UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
CÂMPUS DE ROSANA

LEONARDO FERNANDO FINI

Análise Preditiva do Preço de Liquidação das Diferenças no Mercado de Energia Elétrica via Algoritmos de *Machine Learning*

Rosana - SP
2023

Leonardo Fernando Fini

Análise Preditiva do Preço de Liquidação das Diferenças no Mercado de Energia Elétrica via Algoritmos de *Machine Learning*

Trabalho de Conclusão de Curso apresentado à Coordenadoria de Curso de Engenharia de Energia do Campus de Rosana, Universidade Estadual Paulista, como parte dos requisitos para obtenção do diploma de Graduação em Engenharia de Energia.

Orientador(a): Kleber Rocha de Oliveira

Rosana - SP
2023

F498a

Fini, Leonardo Fernando

Análise Preditiva do Preço de Liquidação das Diferenças no Mercado de Energia Elétrica via Algoritmos de Machine Learning / Leonardo Fernando Fini. -- Rosana, 2023

45 p.

Trabalho de conclusão de curso (Bacharelado - Engenharia de Energia) - Universidade Estadual Paulista (Unesp), Faculdade de Engenharia e Ciências, Rosana

Orientador: Kleber Rocha de Oliveira

1. Machine Learning. 2. Predição. 3. Preços. 4. Energia. 5. Mercado de Energia Elétrica. I. Título.

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
CÂMPUS DE ROSANA

LEONARDO FERNANDO FINI

ESTE TRABALHO DE GRADUAÇÃO FOI JULGADO ADEQUADO COMO PARTE DO REQUISITO PARA A OBTENÇÃO DO DIPLOMA DE "**GRADUADO EM ENGENHARIA DE ENERGIA**"

APROVADO EM SUA FORMA FINAL PELO CONSELHO DE CURSO DE GRADUAÇÃO EM ENGENHARIA DE ENERGIA

Prof. Dr. Leandro Ferreira Pinto
Coordenador

BANCA EXAMINADORA:

Prof. Dr. Kleber Rocha de Oliveira
Orientador/UNESP-Rosana

Prof. Dr. Renivaldo José dos Santos
UNESP-Rosana

Prof. Dra. Claudia Gonçalves de Azevedo
UNESP-Rosana

Junho de 2023

Dedico este trabalho,
de modo especial,
à minha família

AGRADECIMENTOS

Primeiramente, gostaria de expressar os meus sinceros agradecimentos a todos aqueles que contribuíram de alguma forma para chegar até onde eu cheguei. Agradeço grandemente aos meus pais, que nunca mediram esforços para que eu conseguisse alcançar meus sonhos e objetivos. Tudo o que conquistei e o que me tornei foram reflexos do trabalho árduo deles.

Ao meu orientador, professor e amigo Kleber, pela sua expertise e dedicação demonstradas em aulas, o que possibilitou o aperfeiçoamento da minha pessoa para enfrentar desafios dentro e fora do ambiente acadêmico, além do auxílio na conclusão deste trabalho.

Aos meus orientadores de iniciação científica, Marilaine e Wallace, por tudo o que vocês me proporcionaram: amizade, apoio, ensinamentos e muitas oportunidades, além do acolhimento no grupo VISER. Deixo também o meu agradecimento a todos os integrantes que colaboraram para a execução dos meus dois projetos durante a graduação.

Aos meus amigos que fiz durante a graduação, em especial aqueles que moraram comigo nas repúblicas, do Centro Acadêmico e da minha turma de sala, pela amizade, companheirismo e aprendizados. Vocês tornaram o período da graduação gostoso e divertido.

Por fim, agradeço a todos os colaboradores e docentes da UNESP/Rosana, pelo trabalho e dedicação em fornecer um bom ambiente de estudo para nós, alunos.

“A imaginação é mais importante que o conhecimento. O conhecimento é limitado. A imaginação envolve o mundo.”

Albert Einstein.

RESUMO

No mercado de energia elétrica, onde a especulação e a competição são intensas, prever o preço pelo qual a energia será negociada em um determinado período é um desafio que pode auxiliar na tomada de decisões dos agentes envolvidos. Um indicador utilizado nesse mercado é o Preço de Liquidação das Diferenças (PLD), que leva em consideração uma série de fatores, como oferta e demanda de energia, condições hidrológicas, restrições na transmissão, custos operacionais do sistema elétrico, entre outros. Nesse contexto, este trabalho teve como objetivo estudar o comportamento desses fatores para prever o PLD, do submercado Sudeste/Centro-Oeste, no mercado nacional de energia. Para isso, foram utilizadas ferramentas de Análise Exploratória de Dados (AED) e modelos de *Machine Learning* (ML), como Florestas Randômicas, Máquinas de Vetores de Suporte e Aumento do Gradiente. Os modelos foram desenvolvidos e validados por meio da análise de diferentes conjuntos de dados, incluindo a geração de energia de diversas fontes, dados climáticos e dados hidrológicos. A partir disso, os modelos preditivos obtiveram ótimos resultados, tendo como destaque o algoritmo Aumento do Gradiente, que atingiu o MSE no valor de 165.48 e MAPE de 6.28%, além de apresentar uma boa acurácia considerando novos dados distintos. Com isso, essa abordagem permitiu o desenvolvimento de estratégias e novas soluções computacionais para os agentes do mercado de energia elétrica no Brasil.

PALAVRAS-CHAVE: *Machine Learning*, Predição, Preços, Mercado de Energia Elétrica.

ABSTRACT

In the electricity market, where speculation and competition are intense, predicting the price at which energy will be traded in a given period is a challenge that can assist decision-making by the involved agents. An indicator used in this market is the Preço de Liquidação das Diferenças (PLD), which takes into account a series of factors such as energy supply and demand, hydrological conditions, transmission constraints, operational costs of the electrical system, among others. In this context, the aim of this project was to study the behavior of these factors to predict the PLD for the Sudeste/Centro-Oeste submarket in the national energy market. For this purpose, tools of Exploratory Data Analysis (EDA) and Machine Learning (ML) models such as Random Forests, Support Vector Machines, and Gradient Boosting were employed. The models were developed and validated through the analysis of different datasets, including energy generation from various sources, weather data, and hydrological data. As a result, the predictive models achieved excellent results, with Gradient Boosting algorithm standing out, reaching an MSE of 165.48 and MAPE of 6.28%, while demonstrating good accuracy when considering new distinct data. This approach enabled the development of strategies and new computational solutions for agents in the Brazilian electricity market.

KEYWORDS: Machine Learning, Prediction, Prices, Electric Energy Market.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ACL	Ambiente de Contratação Livre
ACR	Ambiente de Contratação Regulada
AE	Aprendizado Estatístico
AED	Análise Exploratória de Dados
AM	Aprendizado de Máquina
CCEE	Câmara de Comercialização de Energia Elétrica
CD	Ciência de Dados
CEPEL	Centro de Pesquisas de Energia Elétrica
CMO	Custo Marginal de Operação
CV	*Cross Validation*
EAR	Energia Armazenada
ENA	Energia Natural Afluente
GTB	*Gradient Tree Boosting*
IA	Inteligência Artificial
MAPE	*Mean Absolute Percentage Error*
MCP	Mercado de Curto Prazo
ML	*Machine Learning*
MSE	*Mean Squared Error*
ONS	Operador Nacional do Sistema
PCHs	Pequenas Usinas Hidrelétricas
PCTs	Pequenas Usinas Térmicas
PLD	Preço de Liquidação das Diferenças
RF	*Random Forest*
RNA	Redes Neurais Artificiais
SEB	Setor Elétrico Brasileiro
SIN	Sistema Interligado Nacional
SVM	*Support Vector Machines*
SVR	*Support Vector Regression*
UHE	Usina Hidrelétrica
UT	Usina Térmica

SUMÁRIO

1. INTRODUÇÃO

O crescimento da demanda por energia elétrica no Brasil está diretamente relacionado ao aumento das atividades de produção e consumo, conforme evidenciado nas últimas décadas. Embora esse crescimento seja essencial para impulsionar o setor produtivo e o consumo, também acarreta impactos ambientais negativos (SOUSA, 2023). Nesse contexto, a continuidade do uso de fontes não renováveis na geração de energia tem sido um ponto de debate para o crescimento econômico sustentável, além de ser fundamental suprir as necessidades energéticas do mundo de maneira a minimizar os impactos ambientais para as futuras gerações (REIS, 2023).

Com isso, destaca-se a importância das fontes de energia renováveis como uma solução para mitigar os problemas ambientais e promover a sustentabilidade na produção de energia. No mercado de energia elétrica brasileiro, há um esforço em direção a uma matriz energética mais limpa e sustentável, impulsionado por políticas governamentais e incentivos à geração de energia renovável. Fontes como energia solar, eólica e biomassa têm ganhado destaque e participação crescente na matriz energética do país (LAMPIS et al., 2022).

Além disso, o mercado de energia elétrica brasileiro passou por reformas e regulamentações visando promover a concorrência, atrair investimentos e garantir a segurança e a eficiência do fornecimento de energia. A abertura desse mercado possibilitou a participação de diversos agentes, como geradores, distribuidores, comercializadores e consumidores livres, estimulando a diversificação da matriz energética e a adoção de fontes renováveis (ALMEIDA et al., 2021). Em suma, o mercado de energia elétrica brasileiro busca equilibrar o suprimento de energia para atender às demandas de crescimento econômico, considerando a necessidade de preservar o meio ambiente e promover a sustentabilidade. A transição para uma matriz energética mais limpa e a diversificação das fontes renováveis são passos importantes para alcançar um futuro energético sustentável no Brasil (EPE, 2022).

Atrelado ao mercado de energia elétrica, o Aprendizado de Máquina (AM) pode desempenhar um papel significativo ao auxiliar na previsão da demanda e no gerenciamento eficiente da geração e distribuição de energia. Através da análise de grandes conjuntos de dados históricos, algoritmos de AM podem identificar padrões e tendências, permitindo a criação de modelos preditivos precisos para a oferta e demanda de energia. Isso auxilia os agentes do mercado a tomar decisões informadas sobre a compra e venda de energia, otimizando a eficiência operacional e minimizando os riscos financeiros. Em suma, a sua utilização no

mercado de energia elétrica brasileiro pode aumentar a eficiência, a sustentabilidade e a qualidade do fornecimento de energia (SANTOS, 2022; NUNES et al., 2023).

Diante do exposto, este trabalho visa estudar o impacto do cruzamento de diferentes fontes de dados abertos visando a predição do Preço de Liquidação das Diferenças no mercado nacional. Para essa tarefa, serão utilizados modelos de Aprendizado de Máquina (AM) e ferramentas de Análise Exploratória de Dados (AED) de modo a criar metodologias computacionais e estratégias para fins de suporte, para que agentes do setor elétrico possam respaldar decisões com base na análise inteligente dos dados.

2. OBJETIVO GERAL

O presente trabalho tem como intuito estudar metodologias baseadas em AM no cenário de previsão do Preço de Liquidação das Diferenças (submercado Sudeste/Centro-Oeste), no mercado de energia elétrica brasileiro. Essa metodologia será aplicada a dados reais coletados em diferentes repositórios abertos, tendo como "produto final" a criação de sumários de dados como gráficos, análises dos possíveis erros/resíduos, entre outros métodos de avaliação para compor conclusões/tomadas de decisões a partir dos algoritmos preditivos e dos dados analisados.

2.1. OBJETIVOS ESPECÍFICOS

A fim de conduzir este trabalho, serão contemplados os seguintes objetivos específicos:

1. Estudar algoritmos de AM e técnicas de Ciência de Dados (CD) no contexto de preço de liquidação no mercado nacional.
2. Analisar e explorar diferentes conjuntos de dados abertos, de forma a pré processá-los e adequá-los para efetiva utilização dos modelos inteligentes de predição.
3. Explorar as etapas de treinamento de cada um dos modelos de AM que serão adotados, bem como suas sensibilidades quanto a variação de parâmetros, aderência aos dados reais do problema, análise de padrões, grau de separabilidade, etc.
4. Obter as séries históricas de predição do PLD (submercado Sudeste/Centro-Oeste) para o cálculo do valor da energia a ser liquidada no mercado. Pretende-se treinar e validar os modelos a partir de quatro conjuntos de dados independentes, disponibilizadas pelo Operador Nacional do Sistema (ONS) e pela Câmara de Comercialização de Energia Elétrica (CCEE), que serão integrados a fim de melhor relacioná-los dentro de uma perspectiva comum de um *framework* de Inteligência Artificial (IA).

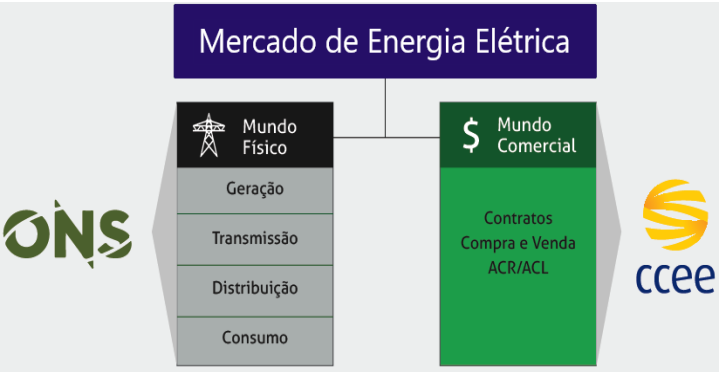
3. REVISÃO BIBLIOGRÁFICA

3.1. O MERCADO DE ENERGIA ELÉTRICA

Com base em um novo modelo institucional, e após as reformas e adaptações do Setor Elétrico Brasileiro (SEB), em 2004 foi inaugurada a CCEE, a qual é responsável pela manutenção e regulação dos Ambiente de Contratação Regulada (ACR) e do Ambiente de Contratação Livre (ACL) (MORAES, 2019).

Há dois cenários que caracterizam o mercado de energia brasileiro: o comercial e o mundo físico. O cenário comercial envolve a parte contratual da comercialização da energia, enquanto o cenário físico corresponde aos parâmetros de geração, transmissão, distribuição e consumo. Assim, atividades como definir o total de energia que será gerado por uma usina são regulamentadas pelo ONS, que atua continuamente na otimização dos recursos energéticos e na coordenação da rede interligada de transmissão, denominada Sistema Interligado Nacional (SIN) (MERCADO LIVRE DE ENERGIA, 2023; CCEE, 2023). A Figura 1 ilustra a organização do mercado de energia elétrica no Brasil.

Figura 1 - Organização do mercado de energia elétrica.



Fonte: Adaptada de (MERCADO LIVRE DE ENERGIA, 2023).

A CCEE viabiliza a comercialização de energia elétrica no SIN de ambos os ambientes de contratação (ACL e ACR), bem como a realização da liquidação financeira e a contabilização das ações realizadas no Mercado de Curto Prazo (MCP). Além disso, a CCEE concebe a ligação entre a capacidade de geração das usinas e os contratos de compra e venda, assim como a mensuração de quanto cada agente vendeu ou consumiu, além da promoção de leilões de compra/venda de energia, apuração do preço da energia que não foi contratada para ser comercializada no MCP, e o controle e fiscalização o mercado a fim de identificar desconformidades com relação à legislação (CCEE, 2023; LIMA, 2019). Portanto, nota-se que

a CCEE desempenha um papel vital no mercado ao passo que promove as transações financeiras e comerciais de energia elétrica no Brasil.

No modelo de mercado atual, têm-se negociações ocorrendo tanto no ACL como no ACR. Entretanto, há alguns fatores particulares em cada ambiente, como por exemplo o tipo de energia contratada nos leilões do ACR (Energia de Reserva, por exemplo), além da energia excedente ou o déficit dela, como uma usina pode gerar uma quantidade de energia que difere do previsto no contrato, assim como um consumidor pode usar mais ou menos energia em relação ao montante acordado. Logo, essas diferenças são liquidadas financeiramente no MCP, com base em um valor de referência denominado Preço de Liquidação das Diferenças (PLD), o qual estabelecido para cada submercado do SIN (Norte, Nordeste, Sul e Sudeste/Centro-Oeste). Além disso, o PLD é também utilizado como base nas negociações realizadas no ACL, servindo de garantia financeira e parâmetro para os agentes que atuam nesse ambiente (SANTOS et al., 2021).

3.2. PREÇO DE LIQUIDAÇÃO DAS DIFERENÇAS

Atualmente, o valor do PLD é calculado diariamente pela CCEE para cada hora do dia seguinte, levando em consideração uma série de informações e restrições.

Um dos principais elementos considerados no cálculo do PLD é o Custo Marginal de Operação (CMO). O CMO é obtido a partir de modelos computacionais criados pelo Centro de Pesquisas de Energia Elétrica (CEPEL) e utilizados pela CCEE, como o NEWAVE, DECOMP e DESSEM, que abrangem diferentes horizontes de tempo (CEPEL, 2023; CCEE, 2023).

O modelo NEWAVE é utilizado para a programação da operação energética de longo prazo, considerando um horizonte de planejamento de cinco anos e discretização mensal. Já o modelo DECOMP é utilizado para a programação da operação energética de curto prazo, abrangendo um horizonte de dois a seis meses, com discretização semanal. Por fim, o modelo DESSEM é utilizado para a programação da operação energética em tempo real, ou seja, para planejar a operação do sistema elétrico brasileiro em tempo real (SANTOS et al., 2022; CCEE, 2023).

A Figura 2 demonstra como é representado a otimização desses modelos.

Figura 2 – Modelos computacionais para o planejamento energético.

Fonte: Adaptada de (CCEE, 2023).

Esses modelos computacionais levam em consideração uma série de informações para calcular o CMO, tais como a situação atual dos reservatórios, a demanda de energia, o preço dos combustíveis nas usinas térmicas, a previsão de chuvas e a entrada de novos projetos de geração de energia. Sabendo disso, são estabelecidos os valores do PLD para cada hora do dia seguinte e são importantes para o mercado de energia elétrica, pois influenciam diretamente o preço da energia para os agentes consumidores e comercializadores (CAMPBELL, 2022; CEPEL, 2023).

Dessa forma, os modelos computacionais utilizados pela CCEE, desempenham um papel crucial na determinação do PLD, fornecendo informações precisas e atualizadas sobre os custos de geração e a operação do sistema elétrico brasileiro. Essas informações são fundamentais para a tomada de decisões dos agentes do mercado e para a gestão eficiente do setor de energia elétrica.

3.3. APRENDIZADO DE MÁQUINA: APLICAÇÕES

A tarefa de estimar a energia futura a ser gerada é um ponto crucial na evolução do mercado, seja qual for o modelo de produção de energia adotado (DUTT et al., 2022). Além disso, cabe ainda ressaltar que há uma grande quantidade de dados que o setor recebe diariamente, tendo a necessidade do emprego de técnicas que possam lidar com padrões complexos utilizando redes neurais, por exemplo (KOPILER, 2019).

Diversas pesquisas vêm sendo realizadas nesse âmbito, como em (RODRIGUES, 2009), onde o autor apresenta um estudo de predição de preço no mercado spot de energia elétrica no Brasil, com a finalidade de apoiar os agentes na tomada de decisão na

comercialização de energia. Já em (CARRIJO, 2019), o autor utilizou a predição do potencial energético por unidade de área em uma determinada região para avaliar a efetividade das Redes Neurais Artificiais (RNA) relacionadas à dados de satélite, analisando assim as relações entre inúmeros índices de vegetação e potencial energético.

No caso do trabalho (ARAUJO et al., 2018), os autores empregaram uma rede neural do tipo *Perceptron* Multicamada, tendo partido da predição da radiação solar global em um determinado município, utilizando modelos de AM e com variáveis exploradas a radiação solar global, velocidade do vento, temperatura, umidade e ponto de orvalho. De maneira análoga, em (JLIDI et al., 2023), os autores realizaram a predição da radiação solar e potencial fotovoltaico em um local de interesse aplicando redes neurais artificiais, onde caracterizaram as vantagens existentes da aplicação de tal método.

Outro método de AM bem estabelecido no contexto de energia é o *Random Forest* (RF), utilizado em (HUANG et al., 2016), a fim de prever a carga elétrica gerada a curto prazo, comparando o resultado alcançado com os modelos *Support Vector Regression* (SVR) e RNA. Esses modelos também foram empregados em (LEME et al., 2020) em adição à técnica *Gradient Boosting* (GB) para predição da carga elétrica no Brasil, tendo os autores constado uma maior acurácia por parte do modelo GB. Já em (MEI et al., 2014), um modelo baseado em RF foi usado para estimar o preço no mercado de eletricidade de Nova York e, em (ABUELLA et al., 2017), os autores combinaram o RF e SVM para criar um modelo inteligente de predição de energia solar.

3.4. TÉCNICAS DE APRENDIZADO DE MÁQUINA ADOTADAS

Diante dos trabalhos mencionados na seção anterior, é possível observar que a resolução de problemas do setor energético pode ser realizada com sucesso a partir de métodos preditivos de AM. Além disso, esses modelos possibilitam o fomento de pesquisas na área, levando ao desenvolvimento de novas tecnologias e implementação de políticas públicas mais efetivas.

3.4.1. ENSEMBLE

Ensembles (ou agrupamento de regressores) são métodos para aperfeiçoar a tarefa de aprendizado computacional. O seu conceito parte de que a melhor decisão é proveniente da opinião de um grupo experiente no assunto do que a de apenas um indivíduo (LIMA et al.,

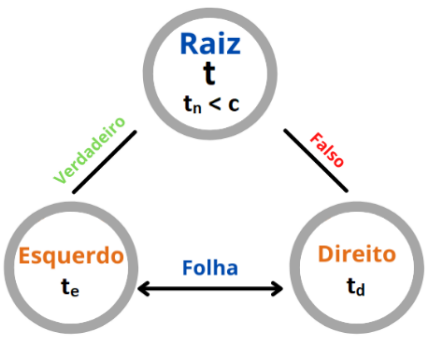
2019). No âmbito dos métodos preditivos, esse “grupo experiente” é descrito como regressores que foram determinados no decorrer do processo de criação do preditor final.

Presentes em várias áreas do conhecimento, bem como na literatura de predição energética, encontram-se inúmeras estratégias de ensemble sendo amplamente empregadas (LEME et al., 2020; PAULA et al., 2020). Dentre elas, destacam-se as abordagens *Bagging, Boosting*, e técnica de subespaço aleatório (RIBEIRO et al., 2020). Cada método, mesmo que possua suas características próprias, busca explorar o bom comportamento local de cada previsor e, dessa forma, ampliar a precisão e a segurança da previsão a partir da combinação de diferentes previsores locais (CARON, 2022). O pipeline usual de ensemble engloba os seguintes passos:

- Cada previsor local obtém os dados de entrada do problema.
- Cada previsor *i* cria um mapeamento F_i: $\mathbb{R}_m \rightarrow \mathbb{R}$ da saída, com início de m entradas.
- Cada saída F_i de cada previsor i é multiplicada por um peso w_i, em que o somatório de todos os pesos obrigatórios às saídas deve ser igual a 1.
- As saídas ponderadas são então somadas para constituir a solução final *ensemble*.

Os modelos *ensembles* que serão adotados neste trabalho tomam como base as árvores de decisão/regressão, que se assemelham à morfologia de uma árvore, onde cada ramificação de um galho (um nó) separa os dados em grupos/valores diferentes partindo de um conjunto de regras específicas de decisão (vide Figura 3). Analisando o topo da árvore, é possível notar o nó raiz, que contém todos os dados de treinamento t, além da regra de subdivisão $t_n < c$. Já os nós resultantes da subdivisão (“folhas”) são tais que t_e e t_c formam as amostras de treinamento para os próximos níveis da árvore (LIER, 2015).

Figura 3 – Representação ilustrativa da árvore de regressão.



Fonte: Adaptada de (LIER, 2015).

3.4.1.1. FLORESTA ALEATÓRIA

Floresta Randômica, ou ainda *Random Forest* (RF), é um método do tipo *ensemble*, ou seja, um conjunto de estimadores que induz à constituição de seus próprios aprendizes e técnicas, onde os aprendizes base são todas as árvores de classificação/regressão. (CAPITAINE et al., 2021). De uma forma mais geral, o método RF baseia-se na realização de três etapas básicas:

- Gerar n conjuntos de amostras da base de dados de treinamento.
- Para cada amostra, formar uma árvore de regressão (sem ajuste) com a seguinte alteração: em cada nó, constitui-se uma amostra aleatória p das variáveis de entrada de toda base de dados de treino no qual escolhe-se a melhor divisão dessas variáveis, com $p < m$, e m caracterizando o número de todas as variáveis da base.
- Prever a nova saída realizando o cálculo da média das saídas de n árvores de regressão quando novas variáveis são introduzidas ao modelo.

3.4.1.2. AUMENTO DO GRADIENTE

O Aumento do Gradiente, também conhecido como *Gradient Tree Boosting* (GTB) é classificado como um método *ensemble* baseado em árvores de regressão. Esse método emprega a estratégia de *boosting* ao invés de *bagging* (como no caso do RF). Com isso, a estratégia do *boosting* é uma melhoria da técnica de *bagging*, onde o *boosting* qual baseia-se em treinar diversos submodelos com subamostras aleatórias no processo de treinamento, e associá-las para conseguir um desempenho menos "particularizado", portanto, com menos sobreajuste (*overfitting*) (PEREIRA, 2018).

O modelo GTB gera uma árvore de regressão simples e utiliza uma variação do gradiente descendente para ir avaliando as árvores nas repetições com a utilização de uma função de custo. Explicitamente, a decisão do GTB é concedida em termos da somatória das estimativas das árvores. Sendo assim, três pontos chaves contendo o princípio clássico do GTB são (BROWNLEE, 2016):

- Uma função de custo a ser otimizada – *loss function* – (combinação da metaheurística do gradiente descendente).
- Modelos preditivos para gerar as predições "*weak learns*" (Exemplo: árvore de decisão).
- Um modelo complementar de "*weak learners*" para minimização da função de custo.

3.4.2. MÁQUINAS DE VETORES DE SUPORTE

O método denominado Máquinas de Vetores de Suporte, ou ainda *Support Vector Machines* (SVM), foi introduzido a partir de uma nova linha de Aprendizado de Máquina, chamada Aprendizado Estatístico (AE). O AE foi concebido para resolver problemas cuja quantidade de dados é pequena ou até mesmo nula, características geralmente presentes em diversas aplicações reais.

A ideia central do SVM é gerar superfícies de decisão para separar instâncias de classes distintas: ele cria um hiperplano ótimo, que potencializa a margem, que é a distância entre os vetores de suporte das classes distintas (vide Figura 4) (VELASCO et al., 2018).

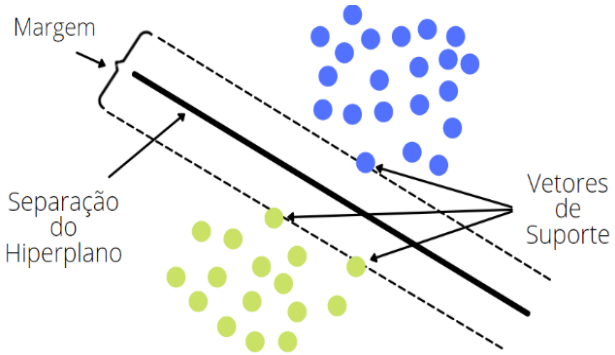
A fim de gerar uma superfície de decisão em problemas que não são capazes de ser separados linearmente, o SVM pode também utilizar funções *kernel*, que geram transformações lineares no vetor de atributos com o objetivo de formar dimensões maiores, em que as classes podem ser mais bem separadas por um hiperplano (TAYLOR, 2014). Com isso, os problemas de regressão podem ser resolvidos a partir de adaptações do SVM clássico, com base na utilização de uma função de perda que, através de um regularizador, é minimizada. Nesses casos de regressão, o SVM é normalmente denominado de SVR (*Support Vector Regression* – Regressão Vetorial de Suporte) (GOMES, 2018).

Dessa maneira, o problema resume-se em descobrir uma função não linear f, conforme Equação (1), com o intuito de diminuir o erro da previsão em relação ao conjunto de treinamento:

$$f(X) = max(0, |f(X_i) - y_i| - \varepsilon), \quad (1)$$

Em que o parâmetro ε controla o erro permitido a mensuração (WANG et al., 2017).

Figura 4 – Representação do hiperplano com a margem de separação.



Fonte: Adaptada de (BORGES, 2023).

4. MATERIAIS E MÉTODOS

Essa seção é destinada à apresentação do procedimento experimental adotado para a realização do presente trabalho, que consiste na previsão do PLD (submercado Sudeste/Centro-Oeste), como descrito na Seção 2.

4.1. REPOSITÓRIO DE DADOS

Para a problemática do trabalho, foram utilizados dados horários do PLD (submercado Sudeste/Centro-Oeste) fornecidos pela CCEE, bem como dados referentes a energia da região Sudeste/Centro-Oeste do país fornecidos pelo ONS como (ONS, 2023):

- Custo Marginal de Operação ($\frac{R\$}{MWh}$): representa o custo unitário de energia necessário para suprir o acréscimo de uma unidade de carga no SIN;
- Carga Horária ($\frac{MWh}{h}$): representa o perfil de consumo de energia elétrica;
- Geração das Usinas Hidráulicas e das Pequenas Centrais Hidráulicas (MWmed): representa a geração de energia provenientes de aproveitamentos hidráulicos;
- Geração das Usinas Térmicas e das Pequenas Usinas Térmicas (MWmed): representa a geração de energia através de fontes térmicas;
- Geração Eólica (MWmed): representa a geração de energia através do aproveitamento dos ventos;
- Geração Solar Fotovoltaica (MWmed): representa a geração de energia através da conversão da luz solar em eletricidade.
- Energia Natural Afluente Bruta (MWmed): corresponde à energia gerada pelo reservatório e é determinada multiplicando as vazões naturais que chegam aos reservatórios pelas produtividades, considerando 65% dos volumes utilizáveis.
- Energia Armazenada (MWmês): corresponde à energia potencial contida no volume de água armazenado nos reservatórios, que pode ser convertida em eletricidade tanto na usina em si quanto em todas as usinas localizadas a montante da sequência de aproveitamentos hidrelétricos.

As unidades de medida MWmed e MWmês correspondem a potência média em *megawatts* ao longo de um determinado período e ao longo de um mês, respectivamente.

4.2. LIMPEZA E PRÉ-PROCESSAMENTO DOS DADOS

Para não ocorrer erros ou incongruências em nossos modelos preditivos, primeiramente houve a necessidade de realizar a técnica *data cleaning*, que consiste no tratamento dos dados. Esse tratamento é realizado através da "limpeza" de dados que possuem valores faltantes ou inconsistentes.

Além disso, existem alguns modelos preditivos como Redes Neurais Artificiais, Regressão Linear/Logística, KNN (K-Nearest Neighbours) e SVM, onde é necessário a realização da padronização dos dados. Esse procedimento consiste em aproximar os dados para uma escala pré-definida (geralmente em um intervalo entre -1 e 1), fazendo com que tenham a mesma ordem de grandeza e possibilitando a otimização do processo e a performance dos algoritmos. Como um dos modelos preditivos escolhidos para este trabalho foi o SVM, utilizou-se a padronização *Standardization*, dada pela fórmula *Z-Score* representada pela Equação (2):

$$z = \frac{x - \mu}{\sigma} \tag{2}$$

Onde x é representado como os dados, μ é a média e σ o desvio padrão desses dados. Além disso, esse algoritmo pode apresentar dificuldades quando aplicados em um grande *database*, tanto em relação ao tempo de treinamento como também no resultado da previsão (DUTRA, 2021).

Como a nossa problemática possui uma grande quantidade de dados, após a realização de diversos testes e com resultados insatisfatórios através do algoritmo SVM, realizou-se a utilização de uma quantidade menor de dados, bem como realizado pela comunidade científica (EXCHANGE, 2023). Sendo assim, para esse modelo preditivo em específico, utilizou-se dados de 01/11/2021 a 30/06/2022 para treinamento e dados de 01/07/2022 a 31/12/2022 para fins de validação do modelo, resultando em um total de 10.224 dados.

4.3. ANÁLISE EXPLORATÓRIA DOS DADOS

Para obter melhores resultados por parte dos modelos preditivos, foi necessário a etapa de AED disponíveis no *database.* Essa técnica consiste em ilustrar o comportamento das variáveis em estudo, bem como melhor relacionar e apresentar padrões que dificilmente seriam notados sem auxílio computacional. Tais análises e visualizações foram compostas por:

- Mapa de calor: analisar o grau de linearidade entre as variáveis de acordo com o coeficiente de *Pearson*, tendo valores entre -1 e 1.
- Gráfico de dispersão: observar graficamente o comportamento dos dados de duas variáveis (a de maior correlação com a variável *target*).
- Análise descritiva: explorar as métricas das variáveis de forma resumida.
- *Box Plot*: representar o comportamento estatístico dos dados e os pontos fora da curva (*outliers*).
- *Line Plot*: analisar graficamente as informações mensais, semanais e horárias da variável *target* "PLD".

4.4. ENGENHARIA DE RECURSOS: CRIAÇÃO DE NOVAS VARIÁVEIS

Com a finalidade de melhorar a acurácia dos modelos preditivos escolhidos, tem sido fundamental a aplicação da engenharia de recursos nas variáveis preditoras, ou seja, criar novas variáveis a partir das já existentes (CHATZIS et al., 2018). Sabendo disso, foram aplicados métodos estatísticos como soma, subtração, divisão, multiplicação e média nas variáveis preditoras, que a correlação com a variável *target* "PLD" dessas novas variáveis (e das já existentes).

4.5. AJUSTE DE HIPERPARÂMETROS E MÉTRICA DE VALIDAÇÃO

Conforme descrito anteriormente, os modelos preditivos estudados foram: *Random Forest*, *Gradient Boosting* e *Support Vector Regression*. Esses algoritmos de AM apresentam diversos hiperparâmetros que podem ser ajustados para que o preditor resultante tenha uma melhor acurácia e aderência aos dados de treino. Com isso, o algoritmo *Random Search* (busca aleatória) tem sido, na prática, uma saída bastante efetiva para realizar combinações em universo de hiperparâmetros. Assim, neste trabalho, utilizou-se esse algoritmo para regulagem de parâmetros, o que possibilitou alcançar uma melhor precisão por parte dos preditores (BERGSTRA et al., 2012).

Para a validação dos resultados, foram realizadas análises visuais e quantitativas a partir dos dados de validação das bases de dados, além da aplicação da métrica *Mean Absolute Percentage Error* (MAPE) e *Mean Squared Error* (MSE), através da Equação (3) e Equação (4), respectivamente:

$$MAPE = \frac{1}{n}\sum_{i=0}^{n}\frac{|Yi - \bar{Y}i|}{|Yi|}\ x\ 100 \quad (3)$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Yi - \bar{Y}i)^2 \quad (4)$$

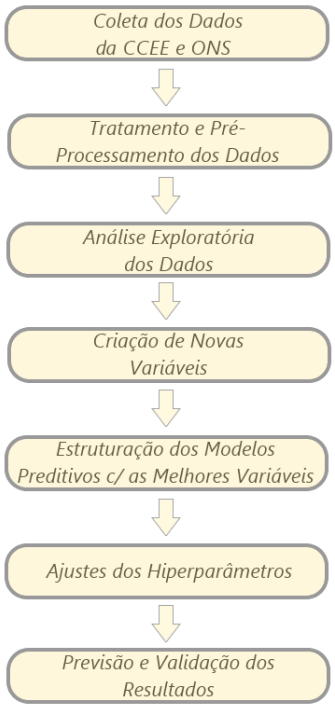
onde *n* é o número de amostras, *Yi* os valores reais e $\bar{Y}i$ os valores preditos (ADHIKARI, 2013).

Os ajustes de hiperparâmetros, bem como a acurácia desses modelos foram detalhados na Seção 5. Resultados e Discussões.

4.6. PIPELINE DE PROCESSAMENTO DE DADOS

Dada a problemática de pesquisa deste trabalho e a discussão sobre os modelos e as técnicas de AM adotadas, a Figura 5 exibe uma representação da metodologia computacional que foi implementada para a tarefa de predição do "PLD".

Figura 5 – Pipeline de processamento dos dados deste trabalho.



Fonte: Elaborado pelo próprio autor.

5. RESULTADOS E DISCUSSÕES

Neste capítulo são apresentados os resultados obtidos referentes as metodologias propostas no trabalho. Além disso, são demonstradas comparações entre os *scores* alcançados pelos modelos adotados.

Os dados utilizados no trabalho foram coletados no período de 01/01/2021 a 31/12/2022, considerando o período abrangido pelo ano completo de 2021 e os primeiros seis meses de 2022, utilizado para o treinamento dos modelos preditivos, e a metade do ano de 2022 para o fim de validação dos modelos pós treinamento. A Figura 6 ilustra uma amostra das variáveis do *data-base* pronto.

Figura 6 – Amostra da base de dados utilizada para a predição do PLD.

Index	CMO	Carga	Geração UHE	Geração PCHs	Geração UT	Geração PCTs	Geração Eólica	Geração Fotov.	ENA Bruta	EAR Verificada	PLD
2021-01-01 00:00:00	209.25	33498.4	24703.4	1319	4517	825	3	0	1715.3	1615.33	204.37
2021-01-01 01:00:00	207.67	33183	24503	1335	4517	825	1	0	1715.3	1615.33	206.82
2021-01-01 02:00:00	206.83	32630.7	24455.1	1337	4517	822	1	0	1715.3	1615.33	204.61
2021-01-01 03:00:00	209.52	31738.5	24978.3	1335	4517	819	0	0	1715.3	1615.33	209.2
2021-01-01 04:00:00	207.78	31233.2	24331.1	1327	4622	819	1	0	1715.3	1615.33	207.19
2021-01-01 05:00:00	201.89	29944.3	23161	1313	4622	822	1	0	1715.3	1615.33	201.99
2021-01-01 06:00:00	185.86	28583.6	23264.9	1313	4562	816	0	69	1715.3	1615.33	195.53
2021-01-01 07:00:00	185.79	28589.8	23206.8	1308	4412	809	1	244	1715.3	1615.33	195.4
2021-01-01 08:00:00	185.73	29052.2	23285.4	1307	4412	801	0	403	1715.3	1615.33	195.5
2021-01-01 09:00:00	191.34	29675.1	23569.5	1307	4312	802	0	518	1715.3	1615.33	198.8

Fonte: Elaborada pelo próprio autor.

Em relação ao tratamento e limpeza dos dados, a Tabela 1 mostra a quantidade de dados nulos ou ausentes de cada variável.

Tabela 1 – Análise dos valores nulos ou inconsistentes das variáveis.

Variáveis	**Quantidade de Valores**	**Porcentagem de Valores (%)**
Custo Marginal de Operação	432	2.47
Carga Horária	0	0
Geração de Usinas Hidráulicas	432	2.47
Geração de Pequenas Usinas Hidráulicas	436	2.49
Geração de Usinas Térmicas	432	2.47
Geração de Pequenas Usinas Térmicas	439	2.51
Geração Eólica	450	2.57
Geração Solar Fotovoltaica	452	2.58
Energia Natural Afluente Bruta	0	0
Energia Armazenada	0	0
Preço de Liquidação das Diferenças	0	0

Fonte: Elaborada pelo próprio autor.

Conforme observado na Tabela 1, a quantidade de valores ausentes de cada variável é bem pequena comparado ao tamanho do *database*, que corresponde a 17.520 dados. Existem diversas técnicas para o "tratamento" desses dados, sendo que para o caso proposto e após a realização de diversos testes, constatou-se que a melhor acurácia dos modelos preditivos foi com o preenchimento pela mediana da variável em questão, conforme ilustrado pela Figura 7.

Figura 7 – Representação do preenchimento dos dados ausentes.

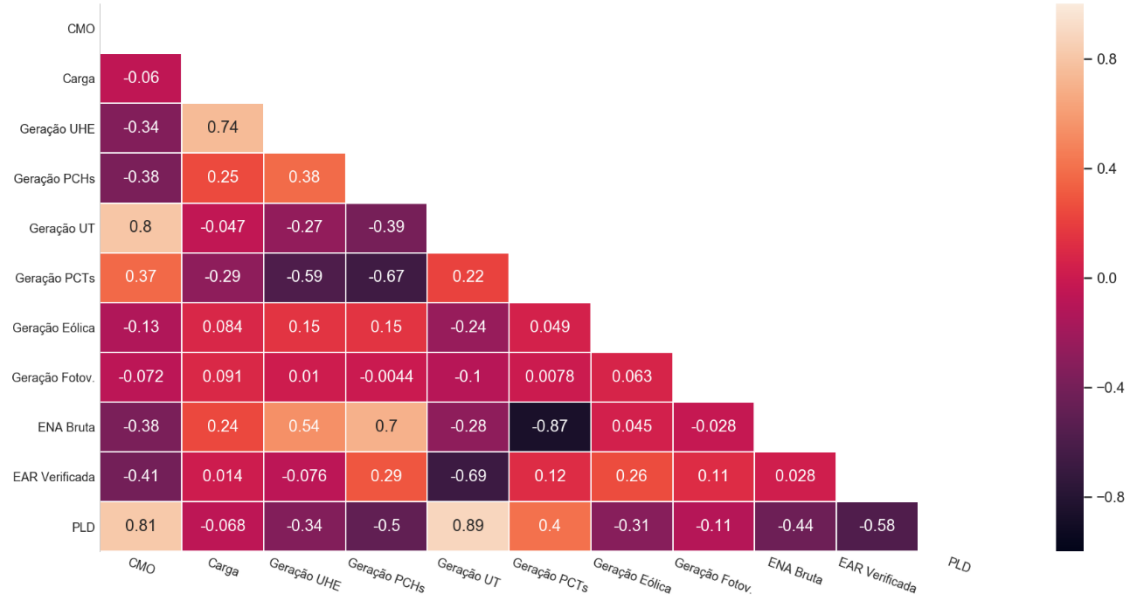
Index	CMO	Carga	Geração UHE	Geração PCHs	Geração UT	Geração PCTs	Geração Eólica	Geração Fotov.	ENA Bruta	EAR Verificada	PLD
2021-11-15 06:00:00	66.24	29197.1	21474.1	nan	4069.17	1967	nan	115	1208.65	1614.65	71.53
2021-11-15 07:00:00	65.96	30431.7	20824.7	nan	4098.4	1961	nan	419	1208.65	1614.65	71.16
2021-11-15 08:00:00	65.92	31810.5	20803.4	1963	4098.4	1941	nan	629	1208.65	1614.65	71.12
2021-11-15 09:00:00	66.36	32999	21584.3	1973	4098.4	1932	nan	731	1208.65	1614.65	71.86
2021-11-15 10:00:00	67.08	34160.4	22406.2	1990	4098.4	1921	nan	778	1208.65	1614.65	72.29
2021-11-15 11:00:00	67.13	35032.5	22813.1	1988	4098.4	1917	nan	803	1208.65	1614.65	72.34
2021-11-15 12:00:00	67.2	35465.8	22823.6	1983	4098.4	1914	nan	791	1208.65	1614.65	72.5
2021-11-15 13:00:00	67.38	35649.6	23115.8	1986	4006.61	nan	nan	744	1208.65	1614.65	72.5
2021-11-15 14:00:00	67.38	35838.8	23330.1	1996	3932.4	nan	nan	689	1208.65	1614.65	72.39

Index	CMO	Carga	Geração UHE	Geração PCHs	Geração UT	Geração PCTs	Geração Eólica	Geração Fotov.	ENA Bruta	EAR Verificada	PLD
2021-11-15 06:00:00	66.24	29197.1	21474.1	1754	4069.17	1967	5	115	1208.65	1614.65	71.53
2021-11-15 07:00:00	65.96	30431.7	20824.7	1754	4098.4	1961	5	419	1208.65	1614.65	71.16
2021-11-15 08:00:00	65.92	31810.5	20803.4	1963	4098.4	1941	5	629	1208.65	1614.65	71.12
2021-11-15 09:00:00	66.36	32999	21584.3	1973	4098.4	1932	5	731	1208.65	1614.65	71.86
2021-11-15 10:00:00	67.08	34160.4	22406.2	1990	4098.4	1921	5	778	1208.65	1614.65	72.29
2021-11-15 11:00:00	67.13	35032.5	22813.1	1988	4098.4	1917	5	803	1208.65	1614.65	72.34
2021-11-15 12:00:00	67.2	35465.8	22823.6	1983	4098.4	1914	5	791	1208.65	1614.65	72.5
2021-11-15 13:00:00	67.38	35649.6	23115.8	1986	4006.61	2722	5	744	1208.65	1614.65	72.5
2021-11-15 14:00:00	67.38	35838.8	23330.1	1996	3932.4	2722	5	689	1208.65	1614.65	72.39

Fonte: Elaborada pelo próprio autor.

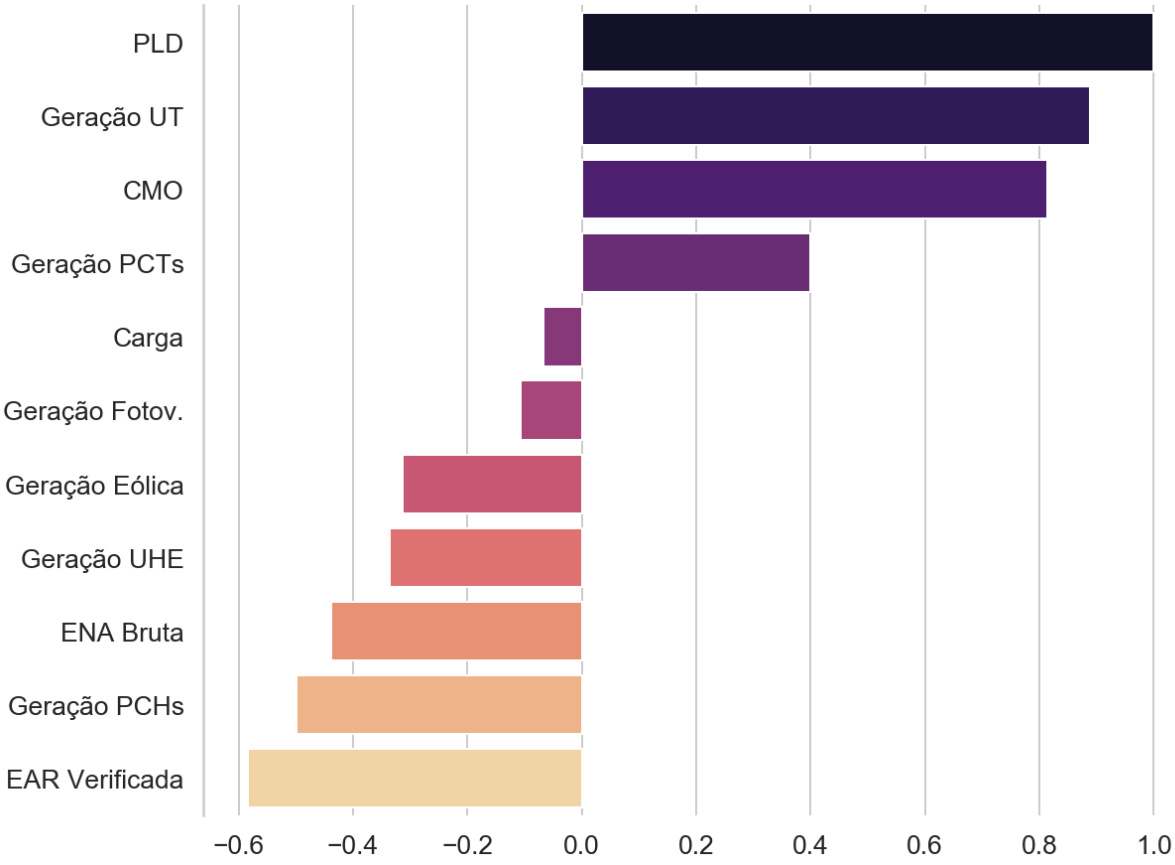
Feito isso, a fim de analisar a correlação das variáveis em questão, realizou-se primeiramente o mapa de calor e um gráfico que evidencia de forma simplificada a correlação das variáveis previsoras com a variável *target* “PLD”, onde são representados pelas Figuras 8 e 9, respectivamente.

Figura 8 – Mapa de calor das variáveis para a predição do PLD.



Fonte: Elaborada pelo próprio autor.

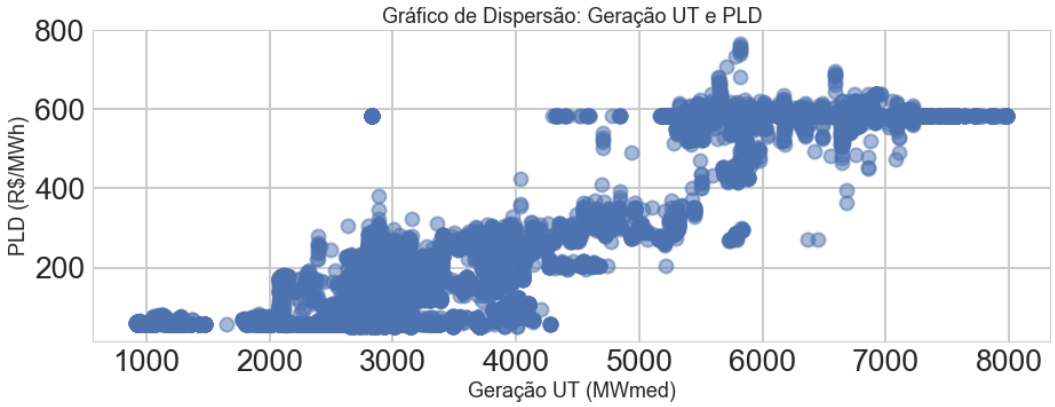
Figura 9 – Mapa de calor das variáveis relacionadas com a variável *target* “PLD”



Fonte: Elaborada pelo próprio autor.

Os valores numéricos do mapa correspondem ao coeficiente de *Pearson* e, através dos dois gráficos ilustrados acima, observa-se que a variável “Geração UT” (Geração de Usinas Térmicas), possui uma forte correlação com a variável *target* “PLD”, assim como “CMO”, o que possibilitam alcançar uma boa acurácia por parte dos modelos preditivos. A correlação da variável “Geração UT” com a variável *target* pode ser exemplificada pelas condições climáticas, em específico, a ausência de precipitação na região sudeste, fazendo com que há um despacho das usinas térmicas movida a combustíveis fósseis e com isso, o aumento do “PLD” e consequentemente da fatura de energia elétrica. Para analisar melhor a correlação dessas duas variáveis, pode-se gerar um gráfico de dispersão, conforme ilustrado pela Figura 10.

Figura 10 – Gráfico de dispersão para análise da correlação “Geração UT” e “PLD”.

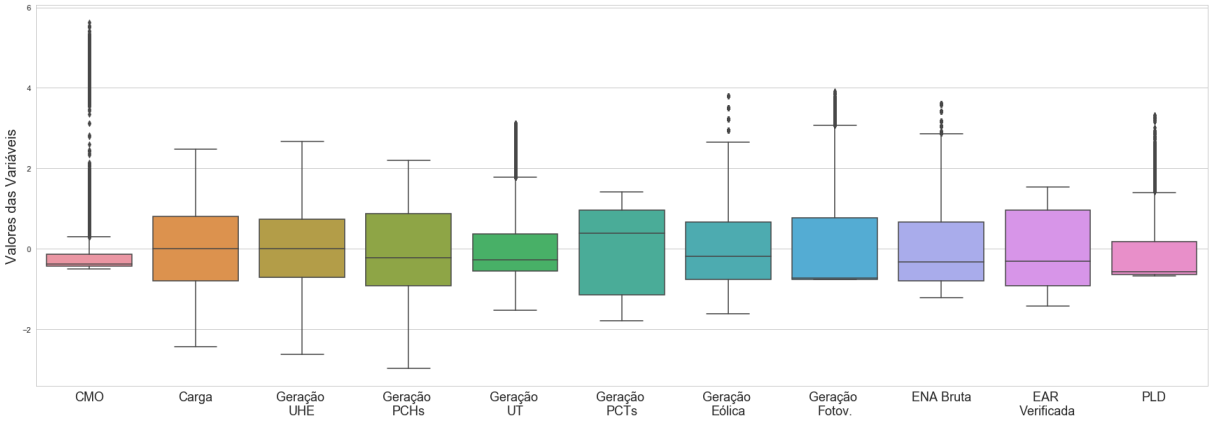


Fonte: Elaborada pelo próprio autor.

Sabendo disso, é possível analisar graficamente o comportamento dos dados, sendo que há uma correlação positiva dos dados, isto é, à medida que “Geração UT” aumenta, é esperado um aumento da variável *target* “PLD”.

Como proposto, também foi realizado a análise descritiva das variáveis e o gráfico *Box Plot*, a fim de entender a estatística dos dados, como a média, desvio padrão, valor máximo e mínimo e os quartis de porcentagem dos dados. Para a realização do gráfico *Box Plot*, foi necessário o escalonamento das variáveis, tendo em vista que as variáveis “ENA Bruta” e “EAR Verificada” apresentaram valores com escala superior as demais variáveis. O gráfico *Box Plot* é representado conforme a Figura 11.

Figura 11 – *Box Plot* das variáveis do *database* para a previsão do PLD.

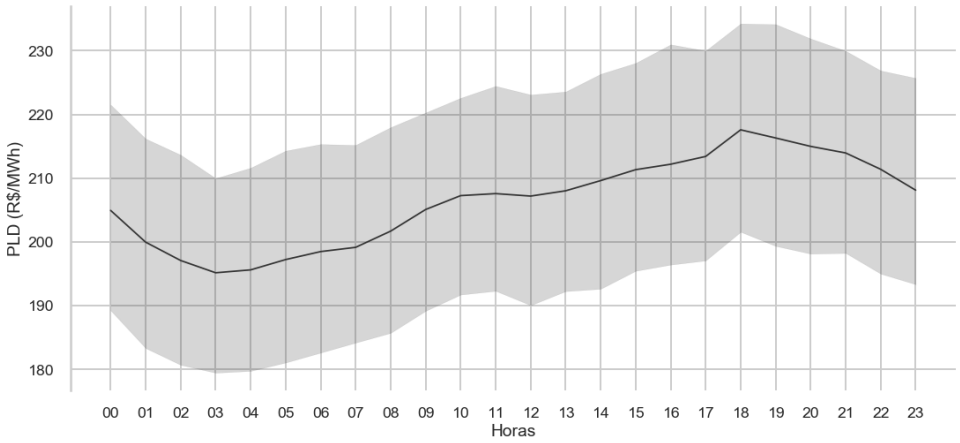


Fonte: Elaborada pelo próprio autor.

Analisando a Figura 11, nota-se os *outliers* das variáveis, que são dados discrepantes e localizados nos pontos extremos do diagrama, conforme analisando as variáveis “CMO”, “Geração UT”, “Geração Eólica”, “Geração Fotov.”, “ENA Bruta” e “PLD”. Para esse caso, os *outilers* foram mantidos, tendo em vista que, após a realização de testes, os mesmos não impactaram negativamente nos resultados.

Por fim, realizou-se análise através dos gráficos *Line Plot*, que consiste em retratar o comportamento da variável *target* de maneira horária, semanal e mensal, sendo representados pelas Figuras 12, 13 e 14, respectivamente.

Figura 12 – Gráfico *Line Plot* horário da variável *target* “PLD”.

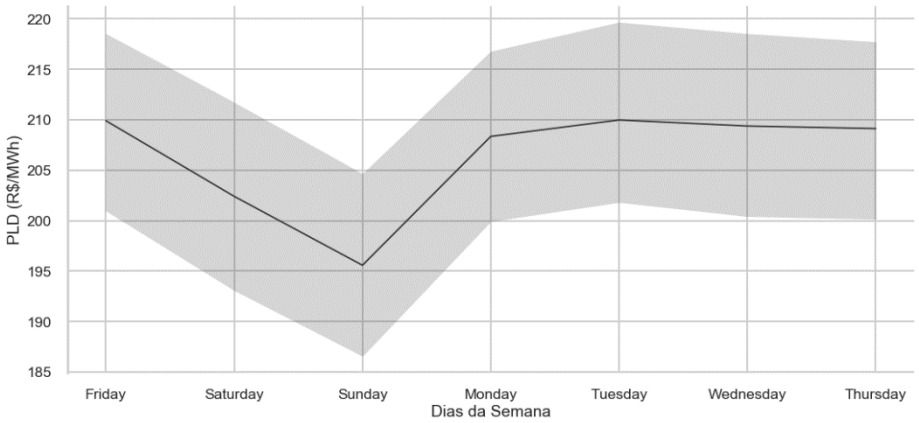


Fonte: Elaborada pelo próprio autor.

Como observado acima, percebe-se que o crescimento acentuado do “PLD” é a partir das 07:00 e a diminuição a partir das 18:00. Pode-se realizar uma analogia com o funcionamento das indústrias e comércios na região, de maneira que há um aumento do consumo de energia

elétrica e consequentemente sua geração e preço. Da mesma forma, tem-se para a série histórica semanal, ilustrada conforme a Figura 13.

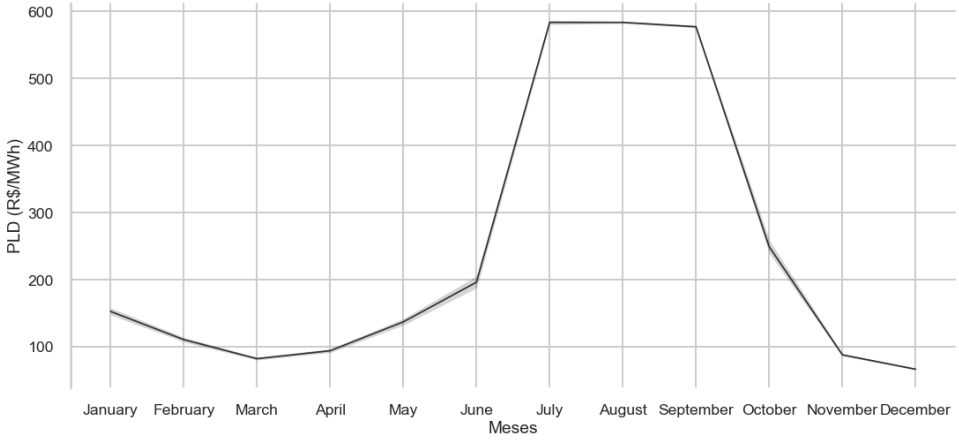
Figura 13 – Gráfico *Line Plot* semanal da variável *target* “PLD”.



Fonte: Elaborada pelo próprio autor.

Nota-se que os valores do “PLD” são maiores nos dias úteis, provavelmente pelo funcionamento das indústrias e comércios durante a semana, assim como o comportamento horário. Por fim, tem-se a representação em forma mensal, conforme representado pela Figura 14.

Figura 14 – Gráfico *Line Plot* mensal da variável *target* “PLD”.



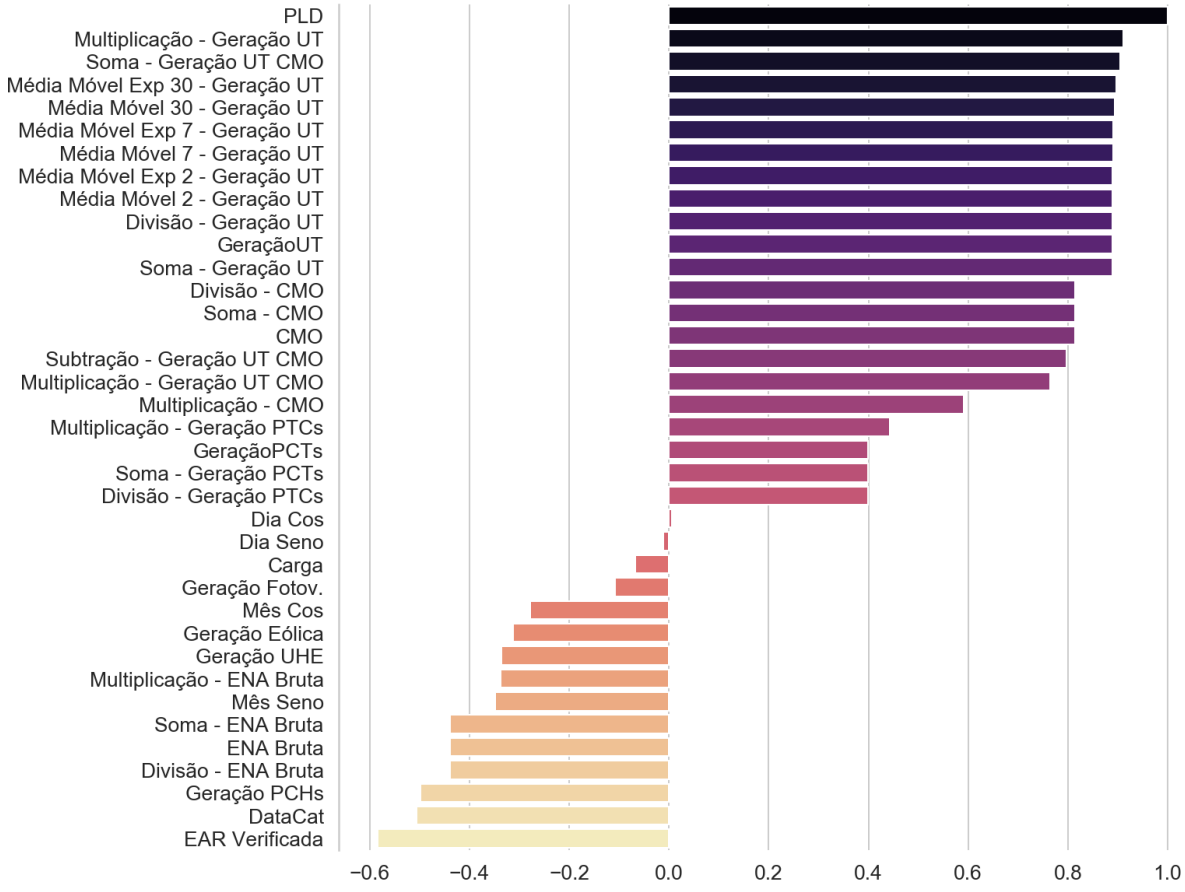
Fonte: Elaborada pelo próprio autor.

Com isso, observa-se um crescimento do “PLD” a partir de abril e uma diminuição a partir de setembro, tendo em vista da estiagem enfrentada na região. Esse evento faz com o que os níveis dos reservatórios fiquem em estado crítico de operação e que, consequentemente, há o despacho das usinas térmicas para suprir ou, ao menos, amenizar o problema da demanda por

energia elétrica. O despacho das usinas térmicas movidas a combustíveis fósseis encarece o valor do “PLD”, levando em conta o preço do barril de petróleo.

Além disso, através da técnica de engenharia de recursos, foram criadas 29 novas variáveis genéricas, totalizando 39 (com as 10 já existentes) para a predição do preço de liquidez, que pode ser observado conforme a Figura 15. Vale ressaltar que cada algoritmo de AM se adaptou com um número maior ou menor de variáveis para a previsão.

Figura 15 – Correlação das novas variáveis (e das já existentes) preditoras com a variável *target*.



Fonte: Elaborada pelo próprio autor.

5.1. FLORESTA ALEATÓRIA

Conduzida todas as etapas anteriores, primeiramente, foi realizado o ajuste de hiperparâmetros do modelo *Random Forest*, com número de iterações (k = 20) e validação cruzada/*cross validation* (cv = 4), e com todas as 13 variáveis preditoras que possuíram maior correlação com a *target*. Com isso, os melhores parâmetros obtidos foram:

- Número de árvores da floresta (*n_estimators*): 55.
- Número mínimo de amostras necessárias para dividir um nó interno (*min_samples_split*): 2.

Logo após, através da Fórmula (3) e (4), foram calculados o MAPE e o MSE do algoritmo preditivo, o qual resultou em 7.08% e 185.35, respectivamente.

5.2. AUMENTO DO GRADIENTE

Em seguida, foi ajustado os parâmetros do modelo *Gradient Boosting*, com número de iterações (k = 20) e validação cruzada/*cross validation* (cv = 5), e com as 14 variáveis preditoras que tiveram maior correlação com a *target*. Sendo assim, os melhores parâmetros obtidos foram:

- Número de recursos a serem considerados ao procurar a melhor divisão (*max_features*): 'auto'.
- Número de árvores da floresta (*n_estimators*): 100.
- Número mínimo de amostras necessárias para dividir um nó interno (*min_samples_split*): 2.

De maneira análoga ao modelo *Random Forest*, também foram calculados o MAPE e o MSE, resultando em um valor de 6.28% e 165.48, respectivamente.

5.3. MÁQUINAS DE VETORES DE SUPORTE

Enfim, foi realizado o ajuste do modelo *Support Vector Regressor* (SVR), com número de iterações (k = 20) e validação cruzada/*cross validation* (cv = 5), e com as 13 variáveis preditoras que possuíram maior correlação com a *target*, pois atingiu a melhor assertividade do modelo. Com isso, os melhores parâmetros encontrados foram:

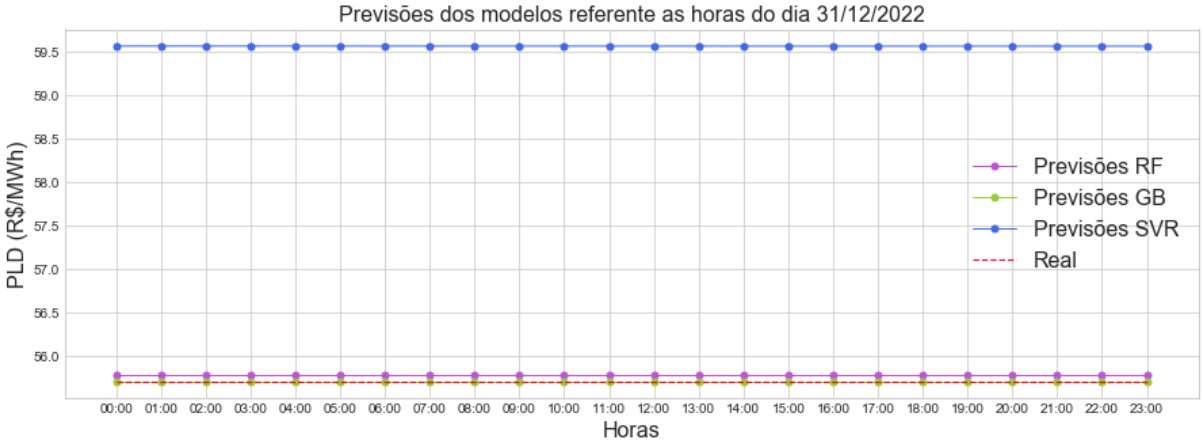
- Função kernel (*kernel*): 'rbf'.
- Coeficiente de kernel (*gamma*): 'auto'.
- Parâmetro de regularização (*C*): 0.0001.
- Número máximo de iterações (*max_iter*): 2000.
- Grau da função kernel (*degree*): 12.

O cálculo do MAPE e MSE obtidos por esse modelo preditivo foi de 9.94% e 238.19, respectivamente.

5.4. PREDIÇÃO DO PREÇO DE LIQUIDAÇÃO DAS DIFERENÇAS

Por fim, para uma melhor visualização, foi realizado a construção de um gráfico em uma amostra menor de dados, e com os três modelos preditivos estudados. Essa amostra contou com 24 pontos (24 horas ou 1 dia) extraídos do subconjunto de teste comparado com os respectivos dados reais, conforme ilustrado na Figura 16.

Figura 16 – Previsões do PLD dos modelos preditivos das horas do dia 31/12/2022.



Fonte: Elaborada pelo próprio autor.

Como observado através da figura acima, as predições dos algoritmos para o subcojunto de teste, de 24 amostras, obtiveram valores previstos próximos aos valores reais. Sendo assim, a Tabela 2 ilustra o MAPE e o MSE de cada modelo preditivo utilizado.

Tabela 2 – MAPE e MSE dos modelos preditivos.

Modelos	**MAPE (%)**	**MSE**
Random Forest	7.08	185.35
Gradient Boosting	6.28	165.48
Support Vector Regressor	9.94	238.19

Fonte: Elaborada pelo próprio autor.

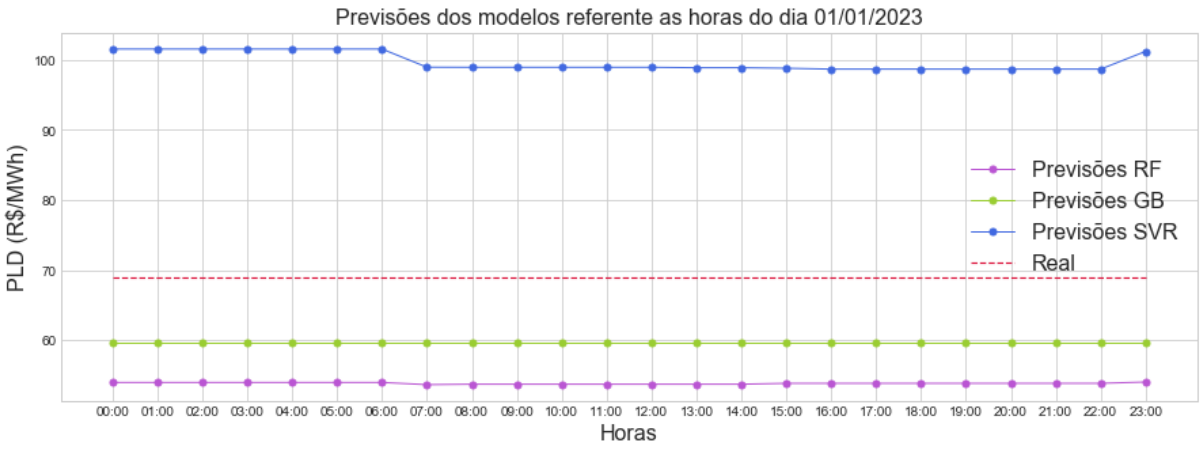
5.5. APLICAÇÃO DE NOVOS DADOS

Concluída a proposta elaborada na Seção 2.1. Objetivos Específicos com a metodologia empregada e com os resultados obtidos, também foi realizado a aplicação de dados distintos da base descrita na Seção 4.1. Repositório de Dados, a fim de retratar e exemplificar dois conceitos muito importantes dentro do contexto de AM. São eles (SOUZA, 2022):

- *Underfitting:* ocorre quando há uma quantidade insuficiente de amostras e a arquitetura do algoritmo durante o treinamento não consegue criar regras específicas para identificar o padrão ideal. Como resultado, o modelo não é capaz de estabelecer regras adequadas para o comportamento dos dados.
- *Overfitting:* ocorre quando os dados se ajustam excessivamente durante o treinamento, resultando em uma baixa capacidade de generalização para a entrada de novos dados.

Sabendo disso, foi utilizado uma pequena base de dados com apenas 24 registros (24 horas) do dia 01/01/2023 para serem aplicados nos modelos preditivos treinados e testados. O resultado foi ilustrado conforme a Figura 17.

Figura 17 – Previsões do PLD dos modelos preditivos treinados e testados com novos dados.



Fonte: Elaborada pelo próprio autor.

Portanto, através da figura acima, constata-se que o modelo preditivo Aumento do Gradiente teve uma acurácia mais elevada que os demais. Para certificar-se, também foi calculado o erro de cada modelo, que pode ser exibido conforme a Tabela 3.

Tabela 3 – MAPE e MSE dos modelos preditivos treinados e testados com novos dados.

Modelos	**MAPE (%)**	**MSE**
Random Forest	21.98	230.38
Gradient Boosting	13.69	89.36
Support Vector Regressor	44.28	936.20

Fonte: Elaborada pelo próprio autor.

6. CONCLUSÃO

A comercialização de energia elétrica é um setor dinâmico e complexo, que envolve a negociação e o comércio de energia entre diferentes participantes do mercado. Nesse contexto, o AM tem se tornado uma ferramenta cada vez mais relevante. Com a quantidade crescente de dados disponíveis, algoritmos de AM podem ser aplicados para analisar e prever tendências, demanda e comportamento dos preços de energia elétrica. Isso possibilita aos participantes do mercado tomar decisões mais informadas e estratégicas, melhorando a eficiência e a lucratividade das operações.

Como apresentado neste trabalho, a finalidade do trabalho foi implementar modelos computacionais para prever o PLD no mercado de comercialização de energia, mais especificamente do submercado Sudeste/Centro-Oeste, utilizando algoritmos de *Machine Learning*: Floresta Aleatória, Aumento do Gradiente e Máquinas de Vetores de Suporte. Com esse objetivo, foi realizado a concatenação de diferentes conjuntos de dados abertos, através dos *websites* da CCEE e ONS, sendo necessário utilizar técnicas para o pré-processamento e "limpeza" de dados incongruentes.

Após a realização da etapa de pré-processamento dos dados, foi feito AED, que possibilitou explorar os dados antes da aplicação dos modelos preditivos adotados, podendo interpretar e compreender o comportamento dos dados e associações entre as variáveis estudadas. Um destaque maior para AED foi a construção do gráfico de mapa de calor, o qual possibilitou a analisar a correlação de todas as variáveis e que, como esperado, a variável "CMO" (custo por unidade de energia produzida para atender a um acréscimo de carga no sistema) e a variável "Geração UT" (geração de energia elétrica proveniente de usinas térmicas), obtiveram uma correlação forte com a *target* "PLD". Vale destacar também a implementação da técnica de engenharia de recursos, que consistiu na criação de novas variáveis, essa técnica foi fundamental para os *scores* satisfatórios obtidos pelos modelos preditivos, assim como a utilização do *Random Search* para a escolha dos melhores hiperparâmetros, que possibilitou o aumento da acurácia dos modelos, como também a diminuição do tempo de testes.

Realizado todas as etapas anteriores e essenciais para a execução dos modelos, os mesmos atingiram resultados oportunos em comparação com os valores reais. Através das métricas de validação adotadas (MAPE e MSE), o modelo preditivo Aumento do Gradiente se mostrou mais apropriado, atingindo o MAPE de 6.28% e MSE de 165.48, em seguida o modelo Floresta Aleatória com MAPE de 7.08% e MSE de 185.35 e por fim, o modelo Máquinas de Vetores de

Suporte com MAPE de 9.94% e MSE de 238.19. Somado a isso, o modelo preditivo Aumento do Gradiente ajustado, se mostrou eficaz com a aplicação de novos dados.

Finalmente, como resultado obtidos por este Trabalho de Conclusão de Curso, foi possível desenvolver modelos preditivos com uma ótima acurácia, que pode servir de suporte para os agentes e gestores do mercado de energia elétrica.

REFERÊNCIAS

ABUELLA, M.; CHOWDHURY, B. Random Forest ensemble of support vector regression models for solar power forecasting. In: **IEEE Power & Energy Society Innovative Smart Grid Technologies Conference**, p. 1-5, 2017.

ADHIKARI, R.; AGRAWAL, R. K. **An introductory study on time series modeling and forecasting**. arXiv preprint arXiv:1302.6613, 2013.

ALMEIDA, Danilo Nichele Ottoni et al. **Análise de viabilidade econômica de adesão ao mercado livre de energia**. 2021.

ARAUJO, E. R. R. M.; LIMA, F. M.; DA SILVA, R. M. Predição da Radiação Solar Global Usando Rede Neural no Município de Seropédica-RJ. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, 6(2), 2018.

BERGSTRA, J.; BENGIO, Y. Random search for hyper-parameter optimization. **Journal of Machine Learning Research**, p. 281–305, 2012.

BORGES, Eduardo Antonio. **Uso de técnicas de aprendizado de máquina na avaliação da qualidade do leite**. 2023.

BROWNLEE, J. **A gentle introduction to the gradient boosting algorithm for machine learning**. 2016. Disponível em: https://machinelearningmastery.com/gentleintroduction-gradient-boosting-algorithmmachine-learning. Acesso em: 04 fevereiro 2023.

CAMPBELL, Marcella Barbosa Brandão da Silva. **Programação Diária Energética com o uso do DESSEM – Visão do Agente de Geração do Setor Elétrico – RJ**. 2022. Dissertação (Mestrado em Engenharia Elétrica) – Universidade Federal de Itajubá, Rio de Janeiro, 2022.

CAPITAINE, Louis; GENUER, Robin; THIÉBAUT, Rodolphe. Random forests for high-dimensional longitudinal data. **Statistical methods in medical research**, v. 30, n. 1, p. 166-184, 2021.

CARON, Alberto; BAIO, Gianluca; MANOLOPOULOU, Ioanna. Shrinkage Bayesian Causal Forests for heterogeneous treatment effects estimation. **Journal of Computational and Graphical Statistics**, v. 31, n. 4, p. 1202-1214, 2022.

CARRIJO, J. V. N. **Inteligência artificial associada à dados de satélite na predição do potencial energético em área de cerradão**. 2019. 63 f., Dissertação (Mestrado em Ciências Florestais), UNB, 2019.

CCEE, Câmara de Comercialização de Energia Elétrica. **Capacitação**. Disponível em: https://www.ccee.org.br/mercado/capacitacao. Acesso em: 04 fevereiro 2023.

CCEE, Câmara de Comercialização de Energia Elétrica. **O que fazemos**. Disponível em: https://www.ccee.org.br. Acesso em: 04 fevereiro 2023.

CCEE, Câmara de Comercialização de Energia Elétrica. **Preços**. Disponível em: https://www.ccee.org.br/. Acesso em: 04 fevereiro 2023.

CEPEL, Centro de Pesquisas de Energia Elétrica. **Manual de Treinamento – NEWAVE e DECOMP**. Disponível em: https://www.cepel.br/. Acesso em: 04 fevereiro 2023.

CHATZIS, Sotirios P. et al. Forecasting stock market crisis events using deep and statistical machine learning techniques. **Expert systems with applications**, v. 112, p. 353-371, 2018.

COELHO, R. A. **O que é o PLD – Preço de Liquidação das Diferenças?** 2017. Disponível em: http://grugeen.eng.br/o-que-e-o-pld-preco-de-liquidacao-das-diferencas/. Acesso em: 04 fevereiro 2023.

DOS SANTOS, Cosme Rodolfo Roque et al. Aplicação de aprendizado de máquina para projeção do preço horário de liquidação das diferenças como suporte às estratégias de comercialização de energia elétrica. **Revista Brasileira de Energia**, v. 28, n. 1, 2022.

DUTRA, Breno. **Importância da normalização e busca dos dados em Machine Learning**. 2021. Parceiro de crescimento da IPNET. Disponível em: https://medium.com/ipnet-growth-partner/padronizacao-normalizacao-dados-machine-learning-f8f29246c12. Acesso em: 04 fevereiro 2023.

DUTT, Vishal; SHARMA, Shweta. Artificial intelligence and technology in weather forecasting and renewable energy systems: emerging techniques and worldwide studies. **Artificial Intelligence for Renewable Energy Systems**, p. 189-207, 2022.

ELÉTRICA, Mercado Livre de Energia. **Conceito.** Disponível em: https://www.mercadolivredeenergia.com.br/consumidores-livres-e-especiais/conceito/. Acesso em: 04 fevereiro 2023.

EPE, Empresa de Pesquisa Energética. **Plano Decenal de Expansão de Energia 2031 / Ministério de Minas e Energia**. Brasília: MME/EPE, 2022.

EXCHANGE, Stack. **Why does training an SVM take so long? How can I speed it up?** Disponível em: https://ai.stackexchange.com/questions/7202/why-does-training-an-svm-take-so-long-how-can-i-speed-it-up. Acesso em: 04 fevereiro 2023.

EXCHANGE, Stack. **Why does the training time of SVMs dramatically decrease after applying dimensionality reduction to the features?** Disponível em: https://ai.stackexchange.com/questions/24100/why-does-the-training-time-of-svms-dramatically-decrease-after-applying-dimensio?rq=1. Acesso em: 04 fevereiro 2023.

GOMES, Jan Luccas de Oliveira. **Estudo de previsão de irradiância solar usando regressão por vetores de suporte.** 2018.

HUANG, N.; LU, G.; XU, D. A permutation importance-based feature selection method for short-term electricity load forecasting using random forest. **Energies**, 9(10), p. 767, 2016.

JLIDI, Mokhtar et al. An Artificial Neural Network for Solar Energy Prediction and Control Using Jaya-SMC. **Electronics**, v. 12, n. 3, p. 592, 2023.

KOPILER, A. A. et al. Redes Neurais Artificiais e suas aplicações no setor elétrico. **Revista de Engenharias da Faculdade Salesiana**, n. 9, p. 27-33, 2019.

LAMPIS, Andrea et al. **Dossier de Energia 2022. Brasil: um foco no setor elétrico**, 2022.

LEME, João Vitor; CASACA, Wallace; COLNAGO, Marilaine; DIAS, Maurício Araújo. Towards Assessing the Electricity Demand in Brazil: data-driven analysis and ensemble learning models. **Energies,** [S.L.], v. 13, n. 6, p. 1407, 18 mar. 2020. MDPI AG. http://dx.doi.org/10.3390/en13061407.

LIER, C. **Applying Machine Learning Techniques to Short Term Load Forecasting**. PhD thesis, University of Groningen, Groningen, Netherlands, 2015.

LIMA, A. dos S. et al. Classificador ensemble: uma abordagem não paramétrica aplicado à detecção de diabetes. **Revista do Seminário Internacional de Estatística com R**, 4(2), 2019.

LIMA, J. K. F. de. **Um estudo sobre a instabilidade causada no ambiente de livre contratação de energia elétrica devido a erros no processo de formação do preço de liquidação das diferenças**. 2019. 69 f. Monografia (Graduação em Engenharia Elétrica), UFC, Fortaleza, 2019.

MEI, Jie et al. A random forest method for real-time price forecasting in New York electricity market. In: **2014 IEEE PES General Meeting Conference & Exposition. IEEE**, 2014. p. 1-5.

MORAES, M. M. M. **Uma análise dos aspectos jurídicos da comercialização de energia elétrica no ambiente de contratação livre**. 2019. 37 f. Monografia (Graduação em Direito), UFC, Fortaleza, 2019.

NUNES, Lucas Renan Maués et al. Uso do ARIMA e SVM para previsão de séries temporais do sistema elétrico brasileiro. **Research, Society and Development**, v. 12, n. 3, p. e8112340438-e8112340438, 2023.

ONS, Operador Nacional do Sistema Elétrico. **Portal de Dados Abertos ONS.** Disponível em: https://dados.ons.org.br/. Acesso em: 04 fevereiro 2023.

PAULA, Matheus; MARILAINE, Colnago; NUNO, Fidalgo Jose; WALLACE, Casaca. Predicting Long-Term Wind Speed in Wind Farms of Northeast Brazil: a comparative analysis through machine learning models. **IEEE Latin America Transactions**, [S.L.], v. 18, n. 11, p. 2011-2018, nov. 2020. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/tla.2020.9398643.

PEREIRA, F. D. **Uso de um método preditivo para inferir a zona de aprendizagem de alunos de programação em um ambiente de correção automática de código**. 2018. 118 f. Dissertação, UFAM, 2018.

REIS, Fernando Simões dos. **Mudanças climáticas e transição energética justa: reflexões sobre a atuação do TCU**. 2023. Monografia (Especialização em Controle da Desestatização e da Regulação) – Instituto Serzedello Corrêa, Escola Superior do Tribunal de Contas da União, Brasília DF. 101f

RIBEIRO, M. H. D. M.; COELHO, L. Abordagem de conjunto baseada em ensacamento, aumento e empilhamento para previsão de curto prazo em séries temporais do agronegócio. **Applied Soft Computing**, v. 86, p. 105837, 2020.

RODRIGUES, A. L. **Redes Neurais Artificiais aplicadas na previsão de preços do mercado spot de energia elétrica.** 2009. Dissertação (Mestrado em Energia), USP, São Paulo, 2009.

SANTOS, André Quites Ordovás et al. Electricity Market in Brazil: A Critical Review on the Ongoing Reform. **Energies**, v. 14, n. 10, p. 2873, 2021.

SANTOS, Matheus Vizzotto dos. **Projeção de demanda elétrica com algoritmos de aprendizado de máquina.** 2022.

SHAWE-TAYLOR, John; SUN, Shiliang. Kernel methods and support vector machines. In: **Academic Press Library in Signal Processing**. Elsevier, 2014. p. 857-881.

SOUSA, Maickson Eduardo Fernandes de. **Fontes energéticas disponíveis para autogeração de energia elétrica na mineração: vantagens e desvantagens.** 2023.

VELASCO, L. C. P. et al. Day-ahead load forecasting using support vector regression machines. **International Journal of Advanced Computer Science and Applications**, 9(3), 2018.

WANG, P.; LI, Y.; REDDY, C. K. **Machine learning for survival analysis: A survey**. arXiv preprint arXiv:1708.04649, 2017.