MODELOS DE APRENDIZADO DE MÁQUINA PARA PREVISÃO DE CURTO PRAZO DA IRRADIÂNCIA GLOBAL HORIZONTAL

Nicolas Moreira Branco[1], *Danilo Silva*[1]

[1]Departamento de Engenharia Elétrica e Eletrônica, Universidade Federal de Santa Catarina, Florianópolis, SC, Brasil.
nicolas.branco@posgrad.ufsc.br, danilo.silva@ufsc.br

RESUMO

Prever com precisão a geração da energia solar fotovoltaica permitiria proporcionar uma maior estabilidade à matriz energética; no entanto, este permanece um problema desafiador devido à variabilidade da geração dos sistemas fotovoltaicos. A irradiância é a variável mais relacionada com a geração solar, então sua previsão é de extrema importância. Este estudo utilizou um conjunto de dados de referência para prever a irradiância global horizontal 15 minutos a frente. Foram utilizados dados radiométricos e ambientais, além de informações sobre a posição solar, valores de irradiâncias de céu claro e previsões do modelo persistente. Com esses dados e valores históricos, os modelos de aprendizado de máquina Ridge, Floresta Aleatória, *Gradient Boosting* e LSTM foram treinados e otimizados, obtendo resultados significativamente superiores ao modelo persistente na métrica de erro quadrático médio. Adicionalmente, o modelo LSTM obteve resultados superiores a todos os demais na métrica de erro absoluto médio.

Palavras-chave – aprendizado de máquina, irradiância global horizontal, previsão de curto prazo, energia solar.

ABSTRACT

Accurately predicting solar photovoltaic generation would allow for greater stability of the energy matrix; however, this remains a challenging problem due to the variability of the photovoltaic plants' generation. Irradiance is the variable most closely related to solar generation, so its prediction is of utmost importance. This study used a reference dataset to predict the horizontal global irradiance 15 minutes ahead. Radiometric and environmental data were used, in addition to information on solar position, clear sky irradiance values, and persistent model predictions. With these data and historical values, the Ridge, Random Forest, Gradient Boosting and LSTM machine learning models were trained and optimized, achieving significantly better results than the persistent model on the mean square error metric. Additionally, the LSTM model outperformed all others in the mean absolute error metric.

Key words – *machine learning, global horizontal irradiance, short-term forecast, solar energy.*

1. INTRODUÇÃO

A geração de energia das usinas de plantas fotovoltaicas não é facilmente previsível e sua intermitência de geração pode causar problemas ao sistema elétrico se não for bem tratada devido a geração de harmônicos e outros problemas relacionados [1]. O fator mais importante para a previsão da geração solar é a irradiância, portanto para prever com precisão a geração solar são necessário bons modelos para prever a irradiância.

Um dos métodos possíveis para esse tipo de previsão é através de dados históricos de telemetria local e com uso de modelos de aprendizado de máquina. Por exemplo, o trabalho de [2] comparou o modelo persistente, redes neurais e floresta aleatória para previsão de 1 a 6 horas a frente utilizando como entrada as irradiâncias global horizontal (GHI), direta normal e difusa global. Já o trabalho de [3] utilizou dados do GHI passado para prever de 10 minutos a 4 horas a frente e comparou os métodos persistente, SVM, LSTM e redes neurais. Já em [4] o GHI também foi previsto 1 hora a frente utilizando 3 tipos de irradiância como entrada junto a dados ambientais.

Nesse trabalho, iremos utilizar dados atuais e históricos radiométricos e ambientais de 5 em 5 minutos, junto a dados da posição solar, irradiâncias de céu claro, e a previsão do modelo persistente para realizar a previsão da GHI 15 minutos a frente. Em particular, os modelos Ridge, Floresta Aletória, *Gradient Boosting* e LSTM serão comparados em relação ao modelo de céu claro e persistente.

2. MATERIAL E MÉTODOS

2.1. Conjunto de dados

Os dados utilizados nesse trabalho são provenientes de uma estação solarimétrica na cidade de Folsom na Califórnia, localizada em 38,642° Norte, 121,148° Oeste e com a altitude de 100 metros [5]. Nessa base estão disponíveis dados radiométricos e ambientais de 2014 a 2016 de minuto a minuto. Nos dados radiométricos estão presentes a Irradiância Global Horizontal (GHI), Irradiância Direta Normal (DNI) e Irradiância Difusa Horizontal (DHI). Já nos dados ambientais estão disponíveis a temperatura e umidade relativa do ar, pressão atmosférica, a velocidade média e máxima do vento, sua direção predominante e a precipitação. Seguindo a sugestão dos autores do conjunto de dados, manteve-se o ano de 2016 como conjunto de teste e definiu-se o ano de 2014 como conjunto de treinamento e o ano de 2015 como conjunto de validação.

2.2. Pré-processamento

Em adição a esses dados disponibilizados, utilizou-se o algoritmo NREL SPA [6] para obter as posições solares (zênite, azimute e equação do tempo) utilizando as coordenadas geográficas como entrada. Considerou-se uma amostra como sendo diurna se o zênite é menor que 85°. Na

Dado	Fonte	Atual	Histórico
GHI	Folsom	x	x
DNI	Folsom	x	
DHI	Folsom	x	
Umidade	Folsom	x	
Pressão	Folsom	x	
VelVento Média	Folsom	x	
VelVento Max	Folsom	x	
DirVento	Folsom	x	
Precipitação	Folsom	x	
Zênite	NRELSPA	x	
Azimute	NRELSPA	x	
Equação do tempo	NRELSPA	x	
GHI céu claro	sSolis	x	x
DNI céu claro	sSolis	x	
DHI céu claro	sSolis	x	
minuto do dia	Folsom	x	
dia do ano	Folsom	x	
K_T	Calculado	x	x
Modelo Persistente	Calculado	x	x
GHI céu claro a frente	sSolis	x	x

Tabela 1: Variáveis de entrada dos modelos. Para o LSTM, todos os históricos são utilizados

sequência, estimou-se as irradiâncias de céu claro (GHI, DNI e DHI) utilizando o modelo de Solis Simplificado (sSolis) [7]. O sSolis foi escolhido por não ter muitos parâmetros como entrada e ser um dos modelos de céu claro com melhor desempenho [8]. Ambos os métodos foram utilizados por meio da biblioteca PVLIB[1]. Foi feito também um ajuste do fuso horário dos dados de UTC para o horário local da estação (UTC-7).

Para todas as variáveis, foi feita uma suavização de 1 em 1 minuto para 5 em 5 minutos, através de uma média móvel atrasada ($x'[n] = (x[n] + \cdots + x[n-4])/5$), mantendo-se somente os dados dos minutos múltiplos de 5. Uma amostra após a suavização foi considerada diurna se e somente se as amostras de todos os 5 minutos correspondentes fossem considerados diurnos. Na sequência, adicionou-se como variáveis os valores do minuto do dia e dia do ano.

Com relação aos dados de entrada, também foi adicionado o GHI de céu claro 15 minutos a frente, o valor calculado do índice de céu claro atual (K_T) e o valor predito do GHI (15 minutos a frente) pelo modelo persistente, cujas definições são dadas na seção 2.3.2. Finalmente, também foram adicionados dados dos 12 últimos instantes de tempo anteriores (correspondentes a um histórico de 60 minutos) das variáveis GHI, GHI de céu claro, GHI de céu claro adiantada de 15 minutos, K_T, e a predição do modelo persistente para 15 minutos a frente. O conjunto de variáveis de entrada é descrito na Tabela 1. Esse conjunto é o considerado para os modelos Ridge, Floresta Aletatória e *Gradient Boosting*; já para o LSTM, todas essas variáveis estão disponíveis com os dados atuais e passados.

Como variável-alvo, foi utilizada a série temporal do GHI de 15 minutos a frente. Foram mantidas apenas as amostras cuja variável-alvo correspondia a um instante considerado diurno (zênite < 85°), sendo as demais excluídas. Porém, essa exclusão não foi aplicada aos dados de entrada, de forma que alguns dados no início do dia possuem nas variáveis históricas valores de GHI noturnos.

[1] https://github.com/pvlib/pvlib-python

conjunto	total	mantidas
treinamento (ano de 2014)	104820	46971
validação (ano de 2015)	105120	48178
teste (ano de 2016)	105133	48105

Tabela 2: Quantidade de amostras totais (em intervalos de 5 minutos) e após a exclusão

A Tabela 2 mostra o número de amostras disponíveis para cada subconjunto de dados após a suavização e exclusão descritas acima.

2.3. Modelos de referência

Os modelos descritos abaixo serão utilizados como ponto de comparação para os modelos de aprendizado de máquina.

2.3.1. Modelo de céu claro

Trata-se de um simples *baseline* que assume todos os dias como sendo de céu claro; portanto, o modelo prevê a GHI como sendo igual à GHI de céu claro no mesmo instante, i.e., $\hat{I}(t + \Delta t) = I_{cs}(t + \Delta t)$, onde $I_{cs}(t)$ denota a GHI de céu claro no instante t.

2.3.2. Modelo persistente

O modelo persistente estima que o próximo instante terá o mesmo K_T que o instante atual [9]. O K_T é a razão entre a irradiância atual real e a irradiância de céu claro atual, $K_T(t) = I(t)/I_{cs}(t)$. Portanto, a previsão de GHI é dada por

$$\hat{I}(t + \Delta t) = I_{cs}(t + \Delta t) \cdot K_T(t). \quad (1)$$

2.4. Métricas

As métricas de avaliação consideradas neste trabalho foram o erro médio absoluto (MAE) e a raiz do erro quadrático médio (RMSE). Para otimização de hiperparâmetros de modelos de aprendizado de máquina, utilizou-se o MAE.

2.5. Modelos de aprendizado de máquina

2.5.1. Ridge

O modelo Ridge corresponde a um modelo de regressão linear por mínimos quadrados com aplicação de regularização L2 (norma quadrática) sobre os coeficientes do modelo.

2.5.2. Floresta aleatória

A floresta aleatória é um modelo de aprendizado de máquina do tipo *ensemble*, o qual combina as predições de diversos modelos mais simples. A floresta aleatória é formada por diversas árvores de decisão treinadas em paralelo. Como esse problema é de regressão, o valor estimado da saída é a média da previsão de cada uma das árvores. Cada uma das árvores usualmente são treinadas com um subconjunto dos

parâmetro	valor	tipo
alpha	$[10^{-10}; 10^{10}]$	loguniform

Tabela 3: Faixa de hiperparâmetros para o modelo Ridge

parâmetro	valor	tipo
n_estimators	100	fixo
max_depth	[10, 50]	randint
min_samples_leaf	[50, 200]	randint
max_features	log2	fixo

Tabela 4: Faixa de hiperparâmetros para o modelo Floresta Aleatória

dados, buscando aumentar a aleatoriedade e independência entre elas.

2.5.3. Gradient boosting

O *Gradient Boosting* é outro modelo do tipo *ensemble* e também tem a árvore de decisão como sua base. Diferentemente da Floresta Aleatória que processa os dados de forma paralela, os componentes do *Gradient Boosting* são treinados sequencialmente, com a entrada de um componente do modelo obtida pela diferença entre a variável alvo e a predição do componente anterior.

2.5.4. LSTM

O modelo *Long Term Short Memory* (LSTM) é um tipo de rede neural recorrente utilizada para séries de dados. Esse é um modelo que foi criado para resolver o problema dissipação dos gradientes e poder trabalhar com séries temporais mais longas. Sua memória é dividida para considerar a entrada atual e também informações da série temporal.

2.6. Treinamento e otimização de hiperparâmetros

Para otimização de hiperparâmetros, os modelos foram treinados no conjunto de treinamento e avaliados no conjunto de validação. Utilizou-se a biblioteca Scikit-learn para todos os modelos de aprendizado de máquina, exceto o modelo LSTM, para o qual foi utilizada a biblioteca PyTorch.

2.6.1. Escolha de hiperparâmetros no Scikit-learn

A busca de hiperparâmetros foi realizada através da função *RandomizedSearchCV* do Scikit-learn, a qual realiza uma amostragem aleatória dos hiperparâmetros a partir de uma distribuição informada.

Para o modelo Ridge, foram avaliadas 10 configurações com os parâmetros da Tabela 3. A melhor configuração encontrada foi {alpha=3.127}.

Para o modelo de Floresta Aleatória, foram avaliadas 25 configurações com os parâmetros da Tabela 4. A melhor configuração encontrada foi {max_depth=37, min_samples_leaf=64}.

Para o modelo de *Gradient Boosting*, foram avaliadas 50 configurações com os parâmetros da Tabela 5. A melhor configuração encontrada foi {learning_rate=0.058, loss='squared_error', max_depth=5, min_samples_leaf=32}.

parâmetro	valor	tipo
n_estimators	100	fixo
loss	'squared_error' 'absolute_error'	discreta
learning_rate	$[10^{-2}, 10^{1}]$	loguniform
max_depth	[2, 8]	randint
min_samples_leaf	[25, 200]	randint
max_features	'log2'	fixo

Tabela 5: Faixa de hiperparâmetros para o modelo do *Gradient Boosting*

2.6.2. Treinamento do modelo LSTM

Foram selecionados como dados de entrada os dados atuais e os 12 últimos instantes de todas as variáveis da Tabela 1. Calculou-se o valor médio e desvio padrão dessas variáveis para o conjunto de treinamento somente para os instantes onde a variável alvo (GHI futuro) era diurno. Utilizando essas métricas escalonou-se os dados de entrada utilizando o método *StandardScaler*. Esses dados foram carregados para serem utilizados no PyTorch através de um *dataloader* com shuffle=True e batch_size=256.

A arquitetura do modelo utilizou uma única camada LSTM com 50 *hidden_units*, uma camada com probabilidade de *dropout* igual a 0.1 e uma camada de saída linear. O modelo LSTM foi treinado com otimizador Adam com taxa de aprendizado 0.01 e função de perda L1 e o melhor resultado foi referente ao modelo obtido na época 73.

3. RESULTADOS E DISCUSSÃO

Os modelos treinados com os melhores hiperparâmetros encontrados foram testados no conjunto de teste e os resultados obtidos estão mostrados na Tabela 6. Percebemos que os modelos possuem desempenhos consideravelmente superiores aos do modelo persistente em relação ao RMSE, mesmo sem uma variação muito significativa entre eles. Por outro lado, em relação ao MAE somente o LSTM obteve uma melhoria significativa em relação ao modelo persistente.

Modelo	MAE	RMSE
Céu Claro	86,7	158,2
Persistente	31,8	74,4
Ridge	31,9	66,1
Floresta Aleatória	31,7	65,4
GradientBoosting	31,4	65,5
LSTM	29,9	65,9

Tabela 6: Resultados dos modelos no conjunto de teste

Na Figura 1 podemos perceber que em dias ensolarados praticamente todos os modelos têm resultados bem similares. Outro ponto interessante a ser observado é que, mesmo nesses dias, o modelo de céu claro não reproduz perfeitamente os valores de GHI, o que pode estar relacionado às limitações do modelo de céu claro utilizado ou também a outros fenômenos meteorológicos de menor intensidade.

Na Figura 2, a qual mostra um exemplo de dia nublado, percebemos uma grande variação entre as previsões dos modelos, especialmente entre os treinados e o persistente.

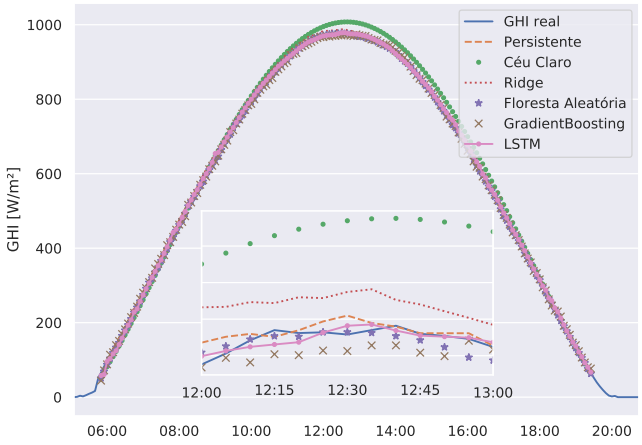

Figura 1: GHI e respostas dos diferentes modelos no dia ensolarado em 23-06-2016

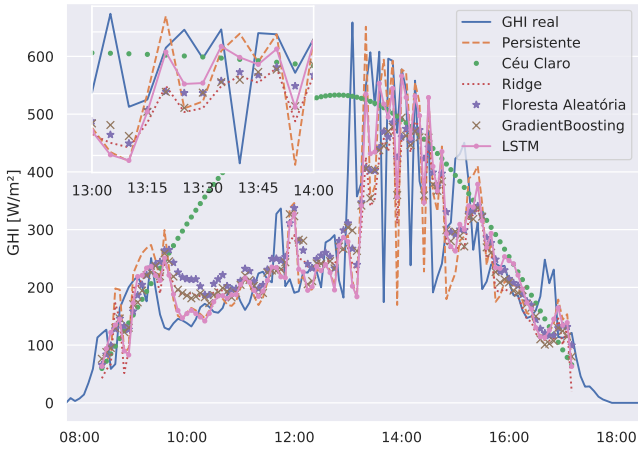

Figura 2: GHI e respostas dos diferentes modelos no dia nublado em 24-01-2016

Além disso, no período observado, percebe-se que eles estão bem próximos entre si e seguem a tendência geral do GHI, porém, o GHI varia mais rápido que a amostragem consegue acompanhar em detalhes.

Os gráficos e a tabela de resultados demonstram que mesmo sem a utilização de imagens do céu é possível obter resultados consideravelmente melhores que o modelo persistente.

4. CONCLUSÕES

Nesse trabalho foi possível demonstrar resultados consideravelmente melhores que o modelo persistente para a previsão de GHI 15 minutos a frente mesmo sem a utilização de imagens. De uma maneira geral, percebemos resultados próximos entre os modelos Ridge, Floresta Aleatória e GradientBoosting, mas o LSTM obteve resultados superiores, especialmente para a métrica MAE.

Para os próximos estudos nesse tema, desejamos incorporar imagens do céu (*whole sky*) além dos dados de sensores. Espera-se que os resultados dos modelos melhorem consideravelmente, já que esta é uma informação muito importante para a previsão de GHI de curto prazo. Outra possibilidade a ser investigada é não realizar a suavização dos dados e trabalhar com os dados de minuto a minuto para poder acompanhar variações metereológicas mais rápidas.

5. REFERÊNCIAS

[1] V. Kumar, A. S. Pandey, and S. K. Sinha. Grid integration and power quality issues of wind and solar energy system: A review. *2016 International Conference on Emerging Trends in Electrical Electronics & Sustainable Energy Systems (ICETEESES)*, pages 71–80, 2016.

[2] L. Benali et al. Solar radiation forecasting using artificial neural network and random forest methods: Application to normal beam, horizontal diffuse and global components. *Renewable Energy*, 132:871–884, March 2019.

[3] S. Gbémou et al. A Comparative Study of Machine Learning-Based Methods for Global Horizontal Irradiance Forecasting. *Energies*, 14(11):3192, May 2021.

[4] Q. Ashfaq et al. Hour-Ahead Global Horizontal Irradiance Forecasting Using Long Short Term Memory Network. In *2020 IEEE 23rd International Multitopic Conference (INMIC)*, pages 1–6, Bahawalpur, Pakistan, November 2020. IEEE.

[5] H. T. C. Pedro, D. P. Larson, and C. F. M. Coimbra. A comprehensive dataset for the accelerated development and benchmarking of solar forecasting methods. *Journal of Renewable and Sustainable Energy*, 11(3):036102, 2019.

[6] I. Reda and A. Andreas. Solar Position Algorithm for Solar Radiation Applications (Revised). Technical Report NREL/TP-560-34302, 15003974, NREL, January 2008.

[7] P. Ineichen. A broadband simplified version of the Solis clear sky model. *Solar Energy*, 82(8):758–762, August 2008.

[8] F. Antonanzas-Torres et al. Clear sky solar irradiance models: A review of seventy models. *Renewable and Sustainable Energy Reviews*, 107:374–387, 2019.

[9] B. Y.H. Liu and R. C. Jordan. The interrelationship and characteristic distribution of direct, diffuse and total solar radiation. *Solar Energy*, 4(3):1–19, July 1960.