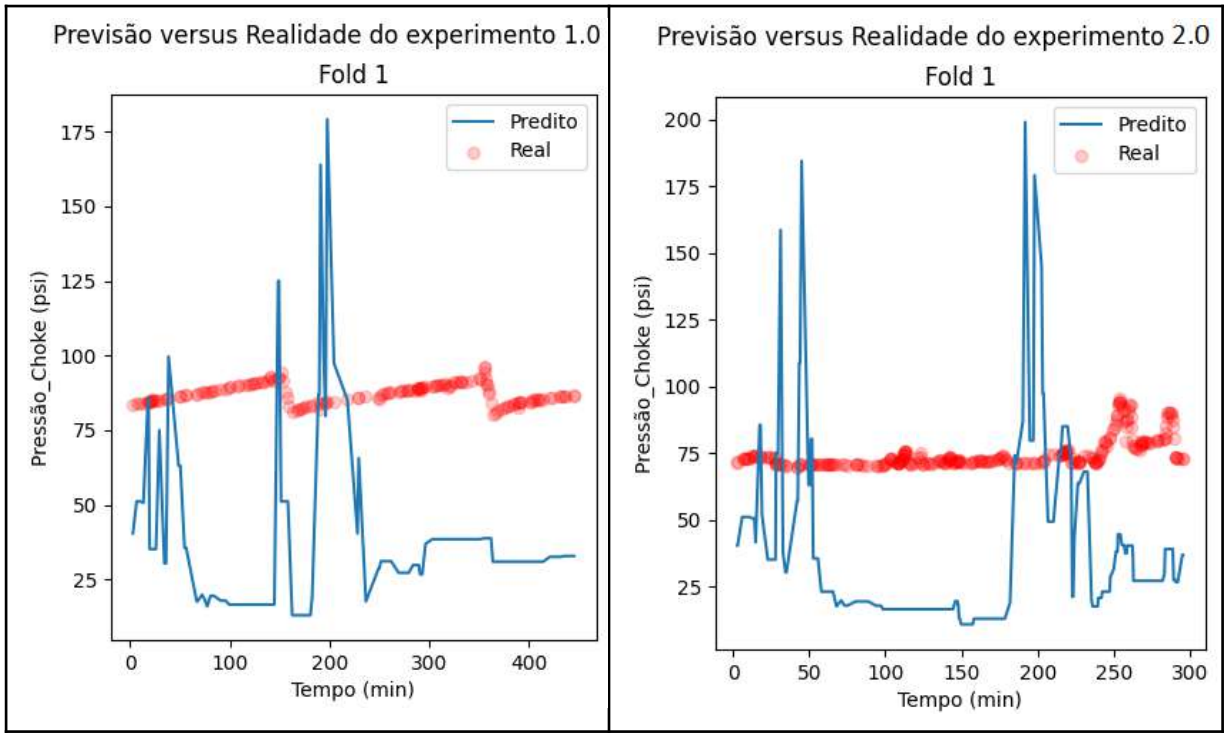

Figura *79* – Previsão versus realidade da validação com dados de poços reais, tratados com a função sig e sem dados passados, para os experimentos 1 e 2. Fonte: A autora.

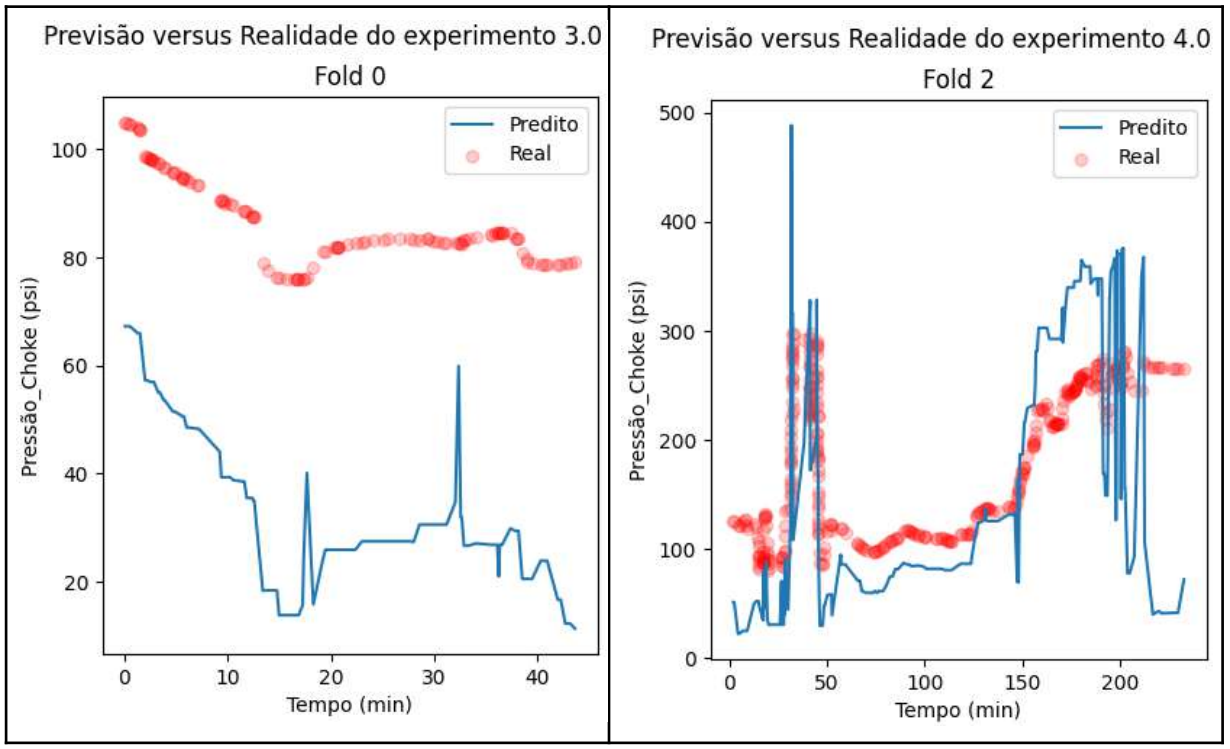

Figura *80* – Previsão versus realidade da validação com dados de poços reais, tratados com a função sig e sem dados passados, para os experimentos 3 e 4. Fonte: A autora.

Com relação à arquitetura do modelo, o modelo treinou 141 árvores, sendo demonstrado na Figura 81 uma parte da árvore. Cada nó contém a decisão de uma variável e as folhas representam a pontuação que será somada a cada amostra para gerar sua previsão ao final. Vale ressaltar que em cada nó os valores das variáveis variam de 0 a 4 por conta da normalização que é realizada antes do treinamento do modelo.

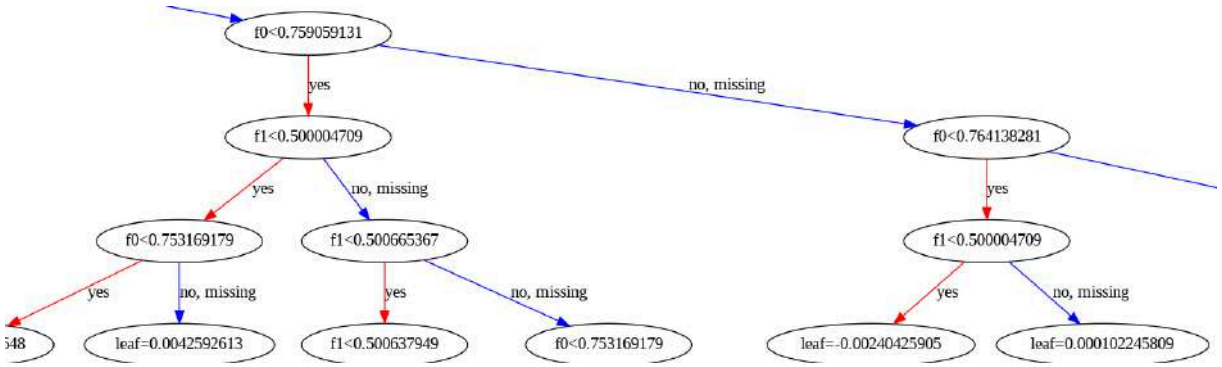

Figura *81* – Parte da árvore gerada pelo modelo com dados de poços reais. Fonte: A autora.

4.2.2 Dados de poços reais tratados com a função sig e com 2 dados passados

Com o uso de 2 dados passados, todas as variáveis são deslocadas, com isso, foram criadas as variáveis para a pressão *choke* em cada passo anterior, conforme mostrado na Figura 82. A vazão possui *outiliers* e foram tratadas com a substituição pelas suas médias.

	Valores_Únicos	Tipo_de_Dado	mean	std	min	25%	50%	75%	max
Pressao_Choke_k-1 (psi)	2308	float64	0.68	0.17	0.5	0.55	0.59	0.90	0.98
Tempo_k (min)	3089	float64	0.74	0.15	0.5	0.59	0.77	0.87	0.98
Vazão_k (bbl/min)	465	float64	0.52	0.10	0.5	0.50	0.50	0.50	0.98
Pressao_Choke_k (psi)	2309	float64	0.68	0.17	0.5	0.55	0.59	0.90	0.98
Pressao_Choke_k+1 (psi)	2311	float64	419.06	545.76	0.0	89.57	181.46	403.59	2001.09
experimento	5	int64	3.03	1.32	1.0	2.00	3.00	4.00	5.00

Figura *82* – Resumo do *dataframe* com dados de poços reais, tratados com a função sig e com 2 dados passados. Fonte: A autora.

Antes dos dados serem treinados pelo XGBoost, é feito a otimização dos hiperparâmetros com o Optuna gerando os seguintes parâmetros: 'n_estimators': 815, 'learning_rate': 0.0137, 'max_depth': 14, 'min_child_weight': 5, 'subsample': 0.6471, 'colsample_bytree': 0.7761, 'gamma': 0.1243, 'reg_alpha': 0.7441 e 'reg_lambda': 0.2376, em 2.49 segundos de execução.

Com os melhores parâmetros gerados pelo Optuna, os dados são treinados, validados e testados pelo XGBoost. As métricas de avaliação do modelo são apresentadas na Figura 83.

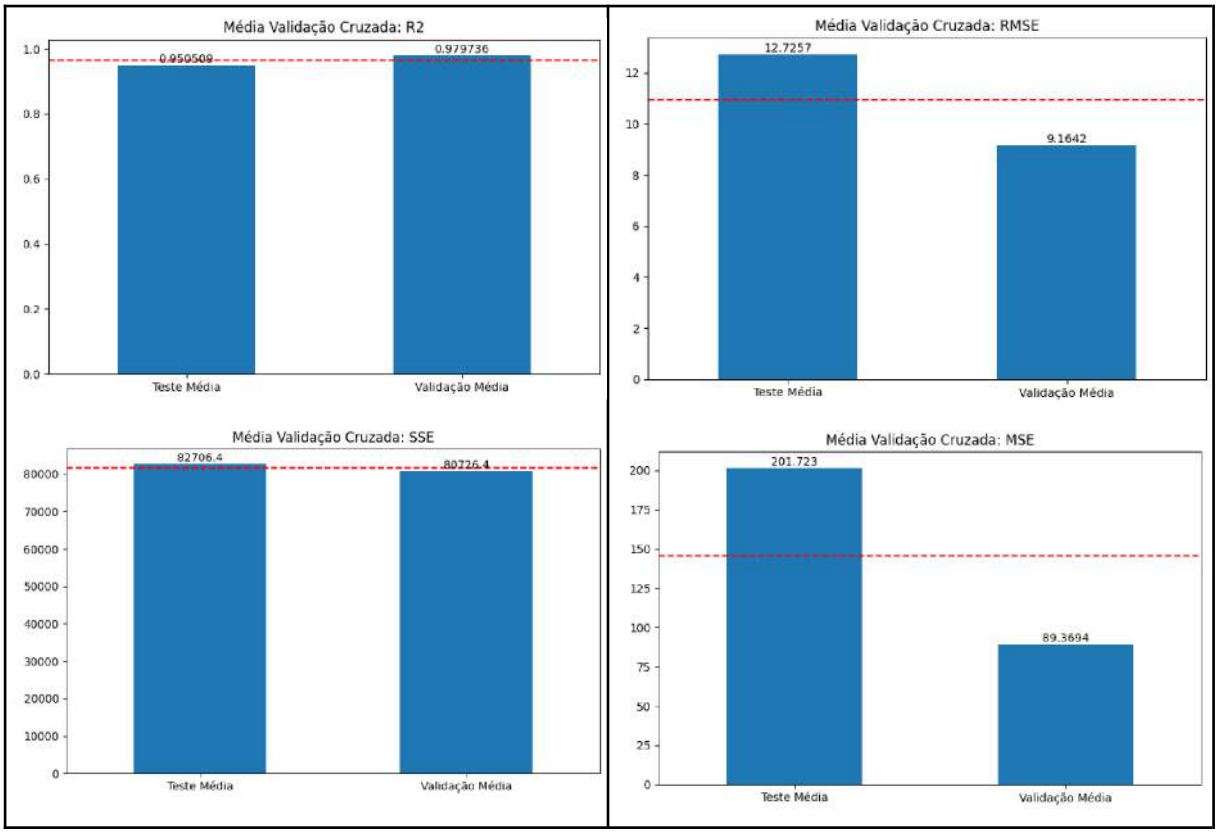

Figura *83* – Métricas de avaliação com dados de poços reais, tratados com a função sig e com 2 dados passados. Fonte: A autora.

São apresentados na Figura 84 os valores da densidade dos dados para a pressão na *choke*, sendo que quanto mais valores preditos de forma correta, mais será observada a presença da cor roxa nos gráficos.

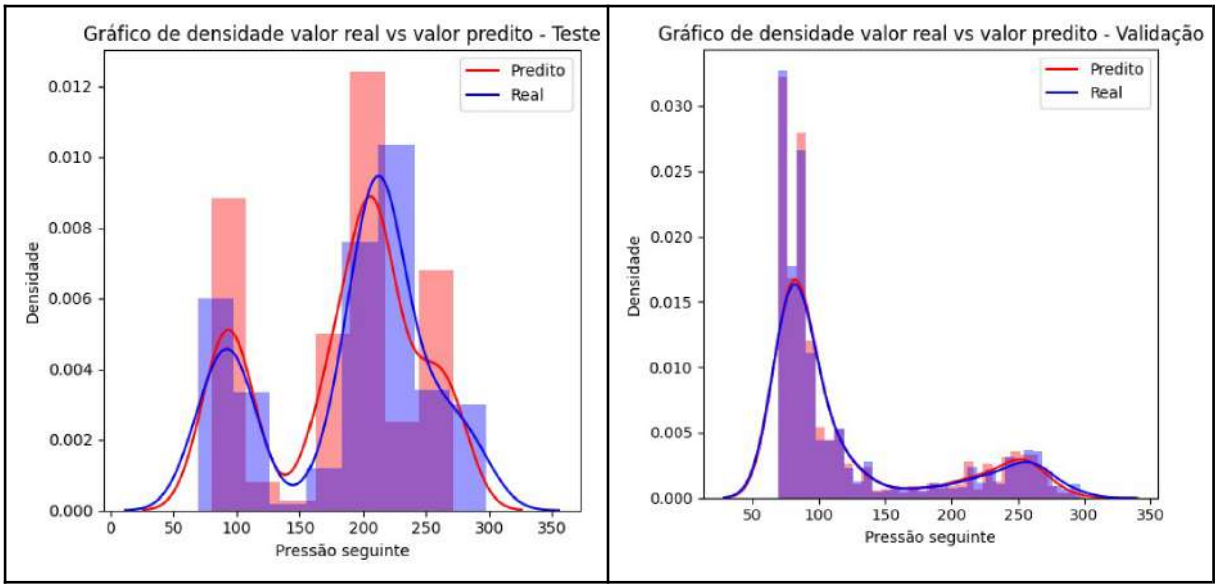

Figura *84* – Gráficos de densidade com dados de poços reais, tratados com a função sig e com 2 dados passados. Fonte: A autora.

São apresentados na Figura 85 os valores da distribuição dos resíduos no teste e na validação do modelo.

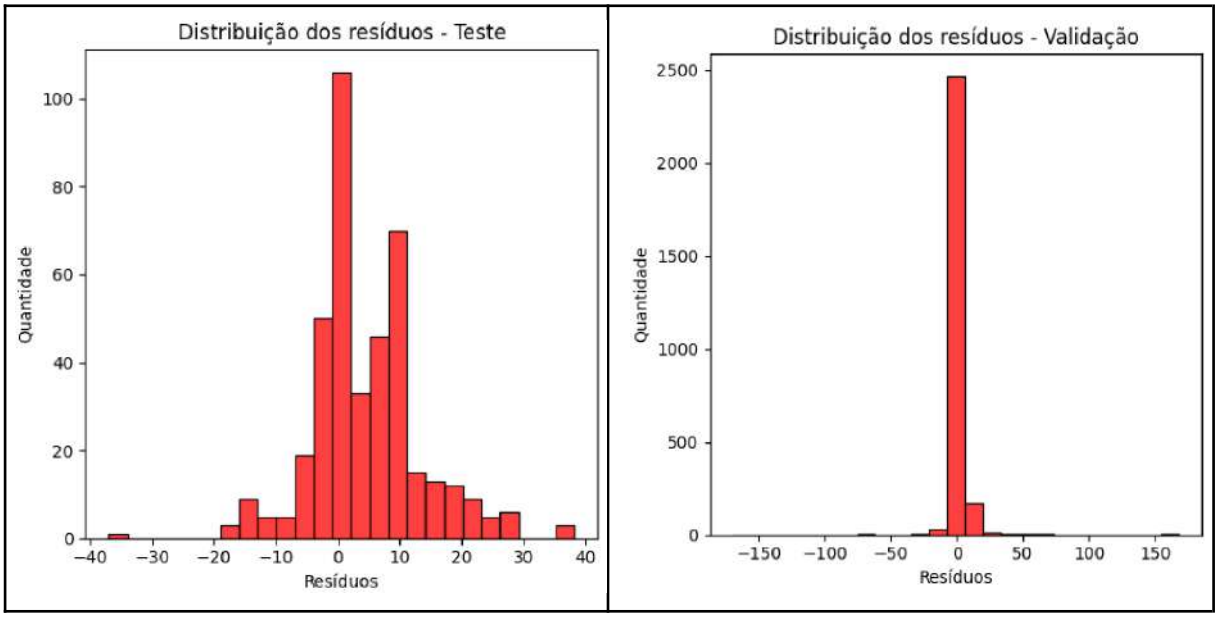

Figura *85* – Distribuição dos resíduos com dados de poços reais, tratados com a função sig e com 2 dados passados. Fonte: A autora.

Na Figura 86 apresenta os pontos na curva de comparação de valor real e valor predito.

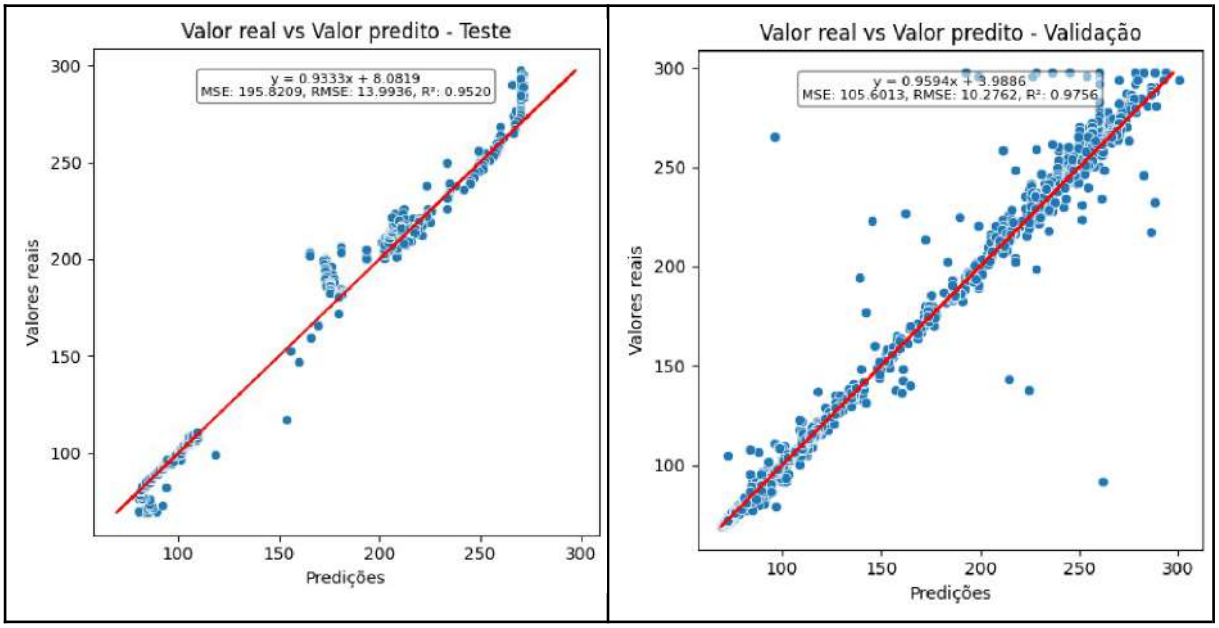

Figura *86* – Gráfico da evolução do modelo com dados de poços reais, tratados com a função sig e com 2 dados passados. Fonte: A autora.

O gráfico da função de perda é apresentado na Figura 87.

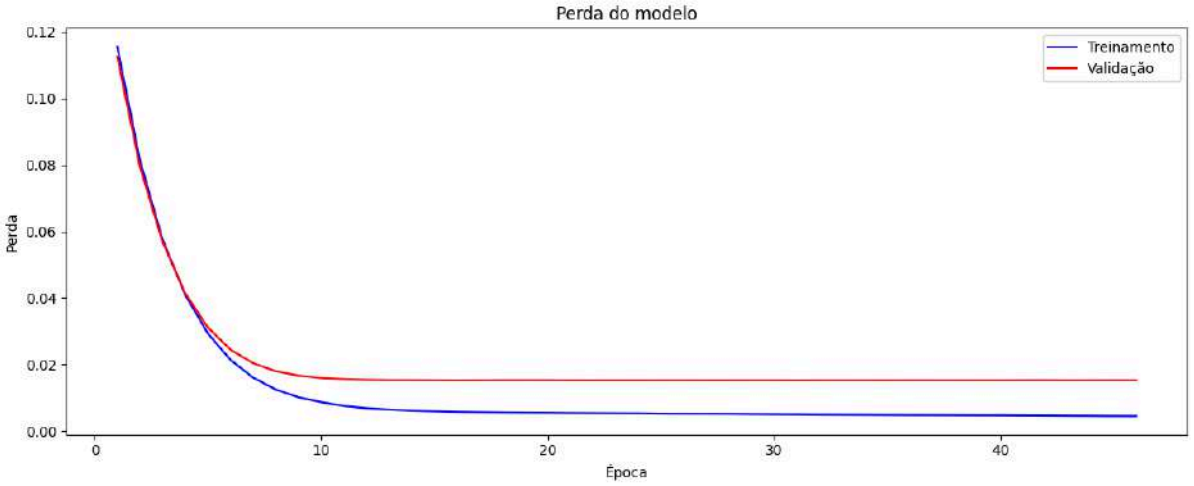

Figura *87* – Curva de perdas função sig e com 2 dados passados. Fonte: A autora.

O gráfico na Figura 88 representa a importância de cada variável para as previsões, sendo que a pressão da *choke* foi a variável mais relevante para a construção do modelo.

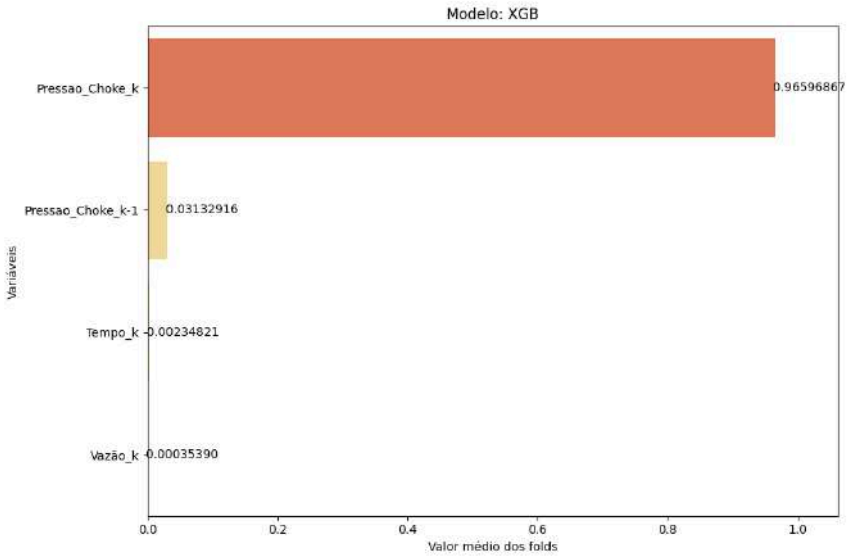

Figura *88* – Importância das variáveis com dados de poços reais, tratados com a função sig e com 2 dados passados. Fonte: A autora.

Os resultados das previsões e realidade do modelo são apresentados nas Figuras 89 a 91, indicando que o modelo não consegue prever a operação de PMCD de forma apropriada.

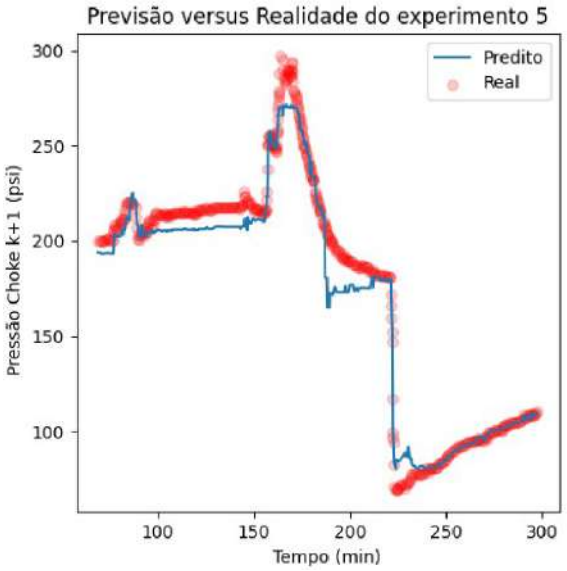

Figura *89* – Previsão versus realidade com dados de poços reais, tratados com a função sig e com 2 dados passados, para o experimento 5. Fonte: A autora.

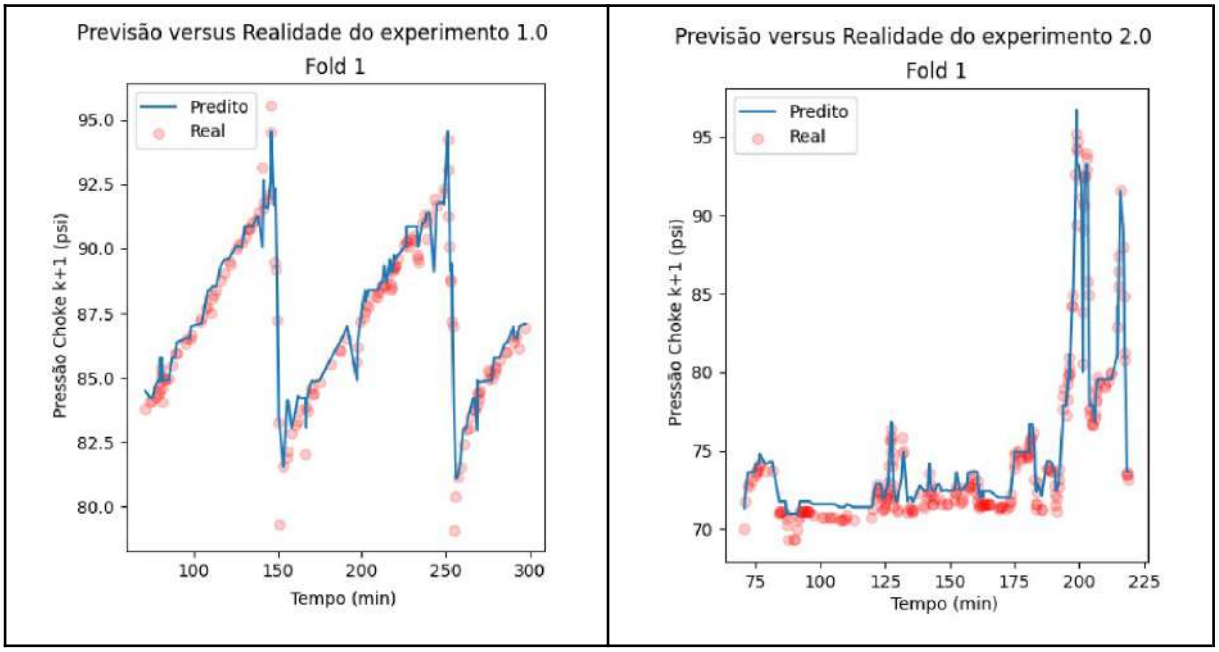

Figura *90* – Previsão versus realidade com dados de poços reais, tratados com a função sig e com 2 dados passados, para os experimentos 1 e 2. Fonte: A autora.

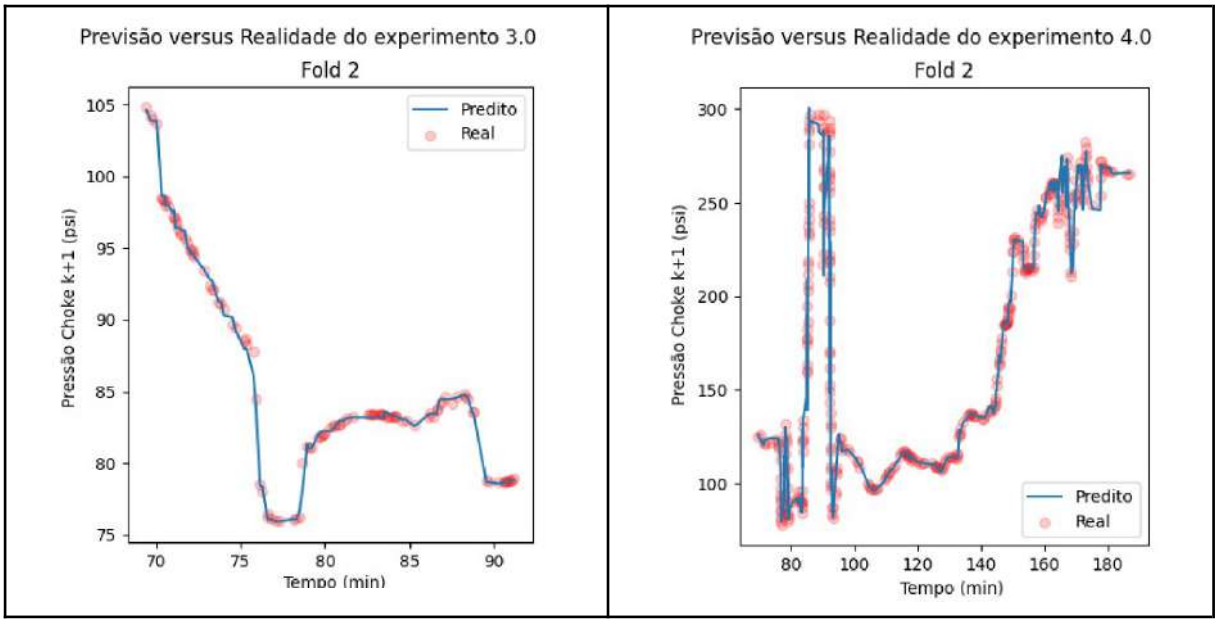

Figura *91* – Previsão versus realidade com dados de poços reais, tratados com a função sig e com 2 dados passados, para os experimentos 3 e 4. Fonte: A autora.

Com relação à arquitetura do modelo, para a configuração utilizando a escala de 0 a 4, junto com a função sig e com 2 dados passados, o modelo treinou 815 árvores. Apesar das métricas de avaliação apresentarem-se superiores, os resultados evidenciam que o uso de 2 dados passados não foi suficiente para a geração de modelos com capacidade preditiva.

4.2.3 Dados de poços reais tratados com a função sig e com 8 dados passados

Com 8 dados passados, todas as variáveis são deslocadas, conforme mostrado na Figura 92. A vazão possui *outliers* que foram tratados com a substituição pelas suas médias.

	Valores_Únicos	Tipo_de_Dado	mean	std	min	25%	50%	75%	max
Pressao_Choke_k-1 (psi)	2296	float64	0.68	0.17	0.5	0.55	0.59	0.90	0.98
Pressao_Choke_k-2 (psi)	2296	float64	0.68	0.17	0.5	0.55	0.59	0.90	0.98
Pressao_Choke_k-3 (psi)	2295	float64	0.68	0.17	0.5	0.55	0.59	0.90	0.98
Pressao_Choke_k-4 (psi)	2296	float64	0.68	0.17	0.5	0.55	0.59	0.90	0.98
Pressao_Choke_k-5 (psi)	2298	float64	0.68	0.17	0.5	0.55	0.59	0.90	0.98
Pressao_Choke_k-6 (psi)	2298	float64	0.68	0.17	0.5	0.55	0.59	0.90	0.98
Pressao_Choke_k-7 (psi)	2297	float64	0.68	0.17	0.5	0.55	0.59	0.90	0.98
Tempo_k	3072	float64	0.74	0.15	0.5	0.59	0.77	0.87	0.98
Vazão_k (m³/h)	465	float64	0.52	0.10	0.5	0.50	0.50	0.50	0.98
Pressao_Choke_k (psi)	2296	float64	0.68	0.17	0.5	0.55	0.59	0.90	0.98
Pressão_Choke_k+1 (psi)	2298	float64	420.27	546.90	0.0	89.66	181.63	402.51	2001.09
experimento	5	int64	3.03	1.32	1.0	2.00	3.00	4.00	5.00

Figura *92* – Resumo do *dataframe* com dados de poços reais, tratados com a função sig e com 8 dados passados. Fonte: A autora.

Antes dos dados serem treinados pelo XGBoost, é feito a otimização dos hiperparâmetros com o Optuna gerando os seguintes parâmetros: 'n_estimators': 696, 'learning_rate': 0.0257, 'max_depth': 5, 'min_child_weight': 8, 'subsample': 0.8748, 'colsample_bytree': 0.7845, 'gamma': 0.8569, 'reg_alpha': 0.4020, e 'reg_lambda': 0.7736, em apenas 3.57 segundos de execução.

Com os melhores parâmetros gerados pelo Optuna, os dados são treinados, validados e testados pelo XGBoost. As métricas de avaliação do modelo são apresentadas na Figura 93.

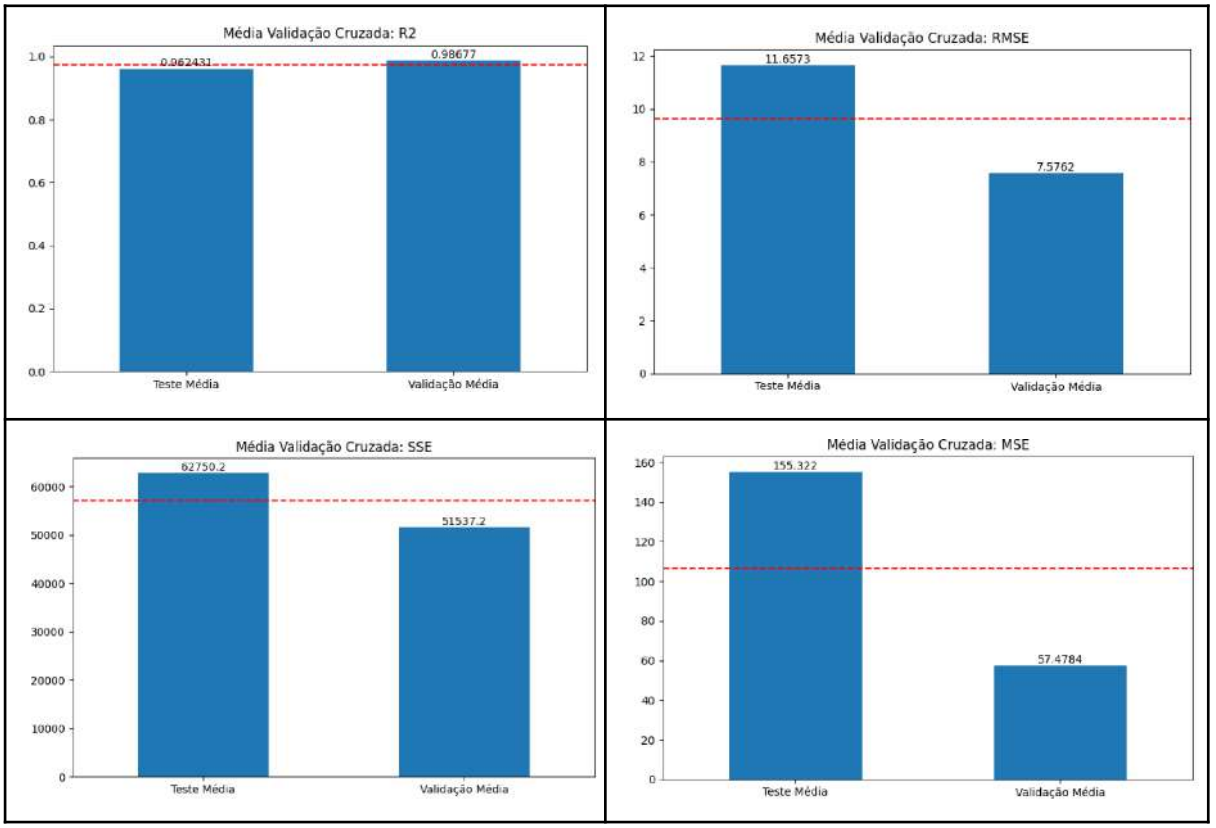

Figura *93* – Métricas de avaliação com dados de poços reais, tratados com a função sig e com 8 dados passados. Fonte: A autora.

São apresentados na Figura 94 os valores da densidade dos dados para a pressão na *choke*, sendo que quanto mais valores preditos de forma correta, mais será observada a presença da cor roxa nos gráficos.

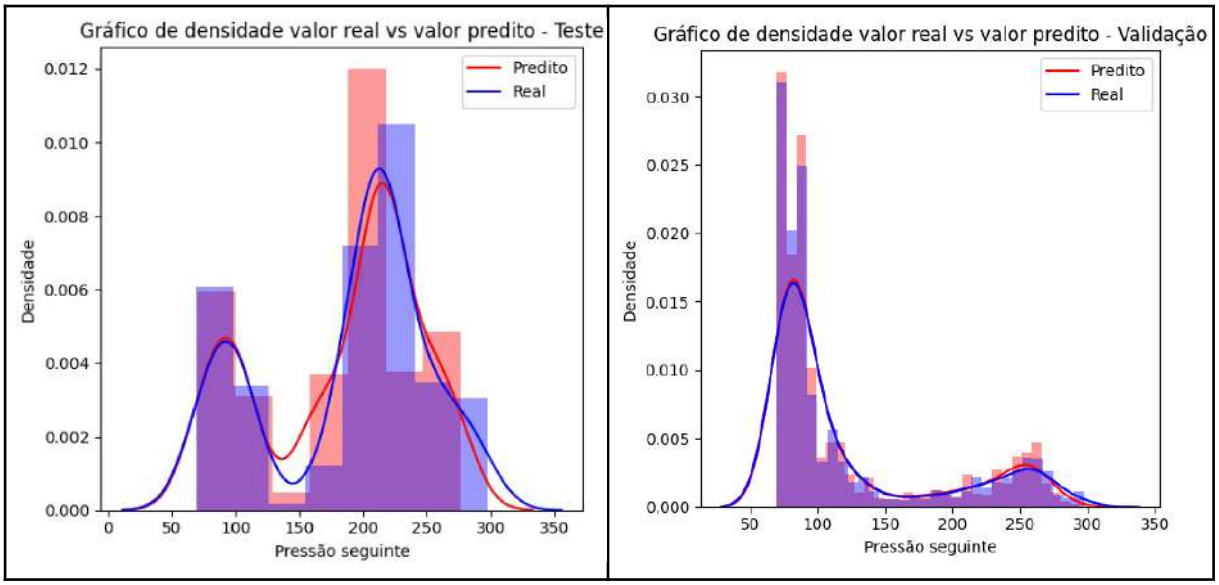

Figura *94* – Gráficos de densidade com dados de poços reais, tratados com a função sig e com 8 dados passados. Fonte: A autora.

São apresentados na Figura 95 os valores da distribuição dos resíduos no teste e na validação do modelo.

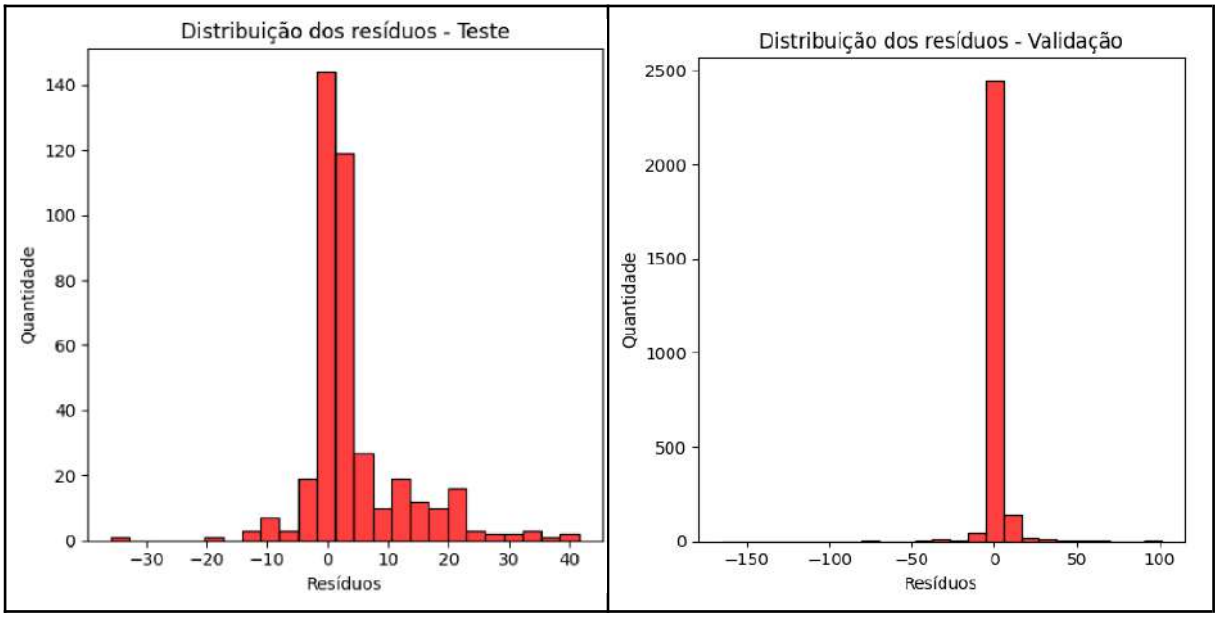

Figura *95* – Distribuição dos resíduos com dados de poços reais, tratados com a função sig e com 8 dados passados. Fonte: A autora.

A Figura 96 apresenta os pontos na curva de comparação de valor real e valor predito.

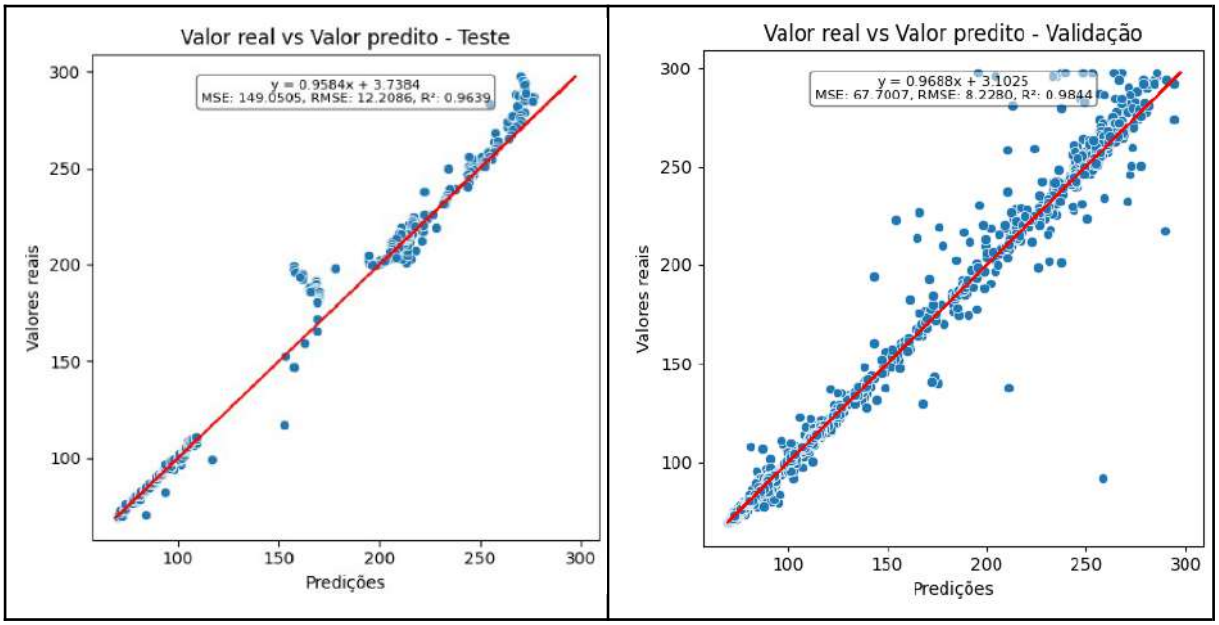

Figura *96* – Gráfico de evolução do modelo com dados de poços reais, tratados com a função sig e com 8 dados passados. Fonte: A autora.

O gráfico da função de perda é apresentado na Figuras 97.

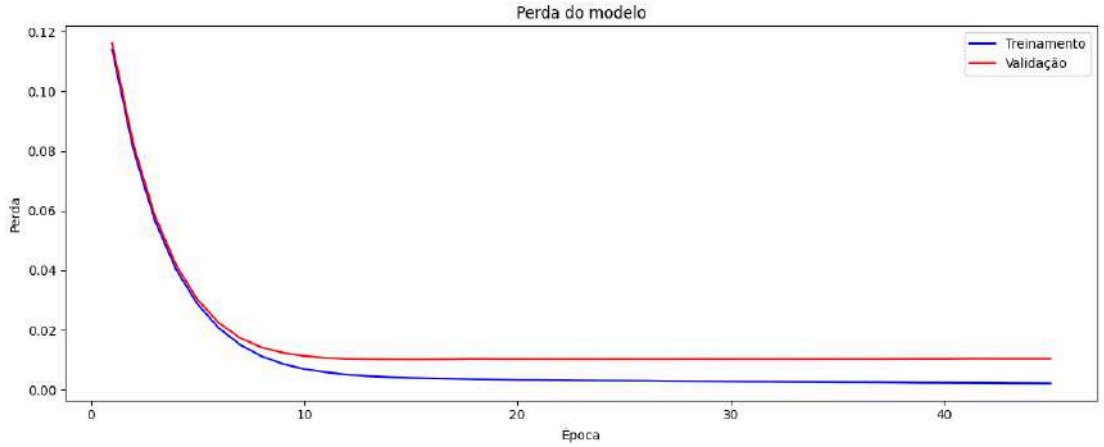

Figura *97* – Curva de perdas função sig e com 8 dados passados. Fonte: A autora.

O gráfico na Figura 98 representa a importância de cada variável para as previsões, seguindo o modelo do XGBoost treinado, sendo que a pressão da *choke* no passo atual, teve uma forte influência nas previsões, seguido da pressão *choke* no passo anterior.

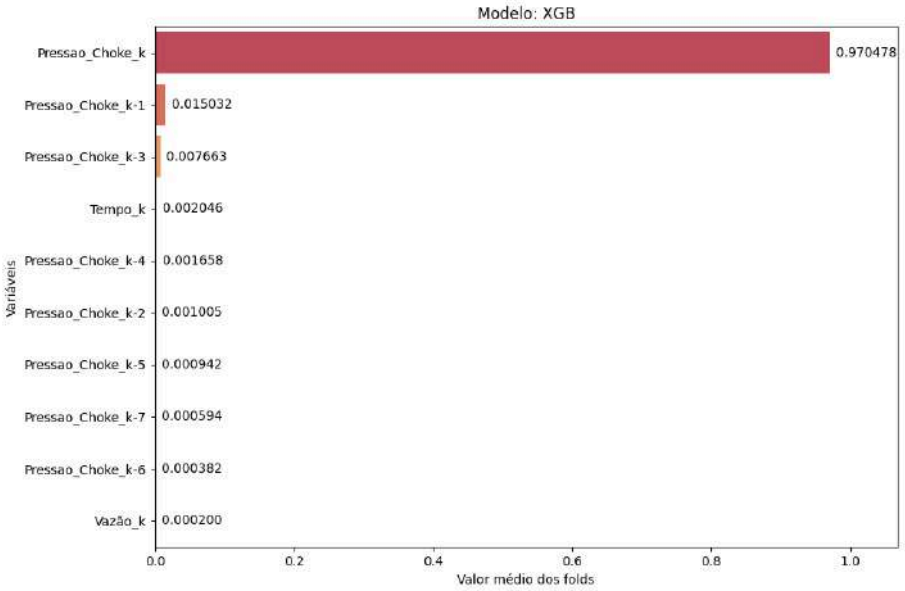

Figura *98* – Importância das variáveis com dados de poços reais, tratados com a função sig e com 8 dados passados. Fonte: A autora.

Os resultados das previsões e realidade do modelo são apresentados nas Figuras 99 a 101, indicando o quanto o modelo consegue prever a operação de PMCD.

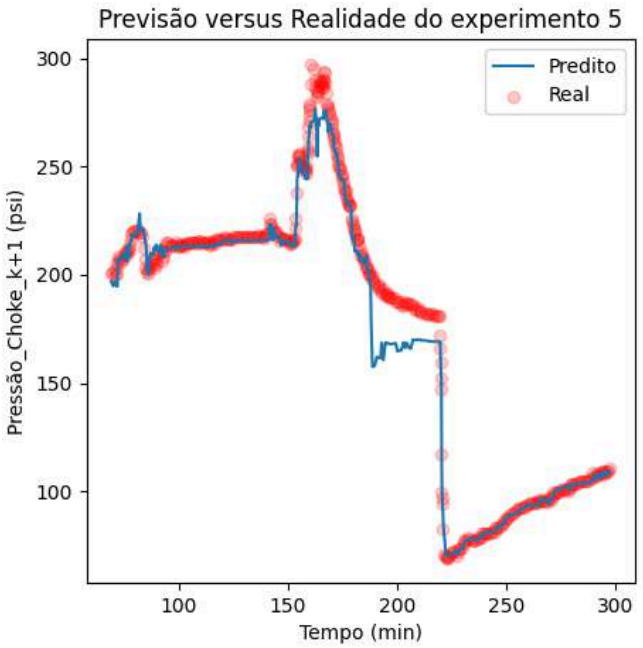

Figura *99* – Previsão versus realidade com dados de poços reais, tratados com a função sig e com 8 dados passados, para o experimento 5. Fonte: A autora.

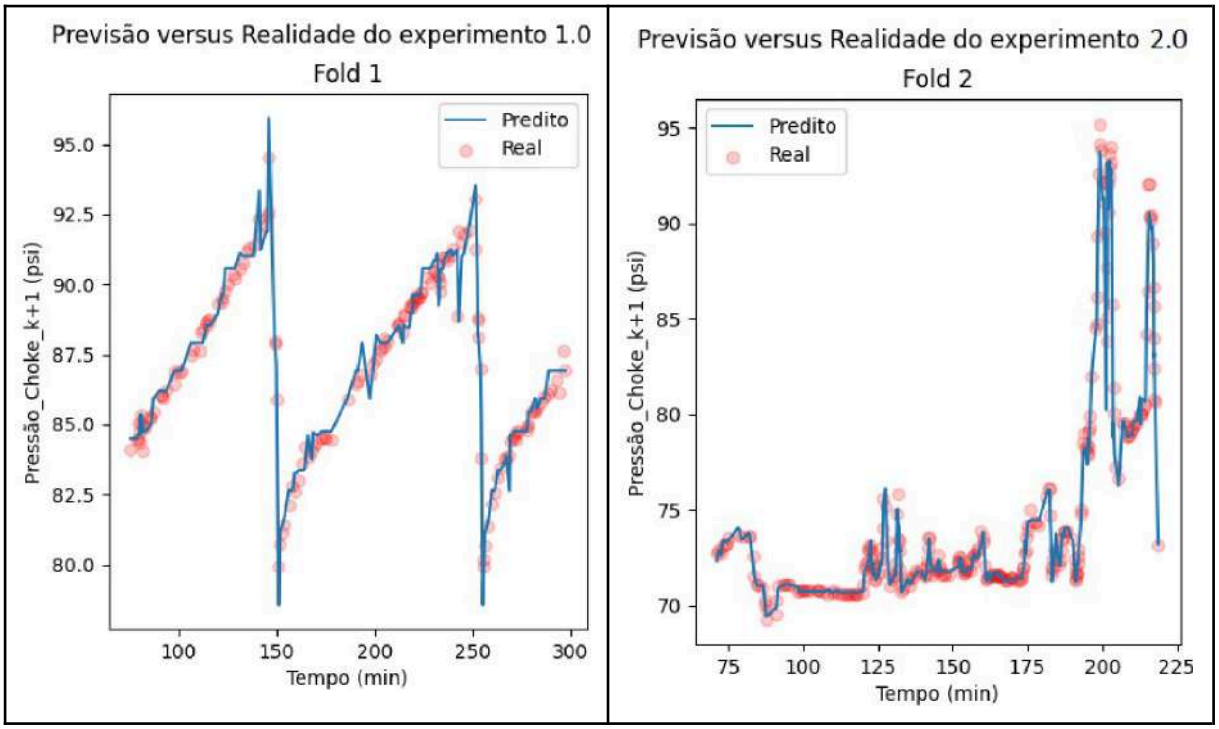

Figura *100* – Previsão versus realidade com dados de poços reais, tratados com a função sig e com 8 dados passados, para os experimentos 1 e 2. Fonte: A autora.

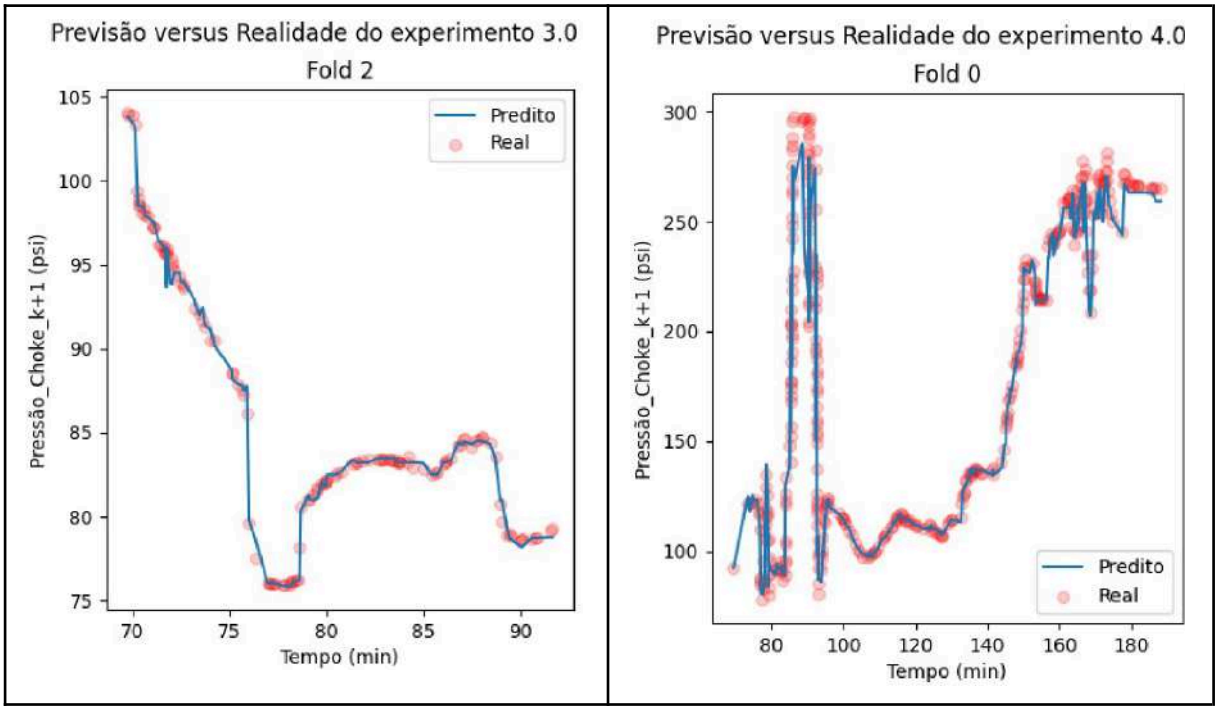

Figura *101* – Previsão versus realidade com dados de poços reais, tratados com a função sig e com 8 dados passados, para os experimentos 3 e 4 Fonte: A autora.

Com relação à arquitetura do modelo, foram treinadas 696 árvores. Todas as métricas de avaliação indicam que houve melhora de desempenho como uso de 8 dados passados.

4.2.4 Dados de poços reais tratados com a função sig e com 20 dados passados

Com 20 dados passados, todas as variáveis são deslocadas, conforme mostrado na Figura 102. A vazão possui *outiliers* que foram tratados com a substituição pelas suas médias.

	Valores_Únicos	Tipo_de_Dado	mean	std	min	25%	50%	75%	max
Pressao_Choke_k-1 (psi)	2260	float64	0.68	0.17	0.5	0.55	0.59	0.90	0.98
Pressao_Choke_k-2 (psi)	2259	float64	0.68	0.17	0.5	0.55	0.59	0.90	0.98
Pressao_Choke_k-3 (psi)	2258	float64	0.68	0.17	0.5	0.55	0.59	0.90	0.98
Pressao_Choke_k-4 (psi)	2260	float64	0.68	0.17	0.5	0.55	0.59	0.90	0.98
Pressao_Choke_k-5 (psi)	2260	float64	0.68	0.17	0.5	0.55	0.60	0.90	0.98
Pressao_Choke_k-6 (psi)	2262	float64	0.68	0.17	0.5	0.55	0.60	0.90	0.98
Pressao_Choke_k-7 (psi)	2263	float64	0.68	0.17	0.5	0.55	0.60	0.90	0.98
Pressao_Choke_k-8 (psi)	2264	float64	0.68	0.17	0.5	0.55	0.60	0.90	0.98
Pressao_Choke_k-9 (psi)	2266	float64	0.68	0.17	0.5	0.55	0.60	0.90	0.98
Pressao_Choke_k-10 (psi)	2268	float64	0.68	0.17	0.5	0.55	0.60	0.90	0.98
Pressao_Choke_k-11 (psi)	2267	float64	0.68	0.17	0.5	0.55	0.60	0.90	0.98
Pressao_Choke_k-12 (psi)	2267	float64	0.68	0.17	0.5	0.55	0.60	0.90	0.98
Pressao_Choke_k-13 (psi)	2268	float64	0.69	0.17	0.5	0.55	0.60	0.90	0.98
Pressao_Choke_k-14 (psi)	2268	float64	0.69	0.17	0.5	0.55	0.60	0.90	0.98
Pressao_Choke_k-15 (psi)	2269	float64	0.69	0.17	0.5	0.55	0.60	0.90	0.98
Pressao_Choke_k-16 (psi)	2270	float64	0.69	0.17	0.5	0.55	0.60	0.90	0.98
Pressao_Choke_k-17 (psi)	2272	float64	0.69	0.17	0.5	0.55	0.60	0.90	0.98
Pressao_Choke_k-18 (psi)	2273	float64	0.69	0.17	0.5	0.55	0.60	0.90	0.98
Pressao_Choke_k-19 (psi)	2272	float64	0.69	0.17	0.5	0.55	0.60	0.90	0.98
Tempo_k	3016	float64	0.75	0.15	0.5	0.59	0.78	0.87	0.98
Vazão_k (m³/h)	443	float64	0.52	0.10	0.5	0.50	0.50	0.50	0.98
Pressao_Choke_k (psi)	2260	float64	0.68	0.17	0.5	0.55	0.59	0.90	0.98
Pressão_Choke_k+1 (psi)	2253	float64	423.47	550.98	0.0	89.47	181.27	403.80	2001.09
experimento	5	int64	4.07	1.63	1.0	3.00	4.00	6.00	6.00

Figura *102* – Resumo do *dataframe* com dados de poços reais, tratados com a função sig e com 20 dados passados. Fonte: A autora.

Antes dos dados serem treinados pelo XGBoost, é feito a otimização dos hiperparâmetros com o Optuna gerando os seguintes parâmetros: 'n_estimators': 650, 'learning_rate': 0.0990, 'max_depth': 8, 'min_child_weight': 9, 'subsample': 0.5208, 'colsample_bytree': 0.7962, 'gamma': 0.4211, 'reg_alpha': 0.1418, 'reg_lambda': 0.4173, em apenas 2.5 segundos de execução, utilizando o mesmo ambiente de execução dos modelos anteriores.

Com os melhores parâmetros gerados pelo Optuna, os dados são treinados, validados e testados pelo XGBoost. As métricas de avaliação do modelo são apresentadas na Figura 103.

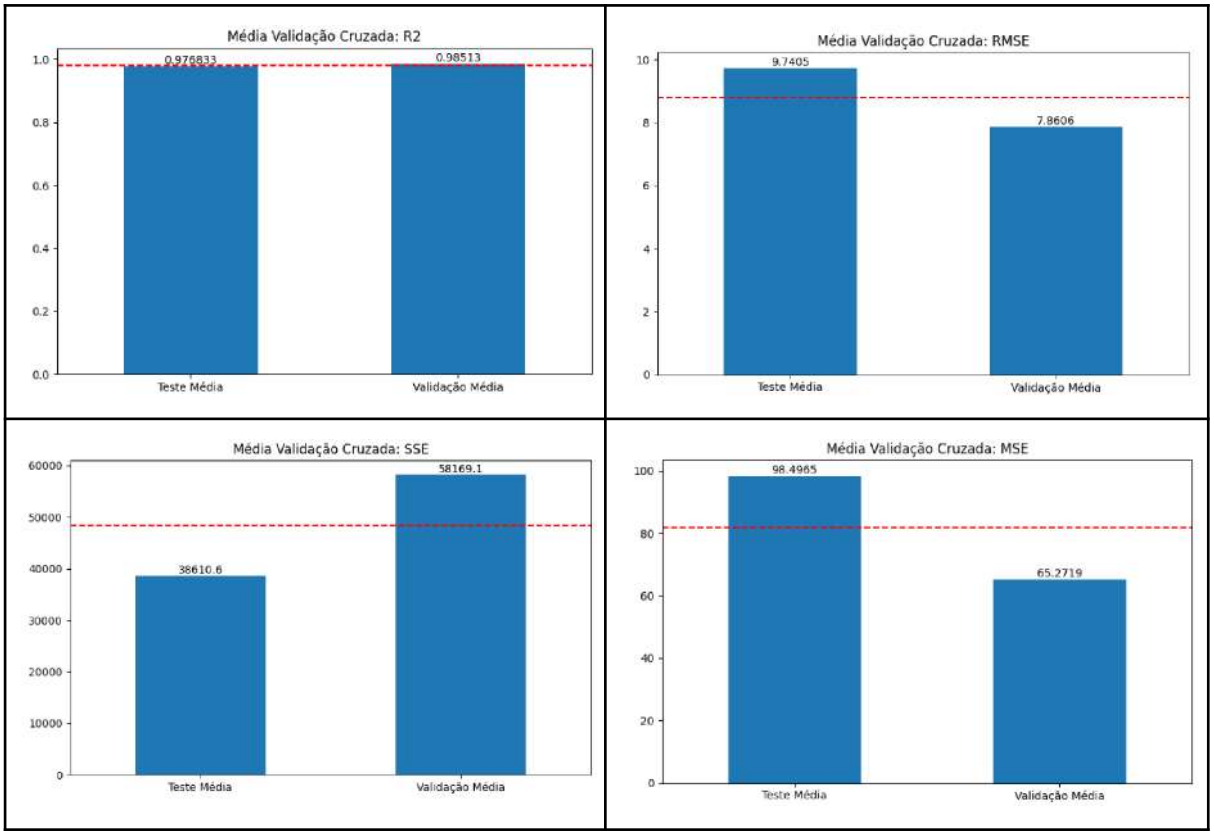

Figura *103* – Métricas de avaliação com dados de poços reais, tratados com a função sig e com 20 dados passados. Fonte: A autora.

São apresentados na Figura 104 os valores da densidade dos dados para a pressão na *choke*, sendo que quanto mais valores preditos de forma correta, mais será observada a presença da cor roxa nos gráficos.

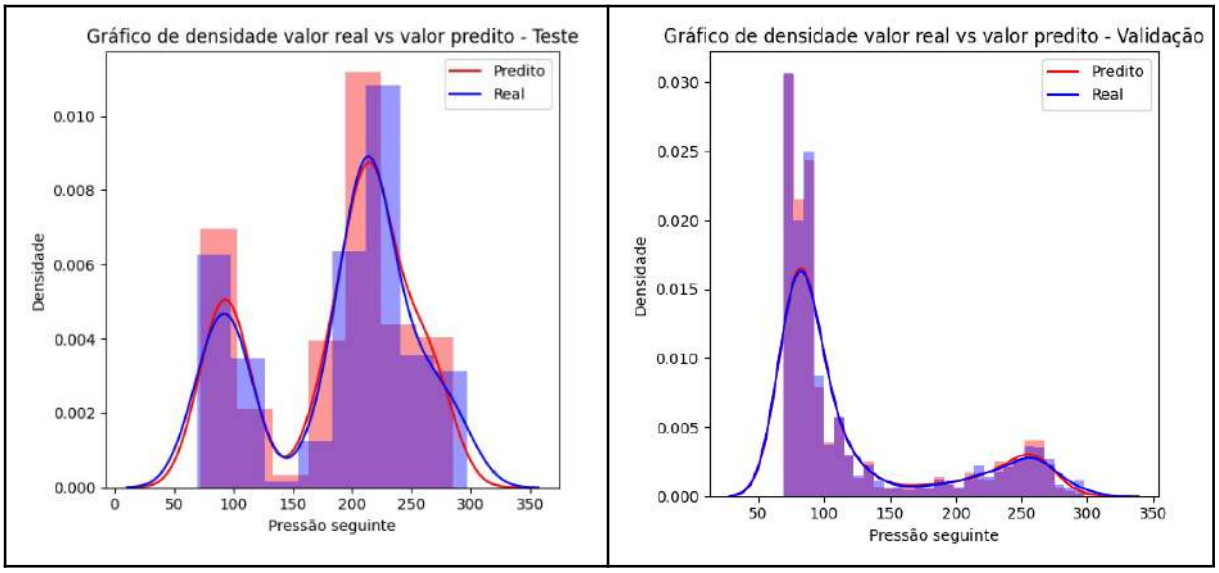

Figura *104* – Gráficos de densidade com dados de poços reais, tratados com a função sig e com 20 dados passados. Fonte: A autora.

São apresentados na Figura 105 os valores da distribuição dos resíduos no teste e na validação do modelo.

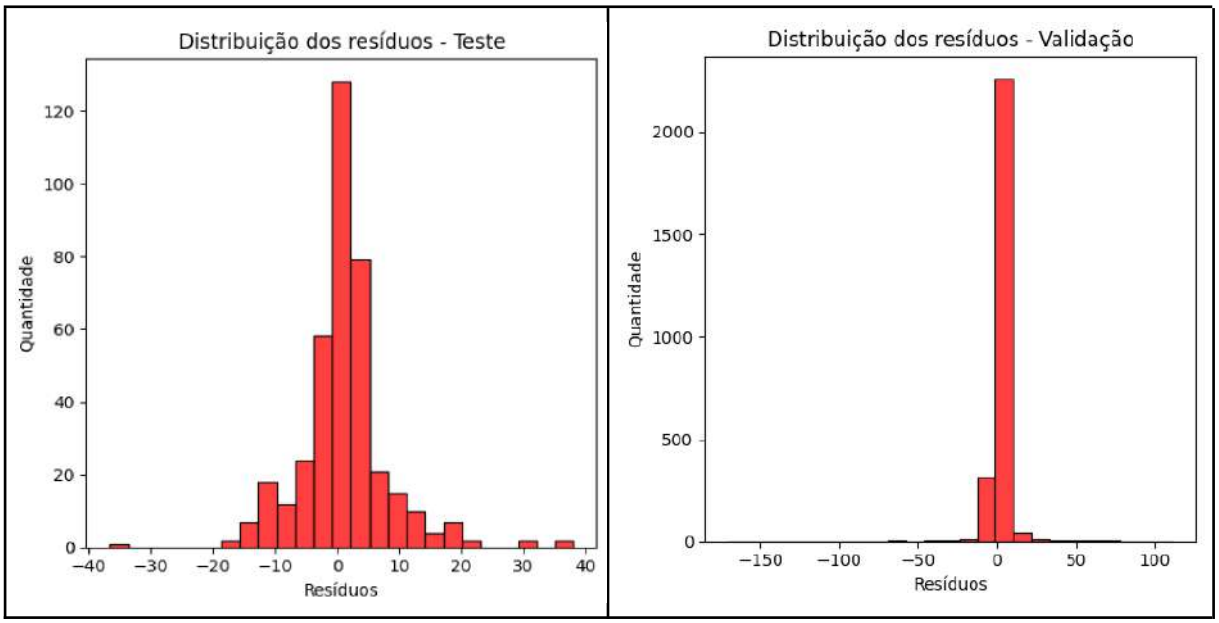

Figura *105* – Distribuição dos resíduos com dados de poços reais, tratados com a função sig e com 20 dados passados. Fonte: A autora.

A Figura 106 apresenta os pontos na curva de comparação de valor real e valor predito.

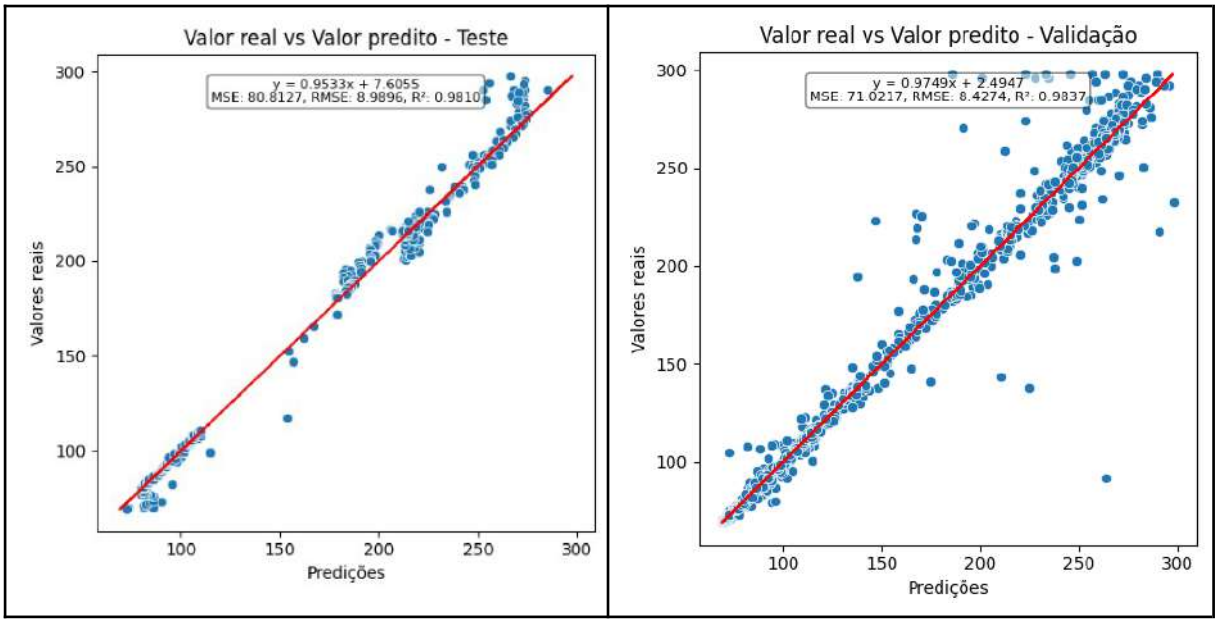

Figura *106* – Gráfico de evolução do modelo com dados de poços reais, tratados com a função sig e com 20 dados passados. Fonte: A autora.

O gráfico da função de perda é apresentado na Figuras 107.

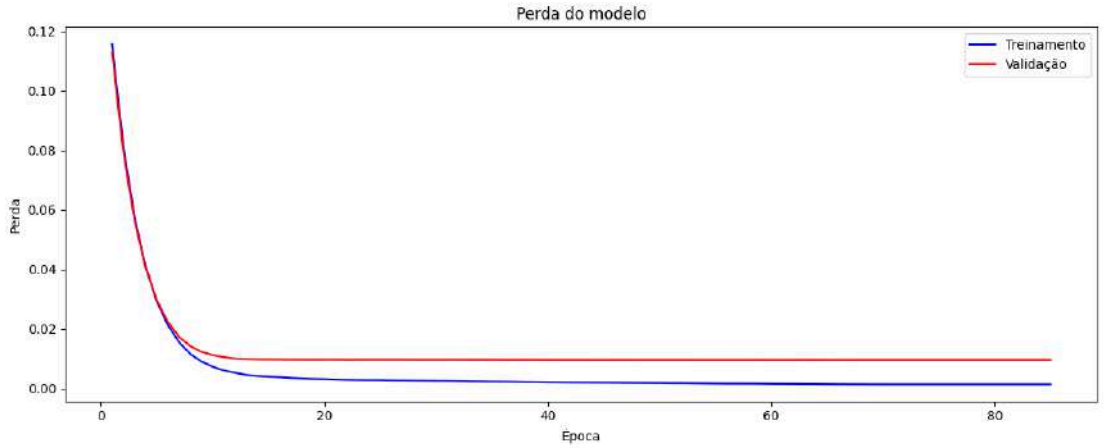

Figura *107* – Curva de perdas função sig e com 20 dados passados. Fonte: A autora.

O gráfico na Figura 108 representa a importância de cada variável para as previsões, seguindo o modelo do XGBoost treinado, sendo que a pressão da *choke* no passo atual sendo a mais importante.

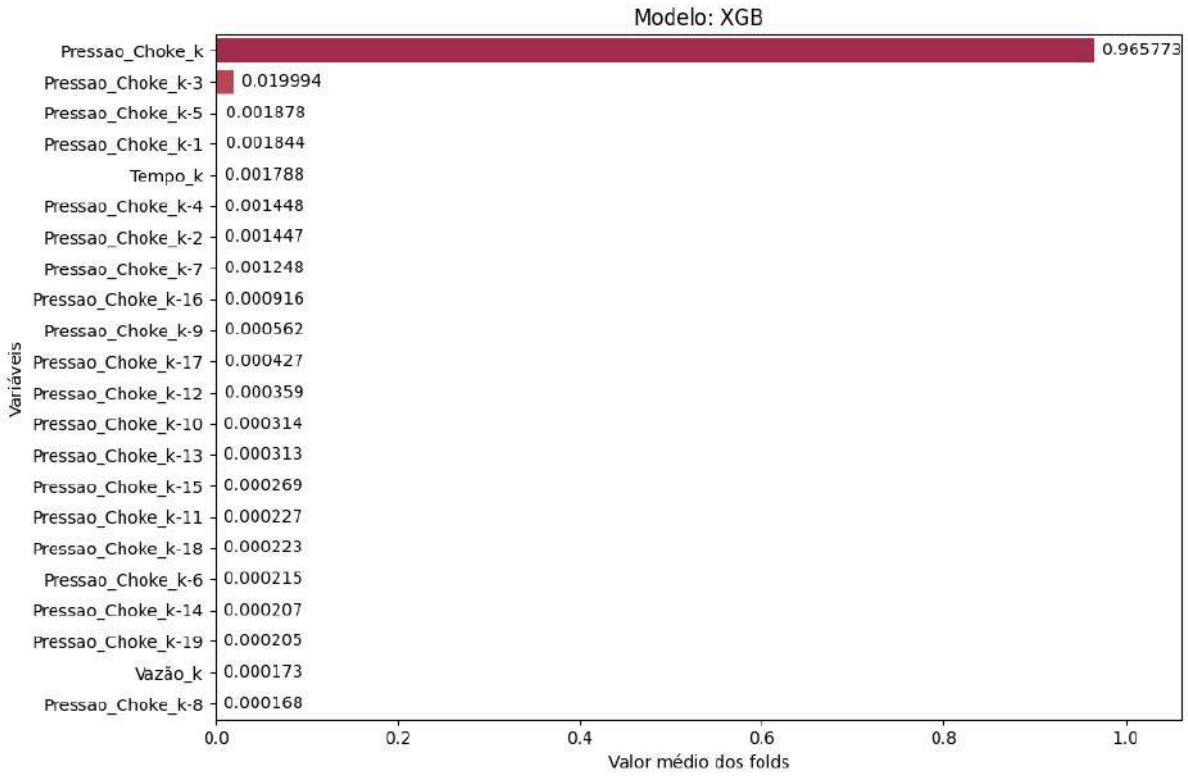

Figura *108* – Importância das variáveis com dados de poços reais, tratados com a função sig e com 20 dados passados. Fonte: A autora.

Os resultados das previsões e realidade do modelo são apresentados nas Figuras 109 a 111, indicando o quanto o modelo consegue prever a operação de PMCD.

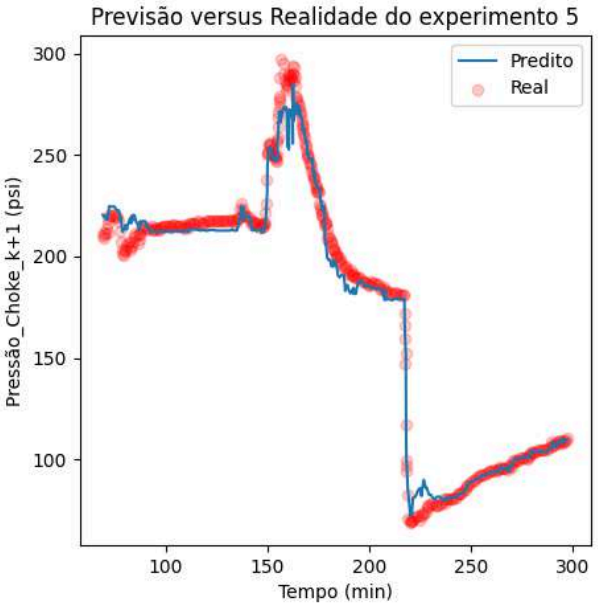

Figura *109* – Previsão versus realidade com dados de poços reais, tratados com a função sig e com 20 dados passados, para o experimento 5. Fonte: A autora.

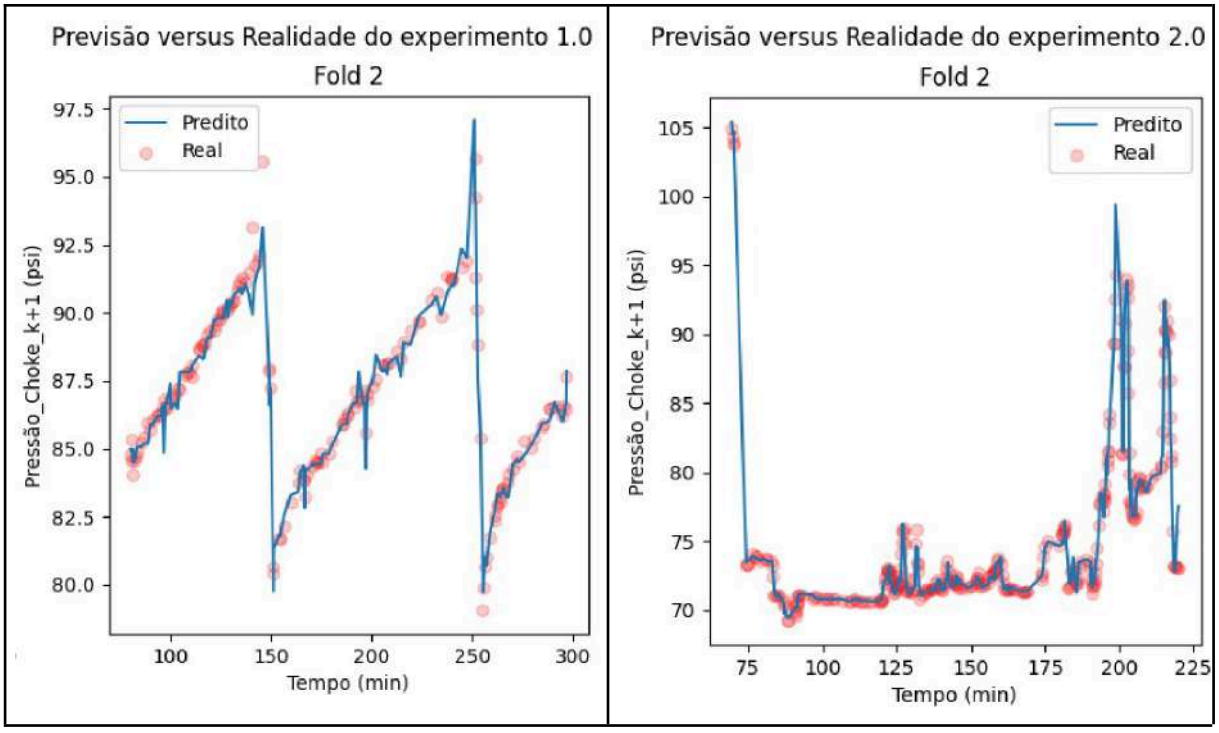

Figura *110* – Previsão versus realidade com dados de poços reais, tratados com a função sig e com 20 dados passados, para os experimentos 1 e 2. Fonte: A autora.

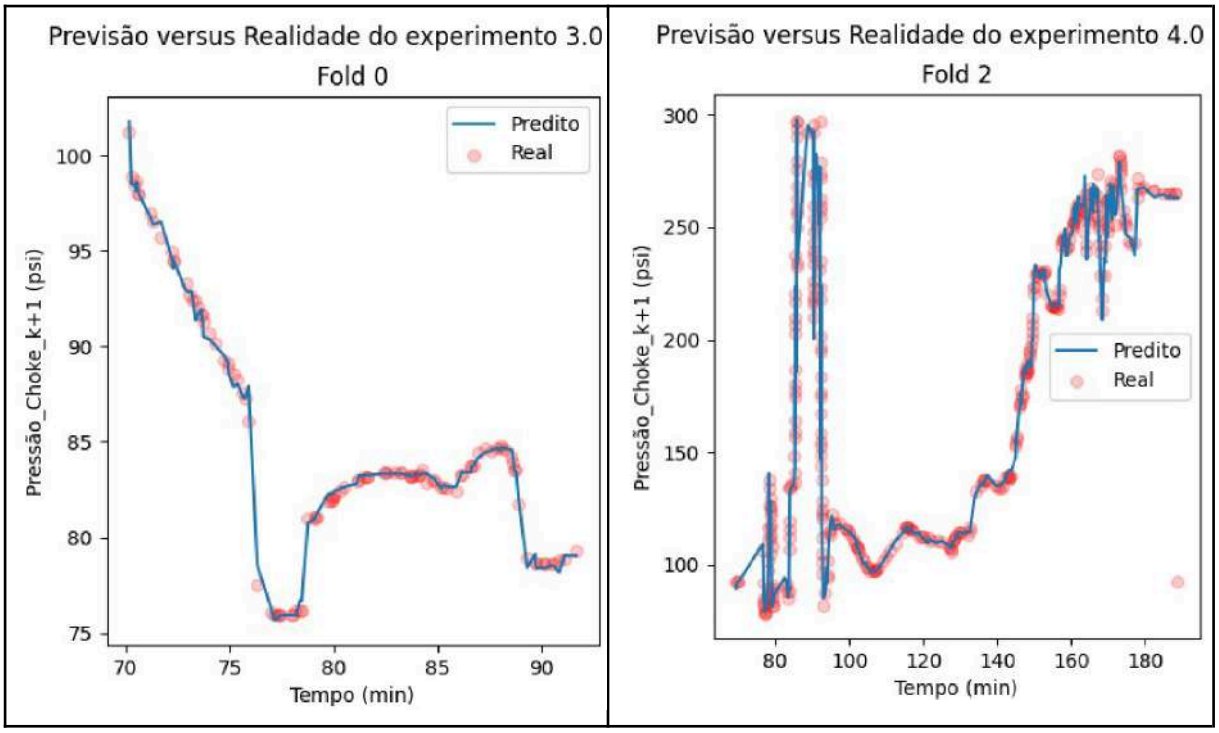

Figura *111* – Previsão versus realidade com dados de poços reais, tratados com a função sig e com 20 dados passados, para os experimentos 3 e 4 Fonte: A autora.

Com relação à arquitetura do modelo, foram treinadas 650 árvores.

A heterogeneidade dos dados referentes a poços reais, comprovada pelo valor do desvio padrão, e a quantidade escassa de dados, conduziram à necessidade de construção de um modelo com mais informações defasadas no tempo.

Para dados referentes a poços reais (Jayah (2013), Zein (2017) e Wattanasuwankorn (2014)), a Tabela 6 apresenta os tempos computacionais requeridos para a construção dos modelos baseados em *machine learning*.

Tabela 6 – Tempos computacionais dos modelos com dados de poços reais com sig. Fonte: A autora.

Modelos com dados de poços reais Sig (0, 4)	Tempo Total Optuna	Tempo Total XGBoost	Tempo Total Optuna e XGBoost
Sem dados passados	1.67 s	3.27 s	4.94 s
Com 2 dados passados	2.49 s	3.78 s	6.27 s
Com 8 dados passados	3.57 s	3.5 s	7.07 s
Com 20 dados passados	3.26 s	6.26 s	9.52 s

Os Anexos H ao K apresentam os resultados para o tratamento dos dados com a função log.

5 CONCLUSÕES E SUGESTÕES PARA TRABALHOS FUTUROS

O desenvolvimento dos modelos matemáticos foi realizado em Python, utilizando a biblioteca XGBoost (Chen & Guestrin, 2016) e o otimizador de hiperparâmetros Optuna (Akiba *et al.*, 2019). A metodologia abordou vários modelos de regressão para operação de perfuração usando a técnica PMCD com dados de poços reais de Jayah *et al.* (2013), Zein *et al.* (2017) e Wattanasuwankorn *et al.* (2014), e com dados experimentais de Carvalho (2018).

O código e as funções foram estruturados de forma semelhante, permitindo uma compreensão unificada do fluxo dos processos. A metodologia demonstrou eficácia e flexibilidade, sendo adaptável para diferentes cenários e tipos de dados, proporcionando resultados promissores em todos os modelos propostos. Tanto o XGBoost como o Optuna se mostraram algoritmos recentes com alto potencial de uso em cenários de perfuração de poços de petróleo. A seguir, serão detalhadas as conclusões quanto a cada modelo:

5.1 O PMCD com dados experimentais

O modelo XGBoost desenvolvido possui natureza preditiva adequada para estimar a operação de PMCD. A utilização de técnicas com o uso de dados passados para incorporar informação transiente, a transformação com as funções sig e log, a padronização da escala, o tratamento de campos nulos e *outliers* asseguram a qualidade da previsão da pressão anular de fundo.

O processo de validação cruzada e o ajuste de hiperparâmetros com o Optuna foram cruciais para otimizar o modelo XGBoost. Os resultados revelaram um desempenho notável, para os modelos com 8 dados passados, com os valores de R^2 acima de 0,9 na validação e no teste, tanto para o modelo com transformação dos dados utilizando a função sig quanto com uso da função log. A análise da distribuição real versus predita, bem como a construção de uma reta de ajuste, reforçou a precisão do modelo.

Comparando os resultados experimento a experimento com o trabalho de Carvalho (2018), observou-se consistência entre ambos os modelos, evidenciando a robustez do método proposto.

5.2 *O PMCD* com dados de poços reais

Nesta modelagem, o foco é a previsão da operação de *bullheading* utilizando o XGBoost a partir de dados de poços reais de Jayah *et al.* (2013), Zein *et al.* (2017) e Wattanasuwankorn *et al.* (2014), conjunto de dados com grande heterogeneidade. A utilização de técnicas como a aplicação de dados passados demonstraram um melhor desempenho para os modelos com 20 dados passados. O processo de validação cruzada e o ajuste de hiperparâmetros com o Optuna foram utilizados para otimizar o modelo.

A análise da distribuição real versus predita, bem como a construção de uma reta de ajuste, reforçou a precisão do modelo. As métricas de avaliação comprovaram o ótimo desempenho para o modelo com 20 dados passados. A função de perda mostra que o modelo enfrenta desafios em generalizar, indicando possível benefício em se utilizar de mais dados para construção do modelo.

5.3 Sugestões para trabalhos futuros

Para utilização de qualquer modelo de aprendizado de máquina é necessária uma quantidade considerável de dados para treinamento, validação e teste. Com isso, recomenda-se o enriquecimento dos algoritmos para apresentados com mais informações do processo.

Para os modelos de regressão prevendo a operação de PMCD, com dados de poços reais seria interessante a inclusão de mais dados da operação.

Nos modelos de regressão prevendo a operação de PMCD a partir de dados experimentais podem ser realizados mais testes, empregando-se novos modelos empíricos para fins de comparação como o uso de redes neuronais, por exemplo.

Além disso, outros fenômenos usuais da perfuração de poços podem ser estudados com o uso de *machine learning*, como mudanças de vazão e perda de circulação, como iniciado pelo trabalho de Carvalho (2018). Posteriormente, estudos podem ser realizados para a criação de um modelo único e robusto o suficiente para prever todos esses diferentes fenômenos.

6 REFERÊNCIAS BIBLIOGRÁFICAS

ABBAS, A. K. et al. **Intelligent decisions to stop or mitigate lost circulation based on machine learning.** Energy, v. 183, p. 1104-1113, 2019.

AGOSTIN, R. et al. **Artificial Intelligence Strategy Minimizes Lost Circulation Non-Productive Time in Brazilian Deep-Water Pre-Salt**. In Offshore Technology Conference Brasil, OTC. 24 out. 2017.

AHMED, A. et al. **Prediction of lost circulation zones using support vector ma-chine and radial basis function**. In: International Petroleum Technology Conference, 2020. IPTC.

AKIBA, T. et al. **Optuna.** Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 25 jul. 2019.

AL-HAMEEDI, A. T.; ALKINANI, H. H.; DUNN-NORMAN, S.; FLORI, R. E.; HILGEDICK, S. A.; AMER, A. S.; ALSABA, M. T. **Using machine learning to predict lost circulation in the Rumaila field, Iraq.** Sociedade de Engenheiros de Petróleo (SPE), 2018.

ALKINANI, H. H., AL-HAMEEDI, A. T. T., DUNN-NORMAN, S., **Data–Driven Decision–Making for Lost Circulation Treatments: A Machine Learning Approach**, Energy and AI, 2020.

ALOUHALI, R., ALJUBRAN, M., GHARBI, S. and Al-YAMI, A., 2018, December. **Drilling through data: automated kick detection using data mining**. In SPE International Heavy Oil Conference and Exhibition. OnePetro.

CALÇADA, L. et al. **Evaluation of suspension flow and particulate materials for control of fluid losses in drilling operation.** Journal of Petroleum Science and Engineering, v. 131, p. 1-10, 2015. ISSN 0920-4105.

CARVALHO, M. A. D.; **Estudos experimentais, de simulação e de controle na perfuração de poços de petróleo utilizando a técnica pressurized mud cap drilling**, Dissertação de Mestrado, Programa de Pós-Graduação em Engenharia Química/ UFRRJ, Seropédica, 2018.

CHEN, T.; GUESTRIN, C. 2016. **XGBoost: A Scalable Tree Boosting System.** Disponível em: <https://dl.acm.org/doi/abs/10.1145/2939672.2939785>. Acesso em: 02 nov. 2024.

CHIEZA, C. P. **Diagnósticos de problemas operacionais durante a perfuração de poços de petróleo**, 2011. 151 p. Dissertação (Mestrado em Engenharia Mecânica). Curso de Pós-Graduação em Engenharia Mecânica. Centro Técnico Científico, Pontifícia Universidade Católica do Rio de Janeiro, RJ. Disponível em: <https://www.maxwell.vrac.puc-rio.br/19161/19161_1.PDF>; Acesso em: 30 set. 2024.

COSTA, F. M. **Implementação de unidade experimental para controle da pressão anular de fundo durante o processo de cimentação de poços de petróleo**. Dissertação de mestrado. Universidade Federal Rural do Rio de Janeiro. [S.l.]. 2016.

FJETLAND, A. K., ZHOU, J., ABEYRATHNA, D. AND GRADVAL, J. E., 2019, June, **Kick detection and influx size estimation during offshore drilling operations using deep learning**. In 2019 14th IEEE Conference on Industrial Electronics and Applications (ICIEA) (pp. 2321-2326). IEEE.

FRANCISCO, MARCELA LOBO; TEIXEIRA, JOSÉ PAULO. **Uma comparação entre os regimes de taxação sobre o petróleo: concessão e partilha.** Rio de Janeiro, 2011. 136p. Tese de Doutorado - Departamento de Engenharia Industrial, Pontifícia Universidade Católica do Rio de Janeiro.

FREITAS, M. G., **Controle da pressão anular de fundo durante a perfuração de poços de petróleo**, Dissertação de Mestrado, Programa de Pós-Graduação em Engenharia Química/ UFRRJ, Seropédica, 2013.

GÉRON, A. **Mãos à obra: Aprendizado de máquina com Scikit-Learn, Keras, and Tensor Flow,** 2° Edição, atualizado com Tensor Flow2: conceitos, ferramentas e técnicas para construção de sistemas inteligentes – RJ, Alta Books, 2021. ISBN: 978-85-5081-548-0.

GHAURI, A. A. **Use of the AUSMV scheme for simulation of gas migration, bullheading and Pressurized Mud Cap Drilling**. Master thesis. University of Stavanger. [S.l.]. 2014.

HELGELAND, L. R. **Drilling of deep-set carbonates using pressurized mud cap drilling**. Master Thesis. Norwegian University of Science and Technology. [S.l.]. 2014.

HUYEN, C. **Designing Machine Learning Systems**. O'Reilly Media, Inc., 2022. ISBN: 9781098107963.

JAYAH, M. N. *et al.* **Implementation of PMCD to explore carbonate reservoirs from semi-submersible rigs in Malaysia results in safe and economical drilling operations.** SPE/IADC Drilling Conference and Exhibition, 2013.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis.** Upper Saddle River, N.J.: Pearson Prentice Hall, 2007.

KAASA, G.. , STAMNES, O. N., IMSLAND, L., and O. M. AAMO. **Intelligent Estimation of Downhole Pressure Using a Simple Hydraulic Model.** Paper presented at the IADC/SPE Managed Pressure Drilling and Underbalanced Operations Conference & Exhibition, Denver, Colorado, USA, April 2011.

KAMYAB, M., SHADIZADEH, S.R., JAZAYERIRAD, H. and Dinarvand, N., 2010, July. **Early kick detection using real time data analysis with dynamic neural network: A case study in iranian oil fields.** In SPE (pp. SPE-136995).

LIU, Y., UPCHURCH, E.R., OZBAYOGLU, E.M., BALDINO, S., ZHENG, D., and WANG, J., 2023. **Design and Calculation of Process Parameters in Bullheading and Pressurized Mud Cap Drilling.** Paper presented at SPE/IADC International Drilling Conferenceand Exhibition, Stavanger, Norway. March 2023.

MARDANIRAD, S., WOOD, D. A., ZAKERI, H., 2021. **The application of deep learning algorithms to classify subsurface drilling lost circulation severity in large oil field datasets.** *SN Applied Sciences*, *3*(9), p.785.

MARTINS, A. L., VEGA, M. P., FERNANDES, L. D. WALDMANN A.T.A., GANDELMAN, R.A., 2019, PETROCONTROL, n° registro BR5120190014642 – INPI Instituto Nacional

NYGAARD, G., NAEVDAL, G., **Nonlinear model predictive control scheme for stabilizing annulus pressure during oil well drilling**. Journal of Process Control, 16, 719-732, 2006.

OLIVEIRA, G. F. M.; VIEIRA, F. R. B.; VEGA, M. P.; COELHO, D. A. F.; RIBEIRO, V. J. S.; **Estudo de controlador com compensação de tempo morto para regular a pressão anular de fundo** – VI Encontro Nacional de Hidráulica de Poços de Petróleo e Gás, 2015.

PATRÍCIO, R. V. **Estudos de controle na perfuração de poços de petróleo em presença de *Kick* de gás.** 2016. 170 f. Dissertação (Mestrado em Química) – Instituto de Tecnologia, Universidade Federal Rural do Rio de Janeiro, Seropédica – RJ, 2016.

RAMALHO, R. V., **Utilização de técnicas de *machine learning* para identificação de perdas de carga na perfuração de poços de petróleo**. Engenharia Química – Instituto de Tecnologia, Universidade Federal Rural do Rio de Janeiro, Seropédica-RJ, 2023.

RIBEIRO, V. de J. da S. **Identificação e controle em linha de processo de perfuração de poços de petróleo utilizando redes neuronais.** 2018. 136 f. Dissertação (Mestrado em Engenharia Química) – Instituto de Tecnologia, Universidade Federal Rural do Rio de Janeiro, Seropédica-RJ, 2018.

ROCHA, L. A. S.; AZEVEDO, C. T., **Projeto de poços de petróleo: Geopressões e Assentamento de Colunas de Revestimento,** 2° ed., Interciência, Rio de Janeiro, 2007. ISBN: 978-85-7193-177-0

ROMUALDO, L. B., de MORAES OLIVEIRA, G.F., RABELLO, G.L. *et al.* **Bullheading optimization study of the PMCD technique.** *Braz. J. Chem. Eng.* **38**, 747–761 (2021).

RUSSANO, E. **Controle de perda de circulação durante a perfuração de poços de petróleo.** 2014. 107 f. Dissertação (Mestrado em Engenharia Química) – Instituto de Tecnologia, Universidade Federal Rural do Rio de Janeiro, Seropédica, 2014.

SABAH, M. et al. **Hybrid machine learning algorithms to enhance lost-circulation prediction and management in the Marun oil field.** Journal of Petroleum Science and Engineering, v. 198, p. 108125–108125, 1 mar. 2021.

SCIKIT-LEARN. **User guide: contents — Scikit-Learn 0.22.1 documentation**. Disponível em: <https://scikit-learn.org/stable/user_guide.html>. Acesso em: 02 nov. 2024.

SHADE, G. ***Azure Machine Learning* - Parte 2**. Disponível em: https://gabrielschade.github.io/2018/01/17/azure-machine-learning-2.html. Acesso em: 29 set. 2024

SILVA, M. C. D. **Estudos de *kick* com solubilidade de gás em fluido de perfuração empregando a técnica pressurized mud cap drilling (PMCD) - Modelagem, controle e experimentos**. Dissertação de mestrado. Universidade Federal Rural do Rio de Janeiro. Seropédica, p. 123. 2019.

THOMAS, J. E., **Fundamentos de Engenharia de Petróleo**, 2ª edição, Interciência, Rio de Janeiro, 2001. ISBN: 857193099-6.

VIEIRA, F. R. B. **Controle da pressão anular de fundo durante a perfuração de poços de petróleo.** 2009. 106 f. Dissertação (Mestrado em Engenharia Química) – Instituto de Tecnologia, Universidade Federal Rural do Rio de Janeiro, Seropédica – RJ, 2009.

WATTANASUWANKORN, R., JIEMSAWAT, N., DUNLOP, T., and KANCHIAK. S., **First Achievement Using Water Shutoff Polymer in Monobore Well Completion, Gulf of Thailand.** Paper presented at the IADC/SPE Asia Pacific Drilling Technology Conference, Bangkok, Thailand, August 2014.

XGBoost Documentation — xgboost 1.5.0-dev documentation. Disponível em: <https://xgboost.readthedocs.io/en/latest/index.html> Acesso em: 02 nov. 2024.

XIE, H., SHANMUGAM, A.K. and ISSA, R.R., 2018. **Big data analysis for monitoring of kick formation in complex underwater drilling projects.** Journal of Computing in Civil Engineering, 32(5), p.04018030.

XU, Y.; GOODACRE, R. **On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning.** Journal of Analysis and Testing, v. 2, n. 3, p. 249–262, jul. 2018.

ZEIN, J., ISKANDAR, Y., DHARMA, D., and ALI A. **Eliminating Non Productive Time While Drilling in Southern Sumatra Field with Pressurized Mud Cap Drilling.** Paper presented at the SPE/IATMI Asia Pacific Oil & Gas Conference and Exhibition, Jakarta, Indonesia, October 2017.

7 ANEXOS

ANEXO A – RESULTADOS DOS DADOS EXPERIMENTAIS SEM DADOS PASSADOS, COM ESCALA DE -4 A 4 E TRANSFORMADOS COM SIG

As Figuras 112 a 116, representam o quanto o modelo consegue prever a operação de PMCD.

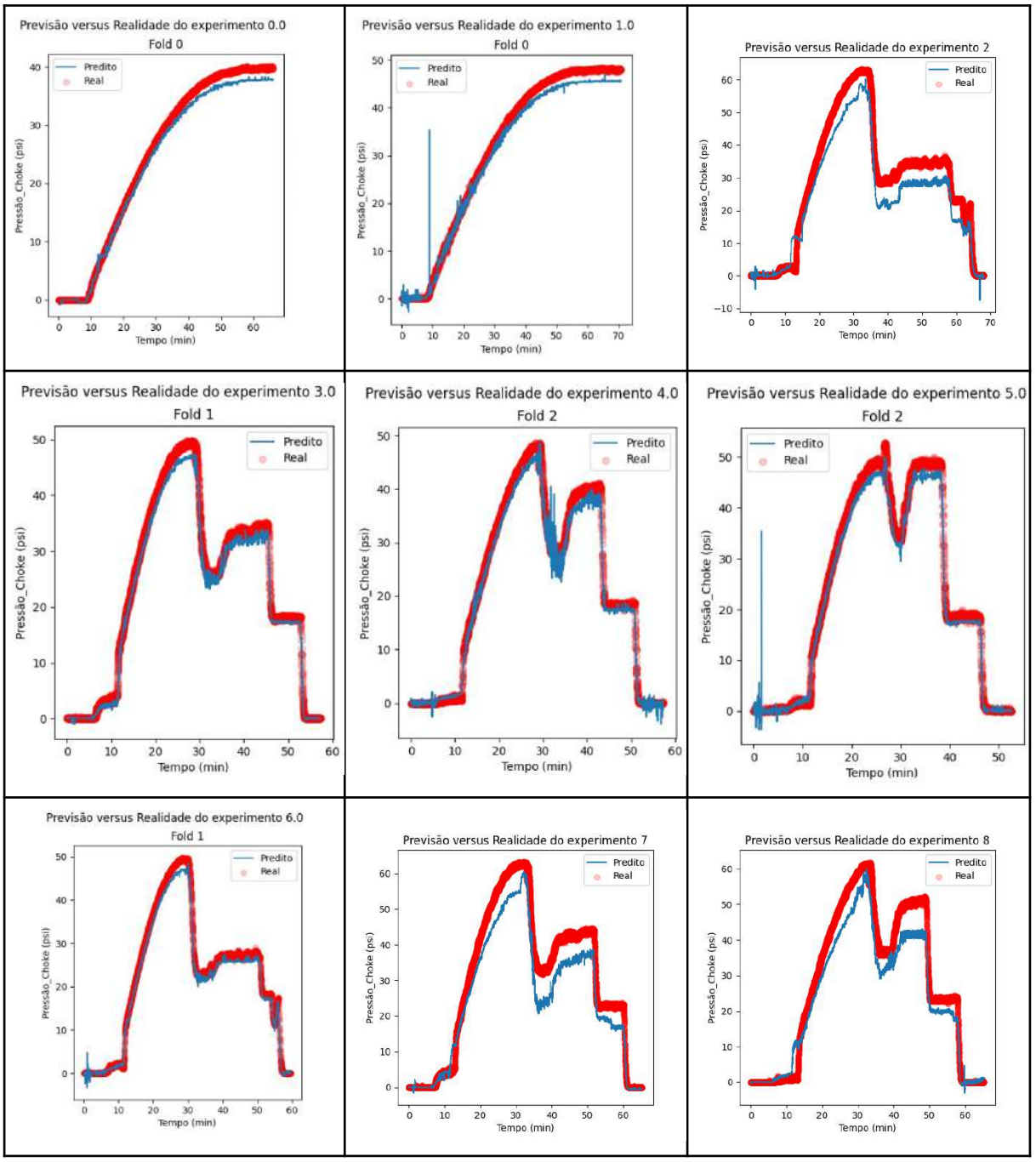

Figura *112* – Previsão e realidade dos experimentos de 0 ao 8 com sig e sem dados passados. Fonte: A autora.

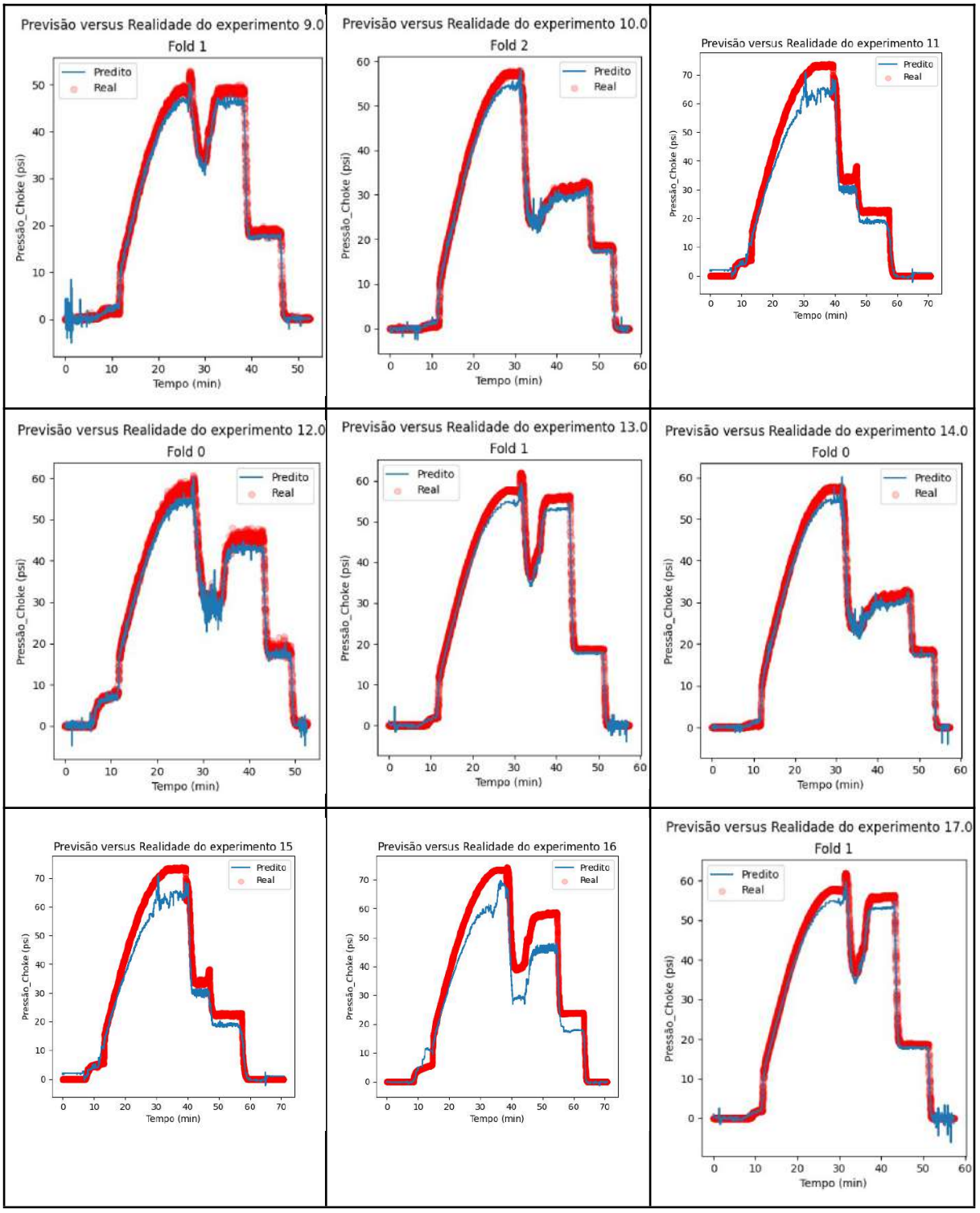

Figura *113* – Previsão e realidade dos experimentos de 9 ao 17 com sig e sem dados passados. Fonte: A autora.

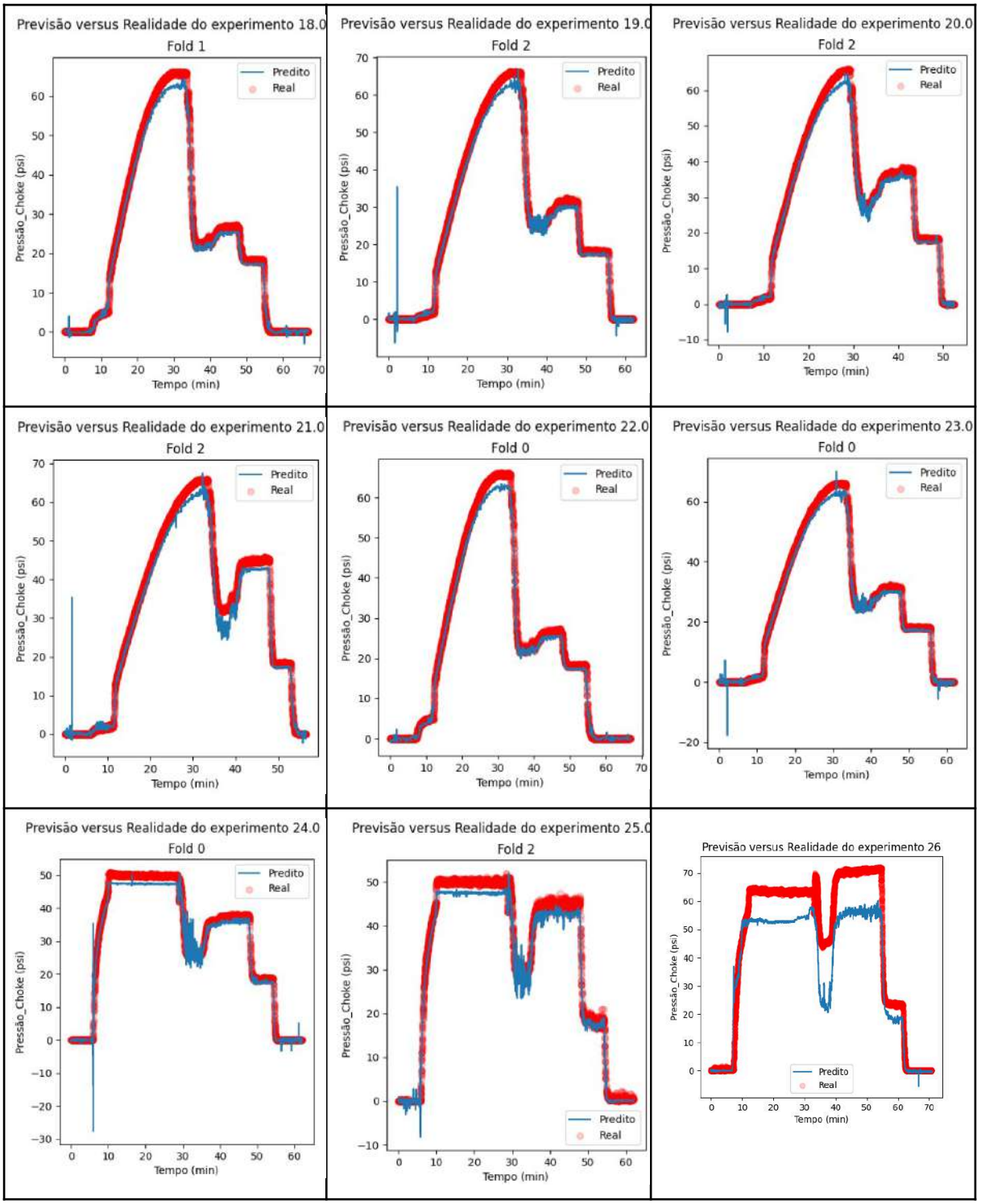

Figura *114* – Previsão e realidade dos experimentos de 18 ao 26 com sig e sem dados passados. Fonte: A autora.

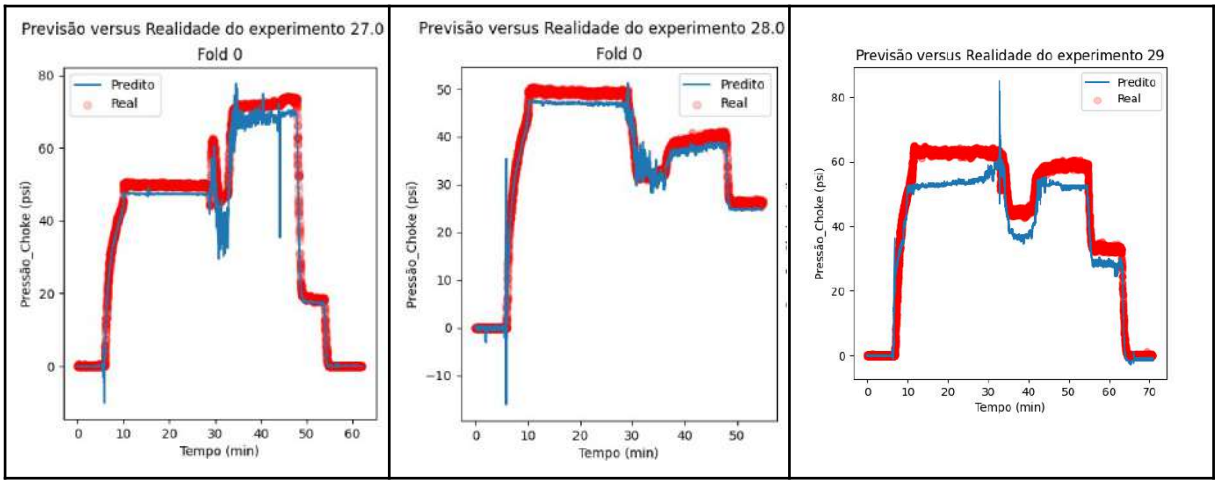

Figura *115* – Previsão e realidade dos experimentos de 27 ao 29 com sig e sem dados passados. Fonte: A autora.

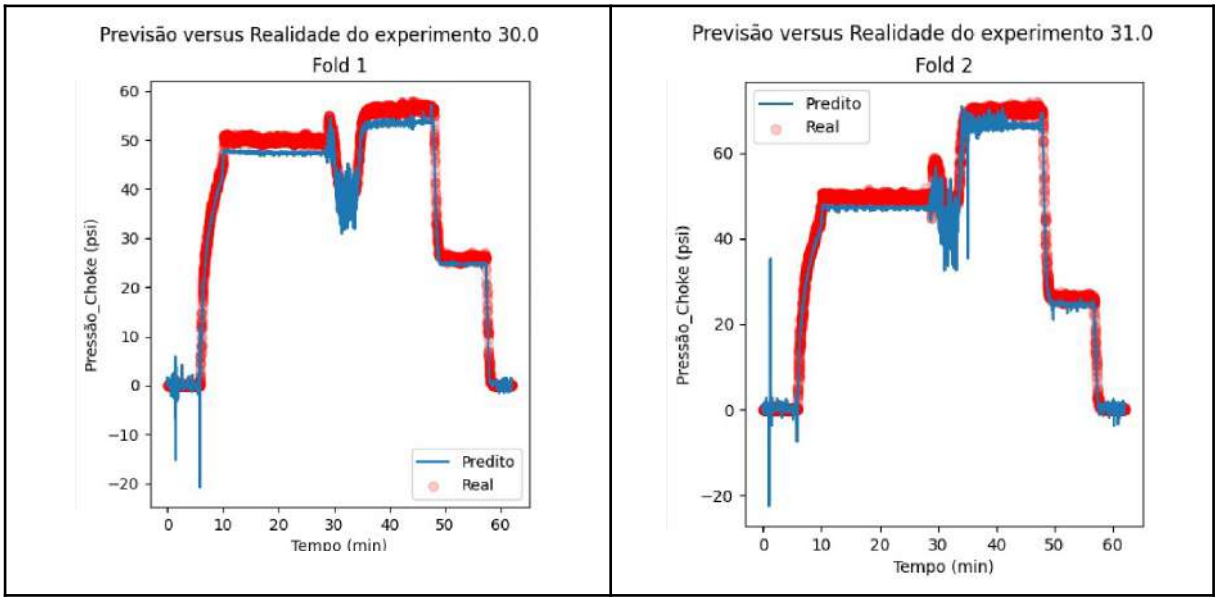

Figura *116* – Previsão e realidade dos experimentos de 30 e 31 com sig e sem dados passados. Fonte: A autora.

ANEXO B – RESULTADOS DOS DADOS EXPERIMENTAIS, COM 2 DADOS PASSADOS, COM ESCALA DE -4 A 4 E TRANSFORMADOS COM SIG

As Figuras 117 a 121, representam o quanto o modelo consegue prever a operação de PMCD.

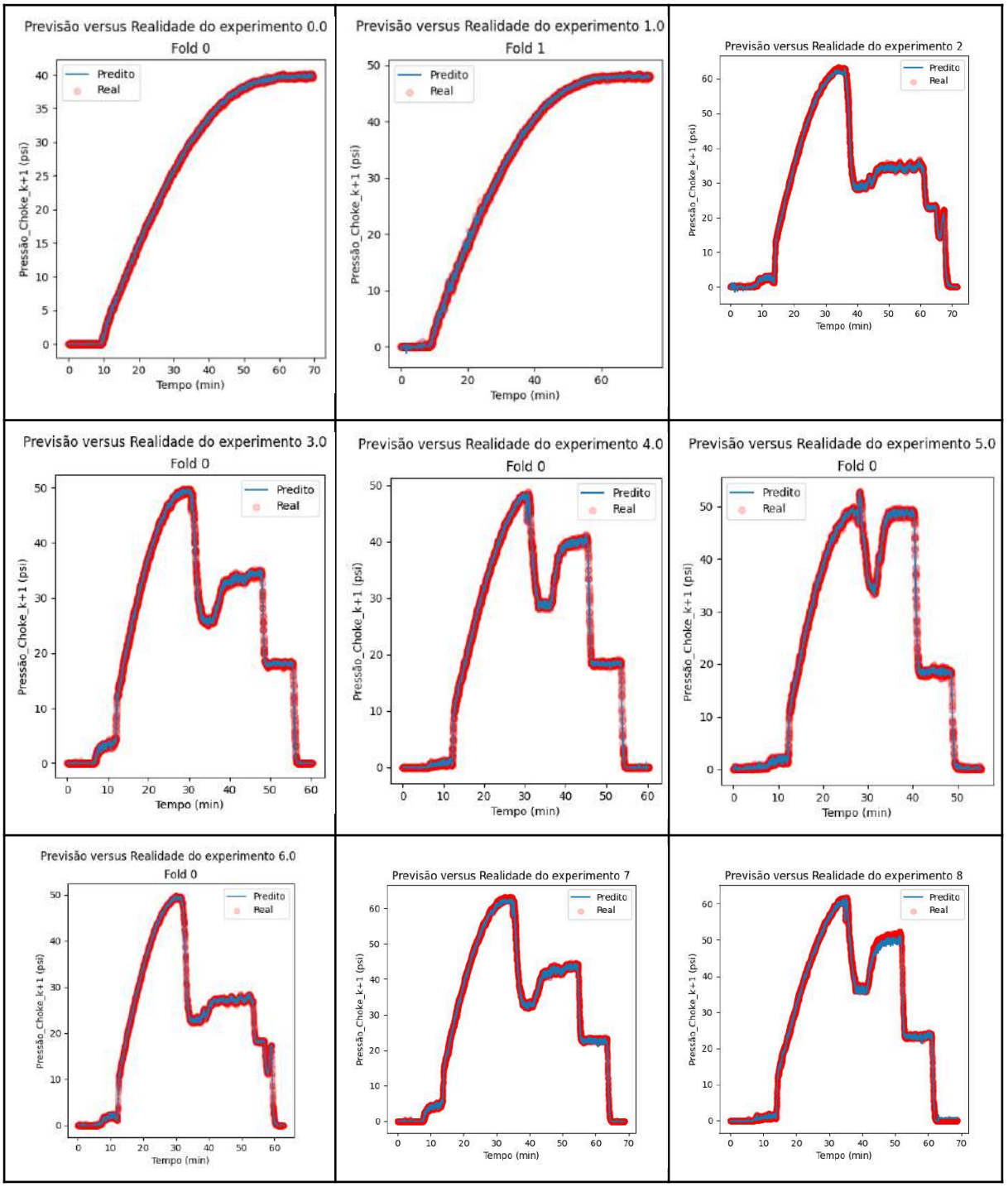

Figura *117* – Previsão e realidade dos experimentos de 0 ao 8 com sig e com 2 dados passados. Fonte: A autora.

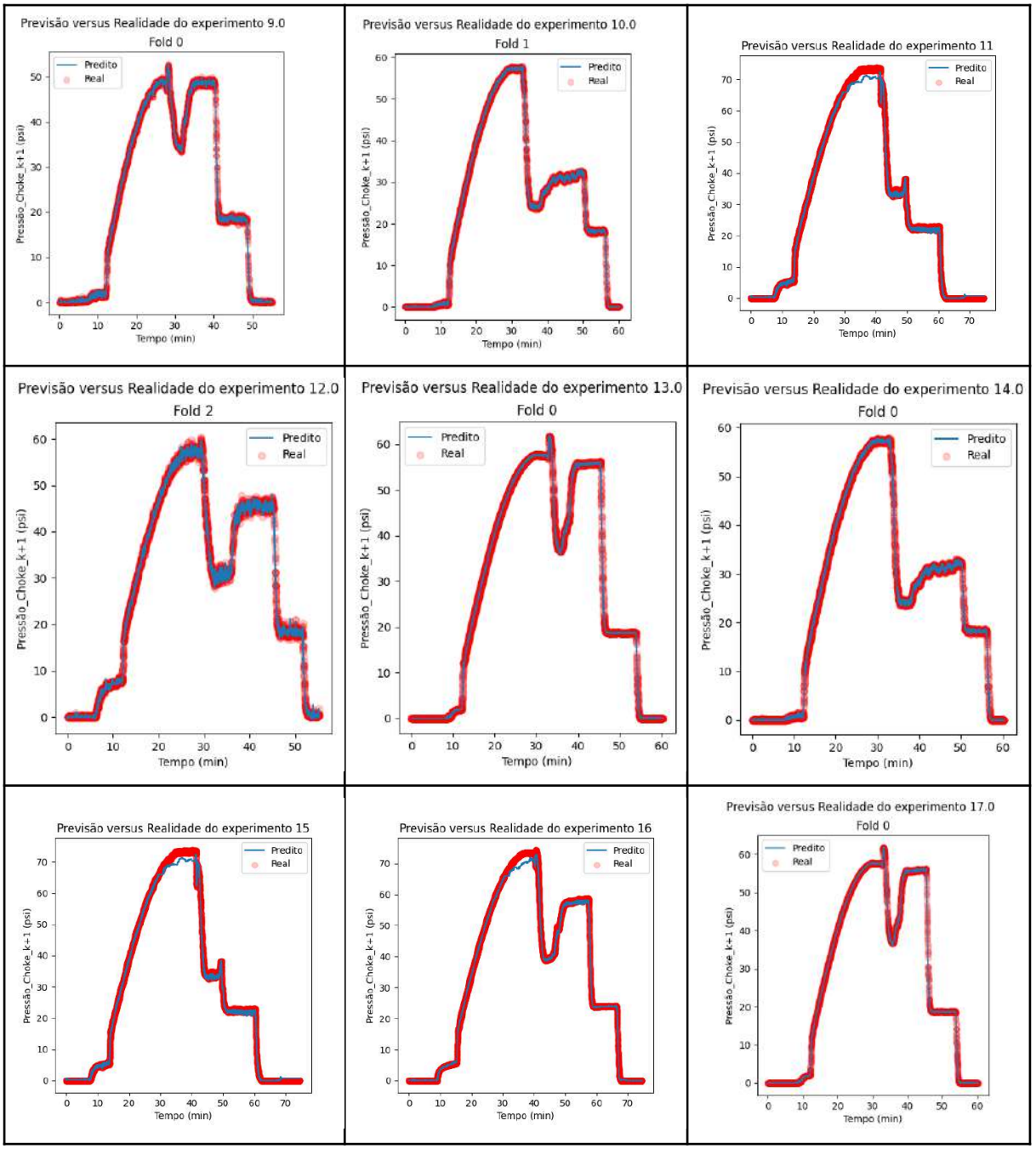

Figura *118* – Previsão e realidade dos experimentos de 9 ao 17 com sig e com 2 dados passados. Fonte: A autora.

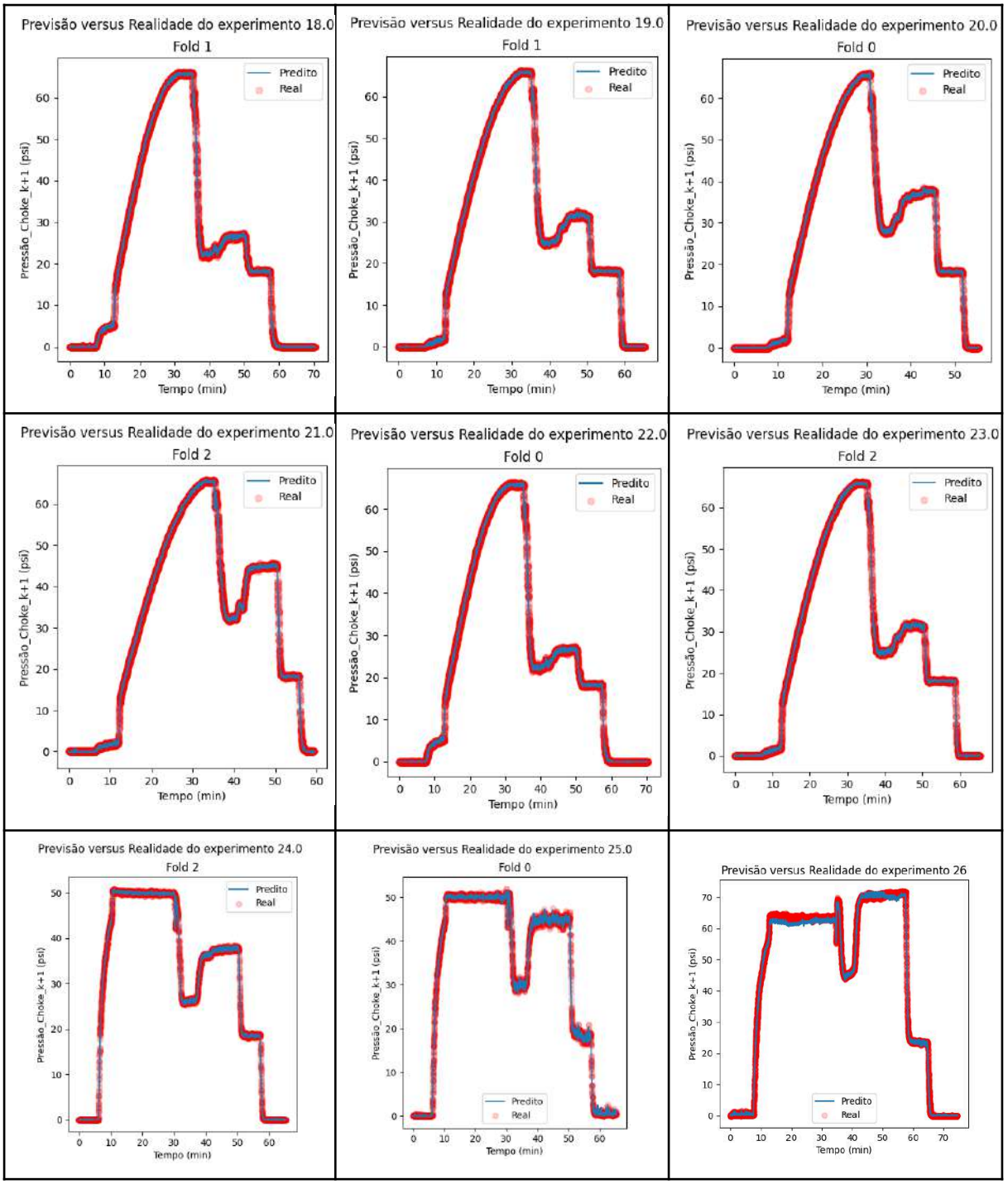

Figura *119* – Previsão e realidade dos experimentos de 18 ao 26 com sig e com 2 dados passados. Fonte: A autora.

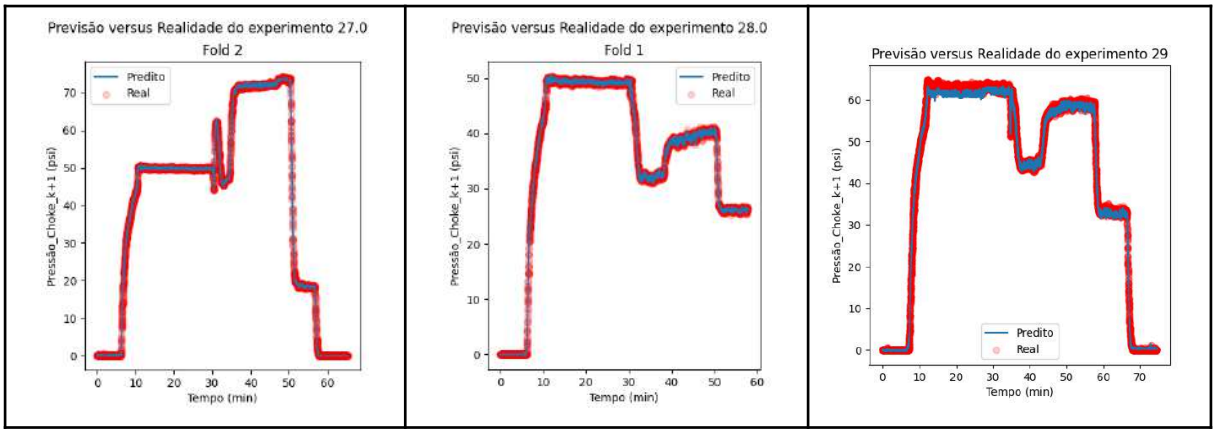

Figura *120* – Previsão e realidade dos experimentos de 27 ao 29 com sig e com 2 dados passados. Fonte: A autora.

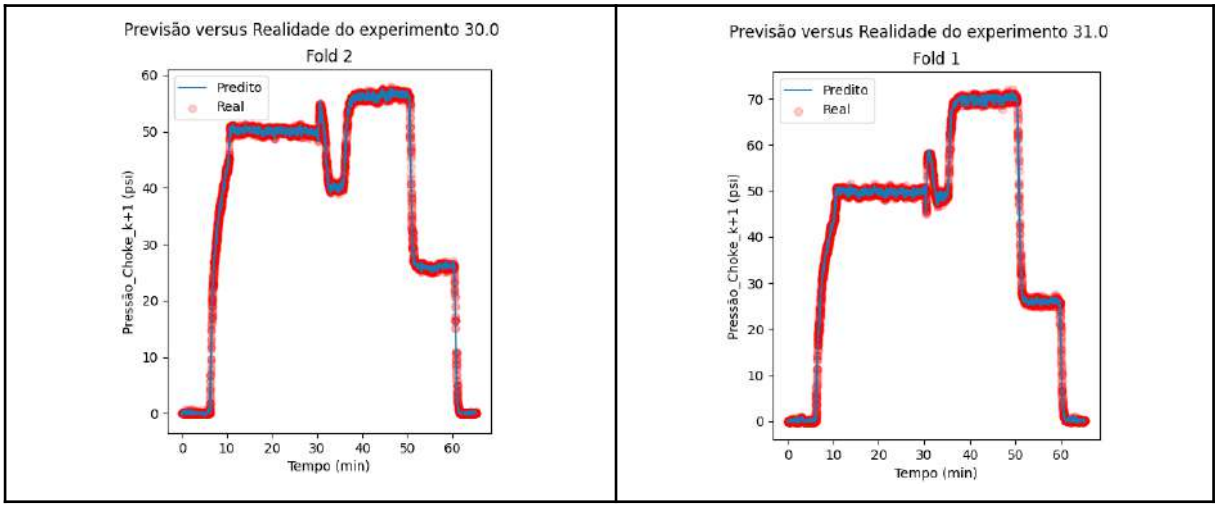

Figura *121* – Previsão e realidade dos experimentos de 30 ao 31 com sig e com 2 dados passados. Fonte: A autora.

ANEXO C – RESULTADOS DOS DADOS EXPERIMENTAIS, COM 8 DADOS PASSADOS, COM ESCALA DE -4 A 4 E TRANSFORMADOS COM SIG

Os resultados das previsões e realidade são apresentados nas Figuras 122 a 126, que representam o quanto o modelo consegue prever a operação de PMCD.

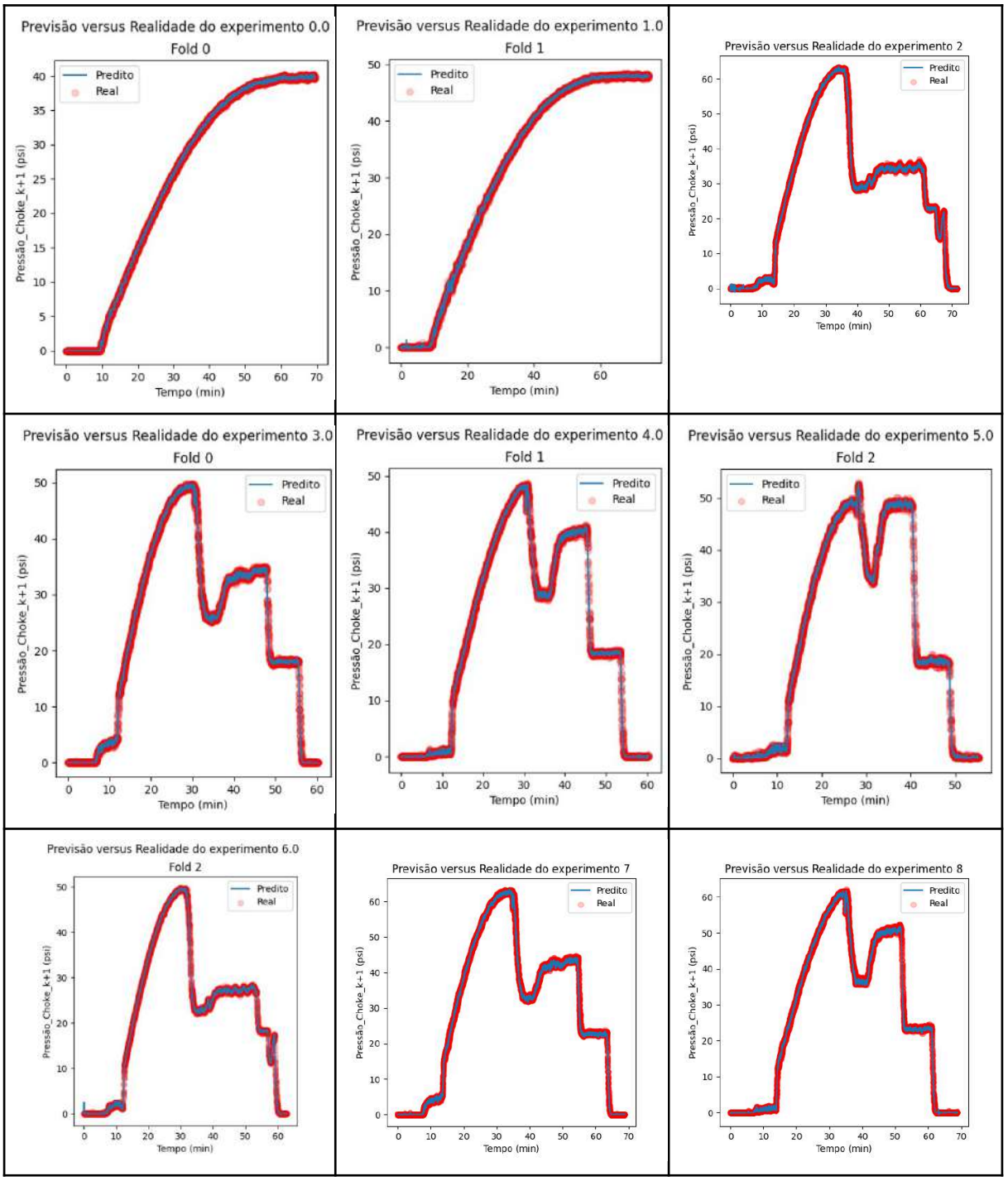

Figura *122* – Previsão e realidade dos experimentos de 0 ao 8 com sig e com 8 dados passados. Fonte: A autora.

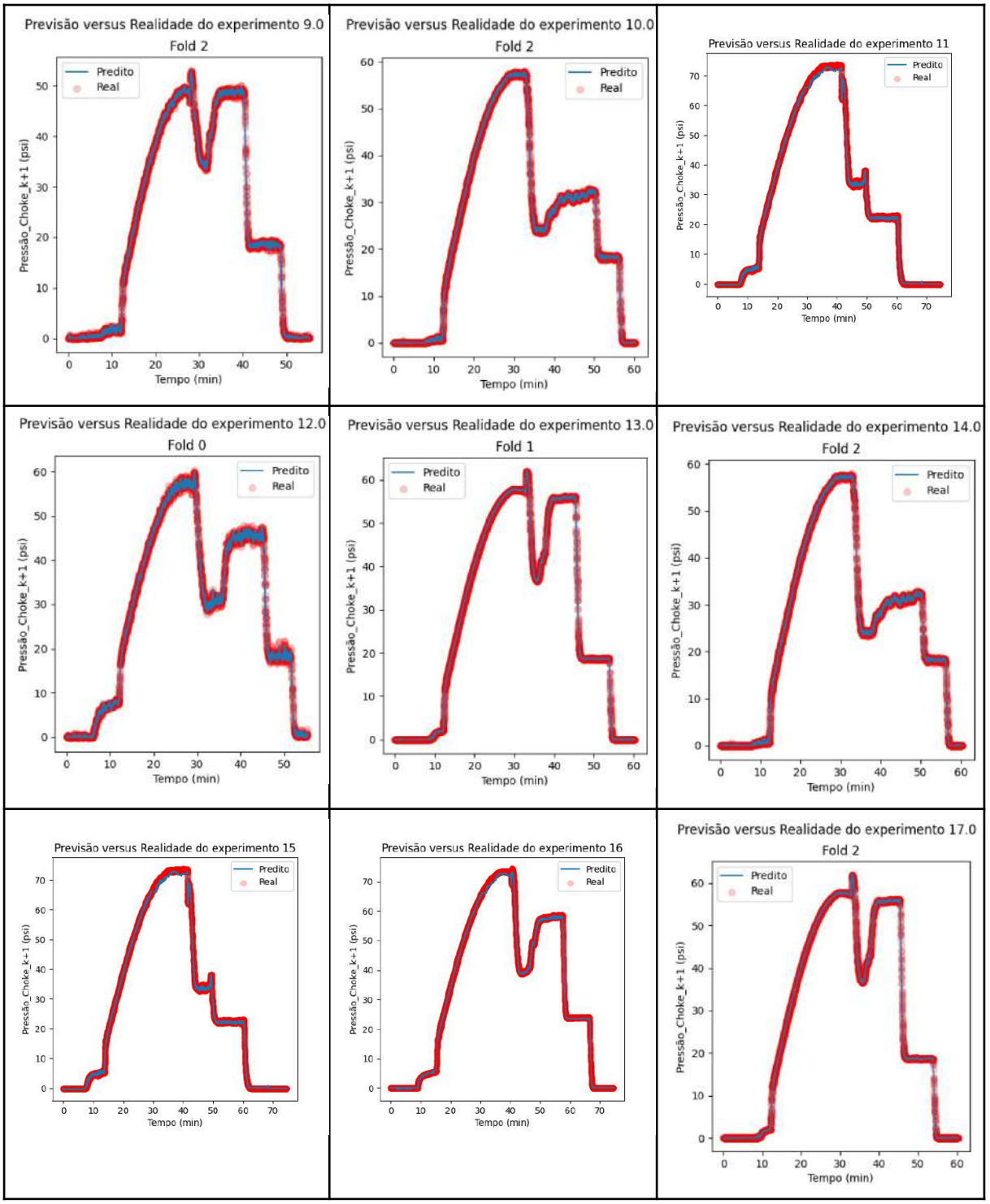

Figura *123* – Previsão e realidade dos experimentos de 9 ao 17 com sig e com 8 dados passados. Fonte: A autora.

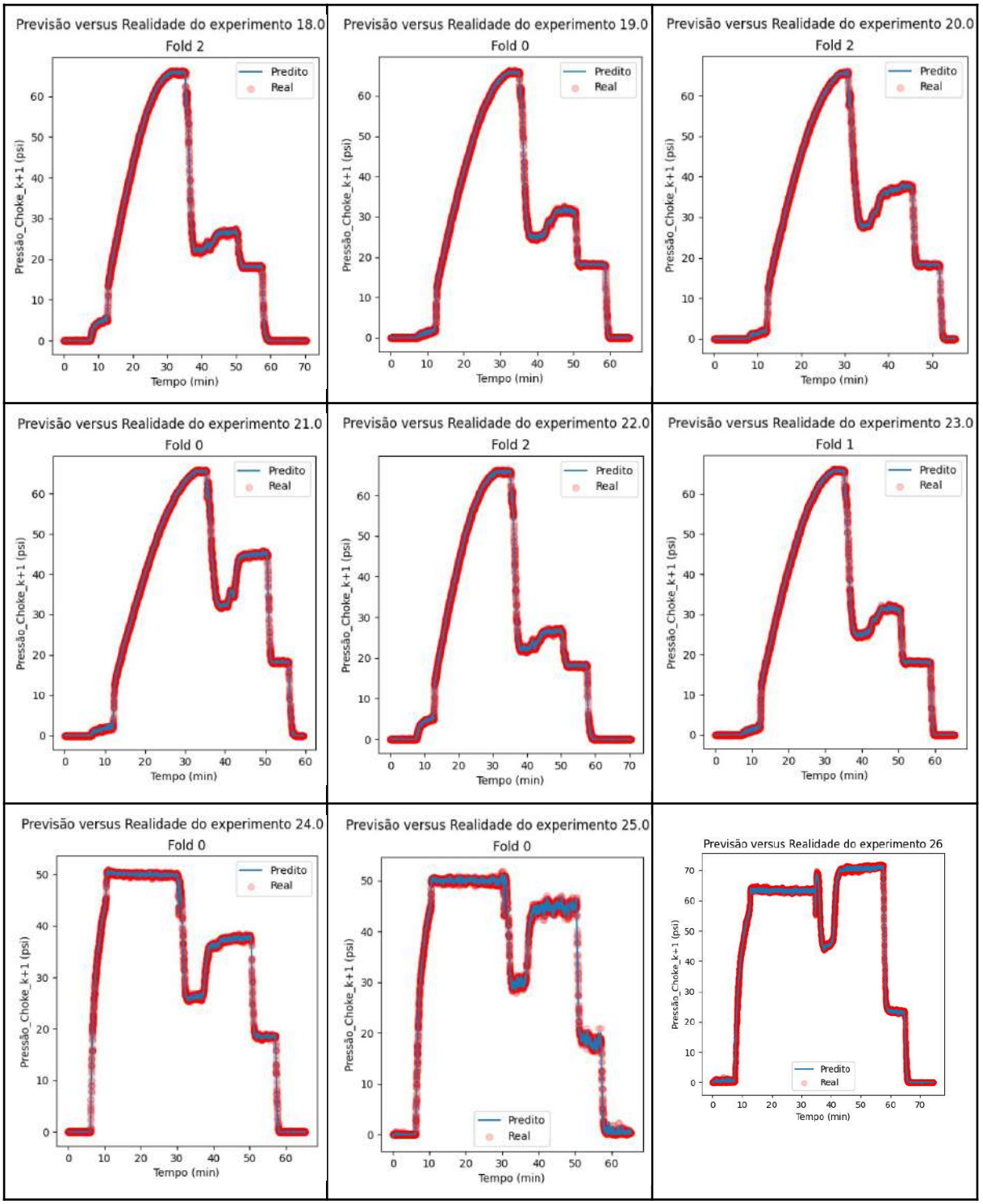

Figura *124* – Previsão e realidade dos experimentos de 18 ao 26 com sig e com 8 dados passados. Fonte: A autora.

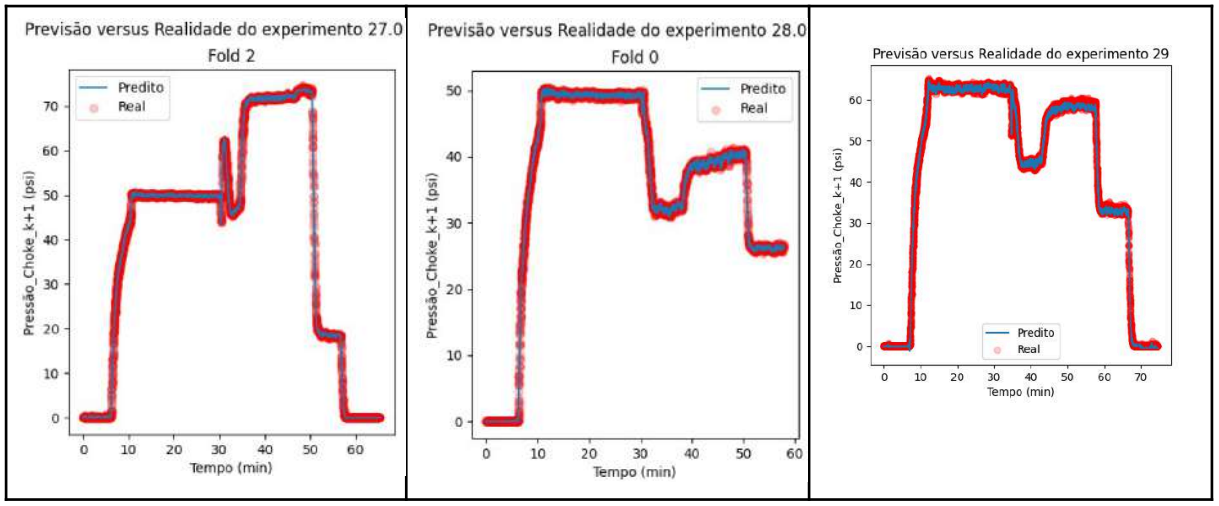

Figura *125* – Previsão e realidade dos experimentos de 27 ao 29 com sig e com 8 dados passados. Fonte: A autora.

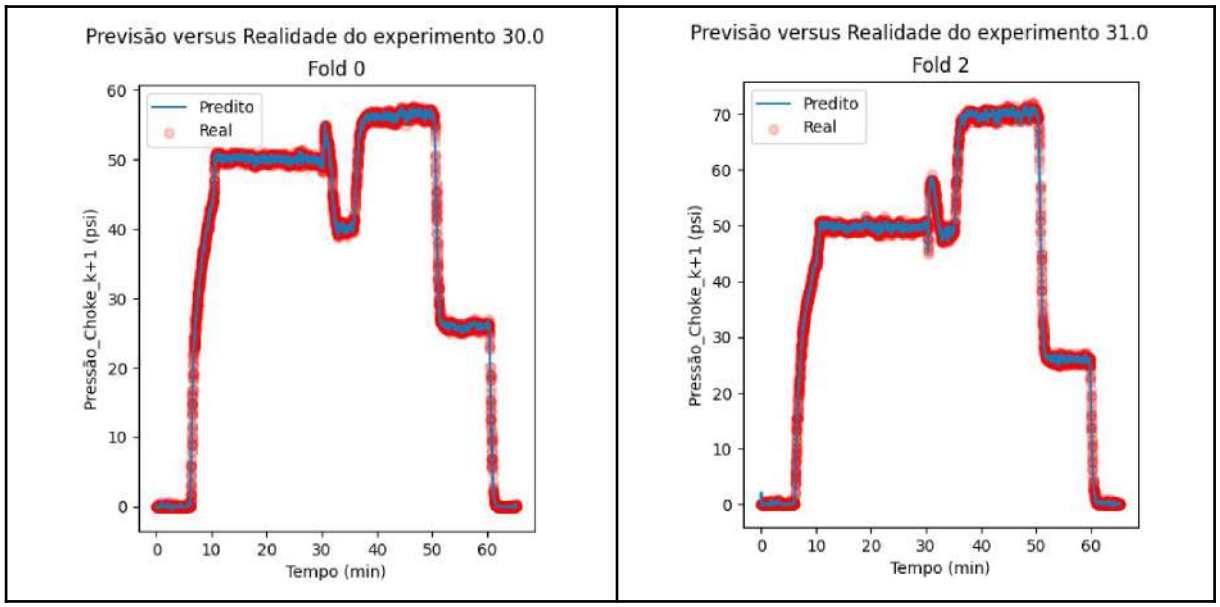

Figura *126* – Previsão e realidade dos experimentos de 30 ao 31 com sig e com 8 dados passados. Fonte: A autora.

ANEXO D – RESULTADOS DOS DADOS EXPERIMENTAIS TRANSFORMADOS COM LOG E SEM DADOS PASSADOS

Não houve necessidade de se criar variáveis defasadas no tempo, foram identificadas as seguintes variáveis contendo *outliers:* vazão e frequência do inversor. São mostrados na Figura 127 o resumo do *dataframe* após o tratamento com a função log e antes de seguirem para o treinamento.

	Valores_Únicos	Tipo_de_Dado	mean	std	min	25%	50%	75%	max
Tempo_poco (min)	169209	float64	0.33	0.18	0.0	0.18	0.33	0.48	0.69
Pressao (psi)	195503	float64	0.35	0.20	0.0	0.20	0.35	0.55	0.69
Vazão (m³/h)	149641	float64	0.05	0.08	0.0	0.00	0.00	0.07	0.69
Freq_Inversor (Hz)	6	float64	0.05	0.14	0.0	0.00	0.00	0.00	0.69
Abertura_choke (%)	198	float64	0.45	0.30	0.0	0.03	0.67	0.67	0.69
Vazão2 (m³/h)	148937	float64	0.05	0.08	0.0	0.00	0.00	0.07	0.69
Abertura_Valvula_Reservatorio (%)	4	float64	0.33	0.29	0.0	0.00	0.24	0.69	0.69
Tempo_tanque (min)	387482	float64	0.33	0.18	0.0	0.18	0.33	0.48	0.69
Pressao_Tanque (psi)	354034	float64	0.43	0.19	0.0	0.23	0.43	0.61	0.69
Pressao_Choke (psi)	155660	float64	35.59	26.45	0.0	4.40	36.30	60.21	94.68
experimento	32	int64	15.61	9.35	0.0	7.00	16.00	24.00	31.00

Figura *127* – Resumo do *dataframe* com dados experimentais, tratados com a função log e sem aplicação de dados passados. Fonte: A autora.

Antes dos dados serem treinados pelo XGBoost, é feito a otimização dos hiperparâmetros com o Optuna: 'n_estimators': 160, 'learning_rate': 0.2205, 'max_depth': 4, 'min_child_weight': 4, 'subsample': 0.6037, 'colsample_bytree': 0.6867, 'gamma': 0.2657, 'reg_alpha': 0.9762 e 'reg_lambda': 0.8346, em apenas 1 minuto e 55 segundos de execução, executado no back-end do Google Compute Engine em Python 3, com 12.7 GB de memória e 107.7 GB disponíveis em disco.

Com os melhores parâmetros gerados pelo Optuna, os dados são treinados, validados e testados pelo XGBoost, gerando um modelo com o melhor aprendizado. Métricas são apresentadas nas Figuras 128 e 129.

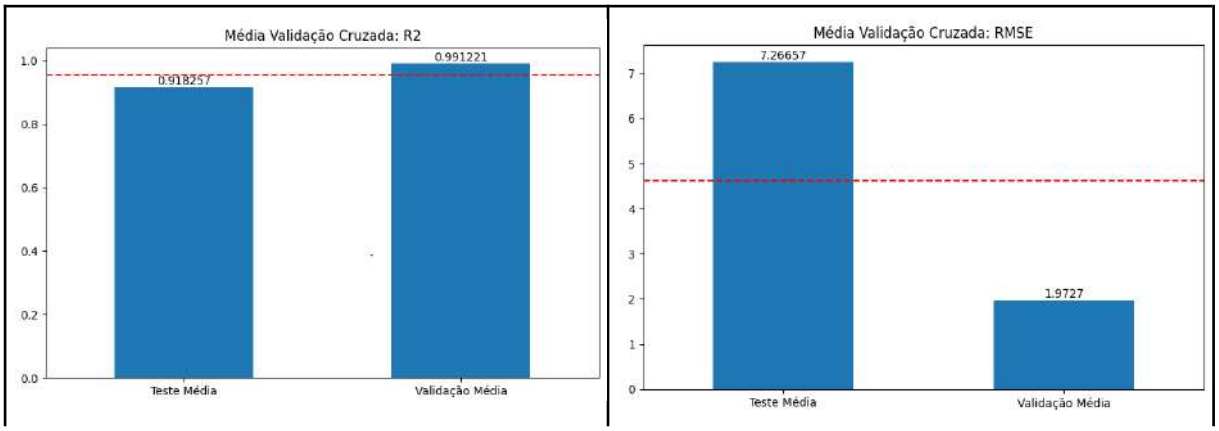

Figura *128* – Métricas de avaliação R² e RMSE com dados experimentais, tratados com a função log e sem dados passados. Fonte: A autora.

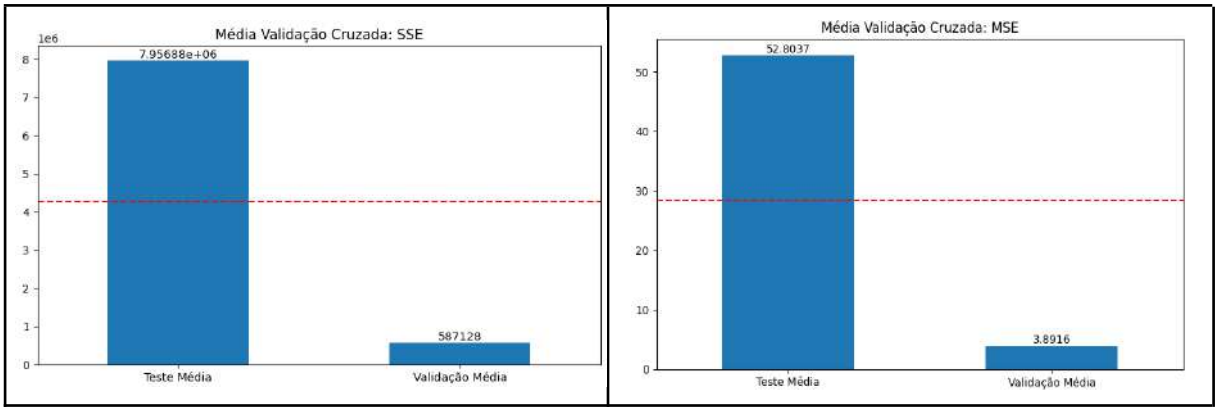

Figura *129* – Métricas de avaliação SSE e MSE com dados experimentais, tratados com a função log e sem dados passados. Fonte: A autora.

São apresentados na Figura 130 os valores da densidade dos dados para a pressão na *choke*, sendo que quanto mais valores preditos de forma correta, ou seja, quando valor predito é igual ao valor real, mais será observada a presença da cor roxa nos gráficos.

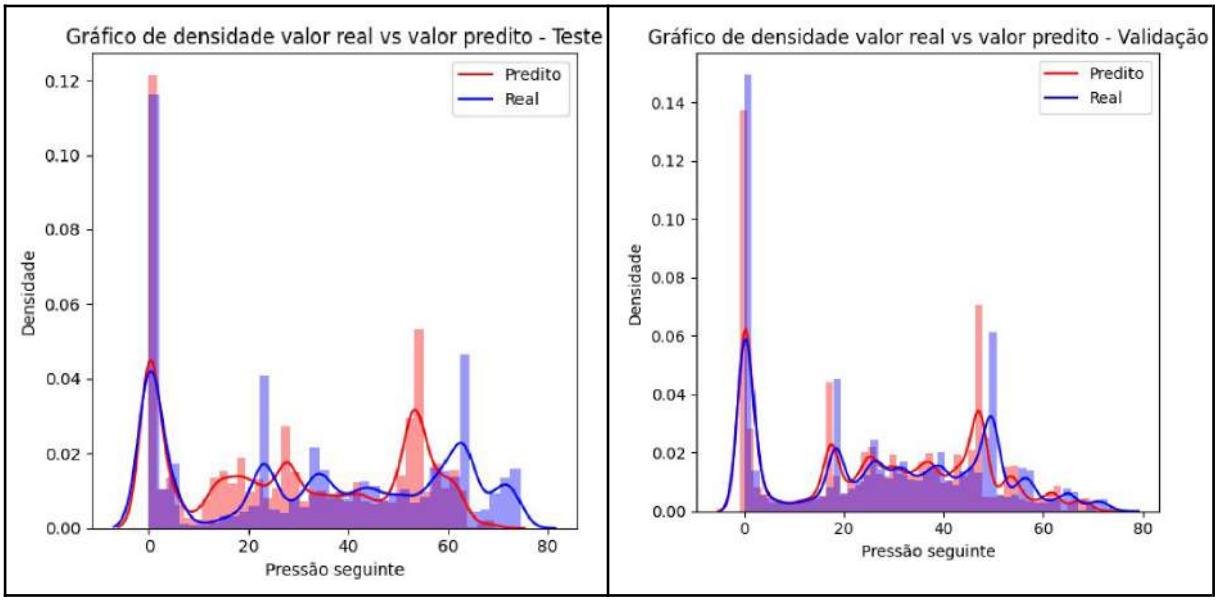

Figura *130* – Gráficos de densidade com dados experimentais, tratados com a função log e sem aplicação de dados passados. Fonte: A autora.

São apresentados na Figura 131 os valores da distribuição dos resíduos no teste e na validação do modelo.

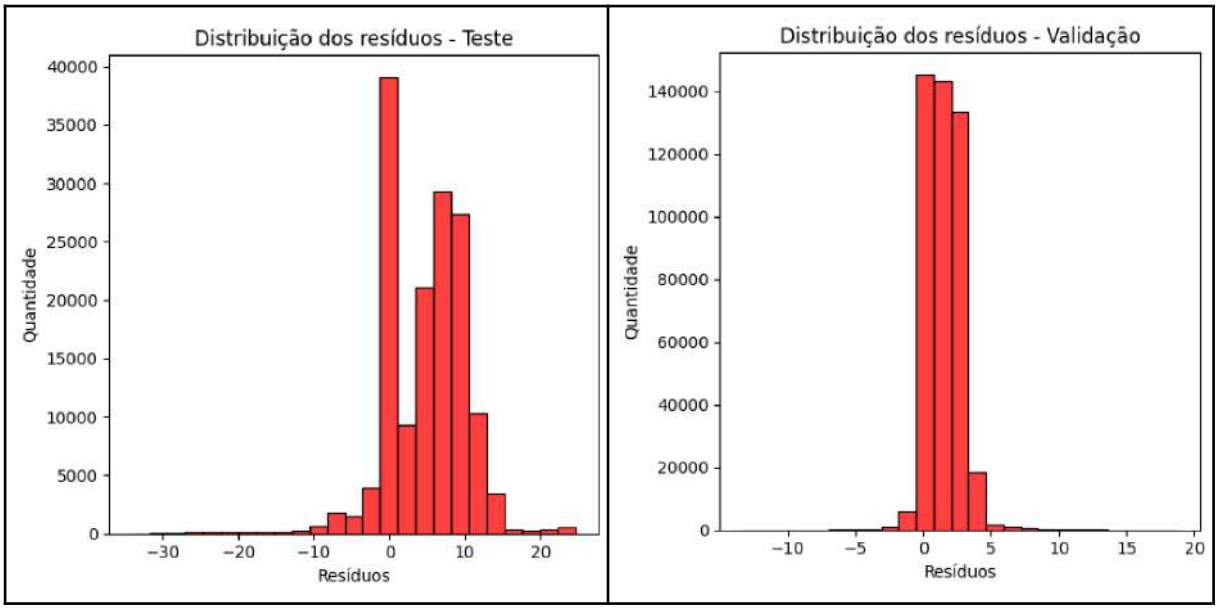

Figura *131* – Distribuição dos resíduos com dados experimentais, tratados com a função log e sem aplicação de dados passados. Fonte: A autora.

A Figura 132 apresenta os pontos na curva de comparação de valor real e valor predito.

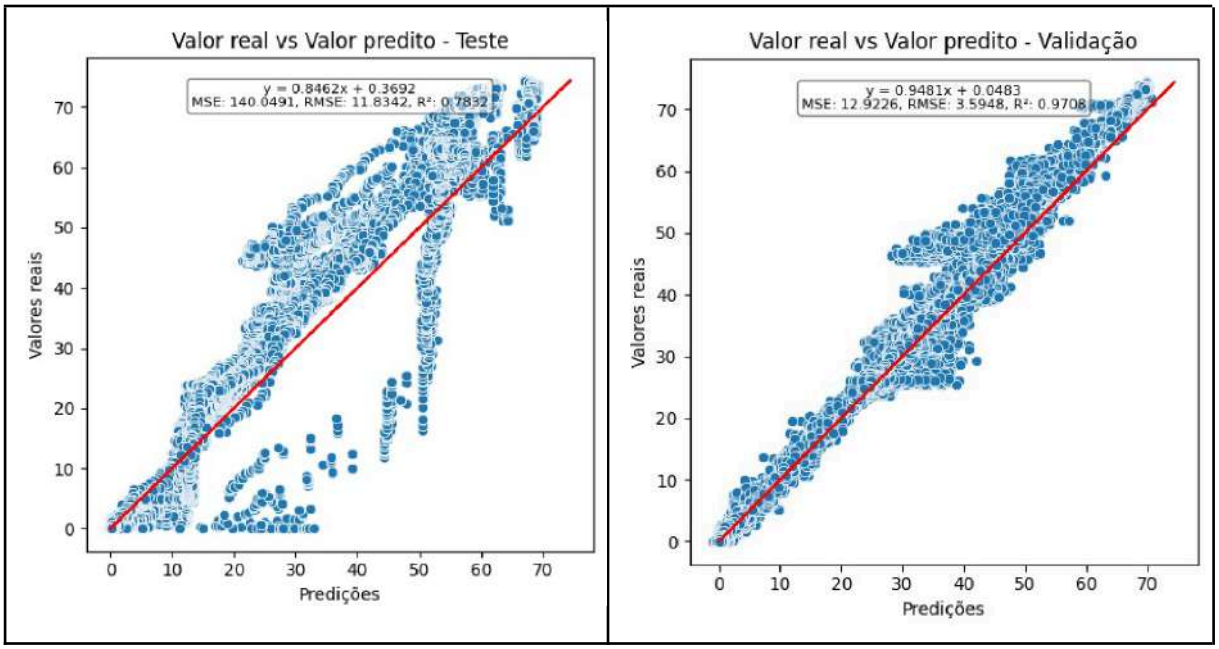

Figura *132* – Gráficos de evolução do modelo com dados experimentais, tratados com a função log e sem aplicação de dados passados. Fonte: A autora.

O gráfico da função de perda é apresentado na Figura 133.

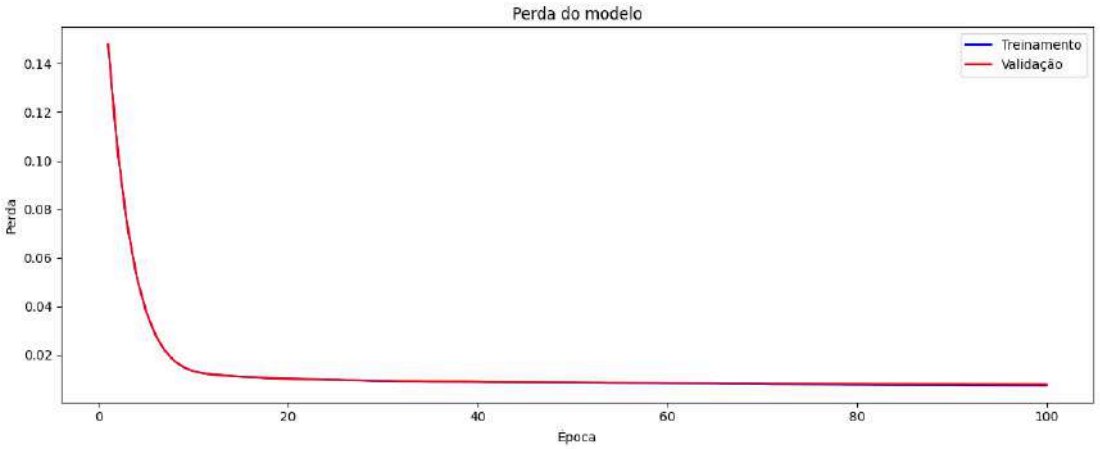

Figura *133* – Curva de perdas função log e sem aplicação de dados passados. Fonte: A autora.

O gráfico na Figura 134 representa a importância de cada variável para as previsões, sendo que a pressão do poço teve uma forte influência nas previsões, seguido da frequência do inversor da bomba e da abertura da válvula *choke.*

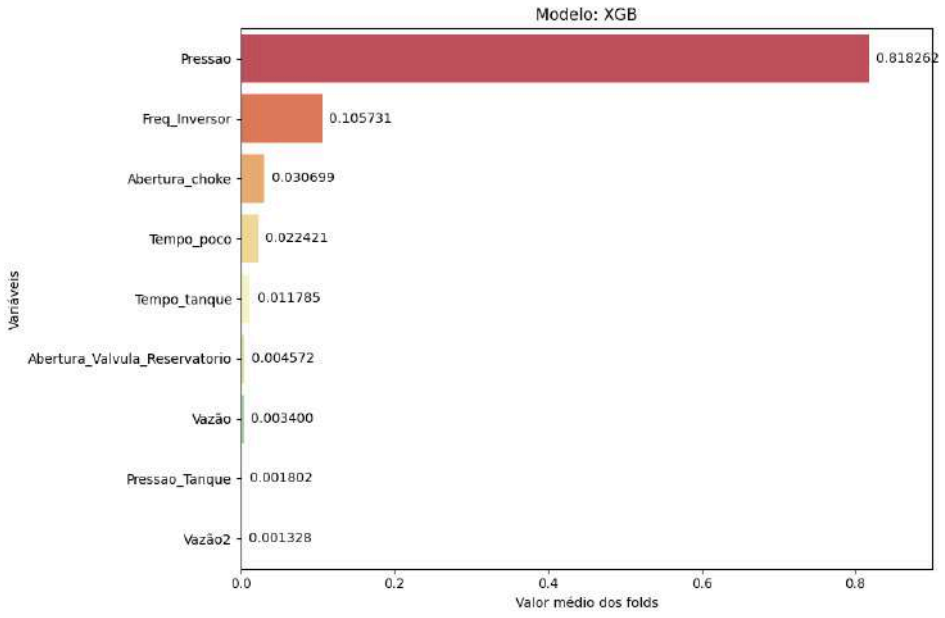

Figura *134* – Importância das variáveis com dados experimentais, tratados com a função log e sem dados passados. Fonte: A autora.

Os resultados das previsões e realidade do modelo são apresentados nas Figuras 135 a 138, que representam o quanto o modelo consegue prever a operação de PMCD.

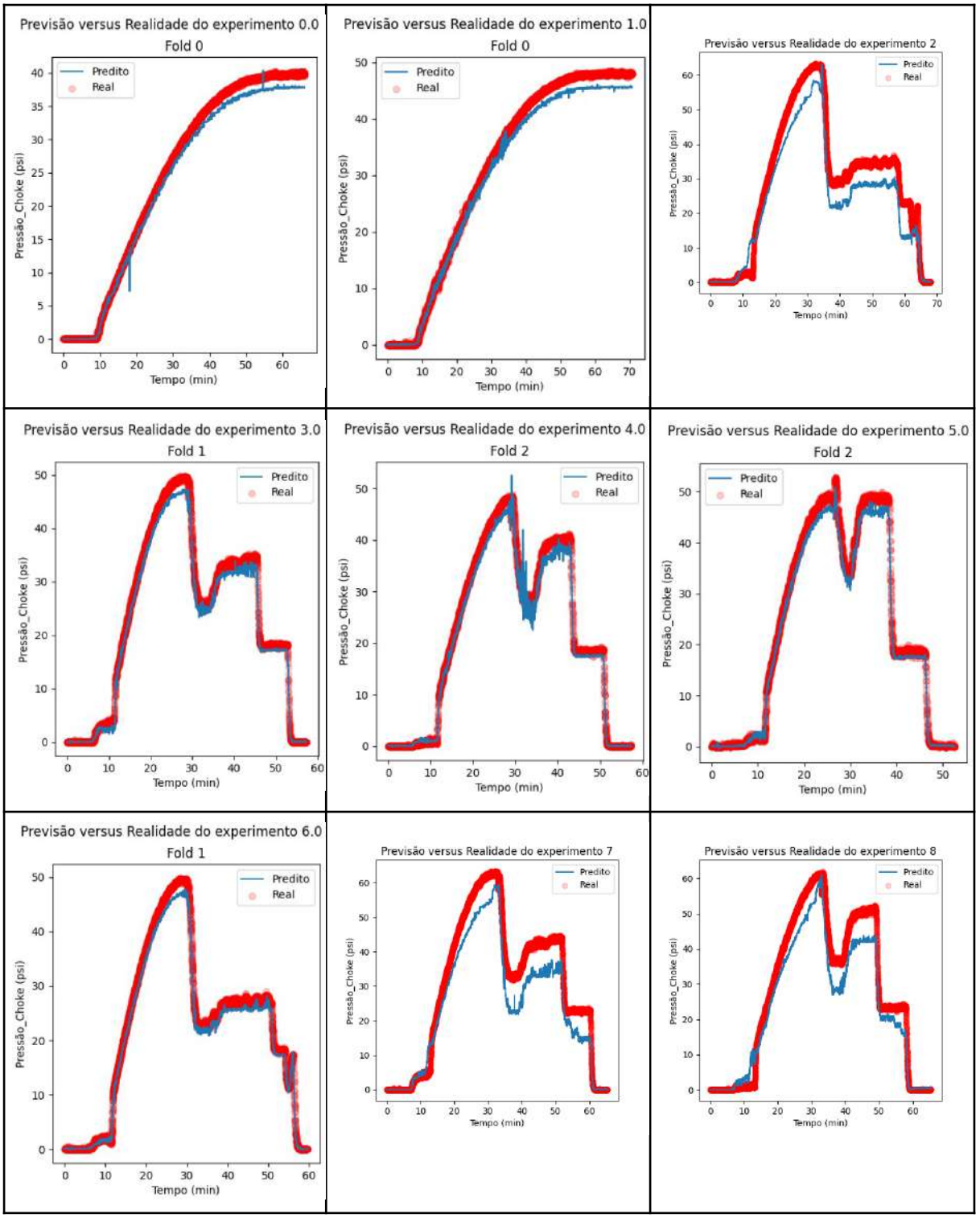

Figura *135* – Previsão e realidade dos experimentos de 0 a 8 com log e sem dados passados. Fonte: A autora.

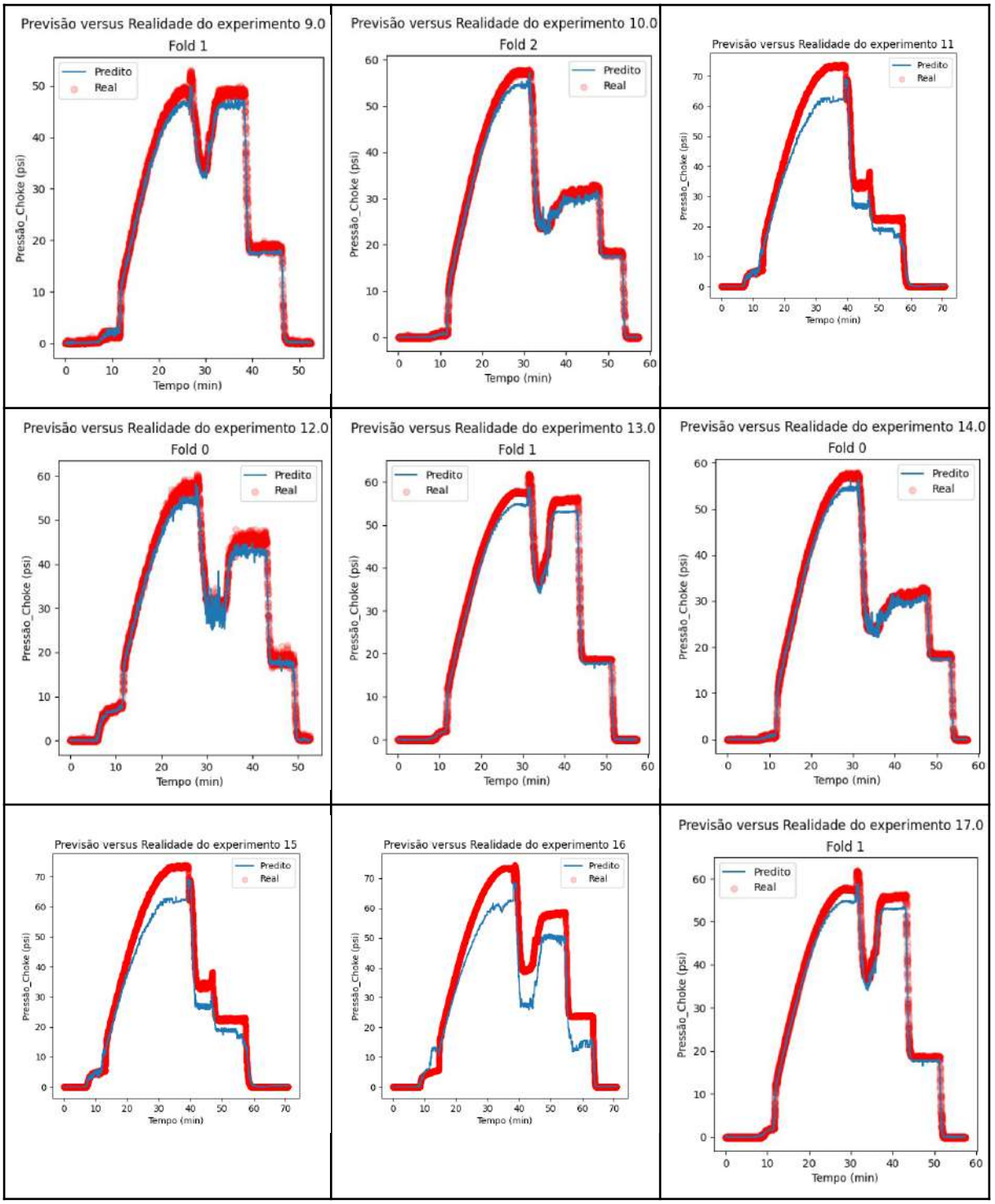

Figura *136* – Previsão e realidade dos experimentos de 9 ao 17 com log e sem dados passados. Fonte: A autora.

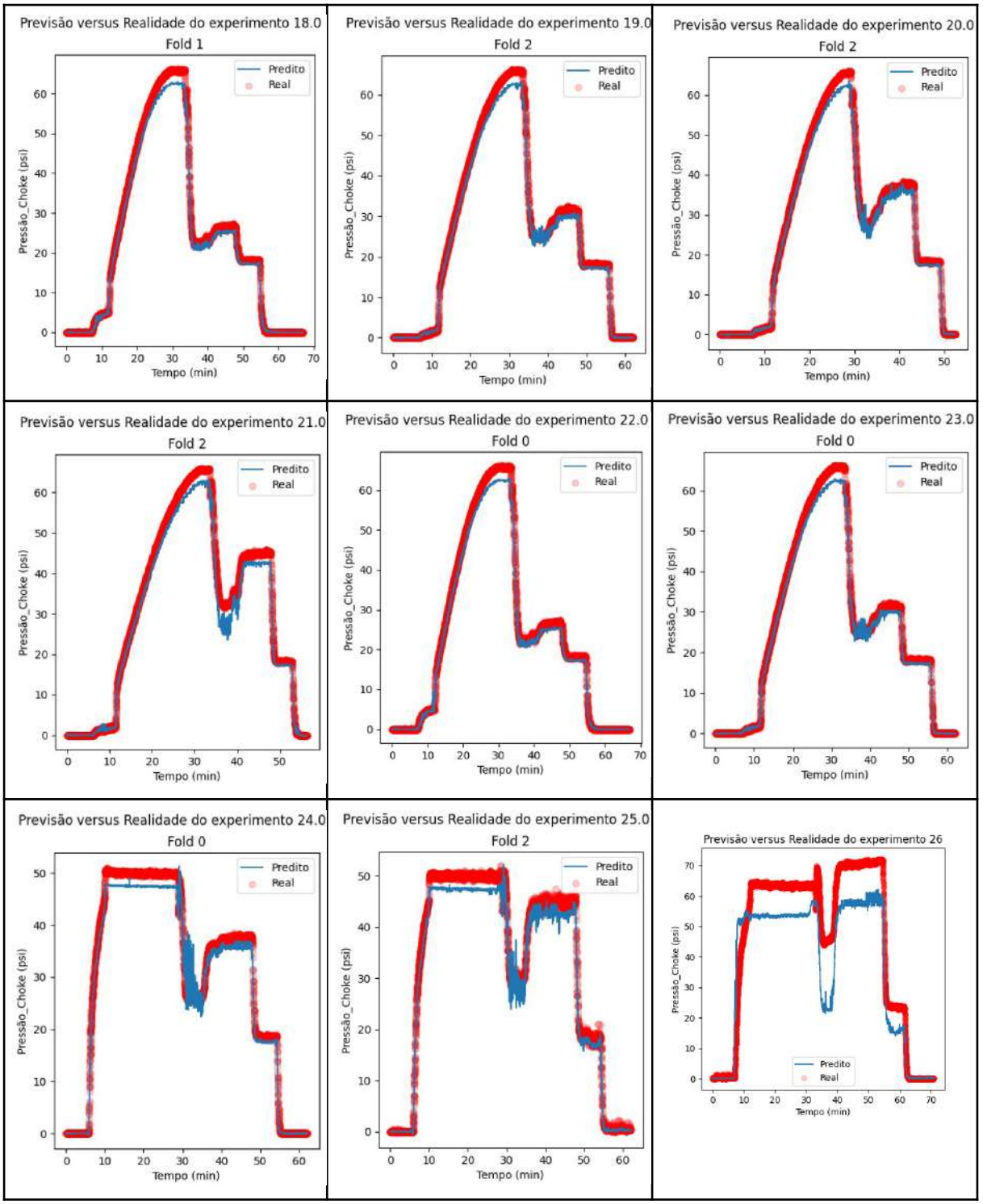

Figura *137* – Previsão e realidade dos experimentos de 18 ao 26 com log e sem dados passados. Fonte: A autora.

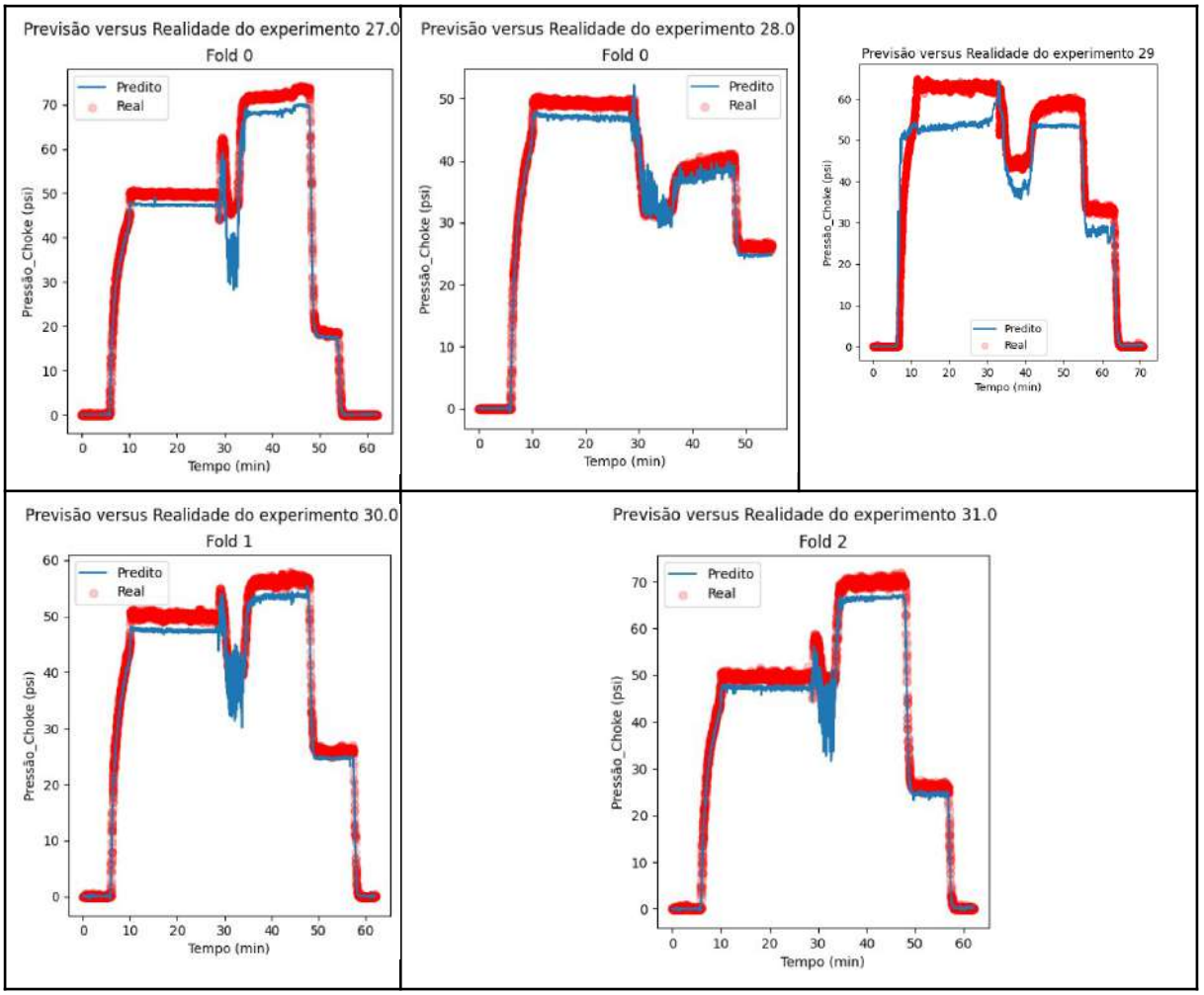

Figura *138* – Previsão e realidade dos experimentos de 27 ao 31 com log e sem dados passados. Fonte: A autora.

Com relação à arquitetura do modelo, as métricas de avaliação do modelo revelam que a ausência de dados passados comprometeu a qualidade do modelo. O modelo treinou 160 árvores.

ANEXO E – RESULTADOS DOS DADOS EXPERIMENTAIS TRANSFORMADOS COM LOG E COM 2 DADOS PASSADOS

O uso de 2 dados passados faz com que as variáveis sejam deslocadas. Os *outiliers* foram tratados com a substituição pelas suas médias (Figura 139).

	Valores_Únicos	Tipo_de_Dado	mean	std	min	25%	50%	75%	max
Pressao_Choke_k-1 (psi)	176138	float64	0.31	0.22	0.0	0.05	0.35	0.51	0.69
Tempo_poco_k (min)	169124	float64	0.33	0.18	0.0	0.18	0.33	0.48	0.69
Pressao_k (psi)	195390	float64	0.35	0.20	0.0	0.20	0.35	0.55	0.69
Vazão_k (m³/h)	149564	float64	0.05	0.08	0.0	0.00	0.00	0.07	0.69
Freq_Inversor_k (Hz)	6	float64	0.05	0.14	0.0	0.00	0.00	0.00	0.69
Abertura_choke_k (%)	198	float64	0.45	0.30	0.0	0.03	0.67	0.67	0.69
Vazão2_k (m³/h)	148858	float64	0.05	0.08	0.0	0.00	0.00	0.07	0.69
Abertura_Valvula_Reservatorio_k (%)	4	float64	0.33	0.29	0.0	0.00	0.24	0.69	0.69
Tempo_tanque_k (min)	387452	float64	0.33	0.18	0.0	0.18	0.33	0.48	0.69
Pressao_Tanque_k (psi)	354017	float64	0.43	0.19	0.0	0.23	0.43	0.61	0.69
Pressao_Choke_k (psi)	176071	float64	0.31	0.22	0.0	0.05	0.35	0.51	0.69
Pressão_Choke_k+1 (psi)	155655	float64	35.60	26.45	0.0	4.40	36.31	60.21	94.68
experimento	32	int64	15.61	9.35	0.0	7.00	16.00	24.00	31.00

Figura *139* – Resumo do *dataframe* com dados experimentais, tratados com a função log e com 2 dados passados. Fonte: A autora.

Antes dos dados serem treinados pelo XGBoost, é realizada a otimização dos hiperparâmetros com o Optuna gerando os seguintes parâmetros: 'n_estimators': 415, 'learning_rate': 0.2430, 'max_depth': 5, 'min_child_weight': 7, 'subsample': 0.5434, 'colsample_bytree': 0.8106, 'gamma': 0.5195, 'reg_alpha': 0.1787 e 'reg_lambda': 0.3034, em apenas 2 minutos e 33 segundos de execução com o mesmo ambiente de execução do modelo anterior.

Com os melhores parâmetros gerados pelo Optuna, os dados são treinados, validados e testados pelo XGBoost. As métricas de avaliação do modelo estão nas Figuras 140 e 141.

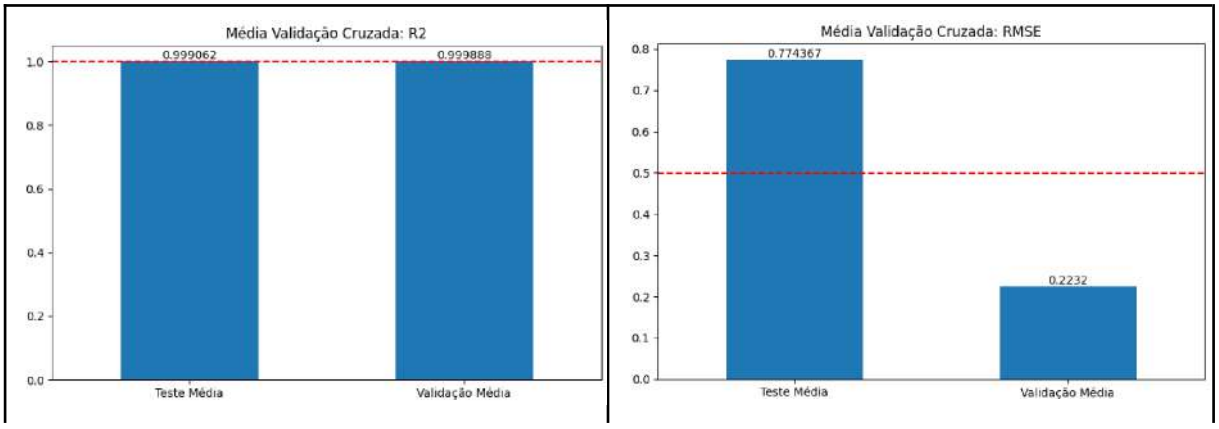

Figura *140* – Métricas de avaliação R² e RMSE, com dados experimentais, tratados com a função log e com 2 dados passados. Fonte: A autora.

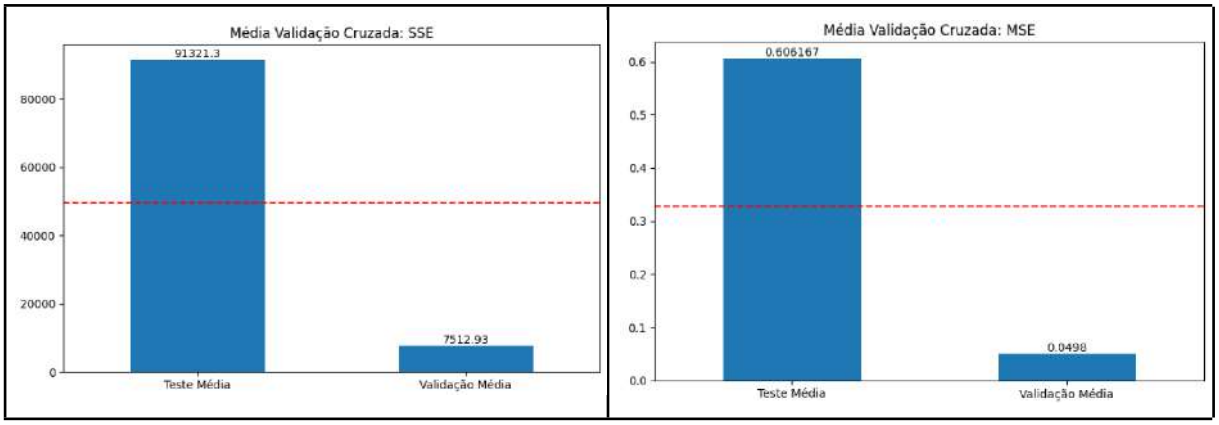

Figura *141* – Métricas de avaliação SSE e MSE, com dados experimentais, tratados com a função log e com 2 dados passados. Fonte: A autora.

São apresentados na Figura 142 os valores da densidade dos dados para a pressão *choke*, sendo que quanto mais valores preditos de forma correta, mais será observada a presença da cor roxa nos gráficos.

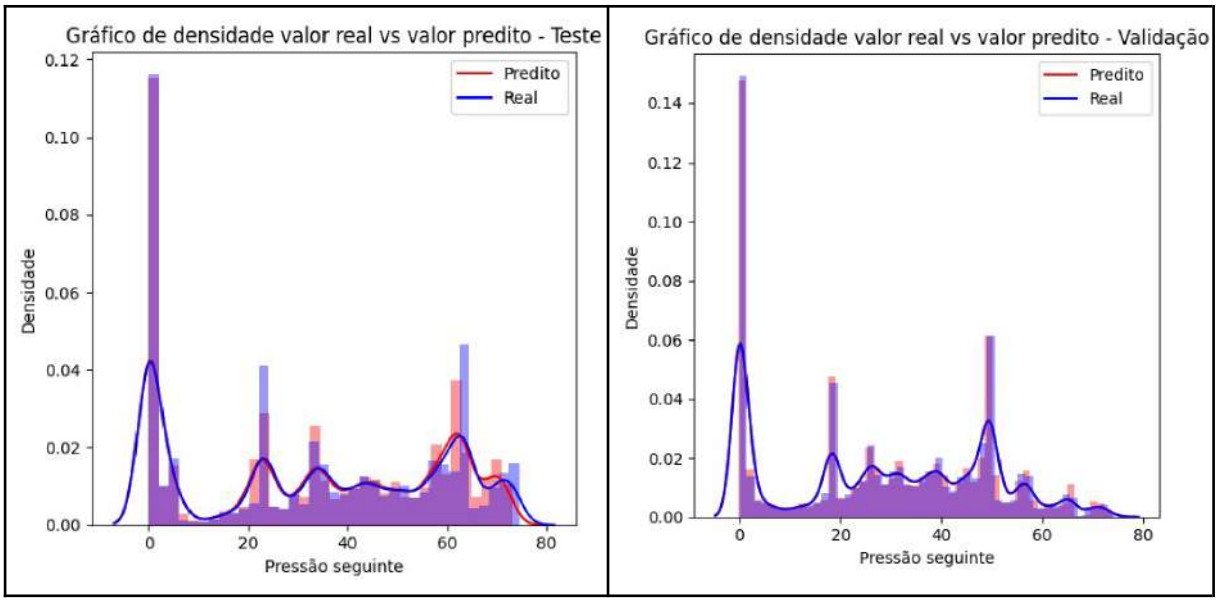

Figura *142* – Gráficos de densidade com dados experimentais, tratados com a função log e com 2 dados passados. Fonte: A autora.

Na Figura 143 os valores da distribuição dos resíduos no teste e na validação do modelo.

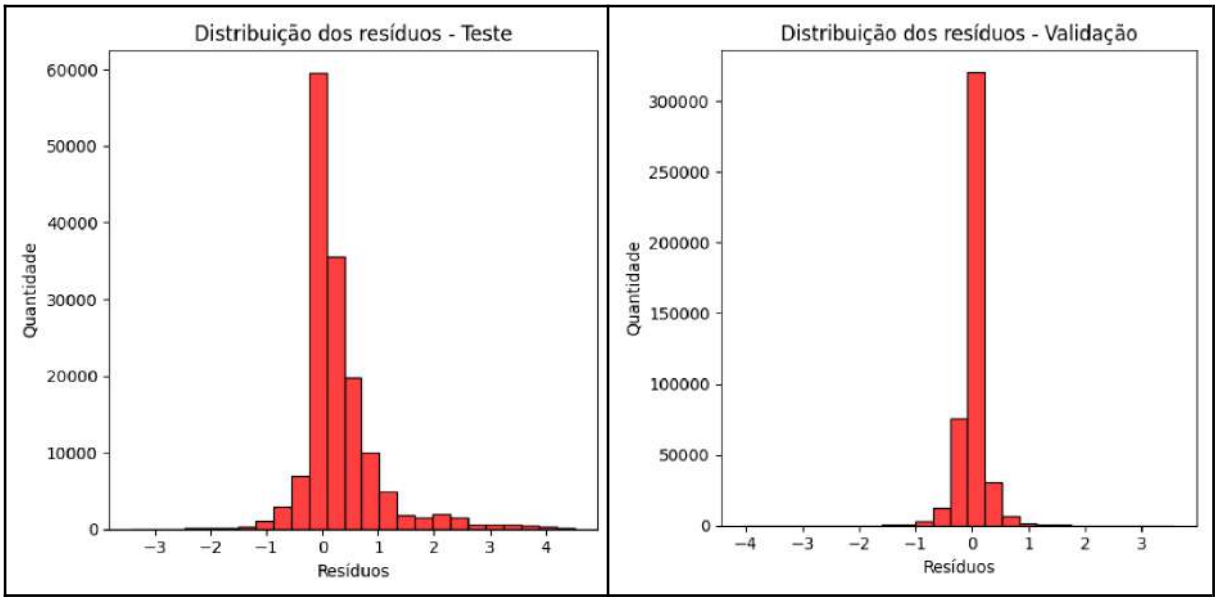

Figura *143* – Distribuição dos resíduos com dados experimentais, tratados com a função log e com 2 dados passados. Fonte: A autora.

A Figura 144 apresenta os pontos na curva de comparação de valor real e valor predito.

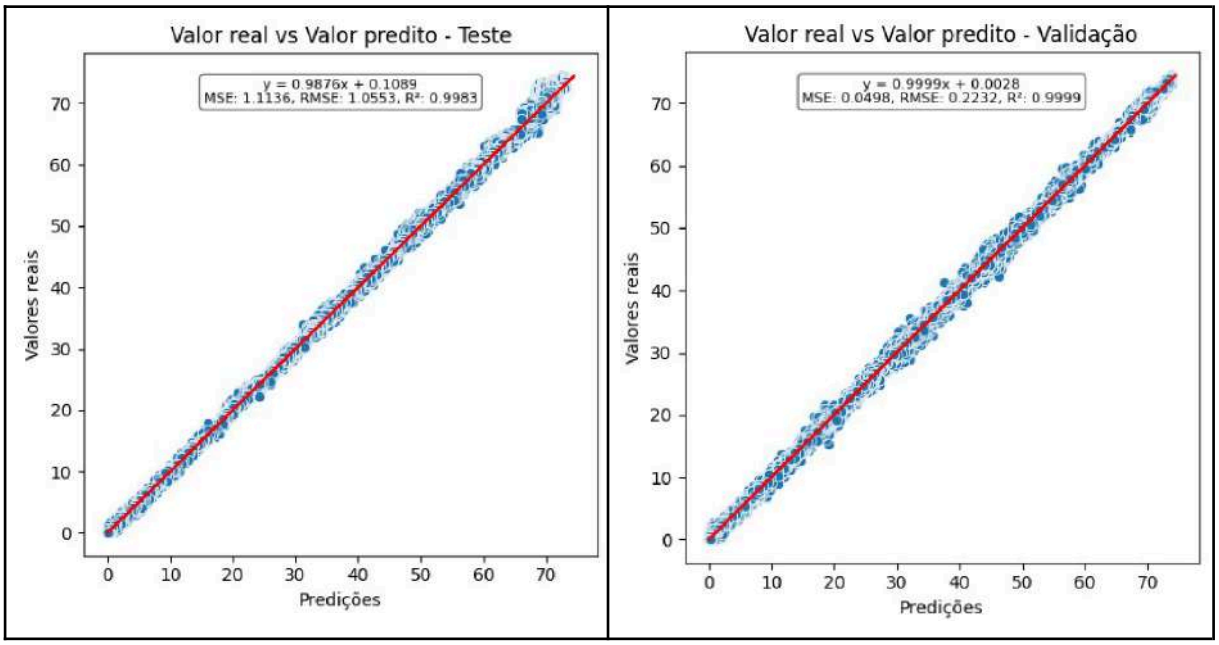

Figura *144* – Gráficos de evolução do modelo com dados experimentais, tratados com a função log e com 2 dados passados. Fonte: A autora.

O gráfico da função de perda apresentado na Figura 145.

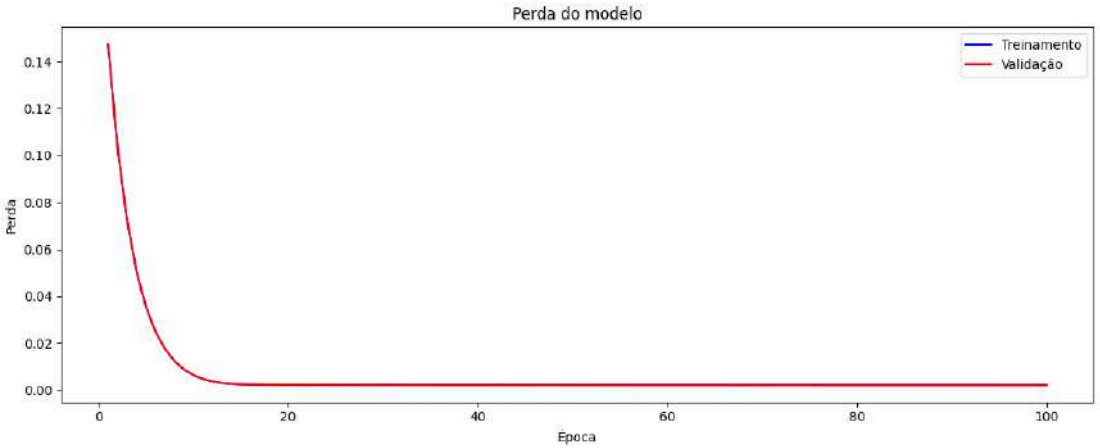

Figura *145* – Curva de perdas com dados experimentais, tratados com a função log e com 2 dados passados. Fonte: A autora.

O gráfico na Figura 146 representa a importância de cada variável para as previsões, sendo que a pressão da *choke* defasados no tempo são as variáveis mais relevantes para a construção do modelo matemático.

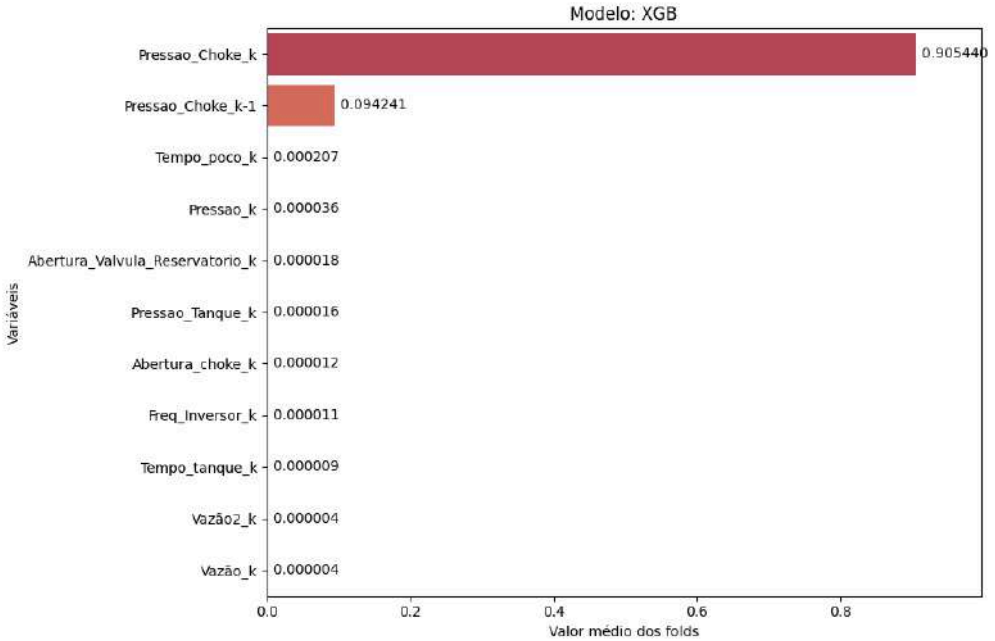

Figura *146* – Importância das variáveis com dados experimentais, tratados com a função log e com 2 dados passados. Fonte: A autora.

Os resultados das previsões e realidade do modelo são apresentados nas Figuras 147 a 150, que representam o quanto o modelo consegue prever a operação de PMCD.

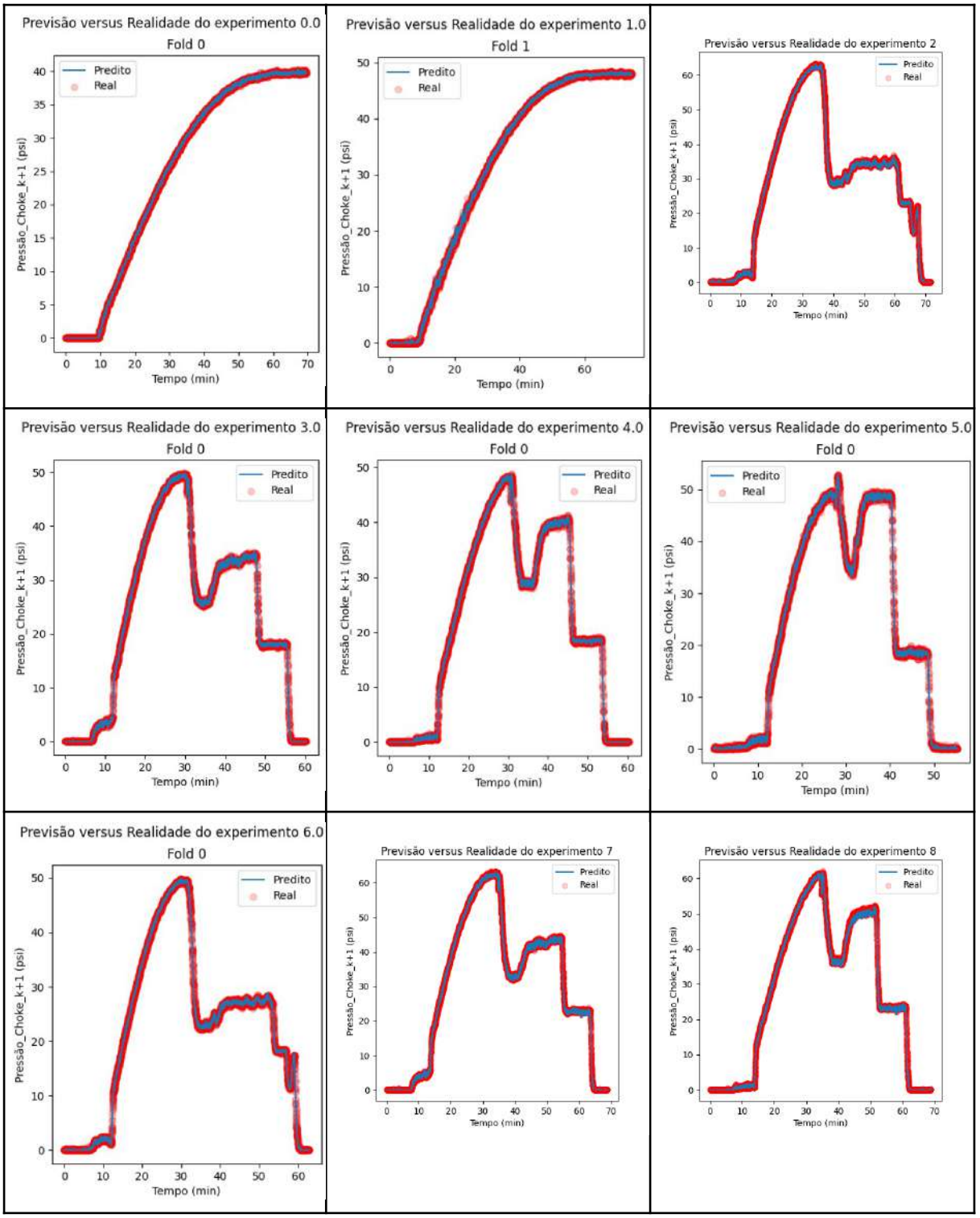

Figura *147* – Previsão e realidade dos experimentos de 0 a 8 com log e com 2 dados passados. Fonte: A autora.

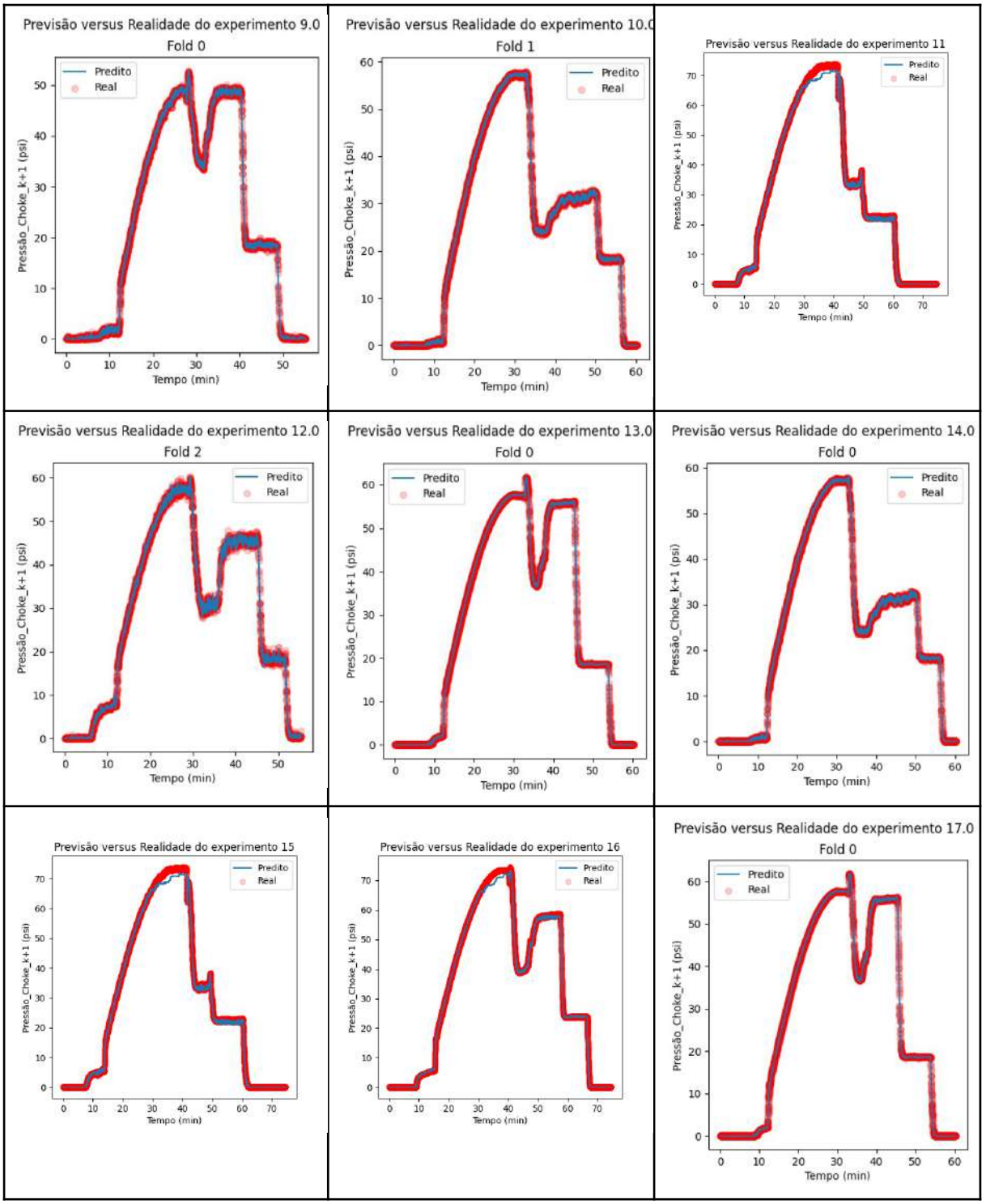

Figura *148* – Previsão e realidade dos experimentos de 9 ao 17 com log e com 2 dados passados. Fonte: A autora.

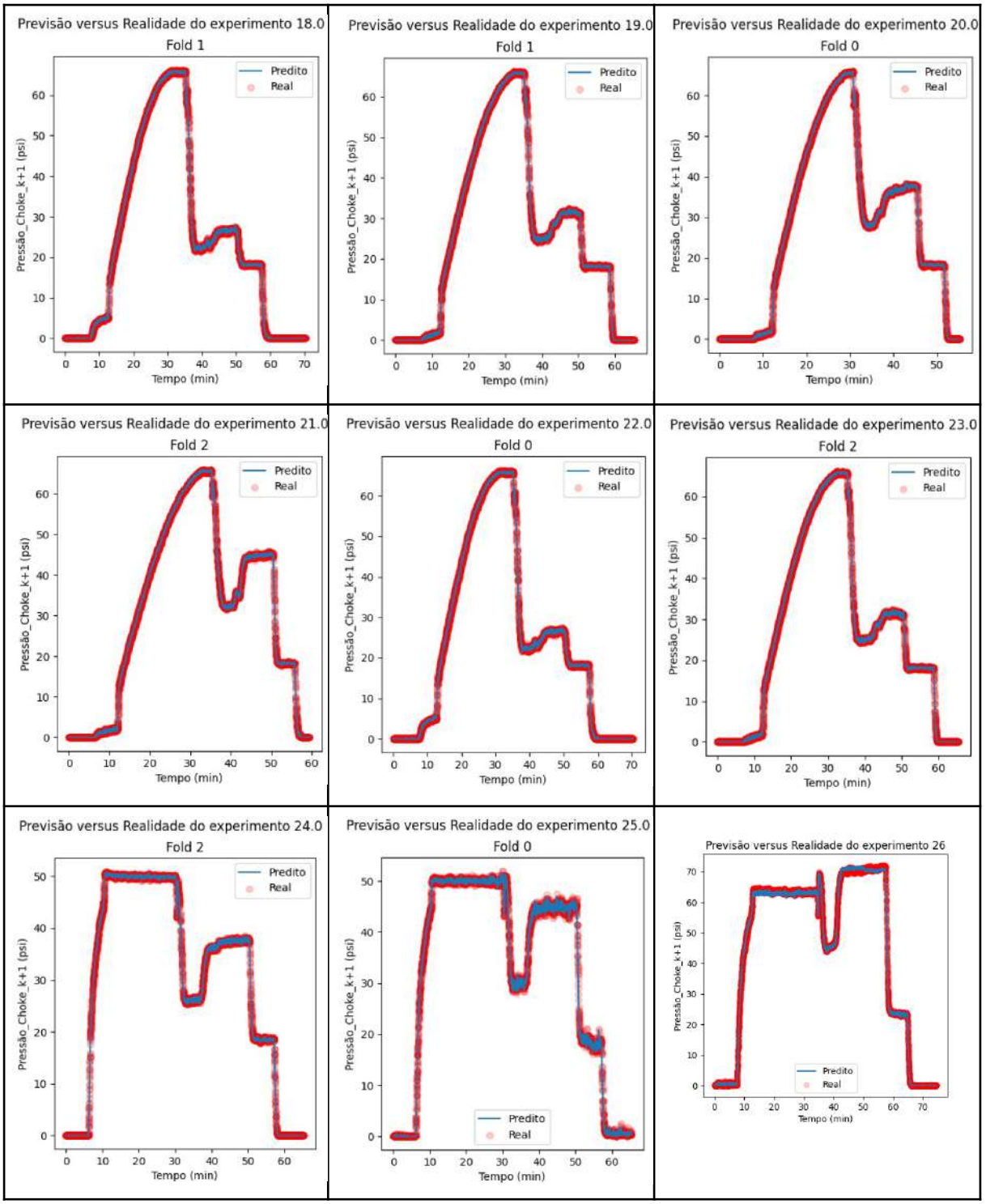

Figura *149* – Previsão e realidade dos experimentos de 18 ao 26 com log e com 2 dados passados. Fonte: A autora.

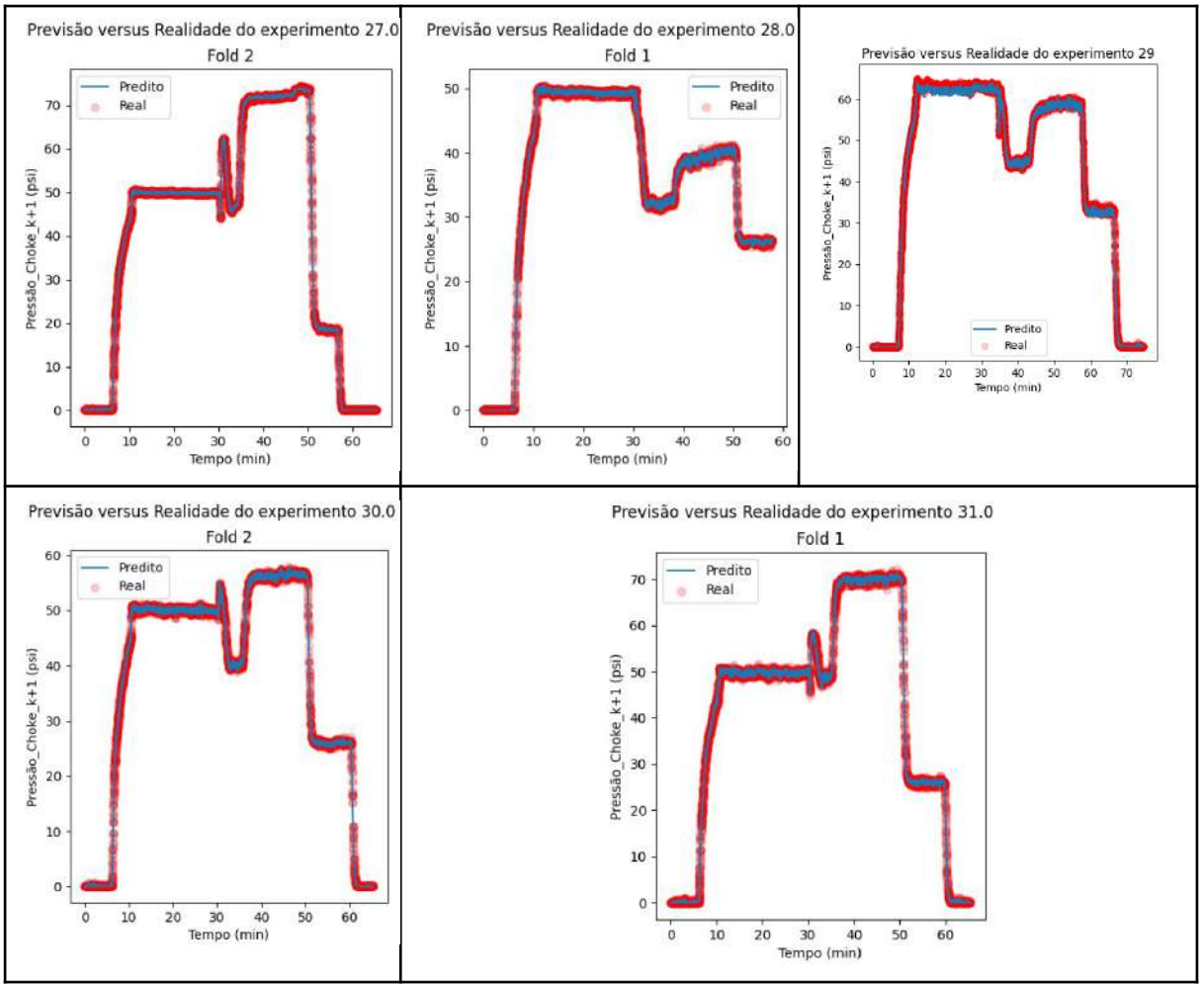

Figura *150* – Previsão e realidade dos experimentos de 27 ao 31 com log e com 2 dados passados. Fonte: A autora.

Com relação à arquitetura do modelo foram empregadas 415 árvores. O uso de 2 dados passados produziu um modelo com métricas de avaliação superiores em comparação ao modelo que não dispõe de informação dinâmica.

ANEXO F – RESULTADOS DOS DADOS EXPERIMENTAIS TRANSFORMADOS COM LOG E COM 8 DADOS PASSADOS

Com o uso de 8 dados passados, todas variáveis são deslocadas. Os *outliers* foram tratados com a substituição pelas suas médias (Figura 151).

	Valores_Únicos	Tipo_de_Dado	mean	std	min	25%	50%	75%	max
Pressao_Choke_k-1 (psi)	176138	float64	0.31	0.22	0.0	0.05	0.35	0.51	0.69
Pressao_Choke_k-2 (psi)	176142	float64	0.31	0.22	0.0	0.05	0.35	0.51	0.69
Pressao_Choke_k-3 (psi)	176136	float64	0.31	0.22	0.0	0.05	0.35	0.51	0.69
Pressao_Choke_k-4 (psi)	176132	float64	0.31	0.22	0.0	0.05	0.35	0.51	0.69
Pressao_Choke_k-5 (psi)	176116	float64	0.31	0.22	0.0	0.05	0.35	0.51	0.69
Pressao_Choke_k-6 (psi)	176115	float64	0.31	0.22	0.0	0.05	0.35	0.51	0.69
Pressao_Choke_k-7 (psi)	176113	float64	0.31	0.22	0.0	0.05	0.35	0.51	0.69
Tempo_poco_k (min)	169067	float64	0.33	0.18	0.0	0.18	0.33	0.48	0.69
Pressao_k (psi)	195323	float64	0.35	0.21	0.0	0.20	0.35	0.55	0.69
Vazão_k (m³/h)	149514	float64	0.05	0.08	0.0	0.00	0.00	0.07	0.69
Freq_Inversor_k (Hz)	6	float64	0.05	0.14	0.0	0.00	0.00	0.00	0.69
Abertura_choke_k (%)	198	float64	0.45	0.30	0.0	0.03	0.67	0.67	0.69
Vazão2_k (m³/h)	148807	float64	0.05	0.08	0.0	0.00	0.00	0.07	0.69
Abertura_Valvula_Reservatorio_k (%)	4	float64	0.33	0.29	0.0	0.00	0.24	0.69	0.69
Tempo_tanque_k (min)	387349	float64	0.33	0.18	0.0	0.18	0.33	0.48	0.69
Pressao_Tanque_k (psi)	353956	float64	0.43	0.19	0.0	0.23	0.43	0.61	0.69
Pressao_Choke_k (psi)	176071	float64	0.31	0.22	0.0	0.05	0.35	0.51	0.69
Pressão_Choke_k+1 (psi)	155655	float64	35.61	26.45	0.0	4.42	36.32	60.22	94.68

Figura *151* – Resumo do *dataframe* com dados experimentais, tratados com a função log e com 8 dados passados. Fonte: A autora.

Antes dos dados serem treinados pelo XGBoost, otimizam-se os hiperparâmetros com o Optuna gerando os seguintes parâmetros: 'n_estimators': 243, 'learning_rate': 0.2293, 'max_depth': 17, 'min_child_weight': 2, 'subsample': 0.7489, 'colsample_bytree': 0.5039, 'gamma': 0.1556, 'reg_alpha': 0.8050 e 'reg_lambda': 0.7125, em apenas 3 minutos e 27 segundos de execução com mesmo ambiente de execução dos modelos anteriores.

Com os melhores parâmetros gerados pelo Optuna, os dados são treinados, validados e testados pelo XGBoost. As métricas de avaliação do modelo são apresentadas nas Figuras 152 e 153.

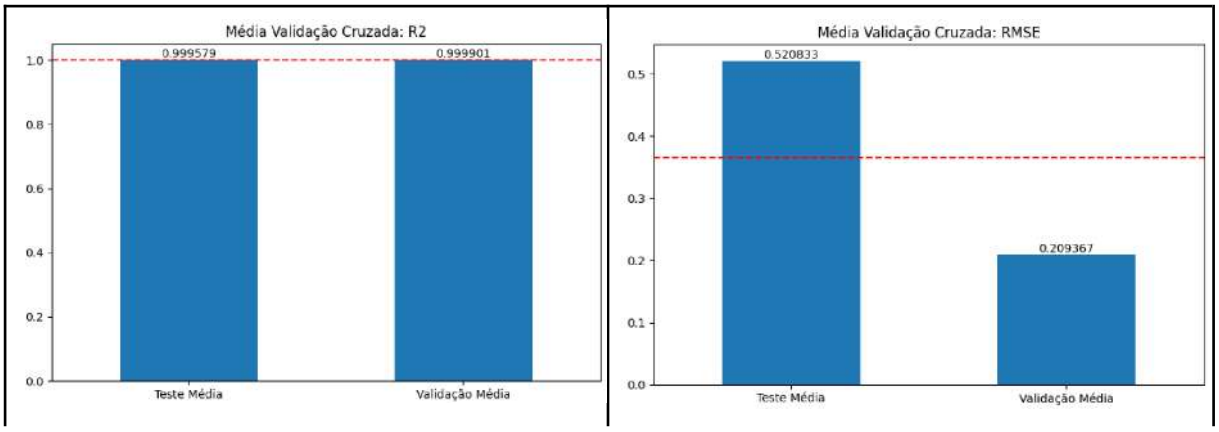

Figura *152* – Métricas de avaliação R^2 e RMSE com dados experimentais, tratados com a função log e com 8 dados passados. Fonte: A autora.

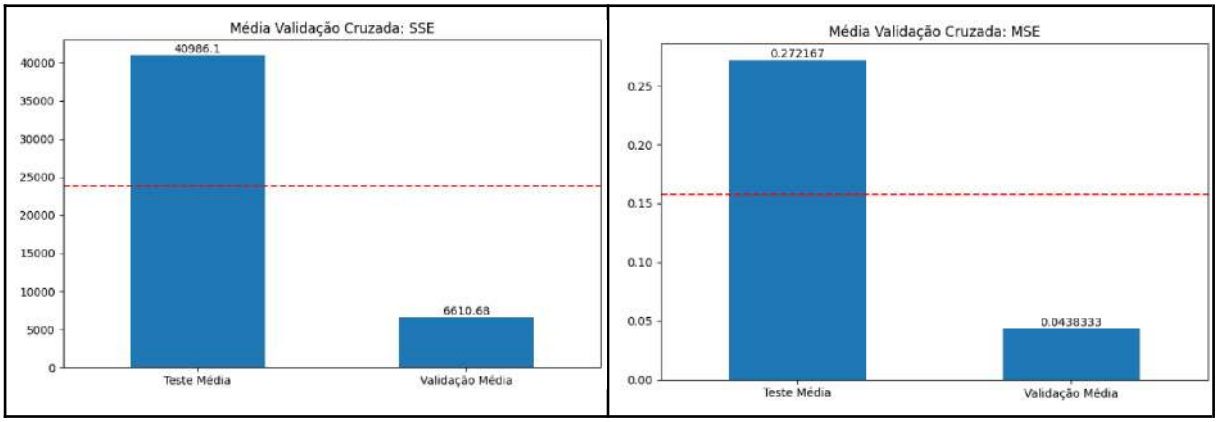

Figura *153* – Métricas de avaliação SSE e MSE com dados experimentais, tratados com a função log e com 8 dados passados. Fonte: A autora.

São apresentados na Figura 154 os valores da densidade dos dados para a pressão *choke*, sendo que quando valor predito é igual ao valor real, mais será observada a presença da cor roxa nos gráficos.

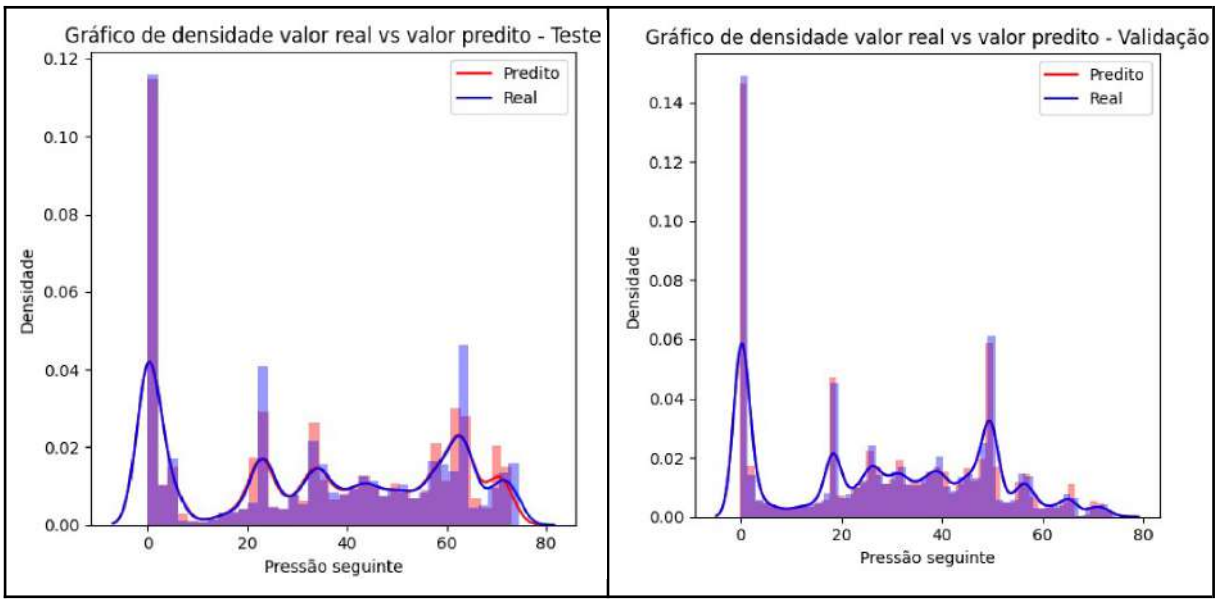

Figura *154* – Gráfico de densidade com dados experimentais, tratados com a função log e com 8 dados passados. Fonte: A autora.

São apresentados na Figura 155 os valores da distribuição dos resíduos no teste e na validação do modelo. Verifica-se que a maior parte dos erros obtidos estão próximos de 0.

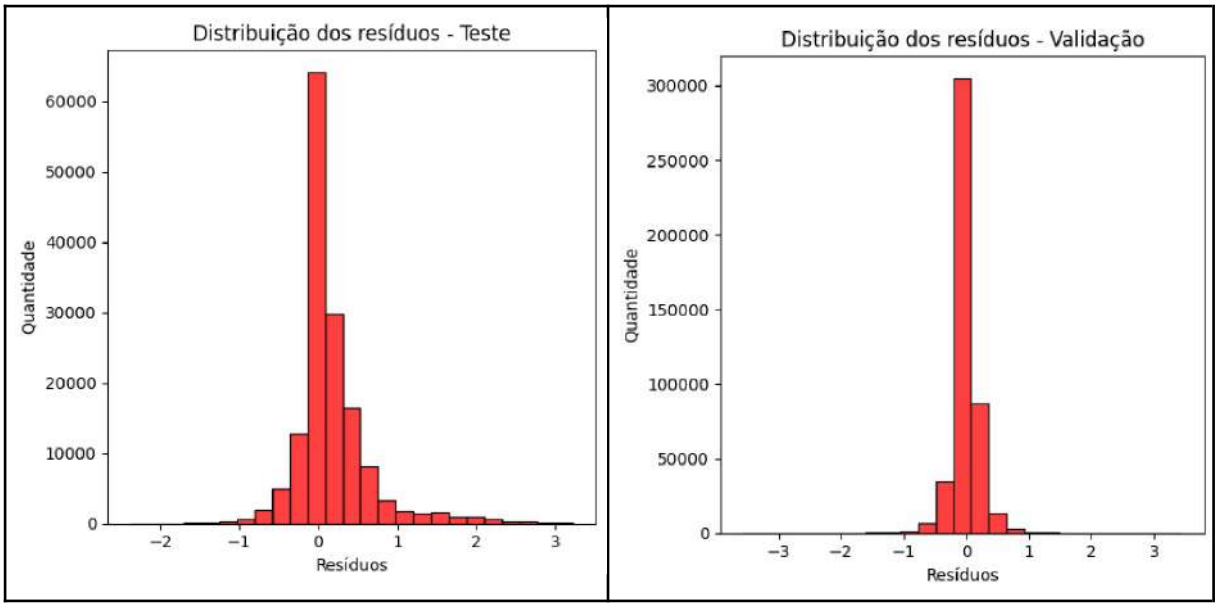

Figura *155* – Distribuição dos resíduos com dados experimentais, tratados com a função log e com 8 dados passados. Fonte: A autora.

A Figura 156 apresenta os pontos na curva de comparação de valor real e valor predito.

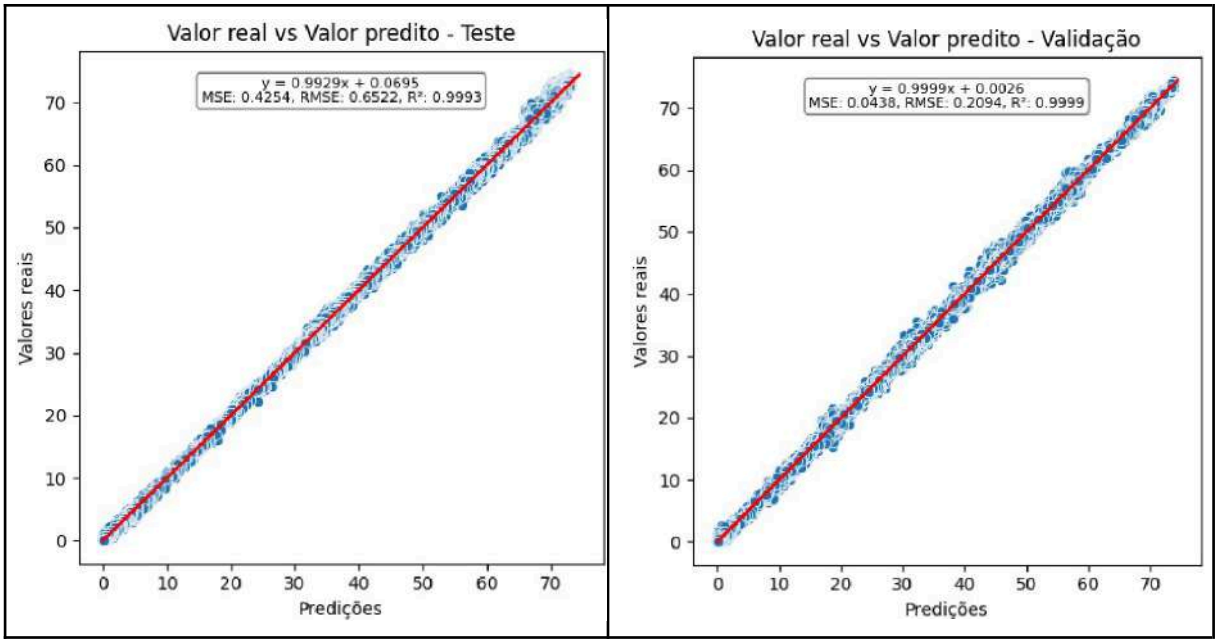

Figura *156* – Evolução do modelo com dados experimentais, tratados com a função log e com 8 dados passados. Fonte: A autora.

O gráfico da função de perda é apresentado na Figura 157. Os resultados do modelo são satisfatórios por se aproximarem de 0 ao longo das épocas.

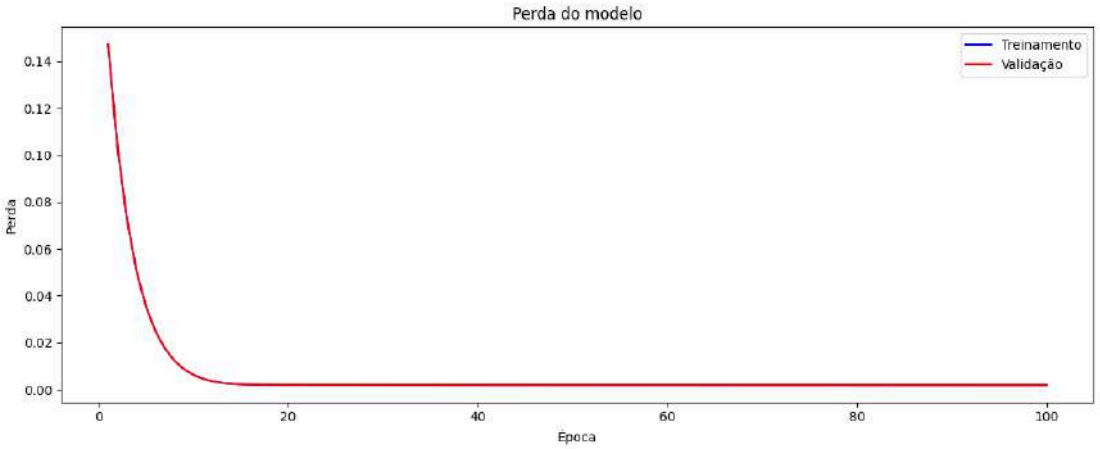

Figura *157* – Curva de perdas com dados experimentais, tratados com a função log e com 8 dados passados. Fonte: A autora.

O gráfico na Figura 158 representa a importância de cada variável para as previsões, sendo que as pressões na *choke* defasadas no tempo são as mais relevantes variáveis para a síntese do modelo.

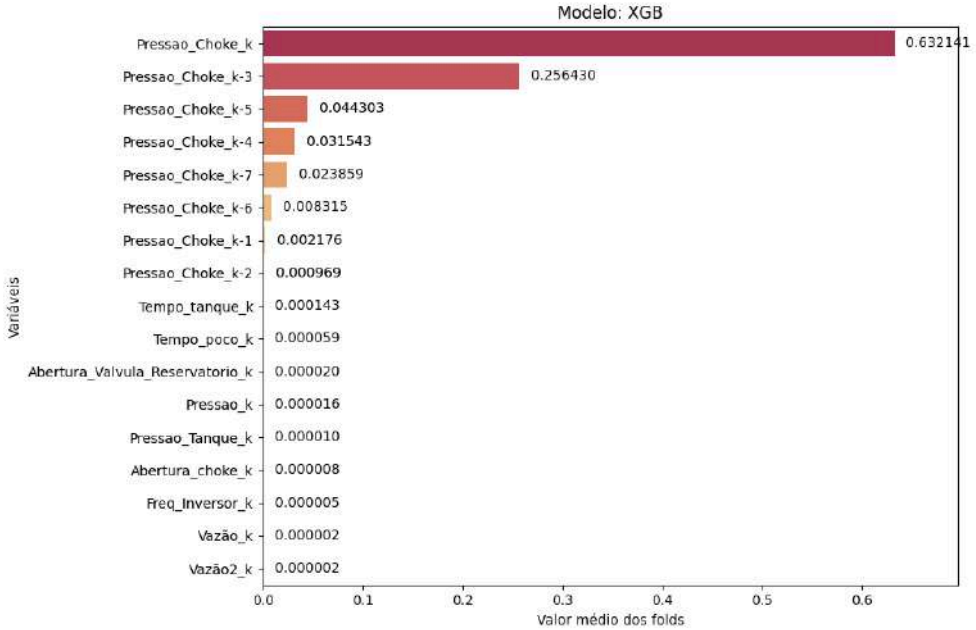

Figura *158* – Importância das variáveis com dados experimentais, tratados com a função log e com 8 dados passados. Fonte: A autora.

Os resultados das previsões e realidade dos testes do modelo são apresentados nas Figuras 159 a 162, indicando o quanto o modelo consegue prever a operação de PMCD.

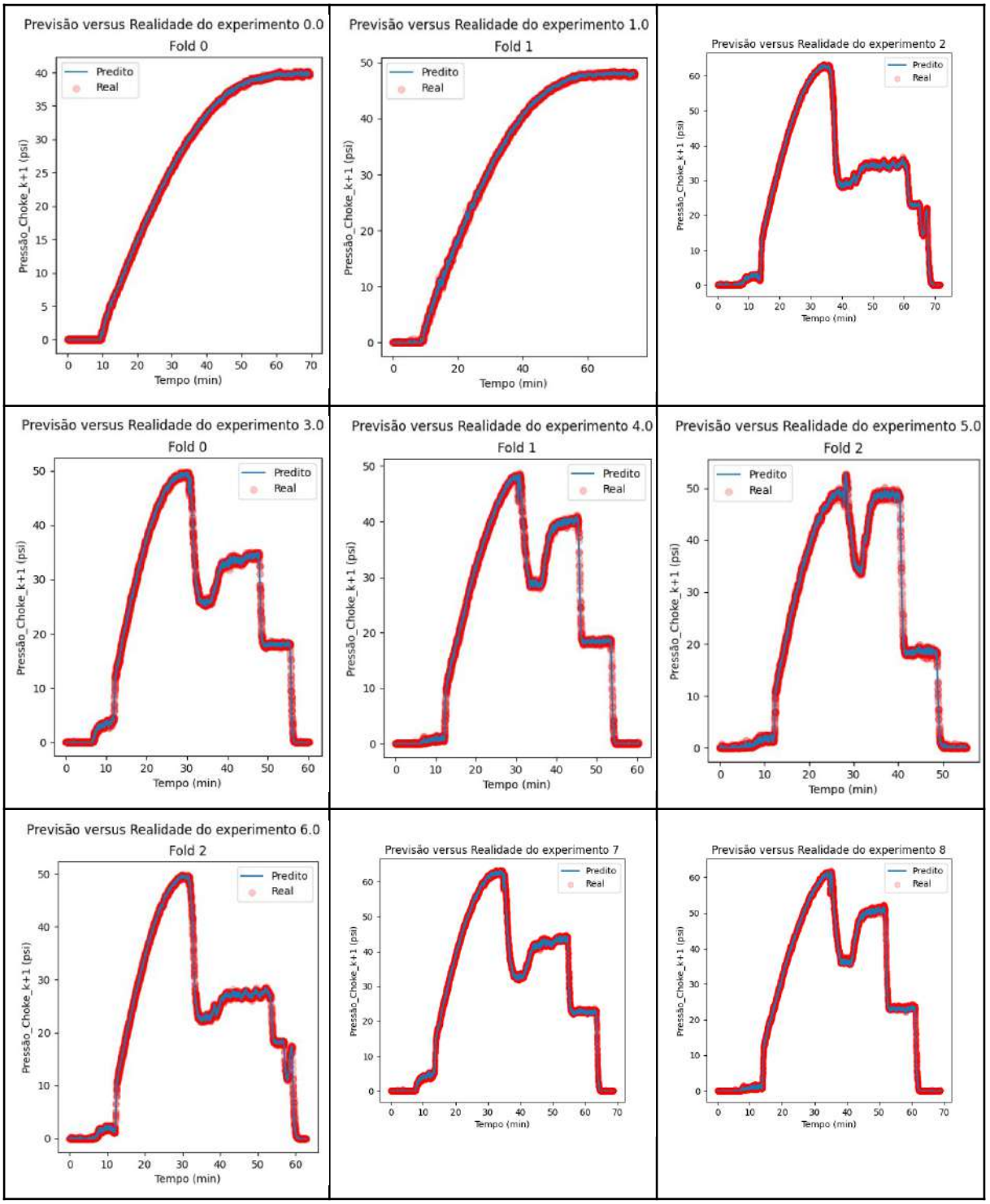

Figura *159* – Previsão e realidade dos experimentos de 0 a 8 com log e com 8 dados passados. Fonte: A autora.

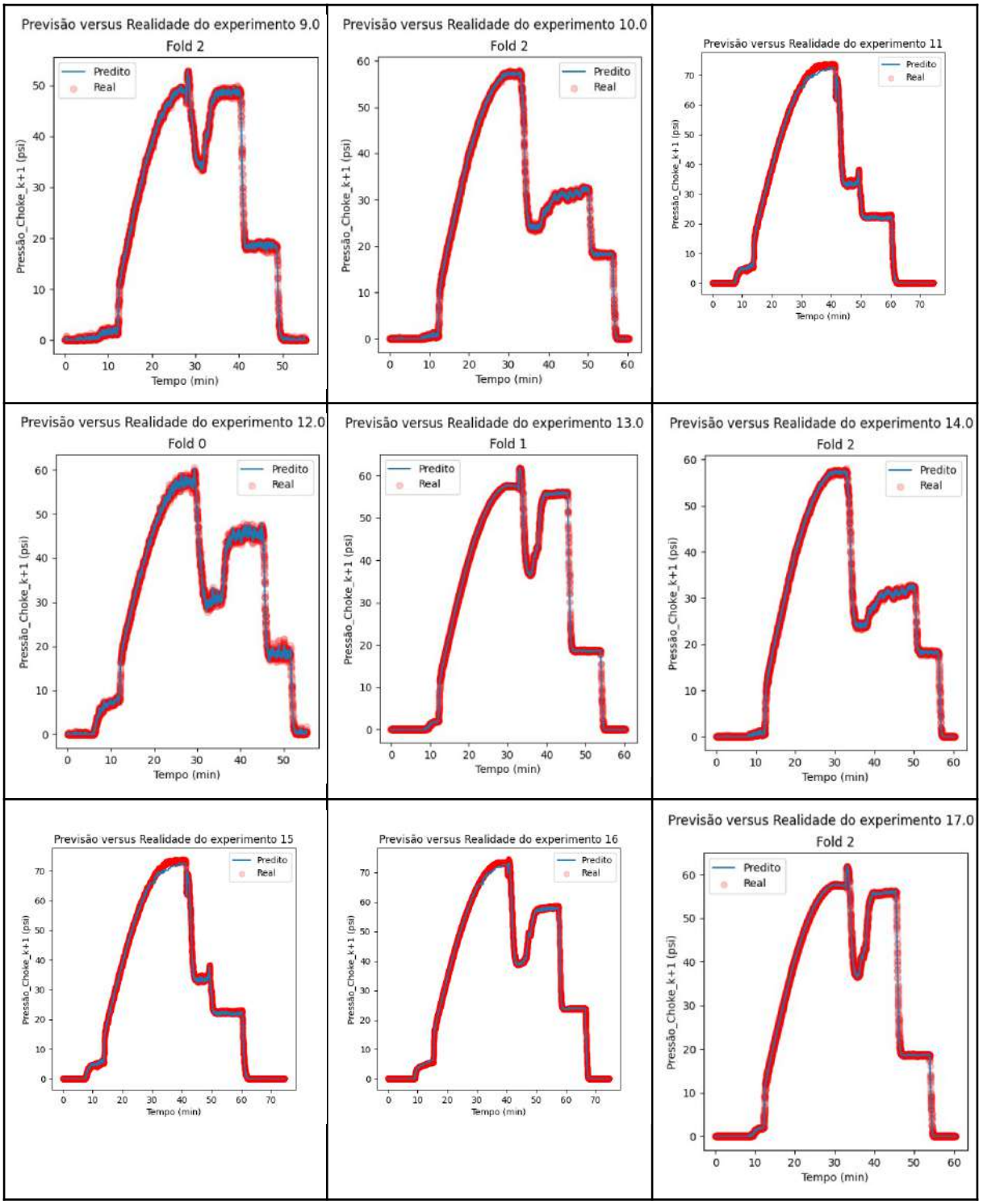

Figura *160* – Previsão e realidade dos experimentos de 9 ao 17 com log e com 8 dados passados. Fonte: A autora.

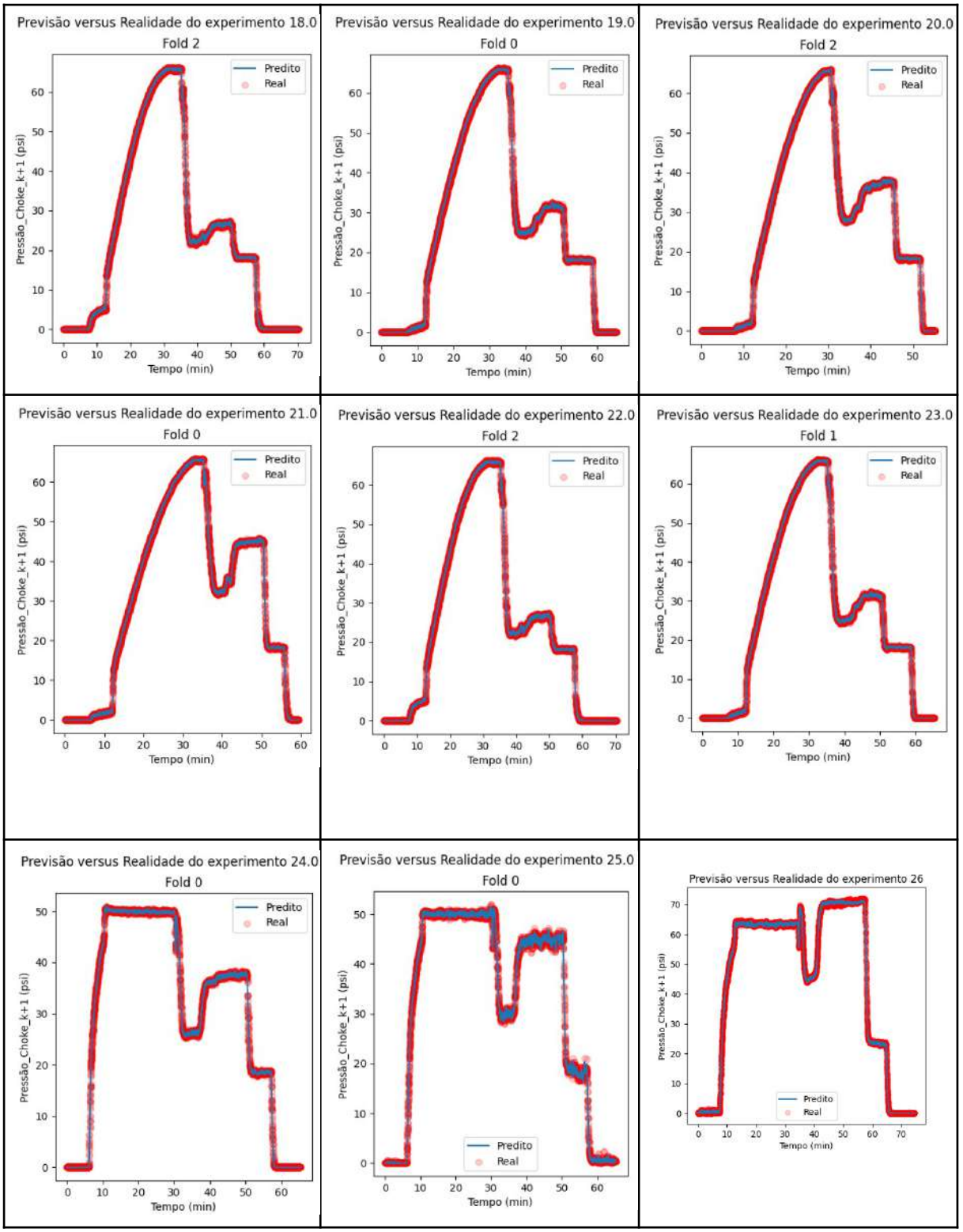

Figura *161* – Previsão e realidade dos experimentos de 18 ao 26 com log e com 8 dados passados. Fonte: A autora.

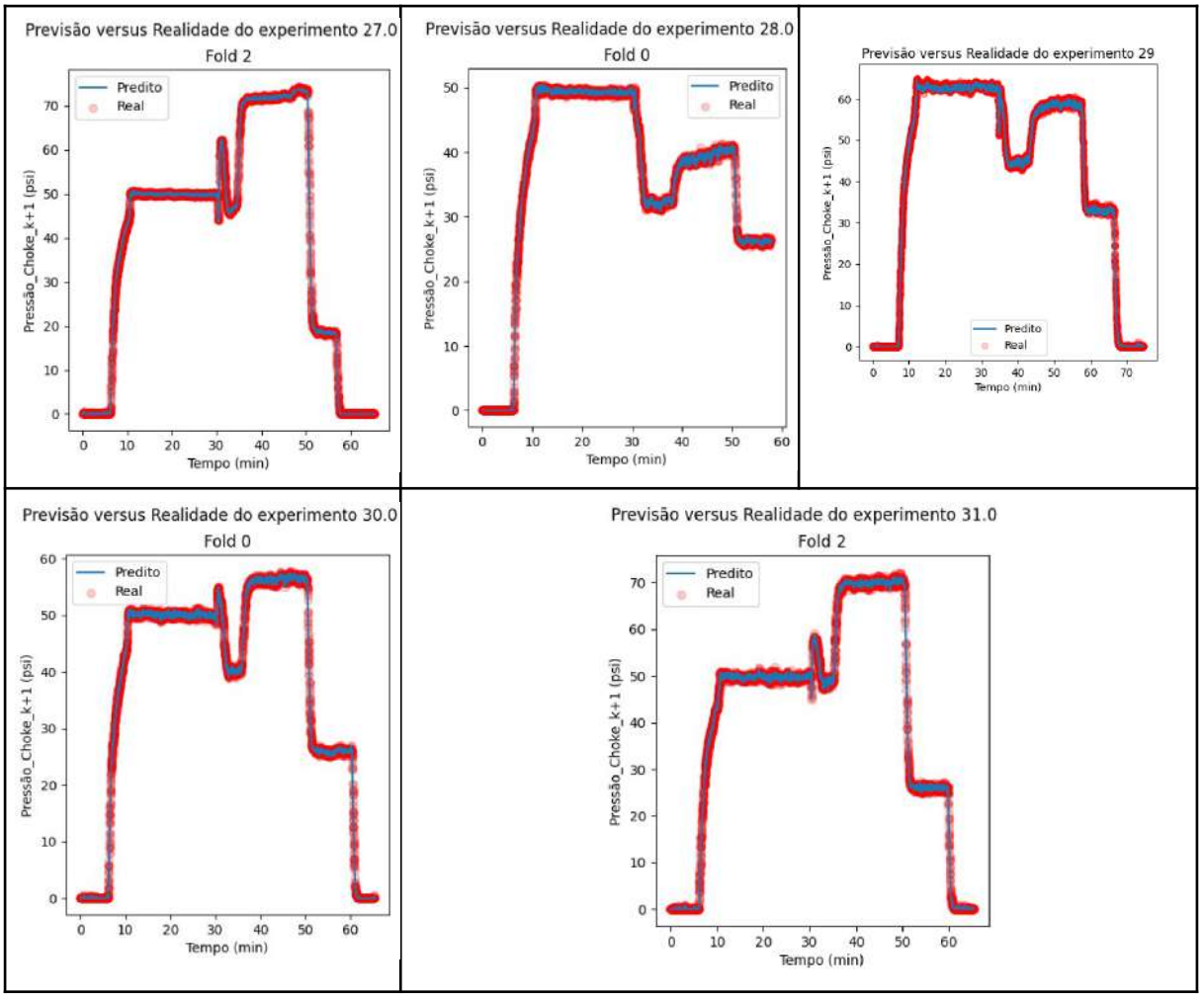

Figura *162* – Previsão e realidade dos experimentos de 27 ao 31 com log e com 8 dados passados. Fonte: A autora.

Com relação à arquitetura do modelo, foram treinadas 243 árvores. O uso de 8 dados passados permitiu a construção do modelo com as melhores métricas de avaliação.

ANEXO G – RESULTADOS DOS DADOS EXPERIMENTAIS TRANSFORMADOS COM LOG, COM 8 DADOS PASSADOS E COM AS ESCALAS DE 0 A 1, DE 0 A 4 E DE 0 A 10

Com base no resultado do modelo tratado com a função log e com 8 dados passados, que apresentou o melhor aprendizado, foram desenvolvidos outros modelos XGBoost, com o objetivo de avaliar se a mudança da padronização da escala interfere na aprendizagem do modelo (Tabelas 7 e 8).

Tabela 7 – Métricas de avaliação da validação dos modelos com 8 dados passados e tratados com a função log em diferentes escalas. Fonte: A autora.

Métricas/Validação	Escala (0, 1)	Escala (0, 4)	Escala (0, 10)
R^2	0.9999	0.9999	0.9999
RMSE	0.2094	0.2103	0.2120
MSE	0.0438	0.0442	0.0450

Tabela 8 – Métricas de avaliação do teste dos modelos com 8 dados passados e tratados com a função log em diferentes escalas. Fonte: A autora.

Métricas/Teste	Escala (0, 1)	Escala (0, 4)	Escala (0, 10)
R^2	0.9993	0.9995	0.9996
RMSE	0.6522	0.5783	0.5300
MSE	0.4254	0.3344	0.2808

Por fim, são mostrados na Figura 163 o resultado da previsão versus realidade do teste com o experimento 11, sendo a imagem mais à esquerda com escala de 0 a 1, do meio com escala de 0 a 4 e à direita com escala de 0 a 10, onde todos os resultados apresentaram desempenho satisfatório.

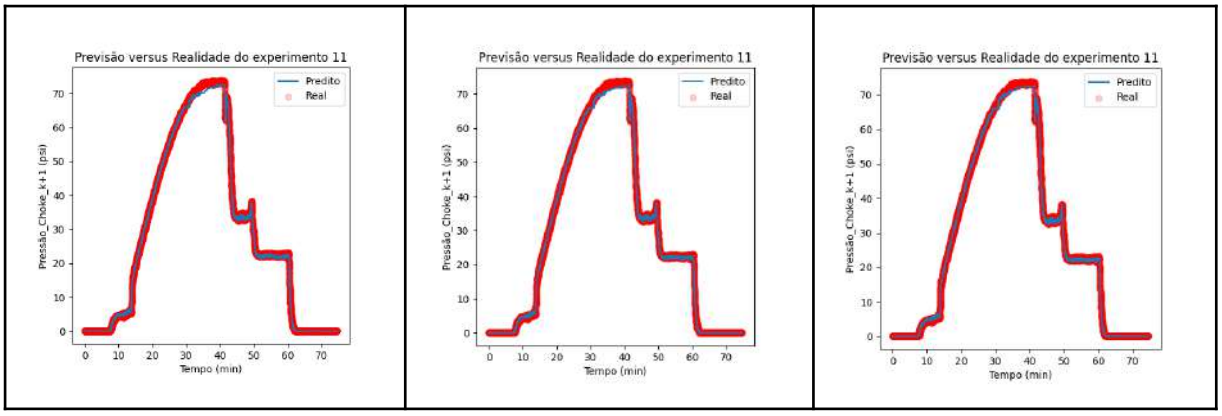

Figura *163* – Previsão versus realidade com dados experimentais, com 8 dados passados e tratados com a função log em diferentes escalas. Fonte: A autora.

O uso de escalas diferentes não alterou significativamente a qualidade dos modelos.

As Tabelas 9 a 11 ilustram os tempos computacionais requeridos para a construção dos modelos baseados em *machine learning* a partir dos dados experimentais de Carvalho (2018).

Tabela 9 – Tempos computacionais dos modelos com dados experimentais com log para execução do Optuna. Fonte: A autora.

Modelos com dados experimentais com log			
Tempo Total Optuna	**Escala (0, 1)**	**Escala (0, 4)**	**Escala (0, 10)**
Sem dados passados	115 s	162 s	135 s
Com 2 dados passados	153 s	175 s	247 s
Com 8 dados passados	207 s	294 s	328 s

Tabela 10 – Tempos computacionais dos modelos com dados experimentais com log para execução do XGBoost. Fonte: A autora.

Modelos com dados experimentais com log			
Tempo Total XGBoost	**Escala (0, 1)**	**Escala (0, 4)**	**Escala (0, 10)**
Sem dados passados	40.2 s	41.1 s	41.2 s
Com 2 dados passados	46.4 s	42.3 s	43.7 s
Com 8 dados passados	56.9 s	53.1 s	53.6 s

Tabela 11 – Tempos computacionais dos modelos com dados experimentais com log para execução do Optuna e o XGBoost. Fonte: A autora.

Modelos com dados experimentais com log			
Tempo Total Optuna e XGBoost	**Escala (0, 1)**	**Escala (0, 4)**	**Escala (0, 10)**
Sem dados passados	155.2 s	203.1 s	176.2 s
Com 2 dados passados	199.4 s	217.3 s	290.7 s
Com 8 dados passados	263.9 s	347.1 s	381.6 s

ANEXO H – RESULTADOS DOS DADOS DE POÇOS REAIS TRANSFORMADOS COM LOG, SEM DADOS PASSADOS, COM ESCALA DE 0 A 10

Como não foram utilizados dados passados, não houve necessidade de criação de novas variáveis (Figura 164). A variável vazão apresentou *outiliers,* que foram tratados com a substituição pela sua média.

	Valores_Únicos	Tipo_de_Dado	mean	std	min	25%	50%	75%	max
Tempo (min)	3095	float64	1.25	0.65	0.0	0.65	1.39	1.76	2.40
Vazão (m³/h)	473	float64	0.12	0.50	0.0	0.00	0.00	0.00	2.40
Pressao_Choke (psi)	2315	float64	418.66	545.39	0.0	89.47	181.27	403.80	2001.09
experimento	5	int64	3.03	1.33	1.0	2.00	3.00	4.00	5.00

Figura *164* – Resumo do *dataframe* com dados de poços reais, tratados com a função log e sem dados passados. Fonte: A autora.

Antes dos dados serem treinados pelo XGBoost, é feito a otimização dos hiperparâmetros com o Optuna gerando os seguintes parâmetros: 'n_estimators': 939, 'learning_rate': 0.0270, 'max_depth': 15, 'min_child_weight': 8, 'subsample': 0.5437, 'colsample_bytree': 0.8194, 'gamma': 0.3454, 'reg_alpha': 0.9377 e 'reg_lambda': 0.4069, em apenas 3.62 segundos de execução.

Com os melhores parâmetros gerados pelo Optuna, os dados são treinados, validados e testados pelo XGBoost, gerando as métricas de avaliação do modelo (Figura 165).

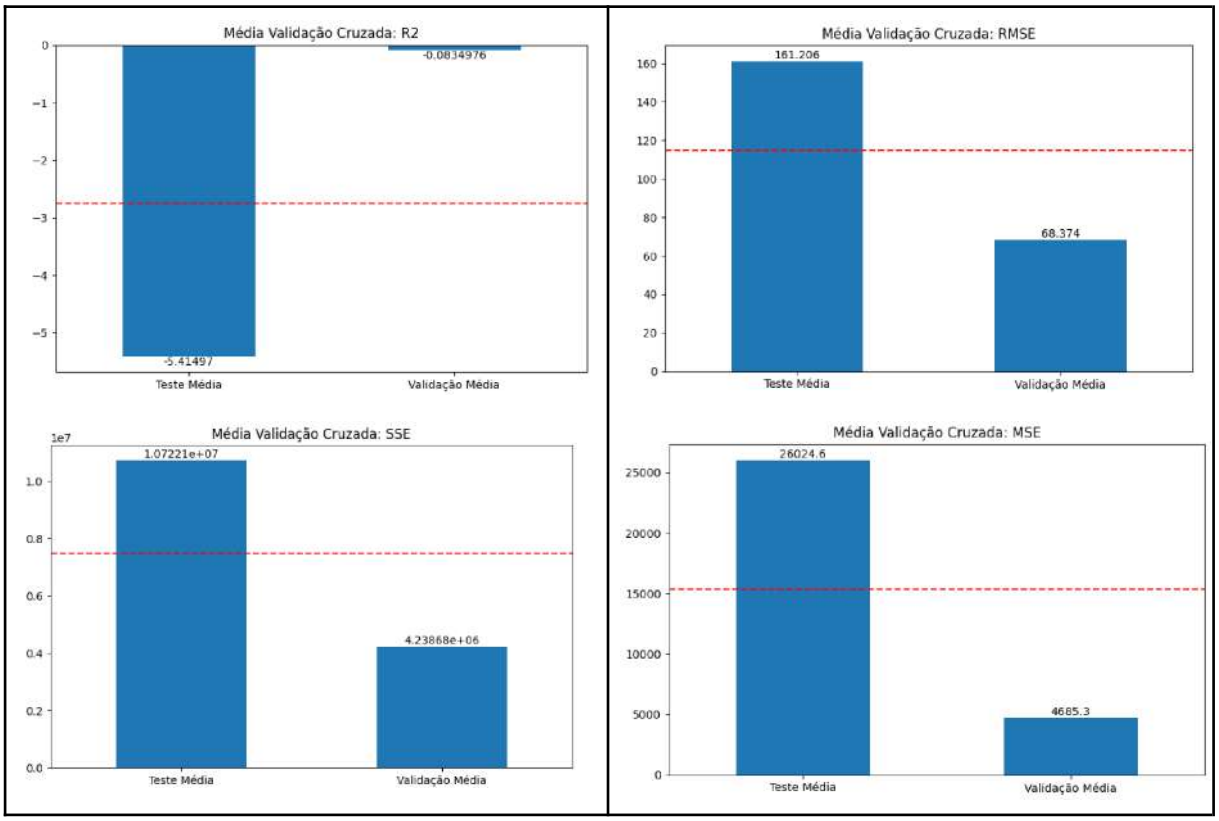

Figura *165* – Métricas de avaliação com dados de poços reais, tratados com a função log e sem aplicação de dados passados. Fonte: A autora.

São apresentados na Figura 166 os valores da densidade dos dados para a pressão *choke*, sendo que quanto mais valores preditos de forma correta, ou seja, quando valor predito é igual ao valor real, mais será observada a presença da cor roxa nos gráficos.

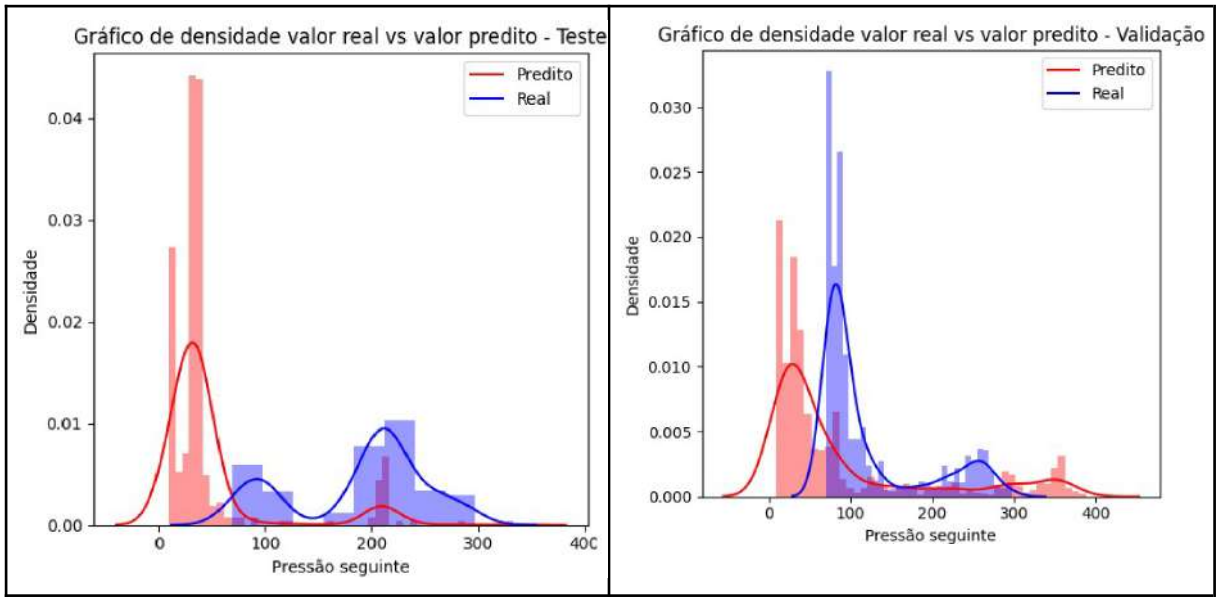

Figura *166* – Gráficos de densidade com dados de poços reais, tratados com a função log e sem dados passados. Fonte: A autora.

São apresentados na Figura 167 os valores da distribuição dos resíduos no teste e na validação do modelo.

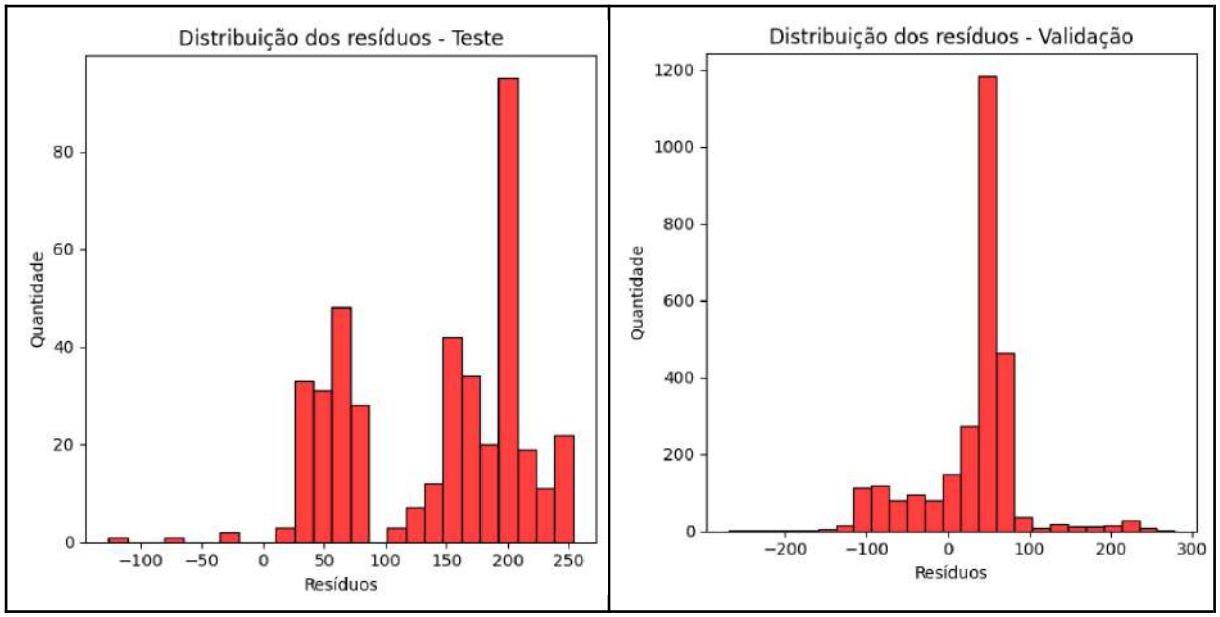

Figura *167* – Distribuição dos resíduos com dados de poços reais, tratados com a função log e sem dados passados. Fonte: A autora.

A Figura 168 apresenta os pontos na curva de comparação de valor real e valor predito.

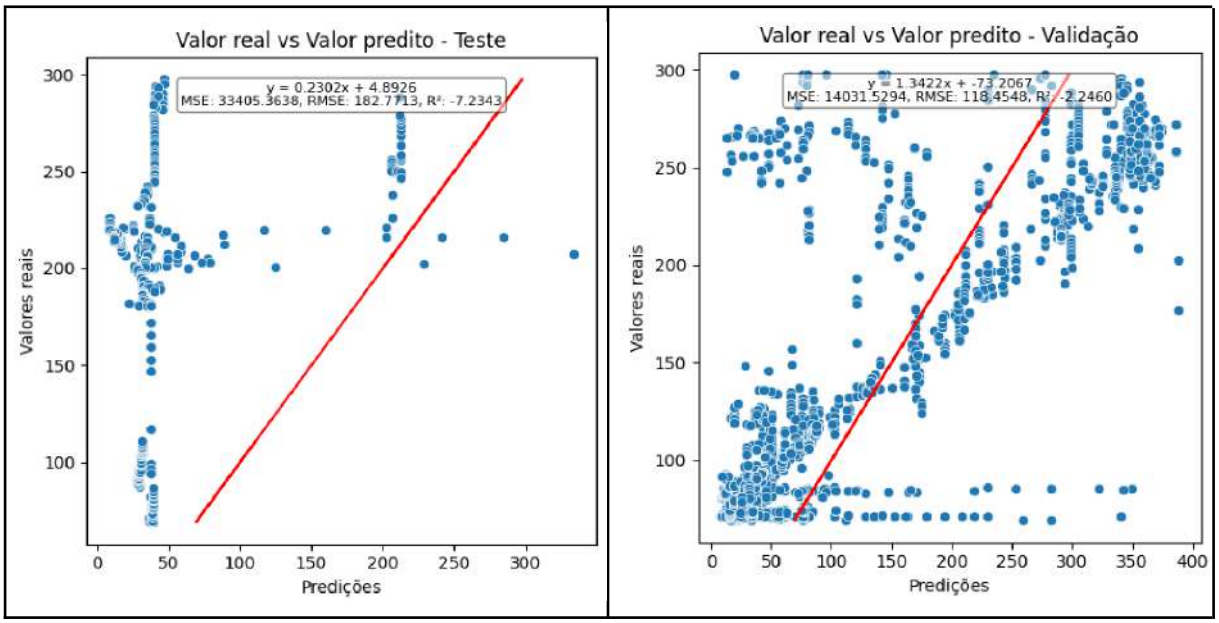

Figura *168* – Gráfico de evolução do modelo com dados de poços reais, tratados com a função log e sem dados passados. Fonte: A autora.

O gráfico da função de perda é apresentado nas Figuras 169.

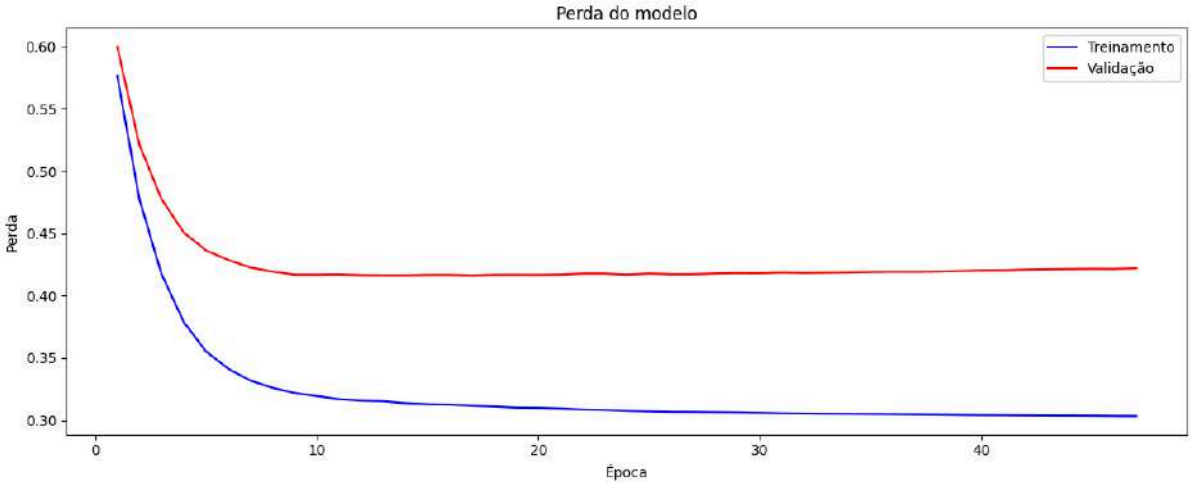

Figura *169* – Curva de perdas função log e sem dados passados. Fonte: A autora.

O gráfico na Figura 170 representa a importância de cada variável para as previsões, sendo que a vazão apresenta a maior importância para o processo.

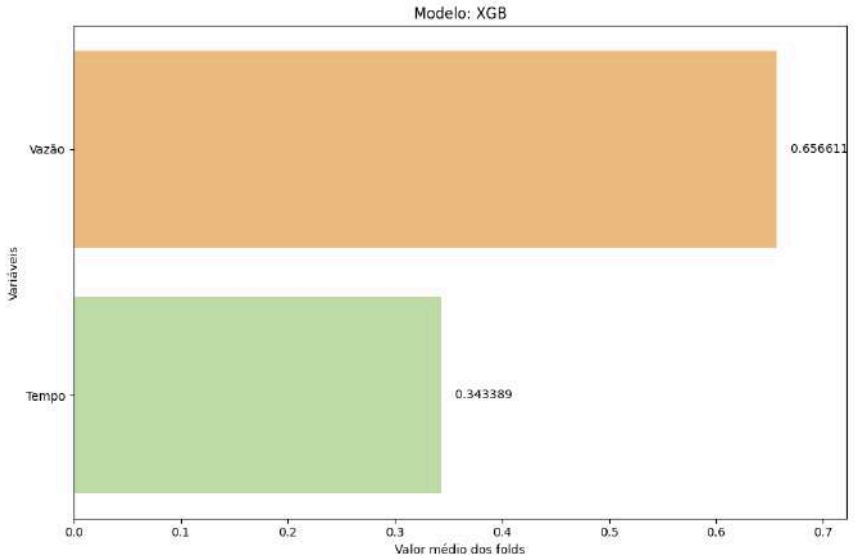

Figura *170* – Importância das variáveis com dados de poços reais, tratados com a função log e sem dados passados. Fonte: A autora.

Os resultados das previsões e realidade dos testes do modelo são apresentados nas Figuras 171 a 173, indicando que o modelo não consegue prever a operação de PMCD.

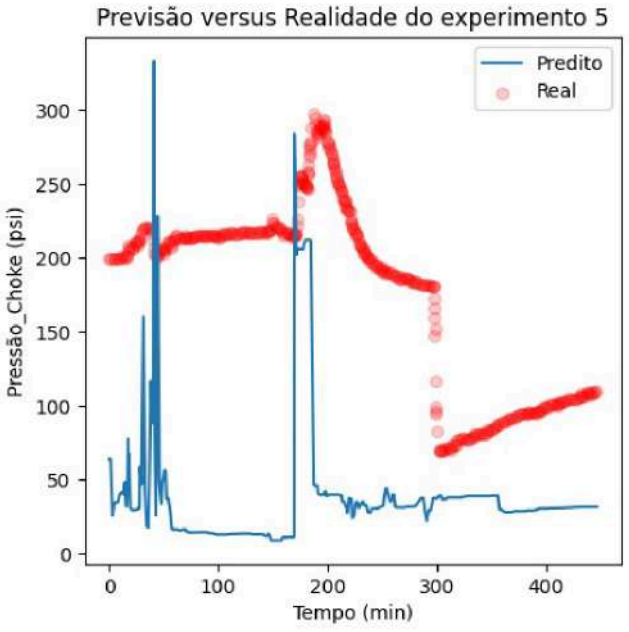

Figura *171* – Previsão versus realidade do teste com dados de poços reais, tratados com a função log e sem dados passados, para o experimento 5. Fonte: A autora.

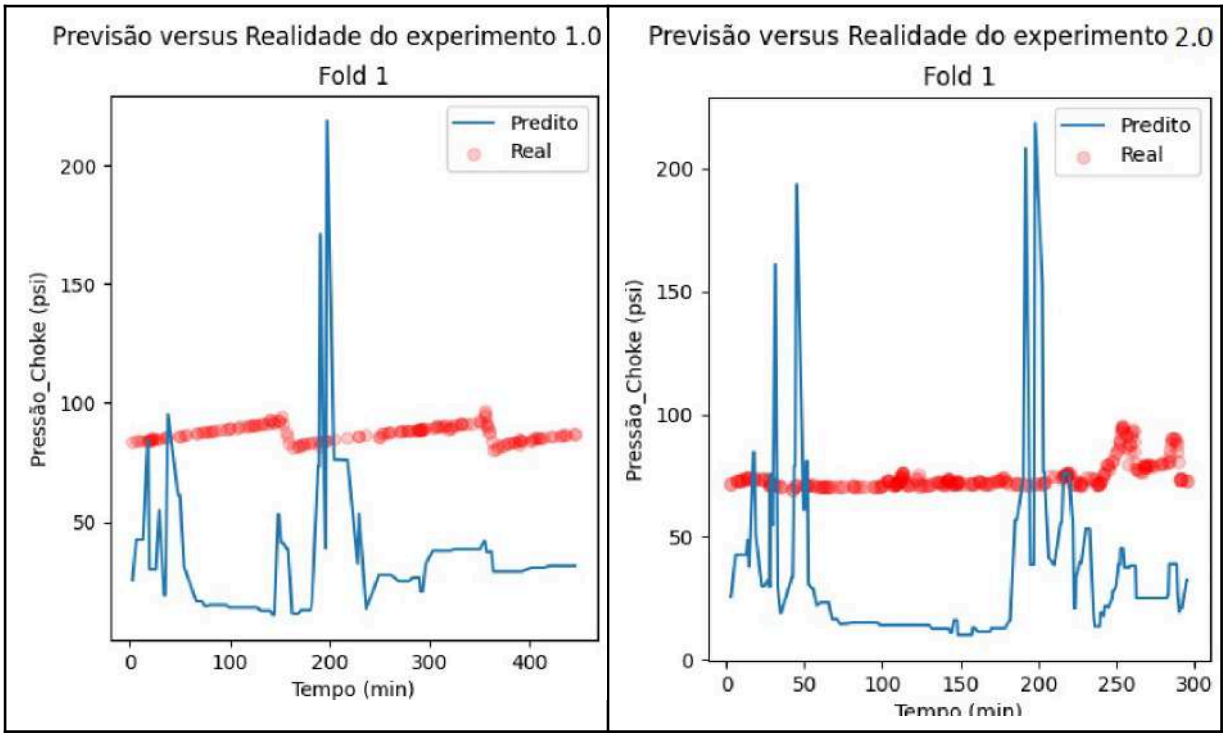

Figura *172* – Previsão versus realidade dos testes com dados de poços reais, tratados com a função log e sem dados passados, para os experimentos 1 e 2. Fonte: A autora.

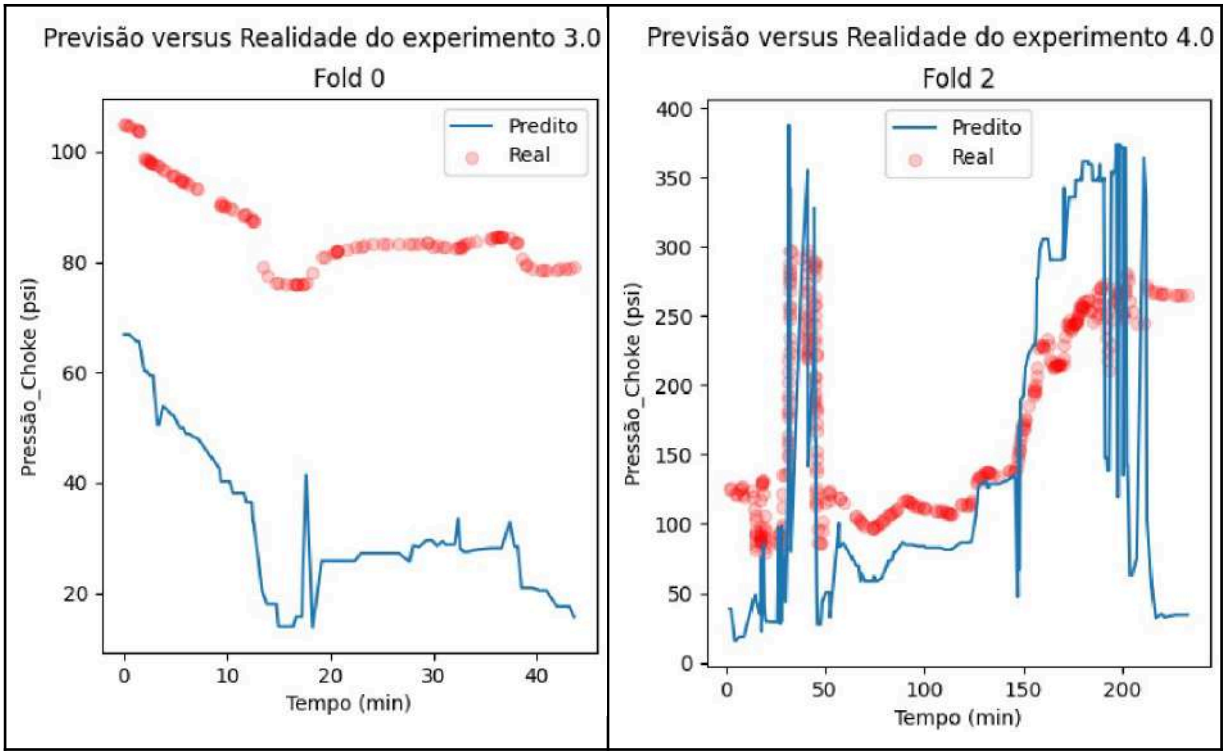

Figura *173* – Previsão versus realidade dos testes com dados de poços reais, tratados com a função log e sem dados passados, para os experimentos 3 e 4. Fonte: A autora.

Com relação à arquitetura do modelo, foram treinadas 939 árvores (Figura 174). Cada nó contém a decisão de uma variável e as folhas representam a pontuação que será somada a cada amostra para gerar sua previsão ao final. Vale ressaltar que em cada nó os valores das variáveis variam de 0 a 10 por conta da normalização que é realizada antes do treinamento do modelo.

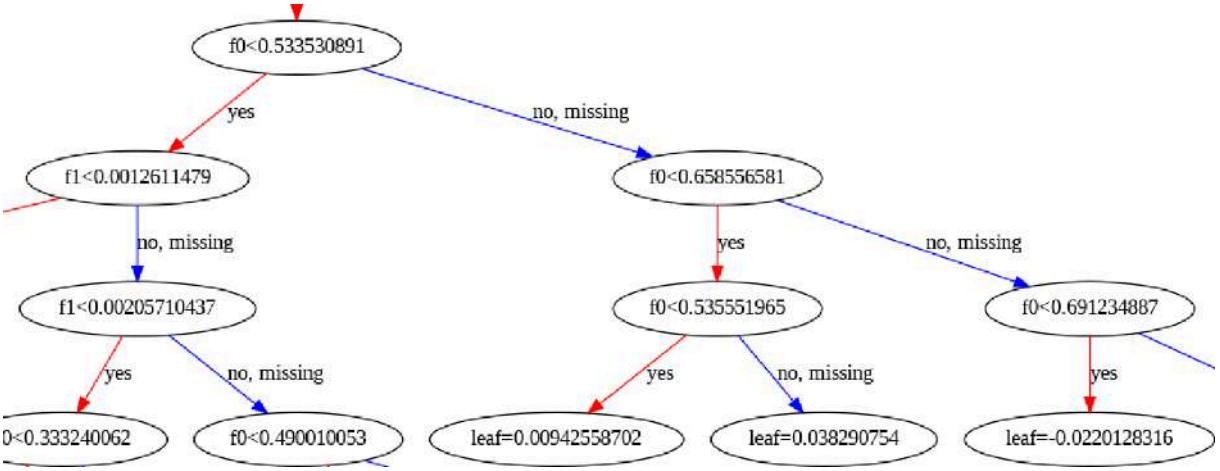

Figura *174* – Parte da árvore gerada pelo modelo com dados de poços reais. Fonte: A autora.

ANEXO I – RESULTADOS DOS DADOS DE POÇOS REAIS TRANSFORMADOS COM LOG, COM 20 DADOS PASSADOS, COM 2 DADOS PASSADOS

Com 2 dados passados, todas as variáveis são deslocadas, com isso, foram criadas as variáveis para a pressão *choke* em cada passo anterior, conforme mostrado na Figura 175. A variável vazão apresentou *outiliers,* que foram tratados com a substituição pela sua média.

	Valores_Únicos	Tipo_de_Dado	mean	std	min	25%	50%	75%	max
Pressao_Choke_k-1 (psi)	2308	float64	0.98	0.77	0.0	0.38	0.67	1.86	2.40
Tempo_k (min)	3089	float64	1.25	0.65	0.0	0.66	1.40	1.76	2.40
Vazão_k (m³/h)	472	float64	0.12	0.50	0.0	0.00	0.00	0.00	2.40
Pressao_Choke_k (psi)	2309	float64	0.98	0.77	0.0	0.38	0.67	1.86	2.40
Pressão_Choke_k+1 (psi)	2311	float64	419.06	545.76	0.0	89.57	181.46	403.59	2001.09
experimento	5	int64	3.03	1.32	1.0	2.00	3.00	4.00	5.00

Figura *175* – Resumo do *dataframe* com dados de poços reais, tratados com a função log e com 2 dados passados. Fonte: A autora.

Antes dos dados serem treinados pelo XGBoost, é feito a otimização dos hiperparâmetros com o Optuna gerando os seguintes parâmetros: 'n_estimators': 746, 'learning_rate': 0.2273, 'max_depth': 14, 'min_child_weight': 10, 'subsample': 0.8732, 'colsample_bytree': 0.8348, 'gamma': 0.02032, 'reg_alpha': 0.1869 e 'reg_lambda': 0.7423, em 2.79 segundos de execução.

Com os melhores parâmetros gerados pelo Optuna, os dados são treinados, validados e testados pelo XGBoost, gerando as métricas de avaliação das Figuras 176 e 177.

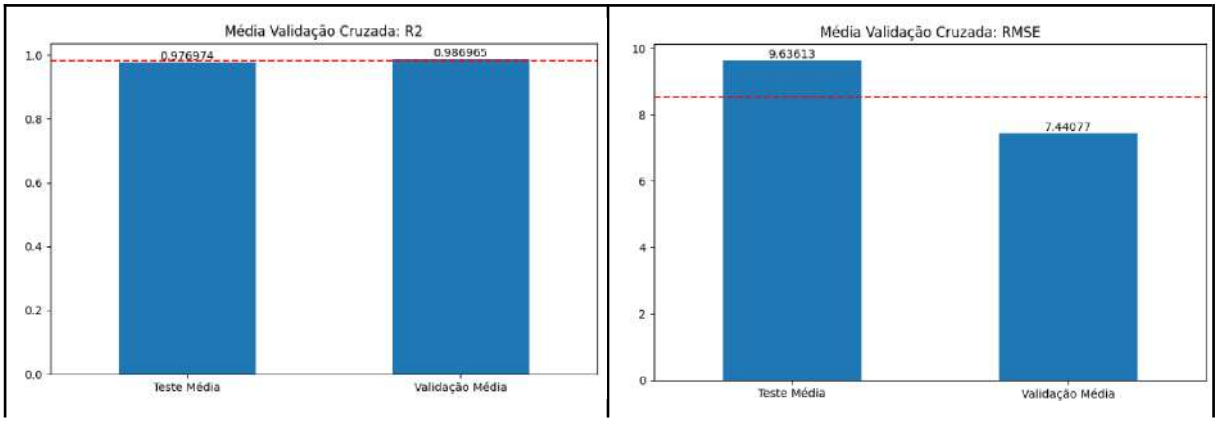

Figura *176* – Métricas de avaliação R² e RMSE com dados de poços reais, tratados com a função log e com 2 dados passados. Fonte: A autora.

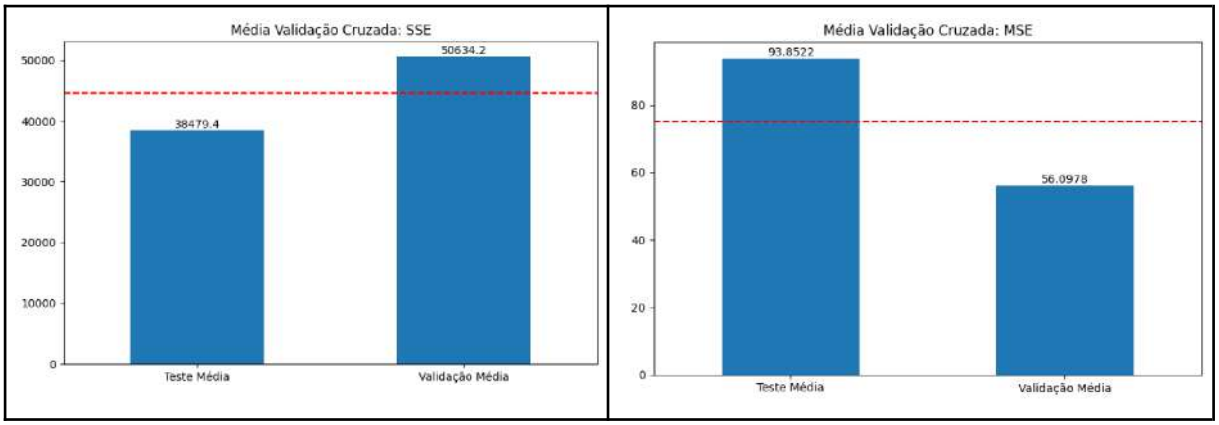

Figura *177* – Métricas de avaliação SSE e MSE com dados de poços reais, tratados com a função log e com 2 dados passados. Fonte: A autora.

São apresentados na Figura 178 os valores da densidade dos dados para a pressão na *choke*, sendo que quanto mais valores preditos de forma correta, ou seja, quando valor predito é igual ao valor real, mais será observada a presença da cor roxa nos gráficos.

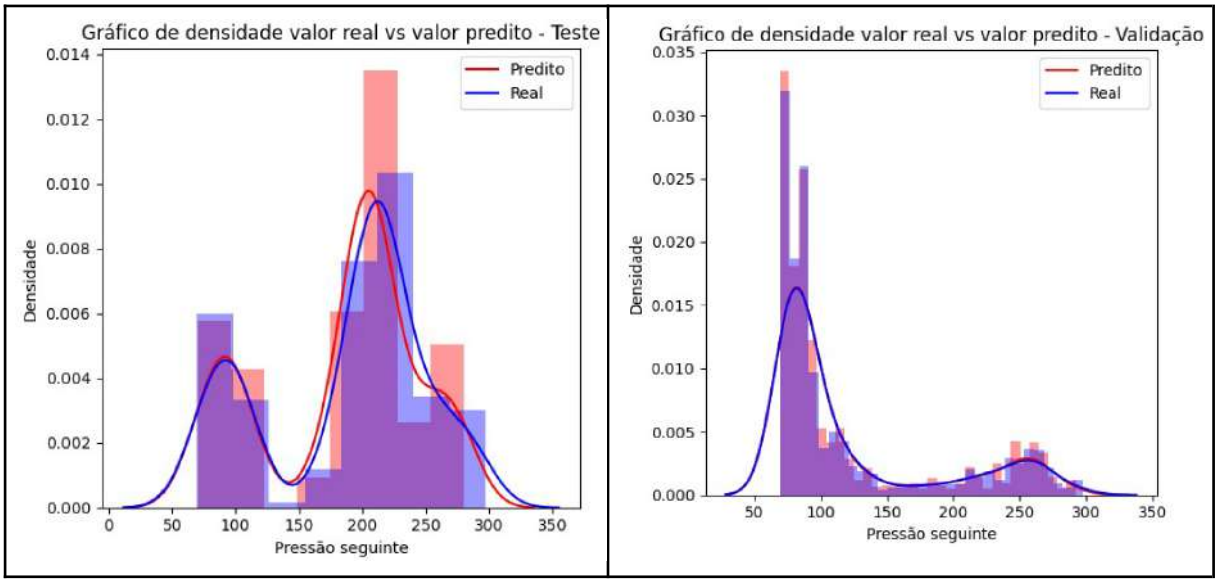

Figura *178* – Gráficos de densidade com dados de poços reais, tratados com a função log e com 2 dados passados. Fonte: A autora.

São apresentados na Figura 179 os valores da distribuição dos resíduos no teste e na validação do modelo.

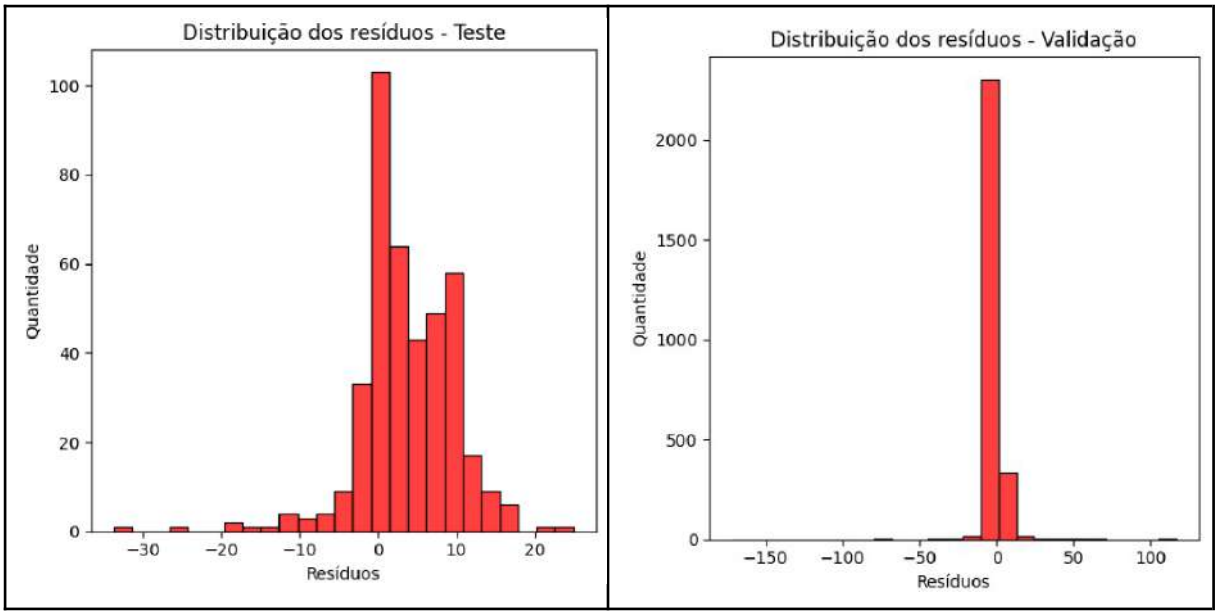

Figura *179* – Distribuição dos resíduos com dados de poços reais, tratados com a função log e com 2 dados passados. Fonte: A autora.

A Figura 180 apresenta os pontos na curva de comparação de valor real e valor predito.

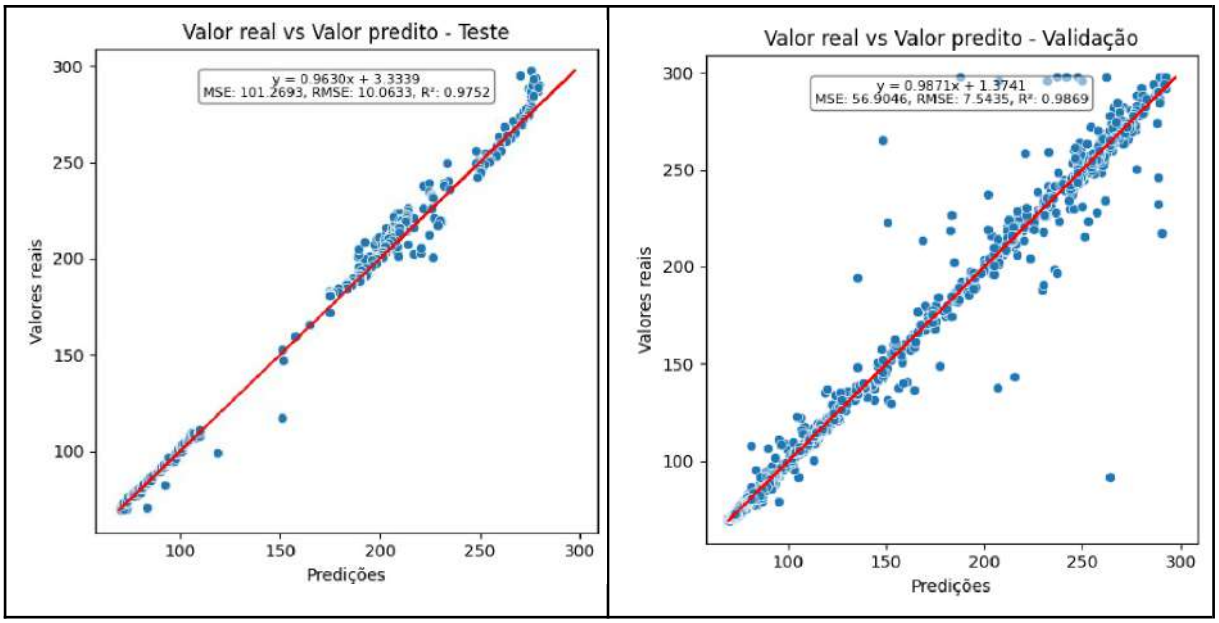

Figura *180* – Gráfico da evolução do modelo com dados de poços reais, tratados com a função log e com 2 dados passados. Fonte: A autora.

O gráfico da função de perda é apresentado na Figura 181.

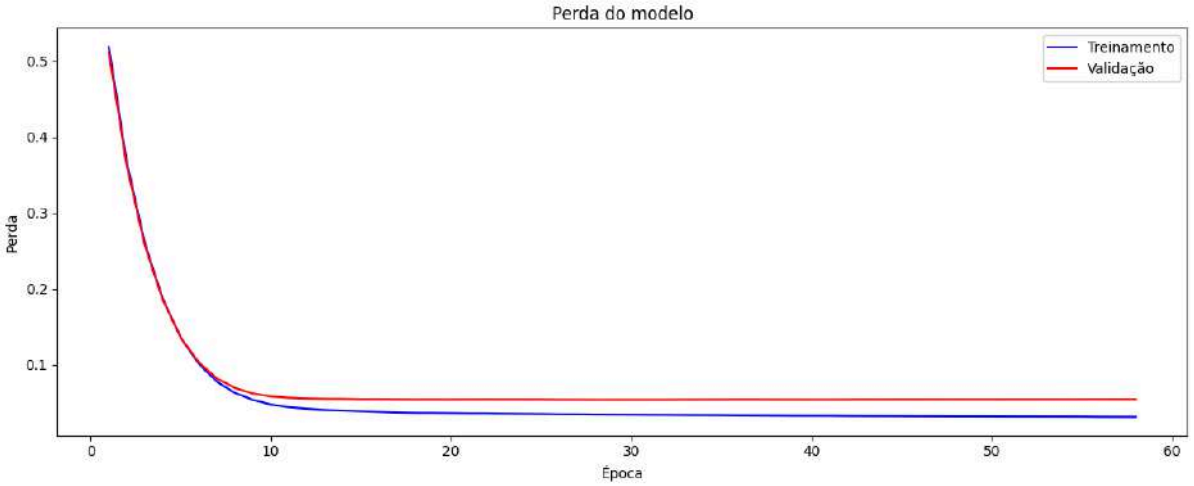

Figura *181* – Curva de perdas função log e com 2 dados passados. Fonte: A autora.

O gráfico na Figura 182 representa a importância de cada variável para as previsões, sendo que a pressão da *choke* foi a variável mais relevante.

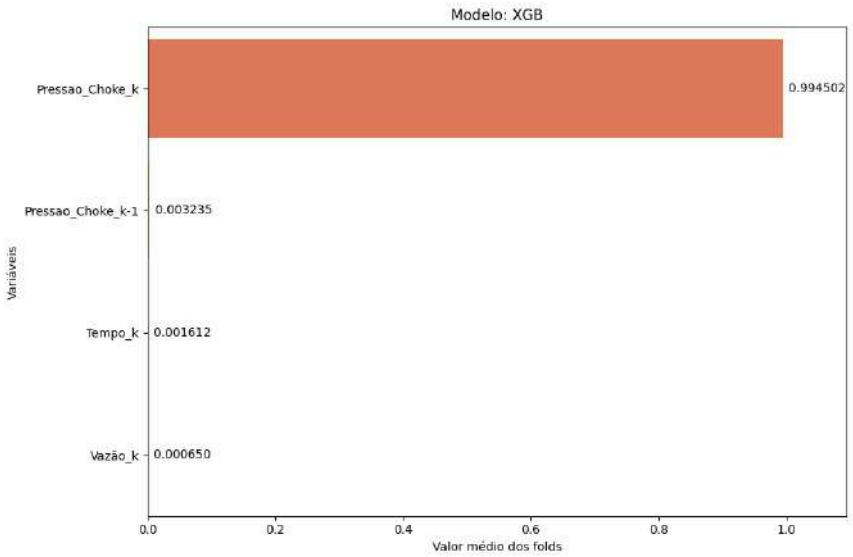

Figura *182* – Importância das variáveis com dados de poços reais, tratados com a função log e com 2 dados passados. Fonte: A autora.

Os resultados das previsões e realidade dos testes do modelo são apresentados nas Figuras 183 a 185, representando o quanto o modelo consegue prever a operação de PMCD.

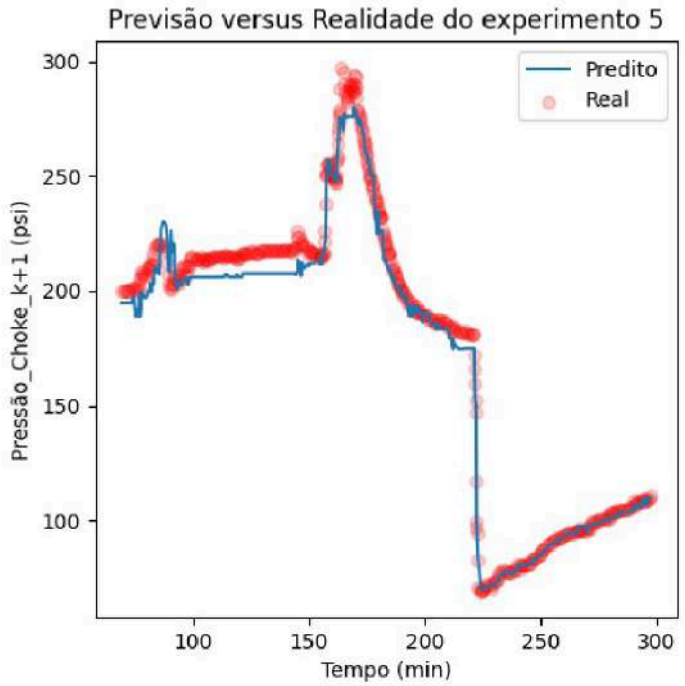

Figura *183* – Previsão versus realidade com dados de poços reais, tratados com a função log e com 2 dados passados, para o experimento 5. Fonte: A autora.

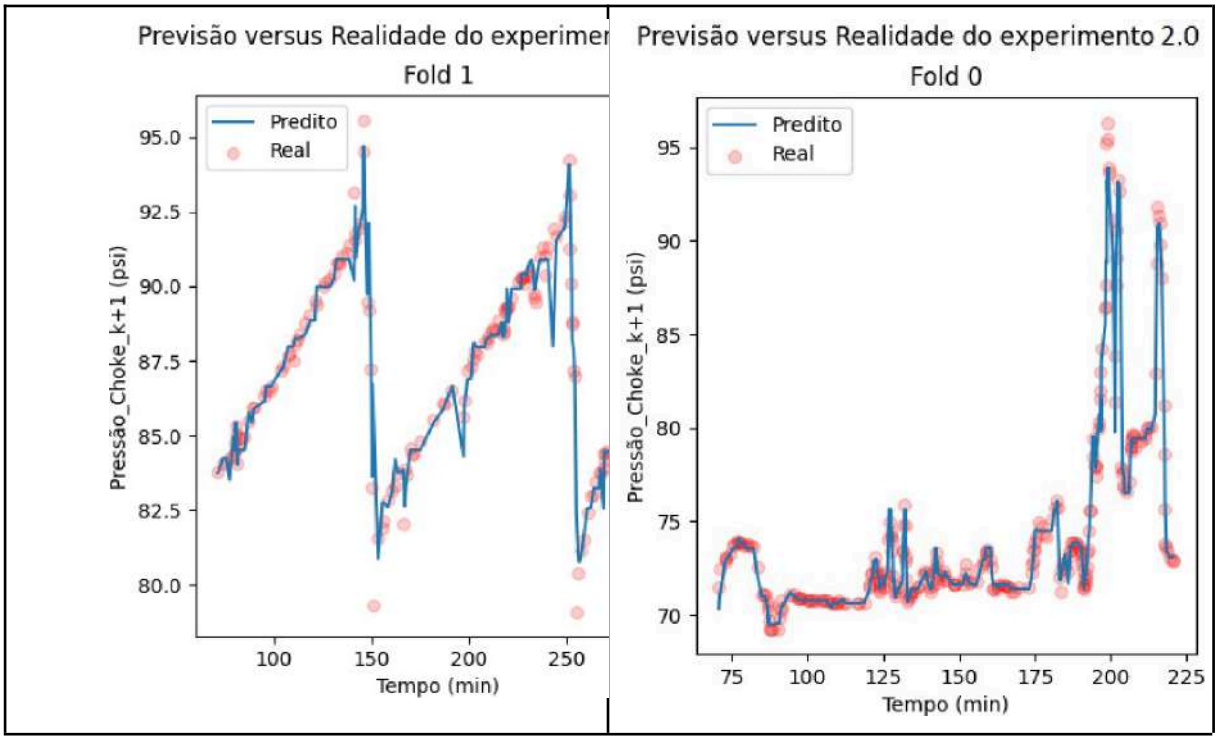

Figura *184* – Previsão versus realidade com dados de poços reais, tratados com a função log e com 2 dados passados, para os experimentos 1 e 2. Fonte: A autora.

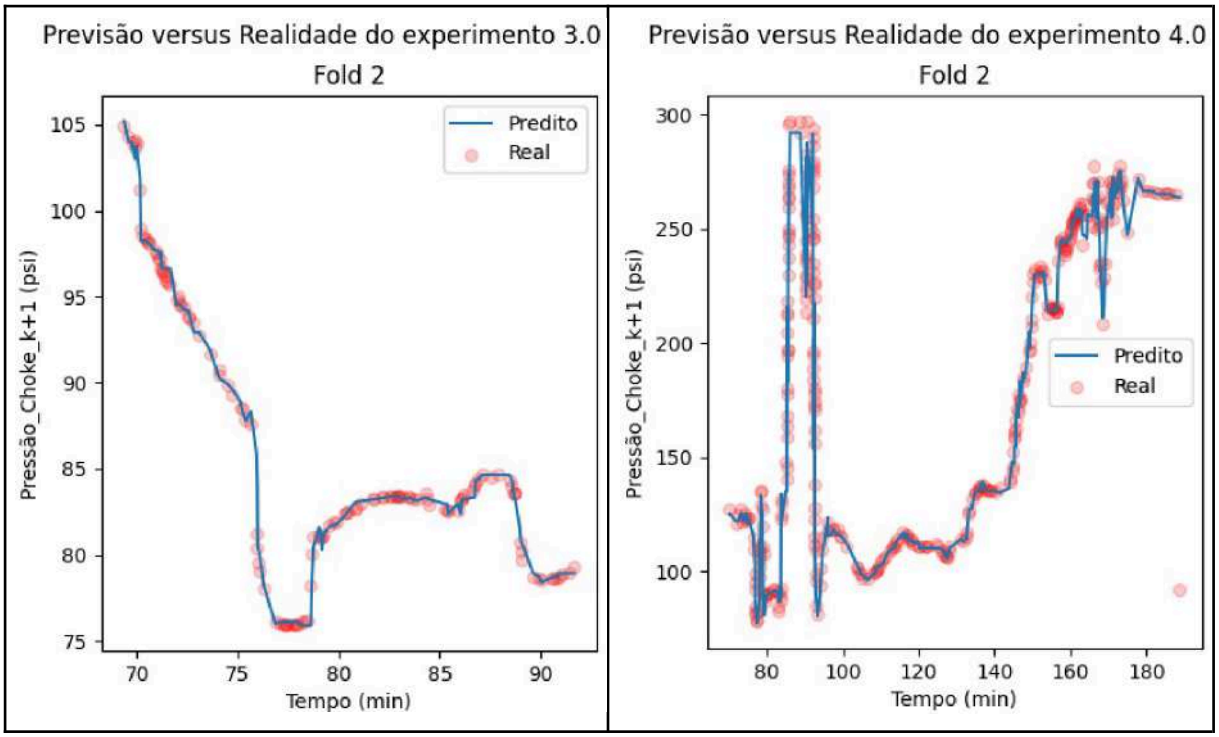

Figura *185* – Previsão versus realidade com dados de poços reais, tratados com a função log e com 2 dados passados, para os experimentos 3 e 4. Fonte: A autora.

Com relação à arquitetura do modelo, foram treinadas 746 árvores. As métricas de avaliação mostram que usar dados passados aprimora o desempenho do modelo matemático.

ANEXO J – RESULTADOS DOS DADOS DE POÇOS REAIS TRANSFORMADOS COM LOG, COM 8 DADOS PASSADOS

Ao aplicar 8 dados passados, todas as variáveis são deslocadas, conforme mostrado na Figura 186. A variável vazão apresentou *outiliers,* que foram tratados com a substituição pela sua média.

	Valores_Únicos	Tipo_de_Dado	mean	std	min	25%	50%	75%	max
Pressao_Choke_k-1 (psi)	2296	float64	0.98	0.77	0.0	0.38	0.67	1.86	2.40
Pressao_Choke_k-2 (psi)	2296	float64	0.98	0.77	0.0	0.38	0.67	1.86	2.40
Pressao_Choke_k-3 (psi)	2295	float64	0.98	0.77	0.0	0.38	0.67	1.86	2.40
Pressao_Choke_k-4 (psi)	2296	float64	0.98	0.77	0.0	0.38	0.67	1.86	2.40
Pressao_Choke_k-5 (psi)	2298	float64	0.98	0.77	0.0	0.38	0.67	1.86	2.40
Pressao_Choke_k-6 (psi)	2298	float64	0.98	0.77	0.0	0.38	0.67	1.86	2.40
Pressao_Choke_k-7 (psi)	2297	float64	0.98	0.77	0.0	0.38	0.67	1.87	2.40
Tempo_k (min)	3072	float64	1.25	0.64	0.0	0.66	1.40	1.76	2.40
Vazão_k (m³/h)	473	float64	0.12	0.50	0.0	0.00	0.00	0.00	2.40
Pressao_Choke_k (psi)	2296	float64	0.98	0.77	0.0	0.38	0.67	1.85	2.40
Pressão_Choke_k+1 (psi)	2298	float64	420.27	546.90	0.0	89.66	181.63	402.51	2001.09
experimento	5	int64	3.03	1.32	1.0	2.00	3.00	4.00	5.00

Figura *186* – Resumo do *dataframe* com dados de poços reais, tratados com a função log e com 8 dados passados. Fonte: A autora.

Antes dos dados serem treinados pelo XGBoost, é feito a otimização dos hiperparâmetros com o Optuna gerando os seguintes parâmetros: 'n_estimators': 826, 'learning_rate': 0.0554, 'max_depth': 15, 'min_child_weight': 3, 'subsample': 0.5296, 'colsample_bytree': 0.7607, 'gamma': 0.6534, 'reg_alpha': 0.8150 e 'reg_lambda': 0.8902, em apenas 2.63 segundos de execução.

Com os melhores parâmetros gerados pelo Optuna, os dados são treinados, validados e testados pelo XGBoost, gerando as métricas de avaliação ilustrados nas Figuras 187 e 188.

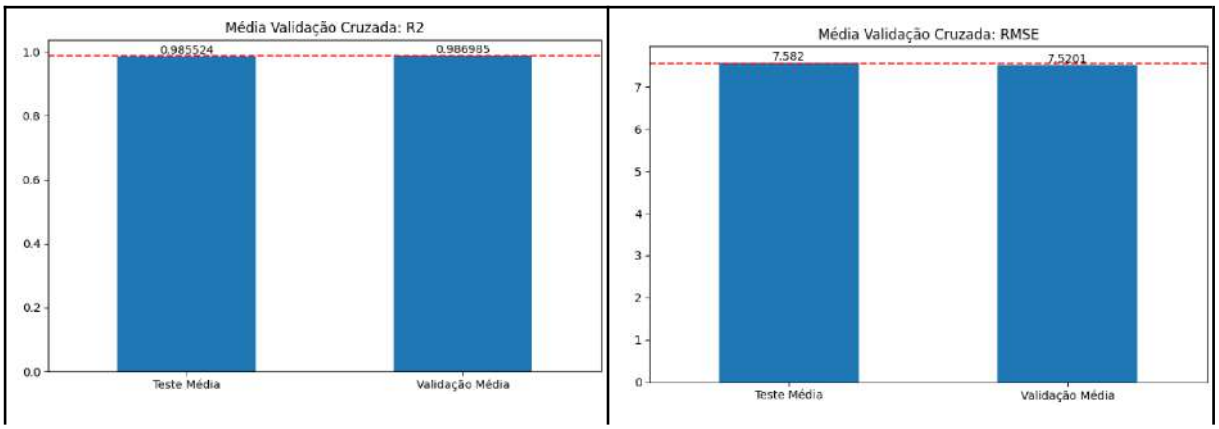

Figura *187* – Métricas de avaliação R² e RMSE com dados de poços reais, tratados com a função log e com 8 dados passados. Fonte: A autora.

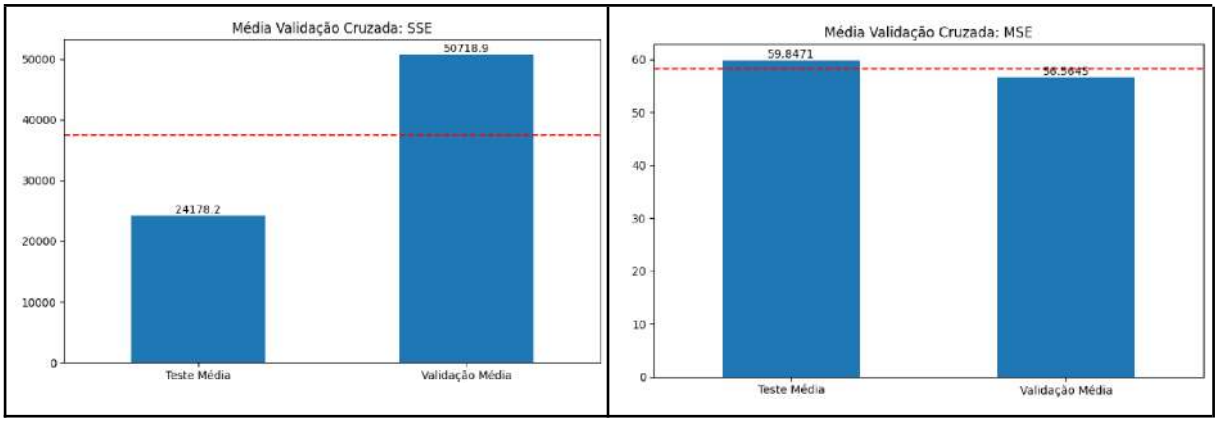

Figura *188* – Métricas de avaliação SSE e MSE com dados de poços reais, tratados com a função log e com 8 dados passados. Fonte: A autora.

São apresentados na Figura 189 os valores da densidade dos dados para a pressão *choke* no teste e na validação do modelo, sendo que quanto mais valores preditos de forma correta, ou seja, quando valor predito é igual ao valor real, mais será observada a presença da cor roxa nos gráficos.

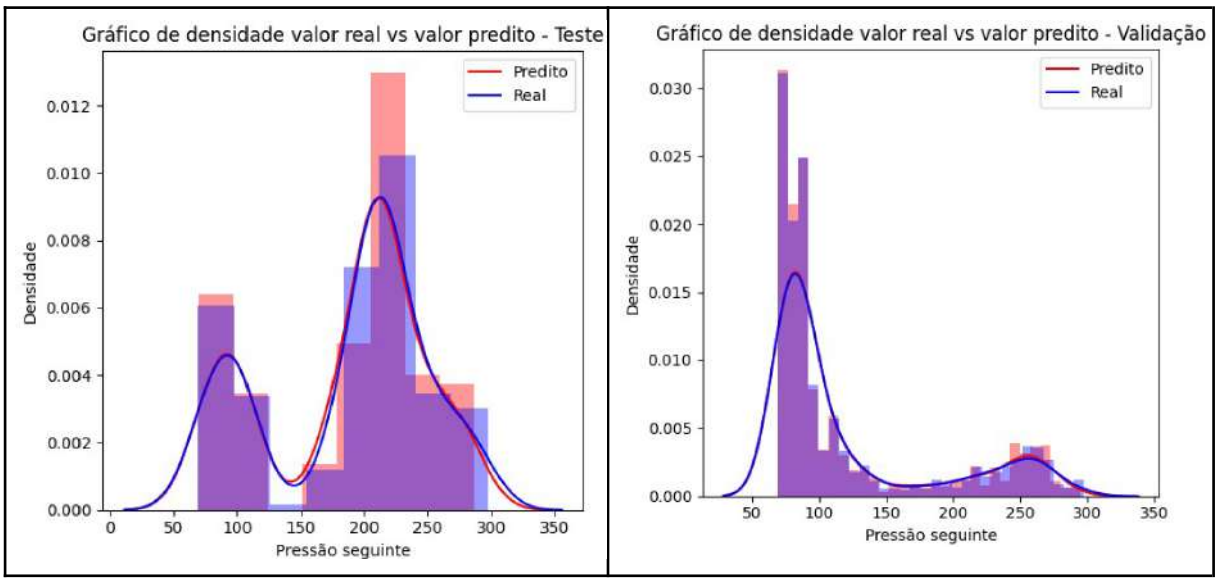

Figura *189* – Gráficos de densidade com dados de poços reais, tratados com a função log e com 8 dados passados. Fonte: A autora.

São apresentados na Figura 190 os valores da distribuição dos resíduos no teste e na validação do modelo.

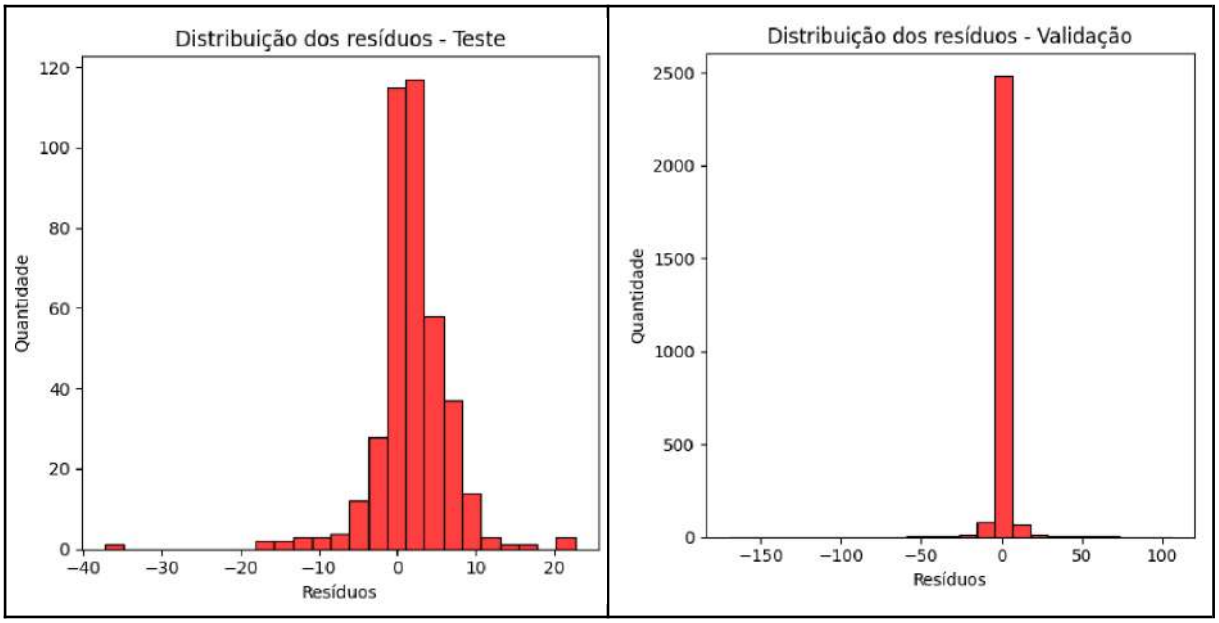

Figura *190* – Distribuição dos resíduos com dados de poços reais, tratados com a função log e com 8 dados passados. Fonte: A autora.

A Figura 191 apresenta os pontos na curva de comparação de valor real e valor predito.

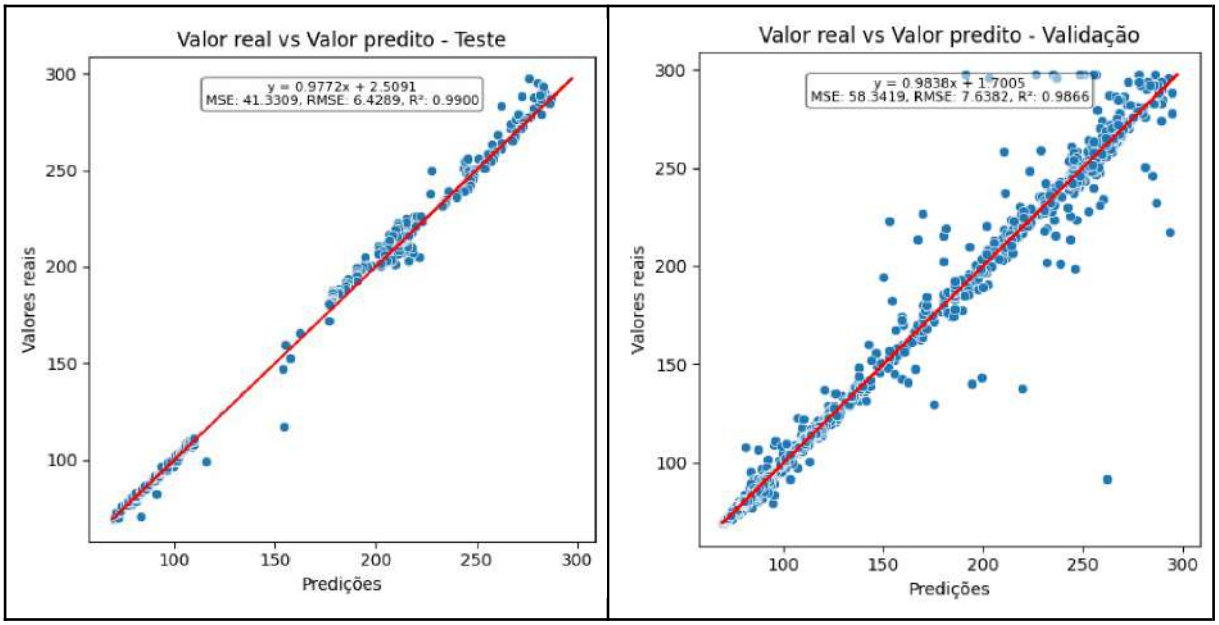

Figura *191* – Gráfico de evolução do modelo com dados de poços reais, tratados com a função log e com 8 dados passados. Fonte: A autora.

O gráfico da função de perda é apresentado na Figura 192.

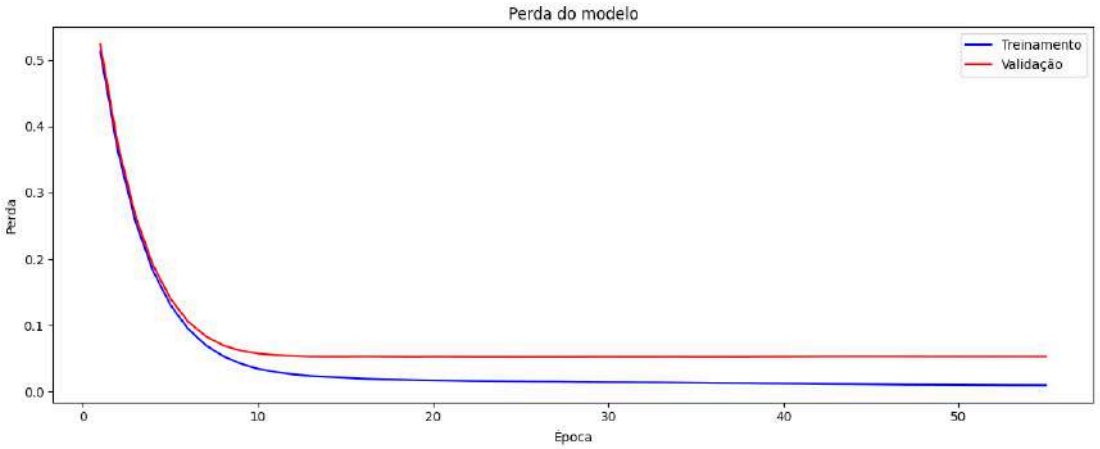

Figura *192* – Curva de perdas função log e com 8 dados passados. Fonte: A autora.

O gráfico na Figura 193 representa a importância de cada variável para as previsões, indicando que a pressão da *choke* é a variável mais relevante para a síntese do modelo.

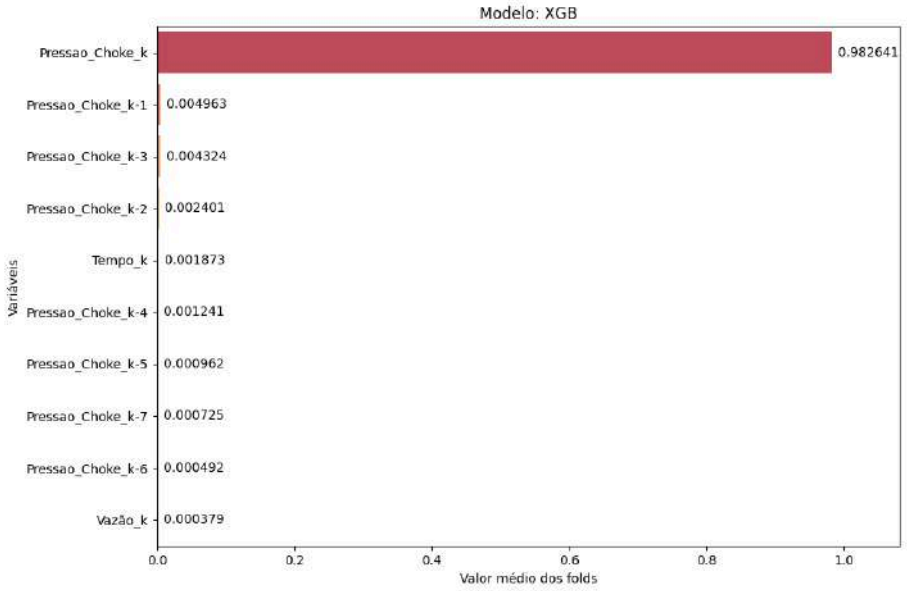

Figura *193* – Importância das variáveis com dados de poços reais, tratados com a função log e com 8 dados passados. Fonte: A autora.

Os resultados das previsões e realidade dos testes do modelo são apresentados nas Figuras 194 a 196, que representam o quanto o modelo consegue prever a operação de PMCD.

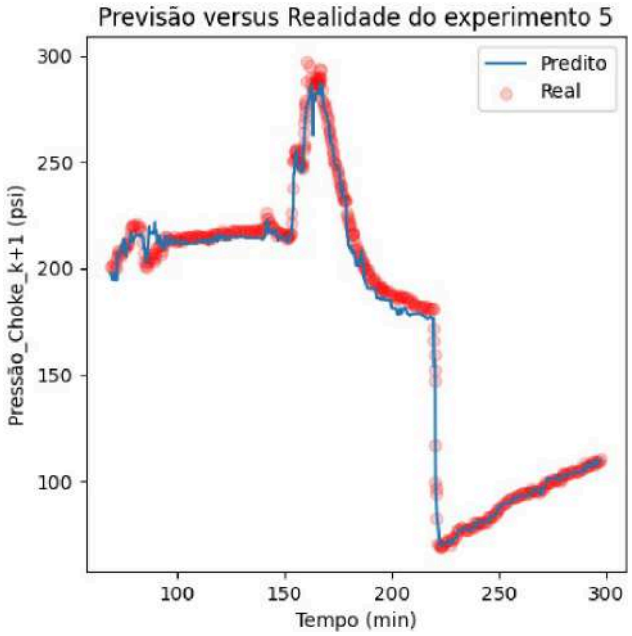

Figura *194* – Previsão versus realidade com dados de poços reais, tratados com a função log e com 8 dados passados, para o experimento 5. Fonte: A autora.

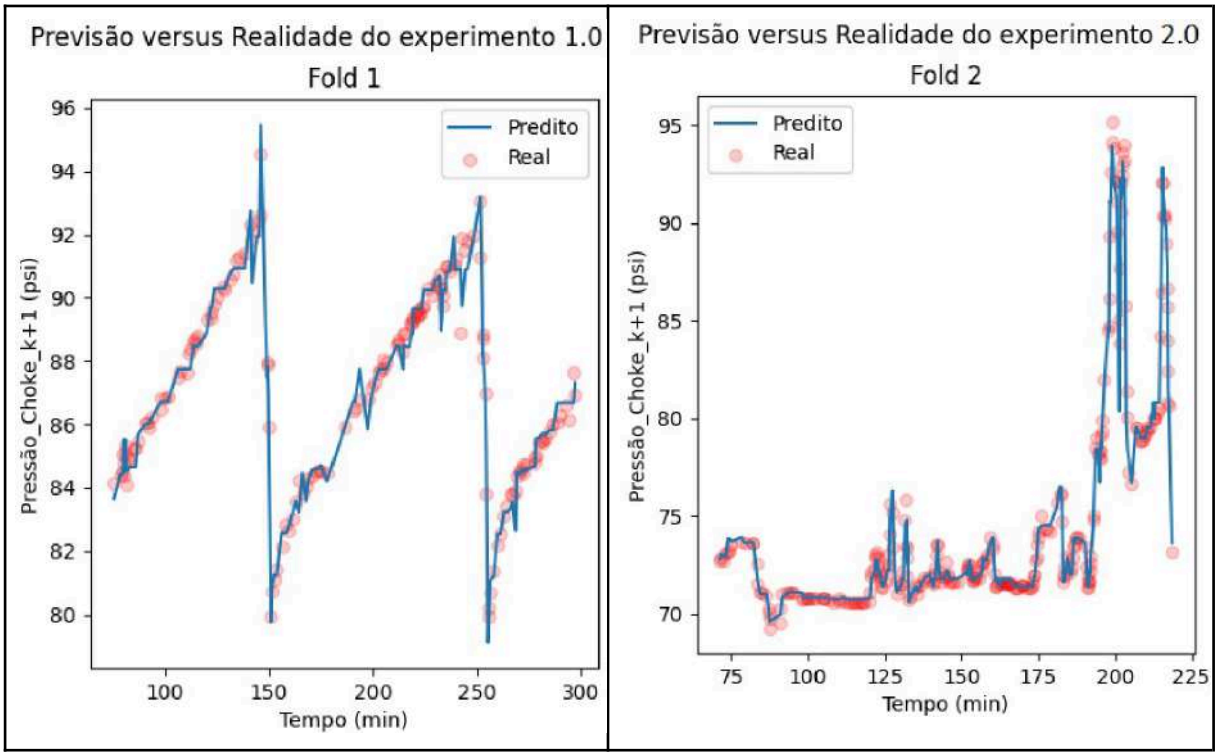

Figura *195* – Previsão versus realidade com dados de poços reais, tratados com a função sig e com 8 dados passados, para os experimentos 1 e 2. Fonte: A autora.

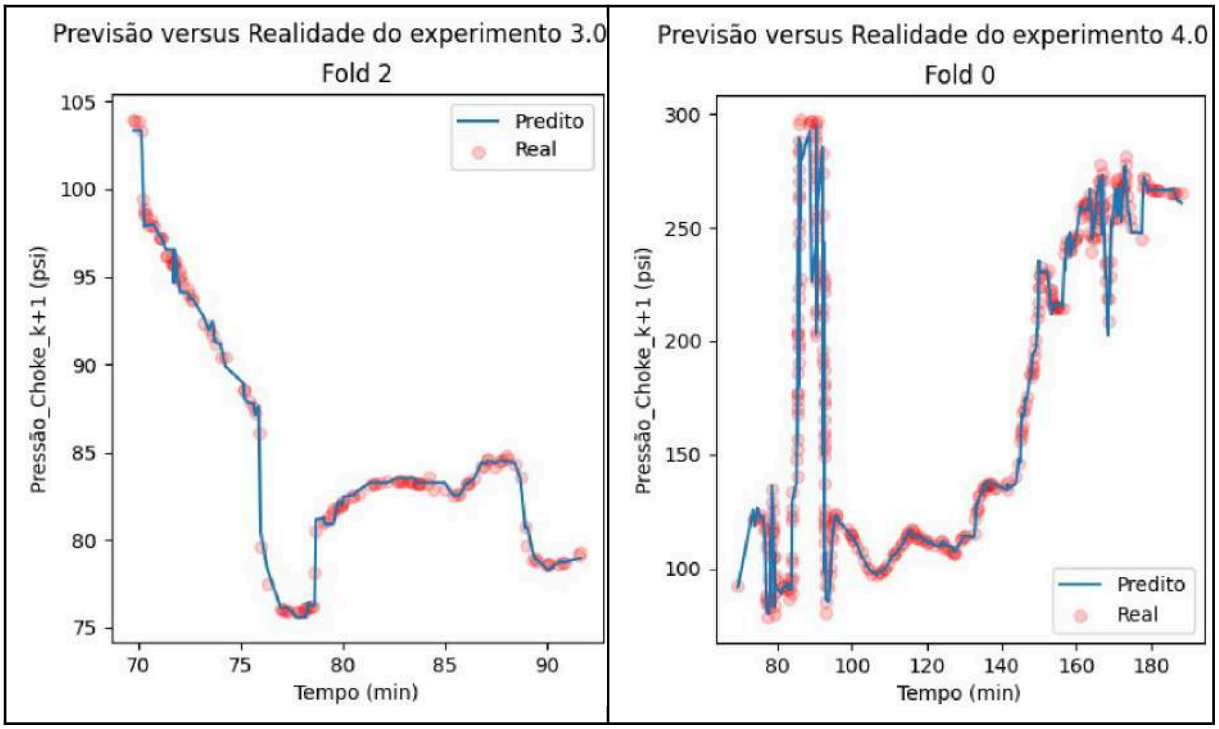

Figura *196* – Previsão versus realidade com dados de poços reais, tratados com a função log e com 8 dados passados, para os experimentos 3 e 4 Fonte: A autora.

Com relação à arquitetura do modelo, foram treinadas 826 árvores.

ANEXO K – RESULTADOS DOS DADOS DE POÇOS REAIS TRANSFORMADOS COM LOG, COM 20 DADOS PASSADOS

Com 20 dados passados, todas as variáveis são deslocadas, conforme mostrado na Figura 197. A vazão possui *outiliers* que foram tratados com a substituição pelas suas médias.

	Valores_Únicos	Tipo_de_Dado	mean	std	min	25%	50%	75%	max
Pressao_Choke_k-1 (psi)	2260	float64	0.98	0.77	0.0	0.38	0.67	1.86	2.40
Pressao_Choke_k-2 (psi)	2259	float64	0.98	0.77	0.0	0.38	0.67	1.86	2.40
Pressao_Choke_k-3 (psi)	2258	float64	0.98	0.77	0.0	0.38	0.67	1.86	2.40
Pressao_Choke_k-4 (psi)	2260	float64	0.98	0.77	0.0	0.38	0.67	1.86	2.40
Pressao_Choke_k-5 (psi)	2260	float64	0.99	0.77	0.0	0.38	0.67	1.87	2.40
Pressao_Choke_k-6 (psi)	2262	float64	0.99	0.77	0.0	0.39	0.67	1.87	2.40
Pressao_Choke_k-7 (psi)	2263	float64	0.99	0.77	0.0	0.39	0.67	1.87	2.40
Pressao_Choke_k-8 (psi)	2264	float64	0.99	0.77	0.0	0.39	0.68	1.87	2.40
Pressao_Choke_k-9 (psi)	2266	float64	0.99	0.77	0.0	0.39	0.68	1.87	2.40
Pressao_Choke_k-10 (psi)	2268	float64	0.99	0.77	0.0	0.39	0.68	1.87	2.40
Pressao_Choke_k-11 (psi)	2267	float64	0.99	0.77	0.0	0.39	0.68	1.88	2.40
Pressao_Choke_k-12 (psi)	2267	float64	0.99	0.77	0.0	0.39	0.68	1.88	2.40
Pressao_Choke_k-13 (psi)	2268	float64	0.99	0.77	0.0	0.39	0.68	1.88	2.40
Pressao_Choke_k-14 (psi)	2268	float64	0.99	0.77	0.0	0.39	0.68	1.88	2.40
Pressao_Choke_k-15 (psi)	2269	float64	0.99	0.77	0.0	0.39	0.68	1.88	2.40
Pressao_Choke_k-16 (psi)	2270	float64	0.99	0.77	0.0	0.39	0.68	1.89	2.40
Pressao_Choke_k-17 (psi)	2272	float64	0.99	0.77	0.0	0.39	0.68	1.89	2.40
Pressao_Choke_k-18 (psi)	2273	float64	0.99	0.77	0.0	0.39	0.68	1.89	2.40
Pressao_Choke_k-19 (psi)	2272	float64	0.99	0.77	0.0	0.39	0.68	1.89	2.40
Tempo_k (min)	3016	float64	1.27	0.63	0.0	0.66	1.42	1.76	2.40
Vazão_k (m³/h)	452	float64	0.12	0.50	0.0	0.00	0.00	0.00	2.40
Pressao_Choke_k (psi)	2260	float64	0.98	0.77	0.0	0.38	0.67	1.86	2.40
Pressão_Choke_k+1 (psi)	2253	float64	423.47	550.98	0.0	89.47	181.27	403.80	2001.09
experimento	5	int64	4.07	1.63	1.0	3.00	4.00	6.00	6.00

Figura *197* – Resumo do *dataframe* com dados de poços reais, tratados com a função log e com 20 dados passados. Fonte: A autora.

Antes dos dados serem treinados pelo XGBoost, é feito a otimização dos hiperparâmetros com o Optuna gerando os seguintes parâmetros: 'n_estimators': 522, 'learning_rate': 0.2000, 'max_depth': 15, 'min_child_weight': 1, 'subsample': 0.5993, 'colsample_bytree': 0.9396, 'gamma': 0.0516, 'reg_alpha': 0.5527, 'reg_lambda': 0.7996, em apenas 5,83 segundos de execução, utilizando o mesmo ambiente de execução dos modelos anteriores.

Com os melhores parâmetros gerados pelo Optuna, os dados são treinados, validados e testados pelo XGBoost. As métricas de avaliação do modelo são apresentadas na Figura 198.

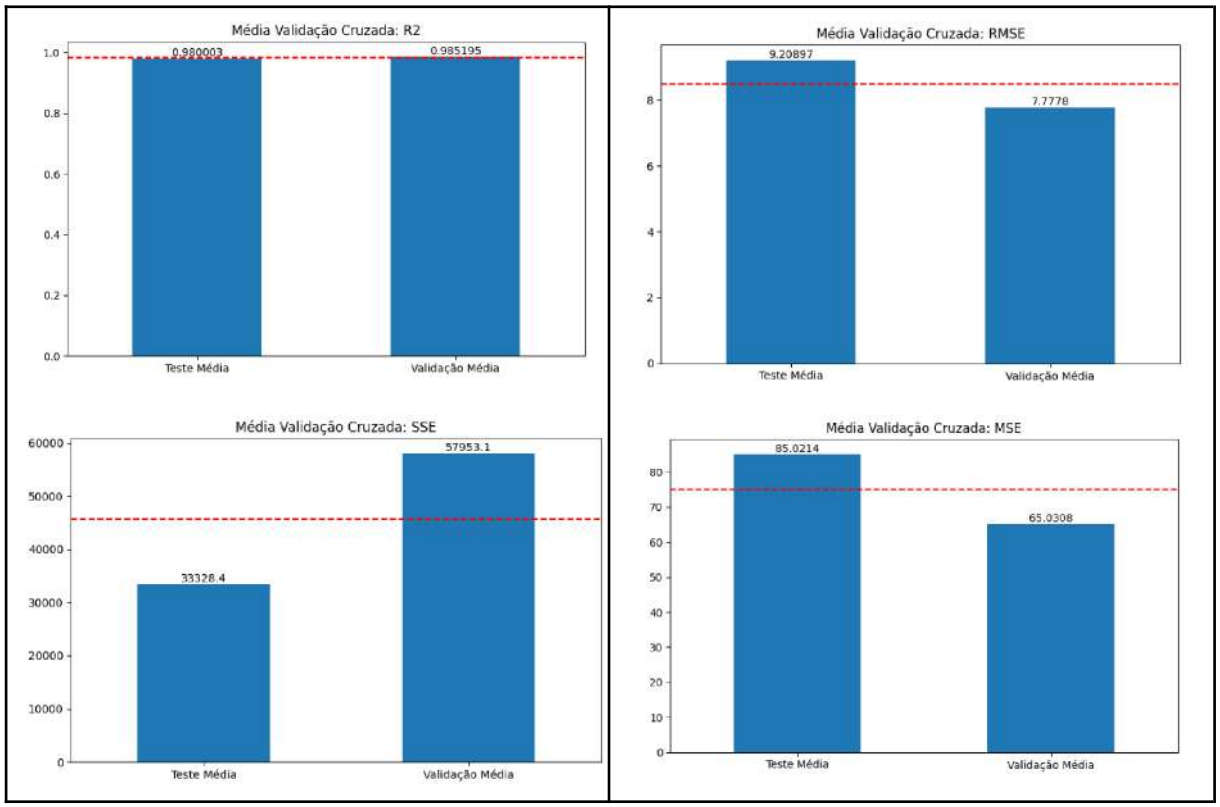

Figura *198* – Métricas de avaliação com dados de poços reais, tratados com a função log e com 20 dados passados. Fonte: A autora.

São apresentados na Figura 199 os valores da densidade dos dados para a pressão na *choke*, sendo que quanto mais valores preditos de forma correta, mais será observada a presença da cor roxa nos gráficos.

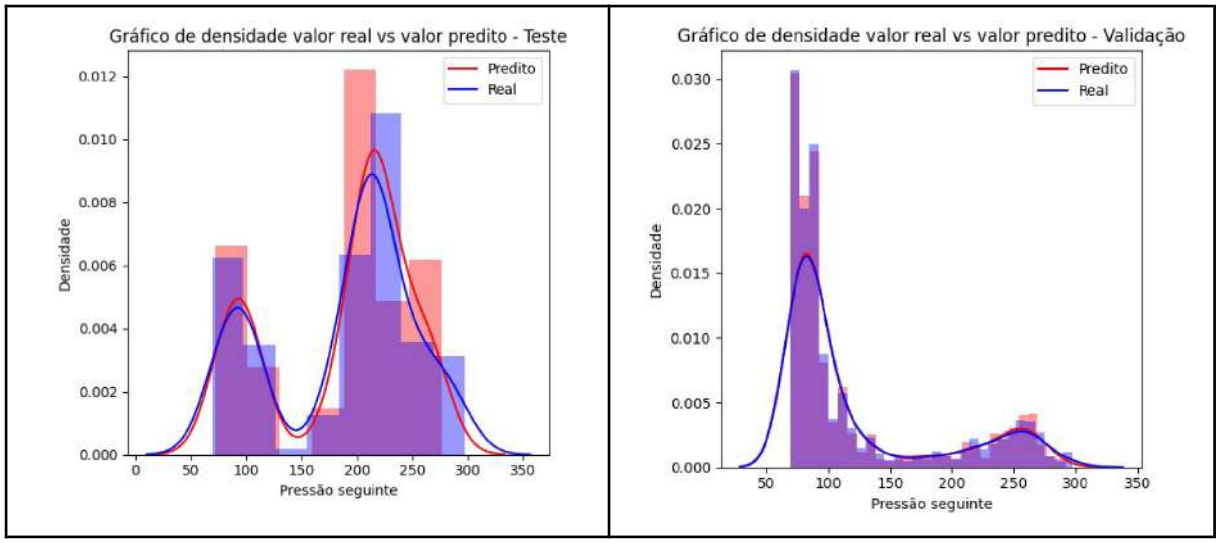

Figura *199* – Gráficos de densidade com dados de poços reais, tratados com a função log e com 20 dados passados. Fonte: A autora.

São apresentados na Figura 200 os valores da distribuição dos resíduos no teste e na validação do modelo.

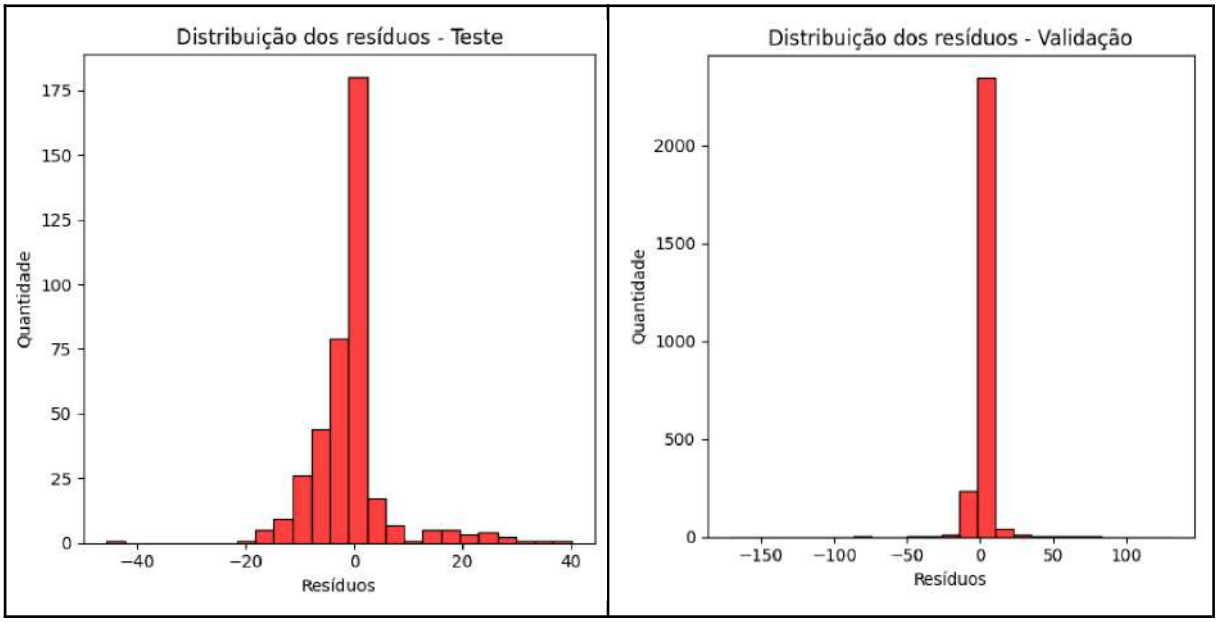

Figura *200* – Distribuição dos resíduos com dados de poços reais, tratados com a função log e com 20 dados passados. Fonte: A autora.

A Figura 201 apresenta os pontos na curva de comparação de valor real e valor predito.

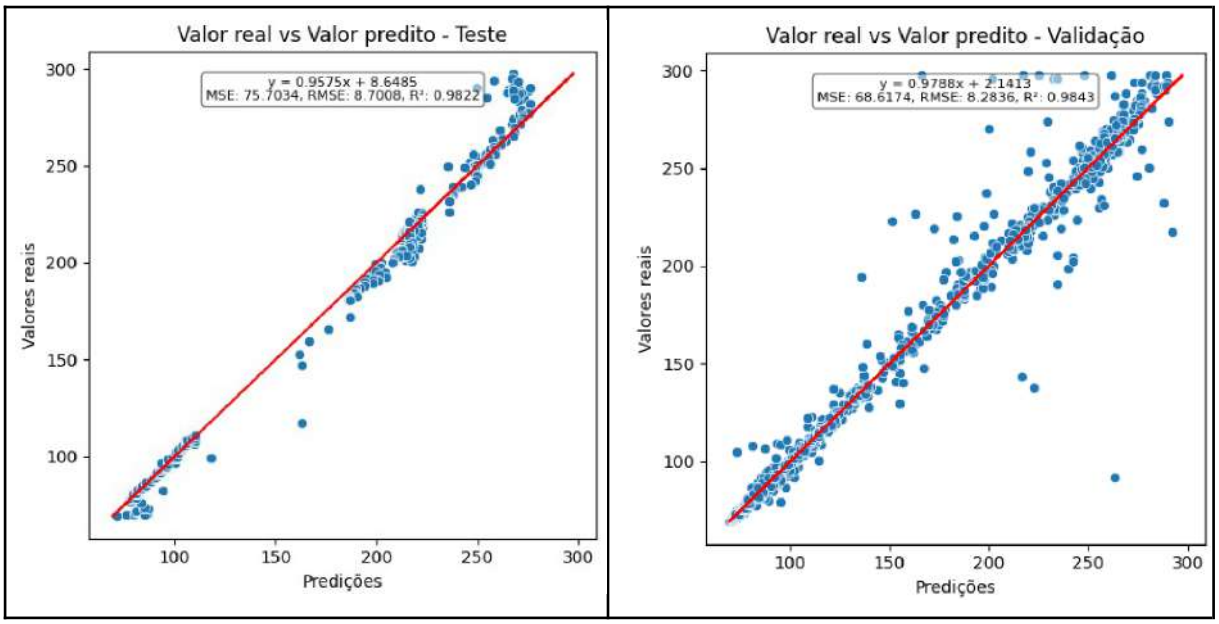

Figura *201* – Gráfico de evolução do modelo com dados de poços reais, tratados com a função log e com 20 dados passados. Fonte: A autora.

O gráfico da função de perda é apresentado na Figuras 202.

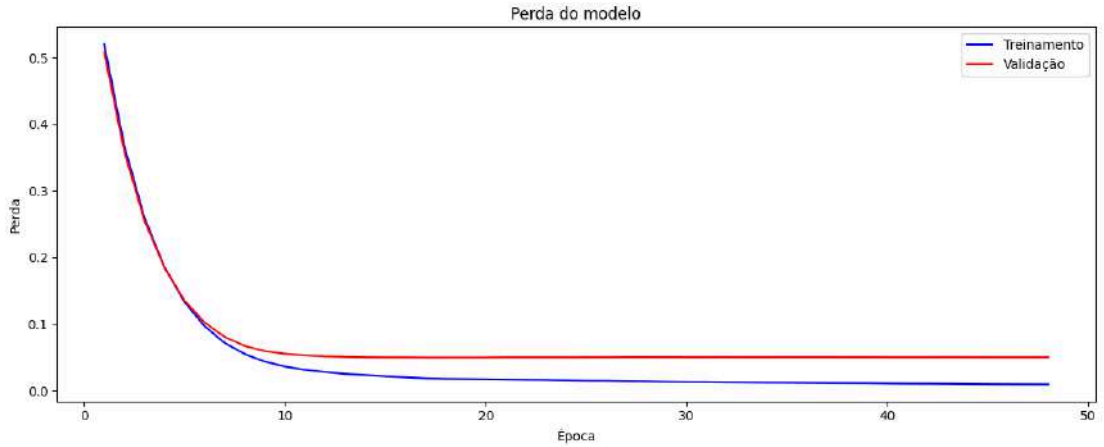

Figura *202* – Curva de perdas função log e com 20 dados passados. Fonte: A autora.

O gráfico na Figura 203 representa a importância de cada variável para as previsões, seguindo o modelo do XGBoost treinado, sendo que a mais importante é a pressão da *choke* no passo atual.

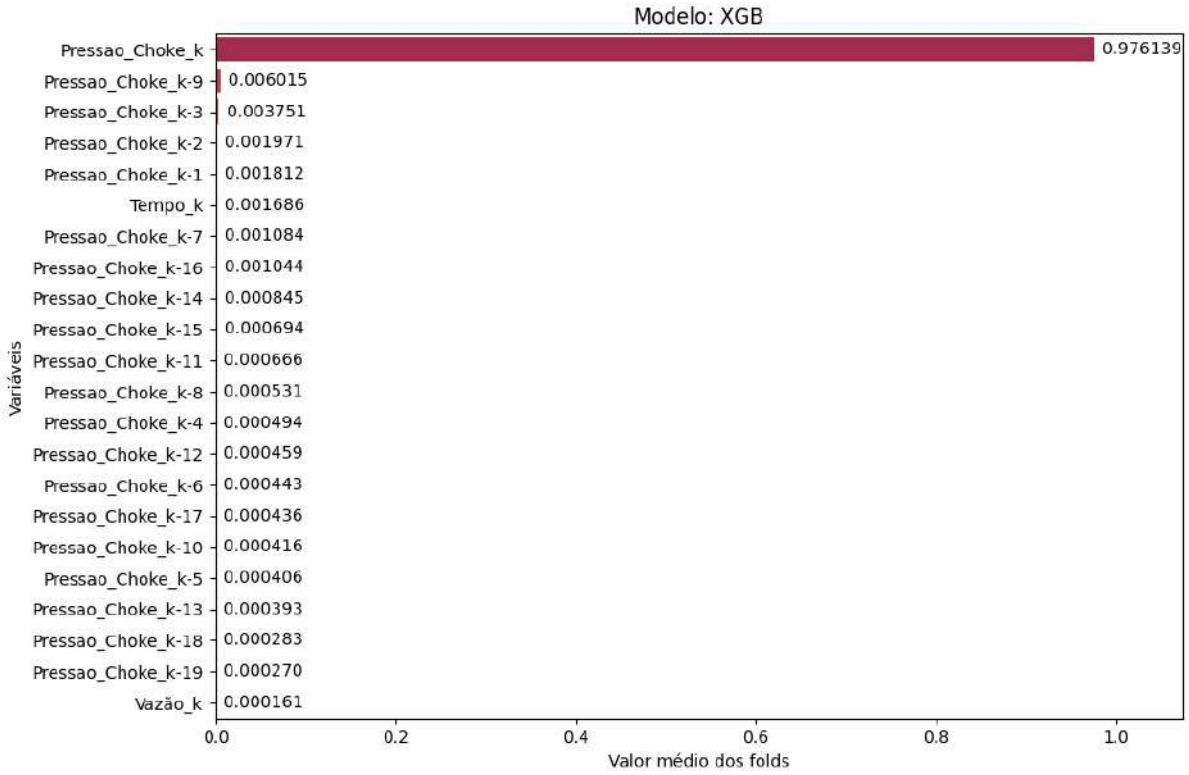

Figura *203* – Importância das variáveis com dados de poços reais, tratados com a função log e com 20 dados passados. Fonte: A autora.

Os resultados das previsões e realidade do modelo são apresentados nas Figuras 204 a 206, indicando o quanto o modelo consegue prever a operação de PMCD.

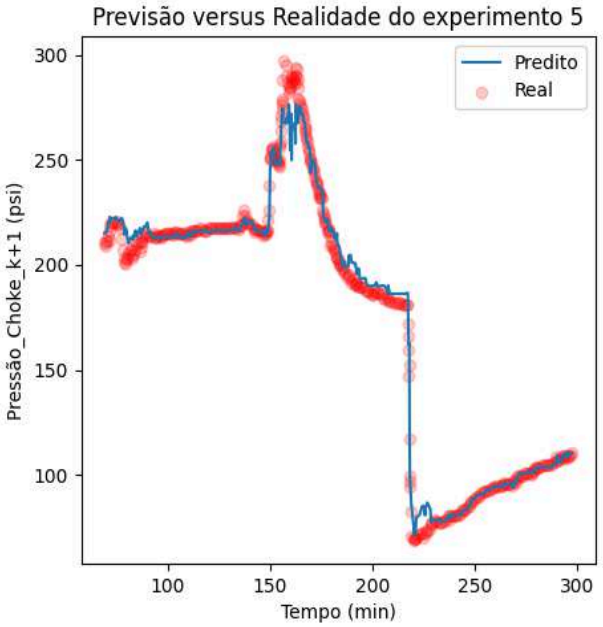

Figura *204* – Previsão versus realidade com dados de poços reais, tratados com a função log e com 20 dados passados, para o experimento 5. Fonte: A autora.

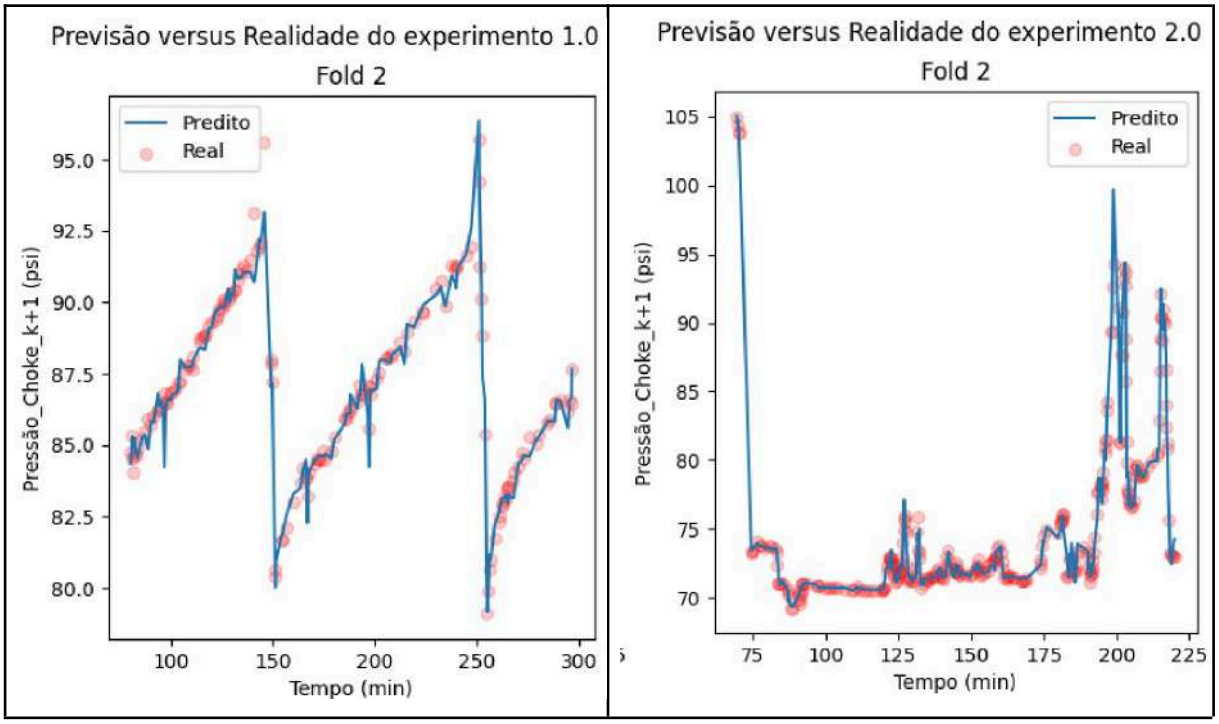

Figura *205* – Previsão versus realidade com dados de poços reais, tratados com a função sig e com 20 dados passados, para os experimentos 1 e 2. Fonte: A autora.

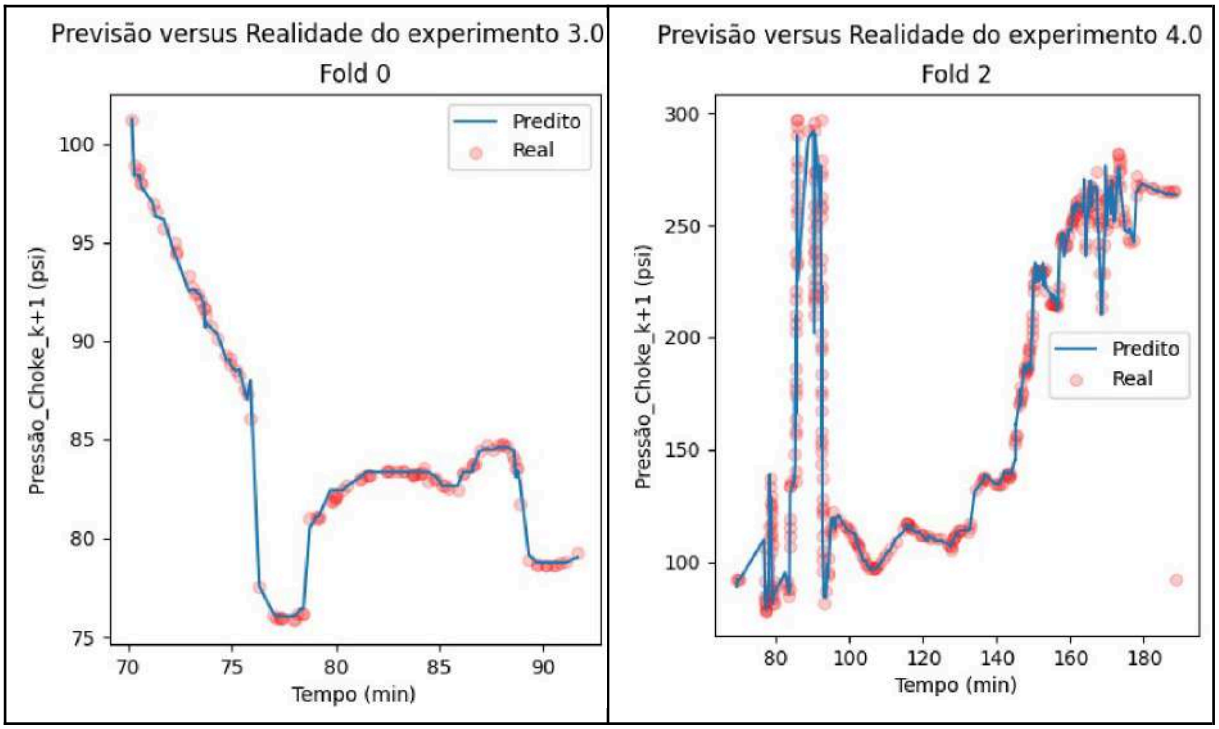

Figura *206* – Previsão versus realidade com dados de poços reais, tratados com a função log e com 20 dados passados, para os experimentos 3 e 4 Fonte: A autora.

Com relação à arquitetura do modelo, foram treinadas 522 árvores.

Para dados referentes a poços reais (Jayah (2013), Zein (2017) e Wattanasuwankorn (2014)), a Tabela 12 apresenta os tempos computacionais requeridos para a construção dos modelos baseados em *machine learning*.

Tabela 12 – Tempos computacionais dos modelos com dados de poços reais com log. Fonte: A autora.

Modelos com dados de poços reais Log (0, 10)	Tempo Total Optuna	Tempo Total XGBoost	Tempo Total Optuna e XGBoost
Sem dados passados	3.62 s	3.57 s	7.19 s
Com 2 dados passados	2.79 s	3.38 s	6.17 s
Com 8 dados passados	2.63 s	3.77 s	6.4 s
Com 20 dados passados	5.83 s	4.41 s	10.24 s

ANEXO L – RESULTADO DOS DADOS EXPERIMENTAIS, EXCLUINDO AS VARIÁVEIS COM MENOR IMPORTÂNCIA NO APRENDIZADO

Foram excluídas as variáveis que apresentaram pouca relevância no aprendizado do modelo com 8 dados passados, com escala de -4 a 4 e transformados com sig, com isso, foram mantidas apenas as variáveis tempo poço no passo atual e as referentes a pressão choke, no passo atual, deslocadas e no passo futuro, sendo mostrados na Figura 207.

	Valores_Únicos	Tipo_de_Dado	mean	std	min	25%	50%	75%	max
Pressao_Choke_k-1 (psi)	176138	float64	0.41	0.35	0.02	0.03	0.33	0.78	0.98
Pressao_Choke_k-2 (psi)	176142	float64	0.41	0.35	0.02	0.03	0.33	0.78	0.98
Pressao_Choke_k-3 (psi)	176136	float64	0.41	0.35	0.02	0.03	0.33	0.78	0.98
Pressao_Choke_k-4 (psi)	176132	float64	0.41	0.35	0.02	0.03	0.33	0.78	0.98
Pressao_Choke_k-5 (psi)	176116	float64	0.41	0.35	0.02	0.03	0.33	0.78	0.98
Pressao_Choke_k-6 (psi)	176115	float64	0.41	0.35	0.02	0.03	0.33	0.78	0.98
Pressao_Choke_k-7 (psi)	176113	float64	0.41	0.35	0.02	0.03	0.33	0.78	0.98
Tempo_poco_k (min)	169067	float64	0.40	0.33	0.02	0.08	0.30	0.72	0.98
Pressao_Choke_k (psi)	176071	float64	0.41	0.35	0.02	0.03	0.33	0.78	0.98
Pressão_Choke_k+1 (psi)	155655	float64	35.61	26.45	0.00	4.42	36.32	60.22	94.68
experimento	32	int64	15.61	9.35	0.00	7.00	16.00	24.00	31.00

Figura *207* – Resumo do *dataframe* com dados experimentais, excluindo as variáveis que apresentaram menor importância no aprendizado. Fonte: A autora.

Antes dos dados serem treinados pelo XGBoost, foi efetuado a otimização dos hiperparâmetros com o Optuna gerando os seguintes parâmetros: 'n_estimators': 512, 'learning_rate': 0.1780, 'max_depth': 16, 'min_child_weight': 3, 'subsample': 0.9304, 'colsample_bytree': 0.7140, 'gamma': 0.2316, 'reg_alpha': 0.6015 e 'reg_lambda': 0.1723, em apenas 231 segundos de execução. Logo em seguida, os dados são treinados, validados e testados pelo XGBoost, gerando as métricas de avaliação ilustrados nas Figuras 208 e 209.

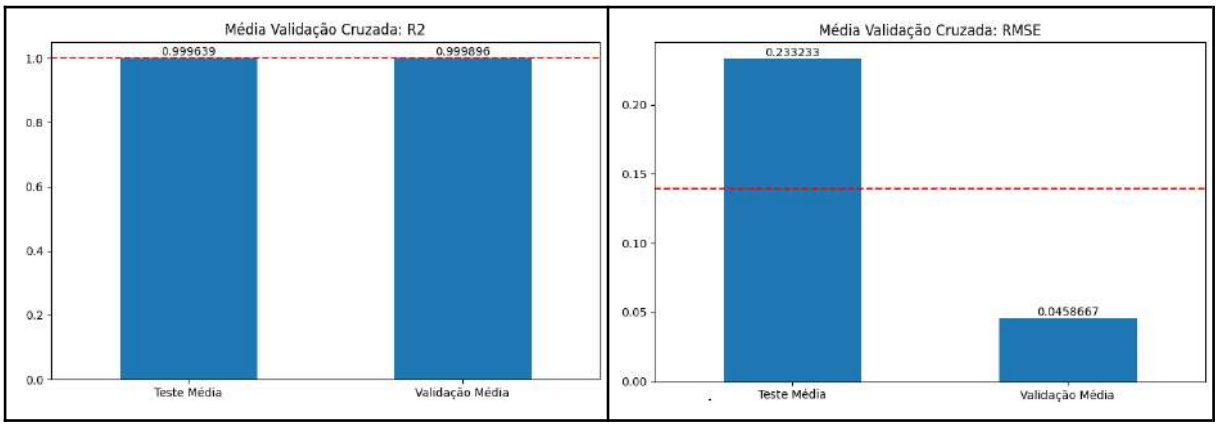

Figura *208* – Métricas de avaliação R² e RMSE com dados experimentais, excluindo as variáveis que apresentaram menor importância no aprendizado. Fonte: A autora.

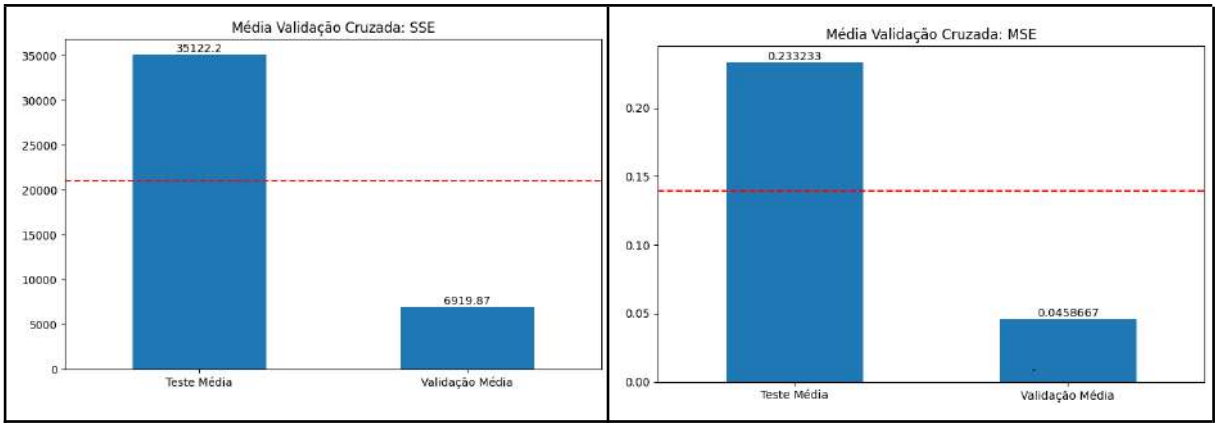

Figura *209* – Métricas de avaliação SSE e MSE com dados experimentais, excluindo as variáveis que apresentaram menor importância no aprendizado. Fonte: A autora.

São apresentados na Figura 210 os valores da densidade dos dados para a pressão na *choke*, no teste e na validação do modelo.

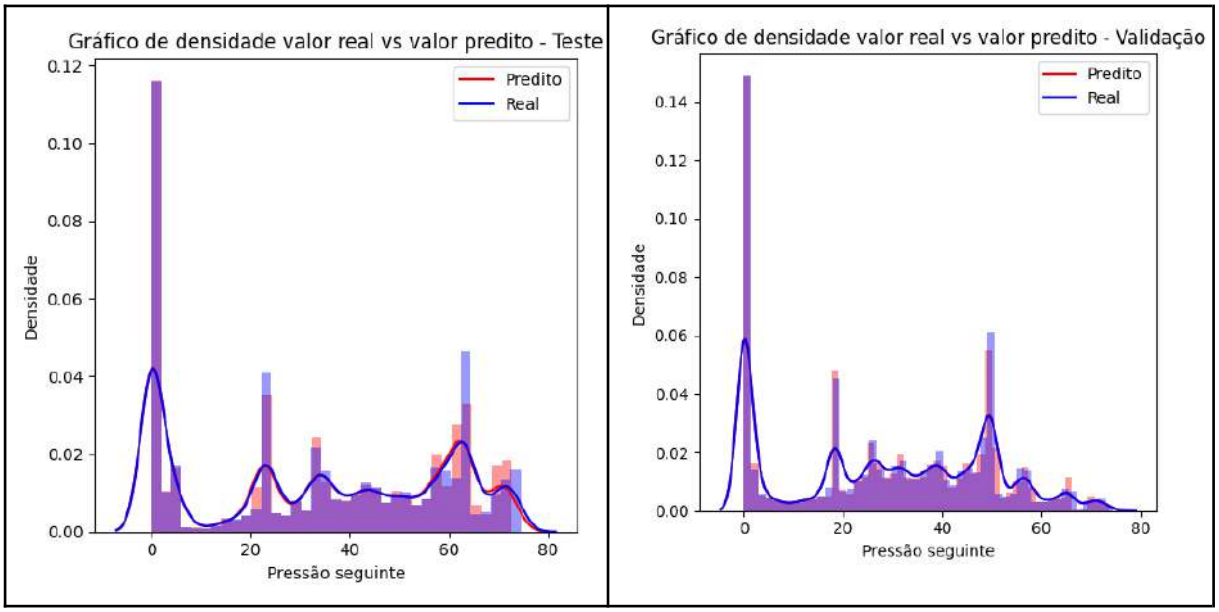

Figura *210* – Gráficos de densidade com dados experimentais, excluindo as variáveis que apresentaram menor importância no aprendizado. Fonte: A autora.

São apresentados na Figura 211 os valores da distribuição dos resíduos no teste e na validação do modelo.

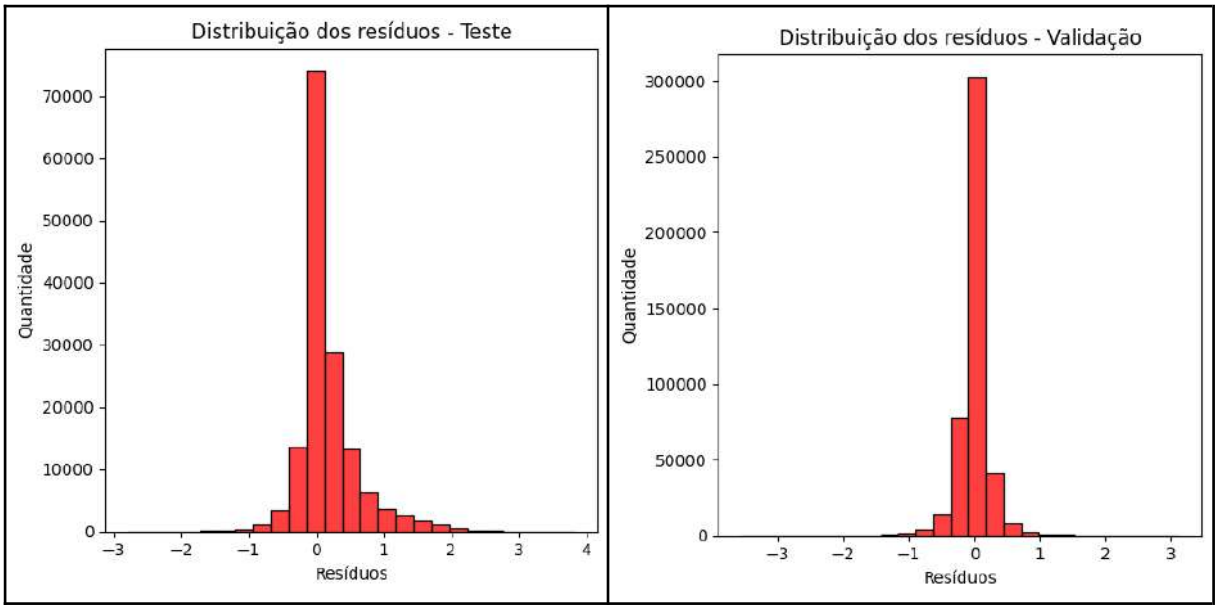

Figura *211* – Distribuição dos resíduos com dados experimentais, excluindo as variáveis que apresentaram menor importância no aprendizado. Fonte: A autora.

A Figura 212 apresenta os pontos na curva de comparação de valor real e valor predito.

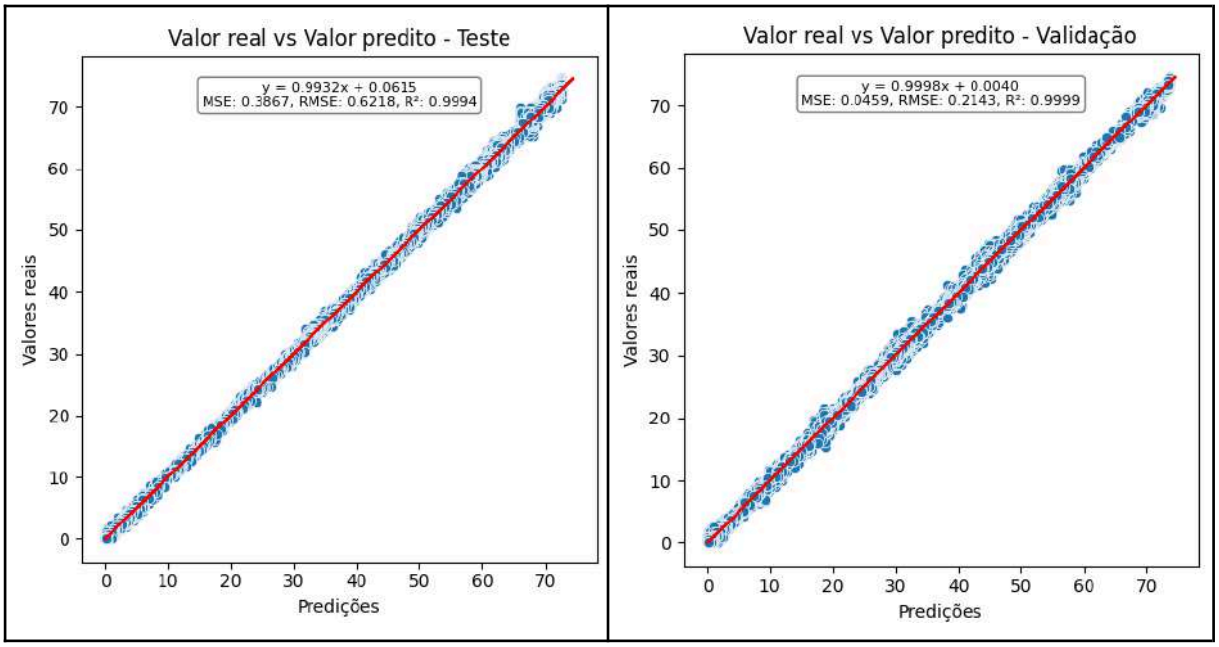

Figura *212* – Gráfico de evolução do modelo com dados experimentais, excluindo as variáveis que apresentaram menor importância no aprendizado. Fonte: A autora.

O gráfico da função de perda é apresentado na Figuras 213.

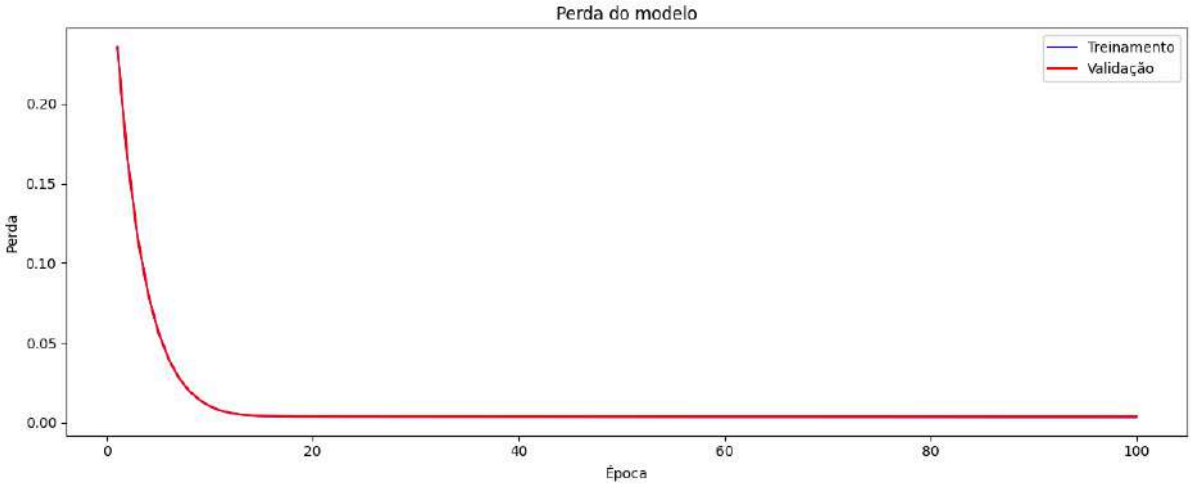

Figura *213* – Curva de perdas do modelo com dados experimentais, excluindo as variáveis que apresentaram menor importância no aprendizado. Fonte: A autora.

O gráfico na Figura 214 representa a importância de cada variável para as previsões, sendo que, a pressão da *choke* no passo atual apresenta maior importância no aprendizado do modelo.

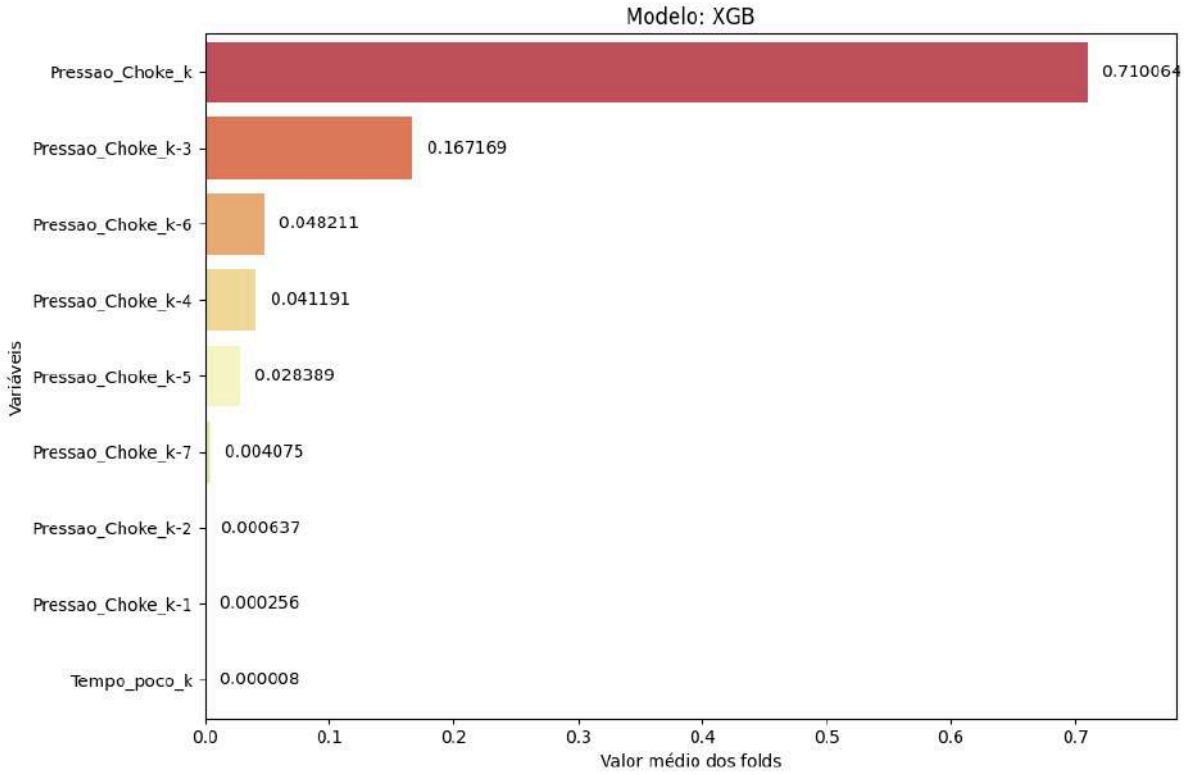

Figura *214* – Importância das variáveis do modelo com dados experimentais, excluindo as variáveis que apresentaram menor importância no aprendizado. Fonte: A autora.

Os resultados das previsões e realidade dos testes do modelo são apresentados nas Figuras 215 a 218, indicando o quanto o modelo consegue prever a operação de PMCD e *bullheading*.

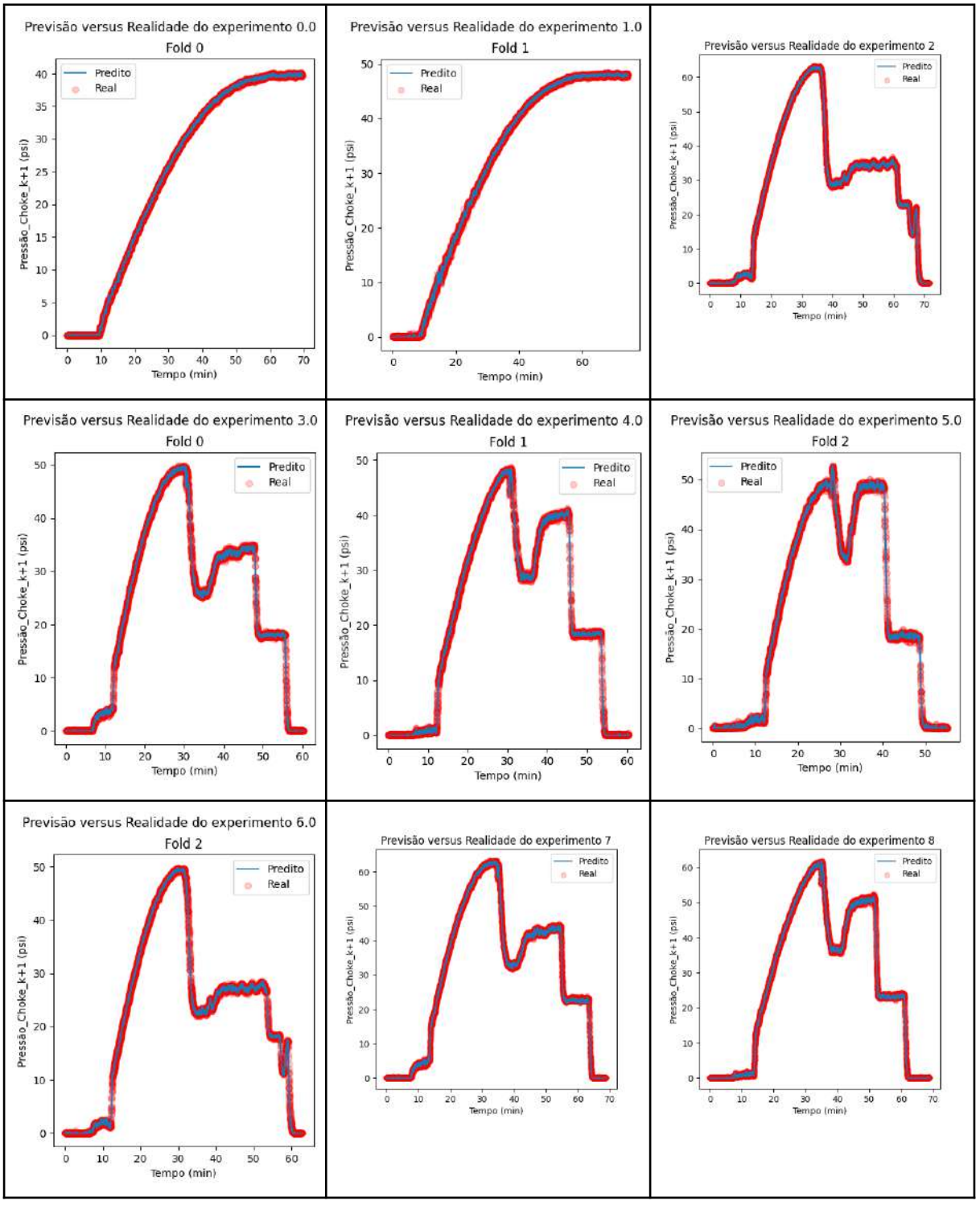

Figura *215* – Previsão e realidade dos experimentos de 0 a 8, excluindo as variáveis que apresentaram menor importância no aprendizado. Fonte: A autora.

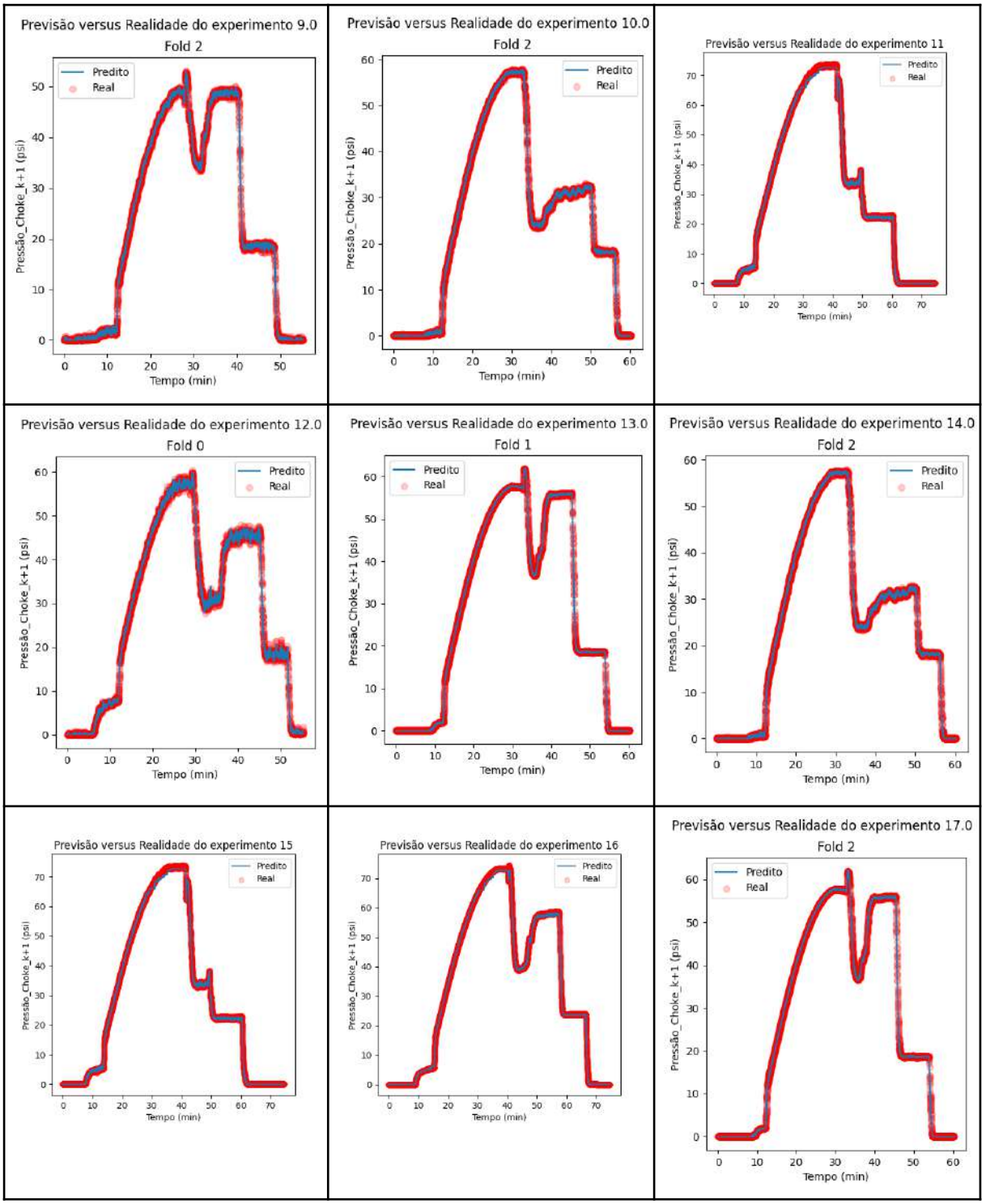

Figura 2*16* – Previsão e realidade dos experimentos de 9 a 17, excluindo as variáveis que apresentaram menor importância no aprendizado. Fonte: A autora.

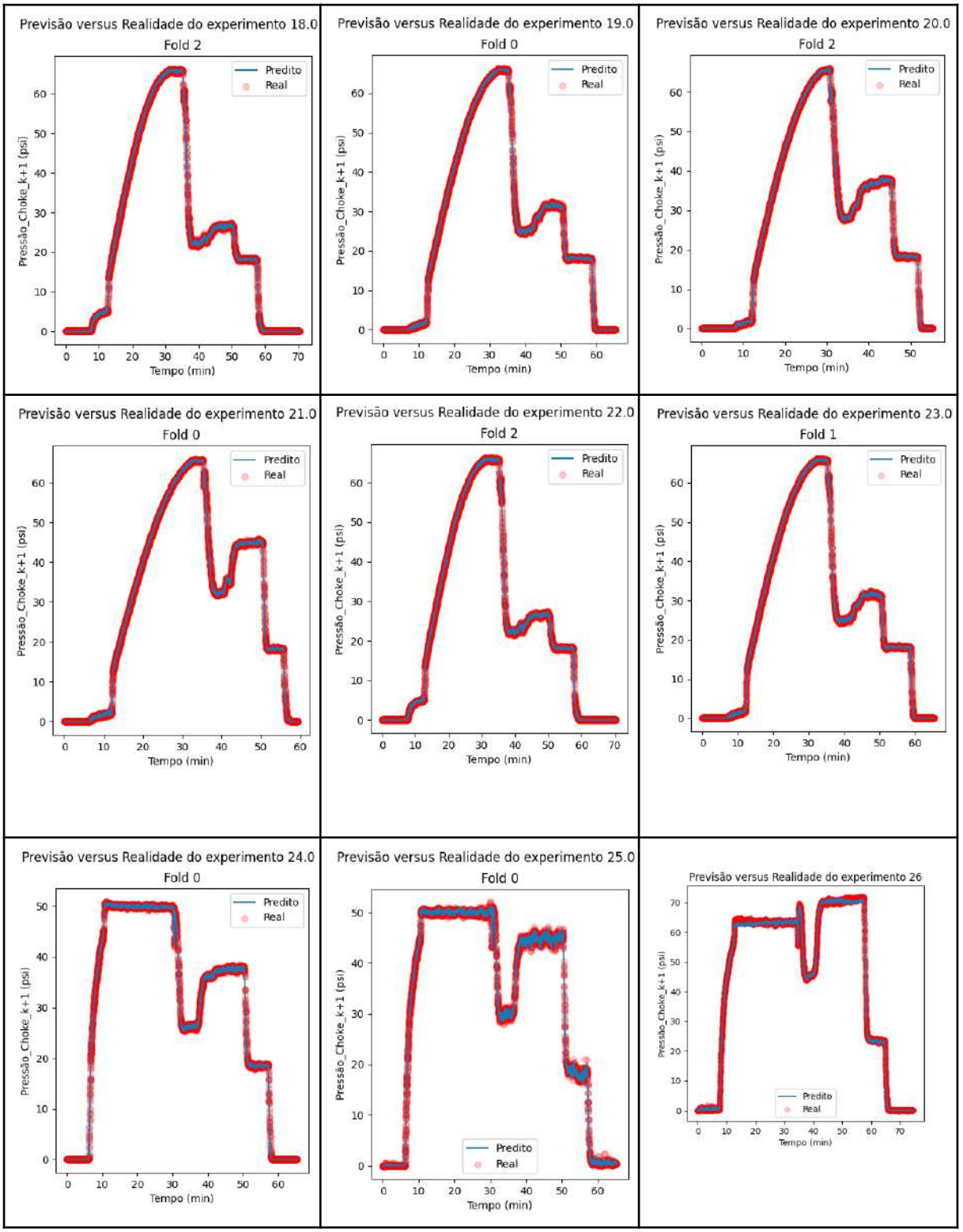

Figura 2*17* – Previsão e realidade dos experimentos de 18 a 26, excluindo as variáveis que apresentaram menor importância no aprendizado. Fonte: A autora.

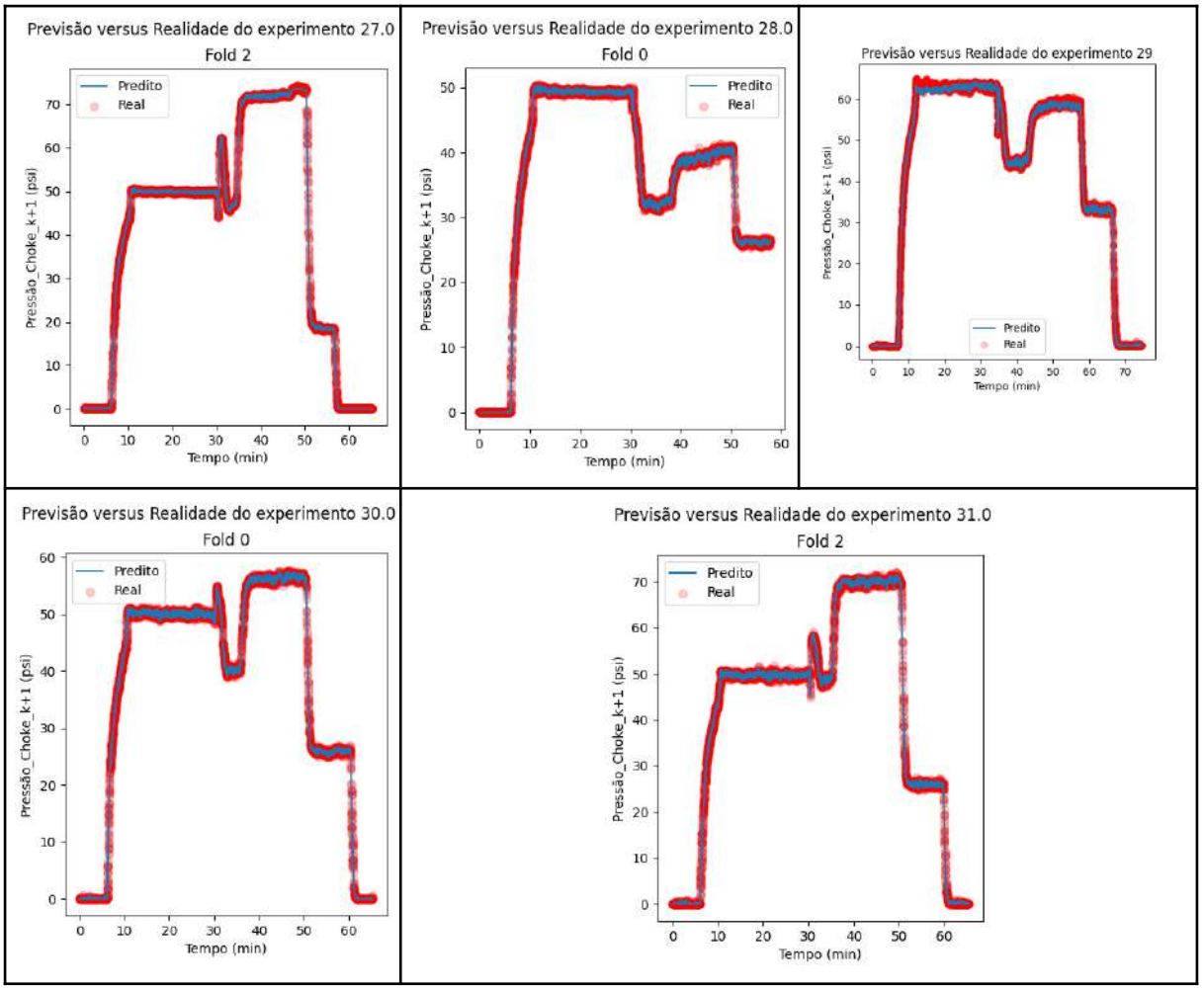

Figura 2*18* – Previsão e realidade dos experimentos de 27 a 31, excluindo as variáveis que apresentaram menor importância no aprendizado. Fonte: A autora.

Com relação à arquitetura do modelo, foram treinadas 512 árvores e os resultados demonstram que mesmo excluindo as variáveis menos importante, o modelo conseguiu prever satisfatoriamente.

ANEXO M – RESULTADO DOS DADOS EXPERIMENTAIS COM DADOS PASSADOS EM TODAS AS VARIÁVEIS

Foram aplicados 2 dados passados em todas as variáveis, sendo apenas o passo futuro na pressão *choke*, com escala de -4 a 4 e transformados com sig, com isso, são mostrados na Figura 219 as variáveis do modelo.

	Valores_Únicos	Tipo_de_Dado	mean	std	min	25%	50%	75%	max
Tempo_poco_k-1 (min)	169108	float64	0.40	0.33	0.02	0.08	0.30	0.72	0.98
Pressao_k-1 (psi)	195374	float64	0.44	0.36	0.02	0.10	0.35	0.87	0.98
Vazão_k-1 (m³/h)	148579	float64	0.04	0.09	0.02	0.02	0.02	0.03	0.98
Freq_Inversor_k-1 (Hz)	6	float64	0.09	0.20	0.02	0.02	0.02	0.02	0.98
Abertura_choke_k-1 (%)	198	float64	0.64	0.44	0.02	0.02	0.97	0.97	0.98
Vazão2_k-1 (m³/h)	147860	float64	0.04	0.09	0.02	0.02	0.02	0.03	0.98
Abertura_Valvula_Reservatorio_k-1 (%)	4	float64	0.39	0.43	0.02	0.02	0.14	0.98	0.98
Tempo_tanque_k-1 (min)	387369	float64	0.40	0.33	0.02	0.08	0.30	0.72	0.98
Pressao_Tanque_k-1 (psi)	353970	float64	0.54	0.38	0.02	0.12	0.58	0.94	0.98
Pressao_Choke_k-1 (psi)	176053	float64	0.41	0.35	0.02	0.03	0.33	0.78	0.98
Tempo_poco_k (min)	169122	float64	0.40	0.33	0.02	0.08	0.30	0.72	0.98
Pressao_k (psi)	195385	float64	0.44	0.36	0.02	0.10	0.35	0.87	0.98
Vazão_k (m³/h)	148588	float64	0.04	0.09	0.02	0.02	0.02	0.03	0.98
Freq_Inversor_k (Hz)	6	float64	0.09	0.20	0.02	0.02	0.02	0.02	0.98
Abertura_choke_k (%)	198	float64	0.64	0.44	0.02	0.02	0.97	0.97	0.98
Vazão2_k (m³/h)	147872	float64	0.04	0.09	0.02	0.02	0.02	0.03	0.98
Abertura_Valvula_Reservatorio_k (%)	4	float64	0.39	0.43	0.02	0.02	0.14	0.98	0.98
Tempo_tanque_k (min)	387392	float64	0.40	0.33	0.02	0.08	0.30	0.72	0.98
Pressao_Tanque_k (psi)	353973	float64	0.54	0.38	0.02	0.12	0.58	0.94	0.98
Pressao_Choke_k (psi)	176066	float64	0.41	0.35	0.02	0.03	0.33	0.78	0.98
Pressão_Choke_k+1 (psi)	155611	float64	35.60	26.45	0.00	4.40	36.31	60.21	94.68
experimento	32	int64	15.61	9.35	0.00	7.00	16.00	24.00	31.00

Figura *219* – Resumo do *dataframe* com dados experimentais e com 2 dados passados em todas as variáveis. Fonte: A autora.

Antes dos dados serem treinados pelo XGBoost, é feito a otimização dos hiperparâmetros com o Optuna gerando os seguintes parâmetros: 'n_estimators': 123, 'learning_rate': 0.1372, 'max_depth': 16, 'min_child_weight': 10, 'subsample': 0.9813, 'colsample_bytree': 0.8244, 'gamma': 0.3857, 'reg_alpha': 0.3826, 'reg_lambda': 0.7780, em apenas 232 segundos de execução, utilizando o mesmo ambiente de execução dos modelos anteriores.

Com os melhores parâmetros gerados pelo Optuna, os dados são treinados, validados e testados pelo XGBoost. As métricas de avaliação do modelo são apresentadas na Figura 220.

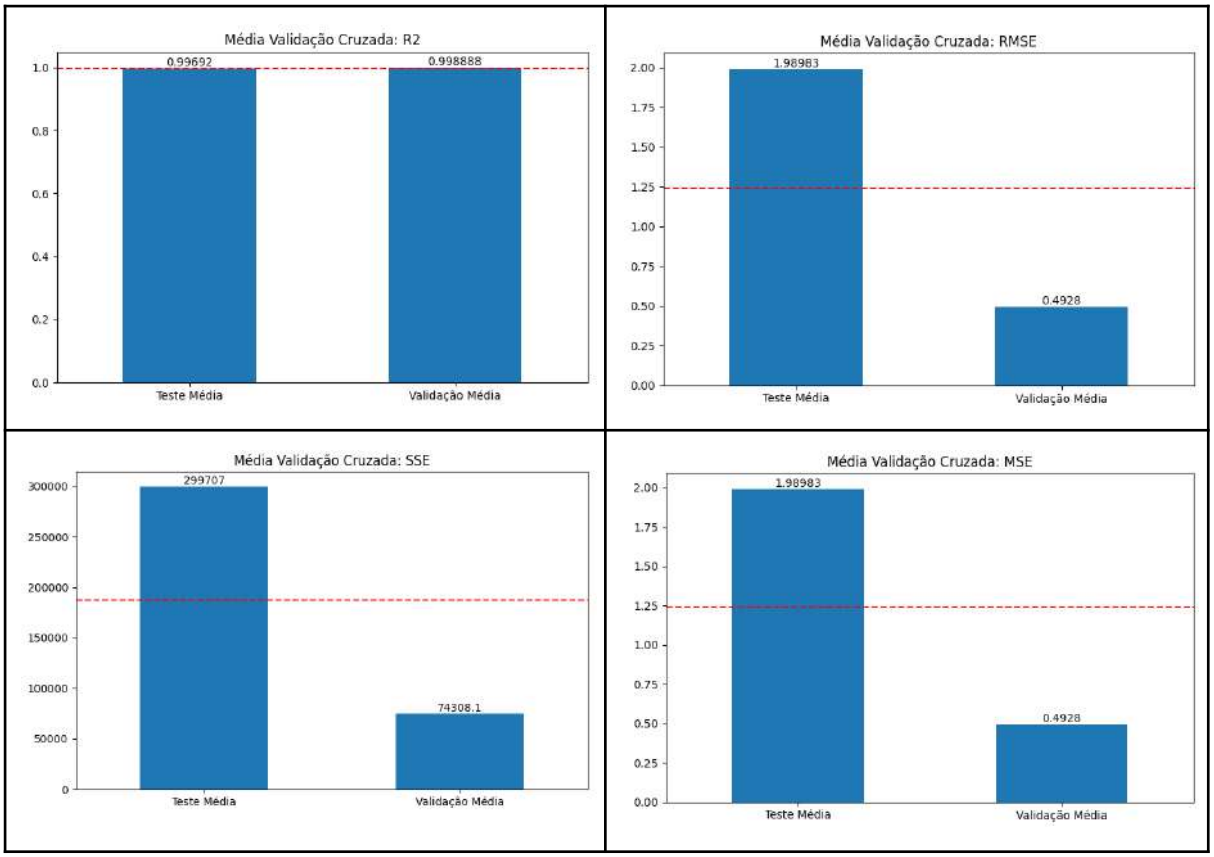

Figura *220* – Métricas de avaliação com dados experimentais e com 2 dados passados em todas as variáveis. Fonte: A autora.

São apresentados na Figura 221 os valores da densidade dos dados no teste e na validação do modelo, onde os dados preditos estão representados na cor vermelha e os dados reais na cor azul.

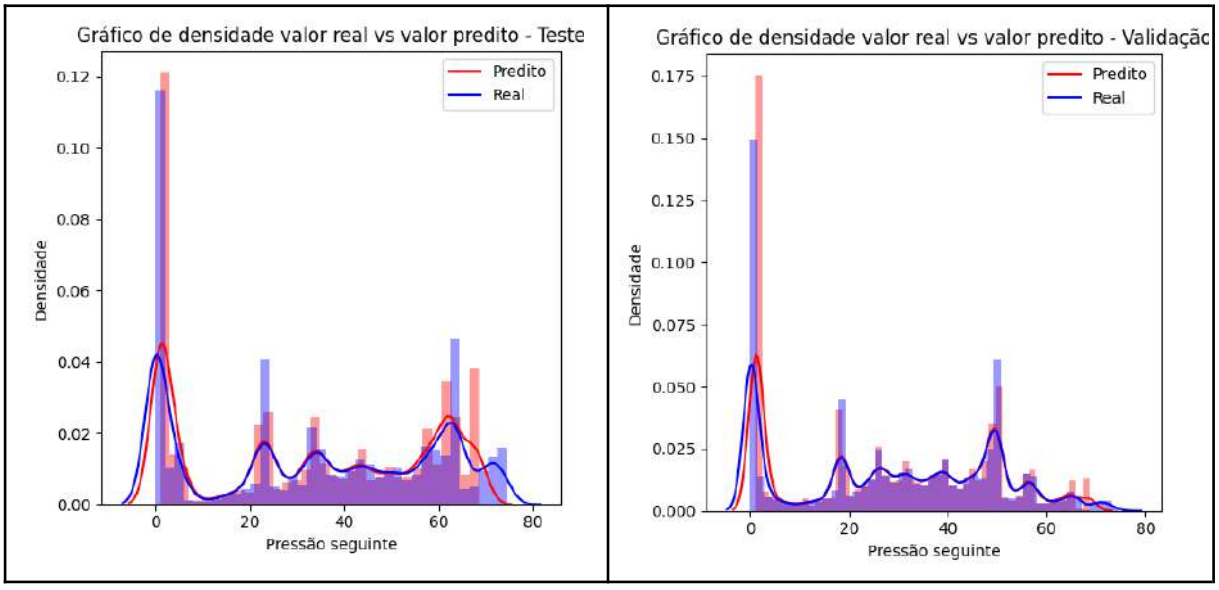

Figura 22*1* – Gráficos de densidade com dados experimentais e com 2 dados passados em todas as variáveis. Fonte: A autora.

São observados os valores das distribuições dos resíduos no teste e na validação do modelo, sendo apresentados na Figura 222.

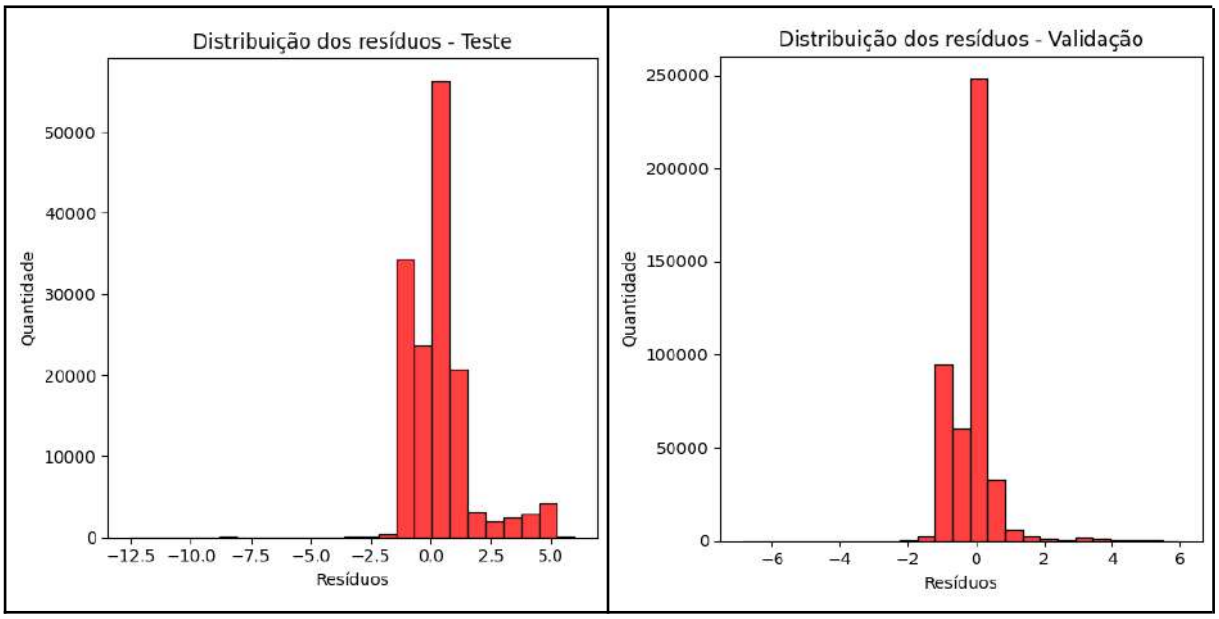

Figura *222* – Distribuição dos resíduos com dados experimentais e com 2 dados passados em todas as variáveis. Fonte: A autora.

São apresentados na Figura 223 os pontos na curva de comparação de valor real e valor predito, bem como os valores dos MSE, RMSE e R^2.

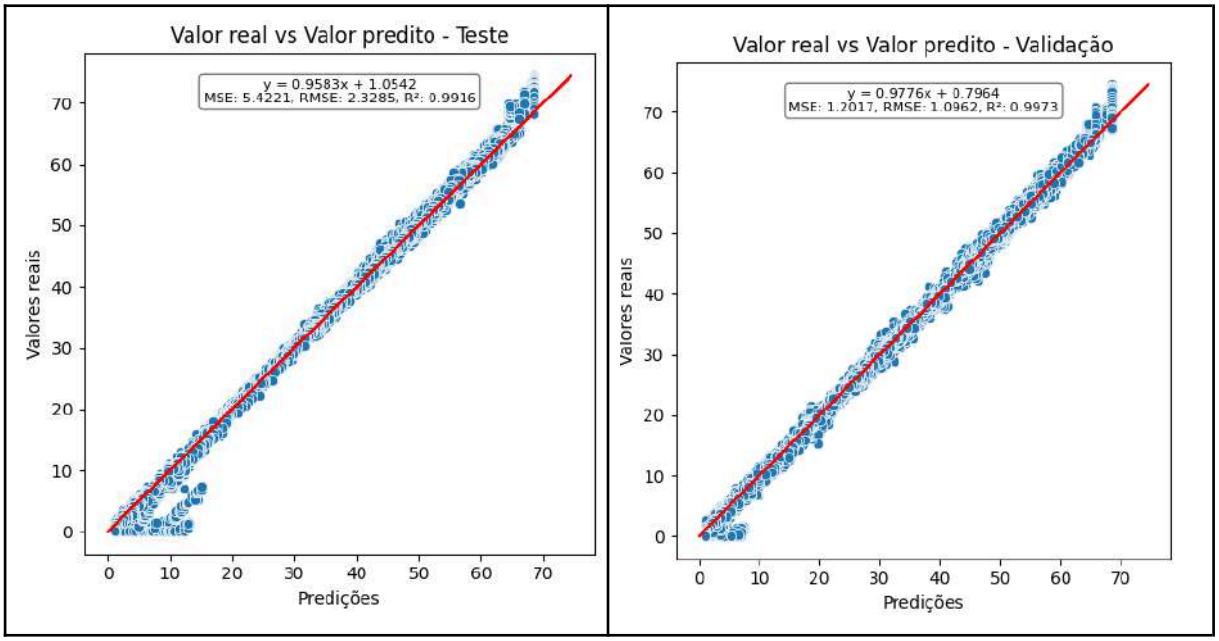

Figura *223* – Gráfico de evolução do modelo com dados experimentais e com 2 dados passados em todas as variáveis. Fonte: A autora.

É apresentado na Figura 224 o gráfico da função de perda do modelo.

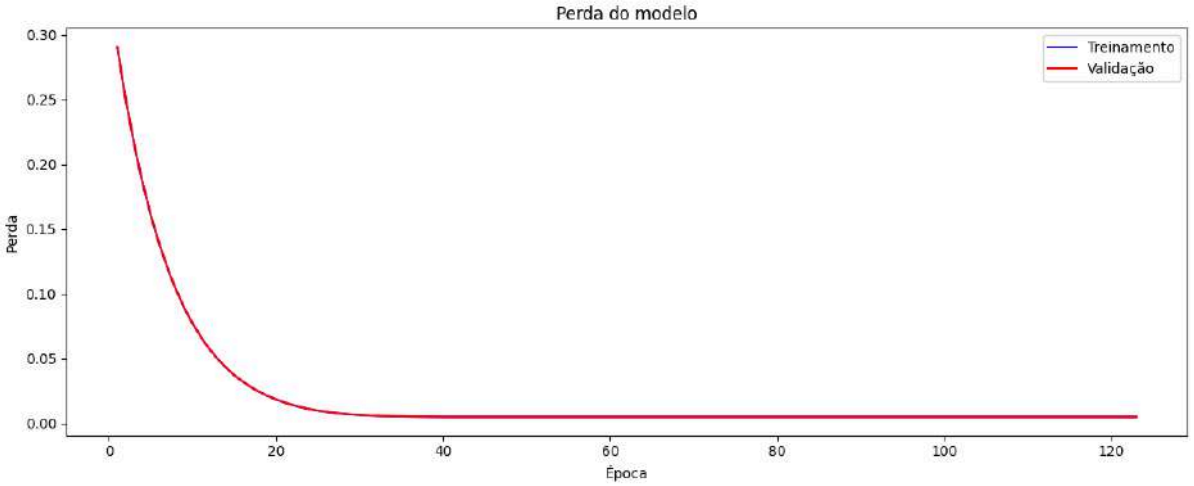

Figura *224* – Curva de perdas do modelo com dados experimentais e com 2 dados passados em todas as variáveis. Fonte: A autora.

As variáveis são mostradas na Figura 225, de acordo com sua importância para a previsão do modelo, sendo que, a pressão da *choke* no passo anterior apresenta maior importância no aprendizado do modelo.

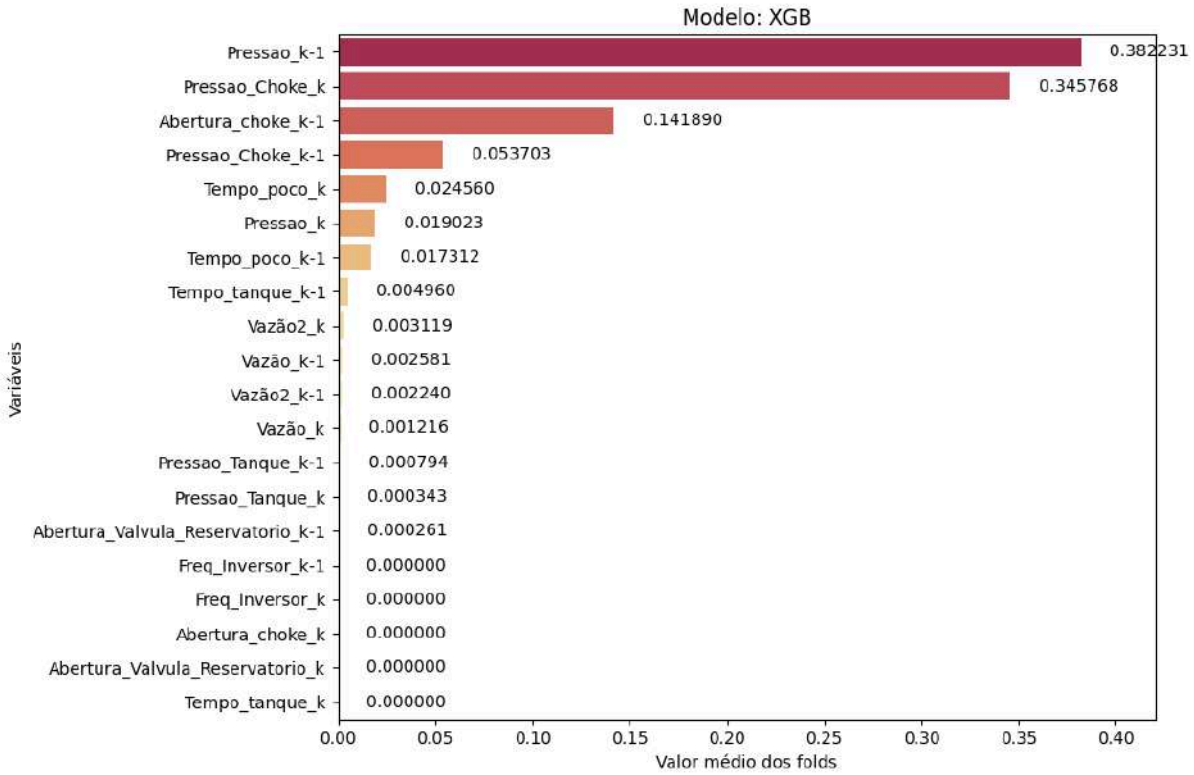

Figura *225* – Importância das variáveis do modelo com dados experimentais e com 2 dados passados em todas as variáveis. Fonte: A autora.

São apresentados nas Figuras 226 a 229, os resultados das previsões e realidade dos testes e validação do modelo, indicando o quanto o modelo consegue prever a operação de PMCD e *bullheading*.

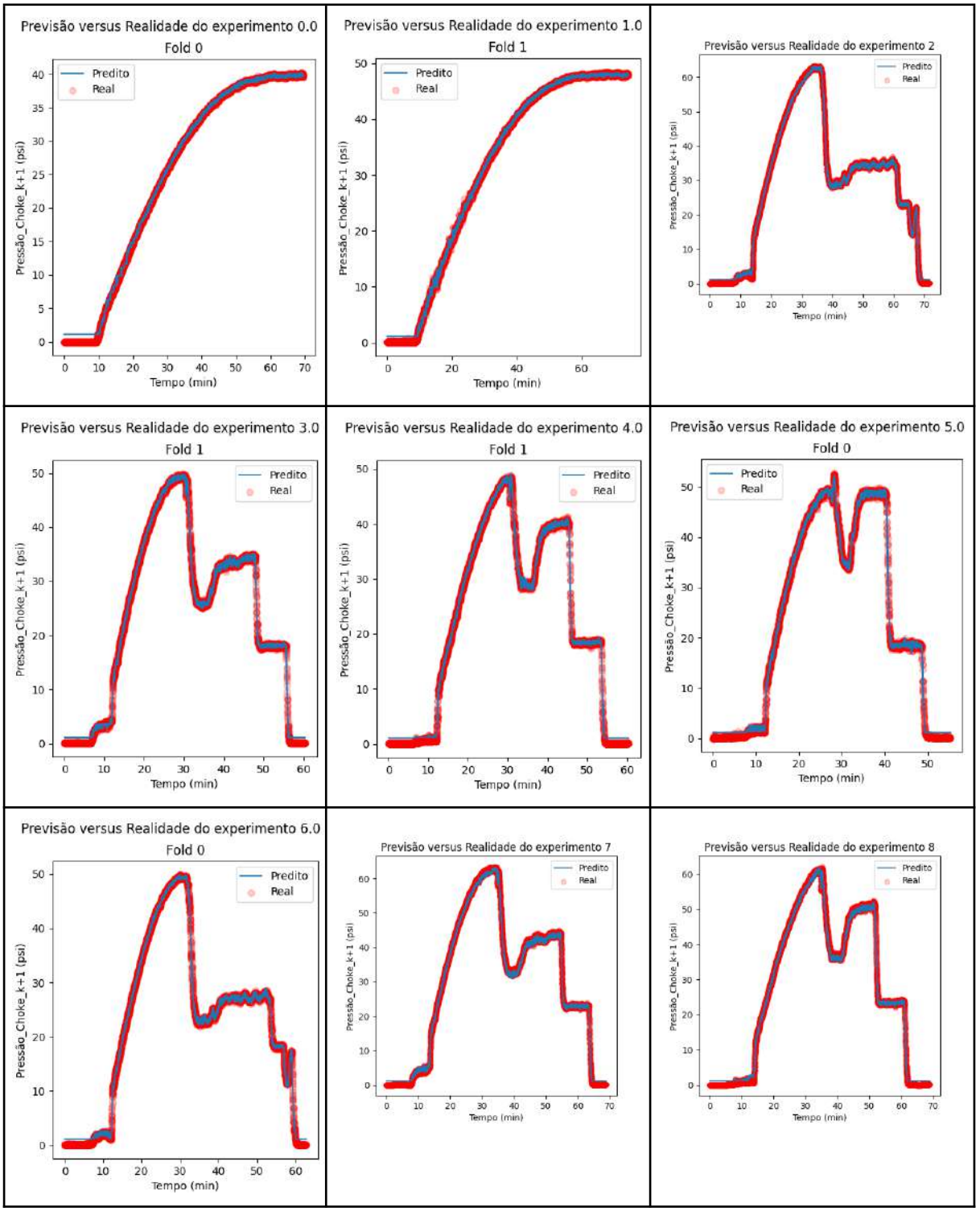

Figura 226 – Previsão e realidade dos experimentos de 0 a 8 do modelo com dados experimentais e com 2 dados passados em todas as variáveis. Fonte: A autora.

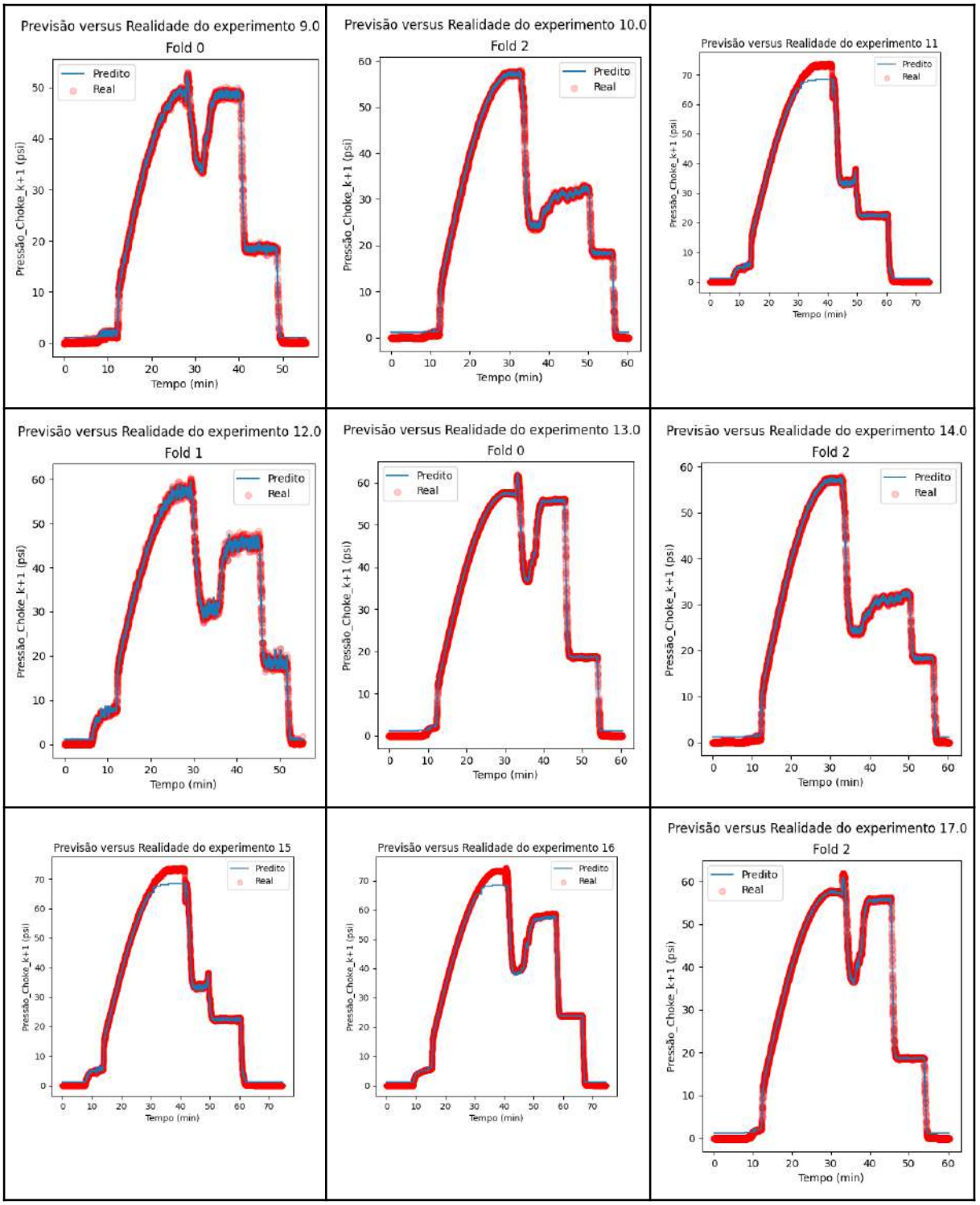

Figura 227 – Previsão e realidade dos experimentos de 9 a 17 do modelo com dados experimentais e com 2 dados passados em todas as variáveis. Fonte: A autora.

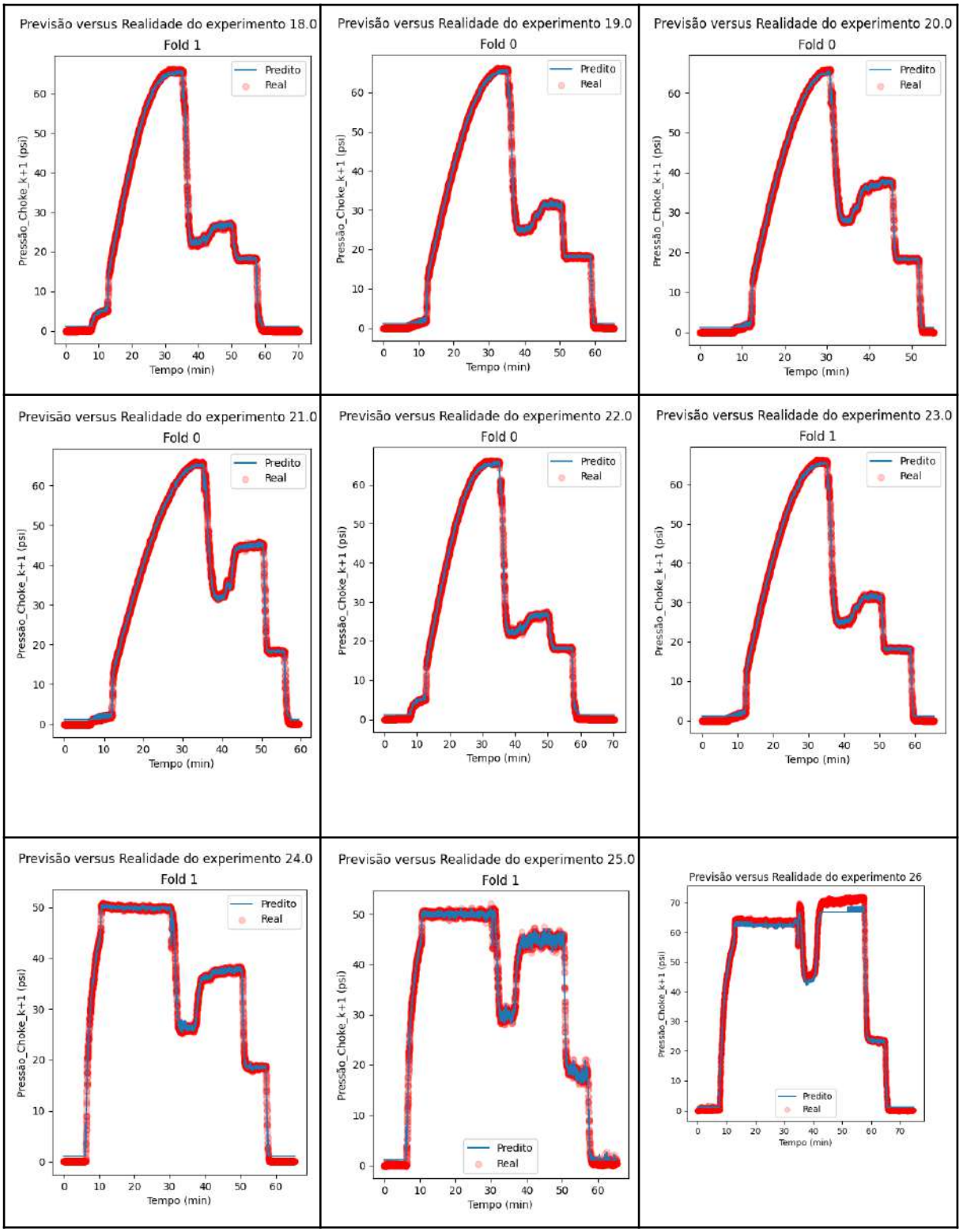

Figura 228 – Previsão e realidade dos experimentos de 18 a 26 do modelo com dados experimentais e com 2 dados passados em todas as variáveis. Fonte: A autora.

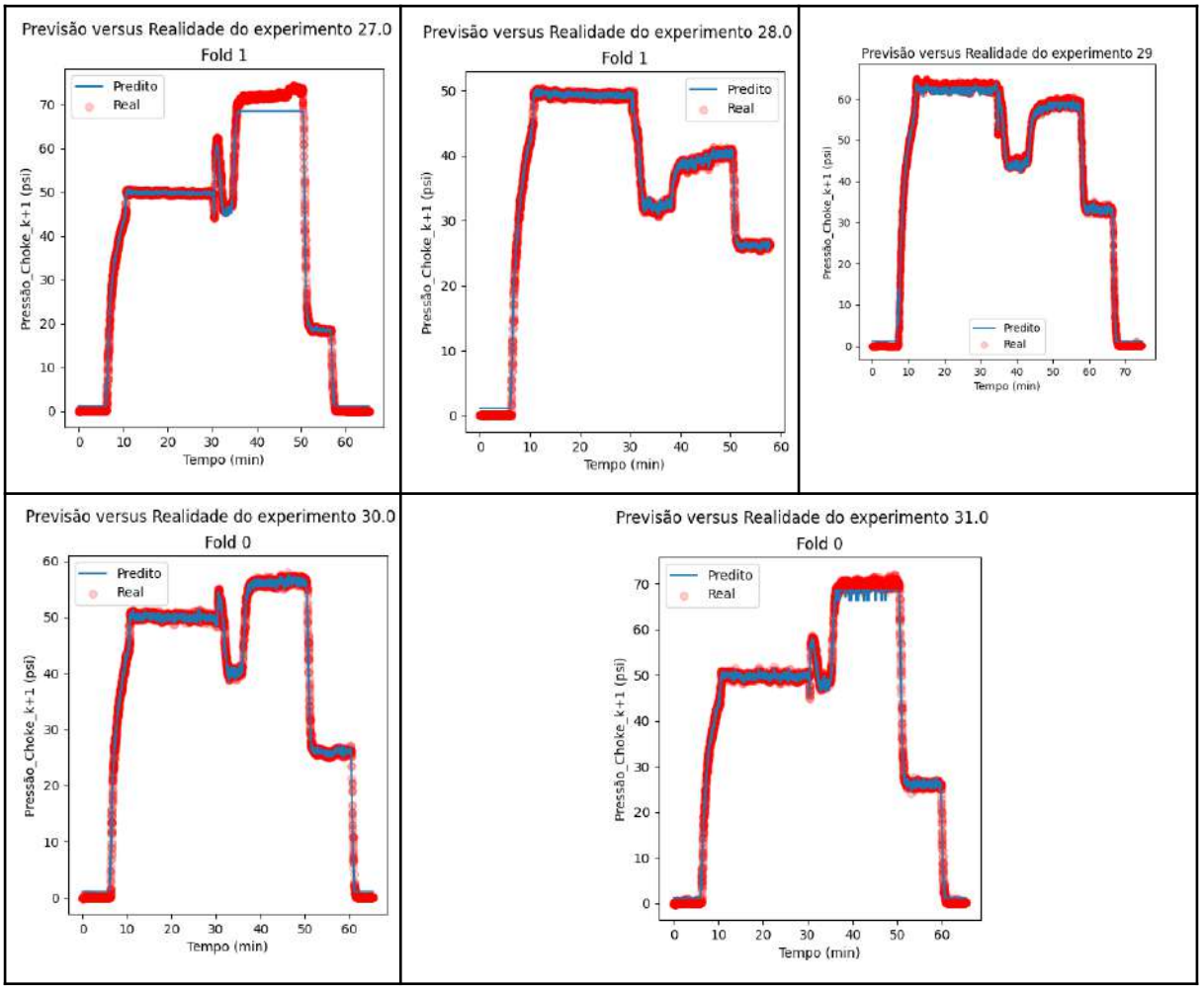

Figura 229 – Previsão e realidade dos experimentos de 27 a 31 do modelo com dados experimentais e com 2 dados passados em todas as variáveis. Fonte: A autora.

Com relação à arquitetura do modelo, foram treinadas 123 árvores e os resultados demonstraram que algumas variáveis defasadas teve influência no aprendizado do modelo, que fez a previsão com erros RMSE de 2.33 psi no teste e de 1.10 psi na validação.