**UNIVERSIDADE FEDERAL DE JUIZ DE FORA**
**INSTITUTO DE CIÊNCIAS EXATAS/FACULDADE DE ENGENHARIA**
**PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL**

**Lucas Pereira Verneck da Silva**

# Desenvolvimento de um modelo para a estimação da carga de radiação solar com base em variáveis climáticas

Juiz de Fora

2021

**Lucas Pereira Verneck da Silva**

# Desenvolvimento de um modelo para a estimação da carga de radiação solar com base em variáveis climáticas

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Modelagem Computacional. Área de concentração: Modelagem Computacional

Orientador: Prof. Dr. Leonardo Goliatt da Fonseca

Juiz de Fora

2021

Ficha catalográfica elaborada através do Modelo Latex do CDC da UFJF
com os dados fornecidos pelo(a) autor(a)

Silva, Lucas.
Desenvolvimento de um modelo para a estimação da carga de radiação solar com base em variáveis climáticas / Lucas Pereira Verneck da Silva. – 2021.
78 f. : il.

Orientador: Leonardo Goliatt da Fonseca
Dissertação (Mestrado) – Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas/Faculdade de Engenharia. Programa de Pós-Graduação em Modelagem Computacional, 2021.

1. energia solar. 2. radiação solar. 3. previsão. I. Goliatt, Leonardo, orient. II. Título.

**Lucas Pereira Verneck da Silva**

**Desenvolvimento de um modelo para a estimação da carga de radiação solar com base em variáveis climáticas**

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Modelagem Computacional. Área de concentração: Modelagem Computacional

Aprovada em 16 de dezembro de 2021.

BANCA EXAMINADORA

**Prof(a)Dr(a) . Leonardo Goliatt da Fonseca** - Orientador
Universidade Federal de Juiz de Fora

**Prof(a)Dr(a) . Luciana Conceição Dias Campos**
Universidade Federal de Juiz de Fora

**Prof(a)Dr(a) . Eliane da Silva Christo**
Universidade Federal Fluminense

Juiz de Fora, 10/12/2021.

Documento assinado eletronicamente por **Luciana Conceicao Dias Campos**, **Professor(a)**, em 16/12/2021, às 17:01, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Leonardo Goliatt da Fonseca**, **Professor(a)**, em 16/12/2021, às 17:02, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **ELIANE DA SILVA CHRISTO**, **Usuário Externo**, em 16/12/2021, às 17:12, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **0613040** e o código CRC **9AB6F655**.

Dedico este trabalho à minha família, pelo apoio incondicional.

## AGRADECIMENTOS

Primeiramente a Deus por me conceder forças durante toda a caminhada. Aos meus pais Manoel e Vera, a meu irmão Felipe e à minha namorada Bárbara por todo apoio e encorajamento.

A todos os colegas do PGMC pela companhia e pelos auxílios prestados durante este período de mestrado. Aos professores do Programa de Pós-Graduação em Modelagem Computacional por seus ensinamentos, em especial ao professor e orientador Leonardo Goliatt pela orientação, incentivo e principalmente, pela paciência durante o desenvolvimento do trabalho.

Por fim, agradeço à Universidade Federal de Juiz de Fora, o PGMC e à agência de fomento CAPES pelo suporte financeiro oferecido durante o período de mestrado.

"The impediment to action advances action. What stands in the way becomes the way." - Marcus Aurelius.

## RESUMO

Dados relacionados à incidência de radiação solar (RS) em uma determinada região desempenham um papel crucial no projeto, modelagem e na operação de sistemas de conversão de energia solar. Além disso, este tipo de informação ajuda nas futuras políticas governamentais de investimento em energia. Contribui para inúmeras outras áreas de estudo e diversas aplicações, como na criação de chaminés térmicas, análise de conforto térmico em edifícios, modelos de crescimento de safras, entre outros. Porém, esses dados não são medidos para todas as regiões devido à falta de equipamentos adequados para a medição da RS nas estações meteorológicas e também aos custos envolvidos na atividade. Embora diferentes métodos alternativos para obtenção da RS tenham sido propostos e utilizados na literatura, ultimamente, a maioria das pesquisas tem focado seus esforços na exploração e utilização de algoritmos de aprendizado de máquina para resolver o problema. Seguindo essa linha, o presente trabalho avalia a aplicação de uma regressão linear com fator de regularização do tipo L2 combinado com uma estratégia de expansão dos dados de entrada para estimar a intensidade da radiação solar diária incidente. O ajuste e avaliação dos modelos são realizados em oito conjuntos de dados diferentes contendo variáveis climáticas facilmente acessíveis pelas estações meteorológicas. Cada conjunto refere-se à uma região diferente, distribuídas ao longo do território da Burkina Faso, país localizado na África subsariana. Para a seleção dos parâmetros do método, o algoritmo de busca exaustiva foi aplicado para encontrar o conjunto de hiperparâmetros que reforçam as capacidades preditivas dos modelos. A avaliação e comparação entre os modelos são realizadas de acordo com a raiz quadrada do erro-médio, erro médio absoluto, coeficiente de determinação e a variação contabilizada. Os resultados obtidos validam a estratégia adotada pelo trabalho, evidenciando o impacto positivo no modelo gerado pela inclusão do termo de regularização. Os experimentos sugerem um melhor desempenho em relação aos valores encontrados na literatura produzidos por modelos de maior complexidade, apresentando redução nos valores de erro em até 50% para algumas estações.



Palavras-chave: Energia solar. Radiação solar. Energia fotovoltaica. Estimação.

## ABSTRACT

Data related to solar radiation (SR) incidence in a given region play a crucial role in the design, modeling, and operation of solar energy conversion systems. Also, this type of information helps in future energy investment policies governments. It contributes to numerous other areas of study and several applications, such as creating thermal chimneys, thermal comfort analysis in buildings, crop growth models, and others. However, these data are not measured for all regions due to the lack of appropriate equipment for the measurement of SR in meteorological stations and the costs involved in the activity. Although different alternative methods for obtaining SR have been proposed and used in the literature, lately, most research has focused its efforts on exploring and using machine learning algorithms to address the problem. Following this line, the present work evaluates applying a linear regression with a regularization factor of type L2 combined with an expansion strategy of the input data to estimate the intensity of incident daily solar radiation. The models' adjustment and evaluation are performed on eight different datasets containing climatic variables easily accessible by the weather stations. Each set refers to the other region distributed throughout Burkina Faso, a country located in Sub-Saharan Africa. For the selection of method parameters, the exhaustive search algorithm was applied to find the set of hyperparameters that reinforce the models' predictive capabilities. The evaluation and comparison between the models are performed according to root mean square error, mean absolute error, coefficient of determination and variance accounted for. The results obtained validate the work's strategy, showing the positive impact on the model generated by the inclusion of the regularization term. The experiments suggest a better performance compared to the values found in the literature produced by models with greater complexity, showing a reduction in the error values up to 50% for some stations.



Keywords: Solar energy. Solar radiation. Photovoltaic energy. Estimation.

## LISTA DE ILUSTRAÇÕES

## LISTA DE TABELAS

## LISTA DE ABREVIATURAS E SIGLAS

| | |
|---|---|
| ANFIS | Adaptive neuro fuzzy inference system |
| ANN | Artificial neural network |
| DE | Differential evolution |
| DL | Deep learning |
| ELM | Extreme learning machine |
| GWh | Gigawatt-hora |
| KWh/m2 | Quilowatt-hora por metro quadrado |
| MAE | Mean absolute error |
| MBE | Mean bias error |
| MJ/m2 | Megajoules por metro quadrado |
| ML | Machine learning |
| Mtep | Milhões de tonelada equivalente de petróleo |
| PSO | Particle swarm optimization |
| R2 | Coeficiente de determinação |
| RS | Radiação solar |
| SOM | Self-organizing maps |
| SVM | Support Vector Machine |
| SVR | Support Vector Regression |
| VAF | Variance accounted for |

**SUMÁRIO**

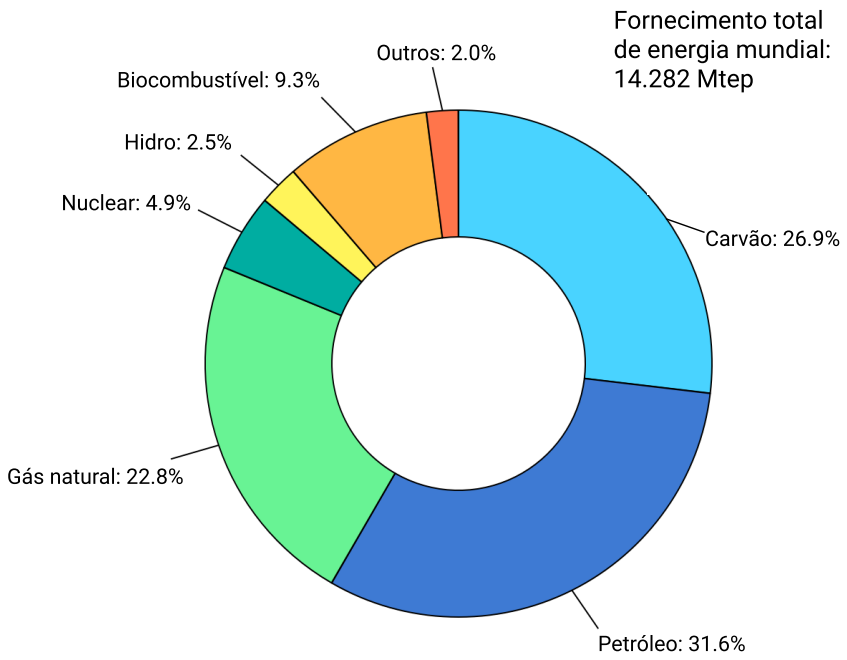
# 1 INTRODUÇÃO

## 1.1 APRESENTAÇÃO DO TEMA E CONTEXTUALIZAÇÂO DO PROBLEMA

Atualmente, o cenário mundial no que diz respeito à oferta total de energia primária mostra que a maior parte da demanda ainda é proporcionada através das fontes de energia não renováveis, baseadas na queima de combustíveis fósseis [1]. Esse tipo de fonte energética gera como subproduto do processo de queima, gases que são nocivos ao meio ambiente, como o Dióxido de Carbono (CO2), Óxido Nitroso (N2O), Metano (CH4), Clorofluorcarbonetos (CFCs), Hidrofluorcarbonetos (HFCs), Perfluorcarbonetos (PFCs) e o Hexafluoreto de Enxofre (SF6). Conhecidos como gases do efeito estufa (GEE) [2], essas substâncias quando presentes na atmosfera, absorvem parte da radiação infravermelha, que é emitida principalmente pela superfície terrestre, e impedem sua saída para o espaço, ocasionando assim no aquecimento da Terra. O efeito estufa é um fenômeno natural e essencial para a manutenção da vida no planeta, porém, devido ao grande aumento na emissão dos GEE ao longo dos anos, esse evento vem se potencializando gerando assim consequências danosas tais como o aumento do nível do mar, problemas de saúde provocados pela mudança climática [3], escassez de recursos naturais e sérios danos aos mais diversos ecossistemas do planeta.

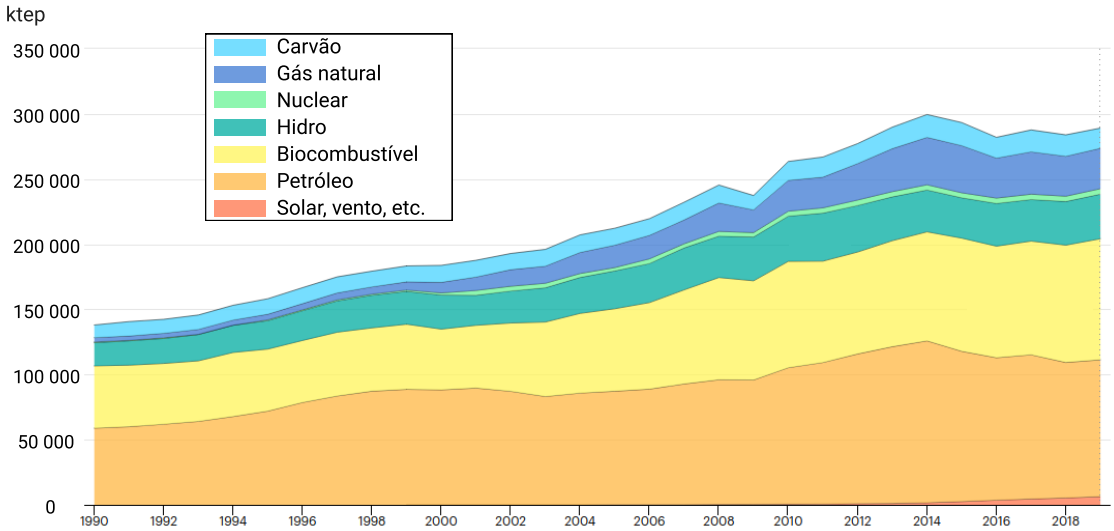
Segundo os dados disponibilizados pela IEA (Agência Internacional de Energia) [4] está claro que apesar das fontes de energia renovável estarem em pauta no mundo todo, ainda há muito trabalho a ser feito para que essas alternativas sustentáveis alcancem um patamar de maior representatividade. Como pode ser observado na Figura 1, no cenário mundial $81,3\%$ do fornecimento total de energia é resultante da queima de combustíveis fósseis. Em contraste ao cenário global o Brasil possui um perfil de fornecimento energético consideravelmente mais limpo, sendo apenas $52,4\%$ do suprimento proveniente de fontes não renováveis, Figura 2. Essa característica do país é um resultado do investimento em fontes renováveis nos últimos anos como a energia solar, eólica, biocombustíveis e principalmente devido ao amplo uso dos recursos hídricos, que por si só é responsável por $11,8\%$ da produção total.

Figura 1 - Participação global do fornecimento total de energia dividido por fonte no ano de 2018.




Fonte: Adaptado de IEA [4].

Figura 2 - Fornecimento total de energia por fonte no Brasil durante o período de 1990 até 2019.



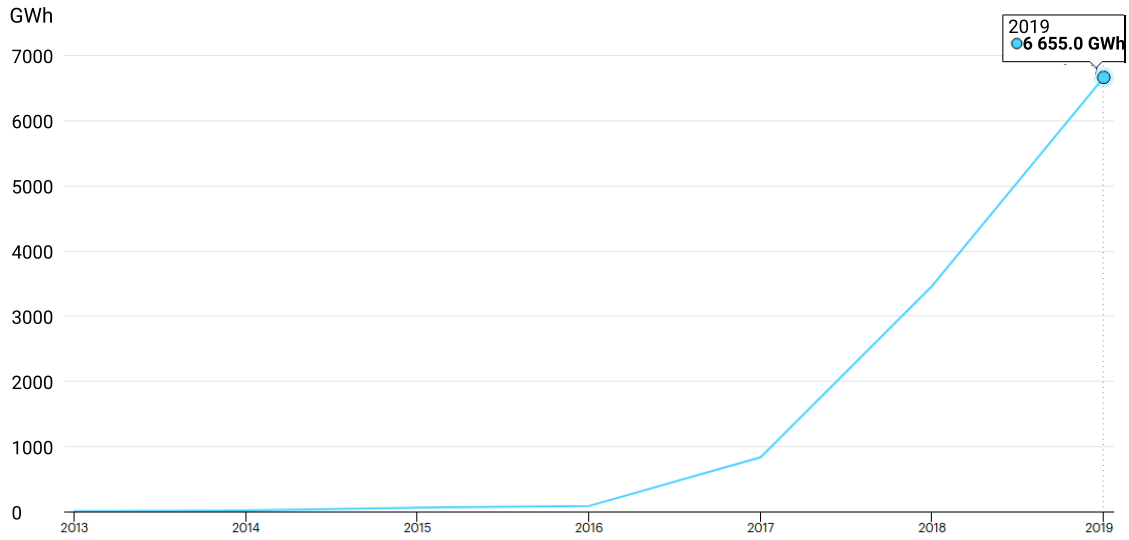

Fonte: Adaptado de IEA [4].

Em meio à essa predominância da matriz energética global por parte das fontes não renováveis e uma vez conhecido os problemas atrelados à esse tipo de produção de energia, o mundo vem mudando em busca da transição para novos meios de produção que sejam sustentáveis ao longo prazo. Nos últimos anos, a popularidade das pesquisas em sistemas de energia renovável tem aumentado significativamente em diferentes campos acadêmicos, envolvendo o empenho de diversas organizações, comunidades mundiais, líderes e gestores de energia [5–7]. Esses sistemas são extremamente benéficos porque ajudam a suprir as necessidades de energia de maneira ambientalmente correta. As principais alternativas renováveis incluem a energia hidráulica, biomassa, geotérmica, eólica e a solar. Em meio às vantagens e desvantagens de cada um desses sistemas de produção, a energia eólica e a solar são consideradas como as alternativas mais poderosas aos sistemas tradicionais de energia para o futuro imediato.

Entre as fontes de energia renováveis, aquela proveniente da luz solar é uma das fontes de energia de maior abundância, limpa, confiável, infinita, amplamente acessível e independente [8]. A energia solar é obtida através de dispositivos comumente modulados em placas produzidas em um material semicondutor. Quando as partículas de luz solar (fótons) incidem sobre a célula fotovoltaica, os elétrons do material semicondutor entram em movimento gerando eletricidade. Entretanto, o desempenho desse tipo de sistema de produção de energia depende intrinsecamente de fatores climáticos como a radiação solar, temperatura, luminosidade, umidade do ar, índices de evaporação e até mesmo da velocidade do vento. Sendo o fator de maior importância a radiação solar, comumente utilizado como variável para a predição da produção dos sistemas de energia solar [9]. Devido à esses fatores, a radiação solar não é distribuída igualmente na mesma quantidade em todos os lugares, ela depende da variabilidade da condição do tempo e também da variabilidade espacial.

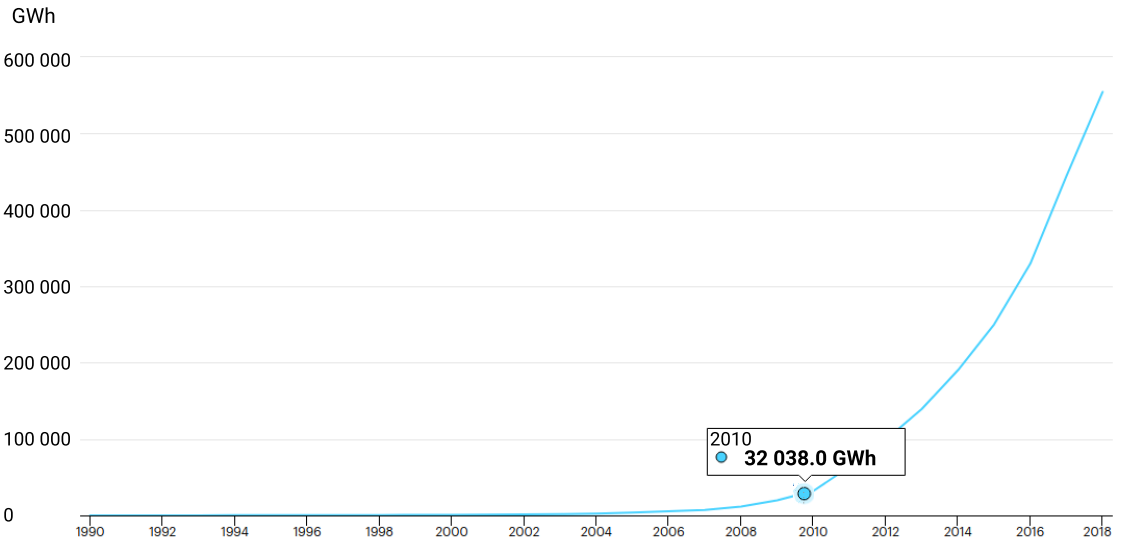
Com a recente redução dos custos dos painéis fotovoltaicos, a acessibilidade e implementação desse tipo de sistema de produção tem se tornado cada vez mais frequente. Ocasionando em um crescimento exponencial da produção de energia tanto no Brasil quanto em todo o mundo por meio desse processo ao longo dos últimos dez anos. Em 2016 a produção total do Brasil por meio da conversão fotovoltaica era de 85 $GWh$, já no ano de 2019 esse número apresentou um expressivo aumento relativo de 7829% totalizando uma produção de 6655 $GWh$ no ano, veja na Figura 3. Enquanto isso, no cenário mundial o aumento da produção cresceu cerca de 1730% entre os anos de 2010 e 2018, presente na Figura 4.

Figura 3 - Geração de energia no Brasil proveniente de operação fotovoltaica durante o período de 2013 a 2019.




Fonte: Adaptado de IEA [4].

Figura 4 - Geração de energia no mundo proveniente de operação fotovoltaica durante o período de 1990 a 2018.

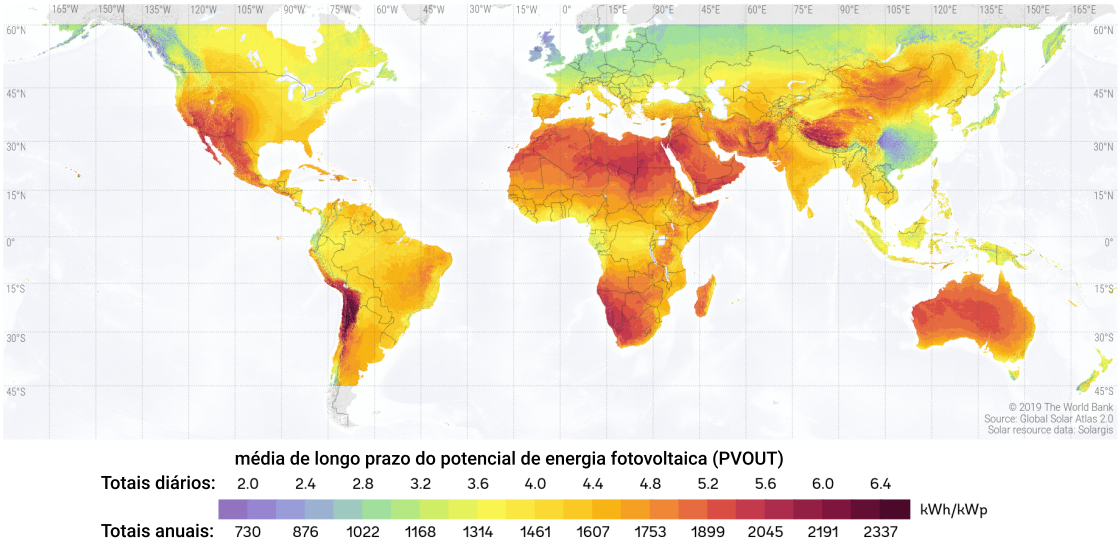



Fonte: Adaptado de IEA [4].

Em conformidade com a variabilidade climática e geográfica o potencial fotovoltaico das diferentes regiões do globo pode ser visto na Figura 5. A capacidade de cada região em captar a radiação solar está diretamente relacionada com a sua proximidade com a linha

do Equador, quanto mais próximo, mais radiação solar pode ser potencialmente captada para a produção de energia. Outra característica importante, é que os países que possuem um alto potencial tendem a apresentar baixa sazonalidade na produção solar fotovoltaica, o que significa que o recurso é relativamente constante entre os diferentes meses do ano.

Figura 5 - Mapa global do potencial de energia fotovoltaica.




Fonte: Adaptado de Global Solar Atlas 2.0 [10].

## 1.2 JUSTIFICATIVA

Dispor do conhecimento da quantidade de radiação solar que incide sobre uma determinada região é uma atividade de extrema importância para os modernos sistemas integrados de energia, já que esses podem operar continuamente por longos períodos [11], além disso, esses dados conseguem auxiliar na gestão da escassez de energia elétrica ocasionada pela própria natureza estocástica que a radiação solar apresenta [12]. Esse tipo de informação desempenha um papel fundamental para integrar fontes de energia solar em uma rede elétrica, dando suporte aos chamados *smart grids* e também às concessionárias de energia, auxiliando na tomada de decisão à respeito do balanceamento de carga da rede e na comutação de transmissões. De uma forma geral, a previsão da carga de radiação solar pode assistir as concessionárias de energia de forma a contribuir para a redução dos custos operacionais, otimização da operação e na melhora da qualidade e confiabilidade da energia fornecida à rede [13].

Devido a sua natureza volátil, conhecer e prever a quantidade de RS torna-se uma atividade de grande significância. Essa prática provem notáveis vantagens em diferentes setores, e não somente contribui para aumentar eficiência operacional dos sistemas fornecedores de energia, mas também contribui para a detecção de locais mais

apropriados para a instalação de sistemas solares, no sentido de se atingir uma maior performance de produção de energia [14]. Olhando do ponto de vista macro, os dados de radiação solar apresentam uma oportunidade para se fazer ou revisar planos futuros de diferentes e importantes setores como por exemplo o setor industrial, a agricultura e até mesmo o setor de turismo.

A RS é um componente crucial não só para o projeto de sistemas solares mas também está presente e desempenha papel importante em diversos outros estudos e em diferentes tipos de aplicações envolvendo energia solar. Alguns exemplos principais são, a análise de carga e conforto térmico em edifícios [15], desenvolvimento de modelos de previsão para o fluxo de ar de chaminés térmicas [16], análise do impacto da intensidade de radiação solar na produtividade de plantações [17] e na qualidade da produção da rizicultura [18], entre muitas outras aplicações importantes [19].

Apesar das inúmeras vantagens e diferentes opções de aplicação dos dados de RS, mensurá-los nem sempre pode ser uma tarefa fácil e direta para cada região. O principal motivo para isso se deve a custos associados aos processos de instalação e manutenção, também aos requisitos de calibração dos dispositivos de medição da radiação [11, 20]. Adicionalmente, na maioria dos casos a região de instalação das plantas de energia solar são distantes dos locais em que a RS é medida, geralmente nas estações meteorológicas.

Diversas regiões, apesar de possuírem uma disposição geográfica favorável para a operação de produção de energia utilizando a luz do sol, essas não possuem incentivos e investimentos necessários para o desenvolvimento e instalação da infraestrutura necessária. Especialmente em locais que apresentam um cenário como esse, é que se faz a compreensão e quantificação da radiação solar incidente se tornar uma atividade ainda mais relevante, saber e prever as respostas às questões de onde, quando e quanta radiação solar incide em determinado local.

Certamente, a melhor e mais apropriada forma de se obter os dados referente à RS é através da utilização do dispositivo radiométrico adequado para realizar a mensuração de forma direta na localidade. Contudo, em função dos problemas pertinentes à esse método, a maioria dos países da África e Ásia possuem dados de RS escassos [21]. Ainda, as localidades favorecidas por estações que conseguem aferir essa informação, estão situadas nas cidades e em grandes centros urbanos, negligenciando as áreas rurais, onde a crise energética se mostra mais proeminente [11].

Burkina Faso é um país exemplo que se encontra imerso nesse cenário, apesar de estar situado na zona equatorial do globo, possuindo um alto potencial para a geração de energia elétrica através da conversão solar, o mesmo possui muitas estações meteorológicas incapazes de inferir dados à respeito da radiação solar que incide na região [22]. E até mesmo os locais em que ainda há dados prontamente disponíveis, os mesmos sofrem deficiência devido à calibração inadequada do equipamento responsável pela mensuração.

Com o intuito de se obter novas formas e métodos de se mensurar a intensidade de RS, inicialmente, foram desenvolvidas inúmeras equações empíricas para tentar realizar a previsão do recurso, tais modelos empíricos eram frequentemente usados para correlacionar a RS com diferentes parâmetros climáticos e com as coordenadas geográficas [23]. Entretanto, o nível de precisão desses modelos empíricos eram questionáveis. Outra forma alternativa de se gerar informação à respeito da radiação solar é através dos modelos numéricos de previsão do tempo, pertencentes à classe de previsões probabilísticas, que em contraste aos modelos determinísticos oferecem informações adicionais sobre a incerteza associada a atividade de previsão do tempo [24]. Entretanto, a aplicação de modelos numéricos envolve simulações com uso de equações matemáticas complexas, resultando em valores simulados que podem diferir daqueles fisicamente observados na superfície do local.

Com o desenvolvimento da ciência e da tecnologia, na tentativa de contornar as deficiências que os métodos numéricos e empíricos apresentam, considerada atenção vem sendo dada à abordagens de previsão de radiação solar baseadas em dados. A aplicação de inteligência artificial e técnicas de aprendizado de máquina para tratar esse tipo de problema vem sendo intensamente exploradas [25]. Diversas técnicas vem sendo empregadas na tentativa de modelar os dados de RS, uso de máquina de vetor suporte para estimar a radiação solar global considerando efeitos de névoa e neblina [26], emprego de redes neurais para a estimação da RS [27], k-ésimo vizinho mais próximo e dentre outros procedimentos [28].

Apesar do grande número de implementações e estudos relacionados da área empregarem os mais diversos modelos complexos e fazerem uso de hibridização para alcançar maior acurácia [11], as redes neurais se destacam por ser um método com grande capacidade de correlação em diferentes campos de pesquisa, sendo responsável pelos resultados de previsão com maior performance em se tratando de dados de RS [29]. Grande parte dessas modelagens possui custo computacional elevado além de faltar com a presença de interpretabilidade do modelo sobre os dados utilizados.

Como já conhecido, possuir acesso à dados de radiação solar de uma região é vital à inúmeras atividades. Devido as adversidades e complicações associadas ao processo de obtenção desse tipo de informação, faz-se necessário a existência de métodos e ferramentas simples, de baixo custo e de fácil acesso para realizar a estimação da quantidade de RS incidente.

## 1.3 OBJETIVOS GERAIS E ESPECÍFICOS

O presente trabalho possui como objetivo principal a aplicação de ferramentas de aprendizado de máquina para desenvolver e determinar, para os modelos criados, a capacidade de estimação da radiação solar horizontal que incide sobre a superfície na Burkina Faso. Para alcançar isso, dados de oito diferentes estações meteorológicas

distribuídas ao longo do país são utilizados para quantificar o impacto de variáveis climáticas na capacidade de estimação do modelo. Oito modelos diferentes serão criados na tentativa de modelar os dados referentes à cada uma das estações consideradas no estudo, com base nas variáveis de umidade máxima e mínima do ar, temperatura máxima e mínima, velocidade do vento, déficit de pressão de vapor e evaporação. Os principais objetivos específicos que são utilizados para orientar o desenvolvimento do trabalho são:

- Explorar o comportamento dos modelos diante do aumento da complexidade da base de dados por meio da geração de novas variáveis com base nas pré-existentes;
- Verificar o impacto da adição e uso do termo de regularização no tratamento do problema estudado;
- Definir o melhor conjunto de hiperparâmetros que maximize o desempenho da estimação de cada modelo referente a cada uma das oito estações meteorológicas;
- Validar e comparar os resultados obtidos nesse trabalho com os dados contidos na literatura;
- Investigar a influência das variáveis sobre a performance dos modelos por meio de seleção e redução de seus números;

## 1.4 ESTRUTURA DO TRABALHO

Este trabalho está dividido em cinco capítulos: Introdução, Revisão da Literatura, Materiais e Métodos, Resultados e Discussão e por fim Conclusões e Trabalhos Futuros.

No Capítulo 2 é apresentado e discutido diversas aplicações e usos da radiação do sol para diferentes contextos, além de apresentar alguns trabalhos semelhantes e os tipos de abordagens que veem sendo empregadas nesse tipo de tarefa.

No Capítulo 3 todo os dados de insumo são apresentados, explicados e discutidos. As metodologias e técnicas utilizadas no presente trabalho são introduzidas e brevemente explicadas de acordo com a formulação teórica e também de acordo com o uso das mesmas.

Já no Capítulo 4, os principais resultados das modelagens são exibidos, discutidos e explicados. Esses, são apresentados por meio de métricas de desempenho, gráficos e tabelas comparativas.

Por fim, no Capítulo 5 há o encerramento do trabalho, depois de uma breve revisão de tudo que foi abordado. As principais contribuições do projeto são ressaltadas assim como a checagem do alcance dos objetivos previamente definidos e encerrando com as sugestões para possíveis continuações em um trabalho futuro.

# 2 REVISÃO DA LITERATURA

Há tempos o mundo vem convergindo para um cenário onde de um lado há a crescente demanda por energia e do lado oposto existe a necessidade de alcançar uma sociedade livre do excesso de carbono, remediando os graves problemas causados pelo aquecimento global. Esse contexto é fruto do acelerado desenvolvimento econômico e social nas últimas décadas, e vem sendo motivo de fomento para atrair a atenção de diversos pesquisadores e outras partes interessadas, na busca pelo estudo e aplicação de fontes de energia alternativas não convencionais. Dentre essas fontes alternativas está em destaque a energia solar, a qual aproveita de forma direta a radiação do sol que incide na superfície da terra para gerar energia elétrica no caso de sistemas de geração fotovoltaicos, e para gerar energia térmica no caso de plantas de aquecimento solar.

A capacidade de geração de energia tanto no caso de plantas fotovoltaicas quanto nas de aquecimento solar, está diretamente ligado à intensidade de radiação solar que incide sobre o local de captação. Essa, por si só é dependente de diversos outros fatores, como a localização geográfica e a variação climática da região. Dados radiométricos referentes à incidência solar é um recurso de grande importância que assiste não só na criação e operação de sistemas de conversão de energia solar, mas que também é considerado essencial para o desenvolvimento de pesquisa em diversas áreas do conhecimento.

A associação de diversas áreas de sistemas da informação, ciências sociais e ciência energética tem gerado uma nova frente de pesquisas que visam melhorar a forma e a eficiência com que a energia é produzida, conservada, distribuída e até mesmo consumida. Grande parte dos trabalhos em torno da gestão de recursos energéticos possui sua base fundamentada no uso intensivo de diversos tipos de dados tanto para a exploração de insights quanto para auxílio da tomada de decisão [30–32].

Uma vez conhecida a relevância dos dados de RS, inúmeros estudos vêm sendo realizados ao longo do tempo, propondo diversas metodologias para a obtenção desse tipo de informação. Inicialmente a quantificação de RS era obtida pelo uso intensivo de diversas equações através de uma abordagem empírica, posteriormente apareceu o uso simulações numéricas para a previsão do tempo. Com o advento dos computadores e com sua crescente capacidade, os métodos de estimação e previsão baseados em dados vêm sendo amplamente utilizados [33]. Sendo os mais recentes fazendo o uso de algoritmos de aprendizado de máquina e inteligência artificial.

## 2.1 UTILIZAÇÃO DA RADIAÇÃO SOLAR

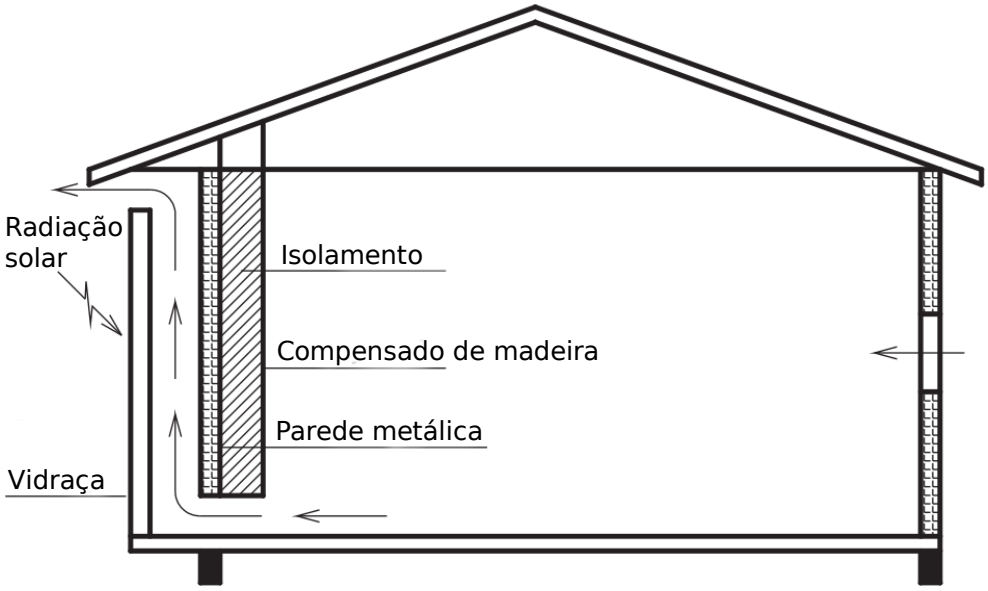
Exercendo grande importância, as informações de incidência solar assumem um papel fundamental em muitas outras diversas aplicações. O nível dessa radiação influência não somente iniciativas criadas pelo homem mas como também inúmeros ecossistemas no

planeta, incluindo projetos arquitetônicos, modelos de crescimento de safra, estimativas de evapotranspiração no projeto de sistemas de irrigação dentre outros [19].

Frequentemente modelos amplamente empregados para a avaliação de ambientes térmicos de padrões internacionais consideram efeitos de apenas seis fatores: temperatura do ar, umidade do ar, velocidade do ar, temperatura média radiante, taxa metabólica humana e o isolamento das roupas. Entretanto essa abordagem é baseada em estudos laboratoriais que fogem ao cenário de um ambiente real. Portando, muitos estudos já demonstraram que o conforto térmico é afetado por outros fatores além desses seis. Devido ao fato de ser cada vez mais comum o emprego de muros de cortina em edifícios modernos, trabalhos mostram a importância de se considerar o efeito da radiação solar para análise do conforto térmico em edificações. Em [15] o autor sugere um novo indicador de conforto térmico, baseado no já existente *predicted mean vote* (PMV) [34], que leva em consideração a radiação solar incidente na edificação para uma melhor estimação, principalmente para os casos com muita incidência solar.

Dentre as muitas formas de se aproveitar a energia proveniente da luz solar, e ainda dentro da área de estudos do conforto térmico em edifícios, existe a aplicação de chaminés solares. Comumente conhecidas pelo nome de chaminé térmica, esse tipo de instalação é uma maneira simples de melhorar a ventilação natural dos ambientes através da convecção do ar aquecido por energia solar de forma passiva Figura 6. Um projeto adequado de uma chaminé térmica requer conhecimento à respeito do fluxo de ar gerado de acordo com a intensidade da radiação solar incidente no edifício. Pesquisas sugerem modelos analíticos para previsão do fluxo de ar que consideram variáveis como temperatura e densidade do ar, já em [16], o trabalho propõem um modelo para se obter a taxa de fluxo de ar de uma chaminé solar, considerando como entradas a incidência de radiação solar e as configurações da chaminé.

Figura 6 - Representação esquemática do funcionamento de uma chaminé térmica passiva para ventilação natural do ambiente.




Fonte: Adaptado de X.Q. Zhai et al. [35].

O setor da agricultura, considerada como uma das principais atividades primárias da economia também tem sido impactado com as mudanças climáticas nas últimas décadas. O crescimento e a produtividade das plantações enfrentam um cenário preocupante. Diante dessa circunstância, consideráveis esforços tem sido empregado na pesquisa e desenvolvimento de soluções inovadoras para contornar tais desafios que o setor enfrenta. Estudos tem levado em consideração a radiação solar para identificar seu impacto na produtividade das culturas. Em [17], usando modelos de ecossistemas o autor investiga o potencial de se usar o tempo de duração da luz solar ao longo do dia como forma de estimativa da radiação solar, bem como o impacto da mesma sobre a *Gross primary productivity* (GPP), que é a taxa na qual a energia solar é capturada em moléculas de açúcar durante a fotossíntese. A interpretação sobre como condições ideais de temperatura e radiação solar para diferentes estágios fenológicos do arroz contribui para o manejo do cultivo e da criação de modelos da cultura é apresentado em [18]. O trabalho faz uso desses dois principais parâmetros para avaliar a resposta em função da qualidade e quantidade da produção de arroz no vale do rio Yangtze na China.

## 2.2 MÉTODOS NÚMERICOS DE PREVISÃO DO TEMPO

Embora a primeira tentativa de uso da previsão numérica do tempo ocorresse na década de 1920, somente mais tarde com o advento da simulação por computador na década de 1950 que as previsões numéricas do tempo produziram resultados realistas. Os

métodos numéricos de previsão do tempo, no inglês *Numerical Weather Prediction* (NWP), são muito utilizados nos dias atuais pelos serviços meteorológicos nacionais e diversos institutos de pesquisa em torno de todo o mundo. Esses modelos, frequentemente consistem no uso de diversos esquemas numéricos que parametrizam diferentes processos físicos, como por exemplo a interação superfície-atmosfera da Terra, turbulência, microfísica de nuvens, processo de convecção e radiação [36]. No entanto, até recentemente a radiação na superfície terrestre não era fornecida como saída do modelo. Até então esse tipo de informação não possuía a devida importância quando comparada em relação à temperatura, vento, umidade ou a precipitação.

Existem um grande número de diferentes modelos para a previsão do tempo baseado em NWP, e em sua maioria eles basicamente resolvem equações fundamentais de dinâmica dos fluídos, diferindo-se apenas na parte das chamadas opções físicas, que são esquemas pertinentes à um processo físico em específico. A modelagem de previsão numérica do tempo requer em grandes quantidades o conhecimento de um especialista, além de saber quais usar e entender as diferenças entre cada uma das opções desses processos físicos, também é necessário levar em consideração a possível interação entre diferentes esquemas dos mesmos [33]. A abordagem de modelagem numérica do tempo é uma tarefa importante e indispensável, mas também pode ser considerada uma tarefa que apresenta um alto nível de complexidade e sujeita a erros [37].

## 2.3 MÉTODOS EMPÍRICOS

Desde a década de 1920 inúmeros modelos de correlação empíricos que visam a estimação da radiação solar vem sendo reportados na literatura. Ja é conhecido que a radiação solar é relativamente afetada por parâmetros meteorológicos, fatores geográficos, astronômicos e geométricos. No período que compreende o final dos anos 70 até o fim dos anos 80 alguns trabalhos foram realizados na tentativa de se obter uma projeção dos valores de RS. Em [57], a radiação solar em uma superfície é dada por uma expressão que é função do angulo zenital do sol (angulo da iluminação na situação hipotética da superfície terrestre ser perfeitamente esférica) e considera algumas constantes, essas que dependem exclusivamente da altitude da superfície e de um modelo de neblina que informa o alcance da visibilidade local. Fazendo uso da cobertura de nuvens como principal parâmetro para a estimação de radiação solar, o trabalho [58] sugere para qualquer latitude que a RS é dada como sendo função da cobertura total de nuvens opacas.

Como forma de contornar os problemas de custos de manutenção e perícia envolvidos na medição de terreno e obtenção de dados derivados de satélite, especialmente em se tratando de áreas rurais e países em desenvolvimento, o autor em [23] estuda inúmeros modelos empíricos de estimação como forma alternativa e viável para regiões da África Ocidental. Os modelos apresentados podem ser classificados em seis principais categorias

diferentes, cada um de acordo com sua própria fundamentação, sendo eles baseados em: luz do sol, nuvens, temperatura, umidade relativa, precipitação e parâmetros híbridos.

Usualmente os modelos baseados na luz solar são os mais frequentemente utilizados devido à disponibilidade e acesso à informação na maioria das estações. Em [38], o autor faz o uso de validação cruzada para melhorar a performance de estimação da radiação solar global de 25 diferentes modelos empíricos baseados na duração da luz solar. Os dados utilizados são referentes ao período de $1986 - 2000$ e são fornecidos pelo departamento meteorológico da Índia, na cidade de Pune. Em [39] o trabalho também apresenta modelos baseados na duração da luz solar, na intenção de realizar tanto a calibração quanto a validação dos mesmos. Adicionalmente o autor propõem dois métodos diferentes de controle de qualidade para rejeitar erros nos dados de radiação solar diário coletado de 2007 a 2017. Os resultados mostraram que para a maior parte das regiões modeladas, a maior correlação entre os valores observados e aqueles estimados, foram detectadas nos modelos baseados em regressão.

## 2.4 MÉTODOS INTELIGENTES BASEADOS EM DADOS

No momento atual a área de aprendizado de máquina talvez seja a abordagem mais popular para a previsão solar. Possuindo diversos métodos com seus inúmeros variantes e suportando uma larga escala de diferentes aplicações, o campo de pesquisa oferece um ferramental conveniente para lidar com problemas dessa natureza. Apesar da diversidade de algoritmos existentes, todos eles são baseados em um mesmo conceito: fazer o uso de dados para extrair parâmetros e aprender padrões de forma a se criar um modelo com capacidade representativa desses mesmos dados [33].

Devido à grande especificidade inerente à natureza da incidência solar para cada região, um grande número de pesquisadores vêm propondo modelos para a previsão da RS tendo como base dados constituídos por parâmetros climáticos. Esses parâmetros geralmente são duração da luz solar, a sazonalidade ao longo dos meses do ano, temperatura ambiente, altitude, longitude, umidade do ar, velocidade do vento, pressão atmosférica, temperatura do solo, evaporação de água dentre outros [23].

O trabalho publicado em [40], descreve um estudo comparativo de diferentes técnicas aplicadas na estimação da radiação solar diária, considerando diferentes combinações das variáveis de entrada. Uma máquina de vetor suporte e uma rede neural artificial são empregados para modelar os dados constituintes do período de $1996 - 2011$. Os resultados de cada abordagem de estimação, são comparados em função dos indicadores de erros MBE, RMSE e $R^2$, e mostram que a máquina de vetor suporte apresenta maior performance quando comparado à rede neural na estimativa da radiação. Em [11], os autores destacam as desvantagens identificadas na literatura associadas à estimação da RS, como a pobre capacidade preditivas dos modelos, incorporação de erros durante a fase de modelagem

e exigências referente à necessidade de dados de longa data. Em contrapartida, para a mitigação desses problemas, é proposto um modelo híbrido que combina evolução diferencial com uma máquina de aprendizado extremo. Devido ao grande potencial e destaque, abordagens mais avançadas e complexas como o aprendizado profundo também vem sendo utilizadas. Em [41] com base em fatores astronômicos, radiação extraterrestre, cobertura de nuvens além de temperatura máxima e mínima e a duração da luz solar são empregados para a estimação da RS. O $R^2$, RMSE e o MAE foram reportados como $0,980$, $0,78\ MJm^{-2}day^{-1}$ e $0,61\ MJm^{-2}day^{-1}$, respectivamente.

Tomando como base na literatura, é possível identificar que parâmetros como o tempo da duração da luz solar, umidade relativa do ar, pressão do ar, temperaturas máxima e mínima do ambiente, latitude e longitude, altitude, e os diferentes meses do ano são frequentemente usados como parâmetros de entrada para aplicações de previsão e estimação de RS.

Diversos e inúmeros outros trabalhos foram e vem sendo realizados objetivando resolver o problema de obtenção dos dados de radiação solar. A Tabela 1, apresenta uma síntese dos trabalhos citados nessa seção, além de apresentar alguns estudos semelhantes adicionais em ordem cronológica, visando ilustrar o cenário de pesquisas da área e bem como sua evolução ao longo dos anos.

Tabela 1 – Síntese de alguns dos trabalhos previamente realizados na área de radiação solar visando sua estimação/previsão.

| **Publicação** | **Método Empregado** | **Ano** | **Resumo** |
|---|---|---|---|
| Hottel, H. C. [57] | Empírico | 1976 | Estimação da RS na superfície em função do angulo zenital do sol e considera constantes definidas de acordo com a altitude e um modelo de neblina. |
| Brinsfield et al. [58] | Empírico | 1984 | RS estimada para qualquer latitude com base fundamentada em função da cobertura total de nuvens opacas da região. |
| Yadav et al. [57] | ANN | 2014 | Síntese geral, com revisão e análise de estudos que visam a previsão de RS utilizando redes neuronais. |
| Dong et al. [57] | SVR + SOM + PSO | 2015 | Abordagem híbrida fazendo uso de regressão de vetor suporte, mapa auto-organizável e otimização de enxame de partículas na previsão de RS. |
| Voyant et al. [57] | ML | 2017 | Revisão geral de estudos da área que fazem o uso de métodos de aprendizado de máquina. |
| Nwokolo et al. [23] | Empírico | 2017 | Revisão quantitativa e classificação de modelos empíricos para previsão da radiação solar na África Ocidental. |
| Da Silva et al. [40] | Empírico SVM ANN | 2017 | Estudo comparativo de diferentes métodos para estimar a radiação solar diária em Botucatu - SP. |
| Yang et al. [33] | — | 2018 | Histórico e tendências na previsão de radiação solar e de energia fotovoltaica. |
| Kaba et al. [41] | DL | 2018 | Método de aprendizado profundo utilizado para estimar a radiação solar em 30 estações localizadas na Turquia. |
| Tao et al. [11] | ELM + DE | 2019 | Utilização de um modelo híbrido para previsão da radiação solar diária utilizando diferentes combinações das variáveis de entrada. |
| Narvaez et al. [63] | DL | 2020 | Proposta de metodologia para o aprimoramento da qualidade dos dados e da previsão da radiação solar utilizando aprendizado profundo. |
| Tao et al. [64] | ANFIS | 2021 | Uso de modelo de inferência adaptativo neuro-fuzzy para a predição da radiação solar com base na temperatura do ar. |

Fonte: Elaborada pelo autor (2021).

## 2.5 REGRESSORES RIDGE

De uma forma generalizada, regressão é um processo estatístico para a estimação das relações entre diferentes variáveis, que pode ser de grande utilidade para a previsão solar em termos da modelagem de variáveis exógenas. Frequentemente através do método dos mínimos quadrados (OLS) é possível estimar o vetor de parâmetros $\beta$, porém para esse método, a acurácia das predições é afetada quando há um grande número de preditores além do modelo perder interpretabilidade. Nesse cenário é fundamental realizar o processo de seleção das variáveis de modo a melhorar a performance entregue pelo modelo [33].

Como solução alternativa, tem-se um tipo específico de regularização de Tikhonov chamada de regressão ridge. Esse tipo de regressão pode ser particularmente útil para mitigar os problemas de multicolinearidade, que comumente ocorre em casos onde há a presença de um grande número de variáveis explicativas. Os modelos de regressão tem sido amplamente utilizados ao longo dos anos na área de eficiência energética, tanto na estimação de radiação solar [33, 42] quanto na previsão de temperatura interna de edifícios [43].

# 3 MATERIAIS E MÉTODOS

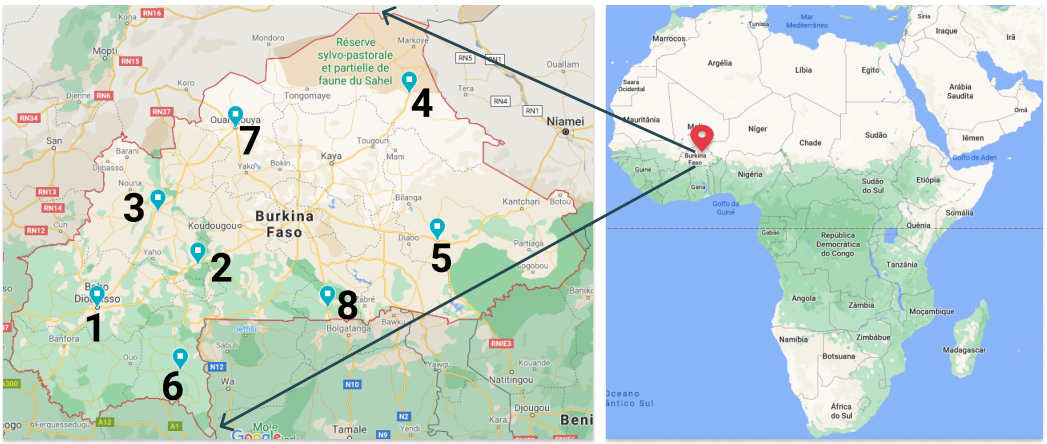
## 3.1 BASE DE DADOS

Para o problema de estimação da radiação solar, a fim de desenvolvimento e avaliação dos modelos presentes nesse trabalho bem como também a validação das ferramentas de aprendizado de máquina empregadas, é proposto a utilização de oito bases de dados disponíveis na literatura por [11]. Detalhes sobre os conjuntos de dados utilizados são descritos na sequência.

A Burkina Faso é um país localizado na África subsariana, que possui uma extensão aproximada de 274.000 $m^2$ e uma população de mais de 20 milhões de habitantes, Figura 7. O país conta com uma infraestrutura geral muito pouco desenvolvida, e em se tratando dos meios de produção de energia, a região apresenta um cenário de uso extensivo de fontes poluentes e pouco eficientes. Cerca de aproximadamente 70% da energia gerada no país é proveniente de fontes de combustíveis termo-fósseis, enquanto os outros 30% provem da energia hidroelétrica. Além de grande parte da matriz energética do país serem fontes geradoras de poluentes, esse tipo de operação possui um alto custo e está vulnerável à instabilidade do preço do petróleo, dada a crescente demanda pelo uso de eletricidade. Adicionalmente, regiões rurais e mais remotas sofrem com a falta de infraestrutura da rede elétrica responsável pelo abastecimento local [11, 23].

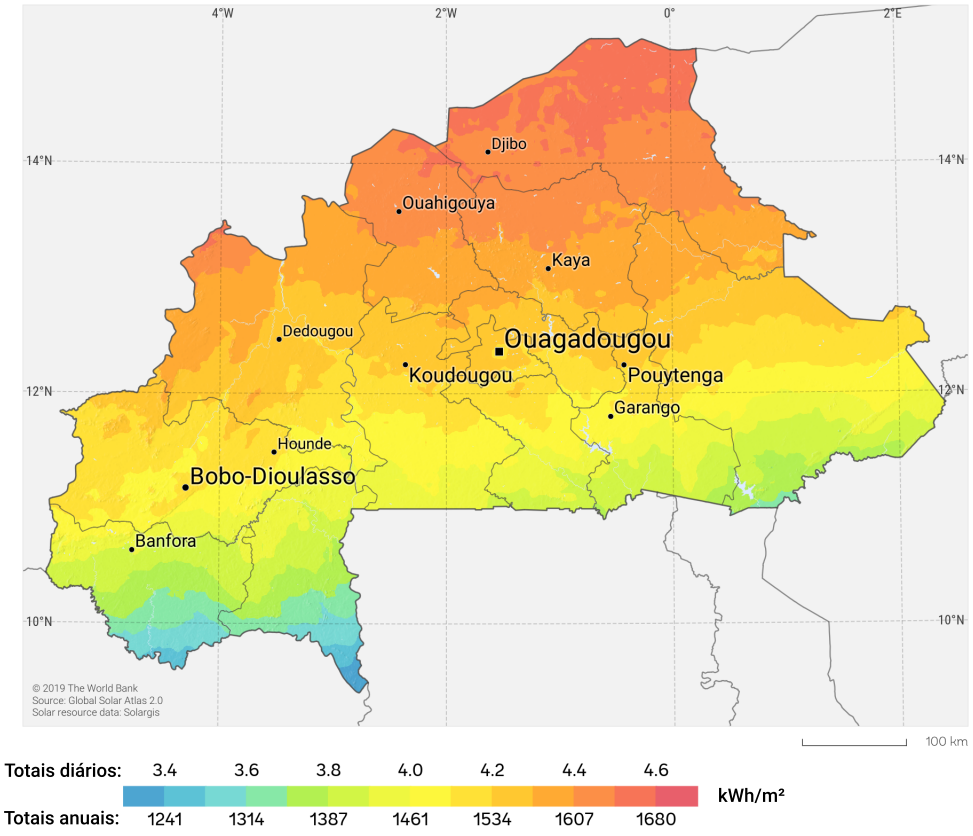
Esse país se encontra geograficamente próxima à região equatorial da Terra, região essa que apresenta um alto grau de incidência solar durante todo o período do ano. Na maior parte de seu território, segundo o Atlas de Potencial de Energia Solar [10] o índice de radiação solar diária é considerado acima de $4,2\ KWh/m^2$, Figura 8.

Figura 7 - Localização geográfica das estações meteorológicas estudadas distribuídas ao longo do território da Burkina Faso na África Ocidental.




Fonte: Adaptado do Google Maps.

Figura 8 - Mapa apresentando a média de longo prazo de radiação normal direta no território da Burkina Faso, durante o período de 1994 – 2018.




Fonte: Adaptado de Global Solar Atlas 2.0 [10].

No presente trabalho, os dados de radiação solar diária média para oito diferentes cidades na Burkina Faso foram estimados usando um modelo em específico para cada uma das regiões. Os dados utilizados foram coletados durante o período de observação entre 1998 e 2012, mensurados por diferentes estações meteorológicas, uma respectiva para cada cidade. A distribuição geográfica das estações ao longo do território do país pode ser vista na Figura 7. Adicionalmente, a Tabela 2 traz informações mais específicas à respeito da localização de cada estação, como as coordenadas geográficas e a altitude, além de apresentar um sistema de numeração que funciona como uma espécie de etiqueta à fim de facilitar a identificação e referenciação ao longo do trabalho.

Tabela 2 – Lista das localidades que compõem os conjuntos de dados utilizados no estudo. Os locais correspondem cada um à uma estação meteorológica distribuídas pelo país.

| **Nº da Estação** | **Latitude** | **Longitude** | **Altitude** | **Período Observado** |
|---|---|---|---|---|
| I | 11°11' N | 4°17' O | 445 m | 1998 - 2012 |
| II | 11°45' N | 2°56' O | 263 m | 1998 - 2012 |
| III | 12°28' N | 3°28' O | 299 m | 1998 - 2012 |
| IV | 14°02' N | 0°02' O | 286 m | 1998 - 2012 |
| V | 12°03' N | 0°22' L | 294 m | 1998 - 2012 |
| VI | 10°19' N | 3°10' O | 339 m | 1998 - 2012 |
| VII | 13°35' N | 2°25' O | 315 m | 1998 - 2012 |
| VIII | 11°10' N | 1°08' O | 305 m | 1998 - 2012 |

Fonte: Elaborada pelo autor (2021).

O conjunto de dados referente a cada uma das estações meteorológicas apresentadas na Tabela 2 é composto por sete variáveis de entrada $(X_1, ..., X_7)$ e apenas uma variável de saída $Y$. A variável dependente $Y$ é a própria radiação solar diária média, e para as variáveis independentes $(X_1, ..., X_7)$ são utilizados parâmetros relacionado ao clima do local. Sendo esses parâmetros, a velocidade do vento, a temperatura máxima e mínima, a umidade máxima e mínima, o deficit de pressão de vapor (definido como a diferença entre a quantidade de umidade no ar e a quantidade de umidade que o ar pode reter quando está saturado) e por último a taxa de evaporação de água, presentes na Tabela 3. Considerando o conjunto de dados por completo, durante o período que compreende a série de observações, para cada estação tem-se 5479 observações diárias, produzindo uma matriz de tamanho $5479 \times 8$.

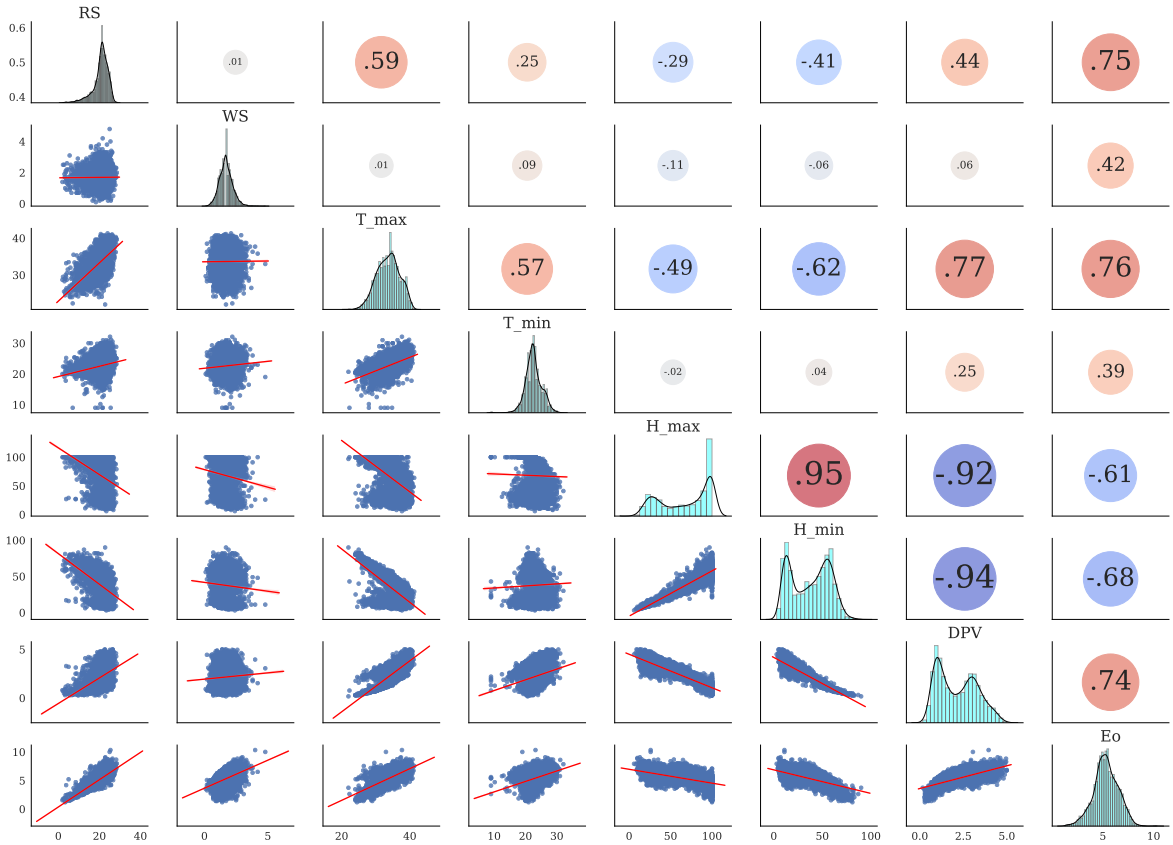
Tabela 3 – Descrição detalhada das variáveis que compõem a base de dados.

| **Nº da Variável** | **Representação Matemática** | **Nome da Variável** | **Tipo** | **Unidade de Medida** |
|---|---|---|---|---|
| 1 | $WS$ | Velocidade do vento | Entrada | $m/s$ |
| 2 | $T_{max}$ | Temperatura máxima | Entrada | $°C$ |
| 3 | $T_{min}$ | Temperatura mínima | Entrada | $°C$ |
| 4 | $H_{max}$ | Umidade máxima | Entrada | $gm/ml^3$ |
| 5 | $H_{min}$ | Umidade mínima | Entrada | $gm/ml^3$ |
| 6 | $DPV$ | Déficit de pressão de vapor | Entrada | $kPa$ |
| 7 | $E_o$ | Evaporação | Entrada | $mm$ |
| 8 | $RS$ | Radiação solar | Saída | $MJ/m^2$ |

Fonte: Elaborada pelo autor (2021).

A Figura 9 abaixo, apresenta um panorama geral à respeito do conjunto de dados referente à estação I. A imagem mostra os gráficos de dispersão bivariados abaixo da diagonal, o histograma mostrando a distribuição de cada variável ao longo da diagonal e exibe a relação existente entre a quantidade de radiação solar com as demais variáveis independentes: velocidade do vento, temperatura máxima e mínima, deficit de pressão de vapor, umidade máxima e mínima e a evaporação. A relação entre as variáveis é apresentada por meio do coeficiente de correlação de Pearson ($\rho$), mostrados acima da diagonal e também é conhecido pelo nome de coeficiente de correlação produto-momento. Esse fator da estatística descritiva mede não somente o grau da correlação entre duas variáveis mas também a direção dessa correlação. Dessa forma, o coeficiente apresenta apenas valores contidos no intervalo de $[-1, 1]$, sendo que $\rho = 1$ significa uma correlação perfeitamente positiva entre duas variáveis, em contrapartida $\rho = -1$ representa uma correlação negativa perfeita, e para $\rho = 0$ sugere-se a não existência de dependência linear entre as variáveis. O coeficiente de correlação de Pearson pode ser calculado pela Equação 3.1, onde $a_1, a_2, ..., a_n$ e $b_1, b_2, ..., b_n$ são os valores medidos para ambas variáveis e $\bar{a}$ e $\bar{b}$ são suas respectivas médias.

Figura 9 - Gráficos de dispersão e de distribuição para cada variável e coeficiente de Pearson apresentando a relação entre as variáveis para a estação de número I.




Fonte: Elaborado pelo autor (2021).

$$\rho = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2} \cdot \sqrt{\sum_{i=1}^{n}(b_i - \bar{b})^2}} = \frac{cov(A, B)}{\sqrt{var(A) \cdot var(B)}} \tag{3.1}$$

A Figura 9 revela a existência de uma correlação positiva e forte (0.75) entre a RS e a quantidade de evaporação. A segunda maior correlação é entre a RS e a temperatura máxima (0.59) que também é positiva. Duas correlações negativas e fortes como esperado, são observadas entre o deficit de pressão de vapor e o nível de umidade do ar tanto máxima quanto mínima, já que por definição o DPV é calculado pela diferença entre a quantidade de umidade presente no ar e a quantidade de umidade que o ar pode reter quando se encontra saturado ($-0.92$) e ($-0.94$). Outro comportamento intuitivo da dinâmica climática que é captado pela análise é a forte correlação positiva entre a temperatura máxima do dia e a quantidade de evaporação (0.76), sugerindo o fato de que em dias mais quentes há uma maior intensidade de evaporação de água. As figuras com os gráficos de dispersão bivariados, histogramas de distribuição e a relação entre as variáveis segundo o coeficiente de correlação de Pearson para o conjunto de dados referente às outras sete estações (II, III, IV, V, VI, VII e VIII) estão no Apêndice A.

## 3.2 MODELOS DE REGRESSÃO

### 3.2.1 RIDGE

A Regressão Ridge [44] é uma técnica utilizada para a análise de dados de regressão múltipla que sofrem com a presença da multicolinearidade, a qual ocorre frequentemente em modelos com um grande número de parâmetros. Regularmente esse método é empregado em casos onde existe a presença de muitos preditores com diferentes coeficientes, e atua de modo a impedir com que os coeficientes do modelo de regressão linear com variáveis correlacionadas sejam mal determinados e resultem em grande variação.

A multicolinearidade é a existência de relações quase-lineares entre as variáveis independentes. A presença desse tipo de relação entre as variáveis pode acarretar na criação de estimativas imprecisas dos coeficientes de regressão, aumentar os erros padrão dos coeficientes e degradar a capacidade preditiva do modelo.

Durante o processo de estimação dos coeficientes a estratégia Ridge fornece um meio de regularização, restringindo o efeito de variáveis não impactantes. Em termos de um modelo de regressão linear típico usando mínimos quadrados ordinários, isso é feito modificando a típica função de perda da soma dos quadrados do erro residual (RSS, do inglês *residual sum of squares*) de modo a adicionar um fator de penalidade para valores de coeficientes de magnitude elevados.

Este tipo de regressão soluciona o problema do tipo:

$$\mathbf{y} = w\mathbf{X} + b, \tag{3.2}$$

em que **y** é o vetor de observações (variável dependente), **X** a matriz de entradas, $w$ é o vetor de coeficientes de regressão e $b$ o vetor de erros residuais.

$$y_i = b + \underbrace{w_1 x_{i1} + w_2 x_{i2} + \cdots + w_p x_{ip}}_{\boldsymbol{w}.x_i} \tag{3.3}$$

A Equação 3.4 representa o i-ésimo resíduo (erro), que é dado pela diferença entre o i-ésimo valor de resposta real $y_i$ e o i-ésimo valor de resposta previsto pelo modelo $\widehat{y_i}$.

$$e_i = y_i - \widehat{y_i} \tag{3.4}$$

O problema recai na busca pelos coeficientes ótimos que otimize de forma a minimizar a função da soma dos quadrados do erro residual, definida como

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2, \tag{3.5}$$

que é equivalente a

$$RSS = \sum_{i=1}^{n} \left( y_i - b - \sum_{j=1}^{p} w_j x_{ij} \right)^2. \tag{3.6}$$

A regressão do tipo Ridge é muito semelhante a um modelo de regressão típico usando a RSS, exceto que, agora existe a adição de um novo parâmetro de redução $\alpha$ que minimiza seu valor, Equação 3.7. O novo parâmetro $\alpha$ também é chamado de "parâmetro de ajuste"ou hiperparâmetro, e é determinado de forma separada, frequentemente através de técnicas de validação cruzada.

$$RSS_{ridge} = \sum_{i=1}^{n} \left( y_i - b - \sum_{j=1}^{p} w_j x_{ij} \right)^2 + \boxed{\alpha \sum_{j=1}^{p} {w_j}^2} \tag{3.7}$$

Supondo que dentre os coeficientes $[w_1, w_2, \cdots, w_p]$ alguns estão assumindo valores já próximos de zero, sugerindo assim atributos (variáveis) que não têm muito impacto na variável de resposta. Com a adição do parâmetro de redução $\alpha$, os valores desses coeficientes que já são baixos tendem a zero. Dessa forma, o segundo termo do lado direito da Equação 3.7 em destaque recebe o nome de termo de redução ou norma-L2. Para o caso onde $\alpha = 0$ o termo de redução desaparece, transformando o modelo em um caso de regressão tradicional, em contrapartida, à medida que $\alpha \to \infty$ a força do termo de regularização se eleva e os coeficientes da regressão Ridge tendem a se aproximar de zero.

A vantagem da regressão Ridge sobre os mínimos quadrados está enraizada na compensação de viés-variância. À medida que $\alpha$ aumenta, a flexibilidade do ajuste da regressão diminui, levando à diminuição da variância mas aumentando o viés.

### 3.2.2 LASSO

Enquanto a regressão Ridge faz uso de um método de regularização que tem como principal objetivo suavizar atributos que sejam relacionados uns aos outros, a regularização Lasso possui o mesmo mecanismo de penalização dos coeficientes com um alto grau de correlação entre si, mas que porém penaliza os coeficientes de acordo com o seu valor absoluto (soma dos valores dos estimadores) usando o mecanismo de minimizar o erro quadrático [62].

Ambos os métodos de regularização são matematicamente semelhantes e possuem uma formulação parecida. A regressão Lasso também faz o uso de um hiperparâmetro $\alpha$ que é responsável pela intensidade da regularização a ser aplicada. Sua formulação pode ser descrita da mesma forma que a Ridge com alterações sendo feitas somente no termo final da equação em destaque, termo também conhecido como norma-L1.

$$RSS_{lasso} = \sum_{i=1}^{n} \left( y_i - b - \sum_{j=1}^{p} w_j x_{ij} \right)^2 + \boxed{\alpha \sum_{j=1}^{p} |w_j|} \tag{3.8}$$

Além de diminuir a variância do modelo, esse tipo de regularização é muito utilizado em algumas abordagens de aprendizado de máquina pela sua capacidade de seleção de atributos/variáveis. Quando há múltiplos atributos altamente correlacionadas ou seja, variáveis que se comportam da mesma maneira, a regularização Lasso seleciona apenas uma dessas e zera os coeficientes das outras. Desse modo, dizemos que esse modelo realiza uma seleção de atributos de forma automatica, gerando vários coeficientes com peso zero, ou seja, que são ignorados pelo modelo. Esse mecanismo facilita a interpretação do modelo, o que frequentemente acaba sendo uma grande vantagem.

## 3.3 ATRIBUTOS POLINOMIAIS

Na área de aprendizado de máquina, buscando a melhoria do desempenho dos modelos frequentemente são utilizados métodos e artifícios para a geração de novos atributos que caracterizem o conjunto de dados trabalhado. Chamado de *feature engineering* (engenharia de atributos), a prática de realizar alterações nos atributos dos dados engloba diferentes abordagens como codificação categórica, geração de atributos (com base na combinação dos dados já existentes), seleção de atributos dentre outros.

Frequentemente, os atributos de entrada para um modelo interagem de maneiras inesperadas e constantemente não lineares. Essas interações podem ser facilmente identifi-

cadas e modeladas por algoritmos de aprendizagem. Uma abordagem mais direta consiste em projetar novos atributos que expõem essas interações diretamente no conjunto de dados, e avaliar se elas contribuem para a melhora de performance do modelo. Além disso, transformações como elevar as variáveis de entrada a uma potência podem contribuir para expor de melhor forma as possíveis relações importantes entre as variáveis de entrada e a variável de saída.

Conhecidos como atributos de interação e atributos polinomiais, os mesmos permitem o uso de algoritmos de modelagem mais simples, já que parte da complexidade de interpretar as variáveis de entrada e suas relações é devolvida ao estágio de pré-processamento dos dados. Em alguns casos, esses atributos podem resultar em melhor desempenho do modelo, apesar do custo de se adicionar inúmeras outras variáveis de entrada a mais.

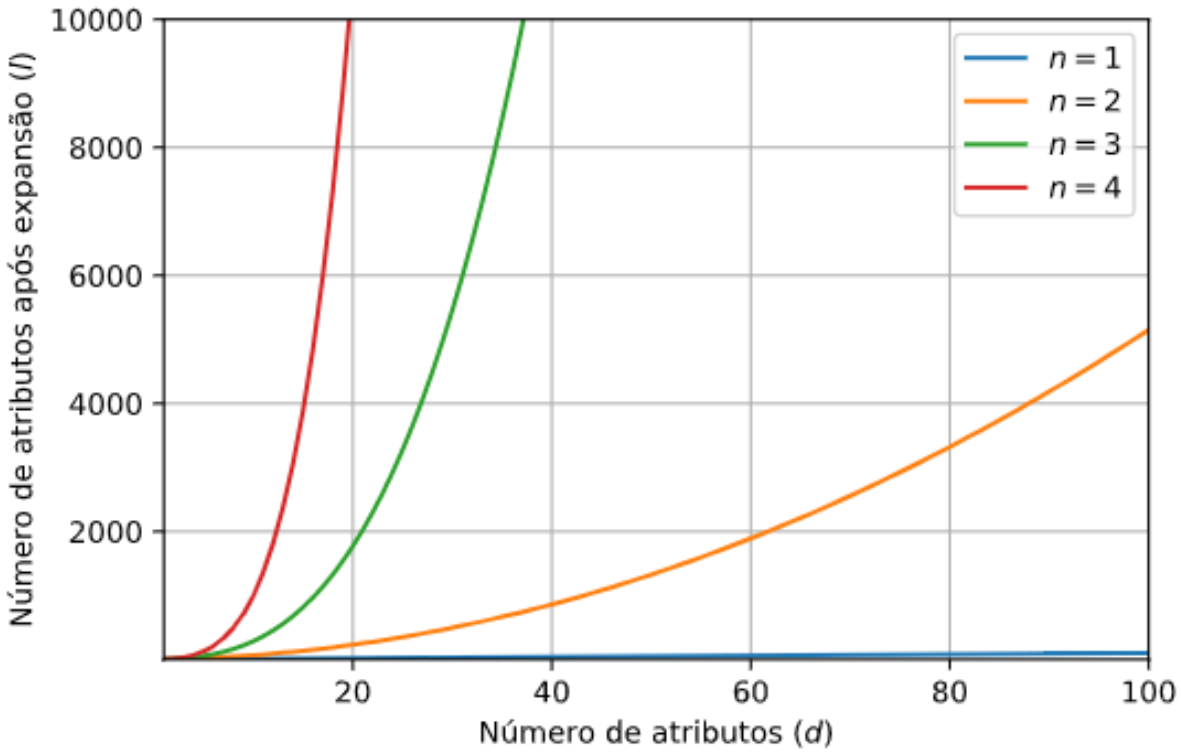
Dada uma única observação representada por $d$ atributos, o vetor das variáveis passado como entrada é definido por $\mathbf{x} = [x_1 \;\; x_2 \;\; ... \;\; x_d]$, e é expandido em termos do polinômio baseado em vetor $\mathbf{p}_n(\mathbf{x})$ onde $n$ é o grau do polinômio. Dessa forma, esse processo permite o mapeamento de um vetor de atributos de dimensão $d$ em um vetor de dimensão $l$ dependente do grau escolhido para o polinômio. Por exemplo, considerando o vetor de entrada $\mathbf{x} = [x_1 \;\; x_2]$ com dimensão $d = 2$, após a expansão para $\mathbf{p}_2(\mathbf{x})$ tem-se

$$\mathbf{p}_2(\mathbf{x}) = [1 \quad x_1 \quad x_2 \quad x_1^2 \quad x_1x_2 \quad x_2^2] \tag{3.9}$$

com dimensão $l = 6$.

O número de atributos gerados pela expansão está diretamente relacionado com o grau do polinômio utilizado, essa relação é dada através de um crescimento exponencial em função não somente do grau mas também do número de variáveis que constituem o conjunto original. A Figura 10 mostra a proporção do aumento no número de atributos gerados para os casos de polinômios com grau $n = 1$, $n = 2$, $n = 3$ e $n = 4$ em função do número de variáveis previamente já existentes. Conhecido o impacto do grau utilizado pelo polinômio sobre a complexidade dos dados, faz-se necessário dosar até que ponto é benéfico explorar a criação de novos atributos representativos de forma à obter mais informações e ao mesmo tempo limitar o número de preditores.

Figura 10 - Variação do número de variáveis de entrada após a utilização da expansão polinomial de acordo com o grau escolhido para o polinômio ($n$).




Fonte: Elaborado pelo autor (2021).

## 3.4 MUTUAL INFORMATION vs F-REGRESSION

Como parte do processo de projetar uma máquina de aprendizado, a etapa de seleção de atributos consiste em realizar análises capazes de selecionar as variáveis de maior importância para serem utilizadas na construção do modelo. Durante a seleção, identificar e descartar as variáveis que pouco contribuem para explicar os dados é igualmente importante à seleção das variáveis que apresentam maior explicação. O processo de seleção tem como premissa central que os dados possuem variáveis redundantes ou até mesmo irrelevantes, e que portanto, podem ser removidas sem incorrer em perda de informação. As técnicas de escolha das variáveis são utilizadas por diversas razões como: simplificação dos modelos, encurtamento do tempo de treinamento, evitar a chamada *curse of dimensionality* e até mesmo para aprimorar a capacidade de generalização dos modelos através da redução do *overfitting*.

Frequentemente uma metodologia para seleção de atributos em um determinado conjunto de dados consiste na seguinte estrutura:

1. Para fins de projeto é escolhido uma função de pontuação (*score function*);

2. Cada variável do conjunto de dados é avaliada segundo a função de pontuação escolhida previamente, e a cada uma delas é atribuída um valor;
3. As variáveis são organizadas em formato de ranking decrescente em função de suas pontuações;
4. O projetista define um número $k$ das top variáveis do ranking a serem incluídas na modelagem e descarta o resto;
5. O modelo é treinado e avaliado segundo os atributos selecionados.

Para fins de realizar a seleção das variáveis, nesse trabalho foi utilizado como função de pontuação a informação mútua (em inglês MI: mutual information). Dentro da área de teoria das probabilidades e teoria da informação, a informação mútua mede a dependência mútua entre duas variáveis aleatórias. Em outras palavras, a MI quantifica a informação que uma variável aleatória contém acerca da outra. Seu conceito está intrinsecamente ligado ao da Entropia de uma variável aleatória, considerada uma noção fundamental da teoria da informação que define a quantidade de informação contida em uma variável aleatória.

A informação mútua entre duas variáveis aleatórias é sempre um valor não-negativo, que assume o valor zero somente se as variáveis são independentes, e para maiores valores implica em uma maior dependência. A função para o cálculo da MI utilizada nesse trabalho se baseia em métodos não-paramétricos baseados na estimativa de entropia das distâncias dos k-vizinhos mais próximos conforme apresentado em [45–47].

Formalmente, a informação mútua pode ser definida para duas variáveis aleatórias discretas $X$ e $Y$ como sendo:

$$MI(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log \left( \frac{p(x,y)}{p(x)p(y)} \right), \tag{3.10}$$

em que $p(x)$ e $p(y)$ são as funções de probabilidade de distribuição marginal de $X$ e $Y$ respectivamente, e $p(x,y)$ é a distribuição de probabilidade conjunta de $X$ e $Y$.

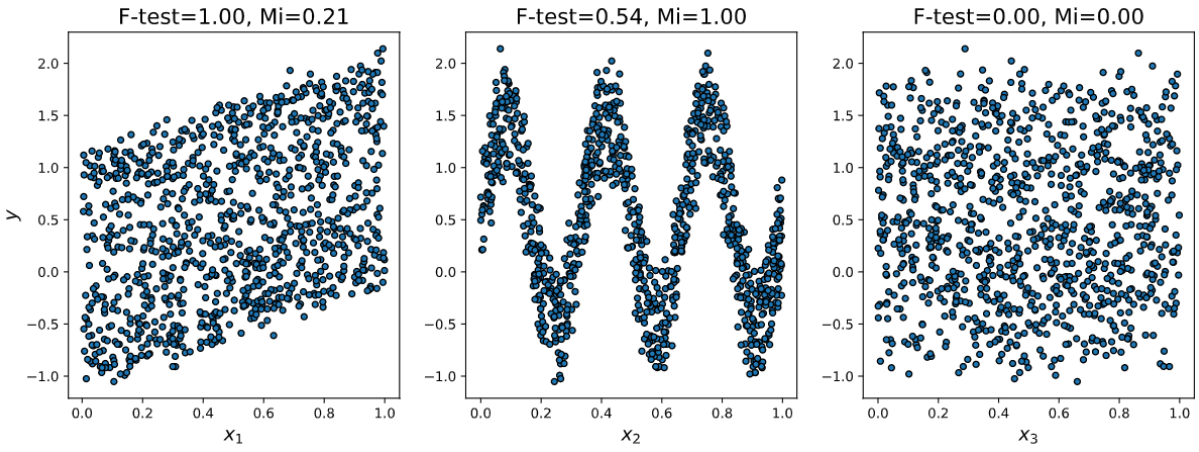
Assim como os testes de regressão linear uni-variados (F-testes) a informação mútua também pode ser utilizada como forma de avaliação e critério de seleção dos atributos. O F-teste consiste na criação de modelos lineares para testar o efeito individual de cada um dos regressores trabalhados. Em um primeiro momento a correlação entre cada regressor e a variável dependente é computada, em seguida os valores são convertidos à uma pontuação e um valor-p.

Enquanto o F-teste possui a capacidade de capturar somente a dependência linear entre variáveis, por outro lado a informação mútua consegue capturar diferentes tipos de dependência entre as variáveis. A título de ilustração dessas características, considerando

três diferentes atributos $x_1$, $x_2$ e $x_3$ distribuídos uniformemente no intervalo $[0, 1]$, sendo a variável dependente dada por $y = x_1 + sen(6\pi x_2) + 0.1$, ou seja a terceira variável é completamente irrelevante.

Nesse contexto, a Figura 11 apresenta a dependência individual de $y$ em função de $x_1$, $x_2$ e $x_3$ e os valores normalizados de estatísticas dos testes-F univariados e as informações mútuas. Para $x_3$, ambos métodos conseguem captar e sinalizar a irrelevância da variável. Por sua característica de captar somente a dependência linear o F-teste elege $x_1$ como a variável de maior importância. Em contrapartida, a informação mútua consegue captar qualquer tipo de dependência e opta por escolher $x_2$ como variável mais discriminativa, o que vai de encontro com a percepção intuitiva à respeito do exemplo.

Figura 11 - Comparação entre F-teste e a informação mútua como critério de seleção de atributos utilizando como base um exemplo com três variáveis independentes.




Fonte: Elaborado pelo autor (2021).

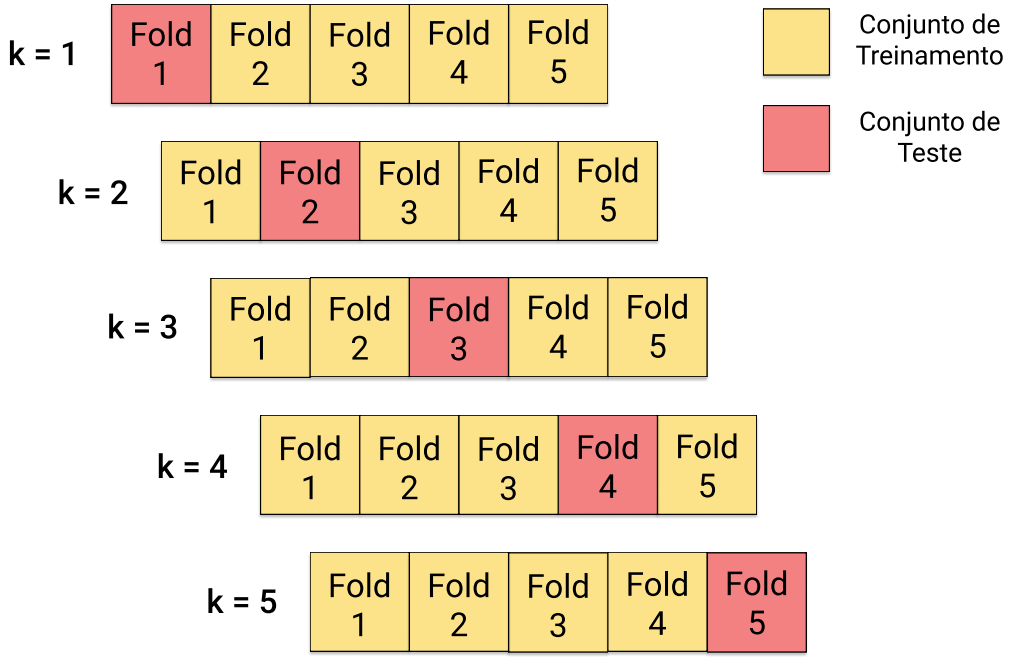
## 3.5 VALIDAÇÃO CRUZADA

Para um algoritmo de aprendizado, obter os seus parâmetros através do treinamento e em seguida testá-lo no mesmo conjunto de dados é considerado um erro de metodologia. Um modelo avaliado em sua fase de testes sobre o mesmo conjunto de dados utilizado para o treinamento pode apresentar uma performance elevada, porém, não seria possível prever ou estimar de maneira satisfatória para um conjunto de dados desconhecido. Essa situação é chamada de sobreajuste ou no inglês *overfitting*. Para evitar as condições de sobreajuste, quando se esta conduzindo experimentos de aprendizado de máquina em abordagem supervisionada, é prática comum dividir o conjunto de dados previamente em grupo de treinamento e grupo de teste. Apesar da estratégia auxiliar na avaliação dos modelos, ela não garante que os modelos não possam vir a sofrer o sobreajuste.

A validação cruzada é uma técnica de amostragem estatística que visa realizar a divisão do conjunto de dados e avaliar a capacidade de generalização de um determinado modelo [48]. Esse tipo de técnica busca estimar o quanto um modelo em específico é preciso para um novo conjunto de dados ainda não apresentado ao mesmo. Dentre os diversos tipos de validação cruzada, a metodologia K-fold é frequentemente a mais utilizada [49].

A Figura 13 apresenta o esquema de funcionamento do método de validação cruzada K-fold, onde no primeiro momento há um particionamento dos dados em $K$ partes e em seguida seleciona essas partes específicas para treinar e testar o modelo. Inicialmente o conjunto de dados original é randomicamente dividido em $K$ subconjuntos, sendo que $K \geq 1$. Posteriormente dos $K$ subconjuntos criados, $K - 1$ são utilizados para a etapa de treinamento e o subconjunto restante é utilizado para o teste do modelo. Esse processo é repetido $K$ vezes, mudando sempre a cada iteração o subconjunto para o teste. Ao final, a avaliação do modelo é feita usando-se o valor médio da métrica obtida ao longo das $k$ iterações.

Figura 12 - Esquema de funcionamento da metodologia de validação cruzada K-fold para o caso de $k = 5$.




Fonte: Elaborado pelo autor (2021).

## 3.6 PADRONIZAÇÃO DOS DADOS

Como parte fundamental do pré-processamento dos dados, faz-se necessário a realização de sua padronização ou normalização. Essa técnica é responsável por alterar os

valores das variáveis presentes no conjunto de dados para uma escala comum, sem distorcer as diferenças em seus respectivos intervalos de valores. Apesar de algumas exceções, a padronização das variáveis é um requisito comum para a maioria dos algoritmos de aprendizado, podendo esses se comportarem de maneira inadequada caso haja atributos não padronizados que não se assemelham com uma distribuição normal padrão.

Muitos elementos utilizados na função objetivo de um algoritmo de aprendizado, assumem que todas variáveis estão centralizadas em torno de 0 e que possuem uma variância de mesma ordem. Caso um atributo possua uma variância de magnitude superior em relação às outras, ela pode dominar a função objetivo e tornar o estimador incapaz de aprender de forma correta com as demais variáveis assim como o esperado.

A padronização dos dados normalmente é feita através da fórmula *z-score*, apresentada abaixo:

$$z = \frac{x - \mu}{\sigma}, \tag{3.11}$$

em que o valor da variável, a sua média e seu desvio padrão são representados por $x$, $\mu$ e $\sigma$, respectivamente.

## 3.7 SELEÇÃO DE MODELOS

Inúmeros modelos de aprendizagem presentes na literatura e até mesmo técnicas e procedimentos realizados em etapa inicial de pré-processamento dos dados envolvem o uso e definição de parâmetros que não podem ser estimados diretamente com base nos dados de entrada. Esses parâmetros são conhecidos como hiperparâmetros e eles devem ser definidos antes da etapa de aprendizado dos algoritmos. O desempenho de um modelo pode ser afetado diretamente de acordo com os hiperparâmetros adotados. Portanto, é uma tarefa de grande importância escolher o conjunto de hiperparâmetros que otimize o desempenho final do modelo, e em contrapartida essa tarefa frequentemente pode se tornar uma atividade com elevada complexidade. Comumente, o conjunto de hiperparâmetros é definido através de experimentações manuais ou por meio de busca exaustiva (*grid search*), onde dependendo do número de hiperparâmetros e do tamanho do espaço de busca pode levar à um alto custo computacional, aumentando em grandes quantidades o tempo de ajuste do modelo.

O presente trabalho considera dois cenários para o ajuste dos modelos. No primeiro, o aprendizado do algoritmo é feito levando em consideração todas as variáveis disponíveis para serem utilizadas como preditoras, gerando assim um modelo com uma configuração ótima de hiperparâmetros para cada estação meteorológica. Já no segundo cenário, utilizando a configuração ótima de hiperparâmetros encontrado anteriormente, um novo lote de modelos são gerados na tentativa de modelar o conjunto de dados referente a cada estação, porém dessa vez, realizando uma seleção de atributos a fim de diminuir a

complexidade através da eliminação de variáveis pouco representativas para o problema em si.

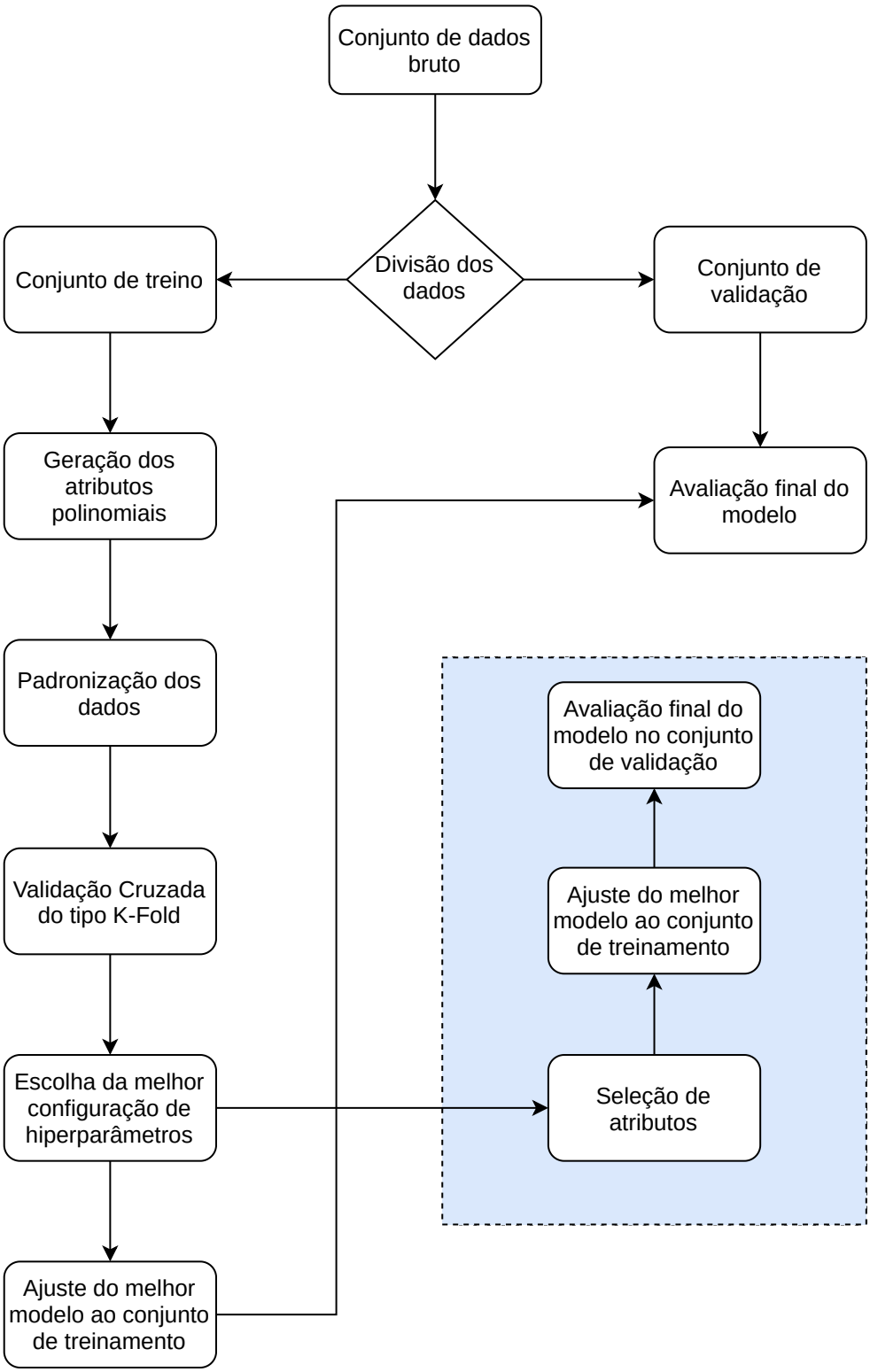
A Figura 13 traz o paradigma da concepção e funcionamento da modelagem proposta no trabalho para realizar a estimação da carga de radiação solar diária tomando como base as variáveis climáticas. O treinamento realizado no trabalho, considera cada conjunto de dados separadamente ou seja, para cada uma das estações meteorológicas estudada um modelo em específico foi criado, portanto, os passos descritos a seguir foram realizados para oito conjuntos diferentes, cada um respectivo à uma região.

Em um primeiro momento os dados em sua forma bruta que estão compreendidos dentro do período de 01/01/1998 a 31/12/2012 totalizando quinze anos de observações diárias é dividido em dois grupos: o conjunto para treinamento $1998 - 2008$ (11 anos) e o conjunto para validação $2008 - 2012$ (4 anos). Realizado a separação inicial dos dados, o conjunto de treinamento é submetido à um *pipeline* onde os atributos polinomiais são criados com base nas variáveis climáticas da Tabela 3, em seguida padronizados e submetidos à validação cruzada utilizando a metodologia K-fold. Todo esse *pipeline* é realizado e testado através de uma busca exaustiva em torno do espaço de hiperparâmetros pré-definido, na intenção de encontrar uma configuração ótima que forneça maiores níveis de performance aos modelos. Depois de escolhido o melhor conjunto de hiperparâmetros por meio de análise de desempenho da métrica RMSE, o modelo é retreinado sobre todo o conjunto de dados e então é avaliado sobre os dados inicialmente separados para validação.

Tomando como base as configurações dos melhores modelos criados, os mesmos hiperparâmetros são reutilizados novamente para uma nova bateria de treinamento em um segundo cenário. Todo o processo é refeito porém agora, também é incluída uma nova etapa de seleção de variáveis com base na análise da informação mútua entre os preditores, na tentativa de reduzir o número de variáveis irrelevantes ou pouco representativas simplificando assim o modelo. A Tabela 4 apresenta os hiperparâmetros utilizados ao longo do *framework* e também os valores empregados para o espaço de busca de cada uma.

Para fins de realização do experimento utilizando o modelo regressor Ridge, os hiperparâmetros envolvidos no processo são detalhados mais à frente, sendo eles: $n\_splits$, $alpha$, $poly\_degree$, $poly\_interaction\_only$, $poly\_include\_bias$ e $features\_select\_k$.

Figura 13 - Fluxograma detalhado das etapas de funcionamento da estratégia proposta no trabalho, desde o conjunto de dados bruto até a consolidação e avaliação final do modelo.




Fonte: Elaborado pelo autor (2021).

Para o uso da metodologia de validação cruzada K-Fold o valor de $k$ (número de compartimentos para divisão) utilizado está definido pelo parâmetro $n\_splits$. O uso do valor $k = 10$ é muito comum na área aplicada de ciência de dados, sendo comumente um bom ponto de partida para análise da relação de viés e variância dos modelos. Quanto ao regressor, o único hiperparâmetro a ser definido é o *alpha*, responsável por definir a intensidade do impacto causado pelo termo de regularização, valores elevados especificam uma forte regularização dos coeficientes. Quanto à criação de atributos polinomiais $poly\_degree$ é responsável por definir o grau do polinômio de expansão dos dados de entrada, $poly\_interaction\_only$ quando definida como $True$ produz apenas atributos onde há a interação entre as variáveis, descartando assim os termos de potência. E por último $poly\_include\_bias$ é encarregado de indicar se deve ou não ser criado uma coluna de atributo na qual todas potências polinomiais são zero. Na etapa de seleção de variáveis, a informação mútua é responsável por pontuar cada variável e $features\_select\_k$ determina o número de preditores do ranking a serem levados em consideração para a construção do modelo.

Tabela 4 – Conjunto dos hiperparâmetros utilizados para a modelagem.

| **Parâmetros** | **Valores** |
|---|---|
| n_splits | $k = 10$ |
| alpha | [0, 1] |
| poly_degree | [1, 2, 3, 4] |
| poly_interaction_only | [True, False] |
| poly_include_bias | [True, False] |
| features_select_k | [30, 60, 90, 120, 150, 180, 210, 240, 270, 300] |

Fonte: Elaborada pelo autor (2021).

## 3.8 MÉTRICAS DE AVALIAÇÃO

Para fins de avaliação e comparação do desempenho entre os modelos desenvolvidos e os presentes na literatura, as seguintes métricas foram utilizadas. A raiz do erro quadrático médio, o erro médio absoluto e o coeficiente de determinação $R^2$ foram utilizados tanto durante a fase de treinamento quanto na fase de avaliação final, enquanto que a variância contabilizada foi utilizada apenas na etapa de avaliação final.

O coeficiente de determinação também conhecido como $R^2$ (R-quadrado) cuja equação é apresentada em 3.12, é uma medida estatística que indica o quão próximos os dados estão da linha de regressão ajustada por um modelo linear generalizado. Em outras palavras, sua definição consiste na porcentagem da variação da variável resposta que é explicada por um modelo linear. Essa métrica de avaliação, assume valores dentro do intervalo de $[0, 1]$, onde, quanto maior o valor assumido, maior capacidade explicativa o

modelo apresenta [50].

$$R^2 = 1 - \frac{\sum\limits_i (y_i - \widehat{y}_i)^2}{\sum\limits_i (y_i - \overline{y})^2} \tag{3.12}$$

A raiz do erro quadrático médio (RMSE, do inglês *Root Mean Squared Error*) é uma métrica de erro amplamente utilizada que fornece a medida das diferenças entre os valores estimados por um modelo e os valores realmente observados. No contexto de estimação da carga de radiação solar, o RMSE é considerada a métrica de avaliação mais frequentemente utilizada, estando presente em diversos trabalhos. Essa métrica indica a adequação absoluta do modelo aos dados, ou seja, o quão próximo as observações estão dos valores previstos pelo modelo, portanto, quanto mais próximo de zero, maior é a performance de estimação do modelo. Considerada como principal métrica de avaliação para o trabalho, o RMSE foi utilizado como métrica prioritária para a escolha de melhor modelo durante a etapa de validação cruzada. Adicionalmente, dada pela raiz quadrada de uma variação, a métrica pode ser interpretada como o desvio padrão da variação e possui a propriedade de estar na mesma unidade que a variável de saída.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2} \tag{3.13}$$

O erro médio absoluto (MAE, do inglês *Mean Absolute Error*), diferentemente do RMSE indica a média de erros absolutos de maneira direta, não introduzindo penalização em casos em que a diferença entre o valor medido e previsto são maiores. Semelhantemente, quanto mais próximo de zero seu valor for apresentado, melhor o resultado entregue pelo modelo avaliado.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \widehat{y}_i| \tag{3.14}$$

A variância contabilizada (VAF, do inglês *Variance Accounted For*) usualmente fornecida em valores percentuais, indica em resumo o quanto da variabilidade dos dados pode ser explicada por um modelo de regressão ajustado [51]. A VAF pode ser calculada pela expressão presente em 3.15.

$$VAF = \left(1 - \frac{var(y_i - \widehat{y}_i)}{var(y_i)}\right) \times 100 \tag{3.15}$$

# 4 RESULTADOS E DISCUSSÃO

Para cada um dos experimentos computacionais apresentados nesse capítulo, os mesmos foram realizados separadamente para os oito diferente conjunto de dados, cada um referente à uma estação meteorológica em específico. Adicionalmente, cada experimento foi executado 30 vezes de forma totalmente independente. O código para análise dos resultados e bem como os algoritmos foram implementados na linguagem Python na sua versão 3.8.3 e fazendo uso de bibliotecas tais como Scikit-learn [53] para aprendizado de máquina, Numpy [54] e Pandas [55] para manipulação dos dados e Matplotlib [56] para gerar figuras e gráficos. Os experimentos foram realizados utilizando o ambiente de computação em nuvem Google Colaboratory com as seguintes especificações de hardware: CPU Intel(R) Xeon(R) (2 núcleos de 2.30GHz e 46MB de Cache), memória RAM de 12GB e sistema operacional Linux Ubuntu.

Os resultados são apresentados de acordo com as duas etapas de ajuste previamente citados. Na etapa (I) o aprendizado do algoritmo é feito levando em consideração todas as variáveis disponíveis para serem utilizadas como preditoras, já na segunda etapa (II) é inserido no pipeline um módulo responsável pela seleção de atributos, afim de tentar diminuir o número de variáveis a serem levadas em conta.

## 4.1 ETAPA I

A Tabela 5 apresenta um resumo geral dos resultados obtidos. É exibido a média e o desvio padrão das métricas estatísticas obtidas ao longo das 30 execuções independentes para cada conjunto de dados referente a cada uma das oito estações. Para fins de análise, para as métricas RMSE e MAE valores menores indicam um melhor desempenho do modelo, enquanto que para R2 os melhores modelos devem apresentar seus valores próximos à 1, e por último VAF próximo do valor de 100% indica maior performance. Pela observação dos valores de cada métrica apresentados na Tabela 5 é possível identificar que a metodologia proposta apresentou melhores resultados para a estação VI. As localidades em I, II, V e VIII tiveram resultados semelhantes, dentro da faixa de $0.5 - 0.6$ de RMSE. E de maneira destoante a estação de número IV demonstrou o pior resultado, apresentando uma baixa performance com mais de 350% em valor de RMSE quando comparado à média das outras estações e 0.58 de coeficiente de determinação R2.

Tabela 5 – Média das métricas estatísticas obtidas ao longo das 30 execuções independentes para cada estação meteorológica. Os valores dos desvios padrão são mostrados entre parênteses.

| **Estação** | **Modelo** | **RMSE** ($MJ/m^2$) | **MAE** ($MJ/m^2$) | **R2** | **VAF** |
|---|---|---|---|---|---|
| **I** | Ridge | 0.571 (0.005) | 0.341 (0.001) | 0.979 (0.0004) | 98.106 (0.047) |
| | Lasso | 0.579 (0.000) | 0.401 (0.000) | 0.975 (0.000) | 97.610 (0.000) |
| | Árvore de Decisão | 1.360 (0.015) | 0.971 (0.007) | 0.861 (0.003) | 86.699 (0.297) |
| | Floresta Aleatória | 0.892 (0.004) | 0.621 (0.003) | 0.940 (0.001) | 94.690 (0.055) |
| **II** | Ridge | 0.503 (0.008) | 0.347 (0.003) | 0.981 (0.0005) | 98.438 (0.063) |
| | Lasso | 0.526 (0.001) | 0.372 (0.000) | 0.980 (0.000) | 98.352 (0.003) |
| | Árvore de Decisão | 1.274 (0.019) | 0.924 (0.010) | 0.884 (0.003) | 88.644 (0.332) |
| | Floresta Aleatória | 0.834 (0.003) | 0.581 (0.002) | 0.950 (0.000) | 95.469 (0.037) |
| **III** | Ridge | 0.767 (0.004) | 0.530 (0.003) | 0.954 (0.0004) | 96.567 (0.071) |
| | Lasso | 0.825 (0.000) | 0.593 (0.000) | 0.948 (0.000) | 95.985 (0.000) |
| | Árvore de Decisão | 1.519 (0.013) | 1.110 (0.007) | 0.823 (0.003) | 84.175 (0.279) |
| | Floresta Aleatória | 1.166 (0.003) | 0.861 (0.003) | 0.895 (0.000) | 91.783 (0.042) |
| **IV** | Ridge | 2.440 (0.005) | 1.862 (0.001) | 0.580 (0.001) | 64.816 (0.375) |
| | Lasso | 2.448 (0.000) | 1.871 (0.001) | 0.578 (0.000) | 64.682 (0.041) |
| | Árvore de Decisão | 3.578 (0.045) | 2.584 (0.024) | 0.098 (0.023) | 17.655 (2.310) |
| | Floresta Aleatória | 2.465 (0.006) | 1.908 (0.005) | 0.572 (0.002) | 64.389 (0.161) |
| **V** | Ridge | 0.679 (0.001) | 0.470 (0.001) | 0.966 (0.0001) | 97.558 (0.013) |
| | Lasso | 0.770 (0.000) | 0.544 (0.000) | 0.957 (0.000) | 96.918 (0.000) |
| | Árvore de Decisão | 1.497 (0.018) | 1.063 (0.012) | 0.837 (0.004) | 84.144 (0.383) |
| | Floresta Aleatória | 0.987 (0.004) | 0.708 (0.003) | 0.929 (0.001) | 93.951 (0.052) |
| **VI** | Ridge | 0.390 (0.006) | 0.262 (0.002) | 0.989 (0.0003) | 99.022 (0.040) |
| | Lasso | 0.413 (0.000) | 0.281 (0.000) | 0.988 (0.000) | 98.927 (0.000) |
| | Árvore de Decisão | 1.126 (0.012) | 0.802 (0.008) | 0.911 (0.002) | 91.215 (0.195) |
| | Floresta Aleatória | 0.692 (0.002) | 0.486 (0.002) | 0.966 (0.000) | 96.805 (0.021) |
| **VII** | Ridge | 1.014 (0.004) | 0.715 (0.006) | 0.912 (0.0008) | 93.816 (0.077) |
| | Lasso | 1.065 (0.000) | 0.780 (0.000) | 0.904 (0.000) | 93.762 (0.000) |
| | Árvore de Decisão | 1.956 (0.020) | 1.404 (0.012) | 0.676 (0.007) | 71.273 (0.565) |
| | Floresta Aleatória | 1.358 (0.004) | 1.043 (0.003) | 0.844 (0.001) | 88.430 (0.062) |
| **VIII** | Ridge | 0.566 (0.0003) | 0.385 (0.0004) | 0.977 (0.0002) | 98.102 (0.005) |
| | Lasso | 0.590 (0.000) | 0.416 (0.000) | 0.975 (0.000) | 97.992 (0.000) |
| | Árvore de Decisão | 1.293 (0.015) | 0.909 (0.009) | 0.881 (0.003) | 88.328 (0.270) |
| | Floresta Aleatória | 0.887 (0.004) | 0.612 (0.003) | 0.944 (0.001) | 94.831 (0.050) |

Fonte: Elaborada pelo autor (2021).

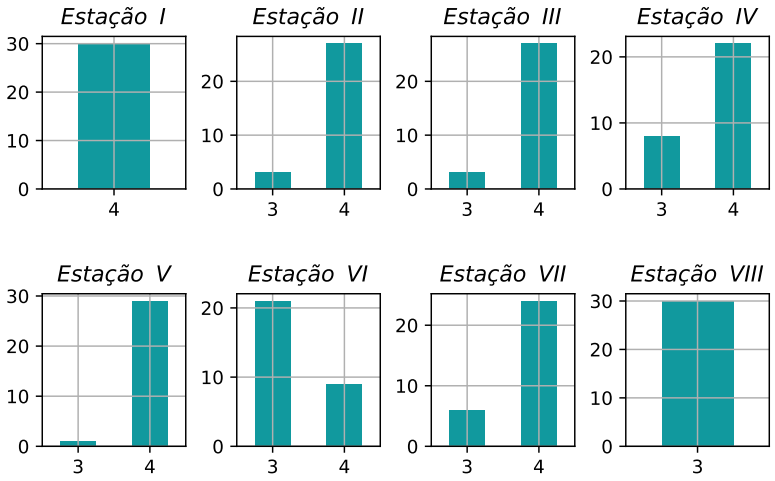
Para fins de comparação com outros modelos e abordagens de aprendizado de máquina, na Tabela 5 também é apresentado o resultado das mesmas métricas para outras três abordagens. A primeira é também uma regressão porém com utilização da norma-L1 como termo regularizador. Além do modelo Lasso ter sido utilizado para fins comparativos, também foi inserido um modelo de árvore de decisão e um modelo de floresta aleatória.

Tais algoritmos subdividem progressivamente os dados em conjuntos cada vez menores e mais específicos, em termos de suas variáveis, até atingirem um tamanho simplificado o bastante para serem rotulados. Esses últimos dois são abordagens comumente utilizadas, principalmente por sua facilidade de interpretação do modelo.

E possível observar que de acordo com os resultados obtidos, o modelo Ridge com seus hiperparâmetros otimizados consegue ter uma performance superior às outras três abordagens para a maioria das regiões. Na sequência, o modelo regularizado Lasso ocupa o segundo lugar no ranking de performance, sendo então seguido pelos modelos baseados em árvores. É possível observar que o método de regularização contribui fortemente para uma melhor generalização do modelo sobre os dados de validação, a prova disso é que a regressão Lasso apresenta métricas de RMSE e MAE até duas vezes menores do que os apresentados pelos modelos baseados em árvore.

As próximas quatro figuras apresentam as distribuições dos parâmetros que compõem o pipeline de ajuste dos modelos para cada uma das estações ao longo das 30 execuções realizadas. A Figura 14 mostra a adoção dos valores para o grau do polinômio a ser utilizado para a expansão dos dados de entrada, o parâmetro *poly_degree*. Na maior parte das execuções e para quase todas as estações o valor retornado pela busca que melhor otimiza o desempenho dos modelos é o valor de grau igual a 4. A única exceção é para a estação VI e VIII, onde na maioria das execuções o valor de grau igual a 3 é selecionado.

Figura 14 - Distribuição do parâmetro *poly_degree* utilizado para a expansão dos dados de entrada ao longo das 30 execuções independentes do procedimento.

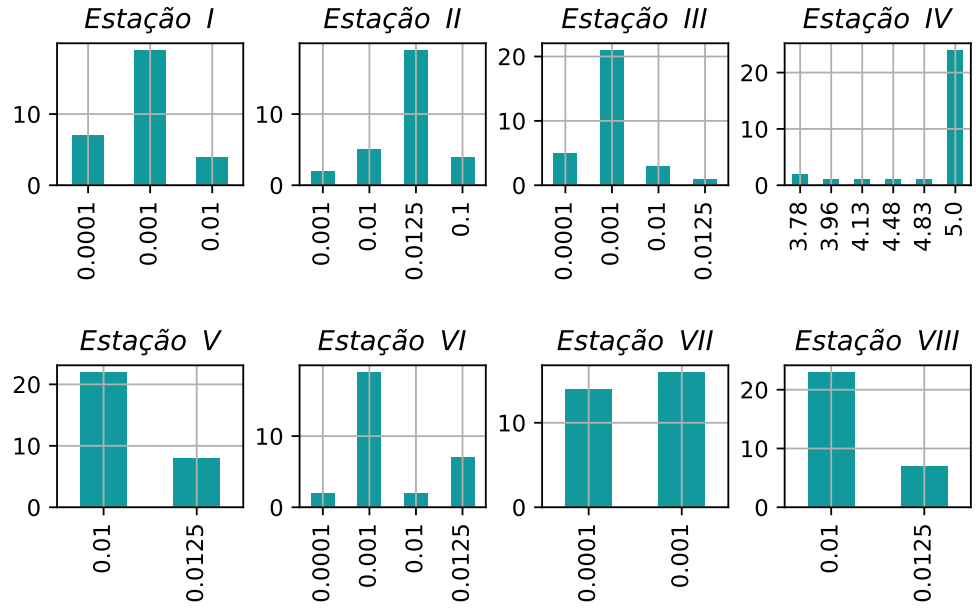



Fonte: Elaborado pelo autor (2021).

Já a Figura 15 apresenta a distribuição do parâmetro *alpha*, responsável por definir

a intensidade da regularização incluída na regressão. Em contraste às demais, que na maioria das execuções selecionaram valores menores e próximos de zero para o coeficiente alfa, a estação IV retornou valores consideravelmente mais elevados indicando a necessidade do uso de de regularização das variáveis de forma mais intensificada.

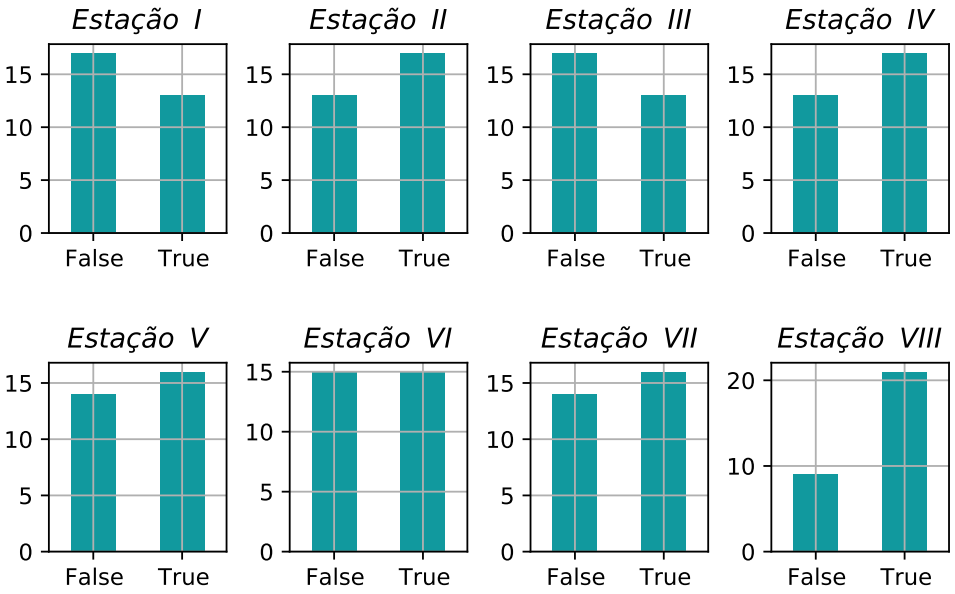
Figura 15 - Distribuição do parâmetro *alpha* utilizado para regular a penalização das variáveis ao longo das 30 execuções independentes do procedimento.




Fonte: Elaborado pelo autor (2021).

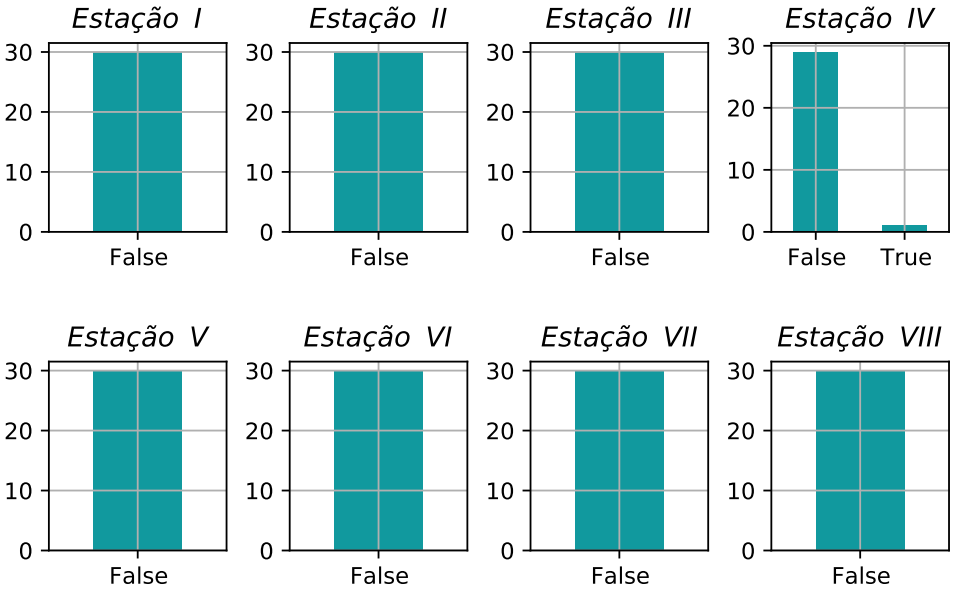
As duas próximas figuras mostram os resultados obtidos nas simulações quanto a escolha dos dois últimos parâmetros. Na Figura 16 tem-se a distribuição do parâmetro *poly_include_bias* e na Figura 17 a distribuição do *poly_interaction_only*. O primeiro apresentou uma frequência entre as duas opções (ligado ou desligado) muito próximas, indicando dessa forma que esse fator possui pouca influencia quanto a sua contribuição à performance do modelo, tornando-se assim indiferente para o estudo. Já no segundo gráfico o parâmetro *poly_interaction_only* unanimemente entre todas as estações foi escolhido como *False*, ou seja considerando a expansão polinomial dos dados de entrada por completo e não somente os termos no qual existe a interação entre duas ou mais variáveis.

Figura 16 - Distribuição do parâmetro *poly_include_bias* ao longo das 30 execuções independentes do procedimento.




Fonte: Elaborado pelo autor (2021).

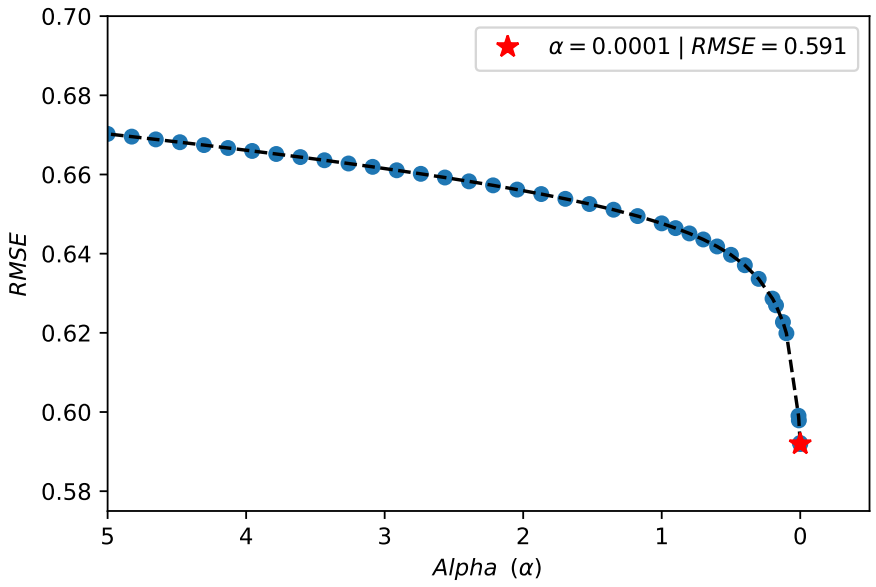
Figura 17 - Distribuição do parâmetro *poly_interaction_only* ao longo das 30 execuções independentes do procedimento.




Fonte: Elaborado pelo autor (2021).

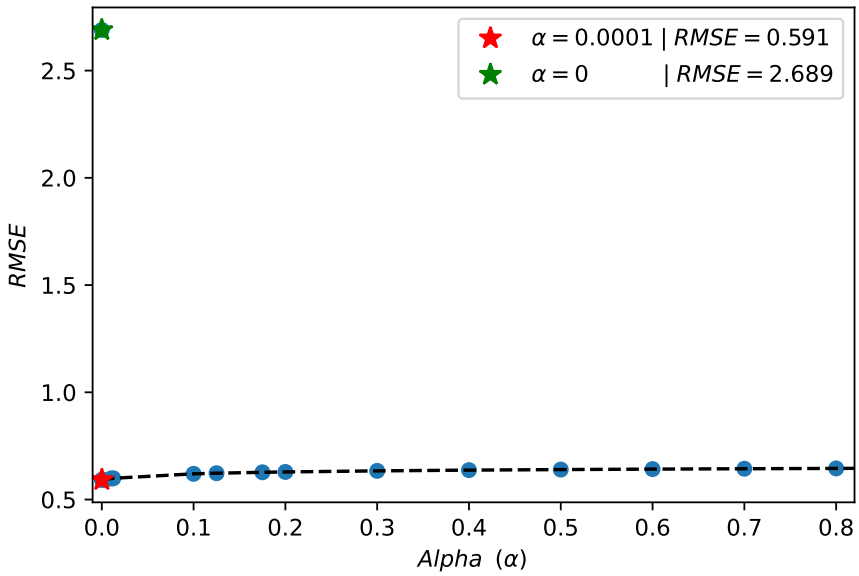
As Figuras 18 e 19 mostram o impacto do parâmetro de regularização sobre o modelo, indicando as variações de valores da métrica estatística de raiz do erro quadrado médio (RMSE) durante a etapa de treinamento em função dos valores assumidos por alfa. Ambos gráficos são referentes à fase de treinamento do modelo para a estação I e os mesmos consideram a variação do parâmetro alfa enquanto os outros parâmetros estão constantes, dessa forma os resultados apresentados na figura capturam somente a contribuição e interferência de alfa no regressor. O resultado apresentado valida a ideia da utilização de regularização para o problema estudado indicando a melhora de performance em função da adição do fator de penalização. A estrela em vermelho indica o ponto com o valor ótimo de alfa encontrado dentre o espaço de busca pré-definido e seus respectivo valor de RMSE, enquanto a estrela em verde na Figura 19 representa o ponto em que $\alpha = 0$, ou seja, quando não há a existência do fator de regularização tornando o modelo em uma regressão linear normal. O fato de existir o termo de penalização consegue prover ao modelo uma melhora significativa, reduzindo o RMSE de 2.689 $MJ/m^2$ para 0.591 $MJ/m^2$ para esse caso em específico. Essa melhora de desempenho devido à inserção do fator de regularização também é observada em todos os outros conjuntos de dados das outras estações.

Figura 18 - Variação da métrica de erro RMSE em função dos diferentes valores assumidos pelo parâmetro de regularização alfa ($\alpha$) para a estação I durante a fase de treinamento do modelo.




Fonte: Elaborado pelo autor (2021).

Figura 19 - Variação do RMSE em função dos diferentes valores assumidos pelo parâmetro de regularização alfa ($\alpha$) para a estação I durante a fase de treinamento do modelo. Diferença entre a regressão com e sem o fator de regularização ($\alpha = 0$).




Fonte: Elaborado pelo autor (2021).

Visando a inserção do módulo responsável por realizar a seleção das variáveis, os valores dos hiperparâmetros encontrados na primeira etapa que garantem melhores valores para as métrica de erro utilizadas são mantidos fixos, e então, na segunda etapa os modelos são treinados novamente porém agora com o novo componente responsável por avaliar a contribuição de cada variável e realizar a redução no número de variáveis. É importante ressaltar que, apesar do modelo considerar diferentes números de variáveis, ao final do processo de ajuste é mantido a configuração que resultou em menor erro. Dessa forma, torna-se função do projetista trabalhar com o custo-benefício entre o erro e o número de variáveis, em outras palavras, decidir em quanto reduzir as variáveis em troca de abrir mão de desempenho.

## 4.2 ETAPA II

Mantendo as configurações de parâmetros de acordo com o encontrado na primeira etapa, a seleção de variáveis utilizando a informação mútua entra no pipeline para dar suporte à análise de interferência do número de variáveis nos modelos. A Tabela 6 mostra os resultados obtidos após a inclusão desse novo módulo. Os resultados apresentados na tabela são muito próximos aos obtidos na primeira etapa. Essa pequena variação das

métricas nesse segundo resultado é explicada pelo fato de que em muitos casos o modelo optou por manter considerando à maioria das variáveis, visto que a performance em se tratando de RMSE seria diminuída. As únicas duas exceções foram na estação I, onde em 20 das 30 execuções independentes o modelo obteve melhores resultados considerando o número das 270 variáveis de maior importância, e em IV onde em 15 execuções o modelo apresentou melhor performance descartando um número maior de variáveis e considerando apenas as 180 mais significativas.

Tabela 6 – Média das métricas estatísticas obtidas ao longo das 30 execuções independentes para cada estação meteorológica após a inserção do módulo de seleção de variáveis. Os valores dos desvios padrão são mostrados entre parênteses.

| **Estação** | **Modelo** | **RMSE (MJ/m²)** | **MAE (MJ/m²)** | **R2** | **VAF** |
|---|---|---|---|---|---|
| **I** | Ridge | 0.517 (0.001) | 0.342 (0.000) | 0.979 (0.000) | 98.113 (0.010) |
| | Lasso | 0.579 (0.000) | 0.401 (0.000) | 0.975 (0.000) | 97.610 (0.000) |
| | Árvore de Decisão | 1.360 (0.015) | 0.971 (0.007) | 0.861 (0.003) | 86.699 (0.297) |
| | Floresta Aleatória | 0.892 (0.004) | 0.621 (0.003) | 0.940 (0.001) | 94.690 (0.055) |
| **II** | Ridge | 0.502 (0.000) | 0.348 (0.000) | 0.981 (0.000) | 98.447 (0.003) |
| | Lasso | 0.526 (0.001) | 0.372 (0.000) | 0.980 (0.000) | 98.352 (0.003) |
| | Árvore de Decisão | 1.274 (0.019) | 0.924 (0.010) | 0.884 (0.003) | 88.644 (0.332) |
| | Floresta Aleatória | 0.834 (0.003) | 0.581 (0.002) | 0.950 (0.000) | 95.469 (0.037) |
| **III** | Ridge | 0.773 (0.001) | 0.531 (0.000) | 0.954 (0.000) | 96.498 (0.012) |
| | Lasso | 0.825 (0.000) | 0.593 (0.000) | 0.948 (0.000) | 95.985 (0.000) |
| | Árvore de Decisão | 1.519 (0.013) | 1.110 (0.007) | 0.823 (0.003) | 84.175 (0.279) |
| | Floresta Aleatória | 1.166 (0.003) | 0.861 (0.003) | 0.895 (0.000) | 91.783 (0.042) |
| **IV** | Ridge | 2.435 (0.004) | 1.860 (0.003) | 0.582 (0.001) | 64.983 (0.159) |
| | Lasso | 2.448 (0.000) | 1.871 (0.001) | 0.578 (0.000) | 64.682 (0.041) |
| | Árvore de Decisão | 3.578 (0.045) | 2.584 (0.024) | 0.098 (0.023) | 17.655 (2.310) |
| | Floresta Aleatória | 2.465 (0.006) | 1.908 (0.005) | 0.572 (0.002) | 64.389 (0.161) |
| **V** | Ridge | 0.675 (0.001) | 0.470 (0.001) | 0.966 (0.000) | 97.624 (0.009) |
| | Lasso | 0.770 (0.000) | 0.544 (0.000) | 0.957 (0.000) | 96.918 (0.000) |
| | Árvore de Decisão | 1.497 (0.018) | 1.063 (0.012) | 0.837 (0.004) | 84.144 (0.383) |
| | Floresta Aleatória | 0.987 (0.004) | 0.708 (0.003) | 0.929 (0.001) | 93.951 (0.052) |
| **VI** | Ridge | 0.402 (0.009) | 0.266 (0.004) | 0.988 (0.000) | 98.944 (0.052) |
| | Lasso | 0.413 (0.000) | 0.281 (0.000) | 0.988 (0.000) | 98.927 (0.000) |
| | Árvore de Decisão | 1.126 (0.012) | 0.802 (0.008) | 0.911 (0.002) | 91.215 (0.195) |
| | Floresta Aleatória | 0.692 (0.002) | 0.486 (0.002) | 0.966 (0.000) | 96.805 (0.021) |
| **VII** | Ridge | 1.021 (0.002) | 0.716 (0.000) | 0.911 (0.000) | 93.724 (0.030) |
| | Lasso | 1.065 (0.000) | 0.780 (0.000) | 0.904 (0.000) | 93.762 (0.000) |
| | Árvore de Decisão | 1.956 (0.020) | 1.404 (0.012) | 0.676 (0.007) | 71.273 (0.565) |
| | Floresta Aleatória | 1.358 (0.004) | 1.043 (0.003) | 0.844 (0.001) | 88.430 (0.062) |
| **VIII** | Ridge | 0.578 (0.005) | 0.398 (0.003) | 0.976 (0.000) | 98.047 (0.050) |
| | Lasso | 0.590 (0.000) | 0.416 (0.000) | 0.975 (0.000) | 97.992 (0.000) |
| | Árvore de Decisão | 1.293 (0.015) | 0.909 (0.009) | 0.881 (0.003) | 88.328 (0.270) |
| | Floresta Aleatória | 0.887 (0.004) | 0.612 (0.003) | 0.944 (0.001) | 94.831 (0.050) |

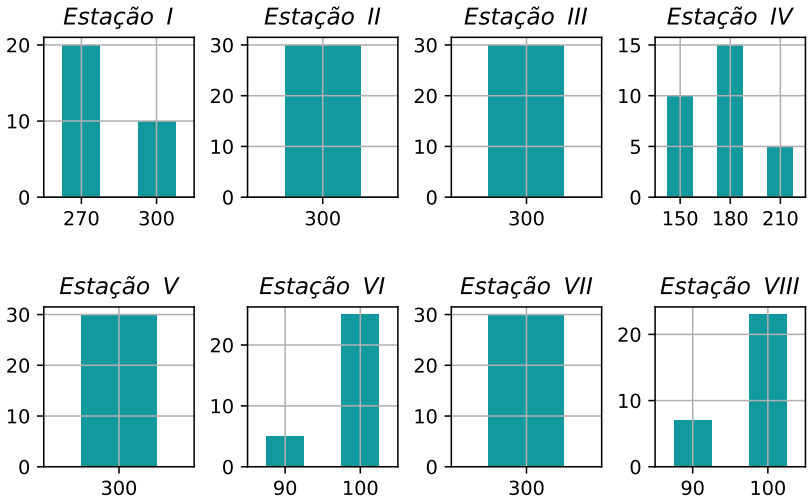
Fonte: Elaborada pelo autor (2021).

A Figura 20 mostra a distribuição do número de variáveis $k$ selecionado por cada modelo referente à cada uma das estações ao longo das 30 execuções independentes. O gráfico mostra o comportamento previamente comentado, onde na maioria dos casos o modelo continua considerando o maior número de variáveis por dessa forma gerar melhores resultados. Mais adiante os gráficos mostram o impacto de uma redução de variáveis mais

severa em função da raiz do erro quadrático médio apresentado pelos modelos. Para as estações VI e VIII o grau do polinômio utilizado para a expansão dos dados de entrada encontrado são de valor 3, diferentemente dos demais. Portanto essas duas estações por natureza já possuem um número de variáveis reduzido em comparação às demais.

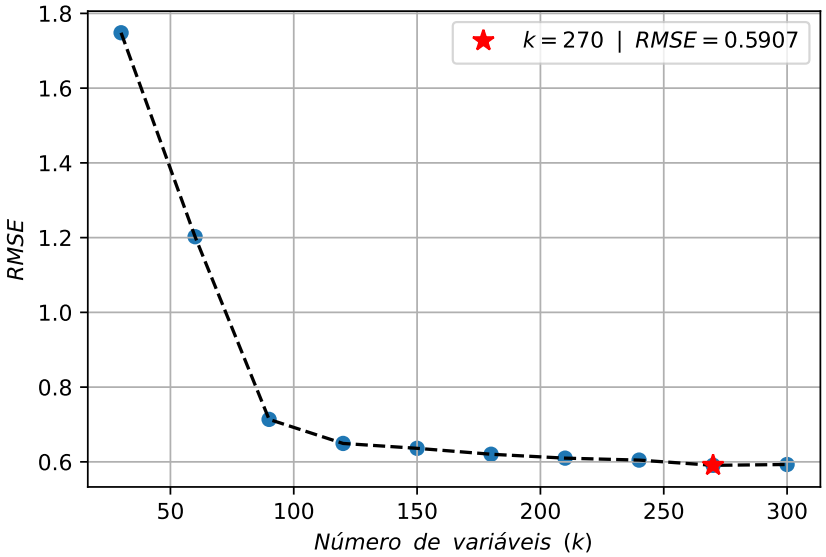
As Figuras 21 a 28 na sequência, apresentam a variação da métrica estatística da raiz do erro quadrado médio em função dos diferentes números de variáveis utilizados em consideração por cada modelo para as oito estações meteorológicas. O ponto nos gráficos representados pela estrela em vermelho indica a configuração que resultado no menor valor de erro. Tomando como base a análise desses gráficos, é possível definir um limiar aceitável de valor para o erro em favor de reduzir ainda mais o número de regressores levados em conta. Essa tarefa em um primeiro momento seria algo totalmente manual e pode variar muito de acordo com as particularidade intrínsecas à necessidade e área de aplicação do uso da RS.

Figura 20 - Distribuição do parâmetro $features\_select\_k$ ao longo das 30 execuções independentes do procedimento.



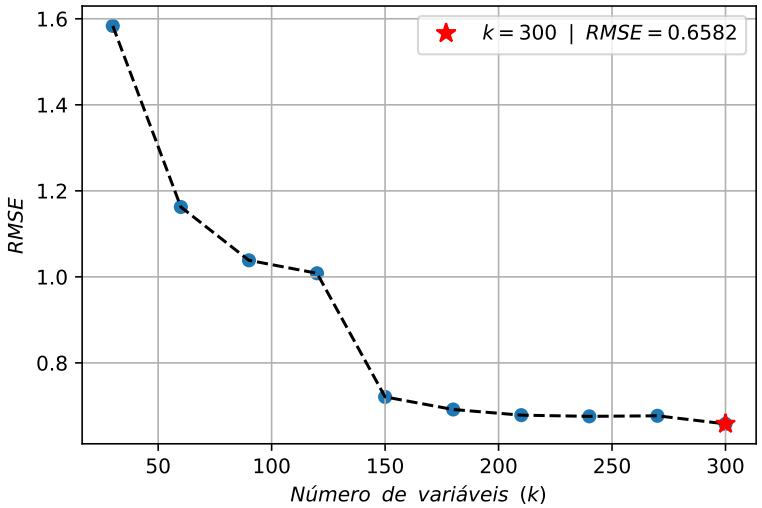

Fonte: Elaborado pelo autor (2021).

Figura 21 - Variação da métrica de erro RMSE de acordo com o número de variáveis levadas em consideração no modelo para o caso da estação I. Em destaque o ponto que apresentou menor valor de erro.



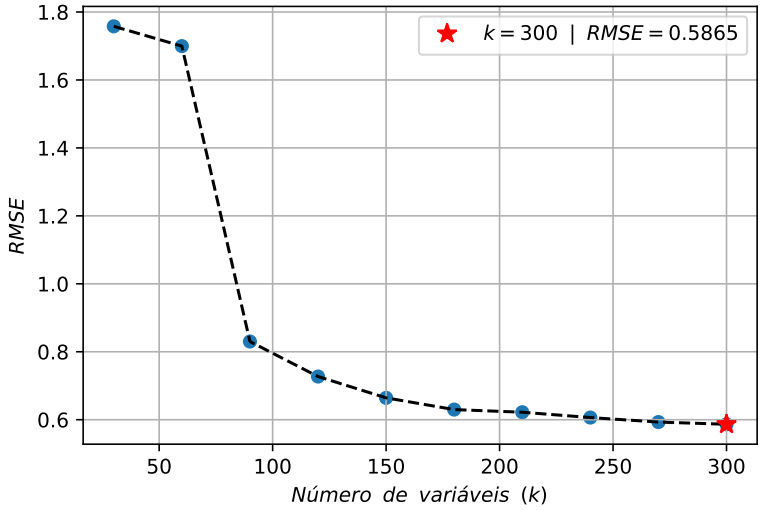

Fonte: Elaborado pelo autor (2021).

Figura 22 - Variação da métrica de erro RMSE de acordo com o número de variáveis levadas em consideração no modelo para o caso da estação II. Em destaque o ponto que apresentou menor valor de erro.



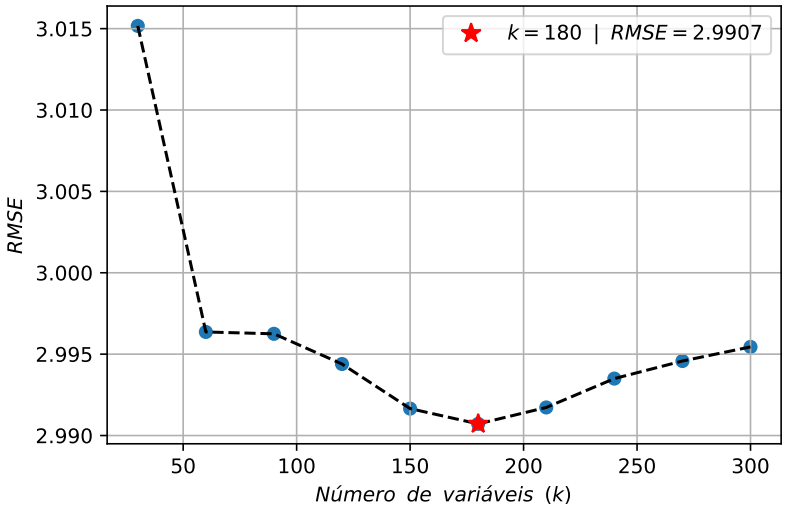

Fonte: Elaborado pelo autor (2021).

Figura 23 - Variação da métrica de erro RMSE de acordo com o número de variáveis levadas em consideração no modelo para o caso da estação III. Em destaque o ponto que apresentou menor valor de erro.



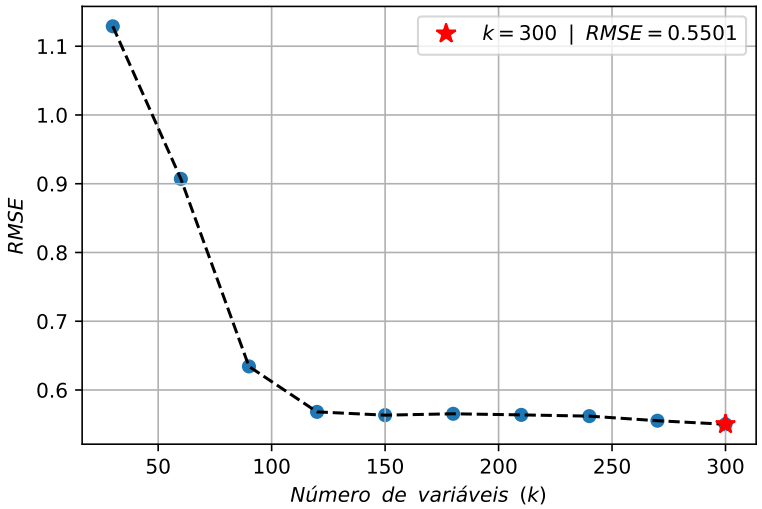

Fonte: Elaborado pelo autor (2021).

Figura 24 - Variação da métrica de erro RMSE de acordo com o número de variáveis levadas em consideração no modelo para o caso da estação IV. Em destaque o ponto que apresentou menor valor de erro.



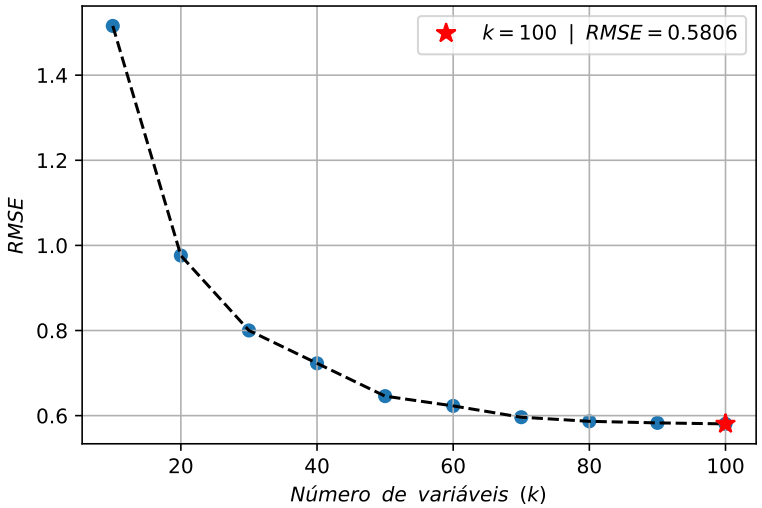

Fonte: Elaborado pelo autor (2021).

Figura 25 - Variação da métrica de erro RMSE de acordo com o número de variáveis levadas em consideração no modelo para o caso da estação V. Em destaque o ponto que apresentou menor valor de erro.



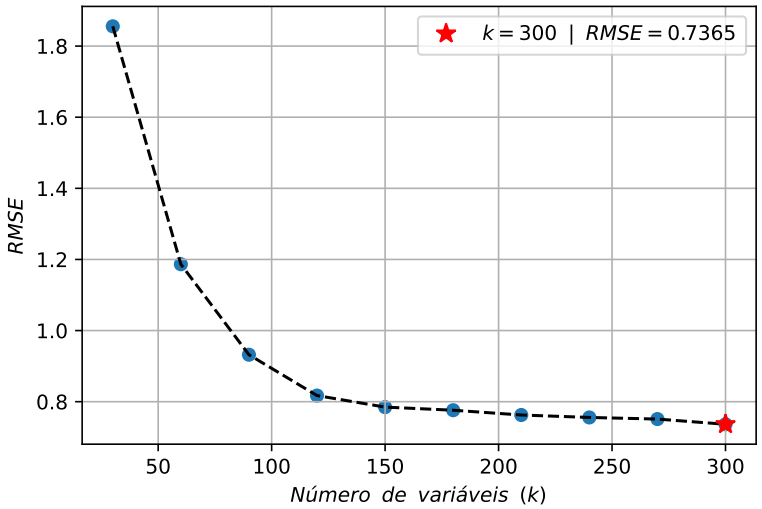

Fonte: Elaborado pelo autor (2021).

Figura 26 - Variação da métrica de erro RMSE de acordo com o número de variáveis levadas em consideração no modelo para o caso da estação VI. Em destaque o ponto que apresentou menor valor de erro.



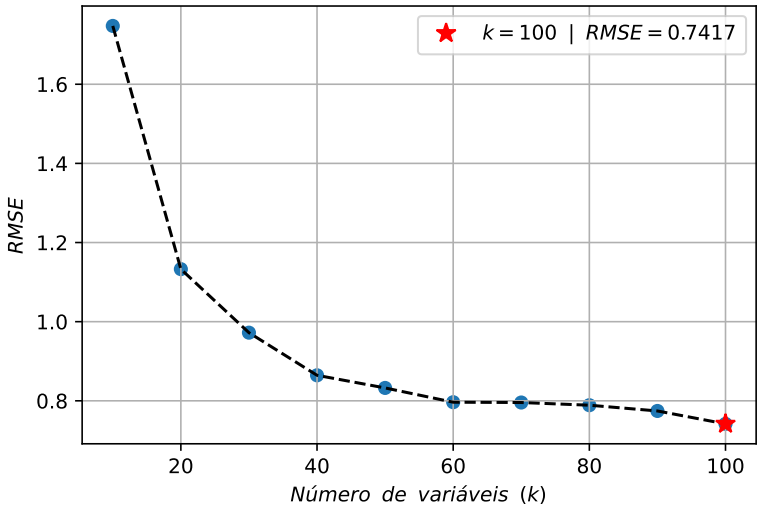

Fonte: Elaborado pelo autor (2021).

Figura 27 - Variação da métrica de erro RMSE de acordo com o número de variáveis levadas em consideração no modelo para o caso da estação VII. Em destaque o ponto que apresentou menor valor de erro.




Fonte: Elaborado pelo autor (2021).

Figura 28 - Variação da métrica de erro RMSE de acordo com o número de variáveis levadas em consideração no modelo para o caso da estação VIII. Em destaque o ponto que apresentou menor valor de erro.




Fonte: Elaborado pelo autor (2021).

Para validação dos resultados encontrados pelos modelos criados no presente trabalho, a Tabela 7 mostra um comparativo com os resultados reportados em estudo previamente publicado em [11]. O estudo usado como comparativo faz uso de um modelo com maior complexidade, uma máquina de aprendizagem extrema evolutiva e auto-adaptativa e também considera o mesmo conjunto de dados de Burkina Faso, porém, reportou os resultados referentes à apenas quatro das oito estações.

Os valores apresentados na tabela abaixo são referentes aos encontrados após a realização da seleção de variáveis da segunda etapa e são apresentados em primeira estância com o desvio padrão dentro de parênteses, abaixo em negrito é mostrado o valor reportado em [11]. Com relação ao coeficiente de determinação R2 os resultados mostram que houve um alto grau de colinearidade entre os valores medidos e os preditos pelo modelo, apresentando valores acima de 0.97, com exceção da estação IV. Com relação a RMSE e MAE os modelos desenvolvidos nesse trabalho apresentaram resultados superiores aos publicados no trabalho prévio, resultando em valores menores de erro, chegando em uma taxa de redução de até 50% dos valores para as estações VI e VIII. Já em contraste, para a estação IV as métricas de performance tiveram valores muito próximos e com pouca variação.

Tabela 7 – Comparativo entre as métricas estatísticas obtidas no trabalho com os valores previamente publicados na literatura em [11]. Os valores dos desvios padrão são mostrados entre parênteses e em negrito os valores encontrados na literatura.

| **Estação** | **RMSE** ($MJ/m^2$) | **MAE** ($MJ/m^2$) | **R2** | **VAF** |
|---|---|---|---|---|
| II | 0.502 (0.000) **[0.722]** | 0.348 (0.000) **[0.544]** | 0.981 (0.000) **[0.982]** | 98.447 (0.003) **[96.492]** |
| IV | 2.435 (0.004) **[2.576]** | 1.860 (0.003) **[1.999]** | 0.582 (0.001) **[0.785]** | 64.983 (0.159) **[61.653]** |
| VI | 0.402 (0.009) **[0.887]** | 0.266 (0.004) **[0.656]** | 0.988 (0.000) **[0.973]** | 98.944 (0.052) **[94.652]** |
| VIII | 0.578 (0.005) **[1.175]** | 0.398 (0.003) **[0.868]** | 0.976 (0.000) **[0.954]** | 98.047 (0.050) **[91.065]** |

Fonte: Elaborada pelo autor (2021).

De um modo geral, comparando as quatro estações é possível concluir que o modelo proposto fazendo uso de expansão polinomial dos dados de entrada e da regularização do tipo L2 sugere uma capacidade melhor de estimação da radiação solar. Esse resultado verifica a estratégia adotada pelo trabalho em se utilizar um modelo simples porém aumentando a complexidade em torno dos dados além de incluir o mecanismo responsável

por realizar a penalização de variáveis cujo pouco têm a contribuir para a capacidade de estimativa final do modelo.

# 5 CONCLUSÕES E TRABALHOS FUTUROS

Ao longo do presente trabalho, a importância de se conhecer a intensidade de radiação solar incidente sobre uma determinada região foi destacada e bem como as suas contribuições e impacto em diversas áreas de estudo ou até mesmo em aplicações práticas como no projeto de chaminés térmicas, análise de conforto térmico em edifícios, projetos de arquitetura e até mesmo modelos de crescimento de safras.

Dada a importância do recurso proveniente do sol, o trabalho propôs uma alternativa simples e de baixo custo para realizar a estimação da intensidade da radiação solar tomando como base variáveis climáticas de maior acessibilidade. A estimação é feita com base em algoritmos e técnicas pertinentes à área de aprendizado de máquina. A estimação da radiação solar foi realizada para oito diferentes localidades ao longo da Burkina Faso, onde cada localidade refere-se à uma estação meteorológica constituindo assim um conjunto de dados em específico. A estratégia aqui utilizada consiste em aumentar o número de variáveis de entrada por meio de uma expansão polinomial, objetivando capturar as interações entre as variáveis. Na sequencia os novos dados são utilizados para o ajuste de um regressor associado à um mecanismo de regularização capaz de penalizar variáveis pouco influentes para a capacidade de estimação do modelo. Os experimentos foram divididos em duas etapas, em que na primeira há um ajuste dos hiperparâmetros e na sequência a segunda etapa adiciona um módulo para seleção de variáveis.

Seguindo as diretrizes definidas através dos objetivos específicos apresentados ainda na introdução do trabalho, foi possível verificar de acordo com os valores das métricas de avaliação, que o aumento da complexidade dos conjuntos de dados junto ao uso da regressão Ridge apresentaram resultados satisfatórios em comparação com os valores encontrados na literatura, tornando-se uma alternativa atrativa para o tratamento desse tipo de problema. Através do uso dos dados em sua forma expandida, foi possível capturar informações extras que contribuem para uma melhor modelagem do problema. Para as quatro estações comparadas, o modelo desenvolvido apresentou melhores resultados, com exceção dos valores de erro encontrados para o caso da estação de número IV.

Logo na primeira etapa foi possível visualizar e validar o efeito positivo agregado pelo fator de penalização proveniente do uso da regularização. Ao se utilizar a regularização do tipo L2 o modelo consegue diminuir consideravelmente os valores de erro, corroborando assim sua utilidade para tratar esse tipo de problema.

A estratégia de otimização dos valores de hiperparâmetros adotada se mostrou funcional e com capacidade de encontrar a configuração ótima dentro do espaço de busca definido. Para diferentes estações a busca exaustiva encontrou diferentes configurações de parâmetros, adaptando de acordo com as particularidades de cada conjunto de dados. Porém, em contrapartida o método apresenta grande sensibilidade ao tamanho do espaço

de busca, podendo estar sujeito à um grande aumento do tempo de computação em função da expansão do espaço.

A seleção de variáveis realizada na segunda etapa apresentou capacidade para reduzir números significativos da quantidade total dos regressores sem a perda de performance apenas para duas estações. Porém, caso seja necessário, o projetista ainda pode por optar uma redução forçada do número de preditores, desde que esteja disposto a abrir mão de desempenho do modelo.

Com base na análise dos resultados obtidos foi possível verificar de forma integral todas hipóteses definidas no início do trabalho, gerando assim ao final um modelo com capacidade competitiva em relação aos encontrados na literatura, e sendo uma opção atrativa para o tratamento do problema em questão.

Como sugestão para trabalhos futuros, é proposto um estudo investigativo e direcionado para a estação de número IV, que não só nesse trabalho mas como na literatura também apresenta valores e comportamento muito divergentes dos outros observados nas demais localidades do país. Além disso, é sugerido o emprego de novas técnicas para a geração de novos atributos e bem como também para a seleção desses, como por exemplo a análise de componentes principais.

O presente trabalho utilizou uma divisão do conjunto total de dados sugerida pela literatura, futuramente com o intuito de melhorar o desempenho da estimação seria interessante utilizar uma nova abordagem quanto à divisão dos dados, de forma à potencializar a etapa de treinamento do modelo por meio do uso de uma amostra maior.

Outra abordagem de grande importância a ser implementada, seria buscar novos meios mais eficientes e otimizados para a realização da estimação dos hiperparâmetros além de desenvolver um modelo de forma integrada, capaz de cruzar as informações referentes às oito unidades distribuídas pelo país levando em consideração a posição geográfica de cada uma delas, gerando-se assim apenas um modelo com capacidade de estimação para qualquer uma das regiões.

# REFERÊNCIAS


[1] BP (British Petrol). Statistical Review of World Energy, 69th ed. 2020. Disponível em: https://www.bp.com/content/dam/bp/business-sites/en/global/corporate/pdfs/energy-economics/statistical-review/bp-stats-review-2020-full-report.pdf. Acesso em: 4 set. 2020.

[2] World Meteorological Organization. WMO Greenhouse Gas Bulletin (GHG Bulletin) - No. 15: The State of Greenhouse Gases in the Atmosphere Based on Global Observations Through 2018. 2019.

[3] Campbell-Lendrum, D.; Prüss-Ustün, A. Climate change, air pollution and noncommunicable diseases. Bulletin of the World Health Organization, 97(2), 160–161, 2018.

[4] Paris. International Energy Agency. Key World Energy Statistics. *In*: Paris. International Energy Agency. Key World Energy Statistics. Paris: International Energy Agency, 2020. Disponível em: https://www.iea.org/reports/key-world-energy-statistics-2020. Acesso em: 4 set. 2020.

[5] Shen, N., Deng, R., Liao, H., & Shevchuk, O. (2020). Mapping renewable energy subsidy policy research published from 1997 to 2018: A scientometric review. Utilities Policy, 64, 101055.

[6] Jurasz, J., Canales, F. A., Kies, A., Guezgouz, M., & Beluco, A. (2020). A review on the complementarity of renewable energy sources: Concept, metrics, application and future research directions. Solar Energy, 195, 703–724.

[7] Bagherian, M. A., & Mehranzamir, K. (2020). A comprehensive review on renewable energy integration for combined heat and power production. Energy Conversion and Management, 224, 113454.

[8] Khanlari A, Sözen A, Şirin C, Tuncer AzimDoğş, Gungor A, Performance enhancement of a greenhouse dryer: Analysis of a cost-effective alternative solar air heater, Journal of Cleaner Production (2020)

[9] Arshi, S., Zhang, L., & Strachan, R. (2019). Prediction Using LSTM Networks. 2019 International Joint Conference on Neural Networks (IJCNN).

[10] Map obtained from the "Global Solar Atlas 2.0, a free, web-based application is developed and operated by the company Solargis s.r.o. on behalf of the World Bank Group, utilizing Solargis data, with funding provided by the Energy Sector Management Assistance Program (ESMAP). For additional information: https://globalsolaratlas.info

[11] Tao, Ebtehaj, Bonakdari, Heddam, Voyant, Al-Ansari, ... Yaseen. (2019). Designing a New Data Intelligence Model for Global Solar Radiation Prediction: Application of Multivariate Modeling Scheme. Energies, 12(7), 1365.

[12] Hou, M., Zhang, T., Weng, F., Ali, M., Al-Ansari, N., & Yaseen, Z. (2018). Global Solar Radiation Prediction Using Hybrid Online Sequential Extreme Learning Machine Model. Energies, 11(12), 3415.

[13] De Freitas Viscondi, G.; Alves-Souza, S.N. A Systematic Literature Review on big data for solar photovoltaic electricity generation forecasting. Sustain. Energy Technol. Assess. 2019, 31, 54–63.

[14] Ando, Y., Oku, T., Yasuda, M., Ushijima, K., Matsuo, H., & Murozono, M. (2020). Dependence of electric power flow on solar radiation power in compact photovoltaic system containing SiC-based inverter with spherical Si solar cells. Heliyon, 6(1), e03094.

[15] Zhang, H., Yang, X., Zheng, W., You, S., Zheng, X., & Ye, T. (2020). The CPMV* for assessing indoor thermal comfort and thermal acceptability under global solar radiation in transparent envelope buildings. Energy and Buildings, 110306.

[16] Duan, S. (2019). A predictive model for airflow in a typical solar chimney based on solar radiation. Journal of Building Engineering, 100916.

[17] Li, D., Ju, W., Lu, D., Zhou, Y., & Wang, H. (2015). Impact of estimated solar radiation on gross primary productivity simulation in subtropical plantation in southeast China. Solar Energy, 120, 175–186.

[18] Deng, N., Ling, X., Sun, Y., Zhang, C., Fahad, S., Peng, S., ... Huang, J. (2015). Influence of temperature and solar radiation on grain yield and quality in irrigated rice system. European Journal of Agronomy, 64, 37–46.

[19] Almorox, J., & Hontoria, C. (2004). Global solar radiation estimation using sunshine duration in Spain. Energy Conversion and Management, 45(9-10), 1529–1535.

[20] Bakirci, K. (2017). Prediction of global solar radiation and comparison with satellite data. Journal of Atmospheric and Solar-Terrestrial Physics, 152-153, 41–49.

[21] Zou, L., Wang, L., Li, J., Lu, Y., Gong, W., & Niu, Y. (2019). Global surface solar radiation and photovoltaic power from Coupled Model Intercomparison Project Phase 5 climate models. Journal of Cleaner Production, 224, 304–324.

[22] Azoumah, Y., Ramdé, E. W., Tapsoba, G., & Thiam, S. (2010). Siting guidelines for concentrating solar power plants in the Sahel: Case study of Burkina Faso. Solar Energy, 84(8), 1545–1553.

[23] Nwokolo, S. C., & Ogbulezie, J. C. (2017). A quantitative review and classification of empirical models for predicting global solar radiation in West Africa. Beni-Suef University Journal of Basic and Applied Sciences.

[24] Perez, R. (Ed.). (2018). Wind Field and Solar Radiation Characterization and Forecasting. Green Energy and Technology.

[25] Mehdizadeh, S., Behmanesh, J., & Khalili, K. (2016). Comparison of artificial intelligence methods and empirical equations to estimate daily solar radiation. Journal of Atmospheric and Solar-Terrestrial Physics, 146, 215–227.

[26] Yao, W., Zhang, C., Hao, H., Wang, X., & Li, X. (2018). A support vector machine approach to estimate global solar radiation with the influence of fog and haze. Renewable Energy, 128, 155–162.

[27] Antonopoulos, V. Z., Papamichail, D. M., Aschonitis, V. G., & Antonopoulos, A. V. (2019). Solar radiation estimation methods using ANN and empirical models. Computers and Electronics in Agriculture, 160, 160–167.

[28] Ağbulut, Ü., Gürel, A. E., & Biçen, Y. (2021). Prediction of daily global solar radiation using different machine learning algorithms: Evaluation and comparison. Renewable and Sustainable Energy Reviews, 135, 110114.

[29] Feng, Y., Gong, D., Zhang, Q., Jiang, S., Zhao, L., & Cui, N. (2019). Evaluation of temperature-based machine learning and empirical models for predicting daily global solar radiation. Energy Conversion and Management, 198, 111780.

[30] He, Q.-C., Yang, Y., Bai, L., & Zhang, B. (2019). Smart Energy Storage Management via Information Systems Design. Energy Economics, 104542.

[31] Zhai, D., Shang, J., Yang, F., & Ang, S. (2018). Measuring Energy Supply Chains' Efficiency with Emission Trading: A Two-Stage Frontier-Shift Data Envelopment Analysis. Journal of Cleaner Production.

[32] Medina-González, S., Shokry, A., Silvente, J., Lupera, G., & Espuña, A. (2018). Optimal management of bio-based energy supply chains under parametric uncertainty through a data-driven decision-support framework. Computers & Industrial Engineering.

[33] Yang, D., Kleissl, J., Gueymard, C. A., Pedro, H. T. C., & Coimbra, C. F. M. (2018). History and trends in solar irradiance and PV power forecasting: A preliminary assessment and review using text mining. Solar Energy, 168, 60–101.

[34] P.O. Fanger, Thermal comfort, Danish Technical Press, Copenhagen, 1970.

[35] Zhai, X. Q., Song, Z. P., & Wang, R. Z. (2011). A review for the applications of solar chimneys in buildings. Renewable and Sustainable Energy Reviews, 15(8), 3757–3767.

[36] Voudouri, A., Khain, P., Carmona, I., Bellprat, O., Grazzini, F., Avgoustoglou, E., ... Kaufmann, P. (2017). Objective calibration of numerical weather prediction models. Atmospheric Research, 190, 128–140.

[37] Larson, V. E. (2013). Forecasting Solar Irradiance with Numerical Weather Prediction Models. Solar Energy Forecasting and Resource Assessment, 299–318.

[38] Saud, S., Jamil, B., Upadhyay, Y., & Irshad, K. (2020). Performance improvement of empirical models for estimation of global solar radiation in India: A k-fold cross-validation approach. Sustainable Energy Technologies and Assessments, 40, 100768.

[39] Naserpour, S., Zolfaghari, H., & Zeaiean Firouzabadi, P. (2020). Calibration and evaluation of sunshine-based empirical models for estimating daily solar radiation in Iran. Sustainable Energy Technologies and Assessments, 42, 100855.

[40] Da Silva, M. B. P., Francisco Escobedo, J., Juliana Rossi, T., dos Santos, C. M., & da Silva, S. H. M. G. (2017). Performance of the Angstrom-Prescott Model (A-P) and SVM and ANN techniques to estimate daily global solar irradiation in Botucatu/SP/Brazil. Journal of Atmospheric and Solar-Terrestrial Physics, 160, 11–23.

[41] Kaba, K., Sarıgül, M., Avcı, M., & Kandırmaz, H. M. (2018). Estimation of daily global solar radiation using deep learning model. Energy, 162, 126–135.

[42] Ibrahim, S., Daut, I., Irwan, Y. M., Irwanto, M., Gomesh, N., & Farhana, Z. (2012). Linear Regression Model in Estimating Solar Radiation in Perlis. Energy Procedia, 18, 1402–1412.

[43] Al-Obeidat, F., Spencer, B., & Alfandi, O. (2018). Consistently accurate forecasts of temperature within buildings from sensor data using ridge and lasso regression. Future Generation Computer Systems.

[44] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1):55–67, 1970.

[45] A. Kraskov, H. Stogbauer and P. Grassberger, "Estimating mutual information". Phys. Rev. E 69, 2004.

[46] B. C. Ross "Mutual Information between Discrete and Continuous Data Sets". PLoS ONE 9(2), 2014.

[47] L. F. Kozachenko, N. N. Leonenko, "Sample Estimate of the Entropy of a Random Vector", Probl. Peredachi Inf., 23:2 (1987), 9-16.

[48] R. Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai, volume 14, pages 1137–1145, 1995.

[49] T. Hastie, R. Tibshirani, and J. Friedman. The elements of statistical learning: data mining, inference, and prediction, springer series in statistics, 2009.

[50] N. J. Nagelkerke et al. A note on a general definition of the coefficient of determination. Biometrika, 78(3):691–692, 1991.

[51] Saito, Akira; Tomita, Aya; Ando, Ryosuke; Watanabe, Kohei; Akima, Hiroshi (2018). Muscle synergies are consistent across level and uphill treadmill running. Scientific Reports, 8(1), 5979.

[52] J.E. Nash; J.V. Sutcliffe (1970). River flow forecasting through conceptual models part I — A discussion of principles. , 10(3), 0–290.

[53] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. Journal of machine learning research, 12(Oct):2825–2830, 2011.

[54] S. Van Der Walt, S. C. Colbert, and G. Varoquaux. The numpy array: a structure for efficient numerical computation. Computing in Science & Engineering, 13(2):22, 2011.

[55] W. McKinney. pandas: a foundational python library for data analysis and statistics. Python for High Performance and Scientific Computing, 14, 2011.

[56] Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, 9(3), 90–95.

[57] Hottel, H. C. (1976). A simple model for estimating the transmittance of direct solar radiation through clear atmospheres. Solar Energy, 18(2), 129–134.

[58] Brinsfield, R., Yaramanoglu, M., & Wheaton, F. (1984). Ground level solar radiation prediction model including cloud cover effects. Solar Energy, 33(6), 493–499.

[59] 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.

[60] Dong, Z., Yang, D., Reindl, T., & Walsh, W. M. (2015). A novel hybrid approach based on self-organizing maps, support vector regression and particle swarm optimization to forecast solar irradiance. Energy, 82, 570–577.

[61] 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.

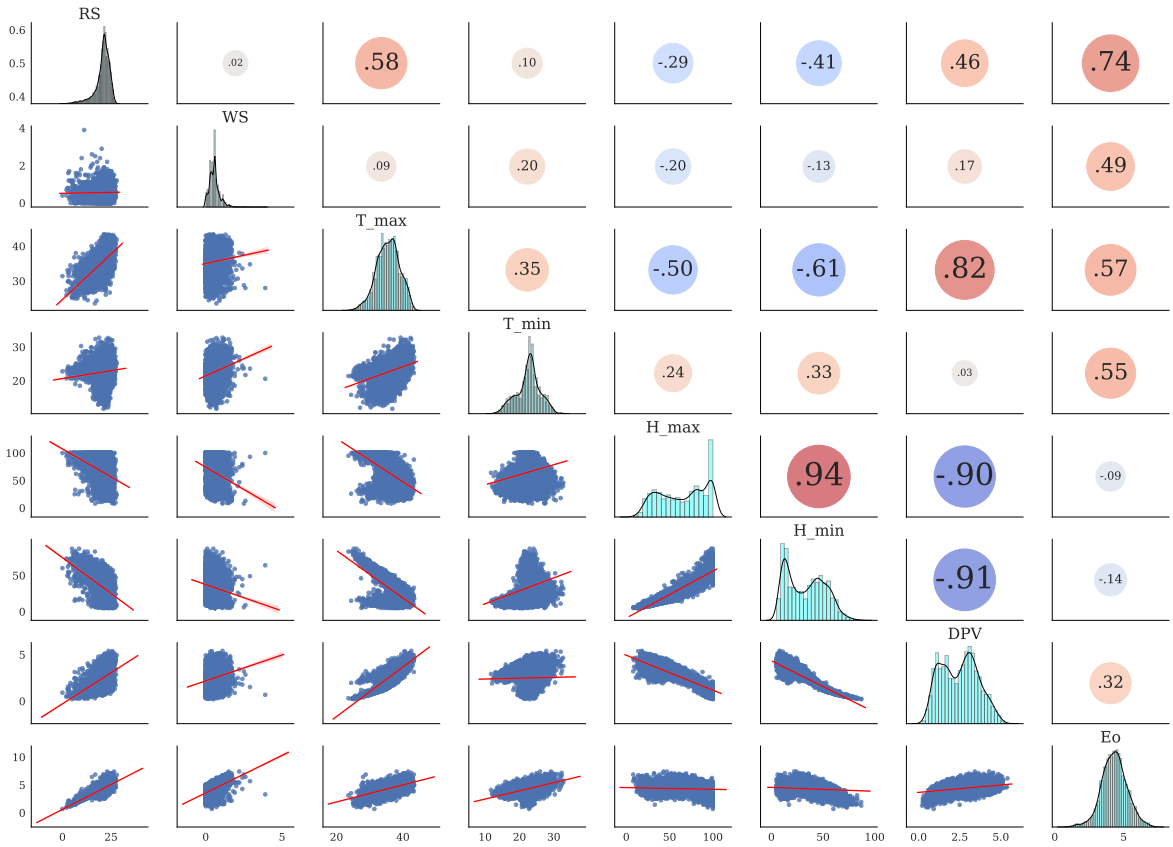
[62] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society, Volume 58, 267–288.

[63] Narvaez, G., Giraldo, L. F., Bressan, M., & Pantoja, A. (2020). Machine Learning for Site-adaptation and Solar Radiation Forecasting. Renewable Energy.

[64] Tao, H., Ewees, A. A., Al-Sulttani, A. O., Beyaztas, U., Hameed, M. M., Salih, S. Q., ... Yaseen, Z. M. (2021). Global solar radiation prediction over North Dakota using air temperature: Development of novel hybrid intelligence model. Energy Reports, 7, 136–157.

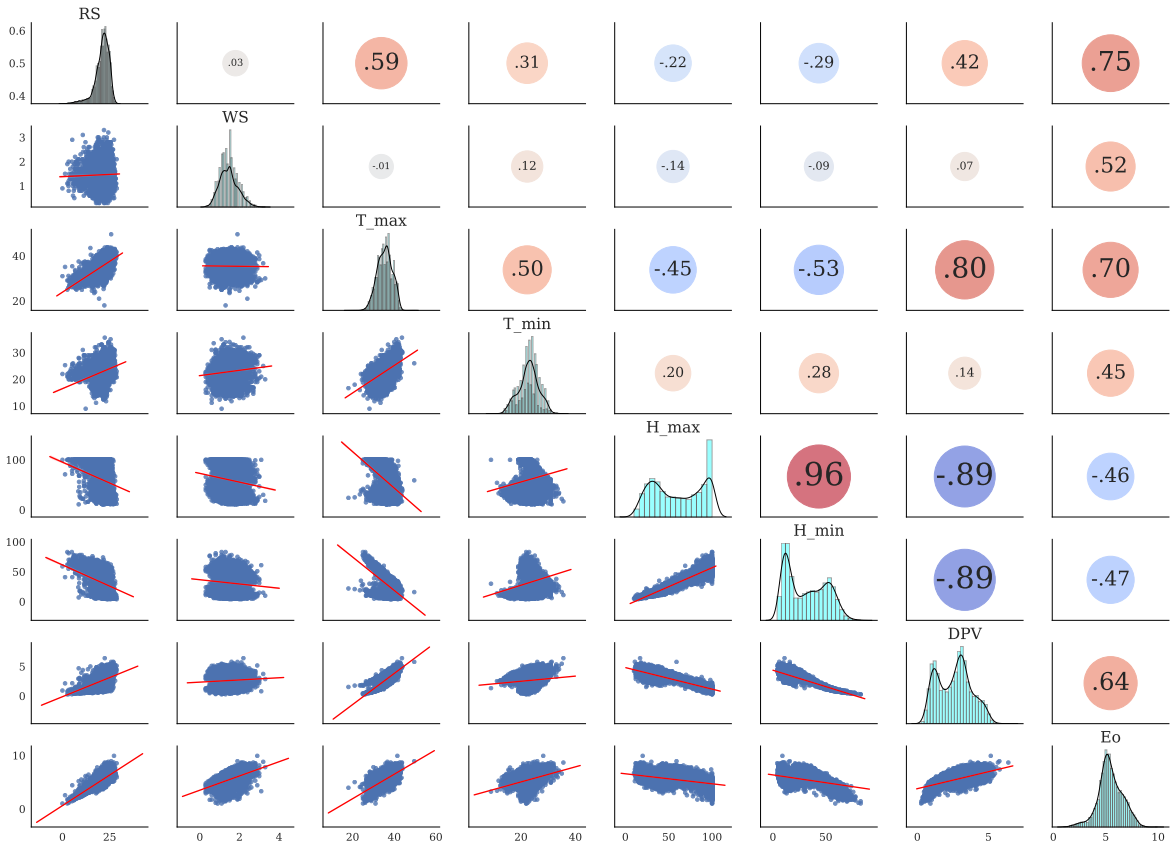
# APÊNDICE A – Gráficos

Figura 29 - Gráficos de dispersão e de distribuição para cada variável e coeficiente de Pearson apresentando a relação entre as variáveis para a estação de número II.



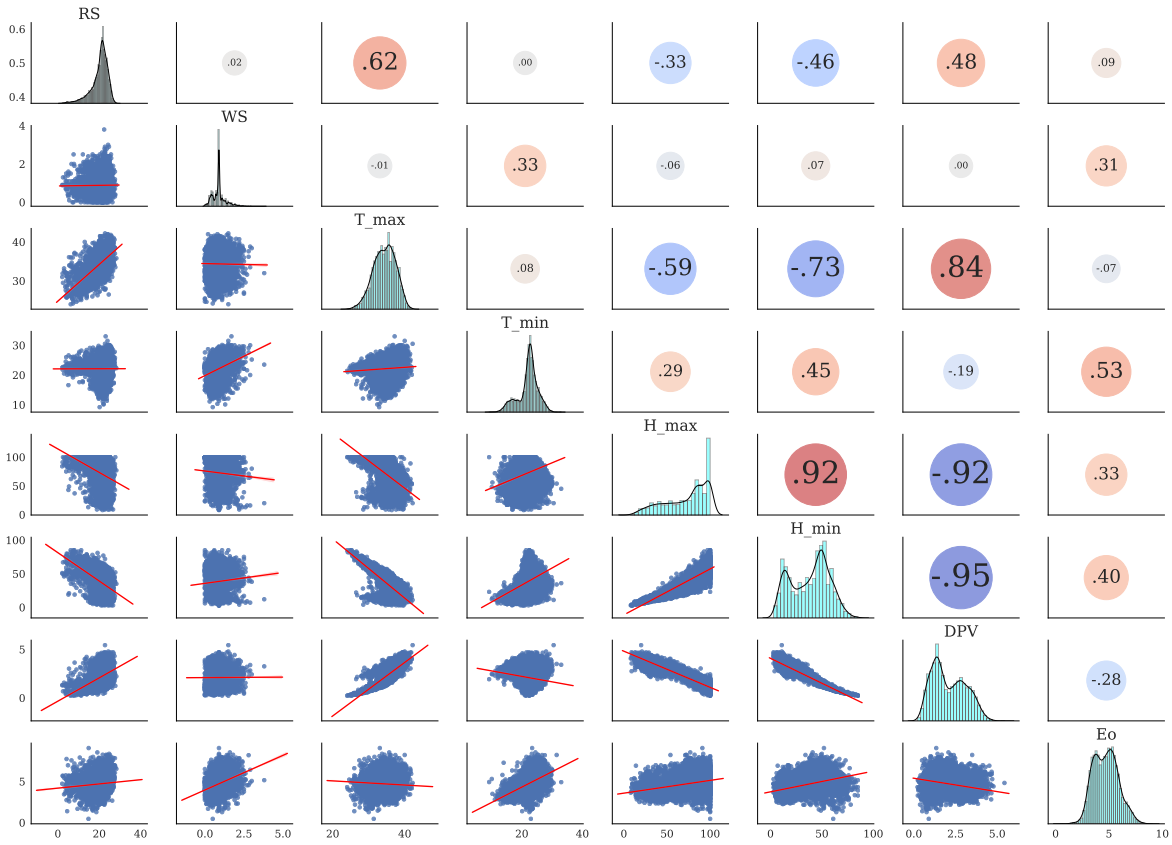

Fonte: Elaborado pelo autor (2021).

Figura 30 - Gráficos de dispersão e de distribuição para cada variável e coeficiente de Pearson apresentando a relação entre as variáveis para a estação de número III.



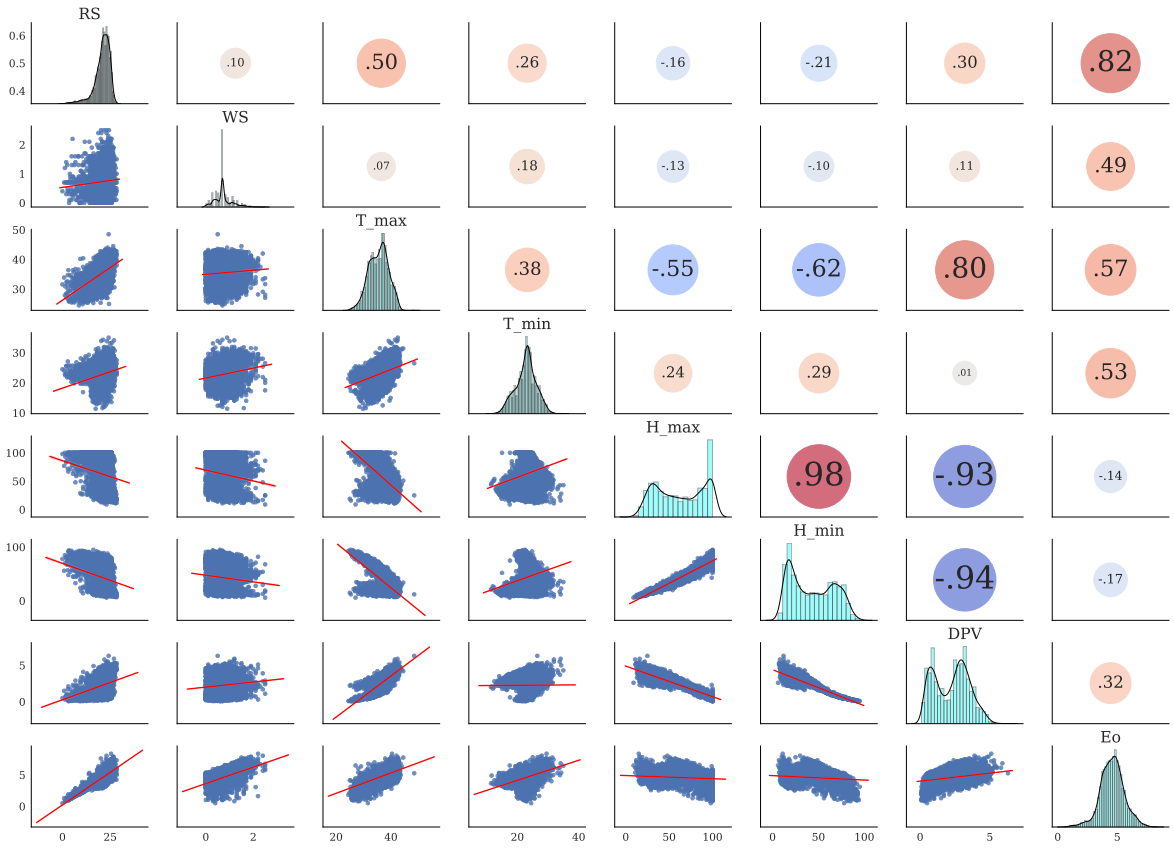

Fonte: Elaborado pelo autor (2021).

Figura 31 - Gráficos de dispersão e de distribuição para cada variável e coeficiente de Pearson apresentando a relação entre as variáveis para a estação de número IV.



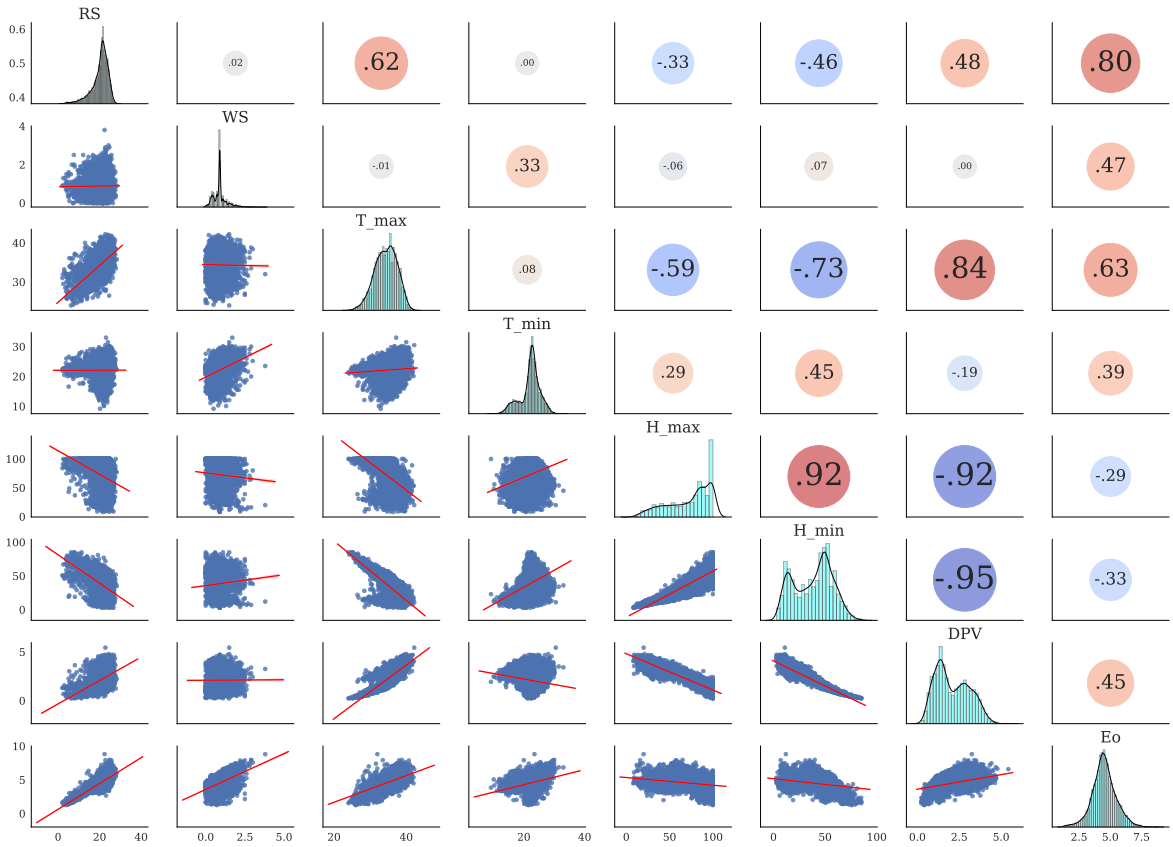

Fonte: Elaborado pelo autor (2021).

Figura 32 - Gráficos de dispersão e de distribuição para cada variável e coeficiente de Pearson apresentando a relação entre as variáveis para a estação de número V.



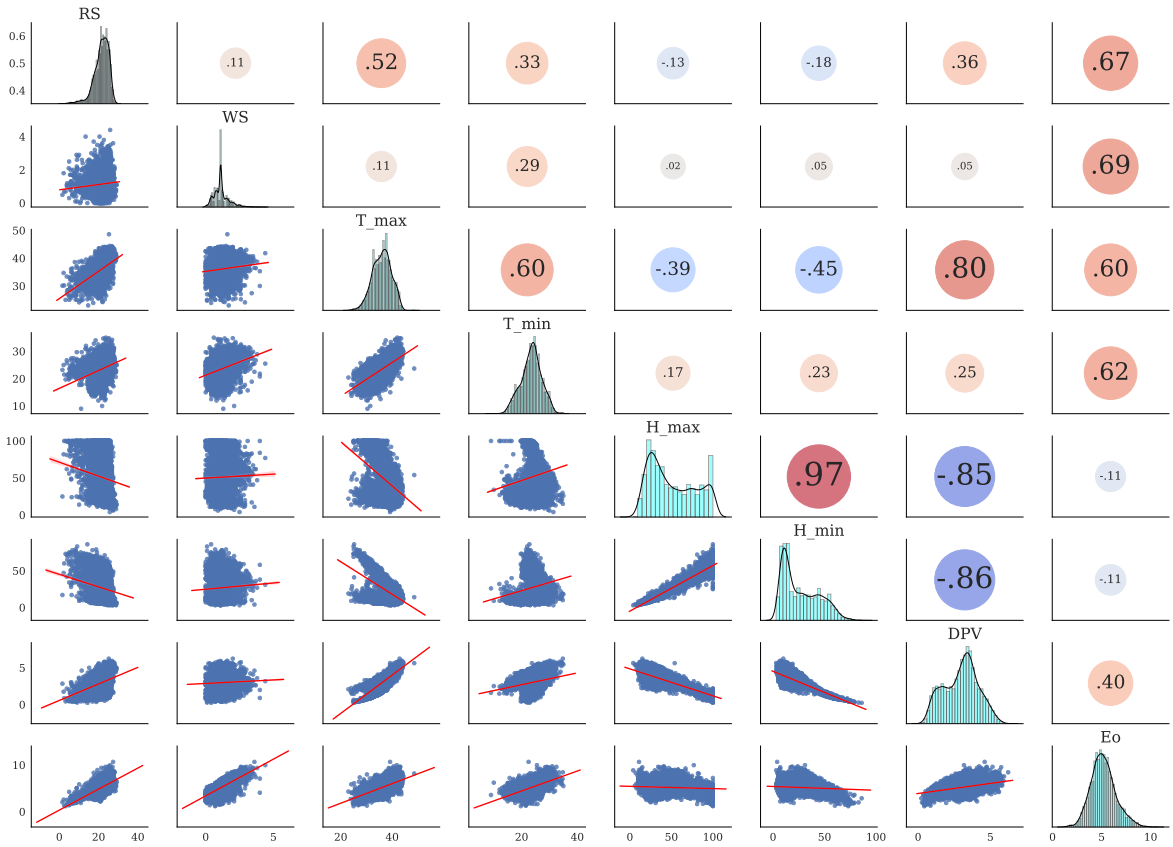

Fonte: Elaborado pelo autor (2021).

Figura 33 - Gráficos de dispersão e de distribuição para cada variável e coeficiente de Pearson apresentando a relação entre as variáveis para a estação de número VI.



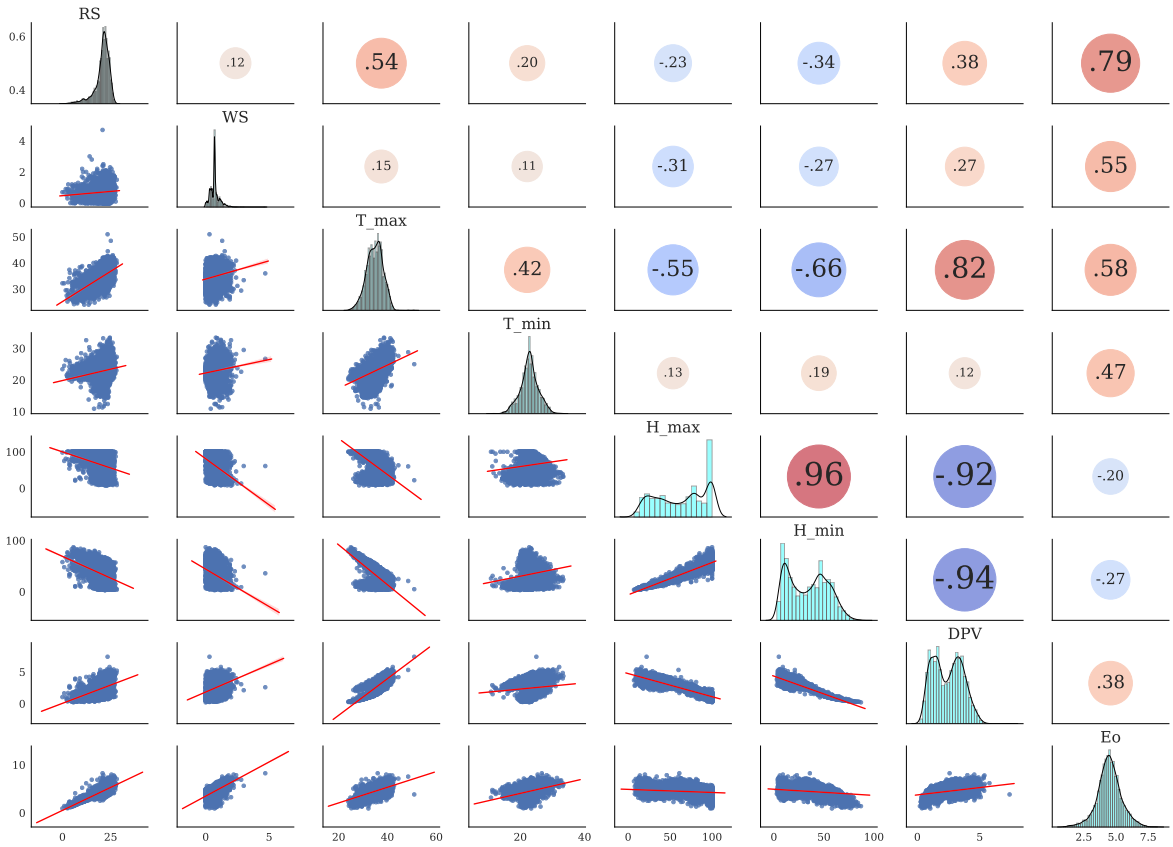

Fonte: Elaborado pelo autor (2021).

Figura 34 - Gráficos de dispersão e de distribuição para cada variável e coeficiente de Pearson apresentando a relação entre as variáveis para a estação de número VII.




Fonte: Elaborado pelo autor (2021).

Figura 35 - Gráficos de dispersão e de distribuição para cada variável e coeficiente de Pearson apresentando a relação entre as variáveis para a estação de número VIII.




Fonte: Elaborado pelo autor (2021).