ADM CONGRESSO INTERNACIONAL DE ADMINISTRAÇÃO
25 a 28 setembro 2024
Campus Central UEPG Ponta Grossa | PR
Explorando as Interseções das Inteligências Artificiais na Sociedade Atual
Realização: UEPG Universidade Estadual de Ponta Grossa; proex Pró-Reitoria de Extensão e Assuntos Culturais; Departamento de Administração
Apoio: Ponta Grossa Prefeitura Turismo; COMTURPG

ALGORITMOS DE MACHINE LEARNING PARA PREVER O PRODUTO INTERNO BRUTO DAS CIDADES BRASILEIRAS

MACHINE LEARNING ALGORITHMS TO PREDICT THE GROSS DOMESTIC PRODUCT OF BRAZILIAN CITIES

ÁREA TEMÁTICA: ADMINISTRAÇÃO DA INFORMAÇÃO

Aline Pacheco Primão, Universidade Federal de Santa Catarina, Brasil, aline.pacheco.pr@gmail.com

Alexandre Marino Costa, Universidade Federal de Santa Catarina, Brasil, marinocad@gmail.com

Resumo

O Produto Interno Bruto (PIB) é um indicador usado para medir a atividade econômica de um país. Esta pesquisa avaliou o PIB das cidades brasileiras de 2010 a 2021, utilizando dados do IBGE e técnicas de Machine Learning. Foram aplicados o Método dos Mínimos Quadrados de Regressão Linear e a Rede Neural Artificial MultiLayer Perceptrons (MLP) para prever o PIB das cidades. Os resultados mostraram que o Método dos Mínimos Quadrados teve o melhor desempenho, com Erro Médio Quadrático (MSE) de 0,2564 no treino e 0,2554 no teste, e um Coeficiente de Determinação (R^2) de 1 para ambos os conjuntos, indicando alta precisão e capacidade de explicar 100% da variabilidade no PIB. O MLP, apesar de também ter obtido R^2 de 1, teve um MSE de 6419,44 no teste, sugerindo menor precisão. Embora mais simples, o Método dos Mínimos Quadrados foi mais eficaz na previsão do PIB das cidades brasileiras comparado ao MLP, sendo a abordagem mais adequada para este problema.

Palavras-chave: *PIB das Cidades Brasileiras; Algoritmos de Machine Learning; MultiLayer Perceptrons; Método dos Mínimos Quadrados.*

Abstract

Gross Domestic Product (GDP) is an indicator used to measure the economic activity of a country. This research evaluated the GDP of Brazilian cities from 2010 to 2021 using data from IBGE and Machine Learning techniques. The study applied the Ordinary Least Squares Linear Regression method and the MultiLayer Perceptrons (MLP) Artificial Neural Network to predict the GDP of the cities. The results showed that the Ordinary Least Squares method performed the best, with a Mean Squared Error (MSE) of 0.2564 for training and 0.2554 for testing, and a Coefficient of Determination (R^2) of 1 for both sets, indicating high precision and the ability to explain 100% of the variability in GDP. Although the MLP also achieved an R^2 of 1, it had an MSE of 6419.44 for the test set, suggesting lower accuracy. Despite being simpler, the Ordinary Least Squares method was more effective in predicting the GDP of Brazilian cities compared to the MLP, making it the most suitable approach for this problem.

Keywords*: GDP of Brazilian Cities; Machine Learning Algorithms; MultiLayer Perceptrons; Ordinary Least Squares Method.*

1. INTRODUÇÃO

O Produto Interno Bruto (PIB) é uma medida importante para avaliar a atividade econômica de um país. É um indicador para monitorar o crescimento econômico ao longo do tempo e pode contribuir na geração de emprego e renda (IBGE, 2024). No contexto global, o PIB também é reconhecido como um indicador significativo para os Objetivos de Desenvolvimento Sustentável (ODS) da ONU, especialmente para monitorar o progresso econômico e social dos países membros (ONU, 2021).

As técnicas de Machine Learning (ML) empregam modelos estatísticos para prever o risco de eventos ocorrerem ou, no caso de modelos de regressão como é o caso desta pesquisa, para estimar variáveis dependentes (PIB). Dentre os diversos algoritmos disponíveis, o Método dos Mínimos Quadrados para Regressão Linear é fundamental em estatística e aprendizado de máquina, buscando ajustar uma linha ou plano aos dados minimizando a soma dos quadrados das diferenças entre valores previstos e reais (Géron, 2019). Já as MultiLayer Perceptrons (MLP) são redes neurais artificiais com múltiplos neurônios organizados em camadas, sendo uma camada de entrada, uma ou mais camadas ocultas para processamento intermediário, e uma camada de saída (Kovács, 2006).

Ao combinar esses dois algoritmos, é possível obter uma análise mais abrangente do PIB das cidades brasileiras, explorando como este indicador pode contribuir para os Objetivos de Desenvolvimento Sustentável (ODS).

Neste sentido, esta pesquisa visa avaliar o PIB das cidades brasileiras utilizando técnicas de Machine Learning. Para isso, analisar dados do IBGE referentes aos anos de 2010 a 2021 de todas as cidades do país. O estudo empregou o Método dos Mínimos Quadrados de Regressão Linear e a Rede Neural Artificial MultiLayer Perceptrons (MLP) para criar os modelos para previsão do PIB das cidades brasileiras.

Justifica-se o uso destes algoritmos, pois eles oferecem abordagens complementares na modelagem do PIB das cidades brasileiras: o Método dos Mínimos Quadrados facilita a compreensão das relações lineares entre variáveis econômicas e o PIB, enquanto a rede neural MLP é capaz de capturar relações não lineares complexas, aumentando a flexibilidade e precisão das previsões, especialmente em dados econômicos complexos.

Este artigo está estruturado da seguinte forma: começando com a introdução, seguido pelo referencial teórico que aborda o PIB das cidades brasileiras e os algoritmos de Machine Learning. Em seguida, na metodologia incluem detalhes sobre a base de dados, o pré-processamento, a criação dos modelos de previsão do PIB e as métricas utilizadas para avaliá-los. Na quarta parte, são apresentados e discutidos os resultados, e por fim, as conclusões da pesquisa são evidenciadas.

2. FUNDAMENTAÇÃO TEÓRICA

Nesta seção será apresenta a fundamentação teórica utilizada para a realização da pesquisa.

2.1. Produto Interno Bruto das cidades brasileiras

Segundo o Instituto Brasileiro de Geografia e Estatísticas (IBGE), o Produto Interno Bruto representa a totalidade dos bens e serviços finais produzidos no território nacional durante um período específico, geralmente de um ano. Ele não é representativo da totalidade da riqueza do país, mas sim um indicador do fluxo de novos bens e serviços finais produzidos, sendo fundamental para a avaliação da atividade econômica e sua evolução ao longo do tempo. O preço final ao consumidor mede os bens e serviços finais que fazem parte do PIB, incluindo os impostos sobre os produtos comercializados (IBGE, 2024).

O PIB oferece várias análises úteis como comparar o PIB de cada ano ao longo do tempo, analisar o tamanho das economias de vários países para efeitos comparativos e examinar o PIB

por habitante, que representa a parte do PIB que cada pessoa receberia se fosse distribuída de forma igual entre todos os habitantes de um país, e outras pesquisas. O PIB é uma síntese da economia, dando detalhes sobre um país, mas não considera fatores como a distribuição de renda, qualidade de vida, educação e saúde (IBGE, 2024).

No caso do Brasil, o próprio IBGE coleta uma variedade de dados para calcular o PIB, como o Índice Nacional de Preços ao Consumidor Amplo (IPCA), Produção Agrícola Municipal (PAM), Pesquisa Mensal de Serviços (PMS), entre outros.

O cálculo do PIB pode variar conforme a abordagem utilizada: oferta ou renda. Ambas levam em consideração fatores como consumo das famílias, investimentos, gastos do governo, exportações e importações. Na ótica da oferta, o PIB é calculado somando o Valor Adicionado Bruto (VAB) aos impostos indiretos e subtraindo os subsídios. Já na ótica da renda, o PIB é calculado somando os salários, o Excedente Operacional Bruto (EOB), os impostos indiretos e subtraindo os subsídios (Silva et al., 2020a).

Independente da abordagem definida, o resultado do PIB será o mesmo. Na abordagem pela oferta, somamos a produção dos três setores econômicos: agropecuária, indústria e serviços (PIB = Agropecuária + Indústria + Serviços), onde serviços é o mais representativo na economia brasileira.

O Valor Adicionado Bruto (VAB) é considerado o valor que cada setor da economia acrescenta ao valor final de tudo que foi produzido em uma região. Neste sentido, se torna uma medida importante para avaliar a contribuição de diversos setores econômicos para o PIB de um país ou região, possibilitando a compreensão da distribuição da produção entre diferentes setores. Compreender a estrutura econômica de uma nação, identificar áreas de crescimento e avaliar a eficácia das políticas econômicas é essencial. No caso dos dados que serão apresentados a seguir o VAB é dado pelo PIB - impostos (Silva et al., 2020a; IBGE, 2024).

O uso de algoritmos de aprendizado de máquina (Machine Learning) pode contribuir significativamente para a avaliação e previsão do PIB. Essas técnicas permitem analisar grandes volumes de dados de maneira eficiente e descobrir padrões complexos que podem melhorar a precisão das previsões econômicas.

2.2. Algoritmos de Machine Learning (ML)

O aprendizado de máquina, Machine Learning (ML), se destacam no processamento de grandes volumes de dados que se tornaram impossíveis para humanos lidarem diretamente. Domingos (2017) ilustra essa questão ao afirmar que “enquanto um cientista talvez passe sua vida inteira criando e testando algumas centenas de hipóteses, um sistema de Machine Learning pode fazer o mesmo em uma fração de segundo” (Domingos, 2017).

O campo de Machine Learning envolve a exploração de muitas hipóteses para encontrar a que melhor se ajusta aos exemplos de treinamento disponíveis e a outras restrições. Para isso, ML consiste na capacidade de desenvolver algoritmos para melhorar seu desempenho em tarefas específicas com base na experiência (Mitchell, 1997).

Para Mitchell (1997), estes algoritmos têm sido amplamente úteis em áreas como a mineração de dados, onde padrões implícitos em grandes conjuntos de dados podem ser descobertos automaticamente; domínios mal compreendidos, onde os humanos podem não possuir o conhecimento necessário para criar algoritmos eficazes; e campos onde o algoritmo deve se adaptar dinamicamente às mudanças nas condições (Mitchell, 1997).

Em modelos supervisionados existem dois problemas básicos: classificação e regressão. Na classificação são definidos rótulos entre os existentes para identificar categorias específicas. Já na regressão, a ideia é prever um alvo de valor numérico a partir de um conjunto de características os quais são chamados de previsores (Géron, 2019).

Atualmente, existe uma diversidade de algoritmos, dentre eles, o Método dos Mínimos Quadrados para regressão linear é uma técnica fundamental em estatística e aprendizado de máquina para encontrar a linha, ou plano em casos de múltiplas variáveis, que melhor se ajusta aos dados observados. Ele busca minimizar a soma dos quadrados das diferenças entre os valores previstos pelo modelo e os valores reais observados. Esse método é amplamente utilizado em problemas de regressão devido à sua simplicidade e eficácia na determinação dos melhores parâmetros para ajustar um modelo linear aos dados observados (Géron, 2019).

As Redes Neurais Artificiais (RNAs), uma forma de Inteligência Artificial conexionista, são usadas para resolver problemas de classificação e regressão, destacando-se em desafios grandes e complexos (Géron, 2019). Inspiradas no cérebro humano, as RNAs foram desenvolvidas por Warren McCulloch e Walter Pitts, que criaram o modelo do neurônio artificial (Géron, 2019).

Dentre as RNAs, as MultiLayer Perceptrons (MLP) são compostas por múltiplos neurônios do tipo discriminadores lineares, organizados em várias camadas. Essas camadas incluem uma "camada de entrada", que recebe as entradas; uma ou mais "camadas ocultas", que realizam o processamento intermediário, recebendo as saídas da camada anterior como entrada; e, por fim, uma "camada de saída", que gera a saída final (Mitchell, 1997; Kovács, 2006).

3. METODOLOGIA

Em relação à filosofia, esta pesquisa pode ser considerada positivista, pois a lógica e a matemática são válidas ao estabelecer as regras da linguagem, sendo consideradas um conhecimento a priori e independente da experiência (Terence; Escrivão Filho, 2006). Quanto ao pensamento, esta pesquisa é dedutiva, seguindo uma abordagem racional na tentativa de confirmar uma hipótese. Em termos de propósito, a pesquisa é explicativa, conforme definido por Gil (2008), buscando identificar fatores que determinam ou contribuem para a ocorrência de um fenômeno, com o intuito de aprofundar o conhecimento da realidade ao explicar as razões por trás dos eventos observados. Quanto à estratégia, adotou-se o método quantitativo, alinhando-se ao paradigma clássico positivista e utilizando métodos estatísticos representados pelos algoritmos de Machine Learning. Por fim, a coleta e análise de dados utilizaram fontes secundárias, com registros extraídos diretamente do banco de dados do IBGE.

A seguir será apresentada a base de dados do IBGE, que traz o PIB das cidades brasileiras utilizada para realizar o trabalho, juntamente com os passos de como foi realizado o pré-processamento dos dados e os modelos criados sem e com RNAs. Foram utilizados modelos para regressão, pois o objetivo é resolver um problema de previsão de valores contínuos (PIB). O modelo sem redes neurais utilizado foi o Método dos Mínimos Quadrados para Regressão Linear e o algoritmo com redes neurais foi MultiLayer Perceptron.

Cabe ressaltar que os algoritmos foram desenvolvidos em Python no Notebook Colaboratory (Colab) do Google.

3.1 Base de dados

A base de dados utilizada neste trabalho é o PIB das cidades brasileiras, disponibilizada pelo IBGE e abrange os anos de 2010 a 2021. Os dados foram extraídos do site do IBGE, especificamente do arquivo "base_de_dados_2010_2021_xlsx.zip" disponível em https://www.ibge.gov.br/estatisticas/downloads-estatisticas.html -> Pib_municipios -> 2021 -> base.

A base de dados extraída continha 66.825 registros e 43 variáveis. Para viabilizar o processamento, o trabalho foi dividido em dois algoritmos: um dedicado ao pré-processamento dos dados e outro à criação dos modelos.

3.2 Pré-processamento dos dados

Inicialmente, a base de dados apresentava algumas colunas com nomes quebrados, o que exigiu correção para permitir a leitura adequada. Foi desenvolvida uma função para resolver essas quebras de linha e aplicada a todas as células do DataFrame. Em seguida, foram analisados os dados em relação a valores faltantes e duplicados. Não foram identificados dados duplicados, porém foram encontrados 337.119 valores ausentes.

Em seguida, foram excluídas 19 colunas que não eram necessárias, contendo códigos como o da unidade da federação e município, entre outros dados que não eram relevantes para a análise, como informações sobre região geográfica imediata e nome de concentração urbana.

Após a exclusão dessas colunas, apenas a variável “Região Metropolitana” permaneceu com dados ausentes. Foi observado que esses dados ausentes correspondiam a áreas que não pertenciam a nenhuma região metropolitana. Para corrigir isso, foi feita uma imputação substituindo os valores NaN pela palavra "Não". Após essa etapa, o DataFrame não continha mais valores ausentes e pôde-se prosseguir com a análise.

Após as correções, o DataFrame resultante continha 66.825 registros e 21 variáveis. As variáveis foram renomeadas devido à extensão e à presença de acentos e caracteres especiais que poderiam causar problemas no processamento. As variáveis que eram do tipo object foram convertidas para o tipo numérico, utilizando o LabelEncoder para a maioria delas e transformação manual para as variáveis que classificavam as atividades de maior valor.

A variável "município" foi ajustada para "municipio_uf" para evitar ambiguidades, já que existem cidades no Brasil com o mesmo nome em diferentes estados.

Durante o processo de limpeza, ao remover caracteres não numéricos da variável “VAB_agropecuaria”, verificou-se que 24 registros continham apenas um '-' e foram transformados em NaN (nulos), exigindo imputação adicional. Foi utilizado o IterativeImputer para essa tarefa.

Por fim, as colunas do DataFrame foram reordenadas para facilitar manipulações futuras. O DataFrame final foi exportado como 'pib.csv' e a exportação foi verificada para garantir a integridade dos dados. A Figura 1 apresenta as 21 variáveis restantes no DataFrame, sendo que a última (PIB) é a variável de saída e as demais as variáveis de entrada.

#	Column	Non-Null Count	Dtype
0	ano	66825 non-null	int64
1	municipio_uf	66825 non-null	int64
2	uf	66825 non-null	int64
3	nome_grande_regiao	66825 non-null	int64
4	regiao_metropolitana	66825 non-null	int64
5	nome_mesoregiao	66825 non-null	int64
6	nome_microregiao	66825 non-null	int64
7	cidade_regiao_sp	66825 non-null	int64
8	semiarido	66825 non-null	int64
9	amazonia_legal	66825 non-null	int64
10	VAB_agropecuaria	66825 non-null	float64
11	VAB_industria	66825 non-null	int64
12	VAB_servicos_sem_adm_publica	66825 non-null	int64
13	VAB_servicos_adm_publica	66825 non-null	int64
14	VAB_total	66825 non-null	int64
15	impostos_subsidios	66825 non-null	int64
16	atividade_primeiro_maior_VAB	66825 non-null	int64
17	atividade_segundo_maior_VAB	66825 non-null	int64
18	atividade_terceiro_maior_VAB	66825 non-null	int64
19	PIB_per_capita	66825 non-null	int64
20	PIB	66825 non-null	int64

Figura 1 – Variável do dataFrame de estudo

O algoritmo desenvolvido no Google Colab para o pré-processamento está acessível para leitura através do link: Trabalho IA - Pre-processamento.ipynb. Para acessá-lo, é necessário baixá-lo no Google Workspace Marketplace.

3.3 Criação dos modelos para prever o PIB

Foi definido criar modelos de regressão, pois para resolver problemas de previsão de valores contínuos, especificamente o PIB, são os ideais. Para isso, foi importada a base de dados processada pelo algoritmo de pré-processamento, que gerou o arquivo 'pib.csv'.

O conjunto de dados foi dividido em duas partes: X, que contém todas as variáveis exceto o PIB, e y, que consiste apenas na variável PIB. Em seguida, os dados foram normalizados usando o StandardScaler e divididos em conjuntos de treinamento e teste, sendo reservados 25% dos dados para teste e utilizando o restante para treinamento. Posteriormente, os dados foram plotados para visualização.

A Figura 2 fornece o gráfico da divisão dos dados de treinamento (em vermelho) e teste (em azul), onde o eixo x apresenta uma das variáveis de entrada e o eixo y a variável de saída (PIB). As variáveis aparecem normalizadas e os dados ficaram com 66.825 linhas, que representam as informações das cidades dentro dos anos estudados (2010 a 2021) e 20 colunas, que representam as variáveis de entrada. O PIB é a 21ª coluna que foi separada, pois representa os dados de saída.

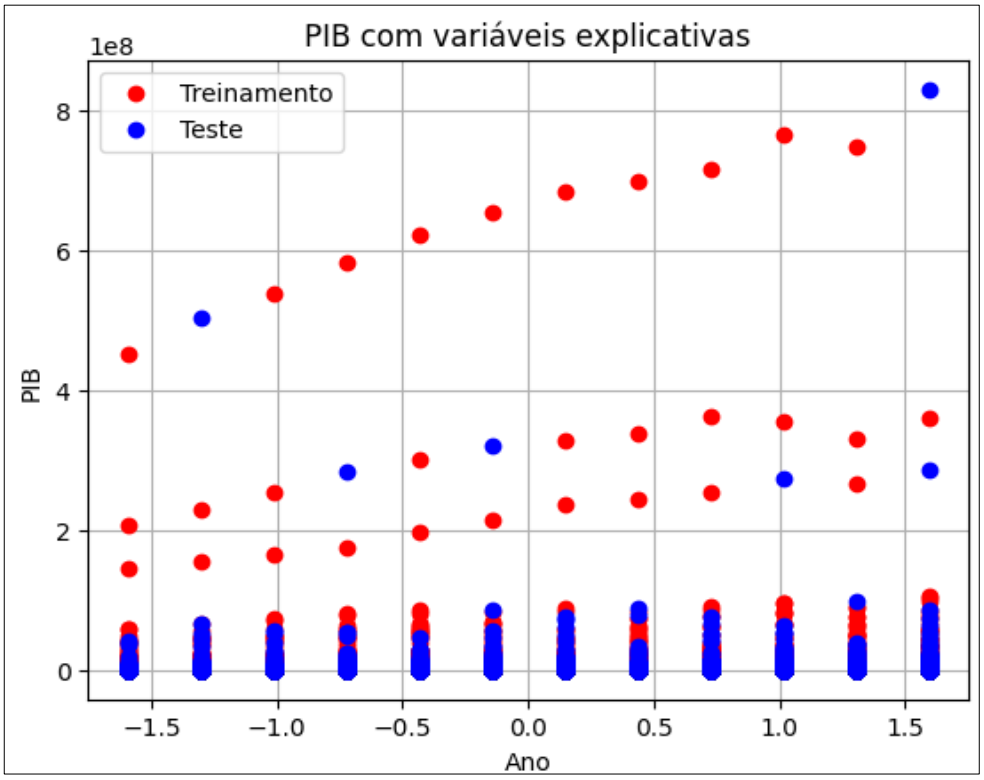

Figura 2 – Divisão das variáveis de treinamento e teste

O algoritmo desenvolvido no Google Colab para o pré-processamento está acessível para leitura através do link: Trabalho IA Modelos.ipynb.

3.4.Métricas para avaliação dos modelos

Para avaliar a eficácia dos modelos de previsão criados, foram selecionadas as seguintes métricas de avaliação: Coeficiente de Determinação (R^2) e o Erro Médio Quadrático (MSE).

O Coeficiente de Determinação (R^2), r2_score, é uma medida na regressão que varia de 0 a 1 e representa a porcentagem da variância do resultado (y) que é explicado pelas variáveis no modelo. Ele indica a qualidade do ajuste do modelo e quão bem ele pode prever novos dados não vistos, através da explicação da variância. A melhor pontuação é 1, e ao contrário da maioria das métricas, esta pode ser negativa, o que representa um modelo arbitrariamente pior (Sklearn, 2024).

O Erro Médio Quadrático (MSE), mean_squared_error, é utilizado para avaliar a precisão do modelo de regressão e calcula a média dos erros ao penalizar de forma mais significativa os erros maiores. Se estiver buscando punir erros significativos, o MSE é uma ferramenta útil. O MSE não fornece uma medida de quão ruim é um modelo, mas possui a capacidade de comparar modelos. Isso é especialmente útil quando os erros seguem uma distribuição normal (Harrison, 2020).

4. APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

Esta seção apresenta os resultados dos modelos desenvolvidos com ambos os algoritmos a partir das métricas utilizadas, os ajustes realizados para melhorá-los.

4.1.Método de Mínimos Quadrados para regressão linear

Primeiramente, para modelar os dados utilizando o Método dos Mínimos Quadrados para Regressão Linear, adicionou-se uma coluna para incluir o termo de viés (bias) para ajustar o modelo corretamente. Em seguida, foi calculado o Theta usando a equação normal, expressa

como (Theta = (X'X)^-1 X'y), onde X representa a matriz de entrada e y a variável dependente (saída).

Após obter os coeficientes do modelo (Theta), foram utilizadas as métricas de desempenho. Calculado o Erro Médio Quadrático (MSE) para os conjuntos de dados de treinamento e teste, obtendo os seguintes resultados: Erro Quadrático Médio (MSE) para os dados de treinamento = 0.2564 e o Erro Quadrático Médio (MSE) para os dados de teste = 0.2554. Os resultados indicam que, em média, os quadrados dos erros entre os valores previstos e os valores reais para treinamento e teste, arredondando, são de 0,26, mostrando bons resultados, onde quanto mais próximo de zero (0) melhor a precisão do modelo de regressão.

Posteriormente, foi calculado o coeficiente de determinação (R^2) para verificar o quão bem o modelo se ajusta aos dados. Assim, o R^2 para os dados de treinamento obteve resultado igual a um (1), e o R^2 para os dados de teste foi igual a um (1). Esses valores indicam que o modelo explica perfeitamente a variabilidade dos dados tanto no conjunto de treinamento quanto no de teste.

Para concluir esta etapa, foram realizadas previsões utilizando os dados de treinamento. Os resultados da regressão foram então apresentados graficamente, proporcionando uma melhor visualização do ajuste do modelo aos dados e demonstrando a precisão das previsões. A Figura 3 fornece os dados previstos e reais com o modelo utilizando o Método dos Mínimos Quadrados.

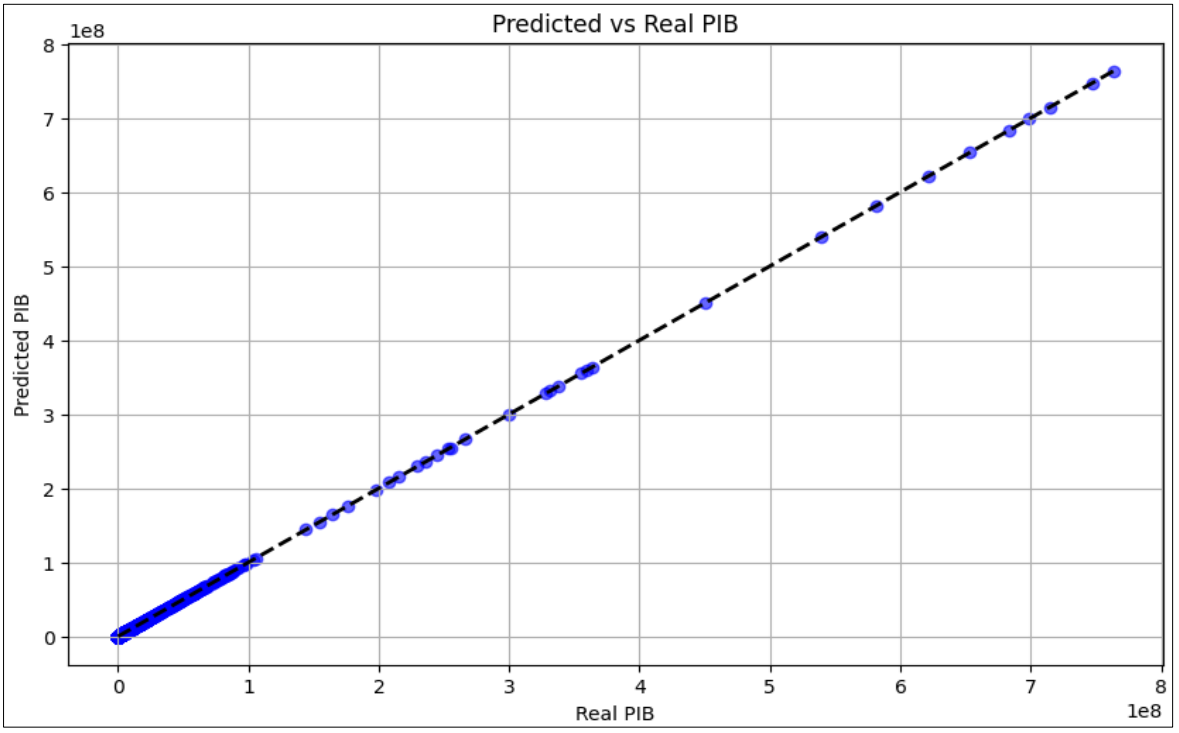

Figura 3 – PIB previsto X PIB real com método dos mínimos quadrados

4.2.MultiLayer Perceptron (MLP)

Inicialmente, foi utilizado o Callback com Early Stopping para evitar o sobreajuste durante o treinamento. A forma de entrada do modelo (input_shape) foi definida para corresponder ao número de colunas em X_train, garantindo que a rede neural soubesse quantas variáveis de entrada esperar. Em seguida, foi inicializado o modelo sequencial. A primeira camada foi configurada com 64 neurônios e a função de ativação ReLU. A função ReLU (Rectified Linear

Unit) ajuda a introduzir não-linearidade no modelo, permitindo que ele aprenda padrões mais complexos. A segunda camada foi configurada com 32 neurônios e utilizou a função de ativação ReLU, permitindo ao modelo capturar relações ainda mais complexas nos dados. A camada de saída foi configurada com um único neurônio, pois trata-se de um problema de regressão. Não foi utilizada nenhuma função de ativação na camada de saída, dessa forma os valores previstos fossem diretamente os valores de saída do neurônio.

Os parâmetros iniciais de treinamento foram definidos como: epochs=100, que é o número máximo de vezes que o algoritmo de treinamento percorre todo o conjunto de dados; learning rate (lr)=0.01, que controla o tamanho dos passos que o modelo dá na direção da minimização da função de perda; momentum=0.9, que é utilizado para acelerar o treinamento e ajudar a evitar que o modelo fique preso em mínimos locais, ajustando a velocidade das atualizações dos pesos; e patience=10, que é o número de épocas que o modelo espera sem melhoria no desempenho de validação antes de parar o treinamento, utilizado pelo Callback de Early Stopping para prevenir o overfitting.

Na compilação do modelo, foi utilizado o otimizador Adam, configurado com uma taxa de aprendizado (learning_rate) e o parâmetro de momentum = beta_1. A função de perda foi definida como o Erro Quadrático Médio. Em seguida, o treinamento foi realizado com um tamanho de lote (batch_size) de 32 e o Callback de Early Stopping previamente configurado. As curvas de aprendizado (Figura 4) foram então plotadas para facilitar a análise visual do processo de treinamento.

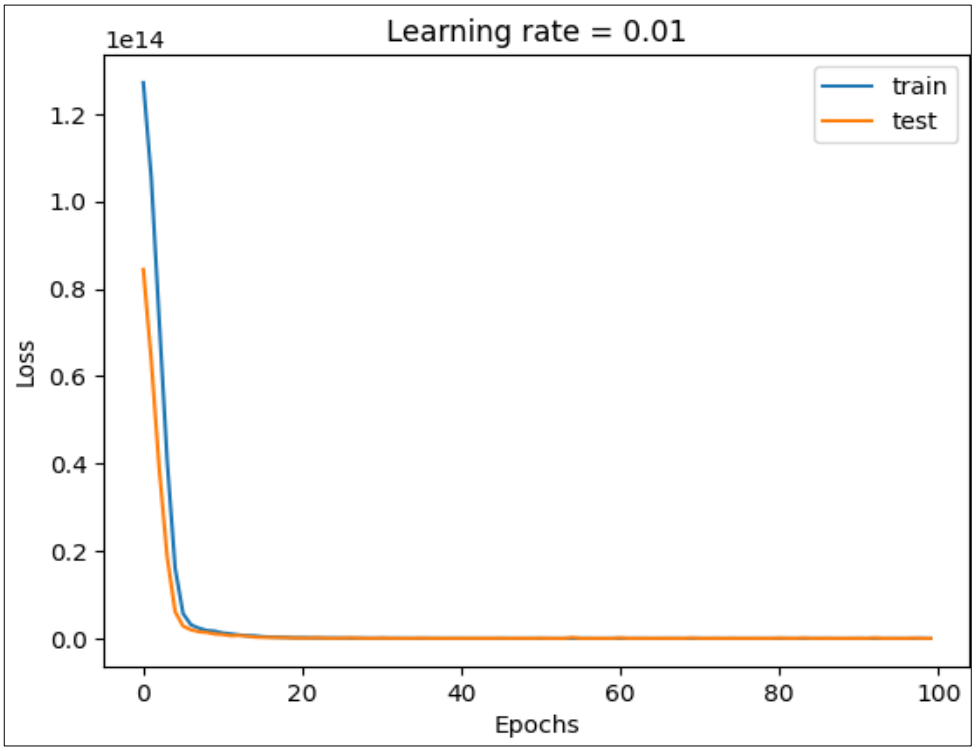

Figura 4 – Curvas de aprendizado do modelo MLP

Posteriormente, foram realizadas previsões para o conjunto de teste, utilizando as métricas mean_squared_error (Erro Quadrático Médio - MSE) e r2_score (Coeficiente de Determinação - R^2). Com um MSE de 172.925.970,99, os resultados indicam uma elevada dispersão entre as previsões do modelo e os valores reais, sugerindo que o modelo tem uma margem significativa de erro. Por outro lado, um valor de R^2 igual a um (1) indica que o modelo de regressão consegue

explicar 100% da variabilidade na variável dependente. Isso significa que as previsões do modelo coincidem exatamente com os valores reais da variável dependente, sem qualquer erro.

Os resultados iniciais dessas métricas estão apresentados no gráfico da Figura 5, que mostra a comparação entre os valores previstos e reais obtidos pelo modelo MLP.

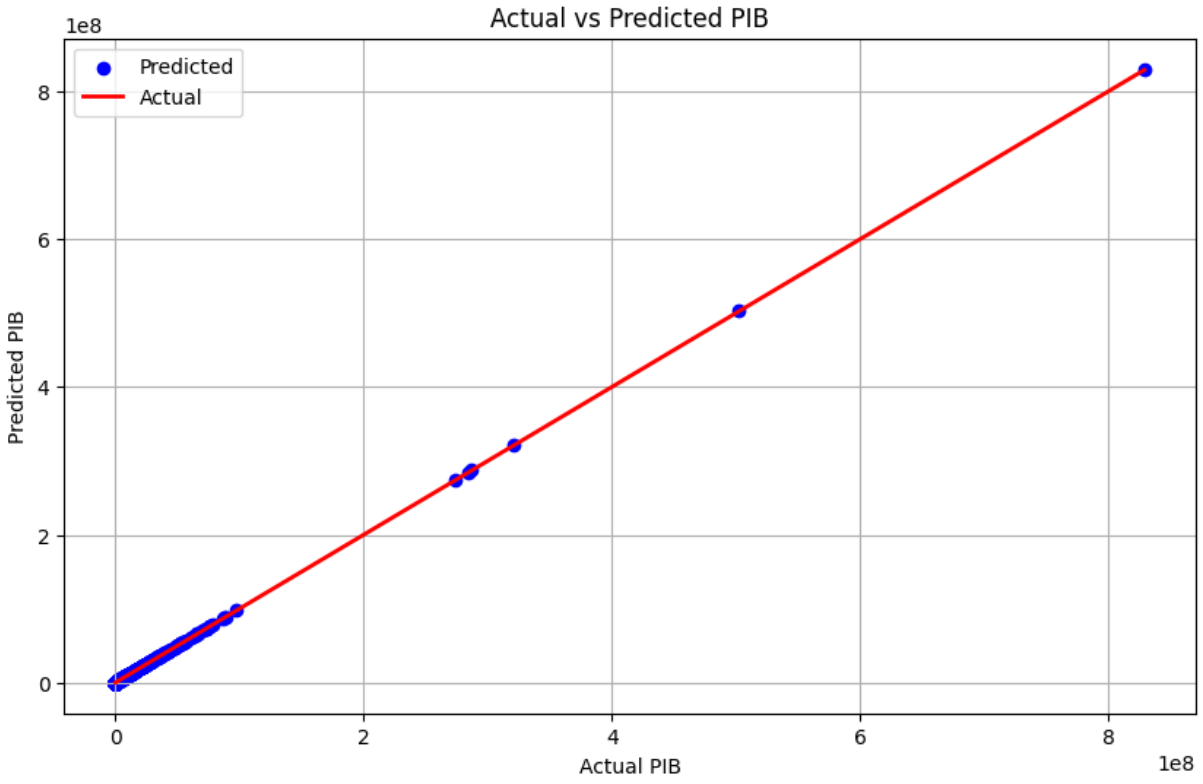

Figura 5 – PIB previsto X PIB real com MLP

Para aprimorar o modelo, foram realizadas várias tentativas de ajustes. Inicialmente, aumentaram-se as épocas (epochs) para 300 e a taxa de aprendizado (lr) para 0,02. O treinamento foi interrompido na época 51, e os resultados mostraram um Erro Quadrático Médio (MSE) de 1.726.769.999,88 e um Coeficiente de Determinação (R^2) de 1,00. Esses resultados indicaram que o modelo ajustado foi menos eficaz do que o modelo inicial, tendo um MSE substancialmente maior do que o modelo inicial.

A seguir, foi testada a função de ativação 'tanh' (hiperbólica tangente), que transforma a entrada em valores entre -1 e 1, com 100 épocas e uma taxa de aprendizado (lr) de 0,02. Os resultados foram insatisfatórios, apresentando um MSE de 92.197.582.848.000,11 e um R^2 de -0,01. Essa configuração mostrou-se inadequada, com um ajuste significativamente pior do que o inicial.

Outra alteração envolveu a troca da função de ativação para 'softplus', que é uma função suave e não-linear semelhante à ReLU, mas que gera valores positivos pequenos para entradas negativas, mantendo 100 épocas e uma taxa de aprendizado de 0,01. Novamente, os resultados foram piores do que os do modelo inicial, com um MSE de 2.955.599.736,40 e um R^2 de 1,00, sugerindo que essa função de ativação não melhorou o desempenho do modelo.

Diante dos resultados, decidiu-se manter a função de ativação ReLU e ajustar outros parâmetros. A quantidade de neurônios foi aumentada para 128 na primeira camada e 64 na segunda camada. O treinamento foi interrompido na época 62 de 100, mas os resultados obtidos não superaram os do modelo inicial, apresentando um MSE de 2.665.352.276,44 e um R^2 de 1,00.

Finalmente, a quantidade de neurônios foi reduzida para 32 na primeira camada e 18 na segunda, com uma taxa de aprendizado diminuída para 0,001. Os resultados obtidos foram um MSE de 446.687.156.015,90 e um R^2 de 1,00.

Apesar das várias tentativas de ajustar o modelo, o modelo inicial com a função de ativação ReLU e parâmetros padrão apresentou as melhores previsões.

CONCLUSÃO

Este trabalho teve como objetivo prever o PIB das cidades brasileiras utilizando algoritmos de Machine Learning. Para isso, foram utilizados os algoritmos Método dos Mínimos Quadrados para Regressão Linear (sem redes neurais) e MultiLayer Perceptrons (MLP), redes neurais, para a modelagem do PIB das cidades brasileiras. A avaliação dos modelos e seus ajustes foi realizada utilizando as métricas Coeficiente de Determinação (R^2) e Erro Médio Quadrático (MSE).

Analisando os modelos, o Método dos Mínimos Quadrados apresentou os melhores resultados com um Erro Médio Quadrático (MSE) de 0,2564 para o conjunto de treino e de 0,2554 para o conjunto de teste, além de um Coeficiente de Determinação (R^2) de 1 tanto para o conjunto de treino quanto para o conjunto de teste, indicando que o modelo foi capaz de prever o PIB com alta precisão e explicando 100% da variabilidade na variável dependente (PIB).

O algoritmo MLP apresentou bons resultados gerais, mas o modelo inicial superou as versões ajustadas com diferentes hiperparâmetros. Embora o R^2 para o modelo inicial tenha sido 1, indicando que ele explicou 100% da variabilidade dos dados, o Erro Médio Quadrático (MSE) para o conjunto de teste foi significativamente mais alto, com um valor de 6419,44. Esse valor é significativamente mais alto em comparação com o MSE obtido pelo modelo usando o Método dos Mínimos Quadrados, o que sugere que, apesar de ter um R^2 perfeito (1), o MLP não foi tão preciso em termos de erro absoluto nas previsões em relação ao modelo usando o Método dos Mínimos Quadrados.

Com base nos resultados obtidos, foi possível concluir que o Método dos Mínimos Quadrados, embora seja um modelo mais simples e sem o uso de redes neurais, apresentou o melhor desempenho para a previsão do PIB das cidades brasileiras. Em comparação com o MultiLayer Perceptron (MLP), que é mais complexo, o Método dos Mínimos Quadrados se mostrou mais eficaz na precisão das previsões, demonstrando ser a melhor abordagem para este problema específico.

REFERÊNCIAS

Domingos, Pedro. O Algoritmo Mestre. Como a busca pelo algoritmo de machine learning definitivo recriará nosso mundo. The Master Algorithm. 1ª Ed. Novac Editora Ltda. 2017. ISBN: 978-85-7522-542-4.

Géron, Aurélien. Mãos à obra: Aprendizado de máquina com Scikit-learn & TensorFlow - conceitos, ferramentas e técnicas para a construção de sistemas inteligentes. Tradução por Rafael Contatori de: Hands-on machine learning with Scikit-learn & TensorFlow. 1ª Ed. Atlas Books. Rio de Janeiro, 2019. ISBN: 978-85-508-0381-4.

Gil, Antonio Carlos. Métodos e Técnicas de Pesquisa Social. 6. ed. São Paulo: Atlas, 2008. ISBN: 978-85-224-5142-5.

Harrison, Matt. Machine learning - guia de referência rápida - trabalhando com dados estruturados em Python. 1ª Ed. Editora Novatec. 2020. ISBN: 978-85-7522-817-3.

Instituto Brasileiro de Geografia e Estatística (IBGE) - Produto interno bruto dos municípios brasileiros. Disponível em: https://www.ibge.gov.br/estatisticas/economicas/contas-nacionais/9088-produto-interno-bruto-dos-municipios.html?t=series-historicas&c=4205407 Acesso em 15/05/2024.

Kovács, Zsolt László. Redes Neurais Artificiais - Fundamentos e Aplicações. 4ª Ed., Editora Livraria da Física, São Paulo – SP. 2006. ISBN: 978-85-8832-514-2.

Mitchell, Tom M. Machine Learning. McGraw-Hill Education. 1ª Ed. 1997. ISBN: 978-00-7042-807-2.

Organização das Nações Unidas (ONU). Agenda 2030 para o desenvolvimento sustentável. Disponível em: https://brasil.un.org/pt-br/91863-agenda-2030-para-o-desenvolvimento-sustent%C3%A1vel. Acesso em: 05/11/2023.

Scikit Learn (SKLEARN) - Metrics and scoring: quantifying the quality of predictions. Disponível em: https://scikit-learn.org/stable/modules/model_evaluation.html. Último Acesso em 22/07/2024.

Silva, Andréa Ferreira da; Almeida, Aléssio Tony Cavalcanti de; Ramalho, Hilton Martins de Brito. Predição do risco de reprovação no ensino superior usando algoritmos de machine learning. Teoria e Prática em Administração. v. 10, n. 2, p. 58-80. DOI. 10.21714/2238-104X2020v10i2-51124. Jul-Dez 2020b.

Silva, Bruno J. B.; Sousa Neto, Pedro B.; Medeiros, Lilian S. de; Medeiros, Elvira H. O. de; Menezes, Andréa M. de; Souza, Pollyanna Thais; Ramalho, Keliane de M. A correlação entre o setor de serviços e o Produto Interno Bruto no Brasil. Research, Society and Development, v.9, n.4, e163943040, 2020a.

Terence, Ana Cláudia Fernandes; Escrivão Filho, Edmundo. Abordagem quantitativa, qualitativa e a utilização da pesquisa-ação nos estudos organizacionais. In: XXVI ENCONTRO NACIONAL DE ENGENHARIA DE PRODUÇÃO – ENEGEP, v. 9, Fortaleza, CE, 2006. Anais [...]. Fortaleza, CE, 2006.