UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE TECNOLOGIA E GEOCIÊNCIAS
DEPARTAMENTO DE ENGENHARIA ELÉTRICA
CURSO DE GRADUAÇÃO EM ENGENHARIA ELÉTRICA

CAIO SÓTER DE BARROS MOTA

PREVISÃO DE POTÊNCIA NECESSÁRIA PARA RESFRIAMENTO E AQUECIMENTO DE AMBIENTES RESIDENCIAIS

Recife
2024

CAIO SÓTER DE BARROS MOTA

PREVISÃO DE POTÊNCIA NECESSÁRIA PARA RESFRIAMENTO E AQUECIMENTO DE AMBIENTES RESIDENCIAIS

Trabalho de Conclusão de Curso apresentado ao Departamento de Engenharia Elétrica da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Engenheiro Eletricista.

Orientador(a): Prof. Dr. Jeydson Lopes da Silva

Recife
2024

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Mota, Caio Sóter de Barros.
Previsão de potência necessária resfriamento e aquecimento de ambientes residenciais / Caio Sóter de Barros Mota. - Recife, 2024.
75 p. : il., tab.

Orientador(a): Jeydson Lopes da Silva
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Tecnologia e Geociências, Engenharia Elétrica - Bacharelado, 2024.
9,2.
Inclui referências, anexos.

1. Eficiência energética. 2. Aprendizado de Máquina. 3. Estatística. 4. Resfriamento. 5. Aquecimento. I. Lopes da Silva, Jeydson. (Orientação). II. Título.

620 CDD (22.ed.)

CAIO SÓTER DE BARROS MOTA

PREVISÃO DE POTÊNCIA NECESSÁRIA PARA RESFRIAMENTO E AQUECIMENTO DE AMBIENTES RESIDENCIAIS

Trabalho de Conclusão de Curso apresentado ao Departamento de Engenharia Elétrica da Universidade Federal de Pernambuco, como requisito parcial para obtenção do grau de Engenheiro Eletricista

Aprovado em: 15/10/2024.

BANCA EXAMINADORA

_____

Prof. Dr. Jeydson Lopes da Silva
Universidade Federal de Pernambuco

_____

Prof. Dr. Calebe Hermann de Oliveira Lima (membro interno)
Universidade Federal de Pernambuco

_____

Eng. M.Sc. Valdemar Moreira Cavalcante Junior (membro externo)
Universidade Federal de Pernambuco

Este trabalho é dedicado a todos que de alguma forma contribuíram para esta construção, para a minha formação como ser humano e eterno aprendiz.

AGRADECIMENTOS

Agradeço primeiramente a meus professores que me mostraram uma forma de pensar a qual não estava habituado, em especial João Gondim, com esta forma, pude estender meus horizontes e aprender a desfrutar do doce sabor do conhecimento. Em seguida, agradeço a minha família e amigos, que sempre me forneceram meios para suprir meus anseios por conhecimento sobre o mundo ao meu redor e por sempre me apoiarem em meus objetivos. Por fim, agradeço a minha companheira Juliana, a qual me fornece uma coisa que os livros nunca puderam e poderão me fornecer, a sensibilidade de ser humano.

Desculpe a carta longa, tivera eu tido mais
tempo, teria escrito uma carta menor.
(BLAISE; PASCAL)

RESUMO

Este trabalho analisou e modelou o problema da potência de resfriamento e aquecimento em estruturas prediais utilizando técnicas estatísticas e Aprendizado de Máquina. O objetivo foi analisar estatisticamente os relacionamentos entre variáveis estruturais dos ambientes prediais e a potência de aquecimento e resfriamento necessárias para cada formato de ambiente, além de construir modelos de Aprendizado de Máquina para fazer previsões acuradas destas variáveis. O trabalho de natureza quantitativa, utilizou método baseado na estrutura de um projeto de Ciência de Dados, onde é composto por coleta de dados, análise exploratória, pré-processamento, ajuste fino e modelagem. Os resultados evidenciaram que é possível modelar, com precisão elevada, as potências com base nas características físicas dos formatos prediais, além de indicar que a modelagem da potência de resfriamento ser mais difícil com relação ao aquecimento. Concluiu-se, portanto, que a utilização de algoritmos de Aprendizado de Máquina, juntamente com ferramentas estatísticas tem grande potencial para resolver problemas dentro da Engenharia Elétrica de maneira precisa e rápida.

Palavras-chave: Eficiência energética; Aprendizado de Máquina; Estatística; Resfriamento; Aquecimento.

ABSTRACT

This work analyzed and modeled the problem of cooling and heating power in building structures using statistical techniques and Machine Learning. The objective was to statistically analyze the relationships between structural variables of building environments and the heating and cooling power required for each environment format, in addition to building Machine Learning models to make accurate predictions of these variables. The quantitative nature work used a method based on the structure of a Data Science project, which is composed of data collection, exploratory analysis, pre-processing, fine-tuning and modeling. The results showed that it is possible to model, with high accuracy, the powers based on the physical characteristics of the building formats, in addition to indicating that the modeling of the cooling power is more difficult in relation to heating. It was concluded, therefore, that the use of Machine Learning algorithms, together with statistical tools, has great potential to solve problems within Electrical Engineering in an accurate and fast manner.

Keywords: Energy Efficiency; Machine Learning; Statistics; Cooling; Heating.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AVAC	Aquecimento, Ventilação e Ar Condicionado
EPE	Empresa de Pesquisa Energética
CART	Classification and Regression Trees
m	metro
m^2	Metro quadrado
kW	quilowatt

LISTA DE SÍMBOLOS

Ω	Espaço amostral
μ	Média populacional
σ	Desvio padrão populacional
s	Desvio padrão amostral
$\bar{x}$	Média amostral

SUMÁRIO

1 INTRODUÇÃO

Com a crescente preocupação em relação à preservação do meio ambiente, a busca por edifícios energeticamente eficientes se tornou uma prioridade. Diversas tecnologias e práticas, como o uso de isolamento térmico e fontes de energia renováveis, oferecem soluções para otimizar o consumo de energia e diminuir o impacto ambiental das construções. Aquecimento e resfriamento são grandes consumidores de energia em prédios. Escolher alternativas mais eficientes, como bombas de calor, aquecimento solar e ar-condicionado com alta classificação energética, automação e dimensionamento adequado de sistemas AVAC, pode diminuir consideravelmente o gasto energético sem atingir o conforto térmico, e contribuir com a diminuição de desperdício de energia.

Uma perspectiva mais local como a informação da Câmara de Comercialização de Energia Elétrica (1), mostra que o Brasil teve um acréscimo de 1,4% no consumo de energia elétrica no primeiro semestre do ano de 2023 em comparação ao mesmo período do ano de 2022. Chegando a uma marca de consumo médio de 66.760 MW. Isto reflete que o país caminha em direção ao crescimento quando se refere a consumo de energia elétrica, dado que há um aumento flutuante da população, industrialização, urbanização e maior acesso a bens de consumo, o que reflete também no desenvolvimento econômico do país. Este acréscimo anual de consumo, deve, portanto, ser estudado para que haja energia suficiente e de qualidade para suprir as necessidades de todos, sem que o ambiente e a própria população estejam à mercê de consequências devastadoras.

Dado a preocupação em diminuir o consumo e melhorar a qualidade da energia gerada, a eficiência energética se mostra uma solução atrativa e de custo relativamente mais baixo em comparação com a integração em larga escala de energias renováveis, que por sua vez possui muitos entraves políticos e econômicos que não torna fácil a implementação. As soluções de eficiência energética podem ser simples como a troca de lâmpadas incandescentes por lâmpadas de LED, até sistemas complexos de automação e utilização de sistemas de Aprendizado de Máquina para previsão e planejamento de consumo.

Estudos indicam que o consumo de energia em construções tem crescido globalmente nas últimas décadas (2, 3), sendo os sistemas de climatização (AVAC)

os principais responsáveis por esse consumo em edifícios (4). Diante disso, uma solução para reduzir a demanda energética é investir em projetos de edifícios mais eficientes, com foco na conservação de energia. Portanto, para projetar edifícios energeticamente eficientes, é fundamental calcular a carga térmica e de refrigeração ou de aquecimento, a fim de dimensionar corretamente os sistemas de climatização e garantir o conforto térmico dos ocupantes. Arquitetos e engenheiros precisam considerar as características do edifício, como o tipo de uso e ocupação, além das condições climáticas, para determinar a capacidade de aquecimento e resfriamento ideal. Cada tipo de edificação, seja residencial ou industrial, possui necessidades específicas que devem ser levadas em conta no projeto.

Finalmente, projetistas utilizam softwares para gerar soluções confiáveis para estimar o impacto do projeto de sistemas em edifícios, entretanto, esta forma vem se mostrado muito exaustiva e requer uma grande especialização dos indivíduos sobre o software. Dado estes problemas, e com o estabelecimento da inteligência artificial no mercado, muitos pesquisadores utilizam técnicas de Aprendizado de máquina para analisar e modelar efeitos dos parâmetros dentro de projetos de edifícios, dado a facilidade de implementação e velocidade de resultado.

1.1 Objetivos

1.1.1 Geral

Analisar dados de características arquitetônicas de vários estabelecimentos simulados pelo Centro de Aprendizado de Máquina e Sistemas Inteligentes localizado em Irvine, Califórnia, para entender o relacionamento entre estas variáveis e o consumo necessário para resfriamento e aquecimento. Além disso, o trabalho tem o objetivo final de desenvolver uma modelagem preditiva sobre as variáveis de consumo. Dessa forma, será proposta uma série de ferramentas estatísticas para análise e tratamento de dados, finalizando com modelagem de uma série de Algoritmos para decisão do modelo que melhor se ajusta ao conjunto de dados proposto.

1.1.2 Específicos

Pretende-se por meio deste trabalho concluir os seguintes objetivos:

- Analisar as variáveis de interesse utilizando ferramentas estatísticas.
- Retirar informações sobre os relacionamentos entre variáveis.
- Identificar o melhor modelo de Aprendizado de Máquina ao conjunto proposto.
- Testar os modelos mais promissores e analisar os resultados.

1.2 Organização do Trabalho

Este trabalho está dividido em cinco capítulos, o primeiro é definido como a introdução e tem a função de contextualizar o assunto escolhido, além de mostrar motivação e objetivos gerais e específicos da escolha deste tema.

O segundo capítulo, tem por finalidade evidenciar as fundações a qual este trabalho se apoia, para garantir que tenha validade científica sob qualquer aspecto. Com isto, conceitos e definições serão apresentados de maneira sucinta para introduzir o leitor as técnicas e informações necessárias.

Em seguida, o terceiro capítulo apresenta a metodologia utilizada para a realização do trabalho, na qual tem o objetivo de relatar o caráter sistemático escolhido para a conclusão do projeto. Além disso, esta etapa consiste na descrição das ferramentas utilizadas durante todo as etapas do projeto.

O quarto capítulo é formado pela análise e desenvolvimento do trabalho, e tem o objetivo de evidenciar os resultados das análises estatísticas e o melhor modelo descrito para o conjunto de dados utilizado.

Finalmente, o último capítulo, a conclusão, é composta da interpretação do resultado obtido, juntamente com proposta de melhorias ou sugestões para continuidade do aperfeiçoamento do projeto.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Eficiência energética

Eficiência energética, em termos formais, caracteriza-se pela utilização otimizada da energia para a realização de determinada função, objetivando a minimização do consumo e a redução de perdas energéticas. Em síntese, consiste em maximizar a produção com o mínimo de recursos energéticos, privilegiando o uso racional e consciente da energia. As estratégias para atingir tal eficiência englobam a adoção de tecnologias inovadoras e de menor consumo, a otimização de processos industriais e a substituição de fontes energéticas tradicionais por fontes renováveis, além da conscientização e mudança nos padrões de consumo da sociedade. A busca pela eficiência energética assume papel crucial na preservação ambiental, na redução de custos e no desenvolvimento econômico sustentável.

2.1.1 Consumo residencial

Segundo os dados da Empresa de Pesquisa Energética (EPE) (5), o consumo da energia elétrica das residências foi de 13.311GWh em janeiro de 2023, um aumento de 1,8% em comparação ao mesmo mês do ano de 2022. A Figura 1 evidencia que aumento do consumo residencial vem se tornado uma parte importante de análise, em consequência do aumento populacional, no qual reflete em sua partição de 30,1% dentro da divisão por consumo de classes de 2021 divulgado pela EPE (6) e que se enquadra como o segundo maior consumo, atrás apenas da Indústria. Além disso, a Figura 1 mostra, também, que o número de consumidores por classe é em sua maioria Residencial, com um número absoluto de 75,2 milhões de consumidores.

Figura 1 – Partição por classe da energia consumida em 2021.

Fonte: Retirado de (6).

2.1.2 Conforto térmico

Os seres humanos passam a maior do tempo em ambientes fechados, como em suas residências, no trabalho ou em estabelecimentos educacionais. A temperatura nesses locais não só afeta o consumo de energia, mas também influencia diretamente nossa qualidade de vida, produtividade, saúde e bem-estar. Por isso, é importante considerar as condições térmicas ao projetar ou reformar edifícios, visando criar espaços mais confortáveis e eficientes para seus ocupantes.

O conceito de conforto térmico é bem definido em (7), como o grau de satisfação de uma pessoa com a temperatura ao seu redor, sem desejar que esteja mais quente ou mais frio. É um estado de bem-estar físico e mental em relação ao ambiente térmico e que é característico de cada indivíduo, além de ser também influenciado pelas variáveis arquitetônicas do local em que se encontra.

2.1.3 Dimensionamento de Carga térmica

A eficiência de um equipamento de climatização depende da quantidade de calor que ele precisa remover ou adicionar ao ambiente, em relação à energia elétrica consumida para realizar essa tarefa. Segundo (8), um dimensionamento correto é fundamental, pois sistemas inadequados podem reduzir a eficiência energética em até

20% e diminuir a vida útil do equipamento em mais de 50%, comparado aos valores padrão do fabricante.

O cálculo da carga térmica em ambientes internos é utilizado para dimensionar sistemas de climatização em edifícios. Modelos numéricos precisos devem reproduzir detalhadamente os fenômenos de transferência de calor, considerando o balanço energético resultante das trocas de calor na envoltória, do uso de equipamentos e da presença de ocupantes (8). Essas trocas térmicas são influenciadas pelas características das superfícies e suas interações com o ambiente.

Em meio a problemas de dimensionamento, torna-se clara a importância de pesquisas que viabilizem soluções sustentáveis e energeticamente eficientes para a climatização dentro de edificações, considerando os requisitos fundamentais para o correto dimensionamento desses sistemas.

2.1.3.1 Norma NBR-16401-1

O cálculo da carga térmica, conforme a norma ABNT NBR 16401-1:2008, é um processo essencial para o dimensionamento adequado de sistemas de ar-condicionado. Ele visa determinar a quantidade de calor que precisa ser removida, no caso de resfriamento, ou adicionada, no caso de aquecimento, para manter as condições de conforto térmico desejadas em um ambiente.

A norma aborda a abrangência e a metodologia do cálculo, enfatizando a necessidade de considerar tanto as cargas térmicas internas geradas por ocupantes, equipamentos e iluminação, além das cargas térmicas externas como transmissão de calor através da envoltória do edifício. Em cima disto, o zoneamento do ambiente é importante para identificar áreas com diferentes perfis de carga térmica, permitindo um controle mais preciso da climatização.

Por fim, o cálculo da carga térmica interna dos recintos envolve a análise da envoltória do edifício (paredes, teto, piso, janelas) e das fontes internas de calor e umidade. A norma fornece orientações para calcular a transmissão de calor através dos elementos da envoltória, considerando suas propriedades térmicas e as condições climáticas externas.

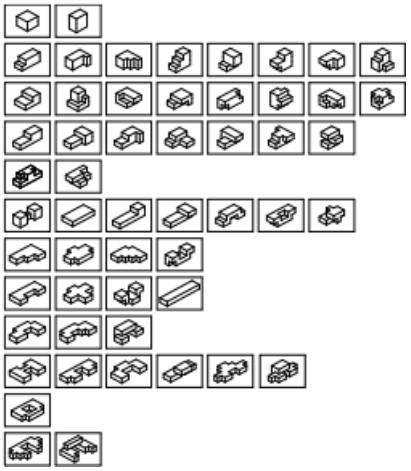
2.2 Dados

Os dados utilizados foram construídos pelo Centro de Aprendizado de Máquina e Sistemas Inteligentes em Irvine, na California e disponibilizados através do repositório UC Irvine Machine Learning (9), em novembro de 2012.

As estruturas prediais foram construídas a partir de 18 cubos com lados de 3,5 m de mesmo material e volume. Os dezoito cubos foram utilizados para simular 12 formatos diferentes de estruturas com volumes iguais de 771.75 m^2 pela instituição através do software Ecotect, na qual é uma ferramenta que tem o objetivo de modelar estruturas arquitetônicas 3D e permitir uma gama de estudos como análise térmica, energética, iluminação, sombreamento e análise de custo.

Um total de 768 estruturas foram geradas, cada uma com suas próprias dimensões e valores estimados de cargas de aquecimento e resfriamento. Informações, mais detalhadas, sobre as estruturas como localização residencial, característica de construção, número de residentes e entre outras características assumidas, estão no artigo (10). Como adição, na Figura 2, pode ser visto um exemplo de como os cubos podem ser modelados para produzir formas diferentes.

Figura 2 – Exemplo de 54 formatos de estrutura.



Fonte: Retirado de (11).

Por fim oito variáveis foram utilizadas e, na Tabela 1, se encontra uma breve descrição que mostra o nome, a dimensão e tipologia dos campos utilizados na modelagem e análise:

Tabela 1 – Descrição dos campos.

Campos	Dimensão	Tipologia
Compactação Relativa	-	Numérico contínuo
Área de Superfície	m²	Numérico contínuo
Área de Parede	m²	Numérico contínuo
Área do Telhado	m²	Numérico contínuo
Altura global	m	Numérico contínuo
Orientação	-	Categórico Nominal
Área de envidraçamento	% de área de chão	Numérico Contínuo
Distribuição da área de envidraçamento	-	Categórico Nominal
Carga de Aquecimento	kW	Numérico contínuo
Carga de Resfriamento	kW	Numérico contínuo

Fonte: O autor.

Em adição a Tabela 1, temos as definições das tipologias para melhor entendimento:

- **Numérico contínuo:** Variáveis com tipologia Numérico contínuo nas quais estão associadas a conjuntos infinitos não contáveis.
- **Categórico Nominal:** Variáveis com tipologia Categórico Nominal assumem valores categóricos que não possuem ordem entre si.

Por fim, uma breve descrição da Tabela 1, o dicionário em seguida traz mais a fundo a definição de cada variável para o melhor entendimento das análises.

- **Compactação relativa:** A compactação relativa é um indicador de quão compacta é uma forma fechada, o que significa que quanto mais compacto for o edifício, menor será o espaço vazio em seu interior que precisa ser aquecido ou resfriado.
- **Área de superfície:** Quantidade total de espaço ocupada pelas superfícies de cada estrutura predial.

- **Área de parede:** Quantidade total de espaço coberto por paredes de cada estrutura predial.
- **Área do telhado:** Quantidade total de espaço composta pelo telhado de cada estrutura predial.
- **Altura global:** É a altura medida do ponto da terra até o ponto mais alto da estrutura.
- **Orientação:** É a direção na qual a estrutura está em face. Este campo é classificado em quatro grupos: Norte, Sul, Leste e Oeste.
- **Área de envidraçamento:** É a percentagem de área do piso na qual é coberta por material envidraçado.
- **Distribuição da área de envidraçamento:** A distribuição da área de envidraçamento é a categorizada em seis grupos que refletem como a área envidraçada é espalhada de acordo com a Orientação de cada estrutura predial. Os grupos são, Uniforme (1), Norte (2), Leste (3), Sul (4), Oeste (5) e sem área envidraçada (0).
    - Uniforme: É definido com 25% de área envidraçada em cada face da estrutura.
    - Norte: É definido com 55% da área envidraçada na face Norte e 15% em cada outra face.
    - Leste: É definido com 55% da área envidraçada na face Leste e 15% em cada outra face.
    - Sul: É definido com 55% da área envidraçada na face Sul e 15% em cada outra face.
    - Oeste: É definido com 55% da área envidraçada na face Oeste e 15% em cada outra face.
- **Carga de aquecimento:** Carga estimada através do software Ecotect para aquecer a estrutura com base nos parâmetros citados anteriormente.
- **Carga de resfriamento:** Carga estimada através do software Ecotect para resfriar a estrutura com base nos parâmetros citados anteriormente.

2.3 Análise Exploratória de Dados

Na etapa de Análise exploratória de dados será abordada as ferramentas estatísticas utilizadas, assim como as bibliotecas as quais estas ferramentas estão implementadas.

2.3.1 Estatística Descritiva

A estatística descritiva tem como objetivo resumir a série de dados para ser possível uma interpretação mais direta do conjunto e, portanto, facilitar explicações e tomadas de decisões. Entre as várias medidas existentes neste ramo da estatística, é possível dividi-las em dois grupos, as Medidas de Posição e Medidas de Dispersão.

2.3.1.1 Medidas de Posição

As medidas de posição são muito utilizadas em análise de dados pois é extremamente útil para resumir séries inteiras com relação a magnitude de seus possíveis valores. Este grupo engloba a Média Aritmética, Mediana e a Moda.

Segundo (12), a Moda é definida como o registro mais frequente do conjunto de valores observados. Como exemplo, temos que dado três lançamentos de um dado balanceado, obtemos no conjunto de resultados o número um com duas vezes e o número quatro apenas uma vez, logo a moda para este conjunto é um.

A definição de mediana segundo (12) é o valor que ocupa a posição central de uma série de observações, dado sua ordenação crescente. Quando o número de observações for par, então a mediana será calculada como a média aritmética entre as duas observações centrais. Finalmente, é possível definir a mediana com o rigor matemático, considerando primeiramente um conjunto de dados em ordem crescente como:

$$x_1 \leq x_2 \leq \cdots \leq x_n \tag{1}$$

Segundo (12), define-se mediana da variável aleatória X, como:

$$md(X) = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & se\ n\ \text{é}\ ímpar \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & se\ n\ \text{é}\ par \end{cases} \quad (2)$$

A última medida, a média aritmética, segundo (12) é definida como a soma de todas observações dividida pelo número de observações. No exemplo anterior a média aritmética será 2, pois a soma das observações é 6 = (1 + 1 + 4) e será dividido por 3 que é o número de jogadas. Com isso, é possível definir a média aritmética em um caráter mais formal como:

$$\bar{x} = \frac{\sum_{i=0}^{n} x_i}{n} \quad (3)$$

Por fim, é importante salientar que a média aritmética é sensível a valores extremos, o que pode fazer com que esta medida não seja a mais indicada para resumir o conjunto de dados.

2.3.1.2 Medidas de Dispersão

É introduzido, dentro da estatística descritiva, as medidas de dispersão como ferramentas matemáticas que quantificam a variabilidade ou dispersão de uma série de dados. Assim como as medidas de posição, as medidas de dispersão têm um importante papel no resumo de informações para interpretação e tomada de decisão. Dentro desta categórica, pode-se incluir a variância e o desvio padrão, que são medidas muito comuns na análise estatística.

A variância pode ser interpretada como a quantificação da dispersão de um conjunto de dados ao redor da média (12). Dito isto, pode-se defini-la matematicamente como:

$$\mathrm{var(X)} = \frac{\sum_{i=0}^{n}(x_i - \bar{x})^2}{n} \quad (4)$$

A Equação 4 evidencia que a dispersão ao redor da média é elevada ao quadrado para evitar cancelamento de termos, dado que o desvio pode ser negativo

ou positivo. Finalmente, divide-se pelo número de observações para facilitar o entendimento da quantidade.

Como a variância eleva os desvios ao quadrado, as unidades de medidas associadas serão também elevadas ao quadrado, fazendo com que seja mais inconveniente a interpretação do resultado. Dito isto, é conveniente definir o Desvio Padrão como sendo a raiz quadrada da variância, para facilitar a interpretação da medida de dispersão. A relação da variância e desvio padrão pode ser exemplificado através de:

$$\mathrm{dp(X)} = \sqrt{\mathrm{var(X)}} = \sqrt{\frac{\sum_{i=0}^{n}(x_i - \bar{x})^2}{n}} \qquad (5)$$

2.3.1.3 Quantis e Percentis

O trabalho exige a definição de medidas que vão além das quantificações de posição e dispersão, visto que estas medidas não são suficientes para informar sobre o grau de simetria do conjunto de dados. Portanto, a definição de (12) explica que um quantil *q*, relacionado a um percentil *p* divide *100*p%* dos dados ordenados de forma crescente, a partir da quantidade *q*. Desta forma o valor *q* estará acima de 100*p% dos dados e abaixo de *(1-p) *100%*.

Os quantis mais comuns dentro da análise de dados são o 1° Quartil, 2° Quartil ou mediana e o 3° Quartil. Estes quartis estão à frente de respectivamente 25%, 50% e 75% dos dados e são muito utilizados pois não são sensíveis a valores extremos, como a média aritmética, além de serem utilizados como indicador de assimetria. A definição matemática é enunciada como:

$$\begin{cases} q(0{,}25) = q_1 \\ q(0{,}50) = q_2 \\ q(0{,}75) = q_3 \end{cases} \qquad (5)$$

Como estas medidas são consideradas resistentes, ou seja, não são sensíveis a uma pequena parcela do conjunto de dados, pode-se definir uma nova medida de dispersão chamada por (12) de distância interquartil ou intervalo interquartil, definida matematicamente como:

$$d_q = q_3 - q_1 \tag{6}$$

Além disso é importante definir duas variáveis matemáticas que representam o primeiro elemento e o último elemento da sequência de observações. Para isto, diz-se que x_1é o primeiro elemento da sequência, enquanto que x_n é o último elemento. Com estes parâmetros é possível ter informação sobre a assimetria do conjunto, pois segundo (12), distribuições aproximadamente simétricas devem estar de acordo com:

$$\begin{cases} q_2 - x_1 \cong x_n - q_2 \\ q_2 - q_1 \cong q_3 - q_2 \\ q_1 - x_1 \cong x_n - q_3 \end{cases} \tag{7}$$

Finalmente, as distâncias entre a mediana e q_1, q_2 devem ser menores do que as distâncias entre q_1 e q_3 com os valores extremos da distribuição.

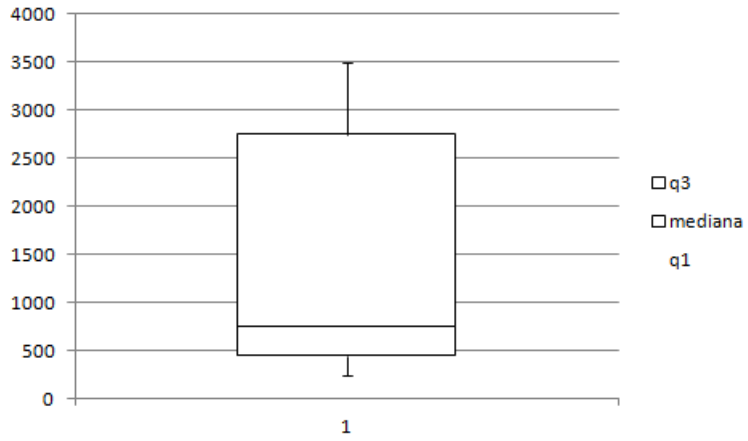
2.3.1.4 Box Plots

Os box plots, nomeados pelo estatístico norte-americano John Wilder Turkey, box plots foram introduzidos no meio acadêmico por diversas pessoas relacionadas a ele nos anos 70 em (13, 14, 15).

O Box plot visto na Figura 3, é um gráfico que mistura conceitos da estatística descritiva para resumir em imagem várias informações úteis. É utilizado como ferramenta para verificação de assimetria, além de ser bastante útil para verificar valores extremos.

Sua construção é feita, primeiramente, estabelecendo um retângulo contendo o primeiro, segundo e terceiro quartis. A partir desta figura, calcula-se o Intervalo Interquartil já definido anteriormente, e constrói linhas partindo do primeiro e terceiro quartis até o dado que não supera, respectivamente, $li = q_1 - 1{,}5 * d_q$ e $ls = q_3 - 1{,}5 * d_q$, nomeados como limite inferior e limite superior. Os valores que superarem estes limites são considerados valores extremos, e poderão ser *Outliers* ou não.

Figura 3 – Exemplar de Box Plot.

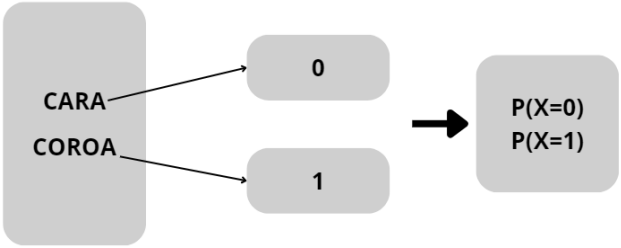


Fonte: Retirado de (16).

2.3.2 Distribuições de Probabilidade

Distribuições são padrões, traduzidos em gráficos, de incerteza associados a variáveis discretas ou contínuas, chamadas de variáveis aleatórias. Estas são assim chamadas pois não possuem um valor determinado, mas podem assumir diversos valores de acordo com natureza estocástica do que está sendo modelado. Em outras palavras, dado um conjunto enumerável chamado de espaço amostral Ω, uma variável aleatória pode ser vista como uma função que associa cada elemento do conjunto de resultados a um número (12), que por sua vez estará relacionado com um conjunto de probabilidades $P(X = x_i)$ como está mostrada na Figura 4.

Figura 4 – Exemplo de Variável aleatória discreta.



Fonte: O autor.

As variáveis aleatórias podem ser divididas em dois grupos, discretas ou contínuas. As primeiras, são assim chamadas pois só podem assumir uma quantidade finita de valores ou valores infinitos enumeráveis, enquanto que as segundas podem

assumir valores pertencentes a um intervalo na reta real (12). Por fim, define-se uma variável aleatória por letras maiúsculas, como *X* ou *Y*.

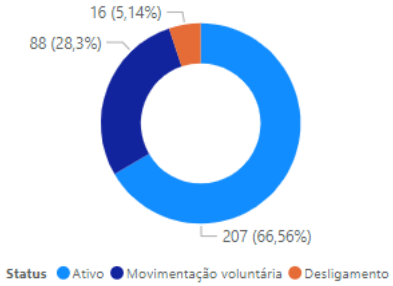
2.3.2.1 Distribuição de frequências

Os gráficos e frequência foram primeiramente introduzidos pelo economista político escocês William Playfair em seu trabalho intitulado *"The comercial and Political Atlas (London 1786)"* (17), onde, de acordo com (18) foi utilizado gráficos de barra para mostrar a importação e exportação entre a Escócia e outros países. Com isto, pode-se entender que deste muito tempo, a análise visual é uma importante fonte de conhecimento para os dados e para a tomada de decisão.

A análise de frequência é uma maneira rápida e eficiente de informar sobre a distribuição de uma variável aleatória. Dentro deste domínio, é possível representar campos categóricos e numéricos utilizando figuras geométricas e eixos cartesianos. Dito isto, possível categorizar dois tipos de gráficos para dois tipos de campos, os gráficos de barras para variáveis categóricas e os histogramas para variáveis numéricas.

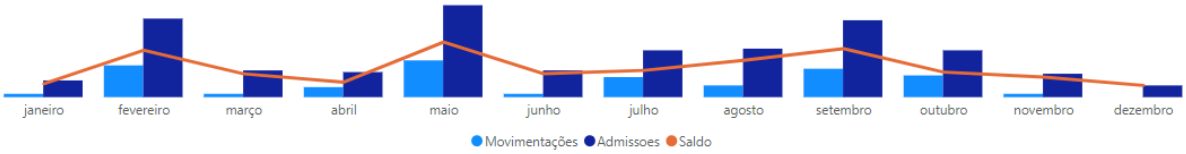
Os gráficos de barras, são estruturas retangulares contidas em dois eixos perpendiculares, e têm o objetivo de agrupar dados das categorias de uma variável qualitativa. Segundo (12), gráficos de barras e também os gráficos de setores são utilizados para representar a frequência destes grupos, em números absolutos ou percentuais. Nas Figuras 5 e 6, é possível visualizar como estes gráficos cumprem seu papel de informar o leitor com relação a frequência das observações.

Figura 5 – Exemplo de gráfico de setores.



Fonte: O autor.

Figura 6 – Exemplo de gráfico de barras vertical.



Fonte: O autor.

Em seguida, para variáveis numéricas, é interessante muitas vezes resumir a frequência através de um gráfico denominado de histograma. Estes gráficos são gráficos de barras verticais contíguas, que compactam a informação em classes para descrever a densidade de frequência de cada uma das classes criadas, sem que haja muita perda de informação, principalmente quando se trata uma variável contínua.

As classes são criadas de maneira que o histograma nem fique tão compacto a ponto de perder informação, nem fique tão discretizado a ponto de não ser possível tirar nenhuma conclusão resumida dos dados. Em seguida, frequência de cada classe é verificada através da área de cada barra e pode ser calculada pelo alcance de uma classe arbitrária multiplicada pela densidade de frequência, ou seja, a altura da barra (12). Como exemplo segue a Tabela 2, que resume a variável contínua Salário dos empregados de um conjunto de dados de 36 observações, retirado do livro (12).

Tabela 2 – Tabela de frequência da variável Salário dos empregados.

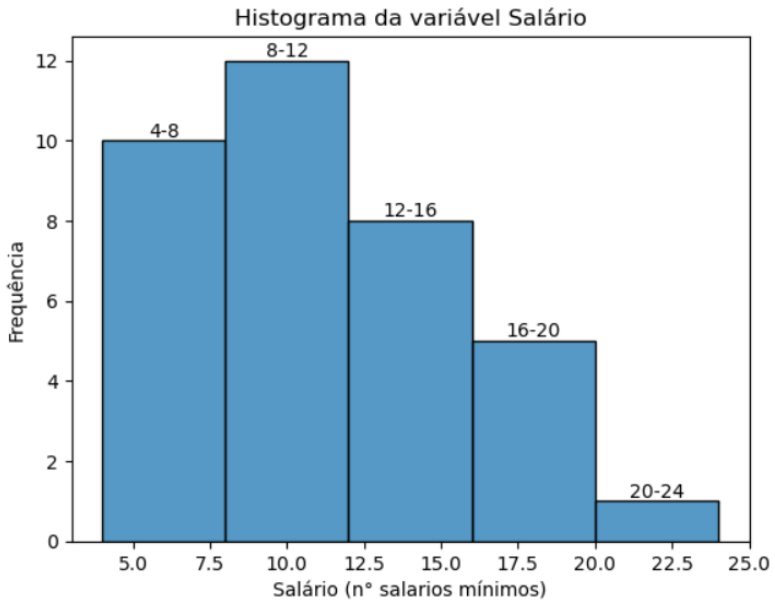
Classes de salários (n° de salários mínimos)	Ponto médio	Frequência (n_i)	Porcentagem % (f_i)
[4,00 – 8,00)	6,00	10	27,78
[8,00 – 12,00)	10,00	12	33,33
[12,00 – 16,00)	14,00	8	22,22
[16,00 - 20,00)	18,00	5	13,89
[20,00 - 24,00)	22,00	1	2,78
Total	-	36	100,00

Fonte: Retirado de (12).

Com a Tabela 2, é possível construir o histograma da seguinte forma, primeiro calcula-se a amplitude de cada classe, para representa uma classe arbitrária representa-se a amplitude da i-ésima classe por Δ_i. Em seguida, para que a área de cada retângulo seja a frequência da respectiva classe, a sua altura deve ser a

densidade de frequência, que será representado pela i-ésima classe como $\frac{f_i}{\Delta_i}$ ou $\frac{n_i}{\Delta_i}$. Portanto, com estas informações é possível construir o histograma da Figura 7 que foi adaptado do livro (12).

Figura 7 – Histograma.



Fonte: Retirado de (12).

2.3.2.2 Distribuição Normal

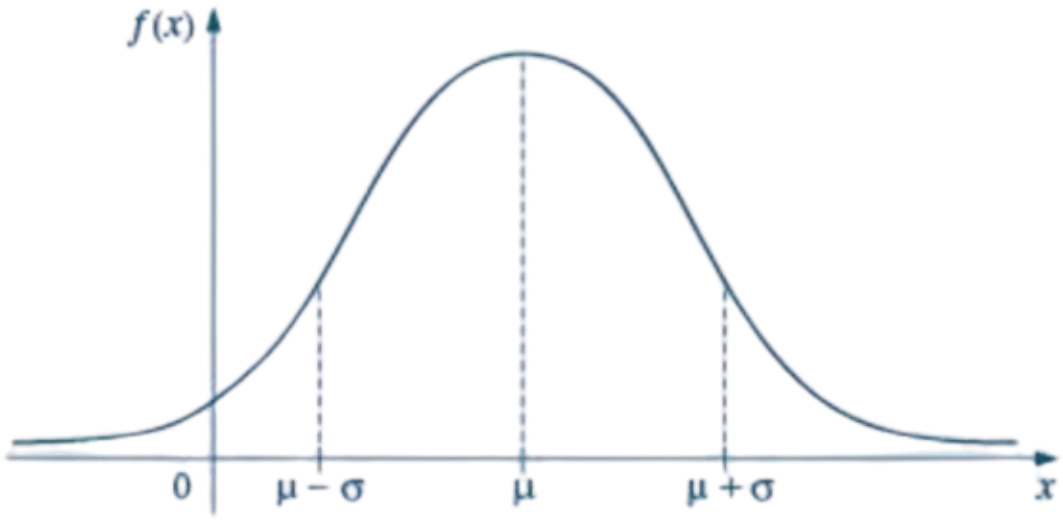
A distribuição normal, foi introduzida por De Moire como uma aproximação da distribuição discreta binomial (19). Sua utilização foi então disseminada com Laplace em 1783 para estudar erros de medição e por Gauss em 1809, quando estudava dados astronômicos, evidenciado em (20).

Define-se formalmente da seguinte maneira, dado uma variável aleatória contínua *X* com parâmetros μ denominada de média populacional, e σ^2 chamada de variância populacional, diz-se que segue uma distribuição normal se a função densidade de probabilidade é da forma:

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty \qquad (8)$$

É evidenciado em (12) que esta função é referida como $X \sim N(\mu, \sigma^2)$ e de que é simétrica com relação a reta $x = \mu$ e, portanto, respeitando $f(\mu + x; \mu, \sigma^2) = f(\mu - x; \mu, \sigma^2)$. Na Figura 8, é possível visualizar uma função densidade de probabilidade normal arbitrária.

Figura 8 – F.d.p normal de parâmetros μ e σ^2.

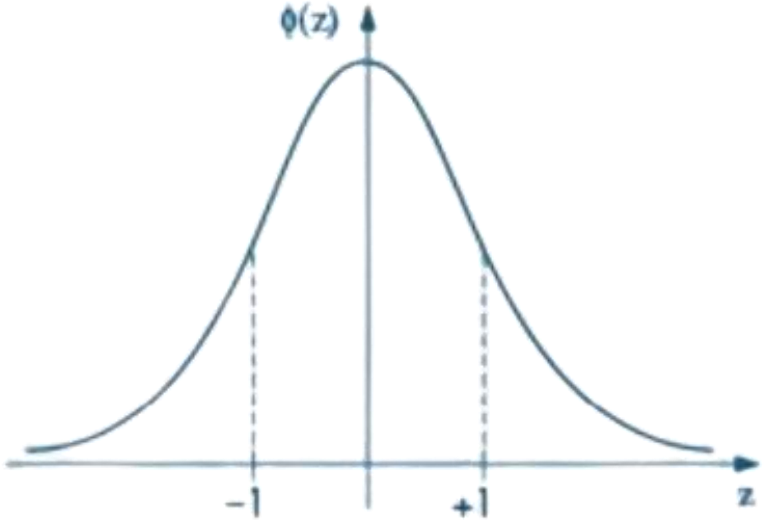


Fonte: Retirado de (12).

Como esta função é analiticamente impossível de calcular as probabilidades, apenas numericamente possível, é necessário padronizar uma função, caso contrário cada combinação de parâmetros exigiria o recalculo das probabilidades. Portanto, diz-se que uma função normal padrão é àquela cujos parâmetros μ e σ^2 são respectivamente iguais a zero e um, logo sendo referenciada como $Z \sim N(1,\ 0)$. A função normal padrão é definida como:

$$\varphi(z; 1, 0) = \frac{1}{\sqrt{2\pi}} e^{\frac{-(z)^2}{2}}, -\infty < z < \infty \qquad (9)$$

Figura 9 – F.d.p normal padrão.



Fonte: Retirado de (12).

Por fim, é importante definir a transformação de uma variável aleatória $X \sim N(\mu, \sigma^2)$ que segue uma distribuição normal qualquer, para uma variável aleatória $Z \sim N(1,\ 0)$ seguindo uma distribuição normal padrão, é por sua vez, segundo (12) da forma:

$$Z = \frac{X - \mu}{\sigma} \qquad (10)$$

2.3.3 Correlações

A história da correlação remete ao físico Francês Auguste Bravais, ao qual foi referida a fala “une correlation”, enquanto trabalhava com a distribuição normal bivariada no final de 1846 (21). Após isto, Francis Galton em seu livro *Natural Inheritance*, de 1889, estendeu e propôs o conceito de correlação (22), enquanto que apenas em 1895, o famoso estatístico Britânico Karl Pearson, definiu a teoria como o “produto-momento” e o relacionamento com a regressão linear. (21)

A análise de correlações pode ser feita utilizando diferentes formas. As mais utilizadas dentro da ciência de dados são, Correlação de Pearson, Correlação de Spearman e Correlação de Kendall. Esta medida, independente da metodologia

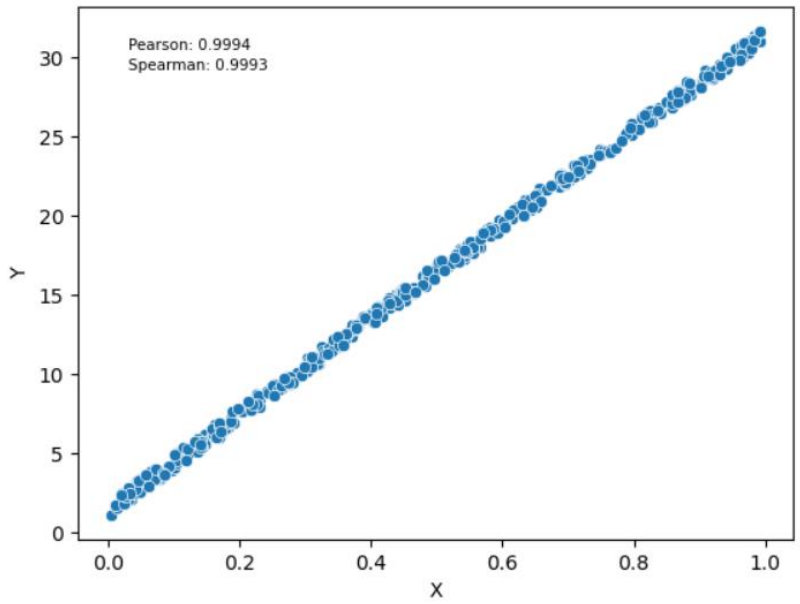
abordada, tem a finalidade de medir direção e força de relacionamento, independente de causalidade, entre variáveis aleatórias o que por sua vez é uma importante etapa, para exploração de variáveis, seleção e modelagem preditiva. Neste trabalho, foi utilizada apenas a Correlação de Spearman, portanto, será a única a qual haverá fundamentação teórica.

2.3.3.1 Correlação de Spearman

A correlação de Spearman é uma abordagem não paramétrica e livre de distribuição, ou seja, não assume normalidade dos dados para quantificar o relacionamento monotônico e arbitrário entre variáveis ranqueadas ou apenas uma variável ranqueada e outra não ranqueada (23). Este coeficiente varia entre -1 e 1, e é um caso do Coeficiente de Correlação de Pearson sobre condições de que os dados estão convertidos em categorias ou ranques antes de qualquer cálculo.

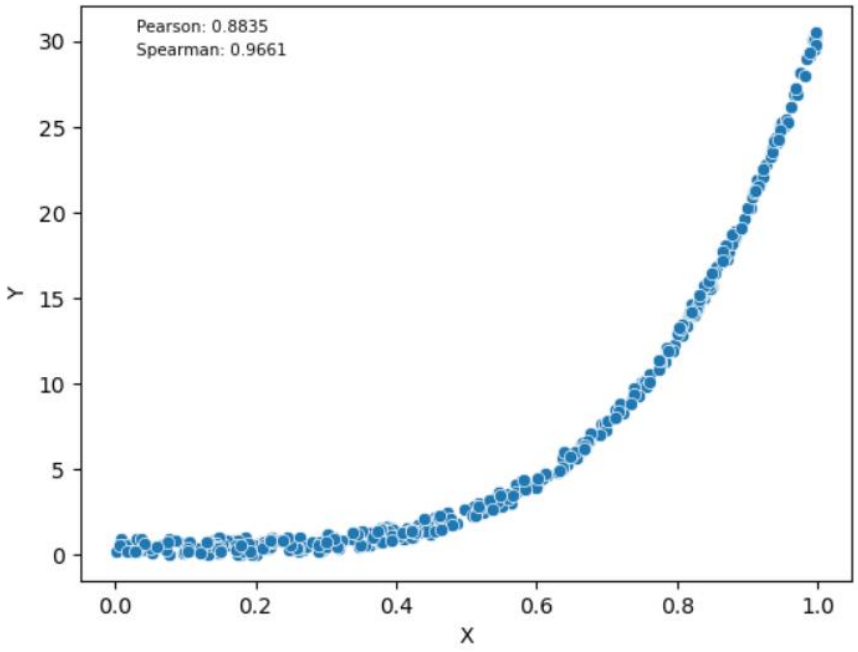
Define-se como relacionamento Monotônico, o tipo de relacionamento que abrange taxa de variação mutável unidirecional entre variáveis aleatórias. Em outras palavras, um conjunto de dados que possui um relacionamento claramente linear como na Figura 10, possui uma taxa de mudança constante e unidirecional (24), enquanto que relacionamentos não lineares como na Figura 11, não possuem este tipo de comportamento, logo pode-se dizer que, correlações baseadas em relações monotônicas são mais abrangentes do que, por exemplo, correlações baseadas em relações lineares como a Correlação de Pearson.

Figura 10 – Relacionamento Monotônico Linear.



Fonte: O autor.

Figura 11 – Relacionamento Monotônico não Linear.



Fonte: O autor.

O Coeficiente de Spearman é calculado, primeiramente, ranqueando os dados de maneira que números maiores recebem um index crescente, ou seja, dado uma lista de dados da forma $[14, 2, 10, 0]$, é construída uma nova lista contendo o valor dos ranques em ordem de posição, $[4, 2, 3, 1]$. Entretanto, vale salientar que, caso haja empate de números na lista original, o ranque será calculado como a média aritmética do ranque do número repetido com o próximo ranque até que não haja número repetido, ou seja se a lista for $[14, 2, 2, 0]$, a lista de ranques será $[4, 2{,}5, 2{,}5, 1]$, portanto 2,5 sendo a média aritmética do ranque 2 com o 3. Em seguida, utilizando os valores ranqueados para das variáveis, calcula-se o coeficiente de Spearman simplesmente utilizando a teoria do Coeficiente de Correlação de Pearson como é evidenciando como:

$$r_s = \rho_{X_r,Y_r} = \frac{COV(X_r, Y_r)}{STD(X_r) * STD(Y_r)} \tag{11}$$

onde $COV(X_r, Y_r)$ significa a covariância entre as variáveis ranqueadas, e é definida na Equação 12, enquanto que $STD(X_r) * STD(Y_r)$ é o produto dos desvios padrões.

$$COV(X_r, Y_r) = \frac{\sum_{i=0}^{n} \left(x_{r_i} - \overline{x_r}\right) * (y_{r_i} - \overline{y_r})}{n - 1} \tag{12}$$

2.4 Transformações

Visto que, dados precisam em geral ser remodelados para serem consumidos por modelos, é interessante analisar o tipo das variáveis para aplicar as transformações adequadas. As transformações podem ser voltadas para variáveis numéricas ou variáveis categóricas.

2.4.1 Transformações de dados numéricos

Transformações de dados numéricos, como a Normalização e Padronização, são bastante utilizadas para enquadrar os dados em um alcance definido e, portanto,

não permitir que variáveis com escalas muito grandes se sobressaiam na modelagem em comparação com variáveis de escalar pequena. Além disso, alguns modelos podem sofrer dificuldades quando variáveis estão em escalas diferentes ou para fazer com que sua distribuição seja mais próxima da distribuição normal.

2.4.1.1 Normalização

Diz-se que uma variável aleatória está normalizada quando os seus dados se enquadram em um alcance entre 0 e 1 e requer que o valor mínimo e valor máximo da variável transformada esteja disponível (25), dado que sua transformação é dada por:

$$y = \frac{x - x_{min}}{(x_{max} - x_{min})} \tag{13}$$

Além disso, é importante destacar que este tipo de transformação não lida com valores extremos e nem modifica a distribuição original da variável, mas apenas muda a sua escala.

2.4.1.2 Padronização

Uma variável aleatória arbitrária está padronizada quando cada elemento é subtraído da média amostral e dividido pelo desvio padrão amostral, como pode ser visto na Equação 14, para que a nova variável tenha média aproximadamente zero e desvio padrão aproximadamente um. Esta transformação centraliza e o formato da distribuição, mas não muda a natureza da mesma. Por fim a padronização assume que os dados seguem uma distribuição normal, apesar de ser possível ser aplicado mesmo quando os dados não seguem esta distribuição, entretanto podendo gerar resultados não confiáveis (25).

$$y = \frac{x - \bar{x}}{s} \tag{14}$$

2.4.2 Transformações de dados categóricos

Modelos de Aprendizado de máquina geralmente não aceitam dados em formato de texto que em geral ocorre por restrições dos algoritmos serem eficientes [26], portanto, é necessário transformar os dados categóricos em números para que seja feito o treinamento e a inferência. Há muitas formas de fazer a codificação de dados categóricos para números, e neste trabalho será abordado o método chamado *One Hot Encoding*.

2.4.2.1 One-Hot Encoding

Variáveis categóricas podem ser divididas em dois grupos, variáveis categóricas ordinais e nominais (26). O primeiro grupo assume que as categóricas possuem um relacionamento de ordem entre si, enquanto que o segundo grupo não possui este tipo de relacionamento. Para ambos os grupos há uma forma específica de se transformar. O *One-Hot Encoding*, por sua vez, está voltado para a transformação de variáveis categóricas nominais, visto que cria uma nova variável binária para cada categórica contida na variável transformada, ou seja, dado uma variável aleatória que representa a cor de carros definida, por:

$$X = \{Verde, Preto, Preto, Branco\} \tag{15}$$

A transformação por *One-Hot Encoding* é do formato visto na Figura 11, no qual o número 0 representa que a observação não é da classe especificado e um que é da classe especificada.

Figura 12 – *One-Hot Encoding* da variável Cor.

	Cor_Branco	Cor_Preto	Cor_Verde
0	0	0	1
1	0	1	0
2	0	1	0
3	1	0	0

Fonte: O autor.

2.5 Modelos

A modelagem estatística é, há muito tempo, uma ferramenta para construir modelos que preveem ou estimam variáveis de natureza aleatória. Com o avanço da tecnologia, surgiu o aprendizado de máquina, um método de análise de dados que automatiza a construção de modelos analíticos. Este ramo da inteligência artificial baseia-se na ideia de que sistemas podem aprender com dados, identificar padrões e tomar decisões com o mínimo de intervenção humana.

Finalmente, em essência, o aprendizado de máquina aprimora a modelagem estatística tradicional, aproveitando o poder computacional para analisar grandes volumes de dados e descobrir padrões complexos que seriam difíceis de identificar manualmente.

2.5.1 Floresta Aleatória

Para o entendimento do modelo de Floresta Aleatória faz-se necessário o entendimento do modelo de Árvore de Decisão, dado que o último nada mais é do que uma aglomeração de Árvores de Decisão na qual são alimentados com subconjuntos aleatórios do conjunto de dados completo e, por fim, aplicando algum tipo de agregação para gerar o resultado final.

2.5.1.1 Algoritmo de treinamento CART

O algoritmo CART (*Classification and Regression Tree*) é um algoritmo ganancioso que busca a melhor divisão de um conjunto de dados e repete recursivamente a mesma lógica. É utilizado em pacotes importantes dentro da análise de dados como no *Scikit-Learn* para treinamento de modelos de Árvore Aleatória para Classificação ou Regressão. Seu funcionamento para Regressão, segundo o livro (27), dá início a partir da divisão do conjunto de treinamento em subconjuntos utilizando um único campo x e um limiar t_x que produzem o subconjunto ponderado pelo número de observações, com base na equação de custo que o algoritmo tem o objetivo de minimizar, como:

$$J(x, t_x) = \frac{n_{esquerda} * MSE_{esquerda}}{n} + \frac{n_{direita} * MSE_{direita}}{n} \tag{16}$$

onde $MSE_{esquerda/direita}$ mede o erro médio quadrático, que pode ser visto na Equação 17, de ambos os subconjuntos, $n_{esquerda/direita}$ significam o número de observações de cada subconjunto, enquanto que n significa o número de observações antes de ser dividido.

Em seguida, o algoritmo continua a dividir o conjunto de treino em subconjuntos recursivamente, utilizando a mesma lógica, até que atinja o valor do hiper parâmetro de profundidade máxima, ou até não conseguir encontrar alguma divisão que diminua o erro médio quadrático do subconjunto.

$$MSE_{nó} = \sum_{i \in nó} (\hat{y}_{nó} - y^{(i)})^2 \tag{17}$$

onde $\hat{y}_{nó}$ é dado pela Equação 18 e $y^{(i)}$ é definido como o valor da variável resposta para a i-ésima observação do nó.

$$\hat{y}_{nó} = \frac{1}{n_{nó}} * \sum_{i \in nó} y^{(i)} \tag{18}$$

Finalmente, em cada nó, o valor estimado da variável resposta é calculado através da Equação 18, até que chegue ao final da árvore, onde será feito a inferência final das observações que se encontrão dentro deles.

2.5.1.2 Combinação de Árvores de Decisão

Um conjunto de modelos Árvores de Decisão é chamado de Floresta Aleatória. Este algoritmo adiciona aleatoriedade no treinamento, pois ao invés de procurar pela melhor variável que dividirá o conjunto em um determinado nó, ele procura pela melhor variável entre um subconjunto de variáveis. Portanto, resultando em uma melhor diversidade de árvores e evitando, assim, o sobreajuste.

O funcionamento do algoritmo, em primeira instância é feito retirando subconjuntos com reposição do conjunto de treinamento. Cada subconjunto, portanto, treinará um estimador de Árvore de Decisão e, portanto, no momento da inferência para um problema de Regressão, o resultado de todos os estimadores é agregado utilizando, geralmente, a média aritmética e gerar a inferência final. O nome deste processo é chamado de *Bagging* por (27).

2.5.2 Extreme gradient Boosting

O algoritmo *Extreme Gradient Boosting* é um modelo baseado em árvores para Classificação ou Regressão e, também, uma versão otimizada do algoritmo *Gradient Boosting*, que tem como objetivo ser mais rápido, portátil e escalável segundo (27).

2.5.2.1 Boosting

A definição de *Boosting* segundo (27) refere-se a um modelo combinado que aglomera vários estimadores fracos, ou seja, que têm baixa capacidade preditiva, para construir um estimador forte. Isso geralmente é feito treinando cada estimador de maneira a corrigir o estimador treinado anteriormente.

2.5.2.2 Gradient Boosting

O algoritmo Gradient Boosting é um modelo que geralmente utiliza Árvores de Decisões como estimadores fundamentas. O seu funcionamento se dá sequencialmente adicionando estimadores, onde cada um deles é treinado para corrigir o anterior, ajustando-os sobre os resíduos do anterior.

Finalmente, a inferência é feita de forma que o resultado da observação através de cada árvore treinada, é somada, resultando em um valor final para a instância.

2.6 Ajuste fino e avaliações

Nesta etapa é importante explicitar os objetivos do ajuste fino e das métricas de avaliação de modelos.

2.6.1 Ajuste fino

Chama-se de ajuste fino, a etapa que compreende a utilização de algum método de verificação de Hiper parâmetros para otimizar o modelo ao problema proposto. Há vários métodos utilizados para isto, mas os mais importantes são o *Grid Search* no qual é uma ferramenta exaustiva para busca da melhor combinação de Hiper parâmetros. O Grid Search funciona treinando um modelo para cada combinação de especificada e retornando, para cada treinamento e teste, uma pontuação dada uma métrica especificada, que é utilizada para retornar o melhor estimador.

Finalmente é importante ressaltar que o conjunto de teste e treino nesta etapa é o próprio conjunto de treino antes definido, que é particionado em um número de subconjuntos previamente definido pelo usuário. Este particionamento é de tal forma que, dado o número de divisões seja k, então $k - 1$ subconjuntos são utilizados para treinamento e o último utilizado para teste (27), sendo repetido este mesmo processo k vezes com todos os subconjuntos passando pelo conjunto de testes uma vez.

2.6.2 Avaliações

As avaliações devem ser analisadas para cada problema no qual está tentando resolver, para este trabalho a utilização do Coeficiente de Determinação simbolizado como R^2, assim como a raiz quadrada do erro médio quadrático simbolizado como $RMSE$.

2.6.2.1 Coeficiente de Determinação

O Coeficiente de Determinação é descrito como medida que quantifica a proporção de variância explicada pelo conjunto de dados previsto em comparação com o conjunto original. Segundo (28), muitos cientistas usam esta métrica para avaliar precisão de modelos lineares e é definida em (28), como:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(y_i - \bar{y})} \tag{18}$$

onde $\hat{y}_i$ é o valor previsto da variável resposta para a i-ésima observação. Por fim, vale ressaltar que a métrica varia entre 0 e um, sendo zero nenhuma variância explicada, e um representando que toda variância pode ser explicada com os dados previstos para a variável resposta.

2.6.2.2 Raiz quadrada do erro médio quadrático

Uma medida muito utilizada para avaliação é a Raiz quadrada do erro médio quadrático ou RMSE. A métrica tem objetivo descrito por (27) de dar uma ideia de quanto erro um sistema geralmente comete em suas previsões, penalizando erros muito grandes devido aos termos quadráticos em:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{19}$$

2.7 Variáveis Desprezíveis

Os campos que serão alimentados no modelo devem ser relevantes o suficiente para produzir um modelo confiável ao problema. Segundo (27), a escolha da informação relevante consiste em três possíveis caminhos que podem ser seguidos simultaneamente.

- Seleção de atributos: Consiste em selecionar as melhores características que explicam o problema, entre todas as existentes.
- Extração de atributo: Este método, consiste em combinar características já existentes no conjunto, para produzir uma mais útil à modelagem.
- Criação de atributos: Consiste em recolher novos dados.

A abordagem utilizada no problema deste presente trabalho diz respeito a seleção de atributos, na qual possui um enorme arsenal de ferramentas. Duas destas ferramentas, a análise de Importância e Eliminação Recursiva de atributos, são formas interessantes de interpretar e selecionar característica para melhorar a qualidade da modelagem.

2.7.1 Importância

A análise de importância, segundo (29), se refere a uma classe de técnicas na qual tem o objetivo de quantificar a relevância das características ao modelar a variável resposta. Esta ferramenta é útil para interpretar quais características influenciam mais no problema e quais não são importantes para descrevê-lo, além de reduzir o número de variáveis para melhorar a eficiência do treinamento e inferência. É calculada tanto para problemas de Regressão quanto para Classificação e é obtida de forma diferente dependendo do tipo de Algoritmo utilizado, embora nem todos os algoritmos permitam a obtenção de tais pontuações, os modelos com Floresta de Aleatória e *Extreme Gradient Boosting* permitem que estas pontuações sejam calculadas em seus interiores.

A obtenção da Importância para modelos implementados com algoritmo CART, segundo (29), como Árvore de Decisão, Floresta Aleatória e Extreme Gradient Boosting são feitos de forma similar e baseados na soma total da redução do tipo de critério utilizado para dividir os conjuntos em cada nó, como Erro médio quadrático (MSE) para problemas de Regressão.

2.7.2 Eliminação recursiva de atributos

O a descrição feita por (30) mostra que a Eliminação Recursiva de Características (RFE) é um método que aprimora modelos de Machine Learning através de um processo iterativo. Em cada etapa, o modelo é treinado e as características menos relevantes são identificadas e removidas, refinando o modelo para as próximas iterações. Essa técnica é particularmente útil para algoritmos que conseguem quantificar a importância de cada característica.

Finalmente, é importante ressaltar que esta técnica pertence a classe de algoritmos de seleção *Wrapper*, como é descrito em (31), visto que encapsula um Algoritmo de Aprendizado de Máquina para utilizar como núcleo da Eliminação recursiva.

3 METODOLOGIA

Este capítulo busca evidenciar a natureza exploratória do trabalho, assim como os motivos das escolhas dos dados, métodos estatísticos e tecnologias. É importante ressaltar que as fontes utilizadas foram de classe secundária, ou seja, buscadas através de livros, sites e artigos científicos, dado a grande quantidade de material de qualidade disponível sobre as tecnologias e o problema em questão.

Esta etapa está dividida em:

- Programação.
- Estruturação da análise e resultados.

3.1 Programação

A linguagem de programação utilizada neste trabalho foi o Python na versão 3.11.8 em conjunto com o Ambiente de desenvolvimento integrado (IDE), *Visual Studio Code*. Esta escolha foi feita dado a sua vasta comunidade e grande quantidade de bibliotecas voltadas para análise estatística e Aprendizado de Máquina.

Finalmente, os pacotes utilizados juntamente com suas descrições estão listados na Tabela 3.

Tabela 3 – Bibliotecas utilizadas.

Biblioteca	Descrição
pandas	Pandas é uma biblioteca para análise e manipulação de dados. É utilizado para leitura, escrita, filtro, transformação, agregação e visualização de DataFrames.
numpy	Numpy é a base da computação científica em Python e é utilizado para manipulação de matrizes e operações matemáticas.
scipy	Scipy é uma coleção de algoritmos e funções matemáticas que complementa o Numpy. É utilizado para tarefas científicas e de engenharia que exigem cálculos numéricos avançados.
statsmodels	Statsmodels é uma biblioteca para análise estatística, na qual fornece funções e classes para estimar e interpretar modelos estatísticos.
matplotlib	Matplotlib é a biblioteca de visualização, na qual permite livre manipulação ao usuário para personalização e criação de gráficos.
sklearn	Scikit-learn é uma biblioteca abrangente para aprendizado de máquina em Python. Esta biblioteca é utilizada para construir, treinar e avaliar modelos eficientemente.

Biblioteca	Descrição
xgboost	XGBoost (*Extreme Gradient Boosting*) é uma biblioteca de código aberto que implementa o algoritmo *Gradient Boosting* de forma otimizada.
joblib	Joblib é uma biblioteca para facilitar o paralelismo em processos computacionais e também utilizada para leitura e escrita de modelos de Aprendizado de Máquina.
os	OS é uma biblioteca que permite a interação com o sistema operacional de maneira flexível e simples.
warnings	Warnings é uma biblioteca utilizada para suprimir avisos e erros desnecessários para aplicação.
seaborn	O Seaborn é uma biblioteca construída em cima do matplotlib na qual facilita a construção de visualizações gráficas.

Fonte: O autor.

3.2 Estruturação da análise e resultados

A análise e previsão de potência foi estruturada em cinco partes, dependentes e ordenadas para simplificar o progresso da pesquisa. A divisão foi feita da seguinte maneira:

1. Importação de Bibliotecas e Carregamento dos Dados.
2. Sistematização da Análise Exploratória de Dados.
3. Descrição do Pré-processamento.
4. Escolha do Ajuste Fino.
5. Modelagem e resultados.

3.2.1 Importação de Bibliotecas e Carregamento dos Dados

Nesta etapa, foram importadas todas as bibliotecas necessárias para a análise e previsão, as quais encontram-se listadas na Tabela 3, assim como a utilização da biblioteca Pandas para o carregamento dos dados em formato *.csv* para um *Dataframe*. Além disso, foi necessário converter os tipos dos dados, nos quais são carregados como *int64* e *float64* para formatos mais eficientes como *int32* e *float32*.

3.2.2 Sistematização da Análise Exploratória de Dados

A análise exploratória é a etapa mais importante do processo, foi feita utilizando as bibliotecas de matemática, visualização e manipulação de dados e tem o objetivo de gerar conhecimento sobre os dados, como conhecimentos sobre distribuição, correlação, relacionamentos não lineares, medidas estatísticas e valores extremos.

Finalmente, nesta etapa, foram utilizados gráficos de barras para análise de média entre grupos, gráficos de dispersão para verificação de relacionamento entre variáveis além de histogramas, Box plots e gráfico de correlações para averiguar o comportamento aleatório e relacional das características.

3.2.3 Descrição do Pré-processamento

O processamento dos dados foi feito utilizando a biblioteca Sklearn, visto que possui uma gama de ferramentas para várias tarefas relacionadas a Aprendizado de Máquina. Esta etapa foi composta de, primeiramente, divisão de amostras com 70% das observações contendo o conjunto de treino e 30% o conjunto de teste. Estes subconjuntos foram obtidos de forma aleatória e sem repetição utilizando função específica do pacote.

Na segunda parte foi feita a conversão das variáveis categóricas para representação numérica utilizando *One-Hot Encoding*, dado que nenhum dos algoritmos utilizados são implementados para receber valores não numéricos.

A terceira etapa foi composta da seleção de variáveis através da ferramenta de Eliminação Recursiva de atributos, além da análise de importância dos modelos *Extreme Gradient Boosting* e Árvore Aleatória.

A quarta etapa, foi composta da transformação dos campos numéricos, utilizando a abordagem Normalização pela função *MinMaxScaler* do pacote Sklearn, visto que os dados possuem distribuições e escalas diferentes e, portanto, podendo afetar negativamente os modelos lineares utilizados.

Finalmente, na quinta etapa, foram avaliados sete algoritmos para seleção daquele que melhor se ajustava aos problemas de previsão de potência de resfriamento e aquecimento. Esta avaliação foi feita utilizando a função *KFold* na qual

treina um modelo arbitrário k vezes utilizando o conjunto de treino particionado, e retorna o valor do Coeficiente de Determinação e a Raiz do Erro médio quadrático, este método é chamado de validação cruzada e foi utilizado um valor de k igual a cinco. Estas métricas foram então plotadas em um gráfico *Box plot*, para avaliar a simetria dos resultados dos sete Algoritmos. Com isto, foi possível selecionar o mais interessante aos problemas.

3.2.4 Escolha do Ajuste Fino

Após a escolha do melhor Algoritmo, foi feito o ajuste fino para determinar o melhor conjunto de Hiper parâmetros. Esta etapa foi feita utilizando a função *GridSearch* do pacote Sklearn, na qual teve o objetivo de buscar exaustivamente os melhores valores entre seis parâmetros, tanto para o problema de resfriamento quanto para o problema de aquecimento, visto que a quantidade de dados não é grande e, portanto, o treinamento não é computacionalmente dispendioso.

3.2.5 Modelagem e resultados

Os modelos, após a escolha dos Hiper parâmetros, foram clonados para agregar ao dicionário de parâmetros, e em seguida foram retreinados no conjunto completo de treinamento. Com os modelos de aquecimento e resfriamento devidamente treinados, foi possível fazer a inferência no conjunto de treinamento e no conjunto de teste para verificar os resultados e possíveis problemas de sobreajuste e subajuste.

Finalmente, é importante ressaltar que os resultados possuem natureza quantitativa e foram formatados em tabelas, sendo diferenciados pela métrica utilizada e pelo conjunto de dados a qual pertencem.

4 ANÁLISE E DESENVOLVIMENTO

Este capítulo está destinado à descrição dos resultados da análise e previsão da Potência de resfriamento e aquecimento de estruturas prediais.

4.1 Descrição estatística das variáveis independentes e dependentes

As descrições estatísticas foram obtidas para trazer familiaridade com as métricas e, portanto, mostrar informação sobre possíveis valores extremos ou erros de coleta, além de ser importante para avaliar a diferença de escala entre o espaço de variáveis. Na Tabela 4, é possível verificar a descrição contendo a média aritmética, desvio padrão amostral, valor mínimo, primeiro quartil, mediana, terceiro quartil e o valor máximo para todas as colunas numéricas.

Tabela 4 – Estatísticas do conjunto de dados numérico.

	Compactação Relativa	Área de superfície	Área de parede	Área e telhado	Altura global	Área de envidraçamento	Carga de Aquecimento	Carga de Resfriamento
Média	0,764	671,708	318,500	176,604	5,250	0,234	22,307	24,588
Desvio Padrão	0,106	88,086	43,626	45,166	1,751	0,133	10,090	9,513
Mínimo	0,620	514,500	245,000	110,250	3,500	0,000	6,010	10,900
1° quartil	0,682	606,375	294,000	140,875	3,500	0,100	12,992	15,620
Mediana	0,750	673,750	318,500	183,750	5,250	0,250	18,950	22,080
3° quartil	0,830	741,125	343,000	220,500	7,000	0,400	31,667	33,133
Máximo	0,980	808,500	416,500	220,500	7,000	0,400	43,100	48,030

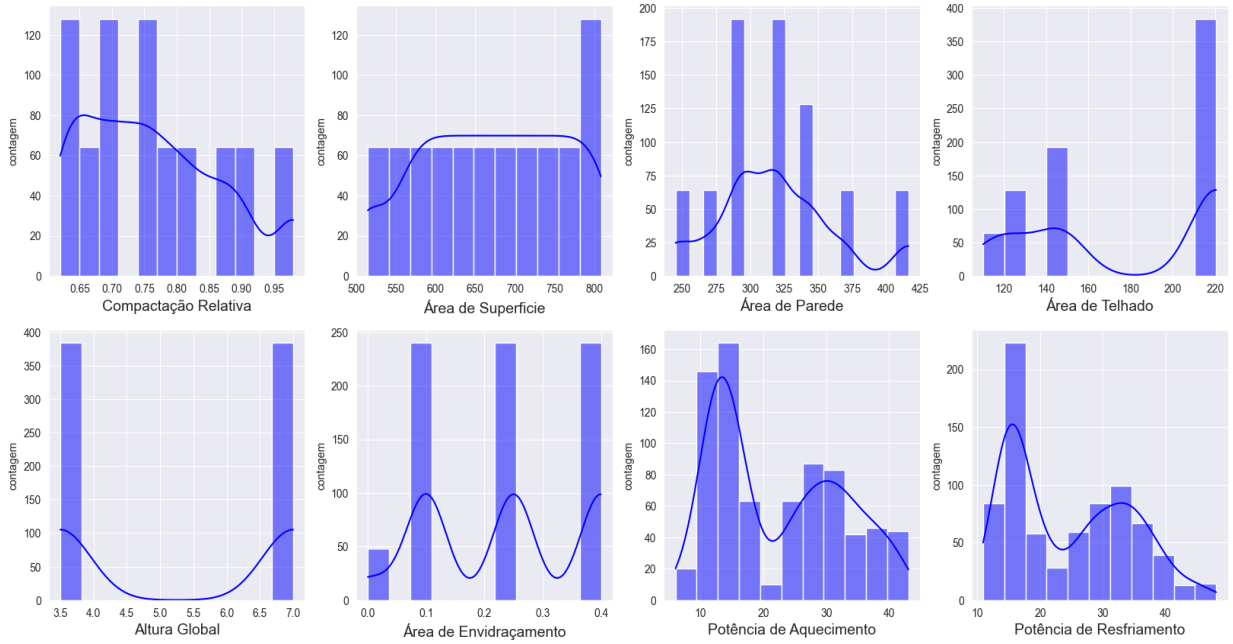
Fonte: O autor.

Estas estatísticas evidenciam que os alcances são desproporcionais. Além disso, é possível notar que a variabilidade nos dados de Potência de aquecimento é sutilmente superior em comparação com a Potência de resfriamento. Finalmente, é possível notar que não há valores errados, ou negativos para nenhuma das variáveis em questão.

4.2 Distribuições experimentais

As distribuições para as variáveis numéricas foram plotadas em uma imagem na qual evidencia tanto os Histogramas quanto a Função Densidade de Probabilidade teórica, gerada com base nas amostras de cada característica. Esta imagem pode ser visualizada na Figura 13.

Figura 13 – Distribuições experimentais.

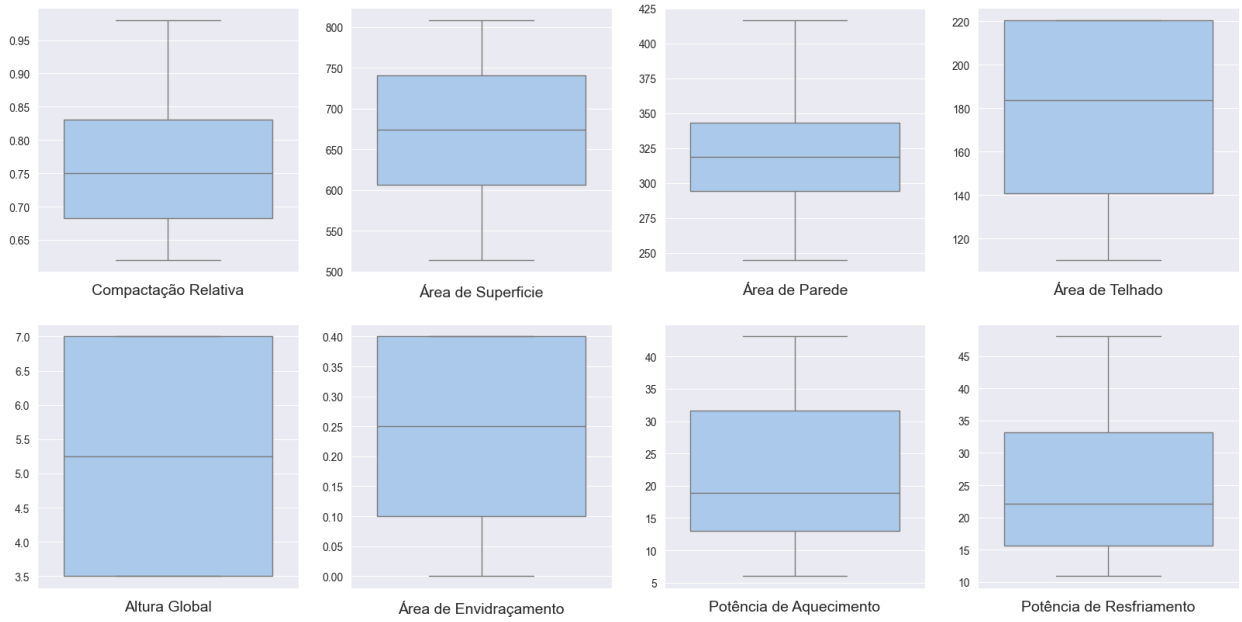


Fonte: O autor.

É possível notar que nenhuma distribuição se assemelha com a distribuição normal, levando a considerar certas ferramentas futuras para análise de correlação de transformação. Em seguida, é possível perceber as Potências de Aquecimento e Resfriamento têm distribuições muito similares, levando a interpretar que o comportamento delas são semelhantes. Além disso, é possível notar que para a variável Altura global, apenas dois valores foram simulados.

Finalmente, foi plotado o conjunto de Box plots para averiguar dados extremos, e para interpretar melhor a assimetria da distribuição dos conjuntos de dados. A Figura 14, mostra estes gráficos para cada variável numérica.

Figura 14 – Box Plots das variáveis numéricas.



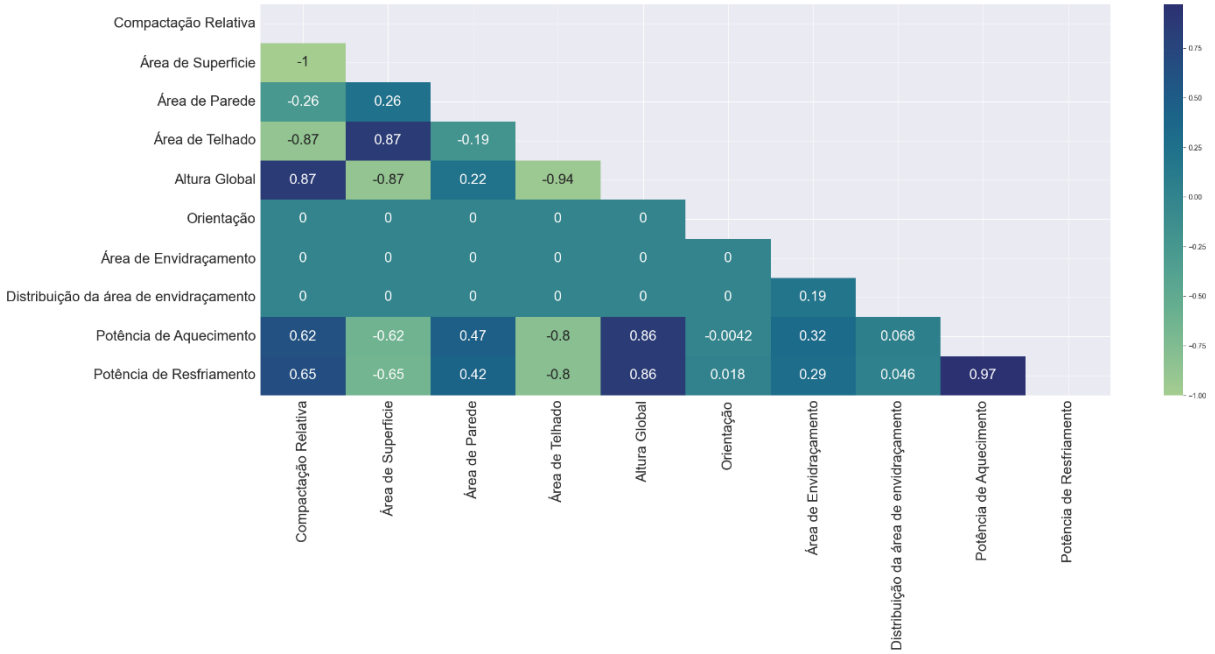
Fonte: O autor.

É possível perceber que, baseado na análise de Box plot, não há dados extremos. Ainda que o grau de simetria entre as Potências de resfriamento e aquecimento sejam muito similares, há um leve deslocamento para cima na distribuição da Potência de resfriamento em comparação com a de aquecimento.

4.3 Visualização das correlações

A escolha da correlação de *Spearman* foi embasada no fato de que as distribuições não são normais, portanto, sendo mais interessante a utilização de ferramentas estatísticas não paramétricas. Com isto foi plotado um mapa de calor que é evidenciado na Figure 15, na qual foi possível verificar todas os pares de correlações entre as variáveis independentes e dependentes, para saber se havia multicolinearidade, além de ser necessário saber o grau de correlação com as Potências envolvidas.

Figura 15 – Correlações de *Spearman*.



Fonte: O autor.

É importante notar na Figura 15, que há multicolinearidade entre as variáveis Altura global, Área de telhado, Compactação Relativa e Área de superfície, dado que as correlações dois a dois são relativamente altas para a consideração de que são variáveis independentes. Em seguida, é possível perceber que as variáveis de potência são extremamente correlacionadas entre si, corroborando com o fato de suas distribuições serem muito similares como vistas no Histograma. Ainda é possível ver que única variável com correlação relativamente alta e negativa, com as variáveis de potência, é a Área de superfície, estabelecendo um relacionamento inversamente proporcional.

Entre as características numéricas, a variável Área de envidraçamento é a que menos tem correlação com as potências, com valores de 0,29 e 0,32, respectivamente para potência de resfriamento e aquecimento, enquanto que a variável Altura global possui a maior correlação com valor de 0,86 para ambas as potências.

Finalmente, com relação as características categóricas, a correlação se mostra muito fraca, tanto para a variável Orientação quanto para a Distribuição da área de envidraçamento.

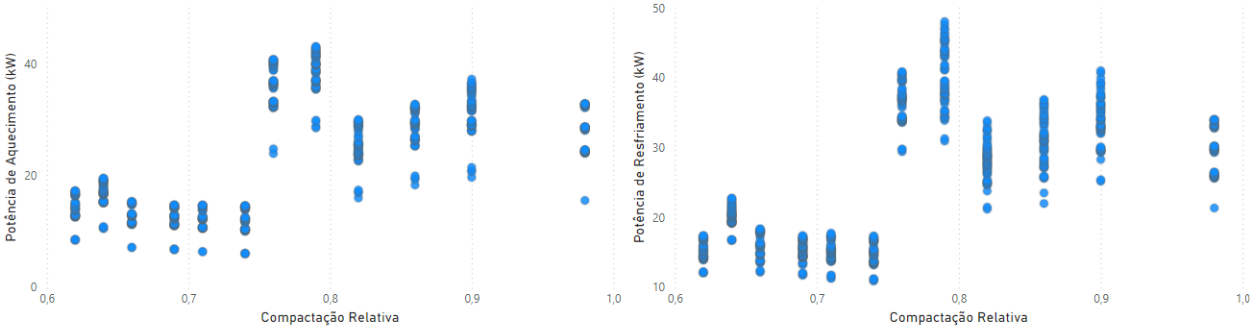
4.4 Relacionamentos entre as variáveis independentes e as variáveis de Potência

O relacionamento entre variáveis é importante, para deixar claro o formato em que elas se relacionam, ou seja, se há linearidade ou não linearidade, ou se não possuem nenhum tipo de padrão ou características singulares entre grupos distintos. análise foi feita para cada variável explicativa a partir de plotagem de gráficos contra as variáveis de potência.

4.4.1 Compactação Relativa versus Potência de Resfriamento e Aquecimento

O relacionamento entre a Compactação Relativa e as potencias não possui uma resposta simples, como pode ser visto na Figura 16. Entretanto, é possível ver que edifícios mais compactos tendem a ser mais eficientes em termos de energia em ambos as situações. Além disso há dois grupos diferentes divididos em aproximadamente um valor de Compactação Relativa de 0,75.

Figura 16 – Relacionamento com a Compactação Relativa.



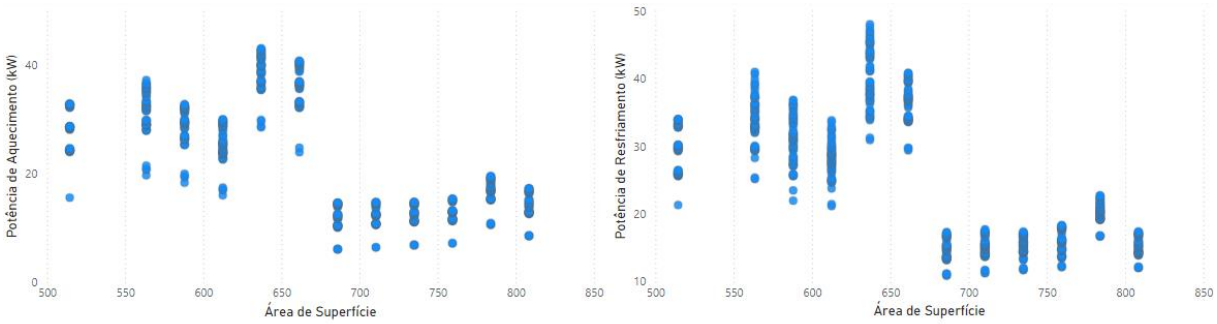
Fonte: O autor.

4.4.2 Área de Superfície versus Potência de Resfriamento e Aquecimento

A plotagem do relacionamento da área de superfície, como vista na Figura 17, mostra que as estrutura prediais estão divididos em dois grupos, com área entre 650 e 700 metros quadrados. Essa divisão sugere uma possível relação entre a área da superfície e as potências. Com base nesta evidência, é possível inferir que as formas

com menor área de superfície tendem a ser menos eficientes em termos de energia. Isso significa que os edifícios do grupo com metragem mais próxima de 700 m^2 provavelmente apresentam um consumo energético para resfriamento e aquecimento menor em comparação com as estruturas do grupo com metragem mais próxima de 650 metros quadrados.

Figura 17 – Relacionamento com a Área de superfície.

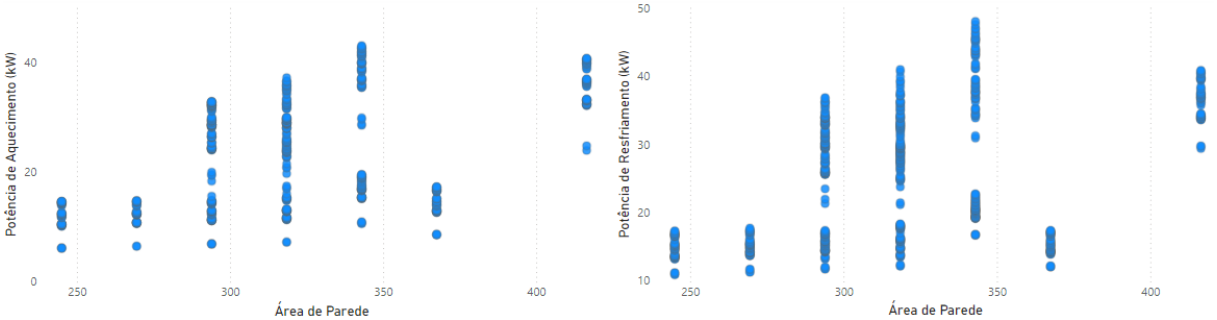


Fonte: O autor.

4.4.3 Área de Parede versus Potência de Resfriamento e Aquecimento

A análise dos dados, revela uma relação não linear entre a área da parede e as potências de aquecimento e resfriamento, com construções de área inferior a 300 se mostrando mais eficientes, como pode ser visto na Figura 18. No entanto, a existência de edificações com áreas entre 350 m^2 e 375 m^2, e alta eficiência energética sugere que outros fatores afetam as variáveis de potências, tornando estas observações mais eficientes.

Figura 18 – Relacionamento com a Área de parede.

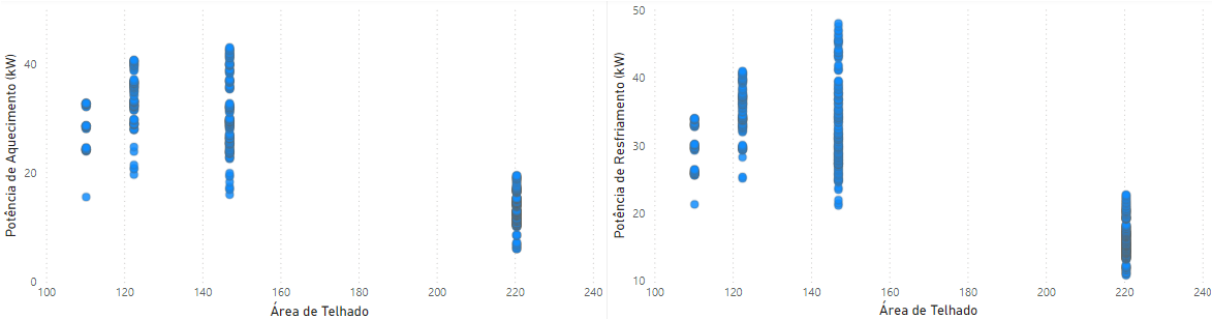


Fonte: O autor.

4.4.4 Área de Telhado versus Potência de Resfriamento e Aquecimento

A análise do relacionamento revela apenas quatro tipos diferentes de área de telhado, como pode ser evidenciado na Figura 19, e sugere uma possível relação inversa entre esta variável e a potência necessária para aquecimento e resfriamento, indicando que estruturas prediais com área de telhado menores podem demandar mais energia. Além disso, a maioria dos edifícios possui área de telhado inferior a 160 m^2, o que pode ter implicações significativas para o consumo energético dessas construções.

Figura 19 – Relacionamento com a Área de telhado.

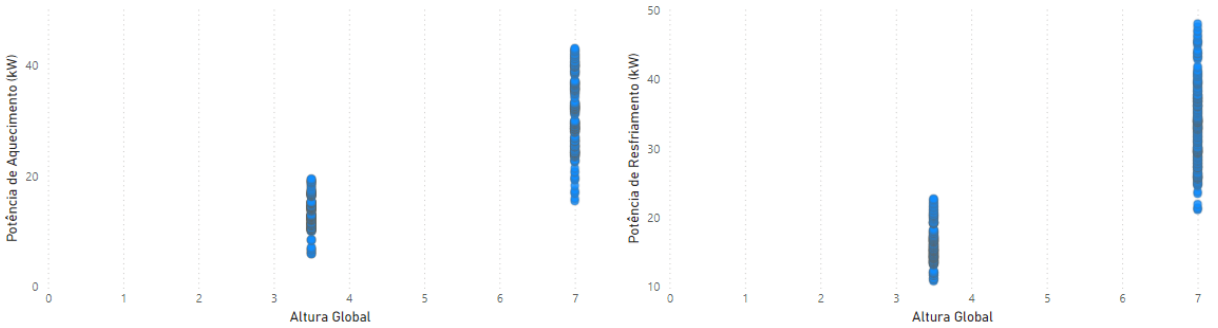


Fonte: O autor.

4.4.5 Altura global versus Potência de Resfriamento e Aquecimento

A análise do relacionamento entre a Altura global e a potência de aquecimento e resfriamento vista na Figura 20, mostra que existem apenas dois tipos de altura para os edifícios. Além disso, a maioria dos edifícios tem uma altura total de 7 m e são menos eficientes em termos de potência necessária. Além disso, a variação nas potências é maior entre as estruturas com altura total de 7 m em comparação com aqueles com altura total de 3,5 m. Isso sugere que, embora estruturas prediais mais altas tendam a ser menos eficientes em termos de energia, há uma maior diversidade em sua eficiência energética, indicando que outros fatores podem estar influenciando o consumo de energia nesses edifícios mais altos.

Figura 20 – Relacionamento com a Altura global.

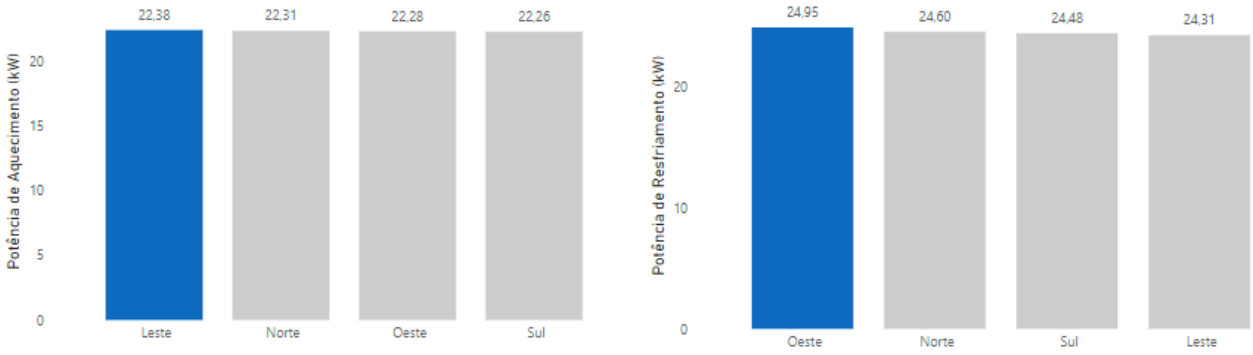


Fonte: O autor.

4.4.6 Orientação versus Potência de Resfriamento e Aquecimento

A análise dos gráficos de barras visto na Figura 21, sugere que a energia média necessária para aquecimento e resfriamento seja semelhante entre as diferentes orientações. Ainda assim, é possível verificar que a orientação leste tende a demandar mais energia para aquecimento com um valor médio de aproximadamente 22,38 kW, enquanto a orientação oeste apresenta maior demanda maior potência para resfriamento com um valor médio de 24,95 kW. Finalmente, pode ser inferido que as estruturas tendem a necessitar de mais energia, independente de Orientação, para resfriamento.

Figura 21 – Média da potência de aquecimento e resfriamento pela Orientação.

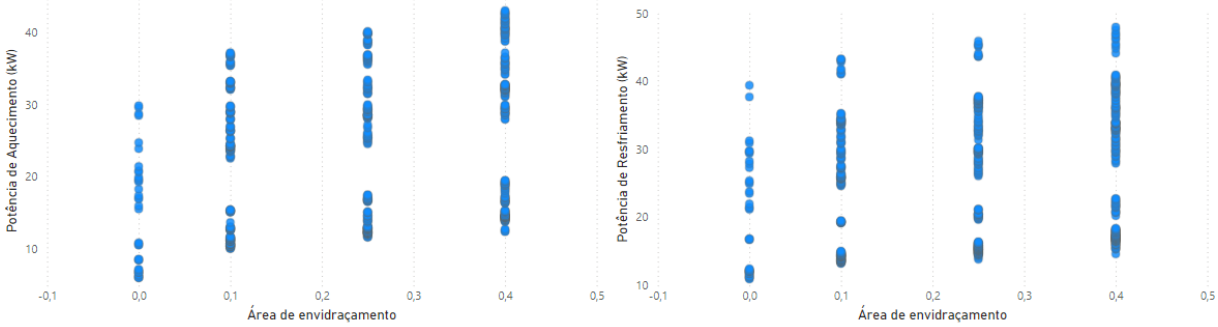


Fonte: O autor

4.4.7 Área de envidraçamento Potência de Resfriamento e Aquecimento

A análise dos gráficos de dispersão da Figura 22, revela uma relação complexa, porém positiva entre a área de envidraçamento e tanto a potência de resfriamento quanto a de aquecimento. Isso indica que, com o aumento da área de envidraçamento, há um aumento na demanda por energia para resfriar e aquecer o ambiente. Finalmente, ainda é possível verificar que a potência para resfriamento é maior em todas as percentagens de envidraçamento em comparação com a potência de aquecimento. Pode ser inferido então, que a incidência solar em ambientes com área de vidro tende a facilitar o aquecimento e, portanto, menos energia será exigido para esquentar o mesmo ambiente.

Figura 22 – Relacionamento com a Área de envidraçamento.



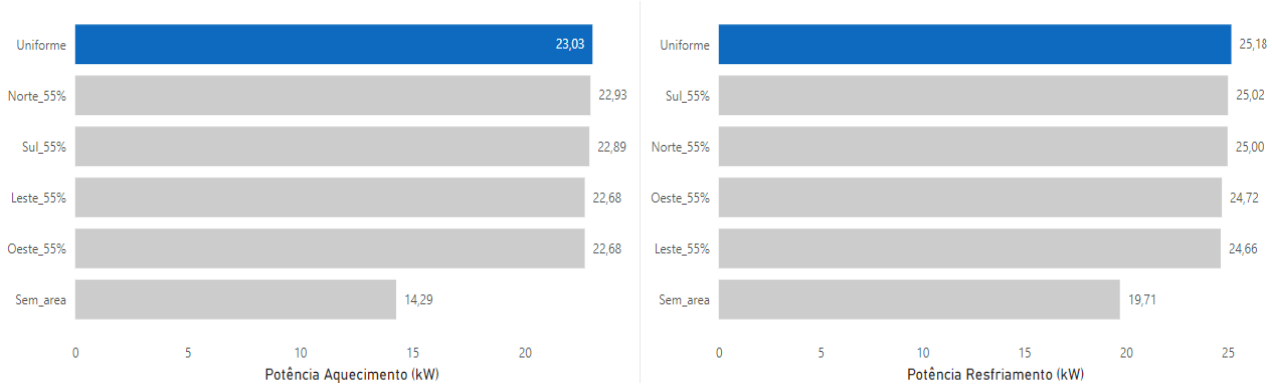
Fonte: O autor

4.4.8 Distribuição da área de envidraçamento versus Potência de Resfriamento e Aquecimento

O gráfico de arras da Figura 23, possui a informação da média da potência de resfriamento e aquecimento, dado os variados tipos de distribuições de área envidraçada. Com isto em mente, é possível perceber que estruturas prediais sem áreas envidraçadas demonstram maior eficiência energética. Além disso, é possível notar que a potência necessária para aquecimento é menor, independente do direcionamento da distribuição, do que a potência necessária para resfriamento.

Por fim, a distribuição de envidraçamento Uniforme, simbolizado pelo número 1, está associada ao maior consumo médio de energia, tanto para aquecimento quanto para resfriamento, indicando que a orientação da distribuição de envidraçamento pode impactar significativamente a eficiência energética das estruturas.

Figura 23 – Média da potência de aquecimento e resfriamento pela Distribuição de área de envidraçamento.



Fonte: O autor

4.5 Análise dos resultados da Seleção de características

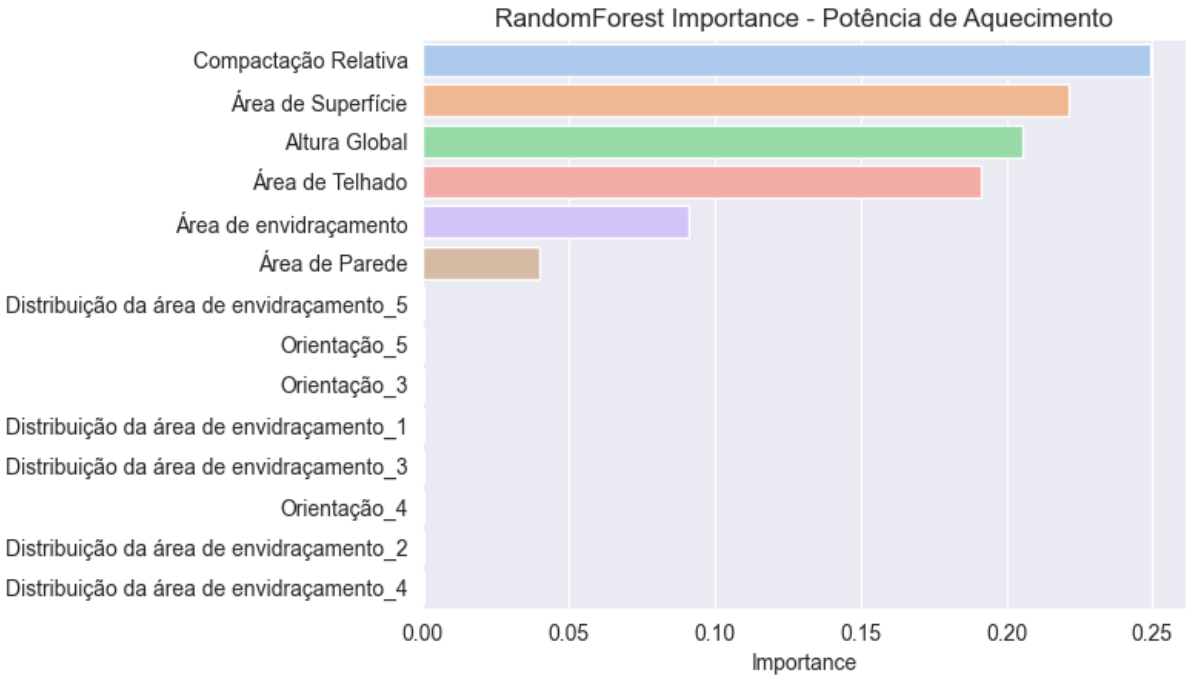
Visto que algumas características testadas, de forma independente, estão correlacionadas com as Potências, foi então verificado o grau de importância do relacionamento conjunto entre todas as variáveis explicativas para entender melhor quais delas têm maior peso para a previsão. Após isto, foi obtido o conjunto de variáveis que trazem os melhores resultados para um conjunto de Algoritmos, através do método Eliminação Recursiva de Características.

4.5.1 Importância do Algoritmo Floresta Aleatória e Extreme Gradient Boosting

Primeiramente, foi treinado um modelo de Floresta Aleatória para previsão tanto da Potência de resfriamento quanto para aquecimento, dado que este algoritmo permite o retorno da medida de importância, assim como o Extreme Gradient

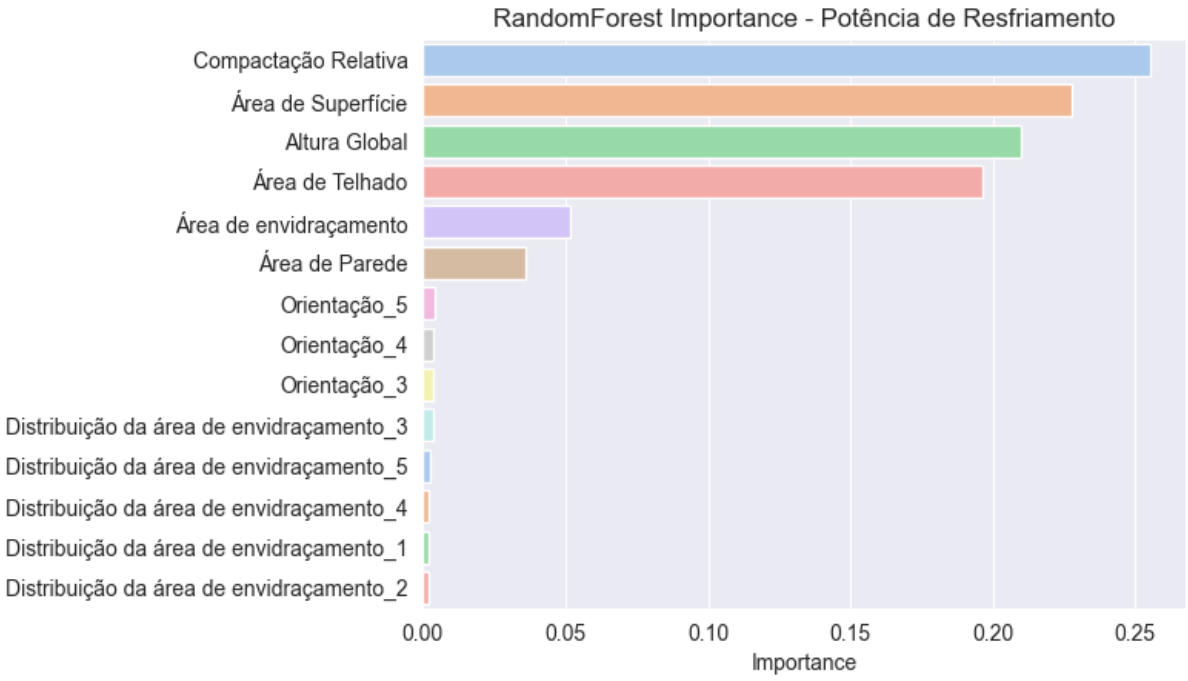
Boosting. As plotagens com respeito ao problema de aquecimento e resfriamento, para o primeiro algoritmo, se encontram respectivamente nas Figuras 24 e 25.

Figura 24 – Importâncias do algoritmo Floresta Aleatória para a Potência de aquecimento.



Fonte: O autor

Figura 25 – Importâncias do algoritmo Floresta Aleatória para a Potência de resfriamento.

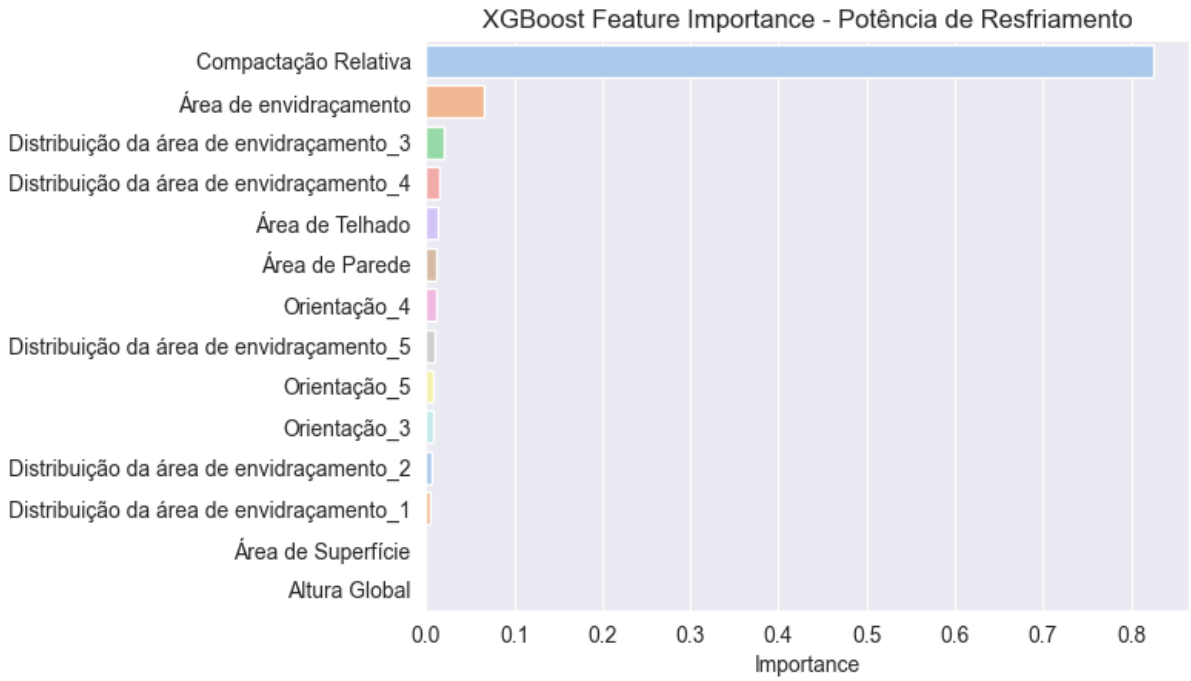


Fonte: O autor

O resultado para problema de aquecimento, visto na Figura 24, mostra que a Compactação Relativa é a característica mais importante para a previsão da Potência, seguida da Área de Superfície. Em contra partida, as variáveis categorias, Orientação e Distribuição da área de envidraçamento não possuem nenhuma importância segundo o Algoritmo utilizado, corroborando com os resultados da Correlação, na qual evidenciou baixo valor. Analisando os resultados para o problema de resfriamento, é possível notar que, as variáveis numéricas possuem a mesma ordem de importância em comparação com o problema de aquecimento, diferenciando apenas no fato em que há uma pequena importância associada às variáveis Orientação e Distribuição da área de envidraçamento.

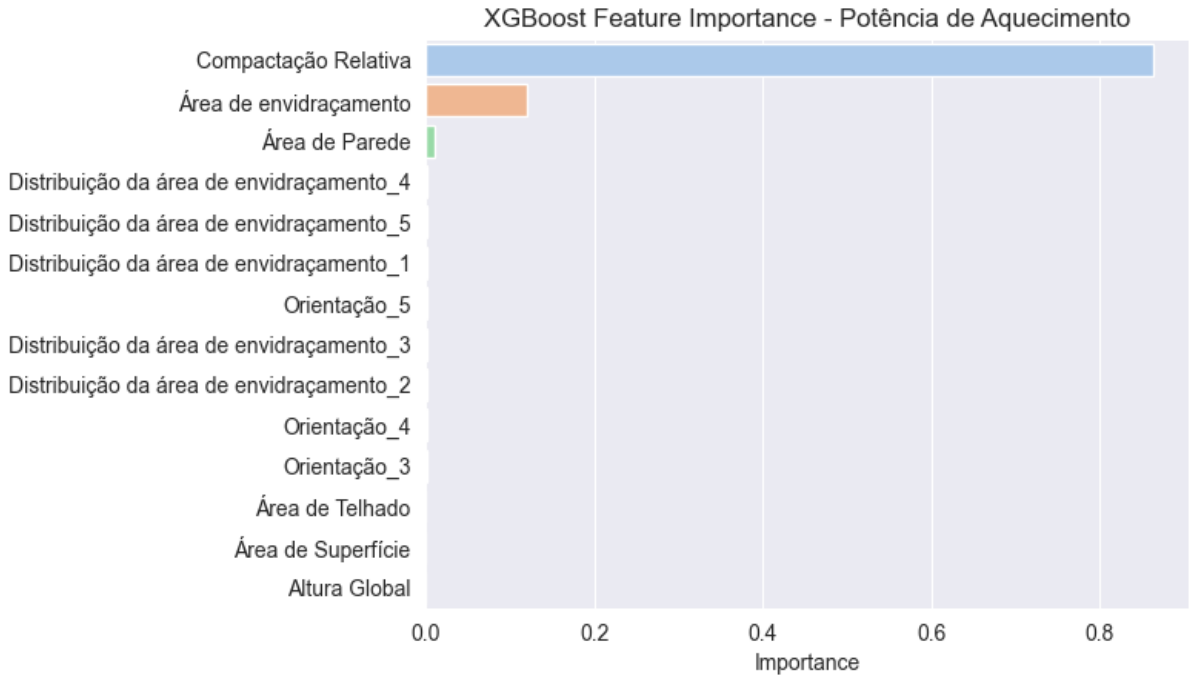
Além das análises de importância com o modelo de Floresta Aleatória, foram obtidas as importâncias com o modelo *Extreme Gradient Boosting*, visto que é um modelo muito rápido e que trouxe bons resultados em muitos problemas de Regressão. Os resultados se encontram nas Figuras 26 e 27, onde respectivamente estão relacionadas com a Potência de resfriamento e aquecimento.

Figura 26 – Importâncias do algoritmo *Extreme Gradient Boosting* para a Potência de resfriamento.



Fonte: O autor

Figura 27 – Importâncias do algoritmo *Extreme Gradient Boosting* para a Potência de aquecimento.



Fonte: O autor

As Figuras 26 e 27, evidenciam que a Compactação Relativa é muito mais importante em comparação com as outras variáveis, para ambos os problemas, seguido da Área de envidraçamento com uma discrepância muito grande com relação ao primeiro lugar.

A diferença entre as duas análises, é mais explícita no fato de que o terceiro lugar do ranque de importância, para o problema de resfriamento, é variável codificada Distribuição de área de envidraçamento Leste, enquanto que para o aquecimento é a Área de parede. Além disso, as variáveis a partir do quarto lugar no ranque para o problema de aquecimento possuem importância aproximadamente zero, enquanto que para o problema de resfriamento apenas as variáveis Área de superfície e Altura global têm este resultado.

Finalmente, é importante enfatizar que os resultados da correlação foram com relação a variável mais importante é diferente em comparação ao resultado da importância pois a correlação é avaliada independentemente das outras variáveis explicativas, enquanto que o cálculo da importância depende de todas as variáveis envolvidas no treinamento dos modelos.

4.5.2 Seleção de variáveis com a Eliminação Recursiva de Características para os problemas de Potência de aquecimento e resfriamento

Os resultados a seguir foram utilizados para selecionar as melhores características para cada problema, separadamente. Esta abordagem foi utilizada com ambos os modelos citados até aqui e selecionaram características diferentes tanto entre modelos, quanto entre problemas.

A Tabela 5, mostra o conjunto de resultados, baseados na raiz do erro médio quadrático, para cada conjunto de características diminuídas a cada iteração do algoritmo para o problema de aquecimento.

Tabela 5 – Resultado da Eliminação Recursiva de Características para o problema de aquecimento.

	Floresta Aleatória	*Extreme Gradient Boosting*
14 características	0,553	0,420
13 características	0,558	0,420
12 características	0,559	0,420
11 características	0,546	0,439
10 características	0,542	0,467
9 características	0,538	0,492
8 características	0,525	0,496
7 características	0,526	0,508
6 características	0,512	0,496
5 características	0,511	0,493
4 características	3,168	0,489
3 características	3,168	0,488
2 características	3,168	0,488
1 características	3,168	3,167

Fonte: O autor

O resultado da Tabela 5, mostra que o algoritmo *Extreme Gradient Boosting* alcança métricas melhores em comparação ao algoritmo Floresta Aleatória, com um resultado em sua melhor iteração de aproximadamente 0,420, entretanto, utiliza 12 das 14 características existentes, enquanto que o melhore modelo de Floresta Aleatória seleciona apenas 5 características, com um resultado de aproximadamente 0,511.

Por fim, visto que o modelo *Extreme Gradient Boosting*, trouxe melhores resultados, e foi selecionado para continuidade da seleção de variáveis. O resultado

após a seleção está compilado na Tabela 6, na qual mostra que as variáveis retiradas para o problema de aquecimento, foram Área de Superfície e Atura global, corroborando com o resultado da Importância da Figura 26.

Tabela 6 – Variáveis selecionadas para o problema de aquecimento.

Variáveis	*Extreme Gradient Boosting*
Compactação Relativa	Manter
Área de superfície	Retirar
Área de parede	Manter
Área de telhado	Manter
Altura global	Retirar
Área de envidraçamento	Manter
Orientação Leste	Manter
Orientação Sul	Manter
Orientação Oeste	Manter
Distribuição da área de envidraçamento Uniforme	Manter
Distribuição da área de envidraçamento Norte	Manter
Distribuição da área de envidraçamento Leste	Manter
Distribuição da área de envidraçamento Sul	Manter
Distribuição da área de envidraçamento Oeste	Manter

Fonte: O autor

A seguir os foram obtidos os mesmos resultados, porém, para o problema de resfriamento. A Tabela 7, evidencia os resultados da Eliminação Recursiva de Características, onde mais uma vez o algoritmo *Extreme Gradient Boosting* aparenta possuir os melhores resultados, enquanto que a Tabela 8 mostra o conjunto das variáveis selecionadas.

Tabela 7 – Resultado da Eliminação Recursiva de Características para o problema de resfriamento.

	Floresta Aleatória	*Extreme Gradient Boosting*
14 características	1,806	1,294
13 características	1,789	1,294
12 características	1,717	1,294
11 características	1,887	1,288
10 características	1,912	1,330
9 características	1,885	1,444
8 características	1,803	1,686

7 características	1,796	1,825
6 características	1,722	1,913
5 características	1,721	1,963
4 características	2,702	1,869
3 características	2,702	1,803
2 características	2,702	1,719
1 características	2,702	2,698

Fonte: O autor

Tabela 8 – Variáveis selecionadas para o problema de resfriamento.

Variáveis	*Extreme Gradient Boosting*
Compactação Relativa	Manter
Área de superfície	Retirar
Área de parede	Manter
Área de telhado	Manter
Altura global	Retirar
Área de envidraçamento	Manter
Orientação Leste	Manter
Orientação Sul	Manter
Orientação Oeste	Manter
Distribuição da área de envidraçamento Uniforme	Retirar
Distribuição da área de envidraçamento Norte	Manter
Distribuição da área de envidraçamento Leste	Manter
Distribuição da área de envidraçamento Sul	Manter
Distribuição da área de envidraçamento Oeste	Manter

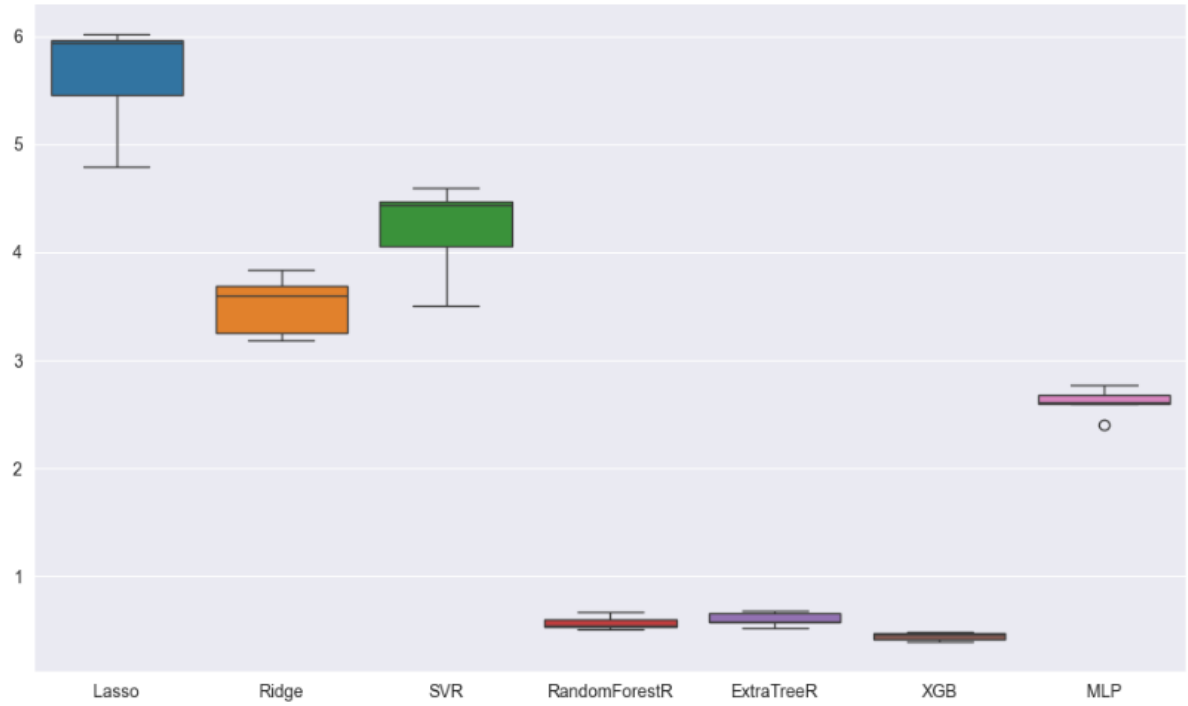
Fonte: O autor

O resultado da Tabela 8 mostra que o problema de resfriamento rejeita, além das duas variáveis rejeitadas no problema de aquecimento, a variável codificada Distribuição da área de envidraçamento Norte como uma variável que traz benefício para a previsão.

4.5.3 Resultados da escolha do melhor algoritmo

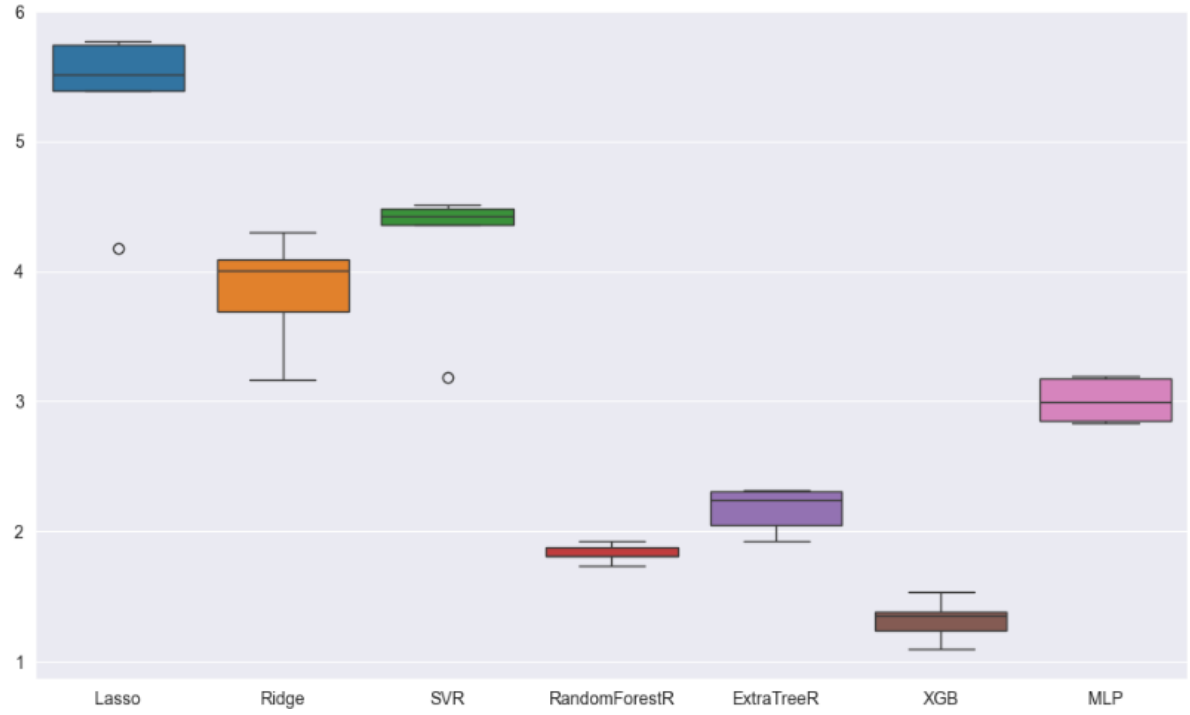
Após o pré-processamento das variáveis categóricas com o *One Hot Encoding* e das variáveis numéricas com a Normalização, além da seleção de variáveis, foi necessário encontrar, de maneira mais confiável do que apenas olhar as métricas da Eliminação recursiva de Características, o algoritmo que melhor se ajusta a ambos os problemas em questão. Para isto, foi feita uma análise de variabilidade dos resultados para checar se, após múltiplos treinamentos e avaliações, qual algoritmo gerou os melhores resultados, em outras palavras, qual modelo gerou em média métricas menores e com baixo desvio padrão da métrica Raiz do erro médio quadrático. Foi possível ainda, construir um gráfico de Box Plot, para cada problema, com o resultado de cada algoritmo. As plotagens, para respectivamente a potência de aquecimento e resfriamento, se encontram nas Figuras 28 e 29.

Figura 28 – Box Plots de sete algoritmos para o problema de aquecimento.



Fonte: O autor

Figura 29 – Box Plots de sete algoritmos para o problema de resfriamento.



Fonte: O autor

As Figuras 28 e 29 evidenciam que o algoritmo *Extreme Gradient Boosting* é realmente o melhor algoritmo para ambos os problemas, enquanto que em segundo lugar se encontra o modelo de Floresta Aleatória. Para uma análise mais minuciosa foram obtidas duas tabelas, nas quais tiveram o objetivo de mostrar claramente os valores de média, desvio padrão, valor mínimo, primeiro quartil, mediana, terceiro quartil e valor máximo para os dois melhores algoritmos, da métrica raiz do erro médio quadrático. Estes dados se encontram, respectivamente para o problema de aquecimento e resfriamento, nas Tabelas 9 e 10.

Tabela 9 – Dados dos estatísticos da Raiz do erro médio quadrático para os dois melhores algoritmos relacionados ao problema de aquecimento.

	Floresta Aleatória	*Extreme Gradient Boosting*
Média	0,568	0,445
Desvio Padrão	0,064	0,039
Valor mínimo	0,504	0,394
1º quartil	0,529	0,414
Mediana	0,542	0,457
3º quartil	0,600	0,470
Valor máximo	0,664	0,488

Fonte: O autor

Tabela 10 – Dados dos estatísticos da Raiz do erro médio quadrático para os dois melhores algoritmos relacionados ao problema de resfriamento.

	Floresta Aleatória	*Extreme Gradient Boosting*
Média	1,834	1,323
Desvio Padrão	0,074	0,162
Valor mínimo	1,735	1,099
1º quartil	1,809	1,244
Mediana	1,815	1,354
3º quartil	1,882	1,385
Valor máximo	1,928	1,531

Fonte: O autor

Para ambos os problemas, em média, a raiz do erro médio quadrático é menor para o algoritmo *Extreme Gradient Boosting*, alcançando um valor de 0,445 para a potência de aquecimento e 1,323 para o resfriamento. Em contra partida, o desvio padrão não é mais baixo em ambos os problemas, sendo o mesmo menor para o algoritmo de Floresta Aleatória no problema de resfriamento.

Finalmente, os valores máximos e mínimos relacionados ao segundo algoritmo, para as duas situações, são inferiores em comparação ao primeiro. Este resultado, portanto, sugere que o algoritmo *Extreme Gradient Boosting* como um modelo muito interessante para ambos os problemas.

4.5.4 Resultado da escolha dos Hiper parâmetros

Após a descoberta do melhor modelo, foi necessário ajustar o conjunto de hiper parâmetros do algoritmo. Para isto, o *GridSearchCV* foi utilizado no conjunto de Hiper parâmetros e seus espaços de valores, assim como uma breve descrição, são evidenciados na Tabela 11.

Tabela 11 – Conjunto de Hiper parâmetros e seus espaços de valores.

	Descrição	Lista de valores
n_estimators	Número de árvores utilizadas.	[100, 150, 200, 250]
max_depth	Profundidade das árvores.	[4, 5, 6]
learning_rate	Taxa com que os passos de atualização são feitos.	[0.01, 0.001, 0.05]
subsample	Define a proporção de amostras na qual cada árvore é treinada.	[0.4, 0.5, 1]

lambda	Parâmetro para regularização e, portanto, controle de sobreajuste.	[3, 4]

Fonte: O autor

Os Hiper parâmetros foram escolhidos para os modelos não sofressem Sobreajuste, entretanto, que fosses relativamente complexos para gerar boas métricas.

Finalmente, foi possível obter os resultados da métrica Raiz do erro médio quadrático do ajuste fino para cada problema e, com o método de validação cruzada ao conjunto de treinamento, gerando, portanto, um valor médio da métrica para o conjunto particionado de treinamento e um conjunto de validação criado automaticamente a partir do primeiro. Os resultados se encontram na Tabela 12, discretizados por tipo de problema e conjunto de dados à qual a métrica foi obtida.

Tabela 12 – Resultados da validação cruzada em conjunto com o ajuste fino, para cada problema.

	Modelo de potência de aquecimento	Modelo de potência de resfriamento
Treinamento	0,318	0,542
Validação	0,464	1,311

Fonte: O autor

Os resultados da Tabela 12, mostram que os modelos não aparentam sofre de sobreajuste. Além disso, pode-se inferir que modelar a potência de resfriamento tende a ser mais difícil, dado o valor mais alto da raiz do erro médio quadrático para ambos os conjuntos.

4.5.5 Resultado da modelagem no conjunto de teste

Ao fim da modelagem, foi utilizado modelos escolhidos e refinados ao conjunto intocado de teste, após o retreinamento no conjunto completo de treinamento. Os resultados para ambos os modelos, utilizando a raiz do erro médio quadrado e o Coeficiente de Determinação, no conjunto intocado e no de treinamento, estão presentes nas tabelas 13 e 14, respectivamente.

Tabela 13 – Resultados da Raiz do erro médio quadrático no conjunto de teste e treinamento.

	Modelo de potência de aquecimento	Modelo de potência de resfriamento
Treinamento completo	0,333	0,526
Teste	0,433	1,240

Fonte: O autor

Tabela 14 – Resultados do Coeficiente de Determinação no conjunto de teste e treinamento.

	Modelo de potência de aquecimento	Modelo de potência de resfriamento
Treinamento completo	0.999	0.997
Teste	0.998	0.983

Fonte: O autor

Os resultados em ambas as tabelas, mostram que o erro é baixo e a quantidade de variância explicada pelos modelos é muito alta, evidenciando que bons modelos foram gerados a partir dos dados. Além disso, não há evidência de sobre ajuste, visto que os valores de resultados para os conjuntos de treino e teste estão razoavelmente próximos. Com isso, é possível indicar que com modelos confiáveis para estabelecer valores concretos de potência, sendo utilizados em locais em que o erro do dimensionamento leva ao desperdício de energia, pode ajudar na redução do custo pago em energia e em reduzir o desperdício de consumo.

Finalmente, é possível inferir que a modelagem da Potência de aquecimento aparenta ser melhor dado o conjunto de dados disponível e variáveis selecionadas para cada problema, em comparação com a Potência de resfriamento, visto que as métricas foram melhores.

5 CONCLUSÕES E PROPOSTAS DE CONTINUIDADE

Em meio ao crescente consumo de energia, o presente trabalho buscou trazer uma nova forma de modelar a potência necessária para resfriamento e aquecimento, baseado em características físicas das estruturas. Mostra-se, portanto, que é possível obter estimativas para ajudar em projetos e melhorar a eficiência energética do local utilizando modelos de Aprendizado de Máquina, visto que é uma alternativa interessante, dada a flexibilidade de adaptação e velocidade de se obter resultados confiáveis.

Além disso, é possível concluir que a modelagem das Potências utilizando o algoritmo *Extreme Gradient Boosting* pode ser feita com precisão utilizando as características físicas das estruturas, podendo ser utilizadas para modelar o comportamento e fazer previsões. Contudo, os resultados evidenciaram que modelar a Potência de Resfriamento aparenta ser sutilmente mais complicado, havendo necessidade de maior exploração para identificar as causas desta discrepância.

Finalmente, apesar das limitações de dados, como quantidade e variabilidade, este trabalho pavimenta o início de um caminho que pode ser aperfeiçoado, para melhorar o conhecimento tanto na área da Engenharia Elétrica como em modelagem de dados. Segue, portanto, que o presente trabalho pode ser seguido adiante com as seguintes propostas:

- Coletar mais dados para uma análise mais robusta.
- Buscar novas variáveis para acrescentar informação.
- Utilizar algoritmos não paramétricos mais complexos.
- Utilizar outros tipos de transformações para normalizar as distribuições.
- Aumentar o número do espaço de Hiper parâmetros no ajuste fino. Além de utilizar métodos mais estocásticos para uma varredura mais eficiente.
- Quantificar se há ganho de custo ao fazer uma previsão utilizando o modelo e ao fazer utilizando outro tipo de software.

REFERÊNCIAS

1. CENTRO DE COMERCIALIZAÇÃO DE ENERGIA ELÉTRICA. **Consumo de energia no Brasil cresceu 1,4% no primeiro semestre de 2023, aponta CCEE**. CCEE, 2023. Disponível em: https://www.ccee.org.br/pt/web/guest/-/consumo-de-energia-no-brasil-cresceu-1-4-no-primeiro-semestre-de-2023-aponta-ccee. Acesso em: 17 mai. 2024.

2. PEREZ-LOMBARD, Luiz; ORTIZ, José; POUT, Christine. **A review on buildings energy consumption information, Energy and Buildings**. Energy and Buildings. p. 394-398. 2008.

3. CAI, W.; WU, Y.; ZHONG, Y.; REN, H. **China building energy consumption: situation, challenges and corresponding measures**. Energy Policy. p. 2054-2059. jun. 2009.

4. YAO, Runming.; LI, Baizhan.; STEEMERS, Koen. **Energy policy and standard for built environment in China, Renewable Energy**, Renewable Energy, p. 1973–1988. 2005.

5. EMPRESA DE PESQUISA ENERGÉTICA. **Resenha Mensal: O consumo nacional de energia elétrica foi de 42.837 GWh em janeiro de 2023, crescimento de 0,6% em comparação com mesmo mês de 2022**. EPE, 2023. Disponível em: https://www.epe.gov.br/pt/imprensa/noticias/resenha-mensal-o-consumo-nacional-de-energia-eletrica-foi-de-42-837-gwh-em-janeiro-de-2023-crescimento-de-0-6-em-comparacao-com-mesmo-mes-de-2022. Acesso em:17 mai. 2024.

6. EMPRESA DE PESQUISA ENERGÉTICA. **Anuário Estatístico de Energia Elétrica 2022**. EPE. 2022. Disponível em: https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-160/topico-168/Fact%20Sheet%20-%20Anu%C3%A1rio%20Estat%C3%ADstico%20de%20Energia%20El%C3%A9trica%202022.pdf. Acesso em 18 mai. 2024.

7. LABORATÓRIO DE EFICIÊNCIA ENERGÉTICA. **Conforto Térmico**. labEEE. Disponível em: https://labeee.ufsc.br/pt-br/linhas-de-pesquisa/conforto-termico. Acessado em: 28 mai. 2024.

8. SILVA, Wesley; DUARTE, Matheus; et al. **Investigação da carga térmica de resfriamento de uma edificação residencial multifamiliar por diferentes algoritmos de transferência de calor**. Engineering and Science. ed 12:2, v. 12. 2023

9. TSANAS, Athanasios; XIFARA, Angeliki. **Energy Efficiency**. UCI Machine Learning Repository, 2012. Disponível em: https://archive.ics.uci.edu/dataset/242/energy+efficiency. Acesso em: 25 mai. 2024.

10. TSANAS, Athanasios; XIFARA, Angeliki. **Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools**. Energy and Buildings, v. 49, p. 560-567, 2012.

11. PRESSENLEHNER, Werner; MAHDAVI, Ardeshir. **A building morphology, transparency, and energy performance**. Eighth International IBPSA Conference Proceedings, Eind-hoven. p. 1025–1032. 11 de ago de 2003.

12. BUSSAB, Wilton de O.; MORETTIN, Pedro A. **Estatística Básica**. 9ª ed. Saraiva: São Paulo, 2017.

13. HOAGLIN, David C; MOSTELLER, Frederick; TURKEY, John W. **Understanding Robust and Exploratory Data Analysis**. Wiley-Blackwell: New York, 1983.

14. TURKEY, John W. **Some Graphic and Semigraphic Displays**. In Statistical Papers in Honor of George W. Snedecor: Ames IA. p. 293–316. 1972.

15. VELLEMAN, Paul; DAVID, Hoaglin. **Applications, Basics, and Computing of Exploratory Data Analysis**. Duxbury Press: Boston, 1 de jan de 1981.

16. VISFERREIRA. **File:Boxplot.png**. Wikimedia Commons, 2016. Disponível em: https://commons.wikimedia.org/wiki/File:Boxplot.png. Acessado em: 25 mai. 2024.

17. PLAYFAIR, William. **THE COMMERCIAL AND POLITICAL ATLAS Representing, by Means of STAINED COPPER-PLATE CHARTS, THE PROGRESS OF THE COMMERCE, REVENUES, EXPENDITURE, AND DEBTS OF ENGLAND, DURING THE WHOLE OF THE EIGHTEENTH CENTURY**. The internet archive, 2018. Disponível em: https://archive.org/details/PLAYFAIRWilliam1801TheCommercialandPoliticalAtlas/page/n89/mode/2up. Acessado em: 25 mai. 2024.

18. LOANNIDIS, Yannis. **The History of Histograms (abridged)**. Proceedings: Berlin. p. 19-30. 2003.

19. WEISSTEIN, Eric W. **Normal Distribution**. MathWorld--A Wolfram Web Resource, s.d. Disponível em: https://mathworld.wolfram.com/NormalDistribution.html. Acessado em 26 mai. 2024.

20. HAVIL, Julian. **Gamma: Exploring Euler's Constant.** Princeton University Press: Princeton, New Jersey, 2003.

21. HAUKE, Juan; KOSSOWSKI, Tomasz. **Comparison of values of Pearson's and Spearman's correlation coefficient on the same sets of data**. Quaestiones Geographicae, Poznań. v. 30. p. 87–93. 2011.

22. STINGLER, Stephen. **Francis Galton's account of the invention of correlation**. Statistical Science, v. 4, n. 2, p. 73-79, mai. 1989.

23. XIAO, Chengwei; YE, Jiaqi; ESTEVES, Rui Máximo; RONG, Chunming. **Using Spearman's correlation coefficients for exploratory data analysis on big dataset**. Concurrency and Computation: Practice and Experience. v. 28, p. 3866-3878. 2015.

24. OLIVEIRA, Nicolas. **Correlação monotônica e não monotônica**. Medium, 2024. Disponível em: https://medium.com/@nicolasfaleiros/correla%C3%A7%C3%A3o-monot%C3%B4nica-e-n%C3%A3o-monot%C3%B4nica-1168780b3363#:~:text=Uma%20rela%C3%A7%C3%A3o%20monot%C3%B4nica%20pode%20ser,para%20identificar%20rela%C3%A7%C3%B5es%20n%C3%A3o%2Dmonot%C3%B4nicas. Acessado em: 25 mai. 2024.

25. BROWNLEE, Jason. **How to use StandardScaler and MinMaxScaler Transforms in Python**. Machine Learning Mastery, 2020. Disponível em: https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/. Acessado em: 25 mai. 2024.

26. BROWNLEE, Jason. **Ordinal and One-Hot Encodings for Categorical Data**. Machine Learning Mastery, 2020. Disponível em: https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/. Acessado em: 25 mai. 2024.

27. GÉRON, Aurélien. **Hands-On Machine Learning with Scikit-learn, Keras & Tensorflow**. 2ª ed. Sebastopol: O'Reilly, 2019.

28. BARRET, James P. **The Coeffcient of Determinations - Some Limitations**. The American Statistician, London, v. 28, p. 19-20, 12 mar. 2012.

29. BROWNLEE, Jason. **How to Calculate Feature Importance with Python**. Machine Learning Mastery, 2020. Disponível em: https://machinelearningmastery.com/calculate-feature-importance-with-python/. Acessado em: 31 mai. 2024.

30. BETTIN, Augusto; SILVA, Felipe; et al. **Studies of attribute selection in forescast models: a behavioral analysis**. Revista FT, Rio de Janeiro, v. 26, n. 127, nov. 2023.

31. BROWNLEE, Jason. **Recursive Feature Elimination (RFE) for Feature Selection in Python**. Machine Learning Mastery, 2020. Disponível em: https://machinelearningmastery.com/rfe-feature-selection-in-python/. Acessando em: 31 mai. 2024.

ANEXOS

ANEXO A – https://github.com/caiosoter/DS-Projects/tree/main/Energy_Efficiency