UNIVERSIDADE FEDERAL DE SANTA CATARINA
DEPARTAMENTO DE ENGENHARIA QUÍMICA E ENGENHARIA DE ALIMENTOS
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA QUÍMICA

BRUNO DEON

GÊMEOS DIGITAIS BASEADOS EM APRENDIZADO DE MÁQUINA PARA APOIO À TOMADA DE DECISÃO OPERACIONAL EM UMA USINA TERMELÉTRICA À COMBUSTÃO INTERNA

FLORIANÓPOLIS - SC
2023

Bruno Deon

GÊMEOS DIGITAIS BASEADOS EM APRENDIZADO DE MÁQUINA PARA APOIO À TOMADA DE DECISÃO OPERACIONAL EM UMA USINA TERMELÉTRICA À COMBUSTÃO INTERNA

Dissertação submetida ao Programa de Pós-Graduação em Engenharia Química da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de Mestre em Engenharia Química.

Orientador: Prof. Dr. Natan Padoin
Coorientadora: Prof.ª Dr.ª Cíntia Soares

Florianópolis - SC
2023

Deon, Bruno
Gêmeos Digitais Baseados em Aprendizado de Máquina para Apoio à Tomada de Decisão Operacional em uma Usina Termelétrica à Combustão Interna / Bruno Deon ; orientador, Natan Padoin, coorientadora, Cíntia Soares, 2023.
113 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Centro Tecnológico, Programa de Pós-Graduação em Engenharia Química, Florianópolis, 2023.

Inclui referências.

1. Engenharia Química. 2. Aprendizado de Máquina. 3. Gêmeos Digitais. 4. Manutenção Preditiva. 5. Usina Termelétrica. I. Padoin, Natan . II. Soares, Cíntia. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia Química. IV. Título.

Bruno Deon

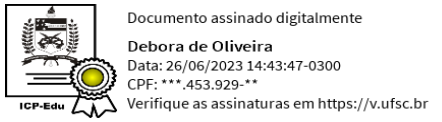
GÊMEOS DIGITAIS BASEADOS EM APRENDIZADO DE MÁQUINA PARA APOIO À TOMADA DE DECISÃO OPERACIONAL EM UMA USINA TERMELÉTRICA À COMBUSTÃO INTERNA

O presente trabalho em nível de Mestrado foi avaliado e aprovado, em 31 de março de 2023, pela banca examinadora composta pelos seguintes membros:

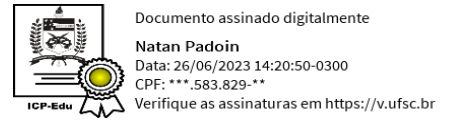
Prof. Maurício de Souza Bezerra, Dr.
Universidade Federal do Rio de Janeiro

Prof. Nicolas Spogis, Dr.
Pontifícia Universidade Católica de Campinas

Certificamos que esta é a versão original e final do trabalho de conclusão que foi julgado adequado para obtenção do título de Mestre em Engenharia Química.



Coordenação do Programa de Pós-Graduação



Prof. Dr. Natan Padoin
Orientador

Florianópolis - SC, 2023

Dedico este trabalho a todos os meus professores,
sendo, meus pais, os maiores entre eles.

AGRADECIMENTOS

Aos meus pais Laércio José Deon e Rosa Maria Magnabosco Deon pela minha criação, educação e amor incondicional, sendo os maiores dos exemplos para mim. Aos demais familiares, pelo incentivo constante.

À Universidade Federal de Santa Catarina, ao Programa de Pós-Graduação em Engenharia Química e ao Laboratório de Materiais e Computação Científica, pela oportunidade e suporte.

Ao meu orientador Prof. Dr. Natan Padoin por todo auxílio, confiança e ensinamentos que abrangeram muito mais do que apenas a minha vida acadêmica.

À minha coorientadora Prof.ª Dr.ª Cíntia Soares por ter contribuído, com paciência invejável, ao trabalho e ao meu crescimento.

A todos os professores do Programa de Pós-graduação em Engenharia Química da UFSC pelos seus ensinamentos e experiências repassadas.

Aos membros da banca de defesa pela disponibilidade em avaliar e contribuir com esta dissertação.

Ao CNPq e CAPES por todo o suporte.

À Radix pela oportunidade, estrutura e suporte técnico, em especial aos meus colegas e amigos Flavio, Kleyton, Robson, Camilla, Freitas e Justino por todo trabalho, conhecimento e risadas que compartilhamos nesta caminhada.

À EPASA pelo suporte financeiro, especialmente ao Carlos e ao Rodrigo que sempre foram muito solícitos.

A todos os amigos e colegas do LabMAC, em especial à Ana Paula, Jessica e Gabi por me guiarem no início dessa caminhada, apesar da distância e do tempo tenho enorme carinho e respeito por vocês.

Ao meu irmão Matheus Magnabosco Deon por todas as curtas conversas que, duravam horas. Aos meus amigos por jamais falharem em alegrar meus dias. À minha companheira Stephanie por tornar leve quaisquer fardos que carreguei.

A todos que, de uma forma ou outra, contribuíram para a elaboração deste trabalho.

A mente que se abre a uma nova ideia
jamais voltará ao seu tamanho original.
(Albert Einstein)

RESUMO

No Brasil, Usinas Termelétricas (UTEs) possuem um caráter estratégico e emergencial devido à flexibilidade de operação e por não dependerem de condições climáticas, de forma que a disponibilidade e a confiabilidade de uma UTE são questões críticas. O desafio de garantir estas características está intrinsicamente associado aos processos de manutenção. Através do programa de Pesquisa e Desenvolvimento da Agência Nacional de Energia Elétrica, a Radix - Engenharia e Desenvolvimento de Software e a Universidade Federal de Santa Catarina, atuando em parceria, desenvolveram um sistema de detecção de anomalias para auxílio à tomada de decisão em manutenções preditivas com abordagem de Gêmeos Digitais para as Unidades Geradoras Diesel (UGDs) das Centrais Elétricas da Paraíba - Epasa. Associando dados de sensores historiados a notas de manutenção, foi possível treinar modelos regressivos nas melhores condições de operação criando, assim, Gêmeos Digitais que mimetizam as principais variáveis de quatro subsistemas integrantes das UGDs: o sistema de admissão de combustível, arrefecimento, lubrificação e ar de admissão e gases de exaustão (sendo os dois últimos combinados, uma vez que estão correlacionados pelo turbocompressor). Para análise e validação destes modelos, as UGDs foram treinadas e testadas em grupos, avaliando o coeficiente de determinação (R^2) e a raiz quadrada do erro quadrático médio (RMSE) em cada caso. Ao normalizar os erros (entre as predições e as variáveis alvo) e submetê-los a uma média móvel, foi possível criar um índice de saúde que reflete o desempenho de cada subsistema e, assim, avaliar a ocorrência de desvios anômalos. Ao correlacionar tais desvios com as notas de manutenção, a partir de janelas temporais, foi possível dar aos modelos regressivos a capacidade de classificação. Dessa forma, métricas como a Acurácia e o *F1-Score* também puderam ser utilizadas. Um algoritmo genético foi utilizado para otimização dos modelos, selecionando as entradas, algoritmos de regressão e hiperparâmetros que maximizassem as métricas de regressão e de classificação, simultaneamente. Dessa forma, foram obtidos como resultados médios finais R^2 de 0,93 e RMSE de 0,072 para as regressões, enquanto para as métricas de classificação uma Acurácia média de 0,86 foi observada e um F1-Score de 0,52 foi alcançado na detecção de anomalias. Por fim, os modelos foram implementados em um sistema SCADA de supervisão com telas personalizadas para acompanhamento em tempo real das operações. A metodologia desenvolvida pode ser replicada para indústrias em geral que utilizam um sistema semelhante de aquisição de dados.

Palavras-chave: Aprendizado de Máquina; Gêmeos Digitais; Manutenção Preditiva; Usina Termelétrica.

ABSTRACT

In Brazil, Thermoelectric Power Plants (TPPs) have a strategic and emergency character due to the flexibility of operation and for not depending on weather conditions, so the TPPs availability and reliability are critical issues. The challenge of ensuring these characteristics is intrinsically associated with maintenance processes. Through the Research and Development program of the National Electric Energy Agency, Radix - Engineering and Software, in partnership with the Federal University of Santa Catarina, developed an anomaly detection system to aid decision-making in predictive maintenance with a Digital Twin approach for the Electric Power Plants of Paraíba (Epasa) Diesel Generating Units (DGUs). By associating historical sensor data with maintenance notes, it was possible to train regressive models in the best operating conditions to create digital twins that mimic the main variables of four subsystems that are part of the DGUs: the fuel admission system, cooling, lubrication, combustion air admission and gases exhaustion (the latter two were combined, since they are correlated by the turbocharger). For analysis and validation, the five DGUs data were combined so it could be trained and tested in groups, evaluating the coefficient of determination (R^2) and the root mean square error (RMSE) of the models. A health index was created by normalizing the errors (between the predictions and the target variables) and submitting them to a moving average, thus reflecting the performance of each subsystem and being able to evaluate the occurrence of anomalous deviations. By correlating such deviations with the maintenance notes, using time windows, it was possible to give the regressive models the ability to classify. Thus, metrics such as accuracy and F1-Score could also be used. A genetic algorithm was used to optimize the models, selecting inputs, regression algorithms and hyperparameters that maximized the regression and classification metrics simultaneously. Resulting in a final average R^2 of 0.93 and RMSE of 0.072 for the regressions, and for the classification metrics an average Accuracy of 0.86 and *F1-Score* of 0.52 for the anomalies detection. Finally, the models were implemented in a SCADA supervisory system with customized dashboards for real-time monitoring. That the methodology can be replicated for industries in general that use a similar data acquisition system.

Keywords: Machine Learning; Digital Twin; Predictive Maintenance; Thermoelectric Power Plants.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANEEL	Agência Nacional de Energia Elétrica
DT	*Digital Twin*
EPASA	Centrais Elétricas da Paraíba
GB	*Gradient Boosting*
HFO	*Heavy Fuel Oil*
IoT	*Internet of Things*
MAE	*Mean Absolute Error*
ML	*Machine Learning*
MLP	*Multilayer Perceptrons*
ONS	Operador Nacional do Sistema Elétrico
P&D	Pesquisa e Desenvolvimento
RF	*Random Forest*
RMSE	*Root Mean Squared Error*
RUL	*Remaining Useful Life*
SAA	Sistema de Água de Arrefecimento
SAC	Sistema de Admissão de Combustível
SAT	Sistema de Arrefecimento de Baixa Temperatura
SBT	Sistema de Arrefecimento de Baixa Temperatura
SIN	Sistema Interligado Nacional
SOL	Sistema de Admissão de Ar e Exaustão
SVM	*Support Vector Machines*
TNE	Termonordeste
TPB	Termoparaíba
UFSC	Universidade Federal de Santa-Catarina
UGD	Unidade Geradora Diesel
UTE	Usina Termelétrica

SUMÁRIO

1 INTRODUÇÃO

No Brasil, a produção e transmissão de energia elétrica é dada por um sistema interligado de grande porte, chamado de Sistema Interligado Nacional (SIN). A coordenação e controle deste sistema se dá pelo Operador Nacional do Sistema Elétrico (ONS) e sua fiscalização e regulamentação está sob competência da ANEEL - Agência Nacional de Energia Elétrica (Sistema Interligado Nacional, s. d.).

O SIN é composto predominantemente pela geração hidrelétrica, e, ultimamente, a geração eólica, assim como a fotovoltaica, tem ganhado espaço com forte crescimento, o que torna a matriz elétrica brasileira uma das mais limpas do mundo (EPE, 2021). Contudo é necessário salientar a importância das Usinas Termelétricas (UTEs). Por estas não dependerem de questões climáticas (precipitação atmosférica, incidência solar e movimento de massas de ar), possuem uma flexibilidade muito maior de geração, apresentando, assim, um caráter mais emergencial e de funções estratégicas, o que permite um controle mais seguro da rede, além da gestão de reservatórios de água das usinas hidrelétricas (O SISTEMA INTERLIGADO NACIONAL, s.d.).

Devido a isto, a disponibilidade, a confiabilidade e o desempenho das máquinas geradoras de uma UTE são questões críticas para maximizar os resultados econômicos e para garantir o atendimento à demanda do setor elétrico. Portanto, o desafio de garantir estas características está associado tanto aos processos de operação quanto aos de manutenção.

Por outro lado, as consequências nocivas para a saúde humana, assim como o impacto ambiental das emissões poluentes, são amplamente conhecidas (KAN, CHEN e TONG, 2012; FAJERSZTAJN, VERAS, *et al.*, 2013; DRISCOLL, BUONOCORE, *et al.*, 2015; MANISALIDIS, STAVROPOULOU, *et al.*, 2020). Além disso, o custo inflacionado dos combustíveis é outro fator que deve ser considerado. Desta forma, apesar dos combustíveis fósseis ainda serem fundamentais como complemento às demais fontes de energia elétrica, garantir o máximo desempenho e um bom funcionamento do maquinário, e consequentemente reduzir os impactos citados, deve ser o cenário almejado.

Com o crescente aumento de conectividade e o uso de sistemas inteligentes, tornou-se possível a previsão de tendências, detecção de padrões de comportamento e correlações por meio de estatísticas ou modelos de aprendizado de máquina

capazes de antecipar falhas, melhorando o processo de tomada de decisão para atividade de manutenção (ZONTA, DA COSTA, *et al.*, 2020). Desta maneira, metodologias em ascensão como as de manutenções de caráter preditivo se consolidaram, as quais não visam apenas a antecipação de falhas, mas também uma operação eficiente, melhora da segurança, qualidade do produto, confiabilidade, disponibilidade e redução de custos (SELCUK, 2017; ZONTA, DA COSTA, *et al.*, 2020).

Vale ainda citar a recente utilização de sistemas que mimetizam o comportamento dos equipamentos, denominadas de Gêmeos Digitais (*Digital Twins* - DT), cuja aplicação mais popular é justamente atrelada a prognósticos e gestão da saúde de equipamentos, o que nada mais são do que formas de manutenções preditivas (LIU, FANG, *et al.*, 2021; TAO, ZHANG, *et al.*, 2018)

Através do Programa de Pesquisa e Desenvolvimento (P&D), a ANEEL busca estimular o desenvolvimento tecnológico das empresas de energia com projetos que demonstrem originalidade, relevância e viabilidade. Por lei, estas empresas devem aplicar um percentual de sua Receita Operacional Líquida no Programa de P&D. É de competência da ANEEL a administração e a alocação destes recursos, incentivando a inovação e o aprimoramento de produtos ou processos do setor, contribuindo, dessa maneira, para a segurança e a confiabilidade do fornecimento de energia, assim como a redução do impacto ambiental do setor (Programa de Pesquisa e Desenvolvimento Tecnológico, s.d.).

Localizada em João Pessoa, no Estado da Paraíba, a EPASA - Centrais Elétricas da Paraíba S.A é uma empresa Produtora Independente de Energia. Composta pela Termonordeste (TNE) e Termoparaíba (TPB), essas duas UTEs possuem capacidade de geração instalada de 171 MW cada, e combinadas resultam na segunda maior planta termoelétrica da categoria do país (Epasa - Geração de Energia, s.d.).

Desta forma, utilizando dos recursos disponíveis no Programa de P&D da ANEEL, por meio de contrato com a Radix - Engenharia e Software e em parceria com a Universidade Federal de Santa-Catarina (UFSC), a EPASA busca maximizar a confiabilidade e disponibilidade de suas plantas termelétricas utilizando tecnologia de inteligência artificial com abordagem de Gêmeos Digitais. Apesar de possuir uma grande quantidade de sensores, a EPASA conta apenas com alarmes de limites e

manutenções preventivas para garantir o bom funcionamento e a confiabilidade de seus equipamentos.

Neste contexto, este trabalho visa a elaboração de uma ferramenta de auxílio à tomada de decisão operacional relativa a manutenções de caráter preditivo, integrada ao Sistema de Supervisão e Aquisição de Dados (SCADA) e com execução em tempo real. Além do mais, é importante destacar que os procedimentos desenvolvidos são aplicáveis nos mais diversos setores industriais que utilizem sistemas similares de supervisório, incluindo demais indústrias geradoras de energia, além de indústrias de processos químicos.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

O objetivo geral deste trabalho é a maximização da confiabilidade e da disponibilidade das UTEs da EPASA através da utilização da tecnologia de inteligência artificial com abordagem de gêmeos digitais, ao desenvolver um sistema de apoio à tomada de decisão via predição de tendências e desvios operacionais.

1.1.2 Objetivos Específicos

Para atender o objetivo geral apresentado na Seção 1.1.1, esta dissertação pauta-se nos seguintes objetivos específicos:

- analisar as condições operacionais da planta, a partir dos diagramas de processo e dos dados historiados (das variáveis de processo e notas de manutenção);
- selecionar as principais variáveis de processo e diferentes condições operacionais, a partir da análise dos dados historiados;
- construir modelos regressivos que mimetizem as principais variáveis de processo nas condições ideais de operação, em função das demais variáveis selecionadas, e propor um índice que reflita o desempenho dos equipamentos (índice de saúde);

- construir modelos que classifiquem as diferentes condições operacionais, e avaliar a utilização do índice de saúde como ferramenta de suporte a esses modelos;
- validar e otimizar os modelos de regressão e classificação, nas condições estabelecidas, e implementar no supervisório da planta aqueles que melhor performarem.

2 REVISÃO BIBLIOGRÁFICA

Os custos com manutenção podem representar a parte majoritária dos gastos industriais. Dependendo da indústria, esses variam entre 15% e 60% dos custos dos bens produzidos. Além disso, cerca de um terço de todo custo das manutenções é desperdiçado com intervenções desnecessárias ou indevidas (MOBLEY, 2002). Portanto, é indiscutível a necessidade de metodologias de manutenções mais eficientes.

Sobre as diferentes metodologias de manutenção, estas podem ser separadas em três principais categorias, de acordo com o seu nível de complexidade e eficácia. As mais simples políticas de manutenções e, consequentemente, as menos eficientes, são as de natureza corretiva. Neste tipo de manutenção, os equipamentos são levados para reparos apenas quando as intervenções são inevitáveis, o que pode acarretar, em muitas vezes, paradas de máquina inesperadas. Por este motivo, tais intervenções tendem a ser mais custosas, pois além de perdas na produção, estas estão associadas a falhas críticas (SUSTO, SCHIRRU, *et al.*, 2014).

Uma forma de contornar tais problemas é por meio de políticas preventivas de manutenções, uma vez que essas buscam evitar que falhas críticas ocorram. Como o seu próprio nome já diz, previnem. Para tanto, as manutenções são programadas de acordo com um cronograma. Consequentemente, este tipo de manutenção é mais eficaz do que as manutenções corretivas. Entretanto, intervenções desnecessárias acabam sendo realizadas, levando a um uso ineficiente dos recursos e do tempo (SUSTO, SCHIRRU, *et al.*, 2014).

Finalmente, existem manutenções de caráter preditivo, as quais buscam antever as falhas de um sistema, tornando possível otimizar as atividades de manutenção. Por meio do diagnóstico do estado do sistema, é possível detectar os primeiros sinais de falha e, então, planejar as intervenções no momento mais oportuno, evitando tanto falhas críticas e paradas de máquinas, quanto o desperdício de recursos decorrentes de manutenções corretivas e preventivas, respectivamente (SELCUK, 2017).

O conceito de manutenção preditiva é relativamente antigo, datando da década de 40. Em sua versão rudimentar este tipo de abordagem dependia da experiência e intuição dos profissionais responsáveis, sendo pouco embasada em conhecimento estruturado e muito suscetível a erros (SELCUK, 2017). Contudo, com

o aumento de conectividade, da quantidade de dados disponíveis, e do uso de sistemas inteligentes com tecnologias como a Internet das Coisas, *Big Data* e Aprendizado de Máquina (*Machine Learning* - ML), ocorre o advento da Industria 4.0. Desta forma, manutenções preditivas puderam se tornar uma metodologia interdisciplinar baseada em dados históricos e modelos matemáticos executada em tempo real, sendo capaz de prever tendências, padrões de comportamento e correlações, que permitem a melhoria do processo de tomada de decisão de manutenção ao detectar falhas com antecedência (ZONTA, DA COSTA, *et al.*, 2020).

Este tipo de manutenção não busca apenas a previsão de momentos falhos, mas também uma operação mais eficiente aumentando a segurança, a qualidade, a confiança e a disponibilidade dos processos ao monitorar condições de operação e, consequentemente, evitar paradas de máquina desnecessárias (SELCUK, 2017).

Atualmente, a EPASA conta apenas com manutenções corretivas e preventivas. Como comentando, estes tipos de manutenção podem resultar em desperdício de recursos caso executadas com muita antecipação. Em um cenário pior, podem, inclusive, causar a indisponibilidade dos equipamentos, acarretando perdas consideráveis, uma vez que, por se tratar de UTEs, estas encontram-se na categoria de disponibilidade do setor elétrico e, normalmente, as solicitações da ONS para despachos de energia acontecem de forma efêmera.

Entretanto, por possuir um histórico de dados sensoriados que antecedem o ano 2018 e mais de 8.000 sensores dispersos pela planta, além de registros de suas manutenções, a EPASA possui as condições necessárias para a implementação de metodologias que tornem viável a realização de manutenções preditivas, otimizando seus processos como um todo.

2.1 APRENDIZADO DE MÁQUINA

O termo aprendizado pode ser definido como a habilidade de se adaptar, de acordo com estímulos externos, lembrando de parte das experiências passadas. Assim, no aprendizado de máquina, um modelo matemático (algoritmo) é treinado, para ser capaz de, posteriormente, com novas informações, tomar as decisões mais prováveis de serem bem-sucedidas (BONACCORSO, 2017). Desta forma, os algoritmos se adaptam, sem ser explicitamente programados, através de repetição (experiência) para melhor executar suas tarefas (EL NAQA e MURPHY, 2015).

Segundo Kelleher *et al.* (2020), o aprendizado de máquina pode ser definido como um processo automatizado de extração de padrões de dados. Ao criar tendências, é possível extrapolar as instâncias para além dos dados de treino (KELLEHER, MAC NAMEE e D'ARCY, 2020).

Historicamente, pode-se atribuir a Arthur Samuel a origem do termo "aprendizado de máquina" (EL NAQA e MURPHY, 2015). Na década de 50, o autor demonstrou que as máquinas (computadores) poderiam ser programadas para aprender a jogar damas (SAMUEL, 1959). Também na década de 50, Frank Rosenblatt (ROSENBLATT, 1958) desenvolveu o que viria a se tornar o protótipo das redes neurais artificiais modernas, o *perceptron*, um modelo matemático inspirado em ideias sobre o funcionamento do sistema nervoso humano.

Contudo, alguns anos depois, em 1969, os estudos de Minsky e Papert demonstraram limitações na complexidade dos problemas que poderiam ser resolvidos por *perceptrons*, enfatizando que estes não poderiam representar funções lógicas como XOR ou NXOR (MINSKY e PAPERT, 1969), uma vez que estas não eram questões linearmente separáveis. Este estudo culminou no então denominado “primeiro inverno” da área de estudos de inteligência artificial, devido à redução de financiamentos e pesquisas na área até cerca de 1980 (FRADKOV, 2020).

Entretanto, grandes avanços foram alcançado pelos estudos de estruturas e técnicas de aprendizagem de redes neurais multicamadas, como o desenvolvimento de *percetrons* multicamadas (*Multilayer Perceptron* - MLP) por Werbos, em 1975 (WERBOS, 1975), e a proposta de uma rede neural convolucional multicamada hierárquica, conhecida como *Neocognitron*, por Kunihiko Fukushima, em 1982 (FUKUSHIMA e MIYAKE, 1982).

Além disto, um impacto significativo também foi causado pelo aprofundamento da utilização da retropropagação de erros (mais conhecida como *backpropagation*) para ajustar os pesos das camadas ocultas de redes neurais, permitindo o treinamento de redes neurais cada vez mais complexas (RUMELHART, HINTON e WILLIAMS, 1986; WERBOS, 1990).

Paralelamente, outros métodos e algoritmos foram criados, como as árvores de decisão desenvolvidas por Quinlan em 1986 (QUINLAN, 1986), que ganharam grande destaque principalmente pelo seu uso em métodos de conjunto (*ensemble*), os quais incluem os algoritmos de Florestas Aleatórias (*Random Forests* - RF) e de Incremento de Gradiente (*Gradient Boosting* – GB). Ainda vale a pena citar os

algoritmos de Máquinas de Vetor de Suporte (*Support Vector Machines* – SVM) resultantes dos trabalhos de Cortes e Vapnik em 1995, os quais merecem grande destaque pelos seus resultados (CORTES e VAPNIK, 1995).

Com estes avanços, grandes expectativas foram criadas. Porém, os resultados práticos não conseguiram atendê-las, resultando novamente na redução de investimentos na área no início de 1990, no chamado “segundo inverno”. Entretanto, na primeira década do século XXI, a quantidade de dados armazenados e disponíveis (*Big Data*) se tornou tão grande que novas abordagens surgiram por necessidades práticas e não por curiosidade dos cientistas. Associado a isto, a redução de custos de computação e de memória, além do desenvolvimento de algoritmos de aprendizado profundo (*deep learning*) reascenderam o interesse pela área (FRADKOV, 2020).

2.2 TIPOS DE APRENDIZADO DE MÁQUINA

Existem pelo menos três grandes divisões dos métodos de aprendizado de máquina, onde cada uma é utilizada para resolver diferentes tipos de tarefas. Estas divisões são: o aprendizado supervisionado, o aprendizado não supervisionado, e o aprendizado por reforço. A Figura 1 ilustra essas divisões, assim como as principais técnicas e aplicações dessas.

O aprendizado de máquina supervisionado consiste no treinamento dos algoritmos para a criação de um modelo que contenha relações entre um conjunto de características descritivas (entradas) e uma ou mais características alvos (saídas), com base em um conjunto de exemplos históricos (base de dados). Pode-se fazer uso deste modelo para prever as características alvos a partir de novas características descritivas (KELLEHER, MAC NAMEE e D'ARCY, 2020).

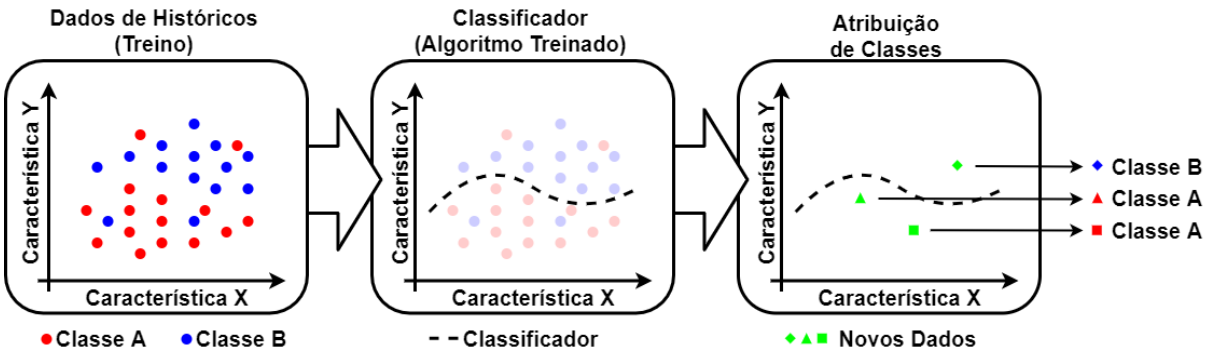
Existem, ainda, dois principais tipos de algoritmos supervisionados: a classificação e a regressão. Na regressão, os modelos são treinados para realizar a previsão de valores do domínio real. Já a classificação busca mapear as saídas dos modelos em classes pré-definidas (NASTESKI, 2017). Basicamente, a diferença destas técnicas se encontra no tipo da variável alvo. Se esta for nominal, ou seja, representar classes, naturalmente o algoritmo se trata de uma classificação. Caso a variável alvo seja continua, os algoritmos são ditos como regressivos.

Figura 1 – Principais subdivisões e aplicações do aprendizado de máquina.

Fonte: Adaptado de Geekstyle (2021).

Na Figura 2 é representado de maneira ilustrativa o processo de classificação, onde através das características (X e Y) de um registro é possível associá-lo (classificá-lo) a uma classe (A ou B), sendo esta classe uma variável nominal, o que caracteriza o processo de classificação.

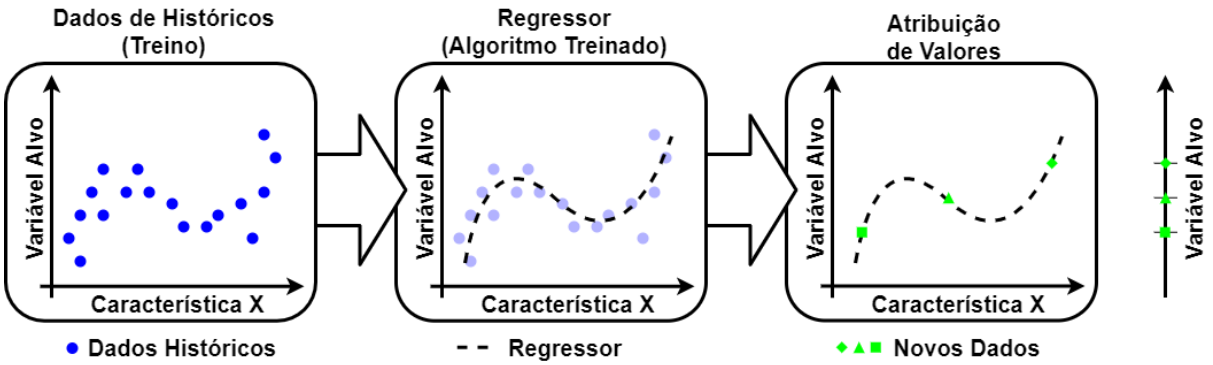
Figura 2 – Ilustração simplificada do funcionamento de algoritmos de classificação.



Fonte: elaborado pelo autor (2023).

Já na Figura 3, é ilustrada uma regressão, onde através dos dados históricos (pontos azuis) uma regra ou tendência pode ser criada correlacionando as características (X) destes dados com a variável alvo. Desta forma, podem ser estimados os valores das variáveis alvos para novos dados (pontos verdes).

Figura 3 – Ilustração simplificada do funcionamento de algoritmos de regressão.



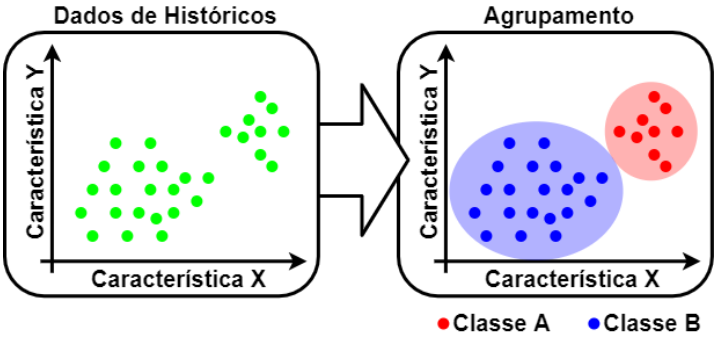
Fonte: elaborado pelo autor (2023).

Aplicações comuns de modelos de aprendizado supervisionado incluem detecção de padrões, detecção de spam, classificação automática de imagens, processamento de linguagem natural, análise de sentimentos, além, é claro, de análises preditivas baseadas tanto em regressões quanto em classificações (BONACCORSO, 2017), como o próprio tema deste trabalho.

O aprendizado de máquina não supervisionado se baseia na ausência de qualquer supervisão, e, consequentemente, medidas de erro. De acordo com Ghahramani (GHAHRAMANI, 2003), o aprendizado não supervisionado é formulado para extrair a estrutura de uma amostra de dados. Desta forma, é majoritariamente utilizado para agrupamentos e reduções de dimensionalidade (MAHESH, 2020).

Assim, o agrupamento é útil quando há a necessidade de aprender como um conjunto de elementos pode ser agrupado de acordo com a sua similaridade (BONACCORSO, 2017). Portanto, os grupos criados pelos métodos não supervisionados serão razoavelmente similares com uma classificação intuitiva (AYODELE, 2010). A Figura 4 busca ilustrar o processo de agrupamento, onde instâncias sem classes pré-definidas, representadas pelos pontos verdes, podem ser agrupadas por similaridade (no caso, por proximidades) em duas classes distintas (classe A e B).

Figura 4 – Ilustração simplificada do funcionamento de algoritmos de agrupamento.

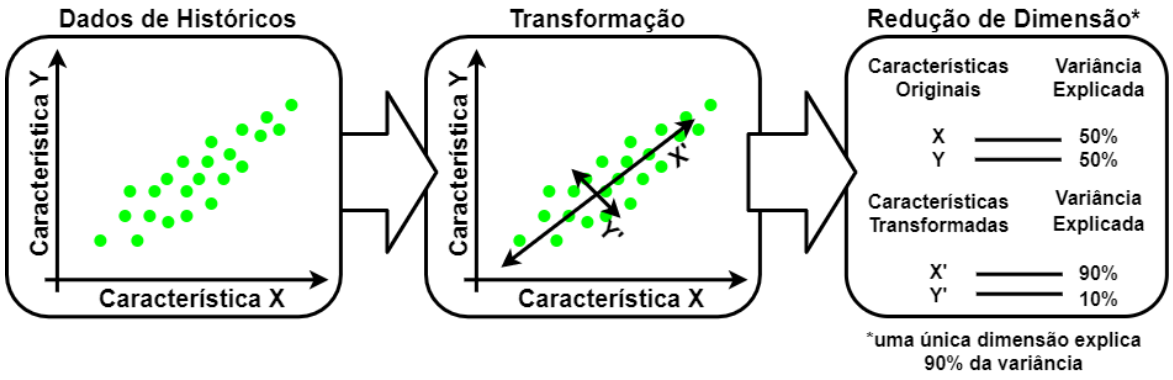


Fonte: elaborado pelo autor (2023).

Por outro lado, a redução da dimensionalidade busca transferir a informação dos dados do seu espaço original de alta dimensão para um novo espaço de dimensão menor, preservando o máximo possível das características essenciais e originais neste novo espaço (HUANG, WU e YE, 2019). Em outras palavras, transfere a informação de um conjunto de diversas variáveis para um conjunto menor, buscando manter a informação, porém reduzindo a quantidade de variáveis.

A Figura 5 ilustra este processo, em que duas dimensões (X e Y) são transformadas em duas novas (X' e Y'). Apesar da dimensão aqui não ser reduzida, é possível notar que apenas uma destas novas variáveis já condensa boa parte da informação presente, anteriormente, em duas variáveis distintas. Obviamente, aqui estamos tratando de poucas dimensões, mas este processo ilustra um exemplo que poderia ser aplicado a uma quantidade muito maior de dimensões, onde de fato se faria necessário tal procedimento.

Figura 5 – Ilustração simplificada do funcionamento de algoritmos de redução de dimensionalidade.



Fonte: elaborado pelo autor (2023).

As aplicações comuns de modelos de aprendizado não supervisionado incluem segmentação de objetos, como, por exemplo, produtos, filmes e músicas, que podem ser utilizados em sistemas de recomendação, assim como a detecção de similaridade e marcação (rotulagem) automática (BONACCORSO, 2017).

2.3 VALIDAÇÃO DE MODELOS

2.3.1 Métricas para Modelos Regressivos

Para a avaliação de modelos regressivos, duas principais métricas são utilizadas: o coeficiente de determinação e a raiz do erro quadrático médio (*Root Mean Squared Error* - RMSE).

O coeficiente de determinação é uma métrica bem estabelecida para avaliação da qualidade do ajuste de modelos regressivos (ZHANG, 2017). Comumente denominado de R-quadrado (R^2), este coeficiente reflete a proporção da variância de uma variável que pode ser explicada por um modelo (ou outra variável) (NAGELKERKE e OTHERS, 1991).

Para tanto, temos que a variação total da variável pode ser definida pela soma dos quadrados totais (SQT), assim como a variação explicada pode ser definida pela soma dos quadrados da regressão (SQR), e, por fim, a variação não explicada pode ser definida pela soma dos quadrados dos erros (SQE). SQT, SQR e SQE são expressas pelas Equações (1), (2) e (3), respectivamente, onde n representa o número total de variáveis avaliadas, e y_i, $\bar{y}$, e $\hat{y}_i$ representam as variáveis de interesse, sua média e o valor predito, respectivamente.

$$SQT = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{1}$$

$$SQE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

$$SQR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \tag{3}$$

Além disso, é válida a relação

$$SQT = SQE + SQR \tag{4}$$

Finalmente, tem-se que

$$R^2 = 1 - \frac{SQE}{SQT} = \frac{SQR}{SQT} \tag{5}$$

Por se tratar de uma proporção, o R-quadrado (Equação (5)), normalmente, expressa um valor entre 0 e 1, que muitas vezes pode ainda ser expresso em termos percentuais. Desta forma, quanto maior seu valor, melhor o ajuste do modelo. Contudo, caso a performance do modelo seja inferior à média dos valores de interesse para predição destes, o SQE será superior ao SQT, resultando em um R-quadrado negativo (HAYASHI, 2011).

Contudo, uma limitação bastante conhecida deste coeficiente é o seu aumento relativo à inserção de variáveis no modelo, sendo essas relevantes ou não (SRIVASTAVA, SRIVASTAVA e ULLAH, 1995). Para tanto, uma correção levando em conta os graus de liberdade pode ser aplicada, a qual é expressa pela Equação (6):

$$R_a^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1} \tag{6}$$

onde R_a^2 é o coeficiente de determinação ajustado e k é o número de variáveis utilizadas para a predição.

Para a avaliação do erro, duas principais métricas são utilizadas: a raiz do erro quadrático médio (RMSE) e o erro médio absoluto (*Mean Absolute Error* – MAE), definidas pelas Equações (7) e (8):

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \tag{7}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \quad (8)$$

É possível notar que ambas as métricas eliminam a influência de valores negativos: a primeira por elevar o erro ao quadrado, e a segunda por extrair o valor absoluto deste. Contudo, o RMSE, diferentemente do MAE, penaliza valores elevados de erro. Assim, o RMSE é, por definição, maior que o MAE (CHAI e DRAXLER, 2014).

Existem algumas divergências na literatura sobre qual destas métricas é a mais aconselhável para a avaliação dos erros de um modelo (CHAI e DRAXLER, 2014; WILLMOTT e MATSUURA, 2005). Entretanto, ambas são amplamente utilizadas.

2.3.2 Métricas para Modelos de Classificação

Para problemas de classificação binária, a avaliação de uma solução ótima pode ser definida com base na matriz de confusão, representada pelo Quadro 1. As colunas da matriz representam os valores preditos, enquanto as linhas representam os valores reais. Em suas intersecções são apresentadas a quantidade das amostras positivas e negativas que foram classificados corretamente ou erroneamente (HOSSIN e SULAIMAN, 2015).

Quadro 1 – Modelo de Matriz de Confusão.

	Classificação Positiva	**Classificação Negativa**
Amostra Positiva	VP	FN
Amostra Negativa	FP	VN

Fonte: elaborado pelo autor (2023).

Desta forma, para problemas de classificação binária existem apenas quatro resultados possíveis:

- **Verdadeiro Positivo (VP):** uma amostra que foi corretamente classificada como positiva.

- **Verdadeiro Negativo (VN):** uma amostra que foi corretamente classificada como negativa.
- **Falso Positivo (FP):** uma amostra que foi erroneamente classificada como positiva.
- **Falso Negativo (FN):** uma amostra que foi erroneamente classificada como negativa.

A matriz de confusão é uma ferramenta de análise que ilustra de maneira detalhada a avaliação de um teste, sendo base para o cálculo de métricas de desempenho (KELLEHER, MAC NAMEE e D'ARCY, 2020), dentre as quais se destacam: a Acurácia, a Precisão, a Sensibilidade e o *F1-Score*. Cada uma destas métricas focam em diferentes análises.

A Acurácia, definida pela Equação (9), reflete a proporção de previsões corretas em relação ao número total de amostras analisadas.

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \tag{9}$$

A Precisão, por sua vez, definida pela Equação (10), está diretamente relacionada com a capacidade de classificar amostras verdadeiramente positivas, evitando a classificação errônea como negativa.

$$Precisão = \frac{VP}{VP + FP} \tag{10}$$

Por outro lado, a Sensibilidade (bastante conhecida por sua denominação em inglês *recall*), definida pela Equação (11), determina a capacidade de detectar amostras verdadeiramente positivas entre todos os potenciais positivos.

$$Sensibilidade = \frac{VP}{VP + FN} \tag{11}$$

Finalmente, o *F1-Score* é uma métrica que leva em conta a Precisão e a Sensibilidade, sendo definida como uma média harmônica destas duas métricas. É expressa pela Equação (12).

$$F1\text{-}Score = 2 \times \frac{Precisão \times Sensibilidade}{Precisão + Sensibilidade} \quad (12)$$

A Acurácia é a métrica mais utilizada para a avaliação de modelos de classificação. No entanto, essa métrica possui várias limitações, uma vez que sua simplicidade pode levar a soluções abaixo do ideal, especialmente ao lidar com dados onde a distribuição de classes é desbalanceada (HOSSIN e SULAIMAN, 2015).

Para classificações que envolvam mais de duas classes (multiclasses), a Matriz de Confusão é expandida e a avaliação dos verdadeiros e falsos positivos e negativos deve ser realizada de forma individual para cada classe. No Quadro 2, observa-se a avaliação da classificação das amostras “A”. É importante ressaltar que todos os valores não estão incluídos na linha ou coluna referentes às classificações e amostras “A” serão, necessariamente, verdadeiros negativos.

Quadro 2 – Exemplo de Matriz de Confusão multiclasses, avaliada para amostras A.

	Classificação A	**Classificação B**	**Classificação C**
Amostra A	VP	FN	FN
Amostra B	FP	VN	VN
Amostra C	FP	VN	VN

Fonte: elaborado pelo autor (2023).

Desta maneira, é possível calcular as mesmas métricas utilizadas na classificação binária para cada uma das classes e combinar os resultados através de diferentes métodos, sendo eles: macro, balanceada e micro (GRANDINI, BAGLI e VISANI, 2020).

Na abordagem macro, as classes são tratadas como elementos básicos do cálculo, não havendo distinção entre essas. Essa metodologia consiste no cálculo da média aritmética simples das métricas de interesse de cada classe considerando que todas as classes possuem o mesmo peso, independentemente do desbalanceamento das amostras.

Sendo K o número de classes, podemos adquirir a Precisão-Macro e a Sensibilidade-Macro, através das Equações (13) e (14):

$$Precisão_{Macro} = \frac{\sum_{k=1}^{K} Precisão_k}{K} \quad (13)$$

$$Sensibilidade_{Macro} = \frac{\sum_{k=1}^{K} Sensibilidade_k}{K} \quad (14)$$

A partir dessas, é possível obter o F1-Macro através da Equação (15).

$$F1\text{-}Score_{Macro} = 2x \frac{Precisão_{Macro}\ x\ Sensibilidade_{Macro}}{Precisão_{Macro} + Sensibilidade_{Macro}} \quad (15)$$

Similar à agregação macro, a abordagem balanceada considera as classes como elementos básicos do cálculo. Contudo, é considerado o desequilíbrio entre essas. Nesse caso, a métrica avaliada de cada classe é multiplicada pela quantidade de amostras correspondentes e dividida pelo número total de amostras. Ou seja, trata-se de uma média ponderada das variáveis de interesse, representada pelas Equações (16) e (17), onde N_k e N representam a quantidade de amostras da classe "k" e a quantidade total de amostras, respectivamente.

$$Precisão_{Balanceada} = \frac{\sum_{k=1}^{K} Precisão_k x N_k}{N} \quad (16)$$

$$Sensibilidade_{Balanceada} = \frac{\sum_{k=1}^{K} Sensibilidade_k x N_k}{N} \quad (17)$$

Na abordagem micro, todas as amostras são consideradas como elementos básicos do cálculo, não havendo distinção entre estas. Assim podemos adquirir a Precisão-Micro e a Sensibilidade-Micro, através das Equações (18) e (19).

$$Precisão_{Micro} = \frac{\sum_{k=1}^{K} VP_k}{\sum_{k=1}^{K} VP_k + FP_k} = \frac{\sum_{k=1}^{K} VP_k}{N} \quad (18)$$

$$Sensibilidade_{Micro} = \frac{\sum_{k=1}^{K} VP_k}{\sum_{k=1}^{K} VP_k + FN_k} = \frac{\sum_{k=1}^{K} VP_k}{N} \quad (19)$$

Uma vez que a Precisão-Micro e a Sensibilidade-Micro possuem as mesmas relações, o F1-Score Micro também apresentará o mesmo resultado (visto que este é a média harmônica dessas duas métricas). Além disso, o somatório dos VP é igual à soma dos elementos da diagonal da matriz de confusão. A razão desse valor pelo total de amostras resulta na Acurácia. Ou seja, a abordagem micro sempre resulta na Acurácia (GRANDINI, BAGLI e VISANI, 2020).

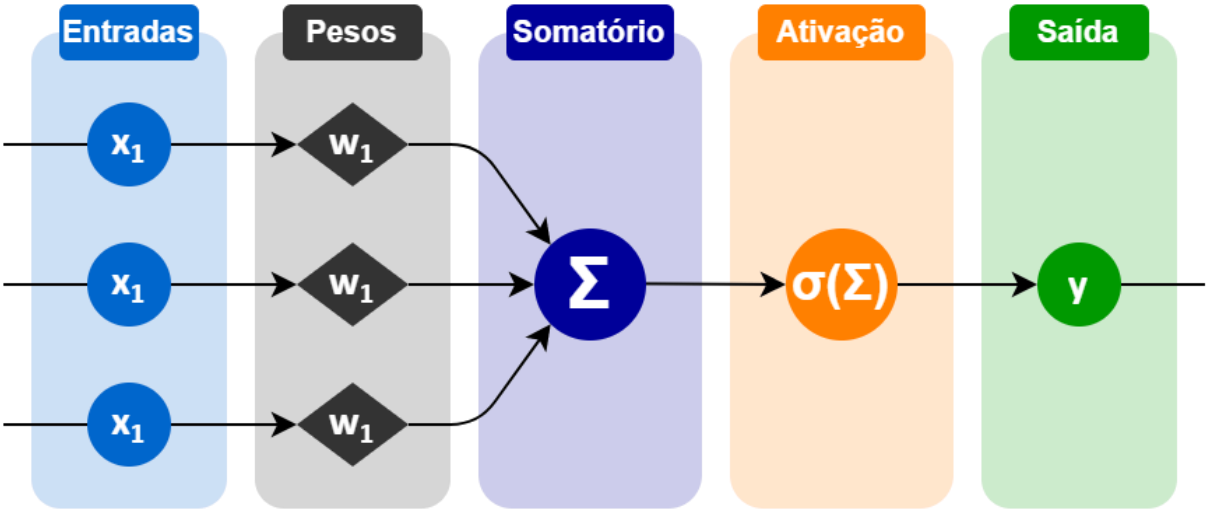
2.4 ALGORITMOS

2.4.1 *Perceptrons* e Redes Neurais

Uma rede neural é um modelo computacional de processamento paralelo e distribuído, composto de unidades simples de processamento (neurônios artificiais), que possui uma propensão natural para armazenar conhecimento experiencial e a disposição para seu uso. Assemelha-se ao cérebro humano em dois aspectos: (i) pelo conhecimento adquirido por meio de um processo de aprendizagem e (ii) pelas conexões entre os neurônios, as quais são utilizadas para armazenar este conhecimento adquirido (HAYKIN, 2001).A Figura 6 representa um único neurônio, elemento unitário para criação de redes neurais, onde x_1, x_2 e x_3 representam as entradas do modelo, e w_1, w_2 e w_3 representam os pesos. O somatório das entradas ponderado pelos pesos é representado por Σ. Finalmente, uma função de ativação σ é aplicada a esse somatório, resultando na saída y.

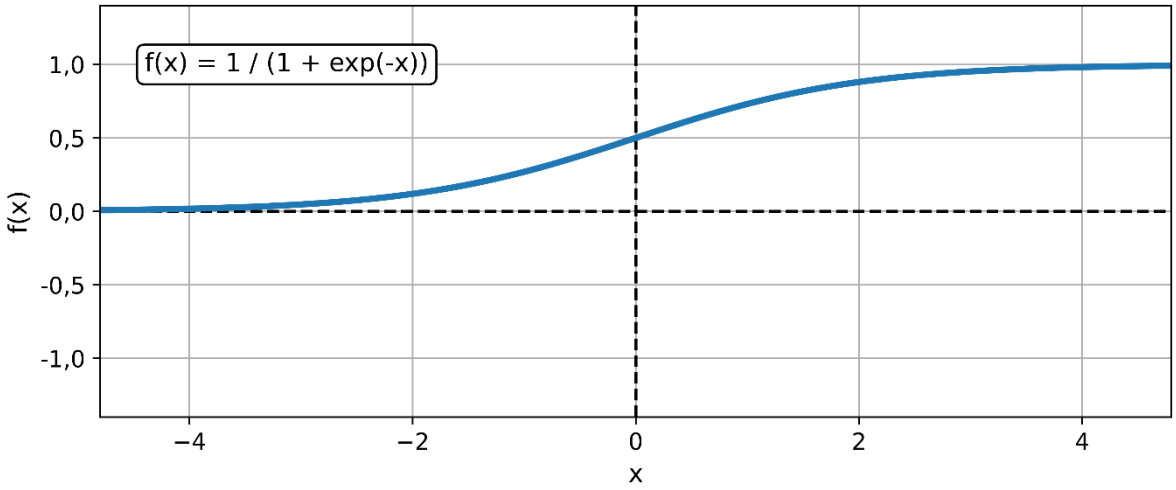
As funções de ativação utilizadas normalmente não são lineares, o que permite que redes neurais performem tarefas mais complexas. Sem a ativação da soma ponderada das entradas, um neurônio (ou rede neural) seria um polinômio de primeiro grau, tendo sua complexidade limitada e, consequentemente, a sua capacidade de aprendizado e reconhecimento de padrões seria afetada. Existem diversas funções de ativação. Entretanto, vale citar a função sigmoide ilustrada na Figura 7. Essa função não linear é a mais utilizada e possui a capacidade de transformar os valores reais dentro de um intervalo de 0 a 1 (SHARMA, SHARMA e ATHAIYA, 2017).

Figura 6 – Ilustração simplificada do funcionamento de um único neurônio de uma rede neural - *Perceptron.*



Fonte: Adaptado de Haykin (2001).

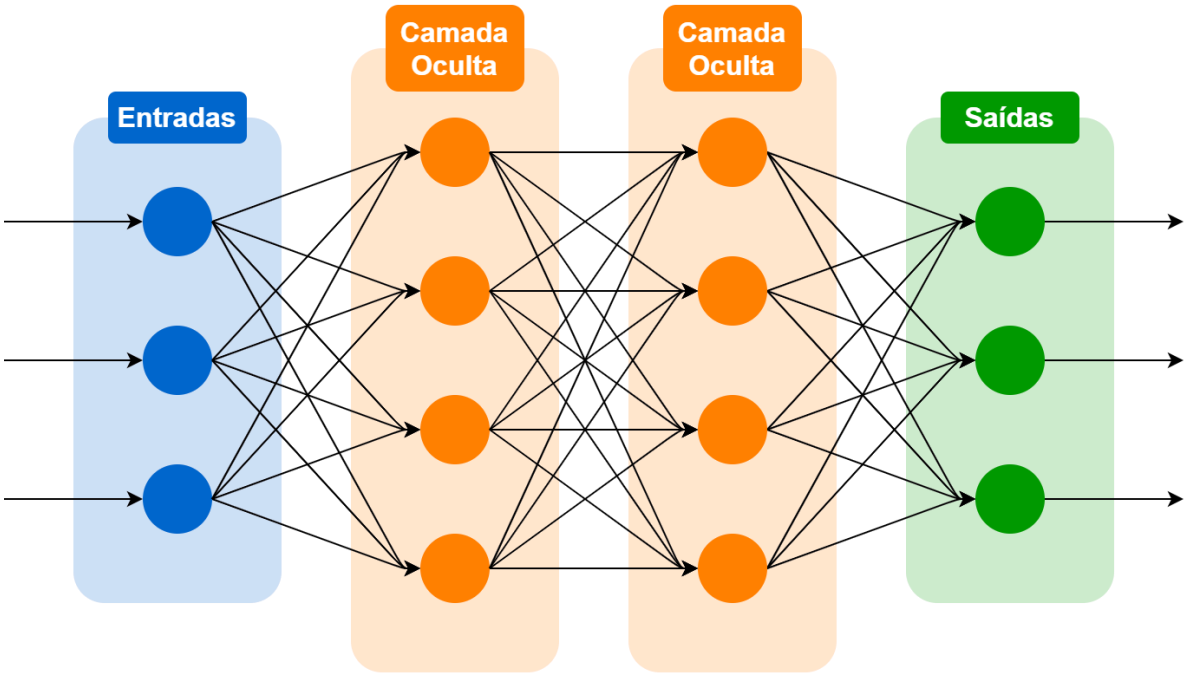
Figura 7 – Exemplificação do comportamento de uma função de ativação - Sigmoide.



Fonte: elaborado pelo autor (2023).

A capacidade computacional de um neurônio é limitada. Contudo, a junção de diversas unidades interligadas é capaz de resolver problemas de alta complexidade, finalmente formando o que é conhecido como redes neurais. A Figura 8 ilustra uma rede neural básica, denominada de *perceptron* multicamada (MLP). Através da retropropagação de erros é possível treinar de maneira supervisionada *perceptrons* multicamadas, os quais têm sido aplicados com sucesso para resolver uma gama de problemas (HAYKIN, 2001).

Figura 8 – Estrutura de uma rede neural multicamadas, com três entradas, duas camadas ocultas compostas de quatro neurônios e três saídas.



Fonte: Adaptado de Haykin (2001).

2.4.2 Métodos de Conjunto

Os métodos de conjunto (*ensemble*) buscam criar diversos modelos com tendências similares - normalmente denominados de modelos fracos (*weak learners*), ou ainda modelos base (*base learners*) - e combinar seus resultados, com o intuito de reduzir a variância. Assumindo que cada modelo gera erros independentes, mas que, de forma geral, estes modelos possuem a mesma tendência, convergindo em seus acertos, o agrupamento dos resultados reduziria o erro total ao calcular, por exemplo, a média dos resultados de cada componente. Vale citar, ainda, que a combinação dos modelos não necessariamente leva a um desempenho superior ao melhor modelo do grupo. Entretanto, reduz a probabilidade de ser selecionado um modelo com um desempenho inferior (POLIKAR, 2012).

Aqui são citadas as árvores de decisão (*decision trees*), sem o aprofundamento do tema, apenas para complemento da compreensão dos modelos de Florestas Randômicas e de Incremento de Gradiente, uma vez que ambos os modelos utilizam (ou normalmente utilizam) este método em sua base e serão

detalhados posteriormente. A Figura 9 ilustra a lógica do funcionamento de uma Árvore de Decisão.

Figura 9 – Exemplo ilustrativo do funcionamento de uma árvore de decisão, representando um processo de tomada de decisão cotidiano, no caso, se é necessário carregar um guarda-chuva.

Fonte: Adaptado de Karkare (2019).

Uma árvore de decisão pode classificar dados através de uma série de questionamentos sobre suas caraterísticas. Cada questionamento está contido em um nó, e todo nó interno aponta para nós filhos de acordo com cada possível resposta. Desta forma, os questionamentos formam uma hierarquia onde há um nó inicial (raiz), sendo este o primeiro questionamento, nós intermediários, e nós sem filhos (folhas), os quais apontam para a saída do modelo, ou resposta (KINGSFORD e SALZBERG, 2008).

As árvores de decisão suportam problemas de classificação com duas ou mais classes e podem ainda ser modificadas para lidar com problemas de regressão (KINGSFORD e SALZBERG, 2008).

As Florestas Aleatórias são uma extensão da ideia de agrupamento por ensacamento (*bagging*) de Breiman (BREIMAN, 1996) e foram desenvolvidas como um concorrente para os modelos de agrupamento sequencial (*boosting*). Assim, estas são essencialmente um conjunto de árvores de decisão treinadas paralelamente com um mecanismo de agrupamento, onde cada árvore depende de um subconjunto dos dados para treinamento (BREIMAN, 2001).

Desta forma, este algoritmo pode ser utilizado para predizer variáveis nominais, buscando a classificação destas, ou ainda variáveis contínuas, sendo assim utilizado como modelos regressivos (POLIKAR, 2012).

Tem-se como ponto de partida uma base de dados $(x_i, y_i)_{i=1}^{n}$, onde x_i representa um vetor de características (variáveis de entrada), y_i a variável alvo (saída) referente a este vetor, e n a quantidade de vetores nesta base de dados. Com subconjuntos desta base de dados, podemos treinar J modelos base (*base learners*) representados pelas funções $h_j|_1^J$ para prever y em função de x. Desta forma, é possível obter um comitê de modelos base agregados, definido como $f(x)$.

Para problemas regressivos, podemos obter a resposta (saída) deste conjunto de modelos simplesmente ao realizar a média dos modelos individuais, definida na Equação (20).

$$f(x) = \frac{1}{J}\sum_{j=1}^{J} h_j(x) \tag{20}$$

Por outro lado, para a classificações, $f(x)$ representa a classe mais frequentemente prevista (“votação”):

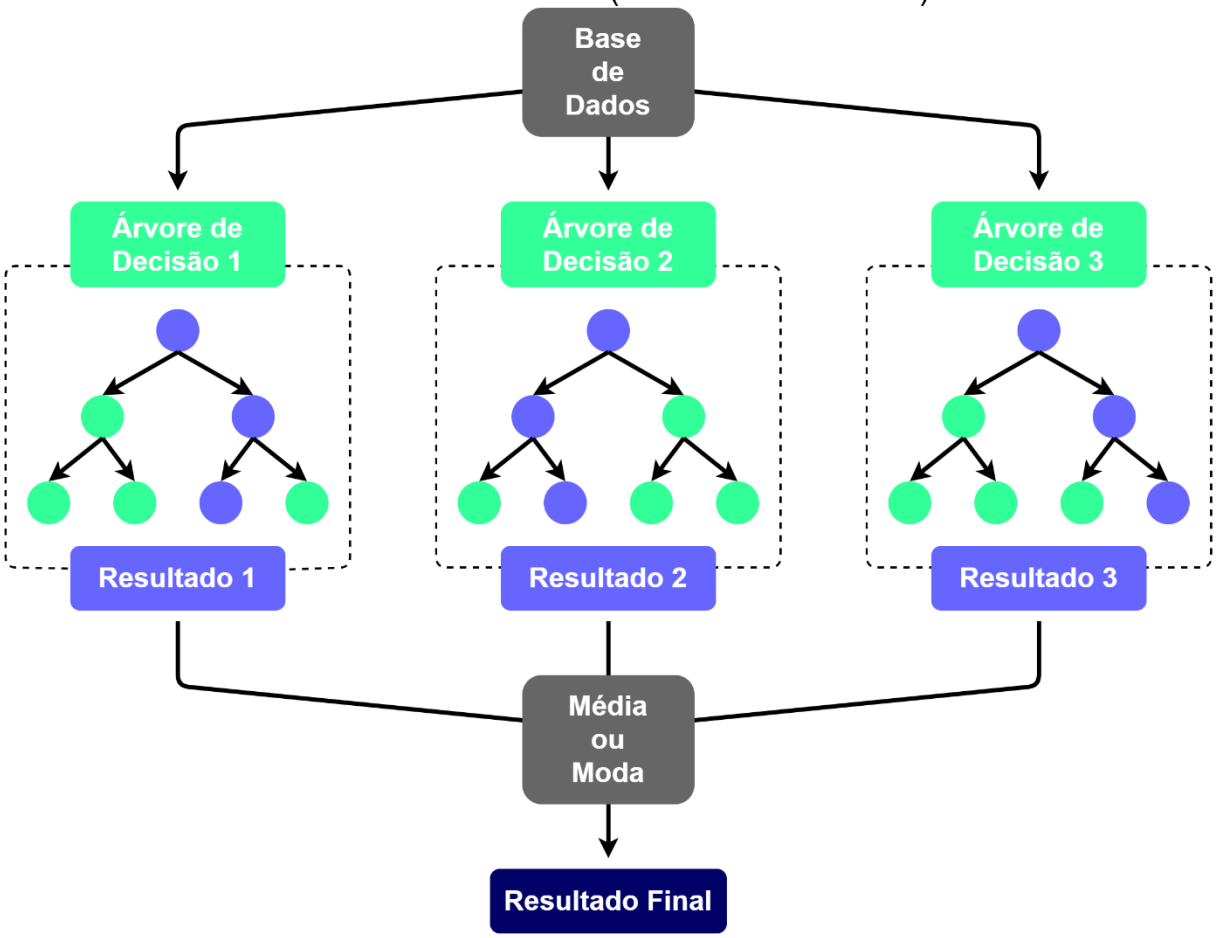
$$f(x) = argmax_y \sum_{j=1}^{J} \delta(y, h_j(x)) \tag{21}$$

onde,

$$\delta(y, h_j(x)) = \begin{cases} 1, & se\ y = h_j(x) \\ 0, & se\ y \neq h_j(x) \end{cases} \tag{22}$$

A Figura 10 busca representar de maneira simplificada como é constituído um modelo FR.

Figura 10 – Exemplo ilustrativo da lógica de criação (treinamento) de algoritmos de Florestas Randômicas (*Random Forest* – RF).



Fonte: Adaptado de TIBCO (s. d.).

Técnicas de incremento sequencial (*boosting*) foram introduzidas por Schapire, em 1990 (SCHAPIRE, 1990), no trabalho de "*The strength of weak learnability*". Trata-se de uma abordagem iterativa para gerar um modelo robusto, capaz de alcançar um erro de treinamento arbitrariamente baixo, a partir de um conjunto de modelos simples. Diferentemente das técnicas de *bagging*, onde os dados selecionados para o treinamento dos modelos de base são um subconjunto dos dados totais de treinamento, nas técnicas de *boosting*, os modelos são sequenciais, e o conjunto de dados de treinamento para cada modelo subsequente se baseia no erro gerado pelos modelos precedentes (POLIKAR, 2012).

Novamente, parte-se de uma base de dados $(x_i, y_i)_{i=1}^{n}$, onde x_i representa um vetor de características (variáveis de entrada), e y_i a variável alvo (saída) referente a

este vetor. Define-se uma função custo $L(y_i, h_j(x_i))$, sendo que $h_j(x_i)$ representa um modelo preditivo. Assim, a inicilização do algoritmo de GB pode ser dada por uma constante, definida na Equação (23).

$$h_0(x) = argmax_k \sum_{i=1}^{n} L(y_I, k) \tag{23}$$

É possível, então, construirmos uma nova base de dados $(x_i, r_{ij})_{i=1}^{n}$, a partir dos denominados pseudo-residuais (r_{ij}) obtidos através da Equação (24).

$$r_{ij}(x) = \left[-\frac{\partial L\big(y_i, h(x_i)\big)}{\partial h(x_i)} \right]_{h(x)=h_{j=1}} \quad para\ i = 1, \dots, n \tag{24}$$

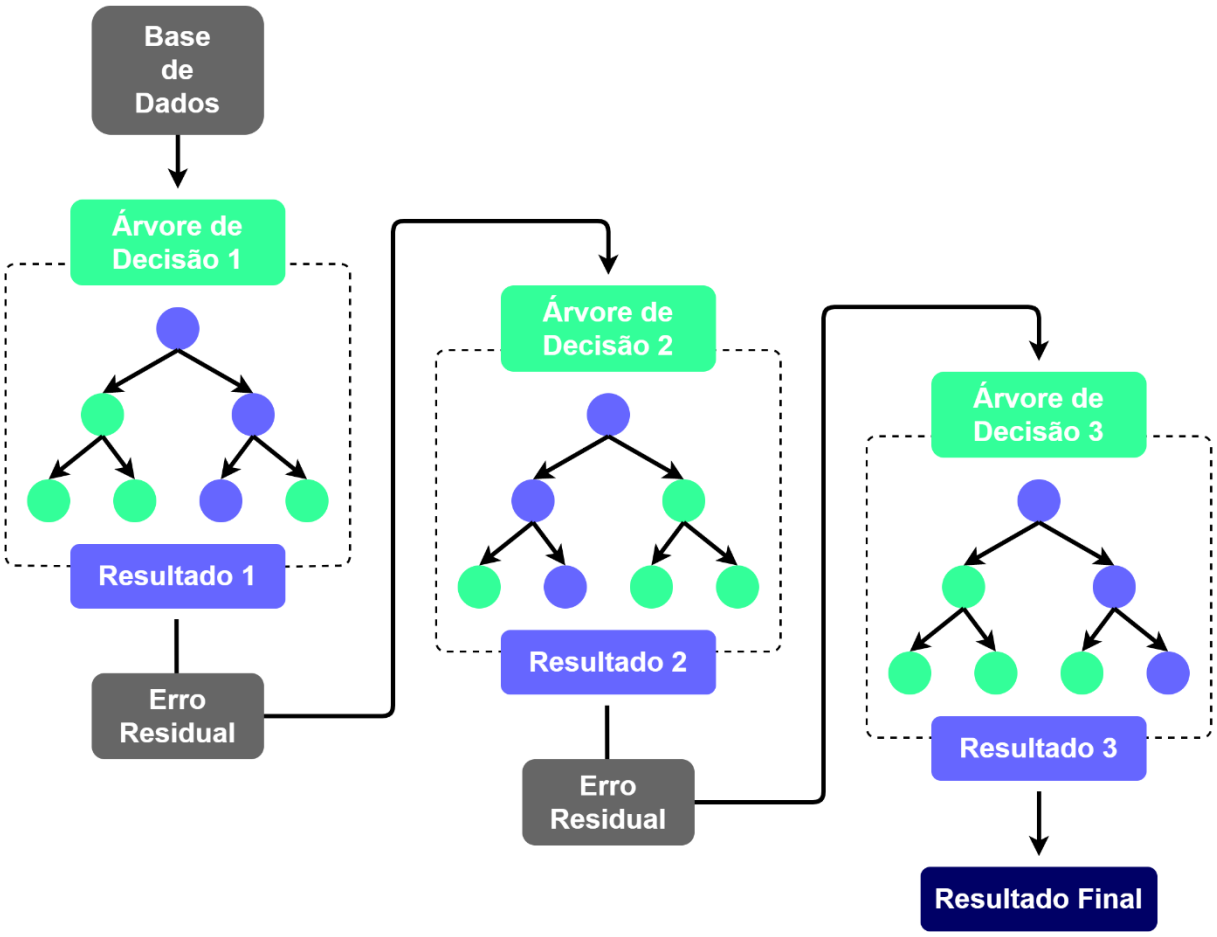
Essa nova base de dados permite a realização do treinamento de um modelo de predição h_j. Assim, podemos obter um modelo final $f(x)$ ao agrupar recursivamente J modelos de base, a partir da mesma abordagem, representado pela Equação (25).

$$f(x) = h_0 + \gamma \sum_{j=1}^{J} h_j(x) \tag{25}$$

Vale comentar a adição da variável γ na Equação (25). Este parâmetro controla a taxa de aprendizado do procedimento, e empiricamente determinou-se que valores pequenos ($\gamma \leq 0{,}1$) levam a uma generalização muito melhor do erro (FRIEDMAN, 2001).

Este processo aqui descrito é uma adaptação simplificada do trabalho de Friedman (FRIEDMAN, 2002), buscando favorecer o entendimento do processo, sem perder sua essência. A Figura 11 ilustra o processo como um todo.

Figura 11 – Exemplo ilustrativo da lógica de criação (treinamento) de algoritmos de Incremento do Gradiente (*Gradient Boosting* – GB).



Fonte: Adaptado de TIBCO (s. d.).

2.5 ALGORITMOS GENÉTICOS

A heurística é uma estratégia para solução de problemas de otimização, que busca produzir soluções aceitáveis para um problema complexo em uma escala de tempo razoável. Dada a complexidade de um problema de interesse, torna-se inviável buscar todas as combinações ou soluções possíveis, motivo pelo qual o foco destas estratégias é determinar, em uma escala de tempo prática, soluções boas, onde dentre estas pode-se esperar que algumas sejam quase ótimas (YANG, 2010).

Desta forma, as heurísticas são algoritmos dependentes do problema, que são desenvolvidas ou adaptados às particularidades de um determinado problema de otimização ou instância do problema. Por outro lado, a meta-heurísticas abrange algoritmos de otimização mais generalistas, independentes de problemas.

Provavelmente um dos mais populares algoritmos de meta-heurística é são os Algoritmos Genéticos (*Genetic Algorithms* - GAs), desenvolvido por John Holland na década de 1970 (HOLLAND, 1975/1992). Um Algoritmo Genético é um modelo computacional que mimetiza processos de evolução biológica para resolução problemas (MITCHELL, 1995). Desde sua criação, este tipo algoritmo se tornou tão bem-sucedido na solução de uma ampla gama de problemas de otimização que milhares de artigos de pesquisa e centenas de livros foram escritos (YANG, 2010).

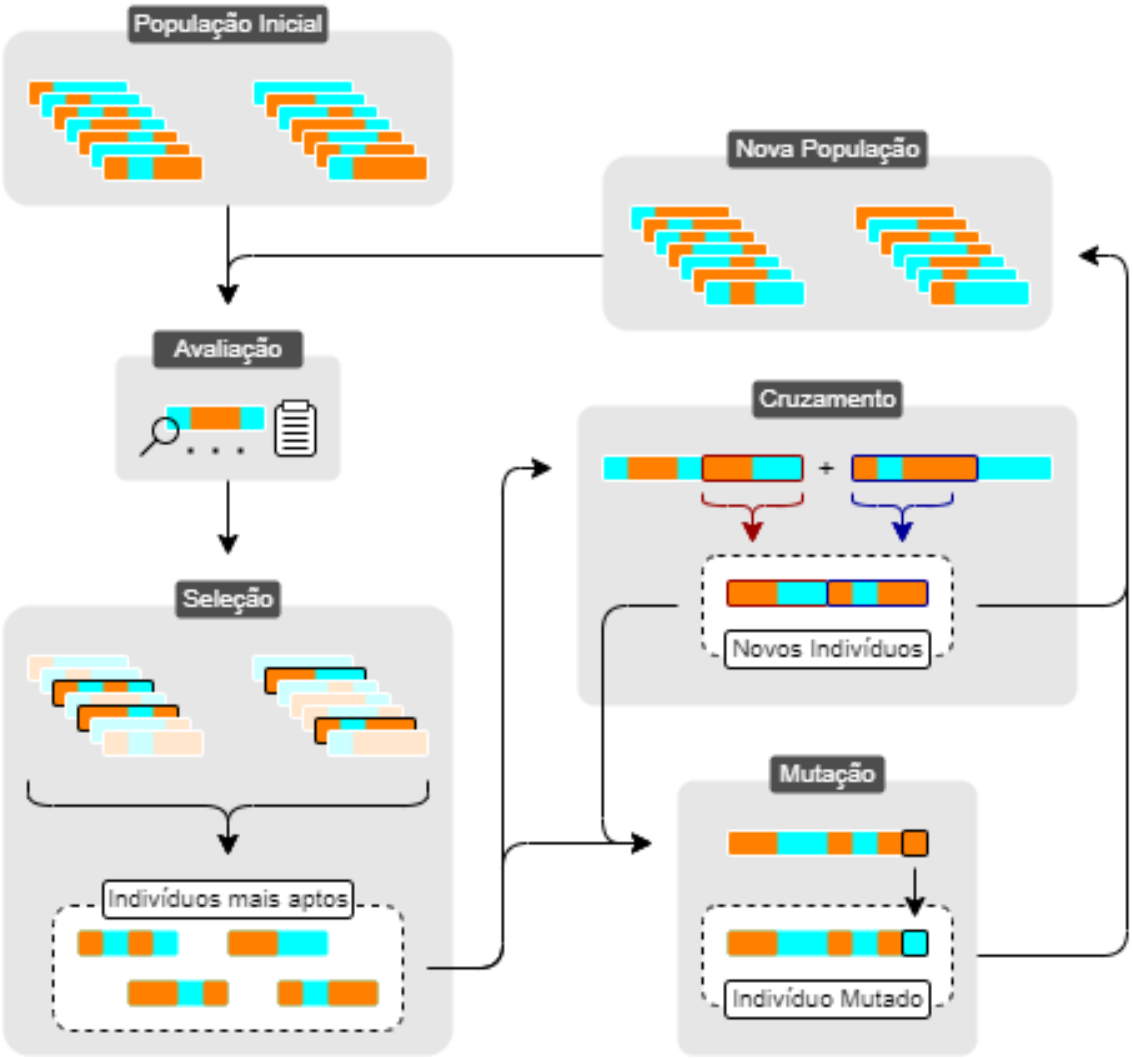
Para tanto, inicialmente é criada uma população de indivíduos de maneira aleatória. Estes indivíduos são representados por cromossomos (por exemplo, uma sequência de bits), onde cada um apresenta uma solução para o problema de interesse, e cada cromossomo é constituído por genes (por exemplo, bits) que representam suas variáveis. Dentre os diversos indivíduos é possível selecionar os mais aptos, ou seja, aqueles que apresentam as melhores soluções, para criar uma nova população. Esta etapa conhecida como seleção.

A nova população é criada a partir do cruzamento dos indivíduos selecionados, onde ocorre troca de partes correspondentes dos cromossomos de dois indivíduos. Existe ainda a possibilidade de mutações (inversão aleatória de bits) ou outras alterações nos cromossomos. A Figura 12 demonstra o mecanismo de um algoritmo genético.

Ao transformar o conjunto anterior de bons indivíduos em um novo conjunto, idealmente há uma chance melhor do que a média de este novo conjunto apresentar boas soluções. Ao repetir este ciclo de avaliação, seleção e operações genéticas, a aptidão da população tende a melhorar, e os indivíduos passam a apresentar soluções otimizadas para qualquer problema proposto (FORREST, 1996; YANG, 2010).

Existem diversas variantes dos GAs, dentre as quais vale ressaltar os Algoritmos Genéticos Multiobjetivo (*Multiobjective Genetic Algorithms* - MOGA) que são uma versão modificada dos GAs simples, principalmente em relação à atribuição da função objetivo. O primeiro MOGA foi desenvolvido por Fonseca e Fleming (1993), utilizando do conceito da eficiência de Pareto. Dessa forma, busca-se soluções no espaço objetivo de modo que não seja possível aprimorar ainda mais uma função objetivo sem prejudicar as demais (KATOCH, CHAUHAN e KUMAR, 2021).

Figura 12 – Exemplo ilustrativo da lógica da otimização realizada pelo algoritmo genético.



Fonte: elaborado pelo autor (2023).

Srinivas e Deb (1994) desenvolveram um Algoritmo Genético de Ordenação não Dominada (*Non-dominated Sorting Genetic Algorithm* - NSGA). No entanto, esse algoritmo apresenta falta de elitismo, necessidade de compartilhamento de parâmetros e alta complexidade computacional. Para aliviar esses problemas, Deb, Pratap, *et al.* (2002) desenvolveram um algoritmo genético de classificação não dominada elitista rápido (NSGA-II).

Vale destacar também o MOGA-II, uma versão aprimorada do MOGA proposto por Poloni e Mosetti (1993). Embora compartilhe o mesmo acrônimo do modelo inicialmente proposto por Fonseca e Fleming, trata-se de uma abordagem distinta. MOGA-II usa um elitismo inteligente de busca múltipla para robustez e

cruzamento direcional para convergência rápida. Sua eficiência é regida por seus operadores (*crossover* clássico, *crossover* direcional, mutação e seleção) e pelo uso de elitismo (RIGONI e POLES, 2005).

A variedade dos GAs vai além da função objetivo. Existem algoritmos genéticos paralelos, que buscam aprimorar o tempo de processamento através da distribuição dos indivíduos, como o *Master-Slave Parallel* GA (MS-PGA). Além disso, outras modificações são propostas, como o uso de sistemas caóticos utilizados para evitar convergências prematuras, resultando em uma grande diversidade de algoritmos (KATOCH, CHAUHAN e KUMAR, 2021).

2.6 GÊMEOS DIGITAIS

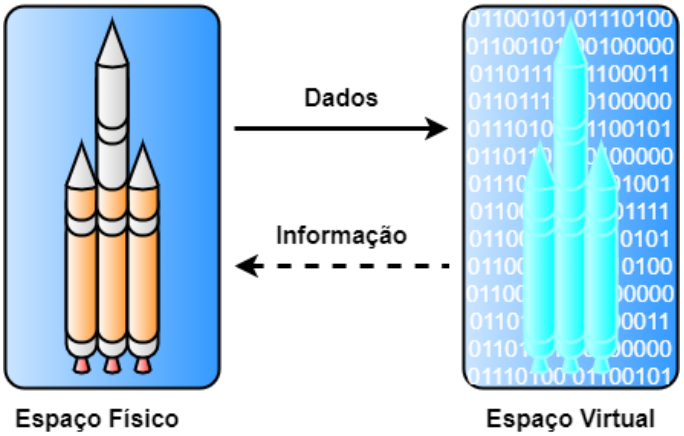
O conceito de Gêmeos Digitais (*Digital Twins* – DT) é relativamente recente. Sua terminologia foi introduzida pela primeira vez, como um equivalente digital de um produto, por Michael Grieves na Universidade de Michigan em 2003 (LIU, FANG, *et al.*, 2021). Poucos artigos foram publicados até 2011. Contudo, o rápido desenvolvimento de tecnologias como a Internet das Coisas (IoT), análise de dados, tecnologia de sensoriamento e de simulação contribuíram para a ascensão do uso de Gêmeos Digitais (TAO, ZHANG, *et al.*, 2018).

Em 2011, Tuegel *et al*. publicaram o primeiro artigo em periódico científico detalhando a utilidade do uso de Gêmeos Digitais para previsão da vida estrutural de aeronaves (TUEGEL, INGRAFFEA, *et al.*, 2011). Mais tarde, em 2012, a NASA formalizou a definição de Gêmeos Digitais e vislumbrou suas perspectivas na indústria aeroespacial (GLAESSGEN e STARGEL, 2012).

Grieves, em 2014 (GRIEVES, 2014), propôs um modelo conceitual contendo três partes principais: produtos físicos no espaço real, produtos virtuais no espaço virtual, e a conexões de dados e informações que unem os produtos virtuais e reais. Este modelo pode ser observado na Figura 13.

Pouco depois, em 2018, Tao *et al.* (TAO, ZHANG, *et al.*, 2018) propuseram cinco dimensões para a arquitetura de Gêmeos Digitais em equipamentos complexos, utilizando um aerogerador como caso de estudo. Suas dimensões estendem aquelas propostas por Grieves ao levar em conta os serviços e os dados dos próprios Gêmeos Digitais.

Figura 13 – Ilustração de um gêmeo digital, onde a entidade virtual se conecta à entidade física por meio da troca de dados e informações. Modelo proposto por Grieves (2014).



Fonte: Adaptado de Grieves (2014).

Desta forma, a ideia central de um gêmeo digital é simples, se tratando de uma entidade digital que reflete o comportamento de uma entidade física de forma precisa e em tempo real, onde estas estão conectadas. Entretanto, é difícil definir sua arquitetura conceitual, uma vez que diversas pesquisas o fazem através de diferentes perspectivas, e ainda assim cada descrição possui coerência (LIU, FANG, *et al.*, 2021).

Contudo, a aplicação mais popular para o uso de Gêmeos Digitais, tanto na academia quanto na indústria, desde o início do seu desenvolvimento até o presente, é a manutenção preditiva (LIU, FANG, *et al.*, 2021). Assim, a maioria das aplicações de Gêmeos Digitais estão relacionadas ao prognóstico e gestão da saúde de equipamentos (TAO, ZHANG, *et al.*, 2018).

2.7 TRABALHOS CORRELATOS

Aivaliotis, Georgoulias e Chryssoloris, em 2019, propuseram uma metodologia para realização de manutenções preditivas por meio de simulações baseadas em modelos físicos, aplicando o conceito *Digital Twin*. Com isto, apresentam uma metodologia para cálculo da Vida Útil Remanescente (*Remaining Useful Life* - RUL) de máquinas e equipamentos, descrita em quatro fases. A primeira fase consiste na modelagem física das máquinas, onde os parâmetros e propriedades do maquinário são utilizados para modelagem em um ambiente virtual capaz de simular o

comportamento da máquina real. Em seguida, na segunda fase, são coletados dados de sensores e controladores, para a realização de ajustes nas simulações, de forma síncrona. A terceira fase consiste na simulação dos modelos físicos, onde processos reais são utilizados como entrada dos modelos. Por fim, a quarta fase inclui a combinação do resultado da simulação e o monitorado dos equipamentos, com o intuito de calcular o RUL. Para validação, um estudo de caso corrobora a metodologia proposta com a previsão do RUL de um robô industrial (AIVALIOTIS, GEORGOULIAS e CHRYSSOLOURIS, 2019).

Negri *et al.* (2019) propuseram uma estrutura para incluir previsões de integridade do equipamento na atividade de planejamento, ao incorporar um Indicador de Integridade do Equipamento síncrono na simulação de DT, utilizando um Algoritmo Genético como abordagem meta-heurística para a otimização do planejamento. O Algoritmo Genético busca otimizar o planejamento de operações, sendo que existem diversos meios alternativos para esta. Para cada planejamento é possível calcular o desempenho da produção, assim como a previsão de falhas ou alertas. Além disto, o modelo é sincronizado, atualizando a taxa de falha no tempo para qualquer instante. Quando um ótimo local for alcançado (por critérios de convergência do Algoritmo Genético) os melhores indivíduos podem ser identificados (ou seja, o agendamento alternativo que forneceram as melhores pontuações de performances de produção) e o cronograma de produção pode ser considerado concluído. O artigo também propõe uma prova de conceito prática, ambientado em uma linha de montagem laboratorial (NEGRI, ARDAKANI, *et al.*, 2019).

Min *et al*. (2019) propõem uma abordagem para a construção de um Gêmeo Digital baseado em IoT de indústrias petroquímicas, utilizando aprendizado de máquina e um ciclo de prática para troca de informações entre a fábrica física e seu modelo de gêmeo digital. A modelagem considera três elementos: a fábrica física, a fábrica digital e a correspondência entre estas. De acordo com a estrutura básica das instalações e o conhecimento especializado, a estrutura digital é construída. A abordagem utiliza aprendizado de máquina para treinar os modelos de Gêmeos Digitais. Combinando informações de demanda do mercado e a solução ideal simulada é possível orientar a otimização do controle de produção. Além disto, os modelos serão treinados e otimizados iterativamente, com base em dados continuamente atualizados, para se adaptar dinamicamente. As abordagens foram avaliadas aplicando-as em uma unidade petroquímica, onde o modelo foi treinado com

dados industriais para realizar controle de produção inteligente, em tempo real, mostrando a eficácia da abordagem (MIN, LU, *et al.*, 2019).

Como comentando anteriormente, Tao *et al*., em 2018, propuseram uma arquitetura de cinco dimensões de Gêmeos Digitais em equipamentos complexos. Para tanto, três estágios no fluxo de trabalho são propostos: observação, análise e decisão. No primeiro ocorre a modelagem e calibração do DT, além da simulação e interação do modelo em tempo real com as condições de trabalho. Com isto é possível julgar se as entidades física e virtual apresentam diferenças toleráveis ou não. Em caso de inconsistência, tem-se o segundo estágio, que trata da identificação e previsão da causa da falha discriminando-as em gradual e abrupta. Por fim, o terceiro estágio aborda as estratégias de manutenção, buscando executá-las na entidade virtual primeiro para validação e, em seguida, na entidade física. A eficácia do método proposto é ilustrada através de um estudo de caso de uma turbina eólica (TAO, ZHANG, *et al.*, 2018).

Cai *et al*. (2017) apresentam uma abordagem para classificação de falhas de motores a diesel similares aos utilizados pela EPASA. Para tanto, os motores a diesel são divididos em quatro subsistemas de acordo com sua estrutura e características de falha. Algoritmos de *Support Vector Machine* (SVM) são treinados utilizando dados históricos e empregados na classificação de falhas. Além disso, algoritmos de mineração de dados são utilizados para extrair características das falhas e analisar relações implícitas entre as falhas e os subsistemas.

O diagnostico de falha de ignição para motores a diesel, utilizando *Gradient Boosting* e XGBoost, juntamente com informações de tempo-frequência de alta precisão de sinais de vibração, é apresentado por Tao *et al*. (2019). Testes de falha de ignição são realizados em diferentes velocidades do eixo, e os correspondentes sinais de vibração são adquiridos. Características dos sinais no domínio do tempo são obtidas utilizando métodos estatísticos, enquanto para sinais no domínio tempo-frequência são obtidas por meio da Transformada de Multissincronização (Multisynchrosqueezing Transform - MSST). Além disso, o método de incorporação linear local (Locally Linear Embedding - LLE) é utilizado para redução de dimensionalidade dos dados.

Aguilar *et al*. (2010) descrevem uma metodologia para monitoramento de vibração e diagnóstico de falhas em motores a diesel de grande porte, utilizando janelas temporais e análise de frequência. Através de um sistema especialista, a

abordagem tem foco em manutenções preditivas dos principais elementos mecânicos de cada cilindro dos motores: válvula de escape, cilindro e anéis de pistão, injetores e mancais principais do virabrequim. O sistema também conta com ferramentas para auxílio na identificação de anomalias referentes à combustão. A eficácia é demonstrada com base nos resultados de dois anos de operação em uma usina termelétrica com unidades de geradores a diesel localizada em Mahon, Espanha.

Gêmeos digitais para o sistema de refrigeração de água auxiliar das unidades geradoras foram desenvolvidos por Alves de Araujo Junior *et al*. (2021), visando otimizar o desempenho e a eficiência dos radiadores ao determinar o número adequado de ventiladores operantes. O modelo é baseado em um algoritmo de extração automática de regras *fuzzy*. Os gêmeos digitais podem atualizar as regras difusas no caso de novos eventos, como operação em estado estacionário e transiente (instabilidade e rampas de partida e parada). Em todos os cenários, o erro percentual médio foi inferior a 5% e o erro absoluto médio da temperatura foi inferior a 3 °C.

3 MATERIAIS E MÉTODOS

3.1 A EPASA

Em operação desde 2010, as Centrais Elétricas da Paraíba são compostas por duas Usinas Termelétricas (TNE e TPB), que combinadas possuem capacidade instalada de 342 MW. Estas duas UTEs se encontram em João Pessoa, na mesma localidade, e inclusive compartilham alguns equipamentos e processos, como tanques de armazenamento, estações de tratamentos e salas de controle. Contudo, as Unidades Geradoras e os principais sistemas atrelados a estas são segmentados.

Figura 14 – Visão aérea das Centrais Elétricas da Paraíba (EPASA).

Fonte: EBRASIL (s. d.).

Cada UTE conta com 20 motores diesel sobrealimentados, de 4 tempos, da fabricante MAN/STX, também denominadas de Unidades Geradoras Diesel (UGDs). Destes motores, 19 são do tipo 18V32/40 (18 pistões dispostos em V, com dimensões de 32 cm de diâmetro e 40 cm de curso) e 1 motor do tipo 9L32/40 (9 pistões com disposição linear, de dimensões de 32 cm de diâmetro e 40 cm de curso), totalizando 40 UGDs.

Os motores 18V possuem capacidade de geração de energia de 8,76 MW, enquanto os motores 9L possuem capacidade de 4,38 MW. Ambos os tipos de

motores possuem rotação de 720 rpm e são alimentados principalmente com óleo combustível (OCB1), e suplementarmente com óleo diesel. O consumo específico é de 212 kg/MWh, contudo este valor varia, principalmente, de acordo com o poder calorífico do combustível utilizado. Cada motor possui atrelado a ele um gerador da fabricante Hyundai do Tipo HSGD 1011K10, com tensão de 13.800 V, corrente de 458,3 A e frequência de 60 Hz.

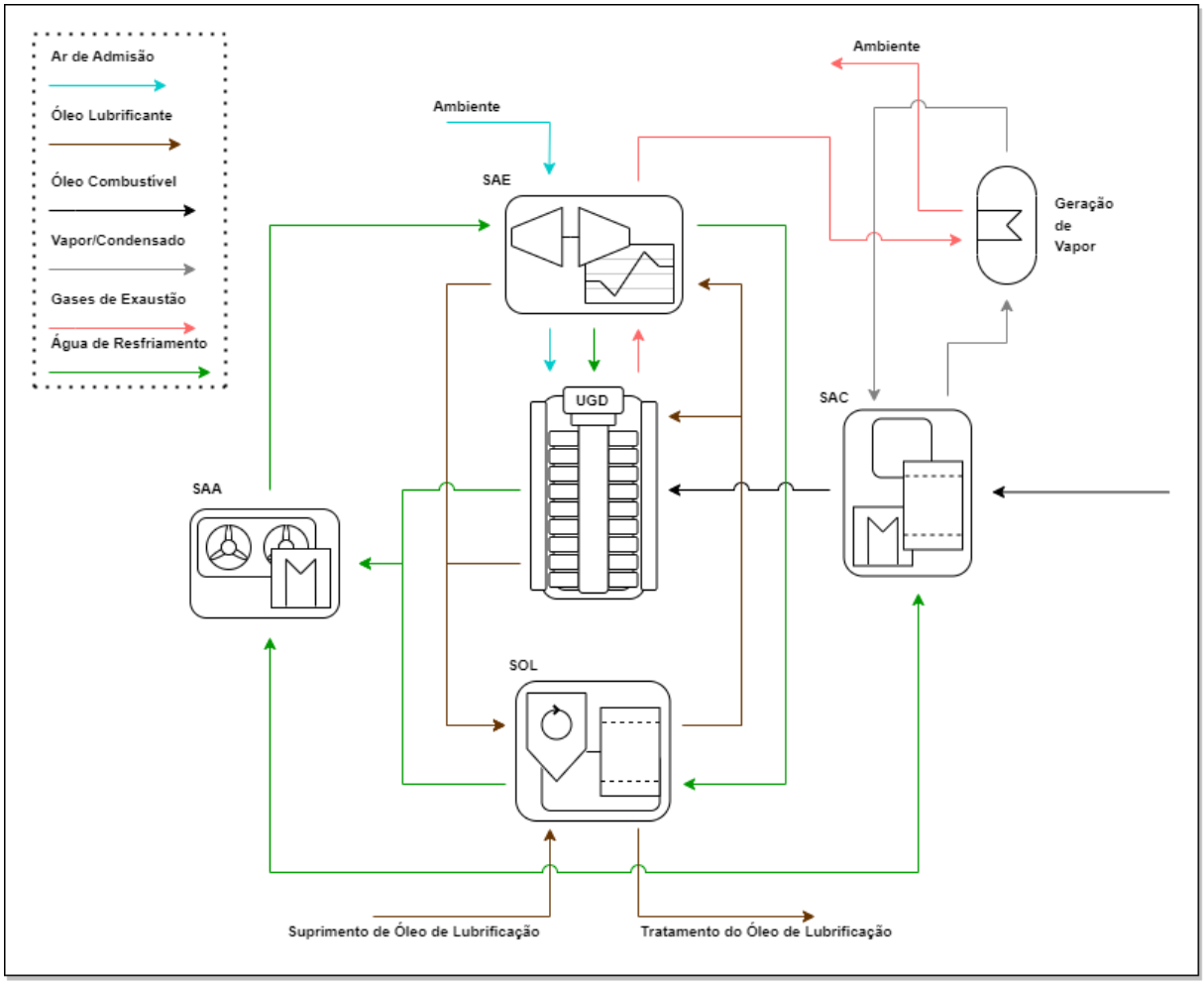
Figura 15 – Interior das instalações da EPASA - UGDs atreladas aos seus respectivos geradores, dispostas lado a lado.

Fonte: Epasa - Geração de Energia (s.d.).

Os principais sistemas auxiliares das unidades geradoras podem ser divididos em pelo menos quatro grupos: O Sistema de Admissão de Combustível (SAC), o Sistema de Água de Arrefecimento (SAA), o Sistema de Óleo de Lubrificação (SOL) e

o Sistema de Admissão de Ar e Exaustão de Gases (SAE), como é ilustrado pela Figura 16.

Figura 16 – Diagrama Geral dos principais processos auxiliares referentes a cada UGD.



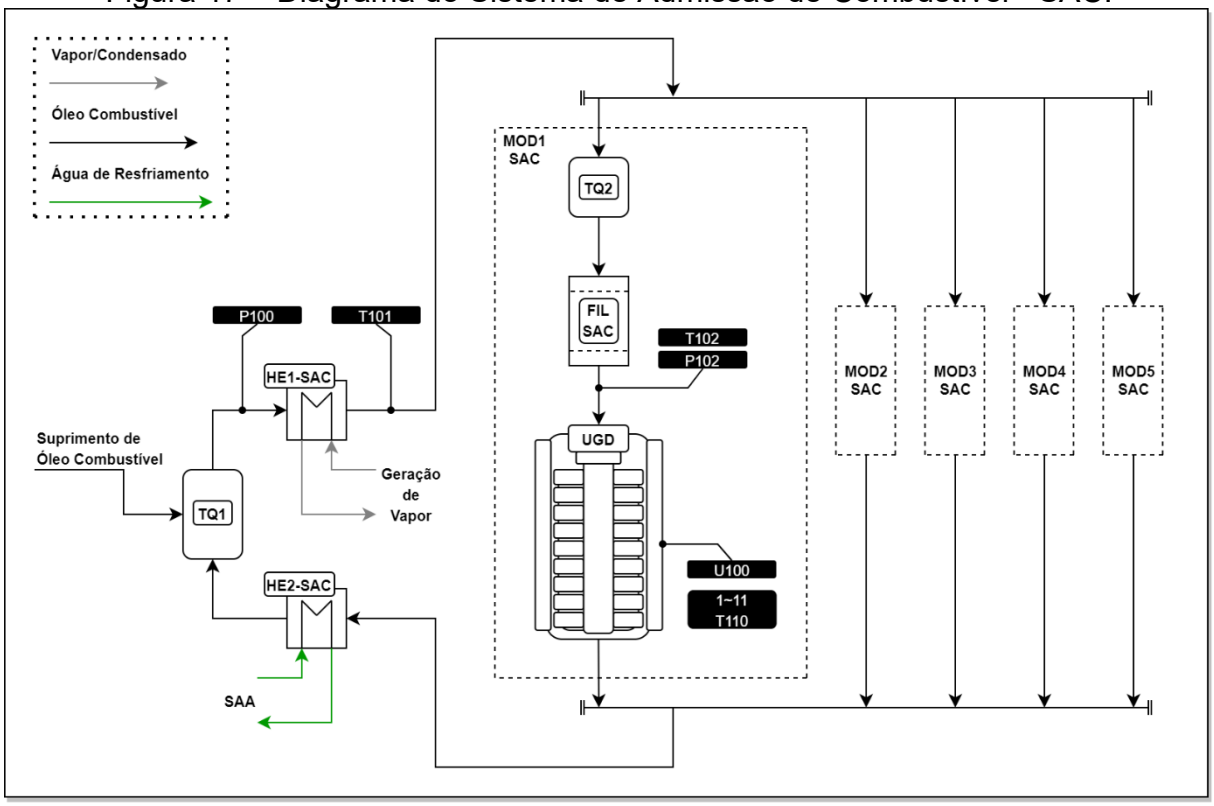
Fonte: elaborado pelo autor (2023).

3.1.1 Sistema de Admissão de Combustível (SAC)

O Sistema de Admissão de Combustível (SAC) é composto por unidades propulsoras denominadas de *boosters*. Esses equipamentos são responsáveis por alimentar cinco unidades geradoras paralelas. Nessas unidades propulsoras, o combustível é recirculado, de modo a garantir uma pressão constante nas tubulações antes de ser injetado nas câmaras de admissão. Para tanto, os equipamentos compartilhados são basicamente três: um tanque pulmão (TQ1), um trocador de calor para aquecimento (HE1-SAC), e um trocador de calor para resfriamento (HE2-SAC).

Há ainda um segundo tanque pulmão (TQ2) e um filtro (FIL-SAC), para cada unidade geradora, como pode ser observado na Figura 17. A Tabela 1 traz informações complementares sobre o sensoriamento desse subsistema.

Figura 17 – Diagrama do Sistema de Admissão de Combustível - SAC.



Fonte: elaborado pelo autor (2023).

Tabela 1 – Principais sensores de cada UGD relativos ao SAC.

Sensor	Descrição
P100	Pressão do combustível na entrada do HE1-SAC
T101	Temperatura do combustível na saída do HE1-SAC
P102	Pressão do combustível na entrada da UGD
T102	Temperatura do combustível na saída da UGD
U100	Potência gerada pela UGD
1~11 T110	Temperatura do mancal interno da UGD 1~11

Fonte: elaborado pelo autor (2023).

O principal combustível utilizado é o óleo pesado (*Heavy Fuel Oil* - HFO), o qual possui alta viscosidade. O óleo diesel também é utilizado ocasionalmente. Dada a sua viscosidade inferior, esse combustível é empregado com o intuito de limpar as tubulações ao final das operações, ou ainda para auxiliar na partida das máquinas. Dessa forma, o uso de óleo diesel é pontual e efêmero, não gerando impacto significativo no desempenho geral dos equipamentos. O tanque pulmão TQ1 permite

a realização da troca entre estes combustíveis, além de evitar grandes oscilações no fluxo dos mesmos.

O primeiro trocador de calor do ciclo (HE1-SAC) possui função de aquecimento do combustível e é utilizado quando há fluxo de HFO, com a finalidade de reduzir sua viscosidade, o que permite a pulverização subsequente deste nas câmaras de combustão. Após o aquecimento do óleo combustível, este é ramificado e direcionado para as UGDs operantes.

Antes de ser alimentado às unidades geradoras, cada uma das ramificações ainda conta com um tanque pulmão (TQ2) e um filtro (FIL-SAC). Apesar do combustível já ser tratado e limpo de impurezas, este último filtro garante que nenhuma partícula sólida seja injetada nas câmaras. Vale citar a existência de alguns componentes menores, internos aos motores e específicos para cada cilindro: as bombas injetoras, os tubos de alta pressão e os bicos injetores. Esses últimos possuem um sistema de arrefecimento separado, garantindo o controle de temperatura do óleo combustível.

Por fim, o combustível não injetado retorna aos *boosters*, passando por um último trocador de calor para resfriamento, o HE2-SAC, que possui como finalidade evitar o superaquecimento dos combustíveis, principalmente ao se tratar de óleo diesel. A água de resfriamento deste trocador é oriunda do SAA, mais especificamente do Sistema de Arrefecimento de Baixa Temperatura (SBT), descrito a seguir.

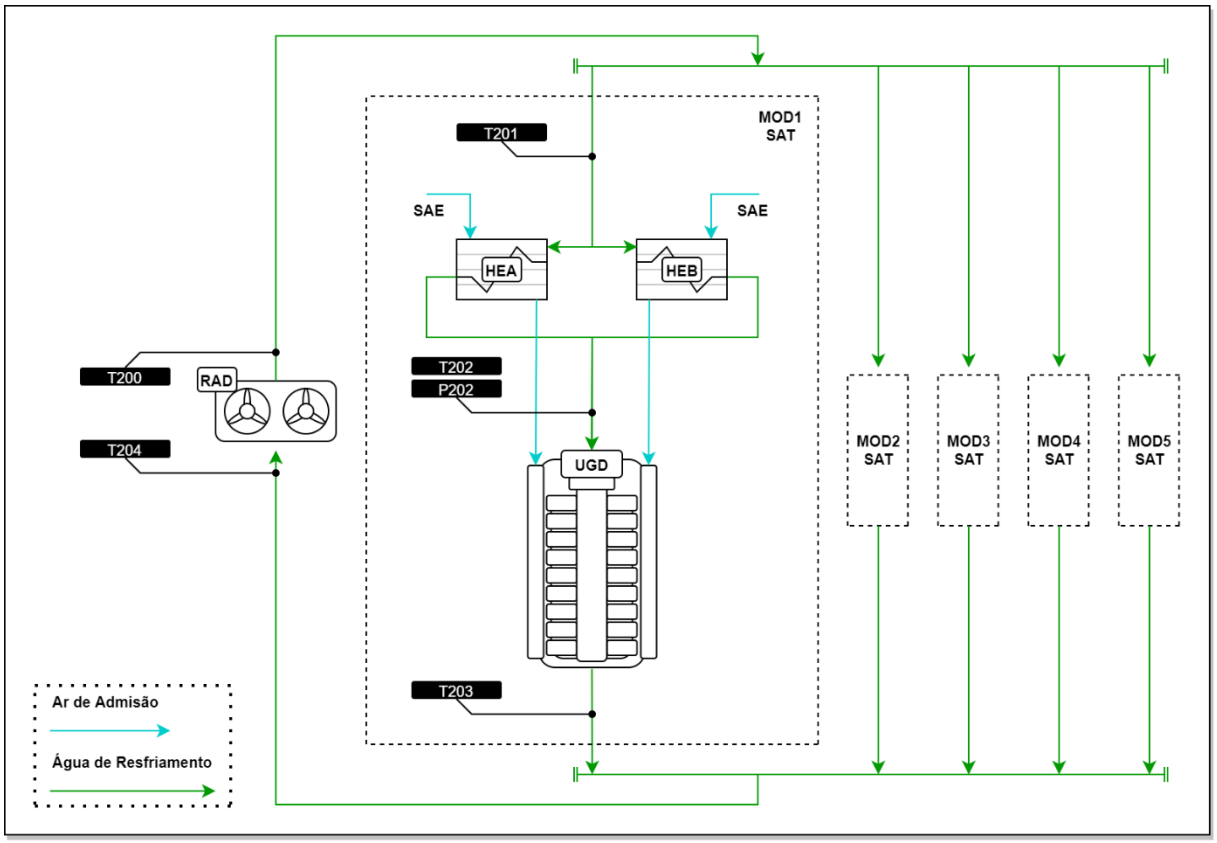
3.1.2 Sistema de Água de Arrefecimento (SAA)

Do mesmo modo que o SAC é compartilhado por cinco UGDs, o SAA também compartilha radiadores para cada grupamento de cinco UGDs. O sistema de arrefecimento ainda conta com duas correntes líquidas independentes, que apesar de compartilharem o mesmo radiador, podem ser subdivididas no Sistema de Arrefecimento de Alta Temperatura (SAT) e no Sistema de Arrefecimento de Baixa Temperatura (SBT), ilustrados pelas Figura 18 e Figura 19, respectivamente. Nessa mesma ordem, a Tabela 2 e a Tabela 3 trazem informações complementares desses sistemas.

No SAT, a água é resfriada por radiadores (RAD) e ramificada entre as UGDs, onde, para cada uma das unidades, é utilizada para o resfriamento do ar de admissão e da carcaça dos motores. O ar de admissão, por ser comprimido, sofre aquecimento

e, para se obter uma melhor eficiência das UGDs, é necessário seu resfriamento nos *intercoolers* (HEA e HEB). Os *intercoolers* são trocadores de calor de dois estágios, internos às unidades geradoras, onde o SAT é responsável pelo resfriamento do primeiro estágio. Essa mesma corrente líquida é posteriormente direcionada para a carcaça, onde promove o controle de temperatura dos motores de maneira geral. Por fim, a água de resfriamento torna aos radiadores completando o ciclo.

Figura 18 – Diagrama do Sistema de Água de Arrefecimento de Alta Temperatura - SAT.



Fonte: elaborado pelo autor (2023).

Tabela 2 – Principais sensores de cada UGD relativos ao SAT.

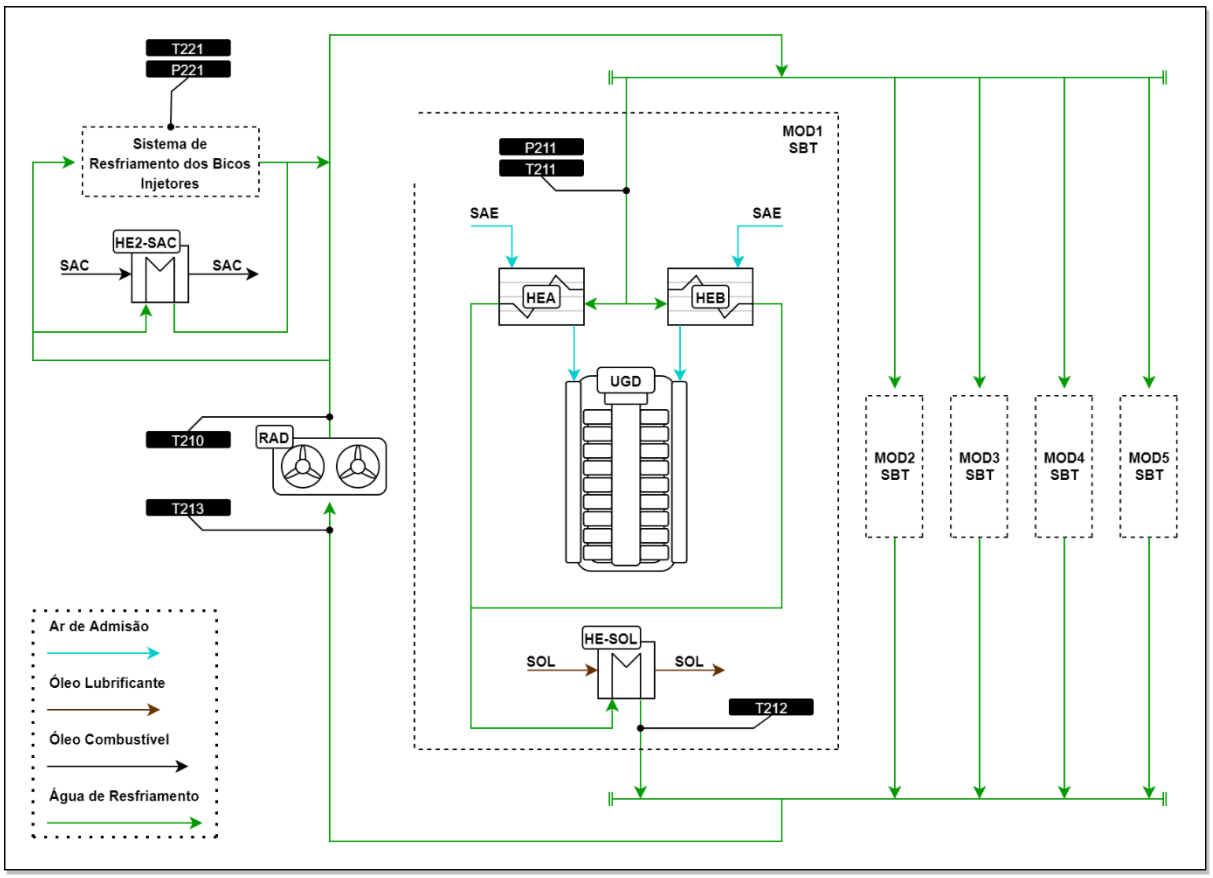
Sensor	Descrição
T200	Temperatura da água de arrefecimento de alta temperatura na saída do RAD
T201	Temperatura da água de arrefecimento de alta temperatura na entrada do HEA e HEB
P202	Pressão da água de arrefecimento de alta temperatura na entrada da UGD
T202	Temperatura da água de arrefecimento de alta temperatura na entrada da UGD
T203	Temperatura da água de arrefecimento de alta temperatura na saída da UGD
T204	Temperatura da água de arrefecimento de alta temperatura na entrada do RAD

Fonte: elaborado pelo autor (2023).

A corrente líquida do SBT é resfriada nos mesmos radiadores que o SAT. Entretanto, essa corrente participa de mais processos e, antes mesmo de entrar nos módulos de cada motor, ocorrem duas ramificações. A primeira é responsável pelo

controle de temperatura do retorno de combustível, já a segunda pelo resfriamento dos bicos injetores (Sistema de Arrefecimento dos Bicos Injetores – SAI). Em seguida, essas correntes voltam à tubulação principal onde posteriormente são ramificadas para as UGDs. Em cada módulo, é utilizada inicialmente no resfriamento do segundo estágio dos *intercoolers*, posteriormente, para o resfriamento do óleo lubrificante, e em seguida, retorna para a tubulação principal, com destino aos radiadores, finalizando o ciclo.

Figura 19 – Diagrama do Sistema de Água de Arrefecimento de Baixa Temperatura – SBT.



Fonte: elaborado pelo autor (2023).

Tabela 3 – Principais sensores de cada UGD relativos ao SAB.

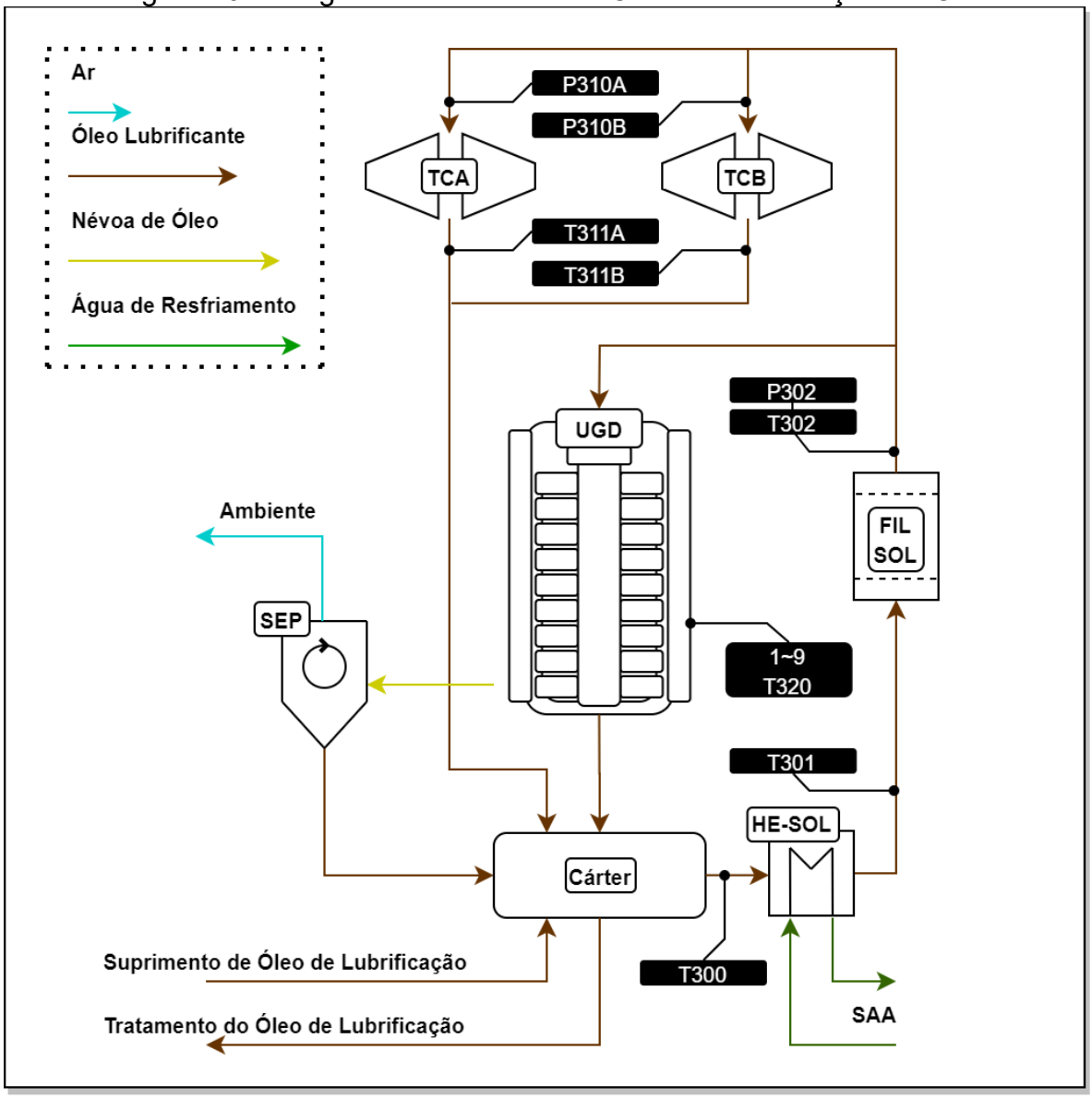
Sensor	Descrição
T210	Temperatura da água de arrefecimento de baixa temperatura na saída do RAD
P211	Pressão da água de arrefecimento de baixa temperatura na entrada do HEA e HEB
T211	Temperatura da água de arrefecimento de baixa temperatura na entrada do HEA e HEB
T212	Temperatura da água de arrefecimento de baixa temperatura na saída do HE-SOL
T213	Temperatura da água de arrefecimento de baixa temperatura na entrada do RAD
P220	Pressão da água de arrefecimento de baixa temperatura no SAI
T220	Temperatura da água de arrefecimento de baixa temperatura no SAI

Fonte: elaborado pelo autor (2023).

3.1.3 Sistema de Óleo de Lubrificação (SOL)

A lubrificação é responsável pela redução do atrito de componentes móveis das UGDs e dos turbocompressores, evitando desgastes excessivos e consequentes superaquecimentos. Armazenado no cárter, o óleo lubrificante é resfriado pelo trocador de calor HE-SOL e filtrado pelo FIL-SOL antes de ser distribuído pelo maquinário. Após o processo de filtração, ocorre a ramificação do óleo lubrificante, onde uma fração é destinada aos turbocompressores TCA e TCB, e a parte majoritária para o próprio motor.

Figura 20 – Diagrama do Sistema de Óleo de Lubrificação – SOL.



Fonte: elaborado pelo autor (2023).

Tabela 4 – Principais sensores de cada UGD relativos ao SOL.

Sensor	Descrição
T300	Temperatura do óleo de lubrificação na entrada do HE-SOL
T301	Temperatura do óleo de lubrificação na saída do HE-SOL
P302	Pressão do óleo de lubrificação na entrada da UGD
T302	Temperatura do óleo de lubrificação na entrada da UGD
P310 A/B	Pressão do óleo de lubrificação na entrada do TCA/B
T311 A/B	Temperatura do óleo de lubrificação na saída do TCA/B
1~9 T221	Temperatura do óleo de lubrificação nos mancais móveis 1~9

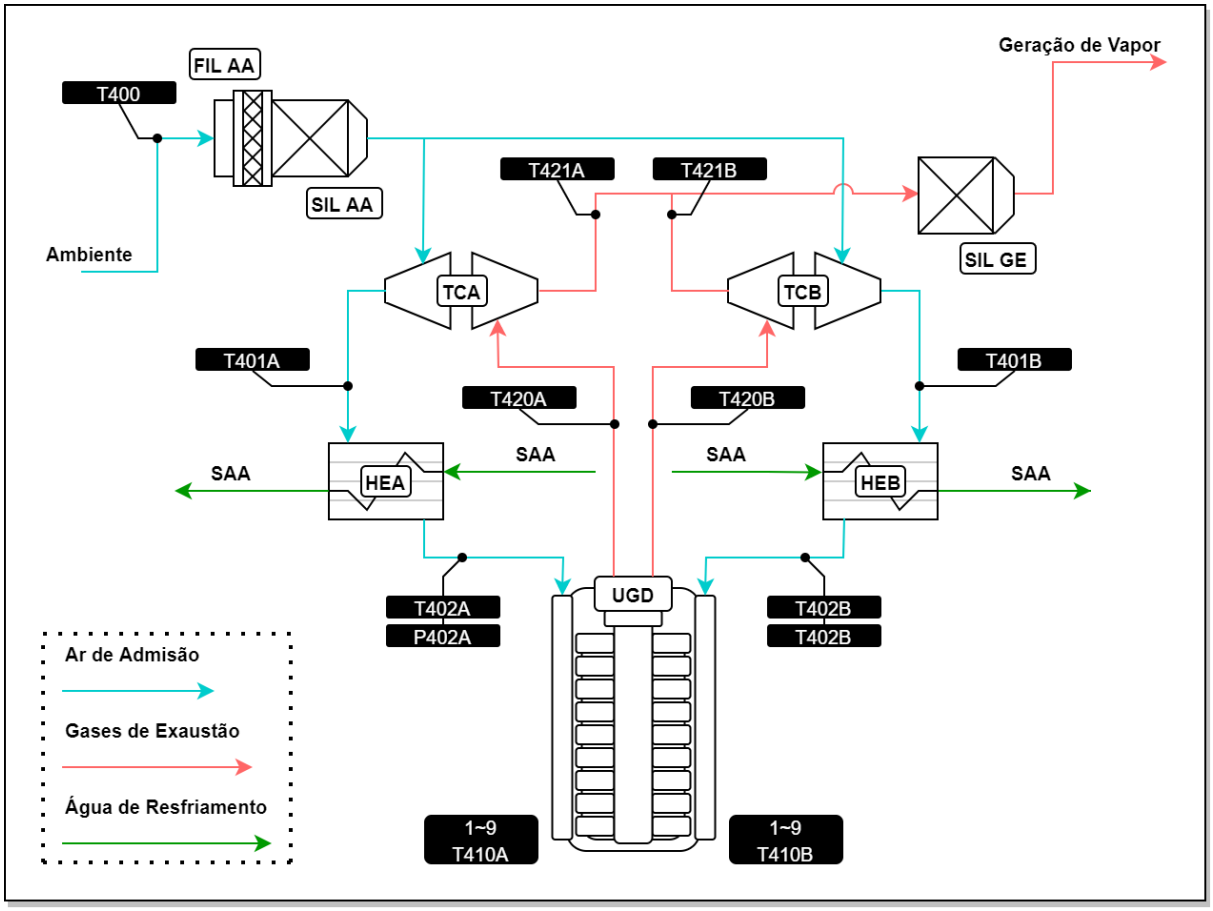
Fonte: elaborado pelo autor (2023).

Os turbocompressores são lubrificados na conexão de suas seções frias (admissão) com suas seções quentes (exaustão), região na qual o eixo é sustentado por mancais, e, consequentemente, onde o atrito e desgaste são mais evidenciados. A lubrificação é continua, e sua saída direcionada ao cárter. No interior dos motores geradores, a lubrificação tem como foco os pistões e o virabrequim. Os pistões necessitam a redução de atrito na região de contato com a câmara de combustão, já o virabrequim conta com eixos fixos (extremidades do motor) e eixos móveis (bielas) que necessitam lubrificação. Como nos turbocompressores, a lubrificação é continua e possui retorno direcionado ao cárter. Contudo, como o óleo é aquecido no interior dos motores, parte deste é vaporizado formando a denominada Névoa de Óleo, aumentando a pressão interna. Para alívio da pressão, esta névoa é direcionada a um filtro (Separador de Névoa - SEP), o qual remove o óleo lubrificante do ar, e o retorna ao cárter.

3.1.4 Sistema de Admissão de Ar e Exaustão de Gases (SAE)

O Sistema de Admissão de Ar e de Exaustão de Gases (SAE) é combinado devido ao compartilhamento de um equipamento fundamental para o funcionamento eficiente das UGDs: o turbocompressor. Antes de entrar nas câmaras de combustão, o ar de admissão proveniente do ambiente é direcionado para um filtro (FIL AA) e um silenciador (SIL AA), que buscam remover impurezas e diminuir os ruídos, respectivamente. Nas UGDs do tipo 18V ocorre uma ramificação da corrente, uma vez que há duas fileiras de 9 cilindros nos motores. Consequentemente, há dois turbocompressores (TCA e TCB). Nas UGDs do tipo 9L, essa ramificação não ocorre, pois estes motores são lineares e contam apenas com o turbocompressor TCA.

Figura 21 – Diagrama do Sistema de Admissão de Ar e de Exaustão de Gases – SAE.



Fonte: elaborado pelo autor (2023).

Tabela 5 – Principais sensores de cada UGD relativos ao SAE.

Sensor	Descrição
T400	Temperatura do ar ambiente
T401 A/B	Temperatura do ar de admissão na saída do TCA/B
P402 A/B	Pressão do ar de admissão na saída do HEA/B
T402 A/B	Temperatura do ar de admissão na saída do HEA/B
1~9 T410 A/B	Temperatura dos gases de exaustão na saída do cilindro 1~9 A/B
T420 A/B	Temperatura dos gases de exaustão na entrada do TCA/B
T421 A/B	Temperatura dos gases de exaustão na saída do TCA/B

Fonte: elaborado pelo autor (2023).

Nos turbocompressores, o ar de admissão é comprimido aumentando, assim, a densidade molar deste gás. Esta compressão permite elevar a eficiência volumétrica dos motores, ou seja, para um mesmo tamanho das câmaras, mais combustível pode ser injetado para queima. Devido à compressão do ar, ocorre consequente aumento de temperatura deste, o qual afeta tanto a eficiência volumétrica, quanto a própria eficiência térmica dos motores. Para tanto, antes de ser admitido pelos motores ainda

existem trocadores de calor, denominados de *intercoolers* (HEA e HEB), responsáveis pelo resfriamento do ar.

A corrente de ar é distribuída pelas câmaras, onde juntamente com o combustível, ocorre a explosão e geração dos gases de exaustão. Estes gases de exaustão são direcionados aos turbocompressores, onde parte de sua energia térmica e cinética é cedida para a compressão do ar de admissão. Posteriormente, as saídas dos gases das turbinas são conectadas e passam por um único silenciador (SIL GE).

De forma geral, este é o ciclo completo do SAE. Todavia, algumas unidades geradoras são acopladas a caldeiras de recuperação para geração de vapor. O principal uso de vapor das UTEs está ligado ao aquecimento do HFO. Por fim, cada termelétrica possui uma única chaminé, onde os gases de exaustão de cada unidade geradora são expelidos em conjunto.

3.2 OS DADOS

Para estudo e criação de modelos, duas bases de dados da EPASA foram utilizadas: (i) os dados históricos dos sensores e (ii) as notas de manutenção. A primeira base de dados contém todos os sensores da planta dispostos em séries temporais, pelos quais são registradas variáveis como temperaturas e pressões de operação. Por outro lado, as notas de manutenção apresentam informações das condições nas quais os maquinários se encontravam, como falhas e vazamentos diversos, revisões gerais e até mesmo possíveis erros de sensoriamento.

3.2.1 Variáveis Sensoriadas

Os sensores são responsáveis pela medição direta ou indireta de variáveis do processo em determinados pontos, armazenando seus valores em uma base de dados histórica. O sensoriamento das UGDs é idêntico, com raras exceções. Deste modo, cada unidade geradora conta com pelo menos 88 variáveis de processo comuns, com exceção dos motores lineares (UGD20 e UGD40), que por não apresentarem dois lados distintos possuem uma quantidade menor de sensores, que no caso são 71.

Durante o decorrer das análises, duas extrações de dados foram realizadas. A primeira, no início do projeto compreendeu dados de primeiro de janeiro de 2018 até final de 2019 (31 de dezembro de 2019). Apesar de todas as variáveis possuírem a mesma frequência de registro de 1 s, o instante destes registros difere entre os sensores. Portanto, uma reamostragem dos valores foi necessária, para que cada instante de tempo definido possuísse registro de todos os sensores

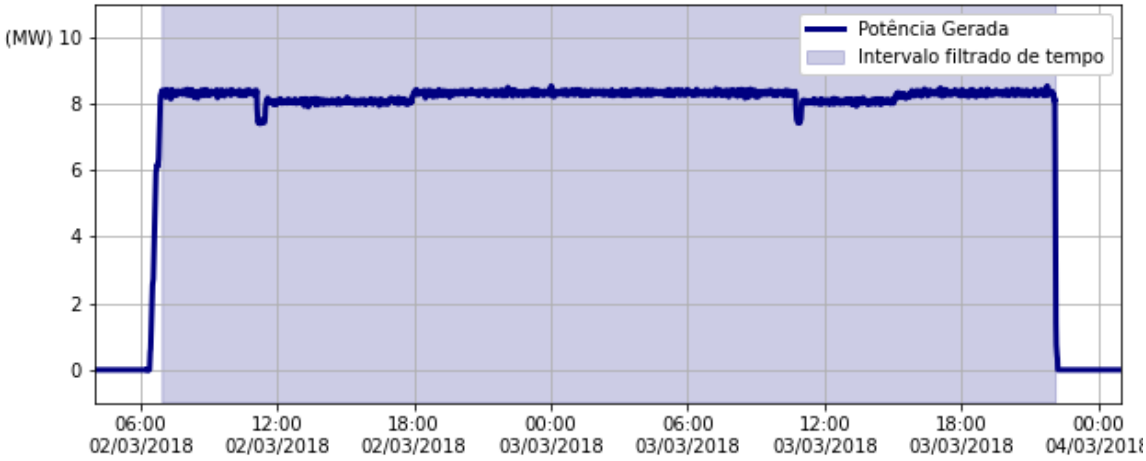
Na primeira extração, foi efetuada uma reamostragem dos dados para uma frequência de 1 min, utilizando a média dos valores compreendidos neste período. Gerou-se um total de 1.049.760 registros, contendo 3.486 variáveis de processo.

A segunda extração englobou dados mais recentes, estendendo o intervalo de análise para 26 de fevereiro de 2021, com o intuito de refinar análises e modelos iniciais já existentes. Portanto, já havia uma expertise sobre os processos e variáveis. A principal mudança (exceto o intervalo temporal) foi a frequência com a qual os dados foram reamostrados, passando para um período de 10 min. O motivo para tanto decorre do fato de que os desvios de tendências e sinais de anomalias poderiam ocorrer em intervalos de tempo superiores a dias de operação. Assim, uma frequência alta apenas acarretava mais tempo de processamento e não agregava informações relevantes. Nesta última extração, gerou-se um total de 165.888 registros com as mesmas 3.486 variáveis de processo. Vale comentar que não houve interpolação ou extrapolação de dados.

Os dados ainda brutos (apenas com reamostragem) foram pré-processados utilizando basicamente dois filtros. O primeiro filtro removia intervalos de tempo em que havia subidas e descidas de máquinas. Estes intervalos apresentam um regime altamente transiente e muito efêmero se comparado à própria geração.

Para tanto, uma abordagem simples, mas eficiente, contava com uma *flag* (variável de status) historiada, que sinalizava o início e o fim do despacho. A Figura 22 representa a remoção destes momentos e passa uma ideia geral de como é o comportamento das unidades geradoras.

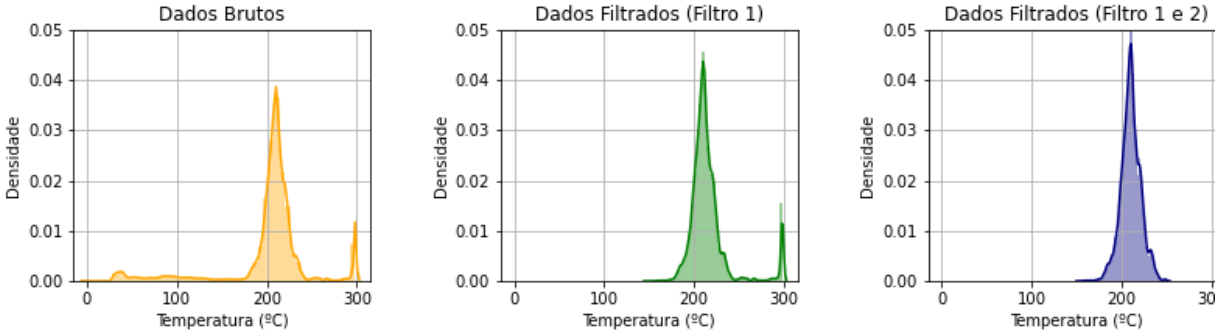
Figura 22 – Exemplo de despacho. Série temporal da potência da UGD 38 no início de março de 2018.



Fonte: elaborado pelo autor (2023).

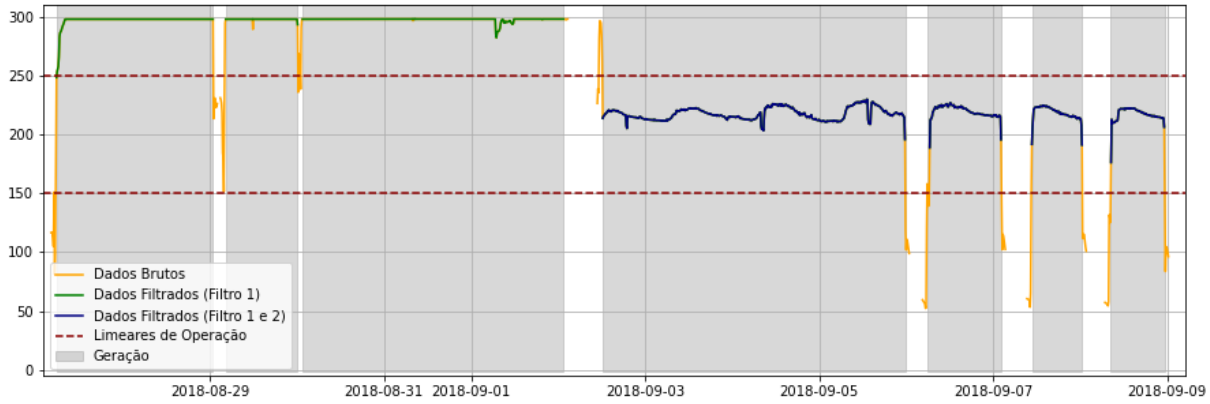
Posteriormente, um segundo filtro foi aplicado, buscando remover variáveis que se encontravam fora dos patamares de operação. Isto se fez necessário dado que vários sensores, devido a falhas, apresentavam valores fisicamente incoerentes. Este filtro foi implementado com base nos alarmes definidos pela operação e limiares. A Figura 23 ilustra a distribuição de uma mesma variável ao passar pelos dois filtros comentados. Por sua vez, a Figura 24 representa a mesma variável, submetida aos mesmos filtros, porém disposta em série temporal, com enfoque no período em que ocorreu fuga dos patamares de operação.

Figura 23 – Comparação entre as distribuições da temperatura do ar de admissão da UGD 31, antes e após a aplicação dos filtros.



Fonte: elaborado pelo autor (2023).

Figura 24 – Comparação entre as séries temporais da temperatura do ar de admissão da UGD 31, antes e após a aplicação dos filtros.



Fonte: elaborado pelo autor (2023).

3.2.1 Notas de Manutenção

A segunda base de dados analisada compreendia as notas de manutenção. Esta base de dados apresentava em suas colunas informações como: a descrição da manutenção, o tipo de manutenção (corretiva ou preventiva), o equipamento referido, e as datas de início e fim. Cada linha desta base representava um único registro.

Contudo, não havia uma padronização destes registros. A descrição é um campo aberto, onde o mesmo tipo de operação era registrado de formas diferentes pelos operados, contendo inclusive erros de gramática. O campo relativo ao equipamento de referência poderia contar com um item mais específico, como um trocador de calor, ou citar simplesmente a unidade geradora. Até mesmo o tipo da manutenção registrada muitas vezes não era coerente, sendo que trocas de filtro, descritas da mesma maneira, poderiam ser consideradas hora como manutenções corretivas, hora como manutenções preventivas.

Além disto, muitas atividades realizadas nas UTEs também eram registradas no mesmo banco de dados, de forma que havia registros de manutenções de ar-condicionados, banheiros, iluminação, entre outras, que para os propósitos aqui definidos são irrelevantes.

Contando com 41.921 registros desde o início de 2018, houve um exaustivo trabalho de filtragem e padronização destes dados. Inicialmente foram selecionadas as manutenções que de alguma forma poderiam ser indexadas às unidades

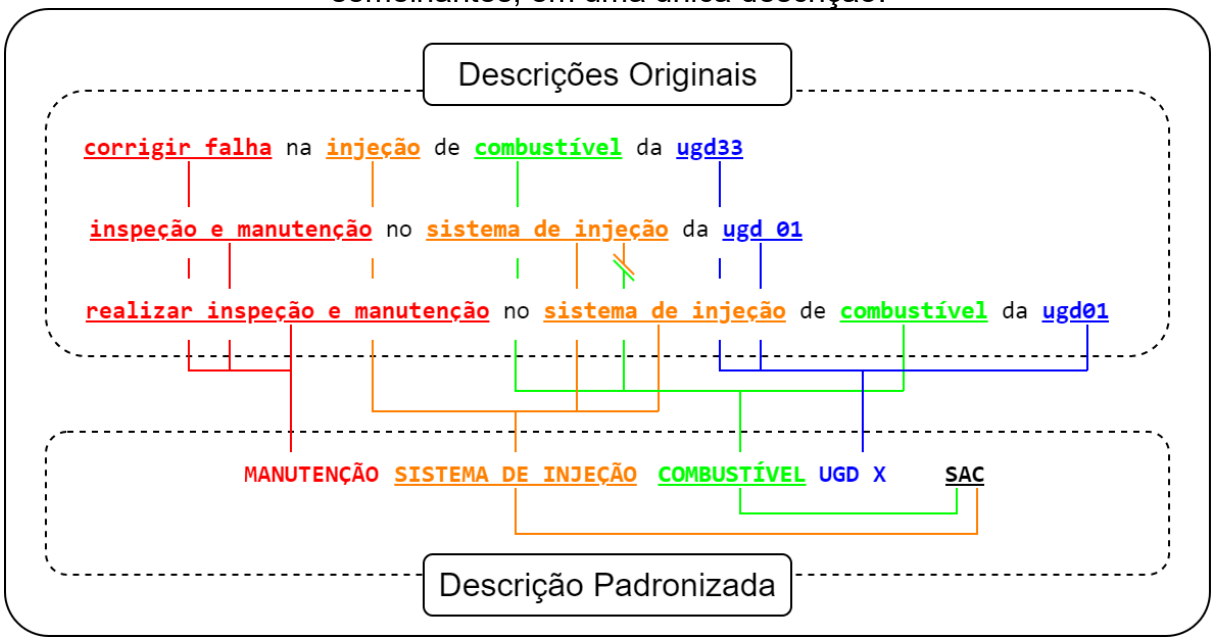
geradoras, seja pela descrição ou pelo equipamento. Apenas este filtro reduziu a quantidade de registros para 15.143.

Posteriormente, utilizando palavras-chave, alguns filtros mais refinados foram aplicados, descartando atividades que fugissem do escopo, como, por exemplo: manutenções menores, como reaperto de parafusos e instalações de braçadeiras; atividades relacionadas à estrutura, como instalação e/ou remoção de andaimes; manutenções em sistemas secundários, como ar de serviço, entre outras. Reduziu-se, assim, o número total de manutenções para 8.340.

Em seguida, foi possível padronizar as informações presentes nas descrições e o índice do equipamento. Para tanto, erros de gramática e/ou digitação foram corrigidos, assim como sinônimos foram mesclados.

As informações presentes nas descrições foram fragmentadas em alguns subitens, como unidade geradora, equipamento principal, fluído e subsistema. Não necessariamente todos os itens estariam presentes em todas as descrições, porém nem todos se faziam necessários, e em alguns casos, por associação, era possível definir itens faltantes. A Figura 25 ilustra um exemplo do processo de padronização destes itens, de forma que três manutenções semelhantes podem ser agrupadas em uma única descrição.

Figura 25 – Exemplo do processo de padronização de três notas de manutenção semelhantes, em uma única descrição.



Fonte: elaborado pelo autor (2023).

Só então, com as descrições padronizadas e filtradas, foi possível definir, juntamente com especialistas da própria EPASA, as ocorrências que de fato se apresentavam anômalas e poderiam ser detectadas pelo sensoriamento das instalações. Totalizando 2.497 ocorrências, essas puderam ser agrupadas de acordo com os subsistemas em pelo menos dois grandes grupos: manutenções e vazamentos.

As notas de manutenção possuem duas principais funções para a criação dos modelos. A primeira é justamente poder definir quando e quais ocorrências foram detectadas pelos operadores, permitindo a avaliação do desempenho dos modelos. A segunda função é poder definir em quais intervalos de tempo as unidades geradoras possuíam melhor funcionamento, para que nestes intervalos os modelos pudessem ser treinados e os Gêmeos Digitais estabelecidos, de forma a refletir um estado ótimo de operação.

Vale comentar que o processo de seleção e padronização das notas de manutenção não foi linear e, com a crescente expertise do processo, muito retrabalho foi efetuado, de forma a refinar as ocorrências mais relevantes.

3.3 OS MODELOS

Foram adotadas duas abordagens complementares para cada subsistema das unidades geradoras, utilizando modelos de regressão e classificação. Modelos de aprendizado de máquina regressivos foram empregados para simular as principais variáveis de cada subsistema nas melhores condições de operação, resultando nos Gêmeos Digitais. Através da comparação entre resultados das simulações e as variáveis reais, é possível detectar desvios operacionais e avaliar o desempenho de cada subsistema. Com o intuito de obter um índice que reflita o desempenho desses sistemas, foram propostas transformações dos erros dos modelos. Em seguida, utilizando janelas temporais, avaliou-se a capacidade de detecção de anomalias com base nesse índice. Em outras palavras, o erro dos modelos regressivos é utilizado para uma classificação binária, sendo importante ressaltar que nesse contexto não há modelos de aprendizado de máquina de classificação, mas sim uma comparação entre grandes desvios dos modelos de regressão e anomalias.

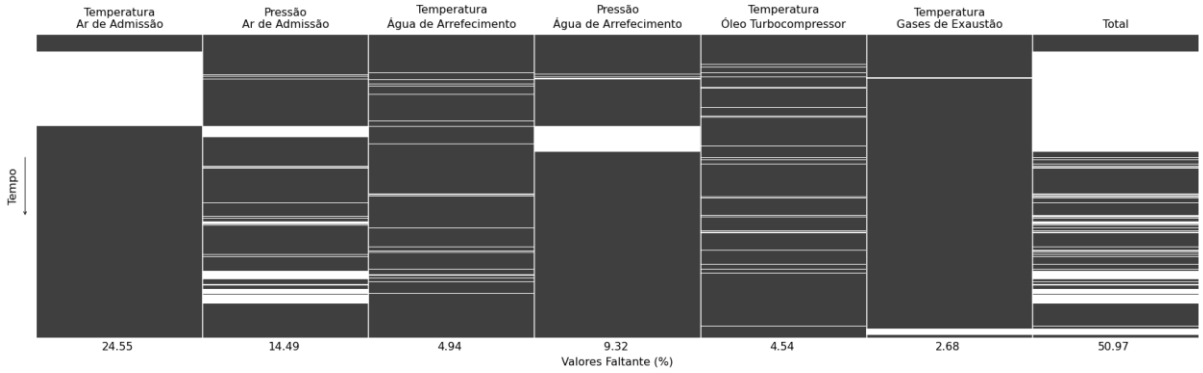
Por outro lado, modelos de aprendizado de máquina de classificação foram utilizados para a classificação não binária das falhas e geração de alarmes. Os

modelos regressivos foram utilizados como um filtro nessas classificações. Em outras palavras, os modelos de classificação podem utilizar a detecção de anomalias dos Gêmeos Digitais como suporte para classificar a falha mais provável.

3.3.1 Treinamento

O pré-tratamento dos dados é essencial para análises mais assertivas, contudo este tratamento acarreta na redução da quantidade dados. A média dos valores removidos corresponde a 2,51%. Apesar desse valor não ser expressivo, algumas variáveis apresentaram uma redução de mais da metade de seus registros. Além disto, ao concatenar as variáveis, a quantidade de valores removidos sofre um rápido crescimento, uma vez que para um intervalo de tempo ser válido para as análises, todas as variáveis precisam conter registros. A Figura 26 ilustra a quantidade de valores faltantes de seis variáveis e a combinação destas, sendo possível perceber que, apesar de poucas variáveis analisadas, ocorre uma redução de mais da metade da quantidade de registros.

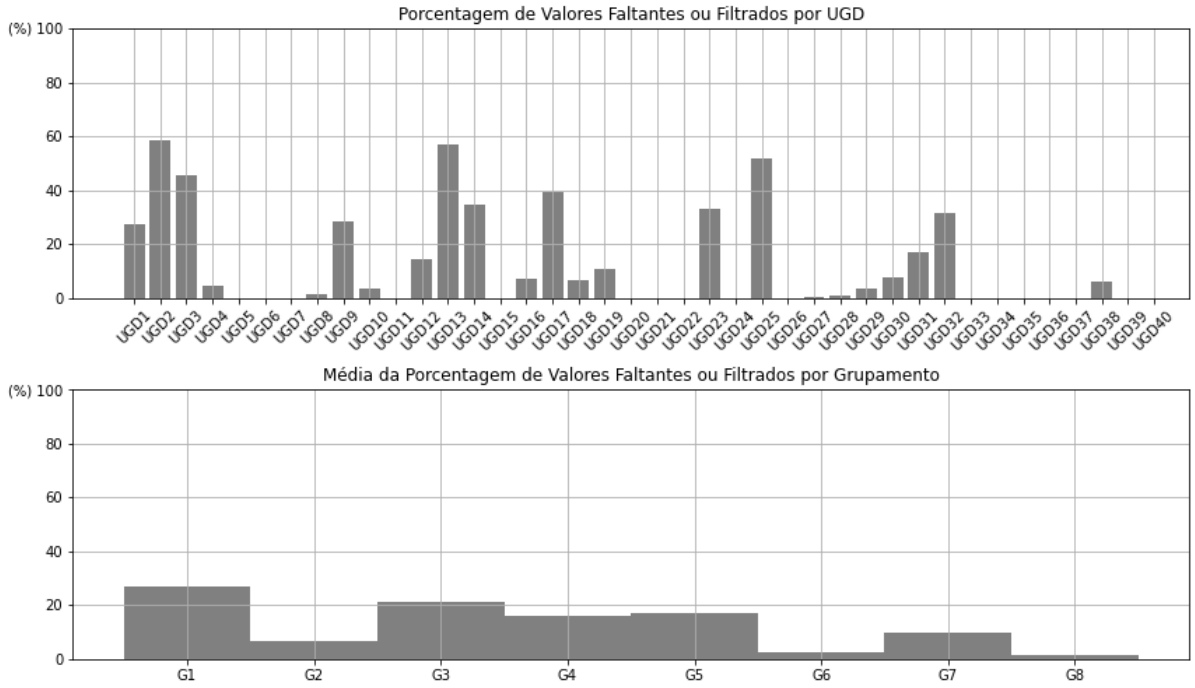
Figura 26 – Exemplo do crescente número de registros inválidos ao combinar seis variáveis de processo. Cada coluna representa uma variável, sendo que a última representa a combinação destes registros. Em cinza são representados os valores úteis e, em branco, os valores faltantes.



Fonte: elaborado pelo autor (2023).

Com isto, analisar todas as unidades geradoras com uma única abordagem (mesma seleção de variáveis) se tornou inviável, uma vez que a disponibilidade de uma única variável poderia comprometer a análise de uma ou mais unidades geradoras, como é ilustrado pela Figura 27.

Figura 27 – Valores faltantes por unidade geradora e por grupamento.



Fonte: elaborado pelo autor (2023).

Por outro lado, analisar as UGDs de forma individual acarretaria em modelos pouco generalistas, uma vez que, diferentemente de uma abordagem convencional, onde os dados de treino são escolhidos de maneira aleatória e podem compreender até 80% da totalidade dos dados, para o treinamento dos Gêmeos Digitais apenas intervalos "saudáveis" (momentos sem manutenções agendadas) foram selecionados. Além de tornar fixo o intervalo de treino, esse procedimento também o tornou escasso, correspondendo, em média, a apenas 32,39% da totalidade dos dados para cada unidade geradora.

Para tanto, avaliar os grupamentos de 5 unidades geradoras se apresentou a opção mais coerente e viável, uma vez que a seleção de variáveis seria mais flexível, além de permitir uma maior disponibilidade de dados de treino e possibilitar a validação cruzada dos modelos. Em outras palavras, cada modelo de regressão foi treinado com os dados “saudáveis” das 5 UGDs e testado com os demais dados, validando os resultados entre as unidades geradoras. Os grupamentos são os mesmos para todos os subsistemas e foram definidos de acordo com subsistemas que possuíam UGDs compartilhadas (SAC e SAA).

Entretanto, havia duas exceções: as unidades geradoras lineares (UGD20 e UGD40), que não pertenciam ao mesmo grupo, mas devido as suas características

semelhantes foram treinadas e validadas em conjunto. Sendo assim, os grupamentos 1, 2, 3, 5, 6, e 7 possuíam 5 UGDs, os grupamentos 4 e 8 possuíam 4 UGDs, e um novo grupamento, denominado de "L", foi criado, possuindo apenas 2 UGDs.

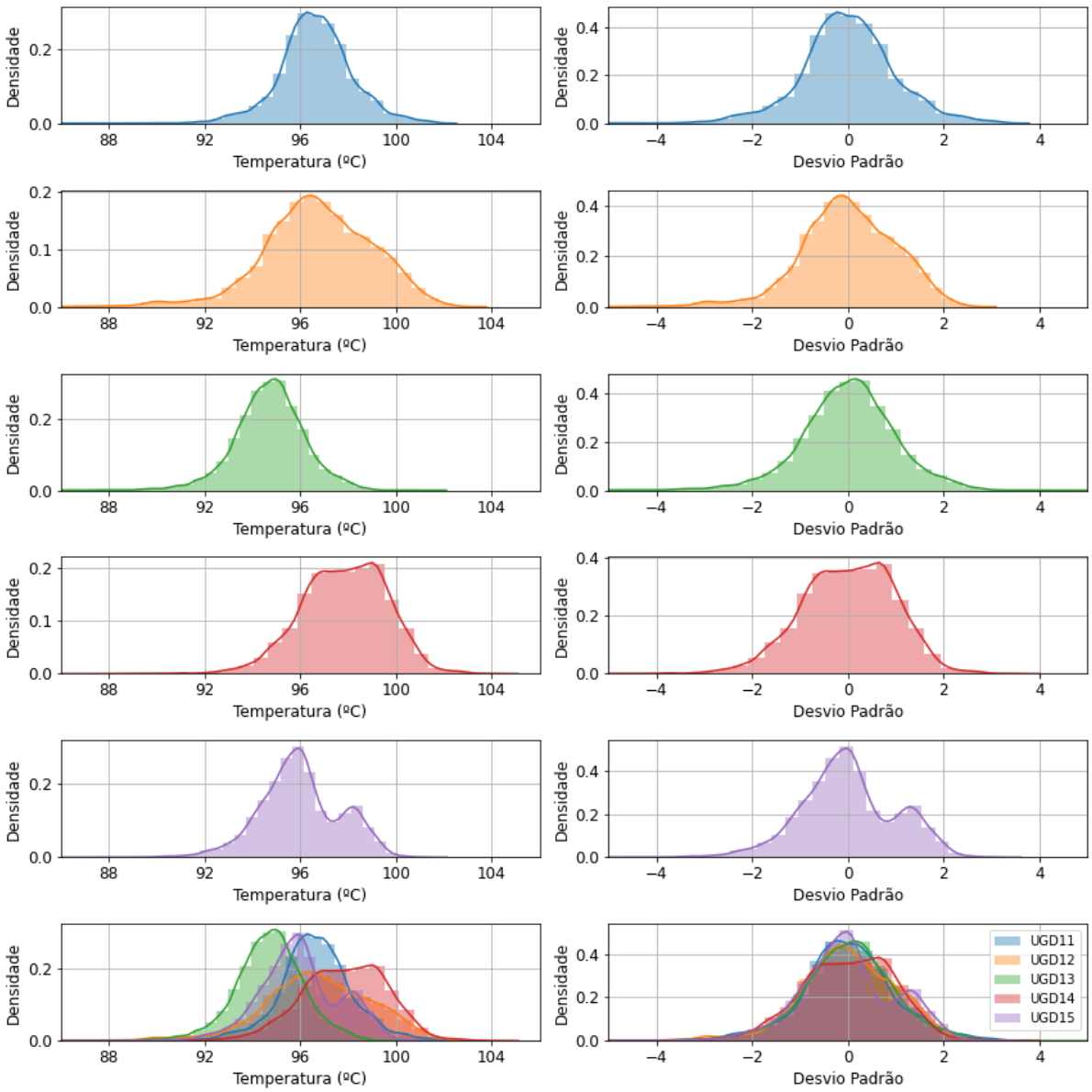
Ainda antes do treinamento, mais um processamento dos dados foi efetuado, a estandardização. A estandardização dos dados busca transformar os valores de uma variável, de modo que a média da variável corresponda a zero, e cada desvio padrão desta corresponda a um valor unitário. A estandardização é efetuada com base na função da Equação (26), onde x, u, s e z são, respectivamente, a variável alvo, sua média, seu desvio padrão e o valor normalizado.

$$z = \frac{(x - u)}{s} \tag{26}$$

Este processo se faz necessário por dois principais motivos. O primeiro decorre do fato que alguns algoritmos de aprendizado de máquina podem ser influenciados pela ordem de grandeza das variáveis. Assim, ao normalizar os valores, todas as variáveis se compreendem no mesmo intervalo, não havendo grande distinção entre elas.

O segundo motivo é que, ao normalizar individualmente uma mesma variável de UGDs diferentes, estas se sobrepõem, além de manter suas características originais. Como os patamares de operação são semelhantes, porém não idênticos, possuindo suas particularidades, o estado normal de algumas variáveis de processo podem ser considerados altos se comparados a outras unidades geradoras. A Figura 28 exemplifica o processo de estandardização de uma mesma variável para todas as unidades geradoras de um grupamento.

Figura 28 – Comparação entre valores reais, à esquerda, e suas respectivas estandardização, à direita, das UGDs 1, 2, 3, 4 e 5. A última linha compreende todas as UGDs analisadas.



Fonte: elaborado pelo autor (2023).

3.3.2 Gêmeos Digitais e o Índice de Saúde

Para cada UGD, foram criados quatro modelos de regressão, um para cada subsistema atrelado à unidade geradora. Estes modelos buscavam simular o comportamento das variáveis mais relevantes dos subsistemas (variáveis alvo) através da utilização das demais variáveis de processo. As variáveis alvo foram definidas juntamente com a equipe especialista da EPASA e podem ser observadas na Tabela 6.

Tabela 6 – Variáveis mais relevantes de cada subsistema, definidas como alvo dos modelos regressivos.

Sistema	Variável alvo
SAC	U100
SAA	T203
SOL	T302
SAE	P402

Fonte: elaborado pelo autor (2023).

Como já comentando, as notas de manutenção permitem definir os intervalos de tempo em que não houve falhas na operação. Com isto, é possível treinar os modelos nestes intervalos de tempo, garantindo que as simulações reflitam as melhores condições de operação.

Desta forma, grandes desvios entre os resultados das simulações e os valores reais sugerem divergência de comportamento, indicando possíveis anomalias e/ou falhas na operação. Quanto mais próximo de zero a diferença destes valores, melhor seria o desempenho do sistema avaliado, pois a simulação estaria coerente com o processo. De maneira análoga, quanto maior a diferença absoluta, pior seria o desempenho dos equipamentos.

Contudo, uma avaliação direta do erro pode se tornar um tanto quanto confusa, dado que cada modelo pode apresentar ordens de grandeza diferentes para esta métrica. Desta forma, é proposta uma transformação do erro para a padronização de seus limites.

Dado que quanto menor o erro, melhor seria o desempenho dos equipamentos, e consequentemente sua "saúde", temos que, quando o valor do erro quadrático tender a zero, a saúde deve expressar seu valor máximo, definido aqui como 1. De modo análogo, o aumento do erro quadrático representaria uma queda no desempenho e, consequentemente, na sua "saúde". Entretanto, como o erro não possui limites superiores, temos que, quando o valor do erro quadrático tender ao infinito, a saúde deve expressar seu valor mínimo, definido aqui como igual a 0. Além disto, o índice que representa a saúde dos equipamentos deve ser representado por uma função decrescente para os valores compreendidos no intervalo $[0, \infty)$.

Desta forma, é necessária uma função que atenda aos requisitos das Equações (27), (28) e (29):

$$\lim_{x \to 0} f(x) = 1 \tag{27}$$

$$\lim_{x\to\infty} f(x) = 0 \tag{28}$$

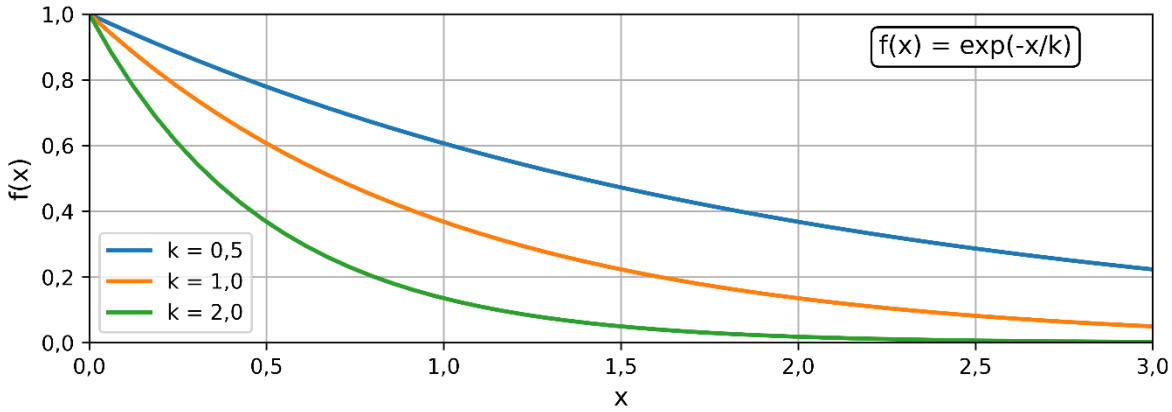
$$\frac{df(x)}{dx} < 0 \ \ \forall\, x > 0 \tag{29}$$

Uma função de decaimento exponencial representada pela Equação (30) atende a todos os requisitos listados acima e, portanto, foi utilizada para compreender o erro dentro de limites padronizados.

$$f(x) = e^{-x/k} \tag{30}$$

Na Equação (30), o número de Euler é representado pela letra e, e o erro elevado ao quadrado é representado por x. Já a constante k é uma variável de ajuste relacionada ao erro inerente do modelo. Desta forma, para modelos mais precisos, a função pode assumir um caráter mais sensível ao reduzir o valor de k, sendo a recíproca verdadeira, o que é ilustrado pela Figura 29.

Figura 29 – Exemplo do perfil da Equação (23) submetida a diferentes valores para a constante k.

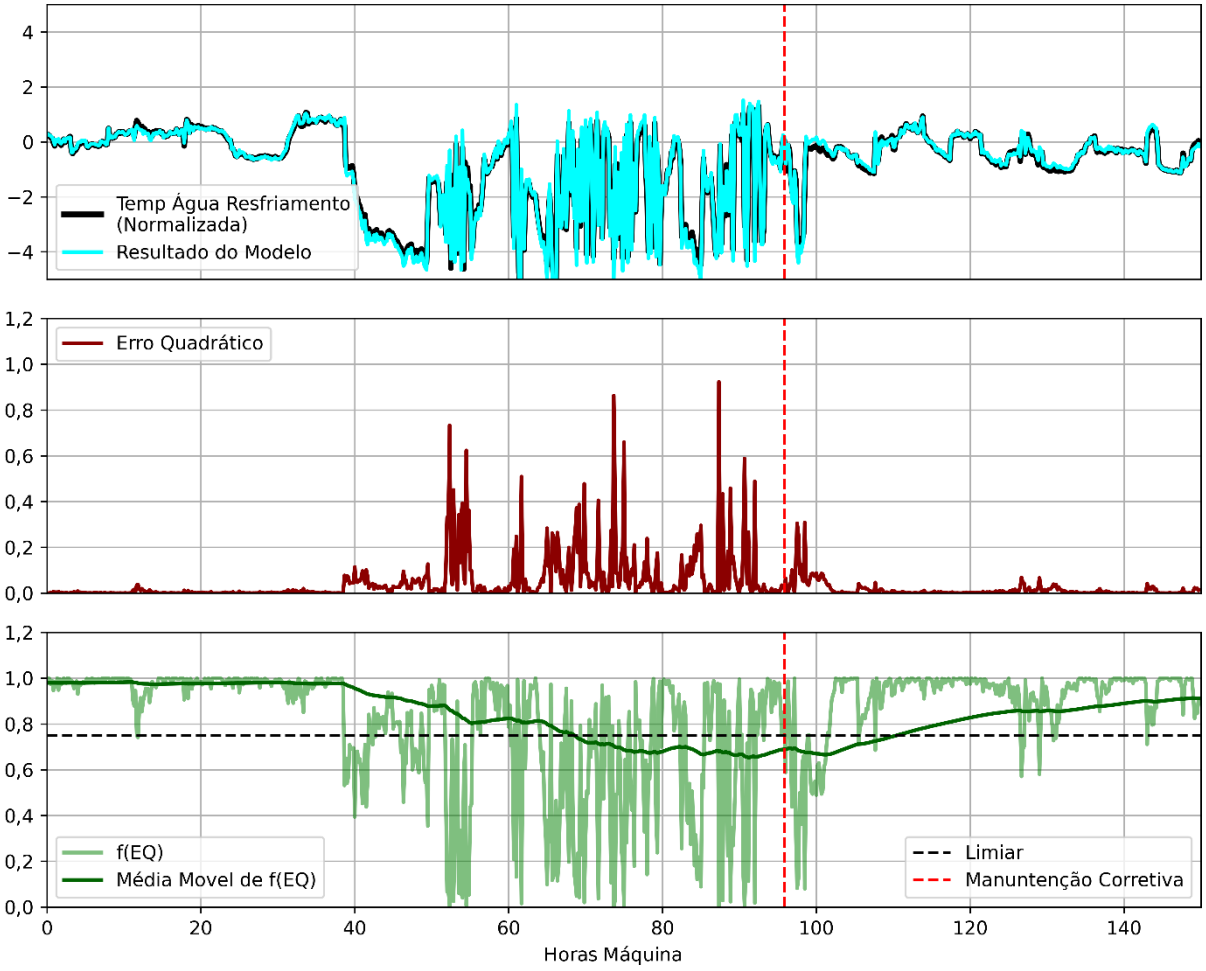


Fonte: elaborado pelo autor (2023).

Entretanto, ainda existe uma problemática. Oscilações pontuais no valor das variáveis de processo podem resultar em erros consideráveis. Como o intuito é a análise em tempo real dos equipamentos, estas oscilações acarretariam uma instabilidade não desejada. Para tanto, com o simples uso de uma média móvel seria

possível contornar tal empecilho, amenizando grandes oscilações. A Figura 30 representa todo o processo de tratamento dos resultados dos modelos, contemplando o resultado dos modelos regressivos, o erro e a transformação do erro que dá origem ao Índice de Saúde.

Figura 30 – Comparação do modelo regressivo com a variável alvo, assim como o tratamento para criação do índice de saúde.



Fonte: elaborado pelo autor (2023).

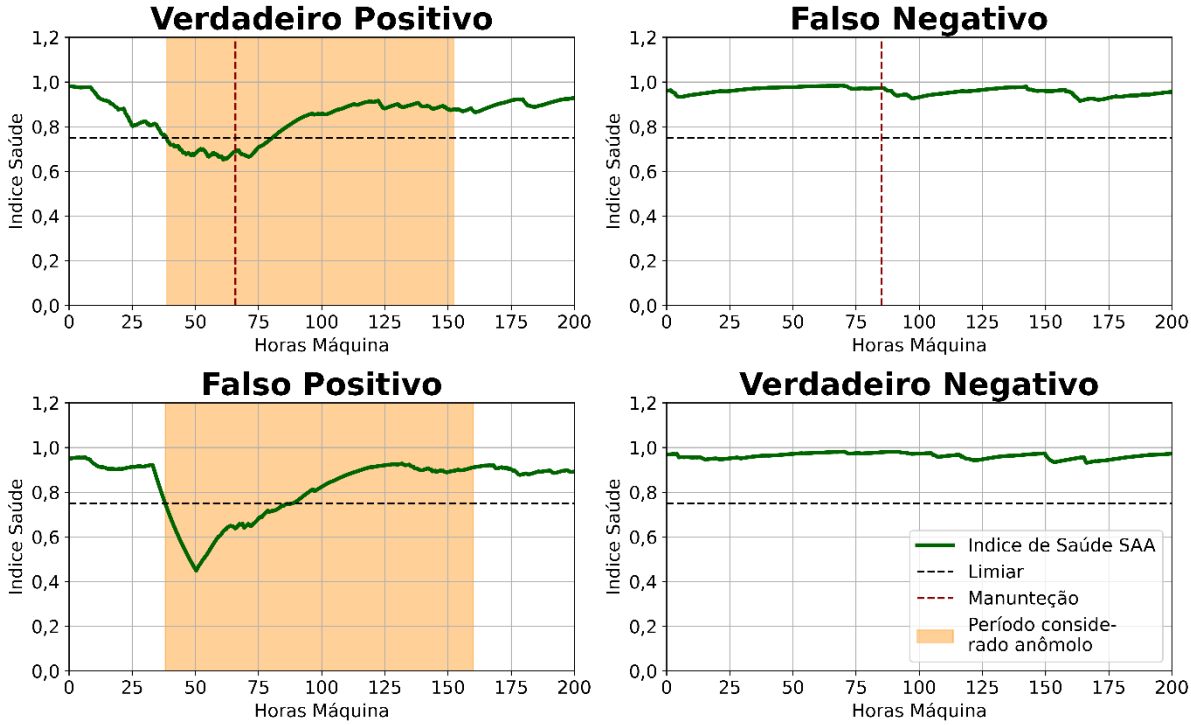
Vale comentar que todo este processo de manipulação do erro foi realizado de forma dinâmica, onde a análise dos erros e a avaliação da capacidade de classificação dos modelos sofrem diversas adaptações e otimizações para só então ser definida uma abordagem final a ser seguida.

Finalmente, após este tratamento da relação entre a variável alvo e a variável predita, foi possível utilizar uma abordagem única para avaliar o potencial de classificação dos modelos regressivos. Ou seja, se grandes desvios entre os resultados dos modelos realmente poderiam indicar anomalias. Para tanto, após a

criação dos protótipos dos Gêmeos Digitais, definiu-se um limiar que, se ultrapassado, indicaria condições anômalas dentro de uma janela temporal. Após a avaliação dos períodos de 1 a 4 dias para as janelas temporais, verificou-se que os melhores resultados foram obtidos com um intervalo fixo de 3 dias, sendo este adotado como o período padrão para as avaliações. A avaliação desses intervalos também define o tempo máximo para a predição da falha, ou seja, caso uma anomalia seja detectada, ela ocorrerá em até 3 dias.

A Figura 31 exemplifica como se dá a classificação de anomalias utilizando o Índice de Saúde. Caso o índice ultrapasse o limiar definido e ocorra alguma falha ou manutenção em até três dias, pode-se considerar que a detecção foi correta, ou seja, houve um verdadeiro positivo. De forma análoga, é possível obter verdadeiros negativos, falsos negativos e falsos positivos.

Figura 31 – Metodologia para avaliação da classificação dos modelos regressivos.



Fonte: elaborado pelo autor (2023).

Com as métricas de regressão e classificação é possível avaliar quantitativamente quais modelos apresentaram o melhor desempenho. Para tanto, um algoritmo genético foi implementado para selecionar os melhores algoritmos de regressão e seus respectivos hiperparâmetros, assim como as variáveis utilizadas como entrada dos modelos ao maximizar a função objetivo, definida como a soma do

coeficiente de correlação (R^2) com a Acurácia e o *F1-Score*. O processo de criação dos Gêmeos Digitais é ilustrado na Figura 32 de forma ampla.

Foram utilizados algoritmos regressivos do pacote *Scikit-learn* e implementados em python. Os algoritmos de aprendizado de máquina selecionados buscaram abranger diversos tipos de modelos de regressão, tais como: modelos lineares, máquinas de vetores de suporte, métodos de conjunto e redes neurais. Entretanto, em testes iniciais, o algoritmo *Support Vector Regression* (um tipo de máquina de vetores de suporte) foi descartado devido ao seu grande tempo de treinamento.

Desta forma, os algoritmos que de fato foram avaliados foram a Regressão Linear (LR) além de suas regularizações Ridge e Lasso, o *Gradient Boosting Regressor* (GRB) e o *Multilayer Perceptron* (MLP). Para o GBR, foram variados os parâmetros de taxa de aprendizagem, números de estimadores e profundidade máxima, respectivamente em (0,05, 0,10, 0,15), (50, 100, 150) e (2,3,4). Da mesma forma, para o MLP o número de camadas ocultas testadas foi de (2, 3) com 6 *perceptrons* cada, e para a ativação foram avaliadas as funções tangente hiperbólica e sigmoide.

Sobre as variáveis de entrada selecionadas, além da utilização das variáveis dos sensores algumas relações entre esses dados também foram efetuadas, com o intuito de extrair informações latentes. Pode-se citar como exemplo a diferença de temperatura entre os gases de exaustão e o ar de admissão, ou ainda, a média das temperaturas dos mancais internos das UGDs. Todas as relações utilizadas estão disponíveis na Tabela A.1 apresentada no APÊNDICE A.

A seleção das variáveis de entrada foi realizada através do algoritmo de otimização e, pelo fato de esse algoritmo ser estocástico, cada modelo possui suas próprias variáveis de entrada. As Tabelas A.2, A.3, A.4 e A,5 no APÊNDICE A apresentam todas as variáveis utilizadas por cada um dos modelos regressivos. Entretanto, vale destacar aquelas que são comuns a maioria dos modelos.

Tabela 7 – Principais sensores utilizados na criação dos modelos regressivos do SAC.

Sensores	Descrição
T402	Temperatura do ar de admissão na saída do HEA/B
P402	Pressão do ar de admissão na saída do HEA/B
T220	Temperatura da água de arrefecimento de baixa temperatura no SAI
T203	Temperatura da água de arrefecimento de alta temperatura na saída da UGD

Fonte: elaborado pelo autor (2023).

Os principais sensores utilizados para criação dos modelos regressivos são apresentados na Tabela 7. A seleção dessas variáveis para simular a potência gerada pelas UGDs condiz com o esperado, dado que a quantidade de ar injetada nos motores é função da temperatura e pressão do ar de admissão e está intrinsecamente correlacionada com a potência gerada. Além disso, as temperaturas da água de resfriamento dos bicos injetores e na saída da UGD trazem informações sobre possíveis superaquecimentos e consequente redução da eficiência dos motores.

A Tabela 8 apresenta os principais sensores utilizados na criação dos gêmeos digitais (modelos de regressão) relacionados ao SAA.

Tabela 8 – Principais sensores utilizados na criação dos modelos regressivos do SAA.

Sensor	Descrição
T201	Temperatura da água de arrefecimento de alta temperatura na entrada do HEA/HEB
T202	Temperatura da água de arrefecimento de alta temperatura na entrada da UGD
P202	Pressão da água de arrefecimento de alta temperatura na entrada da UGD
T402	Temperatura do ar de admissão na saída do HEA/B

Fonte: elaborado pelo autor (2023).

Novamente, a seleção dessas variáveis se mostrou coerente, uma vez que as condições que definem a temperatura da água de resfriamento na saída das UGDs dependem da troca térmica e das condições nas quais esse fluido se encontra inicialmente (temperaturas e pressões de entrada). Assim, ao avaliar as condições de entrada, é possível estimar as condições finais (T203). Caso essa relação não se mantenha, ou o calor cedido pela UGD sofreu alterações, ou a eficiência da troca térmica foi afetada. Ambos os casos podem ser relacionados a falhas e necessidade de manutenção.

Tabela 9 – Principais sensores utilizados na criação dos modelos regressivos do SOL.

Sensor	Descrição
T320	Temperatura do óleo de lubrificação nos mancais móveis
T311	Temperatura do óleo de lubrificação na saída do TCA/B
T203	Temperatura da água de arrefecimento de alta temperatura na saída da UGD
T201	Temperatura da água de arrefecimento de alta temperatura na entrada do HEA/HEB

Fonte: elaborado pelo autor (2023).

Na Tabela 9 se encontram os principais sensores utilizados nos modelos de regressão relacionados ao SOL. Como é avaliada a temperatura de entrada do óleo lubrificante nesse sistema, a seleção das variáveis de entrada se apresenta coesa. As temperaturas do óleo lubrificante nos mancais móveis e nos turbocompressores são consequências das propriedades do óleo, das condições nas quais esse está submetido e da sua temperatura inicial. Além disso, as temperaturas referentes ao arrefecimento trazem informações do aquecimento da unidade geradora. Dessa forma, desvios nos resultados dos modelos podem indicar, por exemplo, que o óleo lubrificante se encontra fora das especificações.

Para os modelos referentes ao SAE, as variáveis que majoritariamente foram utilizadas são apresentadas na Tabela 10.

Tabela 10 - Principais sensores utilizados na criação dos modelos regressivos do SOL.

Sensor	Descrição
T212	Temperatura da água de arrefecimento de baixa temperatura na saída do HE-SOL
T311	Temperatura do óleo de lubrificação na saída do TCA/B
T402	Temperatura do ar de admissão na saída do HEA/B
U100	Potência gerada pela UGD

Fonte: elaborado pelo autor (2023).

Como comentando anteriormente, a quantidade de ar injetada nos motores é função da temperatura e da pressão do ar de admissão e está intrinsecamente correlacionada com a potência gerada. Logo, o uso da temperatura do ar de admissão e da potência gerada para prever a pressão do ar de admissão é essencial. Por outro lado, as temperaturas do óleo de lubrificação nos turbocompressores e da água de arrefecimento podem ter sido selecionadas pela capacidade de identificação de falhas.

Sobre a função objetivo, a utilização do coeficiente de correlação tinha como objetivo quantificar a fidelidade dos modelos de regressão. Entretanto, para a

classificação se utilizou uma abordagem binária. Como somente as manutenções relacionadas ao subsistema avaliado foram consideradas, a base de dados se tornou consideravelmente desbalanceada, ou seja, a proporção de intervalos sem manutenções pertinentes era muito superior aos intervalos com manutenções. Portanto, a Acurácia tende a apresentar valores elevados.

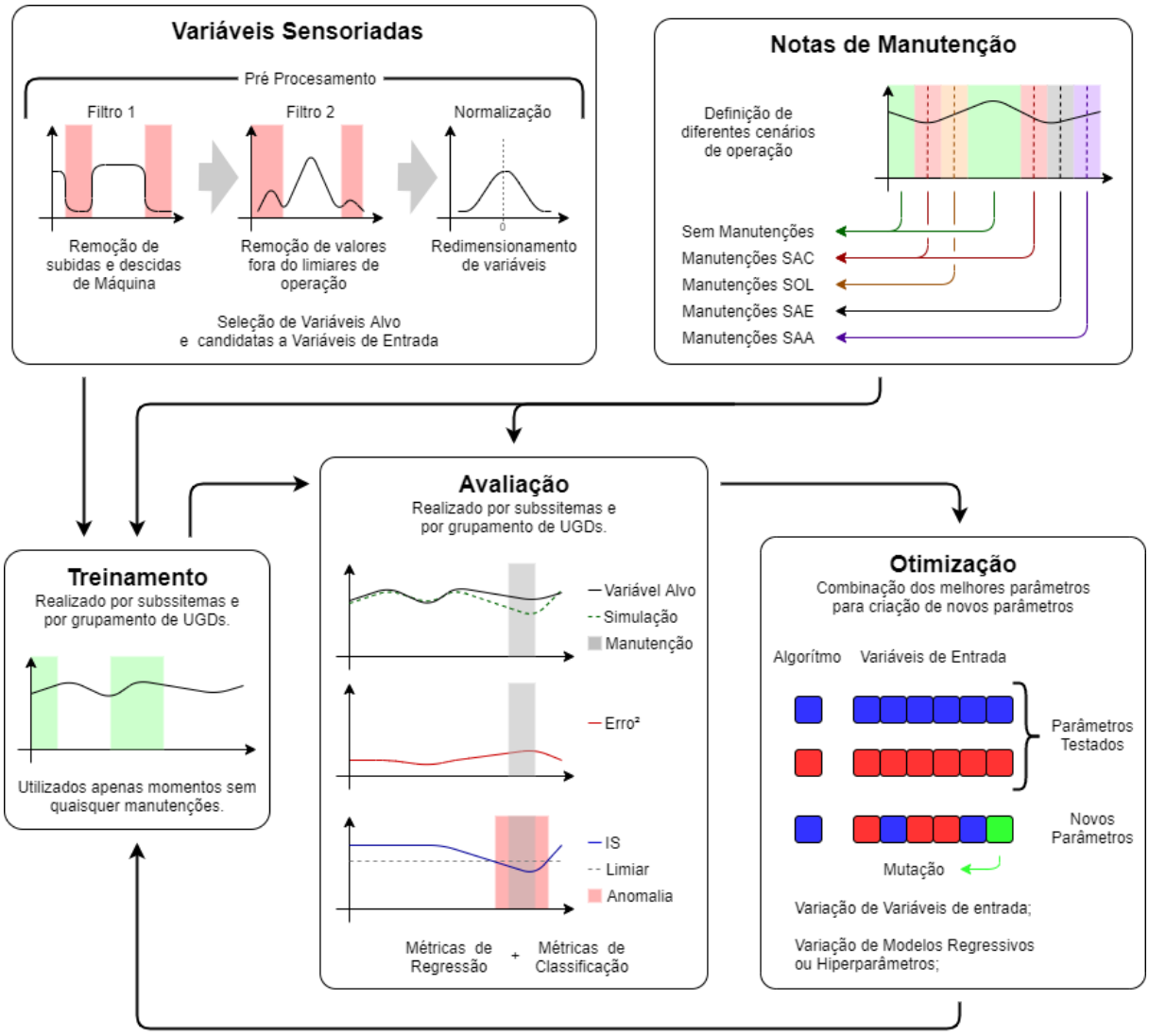
Isto é devido ao fato de a Acurácia simplesmente representar a proporção dos acertos. Por exemplo, se uma base de dados possuísse uma proporção entre classes de 9 para 1, um algoritmo que classificasse todos os dados com a classe dominante já apresentaria uma Acurácia de 0,90 ou 90%.

Por outro lado, a Sensibilidade e a Precisão são ótimas métricas para determinar a capacidade de um algoritmo identificar verdadeiros, uma vez que a Sensibilidade leva em conta falsos positivos, e a Precisão considera verdadeiros negativos. Como o *F1-Score* corresponde a uma média harmônica destas duas métricas, naturalmente também possui esta característica.

Contudo, o uso exclusivo do *F1-Score* (para dados desbalanceados) pode apresentar uma tendência de gerar falsos positivos, uma vez que a Precisão do modelo pode crescer rapidamente. Na prática, estes falsos positivos representam alarmes falsos, o que afeta diretamente a credibilidade dos resultados.

Em testes iniciais, a otimização tinha como função objetivo a soma do coeficiente de correlação com o *F1-Score*. Entretanto, pelos motivos citados acima, uma quantidade significativa de falsos positivos era gerada. Buscando reduzir estes alarmes falsos, a Acurácia também foi levada em consideração, uma vez que pondera verdadeiros positivos e negativos de forma homogênea.

Figura 32 – Fluxograma demonstrativo da metodologia aplicada para geração dos Gêmeos Digitais.



Fonte: elaborado pelo autor (2023).

3.3.3 Alarmes de Falhas

Apesar dos Gêmeos Digitais possuírem a capacidade de detecção de falhas, o sistema desenvolvido ainda conta com modelos de aprendizado de máquina para a classificação das falhas, utilizando os resultados dos Gêmeos Digitais para refinar seus resultados. Os alarmes gerados pelos classificadores são filtrados a partir dos Gêmeos Digitais.

Cada unidade geradora conta com um modelo de classificação por subsistemas. Na busca por uma classificação mais específica das falhas, o Sistema

de Admissão de Ar e Exaustão de Gases (SAE) foi segregado em dois: o Sistema de Admissão de Ar (SAd) e o Sistema de Gases de Exaustão (SEx). Além disto, um novo conjunto de falhas foi avaliado, o Sistema Mecânico, que utiliza os resultados do gêmeo digital do SAC. Há que se ressaltar, ainda, que os algoritmos de classificação utilizados não são binários, mas sim multiclasse, compreendendo diferentes falhas e/ou manutenções de um mesmo sistema.

Da mesma forma que para os modelos regressivos (Gêmeos Digitais), ocorre a remoção dos dados referentes a subidas e descidas de máquina, assim como valores fora dos intervalos de especificação, e normalização das variáveis. Contudo, para a criação dos modelos de classificação, ainda há extração de variáveis a partir da transformada de *wavelet*. Através dos coeficientes resultantes da transformada são obtidas estatísticas tais quais a entropia, quantis (0.05, 0.25, 0.75 e 0.95), média, mediana, desvio padrão, raiz quadrática média e variância para cada variável do conjunto de dados do subsistema.

Ademais, a divisão dos dados para treino e teste foi realizada de maneira distinta, através de uma divisão semi-aleatória. Momentos em que o maquinário estava operante de forma contínua foram separados, garantindo que dados de um mesmo despacho se mantivessem unidos, e destes foram selecionados 80% dos conjuntos para o treinamento dos algoritmos e 20% para teste. Esta divisão foi realizada desta maneira pois, na classificação das falhas, não se buscava recriar um estado de funcionamento ideal do maquinário, mas sim classificá-lo em estados operacionais normais ou suscetíveis a falha.

As falhas foram agregadas em classes, de acordo com a similaridade de suas descrições. Essas agregações, as principais descrições das falhas e a quantidade total de ocorrências (para todas as unidades geradoras) estão dispostas nas Tabela 11-16.

Tabela 11 – Classes de falhas, descrições e quantidade total de ocorrências do SAC.

Classe	Descrição	Qtd.
Falhas no Sistema de Injeção	Vazamentos e manutenções corretivas nas válvulas dos cilindros, bombas injetoras e bicos injetores	376
Falhas no FIL-SAC	Vazamentos e manutenções corretivas no FIL-SAC	111

Fonte: elaborado pelo autor (2023).

Tabela 12 – Classes de falhas, descrições e quantidade total de ocorrências do SAA.

Sensor	Descrição	Qtd
Falhas no SAT	Vazamentos de água de arrefecimento nas principais tubulações e conexões do SAT, além de manutenções corretivas nas válvulas moduladoras	91
Falhas no SBT	Vazamentos de água de arrefecimento nas principais tubulações e conexões do SBT, além de manutenções corretivas nas válvulas moduladoras	16
Falhas sistema SRI	Vazamento e manutenções corretivas no sistema de resfriamento dos bicos injetores	26

Fonte: elaborado pelo autor (2023).

Tabela 13 – Classes de falhas, descrições e quantidade total de ocorrências do SOL.

Sensor	Descrição	Qtd
Falhas no FIL-SOL	Vazamentos e manutenções corretivas no FIL-SOL	382
Falhas no SOL	Vazamentos e manutenções corretivas nas bombas de óleo lubrificante, tubulações e conexões adjacentes	185

Fonte: elaborado pelo autor (2023).

Tabela 14 – Classes de falhas, descrições e quantidade total de ocorrências do SAd.

Sensor	Descrição	Qtd
Vazamentos no SAd	Vazamentos do ar de admissão nos TCA/B, HEA/B e em tubulações e conexões adjacentes	64
Falhas no FIL-SAE	Manutenções corretivas no FIL-SAE e em tubulações e conexões adjacentes	45

Fonte: elaborado pelo autor (2023).

Tabela 15 – Classes de falhas, descrições e quantidade total de ocorrências do SEx.

Sensor	Descrição	Qtd
Vazamentos no SEx	Vazamentos de gases de exaustão nos TCA/B e em tubulações e conexões adjacentes	20
Vazamento de gases nos cilindros	Vazamentos de gases de exaustão nos componentes dos cilindros	29

Fonte: elaborado pelo autor (2023).

Tabela 16 – Classes de falhas, descrições e quantidade total de ocorrências do Mec.

Sensor	Descrição	Qtd
Falhas nos cabeçotes	Vazamentos e manutenções corretivas nos cabeçotes e conexões das unidades geradoras	140
Falhas nos TCA/B	Vazamentos diversos e manutenções corretivas nos turbocompressores	96
Falhas nos eixos de comando de válvulas	Vazamentos e manutenções corretivas nos eixos de comando de válvulas e nas válvulas das unidades geradoras	70

Fonte: elaborado pelo autor (2023).

Devido ao desbalanceamento entre momentos falhos e não falhos, utilizou-se uma abordagem de *undersampling* dos dados. Esta abordagem consiste basicamente na redução da quantidade de dados correspondentes à classe predominante, no caso, momentos não falhos. Para tanto, dentre esses dados, amostras percentuais foram extraídas de todo o período de registros, de tal forma que apresentem aos modelos amostras com todos os patamares de operação das variáveis que as máquinas possuíram ao longo do tempo.

Além disto, foi aplicada a técnica de *Stratified k-fold* para validação dos modelos, onde os dados das unidades geradoras foram mesclados, e não validados entre si. O motivo para esse procedimento decorre das distribuições das falhas, que normalmente eram desbalanceadas entre as UGDs, ou seja, algumas unidades geradoras continham uma quantidade mais significativa de momentos falhos que outras.

A Tabela 17 apresenta a porcentagem dos dados relacionados às classes de falhas dominantes em cada subsistema. É importante destacar que a quantidade das falhas não está diretamente relacionada à disponibilidade dos dados, uma vez que cada sistema avaliado possui diferentes variáveis de entrada e, consequentemente, diferentes volumes de dados disponíveis. Além disso, na modelagem, as classes correspondem a três dias anteriores à necessidade de manutenções, porém cada falha possui um intervalo de tempo particular até ser sanada, o que também impacta na quantidade de dados disponíveis.

Tabela 17 – Porcentagem dos dados disponíveis das classes predominantes de cada subsistema.

Subsistema	Porcentagem dos dados disponíveis das classes predominantes
SAC	8,16 %
SAA	7,25 %
SOL	18,43 %
Sad	1,97 %
Sex	10,37 %
MEc	12,10 %

Fonte: elaborado pelo autor (2023).

Nos modelos de classificação, como não é necessário considerar os intervalos de operação ideais (onde não há qualquer tipo de manutenções ou falhas) utilizados nos Gêmeos Digitais, não houve a necessidade de uma seleção específica das variáveis de entrada para cada um dos grupamentos dos subsistemas. Portanto, as

mesmas variáveis de entrada foram utilizadas nos modelos de classificação de cada um dos sistemas avaliados. A partir destas mesmas variáveis, são extraídas variáveis latentes com a transformada de *wavelet.* As Tabelas 18 a 23 apresentam as variáveis utilizadas como entradas nos modelos de classificação de falhas.

Tabela 18 – Variáveis utilizadas como entradas na classificação de falhas do SAC.

Sensor	Descrição
U100	Potência gerada pela UGD
P102	Pressão do combustível na entrada da UGD
T102	Temperatura do combustível na saída da UGD
T400	Temperatura do ar ambiente
P402 A/B	Pressão do ar de admissão na saída do HEA/B
T421 A/B	Temperatura dos gases de exaustão na saída do TCA/B

Fonte: elaborado pelo autor (2023).

Tabela 19 – Variáveis utilizadas como entradas na classificação de falhas do SAA.

Sensor	Descrição
U100	Potência gerada pela UGD
T201	Temperatura da água de arrefecimento de alta temperatura na entrada do HEA/B
P202	Pressão da água de arrefecimento de alta temperatura na entrada da UGD
T202	Temperatura da água de arrefecimento de alta temperatura na entrada da UGD
T203	Temperatura da água de arrefecimento de alta temperatura na saída da UGD
P211	Pressão da água de arrefecimento de baixa temperatura na entrada do HEA/B
T211	Temperatura da água de arrefecimento de baixa temperatura na entrada do HEA/B
T400	Temperatura do ar ambiente
P402 A/B	Pressão do ar de admissão na saída do HEA/B
T402 A/B	Temperatura do ar de admissão na saída do HEA/B

Fonte: elaborado pelo autor (2023).

Tabela 20 – Variáveis utilizadas como entradas na classificação de falhas do SOL.

Sensor	Descrição
U100	Potência gerada pela UGD
P302	Pressão do óleo de lubrificação na entrada da UGD
T302	Temperatura do óleo de lubrificação na entrada da UGD
T311 A/B	Temperatura do óleo de lubrificação na saída do TCA/B
1~9 T221	Temperatura do óleo de lubrificação nos mancais móveis 1~9
T400	Temperatura do ar ambiente

Fonte: elaborado pelo autor (2023).

Tabela 21 – Variáveis utilizadas como entradas na classificação de falhas do SAd.

Sensor	Descrição
U100	Potência gerada pela UGD
T400	Temperatura do ar ambiente
T401 A/B	Temperatura do ar de admissão na saída do TCA/B
P402 A/B	Pressão do ar de admissão na saída do HEA/B
T402 A/B	Temperatura do ar de admissão na saída do HEA/B

Fonte: elaborado pelo autor (2023).

Tabela 22 – Variáveis utilizadas como entradas na classificação de falhas do SEx.

Sensor	Descrição
U100	Potência gerada pela UGD
T400	Temperatura do ar ambiente
T420 A/B	Temperatura dos gases de exaustão na entrada do TCA/B
T421 A/B	Temperatura dos gases de exaustão na saída do TCA/B

Fonte: elaborado pelo autor (2023).

Tabela 23 – Variáveis utilizadas como entradas na classificação de falhas do Mec.

Sensor	Descrição
U100	Potência gerada pela UGD
P310 A/B	Pressão do óleo de lubrificação na entrada do TCA/B
T311 A/B	Temperatura do óleo de lubrificação na saída do TCA/B
T400	Temperatura do ar ambiente

Fonte: elaborado pelo autor (2023).

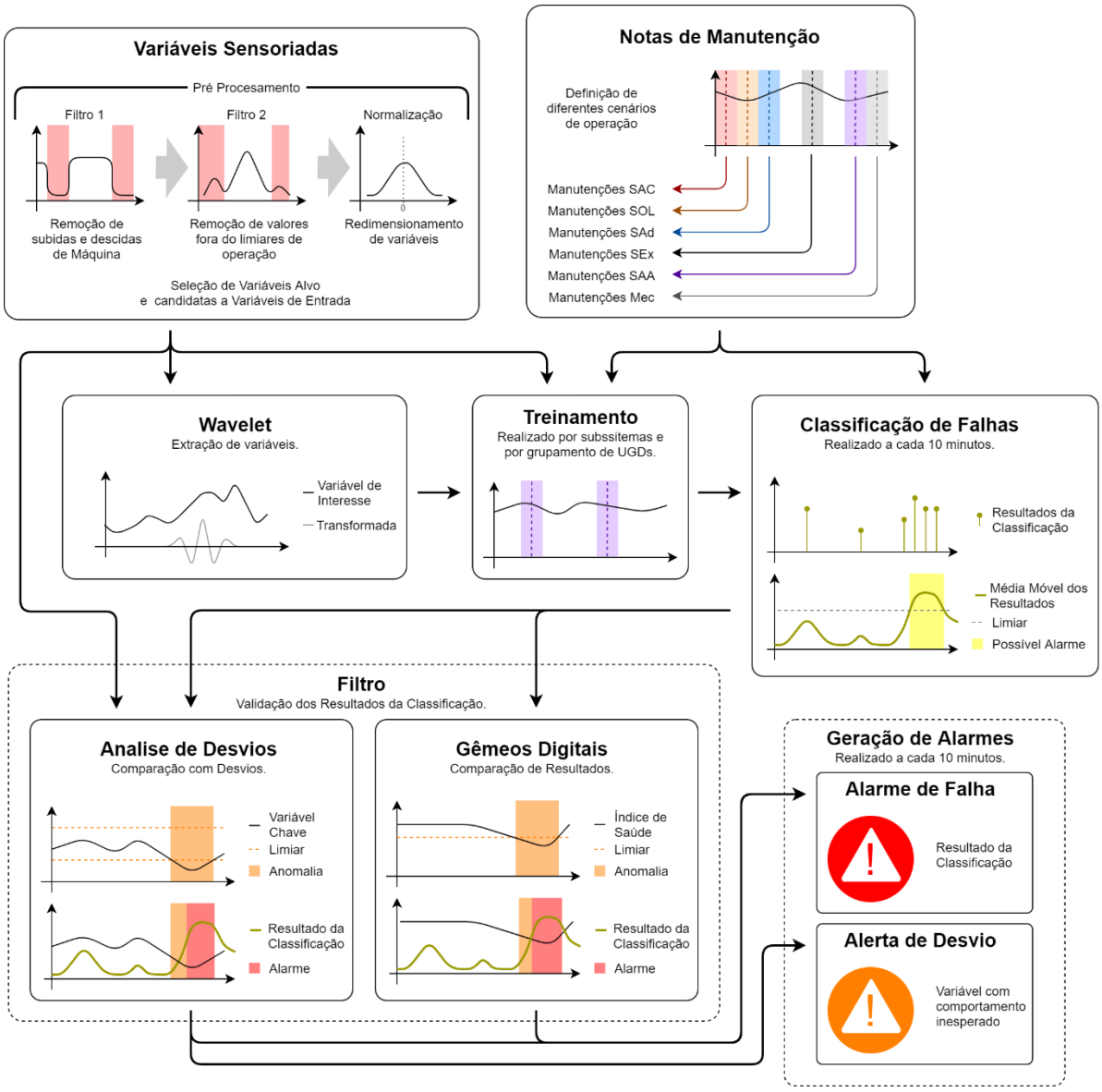
Os algoritmos de classificação empregados permitem obter a probabilidade de ocorrência de cada classe analisada. Por padrão, os algoritmos consideram limiares de 50% para a definição das classes. Contudo, ao analisar a Sensibilidade dos modelos, optou-se por adicionar um critério de severidade na criação dos alarmes. Desta forma, a severidade dos alarmes de falha é dada de acordo com a probabilidade dos modelos de classificação, sendo definidas como severidade baixa probabilidades de até 70%, e alta severidade para probabilidades superiores. Do mesmo modo que nos modelos regressivos, uma média móvel é aplicada nos resultados. Esta abordagem permite tanto o acompanhamento dos alarmes gerados, quanto um refino de falsos positivos.

Por fim, ocorre a validação dos alarmes gerados com momentos considerados atípicos. Estes momentos são resultantes de dois filtros implementados. Assim, um alarme poderá ser criado somente quando o índice de saúde do subsistema específico se encontra abaixo do limiar, ou quando variáveis chaves não se encontrarem compreendidas em até 2 desvios padrões acima ou abaixo de seu valor de referência, definido com base em análise histórica e *setpoints* de sensores. As variáveis chaves selecionadas são as mais estáveis de cada subsistema e que apresentam maior Sensibilidade a momentos registrados de falhas. Essa abordagem permitiu a redução efetiva da quantidade de alarmes falsos.

Este recurso busca criar rastreabilidade sobre os alarmes de falhas gerados, como um identificador de qual desvio de comportamento provocou o problema. Além disso, permite ao operador validar se as leituras dos sensores estão corretas, visto

que, ao longo dos dados históricos, os sensores apresentam momentos de leitura incorreta. Para tanto, também são criados alarmes referentes aos desvios das variáveis, dados pelo erro percentual desta. Todo o processo é ilustrado pelo fluxograma expresso na Figura 33.

Figura 33 – Fluxograma demonstrativo da metodologia aplicada para criação e tratamento dos resultados dos modelos de classificação – geração de alarmes.



Fonte: elaborado pelo autor (2023).

Definida a abordagem final, foram treinados diversos classificadores para cada *booster*, variando os parâmetros: *input* dos modelos, utilização de *wavelet*, tempo mínimo de dados móveis, probabilidade mínima para falha e utilização de filtro. A seleção dos modelos finais se deu pelo *F1-Score* balanceado. Vale comentar que para

elaboração da abordagem final do processo foram realizados testes com diversas abordagens, como a avaliação da divisão dos sistemas, da forma de normalização, da técnica de extração de variáveis latentes, da forma de balanceamento, da divisão dos dados para treino e teste, dos algoritmos de classificação, além da criação dos filtros.

4 RESULTADOS E DISCUSSÃO

Todos os modelos foram criados em máquinas Intel® Core (TM) i7-8565U CPU @ 1.80GHz com 16 GB de RAM com sistema operacional Windows 10 Pro. Utilizando as abordagens anteriormente mencionadas, foram criados 36 modelos regressivos utilizados como Gêmeos Digitais dos 4 subsistemas das 40 UGDs, além de 54 modelos de classificação específicos para as falhas. Totalizando assim 160 aplicações.

4.1 GÊMEOS DIGITAIS

O algoritmo genético foi utilizado para selecionar as variáveis de entrada de cada gêmeo digital, os melhores modelos, e seus respectivos hiper parâmetros. Este algoritmo tinha como parâmetros uma população de 50 indivíduos (constante desde a primeira geração), seleção de indivíduos mais adaptados por meio de "torneios" entre as 4 gerações mais atuais, mutação de 10% dos novos indivíduos, e como critério de convergência a estabilização dos resultados.

Os resultados de cada grupamento de motores (*boosters*) se encontram expressos nas Tabelas 24-27. O grupamento denominado de “L” é referente ao modelo aplicado às unidades geradoras lineares de modelo 9L32/40, ou seja, as UGDs 20 e 40.

Tabela 24 – Resultados dos Gêmeos Digitais do SAC por grupamento.

G	Alg.	R^2	RMSE	Acurácia	F1-Score
1	GBR	0,858 ± 0,045	0,120 ± 0,022	0,878 ± 0,075	0,561 ± 0,195
2	MLP	0,903 ± 0,022	0,098 ± 0,022	0,873 ± 0,100	0,657 ± 0,263
3	MLP	0,919 ± 0,043	0,098 ± 0,062	0,875 ± 0,060	0,435 ± 0,099
4	GBR	0,928 ± 0,031	0,067 ± 0,033	0,873 ± 0,075	0,529 ± 0,163
5	MLP	0,924 ± 0,019	0,085 ± 0,018	0,833 ± 0,077	0,437 ± 0,137
6	MLP	0,932 ± 0,037	0,066 ± 0,028	0,833 ± 0,077	0,437 ± 0,137
7	MLP	0,893 ± 0,104	0,131 ± 0,090	0,779 ± 0,117	0,562 ± 0,079
8	GBR	0,901 ± 0,116	0,096 ± 0,107	0,893 ± 0,066	0,644 ± 0,113
L	MLP	0,950 ± 0,029	0,045 ± 0,024	0,921 ± 0,004	0,762 ± 0,076

Fonte: elaborado pelo autor (2023).

Tabela 25 – Resultados dos Gêmeos Digitais do SAA por grupamento.

G	Alg.	R^2	RMSE	Acurácia	F1-Score
1	MLP	0,991 ± 0,034	0,012 ± 0,029	0,804 ± 0,078	0,543 ± 0,155
2	MLP	0,913 ± 0,039	0,083 ± 0,037	0,821 ± 0,102	0,564 ± 0,221
3	MLP	0,958 ± 0,031	0,040 ± 0,021	0,813 ± 0,125	0,479 ± 0,096
4	LR	0,958 ± 0,007	0,037 ± 0,005	0,817 ± 0,027	0,469 ± 0,128
5	MLP	0,866 ± 0,126	0,180 ± 0,433	0,842 ± 0,111	0,475 ± 0,069
6	MLP	0,969 ± 0,013	0,030 ± 0,010	0,860 ± 0,064	0,598 ± 0,193
7	GBR	0,957 ± 0,033	0,044 ± 0,023	0,856 ± 0,029	0,570 ± 0,163
8	MLP	0,982 ± 0,006	0,019 ± 0,004	0,860 ± 0,028	0,612 ± 0,157
L	MLP	0,968 ± 0,013	0,029 ± 0,008	0,963 ± 0,020	0,879 ± 0,089

Fonte: elaborado pelo autor (2023).

Tabela 26 – Resultados dos Gêmeos Digitais do SOL por grupamento.

G	Alg.	R^2	RMSE	Acurácia	F1-Score
1	MLP	0,901 ± 0,068	0,124 ± 0,063	0,908 ± 0,057	0,712 ± 0,206
2	MLP	0,882 ± 0,092	0,110 ± 0,069	0,856 ± 0,100	0,403 ± 0,125
3	GBR	0,917 ± 0,080	0,067 ± 0,063	0,862 ± 0,172	0,470 ± 0,128
4	MLP	0,939 ± 0,034	0,060 ± 0,030	0,894 ± 0,065	0,577 ± 0,076
5	MLP	0,946 ± 0,037	0,052 ± 0,036	0,828 ± 0,049	0,603 ± 0,088
6	MLP	0,954 ± 0,058	0,063 ± 0,046	0,808 ± 0,085	0,417 ± 0,224
7	MLP	0,927 ± 0,082	0,064 ± 0,067	0,925 ± 0,068	0,752 ± 0,134
8	MLP	0,916 ± 0,029	0,084 ± 0,027	0,888 ± 0,056	0,646 ± 0,271
L	MLP	0,923 ± 0,075	0,076 ± 0,067	0,830 ± 0,003	0,609 ± 0,116

Fonte: elaborado pelo autor (2023).

Tabela 27 – Resultados dos Gêmeos Digitais do SAE por grupamento.

G	Alg.	R^2	RMSE	Acurácia	F1-Score
1	MLP	0,872 ± 0,079	0,110 ± 0,078	0,844 ± 0,071	0,503 ± 0,160
2	MLP	0,929 ± 0,033	0,075 ± 0,031	0,843 ± 0,052	0,459 ± 0,053
3	LR	0,913 ± 0,102	0,084 ± 0,114	0,829 ± 0,093	0,423 ± 0,114
4	MLP	0,945 ± 0,055	0,056 ± 0,049	0,803 ± 0,086	0,414 ± 0,237
5	MLP	0,934 ± 0,099	0,071 ± 0,098	0,850 ± 0,063	0,400 ± 0,210
6	MLP	0,948 ± 0,056	0,041 ± 0,061	0,885 ± 0,126	0,560 ± 0,157
7	LR	0,891 ± 0,028	0,085 ± 0,018	0,901 ± 0,034	0,463 ± 0,151
8	GBR	0,929 ± 0,027	0,066 ± 0,023	0,939 ± 0,170	0,713 ± 0,197
L	MLP	0,956 ± 0,013	0,028 ± 0,006	0,862 ± 0,066	0,604 ± 0,075

Fonte: elaborado pelo autor (2023).

A Tabela 28, por sua vez, expõe os mesmos resultados de maneira mais condensada, agrupando todas as unidades geradoras de acordo com os sistemas avaliados.

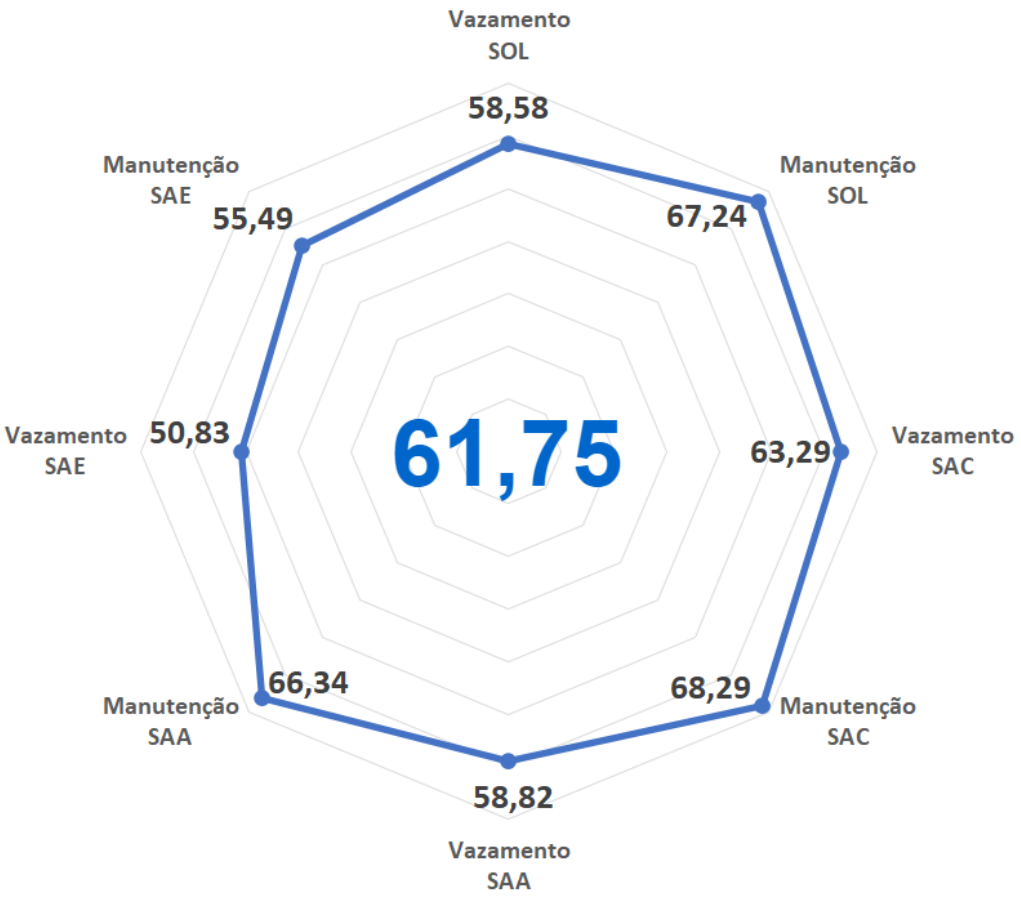
Tabela 28 – Resultados dos Gêmeos Digitais por sistema avaliado.

Sistema	R^2	RMSE	Acurácia	F1-Score
SAC	0,946 ± 0,038	0,058 ± 0,053	0,854 ± 0,048	0,581 ± 0,133
SAA	0,918 ± 0,019	0,086 ± 0,026	0,860 ± 0,044	0,558 ± 0,121
SOL	0,926 ± 0,022	0,072 ± 0,019	0,861 ± 0,039	0,560 ± 0,121
SAE	0,931 ± 0,021	0,063 ± 0,020	0,864 ± 0,043	0,505 ± 0,111

Fonte: elaborado pelo autor (2023).

Por fim, a Figura 34 apresenta os desvios operacionais (falhas e/ou manutenções) detectadas de todas as unidades geradoras nos testes dos algoritmos, sendo 61,75% a média de detecções das classificações.

Figura 34 – Desvios operacionais detectados pelos Gêmeos Digitais em porcentagem.



Fonte: elaborado pelo autor (2023).

4.2 ALARMES DE FALHAS

Assim como para os Gêmeos Digitais, os algoritmos de aprendizado de máquina testados para a classificação buscavam abranger diversos tipos de métodos, sendo estes: *Random Forest*, *Gradient Boosting*, *Support Vector Machines* e *Multilayer Perceptrons*. Contudo, após testes preliminares de comparação, o *Gradient Boosting* foi selecionado devido ao seu melhor desempenho. Além do mais, também foi testada a capacidade de algoritmos não supervisionados de agrupamento na

classificação, sendo utilizado principalmente o *K-Means*. Entretanto, esta abordagem não apresentou resultados satisfatórios.

Analisando a métrica *F1-Score* balanceado dos modelos, foram selecionados os melhores dentre estes variando parâmetros da abordagem: as entradas (variáveis) dos modelos; a utilização das variáveis brutas (*baseline*) com utilização de cálculos e relações de engenharia, ou da transformada de *wavelet;* e a utilização dos filtros de alarmes. As Tabelas 29 e 30 apresentam o percentual de modelos que utilizam tratamentos em suas variáveis de entrada e a utilização de filtros na criação de alarmes, respectivamente.

Tabela 29 – Tratamentos de variáveis utilizados nos modelos de classificação.

Abordagem	%
Wavelet	45,83
Calculado	31,25
Baseline	22,92

Fonte: elaborado pelo autor (2023).

Tabela 30– Utilização de filtros nos modelos de classificação.

Filtro	%
Sim	52,08
Não	47,92

Fonte: elaborado pelo autor (2023).

Os resultados da classificação de cada sistema, por grupamento de motores, se encontram expressos nas Tabelas 31-36. Nestas tabelas também estão contidas a abordagem utilizada no tratamento das variáveis de entrada e a utilização ou não do filtro de alarmes.

Tabela 31 – Resultados dos classificadores do SAC por grupamento.

G	**Abordagem**	**Filtro**	***F1-Score***		
			Micro	**Macro**	**Balanceado**
1	Calculado	Não	0,944 ± 0,014	0,359 ± 0,011	0,936 ± 0,017
2	Calculado	Não	0,905 ± 0,011	0,344 ± 0,013	0,915 ± 0,013
3	Calculado	Não	0,901 ± 0,025	0,388 ± 0,034	0,892 ± 0,018
4	Calculado	Sim	0,918 ± 0,007	0,359 ± 0,034	0,905 ± 0,009
5	*Baseline*	Não	0,934 ± 0,019	0,357 ± 0,054	0,915 ± 0,017
6	Calculado	Não	0,916 ± 0,011	0,374 ± 0,018	0,890 ± 0,017
7	*Wavelet*	Não	0,842 ± 0,037	0,324 ± 0,028	0,874 ± 0,015
8	*Wavelet*	Não	0,945 ± 0,009	0,358 ± 0,060	0,925 ± 0,010

Fonte: elaborado pelo autor (2023).

Tabela 32 – Resultados dos classificadores do SAA por grupamento.

G	Abordagem	Filtro	*F1-Score*		
			Micro	**Macro**	**Balanceado**
1	*Wavelet*	Sim	0,949 ± 0,009	0,209 ± 0,022	0,929 ± 0,010
2	*Wavelet*	Sim	0,918 ± 0,012	0,193 ± 0,024	0,912 ± 0,015
3	*Wavelet*	Sim	0,972 ± 0,005	0,264 ± 0,033	0,961 ± 0,007
4	*Wavelet*	Sim	0,935 ± 0,008	0,217 ± 0,026	0,904 ± 0,001
5	*Baseline*	Sim	0,947 ± 0,004	0,234 ± 0,024	0,939 ± 0,006
6	*Baseline*	Não	0,958 ± 0,005	0,229 ± 0,024	0,946 ± 0,003
7	Calculado	Sim	0,919 ± 0,002	0,216 ± 0,022	0,894 ± 0,000
8	*Baseline*	Sim	0,913 ± 0,009	0,242 ± 0,005	0,885 ± 0,002

Fonte: elaborado pelo autor (2023).

Tabela 33 – Resultados dos classificadores do SOL por grupamento.

G	Abordagem	Filtro	*F1-Score*		
			Micro	**Macro**	**Balanceado**
1	*Baseline*	Não	0,883 ± 0,008	0,230 ± 0,017	0,857 ± 0,007
2	*Wavelet*	Sim	0,820 ± 0,012	0,210 ± 0,026	0,767 ± 0,033
3	*Wavelet*	Não	0,905 ± 0,002	0,248 ± 0,018	0,880 ± 0,001
4	Calculado	Sim	0,805 ± 0,021	0,240 ± 0,010	0,746 ± 0,032
5	*Baseline*	Não	0,807 ± 0,009	0,199 ± 0,003	0,741 ± 0,021
6	*Wavelet*	Sim	0,796 ± 0,006	0,250 ± 0,027	0,729 ± 0,016
7	Calculado	Sim	0,819 ± 0,015	0,210 ± 0,019	0,778 ± 0,020
8	*Wavelet*	Não	0,823 ± 0,022	0,223 ± 0,020	0,768 ± 0,025

Fonte: elaborado pelo autor (2023).

Tabela 34 – Resultados dos classificadores do Sex por grupamento.

G	Abordagem	Filtro	*F1-Score*		
			Micro	**Macro**	**Balanceado**
1	*Wavelet*	Sim	0,982 ± 0,005	0,560 ± 0,105	0,979 ± 0,007
2	*Wavelet*	Sim	0,989 ± 0,000	0,548 ± 0,111	0,986 ± 0,001
3	*Wavelet*	Não	0,994 ± 0,001	0,606 ± 0,051	0,991 ± 0,001
4	*Wavelet*	Sim	0,981 ± 0,005	0,576 ± 0,050	0,977 ± 0,009
5	*Baseline*	Não	0,986 ± 0,005	0,566 ± 0,068	0,981 ± 0,006
6	Calculado	Não	0,978 ± 0,010	0,561 ± 0,089	0,969 ± 0,014
7	*Wavelet*	Sim	0,953 ± 0,002	0,522 ± 0,086	0,964 ± 0,002
8	*Wavelet*	Sim	0,994 ± 0,002	0,572 ± 0,046	0,992 ± 0,001

Fonte: elaborado pelo autor (2023).

Tabela 35 – Resultados dos classificadores do Sad por grupamento.

G	Abordagem	Filtro	*F1-Score*		
			Micro	**Macro**	**Balanceado**
1	*Baseline*	Sim	0,951 ± 0,001	0,347 ± 0,009	0,930 ± 0,002
2	Calculado	Não	0,962 ± 0,004	0,428 ± 0,038	0,952 ± 0,007
3	*Baseline*	Sim	0,967 ± 0,001	0,466 ± 0,083	0,961 ± 0,003
4	*Wavelet*	Sim	0,899 ± 0,007	0,324 ± 0,001	0,913 ± 0,002
5	*Wavelet*	Sim	0,821 ± 0,002	0,288 ± 0,040	0,850 ± 0,004
6	Calculado	Não	0,865 ± 0,029	0,254 ± 0,001	0,811 ± 0,039
7	*Baseline*	Sim	0,973 ± 0,006	0,437 ± 0,125	0,975 ± 0,007
8	Calculado	Sim	0,978 ± 0,001	0,371 ± 0,050	0,969 ± 0,001

Fonte: elaborado pelo autor (2023).

Tabela 36 – Resultados dos classificadores do MEC por grupamento.

G	Abordagem	Filtro	*F1-Score*		
			Micro	**Macro**	**Balanceado**
1	*Wavelet*	Não	0,276 ± 0,023	0,205 ± 0,031	0,244 ± 0,068
2	Calculado	Sim	0,710 ± 0,105	0,210 ± 0,015	0,599 ± 0,132
3	*Wavelet*	Sim	0,832 ± 0,019	0,227 ± 0,003	0,756 ± 0,026
4	Calculado	Sim	0,778 ± 0,059	0,242 ± 0,024	0,695 ± 0,085
5	*Baseline*	Não	0,527 ± 0,124	0,426 ± 0,063	0,500 ± 0,114
6	Calculado	Não	0,367 ± 0,043	0,299 ± 0,060	0,360 ± 0,000
7	*Wavelet*	Não	0,234 ± 0,083	0,143 ± 0,056	0,170 ± 0,071
8	*Wavelet*	Não	0,723 ± 0,034	0,258 ± 0,026	0,627 ± 0,046

Fonte: elaborado pelo autor (2023).

Por fim, a Tabela 37 expõe os mesmos resultados expostos acima, porém de forma condensada, agrupando-os de acordo com os sistemas avaliados. Também estão presentes a porcentagem de modelos que fazem uso dos filtros e da transformada de *wavelet* no tratamento das variáveis de entrada.

Tabela 37 – Resultados dos classificadores por sistema avaliado.

Sist.	Wavelet	Filtro	*F1-Score*		
			Micro	**Macro**	**Balanceado**
SAC	25,0%	12,5%	0,913 ± 0,033	0,358 ± 0,019	0,907 ± 0,020
SAA	50,0%	87,5%	0,939 ± 0,021	0,226 ± 0,022	0,921 ± 0,027
SOL	50,0%	50,0%	0,832 ± 0,039	0,226 ± 0,019	0,783 ± 0,055
Sex	75,0%	62,5%	0,982 ± 0,013	0,564 ± 0,024	0,980 ± 0,010
Sad	25,0%	75,0%	0,927 ± 0,058	0,364 ± 0,075	0,920 ± 0,060
MEC	50,0%	37,5%	0,556 ± 0,238	0,251 ± 0,084	0,494 ± 0,215

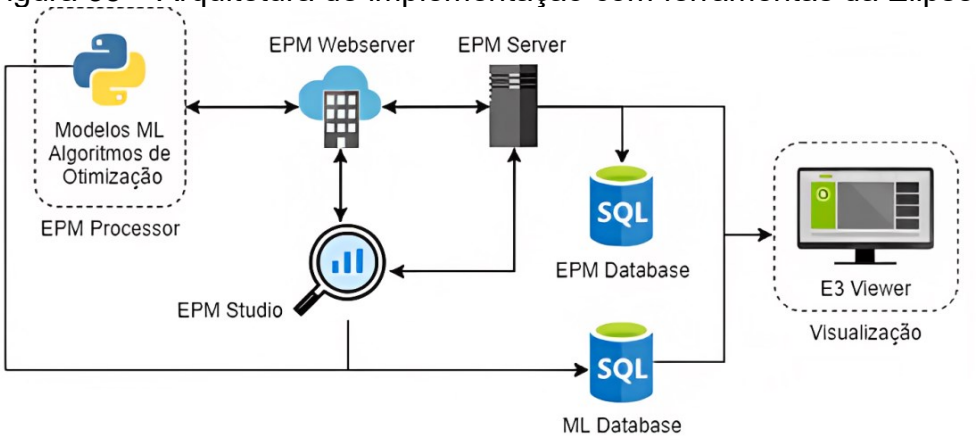
Fonte: elaborado pelo autor (2023).

5 IMPLEMENTAÇÃO

Para a execução dos modelos em tempo real, utilizou-se uma aplicação *on-premise*, onde as soluções são executadas em intervalos de 10 min, e os algoritmos de regressão e classificação são defasados em 5 min, de modo a atendar à necessidade de tomada de decisão da operação. Além disto, uma arquitetura foi integrada ao sistema supervisório da indústria, do tipo SCADA, utilizando um conjunto de ferramentas do *framework* Elipse (*Elipse Plant Manager* - EPM).

Dentro destas ferramentas, temos o EPM *Processor* utilizado para a execução dos modelos gerados pelo projeto. Por meio do EPM *Webserver*, estes processos podem realizar consultas de valores registrados pelos sensores no banco de dados EPM *Database*. Esse banco de dados, por sua vez, é gerenciado pelo EPM *Server*. Ainda é possível registrar novas variáveis no banco de dados do EPM, armazenando um valor, uma marca de tempo e um código de qualidade. Por fim, foram desenvolvidas interfaces utilizando o *Elipse Power* para apresentação dos resultados em telas, que se encontram disponíveis no Elipse E3 *Viewer*.

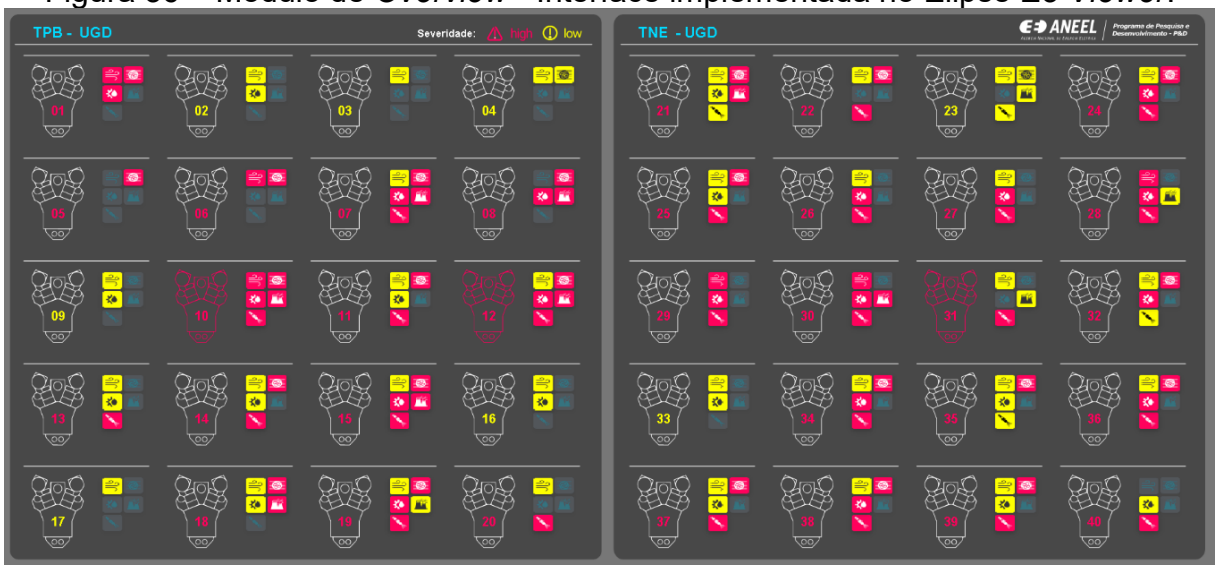
Figura 35 – Arquitetura de implementação com ferramentas da Elipse.



Fonte: elaborado pelo autor (2023).

Foram idealizados e desenvolvidos três painéis para monitoramento dos resultados dos modelos, criados de forma colaborativa,sendo estas: o Módulo de *Overview*, o Módulo de Monitoramento dos Motores e o Módulo de Gerenciamento de Alarmes, ilustrados pelas Figuras 36-38, respectivamente.

Figura 36 – Módulo de *Overview* - Interface implementada no Elipse E3 *Viewer*.



Fonte: elaborado pelo autor (2023).

O módulo de *overview*, representado acima, permite a visualização das 40 UGDs, onde para cada uma destas existem ícones correspondentes aos sistemas. A coloração destes reflete a ocorrência de alarmes e sua respectiva severidade. É possível acessar o Módulo de Monitoramento dos Motores específico de cada unidade geradora (representado abaixo) ao clicar nos respectivos ícones.

Figura 37 – Módulo de Monitoramento dos Motores - Interface implementada no Elipse E3 *Viewer*.

Fonte: elaborado pelo autor (2023).

Neste módulo estão contidas informações referentes à análise de variáveis, aos índices de saúde e aos alarmes. A análise das variáveis ocorre a partir de gráficos temporais (seção superior direita) e de correlação (seção inferior central), que permitem o usuário alterar as variáveis e explorar o funcionamento dos maquinários. O índice de saúde é exposto na forma de série temporal com uma semana de registro (seção superior central), além do valor instantâneo destes (seção inferior esquerda). Já os alarmes são dispostos tanto na seção inferior direta, quanto na seção superior esquerda e permitem o usuário acessar o Módulo de Gerenciamento de Alarmes, ilustrado na Figura 38.

O Módulo de Gerenciamento de Alarmes apresenta a lista dos alarmes gerados, diferenciando esses de acordo com sua severidade através da coloração dos mesmos. Esta listagem pode, ainda, ser filtrada de acordo com o sistema avaliado. É possível também classificar o alarme gerado como procedente ou não, auxiliando na compreensão e validação dos modelos. Além do mais, ainda há o monitoramento das variáveis chaves no momento da ocorrência do alarme, assim como a visualização atualizada de variáveis do processo.

Figura 38 – Módulo de Gerenciamento de Alarmes - Interface implementada no Elipse E3 *Viewer*.

Fonte: elaborado pelo autor (2023).

A execução dos modelos é realizada em tempo real, com consultas dos dados pertinentes das últimas 24 h no banco de dados (EPM *Database*). Todo pré-processamento utilizado na criação dos modelos é replicado no EPM *Processor*, de

modo a manter fidelidade à metodologia desenvolvida. Para tanto, são removidos desta consulta dados que correspondem a subidas e descidas de máquina, além dos que se encontram fora dos patamares de operação (*outliers*). Esses dados ainda são normalizados de acordo com suas médias e desvios padrões históricos.

Para os modelos regressivos, os resultados são comparados com os valores reais das variáveis alvos e transformados de acordo com a Equação (30). Finalmente, é extraída a média móvel destes valores e atualizado o índice de saúde dos equipamentos. Já para os modelos de classificação, podem ser aplicadas as transformadas de *wavelet* e a validação destes resultados com os filtros aplicados. Na classificação também é utilizada uma média móvel dos resultados.

Portanto, as técnicas de aprendizado de máquina, a metodologia e a arquitetura do projeto podem ser expandidas ou adaptadas para quaisquer setores industriais que utilizem supervisórios do tipo SCADA. Convém salientar que esses resultados geraram a publicação do artigo “Digital twin and machine learning for decision support in thermal power plant with combustion engines” (DEON, COTTA, *et al.*, 2022).

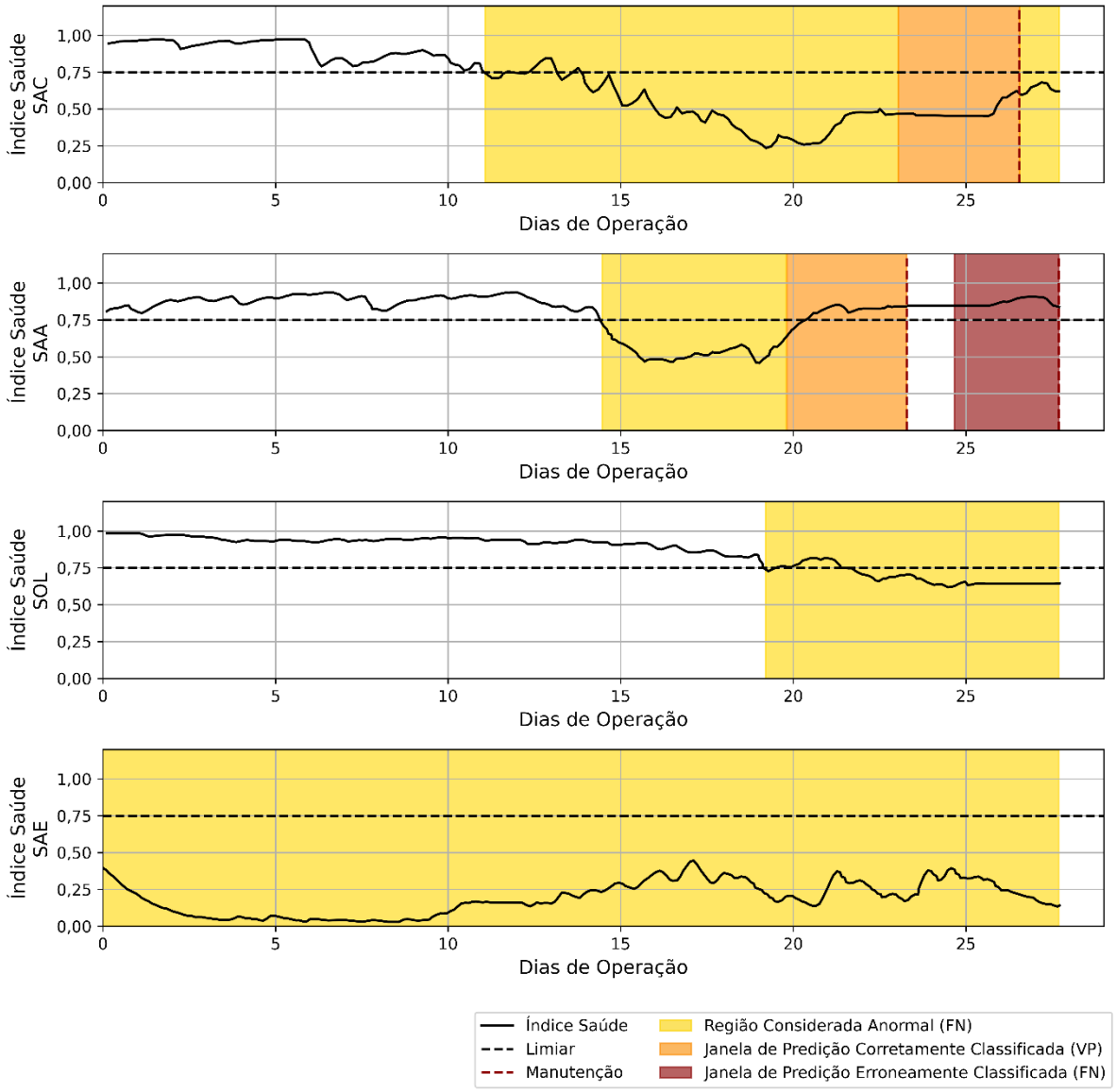
6 OPERAÇÃO ASSITIDA

A implementação final da ferramenta ocorreu em novembro de 2021, e desde então os modelos são processados no EPM *Processor*. Portanto, foi possível avaliar o desempenho da execução dos modelos em tempo real. Esta análise baseia-se na comparação entre os dados gerados pela ferramenta, tanto de índice de saúde quanto classificação, e os registros de eventos da usina, tornando possível correlacionar os dados entre tabelas a partir da máquina e da data.

Para os modelos regressivos, a avaliação necessita da extração dos valores do índice de saúde registrados, além das notas de manutenção. O índice de saúde em si não necessita de nenhum tratamento, contudo para sua avaliação as notas de manutenção necessitam passar pelos mesmos filtros utilizados na criação dos modelos, além de que as janelas temporais utilizadas para a criação das métricas de classificação precisam ser recriadas. A Figura 39 ilustra o histórico dos índices de saúde gerados entre os meses de novembro e dezembro de 2021 da unidade geradora número 7. Nesta figura também estão contidas as manutenções/falhas

associadas a cada índice de saúde, além das janelas temporais respectivas à predição das falhas.

Figura 39 – Avaliação dos Gêmeos Digitais referentes a UGD7 entre os meses de novembro de dezembro de 2021.



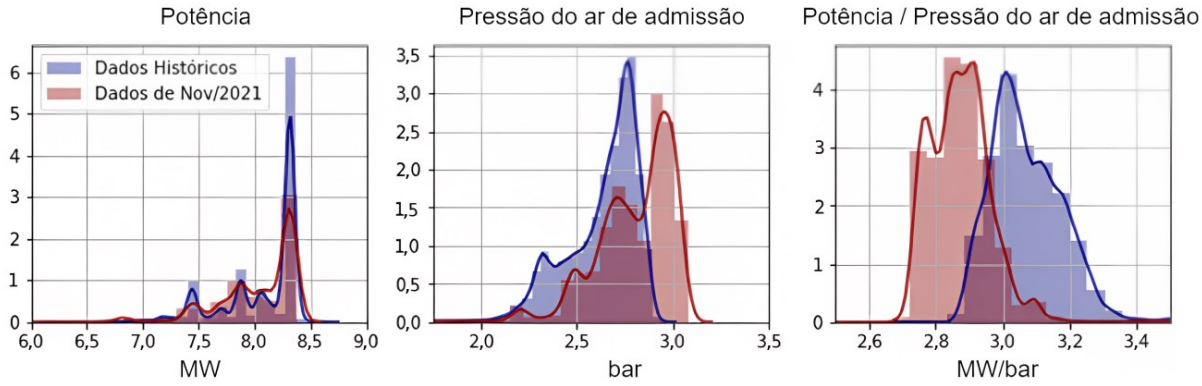
Fonte: elaborado pelo autor (2023).

Sobre as métricas obtidas, é importante ressaltar alguns pontos, como o tempo de avaliação e o método. O período avaliado é consideravelmente efêmero, e tem implicação direta nas métricas. A Acurácia é afetada pela baixa quantidade de verdadeiros positivos (momentos sem manutenções e/ou falhas). Normalmente, estes períodos são bastante vastos, podendo ser superiores a meses de operação. Já o *F1-Score* depende essencialmente da existência de falhas, e uma vez que essas não

ocorram seu valor será igual a zero. Além disto, como são avaliadas janelas temporais (do mesmo modo que na elaboração dos treinos e testes), falsos positivos podem ser criados mesmo quando esses antecipam falhas, como é o caso das duas manutenções preditas, que já haviam sido sinalizadas alguns dias antes da janela temporal.

Ainda há mais uma ressalva, a possível ocorrência de descalibração de sensores e/ou do *concept drift* – conceito no qual os patamares de operação são alterados. Neste caso, a performance dos modelos regressivos pode ser comprometida, e possivelmente é o que ocorre no índice de saúde do SAE da UGD7. A Figura 40 ilustra alteração dos patamares operacionais da pressão do ar de admissão nos meses avaliados. Essa pressão possui uma correlação alta com a potência das unidades geradoras (superior a 0,90).

Figura 40 – Comparação da distribuição das variáveis de pressão do ar de admissão e a potência da UGD7, nos meses de avaliados (11/2021 ~12/2021), com os dados históricos.



Fonte: elaborado pelo autor (2023).

Levando isto em consideração, a Tabela 38 apresenta os resultados obtidos da operação assistida referente à UGD7.

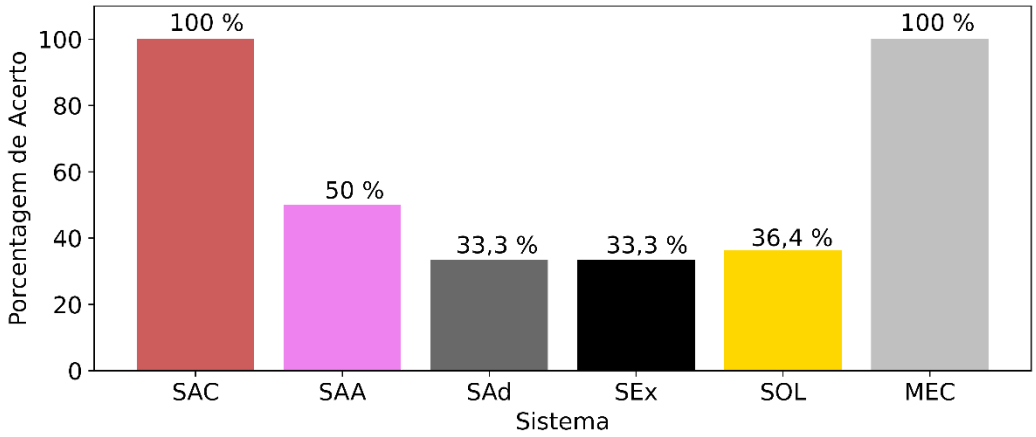
Tabela 38 – Resultados da operação assistida dos Gêmeos Digitais.

Sistema	Acurácia	F1-Score
SAC	0,526	0,348
SAA	0,699	0,455
SOL	0,693	0,000
SAE	0,000	0,000

Fonte: elaborado pelo autor (2023).

Para os modelos de classificação, é possível realizar uma avaliação mais direta e eficaz ao correlacionar os alarmes gerados com os desligamentos automáticos das máquinas. Nesta avaliação, pode-se constatar que aproximadamente 52% dos desligamentos automáticos foram previamente alertados pela ferramenta de predição de falhas. Na Figura 41 é ilustrado o resultado percentual de paradas de máquinas identificadas de acordo com sistemas.

Figura 41 – Desligamentos automáticos detectados pela geração de alarmes.



Fonte: elaborado pelo autor (2023).

7 CONCLUSÃO

Através deste estudo foi possível explorar a utilização de modelos de aprendizado de máquina com abordagem de Gêmeos Digitais para elaboração de um sistema de apoio à tomada de decisão no que diz respeito à execução de manutenções preditivas.

Sobre os resultados dos Gêmeos Digitais, as métricas de regressão apresentam ótimos resultados, sendo que a média dos coeficientes de determinação (R^2) chega a 0,928, o que indica uma correlação alta. A média dos RMSE dos modelos é de 0,072. Levando em conta que este valor é baseado nas variáveis normalizadas e, desta forma, são referentes aos desvios padrões destas, este resultado expressa um erro médio irrisório. Portanto, pode-se concluir que os Gêmeos Digitais se mostram fiéis ao processo.

A capacidade de classificação dos modelos regressivos apresenta uma Acurácia média de 0,861, o que na prática significa que 86,1% do tempo o modelo está correto em suas classificações. Contudo, a média do *F1-Score* corresponde a 0,526. Em uma interpretação simplista deste resultado, pode-se assumir que para cada falha corretamente classificada, uma falha não é reconhecida e um alarme falso é gerado.

Os modelos de classificação apresentam um *F1-Score* balanceado médio de 0,834, superior à capacidade de classificação realizada pelos Gêmeos Digitais. Contudo, vale lembrar que estes modelos fazem uso do índice de saúde para redução de falsos positivos. O *F1-Score* macro, que basicamente se trata de uma média desta métrica entre as classes, possui um valor inferior, sendo 0,332 sua média. Porém, isso não é um empecilho, visto que a sinalização de alguma anomalia presente no sistema é melhor expressa pelo *F1-Score* balanceado.

As notas de manutenção talvez tenham sido o maior empecilho para a criação dos modelos, uma vez que estas não apresentavam de forma clara e padronizada as manutenções realizadas. O uso de processamento de linguagem natural (uma técnica de aprendizado de máquina) combinada com técnicas de agrupamento poderiam ser aplicadas aqui para otimizar os momentos selecionados como “saudáveis” e falhos. Outra alternativa seria a criação de um sistema de registro mais eficiente, com o devido treinamento dos colaboradores.

Além disto, vale ressaltar que alguns sensores possuíam uma quantidade significativa de valores fora de especificação, tornando o uso destes inviável no treinamento dos modelos. Existe também a possibilidade de algumas variáveis apresentarem *concept drift*, ou seja, os patamares das variáveis podem variar com o tempo. Esta questão pode ser sanada ao realizar o retreino dos modelos periodicamente.

Ainda, poderia ser avaliado o uso de técnicas de redução de dimensionalidade em algumas variáveis utilizadas como entradas dos modelos. A investigação de outras funções para a normalização dos desvios dos modelos em relação à variável alvo pode ser considerado outro ponto de aperfeiçoamento. Além disso, as abordagens poderiam ser mescladas, ou seja, os Gêmeos Digitais poderiam contar com a transformada de *wavelet*, e os modelos de classificação poderiam ser otimizados pelo uso do algoritmo genético.

Por fim, pode-se concluir que os objetivos levantados no início do trabalho foram atendidos. Os modelos regressivos são robustos e fiéis aos processos, além de apresentarem uma capacidade de classificação de falhas considerável, permitindo o refinamento dos modelos de classificação. Assim, esses modelos, juntamente com o desenvolvimento das telas de monitoramento, servem como uma ferramenta de auxílio à tomada de decisão na manutenção preditiva bastante completa.

8 REFERÊNCIAS

AGUILAR, F. J. E. et al. Predictive Maintenance System for 2 Stroke Diesel Engines. **ASME International Mechanical Engineering Congress and Exposition**, v. 44458, p. 1105-1114, 2010.

AIVALIOTIS, P.; GEORGOULIAS, K.; CHRYSSOLOURIS, G. The use of Digital Twin for predictive maintenance in manufacturing. **International Journal of Computer Integrated Manufacturing**, v. 32, p. 1067-1080, 2019.

ALVES DE ARAUJO JUNIOR, C. A. et al. Digital twins of the water cooling system in a power plant based on fuzzy logic. **Sensors**, v. 21, p. 6737, 2021.

AYODELE, T. O. Types of machine learning algorithms. **New advances in machine learning**, v. 3, p. 19-48, 2010.

BONACCORSO, G. **Machine learning algorithms**. [S.l.]: Packt Publishing Ltd, 2017.

BREIMAN, L. Bagging predictors. **Machine learning**, v. 24, p. 123-140, 1996.

BREIMAN, L. Random forests. **Machine learning**, p. 5-32, 2001.

CAI, C.; WENG, X.; ZHANG, C. A novel approach for marine diesel engine fault diagnosis. **Cluster computing**, v. 20, n. 1691-1702, 2017.

CHAI, T.; DRAXLER, R. R. Root mean square error (RMSE) or mean absolute error (MAE)?--Arguments against avoiding RMSE in the literature. **Geoscientific model development**, v. 7, p. 1247-1250, 2014.

CORTES, C.; VAPNIK, V. Support-vector networks. **Machine learning**, v. 20, p. 273-297, 1995.

DEB, K. et al. A fast and elitist multiobjective genetic algorithm: NSGA-II. **IEEE transactions on evolutionary computation**, v. 6, p. 182-197, 2002.

DEON, B. et al. Digital Twin and Machine Learning for Decision Support. **Knowledge-Based Systems**, v. 253, p. 109578, 2022.

DRISCOLL, C. T. et al. US power plant carbon standards and clean air and health co-benefits. **Nature Climate Change**, v. 5, p. 535-540, 2015.

EBRASIL Energia. **EPASA**, s. d. Disponivel em: <EBRASIL Energia, "Epasa", https://ebrasilenergia.com.br/empresa/epasa/.>. Acesso em: 1 maio 2023.

EL NAQA, I.; MURPHY, M. J. What is machine learning? In: **machine learning in radiation oncology**. [S.l.]: Springer, 2015. p. 3-11.

EPASA - Geração de Energia. **Epasa (Centrais Elétricas da Paraíba)**, s.d. Disponivel em: <https://www.epasa.online/geracao-de-energia>. Acesso em: 11 Janeiro 2022.

EPE. **Balanço Energético Nacional 2021, Ano base 2020**. EPE [Empresa de Pesquisa Energética]. Rio de Janeiro. 2021.

FAJERSZTAJN, L. et al. Air pollution: a potentially modifiable risk factor for lung cancer. **Nature Reviews Cancer**, v. 13, p. 674-678, 2013.

FONSECA, C. M.; FLEMING, P. J. Genetic algorithms for multiobjective optimization: formulationdiscussion and generalization, 1993.

FORREST, S. Genetic algorithms. **ACM Computing Surveys (CSUR)**, v. 28, p. 77-80, 1996.

FRADKOV, A. L. Early history of machine learning. **IFAC-PapersOnLine**, v. 53, p. 1385-1390, 2020.

FRIEDMAN, J. H. Greedy function approximation: a gradient boosting machine. **Annals of statistics**, p. 1189-1232, 2001.

FRIEDMAN, J. H. Stochastic gradient boosting. **Computational statistics & data analysis**, v. 38, p. 367-378, 2002.

FUKUSHIMA, K.; MIYAKE, S. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In: **Competition and cooperation in neural nets**. [S.l.]: Springer, 1982. p. 267-285.

GEEKSTYLE. **Business Intelligence:** Its Relationship with Big Data, 2021. Disponivel em: <https://www.linkedin.com/pulse/business-intelligence-its-relationship-big-data-geekstyle/>. Acesso em: 01 abr. 2023.

GHAHRAMANI, Z. Unsupervised learning. In: **Summer school on machine learning**. [S.l.]: Springer, 2003. p. 82-112.

GLAESSGEN, E.; STARGEL, D. The digital twin paradigm for future NASA and US Air Force vehicles. In: **53rd AIAA/ASME/ASCE/AHS/ASC structures, structural dynamics and materials conference 20th AIAA/ASME/AHS adaptive structures conference 14th AIAA**. [S.l.]: [s.n.], 2012. p. 1818.

GRANDINI, M.; BAGLI, E.; VISANI, G. Metrics for multi-class classification: an overview. **arXiv preprint arXiv:2008.05756**, 2020.

GRIEVES, M. Digital twin: manufacturing excellence through virtual factory replication. **White paper**, v. 1, p. 1-7, 2014.

HAYASHI, F. **Econometrics**. [S.l.]: Princeton University Press, 2011.

HAYKIN, S. **Redes neurais:** princípios e prática. [S.l.]: Bookman Editora, 2001.

HOLLAND, J. H. **Adaptation in natural and artificial systems:** an introductory analysis with applications to biology, control, and artificial intelligence. [S.l.]: MIT press, 1975/1992.

HOSSIN, M.; SULAIMAN, M. N. A review on evaluation metrics for data classification evaluations. **International journal of data mining \& knowledge management process**, v. 5, p. 1, 2015.

HUANG, X.; WU, L.; YE, Y. A review on dimensionality reduction techniques. **International Journal of Pattern Recognition and Artificial Intelligence**, v. 33, p. 1950017, 2019.

KAN, H.; CHEN, R.; TONG, S. Ambient air pollution, climate change, and population health in China. **Environment international**, v. 42, p. 10-19, 2012.

KARKARE, P. KDnuggets. **Decision Trees - An introduction**, 2019. Disponivel em: <https://www.kdnuggets.com/2019/02/decision-trees-introduction.html>. Acesso em: 01 abr. 2023.

KATOCH, S.; CHAUHAN, S. S.; KUMAR, V. A review on genetic algorithm: past, present, and future. **Multimedia Tools and Applications**, v. 80, p. 8091-8126, 2021.

KELLEHER, J. D.; MAC NAMEE, B.; D'ARCY, A. **Fundamentals of machine learning for predictive data analytics:** algorithms, worked examples, and case studies. [S.l.]: MIT press, 2020.

KINGSFORD, C.; SALZBERG, S. L. What are decision trees? **Nature biotechnology**, v. 26, p. 1011-1013, 2008.

LIU, M. et al. Review of digital twin about concepts, technologies, and industrial applications. **Journal of Manufacturing Systems**, v. 58, p. 346-361, 2021.

MAHESH, B. Machine learning algorithms-a review. **International Journal of Science and Research (IJSR).[Internet]**, v. 9, p. 381-386, 2020.

MANISALIDIS, I. et al. Environmental and health impacts of air pollution: a review. **Frontiers in public health**, p. 14, 2020.

MIN, Q. et al. Machine learning based digital twin framework for production optimization in petrochemical industry. **International Journal of Information Management**, v. 49, p. 502-519, 2019.

MINSKY, M.; PAPERT, S. An introduction to computational geometry. **Cambridge tiass., HIT**, v. 479, p. 480, 1969.

MITCHELL, M. Genetic algorithms: An overview. In: **Complex**. [S.l.]: Citeseer, v. 1, 1995. p. 31-39.

MOBLEY, R. K. **An introduction to predictive maintenance**. [S.l.]: Elsevier, 2002.

NAGELKERKE, N. J.; OTHERS. A note on a general definition of the coefficient of determination. **Biometrika**, v. 78, p. 691-692, 1991.

NASTESKI, V. An overview of the supervised machine learning methods. **Horizons. b**, v. 4, p. 51-62, 2017.

NEGRI, E. et al. A digital twin-based scheduling framework including equipment health index and genetic algorithms. **IFAC-PapersOnLine**, v. 52, p. 43-48, 2019.

O SISTEMA INTERLIGADO NACIONAL. **ONS - Operador Nacional do Sistema Elétrico**, s.d. Disponivel em: <http://www.ons.org.br/paginas/sobre-o-sin/o-que-e-o-sin>. Acesso em: 11 Janeiro 2022.

POLIKAR, R. **Ensemble learning**. [S.l.]: Springer, 2012.

POLONI, C.; MOSETTI, G.; OUTROS. Aerodynamic shape optimization by means of a genetic algorithm. In: **the 5th international symposium on computational fluid dynamics**. [S.l.]: [s.n.], 1993. p. 273-284.

PROGRAMA de Pesquisa e Desenvolvimento Tecnológico. **ANEEL (Agência Nacional de Energia Elétrica)**, s.d. Disponivel em: <https://www.gov.br/aneel/pt-br/assuntos/pesquisa-e-desenvolvimento>. Acesso em: 11 Janeiro 2022.

QUINLAN, J. R. Induction of decision trees. **Machine learning**, v. 1, p. 81-106, 1986.

RIGONI, E.; POLES, S. NBI and MOGA-II, two complementary algorithms for multi-objective optimizations. In: **Dagstuhl seminar proceedings**. [S.l.]: [s.n.], 2005.

ROSENBLATT, F. The perceptron: a probabilistic model for information storage and organization in the brain. **Psychological review**, 1958.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. **Nature**, v. 323, p. 533-536, 1986.

SAMUEL, A. L. Some studies in machine learning using the game of checkers. **IBM Journal on Research and Development**, 1959.

SCHAPIRE, R. E. The strength of weak learnability. **Machine learning**, v. 5, p. 197-227, 1990.

SELCUK, S. Predictive maintenance, its implementation and latest trends. **Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture**, v. 231, p. 1670-1679, 2017.

SHARMA, S.; SHARMA, S.; ATHAIYA, A. Activation functions in neural networks. **towards data science**, v. 6, p. 310-316, 2017.

SISTEMA Interligado Nacional. **ANA - Agência Nacional de Águas**, s. d. Disponivel em: <>. Acesso em: 11 Janeiro 2022.

SRINIVAS, N.; DEB, K. Muiltiobjective optimization using nondominated sorting in genetic algorithms. **Evolutionary computation**, v. 2, p. 221-248, 1994.

SRIVASTAVA, A. K.; SRIVASTAVA, V. K.; ULLAH, A. The coefficient of determination and its adjusted version in linear regression models. **Econometric reviews**, v. 14, p. 229-240, 1995.

SUSTO, G. A. et al. Machine learning for predictive maintenance: A multiple classifier approach. **IEEE transactions on industrial informatics**, v. 11, p. 812-820, 2014.

TAO, F. et al. Digital twin driven prognostics and health management for complex equipment. **Cirp Annals**, v. 67, p. 169-172, 2018.

TAO, F. et al. Digital twin in industry: State-of-the-art. **IEEE Transactions on Industrial Informatics**, v. 15, p. 2405-2415, 2018.

TAO, J. et al. Intelligent fault diagnosis of diesel engines via extreme gradient boosting and high-accuracy time--frequency information of vibration signals. **Sensors**, v. 19, p. 3280, 2019.

TIBCO. **What is a Random Forest?**, s. d. Disponivel em: <https://www.tibco.com/reference-center/what-is-a-random-forest.>. Acesso em: 01 abr. 2023.

TUEGEL, E. J. et al. Reengineering aircraft structural life prediction using a digital twin. **International Journal of Aerospace Engineering**, v. 2011, 2011.

WERBOS, P. J. **Beyond Regression:** New Tools for Prediction and Analysis in the Behavioral Sciences. [S.l.]: Harvard University, 1975.

WERBOS, P. J. Backpropagation through time: what it does and how to do it. **Proceedings of the IEEE**, v. 78, p. 1550-1560, 1990.

WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. **Climate research**, v. 30, p. 79-82, 2005.

YANG, X.-S. **Engineering optimization:** an introduction with metaheuristic applications. [S.l.]: John Wiley & Sons, 2010.

ZHANG, D. A coefficient of determination for generalized linear models. **The American Statistician**, v. 71, p. 310-316, 2017.

ZONTA, T. et al. Predictive maintenance in the Industry 4.0: A systematic literature review. **Computers & Industrial Engineering**, v. 150, p. 106889, 2020.

APÊNDICE A – VARIÁVEIS DE ENTRADA DOS MODELOS DE REGRESSÃO

Tabela A.1 – Variáveis adicionais calculadas com base nas variáveis sensoriadas.

Relação	Descrição	Cálculo
Média T110	Média das temperaturas dos mancais internos da UGD	MED(1~11 T110)
Desvio T110	Desvio padrão das temperaturas dos mancais internos da UGD	STD(1~11 T110)
Média T320	Média das temperaturas do óleo de lubrificação nos mancais móveis	MED(1~9 T320)
Desvio T320	Desvio padrão das temperaturas do óleo de lubrificação nos mancais móveis	STD(1~9 T320B)
Média T410	Média das temperaturas dos gases de exaustão na saída dos cilindros	MED(1~9 T410)
Desvio T410	Desvio padrão das temperaturas dos gases de exaustão na saída dos cilindros	STD(1~9 T410)
Média T410A	Média das temperaturas dos gases de exaustão na saída dos cilindros – Lado A	MED(1~9 T410A)
Desvio T410A	Desvio padrão das temperaturas dos gases de exaustão na saída dos cilindros – Lado A	STD(1~9 T410A)
Média T410B	Média das temperaturas dos gases de exaustão na saída dos cilindros – Lado B	MED(1~9 T410B)
Desvio T410B	Desvio padrão das temperaturas dos gases de exaustão na saída dos cilindros – Lado B	STD(1~9 T410B)
dT SAT	Diferença total de temperatura da água de resfriamento do SAT	T203 - T201
dT SAT HEA/B	Diferença de temperatura da água de resfriamento do SAT no HEA/B	T202 - T201
dT SAT UGD	Diferença de temperatura da água de resfriamento do SAT na UGD	T203 - T202
dT SBT	Diferença total de temperatura da água de resfriamento do SBT	T203 - T201
dT SAd TCA	Diferença de temperatura do ar de admissão no TCA	T402A - T401A
dT SAd TCB	Diferença de temperatura do ar de admissão no TCB	T402B - T401B
dT SEx TCA	Diferença de temperatura dos gases de exaustão no TCA	T420A - T421A
dT SEx TCA	Diferença de temperatura dos gases de exaustão no TCA	T420A - T421A
dT SEx TCB	Diferença de temperatura dos gases de exaustão no TCB	T420B - T421B
dT SAE A	Diferença de temperatura entre os gases de exaustão e o ar de admissão – Lado A	T420A - T402A
dT SAE B	Diferença de temperatura entre os gases de exaustão e o ar de admissão – Lado B	T420B - T402B

Fonte: elaborado pelo autor (2023).

Tabela A.2 – Variáveis de entrada, por grupamento, utilizadas nos modelos de regressão do SAC.

Variáveis de entrada	G1	G2	G3	G4	G5	G6	G7	G8	GL	Total
P102	0	1	0	1	1	0	1	0	0	4
T102	1	1	0	0	1	1	1	0	0	5
T201	1	1	0	0	1	0	0	1	0	4
T202	0	0	0	1	0	1	1	0	0	3
T203	1	1	1	0	1	1	1	0	1	7
T211	1	0	1	1	0	0	1	1	0	5
T220	1	1	1	1	0	1	1	1	1	8
T400	0	0	1	0	1	1	0	0	0	3
T401A	0	0	0	0	0	0	0	0	1	1
P402A	1	1	1	1	0	1	1	1	1	8
T402A	1	0	0	0	0	1	0	0	1	3
P402B	1	1	0	1	0	0	1	0	0	4
T402B	1	1	1	1	1	1	1	1	0	8
T420A	0	1	0	0	1	0	1	0	0	3
T420B	0	1	0	1	0	1	0	1	0	4
T421A	0	1	1	0	1	0	0	0	0	3
T421B	0	0	0	0	0	0	1	0	0	1
Desvio T110	0	1	0	0	0	0	0	0	0	1
Média T110	0	1	0	0	1	0	0	0	0	2
Desvio T320	0	0	0	0	0	0	1	1	1	3
Média T320	0	0	1	1	1	0	0	0	0	3
Média T410	0	0	0	0	0	1	1	1	0	3
Média T410A	0	0	0	1	0	1	0	0	0	2
Desvio T410B	0	0	0	0	1	0	0	0	0	1
Média T410B	0	0	0	0	1	0	1	0	0	2
dT SAT	1	0	0	0	0	0	0	0	0	1
dT SAT HEA/B	0	1	0	0	0	1	1	0	0	3
dT SAT UGD	0	0	0	0	0	0	1	0	0	1
dT SAT IC	0	1	0	0	0	1	1	0	0	3
dT SAT UGD	0	0	0	0	0	0	1	0	0	1
dT SAd TCB	0	0	0	0	0	1	0	0	0	1
dT SEx TCA	0	1	0	0	0	0	0	0	1	2
dT SEx TCB	0	1	1	0	0	1	1	0	0	4
dT SAE A	0	0	0	0	0	0	1	0	0	1
dT SAE B	0	0	0	0	1	1	0	0	0	2

Fonte: elaborado pelo autor (2023).

Tabela A.3 – Variáveis de entrada, por grupamento, utilizadas nos modelos de regressão do SAA.

Variáveis de entrada	G1	G2	G3	G4	G5	G6	G7	G8	GL	Total
P102	0	0	0	1	0	0	0	0	0	1
T102	0	1	1	1	1	1	0	0	0	5
T201	1	1	1	0	1	1	1	1	1	8
P202	1	0	1	1	1	1	0	1	1	7
T202	1	1	1	1	1	1	1	1	0	8
P211	1	0	1	1	0	0	0	0	1	4
T211	0	0	1	0	0	0	0	1	1	3
T212	0	1	0	1	1	0	0	1	1	5
P220	0	0	0	0	1	0	0	0	0	1
T220	1	1	0	0	0	0	0	1	0	3
T400	1	0	0	0	0	0	0	0	0	1
T401A	0	0	0	0	0	0	0	1	1	2
T401B	0	0	0	0	0	1	0	0	0	1
P402A	1	0	0	0	0	1	0	1	1	4
T402A	0	1	1	1	1	0	0	0	1	5
P402B	0	1	0	0	0	0	0	0	0	1
T402B	0	1	0	0	0	1	1	0	0	3
T420A	1	0	0	0	1	0	1	0	1	4
T420B	1	0	1	1	1	0	1	0	0	5
T421A	0	0	0	0	1	0	1	0	1	3
T421B	1	0	1	0	1	1	0	0	0	4
Média T110	0	0	0	1	1	1	0	1	0	4
Desvio T320	0	0	0	1	0	0	0	0	0	1
Média T320	0	1	1	0	0	0	0	0	1	3
Desvio T410	1	0	0	0	0	1	0	1	0	3
Desvio T410A	0	0	0	0	0	1	0	0	0	1
Média T410A	1	0	0	0	0	0	0	0	0	1
Desvio T410B	0	0	0	0	1	0	1	0	0	2
Média T410B	0	0	1	0	1	0	0	0	0	2
dT SAd TCA	0	0	0	0	0	0	0	1	1	2
dT SAd TCB	0	0	0	0	0	1	0	1	0	2
dT SEx TCA	0	0	0	0	0	0	0	0	1	1
dT SEx TCB	0	1	0	0	0	0	0	0	0	1
dT SAE A	0	1	0	0	0	0	1	0	0	2
dT SAE B	0	0	0	0	1	0	0	0	0	1

Fonte: elaborado pelo autor (2023).

Tabela A.4 – Variáveis de entrada, por grupamento, utilizadas nos modelos de regressão do SOL.

Variáveis de entrada	G1	G2	G3	G4	G5	G6	G7	G8	GL	Total
P102	0	0	1	0	1	1	0	0	0	3
T102	1	0	1	0	1	0	0	1	0	4
T201	1	1	0	1	0	1	1	0	1	6
P202	0	0	0	0	1	0	1	1	0	3
T202	0	1	0	0	0	1	0	0	0	2
T203	1	1	1	1	1	0	1	1	0	7
P211	0	0	0	1	0	1	1	0	1	4
T211	0	0	1	1	0	0	0	1	1	4
T212	1	1	0	0	1	0	0	1	1	5
P220	0	1	0	0	0	0	0	0	0	1
T220	1	0	1	1	1	1	0	1	0	6
T300	1	0	0	0	0	0	1	0	0	2
T301	0	0	1	0	0	0	1	0	0	2
P302	1	1	0	0	0	0	0	0	1	3
P310	1	1	0	1	0	1	0	0	1	5
T311A	1	1	1	0	1	0	0	1	0	5
T311B	1	1	0	1	1	1	1	1	0	7
T400	0	1	0	0	0	0	0	0	1	2
P402A	0	1	1	0	0	1	1	1	0	5
T402A	0	1	0	0	0	0	1	0	1	3
T402B	0	0	1	0	0	1	0	1	0	3
Desvio T110	0	1	1	0	0	0	0	0	0	2
Média T110	0	0	0	1	1	1	0	0	0	3
Desvio T320	1	0	0	0	0	0	1	0	1	3
Média T320	1	1	1	1	1	1	1	1	1	9
Desvio T410	0	1	0	0	0	0	1	1	0	3
Média T410	1	0	0	0	0	0	0	0	0	1
Desvio T410A	0	0	0	0	0	0	1	1	0	2
Média T410A	0	0	0	1	0	0	0	1	0	2
Desvio T410B	0	0	0	0	0	0	0	1	0	1
Média T410B	0	1	1	0	1	0	0	0	0	3
U100	0	0	0	0	0	0	1	0	0	1
dT SAT	0	0	0	1	0	1	0	0	0	2
dT SAT HEA/B	0	0	0	1	0	0	1	0	0	2
dT SAT UGD	0	0	1	1	0	0	0	0	0	2
dT SAT IC	0	0	0	1	0	0	1	0	0	2
dT SATUGD	0	0	1	1	0	0	0	0	0	2
dT SBT	0	0	0	0	1	1	0	0	0	2
dT Sad TCA	0	0	0	0	0	0	0	1	0	1
dT Sad TCB	0	0	0	0	0	1	0	0	0	1
dT Sex TCA	0	1	1	0	0	0	0	0	0	2
dT SAE A	0	0	0	0	0	0	1	0	1	2

Fonte: elaborado pelo autor (2023).

Tabela A.5 – Variáveis de entrada, por grupamento, utilizadas nos modelos de regressão do SAE.

Variáveis de entrada	G1	G2	G3	G4	G5	G6	G7	G8	GL	Total
T201	1	0	0	1	1	0	1	1	0	5
P202	0	0	0	0	0	0	0	1	0	1
T202	1	1	1	1	0	1	0	0	0	5
T203	0	0	1	0	0	0	0	0	0	1
P211	0	0	1	0	0	0	0	1	1	3
T211	0	0	1	1	0	0	1	1	0	4
T212	1	1	1	1	1	1	1	1	0	8
P220	0	0	0	1	0	0	0	1	0	2
T220	0	0	0	0	1	0	0	1	0	2
P310	0	1	0	0	0	1	0	0	0	2
T311A	1	1	1	1	0	1	1	1	1	8
T311B	1	1	0	0	1	1	1	1	0	6
T400	0	0	0	0	0	1	0	0	0	1
T401A	0	0	0	0	0	0	0	1	1	2
T401B	0	0	0	0	0	1	0	1	0	2
T402A	1	1	1	0	1	1	0	1	0	6
T402B	0	0	0	0	1	1	0	0	0	2
T420A	1	1	0	0	0	0	1	0	1	4
T420B	0	0	0	0	1	1	1	0	0	3
T421A	1	1	0	1	0	0	0	1	0	4
T421B	0	1	1	0	0	0	1	0	0	3
Desvio T110	0	0	1	0	0	0	1	0	1	3
Média T110	0	0	0	0	0	1	1	0	0	2
Desvio T320	0	0	1	0	0	0	1	0	0	2
Média T320	0	0	1	0	1	0	0	0	1	3
Desvio T410	0	0	1	0	0	1	0	1	0	3
Média T410	0	0	0	0	0	0	1	0	0	1
Desvio T410A	0	0	0	0	0	1	0	0	0	1
Média T410A	0	0	0	0	1	0	0	0	0	1
Desvio T410B	0	0	0	0	1	0	0	0	0	1
Média T410B	0	0	1	0	1	0	1	0	0	3
U1	1	0	1	1	0	1	0	1	0	5
dT SAT	0	0	0	0	1	1	0	1	0	3
dT SAT HEA/B	0	0	0	1	0	0	1	0	0	2
dT SATIC	0	0	0	1	0	0	1	0	0	2
dT SBT	0	0	0	1	0	0	0	0	0	1
dT SEx TCA	0	1	0	0	1	0	0	1	0	3
dT SEx TCB	0	0	0	0	0	0	1	0	0	1
dT SAE A	0	0	0	0	0	1	0	0	0	1
dT SAE B	1	0	0	1	0	0	1	0	0	3

Fonte: elaborado pelo autor (2023).