UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CAMPUS DE DOIS VIZINHOS
CURSO DE ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS

HENRIQUE MARQUES TURQUETI

MODELOS DE APRENDIZADO DE MÁQUINA PARA PREVISÃO DO PREÇO DO ÓLEO DIESEL NA REGIÃO SUDESTE DO BRASIL

TRABALHO DE CONCLUSÃO DE CURSO DE ESPECIALIZAÇÃO

DOIS VIZINHOS
2022

HENRIQUE MARQUES TURQUETI

MODELOS DE APRENDIZADO DE MÁQUINA PARA PREVISÃO DO PREÇO DO ÓLEO DIESEL NA REGIÃO SUDESTE DO BRASIL

MACHINE LEARNING MODELS FOR DIESEL OIL PRICE PREDICTION IN SOUTHEASTERN BRAZIL

Trabalho de Conclusão de Curso de Especialização apresentado ao Curso de Especialização em Ciência de Dados da Universidade Tecnológica Federal do Paraná, como requisito para a obtenção do título de Especialista em Ciência de Dados.

Orientador: Prof. Dr. Rafael Gomes Mantovani
Coorientador: Prof. Dr. Francisco Carlos Monteiro Souza

DOIS VIZINHOS
2022

HENRIQUE MARQUES TURQUETI

MODELOS DE APRENDIZADO DE MÁQUINA PARA PREVISÃO DO PREÇO DO ÓLEO DIESEL NA REGIÃO SUDESTE DO BRASIL

Trabalho de Conclusão de Curso de Especialização apresentado ao Curso de Especialização em Ciência de Dados da Universidade Tecnológica Federal do Paraná, como requisito para a obtenção do título de Especialista em Ciência de Dados.

Data de aprovação: 09/novembro/2022

Rafael Gomes Mantovani
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Apucarana

Jefferson Tales Oliva
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Pato Branco

Luiz Fernando Carvalho
Doutorado
Universidade Tecnológica Federal do Paraná - Câmpus Apucarana

DOIS VIZINHOS
2022

AGRADECIMENTOS

Agradeço primeiramente à Nathalia, minha parceira há quase 10 anos, pela companhia nos momentos que mais precisei e pelos conselhos que foram fundamentais para minha evolução como pessoa.

Também a meus pais, avós e outros familiares que de alguma forma me deram suporte desde a minha infância até agora.

Ao Jake, meu filho de quatro patas que tive a oportunidade de ter comigo esse ano e que me acompanhou durante muitas manhãs e noites de estudos.

A todos os funcionários da UTFPR, especialmente aos meus professores orientadores Rafael Mantovani e Francisco Carlos, por colaborarem com a disseminação do conhecimento em nossa sociedade por meio do ensino público de qualidade.

Por fim, agradeço aos meus amigos e colegas de profissão pela troca de informações e experiências proporcionada nesses últimos anos.

Por vezes sentimos que aquilo que fazemos não é senão uma gota de água no mar. Mas o mar seria menor se lhe faltasse uma gota. (Madre Teresa de Calcutá).

RESUMO

Diversos fenômenos quantitativos de interesse econômico variam ao longo do tempo, podendo ser representados por meio de séries temporais, como no caso variação de preço de ações, combustíveis e criptomoedas. Na ciência de dados é comumente utilizado como técnica para prever seus valores futuros as redes neurais recorrentes, como aquelas com células LSTM (*Long Short Term Memory*), e outros modelos estatísticos clássicos, como ARIMA (*Autoregressive Integrated Moving Average*). Dessa forma, o presente trabalho propõe uma análise exploratória prévia do histórico de preços de combustíveis divulgada pelo governo agrupadas por região administrativa, semana e tipo de combustível. A análise exploratória mostra que o óleo diesel possui uma distribuição de preços semelhante nas cinco diferentes regiões e com uma menor volatilidade do que os demais combustíveis. Além disso, ele é responsável por quase metade do volume vendido no ano de 2021. Por ser a região com maior quantidade vendida nesse mesmo ano, o sudeste foi escolhido para ter os dados utilizados no treinamento e validação dos modelos preditivos. Após a escolha do ARIMA como algoritmo base dos modelos clássicos de aprendizado de máquina, ele foi otimizado com e sem variáveis exógenas (preço do barril de petróleo *Brent* e cotação do dólar em reais) mas apresentou previsões lineares e não condizentes com a volatilidade do histórico de preços. Já o modelo de LSTM otimizado por meio de busca aleatória de parâmetros com validação cruzada obteve erros em um período de teste de 8 semanas iguais a $MSE = 0{,}2908$, $RMSE = 0{,}5393$ e $MAE = 0{,}4568$. Apesar de eles serem superiores aos do ARIMA ($MSE = 0{,}1570$, $RMSE = 0{,}3962$ e $MAE = 0{,}3527$), o modelo de redes neurais recorrentes com LSTM se adaptou à dinâmica de preços, fornecendo resultados em sequências não-lineares, o que condiz com o problema estudado.

Palavras-chave: óleo diesel, combustíveis, aprendizado de máquina, LSTM, ARIMA.

ABSTRACT

Several quantitative phenomena of economic interest vary over time, and can be represented through time series, as in the case of changes in the price of stocks, fuels and cryptocurrencies. In data science, recurrent neural networks, such as those with LSTM cells, and other classical statistical models, such as ARIMA, are commonly used as a technique to predict their future values. In this way, the present work proposes a preliminary exploratory analysis of the history of fuel prices published by the government grouped by administrative region, week, and type of fuel. The exploratory analysis shows that diesel oil has a similar price distribution in the five different regions and with lower volatility than other fuels. In addition, it is responsible for almost half of the volume sold in 2021. As it is the region with the highest volume sold in that same year, the Southeast was chosen to have the data used in the training and validation of predictive models. After choosing ARIMA as the base algorithm of classical machine learning models, it was optimized with and without exogenous variables (Brent barrel price of oil and dollar exchange rate in reais) but presented linear predictions and not consistent with the volatility of the price history. The LSTM model optimized through random parameter search with cross-validation had errors in an 8-week test period equal to $MSE = 0.2908$, $RMSE = 0.5393$ and $MAE = 0.4568$. Although they are superior to those of ARIMA ($MSE = 0.1570$, $RMSE = 0.3962$ and $MAE = 0.3527$), the recurrent neural networks model with LSTM adapted to the price dynamics, providing results in non-linear sequences, which is consistent with the problem studied.

Keywords: diesel oil, fuels, machine learning, LSTM, ARIMA.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANP	Agência Nacional do Petróleo, Gás Natural e Biocombustíveis
API	*American Petroleum Institute*
ARIMA	*Autoregressive Integrated Moving Average*
LSTM	*Long Short Term Memory*
MAE	*Mean Absolute Error*
MAPE	*Mean Absolute Percentage Error*
MSE	*Mean Squared Error*
MLSTM	*Multivariate Long Short Term Memory*
PPI	Preços de Paridade de Importação
RNA	Redes Neurais Artificiais
RMSE	*Root Mean Squared Error*
SARIMAX	*Seasonal Autoregressive Integrated Moving Average with eXogenous factors*
TKU	Tonelada-Quilômetro Útil

SUMÁRIO

1 INTRODUÇÃO

Especificamente em 2019, cerca de 61% do transporte de cargas dentro do Brasil foi feito via rodovias, considerando-se os TKUs (tonelada-quilômetro útil). Esse valor chega a ser de duas a três vezes maior do que a quantidade realizada por outros países como Austrália e Canadá, respectivamente (ALVARENGA, 2020). Devido a essa dependência do modal rodoviário, a mudança de preço nos combustíveis derivados do petróleo possui um grande impacto no preço final de diversos produtos ao consumidor. Segundo ALMEIDA e ARAÚJO (2022), um exemplo dessa influência foi o repasse de aproximadamente 10% no preço dos alimentos em maio de 2022 devido ao encarecimento dos fretes pelo aumento do preço dos combustíveis.

No Brasil, o preço de revenda dos combustíveis é fortemente influenciado por fatores externos como taxa de juros americanos, maior entrada de turistas no país (LAFRATTA, 2020) e eventos que levem a alteração de demanda como crises econômicas e paralisação de extrações por eventos climáticos extremos (SOMMA Investimentos, 2021). Isso ocorre devido a este valor estar vinculado ao preço do barril tipo *Brent*, cujo valor é atrelado à cotação do dólar. Esse método de precificação foi iniciado em 2016 com a chamada política de Preços de Paridade de Importação (PPI) (VILELA, 2022). Segundo BBC News Brasil (2021), a PPI foi uma resposta da Petrobras à política de controle de preços dos combustíveis que estava em vigor durante o governo de Dilma Rousseff e a consequente deterioração contábil da empresa.

Segundo a Agência O Globo (2021), desde o início de 2021, tanto a gasolina como o óleo diesel tiveram um aumento de aproximadamente 50% de seu valor, e a previsão é que a Petrobras proponha novos aumentos. Em outras palavras, na prática, estes valores podem dobrar ou até quadruplicar a partir de 1º de janeiro de 2022. Com a frequente mudança na política de reajustes do custo dos combustíveis, brasileiros que moram próximo à divisa com outros países, como a Argentina, preferem cruzar a fronteira e abastecer os veículos por lá, onde os preços praticados são em média a metade daqueles executados no país (FOLHAPRESS, 2021).

Com a instabilidade do valor do dólar americano no mercado brasileiro e, com a grande variação do preço do barril *Brent* comercializado internacionalmente, atualmente não é possível que empresas de logística e transporte consigam precificar adequadamente seus serviços em janelas de tempo superiores a poucas semanas. Dessa forma, o presente trabalho visa a construção de um modelo para previsão do preço de revenda dos diferentes tipos de combustíveis dentro do país para ser possível também antever custos de empresas de transporte e logística, assim como o impacto em demais índices da econômica, como preço de alimentos.

1.1 Objetivos

Sendo assim, o objetivo principal deste trabalho é desenvolver um modelo preditivo para o preço de revenda de óleo diesel com a dinâmica de reajustes na região Sudeste do Brasil em um horizonte de curto prazo.

Especificamente, deseja-se:

- Definir uma fonte de dados aberta e mantida por uma entidade governamental para o treinamento e validação do modelo;
- Selecionar critérios de avaliação do algoritmo que possibilitem interromper seu treinamento e otimização apenas quando ela estiver apta a realizar as previsões de forma adequada;
- Avaliar os resultados gerados por meio de comparação com dados reais e outros modelos feitos para os combustíveis;
- Identificar possíveis padrões de aumento e diminuição de preços do óleo diesel.

1.2 Justificativa

O tema de pesquisa em questão se justifica devido à atualidade do assunto na mídia brasileira e seu impacto na população na totalidade, podendo influenciar de forma direta preços de alimentos e outros bens de consumo devido ao encarecimento dos fretes rodoviários feitos por caminhões. Além disso, quando concluído, o projeto pode contribuir com a iniciativa privada para uma melhor previsibilidade dos seus gastos no transporte de cargas pelo país.

1.3 Organização do Trabalho

A estrutura do presente trabalho apresenta-se da seguinte forma: no Capítulo 2 são fundamentados os conhecimentos teóricos necessários para entendimento do tema estudado, da dinâmica dos dados temporais e do uso dos modelos preditivos. No Capítulo 3 são apresentados os resultados de estudos semelhantes divulgados por outros autores, analisando-se seus pontos fortes e lacunas. Já no Capítulo 4 estão presentes as informações tanto das bibliotecas utilizadas quanto os métodos seguidos para a obtenção dos resultados. A análise exploratória de dados e os resultados obtidos dos modelos preditivos são apresentados no Capítulo 5. Por fim, são apresentados uma conclusão do estudo com suas limitações encontradas e sugestões para trabalhos futuros no Capítulo 6.

2 REFERENCIAL TEÓRICO

Neste capítulo são apresentados os principais conceitos teóricos necessários para a compreensão e acompanhamento do presente estudo.

2.1 Combustíveis derivados do petróleo

O petróleo, palavra latina que significa óleo de pedra, é uma mistura de hidrocarbonetos de ocorrência natural que possui alguns contaminantes, tais como: o enxofre, oxigênio, metais e outros elementos. O petróleo pode ser encontrado nos estados: sólido, denominado asfalto; semissólido, chamado de betume; líquido, chamado de óleo bruto se for negro e pegajoso, e de condensado se for claro e volátil; e gasoso, referido como gás natural (GAUTO et al., 2016, p. 5).

Além da já conhecida importância do petróleo no fornecimento direto de energia para motores à combustão, devido ao uso de seus combustíveis derivados, ele também é uma matéria-prima essencial à sociedade moderna dada a sua utilização na produção de lubrificantes, plásticos, tecidos sintéticos, tintas, dentre outros (GAUTO; ROSA, 2013, p. 52).

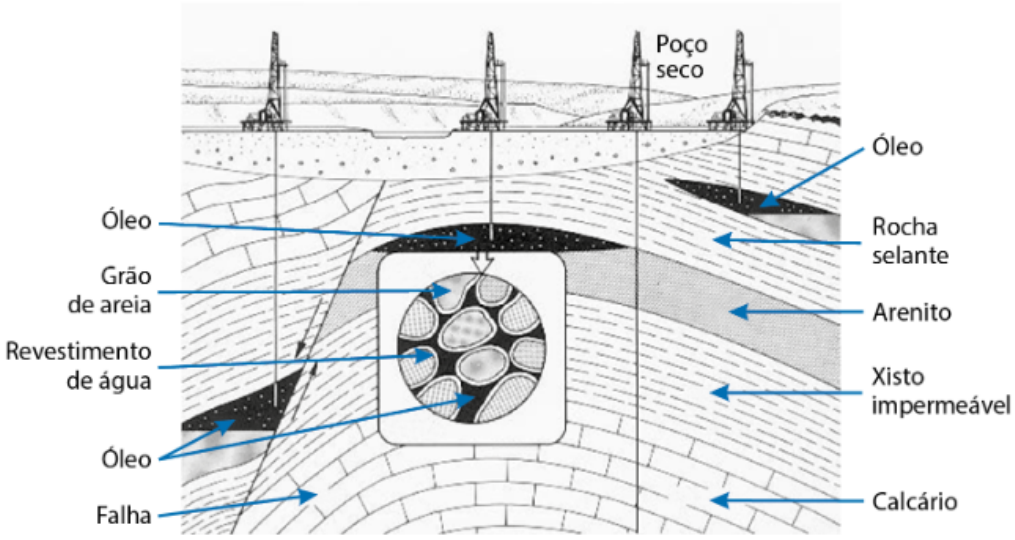
2.1.1 Formação e origem

Segundo Gauto e Rosa (2013, p. 53), o petróleo é formado como resultado da própria ação da natureza, onde o material orgânico de restos de animais e vegetais, depositados há milhões de anos no fundo de lagos e mares, são soterrados por novas camadas, ocasionando um aumento da pressão e calor no local, transformando a matéria orgânica ali existente em petróleo.

Além desses fatores é necessário também que esses elementos estejam em contato com diferentes tipos de rochas, às vezes simultaneamente, como as geradoras (ricas em matérias orgânicas), carreadoras (porosas, permeáveis e capeadas), reservatórios (porosas, permeáveis, capeadas e trapeadas), selantes (baixa permeabilidade) e de sobrecarga (sobrejacentes e que exerçam pressão litostática) (GAUTO et al., 2016, p. 75), conforme é mostrado em um exemplo na Figura 1.

A ausência de pelo menos um desses fatores originadores, ou rochas do sistema petrolífero, eliminou a possibilidade de existência de acúmulo de petróleo em muitas áreas sedimentares pelo mundo. Como consequência, em dois ou três séculos de consumo acelerado o homem possivelmente esgotará os recursos que levaram até 400 milhões de anos para serem produzidos pela natureza (GAUTO; ROSA, 2013, p. 53).

Figura 1 – Estrutura de rochas do sistema petrolífero



Fonte: Gauto et al. (2016, p. 76).

2.1.2 Métodos de exploração e extração

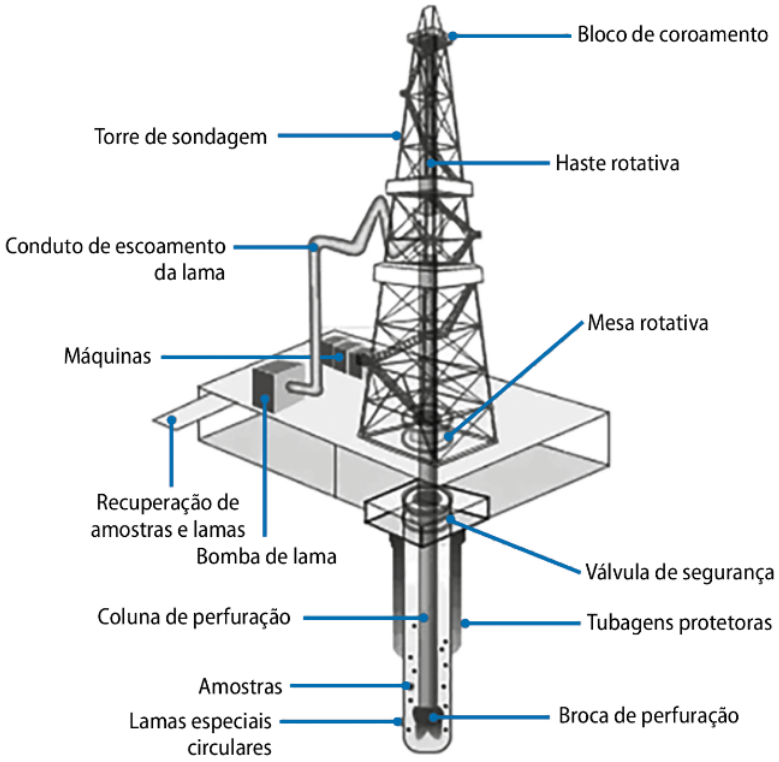
Após a seleção e validação da área de interesse por meio da análise de dados de caráter geográfico (topografia, vegetação, solo, sazonalidade de chuvas, correntes oceânicas e infraestrutura da região), geológico (busca por formações geológicas semelhantes a de outras regiões com petróleo já confirmado) e geofísico (inferência da distribuição de propriedades da subsuperfície utilizando propriedades físicas da superfície), é necessária a perfuração de um poço pioneiro para confirmar ou não a descoberta do petróleo. Um fator crítico nessa análise é que em aproximadamente 80% dos poços pioneiros não são feitas descobertas comercialmente viáveis, ou pela ausência do petróleo, ou devido à presença em pouca quantidade (GAUTO et al., 2016, p. 88-89, 93, 97).

Devido à limitação de profundidade imposta pela sondagem à percussão, usa-se geralmente a sondagem rotativa. A primeira caracteriza-se pela aplicação de golpes sucessivos às rochas, gerando fragmentos e fazendo com que o processo seja pausado frequentemente para retirada deles. Já a última tem a vantagem dos fragmentos serem trazidos automaticamente à superfície, já que é utilizada uma coluna de perfuração rotativa em conjunto com injeção e bombeamento de fluidos de perfuração.

A perfuração de poços pode ser conduzida tanto em terra quanto em mar, sendo que operações em terra tendem a ser muito mais simples. Já a complexidade das operações marítimas aumenta conforme a perfuração se afasta da costa (GAUTO et al., 2016, p. 95).

As sondas de perfuração terrestre, ilustradas na Figura 2, são instaladas no local onde se deseja perfurar sendo fixas. Ela permanece instalada enquanto houver atividade e ao término da operação ela é desativada, tamponada com cimento e abandonada lá, já que os poços ainda podem servir para fornecimento de importantes indicadores para prosseguimento de pesquisas

Figura 2 – Sonda de perfuração terrestre



Fonte: Gauto et al. (2016, p. 97).

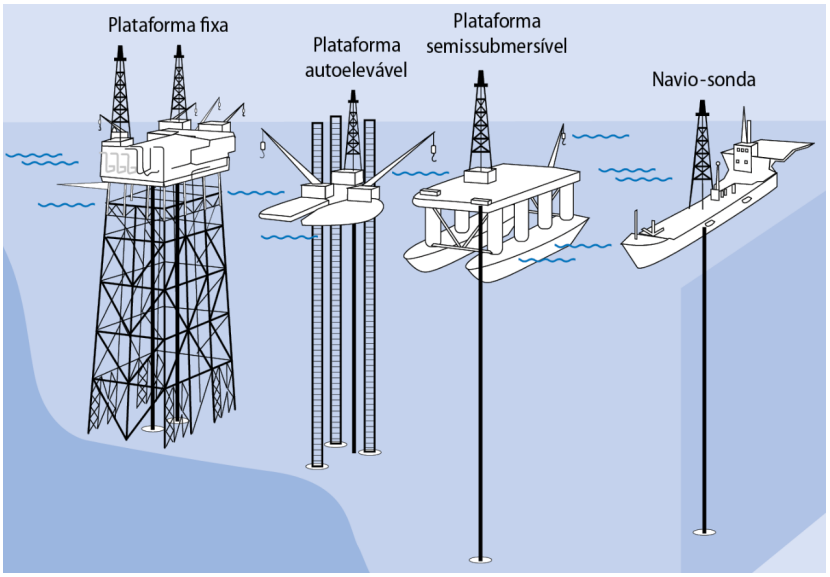
sobre o assunto (GAUTO et al., 2016, p. 97). Já as sondas marinhas, ilustradas na Figura 3, podem ser ou fixas, comportando explorações permanentes ou temporárias, ou móveis, tanto por meio de reboque quanto por meio de locomoção independente (GAUTO et al., 2016, p. 98).

Os navios-sonda são utilizados principalmente em águas profundas e em áreas sob severas condições marítimas. O Brasil é um dos poucos países que domina todo o ciclo de perfuração submarina em águas profundas e ultraprofundas, quando operam em profundidades maiores do que 2000 metros de lâmina d'água (GAUTO; ROSA, 2013, p. 57). Esse fato ficou consolidado depois do êxito na perfuração da Bacia de Campos na década de 70, já que já haviam sido perfurados diversos poços rasos naquela região, mas todos foram dados como secos, tendo também a dificuldade em prosseguir com a perfuração devido à composição das rochas (UDOP, 2021). Devido ao sucesso desse caso e o grande potencial existente em águas profundas e ultraprofundas, esse tipo de exploração representou 95% da produção total de 2021 (PETROBRAS, 2022).

2.1.3 Processos de tratamento e refino

O petróleo bruto extraído das jazidas contém diferentes tipos de hidrocarbonetos e impurezas, necessitando passar por processos de tratamento e refino. O objetivo desse trabalho é separar as diversas frações de interesse por meio de operações unitárias (físicas) e conversões

Figura 3 – Plataformas marítimas para perfuração de poços



Fonte: Gauto et al. (2016, p. 98).

químicas (GAUTO et al., 2016, p. 153).

Conforme explicitado por Gauto e Rosa (2013, p. 63), após a caracterização do petróleo por diversos critérios, como o grau de densidade API (*American Petroleum Institute*), teor de enxofre e razão dos componentes químicos presentes (parafínicos, naftênicos, asfálticos, etc.), ele é submetido ao processo de refino. Estas operações de refino estão resumidas na Figura 4.

Primeiramente o petróleo passa por processos físicos de separação, como destilação e extração a solvente, para obter componentes com diferentes pontos de ebulição e solubilidade. Diferentemente do processo anterior, onde a estrutura química permanece inalterada, nos processos de conversão ocorrem reações químicas com alteração de estrutura molecular para a transformação do produto bruto em outros mais rentáveis. Com o produto de interesse preparado, ele é refinado ainda mais para serem retiradas impurezas por meio de processos químicos ou físicos.

Para que essas três etapas sejam possíveis, as plantas petroquímicas possuem importantes processos auxiliares. Um deles é o de fornecimento de utilidades, como água e vapor d'água, utilizados respectivamente em torres de resfriamento e caldeiras, por exemplo. Existem também o tratamento dos efluentes gerados no próprio processo de refino do petróleo, como a neutralização de solventes por meio de ácidos, bases ou outros componentes específicos antes deles serem descartados na rede pública de tratamento de água e esgoto.

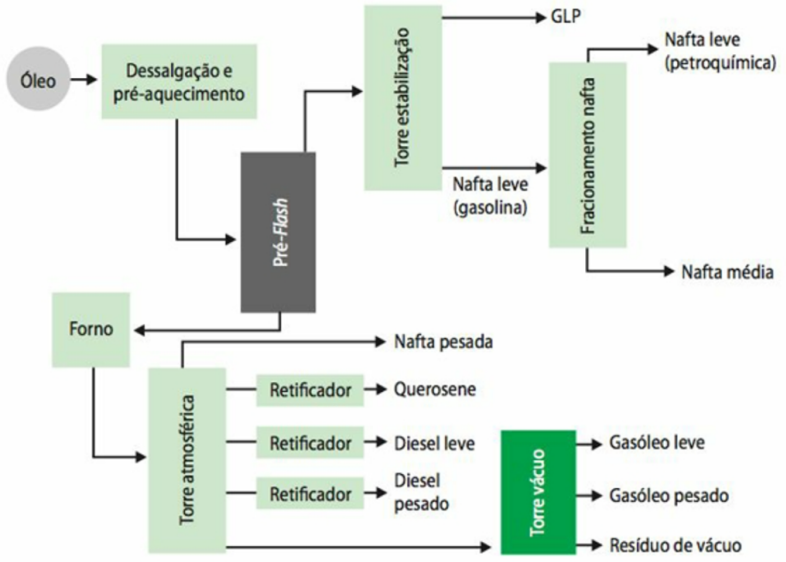
Para ilustrar a separação do produto de interesse desse estudo, ou seja, o óleo diesel, é apresentado na Figura 5 um diagrama de blocos com a saída das principais frações do óleo cru após o processo de destilação.

O óleo cru passa primeiramente por um processo para retirada do sal e pré-aquecimento

Figura 4 – Diagrama com classes e processos para refino do petróleo

Fonte: Autoria própria.

Figura 5 – Diagrama de blocos da destilação do óleo cru



Fonte: Gauto e Rosa (2013, p. 65).

necessário para a próxima etapa. Ao entrar no tanque de *flash*, o conteúdo é separado devido à redução de pressão (Gestra, 2022) em dois componentes: o mais volátil transforma-se em vapor na corrente que segue para a torre de estabilização e o menos volátil segue no estado líquido (condensado) na corrente com destino ao forno. Após o aquecimento no forno, o conteúdo passa por uma torre de destilação atmosférica com diversos pratos para separar os compostos pelos seus diferentes pontos de ebulição, sendo que os mais voláteis são liberados nas partes mais altas da torre.

2.2 Séries temporais

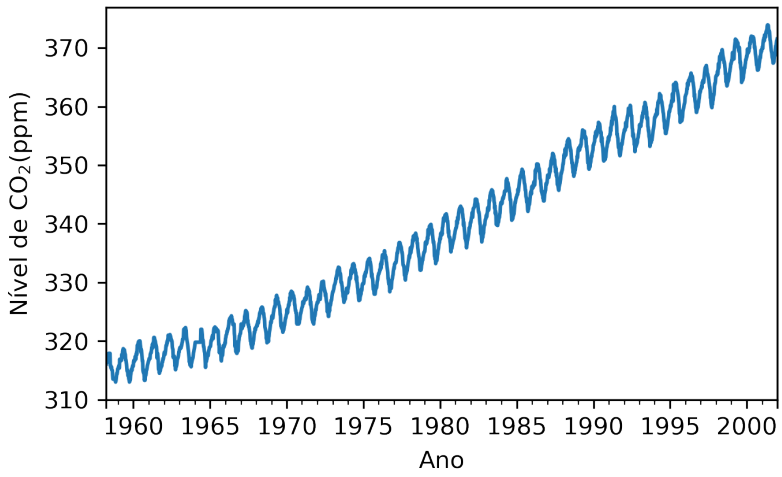
Segundo Sousa et al. (2021, p. 11), as séries temporais são conjuntos de observações ordenadas no tempo, e que podem ser úteis para descrever e acompanhar fenômenos em diversas áreas da ciência. A investigação delas tem como pretensão realizar previsões para períodos futuros. Atualmente empresas de diversos segmentos utilizam dados de séries temporais visando monitorar processos, detectar instabilidades, descobrir tendências e sazonalidades, e obter relação de causa e efeito para, por exemplo, descobrir fatores que podem levar ao aumento das vendas de um determinado produto (SOUSA et al., 2021, p. 12).

2.2.1 Decomposição aditiva e multiplicativa

Durante a análise exploratória de dados, etapa anterior à construção de um modelo de previsão, uma das principais tarefas ao se trabalhar com séries temporais é decompô-las em tendência (T), sazonalidade (S) e resíduos (R) (ATWAN, 2022, p. 300). Segundo Atwan (2022, p. 300), a tendência é responsável por dar a direção de longo prazo para a série, indicando se ela terá um acréscimo ou decréscimo de seus valores, ou se eles permanecerão constantes. Já a sazonalidade é representada por repetidos padrões ao longo do tempo, como o aumento de vendas de um produto próximo à data do Natal ao longo dos anos. Por fim, os resíduos são as partes da série que não podem ser explicadas após a extração dos dois primeiros componentes.

Quando o componente sazonal de uma série temporal não se altera ao longo do tempo, situação exemplificada na Figura 6, é um indicativo de que essa é uma série que pode ser decomposta pelo método aditivo (ou seja, somando-se os dados dos componentes de tendência, sazonalidade e resíduos) (ATWAN, 2022, p. 300). Na figura, a série temporal representa dados do nível de CO_2 em partes por milhão (ppm) fornecidos pela biblioteca `statsmodel` de Seabold e Perktold (2010).

Figura 6 – Série temporal de modelo aditivo

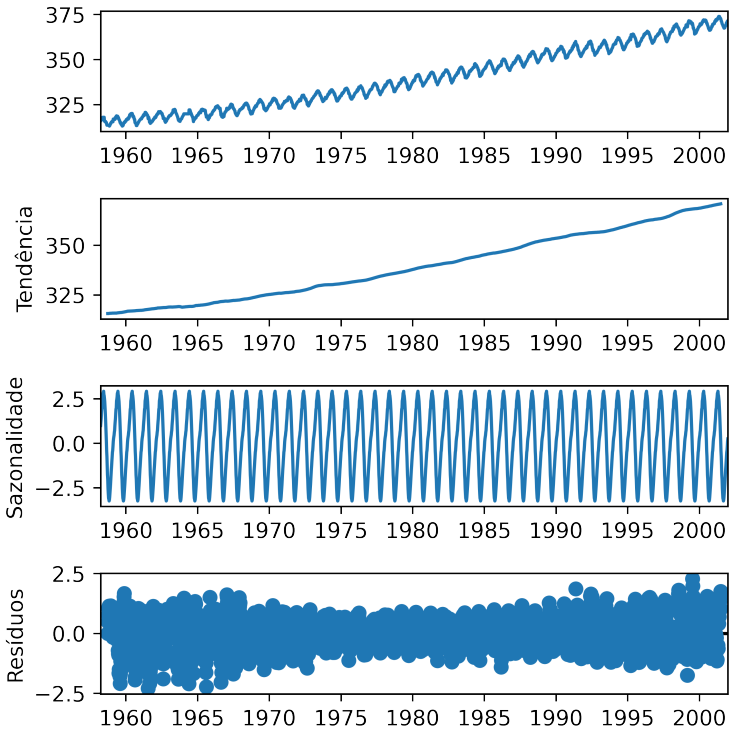


Fonte: Autoria própria.

A Equação (1) mostra que após a decomposição de uma série temporal (Figura 7), nos componentes de tendência, sazonalidade e resíduos, a série pode ser reconstruída adicionando-se estes três componentes novamente (ATWAN, 2022, p. 300).

$$y_t = T_t + S_t + R_t \tag{1}$$

Figura 7 – Decomposição de série temporal de modelo aditivo



Fonte: Autoria própria.

Já quando há uma alteração das variações da sazonalidade de uma série temporal, como são mostrados nos dados de embarques de uma companhia aérea Seabold e Perktold (2010), presentes na Figura 8, ela pode ser decomposta pelo método multiplicativo e reconstruída multiplicando-se os três componentes da Equação (2) (ATWAN, 2022, p. 300).

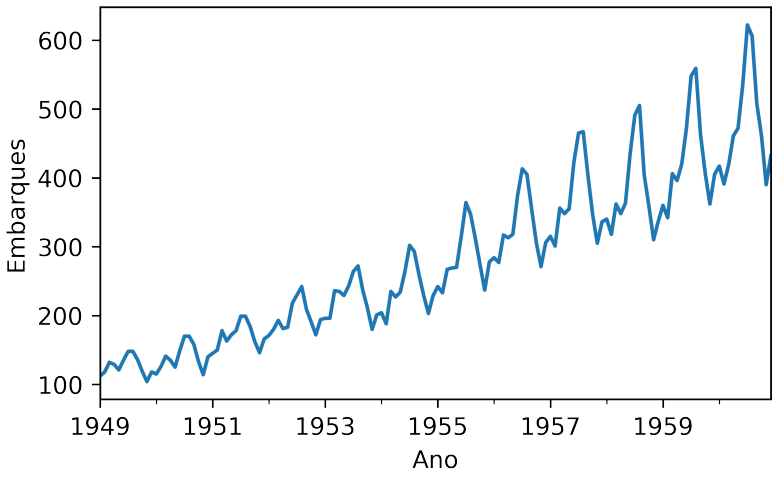
$$y_t = T_t \times S_t \times R_t \tag{2}$$

A decomposição da série temporal da Figura 8 em seus três componentes principais é ilustrada na Figura 9 a seguir.

2.2.2 Estacionariedade

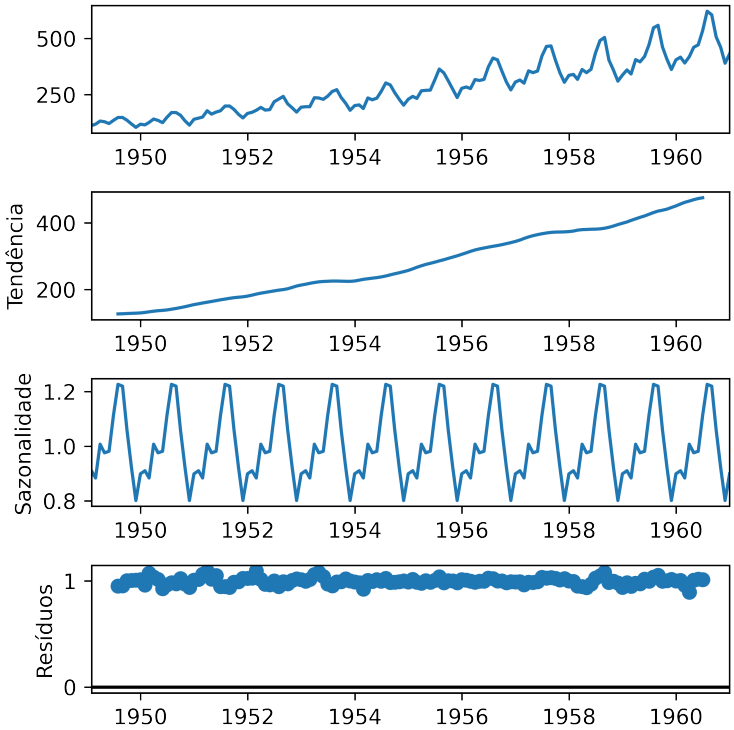
Diversos modelos de previsão de séries temporais assumem que as séries são estacionárias. Logo, esta é uma característica muito importante a ser verificada durante a análise exploratória. A estacionariedade é uma característica que implica na não mudança de certas

Figura 8 – Série temporal de modelo multiplicativo



Fonte: Autoria própria.

Figura 9 – Decomposição de série temporal de modelo multiplicativo



Fonte: Autoria própria.

propriedades ao longo do tempo, ou seja, a série é *estacionária*. Para efeitos práticos, uma série temporal pode ser considerada estacionária se a sua média (μ), variância (σ^2) e covariância (ou autocorrelação) entre períodos igualmente espaçados permanecem constantes no período analisado (ATWAN, 2022, p. 308).

Segundo Atwan (2022, p. 308), apesar de ser possível identificar a não-estacionariedade

de uma série temporal apenas com o uso de um gráfico de linha ou com a presença de um componente sazonal e de tendência, uma opção disponível também é identificar de forma numérica utilizando testes estatísticos de hipóteses, como o teste ADF (*Augmented Dickey-Fuller*) e o teste KPSS (*Kwiatkowski-Phillips-Schmidt-Shin*). Ambos são baseados no modelo de regressão linear e checam a presença de uma raiz unitária em uma série univariada, o que geralmente indica a não-estacionariedade. Conforme explicitado por Woodward, Sadler e Robertson (2022, p. 276), a existência de uma raiz unitária significa que uma ou mais raízes da equação característica autorregressiva da série temporal estão contidas no círculo unitário (ou seja, com um raio igual a 1).

O teste ADF tem como hipótese nula (H^0) que dada uma série temporal, está possui uma raiz unitária e é, portanto, não estacionária. Já o teste KPSS possui uma hipótese nula oposta, assumindo que a série é estacionária (ATWAN, 2022, p. 309). Para considerar uma série temporal estacionária ou não, é necessário comparar o p-valor retornado pelos testes[1] e compará-lo com o nível de significância (α) definido na análise do problema, sendo usualmente utilizado o valor $\alpha = 0.05$ (5%) (WOODWARD; SADLER; ROBERTSON, 2022, p. 339).

Contudo, segundo Nielsen (2019, p. 85), utilizar apenas o teste ADF para confirmar a estacionariedade de uma série temporal pode levar a algumas interpretações errôneas. Isto ocorre devido ao seu baixo poder estatístico em distinguir situações próximas à raiz unitária de fato, e aos falsos positivos comumente encontrados ao se utilizar uma amostra de tamanho insuficiente.

Para modelos onde é necessária estacionariedade da série temporal, Atwan (2022, p. 313) demonstra alguns métodos para torná-la estacionária:

- **Diferenciação de primeira ordem:** também conhecido em inglês como *detrending*, esse método busca subtrair de uma observação no tempo t o valor no tempo $t-1$ (ou seja, $y_t - y_{t-1}$);
- **Diferenciação de segunda ordem:** nos casos em que apenas a diferenciação de primeira ordem não tenha funcionado, ou que a série temporal possua também sazonalidade, é possível diferenciá-la duas vezes. Isto é feito fazendo primeiramente a diferenciação dos valores no período sazonal ($y_t - y_{t-12}$ no caso de uma base de dados mensal com sazonalidade anual, por exemplo) e depois a diferenciação de primeira ordem (valor atual menos anterior);
- **Subtração da média móvel (janela deslizante):** no caso de uma base de dados com granularidade mensal e uma sazonalidade anual, seria necessário utilizar uma janela deslizante igual a 52 semanas (um ano) e subtraí-la dos dados originais para que assim seja eliminado o seu componente sazonal;
- **Transformação logarítmica:** em alguns casos simplesmente calcular o logaritmo de todos os valores da série já é suficiente para estabilizar a variância e torná-la estacionária;
- **Remoção do componente de tendência da série:** conforme mostrado no item 2.2.1, é

[1] ambos disponíveis no pacote Python chamado statsmodels de Seabold e Perktold (2010)

possível decompor uma série em seus componentes de tendência, sazonalidade e resíduos. Nessa técnica basta subtrair dos dados originais a componente de tendência caso a decomposição seja do tipo aditiva.

O conceito de janela deslizante apresentado no item anterior também pode ser utilizado durante a validação cruzada dos resultados do modelo. Ao se definir um tamanho de janela igual a 3, por exemplo, em t_0 o valor da média móvel é igual a $\frac{t_0 - t_{-1} - t_{-2}}{3}$ e assim por diante, conforme é mostrado na Figura 10.

Figura 10 – Exemplo de cálculo de média móvel (janela deslizante)

t	Preço (R$)	Média móvel do preço (R$)
1	6,10	
2	6,28	
3	6,16	6,18
4	6,28	6,24
5	6,78	6,41
6	7,39	6,82
7	8,06	7,41
8	8,54	8,00
9	8,80	8,47
10	9,50	8,95

Fonte: Autoria própria.

Assim como ilustrado na Figura 10, definir uma janela com tamanho igual a três impossibilita o seu cálculo nos primeiros dois tempos da série temporal devido à ausência de valores anteriores para realizar a média.

2.2.3 Autocorrelação

Uma característica importante das séries temporais é que elas frequentemente possuem forte correlação entre as observações ao longo do tempo. Um dos métodos disponíveis para quantificar esse fenômeno é a autocorrelação (CIPRA, 2020, p. 125). De acordo com Cipra (2020, p. 125), a função de autocovariância γ_k para um dado atraso de tempo k (chamado também de parâmetro de deslocamento) pode ser definida em função dos valores esperados da observação no tempo t (y_t), da observação no tempo $t - k$ (y_{t-k}) e da média populacional μ, conforme é mostrado na Equação (3).

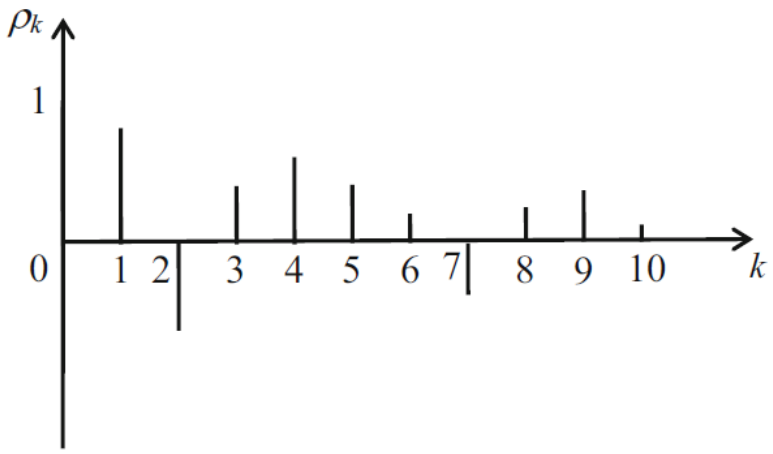
$$\gamma_k = cov(y_t, y_{t-k}) = \mathbf{E}(y_t - \mu)(y_{t-k} - \mu), \qquad k = ..., -1, 0, 1, ... \tag{3}$$

Analogamente, Cipra (2020, p. 125) define a função de autocorrelação ρ_k como a divisão da autocovariância com atraso k, γ_k, pela autocovariância sem atraso, γ_0, segundo a Equação (4).

$$\rho_k = \frac{\gamma_k}{\gamma_0} = \frac{\gamma_k}{\sigma_y^2}, \qquad k = ..., -1, 0, 1, ... \tag{4}$$

O domínio das Equações (3) e (4) pode ser reduzido para $k >= 0$ e sua imagem é limitada a $|\rho_k| <= 1$ (com $\rho_0 = 1$) (CIPRA, 2020, p. 125). Além disso, o gráfico que descreve a dinâmica de curto prazo de uma série temporal por meio de diversos valores de k é denominado correlograma, e pode ser representado conforme ilustrado na Figura 11.

Figura 11 – Exemplo de um correlograma



Fonte: Cipra (2020, p. 126).

No correlograma da Figura 11, caso ele tivesse sido construído a partir de uma série temporal com uma frequência semanal, poderia ter as seguintes interpretações:

- Em $k = 0$, é mostrado um $\rho_k = 1$ porque é dividido o valor da autovariância nesse tempo k por ele mesmo, conforme mostrado na Equação (4);
- Em $k = 1$, existe uma autocovariância positiva entre os valores pares $(y_t,\ y_{t-1})$, mas ela ainda é menor que a autocovariância de um valor comparado a ele mesmo, por isso todos demais valores no correlograma têm sempre seu módulo menor que 1.

Nos casos em que a série temporal é estacionária, as Equações (3) e (4) podem ser reescritas respectivamente como:

$$c_k = \frac{1}{n} \sum_{t=k+1}^{n} (y_t - \overline{y})(y_{t-k} - \overline{y}), \qquad k = 0, 1, ..., n-1 \tag{5}$$

$$r_k = \frac{c_k}{c_0}, \qquad k = 0, 1, ..., n-1 \tag{6}$$

onde,

$$\overline{y} = \frac{1}{n} \sum_{t-1}^{n} y_t \tag{7}$$

Outro conceito utilizado é o da função de autocorrelação parcial, a qual é uma correlação condicional. Ela pode ser definida como a correlação entre duas variáveis sob a suposição de que consideramos os valores de algum outro conjunto de variáveis (LAZZERI, 2020, p. 119). Sendo ρ_{kk} o coeficiente de correlação parcial entre y_t e y_{t-k} sob os valores fixos $y_{t-k+1},...,y_{t-1}$, onde $\rho_{00} = 1$ e $\rho_{11} = \rho_1$, Cipra (2020, p. 128) utiliza o algoritmo recursivo de Durbin-Levinson para o cálculo sequencial dos valores de correlação parcial conforme explicitado nas Equações (8) e (9).

$$r_{11} = r_1, \qquad r_{kk} = \frac{r_k - \sum_{j=1}^{k-1} r_{k-1,j}.r_{k-j}}{1 - \sum_{j=1}^{k-1} r_{k-1,j}.r_j} \qquad \text{para } k > 1 \tag{8}$$

$$r_{kj} = r_{k-1,j} - r_{kk}.r_{k-1,k-j} \qquad \text{para } j = 1,...,k-1 \tag{9}$$

2.2.4 Métricas de erro

É possível comparar duas séries temporais distintas e analisar a semelhança delas por meio do cálculo de erros. Essa tarefa é especialmente importante ao se validar os dados gerados por um modelo e compará-los aos dados reais de teste. Na área de ciência de dados algumas métricas são já bem difundidas e aceitas. A primeira a ser descrita por Géron (2019, p. 71) é a RMSE (do inglês, *Root Mean Square Error*) que fornece a soma dos erros quadráticos médios, conforme a Equação (10).

$$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(\mathbf{x}^{(i)}) - y^{(i)})^2} \tag{10}$$

onde: m é a quantidade de registros da série temporal, $\mathbf{x}^{(i)}$ é o vetor com os valores a serem testados do i-ésimo registro e $y^{(i)}$ é o valor desejado (ou seja, o valor real da série temporal), $\mathbf{X}$ é a matriz contendo todos os valores das variáveis a serem testadas e h é a função do sistema preditivo (também chamado de hipótese). Outra métrica disponível para uso é o MAE (do inglês, *Mean Absolute Error*) cujo valor calculado é a média do erro absoluto, conforme mostrado na Equação (11).

$$\text{MAE}(\mathbf{X}, h) = \frac{1}{m}\sum_{i=1}^{m}|h(\mathbf{x}^{(i)}) - y^{(i)}| \tag{11}$$

Ambas as Equações (10) e (11) são formas de se medir a distância entre dois vetores: o de previsões e o de valores alvo. Enquanto o RMSE corresponde à norma Euclideana, também conhecida como norma l_2, e mais sensível à presença de *outliers*, o MAE corresponde à norma de Manhattan, conhecida como norma l_1, e sendo mais indicado na comparação de séries com uma maior probabilidade de possuírem *outliers* (GÉRON, 2019, p. 74). Segundo afirma Cipra (2020, p. 8), o *outlier* pode ser definido como uma irregularidade não esperada no valor

de uma série temporal, usualmente causada por falha humana, fraudes, indisponibilidade de sistemas, etc.

2.3 Modelos preditivos

Nessa seção estão contidas as informações gerais sobre os modelos preditivos avaliados neste estudo. Cada um deles possui diferentes premissas para predição de valores futuros, como: a série temporal ser univariada ou multivariada, sazonal ou não-sazonal, estacionária ou não-estacionária e linear ou não-linear.

2.3.1 SARIMAX

SARIMAX (do inglês, *seasonal autoregressive integrated moving average model with exogenous variables*) é um modelo estatístico com componentes sazonais (S), autorregressivos (AR), integrativos (I), de média móvel (MA) com variáveis exógenas (X) (PEIXEIRO, 2022, p. 180). Nos próximos tópicos serão analisados cada um dos seus componentes e suas limitações.

2.3.1.1 MA - média móvel

Um modelo de média móvel de ordem q, representado por MA(q), afirma que o valor atual é linearmente dependente dos termos de erro atual e anteriores, sendo esses termos de erro mutualmente independentes e normalmente distribuídos (PEIXEIRO, 2022, p. 63). Sendo o valor presente denotado por y_t, a média da série temporal por μ, o termo do erro presente por ϵ_t, os termos de erros passados por ϵ_{t-q} e a magnitude do impacto dos erros passados por θ_q, Peixeiro (2022, p. 63) expressa a equação geral de um modelo de média móvel de ordem q conforme mostrado na Equação (12).

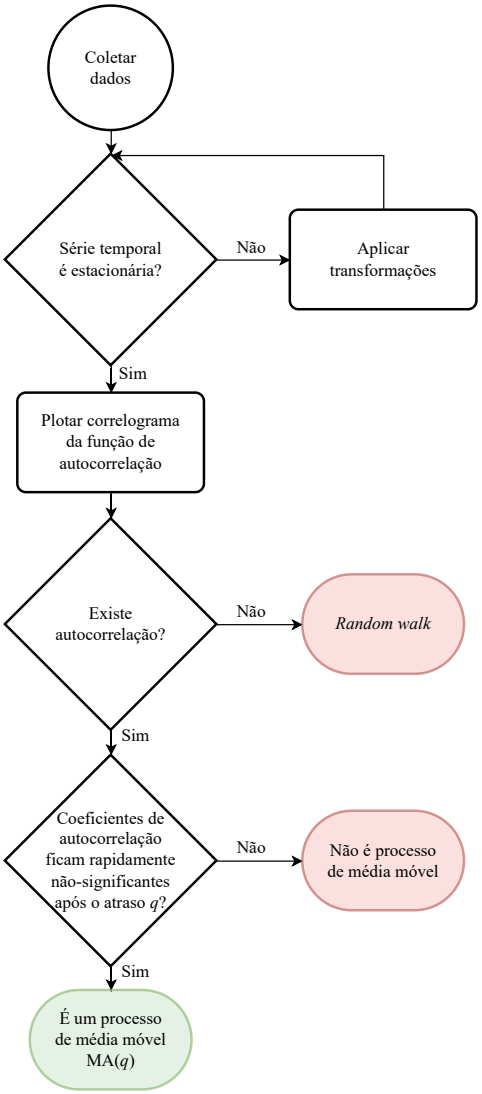
$$y_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + ... + \theta_q\epsilon_{t-q} \tag{12}$$

A ordem q do modelo de média móvel informa a quantidade de termos de erros passados que influenciam no valor atual predito, e é importante que q seja determinado para ser ajustado o modelo de forma correta (PEIXEIRO, 2022, p. 64).

No caso da série temporal não ser estacionária, é necessário aplicar transformações até que essa condição seja verdadeira. Após isso, deve ser construído o correlograma para os essa base de dados e verificar se existe alguma mudança abrupta dos coeficientes de autocorrelação após algum atraso q. Em caso positivo, existe um processo de média móvel MA(q). Os passos necessários para essa análise estão resumidos na Figura 12.

Na Figura 13 é exemplificado o correlograma de uma série temporal (já diferenciada para torná-la estacionária) que pode ser modelada por meio de médias móveis de ordem 2, já que após o atraso $k = 2$ a autocorrelação entra na zona de não-significância em azul. A zona em azul é definida pelo nível de significância α escolhido e o intervalo de confiança gerado.

Figura 12 – Diagrama com passos para determinação da ordem do processo de média móvel



Fonte: Adaptado de Peixeiro (2022, p. 65).

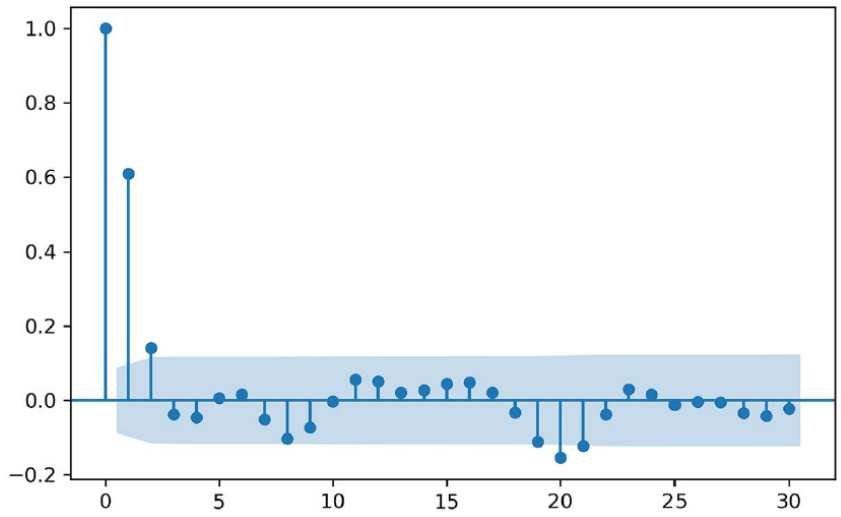
Por padrão do pacote `statsmodel` de Seabold e Perktold (2010) tem-se um $\alpha = 0{,}05$ e um intervalo de confiança igual a 95%.

2.3.1.2 AR - autorregressivo

De acordo com Peixeiro (2022, p. 84), um processo autorregressivo AR(p) de ordem p estabelece que a variável resposta atual é linearmente dependente de seus próprios valores passados. Em outras palavras, é realizada a regressão de uma variável com ela mesma.

Seja y_t a variável resposta no tempo atual, C uma constante, ϵ_t o termo de erro presente, y_{t-p} os valores passados da série temporal e ϕ_p a magnitude da influência dos valores passados no presente, a expressão geral de um modelo autorregressivo de ordem p AR(p) pode ser representada pela Equação (13).

Figura 13 – Correlograma da função de autocorrelação para um modelo de média móvel



Fonte: Peixeiro (2022, p. 69).

$$y_t = C + \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... \ \phi_p y_{t-p} + \epsilon_t \tag{13}$$

De forma semelhante ao modelo de médias móveis, também é necessário encontrar a ordem correta para a série temporal analisada. Para um processo autorregressivo é necessário utilizar também, além da função de autocorrelação, a função de autocorrelação parcial, segundo as etapas desenvolvidas por Peixeiro (2022, p. 85).

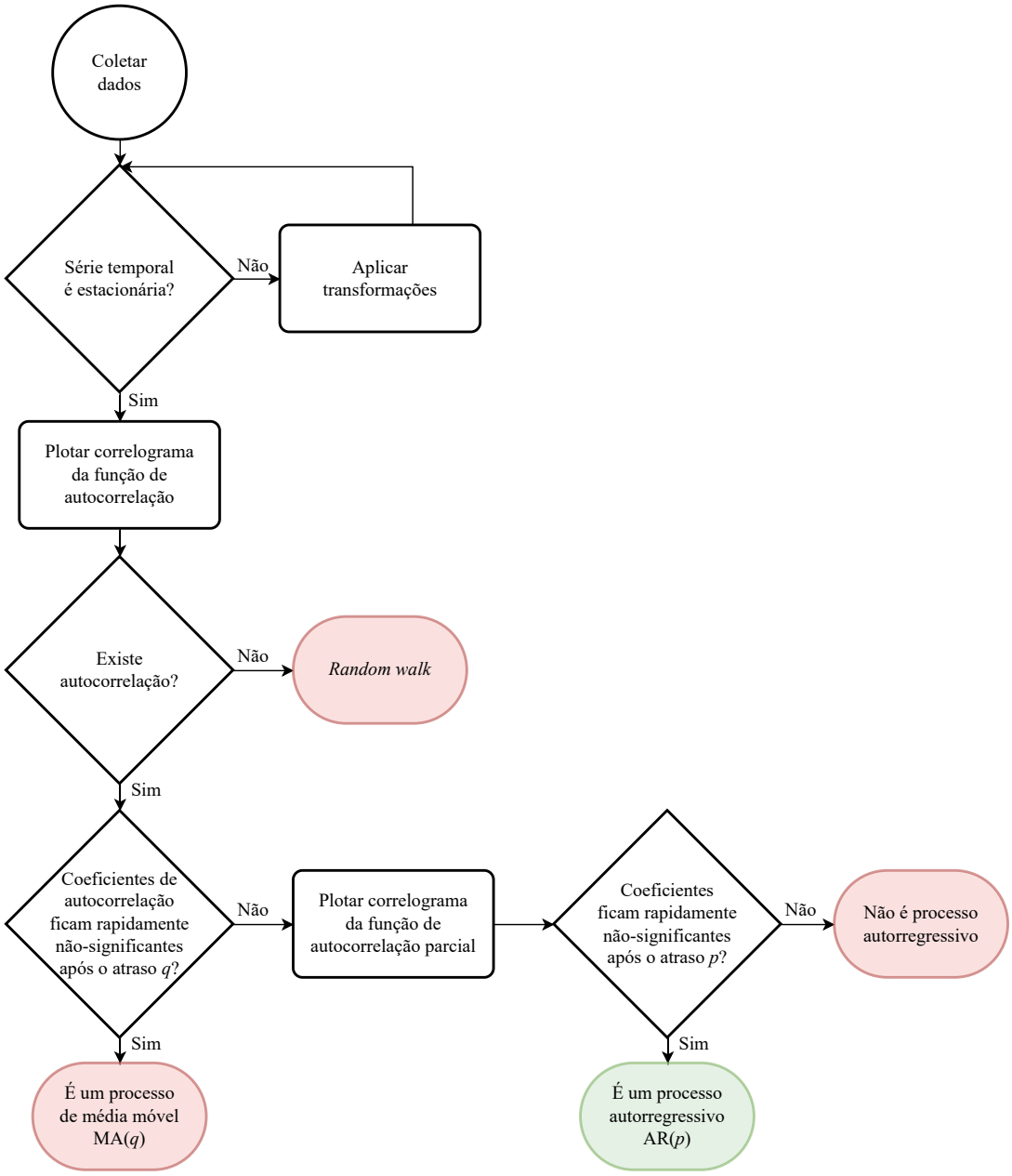
Assim como na Figura 12, é necessário transformar a série temporal em uma estacionária e analisar seu correlograma de autocorrelação. Entretanto, nesse correlograma deve ser constatada uma autocorrelação, mas sem mudanças abruptas para regiões não-significantes. Além disso, é necessária a construção do correlograma da função de autocorrelação parcial e a constatação de uma mudança repentina dos coeficientes para valores não-significantes após um atraso *p*. Feito isso, é possível confirmar um processo autorregressivo AR(p). Esse procedimento é ilustrado e resumido no diagrama da Figura 14 a seguir.

Na Figura 15 é exemplificado um correlograma da função de autocorrelação típico de um processo autorregressivo, já que há coeficientes significativos para atrasos maiores que 0 e não há uma queda abrupta desses coeficientes, mas sim uma queda exponencial. Já na Figura 16 é mostrado o correlograma da função de autocorrelação parcial para um processo autorregressivo, onde após $p = 2$ há uma diminuição abrupta do coeficiente para a zona de não-significância.

2.3.1.3 I - integrativo

A porção integrativa do modelo geral do SARIMAX diz respeito apenas a quantas vezes a série temporal precisou ser diferenciada para se tornar estacionária (PEIXEIRO, 2022,

Figura 14 – Diagrama com passos para determinação da ordem do processo de autorregressão



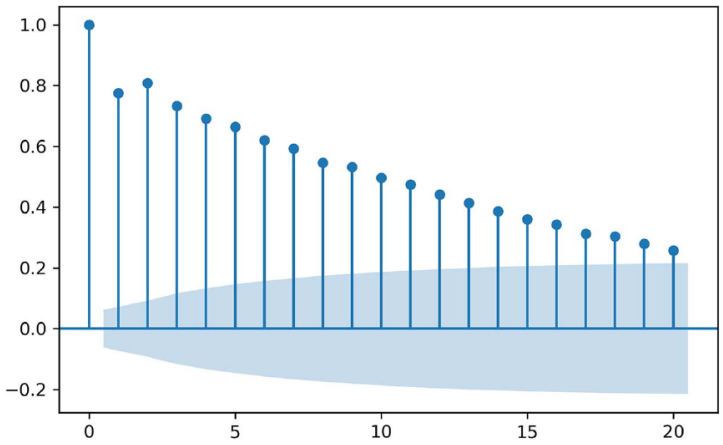
Fonte: Adaptado de Peixeiro (2022, p. 85).

p. 142). Um modelo autorregressivo, integrativo e de médias móveis ARIMA(*p,d,q*) pode ser expresso pela Equação (14)

$$y_t^{'} = C + \varphi_1 y_{t-1}^{'} + ... \ \varphi_p y_{t-p}^{'} + \theta_1 \epsilon_{t-1}^{'} + ... + \theta_q \epsilon_{t-q}^{'} + \epsilon_t \tag{14}$$

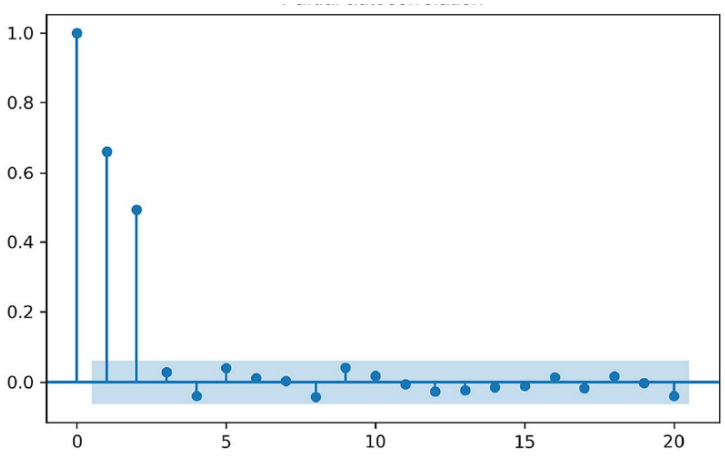
onde $y_t^{'}$ representa o valor presente da série temporal diferenciada *d* vezes, C é uma constante, $\varphi_p y_{t-p}^{'}$ são os valores passados e $\theta_q \epsilon_{t-q}^{'}$ são os termos de erro passados, ambos da série temporal diferenciada, e ϵ_t o termo de erro presente.

Figura 15 – Correlograma da função de autocorrelação para um modelo de autorregressão



Fonte: Peixeiro (2022, p. 89).

Figura 16 – Correlograma da função de autocorrelação parcial para um modelo de autorregressão



Fonte: Peixeiro (2022, p. 91).

2.3.1.4 S - sazonal

O modelo SARIMA(p,d,q)(P,D,Q)$_m$ introduz quatro novos parâmetros que permitem contabilizar o efeito sazonal em uma série temporal, expandindo o já conhecido modelo ARIMA(p,d,q). Os três primeiros parâmetros P, D e Q possuem o mesmo significado explicado anteriormente, mas agora referem-se ao componente sazonal da série. Já o componente m refere-se à frequência, ou seja, um número de observações por ciclo. Para dados registrados a cada ano, trimestre, mês ou semana, considera-se geralmente o ciclo de um ano, obtendo-se respectivamente frequências m iguais a 1 (um registro por ano), 4 (quatro trimestres ano), 12 (12 meses em ano) e 52 (52 semanas em um ano). Nos dados mais granulares, onde, por exemplo, os dados são registrados a cada hora e o ciclo tem uma duração de uma semana, o parâmetro m seria igual a 168, já que existem 168 horas dentro de uma semana (PEIXEIRO,

2022, p. 157-158).

Agora que o parâmetro de frequência foi definido para a série, Peixeiro (2022, p. 159) exemplifica a situação que se $m = 12$ e $P = 2$, deveria ser incluída na equação geral do modelo SARIMA dois termos passados da série com atrasos iguais aos múltiplos de *m* iniciando por 1, y_{t-12} e y_{t-24}. De forma similar, caso $D = 1$, deveria ser feita a diferença sazonal $y_t^{'} = y_t - t_{t-12}$ para torná-la estacionária. Por fim, no caso em que $Q = 2$, deveriam ser incluídos os termos de erros passados ϵ_{t-12} e ϵ_{t-24}.

2.3.1.5 X - variáveis exógenas

Por fim, o componente exógeno da equação geral do modelo SARIMAX(*p,d,q*)(P,D,Q)$_m$ nada mais é do que a combinação linear de diferentes variáveis exógenas que possuem influência na variável resposta a ser calculada pelo modelo (PEIXEIRO, 2022, p. 183). Portanto, sua fórmula geral pode ser resumida na Equação (15) a seguir.

$$y_t = \text{SARIMA}(p,d,q)(P,D,Q)_m + \sum_{i=1}^{n} \beta_i X_t^i \tag{15}$$

onde β_i é o coeficiente de importância da variável exógena *i* e X_t^i é a variável exógena *i* no tempo igual a *t*. Como bem afirma Peixeiro (2022, p. 182), as variáveis exógenas podem ser também categóricas desde que sejam transformadas em *flags* binárias ou sejam codificadas em números por alguma lógica aplicável a ela.

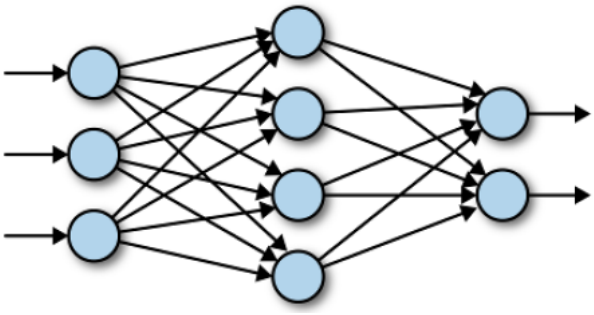
2.3.2 LSTM

O Aprendizado Profundo, também conhecido como *Deep Learning* (DL), tem se mostrado promissor em diversas áreas devido a sua grande flexibilidade de adaptação e possibilidade de identificar comportamentos altamente não-lineares sem necessitar especificar uma fórmula funcional para aquele fenômeno. De forma resumida, o DL pode ser definido como um ramo do aprendizado de máquina no qual células (denominadas também como neurônios) que realizam cálculos matemáticos são densamente conectadas umas às outras (NIELSEN, 2019, p. 289), conforme ilustrado na Figura 17. O resultado dos cálculos, geralmente não lineares devido a funções de ativação com esse comportamento, em uma célula é transferido como valor de entrada para as próximas células conectada conectadas multiplicado por peso individual para cada vértice conectando essas duas células.

Como afirma Nielsen (2019, p. 290), no DL não são necessários, por exemplo, a estacionariedade na série temporal e a busca cuidadosa por certos parâmetros como no modelo SARIMAX (p, d, q, P, D, Q e m). Entretanto, ainda é necessário um pré-processamento dos dados, já que são melhores aceitos pelos algoritmos valores numéricos de entrada em um intervalo [-1, 1].

Segundo Géron (2019, p. 649), no processamento de sequências, como em séries temporais ou em linguagem natural, é muito comum o uso de redes neurais recorrentes (RNNs,

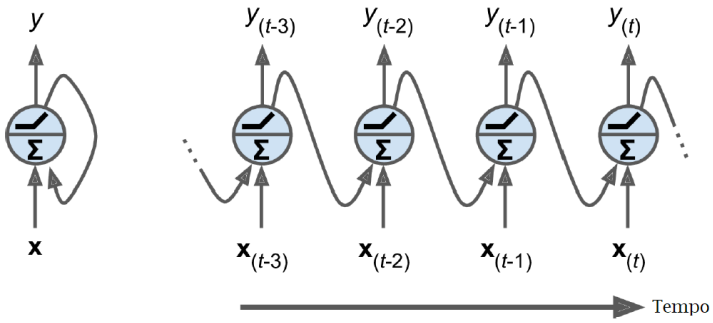
Figura 17 – Estrutura de uma rede neural artificial



Fonte: Adaptado de Nielsen (2019, p. 293).

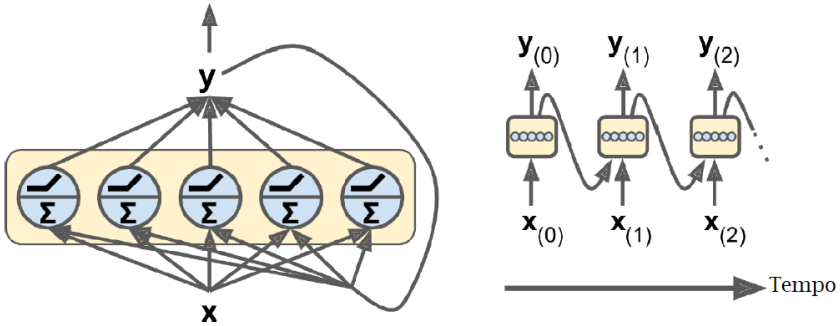
do inglês *recurrent neural networks*) para prever algum dado futuro utilizando células de memória (ilustrada na Figura 18). Diferentemente das células de uma rede neural artificial ilustradas na Figura 17, em um dado tempo t as células de memória da Figura 18 recebem tanto o vetor de entradas $\mathbf{x}_{(t)}$ quanto a sua própria saída no tempo anterior, $y_{(t-1)}$. Quando é formada uma camada de células de memória, em um dado tempo t cada neurônio recebe agora um vetor de entradas $\mathbf{x}_{(t)}$ e outro vetor de saídas do momento anterior $\mathbf{y}_{(t-1)}$, conforme ilustrado na Figura 19.

Figura 18 – Célula de memória de uma rede neural recorrente



Fonte: Adaptado de Géron (2019, p. 650).

Figura 19 – Camada de células de memória de uma rede neural recorrente

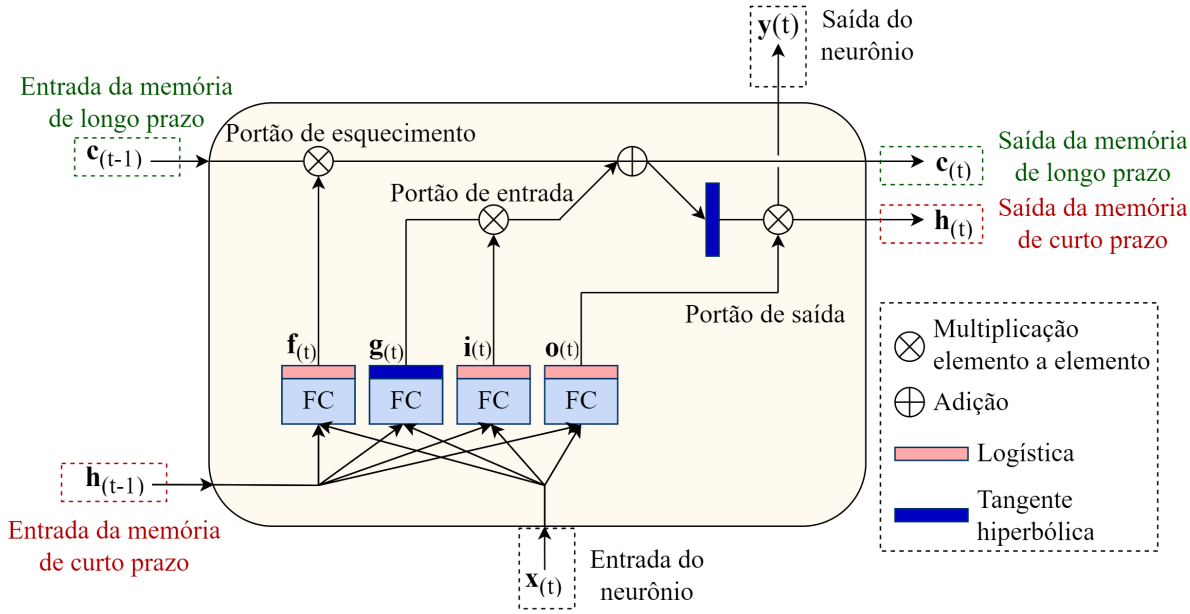


Fonte: Adaptado de Géron (2019, p. 650).

Contudo, como alerta Géron (2019, p. 648), as redes neurais recorrentes têm alguns

problemas conhecidos como: gradientes descendentes instáveis durante a atualização dos parâmetros, e uma memória de curto prazo limitada. Ou seja, existe uma grande possibilidade da otimização dos pesos do modelo ficar presa em um ponto ótimo local e, caso haja algum padrão reconhecido em tempos muito distantes do tempo atual t, ele não será lembrado e considerado no cálculo.

Figura 20 – Arquitetura de uma célula LSTM



Fonte: Adaptado de Géron (2019, p. 672).

Para solucionar o obstáculo da memória de curto prazo limitada, Hochreiter e Schmidhuber (1997) propôs o modelo de célula chamada LSTM (do inglês, *Long Short-Term Memory*) com fluxos de informações separados para uma memória de longo prazo e outra de curto prazo. A memória de longo prazo é controlada principalmente pelo portão de esquecimento, que retira as informações de longo prazo que não são mais necessárias, e pelo portão de entrada, que escolhe memórias do curto prazo para integrar também as memórias de longo prazo. A arquitetura de uma LSTM é ilustrada na Figura 20.

Géron (2019, p. 672) define o vetor $\mathbf{c}_{(t)}$ como o vetor de saída da memória de longo prazo e $\mathbf{h}_{(t)}$ como o vetor de saída da memória de curto prazo da célula. A ideia com essa estrutura é que a rede neural consiga não só aprender e armazenar padrões antigos, mas também esquecer aqueles que não são mais úteis.

Analisando o fluxo de informações vindo da célula de um tempo anterior $t-1$, o vetor de entrada da memória de longo prazo $\mathbf{c}_{(t-1)}$ passa primeiramente pelo portão de esquecimento onde algumas memórias são apagadas, depois outras memórias são acrescentadas via operação de adição com o vetor resultante do portão de entrada e por fim o vetor $\mathbf{c}_{(t)}$ é enviado para fora da célula sem mais transformações (GÉRON, 2019, p. 672-673). Além disso, uma cópia dessas informações é levada a uma função de tangente hiperbólica e filtradas pelo portão de

saída, produzindo a saída da memória de curto prazo $\mathbf{h}_{(t)}$ e a saída do neurônio $\mathbf{y}_{(t)}$.

Ainda analisando o fluxo de novas memórias para a célula, Géron (2019, p. 673) destaca os vetores de entrada do neurônio $\mathbf{x}_{(t)}$ e de entrada da memória de curto prazo $\mathbf{h}_{(t-1)}$ totalmente conectados (FC) a quatro diferentes camadas de processamento, cada uma servindo a um propósito específico:

- A camada principal é a que gera o vetor $\mathbf{g}_{(t)}$. Ela possui o papel de processar as entradas atuais e as memórias de curto prazo passadas;
- As outras três camadas são as responsáveis por fazer o controle dos portões ilustrados na Figura 20. Para tal é utilizada nelas a função de ativação logística, cuja saída encontra-se no intervalo de 0 a 1. Ou seja, ao possuir um valor muito próximo a 0 pode se dizer que o portão é fechado e as informações não fluem para as próximas etapas. Da mesma forma, quando o valor de saída é muito próximo a 1 o portão é aberto e as informações são passadas adiante;
  - O portão de esquecimento, controlado por $\mathbf{f}_{(t)}$, é responsável por escolher quais partes da memória de longo prazo devem ser apagadas;
  - O portão de entrada, controlado por $\mathbf{i}_{(t)}$, é responsável por escolher quais partes de $\mathbf{g}_{(t)}$ devem ser adicionadas à memória de longo prazo;
  - O portão de saída, controlado por $\mathbf{o}_{(t)}$, é responsável por escolher quais partes da memória de longo prazo devem ser repassadas adiante para $\mathbf{h}_{(t)}$ e $\mathbf{y}_{(t)}$.

As operações matemáticas que ocorrem na célula LSTM são descritas por seis equações distintas, todas apresentadas da Equação (16) a (21).

$$\mathbf{i}_{(t)} = \sigma(\mathbf{W}_{xi}^T \mathbf{x}_{(t)} + \mathbf{W}_{hi}^T \mathbf{h}_{(t-1)} + \mathbf{b}_i) \tag{16}$$

$$\mathbf{f}_{(t)} = \sigma(\mathbf{W}_{xf}^T \mathbf{x}_{(t)} + \mathbf{W}_{hf}^T \mathbf{h}_{(t-1)} + \mathbf{b}_f) \tag{17}$$

$$\mathbf{o}_{(t)} = \sigma(\mathbf{W}_{xo}^T \mathbf{x}_{(t)} + \mathbf{W}_{ho}^T \mathbf{h}_{(t-1)} + \mathbf{b}_o) \tag{18}$$

$$\mathbf{g}_{(t)} = tanh(\mathbf{W}_{xg}^T \mathbf{x}_{(t)} + \mathbf{W}_{hg}^T \mathbf{h}_{(t-1)} + \mathbf{b}_g) \tag{19}$$

$$\mathbf{c}_{(t)} = \mathbf{f}_{(t)} \otimes \mathbf{c}_{(t-1)} + \mathbf{i}_{(t)} \otimes \mathbf{g}_{(t)} \tag{20}$$

$$\mathbf{y}_{(t)} = \mathbf{h}_{(t)} = \mathbf{o}_{(t)} \otimes tanh(\mathbf{c}_{(t)}) \tag{21}$$

onde:

- $\mathbf{W}_{xi}$, $\mathbf{W}_{xf}$, $\mathbf{W}_{xo}$ e $\mathbf{W}_{xg}$ são as matrizes de pesos das quatro camadas com sua conexão com o vetor de entrada do neurônio $\mathbf{x}_{(t)}$;
- $\mathbf{W}_{hi}$, $\mathbf{W}_{hf}$, $\mathbf{W}_{ho}$ e $\mathbf{W}_{hg}$ são as matrizes de pesos das quatro camadas com sua conexão com o vetor de entrada da memória de curto prazo $\mathbf{h}_{(t-1)}$;

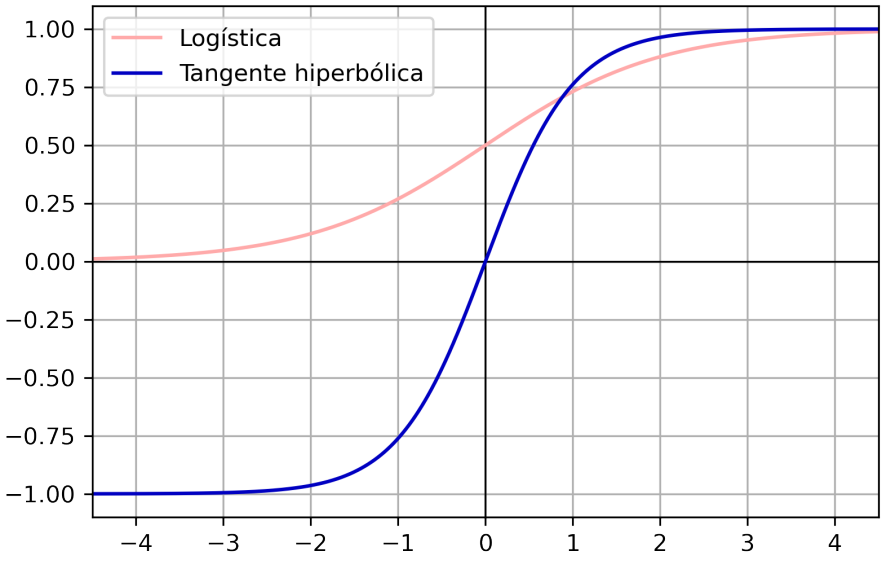
- $\mathbf{b}_i$, $\mathbf{b}_f$, $\mathbf{b}_o$ e $\mathbf{b}_g$ são as matrizes de *bias* das quatro camadas.

Por fim, as funções de ativação não lineares explicitadas no primeiro parágrafo da Subseção 2.3.2 utilizadas na LSTM são a logística σ (também conhecida como sigmóide) e a tangente hiperbólica $tanh$, representadas respectivamente pelas Equações (22) e (23), conforme referência da biblioteca `Keras` (CHOLLET et al., 2015), e ilustradas na Figura 21.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{22}$$

$$tanh(x) = \frac{senh(x)}{cosh(x)} = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{23}$$

Figura 21 – Funções de ativação logística e tangente hiperbólica



Fonte: Autoria própria.

A função de tangente hiperbólica produz sempre resultados entre -1 e 1, enquanto a função logística produz resultados entre 0 e 1.

3 TRABALHOS RELACIONADOS

Neste capítulo são apresentados os trabalhos relacionados que buscaram por meio de algoritmos de Aprendizado de Máquina solucionar problemas semelhantes ao investigado nesse trabalho. Um resumo dos trabalhos analisados foi consolidado e apresentado no Quadro 1. As próximas seções explicam estes trabalhos com mais detalhes.

Quadro 1 – Trabalhos relacionados.

Referência	**Problema a ser Resolvido**	**Algoritmo Base**	**Técnicas de Otimização**	**Função Custo**	**Intervalo de Tempo**	**Fontes de Dados**	**Métricas de Performance**
Abdollahi e Ebrahimi (2020)	prever preços do petróleo cru	ANFIS, ARFIMA, e Markov-switching	diminuição de peso do modelo por aumento do erro e algoritmo genético	RMSE, MAE, MAPE e erro médio	7 anos e 7 meses (2009-12 à 2017-06)	Macro trends	RMSE, MAE e MAPE
Gupta e Nigam (2020)	prever preços do petróleo cru	RNA (backpropagation)	gradiente descendente		5 anos e 11 meses (2014-01 à 2019-09)	investing.com	RMSE
Urolagin, Sharma e Datta (2021)	prever preços do petróleo cru	MLSTM	Adam	MSE	19 anos e 6 meses (2000-01 à 2019-06)	investing.com	MSE, MAE e R^2
Carassai et al. (2021)	prever preço médio do etanol hidratado vendido nos postos de combustíveis brasileiros	regressão linear múltipla	método dos mínimos quadrados		4 anos e 1 mês (2014-01 à 2018-01)	ANP, IBGE, Ministério da Agricultura, Pecuária e Abastecimento, CONAB, CEPEA, ÚNICA, FGV	R^2

Fonte: Autoria própria.

3.1 Trabalhos que buscam prever o preço do petróleo cru

Os artigos explorados nessa seção utilizaram diversas técnicas visando prever preços futuros do petróleo cru, analisando também diferentes janelas de tempo.

Abdollahi e Ebrahimi (2020) elaboraram um modelo híbrido composto pelos algoritmos *Adaptive Neuro Fuzzy Inference System* (ANFIS), *Autoregressive Fractionally Integrated Moving Average* (ARFIMA) e *Markov-switching* para preverem o preço do barril de petróleo *Brent*. Combinando estes três algoritmos permitiria capturar diferentes características da série temporal, como sua não-linearidade e inter-relações de mercado. O preço do ativo em um determinado tempo foi testado em três diferentes cenários:

1. Igual peso (importância) para as previsões de todos os algoritmos;
2. Maior peso para algoritmos com menores erros (testados individualmente e em conjunto por meio de média aritmética simples);
3. Pesos otimizados via algoritmos genéticos.

Para as medidas de erros foram utilizados *root mean square error* (RMSE), *mean absolute error* (MAE) e *mean absolute percentage error* (MAPE). O cenário capaz de obter os menores valores de erro foi o de otimização via algoritmos genéticos com um RMSE, MAE e MAPE respectivamente iguais a 7,8574, 6,1803 e 0,0067. Esses baixos valores de erros foram obtidos devido à curva de valores gerados pelo modelo híbrido acompanhar as mesmas tendências dos valores reais na maior parte da janela de tempo prevista (30 iterações).

Também para se prever o preço futuro do petróleo cru, Gupta e Nigam (2020) usaram uma Rede Neural Artificial (RNA) cujos neurônios na camada de entrada recebiam um vetor de preços defasados, ou seja, para se prever o preço do petróleo no tempo t (y_t) o modelo necessitava das informações [y_{t-1}, y_{t-2}, ..., y_{t-n}], onde n é a quantidade de últimos preços. Os preços foram normalizados no intervalo [0,001, 0,005] antes de serem usados no treinamento do modelo. Como função de ativação da rede foi utilizada a função sigmoide, com algoritmo de treinamento *backpropagation* (BP). A quantidade de últimos preços (n) utilizada que gerou uma previsão mais correta foi igual a três, considerando como métrica de erro a RMSE cujo valor obtido foi de 7,68.

Diferentemente dos dois trabalhos anteriores, onde foram utilizadas apenas séries temporais univariadas, Urolagin, Sharma e Datta (2021) consideraram também como entradas para seus modelos *Multivariate Long Short Term Memory* (MLSTM) as seguintes variáveis: preço do ouro, índice S&P 500, índice de dólar americano, rendimentos dos títulos do Tesouro dos Estados Unidos em 10 anos e o índice de utilidades Dow Jones. Após analisada a importância das cinco variáveis, foram treinados dois modelos MLSTM, um com todas as variáveis e outro apenas com as três mais significativas (índice de dólar americano, preço do ouro e rendimentos dos títulos do Tesouro dos Estados Unidos em 10 anos selecionados pelo maior *F-Score* calculados pelo *selectKBest* do `scikit-learn`). Como os melhores resultados foram obtidos pelo último modelo, que usou apenas três *inputs*, foram treinados dois modelos adicionais apenas com essas três variáveis. O primeiro com os dados passando por uma transformação do

tipo *Z-score* e o segundo com os dados passando por uma transformação de Mahalanobis. Por fim, foram treinados outros dois modelos com as transformações *Z-score* e de Mahalanobis, mas agora excluindo dos dados de entrada os *ouliers*. O modelo de três variáveis com transformação *Z-score* e retirada de *ouliers* foi o melhor dentre os seis, apresentando um RMSE igual a 0,212.

3.2 Trabalhos que buscam prever o preço final de revenda

Nessa seção o único trabalho encontrado nas buscas desenvolveu modelos cujo objetivo foi prever o preço de revenda final do etanol hidratado.

Carassai et al. (2021) buscaram desenvolver um modelo multivariado para a previsão do preço de revenda do etanol hidratado no território brasileiro. As quatro variáveis exógenas utilizadas pelos autores no modelo final foram: o volume de produção de petróleo, o preço médio de revenda da gasolina, o câmbio dólar americano e real brasileiro, e o preço médio do açúcar no mercado nacional. Essa escolha foi feita após a análise de 26 variáveis e a adição delas ao modelo de regressão linear em ordem decrescente de correlação até que o valor da estatística F (teste de significância da existência da regressão) mostrasse que a variável resposta obtida pela equação anterior deixasse de ser significativa (técnica *Stepwise*). Com a retirada de *outliers* dos conjuntos de dados, o modelo de regressão linear múltipla obteve um erro médio percentual absoluto (métrica MAPE dos trabalhos analisados anteriormente) igual a 0,0312 e um R^2 igual a 0,9867.

3.3 Lacunas nos trabalhos analisados

Dos quatro trabalhos analisados na área de previsão de preços de combustíveis, três têm o foco na predição do preço do petróleo cru. Apesar dele de ter um impacto no preço final de revenda, não possui valores de fácil compreensão pela população em geral.

O único dos quatro trabalhos analisados que possui como resposta o preço final de revenda em um tempo futuro faz a previsão para o etanol hidratado. Apesar de comum em carros e motos, ele ainda é pouco representativo no volume de vendas em território nacional, conforme mostram os dados da ANP (Agência Nacional do Petróleo, Gás Natural e Biocombustíveis) (2022b) sumarizados na Tabela 1.

Além da questão do etanol hidratado de abranger apenas 12% do volume vendido dentro do país em 2021, o principal combustível usado no transporte de carga e coletivo de passageiros é o óleo diesel (CARVALHO, 2018). Isso significa que para se analisar a relação com fretes de produtos e preço finais de mercadorias este é o combustível mais adequado a ser estudado.

Tabela 1 – Volume relativo vendido em 2021 no Brasil.

Combustível	Volume relativo (m^3/m^3)
Óleo Diesel	44,52%
Gasolina C	28,18%
Etanol Hidratado	12,04%
GLP	9,65%
Querosene De Aviação	3,14%
Óleo Combustível	2,43%
Gasolina De Aviação	0,03%
Querosene Iluminante	0,00%

Fonte: Autoria própria.

4 MATERIAIS E MÉTODOS

Neste capítulo é descrita a metodologia experimental utilizada para análise e predição dos preços de revenda do óleo diesel no sudeste brasileiro.

4.1 Datasets

Foram utilizadas nesse trabalho três diferentes bases de dados públicas, descritas a seguir.

4.1.1 Preços médios de revenda de combustíveis

A média de preços dos combustíveis agrupados por região administrativa do Brasil disponibilizada pela ANP (Agência Nacional do Petróleo, Gás Natural e Biocombustíveis) (2022a) é um dado que é utilizado tanto na análise quanto no treinamento do modelo preditivo. Devido à Lei do Petróleo (Lei nº 9478/1997, artigo 8º) a ANP faz o acompanhamento dos preços de revenda por postos autorizados e os dados são coletados semanalmente por empresa terceirizada, abrangendo hoje mais de 400 municípios.

Dela foram usadas as informações de data, região administrativa, tipo de combustível, número de postos pesquisados, unidade de medida e preço médio de revenda, respectivamente, nas colunas 'DATA FINAL', 'REGIÃO', 'PRODUTO', 'NÚMERO DE POSTOS PESQUISADOS', 'UNIDADE DE MEDIDA' e 'PREÇO MÉDIO REVENDA'.

4.1.2 Volume de combustíveis comercializados

Outra base de dados importante e utilizada é a de volume de combustível comercializado por mês divulgado pela ANP (Agência Nacional do Petróleo, Gás Natural e Biocombustíveis) (2022b). Dessa forma é possível escolher o combustível e a região mais relevantes para induzir os modelos.

Nesse *dataset* as informações utilizadas foram as de ano, mês, região administrativa, tipo de combustível e volume vendido. Elas foram disponibilizadas, respectivamente, nas colunas 'ANO', 'MÊS', 'GRANDE REGIÃO', 'PRODUTO' e 'VENDAS'.

4.1.3 Preço do barril *Brent* e cotação do dólar

O último conjunto de dados utilizado refere-se aos preços em dólar do barril de petróleo do tipo *Brent* e a cotação do dólar em reais, ambos obtidos diretamente via biblioteca `yfinance` (AROUSSI et al., 2017). Com ela é possível obter diversos dados econômicos diretamente da API do Yahoo! Finance de uma forma simples e online.

Em ambos os casos é retornada uma tabela com diversas informações, mas as únicas utilizadas foram a de data e de preço de fechamento (último preço daquele dia), nas colunas 'Date' e 'Close', respectivamente.

4.2 Algoritmos utilizados

O primeiro algoritmo treinado e validado foi o SARIMA, cujos componentes sazonais, autorregressivos, integrativos e de médias móveis são ajustados aos dados históricos por meio da biblioteca `pmdarima`. Ele foi escolhido dentre os demais tradicionais de aprendizado de máquina (ou seja, excluindo-se os algoritmos de DL) porque ele gerou menores valores de erro no experimento criado pela biblioteca `pycaret` em comparação com todos os outros 27 algoritmos.

Ainda utilizando a mesma biblioteca, outro algoritmo treinado e validado foi o SARIMAX. Além dos componentes do SARIMA ele possui também variáveis exógenas, as quais são outras séries relacionadas à série temporal que se deseja prever.

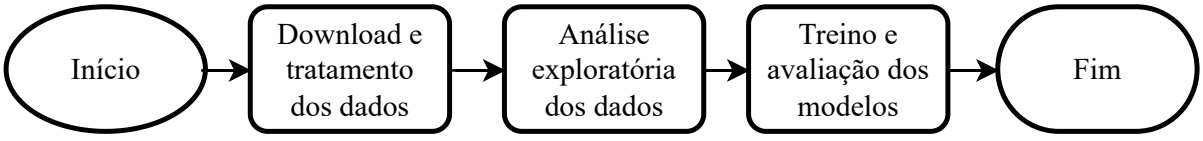
O último modelo treinado e validado foi o de rede neural recorrente LSTM. Esse tipo de célula se adapta bem às séries temporais e outros problemas de previsão de sequências devido à sua memória de curto e longo prazo capazes de aprender e esquecer padrões por grandes períodos. No primeiro treinamento de DL, foi construída uma rede neural recorrente com apenas uma camada e uma célula LSTM utilizando a biblioteca `Keras`.

Visando aumentar o desempenho das previsões geradas, a rede neural recorrente teve sua quantidade de camadas e de neurônios por camada e taxa de *dropout* otimizadas por meio de uma busca aleatória com validação cruzada.

4.3 Pipeline

Nessa subseção é descrita a sequência de métodos utilizados no estudo, permitindo o seu melhor entendimento e possibilidade de reprodutibilidade. O resumo dos passos seguidos é apresentado na Figura 22 a seguir.

Figura 22 – Diagrama da metodologia do estudo



Fonte: Autoria própria.

4.3.1 *Download* e preprocessamento dos dados

Os dados fornecidos pela ANP citados na Seção 4.1 foram baixados em sua extensão original '.csv' (valores separados por vírgulas) e carregados em *DataFrames* do `Pandas`. A base

de histórico de preços de revenda foi filtrada para considerar todos os dados desde a primeira semana disponível em janeiro de 2013 até a última semana do mês de agosto de 2022. Já a base com volume de combustíveis revendidos no Brasil foi filtrada para serem analisados apenas os dados de 2021.

As informações obtidas via `yfinance` foram carregados diretamente no ambiente Python, filtrando como data inicial a mesma data de início da disponibilização dos dados de preços históricas (janeiro de 2013), não havendo a necessidade de *download* de arquivos.

A base de dados com preços médios de revenda dos combustíveis possui algumas semanas com valores ausentes. Pelo fato dos preços não seguirem uma distribuição normal, não foi possível utilizar a média para preencher os valores ausentes. Portanto, a técnica utilizada nesse caso foi de repetir o último preço disponível nas semanas sem essa informação.

Como os dados foram utilizados para indução de modelos preditivos, foi necessária a divisão deles em dois conjuntos distintos, o de treino e de teste. Nesse último há a subdivisão de entrada de teste e de saída de teste. As datas limites para cada um deles são apresentadas na Tabela 2.

Tabela 2 – Datas utilizadas na divisão dos dados nos conjuntos de treino e teste.

Tipo de uso	**Data início**	**Data final**	**Quantidade de semanas**
Treino	05/01/2013	07/05/2022	488
Entrada de teste	14/05/2022	02/07/2022	8
Saída de teste	09/07/2022	27/08/2022	8

Fonte: Autoria própria.

Essa subdivisão foi necessária porque no modelo LSTM é necessário fornecer uma parte dos preços históricos para serem gerados os valores futuros. Para este trabalho, a rede neural recorrente requer as últimas 4 semanas (t_{-3}, t_{-2}, t_{-1} e t_0) para fornecer os preços das próximas 4 semanas (t_1, t_2, t_3 e t_4). Já nos modelos SARIMA e SARIMAX, o modelo requer apenas o número de semanas a ser gerada a previsão. No caso de serem usados os dados do final do conjunto de treinamento como entrada do modelo LSTM, poderia ocasionar o que é chamado de *data leakage* (traduzido, vazamento de dados), caracterizado por usar no teste do modelo informações que deveriam estar restritas apenas à etapa de seu treinamento.

Além disso, foi criado um objeto *scaler* utilizando os preços médios de revenda de treino e aplicada a transformação aos preços de treino e teste, já que os modelos de DL possuem um melhor desempenho normalizados dentro de um intervalo em comum. Dessa forma, essa transformação faz com que o menor valor da série fique com um valor igual a 0 e, o maior, igual a 1. Como os modelos estatísticos SARIMA e SARIMAX não necessitam de dados nessa escala, foi criado um atributo de dados escalados para ser utilizado apenas pela LSTM.

4.3.2 Análise exploratória de dados

O primeiro passo da análise exploratória de dados foi a Verificação de datas e preços mínimos e máximos por região e tipo de combustível. Esse procedimento inicial é importante para se obter uma verificação rápida de distorções e possíveis erros na base, como preços negativos, datas futuras, etc.

O próximo passo foi a Visualização gráfica de todo o histórico de preço dos combustíveis por meio de gráficos de linhas da biblioteca `Seaborn`. Dessa forma, foi possível a identificação de tendências de aumento e queda dos preços e diferenças entre as regiões administrativas.

Visando verificar um possível motivo de distorções nas curvas de preços de alguns tipos de combustíveis, foi feito um gráfico de violino (semelhante ao conhecido *box plot*) da quantidade de postos pesquisados por combustível, também utilizando o `Seaborn`. Essa análise é importante para tentar relacionar mudanças repentinas no histórico de preços médios à presença de possíveis *outliers* pesquisados.

Por último, as séries temporais de preços do óleo diesel nas 5 regiões foi decomposta de forma aditiva e os componentes de tendência, sazonalidade e resíduos foram plotados de forma separada. Dessa forma foi possível identificar se a variação de preços desse combustível se comporta de maneiras distintas pelas regiões do Brasil. Os resultados e interpretações dessas análises são apresentados na Seção 5.1.

4.3.3 Avaliação dos modelos

Usando a biblioteca `pycaret`, 28 modelos diferentes de aprendizado de máquina para previsão de séries temporais foram treinados e testados utilizando o método de validação cruzada. Após o cálculo das métricas de erro MAE e RMSE, escolheu-se o modelo SARIMA como o melhor dentre eles devido aos menores valores de erro.

Visando tentar melhorar o resultado obtido pelo `pycaret`, para os modelos SARIMA e SARIMAX (usando nesse último os preços do barril *Brent* de petróleo e cotação do dólar como variáveis exógenas) foi utilizada a biblioteca `pmdarima` na busca pelos coeficientes p, d, q, P, Q, D e m que melhor se ajustavam aos dados de treino usando como critério o AIC (do inglês, *Akaike Information Critera*). Em vez de se performar uma busca exaustiva por todas as combinações, utilizou-se o algoritmo *stepwise* que faz uma busca mais performática e rápida.

Aplicando-se agora os métodos da biblioteca `Keras`, foi construído um modelo simples de rede neural recorrente com apenas uma camada e uma célula LSTM. Na tentativa de obter menores valores de erro, tanto a quantidade de camadas e neurônios quanto o valor de *dropout* após cada camada foram otimizados via método *RandomizedSearchCV* da biblioteca `scikit-learn` com os intervalos apresentados na Tabela 3. Com os melhores hiperparâmetros encontrados, os modelos foram treinados com todos os dados do conjunto de treino na busca por melhores pesos e *bias*.

No caso dos modelos estatísticos (SARIMA e SARIMAX) foram geradas previsões na

Tabela 3 – Intervalos utilizados na otimização da LSTM por *RandomizedSearchCV*.

Hiperparâmetro	**Intervalo**	**Domínio**
Quantidade de camadas	[1, 50]	$\mathbb{Z}_+^*$
Quantidade de neurônios por camada	[1, 100]	$\mathbb{Z}_+^*$
Dropout	[0,00, 0,95]	$\mathbb{R}_+$

Fonte: Autoria própria.

quantidade igual à soma das semanas de entrada e saída de teste (conforme Tabela 2), mas os erros foram calculados apenas para o intervalo de saída de teste para ser comparável aos erros do modelo LSTM. Já para o LSTM foram usados como *inputs* os preços do conjunto de entrada de teste e geradas as previsões para o mesmo intervalo de data da saída de teste, sendo também esse o intervalo usado no cálculo das métricas de erro.

4.4 Detalhes para Reprodutibilidade do Trabalho

Todas as análises e modelos foram codificados utilizando a linguagem de programação Python em sua versão 3.9 (ROSSUM; DRAKE, 2009). Por ser uma linguagem de sintaxe simples e grande presença no mundo de ciência de dados, existe para ela uma vasta variedade de bibliotecas para os mais diversos fins. As principais bibliotecas utilizadas nesse estudo com uma breve descrição de cada uma delas está disponível a seguir:

- `Pandas` (The pandas development team, 2020): usada para carregamento dos dados em memória e tratamento deles;
- `Numpy` (HARRIS et al., 2020): usada na conversão dos dados em *arrays*;
- `scikit-learn` (PEDREGOSA et al., 2011): usada em diversas etapas desde o treino (divisão de *datasets* em porções de treino e teste) até a validação de modelos (com métricas de desempenho, como MSE e MAE);
- `Matplotlib` (HUNTER, 2007): base de métodos para criação de gráficos visualização de dados;
- `Seaborn` (WASKOM, 2021): outra biblioteca que disponibiliza métodos para a construção de gráficos;
- `statsmodel` (SEABOLD; PERKTOLD, 2010): possui métodos que permitem a decomposição de uma série temporal em seus componentes de tendência, sazonalidade e resíduos, além de também gerar correlogramas de autocorrelação e autocorrelação parcial;
- `pycaret` (ALI, 2020): considerada uma biblioteca de aprendizado de máquina *low code*, ela permite o treino e comparação de diversos modelos preditivos de uma só vez com poucas linhas de código;
- `pmdarima` (SMITH et al., 2017–): permite o ajuste de um modelo SARIMAX aos dados de séries temporais e a previsão de valores futuros;
- `Keras` (CHOLLET et al., 2015): uma das bibliotecas mais difundidas quando se trata de

criação e aperfeiçoamento de modelos de aprendizagem profunda, como redes neurais LSTM.

Durante o treinamento e otimização dos modelos não foi utilizado um valor de *seed* para controlar a aleatoriedade do processo. Essa escolha foi feita para tentar obter melhores resultados das escolhas aleatórias das combinações de quantidade de camadas, neurônios por camada e taxa de *dropout*. Os resultados obtidos nesse experimento podem ser replicados por meio dos modelos salvos nos formatos '.tf' e '.h5' (ambos da biblioteca `Keras`) disponibilizados com o restante dos códigos e base de dados no GitHub (TURQUETI, 2022).

5 RESULTADOS

Este capítulo apresenta os resultados obtidos durante a execução do trabalho, tanto a parte da análise exploratória dos dados quanto os modelos e os resultados obtidos por cada um deles.

5.1 Análise exploratória de dados

A ANP disponibiliza em seu site a média histórica semanal de sete tipos de combustíveis de postos em território brasileiro agrupados por região administrativa. Na Tabela 4 a seguir estão resumidos os preços mínimo e máximo do período das informações (limitado ao último dia de agosto de 2022) disponibilizadas pela agência:

Tabela 4 – Principais informações disponibilizadas por tipo de combustível.

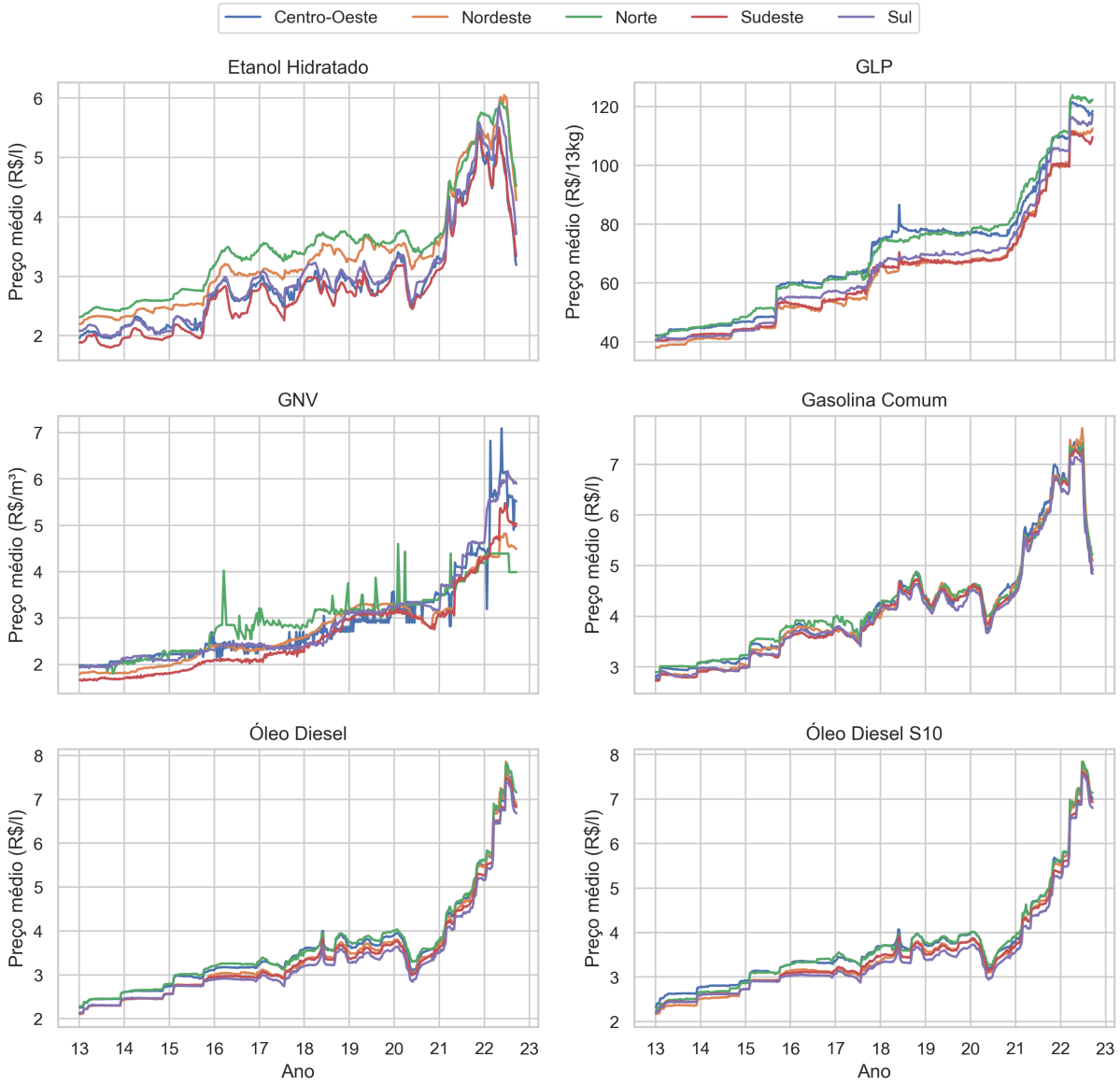
Combustível	Data do primeiro registro	Data do último registro	Preço mínimo	Preço máximo	Unidade de medida
Etanol Hidratado	05/01/2013	27/08/2022	1,800	6,050	R$/l
GLP	05/01/2013	27/08/2022	38,056	123,906	R$/13kg
GNV	05/01/2013	27/08/2022	1,645	7,090	R$/$m^3$
Gasolina Aditivada	24/10/2020	27/08/2022	4,406	7,870	R$/l
Gasolina Comum	05/01/2013	27/08/2022	2,724	7,710	R$/l
Óleo Diesel	05/01/2013	27/08/2022	2,102	7,860	R$/l
Óleo Diesel S10	05/01/2013	27/08/2022	2,164	7,840	R$/l

Fonte: Autoria própria.

A coleta das informações dos preços da gasolina aditivada foi iniciada apenas no final de 2020, diferentemente dos demais combustíveis que começaram a ter os preços analisados no início de janeiro de 2013. Além disso, tanto o GLP quanto o GNV possuem unidades de medida de preço diferentes dos demais. O primeiro tem o preço tabelado por 13kg do gás, enquanto o segundo possui o preço pelo volume ocupado pelo gás.

Visando-se observar as diferenças de preços por região dos combustíveis, é mostrada na Figura 23 o histórico de preço de revenda desde o ano de 2013, excluindo -se os dados referentes à gasolina aditivada. Na figura são apresentados dois grupos de combustíveis: i) aqueles com diferença significativa do preço médio entre as regiões (etanol hidratado, GLP e GNV) e; ii) aqueles com uma média regional com um comportamento mais próximo da média nacional (gasolina comum, óleo diesel e óleo diesel S10). No caso do GNV essa diferença de preço é explicada pela pequena quantidade de postos pesquisados, conforme mostra a Figura 24, favorecendo distorções por *outliers* durante a coleta dos dados. Já para os produtos etanol hidratado e GLP as curvas de preço com comportamentos diferentes entre as regiões são

Figura 23 – Histórico do preço de combustíveis por região do Brasil



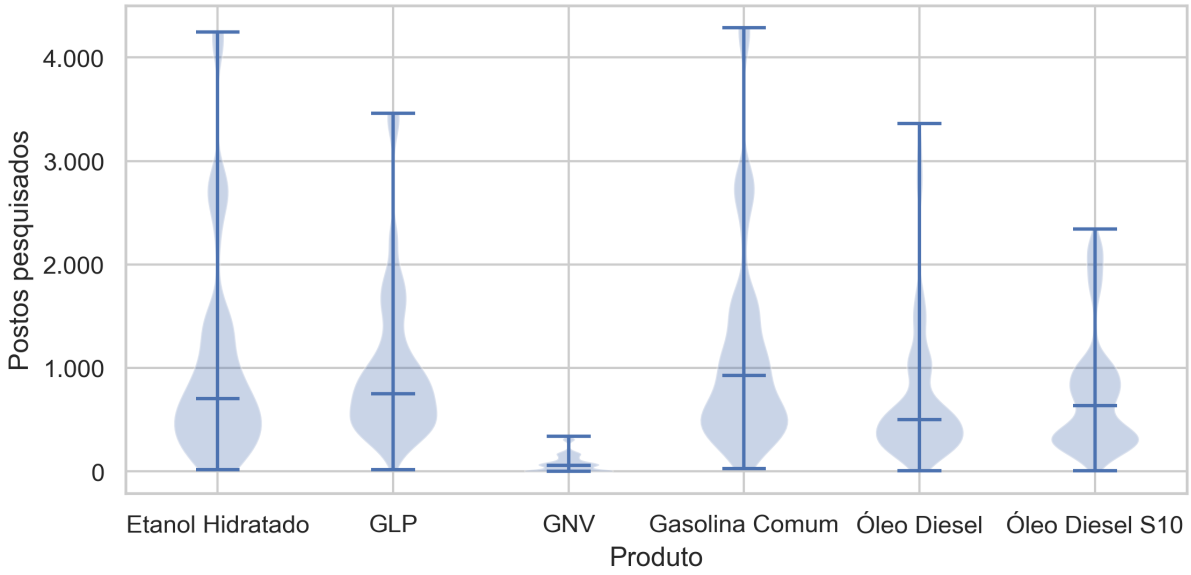
Fonte: Autoria própria.

resultados de diferentes margens médias de revenda (diferença entre os preços de distribuição e revenda).

Outro fato observável na Figura 23 é a constante queda dos preços dos combustíveis nas últimas semanas. Segundo Tavares (2022), esse resultado foi obtido após posicionamento da Comissão Europeia objetivando se fazer uma intervenção emergencial no mercado em um momento de crise energética na Europa. Desde o começo da guerra na Ucrânia, o preço do barril de petróleo do tipo *Brent* atingiu o valor de U$D 130, chegando a U$D 95 em 1º de setembro de 2022. Devido à Política de Preços de Paridade de Importação, essa variação de preço tem forte influência no valor final de revenda no Brasil.

Além disso, houve uma queda ainda mais acentuada para o etanol hidratado e a gasolina comum devido à promulgação da Lei Complementar 194/2022, cujos principais fatores

Figura 24 – Quantidade de postos pesquisados por produto agrupados por semana e região



Fonte: Autoria própria.

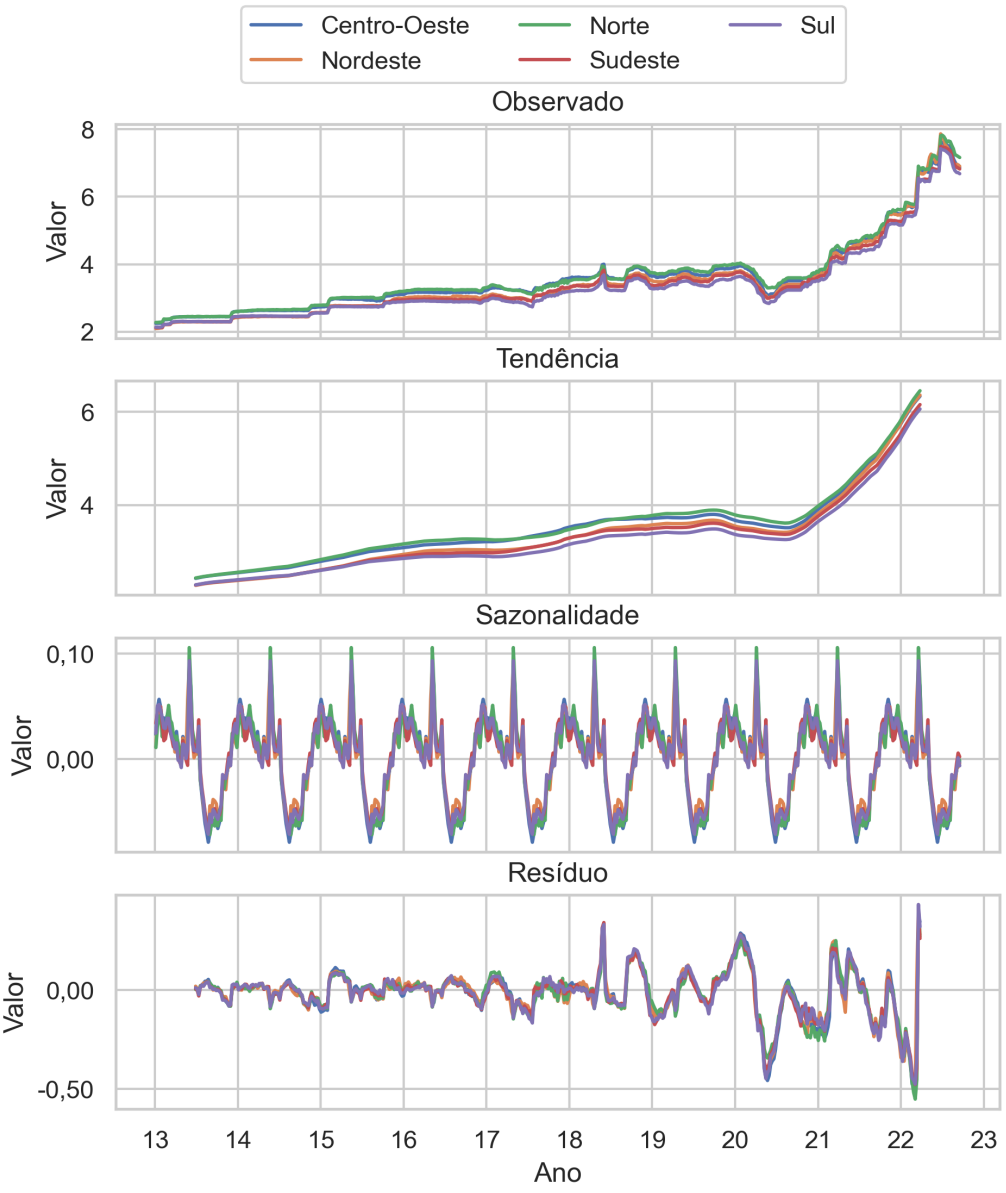
influenciadores são a desoneração dos impostos federais para ambos e a fixação do teto da alíquota de ICMS (17 a 18%) sobre combustíveis no geral e energia elétrica (RAMALHO, 2022). Ramalho (2022) também menciona como fatores a redução da base de cálculo do ICMS pelos estados com a cobrança do imposto sobre a média móvel de 60 meses para o óleo diesel, gasolina e GLP e a redução de preços da Petrobras nas refinarias, chegando a 13% para gasolina e 7,3% para o óleo diesel.

Devido às diferenças de comportamentos entre os produtos, seria necessário a construção, treinamento e validação de um modelo diferente para cada um deles. Entretanto, buscando entender qual o combustível de maior impacto, são analisados e modelados apenas os dados do óleo diesel, já que ele representa quase metade do volume vendido em 2021, conforme Tabela 1.

Uma segunda análise realizada foi referente ao comportamento da série temporal dos valores de óleo diesel. Para tal análise, o preço observado foi decomposto em: tendência, sazonalidade e resíduos. Tal decomposição foi realizada considerando-se uma série temporal aditiva com período de sazonalidade igual a 51 semanas (1 ano), cujos resultados são apresentados na Figura 25.

Como pode ser observado na Figura 25, há uma forte tendência de alta dos preços após o 2º semestre de 2020 e aumento da amplitude dos valores dos resíduos. Além disso, o componente sazonal do preço tem dois comportamentos distintos por semestre do ano: no 1º há uma queda seguida por uma forte alta relativa, enquanto no 2º há uma constante queda até a metade do período e uma recuperação desse preço na segunda metade. Apesar disso, a amplitude dos valores da sazonalidade é baixa (no máximo aproximadamente R$ 0,10) se comparada aos preços observados. Adicionalmente, percebe-se um comportamento semelhante

Figura 25 – Decomposição dos dados dó óleo diesel em tendência, sazonalidade e resíduos



Fonte: Autoria própria.

para o preço do óleo diesel nas cinco regiões brasileiras, variando apenas alguns centavos entre elas.

Para uso no treinamento e validação dos modelos preditivos serão utilizados os dados da região com maior volume de vendas desse combustível no território brasileiro em 2021. Segundo dados disponibilizados pela ANP (Agência Nacional do Petróleo, Gás Natural e Biocombustíveis) (2022b) e consolidados na Tabela 5, a região com maior representatividade é a sudeste.

Tabela 5 – Volume absoluto e relativo de óleo diesel vendido em 2021 no Brasil.

Região	**Volume absoluto**	**Volume relativo** (m^3/m^3)
Centro-Oeste	8.891.915 m^3	14,32%
Nordeste	9.834.542 m^3	15,83%
Norte	6.686.328 m^3	10,77%
Sudeste	23.631.275 m^3	38,05%
Sul	13.067.506 m^3	21,04%

Fonte: Autoria própria.

5.2 Desempenho dos Modelos preditivos

Uma vez analisados os dados de interesse, realizou-se um conjunto de experimentos com modelos preditivos treinados e testados com o histórico de preços do óleo diesel na região sudeste. Estes experimentos analisam resultados do cenário geral ao mais específico, descritos nas próximas seções.

5.2.1 Comparação de modelos via pycaret

Os experimentos foram iniciados considerando uma grande quantidade de algoritmos de previsão de séries temporais. Foi desenvolvido um experimento com hiperparâmetros padrão do `pycaret`, com horizonte de previsão igual a 8 semanas, 5 repetições na validação cruzada e com a avaliação de todos os 28 modelos disponíveis[1] na biblioteca `pycaret`. Os resultados da validação cruzada assim como o tempo gasto no treino e teste do modelo são apresentados na Tabela 6.

Avaliando-se as métricas de MAE e RMSE, o modelo com os melhores valores é o Auto ARIMA com, respectivamente, 0,3066 e 0,3345. Além disso, foi o único algoritmo responsável por aproximadamente 32% do tempo de execução pelo `pycaret`, evidenciando seu alto custo computacional. Com esta evidência, testaram-se também algumas variações do Auto ARIMA, com modelos otimizados considerando ou não a presença de variáveis exógenas.

5.2.2 SARIMA e SARIMAX

Por ser o algoritmo com os menores valores de erro nos conjuntos de treino e teste quando executado pelo `pycaret`, o SARIMA foi escolhido para tentar ser otimizado em sua versão com uso apenas do histórico de preços e também treinado com variáveis exógenas originando o SARIMAX. Com o uso da biblioteca `pmdarima` foram primeiramente encontrados os coeficientes p, d, q, P, D, Q e m que melhor se ajustavam aos dados históricos do preço do óleo diesel. Após o cálculo dos coeficientes o modelo foi treinado com os dados no intervalo já descrito na Tabela 2.

[1] Modelos indicados por * possuem dessazonalização condicional e diferenciação como parte do tratamento automático dos dados.

Tabela 6 – Erros obtidos via modelos tradicionais de aprendizado de máquina.

Modelo	**Nome do modelo**	**MAE**	**RMSE**	**Tempo (s)**
auto_arima	Auto ARIMA	0,3066	0,3345	0,68
exp_smooth	Exponential Smoothing	0,3102	0,3409	0,02
ets	ETS	0,3102	0,3410	0,02
lar_cds_dt	Least Angular Regressor*	0,3328	0,3615	0,03
omp_cds_dt	Orthogonal Matching Pursuit*	0,3328	0,3615	0,03
lr_cds_dt	Linear*	0,3328	0,3615	0,03
br_cds_dt	Bayesian Ridge*	0,3330	0,3616	0,02
theta	Theta Forecaster	0,3769	0,4050	0,02
huber_cds_dt	Huber*	0,3816	0,4098	0,03
naive	Naive Forecaster	0,3856	0,4147	0,01
arima	ARIMA	0,3879	0,4176	0,07
gbr_cds_dt	Gradient Boosting*	0,3924	0,4212	0,06
catboost_cds_dt	CatBoost Regressor*	0,3926	0,4213	0,51
et_cds_dt	Extra Trees*	0,3971	0,4260	0,14
dt_cds_dt	Decision Tree*	0,3974	0,4263	0,03
xgboost_cds_dt	Extreme Gradient Boosting*	0,3986	0,4266	0,04
rf_cds_dt	Random Forest*	0,4010	0,4296	0,11
knn_cds_dt	K Neighbors*	0,4042	0,4327	0,03
ada_cds_dt	AdaBoost*	0,4130	0,4418	0,03
ridge_cds_dt	Ridge*	0,4233	0,4578	0,04
lightgbm_cds_dt	Light Gradient Boosting*	0,7048	0,7204	0,04
croston	Croston	0,7067	0,7235	0,01
polytrend	Polynomial Trend Forecaster	1,6143	1,6214	0,01
en_cds_dt	Elastic Net*	1,6143	1,6215	0,02
lasso_cds_dt	Lasso*	1,6143	1,6215	0,03
llar_cds_dt	Lasso Least Angular Regressor*	1,6143	1,6215	0,03
grand_means	Grand Means Forecaster	2,6987	2,7033	0,01
par_cds_dt	Passive Aggressive*	6,4518	8,6620	0,04

Fonte: Autoria própria.

Utilizando agora como variáveis exógenas os históricos do preço do barril de petróleo *Brent* e da cotação do dólar americano, foi seguido o mesmo procedimento para o modelo SARIMAX. Os coeficientes calculados e os valores de erro para ambos os modelos estão, respectivamente, na Tabela 7 e Tabela 8.

Tabela 7 – Coeficientes ajustados para os modelos SARIMA e SARIMAX.

Modelo	p	d	q	P	D	Q	m
SARIMA	0	2	2	0	0	0	52
SARIMAX	1	1	0	0	0	0	52

Fonte: Autoria própria.

Tabela 8 – Erros calculados no teste dos modelos SARIMA e SARIMAX.

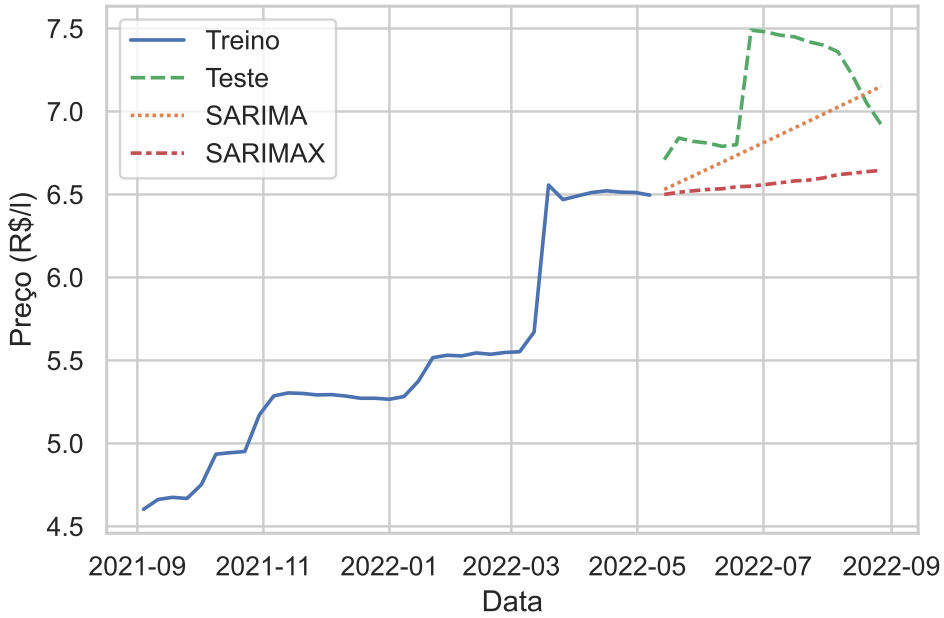
Modelo	MSE	RMSE	MAE
SARIMA	0,1570	0,3962	0,3527
SARIMAX	0,5033	0,7094	0,6766

Fonte: Autoria própria.

Conforme observado na Tabela 8, o modelo com as menores métricas de erro é o SARIMA indicando seu melhor ajuste aos dados históricos de preços. Ele possui apenas os componentes integrativo e de médias móveis, ambos de ordem 2, diferentemente do SARIMAX cujos componentes são o autorregressivo e integrativo (ambos de ordem 1). Uma possível justificativa para uma melhor ajuste de um componente integrativo de segunda ordem em vez de apenas primeira ordem é o componente não linear da tendência de preço, já ilustrada na Figura 25. Caso a tendência fosse uma função de 1º grau, uma única diferenciação conseguiria obter bons resultados.

Outro ponto destacado a ser analisado é que os coeficientes da componente sazonal (P, D, Q) de ambos os modelos foram ajustados com o valor 0. Como já mostrado na Figura 25 o componente sazonal possui uma amplitude de valor muito baixa se comparada ao preço observado, podendo ser desconsiderada pelos modelos.

Figura 26 – Resultados previstos pelos modelos SARIMA e SARIMAX



Fonte: Autoria própria.

Apesar de serem usados para treinamento dos modelos todos os dados desde o período inicial relatado na Tabela 2, para uma melhor comparação dos modelos nos dados de teste, a Figura 26 apresenta apenas os dados de treino a partir de set/2021 e aqueles previstos pelos

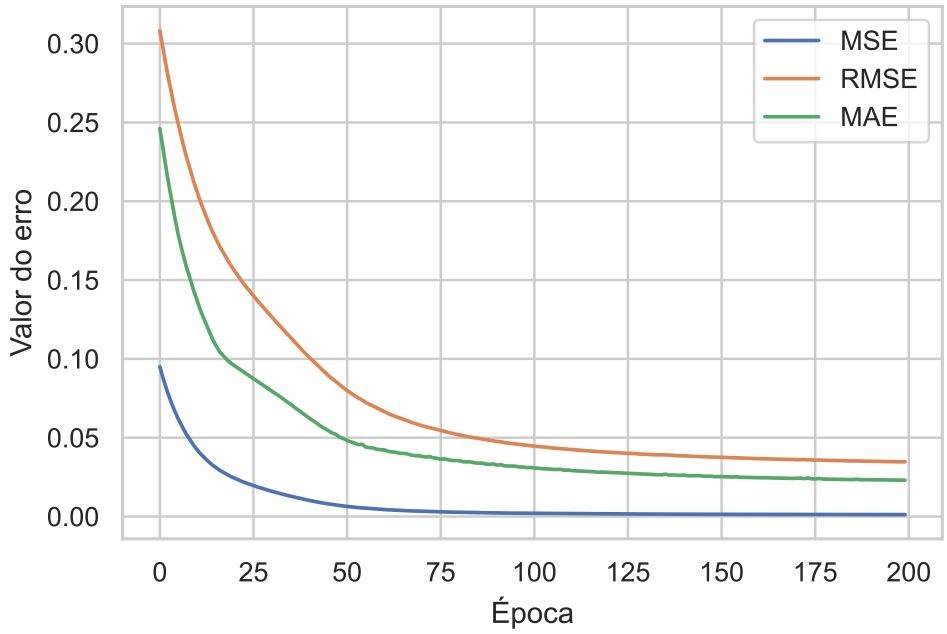
modelos. Corroborando os erros da Tabela 8, a Figura 26 mostra uma maior aproximação das previsões do modelo SARIMA com dados reais de teste para o período. Além disso, é possível verificar que para o SARIMAX existem pequenas variações de preço ao longo das semanas devido às variáveis exógenas, enquanto para o SARIMA os resultados possuem o comportamento linear.

5.2.3 Modelos de aprendizado profundo recorrente - LSTMs

Como já explicitado anteriormente na Subseção 2.3.2, para problemas de processamento de dados sequenciais é comum o uso de redes neurais recorrentes. Dessa forma, foi avaliado o uso de um modelo de LSTM com uma única célula LSTM, e camada de saída densa com 8 neurônios (um neurônio para cada valor futuro previsto). A camada de saída tem esse tamanho porque corresponde à quantidade de dados que o modelo irá prever: um intervalo de janela de 8 semanas.

A LSTM foi codificada através a biblioteca `Keras`, que implementa algoritmos de DL em Python. Esse modelo foi treinado por 200 épocas, usando todos os hiperparâmetros com valores padrão (*default*), visando reduzir o erro tipo MSE. A evolução dos erros ao longo das épocas é mostrada na Figura 27.

Figura 27 – Evolução dos erros no treinamento da LSTM simples

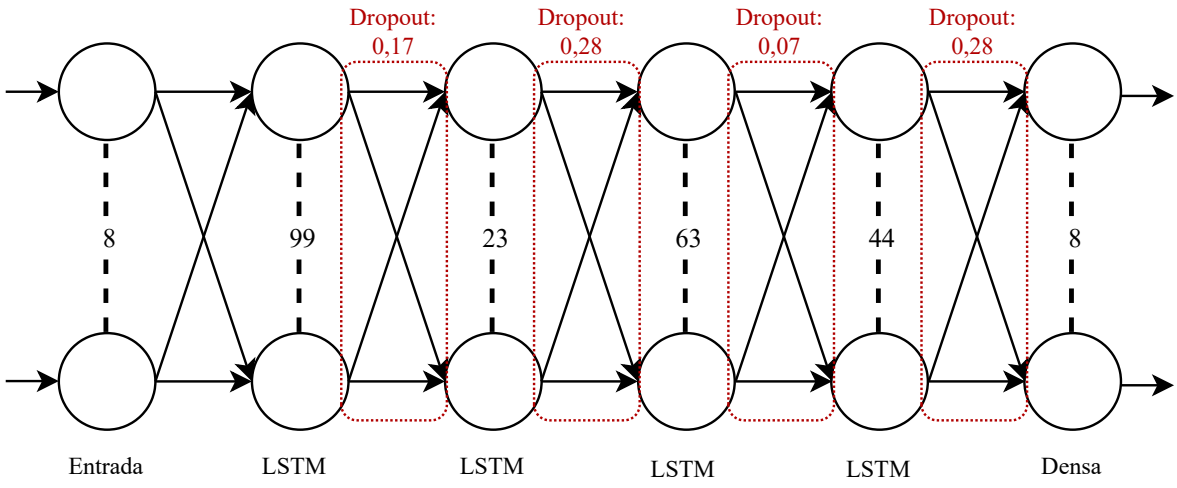


Fonte: Autoria própria.

Conforme observado na figura, a partir da 100ª época a redução do MSE passa a ser quase imperceptível. Com o prosseguimento do treinamento até o final das 200 épocas é possível que esse modelo sofra com um problema de sobreajuste (ou *overfitting*), não conseguindo prever corretamente os preços no intervalo de tempo de teste

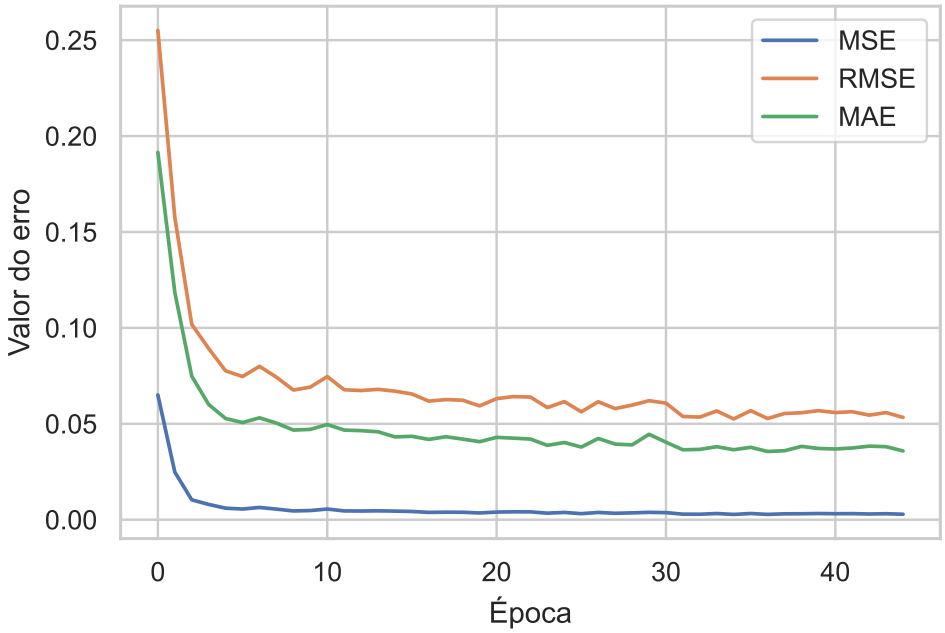
Desta maneira, para se obter melhores resultados, foi treinado outro modelo com um processo um pouco mais robusto. O critério de parada para treinamento foi definido como inalteração ou aumento do erro. Além disso, realizou-se um processo de ajuste de hiperparâmetros via *Random Search* (RS) com validação cruzada visando reduzir o erro do tipo MSE, cuja estrutura final com a quantidade de células, número de camadas, e taxas de *droupout* selecionadas pode ser visualizada na Figura 28.

Figura 28 – Estrutura final do modelo LSTM otimizado



Fonte: Autoria própria.

Figura 29 – Evolução dos erros no treinamento da LSTM otimizada

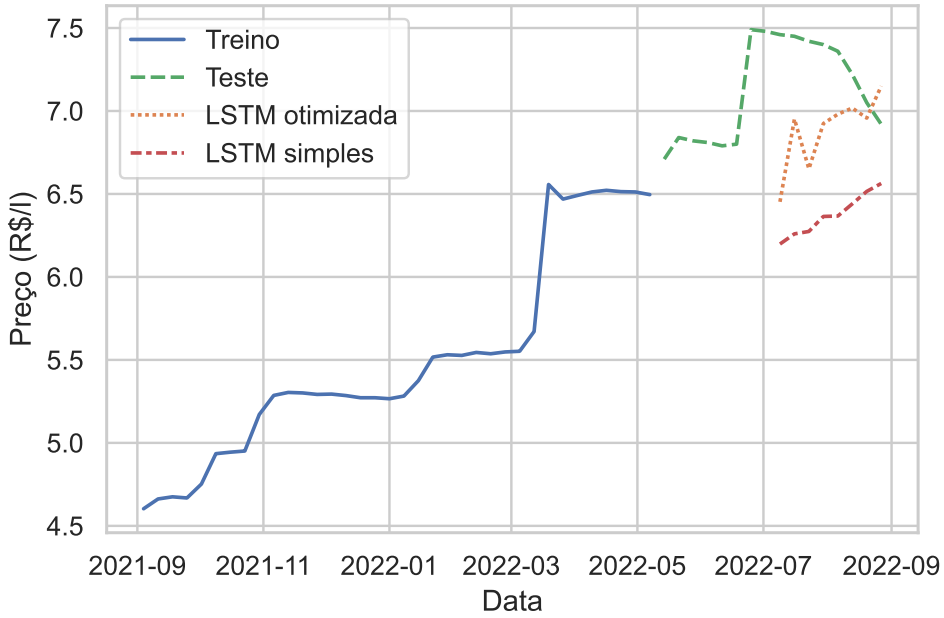


Fonte: Autoria própria.

Assim como no modelo de LSTM simples, na indução do modelo de LSTM ajustado (otimizado) buscou-se reduzir o erro MSE durante o treinamento. A evolução das três métricas de erro neste cenário é apresentada na Figura 29. Conforme ilustra a Figura 29, após uma certa quantidade de épocas sem diminuição do erro MSE o algoritmo interrompe o treinamento da rede neural recorrente para não haver sobreajuste com o conjunto de dados de treino. Além de melhorar os resultados, esse procedimento torna também o treinamento da LSTM mais rápido devido à menor quantidade de épocas.

Os dois modelos de LSTM: i) simples, com os hiperparâmetros padrão; e ii) otimizada, com os hiperparâmetros ajudados; foram comparadas realizando predições em um mesmo conjunto de dados de teste. Foram usadas como entrada dados com 8 semanas de preço do óleo diesel no sudeste para serem preditos os preços das próximas 8 semanas. Os resultados dessa comparação são apresentados na Figura 30.

Figura 30 – Resultados previstos pelos modelos simples e otimizado de LSTM



Fonte: Autoria própria.

Como pode ser constatado pelos resultados da Figura 30, a LSTM simples apresentou um comportamento muito aquém da LSTM otimizada, tanto pela distância maior da curva predita em relação aos valores reais de teste, quanto o comportamento mais linear e invariável às mudanças recentes usadas como *inputs*. Essa melhora após a otimização pode ser atribuída não só pelo aumento das unidades de processamento (camadas e células) mas também pela interrupção do treino antes do fim das 200 épocas, evitando assim que o modelo perdesse a sua capacidade de generalização. Ratificando o gráfico com os preços calculados por ambos os modelos, a Tabela 9 apresenta os três tipos de erro para as 8 semanas de teste com a LSTM otimizada obtendo erros menores dos três casos. Logo,

Tabela 9 – Erros calculados no teste dos modelos simples e otimizado LSTM.

Modelo	MSE	RMSE	MAE
LSTM simples	0,9252	0,9618	0,9122
LSTM otimizada	0,2908	0,5393	0,4568

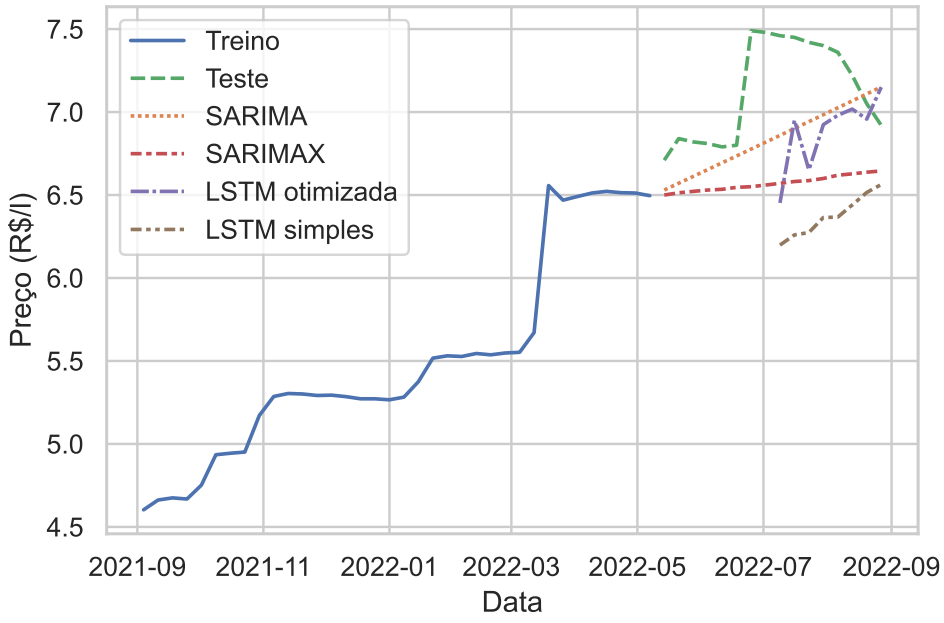
Fonte: Autoria própria.

5.3 Escolha do melhor modelo

Visando auxiliar a escolha do melhor modelo, foram colocadas as previsões dos quatro modelos treinados na Figura 31 assim como os valores de seus erros na Tabela 10. Entretanto, para se adotar um modelo como passível de solucionar um problema, ele não deve ter apenas os menores valores de erro, mas também deve conseguir prever a dinâmica do estudo.

Analisando exclusivamente os erros, o melhor modelo seria o SARIMA já que ele apresenta um menor valor para as três métricas analisadas. Porém, como já mostrado na Figura 31 as previsões feitas por ele apresentam um comportamento linear, destoando completamente do comportamento real dos dados históricos.

Figura 31 – Resultados previstos por todos os modelos



Fonte: Autoria própria.

Após o SARIMA, o modelo com os menores valores de erros é a LSTM otimizada. Analisando suas previsões, percebe-se que elas começam com uma distância de aproximadamente R$ 1,00 dos dados de teste e ao final das 8 semanas os preços convergem para valores muito próximos. Além disso, ele conseguiu obter resultados linearmente independentes, característica importante ao lidar com séries temporais com alta volatilidade. Por fim, apesar de ainda não

Tabela 10 – Erros calculados no teste de todos os modelos.

Modelo	**MSE**	**RMSE**	**MAE**
SARIMA	0,1570	0,3962	0,3527
SARIMAX	0,5033	0,7094	0,6766
LSTM simples	0,9252	0,9618	0,9122
LSTM otimizada	0,2908	0,5393	0,4568

Fonte: Autoria própria.

ser possível utilizar seus resultados para tomada de decisões ou previsão precisa de custos, o melhor modelo treinado foi o de LSTM otimizado por *RandomizedSearchCV*.

6 CONCLUSÃO

Esse trabalho buscou induzir modelos para predição dos preços futuros do óleo diesel na região sudeste do Brasil a partir, principalmente, do uso de dados de preços históricos fornecidos pela ANP. Além disso, explorou-se a dinâmica de preços dos combustíveis vendidos no país.

Os resultados obtidos pelas soluções propostas apresentaram futuros preços plausíveis, mas distantes dos preços reais. O modelo com melhor adaptação aos dados históricos foi o da LSTM otimizada via *RandomizedSearchCV*, cujos erros MSE, RMSE e MAE para um período de 8 semanas foram, respectivamente, 0,2908, 0,5393 e 0,4568. Apesar dos erros não tão altos, o modelo ainda deve ser melhorado para ser utilizado para previsões assertivas de gastos logísticos por empresas.

6.1 Limitações

Mesmo com os avanços nas detecções dos padrões pelos modelos, eles ainda devem ser melhorados para preverem preços mais próximos dos reais. No método *RandomizedSearchCV*, por exemplo, não é possível escolher uma quantidade muito grande de iterações, já que ao definir um número muito limitado de *workers* para processar os dados o tempo decorrido aumenta muito e ao ilimitar a quantidade de *workers* ocorre um consumo crescente de memória RAM do computador (chegando a ser superior a 20 GB e interrompendo o processamento em alguns casos). Pelo fato desse método escolher combinações aleatórias de quantidade de camadas e seus neurônios e de taxas de *dropout* pode-se dizer que em parte é necessário sorte para que os N valores escolhidos pelo algoritmo sejam de fato ótimos para o problema estudado.

Outra dificuldade enfrentada para a conclusão do estudo foram os diversos problemas de incompatibilidade durante a tentativa de uso da biblioteca `pycaret`. Mesmo criando um ambiente virtual limpo, e instalando apenas as dependências indicadas no arquivo "requirements.txt", elas ainda conflitavam entre si devido a restrições de versão. Por ser uma biblioteca nova, ela ainda está em constante evolução, podendo atingir maiores níveis de maturidade daqui a algum tempo.

6.2 Trabalhos Futuros

Além da busca de novas combinações aleatórias de número de camadas e células da LSTM com diferentes taxas de *dropout*, uma abordagem possível seria utilizar a análise multivariada em conjunto com o aprendizado profundo recorrente. Algumas variáveis exógenas sugeridas para isso são:

- Preço do barril de petróleo *Brent*;

- Cotação do dólar em reais;
- Preço das ações ordinárias e preferenciais da Petrobras (respectivamente PETR3 e PETR4);
- Indicadores de troca de gestão Petrobras;
- Indicadores de eventos políticos do governo;
- indicadores de crise econômica no Brasil e no mundo.

É possível também tentar utilizar diferentes períodos de treino e teste e variar a quantidade de semanas utilizadas como entrada do modelo e quantas semanas o modelo deve prever.

Trabalhos na área de AM também utilizam o processamento de linguagem natural de dados obtidos em redes sociais para servirem de variáveis de modelos preditivos, por exemplo, uma análise de sentimento de *tweets* sobre preços de criptomoedas (KRAAIJEVELD; SMEDT, 2020; NAEEM et al., 2021). Nesse caso poderia ser feito algo semelhante para se obter mais uma variável exógena.

Outra possibilidade para trabalhos futuros seria tentar utilizar outros métodos de otimização de modelos além da busca aleatória com validação cruzada, como algoritmo genético, inteligência de enxame, ou otimização bayesiana. Por fim, outros algoritmos clássicos de aprendizado de máquina apresentados na Subseção 5.2.1 além do ARIMA poderiam ser também melhores otimizados para se obter previsões mais precisas.

6.3 Considerações Finais

Por fim, o presente trabalho pode ser considerado um primeiro passo no objetivo de se obter preços futuros de óleo diesel no país. Com resultados parcialmente positivos é possível otimizar ainda mais as soluções aqui apresentadas.

Referências

ABDOLLAHI, H.; EBRAHIMI, S. B. A new hybrid model for forecasting brent crude oil price. **Energy**, v. 200, p. 117520, 2020. ISSN 0360-5442. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0360544220306277>. Citado na página 39.

Agência O Globo. **Preço da gasolina sobe pela sexta semana seguida e mais lugares têm litro a R$ 8**. 2021. Disponível em: <https://economia.ig.com.br/2021-11-12/preco-gasolina-sobe-sexta-semana-seguida.html>. Acesso em: 12 de novembro de 2021. Citado na página 13.

ALI, M. **PyCaret: An open source, low-code machine learning library in Python**. Canada, 2020. PyCaret version 1.0.0. Disponível em: <https://www.pycaret.org>. Citado na página 46.

ALMEIDA, P.; ARAÚJO, T. **Alta do diesel faz preços subirem 10% nas Centrais de Abastecimentos | CNN Brasil**. 2022. Disponível em: <https://www.cnnbrasil.com.br/business/alta-do-diesel-faz-precos-subirem-10-nas-centrais-de-abastecimentos/>. Acesso em: 18 de outubro de 2022. Citado na página 13.

ALVARENGA, H. **Matriz de transportes do Brasil à espera dos investimentos**. 2020. Disponível em: <https://www.ilos.com.br/web/tag/matriz-de-transportes>. Acesso em: 13 de novembro de 2021. Citado na página 13.

ANP (Agência Nacional do Petróleo, Gás Natural e Biocombustíveis). **Série histórica do Levantamento de Preços**. 2022. Disponível em: <https://www.gov.br/anp/pt-br/assuntos/precos-e-defesa-da-concorrencia/precos/precos-revenda-e-de-distribuicao-combustiveis/serie-historica-do-levantamento-de-precos>. Acesso em: 15 de maio de 2022. Citado na página 42.

ANP (Agência Nacional do Petróleo, Gás Natural e Biocombustíveis). **Vendas de derivados petróleo e etanol (metros cúbicos) 1990-2022**. 2022. Disponível em: <https://www.gov.br/anp/pt-br/centrais-de-conteudo/dados-abertos/arquivos/vdpb/vendas-derivados-petroleo-e-etanol/vendas-derivados-petroleo-etanol-m3-1990-2022.csv>. Acesso em: 19 de julho de 2022. Citado 3 vezes nas páginas 40, 42 e 51.

AROUSSI, R. et al. **yfinance: Download market data from Yahoo! Finance's API**. 2017. Disponível em: <https://github.com/ranaroussi/yfinance>. Acesso em: 07 de outubro de 2022. Citado na página 42.

ATWAN, T. **Time Series Analysis with Python Cookbook: Practical Recipes for Exploratory Data Analysis, Data Preparation, Forecasting, and Model Evaluation**. London: Packt Publishing, Limited, 2022. ISBN 9781801075541. Citado 4 vezes nas páginas 20, 21, 22 e 23.

BBC News Brasil. **Preço da gasolina: o que entra nessa conta?** 2021. Disponível em: <https://www.bbc.com/portuguese/internacional-58270223>. Acesso em: 18 de outubro de 2022. Citado na página 13.

CARASSAI, A. F. et al. Análise multivariada aplicada ao preço do etanol hidratado praticado no brasil. **Revista Científica Agropampa**, v. 1, n. 1, p. 22–35, 2021. Citado na página 40.

CARVALHO, I. **Por que o diesel é proibido para veículos de passeio no Brasil?** 2018. Disponível em: <https://quatrorodas.abril.com.br/noticias/por-que-o-diesel-e-proibido-para-veiculos-de-passeio-no-brasil/>. Acesso em: 24 de julho de 2022. Citado na página 40.

CHOLLET, F. et al. **Keras**. 2015. <https://keras.io>. Acesso em: 07 de outubro de 2022. Citado 2 vezes nas páginas 36 e 46.

CIPRA, T. **Time Series in Economics and Finance**. Switzerland: Springer International Publishing, 2020. ISBN 9783030463472. Citado 3 vezes nas páginas 24, 25 e 26.

FOLHAPRESS. **Gasolina por metade do preço na Argentina gera fila de brasileiros**. 2021. Disponível em: <https://www.jj.com.br/economia/2021/11/140047-gasolina-por-metade-do-preco-na-argentina-gera-fila-de-brasileiros.html>. Acesso em: 12 de novembro de 2021. Citado na página 13.

GAUTO, M. A. et al. **Petróleo e Gás: Princípios de Exploração, Produção e Refino**. Bookman Editora, 2016. (Tekne). ISBN 9788582604021. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/9788582604021/>. Acesso em: 03 de outubro de 2022. Citado 4 vezes nas páginas 15, 16, 17 e 18.

GAUTO, M. A.; ROSA, G. **Química Industrial: Série Tekne**. Bookman Editora, 2013. (Tekne). ISBN 9788565837613. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/9788565837613/>. Acesso em: 03 de outubro de 2022. Citado 4 vezes nas páginas 15, 17, 18 e 19.

GÉRON, A. **Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems**. O'Reilly Media, 2019. ISBN 9781492032618. Disponível em: <https://books.google.com.br/books?id=HHetDwAAQBAJ>. Citado 5 vezes nas páginas 26, 32, 33, 34 e 35.

Gestra. **Tanque Flash**. 2022. Disponível em: <https://www.gestra.com/global/pt-GES/products/system-and-packaged-solutions/flash-vessel>. Acesso em: 20 de outubro de 2022. Citado na página 19.

GUPTA, N.; NIGAM, S. Crude oil price prediction using artificial neural network. **Procedia Computer Science**, v. 170, p. 642–647, 2020. ISSN 1877-0509. The 11th International Conference on Ambient Systems, Networks and Technologies (ANT) / The 3rd International Conference on Emerging Data and Industry 4.0 (EDI40) / Affiliated Workshops. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1877050920305913>. Citado na página 39.

HARRIS, C. R. et al. Array programming with NumPy. **Nature**, Springer Science and Business Media LLC, v. 585, n. 7825, p. 357–362, set. 2020. Disponível em: <https://doi.org/10.1038/s41586-020-2649-2>. Citado na página 46.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. **Neural computation**, MIT Press, v. 9, n. 8, p. 1735–1780, 1997. Citado na página 34.

HUNTER, J. D. Matplotlib: A 2d graphics environment. **Computing in Science & Engineering**, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007. Citado na página 46.

KRAAIJEVELD, O.; SMEDT, J. D. The predictive power of public twitter sentiment for forecasting cryptocurrency prices. **Journal of International Financial Markets, Institutions and Money**, Elsevier, v. 65, p. 101188, 2020. Citado na página 61.

LAFRATTA, C. **Variação do dólar: entenda por que ele sobe e desce tanto**. 2020. Disponível em: <https://blog.nubank.com.br/por-que-dolar-sobe-e-desce/>. Acesso em: 18 de outubro de 2022. Citado na página 13.

LAZZERI, F. **Machine Learning for Time Series Forecasting with Python**. United States of America: Wiley, 2020. ISBN 9781119682363. Citado na página 26.

NAEEM, M. A. et al. Does twitter happiness sentiment predict cryptocurrency? **International Review of Finance**, Wiley Online Library, v. 21, n. 4, p. 1529–1538, 2021. Citado na página 61.

NIELSEN, A. **Practical Time Series Analysis: Prediction with Statistics and Machine Learning**. United States of America: O'Reilly Media, 2019. ISBN 9781492041627. Citado 3 vezes nas páginas 23, 32 e 33.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Citado na página 46.

PEIXEIRO, M. **Time Series Forecasting in Python**. United States of America: Manning, 2022. ISBN 9781617299889. Citado 6 vezes nas páginas 27, 28, 29, 30, 31 e 32.

PETROBRAS. **Exploração e Produção de Petróleo e Gás**. 2022. Disponível em: <https://petrobras.com.br/pt/nossas-atividades/areas-de-atuacao/exploracao-e-producao-de-petroleo-e-gas/>. Acesso em: 20 de outubro de 2022. Citado na página 17.

RAMALHO, A. **Gasolina mais barata gera voto? Como queda dos preços afeta Bolsonaro nas pesquisas eleitorais**. 2022. Disponível em: <https://epbr.com.br/gasolina-mais-barata-gera-voto-como-queda-dos-precos-afeta-bolsonaro-nas-pesquisas-eleitorais/>. Acesso em: 28 de setembro de 2022. Citado na página 50.

ROSSUM, G. V.; DRAKE, F. L. **Python 3 Reference Manual**. Scotts Valley, CA: CreateSpace, 2009. ISBN 1441412697. Citado na página 46.

SEABOLD, S.; PERKTOLD, J. statsmodels: Econometric and statistical modeling with python. In: **9th Python in Science Conference**. [S.l.: s.n.], 2010. Citado 5 vezes nas páginas 20, 21, 23, 28 e 46.

SMITH, T. G. et al. **pmdarima: ARIMA estimators for Python**. 2017–. Disponível em: <http://www.alkaline-ml.com/pmdarima>. Acesso em: 15 de maio de 2022. Citado na página 46.

SOMMA Investimentos. **Quais são os fatores que influenciam o preço do petróleo?** 2021. Disponível em: <https://www.sommainvestimentos.com.br/quais-sao-os-fatores-que-influenciam-o-preco-do-petroleo/>. Acesso em: 18 de outubro de 2022. Citado na página 13.

SOUSA, A. R. d. S. et al. **Análise de Séries Temporais**. Grupo A, 2021. (Universitária). ISBN 9786556902876. Disponível em: <https://integrada.minhabiblioteca.com.br/#/books/9786556902876/>. Acesso em: 04 de outubro de 2022. Citado na página 20.

TAVARES, N. **Petrobras anuncia nova redução de R$ 0,25 no preço da gasolina**. 2022. Disponível em: <https://motor1.uol.com.br/news/607827/petrobras-reducao-preco-gasolina-setembro/>. Acesso em: 25 de setembro de 2022. Citado na página 49.

The pandas development team. **pandas-dev/pandas: Pandas**. Zenodo, 2020. Disponível em: <https://doi.org/10.5281/zenodo.3509134>. Citado na página 46.

TURQUETI, H. M. **Diesel fuel price forecasting**. 2022. Disponível em: <https://github.com/hturqueti/diesel-fuel-price-forecasting>. Acesso em: 23 de outubro de 2022. Citado na página 47.

UDOP. **Bacia de Campos: Onde o Brasil virou referência mundial em águas profundas**. 2021. Disponível em: <https://www.udop.com.br/noticia/2021/09/24/bacia-de-campos-onde-o-brasil-virou-referencia-mundial-em-aguas-profundas.html>. Acesso em: 20 de outubro de 2022. Citado na página 17.

UROLAGIN, S.; SHARMA, N.; DATTA, T. K. A combined architecture of multivariate lstm with mahalanobis and z-score transformations for oil price forecasting. **Energy**, v. 231, p. 120963, 2021. ISSN 0360-5442. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0360544221012111>. Citado na página 39.

VILELA, P. R. **Presidente critica política que atrela preço dos combustíveis ao dólar**. 2022. Disponível em: <https://agenciabrasil.ebc.com.br/economia/noticia/2021-10/presidente-critica-politica-que-atrela-preco-dos-combustiveis-ao-dolar>. Acesso em: 03 de abril de 2022. Citado na página 13.

WASKOM, M. L. seaborn: statistical data visualization. **Journal of Open Source Software**, The Open Journal, v. 6, n. 60, p. 3021, 2021. Disponível em: <https://doi.org/10.21105/joss.03021>. Citado na página 46.

WOODWARD, W. A.; SADLER, B. P.; ROBERTSON, S. **Time Series for Data Science: Analysis and Forecasting**. London: CRC Press, 2022. (Chapman & Hall/CRC Texts in Statistical Science). ISBN 9780367537944. Citado na página 23.