



UNIVERSIDADE FEDERAL RURAL DO SEMI-ÁRIDO
PRÓ-REITORIA DE GRADUAÇÃO
CENTRO MULTIDISCIPLINAR DE CARAÚBAS
CURSO DE ENGENHARIA ELÉTRICA

LUIZ FELIPE COSTA DANTAS DE SENA

**DESENVOLVIMENTO DE FERRAMENTA COMPUTACIONAL PARA ANÁLISE DE
DADOS DO BALANÇO ENERGÉTICO NACIONAL**

CARAÚBAS-RN

2024

LUIZ FELIPE COSTA DANTAS DE SENA

**DESENVOLVIMENTO DE FERRAMENTA COMPUTACIONAL PARA ANÁLISE DE
DADOS DO BALANÇO ENERGÉTICO NACIONAL**

Monografia apresentada à Universidade Federal Rural do Semi-Árido como requisito para obtenção do título de Bacharel em Engenharia Elétrica.

Orientadora: Érica Mangueira Lima

Coorientador: Adriano Costa de Oliveira

CARAÚBAS-RN

2024

SS474 Sena, Luiz Felipe Costa Dantas de.
d DESENVOLVIMENTO DE FERRAMENTA
COMPUTACIONAL PARA ANÁLISE DE DADOS DO BALANÇO
ENERGÉTICO
NACIONAL / Luiz Felipe Costa Dantas de Sena. - 2024.
50 f. : il.

Orientadora: Érica Mangueira Lima. Coorientador: Adriano
Costa de Oliveira. Monografia (graduação) -
Universidade Federal
Rural do Semi-árido, Curso de Engenharia
Elétrica, 2024.

1. BEN. 2. Python. 3. Análise de dados. 4.
ARIMA. 5. Energética. I. Lima, Érica Mangueira,
orient. II. Oliveira, Adriano Costa de , co-
orient. III. Título.

Biblioteca Campus Caraúbas
Bibliotecária: Dalvanira Brito Rodrigues
CRB: 15/700

O serviço de Geração Automática de Ficha Catalográfica para Trabalhos de Conclusão de Curso (TCC's) foi desenvolvido pelo Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (USP) e gentilmente cedido para o Sistema de Bibliotecas da Universidade Federal Rural do Semi-Árido (SISBI-UFERSA), sendo customizado pela Superintendência de Tecnologia da Informação e Comunicação (SUTIC) sob orientação dos bibliotecários da instituição para ser adaptado às necessidades dos alunos dos Cursos de Graduação e Programas de Pós-Graduação da Universidade.

LUIZ FELIPE COSTA DANTAS DE SENA

DESENVOLVIMENTO DE FERRAMENTA COMPUTACIONAL PARA ANÁLISE DE DADOS DO BALANÇO ENERGÉTICO NACIONAL

Monografia apresentada à
Universidade Federal Rural do Semi-
Árido como requisito para obtenção do
título de Bacharel em Engenharia
Elétrica.

Orientador: Érica Mangueira Lima

Coorientador: Adriano Costa de Oliveira

Defendida em: 17/04/2024

BANCA EXAMINADORA

Documento assinado digitalmente
 ERICA MANGUEIRA LIMA
Data: 20/04/2024 14:33:25-0300
Verifique em <https://validar.iti.gov.br>

Érica Mangueira Lima,
Prof. Dra. (UFERSA)
Presidente

Documento assinado digitalmente
 ADRIANO COSTA DE OLIVEIRA
Data: 19/04/2024 23:22:33-0300
Verifique em <https://validar.iti.gov.br>

Adriano Costa de Oliveira,
Prof. Dr. (UFERSA)
Membro Examinador

Documento assinado digitalmente
 RODRIGO DE ALMEIDA COELHO
Data: 19/04/2024 23:42:13-0300
Verifique em <https://validar.iti.gov.br>

Rodrigo de Almeida Coelho,
Prof. Dr. (UFERSA)
Membro Examinador

Documento assinado digitalmente
 ISRAEL AIRES COSTA LEAL
Data: 20/04/2024 14:28:40-0300
Verifique em <https://validar.iti.gov.br>

Israel Aires Costa Leal,
Prof. Dr. (UNINASSAU)
Membro Examinador

AGRADECIMENTOS

Inicialmente, agradeço a todas as forças espirituais que regem o universo, pois em suas teias e tramas do destino me agraciaram com capacidade de batalhar por tudo que me lustrou os olhos, e por sempre me apresentarem um local de refúgio repleto de amor, compaixão e compreensão.

Agradeço de forma especial aos meus pais, que sempre depositaram toda fé existente em seu filho, aos meus avós que sempre contribuíram como esse sonho, a minha irmã por todo apoio prestado.

Agradeço aos meus “ismões” por serem colaboradores de todas minhas conquistas, por serem meu alicerce acadêmico, por terem me emprestado cada traço bom de suas personalidades. Ana Carla, agradeço a você por ter me dado todo o seu senso de justiça, sua proteção para com os seus amigos e sua fidelidade, sua amizade ultrapassa as barreiras do físico e transcendem para o espiritual, obrigado por dividir sua vida comigo. Pablo Vinicius, agradeço imensamente por aparecer na minha vida e ter colocado toda sua paixão em tudo que você toca, por sempre me acompanhar em inúmeros momentos de alegrias e dificuldades. Josilândia Santos, agradeço imensamente por sempre inundar os espaços ao seu redor com humildade, amor e serenidade, com certeza todo minério que você toca se torna uma pedra preciosa.

Agradeço também aos meus amigos Joyce Oliveira, Lamarque Braga, Letícia Benevides, André Stanley e Eulinia Queiroz, por sempre me resgatarem para o mundo real, me apoiando com conversas, momentos únicos e inúmeros sorrisos. Agradeço ao meu namorado Pedro Lucas por sempre me proporcionar momentos únicos de aconchego e apoio, por sempre me incentivar a atingir meus objetivos, sempre me amparar em momentos de dificuldades e me ajudar a ter chegado até este estágio da minha vida, deixo aqui meu amor, paixão e admiração por você.

Agradeço à minha orientadora Érica Manguiera Lima, por toda paciência, disponibilidade e apreço com as orientações. Suas contribuições foram essenciais para a edificação deste trabalho, e minha admiração por uma mulher tão incrível e forte só aumentaram com essa experiência.

RESUMO

A dinâmica energética do Balanço Energético Nacional (BEN) é marcada por uma transição gradual em direção a fontes renováveis, embora ainda seja dominada por matrizes não renováveis. Esta pesquisa abordou a análise do BEN utilizando técnicas de ciência de dados e modelagem estatística, com o objetivo de compreender as tendências atuais e projetar cenários futuros. Os dados revelaram um aumento significativo na produção de energia de fontes renováveis em 2022, mas ainda abaixo das fontes não renováveis. Uma análise visual utilizando gráficos gerados com Matplotlib e Seaborn destacou a relação entre a produção de energia renovável e a oferta interna total de energia, evidenciando um possível desequilíbrio entre produção e consumo para fontes não renováveis, potencialmente devido à importação generalizada de energia por muitos países. Além disso, a aplicação de modelos de previsão, como ARIMA, indicou um aumento esperado no consumo final energético entre 2022 e 2023, com potenciais aumentos na oferta de energia para matrizes não renováveis e uma diminuição para fontes renováveis. A análise do BEN reforça a importância da transição para fontes de energia mais sustentáveis, destacando a necessidade de políticas públicas e estratégias de investimento para mitigar os impactos das fontes não renováveis.

Palavras-chave: BEN, Python, Análise de dados, ARIMA, Energética.

ABSTRACT

The energy dynamics of the National Energy Balance (BEN) are marked by a gradual transition towards renewable sources, although it is still dominated by non-renewable matrices. This research addressed the analysis of BEN using data science techniques and statistical modeling, aiming to understand current trends and project future scenarios. The data revealed a significant increase in the production of energy from renewable sources in 2022, but still below non-renewable sources. Visual analysis using graphs generated with Matplotlib and Seaborn highlighted the relationship between renewable energy production and total internal energy supply, indicating a possible imbalance between production and consumption for non-renewable sources, potentially due to widespread energy importation by many countries. Additionally, the application of prediction models, such as ARIMA, indicated an expected increase in final energy consumption between 2022 and 2023, with potential increases in energy supply for non-renewable matrices and a decrease for renewable sources. The analysis of BEN reinforces the importance of transitioning to more sustainable energy sources, emphasizing the need for public policies and investment strategies to mitigate the impacts of non-renewable sources.

Palavras-chave: BEN, Python, Data Analysis, ARIMA, Energetic.

LISTA DE FIGURAS

Figura 1 - Esquemático de SGBD.....	12
Figura 2 - Ciclo da vida dos dados.....	13
Figura 3 - Célula referente a importação de bibliotecas virtuais.....	25
Figura 4 - Célula referente a função de plotar gráficos de linhas.....	26
Figura 5 - Célula referente a função de plotar gráficos de barras.....	27
Figura 6 - Célula referente a manipulação do DataFrame.....	28

LISTA DE GRÁFICOS

Gráfico 1: Produção de Energia Primária Renovável no Brasil.....	31
Gráfico 2: Produção de Energia Primária Renovável no Brasil em porcentagem.....	32
Gráfico 3: Produção de Energia Primária Não Renovável no Brasil.....	33
Gráfico 4: Produção de Energia Primária Não Renovável no Brasil em Porcentagem.....	34
Gráfico 5: Oferta de Energia Primária Renovável no Brasil.....	35
Gráfico 6: Oferta de Energia Primária Renovável no Brasil em Porcentagem.....	35
Gráfico 7: Oferta de Energia Primária Não Renovável no Brasil.....	36
Gráfico 8: Oferta de Energia Primária Não Renovável no Brasil em Porcentagem.....	36
Gráfico 9: Consumo Final Energético por Setor no Brasil.....	37
Gráfico 10: Consumo Final Energético por Setor no Brasil em Porcentagem.....	37
Gráfico 11: Previsão da Produção de Energia Renovável.....	39
Gráfico 12: Previsão da Produção de Energia Não Renovável.....	39
Gráfico 13: Previsão da Oferta Interna de Energia Renovável.....	40
Gráfico 14: Previsão da Oferta Interna de Energia Não Renovável.....	41
Gráfico 15: Previsão do Consumo Final Energético.....	42

SUMÁRIO

1. INTRODUÇÃO.....	10
1.1. Justificativa.....	10
1.2. Objetivo Geral.....	11
1.3. Objetivos Específicos.....	11
2. REFERENCIAL TEÓRICO.....	12
2.1. Dados e sua ciência.....	14
2.2. Introdução à ciência de dados.....	16
2.3. Aprendizado de máquina.....	17
2.4. Análise de dados.....	18
2.5. Estatística computacional.....	19
2.5.1. Distribuição de frequências.....	20
2.5.2. Regressão linear.....	21
2.5.3. Média móvel.....	22
2.5.4. Modelo ARIMA.....	22
3. METODOLOGIA.....	24
3.1. Estudo do caso.....	24
3.2. Pesquisa bibliográfica.....	24
3.3. Desenvolvimento e avaliação dos dados.....	27
4. Resultados e Discussões.....	32
5. Considerações Finais.....	46
REFERÊNCIAS.....	48

1. INTRODUÇÃO

Quando aborda-se os elementos essenciais para a manutenção da vida contemporânea, é incontestável que o setor energético se configura como um dos pilares fundamentais para a sustentabilidade de uma nação já que sua importância reside na capacidade de fornecer a energia necessária para impulsionar o desenvolvimento do país (GOLDEMBERG, MOREIRA, 2005). Com isso, Tolmasquim (2012) afirma que o Brasil está alinhado nas responsabilidades da área energética, se tornando referência internacional na produção devido a sua extensa oferta de matrizes energéticas.

Dessa forma, o reconhecimento do Brasil como uma potência global tanto em termos energéticos quanto ambientais não é uma afirmação exagerada nos dias atuais. A nação efetivamente se destaca por sua abundância de opções de produção provenientes de uma gama de fontes contando com a habilidade de realizar produções em grande escala que servem como modelos inspiradores para diversas nações (TOLMASQUIM, 2012).

De acordo com Gehrke, Goretti e Avila (2021), o mundo está imerso em uma era em que a busca por uma matriz energética sustentável torna-se imperativa, especialmente no Brasil. Nesse contexto desafiador, o BEN assume posição crucial como ferramenta indispensável para compreender e planejar nosso setor energético (EPE, 2022).

Portanto, a presente pesquisa busca analisar minuciosamente os dados do BEN, aprofundando-se nas informações tradicionais relativas à oferta e consumo de energia no Brasil. A investigação abrangerá as atividades de extração de recursos energéticos primários, sua conversão em formas secundárias, importação e exportação, distribuição e o uso final da energia. A partir dessa análise, será possível, mediante a construção de um modelo de análise e previsão de dados, mapear o comportamento da produção de energia primária, oferta interna de energia, consumo final por setor e, por fim, estimar seus respectivos valores para um período futuro de um ano.

1.1. Justificativa

O Brasil, enquanto figura central no cenário global de energia, enfrenta o desafio de desenvolver estratégias alinhadas com pesquisas avançadas sobre matrizes energéticas sustentáveis em harmonia com as demandas atuais e as responsabilidades ambientais. Esse contexto coloca o Brasil diante da complexa tarefa de equilibrar o desenvolvimento econômico com a preservação ambiental, demandando uma abordagem estratégica fundamentada em conhecimento aprofundado em dados (LUCENA, 2017).

Contudo, a predominância do setor privado nas atividades energéticas destaca a falta crítica de uma governança eficaz capaz de explicar detalhadamente as variáveis que influenciam a produção, distribuição e consumo de energia (GOMES, VIEIRA, 2009). Nesse cenário desafiador, o BEN surge como fonte vital de informações, fornecendo *insights* cruciais para fundamentar decisões e políticas no âmbito energético (EPE, 2022).

Dada a importância de estudar métodos de análise e mineração de dados como alternativa para identificar problemas de forma mais tangível, este trabalho apresentará teoricamente os processos para uma análise eficaz de bancos de dados, destacando a exploração de um banco de dados do BEN.

1.2. Objetivo Geral

Apresentar uma análise dos dados do BEN a fim de avaliar as principais variáveis com relação à oferta energética no Brasil.

1.3. Objetivos Específicos

- Apresentar técnicas de análise de dados adequadas para a análise necessária.
- Construir um modelo estatístico para prever dados com base em séries temporais.
- Especificar e caracterizar as variáveis relacionadas à geração, transmissão e consumo de energia.

- Aplicar o modelo de previsão e mostrar as previsões futuras das variáveis de geração, transmissão e consumo de energia nos próximos anos.
- Criar infográficos e mapas que mostrem todos os resultados da pesquisa.

2. REFERENCIAL TEÓRICO

O surgimento da utilização de dados em nível tecnológico é tido durante o período pós-guerra, uma vez que com o aumento exponencial do avanço tecnológico e o surgimento da internet, a quantidade de dados se expandiu. Com isso novas necessidades foram criadas, tanto no cenário social quanto no econômico, o mundo necessitava armazenar informações pontuais do que estava acontecendo e economia precisava adquirir a maior quantidade de insumos numéricos para compreender melhor todo cenário global econômico (Lóscio, 2011).

A partir disso, novas possibilidades foram sendo desenvolvidas com o decorrer do tempo, possibilidades estas correlacionadas com processamento de uma gama gigantesca de dados, previsões estatísticas, análises quali-quantitativas, novas técnicas computacionais modernas e a implementação dos dados de forma visual (Dos Passos, 2016).

De forma explícita, um dado pode possuir dois tipos, especialmente para computação. O primeiro tipo de dados são os dados quantitativos, que são representados por valores numéricos de tal forma que apenas seu valor e sua grandeza são relevantes para seu armazenamento. Em contrapartida existem os dados qualitativos que se agregam a valores nominais e ordinais (categorias), esses tipos de dados se tornam valiosos uma vez que seu valor está relacionado a sua semântica e ao reflexo do seu significado ao mundo real (Camilo, 2009).

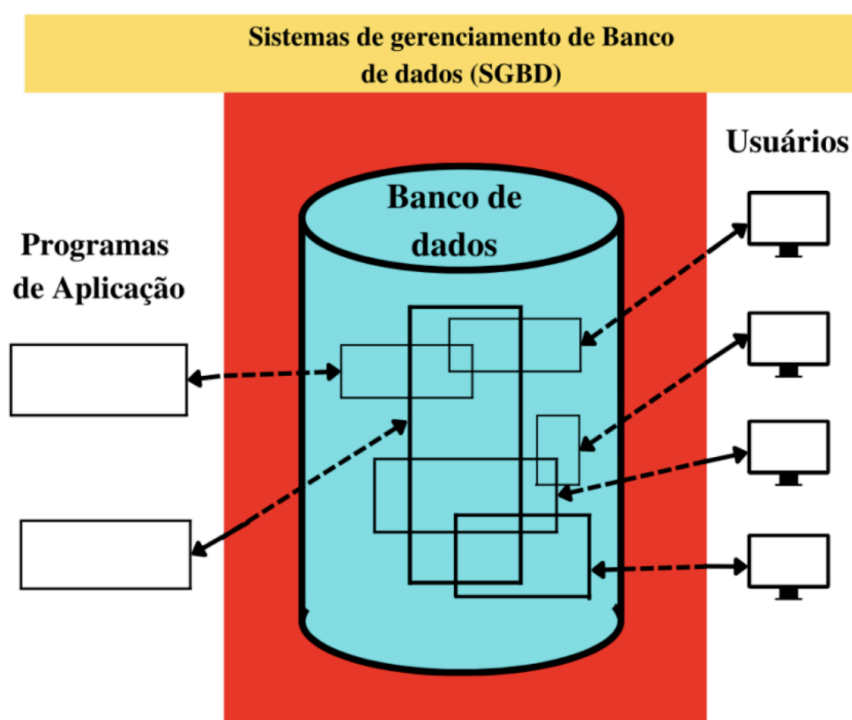
No que diz respeito ao estruturamento dos dados, eles se dividem em estruturados, semiestruturados e não estruturados. Para dados estruturados é importante seus conceitos de forma que seja possível vislumbrar que um dado estruturado esteja construído adotando uma organização padrão e apropriado de tabelas ou arquivos que utilizem valores separados por separadores. Já os dados semi estruturados estão entre entre estruturados e não estruturados, em que os

dados são estritamente tipados e não apresentam estrutura tabeladas ou separados. Por fim, entende-se como dados não estruturados aqueles que sua única importância está relacionada à quantidade de dados armazenados, como dados que acabaram de ser obtidos por sensores, sem se importar com sua organização (Elmasri e Navathe, 2015).

De acordo com Elmasri e Navathe (2015, p.3, grifo do autor), “Um banco de dados é uma coleção de dados relacionados. Com dados, quer se dizer fatos conhecidos que podem ser registrados e possuem significado implícito.” Por ser um reflexo de eventos analógicos, um banco de dados é projetado com o propósito de digitalizar esses fatos de tal forma que se torne possível a manipulação dessas informações para analisar o atual cenário de determinado evento, entender acontecimentos passados e principalmente como é utilizado atualmente, para prever novas situações.

A partir da necessidade de analisar os dados comportados dentro do banco de dados, fazendo com que eles se tornem relevantes e sejam bem trabalhados, surgiu o sistema de gerenciamento de bancos de dados que busca facilitar a interação de requisições do usuário para o banco de dados (Santos, 2007).

De acordo com Rob e Coronel (2011), “sistema de gerenciamento de bancos de dados (SGBD) é um conjunto de programas que gerenciam a estrutura do banco de dados e controlam o acesso aos dados armazenados”. O que se pode entender por um banco de dados é que, por meio de um software robusto, serve como organizador do *DataBase* e ponte entre os metadados e o usuário. O SGBD é dotado de uma estrutura que armazena o banco de dados possibilitando que o único serviço que possa atuar sobre os dados contidos ali seja o seu próprio sistema, apresenta-se na Figura 1 uma forma mais compreensível entre a interação dos usuários finais até os dados requisitados pelos usuários.

Figura 1 - Esquemático de SGBD

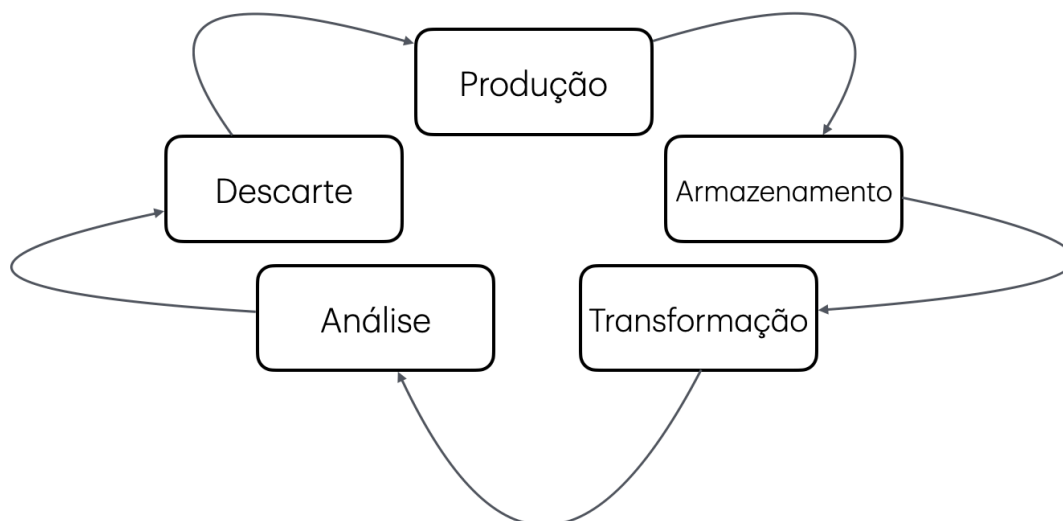
Fonte: Baseado em Date 2022.

2.1. Dados e sua ciência

Pode-se entender que o funcionamento da ciência de dados depende primordialmente de sua matéria prima bruta, os dados. Entretanto, eles correspondem apenas à base de toda pesquisa, o ponto de ignição para a correlacionar o mundo externo com o conhecimento científico pontual. (Cruz, 2018).

Por mais que extremamente indispensável, um dado não é suficiente para um bom entendimento da realidade inserida. Durante o início da análise de dados, deve-se compreender o que aqueles dados representam, se estão bem definidos, se serão de boa serventia para a pesquisa e se correlacionam entre si.

Por muitas vezes, a ciência de dados é ligada apenas a parte estatística dos dados e a apresentação dos dados de forma menos abstrata. Entretanto, a ciência dos dados está intimamente ligada a compreensão com o ciclo da vida do dado, que pode ser compreendido por meio da Figura 2.

Figura 2 - Ciclo da vida dos dados

Fonte: Baseado em Amaral, F. (2016, p. 06).

Para melhor compreender o ciclo da vida dos dados, será tomado como referência a Figura 2, que apresenta um ciclo infinito partindo da sua produção até o seu descarte. Vale salientar que nem todos os dados necessariamente passarão por todas as etapas. É possível que um dado seja descartado na fase de transformação ou até mesmo não precise ser transformado. Produção de dados refere-se a qualquer ação que gere conhecimento sobre algum fato da realidade, como sensores (produzindo dados de variações de inúmeras grandezas), câmeras (obtendo padrões de imagem), informações pessoais e jurídicas (dados obtidos pelo governo) (Amaral, 2016).

Logo após, esses dados devem ser persistidos em uma localização física de um dispositivo volátil com memória RAM. Nesta etapa encontra-se o estágio do armazenamento, em que alguns dados podem ser consumidos assim que obtidos, sem a necessidade de passar por etapas subsequentes. No entanto, para efeitos de continuidade, pressupõe-se que os dados seguirão para a etapa seguinte, a transformação. Nesta etapa, os dados sofrerão ações para alterar seu formato, de modo que possam ser utilizados da melhor forma quando analisados. Por exemplo, alguns dados na forma de agrupamento são comprimidos para otimizar seu

armazenamento. No entanto, todo material que sofre compressão acaba perdendo atributos, como qualidade e resolução (Amaral, 2016).

Em seguida, chega-se à etapa mais conhecida da ciência de dados: a análise. Esta etapa é fundamental, pois consiste na produção de conhecimento a partir dos dados. Esse conhecimento pode se manifestar em estatísticas, dados amostrais como tabelas, gráficos, mapas geográficos e modelos matemáticos de informações. Por fim, quando os dados não forem mais relevantes para o cientista, eles serão descartados (Porto E Ziviani, 2014).

Ao longo de todo o processo de pesquisa e desenvolvimento da tecnologia da informação, a área de ciência de dados se destaca como uma das mais importantes. Nela, expressa-se de forma relevante e modelada o material estudado, por meio da integração do analista com sua base de dados (Cruz, 2018).

Com o objetivo de tornar esse processo de interação entre base de dados e analista mais fluido e veloz, as técnicas de criação de algoritmos propõem uma vasta gama de possibilidades de manutenção e exploração desses dados (Formigoni, 2021).

Dessa forma, para obter essas análises, é necessário compreender a definição de alguns elementos, técnicas de análise e os processos que nelas estão inseridos, como *Data Science*, *Machine Learning* e *Data Analytics*.

2.2. Introdução à ciência de dados

A ciência de dados (Data Science) é definida por Gutierres Castanha (2021, p.1) quando ele discorre:

Entende-se a ciência de dados como um domínio de conhecimento multidisciplinar, que utiliza o método científico e técnicas de análise de dados, *machine learning* e inteligência artificial para resolver problemas reais, trazendo à luz conhecimento estratégico e *insights* acionáveis para o negócio.

Logo, em um mundo vivenciando um aumento quantitativo de dados, a ciência de dados age de forma norteadora, mostrando um caminho de passos para

adquirir, tratar, analisar e aplicar os dados obtidos, a fim de tornar a busca por soluções de problemas mais eficiente (Formigoni, 2021).

Devido ao volume descomunal de dados que são gerados diariamente, de forma crescente e em uma velocidade absurda, a busca pela criação de métodos para desmembrar esses dados e maximizar sua eficiência para a análise fez com que fossem desenvolvidas teorias computacionais que viabilizassem a extração dessas informações menores e mais úteis (Porto, 2014).

A partir dessas teorias aplicadas na prática comprovou-se a eficácia da ciência de dados para o estudo variado de dados. Com isso, amparando-se nesse raciocínio de *Data Science*, pode-se dizer que a ciência envolta dos dados é a melhor forma de analisar e prever um futuro a partir da progressão que esses dados apresentaram em sua análise (Formigoni, 2021).

O âmbito de estudo voltado à ciência de dados possui um caráter multidisciplinar, devido às suas áreas de abordagem, já que pode ser aplicada em qualquer área de conhecimento. Dentro de cada assunto abordado, será criada uma intersecção do tema com artifícios computacionais, permitindo apresentar, por fim, uma conclusão bem elaborada.

Diante dessa realidade, a ciência de dados pode ser aplicada na compreensão do comportamento da geração e oferta de energia com base no BEN. Utilizando bancos de dados, é possível extrair informações sobre os padrões da produção de energia em diferentes fontes de energia e aplicar manobras para exemplificar de forma quantitativa suas variáveis em séries temporais.

2.3. Aprendizado de Máquina

A partir dessa perspectiva, é importante mencionar o árduo trabalho do aprendizado de máquina, uma vez que é categorizado em um conjunto de tarefas que tem como principal objetivo realizar previsões de uma massiva quantidade de dados, possibilitando compreender e prever aspectos de qualquer situação exposta (Souza, 2021).

Tudo isso se tornou possível em 1959, quando o cientista da computação Arthur Samuel escreveu o primeiro algoritmo de aprendizado de máquina, no qual

ele mesmo conceituou o machine learning como uma área da computação que apresenta aos computadores tarefas que possibilitem o aprendizado sem a programação explícita do programa (Silveira, 2022).

Após ter esse breve entendimento do que é o aprendizado de máquina e de onde ele surgiu, deve-se atentar aos tipos de aprendizado de máquina, subdividindo-os em dois: o supervisionado e o não supervisionado. O aprendizado de máquina supervisionado está intimamente relacionado à prévia obtenção dos resultados. Dentro desse modelo, será estritamente necessário já conhecer seus dados finais, pois a partir deles o cientista irá analisar se os resultados que o aprendizado de máquina está apresentando são compatíveis com os que já foram obtidos (Silveira, 2022).

Em contrapartida, o tipo não supervisionado se torna mais denso e mais complexo, uma vez que não se conhece previamente os dados finais obtidos. Logo, deve-se analisar toda a estrutura para o caso e os tipos de previsões futuras (Silveira, 2022).

Portanto, entende-se que o aprendizado de máquina é um elemento da ciência de dados que eleva o nível da análise de dados para uma análise específica repleta de artifícios estatísticos e matemáticos, para que ultrapasse a visão analista de um humano, trazendo perspectivas e projeções únicas e indispensáveis (Cruz, 2018).

2.4. Análise de Dados

Com todo o conhecimento dos processos de apreciação de dados, um dos processos mais marcantes é a análise de dados, pois durante este momento o analista encontra-se com a visão humanitária dos dados produzida pelo cientista (Júnior, 2019).

Entende-se como visão humanitária a perspectiva que o analista deseja passar de forma fidedigna dos dados obtidos para os clientes. Com isso, o analista de dados se apodera de recursos computacionais para potencializar o seu processo de análise (Júnior, 2019).

Dessa forma, toda análise deve ser realizada a partir de critérios, preparações de material, sondagem dos dados, resultados e, por fim, sua apresentação. Transformando isso em etapas torna a análise mais fluida e sistemática. Partindo dos critérios, toda análise deve ser construída em torno de perguntas, uma vez que esta análise tem como objetivo vislumbrar a situação de algum problema ou até mesmo encontrar sua solução. Dessa maneira, definir perguntas iniciais ajuda o cientista a traçar o caminho de sua pesquisa até o encontro com o seu resultado (Pimentel et al., 2021).

Assim como foi citado no tópico 2.1 deste documento, o dado necessita passar por processos antes da análise. Um desses processos é a preparação, que conta com etapas de reunir os dados, avaliar e limpar. Tudo isso é feito para que os dados que são utilizados na análise sejam do mesmo padrão e possuam alta qualidade. Durante esse processo, alguns dados podem ser removidos ou substituídos (Camilo & Silva, 2009).

Após ter posse dos dados preparados e qualificados para a análise, deve-se iniciar sua sondagem, abordando estratégias de agrupamento e separação de dados, na tentativa de iniciar uma criação de perfil de dados. Dentro daqueles grupos, os dados possuem o mesmo padrão com o objetivo comum de poder transmitir a mensagem que reflita a realidade analisada (Tableau, 2022).

As últimas etapas conversam sempre entre si, uma vez que, com a obtenção dos resultados, os mesmos precisam ser comunicados aos clientes/espectadores da pesquisa. Nessa etapa, deve-se conglomerar os resultados em planilhas, gráficos, redes neurais, mapas e fluxogramas (Tableau, 2022).

2.5. Estatística Computacional

Como mencionado anteriormente, um cientista de dados tem como obrigação utilizar uma variedade de ferramentas estatísticas, matemáticas e computacionais para ser capaz de construir análises eficazes. Sabendo dessa importância, o atual tópico busca apresentar técnicas de estatística computacional que serão implementadas ao longo do processo exploratório e analítico.

Com isso, em qualquer análise de banco de dados, é necessário obter o conhecimento do comportamento dessas variáveis, para isso deve-se apropriar das

técnicas de estatísticas como distribuição de frequências e a apresentação de histogramas para ser possível visualizar como os elementos se comportam em relação ao tempo e por fim, compreender o conceito de previsão de séries temporais. Além disso, o método de aplicação do modelo de previsão ARIMA, para obter a previsão do caso estudado.

2.5.1. Distribuição de frequências

Nessa etapa todas as ações irão girar em volta da quantidade de amostras dentro de determinada categoria, possibilitando distribuição de frequências (F_i), que nada mais é do que o somatório do número de observações dentro de uma classe. Essa frequência é denominada simples, e o mais comum para compreensão geral é dada pela distribuição de frequência percentual ($F_i\%$), que se dá pela razão entre a frequência simples e número de observações total da classe, tudo isso multiplicado por uma constante de valor 100. Logo, o somatório de todas as frequências percentuais deve ser igual a 100% (Medeiros, [s.d]).

Os histogramas são úteis para entender a forma e a dispersão dos dados, identificar padrões e tendências, detectar *outliers* e comunicar resultados de maneira visual e intuitiva. Pois, são uma representação gráfica geralmente usada para visualizar a distribuição de frequências de uma variável quantitativa em intervalos, conhecidos como classes.

A construção de um histograma se inicia com a divisão dos dados em classes, em que os intervalos podem ser definidos manualmente ou utilizando métodos automáticos, como a regra de Sturges, que consiste em dividir o número de classes de um histograma de acordo com a Equação 1,

$$k = 1 + 3.322 \log(n), \quad (1)$$

sendo k é o número de classes e n é o tamanho da amostra. Depois disso, para cada classe, é contado o número de observações que se encaixem no intervalo. No gráfico os intervalos das classes são representados ao longo do eixo horizontal, enquanto as frequências correspondentes são representadas ao longo do eixo vertical, dessa forma cada classe é geralmente representada por uma barra vertical em que a altura indica a frequência daquela classe (ALURA, 2023).

A interpretação de um histograma é crucial para entender a distribuição dos dados. Com um histograma é possível revelar se os dados estão distribuídos de forma simétrica ou assimétrica, qual tendência a curva está apresentando, se apresentam outliers, ou outros padrões importantes (ALURA,2023). Portanto, está é uma ferramenta essencial na análise exploratória de dados, sendo amplamente utilizados em diversas áreas, como ciência de dados, estatística e engenharia.

2.5.2. Regressão linear

Compreende-se por regressão linear simples, um artifício matemático cujo objetivo seja obter análises preditivas por meio de uma reta que seja dependente do diagrama de dispersão de dados. Esse conjunto de dados, deverá apresentar duas variáveis, uma é preditora e uma variável dependente (Duarte, 2022).

De forma generalista, a reta da regressão linear (Y) é dada pelo somatório de componentes equivalentes ao intercepto no valor 0 (zero) dos dados (A), o coeficiente angular (B) multiplicado pela variável explicativa (x) somada com o erro aleatório (e) (Duarte, 2022). Logo, essa função é matematicamente representada como a Equação 2,

$$Y = A + Bx + e \quad (2)$$

em que:

- Y é a variável dependente que se procura prever;
- x é a variável preditora;
- A é o intercepto da reta de regressão, representando o valor de Y quando x é igual a zero;
- B é o coeficiente angular da reta de regressão, indicando a taxa de mudança em Y para cada unidade de mudança em x ;
- e é o termo de erro, que representa a diferença entre os valores observados de Y e os valores previstos pela reta de regressão.

Essa equação nos permite prever o valor de Y para um dado valor de x com base nos parâmetros estimados A e B e no erro aleatório e , o objetivo da análise de regressão é encontrar os melhores valores de A e B que minimizem a soma dos

quadrados dos resíduos, ou seja, a diferença entre os valores observados de Y e os valores previstos pela reta de regressão.

2.5.3. Média móvel

A média móvel é uma ferramenta estatística amplamente utilizada na análise de séries temporais para suavizar flutuações aleatórias e identificar tendências ao longo do tempo. Matematicamente, é calculada como a média dos valores em um intervalo de tempo específico, deslizando-se essa janela ao longo da série temporal. Essa técnica é útil para filtrar ruídos e destacar mudanças de longo prazo nos dados (Hyndman & Athanasopoulos, 2018).

Por exemplo, ao analisar vendas ao longo de vários anos, podemos calcular a média móvel de três meses, tirando a média das vendas de janeiro, fevereiro e março, depois de fevereiro, março e abril, e assim por diante. Isso nos dá uma série suavizada que nos ajuda a entender melhor as tendências subjacentes nas vendas.

No entanto, é importante considerar algumas limitações. O tamanho da janela de observação pode afetar a sensibilidade da técnica às mudanças nos dados, e uma janela muito pequena pode capturar apenas variações de curto prazo, enquanto uma janela muito grande pode suavizar demais a série. Além disso, a média móvel pode introduzir um atraso nas respostas às mudanças nos dados originais. Portanto, é essencial ajustar a janela de observação de acordo com os objetivos analíticos específicos da análise.

2.5.4. Modelo ARIMA

O modelo ARIMA foi sistematizado em 1976 pelos estatísticos George Box e Gwilym Jenkins, seu nome consiste em ser um acrônimo, ou seja, é um aglomerado de letras que possui o primeiro caractere de uma palavra. No caso de ARIMA o AR está indicando a utilização de um modelo autorregressivo, que se conceitua em um modelo que apresenta funções de regressão linear e que auto correlacionam com a declinação dos pontos na direção do valor nulo (zero), a partir disso se torna possível mostrar um padrão dos dados analisados (ALURA, 2022).

O seu caracter I vai indicar que o modelo é integrado, basicamente irá definir quantas vezes essa série temporal irá ser diferenciada, tornando-a uma série estacionária. Por fim, o MA apresenta que o modelo contará com uma média móvel que possibilita vislumbrar a tendência desse conjunto de dados analisados derivados de uma série temporal (UFV, 2020).

Para cada característica do acrônimo ARIMA é adotada uma variável, que se encontra a correlação entre seus indicadores e variáveis, com uma adição no modelo SARIMA, o S representa a sazonalidade para séries temporais bem definidas com seus períodos sazonais. Em ARIMA, o AR corresponde ao parâmetro p , o I ao parâmetro d , e o MA ao parâmetro q .

Uma vez que o conceito do modelo ARIMA é familiar, é necessário compreender algumas regras, a primeira se dá quando for necessário obter previsões de séries temporais e a amostra possui um elevado número de observações, então é necessário o ajuste de modelos autorregressivos. Entretanto, se o número de observações for baixo, essa análise não é apropriada, uma vez que a hipótese básica de independência dos resíduos não será validada, caso isso ocorra deve-se partir para métodos de alisamento (Roger, 2015).

No instante em que for requisitado a análise de modelos paramétricos, um método recomendado é o método de Box e Jenkins, que consiste em ajustar os modelos autorregressivos integrados com médias móveis (ARIMA) a um conjunto de dados. Esse ajuste é realizado dentro de um ciclo iterativo, cujo seus estágios são divididos em especificação, será considerado uma classe geral de modelo, identificação, etapa correspondente por analisar autocorrelações e autocorrelações parciais, estimação, instante que será realizado a estimação dos parâmetros do modelo e por fim, a verificação do modelo ajustado por meio de uma análise de resíduos para compreender se está adequado (Roger, 2015).

3. METODOLOGIA

3.1. Estudo do caso

O foco central deste trabalho consiste em desenvolver uma Análise de Dados voltado para um banco de dados do BEN, no qual foi disponibilizado pela própria organização de forma a possuir um livre acesso para cumprir com os objetivos traçados. Dentro da realização dessa análise, será necessário encontrar um conjunto de ferramentas de análise de dados para obter conhecimento no âmbito das ciências de dados, referente a modelagem de modelos de aprendizado de máquina, para que esse banco de dados possa ser explorado e apresentar resultados de forma clara e explicativa.

Nesse caso, o banco de dados possui um modelo mediano em relação a organização de dados uma vez que por mais que seus arquivos sejam disponibilizados em formato xlsx, não está organizado da melhor forma. Além disso, o banco de dados possui dados referentes à produção, oferta e consumo de energia entre o período de tempo de 1970 a 2022. Dessa forma, encontra-se a necessidade de organizar os dados, e explorá-los de forma que seja possível a compreensão do comportamento dos casos apresentados.

3.2. Pesquisa Bibliográfica

Inicialmente, foi realizado um levantamento de caráter bibliográfico, com o intuito de explorar conteúdos em torno à mineração de dados que possam ser aplicados em um banco de dados do BEN, que trata dos estudos e análises orientadas para o planejamento do setor energético, e a partir desses dados desenvolver um algoritmo exploratório que cumpra os objetivos traçados. Com isso, foram abordados vários materiais acadêmicos, como livros didáticos e artigos sobre análise e mineração de dados, dando ênfase à aplicação dessas técnicas em um banco de dados.

Logo após, teve-se como material abordado manuais e apostilas referentes à linguagem de programação Python, que é classificada como uma linguagem de

programação para casos gerais; entretanto, possui uma sintaxe clara proporcionando ao usuário maior legibilidade do algoritmo, assim como um vasto acervo de bibliotecas voltadas para análise e mineração de dados (Borges, 2014).

Por se tratar de uma das linguagens mais bem avaliadas no âmbito científico e por condensar tantas qualidades e variedades de aplicações, esta linguagem foi a escolhida, e com isso foram determinadas algumas bibliotecas que possuem a capacidade de criar um ambiente virtual voltado para o desenvolvimento de aplicações voltadas à ciência de dados (Manzano, 2018).

Referente ao ambiente virtual criado, o terminal que possibilita isso é o terminal interativo do IPython; nele será possível utilizar de forma livre a linguagem Python para resultados rápidos e programas relativamente pequenos e pontuais. Ou seja, é um ambiente que está voltado para apresentar resultados de cada célula escrita e armazenar o código em partes fragmentadas (IPython, 2022).

Para a utilização do IPython, existe uma ferramenta chamada Jupyter Notebook, na qual foi apresentada uma maneira de mesclar o algoritmo codificado e sua documentação, possibilitando a criação de um arquivo narrado com inúmeros elementos de códigos independentes, como funções, scripts e declarações. Durante a utilização do Jupyter Notebook, o usuário pode tanto focar em explicar ao computador o que deve ser feito, como também mostrar de forma explicativa ao ser humano o que aquela célula algorítmica representa a outros seres humanos (Notebook, 2022).

Com todas as ferramentas para a criação do ambiente virtual já bem definidas, foram estudadas algumas bibliotecas referentes à linguagem Python, todas voltadas para análise de grandes quantidades de dados, matemática computacional, visualização de gráficos e análises preditivas com regressões lineares, o que se mostra necessário para as técnicas que serão implementadas.

Quando se refere à análise de um *DataFrame* em Python, a biblioteca pandas apresenta todos os recursos necessários para o desenvolvimento de sua atividade, uma vez que a partir dela torna-se possível trabalhar com dados tabulares, dados temporais, matrizes ou qualquer conjunto de dados. Dessa forma, foi escolhida devido à facilidade de utilização, uma vez que permite uma vasta leitura de conjunto

de dados e possui a possibilidade de usar operações aritméticas entre colunas e linhas (Pandas, 2022).

Sabendo da necessidade de cálculos computacionais para *arrays*, o pacote da biblioteca NumPy possibilita o cálculo de funções matemáticas e cálculos estatísticos. Dessa maneira, foi escolhido por meio da afinidade da biblioteca com a execução de cálculos numéricos, sua rapidez quando apresentado a grandes quantidades de dados e sua economia de memória ao realizar suas funções (Numpy, 2022).

Como todo o intuito do trabalho está voltado para *DataFrame* e *Arrays* para a visualização de boa parte dos resultados, é necessário a geração de gráficos bidimensionais; para isso remeteu-se ao MATLAB e a sua integração com Python, a biblioteca Matplotlib. Vale ressaltar que atrelada a esta biblioteca está a anteriormente citada NumPy, já que necessita da apropriação de operações com arrays para o dimensionamento dos gráficos propostos (Matplotlib, 2022).

Seaborn foi desenvolvido com base no Matplotlib e no Pandas, com foco na plotagem de gráficos. Embora compartilhe semelhanças com o Matplotlib, possui recursos exclusivos que o distinguem. *Seaborn* é especialmente orientado para a plotagem de gráficos estatísticos, permitindo extrair informações significativas diretamente do *DataFrame* graças à sua integração com o Pandas. Elevando o nível da visualização de dados, surgiu a necessidade de apropriar-se da biblioteca Plotly Python, que contém funções voltadas à compreensão de dados de forma simples. Ademais, o retorno da função nos proporciona uma interface de visualização de alto nível, possibilitando interações com o display apresentando dados pontuais (PLOTLY, 2022).

Finalizando a parte de visualização e análise de dados, a etapa seguinte foi adotar uma biblioteca para modelagem de algoritmos de *machine learning*, logo foi adotado como base a biblioteca *scikit-learn*, que dispõe de ferramentas para análise preditiva de dados (Scikit-learn, 2022).

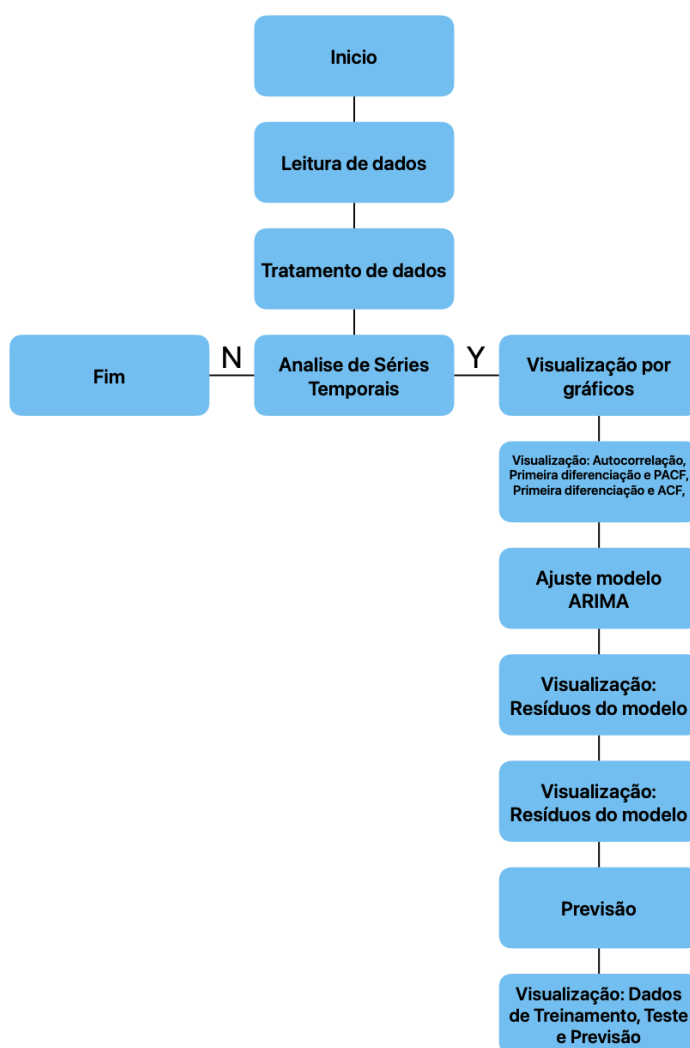
Contudo, um dos objetivos é a previsão dos dados, e para isso mais uma biblioteca será utilizada, a *statsmodels* que tem como foco ser uma ferramenta voltada para modelos estatísticos de previsões de dados temporais (Statsmodels, 2022).

Portanto, todo o processo analítico e experimental desse trabalho estará contido na divisão de estágios entre o desenvolvimento e a avaliação de dados. Com isso, o *DataBase* será tratado e nos resultados e discussão será apresentado o retorno do algoritmo e toda discussão em volta ao assunto.

3.3. Desenvolvimento e avaliação dos dados

Antes de iniciar o desenvolvimento, foi feito um fluxograma que ilustra como foi realizado o todo o processo analítico. O fluxograma está representado na Figura 3.

Figura 3 - Fluxograma de desenvolvimento de análise e previsão de dados

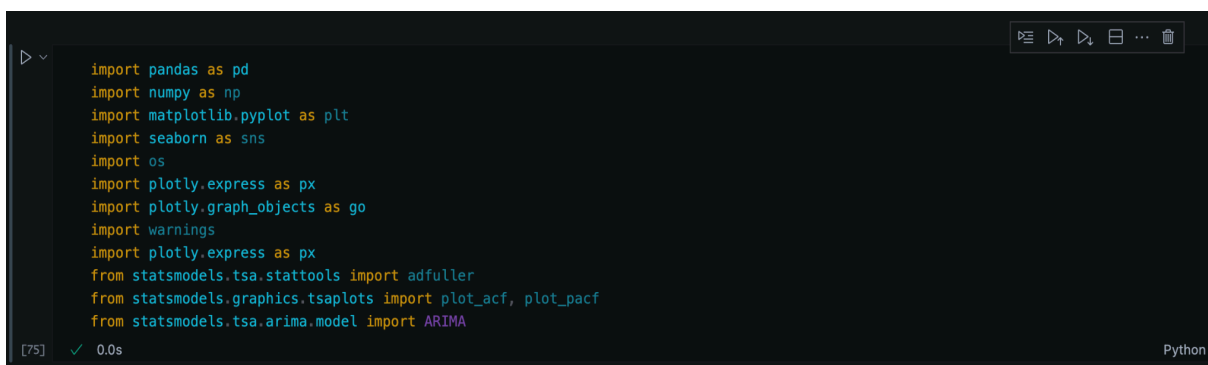


Fonte: Autoria Própria, 2024.

O início do desenvolvimento do processo analítico é dado a partir do momento da coleta dos dados, em que foi adquirido uma base de dados do BEN,

base essa disponível de forma *open-source*. Logo após adquirir o banco de dados, a ação seguinte foi importá-lo para o ambiente virtual do Jupyter Notebook, junto com a importação das bibliotecas futuramente utilizadas, onde foi possível observar os dados contidos no *DataFrame* e seus respectivos tipos de dados, como representado na Figura 4.

Figura 4 - Célula referente a importação de bibliotecas virtuais.

A screenshot of a Jupyter Notebook cell with a dark background. The cell contains a series of Python import statements for data science and visualization libraries. The code is as follows:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import os
import plotly.express as px
import plotly.graph_objects as go
import warnings
import plotly.express as px
from statsmodels.tsa.stattools import adfuller
from statsmodels.graphics.tsaplots import plot_acf, plot_pacf
from statsmodels.tsa.arima.model import ARIMA
```

The cell status bar at the bottom left shows "[75]" and a green checkmark with "0.0s". The bottom right corner of the notebook interface shows the word "Python".

Fonte: Autoria Própria, 2024.

Em seguida surgiu a necessidade de construir funções para que toda a exibição da análise seja feita de forma modular e padronizada, para isso é desenvolvida funções para plotagem de gráficos de linhas e barras de forma que sua apresentação seja iterativa, ou seja, ao passar o cursor seja possível visualizar etiquetas informando os valores pontualmente.

A função de plotar gráficos de linhas foi desenvolvida para proporcionar uma maneira eficiente de plotar gráficos de linhas a partir de um *DataFrame* fornecido como entrada. Este método foi construído com base na biblioteca Plotly Express, que oferece uma interface de alto nível para criar visualizações interativas de dados. Para utilizar essa função, é necessário fornecer três parâmetros obrigatórios: o *DataFrame* (df), o título do gráfico (title) e as colunas que representam os eixos x e y do gráfico (x='Ano' e y='Valor', respectivamente). Além dos parâmetros obrigatórios, a função *create_lineplot* também oferece dois parâmetros opcionais: *width* e *height*, que controlam a largura e altura da visualização do gráfico, respectivamente. Por padrão, esses parâmetros estão definidos como *width*=1600 e *height*=600, mas podem ser ajustados de acordo com as preferências do usuário para otimizar a apresentação visual.

Internamente, a função cria um objeto de figura (fig) usando a função `px.line()` da Plotly Express. Esta função recebe o *DataFrame* fornecido, especifica as colunas para os eixos x e y, e opcionalmente, a coluna de cores ('FONTES'), que permite visualizar diferentes séries de dados com cores distintas no gráfico. Além disso, são definidos o título do gráfico e rótulos personalizados para os eixos x e y. Posteriormente, a função ajusta o layout da figura (`fig.update_layout()`), definindo a largura e altura conforme especificado pelos parâmetros opcionais *width* e *height*. Por fim, o gráfico deverá ser exibido na saída utilizando o método `fig.show()`, permitindo ao usuário visualizar e interagir com o gráfico gerado diretamente no ambiente de execução. O código utilizado está descrito na Figura 5.

Figura 5 - Célula referente a função de plotar gráficos de linhas.



```
def create_lineplot(df, title, width=1600, height=600,):
    fig = px.line(df, x='Ano', y='Valor', color='FONTES', title=title, labels={'Valor': 'Energia (106 tep)'})
    fig.update_layout(width=width, height=height)
    fig.show()
```

Fonte: Autoria Própria, 2024.

Em seguida, foi construída uma função que permite gerar gráficos de barras empilhadas a partir de um *DataFrame* fornecido, representando a distribuição percentual de diferentes categorias ao longo do tempo. Internamente, os dados são organizados utilizando o método `pivot` do *DataFrame* para reorganizar as categorias e períodos. Em seguida, as barras empilhadas são criadas utilizando a biblioteca Plotly, em que cada barra representa uma categoria e sua altura reflete a proporção percentual em relação ao total. A função `create_stacked_bar_chart` requer três parâmetros obrigatórios: o *DataFrame* (*df*), que contém os dados a serem visualizados, o título do gráfico (*title*) e as colunas que representam os períodos de tempo e as categorias. Além disso, há dois parâmetros opcionais para personalizar a largura e altura do gráfico (`width=1600` e `height=800`, respectivamente).

O retorno da função é a exibição do gráfico de barras empilhadas, que representa a distribuição percentual das categorias ao longo do tempo. Esse gráfico é visualizado diretamente no ambiente de execução após a chamada da função. Ele

mostra a contribuição de cada categoria em relação ao total em cada período, permitindo uma análise clara das tendências e padrões ao longo do tempo.

Figura 6 - Célula referente a função de plotar gráficos de barras.



```
def create_stacked_bar_chart(df, title, width=1600, height=800):
    df_pivot = df.pivot(index='Ano', columns='FONTES', values='Valor')
    df_percent = df_pivot.divide(df_pivot.sum(axis=1), axis=0)

    df_percent.reset_index(inplace=True)
    fig = go.Figure(data=[
        go.Bar(name=col, x=df_percent['Ano'], y=df_percent[col]) for col in df_percent.columns if col != 'Ano'
    ],
    layout=dict(width=width, height=height))

    fig.update_layout(barmode='stack', xaxis_title='Ano', yaxis_title='Porcentagem de Produção', title=title)

    fig.show()
```

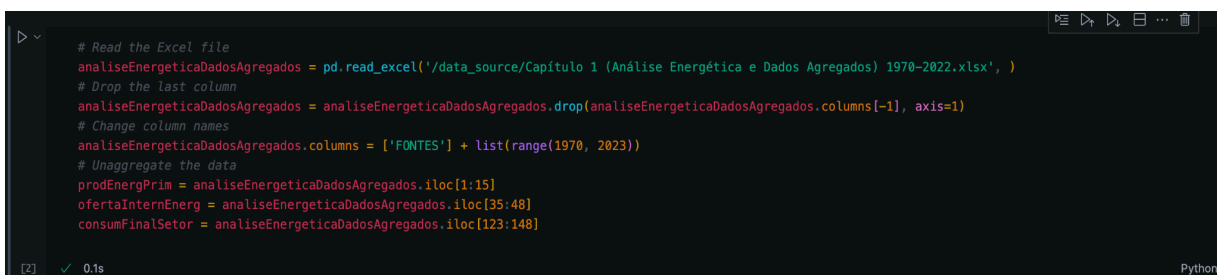
[1] ✓ 4.1s Python

Fonte: Autoria Própria, 2024.

Com as funções já definidas, só é necessário chamar a função passando os argumentos requisitados por cada função e assim é possível obter *insights* relevantes sobre a base de dados analisada.

Portanto, um dos parâmetros em comum de toda função desenvolvida para a análise é o *DataFrame*, com isso ao analisar o conjunto de dados obtido, é visto a necessidade de separar toda a base de dados em pequenas partições e alocar em variáveis, para ter um nível superior de organização e padronização. Essa manobra de engenharia de dados é de extrema importância para que seja possível limpar todas as informações inadequadas para a análise e no fim obter um produto coeso. Essas modificações estão ilustradas na Figura 7.

Figura 7 - Célula referente a manipulação do *DataFrame*.



```
# Read the Excel file
analiseEnergeticaDadosAgregados = pd.read_excel('/data_source/Capítulo 1 (Análise Energética e Dados Agregados) 1970-2022.xlsx', )
# Drop the last column
analiseEnergeticaDadosAgregados = analiseEnergeticaDadosAgregados.drop([analiseEnergeticaDadosAgregados.columns[-1]], axis=1)
# Change column names
analiseEnergeticaDadosAgregados.columns = ['FONTES'] + list(range(1970, 2023))
# Unaggregate the data
prodEnergPrim = analiseEnergeticaDadosAgregados.iloc[1:15]
ofertaInternEnerg = analiseEnergeticaDadosAgregados.iloc[35:48]
consumFinalSetor = analiseEnergeticaDadosAgregados.iloc[123:148]
```

[2] ✓ 0.1s Python

Fonte: Autoria Própria, 2024.

Por fim, com as funções de plotagem de gráficos e com as bases de dados organizadas é possível abordar a construção da função responsável por analisar e treinar o modelo ARIMA para cada base de dados.

Inicialmente, os dados são preparados por meio da filtragem do *DataFrame* para o campo de interesse e da conversão dos valores para o formato numérico, eliminando quaisquer entradas inválidas. A série temporal é então indexada pelo ano para facilitar a análise ao longo do tempo.

Uma análise exploratória de dados (EDA) é realizada para entender a estrutura dos dados. Isso inclui a plotagem da série temporal original para visualizar tendências e padrões sazonais, além de um gráfico de autocorrelação para identificar possíveis correlações entre os valores em diferentes intervalos de tempo.

A identificação dos componentes temporais do modelo ARIMA é realizada por meio da diferenciação da série temporal e da análise da função de autocorrelação parcial (PACF). A diferenciação é aplicada para remover tendências e sazonalidades, tornando a série estacionária. A PACF é uma ferramenta que ajuda a determinar a ordem dos termos autorregressivos no modelo ARIMA. Os picos significativos na PACF indicam os possíveis lags nos quais os efeitos autorregressivos são mais pronunciados, representando os intervalos de tempo anteriores em que há uma correlação considerável entre a observação atual e aquelas em lags anteriores, mesmo após remover o efeito da correlação nos lags intermediários.

Quando ajustamos o modelo ARIMA aos dados utilizando a classe ARIMA do pacote statsmodels, é importante avaliar a qualidade do modelo por meio da análise dos resíduos. Os resíduos são a diferença entre os valores observados e os valores previstos pelo modelo. Um bom ajuste do modelo é indicado por resíduos que não mostram nenhum padrão discernível quando plotados ao longo do tempo e uma distribuição normal quando plotados em um histograma. Resíduos normalmente distribuídos e com variância constante indicam que o modelo capturou adequadamente a estrutura dos dados.

Para realizar previsões futuras, os dados são divididos em conjuntos de treinamento e teste. O modelo ARIMA é ajustado novamente aos dados de treinamento, e as previsões são geradas para um determinado número de períodos futuros. Essas previsões são acompanhadas por intervalos de confiança para avaliar a incerteza associada às previsões. Portanto, com essas funções e manipulações de dados realizadas, é possível adquirir informações valiosas para orientar decisões em diversas áreas de aplicação.

4. Resultados e Discussões

Como descrito no corpo deste trabalho, seu objetivo percorre em torno da análise dos dados do BEN, para que por meio dessa análise fossem apresentadas informações gráficas para apresentar o atual cenário nacional referente a produção, oferta e consumo de energia, assim como a implementação de um modelo estatístico capaz de estimar valores futuros e prever a situação nacional em um cenário futuro de n anos.

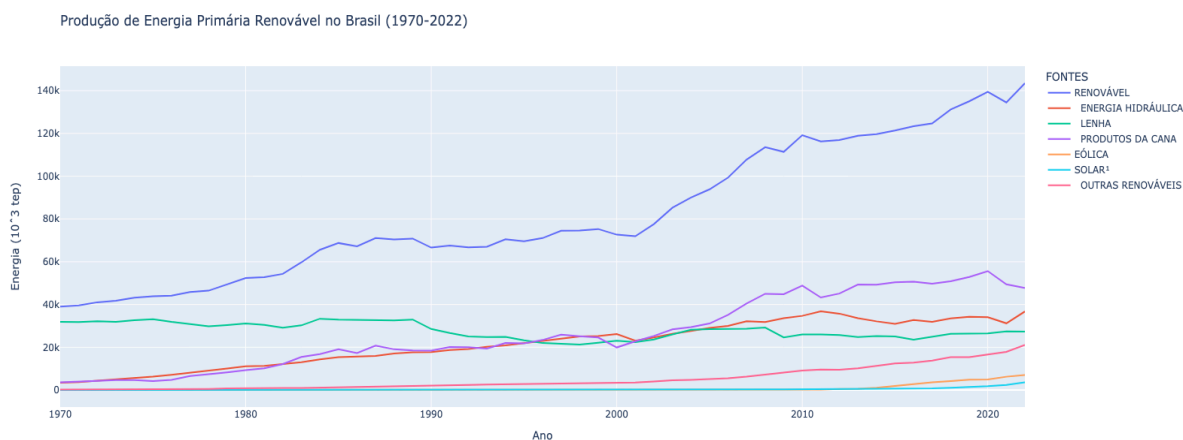
Os dados foram coletados e avaliados, passando por um processo de quali-quantificação, transformando dados brutos e consideravelmente desorganizados, em informações prontas para serem consumidas por algoritmos diversos de *DataVisualization* e *ModelStatistics*, que consistem respectivamente em

técnicas de aprimoramento de visualização gráfica e aprendizado de máquina.

Logo, a primeira análise foi realizada em torno da quantificação, para isso houve a utilização da função *create_lineplot* que analisou toda a variável responsável pelos dados de produção de energia renovável, e por fim gerou o Gráfico 1.

No gráfico, nota-se uma linha azul escura representando a energia renovável total, a soma de todas as fontes. Observa-se um crescimento significativo após os anos 2000. Considerando o período de 1970 a 2022, essa matriz energética experimentou um aumento de mais de 100 milhões de toneladas equivalentes de petróleo (tep). Um tep, ou tonelada equivalente de petróleo, é uma unidade de medida que representa a quantidade de energia liberada pela queima de uma tonelada de petróleo bruto. É uma medida comumente utilizada para comparar diferentes fontes de energia.

Gráfico 1: Produção de Energia Primária Renovável no Brasil.



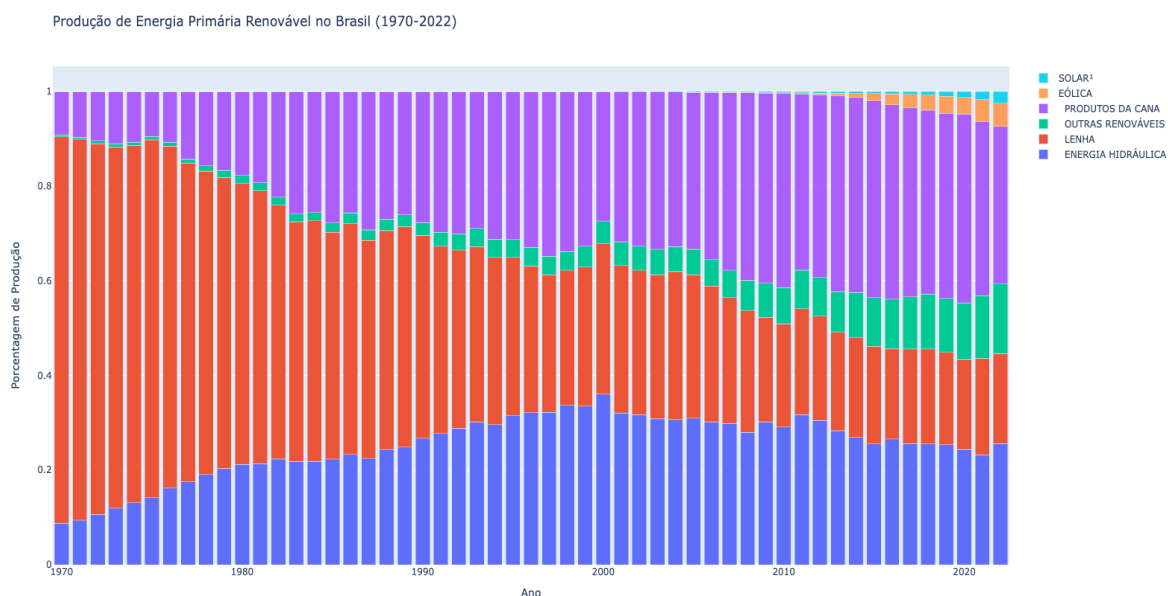
Fonte: Autoria Própria, 2024.

Além disso, é possível visualizar outras fontes energéticas, como a lenha que se mostrou ser a primeira indicada como maior produção energética renovável até metade da década de 90, quando foi superada pelo aumento da produção energética por meio dos produtos da cana, e energia hidráulica e por fim a produção a partir dos produtos da cana se mostra o maior produtor acima de todas as outras fontes, além disso é possível notar o surgimento da energia eólica e solar nos

últimos anos analisados.

Para uma visualização mais detalhada de todas as fontes energéticas, será utilizado a função `create_stacked_bar_chart` para plotar o Gráfico 2, em que é apresentado a porcentagem de cada fonte energética em relação à totalidade produzida de energia renovável. Mostrando assim a contribuição de cada fonte para com o todo.

Gráfico 2: Produção de Energia Primária Renovável no Brasil em porcentagem.



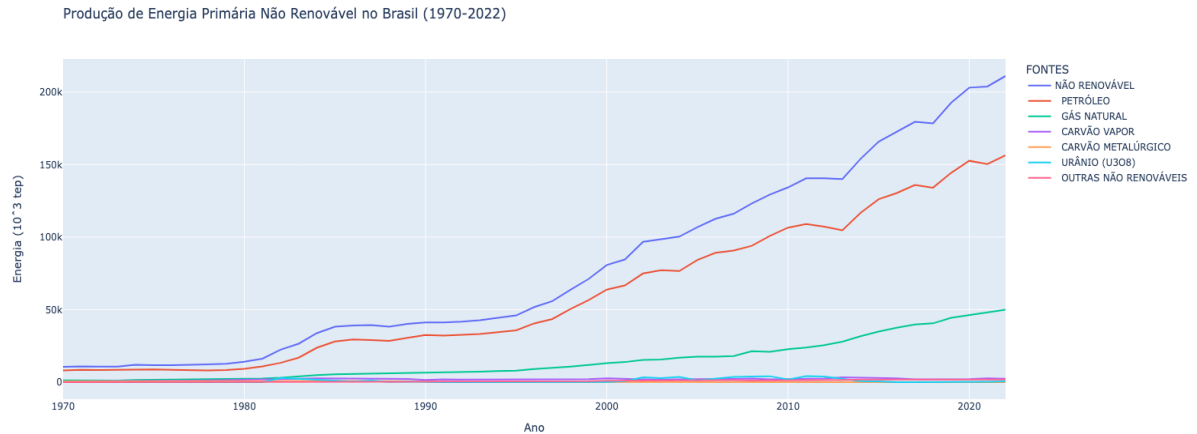
Fonte: Autoria Própria, 2024.

Nesse gráfico de barras é possível ver um funil na região de cor vermelha correspondente a lenha, que em seu início (1970) possuía 81,15% da totalidade de energias renováveis e no ano de 2022 ela apresenta apenas 19,01% da grandeza, apresentando uma queda na sua utilização. Em contrapartida, tem-se os produtos da cana crescendo 24,05% desde 1970, além da cana é possível encontrar a onda da energia hidráulica que em 1970 possuía 8,76%, em 2000 atingiu o seu máximo de produção com 36,03% e se estabilizou por volta dos 25%. É interessante notar o surgimento de novas fontes a partir de 2010, em que a energia solar e eólica iniciou seu processo de tomada de espaço dentre as outras fontes renováveis.

A segunda análise foi também realizada em torno da quantificação. Usando a função `create_lineplot`, que analisou toda a variável responsável pelos dados de

produção de energia não renovável resultando assim o Gráfico 3.

Gráfico 3: Produção de Energia Primária Não Renovável no Brasil.

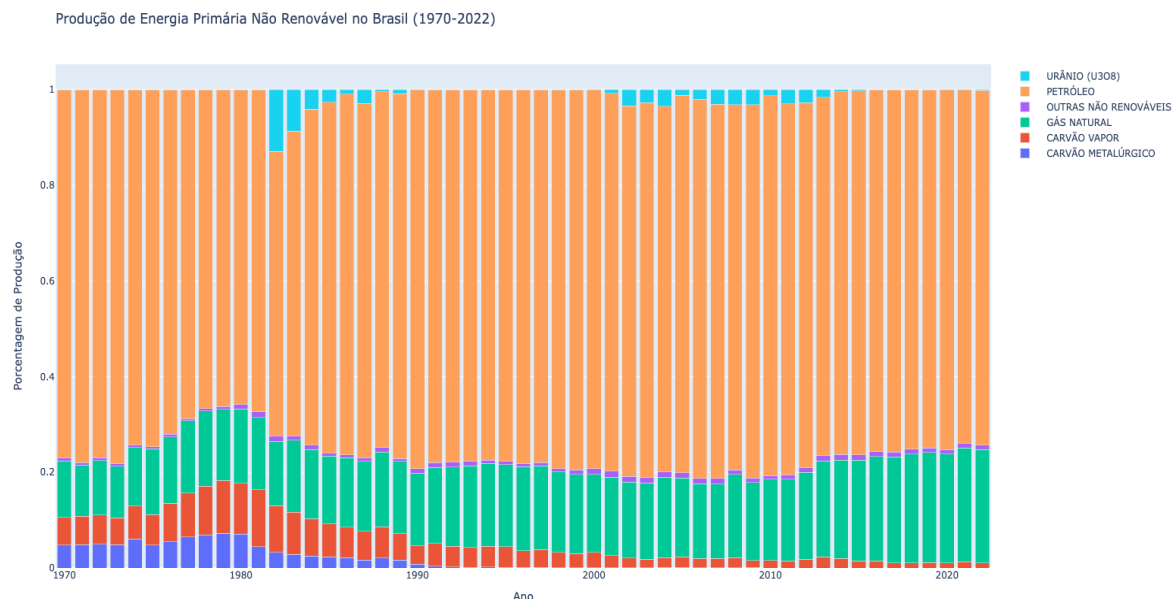


Fonte: Autoria Própria, 2024.

No gráfico é possível observar uma linha azul escuro que representa a totalidade da energia não renovável, sendo o somatório de todas as fontes não renováveis. Ao acompanhar o desenrolar dessa linha é visto que houve um crescimento considerável após 1990 devido ao aumento da produção energética a partir do uso do petróleo, nessa época é dado início a crescimento lento porém constante do uso de gás natural se tornando o segundo maior produtor energético do país.

Agora para uma análise mais abrangente das diferentes fontes de energia, foi utilizada a função `create_stacked_bar_chart` para criar o Gráfico 4. Esse gráfico mostra a proporção de cada fonte de energia em relação à produção total de energia não renovável. Dessa forma, é possível visualizar a contribuição de cada fonte para o conjunto geral de energia não renovável.

Gráfico 4: Produção de Energia Primária Não Renovável no Brasil em Porcentagem.



Fonte: Autoria Própria, 2024.

Nesse gráfico é notório a presença do petróleo como maior fonte energética não renovável, além dele encontra-se o gás natural tendo o seu início com 11,84% e chegando a 2022 com 23,67%. É interessante notar o comportamento do carvão metalúrgico que se extinguiu em 1994 sendo que antes chegou a obter uma produção de 7,25% da totalidade, outra fonte que possui um comportamento semelhante é o carvão vapor, quando em 1970 possuía 5,77% da totalidade, posteriormente atingiu 11,04% em 1979. Porém com o decorrer dos anos está diminuindo chegando em 2022 com 1,09%.

Além da análise detalhada das fontes energéticas não renováveis, é crucial considerar também os dados da oferta interna de energia para uma compreensão abrangente do panorama energético. O Gráfico 4 revela informações necessárias para a compreensão do panorama nacional da oferta de energia para fontes renováveis. Já o Gráfico 5, é apresentado a composição das fontes de energia renováveis em termos de porcentagem, assim como já foi visto anteriormente.

Gráfico 5: Oferta de Energia Primária Renovável no Brasil.

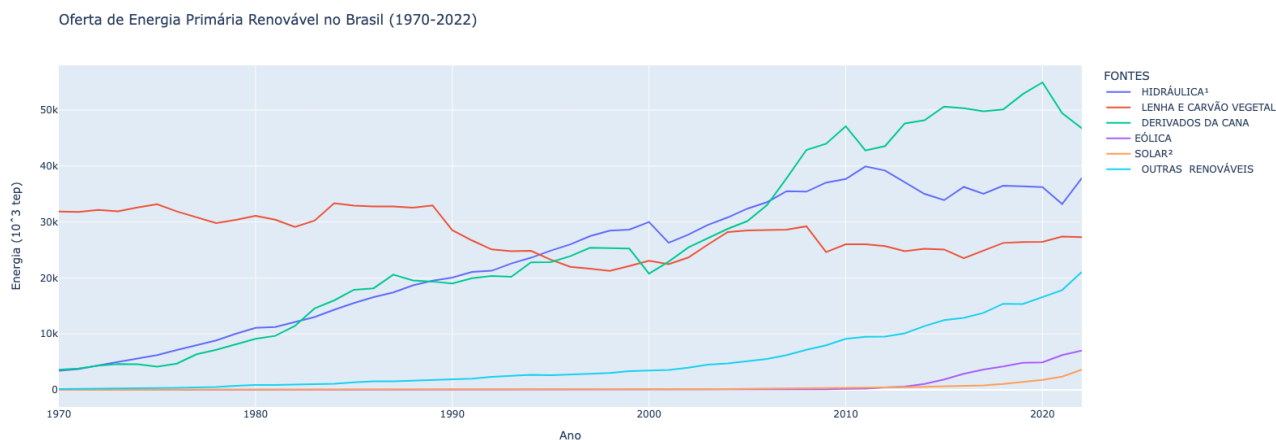
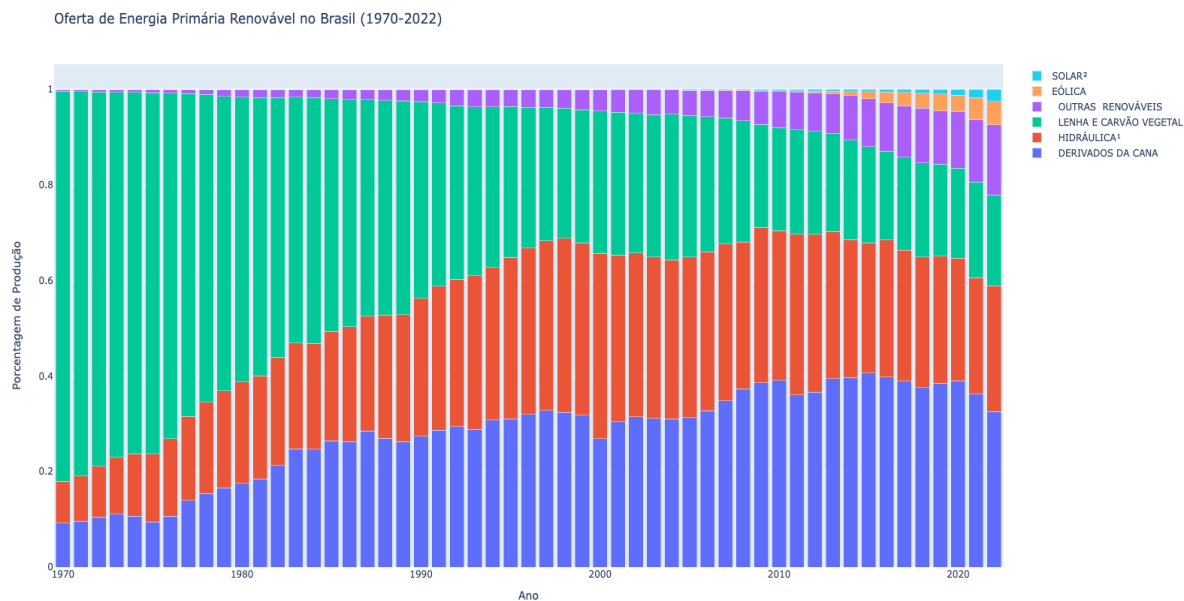


Gráfico 6: Oferta de Energia Primária Renovável no Brasil em Porcentagem.

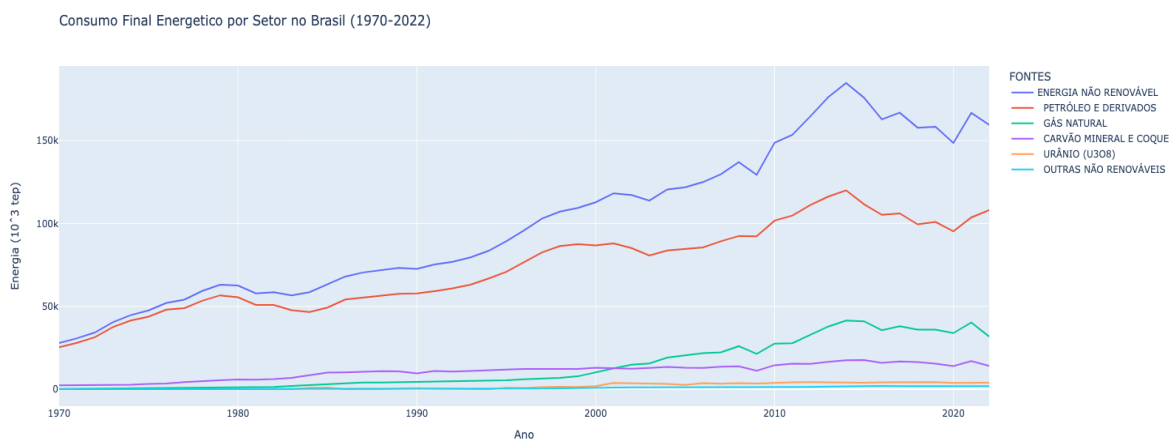


Ao analisar os gráficos de oferta é possível notar a semelhança entre os valores de oferta e produção de energia. É um testemunho da robustez e estabilidade do sistema energético. Esta consistência sugere uma harmonia entre a capacidade de produção e a demanda de energia, indicando uma gestão eficaz dos recursos energéticos. Ademais, isso pode refletir investimentos significativos em infraestrutura e tecnologias de produção de energia, bem como práticas sustentáveis

de gerenciamento de recursos. Essa sincronia entre oferta e produção não apenas assegura um suprimento confiável de energia, mas também sugere uma abordagem responsável e resiliente para atender às necessidades energéticas da sociedade.

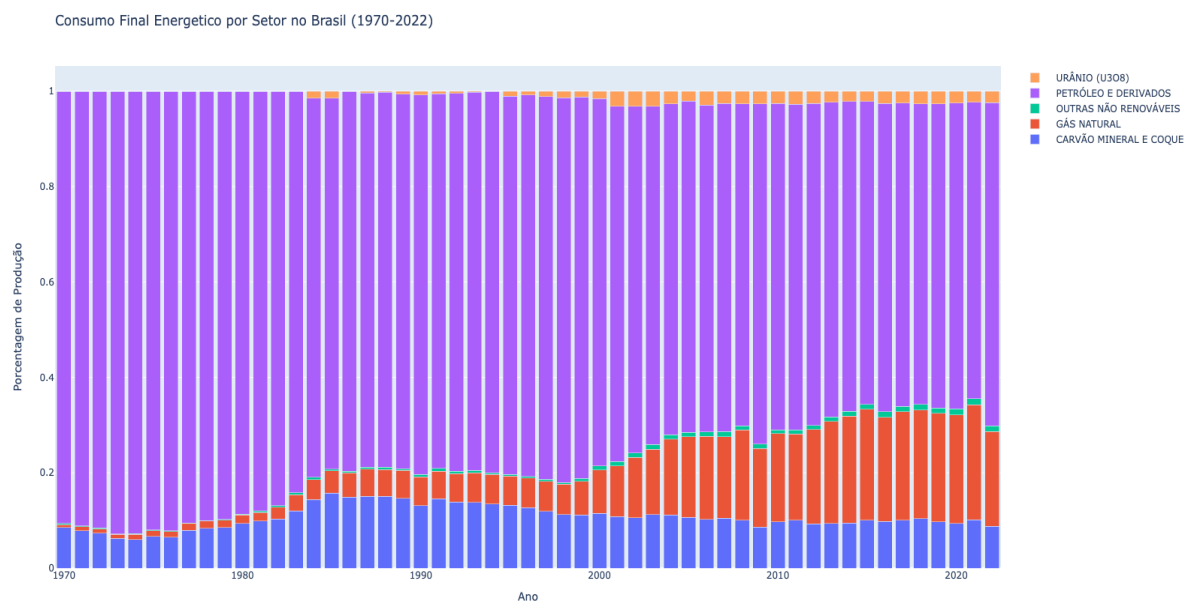
Realizando o mesmo procedimento para fontes de energia não renováveis, obtêm-se os Gráficos 7 e 8.

Gráfico 7: Oferta de Energia Primária Não Renovável no Brasil.



Fonte: Autoria Própria, 2024.

Gráfico 8: Oferta de Energia Primária Não Renovável no Brasil em Porcentagem.



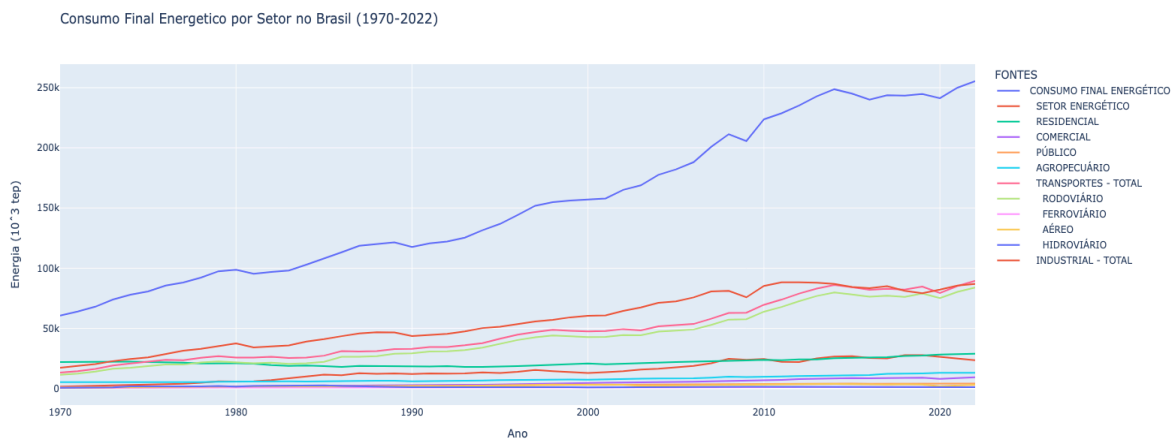
Fonte: Autoria Própria, 2024.

Diferente da situação da energia renovável, em que a disponibilidade muitas vezes está diretamente relacionada aos recursos naturais do país, a dinâmica da energia não renovável frequentemente envolve uma relação mais complexa entre

oferta interna e produção total. Essa complexidade é evidenciada pela significativa discrepância entre a oferta interna de energia e a produção total, uma disparidade que muitas vezes é explicada pela prática generalizada de importação de energia por muitos países.

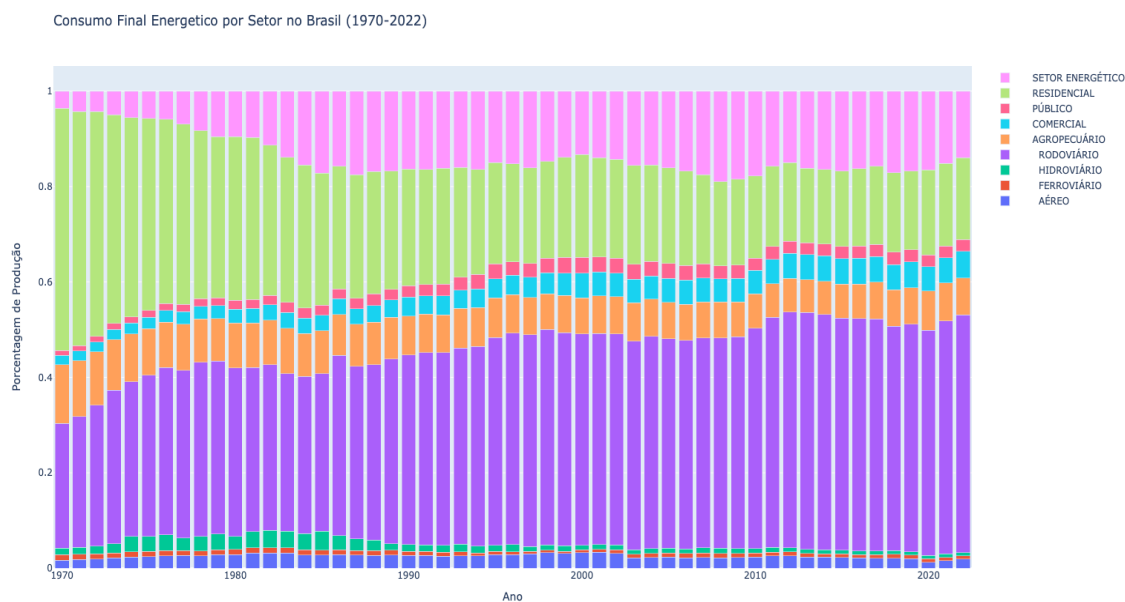
Agora com o entendimento de produção e oferta, nos resta compreender o panorama referente ao consumo final de energia por setor. Para isso será analisado o consumo final energético. Seguindo a mesma estratégia, plotando os Gráficos 9 e 10.

Gráfico 9: Consumo Final Energético por Setor no Brasil.



Fonte: Autoria Própria, 2024.

Gráfico 10: Consumo Final Energético por Setor no Brasil em Porcentagem.



Fonte: Autoria Própria, 2024.

Em relação ao Gráfico 9, tem-se a linha lilás, em que representa a totalidade do consumo final energético, mostrando o seu crescimento chegando a ultrapassar 250k 10³tep. Além disso, tem-se o Gráfico 10, que apresenta a participação de cada setor mostrando uma grande presença do setor rodoviário, setor secundário dentro do setor transporte, que em 2022 representou 49,83% do consumo.

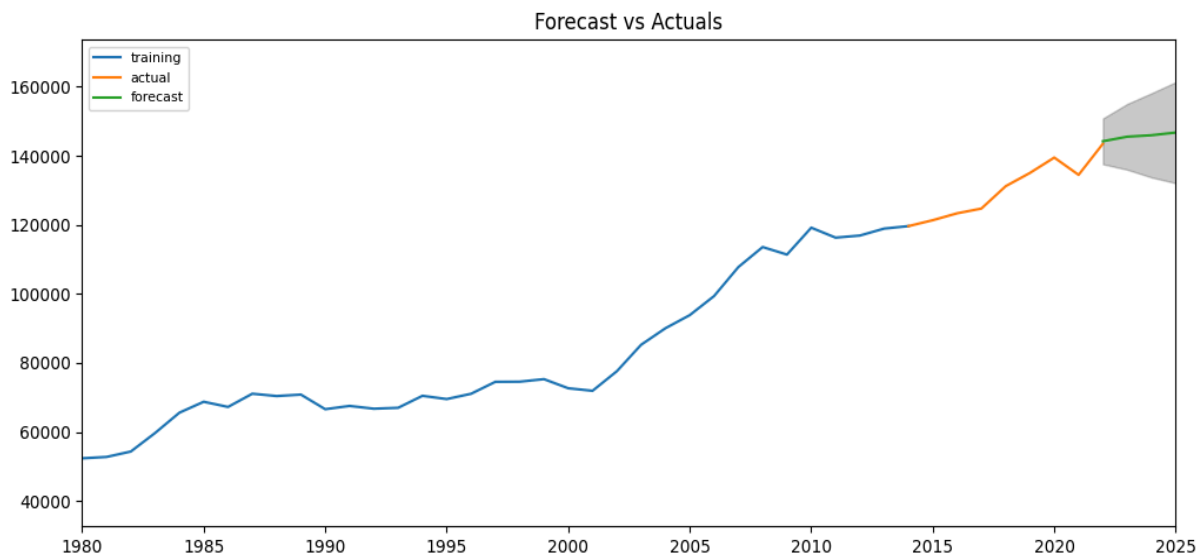
Uma vez compreendido como se comporta a produção, oferta e consumo energético no Brasil é possível avançar para o próximo objetivo, aplicar o modelo ARIMA a partir da função desenvolvida chamada de *plot_arima_forecast*.

Para todas as chamadas dessa função, foi realizada uma verificação detalhada dos dados para garantir que apenas as informações relevantes para a fonte de energia especificada fossem incluídas na análise. Primeiro, os dados foram filtrados com base no campo específico de interesse. Em seguida, foram realizadas etapas de limpeza de dados para garantir que apenas valores numéricos válidos fossem considerados, descartando quaisquer valores ausentes ou inválidos. Após a preparação dos dados, uma série de gráficos foi gerada para fornecer percepções visuais sobre a série temporal em questão. Isso incluiu a plotagem da série temporal original, a função de autocorrelação (ACF) e a função de autocorrelação parcial (PACF), que são cruciais para a identificação de padrões e estruturas nos dados.

A análise desses gráficos permitiu identificar a presença de sazonalidade, tendências e correlações significativas, orientando a escolha dos parâmetros do modelo ARIMA. Com base nessas análises preliminares, um modelo ARIMA foi ajustado aos dados para realizar previsões futuras. O modelo ajustado foi então utilizado para gerar previsões para um período específico, juntamente com intervalos de confiança para essas previsões. Além disso, foram examinados os resíduos do modelo e sua densidade para avaliar a adequação do modelo aos dados. Essas previsões foram visualizadas juntamente com os dados reais para avaliar a precisão do modelo e apresentar informações sobre as tendências futuras da série temporal.

Com isso em mente, é possível partir para a primeira previsão, referente a produção de energia renovável até o ano de 2025. Essa previsão está contida no Gráfico 11.

Gráfico 11: Previsão da Produção de Energia Renovável.

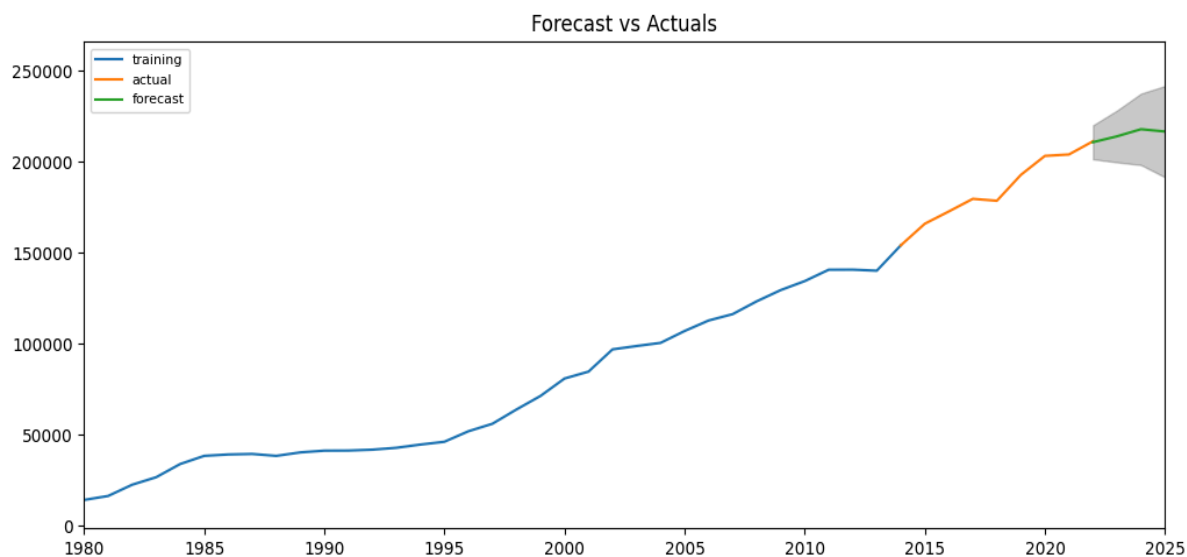


Fonte: Autoria Própria, 2024.

No Gráfico 11, destaca-se uma previsão estável representada pela linha verde, projetada para os anos seguintes. Adicionalmente, observa-se a representação de uma ampla faixa de erro, essencial para a análise cuidadosa dos resultados previstos.

Continuando com a previsão para os dados de produção, no Gráfico 12 é ilustrado a estimativa para os valores de produção energética para as fontes não renováveis.

Gráfico 12: Previsão da Produção de Energia Não Renovável.



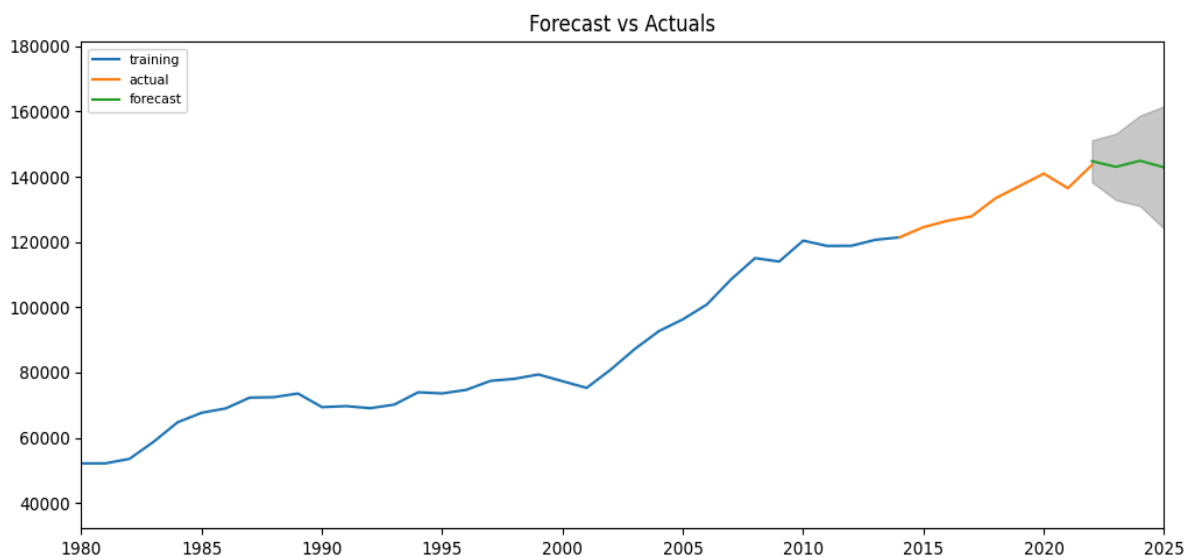
Fonte: Autoria Própria, 2024.

Ao analisar o Gráfico 12, é possível notar que os valores previstos exibem uma tendência geral de crescimento ao longo dos próximos anos, refletindo uma perspectiva positiva em relação à produção de energia. No entanto, é interessante observar uma leve queda no final do período previsto.

Essa diminuição sutil pode sugerir uma variação sazonal ou uma mudança nas condições que influenciam a produção de energia. Investigar essa queda pode ser crucial para entender melhor as dinâmicas do mercado energético e ajustar estratégias conforme necessário.

Seguindo a sequência da análise quantitativa de dados, agora será abordado a previsão dos dados referentes a oferta interna de energia, tanto renováveis no Gráfico 13, quanto não renováveis no Gráfico 13.

Gráfico 13: Previsão da Oferta Interna de Energia Renovável.

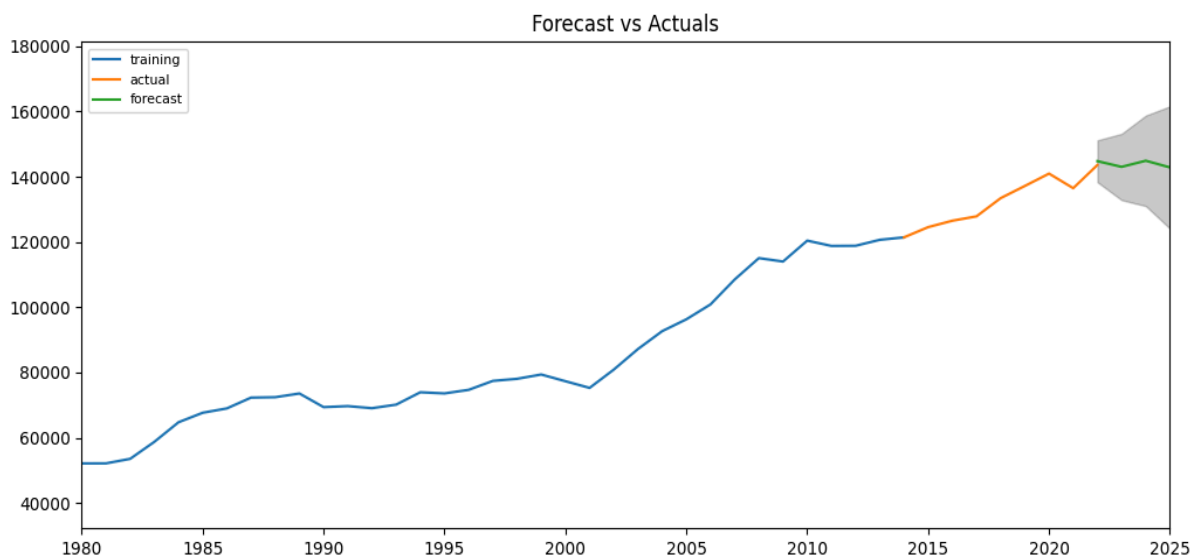


Fonte: Autoria Própria, 2024.

Ao considerar os dados de previsão para os próximos anos, conforme apresentado no Gráfico 13 sobre a oferta interna de energia de fontes renováveis, percebe-se uma tendência geral de estabilidade, seguida por uma ligeira variação ao longo do período previsto. Inicialmente, foi observado uma pequena queda em 2023, seguida por um aumento gradual até 2025.

Posteriormente, entre 2024 e 2025, os valores previstos permanecem relativamente consistentes, sugerindo uma certa estabilidade na oferta interna de energia proveniente de fontes renováveis durante esse período. Essa análise dos dados de previsão nos fornece uma visão valiosa sobre as tendências futuras da oferta interna de energia renovável, permitindo uma melhor compreensão do panorama energético e auxiliando na formulação de estratégias para o uso sustentável dos recursos energéticos.

Gráfico 14: Previsão da Oferta Interna de Energia Não Renovável.



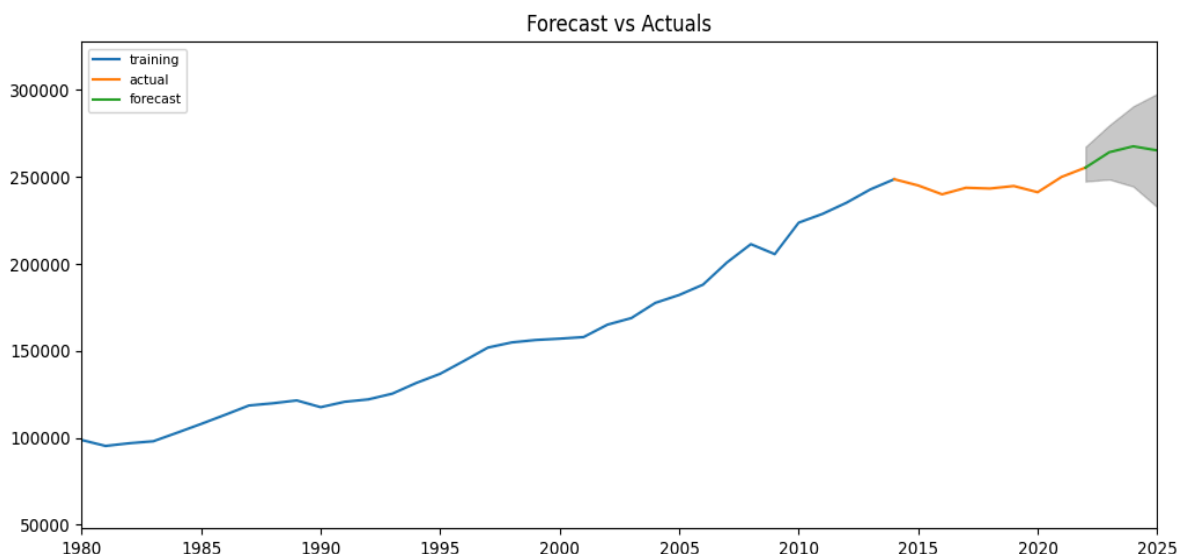
Fonte: Autoria Própria, 2024.

Ao analisar os dados de previsão para os próximos anos, conforme apresentado no Gráfico 14 sobre a oferta interna de energia de fontes não renováveis, percebe-se uma tendência de variação ao longo do período previsto. Observando assim, um aumento significativo em 2024, seguido por uma ligeira queda em 2025.

Essa variação sugere uma dinâmica flutuante na oferta interna de energia proveniente de fontes não renováveis durante o período em questão. Essa análise dos dados de previsão nos fornece informações valiosas sobre as tendências futuras da oferta interna de energia não renovável, permitindo uma melhor compreensão do cenário energético e facilitando o desenvolvimento de estratégias para o uso eficiente e sustentável dos recursos energéticos disponíveis.

Por fim, tem-se a previsão referente ao consumo final energético, em que o Gráfico 15 mostra uma representação visual das projeções futuras da base.

Gráfico 15: Previsão do Consumo Final Energético.



Fonte: Autoria Própria, 2024.

Ao analisar os dados de previsão para o consumo final energético nos próximos anos, observa-se uma tendência que sugere uma onda ao longo do período previsto. Inicialmente, há um aumento em 2023, seguido por um pico em 2024 e uma leve queda em 2025. Essa oscilação pode refletir variações sazonais ou mudanças nas condições econômicas e ambientais que influenciam o consumo de energia.

Essa observação dos dados projetados nos oferece uma visão detalhada das flutuações esperadas no consumo de energia nos próximos anos. A identificação desses padrões de ondas pode desencadear reflexões sobre os fatores subjacentes que impulsionam essas variações, sejam elas sazonais, econômicas ou tecnológicas. Essa compreensão mais profunda é fundamental para direcionar iniciativas e políticas que visam otimizar o uso de recursos energéticos, além de impulsionar a transição para fontes mais sustentáveis e renováveis de energia.

5. Considerações Finais

Este estudo teve como objetivo explorar as técnicas de análise e ciência de dados para compreender o cenário do BEN, utilizando um algoritmo baseado nos modelos ARIMA, implementado em Python. Para atingir esse propósito, foram definidos dois objetivos finais.

Primeiramente, foi abordada a manipulação de banco de dados, que se constatou que em 2022 houve um aumento considerável na produção de energia de fontes renováveis, porém, ainda abaixo de fontes não renováveis, sendo necessário um aumento de 47.13% da produção energética de fontes renováveis para poder ultrapassar as matrizes não renováveis. Visualizou-se, por meio de gráficos gerados com as bibliotecas Matplotlib e Seaborn, que a produção de energia renovável está acompanhando o crescimento da oferta interna de energia, porém foi possível notar que em casos de fontes não renováveis a oferta de energia está bastante superior à produção, indicando que provavelmente pela prática generalizada de importação de energia por muitos países impacta esse resultado.

O segundo objetivo envolveu a aplicação de operações matemáticas e estatísticas em matrizes computacionais para desenvolver um modelo de previsão de casos. Utilizando os modelos ARIMA, foram implementados conceitos de autorregressão e média móvel para obter curvas de previsão. Concluiu-se que o modelo de previsão proporcionou um aumento de 2.64% entre o ano 2022 e 2023 para o consumo final energético, além de indicar um possível aumento de 1.76% da oferta de energia para matrizes não renováveis e um decréscimo de 0.3957% para fontes renováveis. Por fim, foi visto na base de produção energética um aumento de 1.294% para matrizes não renováveis, um crescimento de 1.429%.

Os resultados obtidos na análise do BEN reforçam a suposição de que a dinâmica energética contemporânea está fortemente ligada à necessidade de transição para fontes renováveis. As tendências quantitativas e qualitativas extraídas do BEN corroboram essa conclusão, revelando uma crescente demanda por fontes de energia não renováveis e uma estagnação, ou até mesmo um declínio, na produção e consumo de energia proveniente de fontes renováveis. Essa disparidade

destaca os desafios enfrentados na busca por um sistema energético mais sustentável e ambientalmente responsável.

A análise do BEN não apenas evidencia a urgência de ações concretas para promover a adoção de energias renováveis, mas também destaca a necessidade de políticas públicas e estratégias de investimento que visem mitigar os impactos negativos das fontes de energia não renováveis, especialmente em relação à segurança energética e às emissões de gases de efeito estufa.

Apesar das análises realizadas, o estudo identificou que o banco de dados apresenta algumas limitações, como a falta de organização e a falta de dados mais frequentes, como mensais ou semestrais. Recomenda-se, portanto, em futuras pesquisas, buscar por materiais mais atualizados e estruturados, além de explorar a aplicação de novos modelos estatísticos para aprimorar a compreensão e a análise desses dados em situações específicas.

REFERÊNCIAS

AMARAL, Fernando. **Introdução à Ciência de Dados: Mineração de Dados e Big Data**. Alta Books Editora, 2016.

BORGES, Luiz Eduardo. **Python para Desenvolvedores: Aborda Python 3.3**. Novatec Editora, 2014.

CAMILO, Cássio Oliveira; Silva, João Carlos da. **Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas**. Universidade Federal de Goiás (UFG), v. 1, n. 1, p. 1-29, 2009.

CASTANHA, Renata Cristina Gutierrez. **A Ciência de Dados e a Cientista de Dados**. AtoZ: Novas Práticas em Informação e Conhecimento, v. 10, n. 2, p. 1-4, 2021.

CRUZ, Leandro César. **Data Science: Desenvolvimento de Aplicação para Análise de Dados**. Assis/SP, 2018.

DATE, Christopher J. **Introdução a Sistemas de Bancos de Dados**. Elsevier Brasil, 2004.

Define Analytics: The changing role of BI's favorite catch-all term. Disponível em: <https://www.tableau.com/learn/whitepapers/define-analytics#form>. Acesso em: 2 abril. 2024.

DOS PASSOS, Danielle Sandler. Big Data, Data Science e seus Contributos para o Avanço no Uso da Open Source Intelligence. **Sistemas & Gestão**, v. 11, n. 4, p. 392-396, 2016.

DUARTE, A. **Desvendando a Regressão Linear**. Disponível em: <https://www.alura.com.br/artigos/desvendando-a-regressao-linear>. Acesso em: 4 abril. 2024.

ELMASRI, Ramez; Navathe, Shamkant. **Sistemas de Banco de Dados**. 6. ed. São Paulo, 2005.

Empresas de Pesquisa Energética (EPE). **Balanco Energético Nacional 2022**. Disponível em: <https://www.epe.gov.br/pt/publicacoes-dados-abertos/publicacoes/balanco-energetico-nacional-2022>. Acesso em: 8 de dezembro de 2023.

Formigoni, Philipe de Araújo Fernandes. **Python na Análise de Dados: Estudo de Caso com Dados de Acidentes Aéreos no Brasil**. 2021.

GOLDEMBERG, José; Moreira, José Roberto. **Política Energética no Brasil**. Estudos Avançados, São Paulo, v. 19, n. 55, p. 215, 2005.

GOMES, João Paulo Pombeiro; Vieira, Marcelo Milano Falcão. **O Campo da Energia Elétrica no Brasil de 1880 a 2002**. Revista de Administração Pública (RAP), Rio de Janeiro, v. 43, n. 2, p. 295-321, mar./abr. 2009.

IPHYTHON. **Site Oficial do Pacote IPython**. Disponível em <https://ipython.org/> Acesso em: 4 de abril de 2024.

JÚNIOR, R. S. **Estudo de Dados de Negócio com Data Science**. Revista Gestão & Tecnologia, v. 1, n. 28, p. 147–163, 1 jul. 2019.

Lóscio, Bernadette Farias; Oliveira, HR de; Pontes, JC de S. NoSQL no Desenvolvimento de Aplicações Web Colaborativas. VIII Simpósio Brasileiro de Sistemas Colaborativos, v. 10, n. 1, p. 11, 2011.

LUCENA, Matheus Dantas de et al. **Análise Histórica das Matrizes Energéticas do Brasil e Suas Projeções**. 2017.

MANZANO, José Augusto NG. Introdução à Linguagem Python. Novatec Editora, 2018.

UFV. **Mas afinal, o que é Média Móvel?**. Disponível em <https://indicesdee.ufv.br/ict-ipt/mas-afinal-o-que-e-media-movel/> Acesso em: 4 de abril de 2024.

MATPLOTLIB. **Site Oficial do Pacote Matplotlib**. Disponível em <https://matplotlib.org/gallery/api/logos2.html> Acesso em: 5 de abril de 2024.

NOTEBOOK; Jupyter. **Site Oficial da IDE de Jupyter Notebook**. Disponível em <http://jupyter.org/> Acesso em: 5 de abril de 2024.

NUMPY. **Site Oficial do Pacote NumPy**. Disponível em <http://www.numpy.org/> Acesso em: 5 de abril de 2024.

PANDAS. **Site Oficial do Pacote Pandas**. Disponível em <https://pandas.pydata.org/> Acesso em: 5 de abril de 2024.

PIMENTEL, João Felipe et al. **Ciência de Dados com Reprodutibilidade Usando Jupyter**. Sociedade Brasileira de Computação, 2021.

PLOTLY-PYTHON. **Site Oficial do Pacote Plotly.** Disponível em <https://plotly.com/python/> Acesso em: 5 de abril de 2024.

PYTHON. **Site Oficial da Linguagem de Programação Python.** Disponível em <https://www.python.org/community/logos/> Acesso em: 5 de abril de 2024.

ROB, Peter; Coronel, Carlos. **Sistemas de Banco de Dados.** Projeto, Implementação e, 2011.

SANTOS, David Moises B. et al. **Integrando as Disciplinas de Engenharia de Software, Análise e Projeto de Sistemas e Banco de Dados Utilizando PBL.** In: XV Workshop sobre Educação em Computação—Anais do XXVII Congresso da Sociedade Brasileira de Computação. 2007. p. 66-75.

SCIKIT-LEARN. **Site Oficial do Pacote Scikit-Learn.** Disponível em <http://scikitlearn.org/stable/> Acesso em: 22 de abril de 2024.

SILVEIRA, G. **Machine Learning: Classificação por Trás dos Panos.** Disponível em: <https://cursos.alura.com.br/course/machine-learning-validando-modelos>. Acesso em: 10 abril. 2024.

SOUZA, Vanessa Faria de. **Mineração de Dados Educacionais Com Aprendizagem de Máquina.** Revista Educar Mais, 2021, pp. 766 à 787.

STATSMODELS. **Site Oficial do Pacote Statsmodels.** Disponível em <https://www.statsmodels.org/stable/index.html/> Acesso em: 2 de abril de 2024.