ARTIGO ORIGINAL

MODELAGEM E PREVISÃO DE FENÔMENOS SOCIOECONÔMICOS: UMA ANÁLISE COMPARATIVA ENTRE MÉTODOS CLÁSSICOS E CONTEMPORÂNEOS

MODELING AND FORECASTING SOCIOECONOMIC PHENOMENA: A COMPARATIVE ANALYSIS BETWEEN CLASSICAL AND CONTEMPORARY METHODS

Carlos Roberto Souza Carmo[1]

Gabriel Pereira Lopacinski [2]

RESUMO

A partir de uma amostra relativamente pequena, porém, com uma significativa quantidade variáveis explicativas, esta investigação teve por objetivo avaliar comparativamente o processo de modelagem voltado para análise e predição da remuneração média mensal paga aos empregados nas 27 unidades da federação brasileira, com base em duas metodologias distintas, isto é: uma mais clássica e direcionada pela perspectiva estatística, ou seja, a análise de regressão linear; e, outra mais moderna e direcionada pela perspectiva tecnológica e computacional, portanto, rede neural artificial. Mediante o levantamento das informações referentes ao ano de 2019, foram investigados e identificados dois modelos que capazes de permitir a análise e a previsão da remuneração média mensal paga aos empregados brasileiros (variável dependente) em função da quantidade total de empresas atuantes em 21 segmentos econômicos distintos, em cada unidade da federação (variáveis independentes). Em seguida, cada um desses modelos foi utilizado para prever o valor da remuneração média mensal paga aos referidos empregados no ano de 2020, e ainda, foi realizada a respectiva análise de precisão com base nos valores reais observados para esse ano. Ao final, foi possível concluir que, do ponto de vista analítico, a análise de regressão linear apresentou maior facilidade interpretativa; contudo, a rede neural artificial pôde ser considerada mais eficiente no processo de estimativa da remuneração média mensal, no ano de 2020.

Palavras-chave: análise; regressão; RNA; multicamadas.

ABSTRACT

Based on a relatively small sample, however, with a significant number of explanatory variables, this investigation aimed to comparatively evaluate the modeling process aimed at analyzing and predicting the average monthly remuneration paid to Brazilian employees, in the 27 units of the Brazilian federation, with based on two different methodologies, that is: a

[1] Doutor em Agronomia pela UNESP (campus Botucatu). Mestre em Ciências Contábeis pela PUC-SP. Professor da Faculdade de Ciências Contábeis da Univ. Federal de Uberlândia (FACIC-UFU). carlosjj2004@hotmail.com. https://orcid.org/0000-0002-3806-9228.

[2] Graduando em Ciências Contábeis pela Universidade Federal de Uberlândia (FACIC-UFU). gplopacinski@gmail.com. https://orcid.org/0009-0002-6520-3215.

more classical one guided by the statistical perspective, that is, the linear regression analysis; and another more modern and directed by the technological and computational perspective, therefore, artificial neural network. Through the survey of information referring to the year 2019, two models were investigated and identified that are capable of allowing the analysis and forecast of the average monthly remuneration paid to Brazilian employees (dependent variable) according to the total number of companies operating in 21 economic segments different, in each federative unit (independent variables). Then, each of these models was used to predict the value of the average monthly remuneration paid to employees in each unit of the Brazilian federation in the year 2020, and the respective accuracy analysis was performed based on the actual values observed for that year. In the end, it was possible to conclude that, from the analytical point of view, the linear regression analysis was easier to interpret; however, the artificial neural network could be considered more efficient in the process of estimating the average monthly remuneration in the year 2020.

Keywords: analysis; regression; ANN; multilayer.

1 Introdução

A utilização de modelos estatísticos e/ou computacionais no processo analítico-preditivo envolvendo fenômenos sociais contemporâneos é cada vez mais frequente, caracterizando-se, de certa forma, como uma realidade que se impõe naturalmente em decorrência dos avanços tecnológicos experimentados recentemente. Contudo, é preciso desenvolver certa compreensão conceitual acerca dessas metodologias para que se possa avançar de forma contínua e segura, tornando possível tirar maior proveito desse tipo de tecnologia.

Dentre as várias técnicas utilizadas para modelagem, análise e previsão de fenômenos que cercam a sociedade de uma maneira geral, esta investigação abordou duas técnicas distintas, em que, uma pode ser considerada mais clássica e a outra pode ser considerada mais contemporânea, portanto, a análise de regressão linear e as redes neurais artificiais, respectivamente. A primeira caracteriza-se como uma das técnicas amplamente aplicadas na investigação de relações lineares entre uma única variável objeto de interesse e uma ou mais variáveis independentes, podendo ser considerada como “a base da teoria da aprendizagem estatística” (HUANG, 2023, p. 548). A segunda tem encontrado espaço para aplicação em diversas áreas do conhecimento humano, indo desde manufatura e finanças até a medicina, com desempenhos que podem ser consideradas excelentes em diversas áreas de aplicação (KANG; KANG, 2023)

No campo da estatística, a análise de regressão linear é utilizada para descrever a relação entre o comportamento de uma variável de estudo, ou dependente, em função do comportamento de uma ou mais variáveis explicativas, ou independentes, e já foi

rigorosamente estuda, contando com extensas aplicações de cunho prático (SHEWA; UGWUOWO, 2023). No entanto, a dependência de uma relação linear entre as variáveis envolvidas e a quantidade de pressupostos necessários para sua validação, fazem com que o desempenho da análise de regressão seja negativamente afetado (MOSTOUFI; CONSTANTINIDES, 2023; SHEWA; UGWUOWO, 2023), além do fato de que a maioria dos modelos matemáticos representativos da realidade tendem a ser baseados em parâmetros não-lineares (MOSTOUFI; CONSTANTINIDES, 2023).

Na era "pré-computacional", técnicas baseadas na reorganização de equações e no reagrupamento de variáveis eram alguns dos procedimentos comuns na tentativa de linearizar modelos experimentais baseados na análise de regressão linear (MOSTOUFI; CONSTANTINIDES, 2023, p. 403). Adicionalmente, aquelas possíveis variáveis explicativas (preditores) que não apresentam correlação linear com outras variáveis explicativas (multicolinearidade) ou com a variável de estudo têm coeficientes de regressão com valor igual a zero e/ou são excluídas dos modelos baseados em análise de regressão linear com mais de uma variável independente (multivariada) (BAUER; DRABANT, 2023), o que nem sempre é desejável do ponto de vista analítico-preditivo.

Com o avanço dos métodos computacionais, as análises e estimativas baseadas em redes neurais artificiais têm sido amplamente utilizadas no processo de aprendizagem das complexas relações entre variáveis de estudo e suas possíveis variáveis explicativas, em diversos campos de pesquisa (TRETIAK; SCHOLLMEYER; FERSON, 2023). E, devido à sua capacidade aproximativa da realidade e à expressividade dos resultados alcançados, as redes neurais artificiais e suas aplicações têm se tornado uma área de pesquisa muito ativa nas últimas décadas (JIAO; WANG, YANG, 2023)

De uma maneira geral, os modelos analítico-preditivos baseados em redes neurais são treinados com o objetivo de minimizar o erro em relação aos dados observados em uma amostra, admitindo-se estimativas de viés baseadas na teoria da aproximação, levando-se em conta o quanto o modelo pesquisado pode ser generalizado a partir de amostras finitas de dados (JIAO; WANG, YANG, 2023).

A utilidade comprovada das análises e previsões baseadas em dados fez com que o aprendizado de máquina (*machine learning*) e as redes neurais artificiais se caracterizassem como ferramentas importantes para pesquisa, para a produção e para a sociedade em geral (TOHME; VANSLETTE; YOUCEF-TOUMI, 2023). Dessa forma, a principal diferença entre esses métodos computacionais e as técnicas tradicionais de modelagem estatística reside na

capacidade de realizar análises tratáveis em grandes conjuntos de dados e com uma densidade de informações cada vez mais significativa; sendo que, à medida que o volume de dados produzidos pela sociedade cresce diuturnamente, eleva-se também a expectativa acerca do crescimento e da evolução das técnicas baseadas em métodos computacionais (TOHME; VANSLETTE; YOUCEF-TOUMI, 2023).

Diante desse quadro, esta investigação teve por objetivo geral avaliar comparativamente o processo de modelagem estatístico-computacional voltado para análise e predição da remuneração média mensal paga aos empregados brasileiros, nas 27 unidades da federação brasileira, com base em duas metodologias distintas, ou seja, a análise de regressão linear e em redes neurais artificiais.

Assim, a partir de uma amostra composta por observações referentes à 27 unidades da federação, e com uma quantidade significativa de possíveis variáveis explicativas, composta pelas quantidades totais de unidades operacionais (matriz, filiais, sucursal, etc.) de empresas ativas classificadas de acordo com as 21 seções de atividades econômicas descritas no Cadastro Nacional de Atividades Econômicas ou CNAE 2.0 (IBGE, 2015), a presente investigação buscou identificar as vantagens e desvantagens de cada uma daquelas duas metodologias analítico-preditivas.

2 Referencial Teórico

Esta seção foi desdobrada em duas seções secundárias com o objetivo de delimitar a abordagem de cada uma das metodologias analítico-preditivas avaliadas nesta pesquisa científica. Para tanto, a primeira seção secundária foi destinada à temática relacionada à análise de regressão linear, e a segunda foi direcionada à temática relacionada ao processo de aprendizagem de máquina baseado em redes neurais artificiais.

2.1 Análise de Regressão Linear

Um modelo matemático baseado na análise de regressão se caracteriza pela correlação linear entre o comportamento de uma variável "*Y*" em função do comportamento de uma ou mais variáveis "*x*", na qual, as médias dessa relação de probabilidade variam de alguma forma sistemática (AZEVEDO, 1997).

A partir de uma função matemática, como aquela definida na Equação 1, a análise de regressão linear permite avaliar o comportamento de uma variável dependente "*Y*" em função do comportamento de uma variável independente "*x*", estimando-se os coeficientes "*a*" e "*b*";

onde, "*a*" caracteriza-se como um termo constante ou intercepto, "*b*" indica a intensidade e o sentido da influência do comportamento de "*x*" sobre o comportamento de "*Y*", sendo que, "*e*" é o termo de erro entre valores previstos "$\hat{Y}$" e valores observados (portanto: e = "$\hat{Y}$" – Y) (BRUNI, 2013).

$$\hat{Y} = a + bx + e \quad (1)$$

A análise de regressão linear "[...] se constitui num conjunto de métodos e técnicas para o estabelecimento de fórmulas empíricas que interpretem a relação funcional entre variáveis com boa aproximação" (FONSECA; MARTINS; TOLEDO, 1982, p. 79). Nesse sentido, além da regressão linear simples, já definida pela Equação 1, a análise de regressão linear permite estudar o comportamento da variável dependente "*Y*" em função de mais de uma variável explicativa "*x*", conforme demonstra a Equação 2.

$$\hat{Y} = a + b_1x_1 + b_2x_2 + \cdots + b_nx_n + e \quad (2)$$

Por se tratar de uma modelagem linear, como o próprio nome indica, é necessário admitir a linearidade no relacionamento do comportamento das variáveis estudadas, conforme descrito pelas Equações 1 e 2, nas quais, o "$\hat{Y}$" representa os valores previstos para a variável dependente ou de estudo (BRAULE, 2001). Dessa maneira, tanto na análise do comportamento da variável dependente (*Y*), quanto da variável independente (*x*) ou explicativa, são utilizados dados históricos para que se possa compreender o comportamento da primeira, em função do comportamento da segunda. E, assim, um coeficiente angular negativo atribuído a uma variável "*x*" indica um comportamento inverso à variável de estudo "*Y*", e, em caso seja positivo, denota-se um comportamento no mesmo sentido da variável de estudo (DOWNING; CLARK, 2006).

Em resumo, a função da regressão linear simples é encontrar coeficientes "*a*" e "*b*", tal que, a partir de sua análise, seja possível estimar um valor previsto "$\hat{Y}$" para a variável de estudo "*Y*" com base nos valores da variável independente "*x*" (FIELD, 2009). E, no caso de uma regressão linear múltipla, a lógica aplicada é a mesma da regressão linear simples, porém, existe mais de uma variável explicativa ou independente na equação pesquisada (FÁVERO; BELFIORE, 2017).

Ao abordar os pressupostos de uma modelagem obtida a partir da análise de regressão linear, Fávero *et al.* (2009) apontam três problemas que normalmente podem advir do emprego dessa técnica estatística: a autocorrelação dos resíduos; a heterocedasticidade; e a multicolinearidade.

Conforme Fávero *et al.* (2009), a autocorrelação dos resíduos diz respeito à existência de correlação entre os termos de erro, caracterizados pela diferença entre os valores estimados pelo modelo e os valores reais observados para a variável de estudo ($e = Y - \hat{Y}$). Esse problema tende a ocorrer quando uma ou mais variáveis explicativas relevantes não foram consideradas nos dados em análise. Assim, os resíduos incorporam efeitos dessas variáveis ausentes, direcionando os coeficientes e influenciando também o erro da estimativa (BROOKS, 2002), que assume um comportamento semelhante ao da variável dependente.

A heterocedasticidade está relacionada à ausência de homoscedasticidade nos resíduos, ou seja, a ausência de variância constante na diferença entre os valores estimados pelo modelo e os valores reais observados para a variável de estudo (FÁVERO *et al*, 2009). A heterocedasticidade surge quando existe erro de especificação ou a omissão de uma variável explicativa relevante, o que leva à estimativa ($\hat{Y}$) de valores enviesados, portanto, inúteis em relação à variável de estudo (Y), podendo gerar resultados tendenciosos (BRUNI, 2013).

A multicolinearidade surge quando as variáveis independentes ou explicativas apresentam correlações cruzadas muito fortes (BRUNI, 2013), ou ainda, quando essas variáveis apresentam comportamentos semelhantes (FÁVERO *et al.*, 2009). Esse tipo de ocorrência é considerado um problema, pois também leva a estimativa de valores enviesados para variável de estudo (Y), bem como, a erros de interpretação das possíveis variáveis explicativas (x) de um modelo baseado em regressão linear múltipla, uma vez que os estimadores dos mínimos quadrados foram usados de forma imprecisa (AZEVEDO, 1997).

A ausência desses três problemas constitui o conjunto de pressupostos básicos para estimativa e utilização de um modelo matemático baseado em análise de regressão linear. Isso se deve ao fato de que, se os três não forem satisfeitos, podem ser produzidas análise s e conclusões incorretas acerca do respectivo objeto de estudo, já que as variáveis explicativas tendem a assumir comportamento enviesado, direcionando o resultado da análise de maneira equivocada.

Contudo, é importante salientar que mesmo nos casos em que aqueles pressupostos básicos de uma modelagem matemática analítico-preditiva sejam cumpridos, o modelo pesquisado com base na análise de regressão linear sempre apresentará, em alguma medida,

erro em relação aos valores reais observados ao longo do processo de estimativa e análise (FIELD, 2009). Aliás, esse é um fato que faz parte do processo de pesquisa baseado nesse tipo de análise, o que pode ser observado pela própria composição da equação resultante desse tipo de modelagem, conforme foi descrito pelos termos "*e*" presentes nas Equações 1 e 2. Por outro lado, o cumprimento dos pressupostos básicos do processo da análise de regressão linear tem por finalidade evitar que os erros, que ocorrem naturalmente em qualquer processo de estimativa, apresentem algum tipo de viés capaz de prejudicar tanto a estimativa quanto a análise baseada nesse tipo de metodologia.

2.2 Aprendizagem de Máquina Baseada em Redes Neurais Artificiais

Desde o nascimento, o cérebro humano cria regras de compreensão e de funcionamento que se traduzem pela experiência na execução de tarefas cujo resultado obtido é o aprendizado, independentemente do sucesso ou fracasso nesse processo (HAYKIN, 2007). Para uma máquina, o aprendizado consiste no treinamento por meio de um algoritmo ou modelo, para que ela possa criar regras que relacionam os dados de entrada (atributos previsores) com os dados de saída (atributos alvo), permitindo a realização de tarefas como classificação, previsão e agrupamento de dados, entre outras possibilidades (LENZ *et al.*, 2020).

Nesse sentido, a máquina busca reconhecer padrões e/ou encontrar semelhanças entre as características de diferentes instâncias acerca de determinado conjunto de dados (LENZ *et al.*, 2020). A partir desse reconhecimento, a máquina cria ou adapta parâmetros de acordo com a necessidade indicada a partir de um algoritmo (HAYKIN, 2007). Sendo que, o processo de aprendizagem de máquina se dá por meio de quatro formas: supervisionada; não supervisionada; semi-supervisionada; e reforço.

A aprendizagem supervisionada constitui na ação da máquina em ler dados previamente inseridos e buscar identificar padrões, e por isso é chamada de supervisionada, de tal forma que se possa usar tais padrões para realizar previsões (LENZ *et al.*, 2020).

A aprendizagem não supervisionada ocorre em duas etapas, isto é, uma voltada para o treinamento e outra para testes a partir da identificação de certos eventos e/ou padrões. A máquina busca padrões em uma parcela de um conjunto previamente definido de dados, a partir dos quais se dá o treinamento do algoritmo de aprendizado; e, após identificar padrões de comportamento nos dados que foram utilizados na fase de treinamento, ela os compara

com a outra parte dos dados, procedendo à verificação (teste) do modelo de previsão pesquisado, a fim de avaliar sua funcionalidade (NETTO; MACIEL, 2021).

A aprendizagem semi-supervisionada une as funcionalidades da aprendizagem supervisionada e da não supervisionada, a partir de etapas. Em um primeiro momento são inseridos alguns dados no processamento da máquina de tal forma que ela possa verificar e entender o padrão. Em um segundo momento, os demais dados a serem verificados são adquiridos de forma não supervisionada, onde ela coleta e termina de interpretar os padrões conforme o que aprendeu a partir dos primeiros dados inseridos em seu processamento (LENZ *et al.*, 2020).

A aprendizagem por reforço implica em explicitar para a máquina a importância, ou não, de certos padrões atribuindo crédito ou penalidades de acordo com a decisão tomada (HAYKIN, 2007). Isso ocorre por meio de condicionamento, no qual, quando um parâmetro importante é encontrado, há uma recompensa, e quando é encontrado um parâmetro sem relevância, é aplicada uma penalidade (LENZ *et al.*, 2020).

A aprendizagem de máquina é utilizada para proporcionar autonomia de trabalho à máquina, de tal forma que facilite o trabalho humano ou o faça. Nesse contexto, um dos recursos utilizados são as redes neurais artificiais (RNA). Inicialmente, uma RNA pode ser entendida como um processador paralelamente distribuído de forma massiva e constituído de unidades de processamento simples que utilizam a aprendizagem de máquina para armazenar conhecimento e torná-lo passível de uso/aplicação (HAYKIN, 2007).

A RNA é utilizada para replicar o funcionamento do pensamento humano artificialmente e encontrar respostas ou hipóteses tal qual um humano faria, porém, evitando seus erros e agilizando processos que podem demorar muito tempo quando biologicamente realizados. O processo de aprendizado da RNA é denominado algoritmo de aprendizagem e esse algoritmo interpreta os dados recebidos e os testa repetidamente, e a cada repetição atribuem-se pesos aos dados e suas relações, de tal forma que sejam identificados padrões de comportamento voltados para a previsão. Essa forma de previsão é chamada de generalização, o que permite gerar repostas ou hipóteses por meio de dados que não existiam até o momento do aprendizado (HAYKIN, 2007).

Nesse contexto, um *perceptron* é a forma mais simples de RNA, sendo constituído basicamente por um único neurônio em uma única camada de processamento, com capacidade de interpretação e ajuste de acordo com os pesos atribuídos a cada variável que se investiga e se classifica conforme um objetivo proposto (LENZ *et al.*, 2020). O *perceptron* é utilizado na

classificação de padrões linearmente separáveis, e quando possui apenas um neurônio é limitado a separar sua interpretação em apenas duas classes ou hipóteses, mas se tiver mais de um neurônio, podem ocorrer duas ou mais separações (HAYKIN, 2007).

Os *perceptrons* de múltiplas camadas (*multilayer perceptron* ou MLP) funcionam da mesma forma que o *perceptron* de uma camada. Porém, por trabalharem com mais de uma camada, possuem diversos nós computacionais (HAYKIN, 2007). Devido à sua estrutura, uma RNA MLP possui um poder de avaliação maior que os *perceptrons* de apenas um neurônio (HAYKIN, 2007), podendo ser entendida como uma rede neural profunda, pois é uma rede neural com capacidade de interpretação e resolução de problemas bem mais complexos (SILVA, 2021).

A RNA MLP tem seu fluxo de aprendizado com base em dados ou informações obtidos de um ambiente de processamento e/ou inseridos por um ser humano, por meio de no mínimo três camadas (HAYKIN, 2007), ou seja, um conjunto de unidades sensoriais de entrada, outro formado por unidades ocultas ou de processamento, e ainda, as unidades de saída (LENZ *et al.*, 2020). A camada de entrada recebe as informações, as camadas ocultas processam a informação e aprendem com ela, e a camada de saída apresenta os resultados (HAYKIN, 2007). Tanto as camadas ocultas quanto as de saída são constituídas por nós computacionais onde os neurônios estão situados (HAYKIN, 2007), ao passo que a camada de entrada é responsável por receber os sinais ou dados que vêm do ambiente no qual a RNA está inserida (LENZ *et al.*, 2020).

Nas camadas ocultas, onde ocorre o processamento da RNA, os respectivos neurônicos atribuem pesos a cada um dos sinais recebidos da camada de entrada, e tais pesos determinam o quão importante é o dado, enviando um sinal para as demais camadas ocultas e/ou diretamente para a camada de saída, se a RNA contar com apensas uma camada oculta para processamento (LENZ *et al.*, 2020). As camadas ocultas utilizam cada neurônio para interpretar os sinais vindos da primeira camada, atribuindo-lhes pesos (HAYKIN, 2007). Com essas informações, os neurônios geram uma função do sinal de entrada e dos pesos sinápticos associados aos sinais vindos da camada anterior para, em seguida, calcular seu vetor gradiente e assim propagar e retropropagar a informação aprendida pela RNA (HAYKIN, 2007).

A camada de saída apresenta o resultado proveniente de uma operação chamada junção aditiva (LENZ *et al.*, 2020). A junção aditiva pode ser entendida como o resultado proveniente da soma dos resultados encontrados nas camadas ocultas, após suas multiplicações pelos respectivos pesos (LENZ *et al.*, 2020).

A propagação (*propagation*) ocorre assim que a camada de entrada recebe os sinais de entrada e os envia para que os neurônios façam suas análises (LENZ *et al.*, 2020). A retropropagação (*backpropagation*) ocorre quando é feita a operação contrária, aplicando uma função aos resultados encontrados de tal forma que se chegue ao valor originalmente analisado antes de suas alterações e, nesse processo, os pesos são reajustados, sendo que, caso existam dois ou mais caminhos, os pesos são somados dentro de um nó (GOODFELLOW; BENGIO; COURVILLE, 2016).

Nesse processo analítico, utiliza-se uma função de ativação que é programada dentro da RNA para determinar a forma como o resultado será gerado (LENZ *et al.*, 2020). E, esse processo de determinação ocorre por meio da limitação ou restrição da amplitude de saída do sinal atribuído aos neurônios (HAYKIN, 2007).

As conexões que acontecem entre os neurônios da RNA simulam as sinapses do cérebro humano e são caracterizadas por um peso ou força própria (HAYKIN, 2007). Tais sinapses podem ser excitatórias, quando com um valor positivo, ou inibitórias, quando com um valor negativo (LENZ *et al.*, 2020).

Tanto a análise de regressão linear quanto as RNA utilizam técnicas de modelagem que buscam estabelecer relações matemáticas entre uma variável dependente e uma ou mais variáveis independentes ou explicativas. A análise de regressão linear usa uma função linear para modelar tais relações, enquanto a RNA MLP utiliza algoritmos de aprendizagem de máquina para entender e modelar essa relação de forma linear e/ou não-linear. Nesse contexto, se a análise de regressão utiliza o método dos mínimos quadrados para encontrar os coeficientes de um modelo analítico-preditivo, as RNA utilizam funções específicas que permitem realizar previsões mediante a atribuição de pesos, que são medidos e alterados por uma função de ativação que busca diminuir o erro de um modelo, que tende a ser muito mais preditivo do que analítico.

3 Metodologia

A base de dados utilizada como amostra desta pesquisa foi gerada a partir de consultas realizadas junto ao banco de tabelas estatísticas do Sistema de Recuperação Automática de dados (SIDRA), disponibilizado pelo Instituto Brasileiro de Geografia e Estatística (IBGE). Inicialmente, foram levantadas as informações referentes ao total de salários e outras remunerações pagas anualmente a todas as pessoas ocupadas em cada uma das 27 unidades da federação brasileira (IBGE, 2022), segundo 21 tipos de atividades econômicas classificadas

de acordo com o Cadastro Central de Empresas (CEMPRE). E, na sequência, procedeu-se ao cálculo da remuneração média mensal por pessoa ocupada, em cada uma das 27 unidades da federação, conforme descrito na Equação 3.

$$Remun.media\ mensal = \frac{\left(\frac{Salários\ e\ outras\ remunerações\ (R\$1000)}{Pessoal\ ocupado\ total\ (Pessoas)}\right)}{13\ (meses)} \quad (3)$$

Ainda segundo as informações disponibilizadas para consulta no SIDRA, com base no CEMPRE (IBGE, 2022), foram identificadas as informações referentes a quantidade total de unidades operacionais (matriz, filiais, sucursal, etc) de empresas ativas em cada uma das 27 unidades da federação, devidamente classificadas de acordo com as 21 seções de atividades econômicas descritas no Cadastro Nacional de Atividades Econômicas (CNAE 2.0), ou seja: agricultura, pecuária, produção florestal, pesca e aquicultura; indústrias extrativas, indústrias de transformação; eletricidade e gás; água, esgoto, atividades de gestão de resíduos e descontaminação; construção; comércio; transporte, armazenagem e correio; alojamento e alimentação; informação e comunicação; atividades financeiras, de seguros e serviços relacionados; atividades imobiliárias; atividades profissionais, científicas e técnicas; atividades administrativas e serviços complementares; administração pública, defesa e seguridade social; educação; saúde humana e serviços sociais; artes, cultura, esporte e recreação; outras atividades de serviços; serviços domésticos; organismos internacionais e outras instituições extraterritoriais (IBGE, 2015).

A partir do levantamento das informações referentes ao ano de 2019 (remuneração média mensal e quantidade de empresas por segmento econômico, em cada unidade da federação), foram investigados e identificados dois modelos que foram capazes de permitir a análise e a previsão da remuneração média mensal paga aos empregados brasileiros (variável de estudo), em função da quantidade total de empresas atuantes naqueles 21 segmentos econômicos do CNAE 2.0 de cada unidade da federação (variáveis explicativas).

Com base nos dados do ano de 2019, o primeiro modelo objeto deste estudo foi identificado mediante o emprego da análise de regressão linear múltipla, e o segundo foi identificado com auxílio de uma rede neural artificial (RNA) do tipo *perceptron* de múltiplas camadas (MLP). Em seguida, cada um desses modelos foi utilizado para prever o valor da remuneração média mensal paga aos empregados de cada unidade da federação brasileira no

ano de 2020, levando-se em conta as respectivas quantidades de empresas atuantes naqueles 21 segmentos econômicos descritos pelo CNAE 2.0 nesse ano.

Depois de identificado o valor real da remuneração média mensal paga aos empregados de cada unidade da federação no ano de 2020, procedeu-se à análise comparativa da qualidade preditiva de cada um dos modelos pesquisados segundo as duas metodologias analisadas no presente estudo (regressão linear múltipla e RNA MLP). Para tanto, foi utilizado o conjunto de métricas descrito na Figura 1.

Figura 1 – Métricas de precisão utilizadas para avaliar previsões

Descrição	**Fórmula**	**Unidade de medida (un) e parâmetro de decisão**
Determination coefficient ou coeficiente de determinação (R^2)	$R^2\ (y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$	$0 < R^2 < 1$ **ou** $0\% < R^2 < 100\%$ sendo que: quanto mais próximo de 1 melhor
Mean absolut error ou erro absoluto médio (*MAE*)	$MAE\ (y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	*un do MAE = un de y* sendo que: quanto menor melhor
Median absolut error ou erro absoluto mediano (*MdAE*)	$MdAE\ (y,\hat{y}) = \lvert y_i - \hat{y}_i\rvert$	*un do MAE = un de y* sendo que: quanto menor melhor
Mean absolute percentage error ou erro percentual médio absoluto (*MAPE*)	$MAPE\ (y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert y_i - \hat{y}_i\rvert}{\lvert y_i\rvert}$	$0\% < MAPE < 100\%$ sendo que: quanto menor melhor
Median absolute percentage error ou erro percentual absoluto mediano (*MdAPE*)	$MdAPE\ (y,\hat{y}) = \frac{\lvert y_i - \hat{y}_i\rvert}{\lvert y_i\rvert}$	$0\% < MdAPE < 100\%$ sendo que: quanto menor melhor
Symmetric mean absolute percentage error ou erro percentual médio absoluto simétrico (*SMAPE*)	$SMAPE\ (y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert y_i - \hat{y}_i\rvert}{(\lvert y_i\rvert + \lvert\hat{y}_i\rvert)/2}$	$0\% < SMAPE < 100\%$ sendo que: quanto menor melhor
Median symmetric absolute percentage error ou erro percentual absoluto simétrico mediano (*MdSAPE*)	$MdSAPE\ (y,\hat{y}) = \frac{\lvert y_i - \hat{y}_i\rvert}{(\lvert y_i\rvert + \lvert\hat{y}_i\rvert)/2}$	$0\% < MdSAPE < 100\%$ sendo que: quanto menor melhor
Weighted mean absolute percentage error ou erro percentual médio absoluto ponderado (*WMAPE*)	$WMAPE\ (y,\hat{y}) = \frac{\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert}{\sum_{i=1}^{n}\lvert y_i\rvert}$	$0\% < WMAPE < 100\%$ sendo que: quanto menor melhor
Mean square error ou erro quadrático médio (*MSE*)	$MSE\ (y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	*un do MSE = (y – ŷ)²* sendo que: quanto menor melhor
Median square error ou erro quadrático mediano (*MdSE*)	$MdSE\ (y,\hat{y}) = (y_i - \hat{y}_i)^2$	*un do MdSE = (y – ŷ)²* sendo que: quanto menor melhor
Root mean square error ou raiz do erro quadrático médio (*RMSE*)	$RMSE\ (y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$	*un do RMSE = un de y* sendo que: quanto menor melhor
Legenda: *y* = valor da variável analisada;		

$\hat{y}$ = valor previsto para a variável analisada com base em um modelo;
n = quantidade total de observações referentes a y e/ou $\hat{y}$; e
i = cada observação específica de y e/ou $\hat{y}$.

Fonte: Carmo e Silva (2023, p. 131-132).

O R^2 identifica a proporção das observações reais (y_i) referentes à variável de estudo que podem ser explicadas pelas estimativas ($\hat{y}_i$) realizadas a partir da respectiva modelagem analítico-preditiva (McKEAN; SIEVERS, 1987).

O *MAE* representa uma medida do erro absoluto (y_i - $\hat{y}_i$) obtido a partir da soma de todos os erros de previsão dividida pela quantidade de observações (HYNDMAN; KOEHLER, 2006; KARUNASINGHA,2022). O *MdAE* também é uma medida de erro absoluto, porém, caracteriza-se como uma alternativa ao *MAE*, podendo ser obtido a partir da mediana dos erros absolutos ocorridos a cada comparação pontual entre valores reais (y_i) e valores previstos ($\hat{y}_i$) (HYNDMAN; KOEHLER, 2006).

Caracterizando-se como uma medida de erro relativo, o *MAPE* pode ser obtido a partir da média dos erros percentuais ({[y_i - $\hat{y}_i$] / y_i}.100), ou decimais ([y_i - $\hat{y}_i$] / y_i) (HYNDMAN; KOEHLER, 2006; HUANG *et al*., 2023). De forma análoga à relação descrita entre o *MAE* e o *MdAE*, o *MdAPE* caracteriza-se pela mediana dos erros percentuais observados a cada comparação entre real e previsto (HYNDMAN; KOEHLER, 2006; HUANG *et al*., 2023). Dessa maneira, *MdAE* e *MdAPE* servem como uma medida de apoio para calibrar o *MAE* e o *MAPE*, respectivamente.

Levando em conta tanto os valores reais quanto valores previstos, o SMAPE indica o erro percentual médio absoluto simétrico (HYNDMAN; KOEHLER, 2006). Assim como acontece com *MdAE* e com *MdAPE*, respectivamente em relação ao *MAE* e o *MAP*E, o *MdSPE* caracteriza-se pela mediana do *SMAPE*. Já o *WMAPE* é uma medida de erro percentual (ou decimal) médio absoluto, contudo, ele pondera a quantidade de observações e os seus respectivos montantes, (BUSARI; LIM, 2021; FORBES, 2023).

Por fim, o *MSE* e o *RMSE* caracterizam-se pelo erro quadrático médio e por sua raiz quadrada, respectivamente; e, uma vez que o *MSE* é expresso em uma unidade de leitura de difícil compreensão, o cálculo da sua raiz quadrada gera o *RMSE*, que faz com que o MSE seja expresso na mesma unidade de medida do objeto da previsão realizada (DAM *et al.*, 2022; KARUNASINGHA, 2022). E, semelhante ao que acontece com as demais medidas de erro mediano abordadas até aqui (*MdAE*, *MdAPE* e *MdSPE*), o *MdSE* representa a mediana do MSE (HYNDMAN; KOEHLER, 2006).

Assim, considerando o seu objeto de estudo, a respectiva metodologia analítica e a base de dados utilizada, esta pesquisa pode ser classificada como uma investigação científica de caráter qualitativo e natureza empírica, baseada em métodos quantitativos e em métodos computacionais, ambos aplicados à modelagem, análise e previsão de fenômenos do contexto socioeconômico brasileiro.

4 Análise dos Dados e Resultados

Inicialmente, procedeu-se à análise de regressão linear múltipla baseada no método *stepwise*, segundo o qual, as possíveis variáveis explicativas (*x*) são pesquisadas adicionando-se uma a uma à respectiva modelagem; e, a cada nova adição, realiza-se um teste para remoção da variável explicativa ("*x*") menos significativa, de tal forma que, ao final, restam somente aquelas variáveis capazes de explicar o comportamento da variável de estudo ou variável independe ("*Y*") (FIELD, 2009). Portanto, identificam-se somente aquelas variáveis explicativas cuja combinação linear é significativa. A justificativa para escolha do método *stepwise* se deve ao fato de que o processo analítico ora proposto não levou em conta nenhum pressuposto teórico acerca das variáveis envolvidas, almejando-se apenas identificar a modelagem matemática baseada em regressão linear que melhor explicasse e pudesse prever o comportamento da variável independente.

Conforme pode-se observar na Tabela 1, a análise de regressão linear identificou um termo constante (*a*) com um valor de R$ 2.392,971, e os coeficientes (*b*) referentes a duas variáveis explicativas, ou seja: a quantidade de empresas e unidades operacionais referentes a "organismos internacionais e outras instituições extraterritoriais", com valor de R$ 28,102 por empresa/unidade operacional); e, a quantidade de empresas e unidades operacionais atuantes no "comércio", com um valor de -R$ 0,003 por empresa/unidade.

Tabela 1 – Resumo dos parâmetros da análise de regressão linear múltipla, pelo método *stepwise*

Modelo[a, b, c]	Coeficientes	Est. *t*	Sig.	*Tolerância*	*VIF*	Est. *f*	Sig.
Constante	2392,971	34,427	0,000			71,258	0,000
Organismos internacionais e outras instituições extraterritoriais	28,102	11,935	0,000	0,863	1,159		
Comércio	-0,003	-4,185	0,000	0,863	1,159		

(a) Estatística Durbin-Watson = 1,162
(b) Teste de Pesarán-Pesarpán: Est. *f* = 0,32; Sig. = 0,577
(c) Teste de Kolmogorov-Smirnov : Est. *K-S* = 0,074; Sig. = 0,200

Fonte: elaborado pelos autores com base nos dados da pesquisa.

Acerca dos pressupostos do processo de modelagem baseado na análise de regressão linear, Carmo e Carmo (2014) resumem os seguintes parâmetros aplicáveis às métricas utilizadas nesse tipo de avaliação: a estatística *t* de cada coeficiente deve apresentar uma significância menor que 0,05 para afastar a possibilidade de que o modelo pesquisado apresente tendência ao valor zero; as estatísticas de *Tolerância* e *VIF* devem apresentar valores maiores que 0,20 e menores que 5,00, respectivamente, para se descartar a presença de problemas de multicolinearidade; a estatística *f* do modelo deve apresentar significância menor que 0,05, indicando que a combinação linear das variáveis explicativas (*x*) é significativa; a estatística de Durbin-Watson deve apresentar um valor entre 1,00 e 3,00 para que seja descartada a existência de problemas relacionados à autocorrelação residual; a estatística *f* do teste de Pesarán-Pesarán deve apresentar significância maior que 0,05, de forma a permitir descartar a presença de problemas com heterocedasticidade; e, finalmente, o teste de Kolmogorov-Smirnov deve apresentar uma significância maior que 0,05 para que seja comprovada a normalidade dos termos de erro ou resíduos. Assim, pode-se afirmar que foram satisfeitos todos os pressupostos necessários para a validação de uma modelagem analítico-preditiva baseada na análise de regressão linear, segundo demonstram as informações contidas na Tabela 1.

Analiticamente, o modelo explicativo da variável independente "remuneração média mensal" permitiu realizar o estudo dos valores e dos sinais dos coeficientes pesquisados por meio da análise de regressão linear. Com base na amostra de dados referentes ao ano de 2019, pôde-se inferir que devido à pouca ocorrência desse tipo empresa, as unidades da federação com "organismos internacionais e outras instituições extraterritoriais" têm sua remuneração média mensal por empregado elevada, isto é, R$ 28,10, por cada unidade operacional presente no seu território. E, de forma inversa, pôde-se inferir que, provavelmente devido à grande ocorrência de empresas desse tipo de atividade econômica, as unidades da federação com empreendimentos comerciais têm sua remuneração média mensal por empregado reduzida em R$ 0,003, para cada unidade operacional presente no seu território. Sendo que, independentemente do tipo de atividade econômica e da quantidade de empresas e/ou unidades operacionais, a remuneração média mensal por empregado tem um valor constante de R$ 2.392,97, em todas as unidades da federação brasileira.

Mesmo sem realizar a análise desse conjunto de evidências à luz da teoria adjacente, pôde-se perceber que a análise de regressão linear é capaz de fornecer subsídios para a inferência qualitativa acerca do comportamento da variável de estudo. Adicionalmente,

destaca-se a possibilidade de utilização dos respectivos coeficientes (R$ 2.392,9; R$ 28,10; e, -R$ 0,003) para realizar previsões do valor da variável de estudo em questão, ou seja, da "remuneração média mensal" no distrito federal e nos 26 estados brasileiros.

Na sequência, ainda com base nos dados do ano de 2019, procedeu-se à estimativa do segundo modelo analítico-preditivo objeto de estudo desta pesquisa, todavia, com auxílio de uma rede neural artificial (RNA) do tipo *perceptron* de múltiplas camadas (MLP), conforme o resumo dos parâmetros descritos na Figura 2.

Figura 2 – Resumo dos parâmetros e processamento da RNA MLP

Camada de entrada	Covariáveis	1	A Agricultura
		2	B Indústria extrativa
		3	C Indústria de transf.
		4	D Eletricidade
		5	E Água, esgoto e ativ.
		6	F Construção
		7	G Comércio
		8	H Transporte
		9	I Alojamento
		10	J Informação
		11	K Atividade financ.
		12	L Atividade imobil.
		13	M Atividade prof., cient.
		14	N Atividade adm. e serv.
		15	O Administração publ.
		16	P Educação
		17	Q Saúde humana
		18	R Artes, cultura
		19	S Outras ativ. de serv.
		20	U Organismo internac. e
	Número de unidades		20
	Método de reescalonamento para covariáveis		Padronizado
Camadas ocultas	Número de camadas ocultas		1
	Número de Unidades na Camada Oculta 1[a]		6
	Função de ativação		Tangente hiperbólica
Camada de saída	Variáveis dependentes	1	Remun. média mês_2019
	Número de unidades		1
	Método de reescalonamento para dependentes de escala		Padronizado
	Função de ativação		Identidade
	Função de erro		Soma dos Quadrados
Resumo de processamento do caso[b]	Amostra	N	Porcentagem
	Treinamento	21	77,8%
	Testes	6	22,2%
	Válido	27	100,0%
	Excluídos	0	0
	Total	27	100,0%

(a) Excluindo a unidade de viés
(b) Tempo de treinamento 0:00:00,03

Fonte: elaborado pelos autores com base nos dados da pesquisa.

A RNA MLP proposta considerou uma parcela da amostra (77,8% ou 21 unidades federativas) para treinamento, e uma parcela menor (22,20% ou 6 unidades federativas) para os testes de precisão, conforme informado na Figura 2. Nesse sentido, é possível observar a primeira diferença metodológica em relação a análise de regressão linear, que utiliza toda a amostra para pesquisar seu modelo analítico-preditivo. Ou seja, nos casos de uma amostra com poucas observações, a RNA MLP pode deixar de considerar observações relevantes no processo de aprendizagem e, depois, utilizar tais observações no processo de testagem, o que pode comprometer a qualidade preditiva da modelagem pesquisada.

Diferentemente da análise de regressão linear, a RNA MLP não apresenta coeficientes referentes ao modelo propriamente dito; contudo, ela permite salvar as respectivas estimativas de ponderações sinápticas. Dessa forma, com o auxílio do código de programação na linguagem *eXtensible Markup Language* (*XML*), é possível utilizar as ponderações sinápticas da RNA MLP para realizar estimativas/previsões futuras.

Diante de tal fato, observou-se que, do ponto de vista analítico, a análise de regressão linear é mais dinâmica e informativa que a RNA MLP. Pois, ao identificar os coeficientes do modelo de regressão e seus sinais, é possível compreender o impacto de cada variável independente significativa na explicação do comportamento da variável de estudo. Ao passo que, a RNA MLP somente permite avaliar qual o grau de importância das variáveis explicativas do modelo, conforme demonstrado na Figura 3.

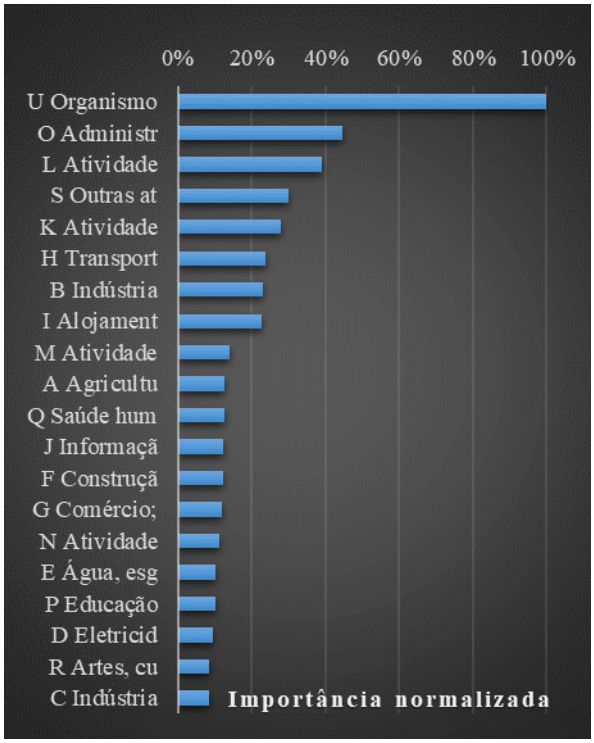
A análise de importância proporcionada pela RNA MLP, descrita na Figura 3, evidenciou que a variável explicativa mais importante é a quantidade de empresas do segmento de “organismos internacionais e outras instituições extraterritoriais”. Esse achado permitiu constatar que essa foi a única informação comum às duas metodologias analítico-preditivas em comparação nesta pesquisa científica. Pois, a segunda variável significativa identificada pela análise de regressão linear (quantidade de empresas do segmento comercial) só ocupou 14º lugar na escala de importância da RNA MLP. Sendo que, a explicação para tanto se deve ao fato de que a RNA MLP considera todas as variáveis explicativas no seu processo de modelagem, ao passo que, na análise de regressão foram consideradas somente aquelas variáveis estatisticamente significativas.

Nesse sentido, ao classificar as variáveis explicativas de acordo com combinação linear mais significativa e ao realizar a análise de tendência a zero dos respectivos coeficientes, com base nas estatísticas *f* e *t* respectivamente, a análise de regressão implica na remoção das variáveis não significativas. Isso, por sua vez, pode tornar a aplicação preditiva

desse modelo menos compreensível do ponto de vista empírico. Por outro lado, uma vez que não realiza a remoção de variáveis não significativas, atribuindo-lhes apenas aquela classificação de importância demonstrada na Figura 3, a RNA MLP torna o processo de aplicação preditiva mais compreensível, do ponto de vista empírico; pois todas as unidades da federação apresentaram, em maior ou menor quantidade, empresas de todos segmentos econômicos descritos pelo CNAE 2.0, segundo a base de dados utilizadas como amostra desta pesquisa.

Figura 3 – Análise de importância das variáveis explicativas do modelo elaborado pela RNA MLP

Variáveis explicativas	Importância relativa	Importância normalizada
U Organismo	0,225	100,00%
O Administr	0,100	44,50%
L Atividade	0,088	39,00%
S Outras at	0,067	29,90%
K Atividade	0,063	27,80%
H Transport	0,053	23,60%
B Indústria	0,052	23,10%
I Alojament	0,051	22,60%
M Atividade	0,031	14,00%
A Agricultu	0,029	12,70%
Q Saúde hum	0,028	12,60%
J Informaçã	0,028	12,40%
F Construçã	0,027	12,20%
G Comércio;	0,027	11,90%
N Atividade	0,025	11,20%
E Água, esg	0,023	10,10%
P Educação	0,023	10,10%
D Eletricid	0,021	9,50%
R Artes, cu	0,019	8,50%
C Indústria	0,019	8,40%



Fonte: elaborado pelos autores com base nos dados da pesquisa.

Na sequência, a partir dos dois modelos pesquisados, e ainda, levando-se em conta a quantidade de empresas atuantes naqueles 21 segmentos econômicos descritos pelo CNAE 2.0 em cada unidade da federação, no ano de 2020, foi calculado o valor previsto da remuneração média mensal paga aos empregados nesse ano. E, tomando por base o valor real observado para remuneração média mensal paga aos empregados no ano de 2020, foi desenvolvida a análise comparativa entre valores previstos e os valores realmente observados, para cada uma das 27 unidades da federação. Sendo que, para avaliação acerca da qualidade preditiva de cada modelo (análise de regressão *versus* RNA MLP), foi utilizado aquele

conjunto de métricas de precisão descritos anteriormente na Figura 1, cujos valores estão detalhados na Tabela 2.

Tabela 2 – Métricas de precisão aplicadas às previsões realizadas com base nos dados de 2020[a]

Parâmetros	Regressão [b]	RNA MLP [b]
Coeficiente de determinação (R^2) [c]	0,87	0,92
Erro absoluto médio (*MAE*)	215	160
Erro absoluto mediano (*MdAE*)	204	107
Erro percentual médio absoluto (*MAPE*)	8,76%	6,17%
Erro percentual absoluto mediano (*MdAPE*)	9,02%	4,54%
Erro percentual médio absoluto simétrico (*SMAPE*)	8,83%	6,26%
Erro percentual absoluto simétrico mediano (*MdSAPE*)	9,31%	4,52%
Erro percentual médio absoluto ponderado (*WMAPE*)	8,53%	6,36%
Erro quadrático médio (*MSE*)	81.100	48.635
Erro quadrático mediano (*MdSE*)	41.762	11.497
Raiz do erro quadrático médio (*RMSE*)	285	221

(a) Estimativas realizadas com base nos dados reais observados para 2020, nos coeficientes identificados a partir do modelo de regressão linear descrito na Tabela 1 desta seção, e ainda, com base nas estimativas de ponderações sinápticas salvas em código *XML*.

(b) As respectivas medidas de erro foram apresentadas em unidades de medida variadas, conforme descrito na Figura 1, apresentada na seção 3 deste artigo.

(c) Como medida alternativa ao R^2, foi calculado o percentual das observações que não pode ser explicado pelos respectivos modelos, portanto: 1-R^2. Assim, foram observados os valores de 0,13 para a previsão realizada com base na regressão, e 0,08 para a previsão realizada com base na RNA MLP.

Fonte: elaborado pelos autores com base nos dados da pesquisa.

Conforme pode ser observado nas informações descritas na Tabela 2, a RNA MLP apresentou melhores indicadores de precisão para todas as medidas de erro analisadas, comparativamente à análise de regressão linear. Essa evidência indica que, do ponto de vista preditivo, a RNA MLP apresentou um desempenho melhor do que a análise de regressão linear.

Ao considerar a amostra dos dados utilizadas nesta pesquisa, a RNA MLP foi mais eficiente no processo de previsão não só por errar menos, mas, também por considerar a quantidade total de empresas atuantes naqueles 21 segmentos econômicos do CNAE 2.0, em cada unidade da federação. Ou seja, isso é bem mais intuitivo do que levar em conta somente as variáveis explicativas consideradas significativas, como aconteceu no caso da análise de regressão linear.

Adicionalmente, deve-se considerar que uma RNA MLP pode realizar previsões com base em modelos não lineares também, o que não acontece no caso da análise de regressão linear. Porém, conforme demonstrado inicialmente, a interpretação analítica proporcionada pela regressão é mais fácil, desde que a relação entre as variáveis explicativas e a variável de estudo seja linear e passível de resumo em uma equação de igual natureza (linear).

Dessa forma, conclui-se que análise de regressão linear é mais fácil de se compreender e interpretar, comparativamente à RNA-MLP. Contudo, ela demanda a pressuposição de relação linear entre as variáveis envolvidas e o cumprimento de uma série de requisitos/pressupostos estatísticos para validação do respectivo modelo de pesquisa, o que pode inviabilizar a sua aplicação, do ponto de vista metodológico.

As RNA-MLP podem ser consideradas mais flexíveis, uma vez que não demandam relação linear entre as variáveis envolvidas no processo de análise e previsão. Por utilizarem técnicas analíticas baseadas no trabalho computacional de maior volume, as RNA-MLP dispensam o rigor metodológico-estatístico requerido na análise de regressão linear. Além disso, as RNA-MLP são capazes de realizar previsões com base em modelos cujos dados não estejam completos, o que nem sempre é possível com a análise de regressão linear. Contudo, devido à complexidade computacional das suas sinapses, as RNA-MLP dificultam a interpretação dos resultados encontrados, sobretudo em processos de pesquisa de caráter mais analítico que preditivo.

5 Considerações finais

Comparativamente, as duas técnicas avaliadas nesta investigação apresentaram vantagens e desvantagens próprias de cada uma, mas, pode-se afirmar que ambas apresentaram boa capacidade preditiva, apesar da superioridade da RNA-MLP.

Do ponto de vista analítico, a análise de regressão linear apresentou maior facilidade interpretativa. Contudo, deve-se ressaltar a complexidade inerente aos parâmetros estatísticos necessários à validação de um modelo dessa natureza, sem perder de vista que ele se limita a relações de natureza linear.

Ao considerar sua flexibilidade, uma vez que essa metodologia não está limitada a relações exclusivamente lineares, e ainda, que ela é capaz de superar problemas referentes a dados incompletos e/ou ruidosos, a RNA-MLP pôde ser considerada mais eficiente no processo de análise e estimativa da remuneração média mensal em função da quantidade total de empresas atuantes nos 21 segmentos econômicos do CNAE 2.0, de cada unidade da federação.

Deve-se ponderar que os resultados observados neste estudo se limitam à respectiva amostra de pesquisa. Porém, dada a natureza analítica das metodologias analisadas comparativamente (estatística *versus* computacional), pode-se inferir que, mesmo diante de bases de dados de naturezas diversas, os resultados gerais seriam distintos, pois, conforme já

dito, a RNA-MLP superaria facilmente as limitações próprias do método da análise de regressão linear.

Para continuidade deste estudo sugere-se a aplicação das metodologias ora analisadas, porém, em bases de dados com dimensões e naturezas diferentes daquelas avaliadas pela presente pesquisa.

Referências

AZEVEDO, P. R. M. de. **Modelos de regressão linear**. Natal: EDUFRN, 1997.

BAUER, J. O.; DRABANT, B.. Regression based thresholds in principal loading analysis. **Journal of Multivariate Analysis**, [*s. l.*], v. 193, e-article 105103, 2023. Disponível em: https://www.sciencedirect.com/science/article/pii/S0047259X2200094X. Acesso em: 11 abr. 2023.

BOJER, C. S.; MELDGAARD, J. P.. Kaggle forecasting competitions: an overlooked learning opportunity. **International Journal of Forecasting**, [*s. l.*], v. 37, issue 2, p. 587-603, Apr.-Jun. 2021. Disponível em: https://doi.org/10.1016/j.ijforecast.2020.07.007. Acesso em: 29 nov. 2022.

BRAULE, R.. **Estatística aplicada com Excel**: para cursos de administração e economia. Rio de Janeiro: Campus, 2001.

BROOKS, C.. **Introductory econometrics for finance**. New York: Cambridge University Press, 2002.

BRUCE, P.; BRUCE, A.. **Estatística prática para cientistas de dados**: 50 conceitos essenciais. Rio de Janeiro-RJ: Alta Books, 2019.

BRUNI, A. L.. **Estatística aplicada à gestão empresarial**. 4. ed. São Paulo: Atlas, 2013.

BUSARI, G. A.; LIM, D. H.. Crude oil price prediction: a comparison between AdaBoost-LSTM and AdaBoost-GRU for improving forecasting performance. **Computers & Chemical Engineering**, [*s. l.*], v. 155, e-article 107513, December 2021. Disponível em: https://doi.org/10.1016/j.compchemeng.2021.107513. Acesso em: 27 dez. 2022.

CARMO, C. R. S.; CARMO, R. de O. S.. Motivação para aprendizagem no ensino superior: um estudo envolvendo o estágio curricular, alunos da modalidade presencial e alunos do curso a distância. **Cadernos da Fucamp**, [*s. l.*], v.13, n.18, p. 70-90, 2014. Disponível em: https://revistas.fucamp.edu.br/index.php/cadernos/article/view/363. Acesso em: 27 mar. 2023.

CARMO, C. R. S.; SILVA. J. R. de M.. Aprendizado de máquina e prestação de serviços de armazenamento de dados: métricas para análise e validação de algoritmos previsores. **Gestão, Tecnologia e Ciências,** [*s. l.*], v.12, n.38, p. 123-144, 2023. Disponível em: https://revistas.fucamp.edu.br/index.php/getec/article/view/2895. Acesso em: 29 mar. 2023.

DAM, R. S. de F.; SALGADO, W. L.; SCHIRRU, R.; SALGADO, C. M.. Application of radioactive particle tracking and an artificial neural network to calculating the flow rate in a two-phase (oil–water) stratified flow regime. **Applied Radiation and Isotopes**, [s. l.], v. 180, e-article 110061, February 2022. Disponível em: https://doi.org/10.1016/j.apradiso.2021.11006. Acesso em: 23 dez. 2022.

DOWNING, D.; CLARK, J.. **Estatística aplicada**. 2.ed. São Paulo: Saraiva, 2006.

FÁVERO, L. P.; BELFIORE, P.; SILVA, F. L. da; CHAN, B. L.. **Análise de dados:** modelagem multivariada para tomada de decisões. Rio de Janeiro: Elsevier, 2009.

FÁVERO, L.P.; BELFIORE, P. **Análise de dados:** estatística e modelagem multivariada com Excel, SPSS e Stata. Rio de Janeiro: Elsevier, 2017.

FIELD, A.. **Descobrindo a estatística usando o SPSS.** 2. ed. Porto Alegre: Artmed, 2009

FONSECA, J. S.; MARTINS, G. de A.; TOLEDO, G. L.. **Estatística aplicada**. 2. ed. São Paulo: Atlas, 1982.

FORBES, K. F.. Demand for grid-supplied electricity in the presence of distributed solar energy resources: Evidence from New York City. **Utilities Policy**, [*s. l.*], v. 80, e-article 101447, February 2023. Disponível em: https://doi.org/10.1016/j.jup.2022.101447. Acesso em 27 dez. 2022.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep learning**. Cambridge: MIT Press, 2016.

HAYKIN, S. **Redes neurais**: princípios e práticas. 2. ed. Porto Alegre: Bookman, 2007.

HYNDMAN, R. J.; KOEHLER, A. B.. Another look at measures of forecast accuracy. **International Journal of Forecasting**, [*s. l.*], n. 22, issue 4, p. 679-688, Oct.–Dec. 2006. Disponível em: https://doi.org/10.1016/j.ijforecast.2006.03.001. Acesso em: 29 nov. 2022.

HUANG, S.. Linear regression analysis. *In*: TIERNEY, R. J.; RIZVI, F.; ERCIKAN, K.. I**nternational encyclopedia of education**. 4th ed. [*S. l.*]: [*s. n.*], 2023. p. 548-557. Disponível em: https://www.sciencedirect.com/science/article/pii/B9780128186305100673. Acesso em: 09 abr. 2023.

HUANG, S. H. Z.; HIRAOKA, T. ; CHÁVEZ, A. P. de L.; ALA-NISSILA, T; LESKELÄ, L.; KIVELÄ, M.; SARAMÄKI, J..Estimating inter-regional mobility during disruption: Comparing and combining different data sources. **Travel Behaviour and Society**, [*s. l.*], v. 31, p. 93-105, April 2023. Disponível em: https://doi.org/10.1016/j.tbs.2022.11.005. Acesso em: 27 dez. 2022.

IBGE, Instituto Brasileiro de Geografia e Estatística. **Classificação nacional de atividades econômicas - CNAE**: versão 2.0. 2. ed. IBGE: Rio de Janeiro, 2015. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv93009.pdf. Acesso em: 11 set. 2022.

IBGE, Instituto Brasileiro de Geografia e Estatística. **Tabela 6449**: empresas e outras organizações, pessoal ocupado total, pessoal ocupado assalariado, salários e outras remunerações, por seção, divisão, grupo e classe da classificação de atividades (CNAE 2.0). Cadastro Central de Empresas (CEMPRE) - Banco de Metadados: IBGE, 2022. Disponível em: https://sidra.ibge.gov.br/tabela/6449#/n3/all/v/allxp/p/last%202/c12762/116830,116880,116910,117296,117307,117329,117363,117484,117543,117555,117608,117666,117673,117714,117774,117788,117810,117838,117861,117888,117892,117897/d/v662%200/l/v,p+c12762,t/cfg/cod,/resultado. Acesso em: 10 set. 2022-17:54h.

JIAO, Y.; WANG, Y; YANG, Y.. Approximation bounds for norm constrained neural networks with applications to regression and GANs. **Applied and Computational Harmonic Analysis**, [*s. l.*], v. 65, p. 249-278, July 2023. Disponível em: https://www.sciencedirect.com/science/article/pii/S1063520323000258. Acesso em: 10 abr. 2023.

KANG, M.; KANG, S.. Surrogate approach to uncertainty quantification of neural networks for regression. **Applied Soft Computing**, [*s. l.*], v. 139, e-article 110234, May 2023. Disponível em: https://www.sciencedirect.com/science/article/pii/S1568494623002521. Acesso em: 10 abr. 2023

KARUNASINGHA, D. S. K.. Root mean square error or mean absolute error? Use their ratio as well. **Information Sciences**, [*s. l.*], v. 585, p. 609-629, March 2022. Disponível em: https://doi.org/10.1016/j.ins.2021.11.036. Acesso em 26 dez. 2022.

LENZ, M. L.; NEUMAN, F. B.; SANTARELLI, R.; SALVADOR, D. **Fundamentos de aprendizagem de máquina**. Porto Alegre: SAGAH, 2020.

MARTINS, M. E.G. Coeficiente de determinação. **Rev. Ciência Elem.**, [*s. l.*], v. 6, n. 01, p. 01, mar. 2018. Disponível em: http://doi.org/10.24927/rce2018.024. Acesso em: 25 dez. 2022.

McKEAN, J. W.; SIEVERS, G. L.. Coefficients of determination for least absolute deviation analysis. **Statistics & Probability Letters,** [*s. l.*], v. 5, issue1, p. 49-54 January 1987. Disponível em: https://doi.org/10.1016/0167-7152(87)90026-5. Acesso em: 27 dez. 2022.

MOSTOUFI, N.; CONSTANTINIDES, A.. Linear and nonlinear regression analysis. In: MOSTOUFI, N.; CONSTANTINIDES, A.. **Applied numerical methods for chemical engineers**. [*S. l.*]: Academic Press, 2023. Chapter 8. p. 403-476. Disponível em: https://www.sciencedirect.com/science/article/pii/B978012822961300008X. Acesso em: 08 abr. 2023.

NETTO, A.; MACIEL, F. **Python para *data science* e *machine learning***: descomplicado. Rio de Janeiro: Alta Books, 2021.

SHEWA, G. A.; UGWUOWO, F. I.. A new hybrid estimator for linear regression model analysis: computations and simulations. **Scientific African**, [*s. l.*], v. 19, e-article 01441, 2023. Disponível em: https://www.sciencedirect.com/science/article/pii/S2468227622003465. Acesso em: 11 abr. 2023

SILVA, R. F. **Deep learning**. São Paulo: Platos Soluções Educacionais, 2021.

TOHME, T.; VANSLETTE, K; YOUCEF-TOUMI, K.. Reliable neural networks for regression uncertainty estimation. **Reliability Engineering & System Safety**, [*s. l.*], v. 229, e-article 108811, January 2023. Disponível em: https://www.sciencedirect.com/science/article/pii/S0951832022004306. Acesso em: 11 abr. 2023.

TRETIAK, K.; SCHOLLMEYER, G.; FERSON, S.. Neural network model for imprecise regression with interval dependent variables. **Neural Networks**, [*s. l.*], v. 161, p. 550-564, April 2023. Disponível em: https://www.sciencedirect.com/science/article/pii/S0893608023000680. Acesso em: 10 abr. 2023.