# SELEÇÃO DE VARIÁVEIS E MODELAGEM PREDITIVA DE RADIAÇÃO SOLAR

***Resumo:*** *O enorme avanço científico e tecnológico dos últimos anos permitiu que produção de energia solar se tornasse cada vez mais eficiente e acessível. Como consequência disso, os custos de instalação de painéis solares foram drasticamente reduzidos, e um número muito maior de residências e edifícios passaram a produzir energia dessa forma. Entretanto, o fato desse tipo de geração ser altamente intermitente e não controlável provou-se um desafio para as concessionárias de energia, visto que estas precisam a todo momento buscar igualar a produção de energia ao seu consumo. O objetivo deste artigo é propor uma metodologia que utiliza dados históricos de radiação solar para estimar a produção de energia solar futura, ajudando, dessa forma, a integração da geração distribuída à rede elétrica. Para isso, neste artigo, será apresentado um método de seleção de variáveis e um modelo de aprendizagem de máquina (LightGBM), dois métodos frequentemente usados para a previsão de séries temporais com dados estruturados.*

***Palavras-chave:*** *Energia solar, modelagem preditiva, aprendizagem de máquina.*

## 1 INTRODUÇÃO

Vimos nas últimas décadas um enorme investimento em tecnologias de produção de energia solar, tornando-a bem mais barata e acessível. Como consequência disso, os custos de instalação de painéis solares foram drasticamente reduzidos, e um número muito maior de residências e edifícios passaram a produzir energia dessa forma. De 2010 a 2020, a produção mundial de energia fotovoltaica cresceu de 37 gigawatts para 770 gigawatts, um aumento anual de 35,4% (International Energy Agency; 2016).

Esse cenário, apesar de representar um enorme passo na redução da dependência de energias provenientes de combustíveis fósseis, representa um imenso desafio para as distribuidoras de energia. Isso se deve ao fato da energia solar ser altamente intermitente (variando ao longo do dia, e extremamente dependente de condições climáticas) e

incontrolável. As concessionárias buscam a todo momento igualar a produção de energia ao seu consumo. O consumo é altamente regular, e por isso já não representa um desafio tão grande. Porém, com o aumento de geração de energia provinda de fontes intermitentes, como a solar, aumenta a volatilidade da produção, tornando-a mais difícil de ser prevista.

A complexidade de prever a produção de energia solar aumenta seus custos de integração à rede elétrica, fazendo com que as concessionárias repassem esses custos aos consumidores, dificultando sua adoção (Sharma, N.; Sharma, P.; Irwin, D.; Shenoy P.; 2011). Dessa forma, torna-se necessária uma metodologia que consiga estimar de forma precisa tal produção.

Por isso, ao longo deste artigo será feita a apresentação de uma técnica de seleção de variáveis e de uma modelo de aprendizagem profunda (*LightGBM*), dois métodos frequentemente usados para a previsão de séries temporais com dados estruturados. O objetivo do modelo é, a partir de dados passados (horários) de radiação solar, estimar os valores para as próximas 24 horas. A geração de energia solar é diretamente proporcional à radiação solar, e por isso ela será usada aqui como um indicador da produção e variável a ser prevista.

Na seção 2 é feita a análise da seleção de variáveis, na qual é feita uma apresentação dos dados utilizados, que são em seguida analisados, e por fim utilizados para produzir novas variáveis com maior valor preditivo. Na seção 3, será apresentado o processo de desenvolvimento do modelo preditivo. Por fim, na seção 4 é feita a análise de resultados obtidos e as conclusões tiradas sobre o problema.

## 2 ENGENHARIA DE VARIÁVEIS

### 2.1 Aquisição e pré-processamento dos dados

A primeira etapa do processo foi a aquisição dos dados. Estes foram obtidos a partir do Banco de Dados Meteorológicos para Ensino e Pesquisa (BDMEP), disponibilizado pelo Instituto Nacional de Meteorologia (INMET) em seu site. O INMET disponibiliza dados de várias cidades brasileiras e, para esta análise, foram utilizados os dados da cidade de Natal-RN, do período de 01 de janeiro de 1961 à 31 de Dezembro de 2019.

O conjunto contém as seguintes variáveis:

- Data ;
- Precipitação (mm) ;
- Temperatura máxima (°C) ;
- Temperatura média (°C) ;
- Temperatura mínima (°C) ;
- Umidade relativa do ar (%) ;
- Velocidade do vento (Km/h) ;
- Radiação (J/m2).

Em estudos prévios, foi observado que as variáveis de temperatura, precipitação, umidade e velocidade do vento, apesar de possuírem correlação com os valores presentes da radiação, tem pouco valor preditivo para a radiação futura, quando comparadas aos valores passados da própria radiação. Por isso, para o resto da análise são utilizados apenas os dados referentes a radiação, a data e o horário da medição.

Após a obtenção dos dados, eles foram processados para se adequar ao formato tabular que foi utilizado ao longo da análise. Uma amostra desses dados pode ser visualizada na Figura 1.

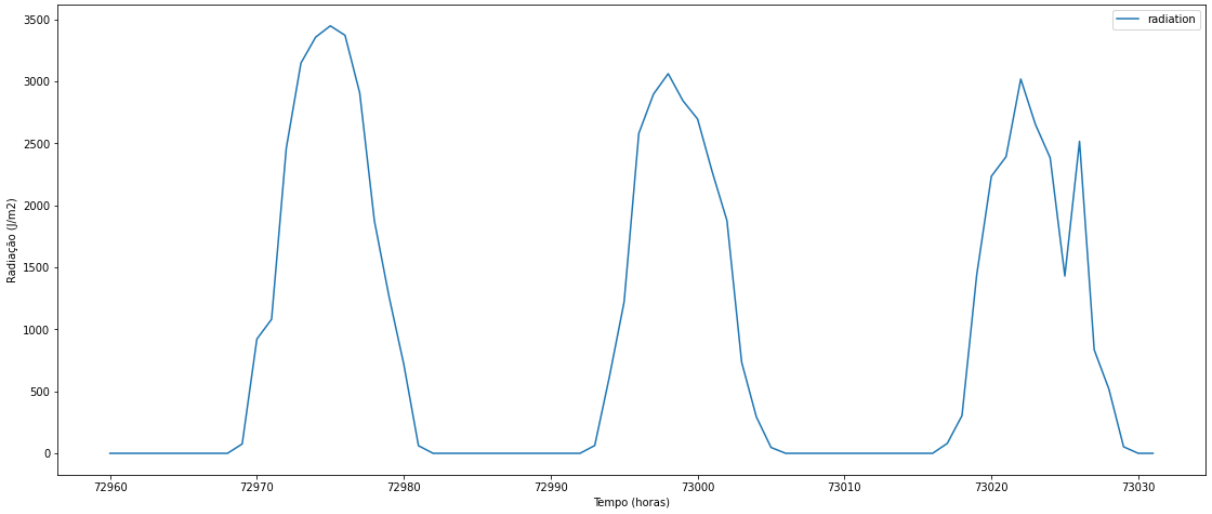
Figura 1 - Amostra do conjunto de dados após a formatação inicial.

| | date | hour | radiation |
|---|---|---|---|
| **0** | 2011-02-09 | 0 | 0.0 |
| **1** | 2011-02-09 | 1 | 0.0 |
| **2** | 2011-02-09 | 2 | 0.0 |
| **3** | 2011-02-09 | 3 | 0.0 |
| **4** | 2011-02-09 | 4 | 0.0 |

Fonte: Elaborado pelo autor (2020)

A análise foi feita utilizando a linguagem de programação Python e algumas bibliotecas, entre elas: Pandas, Numpy - Manipulação e visualização de dados. Matplotlib - Visualização dos dados Scikit-learn/LightGBM - Desenvolvimento e avaliação do modelo preditivo Os dados possuíam algumas falhas, causados por falhas de medição, o que resultava em algumas medições não terem valores definidos para algumas horas. Dado que o número de falhas bastante pequeno em relação ao tamanho do conjunto de dados, foi utilizada uma interpolação linear para preencher os dados faltantes. Uma amostra dos dados, a medida de radiação de 5 dias, é observada na Figura 2.

Figura 2 - Amostra dos dados de radiação.




Fonte: Elaborado pelo autor (2020)

## 2.2 Seleção de variáveis

Partindo das variáveis de data/hora e radiação, é possível produzir novas variáveis com valor preditivo que contribuam para o objetivo do nosso modelo. Essa variáveis podem ser temporais como hora, dia do ano, mês e ano, e também quantitativas como os valores passados de radiação, médias móveis de diferentes janelas temporais, etc. A seguir serão apresentadas as variáveis utilizadas no estudo.

### 2.2.1. Variáveis temporais

A radiação solar, assim como outras medidas de natureza meteorológica, possuem comportamentos sazonais, ou seja, que tendem a se repetir em certos intervalos de tempo. De forma geral, existem duas sazonalidades dominantes, a diária, causada pelo movimento de rotação da Terra em torno do próprio eixo, e anual, causada pelo movimento de translação da Terra ao redor do Sol.

Para modelar esse comportamento sazonal, são utilizadas variáveis temporais, obtidas a partir do horário e data da medição. Essas variáveis são:

- Hora do dia (0 à 23)
- Dia do ano (1 - 366)
- Mês (1 - 12)
- Ano (2000 - 2019)

Apesar de existirem abordagens mais complexas para a análise de sazonalidade, que muitas vezes envolvem a utilização de modelos estatísticos para estimar o comportamento sazonal e isolá-lo, a utilização de variáveis temporais tem sido utilizada com muito sucesso em aplicações, tanto na academia quanto na indústria.

Dessa forma, a partir de funções da biblioteca *Pandas*, as variáveis descritas acima foram adicionadas à tabela de dados.

### 2.2.2. Variáveis quantitativas

Com o objetivo de estimar os valores da radiação solar para as 24 horas futuras, é necessário observar o comportamento passado dessa variável. A forma mais simples de fazê-lo é observar diretamente as medições anteriores ao período que deseja prever (t -1, t - 2, ...), dentro de uma janela de tempo. Esses valores passados são chamados de *Lag* (atraso), e são definidos pela Equação 1, onde *x* é o sinal em questão e *k* é o atraso para qual deseja-se calcular o *lag*.

$$Lag_k(t) = x[t-k] \quad (1)$$

Porém, a partir desses mesmos dados, é possível gerar métricas adicionais que podem aumentar a convergência do modelo preditivo, provocando uma diminuição no erro das previsões. Entre essas métricas, as mais utilizadas são as diferenças de *lag* e as médias móveis (de diferentes janelas de tempo).

A diferença de *lag* consiste em calcular a diferença entre o valor de uma variável no tempo 't' e algum outro momento anterior 't - k'. Isso permite transmitir ao modelo uma medida de variação (semelhante a derivada de uma variável), podendo contribuir para a melhora de seu desempenho. O cálculo da diferença de *lag* é descrito pela Equação 2.

$$Diff_k(t) = x[t] - x[t-k] \tag{2}$$

A média móvel, como o próprio nome sugere, consiste em valor médio de uma variável dos últimos x intervalos de tempo. Isso permite suavizar variações mais bruscas nos dados, reduzindo o efeito de possíveis *outliers* no resultado do modelo. Além disso, a média móvel também pode indicar uma tendência de mais longo prazo nos dados, como um aumento na radiação em um período específico do ano. A média móvel é determinada a partir da Equação 3.

$$MA_N = \frac{1}{N}\sum_{k=0}^{N} x[t-k] \tag{3}$$

Na qual $N$ é a janela de tempo para qual deseja-se calcular a média.

Um parâmetro importante a ser escolhido é a janela de tempo passada que será usada para para prever os valores futuros. Dada a natureza horária dos dados, e os resultados obtidos empiricamente a partir de testes, nos quais essa janela foi variada de 12 horas a 72 horas, foi observado que uma janela de 48 horas produziu os melhores resultados, e por isso foi escolhida.

Dessa forma, as variáveis criadas foram:

- *Lag*: últimas 48 horas;
- Diferença de *lag*: últimas 48 horas;
- Médias móveis: com janelas de 3, 6, 12, 24 e 48 horas.

Juntando as variáveis originais com as novas variáveis obtidas, tem-se um total de 105 variáveis, que serão utilizadas para estimar o valor de outras 24 (a radiação para as 24 horas seguintes).

## 3 MODELAGEM PREDITIVA

Após a análise e tratamento dos dados da série temporal, passamos para a etapa de modelagem do problema. Como dito na introdução, o objetivo é desenvolver um modelo que preveja a radiação solar a partir dos valores passado e das variáveis criadas. Visto que a radiação é uma variável numérica, trata-se de um problema de regressão. Como utilizaremos múltiplas variáveis para a previsão, esse é um problema de regressão multivariado.

Dada as características descritas acima, a métrica escolhida para avaliar os resultados do modelo foi a raiz do erro médio quadrático (RMSE), que é a métrica mais utilizada para avaliação de modelos de previsão de séries temporais (CHAI, T.; Draxler, R.; 2014).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}} \tag{4}$$

O conjunto de dados foi separado em conjunto de treino (75%), no qual treinamos os modelos e conjunto de teste (25%), no qual testamos seus desempenhos.

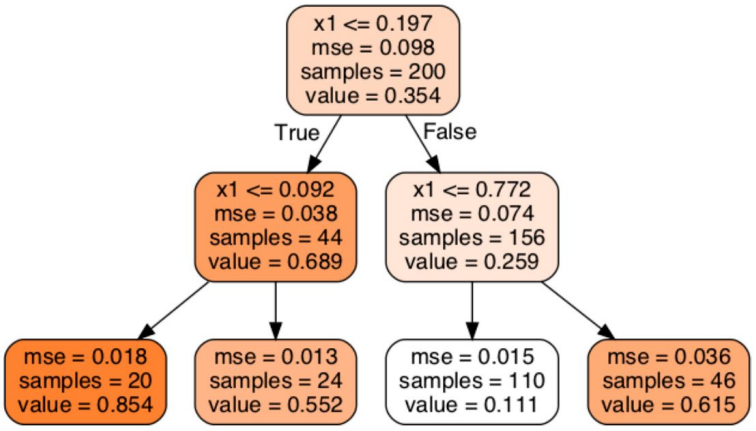
Como apresentado anteriormente, o objetivo é analisar o desempenho do método de aprendizagem de máquina *LightGMB*. Para servir de base de comparação para o modelo, calculamos qual seria o RMSE se apenas repetíssemos os valores de radiação das 24 horas anteriores. Este método é conhecido como *Naive method* (Método ingênuo), devido à simplicidade com que se estima os valores seguintes da série. Utilizando esse método obtivemos:

$$\text{RMSE}_{\text{Naive}} = 0.4539 \tag{5}$$

O modelo de aprendizagem de máquina escolhido para este problema foi o *LightGBM*. *LightGBM* (KE, Guolin. et al.; 2016) é uma implementação do tipo *Gradient Boosting Decision Tree* (árvore de decisão com impulsão de gradiente). Esse modelo aplica uma série de otimizações que reduz seu tempo de processamento em até 20 vezes quando comparado a implementações anteriores de GBDT, como a também muito utilizada *XGBoost* (CHEN, Tianqi.; GUESTRIN, Carlos.; 2016).

Uma árvore de decisão é uma estrutura do tipo fluxograma na qual cada nó interno representa um "teste" em um atributo, cada ramificação representa o resultado do teste e cada nó folha representa uma saída (decisão tomada após calcular todos os atributos)(BRID, Rajesh; 2018). Uma ilustração da estrutura do modelo pode ser observado na Figura 3:

Figura 3 - Exemplo de árvore de decisão.



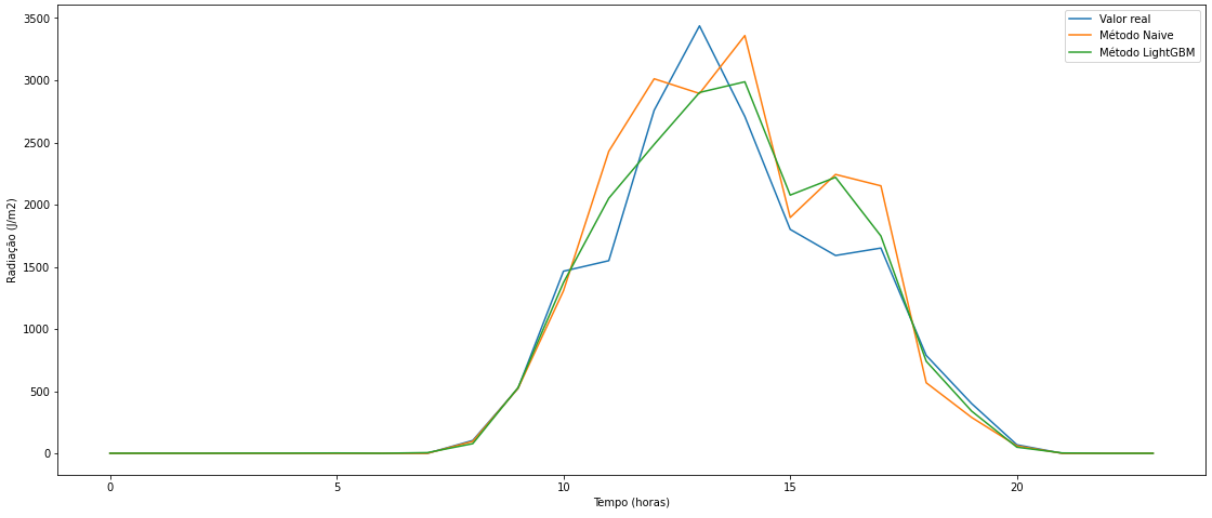

Fonte: GERON, Aurélien; 2019.

O modelo foi então aplicado aos dados de teste, e o seguinte resultado foi obtido:

$$RMSE_{LightGBM} = 0.7133 \quad (6)$$

Um exemplo das previsões são mostrados a seguir na Figura 4:

Figura 4 - Exemplo de resultado dos modelos preditivos.




Fonte: Elaborado pelo autor (2020)

## 4 CONCLUSÃO

Os resultados obtidos, resumido na Tabela 1 demonstram que, apesar da grande quantidade de dados disponíveis, o modelo *LightGBM* não conseguiu superar o desempenho do método *Naive*.

Tabela 1 - Resultados dos modelos preditivos

| Método | RMSE |
|---|---|
| *Naive* | 0.4539 |
| *LightGBM* | 0.7133 |

Fonte: Elaborado pelo autor (2020)

A principal hipótese para esse resultado é que, devido a enorme regularidade nos dados de radiação solar, o método *Naive*, apesar de simples, torna-se um preditor muito preciso.

Possíveis próximos estudos podem analisar a diferença no comportamento dos dois métodos em conjuntos de dados com diferentes regularidades, para verificar se o modelo *LightGBM* supera o método *Naive* em dados com altas variações.

## REFERÊNCIAS

BRID, Rajesh. Decision Trees - A simple way to visualize a decision. Disponível em: https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb#:~:text=A%20decision%20tree%20is%20a%20flowchart%2Dlike%20structure%20in%20which,taken%20after%20computing%20all%20attributes). Acesso em 25 de Julho de 2020.

CHAI, T.; Draxler, R.; "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature", **Geoscientific Model Development,** 2014.

CHEN, Tianqi.; GUESTRIN, Carlos. XGBoost: A Scalable Tree Boosting System. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**, 2016.

GERON, Aurélien. **Hands On Machine Learning with Scikit-learn and Tensorflow.** 2ª edição, Sebastopol, CA, EUA, O'Reilly Media, 2019.

KE, Guolin.; MENG, Qi.; FINLEY, Thomas.; WANG, Taifeng.; CHEN, Wei.; MA, Weidong.; YE, Qiwei.; LIU, Tie-Yan.; "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", **Advances in Neural Information Processing Systems**, 2017.

SHARMA, N.; SHARMA, P.; IRWIN, D.; SHENOY, P.; "Predicting Solar Generation from Weather Forecasts Using Machine Learning", I**EEE International Conference on Communications, Control, and Computing Technologies for Smart Grids**, 2011.

Banco de Dados Meteorológicos do INMET. Disponível em: https://bdmep.inmet.gov.br/. Acesso em: 11 de Maio de 2020.

International Energy Agency, **"Snapshot of Global Photovoltaic Markets"**, 2016.

# FEATURE SELECTION AND PREDICTIVE MODELING OF SOLAR RADIATION

***Abstract:*** *The enormous scientific and technological advance of the last few years has allowed the production of solar energy to become increasingly efficient and accessible. As a result, the costs of installing solar panels have been drastically reduced, and a much larger number of homes and buildings have started to produce energy in this way. However, the fact that this type of generation is highly intermittent and uncontrollable has proved to be a challenge for energy utilities, as they must constantly seek to match energy production to consumption. The purpose of this article is to propose a methodology that uses historical solar radiation data to estimate future solar energy production, thus helping the integration of distributed generation into the electricity grid. To this end, in this article, a method for selecting variables and a machine learning model (LightGBM) will be presented, which are currently the state of the art for forecasting time series.*

***Keywords:*** *Solar energy, predictive modeling, machine learning.*