

**UNIVERSIDADE FEDERAL DE MATO GROSSO
INSTITUTO DE FÍSICA
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA AMBIENTAL**

**PREVISÃO DE RADIAÇÃO SOLAR COM MÉTODOS DE MACHINE
LEARNING EM SÉRIES TEMPORAIS**

MARLON CAPELARI

DANIELA DE OLIVEIRA MAIONCHI
Orientadora

**Cuiabá, MT
Maio, 2024**

**UNIVERSIDADE FEDERAL DE MATO GROSSO
INSTITUTO DE FÍSICA
PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA AMBIENTAL**

**PREVISÃO DE RADIAÇÃO SOLAR COM MÉTODOS DE MACHINE
LEARNING EM SÉRIES TEMPORAIS**

MARLON CAPELARI

**Dissertação apresentada ao Programa de
Pós-graduação em Física Ambiental da
Universidade Federal de Mato Grosso, como
parte dos requisitos para obtenção do título
de Mestre em Física Ambiental.**

DANIELA DE OLIVEIRA MAIONCHI
Orientadora

Cuiabá, MT
Maior, 2024

Dados Internacionais de Catalogação na Fonte.

C238p Capelari, Marlon.

Previsão de radiação solar com métodos de machine learning em séries temporais [recurso eletrônico] / Marlon Capelari. -- Dados eletrônicos (1 arquivo : 39 f., il. color., pdf). -- 2024.

Orientador: Daniela de Oliveira Maionchi.

Dissertação (mestrado) - Universidade Federal de Mato Grosso, Instituto de Física, Programa de Pós-Graduação em Física Ambiental, Cuiabá, 2024.

Modo de acesso: World Wide Web: <https://ri.ufmt.br>.

Inclui bibliografia.

1. Machine learning. 2. GRU. 3. MLP. 4. Python. 5. Redes neurais recorrentes. I. Maionchi, Daniela de Oliveira, *orientador*. II. Título.

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Permitida a reprodução parcial ou total, desde que citada a fonte.



MINISTÉRIO DA EDUCAÇÃO

UNIVERSIDADE FEDERAL DE MATO GROSSO

PRÓ-REITORIA DE ENSINO DE PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM FÍSICA AMBIENTAL

FOLHA DE APROVAÇÃO

TÍTULO: PREVISÃO DE RADIAÇÃO SOLAR COM MÉTODOS DE MACHINE LEARNING EM SÉRIES TEMPORAIS

AUTOR: MESTRANDO MARLON CAPELARI

Dissertação defendida e aprovada em **24 de maio de 2024**.

COMPOSIÇÃO DA BANCA EXAMINADORA

1. Profa. Dra. Daniela de Oliveira Maionchi (Presidente da Banca / **Orientadora**)

INSTITUIÇÃO: Universidade Federal de Mato Grosso

2. Prof. Dr. Sérgio Roberto de Paulo (Membro Interno)

INSTITUIÇÃO: Universidade Federal de Mato Grosso

3. Prof. Dr. Junior Gonçalves da Silva (Membro Interno)

INSTITUIÇÃO: Universidade Federal de Mato Grosso

4. Prof. Dr. Saulo Davi Soares e Reis (Membro Externo)

INSTITUIÇÃO: Universidade Federal do Ceará

Cuiabá, 24/05/2024.



Documento assinado eletronicamente por **MARCELO SACARDI BIUDES**, **Coordenador(a) de Programas de Pós-Graduação em Física Ambiental - IF/UFMT**, em 29/05/2024, às 12:33, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Junior Gonçalves da Silva**, **Usuário Externo**, em 29/05/2024, às 13:37, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **DANIELA DE OLIVEIRA MAIONCHI**, **Docente da Universidade Federal de Mato Grosso**, em 29/05/2024, às 16:22, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **Saulo Davi Soares e Reis, Usuário Externo**, em 29/05/2024, às 16:30, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



Documento assinado eletronicamente por **SERGIO ROBERTO DE PAULO, Docente da Universidade Federal de Mato Grosso**, em 29/05/2024, às 19:03, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do [Decreto nº 10.543, de 13 de novembro de 2020](#).



A autenticidade deste documento pode ser conferida no site http://sei.ufmt.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **6886124** e o código CRC **E9D6F29F**.

Dedicatória

**A Deus,
aos meus pais,
professores e amigos.
Dedico com profunda gratidão.**

Agradecimentos

À Professora Dra. Daniela de Oliveira Maionchi, pela sua notável orientação, confiança, apoio e paciência incansável ao longo do processo de desenvolvimento deste estudo. Sua orientação foi fundamental para o sucesso deste trabalho e para o meu crescimento acadêmico e pessoal.

A todos os professores do Programa de Pós-Graduação em Física Ambiental, pela incansável dedicação e empenho na formação de cada aluno, além de desempenharem papéis únicos em minha jornada acadêmica. Em especial ao professor Dr. Sérgio Roberto de Paulo, pela compreensão, direcionamento e apoio.

Aos meus queridos colegas do Programa de Pós-Graduação em Física Ambiental, os quais se tornaram grandes amigos, pelos ótimos momentos compartilhados e por desempenharem um papel essencial nesta jornada. Em especial a Rafaely Schwaab Sampaio, que compartilhou comigo não apenas uma jornada de aprendizado mas também momentos de alegria e amor. Sua presença foi uma fonte constante de apoio, conhecimento e encorajamento nos momentos mais desafiadores.

Epígrafes, Frases ou Pensamentos

**Com grandes poderes,
vêm grandes responsabilidades.
Tio Ben**

SUMÁRIO

LISTA DE FIGURAS	ii
LISTA DE TABELAS	iii
LISTA DE ABREVIACÕES E/OU SÍMBOLOS	iv
RESUMO	v
ABSTRACT	vi
1 INTRODUÇÃO	1
1.1 OBJETIVO	3
1.1.1 Objetivo Geral	3
1.1.2 Objetivos Específicos	3
2 MATERIAIS E MÉTODOS	4
2.1 Área de Estudo	4
2.2 Coleta de Dados	5
2.3 Modelagem com Machine Learning	6
2.3.1 Tratamento de Dados	9
2.3.2 Redes Neurais	11
2.3.3 Multi-Layer Perceptron	11
2.3.4 Gated Recurrent Unit (GRU)	12
2.4 Métricas	14
3 RESULTADOS E DISCUSSÕES	17
3.1 Seleção de Variáveis	17

3.2	Resultados Gerais	18
3.3	Sinop	20
3.4	Alta Floresta	27
4	CONCLUSÃO	33
5	REVISÃO BIBLIOGRÁFICA	35

Lista de Figuras

1	Biomass de Mato Grosso [10], [11]. Fonte: Autor, 2024	4
2	Representação dos possíveis ajustes de modelo: Underfitting (subajuste), Overfitting (sobreajuste) e Right fit (Ajuste Adequado) Fonte: Autor, 2024	8
3	Resultado da separação dos dados de Sinop para treinamento e teste. Fonte: Autor, 2024	10
4	Resultado da separação dos dados de Alta Floresta para treinamento e teste. Fonte: Autor, 2024	10
5	Achatamento (flattening) dos dados para entrada na rede neural. Fonte: Autor, 2024	12
6	Arquitetura <i>GRU</i> . Fonte: Autor, 2024	14
7	Gráfico do treinamento da rede neural MLP para os dados de Sinop. Fonte: Autor, 2024	21
8	Resultado do teste da rede neural MLP para os dados de Sinop. Fonte: Autor, 2024	21
9	Gráficos da relação entre dados medidos e previstos para o teste utilizando MLP para os dados de Sinop. Fonte: Autor, 2024	21
10	Resultado dos testes da rede neural MLP para os dados de Sinop em diferentes intervalos de tempo. Fonte: Autor, 2024	22
11	Gráficos da relação entre dados medidos e previstos utilizando MLP para os dados de Sinop em diferentes intervalos de tempo. Fonte: Autor, 2024	23
12	Gráfico de treinamento da rede neural GRU para os dados de Sinop. Fonte: Autor, 2024	23
13	Resultado do teste da rede neural GRU para os dados de Sinop. Fonte: Autor, 2024	24
14	Gráficos da relação entre dados medidos e previstos para o teste utilizando GRU para os dados de Sinop. Fonte: Autor, 2024	24
15	Resultado dos testes da rede neural GRU para os dados de Sinop em diferentes intervalos de tempo. Fonte: Autor, 2024	25
16	Gráficos da relação entre dados medidos e previstos utilizando GRU para os dados de Sinop em diferentes intervalos de tempo. Fonte: Autor, 2024	25

17	Gráfico do treinamento da rede neural MLP para os dados de Alta Floresta. Fonte: Autor, 2024	27
18	Resultado do teste da rede neural MLP para os dados de Alta Floresta. Fonte: Autor, 2024	27
19	Gráficos da relação entre dados medidos e previstos para o teste utilizando MLP para os dados de Alta Floresta. Fonte: Autor, 2024	27
20	Resultado dos testes da rede neural MLP para os dados de Alta Floresta em diferentes intervalos de tempo. Fonte: Autor, 2024	28
21	Gráficos da relação entre dados medidos e previstos utilizando MLP para os dados de Alta Floresta em diferentes intervalos de tempo. Fonte: Autor, 2024 .	29
22	Gráfico do treinamento da rede neural GRU para os dados de Alta Floresta. Fonte: Autor, 2024	29
23	Resultado do teste da rede neural GRU para os dados de Alta Floresta. Fonte: Autor, 2024	30
24	Gráficos da relação entre dados medidos e previstos para o teste utilizando GRU para os dados de Alta Floresta. Fonte: Autor, 2024	30
25	Resultado dos testes da rede neural GRU para os dados de Alta Floresta em diferentes intervalos de tempo. Fonte: Autor, 2024	31
26	Gráficos da relação entre dados medidos e previstos utilizando GRU para os dados de Alta Floresta em diferentes intervalos de tempo. Fonte: Autor, 2024 .	31

Lista de Tabelas

1	Descrição das bibliotecas <i>Python</i>	9
2	Hiperparâmetros de treinamento das redes neurais.	17
3	Resultado da aplicação da correlação de Pearson entre variáveis do conjunto de dados de Sinop.	18
4	Resultado da aplicação da correlação de Pearson entre variáveis do conjunto de dados de Alta Floresta.	18
5	Métricas de avaliação de desempenho da rede neural MLP aplicadas aos dados de Sinop.	18
6	Métricas de avaliação de desempenho da rede neural GRU aplicadas aos dados de Sinop.	19
7	Métricas de avaliação de desempenho da rede neural MLP aplicadas aos dados de Alta Floresta.	19
8	Métricas de avaliação de desempenho da rede neural GRU aplicadas aos dados de Alta Floresta.	19
9	Métricas de avaliação de desempenho das redes neurais MLP e GRU aplicadas aos dados de Sinop utilizando umidade relativa a 36 metros de altura	20
10	Métricas de avaliação de desempenho das redes neurais MLP e GRU aplicadas aos dados de Sinop utilizando umidade relativa a 41 metros de altura	20
11	Métricas de avaliação de desempenho das redes neurais MLP e GRU aplicadas aos dados de Sinop.	26
12	Métricas de avaliação de desempenho das redes neurais MLP e GRU aplicadas aos dados de Alta Floresta.	32

LISTA DE ABREVIACÕES E/OU SÍMBOLOS

IBGE	Instituto Brasileiro de Geografia e Estatística
INMET	Instituto Nacional de Meteorologia
GRU	Gated Recurrent Unit
LSTM	Long short-term memory
MAE	Erro Médio Absoluto
MLP	Multi-layer Perceptron
MSE	Erro Quadrático Médio
OMM	Organização Meteorológica Mundial
PPT	Precipitação
R²	Coefficiente de Determinação
R_n	Radiação Líquida
RNA	Rede Neural Artificial
RNN	Rede Neural Recorrente
RMSE	Raiz do Erro Quadrático Médio
R_{sd}	Radiação Solar Direta
UR	Umidade Relativa
T	Temperatura

RESUMO

CAPELARI, M. Previsão de Radiação Solar com Métodos de Machine Learning em Séries Temporais. Dissertação (Mestrado em Física Ambiental) - Instituto de Física, Universidade Federal de Mato Grosso.

Este trabalho tem como objetivo desenvolver modelos de rede neural, utilizando arquiteturas de *machine learning*, que realize a estimativa da radiação solar a fim de sanar a escassez de dados ocasionada, por exemplo, por panes em estações meteorológicas, erros de leitura (lacunas) e/ou carência de equipamentos de coleta em algumas localidades. Os modelos de redes neurais utilizados para efetuar a estimativa de radiação foram, especificamente, redes neurais do tipo *Gated Recurrent Unit* (GRU) e *Multi-Layer Perceptron* (MLP). Os conjuntos de dados de Sinop e Alta Floresta, utilizados neste trabalho, foram coletados de uma torre micrometeorológica e de uma estação meteorológica, respectivamente. Foram utilizadas como variáveis preditoras a temperatura, umidade relativa, precipitação e umidade do solo para Sinop, enquanto para Alta Floresta foram utilizadas a temperatura, umidade relativa, precipitação e velocidade do vento. O algoritmo foi desenvolvido em *Python*, empregando principalmente bibliotecas como *pandas* para manipulação dos dados, *scikit-learn* para normalização dos dados, *Keras* e *TensorFlow* para construção das redes neurais, *scipy* para análise estatística dos dados e avaliação dos modelos, entre outras. Para a validação do algoritmo foram utilizadas métricas como Erro Médio Absoluto (MAE), Erro Médio Quadrático (RMSE), Coeficiente de Determinação (R^2), Índice de Willmott e as Correlações de Spearman e Pearson. Estas métricas foram aplicadas a diferentes intervalos temporais, tais como média horária, média diária e média mensal. O estudo demonstrou que ambas as redes neurais apresentaram resultados satisfatórios. Para o conjunto de Sinop o coeficiente de determinação (R^2) teve um mínimo de 88% para médias mensais, alcançando 95% de ajuste dos modelos para a rede GRU e MLP para médias horárias. A correlação de Pearson apresentou um mínimo de 94% de ajuste para médias mensais e chegando a atingir próximo de 98% para médias horárias. Para o conjunto de dados de Alta Floresta ambas as redes obtiveram bons resultados para intervalos horários alcançando coeficiente de determinação de 92% para MLP e 93% para GRU, porém devido aos erros de leitura apresentados a rede não obteve bons resultados para médias diárias e mensais, identificando assim que a qualidade dos dados impacta substancialmente na qualidade e ajuste final do modelo.

Palavras-chave: Machine Learning, GRU, MLP, Python, Redes Neurais Recorrentes, Séries Temporais, Variáveis Ambientais, Radiação Solar

ABSTRACT

CAPELARI, M. Solar Radiation Forecasting with Machine Learning Methods in Time Series. Dissertation (Master's degree in Environmental Physics) - Institute of Physics, Federal University of Mato Grosso.

This work aims to develop neural network models, using *machine learning* architectures, that estimate solar radiation in order to remedy the lack of data caused, for example, by breakdowns in meteorological stations, reading errors (gaps) and/or lack of collection equipment in some locations. The neural network models used to estimate radiation were, specifically, neural networks of the type *Gated Recurrent Unit* (GRU) and *Multi-Layer Perceptron* (MLP). The Sinop and Alta Floresta datasets used in this work were collected from a micrometeorological tower and a meteorological station, respectively. Temperature, relative humidity, precipitation and soil moisture were used as predictor variables for Sinop, while for Alta Floresta temperature, relative humidity, precipitation and wind speed were used. The algorithm was developed in *Python*, mainly using libraries such as *pandas* for data manipulation, *scikit-learn* for data normalization, *Keras* and *TensorFlow* for construction neural networks, *scipy* for statistical analysis of data and evaluation of models, among others. To validate the algorithm, metrics such as Mean Absolute Error (MAE), Mean Square Error (RMSE), Coefficient of Determination (R^2), Willmott Index and Spearman and Pearson Correlations were used. These metrics were applied to different time intervals, such as hourly average, daily average and monthly average. The study demonstrated that both neural networks presented satisfactory results. For the Sinop group, the coefficient of determination (R^2) had a minimum of 88% for monthly averages, reaching 95% model adjustment for the GRU and MLP network for hourly averages. The Pearson correlation presented a minimum of 94% adjustment for monthly averages and reaching close to 98% for hourly averages. For the Alta Floresta data set, both networks obtained good results for hourly intervals, reaching a coefficient of determination of 92% for MLP and 93% for GRU, however, due to the reading errors presented, the network did not obtain good results for daily averages. and monthly, thus identifying that the quality of the data substantially impacts the quality and final adjustment of the model.

Keywords: Machine Learning, GRU, MLP, Python, Recurrent Neural Networks, Time Series, Environmental Variables, Solar Radiation.

1 INTRODUÇÃO

A radiação solar é o fluxo de energia emitida pelo sol e transmitida sob a forma de radiação eletromagnética. Esta é considerada a fonte de energia primária da Terra, sendo fundamental para manutenção da vida no planeta. Além disso, é capaz de manter os processos físico-químicos da biosfera, tornando-se imprescindível para uma série de atividades antrópicas e aplicações, como por exemplo, meteorologia, engenharia, agricultura, geração de energia, saúde e em muitos campos das ciências ambientais[1] [2]. O Sol age de forma semelhante a de um corpo negro, com temperatura de superfície em torno de 5800K e com irradiância solar média que chega ao topo da atmosfera terrestre sendo de 1367 W/m^{-2} [3]. Inesgotável e limpa, a radiação solar é considerada uma das fontes de energia renováveis mais promissoras para a geração de energia em grande escala [4].

A falta de consistência ou a completa ausência de séries históricas de dados meteorológicos em determinadas regiões, somadas à escassa distribuição de estações meteorológicas no Brasil, emerge como um dos principais obstáculos para a compreensão de fenômenos climáticos no país. Considerando isso, tem-se buscado cada vez mais outros mecanismos para obtenção de dados meteorológicos [5].

É indispensável para a sociedade entender as condições que desencadeiam eventos climáticos e ser capaz de implementar medidas preventivas de forma oportuna para mitigar o impacto negativo de diversos eventos meteorológicos [6]. Devido às mudanças ambientais, a compreensão do comportamento climático emergiu como uma questão fundamental em diversas áreas de estudo, tornando-se essencial para a humanidade. No entanto, essa compreensão pode ser desafiadora, especialmente em áreas como a meteorologia, onde a disponibilidade de dados desempenha um papel crucial. A escassez ou a qualidade inadequada dos dados meteorológicos pode aumentar significativamente a dificuldade do processo de previsão temporal [7]. No Brasil a realidade não é diferente. De acordo com o Instituto Nacional de Meteorologia (INMET), até março de 2024, verificou-se que, das 566 estações automáticas monitoradas, 66 estão em estado de pane, ou seja, representando um pouco mais de 11% do total, comprometendo a disponibilização de dados em várias regiões do país.

É observável uma variedade de estudos nos quais a análise da radiação desempenha funções fundamentais na geração de energia, no crescimento vegetal e nas alterações climáticas, com o *machine learning* frequentemente empregado para a estimação de sua incidência [6]. Rudin e Wagstaff (2014) [8] relatam como a adoção de técnicas computacionais, como o *machine learning*, emergiu como uma ferramenta imprescindível na resolução de uma variedade de problemas em diversos campos, incluindo análises meteorológicas. A eficácia das redes neurais é comprovada em sua capacidade de aprender representações complexas dos dados de treina-

mento e estabelecer conexões significativas com a variável a ser prevista [9]. Essa abordagem pode ser aplicada na estimativa de dados meteorológicos, como a variável de radiação solar, utilizando um conjunto de dados climáticos ao longo de vários anos (séries temporais climáticas) como informação de entrada.

1.1 OBJETIVO

1.1.1 Objetivo Geral

Desenvolver modelos de *machine learning* utilizando arquiteturas *Gated Recurrent Unit* (GRU) e *Multi-Layer Perceptron* (MLP) com o objetivo de estimar dados e realizar o preenchimento das lacunas de radiação solar. As variáveis ambientais, como temperatura, umidade do ar, umidade do solo, precipitação e velocidade do vento, serão utilizadas como entradas nos modelos. A análise comparativa do desempenho de ambas redes neurais foi conduzida em diferentes intervalos de tempo.

1.1.2 Objetivos Específicos

- Implementar e treinar modelos de redes neurais do tipo GRU e MLP utilizando os dados coletados de dois ecossistemas diferentes.
- Avaliar o desempenho dos modelos treinados utilizando métricas de avaliação, como por exemplo, Erro Médio Absoluto (MAE), Raiz do Erro Quadrático Médio (RMSE), coeficiente de determinação (R^2), índice de *Willmott*, coeficiente de correlação de *Spearman* e coeficiente de correlação de *Pearson*.
- Comparar os resultados obtidos pelos modelos em diferentes intervalos de tempo.
- Analisar e interpretar os resultados identificando possíveis áreas para melhorias ou refinamentos.

2 MATERIAIS E MÉTODOS

2.1 Área de Estudo

O estudo foi conduzido utilizando conjuntos de dados retirados de dois ecossistemas localizados na região norte estado de Mato Grosso, Brasil, ambos localizados no bioma amazônico (Figura 1).

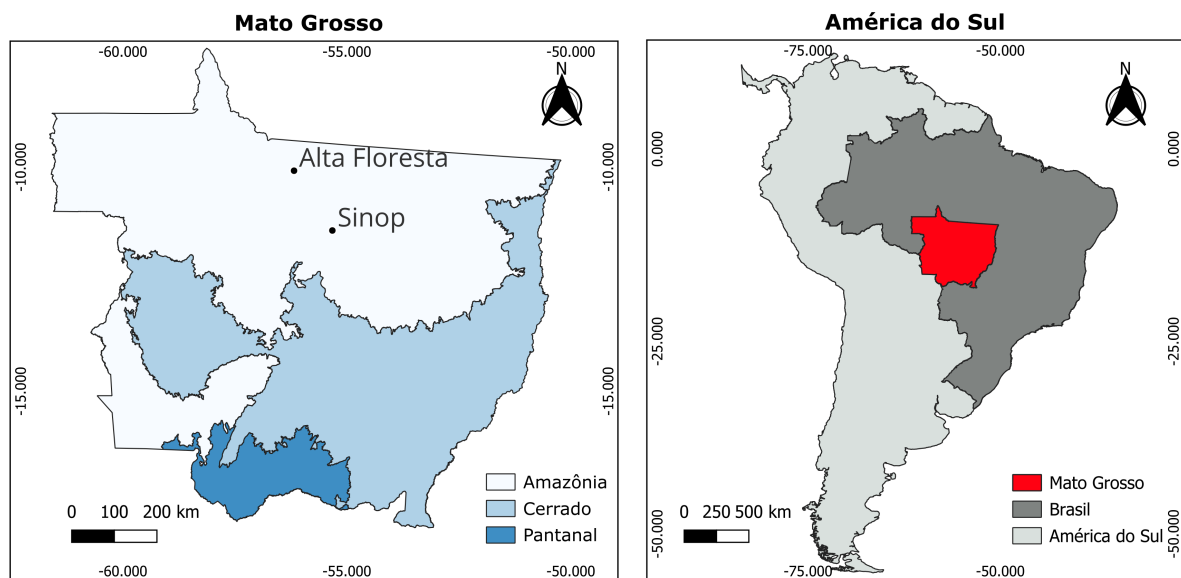


Figura 1: Biomas de Mato Grosso [10], [11].

Fonte: Autor, 2024

Localizado na região de Sinop, o primeiro ecossistema está situado na Fazenda Maracaí, a noroeste de Sinop ($11^{\circ} 24.75' S$ e $55^{\circ} 19.50' W$). A torre micrometeorológica possui 42 metros de altura e está inserida em uma floresta de transição Amazônica-Cerrado. De acordo com a classificação climática de Köppen-Geiger, o clima da região está classificado como Aw, caracterizado como tropical chuvoso, quente e úmido [12], com uma estação chuvosa de outubro a abril e uma estação seca de maio a setembro [13]. A precipitação média anual é de 1975 mm, e a temperatura média anual é de $24.7^{\circ}C$ [14]. As espécies arbóreas dominantes na floresta são *Brosimum lactescens*, *Qualea paraensis* e *Tovomita schomburkii*. O solo é caracterizado como pobre em nutrientes, altamente poroso e com rápida drenagem após eventos de chuva [15].

Já o segundo ecossistema está localizado na região de Alta Floresta. Os dados foram coletados de uma estação meteorológica localizada as margens da rodovia MT-325, a 23km de distância ao sul da cidade ($10^{\circ} 4' 45.43''S$ e $56^{\circ} 10' 49.70''W$). O clima da região é classificado como Aw seguindo classificação climática de Köppen-Geiger, caracterizada por um clima pre-

dominantemente tropical chuvoso, quente e úmido [12] com estação seca de junho a setembro. A precipitação média anual é de 2230 mm/ano⁻¹ e a temperatura média anual é de 25.7 °C [16].

É importante ressaltar a distinção entre torre micrometeorológica, como é o caso de Sinop, e estação meteorológica, como é o caso de Alta Floresta. Uma torre micrometeorológica é uma estrutura de variada estatura, projetada para medir e registrar uma ampla gama de variáveis meteorológicas e ambientais em várias alturas ao longo de sua extensão. Geralmente é instalada em áreas de interesse específico, como campos agrícolas, florestas ou ecossistemas naturais, e é equipada com vários sensores e instrumentos de alta precisão. A frequência de amostragem desses dados pode variar de segundos a horas, proporcionando uma visão detalhada das condições microclimáticas em diferentes altitudes.

Por outro lado, uma estação meteorológica é um local estruturado que tem por objetivo coletar dados sobre o clima de determinada região, tais como temperatura, umidade do ar, radiação solar, velocidade do vento, entre outros, e pode ser dividida em estações convencionais e automáticas. As convencionais exigem a presença diária de um observador meteorológico para a coleta de dados, já as automáticas possuem um conjunto de sensores que enviam as informações coletadas para serem armazenadas em um dispositivo chamado *datalogger*. As leituras podem ser configuradas para ocorrerem em intervalos de tempo determinados [17]. As estações meteorológicas normalmente são constituídas por conjuntos de dispositivos dispostos dentro de uma área cercada, concebida e operada seguindo normas internacionais. Elas abrigam todos os dispositivos essenciais para a análise do clima e das condições atmosféricas, adotando um padrão específico de infraestrutura. Por exemplo, os instrumentos geralmente são pintados de branco para mitigar o aquecimento, enquanto o solo ao redor é mantido gramado para evitar influências térmicas. Essas e outras medidas garantem que os dados coletados sejam comparáveis entre diferentes estações, um requisito fundamental para a validade científica [18].

A Organização Meteorológica Mundial (OMM) estabeleceu que cada país é responsável por gerenciar suas próprias redes de estações meteorológicas. No Brasil, esta responsabilidade fica a cargo do Instituto Nacional de Meteorologia (INMET) [17].

2.2 Coleta de Dados

Para o conjunto de dados de Sinop, dentre as variáveis ambientais disponíveis, foram selecionadas as seguintes: temperatura a 40 metros (T), umidade relativa do ar a 36 metros (UR), precipitação (PPT), conteúdo volumétrico de umidade do solo (Soil H₂O) e radiação solar direta (Rsd). O conjunto de dados abrange o período de Janeiro a Dezembro de 2002, com a coleta ocorrendo em intervalos de meia hora, totalizando 17.521 registros. Após serem desconside-

radas as falhas de leitura em uma ou mais variáveis, obteve-se um aproveitamento de 11.235 dados úteis, cerca de 64.1% dos dados originais.

Já para o conjunto de dados de Alta Floresta o banco de dados do Instituto Nacional de Meteorologia (INMET) oferece várias configurações, das quais foram selecionadas as seguintes variáveis ambientais para este estudo: temperatura (T), umidade relativa do ar (UR), precipitação (PPT), velocidade do vento (WIND) e radiação líquida (Rn). O conjunto de dados abrange o período de maio de 2007 a dezembro de 2022, com a coleta ocorrendo a cada hora, totalizando 136.872 registros. Após serem desconsideradas as falhas de leitura em uma ou mais variáveis, obteve-se um aproveitamento de 66.634 dados úteis, cerca de 48.6% dos dados originais.

2.3 Modelagem com Machine Learning

A área de *machine learning*, uma subárea da Inteligência Artificial, permite que computadores aprendam por meio de algoritmos, de forma semelhante ao aprendizado humano. Os algoritmos são alimentados com grandes quantidades de dados, possibilitando-lhes identificar padrões e correlações entre variáveis, permitindo assim realizar previsões ou tomar decisões de forma autônoma [19].

Este estudo visa a criação de um modelo de estimativa que tenha a radiação como variável de saída, utilizando outras quatro variáveis ambientais como entradas. Para alcançar este resultado, foram implementados algoritmos de regressão supervisionada. Algoritmos de regressão indicam que a variável de saída que se pretende estimar é contínua, ou seja, o objetivo é estimar um valor numérico [20]. Já o aprendizado supervisionado significa que o algoritmo recebe o conjunto de treinamento, que é composto tanto por variáveis de entrada (variáveis independentes), quanto pela variável de saída (variável dependente) [21]. Para algoritmos supervisionados o aprendizado acontece por retropropagação do erro (*backpropagation*). Isto significa que ao estimar uma saída, o erro entre o valor previsto e o observado é calculado utilizando uma função de perda (*loss*), como a função de erro quadrático médio (MSE) para problemas de regressão. Este erro é propagado de volta para as camadas ocultas, camada por camada, calculando a contribuição de cada peso no erro total. Isso é feito usando o gradiente descendente, que ajusta os pesos e os vieses da rede com o intuito de minimizar o erro [22].

O processo de estimativa no *machine learning* ocorre em duas etapas distintas. Na fase de treinamento, o algoritmo é alimentado com os dados de entrada junto com as saídas esperadas. Já na fase de testes, apenas os dados de entrada são fornecidos, cabendo ao algoritmo produzir saídas, com base nos padrões aprendidos durante cada iteração temporal (época) do treinamento. Em ambas etapas, o desempenho pode ser avaliado por meio da análise de métricas,

com o objetivo de verificar a precisão do algoritmo e evitar problemas como *overfitting* e *underfitting*.

O *underfitting* (subajuste), Figura 2a, ocorre quando um modelo de rede neural é muito simples para capturar a complexidade dos dados resultando em um desempenho ruim tanto nos dados de treinamento quanto nos dados de teste. Um modelo subajustado pode ser o resultado de uma baixa complexidade do modelo, hiperparâmetros mal ajustados ou dados de treinamento insuficientes para representar a complexidade do problema [23].

Por outro lado, o *overfitting* (sobreajuste), Figura 2b, ocorre quando um modelo de rede neural se ajusta excessivamente aos dados de treinamento, capturando não apenas os padrões, mas também o ruído nos dados. Isso geralmente resulta em um desempenho excelente nos dados de treinamento, mas um desempenho ruim em dados de teste [24]. Um modelo superajustado pode ser muito complexo em relação à quantidade de dados disponíveis, o que pode levar a uma memorização dos exemplos de treinamento em vez de aprender os padrões que generalizariam bem para novos dados [25].

O *right fit* (ajuste adequado), Figura 2c, é o modelo que melhor representa um bom ajuste tanto na fase de treinamento quanto na fase de teste. Este modelo indica que a rede neural capturou a relação dos dados durante o treinamento de forma satisfatória e que também pode generalizar bem para novos dados.

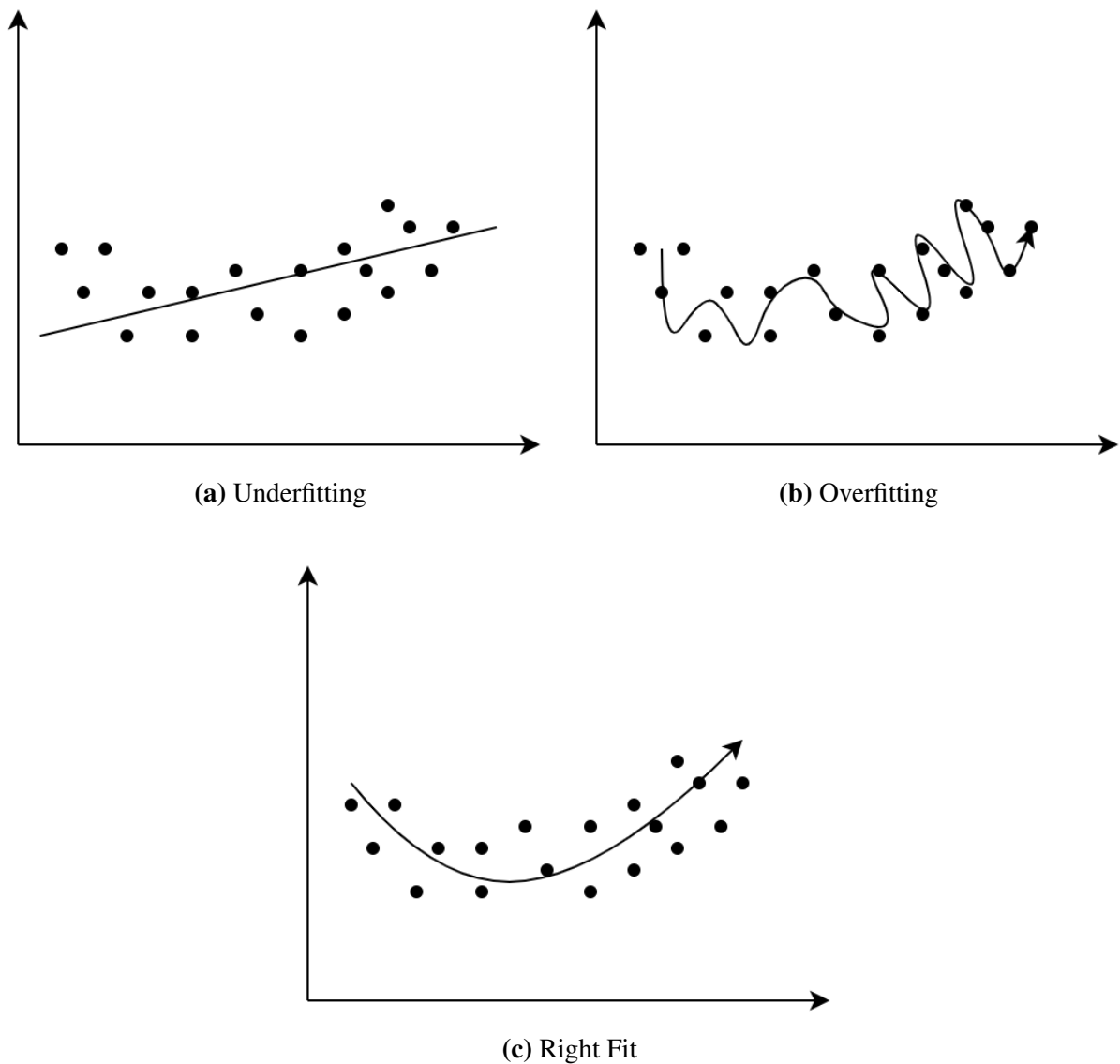


Figura 2: Representação dos possíveis ajustes de modelo: Underfitting (subajuste), Overfitting (sobreajuste) e Right fit (Ajuste Adequado)

Fonte: Autor, 2024

Durante o treinamento da rede neural, ajustes nos hiperparâmetros podem ser necessários para aprimorar o desempenho do modelo. Neste estudo, os ajustes foram realizados principalmente nos hiperparâmetros básicos, como o *batch_size* (tamanho dos lotes de dados) e *epochs* (número de vezes que o conjunto de dados, completo ou dividido em lotes, passa pela rede neural durante o treinamento [22]). Esses ajustes podem ser feitos manualmente, com base na análise das métricas de desempenho, alterando os valores dessas variáveis conforme necessário.

Em situações mais complexas, é possível utilizar uma ferramenta chamada *GridSearch*, incluída na biblioteca Scikit-learn [26]. O *GridSearch* permite especificar os hiperparâmetros que se deseja ajustar e os valores possíveis para cada um deles. A ferramenta então testa todas

as combinações possíveis desses hiperparâmetros e informa qual conjunto de valores produz o melhor desempenho. Isso simplifica o processo de otimização dos hiperparâmetros, especialmente em casos onde há múltiplos parâmetros a serem ajustados e suas interações precisam ser avaliadas.

O desenvolvimento do algoritmo foi realizado em *Python*, utilizando a ferramenta *Jupyter Notebook*. As principais bibliotecas utilizadas incluíram *pandas*, *numpy*, *matplotlib*, *datetime*, *tensorflow*, *scikit-learn*, *parametrics* e *scipy*, conforme mostrado na Tabela 1

Tabela 1: Descrição das bibliotecas *Python*.

Biblioteca	Descrição
<i>Pandas</i>	Manipulação de dados tabulares.
<i>NumPy</i>	Manipulação de arrays incluindo operações matemáticas e estatística básica.
<i>Matplotlib</i>	Visualização de dados.
<i>Datetime</i>	Manipulação de datas e horários.
<i>TensorFlow</i>	Desenvolvimento de modelos de aprendizado de máquina e <i>Deep Learning</i> .
<i>Scikit-learn</i>	Ferramenta eficiente para previsão e análise de dados.
<i>Parametrics</i>	Ferramenta com funções estatísticas para análise de dados.
<i>Scipy</i>	Ferramenta com funções estatísticas para análise de dados.

2.3.1 Tratamento de Dados

Os dados foram pré-processados e preparados para aplicação de técnicas de *machine learning*. Inicialmente, os conjuntos de dados foram manipulados utilizando a biblioteca *pandas* do *Python*, estruturando-os em *DataFrames*. Posteriormente, foi realizada a limpeza dos dados, onde linhas contendo valores em branco ou nulos foram removidas utilizando a função `dropna()`, e a coluna de datas foi definida como índice do *DataFrame* utilizando a função `set_index()`.

Como as variáveis independentes e dependentes tanto dos conjuntos de dados de Sinop quanto de Alta Floresta estavam em escalas diferentes, foi aplicada uma técnica de normalização chamada `MinMaxScaler()` [27]. Essa técnica garante que todos os dados estejam na mesma escala, facilitando o processo de treinamento dos modelos de *machine learning*. A equação para o `MinMaxScaler()` é dada por:

$$X_{\text{scaled}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}, \quad (2.1)$$

onde X é o valor original, $X_{\min}$ é o valor mínimo observado, $X_{\max}$ é o valor máximo observado,

e X_{scaled} é o valor escalonado.

Além disso, foi necessário ajustar a granularidade temporal dos dados, uma vez que as leituras em Sinop foram realizadas com intervalo de meia hora, enquanto em Alta Floresta, a cada hora. Para isso, foi empregada uma técnica de agrupamento de dados, onde as leituras foram agrupadas em intervalos específicos. O tamanho desses intervalos foi definido pela variável n_size , que recebeu os valores 48 e 24 para os algoritmos de Sinop e Alta Floresta, respectivamente. A entrada de dados na rede neural de forma agrupada serviu para reforçar o aprendizado.

Os dados foram então divididos em conjuntos de treinamento e teste utilizando a função `train_test_split` da biblioteca *scikit-learn* [28] utilizando o hiperparâmetro `train_size=0.8`, o que indica que 80% os dados foram utilizados para treinamento, e consequentemente 20% para teste, como pode ser visto nas Figuras 3 e 4. Importante ressaltar que, para respeitar a natureza de séries temporais dos dados, optou-se por não embaralhar os dados durante essa divisão, utilizando o hiperparâmetro `shuffle=False`. O não embaralhamento dos dados tem por objetivo respeitar o conceito de séries temporais que se referem a conjuntos de dados organizados cronologicamente, onde as observações passadas são utilizadas para prever eventos futuros.

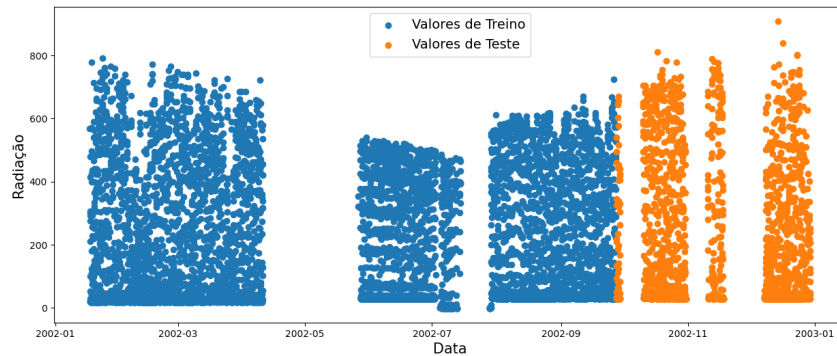


Figura 3: Resultado da separação dos dados de Sinop para treinamento e teste.
Fonte: Autor, 2024

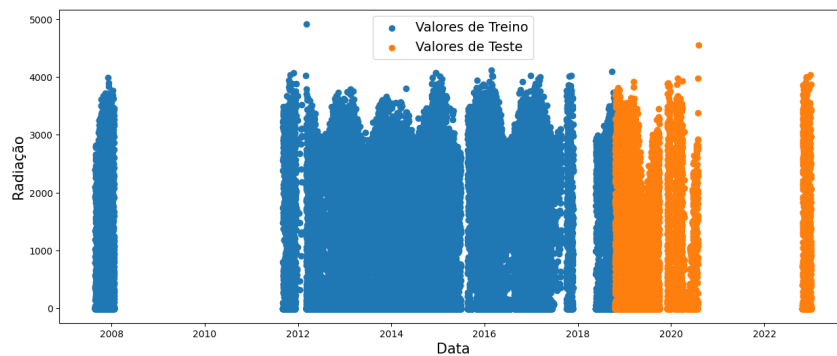


Figura 4: Resultado da separação dos dados de Alta Floresta para treinamento e teste.
Fonte: Autor, 2024

2.3.2 Redes Neurais

Redes neurais artificiais são baseadas no funcionamento do córtex cerebral de mamíferos [29], e podem ser compreendidas como uma coleção de unidades interconectadas entre si que processam as informações de entradas para gerar saídas [30]. São organizadas em camadas de entrada, camadas intermediárias (também chamadas de camadas ocultas) e uma camada de saída. Para cada conexão entre os neurônios há pesos associados [29] que determinam a força dessa conexão entre eles. Durante a fase de treinamento, a rede ajusta os pesos para que as saídas produzidas pela rede se aproximem o máximo possível das saídas desejadas.

Os modelos de *machine learning* foram treinados utilizando redes neurais *Gated Recurrent Unit* (GRU) e *Multi-Layer Perceptron* (MLP), aplicados aos conjuntos de dados de Sinop e Alta Floresta.

2.3.3 Multi-Layer Perceptron

Redes neurais Artificiais (RNA), tem por função receber os dados de entrada, processá-los e gerar uma saída. *Perceptron*, proposto por Rosenblatt (1958) [31] é um dos modelos de rede neural mais antigos e simples, constituída de uma única camada de neurônios onde cada neurônio está conectado a todas as entradas. Neste trabalho, foi utilizado uma rede neural um pouco mais complexa, chamada *Multi-Layer Perceptron* (MLP). A MLP é um tipo de rede neural, composta por 3 tipos de camadas de neurônios densamente conectados: entrada, oculta e saída. A camada de entrada responsável por receber os dados de entrada. As camadas ocultas que são responsáveis por processar, extrair as características e a relação entre os dados. A camada de saída é responsável por apresentar o resultado do processamento. A MLP possui uma característica chamada *feedforward* que significa que no fluxo de processamento dos dados partem da camada de entrada, passando pelas camadas ocultas até chegarem na camada de saída sem qualquer tipo de recursividade ou retorno de dados de saída para camadas anteriores. O ajuste dos pesos da rede se dá através do algoritmo de *backpropagation* [32].

Para entrada da dados na rede MLP, foi empregada uma camada de aplanamento de dados conhecida como “*Flatten*”. Essa camada tem a função de converter dados multidimensionais (matrizes) em unidimensionais (vetores). Sua principal finalidade é aplanar os dados de entrada, garantindo compatibilidade com as camadas subsequentes da rede sem alterar seus valores, apenas reorganizando sua estrutura possibilitando a captura de relacionamentos complexos entre os dados de entrada (Figura 5).

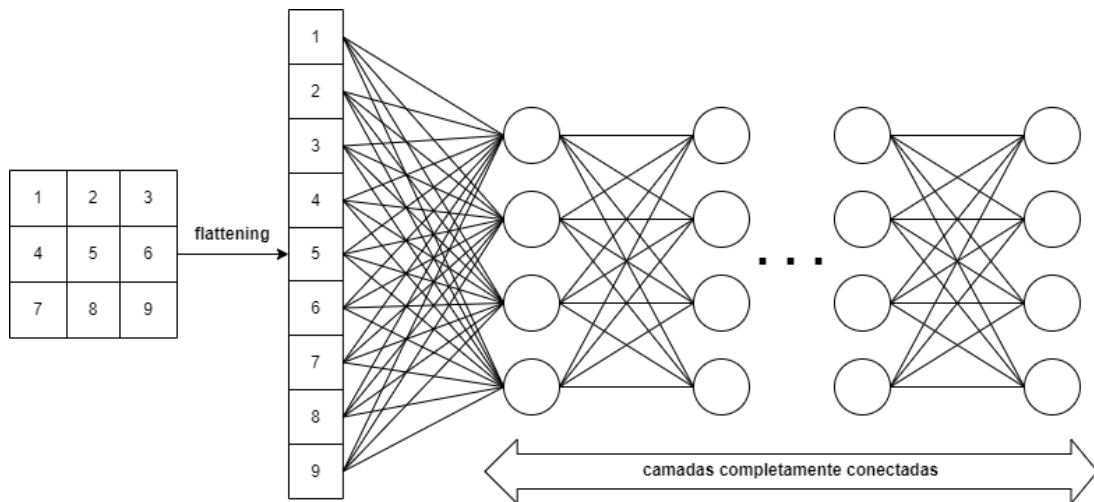


Figura 5: Achatamento (flattening) dos dados para entrada na rede neural.

Fonte: Autor, 2024

Na rede neural apresentada abaixo, há um total de 9 camadas. A primeira é do tipo *Input*. Sua função é especificar o formato dos dados de entrada que serão alimentados na rede neural. Em seguida, há uma camada “*Flatten*” e três camadas “*Dense*”, incluindo a camada de saída. Entre as camadas “*Dense*”, foram incluídas camadas de “*Dropout*” com uma taxa de 20%. A camada de “*Dropout*”, segundo Srivastava, Hinton, Krizhevsky et al. [33] funciona desligando aleatoriamente um determinado número de neurônios durante o treinamento. Isso impede que os neurônios se tornem excessivamente dependentes uns dos outros, tornando a rede neural mais resiliente e capaz de generalizar melhor para novos exemplos, reduzindo assim o risco de *overfitting*.

```
keras.layers.Input(shape=(48,4)),
keras.layers.Flatten(), # Aplainar a entrada
keras.layers.Dense(200, activation='relu'),
tf.keras.layers.Dropout(0.2),
keras.layers.Dense(200, activation='relu'),
tf.keras.layers.Dropout(0.2),
keras.layers.Dense(200, activation='relu'),
tf.keras.layers.Dropout(0.2),
keras.layers.Dense(48)
```

2.3.4 Gated Recurrent Unit (GRU)

GRU (Gated Recurrent Unit) se refere a um tipo de rede neural chamada de recorrente. Redes neurais do tipo recorrente são redes projetadas para lidar com dados sequenciais, como por

exemplo, dados de séries temporais, sequência de eventos entre outros [34]. As redes neurais recorrentes têm estruturas recursivas o que permite armazenar informações de etapas anteriores do processamento. Isso significa que a saída da etapa $t-1$ é realimentada na rede para influenciar o resultado da etapa t e assim por diante para cada etapa subsequente. Portanto, esse tipo de rede tem obtido sucesso no aprendizado de dados sequenciais [35].

A arquitetura GRU foi introduzida por Cho, Merriënboer, Gulcehre et al. [36] e apresentou um tipo de unidade oculta inovadora, inspirada nas camadas ocultas da *Long short-term memory* (LSTM), mas com uma implementação e processamento mais simplificados. A arquitetura LSTM proposta por Hochreiter e Schmidhuber (1997) [37] é constituída de 3 portões: (i) portão de entrada, que é responsável por atualizar o estado da célula. (ii) portão de esquecimento, que é responsável por decidir quais informações devem ser descartadas ou guardadas. (iii) e por fim temos o portão de saída, que decide quanto do estado da célula de memória deve ser enviado como saída. Já a arquitetura GRU implementa apenas dois portões, um portão de reinício e um portão de atualização, para os quais uma função logística sigmoide é aplicada,

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2.2)$$

recebendo como valores de entrada na função o estado oculto anterior (h_{t-1}) e à entrada no tempo atual (Figura 6). Quando o portão de reinício é fechado (valor 0), o estado oculto é forçado a ignorar completamente o estado oculto anterior, reiniciando-se apenas com a entrada atual. Isso efetivamente permite que o estado oculto descarte informações consideradas irrelevantes para futuras etapas, resultando em uma representação mais compacta. Por outro lado, o portão de atualização regula a quantidade de informação que será transferida de um estado oculto para o estado oculto posterior [36].

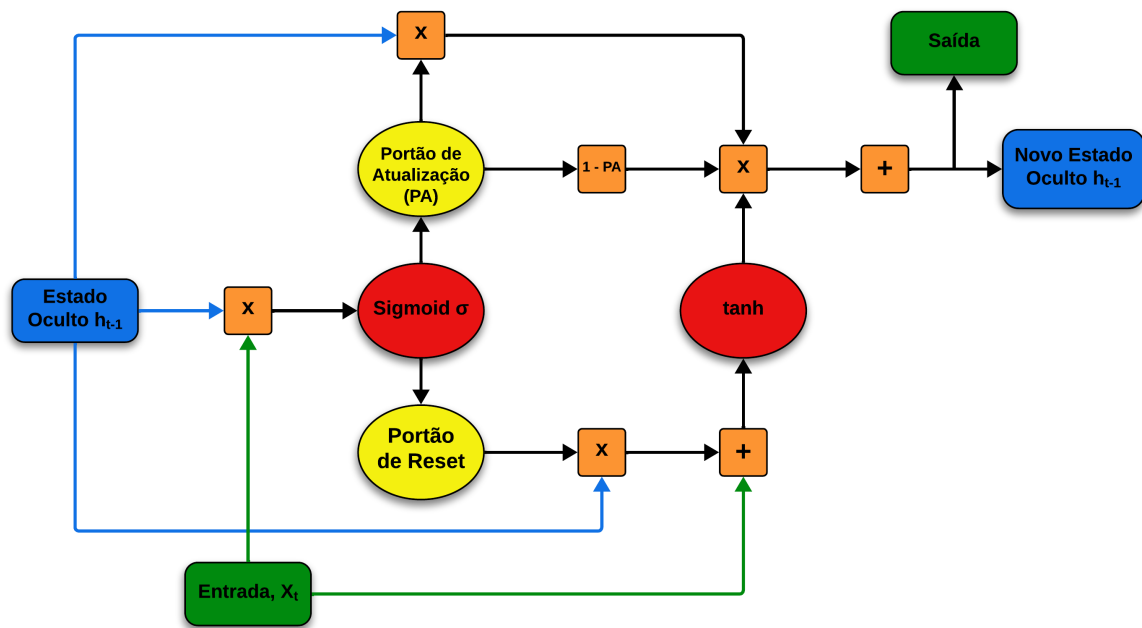


Figura 6: Arquitetura *GRU*.

Fonte: Autor, 2024

A rede neural, como pode ser visto abaixo, possui ao todo 9 camadas, sendo 5 camadas do tipo *GRU* e 1 camada de saída do tipo *Dense*. Entre as camadas do tipo *GRU* foram utilizadas camadas do tipo *Dropout* com taxa de 20%.

```
tf.keras.layers.GRU(200, return_sequences = True,
    input_shape = (48, 4)),
tf.keras.layers.Dropout(0.2),
tf.keras.layers.GRU(200, return_sequences=True),
tf.keras.layers.Dropout(0.2),
tf.keras.layers.GRU(200, return_sequences=True),
tf.keras.layers.Dropout(0.2),
tf.keras.layers.GRU(200),
tf.keras.layers.Dropout(0.2),
tf.keras.layers.Dense(48)
```

2.4 Métricas

A avaliação de modelos de *machine learning* é uma etapa fundamental no processo de criação e treinamento das redes neurais. A precisão das estimativas ou classificações é es-

sencial para garantir o desempenho satisfatório do modelo, quando aplicado a dados reais. O processo de refinamento e otimização dos algoritmos de aprendizado de máquina frequentemente depende do resultado apresentado nas métricas avaliadas.

Durante o treinamento de redes neurais, a “perda”, ou “*loss*”, é uma medida que quantifica o quão boa é a performance do modelo. A perda é uma função que calcula a diferença entre as saídas previstas pelo modelo e as saídas reais nos dados de treinamento [38], como pode ser visto na Figura 7. É possível informar qual é a métrica que o algoritmo deverá utilizar para calcular a perda durante o processo de treinamento da rede neural para realizar o ajustes dos pesos.

No exemplo abaixo, temos um trecho de código utilizado para compilar a rede neural. Nele vemos dois hiperparâmetros “*loss*” e “*metrics*”. Ambas os hiperparâmetros são usados para avaliar o desempenho do modelo durante o treinamento e/ou avaliação, mas somente o “*loss*” será usado diretamente no processo de otimização através de algoritmo de retropropagação.

```
model.compile(optimizer='adam', loss='mse', metrics=['mse', 'mae'])
```

Abaixo temos a fórmula que representa o erro quadrático médio (MSE), onde n é o número de exemplos de treinamento, y_i é a saída real e $\hat{y}_i$ é a saída prevista para o exemplo i .

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (2.3)$$

Para este estudo também foram utilizadas as seguintes métricas para avaliação dos testes: Erro Médio Absoluto (*MAE*), Raíz do Erro Médio Quadrático (*RMSE*), coeficiente de determinação (R^2), índice de *Willmott*, coeficiente de correlação de *Spearman* e coeficiente de correlação de *Pearson*.

1. MAE (Erro Médio Absoluto):

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (2.4)$$

2. RMSE (Raiz do Erro Médio Quadrático):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2.5)$$

3. Coeficiente de Determinação (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (2.6)$$

4. Índice de Willmott:

$$d_w = 1 - \frac{\sum_{i=1}^n (|\hat{y}_i - y_i|)}{\sum_{i=1}^n (|\hat{y}_i - \bar{y}| + |y_i - \bar{y}|)} \quad (2.7)$$

5. Coeficiente de Correlação de Spearman:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (2.8)$$

6. Coeficiente de Correlação de Pearson:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (2.9)$$

onde, y_i e $\hat{y}_i$ são valores observados e previstos, $\bar{y}$ representa a média dos valores observados, d_i é a diferenças nas posições das observações nos rankings de duas variáveis e n é o total de número de observações.

O RMSE e o MAE são métricas amplamente utilizadas em avaliações de modelos e que quantificam a discrepância entre os valores previstos pelo modelo e os valores observados. Quanto menores forem os resultados melhor é o modelo pois indica um melhor ajuste para os dados [39]. Já o R^2 é a relação entre a variação explicada pelo modelo linear e variação total [40]. Quanto mais próximo de 1 o R^2 , melhor o ajuste do modelo aos dados, indicando que uma quantidade significativa da variabilidade nos dados é capturada pelas variáveis independentes do modelo. Por outro lado, um R^2 próximo a 0 sugere que o modelo não é capaz de explicar a variabilidade dos dados de forma eficaz. O índice de *Willmott* indica uma correspondência perfeita entre o modelo e os dados observados para valores próximos a 1, enquanto valores mais baixos, próximos a -1, indicam falta de correspondência entre os valores observados e previstos [41]. O coeficiente de correlação de *Spearman*, por sua vez, avalia o quão bem a relação entre as variáveis pode ser descrita por uma função monotonicamente crescente ou decrescente, ou seja, sem que necessariamente haja uma relação linear entre elas. Um coeficiente de *Spearman* próximo a 1 indica uma forte correlação crescente entre as variáveis previstas e observadas, enquanto valores próximos a -1 indicam uma forte correlação decrescente. Já um coeficiente próximo a 0 sugere ausência de correlação entre as variáveis [42]. Por último, o coeficiente de correlação de *Pearson* é um indicador muito comum de interdependência entre as variáveis e é frequentemente usada para determinar como duas variáveis estão relacionadas linearmente uma com a outra [40].

3 RESULTADOS E DISCUSSÕES

O objeto de estudo é o desenvolvimento de um modelo para realizar a estimativa de radiação solar utilizando outras variáveis ambientais disponíveis como entrada. Para esse fim, utilizamos redes neurais como a principal ferramenta.

O conjunto de dados de Sinop possui como variáveis ambientais preditoras a umidade relativa do ar a 36 metros, a temperatura do ar a 40 metros, a umidade do solo e a precipitação, tendo como variável alvo a radiação solar direta. Já o conjunto de dados de Alta floresta possui como variáveis ambientais a temperatura, umidade relativa do ar, precipitação e velocidade do vento, tendo como variável alvo a radiação solar líquida. A Tabela 2 demonstra os hiperparâmetros utilizados, e os valores atribuídos a eles, para o treinamento de ambas redes neurais nos conjuntos de dados de Sinop e Alta Floresta.

Tabela 2: Hiperparâmetros de treinamento das redes neurais.

Cidade	Rede Neural	Intervalo	n_size	epochs	batch_size
Sinop	MLP	meia hora	48	300	300
	GRU	meia hora	48	150	300
Alta Floresta	MLP	horário	24	200	400
	GRU	horário	24	50	300

Os hiperparâmetros foram sendo ajustados baseando-se nos resultados das métricas de avaliação apresentadas até atingir um ótimo nível de refinamento. As métricas finais podem ser vistas nas Tabelas 11 e 12 mais abaixo.

3.1 Seleção de Variáveis

A seleção de características (ou variáveis) em um modelo de rede neural é importante para o desempenho do aprendizado de máquina, visto que minimiza o tempo de processamento, reduz a complexidade computacional e a chance de ocorrer overfitting. Para isso, foi utilizada a correlação de Pearson, que, como mencionada na sessão 2.4, indica a correlação linear entre as variáveis, onde 0 indica ausência de correlação, e valores próximos de 1 e -1 indicam correlações positivas e negativas perfeitas, respectivamente.

Para a estimativa de radiação solar direta utilizando o conjunto de dados de Sinop além das variáveis de temperatura, precipitação e umidade do solo, a variável de umidade relativa (UR)

possui medição em duas alturas diferentes: uma a 36 metros e outra a 41 metros. Ambas as variáveis foram levadas em consideração no momento da aplicação do método de correlação de Pearson.

Como pode ser vista na Tabela 3 temos que a temperatura a 40 metros é a variável que melhor apresenta correlação positiva enquanto a umidade relativa a 36 metros apresenta melhor correlação negativa, ainda que muito próxima do valor obtido para a umidade relativa a 41 metros.

Tabela 3: Resultado da aplicação da correlação de Pearson entre variáveis do conjunto de dados de Sinop.

	T (40m)	UR (36m)	UR 41 (m)	ppt (mm)	Soil H ₂ O m ³ /m ³
Pearson	0.6443	-0.6652	-0.6407	-0.0444	-0.0354

Para o conjunto de Alta Floresta, conforme Tabela 4, a correlação de Pearson indica que as variáveis com maiores correlações positivas e negativas são a temperatura e umidade relativa respectivamente.

Tabela 4: Resultado da aplicação da correlação de Pearson entre variáveis do conjunto de dados de Alta Floresta.

	T	UR	PPT	WIND
Pearson	0.8000	-0.7540	-0.0730	0.5580

3.2 Resultados Gerais

Considerando os resultados apresentados na subseção 3.1, foram efetuados testes com os conjuntos de variáveis com o intuito de verificar o desempenho dos modelos mediante a redução do número de variáveis. As Tabelas 5 e 6 representam os resultados obtidos para o conjunto de Sinop aplicado as redes MLP e GRU. Já as Tabelas 7 e 8 representam os resultados obtidos para o conjunto de Alta Floresta aplicado as redes MLP e GRU.

Tabela 5: Métricas de avaliação de desempenho da rede neural MLP aplicadas aos dados de Sinop.

Dados	Variáveis	MAE	RMSE	R ²	Willmott	Spearman	Pearson
Teste	T (40m) e RH (36m)	47,9272	76,2542	0,8721	0,9642	0,8612	0,9341
Meia hora	T (40m) e RH (36m)	30,7912	55,3336	0,9226	0,9798	0,9064	0,9607
Média horária	T (40m) e RH (36m)	27,1193	47,9789	0,9402	0,9846	0,9139	0,9698
Média diária	T (40m) e RH (36m)	12,8658	18,1706	0,7805	0,9382	0,7752	0,8857
Média mensal	T (40m) e RH (36m)	5,9324	7,2551	0,7547	0,9229	0,8273	0,8739

Tabela 6: Métricas de avaliação de desempenho da rede neural GRU aplicadas aos dados de Sinop.

Dados	Variáveis	MAE	RMSE	R ²	Willmott	Spearman	Pearson
Teste	T (40m) e RH (36m)	47,6199	73,3688	0,8816	0,9668	0,8955	0,9398
Meia hora	T (40m) e RH (36m)	31,4415	53,8754	0,9266	0,9808	0,8993	0,9626
Média horária	T (40m) e RH (36m)	27,4966	46,5847	0,9436	0,9854	0,9062	0,9714
Média diária	T (40m) e RH (36m)	11,7018	16,7356	0,8138	0,9494	0,7944	0,9033
Média mensal	T (40m) e RH (36m)	6,8696	8,4270	0,6690	0,8972	0,7909	0,8213

Tabela 7: Métricas de avaliação de desempenho da rede neural MLP aplicadas aos dados de Alta Floresta.

Dados	Variáveis	MAE	RMSE	R ²	Willmott	Spearman	Pearson
Teste	T e RH	226,2296	414,0655	0,8161	0,9545	0,8907	0,9186
Horário	T e RH	158,1744	292,5767	0,9182	0,9782	0,8810	0,9583
Média diária	T e RH	82,5733	115,4103	0,7112	0,9098	0,7485	0,8460
Média mensal	T e RH	51,2487	86,0514	0,2830	0,7129	0,6412	0,5667

Tabela 8: Métricas de avaliação de desempenho da rede neural GRU aplicadas aos dados de Alta Floresta.

Dados	Variáveis	MAE	RMSE	R ²	Willmott	Spearman	Pearson
Teste	T e RH	208,6695	388,5672	0,8380	0,9600	0,8484	0,9274
Horário	T e RH	150,0404	274,9053	0,9278	0,9810	0,8664	0,9632
Média diária	T e RH	76,0182	109,6054	0,7395	0,9244	0,7807	0,8610
Média mensal	T e RH	53,8499	86,6100	0,2737	0,7341	0,6475	0,5666

Posteriormente, para o conjunto de Sinop, foram efetuados testes utilizando as redes neurais MLP e GRU, com as variáveis de temperatura, precipitação e umidade do solo. Desta vez, alternaram-se as variáveis de umidade relativa a 36 metros e 41 metros, a fim de verificar o desempenho da rede neural. Os resultados podem ser vistos nas Tabelas 9 e 10.

Analisando os resultados, observou-se que a umidade relativa a 36 metros apresenta uma vantagem, ainda que leve, no coeficiente de determinação, especialmente para a média mensal. Além disso, essa configuração resultou em menores erros (MAE e RMSE), maior concordância (índice de *Willmott*) e correlações mais fortes (*Spearman* e *Pearson*).

Tabela 9: Métricas de avaliação de desempenho das redes neurais MLP e GRU aplicadas aos dados de Sinop utilizando umidade relativa a 36 metros de altura

Rede Neural	Dados	MAE	RMSE	R ²	Willmott	Spearman	Pearson
MLP	Teste	47,9396	75,4575	0,8747	0,9629	0,8874	0,9398
	Meia hora	29,3190	51,7802	0,9322	0,9821	0,9262	0,9655
	Média horária	25,4358	43,8083	0,9501	0,9870	0,9323	0,9748
	Média diária	10,2446	13,3423	0,8817	0,9656	0,8927	0,9412
	Média mensal	4,0314	4,7691	0,8840	0,9672	0,8455	0,9565
GRU	Teste	43,7501	69,8079	0,8928	0,9696	0,8899	0,9466
	Meia hora	28,7518	48,9201	0,9395	0,9843	0,8979	0,9693
	Média horária	24,7892	40,8037	0,9568	0,9889	0,9058	0,9782
	Média diária	9,1217	12,1471	0,9019	0,9739	0,9108	0,9502
	Média mensal	4,2793	4,9892	0,8840	0,9650	0,8364	0,9477

Tabela 10: Métricas de avaliação de desempenho das redes neurais MLP e GRU aplicadas aos dados de Sinop utilizando umidade relativa a 41 metros de altura

Rede Neural	Dados	MAE	RMSE	R ²	Willmott	Spearman	Pearson
MLP	Teste	51,7625	79,2543	0,8618	0,9601	0,8777	0,9292
	Meia hora	30,4224	53,9677	0,9264	0,9801	0,9244	0,9628
	Média horária	26,5252	46,1491	0,9447	0,9853	0,9295	0,9722
	Média diária	9,9868	13,1374	0,8853	0,9672	0,8962	0,9435
	Média mensal	4,6360	5,9675	0,8340	0,9428	0,9000	0,9503
GRU	Teste	45,5786	72,4478	0,8845	0,9658	0,8812	0,9451
	Meia hora	28,9835	49,7762	0,9373	0,9832	0,9060	0,9685
	Média horária	25,0669	41,7982	0,9546	0,9880	0,9135	0,9774
	Média diária	9,0329	12,9693	0,8882	0,9699	0,8973	0,9431
	Média mensal	3,8917	5,7819	0,8442	0,9506	0,8182	0,9385

3.3 Sinop

Na Figura 7, o gráfico representa o resultado do treinamento da rede neural, porém nas Figuras 8 e 9 podemos observar o resultado do teste realizado utilizando os 20% restantes dos dados separados após o treinamento da rede neural MLP, aplicando a ela os dados de Sinop. Podemos observar que a rede neural teve um bom comportamento durante seu treinamento, atingindo um coeficiente de determinação (R²) de aproximadamente 87%, conforme pode ser visto na tabela 11 mais abaixo.

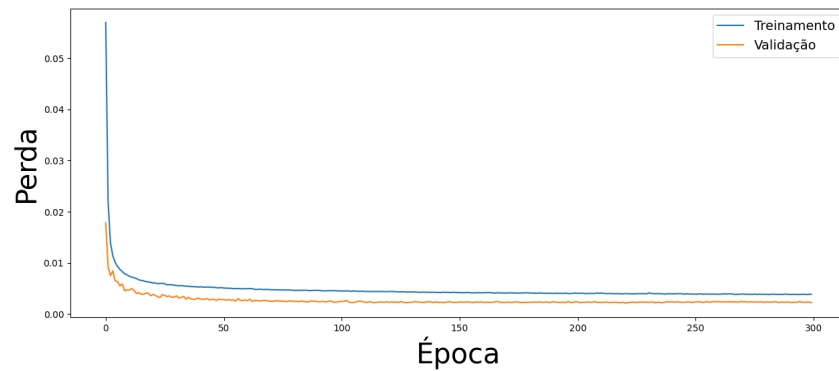


Figura 7: Gráfico do treinamento da rede neural MLP para os dados de Sinop.
Fonte: Autor, 2024

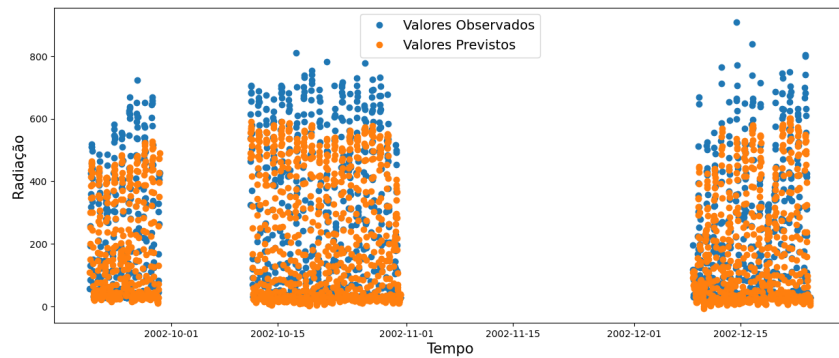


Figura 8: Resultado do teste da rede neural MLP para os dados de Sinop.
Fonte: Autor, 2024

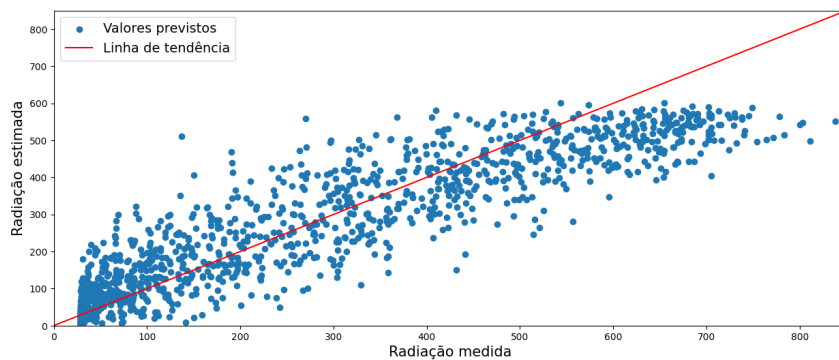


Figura 9: Gráficos da relação entre dados medidos e previstos para o teste utilizando MLP para os dados de Sinop.

Fonte: Autor, 2024

Após o treinamento da rede, durante a fase de validação do modelo, foram utilizados dados com diferentes intervalos temporais. A primeira leitura consistiu nos dados originais, coletados em intervalos de meia hora, seguida por médias horárias, médias diárias e, por fim, médias mensais. Na Figura 10, pode-se observar falhas nas leituras dos dados, resultando em uma

taxa de aproveitamento de apenas 64,1% após aplicado o tratamento de dados. No entanto, isso não impediu que o modelo demonstrasse um bom desempenho. O modelo mostrou-se bem ajustado, com forte correlação entre as variáveis, apresentando um R^2 superior a 88%, um índice de *Willmott* superior a 96% e um coeficiente de *Pearson* superior a 94%. O modelo apresentou para as variações temporais, taxa de RMSE abaixo do que foi apresentado no teste, indicando que o modelo realiza boas estimativas e não apenas uma simples estimativa média.

Na Figura 11, os gráficos de dispersão revelam uma tendência na relação entre os dados medidos e previstos, com os pontos agrupados ao redor da linha diagonal. Isso reforça a capacidade do modelo em capturar padrões de relacionamento.

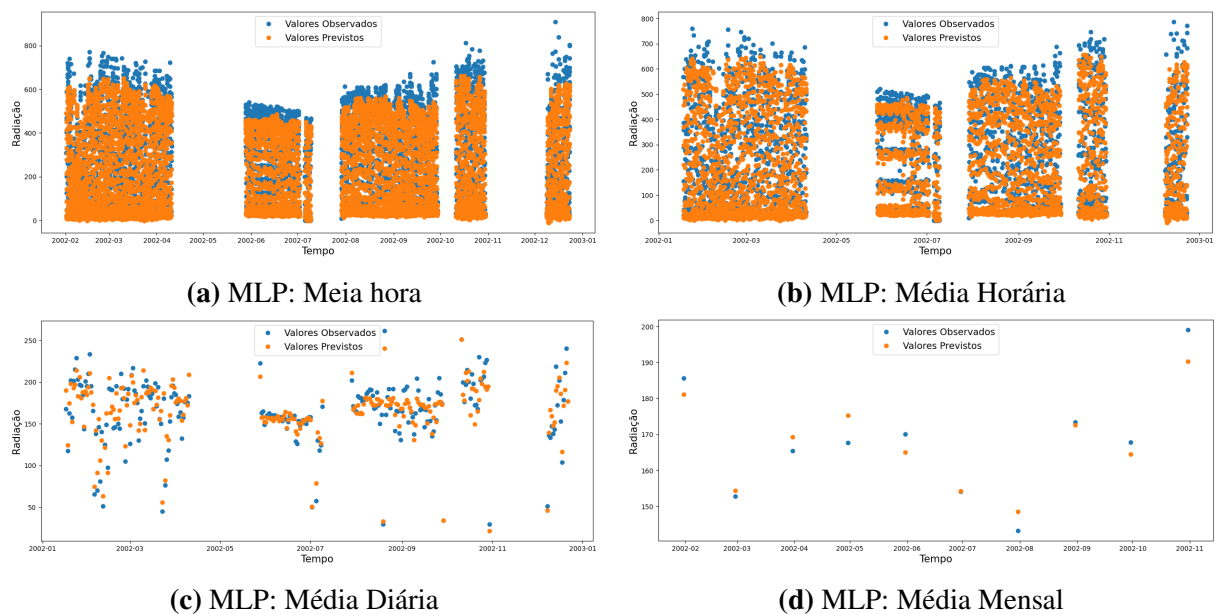


Figura 10: Resultado dos testes da rede neural MLP para os dados de Sinop em diferentes intervalos de tempo.

Fonte: Autor, 2024

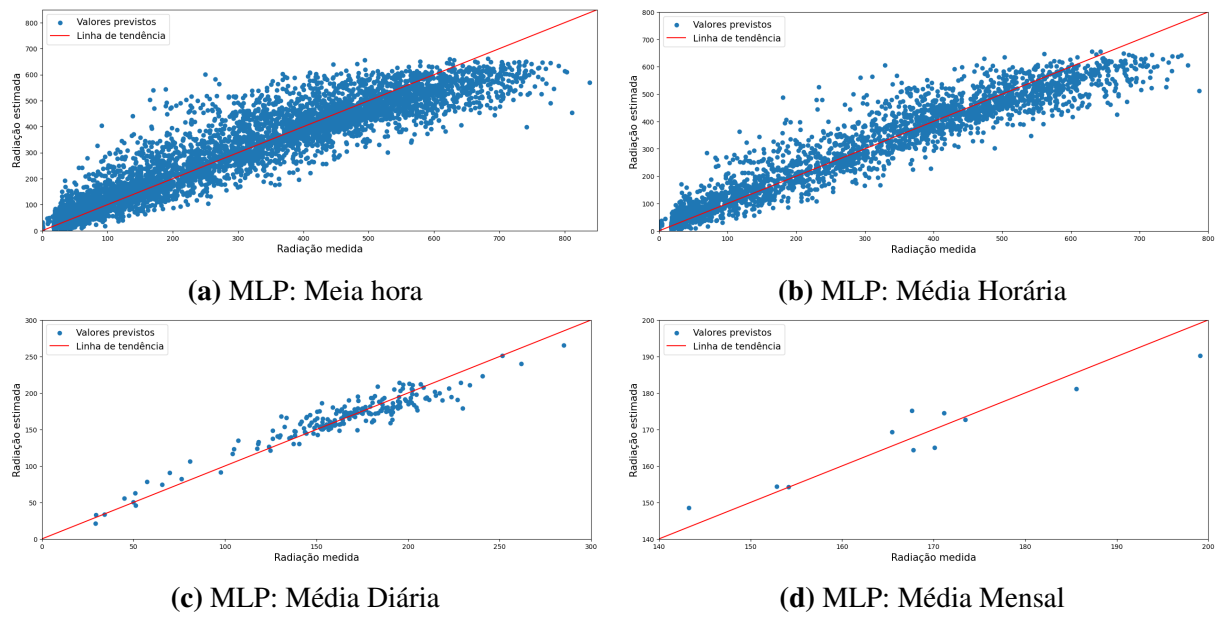


Figura 11: Gráficos da relação entre dados medidos e previstos utilizando MLP para os dados de Sinop em diferentes intervalos de tempo.

Fonte: Autor, 2024

A Figura 12 temos o gráfico que representa o resultado do treinamento da rede neural para a arquitetura *GRU*. As Figuras 13 e 14 demonstram que a rede neural provou ser ligeiramente melhor que a rede MLP, apresentando R^2 de 89%.

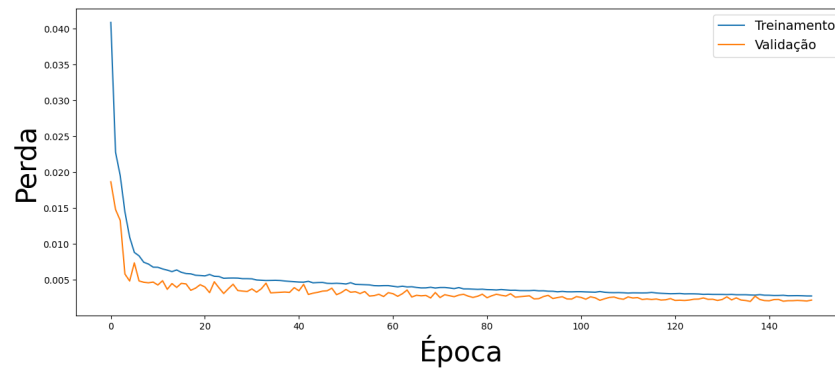


Figura 12: Gráfico de treinamento da rede neural GRU para os dados de Sinop.

Fonte: Autor, 2024

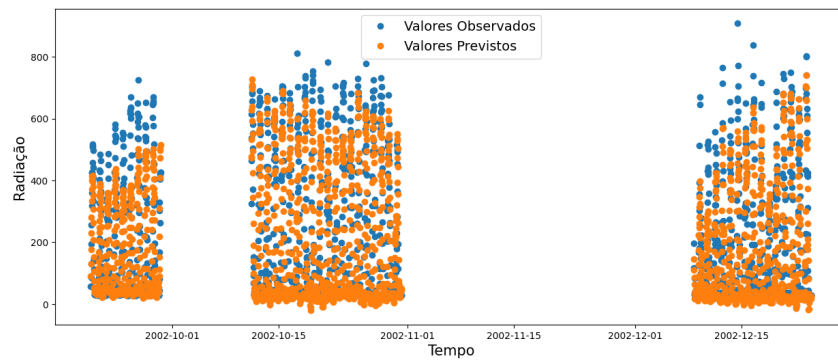


Figura 13: Resultado do teste da rede neural GRU para os dados de Sinop.

Fonte: Autor, 2024

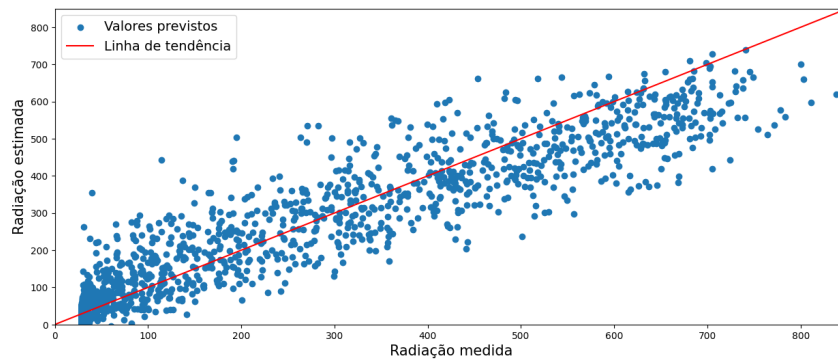


Figura 14: Gráficos da relação entre dados medidos e previstos para o teste utilizando GRU para os dados de Sinop.

Fonte: Autor, 2024

O modelo também mostrou-se bem ajustado, com forte correlação entre as variáveis, apresentando um R^2 superior a 88%, um índice de *Willmott* superior a 96% e um coeficiente de *Pearson* superior a 94%. Analisando as métricas resultantes das estimativas extraídas das variações temporais o modelo apresentou taxa de RMSE abaixo do que pode ser vista no teste, e ainda abaixo do que foi apresentado para a rede neural MLP, mostrando que realizou estimativas melhores em comparação com cada intervalo temporal.

A arquitetura GRU se apresentou ser mais eficiente na estimativa de valores máximos, conforme pode ser visto nas Figuras 15 e 10.

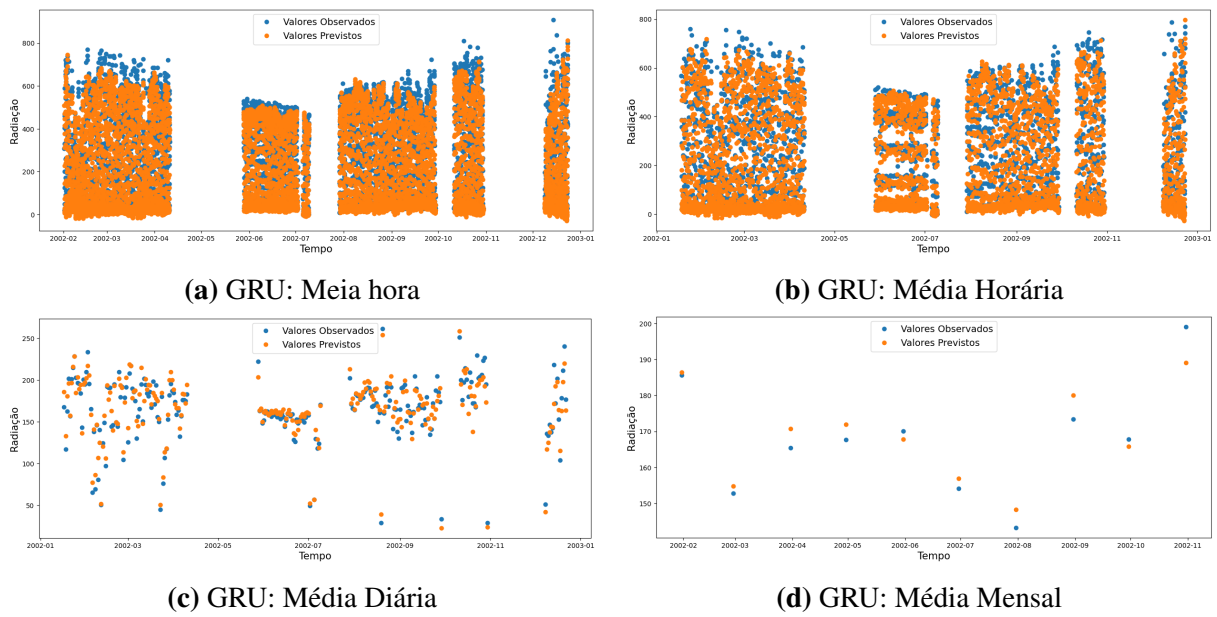


Figura 15: Resultado dos testes da rede neural GRU para os dados de Sinop em diferentes intervalos de tempo.

Fonte: Autor, 2024

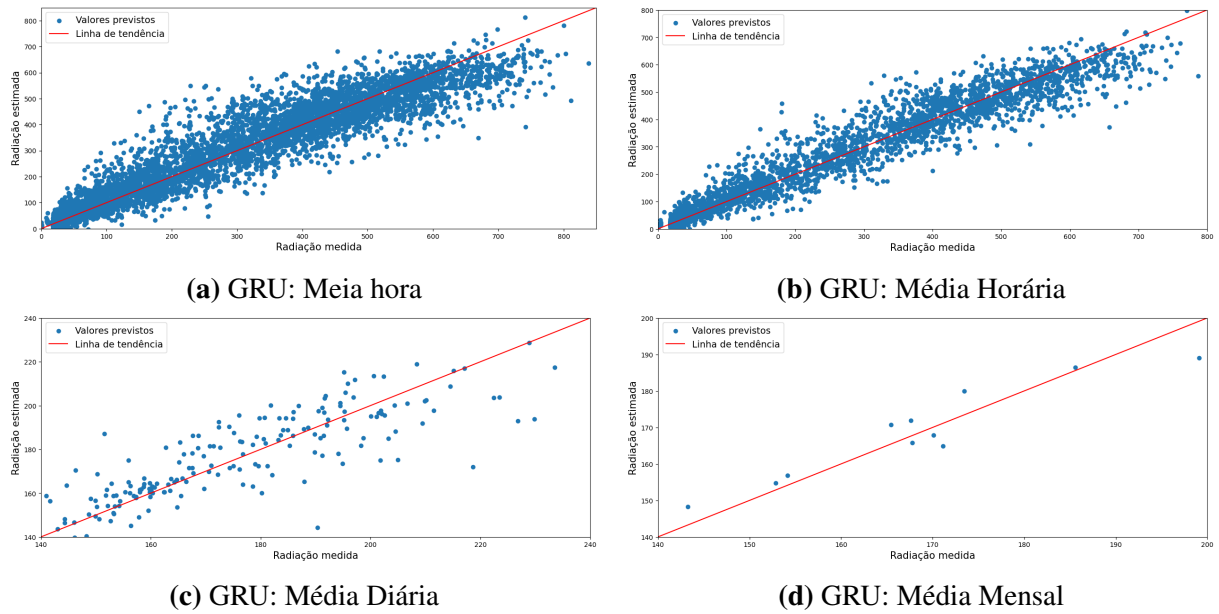


Figura 16: Gráficos da relação entre dados medidos e previstos utilizando GRU para os dados de Sinop em diferentes intervalos de tempo.

Fonte: Autor, 2024

Tabela 11: Métricas de avaliação de desempenho das redes neurais MLP e GRU aplicadas aos dados de Sinop.

Rede Neural	Dados	MAE	RMSE	R ²	Willmott	Spearman	Pearson
MLP	Teste	47,9396	75,4575	0,8747	0,9629	0,8874	0,9398
	Meia hora	29,3190	51,7802	0,9322	0,9821	0,9262	0,9655
	Média horária	25,4358	43,8083	0,9501	0,9870	0,9323	0,9748
	Média diária	10,2446	13,3423	0,8817	0,9656	0,8927	0,9412
	Média mensal	4,0314	4,7691	0,8840	0,9672	0,8455	0,9565
GRU	Teste	43,7501	69,8079	0,8928	0,9696	0,8899	0,9466
	Meia hora	28,7518	48,9201	0,9395	0,9843	0,8979	0,9693
	Média horária	24,7892	40,8037	0,9568	0,9889	0,9058	0,9782
	Média diária	9,1217	12,1471	0,9019	0,9739	0,9108	0,9502
	Média mensal	4,2793	4,9892	0,8840	0,9650	0,8364	0,9477

3.4 Alta Floresta

Considerando os resultados apresentados na Tabela 12 para os dados de Alta Floresta aplicadas a MLP, observamos um resultado de treinamento satisfatório, porém inferior quando se comparado ao conjunto de dados de Sinop.

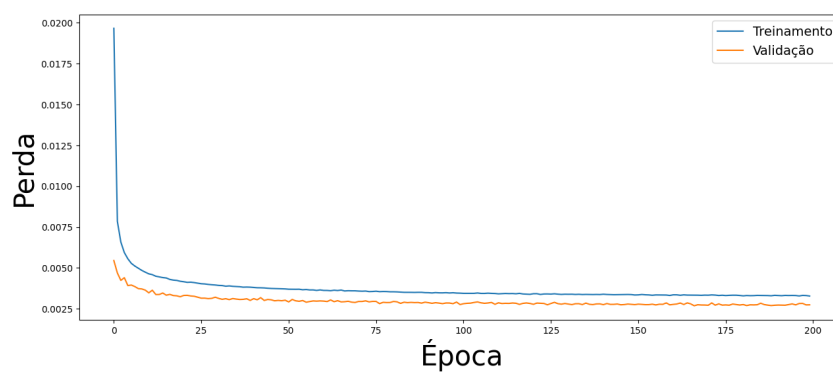


Figura 17: Gráfico do treinamento da rede neural MLP para os dados de Alta Floresta.
Fonte: Autor, 2024

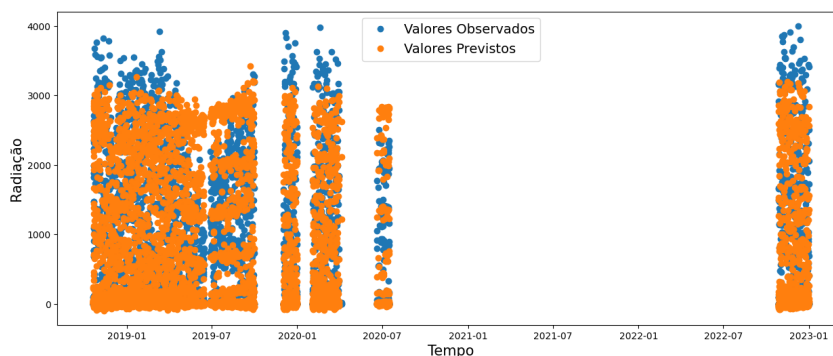


Figura 18: Resultado do teste da rede neural MLP para os dados de Alta Floresta.
Fonte: Autor, 2024

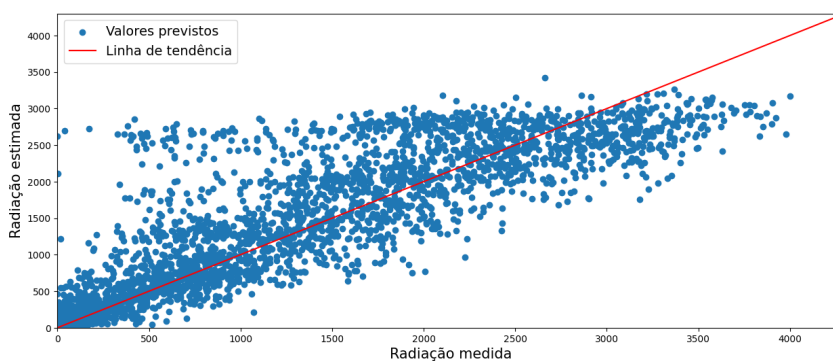
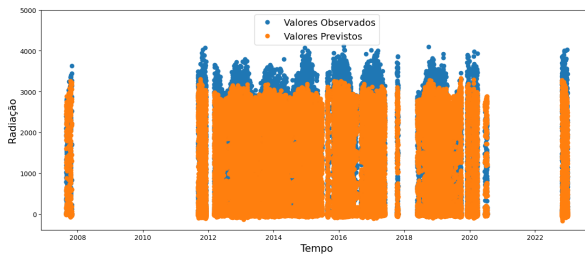


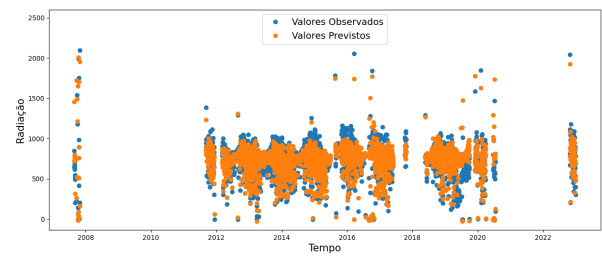
Figura 19: Gráficos da relação entre dados medidos e previstos para o teste utilizando MLP para os dados de Alta Floresta.

Fonte: Autor, 2024

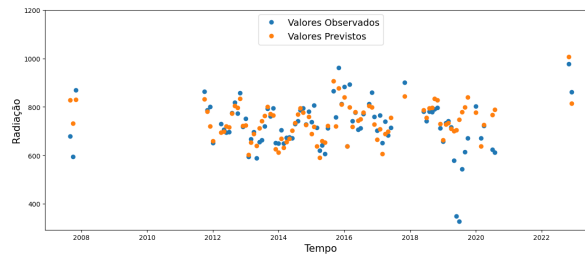
A taxa aprendizado para o intervalo horário apresentou R^2 de 92%, contra 84% da fase de teste, mostrando que o modelo generalizou bem para novos dados. O índice de *Willmott* e o coeficiente de *Pearson* apresentaram alto grau de correlação das variáveis, com 98% e 96% respectivamente. Porém há uma queda considerável na qualidade do modelo, tanto para médias com intervalo diário e mensal, mas principalmente para a última, indicando que os dados se ajustaram melhor ao treinamento, configurando a presença de *overfitting*.



(a) MLP: Horário



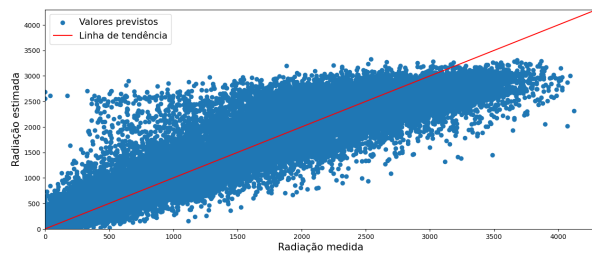
(b) MLP: Média Diária



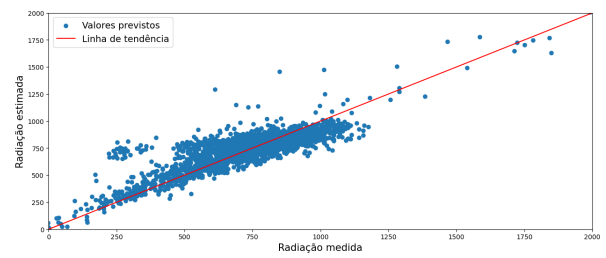
(c) MLP: Média Mensal

Figura 20: Resultado dos testes da rede neural MLP para os dados de Alta Floresta em diferentes intervalos de tempo.

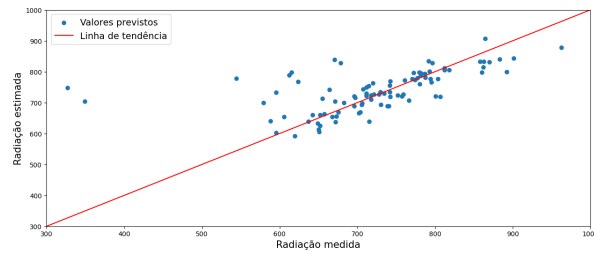
Fonte: Autor, 2024



(a) MLP: Horário



(b) MLP: Média Diária



(c) MLP: Média Mensal

Figura 21: Gráficos da relação entre dados medidos e previstos utilizando MLP para os dados de Alta Floresta em diferentes intervalos de tempo.

Fonte: Autor, 2024

Nota-se que a quantidade de épocas utilizadas para o treinamento da rede GRU foi de apenas 25% em relação a quantidade épocas utilizadas para MLP, como pode ser visto na Tabela 2. Na Figura 22 podemos ver que ao final da última época há uma tendência de que a linha de treino ultrapasse a linha de validação, caracterizando assim uma possibilidade de *overfitting*. Portanto fica evidente a importância do controle sobre os hiperparâmetros.

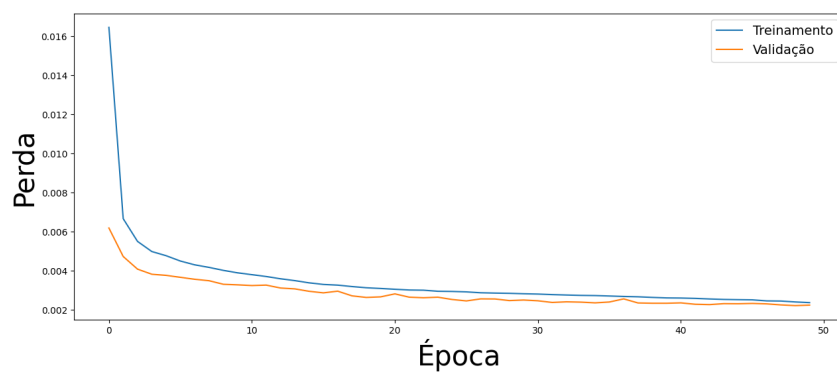


Figura 22: Gráfico do treinamento da rede neural GRU para os dados de Alta Floresta.

Fonte: Autor, 2024

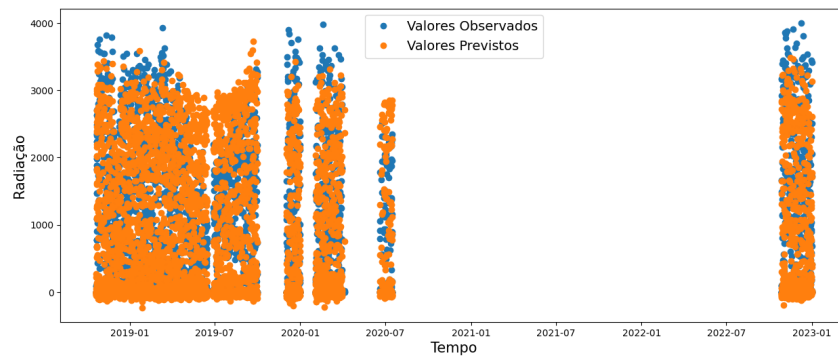


Figura 23: Resultado do teste da rede neural GRU para os dados de Alta Floresta.
Fonte: Autor, 2024

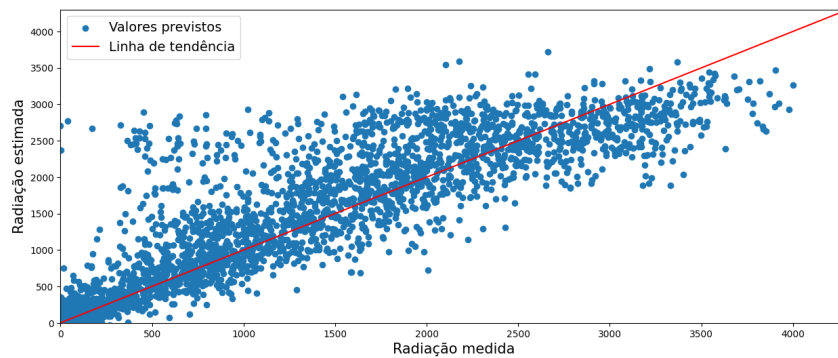


Figura 24: Gráficos da relação entre dados medidos e previstos para o teste utilizando GRU para os dados de Alta Floresta.

Fonte: Autor, 2024

Como pode ser visto na Figura 25a, para dados de intervalos horários, a rede neural obteve um bom desempenho, com R^2 de 93%, e índice de *Willmott* e coeficientes de *Separman* e *Pearson* de 98%, 87% e 96% respectivamente. Ainda se observa um comportamento melhor na estimativa de valores máximos em comparação a estimativa horária da rede MLP (Figura 20a).

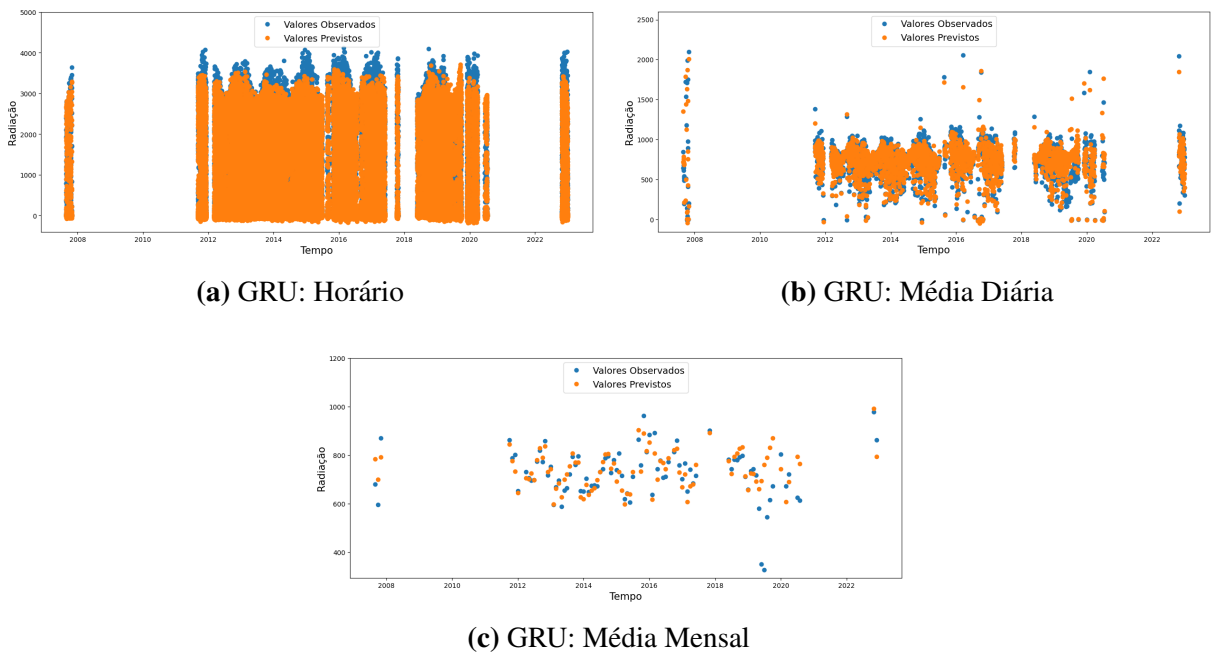


Figura 25: Resultado dos testes da rede neural GRU para os dados de Alta Floresta em diferentes intervalos de tempo.

Fonte: Autor, 2024

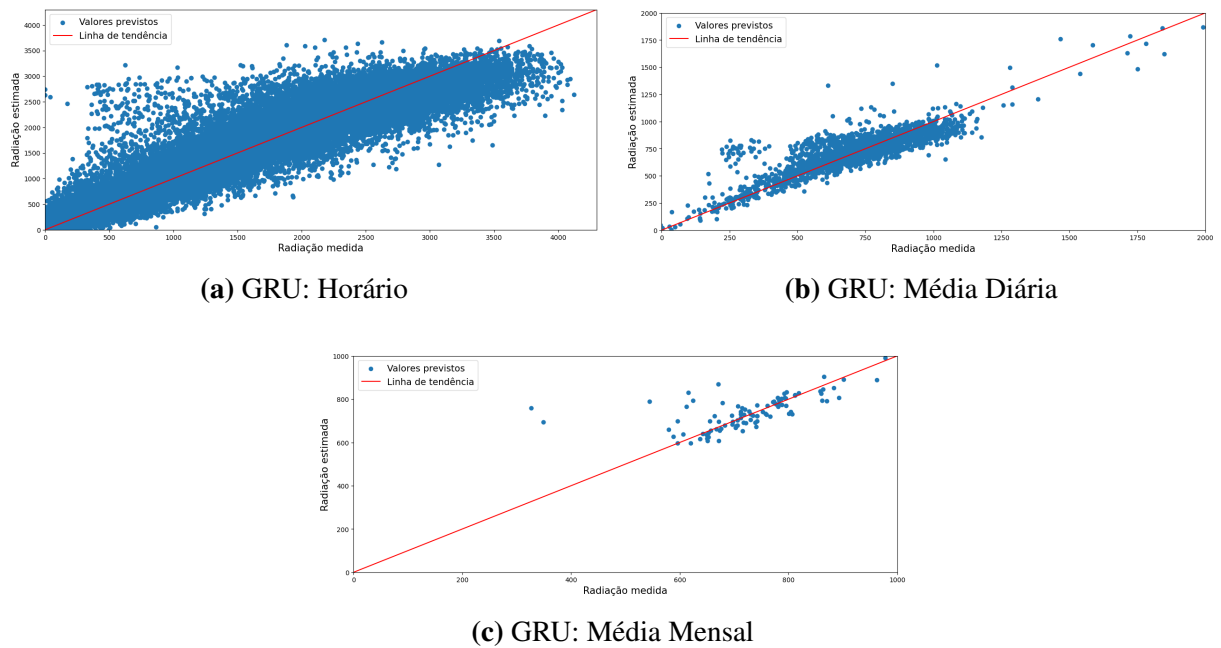


Figura 26: Gráficos da relação entre dados medidos e previstos utilizando GRU para os dados de Alta Floresta em diferentes intervalos de tempo.

Fonte: Autor, 2024

Porém, apesar do controle sobre as épocas de treinamento, o *overfitting* também se evidenciou neste modelo, principalmente para os intervalos diário e mensal, onde o R^2 caiu para 77% e 36% respectivamente, em comparação com a mesma métrica de teste (85%). Este declínio é

significativo. O índice de *Willmott*, e os coeficientes de *Spearman* e *Pearson* também apresentaram uma queda considerável.

Tabela 12: Métricas de avaliação de desempenho das redes neurais MLP e GRU aplicadas aos dados de Alta Floresta.

Rede Neural	Dados	MAE	RMSE	R ²	Willmott	Spearman	Pearson
MLP	Teste	197,3269	378,4203	0,8464	0,9611	0,8728	0,9268
	Horário	148,6143	277,7424	0,9263	0,9806	0,8847	0,9625
	Média diária	75,8355	108,2406	0,7459	0,9230	0,7895	0,8655
	Média mensal	47,3567	82,2839	0,3444	0,7502	0,6871	0,6134
GRU	Teste	188,0853	362,9829	0,8586	0,9654	0,8697	0,9356
	Horário	137,7382	254,6439	0,9380	0,9841	0,8715	0,9688
	Média diária	68,5605	102,2810	0,7731	0,9352	0,8181	0,8822
	Média mensal	44,3045	80,9808	0,3650	0,7727	0,7231	0,6388

4 CONCLUSÃO

Este estudo demonstrou o uso de dois tipos de modelos de redes neurais (GRU e MLP), com ajuste de parâmetros específicos e um tratamento padronizado para cada conjunto de dados. O objetivo foi alcançar o melhor desempenho possível e evitar problemas de estimativa, como o *overfitting* e o *underfitting*.

Os resultados apresentados após a redução de variáveis se apresentaram eficazes, principalmente se considerarmos os resultados apresentados nas Tabelas 5 e 6 para Sinop (testes de meia hora e média horária) e Tabelas 7 e 8 para Alta Floresta (testes horários). Porém podemos observar que conforme a variação temporal aumenta, os resultados decaem, podendo levar a conclusão, levando em consideração os resultados apresentados nas Tabelas 11 e 12, que a utilização de mais variáveis podem ajudar a melhorar o modelo.

Ainda considerando os resultados apresentados nas Tabelas 11 e 12, observamos que ambos modelos alcançaram resultados satisfatórios, principalmente para médias horárias. Em ambos conjuntos de dados, a rede neural baseada em dados sequenciais (GRU) apresentou um desempenho ligeiramente superior à rede MLP, tanto para a leitura original quanto para os demais intervalos por média simples. Portanto, não se pode descartar a utilização de uma rede não recorrente para a estimativa de dados temporais principalmente se formos levar em consideração que o tempo de processamento da rede MLP é menor que da rede GRU. Um comportamento padrão que pode ser percebido em ambos conjuntos de dados é que a eficiência das redes GRU e MLP diminuem à medida que se amplia o intervalo temporal para calcular as médias, conforme indicado pelas métricas de avaliação.

Para o conjunto de dados de Sinop, ambas redes apresentaram um coeficiente R^2 acima de 88%, indicando um bom ajuste dos dados. Além disso, os modelos mostraram uma forte correlação entre as variáveis, evidenciada pelas medidas estatísticas de *Spearman* e *Pearson*, com valores acima de 84% e 95% para MLP e 83% e 94% para GRU, respectivamente.

No conjunto de dados de Alta Floresta, considerando os dados horários, obtivemos R^2 de 92% para MLP e de 93% para GRU, indicando também um bom ajuste dos dados. A forte correlação entre as variáveis também foi evidenciada pelas correlações de *Spearman* e *Pearson*, com valores de 88% e 96% para MLP e 87% e 96% para GRU. No entanto, para médias com intervalos diários e mensais, as redes não alcançaram resultados expressivos. Pode-se dizer que as grandes quantidades de falhas de leitura (51.4%) impediram que as redes neurais capturassem de forma precisa os padrões, relações temporais ou mesmo correlações significativas entre as variáveis.

Com base nos resultados apresentados, podemos concluir que o uso dos modelos de redes neurais GRU e MLP demonstraram ser abordagens eficazes para a estimativa de dados temporais. Os ajustes personalizados dos parâmetros e o tratamento padronizado dos conjuntos de dados foram fundamentais para garantir o melhor desempenho dos modelos e evitar problemas como *overfitting* e *underfitting*. No entanto, é importante observar que, apesar dos resultados satisfatórios, os modelos podem enfrentar limitações em cenários onde os dados contenham uma quantidade significativa de falhas. Isso ressalta a importância de assegurar a qualidade e a integridade dos dados para garantir o bom desempenho dos modelos.

Embora este trabalho tenha contribuído significativamente para a estimativa de radiação solar por meio do uso de redes neurais, ainda há possíveis aprimoramentos a serem explorados. Uma melhoria a ser considerada é a adoção da rede GRU não apenas para prever a radiação no momento atual da leitura das variáveis preditoras, mas também para capturar e aprender os padrões de comportamento dessas variáveis, permitindo previsões para tempos futuros. Essa abordagem pode aumentar a robustez do modelo, oferecendo *insights* valiosos para aplicações práticas.

5 REVISÃO BIBLIOGRÁFICA

- [1] Badescu, V., **Modeling Solar Radiation at the Earth's Surface: Recent Advances**, p. 325, 2014. DOI: 10.1007/978-3-540-77455-6.
- [2] Rodríguez Gómez, J. M., Carlesso, F., Vieira, L. E. e Da Silva, L., **A irradiância solar: conceitos básicos**, *Rev. Bras. Ensino Fís.*, v. 40, n. 3, 2018. DOI: 10.1590/1806-9126-RBEF-2017-0342.
- [3] Domingos, S. F., Monteiro, L. G. e Boaventura, W. d. C., **Estado da arte para previsão da radiação solar**, *Anais Congresso Brasileiro de Energia Solar - CBENS*, nov. de 2020. DOI: 10.59627/cbens.2020.742.
- [4] Wan, C., Zhao, J., Song, Y., Xu, Z., Lin, J. e Hu, Z., **Photovoltaic and solar power forecasting for smart grid energy management**, *CSEE Journal of Power and Energy Systems*, v. 1, n. 4, pp. 38–46, 2015. DOI: 10.17775/CSEEJPES.2015.00046.
- [5] Martins Filho, J. B., Rodrigues, L. N. e Hernandez, F. B. T., **Validação de dados meteorológicos diários da plataforma NASA POWER para o município de Paranapuã-SP**, em *11ª Jornada Científica e Tecnológica da FATEC de Botucatu*, Botucatu, São Paulo, nov. de 2022.
- [6] Carvalho, W. K. M., Laboissiere, L. A. e Fernandes, R. A. S., **Previsão de irradiação solar e temperatura do ar utilizando Redes Neurais Artificiais para a cidade de São Luís-MA**, *9º Congresso Luso-Brasileiro para o Planejamento Urbano, Regional, Integrado e Sustentável (PLURIS 2021 Digital)*, pp. 1–13, 2021.
- [7] Anochi, J. A. e Silva, J. D. S., **Uso de Redes Neurais Artificiais e Teoria de Conjuntos Aproximativos no estudo de padrões climáticos sazonais**, *Learning Nonlinear Models - Revista da Sociedade Brasileira de Redes Neurais*, v. 7, n. 2, pp. 83–91, 2009.
- [8] Rudin, C. e Wagstaff, K. L., **Machine learning for science and society**, *Machine Learning*, v. 95, n. 1, pp. 1–9, abr. de 2014, ISSN: 1573-0565. DOI: 10.1007/s10994-013-5425-9. endereço: <https://doi.org/10.1007/s10994-013-5425-9>.
- [9] Taoufik, N., Boumya, W., Achak, M., Chennouk, H., Dewil, R. e Barka, N., **The state of art on the prediction of efficiency and modeling of the processes of pollutants removal based on machine learning**, *Science of The Total Environment*, v. 807, p. 150554, 2022, ISSN: 0048-9697. DOI: <https://doi.org/10.1016/j.scitotenv.2021.150554>. endereço: <https://www.sciencedirect.com/science/article/pii/S004896972105631X>.

- [10] IBGE (Instituto Brasileiro de Geografia e Estatística). **Arquivo zip contendo dados vectoriais dos biomas do Brasil**. (2014), endereço: http://geoftp.ibge.gov.br/informacoes_ambientais/estudos_ambientais/biomas/vetores/Biomas_250mil.zip. Acessado em: 30-01-2024.
- [11] IBGE (Instituto Brasileiro de Geografia e Estatística). **Arquivo zip contendo malhas territoriais do estado de Mato Grosso, Brasil**. (2014), endereço: https://geoftp.ibge.gov.br/organizacao_do_territorio/malhas_territoriais/malhas_municipais/municipio_2022/UFs/MT/MT_UF_2022.zip. Acessado em : 30-01-2024.
- [12] Alvares, C. A., Stape, J. L., Sentelhas, P. C., Moraes Gonçalves, J. L. e Sparovek, G., **Köppen's climate classification map for Brazil**, *Meteorologische Zeitschrift*, v. 22, n. 6, pp. 711–728, 2013.
- [13] Biudes, M. S., Vourlitis, G. L., Machado, N. G., **Patterns of energy exchange for tropical ecosystems across a climate gradient in Mato Grosso, Brazil**, *Agricultural and Forest Meteorology*, v. 202, pp. 112–124, 2015.
- [14] Souza, A., Mota, L., Zamadei, T., Martin, C., Almeida, F. e Paulino, J., **Classificação climática e balanço hídrico climatológico no estado de Mato Grosso**, *Nativa*, v. 1, n. 1, pp. 34–43, 2013. DOI: 10.31413/nativa.v1i1.1334.
- [15] Vourlitis, G. L., Souza Nogueira, J., Almeida Lobo, F. e Pinto Jr, O. B., **Variations in evapotranspiration and climate for an Amazonian semi-deciduous forest over seasonal, annual, and El Niño cycles**, *International Journal of Biometeorology*, v. 59, n. 2, pp. 217–230, 2015.
- [16] Dubreuil, V., Debortoli, N., Funatsu, B., Nédélec, V. e Durieux, L., **Impact of landcover change in the Southern Amazonia climate: a case study for the region of Alta Floresta, Mato Grosso, Brazil**, *Environmental Monitoring and Assessment*, v. 184, pp. 877–891, 2012. DOI: 10.1007/s10661-011-2006-x.
- [17] Silva, J. G. d., Bastos, V. e Alves, J. M. d. S., **Climatologia e Meteorologia: Livro Didático**, 3ª ed. Palhoça, Santa Catarina: Universidade do Sul de Santa Catarina, Campus UnisulVirtual, 2011.
- [18] Vianello, R. L., **A Estação Meteorológica e Seu Observador**, Instituto Nacional de Meteorologia, rel. técn., 2011. endereço: <https://portal.inmet.gov.br/uploads/publicacoesDigitais/aestacaometeorologicaeseuobservador.pdf>, Acessado em: 22-03-2024.
- [19] Helm, J. M., Swiergosz, A. M., Haeberle, H. S., **Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions**, *Current Reviews in Musculoskeletal Medicine*, v. 13, n. 1, pp. 69–76, fev. de 2020, ISSN: 1935-9748. DOI: 10.

- 1007/s12178-020-09600-8. endereço: <https://doi.org/10.1007/s12178-020-09600-8>.
- [20] Devulapalli, K., **Neural Networks for Classification and Regression**, *Biom Biostat Int J*, v. 2, n. 6, pp. 183–185, 2015. DOI: 10.15406/bbij.2015.02.00046. endereço: <https://doi.org/10.15406/bbij.2015.02.00046>.
- [21] Vicentinia, E. D. e Oliveira-Filho, A. G. S. d., **Introdução às Redes Neurais para Regressões Não-Lineares: Ajuste de Superfícies de Energia Potencial**, *Química Nova*, v. 44, n. 2, pp. 229–234, 2021. DOI: 10.21577/0100-4042.20170650. endereço: <http://dx.doi.org/10.21577/0100-4042.20170650>.
- [22] Fiorin, D. V., Martins, F. R., Schuch, N. J. e Pereira, E. B., **Aplicações de redes neurais e previsões de disponibilidade de recursos energéticos solares**, *Revista Brasileira de Ensino de Física*, v. 33, n. 1, mar. de 2011. DOI: 10.1590/S1806-11172011000100009. endereço: <https://doi.org/10.1590/S1806-11172011000100009>.
- [23] Prajapati, A. e Singh, U., **An Optimal Solution to the Overfitting and Underfitting Problem of Healthcare Machine Learning Models**, *Journal of Systems Engineering and Information Technology (JOSEIT)*, v. 2, pp. 77–84, out. de 2023. DOI: 10.29207/joseit.v2i2.5460.
- [24] Ying, X., **An Overview of Overfitting and its Solutions**, *Journal of Physics: Conference Series*, v. 1168, 2019. DOI: 10.1088/1742-6596/1168/2/022022.
- [25] Abhishek, K., Singh, M., Ghosh, S. e Anand, A., **Weather Forecasting Model using Artificial Neural Network**, *Procedia Technology*, v. 4, pp. 311–318, 2012, 2nd International Conference on Computer, Communication, Control and Information Technology (C3IT-2012) on February 25 - 26, 2012, ISSN: 2212-0173. DOI: <https://doi.org/10.1016/j.protcy.2012.05.047>. endereço: <https://www.sciencedirect.com/science/article/pii/S221201731200326X>.
- [26] scikit-learn. **GridSearchCV**. (2024), endereço: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html. Acessado em: 18-03-2024.
- [27] scikit-learn. **MinMaxScaler**. (2024), endereço: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>. Acessado em: 26-03-2024.
- [28] scikit-learn. **train_test_split**. (2024), endereço: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html. Acessado em: 26-03-2024.
- [29] Khan, S., Rahmani, H., Ali Shah, S. A. e Bennamoun, M., **A Guide to Convolutional Neural Networks for Computer Vision**. Williston, USA: Morgan & Claypool Publishers, 2018, pp. 31–33, 43–44, 1.

- [30] Franjić, S. e Galić, D., **In Shortly about Neural Networks**, *Advances in Computer and Communication*, 2021. DOI: 10.26855/acc.2021.08.001.
- [31] Rosenblatt, F., **The perceptron: A probabilistic model for information storage and organization in the brain**, *Psychological Review*, v. 65, n. 6, pp. 386–408, 1958. DOI: 10.1037/h0042519.
- [32] Cinar, A. C., **Training Feed-Forward Multi-Layer Perceptron Artificial Neural Networks with a Tree-Seed Algorithm**, *Arabian Journal for Science and Engineering*, v. 45, n. 12, pp. 10915–10938, 2020, ISSN: 2191-4281. DOI: 10.1007/s13369-020-04872-1. endereço: <https://doi.org/10.1007/s13369-020-04872-1>.
- [33] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. e Salakhutdinov, R., **Dropout: a simple way to prevent neural networks from overfitting**, *J. Mach. Learn. Res.*, v. 15, pp. 1929–1958, 2014. DOI: 10.5555/2627435.2670313.
- [34] Janiesch, C., Zschech, P. e Heinrich, K., **Machine learning and deep learning**, *Electronic Markets*, v. 31, n. 3, pp. 685–695, set. de 2021, ISSN: 1422-8890. DOI: 10.1007/s12525-021-00475-2. endereço: <https://doi.org/10.1007/s12525-021-00475-2>.
- [35] Le, X.-H., Ho, H. V., Lee, G. e Jung, S., **Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting**, *Water*, v. 11, n. 7, 2019, ISSN: 2073-4441. DOI: 10.3390/w11071387. endereço: <https://www.mdpi.com/2073-4441/11/7/1387>.
- [36] Cho, K. van, Merriënboer, B. van, Gulcehre, C., **Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation**, *arXiv preprint arXiv: 1406.1078*, set. de 2014. endereço: <https://arxiv.org/abs/1406.1078>.
- [37] Hochreiter, S. e Schmidhuber, J., **Long Short-Term Memory**, *Neural Computation*, v. 9, n. 8, pp. 1735–1780, nov. de 1997, ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. eprint: <https://direct.mit.edu/neco/article-pdf/9/8/1735/813796/neco.1997.9.8.1735.pdf>. endereço: <https://doi.org/10.1162/neco.1997.9.8.1735>.
- [38] Tian, Y., Su, D., Lauria, S. e Liu, X., **Recent advances on loss functions in deep learning for computer vision**, *Neurocomputing*, v. 497, pp. 129–158, 2022, ISSN: 0925-2312. DOI: <https://doi.org/10.1016/j.neucom.2022.04.127>. endereço: <https://www.sciencedirect.com/science/article/pii/S0925231222005239>.
- [39] Hodson, T. O., **Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not**, *Geoscientific Model Development*, v. 15, n. 14, pp. 5481–5487, 2022. DOI: 10.5194/gmd-15-5481-2022. endereço: <https://gmd.copernicus.org/articles/15/5481/2022/>.

- [40] Lira, S. e Chaves Neto, A., **Coeficientes de correlação para variáveis ordinais e dicotômicas derivados do coeficiente linear de Pearson**, v. 15, jun. de 2008.
- [41] Willmott, C. J., **On the Validation of Models**, *Physical Geography*, v. 2, n. 2, pp. 184–194, 1981. DOI: 10.1080/02723646.1981.10642213. eprint: <https://doi.org/10.1080/02723646.1981.10642213>. endereço: <https://doi.org/10.1080/02723646.1981.10642213>.
- [42] Bunchaft, G. e Kellner, S., **Statistics without Mysteries**, 2nd. Petrópolis: Vozes, 1999, vol. 2.