# Classificação De Séries Temporais Por Meio De Random Forest Para Previsão Em Curto Prazo De Fenômenos De Sobreirradiância Em Natal-RN

Arthur Diniz Fernandes
*Programa de Pós Graduação em Engenharia Elétrica e de Computação*
*Universidade Federal do Rio Grande do Norte*
Natal, Brasil
arthur.torquato.111@ufrn.edu.br

Adriao Duarte Dória Neto
*Departamento de Engenharia de Computação e Automação*
*Programa de Pós Graduação em Engenharia Elétrica e de Computação*
*Universidade Federal do Rio Grande do Norte*
Natal, Brasil
adriao@dca.ufrn.br

Samira Emiliavaca
*Laboratório de Energia Solar*
*Instituto SENAI de Inovação em Energias Renováveis*
Natal, Brasil
samira@isi-er.com.br

Jean Reis
*Departamento de Meteorologia*
*Instituto SENAI de Inovação em Energias Renováveis*
Natal, Brasil
jeanreis@isi-er.com.br

***Abstract*—A Sobreirradiância é um fenômeno meteorológico que ocorre quando a radiação solar incidente excede significativamente os níveis normais esperados para uma determinada região ou período de tempo. Esse evento vem despertado interesse acadêmico devido aos possíveis danos econômicos causados em Unidades de Geração Fotovoltaicas de larga escala. O presente artigo apresenta um estudo inédito acerca da previsão em curto prazo de tal fenômeno.**

**Durante essa análise se adotou como procedimento principal um esquema de classificação de séries temporais alimentado por um Dataset de variáveis climatológicas colhido na cidade de Natal-RN por meio de uma estação Solarimetrica localizada no Instituto SENAI de inovação em energias renováveis por um período de 7 anos e com uma resolução de 1 minuto**

***Index Terms*—Random Forest, Sobreirradiância, Energia Solar, Machine Learning, Iradiancia**

## I. Introdução

A geração de energia fotovoltaica está passando por um período de notável expansão no Brasil, com a previsão de alcançar uma capacidade instalada entre 27 GW e 90 GW até 2050, conforme projeções da Empresa de Pesquisa Energética (EPE) [1]. Contudo, para fazer frente a essa expansão, é necessário abordar uma série de desafios que impactam o setor, entre os quais se destaca a influência de fatores climáticos na produção de energia. Um fenômeno meteorológico recentemente observado, denominado Sobreirradiância, tem despertado interesse nesse contexto.

A Sobreirradiância ocorre em dias parcialmente nublados, quando a quantidade de radiação que atinge a superfície supera os valores típicos de dias ensolarados [2]. Em circunstâncias extremas, esse excesso de radiação pode inclusive exceder os níveis de operação nominal dos painéis fotovoltaicos, acarretando potenciais danos, como demonstrado em um estudo de Nascimento [3]. Isso pode resultar em falhas nos componentes, perdas de eficiência nos inversores devido à sobrecarga e comprometimento da eficiência do Maximum Power-Point Tracker (MPPT).

Diversas pesquisas têm sido dedicadas à observação e monitoramento desse fenômeno, documentando casos notáveis em várias regiões, incluindo Havaí, Chipre, São Paulo e Natal [4] [5] [6] [7] [8]. No entanto, a literatura carece de abordagens direcionadas à previsão antecipada desse fenômeno, conforme evidenciado por uma busca nas bases de dados científicas.

Nesse contexto, o presente estudo propõe uma abordagem baseada em Aprendizado de Máquinas para a previsão da ocorrência de Sobreirradiância. O método empregado envolve a utilização do algoritmo de Random Forests em uma estrutura de classificação de séries temporais, utilizando dados climatológicos coletados em Natal-RN, os quais foram previamente empregados em estudo conduzido por Costa [8]. Adicionalmente, este estudo inova ao elaborar uma lista das variáveis mais significativas para a previsão do fenômeno.

Esse trabalho foi parcialmente financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

A pesquisa é organizada em seções distintas: a seção de Materiais e Métodos expõe a metodologia empregada na aquisição e processamento dos dados, no treinamento e análise do algoritmo, e na geração do ranking das variáveis relevantes. Segue-se uma seção dedicada à apresentação e análise dos resultados obtidos, culminando em conclusões fundamentais derivadas deste estudo.

## II. Materiais e Métodos

A seção a seguir delineia a metodologia central empregada na coleta de dados e no treinamento do algoritmo de Random Forests. Esta seção é subdividida em três subseções distintas: a primeira delas discute a instrumentação adotada para a construção do conjunto de dados (Dataset), esclarecendo as ferramentas e dispositivos empregados nesse processo além de explananar como a Sobreirradiância é mensurada. A segunda subseção detalha minuciosamente todas as etapas de pré-processamento aplicadas aos dados coletados, abordando aspectos relacionados ao tratamento, à limpeza e à formatação dos dados, que os prepararam para análises subsequentes. Por fim, a terceira subseção oferece uma exposição acerca do procedimento de treinamento empregado no algoritmo em foco.

### A. Coleta de Dados e Cálculo de Sobreirradiância

O Dataset empregado nesta pesquisa foi obtido e processado a partir de registros obtidos entre os anos de 2015 e 2019. Esses registros foram adquiridos por meio de uma estação Solarimétrica localizada na cidade de Natal-RN, situada nas dependências do instituto SENAI de inovação em Energias Renováveis. A Tabela I apresenta uma relação dos sensores que foram utilizados nesse procedimento. Devido às distintas resoluções asseguradas pelos fabricantes dos dispositivos, optou-se por adotar uma taxa de integração dos dados de um minuto.

A partir das variáveis coletadas, conduziu-se uma análise para determinar a incidência de Sobreirradiância em cada conjunto de dados. Para tal, procedeu-se à comparação entre os valores de Irradiância Global Horizontal ($I_g$) e os valores estimados de Irradiância Extraterrestre ($I_0$). Nesse contexto, os valores de $I_g$ que superaram os valores correspondentes de $I_0$ foram identificados como casos de Sobreirradiância. A obtenção dos valores de $I_0$ foi realizada através da aplicação da Equação 1, sendo que as variáveis $e_0$ e Cos($\Theta_z$) foram previamente calculadas e extraídas de um estudo conduzido por Costa [8], que forneceu os dados empregados nessa pesquisa.

$$I_{sc} * e_0 * cos(\Theta_z) = I_0 \quad (1)$$

- $I_{sc}$ = constate solar com valor de 1367 w/m²
- $e_0$ = excentricidade da órbita da terra
- $\Theta_z$ = ângulo zenital

TABLE I: Instrumentos utilizados na coleta do Dataset utilizado nesse estudo

| INSTRUMENTO | MODELO | FABRICANTE | VARIÁVEL MEDIDA OU FUNÇÃO DESEMPENHADA |
|---|---|---|---|
| Datalogger | CR3000 | Campbell Scientific | Aquisição de dados |
| Rastreador | Solys 2 | Kipp & Zonen | Suporte e Rastreio |
| Piranômetro | CPM 22 | Kipp & Zonen | Irradiância Global Horizontal |
| Piranômetro | CPM 22 | Kipp & Zonen | Irradiância Difusa Horizontal |
| Pireliômetro | CHP 1 | Kipp & Zonen | Irradiância Direta Normal |
| Pirgeômetro | CGR 4 | Kipp & Zonen | Radiação de Onda Longa Descendente |
| Barômetro | PTB 110 | Vaisala | Pressão Atmosférica |
| Termohigrômetro | 41382VC | R. M. Young | Temperatura e Umidade Relativa do ar |
| Pluviômetro | TB4-L | Campbell Scientific | Precipitação |
| Anemômetro | Windsonic | Campbell Scientific | Intensidade e Direção do Vento |

### B. Pré-Processamento de Dados

Para realizar o pré-processamento dos dados, foi implementado um procedimento padrão com etapas delineadas na Figura 1. Os próximos tópicos abordarão cada uma dessas etapas de maneira individual.

*1) Remoção de colunas incompletas, Consolidação dos Dados e Seleção de Horários de Interesse:* Devido ao grande período de coleta dos dados empregados nesta pesquisa, é fundamental exercer uma atenção especial ao utilizá-los para treinar um algoritmo de aprendizado de máquinas. Nesse contexto, torna-se imperativo realizar um estudo para identificar quais variáveis medidas são apropriadas para inclusão no treinamento.

Inicialmente, foi conduzida uma análise de completude dos dados, categorizada por ano de coleta. Esse procedimento revelou que ocorrem lacunas na coleta de dados em anos específicos (conforme exemplificado na Figura 2). Para consolidar a análise ao longo dos anos, é necessário estabelecer um denominador comum mínimo de completude. Isso ocorre em razão da natureza das falhas de coleta que podem variar ao longo do tempo.

Consequentemente, optou-se por estabelecer um critério no qual, ao consolidar os dados, uma variável em análise deve estar presente em pelo menos 90% dos registros para ser considerada relevante para a análise do fenômeno. Caso contrário, será excluída da análise. Com base nesse critério, as grandezas seguintes foram selecionadas como variáveis para

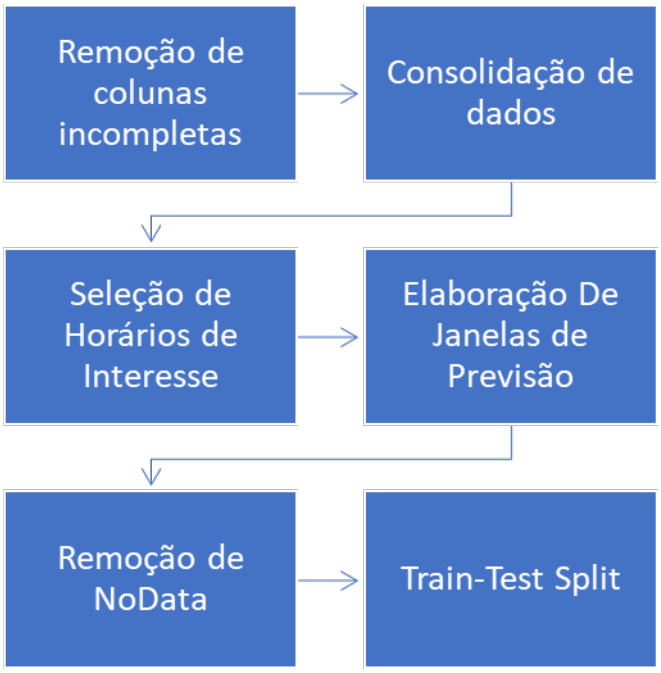


Fig. 1: Esquemático da Etapa de Pré-Processamento de Dados

análise:

- Temperatura do Ar
- Umidade do Ar
- Pressão
- Precipitação Total
- Irradiância Global Média
- Irradiância Direta Média
- Irradiância Difusa Média
- Irradiância de Onda Longa
- Hora de ocorrência da medição do dado
- Mês de medição do dado

Após a seleção das variáveis, os dados que originalmente estavam segmentados por anos são unificados em um único conjunto. Adicionalmente, são eliminados dessa coleção os registros adquiridos em horários nos quais a incidência de irradiância solar é inexistente. Dado que a localização em questão se situa em uma zona equatorial, caracterizada por variações temporais reduzidas na luz solar, determinou-se que somente os dados coletados no intervalo das 6h da manhã até as 18h serão considerados para este estudo.

*2) Elaboração da Janela de Previsão e Regressão:* Após as etapas iniciais, cada linha de dados passa por um processo no qual observações anteriores ao dado em questão são incorporadas. A Figura 3 ilustra um exemplo desse procedimento utilizado para a variável de Temperatura do Ar do conjunto de dados.

Além disso, nessa etapa, é definida a variável alvo. No decorrer desse procedimento, verifica-se a ocorrência dos fenômenos de Cloud-Enhancement e Sobreirradiância até 5 minutos após a coleta do dado em análise. Na eventualidade da constatação de algum desses eventos, atribui-se o valor 1 para Cloud-Enhancement e 2 para Sobreirradiância na coluna correspondente à variável alvo. A Figura 4 exemplifica esse processo, destacando a presença do fenômeno de Cloud-Enhancement e o subsequente registro na coluna de variável alvo.

A tabela II abaixo apresenta a definição das variáveis Y.

TABLE II: Definição das variáveis alvo

| Y | Definição |
|---|---|
| 0 | Cenário normal sem a ocorrência de evento de Sobreirradiância ou cloud-enhancement |
| 1 | Cenário com ocorrência de Cloud-enhancement |
| 2 | Cenário com ocorrência de Sobreirradiância acima de 1000 w/m² |

*3) Remoção de NoData e Train-Test Split:* Por fim, ocorre a exclusão das linhas que contenham valores No-Data, seguida pela divisão dos dados em conjuntos de treinamento, teste e validação. Para o conjunto de validação, os dados balanceados do ano de 2019 foram utilizados. Para a preparação do treinamento e teste do algoritmo, os dados de 2015 a 2018 foram utilizados. Inicialmente, devido ao desequilíbrio acentuado nos dados, ocorreu uma etapa de balanceamento. Isso resultou em uma proporção equitativa de 1/3 de dados sem a ocorrência dos fenômenos em análise, 1/3 com ocorrência de Cloud-Enhancement e 1/3 com ocorrência de Sobreirradiância. Após esse procedimento, os dados foram divididos aleatoriamente, com 70% destinados ao treinamento e 30% para os testes. Com essa etapa concluída, os dados estão prontos para serem empregados no algoritmo de Random Forest.

### *C. Treinamento de Algoritmo*

Para o treinamento do modelo se escolheu um algoritmo de Random Forest, essa escolha se deu devido a possibilidade de produção de indicadores de importância das variáveis avaliadas [9], que serão utilizados para munir estudos posteriores acerca do tema, a sua rápida implementação e a sua boa capacidade de generalização de resultado.

O algoritmo de Random Forest foi apresentado por Breiman no ano de 2001 e é definido como: “uma combinação de preditores de árvore, de forma que cada árvore depende dos valores de um vetor aleatório amostrado de forma independente e com a mesma distribuição para todas as árvores na floresta” [10]. Ele opera a partir do instanciamento do dataset e do treinamento de um conjunto de preditores fracos que ao serem associados a uma função de seleção conseguem entregar resultados com melhor acurácia e generalização. A Figura 5 apresenta um esquemático contendo o funcionamento simplificado do algoritmo

O treinamento do algoritmo foi conduzido utilizando a implementação padrão do Classificador de Random Forests da biblioteca SciKit-Learn [9]. Nesta fase do estudo, o objetivo não é explorar a otimização do algoritmo, mas sim avaliar a viabilidade de prever a ocorrência de Sobreirradiância. Portanto, foram adotados os hiperparâmetros padrões da biblioteca, sendo os principais:

- Numero de Estimadores:100

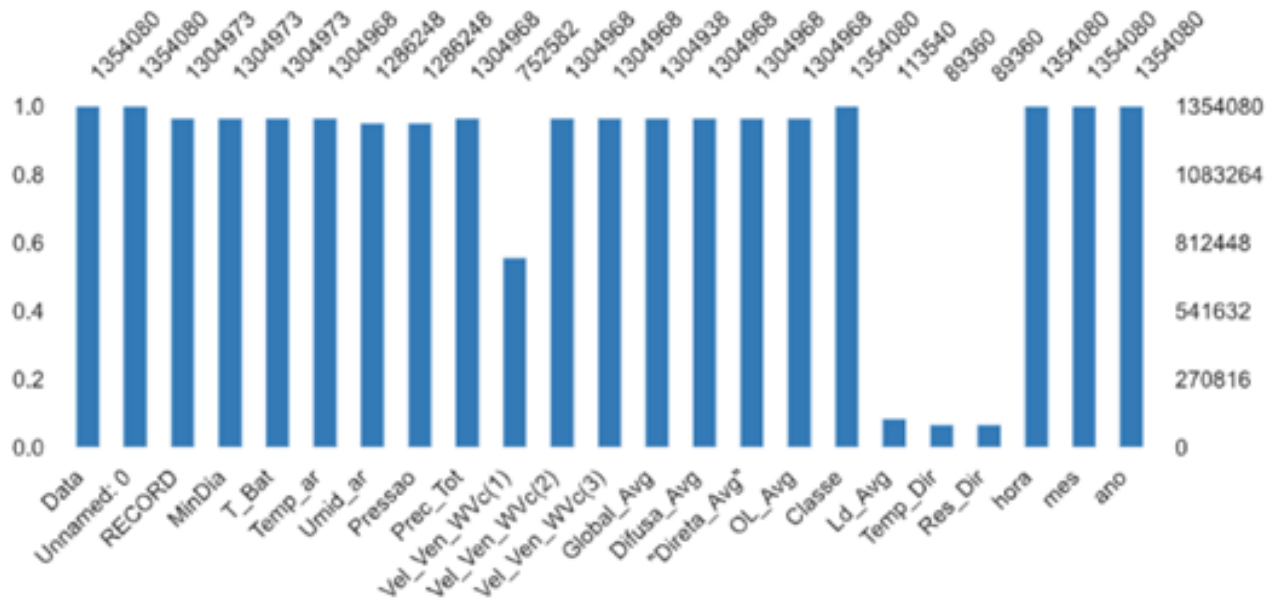


Fig. 2: Tabela de completude de dados do Dataset.

| | Temp_ar | Temp_ar-1 | Temp_ar-2 | Temp_ar-3 | Temp_ar-4 | Temp_ar-5 |
|---|---|---|---|---|---|---|
| **18302** | 28.63 | 28.66 | 28.55 | 28.55 | 28.55 | 28.52 |
| **511960** | 29.87 | 29.70 | 29.96 | 30.10 | 29.83 | 29.77 |
| **209773** | 28.96 | 28.98 | 28.89 | 28.89 | 28.88 | 28.87 |
| **441177** | 28.78 | 28.67 | 28.49 | 28.40 | 28.33 | 28.22 |
| **327533** | 28.20 | 28.28 | 28.45 | 28.28 | 28.06 | 27.96 |

5 rows × 50 columns

Fig. 3: Exemplo de elaboração de série temporal para classificação com variável de temperatura do Ar.

| y+1 | y+2 | y+3 | y+4 | y+5 | y_combine |
|---|---|---|---|---|---|
| 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 1 |
| 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 1 |
| 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 1 |
| 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 1 |
| 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1 |

Fig. 4: Criação de dado para variável alvo.

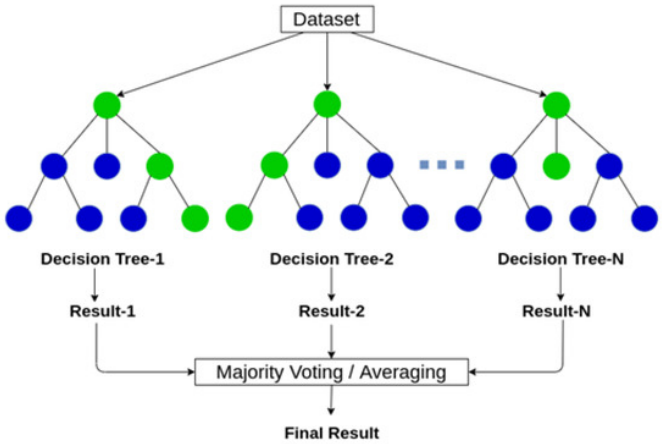


Fig. 5: Esquema simplificado de Árvore de Decisão.

Fonte: Retiradada de: [11]

- Critério: Gini
- Profundidade Máxima: Nenhuma

## III. Resultados

A seguir, serão apresentados os resultados obtidos neste estudo. A seção A aborda os resultados da análise de importância das variáveis, detalhando o modelo empregado para essa avaliação. Por outro lado, a seção B concentra-se na performance do algoritmo, considerando a utilização de dados de teste e validação. A exposição dos resultados será realizada por meio das métricas de acurácia, matriz de confusão, recall global, precisão global e F1 global.

### A. *Análise de Importância de Variáveis*

Para avaliar a importância das variáveis, foi empregado o método ”mean decrease in impurity”, conforme recomendado na documentação da biblioteca Sci-Kit Learn [9].

TABLE III: Importância das variáveis por Mean Decrease in Impurity

| Variavel | Mean decrease in Impurity |
|---|---|
| Irradiância Difusa Média no instante T=0 minutos | 0.082638 |
| Irradiância Difusa Média no instante T=-1 minutos | 0.079477 |
| Valor da Hora no instante T=0 | 0.066243 |
| Irradiância Difusa Média no instante T=-2 minutos | 0.052760 |
| Irradiância Difusa Média no instante T=-3 minutos | 0.048480 |

### B. *Performance de Algoritmo*

O algoritmo produz um resultado numérico que sinaliza a possibilidade de um dos três cenários a seguir:

- Se o resultado for 0, o algoritmo indica que não se espera a ocorrência de nenhum fenômeno meteorológico avaliado nos próximos 5 minutos.
- Se o resultado for 1, o algoritmo aponta a possibilidade de o fenômeno de cloud-enhancement ocorrer nos próximos 5 minutos.
- Se o resultado for 2, o algoritmo sugere a possibilidade de o fenômeno de Sobreirradiância ocorrer nos próximos 5 minutos, com um valor superior a 1000 W/m².

Quando consideramos os dados de teste, a acurácia global foi aproximadamente 90%, um valor considerado satisfatório para prever o fenômeno em estudo. Entretanto, é relevante destacar a significativa prevalência de aproximadamente 17% de falsos negativos para a classe normal. Além disso, a acurácia na detecção de Sobreirradiância atingiu cerca de 95%. Esses resultados podem ser visualizados na matriz de confusão apresentada na Figura 6.

Ao analisar os dados de validação, nota-se uma diminuição no desempenho do algoritmo, resultando em uma acurácia global de aproximadamente 79,4%. Além disso, é relevante notar que mesmo com essa redução no desempenho, o algoritmo ainda alcança uma acurácia de 85% na previsão de Sobreirradiância, conforme pode ser observado na matriz de confusão apresentada na Figura 7.

A Tabela IV fornece uma síntese das métricas de desempenho para os conjuntos de dados de teste e validação, tendo em mente a distribuição uniforme das classes. No contexto dos dados de teste, as métricas de precisão, recall e F1-score exibem valores consistentemente elevados, em torno de 0,88, indicando uma capacidade sólida do modelo de generalizar seus resultados para novos dados. No entanto, é essencial destacar que esses resultados são baseados em uma distribuição uniforme das classes e podem não se traduzir diretamente para cenários reais. Em relação aos dados de

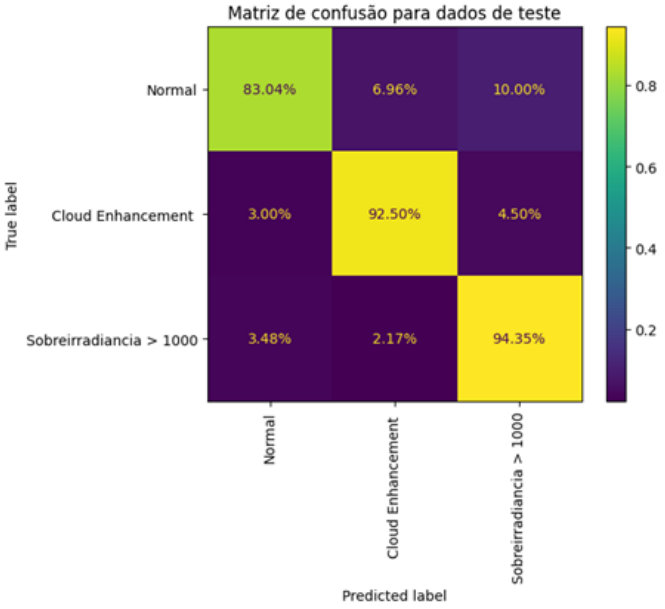


Fig. 6: Matriz de confusão para conjunto de testes.

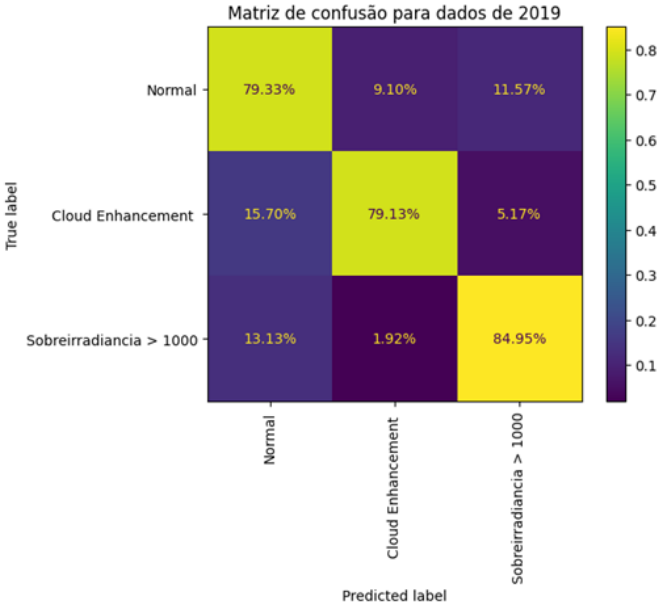


Fig. 7: Matriz de confusão para dados de validação (ano de 2019).

validação, enquanto a precisão permanece alta, aproximando-se de 0,99, o recall apresenta uma leve diminuição, sugerindo que o modelo pode encontrar mais dificuldades em identificar corretamente casos positivos. Esse cenário resulta em um F1-score de 0,87, indicando que o modelo ainda apresenta um desempenho razoável, mas poderia ser otimizado para melhor sensibilidade em relação aos casos positivos em contextos mais reais.

TABLE IV: Importância das variáveis por Mean Decrease in Impurity

| Conjunto | Precisão | Recall | F1 |
|---|---|---|---|
| Dados de Teste (Considerando Balanceamento dos Dados) | 0.88 | 0.88 | 0.88 |
| Dados de Validação (Considerando Balanceamento dos Dados) | 0.99 | 0.79 | 0.87 |

## IV. Conclusões

Ao se analisar os resultados se observa que a abordagem escolhida foi bem-sucedida para o problema em questão e demonstram a possibilidade de previsão do fenômeno de sobre irradiância em um curto prazo por meio de técnicas de aprendizagem de máquina, entretanto destaca-se que o modelo foi testado somente com dados balanceados de forma que ele não é operacionalizado em cenários reais, dessa forma futuros estudos devem se pautar na utilização de diferentes algoritmos de aprendizagem de máquinas para analisar a performance quando comparados a Random Forests, dentre os quais pode-se mencionar:

- Redes Neurais de tipo Long Short Term Memory (LSTM)
- Maquinas de Vetor de Suporte (SVM)
- XGBoost

## V. Agradecimentos

Esse trabalho foi feito com o apoio de dados concedidos pelo Instituto SENAI de Inovação em Energias Renováveis e coletados a partir de projeto de pesquisa realizado em parceria com a China Three Gorges Corporation (CTG). Sua realização foi feita como parte de um projeto de Mestrado financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico.

## References

[1] M. de Minas e Energia (MME). e EPE, *Plano nacional de energia PNE 2050*. Brasília:: MME, EPE,, 2020. [Online]. Available: http://biblioteca.olade.org/opac-tmpl/Documentos/cg00877.pdf

[2] M. Järvelä, K. Lappalainen, and S. Valkealahti, "Characteristics of the cloud enhancement phenomenon and pv power plants," *Solar Energy*, vol. 196, pp. 137–145, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038092X19311909

[3] L. R. do Nascimento, T. de Souza Viana, R. A. Campos, and R. Rüther, "Extreme solar overirradiance events: Occurrence and impacts on utility-scale photovoltaic power plants in brazil," *Solar Energy*, vol. 186, pp. 370–381, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038092X19304530

[4] R. D. Piacentini, G. M. Salum, N. Fraidenraich, and C. Tiba, "Extreme total solar irradiance due to cloud enhancement at sea level of the ne atlantic coast of brazil," *Renewable Energy*, vol. 36, no. 1, pp. 409–412, 2011. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S096014811000265X

[5] R. H. Inman, Y. Chu, and C. F. Coimbra, "Cloud enhancement of global horizontal irradiance in california and hawaii," *Solar Energy*, vol. 130, pp. 128–138, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038092X16001079

[6] M. P. Almeida, R. Zilles, and E. Lorenzo, "Extreme overirradiance events in são paulo, brazil," *Solar Energy*, vol. 110, pp. 168–173, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0038092X14004423

[7] R. Andrade and C. Tiba, "Extreme global solar irradiance due to cloud enhancement in northeastern brazil," *Renewable Energy*, vol. 86, 10 2015.

[8] M. E. N. R. da Costa, S. d. A. S. Emiliavaca, I. T. A. de Freitas, P. A. A. de Araújo, and P. R. Mutti, "AnÁlise da ocorrÊncia de eventos de sobreirradiÂncia em natal-rn, brasil," in *2021 21 Congresso Brasileiro de Meteorologia*, 2021.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[10] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct 2001. [Online]. Available: https://doi.org/10.1023/A:1010933404324

[11] L. Hao, J. Kim, S. Kwon, and I. D. Ha, "Deep learning-based survival analysis for high-dimensional survival data," *Mathematics*, vol. 9, no. 11, 2021. [Online]. Available: https://www.mdpi.com/2227-7390/9/11/1244