MACHINE LEARNING E PEQUENAS AMOSTRAS DE DADOS: UM ESTUDO DE CASO COM PROFISSIONAIS DE CONTABILIDADE E APLICAÇÃO DO MÉTODO DE *BOOTSTRAP*

CARLOS ROBERTO SOUZA CARMO[1]

RESUMO

Seja pelo elevado custo da aquisição de dados, seja pela presença de ruídos amostrais, as dificuldades relacionadas à obtenção de dados úteis, mesmo no contexto do *big data*, podem inviabilizar pesquisas no campo do *machine learning*. Pois, apesar de reproduzir o sistema de aprendizado humano, as máquinas demandam maiores quantidades de casos (observações) para reconhecer padrões implícitos. Nesse contexto, esta investigação buscou avaliar como o método *bootstrap* poderia auxiliar na melhora do desempenho do aprendizado de máquina baseado em redes neurais artificiais (RNA). Para tanto, foi desenvolvido um estudo de caso único acerca de certo fenômeno social, na área de contabilidade, composto por 1 variável dependente e 13 possíveis variáveis explicativas; por meio do qual se aplicou o método de reamostragem, com sorteio por randomização e reposição de dados em uma amostra inicialmente formada por 44 observações, cuja representatividade era inferior a 1% da respectiva população de interesse. Após construir e avaliar o desempenho de seis RNA, uma delas pesquisada com base naquela "amostra original" e outras cinco com base em amostras de *bootstrap* com tamanhos diversos, foi possível constatar que a metodologia em questão se mostrou uma ferramenta muito promissora. Assumindo especial relevância naquelas situações em que a obtenção de grandes amostras pode constituir-se em um fator de insucesso; além da possibilidade de sua utilização para a detecção de problemas relacionados à presença de *overfitting*, assim como, para a seleção de variáveis relevantes para o processo de *machine learning* e/ou para eliminação de parâmetros de entradas inexpressivos.

Palavras-chave: métodos computacionais *overfitting*; reamostragem; *bootstrap*.

[1] Doutor em Agronomia com ênfase em Energia na Agricultura pela Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP). Mestre em Ciências Contábeis pela PUC-SP. Especialista em Ciência de Dados e Big Data Analytics (2024), Data Mining (2024), Análise e Desenvolvimento de Sistemas em Python (2023). MBA em Controladoria e Finanças (2001). Professor adjunto da Faculdade de Ciências Contábeis da Universidade Federal de Uberlândia (FACIC-UFU). https://orcid.org/0000-0002-3806-9228. carlosjj2004@hotmail.com.

MACHINE LEARNING AND SMALL DATA SAMPLES: A CASE STUDY WITH ACCOUNTING PROFESSIONALS AND APPLICATION OF THE BOOTSTRAP METHOD

ABSTRACT

Whether due to the high cost of data acquisition or the presence of sample noise, the difficulties related to obtaining useful data, even in the context of big data, can make research in the field of machine learning unfeasible. Because, despite reproducing the human learning system, machines require greater quantities of cases (observations) to recognize implicit patterns. In this context, this investigation sought to evaluate how the bootstrap method could help improve the performance of machine learning based on artificial neural networks (ANN). To this end, a single case study was developed about a certain social phenomenon, in the area of accounting, consisting of 1 dependent variable and 13 possible explanatory variables; through which the resampling method was applied, with randomization and data replacement in a sample initially formed by 44 observations, whose representation was less than 1% of the respective population of interest. After building and evaluating the performance of six ANN, one of which was researched based on that "original sample" and another five based on bootstrap samples of different sizes, it was possible to verify that the methodology in question proved to be a very promising tool. Assuming special relevance in those situations in which obtaining large samples may constitute a factor of failure; in addition to the possibility of using it to detect problems related to the presence of overfitting, as well as to select relevant variables for the machine learning process and/or to eliminate insignificant input parameters.

Keywords*: computational overfitting methods; resampling; bootstrap.*

1 Introdução

A base do método estatístico concentra-se na realização de inferências sobre o comportamento de populações desconhecidas, a partir de amostras de dados conhecidas, mediante o uso de testes, parâmetros e modelos analítico-preditivos. Contudo, a pouca disponibilidade de dados para composição das amostras pode inviabilizar o processo de extração do conhecimento acerca de determinado objeto de estudo, tornando o processo de amostragem um fator crítico para o sucesso de pesquisas (Carmo; Lima, 2018) e da tomada de decisões baseada no conhecimento obtido mediante a análise de dados.

De forma análoga ao que acontece no ramo da estatística, a amostra assume especial relevância no contexto do aprendizado de máquina (*machine learning*) baseado em redes neurais artificiais (RNA); pois, a partir de uma única amostra extraem-se dois conjuntos de dados, sendo, um é destinado ao treinamento enquanto o outro é utilizado para testes acerca do aprendizado realizado pelas RNA, o que torna ambos essenciais ao bom desempenho do algoritmo utilizado (Pereira; Centeno, 2017).

Porém, por quaisquer que sejam os motivos, a impossibilidade de se obter amostras com uma quantidade de observações (*n*) suficiente para o processo de *machine learning* pode levar a erros de generalização caracterizados pelo *overfitting*, que é o fenômeno pelo qual uma RNA se ajusta muito bem à amostra de dados utilizada para treinamento, entretanto, ela não consegue generalizar o aprendizado quando aplicada ao conjunto de dados utilizados para teste. Sendo que, a explicação mais simples, porém lógica, para esse tipo ocorrência reside no fato de que as máquinas demandam quantidades de exemplos muito maiores que os seres humanos para que possam reconhecer padrões de comportamentos.

Assim, diante de possíveis dificuldades relacionadas à obtenção de dados para a composição de grandes amostras de pesquisa, mesmo no contexto do *big data*, esta investigação teve por objetivo avaliar como o método *bootstrap* poderia auxiliar na melhora do desempenho do *machine learning* baseado em redes RNA.

Nesse sentido, inicialmente, foi composta a base teórica desta investigação, segundo a qual foram abordados os seguintes assuntos: aspectos referentes ao processo de *machine learning* diante dos problemas relacionados à obtenção de grandes amostras de dados; as possíveis alternativas apresentadas para a solução desse tipo problema; e ainda, a possibilidade de utilização do método de *bootstrap* para solução de problemas dessa natureza. Na sequência, foi desenvolvido um estudo de caso único acerca de certo fenômeno social, no qual, se contava com uma amostra composta por apenas 44 observações, cuja representatividade era inferior a 1% da respectiva população de interesse, em que, foram utilizadas RNA no processo de *machine learning* sobre o comportamento de 1 variável dependente e 13 possíveis variáveis explicativas. Nesse mesmo estudo de caso foi aplicado o método de *bootstrap* com o objetivo de se promover possíveis melhoras no processo de aprendizagem daquelas RNA, bem como, a identificação e solução de problemas relacionados à presença de *overfitting*.

2 Referencial Teórico

O atual sucesso da Inteligência Artificial (IA) advém da combinação entre a grande disponibilidade de dados e a alta capacidade computacional das máquinas modernas (Ludemir, 2021). Contudo, para um bom desempenho no processo de *machine learning*, a quantidade de exemplos demandados é muito maior que a quantidade requerida pelos seres humanos para aprender (Ludemir, 2021).

Adicionalmente, destaca-se o fato de que o *machine learning* baseado em RNA é extremamente dependente do arranjo dos dados utilizados para treinamento e teste, o que leva a três fontes de incertezas, ou seja: a incerteza relacionada à qualidade e à representatividade dos dados; a incerteza relacionada à estruturação do modelo em si; e, a incerteza envolvendo a parametrização do modelo pesquisado (Tiwari; Chatterjee, 2010).

Especificamente em relação às incertezas referentes à qualidade e à representatividade dos dados, soma-se uma nova variável que pode ser considerada

crítica, isto é, devido ao rápido avanço do *machine learning*, o que levou à sociedade a experimentar aplicações de IA em larga escala, é cada vez mais recorrente a implementação de modelos orientados por dados cuja disponibilidade nem sempre é suficiente (Li *et al.*, 2018; Zhu *et al.*, 2023). Assim, mesmo na era do *big data*, o problema relacionado às pequenas amostras de dados é uma questão que não pode ser ignorada, tornando-se uma área de pesquisa relevante, para a qual as diferentes alternativas de soluções propostas pela literatura podem ser agrupadas em três categorias básicas (Zhu *et al.*, 2023):

a) a primeira é a dos modelos baseados em lógica *fuzzy*, por exemplo as *Grey Forest Model*, contudo, seu melhor desempenho está condicionado à baixa dimensão e/ou linearidade da distribuição dos dados, o que pode ser considerado contraditório pois pequenas amostras tendem a ser multidimensionais e/ou não lineares;

b) a segunda é baseada em modelos de *machine learning*, por exemplo o *Support Vector Machine*, contudo, a presença de *outliers* individuais em uma amostra pequena tende a afetar significativamente o desempenho desse tipo de algoritmo; e

c)a terceira consiste na geração de amostras virtuais, nas quais, estudam-se os padrões de distribuição de uma amostra de dados originais, com uma pequena quantidade de observações, e realiza-se a geração de amostras virtuais com os mesmo perfil observado na “amostra original”.

Especificamente em relação ao processo de geração de amostras virtuais, destacam-se os estudos voltados para o método de *bootstrap*, que consiste na reamostragem intensiva com reposição, sendo caracterizada como uma das abordagens mais simples de se implementar, uma vez que não demanda cálculos complexos, exigindo somente a randomização do processo de reamostragem em si (Tiwari; Chatterjee, 2010).

São muitas as aplicações do método de *bootstrap*: previsão de séries temporais, permitindo a estimação de intervalos de confiança para a previsão e a avaliação da adequação de modelos aplicados (Diciccio; Efron, 1996; Künsch, 1989);

a análise de dados de sequenciamento de DNA (Efron; Halloran; Holmes, 1996); modelagem de equações estruturais (Hu; Bentler, 1999); análise de dados de expressão gênica, permitindo a identificação de genes significativos (Efron; Tibshirani, 2002), entre outras.

Ao buscar avaliar como o método *bootstrap* poderia auxiliar na melhora do desempenho do aprendizado de máquina baseado em RNA, esta pesquisa científica emprega a alta capacidade computacional das máquinas modernas, observada por Ludemir (2021), e ainda, combina duas técnicas distintas (*machine learning* e geração de amostras virtuais) consideradas como possíveis alternativas para solucionar problemas relacionados às pequenas amostras de dados, conforme aquela classificação proposta por Zhu *et al.* (2023).

Em geral, as abordagens acerca do processo de modelagem baseada em dados podem ser agrupadas em duas categorias distintas: as paramétricas e as não paramétricas; sendo que, a maioria das RNA são caracterizadas como não paramétricas, uma vez que não requerem pressupostos em relação às respectivas funções de mapeamento, dispensando o conhecimento prévio envolvendo a natureza das distribuições dos dados integrantes das amostras de pesquisa utilizadas para treinamento dessas redes (Gu; Wei, 2018). Contudo, ainda assim, o processo de modelagem para *machine learning* orientado a dados requer uma quantidade de dados elevada para que se possa extrair o conhecimento contido nos respectivos objetos de estudo e modelagem (Shen; Qian, 2022).

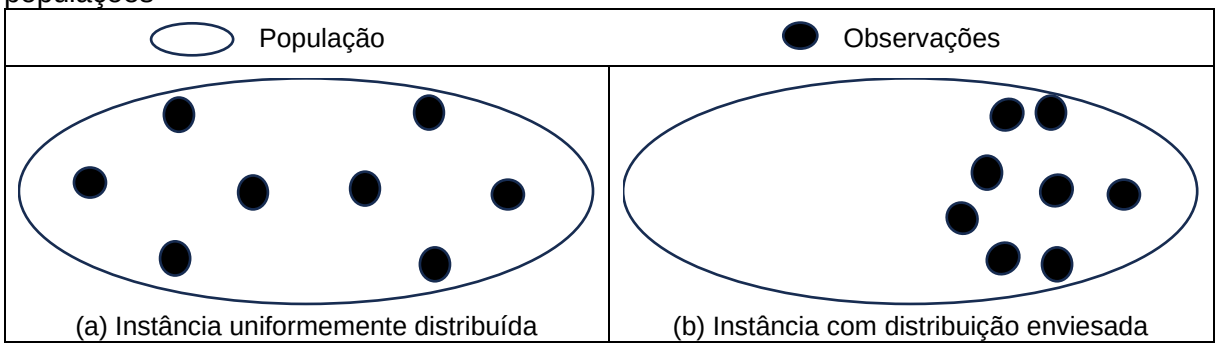
Ao realizar algumas suposições específicas, e ainda, sob certas condições, a maior parte dos métodos de modelagem baseada em dados funciona bem e confiavelmente para a maioria das aplicações utilizadas no processo de *machine learning*, entretanto, quando a quantidade de exemplos integrantes das amostras utilizadas no processo de modelagem baseado em RNA é muito pequena, a incerteza gerada influencia negativamente o desempenho e a acurácia dos modelos pesquisados (Gu; Wei, 2018). Isso acontece porque, a despeito da proposição de uma expressiva quantidade de algoritmos de aprendizagem voltados para a extração de conhecimento com base em dados nas últimas décadas, a maioria deles toma com

suposição básica o fato de que as amostras utilizadas para treinamento, de RNA por exemplo, trazem consigo as características das respectivas populações, e isso nem sempre acontece, especialmente, quando essas amostras apresentam poucas observações (Li *et al*., 2018).

Apesar da capacidade informacional dos modelos de *machine learning* baseados em dados, bem como, da sua relevância no processo de suporte à tomada de decisões, nem sempre é possível obter amostras com um grande número de observações, quer seja por restrições relacionadas a custos e tempo, quer seja por limitações inerentes à própria natureza dos dados que se pretende analisar, por exemplo: previsão precoce de doenças; diagnóstico de doenças raras; desenvolvimento de novos produtos; reconhecimento de padrões de pixels limitados (Li *et al*., 2018); contratação de funcionários; peças para manutenção preventiva (Lin *et al*., 2023); entre tantas outros situações possíveis.

Uma boa explicação para tanto pode ser observada no exemplo descrito pela Figura 1, no qual, a Figura 1(a) apresenta uma instância com poucas observações uniformemente distribuídas, e a Figura 1(b) também representa uma instância com poucas observações, porém, bem mais concentradas. Por isso, na Figura 1(b) podem ser constatadas lacunas de informações, já que, além de serem concentradas, suas observações não estão uniformemente distribuídas entre si. Dessa maneira, a maior parte dos algoritmos de *machine learning* falha na identificação de padrões com base em situações cuja composição é baseada em pequenos conjuntos de dados (Li *et al*., 2018), cujo comportamento tende a seguir o descrito pela Figura 1(b).

Figura 1 - Possíveis distribuições de dois pequenos conjuntos de dados em relação às suas populações



Fonte: adaptado de Li *et al*. (2018).

Enquanto uma das ferramentas de *machine learning* mais utilizadas, as RNA podem apresentar desempenhos bem satisfatórios no processo de modelagem de problemas complexos e de elevado dimensionamento (Gu; Wei, 2018). Contudo, diante da impossibilidade de se utilizar uma amostra com a quantidade de dados suficientes para seu treinamento, os algoritmos de *machine learning* baseado em RNA podem desenvolver problemas relacionados ao *overfitting* (Collins *et al*., 2021).

O *overfitting* é um sério problema ocorrido no treinamento de RNA do tipo perceptron multicamadas (MLP), caracterizado pelo sobreajuste no processo de aprendizagem (Murakoshi, 2005), o que leva a erros de generalização no processo de teste e/ou aplicação da rede (Jia; Culver, 2006). Ou seja, a RNA se ajusta muito bem à amostra de dados utilizadA para aprendizagem, contudo, ela é incapaz de generalizar esse aprendizado quando aplicada ao conjunto de dados utilizados para teste, denotando baixos índices de erros no treinamento e, em contrapartida, elevados índices de erros no processo de teste da RNA MLP (Wang *et al.*, 2023).

Para solucionar esse tipo de problema, o método de *bootstrap* pode caracterizar-se como uma alternativa muito útil. Ou seja, após identificar o perfil da "amostra original", a partir de determinados parâmetros de interesse, e aplicar o processo de reamostragem por sorteio randomizado com reposição, o método de *bootstrap* permite reproduzir *n* amostras com *n* quantidades de observações, denominadas de amostras de *bootstrap* (Ma; Leng; Wang, 2024). Por sua vez, as amostras de *bootstrap* podem ser utilizadas para treinar e estimar os erros de

generalização das respectivas RNA, identificando-se aquelas com maior precisão preditiva (Jia; Culver, 2006), e ainda, cujas métricas avaliativas dos índices de erro se apresentarem mais homogêneas tanto na fase de treinamento quanto na fase teste, afastando a possibilidade de ocorrência do *overfitting*.

Nesse sentido, ao levar em conta que apesar de reproduzir a sistema de aprendizado humano, as máquinas demandam maiores quantidades de casos (observações) para reconhecer padrões implícitos, e ainda, diante de possíveis dificuldades relacionadas à obtenção de dados para a composição de grandes amostras de dados, mesmo no contexto do *big data*, o que poderia inviabilizar pesquisas que utilizassem o *machine learning* baseando em RNA, vislumbra-se a possibilidade desta investigação permitir avaliar como o método *bootstrap* poderia auxiliar na melhora do desempenho do aprendizado de máquina baseado em RNA MLP, assim como foi proposto por Carmo e Lima (2018) em um estudo de natureza análoga a esta pesquisa, porém, aplicado ao contexto exclusivamente estatístico.

3 Metodologia de Pesquisa

Para realização desta pesquisa foi desenvolvido um estudo de caso único baseado em uma amostra de dados coletada junto a um grupo de profissionais de contabilidade da cidade Uberlândia-MG, mediante a utilização de um questionário composto por 14 afirmativas, para as quais cada respondente deveria atribuir qualquer nota entre 0 (zero) e 10 (dez) que indicasse o seu grau de concordância em relação às respectivas proposições. O instrumento de coleta utilizado apresentava uma proposição inicial de caráter geral, na qual afirmava-se que a legislação e a evolução tecnológica observadas ao longo dos anos teriam impactado o exercício da profissão contábil, provocando mudanças no “fazer contábil”; entretanto, tal afirmativa inicial não fazia menção a qualquer tipo de variável específica (fato ou evento) que pudesse influenciar tal fenômeno. A seguir, o mesmo instrumento de coleta apresentava proposições que realizavam afirmações específicas sobre um conjunto de 13 variáveis e períodos de ocorrência que teriam influenciado pontualmente o exercício contábil-

profissional, ainda que em forma e intensidade distintas, ao longo dos anos compreendidos entre 2007 (inclusive) e 2018 (inclusive).

Na pesquisa em questão, foram coletados os dados referentes às notas atribuídas por 44 respondentes, o que representou cerca 0,85% do total profissionais de contabilidade vinculados à cidade em questão, uma vez que consultas realizadas junto ao Conselho Regional de Contabilidade de Minas Gerais (CRC-MG) sinalizaram a existência de 5159 profissionais de contabilidade vinculados ao município de Uberlândia-MG (CRC-MG, 2023). Ao buscar avaliar como esses profissionais percebiam os impactos da legislação tributária e da evolução tecnológica sobre o exercício da profissão contábil como um todo (variável dependente), e ainda, como o conjunto formado por 13 fenômenos ocorridos ao longo dos anos 2007 a 2018 (variáveis independentes) poderia explicar tal processo, observou-se o perfil descritivo resumido na Tabela 1, para a amostra de dados utilizada nesta pesquisa como "amostra mestre".

Tabela 1 – Perfil da "amostra mestre", para um total de 44 observações (*n* = 44)

Variáveis (ano de ocorrência)	Tipo de variável	Média	Limite. inferior	Limite superior	Desvio padrão	Mediana	Máximo	Mínimo
Evolução tecnológica (2007-2018)	Dependente	5,80	5,04	6,55	2,57	6,50	9,00	1,00
Instituição do SPED (2007)	Independente	5,95	4,78	7,13	3,96	7,00	10,00	0,00
Util. de ERP p/ SPED (2008)	Independente	6,41	5,59	7,22	2,76	6,50	10,00	0,00
Obrig. da Nfe (2008)	Independente	6,18	5,09	7,27	3,68	7,00	10,00	0,00
SPED p/ lucro real (2010)	Independente	8,95	8,59	9,32	1,24	9,00	10,00	4,00
Obrig. da ECD (2011)	Independente	6,66	5,74	7,58	3,12	8,00	10,00	0,00
Startup cont. (2011)	Independente	5,98	5,10	6,85	2,95	5,00	10,00	0,00
Contabilidade *on-line* (2012)	Independente	5,80	4,98	6,61	2,75	6,00	10,00	0,00
Obrig. da EFD (2014)	Independente	6,70	5,74	7,67	3,27	6,50	10,00	0,00
Institit. do eSocial (2014)	Independente	5,55	4,53	6,56	3,42	5,50	10,00	0,00
Surg. Contab. Digital (2015)	Independente	5,82	4,97	6,66	2,86	5,50	10,00	0,00
Asses. e serv. Contab. (2015)	Independente	5,57	4,70	6,44	2,95	5,00	10,00	0,00
Implem, Cont. Digital (2017)	Independente	5,95	4,96	6,95	3,38	7,00	10,00	0,00
Obrigat, do eSocial (2018)	Independente	5,84	4,76	6,92	3,65	6,50	10,00	0,00

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

Vale destacar que a pesquisa que originou a “amostra mestre” utilizada para composição do estudo de caso cujos dados serviram de base para apresente investigação não chegou a ser realizada devido à pequena quantidade de respondentes. Sendo que, a partir do estudo de caso ora proposto, foi avaliado como o método de *bootstrap* poderia ser utilizado para viabilizar estudos baseados em RNA MLP, mesmo com base em uma amostra considerada pequena, ou seja, menos de 1% (0,85%) do total de profissionais em questão.

A partir daquela “amostra mestre” composta por 44 observações (n=44) foi implementado o método de reamostragem, com sorteio por randomização e reposição de dados, e foram geradas 5 amostras de *bootstrap* com quantidades de observações distintas, porém, todas com o mesmo perfil da “amostra original” descrito na Tabela 1. Dessa maneira, formou-se uma base de dados com 5 amostras de *bootstrap* que apresentaram as seguintes composições: “*bootstrap* 190” com 190 observações (n=190); “*bootstrap* 475” com 475 observações (n=475); “*bootstrap* 950” com 950 observações (n=950); “*bootstrap* 1425” com 1425 observações (n=1425); e, “*bootstrap* 1900” com 1900 observações (n=1900). E, a seguir, foram implementados os processos de treinamento e teste das respectivas RNA MLP, além de uma RNA MLP baseada na “amostra original” (n=44).

A definição da amostra de *bootstrap* inicial, portanto, aquela com 190 observações (n=190) teve como base as propostas de Effron e Tibshirani (1993), e ainda, Carmo e Lima (2018). As demais amostras de *bootstrap* são múltiplos de 190 e 2,5 (n = 475 observações = 190 x 2,5; n = 950 observações = 190 x 5,0; n = 1425 observações = 190 x 7,5; e, n = 1900 observações = 190 x 10,0).

Effron e Tibshirani (1993) ponderam que a quantidade ideal de reamostras não precisa ser maior que 200 replicações amostrais aleatórias ($n \leq 200$). Por sua vez, ao testarem os parâmetros propostos por Montgomey, Peck e Vining (2001), comparativamente a Effron e Tibshirani (1993), ou seja, se a geração de amostras de *bootstrap* deveria ocorrer até que fosse constatada a estabilização das respetivas médias e os correspondentes desvios, Carmo e Lima (2018) observaram que amostras entre 44 e 55 observações ($44 \leq n \leq 55$) já seriam suficientes. Assim, uma

vez que a “amostra original” já contava com 44 observações, optou-se por produzir uma amostra de *bootstrap* inicial ligeiramente inferior à quantidade proposta por Effron e Tibshirani (1993) (190 < 200), e ir aumentando progressiva e arbitrariamente na ordem de 2,5 vezes.

No contexto em análise, a implementação das RNA do tipo *perceptron* de múltiplas camadas (RNA MLP) teve por objetivo realizar a aprendizagem sobre como a percepção dos contadores acerca daqueles 13 fatores identificados como variáveis independentes poderiam explicar a sua percepção geral sobre o impacto da legislação e da evolução tecnológica na profissão contábil como um todo, aqui identificada como variável dependente, segundo as notas fornecidas pelos 44 respondentes integrantes da pesquisa cuja série de dados apresentou o perfil descrito na Tabela 1. Uma RNA MLP possui uma estrutura composta por, no mínimo, três camadas, das quais: a primeira delas é denominada camada de entrada; uma ou mais camadas ocultas, onde ocorrem as sinapses da rede; e, uma terceira camada denominada camada de saída (CARNEIRO JÚNIOR; SOUZA, 2019). Na camada de entrada, ocorre a inserção dos parâmetros das variáveis preditoras no processo de aprendizagem; nas camadas ocultas ou intermediárias, ocorre o processamento a partir dos neurônios da RNA, e ainda, os ajustes dos pesos sinápticos e funções; e, na camada de saída, identificam-se os parâmetros previstos (CARNEIRO JÚNIOR; SOUZA, 2019).

Além da análise de desempenho de cada uma das 6 RNA MLP implementadas a partir de conjuntos de dados com tamanhos amostrais diferentes (5 amostras de *bootstrap* com *n* diversos e 1 “amostra original” com *n*=44 observações), todas essas RNA foram utilizadas para a estimativa da variável dependente a partir de uma amostra totalmente independente, composta pelas respostas fornecidas por 19 profissionais contábeis diferentes daqueles 44 contadores iniciais. Tal procedimento teve por finalidade avaliar o desempenho de todas as RNA em relação a uma amostra distinta daquela “amostra original” ou amostra mestre. Nessa etapa, foi utilizado o conjunto de métricas para análise e validação de algoritmos previsores propostas por Carmo e Silva (2023), descrito na Figura 2.

Figura 2 – Relação das métricas de precisão utilizadas para avaliação de desempenho das RNA MLP

Descrição	**Fórmula**	**Unidade de medida (un) e parâmetro de decisão**
Determination coefficient ou coeficiente de determinação (R^2)	$R^2\ (y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$	$0 < R^2 < 1$ 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*)	$MAPE\ (y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert y_i - \hat{y}_i\rvert}{\lvert y_i\rvert}$	$0 < MAPE < 1$ sendo que: quanto menor melhor
Median absolute percentage error ou erro percentual absoluto mediano (*MdAPE*)	$MdAPE\ (y,\hat{y}) = \frac{\lvert y_i - \hat{y}_i\rvert}{\lvert y_i\rvert}$	$0 < MdAPE < 1$ sendo que: quanto menor melhor
Symmetric mean absolute percentage error ou erro percentual médio absoluto simétrico (*SMAPE*)	$SMAPE\ (y,\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\lvert y_i - \hat{y}_i\rvert}{(\lvert y_i\rvert + \lvert\hat{y}_i\rvert)/2}$	$0 < SMAPE < 1$ sendo que: quanto menor melhor
Median symmetric absolute percentage error ou erro percentual absoluto simétrico mediano (*MdSAPE*)	$MdSAPE\ (y,\hat{y}) = \frac{\lvert y_i - \hat{y}_i\rvert}{(\lvert y_i\rvert + \lvert\hat{y}_i\rvert)/2}$	$0 < MdSAPE < 1$ sendo que: quanto menor melhor
Weighted mean absolute percentage error ou erro percentual médio absoluto ponderado (*WMAPE*)	$WMAPE\ (y,\hat{y}) = \frac{\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert}{\sum_{i=1}^{n}\lvert y_i\rvert}$	$0 < WMAPE < 1$ 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; *ŷ* = valor previsto para a variável analisada com base em um modelo; *n* = quantidade total de observações referentes a *y* e/ou *ŷ* ; e *i* = cada observação específica de *y* e/ou *ŷ* .		

Fonte: adaptado de Carmo e Silva (2023).

Em função da quantidade e diversidade das métricas compiladas por Carmo e Silva (2023), resumidas na Figura 2, a avaliação acerca da possibilidade de utilização do método de *bootstrap* para detecção de problemas relacionados à presença de *overfitting* adotou como parâmetro somente o *Mean absolute percentage error* (*MAPE*) - ou erro percentual médio absoluto, ou ainda, erro relativo médio – conforme descrito pela Equação 1. Tal decisão se justifica pelo fato de que, além de avaliar o quanto os valores previstos se distanciam dos respectivos valores reais, assim como acontece nas demais métricas, o *MAPE* é calculado como uma unidade de medida relativa de fácil leitura e compreensão (percentual de erro). Por isso, é possível utilizá-lo nas comparações de desempenho entre modelos cujos variáveis sejam expressas em unidades de medidas diferentes (KARAMIRAD *et al.*, 2013). Além disso, o *MAPE* é uma métrica que não sofre influência de grandes quantidades observações (*n*); pelo contrário, ele tende a apresentar-se enviesado somente quando os valores previstos forem muito pequenos (HYNDMAN; KOEHLER, 2006).

$$MAPE\ (y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|}\ 100 \qquad (1)$$

Segundo a formulação descrita na Equação 1, que apresenta uma pequena variação em relação ao *MAPE* descrito na Figura 2, conforme o compilado de métricas realizado por Carmo e Silva (2023), seu valor varia de 0% a 100%, e quanto menor melhor. Contudo, um *MAPE* de até 10% já indica uma alta precisão de previsão, e ainda, um MAPE entre 10% e 20% pode indicar boa capacidade preditiva (YADAV; MALIK; CHANDEL, 2014).

Por fim, a partir da análise de sensibilidade, foi avaliada a importância de cada uma das variáveis independentes na composição das RNA implementadas com base em amostras distintas. A análise de sensibilidade permite quantificar a importância relativa de cada parâmetro de entrada do modelo (variáveis independentes) na determinação do valor da variável de saída atribuída (HOMMA; SALTELLI, 1996). Para tanto, modificam os valores das variáveis independentes uma a uma e mede-se

como a resposta do modelo baseado em RNA se comporta (JECZMIONEK; KOWALSKI, 2022).

Dessa forma, ao considerar seu objeto de estudo, a metodologia analítica e a base de dados utilizada, esta investigação classifica-se como uma pesquisa científica de natureza empírica, baseada em métodos quantitativos e computacionais, aplicada à análise de fenômenos sociais, mediante o uso de algoritmos previsores.

4 Análise de Dados e Apresentação dos Resultados da Pesquisa

Após implementar cada uma das RNA MLP com base naqueles 6 conjuntos de dados com tamanhos amostrais distintos, ou seja, 1 RNA com base na "amostra original" com *n*=44 observações e outras 5 RNA com base nas amostras de *bootstrap* com *n* diversos, procedeu-se à análise de desempenho global de cada uma delas, conforme pode ser observado na Tabela 2. Pôde-se constatar que, de uma forma geral, todas as RNA MLP pesquisadas com base nas amostras de *bootstrap* apresentaram um desempenho superior ao da RNA MLP pesquisada com base na "amostra original".

Tabela 2 – Avaliação de desempenho das RNA MLP por amostra

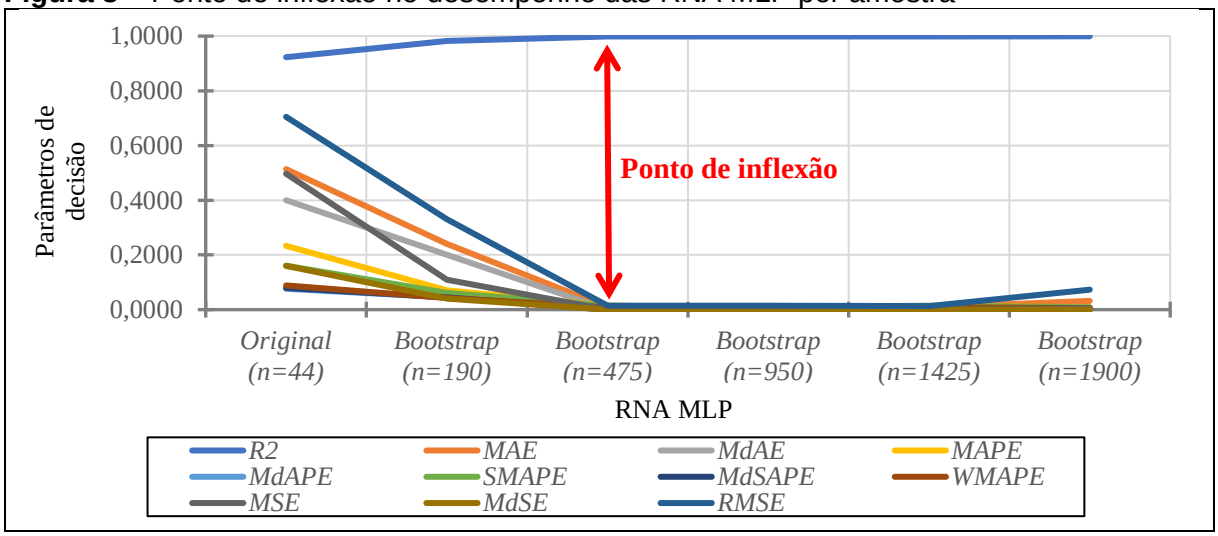
Amostra / Parâmetros	*Original (n=44)*	*Bootstrap (n=190)*	*Bootstrap (n=475)*	*Bootstrap (n=950)*	*Bootstrap (n=1425)*	*Bootstrap (n=1900)*
R2	0,9228	0,9822	0,9989	0,9990	0,9991	0,9992
MAE	0,5136	0,2405	0,0089	0,0063	0,0086	0,0317
MdAE	0,4000	0,2000	0,0000	0,0000	0,0053	0,0000
MAPE	0,2333	0,0702	0,0149	0,0085	0,0139	0,0090
MdAPE	0,0764	0,0444	0,0000	0,0000	0,0049	0,0000
SMAPE	0,1608	0,0614	0,0148	0,0086	0,0142	0,0090
MdSAPE	0,0794	0,0455	0,0000	0,0000	0,0049	0,0000
WMAPE	0,0886	0,0426	0,0089	0,0063	0,0086	0,0055
MSE	0,4968	0,1088	0,0000	0,0000	0,0000	0,0054
MdSE	0,1600	0,0400	0,0000	0,0000	0,0000	0,0000
RMSE	0,7049	0,3298	0,0146	0,0140	0,0130	0,0732

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

Ainda segundo aquele conjunto métricas de desempenho descrito na Tabela 2, pôde-se observar a ocorrência de uma melhora de desempenho global das RNA MLP

pesquisadas com base nas amostras de *bootstrap* à medida que a quantidade de observações (*n*) aumentava. Tal inferência fica mais evidente a partir da análise do ponto de inflexão destacado no gráfico descrito na Figura 3, ou seja, à medida que a quantidade de observações (*n*) cresce (veja o eixo das abscissas), todas as métricas de desempenho (com valores no eixo das ordenadas) apresentam uma melhora significativa até *n*=475 e, a partir daí, tal desempenho tende a se estabilizar.

Figura 3 – Ponto de inflexão no desempenho das RNA MLP por amostra



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

Esse primeiro conjunto de evidências corrobora com o que foi observado por Ludemir (2021), ou seja, as máquinas demandam uma quantidade maior de dados para reconhecer padrões de promover o *machine learning* de forma satisfatória. E, assim sendo, o método de *bootstrap* permite solucionar problemas relacionados à pouca disponibilidade dados para o processo de aprendizagem das máquinas, com especial atenção ao *machine learning* baseado em RNA MLP, conforme sugerido por Tiwari e Chatterjee (2010).

Na sequência, ao utilizar o *MAPE* percentual (conforme já descrito na Equação 1) para detectar problemas relacionados à presença de *overfitting*, com base naquelas 6 amostras de pesquisa e nas respectivas RNA MLP, o método de *bootstrap* também se mostrou promissor. Pois, conforme as informações resumidas na Tabela 3: na

"amostra original", o *MAPE* é maior na RNA MLP cujas observações (*n*) foram utilizadas para treinamento, comparativamente às observações utilizadas para teste; a seguir, à medida que a quantidade de observações aumenta e o *MAPE* é calculado para a amostra de *bootstrap* com *n*=190, o erro nas observações utilizadas para treinamento cai, contudo, ele é maior na amostra de teste, revelando a possível presença de *overfitting* para essa RNA MLP; finalmente, quando o *MAPE* é calculado para o treinamento e para as previsões realizadas pelas RNA MLP pesquisadas com base na amostra de *bootstrap* com *n*=475, *n*=950, *n*=1425 e *n*=1900, o erro percentual médio absoluto se estabiliza e, principalmente, se iguala tanto nas observações utilizadas nos processos de treinamentos quanto naquelas utilizadas para testes das respectivas RNA.

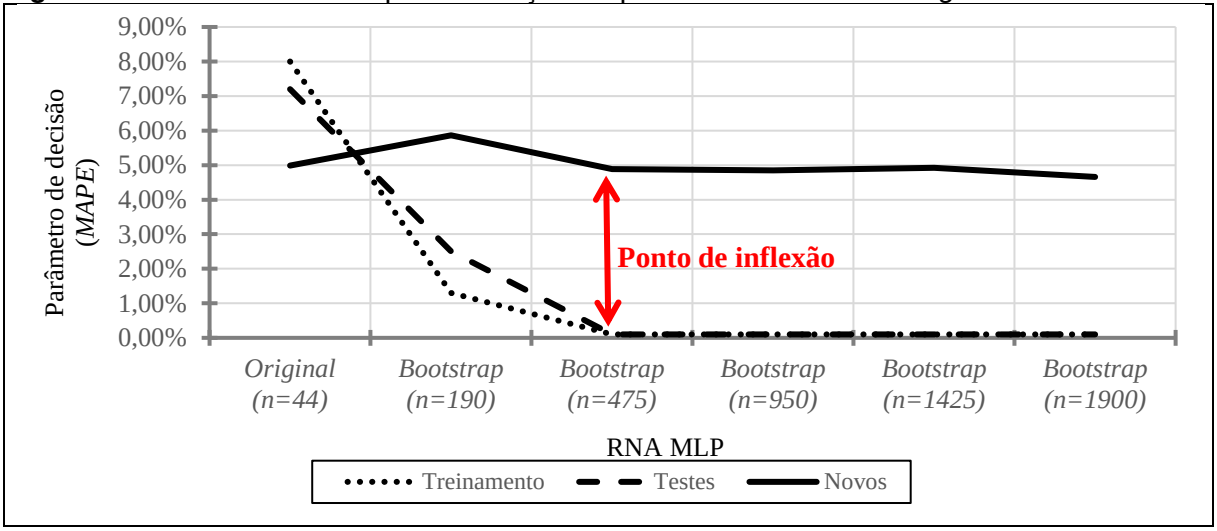
Tabela 3 – Análise do método de *bootstrap* para detecção de problemas com *overfitting*

Parâmetros \ Amostra		*Original (n=44)*	*Bootstrap (n=190)*	*Bootstrap (n=475)*	*Bootstrap (n=950)*	*Bootstrap (n=1425)*	*Bootstrap (n=1900)*
n	Treinamento	35	133	325	660	992	1324
	Testes	9	57	150	290	433	576
	Novos	19	19	19	19	19	19
MAPE	Treinamento	8,00%	1,30%	0,10%	0,10%	0,10%	0,10%
	Testes	7,20%	2,50%	0,10%	0,10%	0,10%	0,10%
	Novos	4,99%	5,86%	4,88%	4,85%	4,92%	4,66%

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

Dessa forma, segundo os dados descritos na Tabela 3, foi constatado que a utilização de amostras geradas a partir do método de *bootstrap* permite realizar inferências voltadas para a detecção do *overfitting* em RNA MLP implementas a partir de amostras com o mesmo perfil descritivo, porém, com diferentes quantidades de observações. Essa inferência pode ser melhor observada mediante a análise do ponto de inflexão destacado no gráfico descrito na Figura 4.

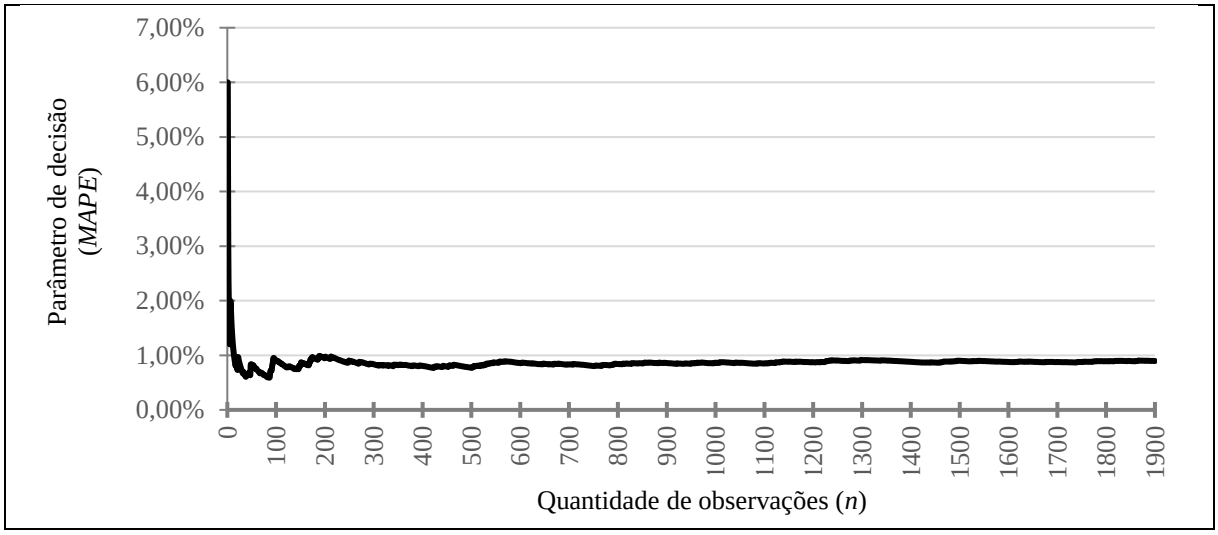
Figura 4 – Ponto de inflexão para detecção de problemas com *overfitting*



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

Esse segundo conjunto de evidências corrobora com o que foi sugerido por Tiwari e Chatterjee (2010), Lee, Ullah e Wang (2020) e Ludemir (2021), acerca da utilização de amostras de *bootstrap* para detecção do *overfitting*. Além disso, ao analisar de forma exploratória o *MAPE* (plotado no eixo das ordenadas) em função das quantidades de observações (plotadas no eixo das abscissas), exclusivamente naquela RNA MLP com maior quantidade de observações (implementada a partir da amostra de *bootstrap* com n=1900), percebe-se que ele decresce à medida que a quantidade de observações n se eleva, apresentado certo grau de estabilização a partir de um n entre 400 e 500 observações ($400 < n < 500$), conforme descrito no gráfico detalhado na Figura 5.

Figura 5 – *MAPE* em função de cada *n* observação inserida na RNA MLP *bootstrap* com *n* = 1900



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

O comportamento descrito no gráfico apresentado na Figura 5 demonstra um comportamento divergente daquele observado nas pesquisas realizadas por Effron e Tibshirani (1993), segundo o qual a quantidade ideal de reamostras não precisaria ser maior que 200 replicações amostrais aleatórias ($n \leq 200$). Isso também acontece em relação à pesquisa realizada por Carmo e Lima (2018), ao ponderarem que amostras entre 44 e 50 observações ($44 \leq n \leq 55$). Contudo, deve-se lembrar que a aplicação proposta nesta investigação é diferente daquelas realizadas por Effron e Tibshirani (1993) e Carmo e Lima (2018), ou seja, num contexto em que o método de *bootstrap* for aplicado ao *machine learning* baseado em RNA, provavelmente, a quantidade mínima de observações demandadas tende a ser maior, conforme a hipótese já levantada por Ludemir (2021).

Por fim, procedeu-se à análise de relevância das variáveis explicativas de acordo com as amostras utilizadas para implementação das respectivas RNA MLP, conforme pode ser observado na descrição contida na Figura 6. O que se percebeu foi que perfil das variáveis consideradas relevantes na implementação da RNA MLP pesquisada com base na amostra original (n=44) teve um comportamento distinto do perfil observado nas demais RNA.

Figura 6 – Relevância das variáveis explicativas de acordo com as amostras utilizadas pelas RNA

Original (n=44)	*Bootstrap* (n=190)	*Bootstrap* (n=475)	*Bootstrap* (n=950)	*Bootstrap* (n=1425)	*Bootstrap* (n=1900)
20,98%	14,18%	16,13%	14,45%	16,64%	16,5%
18,02%	13,94%	13,03%	13,07%	14,03%	13,48%
11,98%	11,04%	10,62%	12,21%	11,65%	11,95%
11,65%	10,83%	8,78%	11,45%	10,01%	10,81%
7,02%	9,42%	8,43%	8,98%	9,87%	8,83%
6,51%	8,42%	7,89%	8,09%	7,92%	8,40%
5,89%	7,44%	7,05%	7,1%	6,25%	6,41%
4,61%	6,35%	6,71%	5,09%	5,73%	4,90%
4,45%	6,09%	6,45%	4,69%	4,44%	4,88%
4,04%	3,95%	4,39%	4,69%	4,36%	4,28%
3,55%	3,61%	3,97%	3,78%	3,48%	3,62%
1,11%	2,77%	3,50%	3,67%	3,26%	3,61%
0,18%	1,97%	3,07%	2,74%	2,34%	2,33%

Legenda:

Instituição do SPED (2007)	Obrig. da EFD (2014)
Util. de ERP p/ SPED (2008)	Institit. do eSocial (2014)
Obrig. da Nfe (2008)	Surg. Contab. Digital (2015)
SPED p/ lucro real (2010)	Asses. e serv. Contab. (2015)
Obrig. da ECD (2011)	Implem, Cont. Digital (2017)
Startup cont. (2011)	Obrigat, do eSocial (2018)
Contabilidade *on-line* (2012)	

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

Ainda segundo a análise de relevância das variáveis explicativas, descrita na Figura 6, apesar da RNA pesquisada com base na amostra de *bootstrap* com *n*=190 apresentar um comportamento muito mais alinhado com as demais amostras de *bootstrap*, comparativamente à "amostra original", o perfil da relevância atribuída às variáveis explicativas mostrou-se mais alinhado entre as RNA MLP implementas a partir de uma quantidade observações (*n*) maior ou igual a 475 ($n \geq 475$). Essa evidência corrobora com o comportamento observado até aqui em relação à avaliação global de desempenho das RNA e em relação à análise para a detecção de *overfitting*, isto é, segundo a "amostra original" utilizada para aplicação do método de *bootstrap* e implementação das respectivas RNA, com $n \geq 475$ as RNA MLP implementadas apresentaram melhores desempenhos.

Essa última evidência demonstra que o método de *bootstrap* também pode ser utilizado para a seleção e, se for o caso, a redução da quantidade de parâmetros de

entrada utilizados no processo de estimação de um modelo de *machine learning* baseado em RNA, conforme sugerido por Zhu *et al.* (2023).

5 Considerações Finais

Ao avaliar como o método bootstrap poderia auxiliar na melhora do desempenho do *machine learning* baseado em redes RNA, a presente pesquisa permitiu inferir que, segundo a amostra dos dados utilizados para composição do respectivo estudo de caso, as RNA MLP implementadas com base nas amostras de *bootstrap* apresentaram um desempenho superior ao desempenho observado para a RNA MLP implementada a partir da amostra original.

Adicionalmente, segundo a amostra dos dados utilizados para composição do respectivo estudo de caso, as evidências coletadas a partir deste trabalho levaram a constatar que a utilização de amostras geradas com o uso do método de *bootstrap* permitem realizar inferências voltadas para a detecção do *overfitting* em RNA MLP.

Por fim, ainda segundo a amostra dos dados utilizados para composição do estudo de caso realizado nesta investigação científica, foi possível observar que o método de *bootstrap* também pode ser utilizado para a seleção e, se for o caso, a redução da quantidade de parâmetros de entrada utilizados no processo de estimação de um modelo de *machine learning* baseado em RNA MLP.

Cabe destacar que, a despeito do rigor metodológico adotado na avaliação dos resultados observados, as evidências coletadas a partir desta pesquisa não podem ser consideradas conclusivas, uma vez que foram levantadas mediante a realização de um estudo de caso único e específico. Assim, sugere-se a continuidade desta investigação mediante a aplicação da metodologia analítica aqui proposta, porém, aplicada a amostras de dados compostas a partir de outros estudos de casos, com dados de diferentes naturezas.

REFERÊNCIAS

CARMO, C. R. S.; LIMA, A. D. de. Métodos quantitativos e pesquisa contábil: um estudo de caso relacionado a pequenas amostras de dados. **CONTABILOMETRIA - Brazilian Journal of Quantitative Methods Applied to Accounting**, [*s. l.*], v. 5, n. 1, p. 92-109, jan.-jun./2018. Disponível em: https://revistas.fucamp.edu.br/index.php/contabilometria/article/view/1025. Acesso em 03 out. 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 set. 2023.

CARNEIRO JÚNIOR, J. B. A.; SOUZA, C. S. de. Aplicação de redes neurais artificiais na previsão do produto interno bruto do Mato Grosso do Sul em função da produção de cana-de-açúcar, açúcar e etanol. **Revista Ibero-Americana de Ciências Ambientais**, [*s. l.*], v. 10, n. 5, p. 218-230, ago.-set. 2019. Disponível em: https://doi.org/10.6008/CBPC2179-6858.2019.005.0019. Acesso em: 08 ago. 2023.

COLLINS, S. D.; PEEK, N.; RILEY, R. D.; MARTIN, G. P.. Sample sizes of prediction model studies in prostate cancer were rarely justified and often insufficien. **Journal of Clinical Epidemiology**, [*s. l.*], v. 133, p. 53-60, May 2021. Disponível em: https://www.sciencedirect.com/science/article/pii/S0895435620312154. Acesso em: 25 abr. 2024.

CRC-MG, Conselho Regional de Contabilidade de Minas Gerais. **Serviços *Online*/Acesso Público/Consulta Cadastral**. Belo Horizonte-MG, consulta *on-line* realizada em 25 out. 2023 – 16:04h. Parâmetros da consulta:" Informe o tipo de pesquisa" = profissional; "Cidade" = Uberlândia.Disponível em: https://cadastro.crcmg.org.br/SPW/ConsultaCadastral/TelaConsultaPublicaCompleta.aspx?_gl=1*1iz4kla*_ga*MTE4ODU3OTUxMS4xNjk2OTUyODc2*_ga_1JTLYCSM7M*MTY5ODI2MDAzOC4yLjAuMTY5ODI2MDAzOC4wLjAuMA..*_ga_CW7MZRNEF3*MTY5ODI2MDAzOC4yLjAuMTY5ODI2MDAzOC4wLjAuMA..&_ga=2.104688512.1600827541.1698260038-1188579511.1696952876.

DICICCIO, T. J.; EFRON, B.. Bootstrap confidence intervals. **Statist. Sci.** , [*s. l.*], v. 11, n. 3, p. 189 - 228, August 1996. Disponível em: https://doi.org/10.1214/ss/1032280214. Acesso em: 14 out. 2023.

EFRON, B.; HALLORAN, E.; HOLMES, S.. Bootstrap confidence levels for phylogenetic trees. **Proc.Natl.Acad.Sci. USA**, [*s. l.*], v. 93, n. 14, p. 7085-7090, July

9, 1996. Disponível em: https://doi.org/10.1073/pnas.93.14.7085. Acesso em: 07 out. 2023.

EFRON, B.; TIBSHIRANI, R. J.. **An introduction to the bootstrap**. New York: Chapman & Hall, 1993.

EFRON, B.; TIBSHIRANI, R.. empirical bayes methods and false discovery rates for microarrays**. Genetic Epidemiology**, [*s. l.*], v. 23, issue 1, p. 70–86, June 2002. https://doi.org/10.1002/gepi.1124. Acesso em: 11 out. 2023.

GU, Y.; WEI, H-L.. A robust model structure selection method for small sample size and multiple datasets problems. **Information Sciences**, [*s. l.*], v, 451–452, p. 195-209, July 2018. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0020025518302627?via%3Dihub. Acesso em: 24 abr. 2024.

HOMMA, T.; SALTELLI, A.. Importance measures in global sensitivity analysis of nonlinear models. **Reliability Engineering & System Safety**, [s. l.], v. 52, issue 1, p. 1-17, 1996. Disponível em: https://doi.org/10.1016/0951-8320(96)00002-6. Acesso em: 01 nov. 2023.

HU, L-t.; BENTLER, P. M..Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria *versus* new alternatives. **Structural Equation Modeling: A Multidisciplinary Journal**, [*s. l.*], v. 6, issue 1, p. 1-55, 1999. Disponível em: https://doi.org/10.1080/10705519909540118. Acesso em: 05 out. 2023.

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: 19 out. 2023.

JECZMIONEK, E; KOWALSKI, P. A.. Input reduction of convolutional neural networks with global sensitivity analysis as a data-centric approach. **Neurocomputing**, [*s. l.*], v. 506, p. 196-205, 2022. Disponível em: https://doi.org/10.1016/j.neucom.2022.07.027. Acesso em: 02 nov. 2023.

JIA, Y.; CULVER, T B.. Bootstrapped artificial neural networks for synthetic flow generation with a small data sample**. Journal of Hydrology**, [*s. l.*], v. 331, issues 3–4, p. 580-590, 15 December 2006. Disponível em: https://doi.org/10.1016/j.jhydrol.2006.06.005. Acesso em: 25 abr. 2024.

KARAMIRAD, M.; OMID; M.; ALIMARDANI, R.; MOUSAZADEH, H.; HEIDARI, S. N.. ANN based simulation and experimental verification of analytical four- and five-parameters models of PV modules. **Simulation Modelling Practice and Theory**, [*s.*

l.], v. 34, p. 86-98, May 2013. Disponível em: https://doi.org/10.1016/j.simpat.2013.02.001. Acesso em: 17 out. 2023.

KÜNSCH, H. R..The jackknife and the bootstrap for general stationary observations. **The Annals of Statistics**, [*s. l.*], v. 17, n. 3, p. 1217–1241, September, 1989. Disponível em: https://doi.org/10.1214/aos/1176347265. Acesso em: 16 out. 2023.

LEE, T-H., ULLAH, A., WANG, R.. Bootstrap aggregating and random forest. *In*: FULEKY, P. (eds) **macroeconomic forecasting in the era of big data**. Advanced Studies in Theoretical and Applied Econometrics, v 52. [*S. l.*]: Springer Nature Switzerland, 2020. Disponível em: https://doi.org/10.1007/978-3-030-31150-6_13. Acesso em: 20 out. 2023.

LI, D.-C.; LIN, W.-K.; CHEN, C.-C.; CHEN, H.-Y.; LIN, L.-S.. Rebuilding sample distributions for small dataset learning. **Decision Support Systems**, [*s.l.*], v. 105, p 66-76, January 2018. Disponível em: https://www.sciencedirect.com/science/article/pii/S0167923617302014. Acesso em: 24 abr. 2024.

LIN, L.-S.; LIN, Y.-S.; LI, D.-C.; LIU, Y.-H.. Improved learning performance for small datasets in high dimensions by new dual-net model for non-linear interpolation virtual sample generation. **Decision Support Systems**, [*s. l.*], v 172, e-article 113996, September 2023. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0167923623000714?via%3Dihub. Acesso em: 24 abr. 2024.

LUDERMIR, T. B.. Inteligência artificial e aprendizado de máquina:estado atual e tendências. **Estudos Avançados**, [*s. l.*], v. 35, n. 101, p. 85-94, 2021. Disponível em: https://doi.org/10.1590/s0103-4014.2021.35101.007. Acesso em: 16 out. 2023.

MA, Y.; LENG, C.; WANG, H.. Optimal subsampling bootstrap for massive data. **Journal of Business & Economic Statistics**, [*s. l.*], v. 42, n. 1, p. 174–186, 2024. Disponível em: https://doi.org/10.1080/07350015.2023.2166514. Acesso em: 25 abr. 2024.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. 3rd ed. New York: Wiley, 2001.

MURAKOSHI, K.. Avoiding overfitting in multilayer perceptrons with feeling-of-knowing using self-organizing maps. **Biosystems**, [*s. l.*], v. 80, Issue 1, p. 37-40, 2005. Disponível em: https://doi.org/10.1016/j.biosystems.2004.09.031. Acesso em: 25 abr. 2024.

PEREIRA, G. H. de A.; CENTENO, J. A. S.. Avaliação do tamanho de amostras de treinamento para redes neurais artificiais na classificação supervisionada de imagens

utilizando dados espectrais e *laser scanner*. **Boletim de Ciências Geodésicas**, [*s. l.*], v. 23, n. 2, p. 268-283, abr.-jun. 2017 Disponível em: https://doi.org/10.1590/S1982-21702017000200017. Acesso em: 03 out. 2023.

SHEN, L.; QIAN, Q.. A virtual sample generation algorithm supporting machine learning with a small-sample dataset: A case study for rubber materials. **Computational Materials Science**, [*s. l.*], v. 211, e-article 111475, August 2022. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0927025622002385?via%3Dihub. Acesso em: 24 abr. 2024.

TIWARI, M. K.; CHATTERJEE, C.. Development of an accurate and reliable hourly flood forecasting model using wavelet–bootstrap–ANN (WBANN) hybrid approach. **Journal of Hydrology**, [*s. l.*], v. 394, issues 3–4, p. 458-470, 26 November 2010. Disponível em: https://doi.org/10.1016/j.jhydrol.2010.10.001. Acesso em: 08 out. 2023.

WANG, M.; HUI, G.; PANG, Y.; WANG, S.; CHEN, S.. Optimization of machine learning approaches for shale gas production forecast. **Geoenergy Science and Engineering**, [*s. l.*], v. 226, e-article 211719, 2023. Disponível em: https://doi.org/10.1016/j.geoen.2023.211719. Acesso em: 25 abr. 2024.

YADAV, A. K.; MALIK, H.; CHANDEL, S. S.. Selection of most relevant input parameters using WEKA for artificial neural network based solar radiation prediction models. **Renewable and Sustainable Energy Reviews**, [*s. l.*], v. 31, p. 509-519, March 2014.
Disponível em: https://doi.org/10.1016/j.rser.2013.12.008. Acesso em: 18 out. 2023.

ZHU, Q-X.; ZHANG, H.-T.; TIAN, Y.. ZHANG, N.; XU, Y.; HE, Y.-L.. Co-training based virtual sample generation for solving the small sample size problem in process industry. **ISA Transactions**, [*s. l.*], v. 134, p. 290-301, 2023. Disponível em: https://www.sciencedirect.com/science/article/pii/S0019057822004244. Acesso em: 11 out. 2023.