ANÁLISE COMPARATIVA DAS TÉCNICAS DE REGRESSÃO PARA ESTIMAÇÃO DA GERAÇÃO DE ENERGIA SOLAR

Kevin A. T. de Almeida – kevinbdf1@gmail.com
Universidade de Brasília, Campus Darcy Ribeiro
Fernando C. Melo
Universidade de Brasília, Campus Darcy Ribeiro

Resumo. *Este trabalho apresenta uma análise comparativa das técnicas de regressão utilizadas para estimação da geração de energia solar. A utilização de fontes renováveis ao redor do mundo está em grande expansão devido aos incentivos governamentais, industriais e sociais para tornar a matriz energética mundial limpa e renovável. Apesar dos benefícios gerados por essas fontes de energia, diversos desafios ainda permeiam essa área. Entre os principais desafios é possível citar a falta de inércia dessas fontes, o que pode ocasionar perda de estabilidade do sistema, dificuldade em prever a geração desses sistemas, e nos casos de micro e minigeração, a dificuldade em prever a carga efetivamente consumida pelo sistema e a injeção de distorções harmônicas na rede básica. Nesse contexto, a estimação da geração dessas fontes é de extrema necessidade para os órgãos de regulação, distribuição e transmissão de energia elétrica. Os principais modelos de estimação utilizam dados de projeto das usinas eólicas e fotovoltaicas, no entanto, para questões de regulação, previsão de carga e constrained-off, dada a quantidade de usinas instaladas, torna-se muito complexo adquirir os dados de projeto das usinas instaladas. Assim, os modelos de caixa preta, que são modelos capazes de relacionar a temperatura, irradiância e geração, sem a necessidade dos dados de projetos, são essenciais para essas aplicações. Dessa forma, este trabalho visa apresentar algumas técnicas de regressão capazes de encontrar uma função de produtividade para estimar a geração de energia elétrica utilizando os dados de irradiância, temperatura e geração coletados da usina fotovoltaica da Faculdade de Tecnologia da Universidade de Brasília – FT-UnB.*

Palavras-chave: *Aprendizado de máquina, constrained-off solar, energia Solar, estimação de geração solar, técnicas de regressão.*

1. INTRODUÇÃO

Dada a crescente inserção de fontes intermitentes no Brasil e ao redor do mundo, com o objetivo de efetuar uma transição para uma matriz energética mais limpa e sustentável, o Sistema Interligado Nacional (SIN) passou a contar com desafios mais complexos no que diz respeito aos métodos de operação, planejamento da operação e pós-operação. Em relação ao planejamento da operação é possível destacar a dificuldade em estimar e prever a quantidade de energia que pode ser gerada por essas fontes de energia, na operação em tempo real, dada a falta de inércia dessas fontes e a diferença entre a geração prevista e a real, a operação do sistema elétrico torna-se mais complexa e pode exigir o corte ou frustração da geração (*constrained-off*) dessas fontes. E, para a pós-operação, o desafio é estimar a quantidade de energia frustrada para efeitos de ressarcimento ao agente, conforme as resoluções ANEEL nº927/2020 e ANEEL nº1073/2023.

Nesse contexto, a estimação e previsão da geração de energia por fontes intermitentes é um dos maiores desafios do setor elétrico atual, pois afeta a geração, transmissão e distribuição em vários aspectos. Dessa forma, desenvolver técnicas de estimação e previsão da geração dessas fontes garante uma melhor operação do sistema, menores níveis de cortes de geração e maior segurança e confiabilidade ao SIN. No entanto, os modelos de previsão e estimação da geração mais assertivos utilizam dados de projeto das usinas fotovoltaicas (quantidade de módulos fotovoltaicos, tipo dos módulos, quantidade de unidades condicionadoras de potência, potência nominal, entre outros), além dos dados meteorológicos (velocidade do vento, temperatura ambiente, temperatura de operação do módulo fotovoltaico e irradiância). Mas para os órgãos de operação do sistema elétrico, não é plausível simular todas as usinas com dados reais de projeto, devido à dificuldade de ter acesso a esses dados, seja por questões de segurança ou proteção de dados do agente, ou pela quantidade de dados. Assim, para ter essas estimações e previsões assertivas sem a utilização dos dados de projeto das usinas fotovoltaicas, torna-se necessária a utilização de modelos de caixa preta, que não consideram os dados de projeto da usina. Esses modelos utilizam apenas as entradas e saídas das usinas, e através de uma função de transferência ou produtividade conseguem estimar ou prever a geração.

Diante do exposto, este trabalho busca apresentar algumas técnicas de regressão que podem ser utilizadas para estimar a geração de energia solar de usinas fotovoltaicas utilizando os dados de entrada (irradiância e temperatura ambiente) e os dados de saída (energia gerada), com base em dados históricos. As técnicas utilizadas nesse artigo para o desenvolvimento da função de transferência são: regressão linear, regressão polinomial, árvore de decisão, máquina de vetores de suporte, *gradient boosting* e floresta aleatória.

2. REVISÃO BIBLIOGRÁFICA

Essa seção apresenta a revisão bibliográfica dos conceitos necessários para o entendimento do artigo. São apresentados os conceitos básicos das usinas fotovoltaicas, conceitos meteorológicos e as técnicas de regressão utilizadas no desenvolvimento do trabalho.

2.1 Energia Solar Fotovoltaica

A energia solar fotovoltaica consiste na conversão da radiação solar em energia elétrica através de dispositivos semicondutores, denominados módulos fotovoltaicos. De forma simples, um sistema fotovoltaico consiste em um conjunto de módulos fotovoltaicos conectados entre si em uma série fotovoltaica. Um conjunto de séries fotovoltaicas é conhecido como arranjo fotovoltaico. Esse arranjo fotovoltaico é conectado à uma unidade de condicionamento de potência, que é o dispositivo que converte a energia em corrente contínua gerada pelos módulos fotovoltaicos em energia em corrente alternada, na frequência e tensão da rede básica, possibilitando a sua utilização.

A energia gerada por uma usina fotovoltaica está diretamente ligada à quantidade de módulos ou potência disponível dos módulos, potência nominal das unidades de condicionamento de potência e a disponibilidade de irradiação solar. Essa relação possibilita o desenvolvimento de modelos de estimação da geração de energia solar utilizando esses dados, softwares como *PVSyst* e *System Advisor Model* (Darwish, 2021; Viana, 2020), utilizam esses dados para efetuar estimações da geração.

2.1.1 Irradiância Solar

A irradiância é o fluxo de energia radiante instantâneo incidente em uma superfície por unidade de área, e é expressa em watts por metro quadrado (W/m²). É uma variável meteorológica muito importante na avaliação do potencial de geração de energia elétrica, por representar a quantidade de energia solar disponível e que pode ser convertida em energia elétrica através dos módulos fotovoltaicos (Shi, 2021). A avaliação da irradiância consiste em efetuar medições da irradiância solar total e suas componentes, a irradiância direta e a irradiância difusa. A medida de irradiância mais utilizada para efetuar avaliações do potencial de geração é a irradiância global horizontal (IGH), que é a irradiância solar total medida em um plano paralelo ao solo.

2.1.2 Temperatura do Módulo

A temperatura do módulo fotovoltaico é considerada uma variável meteorológica por possuir uma relação inversamente proporcional com a geração de energia solar. Essa variação, apesar de pequena, pode afetar de maneira considerável a geração de energia solar (Shi, 2021). O valor de perda de eficiência do módulo fotovoltaico por temperatura em média é de 0.05 %/ºC. Em todos os principais modelos de estimação e previsão da geração de energia solar as medidas de temperatura de operação são utilizadas, seja através de modelos de estimação da temperatura do módulo a partir da temperatura ambiente ou de medições em campo.

A temperatura de operações dos módulos fotovoltaicos também pode ser estimada a partir da temperatura ambiente e da irradiância, através de equações empíricas. Existem diversas equações empíricas para efetuar essas estimativas. Uma das equações de mais fácil implementação e com bons resultados é apresentada na Eq. 1 (Araneo, 2014).

$$T_{cel} = T_{ref} + (NOCT - 20).\frac{I_{rr}}{800} \quad (1)$$

Em que, T_{cel} é a temperatura da célula, T_{ref} é a temperatura ambiente, $NOCT$ é a temperatura de operação nominal do módulo e I_{rr} é a irradiância incidente no módulo.

2.2 Técnicas de Regressão para Estimação de Energia Solar

As principais técnicas de estimação da geração solar utilizam os dados de projeto em conjunto com as equações físicas dos módulos fotovoltaicos, unidades de condicionamento de potência e alguns valores de perdas (perdas nos cabos, sombreamento, reflexão e descasamento de potência entre os módulos). Esses são os modelos disponíveis com o maior nível de assertividade. No entanto, exigem uma grande quantidade de dados que pode não estar disponível.

As técnicas de regressão também são bastante utilizadas na estimação da geração de energia solar quando os dados de projeto não estão disponíveis, pois são capazes de relacionar os dados de entrada e saída e criar uma função de transferência capaz de apresentar estimativas de geração a partir dos dados de entrada (Verma, 2016; Jawaid, 2016). Entre as técnicas de regressão, as mais comuns são a regressão linear, regressão polinomial, árvore de decisão, máquina de vetores de suporte, *gradient boosting* e floresta aleatória.

2.2.1 Regressão Linear

A técnica de regressão linear é um método estatístico que consiste em modelar a relação entre uma variável dependente e uma ou mais variáveis independentes. A regressão linear busca, basicamente, estimar uma equação linear que consiga descrever o comportamento do conjunto de dados de entrada (Anuradha, 2021). A regressão pode ser simples, quando existe apenas uma variável independente ou multivariada quando possui mais de uma variável independente. A Eq. 2, apresenta a regressão linear simples e a Eq. 3, a regressão linear multivariada.

$$Y = \alpha + \beta \cdot X + \varepsilon_i \quad (2)$$

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_n \cdot X_n + \varepsilon_i \quad (3)$$

Em que, Y é a variável dependente, X é a variável independente, X_n é o n-ésimo termo de uma regressão multivariada, β_0 é o intercepto, β_n é o coeficiente angular que representa o efeito da variável independente na variável dependente e ε_i é a variabilidade que não pôde ser explicada pelo modelo, conhecida como termo de erro.

2.2.2 Regressão Polinomial

A regressão polinomial ou regressão não linear é uma técnica utilizada para modelar o comportamento dos dados que não possuem uma relação linear, o que ocorre na maioria dos fenômenos. Assim como a regressão linear, a regressão polinomial busca ajustar uma equação que relacione de forma polinomial as variáveis independentes às variáveis dependentes dos dados utilizados como dados de treinamento (Verma, 2016). Além de conseguir modelar comportamentos não lineares, a regressão polinomial permite que o grau da equação seja ajustado conforme a necessidade, partindo de um grau 2 ao grau n. Dessa forma, os modelos polinomiais conseguem capturar melhor o comportamento dos dados e apresentar uma resposta mais aderente e com menor erro. A Eq. 4 apresenta um polinômio de grau n.

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \cdots + \beta_n \cdot X^n + \varepsilon_i \quad (4)$$

Em que, β_1 é o termo linear da equação, β_2 o termo quadrático e β_n o n-ésimo termo da regressão polinomial.

Apesar de serem mais indicados para modelar conjuntos não lineares, a regressão polinomial exige um cuidado adicional ao definir o n-ésimo termo da equação, pois ao aumentar o grau da regressão é possível gerar um caso de *overfitting*, que ocorre quando a regressão adquire todas as características dos dados de treinamento, mas não consegue efetuar boas previsões com dados que não pertenciam ao conjunto de treinamento.

2.2.3 Árvore de Decisão

Árvore de decisão é uma técnica de aprendizado de máquina que pode ser utilizada especialmente em problemas de regressão e classificação. A estrutura da técnica é semelhante a um fluxograma em formato de árvore, no qual cada raiz representa uma decisão e cada folha o resultado dessa decisão. O algoritmo funciona por meio de teste de "se-então", onde cada ramo gera uma pergunta. Caso a necessidade do "se-então" seja atendida, o algoritmo testa uma nova raiz ou gera um resultado (Jumin, 2021; Mahmud, 2021).

O primeiro passo de uma árvore de decisão é definir o atributo de separação dos dados em conjuntos mais homogêneos em termos de valores ou classes, para casos de regressão e classificação, respectivamente. Após a separação são criadas raízes ou ramos, que são bifurcados em novos ramos. Esse processo é feito de forma recursiva, respeitando os limites de profundidade máxima da árvore, número mínimo de amostras em um ramo ou a condição desejada seja atendida. Nos problemas de regressão, a predição é obtida a partir da média ou mediana das amostras da folha correspondente.

2.2.4 Máquinas de Vetores de Suporte

Support Vector Machine – SVM ou Máquinas de Vetores de Suporte também são uma técnica de aprendizado de máquina utilizada na solução de problemas de classificação e regressão. Essa técnica busca identificar um hiperplano que separe os dados com o objetivo de maximizar a distância entre as classes, sendo mais comumente utilizado em problemas de classificação (Lima, 2021). Em casos de regressão, o objetivo da técnica é identificar o hiperplano que minimize a função de perda com uma margem de erro dentro de uma margem aceitável. Essa variação do SVM é conhecida como *Support Vector Regression* – SVR.

Nas aplicações do SVR é utilizada uma equação de regressão, margens e tubos de suportes e uma função de perda. O objetivo das margens e dos tubos é criar uma zona que tolere a variação da margem em função da proximidade da previsão. A Eq. 5 apresenta a regressão e a Eq. 6 apresenta a equação de perda quadrática.

$$\boldsymbol{f(x) = \omega x + b} \quad (\mathbf{5})$$

$$\boldsymbol{L(y, f(x)) = \| y - f(x) \|^2} \quad (\mathbf{6})$$

Em que, $f(x)$ é a saída prevista, ω é o vetor de pesos, x é o vetor de característica, b é o viés e y é o valor real.

Além desses parâmetros também temos a equação da margem épsilon ($\boldsymbol{\epsilon}$) (Eq. 7) que é a diferença entre o valor previsto e o valor real, onde o erro é tolerado dentro de uma margem e não contribuem para a função de perda, diferente do SVM.

$$\boldsymbol{| y - f(x) | \leq \epsilon} \quad (\mathbf{7})$$

2.2.5 Gradient Boosting

É uma técnica de aprendizado de máquina que constrói um modelo preditivo forte utilizando modelos mais fracos combinados. O modelo é implementado a partir de uma árvore de decisão rasa (modelo fraco). A partir desses modelos fracos são calculados os resíduos (diferença entre o valor real e o previsto) e para otimizar o modelo, a técnica calcula o gradiente desses resíduos (Persson, 2017; Wang, 2018). Assim, um novo modelo é criado a partir do modelo anterior, com o objetivo de corrigir os problemas e erros do modelo anterior e assim por diante.

A previsão é dada pela equação 7, que é a soma das previsões dos modelos fracos com uma ponderação baseada na taxa de aprendizado do modelo.

$$\boldsymbol{f(x) = f_{ant}(x) + Tx_{aprend} \times f_{novo}(x)} \quad (\mathbf{8})$$

Em que, $f(x)$ é a saída prevista, $f_{ant}(x)$ é a saída do modelo previsto anteriormente, Tx_{aprend} é a taxa de aprendizado ponderada e $f_{novo}(x)$ é a saída do novo modelo calculado**.**

Assim como os outros modelos, para evitar o *overfitting* é necessário limitar a profundidade da árvore de decisão fracas e/ou ajustar a taxa de aprendizado.

2.2.6 Floresta Aleatória

Floresta aleatória é uma técnica de aprendizado de máquina, baseada no método *Ensemble*, que cria um conjunto de árvores de decisões, podendo ser aplicada em problemas de regressão e classificação (Chiteka, 2020). O modelo consiste em criar várias árvores de decisões e efetuar o treinamento, onde cada árvore é ligeiramente diferente. Para isso, a técnica de *Bootstrap Aggregating* é utilizada, essa técnica cria subconjuntos de treinamento diferentes para cada árvore a partir do conjunto original de treinamento. Essa técnica permite diminuir a correlação entre as árvores, tornando o modelo mais robusto e menos suscetível ao *overfitting*.

Nos problemas de classificação cada uma das árvores criadas possui um voto e o resultado é obtido por voto majoritário. Em problemas de regressão o resultado é obtido a partir da média das saídas obtidas por cada árvore. Essa forma de definir a saída do modelo o torna mais robusto e permite que o modelo consiga lidar melhor com dados faltantes e *outliers*, pois a previsão é dada por voto majoritário ou média das saídas individuais.

Os parâmetros importantes em uma floresta aleatória são: número de árvores, profundidade máxima de cada árvore, número mínimo de amostras em folhas e o número mínimo de amostras para divisão.

3. METODOLOGIA

Os principais modelos de previsão e estimação da geração de energia solar utilizam modelos físicos e os dados de projeto das usinas fotovoltaicas. No caso das distribuidoras, órgãos de regulação e de controle do sistema elétrico, esses dados não estão disponíveis para acesso, além da quantidade de usinas instaladas no país tornar inviável a utilização de modelos físicos que utilizam os dados de projeto, o que torna ainda mais complexo o desenvolvimento de modelos de previsão. Nesta seção são apresentadas algumas das principais técnicas de regressão capazes de prever a geração de energia solar a partir dos dados de entrada, como irradiância e temperatura do módulo, e os dados de saída, como a geração verificada. Esses modelos funcionam como modelos de caixa preta, onde apenas os dados de entrada e saída são conhecidos e uma função de transferência/produtividade é obtida a partir dos dados e dos modelos implementados.

3.1 Coleta dos Dados

Os dados utilizados neste trabalho foram coletados a partir de um piranômetro (Estrela 240-8101) e um sensor de temperatura (DSB1820) instalados no laboratório do bloco SG-11 da Faculdade de Tecnologia da Universidade de Brasília – FT/UnB. Os dados são coletados a cada 5 minutos, no entanto, são integralizados para intervalos de 30 minutos. Os dados de geração de energia solar são provenientes da usina fotovoltaica da Faculdade de Tecnologia. A tabela 1 apresenta algumas informações de projeto da usina, mas que não foram considerados no desenvolvimento dos modelos de regressão estudados.

Tabela 1. Dados de projeto da usina da Faculdade de Tecnologia – FT/UnB.

Dados da Usina da Faculdade de Tecnologia – FT/UnB	
Quantidade de Módulos	450
Modelo dos Módulos Fotovoltaicos	Canadian Solar CS6U-335P
Quantidade de UCP’s	3
Modelo das UCP’s	ABB TRIO TM-50.0-400
Potência Instalada	150,75 kWp
Inclinação dos Painéis	15º

A temperatura de operação dos módulos fotovoltaicos foi estimada utilizando a equação empírica apresentada na Eq. 1.

3.2 Tratamento dos Dados

O tratamento inicial dos dados de irradiância consistiu em eliminar os dados em que os valores estavam abaixo do limite de incerteza do piranômetro (20w/m²). Além desse tratamento, os dados de irradiância, temperatura e geração foram tratados em função de um limite máximo e mínimo. Esses limites foram estabelecidos utilizando um intervalo interquartil, sendo o interquartil a diferença entre o primeiro e o terceiro quartil. A partir do valor interquartil, dos valores dos quartis (1º e 3º) e de uma taxa de tolerância, adotada como 1.5, é possível identificar e eliminar os *outliers*.

$$Q_1 - 1{,}5 \cdot IQR \tag{9}$$

$$Q_3 + 1{,}5 \cdot IQR \tag{10}$$

Em que, Q_1 é o primeiro quartil, Q_3 é o terceiro quartil e IQR é o intervalo interquartil.

Além dos tratamentos mencionados, os dados utilizados nesse estudo passaram por um processo de normalização, o processo utilizado é o de máximo e mínimo, apresentando na Eq. 11.

$$X_{Norm} = \frac{X - X_{Min}}{X_{Max} - X_{Min}} \tag{11}$$

Em que, X_{Norm} é o dado normalizado, X é o dado original, X_{Min} é o menor valor do conjunto de dados e X_{Max} é i maior valor do conjunto de dados.

3.3 Características Iniciais da Estimação

As regressões apresentadas nesse estudo foram desenvolvidas com o objetivo de estimar a geração de energia fotovoltaica a partir dos dados de irradiância observados, para verificação e diagnóstico da geração real de uma usina fotovoltaica. Os testes dos modelos foram feitos utilizando 4 meses de dados (11/2022, 12/2022, 01/2023 e 02/2023), os conjuntos de dados foram separados em dois conjuntos, conjunto de treinamento (11/2022, 12/2022 e 01/2023) e conjunto de teste (02/2023).

Neste teste os dados de validação e testes pertencem ao mesmo conjunto, pois os hiperparâmetros foram definidos de forma arbitrária e não foram alterados ao longo do estudo.

3.4 Desenvolvimento dos Modelos

Os modelos escolhidos para implementação são: regressão linear, regressão polinomial, árvore de decisão, floresta aleatório, *gradient boosting* e máquina de vetores de suporte. As bibliotecas e os parâmetros de implementação são apresentados abaixo:

- Regressão Linear: consiste em um modelo linear implementado em R utilizando a função "lm()", esse modelo não necessita de parâmetros adicionais. Os dados de treinamento são inseridos ao modelo para ajustar uma equação linear que consiga representar o comportamento dos dados.
- Regressão Polinomial: assim como o modelo de regressão linear, a implementação desse modelo utiliza a função "lm()", no entanto, é necessário ajustar o grau do modelo. Na implementação utilizada neste trabalho o grau da equação é 2.
- Árvore de Decisão: este modelo foi implementado através da biblioteca "rpart", por ser um modelo mais complexo, alguns parâmetros de ajuste são necessários. Por exemplo, a profundidade máxima definida para o modelo é de 15, o número de amostras por divisão é 10 e o número mínimo de amostras por terminal é de 5.
- Máquina de Vetores de Suporte: a implementação deste modelo utiliza a função "svm" e utiliza um único parâmetro de ajustes, que é o *kernel* do modelo, ou seja, o tipo de núcleo utilizado. Neste trabalho, o kernel utilizado é do tipo "radial".
- Gradient Boosting: para implementar o modelo de *gradient boosting* foi utilizada a biblioteca "gbm" e a função gbm. A quantidade de árvores do modelo foi definida como 20, o nível de profundidade da interação é 3 e a função de distribuição utilizada foi a gaussiana.
- Floresta Aleatória: sua implementação utiliza apenas dois parâmetros, o número de árvores e o mtry, que é a quantidade de dados de entrada que cada árvore de decisão pode utilizar. Na implementação do modelo o mtry foi definido como a raiz quadrada da quantidade de dados menos 1 e o número de árvores foi definido como 20.

Todos os modelos foram implementados em R, utilizando o *R Studio* e suas bibliotecas. Para efetuar o treinamento dos modelos os dados foram separados em conjunto de treinamento e conjunto de teste, sendo que o conjunto de treinamento possui três meses de dados (11/2022, 12/2022 e 01/2023) e o conjunto de testes possui um mês (02/2023) para todos os modelos.

Após o treinamento dos modelos é utilizada a função “predict” do R para efetuar as previsões de geração para cada um dos modelos.

3.5 Avaliação dos Modelos

Os modelos implementados foram avaliados utilizando as seguintes métricas de erro: raiz do erro quadrático médio (RMSE), erro médio absoluto (MAE) e o erro percentual absoluto médio (MAPE). Além disso, são avaliados o erro global, que é o erro do modelo ao efetuar as estimações utilizando o conjunto de treinamento e o erro real, que é o erro do modelo ao efetuar as estimações do conjunto de teste. O erro global tende a ser menor que o erro real, pois o modelo prevê os dados que são conhecidos pelo modelo, apesar de não refletir o erro real do modelo, o erro global é importante para a verificação do *overfitting*. Um erro global muito baixo demonstra que o modelo está se ajustando aos dados de forma excessiva, o que causa uma dificuldade em efetuar boas estimações para dados que não estão compreendidos no conjunto de teste.

3.6 Avaliação do Impacto da Temperatura de Operação nas Estimações

Outra análise efetuada neste trabalho é a verificação da influência do dado de temperatura de operação do módulo nas estimações por regressões. É amplamente conhecido o efeito da temperatura em modelos físicos, no entanto, em modelos regressivos, mesmo sem o dado de temperatura de operação, o modelo pode conseguir encontrar uma relação entre a irradiância e a geração de energia que consiga efetuar as estimações com bons resultados e com uma complexidade de implementação menor dos que os modelos regressivos multivariados.

Dessa forma, todos os modelos foram treinados de duas formas distintas. O primeiro conjunto de modelos foi treinado utilizando o conjunto de dados de irradiância e geração verificada. O segundo conjunto de modelos foi treinando utilizando a irradiância verificada, a geração verificada, e a temperatura de operação dos módulos, que foi estimada a partir dos valores de temperatura ambiente verificada.

4. RESULTADOS

Este capítulo apresenta os resultados das implementações e avaliações dos resultados obtidos em cada modelo implementado, utilizando as duas variáveis (temperatura de operação do módulo e irradiância) e utilizando somente uma variável (irradiância).

4.1 Resultados dos Modelos de Regressão Considerando Irradiância e Temperatura

Nesta seção são apresentadas as métricas de erro global (erro do conjunto de treinamento) e real (erro do conjunto de teste) utilizando a métrica de erro médio absoluto (MAE), a raiz do erro quadrático médio (RMSE) e o erro percentual absoluto médio (MAPE) das estimações efetuadas pelos modelos estudados.

Tabela 2. Raiz do erro quadrático médio do conjunto global (treino) e real (teste).

Modelo	**Erro Global (RMSE)**	**Erro Real (RMSE)**
Regressão Linear	7360,54	8356,25
Regressão Polinomial	7163,07	8304,33
Árvore de Decisão	5511,74	9418,14
Floresta Aleatória	3760,22	8995,15
Máquinas de Vetor de Suporte	7111,72	8417,27
Gradient Boosting	7609,41	9576,42

Tabela 3. Erro médio absoluto do conjunto global (treino) e real (teste).

Modelo	**Erro Global (MAE)**	**Erro Real (MAE)**
Regressão Linear	3765,24	4437,88
Regressão Polinomial	3642,06	4376,15
Árvore de Decisão	2851,74	4990,10
Floresta Aleatória	1902,30	4703,83
Máquinas de Vetor de Suporte	3608,29	4369,14
Gradient Boosting	3897,41	5019,89

Tabela 4. Erro percentual absoluto médio do conjunto global (treino) e real (teste).

Modelo	**Erro Global (MAPE)**	**Erro Real (MAPE)**
Regressão Linear	15,89	19,38
Regressão Polinomial	15,88	22,42
Árvore de Decisão	11,42	24,07
Floresta Aleatória	7,96	24,21
Máquinas de Vetor de Suporte	15,61	21,87
Gradient Boosting	15,50	21,05

Apesar dos resultados da raiz do erro quadrático médio e erro médio absoluto apresentarem valores altos, na casa dos milhares, é importante ressaltar que estamos analisando uma usina com 150.75 kWp, ou seja, um erro de 10000 Watts, representa apenas 6.66% da capacidade total da usina. Outro detalhe é que o modelo de floresta aleatório apresentou uma forte tendência ao *overfitting*, apresentando um erro global muito abaixo ao erro verificado no conjunto de teste (erro real)

Analisando os resultados individuais, é possível notar uma forte tendência dos modelos de árvore de decisão e floresta aleatória ao *overfitting*, apresentam resultados excelentes no conjunto de treino e resultados semelhantes aos outros modelos no conjunto de testes. A regressão linear multivariada apresentou resultados consistentes e até melhores que os outros modelos.

4.2 Resultados dos Modelos de Regressão Considerando Irradiância

Os modelos agora são avaliados utilizando somente a variável irradiância, mantendo todas as configurações e parâmetros utilizados no teste anterior. O intuito dessa análise é avaliar o impacto da temperatura em modelos de regressão e se o aumento da complexidade do modelo é justificado pelo possível ganho de performance do modelo.

Tabela 6. Raiz do erro quadrático médio do conjunto global (treino) e real (teste) utilizando apenas irradiância.

Modelo	**Erro Global (RMSE)**	**Erro Real (RMSE)**
Regressão Linear	7371,11	8389,67
Regressão Polinomial	7305,14	8645,55
Árvore de Decisão	6152,18	9484,07
Floresta Aleatória	4451,05	9598,47
Máquinas de Vetores de Suporte	7263,31	8522,52
Gradient Boosting	7721,79	10127,30

Tabela 7. Erro médio absoluto do conjunto global (treino) e real (teste) utilizando apenas irradiância.

Modelo	**Erro Global (MAE)**	**Erro Real (MAE)**
Regressão Linear	3771,71	4459,21
Regressão Polinomial	3734,21	4627,07
Árvore de Decisão	3150,12	4949,89
Floresta Aleatória	2283,25	5025,67
Máquinas de Vetores de Suporte	3697,71	4518,71
Gradient Boosting	3976,58	5502,63

Tabela 8. Erro percentual absoluto médio do conjunto global (treino) e real (teste).

Modelo	**Erro Global % (MAPE)**	**Erro Real % (MAPE)**
Regressão Linear	15,92	19,25
Regressão Polinomial	16,45	20,19
Árvore de Decisão	13,16	21,37
Floresta Aleatória	9,38	21,52
Máquinas de Vetores de Suporte	15,96	19,49
Gradient Boosting	16,00	20,46

Em relação ao RMSE e MAE, houve um pequeno aumento em relação aos modelos que consideram a temperatura. No entanto, para o erro MAPE, em todos os modelos é possível identificar uma pequena diminuição do erro.

4.3 Apresentação das Curvas Obtidas

A Fig. 1 apresenta as curvas de geração dos modelos apresentados para o dia 15 de fevereiro de 2022 considerando apenas a irradiância como variável de entrada do modelo.

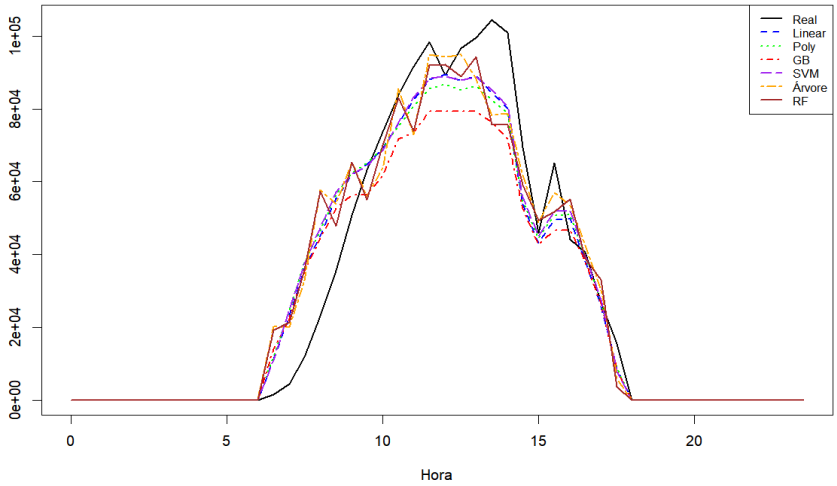

Figura 1 - Curvas de geração (W) considerando apenas a irradiância como variável de entrada.

A Fig. 2 apresenta o gráfico dos modelos estudados considerando a temperatura e a irradiância como variável de entrada. Apesar de não apresentar ganhos em termos percentuais, é possível notar que ao fim do dia os modelos que consideram a temperatura apresentam uma relação melhor com a geração real do que os modelos que consideram apenas a irradiância.

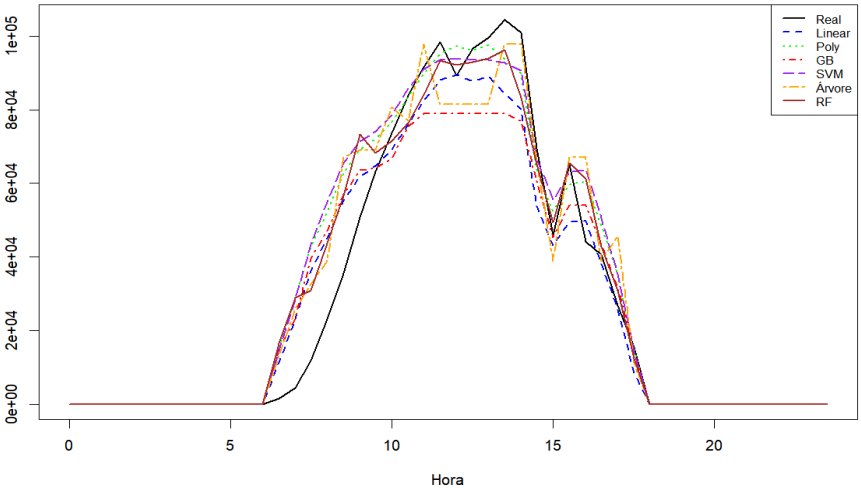

Figura 2 - Curvas de geração (W) considerando a irradiância e temperatura como variável de entrada.

A série temporal completa do mês de fevereiro de 2023 com os dados de geração observada e os dados estimados pelos modelos mostram que as estimações realizadas pelo modelo apresentam comportamentos semelhantes aos dados observados, garantindo a reprodutibilidade dos modelos para dados um mês a frente. As séries temporais das Fig. 3, Fig. 4 e Fig. 5 foram obtidas utilizando os dados de saída (geração estimada) dos modelos que consideram apenas a irradiância como variável de entrada.

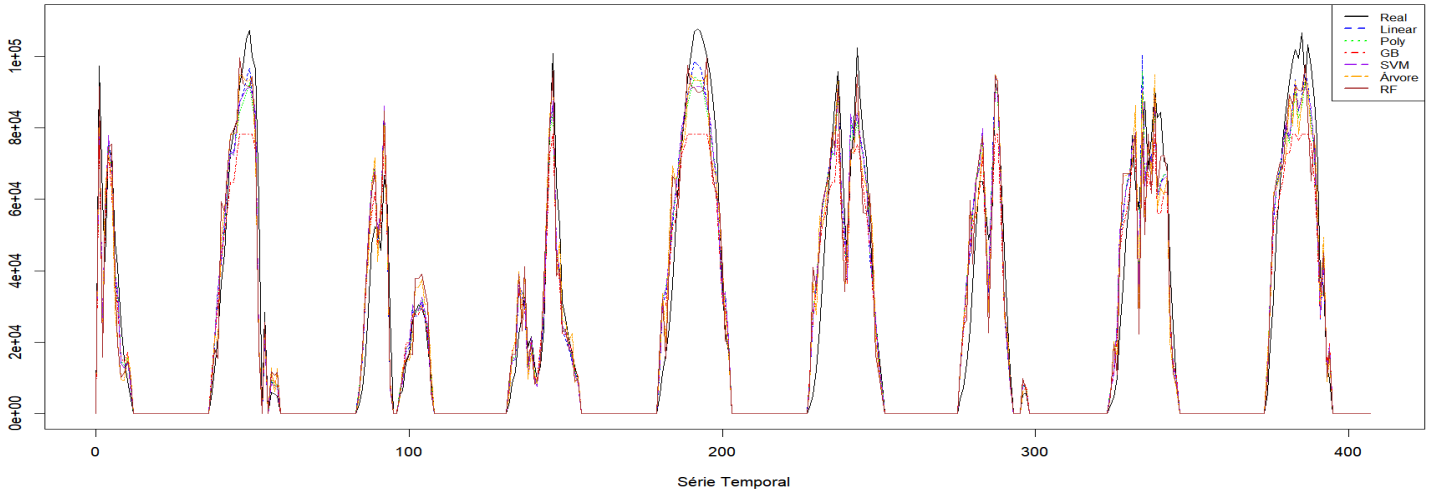

Figura 3 – Série temporal da geração verificada e estimada (W) para os dias 01/02/2023 ao dia 09/02/2023.

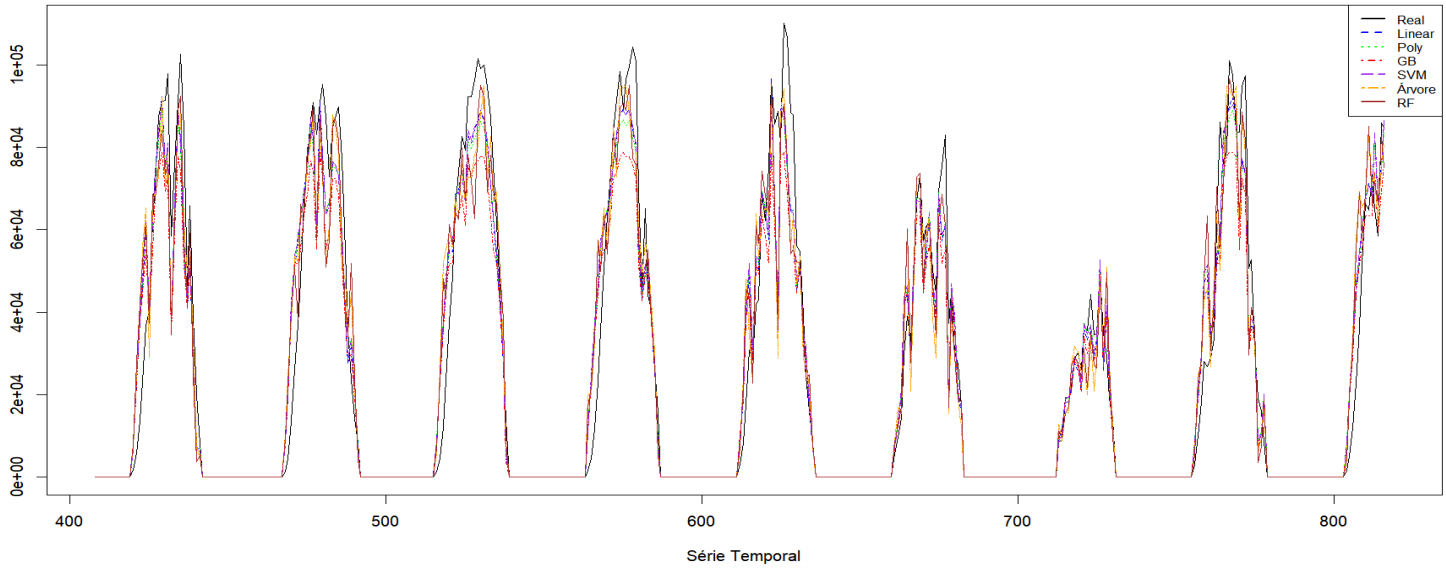

Figura 4 – Série temporal da geração verificada e estimada (W) para os dias 10/02/2023 ao dia 18/02/2023.

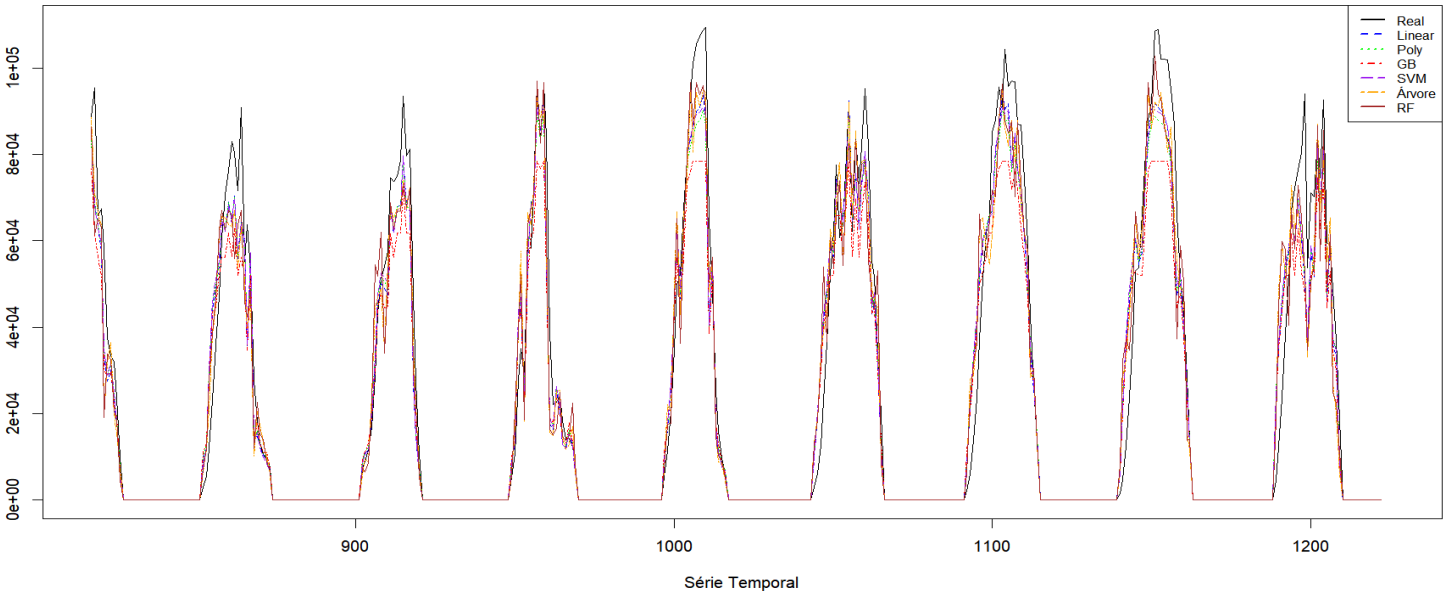

Figura 5 – Série temporal da geração verificada e estimada (W) para os dias 18/02/2023 ao dia 26/02/2023.

5. CONCLUSÕES

As regressões utilizadas neste estudo apresentaram bons resultados de estimação da geração real de energia elétrica de uma usina fotovoltaica. Os modelos de regressão linear e de máquinas de vetores de suporte apresentaram resultados extremamente próximos da geração real, no entanto, todos os modelos estudados conseguiram reproduzir o comportamento da geração observada ao longo do mês estimado. Diante disso, os modelos mostraram que são capazes de estimar a geração total mensal de uma usina fotovoltaica, o que pode ser amplamente aproveitado nas aplicações de *constrained-off*.

Outra análise interessante desses modelos é a avaliação da influência da temperatura do módulo, que em todos os casos, além de não melhorar a previsão, diminui ligeiramente a precisão dos modelos.

Dentre os modelos estudados, observa-se que tanto a técnica de árvore de decisão quanto a técnica de máquinas de vetores de suporte apresentam baixos níveis de erro. No entanto, há uma alta possibilidade de overfitting, pois, apesar do baixo erro global, ao estimarem dados desconhecidos (conjunto de teste), o erro aumenta consideravelmente, aproximando-se dos níveis observados nos outros modelos estudados. O modelo linear apresentou resultados consistentes e não possui uma volatilidade muito alta, tendo seu comportamento com baixas variações, ou seja, ao aumentar a irradiância a geração sempre aumenta, o que não ocorre nos modelos de aprendizado de máquina.

Por fim, dados os resultados obtidos neste trabalho, é possível concluir que os modelos de regressão estudados podem ser utilizados para fins de verificação e diagnóstico de usinas fotovoltaicas, para aplicações em *constrained-off* e avaliação do potencial de geração de energia solar. Diante das análises, é possível concluir que a escolha do modelo de regressão e das variáveis de entrada dependem do tipo de aplicação necessária e da disponibilidade de dados.

Agradecimentos

Os autores agradecem ao Departamento de Engenharia Elétrica da Universidade de Brasília, à Faculdade de Tecnologia da Universidade de Brasília e ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade de Brasília.

REFERÊNCIAS

M. A. Aly Darwish, "Design of a photovoltaic system using SAM and ETAP software," 2021 International Conference on Green Energy, Computing and Sustainable Technology (GECOST), Miri, Malaysia, 2021.

Viana, Zulkner Cruz, et al. "Accuracy Analysis of Pvsyst Software for Estimating the Generation of a Photovoltaic System at the Polo de Inovação Campos dos Goytacazes." 2020 IEEE PES Transmission & Distribution Conference and Exhibition-Latin America (T&D LA). IEEE, 2020.

Ying Shi, Yue Sun, Junliang Liu, Xiong Du, Model and stability analysis of grid-connected PV system considering the variation of solar irradiance and cell temperature,International Journal of Electrical Power & Energy Systems, Volume 132, 2021.

Araneo, Rodolfo, Umberto Grasselli, and Salvatore Celozzi. "Assessment of a practical model to estimate the cell temperature of a photovoltaic module." International Journal of Energy and Environmental Engineering 5 (2014).

T. Verma, A. P. S. Tiwana, C. C. Reddy, V. Arora and P. Devanand, "Data Analysis to Generate Models Based on Neural Network and Regression for Solar Power Generation Forecasting," 2016 7th International Conference on Intelligent Systems, Modelling and Simulation (ISMS), Bangkok, Thailand, 2016, pp. 97-100, doi: 10.1109/ISMS.2016.65.

F. Jawaid and K. NazirJunejo, "Predicting daily mean solar power using machine learning regression techniques," 2016 Sixth International Conference on Innovative Computing Technology (INTECH), Dublin, Ireland, 2016.

Lima, M. A. F. B., Fernández Ramírez, L. M., Carvalho, P. C. M., Batista, J. G., and Freitas, D. M. (August 16, 2021). "A Comparison Between Deep Learning and Support Vector Regression Techniques Applied to Solar Forecast in Spain." ASME. J. Sol. Energy Eng. February, 2022.

Chiteka, Kudzanayi, Rajesh Arora, and S. N. Sridhara. "A method to predict solar photovoltaic soiling using artificial neural networks and multiple linear regression models." Energy Systems 11.4 (2020).

Anuradha, K., et al. "Analysis of solar power generation forecasting using machine learning techniques." E3S Web of Conferences. Vol. 309. EDP Sciences, 2021.

Jumin, Ellysia, et al. "Solar radiation prediction using boosted decision tree regression model: A case study in Malaysia." Environmental Science and Pollution Research 28 (2021).

Mahmud, Khizir, et al. "Machine learning based PV power generation forecasting in alice springs." IEEE Access 9 (2021).

Persson, Caroline, et al. "Multi-site solar power forecasting using gradient boosted regression trees." Solar Energy (2017).

Wang, Jidong, et al. "A short-term photovoltaic power prediction model based on the gradient boost decision tree." Applied Sciences 8.5 (2018).

Agência Nacional de Energia Elétrica - ANEEL. Resolução Normativa nº927/2020, de 22 de março de 2021.

Agência Nacional de Energia Elétrica - ANEEL. Resolução Normativa nº1073/2023, de 12 de setembro 2023.

COMPARATIVE ANALYSIS OF REGRESSION TECHNIQUES FOR SOLAR ENERGY GENERATION ESTIMATION

Abstract. *This paper presents a comparative analysis of regression techniques used to estimate solar energy generation. The global expansion of renewable sources is significantly driven by government, industrial, and social incentives to make the world's energy matrix clean and renewable. Despite the benefits derived from these energy sources, numerous challenges persist in this field. Among the main challenges are the lack of inertia in these sources, which can lead to system instability, difficulty in predicting their generation, and, in the case of micro and minigeneration, challenges in forecasting the load effectively consumed by the system and injecting distortions into the basic grid. In this context, predicting the generation of these sources is crucial for electricity regulation, distribution, and transmission authorities. While major forecasting models utilize project data from wind and photovoltaic plants, acquiring such data for installed plants becomes complex due to their quantity. Hence, black-box models, capable of relating temperature, irradiance, and generation without the need for project data, are essential for these applications. Therefore, this work aims to present some regression techniques capable of estimating a productivity function to predict electrical energy generation using irradiance, temperature, and generation data collected from the photovoltaic plant at the Faculty of Technology of the University of Brasília – FT-UnB.*

Keywords: *Solar Energy, solar generation estimation, regression techniques, constrained-off solar, generation forecasting.*