



MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

ASSIMILAÇÃO DE DADOS POR REDES NEURAIIS NO MODELO WRF-NCAR

Gerônimo Gallarreta Zubiaurre Lemos

Relatório de atividades Iniciação Científica, orientada pelos Drs. Haroldo Fraga de Campos Velho e Juliana Aparecida Anochi.

URL do documento original:

<<http://urlib.net/xx/yy>>

INPE
São José dos Campos
2023

PUBLICADO POR:

Instituto Nacional de Pesquisas Espaciais - INPE
Coordenação de Ensino, Pesquisa e Extensão (COEPE)
Divisão de Biblioteca (DIBIB)
CEP 12.227-010
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/7348
E-mail: pubtc@inpe.br

CONSELHO DE EDITORAÇÃO E PRESERVAÇÃO DA PRODUÇÃO INTELLECTUAL DO INPE - CEPPII (PORTARIA Nº 176/2018/SEI-INPE):

Presidente:

Dra. Marley Cavalcante de Lima Moscati - Coordenação-Geral de Ciências da Terra (CGCT)

Membros:

Dra. Ieda Del Arco Sanches - Conselho de Pós-Graduação (CPG)
Dr. Evandro Marconi Rocco - Coordenação-Geral de Engenharia, Tecnologia e Ciência Espaciais (CGCE)
Dr. Rafael Duarte Coelho dos Santos - Coordenação-Geral de Infraestrutura e Pesquisas Aplicadas (CGIP)
Simone Angélica Del Ducca Barbedo - Divisão de Biblioteca (DIBIB)

BIBLIOTECA DIGITAL:

Dr. Gerald Jean Francis Banon
Clayton Martins Pereira - Divisão de Biblioteca (DIBIB)

REVISÃO E NORMALIZAÇÃO DOCUMENTÁRIA:

Simone Angélica Del Ducca Barbedo - Divisão de Biblioteca (DIBIB)
André Luis Dias Fernandes - Divisão de Biblioteca (DIBIB)

EDITORAÇÃO ELETRÔNICA:

Ivone Martins - Divisão de Biblioteca (DIBIB)
André Luis Dias Fernandes - Divisão de Biblioteca (DIBIB)



MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

ASSIMILAÇÃO DE DADOS POR REDES NEURAIIS NO MODELO WRF-NCAR

Gerônimo Gallarreta Zubiaurre Lemos

Relatório de atividades Iniciação Científica, orientada pelos Drs. Haroldo Fraga de Campos Velho e Juliana Aparecida Anochi.

URL do documento original:

[<http://urlib.net/xx/yy>](http://urlib.net/xx/yy)

INPE
São José dos Campos
2023

RESUMO

O presente trabalho propõe o uso de método de aprendizado de máquina para o problema de assimilação de dados (DA: Data Assimilation) aplicado ao modelo de meso-escala WRF. Para o desenvolvimento do projeto, foi estruturado um banco de dados contendo análises do sistema 3D-Var, previsões do modelo WRF (Weather Research and Forecasting System) e um conjunto de observações. Para gerar essas análises, foram obtidos dados de previsão do modelo GFS-NOAA (Global Forecast System) e dados de observações para execução do módulo de assimilação do modelo WRF (WRF-DA). Os dados para o experimento numérico foram coletados dos meses de Junho, Julho e Agosto dos anos de 2015, 2016, 2017 e 2018. O domínio estudado abrange a região sul do Brasil, Rio Grande do Sul, Santa Catarina e Paraná, além do Uruguai e parte da Argentina e sul do Paraguai. O método de aprendizado de máquina é chamado de árvores de decisão e usou-se a codificação da biblioteca XGBoost, que implementa uma versão otimizada do algoritmo Gradient Boosting, um algoritmo supervisionado. Para configurar a biblioteca XGBoost para a aplicação, o conjunto de dados foi subdividido em dois conjuntos distintos: conjunto de treinamento e conjunto de teste. O conjunto de treinamento é constituído pelos dados dos anos de 2015, 2016 e 2017, enquanto o conjunto de teste possui os dados do ano de 2018. A técnica de K-fold Cross-Validation foi utilizada para o treinamento do modelo, com $k = 3$, sendo os conjuntos de validação para cada iteração um dos anos do conjunto de treinamento. Ou seja, o modelo XGBoost-DA foi configurado por um procedimento em etapas: inicialmente, treinado com os anos de 2015 e 2016 e avaliado no ano de 2017; após, treinado com os anos de 2015-2017 e avaliado no ano de 2017 e, por fim, o último modelo foi treinado nos anos de 2016 e 2017 e avaliado no ano de 2015. Ao fim desse processo, foi escolhido o modelo com menor erro sobre o conjunto de validação. Para avaliação final do modelo, foi executado um exemplo de previsão de 24 horas no ano de 2018, com análises geradas pelo módulo de 3D-Var (nativo no WRF) e análise geradas pelo XGBoost-DA. Previsões geradas com as duas análises se mostraram muito similares, evidenciando a aplicação do XGBoost como uma ferramenta promissora para assimilação de dados.

Palavras-chave: Inteligência Artificial. Aprendizado de Máquina. Assimilação de Dados. Previsão Numérica do Tempo. WRF.

LISTA DE FIGURAS

	<u>Pág.</u>
2.1 Representação gráfica do domínio estudado. Fonte: O Autor.	3
2.2 Distribuição dos dados de observações Fonte: O Autor.	5
3.1 Exemplo de utilização de Árvore de Decisão. Fonte: scikit-learn contri- butors (a)	6
3.2 Representação gráfica do algoritmo GB. Fonte: Amazon Web Services, Inc. (2023)	7
4.1 Representação gráfica de <i>K-Cross Validation</i> . Fonte: scikit-learn contri- butors (b).	9
4.2 Séries Temporais dos dados de referência e valores preditos. Fonte: O Autor.	10
4.3 Simulação gerada pelo modelo GFS. Fonte: O Autor	11
4.4 Análises geradas por métodos de DA. Fonte: O Autor	12
4.5 Previsões geradas a partir de análises geradas por meio de DA. Fonte: O Autor	13
4.6 Diferença absoluta entre as previsões geradas. Fonte: O Autor	13

SUMÁRIO

	<u>Pág.</u>
1 INTRODUÇÃO	2
2 DADOS ANALISADOS	3
2.1 O Domínio	3
2.2 Dados utilizados	4
3 Algoritmos de Aprendizado de Máquina	4
3.1 Aprendizado Supervisionado	5
3.2 Árvores de Decisão	5
3.3 <i>Gradient Boosting</i>	7
4 Metodologia de Treinamento	7
4.1 Divisão dos dados e <i>K-Cross Validation</i>	8
4.2 Análise dos resultados	9
4.3 Comparação de Previsões Numéricas e discussão dos resultados	11
5 Discussão dos Resultados	13
REFERÊNCIAS BIBLIOGRÁFICAS	15

1 INTRODUÇÃO

A técnica de Assimilação de Dados (DA: Data Assimilation), no contexto da meteorologia, consiste na utilização de dados observados para gerar uma melhor Previsão Numérica do Tempo (PNT). A utilização dessa técnica possibilita uma melhoria significativa nos resultados de previsões numéricas. Para que os dados de observações possam ser utilizados nos modelos matemáticos de previsão, é necessário que as observações passem por uma etapa de pré-processamento, a fim de garantir a integridade das previsões. Atualmente, técnicas baseadas em métodos variacionais, como o 3D-Var e 4D-Var constituem o estado da arte em DA.

Com o recente avanço de pesquisas na área de Aprendizado de Máquina (AM), bem como o aumento da disponibilidade de dados de observações, surge a possibilidade da utilização de técnicas de AM no problema de DA. Além disso, os bons resultados alcançados por essas técnicas em áreas análogas, como previsão climática sazonal (ANOCHI; VELHO, 2014) e previsão de precipitação (MONEGO et al., 2022) corroboram com a utilização de tais técnicas para o problema de DA.

Sendo assim, o presente relatório descreve o andamento de projeto de pesquisa, o qual utiliza AM supervisionado para o problema de DA de dados de temperatura. O restante deste documento está organizado da seguinte forma:

- Seção 2: Dados Analisados. Aqui, são apresentadas as características dos dados utilizados durante a pesquisa. São apresentadas informações sobre o domínio escolhido para a realização da experimentação e natureza dos dados;
- Seção 3: Descrição dos algoritmos de AM utilizados;
- Seção 4: Metodologia de treinamento e avaliação do modelo desenvolvido. Introdução do campo de temperatura gerado pelo modelo desenvolvido no meso-escala WRF e geração de uma previsão de vinte e quatro horas;
- Seção 5: Discussão dos resultados obtidos e trabalhos futuros.

2 DADOS ANALISADOS

Esta seção aborda as características do domínio escolhido para experimentação durante o desenvolvimento do projeto.

2.1 O Domínio

O domínio estudado na presente pesquisa tem coordenadas de latitude(-36.19862, -24.9365) e longitude (-59.6492, -47.7399). Essa região compreende a região sul do Brasil, contendo os estados do Rio Grande do Sul, Santa Catarina e Paraná, bem como os territórios de Uruguai, Argentina e Paraguai. A Figura 3.2 apresenta a representação gráfica do domínio de estudo.



Figura 2.1 - Representação gráfica do domínio estudado. Fonte: O Autor.

A região selecionado possui 17931 pontos, formando um *grid* de pontos de dimensão (129x139). A distância entre pontos de grade é de 9Km.

2.2 Dados utilizados

A variável de interesse nesta etapa de pesquisa é a Temperatura a 2 metros da superfície (T2). O conjunto de dados utilizado contém as seguintes variáveis:

- Latitude: Valor de Latitude;
- Longitude: Valor de Longitude;
- Ponto de Grade X: Linha da grade em que o ponto se encontra
- Ponto de Grade Y: Coluna da grade em que o ponto se encontra;
- Valor de Simulação: Valor de previsão do modelo GFS-NOAA;
- Valor de Observação: Dados de observações;
- *Target*: Análise gerada pelo sistema 3D-Var, tendo como entrada o valor de simulação e os dados observados. Este é o valor de referência para o treinamento do modelo de AM.

Na prática, apenas as variáveis de Valor de Simulação, Valor de Observação e *Target* são utilizadas no treinamento do modelo, sendo as outras importantes apenas para construção das visualizações de resultados.

Os dados obtidos para realização do experimento numérico compreendem os meses de Junho, Julho e Agosto, dos anos de 2015, 2016, 2017 e 2018. Para cada dia contido no conjunto de dados, há um total de 56 observações obtidas. A Figura 2.2 contém a distribuição espacial das observações no domínio de estudo.

3 Algoritmos de Aprendizado de Máquina

Aprendizado de Máquina (AM) é a subárea do campo de estudo da Inteligência Artificial (IA) que se preocupa em desenvolver modelos capazes de aprender padrões em um conjunto de dados. Dentre essa área de estudo, existem três subcategorias de AM: aprendizado supervisionado, aprendizado não-supervisionado e aprendizado por reforço. A utilização de cada uma das técnicas mencionadas depende do problema abordado e da característica dos dados disponíveis. Durante o desenvolvimento desta pesquisa, foram utilizados apenas algoritmos de aprendizado de máquina supervisionado. Portanto, será dado foco apenas neste tipo de técnica.

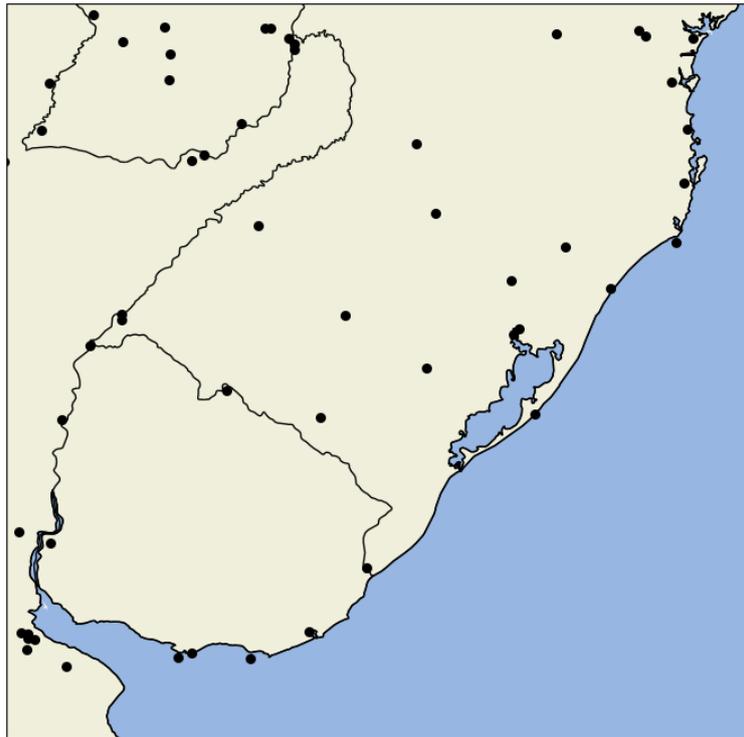


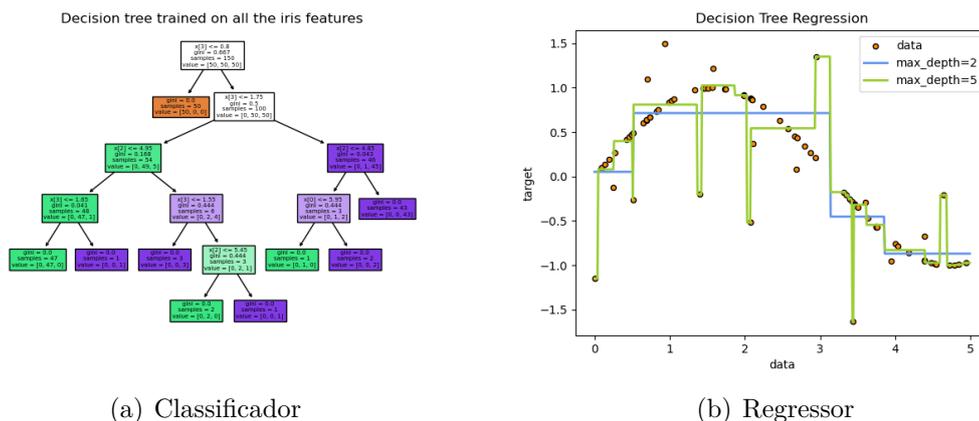
Figura 2.2 - Distribuição dos dados de observações Fonte: O Autor.

3.1 Aprendizado Supervisionado

Aprendizado Supervisionado (do inglês, *Supervised Learning*) é o tipo de aprendizado de máquina em que o modelo aprende a partir de um sinal supervisorio. Ou seja, durante o treinamento, para cada elemento do conjunto de treinamento, há uma informação sobre a saída desejada para aquela entrada. Sendo assim, há duas classes de problemas que podem ser solucionados por modelos de AM supervisionado: classificação e regressão. A diferença entre essas duas classes está justamente no sinal supervisorio. Em problemas de classificação, as saídas esperadas pelo modelo são valores discretos, ou classes. Já em problemas de regressão, as saídas esperadas pelo modelo são constituídas de valores contínuos.

3.2 Árvores de Decisão

Inicialmente proposto por (QUINLAN, 1986) como um algoritmo apenas de classificação e suportando apenas variáveis categóricas, originalmente chamado de ID3, as Árvores de Decisão (do inglês, Decision Trees: DTs), buscam, em cada nó da árvore, encontrar o atributo de entrada e o *threshold* que representam o maior ganho de informação. Sendo assim, cada nó da árvore representa o maior "corte" naquela altura



(a) Classificador

(b) Regressor

Figura 3.1 - Exemplo de utilização de Árvore de Decisão. Fonte: [scikit-learn contributors](#) (a)

da árvore. Ao final, os nós folhas, aqueles que não possuem nenhum nível abaixo, representam as previsões do modelo. Para problemas de classificação, a classe predita será aquela que possui o maior número de exemplares na folha. Já para problemas de regressão, a predição será a média aritmética do nó folha.

Ao contrários de RNAs, DTs são consideradas um modelo de fácil interpretação, ou "caixa branca", justamente por ser possível observar cada corte realizado pelo modelo em cada um dos níveis da árvore.

Com o desenvolvimento da área, outros algoritmos de Árvores de Decisão foram sendo apresentados, inclusive dando suporte a resolução a problemas de regressão. Alguns exemplos são os algoritmos C4.5 e CART ([BUTINCK et al., 2013](#)). A figura 3.1 apresenta uma representação gráfica de Árvores de Decisão para as duas classes de problemas.

Embora DTs sejam um modelo de fácil interpretação e que é capaz de resolver um grande número de problemas, especialmente aqueles em que as entradas se tratam de dados estruturados (i.e dados tabulares), esse algoritmo tende a sobreajustar (*overfit*) os dados de treinamento com o aumento da altura da árvore. Isso se explica pois, quanto maior for a árvore treinada, mais divisões dos dados serão feitas. Extrapolando essa ideia, é possível treinar uma DT que possui um nó folha para cada exemplo no conjunto de treinamento. Para contornar esse problema, são propostos alguns outros algoritmos baseados em DTs, como *Random Forests* (RF) e *Gradient Boosting* (GB).

3.3 Gradient Boosting

O algoritmo Gradient Boosting (GB) é baseado em DTs. Nele, vários modelos DTs, aqui chamados de *weak learners* são treinados. O treinamento se dá da seguinte forma. Inicialmente, é definida uma função de perda (*loss function*), de forma análoga ao treinamento de RNAs. Em seguida, o primeiro *weak learner* é treinado em uma amostra do conjunto de treinamento. Após, é calculado o erro deste modelo. Em seguida, o próximo *weak learner* será treinado, com uma outra amostra do conjunto de treinamento e recebendo também como entrada o erro do modelo anterior, chamado de resíduo. Ao propagar o resíduo de cada um dos modelos treinados, ao treinar o próximo *weak learner*, temos o ponto da curva de erro que o modelo está naquele estágio de treinamento. Sendo assim, podemos treinar o próximo modelo de modo a diminuir o o resíduo. Podemos traçar um paralelo desta estratégia com o algoritmo de descida de gradiente, utilizado para treinar RNAs.

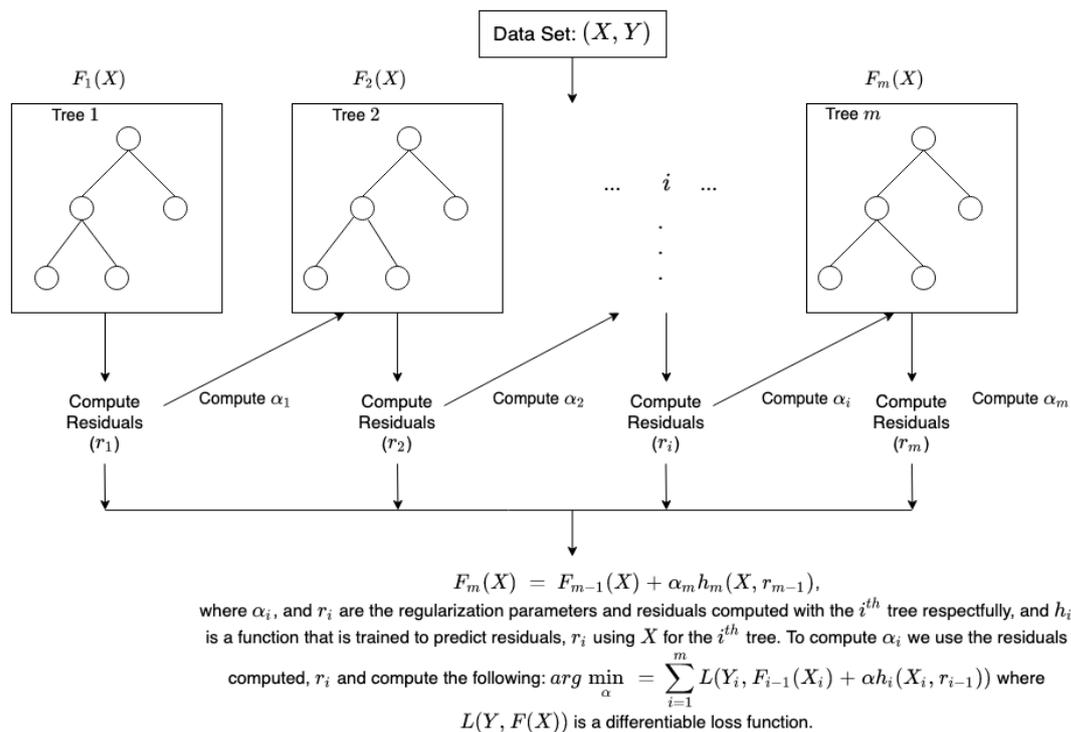


Figura 3.2 - Representação gráfica do algoritmo GB. Fonte: Amazon Web Services, Inc. (2023)

4 Metodologia de Treinamento

Para o desenvolvimento da estratégia de DA por aprendizado de máquina, escolheu-se o algoritmo GB. Mais especificamente, a implementação da biblioteca XGBoost desse algoritmo, a qual apresenta uma versão otimizada do algoritmo (CHEN; GUESTRIN, 2016). Inicialmente, era proposta a utilização de RNAs para o problema abordado. Entretanto, a percepção dos bons resultados de GB para problemas com dados estruturados, especialmente os resultados obtidos por (MONEGO et al., 2022) determinaram a mudança de escolha do algoritmo de treinamento do modelo.

4.1 Divisão dos dados e *K-Cross Validation*

Inicialmente, o conjunto de dados foi dividido entre conjunto de treinamento e conjunto de teste. O conjunto de treinamento é constituído pelos dados dos anos de 2015, 2016 e 2017, enquanto o conjunto de teste é constituído pelos dados do ano de 2018.

Após a divisão dos dados nesses dois conjuntos, deu-se início o processo de treinamento do modelo. Para obter um resultado com maior confiabilidade, utilizou-se a estratégia *K-Fold Cross-Validation* para avaliar a estratégia escolhida. Essa técnica consiste na divisão do conjunto de treinamento em K conjuntos de validação distintos. Assim, K modelos serão treinados, cada um deles sendo treinados e avaliados com diferentes partes do conjunto de treinamento. Assim, é possível avaliar o desempenho do modelo em uma maior variedade de dados, aumentando, portanto a confiabilidade da metodologia definida. A Figura 4.1 demonstra graficamente o processo de divisão dos dados da técnica de *K-Cross Validation*.

A estratégia de validação cruzada utilizada foi a seguinte: utilizou-se $K = 3$ *folds* distintos, cada um utilizando um dos anos do conjunto de treinamento como conjunto de avaliação. Ou seja, inicialmente, foi treinado um modelo com os anos de 2015 e 2016 e avaliado no ano de 2017; após, um modelo foi treinado com os anos de 2015-2017 e avaliado no ano de 2016 e, por fim, o último modelo foi treinado nos anos de 2016 e 2017 e avaliado no ano de 2015.

Os principais hiperparâmetros escolhidos para o treinamento do modelo XGBoost-DA estão contidos na Tabela 4.1. *Objective* diz respeito à função de perda utilizada pelo modelo. Nesse caso, está sendo utilizada uma função quadrática. *N_estimators* é o número máximo de *weak learners* que podem ser treinados. Por fim, *max_depth* é a altura máxima que cada um dos estimadores poderá ter.

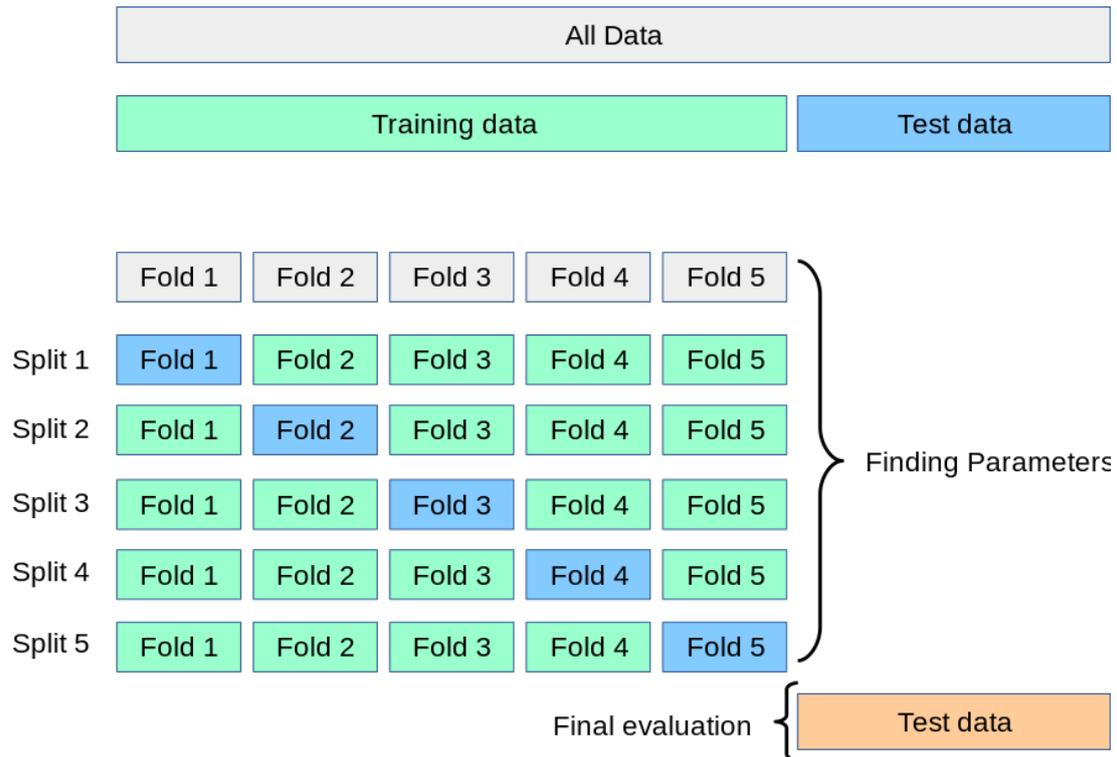


Figura 4.1 - Representação gráfica de *K-Cross Validation*. Fonte: [scikit-learn contributors \(b\)](#).

Tabela 4.1 - Configuração do treinamento.

Hiperparâmetro	Valor
<i>objective</i>	<i>reg:squarederror</i>
<i>n_estimators</i>	2000
<i>max_depth</i>	3

4.2 Análise dos resultados

Ao fim do processo de validação cruzada, três modelos distintos foram produzidos. A tabela 4.2 apresenta o erro quadrático de cada um dos modelos sobre o seu respectivo conjunto de validação. Observando a tabela, é possível perceber a proximidade dos erros de cada um dos modelos, o que indica que a metodologia de utilização do algoritmo GB para DA é adequada. Entretanto, só é possível realizar tal afirmação após a avaliação do modelo final no conjunto de teste.

Após a etapa de treinamento, foi escolhido o modelo com o menor erro quadrático dentre os modelos treinados durante a etapa de validação cruzada. O modelo escolhido foi utilizado para realizar previsões sobre o conjunto de teste previamente

Tabela 4.2 - Erro por *fold*.

<i>K-Fold</i>	Valor
1	0.22377
2	0.22615
3	0.22850
Média	0.22614

selecionado, alcançando um erro absoluto de 0.01412, e um desvio padrão de 0.47153. A Figura 4.2 apresenta as séries temporais dos dados de referência e das ativações do modelo sobrepostas. Percebe-se que, em alguns casos, o modelo subestima o valor de referência. Entretanto, para boa parte da série, os valores preditos são próximos aos dados de referência.

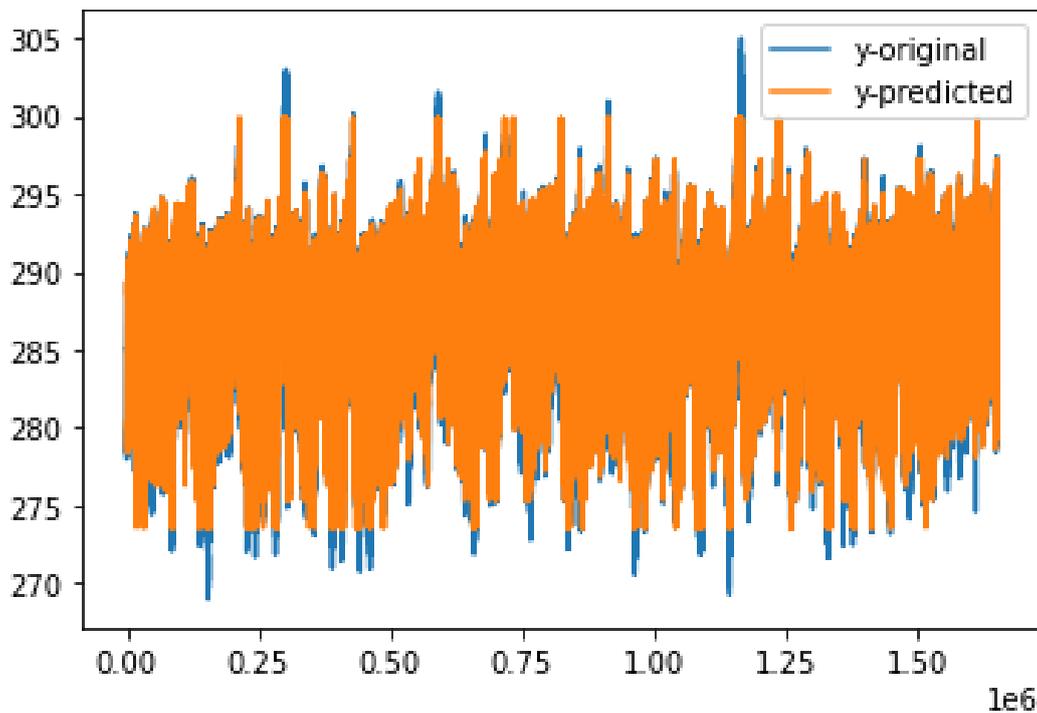


Figura 4.2 - Séries Temporais dos dados de referência e valores preditos. Fonte: O Autor.

Após análise dos resultados sobre o conjunto de teste de forma geral, foi selecionado um dia do conjunto de teste, para ser realizada uma avaliação do campo de temperatura gerado. O dia escolhido para essa análise foi 06/07. Assim, foram comparados os campos de temperatura gerados pelo modelo de AM e o 3D-Var. Os campos de

Temperatura gerados pelos dois métodos estão presentes na Figura 4.4 e a condição inicial do GFS é apresentada na Figura 4.3.

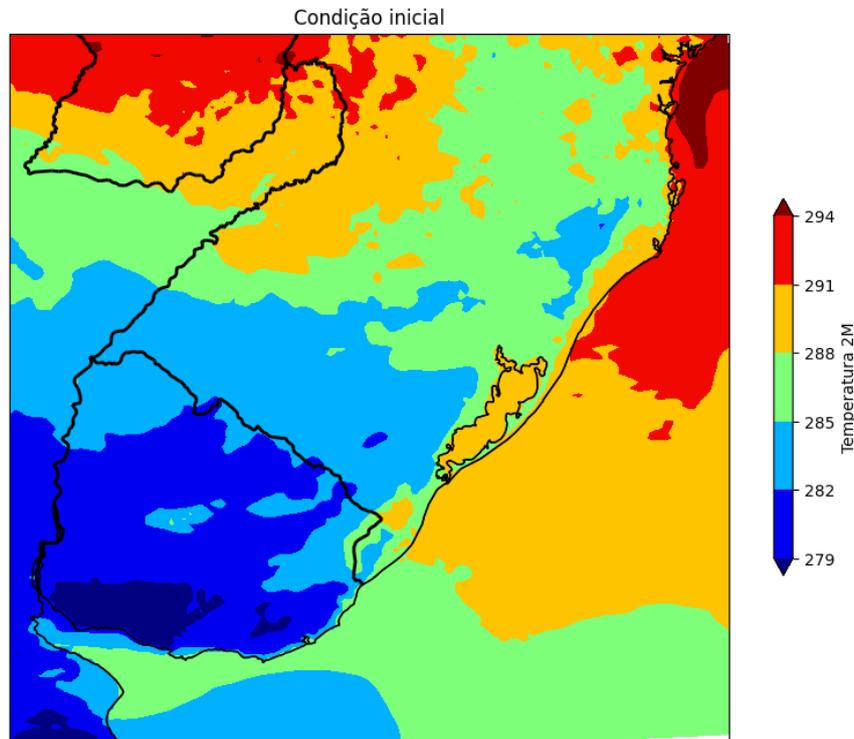


Figura 4.3 - Simulação gerada pelo modelo GFS. Fonte: O Autor

Como mostra a Figura 4.4, as condições iniciais geradas pelos dois métodos são bastante parecidas. O próximo passo, portanto, é avaliar a qualidade da previsão numérica obtida através da utilização da condição inicial gerada pelo modelo XGBoost-DA. Esse procedimento é descrito a seguir.

4.3 Comparação de Previsões Numéricas e discussão dos resultados

Após a geração de condições iniciais para um mesmo dia, foram realizadas duas previsões de vinte e quatro horas com o modelo WRF. Em uma delas, foi utilizada a condição inicial gerada pelo sistema 3D-Var. A outra, utilizando a condição inicial gerada pelo XGBoost-DA. A Figura 4.5 apresenta os campos de temperatura previstos. É possível perceber pela Figura que os campos previstos são bastante similares. Esse resultado é esperado, se analisarmos a Figura 4.4, pois os campos de condição inicial são bastante parecidos também. Observando-se o campo de diferenças entre as duas previsões, conseguimos perceber o quão boa, de fato, é a previsão gerada a

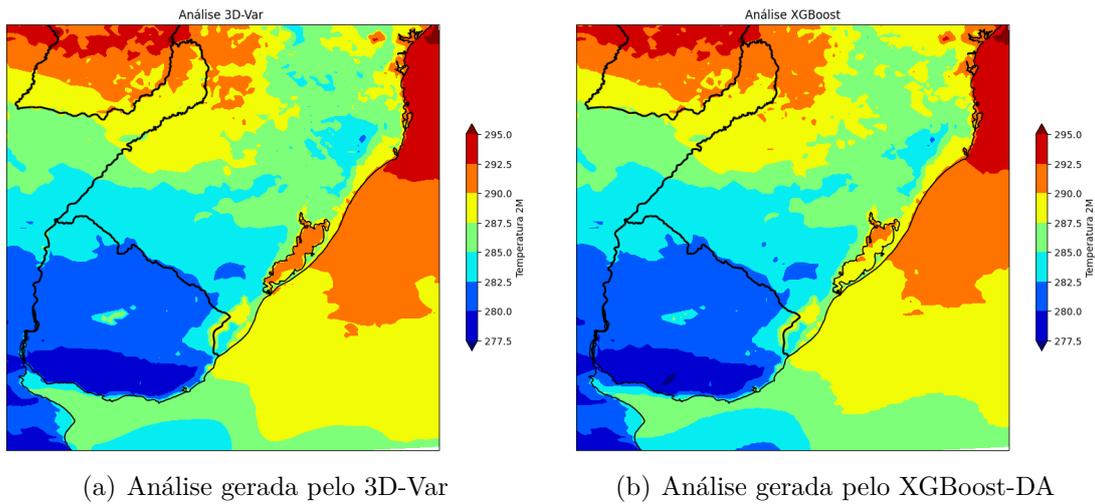
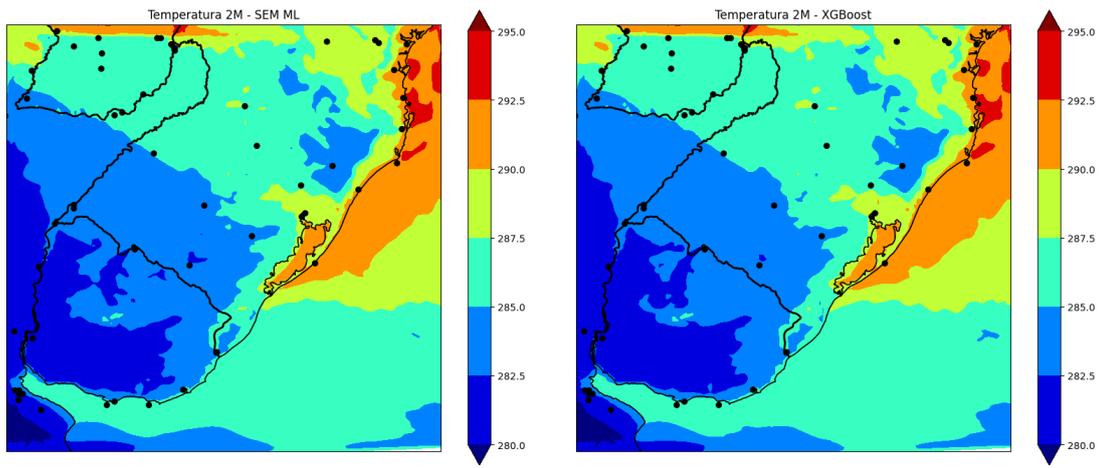


Figura 4.4 - Análises geradas por métodos de DA. Fonte: O Autor

partir da condição inicial gerada pelo modelo XGBoost-DA, uma vez ela se aproxima muito da gerada pela condição inicial do 3D-Var: os pontos de maior diferença são de menos de 2 graus. Sendo assim, é possível afirmar que a previsão gerada a partir da condição inicial gerada pelo modelo XGBoost-DA alcança resultados tão bom quanto aquela gerada pelo sistema 3D-Var. Sendo assim, a utilização de técnicas de AM supervisionado se torna uma ferramenta promissora para a resolução do problema de assimilação de dados.



(a) Análise gerada pelo 3D-Var

(b) Análise gerada pelo XGBoost-DA

Figura 4.5 - Previsões geradas a partir de análises geradas por meio de DA. Fonte: O Autor

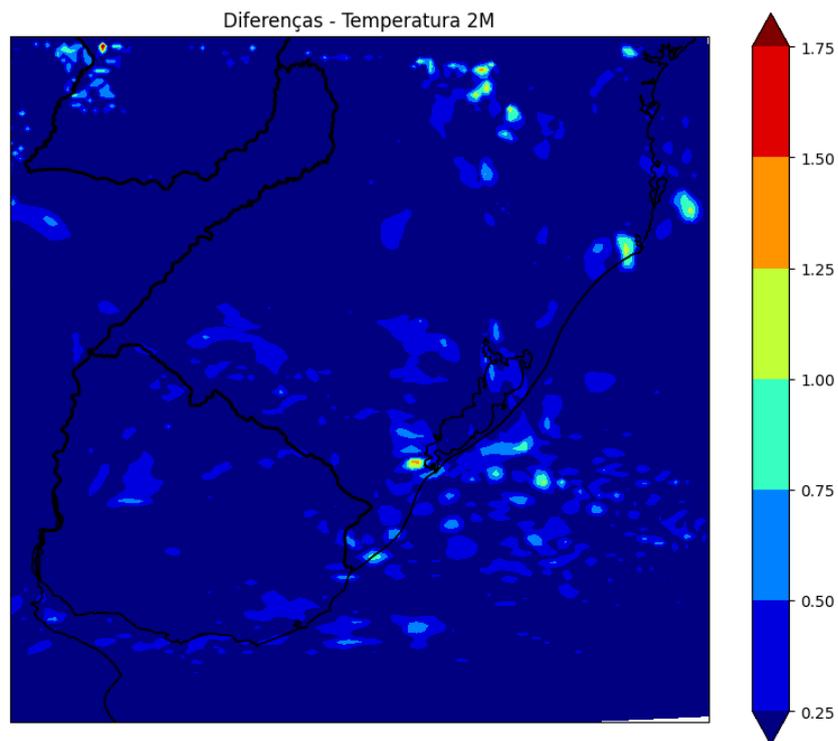


Figura 4.6 - Diferença absoluta entre as previsões geradas. Fonte: O Autor

5 Discussão dos Resultados

Durante o desenvolvimento desta pesquisa, foram estudados algoritmos de AM supervisionado e sua aplicação no problema de DA de dados de temperatura. Foi utilizado um conjunto de dados contendo dados de simulação do modelo GFS, bem como um conjunto de dados de observações e um conjunto de análises do sistema 3D-Var, os quais servira de referência para o treinamento do modelo de AM. Para o desenvolvimento do modelo, escolheu-se o algoritmo de treinamento *Gradient Boosting*. Para obter maior confiabilidade nos resultados, foi utilizada a técnica de *K Cross-Validation*. Ao término do treinamento, foi escolhido um dia do conjunto de teste para servir como condição inicial para a realização de uma previsão de vinte e quatro horas. Os dados de previsão obtidos demonstram que a utilização do modelo de AM desenvolvido consegue alcançar resultados similares às previsões geradas a partir da assimilação pelo sistema 3D-Var, tornando-se, portanto, uma ferramenta promissora para a resolução deste tipo de problema.

Trabalhos futuros incluem a utilização das técnicas aplicadas durante o desenvolvimento dessa pesquisa para outros domínios, bem como a assimilação de outras variáveis do modelo WRF. Futuramente também será estudada a diferença de custo computacional entre as técnicas utilizadas nesta pesquisa.

REFERÊNCIAS BIBLIOGRÁFICAS

Amazon Web Services, Inc. **Amazon SageMaker - XGBoost How It Works**. 2023. <<https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html>>. Acesso em: 11 de Agosto 2023. iii, 7

ANOCHI, J. A.; VELHO, H. F. de C. Optimization of feedforward neural network by multiple particle collision algorithm. In: IEEE. **2014 IEEE Symposium on Foundations of Computational Intelligence (FOCI)**. [S.l.], 2014. p. 128–134. 2

BUITINCK, L.; LOUPPE, G.; BLONDEL, M.; PEDREGOSA, F.; MUELLER, A.; GRISEL, O.; NICULAE, V.; PRETTENHOFER, P.; GRAMFORT, A.; GROBLER, J.; LAYTON, R.; VANDERPLAS, J.; JOLY, A.; HOLT, B.; VAROQUAUX, G. API design for machine learning software: experiences from the scikit-learn project. In: **ECML PKDD Workshop: Languages for Data Mining and Machine Learning**. [S.l.: s.n.], 2013. p. 108–122. 6

CHEN, T.; GUESTRIN, C. Xgboost: A scalable tree boosting system. In: **Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining**. [S.l.: s.n.], 2016. p. 785–794. 8

MONEGO, V. S.; ANOCHI, J. A.; VELHO, H. F. de C. South america seasonal precipitation prediction by gradient-boosting machine-learning approach. **Atmosphere**, MDPI, v. 13, n. 2, p. 243, 2022. 2, 8

QUINLAN, J. R. Induction of decision trees. **Machine learning**, Springer, v. 1, p. 81–106, 1986. 5

scikit-learn contributors. **Decision Trees - scikit-learn Documentation**. <<https://scikit-learn.org/stable/modules/tree.html>>. Acesso em: 11 de Agosto de 2023. iii, 6

_____. **KFold - scikit-learn Documentation**. <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html>. Acesso em: 11 de Agosto de 2023. iii, 9