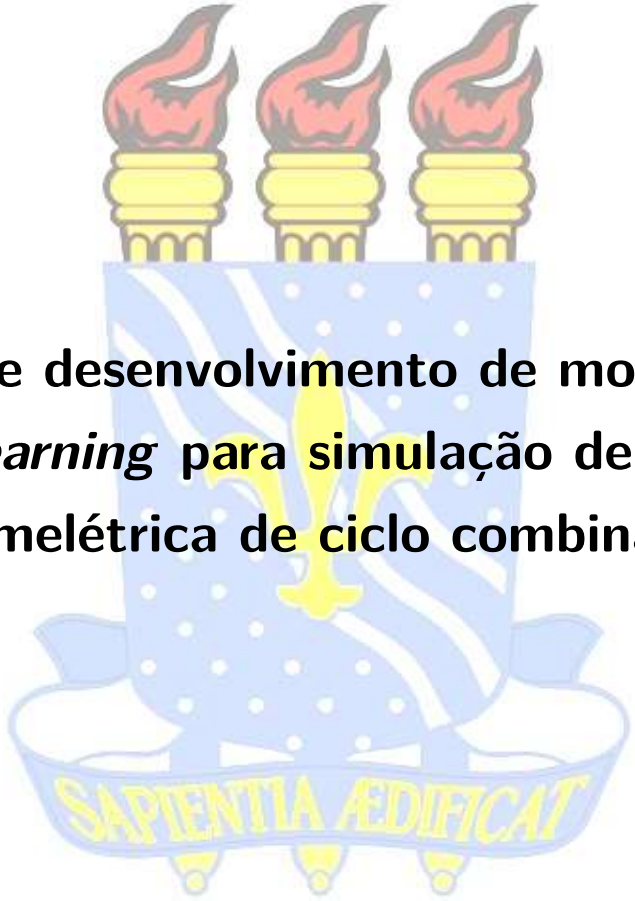
UNIVERSIDADE FEDERAL DA PARAÍBA
CENTRO DE ENERGIAS ALTERNATIVAS E RENOVÁVEIS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

Filipe Bulhões Froes



Análise e desenvolvimento de modelos de *machine learning* para simulação de uma usina termelétrica de ciclo combinado

João Pessoa
2026

Filipe Bulhões Froes

Análise e desenvolvimento de modelos de *machine learning* para simulação de uma usina termelétrica de ciclo combinado

Dissertação apresentada ao Programa de Pós Graduação em Engenharia Elétrica da Universidade Federal da Paraíba como exigência para a obtenção do título de Mestre em Engenharia Elétrica.

Universidade Federal da Paraíba

Centro de Energias Alternativas e Renováveis

Programa de Pós Graduação em Engenharia Elétrica

Orientador: Prof. Dr. Euler Cássio Tavares de Macedo

Coorientador: Prof. Dr. Juan Moises Maurício Villanueva

João Pessoa

2026

Catalogação na publicação
Seção de Catalogação e Classificação

F926a Froes, Filipe Bulhões.
Análise e desenvolvimento de modelos de machine learning para simulação de uma usina termelétrica de ciclo combinado / Filipe Bulhões Froes. - João Pessoa, 2026.
76 f.

Orientação: Euler Cássio Tavares de Macedo, Coorientação: Juan Moisés Maurício Villanueva.
Dissertação (Mestrado) - UFPB/CEAR.

1. Aprendizado de máquina. 2. Comparação de modelos. 3. Sistemas gêmeos digitais. 4. Usina termelétrica. I. Macedo, Euler Cássio Tavares de. II. Villanueva, Juan Moisés Maurício. III. Título.

UFPB/BC CDU 621.311:004.85(043)

Elaborado por KAROLLYNE FERREIRA DANTAS - CRB-4/2319

UNIVERSIDADE FEDERAL DA PARAÍBA – UFPB

CENTRO DE ENERGIAS ALTERNATIVAS E RENOVÁVEIS – CEAR

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA – PPGEE

A Comissão Examinadora, abaixo assinada, aprova a Dissertação

ANÁLISE E DESENVOLVIMENTO DE MODELOS DE MACHINE LEARNING PARA SIMULAÇÃO DE UMA USINA TERMELÉTRICA DE CICLO COMBINADO

Elaborada por

FILIPE BULHÕES FROES

como requisito parcial para obtenção do grau de

Mestre em Engenharia Elétrica.

COMISSÃO EXAMINADORA

Documento assinado digitalmente
gov.br EULER CASSIO TAVARES DE MACEDO
Data: 02/02/2026 14:32:20-0300
Verifique em https://validar.iti.gov.br

PROF. DR. EULER CÁSSIO TAVARES DE MACÊDO
Orientador – UFPB

Documento assinado digitalmente
gov.br JUAN MOISES MAURICIO VILLANUEVA
Data: 03/02/2026 00:56:19-0300
Verifique em https://validar.iti.gov.br

PROF. DR. JUAN MOISES MAURICIO VILLANUEVA
Coorientador – UFPB

Documento assinado digitalmente
gov.br CICERO DA ROCHA SOUTO
Data: 03/02/2026 10:48:46-0300
Verifique em https://validar.iti.gov.br

PROF. DR. CÍCERO DA ROCHA SOUTO
Examinador Interno – UFPB

Documento assinado digitalmente
gov.br LUIZ AFFONSO HENDERSON GUEDES DE OLIVEI
Data: 03/02/2026 09:14:46-0300
Verifique em https://validar.iti.gov.br

PROF. DR. LUIZ AFFONSO HENDERSON GUEDES DE OLIVEIRA
Examinador Externo – UFRN

João Pessoa/PB, 30 de janeiro de 2026

Dedico este trabalho à minha família e amigos.

AGRADECIMENTOS

Agradeço a Deus por proporcionar as condições necessárias para a realização deste trabalho, apesar de todos os percalços, do cansaço, das preocupações e dos novos desafios.

Agradeço também à minha família pelo apoio. Em especial à minha mãe, por ser minha base forte e inabalável; aos meus tios, tias, primos e avós, pelo apoio incondicional; ao meu pai, pela inspiração; e à minha madrinha, minha segunda mãe.

Agradeço a todos os meus amigos, novos e de longa data, pela motivação e pela paciência, que desta vez foram mais necessárias do que nunca. Àqueles que estão longe há tanto tempo, mas que sempre se fizeram presentes; aos que conheci durante a graduação; aos que conheci na pós-graduação; e aos que me acolheram como parte da família após a minha mudança.

Agradeço à instituição de ensino, à qual tenho orgulho de pertencer, e a todos que fazem parte dela. A todos os professores que acreditaram no meu trabalho e potencial, e que tiveram paciência para compreender as mudanças pelas quais passei no último ano.

Finalmente, agradeço a todos aqueles que contribuíram, direta ou indiretamente, para a realização deste trabalho: colegas e professores do GICA, meus orientadores, amigos e colegas de trabalho que incentivaram, trocaram ideias e me apoiaram quando mais precisei.

A todos, muito obrigado!

—

"Não podemos resolver problemas usando o mesmo tipo de pensamento que usamos quando os criamos."

(Albert Einstein)

RESUMO

A inteligência artificial (IA) tem se mostrado uma ferramenta poderosa para enfrentar os desafios da transição energética e da busca por maior eficiência nos sistemas de geração elétrica. No setor elétrico brasileiro, marcado pela crescente participação das usinas termelétricas a gás natural, técnicas de aprendizado de máquina (ML) podem auxiliar na otimização da operação dessas plantas, contribuindo para maior eficiência e redução de emissões. Nesse contexto, os sistemas gêmeos digitais (SGD) surgem como uma tecnologia promissora para simulação e apoio à tomada de decisão.

O presente trabalho teve como objetivo desenvolver e comparar diferentes modelos de aprendizado de máquina aplicados à simulação de uma usina termelétrica de ciclo combinado, em preparação para sua integração a um SGD. Foram exploradas metodologias de processamento de dados, treinamento, avaliação e comparação de modelos, incluindo algoritmos individuais e técnicas de conjunto como *bagging*, *boosting*, *stacking* e *blending*.

Os resultados obtidos demonstraram que modelos de regressão como SVR e GPR, bem como suas variantes em técnicas de conjunto, alcançaram melhor desempenho em termos de erro médio absoluto normalizado e coeficiente de determinação. Além disso, verificou-se que combinações de modelos podem fornecer maior robustez e precisão, indicando que abordagens híbridas são alternativas promissoras para representar o comportamento da usina em diferentes condições operacionais.

Palavras-chave: Aprendizado de máquina; Comparação de modelos; Sistemas gêmeos digitais; Técnicas de conjunto; Usinas termelétricas a gás natural.

ABSTRACT

Artificial intelligence (AI) has proven to be a powerful tool for addressing the challenges of energy transition and the search for greater efficiency in power generation systems. In the Brazilian electricity sector, marked by the growing participation of natural gas-fired thermoelectric power plants, machine learning (ML) techniques can help optimize the operation of these plants, contributing to greater efficiency and reduced emissions. In this context, digital twin systems (DTS) emerge as a promising technology for simulation and decision support.

The objective of this study was to develop and compare different machine learning models applied to the simulation of a combined cycle thermal power plant, in preparation for its integration into a DTS. Data processing, training, evaluation, and model comparison methodologies were explored, including individual algorithms and ensemble techniques such as bagging, boosting, stacking, and blending.

The results obtained demonstrated that regression models such as SVR and GPR, as well as their variants in ensemble techniques, achieved better performance in terms of normalized mean absolute error and coefficient of determination. In addition, it was found that combinations of models can provide greater robustness and accuracy, indicating that hybrid approaches are promising alternatives for representing the behavior of the plant under different conditions.

Keywords: Digital twin systems; Ensemble techniques; Machine learning; Model comparison; Natural gas thermoelectric power plant.

LISTA DE ILUSTRAÇÕES

Figura 1 – Topologia geral de um modelo *bagging*. 19
Figura 2 – Topologia geral de um modelo *boosting*. 20
Figura 3 – Topologia geral de um modelo *stacking*. 21
Figura 4 – Construção dos dados de treinamento do modelo meta: *stacking*. 22
Figura 5 – Topologia geral de um modelo *blending*. 23
Figura 6 – Construção dos dados de treinamento do modelo meta: *blending*. 24
Figura 7 – Sequência da Metodologia. 28
Figura 8 – Representação gráfica de uma usina de ciclo combinado. 29
Figura 9 – Processamento de dados. 33
Figura 10 – Esquema geral do modelo da potência da turbina a gás. 40
Figura 11 – Esquema geral do modelo da potência da turbina a vapor. 41
Figura 12 – Esquema geral do modelo da potência líquida. 41
Figura 13 – Matriz de correlação: Potência da turbina a gás. 42
Figura 14 – Matriz de correlação: Potência da turbina a vapor. 44
Figura 15 – Matriz de correlação: Potência líquida. 45
Figura 16 – Topologia do modelo hierárquico com ponderação adaptativa - modelo meta. 47
Figura 17 – Erro dos modelos: Potência da TG. 49
Figura 17 – Erro dos modelos: Potência da TG. 50
Figura 17 – Erro dos modelos: Potência da TG. 51
Figura 17 – Erro dos modelos: Potência da TG. 52
Figura 17 – Erro dos modelos: Potência da TG. 53
Figura 18 – Erro dos melhores modelos: Potência da TG. 54
Figura 19 – Erro dos modelos: Potência da TV. 56
Figura 19 – Erro dos modelos: Potência da TV. 57
Figura 19 – Erro dos modelos: Potência da TV. 58
Figura 19 – Erro dos modelos: Potência da TG. 59
Figura 19 – Erro dos modelos: Potência da TV. 60
Figura 20 – Erro dos melhores modelos: Potência da TV. 61
Figura 21 – Erro dos modelos: Potência líquida. 62
Figura 21 – Erro dos modelos: Potência líquida. 63
Figura 21 – Erro dos modelos: Potência líquida. 64
Figura 21 – Erro dos modelos: Potência líquida. 65
Figura 21 – Erro dos modelos: Potência líquida. 66
Figura 22 – Erro dos melhores modelos: Potência líquida. 67

LISTA DE TABELAS

Tabela 1 – Artigos que comparam modelos de ML: Regressão. 24
Tabela 2 – Resumo das métricas de avaliação de modelos. 27
Tabela 3 – Técnicas de ML testadas - Potência da TG 36
Tabela 4 – Descrição das variáveis selecionadas. 39
Tabela 5 – Potência da TG: Métricas dos modelos. 48
Tabela 6 – Potência da TV: Métricas dos modelos. 55
Tabela 7 – Potência Líquida: Métricas dos modelos. 62

SUMÁRIO

Lista de tabelas **8**

1 INTRODUÇÃO **11**
1.1 Objetivos **12**
1.2 Estrutura e organização do trabalho **13**

2 FUNDAMENTAÇÃO TEÓRICA **14**
2.1 Sistemas gêmeos digitais **14**
2.2 Tipos de modelos de ML **17**
2.2.1 Modelos simples 17
2.2.2 Modelos homogêneos 18
2.2.2.1 *Bagging* 18
2.2.2.2 *Boosting* 19
2.2.3 Modelos Heterogêneos 20
2.2.3.1 *Stacking* 20
2.2.3.2 *Blending* 22
2.3 Métricas e métodos de avaliação de modelos **24**
2.3.1 Métricas 25
2.3.2 Validação cruzada 27

3 METODOLOGIA **28**
3.1 Estudo de caso **28**
3.1.1 O ciclo combinado 29
3.1.2 Premissas 30
3.1.2.1 Caracterização do sistema 30
3.2 Seleção de algoritmos **31**
3.3 Processamento de dados **32**
3.4 Otimização de hiperparâmetros **35**
3.5 Comparação do desempenho **37**

4 RESULTADOS **38**
4.0.1 Análise do banco de dados 38
4.0.2 Correlação entre as variáveis 42
4.1 Técnicas de ML utilizadas **46**
4.1.1 O modelo de média dinâmica - Modelo hierárquico com ponderação adaptativa 46
4.2 Comparação de desempenho dos modelos **47**

4.2.1 Potência da TG 47
4.2.2 Potência da TV 54
4.2.3 Potência líquida 61

5 DISCUSSÃO E CONCLUSÕES 68
5.1 Trabalhos futuros 69

REFERÊNCIAS 70

1 INTRODUÇÃO

A crescente preocupação com o aquecimento global tem impulsionado a busca pelo aumento da eficiência da exploração dos recursos naturais, bem como a transição energética para fontes renováveis de energia. O relatório IPCC (2023) indica que no setor energético, a transição para sistemas de baixa emissão de carbono trará múltiplos benefícios, incluindo melhorias na qualidade do ar e na saúde. Existem sinergias potenciais entre o desenvolvimento sustentável e, por exemplo, a eficiência energética e as energias renováveis. Neste cenário, tecnologias como painéis fotovoltaicos, turbinas de geração eólica e termelétricas a biomassa têm aumentado significativamente sua participação na matriz energética no Brasil e no mundo.

Contudo, a inserção de fontes de energia solar e eólica na rede elétrica tem gerado problemas na operação dos sistemas de transmissão e distribuição de energia. Sendo fontes intermitentes, elas aumentam a complexidade operacional, dada a necessidade de transferências de grandes blocos de energia a longas distâncias. Uma solução para aumentar a segurança do nível de penetração de fontes renováveis intermitentes em uma determinada região ou sistema é a instalação de sistemas de armazenamento de energia (OLIVEIRA, 2020).

Infelizmente, esta solução ainda não pode ser adotada em larga escala, principalmente o armazenamento de energia em baterias, devido ao seu alto custo de desenvolvimento e instalação. Além disso, há a necessidade do aprimoramento de toda a sua cadeia produtiva para garantir que ela também seja sustentável (ANJOS, 2023).

Dessa forma, é notável que existe uma urgência por soluções que possam resolver ou mitigar os problemas enfrentados hoje pelo setor elétrico. Contraintuitivamente, o aumento da participação das fontes renováveis não suplanta a necessidade de utilização das termelétricas. Embora sejam mais poluentes, essas usinas têm sido fundamentais para garantir a segurança energética do país, não por conta da capacidade de geração das fontes renováveis, mas pela sua intermitência (G1, 2023; CAMPOS JR., 2024). Esse cenário não tem perspectiva de mudança em um futuro próximo, pois depende do desenvolvimento, amadurecimento e implementação de sistemas que possam substituí-las nesta função, como é o caso de sistemas de armazenamento de energia em baterias. Uma alternativa que tem gerado muito interesse, em especial no Brasil, é a substituição gradativa de combustíveis como carvão e díesel pelo gás natural, menos poluente. A previsão é que até 2050 a participação do gás natural na matriz elétrica brasileira aumente, muito em função da expectativa de sua maior disponibilidade a preços competitivos (BRASIL, 2020).

Além da substituição da tecnologia, adotando o gás natural como combustível, uma

alternativa para reduzir ainda mais as emissões de carbono é a adoção de sistemas de ciclo combinado em detrimento de sistemas convencionais (TEIXEIRA et al., 2021). Outra forma de aumentar a eficiência desses sistemas de geração é por meio de uma operação otimizada. Isso porque a eficiência desses sistemas é variável, alcançando seus índices máximos em condições nominais de carga e temperatura ambiente. O relatório IPCC (2023) aponta que tecnologias como inteligência artificial (IA), aprendizado de máquina (*machine learning* - ML) e internet das coisas podem auxiliar nessa tarefa.

Estando entre uma das tecnologias com maior potencial de transformação para as mais diversas áreas, os Sistemas Gêmeos Digitais (SGD) foram listados como uma das dez tendências tecnológicas nos anos de 2017 a 2019 (SHARMA et al., 2022). De 2019 em diante, o termo desapareceu da lista; contudo, termos como IA adaptativa, engenharia de IA, aplicações inteligentes, inteligência de decisão e hiperautomação têm sido citados nesta lista até os dias de hoje (PANETTA, 2019; PANETTA, 2020; COSTELLO; RIMOL, 2021; RIMOL; HOWLEY, 2022; MCCARTNEY, 2023), sendo tecnologias intimamente ligadas aos SGDs.

Assim, os SGDs têm um potencial muito grande para auxiliar na redução das emissões de carbono no curto e médio prazo, otimizando a operação de sistemas de geração termelétrica. Contudo, esta tecnologia ainda carece de procedimentos de desenvolvimento de modelos aplicados a simulações bem estruturados, principalmente para SGDs que utilizam técnicas de IA (SHARMA et al., 2022).

1.1 OBJETIVOS

Pelo exposto, o objetivo deste trabalho é desenvolver um estudo comparativo de desempenho de modelos virtuais baseados em ML para aplicação em um SGD de uma usina termelétrica de ciclo combinado.

Além do objetivo principal, o presente trabalho tem os objetivos específicos de:

- Explorar uma metodologia de desenvolvimento e comparação de modelos de *machine learning* aplicados aos SGDs.
- Identificar as características dos sistemas que os modelos virtuais representam e suas restrições.
- Revisar e definir as métricas e métodos utilizados para comparar o desempenho dos modelos de ML.
- Explorar um novo método de agregação para modelos de conjunto: médias ponderadas com pesos dinâmicos utilizando Redes Neurais Artificiais.

1.2 ESTRUTURA E ORGANIZAÇÃO DO TRABALHO

O presente trabalho está organizado da seguinte forma:

- No Capítulo 1, está disposta a problemática e as alternativas que serão abordadas no estudo para solucioná-la. Além disso, são apresentados os objetivos principais e secundários do trabalho desenvolvido;
- No Capítulo 2, são abordados os principais conceitos que serão utilizados ao longo do desenvolvimento do trabalho. Neste capítulo, também é feita a análise da usina e dos dados disponibilizados por ela, além da caracterização dos sistemas que a compõe;
- No Capítulo 3, são descritos os procedimentos e ferramentas adotadas no desenvolvimento do trabalho;
- No Capítulo 4, são expostos os resultados obtidos no estudo, assim como, a avaliação destes resultados;
- Finalmente, no Capítulo 5, apresentam-se as considerações finais do presente estudo e sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Para nortear o presente estudo, foram analisados trabalhos relacionados aos temas SGD, aprendizado de máquina e desempenho de modelos de regressão. Além disso, foi feita uma revisão a respeito da usina emulada pelos modelos desenvolvidos ao longo deste trabalho.

2.1 SISTEMAS GÊMEOS DIGITAIS

A primeira vez que o conceito de sistemas gêmeos digitais (SGD), popularmente chamados apenas de gêmeos digitais (GD), surgiu na literatura foi em 2003, sendo na época um conceito bastante limitado, visto que era descrito apenas como a representação virtual de um objeto no mundo real. Na época, a coleta de dados era pouco automatizada e majoritariamente feita de forma manual (GRIEVES, 2015). Com o advento da era digital e o avanço nos sistemas de coleta, bem como na capacidade de armazenamento e processamento de dados, o conceito de GD tem se desenvolvido e se espalhado, sendo aplicado em diversas áreas atualmente (LIU et al., 2021).

De acordo com Kreuzer, Papapetrou e Zdravkovic (2024), SGDs são representações virtuais de sistemas físicos reais que replicam o comportamento desses sistemas de forma acurada e precisa no ambiente virtual. Estes fazem parte de sistemas chamados Sistemas gêmeos digitais (SGD), sendo eles a parte mais importante desses sistemas. Esses sistemas são desenvolvidos com o objetivo de monitoramento, simulação e otimização de processos, detecção de falhas, entre outras funcionalidades.

Uma definição mais ampla é fornecida no trabalho de Wright e Davidson (2020), que indica as três principais características de um SGD, sendo elas:

- Um ou mais modelos virtuais.
- Um conjunto de dados, relacionados ao objeto modelado, que seja dinâmico e em expansão.
- Uma forma de atualizar os modelos virtuais por meio dos dados coletados para que o sistema se mantenha fiel à sua contraparte real.

Ainda segundo os autores Wright e Davidson (2020), o que diferencia um modelo virtual de um SGD é a sua capacidade de adaptar seu comportamento para que permaneça fiel ao modelo real por meio dos dados e a possibilidade de conter diversos modelos que irão

compor o SGD. Isso facilita a sua concepção e implementação, além de tornar a solução customizável e adaptável, de acordo com a necessidade.

Aprofundando um pouco mais a respeito dos SGDs, é possível citar três níveis de integração: os modelos digitais, as sombras digitais e os gêmeos digitais. Os modelos digitais são atualizados de forma manual, de forma que não há uma conexão direta dos dados com o modelo. As sombras digitais, por sua vez, têm um fluxo de dados automático do sistema físico para o ambiente virtual, porém o fluxo de dados do sistema virtual para o sistema físico é feito manualmente. Finalmente, os gêmeos digitais têm o fluxo de dados entre o modelo físico e o virtual de forma totalmente automática. Assim, os SGDs podem se manter atualizados e atuar na planta física sem a necessidade da interferência humana (KRITZINGER et al., 2018).

Essa nomenclatura causa uma certa confusão na literatura, sendo conflitante a primeira vista, a exemplo dos trabalhos dos autores supracitados. Contudo, o que há em comum na definição dos SGDs é a sua capacidade de atualização, mantendo-se fiéis aos modelos reais. Isso ainda é válido para as três categorias citadas por Kritzinger et al. (2018), diferindo entre si apenas pela forma de atualização e atuação no sistema físico.

Os SGDs são soluções bastante particulares, customizadas especificamente para o problema que desejam solucionar, com aplicações na indústria e robótica (ATANASYAN et al., 2022; ZHENG; YANG; CHENG, 2018; MENDI, 2022), logística (MARMOLEJO-SAUCEDO, 2020), agropecuária (VERDOUW et al., 2021), além de diversas outras.

No âmbito dos sistemas de geração termelétrica, Lim et al. (2021) desenvolveram um sistema gêmeo digital capaz de monitorar o desempenho, bem como prever carga e desempenho, detectar e prever falhas. Isso, de forma integrada ao sistema de monitoramento e diagnóstico existente na usina. Uma limitação da abordagem deste trabalho é que a implementação da metodologia apresentada necessita que os equipamentos possuam modelos virtuais (modelos físicos) pré-existentes.

De forma similar, o trabalho de Panov e Cruz-Manzo (2020) também utiliza modelos dinâmicos não lineares no SGD desenvolvido. Este foi integrado ao sistema de controle da planta, via controlador lógico programável (CLP), para auxiliar no controle, monitoramento e detecção de falhas da planta.

A fim de reduzir o tempo de desenvolvimento dos SGDs, eliminando a necessidade do desenvolvimento de modelos baseados em sistemas de equações não lineares, foram desenvolvidos sistemas que utilizam modelos de aprendizado de máquina, ao invés de modelos físicos. Algumas das vantagens desta abordagem são o reconhecimento de padrões, o auxílio à tomada de decisão e a capacidade de ajuste aos dados sem a necessidade de fornecer um modelo físico detalhado. Contudo, também possui alguns desafios em seu desenvolvimento e utilização, sendo os principais deles o comprometimento da interpre-

tabilidade, a robustez e as simplificações e premissas na criação dos modelos (SHEN; ARRAÑO-VARGAS; KONSTANTINOU, 2023).

No trabalho de Junior et al. (2021), foi desenvolvido um SGD para otimizar a operação do sistema de arrefecimento de uma termelétrica utilizando um sistema de inferência *Fuzzy* com geração automática de regras. O sistema conseguiu alcançar uma média de erro de predição de menos de cinco por cento, cerca de três graus celsius, sendo atualizado automaticamente durante a operação.

Outra implementação de SGD, desta vez com redes neurais artificiais, foi o trabalho de Huang et al. (2022), que desenvolveu um SGD para um Sistema de Cogeração de Energia Elétrica, Térmica e de Refrigeração com Recuperação de Energia de Refrigeração. O objetivo do SGD foi de emular a degradação do sistema real a fim de fornecer as melhores condições de operação para o estado atual do sistema. Com a solução desenvolvida, os autores conseguiram um aumento na taxa diária de economia primária de energia de respectivamente 2,23%, 0,35% e 1,53% para os períodos do inverno, verão e dias de transição, quando o compressor e a turbina a gás operam com uma eficiência reduzida em até 2%.

Em suma, o uso de sistemas gêmeos digitais para monitorar e otimizar sistemas tem se mostrado cada vez mais relevante, impulsionando uma crescente pesquisa na área. No entanto, a falta de uma definição universal e de padrões de desenvolvimento claros ainda representa um obstáculo para a ampla adoção dessa tecnologia (SHARMA et al., 2022). Na geração de energia elétrica, isso se torna ainda mais evidente, principalmente para sistemas que utilizam modelos de aprendizado de máquina em suas arquiteturas, sendo necessária a implementação de uma etapa de validação da operação do sistema para garantir a robustez da solução (SHEN; ARRAÑO-VARGAS; KONSTANTINOU, 2023).

Sendo assim, neste trabalho será adotado o termo Sistema Gêmeo Digital em referência a qualquer uma das três classes: modelo, sombra ou gêmeo digital, seguindo a nomenclatura adotada por Kritzinger et al. (2018) e Kreuzer, Papapetrou e Zdravkovic (2024). Além disso, vale ressaltar a diferença entre o modelo virtual e o modelo digital. Serão adotadas a partir daqui as nomenclaturas modelo virtual em referência aos modelos matemáticos ou de ML, utilizados para modelar um sistema físico no ambiente virtual, e modelo digital em referência à classe de SGDs cuja interface com o sistema físico real acontece de forma totalmente manual. Observe que o primeiro pode ser parte integrante do segundo, como citado por Wright e Davidson (2020). Assim, o termo gêmeo digital será utilizado apenas em referência à classe de SGDs que têm sua interface com o modelo real totalmente automatizada.

Técnicas de ML e IA são indispensáveis para liberar o completo potencial dos Sistemas Gêmeos Digitais. Elas podem ser utilizadas tanto para modelagem e simulação

de sistemas físicos quanto na assistência à tomada de decisão, aplicadas à análise de dados (EMMERT-STREIB, 2023). As principais aplicações de ML em SGDs são a simulação e a classificação para tomada de decisão, enquanto suas principais características são a capacidade de aprendizado contínuo e a representação de sistemas complexos. Biller e Biller (2023), por exemplo, implementaram em seu trabalho dois SGDs, sendo eles os SGDs de uma fábrica e de uma cadeia de suprimentos. O primeiro utiliza técnicas de ML para simulação, enquanto o segundo utiliza aprendizado por reforço para otimização de processos.

Outro exemplo de implementação de um SGD é o trabalho de Perno, Hvam e Haug (2023). Neste trabalho, os autores desenvolvem um SGD baseado em ML para prever parâmetros críticos de processos de uma linha de produção, em tempo real.

Para que o SGD seja fiel ao sistema real, é necessário que os modelos que o compõem repliquem corretamente as relações matemáticas entre as variáveis medidas e as preditas, cujo comportamento se deseja replicar. Assim, fica evidente a importância da exatidão dos modelos na implementação dos SGD. Portanto, parte importante do processo de desenvolvimento de um SGD baseado em IA é a escolha da técnica de ML utilizada, o que impactará diretamente no seu desempenho e forma de utilização.

2.2 TIPOS DE MODELOS DE ML

Para simulação, há uma grande quantidade de técnicas de modelagem disponíveis que podem ser utilizadas tanto para classificação quanto para regressão linear. Tais técnicas podem ser divididas em três grupos principais: modelos simples, modelos homogêneos e modelos heterogêneos, sendo as duas últimas classes formadas por uma combinação de modelos, popularmente conhecidas como *ensemble models* - modelos de conjunto, em tradução livre (KOTU; DESHPANDE, 2019).

Os modelos de conjunto, contrariando a Navalha de Occam, também conhecida como “Lei da Parcimônia”, que postula que a melhor solução é sempre a mais simples, superam em desempenho os modelos simples na grande maioria dos estudos comparativos (ELDER, 2018).

2.2.1 MODELOS SIMPLES

São algoritmos que utilizam um banco de dados para gerar um modelo lógico ou matemático que captura e replica as relações entre as variáveis presentes neste banco de dados, sem ser programado diretamente para esta tarefa (SARKER, 2021). Alguns algoritmos representantes da classe de modelos simples são as redes neurais artificiais (RNA), as máquinas de vetor de suporte (*Support Vector Machine* - SVM e *Support Vector*

Regression - SVR), árvores de decisão (*Decision Trees* - DT) e modelos probabilísticos como modelos de misturas gaussianas (*Gaussian Mixture Models* - GMM) e regressor de processo gaussiano (*Gaussian Process Regressor* - GPR).

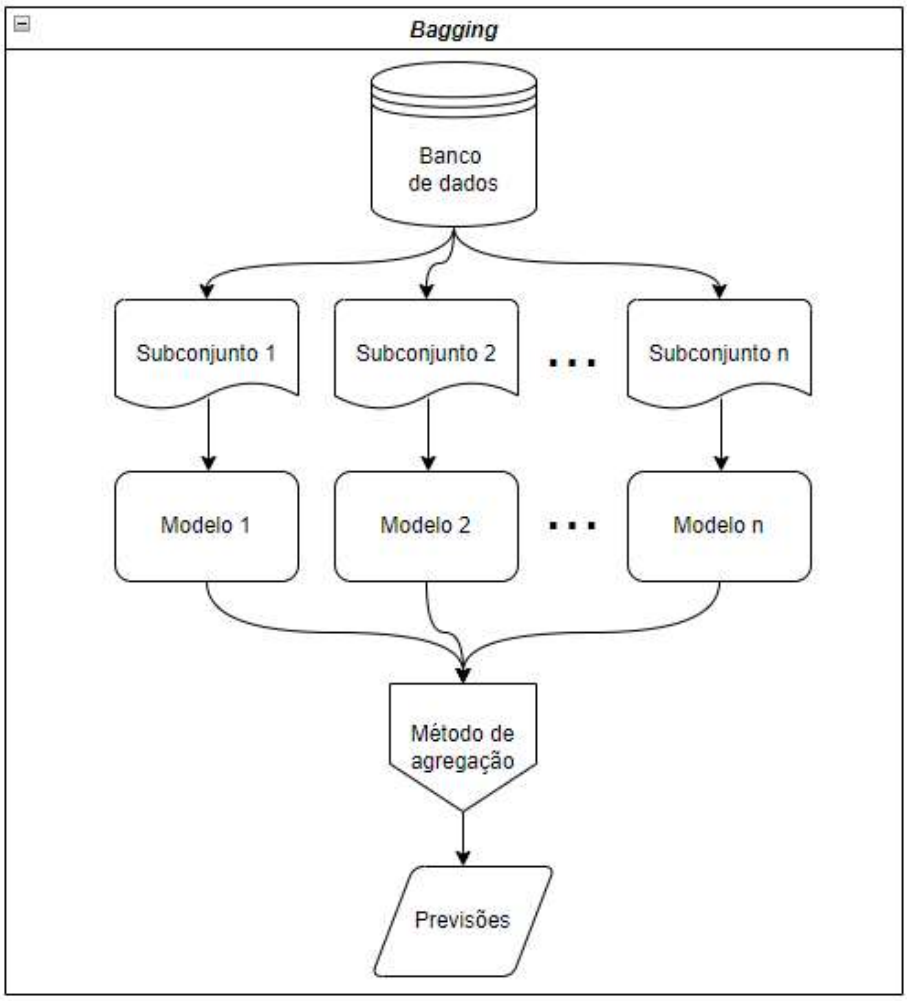
2.2.2 MODELOS HOMOGÊNEOS

Modelos de conjunto homogêneos são bastante famosos e amplamente utilizados. Em sua grande maioria, os modelos homogêneos utilizam árvores de decisão como modelos base, porém isso não é uma regra. São subdivididos nas classes *Bagging* e *Boosting*.

2.2.2.1 *BAGGING*

Bagging é uma técnica que utiliza vários modelos simples, de mesma arquitetura, em paralelo. Cada modelo recebe uma amostra de dados aleatória, conhecida como amostra de *Bootstrap*, durante o treinamento, de forma que cada um deles aprenderá nuances diferentes dos dados. A previsão final é dada por meio de algum método de agregação como votação simples, para modelos de classificação ou média simples, para modelos de regressão. Esta técnica tende a reduzir o *overfitting* e a variância dos dados ao combinar as previsões dos modelos treinados em diferentes subconjuntos de dados (CHAURASIA; PANDEY; PAL, 2022). Na Figura 1, está disposta a topologia que exemplifica o funcionamento desta técnica. As Florestas aleatórias (*Random Forests* - RF) são representantes bem conhecidos desta classe de modelos.

Figura 1 – Topologia geral de um modelo *bagging*.

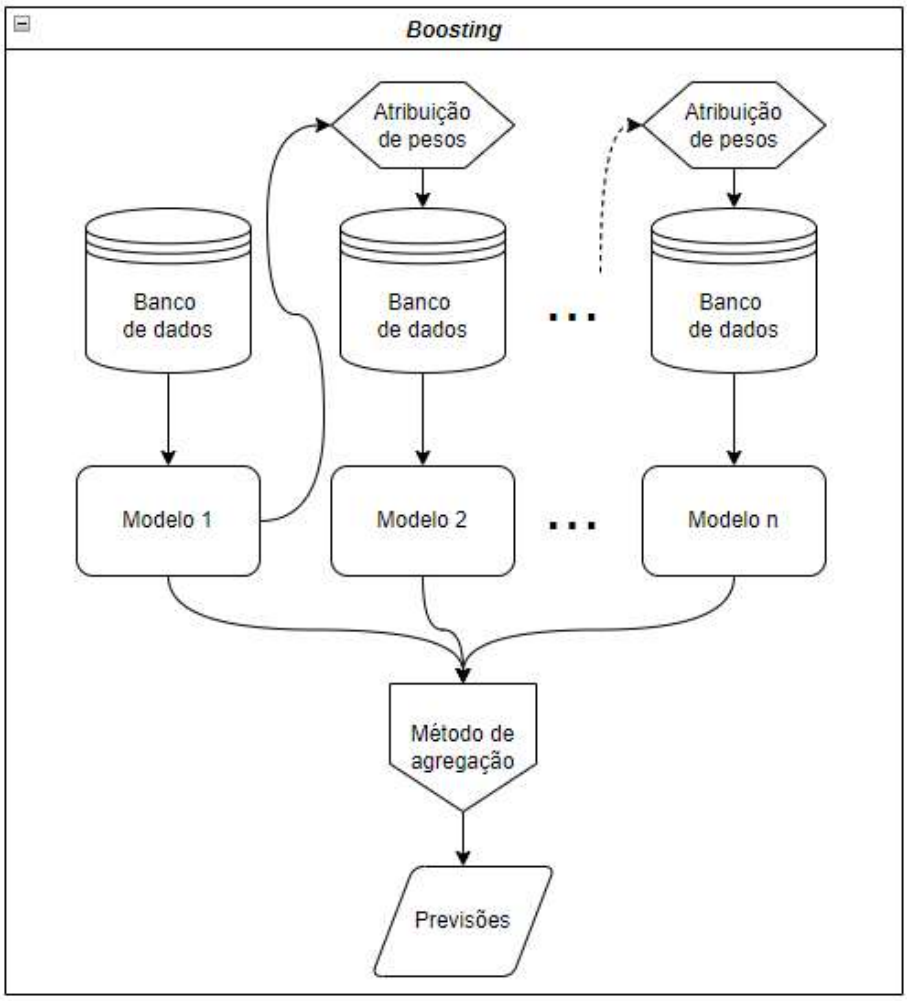


Fonte: De autoria própria.

2.2.2.2 *BOOSTING*

Boosting, por sua vez, é uma técnica que utiliza vários modelos simples, porém treinando-os de forma sequencial. A cada modelo treinado, são atribuídos pesos aos dados de forma que os modelos seguintes se especializem na porção dos dados que os modelos anteriores não tiveram um bom desempenho, melhorando assim, o desempenho do conjunto como um todo. Os métodos de agregação mais utilizados com esta técnica são a votação e a média ponderadas, para classificação e regressão, respectivamente (MAHAJAN et al., 2024). Na Figura 2 é possível observar o esquemático que ilustra o funcionamento desta classe. Existem vários *frameworks* bastante conhecidos que utilizam esta técnica, como XGBoost, CatBoost, LightGBM e AdaBoost.

Figura 2 – Topologia geral de um modelo *boosting*.



Fonte: De autoria própria.

2.2.3 MODELOS HETEROGÊNEOS

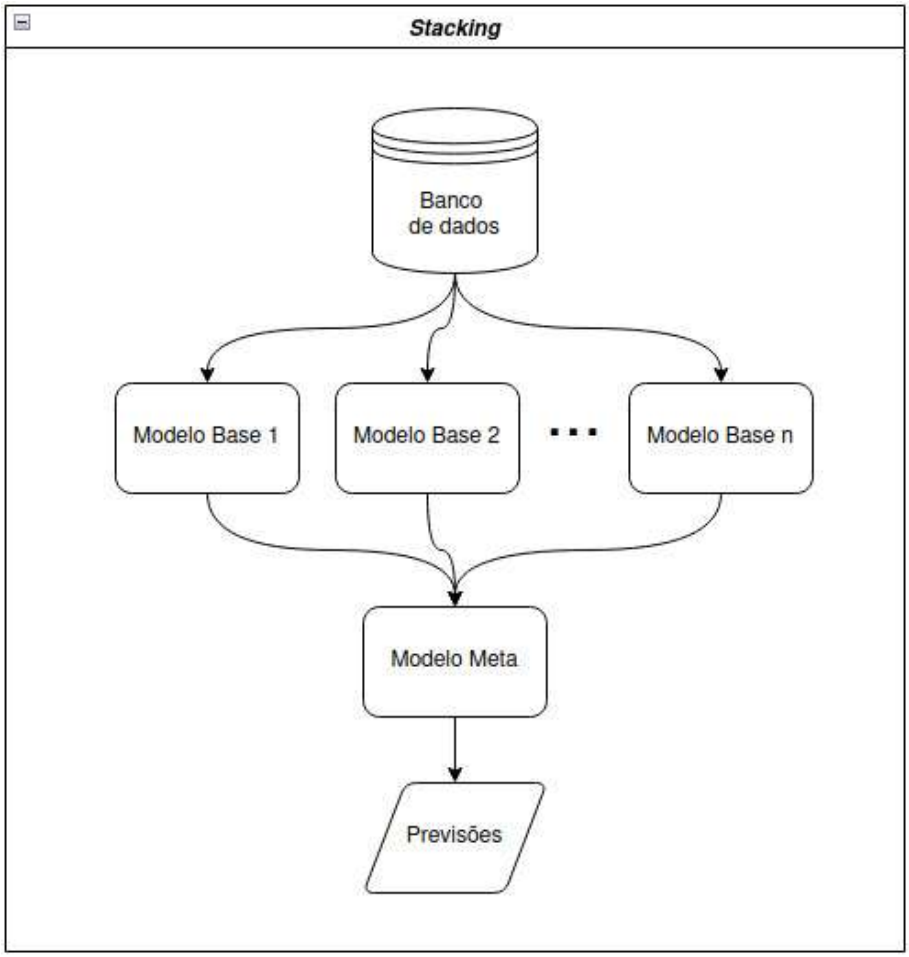
Finalmente, modelos de conjunto heterogêneos, embora mais complexos, têm o potencial para ser a classe de modelos com melhor desempenho dentre as três citadas. Eles combinam as previsões feitas por vários modelos simples de diferentes algoritmos (modelos base) por meio de um preditor final (modelo meta) (HASAN; HASAN, 2019). Esta classe de algoritmos não tem uma fórmula pronta, ficando a cargo do desenvolvedor escolher os modelos base e meta que serão utilizados. O treinamento e desenvolvimento do modelo pode ser feito de duas formas diferentes, sendo elas: *Stacking* e *Blending*.

2.2.3.1 *STACKING*

São modelos que utilizam apenas as previsões dos modelos base como entradas do modelo meta. Dessa forma, os modelos stacking também podem ser descritos como o empilhamento de modelos. A topologia geral de um modelo *Stacking* é bem parecida com

a de um modelo *Bagging*. O que os diferencia é que não é feita a amostragem de Bootstrap, os modelos base são algoritmos diferentes entre si e o método de agregação é outro modelo de ML. Esta topologia está representada na Figura 3.

Figura 3 – Topologia geral de um modelo *stacking*.



Fonte: De autoria própria.

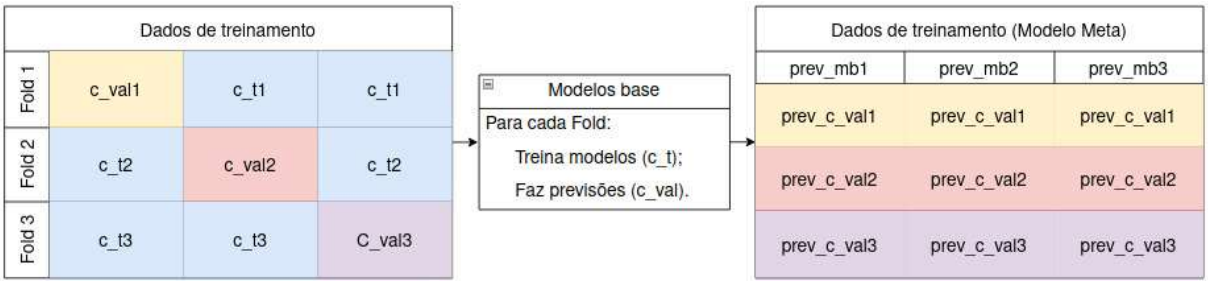
Estes modelos são treinados da seguinte forma:

- É feita a divisão entre conjunto de treinamento e teste;
- No conjunto de treinamento, é utilizada alguma forma de validação cruzada, como por exemplo o k-fold, para treinar os modelos base e fazer previsões com estes modelos para o conjunto de validação a cada rodada da validação cruzada;
- As previsões feitas pelos modelos base da etapa anterior irão compor o conjunto de treinamento do modelo meta, junto com a variável objetivo, sendo estas previsões as variáveis de entrada (*features*) do modelo meta;

- Por fim, os modelos base são treinados com todo o conjunto de dados de treinamento e o modelo meta é treinado com as previsões dos modelos base na etapa de validação cruzada.

A construção do conjunto de dados de treinamento do modelo meta está ilustrada na Figura 4, exemplificada com a técnica k-fold, com três dobras. Onde *c_t* são os dados de treinamento e *c_val* são os dados de validação e *prev_c_val* são as previsões de cada modelo base para cada dobra da validação cruzada.

Figura 4 – Construção dos dados de treinamento do modelo meta: *stacking*.



Fonte: De autoria própria.

2.2.3.2 *BLENDING*

São modelos similares aos modelos *stacking*, visto que também há o empilhamento de modelos nesta classe. Contudo, o modelo meta não recebe apenas as predições dos modelos base como entradas, mas também as entradas destes modelos. (KHYANI et al., 2021), difere os modelos *blending* dos modelos *stacking* por meio da forma de treinamento, sendo ele descrito pelo seguinte procedimento:

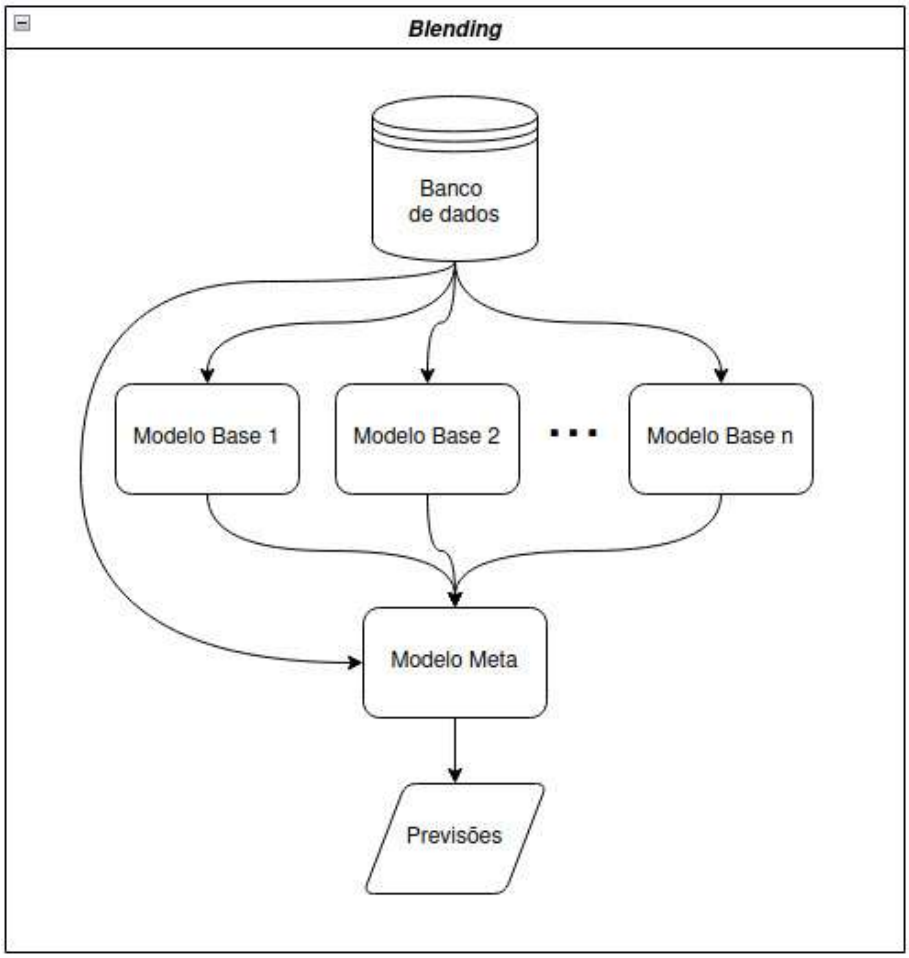
- Separa os dados nos conjuntos de treinamento do modelo base, validação ou *hold-out*, também chamado de conjunto de extensão e testes;
- Treina os modelos base utilizando o primeiro conjunto;
- Utiliza os modelos base treinados para realizar predições no conjunto de validação;
- Acrescenta as previsões feitas pelos modelos base ao conjunto de dados de validação e o utiliza para treinar o modelo meta.

Note que esse procedimento gera um problema. A quantidade de dados utilizados para treinar o modelo meta é muito pequena, visto que normalmente o conjunto de validação corresponde de 10% a 20% do total dos dados. Por este motivo, uma abordagem mais próxima da realizada nos modelos *stacking* pode ser adotada:

- É feita a divisão entre conjunto de treinamento, extensão (*hold-out*) e teste;
- No conjunto de treinamento, é utilizada alguma forma de validação cruzada, como por exemplo o k-fold, para treinar os modelos base e fazer previsões com estes modelos para o conjunto de validação a cada rodada da validação cruzada;
- Os modelos base são treinados com todo o conjunto de dados de treinamento;
- Com os modelos base, são feitas previsões no conjunto de dados de extensão;
- As previsões feitas durante a validação cruzada e no conjunto de extensão são adicionadas aos conjuntos de dados de entrada destas duas tabelas, que são unificadas para compor o conjunto de dados de treinamento do modelo meta;
- Finalmente, o modelo meta é treinado com as variáveis de entrada originais e as previsões dos modelos base.

Na Figura 5 é possível visualizar a representação gráfica da topologia geral dos modelos *Blending*.

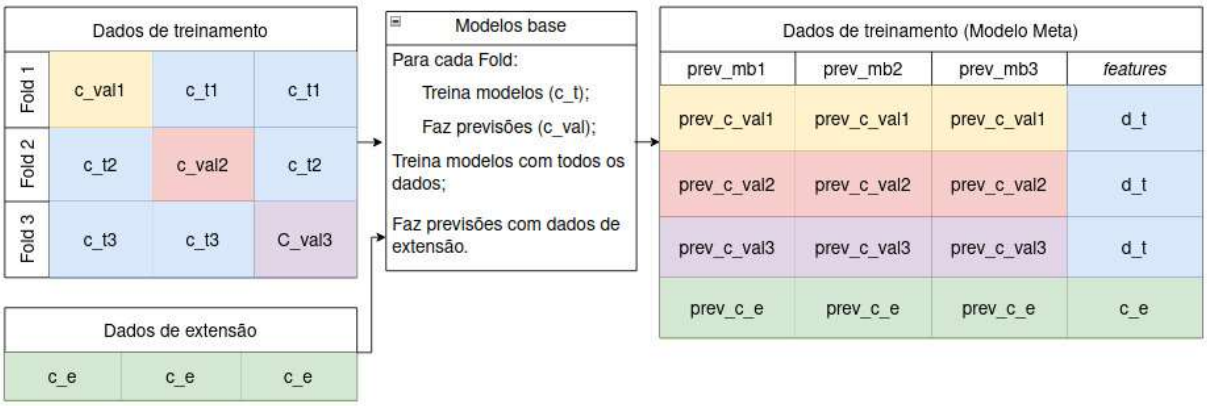
Figura 5 – Topologia geral de um modelo *blending*.



Fonte: De autoria própria.

Na Figura 6 representa-se o resumo do processo de construção do conjunto de treinamento do modelo meta. Note que as únicas diferenças entre o treinamento desta classe de modelos e os modelos *Stacking* é a presença do conjunto de extensão e a adição das variáveis de entrada (*features*) dos dados de treinamento e do conjunto de extensão no conjunto de dados de treinamento do modelo meta.

Figura 6 – Construção dos dados de treinamento do modelo meta: *blending*.



Fonte: De autoria própria.

2.3 MÉTRICAS E MÉTODOS DE AVALIAÇÃO DE MODELOS

Para escolher o modelo mais adequado para cada aplicação, é necessário ter uma forma de comparar o desempenho desses modelos entre si. Foram analisados mais de cem artigos, dos quais 11 foram selecionados para uma leitura mais aprofundada. Esses artigos comparam modelos de ML em tarefas de regressão. As informações obtidas foram organizadas na Tabela 1.

Tabela 1 – Artigos que comparam modelos de ML: Regressão.

Artigo	Melhor modelo	Métricas
Matitopanum et al. (2023)	Heterogêneo	RMSE
Ang et al. (2022)	Homogêneo	RMSE, MAE
Alam et al. (2023)	Homogêneo	MAE, RMSE, MAPE, MSLE, R^2
Lin, Lin e Lane (2021)	Homogêneo	RMSE
Radočaj, Jurišić e Tadić (2023)	Heterogêneo	R^2, RMSE
Bashir et al. (2023)	Heterogêneo	R^2, RMSE, MAE, MAPE
Fleming et al. (2021)	Heterogêneo	RPSS, R^2, RMSE
Sharghi, Nourani e Behfar (2018)	Heterogêneo	R^2, RMSE
Sarmas et al. (2024)	Heterogêneo	NMBE, MAE, RMSE
Sarreshtedar et al. (2022)	Heterogêneo	R^2, RMSE
Chao et al. (2023)	Heterogêneo	R^2, RMSE, MAPE

Fonte: De própria autoria.

Nos trabalhos apresentados na Tabela 1, em todos os casos algum modelo em conjunto obteve o melhor desempenho. Em 27% deles, o melhor desempenho foi de um modelo homogêneo, enquanto em 73% os modelos heterogêneos alcançaram melhores métricas.

2.3.1 MÉTRICAS

Ao analisar a Tabela 1, observa-se que as métricas mais populares entre os trabalhos avaliados são RMSE, MAE e R^2.

De acordo com Terven et al. (2025), o erro percentual absoluto médio (*Mean Absolute Percentage Error* - MAPE) e o coeficiente de determinação (R^2) condensam as informações do sinal medido e do sinal de referência, sendo ambas métricas adimensionais. O MAPE (Eq. 2.1) é uma métrica que mede a precisão de modelos de previsão, expressando o erro médio em termos percentuais da diferença entre valores reais e previstos. Entretanto, essa métrica é sensível a valores reais próximos de zero, o que pode fazê-la tender ao infinito quando o banco de dados contém valores próximos a esse limiar.

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{|y_i - \hat{y}_i|}{y_i} \times 100\right) \tag{2.1}$$

Por esse motivo, o MAPE foi substituído neste estudo pelo erro médio absoluto normalizado (*Normalized Mean Absolute Error* - NMAE) (Eq. 2.2). Essa métrica possui interpretação semelhante ao MAPE, mas é menos sensível a valores reais próximos de zero, exceto quando todo o conjunto de dados se encontra próximo desse valor.

$$\text{NMAE} = \frac{\text{MAE}}{\overline{y}} = \frac{\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|}{\frac{1}{n}\sum_{i=1}^{n} y_i} \tag{2.2}$$

Em todas as fórmulas de erro aqui descritas, n representa o número de amostras, y_i o valor real de uma amostra, $\hat{y}_i$ o valor predito e $\bar{y}$ a média dos valores reais do conjunto de amostras.

O R^2 (Eq. 2.3), por sua vez, é um índice que quantifica a proporção da variabilidade total da variável resposta explicada pelo modelo. Em geral, valores mais elevados de R^2 indicam maior capacidade explicativa do modelo. Entretanto, a interpretação do R^2 deve considerar o contexto específico do problema. Por exemplo, um valor reduzido de R^2 pode ser aceitável quando os dados apresentam elevada variabilidade intrínseca e apenas previsibilidade parcial.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.3}$$

No estudo de Terven et al. (2025), é apresentada a métrica R^2 ajustado, que consiste em uma variação do R^2. Segundo os autores, essa métrica é mais adequada para a análise de desempenho de modelos com múltiplos preditores, como ocorre nos conjuntos de modelos avaliados neste trabalho. O objetivo do ajuste é penalizar modelos com muitos preditores, favorecendo modelos mais simples. Essa penalização busca maximizar a relação esforço-retorno entre o treinamento dos modelos e sua exatidão. Entretanto, considerando que o objetivo deste estudo é identificar os melhores modelos em termos de exatidão, penalizando apenas aqueles cujo desempenho seja significativamente inferior à média em relação ao tempo de treinamento, optou-se por manter a métrica R^2, em vez de utilizar sua versão ajustada.

Ainda segundo Terven et al. (2025), o erro médio absoluto (*Mean Absolute Error* - MAE) e a raiz do erro quadrático médio (*Root Mean Squared Error* - RMSE) são duas métricas bastante populares. Ambas são sensíveis à escala da variável avaliada. O MAE (Eq. 2.4) representa a média das diferenças absolutas entre os valores previstos e os valores reais. Trata-se de uma medida não negativa que penaliza linearmente erros maiores, sendo menos sensível a *outliers* do que o erro quadrático médio (*Mean Squared Error* - MSE), por exemplo.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.4}$$

O RMSE (Eq. 2.5) representa a média da distância euclidiana entre as previsões e os valores reais, atribuindo maior peso a resíduos maiores. Isso significa que essa métrica é mais sensível a *outliers*.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2.5}$$

Na Tabela 2, estão resumidas as principais características das métricas escolhidas para este estudo.

Tabela 2 – Resumo das métricas de avaliação de modelos.

Métrica	**Nome**	**Características**	**Aplicação**
MAE	Erro Absoluto Médio	Mede a média das diferenças absolutas. Penaliza linearmente erros maiores.	Indicado quando se deseja interpretar o erro médio de forma direta, em problemas de regressão com poucos *outliers*.
RMSE	Raiz do Erro Quadrático Médio	Mantém a unidade original dos dados. Mais sensível a *outliers*.	Comum em regressão e séries temporais, quando se deseja interpretar o erro na mesma escala da variável prevista.
NMAE	Erro Absoluto Médio Normalizado	Similar ao MAPE, mas menos sensível a valores próximos de zero.	Adequado para comparar modelos em diferentes escalas ou bases de dados heterogêneas, evitando distorções quando há valores próximos de zero.
R^2	Coeficiente de Determinação	Indica a proporção da variabilidade da variável resposta explicada pelo modelo. Valores mais altos sugerem maior capacidade explicativa.	Usado para avaliar o ajuste global de modelos de regressão. Útil em contextos onde se deseja medir o quanto o modelo explica a variabilidade dos dados.

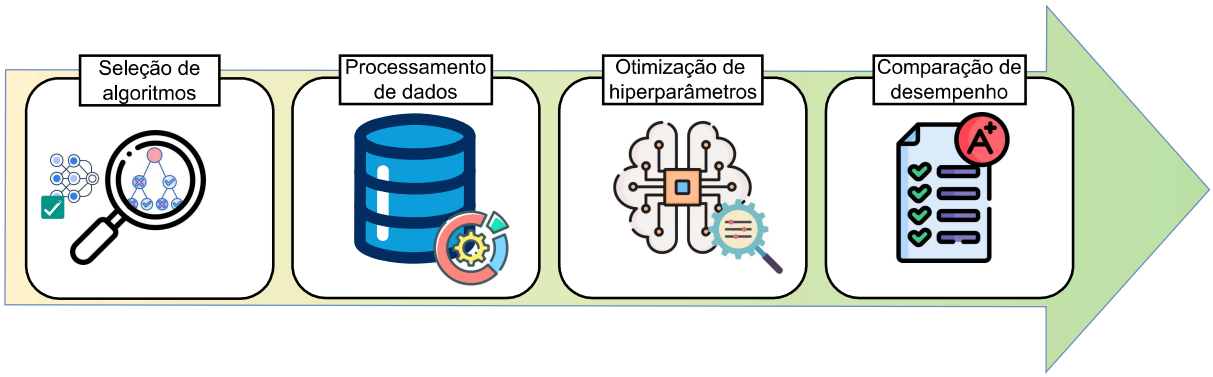
2.3.2 VALIDAÇÃO CRUZADA

Analisando os trabalhos apresentados na Tabela 1, nota-se que a validação cruzada é um ponto em comum a todos eles. Essa técnica é utilizada para garantir que a comparação entre os modelos seja justa, tanto no momento da escolha de seus hiperparâmetros quanto na comparação de diferentes técnicas de ML para uma mesma tarefa. A validação cruzada reduz a dependência do desempenho dos modelos em relação aos conjuntos de dados de validação e treinamento, utilizando todo o conjunto de treinamento para treinar e avaliar os modelos (IYENGAR; LAM; WANG, 2024).

3 METODOLOGIA

A sequência da metodologia proposta para o desenvolvimento e a seleção dos modelos virtuais está representada de forma resumida na Figura 7:

Figura 7 – Sequência da Metodologia.



Fonte: Elaboração própria.

Cada uma dessas etapas foi realizada de forma sequencial para cada sistema modelado, exceto pela escolha das técnicas de aprendizado de máquina (ML), que ocorreu previamente. Assim, para cada técnica selecionada, foram desenvolvidos três modelos, representando cada uma das variáveis simuladas.

A implementação da metodologia descrita foi realizada utilizando a linguagem Python e seus recursos (VIRTANEN et al., 2020; NOGUEIRA, 2014–; PEDREGOSA et al., 2011; ABADI et al., 2015; DOROGUSH; ERSHOV; GULIN, 2018).

Porém, antes de explorar os métodos utilizados para o desenvolvimento do presente trabalho, é importante compreender o estudo de caso ao qual ele está vinculado.

3.1 ESTUDO DE CASO

Os dados utilizados para a realização do presente estudo foram cedidos por uma empresa responsável pelo gerenciamento de uma usina termelétrica de ciclo combinado de capacidade instalada de 40 MW, localizada no Brasil, operando de forma isolada do Sistema Interligado Nacional. Por suprir sozinha a carga da cidade onde está instalada, sua operação torna-se bastante desafiadora. Isso ocorre porque a demanda da cidade varia constantemente, forçando a planta a operar em condições de menor eficiência em troca da manutenção do funcionamento do sistema em seus parâmetros normais. Neste caso, a ferramenta escolhida para otimizar a operação desta UTE foi um SGD. Os modelos

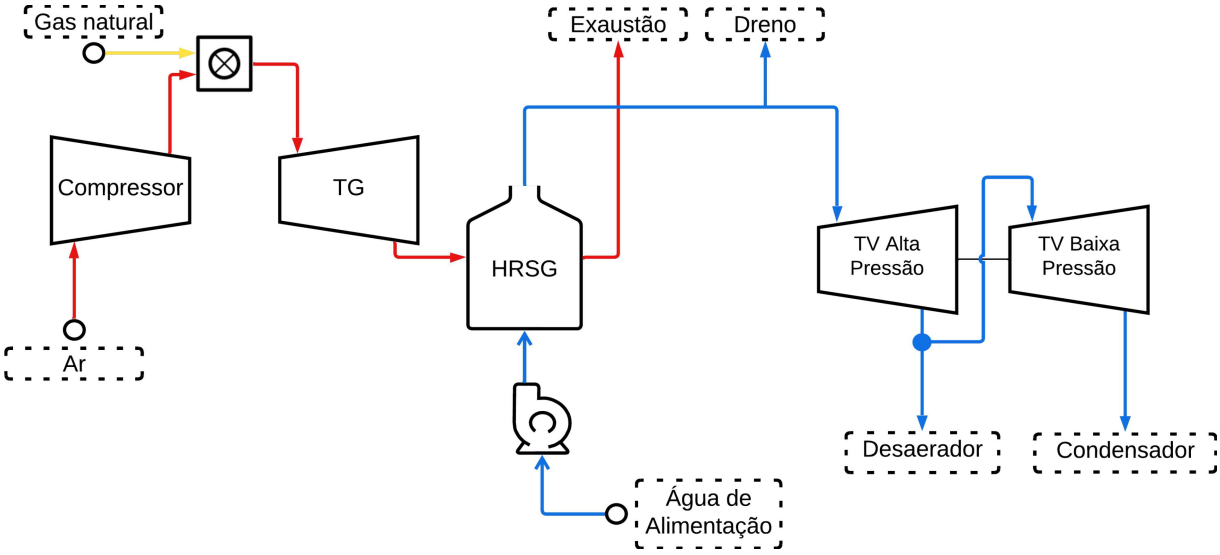
utilizados para replicar o comportamento de algumas das variáveis mais importantes desta operação são o objeto de estudo do presente trabalho.

3.1.1 O CICLO COMBINADO

A usina é composta por dois conjuntos compostos de gerador e turbina a gás (TG) - Ciclo Brayton -, com potência nominal de 14 MW cada, e um conjunto composto por gerador e turbina a vapor (TV) - Ciclo Rankine, com potência nominal de 12 MW. Além disso, a planta também conta com um sistema de condicionamento de ar, acoplado às entradas de ar das turbinas a gás, uma caldeira de recuperação de calor (*Heat recovery steam generator* - HRSG) e um queimador auxiliar (QA), que injeta calor diretamente no sistema da TV, para aumentar a potência gerada pelo seu conjunto.

Note que, embora a planta contenha dois conjuntos de TGs, apenas um deles fica ativo em operação normal, enquanto isso, o outro passa por manutenção ou atua como *backup* em caso de falhas. Na Figura 8, é possível observar uma representação gráfica do esquema geral da usina.

Figura 8 – Representação gráfica de uma usina de ciclo combinado.



Fonte: De autoria própria.

Dessa forma, a operação da planta consiste em ajustar os parâmetros de entrada destes equipamentos, principalmente as vazões de gás que alimentam a TG e o QA e a temperatura do ar que vai para a TG ativa, por meio do sistema de condicionamento de ar.

3.1.2 PREMISSAS

As seguintes premissas foram consideradas para o desenvolvimento dos modelos para o presente estudo:

1. Apenas uma TG estará ativa por vez;
2. O sistema desenvolvido não será integrado ao sistema supervisório da planta;
3. As entradas dos modelos devem ser controláveis ou representar condições ambientais;
4. Os transientes não serão levados em consideração, apenas o comportamento em regime permanente será emulado.

Sendo assim, apenas o comportamento normal, excetuando-se situações de falha e períodos de manutenção, nos quais a operação foge à normalidade, será replicado pelos modelos. Note que os modelos ainda replicarão o comportamento dos sistemas simulados em função do seu desgaste e das condições ambientais. Assim, será possível realizar experimentos na planta, no ambiente virtual, obtendo informações fidedignas a respeito da sua operação e eficiência.

Quanto ao sistema não considerar transientes, esta decisão foi tomada para reduzir a complexidade dos modelos e da solução final. Uma vez que a finalidade do SGD é treinar novos operadores e auxiliar na tomada de decisão, analisando cenários históricos de operação, não há a necessidade de integrar a solução final ao supervisório da usina. Além disso, esta característica confere uma maior segurança à solução final, reforçando medidas de cybersegurança.

Ademais, evitar transientes reforça a coerência dos dados. Isso porque momentos de transiente podem gerar dados instantâneos incoerentes com o sistema físico teórico, como por exemplo uma eficiência de mais de 50%, dado que a eficiência máxima do equipamento emulado é de aproximadamente 45%. Isso pode ocorrer devido à inércia térmica, que mantém uma energia residual no sistema e que em momentos de redução de carga, pode acarretar em eficiências instantâneas acima do normal. Portanto, para contemplar estas características, seria necessário realizar uma análise em janela temporal. Este comportamento é indesejado no SGD em produção e, por isso, não será levado em consideração. Pelo exposto, o sistema desenvolvido será um SGD do tipo modelo digital, não havendo interesse no comportamento transiente, apenas no comportamento em regime permanente.

3.1.2.1 CARACTERIZAÇÃO DO SISTEMA

A premissa da modelagem baseada em dados é que o modelo gere dados de saída o mais próximo possível dos valores que seriam obtidos ao alimentar a planta original com

os mesmos valores de entrada. Dessa forma, não seria mais necessário fazer experimentos reais com a planta, bastando apenas alimentar o modelo com dados de controle para obter as respostas desejadas. Portanto, tais modelos geram sinais de saída a partir de sinais de entrada, sendo esta a definição de um sistema, de acordo com Oppenheim. e Willsky (2010) e Lathi (2007).

Sendo assim, é possível estipular suas características de acordo com o comportamento dos dados que os definem e o modelo real o qual ele representa, uma vez que, depois de treinado, o modelo irá reproduzir esse comportamento.

Em Lyantsev, Kazantsev e Abdulnagimov (2017), o autor caracteriza uma turbina a gás como um sistema não linear. Quanto à invariância no tempo, é possível considerar que uma turbina a gás representa um sistema invariante no tempo para um curto período, visto que, no longo prazo, tais equipamentos perdem eficiência e passam por manutenções, modificando suas características operacionais. Sendo assim, será necessário retreinar o modelo digital sempre que seu comportamento divergir do comportamento dos dados da planta real.

Quanto à estabilidade, tais sistemas são estáveis, dentro dos limites operacionais. Uma vez que o modelo digital seguirá o comportamento da planta a partir dos dados de operação, é possível garantir que o modelo digital também será estável dentro da mesma faixa de operação. Assim, operações fora desta faixa podem gerar resultados inesperados, por exemplo, uma potência de saída maior que a potência máxima do equipamento real.

Com relação à memória, uma vez que o projeto não contempla análise de transientes, é possível considerar que o sistema não possui memória. Desta forma, é possível representar o modelo digital como um sistema em malha aberta. Isso irá evitar que ocorra o acúmulo do erro na resposta do modelo, por conta da realimentação.

A mesma análise se aplica à turbina a vapor, com a diferença de que a sua inércia térmica é muito maior. Dessa forma, a taxa de variação da sua curva de geração de energia é muito mais lenta em comparação com a TG, sendo necessário muito mais tempo entre o estímulo - mudança no ponto de operação - e a estabilização do sistema.

3.2 SELEÇÃO DE ALGORITMOS

A seleção dos algoritmos teve como objetivo definir as técnicas de aprendizado de máquina (ML) utilizadas no estudo comparativo. Essa escolha foi realizada com base nos seguintes critérios:

- Técnicas presentes nos trabalhos analisados;
- Técnicas adequadas à solução do problema;

- Técnicas pouco exploradas, mas com desempenho promissor.

Note que não foram selecionadas técnicas como *Long Short-Term Memory* (LSTM), por exemplo, pois, apesar de seu reconhecido desempenho, esse tipo de abordagem não era adequado ao tipo de simulação realizada pelos modelos desenvolvidos neste trabalho. Técnicas dessa natureza utilizam informações residuais de previsões ou variáveis de entrada anteriores para estimar a saída atual. Isso significa que esses modelos possuem memória e dependem de informações sobre um estado anterior para estimar o estado atual.

No entanto, a simulação realizada pelos modelos desenvolvidos considerou apenas as variáveis de estado atuais para estimar a variável de saída. As simulações foram individuais e independentes entre si. Dessa forma, modelos com memória não se mostraram apropriados ao problema.

3.3 PROCESSAMENTO DE DADOS

O processamento de dados teve como objetivo analisar e preparar os dados para a utilização no treinamento dos modelos. Esse processo foi dividido nas etapas ilustradas na Figura 9.

Na primeira etapa, foi realizada a análise dos dados com o objetivo de identificar as variáveis, verificar as taxas de amostragem, detectar valores faltantes, inconsistentes ou inválidos, selecionar as variáveis de interesse e avaliar as correlações entre as variáveis de entrada e as variáveis de saída.

Após a análise do banco de dados, realizou-se a consolidação dos dados, compilando-os em uma única tabela e corrigindo possíveis assincronias.

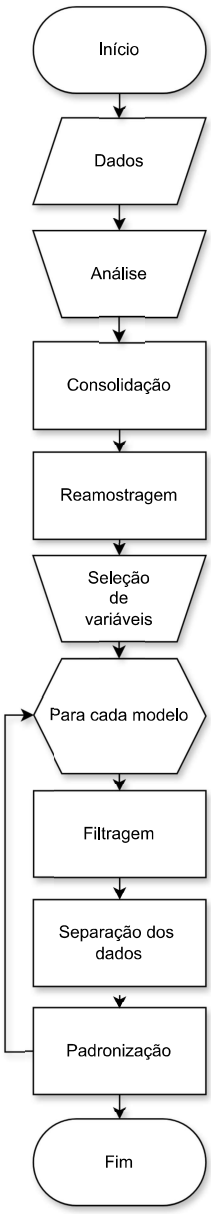
Na etapa seguinte, os dados foram reamostrados utilizando a média móvel para períodos de dez minutos, com os objetivos de reduzir a influência de transientes e inércia, filtrar ruídos de alta frequência e homogeneizar os dados, uma vez que foram disponibilizados em arquivos com taxas de amostragem distintas.

Em seguida, foi realizada a seleção das variáveis de entrada dos modelos. Para que uma variável fosse selecionada, ela deveria atender a um ou mais dos seguintes critérios:

- A variável deveria ser controlável; Ex.: Quantidade de gás injetado na turbina a gás.
- A variável deveria ser um parâmetro ambiental; Ex.: Temperatura ambiente.
- A variável deveria apresentar boa correlação com a variável objetivo.

A correlação utilizada foi a correlação de Spearman, que indica relações monotônicas independentemente da linearidade (KING; ECKERSLEY, 2019).

Figura 9 – Processamento de dados.



Fonte: Elaboração própria.

A separação do banco de dados para as etapas do trabalho foi realizada considerando a quantidade de dados disponível e a quantidade necessária para cada etapa.

Note que nem sempre vale a máxima: quanto mais dados, melhor! Por isso, para o treinamento dos modelos, foi utilizado um período de dados considerado estável. O comportamento das entidades físicas, que os modelos deveriam replicar, variava ao longo do tempo, mas permanecia estável no curto prazo. Dessa forma, o tamanho da janela de dados para treinamento e testes precisou ser ideal, a fim de mitigar esse efeito e, ainda assim, garantir dados suficientes para um bom desempenho dos modelos.

Em seguida, foi realizada a filtragem dos dados. Essa etapa ocorreu individualmente antes do treinamento de cada modelo, evitando a perda de informações ao eliminar registros apenas do subconjunto destinado ao treinamento e aos testes. Foram removidos do banco de dados os registros contendo:

- **Dados faltantes:** Registros que apresentavam uma ou mais variáveis sem valores.
- **Dados inválidos:** Registros fora das premissas e regras de operação padrão da usina, como períodos em que uma turbina estava em manutenção.
- **Dados espúrios:** Registros com relações entre variáveis fora do padrão em relação ao restante do conjunto de dados (*outliers*).

A filtragem de *outliers* foi realizada por meio da técnica *Local Outlier Factor* (LOF) (BREUNIG et al., 2000), uma abordagem não supervisionada. Essa técnica mede o quanto uma amostra se distancia de suas vizinhas em termos de densidade. A vizinhança é definida pelos k pontos mais próximos (*k-nearest neighbors*), e a distância entre esses pontos é utilizada para calcular a densidade local. Comparando a densidade local de uma amostra com a densidade de seus vizinhos, é possível identificar amostras significativamente menos densas, que são consideradas *outliers* e removidas do banco de dados.

Esta técnica foi escolhida em detrimento de outras como DBSCAN, por exemplo. Isso porque, apesar de ser mais sensível ao ruído, é mais indicada para lidar com regiões esparsas de dados. Além disso, é ideal para conjuntos de dados com poucas variáveis (*features*) (BREUNIG et al., 2000; GOLDSTEIN; UCHIDA, 2016).

A análise foi realizada registro a registro, considerando todas as variáveis de entrada e a variável objetivo. Além disso, a implementação do algoritmo permitiu especificar ou não o percentual de contaminação dos dados. Para este trabalho, esse valor não foi definido, sendo otimizado pelo próprio algoritmo.

Após a filtragem, os dados foram divididos em conjuntos de treinamento e teste, respeitando a proporção de 80-20 para cada conjunto, respectivamente. O conjunto de treinamento pôde ser subdividido conforme a técnica de aprendizado de máquina utilizada; o detalhamento dessa subdivisão é apresentado no Capítulo 4.

Finalmente, os dados foram padronizados, normalizados para o intervalo $[0, 1]$, a fim de evitar prejuízos em técnicas sensíveis a dados não padronizados e permitir comparações diretas com técnicas que não apresentam essa limitação. Para isso, foi utilizado o estimador *MinMaxScaler*, que preserva a distribuição original dos dados, apenas transformando-os para o novo domínio (PEDREGOSA et al., 2011). Observe que outras técnicas de padronização foram testadas, mas esta acabou sendo a mais eficaz em testes preliminares; por isso, foi a adotada neste trabalho.

3.4 OTIMIZAÇÃO DE HIPERPARÂMETROS

Para a otimização de hiperparâmetros, em cada um dos modelos virtuais testados, foi realizada uma busca utilizando a técnica de otimização bayesiana (NOGUEIRA, 2014–; DELOUILLE et al., 2018). Essa abordagem é eficaz para funções ruidosas, de natureza caixa-preta ou com alto custo computacional, sendo bastante adequada para este contexto. Sua principal vantagem é realizar uma busca guiada por pontuação, semelhante a técnicas populacionais como *Particle Swarm Optimization* (PSO) e Algoritmos Genéticos (AG), porém testando apenas uma configuração por iteração. Isso faz com que sejam alcançados resultados equivalentes aos obtidos pelas técnicas populacionais, porém utilizando bem menos simulações por iteração, mesmo que isso aumente um pouco o número de iterações. Dessa forma, é possível alcançar resultados satisfatórios com menor esforço computacional, reduzindo significativamente o tempo da etapa de busca de hiperparâmetros.

O algoritmo utilizou as informações das configurações previamente avaliadas para estimar, de forma iterativa, a configuração que se aproximava do mínimo global. Os limites dos hiperparâmetros, bem como os valores utilizados para cada modelo, foram definidos conforme as recomendações da documentação de cada arquitetura testada.

A avaliação do desempenho das configurações dos modelos em cada etapa da otimização foi realizada por meio da média do RMSE obtida na validação cruzada *k-fold*. Inicialmente, foram testadas 150 configurações de hiperparâmetros para cada modelo. Contudo, esse número pôde ser ajustado em casos de treinamento mais lento ou quando a melhor configuração foi encontrada nas etapas iniciais, maximizando o resultado final e minimizando o tempo e o esforço computacional para concluir essa fase.

Quanto aos hiperparâmetros, foram escolhidos os recomendados pela documentação das bibliotecas utilizadas para a implementação dos algoritmos ou os utilizados nos estudos avaliados no Capítulo 2. Dessa forma, os algoritmos escolhidos para este trabalho e seus respectivos espaços de busca de hiperparâmetros estão dispostos na Tabela 3. É importante ressaltar que, para os modelos heterogêneos, o objetivo da busca de hiperparâmetros foi otimizar apenas o modelo meta, visto que os hiperparâmetros dos modelos base foram os mesmos encontrados nas buscas de hiperparâmetros dos modelos simples e homogêneos utilizados.

Tabela 3 – Técnicas de ML testadas - Potência da TG

MODELO	TIPO	HIPERPARÂMETROS
ADABOOST	HOMOGÊNEO: BOOSTING	learning_rate: [0,1, 5] n_estimators: [20, 200]
BLENDING FFNN - MÉDIA DINÂMICA	HETEROGÊNEO	n_hidden_layers: [2, 5] neurons_in_layer_0: [3,100] neurons_in_layer_1: [3,100] neurons_in_layer_2: [3,100] neurons_in_layer_3: [3,100] neurons_in_layer_4: [3,100]
BLENDING FFNN - PADRÃO	HETEROGÊNEO	n_hidden_layers: [2, 5] neurons_in_layer_0: [3,100] neurons_in_layer_1: [3,100] neurons_in_layer_2: [3,100] neurons_in_layer_3: [3,100] neurons_in_layer_4: [3,100]
BLENDING GPR	HETEROGÊNEO	alpha : [0,001, 1] kernel: rbf' RBF: length_scale [1, 20]
BLENDING RIDGE_CV	HETEROGÊNEO	cv: [1, 10]
BLENDING SVR	HETEROGÊNEO	kernel: 'rbf' C: [0,1, 2] epsilon: [0,01, 5] gamma: [0,1, 2]
CATBOOST	HOMOGÊNEO: BOOSTING	bagging_temperature: [0, 100] depth: [2, 10] iterations: [500,5000] l2: (0, 5] random_strength: [0, 10]
FFNN	SIMPLES	n_hidden_layers: [2, 5] neurons_in_layer_0: [3,100] neurons_in_layer_1: [3,100] neurons_in_layer_2: [3,100] neurons_in_layer_3: [3,100] neurons_in_layer_4: [3,100]
GPR	SIMPLES	alpha : [0,001, 1] kernel: rbf' RBF: length_scale = [1, 20]
RANDOM FOREST	HOMOGÊNEO: BAGGING	max_depth: [2, 5] min_samples_split: [2, 10] n_estimators: [20, 200]
STACKING GPR	HETEROGÊNEO	alpha: [0,001, 1] kernel: rbf' RBF: length_scale = [1, 20]
STACKING RIDGE_CV	HETEROGÊNEO	cv: [1, 10]
SVR	SIMPLES	kernel: 'rbf' C: [0,1, 2] epsilon: [0,01, 5] gamma: [0,1, 2]

Fonte: De autoria própria.

3.5 COMPARAÇÃO DO DESEMPENHO

Após a otimização dos hiperparâmetros, os modelos foram treinados com todo o conjunto de dados de treinamento, respeitando as especificidades de cada um e utilizando os hiperparâmetros definidos na etapa anterior. Uma vez treinados, os modelos foram avaliados com o conjunto de teste, por meio das métricas selecionadas: RMSE, MAE, R^2 e NMAE, compiladas em uma tabela.

Além disso, foi realizada a análise gráfica do erro dos modelos para cada variável objetivo. Com base nessas informações, foi escolhido um modelo virtual para cada uma delas.

4 RESULTADOS

Este capítulo descreve os resultados obtidos a partir da aplicação da metodologia proposta no Capítulo 3.

4.0.1 ANÁLISE DO BANCO DE DADOS

Os sistemas a serem virtualizados pelos modelos são os conjuntos de turbinas a gás e a vapor, além do funcionamento geral da usina. Dessa forma, como deseja-se obter a informação a respeito da eficiência da operação da usina, é necessário modelar as potências geradas por cada uma das turbinas e a potência líquida, entregue à rede, que é a soma de toda a potência gerada pela usina, descontada do consumo interno.

De forma semelhante, as temperaturas internas das turbinas constituem informações relevantes e que deverão ser incorporadas na versão final do SGD. Entretanto, considerando que, para cada variável modelada, foram avaliados 13 modelos de aprendizado de máquina, optou-se, neste estudo, por não incluir as temperaturas internas, a fim de reduzir a complexidade computacional da análise. Além disso, como as turbinas a gás (TG) são do mesmo modelo, os testes foram realizados utilizando-se apenas uma delas. Dessa forma, a mesma topologia de modelagem foi empregada para ambas, diferenciando-se apenas os conjuntos de dados utilizados no treinamento, de modo que as informações associadas ao desgaste permaneçam específicas de cada TG. Assim, as variáveis efetivamente modeladas neste trabalho são as potências da TG (uma unidade), da turbina a vapor (TV) e a potência líquida do sistema.

Um dos desafios foi estimar a potência líquida sem explicitar o consumo interno como variável de entrada, visto que essa informação, embora esteja presente no conjunto de dados disponibilizado, mostrou-se inconsistente após a análise dos dados, quando comparada com as demais informações fornecidas.

Os dados foram disponibilizados em pacotes mensais ou quinzenais. Até a data de elaboração deste documento, o banco de dados conta com dados da operação da usina, datados de 14 de abril de 2023 à 28 de setembro de 2024.

Os dados disponibilizados foram coletados de quatro interfaces homem-máquina, cada uma pertencente a um sistema isolado, sem sincronia ou comunicação de dados entre si. Além disso, os dados foram armazenados com taxas de amostragem diferentes entre si, sendo eles registrados a cada minuto ou a cada segundo. Estes registros foram sincronizados e reamostrados para o período de dez em dez minutos, utilizando a média móvel para filtrar as amostras dentro do período. Dessa forma, são eliminados ruídos de

alta frequência e transientes. Este período foi escolhido levando em consideração o tempo de resposta da TG e da TV, bem como a variação da demanda da usina, que fica bem próximo deste intervalo.

Foram 95 variáveis coletadas, contudo apenas algumas delas serão utilizadas para a elaboração deste trabalho. Na Tabela 4, estão descritas as variáveis utilizadas no desenvolvimento dos modelos.

Ao todo, o banco de dados contendo apenas as variáveis utilizadas neste estudo apresentava 61.057 registros antes do tratamento. Após uma primeira remoção de registros com valores inválidos e *outliers*, a quantidade final foi reduzida para 8.403 registros. Essa redução significativa ocorreu devido à grande quantidade de registros parcialmente vazios, que, conforme a política de descarte adotada, foram eliminados por não atenderem aos critérios de validade.

Para mitigar esse efeito, o banco de dados original foi subdividido em três conjuntos, cada um correspondente a uma variável objetivo, acompanhado de suas respectivas variáveis de entrada. Em seguida, o tratamento de dados foi feito para cada um dos subconjuntos de dados. Assim, para a potência líquida, o banco final contém 35.766 registros. Para TG e TV, os bancos de dados possuem, respectivamente, 31.193 e 13.512 registros. No entanto, considerando o objetivo de utilizar os dados disponíveis para o desenvolvimento de um Sistema de Gestão de Dados (SGD), é necessário reservar uma parte dos registros para testes e retreinamento dos modelos.

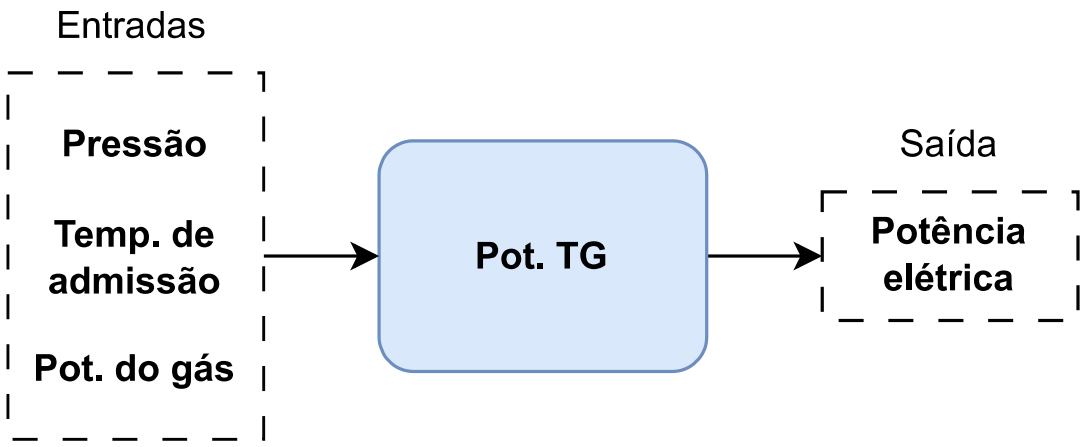
Dessa forma, apenas dados de operação do ano de 2023 foram utilizados. Sendo assim, os conjuntos de dados utilizados para treinamento e testes dos modelos correspondem a 16.330, 8.577 e 1.911 registros, respectivamente, para a potência líquida e potências da TG e TV.

Tabela 4 – Descrição das variáveis selecionadas.

VARIÁVEL (UND)	DESCRIÇÃO
POTÊNCIA DA TG (MW)	Potência ativa gerada pela TG.
POTÊNCIA LÍQUIDA (MW)	Potência ativa gerada pela usina, descontando o consumo interno.
POTÊNCIA BRUTA (MW)	Potência ativa bruta gerada pela usina, sem descontar o consumo interno.
POTÊNCIA DA TV (MW)	Potência ativa gerada pela TV.
POTENCIA DO GÁS (MW)	Potencial químico do gás fornecido à TG, calculado por meio da vazão do gás.
VAZÃO DO GÁS QA (m^3/s)	Vazão do gás que alimenta o queimador auxiliar.
TEMPERATURADO AR (°C)	Temperatura do ar de admissão da TG, pós condicionamento de ar.
TEMPERATURA AMBIENTE (°C)	Temperatura do ar antes do condicionamento de ar.
PRESSÃO (Bar)	Pressão do ar de admissão da TG.

Assim, as potências ativas líquida, da TG e da TV foram as variáveis objetivo dos três modelos desenvolvidos, respectivamente. Para o modelo da TG, foram utilizadas a temperatura do ar de admissão, a pressão e a potência do gás. A representação do modelo pode ser visualizada na Figura 10. Note que a potência do gás é fornecida no banco de dados em MW, sendo calculada pelo sistema da usina em função da vazão do gás.

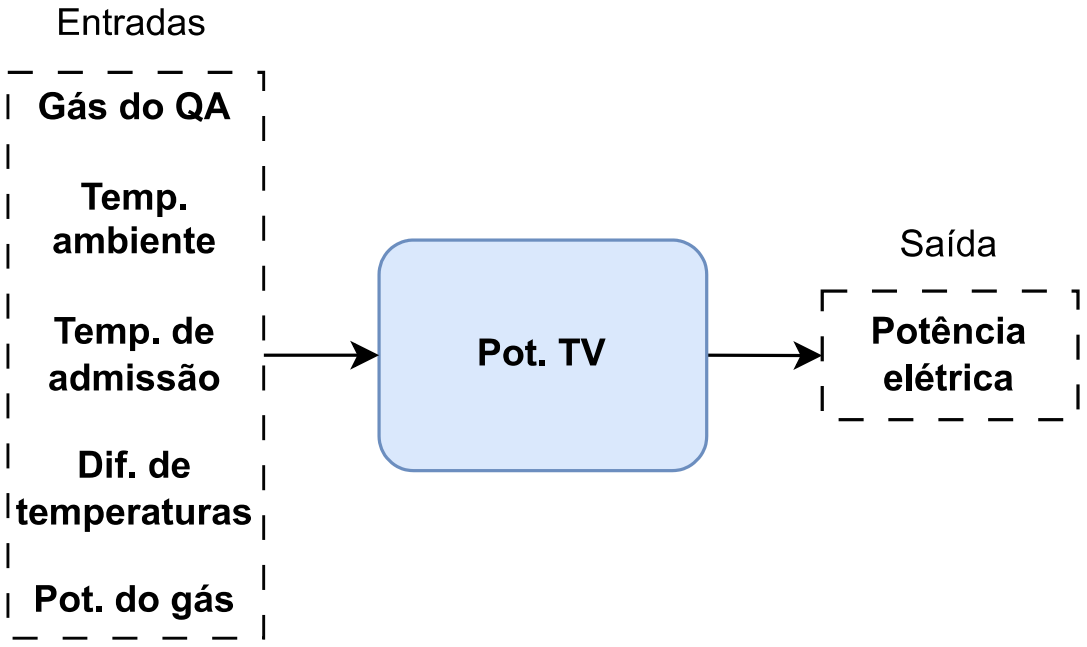
Figura 10 – Esquema geral do modelo da potência da turbina a gás.



Fonte: De autoria própria.

Para o modelo da TV, foram utilizadas as variáveis de entrada das TGs, exceto a pressão, unificadas em uma única variável. Dessa forma, o modelo da TV, ilustrado na Figura 11, utiliza as informações de potência do gás e temperatura de admissão das duas turbinas. Isso foi feito porque a TV opera com as duas TGs em regime complementar, não operando simultaneamente. Poderia ter sido utilizada também a temperatura interna das TGs como variável de entrada para este modelo. Contudo, isso implicaria em um modelo dependente da saída de outro, o que impediria a realização de testes com a TV de forma isolada e poderia causar a propagação de erros de um modelo para o outro, possivelmente aumentando o erro de estimativa da predição deste modelo ao utilizá-lo em conjunto com os demais. Ademais, a informação a respeito da temperatura estava sendo propagada por meio da potência do gás e da temperatura de admissão, visto que estas se correlacionam com a potência e com a temperatura geradas pelas TGs. Além disso, também foram utilizadas a temperatura ambiente e a diferença entre as temperaturas ambiente e a temperatura de admissão das TGs como entradas deste modelo.

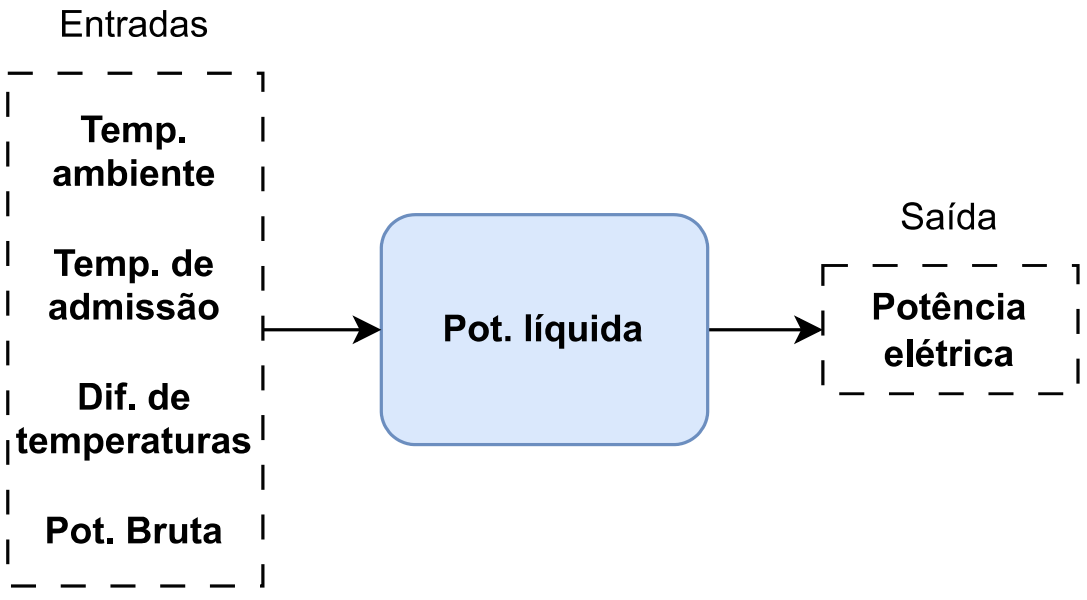
Figura 11 – Esquema geral do modelo da potência da turbina a vapor.



Fonte: De autoria própria.

Finalmente, para o modelo da potência líquida, as variáveis de entrada utilizadas foram a temperatura ambiente, a temperatura do ar de admissão, a potência bruta gerada pela usina, sendo ela a soma das potências geradas individualmente pelas TG e TV e a diferença entre a temperatura ambiente e a temperatura de admissão da TG ativa, como ilustrado na Figura 12.

Figura 12 – Esquema geral do modelo da potência líquida.

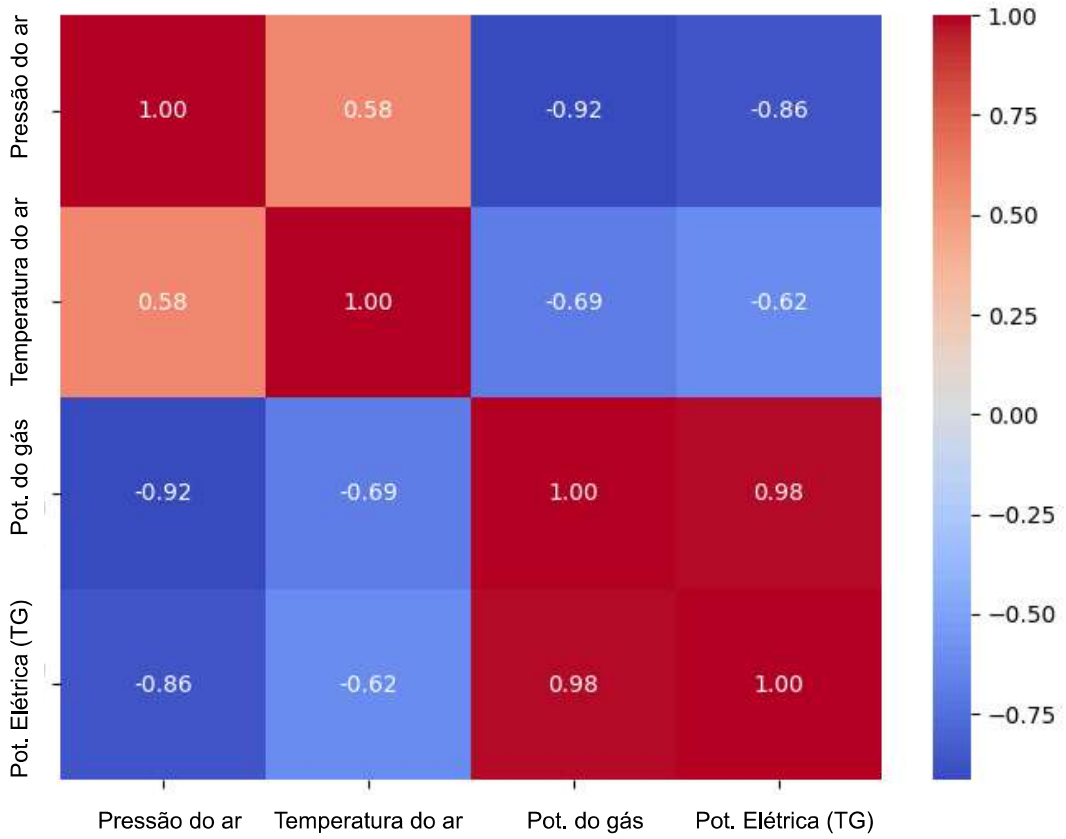


Fonte: De autoria própria.

4.0.2 CORRELAÇÃO ENTRE AS VARIÁVEIS

A análise da matriz de correlação é essencial para compreender as interdependências entre as variáveis e orientar a seleção dos atributos mais relevantes na construção dos modelos de regressão. Nesse contexto, o cálculo da correlação de Spearman entre todas as variáveis dos conjuntos de dados escolhidos permite identificar não apenas a relação direta entre variáveis de entrada e a variável objetivo, mas também as interações existentes entre os próprios atributos, fornecendo uma visão mais abrangente das dependências que influenciam o desempenho da turbina. Os resultados estão expostos nas Figuras 13, 14 e 15.

Figura 13 – Matriz de correlação: Potência da turbina a gás.



Fonte: De autoria própria.

A Figura 13 apresenta a matriz de correlação entre as variáveis envolvidas na geração de potência da turbina a gás. Observa-se que a potência do gás apresenta uma correlação fortemente positiva com a potência elétrica da TG ($\rho = 0{,}98$), indicando que ambas as variáveis estão altamente associadas e tendem a variar de forma conjunta. Essa relação é esperada, dado que a potência elétrica gerada depende diretamente da energia liberada na combustão do gás.

Por outro lado, tanto a pressão do ar quanto a temperatura do ar apresentam correlações negativas com a variável objetivo. Apesar de serem negativas, seus altos valores indicam forte associação entre as variáveis, sugerindo que aumentos na pressão ou temperatura de entrada estão associados à redução da potência gerada.

Com relação à redução da potência com o aumento da temperatura, existem alguns pontos que devem ser levados em consideração para a interpretação desta correlação.

- **Densidade do ar:** De acordo com a Lei dos Gases Ideais, a densidade do ar é inversamente proporcional à sua temperatura. Assim, quanto mais denso o ar, maior será a quantidade de oxigênio disponível por unidade de volume, favorecendo a combustão do combustível na câmara de combustão — desde que respeitadas as proporções estequiométricas (ÜNVER et al., 2018). Portanto, temperaturas mais elevadas reduzem a densidade do ar, o que pode comprometer a eficiência da turbina.
- **Limites operacionais:** As turbinas possuem limites térmicos internos definidos pelos fabricantes, especialmente na câmara de combustão. Esses limites estão relacionados a fatores construtivos, como dilatação térmica, temperatura de fusão dos materiais, degradação de fluidos lubrificantes, entre outros. Ao atingir esse limite, torna-se necessário resfriar o ar de entrada para permitir o aumento da potência gerada, sem comprometer a integridade do sistema (GUERCIO; BINI, 2017). Esse comportamento foi observado nos dados e contribui para a correlação negativa entre a temperatura do ar e a variável objetivo.

Dessa forma, levando esses fatores em consideração, observa-se a correlação negativa entre a temperatura do ar e a variável objetivo.

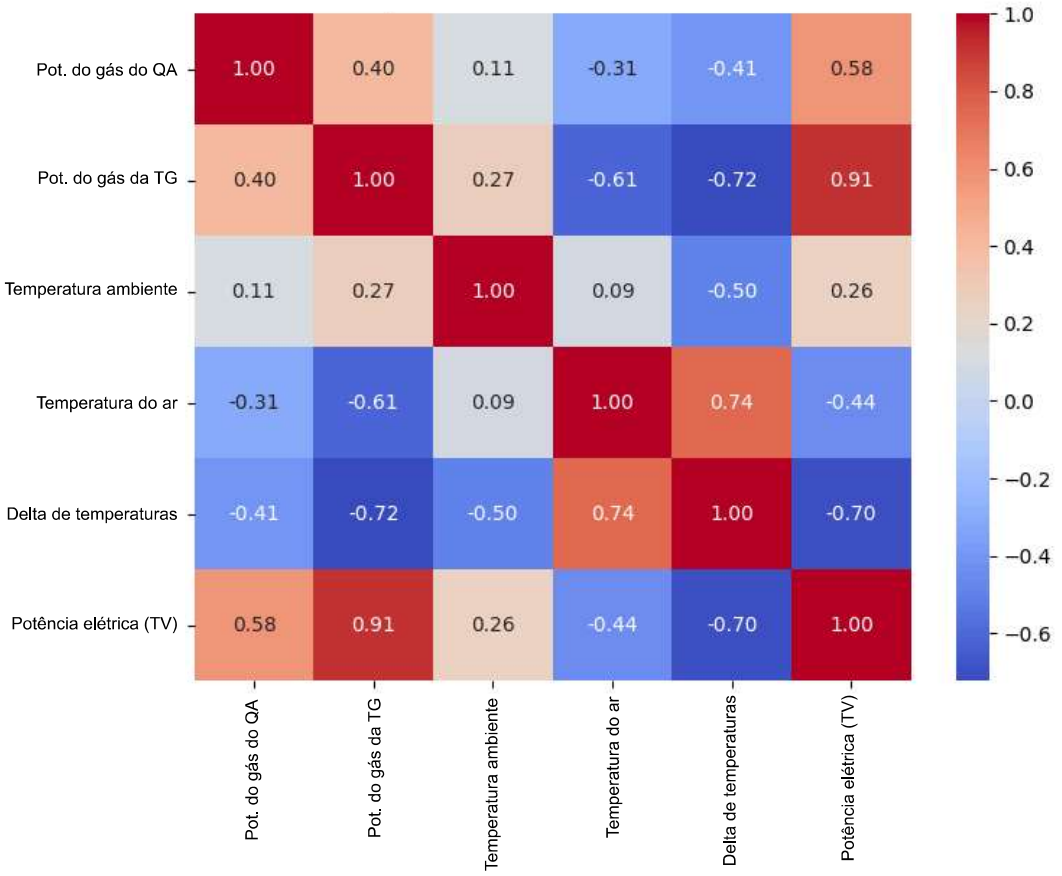
A pressão do ar, por outro lado, é um efeito ocasionado pela sucção do ar para a turbina, ocasionada pela operação da mesma. O que explica a correlação negativa, visto que a pressão de um fluido é inversamente proporcional à sua velocidade, de acordo com o princípio de Bernoulli, apesar das pequenas variações meteorológicas.

Além disso, a correlação relativamente baixa entre a temperatura e a pressão do ar sugere que essas duas variáveis fornecem informações complementares ao modelo. Esse resultado reforça que as condições ambientais exercem influência significativa sobre o desempenho da turbina, constituindo fatores essenciais a serem considerados na modelagem preditiva.

A Figura 14 apresenta a matriz de correlação entre as variáveis relacionadas à geração de potência da turbina a vapor. Observa-se que a potência do gás da TG apresenta uma alta correlação, diretamente proporcional à potência elétrica da TV, indicando que ambas variáveis estão altamente associadas. Esse resultado é coerente com o funcionamento

físico do sistema, no qual a energia térmica proveniente da turbina a gás influencia diretamente o desempenho da turbina a vapor.

Figura 14 – Matriz de correlação: Potência da turbina a vapor.



Fonte: De autoria própria.

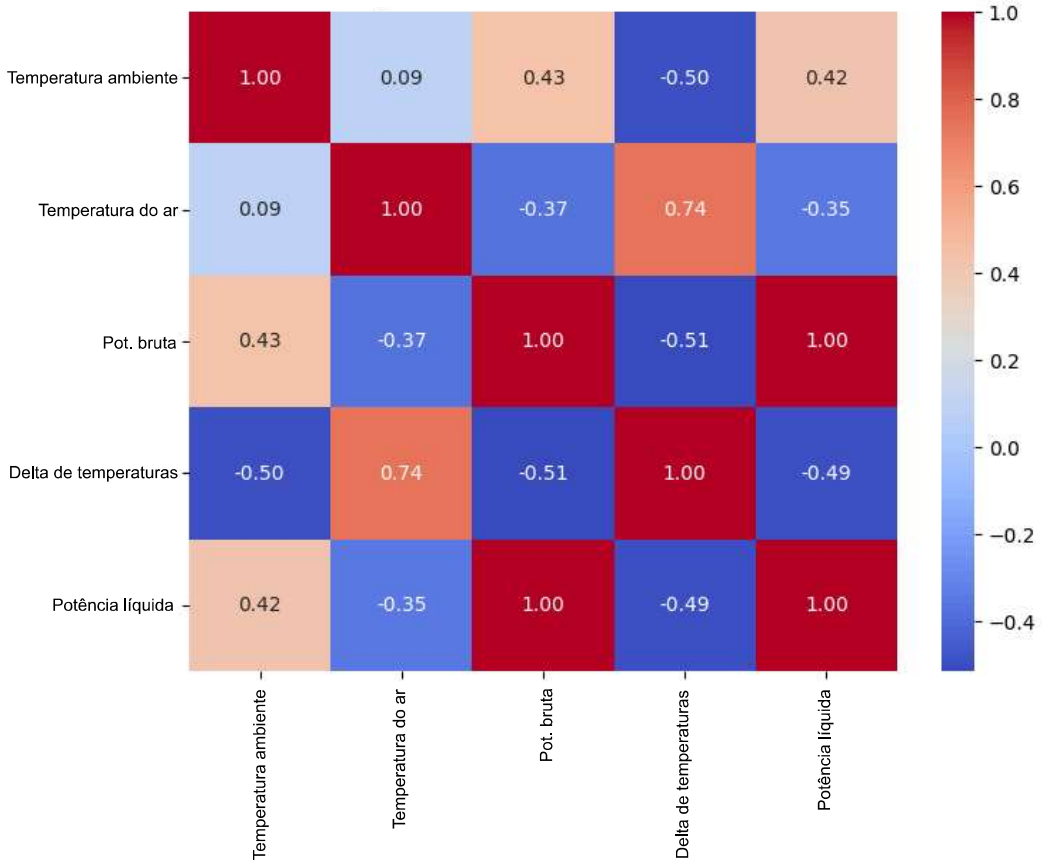
Por outro lado, variáveis térmicas como a temperatura do ar e o delta de temperaturas apresentam correlações negativas com a potência elétrica da TV, com coeficientes de $\rho = -0{,}44$ e $\rho = -0{,}70$, respectivamente. Isso sugere que aumentos nessas variáveis estão associados à redução da potência gerada, o que pode estar relacionado à menor densidade do ar em temperaturas mais elevadas, à redução da eficiência térmica do ciclo e aos limites operacionais da câmara de combustão, conforme discutido anteriormente.

A potência do gás do QA também apresenta correlação positiva com a potência elétrica da TV ($\rho = 0{,}58$), embora de forma menos intensa. Isso ocorre porque, embora o QA forneça calor diretamente para o sistema da TV, em muitos momentos ele se encontra desligado. Note que, neste ponto, o QA será relevante para o modelo, mas apenas quando este se encontrar ligado. Uma possível análise futura para este estudo seria estudar mais a fundo o sistema da TV, desmembrando o modelo em dois: um para o QA desligado e outro para o QA ligado, a fim de verificar se modelos especializados para cada estado de

operação trariam melhorias significativas de desempenho.

Finalmente, a Figura 15 apresenta a matriz de correlação entre as variáveis que influenciam na potência líquida do sistema. Observa-se que a potência bruta apresenta correlação virtualmente igual à máxima possível, com relação à potência líquida, o que é esperado, dado que a potência líquida é derivada diretamente da potência bruta, descontando perdas operacionais.

Figura 15 – Matriz de correlação: Potência líquida.



Fonte: De autoria própria.

Por outro lado, variáveis térmicas como a temperatura do ar e o delta de temperaturas apresentam correlações negativas com a potência líquida. A temperatura ambiente apresenta correlação positiva moderada com a potência líquida, o que sugere uma influência indireta sobre o desempenho do sistema, possivelmente relacionada a variações sazonais ou condições externas que afetam o rendimento global. Além disso, podem também indicar o comportamento da carga a qual a usina abastece, possivelmente indicando maior consumo em épocas de temperaturas mais elevadas.

Além disso, a exemplo do que ocorre com a pressão ar e a temperatura do ar para a potência da TV, a correlação relativamente baixa entre a temperatura ambiente e a

temperatura do ar indica que essas variáveis fornecem informações complementares ao modelo.

4.1 TÉCNICAS DE ML UTILIZADAS

Foram testadas técnicas simples e de conjunto, tanto homogêneas quanto heterogêneas. Alguns dos modelos simples foram utilizados como base para os conjuntos de modelos heterogêneos. Todas as técnicas avaliadas estão listadas na Tabela 3, apresentada no Capítulo 3. Nessa tabela, também estão descritos os hiperparâmetros utilizados e seus respectivos valores mínimos e máximos, empregados na otimização dos modelos. Para determinar esses hiperparâmetros, foi utilizado o otimizador bayesiano (NOGUEIRA, 2014–), considerando a média do RMSE obtido pela validação cruzada (*K-fold*), com cinco dobras como pontuação. Para todos os modelos, foram testadas 150 configurações diferentes, exceto nos modelos baseados em redes neurais, pois foi observado que, após 30 etapas, seu desempenho não apresentava melhorias significativas. Dessa forma, foram utilizadas 35 etapas para os modelos desse tipo, o que também compensa seu longo tempo de treinamento em comparação com as demais técnicas.

Para os modelos heterogêneos, os hiperparâmetros listados são referentes ao modelo meta. Para os modelos base, foram utilizados os melhores hiperparâmetros encontrados para cada uma das técnicas empregadas, sendo elas: catboost, adaboost, random forest e SVR.

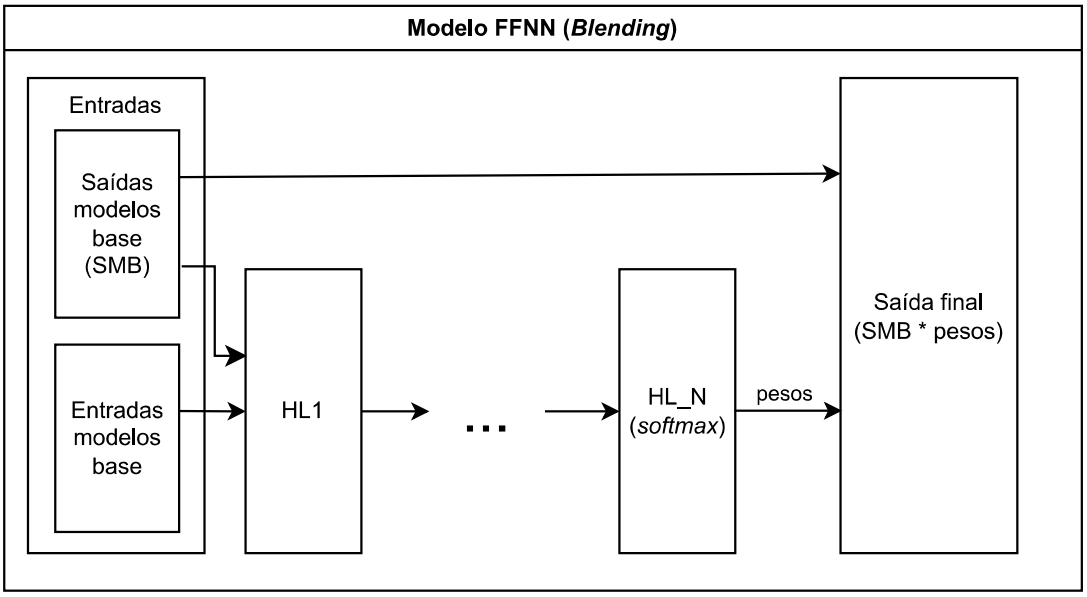
Com relação à divisão dos dados, para os modelos simples e homogêneos, foram utilizados 80% dos dados para treinamento e 20% para teste, a mesma divisão foi feita para os modelos heterogêneos, com a diferença que 12,5% do conjunto de treinamento (10% do total) foi separado para ser utilizado como o conjunto de extensão (*hold-out*). A divisão entre treinamento e testes foi feita de forma ordenada, utilizando dados mais recentes para testes e mais antigos para treinamento, já a divisão entre o conjunto de treinamento e extensão foi feita de forma aleatória.

4.1.1 O MODELO DE MÉDIA DINÂMICA - MODELO HIERÁRQUICO COM PONDERAÇÃO ADAPTATIVA

O modelo hierárquico com ponderação adaptativa, ou simplesmente modelo de média dinâmica, é um modelo *blending* que, até onde o autor tem conhecimento, acredita-se ser uma das principais inovações do presente estudo. Esse modelo surgiu da tentativa de mesclar as características de um modelo do tipo *blending* com o método de agregação baseado em média ponderada. Diferentemente de utilizar uma média ponderada das saídas dos modelos base com pesos fixos para gerar a saída final, o modelo emprega uma rede

neural para calcular os pesos, com base nas entradas e saídas dos modelos base. Esses pesos são obtidos por meio de uma camada com função de ativação *softmax* e são multiplicados pelas saídas dos modelos base, por meio de uma camada de multiplicação escalar (*dot*). Na Figura 16, é possível observar a topologia geral do modelo meta proposto.

Figura 16 – Topologia do modelo hierárquico com ponderação adaptativa - modelo meta.



Fonte: De autoria própria.

4.2 COMPARAÇÃO DE DESEMPENHO DOS MODELOS

Após encontrar os melhores hiperparâmetros para todos os modelos testados, os modelos foram utilizados para fazer previsões no conjunto de testes.

4.2.1 POTÊNCIA DA TG

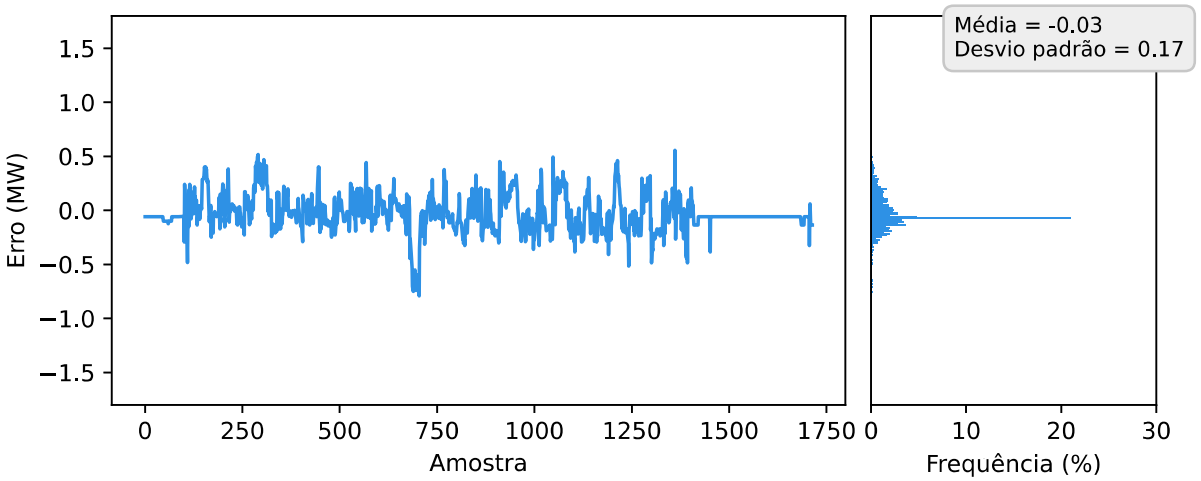
Na Tabela 5, estão dispostas as métricas obtidas pela comparação entre os valores reais e as previsões realizadas por cada um dos modelos testados para a turbina a gás.

Tabela 5 – Potência da TG: Métricas dos modelos.

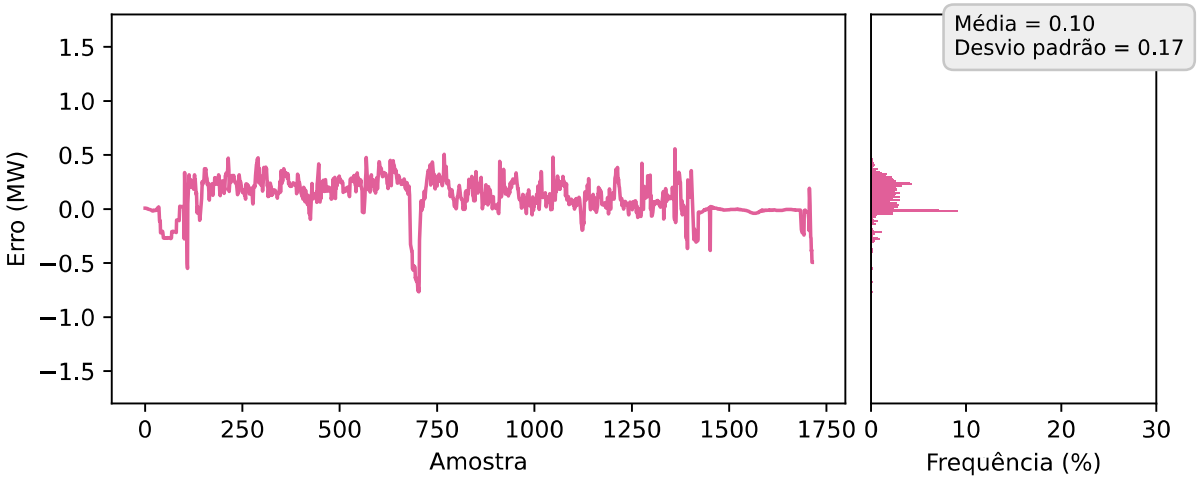
MODELO	RMSE	MAE	R2	NMAE (%)
ADABOOST	0,168530	0,126351	0,998849	1,469867
BLENDING FFNN - MÉDIA DINÂMICA	0,197859	0,153226	0,998414	1,782498
BLENDING FFNN - PADRÃO	0,191876	0,128670	0,998508	1,496835
BLENDING GPR	0,130488	0,088494	0,999310	1,029459
BLENDING RIDGE_CV	0,135736	0,092446	0,999253	1,075438
BLENDING SVR	0,128939	0,095770	0,999326	1,114109
CATBOOST	0,241075	0,146928	0,997645	1,709235
FFNN	0,320360	0,247371	0,995841	2,877711
GPR	0,193916	0,120715	0,998476	1,404297
RANDOM FOREST	0,154784	0,103339	0,999029	1,202164
STACKING GPR	0,147524	0,096881	0,999118	1,127037
STACKING RIDGE_CV	0,146784	0,096534	0,999127	1,122991
SVR	0,197049	0,137076	0,998427	1,594628

Na Figura 17, é possível observar os gráficos do erro em função do tempo, para o conjunto de testes, bem como sua distribuição, para o modelo da potência da TG para cada modelo.

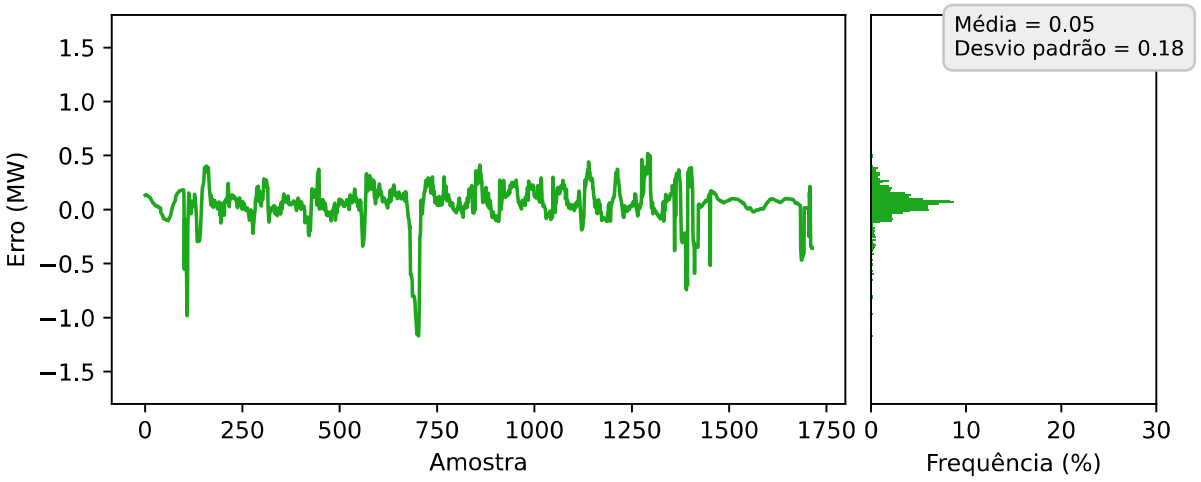
Figura 17 – Erro dos modelos: Potência da TG.



(a) Adaboost.

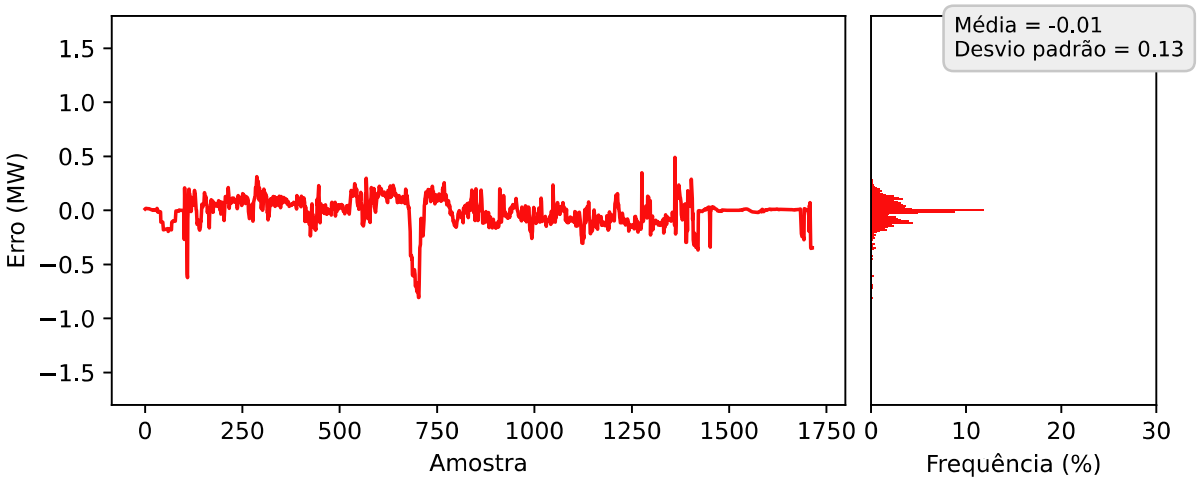


(b) Blending FFNN - Média dinâmica.

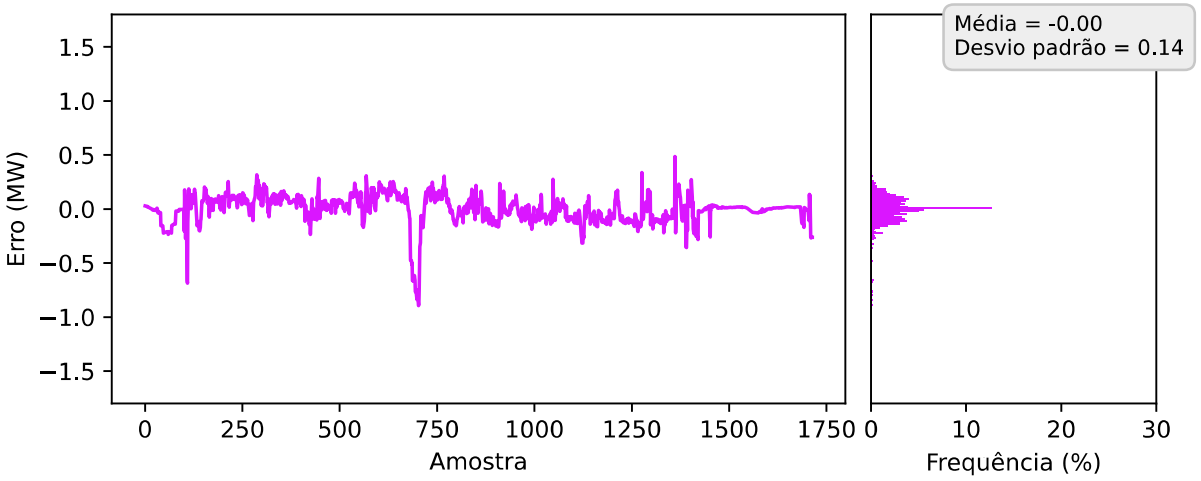


(c) Blending FFNN.

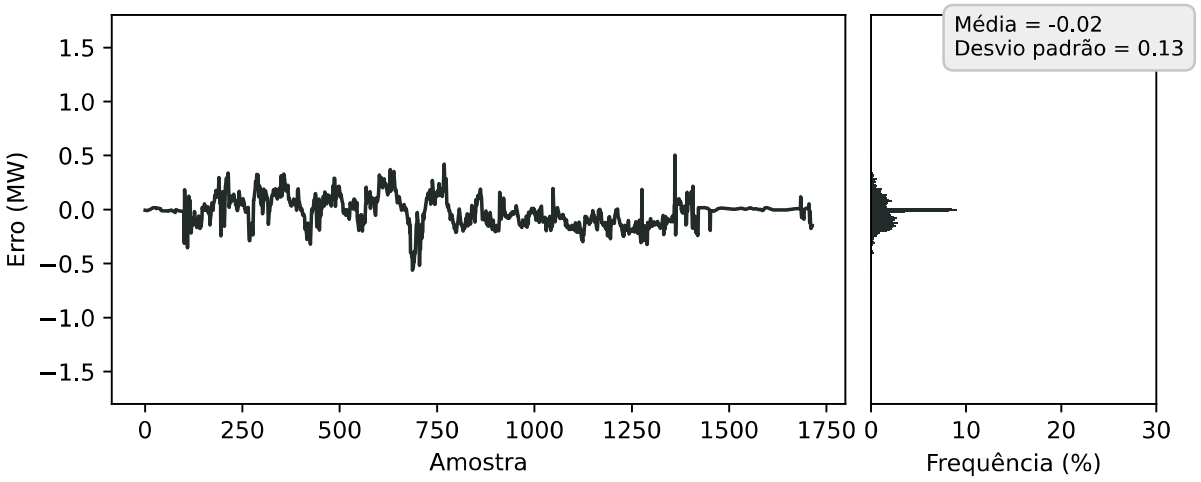
Figura 17 – Erro dos modelos: Potência da TG.



(d) Blending GPR.

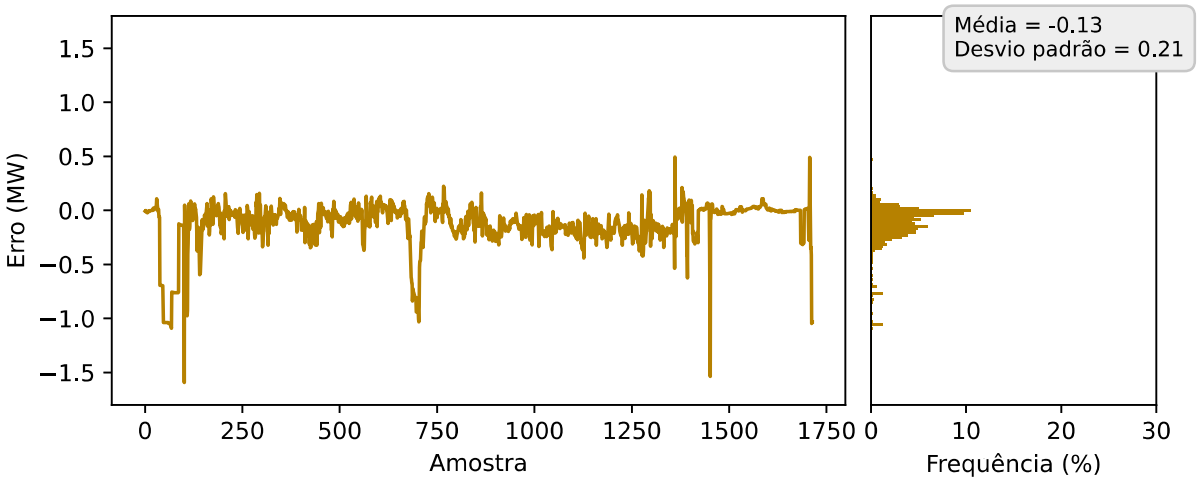


(e) Blending Ridge CV.

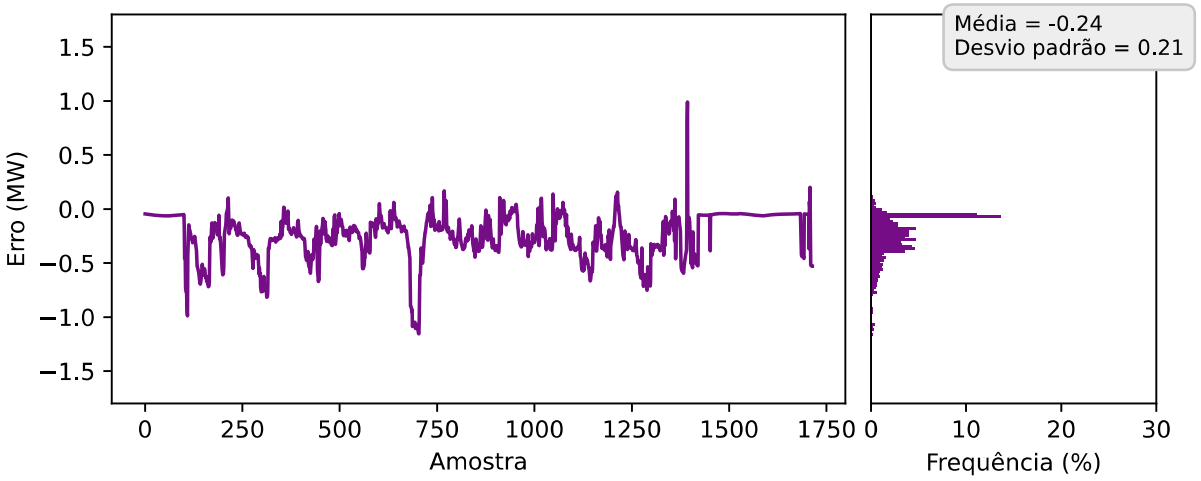


(f) Blending SVR.

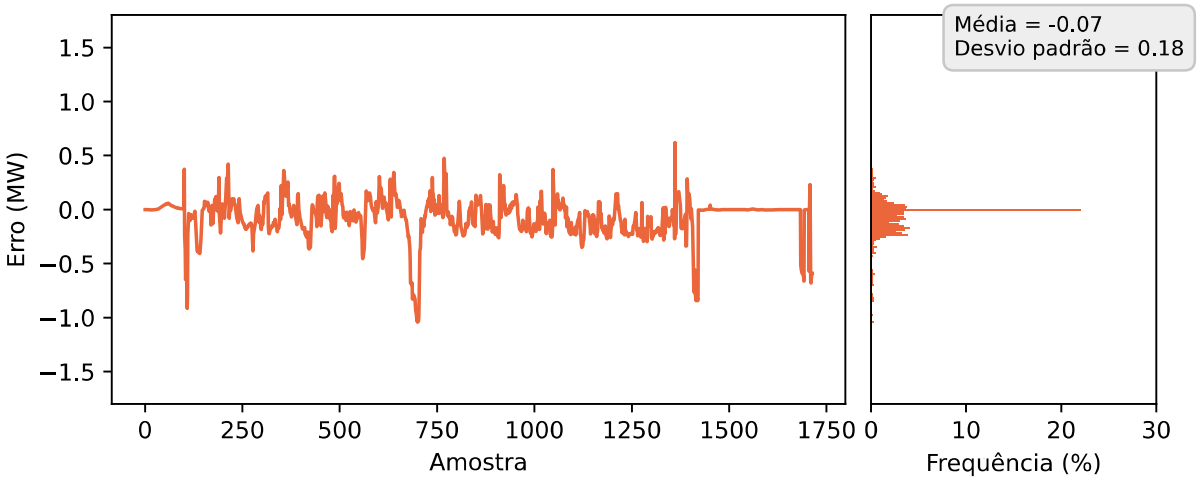
Figura 17 – Erro dos modelos: Potência da TG.



(g) Catboost.

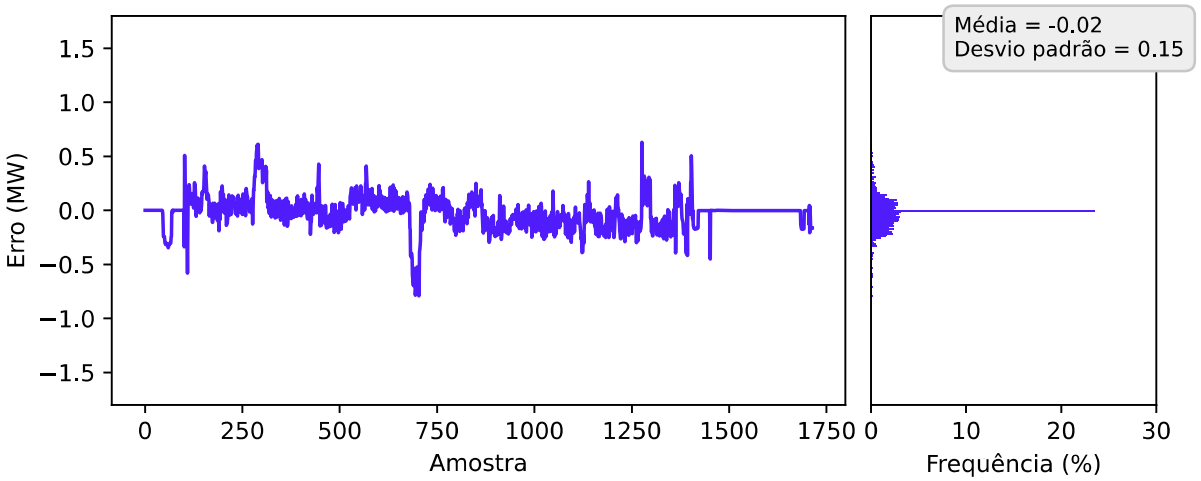


(h) FFNN.

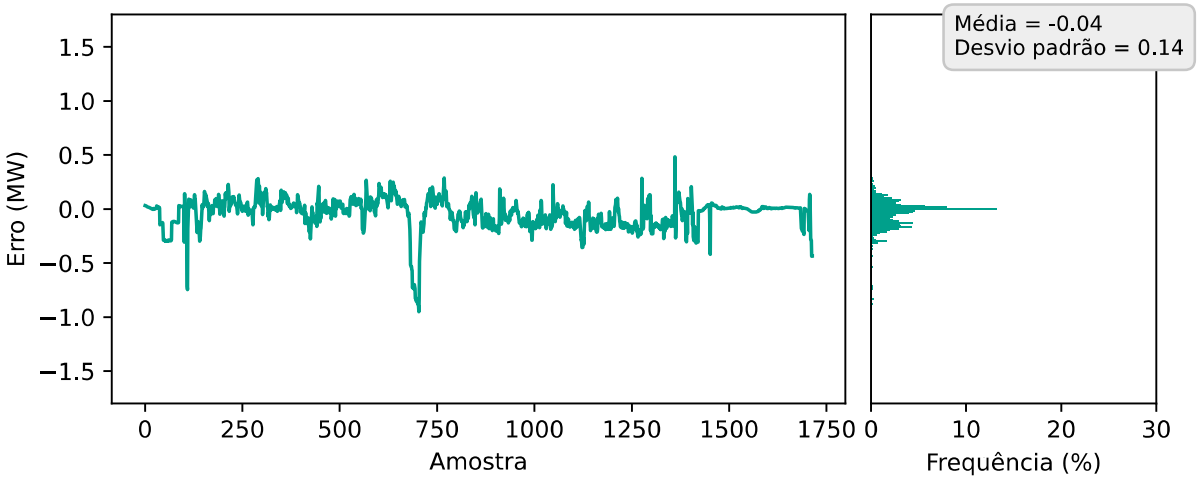


(i) GPR.

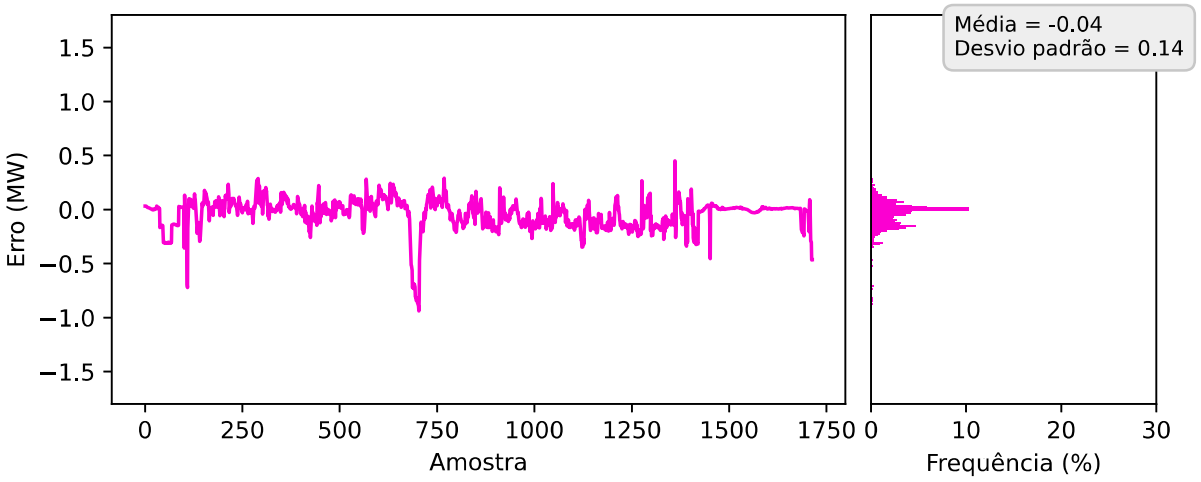
Figura 17 – Erro dos modelos: Potência da TG.



(j) RF.

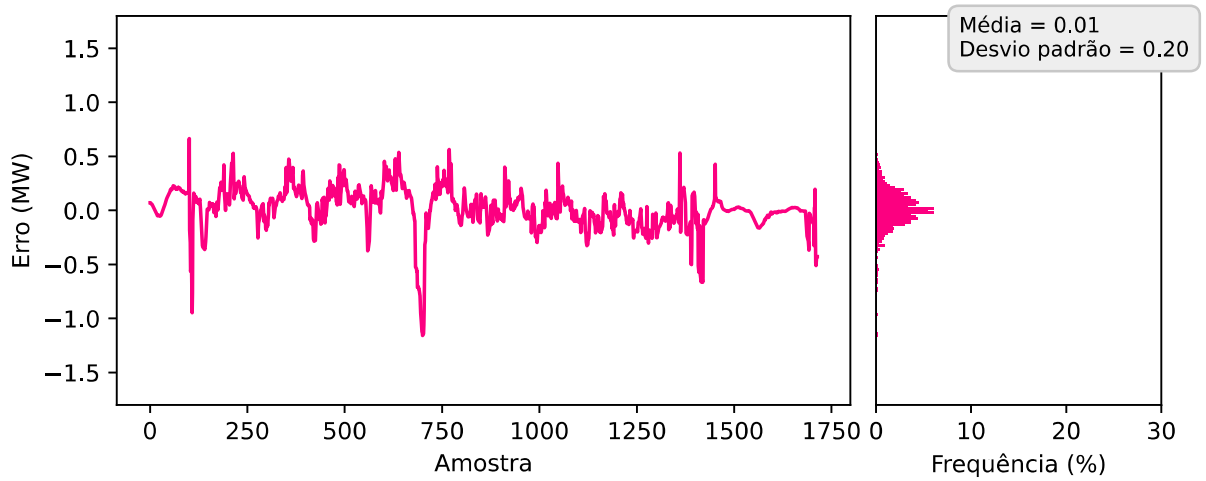


(k) Stacking GPR.



(l) Stacking Ridge CV.

Figura 17 – Erro dos modelos: Potência da TG.



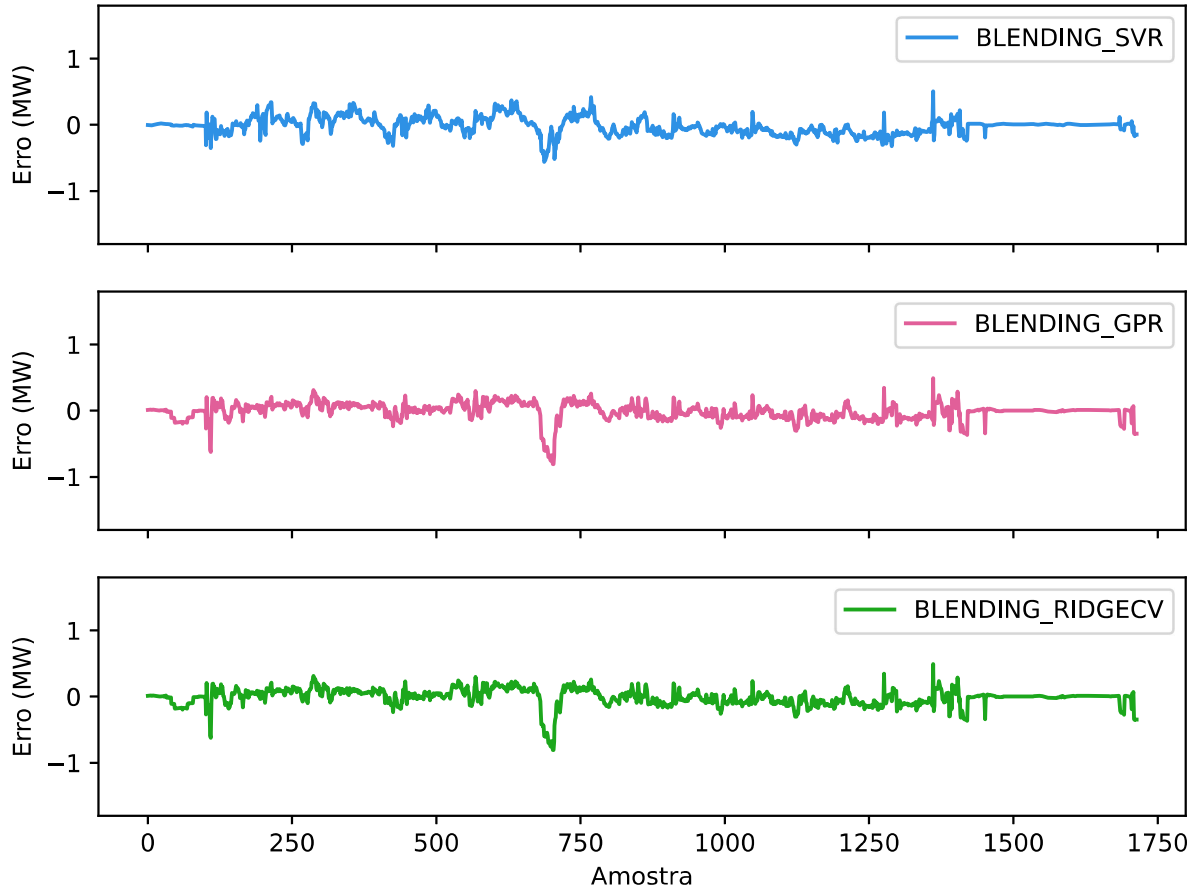
(m) SVR.

Fonte: De autoria própria.

Analisando os dados da Tabela 5 e a Figura 17, é possível notar que em geral os modelos de conjunto têm um melhor desempenho que os modelos simples e os modelos base que os compõe, no caso dos heterogêneos. Os modelos de melhor desempenho foram os modelos *blending* com os estimadores finais SVR, GPR e Ridge CV, respectivamente. Na Figura 18, é possível visualizar os sinais dos três melhores modelos para a potência da TG.

Note que, devido à característica do GPR, é possível fornecer um intervalo de confiança dinâmico para cada predição realizada por este modelo, visto que há a opção de fazê-lo retornar também o desvio padrão da média das previsões feitas por ele (previsão final do estimador), sendo esta uma vantagem que os diferencia dos outros modelos. É importante destacar também a distribuição do erro destes modelos, que se aproximam bastante de uma distribuição normal, com médias muito próximas de zero e desvio padrão menor que 0,15 MW.

Figura 18 – Erro dos melhores modelos: Potência da TG.



4.2.2 POTÊNCIA DA TV

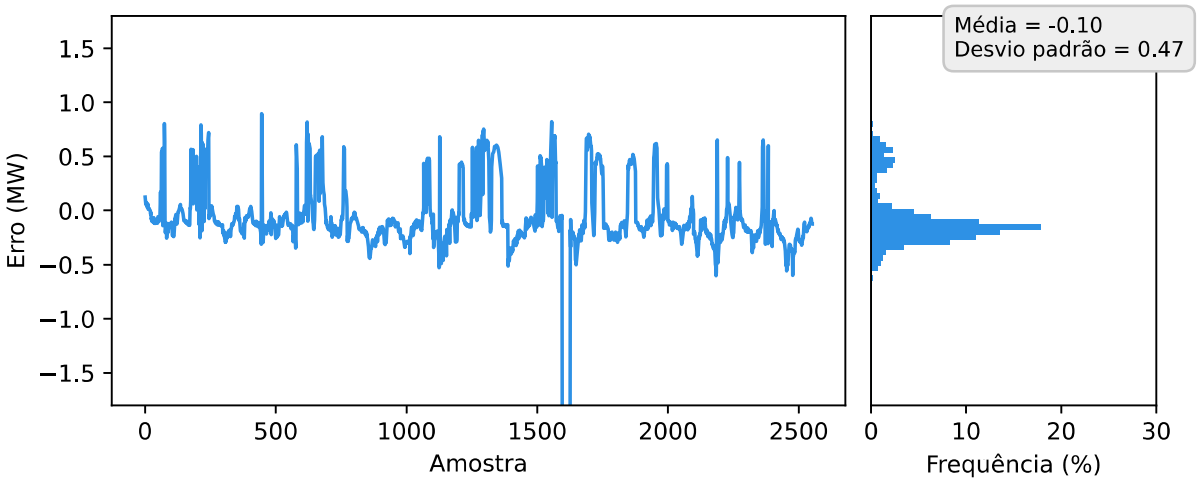
De modo semelhante, na Tabela 6, estão dispostas as métricas obtidas pela comparação entre os valores reais e as previsões realizadas por cada um dos modelos testados para a turbina a vapor.

Tabela 6 – Potência da TV: Métricas dos modelos.

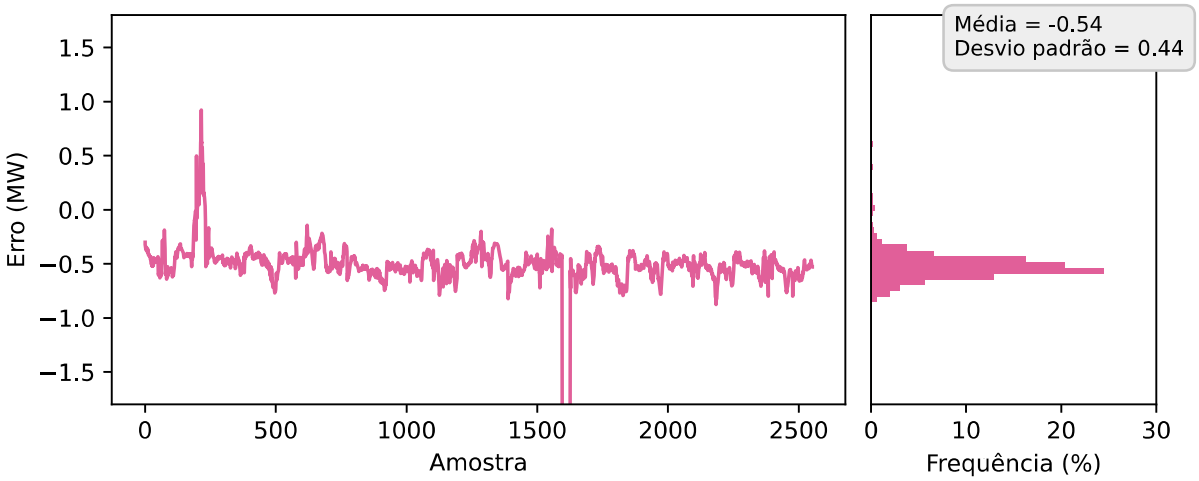
MODELO	RMSE	MAE	R2	NMAE (%)
ADABOOST	0,4753	0,253238	0,29735	6,6554
BLENDING FFNN - MÉDIA DINÂMICA	-	-	-	-
BLENDING FFNN - PADRÃO	0,700966	0,549978	-0,542649	14,4542
BLENDING GPR	0,835698	0,714018	-1,192663	18,7654
BLENDING RIDGE_CV	0,838096	0,715688	-1,205267	18,8093
BLENDING SVR	1,012132	0,913146	-2,216233	23,9988
CATBOOST	0,464205	0,152389	0,323461	4,005
FFNN	0,467175	0,190974	0,314776	5,0191
GPR	0,447631	0,131425	0,370908	3,454
RANDOM FOREST	0,453207	0,152443	0,355138	4,0064
STACKING GPR	0,439374	0,111412	0,393904	2,9281
STACKING RIDGE_CV	0,439466	0,111701	0,39365	2,9357
SVR	0,446628	0,118588	0,373726	3,1167

Na Figura 19, é possível observar o gráfico do erro em função do tempo, para o conjunto de testes, bem como sua distribuição, para o modelo da potência da TV considerando os diferentes modelos utilizados.

Figura 19 – Erro dos modelos: Potência da TV.

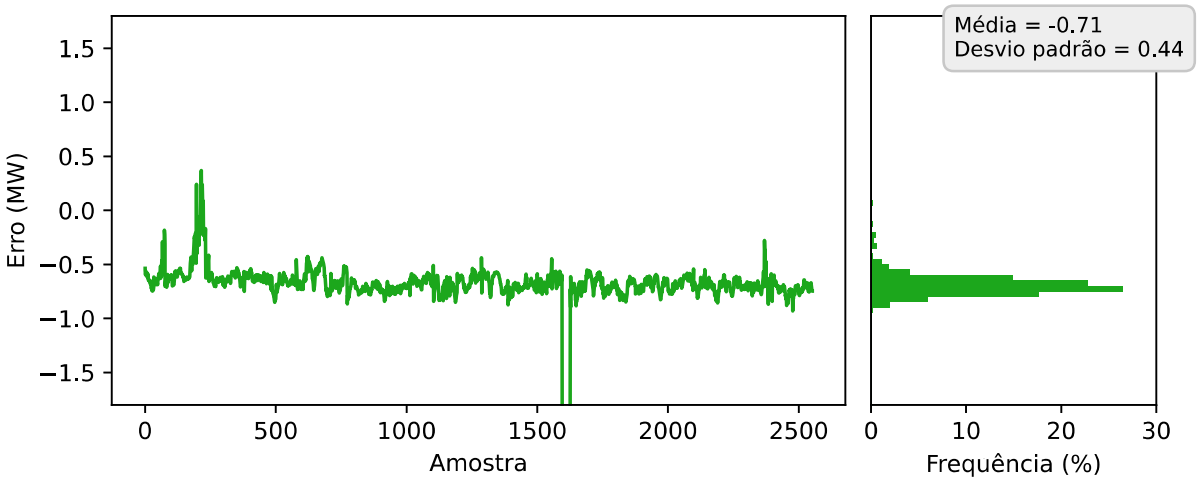


(a) Adaboost.

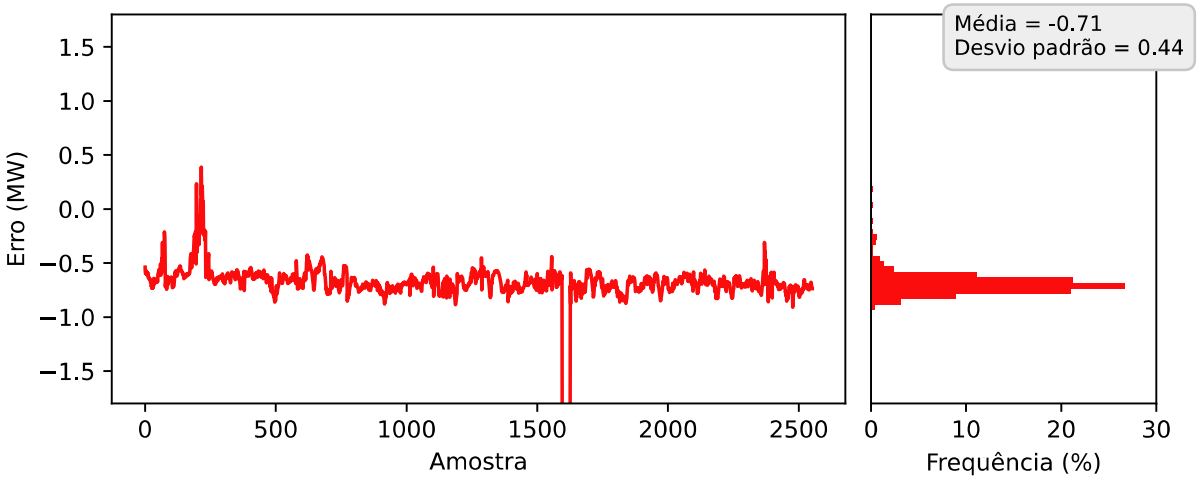


(b) Blending FFNN.

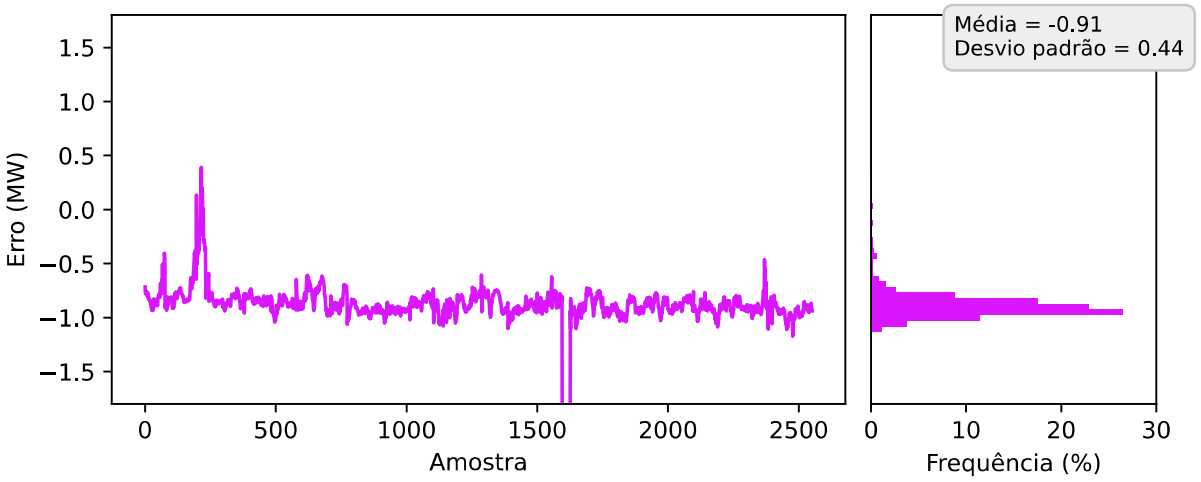
Figura 19 – Erro dos modelos: Potência da TV.



(c) Blending GPR.

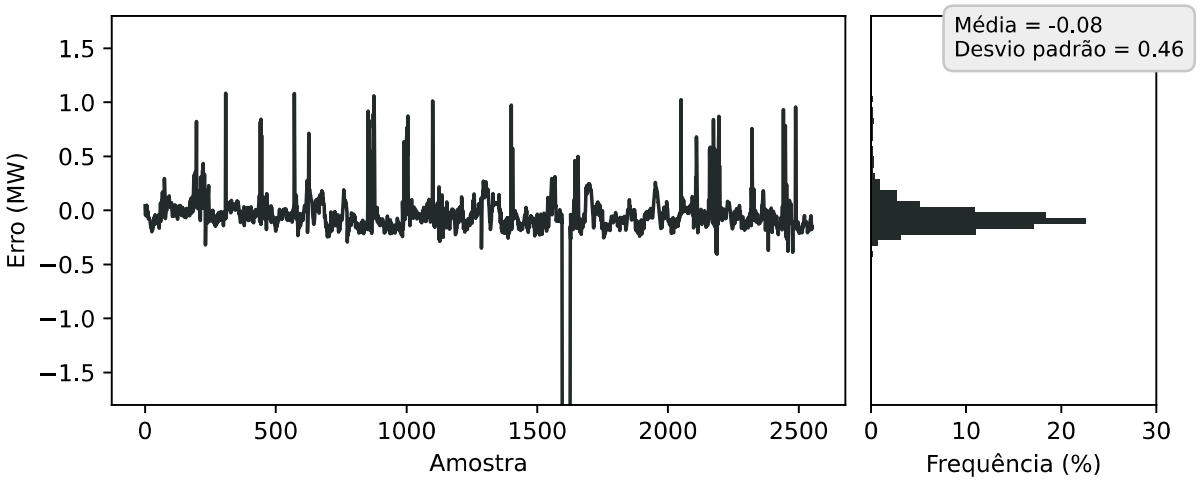


(d) Blending Ridge CV.

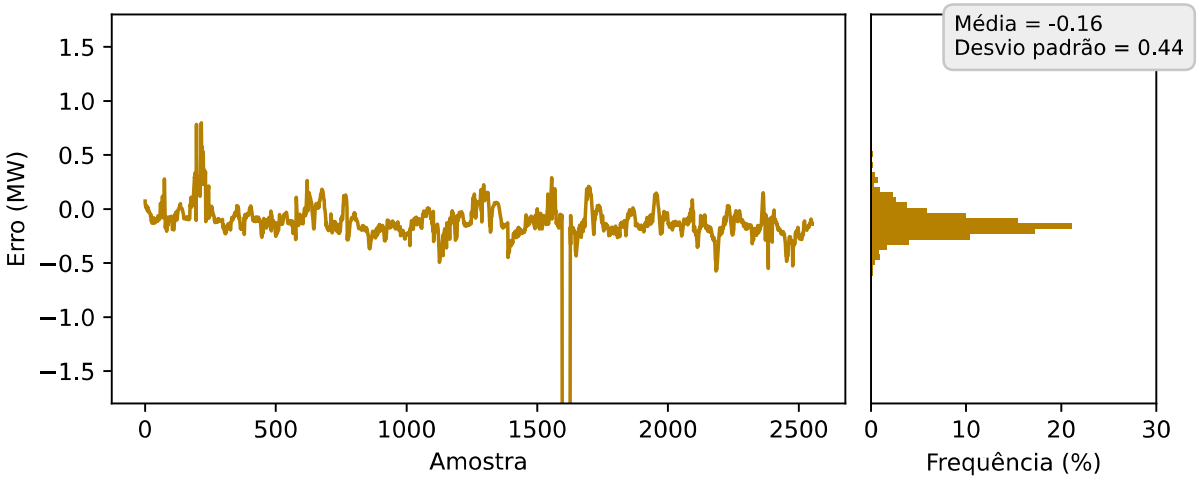


(e) Blending SVR.

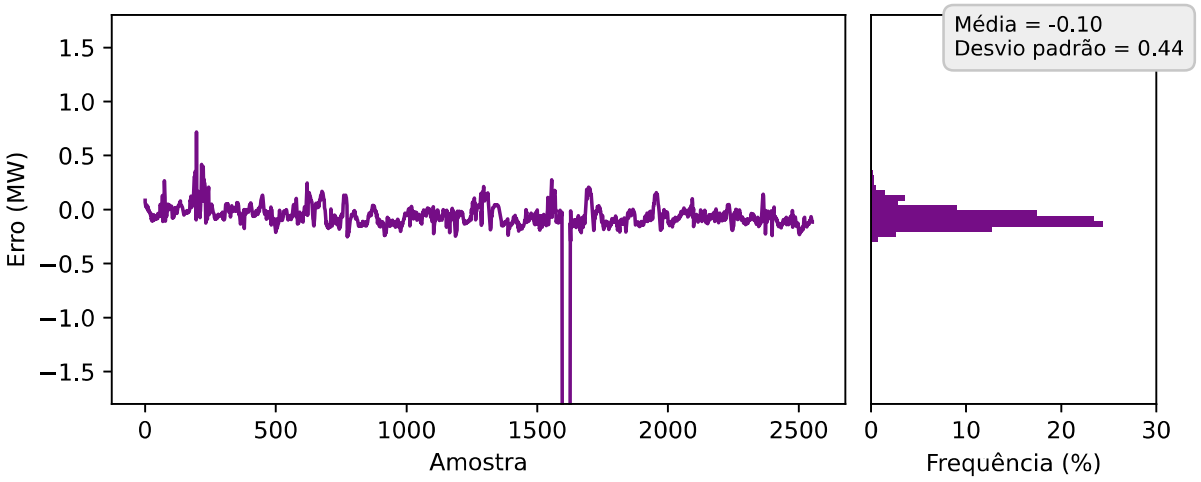
Figura 19 – Erro dos modelos: Potência da TV.



(f) Catboost.

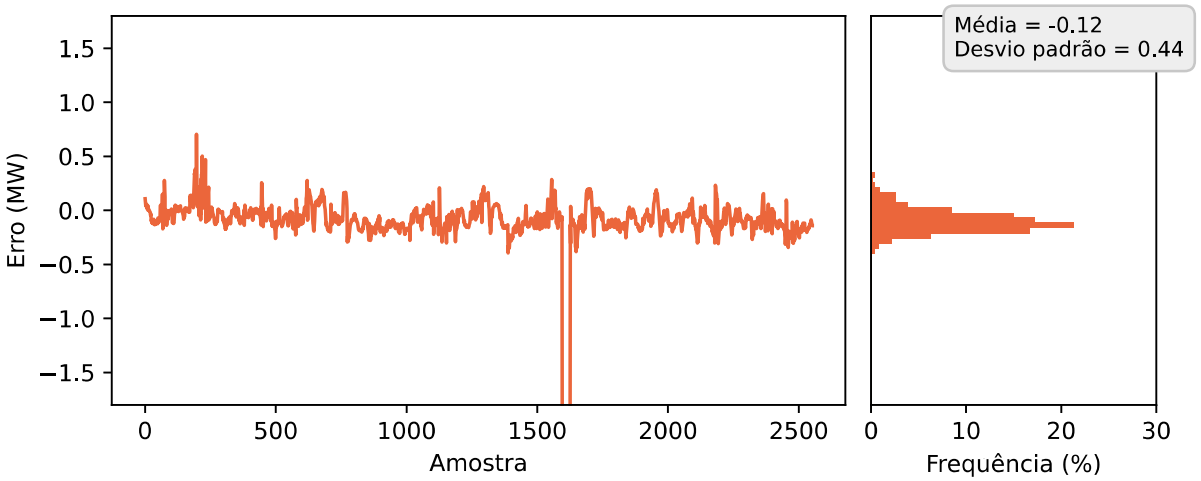


(g) FFNN.

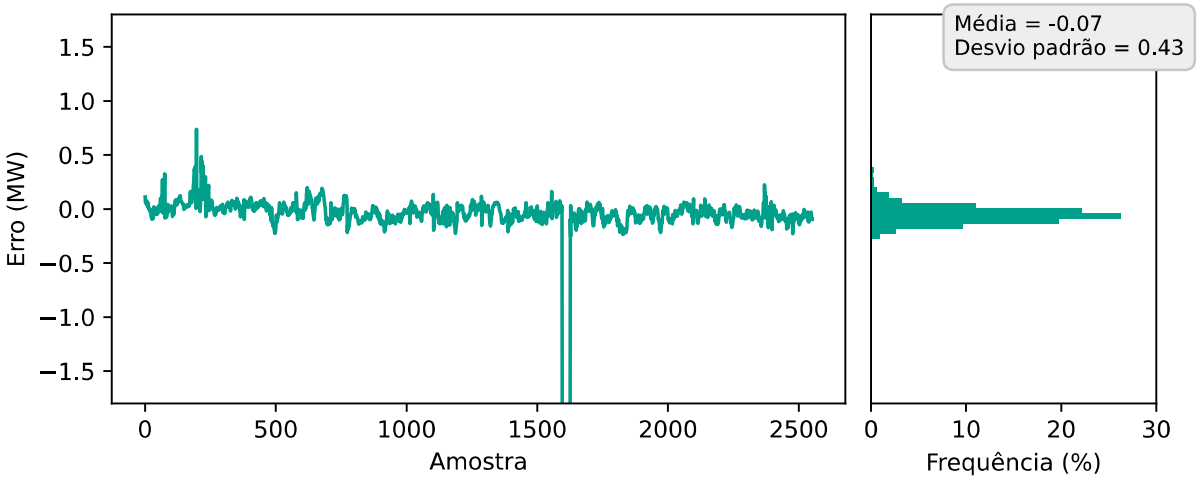


(h) GPR.

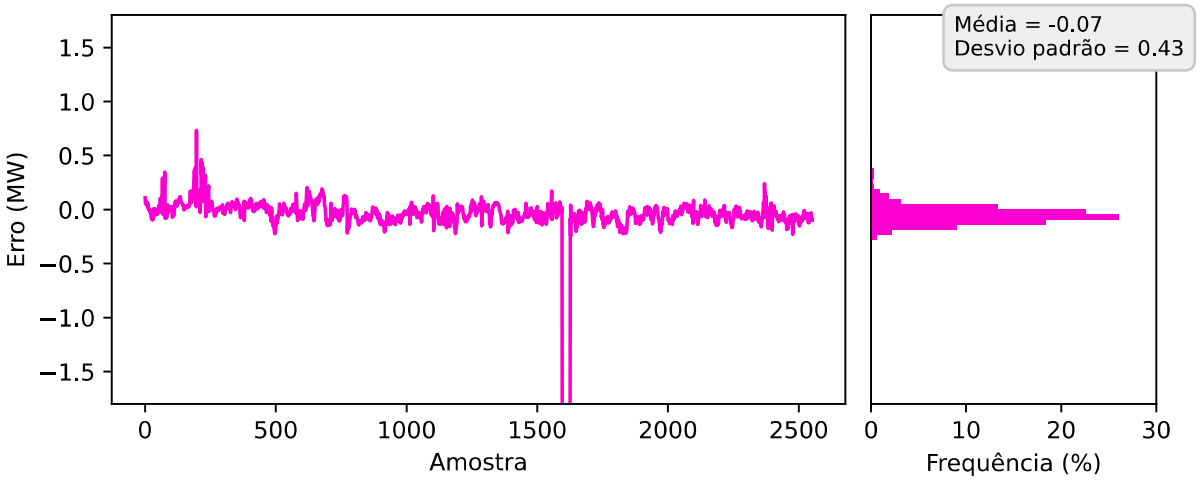
Figura 19 – Erro dos modelos: Potência da TG.



(i) RF.

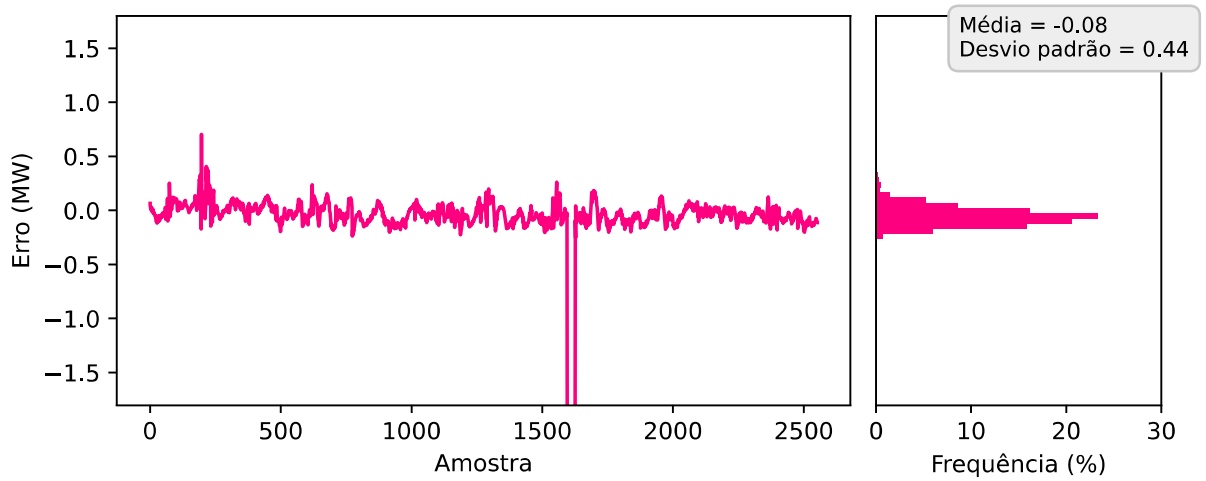


(j) Stacking GPR.



(k) Stacking Ridge CV.

Figura 19 – Erro dos modelos: Potência da TV.



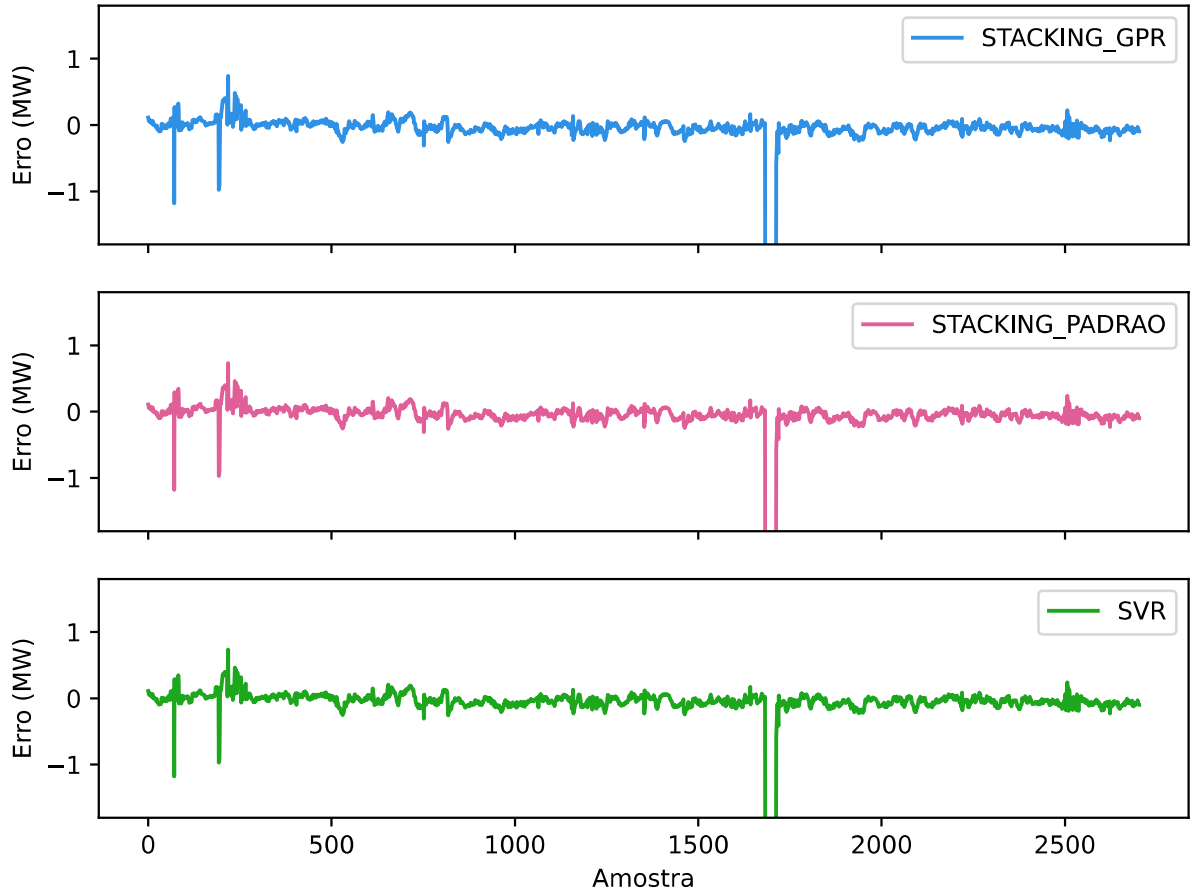
(l) SVR.

Fonte: De autoria própria.

Analisando os dados da Tabela 6 e a Figura 19, é possível concluir que, assim como para os modelos da TG, os conjuntos de modelos obtiveram melhores desempenhos também para os dados da TV. Note que, para este conjunto de dados, a métrica R^2 não atinge valores altos, como ocorreu com os modelos da TG, mesmo para o melhor modelo, que alcançou um erro percentual $NMAE(\%) = 2,92$. Isso ocorreu porque os dados da TV apresentam uma variabilidade intrínseca maior que os dados da TG, além de alguns *outliers* terem sido incluídos no conjunto final de dados. Isso, como indicado no Capítulo 2, impacta negativamente no coeficiente de determinação (R^2).

Para esse conjunto de dados, o modelo *blending* de média dinâmica não convergiu. O experimento foi cancelado após quatro horas, sem resultados. Os modelos de melhor desempenho foram os modelos *stacking* GPR, *stacking* ridge cv e o modelo simples SVR, respectivamente. Na Figura 20, é possível visualizar os sinais dos três melhores modelos para a potência da TV.

Figura 20 – Erro dos melhores modelos: Potência da TV.



4.2.3 POTÊNCIA LÍQUIDA

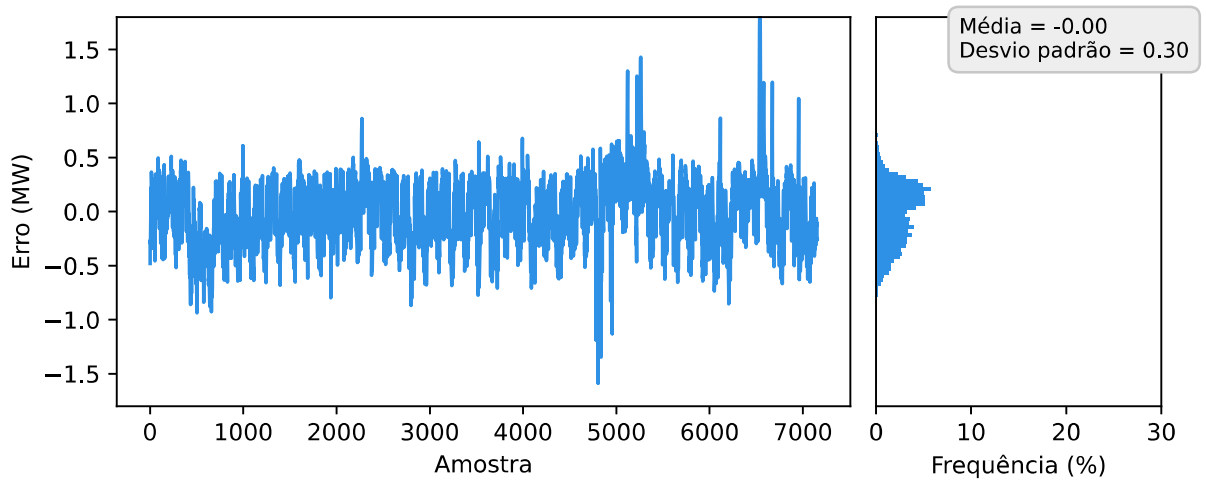
Finalmente, na Tabela 7, estão dispostas as métricas obtidas pela comparação entre os valores reais e as previsões realizadas por cada um dos modelos testados para a potência líquida.

Tabela 7 – Potência Líquida: Métricas dos modelos.

MODELO	RMSE	MAE	R2	NMAE(%)
ADABOOST	0,301618	0.243754	0,986657	1,4443
BLENDING FFNN - MÉDIA DINÂMICA	-	-	-	-
BLENDING FFNN - PADRÃO	-	-	-	-
BLENDING GPR	0,325616	0,273169	0,984449	1,6186
BLENDING RIDGE_CV	0,316419	0,259056	0,985315	0,01535
BLENDING SVR	0,382095	0,315942	0,978586	1,872
CATBOOST	0,150086	0,097869	0,996696	0,5799
FFNN	0,733729	0,681594	0,921037	4,0386
GPR	0,202449	0,112756	0,993988	0,6681
RANDOM FOREST	0,285543	0,228816	0,988041	1,3558
STACKING GPR	-	-	-	-
STACKING RIDGE_CV	0,174931	0,131129	0,995512	0,7770
SVR	0,147610	0,109327	0,996804	0,6478

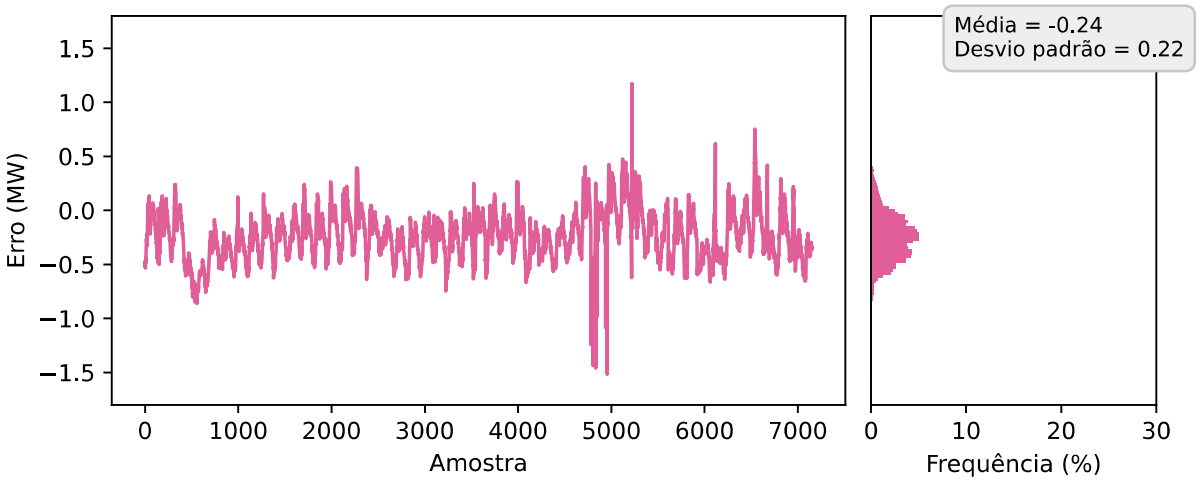
Na Figura 21, é possível observar o gráfico do erro em função do tempo, para o conjunto de testes, bem como sua distribuição, para o modelo da potência líquida.

Figura 21 – Erro dos modelos: Potência líquida.

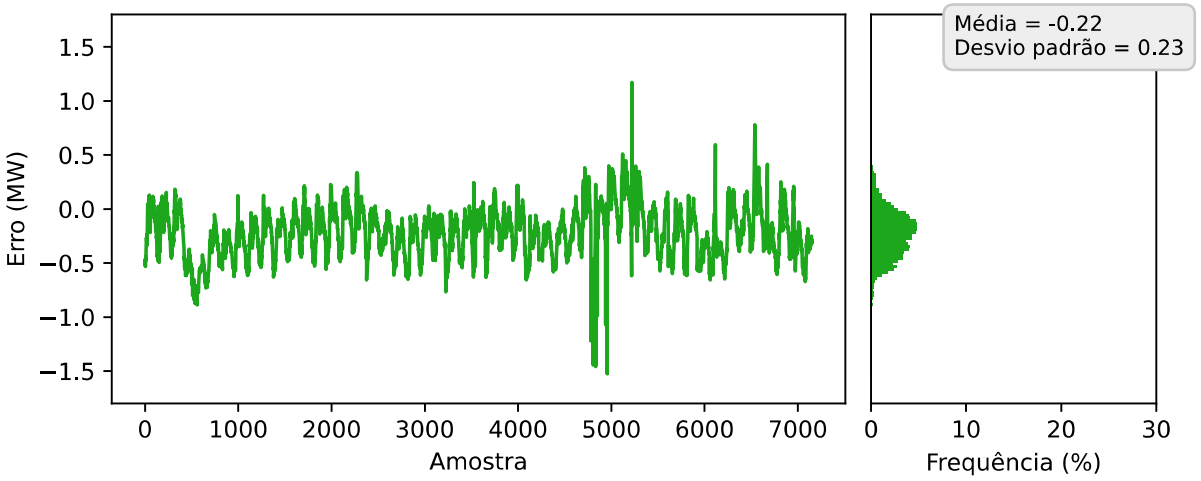


(a) Adaboost.

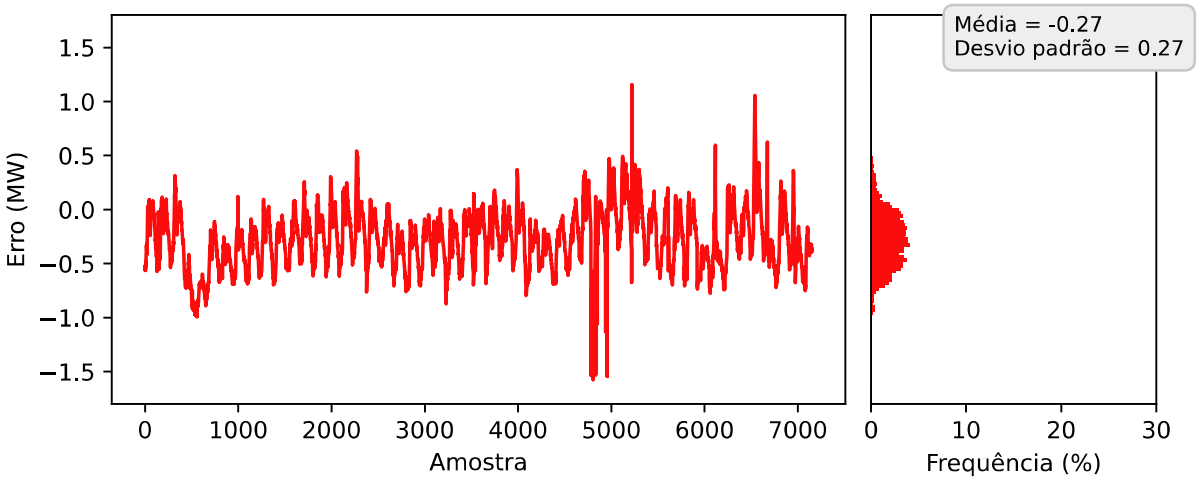
Figura 21 – Erro dos modelos: Potência líquida.



(b) Blending GPR.

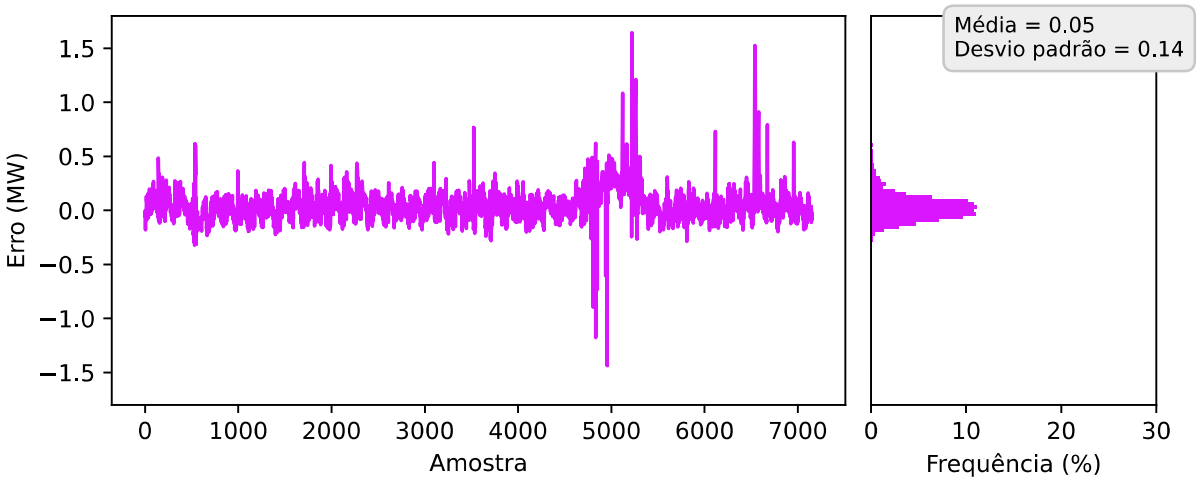


(c) Blending Ridge CV.

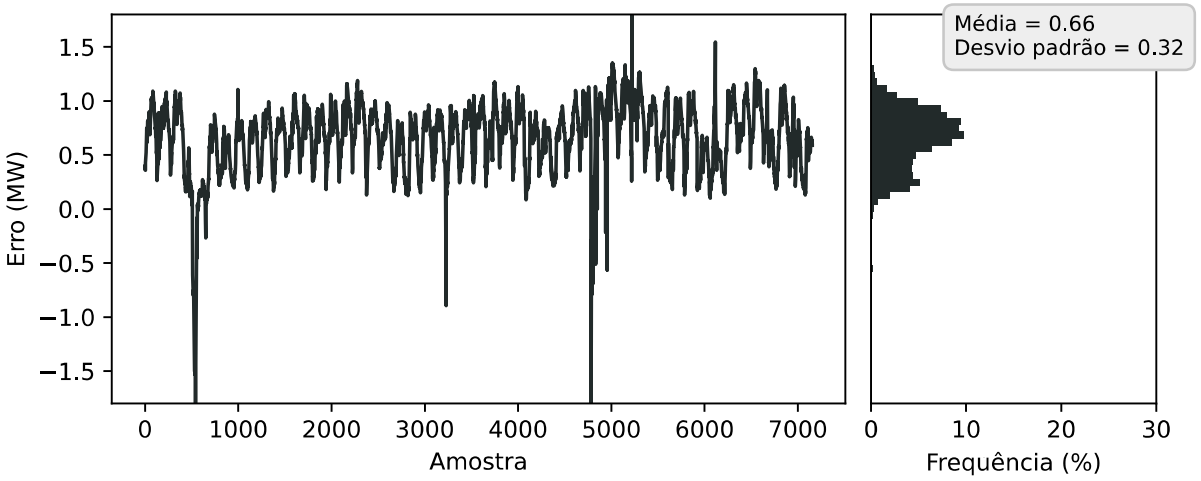


(d) Blending SVR.

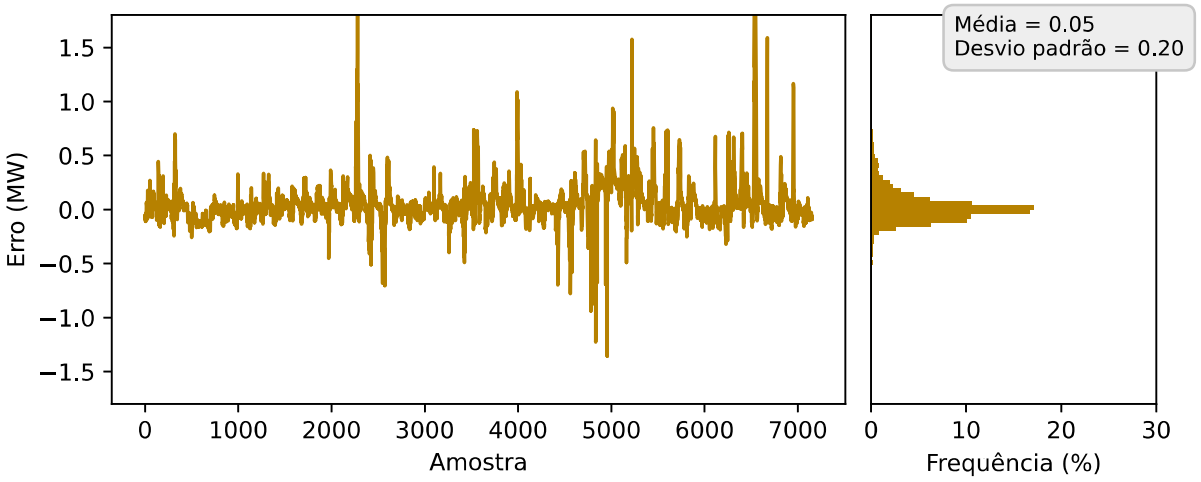
Figura 21 – Erro dos modelos: Potência líquida.



(e) Catboost.

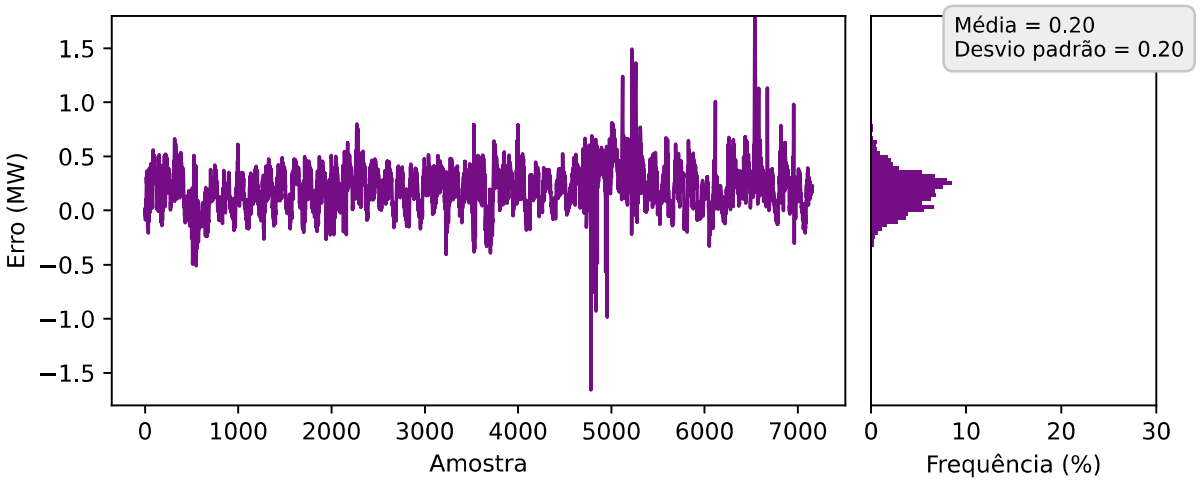


(f) FFNN.

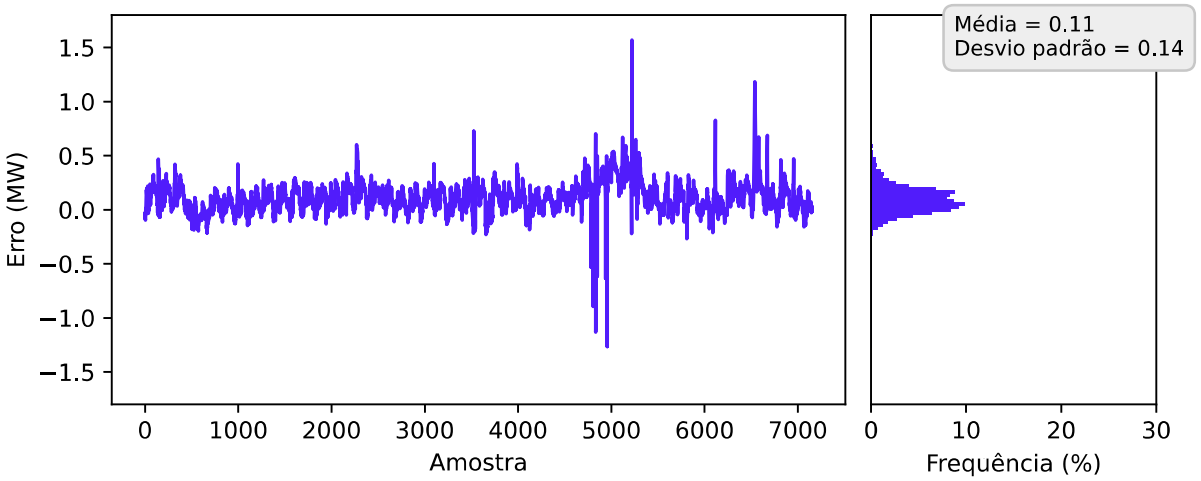


(g) GPR.

Figura 21 – Erro dos modelos: Potência líquida.

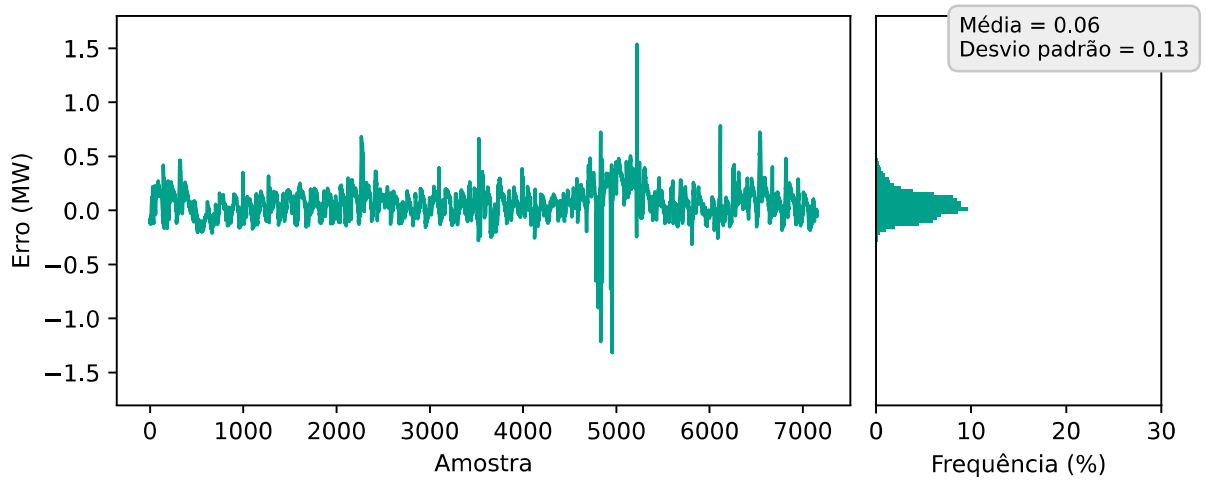


(h) RF.



(i) Stacking Ridge CV.

Figura 21 – Erro dos modelos: Potência líquida.



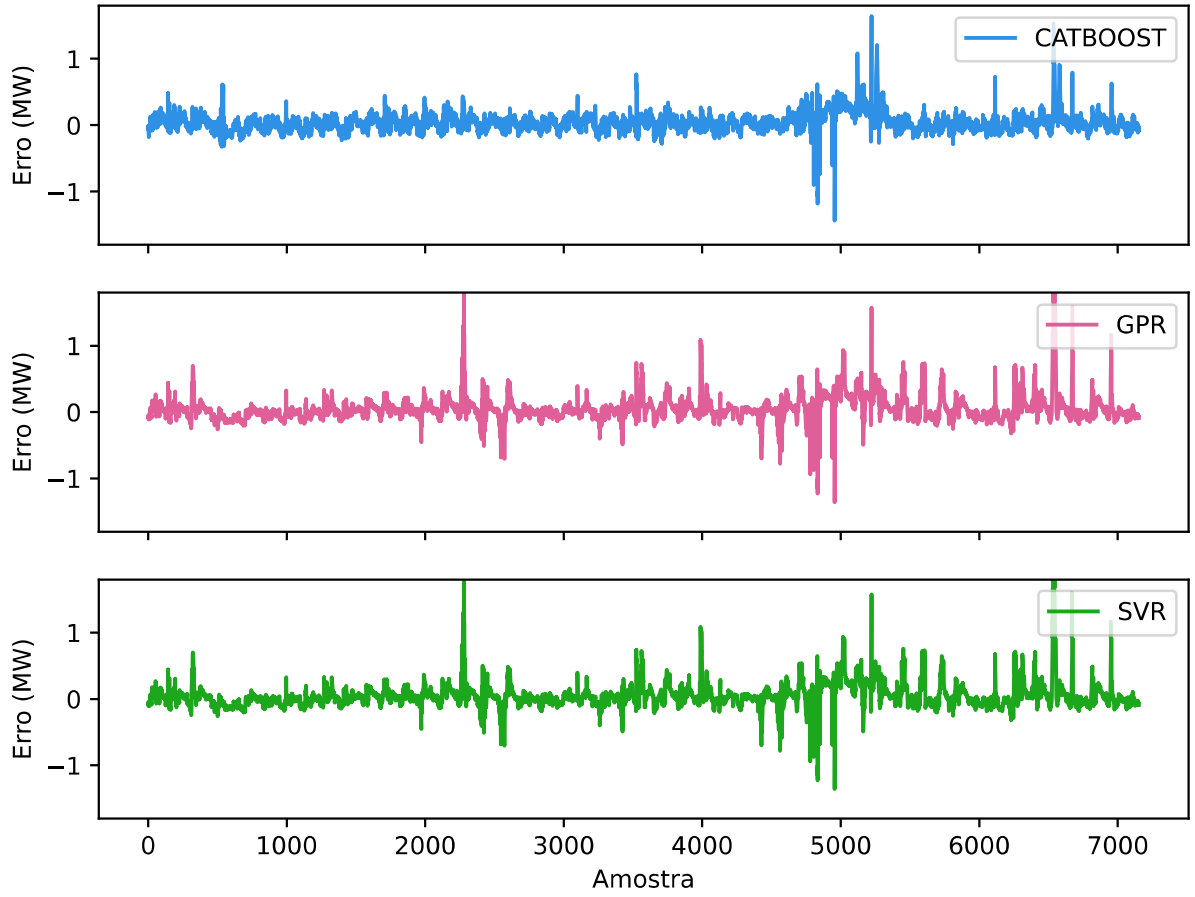
(j) SVR.

Fonte: De autoria própria.

De forma semelhante ao realizado para os modelos das potências da TG e da TV, analisando os dados da Tabela 7 e a Figura 21, é possível concluir que os conjuntos de modelos obtiveram melhores desempenhos também para os dados da potência líquida. Os modelos *blending* de média dinâmica, *blending* FFNN padrão e *stacking* GPR não convergiram. De forma semelhante ao realizado para os modelos da TV, o experimento foi cancelado após quatro horas, sem resultados para os modelos supracitados.

Os modelos de melhor desempenho foram os modelos *catboost*, e os modelos simples GPR e SVR, respectivamente. Na Figura 22, é possível visualizar os sinais dos três melhores modelos para a potência líquida. Para esta variável, dois dos três melhores modelos foram modelos simples. Apesar disso, o modelo *stacking ridge* CV obteve um desempenho bem próximo aos primeiros colocados. Outro ponto observado foi o desempenho superior do *catboost* para essa variável, com relação aos demais, visto que para as outras variáveis ele obteve resultados medianos, porém consistentes.

Figura 22 – Erro dos melhores modelos: Potência líquida.



5 DISCUSSÃO E CONCLUSÕES

Ao longo deste documento, foi explorada uma metodologia aplicada ao desenvolvimento e escolha de modelos de *machine learning* no contexto de sistemas gêmeos digitais de uma usina termelétrica a gás de ciclo combinado.

Foi feita uma revisão a respeito das métricas mais utilizadas para regressão linear na literatura. Para a comparação dos modelos, foram escolhidos o erro absoluto médio (MAE), a raíz do erro quadrático médio (RMSE), o coeficiente de determinação (R^2) e o erro absoluto médio normalizado (NMAE). Com relação às métricas, o NMAE foi o maior diferencial do presente trabalho, substituindo o erro médio absoluto percentual (MAPE).

Com relação ao desenvolvimento dos modelos, foram escolhidas as técnicas de ML simples e de conjunto mais utilizadas na literatura, além de um modelo próprio, baseado em médias ponderadas com pesos dinâmicos. Os resultados obtidos demonstraram que técnicas de conjunto (*ensemble*), em especial os modelos de *stacking* e *blending*, apresentaram desempenho superior em relação aos modelos individuais, confirmando a relevância da combinação de estimadores para aumentar a robustez e a precisão das previsões. Os modelos testados demonstraram um bom desempenho, mesmo utilizando uma pequena quantidade de variáveis de entrada para cada um dos modelos.

Apesar de ter sido testado o modelo de mêdias dinâmicas ponderadas, este modelo não atingiu os resultados esperados. Em dois dos três testes realizados, o modelo não convergiu. Comparando-o diretamente com o modelo similar, o único modelo que convergiu alcançou um desempenho comparável ao modelo com estimador final FFNN. Em termos de tempo de compilação, o modelo também não apresentou melhora com relação à sua contraparte tradicional. Contudo, o conceito é interessante e, embora tenha atingido um desempenho mediano com relação aos demais modelos, mais testes e ajustes na arquitetura são necessários para dar um veredito final a respeito dessa topologia.

A técnica GPR foi uma das técnicas de modelos simples que obteve melhores resultados, aprimorando seu desempenho quando utilizada como estimador final para conjuntos de modelos. Além disso, dada à sua característica de fornecer de forma nativa um intervalo de confiança para cada uma de suas previsões, este modelo é o mais indicado para a representação virtual dos sistemas que compõem a usina termelétrica de ciclo combinado.

Esses resultados reforçam a importância da utilização de técnicas avançadas de ML na construção de gêmeos digitais voltados para a otimização da eficiência energética e redução de emissões. Dessa forma, este trabalho contribui para o avanço da aplicação de

ML em sistemas gêmeos digitais, oferecendo subsídios para o desenvolvimento de soluções mais eficientes e sustentáveis no setor energético.

5.1 TRABALHOS FUTUROS

Apesar dos avanços alcançados, algumas limitações foram identificadas, como a dependência da qualidade e abrangência dos dados disponíveis, além da necessidade de maior interpretabilidade dos modelos para aplicação em ambientes industriais. Esses aspectos abrem espaço para futuras pesquisas voltadas à integração de modelos híbridos, que combinem aprendizado de máquina com modelos físicos, bem como à aplicação de técnicas de explicabilidade em inteligência artificial.

Entre as perspectivas futuras, destaca-se a validação dos modelos desenvolvidos em dados reais de operação de usinas, a exploração de arquiteturas mais sofisticadas de redes neurais e a investigação de métodos de otimização adaptativa para diferentes cenários de carga e condições ambientais. Outra linha promissora consiste na integração dos modelos ao Sistema Gêmeo Digital e, posteriormente, ao supervisório da planta, possibilitando que o próprio SGD auxilie na operação do sistema em tempo real. Outras possibilidades são o desenvolvimento de um módulo otimizador para o SGD, focado em maximizar a eficiência da usina; e um estudo de estabilidade dos modelos, focado na atualização dos modelos e na janela ideal de dados para manutenção da exatidão dos modelos, evitanto *overfitting*.

Adicionalmente, recomenda-se uma análise mais aprofundada do modelo de média dinâmica, avaliando seu desempenho em diferentes bases de dados e conjuntos de modelos base, de modo a verificar sua robustez e generalização em contextos variados.

REFERÊNCIAS

ABADI, M. et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. Software available from tensorflow.org. Disponível em: <https://www.tensorflow.org/>. Citado na página 28.

ALAM, M. S. et al. Ensemble machine-learning models for accurate prediction of solar irradiation in bangladesh. *Processes*, v. 11, n. 3, 2023. ISSN 2227-9717. Disponível em: <https://www.mdpi.com/2227-9717/11/3/908>. Citado na página 24.

ANG, Y. et al. A novel ensemble machine learning and time series approach for oil palm yield prediction using landsat time series imagery based on ndvi. *Geocarto International*, Taylor & Francis, v. 37, n. 25, p. 9865–9896, 2022. Disponível em: <https://doi.org/10.1080/10106049.2022.2025920>. Citado na página 24.

ANJOS, L. G. *Armazenamento de Energia: Uma Solução para o Futuro Sustentável*. 2023. <https://welcome.atlasgov.com/blog/esg/armazenamento-de-energia/>. [Accessed 06-11-2024]. Citado na página 11.

ATANASYAN, A. et al. The fischertwin: An experimentable digital twin case study. In: *Proceedings of the 12th International Conference on Simulation and Modeling Methodologies, Technologies and Applications*. SCITEPRESS - Science and Technology Publications, 2022. p. 280–287. Disponível em: <http://dx.doi.org/10.5220/0011277800003274>. Citado na página 15.

BASHIR, R. N. et al. Smart reference evapotranspiration using internet of things and hybrid ensemble machine learning approach. *Internet of Things*, v. 24, p. 100962, 2023. ISSN 2542-6605. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2542660523002858>. Citado na página 24.

BILLER, B.; BILLER, S. Implementing digital twins that learn: Ai and simulation are at the core. *Machines*, v. 11, n. 4, 2023. ISSN 2075-1702. Disponível em: <https://www.mdpi.com/2075-1702/11/4/425>. Citado na página 17.

BRASIL. Ministério de Minas e Energia. Empresa de Pesquisa Energética. *Plano Nacional de Energia 2050*. Brasília, DF, 2020. Citado na página 11.

BREUNIG, M. M. et al. Lof: identifying density-based local outliers. *ACM SIGMOD Record*, Association for Computing Machinery (ACM), v. 29, n. 2, p. 93–104, maio 2000. ISSN 0163-5808. Disponível em: <http://dx.doi.org/10.1145/335191.335388>. Citado na página 34.

CAMPOS JR., G. *Uso de térmicas deve seguir em alta para assegurar sistema elétrico*. 2024. <https://www.poder360.com.br/poder-energia/uso-de-termicas-deve-seguir-em-alta-para-assegurar-sistema-eletrico/>. [Accessed 06-11-2024]. Citado na página 11.

CHAO, Z. et al. Artificial intelligence algorithms for predicting peak shear strength of clayey soil-geomembrane interfaces and experimental validation. *Geotextiles and*

Geomembranes, v. 51, n. 1, p. 179–198, 2023. ISSN 0266-1144. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0266114422001169>. Citado na página 24.

CHAURASIA, V.; PANDEY, M. K.; PAL, S. Chronic kidney disease: a prediction and comparison of ensemble and basic classifiers performance. *Hum.-Intell. Syst. Integr.*, Springer Science and Business Media LLC, v. 4, n. 1-2, p. 1–10, jun. 2022. Citado na página 18.

COSTELLO, K.; RIMOL, M. *Gartner Identifies the Top Strategic Technology Trends for 2022*. 2021. <https://www.gartner.com/en/newsroom/press-releases/2021-10-18-gartner-identifies-the-top-strategic-technology-trends-for-2022>. [Accessed 14-11-2024]. Citado na página 12.

DELOUILLE, V. et al. Chapter 15 - coronal holes detection using supervised classification. In: CAMPOREALE, E.; WING, S.; JOHNSON, J. R. (Ed.). *Machine Learning Techniques for Space Weather*. Elsevier, 2018. p. 365–395. ISBN 978-0-12-811788-0. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128117880000159>. Citado na página 35.

DOROGUSH, A. V.; ERSHOV, V.; GULIN, A. *CatBoost: gradient boosting with categorical features support*. arXiv, 2018. Disponível em: <https://arxiv.org/abs/1810.11363>. Citado na página 28.

ELDER, J. Chapter 16 - the apparent paradox of complexity in ensemble modeling*. In: NISBET, R.; MINER, G.; YALE, K. (Ed.). *Handbook of Statistical Analysis and Data Mining Applications (Second Edition)*. Second edition. Boston: Academic Press, 2018. p. 705–718. ISBN 978-0-12-416632-5. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780124166325000165>. Citado na página 17.

EMMERT-STREIB, F. What is the role of ai for digital twins? *AI*, v. 4, n. 3, p. 721–728, 2023. ISSN 2673-2688. Disponível em: <https://www.mdpi.com/2673-2688/4/3/38>. Citado na página 17.

FLEMING, S. W. et al. Assessing the new natural resources conservation service water supply forecast model for the american west: A challenging test of explainable, automated, ensemble artificial intelligence. *Journal of Hydrology*, v. 602, p. 126782, 2021. ISSN 0022-1694. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0022169421008325>. Citado na página 24.

G1. *ONS aciona usinas térmicas após aumento na demanda de energia causada pelo calor; especialista alerta sobre gastos e poluição*. [S.l.]: Jornal da Globo, 2023. <https://g1.globo.com/jornal-da-globo/noticia/2023/11/17/onda-de-calor-aumento-na-demanda-por-energia-eletrica-faz-ons-acionar-usinas-termicas.ghtml>. [Accessed 06-11-2024]. Citado na página 11.

GOLDSTEIN, M.; UCHIDA, S. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLOS ONE*, Public Library of Science (PLoS), v. 11, n. 4, p. e0152173, abr. 2016. ISSN 1932-6203. Disponível em: <http://dx.doi.org/10.1371/journal.pone.0152173>. Citado na página 34.

GRIEVES, M. *Digital Twin: Manufacturing Excellence through Virtual Factory Replication*. [S.l.], 2015. Citado na página 14.

GUERCIO, A.; BINI, R. Biomass-fired organic rankine cycle combined heat and power systems. In: *Organic Rankine Cycle (ORC) Power Systems*. [S.l.]: Elsevier, 2017. p. 527–567. Citado na página 43.

HASAN, K. M. Z.; HASAN, M. Z. Performance evaluation of ensemble-based machine learning techniques for prediction of chronic kidney disease. In: *Emerging Research in Computing, Information, Communication and Applications*. Singapore: Springer Singapore, 2019, (Advances in intelligent systems and computing). p. 415–426. Citado na página 20.

HUANG, Z. et al. Digital twin driven life-cycle operation optimization for combined cooling heating and power-cold energy recovery (cchp-cer) system. *Applied Energy*, Elsevier BV, v. 324, out. 2022. ISSN 0306-2619. Disponível em: <http://dx.doi.org/10.1016/j.apenergy.2022.119774>. Citado na página 16.

IPCC. *Section 3 - Long-Term Climate and Development Futures. In: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Geneva, Switzerland: IPCC, 2023. 35–115 p. Citado 2 vezes nas páginas 11 e 12.

IYENGAR, G.; LAM, H.; WANG, T. *Is Cross-Validation the Gold Standard to Evaluate Model Performance?* 2024. Disponível em: <https://arxiv.org/abs/2407.02754>. Citado na página 27.

JUNIOR, C. A. Alves de A. et al. Digital twins of the water cooling system in a power plant based on fuzzy logic. *Sensors*, MDPI AG, v. 21, n. 20, out. 2021. ISSN 1424-8220. Disponível em: <http://dx.doi.org/10.3390/s21206737>. Citado na página 16.

KHYANI, D. et al. An interpretation of stacking and blending approach in machine learning. *International Research Journal of Engineering and Technology (IRJET)*, v. 08, n. 07, p. 3117–3120, jul. 2021. Citado na página 22.

KING, A. P.; ECKERSLEY, R. J. Chapter 2 - descriptive statistics ii: Bivariate and multivariate statistics. In: KING, A. P.; ECKERSLEY, R. J. (Ed.). *Statistics for Biomedical Engineers and Scientists*. Academic Press, 2019. p. 23–56. ISBN 978-0-08-102939-8. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780081029398000116>. Citado na página 32.

KOTU, V.; DESHPANDE, B. Chapter 2 - data science process. In: KOTU, V.; DESHPANDE, B. (Ed.). *Data Science (Second Edition)*. Second edition. Morgan Kaufmann, 2019. p. 19–37. ISBN 978-0-12-814761-0. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128147610000022>. Citado na página 17.

KREUZER, T.; PAPAPETROU, P.; ZDRAVKOVIC, J. Artificial intelligence in digital twins—a systematic literature review. *Data & Knowledge Engineering*, v. 151, p. 102304, 2024. ISSN 0169-023X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0169023X24000284>. Citado 2 vezes nas páginas 14 e 16.

KRITZINGER, W. et al. Digital twin in manufacturing: A categorical literature review and classification. *IFAC-PapersOnLine*, v. 51, n. 11, p. 1016–1022, 2018. ISSN 2405-8963. 16th IFAC Symposium on Information Control Problems in Manufacturing INCOM 2018. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2405896318316021>. Citado 2 vezes nas páginas 15 e 16.

LATHI, B. Sinais e sistemas lineares. In: _____. *Sinais e sistemas*. 2. ed. Porto Alegre, RS: Bookman, 2007. cap. 1. Citado na página 31.

LIM, J. et al. The epri gas turbine digital twin – a platform for operator focused integrated diagnostics and performance forecasting. In: *Volume 4: Controls, Diagnostics, and Instrumentation; Cycle Innovations; Cycle Innovations: Energy Storage; Education; Electric Power*. American Society of Mechanical Engineers, 2021. (GT2021). Disponível em: <http://dx.doi.org/10.1115/GT2021-59572>. Citado na página 15.

LIN, E.; LIN, C.-H.; LANE, H.-Y. Prediction of functional outcomes of schizophrenia with genetic biomarkers using a bagging ensemble machine learning method with feature selection. *Scientific Reports*, v. 11, n. 1, p. 10179, May 2021. ISSN 2045-2322. Disponível em: <https://doi.org/10.1038/s41598-021-89540-6>. Citado na página 24.

LIU, M. et al. Review of digital twin about concepts, technologies, and industrial applications. *Journal of Manufacturing Systems*, Elsevier BV, v. 58, p. 346–361, jan. 2021. ISSN 0278-6125. Disponível em: <http://dx.doi.org/10.1016/j.jmsy.2020.06.017>. Citado na página 14.

LYANTSEV, O.; KAZANTSEV, A.; ABDULNAGIMOV, A. Identification method for nonlinear dynamic models of gas turbine engines on acceleration mode. *Procedia Engineering*, Elsevier BV, v. 176, p. 409–415, 2017. ISSN 1877-7058. Disponível em: <http://dx.doi.org/10.1016/j.proeng.2017.02.339>. Citado na página 31.

MAHAJAN, P. et al. A comparative evaluation of machine learning ensemble approaches for disease prediction using multiple datasets. *Health Technol. (Berl.)*, Springer Science and Business Media LLC, v. 14, n. 3, p. 597–613, maio 2024. Citado na página 19.

MARMOLEJO-SAUCEDO, J. A. Design and development of digital twins: a case study in supply chains. *Mobile Networks and Applications*, Springer Science and Business Media LLC, v. 25, n. 6, p. 2141–2160, jun. 2020. ISSN 1572-8153. Disponível em: <http://dx.doi.org/10.1007/s11036-020-01557-9>. Citado na página 15.

MATITOPANUM, S. et al. Prediction of the ultimate tensile strength (uts) of asymmetric friction stir welding using ensemble machine learning methods. *Processes*, v. 11, n. 2, 2023. ISSN 2227-9717. Disponível em: <https://www.mdpi.com/2227-9717/11/2/391>. Citado na página 24.

MCCARTNEY, A. *Gartner Top 10 Strategic Technology Trends for 2024*. 2023. <https://www.gartner.com/en/articles/gartner-top-10-strategic-technology-trends-for-2024>. [Accessed 14-11-2024]. Citado na página 12.

MENDI, A. F. A digital twin case study on automotive production line. *Sensors*, MDPI AG, v. 22, n. 18, set. 2022. ISSN 1424-8220. Disponível em: <http://dx.doi.org/10.3390/s22186963>. Citado na página 15.

NOGUEIRA, F. *Bayesian Optimization: Open source constrained global optimization tool for Python*. 2014–. Disponível em: <https://github.com/bayesian-optimization/BayesianOptimization>. Citado 3 vezes nas páginas 28, 35 e 46.

OLIVEIRA, F. J. A. de. *O planejamento da operação energética no sistema : conceitos, modelagem matemática, previsão de geração e carga*. São Paulo: Artliber, 2020. ISBN 978-65-86443-02-8. Citado na página 11.

OPPENHEIM., A.; WILLSKY, A. Sinais e sistemas. In: _____. *Sinais e sistemas*. 2. ed. [S.l.]: Pearson Prentice Hall, 2010. cap. 1. Citado na página 31.

PANETTA, K. *Gartner Top 10 Strategic Technology Trends For 2020*. 2019. <https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2020>. [Accessed 14-11-2024]. Citado na página 12.

PANETTA, K. *Gartner Top Strategic Technology Trends for 2021*. 2020. <https://www.gartner.com/smarterwithgartner/gartner-top-strategic-technology-trends-for-2021>. [Accessed 14-11-2024]. Citado na página 12.

PANOV, V.; CRUZ-MANZO, S. Gas turbine performance digital twin for real-time embedded systems. In: *Volume 5: Controls, Diagnostics, and Instrumentation; Cycle Innovations; Cycle Innovations: Energy Storage*. American Society of Mechanical Engineers, 2020. (GT2020). Disponível em: <http://dx.doi.org/10.1115/GT2020-14664>. Citado na página 15.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado 2 vezes nas páginas 28 e 34.

PERNO, M.; HVAM, L.; HAUG, A. A machine learning digital twin approach for critical process parameter prediction in a catalyst manufacturing line. *Computers in Industry*, v. 151, p. 103987, 2023. ISSN 0166-3615. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0166361523001379>. Citado na página 17.

RADOčAJ, D.; JURIšIć, M.; TADIć, V. The effect of bioclimatic covariates on ensemble machine learning prediction of total soil carbon in the pannonian biogeoregion. *Agronomy*, v. 13, n. 10, 2023. ISSN 2073-4395. Disponível em: <https://www.mdpi.com/2073-4395/13/10/2516>. Citado na página 24.

RIMOL, M.; HOWLEY, C. *Gartner Identifies the Top 10 Strategic Technology Trends for 2023*. 2022. <https://www.gartner.com/en/newsroom/press-releases/2022-10-17-gartner-identifies-the-top-10-strategic-technology-trends-for-2023>. [Accessed 14-11-2024]. Citado na página 12.

SARKER, I. H. Machine learning: Algorithms, real-world applications and research directions. *SN Comput. Sci.*, Springer Science and Business Media LLC, v. 2, n. 3, p. 160, mar. 2021. Citado na página 17.

SARMAS, E. et al. Baseline energy modeling for improved measurement and verification through the use of ensemble artificial intelligence models. *Information Sciences*, v. 654, p. 119879, 2024. ISSN 0020-0255. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0020025523014640>. Citado na página 24.

SARRESHTEDAR, A. et al. Investigation of quantitative and qualitative changes in groundwater of ardebil plain using ensemble artificial intelligence-based modeling. *Water Supply*, IWA Publishing, v. 22, n. 9, p. 7140–7157, jul. 2022. ISSN 1607-0798. Disponível em: <http://dx.doi.org/10.2166/ws.2022.273>. Citado na página 24.

SHARGHI, E.; NOURANI, V.; BEHFAR, N. Earthfill dam seepage analysis using ensemble artificial intelligence based modeling. *Journal of Hydroinformatics*, v. 20, p. 1071–1084, 2018. Disponível em: <https://api.semanticscholar.org/CorpusID:64487583>. Citado na página 24.

SHARMA, A. et al. Digital twins: State of the art theory and practice, challenges, and open research questions. *Journal of Industrial Information Integration*, Elsevier BV, v. 30, nov. 2022. ISSN 2452-414X. Disponível em: <http://dx.doi.org/10.1016/j.jii.2022.100383>. Citado 2 vezes nas páginas 12 e 16.

SHEN, Z.; ARRAÑO-VARGAS, F.; KONSTANTINOU, G. Artificial intelligence and digital twins in power systems: Trends, synergies and opportunities. *Digital Twin*, F1000 Research Ltd, v. 2, p. 11, mar. 2023. ISSN 2752-5783. Disponível em: <http://dx.doi.org/10.12688/digitaltwin.17632.2>. Citado na página 16.

TEIXEIRA, C. A. N. et al. Gás natural – um combustÍvel-chave para uma economia de baixo carbono. *Petróleo & Gás*, BNDES, v. 27, n. 53, p. 131–175, 2021. Citado na página 12.

TERVEN, J. et al. A comprehensive survey of loss functions and metrics in deep learning. *Artif. Intell. Rev.*, Springer Science and Business Media LLC, v. 58, n. 7, abr. 2025. Citado 2 vezes nas páginas 25 e 26.

ÜNVER, Ü. et al. Design of an inlet air-cooling system for a gas turbine power plant. In: *Green Energy and Technology*. Cham: Springer International Publishing, 2018, (Green Energy and Technology). p. 1089–1100. Citado na página 43.

VERDOUW, C. et al. Digital twins in smart farming. *Agricultural Systems*, Elsevier BV, v. 189, abr. 2021. ISSN 0308-521X. Disponível em: <http://dx.doi.org/10.1016/j.agsy.2020.103046>. Citado na página 15.

VIRTANEN, P. et al. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, v. 17, p. 261–272, 2020. Citado na página 28.

WRIGHT, L.; DAVIDSON, S. How to tell the difference between a model and a digital twin. *Advanced Modeling and Simulation in Engineering Sciences*, v. 7, n. 1, p. 13, Mar 2020. ISSN 2213-7467. Disponível em: <https://doi.org/10.1186/s40323-020-00147-4>. Citado 2 vezes nas páginas 14 e 16.

ZHENG, Y.; YANG, S.; CHENG, H. An application framework of digital twin and its case study. *Journal of Ambient Intelligence and Humanized Computing*, Springer Science and Business Media LLC, v. 10, n. 3, p. 1141–1153, jun. 2018. ISSN 1868-5145. Disponível em: <http://dx.doi.org/10.1007/s12652-018-0911-3>. Citado na página 15.