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

Trabalho de Conclusão de Curso

Análise Preditiva da Geração Fotovoltaica via Algoritmos de Inteligência Computacional: Modelagem e Estudo de Caso da Usina Solar Bom Jesus da Lapa - BA

HENRIQUE POSTINGEL BERGAMO

LUCAS GOMES DOS RAMOS

Rosana - SP

Janeiro de 2022

HENRIQUE POSTINGEL BERGAMO

LUCAS GOMES DOS RAMOS

Análise Preditiva da Geração Fotovoltaica via Algoritmos de Inteligência Computacional: Modelagem e estudo de caso da Usina Solar Bom Jesus da Lapa - BA

Trabalho de Conclusão de Curso apresentado à UNESP, Câmpus de Rosana, Alunos: Henrique Postingel Bergamo e Lucas Gomes dos Ramos. Orientadores: Wallace Correa de Oliveira Casaca e Marilaine Colnago.

Rosana - SP

Janeiro de 2022

B493

Bergamo, Henrique Postingel

Análise Preditiva da Geração Fotovoltaica via Algoritmos de Inteligência Computacional: Modelagem e Estudo de Caso da Usina Solar Bom Jesus da Lapa - BA / Henrique Postingel Bergamo. -- Rosana, 2022

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

Orientador: Wallace Correa de Oliveira Casaca

Coorientadora: Marilaine Colnago

1. Gomes dos Ramos. 2. Lucas. 3. Ciência de Dados. 4. Energia.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Engenharia e Ciências, Rosana. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

DEDICATÓRIA

Dedicamos este trabalho a nossas famílias, orientadores e ao campus experimental da UNESP Rosana, que sempre estiveram ao nosso lado apoiando e incentivando a realizá-lo.

AGRADECIMENTOS

Primeiramente gostaríamos de agradecer nossos respectivos pais, que nos apoiaram a cada momento de nossas vidas e nos deram base para concluir esta graduação.

Aos nossos professores orientadores, que durante 3 anos nos acompanharam durante nossa graduação e pesquisa, apresentando-nos ao mundo acadêmico, além de fornecerem todo o auxílio para a elaboração deste trabalho.

Aos professores do curso de Engenharia de Energia do campus experimental da Unesp Rosana, que através dos seus ensinamentos permitiram que pudéssemos evoluir de diversas formas.

E as nossas respectivas namoradas (Henrique e Karla | Lucas e Laysla), por nos dar forças estando sempre ao nosso lado nos apoiando e incentivando durante todos os momentos da graduação.

RESUMO

O Brasil passa, atualmente, por um momento de diversificação de sua matriz energética. De um lado, ainda temos grande incerta com relação ao reestabelecimento anual dos reservatórios, e de outro vemos um crescimento e diversificação em ritmo acelerado das mais variadas fontes de energia alternativas. Neste cenário, estudos que possam diminuir estas incertezas, e ainda alavancar a confiabilidade dessas novas fontes de energia são de grande importância para o setor elétrico brasileiro. Neste cenário, em que se busca compreender e predizer dados futuros, modelos matemáticos computacionais têm sido aplicados com sucesso em diferentes áreas de conhecimento, com destaque especial para o setor elétrico. Desta forma, este trabalho objetiva estudar o comportamento da geração fotovoltaica de uma usina solar, definindo parâmetros que possam subsidiar estudos em outras usinas similares. Tal tarefa foi realizada a partir do uso de ferramentas de Análise Exploratória de Dados (AED) e de modelos de Aprendizado de Máquina (AM). A fim de viabilizar a metodologia proposta e conseguir validá-la em um cenário real, este trabalho combinou dados da usina solar de Bom Jesus da Lapa com dados meteorológicos da referida região, promovendo assim um estudo de previsibilidade da geração da energia elétrica. Para o desenvolvimento da pesquisa, foram exploradas quatro diferentes metodologias de AM: Florestas Aleatórias (*Random Forest*), Máquinas de Vetores de Suporte (*Support Vector Machine*), Aumento Extremo do Gradiente (XGBoost – *Extreme Gradient Boosting*), e Rede Perceptron Multicamadas (Multilayer Perceptron). Os modelos foram codificados e validados a partir de bases de dados reais de geração de energia fotovoltaica em conjunto com dados meteorológicos, cujo arcabouço computacional desenvolvido poderá servir para fomentar planos de eficiência energética tanto em âmbito governamental como no setor elétrico brasileiro em geral.

Palavras-chave: Matriz energética; Fontes de energia alternativas; Modelos de Aprendizado de Máquina.

ABSTRACT

Brazil is currently going through a sensitive moment in the maturing of its energy matrix. On one hand, we still have great uncertainty regarding the annual reestablishment of the reservoirs, and on the other we see a growth and diversification at an accelerated pace of the most varied alternative energy sources. In this scenario, studies that can reduce these uncertainties, and also leverage the reliability of these new energy sources are of great importance to the Brazilian electricity sector. In this scenario, in which one seeks to understand and predict future data, computational mathematical models have been successfully applied in different areas of knowledge, with special emphasis on the electricity sector. Thus, this work aims to study the behavior of photovoltaic generation in a solar plant, defining parameters that can subsidize studies in other similar plants. This task was accomplished through the use of Exploratory Data Analysis (EDA) tools and Machine Learning (ML) models. In order to make the proposed methodology viable and to validate it in a real scenario, this work combined data from the Bom Jesus da Lapa solar plant with meteorological data from that region, thus promoting a study of the predictability of electricity generation. To develop the research, four different ML methodologies were explored: Random Forest, Support Vector Machine, Extreme Gradient Boosting, and Multilayer Perceptron Network. The models were coded and validated from real databases of photovoltaic power generation in conjunction with meteorological data, whose developed computational framework may serve to foster energy efficiency plans both at the governmental level and in the Brazilian electricity sector in general.

Keywords: Energy Matrix; Alternative Energy Sources; Machine Learning Models.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

Ganhando cada vez mais espaço na matriz energética mundial, as fontes de energia renováveis tornam-se cada vez mais importantes tanto no cenário econômico como no ambiental, levando à uma substituição gradual dos combustíveis fósseis (VIEIRA et al, 2020).

Com a crescente utilização de combustíveis fósseis após a primeira e segunda revolução industrial, atingiu-se um avanço extremamente rápido, onde os processos e tecnologias mantiveram-se em constante evolução até os dias atuais (ATKESON et al, 2001). Em contrapartida ao avanço proporcionado, os combustíveis fósseis são altamente nocivos ao meio ambiente, mesmo possuindo intervalo de utilização em larga escala relativamente pequeno (menos de 300 anos). Combustíveis fósseis acumulam diversos impactos negativos ao planeta Terra, como por exemplo: as altas taxas de dióxido de carbono emitidas na atmosfera, que acabam por gerar problemas como o efeito estufa e doenças como câncer (JIA et al, 2019).

Logo, a substituição dos combustíveis fósseis por outra forma de energia é um dos maiores desafios enfrentados atualmente, porém, tentar frear o consumo desse tipo de combustível é uma saída inviável na prática, uma vez que representam a base de diversos setores ao redor do mundo (JOHNSSON et al, 2019). Assim, o que nos resta fazer é buscar uma substituição gradativa da referida fonte de energia por meio de fontes renováveis de energia, o que a longo prazo representaria uma transição segura e extremamente importante para o desenvolvimento contínuo da sociedade.

Outro fator importante que corrobora e acelera esta inserção das fontes renováveis na matriz energética mundial é a busca pela sua diversificação, i.e., a busca por evitar a dependência de uma única fonte (PAIM et al, 2019). Problemas como este são frequentes no Brasil, já que o país depende fortemente de fontes hídricas de energia (atualmente representa aproximadamente 60% da sua geração elétrica), o que implica em dificuldades no controle dos preços da energia elétrica, além de racionamentos frequentes (EPE, 2021). Além disso, este problema não é exclusivo do Brasil, a crise do petróleo que atingiu o mundo no começo da década de 70 introduziu uma preocupação imediata em diversos países em relação a dependência majoritária de uma fonte energética.

Tomando como base os problemas citados, as principais potências mundiais buscam uma diversificação de suas fontes, com foco em energias renováveis (TUFAIL et al, 2018). Na medida que novas fontes de energia foram se desenvolvendo, por meio de pesquisas

abrangentes, sua aplicação tornou-se cada vez mais acessível, intensificando seu uso e acarretando impactos positivos em diversos setores (SAKER et al, 2018).

Com o crescimento da aplicação destas novas fontes, algumas preocupações começam a ser pontuadas. Por possuírem um comportamento diferente das fontes convencionais, o aumento de representatividade na matriz energética por fontes como Solar e Eólica necessita de um cuidado maior em relação à variáveis antes ignoradas neste setor, como por exemplo variáveis relacionadas a irradiação (ZIANE et al, 2021).

Em cenários como este, a aplicação de aparatos computacionais, em especial técnicas de Aprendizado de Máquina baseadas em modelos matemáticos e estatísticos, ganham cada vez mais espaço, conseguindo atingir resultados robustos de alto desempenho (CHANDRAN et al, 2021) (PALLONETTO et al, 2019). No estudo apresentado em (LEME et al, 2020), são utilizados três distintos modelos de AM para predição da Carga Elétrica do SIN (Sistema Interligado Nacional), onde todos os modelos apresentaram resultados excelentes, já em (RAMOS et al, 2022, são utilizados três diferentes modelos de AM para prever a carga de energia gerada/exportada de energia fotovoltaica distribuída em Queensland, Australia. Tal crescimento pode ser justificado em razão do uso de modelos preditivos em dados de energia elétrica, tendo proporcionado uma melhor previsão do comportamento de geração e, consequentemente, ampliando a confiabilidade e desenvolvimento destas fontes.

Perante o que foi exposto, este estudo tem como principal objetivo estudar o impacto da utilização combinada de diferentes fontes de dados, visando assim prever a geração de energia elétrica do parque solar de Bom Jesus da Lapa - BA, utilizando dados de geração histórica em conjunto com dados meteorológicos da região. Para tornar factível a tarefa almejada, serão implementados diferentes modelos de Aprendizado de Máquina (AM), além de ferramentas para Análise Exploratória de Dados (AED) de modo a criar metodologias e modelos computacionais que permitam predizer curvas de geração elétrica, dando aos agentes do setor elétrico base para fundamentar melhor suas decisões de operação e manutenção.

1.1. OBJETIVOS

A fim de se cumprir com o objetivo geral proposto, procurou-se desenvolver as seguintes etapas:

1. Combinar dois bancos de dados, sendo eles: o banco de dados meteorológicos, referentes à cidade de Bom Jesus da Lapa - BA, disponibilizado no site do Instituto Nacional de Meteorologia (INMET), e o banco coletado no site do Operador Nacional do Sistema (ONS), que é composto por dados de geração da energia fotovoltaica, do parque solar de mesmo nome.

2. Efetuar a limpeza e tratamento dos dados, tornando possível a utilização da base por parte dos modelos de AM estudados.

3. Realizar a Análise Exploratória de Dados (AED), que tem como principal objetivo estudar/entender comportamentos presentes nos dados e correlações entre as variáveis, a fim de melhorar a acurácia dos modelos preditivos.

4. Validar e analisar os resultados dos respectivos modelos de Aprendizado de Máquina, realizando análises gráficas e quantitativas a partir dos dados segregados para a validação da resposta, com base em métricas de qualidade da área estatística.

2. REFERENCIAL TEÓRICO

Nesta seção, serão abordados os principais conceitos empregados no estudo, desde a importância da geração fotovoltaica para a matriz energética mundial e dos modelos de aprendizado de máquinas para o setor elétrico, bem como artigos e pesquisas que obtiveram resultados satisfatórios em áreas semelhantes de pesquisa, que serviram de motivação e base para o desenvolvimento deste trabalho.

2.1. GERAÇÃO SOLAR FOTOVOLTAICA

Nos últimos anos, o uso de energia proveniente de fontes renováveis tem crescido de maneira acelerada. Isto se deve a incessante busca pela diversificação da matriz energética e pela procura pela substituição gradual dos combustíveis fósseis, visto que os impactos ambientais enfrentados hoje, estão fortemente correlacionados com o crescimento desenfreado do setor industrial e a utilização em larga escala de fontes de energias não renováveis, as quais vêm sendo utilizadas de forma sistemática nos processos de industrialização e expansão da economia mundial a décadas (DESTEK et al, 2020) (ZAFAR et al, 2019).

Dentre todas as fontes de energia renováveis disponíveis no momento, podemos destacar a fonte fotovoltaica, que vem ganhando cada vez mais espaço na matriz energética global (LAUGS et al, 2020). No Brasil, podemos afirmar que o cenário se repete, apresentando crescimento quase que de maneira exponencial, visto que entre os anos de 2013 e 2019, a Geração Distribuída (GD) cresceu aproximadamente 230% ao ano; tal crescimento possibilitou que o país atingisse 1 GW de potência instalada no ano de 2019, passando em janeiro de 2020 a 2 GW, e em junho do mesmo ano alcançou 3 GW[1]. Segundo a Agência Nacional de Energia Elétrica (ANEEL), nos últimos sete anos a produção de energia fotovoltaica, tanto em geração distribuída quanto em geração centralizada, cresceu aproximadamente 150% ao ano. A partir do ano de 2017, a geração vem dobrando sua capacidade instalada anualmente, passando de 1 GW no ano de 2017, para 2 GW em 2018, 4 GW em 2019, e em 2020 atingiu 6 GW (ABSOLAR, 2021).

Se tratando do cenário global, a Agência Internacional de Energia (AIE), prevê um crescimento global de 50% na potência instalada de fontes de energia renováveis, entre os anos

[1] Disponível em:https://www.absolar.org.br/noticia/geracao-distribuida-fotovoltaica-cresce-230-ao-ano-no-brasil/. Acesso em: 2 de dez. 2021.

de 2019 e 2024, o que resultará em um crescimento global de 1.200GW, onde 60% desse crescimento será proveniente da energia fotovoltaica (ACADEMIA DO SOL, 2021).

Dentre todos os países do mundo, podemos ao menos destacar cinco grandes potências. No quesito capacidade instalada, são elas: China, Estados Unidos da América, Japão, Alemanha e Índia. A Figura 1 ilustra o ranking global de potência instalada (fotovoltaica) no ano de 2020.

Figura 1: Ranking global de potência instalada.

Ranking	País	Capacidade Instalada [MW] (acumulada em 2020)
1	China	253.884
2	EUA	73.814
3	Japão	68.665
4	Alemanha	53.781
5	Índia	38.983
6	Itália	21.594
7	Austrália	17.342
8	Vietña	16.504
9	Coréia do Sul	14.575
10	Reino Unido	13.462
11	Espanha	11.785
12	França	11.724
13	Países Baixos	10.213
14	Brasil	10.000*

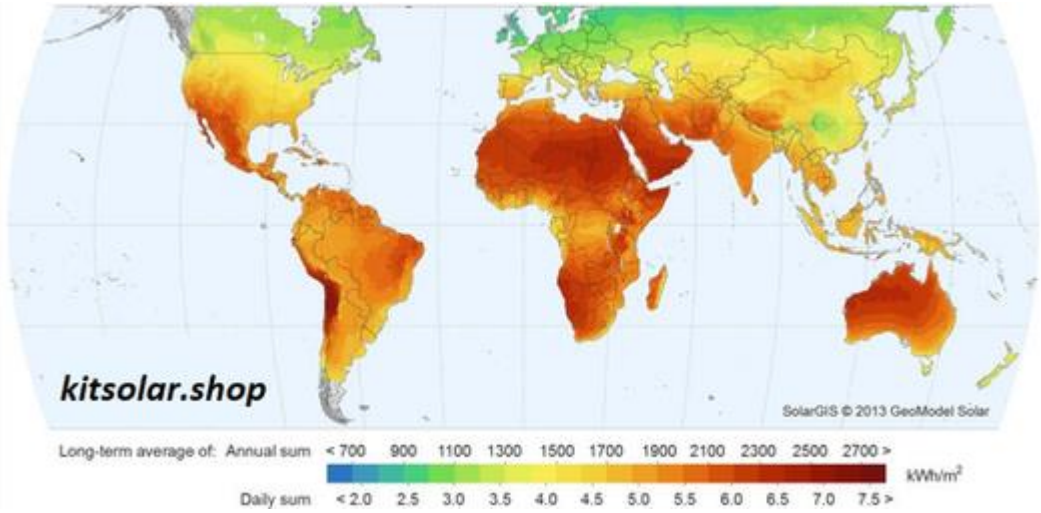
Fonte: (SUNTOP, 2021)

Um ponto de destaque para esse tipo de fonte de energia é a sua facilidade de implementação, visto que os painéis fotovoltaicos podem ser instalados nas mais diferentes localidades, onde fontes de energia convencionais não poderiam ser instaladas, ou não teriam um desempenho viável.

Com base no discorrido, podemos afirmar que a energia solar fotovoltaica é uma das mais promissoras opções energéticas, tanto para o Brasil quanto para o resto do mundo, e um ponto de grande importância é a alta capacidade de produção brasileira, visto que a maior parte do território está localizado dentro da região intertropical, o que proporciona uma alta incidência de radiação solar durante todo o ano (CARVALHO et al, 2017).

A Figura 2 ilustra, através de um mapa de calor, a incidência solar média por hora em KWh/m², em uma escala global, onde podemos ver de maneira clara o potencial brasileiro para a geração. Quando comparamos com outros países como Japão, Alemanha e Estados Unidos da América, fica claro que o Brasil recebe uma maior incidência de radiação solar por m², o que viabiliza ainda mais a implantação de fonte de energia por todo o território brasileiro.

Figura 2: Mapa de calor da incidência da radiação solar.



Fonte: (DESTEK et al, 2020).

2.2. MERCADO LIVRE DE ENERGIA E AS FONTES RENOVÁVEIS

Com o marco regulatório de 1995 sancionando o mercado livre de energia, também conhecido como Ambiente de Contratação Livre (ACL), deu-se início a um ambiente mais competitivo de negociação de energia elétrica, onde os comercializadores e consumidores podem negociar todas as condições referentes ao contrato de energia, como por exemplo, a quantidade, o período e o preço do produto (SILVA et al, 2020).

Com o avanço e amadurecimento das regras deste mercado, definidas em sua maioria pela Câmara de Comercialização de Energia Elétrica, oportunidades para investimentos e construções de usinas de fontes renováveis foram ganhando cada vez mais atenção aos olhos de investidores internos e externos. Este interesse surge em função dos ganhos possíveis com a venda da energia elétrica gerada e pelo enorme potencial do mercado brasileiro.

A usina em estudo, Bom Jesus da Lapa, é administrada pela Enel Green Power (multinacional italiana do ramo de energia elétrica)[2], e assim como outras grandes empresas, enxergou o grande potencial elétrico e financeiro que se teria ao construir plantas de geração ligadas ao ACL. Logo, com a abertura da comercialização de energia para empresas do setor privado, há uma alta implantação de usinas de diferentes fontes renováveis, trazendo ao país uma maior diversificação e segurança energética, ajudando a suprir a demanda crescente de energia no país.

2.3. APRENDIZADO DE MÁQUINA E O SETOR ELÉTRICO

A aplicação de técnicas computacionais tornou-se uma solução extremamente poderosa para diversos problemas em diferentes áreas, o que se deve ao grande aumento da quantidade de dados disponíveis, o que tem viabilizado a modelagem preditiva.

Um exemplo representativo dessa questão é o elaborado por Silveira (SILVEIRA et al, 2021), que apresenta um estudo da aplicação de técnicas de aprendizado de máquina a fim de estimar a capacidade disponível de potência para inserção de Geração Distribuída em pontos da rede elétrica do Ceará. Outro exemplo, agora focado em realizar a estimativa de consumo de energia elétrica por residências familiares, é apresentado por Jui-Sheng (CHOU te al, 2018).

Já nos trabalhos apresentados em (FAN et al, 2018) (OBIORA et al, 2021), temos o uso do modelo Aumento Extremo do Gradiente (XGBoost) para prever a radiação solar incidente, enquanto em (RAMOS et al, 2021) o objetivo foi trabalhar com geração distribuída de energia fotovoltaica. Ainda com base na literatura, podemos destacar os seguintes artigos: em (ZHENG et al, 2019), o autor utiliza o XGBoost na previsão da geração de energia eólica, enquanto em (PAULA et al, 2020), os autores empregaram, além do modelo Aumento do Gradiente, a técnica *Random Forest* para solucionar o mesmo problema anterior, isto é, de previsão de energia eólica.

Podemos também destacar alguns estudos que utilizaram o modelo de RNAs para solucionar diferentes problemas, nas mais distintas áreas, partindo da predição da velocidade do vento, com é apresentado em (SAMADIANFARD et al, 2020). Já em (EHSAN et al, 2017), temos um trabalho totalmente voltado para a previsão da geração elétrica fotovoltaica via

[2] Disponível em:https://www.enelgreenpower.com/pt/midias/news/2017/10/brasil-egp-inaugura-parque-solar-lapa. Acesso em: 3 de Dez de 2021.

RNAs. Nota-se que ambos os trabalhos obtiveram resultados satisfatórios no quesito acurácia de previsão, demonstrando a versatilidade das RNAs para trabalhar com aplicações de diferentes áreas ligadas ao setor de energias renováveis.

De acordo com o que foi exposto acima, vemos que o uso de técnicas de aprendizado de máquina tem sido um recurso bastante efetivo no setor energético, visto que estudos nessa frente possibilitam a implementação de planos energéticos e polícias públicas mais assertivas, garantindo segurança energética e parâmetros para tomadas de decisões.

3. APRENDIZADO DE MÁQUINA: ASPECTOS TEÓRICOS E TÉCNICOS

3.1. EXTREME GRADIENT BOOSTING (XGBOOST)

O *Extreme Gradient Boost* (XGBoost), é um modelo de aprendizado de máquina, amplamente utilizado para solucionar problemas de classificação e regressão, derivado dos modelos *ensemble,* ou seja, produz seus resultados baseado em combinações de árvores de decisão (NGUYEN et al, 2020). Segundo a artigo apresentado em (CHEN et al, 2016), o termo *Boosting* é um processo que busca melhorar os resultados do modelo, de modo que o algoritmo se fundamenta na ideia de conciliar os classificadores genéricos para construir classificadores mais robustos, além da existência de uma função objetivo, que contribui para o classificador alcançar o melhor resultado possível.

De maneira geral, podemos definir a função objetivo através da função de perda somada à uma função de regularização, onde ao decorrer das interações, o modelo buscará construir uma árvore de decisão que minimize a função objetivo (ELAVARASAN et al, 2020).

O funcionamento do XGBoost é constituído pelos seguintes passos:

- O previsor local recebe a entrada dos dados.
- A árvore baseada na função de custo é construída.
- A função de custo da árvore baseada no desempenho atingido é otimizada.
- É gerado o resultado por meio do agrupamento dos resultados atingidos por cada previsor (levando em consideração o peso atribuído a cada um).

3.2. REDES NEURAIS ARTIFICIAIS (MLP)

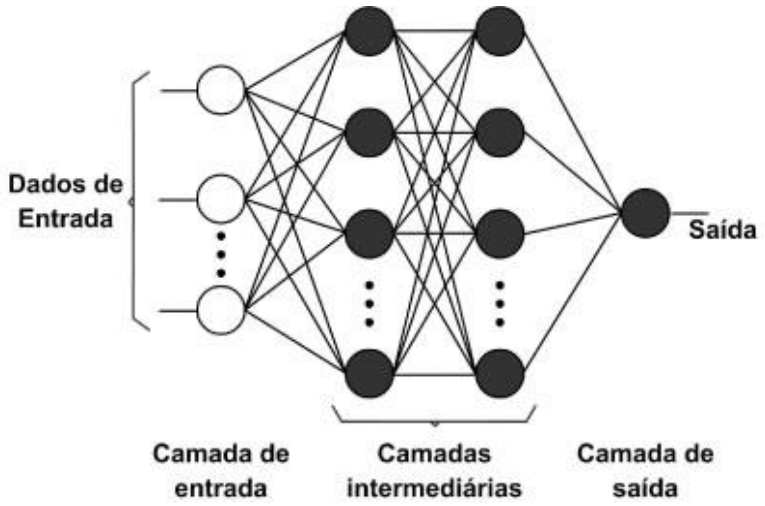
O campo das redes neurais artificiais costuma ser chamado apenas de redes neurais ou perceptron multicamadas. Um perceptor é um modelo de neurônio único, que foi o precursor de redes neurais.

É um campo que investiga como modelos simplificados, de cérebros biológicos, podem ser usados para resolver tarefas computacionais difíceis, como as de modelagem preditiva que vemos no aprendizado de máquina. O objetivo não é criar modelos realistas do cérebro, mas

sim desenvolver algoritmos robustos e estruturas de dados que possamos usar para modelar problemas complexos.

A Figura 3 apresenta a estrutura de uma Rede Neural Artificial, de modo que as RNA's são basicamente compostas por "neurônios", representados pelos círculos, e suas interconexões entre os neurônios são representadas pelos traços, os quais são responsáveis por conectar e indicar "pesos" para o neurônio subsequente. Na parte mais à esquerda da figura, temos a camada de entrada, responsável pelo processamento e transmissão dos dados de entrada para a camada subsequente. Esta, por sua vez, é conhecida como camada oculta/camada intermediária, onde em conjunto com a função de ativação reproduz relações não lineares a fim de melhorar a capacidade da rede de aprender. Já à direita da figura, temos a camada de saída que, por sua vez, é responsável por apresentar o valor final do problema solucionado (NEAGOE et al, 2018).

Figura 3: Estrutura de um modelo RNA.



Fonte: (SILVEIRA et al, 2021)

O poder das redes neurais deriva de sua capacidade de aprender a representação em seus dados de treinamento e como relacioná-los da melhor forma com a variável, a ser predita. Nesse sentido, as redes neurais aprendem um mapeamento. Matematicamente, elas são capazes de aprender qualquer função de mapeamento, e provaram ser um algoritmo de aproximação universal. A capacidade preditiva das redes neurais vem de suas estruturas hierárquicas, ou da inserção de várias camadas internas (ISMAIL et al, 2015).

3.3. FLORESTA ALEATÓRIA (RF)

Floresta Aleatória (*Random Forest* - RF) é considerado um método *ensemble*, isto é, parte de que a melhor decisão é proveniente da opinião de um grupo treinado, e não apenas de um indivíduo. Trata-se de um conjunto de estimadores que induzem a criação de seus próprios aprendizes e estratégias, onde os aprendizes base são todas as árvores de classificação/regressão.

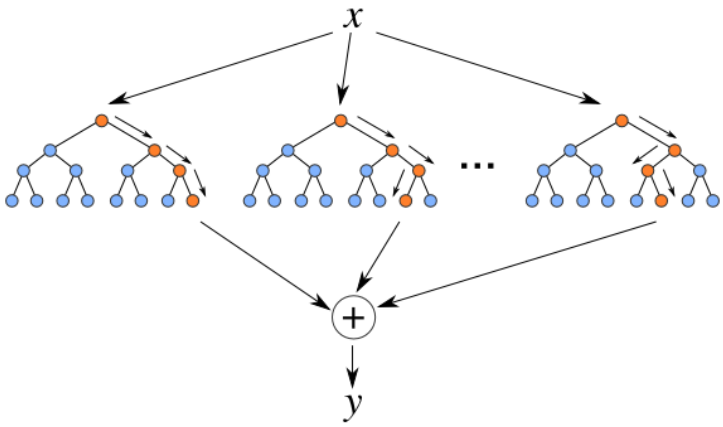
Este método de aprendizado pode tomar duas estratégias: o *Bagging* e o *Boosting*. A principal diferença entre ambas as técnicas é que, em um determinado nó, ao invés de usar todas as variáveis como o *Boosting*, o *Bagging* usa apenas um subconjunto aleatório para selecionar as variáveis no critério de divisão. Desta forma, tal randomização pode reduzir a correlação entre as diferentes árvores e, portanto, melhorar o desempenho da previsão.

O método consiste na execução de três etapas básicas:

- Gerar conjuntos de amostras *bootstrap* da base de dados de treinamento.
- Para cada amostra *bootstrap*, criar uma árvore de regressão (sem ajuste) com a seguinte modificação: em cada nó, gera-se uma amostra aleatória das variáveis de entrada de toda base de dados de treino onde escolhe-se a melhor subdivisão dessas variáveis, com, e representando o número total das variáveis da base.
- Prever a nova saída a partir do cálculo da média das saídas de árvores de regressão quando novas variáveis são inseridas ao modelo.

A Figura 4 mostra, de maneira simplificada, a estrutura de um modelo do tipo *Random Forest*, onde a mesma parte de um banco de dados (x), e em cada nó subsequente o algoritmo toma sua decisão com base em métricas predefinidas de regressão e ou classificação, atingindo a resposta média final (y), para um determinado problema (WU et al, 2017).

Figura 4: Estrutura de um modo RF.



Fonte: (EHSAN et al, 2017).

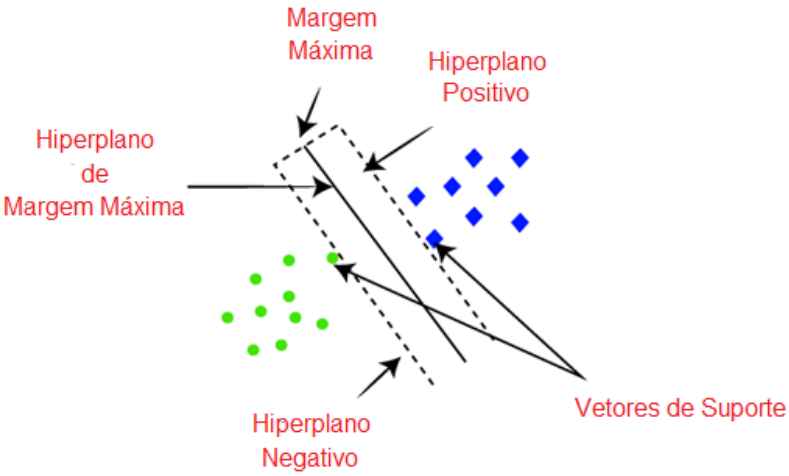
3.4. MÁQUINA DE VETORES DE SUPORTE (SVR)

Máquinas de Vetores de Suporte (*Support Vector Machines* - SVRs) é um método desenvolvido com base no aprendizado estatístico: um novo paradigma na área de AM. O método gera um hiperplano ótimo que maximiza a margem no espaço dos dados, que é a distância entre os vetores de suporte das classes distintas. Esses vetores recebem essa denominação em virtude da sua proximidade com a superfície de decisão, contribuindo de maneira decisiva para a definição de tal superfície (RADHIKA et al, 2009).

Apesar de surgir visando problemas de classificação, pode facilmente ser adaptado para resolver problemas de regressão, a partir do uso de uma função de perda, a qual é minimizada com um regularizador. Assim, os SVRs utilizados para regressão são comumente chamados de SVR (*Support Vector Regression* - Regressão Vetorial de Suporte). Desta forma, adaptando-o para nosso contexto, o problema consiste em encontrar uma função não linear que minimize o erro da previsão com relação ao conjunto de treinamento (YANG et al, 2002).

A Figura 5 nos ilustra, de maneira sucinta, a separação de duas classes. A reta central está representando o hiperplano ótimo de separação, que potencializa a margem, que é a distância entre os vetores de suporte das classes distintas.

Figura 5: Estrutura de um modelo SVR.



Fonte: (FAN et al, 2018).

4. MATERIAIS E MÉTODOS

4.1. REPOSITÓRIO DE DADOS

Os dados explorados neste trabalho foram coletados por meio das plataformas dos órgãos governamentais, sendo elas, Operador Nacional do Sistema Elétrico (ONS) e o Instituto Nacional de Meteorologia (INMET). As bases de dados disponibilizadas por estes repositórios foram concatenadas, contendo informações referentes a:

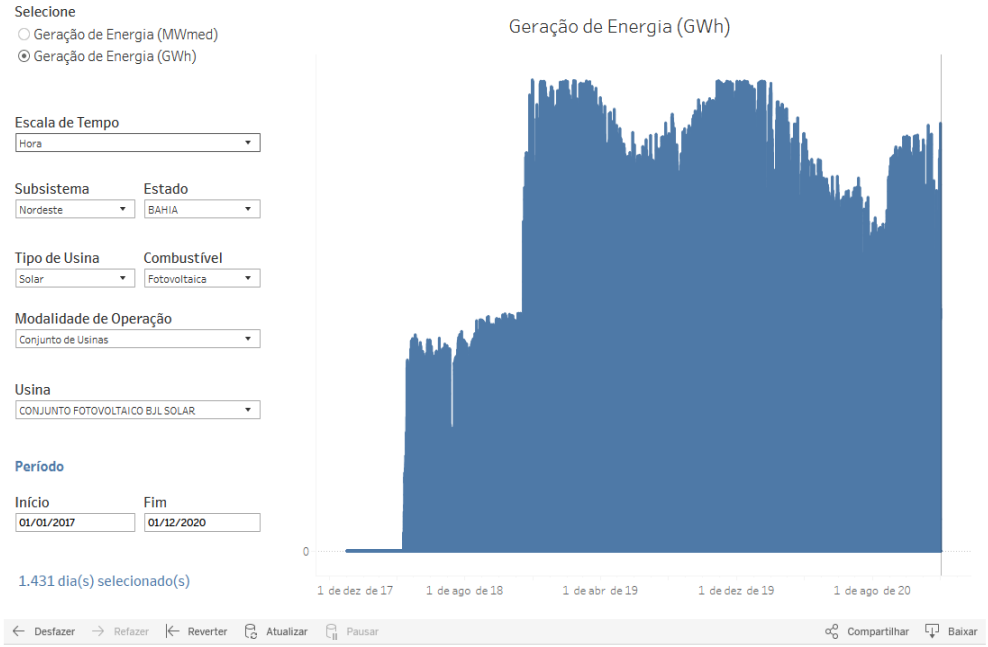
- Temperatura atual;
- Temperatura Máxima e Mínima da hora anterior (ºC);
- Vento(m/s), Radiação Global (kJ/m²);
- Precipitação (mm) e Data (Dia e Hora);
- Geração Elétrica (MWm).

O ONS é o órgão responsável pela operação (coordenação e controle) do Sistema Interligado Nacional (SIN), além de gerir os sistemas isolados do Brasil[3]. Com o intuito de monitorar e de manter um histórico acessível de dados, a plataforma online do ONS mantém um repositório de diversos tipos de dados elétricos, entre eles a de geração de energia (podendo ser filtrada por tempo, local, fonte e usina). A Figura 6 ilustra de maneira sucinta a interface do site.

Já o órgão INMET[4] foi constituído no intuito de analisar, armazenar e disponibilizar dados meteorológicos, organizando e apresentando esses dados de maneira limpa e intuitiva, servindo como uma ferramenta poderosa para se produzir análises detalhadas do comportamento meteorológico brasileiro. A Figura 7 apresenta de maneira resumida a interface do site.

[3] Disponível em: https://dados.ons.org.br/. Acesso em: 4 de dez de 2021.
[4] Disponível em: https://bdmep.inmet.gov.br/. Acesso em: 4 de dez de 2021.

Figura 6: Repositório de dados ONS.



Fonte: ONS.

Figura 7: Repositório de dados INMET.

Data Inicio: 01/01/2017
Data Fim: 31/12/2020

Variáveis:

- PRECIPITACAO TOTAL, HORARIO
- PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO, HORARIA
- PRESSAO ATMOSFERICA REDUZIDA NIVEL DO MAR, AUT
- PRESSAO ATMOSFERICA MAX.NA HORA ANT. (AUT)
- PRESSAO ATMOSFERICA MIN. NA HORA ANT. (AUT)
- RADIACAO GLOBAL
- TEMPERATURA DA CPU DA ESTACAO
- TEMPERATURA DO AR - BULBO SECO, HORARIA
- TEMPERATURA DO PONTO DE ORVALHO
- TEMPERATURA MAXIMA NA HORA ANT. (AUT)
- TEMPERATURA MINIMA NA HORA ANT. (AUT)
- TEMPERATURA ORVALHO MAX. NA HORA ANT. (AUT)
- TEMPERATURA ORVALHO MIN. NA HORA ANT. (AUT)
- TENSAO DA BATERIA DA ESTACAO
- UMIDADE REL. MAX. NA HORA ANT. (AUT)
- UMIDADE REL. MIN. NA HORA ANT. (AUT)
- UMIDADE RELATIVA DO AR, HORARIA
- VENTO, DIRECAO HORARIA (gr)
- VENTO, RAJADA MAXIMA
- VENTO, VELOCIDADE HORARIA

Estações:

- BOM JESUS DA LAPA (A418) - [BA]

Fonte: INMET.

Visto que a base de dados foi escolhida em par (unidade geradora – estação meteorológica), e visando determinar um conjunto ideal, a escolha das bases foi definida pela proximidade do par.

4.2. ANÁLISE DE DADOS: APLICAÇÃO DO PROCESSO DE DESCOBERTA DO CONHECIMENTO (KDD)

O processo de Descoberta de Conhecimento (*Knowledge Data Discovery* - KDD) se baseia na aplicação de técnicas sobre os dados no intuito de buscar informações mais profundas sobre a relação entre as variáveis em estudo, descobrindo novos padrões e amadurecendo hipóteses preliminares. Este processo é uma ferramenta poderosa quando falamos de análise de dados, possui ferramentas robustas capazes de auxiliar tanto a previsão de comportamento futuro baseado na análise dos dados históricos, quanto a apresentação de padrões identificados na análise dos dados. A aplicação deste ferramental pode ser vista como uma atividade multidisciplinar, pois engloba técnicas de visualização e análise exploratória de dados, indo além da concepção clássica de aprendizagem de máquina (UYSAL et al, 1999). Como parte do KDD, alguns tópicos frequentemente abordados são:

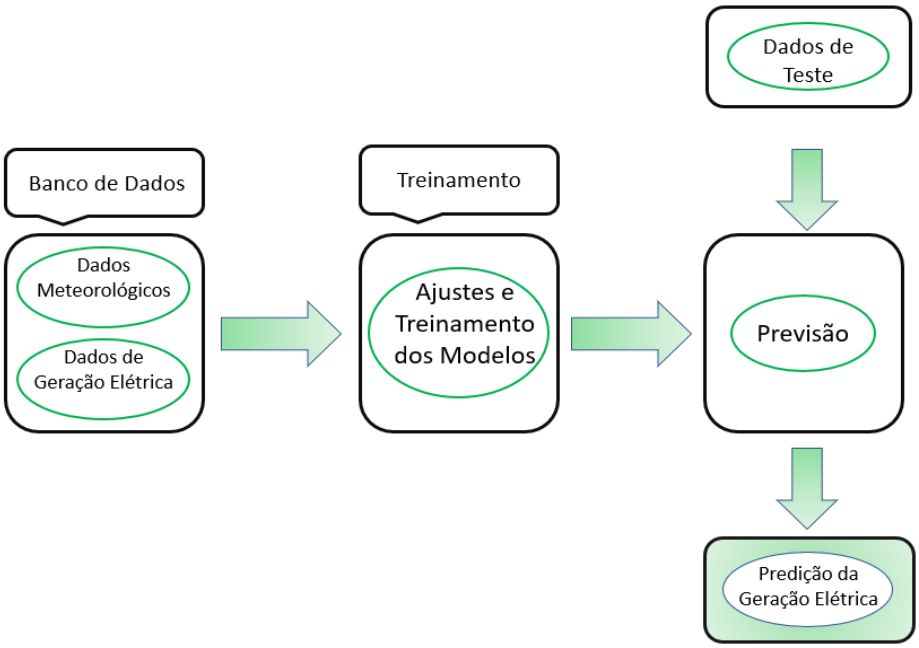
1. Importação e limpeza dos dados;
2. Teste de não linearidade;
3. Análise exploratória dos dados (AED);
4. Engenharia e seleção de recursos.

Em suma, parte do aparato acima foi aplicado nesta pesquisa a fim de embasar nossas análises para uso posterior das técnicas estudadas de AM.

4.3. PROCEDIMENTO METODOLÓGICO EMPREGADO

Considerando a problemática de pesquisa deste trabalho, bem como ainda os métodos de AM já discutidos, a Figura 8 apresenta uma ilustração da metodologia computacional aplicada para cumprir com a tarefa de predição da geração e distribuição da eletricidade, seguida dos passos mais bem detalhados.

Figura 8: Visão geral do protótipo de aprendizado a ser implementado.



Fonte: Próprios autores.

a) Coleta dos dados e criação do banco de dados conjunto (*database*), utilizado na pesquisa, a partir da compilação dos dados obtidos pelas plataformas do INMET (Instituto Nacional de Meteorologia) e o ONS (Operador Nacional do Sistema).

b) Limpeza (*Data Cleaning*) e organização dos dados a fim de mitigar redundâncias e possíveis efeitos degenerativos devido a presença de *outliers*.

c) Análise exploratória de dados, geração de sumários, e uso de ferramentas de KDD a fim de investigar o comportamento das variáveis do problema.

d) Calibragem, ajuste de performance e etapa de treinamento dos modelos preditivos.

e) Análises gráficas e testes quantitativos de validação dos modelos, bem como, comparações envolvendo cada um dos preditores obtidos na etapa anterior.

4.4. AJUSTE HIPERPARÂMETROS

Considerando os algoritmos de AM utilizados no trabalho, apresentados na Seção 3, eles apresentam diversos hiperparâmetros ajustáveis, possibilitando os modelos alcançar a melhor acurácia possível com os dados de treinamento. Entretanto, essa tarefa de otimização é, na maioria dos casos, exploratória e o número possível de ajustes de hiperparâmetros pode crescer exponencialmente.

Existe uma estratégia bastante prática na literatura de AM chamada busca em grade (ou Grid Search), a qual testa todas essas possibilidades e retorna o melhor pipeline de AM. Entretanto, há pesquisas que atestam que a busca aleatória (ou Random Search) de um número limitado de possíveis pipelines de AM podem ser também bastante efetiva e ter propriedades favoráveis quando comparada com a busca em grade (SHAHHOSSEINI et al, 2019).

Na abordagem via Random Search, deve-se estimar um número k de iterações para a construção das melhores configurações de hiperparâmetros. Como resultado, o Random Search retorna o melhor pipeline considerando o espaço original de soluções. Portanto, neste projeto, foi empregado a Random Search na tarefa de ajuste de hiperparâmetros, o que possibilitou atingir uma melhor acurácia por parte dos preditores (BERGSTRA et al, 2012).

5. RESULTADOS E DISCUSSÕES

Neste capítulo, serão apresentados os resultados alcançados referentes às metodologias propostas na pesquisa. Alguns dos procedimentos já mencionados nos capítulos anteriores também foram detalhados nas páginas seguintes.

5.1. COLETA DOS DADOS

Como exibido anteriormente (Seção 4.1), os dados utilizados neste trabalho foram coletados por meio de duas plataformas distintas, a primeira trata-se do INMET e a outra plataforma é a do ONS, o casamento entre ambos os dados formou o banco de dados em estudo.

Através da plataforma do INMET, foram coletados dados horários das variáveis meteorológicas, como Temperatura atual, Temperatura Máxima e Mínima da hora(ºC), Vento(m/s), Radiação Global (Kj/m²), Precipitação (mm) e Data (Dia e Hora). Já na plataforma do ONS, foram coletados dados horários referentes a Geração Elétrica (MWm) do parque solar de Bom Jesus da Lapa- BA, o horizonte que o conjunto de dados contempla é de janeiro de 2017 a dezembro de 2020.

A criação do banco de dados foi primeiramente elaborada por meio do *software* Excel (presente no pacote Office), gerando como resultado um arquivo .csv, assim possibilitando sua utilização na ferramenta *Spyder (*Linguagem *Python*: Presente no Pacote Anaconda).

Na Figura 9 é apresentado o banco de dados concatenado e pronto para a aplicação das técnicas de pré-processamento apresentadas na próxima seção (5.2).

Figura 9: Banco de dados inicial.

DATA	HORA	CHUV_TOTAL	RAD_GLOB	TEMPERATURA	TEMP_MAX	TEMP_MIN	VENTO_RAJ_MAX	VENTO	GERAÇÃO
17/05/2017	00:00	0	0	29	30.4	29	1.5	0	0
17/05/2017	01:00	0	0	27.6	29	27.3	0	0	0
17/05/2017	02:00	0	0	25.7	27.8	25.3	2.7	0	0
17/05/2017	03:00	0	0	26	26	25.1	3.7	0.7	0
17/05/2017	04:00	0	0	26.3	26.8	26	4.1	1.4	0
17/05/2017	05:00	0	0	24.9	26.3	24.7	3.3	0.8	0
17/05/2017	06:00	0	0	25.1	25.1	24.6	3.2	1.7	0
17/05/2017	07:00	0	0	24.7	25.2	24.3	3.7	1.9	0
17/05/2017	08:00	0	0	24.5	24.8	24.3	3.7	1.2	0.14
17/05/2017	09:00	0	0	24.3	24.5	24.3	4.5	1.8	1.27
17/05/2017	10:00	0	185.8	25	25	24.1	4.8	2	2.83
17/05/2017	11:00	0	898.2	27.9	27.9	25	5.4	2.2	3.75
17/05/2017	12:00	0	1616	29.6	30	27.8	5.2	1.6	4.71
17/05/2017	13:00	0	2227.9	31.2	31.2	29.4	4.3	0.6	4.9
17/05/2017	14:00	0	2622.5	32.2	32.5	31	4.3	1.1	3.07

Fonte: Próprios autores.

5.2. PRÉ-PROCESSAMENTO DOS DADOS

5.2.1. LIMPEZA DE DADOS

Com o objetivo de possibilitar utilização da base pelos modelos de AM, foi realizado o tratamento/limpeza do banco de dados retirando inconsistências presentes que prejudicariam de maneira significativa o aprendizado dos modelos.

Primeiramente, utilizou-se metodologia da interpolação linear para substituir incongruências, como por exemplo, o preenchimento ou substituição de dados faltantes/incoerentes únicos (quando representavam apenas uma célula na linha, sem valor). A Figura 10 exibe os passos do procedimento.

Figura 10: Representação de preenchimento em dados ausentes.

DATA	HORA	CHUV_TOTAL	RAD_GLOB	TEMPERATURA	TEMP_MAX	TEMP_MIN	VENTO_RAJ_MAX	VENTO	GERAÇÃO
17/05/2017	10:00	0	185.8	25	25	24.1	4.8	2	0.14
17/05/2017	11:00	0	898.2	27.9	27.9	25	5.4	2.2	1.27
17/05/2017	12:00	0	1616	29.6	30	27.8	5.2	1.6	2.83
17/05/2017	13:00	0	2227.9	31.2	31.2	29.4	4.3	0.6	3.75
17/05/2017	14:00	0	2622.5	32.2	32.5	31	4.3	1.1	4.71
17/05/2017	15:00	0	2638.1	-	33.9	32	3.9	1.3	4.9
17/05/2017	16:00	0	2929.7	34.4	35.5	33.2	5.4	1.7	3.07
17/05/2017	17:00	0	2782.5	34.4	36.1	33.9	5.1	1.6	4.09
17/05/2017	18:00	0	1599.4	34.6	35.7	34.1	4.2	0.8	2
17/05/2017	19:00	0	682.1	32.4	34.6	32.4	6.5	3.9	0.2
17/05/2017	20:00	0	180	32	32.4	32	6.5	1.5	0

DATA	HORA	CHUV_TOTAL	RAD_GLOB	TEMPERATURA	TEMP_MAX	TEMP_MIN	VENTO_RAJ_MAX	VENTO	GERAÇÃO
17/05/2017	10:00	0	185.8	25	25	24.1	4.8	2	0.14
17/05/2017	11:00	0	898.2	27.9	27.9	25	5.4	2.2	1.27
17/05/2017	12:00	0	1616	29.6	30	27.8	5.2	1.6	2.83
17/05/2017	13:00	0	2227.9	31.2	31.2	29.4	4.3	0.6	3.75
17/05/2017	14:00	0	2622.5	32.2	32.5	31	4.3	1.1	4.71
17/05/2017	15:00	0	2638.1	33.3	33.9	32	3.9	1.3	4.9
17/05/2017	16:00	0	2929.7	34.4	35.5	33.2	5.4	1.7	3.07
17/05/2017	17:00	0	2782.5	34.4	36.1	33.9	5.1	1.6	4.09
17/05/2017	18:00	0	1599.4	34.6	35.7	34.1	4.2	0.8	2
17/05/2017	19:00	0	682.1	32.4	34.6	32.4	6.5	3.9	0.2
17/05/2017	20:00	0	180	32	32.4	32	6.5	1.5	0

Fonte: Próprios autores.

Já em casos em que a linha inteira (ou a maior parte) se encontrava vazia, optou-se pela sua remoção definitiva do banco de dados, visto que o preenchimento ocasionaria em uma propagação dos erros de cada célula em cascata (LI et al, 2020).

5.2.2. ENGENHARIA DE RECURSOS: CRIAÇÃO DE NOVAS VARIÁVEIS

A implementação e utilização de novas variáveis em modelos preditivos é uma tarefa de alta importância, uma vez que existem relações entre os dados que ainda não foram esclarecidas ou visualizadas, necessitando assim de um tratamento mais direcionado.

Com base nisso, nesta etapa foram geradas, através de métodos estatísticos, novas variáveis por meio das originais. Como exemplo, temos a média móvel e subtração móvel da Radiação Global, que se baseia nos dados da Radiação Global originais provenientes da base de dados meteorológicos. Ao todo foram criadas 4 novas *features* baseadas nas originais (média

móvel e subtração móvel: (i) Radiação Global e (ii) Temperatura), totalizando 14 variáveis (originais + artificiais).

A Figura 11, ilustra o resultado do banco de dados após a criação das novas variáveis ao banco de dados original.

Figura 11: Banco de dados com as novas variáveis artificias.

DATA	HORA	CHUV_TOTAL	CHOVEU	RAD_GLOB	RAD_GLOB_MM	RAD_GLOB_SUBM	TEMPERATURA	TEMP_MM	TEMP_SUBM	TEMP_MAX_ANT	TEMP_MIN_ANT	VENTO	GE
17/05/2017	00:00	0	NAO	0	0	0	26.3	25.43	-1.40	26.8	26	1.4	0
17/05/2017	01:00	0	NAO	0	0	0	24.9	25.43	0.20	26.3	24.7	0.8	0
17/05/2017	02:00	0	NAO	0	0	0	25.1	24.90	-0.40	25.1	24.6	1.7	0
17/05/2017	03:00	0	NAO	0	0	0	24.7	24.77	-0.20	25.2	24.3	1.9	0
17/05/2017	04:00	0	NAO	0	0	0	24.5	24.50	-0.20	24.8	24.3	1.2	0
17/05/2017	05:00	0	NAO	0	61.93	185.8	24.3	24.60	0.70	24.5	24.3	1.8	0
17/05/2017	06:00	0	NAO	185.8	361.33	712.4	25	25.73	2.90	25	24.1	2	0
17/05/2017	07:00	0	NAO	898.2	900	717.8	27.9	27.50	1.70	27.9	25	2.2	0
17/05/2017	08:00	0	NAO	1616	1580.7	611.9	29.6	29.57	1.60	30	27.8	1.6	0.14
17/05/2017	09:00	0	NAO	2227.9	2155.47	394.6	31.2	31.00	1.00	31.2	29.4	0.6	1.27
17/05/2017	10:00	0	NAO	2622.5	2496.17	15.6	32.2	32.30	1.30	32.5	31	1.1	2.83
17/05/2017	11:00	0	NAO	2638.1	2730.1	291.6	33.5	33.37	0.90	33.9	32	1.3	3.75
17/05/2017	12:00	0	NAO	2929.7	2783.43	-147.2	34.4	34.10	0.00	35.5	33.2	1.7	4.71
17/05/2017	13:00	0	NAO	2782.5	2437.2	-1183.1	34.4	34.47	0.20	36.1	33.9	1.6	4.9
17/05/2017	14:00	0	NAO	1599.4	1688	-917.3	34.6	33.80	-2.20	35.7	34.1	0.8	3.07
17/05/2017	15:00	0	NAO	682.1	820.5	-502.1	32.4	33.00	-0.40	34.6	32.4	3.9	4.09
17/05/2017	16:00	0	NAO	180	294.73	-157.9	32	31.83	-0.90	32.4	32	1.5	2
17/05/2017	17:00	0	NAO	22.1	67.37	-22.1	31.1	30.50	-2.70	32	31.1	2.3	0.2
17/05/2017	18:00	0	NAO	0	7.37	0	28.4	28.70	-1.80	31.1	28.4	3.1	0

Fonte: Próprios autores.

A partir da inserção destas novas variáveis como sugere (AWAN et al, 2019), o comportamento das curvas e a relação entre as variáveis foram mais bem compreendidos pelos modelos, o que consequentemente levou à resultados mais assertivos como apresentado na Seção de Resultados.

5.3. ANÁLISE EXPLORATÓRIA DOS DADOS

A Análise Exploratória de Dados (AED) tem como principal objetivo detectar comportamentos das variáveis que formam o banco de dados em estudo. Busca-se relacionar e apresentar padrões que dificilmente seriam visualizados sem o auxílio computacional (LUEMBA et al, 2021).

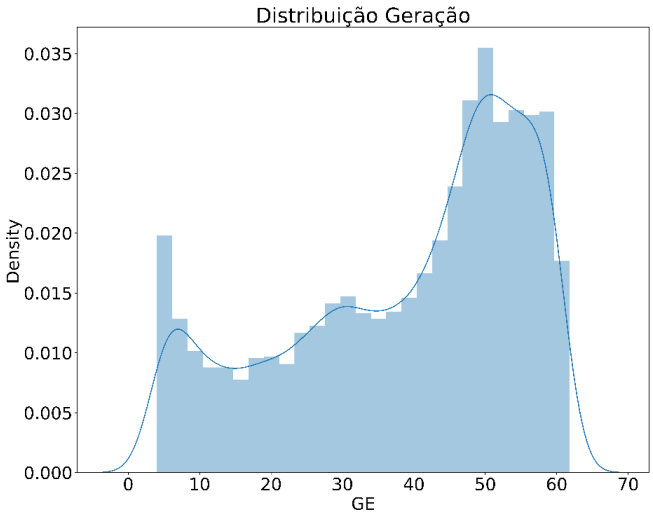
O ferramental computacional utilizado para realização dessas visualizações, foram as bibliotecas *Seaborn* e *Matplotlib.pyplot*, ambos presentes na linguagem de programação Python. Vale ressaltar que ambas as bibliotecas estão sendo muito utilizadas devido à robustez na execução das análises (LEE et al, 2019).

Através da biblioteca do *Seaborn*, foram aplicadas funções como *DistPlot*, *PairPlot* e *BoxPlot* as quais têm a função de apresentar a distribuição das variáveis e de correlacionar as principais variáveis explicitando uma possível correlação.

Nas Figuras 12, 13 e 14, temos o resultado da função *DistPlot* aplicada nas seguintes variáveis: “Geração Elétrica", “Radiação Global” e “Temperatura Média”, tornando a distribuição mais clara.

Na Figura 12 exibe-se os valores de densidade da variável alvo (Geração). Vemos que a distribuição atinge um intervalo entre 0 e 63 MWm. Além disso, percebe-se que a maior densidade se concentra entre a banda de 45 e 60 MWm.

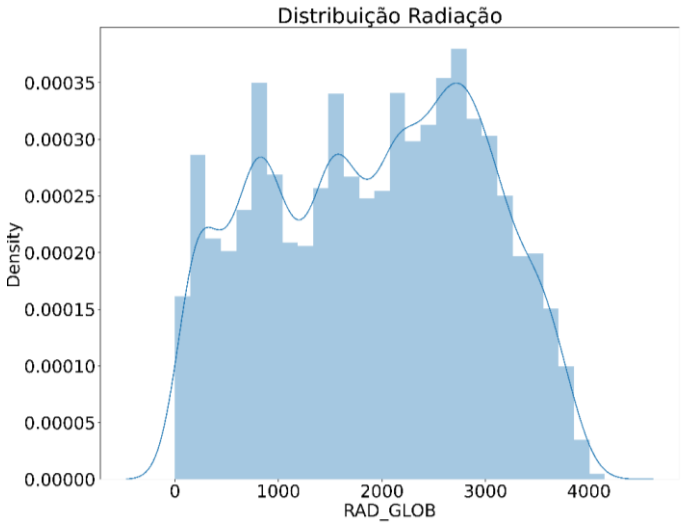
Figura 12: Gráfico da distribuição da GE



Fonte: Próprios autores.

Na Figura 13, apresentamos a distribuição da Radiação Global, que possui intervalo de 0 a 4000 kJ/m², a maior concentração dos dados ocorre entre 1500 kJ/m² e 3500 kJ/m².

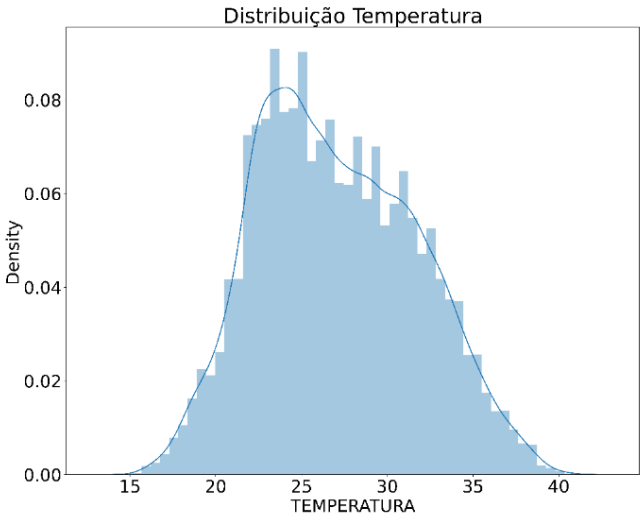
Figura 13: Gráfico de distribuição da Radiação Global



Fonte: Próprios autores.

Já na Figura 14, temos a distribuição dos dados de Temperatura. Nota-se que a maior parte dos dados de geração se concentram entre 20 e 25 °C.

Figura 14: Gráfico da distribuição da Temperatura.

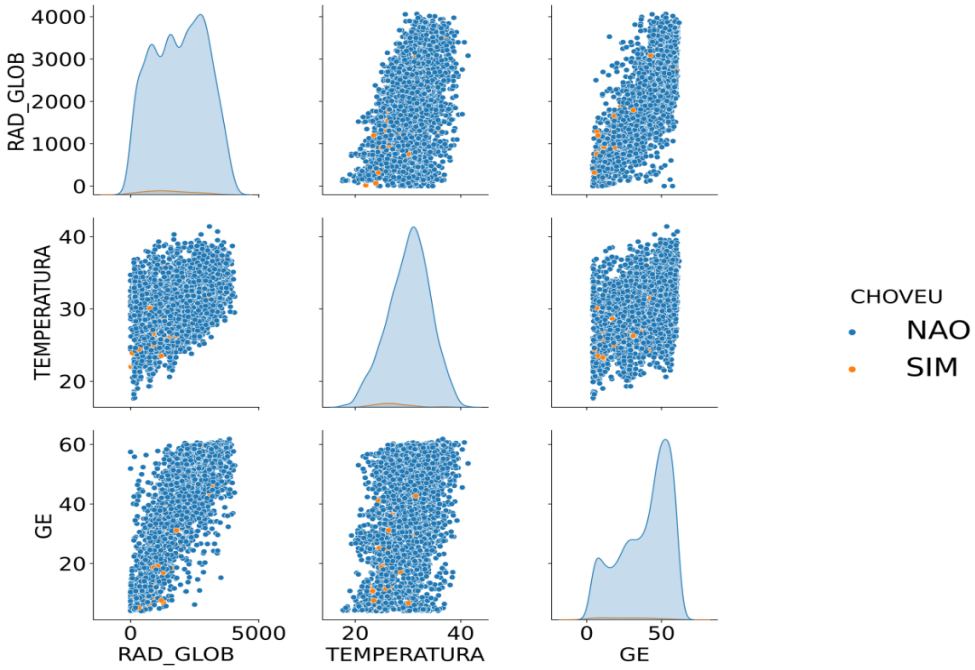


Fonte: Próprios autores.

A Figura 15 apresentada a seguir, nos mostra os resultados obtidos com a aplicação da função *PairPlot*, a fim de elucidar as correlações entre as principais variáveis da base.

Primeiramente, observa-se que a região em estudo apresenta baixa pluviosidade. Assim como esperado[5], esses dias (chuvosos) acarretam na diminuição das variáveis climatológicas Radiação e Temperatura, o que consequentemente prejudica a geração de energia do parque solar. Além disto, podemos notar significativa relação positiva entre a Radiação Global e a Geração Elétrica, algo também esperado, já que a variável preditora representa a incidência de energia por área (WHITAKER et al, 1991).

Figura 15: Aplicação da função PairPlot para análise de dependências entre as variáveis.

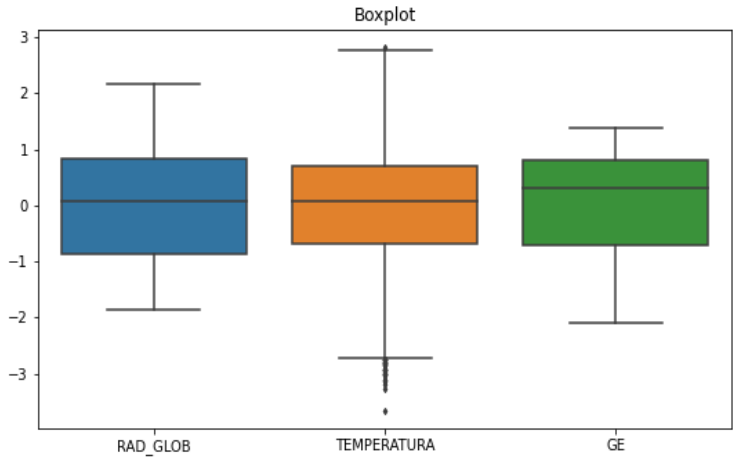


Fonte: Próprios autores.

Em sequência, foi aplicada a função *Boxplot,* que possui a opção de representar tanto o comportamento estatístico dos dados como ainda revelar a presença de *Outliers*. A Figura 16 ilustra o comportamento das principais variáveis estudadas.

[5] Disponível em: https://pt.climate-data.org/america-do-sul/brasil/bahia/bom-jesus-da-lapa-43241/. Acesso em: 4 de dez de 2021

Figura 16: Boxplot das principais variáveis do banco de dados



Fonte: Próprios autores.

Com base no comportamento estatístico apresentado, podemos afirmar que as variáveis “Radiação Global” e “Geração Elétrica - GE” não apresentaram pontos fora da curva, o que expõe um padrão bem definido do ponto de vista estatístico, elevando assim a qualidade do banco de dados. Por outro lado, a variável “Temperatura” apresentou *Outliers*. Assim, mesmo esse fato sendo prejudicial, pode ser desprezado, visto que as dimensões da evasão são pequenas.

Como já mencionado, o banco de dados final utilizado neste trabalho contém, ao todo, 14 variáveis, porém, a análise exploratória dos dados foi direcionada às apenas quatro delas: “Radiação Global”, "Temperatura", "Geração Elétrica” e “Clima “, por possuírem maior peso no desempenho da predição.

Vale ressaltar que a variável “Chuva Total” recebeu uma análise customizada, devido à sua alta volatilidade e baixa padronização, transformando-a em uma variável classificatória "Variáveis Dummies”. Essa utilização foi possível com a separação em duas classes.

5.4. IMPLEMENTAÇÃO DOS MODELOS PREDITIVOS (HIPERPARÂMETROS)

Nesta seção, são apresentados os resultados obtidos com a implementação da técnica *Random Search*, nos quatro modelos preditivos em estudo, sendo eles: XGB, MLP, RF e SVR. De acordo com o planejamento (Seção 3), o uso desta técnica possibilitou a otimização da predição dos modelos ao definir iterativamente os melhores parâmetros para cada um.

5.4.1. EXTREME GRADIENT BOOSTING (XGB)

Primeiramente, temos o *tuning* do modelo Aumento de Gradiente, com número de iterações (k = 100), *cross validation* ($cv = 4$), e 13 *features* previsoras (originais junto às artificiais, que foram agregadas com o uso da Engenharia de Recursos - Seção 5.2.2). Dessa forma, os melhores parâmetros alcançados foram:

- Número de Árvores: 1500;
- Número mínimo de amostras necessárias para dividir um nó interno: 3;
- Profundidade Máxima da Árvore: 6;
- Taxa de aprendizado: 0.01 (Quanto maior a taxa, menor a influência de cada árvore);
- Gamma = 0.1 (Redução de perda mínima necessária para fazer uma partição adicional em um nó);
- Número mínimo de instâncias em cada nó: 2 (Quanto maior, mais conservador será o algoritmo)

5.4.2. REDES NEURAIS ARTIFICIAIS (MLP)

O segundo modelo analisado foi a Redes Neurais Artificiais (*Multilayer Perceptron*), onde foi submetido a todos os processos de ajuste de hiperparâmetros explicados na Seção (4.4).

Foi aplicada a função do *Random Search*, o qual nos retornou a melhor configuração dos parâmetros do modelo para solucionar o problema proposto, as configurações do *Random Search* foram: número interações $k = 100$, *cross validation* ($cv = 4$) e 13 *features* previsoras.

Com isso os melhores parâmetros obtidos para o modelo foram:

- Função de Ativação = logistic;
- Número de camadas ocultas = 20;
- Modo de aprendizado = adaptive;
- Número máximo de iterações = 500;
- Solucionador para otimização de peso= lbfgs;

5.4.3. FLORESTAS ALEATÓRIA (RF)

Em seguida, temos o ajuste do modelo *Random Forest*, com número de iterações $k = 100$, *cross validation* ($cv = 4$) e 13 *features* previsoras. Dessa forma, os melhores parâmetros alcançados para o referido modelo de AM foram os seguintes:

- Número de Árvores da Floresta: 1500;
- Número mínimo de amostras necessárias para dividir um nó interno: 2;
- Profundidade Máxima da Árvore: 16;
- Número de recursos a serem utilizados: “Auto” (todas as *features* do *dataset*);
- Amostras de Autoinicialização: “True” (toda a amostra é utilizada para a criação da árvore).

5.4.4. MÁQUINA DE VETORES DE SUPORTE (SVR)

Finalmente, temos a otimização do modelo Máquinas de Vetores de Suporte, com número de iterações $k = 100$, *cross validation* ($cv = 4$) e 13 *features* previsoras. Dessa forma, os melhores parâmetros alcançados foram:

- Kernel: “rbf” (Especifica o tipo de kernel a ser usado no algoritmo);

- Gamma: “auto” (Coeficiente de kernel, se “auto” = usa 1 / n° features);
- Parâmetro de Regularização: 1;
- Limite máximo de iterações: “-1” (Quando “-1”, não possui limite).

Salienta-se, ainda, que cada modelo possui diferentes parâmetros a serem definidos devido aos seus distintos modos de funcionamento.

5.5. APLICAÇÃO DOS MODELOS NA BASE DE DADOS

Nesta seção serão exibidos os resultados alcançados com a utilização dos 4 modelos de Aprendizado de Máquina na predição da geração de energia (MWm).

A proporção treino/teste de todas as aplicações foi de 3/1 (3 anos de treino, 1 ano de teste), com o período de treino de 01/01/2017 a 31/12/2019, e o período de teste de 01/01/2020 a 31/12/2020. Para a validação dos resultados, foram realizadas análises visuais e quantitativas a partir dos dados segregados para esse fim (de validação), além do emprego de métricas de qualidade da área estatística como o Mean Absolute Error (MAE), Rooted Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) e consequentemente a Accuracy (100-MAPE %) (UYSAL et al, 2019).

O MAE tem como propsota medir a magnitude média dos erros em um determinado conjunto de previsões: é a média sobre a amostra de verificação dos valores absolutos das diferenças entre a previsão e a observação correspondente. Já o RMSE, trata-se de uma medida baseada em uma regra de pontuação quadrática, que mede a magnitude média do erro. Como os erros são elevados ao quadrado antes de serem calculados, o RMSE atribui um peso relativamente alto aos grandes erros.

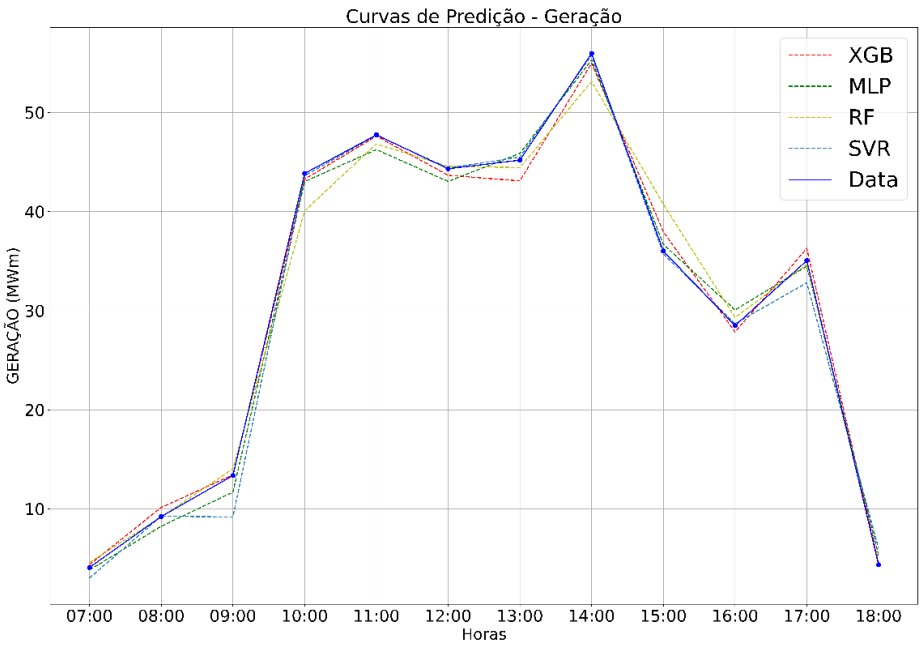
Ambas as funções podem ser usadas em conjunto para diagnosticar a variação dos erros em um conjunto de previsões. O RMSE será sempre maior ou igual ao MAE; quanto maior a diferença entre eles, maior a variância dos erros individuais da amostra, ou seja, se o RMSE for igual ao MAE, então todos os erros são da mesma magnitude. A métrica MAPE é uma medida de quão preciso é um sistema de previsão, basicamente mede a precisão como uma

porcentagem e pode ser calculado como o erro percentual absoluto médio, para cada período previstos menos os valores reais, divididos pelos valores reais.

5.5.1. PREVISÕES FINAIS OBTIDAS

Na Figura 17, são apresentadas as predições dos quatro modelos (XGB, MLP, RF e SVR), no período entre 06:00 às 17:00 do dia 05/01/2020, em comparação com os dados reais de teste do respectivo período.

Figura 17: Predição da Geração Elétrica.



Fonte: Próprios autores.

Primeiramente, observa-se que os resultados obtidos são bastante precisos, exibindo uma alta taxa de precisão em relação aos dados reais. Visualmente, vemos que nenhum modelo se destaca positivo ou negativamente na lacuna entre os dados reais e a predição, mantendo um elevado grau de assertividade.

Na Tabela 2, são apresentadas as métricas de qualidade alcançadas pelos modelos na predição para a base de dados.

Tabela 1: Métricas de qualidade - Predição da Geração Elétrica.

Modelos	MAE (MWm)	RMSE(MWm)	MAPE (%)	Acurácia (%)
RF	1.0241	1.6096	3.2475	96.7524
SVR	0.3756	1.0713	2.6153	97.3846
XGB	0.8264	1.1476	3.0066	96.9336
MLP	0.9455	1.2961	3.8140	96.1859

Fonte: Próprios autores.

Com base nos dados sumarizados acima, podemos ranquear os modelos através dos resultados apresentados. O principal destaque foi o modelo SVR, que melhor se adaptou a problemática proposta neste trabalho, obtendo os menores erros. Em segundo lugar, tem-se o modelo de XGB, seguido pelo RF. Finalmente, o modelo menos acurado foi o MLP. Vale ressaltar que todos os modelos obtiveram alta assertividade, quando os comparamos com a literatura vigente deste trabalho.

5.6. ESTUDO DO IMPACTO DO DESVIO PADRÃO DAS VARIÁVEIS AO DESEMPENHO

Parâmetros relacionados ao desvio padrão elucidam se a variável em estudo mantém um comportamento com muita ou pouca variação entres os dados. Bases de dados com altos desvios costumam afetar de maneira significativa o desempenho de modelos preditivos, dessa forma, quanto menor o desvio (dados mais padronizados), obtemos uma melhor interpretação dos modelos.

Dessa maneira, estes fatores serão estudados avaliando o impacto que podem ocasionar no desempenho das predições.

Os fatores estudados são:

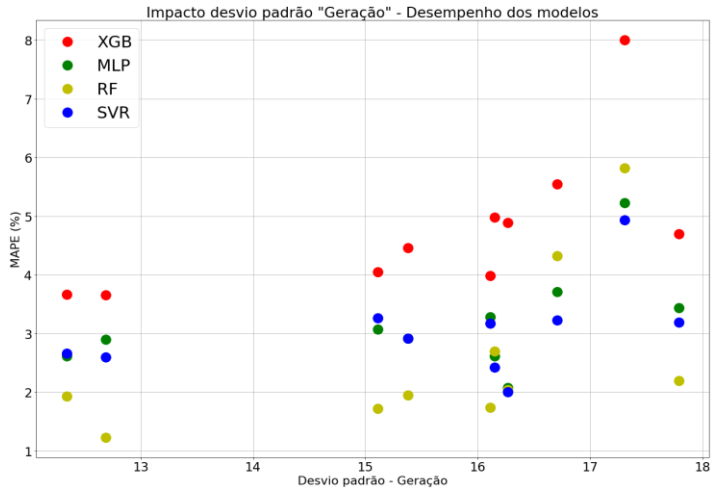
- Desvio padrão da Geração (MWm);
- Desvio padrão da Radiação Global (kJ/m²);

Para realização deste estudo, precisamos comparar períodos para entender como um desvio padrão mais elevado pode afetar o desempenho dos modelos, entretanto, a pesquisa se baseia em apenas uma base de dados. Portanto, houve a necessidade de selecionar amostras aleatórias de mesmo tamanho como fator de comparação.

Assim, do período de teste de 1 ano (¼ do tempo de estudo - Seção 5.5). Foram coletadas 10 amostras de 300 horários cada, o que possibilitou o cálculo do desvio padrão para cada intervalo e seus respectivos impactos (MAPE) para os modelos.

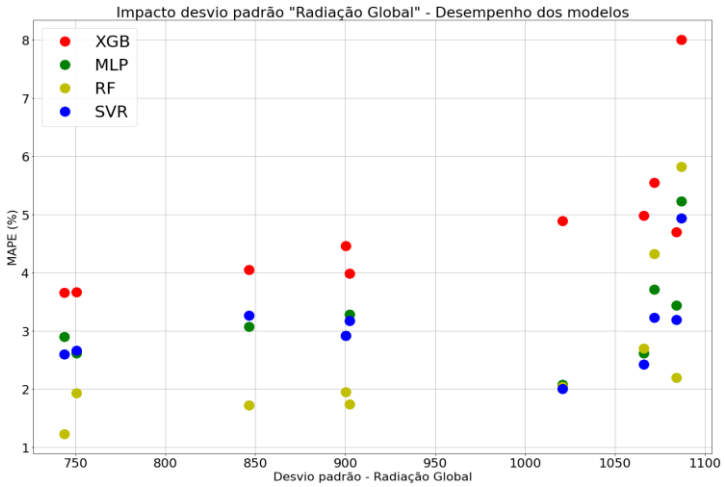
Nas Figuras 18 e 19 são apresentados os impactos ocasionados em cada modelo preditivo pelas variações da Geração (18) e Radiação (19).

Figura 18: Desvio padrão da Geração x MAPE.



Fonte: Próprios autores.

Figura 19: Desvio padrão da Radiação x MAPE.



Fonte: Próprios autores.

Nota-se, que os modelos apresentam um comportamento similar na análise das duas variáveis. É possível observar, portanto, uma boa correlação entre o aumento do desvio padrão com a diminuição do desempenho (aumento do MAPE).

O modelo XGB é o mais afetado pelo desvio de ambas as variáveis em estudo, exibindo erros elevados em função do aumento delas. Exibindo comportamento similar, mas em menor grau de impacto, temos o modelo RF, que aparenta resistência até certo nível de desvio, mas acaba sofrendo alto impacto após um aumento considerativo.

Por outro lado, os modelos SVR e MLP apresentam menor variação em sua assertividade em função do desvio, mantendo uma correlação de baixo impacto.

Em conjunto com as Figuras 18 e 19, a Tabela 3 traz as informações relacionadas ao desvio padrão das variáveis e a acurácia atingida pelos modelos na predição do alvo geração.

Tabela 2: Desvio padrão da Geração e Radiação Solar - Impacto (MAPE %) em cada modelo

GERAÇÃO	Rad. Solar	XGB	MLP	RF	SVR
17.794	1084.118	4.698	3.441	2.201	3.191
16.111	902.423	3.995	3.279	1.743	3.168
15.107	846.660	4.054	3.067	1.725	3.264
12.340	750.627	3.665	2.615	1.927	2.658
12.693	743.735	3.663	2.899	1.231	2.604
15.383	900.335	4.457	2.916	1.949	2.917
16.272	1020.862	4.893	2.076	2.041	2.002
16.150	1066.055	4.979	2.622	2.702	2.420
16.709	1071.798	5.547	3.714	4.318	3.229
17.308	1086.783	7.998	5.231	5.818	4.936

Fonte: Próprios autores.

Confirmando o comportamento exibido no gráfico, vemos que o modelo mais afetado é o XGB, apresentando range de 3.663% (menor desvio da Radiação) a 7.998% (maior desvio) em função do aumento do desvio. Porém, assim como o XGB, os outros 3 modelos apresentaram certo grau de perda de desempenho. O melhor modelo a lidar com esse aumento de desvio foi o RF, apresentando baixa variação de MAPE em todos os períodos.

Com isso, assim como esperado e discutido anteriormente, vemos que a variável Radiação tem um grande impacto na predição dos modelos, podendo ser classificada como a principal preditora da geração solar.

6. CONSIDERAÇÕES FINAIS

Após a aplicação dos métodos preditivos na base construída para este estudo, observou-se que os modelos atingiram resultados de elevada acurácia, mantendo uma proximidade elevada com os dados reais, atingindo valores para a métrica MAPE que estão limitados à faixa de 2% a 4%.

Entre os quatros modelos implementados e treinados, o Máquina de Vetores de Suporte (*Support Vector Machine* - SVR) foi aquele com maior desempenho, apresentando uma acurácia de 97.385%. Em seguida, temos o modelo Aumento Extremo do Gradiente (*Extreme Gradient Boosting* - XGB), que atingiu acurácia de 96.934%. O modelo Florestas Aleatórias (*Random Forest* - RF), com 96.752%. E por último, mas ainda assim com um ótimo nível de assertividade, temos o modelo Perceptron Multicamadas (*Multilayer Perceptron* - MLP), que exibiu uma acurácia de 96.186%.

Além disto, notou-se que no estudo realizado sobre o impacto do desvio padrão das variáveis, os modelos reagem de maneiras distintas à elevação do nível do desvio. Como discutido, os modelos XGB e RF apresentaram maior variação do desempenho, enquanto os modelos SVR e MLP exibiram resultados mais controlados em desvios mais altos.

Outro ponto observado nesta análise é a confirmação da dependência da variável alvo (Geração) com a radiação global, algo que já era esperado devido à natureza de ambas, mas que acabou se sustentando na predição dos modelos, ou seja, a radiação tem um ótimo fator preditivo quando analisamos o comportamento da variável alvo.

Apesar do modelo RF apresentar ótimo desempenho em níveis mais baixos de desvio, dados reais de geração fotovoltaica costumam ser extremamente voláteis (alto desvio), o que desencorajaria em certo grau o seu uso em larga escala.

Dessa forma, devido a sua excelente assertividade, alta precisão e ter baixo nível de impacto com altas variações, o modelo SVR mostrou-se mais robusto e apto para garantir uma predição mais coesa com os dados reais neste tipo de aplicação.

Como forma de dar continuidade a este trabalho, torna-se interessante a realização de uma pesquisa estudando a construção da integração dos modelos para com a usina, com foco em estudo de previsibilidade e analise de padrões de geração.

7. REFERÊNCIAS

ABSOLAR – Associação Brasileira de Energia Solar Fotovoltaica – Geração Distribuída Fotovoltaica Cresce 230% ao ano no Brasil, Disponível em: http://www.absolar.org.br/noticia/noticias-externas/geracao-distribuida-fotovoltaica-cresce-230-ao-ano-no-brasil.html. Acesso em: 12 jul. 2021.

ACADEMIA DO SOL (ed.). Crescimento da Energia Solar no Mundo. In: Crescimento da Energia Solar no Mundo. [S. l.], 2019. Disponível em: http://academiadosol.com.br/blog/crescimento-da-energia-solar-no-mundo/. Acesso em: 12 jul. 2021.

ATKESON, Andrew; KEHOE, Patrick J. The transition to a new economy after the second industrial revolution. 2001.

AWAN, Saqib E. et al. Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death. PloS one, v. 14, n. 6, p. e0218760, 2019.

BERGSTRA, James; BENGIO, Yoshua. Random search for hyper-parameter optimization. Journal of machine learning research, v. 13, n. 2, 2012.

CARVALHO, Monica et al. Potential of photovoltaic solar energy to reduce the carbon footprint of the Brazilian electricity matrix. LALCA: Revista Latino-Americana em Avaliação do Ciclo de Vida, v. 1, n. 1, p. 64-85, 2017.

CHANDRAN, Venkatesan et al. State of charge estimation of lithium-ion battery for electric vehicles using machine learning algorithms. World Electric Vehicle Journal, v. 12, n. 1, p. 38, 2021.

CHEN, Tianqi; GUESTRIN, Carlos. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016. p. 785-794.

CHOU, Jui-Sheng; TRAN, Duc-Son. Forecasting energy consumption time series using machine learning techniques based on usage patterns of residential householders. Energy, v. 165, p. 709-726, 2018.

DESTEK, Mehmet Akif; SINHA, Avik. Renewable, non-renewable energy consumption, economic growth, trade openness and ecological footprint: Evidence from organisation for economic Co-operation and development countries. Journal of Cleaner Production, v. 242, p. 118537, 2020.

EHSAN, R. Muhammad; SIMON, Sishaj P.; VENKATESWARAN, P. R. Day-ahead forecasting of solar photovoltaic output power using multilayer perceptron. Neural Computing and Applications, v. 28, n. 12, p. 3981-3992, 2017.

ELAVARASAN, Dhivya; VINCENT, Durai Raj. Reinforced XGBoost machine learning model for sustainable intelligent agrarian applications. Journal of Intelligent & Fuzzy Systems, n. Preprint, p. 1-16, 2020.

EPE. Matriz Energética e Elétrica. In: Matriz Energética e Elétrica. [S. l.], 2020. Disponível em: https://www.epe.gov.br/pt/abcdenergia/matriz-energetica-e-eletrica. Acesso em: 12 jul. 2021.

FAN, Junliang et al. Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: A case study in China. Energy conversion and management, v. 164, p. 102-111, 2018.

ISMAIL, Mohammad Ridwan et al. A multi-layer perceptron approach for customer churn prediction. International Journal of Multimedia and Ubiquitous Engineering, v. 10, n. 7, p. 213-222, 2015.

JIA, Zhunan et al. Critical review of volatile organic compound analysis in breath and in vitro cell culture for detection of lung cancer. Metabolites, v. 9, n. 3, p. 52, 2019.

JOHNSSON, Filip; KJÄRSTAD, Jan; ROOTZÉN, Johan. The threat to climate change mitigation posed by the abundance of fossil fuels. Climate Policy, v. 19, n. 2, p. 258-274, 2019.

LAUGS, Gideon AH; BENDERS, René MJ; MOLL, Henri C. Balancing responsibilities: Effects of growth of variable renewable energy, storage, and undue grid interaction. Energy Policy, v. 139, p. 111203, 2020.

LEE, Wei-Meng. Python machine learning. John Wiley & Sons, 2019.

LEME, João Vitor et al. Towards assessing the electricity demand in Brazil: Data-driven analysis and ensemble learning models. Energies, v. 13, n. 6, p. 1407, 2020.

LI, Kecheng et al. Research on Power System State Estimation Technology Considering Data Filling Technology. In: 2020 IEEE Sustainable Power and Energy Conference (iSPEC). IEEE, 2020. p. 2570-2576.

LUEMBA, Matias Emir. Análise exploratória e visualização de dados florestais brasileiros a partir do sistema DOF do IBAMA. 2021.

NEAGOE, Victor-Emil; CIOTEC, Adrian-Dumitru; CUCU, George-Sorin. Deep convolutional neural networks versus multilayer perceptron for financial prediction. In: 2018 International Conference on Communications (COMM). IEEE, 2018. p. 201-206.

NEMETH, Martin; BORKIN, Dmitrii; MICHALCONOK, German. The comparison of machine-learning methods XGBoost and LightGBM to predict energy development. In: Proceedings of the Computational Methods in Systems and Software. Springer, Cham, 2019. p. 208-215.

NGUYEN, Lien Thi Kim et al. Using XGBoost and skip-gram model to predict online review popularity. SAGE Open, v. 10, n. 4, p. 2158244020983316, 2020.

OBIORA, Chibuzor N.; ALI, Ahmed; HASAN, Ali N. Implementing Extreme Gradient Boosting (XGBoost) Algorithm in Predicting Solar Irradiance. In: 2021 IEEE PES/IAS PowerAfrica. IEEE, 2021. p. 1-5.

PAIM, Maria-Augusta et al. Evaluating regulatory strategies for mitigating hydrological risk in Brazil through diversification of its electricity mix. Energy Policy, v. 128, p. 393-401, 2019.

PALLONETTO, Fabiano et al. Demand response algorithms for smart-grid ready residential buildings using machine learning models. Applied energy, v. 239, p. 1265-1282, 2019.

PAULA, Matheus et al. Predicting Long-Term Wind Speed in Wind Farms of Northeast Brazil: A Comparative Analysis Through Machine Learning Models. IEEE Latin America Transactions, v. 18, n. 11, p. 2011-2018, 2020.

RADHIKA, Y.; SHASHI, M. Atmospheric temperature prediction using support vector machines. International journal of computer theory and engineering, v. 1, n. 1, p. 55, 2009.

RAMOS, Lucas G.; COLNAGO, Marilaine; CASACA, Wallace. Análise preditiva na geração distribuída de energia fotovoltaica: aplicações e algoritmos inteligentes. Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 8, n. 1, 2021.

RAMOS, Lucas; COLNAGO, Marilaine; CASACA, Wallace. Data-driven analysis and machine learning for energy prediction in distributed photovoltaic generation plants: A case study in Queensland, Australia. Energy Reports, v. 8, p. 745-751, 2022.

SAKER, Nathalie et al. Cost–benefit analysis of rooftop photovoltaic systems based on climate conditions of Gulf Cooperation Council countries. IET Renewable Power Generation, v. 12, n. 9, p. 1074-1081, 2018.

SAMADIANFARD, Saeed et al. Wind speed prediction using a hybrid model of the multi-layer perceptron and whale optimization algorithm. Energy Reports, v. 6, p. 1147-1159, 2020.

SHAHHOSSEINI, Mohsen; HU, Guiping; PHAM, Hieu. Optimizing ensemble weights and hyperparameters of machine learning models for regression problems. arXiv preprint arXiv:1908.05287, 2019.

SILVA, Walquíria do Nascimento; SCHRODER E BRAGA, Luís Gustavo. A evolução do setor elétrico brasileiro e mercado livre de energia. 2020.

SILVEIRA, Gabriel Eugênio de Aguiar. Aprendizado de máquina aplicado à predição de potência de geração distribuída na rede de distribuição de média tensão. 2021.

SUN, TOP (ed.). Geração de Energia Solar cresce no país e coloca o Brasil no TOP 15 mundial. In: Geração de Energia Solar cresce no país e coloca o Brasil no TOP 15 mundial. [S. l.], 23

set. 2021. Disponível em: https://g1.globo.com/sc/santa-catarina/especial-publicitario/top-sun/top-sun-energia-solar/noticia/2021/09/23/geracao-de-energia-solar-cresce-no-pais-e-coloca-o-brasil-no-top-15-mundial.ghtml. Acesso em: 7 jan. 2022.

TUFAIL, Muhamad Mutasim Billah; IBRAHIM, Jafni Azhan; MELAN, Mustakim. Conceptualizing energy security and the role of diversification as the key indicator against energy supply disruption. Journal of Advanced Research in Business and Management Studies, v. 11, n. 1, p. 1-9, 2018.

UYSAL, Ilhan; GÜVENIR, H. Altay. An overview of regression techniques for knowledge discovery. The Knowledge Engineering Review, v. 14, n. 4, p. 319-340, 1999.

VIEIRA, Ana Cândida Ferreira. Energias renováveis e sua eficiência na nova economia energética no Brasil. Revista Brasileira de Gestão Ambiental e Sustentabilidade, v. 8, n. 18, p. 211-223, 2020.

WHITAKER, C. M. et al. Effects of irradiance and other factors on PV temperature coefficients. In: The Conference Record of the Twenty-Second IEEE Photovoltaic Specialists Conference-1991. IEEE, 1991. p. 608-613.

WU, Dazhong et al. A comparative study on machine learning algorithms for smart manufacturing: tool wear prediction using random forests. Journal of Manufacturing Science and Engineering, v. 139, n. 7, 2017.

YANG, Haiqin; CHAN, Laiwan; KING, Irwin. Support vector machine regression for volatile stock market prediction. In: International Conference on Intelligent Data Engineering and Automated Learning. Springer, Berlin, Heidelberg, 2002. p. 391-396.

ZAFAR, Muhammad Wasif et al. From nonrenewable to renewable energy and its impact on economic growth: the role of research & development expenditures in Asia-Pacific Economic Cooperation countries. Journal of cleaner production, v. 212, p. 1166-1178, 2019.

ZHENG, Huan; WU, Yanghui. A xgboost model with weather similarity analysis and feature engineering for short-term wind power forecasting. Applied Sciences, v. 9, n. 15, p. 3019, 2019.

ZIANE, Abderrezzaq et al. Photovoltaic output power performance assessment and forecasting: Impact of meteorological variables. Solar Energy, v. 220, p. 745-757, 2021.