

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE MINAS
GERAIS – *CAMPUS* BAMBUÍ
BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO

Kevenn Henrique de Paula Silva

**REDUÇÃO DE DIMENSIONALIDADE EM DADOS DE
CLIMA COM USO DE *STACKED AUTOENCODERS***

KEVENN HENRIQUE DE PAULA SILVA

**REDUÇÃO DE DIMENSIONALIDADE EM DADOS DE
CLIMA COM USO DE *STACKED AUTOENCODERS***

Trabalho de conclusão de curso apresentado ao Curso de Bacharelado em Engenharia de Computação do Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais – *Campus* Bambuí para obtenção do grau de Bacharel em Engenharia de Computação.

Orientador: Prof. Dr. Ciniro Ap. Leite Nametala

Bambuí - MG

2023

Catálogo na Fonte Biblioteca IFMG - Campus Bambuí

S586r Silva, Kevenn Henrique de Paula.
Redução de dimensionalidade em dados de clima com uso de Stacked Autoencoders. / Kevenn Henrique de Paula Silva. – 2023.
50 f. : il. ; color.

Orientador: Prof. Dr. Ciniro Ap. Leite Nametala.
Trabalho de Conclusão de Curso (graduação) - Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais – Campus Bambuí, MG, Curso Bacharelado em Engenharia de Computação, 2023.

1. Autoencoders. 2. Redução de dimensionalidade. 3. Dados climáticos. I. Nametala, Ciniro Ap. Leite. II. Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais – Campus Bambuí, MG. III. Título.

CDD 001.61

Kevenn Henrique de Paula Silva

REDUÇÃO DE DIMENSIONALIDADE EM DADOS DE CLIMÁ COM USO DE *STACKED AUTOENCODERS*

Trabalho de conclusão de curso apresentado ao Curso de Bacharelado em Engenharia de Computação do Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais – *Campus* Bambuí para obtenção do grau de Bacharel em Engenharia de Computação.

Aprovado em 11 de dezembro de 2023 pela banca examinadora:

Prof. Dr. Ciniro Ap. Leite Nametala – IFMG – *Campus* Bambuí – (Orientador)
Prof. Me. Cláudio Ribeiro de Sousa – IFMG – *Campus* Bambuí
Prof. Dr. Marcos Roberto Ribeiro – IFMG – *Campus* Bambuí



Documento assinado eletronicamente por **Ciniro Aparecido Leite Nametala**, **Professor**, em 11/12/2023, às 08:18, conforme Decreto nº 10.543, de 13 de novembro de 2020.



Documento assinado eletronicamente por **Claudio Ribeiro de Sousa**, **Professor EBTT**, em 11/12/2023, às 08:18, conforme Decreto nº 10.543, de 13 de novembro de 2020.



Documento assinado eletronicamente por **Marcos Roberto Ribeiro**, **Professor**, em 11/12/2023, às 08:19, conforme Decreto nº 10.543, de 13 de novembro de 2020.



A autenticidade do documento pode ser conferida no site <https://sei.ifmg.edu.br/consultadocs> informando o código verificador **1758477** e o código CRC **20FE1139**.

*Dedico este trabalho a todos que, de alguma forma, tenham contribuído, seja com seu apoio, inspiração, orientação, conhecimento, ou simplesmente com seu incentivo e compreensão ao longo dessa jornada. Suas contribuições tornaram este trabalho possível.
Obrigado!!!*

“Quanto mais aumenta nosso conhecimento, mais evidente fica nossa ignorância.”
(John F. Kennedy)

RESUMO

Durante as últimas décadas, as bases de dados vêm crescendo exponencialmente. Esse aumento não ocorre somente na quantidade de amostras de dados, mas também em relação à quantidade de características que descrevem as variáveis, deixando os *datasets* cada vez mais massivos. Devido a isso, torna-se necessária uma simplificação desses conjuntos de dados por meio de uma redução nessas dimensionalidades. Os dados climáticos são exemplos de dados que, geralmente, possuem muitas amostras e características envolvidas. Estes fatores acarretam uma alta dimensionalidade e, por sua vez, esse cenário afeta o custo computacional e a capacidade preditiva, prejudicando a busca por padrões e descoberta de conhecimento. Além disso, a redução de dimensionalidade irá beneficiar a visualização e o armazenamento de grandes conjuntos de dados climáticos. Desse modo, este trabalho propôs utilizar uma Rede Neural Artificial (RNA) com arquitetura *Autoencoder* (AE), especificamente um tipo chamado de *Stacked Autoencoder* (SAE), para realizar a compressão de dados climáticos de entrada, criando uma representação compacta e de menor dimensão. Após isso, com a capacidade do AE de reconstruir os dados de entrada a partir dessa representação latente, verificou-se que a redução de dimensionalidade é boa o suficiente para reconstruir os dados. Estes dados reconstruídos foram avaliados por meio do erro quadrático médio (EQM), obtendo-se um valor de 0.01605. Com esse resultado, é possível confirmar que a RNA cumpriu sua função de reduzir a base de dados climáticos.

Palavras-chave: *Autoencoders*. Redução de dimensionalidade. Dados climáticos.

ABSTRACT

Over the past few decades, databases have been growing exponentially. This increase is not only in terms of the quantity of data samples but also in the number of features describing the variables, making datasets increasingly massive. Due to this, it becomes necessary to simplify these data sets by reducing their dimensionality. Climatic data are examples of data that often have many samples and involved characteristics. These factors result in high dimensionality, which, in turn, affects computational cost and predictive capacity, hindering the search for patterns and knowledge discovery. Furthermore, dimensionality reduction will benefit the visualization and storage of large climatic data sets. Thus, this work aims to use an Artificial Neural Network (ANN) with an Autoencoder (AE) architecture, specifically a type called a Stacked Autoencoder (SAE), to compress input climatic data, creating a compact and lower-dimensional representation. After that, with the AE's ability to reconstruct the input data from this latent representation, it was found that the dimensionality reduction is good enough to reconstruct the data. These reconstructed data were evaluated using the mean squared error (MSE), resulting in a value of 0.01605. With this result, it is possible to confirm that the ANN fulfills its role of reducing the climatic data set.

Keywords: *Autoencoders. Dimensionality reduction. Climate data.*

LISTA DE FIGURAS

Figura 1 – Representação esquemática do AE	13
Figura 2 – Frequência relativa de citações	16
Figura 3 – Modelo de um neurônio	18
Figura 4 – Modelo básico utilizado pelos AEs	20
Figura 5 – Arquitetura básica de um AE	20
Figura 6 – Esquemática da metodologia	25
Figura 7 – Divisão do <i>dataset</i>	27
Figura 8 – Arquitetura de um SAE	31
Figura 9 – Distribuição dos dados de 2020	33
Figura 10 – Convergência do erro com uma representação latente	39
Figura 11 – Convergência do erro com duas representação latente	49
Figura 12 – Convergência do erro com três representação latente	49
Figura 13 – Convergência do erro com quatro representação latente	50
Figura 14 – Convergência do erro com cinco representação latente	50

LISTA DE TABELAS

Tabela 1 – Sumarização do <i>dataset</i> de 2020	34
Tabela 2 – Sumarização do <i>dataset</i> de 2022	35
Tabela 3 – Dados faltantes no <i>dataset</i> de 2020	36
Tabela 4 – Dados faltantes no <i>dataset</i> de 2022	36
Tabela 5 – Divisão dos <i>datasets</i>	37
Tabela 6 – Soma total de cada atributo de entrada	40
Tabela 7 – Soma total de cada atributo reconstruído	40
Tabela 8 – Diferença do total de cada atributo	41
Tabela 9 – EQM de diferentes representações latentes	42

LISTA DE ABREVIATURAS E SIGLAS

AAEs – *Asymmetric Autoencoders*

AE – *Autoencoder*

API – *Application Programming Interface*

EQM – Erro quadrático médio

INMET – Instituto Nacional de Meteorologia

IoT – *Internet of Things*

IPCA – *Independent Principal Component Analysis*

NF-es – Notas Fiscais Eletrônicas

PCA – *Principal Component Analysis*

RNA – Rede Neural Artificial

SAE – *Stacked Autoencoder*

SPCA – *Sparse Principal Component Analysis*

SymAE – *Symmetric Autoencoder*

VAE – *Variational Autoencoder*

SUMÁRIO

1	INTRODUÇÃO	13
1.1	Objetivos	14
1.1.1	<i>Objetivo Geral</i>	14
1.1.2	<i>Objetivos Específicos</i>	14
1.1.3	<i>Justificativa</i>	14
1.1.4	<i>Estrutura do documento</i>	15
2	REFERENCIAL TEÓRICO	16
2.1	Dados climáticos	16
2.2	Redução da dimensionalidade	17
2.3	Redes Neurais	17
2.4	AEs	19
2.5	Levantamento de literatura	21
2.6	Estado-da-arte	21
3	METODOLOGIA	24
3.1	Abordagem de desenvolvimento	24
3.2	Classificação do trabalho	24
3.3	Materiais	25
3.4	Método	25
3.4.1	<i>Etapa 1: Coleta dos dados</i>	26
3.4.2	<i>Etapa 2: Análise exploratória dos dados</i>	26
3.4.3	<i>Etapa 3: Tratamento dos dados</i>	26
3.4.4	<i>Etapa 4: Divisão dos datasets nos conjuntos de treino, validação e teste</i>	27
3.4.5	<i>Etapa 5: Construção da Rede Neural</i>	27
3.4.6	<i>Etapa 6: Treinamento da Rede Neural</i>	29
3.4.7	<i>Etapa 7: Avaliação da reconstrução dos dados</i>	29
4	EXPERIMENTOS E RESULTADOS	31
4.1	Escolha do AE: SAE	31
4.2	Ferramentas para implementação	31
4.3	Coleta dos dados	32
4.4	Análise exploratória dos dados	32
4.4.1	<i>Descobertas iniciais</i>	32
4.4.2	<i>Verificando dados faltantes</i>	34
4.5	Tratamento dos dados	35
4.5.1	<i>Tratamento dos dados faltantes</i>	35

4.5.2	<i>Escalonamento</i>	37
4.6	Divisão dos <i>datasets</i> nos conjuntos de treino, validação e teste . .	37
4.7	Construção da Rede Neural	37
4.8	Treinamento da Rede Neural	38
4.9	Avaliação da reconstrução dos dados	39
4.10	Avaliação da reconstrução dos dados climáticos utilizando dife- rentes tamanhos de dados latentes	41
5	CONCLUSÃO	43
	REFERÊNCIAS	45
	APÊNDICES	49
	APÊNDICE A – GRÁFICOS DE CONVERGÊNCIA DO ERRO . . .	49

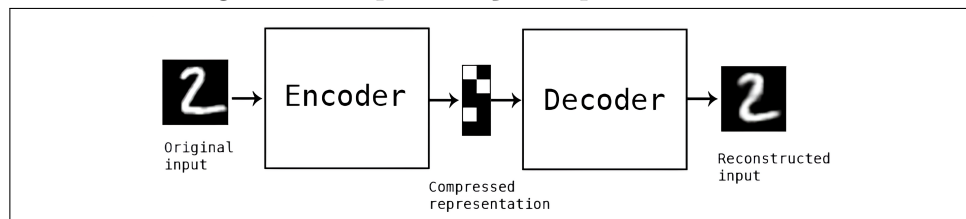
1 INTRODUÇÃO

No decorrer das últimas décadas, as bases de dados vêm crescendo exponencialmente (MENEZES; FREITAS; PARPINELLI, 2016). Esse aumento não ocorre somente na quantidade de amostras de dados, mas também em relação à quantidade de características que descrevem as variáveis. Estas, por sua vez, estimulam acréscimos nas dimensões do espaço matemático, sendo preciso fazer uma redução nessas dimensionalidades, buscando melhorar o desempenho e a simplificação dos resultados da modelagem (CAMARGO, 2010).

Os dados climáticos são exemplos de dados que possuem muitas características envolvidas, acarretando um alto grau de dimensionalidade, o que, por sua vez, afeta o custo computacional e a capacidade preditiva em eventuais tarefas de aprendizado de máquina. Essa característica tende a prejudicar a busca por padrões e a descoberta de conhecimento (CIOFFI, 2022).

Das várias abordagens existentes para se realizar a redução de dimensionalidade, o método designado nesta pesquisa foi o *Autoencoder* (AE), especificamente um tipo chamado de *Stacked Autoencoder* (SAE). Consegue codificar os dados em uma representação latente e, a partir dessa representação, reconstruir os dados de entrada da melhor forma possível. A Figura 1 ilustra esse processo. O AE, em sua fase de aprendizado, tenta chegar em um bom modelo, e, quando o alcança, consegue reconstruir os dados (PIRES, 2017).

Figura 1 – Representação esquemática do AE



Fonte: CHOLLET, 2017.

Uma propriedade valiosa dos AEs é a sua capacidade de treinamento automática com base em exemplos de dados. Isso significa que é possível treinar instâncias especializadas do algoritmo que apresentaram um bom desempenho em um tipo específico de entrada, sendo necessário, para isso, apenas ter dados de treinamento adequados (CHOLLET, 2017). No contexto desta pesquisa, os dados de entrada específicos referem-se aos dados climáticos, criando um modelo especializado na codificação e decodificação destes dados.

Alguns autores, como Alsenan, Al-Turaiqi e Hafez (2020), Banihashemi, Weber e Lang (2022) e Boquet *et al.* (2020), utilizaram AEs em suas pesquisas, alcançando resultados satisfatórios. A eficácia demonstrada por esses estudos indica que a aplicação e a investigação mais aprofundadas dessa técnica, neste contexto específico, podem ser valiosas. Estes estudos são abordados posteriormente, em mais detalhes, na Seção 2.6.

Nesta pesquisa, a quantidade de amostras de dados, com suas respectivas características, foi o problema. Estas amostras são resultantes de coletas de dados em intervalos horários, ao longo do ano, ou seja, matrizes de dados climáticos coletados de hora em hora, ao longo de um ano.

1.1 Objetivos

Os objetivos geral e específicos são tratados a seguir.

1.1.1 *Objetivo Geral*

Este estudo propôs utilizar uma Rede Neural Artificial (RNA), com arquitetura SAE, para realizar a compressão de dados climáticos de entrada, criando uma representação compacta e de menor dimensão. Com esta representação latente, foram reconstruídos os dados climáticos originais. Após isso, avaliou-se a qualidade da reconstrução dos dados multidimensionais de clima - dados estes coletados pelas estações de solo do Instituto Nacional de Meteorologia (INMET) da Região Sudeste do Brasil. A avaliação dessa qualidade de reconstrução foi baseada no erro quadrático médio (EQM) entre os dados de entrada e os reconstruídos.

1.1.2 *Objetivos Específicos*

Os objetivos específicos deste trabalho foram os seguintes:

- coletar e preparar os dados climáticos necessários para o AE;
- identificar o tipo de AE mais adequado para a tarefa de redução de dimensionalidade dos dados climáticos, avaliando algumas arquiteturas e hiper-parâmetros disponíveis;
- buscar hiper-parâmetros que melhor se ajustassem ao AE, permitindo otimizar a capacidade na compressão e reconstrução dos dados climáticos.

1.1.3 *Justificativa*

Considerando-se que os dados climáticos possuem inúmeras características, resultando em uma alta dimensionalidade, é possível observar que a capacidade preditiva e a busca por padrões são prejudicadas (CIOFFI, 2022). A redução de dimensionalidade é uma técnica que pode ser empregada para lidar com esse problema (ATIENZA, 2018), tornando a análise dos dados mais eficiente (BORGES; NIEVOLA, 2006). Dessa forma, o objetivo deste trabalho foi justamente verificar se a redução de dimensionalidade dos dados climáticos, utilizando SAE, pode ser uma estratégia eficaz para melhorar a análise deles.

1.1.4 Estrutura do documento

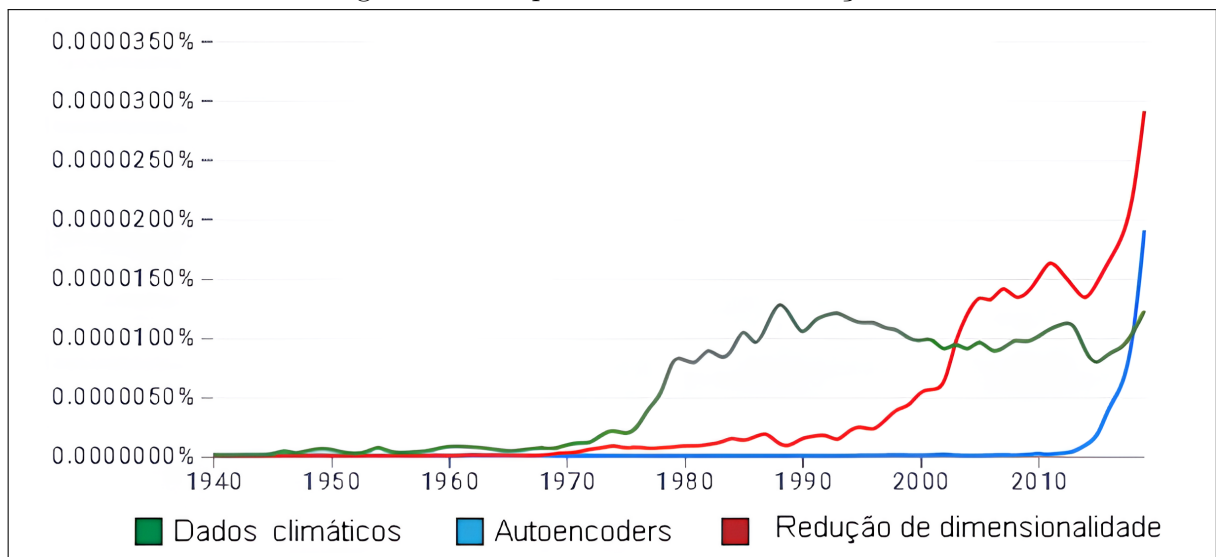
Na sequência deste capítulo, apresenta-se a estrutura deste documento, que está organizada conforme a seguir:

- no Capítulo 2, encontra-se o referencial teórico, com os principais conceitos necessários ao entendimento deste estudo;
- no Capítulo 3, a metodologia aplicada é exposta, juntamente com as métricas de avaliação adotadas;
- no Capítulo 4, os experimentos e os resultados são apresentados;
- por fim, o Capítulo 5 finaliza, com a conclusão.

2 REFERENCIAL TEÓRICO

Neste capítulo, serão abordados conceitos, fundamentos e trabalhos correlatos necessários para se compreender o tema proposto e demais temas relacionados, além de serem referência para a presente monografia. Mas, antes disso, será feita uma breve análise bibliométrica utilizando-se as palavras-chaves da pesquisa, sendo elas: *Autoencoders*, redução de dimensionalidade e dados climáticos.

Figura 2 – Frequência relativa de citações



Fonte: Adaptado de: BOOKS, 2023.

Conforme visto na Figura 2, a partir de 1970, as citações de dados climáticos tiveram um grande aumento na literatura. Juntamente com esse aumento, é possível notar que a redução de dimensionalidade começou a ser citada com mais frequência. Por meio desta análise, é interessante notar que o comportamento da curva da quantidade de citações de dados segue o mesmo movimento que as técnicas de redução de dimensionalidade. Além disso, é possível perceber que, de acordo com a Figura 2, a partir de 2014, os AEs tiveram um exponencial crescimento.

Sabendo-se que, basicamente, AEs são técnicas para se efetuar a redução de dimensionalidade, então pode-se especular que, provavelmente, esta técnica contribuiu para haver também, a partir de 2014, uma curva de crescimento exponencial na quantidade de citações sobre redução de dimensionalidade.

2.1 Dados climáticos

Os dados climáticos são um exemplo de dados que apresentam múltiplas características, resultando em uma alta dimensionalidade. Isso gera implicações negativas no custo computacional, na capacidade preditiva em tarefas de aprendizado de máquina, na busca por padrões e na descoberta de conhecimento (CIOFFI, 2022).

A capacidade de gerenciar e consultar eficientemente dados climáticos multidimensionais em formato de grade ou matriz ainda é um desafio para a maioria dos bancos de dados. A incompatibilidade entre o modelo de dados em matriz e o modelo de dados relacionais tem limitado o desempenho na consulta de dados multidimensionais em bancos de dados tradicionais. Essa limitação é especialmente evidente quando o volume de dados atinge um limite de *hardware*. Até mesmo uma simples recuperação de dados em grandes conjuntos de dados multidimensionais, em um banco de dados relacional, é demorada e requer um espaço de armazenamento significativo (XU *et al.*, 2021).

Com base nessas informações, é interessante considerar que, caso a redução de dimensionalidade utilizando AEs obtenha uma alta taxa de representatividade dos dados originais, será possível armazenar os dados latentes, sendo estes dados mais compactos que os originais.

2.2 Redução da dimensionalidade

Atualmente, a quantidade de dados armazenados é imensa e continua a crescer rapidamente. No entanto, devido à limitação humana na interpretação de tal quantidade de dados, muitas informações e conhecimentos potencialmente úteis podem estar sendo desperdiçados, ficando ocultos dentro das bases de dados (REZENDE, 2003).

Cientistas que lidam com uma grande quantidade de dados complexos, por exemplo, dados climáticos globais, espectros estelares ou distribuições de genes humanos, estão frequentemente enfrentando desafios para a redução da dimensionalidade dos dados. Eles buscam extrair estruturas significativas de dados com alta dimensão e representá-las de forma mais simples, em dimensões inferiores. Essa tarefa é semelhante ao que o cérebro humano faz na percepção cotidiana, onde extrai um número relativamente pequeno de características relevantes a partir de 30.000 fibras nervosas sensoriais de entrada (TENENBAUM; SILVA; LANGFORD, 2000).

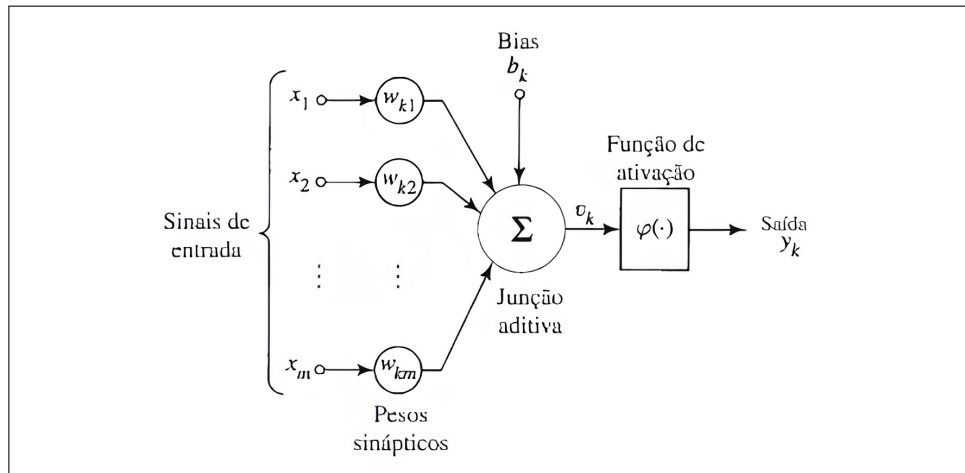
2.3 Redes Neurais

Haykin (2001) define uma RNA como um sistema de processamento de informações inspirado no funcionamento do cérebro humano, sendo ela um conjunto interconectado de unidades de processamento simples, chamadas de neurônios artificiais, que trabalham em conjunto para realizar tarefas de processamento de informações.

Uma RNA é composta por várias camadas de neurônios interconectados, onde cada neurônio recebe um conjunto de entradas. Após isso, realiza um cálculo simples com essas entradas e produz uma saída, possuindo, então, a capacidade de aprender a partir dos dados de entrada. As conexões entre os neurônios têm pesos que controlam a influência de cada um nas saídas dos outros neurônios. Um dos principais mecanismos de

aprendizado das RNAs é o ajuste dos pesos das conexões para otimizar o desempenho da rede em relação a uma determinada tarefa (HAYKIN, 2001).

Figura 3 – Modelo de um neurônio



Fonte: HAYKIN, 2001.

A Figura 3 é a representação de um modelo de um neurônio que forma a base para o projeto de RNAs. Segundo Suzuki (2013), Loy (2019) e Kovács (2002), uma breve apresentação sobre cada um dos seus elementos é exposta a seguir:

- sinais de entrada:

Os sinais de entrada são os dados ou informações que são fornecidos ao neurônio.

- bias:

O bias é um valor constante adicionado aos sinais de entrada antes de serem processados pelo neurônio. Permite ao neurônio ajustar o ponto de operação ou a sensibilidade para diferentes intervalos de entrada.

- pesos sinápticos:

Os pesos sinápticos são valores associados a cada conexão entre os sinais de entrada e o neurônio. Eles representam a força ou a importância relativa de cada sinal de entrada para o neurônio. Os pesos sinápticos são ajustados durante o processo de aprendizado para que o neurônio possa aprender a reconhecer padrões ou realizar determinadas tarefas.

- junção aditiva:

A junção aditiva é a etapa em que os sinais de entrada ponderados pelos pesos sinápticos e o bias são somados. Essa soma representa a combinação linear dos sinais de entrada.

- função de ativação:

A função de ativação é aplicada ao resultado da junção aditiva. A função de ativação introduz não linearidades ao processamento do neurônio e determina se este será ativado ou não, ou seja, se ele enviará um sinal de saída, ou não, com base na entrada recebida. A função de ativação pode ser simples, como a função degrau, ou mais complexa, como a função sigmoide.

- saída:

A saída do neurônio é o resultado final após a aplicação da função de ativação. Dependendo do tipo de problema ou tarefa, a saída pode ter diferentes interpretações ou representações. Por exemplo, em um problema de classificação binária, a saída pode ser interpretada como a probabilidade de pertencer a uma classe específica.

Os autores Suzuki (2013), Loy (2019) e Kovács (2002) descrevem o funcionamento essencial de uma RNA como: inicialmente, para os pesos sinápticos nas conexões entre os neurônios, são atribuídos valores aleatórios. Durante o treinamento, um conjunto de dados de entrada é introduzido na rede. Cada sinal de entrada, incluindo o viés (bias), é multiplicado pelos pesos sinápticos correspondentes. Os resultados dessas multiplicações são somados, representando a junção aditiva. Em seguida, essa soma passa por uma função de ativação, que determina a saída do neurônio.

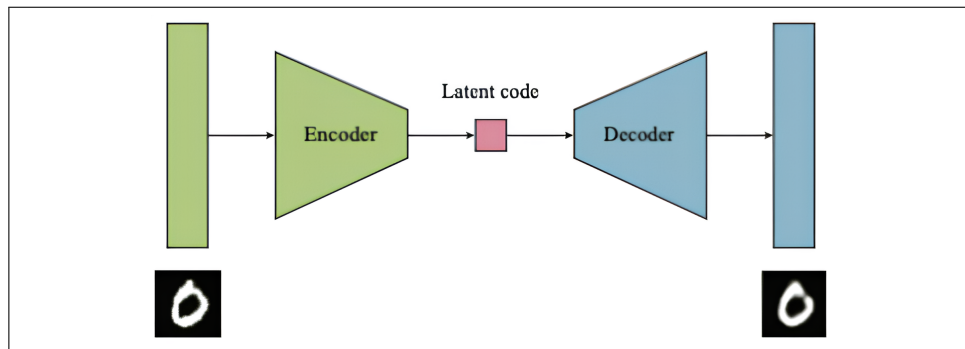
Este processo é repetido para cada neurônio, em uma rede neural. As saídas geradas são comparadas com os valores desejados. Os pesos são, então, ajustados iterativamente para minimizar a diferença entre as saídas reais e as desejadas. Esse processo é repetido para várias amostras de treinamento até que a rede aprenda da melhor forma possível a mapear as entradas para as saídas. Isso permite que a rede processe e transforme informações complexas, adaptando seus pesos sinápticos com base nos padrões presentes nos dados.

2.4 AEs

Os AEs são RNAs, ou seja, RNAs com arquitetura AEs. Possuem a capacidade de comprimir e descomprimir um conjunto de dados, podendo ser utilizados na redução de dimensionalidade (ATIENZA, 2018). Gonçalves (2023) pontua que isso ocorre por que “os AEs conseguem, por meio de seu codificador, retirar apenas as informações úteis de um conjunto de dados. Ademais, o decodificador serve para poder regenerar os dados para dimensionalidade original quando necessário”.

Na Figura 4, é possível visualizar o modelo básico utilizado pelos AEs que permite transformar a entrada da rede em uma representação latente usando um codificador. Em seguida, um decodificador reconstrói a entrada da maneira mais fiel possível ao original (MECHELLI; VIERA, 2019).

Figura 4 – Modelo básico utilizado pelos AEs

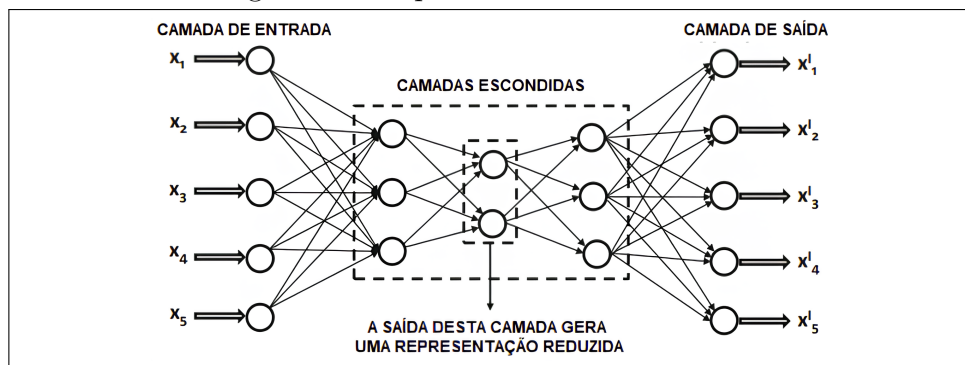


Fonte: MECHELLI; VIERA, 2019.

Devido a essa capacidade de compactação, as redes neurais AEs são opções viáveis para reduzir o tamanho de grandes conjuntos de dados e, posteriormente, caso necessário, reconstruí-los.

A utilização de técnicas de redução de dimensionalidade não apenas remove dados redundantes e irrelevantes, mas também melhora a compreensão dos resultados obtidos, destacando a influência de cada atributo selecionado e seu nível de expressão em relação ao objetivo desejado (BORGES; NIEVOLA, 2006). Na Figura 5, é apresentada a arquitetura básica de um AE com seus respectivos neurônios e camadas.

Figura 5 – Arquitetura básica de um AE



Fonte: MORAES; ROSSI, 2020.

Uma característica que distingue essa rede AE das outras é que o ajuste dos pesos da rede ocorre de forma supervisionada. O modelo é treinado para aprender representações dos dados a partir das entradas, pois as saídas devem ser a reconstrução das entradas. Isso significa que a qualidade dos pesos é avaliada com base na análise da reconstrução dos dados em comparação com os valores de entrada originais. Como dito anteriormente, o processo de aprendizado dos AEs é dividido em uma estrutura codificador-decodificador. O objetivo do aprendizado é ajustar os pesos correspondentes à fase de codificação, de modo a produzir uma representação codificada que possa ser recuperada na fase de decodificação. Os pesos são ajustados nessa fase para possibilitar a reconstrução dos dados inseridos (BOCHIE *et al.*, 2020).

2.5 Levantamento de literatura

Esta seção aborda como foram feitas a pesquisa e a escolha dos trabalhos referenciados, apresentando quais foram os critérios usados para a escolha deles.

Inicialmente, com base no contexto do trabalho, foram definidas as três palavras-chaves: *Autoencoders*, redução de dimensionalidade e dados climáticos. Posteriormente, elas foram utilizadas para realizar a pesquisa na base científica *google scholar*, pois possui abrangência geográfica, linguística, temática, diversidade de documentos e recuperação rápida de informações (VERÍSSIMO, 2012). Optou-se por priorizar os trabalhos científicos que:

- possuíam *datasets* climáticos;
- foram publicadas após o ano de 2019;
- utilizaram RNA com arquitetura AE;
- efetuaram a redução de dimensionalidade em seus dados.

Como resultado da busca, foram obtidos os trabalhos científicos apresentados na seção a seguir.

2.6 Estado-da-arte

Nesta seção, são apresentados trabalhos que exibem alguns aspectos em comum com esta respectiva pesquisa, destacando-se, ao final, algumas destas semelhanças.

Costa (2022) propôs o desenvolvimento de metodologias eficientes para a análise de dados genômicos com alta dimensionalidade, buscando reduzir o tempo e esforço computacional necessários. Foram aplicadas técnicas como *Principal Component Analysis* (PCA), *Sparse Principal Component Analysis* (SPCA), *Independent Principal Component Analysis* (IPCA) e AE para analisar os dados de marcadores moleculares do arroz asiático. Os resultados mostraram que as técnicas estatísticas e o AE tiveram desempenhos semelhantes em termos de matriz de confusão e taxa de acerto. No entanto, o AE é capaz de formar grupos mais compactos, com menor variância interna e maior dissimilaridade entre eles, em comparação com os métodos estatísticos tradicionais.

Alsenan, Al-Turaiki e Hafez (2020) diz, em seu estudo, que AE é um tipo de rede neural que não é totalmente explorado para fins de redução de dimensionalidade. Investigando o impacto do AE em um conjunto de dados de alta dimensão e o desempenho desta técnica em uma análise preliminar, demonstraram-se superior em comparação com a técnica PCA, que é uma técnica de redução de dimensionalidade linear amplamente utilizada.

Nascimento *et al.* (2020) tiveram como objetivo, no estudo realizado, desenvolver sistemas de detecção de falhas em poços de produção de petróleo. Para isso, utilizaram dados de domínio público *3W dataset* e a técnica de AE para redução de dimensionalidade. A classificação foi realizada com base em cinco variáveis de processo, utilizando técnicas como Floresta de Isolamento e Máquinas de Vetores de Suporte. Os resultados mostraram uma melhoria significativa em comparação com outros trabalhos da literatura, com uma diferença de até dezoito pontos percentuais. Os AEs demonstraram eficácia na redução de dimensionalidade e extração de características relevantes para detecção de anomalias em poços de produção de petróleo *offshore*.

Banihashemi, Weber e Lang (2022) diz, em seu trabalho, que o AE pode ser aplicado de forma eficiente para criar uma representação de baixa dimensão dos dados usados na simulação de consumo energético em edifícios. Essa redução permite estimar as necessidades de energia de um edifício com alta precisão, reduzindo drasticamente os tempos de simulação. O resultado é um modelo simplificado, com menos de 1% de erro em relação ao modelo original, tornando mais viáveis as análises paramétricas e otimizações numéricas em edifícios complexos.

Boquet *et al.* (2020) utilizou o *Variational Autoencoder* (VAE) para aprender como os dados de tráfego são gerados. Foi criado um modelo de previsão de tráfego rodoviário com base no pressuposto de que os dados podem ser gerados a partir de um espaço latente com menos dimensões contendo as características do tráfego. O resultado é um modelo que não apenas aprende características úteis de previsão, como também pode ser utilizado para detecção de anomalias.

Silva Gilbert, Campos e Campista (2023) propuseram o uso de *Asymmetric Autoencoders* (AAEs) na pesquisa realizada como uma solução de baixo custo para compressão e redução de dimensionalidade em cenários de *Internet of Things* (IoT). Os dados gerados pelos IoT podem apresentar redundâncias intrínsecas que permitam a sua compressão sem perdas de informação, reduzindo a quantidade de dados transmitidos pela rede, sendo essa uma das tarefas com maior consumo de energia para dispositivos IoT. Os experimentos mostraram melhorias em relação aos *Symmetric Autoencoder* (SymAE), alcançando menores erros de reconstrução.

Schulte (2022) desenvolveu um serviço chamado Elinac, que tem como objetivo agrupar dados de texto curto, que, neste caso, são as Notas Fiscais Eletrônicas (NF-es), utilizando um AE. O Elinac foi criado para auxiliar na auditoria de transações documentadas em NF-e, possibilitando o agrupamento de dados semelhantes com base em descrições curtas de texto, o que facilita a detecção de anomalias em campos numéricos. O diferencial do Elinac é a forma como o AE é modelado, garantindo que os custos computacionais não sejam aumentados significativamente ao lidar com um grande volume de dados de texto curto. Os resultados obtidos demonstraram que o Elinac é eficiente no agrupamento de dados e possui uma execução três vezes mais rápida que as soluções

encontradas na literatura.

Menegolla (2019) teve como objetivo em seu trabalho estudar e aplicar a técnica de AE para o desenvolvimento de inferências em processos industriais. O problema abordado é a dificuldade de medição de algumas variáveis nos processos industriais, devido ao tempo de amostragem e incerteza das medições dos analisadores. Para contornar esse problema, o AE é utilizado para comprimir e descomprimir os dados de entrada, criando um espaço latente de menor dimensão que concentra as informações relevantes. Esse espaço latente é, então, utilizado para desenvolver as inferências.

Para contar o número de pessoas em um ambiente utilizando sinais de radiofrequência, Medeiros Campos *et al.* (2021) propuseram utilizar a redução de dimensionalidade em dados através de um AE, onde, posteriormente, estes dados latentes foram analisados por três classificadores. Os resultados desta pesquisa mostraram que foi possível obter o número de presentes em um ambiente com pelo menos 90% de acurácia.

Algumas semelhanças que o presente trabalho possui em relação a algumas pesquisas apresentadas são:

- assim como em Silva Gilbert, Campos e Campista (2023), os dados climáticos utilizados neste trabalho também apresentam redundâncias e padrões intrínsecos;
- o AE foi modelado de forma a comprimir um grande volume de dados, como o feito na pesquisa de Schulte (2022), reduzindo, assim, o custo computacional para manipular esse novo conjunto de dados;
- foi realizada a extração de características relevantes, semelhantemente ao trabalho de Nascimento *et al.* (2020), para formar um espaço latente de menor dimensionalidade. Este espaço concentra as informações e características relevantes, formando grupos mais compactos e de menor dimensionalidade, assim como no trabalho de Costa (2022);
- o espaço latente foi utilizado para inferências, como na pesquisa de Menegolla (2019). No entanto, as inferências, neste caso, são focadas na recriação dos dados de entrada com base nas características mais relevantes nos espaços latentes;
- assim como em Medeiros Campos *et al.* (2021), os resultados foram avaliados. Neste caso, a avaliação foi em relação aos dados de entrada e os reconstruídos, determinando, assim, o erro obtido.

3 METODOLOGIA

Neste capítulo, discutem-se a abordagem de desenvolvimento, a classificação do trabalho, os materiais que foram utilizados e o método.

3.1 Abordagem de desenvolvimento

Nesta monografia, foi adotada uma abordagem de desenvolvimento que incorporou a técnica de prototipagem, conforme discutida por Pressman (2016).

A abordagem de desenvolvimento iniciou-se com a fase de comunicação, na qual os requisitos foram coletados e discutidos detalhadamente. Em seguida, entrou-se na fase de planejamento, onde foram definidos os objetivos e escopo do protótipo. Posteriormente, a fase de modelagem e projeto que envolveu a criação da estrutura a ser realizada, seguida pela fase de construção do protótipo. Nesta etapa, o protótipo foi desenvolvido com base nos requisitos definidos. Após a construção do protótipo, a fase de *feedback* foi feita avaliando-se os resultados do modelo. Esse ciclo iterativo de criação, avaliação e refinamento foi repetido até que os resultados finais atendessem plenamente às necessidades e expectativas estabelecidas no início da pesquisa.

3.2 Classificação do trabalho

De acordo com a classificação apresentada por Gerhardt e Silveira (2009), a classificação desta pesquisa é a seguinte:

- natureza da pesquisa:

A pesquisa é classificada como aplicada porque busca aplicar e desenvolver técnicas específicas; nesse caso, utilizando RNAs com a arquitetura de AE para lidar com o problema da alta dimensionalidade dos dados climáticos. A pesquisa tem o objetivo de resolver um problema prático e contribuir para a área de estudo.

- objetivo da pesquisa:

A pesquisa é classificada como exploratória porque tem como objetivo explorar e investigar o uso de RNAs com arquitetura de AE na redução de dimensionalidade dos dados climáticos, buscando-se obter conhecimento sobre a aplicação dessa técnica e sua eficácia na reconstrução dos dados.

- procedimentos da pesquisa:

A pesquisa é classificada como experimental porque envolve a coleta de dados climáticos, a aplicação dos procedimentos, utilizando a arquitetura de AE, e a avaliação da qualidade da reconstrução dos dados. A pesquisa segue uma abordagem empírica, testando e experimentando a aplicação da técnica em um contexto específico.

- abordagem da pesquisa:

A pesquisa é classificada como quantitativa porque utiliza métodos e técnicas quantitativas para analisar os dados e avaliar a qualidade da reconstrução dos dados climáticos. A avaliação é baseada em métricas numéricas, como o EQM, entre os dados originais e os dados reconstruídos, buscando resultados objetivos e mensuráveis.

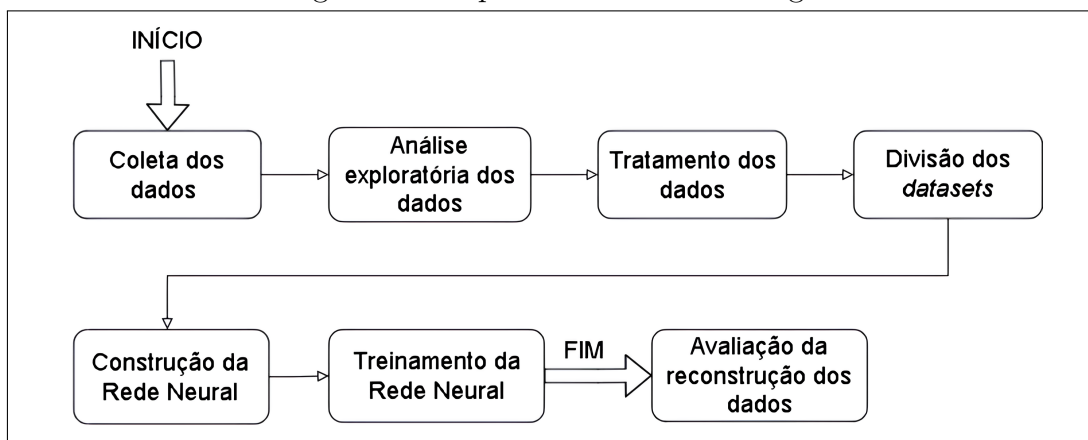
3.3 Materiais

Para a realização do presente trabalho, foi utilizado como *hardware* um computador, com sistema operacional *Windows* 10, de 64 bits, processador *AMD Ryzen™* 5 3500U, com placa gráfica *AMD Radeon™* 540x e memória RAM de 8 GB. A linguagem *Python*, versão 3.11, foi a escolhida para realizar os experimentos.

3.4 Método

Nesta seção, será detalhado todo o planejamento realizado com base nas etapas da Figura 6.

Figura 6 – Esquemática da metodologia



Fonte: Elaborado pelo autor, 2023.

Primeiramente, foi feita a revisão da literatura buscando-se conceitos, fundamentos e trabalhos científicos com algum tipo de semelhança com esta pesquisa. Esta parte é detalhada com mais informações no Capítulo 2.5.

Após a revisão da literatura, foi feito, então, o estudo dos AEs, visando entender como é a lógica dos algoritmos, suas principais características, seus benefícios e quando eles são mais utilizados.

A partir do estudo dos AEs, foi, então, selecionado qual mais se enquadrava para a tarefa de redução de dimensionalidade dos dados climáticos. Isto foi feito avaliando-se algumas arquiteturas e hiper-parâmetros disponíveis.

3.4.1 *Etapa 1: Coleta dos dados*

Os dados climáticos presentes nos *datasets* que foram utilizados para realizar os experimentos deste trabalho foram coletados no site do INMET (2023).

Essa abordagem permitiu a obtenção de um conjunto de dados provenientes de uma fonte confiável e especializada em monitoramento meteorológico.

3.4.2 *Etapa 2: Análise exploratória dos dados*

Inicialmente, foi realizada a análise exploratória dos dados, visando identificar quais tipos de tratamentos de dados poderiam ser realizados posteriormente no *dataset*. A finalidade era obter *insights* iniciais, entendendo a estrutura dos dados e identificando possíveis tratamentos destes. Essa etapa de análise foi constituída por diversas técnicas para se compreender o comportamento dos dados, suas características e seus padrões. As técnicas utilizadas nessa etapa foram:

- a sumarização dos dados:

Onde foram calculadas medidas estatísticas básicas, como média, mediana, desvio padrão, mínimo e máximo, para, assim, se ter uma visão geral dos dados.

- visualização gráfica:

Feita por meio de *plots* de gráficos para representar os dados em uma forma visual, ajudando, desse modo, a identificar tendências, padrões, *outliers* e distribuições dos dados.

- identificação dos dados faltantes:

Verificou-se se o *dataset* possuía dados ausentes ou algum outro tipo de dado que poderia ser considerado como ruído.

3.4.3 *Etapa 3: Tratamento dos dados*

Para um bom funcionamento da rede, é necessário garantir que a base de dados passe por um adequado tratamento. Isto foi possível graças à análise exploratória dos dados, que identificou possíveis problemas e determinou quais tratamentos seriam necessários.

Inicialmente, realizou-se o tratamento dos dados faltantes, onde aqueles ausentes, identificados na seção, foram tratados. Esse tratamento pôde ser efetuado por meio de técnicas de interpolação, preenchendo-se os valores faltantes.

Em seguida, procedeu-se com o escalonamento, onde os dados foram dimensionados para uma escala comum, de 0 até 1. O objetivo foi colocar todas as variáveis em uma escala comparável, eliminando as diferenças de magnitude entre elas. Dessa forma,

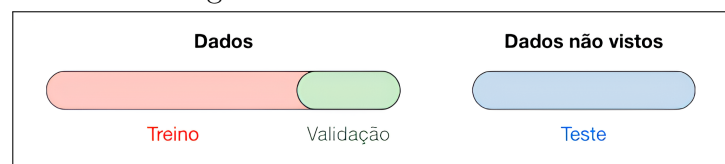
permitiu-se que as diferentes variáveis tivessem uma contribuição equilibrada durante a análise e modelagem de dados. Esse tratamento foi realizado de forma individual, para cada variável, a fim de garantir que todas fossem tratadas adequadamente e seguindo suas particularidades.

Ao se realizar o adequado tratamento da base de dados, a qualidade e a confiabilidade dos dados utilizados para treinar a RNA aumentam, contribuindo para melhorar o desempenho e obter resultados mais precisos durante o treinamento e a posterior aplicação do modelo treinado.

3.4.4 *Etapa 4: Divisão dos datasets nos conjuntos de treino, validação e teste*

Nesta etapa, os dados tratados foram divididos conforme ilustrado na Figura 7, sendo utilizados para realizar o treinamento, a validação (do modelo treinado) e o teste (do modelo treinado e validado). Para evitar vieses e garantir uma avaliação mais confiável do modelo, a divisão do *dataset* em conjuntos de treino e validação foi realizada de forma aleatória.

Figura 7 – Divisão do *dataset*



Fonte: AMIDI; AMIDI, 2018.

3.4.5 *Etapa 5: Construção da Rede Neural*

Nesta etapa, foi construída a rede neural com arquitetura AE. A RNA construída neste ponto foi provisória, pois, na próxima subseção, foi feito o treinamento do modelo, e, em cada instanciação deste, alguns hiper-parâmetros foram substituídos.

A construção da Rede Neural envolveu os seguintes passos:

1. a escolha do número de camadas ocultas:

Nesse passo, foram definidas quantas camadas a rede tem entre as camadas de entrada e saída.

2. a escolha da quantidade de neurônios em cada camada:

Determinou-se a quantidade de neurônios em cada camada oculta. As camadas de entrada e saída possuem a quantidade de neurônios equivalente ao número de atributos do *dataset*, sendo estes recebidos como entrada pela rede.

3. definição da função de ativação:

Nesse passo, foi necessário definir uma função de ativação que introduza não linearidade nas saídas dos neurônios.

4. inicialização dos pesos:

Os pesos da rede neural foram inicializados de forma que sempre fossem os mesmos. Isto para que os valores que indicam desempenho não mudassem em cada instanciação da rede com os mesmos valores, permitindo um monitoramento mais eficaz.

5. definição da função de perda (*loss function*):

Foi definida a função de perda utilizada para medir quão bem o modelo está performando em relação aos dados de treinamento.

6. a escolha do otimizador:

Escolheu-se o otimizador responsável por ajustar os pesos da rede durante o treinamento, utilizando-se um algoritmo de otimização.

7. a escolha da taxa de aprendizagem:

Neste passo, foi definido o tamanho do passo que o algoritmo de otimização toma ao ajustar os pesos da rede durante o processo de treinamento.

8. a escolha da quantidade de épocas máxima:

Nesta etapa, determinou-se o número máximo de iterações ou passagens completas pelo conjunto de treinamento que o algoritmo executou antes de interromper o processo de treinamento.

9. o tamanho do *batch*:

Foi definido, neste passo, o número de exemplos de treinamento processados em paralelo antes de se atualizar os pesos da rede neural. Ou seja, o número de amostras de dados propagados pela rede antes que uma atualização dos pesos seja realizada com base no erro calculado.

10. o valor da paciência:

Nesta etapa, determinou-se o hiper-parâmetro que foi usado na RNA para controlar a parada antecipada do treinamento, com base na melhoria do desempenho do modelo. Se o desempenho estagnar por um número especificado de épocas (paciência), o treinamento é interrompido, assumindo que o modelo já alcançou seu ponto ótimo.

3.4.6 Etapa 6: Treinamento da Rede Neural

Com a RNA parametrizada, o próximo passo foi alimentá-la com os dados de treino e executar o processo de treinamento. Durante o treinamento, os pesos da rede foram atualizados iterativamente com base nas entradas e saídas esperadas. Isto foi feito por meio do otimizador e da função de perda escolhida. O otimizador calculou os gradientes dos pesos em relação à função de perda e realizou os ajustes necessários.

Durante o treinamento, foi necessário ajustar alguns dos hiper-parâmetros da rede já definidos na etapa anterior (Subseção 3.4.5). Esses ajustes foram executados com base no desempenho da rede, no monitoramento da função de perda escolhida e no gráfico de convergência do erro do conjunto de teste em relação ao conjunto de validação. A curva de erro da validação deve se comportar como a curva de erro do treinamento, indicando, assim, que os valores dos erros nos dois conjuntos foram parecidos. Em outras palavras, o modelo foi validado.

Cada ciclo de treinamento e validação foi finalizado quando a rede neural atingiu o critério de parada predefinido. Neste caso, a parada ocorreu quando atingiu o número máximo de épocas ou o limite de melhoria no desempenho (paciência).

E por fim, após essa validação, foi avaliado o desempenho da rede neural em dados não vistos utilizando-se o conjunto de teste. Isso permitiu estimar a capacidade de generalização da rede e verificar se ela estava apta a realizar a tarefa desejada, com boa precisão, em novos dados. Caso esta avaliação apontasse que o modelo não generalizava bem, então ele não poderia ser utilizado em novos dados.

3.4.7 Etapa 7: Avaliação da reconstrução dos dados

Para avaliar a reconstrução dos dados, foi feita a soma por atributo de todas as amostras dos dados esperados e dos dados previstos. Em seguida, eles foram comparados, resultando em uma medida de diferença total entre os valores esperados e os valores previstos. Essa diferença indica, para este trabalho, o grau de discrepância geral entre os dois conjuntos de dados. Quando essas medidas de diferenças totais calculadas são iguais, significa que a previsão do modelo está perfeitamente alinhada com o valor da soma dos valores esperados, sem nenhuma diferença total. Por outro lado, quando a soma dos dados previstos foi maior ou menor que a soma dos dados esperados, indicou que havia uma diferença total positiva ou negativa, respectivamente.

No entanto, como a comparação das somas totais não fornece informações detalhadas sobre as discrepâncias individuais entre os dados, foi feita, também, a avaliação da qualidade da reconstrução dos dados climáticos utilizando-se o EQM, possibilitando, assim, uma visão mais completa da qualidade das previsões do modelo gerado.

A avaliação da qualidade da reconstrução dos dados climáticos, utilizando o EQM, foi feita entre a saída reconstruída e a entrada original. Se o erro entre esses dois

pontos for zero, então a reconstrução foi perfeita, e os dados, reproduzidos exatamente como na entrada. Quanto menor o erro, melhor a capacidade do AE de reconstruir os dados climáticos de entrada de forma precisa.

4 EXPERIMENTOS E RESULTADOS

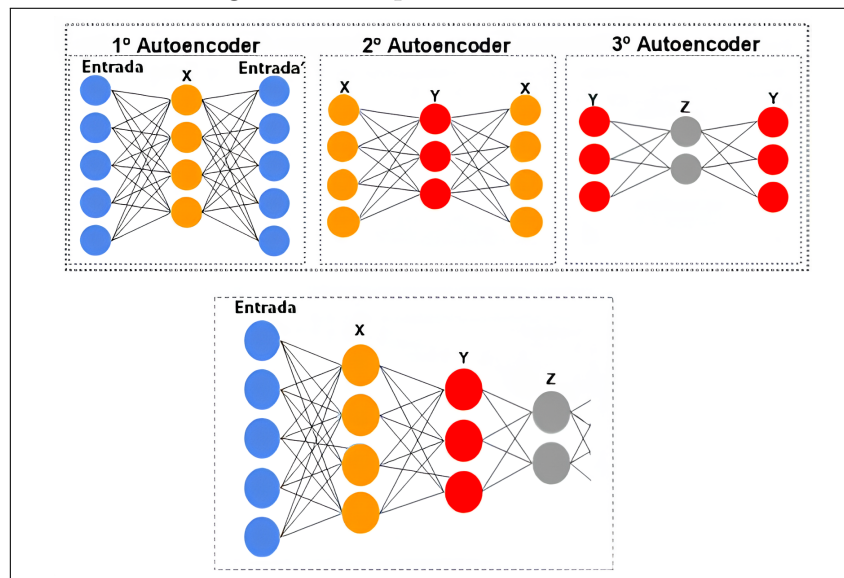
Neste capítulo, serão apresentadas as ferramentas utilizadas para o desenvolvimento do presente trabalho, os experimentos e os resultados gerados durante a aplicação da metodologia.

4.1 Escolha do AE: SAE

O AE escolhido foi o SAE, pois, segundo Géron (2017), é um tipo de rede neural que usa todos os benefícios de uma RNA profunda para aprender dados úteis sobre agrupamento hierárquico. Em outras palavras, o SAE consegue identificar e aprender padrões ou grupos de características em diferentes níveis de abstração. Ao contrário de apenas extrair características simples dos dados, o SAE pode identificar aspectos mais complexos e abstratos, os quais são construídos a partir de combinações de características mais simples. Desse modo, possui a capacidade de aprender padrões, sendo esta a principal característica desta rede neural profunda. As múltiplas camadas desta rede representam os diversos níveis de abstração dos dados.

Conforme mencionado por Santos Mello (2019), a arquitetura do SAE é baseada na conexão entre AE, composto por um *encoder* responsável pela codificação e um *decoder* responsável pela decodificação. A Figura 8 ilustra essa arquitetura.

Figura 8 – Arquitetura de um SAE



Fonte: SANTOS MELLO, 2019.

4.2 Ferramentas para implementação

A linguagem *Python* foi a utilizada para a implementação do código, sendo a biblioteca *tensorflow* a mais importante nesse caso. Essa biblioteca possui uma *Application*

Programming Interface (API) essencial para a implementação da rede, que é a API *Keras*, a qual é uma API de alto nível para construção e treinamento de RNAs, utilizada para realizar uma implementação de forma mais simples, em comparação com a implementação direta usando o *TensorFlow* em baixo nível. Esta escolha facilitou as mudanças que precisaram ser feitas na fase de experimentação nos hiper-parâmetros ao longo do treinamento da rede. O *Keras* permitiu, por exemplo, definir as camadas, o otimizador e a função de perda. Outra importante biblioteca é a *sklearn*, que fornece algoritmos e ferramentas para pré-processamento de dados, divisão dos recursos e avaliação de desempenho.

4.3 Coleta dos dados

Foram coletados dois *dataset* climáticos no site INMET, referentes à cidade de Bambuí. Um deles foi coletado em intervalos de uma hora, ao longo do ano de 2020, com um total de 8784 amostras e 17 atributos, e o outro, em intervalos de uma hora, ao longo do ano de 2022, com um total de 8760 amostras e 17 atributos.

Estes *dataset* climáticos referentes à cidade de Bambuí, coletados no site INMET, foram selecionados por serem os mais recentes, que apresentam um menor índice de dados faltantes em comparação a outros períodos. Estes critérios visam assegurar a representatividade e a integridade das informações climáticas, uma vez que a menor presença de dados faltantes contribui para uma análise mais robusta e precisa, evitando possíveis distorções na modelagem e nas inferências obtidas a partir dos dados climáticos. Essa abordagem maximiza a confiabilidade e a relevância dos resultados obtidos a partir destes dados.

4.4 Análise exploratória dos dados

Nesta seção, será feita uma breve análise dos dados.

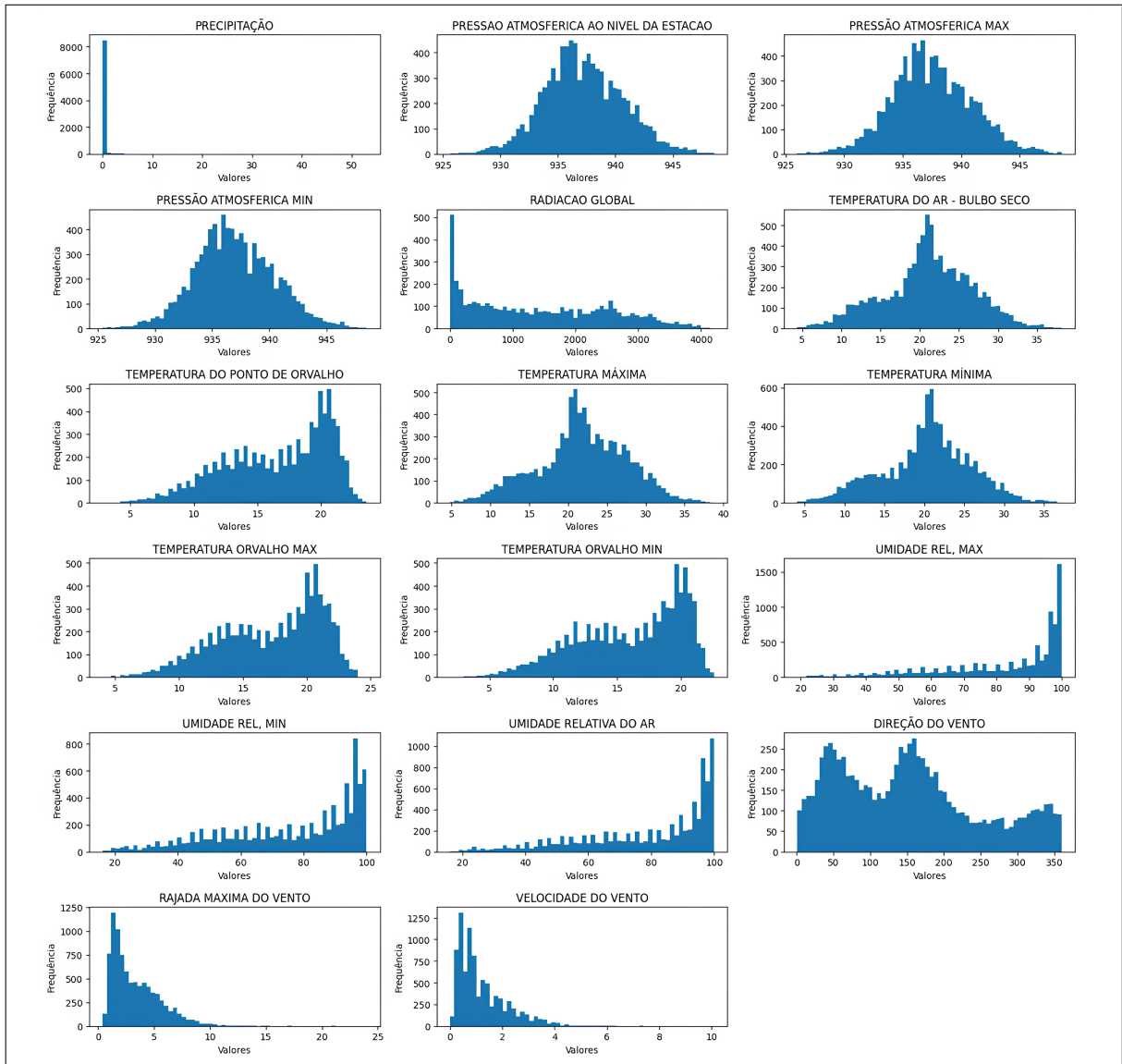
4.4.1 *Descobertas iniciais*

Observou-se, por meio do histograma de cada atributo presente na Figura 9, a distribuição dos dados no *dataset* de 2020.

Verificou-se que algumas técnicas de pré-processamento, como a remoção de *outliers* e a normalização (estabilização de variância), não apresentaram benefícios no treinamento da rede AE para a tarefa de reconstrução dos dados climáticos. Os dados sem estes pré-processamentos foram mais úteis à rede neural para capturar os padrões necessários. Os motivos identificados são os seguintes:

- Remoção de *outliers*: os *outliers*, nestes dados climáticos, representam eventos climáticos considerados atípicos, raros ou extremos, como tempestades severas e

Figura 9 – Distribuição dos dados de 2020



Fonte: Elaborado pelo autor, 2023.

ondas de calor. Sendo estas partes da variabilidade natural do clima, removê-los poderia distorcer a representação real dos dados e eliminaria alguns padrões climáticos, prejudicando o desempenho da rede neural.

- Normalização: como pode ser observado na Figura 9, alguns dos dados climáticos não seguem uma distribuição normal, apresentando características específicas em determinadas épocas do ano. Ao forçar a normalização, pode-se distorcer a representação adequada dos padrões climáticos, levando à perda de informações importantes dos dados, pois características distintas dos eventos climáticos podem ser suavizadas ou completamente perdidas. Posteriormente, caso o modelo da rede neural fosse treinado com estas informações, ela não seria capaz de generalizar para novos dados.

Foi feita, também, a sumarização dos dados verificando a média, mediana, desvio

padrão, mínimo e máximo, para, assim, ter uma visão geral dos dados. A sumarização do *dataset* de 2020 pode ser conferida na Tabela 1, e o de 2022, na Tabela 2.

Tabela 1 – Sumarização do *dataset* de 2020

	Máximo	Mínimo	Média	Mediana	Desvio Padrão
Precipitação	53	0	0.2	0	1.5
Pressão Atmosférica ao Nível da Estação	948.6	925.6	937.2	937	3.4
Pressão Atmosférica Máxima	948.6	926	937.5	937.3	3.4
Pressão Atmosférica Mínima	948.5	925.4	937	936.7	3.5
Radiação Global	4209.3	0	1420.2	1285.1	1102.9
Temperatura do Ar - Bulbo Seco	38.1	4.4	21.1	21.2	5.6
Temperatura do Ponto de Orvalho	23.6	2.8	16.5	17.4	4.1
Temperatura Máxima	38.8	4.8	21.8	21.8	5.7
Temperatura Mínima	37.2	4.1	20.4	20.7	5.5
Temperatura Orvalho Máxima	24.7	4	17.2	18	4
Temperatura Orvalho Mínima	22.6	1.9	16	16.8	4.1
Umidade Real Máxima	100	19	82.4	91	19.2
Umidade Real Mínima	100	16	75.6	82	21.3
Umidade Relativa do Ar	100	16	79.1	87	20.4
Direção do Vento	360	1	152	145	95.9
Rajada Máxima do vento	24	0.4	3.4	2.7	2.3
Velocidade do Vento	10.1	0	1.2	0.9	1

Fonte: Elaborado pelo autor, 2023.

4.4.2 Verificando dados faltantes

Nesta etapa, efetuou-se uma verificação nos *datasets* buscando encontrar dados faltantes. É possível observar os resultados dessa verificação na Tabela 3, referente ao *dataset* de 2020, e na Tabela 4, relativa ao *dataset* de 2022.

Pode-se notar, na terceira coluna das tabelas, que os atributos com dados faltantes representam uma porcentagem muito pequena. Estes dados são insignificantes em relação ao tamanho total da amostra, não devendo afetar negativamente as análises ou modelos gerados.

Já a variável radiação global tem um alto percentual de dados faltantes; contudo, eles ocorrem consistentemente em todos os anos nos *datasets*, apontando que essa inconsistência, em diferentes anos, tem sido um padrão estável. No contexto desta pesquisa, que envolve o treinamento de uma rede neural para compactar e reconstruir dados de entrada a partir de uma representação latente, a inconsistência na falta de dados não afeta adversamente o treinamento da rede. A RNA é capaz de aprender, a partir dos dados

Tabela 2 – Sumarização do *dataset* de 2022

	Máximo	Mínimo	Média	Mediana	Desvio Padrão
Precipitação	30.4	0	0.2	0	1.4
Pressão Atmosférica ao Nível da Estação	946.3	923.7	936.8	936.7	3.3
Pressão Atmosférica Máxima	946.3	924.2	937.1	937	3.3
Pressão Atmosférica Mínima	946	923.7	936.5	936.4	3.3
Radiação Global	4056.3	0	1444.4	1322.5	1100.7
Temperatura do Ar - Bulbo Seco	35.9	3.8	20.7	20.8	5.8
Temperatura do Ponto de Orvalho	23.7	-1.2	16	17.1	4.3
Temperatura Máxima	36.6	4.2	21.5	21.4	5.9
Temperatura Mínima	34.8	3.4	20	20.3	5.7
Temperatura Orvalho Máxima	24.3	0.3	16.6	17.7	4.3
Temperatura Orvalho Mínima	22.8	-2.6	15.4	16.5	4.4
Umidade Real Máxima	100	19	81.7	91	20.4
Umidade Real Mínima	100	16	74.7	80	22.5
Umidade Relativa do Ar	100	16	78.2	86	21.6
Direção do Vento	360	1	151.2	143	98.1
Rajada Máxima do vento	36.5	0.4	3.5	2.9	2.4
Velocidade do Vento	9.3	0	1.3	1	1

Fonte: Elaborado pelo autor, 2023.

disponíveis e, dada a estabilidade desse padrão, a rede consegue identificá-lo e aprendê-lo com eficácia, não afetando as análises e modelos gerados.

Estes dados faltantes são tratados na próxima seção por meio de um preenchimento artificial.

4.5 Tratamento dos dados

Nesta seção, é descrito como o tratamento dos dados foi feito.

4.5.1 *Tratamento dos dados faltantes*

Os dados faltantes, identificados na Subseção 4.4.2, foram tratados dimensão por dimensão através da técnica de interpolação por vizinhos mais próximos. Esta técnica foi executada utilizando-se um método em *Python* chamado *interpolate*, que faz parte da biblioteca *pandas*. Técnica esta que foi utilizada em todos os dados, com exceção dos dois primeiros e últimos dados faltantes identificados que estavam no início ou no fim da coluna. Estes foram substituídos pela média da coluna, pois esta técnica de interpolação não consegue interpolar muito bem os primeiros e os últimos valores ausentes. Isto ocorre

Tabela 3 – Dados faltantes no *dataset* de 2020

	Dados faltantes	Percentual (%)
Precipitação	34	0.387
Pressão Atmosférica ao Nível da Estação	34	0.387
Pressão Atmosférica Máxima	34	0.387
Pressão Atmosférica Mínima	34	0.387
Radiação Global	4082	46.471
Temperatura do Ar - Bulbo Seco	34	0.387
Temperatura do Ponto de Orvalho	34	0.387
Temperatura Máxima	34	0.387
Temperatura Mínima	34	0.387
Temperatura Orvalho Máxima	34	0.387
Temperatura Orvalho Mínima	34	0.387
Umidade Real Máxima	34	0.387
Umidade Real Mínima	34	0.387
Umidade Relativa do Ar	34	0.387
Direção do Vento	71	0.808
Rajada Máxima do vento	35	0.398
Velocidade do Vento	35	0.398

Fonte: Elaborado pelo autor, 2023.

Tabela 4 – Dados faltantes no *dataset* de 2022

	Dados faltantes	Percentual (%)
Precipitação	168	1.918
Pressão Atmosférica ao Nível da Estação	168	1.918
Pressão Atmosférica Máxima	168	1.918
Pressão Atmosférica Mínima	168	1.918
Radiação Global	4090	46.69
Temperatura do Ar - Bulbo Seco	168	1.918
Temperatura do Ponto de Orvalho	168	1.918
Temperatura Máxima	168	1.918
Temperatura Mínima	168	1.918
Temperatura Orvalho Máxima	168	1.918
Temperatura Orvalho Mínima	168	1.918
Umidade Real Máxima	168	1.918
Umidade Real Mínima	168	1.918
Umidade Relativa do Ar	168	1.918
Direção do Vento	196	2.237
Rajada Máxima do vento	168	1.918
Velocidade do Vento	169	1.929

Fonte: Elaborado pelo autor, 2023.

devido à falta de pontos de referência à esquerda (para os primeiros) ou à direita (para os últimos) para calcular os valores interpolados.

Esta técnica de interpolação por vizinhos mais próximos, utilizada nesta pesquisa, consiste em procurar os valores não faltantes mais próximos em relação aos pontos de dados com valores ausentes, verificados os valores disponíveis tanto antes (esquerda)

quanto depois (direita) do ponto com valor ausente. O método, então, seleciona o valor mais próximo encontrado (dando preferência ao valor à esquerda) e o atribui ao ponto com valor ausente, preenchendo, assim, os valores faltantes. Este processo é repetido até não sobraem valores ausentes no conjunto de dados.

4.5.2 Escalonamento

Nesta etapa, foi feito o escalonamento dos dados, colocando dimensão por dimensão dos conjuntos de dados em uma escala de 0 a 1. Este processo foi feito porque as variáveis possuem intervalos de valores distintos (é possível verificar isso nas Tabelas 1 e 2). Os valores de mínimo e máximo de cada atributo diferem. Então, aplicar a mesma transformação a todas as variáveis poderia distorcer a distribuição dos dados e comprometer a interpretação e o desempenho da RNA.

4.6 Divisão dos *datasets* nos conjuntos de treino, validação e teste

O *dataset* contendo os dados de 2020 foi dividido em dois conjuntos, um de treino, contendo 70% dos dados, e outro de validação, contendo 30% dos dados.

O *dataset* contendo as amostras do ano de 2022 foi reservado para realizar o teste no modelo treinado e validado, tendo estes dados como principal objetivo, e também avaliar o desempenho do modelo em dados não vistos previamente, verificando sua capacidade de generalização.

Após a divisão dos dados, eles ficaram com as respectivas quantidades ilustradas na Tabela 5.

Tabela 5 – Divisão dos *datasets*

	Quantidade de amostras	Percentual (%)
Dados de treino	6148	70
Validação	2636	30
Teste	8760	100

Fonte: Elaborado pelo autor, 2023.

4.7 Construção da Rede Neural

Seguindo os passos preestabelecidos na Seção 3.4.5 da metodologia, a RNA foi construída:

1. Escolha do número de camadas ocultas: 7 camadas ocultas;
2. Escolha da quantidade de neurônios em cada camada: as camadas de entrada e de saída foram configuradas com 17 neurônios; a primeira camada oculta 16 neurônios; 8, na segunda; 6, na terceira; 1, na quarta; 6, na quinta; 8, na sexta; e 16, na sétima.

Foi utilizada, também, a camada *dropout* após cada uma das camadas ocultas, com o valor de 10%, fazendo com que, durante o treinamento de cada lote (*batch*) de dados, 10% dos neurônios fossem temporariamente ignorados.

3. Definição da função de ativação: a função de ativação configurada foi a *sigmoid* em ambas as camadas ocultas.
4. Inicialização dos pesos: os pesos da rede neural foram inicializados utilizando-se a função *LecunNormal*, que segue uma distribuição normal.
5. Definição da função de perda (*loss function*): foi utilizada a função de perda EQM para medir o desempenho do modelo em relação aos dados de treinamento.
6. Escolha do otimizador: utilizou-se o otimizador "Adam" para ajustar os pesos da rede durante o treinamento.
7. Escolha da taxa de aprendizagem: foi definida a taxa de aprendizagem como 0.007.
8. Escolha da quantidade de épocas máxima: o número máximo de épocas definido foi 2000.
9. Tamanho do (*batch*): o tamanho do (*batch*) foi definido como 100.
10. Valor da paciência: o valor da paciência foi definido como 100.

A camada latente que foi produzida com essa configuração da topologia do modelo foi um vetor com uma posição. Ou seja, os 17 atributos foram compactados para uma representação com 1 atributo.

4.8 Treinamento da Rede Neural

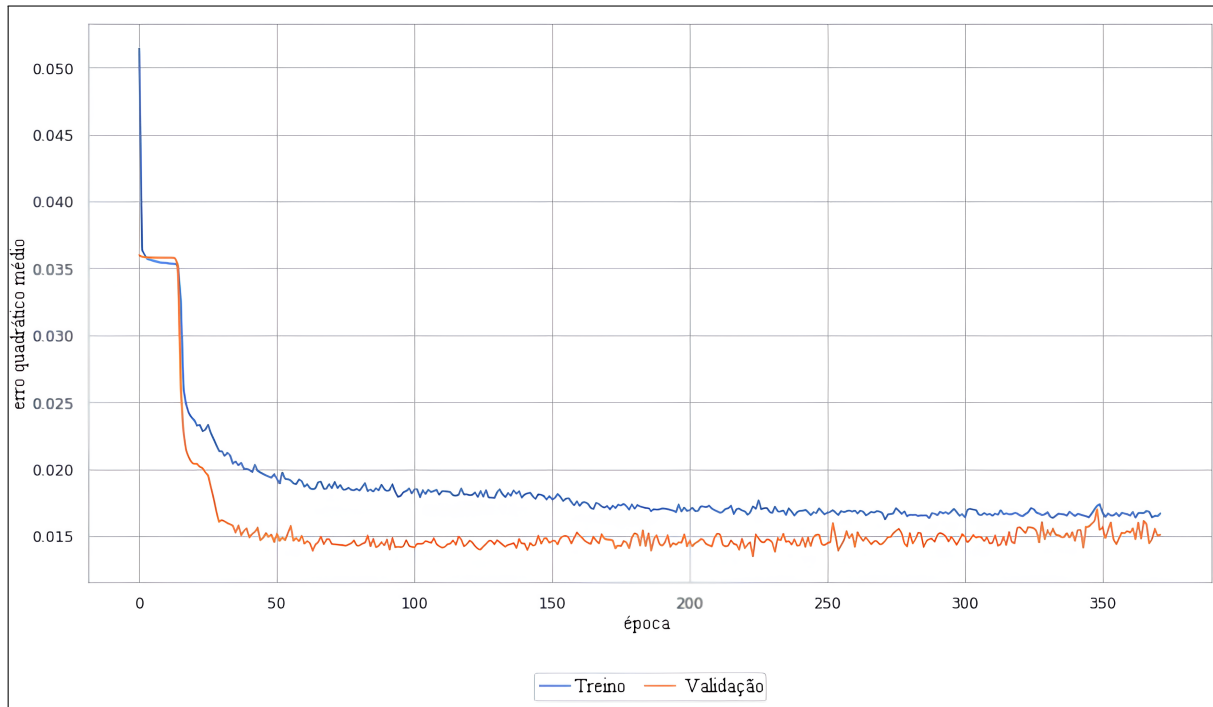
Os valores definidos na seção anterior para a construção da rede são os valores finais, que foram obtidos após as avaliações empíricas com a rede, com o objetivo de produzir o melhor modelo utilizando apenas um valor latente, não representando os hiper-parâmetros iniciais. Os gráficos e as métricas mencionados nesta seção são referentes a estes valores finais, que já foram avaliados.

O treinamento foi feito utilizando-se o conjunto de dados de treino que foi estabelecido na Seção 4.6. Durante este processo, os pesos da rede foram atualizados iterativamente com base nas entradas e nas saídas esperadas. Isso foi feito utilizando-se o otimizador escolhido e a função de perda definida anteriormente. O otimizador calculou os gradientes dos pesos, em relação à função de perda, e realizou os ajustes necessários.

Durante o treinamento, foi necessário ajustar os hiper-parâmetros da rede com base no desempenho, monitorando-se a métrica EQM escolhida e observando-se o gráfico de convergência do erro no conjunto de validação em relação ao conjunto de treinamento.

Idealmente, a curva de erro da validação deve seguir a curva de erro do treinamento, indicando que o modelo foi validado corretamente. A Figura 10 representa a convergência destes erros, referentes à topologia configurada na Seção 4.7 durante a construção do modelo, sendo esta uma configuração que satisfaz o critério.

Figura 10 – Convergência do erro com uma representação latente



Fonte: Elaborado pelo autor, 2023.

Cada ciclo de treinamento e validação foi finalizado quando a rede neural atingiu o critério de parada pré-definido, ou seja, quando atingiu o número máximo de épocas igual a 2000 ou quando a paciência atingiu 100.

Após a validação do modelo, foi feita uma avaliação do desempenho da rede neural utilizando o conjunto de teste, que são os dados não vistos durante o treinamento, ou seja, o conjunto de dados referente ao ano de 2022. Esse processo permitiu estimar a capacidade de generalização da rede e verificar que ela é capaz de realizar a tarefa desejada com precisão em novos dados. O resultado dessa avaliação utilizando o EQM nos dados de teste foi igual a 0.01605. Como esse valor é o previsto conforme pode ser observado no gráfico de convergência da Figura 10, não destoando daquilo que era esperado, então pode-se confirmar que o modelo generaliza bem para estes dados, denotando que ele pode ser utilizado com confiança em novas situações.

4.9 Avaliação da reconstrução dos dados

Para avaliar a reconstrução dos dados de entrada, foi feita a comparação da medida da soma total de cada atributo entre os valores esperados e os valores previstos,

resultando na medida de diferença total.

O valor da medida da soma total de cada atributo dos dados de entrada pode ser conferido na Tabela 6, e essa medida dos dados reconstruídos pode ser conferida na Tabela 7 .

Tabela 6 – Soma total de cada atributo de entrada

	Soma total
Precipitação	9.06604
Pressão Atmosférica ao Nível da Estação	8.62379
Pressão Atmosférica Máxima	9.21286
Pressão Atmosférica Mínima	9.42804
Radiação Global	9.29138
Temperatura do Ar - Bulbo Seco	9.14445
Temperatura do Ponto de Orvalho	8.59845
Temperatura Máxima	8.63282
Temperatura Mínima	8.6608
Temperatura Orvalho Máxima	8.58049
Temperatura Orvalho Mínima	5.80388
Umidade Real Máxima	9.17512
Umidade Real Mínima	9.31435
Umidade Relativa do Ar	9.28119
Direção do Vento	9.55757
Rajada Máxima do vento	9.08618
Velocidade do Vento	9.23634

Fonte: Elaborado pelo autor, 2023.

Tabela 7 – Soma total de cada atributo reconstruído

	Soma total
Precipitação	8.7155
Pressão Atmosférica ao Nível da Estação	8.79077
Pressão Atmosférica Máxima	8.68182
Pressão Atmosférica Mínima	8.71295
Radiação Global	8.69655
Temperatura do Ar - Bulbo Seco	8.68492
Temperatura do Ponto de Orvalho	8.5567
Temperatura Máxima	8.54761
Temperatura Mínima	8.54448
Temperatura Orvalho Máxima	8.5336
Temperatura Orvalho Mínima	8.54907
Umidade Real Máxima	8.65915
Umidade Real Mínima	8.68162
Umidade Relativa do Ar	8.79063
Direção do Vento	8.61709
Rajada Máxima do vento	8.61517
Velocidade do Vento	8.59408

Fonte: Elaborado pelo autor, 2023.

A medida da diferença total obtida ao comparar esses valores pode ser conferida na Tabela 8, sendo que estes valores estão em forma de porcentagem. Essa diferença indica o grau de discrepância geral entre os dois conjuntos de dados. Caso as diferenças totais fossem iguais, indicaria que a previsão do modelo está perfeitamente alinhada com o valor da soma dos valores esperados, sem nenhuma diferença total. Por outro lado, a soma dos dados previstos, quando é maior (acima de 100%) ou menor (abaixo de 100%) que a soma dos dados esperados, indica que há uma diferença total positiva ou negativa, respectivamente.

Tabela 8 – Diferença do total de cada atributo

	Diferença total (%)
Precipitação	96.1335
Pressão Atmosférica ao Nível da Estação	101.936
Pressão Atmosférica Máxima	94.236
Pressão Atmosférica Mínima	92.4153
Radiação Global	93.5981
Temperatura do Ar - Bulbo Seco	94.9748
Temperatura do Ponto de Orvalho	99.5144
Temperatura Máxima	99.0129
Temperatura Mínima	98.6569
Temperatura Orvalho Máxima	99.4535
Temperatura Orvalho Mínima	97.1058
Umidade Real Máxima	94.3764
Umidade Real Mínima	93.2069
Umidade Relativa do Ar	94.7145
Direção do Vento	90.1598
Rajada Máxima do vento	94.9162
Velocidade do Vento	93.0464

Fonte: Elaborado pelo autor, 2023.

No entanto, como a comparação das somas totais não fornece informações detalhadas sobre as discrepâncias individuais entre os dados, para se obter uma visão mais completa da qualidade das previsões do modelo gerado, a avaliação da qualidade da reconstrução dos dados climáticos foi também analisada utilizando-se o EQM já calculado. Essa avaliação foi feita entre a saída reconstruída e a entrada original, gerando um valor de 0.01605. Quanto menor este erro, melhor a capacidade do SAE de reconstruir os dados climáticos de entrada de forma precisa.

4.10 Avaliação da reconstrução dos dados climáticos utilizando diferentes tamanhos de dados latentes

Foi feita a avaliação da reconstrução dos dados climáticos em mais quatro diferentes representações latentes. Para criar essas representações, empregou-se a mesma

configuração da topologia da RNA preestabelecida anteriormente, mudando apenas a quarta camada (a camada do meio), para os valores dois, três, quatro e cinco.

Os gráficos de convergência utilizados para validar estas representações latentes estão disponíveis no Apêndice A. O valor do EQM para cada uma dessas configurações, ao serem testadas nos conjuntos de dados não vistos pela rede, pode ser conferido na Tabela 9. Estes valores são o que foi previsto, conforme pode ser observado em seus respectivos gráficos de convergência, não destoando daquilo que era esperado. Então, pode-se confirmar que o modelo generaliza bem para estes dados, denotando que ele pode ser utilizado com confiança em novas situações caso o respectivo erro possa ser tolerado.

É importante lembrar que dizer que o modelo generaliza bem não significa dizer que os valores dessas generalizações foram os melhores possíveis, por exemplo, no caso da representação com três dados latentes, que generaliza bem, contudo fornece erros elevados se comparada com as outras representações.

Tabela 9 – EQM de diferentes representações latentes

Tamanho da representação latente	EQM
1	0.01605
2	0.01416
3	0.01752
4	0.01389
5	0.01202

Fonte: Elaborado pelo autor, 2023.

5 CONCLUSÃO

A rede SAE cumpre sua função de reduzir a base de dados climáticos de entrada em uma representação latente e, a partir dessa representação, reconstruir os dados de entrada o mais próximo possível, com um EQM de 0.01605 para o pior cenário. Tendo esse cenário ocorrido quando os 17 atributos climáticos de entrada foram compactos para uma representação latente de apenas 1 atributo, a partir deste, realizou-se a reconstrução dos 17 dados de entrada, reconstruindo-os o mais próximo possível da entrada original.

Essa redução na complexidade é de grande importância, pois melhora o desempenho em tarefas que utilizam dados de clima e aceitam um EQM equivalente. Além disso, também beneficia a visualização e o armazenamento de grandes conjuntos de dados climáticos.

Utilizando a topologia especificada para apenas uma representação compacta, a rede efetuou boas reduções nos cinco diferentes tamanhos dos dados latentes (dados comprimidos). Contudo, em especial, a redução para a dimensionalidade de três dados foi problemática, apresentando resultados insatisfatórios, pois um espaço latente de maior dimensionalidade pode capturar mais informações para reconstruir os dados de entrada. Nesse caso, ela apresentou resultados piores que aquela com apenas uma representação latente. Para obter resultados melhores e mais satisfatórios, é necessário desenvolver modelos especializados, que sejam capazes de reduzir cada uma dessas dimensionalidades de forma eficiente e precisa.

Os resultados alcançados refletem diretamente na implementação da metodologia elaborada, a qual foi desenhada estrategicamente para garantir que todos os objetivos fossem alcançados, incluindo os específicos. O primeiro objetivo específico foi atingido por meio de uma fase dedicada à coleta e preparação dos dados climáticos, abordando nuances fundamentais para a qualidade da análise. Ao realizar uma seleção criteriosa do tipo de AE mais apropriado, levando em consideração as particularidades do conjunto de dados em questão, alcançou-se o segundo objetivo específico. Por fim, para concluir o terceiro objetivo específico, realizaram-se a busca e o ajuste dos hiper-parâmetros, visando otimizar o desempenho da Rede Neural SAE. O conjunto desses fatores resultou em uma notável capacidade de compressão e reconstrução dos dados climáticos.

As maiores dificuldades enfrentadas durante a elaboração deste trabalho foram a escolha, a configuração e o ajuste adequado dos hiper-parâmetros para o SAE, pois estes valores influenciam diretamente na capacidade de compressão e reconstrução dos dados climáticos.

Os trabalhos futuros podem direcionar-se a diversas áreas de aprimoramento desta pesquisa. Como uma possível continuação deste trabalho, sugere-se reduzir ainda mais o EQM, mantendo a topologia criada, implicando em uma investigação mais aprofundada de configurações e hiper-parâmetros que otimizem a eficiência da compressão sem comprometer a qualidade da reconstrução.

Uma segunda linha de trabalho futuro envolve a comparação dos resultados obtidos nesta pesquisa com outros existentes na literatura. A análise comparativa permitirá avaliar a eficácia relativa da abordagem proposta em relação a métodos estabelecidos, contribuindo para a contextualização e validação dos resultados.

Além disso, existe margem para se investigar alternativas que possam aperfeiçoar ainda mais a técnica. Isso envolve a exploração de diferentes configurações topológicas, como a variação do número de neurônios nas camadas ocultas, diferentes quantidades de camadas ocultas e ajustes nos hiper-parâmetros do SAE. Essas investigações visam identificar configurações mais robustas e adaptáveis, proporcionando *insights* valiosos para o desenvolvimento futuro de técnicas de compressão de dados climáticos.

Em suma, as sugestões para trabalhos futuros buscam aperfeiçoar a abordagem proposta, contextualizá-la em relação a outras técnicas e explorar possíveis modificações que maximizem sua eficiência.

REFERÊNCIAS

- ALSENAN, S.; AL-TURAIKI, I.; HAFEZ, A. Autoencoder-based dimensionality reduction for QSAR modeling. In: IEEE. 2020 3rd International Conference on Computer Applications & Information Security (ICCAIS). 2020. p. 1–4.
- AMIDI, A.; AMIDI, S. **Dicas e truques para aprendizado de máquina**. 2018. <https://stanford.edu/~shervine/l/pt/teaching/cs-229/dicas-truques-aprendizado-maquina>. Acesso em: 02 jul. 2023.
- ATIENZA, R. **Advanced Deep Learning with Keras: Apply deep learning techniques, autoencoders, GANs, variational autoencoders, deep reinforcement learning, policy gradients, and more**. Packt Publishing Ltd, 2018.
- BANIHASHEMI, F.; WEBER, M.; LANG, W. Model order reduction of building energy simulation models using a convolutional neural network autoencoder. **Building and Environment**, Elsevier, v. 207, p. 108498, 2022.
- BOCHIE, K. *et al.* **Aprendizado profundo em redes desafiadoras: Conceitos e aplicações**. Sociedade Brasileira de Computação, 2020.
- BOOKS, G. Google Ngram Viewer. 2023. Disponível em: <https://11nk.dev/TvdIU>. Acesso em: 24 mai. 2023.
- BOQUET, G. *et al.* A variational autoencoder solution for road traffic forecasting systems: Missing data imputation, dimension reduction, model selection and anomaly detection. **Transportation Research Part C: Emerging Technologies**, Elsevier, v. 115, p. 102622, 2020.
- BORGES, H. B.; NIEVOLA, J. **Redução de Dimensionalidade em Bases de Dados de Expressão Gênica**. 2006. Tese (Doutorado) – Dissertação de Mestrado, PPGIa-PUCPR.
- CAMARGO, S. d. S. Um modelo neural de aprimoramento progressivo para redução de dimensionalidade, 2010.
- CHOLLET, F. **Building Autoencoders in Keras**. 2017. Consultado em 01-03-2017. URL: <https://blog.keras.io/building-autoencoders-in-keras.html>.
- CIOFFI, J. R. Sistema de planejamento de voo autônomo utilizando inteligência artificial. Universidade Estadual Paulista (Unesp), 2022.
- COSTA, J. A. d. Autoencoder, análise via componentes principais e independentes aplicados no reconhecimento de padrões de populações. Universidade Federal de Viçosa, 2022.
- GERHARDT, T. E.; SILVEIRA, D. T. **Métodos de pesquisa**. Plageder, 2009.
- GÉRON, A. **Hands-On Machine Learning with Scikit-Learn & TensorFlow**. O'Reilly Media, 2017. ISBN 978-1491962299. Disponível em: <https://www.oreilly.com/library/view/hands-on-machine-learning/9781491962282/>.

GONÇALVES, H. E. **Comparação da eficiência de modelos de redes neurais artificiais na detecção de intrusões em redes de computadores**. 2023. Repositório Institucional - Universidade Federal de Uberlândia. Disponível em: <http://repositorio.ufu.br/handle/123456789/37246>. Acesso em: 22 abr. 2023.

HAYKIN, S. **Redes neurais: princípios e prática**. Bookman Editora, 2001.

INMET. **Instituto Nacional de Meteorologia - Dados Históricos**. Acessado em 2023. 2023. Disponível em: <https://portal.inmet.gov.br/dadoshistoricos>.

KOVÁCS, Z. L. **Redes neurais artificiais**. Editora Livraria da Física, 2002.

LOY, J. **Neural Network Projects with Python: The ultimate guide to using Python to explore the true power of neural networks through six projects**. Packt Publishing Ltd, 2019.

MECHELLI, A.; VIERA, S. **Machine learning: methods and applications to brain disorders**. Academic Press, 2019.

MEDEIROS CAMPOS, M. M. de *et al.* Aprendizagem Profunda Aplicada ao Problema de Contar o Número de Presentes em um Ambiente. **Brazilian Journal of Development**, v. 7, n. 4, p. 35112–35118, 2021.

MENEGOLLA, H. B. Estudos sobre a aplicação de autoencoder para construção de inferências na indústria química, 2019.

MENEZES, S. L.; FREITAS, R. S.; PARPINELLI, R. S. Mineração em grandes massas de dados utilizando hadoop mapreduce e algoritmos bio-inspirados: Uma revisão sistemática. **Revista de Informática Teórica e Aplicada**, v. 23, n. 1, p. 69–101, 2016.

MORAES, L. G. de; ROSSI, R. G. Estudo de Técnicas de Aprendizado Baseado em Uma Única Classe para o Reconhecimento Facial, 2020.

NASCIMENTO, R. S. F. do *et al.* Detecção de falhas com Stacked Autoencoders e técnicas de reconhecimento de padrões em poços de petróleo operados por gas lift. In: 1. CONGRESSO Brasileiro de Automática-CBA. 2020. v. 2.

PIRES, J. M. N. G. **Aprendizagem profunda: estudo e aplicações**. 2017. Diss. (Mestrado) – Universidade de Évora.

PRESSMAN, R. S. **Software Engineering: A Practitioner's Approach**. 8th: AMGH Editora Ltda, 2016.

REZENDE, S. O. **Sistemas inteligentes: fundamentos e aplicações**. Editora Manole Ltda, 2003.

SANTOS MELLO, V. dos. **AUTOCODIFICADORES PROFUNDOS APLICADOS A DETECÇÃO DE NOVIDADE EM SISTEMAS DE SONAR PASSIVO**. 2019. Tese (Doutorado) – Universidade Federal do Rio de Janeiro.

SCHULTE, J. P. Uma abordagem para agrupar dados de NF-e com base em Autoencoder, 2022.

SILVA GILBERT, M. da; CAMPOS, M. L. R. de; CAMPISTA, M. E. M. Autoencoders Assimétricos para a Compressão de Dados IoT. in XLI Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC 2023), Brasília, 2023.

SUZUKI, K. **Artificial neural networks: methodological advances and biomedical applications**. BoD–Books on Demand, 2013.

TENENBAUM, J. B.; SILVA, V. d.; LANGFORD, J. C. A global geometric framework for nonlinear dimensionality reduction. **science**, American Association for the Advancement of Science, v. 290, n. 5500, p. 2319–2323, 2000.

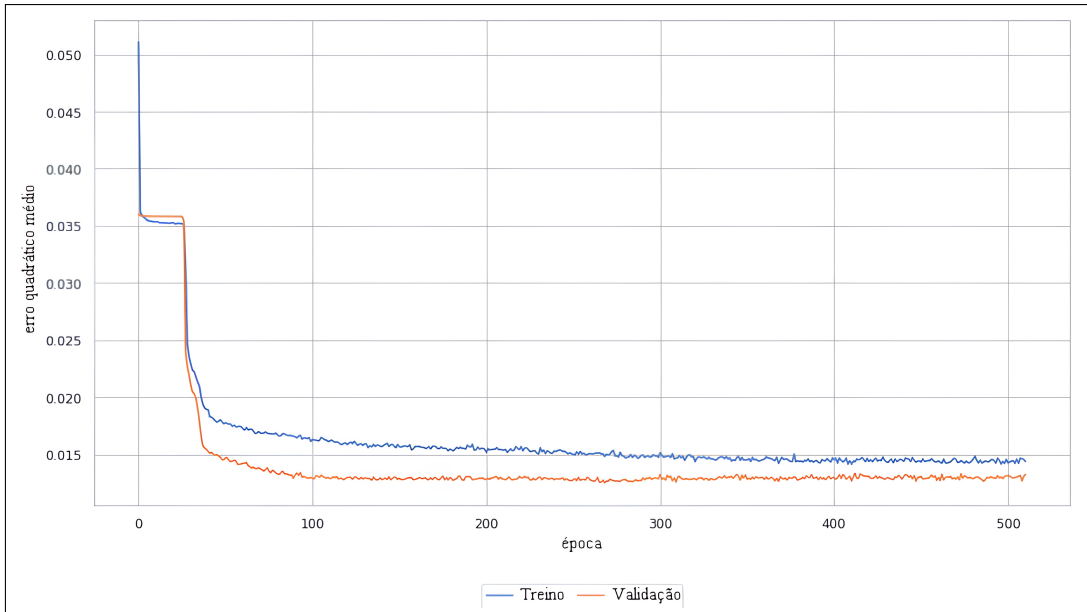
VERÍSSIMO, J. M. D. **As bibliotecas universitárias face ao desafio do Google Scholar: ameaça ou oportunidade?** 2012. Tese (Doutorado).

XU, M. *et al.* Integrating memory-mapping and N-dimensional hash function for fast and efficient grid-based climate data query. **Annals of GIS**, Taylor & Francis, v. 27, n. 1, p. 57–69, 2021.

APÊNDICES

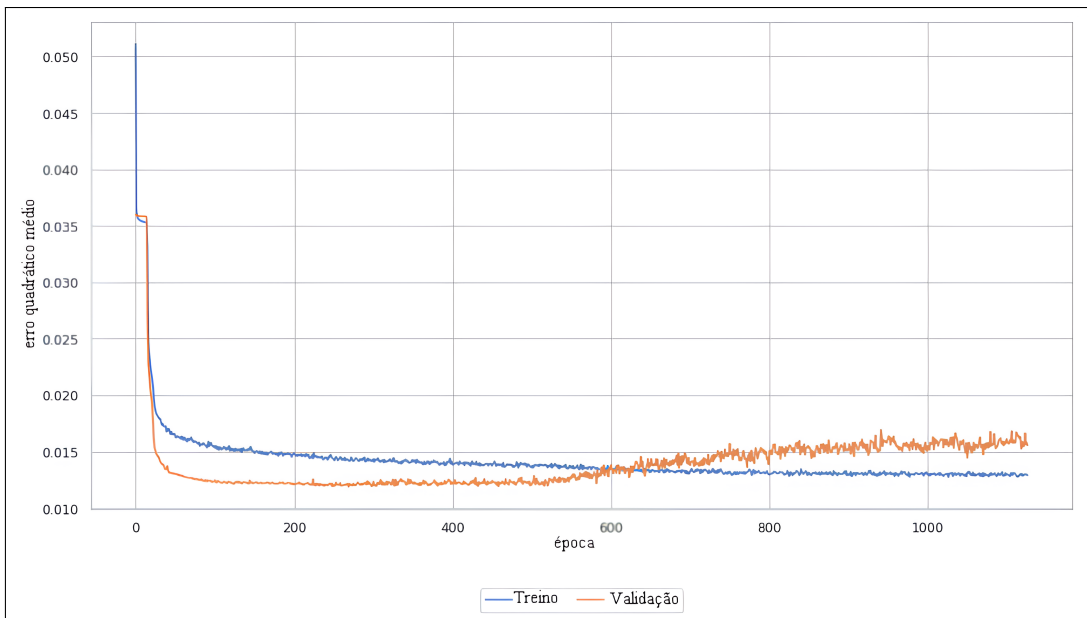
APÊNDICE A – GRÁFICOS DE CONVERGÊNCIA DO ERRO

Figura 11 – Convergência do erro com duas representação latente



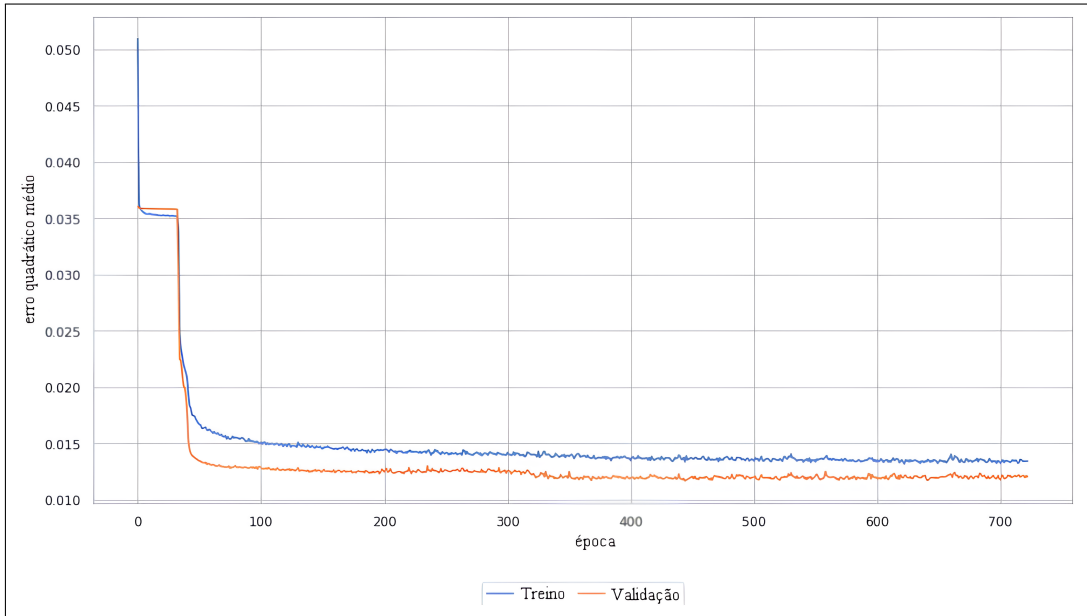
Fonte: Elaborado pelo autor, 2023.

Figura 12 – Convergência do erro com três representação latente



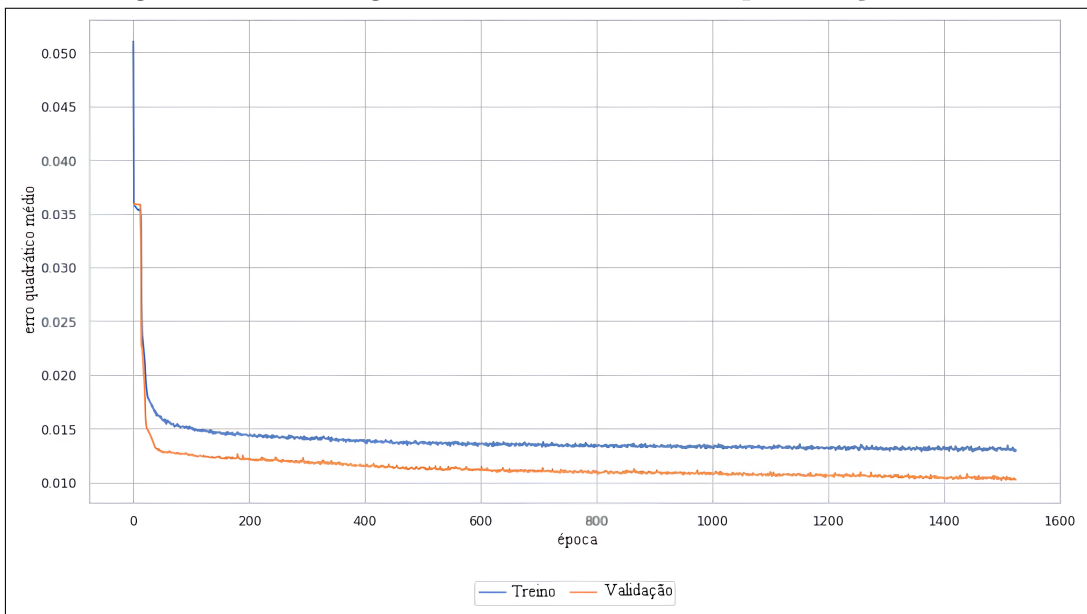
Fonte: Elaborado pelo autor, 2023.

Figura 13 – Convergência do erro com quatro representação latente



Fonte: Elaborado pelo autor, 2023.

Figura 14 – Convergência do erro com cinco representação latente



Fonte: Elaborado pelo autor, 2023.