UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus Experimental de Rosana

SAMANTHA VIEIRA LANZELOTTI

Análise Comparativa de Ferramentas do Aprendizado de Máquina para Predição da Produção de Gasolina e Etanol

Rosana - SP
2022

Samantha Vieira Lanzelotti

Análise Comparativa de Ferramentas do Aprendizado de Máquina para Predição da Produção de Gasolina e Etanol

Trabalho de Conclusão de Curso apresentado à Coordenadoria de Curso de Engenharia de Energia do Campus Experimental de Rosana, Universidade Estadual Paulista, como parte dos requisitos para obtenção do diploma de Graduação em Engenharia de Energia.

Orientador(a): Kleber Rocha de Oliveira

Rosana - SP
2022

L297a

Lanzelotti, Samantha Vieira

Análise comparativa de ferramentas do aprendizado de máquina para predição da produção de gasolina e etanol / Samantha Vieira Lanzelotti. -- Rosana, 2022

43 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Engenharia de Energia) - Universidade Estadual Paulista (Unesp), Faculdade de Engenharia e Ciências, Rosana

Orientador: Kleber Rocha da Silva

1. Combustível. 2. Aprendizado de Máquina. 3. Energia. 4. Predição. 5. Série temporal. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Engenharia e Ciências, Rosana. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Campus Experimental de Rosana

SAMANTHA VIEIRA LANZELOTTI

ESTE TRABALHO DE GRADUAÇÃO FOI JULGADO ADEQUADO COMO
PARTE DO REQUISITO PARA A OBTENÇÃO DO DIPLOMA DE
"**GRADUADO EM ENGENHARIA DE ENERGIA**"

APROVADO EM SUA FORMA FINAL PELO CONSELHO DE CURSO DE GRADUAÇÃO EM ENGENHARIA DE ENERGIA

Prof. Dr. Leandro Ferreira Pinto
Coordenador

BANCA EXAMINADORA:

Prof. Dr. KLEBER ROCHA DE OLIVEIRA
Orientador/UNESP-Rosana

Profª. Drª. LETÍCIA SABO BOSCHI
UNESP-Rosana

Profª. Drª. MARILAINE COLNAGO
Membro Externo

Dezembro 2022

Dedico este trabalho em especial a Deus, aos meus pais e minha irmã.

AGRADECIMENTOS

Primeiramente gostaria de agradecer a Deus que me colocou nesse caminho e esteve presente em todos os momentos. Me deu forças, sabedoria e iluminou toda minha trajetória.

A todos os meus familiares, que sempre se mostraram presentes e me encorajaram a buscar o meu caminho nos estudos, ainda que em uma cidade distante. Minhas duas avós, pilares das duas famílias, símbolos da experiência e sabedoria, vossos amor e dedicação semearam o fruto do que hoje represento e o que conquistei. Em especial, não tenho palavras e nem como retribuir aos meus pais, por conduzirem o ser que me tornei, pela dedicação para fornecer a melhor educação, estudos, apoio, incentivos, suporte, amor e materiais, em que muitas das vezes deixaram seus sonhos de lado, para que eu conquistasse todos os meus objetivos. A minha irmã, que em toda a vida foi a minha melhor amiga e companheira, sempre cuidou, se preocupou e me encorajou em todos os passos. O que sempre me motiva é olhar para trás, recapitular todos os momentos, e ver dois jovens com suas duas filhas, que mesmo em meio a tantas dificuldades, não permitiram faltar nada, principalmente o amor e afeto. Graça, Sidney e Sthefany, minha eterna gratidão, todas as minhas vitórias e conquistas são por vocês.

Aos meus amigos, em especial àqueles que pelo companheirismo, apoio e carinho tornaram essa jornada mais leve e fácil, muitas vezes fazendo papel de um familiar, e que também, estiveram presentes nos momentos de alegria e tristeza. Guardarei para sempre esses anos vivenciados em Primavera, em que vocês tiveram ao meu lado, tanto na sala de aula, quanto nos almoços, churrascos, festas, nos esportes e lazeres. Quero também homenagear e deixar meu apreço aos colegas que participaram junto a mim no Centro Acadêmico, GECET e VISER, são pessoas extraordinárias que compartilharam muitas de suas experiências ao longo da graduação.

Meus queridos orientadores e amigos, Marilaine e Wallace, deixo meu agradecimento e admiração, por terem me acolhido ao grupo e na vida de vocês, que ao longo da graduação depositaram confiança, sinceridade e respeito ao meu trabalho e estudos. Ao meu orientador Kleber, agradeço por sua paciência e ajuda durante o meu trabalho final.

Por fim, não poderia deixar de agradecer e declarar todo o meu carinho aos professores, funcionários e colegas da UNESP do campus de Rosana, que participaram e proporcionaram a minha realização no curso de Engenharia de Energia.

“É possível que haja tempos difíceis, mas as dificuldades que você encontra irão torná-la mais determinada para alcançar os seus objetivos e ganhar contra todas as probabilidades.”

Martha.

RESUMO

O Brasil tem papel de destaque no cenário mundial de combustíveis e biocombustíveis, e influência tanto nas movimentações externas quanto internas do mercado e do setor de energia. O uso da inteligência artificial e seus subcampos são peças fundamentais para tomadas de decisões mais objetivas e eficientes, além de propor estratégias matematicamente mais vantajosas. Dessa forma, este projeto apresentou, por intermédio de metodologias computacionais, um estudo da comparação entre os métodos aplicados para predição da produção média (m^3) dos combustíveis mais utilizados no estado de São Paulo, região de maior demanda e consumo do país. Para tal, foram estudadas e empregadas tanto ferramentas de Análise Exploratória de Dados (AED) como modelos de Aprendizado de Máquina, sendo eles: Florestas Aleatórias (*Random Forest*), Redes Neurais Artificiais e ARIMA. Os modelos foram moldados e validados a partir do cruzamento de diferentes bases de dados, possibilitando o desenvolvimento de estratégias e novo aparato computacional para os agentes do mercado de energia. Os resultados obtidos foram de uma ferramenta computacional consistente, com bons níveis de assertividade pelos modelos, sendo o *Random Forest* o que mais se aproximou dos valores reais da produção (m^3) da gasolina, cujo MAPE foi de 5,67%, enquanto para a produção (m^3) do etanol, o modelo de série temporal - ARIMA - atingiu menor porcentagem de erro, equivalente a 5,88%. Sendo assim, a ferramenta é capaz de dar suporte a novos estudos de análises preditivas de Aprendizado de Máquina e agentes do setor de combustíveis.

PALAVRAS-CHAVE: Energia. Aprendizado de Máquina. Combustíveis. Séries temporais.

ABSTRACT

Brazil has a prominent role in the world scenario of fuels and biofuels, and influence both in the external and internal movements of the market and the energy sector. The use of artificial intelligence and its subfields are fundamental for making more objective and efficient decisions, in addition to proposing mathematically more advantageous strategies. Thus, this project presented, through computational methodologies, a study of the comparison between the methods applied to predict the average production (m^3) of the most used fuels in the state of São Paulo, the region with the highest demand and consumption in the country. For this purpose, both Exploratory Data Analysis (AED) and Machine Learning models were studied and used, namely: Random Forests, Artificial Neural Networks and ARIMA. The models were molded and validated from the intersection of different databases, enabling the development of strategies and new computational apparatus for energy market agents. The results obtained were from a consistent computational tool, with good levels of assertiveness for the models, with Random Forest being the one that came closest to the actual production values (m^3) of gasoline, whose MAPE was 5.67%, while for the production (m^3) of ethanol, the time series model - ARIMA - reached the lowest percentage of error, equivalent to 5.88%. Therefore, the tool is able to support new studies of predictive analysis of Machine Learning and agents in the fuel sector.

KEYWORDS: Energy. Machine Learning. Fuels. Time series.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SÍMBOLOS

ABNT	Associação Brasileira de Normas Técnicas
AED	Análise Exploratória de Dados
AM	Aprendizado de Máquina
ANEEL	Agência Nacional de Energia Elétrica
ANP	Agência Nacional do Petróleo
AR	Auto Regressive
ARIMA	Auto-regresive integrated moving average
BD	Banco de Dados
BEN	Balanço Energético Nacional
BNB	Banco do Nordeste
EPE	Empresa de Pesquisa Energéticas
IA	INTELIGÊNCIA ARTIFICIAL
IBP	Instituto Brasileiro de Petróleo e Gás
KDD	Processo de Descoberta de Conhecimento
MAPE	PERCENTUAL ABSOLUTO MÉDIO
ML	Machine Learning
OPEP	Organização dos Países Exportadores de Petróleo
PETROBRAS	Petróleo Brasileiro S.A.
PIB	Produto Interno Bruto
RF	Random Forest
RNA	Redes Neurais Artificiais
SP	São Paulo

SUMÁRIO

1. INTRODUÇÃO

O Brasil tem papel de destaque no cenário mundial de combustíveis e biocombustíveis, sendo o 9º maior produtor de petróleo (IBP - Instituto Brasileiro de Petróleo e Gás, 2020) e o 2° maior produtor de etanol do mundo (VIDAL, 2020). Em 2020, a indústria de petróleo e gás correspondiam a 13% do PIB nacional e 50% da oferta interna de energia (ANP - Agência Nacional do Petróleo, Gás Natural e Biocombustíveis, 2020). Todos esses fatos são indicadores da importância que o setor possui na economia e nas tomadas de decisões internas e externas do país.

A pandemia da Covid-19 foi um episódio recente que desestabilizou a curva de conduta do mercado mundial. Obteve como resposta da OPEP uma restrição na produção da gasolina a fim de manter os preços praticados. O ocorrido inesperado, somando-se a falta de planejamento de muitos países, gerou um aumento nos preços e queda na produção de combustíveis. O Brasil serve de exemplo, pois como resultado passou por uma das maiores crises de combustível de sua história, e ainda sofre resquícios (OLIVEIRA; LUZ, 2021).

Eventualidades acontecem incessantemente, e por meio da história, ferramentas foram criadas para prevê-las e solucioná-las. A maior parte do mercado de energia se encontra em um contexto benéfico, capaz de apresentar respostas mais eficientes, objetivas e imparciais, além de propor estratégias matematicamente mais vantajosas para os problemas, por meio da Inteligência Artificial (IA), visto que o governo brasileiro e seus agentes disponibilizam um vasto repertório de dados desde a sua geração até a comercialização das fontes de combustíveis (DE TEFFÉ; MEDON, 2020).

As empresas pertencentes ao setor de combustíveis cada vez mais utilizam o Aprendizado de Máquina (ML) para propor ações, passar confiança, compreender o cenário futuro para iniciar estratégias no presente, e também, compreender os momentos que o mercado está passando. Visto isso, contam com a engenharia de dados, cientistas de dados, governança de dados, pesquisadores na área de ML e algoritmos mais robustos.

Diante do exposto, este trabalho objetiva estudar o impacto do cruzamento de diferentes fontes de dados e técnicas de IA visando a predição da produção da gasolina e etanol, tendo como referência os dados do estado de São Paulo, região de maior demanda e consumo do país. Para essa tarefa, foram utilizados modelos de ML e ferramentas de Análise Exploratória de Dados (AED) de modo a criar uma metodologia computacional, bem como estratégias para fins de suporte, para que agentes do setor de energia possam respaldar decisões com base na análise inteligente dos dados.

2. OBJETIVO GERAL

O presente Trabalho de Conclusão de Curso (TCC) possui como objetivo geral realizar uma análise comparativa entre os modelos de regressão linear e série temporal selecionados no estudo para previsão da produção dos combustíveis etanol e gasolina do estado de São Paulo.

2.1. OBJETIVOS ESPECÍFICOS

Este estudo teve como principais objetivos:

- A exploração visual/analítica dos conjuntos de dados disponibilizados pela ANP, de forma a pré-processá-los e adequá-los para efetiva utilização dos modelos preditivos (passos de *Data Cleaning* e Descoberta de Conhecimento);
- Explorar as etapas de treinamento e fase de testes para cada um dos modelos de Aprendizado de Máquina adotados no projeto;
- Obter predições, aferindo-as em cenários práticos de uso, isto é, a partir dos dados reais da ANP. Em particular, pretende-se treinar e validar os modelos computacionais a partir de quatro datasets (especificados na Seção 4.1), disponibilizados pela ANP;
- Criação de sumários de dados como gráficos, tipos distintos de visualizações, e análises de medidas de erros, a fim de auxiliar atores/gestores do setor energético a tomarem decisões com base nos dados e nas tendências geradas pelos modelos.

3. REVISÃO DA LITERATURA

3.1. APRENDIZADO DE MÁQUINA

O primórdio do *Machine Learning* (ML) é sequenciado pelo aperfeiçoamento de estudos na área de sistemas conduzidos a simular o comportamento humano. Um dos exemplos mais associado à sua origem é a máquina "*perceptron*", que alcançou a capacidade de projetar letras do alfabeto e formas geométricas primitivas, advindo da classificação de padrões de quais fotocélulas eram ativadas (EBY; MOSCARDI, 2022). A percussão dessa máquina é referente a muitos trabalhos sobre o funcionamento do sistema nervoso humano do psicólogo Frank Rosenblatt e seu grupo, que anos seguintes, serviu de protótipo para a desenvoltura das redes neurais artificiais (FRADKOV, 2022).

Em virtude de pesquisas nas áreas de ciência da computação, estatística, engenharia, matemática e do avanço da tecnologia, entende-se o aprendizado de máquina como subcampo

da Inteligência Artificial, por ser um sistema hábil a tomar decisões e encontrar padrões por meios de técnicas que aprendam a partir de um grande volume de dados automaticamente, simulando um cérebro humano (LUDERMIR, 2021).

3.2. TIPOS DE APRENDIZAGEM E TÉCNICAS DO APRENDIZADO DE MÁQUINA

O ML possui três tipos de aprendizagem - a supervisionada, quando há a necessidade do humano fornecer a entrada e saída dos dados, transfigurando-se no papel de supervisor. Precisa das características necessárias para o algoritmo identificar a correspondência correta das informações, e também, o retorno da precisão da previsão no processo de treinamento. Dentro da aprendizagem supervisionada existem os métodos de regressão e classificação que são os métodos preditivos. A não supervisionada, em que não há a necessidade de nenhum processo de treinamento, analisa automaticamente os dados e necessita da análise para determinar o significado dos padrões. Detém a associação, agrupamento, detecção de desvios, padrões sequenciais e sumarização como métodos preditivos. Por último, a aprendizagem por reforço, em que o algoritmo aprende com as interações ao ambiente, ou seja, com sua própria experiência (SARAVANAN; SUJATHA, 2018).

Para o presente trabalho utilizaremos os modelos de aprendizagem supervisionada por regressão, como apresentados a seguir.

3.3. MODELOS DE REGRESSÃO

O método preditivo por regressão usufrui da relação entre uma variável dependente (y) e variáveis explanatórias (x), também conhecidas como atributos previsores. Existem dois tipos de análise por regressão: a regressão linear simples, que possui uma variável explanatória. Por fim, a regressão linear múltipla, que requer mais de uma variável previsora (ALITA; PUTRA, 2021). Dos múltiplos modelos de regressão, existem alguns que são destaques em estudos do setor de energia por alcançarem sucessivas previsões, dentre eles temos: Florestas Aleatórias (RF), Redes Neurais Artificiais (RNA), entre outros.

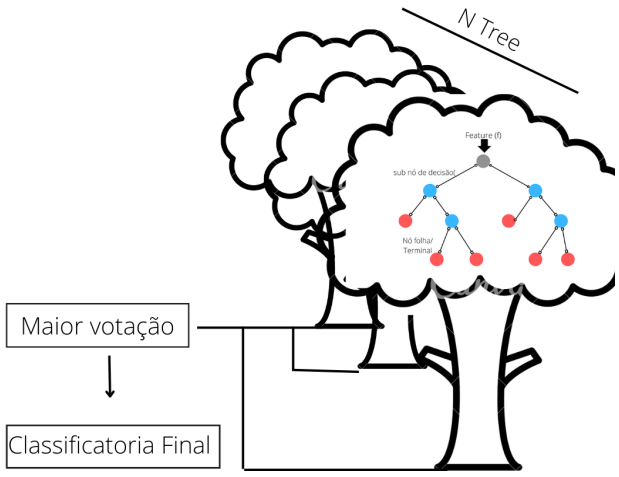
3.3.1. Random Forest

As Florestas Aleatórias (do inglês, *Random Forest*) são utilizadas para resolver problemas de regressão e classificação. A aprendizagem é realizada por conjunto, ou então, da união dos preditores das árvores de decisão, sendo que cada árvore tem como entrada um

vetor aleatório e a mesma distribuição para todas as outras árvores de hipóteses ou dados. A junção de mais uma árvore aumenta a precisão para resolver o mesmo problema, estabelecendo regras, contendo nodos e arestas para tomada de decisões com base nos recursos das instâncias (JÚLIO CESAR et al., 2022).

Portanto, combina suas previsões calculando a média ou por votação dos conjuntos com o objetivo de fornecer soluções para problemas complexos (SANTOS et al., 2021). A Figura 1 exemplifica um fluxograma do modelo de RF.

Figura 1 - Comportamento do modelo de RF



Fonte: Adaptado de Mohammadreza et al (2020).

Trata-se de um modelo vantajoso, porque evita o *overfitting*[1], tendo assim uma boa precisão, pois utiliza o treinamento, validação e teste, construindo modelos mais generalizados, mantendo a complexidade dependente do tamanho da sua base de dados. O erro de generalização para florestas converge à medida que o número de árvores nas florestas aleatórias aumenta e depende da força das árvores individuais e da correlação entre elas, isto é monitorado pelas estimativas internas. Essas estimativas também são usadas para medir a importância da variável, tanto na classificação quanto na regressão (MOREIRA et al., 2022).

3.3.2. Redes Neurais Artificiais

Arquitetada matematicamente, em forma de algoritmos computacionais, as Redes Neurais Artificiais refletem dois aspectos básicos de como o cérebro realiza uma tarefa. A

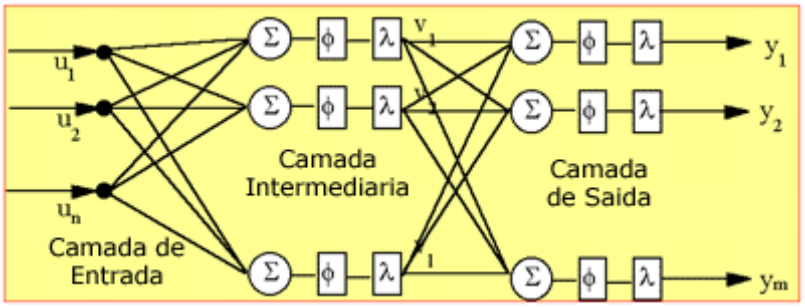
[1] Quando o modelo se adapta ao conjunto de treinamento e torna-se ineficaz para prever novos resultados.

aprendizagem é alcançada por meio do conhecimento adquirido de seu ambiente por tentativa e erro, além disso, a armazenagem do conhecimento é concretizada por meio da força entre as interações dos neurônios, ou pesos sinápticos (FLECK et al., 2016; ANDRE, 2022). Por fim, o bom desempenho sucede da interligação maciça das células computacionais simples, tituladas de neurônio ou unidades de processamento (DIAS, 2013).

Usualmente, o modelo conta com três classes de arquiteturas de rede neural diferentes: redes alimentadas adiante com camada única, redes alimentadas diretamente com múltiplas camadas e redes recorrentes (FLECK et al., 2016). Por esse motivo, a seleção da melhor arquitetura de uma rede neural é um dos maiores desafios do algoritmo, uma vez que esse processo é experimental e demanda um grande tempo de execução.

Compreende-se as RNA's como receptor de sinais de determinado padrão em sua entrada, apto a analisar e então informar sobre a classe à qual eles pertencem. A identificação do padrão só acontece após o treinamento da rede, que pode acontecer de forma supervisionada ou não supervisionada em várias possibilidades de arquitetura e algoritmos de treinamento (DE ALMEIDA et al, 2021). Em resumo, cada elemento de processamento ou neurônio em uma camada está conectado e levando um valor a todos os elementos de processamento na próxima camada, e assim por diante, até que os neurônios na última camada oculta estejam conectados aos neurônios da camada de saída (LIU et al., 2021), vide Figura 2. Importante mencionar que são ativados se o valor de entrada for maior que um número definido.

Figura 2 - Modelo matemático do neurônio multicamadas de uma RNA



Fonte: Adaptado de Gérson do Santos (2019).

3.4. SÉRIE TEMPORAL

As séries temporais são representadas por um conjunto de dados ordenados em forma cronológica, tendo como aplicação principal a previsão de valores futuros por meio do

fenômeno estudado ao longo do tempo. A expressão matemática que descreve uma série temporal é dada por “Y”, em que representa a variável de interesse, e “T”, que retrata o conjunto de índices relacionados aos tempos de medição (SILVA et al., 2021).

A natureza dos valores observados caracteriza a série temporal em dois estilos: em contínua, na qual a variável preditora (*target*) é evidenciada de forma sequencial em um intervalo de tempo; ou discreta, quando são feitas em intervalos de tempo fixos e enumeráveis (SILVA et al., 2021). Além do mais, são quatro componentes que modela o algoritmo, sendo os parâmetros dados por:

- Tendência (T): Análise do comportamento da variável global ao decorrer do intervalo de tempo, podendo ser de crescimento ou decrescimento;
- Ciclo (C): Estudo específico do modo que a variável se comporta em pequenos intervalos de tempo, os quais se repetem com certa periodicidade;
- Sazonalidade: Comportamento específico (padrões de flutuações) do objeto de estudo em cada determinado tempo, como dados da economia e clima;
- Variação Irregular (I): Valores que não podem ser previstos, pois são ocasionados por fatores externos os quais não se tem controle, ou então, não apresentam um padrão identificável (ANDRADE et al., 2021; OLIVEIRA, 2022).

A partir da realização da análise exploratória dos dados, é possível avaliar a aplicabilidade de diferentes modelos para previsão de séries temporais. Dentre estes, o Modelo Autorregressivo de Médias Móveis (ARIMA) é comumente conhecido e utilizado pelas literaturas cujo tema está associado à previsões por série temporal.

3.4.1. Modelo Autorregressivo de Médias Móveis

Diferente dos outros métodos de regressão mencionados anteriormente, o ARIMA não sofre influência de outras variáveis, sendo uma técnica auto regressiva que calcula futuras previsões de curto prazo a partir da análise dos valores defasados de sua própria correlação temporal (FAN et al., 2021).

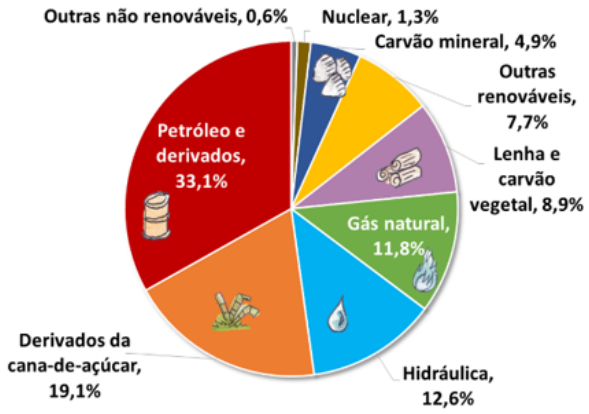
A identificação dos modelos ARIMA são baseados na determinação das ordens de três componentes, também denominados “filtros”: a ordem ‘p’, associada aos parâmetros do filtro auto-regressivo (AR), ordem de defasagem; a ordem ‘q’, associada aos parâmetros do filtro Médias Móveis (MA) tamanho da janela móvel; e ordem ‘d’, associada ao número de diferenciações necessárias para tornar a série estacionária, grau de diferenciação envolvido, ou seja, ao parâmetro filtro Integração (I) (SAULO, 2022). A identificação da estabilidade

serial, ou estacionariedade, é a primeira etapa da modelagem, devido à possibilidade de existir períodos sazonais na série (SILVA et al., 2021).

3.5. MERCADO DE COMBUSTÍVEIS

O Brasil possui sua economia primária marcada por algumas atividades, entre elas, agropecuária e extração do petróleo, em virtude das condições favoráveis do clima e solo. Conjuntura que alavanca o comércio do petróleo e seus derivados, além, da indústria de cana de açúcar. São hoje, 450 usinas de cana de açúcar instaladas e 280 poços produtores no país, respaldando no cenário atual da matriz energética (OPERACAO, 2022; PETROBRAS, 2022), vide Figura 1.

Figura 3 - Matriz Energética Brasileira 2022



Fonte: EPE (2022).

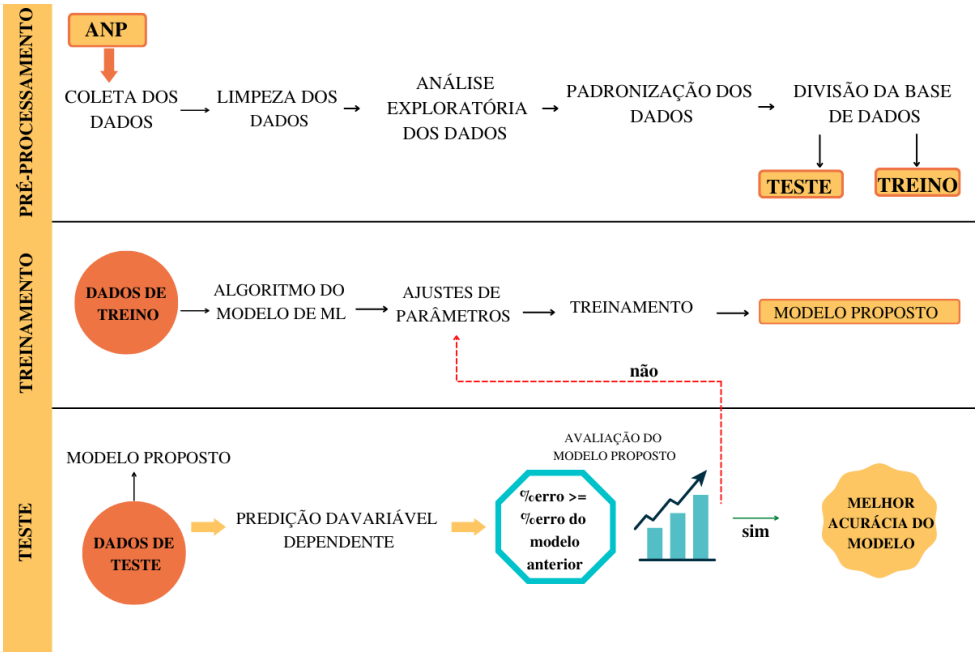
Por tratarem de transações economicamente e ambientalmente significantes, criou- se a Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP), cujo papel é fiscalizar as atividades da indústria do petróleo e a distribuição e revenda de derivados de petróleo e álcool combustível (REPÚBLICA, 1998). O mercado de gasolina também é regulamentado pela Lei Federal 9.478/97, a qual flexibiliza o controle do setor de petróleo e gás natural, até então exercido pela Petrobras, tornando aberto o mercado de combustíveis no país e permitindo que os reajustes nos preços dos combustíveis passassem a caber exclusivamente a cada agente econômico – do poço ao posto revendedor (NUNES, 2021). A tendência a partir dos próximos anos é prosseguir com estratégias para transição energética, e o mercado já vem tomando algumas ações, designadamente em incentivos a novas fontes,

automação e inteligência artificial, captura de carbono, eficiência energética, produção de combustíveis a partir de resíduos (Bioetanol), entre outros, o que irão causar uma mudança no cenário atual.

4. MATERIAIS E MÉTODOS

O primeiro passo se resume à extração dos dados e criação dos bancos de dados (*database*), em seguida, para desenvolvimento do código em linguagem Python usa-se do software Jupyter. As etapas que requerem muita atenção e que também impactam na assertividade do algoritmo estão ligadas à limpeza (*Data Cleaning*) e organização dos dados, a fim de mitigar redundâncias e possíveis efeitos degenerativos da presença de *outliers*. A análise exploratória de dados e geração de sumários, que usufruem de ferramentas KDD a fim de investigar o comportamento e novos padrões que sejam válidos e potencialmente úteis, são necessárias para os modelos encontrarem o melhor padrão. A calibragem, ajuste de performance, etapa de treinamento e teste de cada um dos modelos de aprendizagem supervisionada a serem empregados, constituem a etapa final da construção do modelo. A Figura 4 a seguir, representa o fluxograma das etapas, cujo objetivo é esclarecer as etapas já mencionadas.

Figura 4 - Fluxograma das etapas realizadas no trabalho para construção do modelo computacional.



Fonte: Autor (2022).

4.1. DESCRIÇÃO DAS VARIÁVEIS

Na pesquisa foram utilizados quatro diferentes bases de dados (BD) para previsão da produção (m³) de cada fonte de combustível selecionada (etanol e gasolina), as quais foram acopladas e estão disponíveis no site da Agência Nacional de Petróleo [2]. A escolha do período de tempo foi de acordo com a disponibilidade do site da ANP, nesse caso, intervalo mensal. Em seguida, para os bancos de dados utilizados pelos modelos RF e RNA, a variável data foi transformada em uma nova variável do tipo *float*, pois uma das características do treinamento por regressão é o aprendizado por variáveis numéricas.

As variáveis que foram adicionadas de outras fontes, sem ser da ANP, para melhoramento do modelo, serão referenciadas nos subcapítulos a seguir.

4.1.1. Modelos de Regressão Linear Simples

4.1.1.1. Etanol Hidratado

Foram consideradas as seguintes variáveis explanatórias: data (mensal); produção etanol anidro(m³) em SP; produção diesel (m³) em SP; quantidade de importação do etanol anidro (m³); quantidade exportação do etanol anidro; produção da cana de açúcar (toneladas); exportação do açúcar (tol); exportação etanol hidratado (m³); área da produção de cana (m³); venda do etanol (m³); venda da gasolina C (m³); venda do óleo diesel (m³); preço do dólar (convertido em reais). Detendo como variável meta: produção etanol hidratado (m³) em SP.

A variável do "preço do dólar" foi retirada no site Investing[3]. Já as variáveis da "produção da cana de açúcar", "exportação do açúcar" e "área da produção de cana" foram retirados dos sites da Conab[4] e Observatório da Cana[5].

O modelo inicial possui o artefato de dados de janeiro de 2012 a março de 2021, no qual, e em todas bases trabalhadas, é na etapa do Pré Processamento que o período pode variar de acordo com a necessidade dos dados tratados.

---

[2] Disponível em: <http://w ww.anp.gov.br/> Acesso em: 01 ago, 2022.
[3] Disponível em: <https://br.investing.co m/currencies/usd-brl-historical-data> Acesso em: 01 ago, 2022.
[4] Disponível em: <https://www.conab.gov.br/info-agro/safras/serie-historica-das-safras> Acesso em: 01 set, 2022.
[5] Disponível em: <https://observatoriodacana.com.br/> Acesso em: 01 set. 2022.

4.1.1.2. Gasolina A

O período do repertório de dados da gasolina, foi inicialmente coletado de janeiro de 1990 a julho de 2021, considerando as seguintes variáveis explanatórias: data (mensal); preço do dólar (convertido em reais); processamento do petróleo (m³); produção petróleo em SP (m³); produção etanol anidro em SP (m³); quantidade da venda do etanol em SP (m³); quantidade da venda da gasolina C em SP (m³); quantidade da venda do óleo diesel em SP (m³). Possuindo a variável "produção média da gasolina em São Paulo (m³)" como meta.

4.1.2. Modelos de Regressão por Série Temporal

4.1.2.1. Etanol Hidratado

Foram consideradas as seguintes variáveis: data (mensal) e produção do etanol hidratado (m³) em SP. O modelo inicial possui o artefato de dados de janeiro de 2012 a março de 2021, no qual, e em todas bases trabalhadas, é na etapa do Pré-Processamento que a data pode variar de acordo com o tratamento dos dados.

4.1.2.2. Gasolina A

O período do repertório de dados da gasolina, inicialmente agregado de janeiro de 1990 a julho de 2021. Considerando as seguintes variáveis: data (mensal) e produção média da gasolina em São Paulo (m³).

4.1.3. Engenharia de Recursos

Essa etapa foi utilizada para os modelos de regressão linear simples, pois o desempenho adequado do algoritmo depende, principalmente, de dados representativos e de características expressivas o suficiente para que o aprendizado de máquina descubra excelentes padrões (TATIS et al, 2022). Por essa razão, no presente trabalho foi aplicada a engenharia de recursos nas variáveis preditoras, a qual remete a criação de novas variáveis a partir das já existentes através de métodos estatísticos.

Efetuou-se em todos os modelos o total de 12 novas variáveis para cada, por meio dos cálculos de soma, subtração, divisão, multiplicação, média e exponencial, usando as duas variáveis preditoras de melhor correlação.

4.2 ANÁLISE EXPLORATÓRIA DOS DADOS

Os dados pré-processados trazem informações importantes, além da descrição do comportamento de cada variável, cujo objetivo é resumir os elementos como uma base para posteriormente realizar uma análise crítica da inferência estatística (TATIS et al, 2022). Essa inspeção foi realizada por meio da utilização de gráficos, entre as técnicas escolhidas para todas as bases foram:

- Análise descritiva em formato de tabela, a qual explora as métricas de cada variável de forma resumida;
- Gráfico de histograma do comportamento de cada variável *target*, utilizando a função *distplot* da biblioteca Seaborn;
- Visualização dos dados em relação ao tempo através da biblioteca matplotlib;
- Mapa de calor utilizando a técnica do coeficiente de Pearson, a fim de analisar o grau de correlação entre as variáveis de escala métrica, assumindo valores de -1 a 1;
- Gráfico da acurácia de cada modelo de ML aplicado.

4.3. MODELOS DE APRENDIZADO DE MÁQUINA (AM)

4.3.1. Random Forest (RF)

Para executar o modelo RF, dois parâmetros foram definidos: a quantidade de árvores (Ntree) e o número de recursos selecionados aleatoriamente, pois cada árvore de decisão é treinada em um subconjunto indeterminado de amostras com substituições derivadas dos dados de treinamento juntamente com conjuntos de recursos selecionados aleatoriamente, buscando a melhor característica (*feature*). As previsões individuais resultantes são então submetidas a um processo de combinação (*ensemble*) e por se tratar de um modelo de regressão, a média é considerada a solução final prevista (MAZUMDAR; NETO; PAULOVICH, 2021).

Portanto, as bases de teste foram treinadas pelo modelo com diferentes estimadores, sendo que para a construção de cada um, usou-se a aleatoriedade, e também, por meio da ferramenta Sklearn, mediu-se o grau importância das características e analisou quais nodos das árvores reduziam a impureza da árvore. Características que evitam o *overfitting*.

4.3.2. Redes Neurais Artificiais (RNA)

Através do aprendizado supervisionado, o algoritmo de Rede Neural Artificial de multicamadas na atual pesquisa inicia seu treinamento com pesos aleatórios para cada uma de suas conexões, e repetindo até que o erro fosse satisfatoriamente pequeno, subtraindo as previsões pelos dados reais. A equação de como o algoritmo atualiza os pesos até encontrar o melhor resultado é representado pela equação 6 - *backpropagation* - porém antes dessa solução é realizada a derivada da função (equação 3), em seguida o cálculo do delta (indica melhor direção para o alcance do melhor resultado- equações 4 e 5) e depois o gradiente até o valor mínimo global. O cálculo da função soma, dado pela equação 1, depende da quantidade do número de entradas, que consequentemente torna-se o valor inserido na função de ativação - intermédio do resultado do cálculo gerados pelas unidades, o qual também informa se o neurônio foi ativado em sua camada. Na pesquisa foi utilizada na camada de saída a função sigmóide (retorna valores de 0 a 1), apresentada na equação 2.

$$soma = \sum_{i=1}^{p} xi * wki \quad (1)$$

$$\phi = \frac{1}{1 + e^{-x}} \quad (2)$$

$$d = y * (1 - y) \quad (3)$$

$$delta\ saída = Erro * Derivada\ Sigmoide \quad (4)$$

$$deltacamada\ escondida = Derivada\ sigmoide * peso * delta\ saída \quad (5)$$

$$peso(n + 1) = (peso\ (n) * momento) + (taxa\ de\ aprendizagem * entrada * delta) \quad (6)$$

4.3.3. ARIMA

Para esse estudo, o pacote *'pmdarima'* pertencente a biblioteca pandas do python foi utilizada para previsão de uma série temporal, usufruindo da metodologia ARIMA. Visto que usualmente é aplicada nos casos onde os dados mostram evidências de não estacionariedade, adicionando a noção de integração, visando analisar e solucionar tendências, sazonalidades e elementos aleatórios (DA SILVA et al., 2022).

São três etapas principais que estruturam o modelo ARIMA:

- Identificação do modelo: inicialmente plotou-se os gráficos em relação à variável previsora x tempo, para analisar se os dados estão distribuídos de forma estacionária. A diferenciação é geralmente aplicada sobre os dados para remover as tendências e estabilizar a variância. Logo, é nessa etapa que o parâmetro ‘d’ é determinado.
- Estimativa dos parâmetros: por ser um modelo auto regressivo e usufruindo do código auto arima - ‘model.order’, o modelo identifica os números dos parâmetros corretos (‘p’,’q ‘e ‘d’).
- Verificação diagnóstica: finalizada a previsão por um tempo estimado, realiza-se a verificação da precisão do modelo, por meio do cálculo do Erro Absoluto Médio Percentual (MAPE).

Outra etapa fundamental, na ocasião do pré processamento, é transformar a variável da data em um índice, de modo a criar a série temporal, como mostra a Figura 5 a seguir:

Figura 5 - BD finalizada após o pré-processamento com variável “DATA” como índice.

DATA	PRODUÇÃO
2012-01-01	21710
2012-02-01	2515
2012-03-01	6467
2012-04-01	217299
2012-05-01	730343

Fonte: Autor (2022).

As etapas auto regressivo (AR), integração (I) e médias móveis (MA), podem ser analisadas matematicamente através das equações 7, 8 e 9, respectivamente.

$$AR = \Theta\, Yt - 1 \quad (7)$$

$$\nabla Yt = Yt - Yt - 1 \quad (8)$$

$$MA = at - \theta - at - 1 \quad (9)$$

4.4. AVALIAÇÃO DO ERRO

Ao se construir cada métrica, podem surgir duas categorias de erro. A seguir, segue a explicação teórica de cada tipo.

- o *underfit*: quando há pouca quantidade de amostras, e a *baseline* em análise, na parte de treinamento, a arquitetura do algoritmo não permite a criação de regras específicas para encontrar o melhor padrão, assim, o modelo não consegue criar regras adequadas de comportamento de dados (SOUZA, 2022).
- o *overfit*: refere-se a adaptação dos dados ao treinamento, consequentemente, não generaliza bem a entrada de novos dados (SOUZA, 2022).

Com o propósito de avaliar a acurácia de cada modelo foi considerado o Erro Absoluto Médio Percentual (MAPE), que traz a facilidade da interpretação, pois é gerada a análise percentual da precisão do treinamento. Matematicamente calculado por meio da equação 10.

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \widehat{y}_i|}{y_i} * 100\% \quad (10)$$

4.5. TESTE COM UM NOVO DATA FRAME

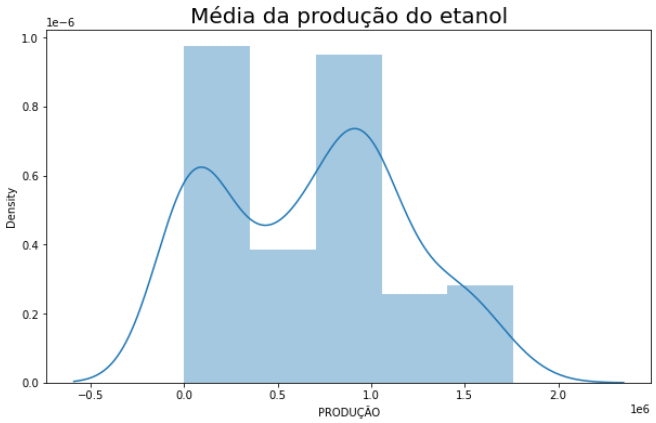
O objetivo dessa fase é aplicar o modelo computacional a um novo data frame similar aos utilizados na etapa de treinamento e teste, porém, com valores distintos. Desse modo, para previsão em curto/médio tempo, é possível avaliar se o modelo ocasionará em futuros erros caso haja a inserção de novos dados. Para tal, avalia-se o resultado pelo cálculo do MAPE.

5. RESULTADOS E DISCUSSÕES

Finalizado a etapa do pré-processamento para os modelos de predição da produção de cada fonte de combustível, decorre o início da visualização da exploração dos dados, cujo objetivo é analisar o comportamento da variável alvo diante do data frame.

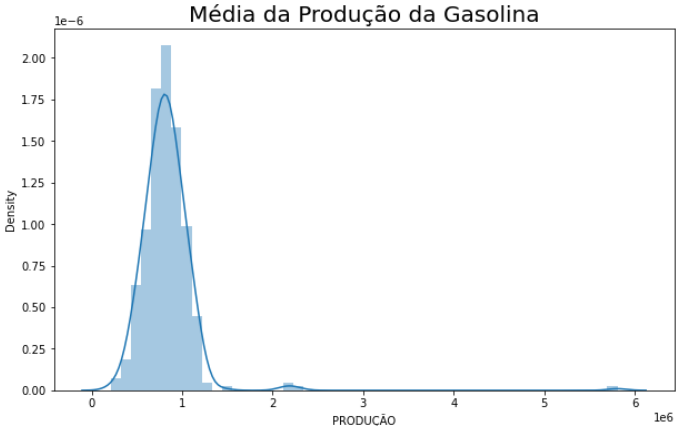
Por meio das Figuras 6 e 7, observa-se que o comportamento dos dados entre os diferentes tipos de combustíveis se diferenciam, podendo interpretar que haverá cenários e variáveis incomuns entre os mesmos. Como a parábola dos gráficos da gasolina e etanol não se concentram no meio, indica que há existência de muitos *outliers*, o que pode influenciar no resultado final e os dados devem ser tratados para que o algoritmo encontre o melhor padrão.

Figura 6 - Gráficos da distribuição das produções médias do estado de São Paulo das bases de dados do Etanol.



Fonte: Autor (2022).

Figura 7 - Gráficos da distribuição das produções médias do estado de São Paulo das bases de dados da Gasolina.



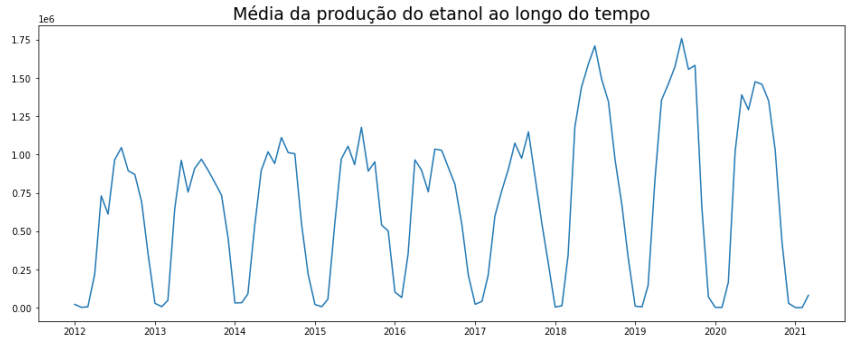
Fonte: Autor (2022).

Nas Figuras 8 e 9, analisa-se ao longo do tempo em qual momento há ocorrência da dispersão dos dados, se as projeções possuem uma tendência ou linearidade. Desse modo, foi possível verificar que a gasolina, no ano de 2012, obteve um pico elevado, como mostra a Figura 9. Por meio de pesquisas, concluímos que trata-se de um fenômeno ocorrido no Brasil, em que precisou importar um volume esporádico, em decorrência do crescimento de demanda e insuficiência das produções nas refinarias, as quais não foram capazes de manter conforme a necessidade do mercado (VEJA, 2013). Contudo, a Figura 8 constata que a produção do etanol sofre picos consideráveis todo o ano e em determinados meses. A justificativa dessas

ocorrências se respalda nos meses da abertura e fechamento do ano safra, que são distintos do método da produção da gasolina.

Essas análises apontam a importância das visualizações, e quais medidas tomar para o modelo encontrar o melhor padrão. Assim sendo, no trabalho tratamos de coletar variáveis que possivelmente influenciaram nas causas detectadas nos gráficos, além das já coletadas da ANP, dentre elas foram: “preço do dólar”, “produção da cana de açúcar”, “exportação do açúcar” e “área da produção de cana”.

Figura 8 - Gráficos do comportamento das produções médias do estado de São Paulo ao longo do tempo das bases de dados do Etanol.



Fonte: Autor (2022).

Figura 9 - Gráficos do comportamento das produções médias do estado de São Paulo ao longo do tempo das bases de dados da Gasolina.

Fonte: Autor (2022).

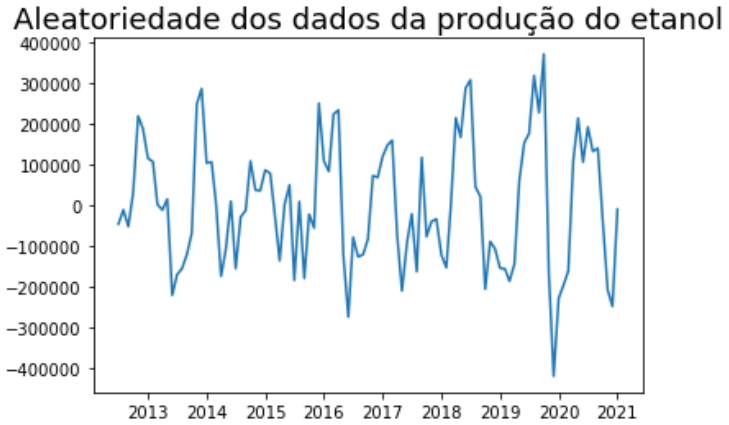
A partir das Figuras 10, 11 e 12, podemos interpretar os casos para a ocorrência da sazonalidade, aleatoriedade e tendência das séries da produção média do etanol. A Figura 10 demonstra efeitos em períodos específicos, e que era esperado que fosse similar ao gráfico 8, pelo fato de ser nas férias e final de ano a ocorrência do aumento exacerbado do consumo,

além da época de colheita ser entre novembro e abril[6]. O mercado de combustível já prepara uma produção em grande escala para esses períodos. A Figura 11 remete uma proporção de aleatoriedade significativa por parte dessa fonte de combustível, e que tem procedência da complexibilidade da produção da matéria-prima utilizada pelo etanol, a cana de açúcar. Pois ela depende do clima, solo, pragas, escala de produção, entre outros fatores. Por último, a Figura 12 detalha a tendência de crescimento de 2018 até o início de 2021, que reflete no cenário econômico do país, uma das influências já citadas na introdução, foi a pandemia da Covid-19 e a tomada de ação da OPEP, em foi preciso reduzir a produção gasolina no país e que impacta diretamente na produção de etanol.

Figura 10 - Gráficos da sazonalidade da série temporal da produção média do etanol no estado de São Paulo.

Fonte: Autor (2022).

Figura 11 - Gráficos da aleatoriedade da série temporal da produção média do etanol no estado de São Paulo.



Fonte: Autor (2022).

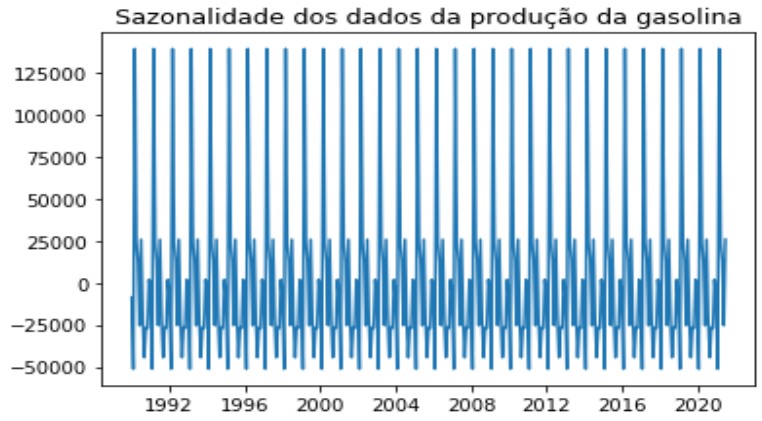
[6] Disponível em: <https://www.embrapa.br> Acesso em: 22 ago, 2022.

Figura 12 - Gráficos da tendência da série temporal da produção média do etanol no estado de São Paulo.

Fonte: Autor (2022).

As Figuras 13, 14 e 15, expressam a sazonalidade, aleatoriedade e tendência das séries da produção média da gasolina. No gráfico de sazonalidade, Figura 13, identifica-se cenários constantes ao percorrer do tempo, que então, diferencia um pouco do etanol. Um dos fatores que explica o fenômeno é que a gasolina não é estaticamente afetada pelos preços, e por hora, é um combustível insubstituível no país (ALVARENGA, 2017). Nos gráficos de tendência e aleatoriedade, Figura 14 e 15, percebe-se uma similaridade entre os mesmos, pois a produção da matéria prima não depende do clima ou umidade, como a cana, visto que o petróleo é uma produção consolidada e estruturada no Brasil.

Figura 13 - Gráficos da sazonalidade da série temporal da produção média da gasolina no estado de São Paulo.



Fonte: Autor (2022).

Figura 14 - Gráficos da tendência da série temporal da produção média da gasolina no estado de São Paulo.

Fonte: Autor (2022).

Figura 15 - Gráficos da aleatoriedade da série temporal da produção média da gasolina no estado de São Paulo.

Fonte: Autor (2022).

A fim de auxiliar os modelos RF e RNA a encontrarem o melhor padrão, também foi medida a correlação de Pearson entre as variáveis.

Geralmente, variáveis de maior correlação tendem a encontrar o melhor padrão para os modelos de algoritmo. Portanto, a Figura 16 exemplifica os mapas de calor gerados em cada data frame, com o propósito de auxiliar na melhor escolha de variáveis. A seguir, a Tabela 1 detalha quais foram utilizadas e que o modelo gerou a menor porcentagem de erro.

Figura 16 - Mapa de calor das variáveis da ferramenta de predição da produção média da gasolina no Estado de São Paulo.

Fonte: Autor (2022).

Tabela 1 - Variáveis utilizadas no modelo RF e RNA para predição da produção da gasolina e etanol.

Gasolina	Etanol
Data	Data
Preço do Dólar	Produção do Etanol Anidro
Processamento de petróleo	Importação Etanol Anidro
Produção de petróleo em SP	Exportação do Etanol Anidro
Venda de Etanol	Produção da cana de açúcar (tol)
Venda da gasolina	Exportação de Açúcar
	Venda do Diesel
	Venda Gasolina C

Fonte: Autor (2022).

Os resultados da análise exploratória mostra o grau de dificuldade em realizar uma previsão assertiva para produção, contudo, como exposto nos tópicos de metodologia, técnicas computacionais foram utilizadas para melhoramento do algoritmo para os métodos de regressão simples, sendo a engenharia de recursos um dos exemplos. Por conseguinte, para

predição da produção de etanol foram utilizadas “Produção Etanol Anidro” e “Produção da Cana de Açúcar” que criaram 12 novas variáveis por meio de cálculos matemáticos da engenharia de recursos, e dessas 10 foram usadas no modelo. Contudo, para predição da produção de gasolina usou “Venda da gasolina” e “Preço do dólar” que criou 12 novas variáveis explanatórias, mas só agregou ao modelo 8 variáveis.

Concretizado todas as etapas metodológicas e construída a ferramenta computacional utilizando as três metodologias de *Machine Learning*: Florestas Aleatórias, Redes Neurais Artificiais e ARIMA, utilizou-se o cálculo do MAPE (Seção 4.4) e análise exploratória para medir a acurácia de cada modelo aplicado.

Dessa forma, o resultado dos algoritmos da produção da gasolina por regressão linear simples de menor MAPE encontrado obteve 85% da base para treinamento, 15% teste e 16 variáveis explanatórias. O *Random Forest* usufruiu de 110 números de árvores da floresta, e o modelo de Rede Neural Artificial constitui de 2 camadas ocultas, função sigmóide, 1000 número de máximo de iterações, 9 de peso na camada oculta.

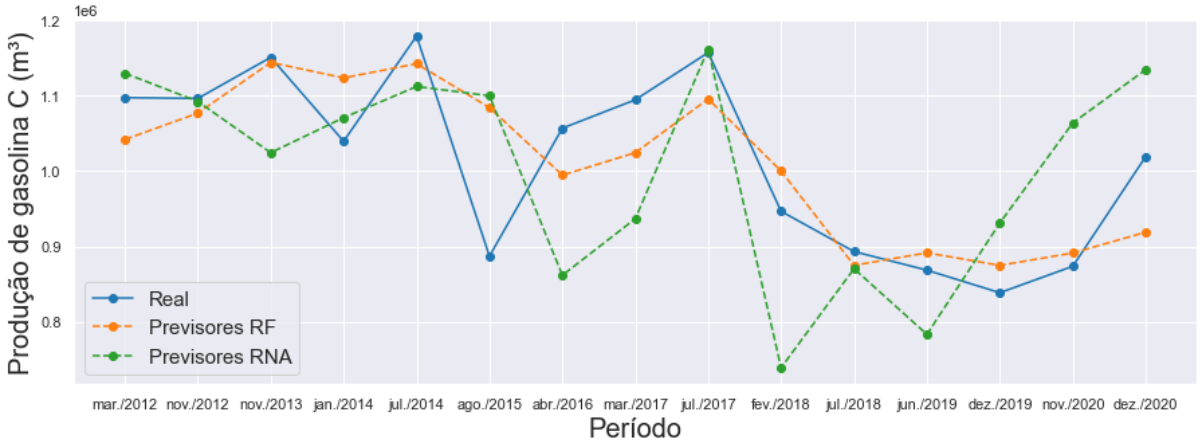
Em relação ao de etanol, o modelo compôs 75% da base para treinamento, 25% teste e 15 variáveis dependentes. O *Random Forest* portou de 110 números de árvores da floresta e na etapa da Rede Neural Artificial seguiu na codificação Função sigmóide, 1500 de máximo de interação, 2 camadas ocultas, 8 no peso da camada oculta, tol = 0.0010, solver = ’adam’ e activation= ’relu’.

Para os algoritmos de série temporal, a previsão da produção média da gasolina com ARIMA usou os parâmetros p=4, q=1 e d=1, além de separar 300 dados para treinamento e 79 para teste, realizando uma previsão em um período de 12 meses. Com relação a previsão do etanol os parâmetros foram p=5, q=0 e d=0, dividindo 100 dados para treinamento e 15 para teste. Nesse caso, foi realizado um previsão com período de 15 meses.

Logo, o erro de cada aprendizado foi mensurado pelo cálculo do erro percentual médio e explícito através das Figuras 17, 18, 19 e 20.

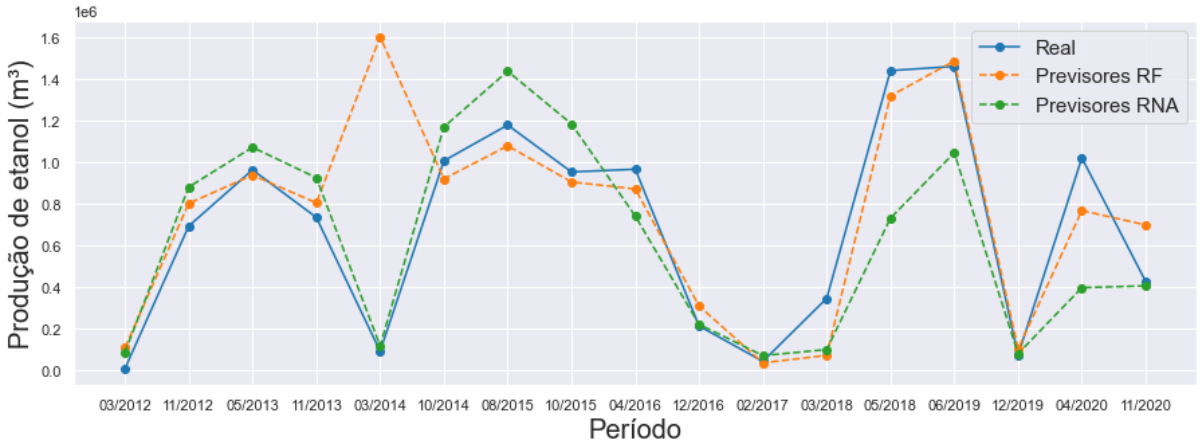
As Figuras 17 e 18, detalham a assertividade dos modelos do RF e RNA, em que resultaram em valores próximos do real, visto que só na Figura 18, o RF teve um erro exagerado, que consequentemente teve impacto no cálculo do MAPE. Na Figura 17, o RNA também apontou alguns erros, prevendo queda na produção, quando na verdade aumentou.

Figura 17 - Assertividade dos modelos RF e RNA para predição da produção média da gasolina no estado de São Paulo.



Fonte: Autor (2022).

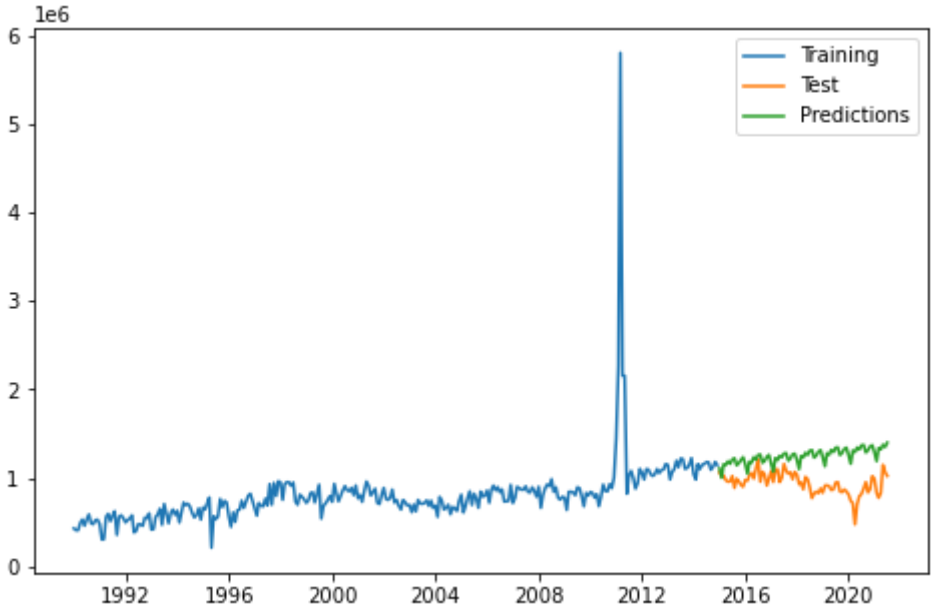
Figura 18 - Assertividade dos modelos RF e RNA para predição da produção média do etanol no estado de São Paulo.



Fonte: Autor (2022).

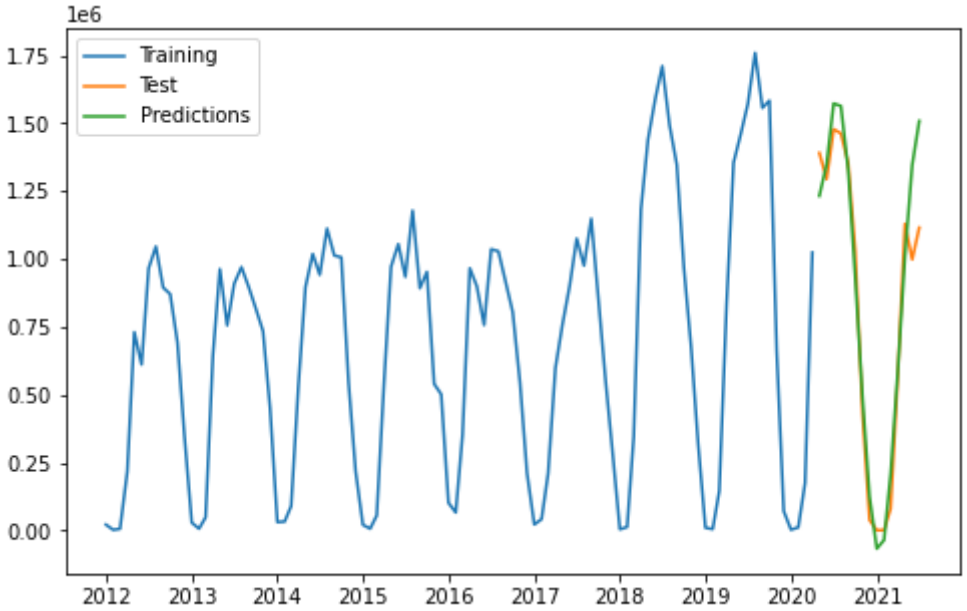
As Figuras 19 e 20, detalham a assertividade do modelo ARIMA, que resultou em valores consideráveis, dado que a Figura 19 expressa uma contradição entre os valores previstos do real. A previsão ocorreu na sazonalidade de 2016 a 2021, após o ano de 2020 o modelo disseminou significamente, que possivelmente está atrelado a eventos esporádicos. Apesar disso, a Figura 20 descreve uma excelente assertividade do valor real.

Figura 19 - Predição da produção média da gasolina no estado de São Paulo do modelo ARIMA.



Fonte: Autor (2022).

Figura 20 - Predição da produção média do etanol no estado de São Paulo do modelo ARIMA.



Fonte: Autor (2022).

Portanto, os modelos apresentaram bons níveis de assertividade e uma análise exploratória detalhada sobre os possíveis acontecimentos que impactaram nos valores finais. Dentre todo o treinamento o RF foi o modelo que mais se aproximou do real na predição da produção média da gasolina, enquanto o ARIMA, foi o modelo com a menor porcentagem de erro na predição da produção do etanol.

A Tabela 2 ilustra o desfecho detalhado da assertividade dos modelos de aprendizado de máquina, completando as análises visuais já discutidas no trabalho (Figuras 17, 18, 19 e 20), pois analisando somente os gráficos, compreenderíamos que o RNA teve maior assertividade que o RF, dado que reflete uma parte das previsões, enquanto o MAPE analisa o todo. Desse modo, pode-se concluir que o modelo escolhido da série temporal encontrou mais facilidade em trabalhar com cenários de maior aleatoriedade e um banco de dados reduzido, resultando na predição da produção do etanol em curto/médio prazo próximos ao dos valores reais, enquanto os modelos por regressão linear encontraram melhor padrão com uma base de dados de grande volume e de menor aleatoriedade, resultando menor erro para produção de gasolina.

Tabela 2 - Resultados dos modelos de aprendizado de máquina

Tipo de combustível	Modelo ML	MAPE
Etanol	RF	41,03%
Etanol	RNA	52,06%
Etanol	ARIMA	5,88%
Gasolina	RF	5,67%
Gasolina	RNA	14,92%
Gasolina	ARIMA	25%

Fonte: Autor (2022).

6. CONCLUSÃO

O setor de combustíveis possui uma gama de dados abertos, o que facilita o uso de ferramentas de aprendizado de máquina para assessorar agentes da área a tomar ações ou enxergar problemas com pouco esforço e tempo. Para o presente projeto, a construção de um modelo computacional utilizando as três técnicas - Random Forest, Redes Neurais Artificiais e ARIMA- teve como objetivo realizar a predição da produção dos combustíveis mais comuns no estado de São Paulo, a fim de visualizar cenários preditivos em um curto/médio tempo, visto que são temas triviais a economia e mudanças de planos políticos do mundo.

Para tal, foi realizada a combinação de diferentes conjuntos de dados, sendo necessário o pré-processamento e "limpeza" de dados inconsistentes, em seguida, a AED forneceu o comportamento e a correlação dos dados antes da aplicação dos modelos. O mapa de calor assegurou as variáveis de maior grau de correlação, em que para os dois data frame utilizados para os métodos de regressão linear simples, suas variáveis *target* não possuíram boa correlação, exclusivamente a produção do diesel teve correlação com a produção da gasolina; produção do etanol com a do diesel. Salienta-se que a implementação da técnica de engenharia de recursos com a criação de novas variáveis, promoveu o aumento da assertividade dos modelos e diminuição do tempo de testes.

Determinado as porcentagens de dados para treinamento e teste, consequentemente, a aplicação dos modelos preditivos, a predição alcançou resultados válidos comparados aos valores reais. Destaca-se que a inserção de uma base de dados novas revelou as predições com *overfitting*, assegurando a confiabilidade da ferramenta computacional. Em vista disso, o *Random Forest* se mostrou o modelo mais adequado para previsão da produção média de gasolina, em seguida, a Rede Neural Artificial foi a segunda melhor, muitas vezes chegando bem próximo ao RF. Já para produção média do etanol, o ARIMA teve melhor assertividade comparado aos outros dois modelos.

Portanto, diante da pesquisa, foi possível criar bons resultados, identificando que para bases de maior histórico de dados e menor quantidade de *outliers*, foram os modelos por regressão linear simples que encontraram o melhor padrão e assertividade, em destaque para o RF. Além disso, bases de uma série sequencial menor e com maior quantidade de casos aleatórios, o modelo ARIMA atingiu maior assertividade. Desta maneira, conclui-se que a ferramenta computacional criada tem a capacidade de dar suporte a novas pesquisas que trabalham com análises preditivas por Aprendizado de Máquina e a agentes do setor de energia.

REFERÊNCIAS

ALITA, Debby; PUTRA, Ade Dwi; DARWIS, Dedi. Analysis of classic assumption test and multiple linear regression coefficient test for employee structural office recommendation. IJCCS (Indonesian Journal of Computing and Cybernetics Systems), v. 15, n. 3, p. 1-5, 2021.

ALVARENGA, Samia Mercado; VIEIRA, Kelmara Mendes; FIALHO, Pedro Pessano. Demanda por gasolina: um estudo de caso para uma rede de postos de combustíveis. **Estudos do CEPE**, n. 46, p. 149-165, 2017

ANDRADE, Fillipe de A. et al. Previsao da Geraçao de Energia Fotovoltaica Utilizando Inteligência Artificial em Séries Temporais. In: Simpósio Brasileiro de Automação Inteligente-SBAI. 2021.

ANP. Especial ANP 20 Anos. Disponível em: https://www.gov.br/anp/pt-br/acesso-a-informacao/institucional/especial-anp-20-anos#:~:text=Hoje%2C%20a%20ind%C3%BAstria%20do%20petr%C3%B3leo,distribui%C3%A7%C3%A3o%20e%20revenda%20de%20combust%C3%ADveis. Acesso em: 8 nov. 2022.

OLIVEIRA, Drielly Galhardo; LUZ, Giovanna Freo da. Indústria petrolífera: o mercado pós impactos da Covid-19. 2021.

Carvalho, A. C. P. L. F.. "Redes Neurais Artificiais". disponível em: https://sites.icmc.usp.br/andre/research/neural. Último acesso: 10/11/2022.

DA SILVA, Ubiratan da Silva Tavares et al. ANÁLISE COMPARATIVA ENTRE OS MODELOS ARIMA E LSTM NA PREVISÃO DE CURTO PRAZO DA DEMANDA DE POTÊNCIA ATIVA. REVISTA DE ENGENHARIA E TECNOLOGIA, v. 14, n. 1, 2022

DE ALMEIDA, Murilo Marcineiro et al. Aplicação de Redes Neurais Artificiais ao diagnóstico de doenças em aves Application of Artificial Neural Networks to the diagnosis of diseases in birds. Brazilian Journal of Development, v. 7, n. 9, p. 94044-94056, 2021.

DE TEFFÉ, Chiara Spadaccini; MEDON, Filipe. Responsabilidade civil e regulação de novas tecnologias: questões acerca da utilização de inteligência artificial na tomada de decisões empresariais. REI-Revista Estudos Institucionais, v. 6, n. 1, p. 301-333, 2020.

DIAS, Bruno Rafael Rodrigues. Sistema inteligente para reconhecimento de timbre. 2013.

EBY, Michael; MOSCARDI, Rafael. Conexões matriciais: Perceptron enquanto diagrama. Das Questões, v. 15, n. 1, 2022.

FAN, Dongyan et al. Well production forecasting based on ARIMA-LSTM model considering manual operations. Energy, v. 220, p. 119708, 2021.

FLECK, Leandro et al. Redes neurais artificiais: Princípios básicos. Revista Eletrônica Científica Inovação e Tecnologia, v. 1, n. 13, p. 47-57, 2016.

FRADKOV, Alexander L. Early history of machine learning. IFAC-PapersOnLine, v. 53, n. 2, p. 1385-1390, 2020.

JÚLIO CESAR, Franke Fagundes et al. Abordagem baseada em Árvores de Decisão para detecção e identificação de intrusões em ambientes da Internet das Coisas baseados em Computação em Nevoeiro. 2022.

LIU, Xin et al. A review of artificial neural networks in the constitutive modeling of composite materials. Composites Part B: Engineering, v. 224, p. 109152, 2021.

LUDERMIR, Teresa Bernarda. Inteligência Artificial e Aprendizado de Máquina: estado atual e tendências. Estudos Avançados, v. 35, p. 85-94, 2021.

IBP. Maiores produtores mundiais de petróleo em 2021. Disponível em: https://www.ibp.org.br/observatorio-do-setor/snapshots/maiores-produtores-mundiais-de-petroleo-em-2020/. Acesso em: 08 nov. 2022.

MAZUMDAR, Dipankar; NETO, Mário Popolin; PAULOVICH, Fernando V. Random Forest Similarity Maps: A Scalable Visual Representation for Global and Local Interpretation. Electronics, v. 10, n. 22, p. 2862, 2021.

MOREIRA, Amanda Silva et al. Classificação de proteínas expostas na superficie com Random Forest. 2022.

NUNES, Gérson dos Santos. O uso dos métodos arima e var-vec no estudo da demanda de energia elétrica no Rio Grande do Sul. 2019. Dissertação de Mestrado.

NUNES, Paulo Henrique de Castro. Análise dos choques de preço e comportamentos anticompetitivos na precificação da gasolina nos postos de Fortaleza. 2021.

OLIVEIRA, Elias Fernandes de. Análise de séries temporais para previsão de demanda no INSS. 2022

OPERACAO. Versatilidade do uso da cana-de-açúcar no mercado brasileiro. Disponível em: https://safras.com.br/versatilidade-da-cana-de-acucar-no-brasil/#:~:text=O%20mercado%20da%20cana%2Dde,55%25%20da%20%C3%A1rea%20nacional%20plantada.. Acesso em: 13 abr. 2022.

PETROBRAS. Bacia de campos. Disponível em: https://more.ufsc.br/homepage/inserir_homepage. Acesso em: 13 abr. 2022.
ENERGÉTICA, Empresa de Pesquisa. Matriz Energética e Elétrica. Disponível em: https://www.epe.gov.br/pt/abcdenergia/matriz-energetica-e-eletrica. Acesso em: 16 ago. 2022.

REPÚBLICA, Presidência da. DECRETO Nº 2.455, DE 14 DE JANEIRO DE 1998. Disponível em: http://www.planalto.gov.br/ccivil_03/decreto/d2455.htm. Acesso em: 16 ago. 2022.

SANTOS, Deyvyd Costa dos et al. Proposta de Ensino à Distância utilizando metodologia ativa de aprendizagem: com uso do Random Forest. 2021.

SARAVANAN, R.; SUJATHA, Pothula. A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification. In: 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2018. p. 945-949.

SHEYKHMOUSA, Mohammadreza et al. Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, v. 13, p. 6308-6325, 2020.

SILVA, Aline Beatriz dos Santos et al. Modelo Autorregressivo Integrado de Médias Móveis (ARIMA): aspectos conceituais e metodológicos e sua aplicabilidade na mortalidade infantil.

SOUZA, Rafael Fernando Silva. Detecção e classificação de falhas em rolamentos de motores elétricos baseado em árvores de decisão. 2022. Trabalho de Conclusão de Curso. Universidade Federal do Rio Grande do Norte.

Redes Neurais. Prof. Paulo Martins Engel. Modelos de neurônios artificiais. Fundamentos da lógica de limiar. Informática. UFRGS. Prof. Paulo Martins Engel.Revista Brasileira de Saúde Materno Infantil, v. 21, p. 647-656, 2021.

TATIS, Ana Flávia Giacondino Soligo Lezcano; CORRENTE, José Eduardo; FUMES-GHANTOUS, Giovana. Análise exploratória gráfica para dados assimétricos com presença de pontos discrepantes. Revista Brasileira de Iniciação Científica, v. 9, p. e022017-e022017, 2022.

VEJA, Da Redação. Importação de gasolina pelo Brasil é recorde. Disponível em: https://veja.abril.com.br/economia/importacao-de-gasolina-pelo-brasil-e-recorde/. Acesso em: 08 nov. 2022.

VIDAL, Maria de Fátima. Produção e mercado de etanol. 2020.

ZAMBIASI, S. P. "Arquitetura das Redes Neurais".https://www.gsigma.ufsc.br/~popov/aulas/rna/arquitetura. Último acesso: 15/11/2022.