UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Instituto de Ciência e Tecnologia
Campus Sorocaba - ICTS/Unesp

Augusto Andreotte Lorençatto

ANÁLISE DO EFEITO DAS CONDIÇÕES CLIMÁTICAS NA PREVISÃO DE CURTO PRAZO DA DEMANDA ENERGÉTICA UTILIZANDO O MÉTODO XGBOOST

Sorocaba/SP
2022

AUGUSTO ANDREOTTE LORENÇATTO

ANÁLISE DO EFEITO DAS CONDIÇÕES CLIMÁTICAS NA PREVISÃO DE CURTO PRAZO DA DEMANDA ENERGÉTICA UTILIZANDO O MÉTODO XGBOOST

Trabalho de Conclusão de Curso de Graduação apresentado como parte dos pré-requisitos para a obtenção do título de Engenheiro Ambiental, à Universidade Estadual Paulista "Júlio de Mesquita Filho"

ORIENTADOR: Prof. Dr. Antonio Cesar Germano Martins

Sorocaba/SP

2022

Lorencatto, Augusto
L868a Análise do efeito das condições climáticas na previsão de curto prazo da demanda energética utilizando o método xgboost / Augusto Lorencatto. -- Sorocaba, 2023
45 p.

Trabalho de conclusão de curso (Bacharelado - Engenharia Ambiental) - Universidade Estadual Paulista (Unesp), Instituto de Ciência e Tecnologia, Sorocaba
Orientador: Antonio Cesar Martins

1. Modelos de Aprendizado de máquina. 2. Dados climáticos. 3. Demanda energética. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Ciência e Tecnologia, Sorocaba. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

AUGUSTO ANDREOTTE LORENÇATTO

ANÁLISE DO EFEITO DAS CONDIÇÕES CLIMÁTICAS NA PREVISÃO DE CURTO PRAZO DA DEMANDA ENERGÉTICA UTILIZANDO O MÉTODO XGBOOST

Trabalho de Conclusão de Curso de Graduação apresentado como parte dos pré-requisitos para a obtenção do título de Engenheiro Ambiental, à Universidade Estadual Paulista "Júlio de Mesquita Filho"

Sorocaba, 3 de dezembro de 2022.

_____

Prof. Dr. Antonio Cesar Germano Martins

Sorocaba/SP

2022

RESUMO

Diante de um cenário onde a demanda energética se encontra em contínua ascensão e a distribuição dos métodos de geração ainda não se afirma sustentável, estudos que visam a entrega de uma previsibilidade são de extrema importância para o meio ambiente e para o setor elétrico brasileiro. Como ferramenta para essa situação, modelos matemáticos computacionais vem ganhando bastante espaço, devido ao seu sucesso na aplicação em diversas áreas do conhecimento, como o setor financeiro e de fraude, por exemplo. Dessa forma, o presente trabalho tem como objetivo entender o comportamento da curva da carga horária no estado de São Paulo, correlacionando principalmente com atributos climáticos de diversas estações meteorológicas automáticas no polígono que contorna o estado, além de informações derivadas do próprio atributo temporal. No primeiro momento, tem-se como premissa o desenvolvimento de um ciclo completo de um projeto envolvendo machine learning, desde a extração dos principais dados até a construção final do modelo preditivo, passando por etapas de maturação da informação. Visando a real aplicabilidade do modelo, o trabalho buscou dados já consolidados tanto dos atributos preditores como do atributo alvo, consultando o Instituto Nacional de Meteorologia (INMET) para dados climáticos e o Operador Nacional do Sistema Elétrico (ONS) para os dados da demanda energética. O modelo desenvolvido foi embasado em técnicas de ensemble, especificamente o *XGboost*, no qual o repositório computacional desenvolvido como modelo para a aplicação em qualquer região do país, além de servir de insumo para o aprofundamento da problemática abordada.

Palavras-Chave: Modelos de Aprendizado de máquina, Dados climáticos, Demanda energética

ABSTRACT

Faced with a scenario where energy demand is continuously rising and the distribution of generation methods is not yet sustainable, studies that aim to deliver predictability are extremely important for the environment and for the Brazilian electricity sector. As a tool for this situation, computational mathematical models have had a strong attraction, due to their successful application in several areas of knowledge, such as the financial and fraud sector, for example. Thus, the present work aims to understand the behavior of the hourly load curve in the state of São Paulo, mainly correlating with climatic attributes of several automatic weather stations in the surroundings of the polygon that surrounds the state, in addition to information derived from the temporal attribute itself. At first, the premise is the development of a complete cycle of a project involving machine learning, from the extraction of the main data to the final construction of the predictive model, going through stages of information maturation. Aiming at the real applicability of the model, the work sought already consolidated data of both the predictor and the target attributes, consulting the National Institute of Meteorology (INMET) for climate data and the National Electric System Operator (ONS) for energy demand data. . The developed model was based on ensemble techniques, specifically XGboost, in which the computational repository developed can be used in any region of the country, in addition to serving as an input for the deepening the addressed problem.

Keywords: Machine Learning Models, Weather Data, Energy Demand

Sumário

32

OBJETIVO

O presente trabalho tem como objetivo principal o desenvolvimento de um sistema para apoio à tomada de decisão na estimativa do consumo de energia elétrica da região Sudeste do território Brasileiro, englobando técnicas de mineração, exploração e tratamento de dados climáticos que servirão de atributos.

OBJETIVOS ESPECÍFICOS

1. Selecionar estações meteorológicas para a coleta de dados climáticos no estado de São Paulo;
2. Realizar a junção das bases de dados climáticas e da demanda energética, selecionando os melhores atributos;
3. Realizar a exploração dos dados, entendendo seu comportamento, tendências e sazonalidade;
4. Criar um modelo para a previsão da curva da carga horária utilizando os dados climáticos das estações meteorológicas, além das próprias características temporais.

1. INTRODUÇÃO

1.1 Motivação e entendimento da área

O setor de energia elétrica é essencial para a economia de um país pois, além do uso em todas as áreas, também está associado à qualidade de vida da população.

Para que seja feito um planejamento de instalação de fontes e organização da distribuição, é importante que sejam realizadas previsões de demanda para se balancear a produção e o consumo.

No Brasil, o setor de energia elétrica teve sua reestruturação nas últimas duas décadas, permitindo a livre negociação da compra e venda de energia, o que habilitou uma forma independente e autônoma de crescimento vertical nas atividades de geração, transmissão e distribuição de energia elétrica.

No contexto do planejamento da operação de curto prazo de sistemas elétricos de potência, a previsão do consumo de energia é fundamental para a montagem do programa de operação dos dias seguintes, sendo que erros resultantes desse processo podem acarretar em uma instabilidade do sistema e um possível não atendimento da demanda. Para isso, os setores de planejamento de concessionárias entenderam que há valor em recuperar as informações da operação no passado e analisá-las para então se obter conhecimento ou padrões necessários para tentar prever comportamentos dinâmicos no futuro, englobando assim uma problemática de séries temporais.

Em termos gerais, pode-se dizer que as séries temporais contemplam informações que foram baseadas em um conjunto de observações ou medidas coletadas em um determinado intervalo de tempo, e podem ter seus valores dependentes um dos outros, permitindo assim uma base que pode ser compreendida no estudo de métodos de análise e/ou previsão, sendo que a precisão da previsão contemplada nesse estudo é de grande importância.

Além do benefício econômico associado, com as análises preditivas é possível entender a crescente demanda energética da população contribuindo para um planejamento mais eficiente da operação da distribuição da energia, de forma mais confiável e principalmente, econômica.

2. REVISÃO DA LITERATURA

2.1 Séries temporais

Uma série temporal é caracterizada por qualquer conjunto de observações ordenadas no tempo, que podem ser subdivididas em discretas e contínuas. De uma forma geral, uma série temporal discreta é consolidada através da amostragem de uma série temporal contínua em intervalos de tempos iguais .

A análise de séries temporais compreende duas abordagens com o objetivo de construir modelos matemáticos de acordo com características pré-estabelecidas, podendo ser realizadas no domínio temporal ou no domínio de frequências.

No domínio de frequências, a análise espectral consiste em obter e analisar a distribuição de frequências presentes no sinal (MORETTIN, 2004).

As séries temporais podem ser classificadas como univariadas ou multivariadas, dependendo do número de variáveis dentro do sistema. As univariadas são obtidas a partir da amostragem de um único padrão de observação, por exemplo, os valores de um sinal dependente do tempo. Já as séries temporais multivariadas, contempladas no presente estudo, são geradas a partir da observação simultânea de duas ou mais variáveis. (PALIT, et al., 2005)

Além disso, uma série temporal possui características que impactam na escolha do método e forma de se analisar o problema estudado. A seguir, estão listados algumas dessas características:

- **Tendência**

A tendência de uma série temporal é definida com base em seu crescimento/decrescimento em um determinado período. Isto é, para esse período de tempo estabelecido, a tendência indica se a série cresce, decresce ou permanece estável (ASSUNÇÃO,2020).

Esse componente se manifesta globalmente ou localmente, através do aumento ou diminuição de valores dos dados, como uma consequência da superposição de valores de série verdadeira e uma perturbação (PALIT, et al., 2005).

Para o estudo da curva de carga de energia, esse componente é extremamente importante a médio e longo prazo, possibilitando o entendimento do comportamento de consumo como um todo, ou até mesmo servindo como ferramenta para validação, considerando a análise de impacto de determinada medida para a melhor distribuição das

fontes geradoras, por exemplo.

- **Autocorrelação**

Em séries temporais, a autocorrelação quando explorada corretamente traz muitos ganhos às análises envolvidas e a capacidade preditiva no modelo, pois está relacionada a associação entre os valores de uma mesma variável, na maioria das vezes ordenada pelo tempo.

Dessa forma, torna-se necessário o conhecimento do comportamento dos atributos entre as observações atuais e as anteriores, os quais podem ser avaliados através de funções de autocorrelação.

- **Sazonalidade**

Está associada a identificação de padrões na série temporal, ou seja, oscilações de subida ou descida que ocorrem em períodos específicos.(ASSUNÇÃO,2020).

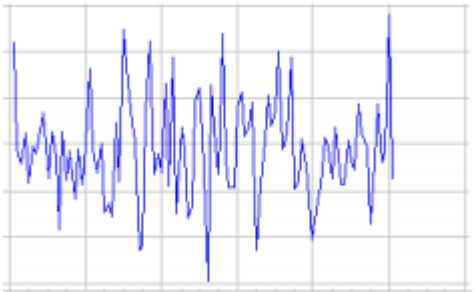
Tratando de dados climatológicos, esse componente é muito presente principalmente em períodos durante o ano, tais como inverno e verão.

- **Estacionaridade**

Uma série temporal é estacionária quando suas características estatísticas (média, variância, autocorrelação, …) são constantes ao longo do tempo, refletindo em uma série que se desenvolve em torno de uma média constante, refletindo alguma forma de estabilidade e equilíbrio estatístico (MADALLA e LAHIRI, 2009).

Para identificação de tal componente, basta verificar se no comportamento da série temporal há uma média constante ou uma variância constante. A figura 1 ilustra o comportamento descrito anteriormente.

Figura 1 : Comportamento estacionário série temporal



Fonte : (MADALLA e LAHIRI, 2009)

2.2 Modelos estatísticos aplicados em séries temporais

2.2.1 Xgboost

Dentro do universo das técnicas de *machine learning* , o *Ensemble Learning* se destaca positivamente e consiste em combinar modelos, entendidos como 'weak learners', que são treinados em conjunto, com a premissa de tentar reduzir o viés e/ou a variância, de modo a resolver o mesmo problema, apresentando melhores resultados (ROCCA,2019). Dessa forma, esta técnica parte do princípio de que juntos os modelos mais fracos produzem um modelo mais robusto, denominado de *strong learners*.

Para combinar esses múltiplos modelos base, a metodologia que mais se destaca é o impulsionamento (*boosting*), que considera a homogeneidade dos modelos base e desenvolve o processo de combinação de forma sequencial, se mostrando extremamente adaptativo, considerando os resultados anteriores.

Assim, esse processo ocorre da seguinte forma :

1. Desenvolvimento de um modelo base weak learner;
2. Agregação desse modelo base no conjunto (Ensemble);
3. Atualização dos conjuntos de dados de treinamento, identificando os pontos fortes e fracos no modelo atual para servir de base para o próximo modelo base.

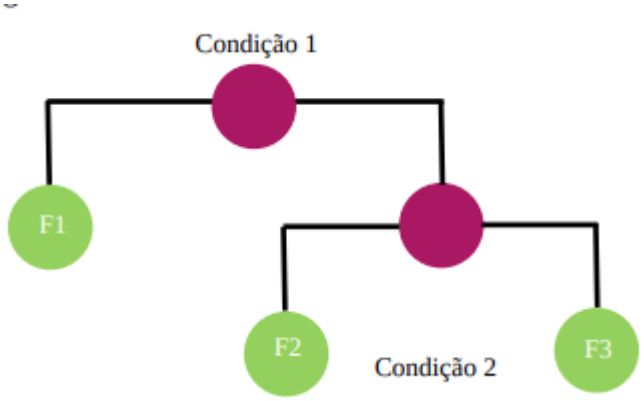
Vale acrescentar que a técnica de boosting fortalece a situação em seu momento atual onde os modelos anteriores tiveram dificuldade no treinamento, tornando a solução mais resistente ao enviesamento.

Entendido os conceitos base, o XGBoost é o nome do algoritmo que implementa um modelo de Ensemble learning, utilizado tanto para regressão quanto classificação, baseado em múltiplas árvores de decisão agregadas pela técnica de boosting.

A árvore de decisão é um modelo voltado para classificação ou regressão, baseado na execução de sucessivas partições binárias de uma amostra, buscando a construção de subamostras internamente mais homogêneas. Cada subamostra particionada recebe o nome de nó e cada resultado final identificado recebe o nome de folha. A Figura 2 ilustra uma árvore para prever uma nova observação. Começa-se do primeiro nó, verificando se a condição nele

imposta é ou não atendida. Em caso afirmativo, prossegue-se à esquerda e em caso negativo, à direita. Tal procedimento é repetido, até que se atinja uma folha, última instância de classificação.

Figura 2 :Estrutura de uma árvore de decisão



Fonte : IZBICKI e SANTOS, 2000.

A construção de uma árvore de decisão é feita a partir da criação dos ramos de acordo com uma métrica de homogeneidade dos dados e posteriormente, a exclusão de alguns ramos com um processo de poda, que visa evitar seu sobreajuste ao conjunto de dados utilizado. O XGBoost utiliza a abordagem de gradiente descendente para aprimorar, por meio do boosting, uma determinada árvore inicialmente construída (MEDEIROS,2021)

2.2.2 Divisão de dados para construção de modelos em séries temporais

Em análises e construção de modelos preditivos baseados em séries temporais, observações costumam não ser independentes, e por isso não se deve simplesmente randomizar os dados, como em construções de modelos de classificação, por exemplo. Ao invés disso, usualmente são divididos conjuntos de observações ao longo de uma sequência.

As características de uma série temporal, como sua natureza autorregressiva, apresentação de tendências, sazonalidades, entre outros, não permite que um embaralhamento aleatório dos dados para seu treinamento seja válido. Como um simples exemplo, se as observações apresentam autocorrelação, uma análise em t no conjunto de dados de teste e outra análise em $t+1$ poderia causar um problema, não representando assim a real natureza do conjunto de dados. Um modelo que entende bem a formação da série temporal de maneira natural apresenta uma capacidade preditiva muito maior (MIYAKI, 2019).

Uma abordagem mais segura para esse tipo de problema é inicialmente dividir os dados de treinamento em múltiplos segmentos, e implementar as seguintes etapas :

1. Treinamento do primeiro segmento com um conjunto de hiperparâmetros;
2. Utilização do segundo segmento para testar os dados;
3. Repetir o processo para todos os segmentos.

Dessa forma, é realizado o treinamento e validação k-1 vezes, sendo k o número de segmentos criados inicialmente. Para a implementação de tal técnica, a biblioteca Scikit-learn traz uma função específica para validação de séries temporais, denominada *TimeSeriesSplit* (MIYAKI,2019).

2.3 Principais trabalhos relacionados a temática

Tendo em vista as técnicas computacionais descritas anteriormente, a aplicação dessas metodologias trazem soluções extremamente poderosas em diversas áreas do conhecimento devido ao grande volume de dados presentes nos negócios e em modelos que representam os fenômenos da natureza, o que leva a uma abordagem preditiva muito mais assertiva.

Para definição de um modelo de previsão de carga elétrica, foi necessário realizar pesquisas, identificando as mais tradicionais metodologias para resolução de tal problema.

Dentre as técnicas, destaca-se as baseadas em redes neurais MLP (Multilayer Perceptron), junto com o algoritmo de Backpropagation, que foi aplicada no trabalho de Miranda (2021), no qual se tinha como principal objetivo a análise da demanda de energia elétrica no horizonte de curto prazo em função da temperatura, utilizando redes neurais artificiais (RNA). Para isso, levantou-se as características mais importantes do comportamento não linear da curva de carga e posteriormente os fatores que influenciam esse comportamento. Em seguida, foram treinadas as RNAs através da arquitetura MLP com resultados muito interessantes para a previsão da curva de carga nos 31 dias posteriores a um mês de referência, comprovando que a técnica escolhida se mostra bastante eficiente para a temática envolvida e pode expandir esta área de atuação, sendo moldável para diversas áreas do conhecimento.

Também utilizando RNAs com MLP, SILVEIRA (2022) traz uma abordagem complementar utilizando lógica fuzzy, através do ANFIS (*Adaptive Network-based Fuzzy Inference System*), explorando a previsão da carga elétrica diária. Além disso, o trabalho visava realizar a combinação dos resultados dos modelos propostos para verificar o fato de

que há uma tendência de se obter um resultado melhor do que aquele obtido individualmente pelo melhor dos previsores. Como resultado dessa proposta, o autor concluiu que não houve ganhos significativos na combinação dos modelos desenvolvidos e que individualmente, estes produziram resultados satisfatórios.

PONTES (2022) em seu trabalho de pós-graduação explora a modelagem para a realização da previsão da produção industrial do Brasil e em paralelo o consumo de energia elétrica residencial, analisando a elasticidade de preço e renda, acrescentando um viés econômico a sua dissertação. O autor considera que a combinação dos modelos utilizados apresentou uma melhor capacidade preditiva e um desempenho bastante superior aos modelos quando aplicados individualmente, e que o estudo pode contribuir para a previsão do consumo de energia elétrica residencial através de seu modelo econométrico.

Em BERGAMO (2022), o ponto focal é a previsão da geração de energia elétrica dentro de um parque fotovoltaico, utilizando dados de geração histórica, ou seja, da própria série temporal da região, somado com atributos meteorológicos. O autor desenvolve as atividades de pré-processamento e exploração envolvida no processo, até de fato chegar na construção dos modelos preditivos, onde utiliza várias opções e dentre elas o *XGboost e* o modelo de Máquina de Vetores de Suporte (SVM). Posteriormente é feita a comparação entre o desempenho dos modelos, sendo constatado que para o problema abordado, o SVM apresentou a maior robustez em sua capacidade preditiva.

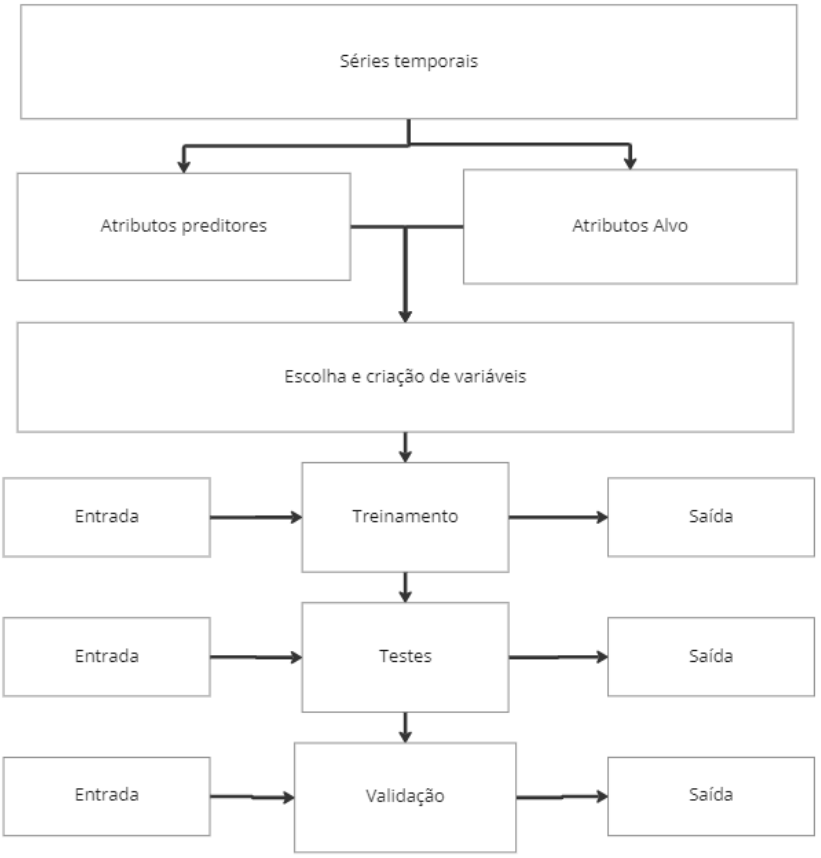
3. MATERIAIS E MÉTODOS

No presente capítulo, serão abordados assuntos que envolvem a metodologia empregada, descrição dos repositórios de dados utilizados, ambiente e ferramentas para o desenvolvimento do projeto.

3.1 Metodologia utilizada no desenvolvimento

Tendo em vista que 0 problema abordado no presente trabalho trata-se de uma série temporal, é necessário garantir que os dados estejam prontos para serem submetidos ao treinamento do modelo, sendo essencial etapas preparatórias que garantam que as informações utilizadas não possibilitem que o treinamento seja enviesado. A Figura 3 mostra a arquitetura básica de um projeto de regressão que envolve séries temporais.

Figura 3 : Arquitetura básica no sistema de séries temporais



Fonte : Autoria própria

Dessa forma, para o desenvolvimento do presente trabalho foi planejado a execução das seguintes etapas :

1) Coleta de dados em repositórios públicos das informações climáticas e da curva da carga horária;
2) Limpeza, organização e junção das bases com a finalidade de normalizar as informações, evitando dados redundantes que podem apresentar um risco para o desempenho do modelo;
3) Engenharia de variáveis, explorando o atributo temporal da base;
4) Análise exploratória, visando entender o comportamento das variáveis do problema;
5) Divisão dos dados e treinamento do modelo;
6) Análises e interpretação dos resultados do modelo.

3.2 Extração, repositório de dados e ambiente para desenvolvimento

A exploração e posterior construção do modelo preditivo foi realizada baseada em duas principais fontes de dados, descritas a seguir.

3.2.1 Curva da carga horária

Contempla os dados do atributo alvo, sendo relativos a curva da carga horária, representando o perfil de consumo de energia elétrica com a discretização horária, coletados diretamente no site do Operador Nacional do Sistema Elétrico do Brasil, através do Portal de Dados Abertos, que foi desenvolvido com o objetivo de facilitar, melhorar e democratizar o acesso de dados históricos da operação do Sistema Interligado Nacional - SIN.

A Figura 4 mostra a interface do sistema, ressaltando a acessibilidade dessas informações.

Figura 4 : Repositório de dados ONS

Fonte : ONS, 2022

Estes dados possuem características que permitem a identificação do subsistema da coleta, o instante da medição e o valor em MegaWatts. A Tabela 1 apresenta uma amostra desses dados,

Tabela 1 : Amostra de dados da Curva da Carga Horária

id_subsistema	nom_subsistema	din_instante	val_cargaenergiahomwmed
N	NORTE	2013-01-01 00:00:00	3799.16
NE	NORDESTE	2013-01-01 00:00:00	8.426.396
S	SUL	2013-01-01 00:00:00	7.263.287
SE	SUDESTE	2013-01-01 00:00:00	29.751.172
N	NORTE	2013-01-01 01:00:00	3720.14
NE	NORDESTE	2013-01-01 01:00:00	807.438.899.999
S	SUL	2013-01-01 01:00:00	7.107.271
SE	SUDESTE	2013-01-01 01:00:00	29.334.006
N	NORTE	2013-01-01 02:00:00	3700.76

Fonte : Autoria própria

Dessa forma, para o presente trabalho foi utilizada a série história da curva de carga horária de 01/01/2012 00:00:00 até 10/03/2022 23:00:00 da região Sudeste.

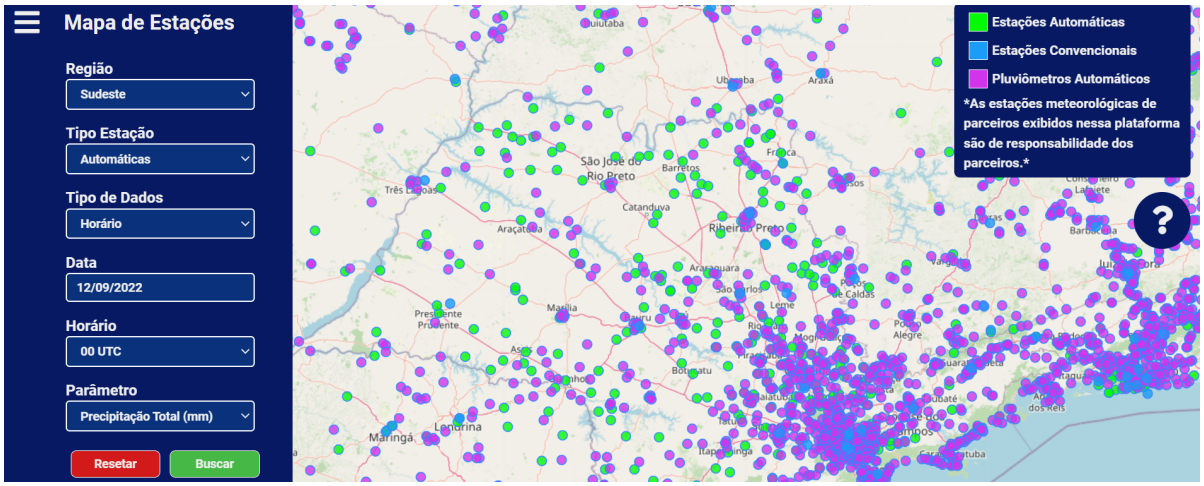
3.2.2 Dados climáticos

Os atributos preditores foram retirados do Instituto Nacional de Meteorologia (INMET), contendo todas as informações climáticas de algumas estações meteorológicas distribuídas na região Sudeste. A missão do INMET, órgão do Ministério da Agricultura, Pecuária e Abastecimento, é prover informações meteorológicas à sociedade brasileira e influir construtivamente no processo de tomada de decisão.

O Sistema de Coleta e Distribuição de Dados Meteorológicos do Instituto (temperatura, umidade relativa do ar, direção e velocidade do vento, pressão atmosférica, precipitação, entre outras variáveis) é dotado de estações de sondagem de ar superior (radiossonda), estações meteorológicas de superfície, operadas manualmente, e a maior rede de estações automáticas da América do Sul.

Os dados coletados por essa rede são disseminados, de forma democrática e gratuita, em tempo real, na página https://portal.inmet.gov.br, tendo aplicação em todos os setores da economia, de modo especial no agropecuário e em apoio à Defesa Civil. A Figura 5 mostra a interface de busca das estações meteorológicas na plataforma e a Figura 6 os detalhes da estação, possibilitando a visualização por gráficos ou o download dos dados.

Figura 5 : Mapa de exploração INMET



Fonte : INMET, 2022

Figura 6 : Zoom por estação meteorológica

[V0428] MONTE APRAZIVEL - SP

Data Inicio 12/09/2022 Data Fim 12/09/2022 Gerar Tabela Baixar CSV

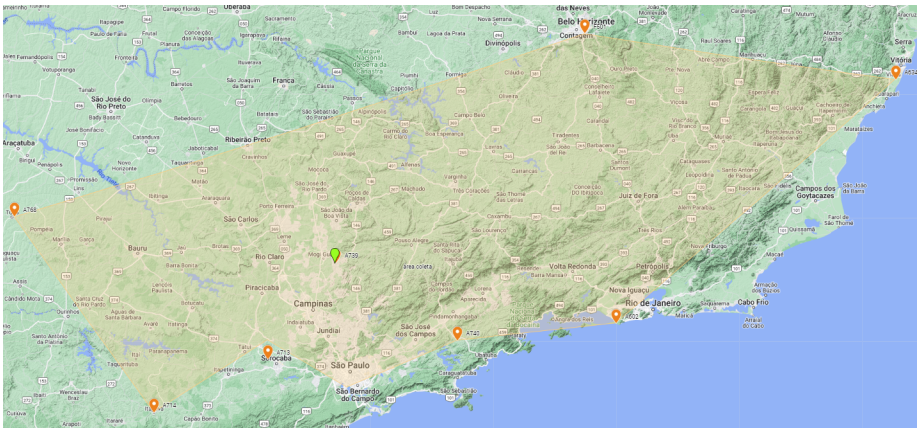
Data	Hora	Temperatura (°C)			Umidade (%)			Pto. Orvalho (°C)			Pressão (hPa)			Vento			Radiação	Chuva
	UTC	Inst.	Máx.	Mín.	Inst.	Máx.	Mín.	Inst.	Máx.	Mín.	Inst.	Máx.	Mín.	Vel. (m/s)	Dir. (°)	Raj. (m/s)	Kj/m²	mm
12/09/2022	0000	24,9			55,0									3,5				0,0
12/09/2022	0100	24,0			55,8									3,5				0,0
12/09/2022	0200																	
12/09/2022	0300	22,7			59,6									4,0				0,0
12/09/2022	0400	21,5			64,0									3,3				0,0
12/09/2022	0500	20,2			67,8									3,8				0,0
12/09/2022	0600	19,0			72,0									4,3				0,0

Fonte: INME,2022

Vale ressaltar que para a aquisição dessas informações não foi possível gerar as tabelas na própria plataforma, devido ao grande volume de dados.

Assim, foram escolhidas estações meteorológicas distribuídas na região Sudeste, na tentativa de abranger as principais áreas que representam o maior consumo de energia, por serem caracterizadas como polos industriais e com maior concentração populacional. Dentre as regiões escolhidas, foram filtradas as estações meteorológicas com um certo grau de consistência dos dados em sua série histórica, e solicitado ao INMET as informações de 01/01/2010 a 31/12/2021. A Figura 7 mostra a distribuição das estações escolhidas (marcações em laranja).

Figura 7 : Distribuição das estações meteorológicas na região Sudeste



Fonte : Autoria própria

Na Tabela 2, pode-se observar na primeira coluna a identificação da estação, seguida por uma breve descrição do local onde a estação está localizada, e por último sua referência geográfica.

Tabela 2 : Posições das estações meteorológicas

estação	descricao local	point
A740	SAO LUIZ DO PARAITINGA - SP	POINT (-45.4198283 -23.2297503)
A714	ITAPEVA - SP	POINT (-48.8900005 -23.9799998)
A768	TUPA - SP	POINT (-50.4900007 -21.9300006)
F501	BELO HORIZONTE - CERCADINHO - MG	POINT (-43.9600019 -19.9800022)
A713	SOROCABA - SP	POINT (-47.5900019 -23.4300019)
A602	RIO DE JANEIRO-MARAMBAIA - RJ	POINT (-43.6000063 -23.0500034)
A634	VILA VELHA - ES	POINT (-40.3999994 -20.4700666)
A739	ITAPIRA - SP	POINT (-46.8096567 -22.4095837)

Fonte: Autoral

Os dados extraídos são compatíveis com a descrição dos metadados conforme descrito no Anexo A.

As Tabelas 4 , 5 e 6 trazem uma amostra dos dados climáticos coletados:

Tabela 4 : Amostra dos dados climáticos 1

Data Medicao	Hora Medicao	PRECIPITACAO TOTAL, HORARIO(mm)	PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO, HORARIA(mB)	PRESSAO ATMOSFERICA REDUZIDA NIVEL DO MAR, AUT(mB)	PRESSAO ATMOSFERICA MAX.NA HORA ANT. (AUT)(mB)	PRESSAO ATMOSFERICA MIN. NA HORA ANT. (AUT)(mB)	RADIACAO GLOBAL(Kj/m²)
2010-01-01	0	0,2	898,3	1011,716525	898,3	897,5	3,974
2010-01-01	100	2,2	899,3	1012,842781	899,3	898,3	-2,122
2010-01-01	200	0,4	899,7	1013,334328	899,8	899,3	0,21
2010-01-01	300	0,6	899	1012,669128	899,7	899	-0,068
2010-01-01	400	0	898,4	1012,075499	899	898,4	-0,05

Fonte : Autoral

Tabela 5 : Amostra dos dados climáticos 2

TEMPERATURA DA CPU DA ESTACAO(°C)	TEMPERATURA DO AR - BULBO SECO, HORARIA(°C)	TEMPERATURA DO PONTO DE ORVALHO(°C)	TEMPERATURA MAXIMA NA HORA ANT. (AUT)(°C)	TEMPERATURA MINIMA NA HORA ANT. (AUT)(°C)	TEMPERATURA ORVALHO MAX. NA HORA ANT. (AUT)(°C)	TEMPERATURA ORVALHO MIN. NA HORA ANT. (AUT)(°C)
22	20,5	18,6	21,3	20,5	19,3	18,6
22	20,5	19,2	20,6	20,2	19,2	18,4
22	20,4	19,4	20,5	20,2	19,4	19,1
21	20,1	19,1	20,4	20,1	19,4	19,1
21	19,9	18,9	20,1	19,8	19,2	18,9
21	19,9	19,0	20,0	19,8	19,1	18,8
21	20,1	19,2	20,1	19,9	19,2	19,0
21	20,1	18,9	20,3	20,1	19,2	18,9

Fonte : Autoral

Tabela 6 : Amostra dos dados climáticos 3

TENSAO DA BATERIA DA ESTACAO(V)	UMIDADE REL. MAX. NA HORA ANT. (AUT)(%)	UMIDADE REL. MIN. NA HORA ANT. (AUT)(%)	UMIDADE RELATIVA DO AR, HORARIA(%)	VENTO, DIRECAO HORARIA (gr)(° (gr))	VENTO, RAJADA MAXIMA(m/s)	VENTO, VELOCIDADE HORARIA(m/s)
12,5	91	87	89	104	4,8	2,5
12,5	92	89	92	33	4,7	1,7
12,4	94	92	94	27	5,9	1,7
12,4	94	94	94	358	4,8	1,1
12,4	94	94	94	277	2,2	0,1
12,4	94	94	94	78	2,3	1,3
12,4	94	94	94	79	2,3	1,0
12,4	94	93	93	55	4,5	2,8
12,3	93	92	92	73	5,0	2,8

Fonte : Autoral

3.2.3 Ambiente computacional

Para o desenvolvimento das análises e modelos vinculados ao presente trabalho, foi utilizada a linguagem Python 2.7.18, rodando no sistema operacional Ubuntu, versão 20.04.2 LTS.

Complementando a linguagem, foram utilizadas bibliotecas Python, sendo estas de código aberto, possibilitando atividades de exploração de dados e construção de modelos estatísticos. Tais bibliotecas são descritas a seguir.

- **Pandas :** Biblioteca frequentemente usada em manipulação e análise de dados. Uma das suas principais funcionalidades é permitir a manipulação de tabelas inteiras e não apenas listas. Permite também uma manipulação mais rápida do que o uso de dicionários em Python porque usa uma otimização interna de seus dados resultando em uma leitura rápida de dados (The Pandas Development team, 2020).
- **XGboost :** Biblioteca usada para implementar o gradient boosting, com a possibilidade de distribuir em um grupo de máquinas o processamento a ser realizado para obter os modelos (Chen e Guestrin, 2016)
- **Numpy :** Biblioteca que permite o uso de cálculos matemáticos para grandes listas (arrays), sendo usada em aprendizado de máquina com grandes quantidades de dados. Esta biblioteca também é muito usada para transformações matemáticas e operações com matrizes (Harris et al., 2020)
- **Scikit-learn :** Desenvolvida especificamente para aplicação prática de machine learning, dispõe de ferramentas simples e eficientes para análise preditiva de dados, é reutilizável em diferentes situações, possui código aberto, sendo acessível a todos e construída sobre os pacotes NumPy, SciPy e matplotilib.

Além disso, foi utilizado o git como ferramenta de versionamento de código e o github como repositório. Todo o projeto se encontra em um repositório público, disponível em : https://github.com/Alorencatto/tcc.

3.3 Normalização, exploração e descoberta de conhecimento nos dados

3.3.1 Pré-processamento

Já com ambas as bases de dados importadas no ambiente de desenvolvimento, foi iniciado o processo de limpeza e tratamento dos dados, com o objetivo de retirar as inconsistências que possam prejudicar no aprendizado do modelo.

Para os dados climáticos, foi obtida uma base por estação, totalizando 8 arquivos csv para o período de 01/01/2010 a 31/12/2021.

A partir de análises preliminares, foram filtrados apenas os dados maiores ou iguais a 2012, pois algumas das estações não tinham informações tão sólidas dos períodos anteriores.

Com o período definido, foram concatenadas todas as colunas de todas as estações, resultando em uma única base de dados com 87762 linhas e 70 colunas, já que cada atributo de cada estação se tornou uma variável. A Tabela 7 ilustra uma amostra dessa junção, representando que na mesma tabela é possível encontrar a variável de precipitação total, porém repetida para todas as estações.

Tabela 7 : Amostra da união das bases de dados climáticos

data-hora	precipitacao_total_ A602	precipitacao_total_A740	precipitacao_total_A739	...
01/01/2012 00:00	2.4	0.2	4.1	...
01/01/2012 01:00	1.6	6.8	2.3	...
01/01/2012 02:00	6.0	0.0	0.2	...
01/01/2012 03:00	1.6	0.2	0.3	...

Fonte : Autoria própria

Além disso, foi necessário o ajuste do metadado em valores do tipo ponto flutuante e por fim a indexação da coluna ‘data-hora’ nessa base de dados, facilitando assim a manipulação com a biblioteca pandas.

Os dados da curva da carga horária são disponibilizados na plataforma da ONS por ano. Dessa forma, foram baixadas as bases de 2012 a 2022, resultando em 11 arquivos csv.

Todas as bases foram concatenadas, resultando em uma única estrutura com 357136 linhas e 4 colunas. A Tabela 8 traz uma amostra do resultado obtido.

Tabela 8: Amostra resultante da junção das bases de curva da carga horária

id_subsistema	nom_subsistema	din_instante	val_cargaenergiahomwmed (MW)
N	NORTE	2012-01-01 00:00:00	3.961.000
NE	NORDESTE	2012-01-01 00:00:00	7.917.469
S	SUL	2012-01-01 00:00:00	7.253.940
SE	SUDESTE	2012-01-01 00:00:00	28.212.460
N	NORTE	2012-01-01 01:00:00	3.865.590

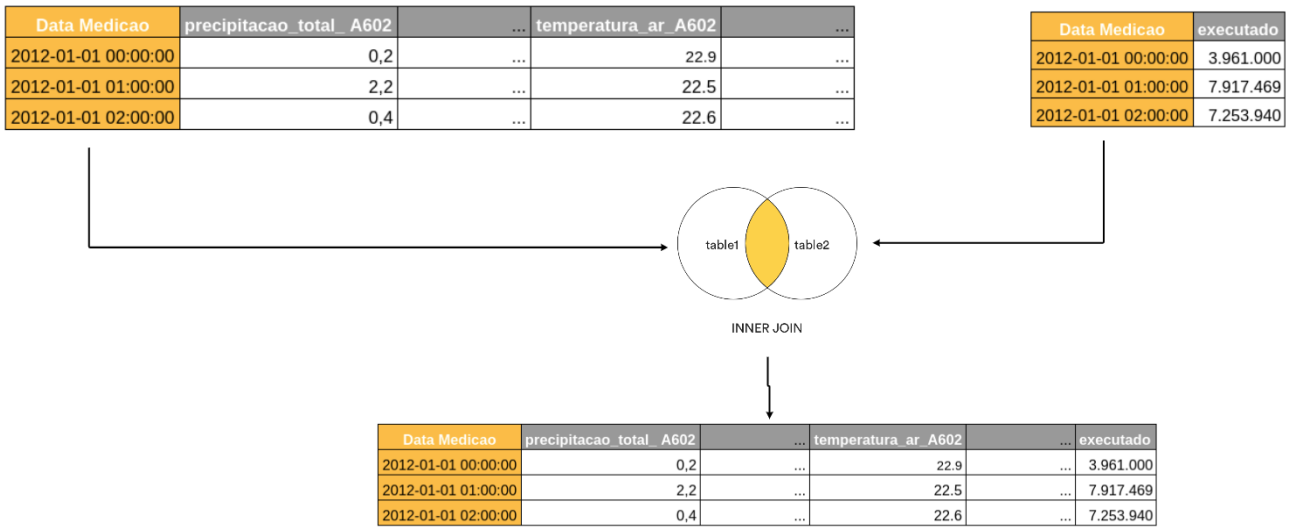
Fonte : Autoria própria

Após a junção das bases, foram filtrados apenas os registros da região sudeste, através da coluna ‘nom_subsistema’.

Assim como nos dados climáticos, foram renomeadas as colunas e foi feita também a indexação da coluna que contém os valores temporais da série, padronizando as duas estruturas de dados.

Por fim, foi realizada a operação de ‘inner join’ entre as bases pela coluna que contém os valores temporais, ou seja, para cada data-hora é obtido todos os atributos climáticos e seus respectivos valores de energia consumida. Esta operação de união obtém somente os valores que estão presentes em ambos os conjuntos, eliminando a possibilidade de haver uma linha com o valor temporal e nenhum dos atributos preditores e alvo. A Figura 8 ilustra o que foi descrito anteriormente.

Figura 8 : Ilustração da operação inner join



Fonte : Autoria própria

Assim, como resultado dessa etapa de pré-processamento foi obtida uma base de dados com frequência de 1 hora, contendo os atributos climáticos de todas as estações meteorológicas utilizadas e os valores de energia consumidos.

3.3.2 Engenharia de variáveis

É de suma importância o conhecimento dos atributos preditores e derivá-los de forma a criar novas formas de entendimento do ambiente de estudo, adicionando insumos para as posteriores tarefas de aprendizado.

No presente estudo, foram adicionadas informações externas às fontes originais, as quais fazem parte do contexto do problema, e variáveis derivadas das originais, expondo relações entre dados, necessitando assim de uma refinamento mais profundo para sua utilização.

A seguir, serão sucintamente descritas as variáveis criadas nesse processo.

- **Variáveis temporais**

Oo atributo temporal é identificado com um objeto do tipo `datetime` dentro da linguagem Python, sendo possível extrair :

1. Ano da medição;
2. Mês da medição;
3. Semana da medição;
4. Dia da medição;
5. Dia da semana da medição;
6. Hora da mediação.

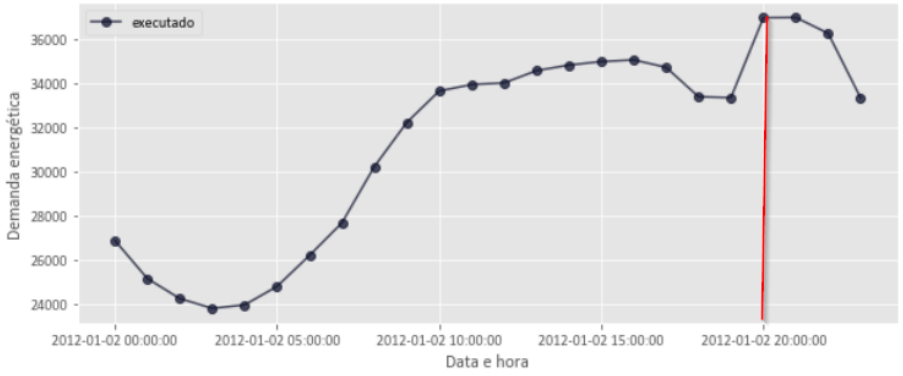
Com isso, pode-se por exemplo entender que o consumo de energia em determinado mês se concentrou mais em dias de finais de semana, ou até mesmo explorar os horários dos finais de semana que o consumo é de fato mais alto.

- **Pico de potência**

Foi criado também uma marcação do pico de potência diário, ou seja, no dia de análise qual foi a hora em que se atingiu o consumo mais alto de energia.

Para chegar nessa informação, foi realizada uma cópia dos dados originais e em seguida realizado o agrupamento pelo dia, hora e valor de energia consumido e obtido o valor máximo a partir do agrupamento. A figura 9 retrata a marcação do pico de potência

Figura 9 : Demanda de Potência amostra (2012/01/02) com marcação de pico de energia



Fonte : Autoria própria

Dessa forma, para cada dia entre 01/01/2012 até 31/12/2021 foi obtida uma hora

específica entendida como pico de energia diário.

- **Feriados nacionais**

O comportamento energético na região Sudeste está diretamente relacionado com os padrões de consumo da população e da indústria como um todo. Portanto, entender se o dia da análise é útil ou não é extremamente relevante.

Para isso, foi utilizada a biblioteca ‘holidays’ do Python que traz um dicionário com todos os feriados nacionais dentro de um período.

Dessa forma, foram consultados através desse recurso todos os feriados de 2012 a 2022 e realizadas as marcações na base de dados, obtendo-se, assim, a informação se no dia analisado foi feriado ou não.

- **El nino e la nina**

Pensando na correlação com os atributos climáticos, foi criado uma variável na base de dados, rotulando se o ano daquela análise estava ocorrendo algum dos fenômenos climáticos El Ninõ ou La Ninã a partir de informações do Instituto Nacional de Pesquisas Espaciais (INPE,2022).

- **Atributos autocorrelacionadas**

Como explorado mais detalhadamente na seção de revisão da literatura, a autocorrelação em uma série temporal é definida como uma observação num determinado instante está relacionada às observações passadas.

No presente estudo, foram trabalhados dois tipos de correlação, onde o procedimento de aplicação se encontra descrito a seguir.

1. Lag

A variável de `lag` nada mais é que o número de períodos de tempo que separam duas séries temporais, ou seja, o *delay* entre elas. Com a biblioteca pandas, pode-se criar essa variável facilmente utilizando o método shift, passando como argumentos o número de período de tempos que são deslocados juntamente com a frequência associada a esse período.

A Figura 10 ilustra a aplicação desse método, onde a variável *df_demanda_potencia* é

a base de dados com a série temporal, dados climáticos e consumo energético.

Figura 10 : Ilustração aplicação do método shift

```
df_demanda_potencia['lag_1'] = df_demanda_potencia['executado'].shift(1, freq = 'D')
df_demanda_potencia['lag_7'] = df_demanda_potencia['executado'].shift(7, freq = 'D')
df_demanda_potencia['lag_14'] = df_demanda_potencia['executado'].shift(14, freq = 'D')
```

Fonte : Autoria própria

Um exemplo do deslocamento da série temporal é ilustrado na Figura 11, onde o valor executado às 2:00AM do dia 01/02/2012 é a variável *`lag_1`* às 2:00AM do dia 02/02/2012, representando assim o deslocamento exato de um dia.

Figura 11 : Ilustração do Lag

data-hora	executado	lag_1
2012-02-01 00:00:00	33738.15	33936.69
2012-02-01 01:00:00	31389.03	31515.64

data-hora	executado	lag_1
2012-02-02 00:00:00	35096.36	33738.15
2012-02-02 01:00:00	32715.68	31389.03

Fonte : Autoria própria

O mesmo deslocamento foi realizado também para 7 e 14 dias, compondo assim 3 variáveis de lag na base de dados.

2. Rolling mean

Complementando as variáveis autocorrelacionadas, foi utilizada também a média dos valores de consumo de energia dentro de uma janela de tempo, sendo a referência a cada momento da série temporal. Esta etapa foi possível através da utilização do método rolling da biblioteca pandas, a qual permite definir uma janela contínua numa série temporal para realizar operações matemáticas. A Figura 12 ilustra a aplicação desse método.

Figura 12 : Aplicação do método rolling

```
df_demanda_potencia['rolling_mean_1'] = df_demanda_potencia['executado'].rolling(window=96).mean()
df_demanda_potencia['rolling_mean_7'] = df_demanda_potencia['executado'].rolling(window=672).mean()
df_demanda_potencia['rolling_mean_14'] = df_demanda_potencia['executado'].rolling(window=1344).mean()
```

Fonte : Autoria própria

Um exemplo dos resultados obtidos com essa aplicação, é ilustrado na Figura 13, onde o valor destacado em amarelo representa a média dos valores de 96 períodos anteriores à 2:00AM do dia 02/02/2012.

Figura 13 : Exemplo rolling

data-hora	executado	rolling_mean_1
2012-02-01 00:00:00	33738.15	34368.257500
2012-02-01 01:00:00	31389.03	34370.400833
2012-02-01 02:00:00	30198.53	34376.110417

Fonte : Autoria própria

Assim como no período de 96, foi realizado o cálculo também para 672 e 1344 períodos.

- Preenchimento de dados nulos

Para esse tipo de tratativa, foi utilizado um método muito comum para o cálculo de valores faltantes, que é a interpolação, a qual basicamente estima pontos desconhecidos entre dois pontos conhecidos. Tendo em vista os métodos de interpolação, foi utilizado o polinomial, o qual preenche os valores alvo com o ponto mais baixo que consiga passar entre os dois pontos conhecidos. A curva de tal função é uma parábola, já que o grau da função polinomial utilizada foi 2.

3.3.3 Análise exploratória

Embora não contemplasse os objetivos gerais do trabalho, a etapa exploratória foi uma etapa possível devido ao pré-processamento realizado, garantido uma qualidade aceitável para as amostras, as quais foram utilizadas como insumos para gerar algumas

análises.

A primeira delas é o comportamento da série temporal como um todo, devido ao amplo período das amostras. Gráficos auxiliam muito nesse processo, sendo a biblioteca *matplotlib* utilizada para suas plotagens, devido a integração desse pacote com a biblioteca *pandas*. Somado a isso, o período contempla um marco importante na década, a pandemia do Covid-19, no qual foi possível se observar o impacto desse evento tão significativo, como fator externo, na demanda energética do estado de SP.

Outra análise interessante foi a distribuição da probabilidade da demanda ao longo dos anos, sendo utilizada a biblioteca *seaborn* com a função *displot* para tal função.

Outra distribuição da probabilidade aplicada foi a do pico de energia, envolvendo o período de 24 horas. Para isso, também foi utilizada a biblioteca *seaborn* com a função *displot*.

3.3.4 Implementação e treinamento do modelo

Garantida a qualidade dos dados, conforme descrito pelas etapas anteriores, a série temporal como um todo foi dividida de forma a 70% da base corresponder ao conjunto de treino, totalizando 61399 registros, os quais contemplam o período de 01/01/2012 00:00:00 até 01/01/2019 02:00:00, e os 30% restantes retratam o período de 1/01/2019 03:00:00 até 31/12/2021 23:00:00, totalizando 26301 registros.

Para o treinamento de fato, foi utilizado a biblioteca Scikit-learn, que possui uma função específica para o treinamento de séries temporais, conforme descrito na revisão de literatura. Dessa forma, a série como um todo foi dividida em 5 grandes segmentos.

Esses segmentos foram submetidos a uma série de hiperparâmetros, os quais são utilizados para controlar o processo de aprendizado no algoritmo, e em seguida utilizado a funcionalidade de *Grid Search*, que é uma ferramenta utilizada para automatizar o processo de ajuste dos parâmetros no processo de aprendizado (UYEKITA,2019). Os tópicos a seguir expõem alguns dos hiperparâmetros utilizados:

1. Learning rate : Taxa de aprendizado para o gradiente. São as quantidades de correções feitas em cada adição de árvore, e quanto menor seu valor, menor a possibilidade de overfitting. Os valores testados foram de 0.1 e 0.05.
2. Gamma : Basicamente, o fator de regularização entre árvores. Compreende o

controle da complexidade do modelo. Os valores testados foram de 0.01, 0.1,0.3,0.5.

3. Max depth: Profundidade máxima da árvore, ou seja, quantidade de nós até a folha. Contempla o ajuste macro do modelo para evitar o overfitting. Foram testados valores de 2,4,7,10.
4. Min Child Lenght : Soma mínima de pesos das observações. Basicamente, a quantidade mínima de amostras para se criar um novo nó. Os valores testados foram de 1,3,5,7.
5. N estimators : Número de árvores. Foram testados os valores de 100, 250 e 500.

Realizada a etapa de escolha dos hiperparâmetros, foram recuperados esses valores e submetidos ao regressor do modelo do XGboost, juntamente com os dados de treino e teste, descritos anteriormente.

3.5 Métricas de erro

Para validação e entendimento da assertividade do modelo criado, foram utilizadas funções que implementam métricas específicas para problemas que envolvem séries temporais, descritas a seguir :

- Mean Absolute Error (MAE) : Essa métrica é calculada a partir da média das diferenças absolutas entre os valores originais, ou seja, os valores de validação e os valores que foram preditos pelo modelo. Vale ressaltar que essa métrica é extremamente simples e apenas é válida para comparação entre séries com a mesma unidade. Além disso, em cenários em que a penalização de outliers é uma premissa, essa métrica não é recomendada.
- Mean Squared Error (MSE) : É considerada uma média entre o quadrado dos erros (diferença entre o valor original e o valor predito). Nesse cenário, quando o erro é elevado ao quadrado, valores maiores são expostos em outliers, penalizando esse tipo de comportamento. Em suma, essa métrica é interessante quando deseja-se focar em grandes erros entre os valores originais e preditos. Pode-se entender como desvantagem dessa métrica, a perda de unidade.
- Root Mean Square Error (RMSE) : Trata-se do MSE porém sem a perda de unidade no erro. Essa métrica, como o MAE, é mais sensível a outliers.
- Mean Absolute Percentage Error (MAPE) : Trata-se de uma das métricas mais

populares em problemas com séries temporais, sendo calculado com a média da diferença absoluta entre os valores originais e os valores preditos dividido pelos valores originais. Essa métrica entra no grupo de métricas de erros percentuais sendo muito simples de ser interpretada, além de independente da escala e do tipo do problema.

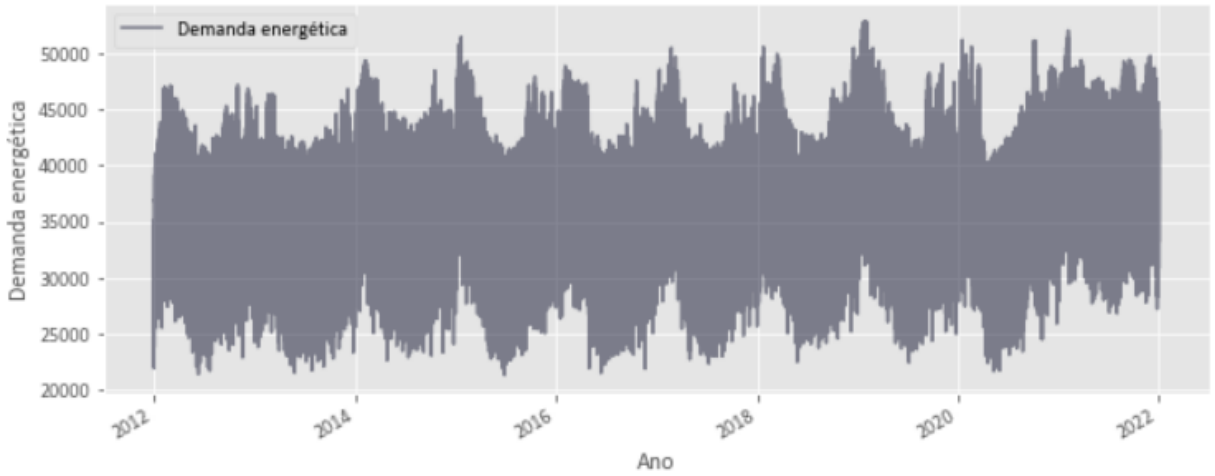
4. RESULTADOS

4.1 Análise exploratória

4.1.1 Comportamento da série histórica

Como já abordado anteriormente, o trabalho tem como espaço amostral de tempo os dados de 2012 a 2021, tanto das informações climáticas quanto da demanda energética. No Gráfico 1 pode ser visualizado o comportamento da série histórica completa da demanda energética, dentro do período citado.

Gráfico 1 : Plotagem da série histórica completa



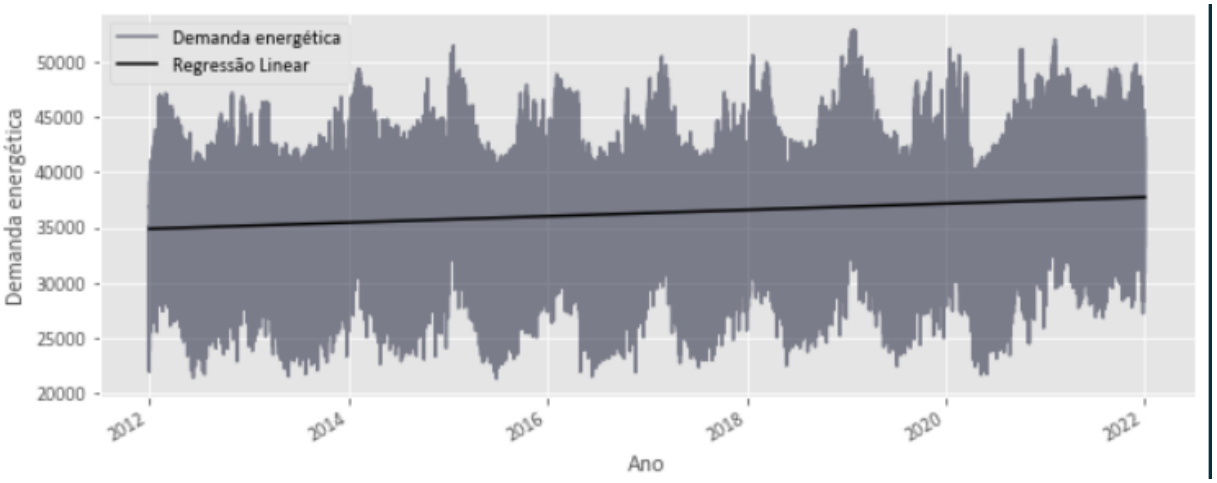
Fonte : Autoria própria

É claro o comportamento sazonal na série histórica, atingindo valores mais altos no começo de cada ano.

Além disso, percebe-se também uma tendência ao aumento dessa demanda, a qual foi modelada por uma regressão linear simples, considerando apenas os atributos temporais e os dados da curva da carga horária, com o auxílio da biblioteca Scikit-Learn, com o resultado

apresentado no Gráfico 2.

Gráfico 2 : Tendência demanda de energética 2012-2020

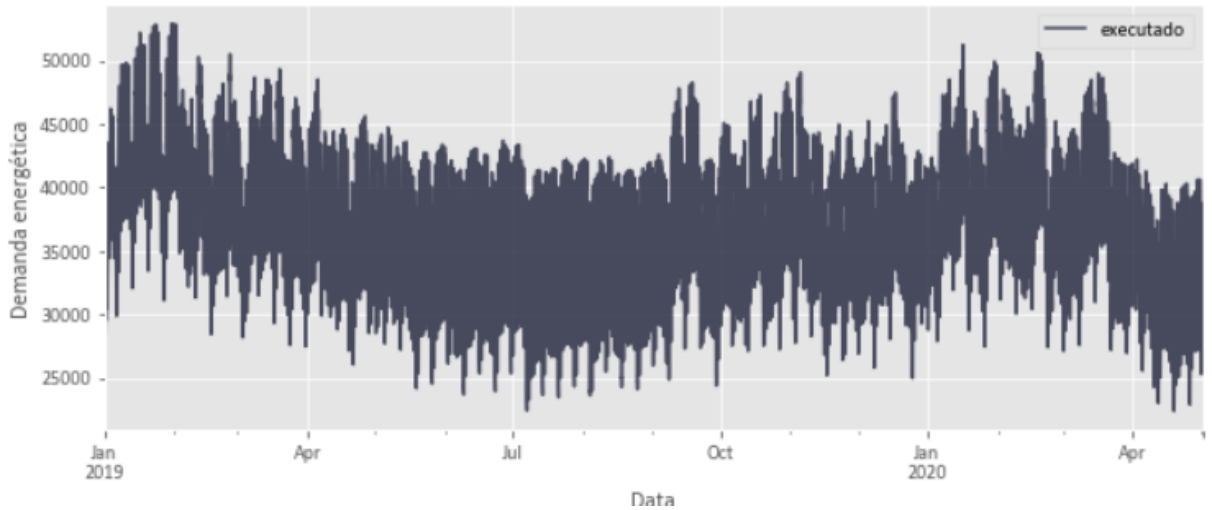


Fonte : Autoria própria

O coeficiente dessa regressão linear é positivo, o que de fato comprova o crescimento da demanda energética ao longo dos anos.

Diante da série temporal abordada, é possível notar também a diferença de comportamento nos períodos em que houveram as maiores restrições durante a pandemia, envolvendo lockdown, diminuição de produção fabril, aumento do desemprego, entre outros fatores. Para melhor visualização dessa situação, o Gráfico 3 apresenta somente o período do começo de 2020 até o final de 2021.

Gráfico 3 : Demanda de Potência 2019 - 2020 (Efeito Covid 19).



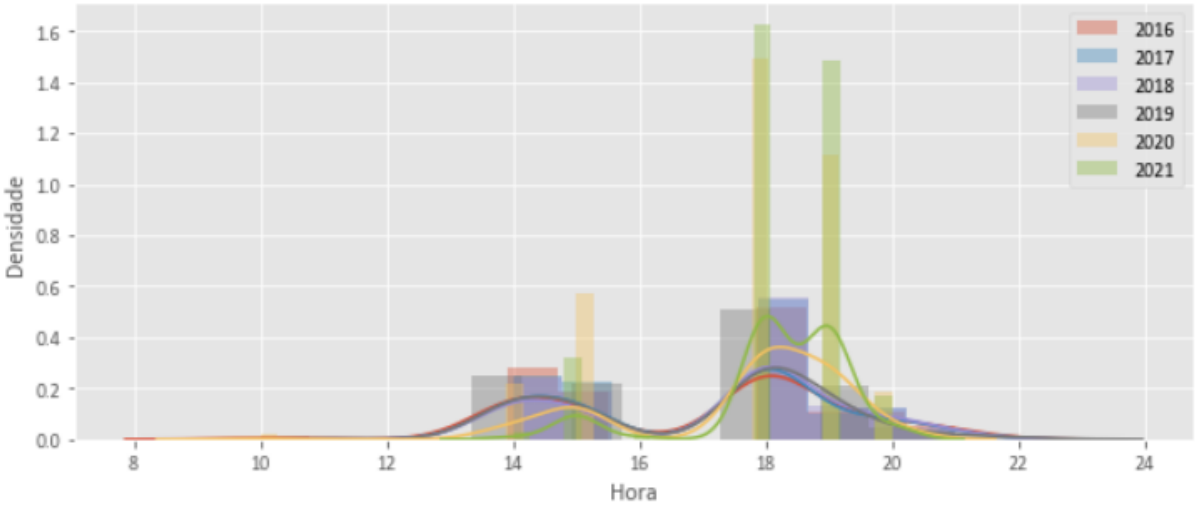
Fonte : Autoria própria

Diante do gráfico apresentado, o comportamento vai de encontro com o estudo de VALLEJOS (2020) que retrata que as medidas restritivas do Covid-19 começaram a ter impacto significativo no setor energético na semana de 23 de março de 2020, totalizando uma queda na demanda do SIN de aproximadamente 14,9%. No Gráfico 3, pode-se notar essa queda nesse período.

4.1.2 Distribuição da probabilidade

Foi feita a análise para a variável referente ao pico de demanda energética ao longo do dia, entre 2016 e 2021, com o auxílio da estimativa de densidade do Kernel (KDE), utilizado para visualizar a densidade de probabilidade de uma variável contínua, o que pode ser visualizado no Gráfico 4.

Gráfico 4 : Distribuição de Probabilidade do pico de energia

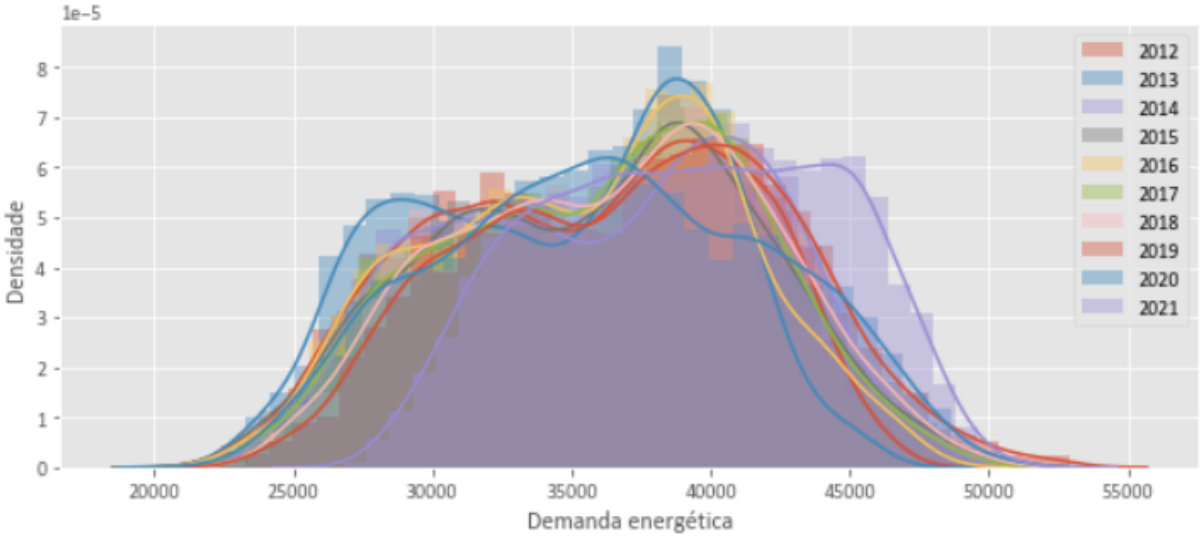


Fonte : Autoria própria

Dessa forma, as linhas sólidas representam as funções de probabilidade, destacando o fato de que nos anos de 2020 e 2021, na maior parte das vezes o pico energético foi as 18:00 horas.

A mesma análise foi realizada também para a variável da curva da carga horária, cujo resultado é apresentado no Gráfico 5 .

Gráfico 5 : Distribuição de Probabilidade da Demanda Energética



Fonte : Autoria própria

Pode-se entender que as curvas de distribuição ao longo dos anos se deslocam para a direita, indicando que há um aumento constante dessa variável. Essa análise só fortalece o que foi modelado com a regressão linear simples.

4.2 Resultados do modelo

Nesta seção serão exibidos os resultados alcançados com a utilização do modelo XGboost para a predição da demanda energética, expresso em gigawatt (GW).

Dentre os hiperparâmetros disponibilizados para o *Grid Search* realizar a escolha no momento do treinamento, foram escolhidos:

1. Learning rate : 0.1
2. Gamma : 0.3
3. Max depth: 10
4. min child lenght : 1
5. N estimators : 500

Com isso, o modelo foi treinado e apresentou as seguintes métricas de resultado :

1. Mean Absolute Error (MAE) : 1.2265.
2. Mean Squared Error (MSE) : 2.6503.
3. Root Mean Square Error (RMSE) : 1.628.

4. Mean Absolute Percentage Error (MAPE) : 3.2428.

Vale destacar que o MSE apresentou um valor significativamente superior ao MAE, expondo que no modelo treinado houve presença de outliers, atuando diretamente na acurácia. Para a métrica de MAPE, o valor obtido se mostrou satisfatório, apresentando aproximadamente 3% de erro percentual, viabilizando a utilização do modelo.

Para o melhor entendimento e materialização dos resultados, a Tabela 9 expõe uma amostra dos dados originais, previsão, instante da previsão, diferença em GW e por fim diferença percentual entre os valores originais e a previsão.

Tabela 9 : Amostra dos resultados da previsão

data-hora	previsao	real	diferenca	diferenca_percentual
2019-01-01 03:00:00	295.313	299.909	4.596	0,016
2019-01-01 05:00:00	274.811	289.441	14.630	0,053
2019-01-01 06:00:00	290.094	275.912	14.182	0,049
2019-01-01 09:00:00	287.421	285.375	2.046	0,007
2019-01-01 11:00:00	321.043	300.959	20.084	0,063
2019-01-01 12:00:00	31.075	30.188	887	0,029
2019-01-01 13:00:00	306.854	297.339	9.515	0,031
2019-01-01 14:00:00	296.081	295.134	947	0,003
2019-01-01 15:00:00	299.561	296.362	3.199	0,011
2019-01-01 16:00:00	299.377	301.027	1.650	0,006

Fonte : Autoria própria

Um ótimo benefício da utilização de ensamble com árvores de decisão é a facilidade de entender e gerar o valor estimado da importância de cada variável na capacidade preditiva do modelo. A métrica utilizada foi de “peso”, que retrata o número de vezes que cada variável foi utilizada na divisão dos dados nos nós das árvores. Dessa forma, a Tabela 10 retrata a pontuação das principais variáveis que resultantes dessa métrica.

Tabela 10 : Importâncias das variáveis na construção do modelo

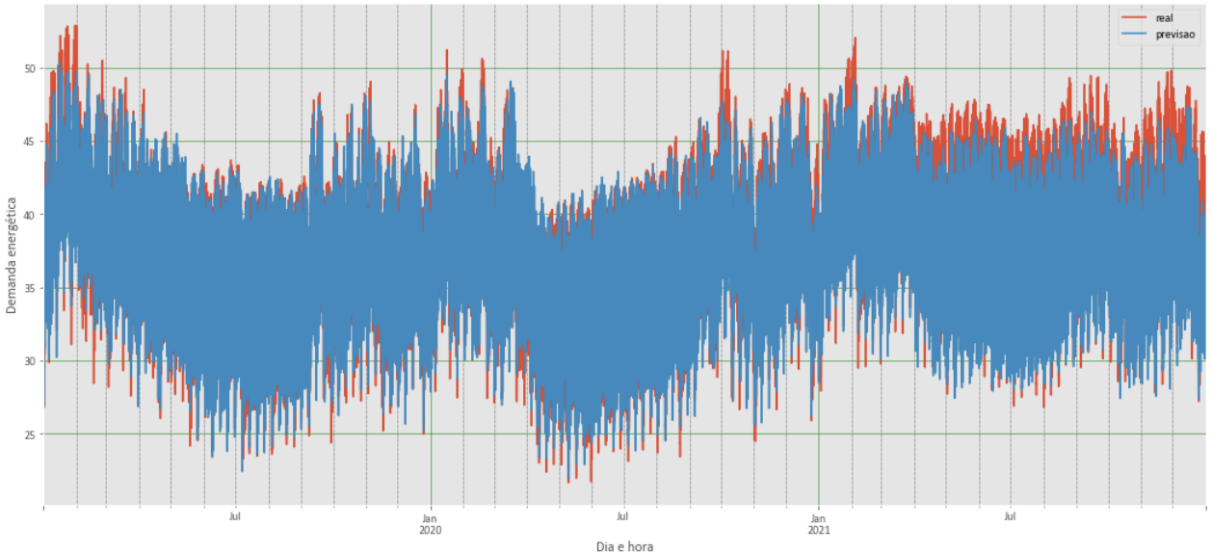
Variável	Descrição da variável	Score de importância
rolling_mean_1	Média dos 96 períodos anteriores	42080
lag_1	Demanda energética deslocada 1 dia	41310
pressao_atmosferica_A602	Pressão atmosférica da estação A602	38510
lag_7	Demanda energética deslocada 7 dias	36300
lag_14	Demanda energética deslocada 14 dias	35640
rolling_mean_14	Média dos 1344 períodos anteriores	34420
temperatura_ponto_orvalho_A602	Temperatura do ponto de orvalho da estação A602	33390
rolling_mean_7	Média dos 672 períodos anteriores	33320
temperatura_ar_A602	Temperatura do ar na estação A602	33050
vento_direcao_A602	Direção do vento na estação A602	32930
dia	Dia da medição	31290

Fonte : Autoria própria

Nota-se que as variáveis de autocorrelação com a série temporal obtiveram uma grande influência no modelo, assim como algumas informações climáticas da estação A602, localizada no Rio de Janeiro.

O Gráfico 6 mostra o ajuste dos valores da previsão em relação aos dados originais, onde pode-se visualizar a efetividade do modelo desenvolvido.

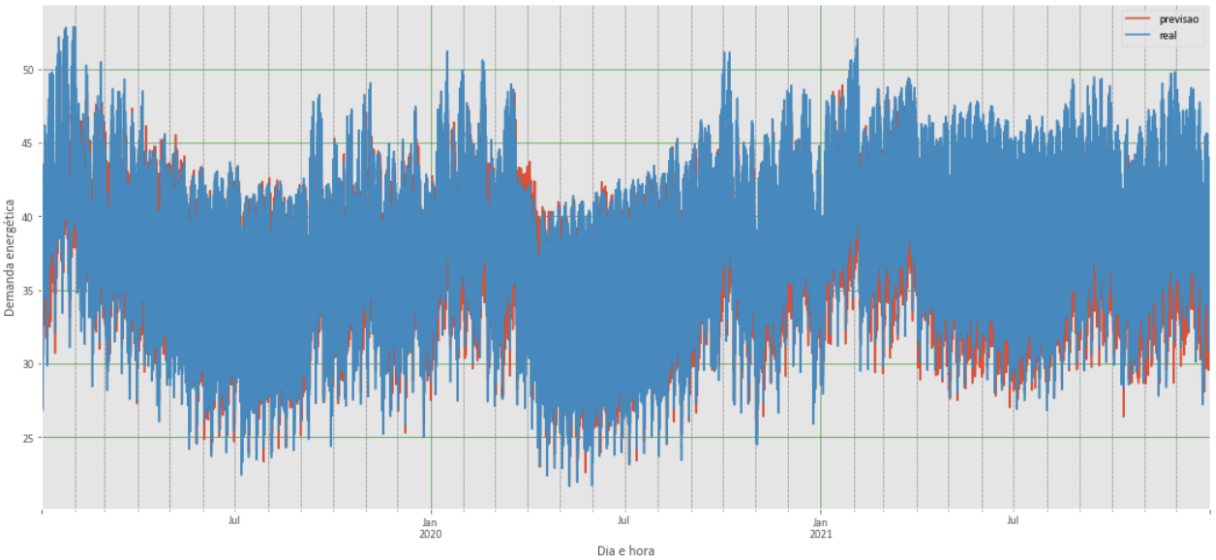
Gráfico 6 : Ilustração com a previsão modelo (em azul) comparado aos dados reais (em vermelho).



Fonte : Autoria própria

Já o gráfico 7, traz a mesma análise porém com os dados originais em primeiro plano, expondo outro ponto de vista os resultados obtidos

Gráfico 7 : Ilustração com a previsão modelo (em vermelho) comparado aos dados reais (em azul)



Fonte : Autoria própria

Nota-se que no modelo apresentado, a solução perde certa eficiência após o ano de 2021, fato que pode ser explorado em trabalhos futuros.

5. CONCLUSÃO

O presente trabalho buscou construir um modelo baseado em IA para tratar séries temporais englobando as principais etapas dentro de um projeto analítico voltado para machine learning, envolvendo conceitos de mineração, exploração e processamento de dados como etapas anteriores ao objetivo principal.

Entende-se que o país possui uma excelente série histórica tanto em aspectos climáticos quanto em curva da carga horária, além de outros componentes que podem ser amplamente explorados, individualmente ou integrados com a presente solução. As interações com esses dados podem ser de extrema importância para uma abordagem mais unificada em termos de gestão.

O modelo desenvolvido apresentou resultados bastante satisfatórios, com desvios em relação aos valores originais muito baixos, comprovando a coerência da solução desenvolvida além da segura aplicabilidade em outras regiões, diante da metodologia estabelecida.

Entende-se que a análise gerada e a construção desse modelo preditivo é o início de uma solução mais completa em termos de gestão de recursos energéticos, podendo ser posteriormente somada a uma análise integrada das principais matrizes geradoras de energia, assim como sua origem, de modo a priorizar a utilização de energias limpas. Além disso, os horários dos picos combinados com as tarifas dinâmicas podem apresentar informações riquíssimas para o planejamento de indústrias em termos de horário de produção, focando principalmente na redução de custos.

Assim, como forma de dar continuidade no trabalho, torna-se interessante a aplicação do que foi descrito no parágrafo anterior além da implementação de outros modelos preditivos, como por exemplo, as redes neurais artificiais.

ANEXOS

ANEXO A - Descrição dos metadados climáticos

Variável	Unidade
PRECIPITACAO TOTAL	milímetro (mm)
PRESSAO ATMOSFERICA AO NIVEL DA ESTACAO, HORARIA	milibar (mb)
PRESSAO ATMOSFERICA REDUZIDA NIVEL DO MAR, AU	milibar (mb)
PRESSAO ATMOSFERICA MAX.NA HORA ANT. (AUT)	milibar (mb)
PRESSAO ATMOSFERICA MIN. NA HORA ANT.	milibar (mb)
TEMPERATURA DA CPU DA ESTACAO	graus celsius (°C)
RADIACAO GLOBAL	kilojoule / m² (kj/m²)
TEMPERATURA DO AR - BULBO SECO, HORARIA	graus celsius (°C)
TEMPERATURA DO PONTO DE ORVALHO	graus celsius (°C)
TEMPERATURA ORVALHO MAX. NA HORA ANT. (AUT)	graus celsius (°C)
TEMPERATURA ORVALHO MIN. NA HORA ANT. (AUT)	graus celsius (°C)
TEMPERATURA MAXIMA NA HORA ANT. (AUT)	graus celsius (°C)
TEMPERATURA MINIMA NA HORA ANT. (AUT)(	graus celsius (°C)
UMIDADE RELATIVA DO AR, HORARIA	percentual (%)
UMIDADE REL. MAX. NA HORA ANT	percentual (%)
UMIDADE REL. MIN. NA HORA ANT.	percentual (%)
TENSAO DA BATERIA DA ESTACAO	Volt (V)
VENTO, DIRECAO HORARIA	
ENTO, RAJADA MAXIMA	metros / segundo (m/s)
VENTO, VELOCIDADE HORARIA	metros / segundo (m/s)

Fonte : Autoria própria

REFERÊNCIAS

ASSUNÇÃO, Jéssica. **Séries temporais - definições e características.** Dadosfera, São paulo, 21 de jun. de 2020. Disponível em : https://medium.com/data-sprints/s%C3%A9ries-temporais-defini%C3%A7%C3%B5es-e-caracter%C3%ADsticas-698d85f4b353#:~:text=Caracter%C3%ADsticas%20de%20uma%20s%C3%A9rie%20temporal,utilizado%2C%20%C3%A9%20dependente%20dessas%20caracter%C3%ADsticas.
Acesso em 28 set. 2022.

BERGAMO, Henrique Postingel. **Análise Preditiva da Geração Fotovoltaica via Algoritmos de Inteligência Computacional: Modelagem e Estudo de Caso da Usina Solar Bom Jesus da Lapa - BA.** Orientadores : Wallace Correa de Oliveira Casaca e Marilaine Colnago.2022.TCC - Universidade Estadual Paulista "Júlio de mesquita filho" (UNESP), 2022. Disponível em : https://repositorio.unesp.br/bitstream/handle/11449/216985/bergamoramos_hplg_tcc_rosa.pdf?sequence=8&isAllowed=y
Acesso em 27 set. 2022.

HYNDMAN, R.J. ATHANASOPOULOS, G. Forecasting: principles and practice, OTexts, Melbourne, Australia, vol. 2, 2018.

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS (INPE),2022. **Condições atuais do ENOS : LA NIÑA.** Disponível em : http://enos.cptec.inpe.br/. Acesso em 2 nov. 2022.

IZBICKI, R.; SANTOS, T. M. **Aprendizado de máquina: uma abordagem estatística**. São Carlos: Rafael Izbicki, 2020.
Acesso em 29 set. 2022.

Kriger, Brunno. **O QUE É GIT: CONCEITOS, PRINCIPAIS COMANDOS E QUAIS AS VANTAGENS?** Disponível em : https://kenzie.com.br/blog/o-que-e-git/. Acesso em 20 nov. 2022.

MADDALA, G S e LAHIRI, Kaja (2009). **Conceitos Básicos de Séries Temporais para Modelagem Macroeconômica.** Disponível em : http://www.icad.puc-rio.br/cfeijo/pdf/revis%C3%A3o%20b%C3%A1sica%20s%C3%A9ries%20temporais_material%20de%20apoio_curso%20teoria%20macroeconomica_PPGE%20UFF.pdf.
Acesso em 28 set. 2022.

MARQUES,André e FLÓRIO,Denise. **Aplicação do método XGBoost em comércio eletrônico de roupas.** Laboratório de ciência de dados MAC-6967.
Acesso em 1 out. 2022.

MEDEIROS,Augusto Santana Veras. **ESTUDO SOBRE O USO DE ANÁLISE TÉCNICA E XGBOOST EM OPERAÇÕES DE DAY-TRADE.** Orientador : Prof. Dr. Marcus Alexandre Nunes**.** Universidade Federal do Rio Grande do Norte,,Natal,2021.
Disponível em :
https://repositorio.ufrn.br/bitstream/123456789/33362/1/Estudosobreousodean%c3%a1liset%c3%a9cnicaexgboostemopera%c3%a7%c3%b5esdeday-trade_Medeiros_2021.pdf
Acesso em 29 set. 2022

MIRANDA, Priscila Bernardeli. **Análise do efeito da temperatura na previsão de curto prazo da demanda de energia elétrica através de redes neurais.** Orientador : Prof. Dra. Laura Lisiane Callai Dos Santos. 2021.TCC - Universidade Federal de Santa Maria (UFSM), Rio grande do Sul, 2021. Disponível em : https://repositorio.ufsm.br/bitstream/handle/1/26056/Miranda_Priscila_Bernardeli_2021_TCC.pdf?sequence=1&isAllowed=y.
Acesso em 27 set. 2022.

MIYAKI,Keita (2019). **Time Series Split with Scikit-learn.** Medium. Disponível em : https://medium.com/keita-starts-data-science/time-series-split-with-scikit-learn-74f5be38489e.
Acesso em 29 set. 2022.

MORETTIN, P. **Análise de Séries Temporais**. Edgard Blucher, São Paulo, 2004

PALIT, A.K.e Popovic, D.2005. **Computational Intelligence in Time Series Forecasting - Theory and Engineering Applications**. 1.s.l : Springer,2005.
Acesso em 28 set. 2022.

PANDAS. Disponível em : https://pandas.pydata.org/docs/ . Acesso em 01 dez. 2022

PONTES, Lucélia Pontes. **Previsão de séries temporais : Produção industrial e demanda de energia elétrica residencial no Brasil.** Orientador : Prof. Dr. Carlos Enrique Carrasco Gutierrez. 2018. Pós-graduação Stricto Sensu em Economia,Brasília, 2018. Disponível em : https://bdtd.ucb.br:8443/jspui/bitstream/tede/2645/2/LuceliaPontesePontesTese2018.pdf
.
Acesso em 27 set. 2022.

RINK, Konstantin. **Time Series Forecast Error Metrics You should Know**. Towards Data Science. Disponível em : https://towardsdatascience.com/time-series-forecast-error-metrics-you-should-know-cc88b8c67f27
Acesso em 1 out. 2022.

ROCCA, J. (2019). **Ensemble methods: bagging, boosting and stacking.** Medium towards data science.Disponível em : https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205
Acesso em 29 set. 2022

SILVEIRA, Tatiana Maria Andrade. **Modelos de previsão de carga elétrica em curto prazo desenvolvidos com redes neurais artificiais e lógica fuzzy considerando a variável temperatura.** Orientador : Ronaldo R.B de Aquino, D.Sc. 2010. Pós-graduação em engenharia elétrica, Universidade federal de pernambuco, Recife, 2010. Disponível em : https://repositorio.ufpe.br/bitstream/123456789/5329/1/arquivo5618_1.pdf. Acesso em 27 set. 2022.

SOLTAU, Samuel Bueno. **Detenção da periodicidade em dados multifrquência de Núcleos Ativos de Galáxias com aprendizagem de máquina (XGboost).** Orientador : Prof.Dr.Luiz Claudio Lima Botti. 2019. Universidade Presbiteriana Mackenzie, São Paulo, 2019. Disponível em : https://dspace.mackenzie.br/bitstream/handle/10899/25803/Tese%20Samuel%20Bueno%20Soultau%20PROTEGIDO.pdf?sequence=1&isAllowed=y Acesso em 28 set. 2022.

UYEKITA, AH, 2019. **Como usar o GridSearchCV** .Disponível em : https://andersonuyekita.github.io/notebooks/blog/2019/03/21/como-usar-o-gridsearchcv/. Acesso em 2 nov. 2022.

VALLEJOS, Claudio Andres Villegas, 2020. **O impacto da Covid-19 sobre a demanda de energia elétrica no Brasil**. Refinitiv. Disponível em : https://www.refinitiv.com/pt/blog/trading/o-impacto-da-covid-19-sobre-a-demanda-de-energia-eletrica-no-brasil/#:~:text=As%20medidas%20restritivas%20come%C3%A7aram%20a,de%20aproximadamente%2014%2C9%25. Acesso em 2 out. 2022.