Técnicas de análise de dados e *machine learning* na análise de vendas de autoveículos no Brasil durante a pandemia da Covid-19

Data analysis and machine learning techniques in the analysis of car sales in Brazil during the Covid-19 pandemic

Bernard Roger Ramos Collin
https://orcid.org/0000-0002-1026-2632
Graduando em Engenharia de Produção. Universidade Federal do Vale do São Francisco (UNIVASF) – Brasil. bernard.collin@hotmail.com

Thiago Magalhães Amaral
https://orcid.org/0000-0003-3642-5054
Doutor em Engenharia de Produção. Universidade Federal do Vale do São Francisco (UNIVASF) – Brasil. prof.thiago.magalhaes@gmail.com

Fernanda Magalhães Amaral
https://orcid.org/0000-0003-3945-1856
Doutora em Engenharia Civil. Universidade Federal de Pernambuco (UFPE) – Brasil. nanda25magalhaes@hotmail.com

RESUMO

O presente estudo tem como objetivo analisar os impactos da Covid-19 nas vendas de autoveículos no Brasil, correlacionando esses fatores com as importações de semicondutores e indicadores econômicos através de modelos de *Machine Learning*. Para isso, ocorreu uma coleta e preparação de dados antes de ser executada a análise exploratória, para em seguida, serem desenvolvidos os modelos preditivos. Este trabalho se caracteriza como uma pesquisa aplicada e exploratória, com uma abordagem quantitativa. A análise exploratória e a criação dos modelos foram executadas na linguagem *Python*. Os dados utilizados no estudo foram coletados de múltiplas fontes de empresas privadas, governamentais, associações e institutos de pesquisa e organizados em intervalos mensais. Foram utilizados três algoritmos de regressão: *Random Forest, Multi-Layer Perceptron* e Regressão Linear Múltipla. A rede neural apresentou os melhores resultados dentre os algoritmos aplicados, alcançando um R^2 de 82,01% no conjunto dos dados de teste. Após a análise exploratória, percebeu-se também o alto impacto que os semicondutores têm nas vendas de autoveículos e os efeitos drásticos ocasionado pela Covid-19 nas variáveis utilizadas no estudo. Com o modelo criado, é possível prever as vendas de autoveículos no Brasil em um determinado mês com base em alguns indicadores econômicos, importação de semicondutores e os óbitos mensais da Covid-19.

Palavras-chave: covid-19; *machine learning*; regressão; autoveículos.

ABSTRACT

This study aims to analyze the impacts of Covid-19 on vehicle sales in Brazil, correlating these factors with semiconductors imports and economic indicators using Machine Learning models. For this, a collection and preparation of data took place before performing the exploratory analysis, and the developing of predictive models. This work is characterized as an applied and exploratory research, with a quantitative approach. The Python language was used to perform exploratory analysis and to create the models. The data used in this study were collected from multiples sources of private companies, government, associations and research institutes. Three regression algorithms were used: Random Forest, Multi-Layer Perceptron and Multiple Linear Regression. The neural network presented the best results among the applied algorithms, reaching an R^2 of 82.01% in the test data set. After the exploratory analysis, we also noticed the high impact that semiconductors have on vehicle sales and the drastic effects caused by Covid-19 on the variables used in the study. With the model created, it is possible to predict the sales of vehicles in Brazil in a given month based on some economic indicators, semiconductor imports and the monthly deaths of Covid-19.
Keywords: Covid-19; machine learning; regression; vehicles.

Recebido em 11/11/2021. Aprovado em 23/12/2021. Avaliado pelo sistema *double blind peer review*. Publicado conforme normas da ABNT.
https://doi.org/10.22279/navus.2022.v12.p01-24.1717

1 INTRODUÇÃO

O termo *Big Data* refere-se à grande massa de dados digitais produzidas pela sociedade e, cada vez mais, requer ferramentas de análises mais sofisticadas (RIAHI; RIAHI, 2018). O *Big Data* é conhecido pela imensa quantidade de dados armazenados, em torno de 4,4 *zettabytes* (Zib) ou 44 trilhões de *gigabytes* e pelo fato de 90% do total dos dados criados pela humanidade terem sido gerados nos 2 últimos anos (MARR, 2015; NOGUEIRA, 2019).

Através da Inteligência Artificial (IA) e da Aprendizagem de Máquina (AM), é possível entender essa grande massa de dados, descobrir padrões e correlações ocultas, proporcionando informações inestimáveis para estudos e estratégias competitivas, inclusive em momentos de incertezas e variações anormais como o período de pandemia ocasionado pela Covid-19 (KIRCSH; HURWITZ, 2018). Com o surgimento do vírus, ocorreram muitas incertezas e mudanças nos índices econômicos, na produção industrial e nas demandas de diversos produtos (BOSQUEEROLLI *et al.,* 2020). O setor automobilístico foi um dos setores mais impactados pela pandemia e, segundo Cilo (2020), demorará 3 anos para retomar os níveis de vendas semelhantes ao período anterior.

Com a grande massa de dados sendo gerada continuamente, modelos mais simples tornam-se inviáveis e custosos para prever a demanda de produtos e serviços. Segundo Teixeira (2019), a IA amplia a capacidade e compreensão científica dos dados, com seus processos aperfeiçoados e capazes de processarem diversas informações em um menor período. Libert e Beck (2018) destacam que dados significativos são mais eficientes do que dados abrangentes e a eficiência de um modelo de AM depende diretamente da qualidade e quantidade dos dados envolvidos no estudo.

Wehle (2017) define a AM como sendo qualquer tipo de programa de computador capaz de aprender sobre um conjunto de dados sem ser programado por um humano. Informações ou dados discrepantes que deveriam ser retirados em modelos mais simples de previsão, são possíveis de serem interpretados através da AM, como a queda de vendas em 28,57% de autoveículos no Brasil em 2020 que interrompeu a sequência de crescimento das unidades vendidas (FENABRE, 2020).

De acordo com o IBGE (2020), logo no início da pandemia durante o mês de março, a produção industrial sofreu uma retração de 9,1% e, segundo o Ministério da Economia (2020), o setor da indústria automotiva foi um dos mais afetados, com jornadas de trabalhos reduzidas e muitos contratos suspensos. Além disso, a produção de autoveículos foi afetada pela falta de componentes essenciais, que de acordo com a ANFAVEA (2021), fez a produção recuar por causa da falta de semicondutores pelo segundo mês consecutivo, ocasionando o fechamento de diversas fábricas.

Só nos 6 primeiros meses do ano de 2021, segundo a ANFAVEA (2021), cerca de 120 mil veículos deixaram de ser produzidos no Brasil por falta de peças. Logo no início da pandemia, as vendas de veículos no Brasil caíram cerca de 40% em relação ao ano anterior, com projeções de vendas equivalentes ao ano de 2019 somente no ano de 2025 (ANFAVEA 2020). O Ministério da Economia (2020) complementa afirmando que o setor de fabricação de veículos foi o 8º mais afetado, seguido pelo comércio de veículos que ficou em 10º.

No momento de instabilidade causada pela Covid-19, assim como as quedas nas vendas de autoveículos, outras variáveis sofreram alterações que contrariaram as tendências normais. Como exemplo, tem-se os indicadores econômicos que também foram afetados pela pandemia e irão manter um cenário negativo temporário, reforçando o contexto de incertezas econômicas (FGV, 2022). Segundo Nogueira (2012), os indicadores econômicos são índices que representam uma determinada variável econômica e servem como parâmetros para avaliar a conjuntura financeira de um país.

O uso da AM é bastante amplo e alguns exemplos práticos de aplicação desses modelos preditivos estão descritos em: Paolanti *et al.* (2018) que desenvolveram um modelo de AM, *Random Forest*, para manutenção preditiva de máquinas. Silveira (2019) criou um modelo de previsão de demanda no setor de varejo através do aprendizado supervisionado de máquina e Peres *et al.* (2019), aplicaram AM no controle de qualidade para prever defeitos em uma linha de montagem automotiva; na indústria automobilística destacam-se os trabalhos de Peres *et al.*(2019), que aplicaram diversos algoritmos de classificação para fazer o controle de qualidade de defeitos em uma linha de montagem; Penumuru (2020) que criou um modelo preditivo de classificação de materiais para contribuir com as habilidades cognitivas dos robôs implementados. Vale ressaltar ainda o trabalho de Pereira (2020), que utilizou AM para prever o desgaste de fresas de topo esférico no setor industrial em processos de fresamento e He *et al.* (2021) que utilizou AM para testar e revisar semicondutores em diversas áreas de fabricação.

Durante o período da pandemia, surgiram vários estudos utilizando a AM como é o caso de Tiwari e Khan (2020) que desenvolveram um modelo preditivo sobre a propagação da COVID-19 em 7 dias, analisando os casos da Índia, Coreia do Sul, China e Itália. Já Zoabi, Rozov e Shomron (2021) utilizaram AM para criar um modelo capaz de prever uma infecção positiva com base nas respostas de 8 perguntas e, Fernandes *et al.* (2021), que criaram modelos de AM para prever o prognóstico negativo de COVID-19.

Assim, com base no conteúdo já exposto, surge o questionamento: De que forma, as variações nos indicadores econômicos, nas importações de semicondutores e o número de óbitos da Covid-19 influenciam nas vendas de autoveículos no Brasil e como prever essa demanda através da AM?

Logo, o presente trabalho tem como objetivo analisar, a relação dos indicadores econômicos, o número de óbitos da Covid-19 e importações de semicondutores com as vendas de autoveículos no Brasil através de modelos de regressão da AM.

2 REFERENCIAL TEÓRICO

As próximas seções abordarão aprendizagem de máquina, análise de dados e indicadores econômicos.

2.1 Aprendizagem de Máquina (AM)

Segundo Sierra (2021), o objetivo da AM é permitir que os computadores reconheçam padrões entre dados através das técnicas de algoritmos e matemática. O autor define ainda a AM como sendo um ramo da inteligência artificial e Cid (2019) ressalta a capacidade de inferência dos modelos por conta do aprendizado baseado em dados históricos. A AM pode ser aplicada em quase todos os setores e permite previsões que antecipam cenários e problemas que podem ser analisados previamente. Kircsh e Hurwitz (2018) afirmam ainda o ganho de importância da AM dentro das organizações que buscam soluções inovadoras para solucionar problemas de negócio através dos dados.

Um modelo preditivo de AM permite ao algoritmo identificar padrões e *insights* quando ele é aplicado a um conjunto de dados, com isso o modelo define uma função matemática para futuras previsões. Seymour (2016) define dois tipos de classificação dos modelos preditivos de AM, sendo eles os modelos supervisionados e não supervisionados. O primeiro se refere aos modelos que utilizam um conjunto de dados para treinar, recebendo entradas e saídas de dados, permitindo ao modelo identificar os padrões e relações ocultas entre as entradas e saídas no conjunto de dados. Já os modelos não supervisionados recebem apenas os dados de entrada, sem a saída, e, com isso, ele identifica padrões relacionados aos dados de entrada. Os modelos não supervisionados são utilizados quando se tem o objetivo de descrever eventos ainda não conhecidos.

Um exemplo de modelo supervisionado e presente no referido estudo, ocorre em situações de regressão, onde o modelo aprende com os dados de entrada (variáveis independentes) e com os respectivos dados de saída (variável dependente) para assim então identificar padrões e, através da aprendizagem anterior, fazer a previsão. Souza (2019) destaca as principais etapas para se realizar um projeto de aprendizagem de máquina, sendo elas:

a) Definição do problema;
b) Preparação dos dados;
c) Avaliação dos algoritmos aplicados;
d) Aperfeiçoamento dos resultados;
e) Apresentação dos resultados.

2.1.1 Modelos de Regressão

De acordo com Harrison (2019), a regressão é um modelo de aprendizado supervisionado, utilizado quando se tem o objetivo de prever uma variável numérica ou um dado quantitativo. Existem vários métodos de AM para resolver problemas de regressão. A seguir são apresentados os métodos utilizados para o presente estudo.

2.1.1.1. Regressão Linear Múltipla

Segundo Silveira (2019), a regressão múltipla busca determinar o valor da variável alvo de acordo com a relação linear com 2 ou mais variáveis independentes. O objetivo é encontrar os coeficientes que minimizam a soma dos erros ao quadrado da reta de regressão. A Equação 1 mostra o que o modelo de regressão linear múltipla de aprendizagem de máquina busca definir:

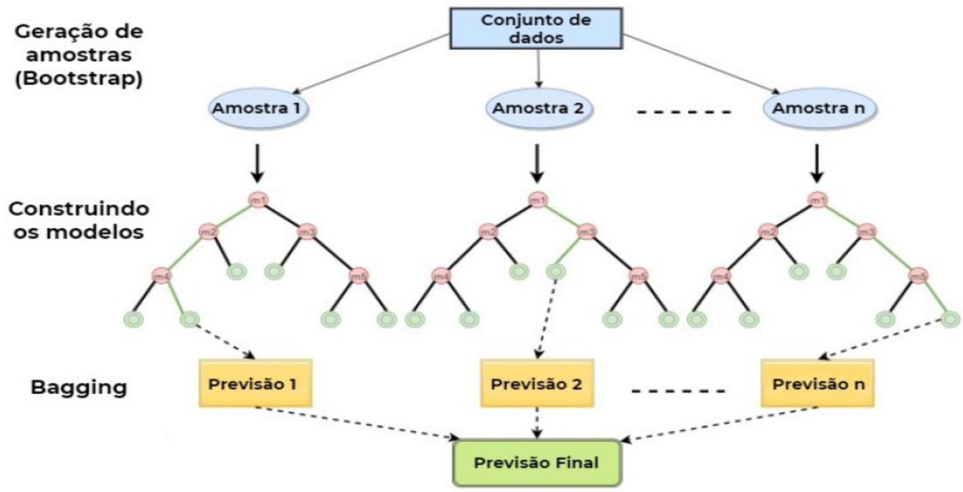
$$Y = bo + b_1X_1 + b_2X_2 + \cdots + b_kX_k \quad (1)$$

2.1.1.2 Random Forest

Géron (2017) destaca a aprendizagem *ensemble* do modelo, ou seja, o algoritmo de *Random Forest* utiliza um agrupamento de diversos outros modelos mais fracos para criar um modelo forte capaz de fazer previsões mais precisas. O modelo utiliza a técnica de *Bagging,* que consiste em combinar o resultado de vários modelos diferentes utilizando diferentes amostras (*Bootstrap*) do mesmo conjunto de dados, a estrutura do modelo pode ser consultada na Figura 1.

Um algoritmo *Random Forest* consegue se adaptar muito bem a conjunto de dados com diversas dimensões diferentes e *outliers* (dados discrepantes e que se diferenciam muito do restante), no referido estudo foi utilizado justamente para entender as *outliers* presentes durante o período da Covid-19.

Figura 1 – Estrutura de *Bagging* presente no algoritmo *Random Forest*



Fonte: Dmitrievsky (2018).

2.1.1.3. Multi-Layer Perceptron Regressor (MLP)

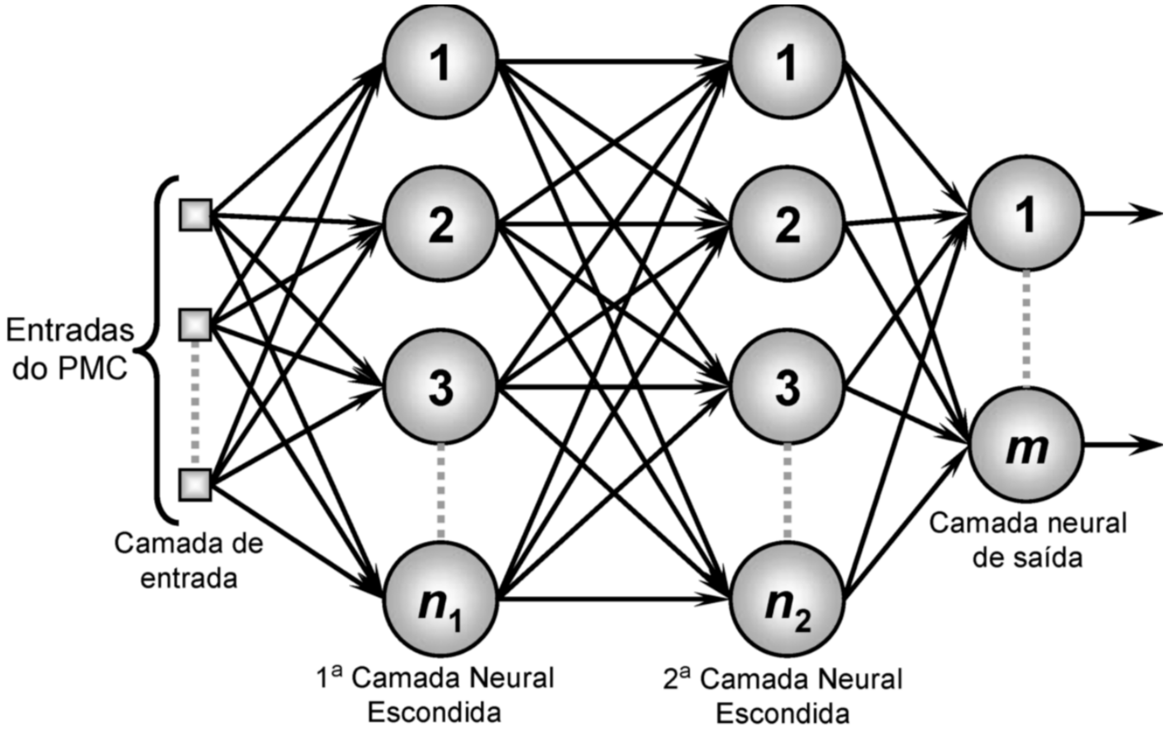
O algoritmo MLP caracteriza-se como uma Rede Neural Artificial (RNA) e possui 3 tipos de camadas principais (SILVA; SPATTI; FLAUZINO, 2016). A primeira é a camada de entrada, seguida pelas camadas ocultas e por fim pela camada de saída. A primeira e segunda camada são compostas e ligadas por neurônios artificiais, também conhecidas como unidades de ativação do modelo.

O MLP compõe vários *Perceptrons,* sendo uma das arquiteturas de RNA, e utiliza um critério de pesos e como aponta Géron (2017), "para cada neurônio de saída que produziu uma previsão incorreta, ele reforçará os pesos de conexão das entradas que contribuíram para a previsão correta". A Equação 2 representa a regra de aprendizado do *Perceptron* através da atualização do peso:

$$W_{i,j}{}^{(próximo\ degrau)} = W_{i,j} + \eta\left(Y_j - \hat{Y}_j\right) X_i \qquad (2)$$

Um exemplo de estrutura de multicamadas do MLP pode ser consultado na Figura 2, onde é possível perceber a camada de entrada, seguida pelas 2 camadas ocultas de neurônios artificias e pôr fim a camada de saída do MLP.

Figura 2 – Estrutura do *Perceptron* Multicamadas (PMC)



Fonte: Silva, Spatti e Flauzino (2016).

2.1.2 Métricas dos modelos de Regressão

A seguir, as principais métricas usadas neste trabalho.

2.1.2.1 Root Mean Squared Error (RMSE)

De acordo com Harrison (2019), a *Root Mean Squared Error* (RMSE) avalia o modelo em termos da variável dependente e serve como forma de se comparar modelos de regressão. Segundo Silveira (2019), "a RMSE é calculada como a raiz quadrada das médias das diferenças quadradas entre a previsão e o dado real".

A escolha de um modelo com base no RMSE se dá em relação aquele que obtiver o menor valor e a Equação 3 representa o seu cálculo:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(Yj - \hat{Y}i\right)^{2}} \qquad (3)$$

2.1.2.2 Coeficiente de determinação (R^2)

O R^2 em modelos de aprendizagem de máquina informa a capacidade do algoritmo de se ajustar a um conjunto de dados, quanto maior seu valor mais o modelo é capaz de se ajustar aqueles dados e fornecer previsões mais precisas. Jain (2019) define o R^2 como sendo uma comparação da soma dos quadrados dos resíduos, fatores aleatórios nos dados que o modelo não conseguiu identificar, com a soma total dos quadrados e quanto mais próximo de 1 melhor é o modelo. A Equação 4 representa o cálculo do R^2:

$$R^2 = \frac{\sum(Yi - Y_{avg})^2}{\sum(Yi - \widehat{Yi})^2} \quad (4)$$

2.2 Análise de dados

Como aponta Bruce (2017), com o aprimoramento da capacidade computacional e, da necessidade de se analisar grandes volumes de dados, surgiram novos softwares que possibilitaram a evolução do escopo do processo de análise exploratória de dados.

Para Lorenzi (2021), o objetivo da análise de dados é encontrar correlações significativas entres as diversas variáveis em um grande conjunto de dados, encontrando *insights* e transformá-los em informação útil que possa ser facilmente visualizada e entendida.

Para se fazer a análise dos dados, a implementação de modelos preditivos de aprendizado de máquina, existem diversas linguagens de programação e *softwares* estatísticos, entre elas vale ressaltar a linguagem *Python* e R (ROCHA, 2020).

2.3 Indicadores econômicos

"Os indicadores econômicos são termos utilizados para medir a performance de uma cidade, região ou pais e até mesmo de uma empresa" (RODRIGUES, 2021). Segundo Stumpf (2019), os indicadores econômicos indicam a situação atual de uma determinada região, em um período determinado, caracterizando se um país ou região está ou não com perspectivas de crescimento econômico.

Amaral (2018) ressalta o uso dos indicadores como parâmetros para tomadas de decisão nas empresas, pois eles fornecem um panorama da situação econômica do país. Ao se entender os indicadores econômicos, as empresas traçam planos financeiros alinhados as tendências econômicas do país.

"Somente por meio da análise dos índices econômicos é possível saber qual é a situação atual, microambiente ou macroambiente, e quais investimentos devem ser feitos" (LACONSKI, 2019). Lemos (2019) complementa dizendo que cada um dos indicadores econômicos demonstra situações diversas e são calculados de formas diferentes.

Os principais indicadores usados neste trabalho estão descritos no Quadro 1.

Quadro 1 - Indicadores Econômicos

Índice Geral de Preços do Mercado (IGPM)	"Calculado pela Faculdade Getúlio Vargas (FGV) calcula a oscilação dos preços para os setores de **atacado e da construção.** É o indicador de confiança dado ao mercado privado, sendo utilizado como indexador de contratos" (STUMPF, 2019).
Índice Nacional de Preços ao Consumidor (INPC)	"Produzido pelo Instituto Brasileiro de Geografia e Estatística (IBGE), o Índice Nacional de Preços ao Consumidor (INPC) mede a variação do custo de vida das famílias com chefes assalariados e com rendimento mensal compreendido entre 1 e 5 salários mínimos mensais" ADVFN (2022).

Índice Nacional de Preços ao Consumidor Amplo (IPCA)	"Este índice é oficial de inflação do Brasil, calculado pelo IBGE. Serve de parâmetro para definir as metas de inflação e para manter o controle com políticas monetárias ou fiscais." (STUMPF, 2019).
Taxa SELIC	"A partir da Selic é possível ter uma noção de quanto dinheiro há no mercado, porque é basicamente isso que ela mostra, já que é a taxa de juros praticada pelos bancos nos empréstimos" (AMARAL, 2018).
Produto Interno Bruto (PIB)	O PIB (Produto Interno Bruto) é um indicador econômico muito usado para demonstrar o crescimento da atividade econômica do país. Ao contrário do que muitos pensam, o PIB não demonstra o estoque de riqueza de um país, mas, na verdade, demonstra o total de produtos e serviços finais que foram produzidos no país durante determinado período." (LEMOS, 2019).
Salário Mínimo	"É o mais baixo valor de salário que os empregadores podem legalmente pagar aos seus funcionários pelo tempo e esforço gastos na produção de bens e serviços" (JUSBRASIL, 2021).
Balança Comercial	"Se refere à diferença entre os valores das exportações e das importações de um país em um certo período, revelando se a nação é compradora ou vendedora nos mercados mundiais de bens e serviços." (LEITE, 2020).
Produção Industrial	Segundo a ADVFN (2021) refere-se à produção da indústria geral de um país e, dentro dos setores produtivos, a indústria é a que exerce um impacto mais significativo no crescimento do produto agregado.

Fonte: Elaborado pelos autores (2021).

Essas definições serviram como base para a fundamentação da análise exploratória desenvolvida no estudo, possibilitando uma melhor compreensão em relação aos resultados e gráficos obtidos.

2.4. Indicadores da Covid-19 e os semicondutores no contexto industrial

"Um indicador de saúde é uma medida projetada para resumir informações sobre um determinado tópico prioritário na saúde da população ou no desempenho do sistema de saúde. Fornecem informações comparáveis e/ou podem exibir o progresso ao longo do tempo entre diferentes fronteiras geog4raficas ou administrativas" (CIHI, 2014).

Alguns dos principais indicadores utilizados para monitorar o avanço da Covid-19, conforme Coelho e Pilecco (2020), são:

- Incidência acumulada: Intensidade com a qual a doença se espalha população;
- Mortalidade: Quociente do número absoluto de óbitos pelo número de indivíduos expostos;
- Letalidade: Gravidade da doença;
- Taxa de positividade: Proporção de testes positivos dentre os testes realizados.

"O efeito indireto da Covid-19 sobre o resultado do governo federal decorre de seu impacto negativo sobre a atividade econômica e da queda da arrecadação de impostos e outras receitas ligadas ao ciclo econômico" (IPEA, 2021).

Para Gonzalez (2021), o maior problema da indústria automotiva brasileira no cenário atual de pandemia não está sendo a falta de demanda e sim a escassez de insumos, como os semicondutores.

Penn (2021) ressalta o fato do cancelamento, por parte da indústria automotiva, de todos os pedidos de semicondutores em março e abril de 2020 e o fato de logo em seguida, quando ressurgiu a necessidade de reabastecimento desses componentes, o lead time já estava em torno de 6 meses.

"Com o início da pandemia os pedidos de automóveis entrarem em colapso, o que resultou na diminuição rápida dos pedidos de semicondutores do setor automotivo. À medida que os pedidos de semicondutores diminuíram para automóveis, os pedidos aumentaram rapidamente para eletrônicos de consumo, à medida que a força de trabalho entrava em quarentena e em um ambiente de trabalho remoto" (Sjoberg, 2021).

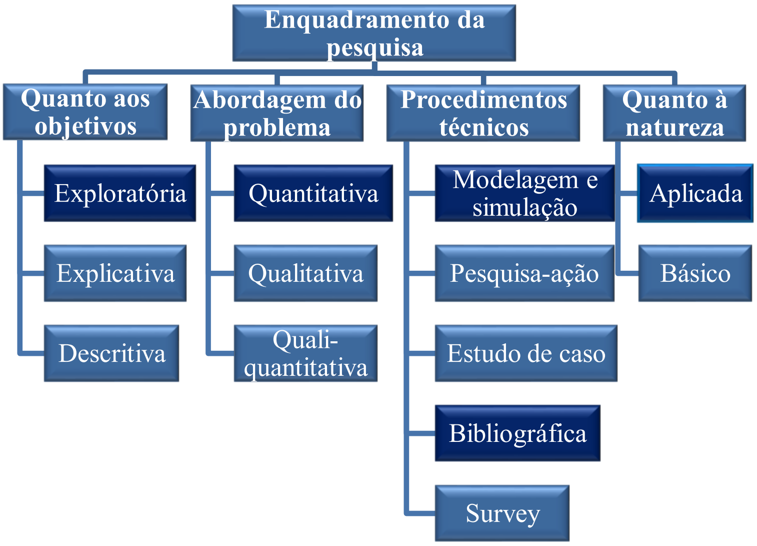
3 PROCEDIMENTOS METODOLÓGICOS

A seguir, serão exibidas a natureza, as etapas e limitações da pesquisa.

3.1 Natureza da pesquisa

Quanto aos objetivos o presente estudo é classificado como de caráter exploratório, devido à análise de dados feita nos indicadores econômicos, no número de importações de semicondutores e com os dados da Covid-19, buscando encontrar padrões e relações com as vendas de autoveículos no Brasil através do AM. A abordagem é quantitativa uma vez que foram utilizados métricas de avaliação, dados estatísticos e análises com base nos resultados dos modelos. Em relação aos procedimentos técnicos, o estudo enquadra-se em modelagem e simulação, pois foi utilizada a linguagem de programação *Python* para se criar algoritmos de AM, assim como a pesquisa se enquadra como bibliográfica visto que foram feitas pesquisas referentes ao estado atual das aplicações de AM e das variáveis utilizadas no estudo. Em relação à natureza, esta pesquisa é aplicada, pois visa manipular variáveis objetivando uma utilidade econômica (MARTINS, 2014; VERGARA, 2000; BERTO; NAKANO, 2000). A classificação do estudo de acordo com a avaliação metodológica está presente na Figura 3, onde as características do presente estudo estão destacadas em azul mais escuro.

Figura 3 – Enquadramento da pesquisa



Fonte: Adaptado de Lacerda, Ensslin e Ensslin (2012).

3.2 Etapas da pesquisa

Antes de coletar os dados e executar as análises foi necessário estabelecer o questionamento a ser respondido. Após a definição do problema, as variáveis foram coletadas em diferentes sites governamentais/institucionais e organizadas em uma planilha da *Microsoft Excel®* totalizando um conjunto de 102 meses (janeiro de 2013 até junho de 2021) com dados de 10 variáveis preditoras e das vendas de autoveículos como variável dependente. As variáveis e suas fontes podem ser consultadas no Quadro 2.

Quadro 2 – Origem dos dados utilizados

Variável	Instituto / organização de origem
Variação mensal do IGPM.	Debit
Variação mensal do INPC.	Instituto Brasileiro de Geografia e Estatística
Variação mensal do IPCA.	Instituto Brasileiro de Geografia e Estatística
Valor em R$ do salário mínimo.	Instituto de Pesquisa Econômica Aplicada
Valor da taxa Selic Over.	Instituto de Pesquisa Econômica Aplicada
Valor em US$ da balança comercial no mês.	Ministério da Economia
Valor em R$ (milhões) do PIB.	Instituto de Pesquisa Econômica Aplicada
Variação da produção industrial.	Instituto de Pesquisa Econômica Aplicada
Novos óbitos mensais da Covid-19.	Ministério da Saúde
Valores em US$ 10.000 da importação brasileira de componentes eletroeletrônicos.	Associação Brasileira da Indústria Elétrica e Eletrônica
Vendas, em unidades, de autoveículos no Brasil.	Associação Nacional dos Fabricantes de Veículos Automotores

Fonte: Elaborado pelos autores (2021)

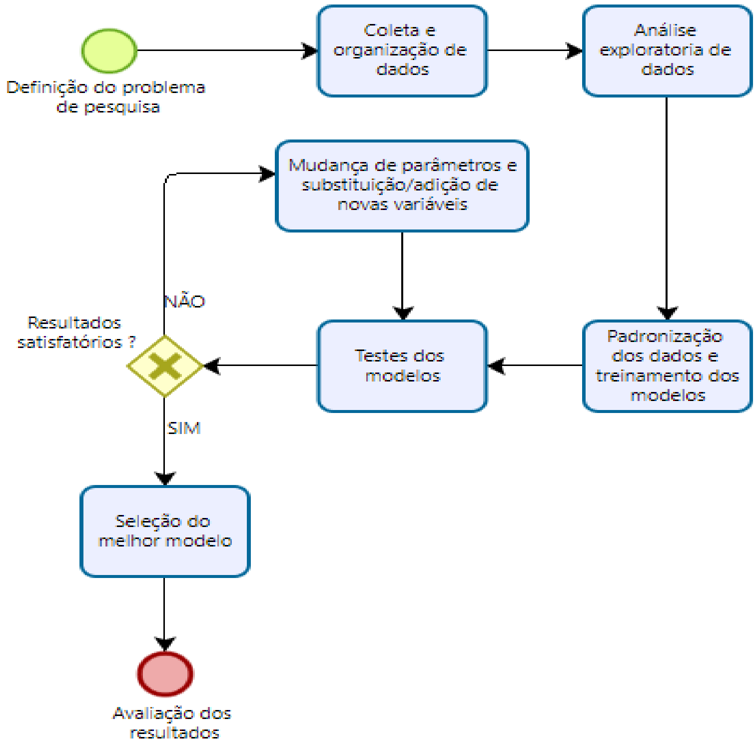
Durante a etapa de pré-processamento (coleta, organização, análise inicial) os dados das variáveis preditoras (indicadores econômicos, óbitos ocasionados pela Covid-19 e importação de componentes eletroeletrônicos) e da variável alvo (vendas de autoveículos no Brasil) foram importados para o ambiente *Python*, onde foi retirada a coluna "Data" do *dataset,* pois a mesma era irrelevante para os modelos de AM.

Com os dados organizados, foi iniciada a etapa de análise para se entender o comportamento e relação entre as variáveis, nesta etapa foram observadas as correlações, características, gráficos e medidas estatísticas dos dados. Feita a análise, a próxima etapa consistiu-se em ajustar os dados para então criar e treinar os modelos de AM, e, por último, a capacidade de previsão com novos dados dos modelos foi testada e avaliada conforme as métricas de desempenho para os modelos de regressão, sendo elas o R^2 e o RMSE. Descreve um pouco os métodos utilizados e coloca da temporalidade e de onde você retirou os dados

O método utilizado para a previsão foi o de aprendizagem supervisionada, utilizando-se da regressão para prever uma variável numérica por meio de diferentes algoritmos e, assim, selecionar o modelo com as previsões mais ajustadas aos valores de vendas mensais reais. Foram utilizados três algoritmos com processos de aprendizagem diferentes. O primeiro algoritmo testado foi o de regressão linear múltipla com o objetivo de se estimar uma equação linear de previsão, posteriormente, foi utilizado um método de aprendizagem ensemble, por meio do algoritmo Random Forest Regressor e, por fim, um modelo com aprendizagem envolvendo neurônios artificiais, a RNA MLPRegressor, foi aplicado.

O conjunto de dados foi separado em 80% para o processo de aprendizagem e 20% para testar a capacidade de previsão dos modelos. Durante o processo de teste algumas variáveis foram substituídas/adicionadas e os parâmetros dos modelos de AM foram ajustados até se alcançar resultados mais precisos, os parâmetros dos modelos e os ajustes nas variáveis ocorreram até que os algoritmos possuíssem métricas de desempenhos satisfatórias para fornecer previsões, com erros baixos e com capacidade de ajuste R^2 em torno de 80% com os dados de teste, caso os resultados não fossem aceitáveis, retornava-se para a etapa de ajuste dos parâmetros nos modelos e das variáveis. Vale ressaltar que no processo de alteração dos parâmetros os modelos eram recriados e os dados separados novamente a fim de se evitar o *Overfitting* (sobre-ajuste) dos modelos com os dados de teste, possibilitando que os modelos fossem testados com dados que não foram utilizados durante a etapa de treinamento. Por fim, o melhor modelo foi selecionado, com base nas métricas referentes as previsões, e seus resultados analisados. As etapas do processo podem ser consultadas na Figura 4.

Figura 4 - Fluxograma das etapas do estudo



Fonte: Elaborado pelos autores (2021).

Vale ressaltar que para o desenvolvimento do estudo foi utilizada a linguagem *Python* (versão 3.7.6) e as principais bibliotecas utilizadas foram:

a) *Pandas* para manipulação do conjunto de dados;
b) *Scikit-learn* para a importação e ajustes dos modelos de AM;
c) *Matplotlib e seaborn* para a criação dos gráficos.

Os modelos foram treinados utilizando um processador i5–7300 HQ 2,5 GHz 12Gb de memória RAM e placa gráfica integrada GTX 1050 de 4Gb, configuração esta que demorou em torno de 2 a 4 minutos para fazer o MLP convergir.

3.3 Limitações

Algumas limitações foram estabelecidas na pesquisa, como o período de análise sendo de janeiro de 2013 até junho de 2021, dados mensais das vendas de autoveículos e valores das variáveis preditoras; visto que antes de janeiro de 2013 não existiam dados para algumas das variáveis utilizadas no estudo. Vale ressaltar ainda que foram feitas análises considerando dois períodos distintos. Os dados foram analisados durante o período anterior e o período de ocorrência Covid-19, a fim de se observar melhor os impactos da pandemia nas vendas de autoveículos.

Destaca-se ainda a inexistência de trabalhos nacionais ou internacionais similares, com aplicações de modelos de AM para prever vendas de autoveículos. Por fim, outra limitação foi a de não serem feitas previsões para os anos seguintes, pois os modelos de AM criados necessitam das variáveis preditoras para prever as vendas de autoveículos em um mês e não foram encontrados dados ou projeções mensais para todas as variáveis selecionadas.

4 RESULTADOS E DISCUSSÃO

A seguir, serão exibidas as fases de pré-processamento dos dados e análise exploratória dos dados, assim como as discussões deste trabalho.

4.1 Pré-processamento dos dados

As diferentes variáveis do estudo foram nomeadas de maneira a facilitar a execução das linhas de códigos. Os nomes atribuídos às variáveis podem sem consultados no Quadro 2:

Quadro 2 – Nomeação das variáveis utilizadas no estudo

Nome da coluna	Variável
Var._IGPM	Variação mensal do IGPM.
Var._INPC	Variação mensal do INPC.
Var._IPCA	Variação mensal do IPCA.
Salário_mín.	Valor em R$ do salário mínimo.
Selic_a_a	Valor da taxa Selic Over.
Balança_comercial_USS	Valor em US$ da balança comercial no mês.
PIB_milhões_RS	Valor em R$ (milhões) do PIB.
Var_produção_industrial_a_a	Variação da produção industrial.
Óbitos_Covid	Novos óbitos da Covid-19 no mês.
Importação_eletro_1	Valores em US$ 10.000 da importação brasileira de componentes eletroeletrônicos.
Unidades_vendidas	Vendas em unidades de autoveículos no Brasil.

Fonte: Elaborado pelos autores (2021).

Vale ressaltar ainda, que foram utilizados os óbitos da Covid-19 como variável pois ela demonstrou uma maior correlação com as vendas de autoveículos do que os casos do vírus. Os casos da Covid-19 foram utilizados anteriormente e ao perceber um melhor ajuste dos modelos com os óbitos a mesma foi substituída, este processo de substituição está presente na etapa "Mudança de parâmetros e substituição/adição de novas variáveis" descrita no processo metodológico.

4.2. Análise exploratória dos dados

O primeiro passo da análise foi identificar as variáveis que não possuíam significância estatística com relação as vendas de autoveículos. Após a análise em Python, verificou-se que as variáveis "var_IGPM", "var_INPC" e "var IPCA" não eram estaticamente significativas (considerando um p-valor = 0,05) sendo então removidas do conjunto de dados. Portanto, foram utilizadas no estudo apenas as variáveis estatisticamente significativas, onde a maioria ficou com o p-valor igual a 0.000, com exceção da variável "Balança_comercial_USS" e "importação_eletro_1" onde o p-valor foi de 0.005 e 0.016 respectivamente.

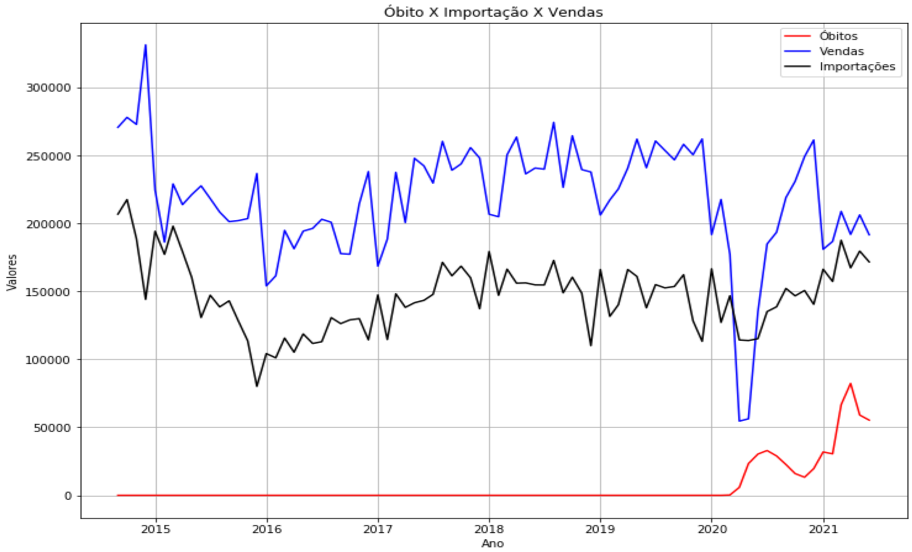
Foi possível identificar também, através do comando "*describe*()", a média das unidades vendidas sendo de 231.506 unidades, o menor valor de venda em 54.580 unidades (período da Covid-19) e o maior valor de unidades vendidas em um mês equivalendo a 332.087 unidades.

Buscou-se analisar também o comportamento de algumas variáveis em relação às vendas, onde foi possível perceber o impacto ocasionado pela Covid-19. Por meio da Figura 5 é possível perceber as quedas nas importações de componentes eletroeletrônicos quando os óbitos da Covid-19 no Brasil começaram a surgir, o que impactou bruscamente nas unidades vendidas de autoveículos durante o início de 2020.

O eixo y da figura 5 representa os novos valores mensais das 3 variáveis (novos óbitos da Covid-19 no mês, vendas mensais de autoveículos e importações de componentes eletroeletrônicos) e o eixo x representa os meses dos anos utilizados no estudo. Os semicondutores são peças fundamentais em diversos outros setores, não somente para a indústria automobilística, setores de games, celulares e dispositivos eletroeletrônicos em geral utilizam os semicondutores como componentes essenciais. Com o confinamento e o aumento dos trabalhos remotos ocasionados pela pandemia da Covid-19, as demandas por semicondutores aumentaram, uma vez que esses fatores levaram a uma maior necessidade de notebooks, computadores, vídeo games e outros dispositivos eletrônicos por conta da nova rotina de confinamento.

As quedas nas importações de semicondutores não foram expressivas, porém ao se manter durante 3 meses com valores baixos, todas as cadeias que demandam dessas peças foram afetadas, dentre elas está a indústria automobilística onde, por meio da Figura 5, é possível perceber grandes impactos nas vendas mensais de autoveículos. Percebe-se ainda a retomada do crescimento das importações de semicondutores nos meses seguintes, chegando a ultrapassar o pico de 2018, devido à alta demanda gerada.

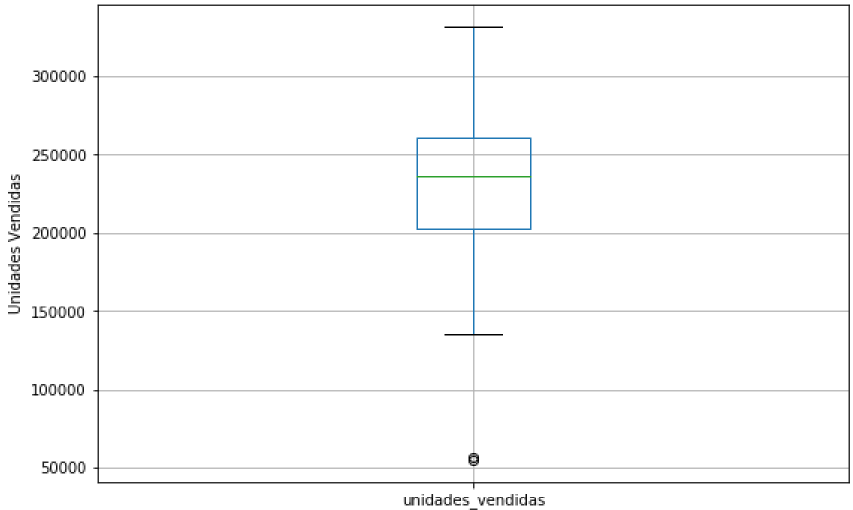
Figura 5 – Óbitos da Covid-19, importação de componentes eletroeletrônicos e vendas de autoveículos no Brasil ao longo dos anos



Fonte: Elaborado pelos autores (2021).

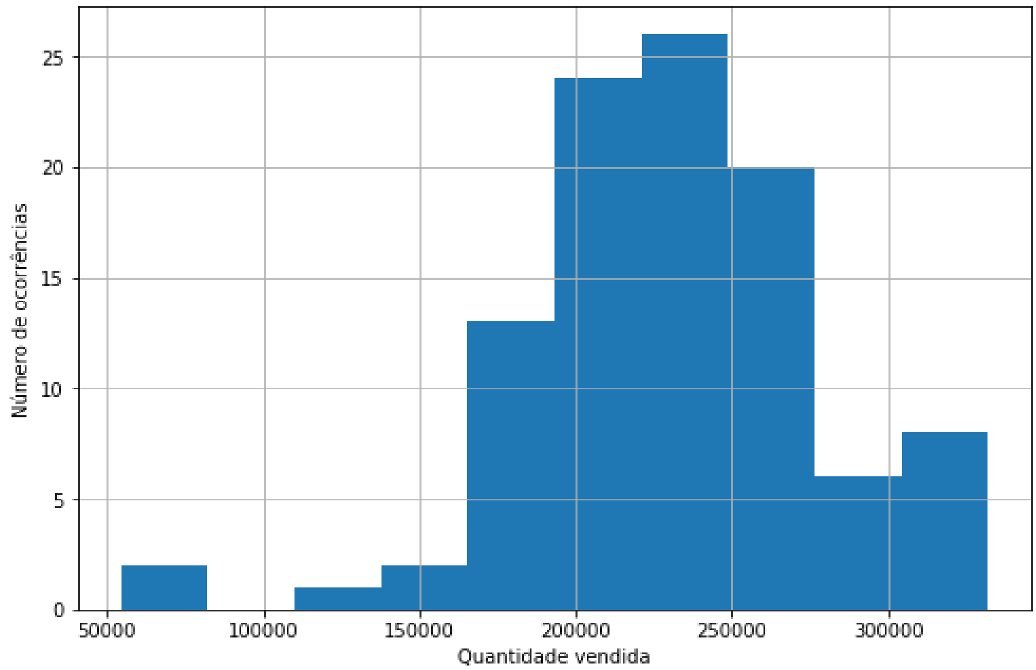
Nas Figuras 6 e 7 estão representados, respectivamente, o *Boxplot* e o Histograma das vendas e, a partir deles, é possível perceber a distribuição assimétrica negativa das vendas (linha da mediana próxima ao terceiro quartil) e a concentração das vendas em torno de 250.000 mil unidades.

Figura 6 – Boxplot das vendas de autoveículos



Fonte: Elaborado pelos autores (2021).

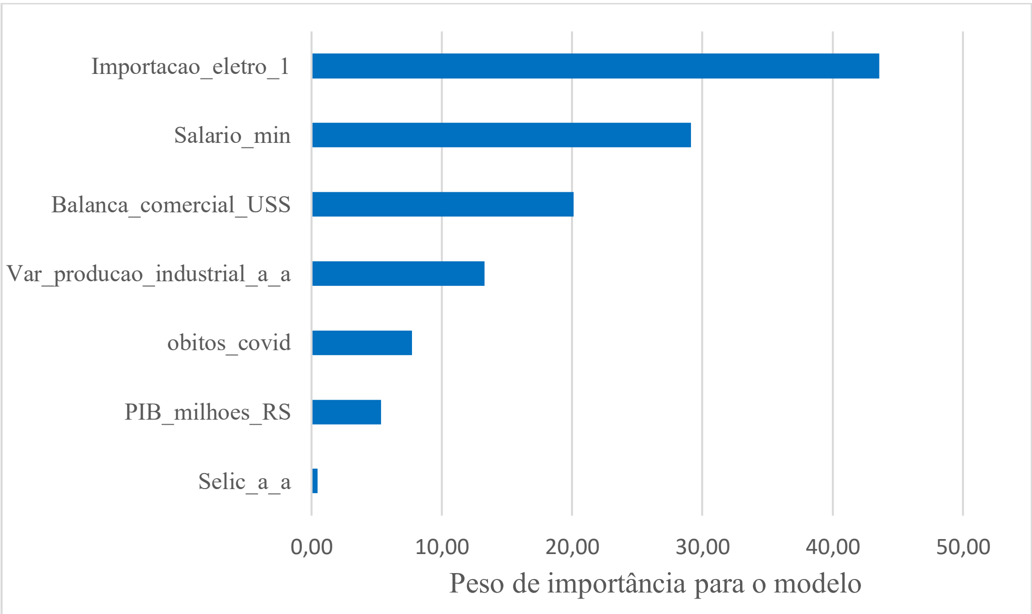
Figura 7 – Histograma das vendas de autoveículos



Fonte: Elaborado pelos autores (2021).

Foi possível identificar também as variáveis mais importantes para os modelos de AM. A Figura 8 indica o nível de importância das variáveis, onde os valores representam os pesos atribuídos pelo modelo de AM *Random Forest* e quanto maior o número, mais importante é aquela variável para o algoritmo aprender sobre os dados.

Figura 8 – Peso de importância das variáveis para o modelo *Random Forest*



Fonte: Elaborado pelos autores (2021).

Com base nos pesos indicados pelo algoritmo de AM, a variável mais importante foi a importação de componentes eletroeletrônicos, isso se dá ao fato da mesma incluir os semicondutores nas importações, como citado anteriormente no estudo, são peças fundamentais nas fabricações de autoveículos e de eletroeletrônicos em geral e a redução dessas importações impactaram seriamente nas vendas de autoveículos. Ao se retirar "importação_eletro_1" do conjunto de dados o desempenho do modelo cai bruscamente, mas ao se retirar a variável "Selic_a_a" o desempenho do modelo é pouco afetado.

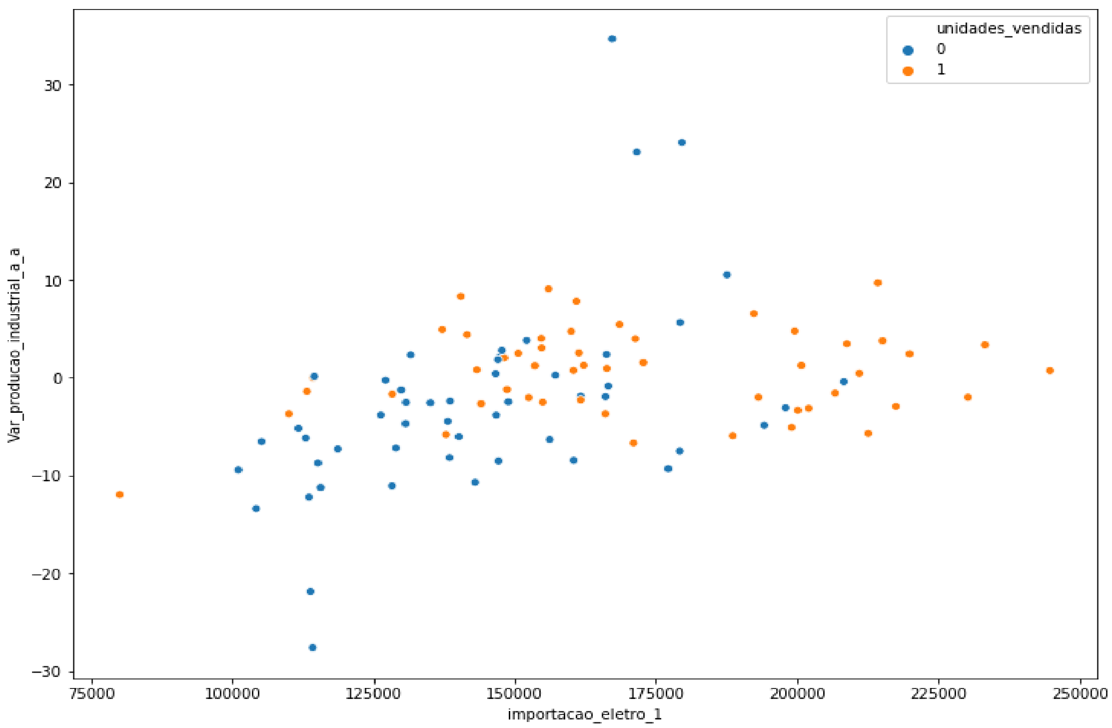
Por fim, ao se identificar as variáveis mais importantes para as previsões e a distribuição assimétrica negativa das vendas, foram feitas análises considerando esses fatores e a mediana. As vendas de cada mês foram classificadas em valores abaixo e acima da mediana (236.545 unidades), onde os valores abaixo da mediana foram classificados com cores azuis e com o valor binário 0 e os valores acima da mediana com valor binário 1 e cores laranjas.

A Figura 9 demonstra um gráfico de dispersão, indicando que quanto mais se faz importação de componentes eletroeletrônicos mais as vendas tendem a ser acima da mediana, é possível perceber também uma tendência de concentração das vendas abaixo da mediana quando a variação da produção industrial está em torno de 0 a -10% e acima da mediana com a variação da produção industrial entre 0 a 10%.

A Figura 10 demonstra as vendas de autoveículos se concentrando quando a importação de componentes eletroeletrônicos está em torno de 1.500.000.000 US$ e a variação da produção industrial entre 0 a 7% aproximadamente.

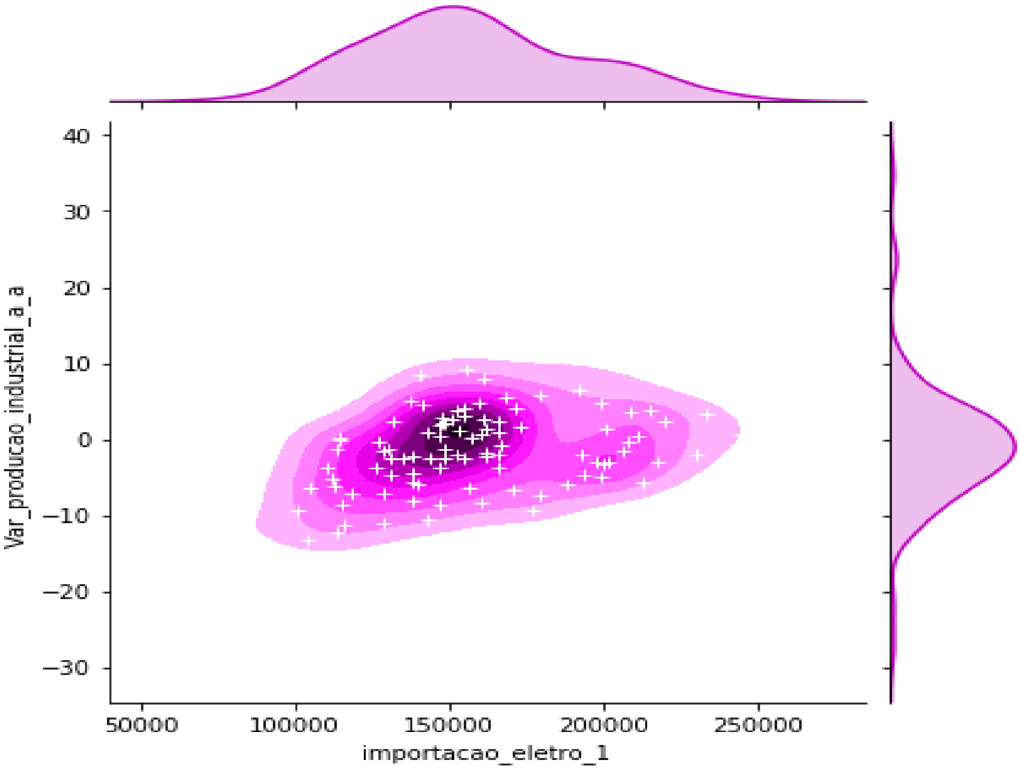
Com o objetivo de se entender as vendas no período da Covid-19 e com base na classificação em torno da mediana já citada anteriormente, foi plotado um gráfico de dispersão, considerando agora os óbitos da Covid-19 no eixo y e as importações de componentes eletroeletrônicos no eito x.

Figura 9 – Diagrama de dispersão das vendas acima e abaixo da mediana



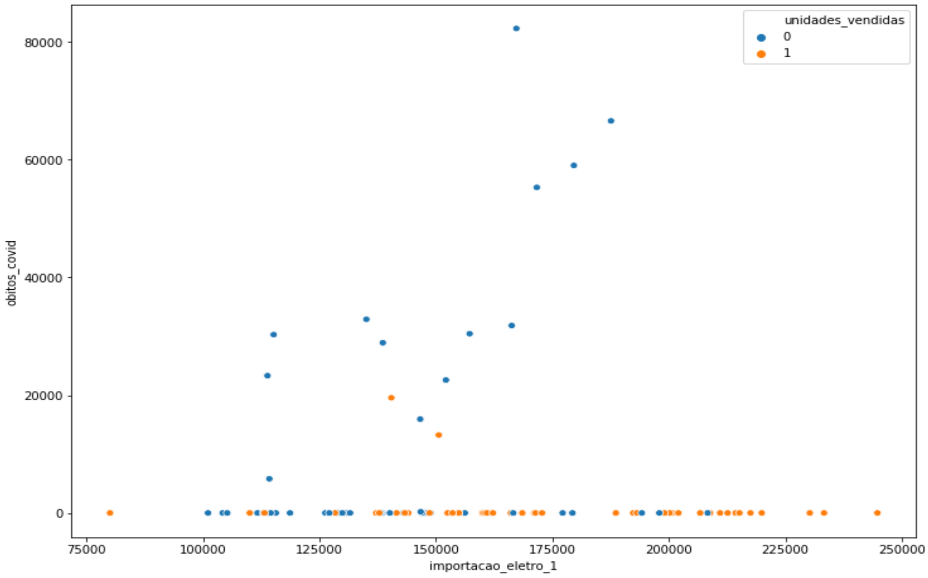
Fonte: Elaborado pelos autores (2021).

Figura 10 – Distribuição bivariada das vendas



Fonte: Elaborado pelos autores (2021).

Figura 11 – Diagrama de dispersão das vendas no período da Covid-19



Fonte: Elaborado pelos autores (2021).

Com a Figura 11, fica evidente, mais uma vez, o impacto da pandemia nas vendas de autoveículos no Brasil ao ser identificado mais pontos azuis (vendas abaixo da mediana) quando os óbitos ocasionados pela Covid-19 começam a surgir (eixo y).

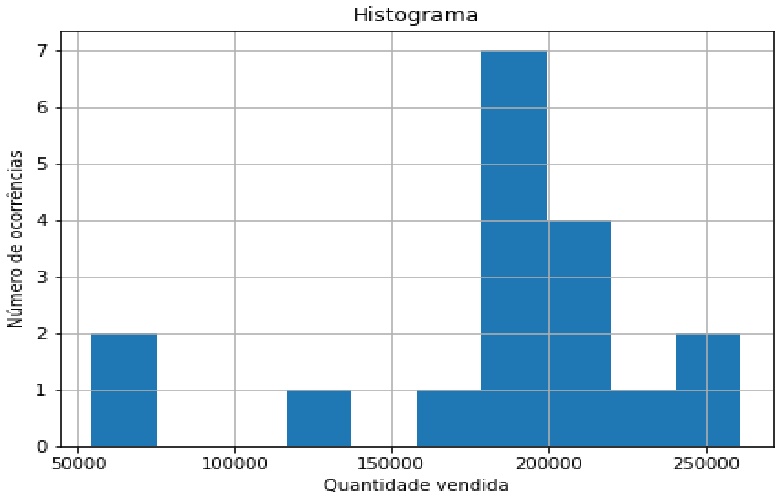
No período anterior a pandemia, quando ainda não existia óbitos ocasionados pela Covid-19, é possível perceber a tendência das vendas acima da mediana quando os valores das importações aumentavam, porém quando os óbitos da Covid-19 começaram a surgir é possível identificar as vendas, em sua grande maioria, com valores abaixo da mediana.

4.2.1. Análise de dados durante o período da Covid-19

Para possibilitar uma melhor visualização dos impactos nas vendas durante o período da Covid-19 eles foram isolados no período de janeiro de 2020 até junho de 2021. Através da Figura 12 é possível perceber a concentração das vendas durante esse período em torno de 200.000 unidades vendidas aproximadamente e algumas observações que se distanciam muito do restante, como os casos em torno de 50.000 e 125.000 unidades vendidas em um mês, pontos estes que ocorreram durante a pandemia.

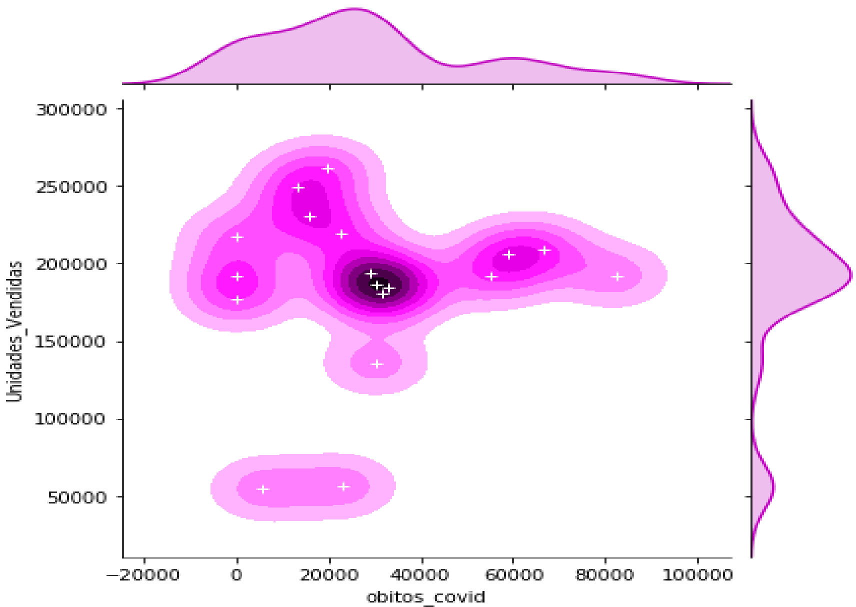
Com a Figura 13 é perceptível um maior número de vendas, em torno de 200.000 unidades, com os óbitos da Covid-19 se concentrando em torno de 30.000 em um mês, além disso, percebe-se que os menores valores de vendas, em torno de 50.000 e 125.000 unidades, ocorreram quando os óbitos começaram a surgir.

Figura 12 – Histograma das vendas durante o período da Covid-19



Fonte: Elaborado pelos autores (2021).

Figura 13 – Distribuição bivariada durante período da Covid-19



Fonte: Elaborado pelos autores (2021).

Foi utilizado o comando "*describe*()" para identificar as principais diferenças estatísticas entre o período anterior e o período da Covid-19. A Tabela 1 indica as informações durante esses períodos, evidenciando a queda das vendas em valores de média, mínimo e máximo em um mês assim como na mediana.

Tabela 1 – Diferenças estatísticas antes e durante a Covid-19

	Período anterior à Covid-19 (2013 - 2020) (unidades vendidas)	Período durante à Covid-19 (2020 - 2021) (unidades vendidas)
Média	240.801	185.400
Menor valor	154.012	54.580
Mediana	238.856	191.778
Maior valor	332.087	261.321

Fonte: Elaborado pelos autores (2021).

4.3 Modelos de *Machine Learning*

Para a criação dos modelos preditivos foram utilizados 3 métodos de regressão, sendo eles a Regressão Linear Múltipla, *Random Forest* e a RNA *Multi-Layer Perceptron Regressor*. Os modelos foram treinados com 80% do conjunto de dados e utilizados para fazerem previsões com os 20% restantes dos dados, o conjunto de teste.

Vale ressaltar que é normal os modelos de AM possuírem diferentes valores de métricas de desempenho para os conjuntos de dados de treino e de teste, pois os modelos com o conjunto de treino fazem previsões com base nos dados utilizados na etapa de aprendizagem, ou seja, dados já conhecidos pelo modelo anteriormente.

Porém, com os dados de teste os valores de desempenho tendem a serem menores, pois o algoritmo faz a previsão com base em dados novos, não conhecidos pelo modelo. Em termos de escolha, são selecionados os modelos com melhores desempenhos para com o conjunto de teste, visto que representam a real capacidade de previsão do modelo para novos dados.

Os 3 modelos foram treinados, testados e as métricas de desempenho foram organizadas na Tabela 2, onde é possível perceber que o modelo de *Random Forest* obteve o melhor desempenho para fazer previsões com base no conjunto dos dados de treino, porém para fins de previsões reais, com novos conjuntos de dados, o melhor modelo foi o *MLP Regressor,* em razão dos melhores resultados obtidos com o conjunto dos dados de teste, representando uma capacidade de ajuste de 82,01% com novos dados e um erro médio de 14.685 mil unidades vendidas nas previsões.

Embora citada no referencial teórico, não foram apresentados os resultados dos coeficientes de regressão múltipla e suas respectivas significâncias.

Tabela 2 – Desempenho dos modelos

	Dados de treino		**Dados de teste**	
	R2 *Score* (%)	RMSE	R2 *Score* (%)	RMSE
Regressão Linear Múltipla	88.25	17.440	77,48	16.432
Random Forest	**95.71**	**10.535**	79,21	15.788
MLP Regressor	90,03	16.060	**82,01**	**14.685**

Fonte: Elaborado pelos autores (2021).

Os valores dos coeficientes encontrados pelo método da Regressão Linear Múltipla podem ser consultados na Tabela 3, o valor da constante foi de 313577.087893:

Tabela 3 – Coeficientes da equação linear encontrada

Variável	**Coeficiente**
Salário Mínimo	-591.763861
Selic	- 3226.229449
Balança Comercial	0,000003
PIB	0.812506
Variação da produção industrial	2730.652521
Óbitos Covid-19	-2.198226
Importação de componentes eletroeletrônicos	0.231096

Fonte: Elaborado pelos autores (2021)

Vale ressaltar ainda que os modelos de Regressão Linear Múltipla e *Random Forest* obtiverem bons desempenhos para serem utilizados em previsões, uma vez que os ajustes de ambos ficaram em torno de 80% com o conjunto de teste, no entanto não foram superiores a RNA.

4.3.1 Multi-layer Perceptron Regressor

Para aplicar os dados no modelo de RNA foi necessária uma padronização, para que as diferentes escalas das variáveis não influenciassem o desempenho do algoritmo, visto que as variações das diferentes variáveis estão em diferentes escalas variando de centenas até milhões com valores em reais e outras variáveis com valores em dólares, diferentes taxas de porcentagem e os óbitos da Covid-19 com variações de 0 até valores próximos de 90.000. Com isso os dados foram transformados, no ambiente *Python*, de forma a ficarem com a distribuição padronizada com um valor médio 0 e desvio padrão de 1.

Para encontrar os melhores parâmetros da rede neural, diversos testes foram realizados até achar um conjunto de parâmetros em que o modelo conseguisse convergir para o menor erro. Os testes foram feitos com os 20% dos dados separados na etapa de aprendizagem e para se evitar o sobreajuste do modelo a cada teste, o conjunto de dados era separado a cada configuração testada e o modelo recriado. Os parâmetros utilizados podem ser consultados na Tabela 4.

Tabela 4 – Configuração de parâmetros do MLP

Nome do parâmetro	Configuração utilizada
Função	"relu"
Nº de camadas ocultas	3
Neurônios por camada	2
Taxa de aprendizagem inicial	0.1
Otimizador de pesos	"lbfgs"
Nº máximo de interações	20.000
Tolerância de otimização	0.0000001

Fonte: Elaborado pelos autores (2021)

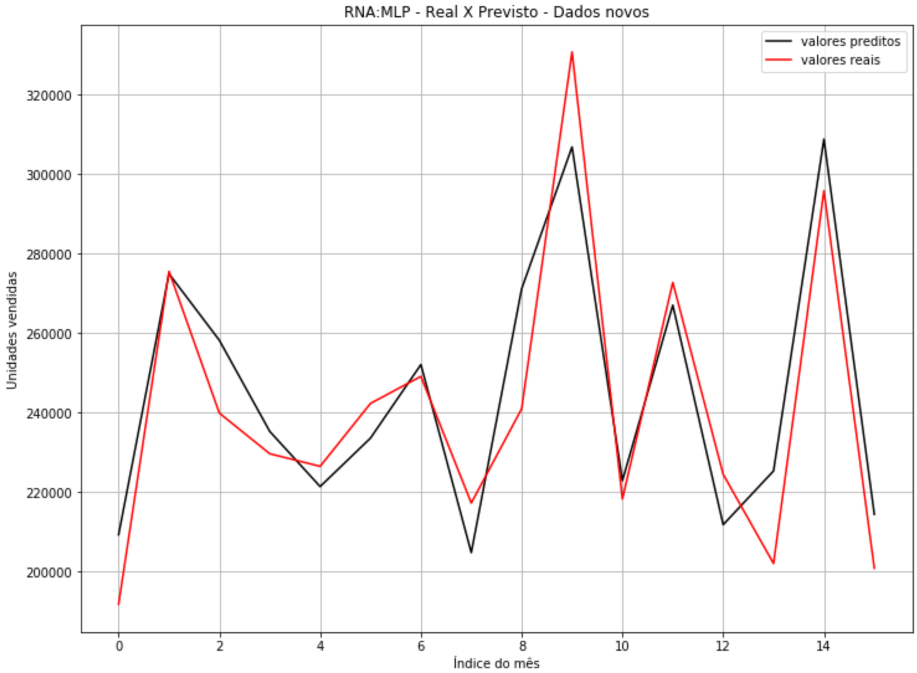
Ou seja, quando o algoritmo não melhorou em 0.0000001, tendo como base o máximo de 20.000 interações, utilizando o restante dos parâmetros configurados, o modelo convergiu e o treinamento foi interrompido.

4.3.1.1 Desempenho da rede neural artificial

Com o objetivo de testar a capacidade de previsão do modelo, ele foi testado com um conjunto de 21 meses, correspondente aos 20% dos dados separados para o teste, que não foram utilizados na fase de treino.

Analisando a Figura 14, é possível perceber o bom desempenho do modelo (já evidenciadas nas métricas) ao fornecer previsões bem ajustadas, por conta do alto valor do R^2, não seguindo apenas tendências de crescimento ou decaimento das vendas. Ao se analisar o RMSE, equivalente em 14.685 unidades, ele corresponde a um erro de 6,34% em relação ao valor médio de vendas mensais.

Figura 14 – Desempenho das previsões da RNA



Fonte: Elaborado pelos autores (2021).

O presente estudo foi comparado com outros trabalhos referentes a modelos de AM. Como comparação, foi utilizada a métrica R^2, que como dito anteriormente, reflete a capacidade do modelo de se ajustar aos dados e quanto mais próximo de 100% mais as previsões tendem a serem precisas. Em aplicações do *MLP Regressor* tem-se Dantas (2018) que aplicou a RNA para prever a quantidade de novos casos de Leptospirose em Recife-PE, utilizando dados históricos do número de casos e dos dados pluviométricos da cidade de 2008 até 2017 e mais duas variáveis climatológicas. Ainda aplicando *MLP Regressor,* é possível citar Santos (2020) que utilizou o algoritmo para prever a produtividade de soja utilizando dados históricos da produtividade e das informações climáticas mensais na fronteira agrícola do MATOPIBA no Brasil.

Por fim, pode-se citar Rozanec *et al* (2021), que de maneira semelhante a este trabalho, aplicaram mais de um algoritmo de AM com o objetivo de selecionar o modelo com a melhor performance para fazer a previsão de demanda de partes e subsistemas de automóveis. Ao todo, os autores aplicaram 21 técnicas de AM, sendo selecionado o modelo *Support Vector Regressor* (SVR) com o melhor desempenho. A Tabela 5 representa o comparativo dos trabalhos com o presente estudo.

Tabela 5 – Comparativo dos melhores resultados

Algoritmo	**Trabalho**	**R^2 (%)**
MLP Regressor	Dantas (2018)	74,00
MLP Regressor	Santos (2020)	86,00
SVR	Rozanec *et al* (2021)	92,12
MLP Regressor	Presente estudo	82,01

Fonte: Elaborado pelos autores (2021).

Ao se analisar a Tabela 5, percebe-se que o presente trabalho apresentou um bom desempenho (R^2 acima de 80%) em relação à literatura, possuindo um R^2 superior ao MLP de Dantas (2019) e próximo ao modelo MLP desenvolvido por Santos (2020), evidenciando que o modelo desenvolvido apresenta uma alta capacidade de ajuste em relação aos dados. Vale ressaltar ainda a importância do trabalho de Rozanec *et al* (2021) onde evidenciaram o impacto positivo em se testar várias técnicas, com diferentes combinações de variáveis e parâmetros para o desenvolvimento de um modelo mais preciso, onde apresentou um R^2 de 92,12% utilizando outro algoritmo.

5 CONSIDERAÇÕES FINAIS

Este estudo abordou técnicas de análise de dados e aplicação de modelos de AM na previsão das vendas de veículos utilizando 3 algoritmos de regressão de aprendizagem supervisionada. A partir dos indicadores econômicos selecionados, importações de componentes eletroeletrônicos e dos óbitos da Covid-19, foi criado um conjunto de dados que foi aplicado aos modelos de AM.

Os modelos foram treinados, testados e foi selecionado o algoritmo *MLP Regressor* com o melhor desempenho para fornecer previsões mais precisas. Durante o treinamento, ocorreram algumas dificuldades na seleção dos parâmetros para ocorrer a convergência da rede neural, sendo solucionados após uma fase de testes de parâmetros. Com os melhores parâmetros encontrados e aplicados no conjunto de teste, o modelo realizou previsões para um conjunto de dados que não foram apresentados durante a fase de treino, onde a RNA criada apresentou um bom desempenho sendo capaz de fornecer previsões bem ajustadas das vendas de autoveículos no Brasil.

O objetivo do trabalho foi alcançado uma vez que foi feita uma análise exploratória dos dados, trazendo ainda informações relevantes sobre os impactos da Covid-19 nas vendas de autoveículos no Brasil e, foram feitos modelos de AM capazes de fornecer previsões das vendas com base nos indicadores econômicos, óbitos da Covid-19 e importação de componentes eletroeletrônicos, evidenciando o potencial preditivo dos algoritmos de AM.

Vale ressaltar ainda, que os resultados mostraram que a variável que refletia a importação de semicondutores foi a mais importante no aprendizado dos modelos de AM, reforçando o que foi encontrado na literatura atual sobre os impactos da falta de semicondutores nas vendas de autoveículos. O salário mínimo obteve um alto peso de importância, uma vez que a mesma reflete o poder aquisitivo da população e com o aumento do poder de compra, o indivíduo consegue adquirir mais bens e serviços. Em terceiro na escala de importância de variáveis está a balança comercial, onde a mesma reflete a desvalorização ou valorização do real em relação com as moedas estrangeiras. Além disso, o algoritmo MLP apresentou ótimos resultados em fornecer previsões precisas mesmo considerando os outliers que ocorreram durante o período da Covid-19.

Algumas limitações foram estabelecidas na pesquisa, como o período de análise sendo de janeiro de 2013 até junho de 2021, dados mensais das vendas de autoveículos e valores das variáveis preditoras; visto que antes de janeiro de 2013 não existiam dados para algumas das variáveis utilizadas no estudo. Vale ressaltar ainda que foram feitas análises considerando dois períodos distintos. Os dados foram analisados durante o período anterior e o período de ocorrência Covid-19, a fim de se observar melhor os impactos da pandemia nas vendas de autoveículos.

Destaca-se ainda a inexistência de trabalhos nacionais ou internacionais similares, com aplicações de modelos de AM para prever vendas de autoveículos. Por fim, outra limitação foi a de não serem feitas previsões para os anos seguintes, pois os modelos de AM criados necessitam das variáveis preditoras para prever as vendas de autoveículos em um mês e não foram encontrados dados ou projeções mensais para todas as variáveis selecionadas.

Por fim, para trabalhos futuros recomenda-se uma nova coleta de dados referentes às variáveis para serem aplicadas ao modelo. Além disso, novas variáveis podem ser adicionadas para substituir as variáveis com menores níveis de importância para os modelos, como o caso da taxa Selic. Recomenda-se também testar outros parâmetros do MLP Regressor ou até mesmo outros modelos de AM para buscar previsões ainda mais precisas.

REFERÊNCIAS

ADVFN. **Portal de investimentos em ações da bolsa de valores do Brasil, com cotações da Bovespa e BM&F**. Disponível em: https://br.advfn.com/. Acessado em: 31 jan. 2022.

AMARAL, Lilian. **8 indicadores econômicos para monitorar e usar no planejamento orçamentário**. [2018]. Disponível em: https://www.treasy.com.br/blog/indicadores-economicos/. Acesso em: 20 ago. 2021.

ANFAVEA - ASSOCIAÇÃO NACIONAL DOS FABRICANTES DE VEÍCULOS AUTOMOTORES. **Anuário**. São Paulo: ANFAVEA, 2020.

ANFAVEA - ASSOCIAÇÃO NACIONAL DOS FABRICANTES DE VEÍCULOS AUTOMOTORES. Produção de automóveis tem nova queda pela falta de semicondutores. Estoques são os menores das últimas duas décadas. São Paulo: ANFAVEA, 2021.

BERTO, Rosa Maria Villares.; NAKANO, Davi Noboru. Um Levantamento de Métodos e Tipos de Pesquisa. Produção Científica nos Anais do Encontro Nacional de Engenharia de Produção. **Produção**, v. 9, n. 2, p. 65-76, 2000.

BOSQUEEROLLI, Arthur Marti; FUJARRA, Bruno Henrique; KESSEY, Getúlio Antônio Brandalise Rodrigues; Colaço, HENRIQUE, Malicheski; OLIVEIRA, Henrique Vinícius de; SANTOS, Laura Carvalho Gomes dos; SARRES, Lucas Silva; ALENCASTRO, Matheus Fiuza de; TAO, Matheus Itiro de Castro; VIEIRA, Natalia Podbevsek; NIRO, Raul de Carvalho. **Brasil e o mundo diante da Covid-19 e da crise econômica**. Curitiba: Universidade Federal do Paraná, 2020.

BRUCE, Peter; BRUCE, Andrew. **Practical Statistics for Data Scientists**. New York, O'Reilly, 2017.

CID, Allend Hector. **Machine Learning**: Catalisador da ciência. Chile: Pontificia Universidad Católica de Valparaíso, 2019.

CIHI – CANADIAN INSTITUTE FOR HEALTH INFORMATION. **Health indicators**. Disponível em: https://www.cihi.ca/en/health-indicators. Acesso em: 19 dez. 2021.

CILO, Nelson. **COVID-19**: Setor automobilístico vai levar três anos para se recuperar. [2020]. Disponível em: https://www.em.com.br/app/noticia/economia/2020/04/27/internas_economia,1142056/Covid -19-setor-automobilistico-vai-levar-tres-anos-para-se-recuperar.shtml. Acesso em: 02 ago.2021.

COELHO, Carolina Gomes; PILECCO, Flávia Bulegon. **Indicadores de saúde e testagem para Covid-19**. Salvador: Universidade Federal da Bahia, 2020.

DANTAS, Elias F. **Redes Neurais Artificias Aplicadas à Previsão de Surtos de Leptospirose**. 2018. TCC (Bacharelado em Sistemas de Informação) — Universidade Federal Rural de Pernambuco, Serra Talhada, 2018.

DMITRIEVSKY, Maxim. **Floresta de decisão aleatória na aprendizagem por reforço**. [2018]. Disponível em: https://www.mql5.com/pt/articles/3856. Acesso em: 20 de ago. de 2021.

FENABRAVE. **Anuário 2020**: O desempenho da Distribuição Automotiva no Brasil. São Paulo: FENABRAVE, 2020. Disponível em: http://www.fenabrave.org.br/anuarios/Anuario2013.pdf. Acessado em: 31 jan. 2022.

FERNANDES, Fernando Timoteo; OLIVEIRA, Tiago Almeida de; TEIXEIRA, Cristiane Esteves; BATISTA, Andre Filipe de Moraes; COSTA, Gabriel Dalla; CHIAVEGATTO FILHO, Alexandre Dias Porto. A multipurpose machine learning approach to predict COVID-19 negative prognosis. **Nature**, 2021. DOI 10.1038/s41598-021-82885-y. Disponível em: https://www.nature.com/articles/s41598-021-82885-y#citeas. Acesso em: 30 set. 2021.

FGV. **Portal FGV - índices**. Disponível em: https://portalibre.fgv.br/?utm_source=portal-fgv&utm_medium=menu-indices&utm_campaign=portal-fgv-menu-indices. Acessado em: 31 jan. 2022.

GÉRON, Aurélien. **Hands-On Machine Learning With Scikit-Learn and TensorFlow**. New York: O'Reilly, 2017.

HARRISON, Matt. **Machine Learning Pocket Reference: Working with structured Data in Python**. New York: O'Reilly, 2019.

HE, Chen; HU, Hanbin; LI, Peng. Applications for Machine Learning in Semiconductor Manufacturing and Test. *In*: IEEE ELECTRON DEVICES TECHNOLOGY & MANUFACTURING CONFERENCE (EDTM), 5., 2021. **Proceedings** [...]. USA/CANADA: IEEE, 2021. DOI 10.1109/EDTM50988.2021.9420935. Disponível em: https://ieeexplore.ieee.org/document/9420935. Acesso em: 30 set. 2021.

IBGE - INSTITURO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Pesquisa Industrial Mensal – Produção Física – Brasil**. Rio de Janeiro: IBGE, 2020.

IPEA – INSTITUO DE PESQUISA ECONÔMICA APLICADA. **Carta de Conjuntura – Impactos da pandemia sobre os resultados recentes das contas públicas**. Brasil: IPEA, 2021.

JAIN, Deepak. **Machine Learning, R-Squared in Regression Analysis**. [2019]. Disponível em: https://www.geeksforgeeks.org/ml-r-squared-in-regression-analysis/. Acesso em: 23 ago. 2021.

JUSBRASIL, **Salário Mínimo**. Disponível em: https://www.jusbrasil.com.br/busca?q=sal%C3%A1rio+m%C3%ADnimo. Acesso em: 23 ago. 2021.

KIRSCH, Daniel; HURWITZ, Judith. **Machine learning for dummies**. Hoboken: IBM, 2018.

LACERDA, Rogério Tadeu de Oliveira; ENSSLIN, Leonardo; ENSSLIN, Sandra Rolim. Uma análise bibliométrica da literatura sobre estratégia e avaliação de desempenho. **Gestão & Produção**, v. 19, n. 1, p. 59-78, 2012.

LACONSKI, Alisson. **Índices econômicos, para que servem e outros**. [2019]. Disponível em: https://www.guiabanco.com.br/indicadores-economicos.html. Acesso em: 20 ago. 2021.

LEITE, Vitor. **O que é balança comercial e o que impacta esse resultado?** [2020]. Disponível em: https://blog.nubank.com.br/balanca-comercial-o-que-e/. Acesso em: 23 ago. 2021.

LEMOS, Arthur D. **Quais são os indicadores econômicos mais importantes**? [2019]. Disponível em: https://empreenderdinheiro.com.br/blog/indicadores-economicos/. Acesso em: 20 ago. 2021.

LIBERT, Barry; BECK, Megan. **The machine learning race is really a data race**. [2018]. Disponível em: https://sloanreview.mit.edu/article/the-machine-learning-race-is-really-a-data-race/. Acesso em: 02 ago. 2021.

LORENZI, Larissa. **Ciência de dados x análise de dados:** quais as diferenças? [2021]. Disponível em: https://blog.indicium.tech/diferencas-entre-ciencia-de-dados-e-analise-de-dados/. Acesso em: 20 ago. 2021.

MARR, Bernard. **20 fatos sobre a internet que você não sabe.** [2015]. Disponível em: https://forbes.com.br/fotos/2015/10/20-fatos-sobre-a-internet-que-voce-provavelmente-nao-sabe. Acesso em: 02 ago. 2021.

MARTINS, Roberto Antônio. **Guia para elaboração de monografia e TCC em Engenharia de produção**. São Paulo: Editora Atlas, 2014.

MINISTÉRIO DA ECONOMIA. Gabinete do Ministro**. Portaria 20.809, de 14 de setembro de 2020**. Lista os setores da economia mais impactados pela pandemia após a decretação da calamidade pública decorrente do Covid-19. 2020. Disponível em: https://www.in.gov.br/en/web/dou/-/portaria-n-20.809-de-14-de-setembro-de-2020-277430324. Acesso em: 31 jan. 2022.

NOGUEIRA, Hugo Clapton. **Indicadores econômicos**: a definição e o uso do índice de movimentação econômica. 2012. Monografia (Graduação) — Universidade Estadual do Sudoeste da Bahia, Vitoria da Conquista, 2012.

NOGUERIA, Olivia. **90% dos dados no mundo foram criados apenas nos dois últimos anos**. [2019]. Disponível em: https://www.jornalopcao.com.br/ultimas-noticias/90-dos-dados-no-mundo-foram-criados-apenas-nos-dois-ultimos-anos-diz-google-207812/. Acesso em: 02 ago. 2021.

PAOLANTI, Marina; MANCINI, Adriano; ROMEO, Luca; FRONTONI, Emanuele. Machine learning approach for predictive maintenance in industry 4.0. *In*: IEEE/ASME INTERNATIONAL CONFERENCE ON MECHTRONIC AND EMBEDDED SYSTEMS AND APPLICATIONS, 14., 2018, Oulu. **Proceedings** [...]. Oulu: IEEE, 2018. DOI 10.1109/MESA.2018.8449150. Disponível em: https://ieeexplore.ieee.org/document/8449150. Acesso em: 20 ago. 2021.

PENN, Malcom. **Why Didi the Covid-19 pandemic lead to semiconductor shortages**? Disponível em: https://newseu.cgtn.com/news/2021-09-05/Why-did-the-COVID-19-pandemic-lead-to-semiconductor-shortages--13gl7OiynEA/index.html. Acesso em: 19 dez.2021.

PENUMURU, Durga Prasad; MUTHUSWAMY, Sreekumar; KARUMBU, Premkumar. Identification and classification of materials using machine vision and machine learning in the context of industry 4.0. **Journal of Intelligent Manufacturing,** n. 31, p. 1229-1242, 2020.

PEREIRA, Eduardo Lacerda. **Aplicação de um modelo de aprendizado de máquina para previsão do desgaste de fresas de topo esférico.** 2020. TCC (Graduação em Engenharia da Produção) — Universidade Federal de Santa Catarina, Florianópolis, 2020.

PERES, Ricardo Silva; BARATA, Jose; LEITAO, Paulo; GARCIA, Gisele. Multistage Quality Control Using Machine Learning in the Automotive Industry. **IEEE Access**, v. 7, 2019. DOI 10.1109/ACCESS.2019.2923405. Disponível em: https://ieeexplore.ieee.org/document/8737933. Acesso em: 20 ago. 2021.

RIAHI, Youssra; RIAHI, Sara. Big Data Analytics: Concepts, Types and Technologies. **International Journal of Research and Engineering**, v. 5, n. 9, p. 524-528, 2018. DOI 10.21276/ijre.2018.5.9.5. Disponível em: https://www.researchgate.net/publication/328783489_Big_Data_and_Big_Data_Analytics_Concepts_Types_and_Technologies. Acesso em: 20 ago. 2021.

ROCHA, Ricardo. R vs Python: **Uma análise desapaixonada**. [2020]. Disponível em; https://www.flai.com.br/ricardo/r-vs-python-uma-analise-desapaixonada/. Acesso em: 19 ago. 2021.

RODRIGUES, Gleyson. **O que são indicadores econômicos? Para que servem?** [2021]. Disponível em: https://guiabancario.com.br/o-que-sao-indicadores-economicos/. Acesso em: 20 ago. 2021.

ROZANEC, Joze M; KAZIC, Blaz; SKRJANC, Maja; FORTUNA, Blaz; MLADENIC, Dunja. Automotive OEM Demand Forecasting: A comparative Study of Forecasring Algoriths and Strategies. **Applied Sciences,** v. 11, n. 15, 2021. DOI 10.3390/app11156787. Disponível em: https://www.mdpi.com/2076-3417/11/15/6787. Acesso em: 20 ago. 2021.

SANTOS, Valter Barbosa. **Estimação e previsão de produtividade de soja por redes neurais no MATOPIBA**. 2020. Dissertação (Mestrado em Agronomia) — Universidade Estadual Paulista, Jaboticabal, 2020.

SEYMOUR, Geisser, **Predictive Inference**: an introduction. New York: Chapman & Hall, 2016.

SIERRA, Jaime Barco. **Como o machine learning é afetado pela Covid-19:** Análise de dados e modelos preditivos em um cenário atípico. Belo Horizonte: Tatic, 2021.

SILVA, Ivan Nunes; SPATTI, Danilo Hernane.; FLAUZINO, Rogério Andrade. **Redes neurais artificias para Engenharia e Ciências Aplicadas**: fundamentos teóricos e aspectos práticos. 6. ed. São Paulo: Artliber, 2016.

SILVEIRA, Ian Vieira. **Modelo de previsão de demanda com o uso de aprendizado supervisionado de máquina**: um estudo de caso em uma empresa de varejo. 2019. TCC (Engenharia de Produção) — Universidade Federal de Santa Catarina, Florianópolis, 2019.

SJOBERG, Mark Ludwikowskie Willian. **Semiconductor shortage and the U.S. auto industry**. [2021]. Disponivel em: https://www.reuters.com/legal/legalindustry/semiconductor-shortage-us-auto-industry-2021-06-22/. Acesso em: 19 dez. 2021.

SOUZA, Alex. **Seu primeiro projeto de Machine Learning em Python**. [2019]. Disponível em: https://medium.com/blog-do-zouza/seu-primeiro-projeto-de-machine-learning-em-python-passo-a-passo-78c5f7bce22d. Acesso em: 23 ago. 2021.

STUMPF, Kleber. **Indicadores Econômicos.** [2019]. Disponível em: https://www.topinvest.com.br/indicadores-economicos/. Acesso em: 20 ago. 2021.

TEIXEIRA, Daniel. **Inteligência artificial aplicada à pesquisa de mercado e comunicação**. Monografia (Especialização em Pesquisa de Mercado Aplicada em Comunicações) — Universidade de São Paulo, São Paulo, 2019.

TIWARI, Upendra Kumar; KHAN, Rijwan. Role of machine learning to predict the outbreak of Covid-19 in India. **Journal of Xi'na University of Architecture & Technology**, Xi'na, n. 12, p. 2663-2669, 2020.

VERGARA, S. **Projetos e Relatórios de Pesquisa em Administração**. São Paulo: Atlas, 2000.

WEHLE, Hans Dieter. **Machine Learning, Deep Learning, and AI: What's the difference?** [2017]. Disponível em: https://www.researchgate.net/publication/328783489_Big_Data_and_Big_Data_Analytics_Concepts_Types_and_Technologies. Acesso em: 31 jan. 2022.

ZOABI, Yazeed; ROZOV, Shira; SHOMRON, Noam. Machine learning based prediction of COVID-19 diagnosis based on symptoms. **Nature**, 2021. DOI 10.1038/s41746-020-00372-6. Disponível em: https://www.nature.com/articles/s41746-020-00372-6. Acesso em: 02 ago. 2021.