

Bolsista: Enzo Laragnoit Fernandes

Previsão de Irradiação Solar para Sistemas Fotovoltaicos utilizando *Machine Learning*

**Uma abordagem de Aprendizagem Supervisionada para Previsão de
Irradiação Solar no Estado de São Paulo**

FAPESP

Fundação de Amparo à Pesquisa do Estado de São Paulo

Processo: 2020/09607-9

Vigência: 01/09/2020 a 31/08/2021

Relatório Científico de Progresso

Período: 01/09/2020 a 10/02/2021

Fevereiro de 2021

1 Resumo do Projeto Proposto

Um dos principais desafios para o século XXI é conciliar a busca por soluções energéticas com o desenvolvimento sustentável sem causar maiores prejuízos ao já afetado meio ambiente. Com a crescente demanda por energia elétrica (EPE, 2017), movida pela progressiva informatização da sociedade moderna, surgem diferentes formas renováveis de geração de energia elétrica em alternativa aos finitos recursos não renováveis como carvão mineral e petróleo.

Uma das alternativas promissoras é a conversão de energia solar em energia elétrica por meio de painéis solares fotovoltaicos que transformam a irradiação solar em energia elétrica que pode ser prontamente utilizada ou armazenada em baterias sob a forma de energia química. Presente em grande abundância e em praticamente todos os lugares, a quantidade de energia incidente na Terra proveniente do Sol é aproximadamente 10 000 vezes maior que a quantidade de energia consumida pela humanidade (SMETS et al., 2015). A energia solar apresenta vantagens quando comparada a outras formas de energia renovável, como a eólica e a hidrelétrica, pela facilidade de instalação, considerando aspectos de construção, danos ambientais, custos operacionais e manutenção. Além disso, apresenta a possibilidade de ser utilizada sem grandes dificuldades nos ambientes urbanos e também oferece como alternativa um modelo de geração de energia descentralizada, no qual consumidores podem gerar sua própria energia. O aumento da eficiência energética e a constante redução dos custos dos sistemas fotovoltaicos justificam sua tendência de crescimento e inserção nas matrizes energéticas (ACHILLES, 2013).

Outra área de grandes avanços e crescentes expectativas é a de Inteligência Artificial, especialmente a subárea de Aprendizado de Máquina (PERRAULT et al., 2019). Possibilitada pelo aprimoramento das técnicas e pelo aumento do poder computacional dos sistemas modernos, sua presença em aplicações utilizadas no cotidiano como reconhecimento de imagens, processamento de linguagem natural, *chatbots* e sistemas de recomendação tornou-se praticamente universal em softwares nas mais diversas plataformas. Grande esperança é depositada nos *insights* e resultados obtidos da aplicação dessas técnicas para trabalhar com questões na fronteira do conhecimento de tópicos fundamentais como energia, saúde e economia em consonância com a computação científica.

Dentre as aplicações possibilitadas pelo uso das técnicas de Aprendizado de Máquina destaca-se, no contexto deste projeto, o uso de Aprendizado de Máquina Supervisionado (*Supervised Machine Learning*) na competência de aferir previsões a partir de um conjunto de dados já conhecidos *a priori*. Essa tarefa é realizada utilizando diferentes algoritmos como em (HAYKIN, 1994; SUYKENS; VANDEWALLE, 1999), cujo propósito geral é obter uma função $f : X \rightarrow Y$ capaz de realizar o mapeamento entre dados de entrada, chamados de atributos (*features*), e as variáveis alvo (*targets*) cujos valores se objetiva determinar. Para este projeto, as informações relacionadas aos dados meteorológicos serão utilizadas como os dados de entrada X da função f e a radiação solar horizontal global será a variável alvo (*target variable*) a ser determinada.

O projeto foi baseado em aplicações similares dessas técnicas em diferentes contextos e localidades como (ALZAHIRANI et al., 2017; LI et al., 2016; ASSOULINE; MOHAJERI; SCARTEZZINI,

2017; LOU et al., 2016; ZENG; QIAO, 2013), na intersecção entre as áreas de Inteligência Artificial e Energia Solar.

2 Objetivos e Cronograma

O objetivo desse projeto consiste em utilizar técnicas de Aprendizado de Máquina a fim de implementar modelos capazes de prever a irradiação solar, medida em kJ/m^2 , em intervalos de antecedência de 60 minutos em pontos específicos no estado de São Paulo. Para tal, foram estipuladas as metas abaixo cujo cronograma é apresentado na Tabela 1.

1. Obtenção do conjunto de dados;
2. Geração do conjunto de dados preprocessados;
3. Seleção dos algoritmos para a implementação dos modelos;
4. Implementação dos modelos de previsão a partir de diferentes algoritmos de aprendizagem supervisionada selecionados;
5. Teste e avaliação da performance dos modelos em relação ao conjunto de dados obtidos *a priori*;
6. Ajuste dos modelos com base nas informações do item 5;
7. Teste e avaliação da performance dos modelos na previsão de irradiação nos locais específicos durante 1 mês;
8. Elaboração dos Relatórios;

Tabela 1 – Cronograma de execução do projeto. As metas marcadas em verde foram concluídas e compõem este Relatório Científico de Progresso, enquanto as em laranja serão desenvolvidas até o final da vigência do projeto.

Metas	Meses											
	1	2	3	4	5	6	7	8	9	10	11	12
1												
2												
3												
4												
5												
6												
7												
8												

Fonte: Elaborada pelo autor.

3 Realizações no Período

3.1 Obtenção do Conjunto de Dados

O processo de obtenção dos dados foi realizado em três etapas: **1:** elencar possíveis fontes de dados meteorológicos dentre empresas e órgãos públicos; **2:** obter conjuntos de coordenadas para seleção de possíveis pontos de coleta de dados; e **3:** realizar o cruzamento entre **1:** e **2:** para auxiliar a seleção.

3.1.1 Busca por Possíveis Fontes de Dados

Destacaram-se como possíveis fontes:

Atlas Brasileiro de Energia Solar elaborado pelo LABREN – Laboratório de Modelagem e Estudos de Recursos Renováveis de Energia¹.

SONDA – Sistema de Organização de Dados Ambientais que disponibiliza dados coletados por estações meteorológicas automáticas.

Climatempo empresa que disponibiliza dados por meio do seu portal Data Clima².

IAC – Instituto Agronômico de Campinas que integra e disponibiliza dados através do portal CIIAGRO – Centro Integrado de Informações Agrometeorológicas³;

Projeto POWER variáveis radiométricas e meteorológicas obtidas via satélite em seu portal⁴;

INMET – Instituto Nacional de Meteorologia que disponibiliza⁵ dados coletados por estações automáticas e convencionais em todo o país;

Solcast empresa⁶ que distribui comercialmente dados coletados por terceiros.

As fontes mais promissoras foram o INMET, o IAC e a Solcast, pela disponibilidade de diversas variáveis meteorológicas em diversos locais, com frequência de amostragem adequada. As demais foram descartadas por não atenderem aos requisitos elencados acima. Os dados do INMET foram obtidos por meio de sua página⁷; os dados do IAC não puderam ser obtidos pois não houve resposta ao pedido realizado; por fim, a empresa Solcast disponibilizou uma quantidade de dados insuficiente e, portanto, foi descartada.

3.1.2 Elaboração dos Conjuntos de Coordenadas

A fim de auxiliar a escolha do conjunto de dados, os seguintes critérios foram adotados:

1. **A proximidade à linhas de transmissão de energia elétrica;**
2. **A proximidade aos centros urbanos mais populosos.**

¹ http://labren.ccst.inpe.br/atlas_2017.html

² <https://www.climatempoconsultoria.com.br/levantamento-de-dados-meteorologicos/>

³ <http://www.ciiagro.sp.gov.br/climasp.html>

⁴ <https://power.larc.nasa.gov/>

⁵ <https://portal.inmet.gov.br/>

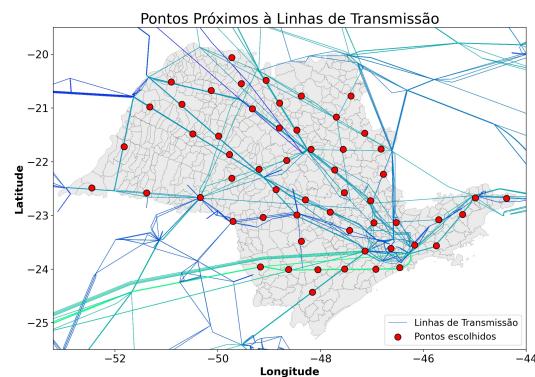
⁶ <https://solcast.com/>

⁷ Disponível em: <https://portal.inmet.gov.br/dadoshistoricos>. Acesso em 09/10/2020.

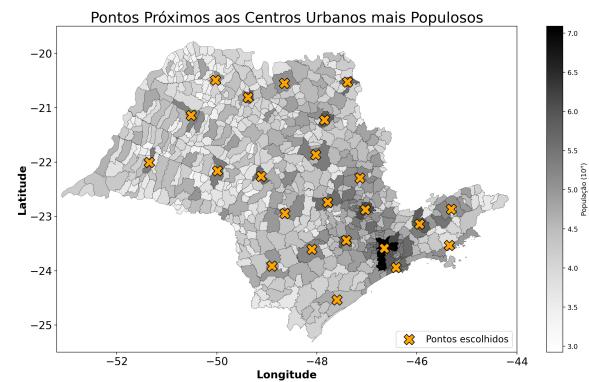
Para o Critério 1 foram obtidas informações sobre a disposição das redes de transmissão da Agência Nacional de Energia Elétrica – ANEEL⁸ (Figura 1a). Para o Critério 2 foram obtidos dados do Instituto Brasileiro de Geografia e Estatística – IBGE⁹ a partir dos quais foi possível determinar os pontos mais próximos aos centros urbanos mais populosos (Figura 1b).

Figura 1 – Conjunto de coordenadas.

(a) Gerado considerando o Critério 1, proximidade a linhas de transmissão.



(b) Gerado considerando o Critério 2, proximidade a centros urbanos.



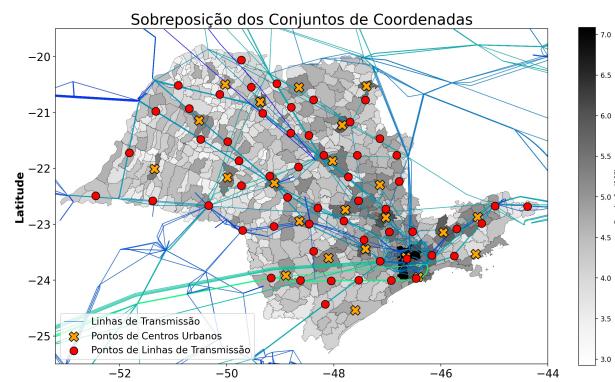
Fonte: Elaborada pelo autor, usando dados obtidos do INMET e do IBGE.

3.1.3 Cruzamento das Informações e Coordenadas

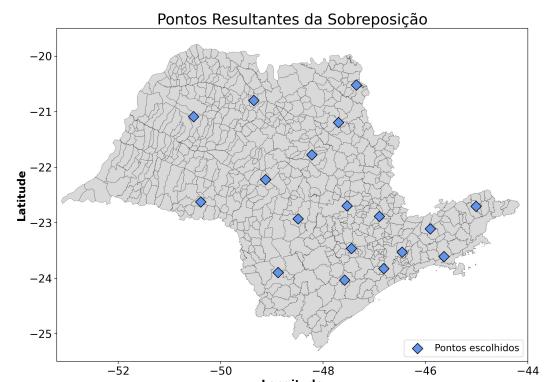
A Figura 2a é o resultado da sobreposição dos pontos a fim de determinar locais cujos pontos estabelecidos pelos Critérios 1 e 2 fossem próximos (ou coincidentes). A Figura 2b apresenta os pontos resultantes da sobreposição dos conjuntos de coordenadas.

Figura 2 – Sobreposição dos Critérios 1 e 2.

(a) Sobreposição dos conjuntos.



(b) Pontos escolhidos após a sobreposição.



Fonte: Elaborada pelo autor, usando dados obtidos da ANEEL e do IBGE.

3.1.4 Escolha da Fonte de Dados

A Figura 3 mostra as localizações dos pontos de coleta de dados do INMET em relação aos pontos determinados a partir dos Critérios 1 e 2.

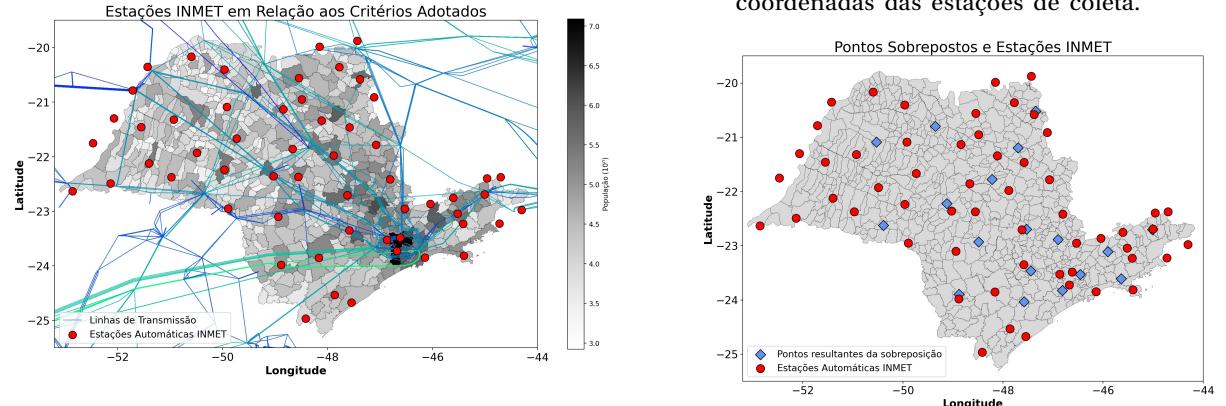
⁸ Disponível em <https://sigel.aneel.gov.br/portal/home/>. Acesso em 12/12/2020.

⁹ Disponível em <https://www.ibge.gov.br/cidades-e-estados/sp.html>. Acesso em 15/12/2020.

Figura 3 – Pontos de coleta de dados do INMET em comparação aos pontos obtidos na Seção 3.1.3.

(a) Estações de coleta, segundo Critérios 1 e 2.

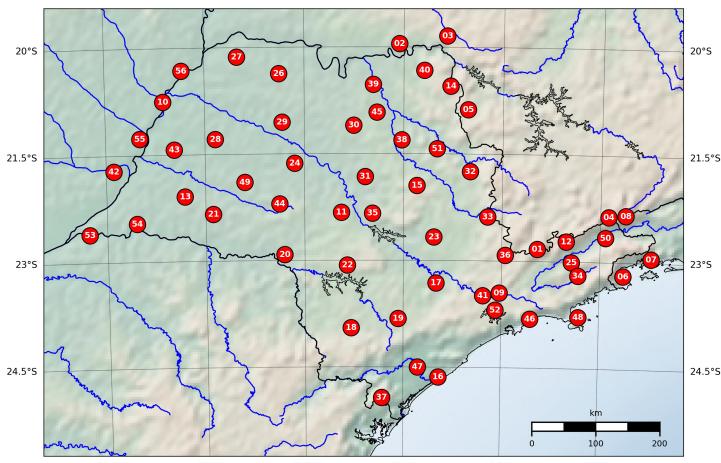
(b) Sobreposição dos pontos gerados com as coordenadas das estações de coleta.



Fonte: Elaborada pelo autor, usando dados obtidos do INMET.

A Figura 4 apresenta os pontos contemplados pelo conjunto de dados obtido. O número faz referência à Tabela 4 no Apêndice A na coluna “Nº Estação”, onde há mais informações sobre cada um dos pontos.

Figura 4 – Localização das estações de coletas de dados utilizadas como conjunto final de dados



Fonte: Elaborada pelo autor.

As observações vão de Janeiro de 2001 a Setembro de 2020 com frequência de amostragem de 1h das variáveis: **RADIAÇÃO SOLAR GLOBAL** (kJ/m^2) total de energia acumulada na hora atual; **TEMPERATURA DO AR** ($^{\circ}\text{C}$) e **PRESSÃO ATMOSFÉRICA** (hPa) instantâneas da hora atual e máximas e mínimas da hora anterior; **UMIDADE RELATIVA** (%) instantânea, máxima e mínima da hora atual; **PONTO DE ORVALHO** ($^{\circ}\text{C}$) instantâneo, máxima e mínima da hora anterior; **PRECIPITAÇÃO** (mm) total de chuva acumulada na hora atual; **VENTO** (m/s e $^{\circ}$) em relação à velocidade instantânea e rajada; **DATA** (DD/MM/AAAA); **HORA** (HHMM, UTC).

3.2 Geração do Conjunto de Dados Pré-processados

Nesta etapa foram removidos dados inconsistentes ou faltantes e eliminados atributos irrelevantes. Para tanto, foram utilizadas as bibliotecas **Pandas** para manipulação do conjunto de dados; **Numpy**

para operações de álgebra linear e **Matplotlib** para representação visual dos dados. As Seções 3.2.1 até 3.2.6 descrevem esta etapa de pré-processamento.

3.2.1 Organização e Junção dos Arquivos de Dados

Os dados foram disponibilizados agrupados por estação para cada período disponível. Foi realizada a filtragem dos dados selecionando as observações relativas ao estado de São Paulo. Também foram incluídos dados de 15 estações nas proximidades do Estado. Todos os dados foram reunidos em um arquivo agregando informações contidas nos metadados: **ALTITUDE** (m) em relação ao nível do mar; **LATITUDE** e **LONGITUDE**, ambas em base decimal; e **ID ESTAÇÃO**, composta por uma letra seguida de três números (código atribuído pelo INMET).

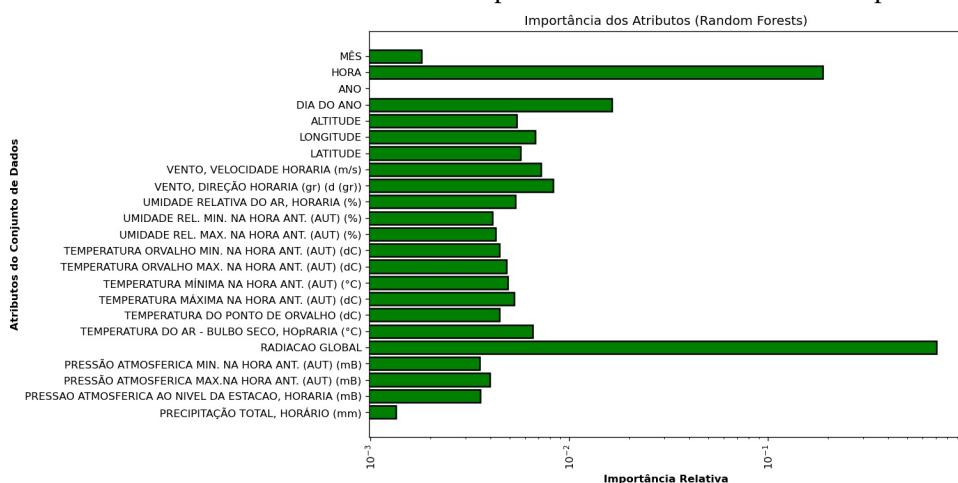
3.2.2 Processamento das Unidades de Tempo

As unidades de tempo foram processadas a partir dos atributos **DATA** e **HORA** por meio da criação de um *timestamp* temporário que foi utilizado para a geração dos atributos **DIA DO ANO**, contado a partir do dia 1º de Janeiro (**DIA DO ANO 1**); **HORA DO DIA**, contando a partir da hora 00:00 (**HORA DO DIA 0**); **MÊS** e **ANO** correspondentes ao mês e ano da observação, respectivamente.

3.2.3 Feature Selection e Feature Extraction

Feature Selection e Feature Extraction (ABE, 2010) consistem na seleção e na combinação dos atributos nos dados, melhorando a eficiência computacional e removendo informações irrelevantes e ruídos eventualmente prejudiciais à capacidade de generalização dos modelos. Para determinar a importância de cada atributo em relação à variável alvo **RADIAÇÃO PRÓXIMA HORA** foi utilizado o algoritmo *Random Forests* (BREIMAN, 2001) (cf. Figura 5). Com o resultado da aplicação

Figura 5 – Importância dos atributos do conjunto de dados estimada pelo algoritmo *Random Forests* no intervalo [0,1[, onde valores próximos a 1 indicam maior importância relativa.



Fonte: Elaborada pelo autor.

deste algoritmo, foi possível remover os atributos **ANO** e **MÊS**. Em seguida foram considerados as

variáveis **PRESSÃO ATMOSFÉRICA**, **UMIDADE RELATIVA DO AR**, **TEMPERATURA DO AR** e **PONTO DE ORVALHO**. Para cada variável foram registradas três medidas independentes que levaram à conclusão de que seria possível utilizar técnicas de Feature Extraction. Foi aplicado o algoritmo PCA (JOLLIFFE; CADIMA, 2016) e os atributos **PRESSÃO ATMOSFÉRICA**, **TEMPERATURA DO AR** e **PONTO DE ORVALHO** foram reduzidos de 3 dimensões para apenas uma; para o atributo **UMIDADE RELATIVA** a redução foi para 2 dimensões.

Na sequência foi aplicado o algoritmo *Sequential Forward Floating Selection* (PUDIL; NOVO-VIOVÁ; KITTNER, 1994) que realiza uma busca pelos subconjuntos de atributos tomando como referência uma métrica de desempenho para determinar o melhor subconjunto de atributos para o treinamento dos modelos. Diante da possibilidade da escolha, foi priorizado o subconjunto responsável pelo menor MSE, com 11 atributos: **RADIAÇÃO GLOBAL**, **VENTO – DIREÇÃO HORÁRIA**, **VENTO – VELOCIDADE HORÁRIA**, **LATITUDE**, **DIA DO ANO**, **HORA**, **PRESSAO ATMOSFERICA**, **UMIDADE_1**, **UMIDADE_2**¹⁰, **TEMPERATURA** e (Temperatura do) **PONTO DE ORVALHO**.

3.2.4 Tratamento dos Valores Faltantes

As observações são realizadas em estações meteorológicas automáticas sujeitas a diversos fatores que podem comprometer o correto funcionamento dos sensores, resultando em observações corrompidas. No conjunto de dados foram observados valores faltantes que puderam ser categorizados como MNAR (*Missing Not at Random*). Nestas circunstâncias a estratégia adotada para contornar o problema foi a Análise de Caso Completo, que consiste na remoção das observações que contenham ao menos um valor faltante em um de seus atributos, sem imputação.

3.2.5 Ajuste do Atributo Alvo

O atributo alvo **RADIAÇÃO PRÓXIMA HORA** (kJ/m^2) foi criado por deslocamento a partir de **RADIAÇÃO GLOBAL**. A Tabela 2, exemplifica a manipulação feita. O conjunto de dados foi agru-

Tabela 2 – Exemplo da manipulação para o ajuste do atributo alvo.

...	ID ESTAÇÃO	HORA	RADIAÇÃO GLOBAL	RADIAÇÃO PRÓXIMA HORA	...
...	A001	10:00	900	1.100	...
...	A001	11:00	1.100	1.300	...
...	A001	12:00	1.300	1.500	...
⋮					

Fonte: Elaborada pelo autor.

pado por **ID ESTAÇÃO** e ordenado por data e hora para evitar a introdução de inconsistências.

¹⁰ No processo de redução de dimensionalidade deste atributo houve a redução de 3 dimensões para 2 dimensões, justificando o segundo atributo.

3.2.6 Tratamento de *Outliers* e Padronização

Além da exclusão de valores absurdos como pressão atmosférica e umidade relativa do ar negativas ou o registro de temperaturas inferiores a -10°C , foi utilizado o algoritmo *Robust Scaler* (DEVELOPERS, 2020) que não sofre grande influência de valores discrepantes, permitindo que a escala dos atributos seja centrada em valores próximos.

3.3 Seleção dos Algoritmos para Implementação dos Modelos

3.3.1 Support Vector Machine

As Máquinas de Vetores de Suporte (SUYKENS; VANDEWALLE, 1999) são modelos que tentam encontrar um hiperplano ótimo que proporcione a melhor separação entre os dados para classificação, ou que melhor se ajuste a eles, para regressão. Para tanto o algoritmo realiza transformações nos dados que ocorrem por meio de funções de kernel, que calculam a relação entre cada um dos pares da amostra de dados.

3.3.2 Random Forests e Extra-Trees

Random Forests (BREIMAN, 2001) é um algoritmo baseado em composição que se utiliza de estimadores implementados por Árvores de Decisão, a partir de diferentes subconjuntos de atributos e amostras escolhidos dentro dos dados de entrada, permitindo um melhor desempenho quando comparada com as Árvores de Decisão isoladas. Adicionalmente, é possível obter mais aleatoriedade nos estimadores selecionando aleatoriamente os separadores para os nós das Árvores de Decisão. Neste o caso, o modelo é denominado *Extremely Randomized Trees* ou *Extra-Trees*. Ambos os modelos são utilizados devido à robustez que apresentam quanto ao *overfitting* e à facilidade de uso.

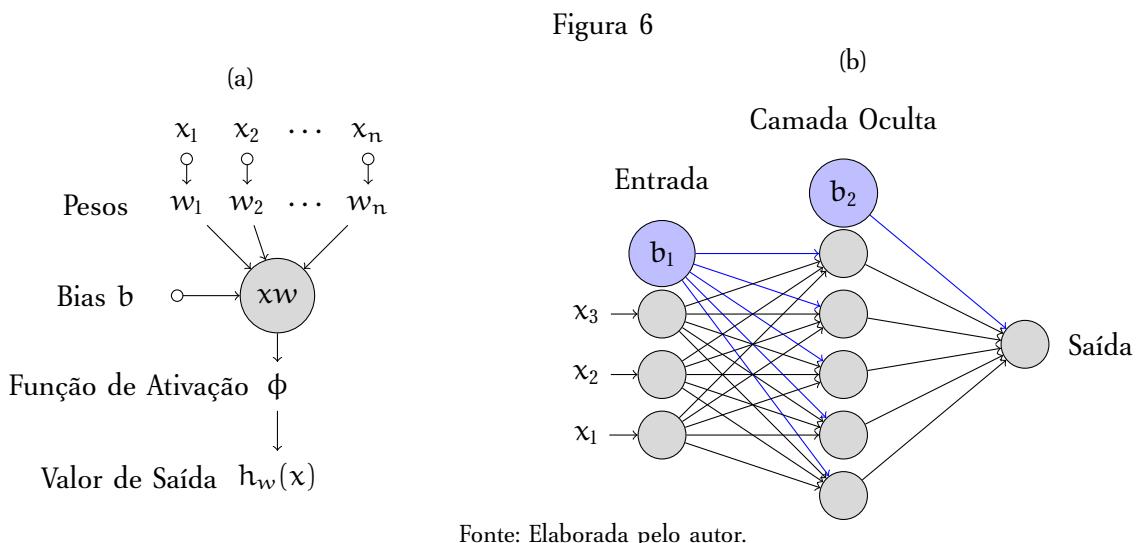
3.3.3 Extreme Gradient Boosting

O algoritmo *Extreme Gradient Boosting* (KEANY, 2020; CHEN; GUESTRIN, 2016) é baseado em composição que se utiliza de estimadores implementados por Árvores de Decisão, de maneira similar ao *Random Forests*. Na composição os estimadores são adicionados iterativamente de modo que cada novo estimador é responsável por se ajustar aos dados cuja generalização não foi realizada de maneira adequada pelos anteriores. Neste contexto, o algoritmo se utiliza de regularização e do Método de Newton (GALÁNTAI, 2000) para obter a convergência dos modelos da composição, além de apresentar características como a tolerância a valores faltantes e a separação dos nós das Árvores de Decisão baseada nos percentis dos atributos, conferindo robustez ao modelo.

3.3.4 Redes Neurais Artificiais

Redes Neurais Multicamada

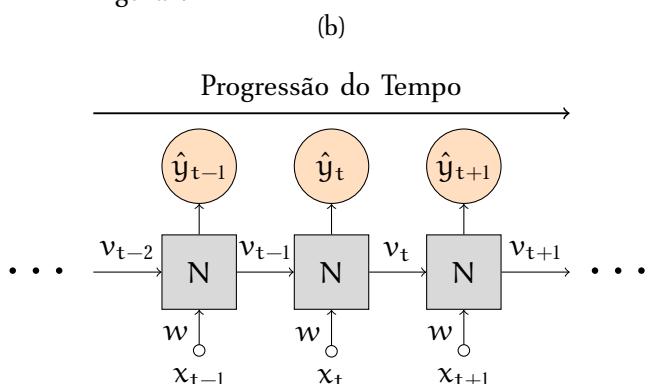
Redes Neurais Artificiais (HAYKIN, 1994), são modelos do comportamento de neurônios biológicos, sendo amplamente utilizadas em aprendizado supervisionado (ADOUNI et al., 2013; PIRDASHTI et al., 2013). Sua organização geral consiste da utilização de unidades de processamento denominadas neurônios (Figura 6a) dispostas em camadas, de modo que cada neurônio da camada n está conectado a todos os neurônios da camada $n + 1$ (Figura 6b).



Redes Neurais Recorrentes

Redes Neurais Artificiais Recorrentes são variações das Redes Neurais Multicamada nas quais cada neurônio utiliza como entrada não apenas os valores vindos do conjunto de dados, mas também o seu valor de saída na iteração anterior. A Figura 7a apresenta a configuração de um neurônio enquanto a Figura 7b ilustra o fluxo das informações em um neurônio ao longo do tempo.

Figura 7



Em virtude destas características, estes modelos são amplamente utilizados em processamento de séries temporais. Em alternativa aos neurônios convencionais presentes nas Redes Neurais

Multicamada, foram desenvolvidas unidades como as LSTMs (*Long Short-Term Memory*) e GRUs (*Gated Recurrent Units*) que otimizam a captura de padrões em séries temporais, tornando-as versáteis para tarefas como reconhecimento de fala (GRAVES; MOHAMED; HINTON, 2013) e previsão de preços de ações no Mercado Financeiro (RAHMAN et al., 2019).

3.3.5 Ensemble Methods

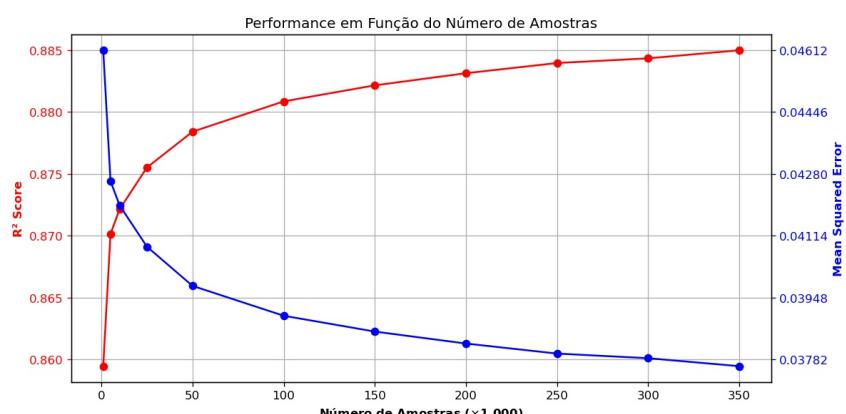
Modelos baseados em composição (DIETTERICH, 2000) são modelos compostos por diversos estimadores que empregam alguma forma de agregação para utilizar os valores propostos por seus estimadores a fim de determinar o valor proposto pela composição. Sua utilização tenta agregar os modelos já treinados para obter uma composição que tenha melhor desempenho quando comparado aos seus estimadores isoladamente, aproveitando a diversificação que essa prática confere aos modelos.

3.4 Execução dos Algoritmos de Treinamento

A execução dos algoritmos deu-se em um cluster com 7 nós configurado utilizando o framework Dask¹¹, ou em serviços como o Google Collaboratory¹² quando apropriado. A implementação foi feita na linguagem Python (versão 3.7) com as bibliotecas **Scikit-Learn**, que implementa algoritmos de treinamento, avaliação de desempenho, *feature selection*, etc; **MLXtend**, que estende funcionalidades do Scikit-Learn; **Tensorflow**, que implementa algoritmos de aprendizagem profunda; e **XGBoost**, que implementa o algoritmo *Extreme Gradient Boosting*.

Foi considerado inicialmente o treinamento de apenas um modelo a partir dos algoritmos selecionados com todas as 200 000 observações disponíveis. No entanto, visando melhorar a eficiência computacional, foi observado o desempenho dos modelos em função do número de amostras no treinamento, como ilustrado na Figura 8. Concluiu-se que utilizar mais do que 350 000 observações não gera ganhos de desempenho sensíveis.

Figura 8 – Desempenho dos modelos com relação às métricas MSE e R² com relação ao número de amostras.



Fonte: Elaborada pelo autor.

¹¹ Disponível em: <https://dask.org/>.

¹² Disponível em: <https://colab.research.google.com/notebooks/intro.ipynb>

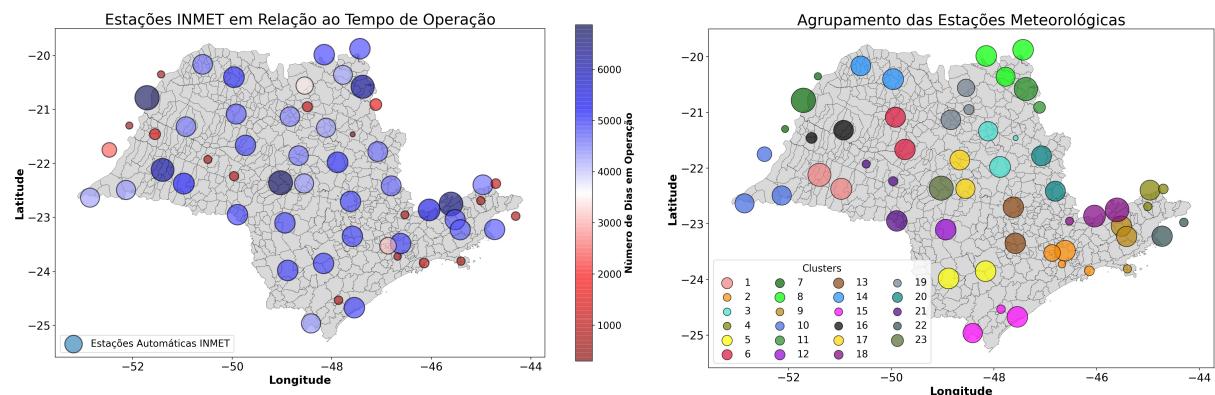
Em seguida foram consideradas as seguintes estratégias:

1. Implementar um modelo utilizando os dados agrupados por ano¹³ adicionando-os iterativamente ao conjunto de treinamento.
2. Implementar diversos modelos, cada um correspondendo a um ponto de coleta de dados e agregá-los em um modelo baseado em composição.
3. Análogo ao anterior, mas utilizando os dados agrupados em clusters de pontos de coleta de dados para o treinamento.

A Estratégia 1 usou a forma mais organizada dos dados, no entanto, já na adição do segundo ano de dados foi ultrapassada a marca de 350 000 observações, inviabilizando a adoção desta estratégia. A Estratégia 2 mostrou-se vantajosa por melhorar a eficiência computacional, entretanto, o treinamento nos pontos de coleta de dados com uma quantidade pequena de observações (*cf.* Figura 9a) poderia prejudicar a capacidade de generalização. Para a Estratégia 3 foram determinados os agrupamentos a fim de que um mesmo cluster não contivesse apenas estações com poucas observações de modo a balancear sua quantidade. Para o agrupamento foi utilizado o algoritmo *K-Means* (ARTHUR; VASSILVITSKII, 2006), utilizando as coordenadas geográficas dos pontos de coleta de dados. O algoritmo foi executado iterativamente incrementando o número de clusters até que seu número ótimo fosse determinado (*cf.* Figura 9b), resultando em **23 clusters**.

Figura 9

(a) Estações meteorológicas em relação ao tempo de operação. (b) Estabelecimento dos grupos de pontos de coleta de dados para o treinamento dos modelos.

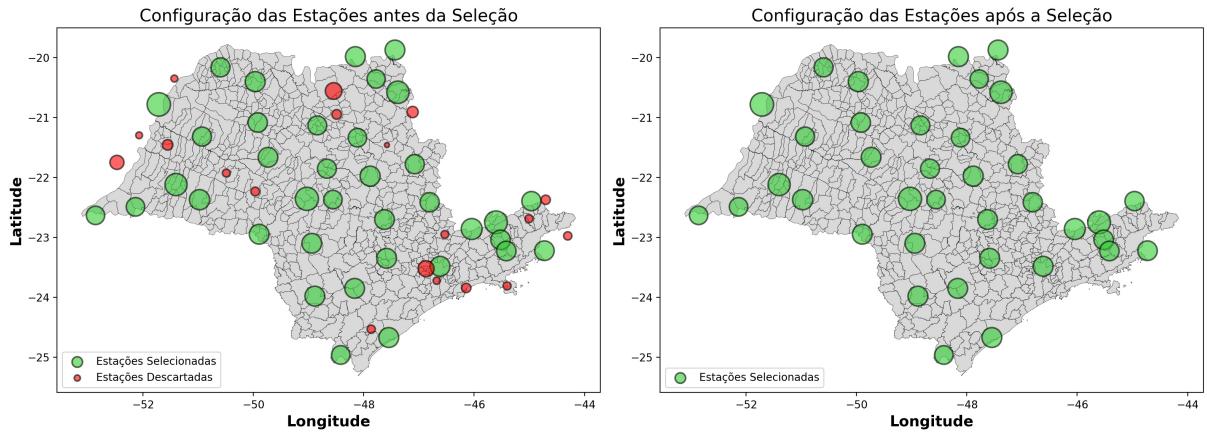


Fonte: Elaborada pelo autor.

Também foi adotada uma variação da Estratégia 2 que consistiu na remoção de 19 estações com uma quantidade de dias de operação inferior a 3500 (*cf.* Figura 10). Assim foi possível utilizar os dados das estações removidas para o teste dos modelos juntamente com uma fração das amostras em cada um das estações selecionadas para treinamento.

¹³ À exceção do ano de 2020, retiradas desse agrupamento por não constituírem um ainda um ano completo.

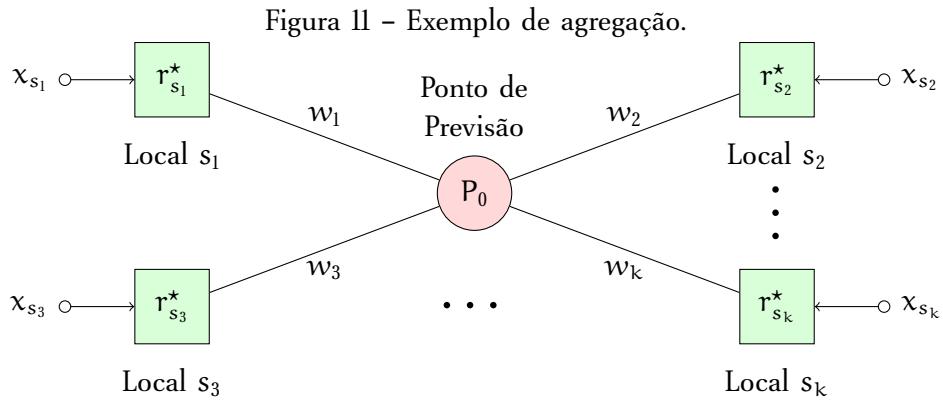
Figura 10 – Resultado da seleção das estações para treinamento na variação da estratégia 2.



Fonte: Elaborada pelo autor.

3.5 Design do Modelo de Agregação

O modelo proposto consiste em uma agregação dos demais modelos treinados nas localidades específicas para realizar a interpolação de cada uma das previsões ao longo do espaço. A técnica de interpolação proposta se utiliza da distância e das métricas de desempenho para determinar os pesos de cada um dos pontos de referência a fim de calcular $\psi(P_0)$. A Figura 11 ilustra o modelo de agregação proposto.



Neste contexto, para cada localização $s_i \in L$, onde L é o conjunto com m localidades, há um conjunto $R_{s_i} = \{r_{s_i}^1, r_{s_i}^2, \dots, r_{s_i}^n\}$ que contém n modelos implementados a partir dos diferentes algoritmos e conjuntos de hiperparâmetros utilizando o conjunto de dados relativo a s . Dentre as m localidades, são selecionados os k locais mais próximos de P_0 e em seguida é determinado, em cada um dos k locais, o modelo com melhor desempenho $r_{s_i}^*$, tal que $r_{s_i}^* \in R_{s_i}$, a ser utilizado na agregação.

No cálculo de $\psi(P_0)$ são utilizadas as coordenadas geográficas de P_0 e os valores de previsão $r_s^*(x_s)$ propostos pelo modelo selecionado em cada um dos k locais, de modo que $r_s^*(x_s) = \hat{y}_s(x_s)$, onde x_s é o conjunto de dados de entrada referente a cada local s . Esses valores propostos são então agregados conforme as Equações (1) e (2). Nestas circunstâncias não há necessidade de se utilizar a variável **LONGITUDE** no treinamento dos modelos, apenas na agregação.

3.5.1 Seleção dos Modelos

A escolha do melhor modelo $r_s^* \in R_s$ consiste em dois processos de seleção: **1:** seleção entre os modelos implementados a partir do mesmo algoritmo; e **2:** seleção dos melhor modelo dentre os pré-selecionados. Em ambos os processos os critérios adotados foram os valores do RMSE, MAE e R^2 Score, respectivamente.

3.5.2 Estabelecimento do Conjunto de Pesos

Para determinar o peso de cada modelo são considerados: **1:** proximidade do ponto de previsão P_0 com o modelo selecionado — quanto menor a distância entre s e P_0 maior é sua influência de r_s^* em $\psi(P_0)$; e **2:** desempenho do modelo selecionado — quanto melhor o desempenho do modelo r_s^* , maior sua influência em r_s^* em $\psi(P_0)$.

O peso w_i para o melhor modelo do i -ésimo local dentre os k escolhidos é dado pela Equação (1), onde d corresponde à distância geográfica entre P_0 e s , p é um parâmetro de suavização e RMSE e MAE são as métricas de desempenho do modelo $r_{s_i}^*$ (*cf.* Apêndice B)

$$w_i(s_i, P_0) = \frac{1}{d(s_i, P_0)^p} \left(\frac{\text{RMSE}(r_{s_i}^*) + \text{MAE}(r_{s_i}^*)}{2} \right)^{-1}. \quad (1)$$

3.5.3 Possibilidades para Agregação

O modelo de agregação é descrito pela Equação (2). Nele é possível utilizar diferentes valores para a quantidade de locais k e para a suavização p , a fim de obter a melhor generalização para as métricas de desempenho adotadas

$$\psi(P_0) = \frac{\sum_{i=1}^k [r_{s_i}^*(x_{s_i}) w_i(s_i, P_0)]}{\sum_{i=1}^k w_i(s_i, P_0)}. \quad (2)$$

A estratégia para a escolha dos parâmetros k e p consistiu de uma busca parametrizada análoga ao *Grid-Search* para determinar a melhor combinação (k, p) (*cf.* Tabela 3).

Tabela 3 – Valores para busca parametrizada no modelo de agregação.

Parâmetro	Intervalo
k	$\{2, 3, 4, \dots, 12\}$
p	$\{\frac{1}{4}, \frac{1}{2}, 2, 4\}$

Fonte: Elaborada pelo autor.

4 Plano de Atividades para o Próximo Período

Conforme o cronograma apresentado na Tabela 1, as próximas etapas a serem realizadas para o andamento do projeto são:

Avaliação do desempenho dos modelos no conjunto de dados Com os dados pré-processados e prontos para serem utilizados nos algoritmos de treinamento, e com a infraestrutura operacional devidamente configurada, haverá o término do treinamento e o devido registro das métricas de desempenho para os respectivos modelos. Considerando a proposta de utilização de um modelo de previsão baseado da agregação de vários estimadores haverá o tabelamento das informações relativas ao desempenho para cada um dos modelos presentes na agregação;

Ajuste dos modelos com base nas informações do item anterior Uma vez obtido o conjunto com todos os estimadores treinados a partir dos hiperparâmetros padrão, com as respectivas métricas de desempenho, será realizada a busca pelos melhores hiperparâmetros de cada um dos modelos presentes na composição, utilizando-se de técnicas como *Grid-Search* ou *Randomized Search*, quando pertinente. Em um segundo momento, também serão considerados na busca os melhores parâmetros para o modelo baseado na agregação dos estimadores, *i.e.* os parâmetros k e p conforme a Seção 3.5;

Teste e avaliação do desempenho dos modelos nos locais específicos durante 1 mês Após a otimização dos modelos por hiperparametrização, considerando o modelo baseado em composição pronto para ser utilizado em um cenário real de previsão, haverá o teste e avaliação do desempenho do modelo utilizando novos dados. Nestas circunstâncias serão obtidos novos dados da mesma fonte, a fim de que o modelo seja testado com dados originais, não presentes nos conjuntos de treinamento e validação.

Em virtude da estratégia de utilização dos dados adotada no treinamento dos modelos, será possível averiguar o desempenho dos modelos, tanto em locais utilizados para o treinamento, quanto para os locais que foram retirados do conjunto de treinamento pelo mesmo critério, mas continuam a registrar observações. Nos locais selecionados há a possibilidade de avaliar o desempenho dos modelos selecionados para a agregação. Nos locais não utilizados no treinamento será possível avaliar o modelo de agregação em si, em localidades distantes dos pontos de observação.

Adicionalmente há a etapa de elaboração dos relatórios que contemplará o registro do desenvolvimento da segunda etapa do projeto.

Referências

- ABE, Shigeo. **Feature selection and extraction**. Springer, 2010. P. 331–341.
- ACHILLES, R. **Energia Solar Paulista: Levantamento do Potencial**. 2013. P. 8.
- ADOUNI, Amel et al. Sensor and actuator fault detection and isolation based on artificial neural networks and fuzzy logic applied on induction motor. In: IEEE. 2013 International Conference on Control, Decision and Information Technologies (CoDIT). 2013. P. 917–922.
- ALZAHRANI, Ahmad et al. Solar Irradiance Forecasting Using Deep Neural Networks. **Procedia Computer Science**, Elsevier, v. 114, p. 304–313, 2017.
- ARTHUR, David; VASSILVITSKII, Sergei. **k-means++: The advantages of careful seeding**. 2006.
- ASSOULINE, Dan; MOHAJERI, Nahid; SCARTEZZINI, Jean-Louis. Quantifying Rooftop Photovoltaic Solar Energy Potential: A Machine Learning Approach. **Solar Energy**, Elsevier, v. 141, p. 278–296, 2017.
- BREIMAN, Leo. Random forests. **Machine learning**, Springer, v. 45, n. 1, 2001.
- CHEN, Tianqi; GUESTRIN, Carlos. Xgboost: A scalable tree boosting system. In: PROCEEDINGS of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. P. 785–794.
- DEVELOPERS, Scikit Learn. **Compare the Effect of Different Scalers on Data with Outliers**. 2020. Disponível em:
https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html. Acesso em: 2 fev. 2021.
- DIETTERICH, Thomas G. Ensemble methods in machine learning. In: SPRINGER. INTERNATIONAL workshop on multiple classifier systems. 2000. P. 1–15.
- EPE, EDPE. Projeção de Demanda de Energia Elétrica-2017-2026. **Ministério de Minas e Energia. Rio de Janeiro**, p. 95, 2017.
- GALÁNTAI, Aurel. The theory of Newton's method. **Journal of Computational and Applied Mathematics**, Elsevier, v. 124, n. 1-2, p. 25–44, 2000.
- GRAVES, A.; MOHAMED, A.; HINTON, G. Speech recognition with deep recurrent neural networks. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. 2013. P. 6645–6649. DOI: 10.1109/ICASSP.2013.6638947.
- HAYKIN, Simon. **Neural Networks: a Comprehensive Foundation**. Prentice Hall PTR, 1994.
- JOLLIFFE, Ian T; CADIMA, Jorge. Principal component analysis: a review and recent developments. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, The Royal Society Publishing, v. 374, n. 2065, p. 20150202, 2016.

- KEANY, Eoghan. **What makes XGBoost so Extreme?** 2020. Disponível em: <<https://medium.com/Analytics-vidhya/what-makes-xgboost-so-extreme-e1544a4433bb>>. Acesso em: 10 jan. 2021.
- LI, Jiaming et al. Machine Learning for Solar Irradiance Forecasting of Photovoltaic System. **Renewable energy**, Elsevier, v. 90, p. 542–553, 2016.
- LOU, Siwei et al. Prediction of Diffuse Solar Irradiance Using Machine Learning and Multivariable Regression. **Applied energy**, Elsevier, v. 181, p. 367–374, 2016.
- PERRAULT, Raymond et al. The AI Index 2019 Annual Report. **AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA**, 2019.
- PIRDASHTI, Mohsen et al. Artificial neural networks: applications in chemical engineering. **Reviews in Chemical Engineering**, De Gruyter, v. 29, n. 4, p. 205–239, 2013.
- PUDIL, Pavel; NOVOVIOVÁ, Jana; KITTLER, Josef. Floating Search Methods in Feature Selection. **Pattern Recognition Letters**, Elsevier, v. 15, n. 11, p. 1119–1125, 1994.
- RAHMAN, Mohammad Obaidur et al. Predicting Prices of Stock Market using Gated Recurrent Units (GRUs) Neural Networks. **INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND NETWORK SECURITY, INT JOURNAL COMPUTER SCIENCE & NETWORK SECURITY-IJCSNS DAE-SANG OFFICE 301**, v. 19, n. 1, p. 213–222, 2019.
- SMETS, Arno HM et al. **Solar Energy: The physics and engineering of photovoltaic conversion, technologies and systems**. UIT Cambridge, 2015.
- SUYKENS, Johan AK; VANDEWALLE, Joos. Least Squares Support Vector Machine Classifiers. **Neural processing letters**, Springer, v. 9, n. 3, p. 293–300, 1999.
- ZENG, Jianwu; QIAO, Wei. Short-term Solar Power Prediction Using a Support Vector Machine. **Renewable Energy**, Elsevier, v. 52, p. 118–127, 2013.

APÊNDICE A - Detalhamento das Estações Meteorológicas

A Tabela 4 apresenta informações sobre as estações meteorológicas automáticas do Instituto Nacional de Meteorologia, das quais foram coletados os dados utilizados neste projeto.

Tabela 4 – Discriminação dos pontos de coletas de dados. Os números das estações correspondem aos do mapa da Figura 4.

Nº	Estação	ID INMET	Latitude	Longitude	Altitude (m)	Início das Observações	Município
1	A509	-22.8614	-46.0433	1500	19/12/2004	Monte Verde	
2	A520	-19.9858	-48.1514	568	18/07/2006	Conceição Das Alagoas	
3	A525	-19.8753	-47.4341	912	19/08/2006	Sacramento	
4	A529	-22.3958	-44.9617	1017	30/05/2007	Passa Quatro	
5	A561	-20.9099	-47.1142	845	17/08/2015	São Sebastião Do Paraíso	
6	A619	-23.2233	-44.7267	3	19/11/2006	Parati	
7	A628	-22.9757	-44.3033	6	25/08/2017	Angra Dos Reis	
8	A635	-22.3739	-44.7031	2450	01/09/2017	Itatiaia	
9	A701	-23.4833	-46.6167	785	25/07/2006	São Paulo - Mirante	
10	A704	-20.7833	-51.7122	313	03/09/2001	Três Lagoas	
11	A705	-22.3581	-49.0289	636	30/08/2001	Bauru	
12	A706	-22.7502	-45.6038	1642	13/03/2002	Campos Do Jordão	
13	A707	-22.1199	-51.4000	431	04/02/2003	Presidente Prudente	
14	A708	-20.5800	-47.3800	1002	12/12/2002	Franca	
15	A711	-21.9797	-47.8833	859	04/09/2006	São Carlos	
16	A712	-24.6717	-47.5459	2	20/07/2006	Iguape	
17	A713	-23.3500	-47.5856	609	22/08/2006	Sorocaba	
18	A714	-23.9814	-48.8853	743	25/07/2006	Itapeva	
19	A715	-23.8514	-48.1644	675	15/08/2006	São Miguel Arcanjo	
20	A716	-22.9486	-49.8942	443	29/08/2006	Ourinhos	
21	A718	-22.3725	-50.9742	398	01/09/2006	Rancharia	
22	A725	-23.0997	-48.9411	775	22/09/2006	Avaré	
23	A726	-22.7028	-47.6231	566	26/09/2006	Piracicaba	
24	A727	-21.6653	-49.7342	450	20/09/2006	Lins	
25	A728	-23.0417	-45.5203	571	20/12/2006	Taubaté	
26	A729	-20.4031	-49.9658	465	04/12/2006	Votuporanga	
27	A733	-20.1650	-50.5950	457	22/08/2007	Jales	
28	A734	-21.3191	-50.9302	374	30/08/2007	Valparaiso	
29	A735	-21.0856	-49.9204	405	03/09/2007	José Bonifácio	
30	A736	-21.1331	-48.8403	525	13/11/2007	Ariranha	
31	A737	-21.8556	-48.6667	492	09/11/2007	Ibitinga	
32	A738	-21.7797	-47.0753	730	25/06/2007	Casa Branca	
33	A739	-22.4150	-46.8053	633	05/11/2007	Itapira	
34	A740	-23.2283	-45.4169	730	01/11/2007	Sao Luis Do Paraitinga	
35	A741	-22.3708	-48.5572	533	24/04/2008	Barra Bonita	
36	A744	-22.9519	-46.5305	891	20/12/2017	Bragançaa Paulista	
37	A746	-24.9628	-48.4164	659	03/07/2008	Barra Do Turvo	
38	A747	-21.3383	-48.1139	540	22/04/2008	Pradópolis	
39	A748	-20.5589	-48.5447	533	19/06/2010	Barretos	
40	A753	-20.3594	-47.7750	600	17/07/2008	Ituverava	
41	A755	-23.5233	-46.8692	776	29/03/2011	Barueri	
42	A759	-21.7501	-52.4706	387	21/03/2013	Bataguassu	
43	A762	-21.4577	-51.5522	383	03/11/2016	Dracena	
44	A763	-22.2352	-49.9650	660	15/05/2017	Marília	
45	A764	-20.9492	-48.4897	590	27/10/2016	Bebedouro	
46	A765	-23.8447	-46.1431	5	01/02/2017	Bertioga	
47	A766	-24.5331	-47.8641	35	09/02/2017	Registro	
48	A767	-23.8107	-45.4025	24	25/10/2017	São Sebastião	
49	A768	-21.9272	-50.4903	498	12/05/2017	Tupã	
50	A769	-22.6889	-45.0054	586	20/10/2017	Cachoeira Paulista	
51	A770	-21.4611	-47.5794	620	03/07/2019	São Simão	
52	A771	-23.7245	-46.6775	771	14/03/2018	São Paulo - Interlagos	
53	A849	-22.6339	-52.8589	362	06/03/2008	Diamante Do Norte	
54	A850	-22.4917	-52.1344	308	04/03/2008	Paranapoema	
55	S705	-21.2983	-52.0689	345	06/04/2018	Brasilândia	
56	S717	-20.3514	-51.4302	374	05/04/2018	Selvíria	

APÊNDICE B – Métricas de Desempenho

A seguir são apresentadas métricas de desempenho empregadas para avaliar os modelos de previsão. Em cada uma das métricas apresentadas, n indica o número total de amostras sobre o qual a métrica é calculada, y_i indica o valor da i -ésima observação e \hat{y}_i é o valor da i -ésima previsão realizada pelo modelo.

- R^2 Score – Correlação entre as previsões e as observações:

$$R^2 = 1 - \frac{\sum_{i=0}^n (y_i - \hat{y}_i)^2}{\sum_{i=0}^n (y_i - \bar{y})^2} \quad \text{onde} \quad \bar{y} = \frac{1}{n} \sum_{i=0}^n y_i.$$

- RMSE – *Root Mean Squared Error* – Raiz do Erro Quadrático Médio:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=0}^n (y_i - \hat{y}_i)^2}.$$

- MAE – *Mean Absolute Error* – Erro Absoluto Médio:

$$\text{MAE} = \frac{1}{n} \sum_{i=0}^n |y_i - \hat{y}_i|.$$

- MAPE – *Mean Average Percentage Error* – Erro Percentual Médio:

$$\text{MAPE} = \frac{1}{n} \sum_{i=0}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100.$$

As métricas de desempenho adotadas serão aplicadas tanto nos modelos que compõe a agregação quanto para a agregação em si.