GABRIEL DE FREITAS VISCONDI

PREDIÇÃO DE RADIAÇÃO SOLAR USANDO ALGORITMOS DE APRENDIZAGEM DE MÁQUINA E PARÂMETROS METEOROLÓGICOS

Dissertação de Mestrado apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do título de Mestre em Ciências

SÃO PAULO

2022

GABRIEL DE FREITAS VISCONDI

PREDIÇÃO DE RADIAÇÃO SOLAR USANDO ALGORITMOS DE APRENDIZAGEM DE MÁQUINA E PARÂMETROS METEOROLÓGICOS

Versão Corrigida

Dissertação de Mestrado apresentada à Escola Politécnica da Universidade de São Paulo para obtenção do título de Mestre em Ciências

Área de concentração: Engenharia da Computação

Orientadora: Solange Nice Alves de Souza

SÃO PAULO

2022

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, 28 de dezembro de 22

Assinatura do autor: 

Assinatura do orientador: 

Catalogação-na-publicação

de Freitas Viscondi, Gabriel
Predição de Radiação Solar Usando Algoritmos de Aprendizagem de Máquina e Parâmetros Meteorológicos / G. de Freitas Viscondi -- versão corr. -- São Paulo, 2022.
97 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais.

1.Aprendizagem de máquina 2.Redes neurais 3.Energia solar fotovoltaica I.Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Computação e Sistemas Digitais II.t.

AGRADECIMENTOS

No início do desenvolvimento dessa dissertação, o meu perfil profissional no universo da engenharia de computação e do trabalho com dados começava a se materializar. Era uma mistura de euforia, anseio e receio por ter escolhido este foco de atuação que rapidamente traria conquistas, frustrações e ainda mais sonhos para minha vida.

Ironicamente, transitando no mais próximo contato com dispositivos na era da inteligência artificial, recentemente ao acordar, dei bom dia à Alexa em minha cozinha que respondeu em bom tom: dê valor aos momentos de dificuldade e ausência de calmaria, pois são exatamente esses que serão eternamente lembrados.

Era para ser um dia como outro qualquer, mas sou subitamente inundado mentalmente por imagens e recordações de alguns episódios de dificuldade enquanto o cheiro de café começava a transitar pelo ambiente. Dentre eles, lembro-me de ser tomado por um sentimento de gratidão ao lembrar da minha transição profissional para o foco em tecnologia que se materializa em tudo que vivi para escrever este trabalho.

Agradeço imensamente todas as pessoas que fizeram parte da longa jornada em direção ao meu sonho de estudar na Escola Politécnica da Universidade de São Paulo.

Primeiramente, sou muito grato aos meus pais, Ronaldo e Rosana, que me conduziram na vida pelo exemplo e são para mim fonte vital de valores, de crença inabalável em minha pessoa e por terem me mostrado desde cedo o papel da dedicação em nossos sonhos.

Agradeço meu irmão Thiago, por ser meu exemplo desde pequeno, por ter me apoiado diversas vezes em meus desafios e por ter me ensinado a importância da consistência e profundidade em minha personalidade.

Agradeço meu irmão Lucas, por me ensinar que a vida pode ser leve, divertida e fluida mesmo em momentos de tormenta e me permitir criar constantemente uma versão melhor e atualizada de mim mesmo.

Agradeço a Luiza, minha amiga, companheira e esposa por me mostrar, iluminar e engrandecer nosso caminho, nos permitindo constantemente lembrar o quanto tudo que estamos construindo tem superado nossas maiores expectativas e sonhos.

Em especial, agradeço minha avó Therezinha por ter me ensinado o valor da generosidade. A toda minha família, obrigado pelo apoio incondicional, pelos momentos compartilhados e por estarem ao meu lado.

A minha orientadora Solange Nice, agradeço a paciência, os ensinamentos e por trazer elementos essenciais para compor a pessoa e profissional que sou hoje.

Por fim, deixo meu agradecimento e celebro todas as pessoas que encontrei nesses anos recentemente vividos presencialmente e virtualmente na Universidade de São Paulo. Guardo todos com muito carinho em minhas recordações e exalto o prazer de nossos caminhos terem se cruzado.

Viver é partir, voltar e repartir.
Partir, voltar e repartir.

É tudo pra ontem
(Emicida e Gilberto Gil)

RESUMO

A penetração de energias renováveis é fundamental para uma transição de matriz energética ambientalmente correta, economicamente próspera e socialmente justa. Dentro deste contexto, a energia solar fotovoltaica deve desempenhar um papel substancial nos próximos anos e, muito embora tenha avançado a passos largos dado sua modularidade e abundância de recurso natural, a variabilidade intrínseca a fonte solar é um dos fatores limitativos basais para a sua integração em sistemas elétricos.

Prever a geração solar fotovoltaica, portanto, torna-se uma das condições chave para reduzir os impactos da variabilidade da fonte, facilitando sua integração na matriz elétrica mundial. Para tal, técnicas de aprendizado de máquina são amplamente conhecidas por sua capacidade de previsão e, nesta condição, foram estudadas por este trabalho de pesquisa para entender o estado da arte dos algoritmos empregados para previsibilidade de geração solar fotovoltaica e aplicar as melhores práticas encontradas na literatura em contexto brasileiro.

Em um estágio inicial do trabalho, realizou-se uma revisão de escopo da literatura para identificar como o problema de variabilidade de geração de energia solar fotovoltaica estava sendo abordado conjuntamente com modelos de aprendizado de máquina. A partir de questões estruturadas de pesquisa, selecionou-se por meio de critérios objetivos 38 artigos para revisão. Como resultado identificou-se que Máquinas de Vetores de Suporte, Redes Neurais Artificiais e Máquinas de Aprendizado Extremo eram os algoritmos mais utilizados para o problema de previsibilidade proposto. Por fim, entendeu-se que a modelagem de dados de parâmetros meteorológicos para a região em estudo era a abordagem mais utilizada dado a próxima relação entre a disponibilidade do recurso natural de irradiação e a geração de eletricidade pela conversão fotovoltaica.

Em sequência, os três algoritmos foram implementados, criando-se modelos de previsão fundamentados em dados de parâmetros meteorológicos medidos na cidade de São Paulo. No cenário proposto, constatou-se que a inclusão de todos os

parâmetros meteorológicos disponibilizados para a região era a configuração de modelagem com melhores resultados. Nesta configuração, os três algoritmos implementados possuem acurácia de previsão próximas, com valores para correlação de Pearson entre irradiação prevista e observada entre 0,87 e 0,89. Entretanto, existe ligeira vantagem para os modelos implementados com o algoritmo SVM quando comparados os resultados nas métricas de erro propostas (RSME e MAE).

Palavras-chave: Aprendizado de máquina, redes neurais, máquinas de aprendizado extremo, máquinas de vetores de suporte, previsão de energia solar fotovoltaica

ABSTRACT

The penetration of renewable energy generation sources is fundamental for a transition to an environmentally correct, economically prosperous, and socially equal electricity mix. Within this context, photovoltaic solar energy must play a substantial role in the coming years and, although it has advanced rapidly given its modularity and abundance of natural resources, the intrinsic variability of the solar source is one of the basic limiting factors for its integration into power systems.

Predicting photovoltaic solar generation, therefore, becomes one of the key conditions to reduce the impacts of the source variability, facilitating its integration into the global electrical mix. Machine learning methods are widely known for their prediction capability, and, in this condition, these techniques were studied by this research project aiming to understand the state-of-art algorithms used for predicting solar photovoltaic energy generation and to apply the best practices in a Brazilian context.

In an initial stage of the work, a literature scoping review was conducted to identify how the problem of variability of photovoltaic solar energy generation was being approached together with machine learning models. Based on structured research questions, 38 articles were selected through objective criteria for review. As a result, it was identified that Support Vector Machines, Artificial Neural Networks and Extreme Learning Machines were the most used algorithms to address the proposed predictability problem. Finally, it was understood that the modeling of meteorological parameters for the region under study was the most used approach given the relationship between the availability of the natural irradiation resource and the generation of electricity by photovoltaic conversion.

In sequence, the three most frequently cited algorithms in the literature were implemented, creating forecast models based on data from meteorological parameters measured in the city of São Paulo. In the proposed scenario, it was found that the inclusion of all meteorological parameters available for the region was the modeling

configuration with the best results. In this configuration, the three implemented algorithms have close prediction accuracy, with values for Pearson's correlation between predicted and observed irradiation ranging from 0.87 to 0.89. However, there is a slight advantage for the models implemented with the SVM algorithm when comparing the results for the proposed error metrics (RSME and MAE).

Keywords: Machine Learning, neural networks, extreme learning machines, support vector machines, solar photovoltaic energy forecasting

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANEEL	Agência Nacional de Energia Elétrica
ANN	Artificial Neural Networks
DAMA	Data Management Association
dirdom	Direção Predominante dos Ventos
dirrajd	Direção Rajada Diária
duração	Precipitação Horária
ELM	Extreme Learning Machines
GB	Gradient Boosting
GEE	Gases de Efeito Estufa
IA	Inteligência Artificial
IAG-USP	Instituto de Astronomia, Geofísica e Ciências Atmosféricas da Universidade de São Paulo
IEMA	Instituto de Energia e Meio Ambiente
insol	Fração Horária de Brilho Solar
IRENA	International Renewable Energy Agency
irrad	Irradiação Solar Total Diária
LCOE	Levelized Cost of Energy
MAE	Mean Absolute Error
MIT	Massachusetts Institute of Technology
ML	Machine Learning
NWP	Numerical Weather Prediction
ONS	Operador Nacional do Sistema
P/R	Precision/Recall
prec	Precipitação Horária
press24	Pressão Atmosférica

QD	Qualidade dos Dados
rajd	Maior Rajada de Vento Diária
rajh	Maior Rajada de Vento Horária
RF	Random Forest
RMSE	Root Mean Square Error
REL	Revisão de Escopo da Literatura
SIN	Sistema Interligado Nacional
SLFN	Single Hidden Layer Feedforward Neural Network
SVM	Support Vector Machines
TAE	Teoria do Aprendizado Estatístico
temperatura	Temperatura do ar
tmax	Temperatura Máxima
tmin	Temperatura Mínima
tseco	Temperatura do Bulbo Seco
tsfc	Temperatura de Superfície
túmido	Temperatura do Bulbo Úmido
UR	Umidade Relativa
vmed	Velocidade Média do Vento
vmedm	Velocidade Média do Vento meridional

SUMÁRIO

1 INTRODUÇÃO

O uso de tecnologia tem sido extremamente efetivo no auxílio à tomada de decisão nos mais variados setores da economia, levando, dentre outros aspectos, ao aumento da produtividade, à redução de custos operacionais e à maior agilidade nas decisões, seja no poder público, no setor privado, na academia ou nas organizações do terceiro setor. A velocidade com que as decisões passam a ser tomadas faz com que os agentes envolvidos utilizem ferramentas e informações que permitam análises fundamentadas, que reduzam o risco de falhas interpretativas e que mostrem alternativas inexploradas.

Essa nova perspectiva tecnológica vem acompanhada de uma dimensão distinta de coleta e interpretação de dados, visto o crescimento acelerado de novos ambientes e fontes de aquisição: redes sociais, sensores, e-mails, transações eletrônicas, dentre outros, que acabam nos forçando a olhar os problemas sob outras perspectivas (McKinsey & Company, 2011). Essa intensa concentração de dados traz consigo variabilidade, velocidade e volume (Zhang, 2013), fazendo com que, neste âmbito, técnicas e tecnologias para *crowd sourcing, machine learning* e *data mining*, por exemplo, surjam para minimizar riscos de análises ou auxiliar a extração de informações em ambientes de *big data*.

Os resultados trazidos por estas ferramentas têm motivado o setor de energia a investir na área. Principalmente no que tange ao setor elétrico, o mundo deve vivenciar nos próximos anos uma revolução na maneira com que enxerga a geração e o consumo de eletricidade. Seja por motivo de desgastes ambientais, desenvolvimento tecnológico ou pressões de mercado, os novos modelos de negócio e regulamentações têm mudado o mercado de energia elétrica, que deve vivenciar intensamente um cenário disruptivo de descentralização da geração de energia elétrica, capitaneado pela utilização de fontes renováveis como solar, eólica e biomassa, bem como uma maior tendência de gestão de demanda por eletricidade (IEA, 2016).

Se por um lado essas fontes renováveis trazem novas perspectivas benéficas do ponto de vista tecnológico, mercadológico, ambiental e de pesquisa, por outro este novo modelo do setor elétrico, que começa a ser configurado com grande participação de fontes não despacháveis[1], traz alguns desafios e incertezas para a operação do Sistema Interligado Nacional (SIN) que impactam diretamente na segurança do fornecimento de energia elétrica (Fürstenwerth et al., 2015). Estes estão principalmente relacionados à variabilidade dos recursos naturais energéticos a elas vinculados – radiação solar e correntes de ventos – que chegam a variar abruptamente sua disponibilidade e consequente geração de eletricidade em horizontes temporais na casa dos segundos (Kelman, 2016).

Para viabilizar e compreender este paradigma de mercado, tecnologias inovadoras como *smart grids, smart meters,* sensores e modelos de previsibilidade dos comportamentos de geração e consumo, passam a ser essenciais visto a variabilidade das fontes renováveis, dependentes de fatores meteorológicos como radiação solar, ventos, sazonalidade das produções agrícolas e os novos comportamentos dos consumidores (International Energy Agency (IEA), 2017). Essas tecnologias trazem a reboque o desenvolvimento de aplicações específicas para a análise de dados em volume, velocidade e variedade - *big data* - até então pouco exploradas (Kleissl et al., 2012).

Nesse contexto, a fonte solar fotovoltaica apresenta papel relevante devido à capacidade de expansão acelerada por meio dos painéis residenciais e comerciais descentralizados. Esta perspectiva integra uma infinidade de agentes de geração e consumo no sistema elétrico, em localizações distintas, impactados por eventos

[1] Fontes de energia que não possuem a habilidade de produzir eletricidade quando determinado pelo operador do sistema elétrico, como no caso das fontes solar e eólica que geram energia quando há radiação incidente ou ventos.

meteorológicos adversos, e com caráter extremamente variável, alimentando o setor com uma grande quantidade de dados para a tomada de decisão (Shuo et al., 2016).

Na perspectiva da operação do sistema elétrico, torna-se fundamental a capacidade de previsão do comportamento desses milhares de novos agentes entrando frequentemente no sistema, seja para fins de planejamento, controle ou adequação das lógicas de operação (Haupt & Kosović, 2017). Algoritmos preditivos treinados com séries históricas de monitoramento dos mais diversos parâmetros meteorológicos são capazes de auxiliar a compreensão do comportamento da radiação solar e, consequentemente, da geração de eletricidade por cada um desses micro agentes distribuídos. Motiva-se daí, o presente trabalho de pesquisa.

1.1 Objetivos

Propõe-se neste trabalho a avaliação de técnicas de aprendizado de máquina para previsão de radiação solar com base em parâmetros meteorológicos, com vistas ao aprimoramento de modelos que buscam melhorar a previsibilidade de geração de energia elétrica em sistemas solares fotovoltaicos distribuídos.

Faz parte da metodologia uma revisão de escopo da literatura para confirmar os desafios de previsão de radiação solar, assim como, identificar as técnicas mais empregadas e com maiores ganhos de previsibilidade.

Em um contexto de aplicação brasileiro, esse trabalho busca avaliar a:

(i) influência do número de parâmetros meteorológicos de entrada nos modelos;
(ii) contribuição de cada parâmetro meteorológico na melhoria de previsibilidade de radiação solar;

1.2 Metodologia

O trabalho descrito por esta dissertação foi organizado metodologicamente da seguinte forma:

1. **Levantamento e análise das referências bibliográficas:** Nesta etapa, foi elaborada uma Revisão de Escopo da Literatura (REL), buscando-se responder as perguntas orientadoras da pesquisa (Kitchenham & Charters, 2007);
2. **Análise do problema proposto e dos algoritmos de solução:** Concentrou-se em analisar os resultados da REL propondo-se os entregáveis deste trabalho. Nesta etapa foram escolhidos os três algoritmos que seriam utilizados para modelagem (de Freitas Viscondi & Alves-Souza, 2019). Também se definiu as simulações para comparação de diferentes cenários.
3. **Levantamento da base de dados a ser utilizada:** Considerou-se a utilização somente de dados meteorológicos para a construção dos modelos de previsão, optando-se como fonte principal a utilização da base de dados fornecida pelo Instituto de Astronomia, Geofísica e Ciências Atmosféricas da Universidade de São Paulo (IAG-USP);
4. **Simulação comparativa do desempenho dos modelos:** Desenvolveu-se os modelos propostos para avaliação dos resultados de previsão. Após treinados, os modelos são avaliados por meio de métricas recorrentemente vistas na REL como *Root Mean Square Error* (RMSE), *Mean Absolute Error* (MAE) e *Precision/Recall* (P/R);
5. **Simulação comparativa entre o número de parâmetros meteorológicos utilizados:** Observou-se na REL, que diversos parâmetros meteorológicos foram utilizados para construir os modelos de previsão como umidade, pluviosidade, cobertura de nuvens, irradiação indireta, dentre outros. A escolha por esses parâmetros estava muito mais relacionada à disponibilidade dos dados do que pela contribuição clara dessas variáveis ao modelo. Assim, nesta etapa, os modelos foram avaliados com variações na quantidade e escolha de diferentes combinações de parâmetros

meteorológicos. Para avaliar quais combinações de parâmetros fornecia melhor previsão da radiação solar (explicam melhor a variável alvo da predição: radiação incidente), empregou-se diferentes métricas como RMSE, MAE e P/R, que possibilitam a comparação da acurácia preditiva dos modelos.

2 CONTEXTUALIZAÇÃO TÉORICA

Este capítulo busca contextualizar teoricamente os conceitos e abordagens utilizados por esta dissertação, trazendo descrições da literatura sobre fonte solar fotovoltaica para a geração de eletricidade e a crescente necessidade por tecnologia de análise e técnicas para o tratamento da qualidade de dados.

Neste contexto, também é apresentada a teoria que fundamenta os modelos de ML utilizados na composição de soluções que ensejam a previsibilidade de geração de energia por sistemas fotovoltaicos. Por fim, discute-se as definições de qualidade de dados pertinentes ao escopo desse trabalho.

2.1 Geração de eletricidade

Sistemas elétricos de potência são definidos como sistemas que contemplam as fases de geração, transmissão e distribuição de energia elétrica. Enquanto as fases finais de transmissão e distribuição são as responsáveis por transportar as correntes elétricas de um local para o outro, a primeira fase é a qual converte-se diferentes formas de energia em eletricidade. Essa conversão se dá em unidades espalhadas pelo território comumente chamadas de usinas (Junior, 2006).

No caso do Brasil, por exemplo, existe um sistema único para gerenciamento dessas fases de disponibilização de eletricidade. O SIN é um sistema elétrico de dimensões continentais, dividido em quatro subsistemas – Sul, Sudeste/Centro-Oeste, Norte e Nordeste, e que contempla toda a infraestrutura de geração e linhas de transmissão de energia elétrica do país (Zambon, 2015).

O SIN é controlado centralizadamente por uma entidade denominada Operador Nacional do Sistema (ONS), a qual é responsável, dentre outras atribuições, por organizar o despacho – usinas que devem entrar em operação em dado instante – da infraestrutura geradora de eletricidade.

Em sistemas elétricos de potência, diferentes fontes de energia são convertidas em eletricidade atuando como um portfólio de opções para um país, caracterizando o que se denomina matriz elétrica. Sendo assim, múltiplos fatores como disponibilidade de combustível, competitividade de preços, sustentabilidade ambiental, impactos sociais, maturidade tecnológica e capacitação de mão-de-obra, definem o perfil das matrizes elétricas no mundo, as quais se configuram com o objetivo de fornecerem energia elétrica aos menores preços possíveis (EPE, 2012).

Usinas popularmente conhecidas como hidroelétricas com reservatório e termoelétricas a carvão, gás natural ou biomassa são classificadas como fontes despacháveis de energia elétrica, possuindo a habilidade de produzir eletricidade quando determinado pelo operador do sistema elétrico. Esta capacidade é intrínseca à natureza da fonte de energia ou tecnologia utilizada na geração. Grosso modo, sendo necessária a introdução de energia elétrica no sistema para o consumo no curto prazo, mais gás natural pode ser adicionado às caldeiras das usinas termoelétricas, assim como novas comportas podem ser abertas para iniciar o funcionamento de turbinas hidráulicas em uma usina hidroelétrica.

No entanto, outras importantes fontes de geração de eletricidade **não são despacháveis**, ou seja, são usinas que não possuem caráter de operação condicionada pela vontade do operador. Neste grupo, podemos enquadrar as fontes de geração eólica e solar fotovoltaica, predominantemente pela natureza da fonte de energia (Soares, 2016). Essas fontes são consideras fontes variáveis de geração de eletricidade, visto que sua operação está condicionada a variações naturais dos recursos energéticos utilizados por suas usinas. Uma usina eólica, por exemplo, está disponível para a geração de eletricidade a todo momento, entretanto, mundialmente, em média, somente em cerca de 30% do tempo os ventos são capazes de gerar eletricidade (Boccard, 2009). No Brasil esse número é relativamente maior – 42% em 2019 – por conta da qualidade e quantidade de ventos nos parques eólicos nas regiões norte, nordeste e sul (ONS, 2019). A esta medida, dá-se o nome de **fator de capacidade**.

Entretanto, mundialmente, apesar de essas fontes renováveis não despacháveis concederem os inúmeros benefícios tecnológicos, econômicos e ambientais citados anteriormente, sua expansão e consequente ganho de representatividade nas matrizes elétricas traz custos de integração e outros problemas nas estruturas dos sistemas elétricos, principalmente devido ao caráter variável dos recursos naturais – sol e vento - a elas atrelados (Fürstenwerth et al., 2015).

De acordo com o Instituto de Energia e Meio Ambiente (IEMA), os principais problemas atrelados às fontes renováveis, solar e eólica, são (Cunha et al., 2016):

1. **Variabilidade:** de acordo com as condições climatológicas, a quantidade e qualidade do sol e do vento varia, implicando em alterações na quantidade de energia elétrica convertida pelos sistemas de geração. Estas variações permeiam muitas escalas de tempo, chegando, em seu limite inferior, a escala dos segundos;
2. **Maiores custos sistêmicos de geração:** a variabilidade e consequente incerteza atrelada a geração de energia elétrica por meio dessas fontes, gera, momentos de excesso ou de escassez de eletricidade no sistema. Como consequência, usinas termoelétricas de elevado custo de geração podem ficar ociosas ou operar intensamente para atender a demanda por eletricidade, gerando alta flutuação nos custos sistêmicos de geração. Este efeito é altamente impactante na tarifa paga pelo consumidor;
3. **Distribuição desigual no território:** Os potenciais de geração de energias eólica e solar estão distribuídos de maneira desigual pelo território nacional. Para a energia solar, em especial, o potencial está espalhado pelo Brasil. Desta forma, é necessário investimento em infraestrutura de transmissão para escoar a energia produzida pelo país, visto que nem sempre a geração está localizada próxima aos centros de carga[2];

[2] Centros de consumo de energia elétrica.

4. **Modularidade:** a possibilidade de modulação dos sistemas de geração solar e eólico é extremamente benéfica para o consumidor que passa a ser um agente ativo do sistema elétrico ao poder gerar sua própria eletricidade, adequando o projeto de geração à sua demanda. Entretanto, a descentralização da geração trará diversos desafios de adequação dos sistemas de transmissão e distribuição, que permeiam necessidades de evolução tecnológica e aplicação de altos investimentos para aprimoramento técnico do SIN;
5. **Patamares de tensão e ausência de sincronismo:** As redes de distribuição do sistema elétrico são construídas pensando na acomodação da demanda por eletricidade de maneira otimizada, na qual os níveis de tensão das redes são ajustados de acordo com a previsão de fluxos médios de eletricidade. Entretanto, com o crescimento da geração distribuída, as redes elétricas terão que se adaptar para os fluxos de energia elétrica em ambos os sentidos (produção e consumo), considerando a sazonalidade diária desses fluxos. Ainda, as fontes solar e eólica não são capazes de manter a qualidade das ondas elétricas em nível adequado (regime permanente), exigindo evolução tecnológica e regulamentação de serviços para além da geração de eletricidade (serviços ancilares[3]).

Impulsionadas pelos custos nulos de matéria-prima, impactos ambientais reduzidos na fase de geração, avanços tecnológicos e possibilidades de instalação modular, as fontes renováveis solar e eólica tem ganhado bastante relevância nos últimos anos. Grandes parques eólicos e solares fotovoltaicos foram instalados no território nacional, aumentando a participação dessas fontes na geração diária de eletricidade de milhares de brasileiros (IRENA, 2019). A essas usinas, damos o nome de usinas centralizadas de geração de eletricidade.

[3] Serviços que têm a finalidade de garantir a segurança e operação do sistema elétrico. Para a fonte solar fotovoltaica, vale ressaltar o papel dos sistemas de armazenamento de energia (baterias) que possuem relevância para o desenvolvimento da fonte dado a intermitência natural do recurso energético.

Em 2012, um marco para o setor, a modularidade de instalação de usinas de geração de eletricidade foi contemplada pela Resolução Normativa nº 482/2012 da agência Nacional de Energia Elétrica (ANEEL), definindo as normativas para sistemas de geração conectados diretamente à rede de distribuição, situados juntos aos pontos de carga (consumo) (ANEEL, 2012). Dá-se início a um novo paradigma do setor elétrico, no qual o consumidor passa a poder ser também um agente de geração por meio da geração distribuída. Este marco regulatório habilita a penetração da fonte solar na matriz elétrica brasileira, principalmente pelos benefícios que essa fonte possui na modalidade distribuída (WWF, 2015).

Vale ressaltar que o marco regulatório foi atualizado em janeiro de 2022 pela lei n°14300 que institui o marco legal da micro e minigeração distribuída, o Sistema de Compensação de Energia Elétrica e o Programa de Energia Renovável Social. Essa lei busca trazer maior segurança jurídica e transparência para o mercado de energia renovável, contribuindo ainda mais para o desenvolvimento acelerado da energia solar fotovoltaica no país (Brasil, 2022).

2.1.1 Geração solar fotovoltaica distribuída

Os sistemas fotovoltaicos são capazes de gerar energia elétrica convertendo radiação solar em corrente elétrica por meio das chamadas células fotovoltaicas. Essas células são distribuídas em painéis com cerca de $2m^2$ de área, os quais são modulares suficientemente para atender as necessidades de projetos residenciais, comerciais, como estádios de futebol, e de grandes usinas centralizadas.

O processo de conversão em energia elétrica é simples. A energia solar incide no painel em forma de radiação, sendo capaz de estimular a troca de elétrons entre as diferentes camadas de seu material interno. O silício extremamente puro e enriquecido no processo de dopagem - adição impurezas químicas elementares como Boro, Índio ou Fósforo - é responsável pela passagem ordenada de elétrons, gerando corrente elétrica para uso ao final do processo (Sampaio et.al, 2019).

Apesar de indiscutivelmente ser uma das fontes mais importantes para o futuro do setor elétrico mundial, as características desses novos sistemas fotovoltaicos, sejam centralizados ou distribuídos, trazem a reboque diversos desafios para a concepção e operação de sistemas elétricos (IRENA, 2019). Além dos desafios tecnológicos e na competitividade de preço, a variabilidade intrínseca do recurso natural é um fator bastante limitante ao seu desenvolvimento, assim como entender e acomodar as dificuldades momentâneas da descentralização da geração de energia elétrica.

Além da natural indisponibilidade sazonal e diária de radiação, os painéis solares são extremamente sensíveis à variação da radiação solar incidente, sendo impactados negativamente com a passagem de nuvens, a presença de animais, sujeira, dentre outros tipos de cobertura direta ou indireta. A variação da energia gerada chega a ser abrupta, reduzindo a quase zero o processo de conversão em módulos do sistema com qualquer tipo de cobertura (Lopes, 2018).

Dessa forma, do ponto de vista do controle centralizado pelo ONS, a variabilidade constante de geração de eletricidade e a descentralização, que aumenta consideravelmente o número de agentes do sistema, trazem necessidades de adaptação dos modelos de gerenciamento que passarão certamente por um ambiente de alta disponibilidade de dados.

Em termos de velocidade, dados são emitidos com frequência pelas unidades de geração de energia fotovoltaica com as variações constantes de parâmetros como radiação incidente, corrente, tensão e energia elétrica produzida. Em volume, os dados são armazenados para que as decisões sobre performance e qualidade dos sistemas sejam orientadas por medições reais. Cresce, também, a variabilidade dos dados pelo número de sensores e medidas atreladas à geração, a fim de monitorar em tempo real os sistemas e orientar a manutenção, operação, consumo e venda de energia elétrica.

Para apoiar a expansão da fonte solar, algoritmos de ML para previsibilidade de geração estão sendo desenvolvidos e aprimorados para trazer mais segurança de

operação pela redução de riscos de falta de eletricidade, mais economia por redução de despacho de usinas mais caras e maior previsibilidade no atendimento da demanda. Esses algoritmos levam em consideração diversos parâmetros de entrada como, principalmente, diversas variáveis meteorológicos (Li et al., 2016; Shao et al., 2016; Aler et al., 2015; Southern et al., 2015).

2.1.2 Dados meteorológicos subsidiando a decisão em fontes variáveis

Tem-se como comum a relação próxima entre o setor elétrico e a meteorologia, visto que a conversão de energia em eletricidade depende majoritariamente de recursos naturais. Em usinas hidroelétricas, por exemplo, os níveis dos reservatórios e, consequente, o volume de água disponível para a geração são afetados por fatores como níveis de chuva, cobertura de nuvens e temperatura (Mouriño et al., 2016). Já para usinas eólicas, as medidas de velocidade e direção dos ventos são fundamentais para entender os padrões sazonais de geração de eletricidade (WWF, 2015).

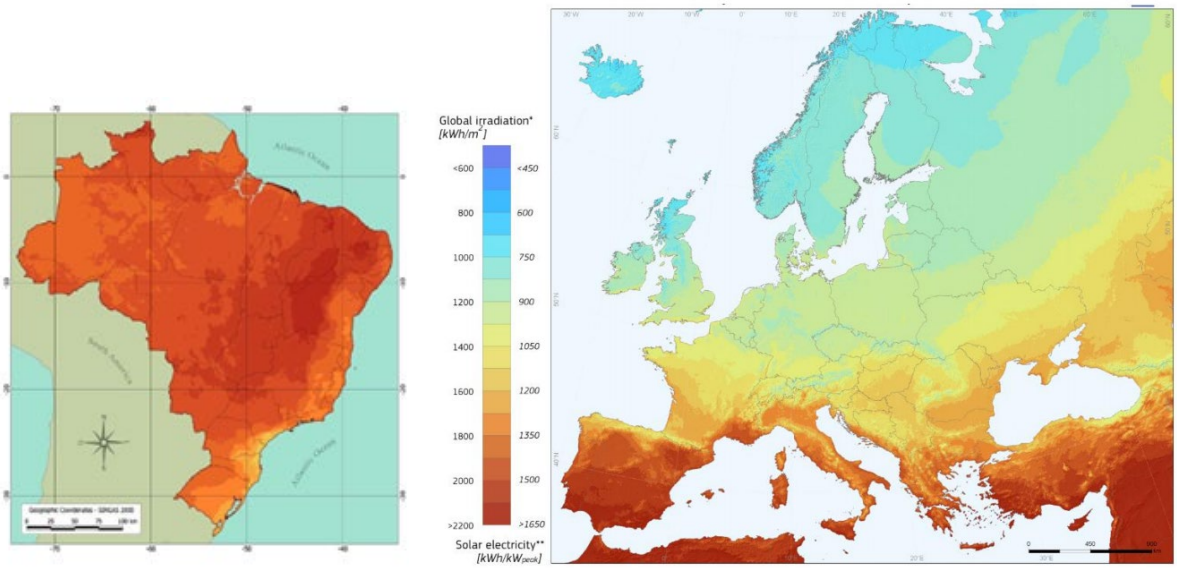
Para usinas solares fotovoltaicas, esse contexto não é diferente. O nível de radiação solar incidente (irradiação/energia) no painel é a métrica chave para compreender a quantidade de energia a ser gerada por painéis solares. Em média, os painéis atuais possuem eficiência de conversão da ordem de 20%, ou seja, para cada 1000W de radiação incidente por metro quadrado, 200W são convertidos em energia elétrica em corrente contínua. Entretanto, existem projetos já fazendo implementação de painéis de terceira geração com eficiência na ordem de 26% dado os materiais utilizados para suas construções (ESMC, 2021).

O Brasil é um país privilegiado em termos de irradiação solar global[4], oferecendo uma boa uniformidade ao longo do território e níveis médios muito maiores que países

[4] A radiação solar global é a soma das radiações solares incidentes diretamente do sol (radiação direta) e a radiação que chega à superfície após reflexões em outros corpos como nuvens (radiação difusa)

que estão se destacando na penetração de energia solar fotovoltaica, como Alemanha e Itália (INPE, 2006), como ilustrado pela Figura 1.

Figura 1 – Comparação da irradiação média global entre Brasil e Europa.



Fonte: (Atlas Brasileiro de Energia Solar, 2006)

Espera-se que com o crescimento da demanda por energia elétrica fruto do desenvolvimento nacional, assim como o crescimento acelerado da fonte solar fotovoltaica e os incentivos que vem recebendo nos últimos anos (IRENA, 2022), cresça também a necessidade de se compreender a disponibilidade da fonte para a geração de eletricidade localmente.

Essa disponibilidade está intimamente ligada às medidas de radiação solar incidente, fonte de energia com extrema influência nos processos atmosféricos, que por sua vez têm relação direta com parâmetros meteorológicos como cobertura de nuvens, velocidade e direção dos ventos, pluviosidade e umidade relativa (De Souza et al., 2008).

Nesse contexto, os algoritmos de ML buscam prever a radiação solar incidente e trazer inteligência para a gestão de recursos energéticos utilizando como dados de entrada dos modelos séries históricas de outros parâmetros meteorológicos mensurados localmente. Evidencia-se em estudos a correlação existente entre

radiação e outras variáveis meteorológicas para tomada de decisão no setor elétrico (Francisco et al., 2019).

Com a intensificação do uso e incorporação de modelos inteligentes no planejamento da operação do sistema elétrico, os custos sistêmicos são reduzidos, as falhas são previstas e consertadas de maneira mais rápida e o planejamento de longo prazo do setor fica mais assertivo (Collaborative., 2013). Os modelos que vêm sendo mais utilizados para este fim serão detalhados nas seções subsequentes.

2.2 Algoritmos de ML para previsibilidade da radiação solar

A literatura sobre o emprego de algoritmos de ML matrizes elétricas é ampla, sendo que as principais aplicações buscam minimizar os impactos da variabilidade deste recurso natural. Diversas pesquisas endereçam desde rastreadores solares, que buscam direcionar os painéis de maneira inteligente para o máximo de radiação solar incidente no local, até soluções de armazenamento (Revankar et al., 2010) (Chia et al., 2015) (Simmham et al., 2013).

Como mencionado anteriormente, aumentar a previsibilidade da geração de eletricidade é uma opção vantajosa, visto que traz maior segurança para o planejamento e redução dos custos totais de geração para atendimento de demanda. Algoritmos de ML tem ajudado a tornar as previsões mais precisas e com maior antecedência para o planejamento (Aybar-Ruiz et al., 2016; Burianek et al., 2016; Shamshirband et al., 2015)

Um modelo de predição é construído a partir dos dados e do algoritmo de ML empregado. De maneira geral, os modelos de predição fornecem as previsões de radiação solar incidente por meio de séries históricas de parâmetros meteorológicos. Os algoritmos mais empregados na literatura por apresentarem melhores resultados são: *Support Vector Machines* (SVM), *Artificial Neural Networks* (ANN) e *Extreme Learning Machines* (ELM) (Lou et al., 2016; Zhaoxuan et al., 2016; Gala et al., 2016).

2.2.1 Support Vector Machines (SVM)

SVM é um algoritmo inicialmente proposto em 1992 e que tem se sido muito empregado (Boser et al., 1992). É um algoritmo que apresenta versatilidade de aplicação, podendo auxiliar diferentes domínios do conhecimento, como análise de imagens, classificação de textos e bioinformática (Tuia et al., 2009) (Yang, Z, 2004) (Sun et al., 2009).

SVM é um algoritmo computacional supervisionado[5] capaz de aprender com exemplos, classificando a partir de seu aprendizado novas amostras fornecidas ao modelo. Simplificadamente, esse algoritmo maximiza uma função matemática para uma determinada amostra de dados. Desta forma, o algoritmo busca classificar conjuntos de dados mapeando-os em um espaço de características multidimensionais por meio do uso de uma função kernel (Lorena et al., 2007).

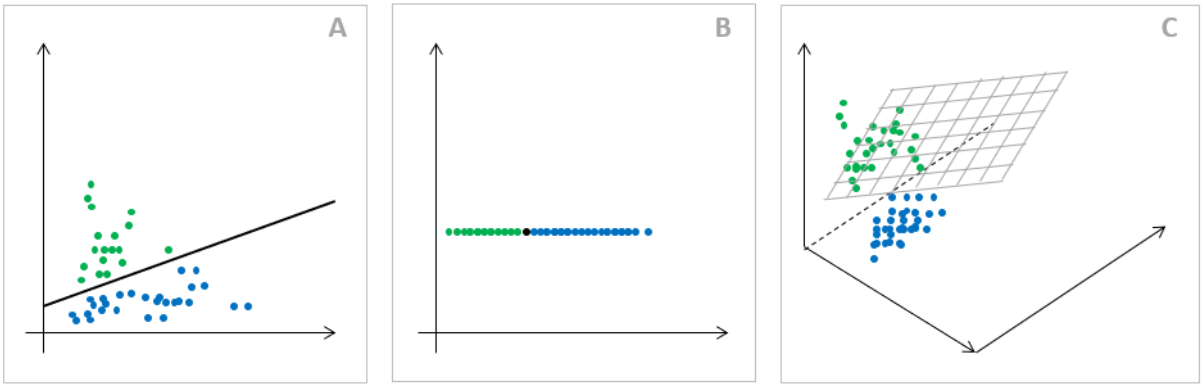
Para compreender o SVM, é necessário ter clareza sobre quatro principais conceitos: (i) o hiperplano de separação, (ii) o hiperplano de máxima margem, (iii) a margem macia e (iv) a função kernel (Steinwart, et al., 2008).

Considerando o problema de classificar dois grupos distintos de dados conforme apresentado pela Figura 2, caso esse problema aconteça em uma única dimensão (B), um ponto (o ponto preto na figura) é capaz de separar os conjuntos de dados em duas classificações distintas – azul e verde. No caso de duas dimensões (A), uma linha separa os dados classificados em dois grupos. Por fim, um plano faz a separação das classes em um espaço tridimensional (C). O termo geral para a superfície divisória em um espaço de elevadas dimensões (acima de três dimensões) é hiperplano e, na essência, o hiperplano é a generalização de uma linha reta que separa os dados em

[5] O aprendizado supervisionado computacional se dá quando é apresentado ao algoritmo as entradas e consequentes saídas desejadas, com o objetivo de que se aprenda uma regra geral de relação entre entradas e saídas.

dois grupos distintos. A esse hiperplano dá-se o nome de hiperplano de separação (Noble, W, 2006).

Figura 2 – Exemplo de hiperplano de separação.



Fonte: (adaptada Noble, W, 2006).

O conceito de separação em espaços multidimensionais não é exclusivo de SVM. Entretanto, o SVM é diferente de outros classificadores que utilizam hiperplanos de separação pela maneira com que esse hiperplano é selecionado, visto que múltiplos planos poderiam separar esses dados nos mesmos dois grupos distintos.

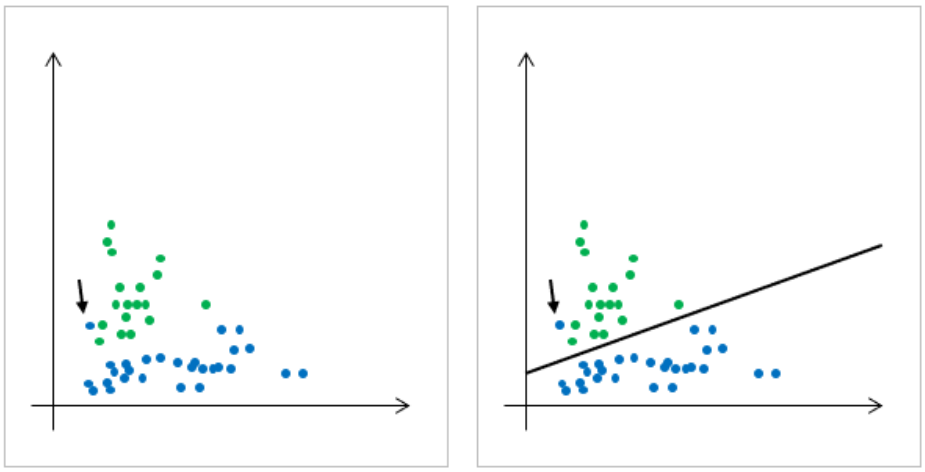
Fundamentado pela Teoria do Aprendizado Estatístico (TAE), os algoritmos de SVM selecionam os hiperplanos que maximizam a habilidade/probabilidade do algoritmo prever uma classificação correta de exemplos ainda não experimentados. Essa maximização se dá por meio da seleção do **hiperplano de máxima margem (**Noble, W, 2006).

Define-se como margem de um hiperplano a distância que separa esse hiperplano do menor vetor de expressão (vetor entre o hiperplano e um ponto de dados). O SVM escolhe o hiperplano com a maior margem possível, justificando a melhor performance do algoritmo.

Entretanto, em exemplos reais de dados, torna-se difícil a separação completa dos dados em dois grupos, sendo considerado um certo erro por parte do classificador. Este erro, denominado margem macia**,** permite que observações de um grupo de dados ultrapassem a margem dos hiperplanos de separação sem afetar o resultado.

O comportamento das margens macias pode ser observado na Figura 3, nesta à esquerda, tem-se o dado em uma região que possivelmente é de outra classificação, e à direita, mesmo com hiperplano definindo limites entre os grupos de dados, a margem macia permite o avanço da observação. Neste caso, a margem macia torna-se um parâmetro definido por quem está fazendo a modelagem, controlando basicamente o número de observações equivocadas permitidas do outro lado do hiperplano de separação (Noble, W, 2006).

Figura 3 – Exemplificação de ocorrência da margem macia.



Fonte: (adaptada Noble, W, 2006).

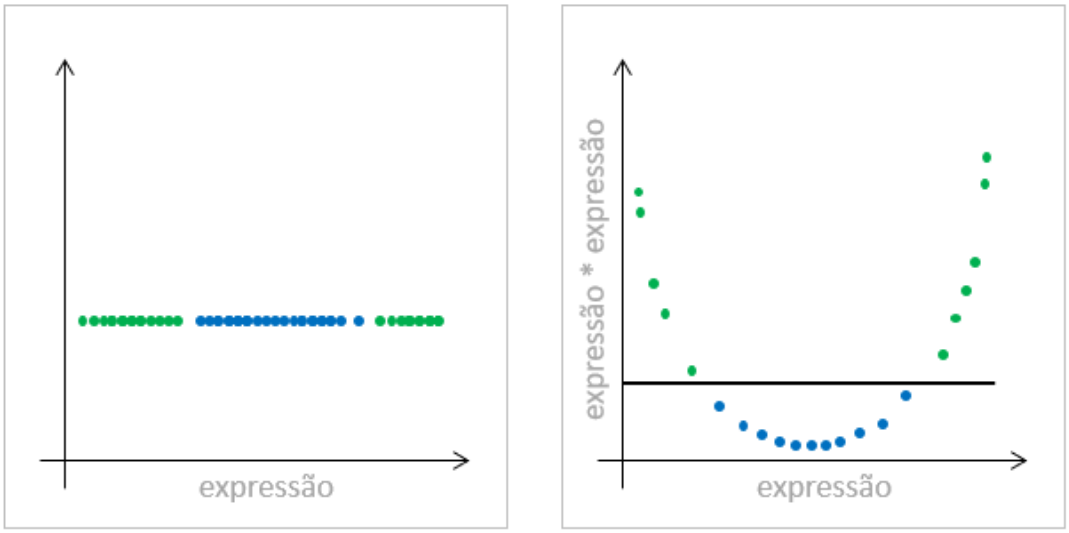
É difícil imaginar que nas dimensões que os dados coexistem naturalmente, seja possível encontrar funções lineares que os separem de maneira satisfatória.

Torna-se fundamental o papel da função kernel que busca adicionar dimensões para o conjunto de dados, tornando possível a separação dos dados em conjuntos por meio de planos. Na essência, a função kernel é um ajuste matemático que permite uma classificação bidimensional em um conjunto de dados que inicialmente é unidimensional. Sendo assim, de maneira geral, a função kernel projeta os dados de um espaço de menor dimensão para um de maior dimensão, no qual os dados são separáveis linearmente (Noble, W, 2006).

Para exemplificar, tem-se a aplicação apresentada pela Figura 4. À esquerda, temos um conjunto de dados unidimensional. Para classificá-los em dois grupos,

verde e vermelho, seria necessária uma função não linear. Entretanto, aplicando uma função kernel, em que todos os dados são elevados ao quadrado, torna-se possível a separação dos dois conjuntos de dados por um hiperplano (direita).

Figura 4 – Exemplo de aplicação da função kernel na definição do hiperplano de separação



Fonte: (adaptada Noble, W, 2006).

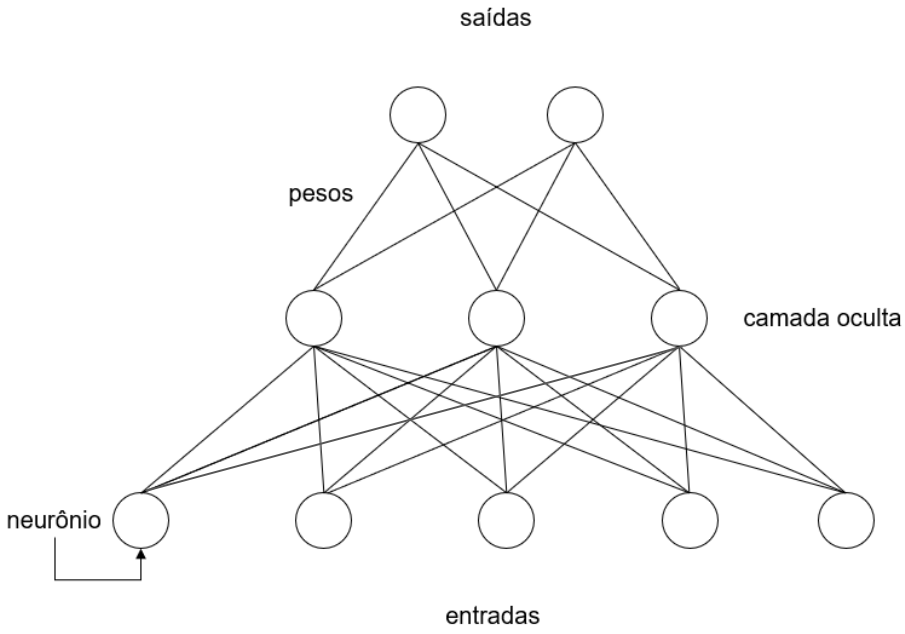
2.2.2 Artificial Neural Networks (ANN)

ANN são altamente inspiradas pelo funcionamento sofisticado dos cérebros humanos, nos quais informações são processadas paralelamente por bilhões de neurônios interconectados. A partir dessa inspiração, redes neurais vêm sendo aplicadas em diversos contextos, como problemas de identificação de imagens (Namba & Zhang, 2006), previsibilidade de serviços financeiros (Odom & Sharda, 1990) e até mesmo para reconhecimento de padrões em DNA (Cherry & Qian, 2018).

Essa versatilidade de aplicações faz das redes neurais uma das técnicas mais discutidas na atualidade, podendo ser construídas para resolver problemas de classificação, clusterização ou predição (regressão) (Wang, 2003).

Uma ANN é composta por uma camada de neurônios de entrada, uma ou mais camadas de neurônios intermediárias – chamadas de camadas ocultas - e uma camada de neurônios de saída do modelo. Esses neurônios são frequentemente chamados de nós ou unidades do modelo (Wang, 2003). A Figura 5 ilustra a integração entre camadas em um modelo de ANN.

Figura 5 – Arquitetura Geral de uma ANN.



Fonte: (Wang, 2003)

Como é possível observar na Figura 5, os neurônios são interconectados, sendo as conexões representadas por linhas que unem as diferentes camadas do modelo. Cada conexão é associada a um número, chamado de “peso” para o modelo.

As entradas para o modelo, assim como suas saídas, podem ser dados binários (sim ou não), elementos numéricos ou até mesmos símbolos (verde, vermelho, ...), o que confere às redes neurais uma alta gama de aplicabilidade.

A saída do modelo (h_i) em cada neurônio (i) na camada oculta pode ser representada pela equação (1):

$$h_i = \sigma\left(\sum_{j=1}^{N} V_{ij} x_j + T_i^{hid}\right), \quad (1)$$

Onde:

- σ é a função ativação que, além de adicionar componentes de não linearidade para a rede neural, acompanha o valor assumido pelo neurônio para que a rede neural não seja paralisada por neurônios divergentes;
- N é o número de neurônios de entrada;
- V_{ij} correspondem aos pesos do modelo;
- x_j são as entradas para os neurônios de entrada;
- T_i^{hid} são as definições de linhas de corte para os neurônios ocultos.

Desenhada a estrutura, um dos elementos mais essenciais para a implementação de uma rede neural é seu treinamento, o qual é projetado de maneira similar ao aprendizado humano. Para realizar um treinamento, separa-se uma amostra de dados de entrada no modelo os quais já se sabe os dados esperados de saída. Sendo assim, o objetivo da etapa de treinamento é ir ajustando os pesos estabelecidos nas conexões entre os neurônios para que uma função erro seja minimizada. Essa função erro, geralmente, é a soma dos quadrados das diferenças entre as saídas obtidas e as saídas já conhecidas no início do treinamento (Ripley, 1996).

O tamanho da amostra de dados para treinamento também deve ser cuidadosamente levado em consideração. A amostra deve ser grande o suficiente para que o modelo memorize elementos e tendências incluídas nessa base de dados. Por outro lado, se muitos elementos desnecessários são incluídos nessa amostra, a rede neural pode gastar recursos para se ajustar aos ruídos dessa amostragem não interessante de dados. A amostragem correta e apurada dos dados para treino é, portanto, fator crucial na definição do sucesso de um modelo de rede neural (Ripley, 1996).

É comum no início de um projeto envolvendo ANN, avaliar diferentes arquiteturas para selecionar a que melhor se aplica ao problema considerado. Neste caso, após a etapa de treinamento, uma amostra de dados de validação é separada e introduzida nas diferentes arquiteturas de redes neurais desenhadas, avaliando-se qual é mais eficiente naquele caso (Ripley, 1996).

2.2.3 Extreme Learning Machines (ELM)

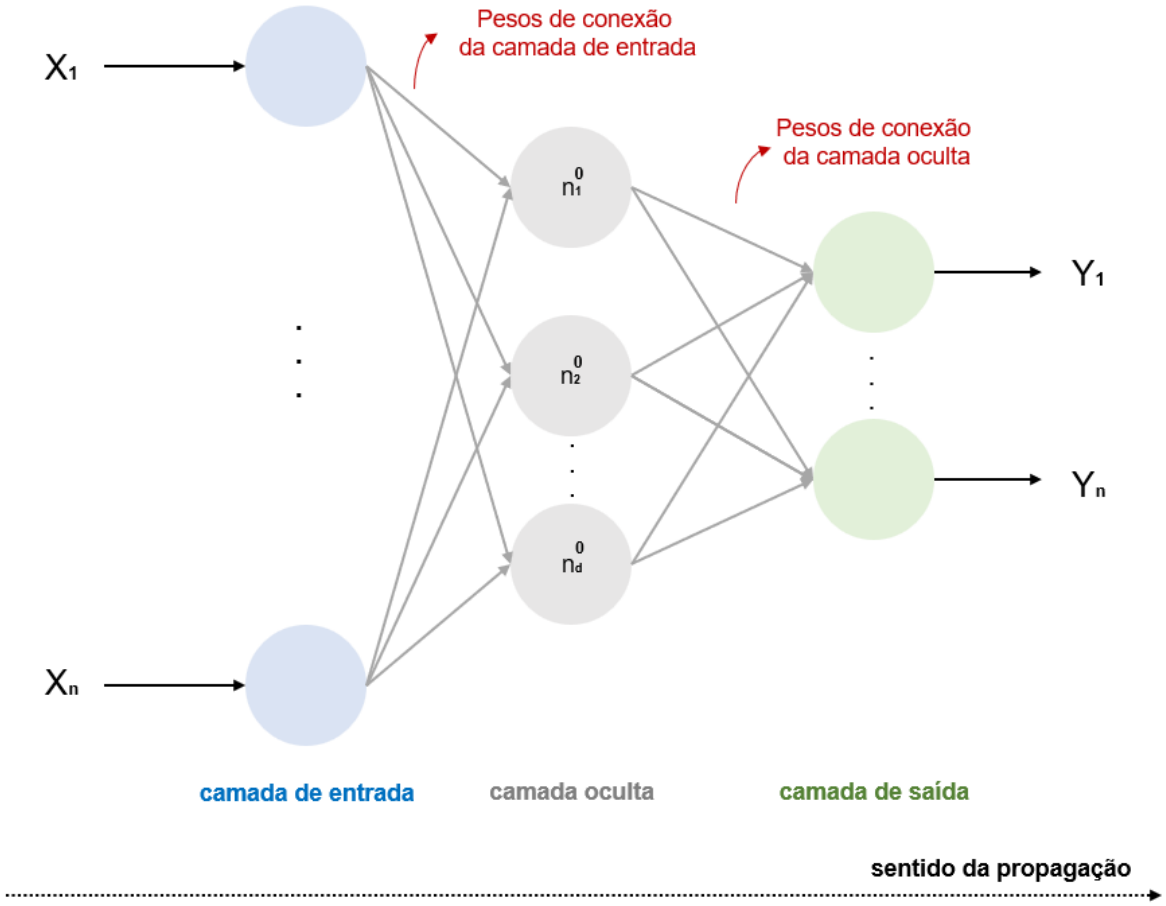
Como discutido anteriormente, é importante o papel SVM e ANN não só para o problema de previsibilidade de radiação solar, mas dentre todas as técnicas de inteligência computacional utilizadas nos últimos anos para diversas aplicações (Yang, Z, 2004) (Sun et al., 2009) (Namba & Zhang, 2006) (Cherry & Qian, 2018) (Odom & Sharda, 1990). Entretanto, essas técnicas também enfrentam desafios como a baixa velocidade de aprendizado, possibilidade/necessidade de intervenção humana no processo, elevado custo computacional de processamento e a necessidade de grandes volumes de amostras para treino (Suka et al., 2007) (Huang et al., 2011).

ELM é um algoritmo de aprendizado para a redes neurais, o qual possui uma única camada oculta de nós, alimentação na primeira camada de nós, escolha aleatória dos parâmetros dos nós ocultos e cálculo computacional dos pesos das saídas. Pode-se dessa forma chamar esse tipo de rede neural de "*single hidden layer feedforward neural network*" (SLFN). Este algoritmo tem sido amplamente estudado nos últimos anos devido sua capacidade de aprendizado rápido, boa generalização e elevada capacidade de aproximação/classificação (Tang, Deng, & Huang, 2016).

Ao contrário do que foi apresentado anteriormente nesta dissertação para os modelos de SVM e ANN, os parâmetros para a camada oculta de nós em arquiteturas de ELM são definidos de maneira aleatória e não precisam ser ajustados ao longo de etapas de treinamento. Ou seja, a camada de nós ocultos da estrutura pode ser definida previamente ao treinamento ou aquisição das amostras para treinamento (Tang, Deng, & Huang, 2016).

De maneira a ilustrar a arquitetura de uma ELM, é possível utilizar a mesma imagem que detalha uma ANN (Figura 5), porém define-se apenas uma camada oculta. A Figura 6 apresenta a arquitetura de uma ELM considerando o caso de k = 1

Figura 6 – Arquitetura de uma ELM



Fonte: (adaptado de Pacheco, A. 2017).

Diversos artigos já demonstraram, na teoria, como ELM tende a apresentar melhores e mais rápidas performances de generalização quando comparados a ANN e SVM (Huang et al., 2012) (Huang, 2014). Huang et al. mostrou que SLFN com uma camada oculta de neurônios aleatoriamente gerada e pesos de saída devidamente ajustados, mantém a capacidade universal de generalização das redes neurais, mesmo se não atualizados os parâmetros das camadas ocultas, além de ser muito mais rápida a definição dos pesos para esses algoritmos (Huang et al., 2006).

2.3 Qualidade de dados

Apesar da consolidação da importância do uso de dados para fundamentar a tomada de decisão, muito impulsionado pelos avanços já citados da indústria de tecnologia da informação desde o início do século XXI, ainda pouca atenção tem sido voltada para qualidade dos dados (QD) e os seus impactos na capacidade de gerar valor independente da área de aplicação (Saha, B., Srivastava, D.,2014).

Num contexto de alto volume, velocidade e variedade de dados, os desafios do mundo operando em *big data* perpassam discussões de qualidade de dados em todas essas características. O **volume** de dados é tremendo, o que torna muito difícil um julgamento preciso de qualidade dos dados num curto período. Os dados mudam com extrema **velocidade** e o período de utilização pode ser bem curto, necessitando elevadíssima capacidade de processamento. Por fim, a **variedade** traz para a cadeia de dados diferentes tipos de dados que podem aumentar a complexidade de integração, tratamento, armazenamento e processamento desses dados (Cai, L., Zhu, Y., 2015).

O conceito de *big data* mostra uma face em que o baixo controle de qualidade dos dados utilizados afeta diretamente o nível confiança na qualidade dos dados que estão sendo utilizados para subsidiar decisões. Em estudo feito em 2004 pela *PricewaterhouseCoopers* com 452 empresas, somente 34% dos respondentes se mostraram muito confiantes com a qualidade dos dados (QD) que estão utilizando (PwC, 2004).

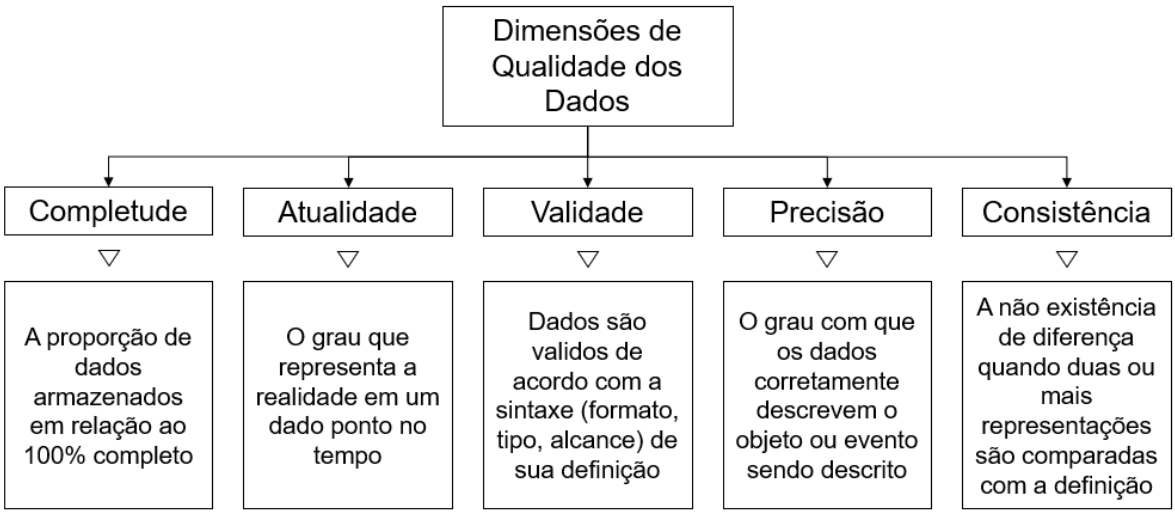
Estima-se que o impacto anual da utilização de dados de baixa qualidade nos negócios norte-americanos seja da ordem de 600 bilhões de dólares (Eckerson, 2002). Em projetos de *data warehousing* – sistemas que integram dados de múltiplas fontes para a tomada de decisão – estima-se que entre 30 e 80% do tempo de desenvolvimento seja gasto em limpeza e outros problemas de qualidade dos dados (Saha, B., Srivastava, D.,2014; Cai, L., Zhu, Y., 2015).

Em 1996, o grupo de pesquisa em *Total Data Quality Management* do Massachusetts Institute of Technology (MIT) apresentou um conceito de QD como “adequado ao uso pretendido”, propondo que quem os utiliza, o consumidor, deve ser o responsável por avaliar sua qualidade. O grupo liderado pelo Professor Richard Y. Wang, também define o conceito de dimensões de qualidade dos dados como um conjunto de atributos que representam diferentes aspectos dos dados (Wang, R. Y. et al., 1996).

Quando se busca melhorar a QD, o objetivo é mensurar e melhorar as dimensões de qualidade dos dados que caracterizam os problemas identificados. A literatura traz um amplo conjunto de dimensões de QD (Woodall et al., 2014) (Taleb et al., 2018) (Batini et al., 2015) (Rao et al., 2015), que foram ampliadas com o contexto de big data. Não há consenso nem quanto número, nem quanto a definição das dimensões (Francisco et al. 2017).

A *International Data Management Association* (DAMA), entretanto, define 6 dimensões para qualidade dos dados as quais são representadas e descritas pela Figura 7 (DAMA, 2013).

Figura 7 - Dimensões de Qualidade dos Dados



Fonte: (DAMA, 2013).

No contexto de *big data* alguns argumentam que, mesmo que não como uma regra, o volume de dados pode compensar a qualidade por meio da diluição dos erros em alguns cenários e aplicações (Ramasamy, A., Chowdhury, S., 2020). Porém, o que se observa é uma preocupação com a qualidade dos dados e com formas de mitigar os problemas identificados (Sadiq and Papotti, 2016) (DAMA International, 2017), pois se uma amostra de dados tem erros, aumentar a quantidade dos dados também pode implicar em aumentar a quantidade de erros.

Na REL elaborada por (Ramasamy, A., Chowdhury, S., 2020), que contempla uma análise detalhada de 17 artigos publicados desde 2013 e que abordam dimensões de dados em ambientes de *big data*, 10 dimensões chave de qualidade de dados são propostas. Essas dimensões são apresentadas pela Tabela 1.

Tabela 1 – Dimensões de qualidade dos dados em ambientes de *Big Data.*

(continua)

Dimensão de qualidade dos dados	**Definição**
Acessibilidade	Acessibilidade e disponibilidade estão relacionadas à capacidade do usuário de acessar dados a partir de sua cultura, estado físico/capacidades e tecnologias disponíveis
Coesão	Consistência, coesão e coerência referem-se à capacidade dos dados de cumprir sem contradições com todas as propriedades da realidade de interesse, conforme especificado em termos de restrições de integridade, edições de dados, regras de negócio e outros formalismos
Confidencialidade	Dimensão de qualidade que determina se os dados certos estão nas mãos certas. Os dados estão seguros?
Credibilidade	Os dados devem vir de organizações especializadas de um país, área ou indústria. Especialistas auditam regularmente e verificam a exatidão do conteúdo dos dados. Os dados existem na faixa de valores conhecidos ou aceitáveis.
Linhagem/Ancestralidade	Esta dimensão auxilia no conhecimento da fonte dos dados para que qualquer inconsistência seja corrigida na fonte e não em outras instâncias.

Tabela 1 – Dimensões de qualidade dos dados em ambientes de Big Data

(continuação)

Legibilidade	Também representada como clareza, simplicidade, facilidade de compreensão, interpretabilidade, compreensibilidade, esta dimensão refere-se à facilidade de compreensão dos dados pelos usuários.
Redundância	Redundância, compactação e concisão referem-se à capacidade de representar a realidade de interesse com o uso mínimo de recursos informativos

Fonte: (adaptado de Ramasamy, A., Chowdhury, S., 2020).

3 RESULTADOS DA REVISÃO DE ESCOPO DA LITERATURA

Uma REL foi inicialmente realizada para identificar qual o estado da arte em relação ao uso de técnicas de big data para previsão de geração de eletricidade fotovoltaica.

A REL resultou na publicação do artigo (de Freitas Viscondi & Alves-Souza, 2019). Os principais aspectos dessa publicação são descritos a seguir.

3.1 Metodologia empregada na Revisão de Escopo da Literatura

A metodologia adotada para REL foi proposta por Kitchenham e Charters (Kitchenham & Charters, 2007). As questões de pesquisa foram estabelecidas de forma a identificar os trabalhos científicos que investigam a relação entre previsibilidade de radiação solar para geração fotovoltaica, modelos fundamentados em inteligência artificial (IA) e grande quantidade de dados. Foram definidas 4 questões de pesquisa:

- **Questão 1** – Onde, por que e por quem esses estudos estão sendo feitos?
- **Questão 2** – Como os modelos fundamentados por grandes quantidades de dados estão ajudando a resolver o problema de previsibilidade de geração solar fotovoltaica?
- **Questão 3** – Quais e que tipos de dados estão sendo utilizados?
- **Questão 4** – Como o desenvolvimento de conhecimento em previsibilidade de geração está conectado com a penetração da fonte solar renovável no mundo?

Com as questões de pesquisa definidas, seguiu-se o protocolo para revisões de escopo da literatura, definindo uma cadeia lógica e estruturada de passos para sistematização e análise dos artigos relacionados com as questões propostas. A REL apresenta de maneira detalhada todos os nove passos considerados de acordo com o protocolo.

As bases de conhecimento consultadas foram: *Web of Science*, *Science Direct*, IEEE e *Google Scholar*. Optou-se por essas bases pois grande parte das publicações acadêmicas estão indexadas neste conjunto de bases e os autores possuem acesso completo pela Universidade de São Paulo.

Utilizou-se as seguintes frases de busca para consultar artigos relevantes à realização da pesquisa: "*Big Data*" *and* "*Solar*"; "*Data Mining*" *and* "*Solar*"; "*Machine Learning*" *and* "*Solar*"; e "*Big Data*" *and* "*Power Forecasting*", sendo adaptadas em cada base de conhecimento, segundo seu mecanismo de busca. Ao todo 95 artigos foram encontrados nesta etapa. Neste grupo de artigos, aplicou-se os critérios de inclusão e exclusão para que a análise seguisse somente com trabalhos estritamente relacionados às questões de pesquisa previamente definidas. Foram eles:

Critérios de Inclusão:

- **Publicação entre 01/2013 e 05/2017:** visando incluir, na data de realização da publicação, somente os trabalhos mais recentes;
- **Artigo apresenta modelo de previsibilidade usando modelos com grandes quantidades de dados:** para evitar artigos que incluem modelos de previsibilidade que não utilizam ML ou que utilizam ML para resolver outros problemas relacionados à geração solar fotovoltaica;
- **Artigo compara diferentes técnicas e modelos para previsão de radiação solar fotovoltaica:** incluir artigos que comparem diferentes modelos de ML.

Critérios de Exclusão:

- **Artigos que não foram escritos em inglês:** incluir somente artigos no idioma considerado internacional;
- **Artigos que são revisões secundárias ou estudos terciários** incluir somente estudos primários;
- **Artigos que utilizam os modelos para resolver outros problemas relacionados à fonte solar fotovoltaica:** excluir artigos que usam

modelos de ML para prever outros problemas relacionados à geração solar fotovoltaica.

Após remoção de duplicatas e aplicação dos critérios supracitados, 38 artigos foram contemplados pela REL. Destes, 10 foram avaliados como mais pertinentes às questões de pesquisas, sendo considerados os artigos selecionados para análise.

3.2 Principais resultados

SVM, ANN, ELM, *gradient boosting* (GB) e *random forest* (RF) são os algoritmos mais citados pelos 10 artigos selecionados. A Tabela 2 apresenta o número de vezes que cada algoritmo foi utilizado para a resolução do problema de previsibilidade de radiação solar.

Tabela 2 - Número de vezes que cada algoritmo de ML ou outro algoritmo foi utilizado pelos 38 artigos revisados.

Algoritmo/Técnica	Número de Utilizações
Support Vector Machine - SVM	17
Artificial Neural Network – ANN	14
Extreme Learning Machine - ELM	7
Gradient Boosting – GB	3
Random Forest – RF	5
Genetic Algorithm	2
Decision Trees	2
Outras técnicas	7
Algoritmo não especificado	4

Fonte: de Freitas Viscondi, G., & Alves-Souza, S. N. (2019)

É significativo o número de artigos que propõem, também, técnicas de preparação dos dados prévia à alimentação dos modelos – como clusterização dos dados - ou algoritmos híbridos que misturam técnicas de ML e modelagem numérica para previsão climática - *numerical weather prediction* (NWP). Alguns artigos também

comparam os resultados dos modelos de ML aos resultados de modelagens lineares (3 artigos) e modelos numéricos (6 artigos).

Entretanto, parece ser um consenso, até então, que ANN constituem a melhor estratégia para a resolução do problema de previsibilidade. Mesmo artigos que não utilizam algoritmos de redes neurais discutem a relevância dessa técnica para o problema. Dessa forma, diversas derivações de redes neurais são também frequentemente propostas, com ênfase para ELM, para as quais são relatadas melhores velocidades de aprendizado e resultados de previsão.

Da perspectiva da alimentação dos modelos com dados que façam sentido para a previsibildiade de geração de eletricidade, surge o questionamento de quais os tipos de dados mais utilizados e como é abordado o tratamento para qualidade pré-ingestão nos modelos.

Considerando os tipos de dados empregados nos trabalhos para a previsibilidade de geração de eletricidade, analisou-se os trabalhos que empregaram dados relacionados à eletricidade – como geração de eletricidade (kWh), corrente (A), tensão (V) ou potência (W). Identificou-se que menos de 20% dos 38 artigos usaram dados desse tipo.

A maioria dos trabalhos concentrou seus esforços em lidar diretamente com o recurso natural - radiação solar. Nos artigos consultados, mais de 73% (28) dos trabalhos desenvolveram suas pesquisas combinando dados de irradiância solar e outros parâmetros meteorológicos. Ainda, 26% dos artigos (10) usaram apenas dados de irradiância para produzir diretamente as previsões. A maioria dos artigos utilizou uma série histórica de dados de 5 a 15 anos – 23 artigos ou 60% dos trabalhos consultados. Porém, também foram encontrados 8 resultados com dados coletados por apenas um ano.

Observou-se que, conforme relatado pelos trabalhos que utilizaram irradiância solar e outras variáveis meteorológicas para treinar os algoritmos, o aumento no número de parâmetros utilizados na sessão de treinamento resultou em uma melhora da previsão.

Apenas 10 dos 38 artigos relataram questionamentos ou ações sobre a qualidade dos dados, inferindo-se que muito pouco tem sido produzido em relação a intersecção entre qualidade dos dados e previsibilidade de geração fotovoltaica distribuída.

4 CONTEXTUALIZAÇÃO E ANÁLISE DO CENÁRIO DE ESTUDO

Este capítulo busca contextualizar o cenário de estudo proposto por este trabalho, no qual o Brasil, assim como outros países do mundo, tem passado pelo desenvolvimento e ganho de maturidade da fonte solar fotovoltaica.

Traz-se o contexto atual da fonte solar no Brasil e como o trabalho busca abordar um dos principais entraves ao desenvolvimento da fonte: sua variabilidade de geração e a necessidade de integração ao SIN.

Para a aplicação em um contexto local e construção dos modelos, utiliza-se dados coletados por uma estação meteorológica na cidade de São Paulo. Este capítulo também descreve os dados utilizados, assim como o processo adotado para limpeza, preparo e aplicação na construção dos modelos propostos.

4.1 Cenário de Estudo

O setor elétrico mundial vem passando por um momento muito propício para inovação e incorporação de novas tecnologias. Motivado por questões ambientais, como emissões de gases de efeito estufa (GEE), emissões de poluentes locais, consumo de água, dentre outros, econômicas, como redução dos custos de geração, ou de política industrial e de desenvolvimento tecnológico, o setor elétrico busca por alternativas para as fontes fósseis de geração de eletricidade, com o intuito de suprir demandas crescentes por energia elétrica (IRENA, 2020).

Considerando as estatísticas recentes de expansão das matrizes elétricas mundiais, pode-se dizer que nas últimas duas décadas o mundo vem apostando nas fontes solar e eólica para garantir a renovabilidade da energia elétrica gerada no futuro. De acordo com a *International Renewable Energy Agency* (IRENA), entre os anos de 2011 e 2021, foram instalados 642 GW de energia eólica e 813 GW de energia solar ao redor do mundo. Ambas as fontes representam 79% da capacidade renovável adicional instalada mundialmente nos últimos 10 anos (IRENA, 2022).

O cenário para as fontes solar e eólica no Brasil não é diferente, mesmo que as fontes enfrentem momentos de desenvolvimento distintos no país. Os planos de expansão governamentais e os compromissos firmados internacionalmente sinalizam um futuro próspero para essas fontes não despacháveis de geração de eletricidade. Segundo o Plano Nacional de Energia 2050, é esperado que ao final do horizonte de 2050 o país esteja com a capacidade instalada para usinas eólicas *onshore*[6] entre 110 e 195 GW e algo entre 27 e 90 GW somente em usinas solares (EPE/MME, 2020). Já no Plano Decenal de Expansão de Energia 2031 (PDE2031), a Empresa de Pesquisa Energética (EPE) espera, nos próximos anos, um salto de 8GW para 37 GW instalados de energia solar fotovoltaica distribuída e a instalação de mais 10 GW de energia eólica *onshore* (EPE/MME, 2022).

O cenário torna-se ainda mais propício para investimento quando os custos nivelados de energia – *levelized cost of energy* (LCOE[7]) para as fontes solar e eólica são analisados ao longo do tempo. Tecnologias tidas como recentes e custosas no passado, têm apresentados custos decrescentes nos últimos anos (Jaradat et al., 2015).

Segundo a IRENA, espera-se que a energia solar fotovoltaica tenha, em 2025, seu LCOE reduzido em 57% quando comparado com os custos mundiais em 2015. Cenário semelhante ocorre para a energia eólica *onshore* e *offshore*[8] para as quais, por estarem em estado de maturidade tecnológica mais avançado no ano de 2015, espera-se uma redução menor nos custos (12 e 15%, respectivamente) (IRENA, 2016).

[6] Usinas eólicas de geração de eletricidade instaladas e fixadas em terra.
[7] O LCOE analisa os custos globais de um sistema (investimento, operação e manutenção) e o montante de energia (medido em quilowatts hora) que ele gera ao longo de sua vida útil. Os valores são trazidos para o valor presente e apresentados em unidades monetárias por unidade de energia (e.g. R$/kWh).
[8] Usinas eólicas de geração de eletricidade instaladas e fixadas em alto mar.

No Brasil, um dos países com os melhores ventos[9] para a geração eólica no mundo (Cresesb, 2001), a energia eólica foi contratada nos leilões de energia nova em 2022 pelo preço médio de 176,3 R$/MWh e chegou a ser contratada em agosto de 2018 por 79 R$/MWh, um dos menores valores históricos para a fonte (EPE/MME, 2018). Da fonte solar, no leilão A-5 de 2022, foram contratados 200 MW de potência total, em 4 projetos, a um preço médio de 171,4 R$/MWh (CCEE, 2022).

Visto que os painéis solares são extremamente modulares, a fonte encontra-se em constante e acelerada expansão no modelo distribuído, intensificando os problemas subjacentes do SIN e trazendo necessidades de melhoria visto o espalhamento geográfico da geração em pequenos sistemas e a variabilidade intrínseca ao recurso solar.

Em um contexto de expansão da fonte solar fotovoltaica no Brasil, exigem-se soluções que conectem de maneira inteligente os sistemas de geração fotovoltaicos ao SIN, reduzindo os impactos da variabilidade de geração a partir da radiação, e que sejam rápidas e integradas suficientemente para contemplar a interação de múltiplos agentes, fator multiplicado exponencialmente por meio da geração distribuída (Singh et al., 2015).

Viu-se na REL apresentada pelo Capítulo 3 deste trabalho, que uma das soluções utilizadas atualmente para a redução da variabilidade e integração de usinas descentralizadas de geração em sistemas elétricos está no aumento da capacidade de previsão do recurso solar por algoritmos de ML.

Este trabalho busca trazer os objetivos descritos em seu Capítulo 1 para um contexto nacional de aplicação, utilizando-se de dados meteorológicos locais para

9 Para produzir **energia eólica**, são necessários bons ventos: estáveis, sem mudanças bruscas de velocidade ou de direção e com a intensidade certa – minimamente 7 a 8 m/s a uma altura de 50m.

avaliar a previsibilidade de radiação solar e consequente geração de energia elétrica em um contexto de operação do nosso sistema elétrico de potência – o SIN.

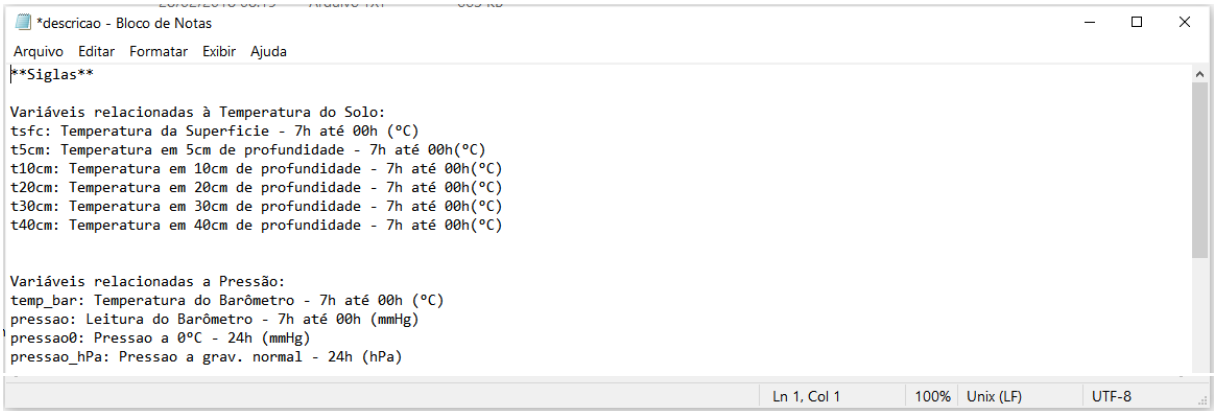
4.2 Descrição da base de dados utilizada

Para o desenvolvimento da pesquisa e alcance dos objetos elencados no Capítulo 1, utilizou-se dados de séries históricas de parâmetros meteorológicos medidos na cidade de São Paulo pelo IAG-USP.

A série histórica de dados meteorológicos utilizados foi escolhida devido ao fácil acesso dentro da Universidade de São Paulo e à capacidade de contribuição científica pela análise de dados num contexto brasileiro de aplicação de modelos de aprendizagem de máquina. Essa também é a mais antiga base de dados de parâmetros meteorológicos no Brasil, com medições datadas desde 1933.

Os dados foram recebidos em 20 arquivos de texto (valores separados por vírgulas), sendo 1 para descrição dos dados contidos nos arquivos e outros 19 contendo os registros para cada parâmetro meteorológico individualmente. A Figura 8 mostra um excerto do arquivo de descrição dos dados recebidos:

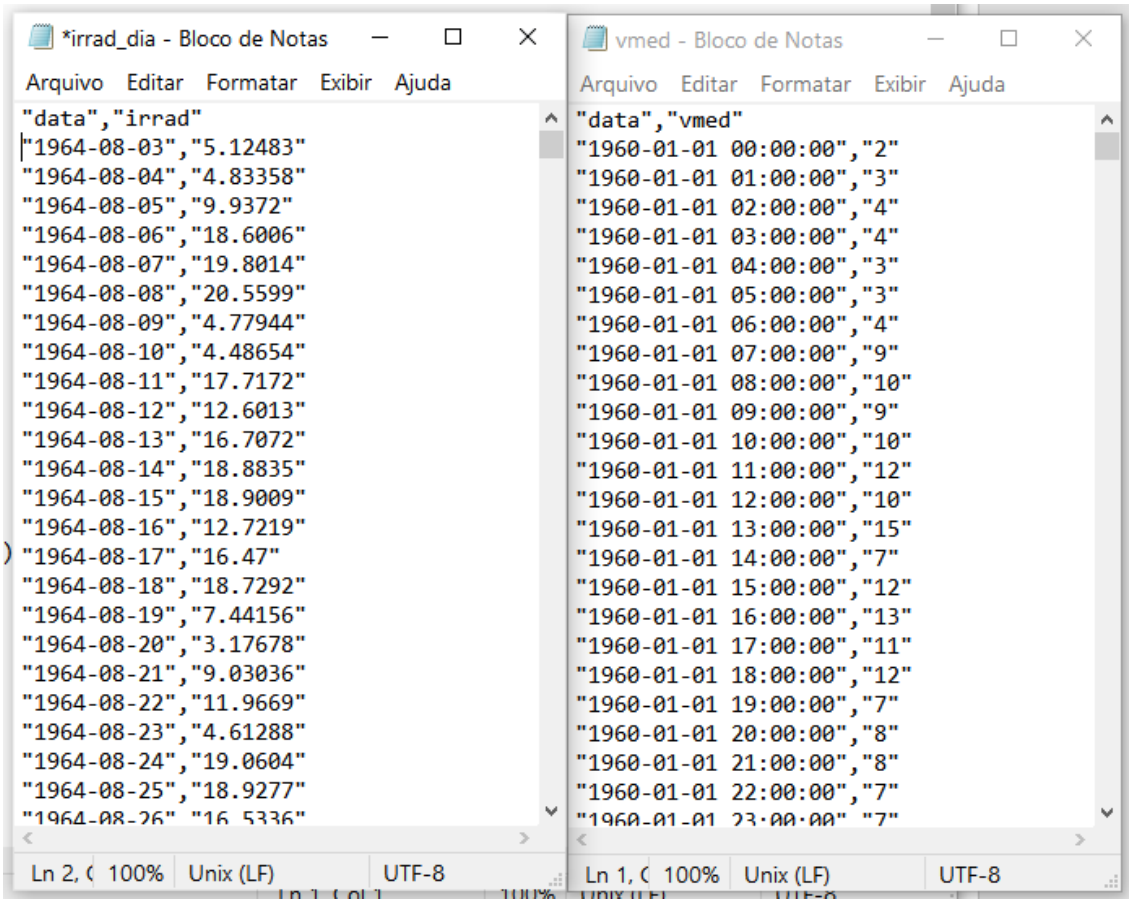
Figura 8 – Descrição dos dados recebidos fornecida pelo IAG-USP



Cada parâmetro meteorológico possui duas colunas de registros de dados, sendo a primeira para o registro da data, podendo ser diária ou horária a depender do

parâmetro, e a segunda com o registro da medição do parâmetro na unidade pertinente para cada medida. A Figura 9 apresenta a visão de dois arquivos recebidos e que contemplam dois diferentes parâmetros meteorológicos medidos acompanhados pelo IAG-USP:

Figura 9 - Arquivo contendo valores registrados para irradiação solar total diária (esquerda) Arquivo contendo valores registrados para velocidade média horária do vento (direita)



Os parâmetros recebidos, o número de registros de medidas para esse parâmetro, assim como a frequência de registro dos dados estão sintetizados na Tabela 3.

Tabela 3 - Parâmetros Meteorológicos Recebidos

Parâmetro meteorológico	Unidade	Número de Registros	Frequência de Registros
Temperatura de Superfície (tsfc)	°C	368.118	Horária das 07h às 24h
Pressão Atmosférica (press24)	bar	692.687	Horária: 24h
Temperatura do Bulbo Seco (tseco)	°C	558.727	Horária das 07h às 24h
Temperatura do ar (temperatura)	°C	785.231	Horária: 24h
Temperatura do Bulbo Úmido (túmido)	°C	545.590	Horária das 07h às 24h
Temperatura Máxima (tmax)	°C	30.288	Diária
Temperatura Mínima (tmin)	°C	30.288	Diária
Umidade Relativa (ur)	%	525.960	Horária: 24h
Precipitação Horária (prec)	mm	745.105	Horária: 24h
Precipitação Horária (duração)	min	745.105	Horária: 24h
Irradiação Solar Total Diária (irrad)	MJ/m^2	31.045	Diária
Fração Horária de Brilho Solar (insol)	fração	745.105	Horária: 24h
Direção Predominante dos Ventos (dirdom)	Direção	508.440	Horária: 24h
Direção Rajada Diária (dirrajd)	Direção	21.185	Diária
Maior Rajada de Vento Horária (rajh)	m/s	508.440	Horária: 24h
Maior Rajada de Vento Diária (rajd)	m/s	21.185	Diária
Velocidade Média do Vento (vmed)	km/h	508.440	Horária: 24h
Velocidade Média do Vento meridional (vmed)	km/h	522.055	Horária: 24h
Características de nuvens – Quantidade e Tipo (tipom e tipoa)	-	558.727	Horária das 07h às 24h

Apesar dos parâmetros meteorológicos terem sido recebidos conjuntamente e com frequência de registros semelhantes, as medidas apresentam intervalo de medição diferente para cada parâmetro. A disponibilidade anual de dados de cada parâmetro recebido pode ser vista na Tabela 4:

Tabela 4 - Série histórica de dados para cada parâmetro meteorológico

Parâmetro meteorológico	Intervalo de Tempo	1930	1940	1950	1960	1970	1980	1990	2000	2010	2020
Temperatura de Superfície (tsfc)	1962-2017										
Pressão Atmosférica (press24)	1934-2014										
Temperatura do ar (tseco)	1933-2017										
Temperatura do ar (temperatura)	1933-2017										
Temperatura do Bulbo Úmido (túmido)	1935-2017										
Temperatura Máxima (tmax)	1935-2017										
Temperatura Mínima (tmin)	1935-2017										
Umidade Relativa (ur)	1958-2017										
Precipitação Horária (prec)	1933-2017										
Precipitação Horária (duração)	1933-2017										
Irradiação Solar Total Diária (irrad)	1961- 2017										
Fração Horária de Brilho Solar (insol)	1933-2017										
Direção Predominante dos Ventos (dirdom)	1960-2017										
Direção Rajada Diária (dirrajd)	1960-2017										
Maior Rajada de Vento Horária (rajh)	1960-2017										
Maior Rajada de Vento diária (rajd)	1960-2017										
Velocidade Média do Vento (vmed)	1960-2017										
Velocidade Média do Vento meridional (vmed)	1960-2017										
Características de nuvens – Quantidade e Tipo	1958-2017										

Visto que os parâmetros meteorológicos possuem diferentes frequências de aquisição de registros e períodos de disponibilidade de dados, uma base de dados final e reduzida, cuja estrutura é apresentada na Tabela 5, foi modelada para a implementação dos algoritmos de ML. A base final possui 19.358 observações, contendo registros diários de dez parâmetros meteorológicos de 1962 a 2014:

Tabela 5 – Estrutura da base de dados modelada a partir dos registros da estação meteorológica da USP para a implementação dos modelos de ML

Nome da Coluna	Conteúdo	Unidade
data	Data de registro da medição	-
irradiation	Irradiação Solar Total Diária	MJ/m²
temp_max	Temperatura Máxima	°C
temp_min	Temperatura Mínima	°C
wind_daily	Maior Rajada de Vento Diária	m/s
humidity	Umidade Relativa	%
prec	Precipitação Diária	mm
pressure	Pressão Atmosférica	atm
clouds_qtb	Quantidade de Nuvens – baixa altitude	-
clouds_qtm	Quantidade de Nuvens – média altitude	-
clouds_qta	Quantidade de Nuvens – elevada altitude	-

Para complementar a base de dados, uma última coluna foi adiciona contendo a estação do ano para cada dia com registro, uma vez que a variável irradiação a ser prevista pelo algoritmo é altamente dependente da sazonalidade anual.

Para fins de compreensão da base de dados utilizada para produzir os resultados apresentados no Capítulo 5 e entendimento da região em estudo, a Tabela 6 apresenta medidas de estatística descrita para todos os parâmetros meteorológicos utilizados:

Tabela 6 – Base de dados modelada para a implementação dos modelos de ML

Parâmetro meteorológico	Unidade	Registro mínimo	1°quartil	Mediana	Média	3° quartil	Registro Máximo
irradiation	MJ/m²	0	11.93	15.99	16.19	20.49	35.56
temp_max	°C	8.60	22.00	25.50	25.08	28.40	37.20
temp_min	°C	-1.10	12.60	15.20	14.95	17.80	23.20
wind_daily	m/s	0	5	6	6.41	8	28
humidity	%	33.87	76.41	82.15	81.06	87.12	99.25
prec	mm	0	0	0.1	0.401	2.4	146
pressure	atm	893.6	923.4	925.7	925.8	928.2	939.6
clouds_qtb	-	0	2.22	4.61	4.82	7.39	10.00
clouds_qtm	-	0	0	0.44	1.32	1.94	9.89
clouds_qta	-	0	0	0.11	0.076	1.00	9.94

Observando as medidas para a variável de saída do modelo – irradiação – temos o valor médio de 16,19 MJ/m2 e registros máximos diários em 36,56 MJ/m^2 para todo o período de análise. Os registros mínimos incluem medidas noturnas e por isso valores iguais a zero aparecem na base de dados.

Segundo o Atlas Brasileiro de Energia Solar, a irradiação média no Brasil varia de 15,99 MJ/m^2 a 19,74 MJ/m^2 e, apesar de apresentar valores altos de irradiação para a produção de energia solar, a região em estudo está próxima do limite inferior para medidas nacionais (Atlas Brasileiro de Energia Solar, 2006). O comportamento sazonal de irradiação solar medida no local de estudo entre 1962 e 2014 é apresentado pela Figura 10:

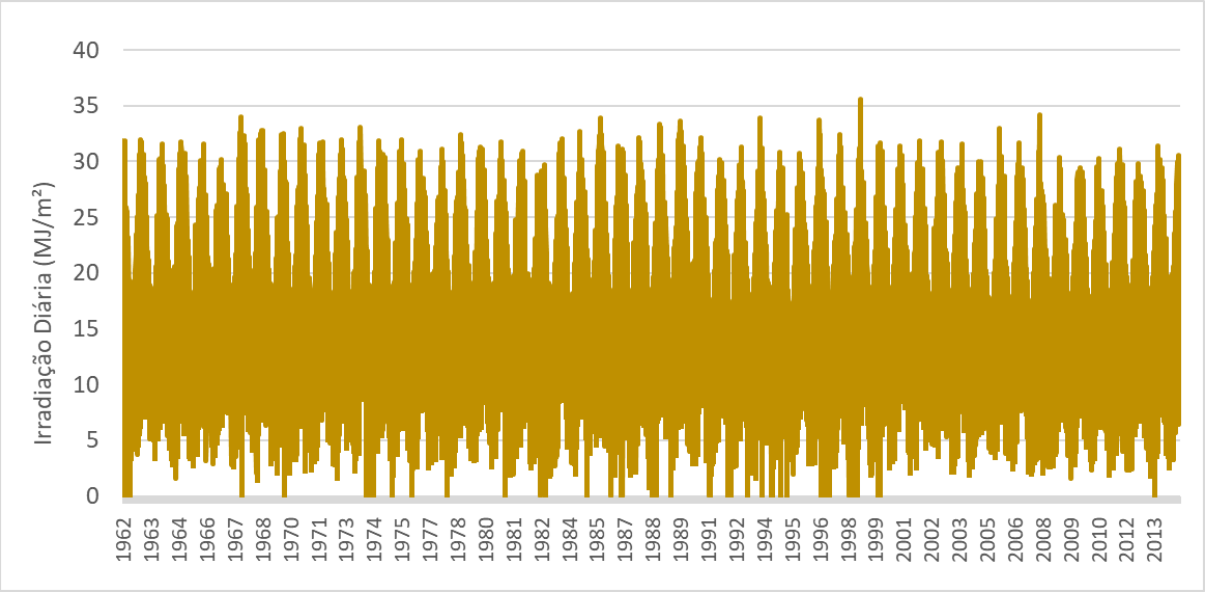


Figura 10 - Variação diária da Irradiação (MJ/m²) medida pela estação da USP

4.3 Análise de Correlação das Variáveis do Estudo

Toda a ingestão dos dados e o desenvolvimento dos modelos de ML foram feitos utilizando o software RStudio e, por meio da linguagem de programação R e suas bibliotecas. Após os dados estarem integrados e prontos para uso no software, realizou-se, como etapa inicial de análise exploratória, a compreensão da correlação das variáveis disponíveis para estudo.

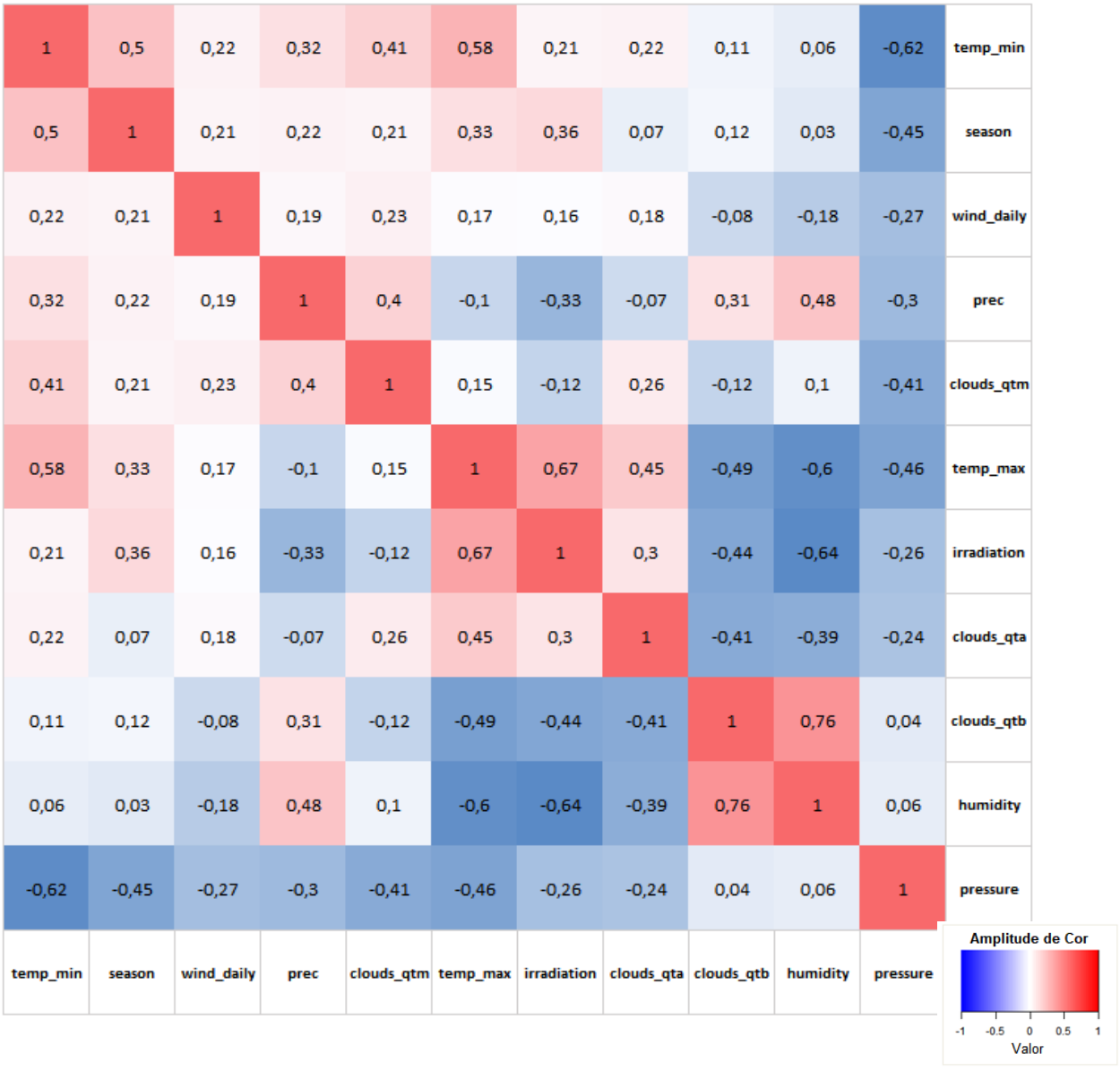
Na área de estudo de ML, entender a correlação e possível causalidade entre as variáveis disponíveis para a construção do modelo é um passo importante, se não fundamental, para evoluções incrementais em sua acurácia. A partir da análise de correlação, entende-se a associação entre duas variáveis e como variam conjuntamente ao longo de séries temporais. No caso deste estudo, busca-se entender como cada um dos parâmetros meteorológicos variam ao longo do tempo conjuntamente com as medições de irradiação, direcionando a utilização das variáveis mais importantes para a construção dos modelos.

Utilizou-se o coeficiente de Spearman para entender as correlações entre os parâmetros meteorológicos e a variável irradiação (variável de saída do modelo). Esse

coeficiente de correlação foi escolhido dado a possível relação monotônica entre as medidas de irradiação e dos parâmetros meteorológicos observados, capturando relações lineares e não-lineares. Relações monotônicas ocorrem em variáveis que apesar de tenderem a variar conjuntamente ao longo do tempo, não fazem isso a uma taxa constante, ao contrário de relações lineares capturadas por outros coeficientes frequentemente utilizados como o de Pearson, por exemplo.

A Figura 11 apresenta um mapa de calor que avalia essa correlação:

Figura 11 – Mapa de Calor: Matriz Spearman de Correlação

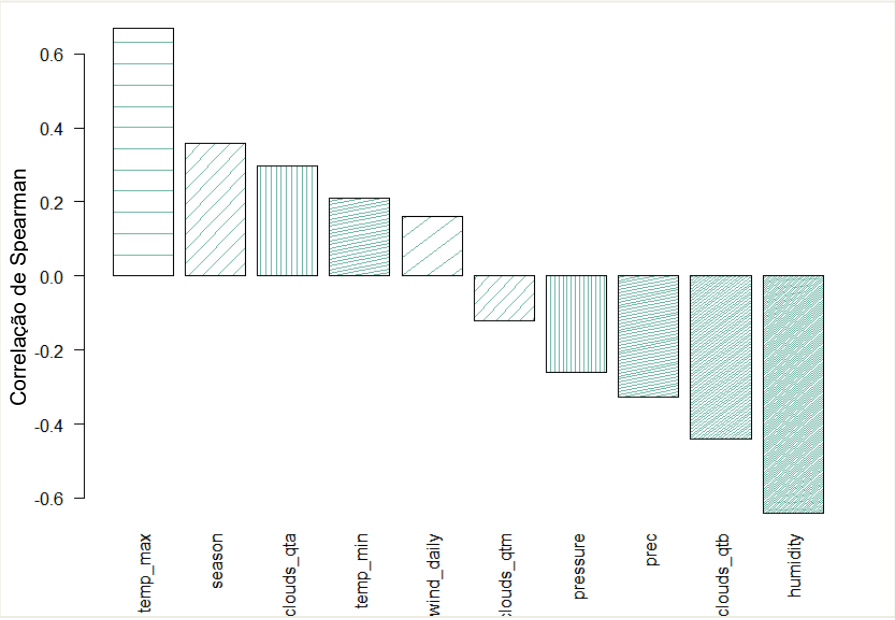


O mapa de calor pode ser lido como uma matriz de correlação entre as variáveis disponíveis para estudo. Dado que a base de estudo possui 11 parâmetros meteorológicos, o mapa de calor é representado por 11 linhas e 11 colunas com todos os parâmetros em cada eixo. Os números na matriz representam a correlação de Sperman entre as variáveis presentes na respectiva linha e coluna, variando entre -1 e 1. A escala de cores auxilia na visualização geral das correlações, sendo tonalidades de azul para correlações menores que 0, tonalidades bancas para correlações iguais a 0 e tonalidades de vermelho para correlações maiores que 0. Por fim, uma diagonal com valores iguais a 1 é formada dado que a correlação entre um parâmetro e ele mesmo é igual a 1.

Pela leitura do mapa de calor, busca-se entender a correlação entre a variável de estudo irradiação e os outros parâmetros meteorológicos disponíveis na base de dados. Em relação a irradiação, a maior correlação positiva está no parâmetro máxima temperatura diária (temp_max) (0,67) e na estação do ano (season) (0,36). Já as maiores correlações negativas estão na umidade do ar (humidity) (-0,64) e índice de cobertura de nuvens de baixa altitude (clouds_qtb) (-0,44). Essas correlações são esperadas dado que a irradiação ao nível do solo está altamente relacionada com o aumento da temperatura e proximidade com o verão. Por outro lado, umidade e cobertura de nuvens atuam como barreiras físicas para a radiação, reduzindo a energia que chega ao solo e, consequentemente, interferindo negativamente na conversão de energia solar em elétrica pelos painéis solares.

Para a implementação dos algoritmos foi necessário definir um critério para a seleção dos grupos de parâmetros a serem utilizados na fase de treinamento dos modelos. Utilizou-se como ponto de partida para a discussão a correlação de Spearman entre cada parâmetro meteorológico e irradiação solar, como apresentado pelo ranking na Figura 12:

Figura 12 – Ranking de correlação de Spearman entre cada variável e irradiação



Como é possível constatar, as variáveis com maior correlação positiva, ordenadas da maior para menor, são: temperatura máxima diária (temp_max), estação do ano (season), cobertura de nuvens de alta altitude (clouds_qta), temperatura mínima diária (temp_min) e maior rajada de vento diária (wind_daily). As variáveis com maiores correlações negativas, ordenadas da maior para menor, são: umidade (humidity), cobertura de nuvens de baixa altitude (clouds_qtb), precipitação (prec), pressão atmosférica (pressure) e cobertura de nuvens de média altitude (clouts_qtm), respectivamente.

Estes valores serão utilizados para orientar a implementação dos algoritmos nas seções seguintes deste trabalho.

4.4 Implementação dos Algoritmos

Os três algoritmos identificados na REL, SVM, ANN e ELM foram implementados no RStudio utilizando as bibliotecas “e1071”, “neuralnet” e “ELMR”, respectivamente. Um notebook equipado com a Microsoft Windows 10, processador Intel Core i7-8565U

CPU @ 1.80 GHz, 16gb de memória RAM DDR-4 e placa de vídeo GeForce GTX1650 4Gb GDDR5 foi utilizado para todas as implementações e simulações.

Os dados foram dimensionados e trabalhados previamente à implementação de cada algoritmo, seguindo o seguinte fluxo:

A base de dados com 19.358 registros foi embaralhada de maneira aleatória pela função “sample” no RStudio, selecionando-se uma semente de aleatoriedade. Esse procedimento visa evitar vieses de amostragem na separação e utilização dos dados nas etapas subsequentes;

i. As colunas de dados foram normalizadas para deixar todos os parâmetros meteorológicos em escala, não havendo assim distorção na construção dos modelos;
ii. Os registros foram divididos em duas diferentes bases de dados para treino e teste dos modelos: base de treino contendo 15.000 registros (77,5% da base inicial) e a base para testar a precisão dos modelos contendo 4.358 registros (22,5% da base inicial).

Foram desenvolvidos modelos para cada um dos algoritmos: SVM, ANN e ELM usando a mesma combinação de parâmetros meteorológicos.

Também com resultado da REL, três métricas foram utilizadas para comparar o desempenho de modelos na predição da variável de saída: MAE, RMSE e a correlação de Pearson entre a radiação real medida presente na base de dados e a radiação prevista pelo modelo. Também foi analisado o tempo de treino de cada modelo como uma variável de performance importante para compreender a possibilidade de aplicação desses modelos em um contexto real de conversão ou modelagem de geração de energia solar fotovoltaica.

MAE e RMSE são amplamente utilizadas como métricas para compreender a acurácia de predição de modelos de ML.

MAE é a média de todos os erros em um grupo de predições. Essa média é calculada com base na diferença entre cada valor real observado contido na base de dados de teste e valor previsto como saída dos modelos, em módulo. Dessa forma, MAE é definido pela equação (2):

$$\text{MAE} = \frac{1}{n}\sum_{j=1}^{n} \left| y_i - \hat{y}_j \right| \qquad (2)$$

Onde:

- N é número total de valores previstos e observados;
- y_i = valor observado;
- $\hat{y}_j$ = valor previsto.

RMSE é também uma métrica que avalia o tamanho do erro em uma amostra. Entretanto, os erros (diferença entre os valores reais observados e valores previstos) são elevados ao quadrado antes de comporem uma média aritmética simples e, finalmente, uma raiz quadrada é executada nesse valor médio. Dessa forma, RMSE aumenta o peso de erros maiores na métrica em comparação ao MAE. RMSE é uma métrica definida pela equação (3):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n} \left| y_i - \hat{y}_j \right|^2} \qquad (3)$$

Onde:

- N é número total de valores previstos e observados;
- y_i = valor observado;
- $\hat{y}_j$ = valor previsto.

Como um problema de regressão no âmbito de ML, os algoritmos de ANN e ELM são primeiramente otimizados em relação aos seus parâmetros de operação. Para ANN, o número de camadas ocultas de neurônios foi o parâmetro ajustado. Para ELM, como uma rede neural de única camada, o número de neurônios e a função de ativação foram selecionados para o melhor ajuste do problema de predição proposto.

Por fim, como parte integrante do Capítulo 5 deste trabalho, os resultados obtidos pelos modelos treinados foram confrontados para avaliar o impacto nas métricas de análise a partir de diferentes combinações e o número de parâmetros meteorológicos utilizados para treinar o modelo. Como apresentado na seção 1.1 buscou-se entender a partir de diferentes combinações dos parâmetros meteorológicos disponíveis a contribuição do número de parâmetros na entrada dos modelos e a

contribuição individual de cada parâmetro na melhoria de previsibilidade de radiação solar.

5 RESULTADOS E DISCUSSÃO

Este capítulo traz os resultados da implementação dos diferentes modelos de ML propostos, assim como uma discussão comparativa sobre a acurácia de cada modelo em prever a variável de saída: irradiação solar.

Os resultados apresentados neste capítulo resultaram na publicação do artigo (de Freitas Viscondi & Alves-Souza, 2021). Em consequência, os principais aspectos dessa publicação são descritos a seguir.

5.1 Comparação dos Modelos de ML

Com base na análise de correlações de Spearman apresentada anteriormente optou-se por seguir com 4 grupos de variáveis para implementação dos algoritmos, baseando-se no ranking de maiores correlações como apresentado pela Tabela 7. O procedimento adotado foi basicamente de agrupar as variáveis com maiores e menores correlações de Spearman, utilizar as variáveis presentes nesses grupos para treinar os modelos e, finalmente, comparar os resultados obtidos em cada um dos algoritmos.

Tabela 7 - Agrupamentos de Parâmetros Meteorológicos

Grupo	Critério	Parâmetros Selecionados
1	Top 1	temp_max + humidity
2	Top 1 e 2	temp_max + humidity + season + clouds_qtb
3	Top 1, 2 e 3	temp_max + humidity + season + clouds_qtb + clouds_qta + prec
4	Todos os Parâmetros	temp_max + humidity + season + clouds_qtb + clouds_qta + prec + temp_min + pressure + wind_daily + clouds_qtm

No Grupo 1, tem-se dois parâmetros meteorológicos que apresentam as maiores correlações positivas e negativas em relação à irradiação solar. No Grupo 2, as primeiras e segundas maiores correlações positiva e negativa. No Grupo 3, adicionou-

se as terceiras correlações positiva e negativa, totalizando 6 parâmetros. E, por fim, no Grupo 4, todos os 10 parâmetros meteorológicos da base de dados foram adicionados para treinar os modelos.

Com os grupos definidos e os algoritmos com os parâmetros de operação previamente ajustados, iniciou-se o treinamento dos modelos pelos algoritmos de SVM. A Tabela 8 sintetiza e compara os resultados para todos os 4 modelos construídos com o algoritmo SVM:

Tabela 8 – Resultados de previsão dos modelos com algoritmo SVM

SUPPORT VECTOR MACHINES (SVM)					
Modelo	**Grupo**	**MAE [MJ/m²]**	**RMSE [MJ/m²]**	**CORRELAÇÃO DE PEARSON**	**TEMPO DE TREINO [s]**
SVM_1	1	3,08	4,15	0,76	29,15
SVM_2	2	2,54	3,43	0,84	28,99
SVM_3	3	2,41	3,24	0,86	28,66
SVM_4	4	2,05	2,78	0,89	35,10

É possível notar um incremento da acurácia de previsão do modelo de acordo com a inserção de mais parâmetros meteorológicos. O modelo SVM_4 é o modelo com menor erro de previsão, apresentando um MAE de 2,05 MJ/m² ou 12,7% da irradiação média na série temporal de dados – 16,2 MJ/m². Com o aumento do número de variáveis de treino, é possível notar quase nenhum impacto no tempo de treinamento do modelo.

O segundo algoritmo a ser treinado foi o ANN. Várias configurações de redes neurais foram testadas e os parâmetros que mais se ajustaram para aplicação nesse problema de previsão foram a tangente hiperbólica como função de ativação, 5 neurônios na primeira camada oculta, 2 camadas ocultas e 0,5 como limite da função de ativação.

O aumento do número de neurônios, camadas ocultas e a redução do limite da função de ativação geraram quase nenhum impacto positivo nos resultados de previsão inicialmente testados, trazendo um grande impacto nos custos

computacionais para a previsão (aumento relevante no tempo de treinamento). Por exemplo, a redução do limite da função de ativação de 0,5 para 0,1, reduziu o MAE de 3,13 para 3,09, às custas de um aumento de 55 vezes o tempo de treino do modelo. A Tabela 9 sintetiza e permite a comparação dos resultados para os modelos de ANN implementados:

Tabela 9 - Resultados de previsão dos modelos com algoritmo ANN

ARTIFICIAL NEURAL NETWORK (ANN)					
Modelo	**Grupo**	**MAE [MJ/m²]**	**RMSE [MJ/m²]**	**CORRELAÇÃO DE PEARSON**	**TEMPO DE TREINO [s]**
ANN_1	1	3,13	4,12	0,76	16,6
ANN_2	2	2,70	3,58	0,83	25,9
ANN_3	3	2,67	3,48	0,83	39,6
ANN_4	4	2,24	2,99	0,88	29,4

As redes neurais artificiais apresentaram um tempo de treinamento semelhante aos modelos com SVM, em um tempo médio de 28 segundos para completar o processo de treinamento. O comportamento em relação ao incremento do número de parâmetros meteorológicos utilizados para treino também é semelhante ao constatado no SVM, com redução dos erros à medida que novos parâmetros são adicionados no treino.

Em relação aos erros – MAE e RMSE – e a correlação de Pearson, os resultados quando comparados os mesmos grupos de parâmetros são menores na implementação de modelos com ANN. O MAE para o modelo ANN_4, o modelo com melhores resultados de previsão para o algoritmo de ANN, é 13,8% da média dos valores de irradiação observados na série temporal da base de dados.

Por fim, implementou-se os modelos com o algoritmo ELM. Assim como feito para ANN, configurou-se o algoritmo de ELM na melhor configuração de parâmetros de operação observada para este problema de regressão. A função seno foi utilizada como ativação, 100 neurônios foram configurados na única camada oculta e os dados foram processados em blocos de 50 unidades. A Tabela 10 sintetiza e permite a comparação dos resultados obtidos pelos modelos de ELM:

Tabela 10 - Resultados de previsão dos modelos com algoritmo ELM

EXTREME LEARNING MACHINE (ELM)					
Modelo	**Grupo**	**MAE [MJ/m²]**	**RMSE [MJ/m²]**	**CORRELAÇÃO DE PEARSON**	**TEMPO DE TREINO [s]**
ELM_1	1	3,31	4,30	0,73	3,27
ELM_2	2	2,84	3,73	0,80	1,35
ELM_3	3	2,77	3,63	0,82	1,19
ELM_4	4	2,35	3,09	0,87	1,15

O tempo de treinamento para os modelos é consistente e razoavelmente inferior quando comparado aos outros algoritmos. O tempo médio de treino para modelos com ELM foi 94,3% menor quando comparado aos modelos com SVM e 93,8% menor quando a comparação é feita com modelos que utilizam o algoritmo de ANN.

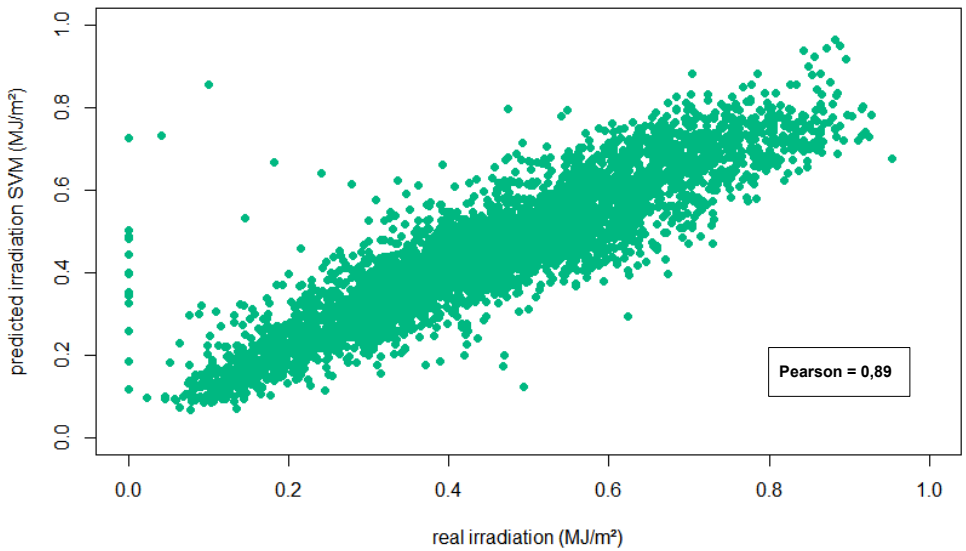
É interessante notar que, diferente do observado nos outros algoritmos, quando novos parâmetros meteorológicos são adicionados na fase de treino do modelo, o tempo total do processo é reduzido. Como nos outros modelos, mais parâmetros representam menores erros e, dessa forma, os modelos com ELM apresentam incremento de acurácia sem quase nenhum incremento de custos computacionais na fase de treino. O MAE para o modelo ELM_4 é 14,5% do valor médio de irradiação observado na série temporal.

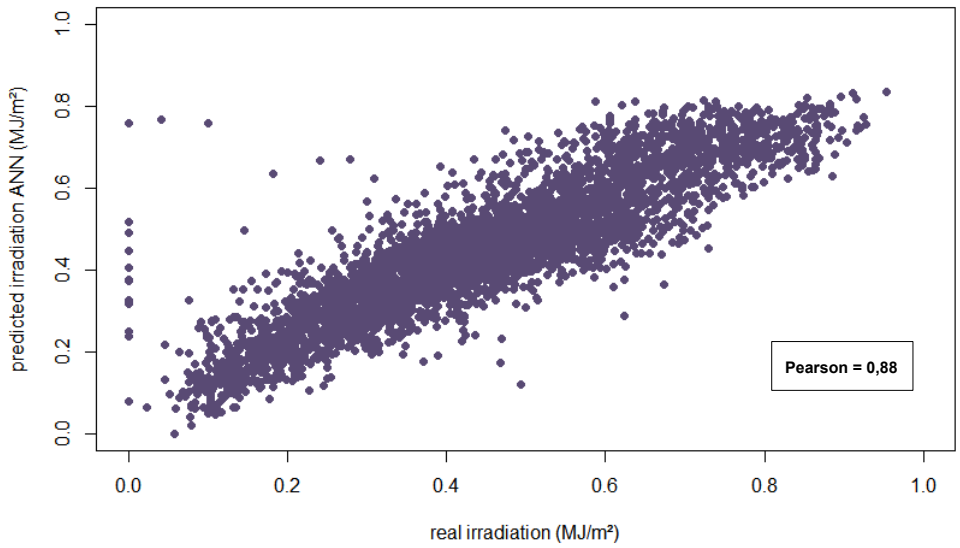
Todos os três algoritmos apresentaram suas melhores acurácias quando todos os parâmetros meteorológicos presentes na base de dados foram adicionados como variáveis de treino, demonstrando que todos contribuem incrementalmente para a previsão da variável de saída, o que responde um dos objetivos chave deste trabalho de pesquisa.

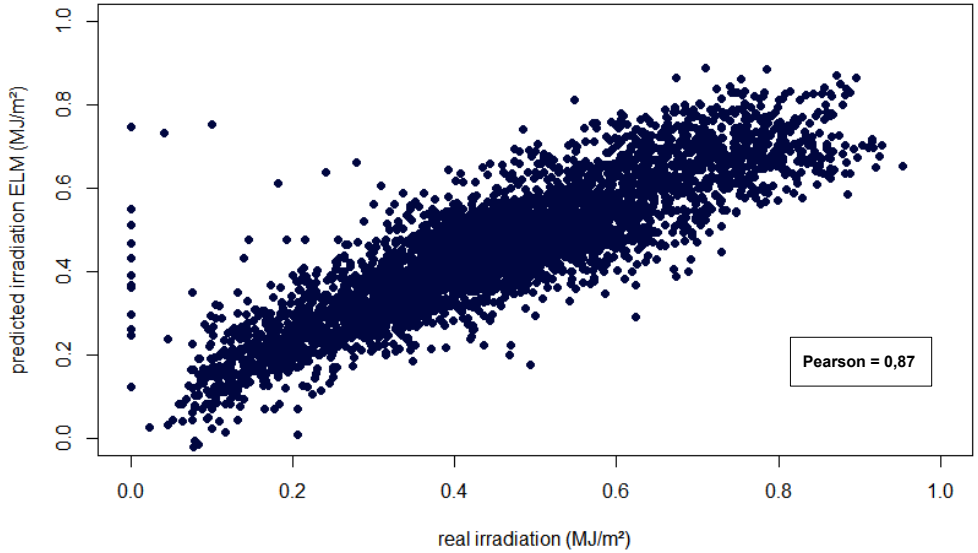
Como apresentado na Figura 13, os resultados de previsão para os modelos implementados com os algoritmos SVM, ANN e ELM são próximos quando considerado o mesmo agrupamento de parâmetros meteorológicos. Entretanto, quando aplicado o Grupo 4 de parâmetros meteorológicos e comparadas as métricas

de MAE, RMSE e correlação de Pearson, existe uma pequena vantagem em termos de precisão para o algoritmo SVM, seguido por ANN e finalmente ELM.

Figura 13 – Irradiação prevista versus irradiação real observada quando todos os parâmetros meteorológicos contidos na base de dados são utilizados (Grupo 4). Algoritmo de SVM (gráfico superior esquerdo), ANN (gráfico superior direito) e ELM (gráfico inferior central)







6 CONCLUSÃO

Embora existam muitas discussões e trabalhos científicos apresentando abordagens para reduzir os impactos negativos da variabilidade de geração de energia solar, incluindo múltiplos avanços centrados em modelos de aprendizado de máquina, é importante ressaltar as contribuições deste trabalho ao tema.

Empregando-se a metodologia de revisão de escopo da literatura, apresentou-se no Capítulo 3 uma revisão detalhada da literatura sobre a aplicação de técnicas focadas em contextos de *big data* para a previsão de geração de energia solar fotovoltaica. Até a data de publicação do artigo referente ao tema, nenhum trabalho de pesquisa envolvendo esta metodologia havia sido disponibilizado na literatura nas bases de pesquisa consultadas e os resultados encontrados durante esta fase constituíram nossas primeiras contribuições ao tema.

Durante o desenvolvimento da REL, entendeu-se após análise minuciosa e sistemática dos 38 artigos encontrados que as técnicas de aprendizado de máquina são as abordagens mais empregadas no contexto de estudo, assim como a prevalência de modelos focam suas iniciativas na previsibilidade do recurso natural – irradiação – mais do que na fase de conversão da energia solar em elétrica. Como discutido nas sessões anteriores deste trabalho, a conversão em energia elétrica nos painéis solares é bastante estável e com eficiência bastante conhecida de conversão. O desafio para a fonte está na compreensão locacional da usina de geração dado que diversos fatores locais e pontuais, como incidência de nuvens, umidade, pluviosidade, dentre outros aspectos, afetam imediatamente a geração de eletricidade, evidenciando-se a necessidade de avaliar os parâmetros meteorológicos locais para entender o comportamento futuro da incidência solar.

Em sequência, estabeleceu-se o estado da arte em relação aos algoritmos utilizados para endereçar o problema, assim como as métricas recorrentemente empregadas para estabelecer réguas comuns de avaliação dos modelos de previsão. Constatou-se que, apesar de inúmeros algoritmos terem sido empregados na

literatura – *random forest, decision trees, gradiente boosting,* dentre outros – *support vector machines* (SVM), *artificial neural networks* (ANN) e *extreme learning machines* (ELM) foram os algoritmos mais citados e constantemente referenciados como os mais precisos quando comparadas as métricas de performance *mean average error* (MAE), *root mean square error* (RMSE) e correlações de Pearson entre os valores previstos e observados.

Como último resultado desta etapa do trabalho, constatou-se que somente 26% trabalhos de pesquisa analisados apresentaram com clareza preocupações em relação ao impacto da qualidade dos dados de entrada nos resultados dos modelos construídos. Neste caso, ratifica-se a necessidade de maior atenção durante a fase de limpeza e modelagem dos dados que acontece previamente à construção de robustos modelos de previsão.

Em um segundo estágio do trabalho, é importante ressaltar a contribuição da implementação dos modelos de aprendizado de máquina em um contexto brasileiro de previsibilidade de geração. Até a data de publicação do artigo referente ao tema, nenhum artigo foi encontrado na literatura apresentando modelos de aprendizado de máquina aplicados em um contexto nacional de previsão, sendo importante para as evoluções locais acerca das discussões de penetração de energia solar fotovoltaica na matriz elétrica brasileira.

Ao utilizar a base de dados fornecida pela USP contendo dados para 10 parâmetros meteorológicos com medidas diárias entre 1962 e 2014, constatou-se que os três algoritmos implementados possuem acurácia de previsão próximas, com valores para correlação de Pearson entre irradiação prevista e observada entre 0,87 e 0,89, nas melhores configurações de implementação. Entretanto, existe ligeira vantagem para os modelos implementados com o algoritmo SVM quando comparados os resultados nas métricas propostas.

Por fim, entendeu-se que existe um aumento da acurácia dos modelos de previsão à medida que novos parâmetros meteorológicos contidos na base de dados são incorporados na fase de treino dos modelos, sugerindo que utilizar somente os

parâmetros com maiores correlações de *Spearman* em relação às medidas de irradiação não é suficiente para uma boa acurácia de previsão localmente.

Entende-se, portanto, que os resultados apresentados por este trabalho de pesquisa ajudam a levantar opções e a direcionar escolhas para endereçar o aumento da previsibilidade da energia solar como fonte de conversão em energia elétrica. Desta forma, os algoritmos implementados neste trabalho podem contribuir consistentemente para a construção de modelos integrados ao SIN, aumentando a previsibilidade e confiabilidade de geração elétrica de forma a auxiliar a operabilidade de um sistema elétrico de potência de dimensões continentais e fundamentado completamente em fontes renováveis de energia primária.

Ressalta-se que parte do referencial bibliográfico utilizado para o desenvolvimento desse trabalho concentra-se entre os anos de 2013 e 2018 devido ao período de seleção de artigos para a elaboração da revisão de escopo da literatura. Entende-se como um próximo passo de evolução, dado a constante evolução de algoritmos de aprendizado de máquina, a atualização desse referencial bibliográfico principalmente com foco em buscar novos algoritmos e abordagens para o problema de previsibilidade do recurso solar fotovoltaico.

Como continuidade deste trabalho, sugere-se avaliar os efeitos da má qualidade dos dados para os modelos de previsibilidade propostos. Como abordado no capítulo da REL, poucos trabalhos abordam os impactos da qualidade dos dados nos resultados de previsão. Dado a constante evolução da pesquisa em algoritmos de aprendizado de máquina, indica-se avaliar, também, a capacidade preditiva e acurácia de outros algoritmos que venham a surgir na literatura, assim como possibilidades de ajustes finos nos algoritmos abordados por este trabalho.

Vale ressaltar a importância de efetuar comparações entre os modelos de aprendizado de máquina propostos por essa dissertação e outros modelos de predição fundamentados em outras técnicas como, por exemplo, modelos de previsão numérica de tempo que continuam fundamentais para abordagens de previsão climáticas. Por fim, dada a proposta de utilização em âmbito nacional dos modelos

construídos neste trabalho, é recomendável a implementação do mesmo procedimento de avaliação em outras regiões do país.

REFERÊNCIAS

AGORA. The integration costs of wind and solar power: an overview of the debate on the effects of adding wind and solar photovoltaic into power systems. Berlin: Agora Energiewende, 2015.

ANEEL. Resolução Aneel n 482 de 17 de abril de 2012. Estabelece as condições gerais para o acesso de microgeração e minigeração distribuída aos sistemas de distribuição de energia elétrica, o sistema de compensação de energia elétrica, e dá outras providências. Disponível em: http://www2.aneel.gov.br/cedoc/ren2012482.pdf. Acesso em 30 de julho de 2020.

Atlas brasileiro de energia solar / Enio Bueno Pereira; Fernando Ramos Martins; Samuel Luna de Abreu e Ricardo Rüther. – São José dos Campos : INPE, 2006

Aybar-Ruiz, A., Jiménez-Fernández, S., L., Cornejo-Bueno, C., Casanova-Mateo, J. Sanz-Justo, P. Salvador-González et al., A novel Grouping Genetic Algorithm–Extreme Learning Machine approach for global solar radiation prediction from numerical weather models inputs, Solar Energy. 132 (2016) 129-142. doi:10.1016/j.solener.2016.03.015.

B.D.Ripley, Pattern Recognition and Neural Networks, Cambridge University Press, Cambridge (1996)

Boccard, N. (2009). *Capacity factor of wind power realized values vs. estimates. Energy Policy, 37(7), 2679–2688.* doi:10.1016/j.enpol.2009.02.046

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. Proceedings of the Fifth Annual Workshop on Computational Learning Theory - COLT ’92. doi:10.1145/130385.130401

Brasil. Lei n° 14.300, de 06 de janeiro de 2022. Institui o marco legal da microgeração e minigeração distribuída, o Sistema de Compensação de Energia Elétrica (SCEE) e o Programa de Energia Renovável Social (PERS); altera as Leis nºs 10.848, de 15 de

março de 2004, e 9.427, de 26 de dezembro de 1996; e dá outras providências. Diário Oficial da União, Brasilia, DF, p. 04, 07 jan. 2022.

Burianek, T., and Stanislav, M. “Solar Irradiance Forecasting Model Based on Extreme Learning Machine.” 2016 IEEE 16th International Conference on Environment and Electrical Engineering (EEEIC) (2016). doi:10.1109/eeeic.2016.7555445.

CCEE. Estatísticas do resultado do 37° Leilão de Energia Nova A-5. Câmara de Comerciailzação de Energia Elétrica. Brasília, 2022. Disponível em: https://static.poder360.com.br/2022/10/Estatisticas_LEN_A-5_2022.pdf. Acesso em 28 de dezembro de 2022.

CETEM – Centro de Tecnologia Mineral. Silício Grau Solar - Uma Revisão das Tecnologias de Produção (2019).

Cherry, K. M., & Qian, L. (2018). Scaling up molecular pattern recognition with DNA-based winner-take-all neural networks. Nature, 559(7714), 370–376. doi:10.1038/s41586-018-0289-6

Chia, Y., Lee, L., Shafiabady, N. and Isa, D.A load predictive energy management system for supercapacitor-battery hybrid energy storage system in solar application using the Support Vector Machine. Applied Energy, 137, pp.588-602. (2015).

Cresesb. Atlas do Potencial Eólico Brasileiro. Centro de Referência para Energia Solar e Eólica. Brasília, 2001.

DAMA International (2017) DAMA-DMBOK 2 | Data Management Body of Knowledge. second. Edited by C. Deborah Henderson, C. Susan Earley, and I. Laura Sebastian-Coleman, CDMP. USA: Technics Publications.

de Freitas Viscondi, G., & Alves-Souza, S. N. (2019). A Systematic Literature Review on big data for solar photovoltaic electricity generation forecasting. *Sustainable Energy Technologies and Assessments*, *31*(November 2018), 54–63. https://doi.org/10.1016/j.seta.2018.11.008.

de Freitas Viscondi, G., & Alves-Souza, S. N. (2021). Solar Irradiance Prediction with Machine Learning Algorithms: A Brazilian Case Study on Photovoltaic Electricity Generation. Energies. 14. 5657. 10.3390/en14185657.

de Souza, J. D., Silva, B. B., Ceballos, J. C. (2008) Estimativa de radiação solar global à superfície usando um modelo estocástico: caso sem nuvens. Revista Brasileira de Geofísica.

EPE. Análise da Inserção da Geração Solar na Matriz Elétrica Brasileira. Rio de Janeiro: Empresa de Pesquisa Energética, 2012.

EPE. Informe dos Resultados da Habilitação Técnica e Vencedores do Leilão A-6 de 2018. Rio de Janeiro: Empresa de Pesquisa Energética, 2018.

EPE. Plano Nacional de Energia 2050 / Ministério de Minas e Energia. Empresa de Pesquisa Energética. Brasília: Ministério de Minas e Energia e Empresa de Pesquisa Energética, 2020

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

ESMC. PV Manufacturing Lessons Learned in Europe. Bruxelas: European Solar Manufacturing Council, 2021.

Francisco, A. C. C., Vieira, H. E. M., Romano, R. R., & Roveda, S. R. M. M. (2019). Influência de parâmetros meteorológicos na geração de energia em painéis fotovoltaicos: um caso de estudo do smart campus Facens, SP

Francisco, Maritza M. C., Solange N. Alves-Souza, Edit G. L. Campos, and Luiz S. De Souza. 2017. “Total Data Quality Management and Total Information Quality Management Applied to Costumer Relationship Management.” In ICIME 2017: 2017 9th International Conference on Information Management and Engineering. Barcelona, Espanha: ACM. https://doi.org/https://doi.org/10.1145/3149572.3149575.

G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, Extreme learning machine for regression and multiclass classification, IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 42, no. 2, pp. 513–529, Apr. 2012.

G.-B. Huang, L. Chen, and C.-K. Siew, Universal approximation using incremental constructive feedforward networks with random hidden nodes, IEEE Trans. Neural Netw., vol. 17, no. 4, pp. 879–892, Jul. 2006.

Gala, Y., et al. “Hybrid Machine Learning Forecasting of Solar Radiation Values.”Neurocomputing, vol. 176 pp. 48–59 (2016). doi:10.1016/j.neucom.2015.02.078.

HAUPT, S. E. Variable generation power forecasting as a big data problem. IEEE Transactions on Sustainable Energy, v. 90, n. 99, p. 1-1, 2016.

HAUPT, S. E.; KOSOVIC, B. Big data and machine learning for applied weather forecasts: Forecasting solar power for utility operations. Proceedings - 2015 IEEE Symposium Series on Computational Intelligence, SSCI 2015. 2016.

Huang, G.-B. (2014). An Insight into Extreme Learning Machines: Random Neurons, Random Features and Kernels. Cognitive Computation, 6(3), 376–390. doi:10.1007/s12559-014-9255-2.

Huang, G.-B., Wang, D. H., & Lan, Y. (2011). Extreme learning machines: a survey. International Journal of Machine Learning and Cybernetics, 2(2), 107–122. doi:10.1007/s13042-011-0019-y

IEA. Next Generation Wind and Solar Power: From Cost to Value. Paris: International Energy Agency, 2016.

IEA. Renewable Energy: Medium-term Market Report 2016. Paris: International Energy Agency, 2016.

IEMA. Prioridades para a integração das fontes renováveis variáveis no sistema elétrico. São Paulo: Instituto de Energia e Meio Ambiente, 2016.

IRENA, Renewable capacity statistics 2019, International Renewable Energy Agency (IRENA), Abu Dhabi 2019.

IRENA, Renewable capacity statistics time series, International Renewable Energy Agency (IRENA), Abu Dhabi 2022. Disponível em: https://www.irena.org/Data/View-data-by-topic/Capacity-and-Generation/Statistics-Time-Series. Acesso em 28 de dezembro de 2022.

IRENA and ADFD (2020), Advancing renewables in developing countries: Progress of projects supported through the IRENA/ADFD Project Facility, International Renewable Energy Agency (IRENA) and Abu Dhabi Fund for Development (ADFD), Abu Dhabi.

IRENA. Global Energy Transformation: A roadmap to 2050 - 2019 Edition. International Renewable Energy Agency (IRENA), Abu Dhab (2019).

IRENA. The Power to Change: Solar and Wind Cost Reduction Potential to 2025. International Renewable Energy Agency (IRENA), Abu Dhab (2016).

IRENA. Future of Solar Photovoltaic: Deployment, investment, technology, grid integration and socio-economic aspects. International Renewable Energy Agency (IRENA), Abu Dhab (2019). Disponível em: https://www.irena.org/-/media/Files/IRENA/Agency/Publication/2019/Nov/IRENA_Future_of_Solar_PV_2019.pdf. Acesso em 28 de dezembro de 2022.

IRENA. Scenarios for the Energy Transition: Experience and Good Practices in Latin America and the Caribbean. International Renewable Energy Agency (IRENA), Abu Dhab (2022).

Jaradat, M., Jarrah, M., Bousselham, A., Jararweh, Y., & Al-Ayyoub, M. (2015). The Internet of Energy: Smart Sensor Networks and Big Data Management for Smart Grid. Procedia Computer Science, 56, 592–597. doi:10.1016/j.procs.2015.07.250.

KELMAN, R. Inserção de Fontes Renováveis na Matriz Elétrica Brasileira. In: PRIORIDADES PARA A INTEGRAÇÃO DE RENOVÁVEIS NA MATRIZ ELÉTRICA BRASILEIRA. São Paulo, 2016.

Kitchenham, B., Charters, S. Guidelines for performing Systematic Literature Reviews in Software Engineering, (2007).

Kleissl, J., Lave, M., Jamaly, M., & Bosch, J. (2012). Aggregate solar variability. 2012 IEEE Power and Energy Society General Meeting. doi:10.1109/pesgm.2012.6344809.

LOPES, Mariana Granzoto. Análise dos impactos técnicos resultantes da variabilidade de geração de curto prazo de sistemas fotovoltaicos. 2015. 1 recurso online (124 p.). Dissertação (mestrado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação, Campinas, SP. Disponível em: <http://www.repositorio.unicamp.br/handle/REPOSIP/259049>. Acesso em: 29 ago. 2018.

Lorena, A. C. L., Carvalho, A. C. P. L. F (2007). Uma introdução às Support Vector Machines. Revista de Informática Aplicada e Teórica. Doi: 10.22456/2175-2745.5690

Lou, S., et al. “Prediction of Diffuse Solar Irradiance Using Machine Learning and Multivariable Regression.” Applied Energy, vol. 181. (2016) pp. 367–374. doi:10.1016/j.apenergy.2016.08.093.

MCKINSEY & COMPANY. Big data: The next frontier for innovation, competition, and productivity. McKinsey Global Institute, n. June, p. 156, 2011.

Mouriño, G. L. de, Assireu, A. T., & Pimenta, F. (2016). Regularization of hydroelectric reservoir levels through hydro and solar energy complementarity. RBRH, 21(3), 549–555. doi:10.1590/2318-0331.011615174

Namba, M., & Zhang, Z. (n.d.). Cellular Neural Network for Associative Memory and Its Application to Braille Image Recognition. The 2006 IEEE International Joint Conference on Neural Network Proceedings. doi:10.1109/ijcnn.2006.1716416

Noble, W. S. (2006). What is a support vector machine? Nature Biotechnology, 24(12), 1565–1567. doi:10.1038/nbt1206-1565

Odom, M. D., & Sharda, R. (1990). A neural network model for bankruptcy prediction. 1990 IJCNN International Joint Conference on Neural Networks. doi:10.1109/ijcnn.1990.137710

ONS. Boletim Mensal de Geração Eólica – Dezembro de 2019. (2019) Disponível em: http://www.ons.org.br/AcervoDigitalDocumentosEPublicacoes/Boletim%20Mensal%20de%20Gera%C3%A7%C3%A3o%20Eolica%202019-12.pdf. Acesso em 30 de julho de 2020.

Pacheco, A. (2017). A máquina de aprendizado extremo – ELM. Disponível em https://computacaointeligente.com.br/algoritmos/maquina-de-aprendizado-extremo/. Acesso em 06 de novembro de 2022.

PwC. Data Quality Survey. Reino Unido: PricewaterhouseCoopers, 2004.

Revankar, P. S., Thosar, A. G. and Gandhare, W. Z., "Maximum Power Point Tracking for PV Systems Using MATALAB/SIMULINK," 2010 Second International Conference on Machine Learning and Computing, Bangalore, (2010) pp. 8-11. doi: 10.1109/ICMLC.2010.54

Sadiq, S. and Papotti, P. (2016) 'Big data quality - whose problem is it?', in 2016 IEEE 32nd International Conference on Data Engineering (ICDE), pp. 1446–1447. doi: 10.1109/ICDE.2016.7498367.

Shamshirband, S., Mohammadi, K., Yee, P., Petković, D., Mostafaeipour, A.,, A comparative evaluation for identifying the suitability of extreme learning machine to predict horizontal global solar radiation, Renewable And Sustainable Energy Reviews. 52 (2015) 1031-1042. doi:10.1016/j.rser.2015.07.173.

Shuo, L., Jun, M., Xiong, M., Hui, H. G., & Wei, H. R. (2016). The platform of monitoring and analysis for solar power data. 2016 China International Conference on Electricity Distribution (CICED). doi:10.1109/ciced.2016.7575906.

Simmhan , Y. et al., "Cloud-Based Software Platform for Big Data Analytics in Smart Grids," in Computing in Science & Engineering, vol. 15, no. 4, pp. 38-47, (2013). doi: 10.1109/MCSE.2013.39

Singh, B., Dwivedi, S., Hussain, I., & Verma, A. K. (2014). Grid integration of solar PV power generating system using QPLL based control algorithm. 2014 6th IEEE Power India International Conference (PIICON). doi:10.1109/34084poweri.2014. 7117785

Smart Grid Consumer Collaborative. Smart Grids: Economic and Environmental Benefits: A Review and Synthesis of Research on Smart Grid Benefits and Costs. 2013

Suka, Machi & Oeda, Shinichi & Ichimura, Takumi & Yoshida, Katsumi & Takezawa, Jun. (2007). Advantages and Disadvantages of Neural Networks for Predicting Clinical Outcomes.. IMECS 2007: International Multiconference of engineers and computer scientists. 839-844.

Sun, A., Lim, E.-P., & Liu, Y. (2009). On strategies for imbalanced text classification using SVM: A comparative study. Decision Support Systems, 48(1), 191–201. doi:10.1016/j.dss.2009.07.011

Tang, J., Deng, C., & Huang, G.-B. (2016). Extreme Learning Machine for Multilayer Perceptron. IEEE Transactions on Neural Networks and Learning Systems, 27(4), 809–821. doi:10.1109/tnnls.2015.2424995

Tuia, D., Ratle, F., Pacifici, F., Kanevski, M. F., & Emery, W. J. (2009). Active Learning Methods for Remote Sensing Image Classification. IEEE Transactions on Geoscience and Remote Sensing, 47(7), 2218–2232. doi:10.1109/tgrs.2008.2010404

Voyant, C., Notton, G., Kalogirou, S., Nivet, M. L., Paoli, C., Motte, F., & Fouilloy, A. (2017). Machine learning methods for solar radiation forecasting: A review. *Renewable Energy*, *105*, 569–582. https://doi.org/10.1016/j.renene.2016.12.095

Wang, S.-C. (2003). Artificial Neural Network. Interdisciplinary Computing in Java Programming, 81–100. doi:10.1007/978-1-4615-0377-4_5

WWF-Brasil. Desafios e Oportunidades para a energia eólica no Brasil: recomendações para políticas públicas. Brasília: World Wildlife Fund, 2015.

Yang, Z. R. (2004). Biological applications of support vector machines. Briefings in Bioinformatics, 5(4), 328–338. doi:10.1093/bib/5.4.328

Yesilbudak, M., Colak, M., & Bayindir, R. (2016). A review of data mining and solar power prediction. *2016 IEEE International Conference on Renewable Energy Research and Applications, ICRERA 2016*, *0*, 1117–1121. https://doi.org/10.1109/ICRERA.2016.7884507

Zambon, R. C. (2015). A operação dos reservatórios e o planejamento da operação hidrotérmica do Sistema Interligado Nacional. Revista USP, (104), 133. doi:10.11606/issn.2316-9036.v0i104p133-144

Zanetta Júnior, L. C. Fundamentos de sistemas elétricos de potência. 1ed. – São Paulo: Editora Livraria Física, 2005.

ZHANG, D. Inconsistencies in big data. Proceedings of the 12th IEEE International Conference on Cognitive Informatics and Cognitive Computing, 2013

Zhaoxuan. L. et al. “A Hierarchical Approach Using Machine Learning Methods in Solar Photovoltaic Energy Production Forecasting.” Energies, vol. 9, no. 12 p. 55 (2016). doi:10.3390/en9010055.

Zhou, K., Fu, C., & Yang, S. (2016). Big data driven smart energy management: From big data to big insights. Renewable and Sustainable Energy Reviews, 56, 215–225. doi:10.1016/j.rser.2015.11.050

ANEXO A - REFERÊNCIAS UTILIZADAS NA REVISÃO DE ESCOPO DA LITERATURA

[1] IRENA. “Renewable capacity statistics 2017”. International Renewable Energy Agency (IRENA), Abu Dhab (2017).

[2] EPE. “Balanço Energético Nacional 2016, Ano-base 2015”. Brasília: Empresa de Pesquisa Energética, (2016)

[3] EPE. “O Compromisso do Brasil no Combate às Mudanças Climáticas: Produção e Uso de Energia”. Rio de Janeiro: Empresa de Pesquisa Energética, (2016).

[4] Agora Energiewende. “The Integration Cost of Wind and Solar Power. An Overview of the Debate on the Effects of Adding Wind and Solar Photovoltaic into Power Systems” (2015).

[5] Shaheen, M., Khan, M. A method of data mining for selection of site for wind turbines, Renewable And Sustainable Energy Reviews. 55 (2016) 1225-1233. doi:10.1016/j.rser.2015.04.015.

[6] Hu, T., Zheng, M., Tan, J., Zhu, L., Miao, W., Intelligent photovoltaic monitoring based on solar irradiance big data and wireless sensor networks, Ad Hoc Networks. 35 (2015) 127-136. doi:10.1016/j.adhoc.2015.07.004.

[7] Yuregir, O., Sagiroglu, C., Solar Energy Validation for Strategic Investment Planning via Comparative Data Mining Methods: An Expanded Example within the Cities of Turkey, International Journal Of Photoenergy. 2016 (2016) 1-16. doi:10.1155/2016/8506193.

[8] Aybar-Ruiz, A., Jiménez-Fernández, S., L., Cornejo-Bueno, C., Casanova-Mateo, J. Sanz-Justo, P. Salvador-González et al., A novel Grouping Genetic Algorithm–Extreme Learning Machine approach for global solar radiation prediction from numerical weather models inputs, Solar Energy. 132 (2016) 129-142. doi:10.1016/j.solener.2016.03.015.

[9] Burianek, T., and Stanislav, M. “Solar Irradiance Forecasting Model Based on Extreme Learning Machine.” 2016 IEEE 16th International Conference on Environment and Electrical Engineering (EEEIC) (2016). doi:10.1109/eeeic.2016.7555445.

[10] Shamshirband, S., Mohammadi, K., Yee, P., Petković, D., Mostafaeipour, A.,, A comparative evaluation for identifying the suitability of extreme learning machine to predict horizontal global solar radiation, Renewable And Sustainable Energy Reviews. 52 (2015) 1031-1042. doi:10.1016/j.rser.2015.07.173.

[11] Bouzgou, H. A fast and accurate model for forecasting wind speed and solar radiation time series based on extreme learning machines and principal components analysis, Journal of Renewable And Sustainable Energy. 6 (2014) 013114. doi:10.1063/1.4862488.

[12] Abadi, Imam, et al. “Extreme Learning Machine Approach to Estimate Hourly Solar Radiation on Horizontal Surface (PV) in Surabaya -East Java.” 2014 The 1st International Conference on Information Technology, Computer, and Electrical Engineering. (2014) doi:10.1109/icitacee.2014.7065774.

[13] Pinto, Tiago, et al. “Solar Intensity Characterization Using Data-Mining to Support Solar Forecasting.” Distributed Computing and Artificial Intelligence, 12th International Conference Advances in Intelligent Systems and Computing (2015) pp. 193–201., doi:10.1007/978-3-319-19638-1_22.

[14] Lou, S., et al. “Prediction of Diffuse Solar Irradiance Using Machine Learning and Multivariable Regression.” Applied Energy, vol. 181. (2016) pp. 367–374. doi:10.1016/j.apenergy.2016.08.093.

[15] Zhaoxuan. L. et al. “A Hierarchical Approach Using Machine Learning Methods in Solar Photovoltaic Energy Production Forecasting.” Energies, vol. 9, no. 12 p. 55 (2016). doi:10.3390/en9010055.

[16] Gala, Y., et al. “Hybrid Machine Learning Forecasting of Solar Radiation Values.”Neurocomputing, vol. 176 pp. 48–59 (2016). doi:10.1016/j.neucom.2015.02.078.

[17] Lauret, P., et al. “A Benchmarking of Machine Learning Techniques for Solar Radiation Forecasting in an Insular Context.” Solar Energy, vol. 112 pp. 446–457 (2015). doi:10.1016/j.solener.2014.12.014.

[18] Kitchenham, B., Charters, S. Guidelines for performing Systematic Literature Reviews in Software Engineering, (2007).

[19] Carvalho, M. M. ; Alves-Souza, S. N. ; CAMPOS, E. G. L. ; de SOUZA, Luiz Sergio . Total Data Quality Management and Total Information Quality Management Applied to Costumer Relationship Management. In: 9th International Conference on Information Management and Engineering, 2017, Barcelona. Proceedings of the 2016 8th International Conference on Information Management and Engineering (2017).

[20] Muntean, M., et al. “Data Mining Methods for Parameters Forecasting of a Small Solar Plant.” Advanced Topics in Optoelectronics, Microelectronics, and Nanotechnologies VIII (2016) doi:10.1117/12.2245896.

[21] Colak, I., et al. “A Data Mining Approach: Analyzing Wind Speed and Insolation Period Data in Turkey for Installations of Wind and Solar Power Plants.” Energy Conversion and Management, vol. 65 pp. 185–197 (2013) doi:10.1016/j.enconman.2012.07.011.

[22] Jimenez-Perez, Pedro F., and Llanos, M. Modeling Daily Profiles of Solar Global Radiation Using Statistical and Data Mining Techniques. Advances in Intelligent Data Analysis XIII Lecture Notes in Computer Science pp. 155–166 (2014). doi:10.1007/978-3-319-12571-8_14.

[23] Assouline, D., Mohajeri, N., Scartezzini, J., Quantifying rooftop photovoltaic solar energy potential: A machine learning approach, Solar Energy. 141 (2017) 278-296. doi:10.1016/j.solener.2016.11.045.

[24] Kausika, Bhavya, et al. “A Big Data approach to the solar PV market design and results of a pilot in the Netherlands.” (2014) doi:10.4229/EUPVSEC20142014-7AV.6.18.

[25] Siyuan, L. et al. “Machine Learning Based Multi-Physical-Model Blending for Enhancing Renewable Energy Forecast - Improvement via Situation Dependent Error Correction.”2015 European Control Conference (ECC). (2015) doi:10.1109/ecc.2015.7330558.

[26] Yuan, L. et al. “A Machine-Learning Approach for Regional Photovoltaic Power Forecasting.”2016 IEEE Power and Energy Society General Meeting (PESGM). (2016) doi:10.1109/pesgm.2016.7741991.

[27] Mayilvahanan, M., and M. Sabitha. “Estimating the Availability of Sunshine Using Data Mining Techniques.” 2013 International Conference on Computer Communication and Informatics. (2013) doi:10.1109/iccci.2013.6466298.

[28] Haupt, S., and Kosovic, B.. “Big Data and Machine Learning for Applied Weather Forecasts: Forecasting Solar Power for Utility Operations.” 2015 IEEE Symposium Series on Computational Intelligence (2015) doi:10.1109/ssci.2015.79.

[29] J. Li, J. Ward, J. Tong, L. Collins, G. Platt, Machine learning for solar irradiance forecasting of photovoltaic system, Renewable Energy. 90 (2016) 542-553. doi:10.1016/j.renene.2015.12.069.

[30] Lauret, P.; David, M.; Tapachès, E. “Machine learning techniques for short term solar forecasting”. 3rd Southern African Solar Energy Conference, South Africa, 11-13 (2015).

[31] M. Şahin, Y. Kaya, M. Uyar, S. Yıldırım, Application of extreme learning machine for estimating solar radiation from satellite data, International Journal of Energy Research. 38 (2013) 205-212. doi:10.1002/er.3030.

[32] Assouline, D., et al. “A machine learning methodology for estimating roof-top photovoltaic solar energy potential in Switzerland.” Proceedings of International

Conference CISBAT 2015 Future Buildings and Districts Sustainability from Nano to Urban Scale, 555-560 (2015).

[33] Cadre, H., et al. “Solar PV Power Forecasting Using Extreme Learning Machine and Information Fusion”. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, (2015).

[34] Martin, R., et al. “Machine Learning Techniques for Daily Solar Energy Prediction and Interpolation Using Numerical Weather Models.” Concurrency and Computation: Practice and Experience, vol. 28, no. 4, pp. 1261–1274, (2015). doi:10.1002/cpe.3631.

[35] Xiaoyan, S. et al. “Solar Radiation Forecast with Machine Learning.” 2016 23rd International Workshop on Active-Matrix Flatpanel Displays and Devices (AM-FPD), (2016) doi:10.1109/am-fpd.2016.7543604.

[36] Aler, R., et al. “A Study of Machine Learning Techniques for Daily Solar Energy Forecasting Using Numerical Weather Models.” Intelligent Distributed Computing VIII Studies in Computational Intelligence pp. 269–278, (2015). doi:10.1007/978-3-319-10422-5_29.

[37] Hassan, M. Khalil, A., Kaseb, S., Kassem, M. Potential of four different machine-learning algorithms in modeling daily global solar radiation, Renewable Energy. 111 52-62 (2017). doi:10.1016/j.renene.2017.03.083.

[38] Loury, Fl., et al., “Data Mining Determination of Sunlight Average Input for Solar Power Plant” International Journal of Computer and Information Engineering Vol:7, No:9, (2013).

[39] Mehra, V. et al. “A Novel Application of Machine Learning Techniques for Activity-Based Load Disaggregation in Rural off-Grid, Isolated Solar Systems.” 2016 IEEE Global Humanitarian Technology Conference (GHTC), (2016) doi:10.1109/ghtc.2016.7857308.

[40] Melzi, F., et al. “Hourly Solar Irradiance Forecasting Based on Machine Learning Models.” 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA) (2016) doi:10.1109/icmla.2016.0078.

[41] Hossain, R. et al. The Combined Effect of Applying Feature Selection and Parameter Optimization on Machine Learning Techniques for Solar Power Prediction, American Journal Of Energy Research. 1 (2013) 7-16. doi:10.12691/ajer-1-1-2.

[42] Senekane, M, Molibeli, B.. “Prediction of Solar Irradiation Using Quantum Support Vector Machine Learning Algorithm.” Smart Grid and Renewable Energy, vol. 07, no. 12 pp. 293–301. (2016), doi:10.4236/sgre.2016.712022.

[43] Jawaid, F., Nazirjunejo, K.. “Predicting Daily Mean Solar Power Using Machine Learning Regression Techniques.” 2016 Sixth International Conference on Innovative Computing Technology (INTECH) (2016) doi:10.1109/intech.2016.7845051.

[44] Haupt, S., and Branko Kosovic. “Variable Generation Power Forecasting as a Big Data Problem.” IEEE Transactions on Sustainable Energy, vol. 8, no. 2, pp. 725–732. (2017) doi:10.1109/tste.2016.2604679.

[45] Zhang, M. Cheng, Y. Liu, D. Li, R. Wu, Short-Term Load Forecasting Based on Big Data Technologies, Applied Mechanics And Materials. 687-691 (2014) 1186-1192. doi:10.4028/www.scientific.net/amm.687-691.1186.

[46] Ovidiu Ivanciuc. “Applications of Support Vector Machines in Chemistry.” In: Reviews in Computational Chemistry, Volume 23, Eds.: K. B. Lipkowitz and T. R. Cundari. Wiley-VCH, Weinheim, pp. 291-400. (2007).

[47] Wang SC. Artificial Neural Network. In: Interdisciplinary Computing in Java Programming. The Springer International Series in Engineering and Computer Science, vol 743. Springer, Boston, MA. (2003).

[48] Hsu, K., Gupta, H. and Sorooshian, S.” Artificial Neural Network Modeling of the Rainfall-Runoff Process”. Water Resources Research, 31(10), pp.2517-2530 (1995).

[49] Huang, G., Huang, G., Song, S. and You, K.. “Trends in extreme learning machines: A review.” Neural Networks, 61, pp.32-48. (2015).

[50] Huang, G., Wang, D. and Lan, Y. “Extreme learning machines: a survey”. International Journal of Machine Learning and Cybernetics, 2(2), pp.107-122. (2011).

[51] Natekin, A. Knoll, A. Gradient boosting machines, a tutorial, Frontiers In Neurorobotics. 7 (2013). doi:10.3389/fnbot.2013.00021.

[52] Breiman, L. Machine Learning 45: 5. (2001) https://doi.org/10.1023/A:1010933404324

[53] Voyant, C. Notton, G. Kalogirou, S. Nivet, M. Paoli, C. Motte, F. et al. Machine learning methods for solar radiation forecasting: A review, Renewable Energy. 105 (2017) 569-582. doi:10.1016/j.renene.2016.12.095.

[54] Mehmet, Y. et al. “A Review of Data Mining and Solar Power Prediction.” 2016 IEEE International Conference on Renewable Energy Research and Applications (ICRERA), (2016) doi:10.1109/icrera.2016.7884507.

[55] Andrew, C. et al. Using machine learning to predict wind turbine power output. Environmental Research Letters. (2013). 8. 024009. 10.1088/1748-9326/8/2/024009.

[56] Heinermann, J., Kramer, O, Machine learning ensembles for wind power prediction, Renewable Energy. 89 (2016) 671-679. doi:10.1016/j.renene.2015.11.073.

[57] Revankar, P. S., Thosar, A. G. and Gandhare, W. Z., "Maximum Power Point Tracking for PV Systems Using MATALAB/SIMULINK," 2010 Second International Conference on Machine Learning and Computing, Bangalore, (2010) pp. 8-11. doi: 10.1109/ICMLC.2010.54

[58] Chia, Y., Lee, L., Shafiabady, N. and Isa, D.A load predictive energy management system for supercapacitor-battery hybrid energy storage system in solar application using the Support Vector Machine. Applied Energy, 137, pp.588-602. (2015).

[59] Simmhan , Y. *et al*., "Cloud-Based Software Platform for Big Data Analytics in Smart Grids," in *Computing in Science & Engineering*, vol. 15, no. 4, pp. 38-47, (2013). doi: 10.1109/MCSE.2013.39

[60] Steinwart, I et al., Support Vector Machines. (2008). *Information Science and Statistics.* doi:10.1007/978-0-387-77242-4

[61] Saha, B., & Srivastava, D. (2014). *Data quality: The other face of Big Data. 2014 IEEE 30th International Conference on Data Engineering.* doi:10.1109/icde.2014.6816764

[62] W. W. Eckerson: Data quality and the bottom line: achieving business success through a commitment to high quality data. *Data Warehousing Institute, 2002*.

[63] Wang, R. Y., & Strong, D. M. (1996) Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems 12(4), pp 5–33.

[64] Cai, L. and Zhu, Y., 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14, p.2. DOI: http://doi.org/10.5334/dsj-2015-002

[65] Woodall, Philip & Borek, Alexander & Gao, Jing & Oberhofer, Martin & Koronios, Andy. (2014). An Investigation of How Data Quality is Affected by Dataset Size in the Context of Big Data Analytics.

[66] Taleb, Ikbal & Serhani, Mohamed & Dssouli, Rachida. (2018). Big Data Quality: A Survey. 10.1109/BigDataCongress.2018.00029.

[67] Batini, C., Rula, A., Scannapieco, M., & Viscusi, G. (2015). *From Data Quality to Big Data Quality. Journal of Database Management, 26(1), 60–82.* doi:10.4018/jdm.2015010103

[68] Rao, D., Gudivada, V. N., & Raghavan, V. V. (2015). Data quality issues in big data. 2015 IEEE International Conference on Big Data (Big Data). doi:10.1109/bigdata.2015.7364065

[69] DAMA, (2013). Defining Data Quality Dimensions. Data Management Association (DAMA)/ UK Working Group. https://is.gd/dama_def_data_quality_dim

[70] Ramasamy, Anandhi & Chowdhury, Soumitra. (2020). BIG DATA QUALITY DIMENSIONS: A SYSTEMATIC LITERATURE REVIEW. 10.4301/S1807-1775202017003.

ANEXO B - REFERÊNCIAS UTILIZADAS NA IMPLEMENTAÇÃO DOS ALGORITMOS PREDIÇÃO E RESULTADOS

[1] IRENA and ADFD (2020), Advancing renewables in developing countries: Progress of projects supported through the IRENA/ADFD Project Facility, International Renewable Energy Agency (IRENA) and Abu Dhabi Fund for Development (ADFD), Abu Dhabi.

[2] IRENA (2019), Renewable capacity statistics 2019, International Renewable Energy Agency (IRENA), Abu Dhabi 2019.

[3] EPE. “O Compromisso do Brasil no Combate às Mudanças Climáticas: Produção e Uso de Energia”. Rio de Janeiro: Empresa de Pesquisa Energética, (2016).

[4] Plano Decenal de Expansão de Energia 2027 / Ministério de Minas e Energia. Empresa de Pesquisa Energética. Brasília: MME/EPE, 2018.

[5] Singh, B., Dwivedi, S., Hussain, I., & Verma, A. K. (2014). Grid integration of solar PV power generating system using QPLL based control algorithm. 2014 6th IEEE Power India International Conference (PIICON). doi:10.1109/34084poweri.2014. 7117785.

[6] Francisco, A. C. C., Vieira, H. E. M., Romano, R. R., & Roveda, S. R. M. M. (2019). Influência de parâmetros meteorológicos na geração de energia em painéis fotovoltaicos: um caso de estudo do smart campus Facens, SP.

[7] Revankar, P. S., Thosar, A. G. and Gandhare, W. Z., "Maximum Power Point Tracking for PV Systems Using MATALAB/SIMULINK," 2010 Second International Conference on Machine Learning and Computing, Bangalore, (2010) pp. 8-11. doi: 10.1109/ICMLC.2010.54.

[8] Chia, Y., Lee, L., Shafiabady, N. and Isa, D.A load predictive energy management system for supercapacitor-battery hybrid energy storage system in solar application using the Support Vector Machine. Applied Energy, 137, pp.588-602. (2015).

[9] Simmhan , Y. et al., "Cloud-Based Software Platform for Big Data Analytics in Smart Grids," in Computing in Science & Engineering, vol. 15, no. 4, pp. 38-47, (2013). doi: 10.1109/MCSE.2013.39.

[10] Aybar-Ruiz, A., Jiménez-Fernández, S., L., Cornejo-Bueno, C., Casanova-Mateo, J. Sanz-Justo, P. Salvador-González et al., A novel Grouping Genetic Algorithm–Extreme Learning Machine approach for global solar radiation prediction from numerical weather models inputs, Solar Energy. 132 (2016) 129-142. doi:10.1016/j.solener.2016.03.015.

[11] Burianek, T., and Stanislav, M. “Solar Irradiance Forecasting Model Based on Extreme Learning Machine.” 2016 IEEE 16th International Conference on Environment and Electrical Engineering (EEEIC) (2016). doi:10.1109/eeeic.2016.7555445.

[12] Shamshirband, S., Mohammadi, K., Yee, P., Petković, D., Mostafaeipour, A.,, A comparative evaluation for identifying the suitability of extreme learning machine to predict horizontal global solar radiation, Renewable And Sustainable Energy Reviews. 52 (2015) 1031-1042. doi:10.1016/j.rser.2015.07.173.

[13] Belaid, S., & Mellit, A. (2016). Prediction of daily and mean monthly global solar radiation using support vector machine in an arid climate. Energy Conversion and Management, 118, 105–118. doi:10.1016/j.enconman.2016.03.082.

[14] Chow, S.K.H.; Lee, E.W.M.; Li, D.H.W. Short-Term Prediction of Photovoltaic Energy Generation by Intelligent Approach. Energy Build. 2012, 55, 660–667.

[15] Yadav, A. K., & Chandel, S. S. (2014). Solar radiation prediction using Artificial Neural Network techniques: A review. Renewable and Sustainable Energy Reviews, 33, 772–781. doi:10.1016/j.rser.2013.08.055.

[16] Khatib, T., Mohamed, A., Sopian, K., & Mahmoud, M. (2012). Assessment of Artificial Neural Networks for Hourly Solar Radiation Prediction. International Journal of Photoenergy, 2012, 1–7. doi:10.1155/2012/946890.

[17] Melzi, F. N., Touati, T., Same, A., & Oukhellou, L. (2016). Hourly Solar Irradiance Forecasting Based on Machine Learning Models. 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA). doi:10.1109/icmla.2016.0078.

[18] Al-Dahidi, Sameer; Ayadi, Osama; Adeeb, Jehad; Alrbai, Mohammad; Qawasmeh, Bashar R. 2018. "Extreme Learning Machines for Solar Photovoltaic Power Predictions" Energies 11, no. 10: 2725. https://doi.org/10.3390/en11102725.

[19] Huang, G.-B.; Zhu, Q.; Siew, C. Extreme Learning Machine: Theory and Applications. Neurocomputing 2006, 70, 489–501.

[20] de Freitas Viscondi, G., & Alves-Souza, S. N. (2019). A Systematic Literature Review on big data for solar photovoltaic electricity generation forecasting. Sustainable Energy Technologies and Assessments, 31(November 2018), 54–63. https://doi.org/10.1016/j.seta.2018.11.008.

[21] Mahesh, Batta. (2019). Machine Learning Algorithms - A Review. 10.21275/ART20203995.

[22] IDC (2021). Worldwide Global DataSphere Forecast, 2021–2025: The World Keeps Creating More Data — Now, What Do We Do with It All? International Data Corporation (IDC), Massachusetts, 2021.

[23] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. Proceedings of the Fifth Annual Workshop on Computational Learning Theory - COLT ’92. doi:10.1145/130385.130401.

[24] Lorena, A. C. L., Carvalho, A. C. P. L. F (2007). Uma introdução às Support Vector Machines. Revista de Informática Aplicada e Teórica.

[25] Steinwart, I et al., Support Vector Machines. (2008). Information Science and Statistics. doi:10.1007/978-0-387-77242-4.

[26] Noble, W. S. (2006). What is a support vector machine? Nature Biotechnology, 24(12), 1565–1567. doi:10.1038/nbt1206-1565.

[27] Namba, M., & Zhang, Z. (n.d.). Cellular Neural Network for Associative Memory and Its Application to Braille Image Recognition. The 2006 IEEE International Joint Conference on Neural Network Proceedings. doi:10.1109/ijcnn.2006.1716416.

[28] Odom, M. D., & Sharda, R. (1990). A neural network model for bankruptcy prediction. 1990 IJCNN International Joint Conference on Neural Networks. doi:10.1109/ijcnn.1990.137710.

[29] Cherry, K. M., & Qian, L. (2018). Scaling up molecular pattern recognition with DNA-based winner-take-all neural networks. Nature, 559(7714), 370–376. doi:10.1038/s41586-018-0289-6.

[30] Wang, S.-C. (2003). Artificial Neural Network. Interdisciplinary Computing in Java Programming, 81–100. doi:10.1007/978-1-4615-0377-4_5.

[31] B.D.Ripley, Pattern Recognition and Neural Networks, Cambridge University Press, Cambridge (1996).

[32] Tang, J., Deng, C., & Huang, G.-B. (2016). Extreme Learning Machine for Multilayer Perceptron. IEEE Transactions on Neural Networks and Learning Systems, 27(4), 809–821. doi:10.1109/tnnls.2015.2424995.

[33] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, Extreme learning machine for regression and multiclass classification, IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 42, no. 2, pp. 513–529, Apr. 2012.

[34] Huang, G.-B. (2014). An Insight into Extreme Learning Machines: Random Neurons, Random Features and Kernels. Cognitive Computation, 6(3), 376–390. doi:10.1007/s12559-014-9255-2.

[35] G.-B. Huang, L. Chen, and C.-K. Siew, Universal approximation using incremental constructive feedforward networks with random hidden nodes, IEEE Trans. Neural Netw., vol. 17, no. 4, pp. 879–892, Jul. 2006.

[36] São Paulo será 6ª cidade mais rica do mundo até 2025. Price Waterhouse & Coopers e BBC Brasil. November 2009.

[37] LAERD Statistics. (2021). Available at: https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php. Access: July 25th of 2021.