UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA ELÉTRICA
CURSO DE ENGENHARIA ELÉTRICA

FRANCISCO MARCOS GUIMARÃES OLIVEIRA FILHO

APRENDIZADO DE MÁQUINA APLICADO À PREDIÇÃO DE VIABILIDADE TÉCNICA PARA CONEXÃO DE CARGAS EM REDE DE DISTRIBUIÇÃO DE MÉDIA TENSÃO

FORTALEZA
2022

FRANCISCO MARCOS GUIMARÃES OLIVEIRA FILHO

APRENDIZADO DE MÁQUINA APLICADO À PREDIÇÃO DE VIABILIDADE TÉCNICA PARA CONEXÃO DE CARGAS EM REDE DE DISTRIBUIÇÃO DE MÉDIA TENSÃO

Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia Elétrica da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Bacharel em Engenharia Elétrica.
Orientadora: Profa. PhD. Ruth Pastôra Saraiva Leão.

FORTALEZA

2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

O47a Oliveira Filho, Francisco Marcos Guimarães.
Aprendizado de máquina aplicado à predição de viabilidade técnica para conexão de cargas em rede de distribuição de média tensão / Francisco Marcos Guimarães Oliveira Filho. – 2022.
71 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Tecnologia, Curso de Engenharia Elétrica, Fortaleza, 2022.
Orientação: Profa. Dra. Ruth Pastôra Saraiva Leão.

1. Aprendizado de máquina. 2. Análise preditiva. 3. Setor elétrico. 4. Rede de distribuição. 5. Viabilidade técnica. I. Título.

CDD 621.3

FRANCISCO MARCOS GUIMARÃES OLIVEIRA FILHO

APRENDIZADO DE MÁQUINA APLICADO À PREDIÇÃO DE VIABILIDADE TÉCNICA PARA CONEXÃO DE CARGAS EM REDE DE DISTRIBUIÇÃO DE MÉDIA TENSÃO

Trabalho de Conclusão de Curso apresentado ao Curso de Engenharia Elétrica da Universidade Federal do Ceará, como requisito parcial à obtenção do título de Bacharel em Engenharia Elétrica.
Orientadora: Profa. PhD. Ruth Pastôra Saraiva Leão.

Aprovada em: __/__/____.

BANCA EXAMINADORA

_____
Profa. PhD. Ruth Pastôra Saraiva Leão (Orientadora)
Universidade Federal do Ceará (UFC)

_____
Eng. Gabriel Eugênio de Aguiar Silveira
FortBrasil

_____
Eng. MSc. Klendson Marques Canuto
Enel Distribuição Ceará

A Deus.
A toda minha família, especialmente meus pais, Marcos e Didia, meu irmão Gabriel, minha avó Sara, e minha noiva Gilvanira.
Eu dedico.

AGRADECIMENTOS

Em primeiro lugar, agradeço a Deus, que me deu força para trilhar o caminho e superar os obstáculos, além de amparar com bençãos e saúde a mim e minha família.

A minha mãe, Didia, minha primeira grande educadora, que me deu todo seu amor e carinho. Gratidão por seu apoio, amor e por suas preocupações. Sem você, eu nada seria.

A meu pai, Marcos, minha referência como homem, honestidade e humildade. Gratidão por todo suor e trabalho para criar seus filhos. Sem você, não chegaria onde estou.

Agradeço meu irmão Gabriel, minha avó Sara, minha noiva Gilvanira, por todo incentivo, afeto e sempre estar ao meu lado.

Agradeço a Universidade Federal do Ceará e todos os profissionais, por proporcionar um ambiente propício para aprendizagem. Agradeço a todo corpo docente do Departamento de Engenharia Elétrica, essencial no meu processo de formação profissional e na formação como cidadão. Deixo um agradecimento especialmente a minha orientadora, professora PhD Ruth Pastôra Saraiva Leão, pela disponibilidade, orientação, dedicação e paciência na elaboração desse trabalho.

Agradeço a todos aqueles que direta ou indiretamente me apoiaram em todas as minhas etapas.

Por fim, agradeço a disponibilização dos dados e materiais pela Enel. Agradeço a todos os profissionais que me auxiliaram desde o meu estágio até minha efetivação. Em especial, deixo meus agradecimentos à equipe de Planejamento AT/MT pelos ensinamentos que foram fundamentais na minha formação.

“Se você quiser descobrir os segredos do Universo, pense em termos de energia, frequência e vibração.”

Nikola Tesla

RESUMO

O sistema elétrico está sendo demandado por um volume crescente de solicitações de conexão por acessantes de geração distribuída, industriais, comerciais e residenciais que tornam complexo o processo de análise, refletindo em um prolongamento no tempo de resposta. Realizar um mapeamento da rede com uma estimativa de potência máxima para o atendimento torna o processo de decisão do usuário mais simples e eficiente, beneficiando o acessante e a distribuidora. O trabalho proposto apresenta um estudo com foco em predizer a potência disponível para conexão de cargas em qualquer ponto da rede elétrica de média tensão do estado do Ceará sem a necessidade de intervenções da distribuidora. A análise preditiva foi realizada através dos algoritmos de *Random Forest (*RF), *Support Vector Machine* (SVM), Redes Neurais Artificiais (RNA) e *Extreme Gradient Boosting* (XGBoost). Os modelos de aprendizado de máquina foram aplicados em uma base de dados do sistema elétrico e em estudos de fluxo de potência de especialistas da Enel Distribuição Ceará. A definição do melhor algoritmo foi embasada por métricas de desempenho, resultando no XGBoost com uma taxa de acerto geral de 90%. Os resultados demonstraram a eficiência do método na estimativa de potência com os preditores disponíveis nas quatro modelagens obtidas. Um protótipo de mapa interativo foi criado com o XGBoost visando facilitar a interface entre usuário e predições, com duas versões, sendo uma apresentando a visão de todas as barras dos alimentadores e outra mais simples somente com as subestações.

Palavras-chave: Aprendizado de máquina; análise preditiva; setor elétrico; rede de distribuição; viabilidade técnica; cargas elétricas.

ABSTRACT

The electrical system is being demanded by a growing volume of connection requests by users of distributed generation, industrial, commercial and residential, which makes the analysis process complex, reflecting in an extension in the response time. Carrying out a network mapping with an estimate of maximum power for the service makes the user's decision process simpler and more efficient, benefiting the user and the distributor. The proposed work presents a study focused on predicting the power available for connection of loads at any point of the medium voltage electrical network in the state of Ceará without the need for interventions by the distributor. Predictive analysis was performed using Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Networks (ANN) and Extreme Gradient Boosting (XGBoost) algorithms. The machine learning models were applied in an electrical system database and in power flow studies by a specialist from Enel Distribuição Ceará. The definition of the best algorithm was based on performance metrics, resulting in XGBoost with an overall hit rate of 90%. The results demonstrated the efficiency of the method in estimating power with the predictors available in the four models obtained. An interactive map prototype was created with XGBoost in order to facilitate the interface between the user and predictions, with two versions, one presenting a view of all the feeder bars and another simpler with only the substations.

Keywords: Machine learning; predictive analysis; electrical sector; distribution network; technical viability; electrical loads.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANEEL	Agência Nacional de Energia Elétrica
AVT	Análise de Viabilidade Técnica
Btu	Unidade Térmica Britânica (*British Thermal Unit*)
CCEE	Câmara de Comercialização de Energia Elétrica
CEMIG	Companhia Energética de Minas Gerais S.A.
EIA	*United States Energy Information Administration*
EPE	Empresa de pesquisa Energética
FIEC	Federação das Indústrias do Estado do Ceará
GD	Geração Distribuída
html	*HyperText Markup Language*
Icc3Φ	Curto-Circuito Trifásico Simétrico
kV	Kilovolt
kW	Kilowatt
MAE	Erro Médio Absoluto (*Mean Absolute Error*)
MME	Ministério de Minas e Energia
MRT	Monofilar com Retorno por Terra
MSE	Erro Quadrático Médio (*Mean Square Error*)
MVA	Megavoltampère
MW	Megawatt
ONS	Operador Nacional do Sistema Elétrico
PRODIST	Procedimentos de Distribuição de Energia Elétrica no Sistema Elétrico Nacional
RF	Floresta Aleatória (*Randon Forest*)
RNA	Redes Neurais Artificiais
s	Segundo
SVM	Máquina de Vetores de Suporte (*Support Vector Machine*)
XGBoost	Aumento de Gradiente Extremo (*Extreme Gradient Boosting*)

SUMÁRIO

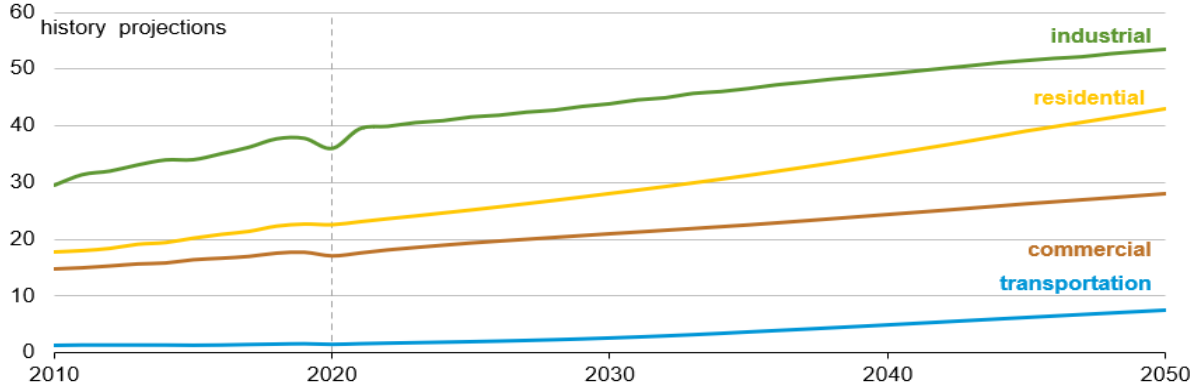
1 INTRODUÇÃO

Este capítulo abrange a descrição do contexto do tema de pesquisa, visando a compreensão do leitor quanto ao propósito do trabalho. A justificativa e a relevância da pesquisa são apresentadas dando o enfoque em todos os envolvidos relacionados com a distribuidora e consumidores. O problema de pesquisa é enunciado de forma a conferir o embasamento prático e intelectual do trabalho dissertativo. Posteriormente, os objetivos gerais do trabalho e seus desdobramentos em objetivos específicos são evidenciados, e finalmente, apresenta-se como a dissertação foi estruturada.

1.1 Contextualização

A sociedade moderna tem apresentado uma tendência de eletrificação, tanto com o crescimento da população mundial, quanto com o aumento da quantidade de equipamentos elétricos utilizados para a melhoria na qualidade de vida. Além disso, a produção de bens e serviços está ligada ao crescimento econômico dos países, então à medida que se desenvolvem há um aumento no consumo de energia (PASTORA *et al.*, 2018).

Figura 1 – Previsão de consumo de energia elétrica em quatrilhão de unidades térmicas Britânicas (Btu).



Fonte: *United States Energy Information Administration, International Energy Outlook* 2021 (IEO2021).

A *United State Energy Information Administration (EIA)* projeta um forte crescimento econômico e populacional, que irá impactar em uma tendência de aumento do consumo de energia em todos os setores até 2050. Nesse estudo, as

residências apresentarão o maior crescimento percentual, seguido de transportes, comércios e por fim indústrias que apesar de percentualmente estar em último, ainda liderará com o consumo total, como mostrado na Figura 1.

As mudanças no consumo de energia são decorrentes de fatores externos específicos em cada setor. As residências serão afetadas por uma expansão na utilização de equipamentos condicionadores de ar, para refrigeração em ambientes quentes e aquecedores em lugares frios. O comércio e as indústrias serão impactados principalmente pelo crescimento populacional, pois naturalmente, todos os bens e serviços precisarão suprir essa nova demanda. Já os transportes são reflexos da eletrificação dos veículos leves e pesados, um novo mercado que deverá ser atendido pelas distribuidoras e que substituirá gradualmente a frota existente (VAN RUIJVEN, DE CIAN e SUE WING *et al*., 2019).

No Brasil, o documento “Previsões de carga para o Planejamento Anual da Operação Energética 2022-2026”, realizado em parceria entre a Empresa de Pesquisa Energética (EPE), o Operador Nacional do Sistema Elétrico (ONS) e a Câmara de Comercialização de Energia Elétrica (CCEE), projeta um crescimento de 17,4% no ciclo entre 2021 e 2026, apresentado na Tabela 1. Segundo o mesmo estudo, o Nordeste apresentará um crescimento de 19,3%, percentual este, maior do que a média nacional para o mesmo período, totalizando 2.217 MW de incremento de carga.

Tabela 1 – Previsão de carga de energia para o ciclo 2022-2026.

Carga de energia (MWmédios) Planejamento Anual 2022-2026						
Subsistema	**2021**	**2022**	**2023**	**2024**	**2025**	**2026**
Norte	5.984	6.413	6.835	7.062	7.262	7.633
Nordeste	11.473	11.791	12.223	12.677	13.158	13.690
Sudeste/CO	39.888	40.782	42.088	43.386	44.773	46.127
Sul	12.130	12.388	12.802	13.230	13.686	14.154
SIN	**69.475**	**71.373**	**73.948**	**76.355**	**78.880**	**81.604**

Fonte: EPE (Previsões de carga para o Planejamento Anual da Operação Energética 2022-2026).

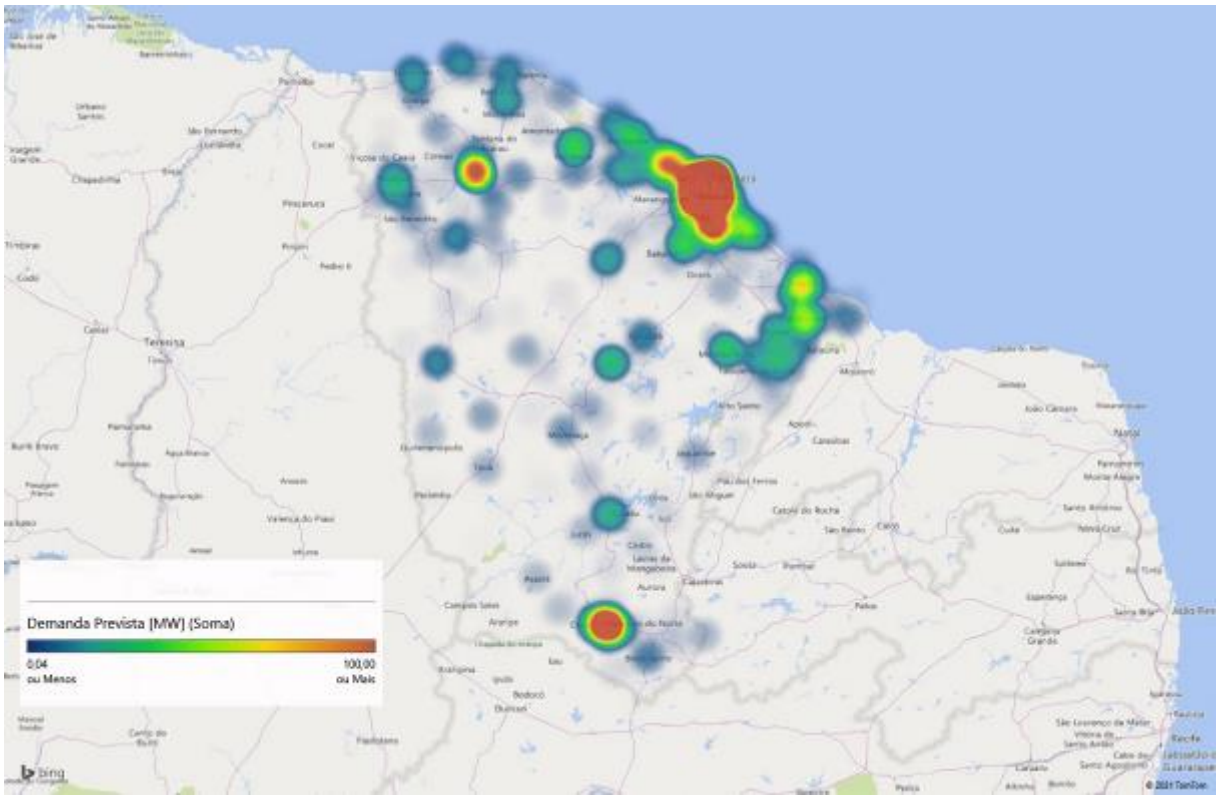
Em 2020, o estado do Ceará apresentou 2,5% de participação no consumo total de energia do Brasil, de acordo com o Anuário Estatístico de Energia Elétrica 2021 publicado pela EPE. Considerando apenas o Nordeste, o estado consumiu cerca

de 14,7% da energia anual, atrás apenas de Pernambuco e Bahia, com percentuais de 17,4% e 30,6%, respectivamente.

No estudo socioeconômico “Rotas Estratégicas Setoriais 2025”, elaborado pela Federação das Indústrias do Estado do Ceará (FIEC), o Ceará possui cenários de potencial de crescimento frente ao nordeste em diversos setores, entre eles o consumo de energia elétrica, impactado pelo crescimento econômico do estado.

A Figura 2 retrata o histórico dos pedidos de ingresso ou acréscimo de novas cargas no estado de Ceará ao setor de Planejamento da Rede AT/MT da Enel Distribuição Ceará, entre os anos de 2014 e 2021. Nessa figura temos a localização e a soma das demandas nas áreas adjacentes.

Figura 2 – Mapa de calor da demanda prevista por região do estado do Ceará.



Fonte: elaborado pelo autor.

O volume dessas solicitações, seja de acréscimo ou novas conexões, somado a mudança no comportamento das cargas e ingresso de gerações, reflete em um aumento na complexidade das análises do sistema elétrico (RAVADANEGH et al., 2014). O crescimento no montante e a dificuldade das análises originou uma demanda exaustiva de estudo do fluxo de potência, impactando o processo de emissão da

Análise de Viabilidade Técnica (AVT). O AVT é um documento que requer um nível de responsabilidade e especialização técnica maior, pois atesta a conexão com ou sem intervenções na rede elétrica, impactando financeiramente o empreendimento e a distribuidora.

Esse processo tem demandado um tempo maior para estudo, impactando diretamente no prazo para emissão, ocasionando transtorno aos clientes, por conseguinte, posterga a decisão de ligação do empreendimento. Ademais, não há dados disponibilizados pelas distribuidoras de forma proativa, de modo que, os clientes submetem incontáveis pontos e variações de cargas, buscando o melhor ponto para conexão, idealmente sem custo algum. Deste modo, existem solicitações desnecessárias ou redundantes, ocasionando em sobrecarga para o fluxo do processo.

Partindo das previsões de crescimento de cargas do Ceará e da complexidade necessária para analisar o impacto na rede, foi possível realizar a contextualização do tema de pesquisa que estabelece uma metodologia de previsão da capacidade de conexão de cargas sem a necessidade de intervenções, em qualquer ponto do sistema elétrico, por meio de algoritmos de aprendizado de máquina. O próximo tópico visa elucidar sobre a relevância da pesquisa desenvolvida e justificar sua importância.

1.2 Justificativa

A entrega em curto prazo em um mercado competitivo melhora a imagem da empresa (ASADZADEH et al., 2011), o que não é possível para as empresas distribuidoras de energia elétrica, visto a complexidade e proporção das solicitações, afetando assim notadamente o tempo de análise e por consequência os processos envolvidos. Pelo lado do acessante, a falta de dados e conhecimento sobre a rede elétrica para detectar os pontos de conexão sem a necessidade de intervenção, ocasiona um aumento das solicitações. Cerca de 25% dos pedidos de AVT, realizados no período entre 2014 e 2021, têm cunho prospectivo, ou seja, para identificar a viabilidade econômica do empreendimento no ponto selecionado.

Verifica-se, desta maneira, que após o cenário exposto sobre conexão de cargas, as distribuidoras apresentam dificuldades em atender com celeridade e qualidade os acessantes. Portanto, esse é um tema de extrema relevância para todas

as classes de consumo e seus impactos afetam diretamente os empreendimentos e a sociedade de consumo, fazendo com que a questão levantada se transforme em um problema de pesquisa para a academia que desenvolve esforços para encontrar soluções e elucidar caminhos para modelar e controlar esse problema. Deste modo, a justificativa deste trabalho foi identificar que a complexidade e interesse do acessante em ter dados sobre a rede poderiam se tornar uma oportunidade de apresentar uma proposta inovadora, com o propósito de contribuir na inserção otimizada de cargas na rede elétrica.

1.3 Definição do Problema

Apesar da importância e dos impactos da disponibilização proativa de dados, pouca pesquisa sistemática tem sido realizada. Existe uma preocupação e um grande esforço dos gestores em reduzir o tempo de emissão das análises de rede, mas pouca pesquisa com foco em estimar antecipadamente a carga ótima para o ponto de conexão. Com o aumento da concorrência no mercado e das complexidades dos estudos de impactos do ingresso de novos tipos de cargas no sistema elétrico, exigindo soluções que sejam cada vez mais precisas e eficientes.

Diante da relevância, dos impactos, e dos métodos propostos para conexão de cargas no sistema elétrico, apresentam-se como problemas de pesquisa traduzidos como perguntas:

1. Como os métodos de aprendizado de máquina podem ajudar na estimativa de potência de conexão de cargas, de tal forma que não exista a necessidade de intervenção na rede elétrica?
2. De que forma é possível utilizar essa estimativa a fim de melhorar o processo de ampliação ou conexão de novas cargas?

Segundo Gil (2002), o problema de pesquisa deve ser claro, preciso, empírico, suscetível à solução e delimitado a uma proporção viável. O problema de pesquisa é delimitado ao estudo de caso da coleta de dados da empresa Enel Distribuição Ceará, o que deixa o problema disponível para a avaliação. Além disso, o problema de pesquisa é empírico e suscetível de solução por propor soluções mediante aplicação do método de aprendizado de máquina.

A solução do problema de pesquisa mostra sua dupla importância, na academia apresenta-se como um estudo específico de uma empresa de nível multinacional que atua em diversas áreas do setor elétrico, sendo um trabalho original que avalia as análises de conexão de cargas realizadas por especialistas da Enel Distribuição Ceará para encontrar padrões por métodos de aprendizado de máquina. Na sociedade em geral, mostra-se como uma importante metodologia de pesquisa para ajudar a tomada de decisões e que tem impactos nos custos de implementação e desenvolvimento dos empreendimentos.

1.4 Objetivos

Objetivo geral do trabalho é desenvolver um algoritmo capaz de predizer a demanda máxima em cada ponto do sistema elétrico de concessão da Enel Distribuição Ceará de forma que não sejam necessárias intervenções para conexão do cliente. O delineamento metodológico aplicado à pesquisa para alcançar o objetivo do trabalho inclui a análise preditiva e o aprendizado de máquina. A metodologia e os resultados são validados a partir de métricas de análise de desempenho dos modelos de aprendizagem propostos.

O objetivo geral se desdobra nos objetivos específicos:

- Determinar as variáveis que mais sensibilizem o algoritmo baseado em aprendizado de máquina e os dados operacionais da rede que serão utilizados no treinamento do algoritmo;
- Explanar conceitos e métodos da análise preditiva e o aprendizado de máquina;
- Verificar os métodos com melhores resultados a partir de análises de desempenho do modelo;
- Evidenciar os resultados encontrados, pela aplicação dos métodos de aprendizado de máquina e análise preditiva, frente a discussões de pesquisa.

1.5 Estrutura do Trabalho

Após conceituar os pressupostos de contextualização, a justificativa e relevância do tema, os objetivos gerais e específicos, a organização da dissertação é apontada:

Incialmente, foi redigida uma breve contextualização com o objetivo de direcionar a respeito da conexão e previsão de novas cargas, e como problema de pesquisa, com foco em predizer a demanda máxima sem a necessidade de intervenção em cada ponto do sistema elétrico de concessão da Enel Distribuição Ceará (Capítulo 1). Busca-se nessa introdução apresentar a relevância da pesquisa desenvolvida e como o método proposto (aprendizado de máquina) e seus impactos auxiliam os especialistas da distribuidora e usuários sobre as tomadas de decisões. Os objetivos gerais e específicos são apresentados visando orientar o processo de pesquisa.

No capítulo 2 é feita uma revisão bibliográfica abordando alguns temas centrais como conexão de consumidores no sistema elétrico e mapas de capacidade da rede, tendo como principal objetivo apresentar estudos de outros autores e suas contribuições, com suas semelhanças e diferenças com esse trabalho.

O capítulo 3 exibe inicialmente as etapas que antecedem a construção de um modelo de aprendizado de máquina. Um foco é dado para quatro métodos, Floresta Aleatória (*Random Forest* - RF), Máquinas de Vetores de Suporte (*Support Vector Machine* - SVM), Redes Neurais Artificiais (RNA) e Aumento de Gradiente Extremo (*Extreme Gradient Boosting* - XGBoost) por apresentar bons resultados com problemas similares ao exposto pela dissertação. As métricas de desempenho junto aos modelos de validação são expostas, além das ferramentas de aplicação do aprendizado de máquina.

O capítulo 4 descreve a metodologia aplicada à pesquisa tendo base os modelos de análise preditiva e de aprendizado de máquinas. O capítulo se propõe a elucidar sobre as decisões tomadas para solucionar o problema de pesquisa.

O capítulo 5 apresenta os resultados e discussões a partir dos procedimentos de análise realizados, tendo a fundamentação teórica da dissertação, os objetivos de pesquisa e a metodologia aplicada como suporte. O capítulo tem a finalidade de apresentar os resultados do modelo preditivo aplicado expondo como as questões direcionadoras da dissertação foram resolvidas.

Por fim, as conclusões da pesquisa onde são exibidas as principais contribuições, os pontos pertinentes do conhecimento adquirido na elaboração do trabalho, as adversidades encontradas e as sugestões para trabalhos futuros. Posteriormente, expõe-se a bibliografia.

2 REVISÃO BIBLIOGRÁFICA

Este capítulo tem o objetivo de abordar artigos, pesquisas e trabalhos acadêmicos de maior relevância utilizados para o desenvolvimento deste estudo. A princípio é feita uma contextualização dos trâmites para conexão de clientes no sistema elétrico. Ao final do capítulo serão apresentados mapas disponibilizados de forma online pelas distribuidoras para avaliação da capacidade para conexão na rede elétrica no mundo.

2.1 Conexão de Consumidores no Sistema Elétrico

Em primeiro de janeiro de 2022 entraram em vigor as resoluções normativas 1.000/2021 e 956/2021, ambas publicadas pela Agência Nacional de Energia Elétrica (ANEEL). A Resolução Normativa 1.000/2021 consolida as principais regras da Agência para a prestação do serviço público de distribuição de energia elétrica, além de dispor os direitos e deveres dos consumidores e demais usuários. Já a Resolução Normativa 956/2021 estabelece os Procedimentos de Distribuição de Energia Elétrica no Sistema Elétrico Nacional – PRODIST.

Para obter acesso ao sistema de distribuição, o consumidor precisa atender aos requisitos mínimos estabelecidos pelo Módulo 3 do PRODIST, entre eles os principais pontos são:

- Padronização nas normas vigentes em âmbito nacional e local;
- Definição da carga prevista do empreendimento;
- Cálculo de demanda;
- Dimensionamento dos condutores;
- Cálculo de Perdas, Variação de tensão, Queda e elevação de tensão no ponto de conexão;
- Sistema de aterramento;
- Para-raios;
- Subestação;
- Definição de Proteções;
- Arranjo geral (Plantas, cortes, detalhes e lista de material);
- Sinalização etc.

Os itens supracitados devem estar presentes em um projeto elétrico, juntamente com as demais definições previstas em normas e com os documentos necessários para ser iniciado o processo de conexão na distribuidora. Conforme estabelecido pela Resolução Normativa 1.000/2021, as etapas e prazos possuem variações de acordo com o nível de tensão de atendimento, apontados na Tabela 2.

Tabela 2 – Etapas e prazos para conexão segundo Resolução Normativa 1.000/2021.

#	ETAPAS / PRAZOS	NÍVEL DE TENSÃO		
		BT	MT	AT
1	Aprovação Prévia de Projeto (conforme NT da distribuidora)	30 dias / 10 dias úteis (reanálise)		
2	Consulta e Entrega de Orçamento Estimado (opcional)	30 dias		
3	Pedido de Conexão	---		
4	Aceite / Rejeição do Pedido e entrega de protocolo	5 dias úteis		
5	Análise Distribuidora (alternativas) Entrega do Orçamento Prévio Técnico Comercial	15 dias: sem obra 30 dias: com obra		45 dias
6	Aprovação do Orçamento	10 dias úteis		
7	Assinatura de Contrato + Pagamento	30 dias		
8	Realização de Obras pela Distribuidora	60 dias	- Até 1km: 120 dias ou - Cronograma com prazo até 1 ano	cronograma
9	Vistoria e Instalação Medição	5 úteis	10 úteis	15 úteis

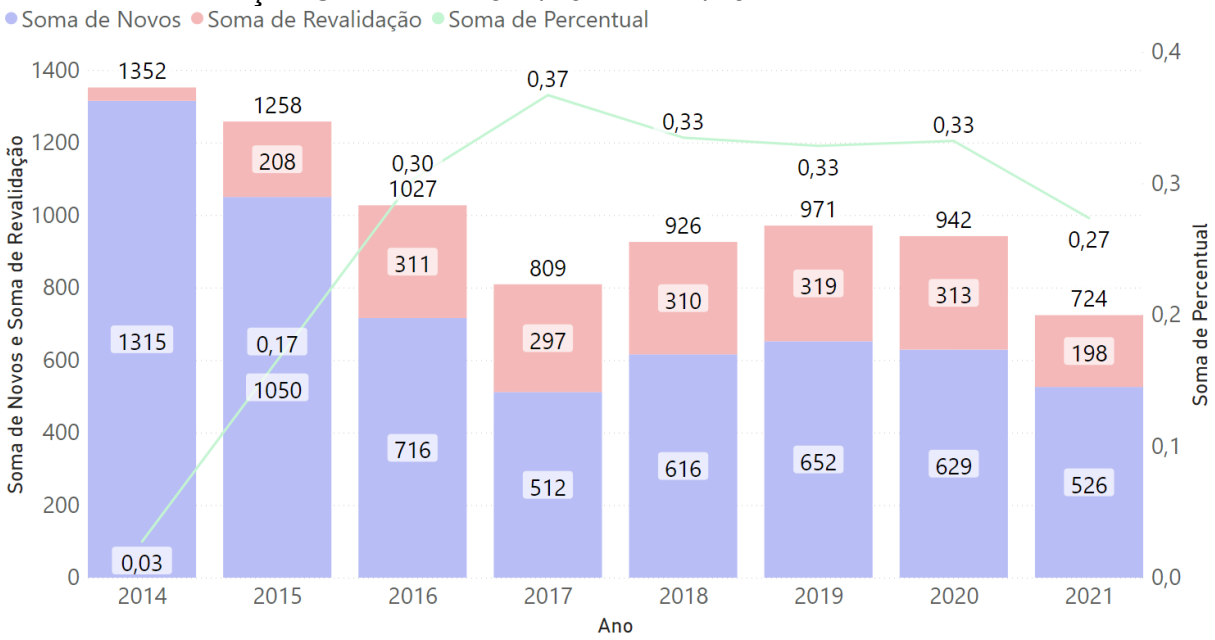
Fonte: elaborado pelo autor.

Antes de firmar contrato com a distribuidora o cliente tem a possibilidade de prospectar o local de conexão, item 2 da Tabela 2, e receber um orçamento estimado para verificar se será preciso intervir na rede (Resolução Normativa 1.000/2021, ANEEL). Em caso positivo poderá impactar financeiramente e aumentar os custos de implantação do empreendimento na região selecionada, dado que o cliente é responsável uma parte do valor da obra e como empreendedor busca o menor investimento para maximizar seus ganhos. Realizará sondagens até encontrar um ponto da rede que suporte sua conexão com o custo mais viável. A Figura 3 apresenta o impacto dessas prospecções na Enel Distribuição Ceará, de clientes que solicitaram a prospecção do mesmo empreendimento em diversos locais ou variações de demanda no mesmo local.

Com a definição do ponto de conexão, documentação, projeto e protocolos, ocorrerá a validação e na condição de estar em acordo com as normas, a aprovação

se dará nos prazos dos tópicos 1 e 4 da Tabela 2. Dando continuidade ao processo, são avaliados os impactos na rede elétrica, proteções e indicadores de qualidade e continuidade, com o ingresso ou acréscimo de carga/geração no ponto de conexão informado, visando sempre o menor custo global tanto para o cliente, distribuidora e sociedade. Na sequência, ocorre as etapas de aprovação do orçamento em campo, assinatura do contrato por ambas as partes, e pagamento conforme estipulado pela participação do cliente na obra. Por fim, com a obra realizada, a conexão é vistoriada e instalada a medição para faturamento (Res.no.1.000/2021, ANEEL).

Figura 3 – Quantidade de solicitações e retrabalho da Enel Distribuição Ceará entre Jan/2014 e Set/2021.



Fonte: elaborado pelo autor.

2.2 Mapas de Capacidade da Rede

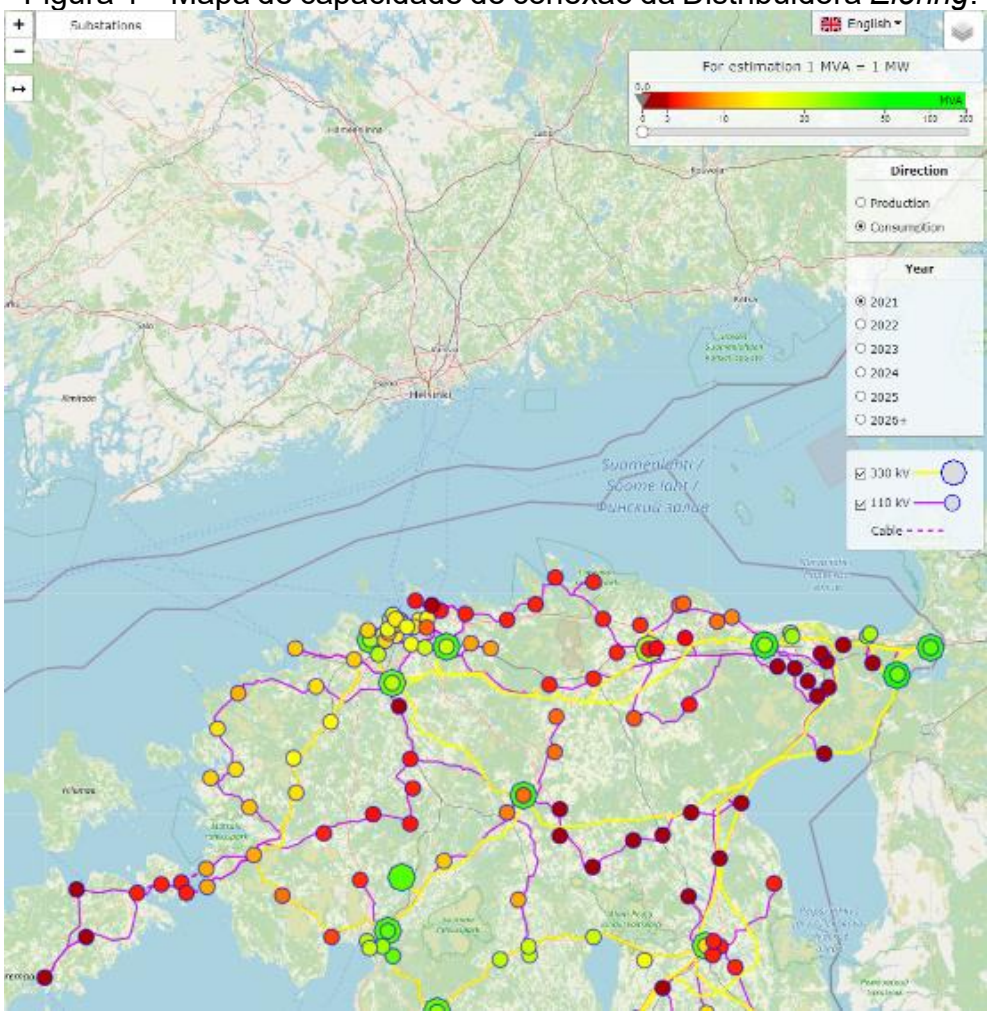
Frente à grande demanda e complexidade das solicitações, as distribuidoras de alguns países estão desenvolvendo e disponibilizando novas ferramentas de análise, com o objetivo de fornecer aos acessantes uma visão geral sobre prováveis obras necessárias, bem como, limitações de potência para suprimento e injeção na rede. Essas ferramentas permitem delimitar e encontrar áreas com maior potencial de conexão e menores restrições, deste modo os clientes tendem a se conectar em regiões que sejam robustas, com alta capacidade de atendimento, por ter acesso a informações sobre o sistema elétrico de forma preliminar. Para a distribuidora, esse

comportamento facilita a previsão dos locais de crescimento da rede, podendo se antecipar e direcionar os clientes para zonas de sua concessão com a construção ou melhoria do sistema de distribuição.

As buscas foram voltadas aos trabalhos que tratam a disponibilidade de dados prévios para a conexão de cargas e/ou gerações, em um escopo global de países ou regiões de concessão, onde quatro fontes de pesquisas publicaram mapas interativos.

Na Estônia, a distribuidora *Elering* desenvolveu uma aplicação chamada "*E-Gridmap*", apresentado na Figura 4, que mostra as capacidades aproximadas de conexão em subestações de 110kV e 330kV, de forma estimada e que não substitui a formalização do pedido de conexão. Além disso, apresenta os custos estimados adicionais para acomodar o produtor ou consumidor de acordo com a capacidade desejada e ano de entrada, mas com a limitação de 100MVA na rede de 110kV e 300MVA e na rede de 330kV.

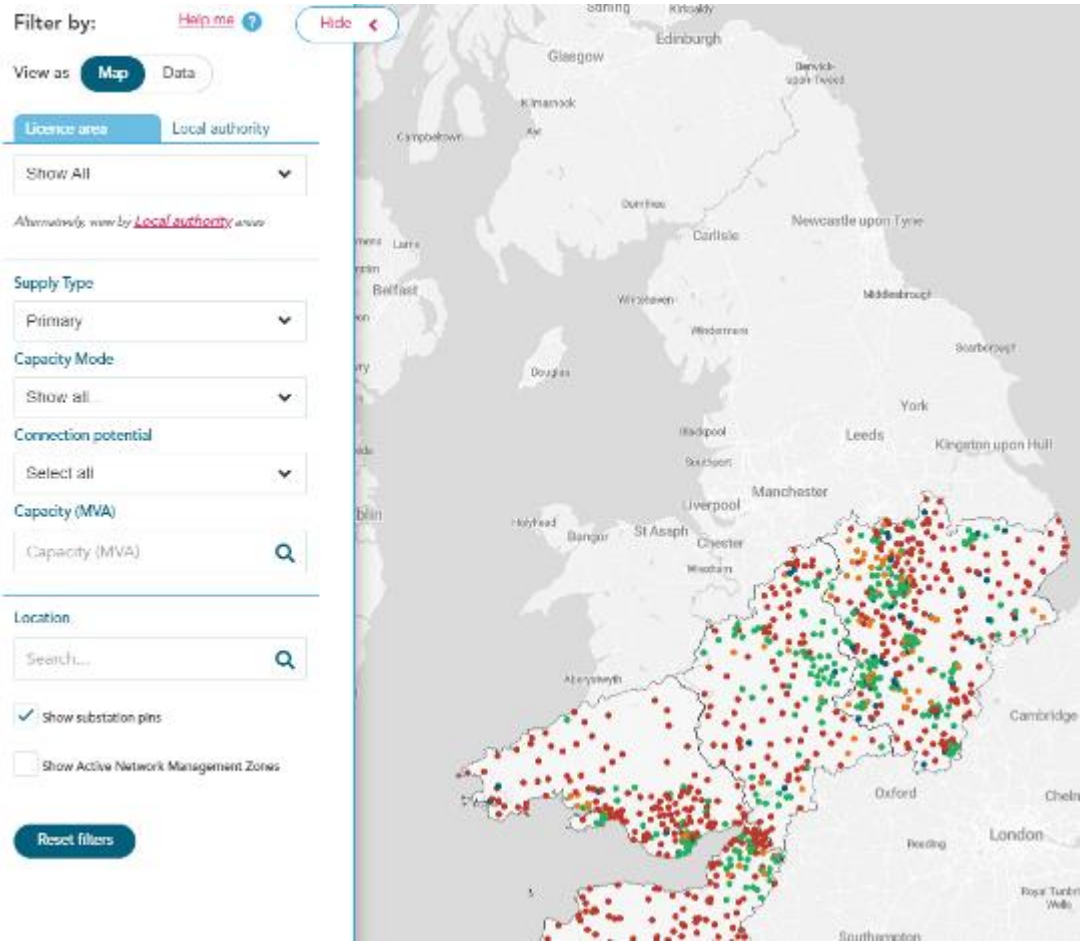
Figura 4 – Mapa de capacidade de conexão da Distribuidora *Elering*.



Fonte: *Elering*, 2021.

Na Inglaterra, a empresa *Western Power Distribution* criou um mapa de capacidade da rede, mostrado na Figura 5, com o objetivo de fornecer uma indicação para conexão de empreendimentos no sistema de transmissão ou distribuição. O mapa apresenta uma gradação de cores que representam o estado em que se encontram as subestações, aquelas na cor azul não é possível suprir o atendimento, em vermelho para as que precisam de intervenções na rede, em amarelo para atender no limite da capacidade e verde para o restante. Outros filtros podem ser utilizados, como a definição do nível de atendimento, capacidade e o tipo (carga ou geração). Semelhantemente, os dados fornecidos não se destinam a ser confiáveis conforme consta no site e o cliente deve ingressar uma solicitação para encontrar os dados corretos da rede.

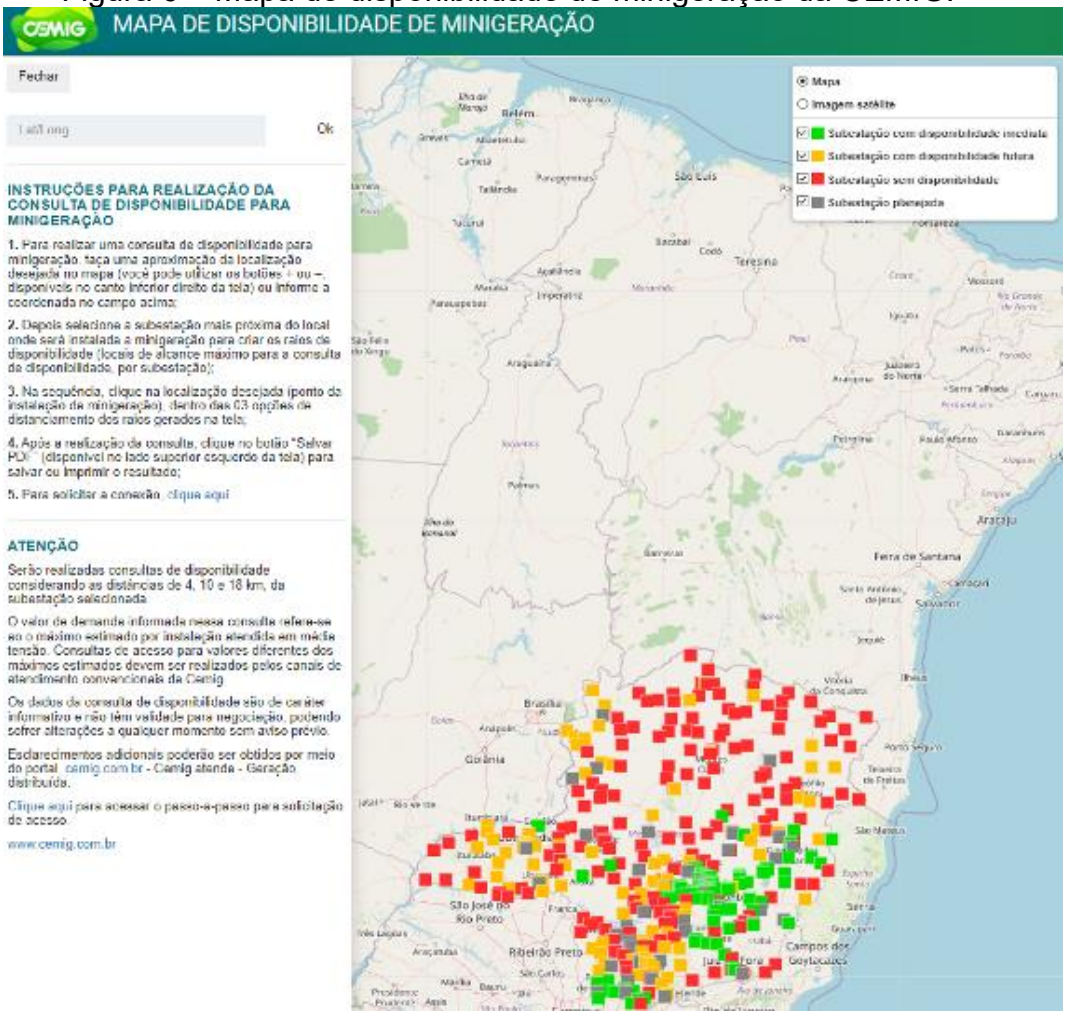
Figura 5 – Mapa de capacidade da rede da *Western Power Distribution*.



Fonte: *Western Power Distribution*, 2022.

No Brasil, a Companhia Energética de Minas Gerais (CEMIG) produziu uma ferramenta em seu portal, exibido na Figura 7, com o intuito de informar a disponibilidade de ligação para novas conexões de minigeração distribuída. O mapa é uma parceria da distribuidora com o Governo do Estado para indicar a capacidade da rede elétrica na área de concessão da CEMIG. A plataforma classifica quatro cores nas subestações de acordo com a disponibilidade de cada uma, de forma que verde descreve subestações com disponibilidade, amarelo está condicionada a uma obra estruturante, vermelha não há capacidade disponível e cinza para subestações planejadas para construção. A ferramenta tem o objetivo de agilizar as conexões, diminuir a quantidade e prazos de execução de obras e identificar o cenário de cada ponto, mas os dados emitidos são de caráter meramente informativo.

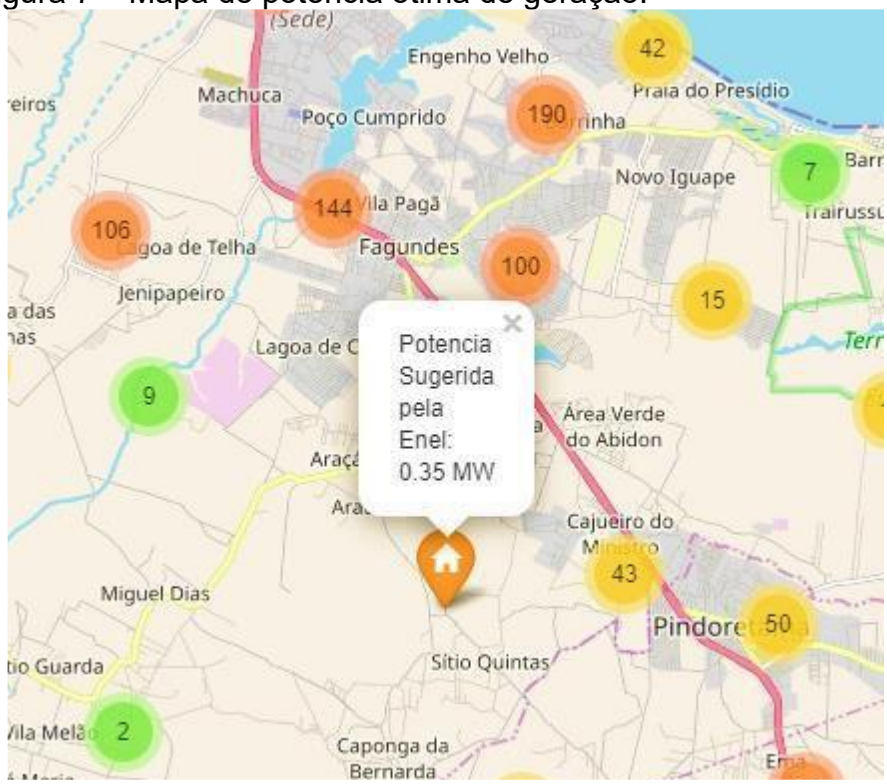
Figura 6 – Mapa de disponibilidade de minigeração da CEMIG.



Fonte: CEMIG, 2022.

No Ceará, o tema de análise prévia de geração foi abordado por Silveira (2021), com um mapa de todas as barras de média tensão em que cada ponto apresenta uma potência ótima para conexão de geração sem a necessidade de intervenção da distribuidora, utilizando aprendizado de máquina. Na Figura 7, podemos observar o resultado do trabalho com uma interface projetada para o consumidor navegar pelo mapa do Ceará e verificar a viabilidade do projeto no ponto escolhido.

Figura 7 – Mapa de potência ótima de geração.



Fonte: SILVEIRA, 2021.

Em outros trabalhos são aplicados estudos de fluxo de potência para prever a capacidade de cada ponto para carga ou geração, mas novas abordagens podem ser propostas para agilizar o processo e tornar a informação mais eficiente com a utilização de aprendizado de máquina aplicado a consumidores. Identificado assim o problema, o próximo capítulo tem o objetivo de aprofundar os conhecimentos técnicos da análise preditiva e dos métodos de aprendizado de máquina. Esse estudo explorará os procedimentos fundamentais da análise preditiva, os algoritmos aplicados a essa dissertação, as ferramentas utilizadas para a aplicação dos métodos e as métricas de desempenho dos modelos propostos.

3 ANÁLISE PREDITIVA

Este capítulo aborda os métodos de aprendizado de máquinas aplicados à análise preditiva, assim como seus procedimentos técnicos. São detalhados os passos de pré-processamento e as técnicas de aprendizado de máquina que abrangem o RF, SVM, RNA e XGBoost. As métricas de desempenho dos modelos expostos, bem como suas ferramentas de aplicação, são apresentadas como instrumentos de avaliação e execução.

3.1 Aprendizado de Máquina e Análise Preditiva

Junto com era da tecnologia digital surgiram diversos problemas para quantificar e entender os dados estruturados e não estruturados. Visando contornar o desafio de extrair informação de interesse a partir de grande volume de dados, foram elaborados modelos de aprendizado de máquina e a criação da computação de alto desempenho. Com o objetivo de realizar previsões, um ramo da inteligência artificial foi aprimorado, o aprendizado de máquina, de forma a adquirir conhecimento de uma base de dados por meio de autoaprendizado (NILSSON, 1998).

A construção de um algoritmo utilizando aprendizado de máquinas supervisionado para tomadas de decisões, deve partir de uma entrada de dados pré-processados, aplicados a algoritmo selecionado para realizar uma predição e por fim avaliar o desempenho do modelo proposto. A primeira etapa, o pré-processamento, engloba a extração e seleção de variáveis, redução de dimensionalidade e exclusão de *outliers*. A segunda etapa, aprendizado, abrange a seleção de modelos, validação cruzada, métricas de desempenho e otimização dos hiperparâmetros. Por fim esse fluxo é rodado de tal forma que a predição do modelo final obtenha a melhor avaliação para o modelo selecionado (RASCHKA e MIRJALILI, 2017).

Além do processo de validação do modelo, o desempenho do modelo pode ser melhorado ainda mais por um especialista no problema, por meio de remoção de dados irrelevantes ou incoerentes, adicionando dados faltantes ou tratando os dados existentes que afetam diretamente a performance do modelo (KUHN e JOHNSON, 2013).

3.2 Pré-Processamento dos Dados

O pré-processamento é a etapa inicial no processo de implementação do algoritmo de aprendizagem, sendo fundamental no impacto no desempenho inicial do modelo, em razão do dado bruto necessitar de uma inclusão, exclusão ou transformação dos dados. A adição de dados considerados relevantes, remoção de ruídos e a transformação do tipo de dados seja categórica ou numérica, pode resultar em uma interação mais rápida e mais precisa do modelo (KUHN e JOHNSON, 2013).

O processo de extração ou montagem da base de dados pode resultar em ausências no conjunto, existindo duas técnicas principais para tratamento dessas inconsistências, a remoção e a interpolação. A primeira técnica remove os dados faltantes e a segunda acrescenta um valor relacionado com os dados presentes. A aplicação delas deve ser escolhida com cuidado, em virtude do impacto da remoção ou adição de um valor em um preditor relevante que resulta em variações na assertividade do modelo (RASCHKA e MIRJALILI, 2017).

A amplitude, variação e quantidade de dados selecionados também afetam diretamente o seu rendimento e caso não sejam tratados ou testados podem enviesar o resultado, logo três ferramentas são utilizadas para resolver essas características, a normalização, padronização e a redução de dados. A normalização transforma conjuntos de dados em valores entre 0 e 1, onde 1 é o maior e 0 o menor valor. A padronização coloca a média dos dados em 0 e o desvio padrão em 1. Já a redução de dados é uma técnica que visa reduzir a quantidade de preditores com alta correlação entre si, ou seja, apresentam características semelhantes e a presença de um único preditor com essas características já seria suficiente (KUHN e JOHNSON, 2013).

Observando as características dos dados, quando um valor de um preditor está drasticamente distante do restante, define-se como um *outlier*. Esse tipo de valor foge da normalidade e pode causar anomalias para a escolha do preditor, interferindo na normalização, padronização e na estruturação da modelagem. Esses pontos fora da curva precisam ser detectados e analisados por meio de ferramentas gráficas ou estatísticas, de modo que permitam avaliar o impacto da remoção frente ao tamanho da amostra e no resultado do modelo (KUHN e JOHNSON, 2013).

Diversas modelagens necessitam que todos os preditores sejam do tipo numérico, porém existem preditores importantes para o modelo que são do tipo

categórico e para esse tipo de situação os dados categóricos são transformados em variáveis fictícias do tipo *dummy*. Esse tipo de variável transforma cada categoria em um valor, e a quantidade de valores disponíveis para aplicação é igual à quantidade de categorias do preditor original menos um (KUHN e JOHNSON, 2013).

Com a base e preditores devidamente tratados, na etapa anterior à aplicação do modelo é aplicada uma separação de forma aleatória dos conjuntos de dados para parte destes servir no processo de treinamento e o restante para teste (RASCHKA e MIRJALILI, 2017). Pela base de treino, o modelo é treinado para obter o menor erro, na sequência a base de teste é aplicada ao modelo e validada, comparando os resultados de modo que a performance em ambos os passos deve ser semelhante. O *overfitting* ocorre quando o modelo apresenta um bom desempenho para a base de treino, mas com outra entrada não é efetivo (HASTIE, TREVOR, TIBSHIRANI e FRIEDMAN, 2017), em outras palavras, é como se o modelo se especializasse nos valores do teste, ficando “sobreajustado” a esses dados e não consegue generalizar para outras entradas.

A validação cruzada é uma técnica utilizada para avaliar se um modelo possui capacidade de generalizar dado um conjunto de informações. A técnica consiste na divisão do conjunto de dados em novos subconjuntos de dados, que pode ser utilizado para treinar o modelo ou testar, de modo que são realizados vários ciclos de validação para que o modelo tenha uma maior generalização. Essa abordagem atesta que problemas de *overfitting* não sensibilize o modelo, além de garantir uma reamostragem para dados considerados insuficientes (HASTIE, TREVOR, TIBSHIRANI e FRIEDMAN, 2017).

3.3 Técnicas de Aprendizagem

Esse tópico tem como objetivo detalhar os quatro modelos de técnicas de aprendizagem, escolhidos para a aplicação desejada.

3.3.1 Random Forest (RF)

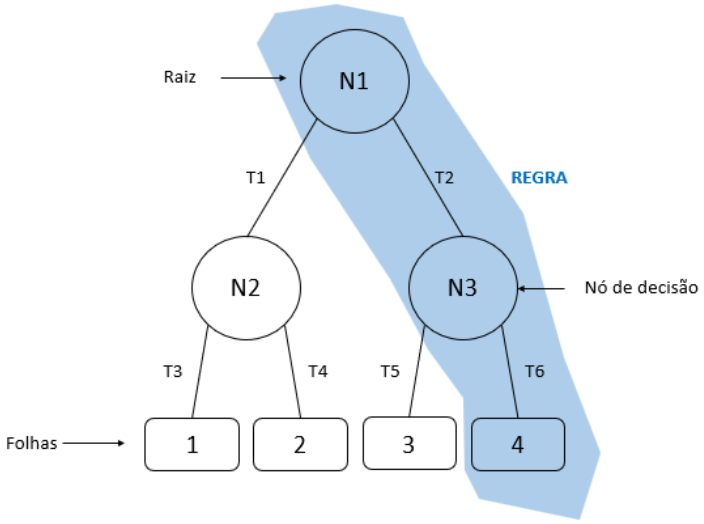
Para entender o algoritmo RF é necessária uma breve explicação sobre árvores de decisão (*Decision Tree*). A árvore de decisão é um algoritmo que se baseia

no entendimento de divisão de grupos homogêneos de modo a tornar um problema complexo em diversos pedaços mais simples para tomar uma decisão de classificação ou regressão.

Na Figura 8, temos as partes que constituem e representam o funcionamento do algoritmo. A raiz é o começo do algoritmo, os nós são testes de condição e as folhas são as respostas, onde os dados saem da raiz e passam pelos nós em direção às folhas. Cada nó é submetido a uma condição, que caso seja atendida, irá para outro nó e em caso contrário irá para o nó que atende, a sequência é repetida até alcançar as folhas, onde estão agrupados dados semelhantes (SHI, 2007, p. 2-5).

O RF é um conjunto de árvores de decisão, formando deste modo, uma floresta, em que cada árvore é única e descrita como um preditor. O algoritmo então distribui conjuntos de dados aleatórios pelas árvores e sua decisão será determinada pela contagem de votos dos componentes preditores. O uso desse algoritmo reduz a variância e tem um maior poder de generalização, evitando assim o *overfitting*, atribuindo precisão e robustez a ruídos se comparado a outros métodos. Por outro lado, quanto maior o número de árvores, maior a capacidade computacional para processar os dados e construir o modelo (BREIMAN, 2001).

Figura 8 – Arquitetura do algoritmo de árvore de decisão.



Fonte: elaborado pelo autor.

Em problemas que utilizam RF para classificação, o retorno do algoritmo é uma previsão da categoria com maior número de contagem de árvores. Em regressão,

a previsão será a média de todas as árvores presentes no modelo (SVETNIK et al., 2003).

3.3.2 Extreme Gradient Boosting (XGBoost)

O XGBoost é conhecido como um dos algoritmos de aprendizado de máquinas de melhor desempenho utilizados para aprendizado supervisionado, sendo o preferido tanto para classificação como para regressão devido à sua alta velocidade de execução.

Chen e Guestrin (2016) sugeriram o XGBoost como um método alternativo para prever uma saída dadas certas variáveis. O ponto principal do algoritmo é que ele constrói árvores de decisão, de modo que cada nova árvore seja treinada usando os resíduos de árvores anteriores. Em outras palavras, o novo modelo corrige os erros do passado para prever a saída.

O XGBoost pode ser entendido como uma agregação e melhoria de diversas técnicas como RF, *Boosting* e o *Gradient Boosting*.

- *Random Forest*: O RF foi um tema abordado anteriormente, e nesse caso sua utilização é semelhante.
- *Boosting*: Essa é uma técnica de aprendizado de máquina que reduz a variação dos dados. O algoritmo atribui pesos maiores para dados classificados corretamente e menos peso para dados classificados incorretamente, ponderando novamente a cada rodada de testes (DHALIWAL; NAHID; ABBAS, 2018, p 7-8).
- *Gradient Boosting*: Esse é algoritmo de reforço em que os erros são minimizados, em que cada bateria de testes os menos qualificados são eliminados (FRIEDMAN, 2002).

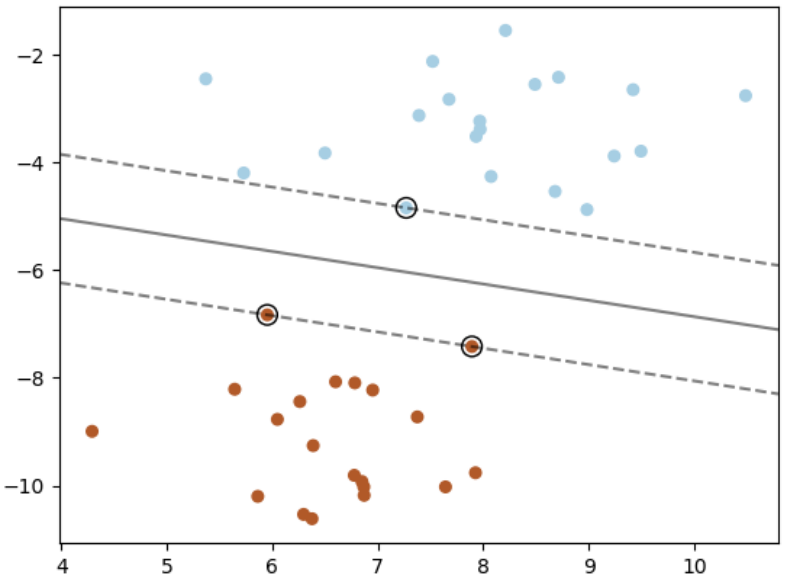
O XGBoost é construído com árvores sequenciais, mas que diferente do RF, em vez de ser paralelo, realiza ciclos em que árvores são podadas, ponderadas e impulsionadas, minimizando os erros e otimizando os recursos computacionais (CHEN e GUESTRIN, 2016).

3.3.3 Support Vector Machine (SVM)

O algoritmo SVM é um conjunto de métodos com uma ampla faixa de aplicação e variação. O SVM é uma técnica supervisionada com finalidade de classificação, regressão e detecção de *outliers*. Sua formulação é determinada pela construção de um ou mais hiperplanos em um espaço de dimensão muito alta, de forma que os pontos dos dados possuam a maior distância possível (HAYKIN, 2009; SVENSÉN e BISHOP, 2007).

Na Figura 9, cada cor de ponto é a representação de uma classe, os pontos circulados em preto são chamados de vetores de suporte e a reta contínua central é o hiperplano. A região entre as retas tracejadas nos vetores de suporte que contém o hiperplano é chamada de margem, quanto maior essa área, menor vai ser o erro associado a generalização do modelo. (SVENSÉN e BISHOP, 2007).

Figura 9 – Função de decisão com vetores de suporte separando duas classes.



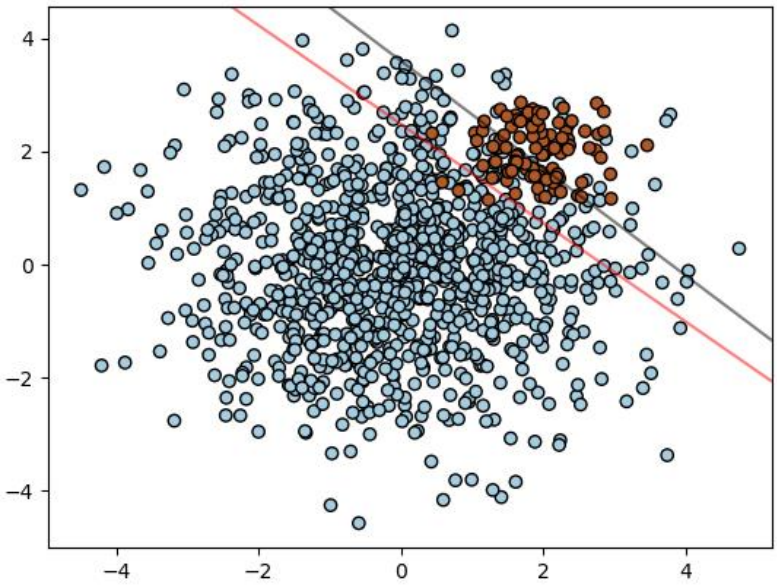
Fonte: Pedregosa et al. (2011)

Alguns cenários não podem ser resolvidos aplicando um hiperplano linear, como mostrado na Figura 10. Nesse caso o SVM transforma o espaço original em um espaço de dimensão superior, adicionando alguma propriedade com base no existente, assim pode ser solucionado por um hiperplano linear ou um hiperplano não linear. Outra solução é permitir que parte dos pontos se encontre no lado errado da

margem, esse é igualmente o método para detecção de *outliers*, mas afeta a precisão do modelo (SVENSÉN e BISHOP, 2007).

A exemplificação tratou casos de classes e sua classificação, mas o SVM pode ser aplicado para realizar a regressão. O método consiste em considerar os pontos mais distantes para traçar um hiperplano entre eles e que mais se aproxime do restante dos dados. A previsão será uma função que medirá a distância para o hiperplano, desconsiderando os dados além da margem (SVENSÉN e BISHOP, 2007).

Figura 10 – Problema não linear de SVM.



Fonte: Adaptado de Pedregosa (2011).

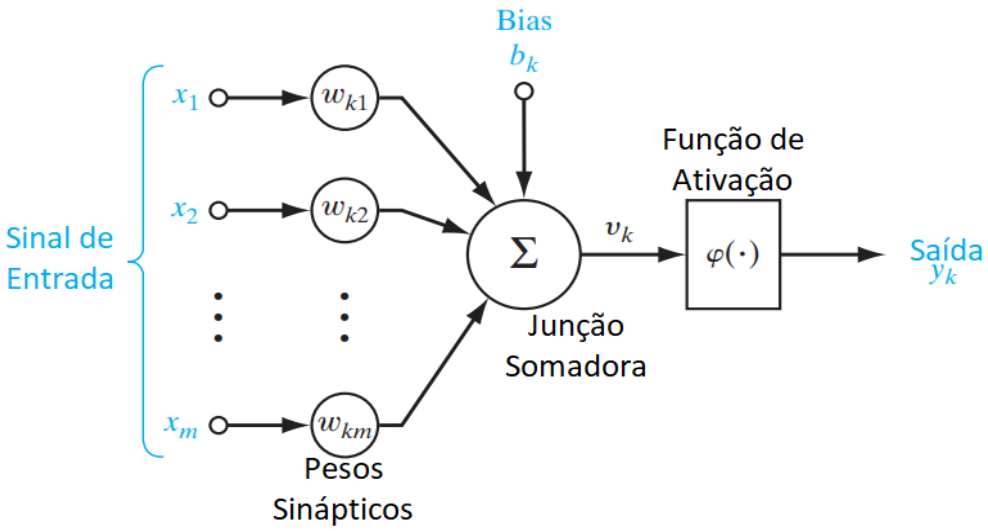
3.3.4 Redes Neurais Artificiais (RNA)

As Redes Neurais Artificiais (RNA) é uma técnica de aprendizado de máquina que representa em miniatura o sistema nervoso biológico do cérebro humano que é composto por um grande número de neurônios, dispostos em camadas, com conexões sinápticas entre si, com pesos correspondentes e aplicam um conhecimento prévio (HAYKIN, 2009).

A RNA padrão consiste em camadas e em cada uma delas tem elementos, chamados de neurônios, ele é fundamental na estrutura e a seleção dos seus parâmetros de entrada é o fator significativo no projeto de RNA. A Figura 11 mostra o

modelo de representação de um neurônio artificial constituído pelo sinal de entrada, pesos sinápticos, junção somadora e função de ativação. Inicialmente, existem X sinais de entrada, que serão ponderados pelos pesos sinápticos e somados pela junção somadora. O resultado desta equação é passado por uma função de ativação que restringe o valor real do neurônio (HAYKIN, 2009).

Figura 11 – Representação de um neurônio artificial.

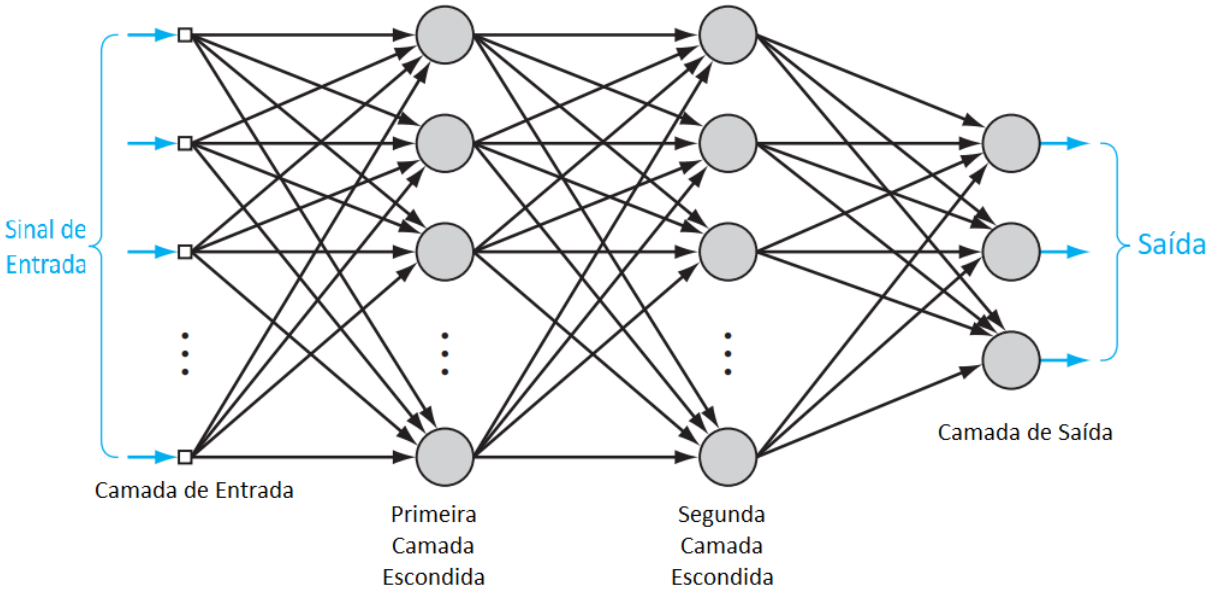


Fonte: Adaptado de Haykin, 2009.

O conjunto de neurônio forma uma rede, e a Multilayer Perceptron (MLP) é um dos tipos de redes neurais amplamente implementadas, que apresenta duas características principais, a não linearidade e a interconectividade massiva. Em importantes problemas de dinâmica não linear e mapeamento de funções, a aplicação da MLP é capaz de realizar aproximações (PRINCIPE *et al.*, 2000).

A Figura 12 exibe as três principais camadas de nós de uma MLP, a de entrada, oculta e de saída. A camada de entrada repassa os sinais para a camada oculta que atribui pesos e realiza um tratamento não linear para a caracterização do sinal de entrada e posteriormente envia para saída. Por fim, a camada de saída produz o resultado da rede neural. O treinamento de uma MLP normalmente é feito pelo algoritmo de *Backpropagation*. Os dados de saída são propagados de volta para a entrada, permitindo que os pesos das camadas ocultas se ajustem para diminuir o erro encontrado, até um nível considerado válido por um especialista (HAYKIN, 2009).

Figura 12 – Representação gráfica de uma MLP com duas camadas escondidas.



Fonte: Adaptado de Haykin, 2009.

3.4 Avaliação de Desempenho

A avaliação de desempenho é uma das tarefas mais fundamentais e críticas para o aprendizado de máquinas, pois as taxas de acertos dos modelos dependem das métricas estipuladas. Uma abordagem correta para comparação de diferentes algoritmos e técnicas contribui para avaliar a melhor solução.

As métricas amplamente utilizadas e com interpretação direta simples são as de erro médio absoluto (MAE – *Mean Absolute Error*) e o erro quadrático médio (MSE – *Mean Square Error*). As métricas de desempenho em aprendizado de máquina são usadas para comparar a previsão com a base de teste e seus resultados podem influenciar na decisão da seleção do algoritmo para implementação (BOTCHKAREV, 2019)

O erro médio absoluto é o módulo da diferença entre as observações, previsão e real. O erro quadrático médio é o erro médio absoluto ao quadrado. Enquanto o erro médio absoluto dá o mesmo peso para todos os tipos de erro, o erro quadrático médio penaliza atribuindo mais peso para valores maiores e diminuindo o peso de valores menores (CHAI e DRAXLER, 2014).

A medida *F-measure* é outra métrica que tenta equilibrar a precisão através de uma média harmônica. Essa métrica utiliza a sensibilidade (*recall*) que é a parte de

previsões corretas sobre o número de exemplos positivo do conjunto, assim como a precisão, onde quanto maior o valor melhor o algoritmo (ZAKI et al., 2014). A Tabela 3 apresenta as etapas para calcular o *F-measure.*

Tabela 3 – Cálculo do F-measure.

F-measure	$\frac{(\beta^2+1)Precision \cdot Recall}{\beta^2 Precision + Recall}$
Recall	$\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i + fn_i)}$
Precisão	$\frac{\sum_{i=1}^{l} tp_i}{\sum_{i=1}^{l}(tp_i + fp_i)}$

Fonte: elaborado pelo autor.

O tempo de treinamento é a velocidade que um algoritmo modela uma previsão, sendo um critério que não pode ser ignorado, pois apesar dos algoritmos apresentar boas métricas e acerto, se o tempo for consideravelmente grande para atingir o resultado, ele se torna indesejável (LIM, LOH e SHIH, 2020). Em caso de classificadores estarem com assertividade muito próxima, o tempo para treinamento poderá ser levado em conta para ser mais um critério de desempate.

O coeficiente de determinação reflete uma relação entre duas variáveis, que ao estar próximo do módulo de 1 é forte e de 0 é fraca, caso seja positiva, ambas crescem proporcionalmente e negativa quando ambas diminuem. Existem muitos tipos de coeficientes, mas os principais são o de Pearson, Spearman e Kendall (CHOK, 2010). A correlação de Pearson é definida como a razão da covariância sobre seus respectivos produtos, e as outras duas são semelhantes, mas com suas próprias considerações na classificação das variáveis do algoritmo e servirão como base para seleção dos preditores.

3.5 Python

O Python é uma linguagem de programação interativa muito popular e é conhecida por ter um grande número de bibliotecas para diversas aplicações. Para as tarefas de aprendizado de máquina a biblioteca *scikit-learn* tem uma das ferramentas mais acessíveis e populares de código aberto (RASCHKA e MIRJALILI, 2017).

A linguagem de programação Python por ser uma licença de código aberto, ter uma estrutura mais simples e com poder de resolver problemas complexos com suas bibliotecas, justificam a vanguarda no segmento, fortificando a comunidade e cresce cada vez mais (PEDREGOSA *et al*., 2011).

Comparando alguns tópicos das linguagens como a conexão com os algoritmos, o suporte da comunidade, a simplicidade e exemplos de aplicação em situações semelhantes de aprendizado de máquina, tornaram ela como a linguagem selecionada para o trabalho.

4 METODOLOGIA

Este capítulo contém a metodologia aplicada à pesquisa, com os procedimentos e técnicas que visam encontrar a solução do problema com a identificação das variáveis de saída e preditoras, seus tratamentos e seleção de melhores parâmetros para aplicação dos algoritmos.

4.1 Procedimentos Metodológicos

As atividades foram desenvolvidas de forma a atingir os objetivos estabelecidos, com o foco em prever com regressão a potência ótima de carga em cada ponto da rede elétrica de média tensão da Enel Distribuição Ceará. As etapas para produzir os algoritmos foram:

1. Estudo dos procedimentos e extração dos dados da empresa;
2. Identificação das variáveis preditoras;
3. Transformação e limpeza dos dados;
4. Exclusão de *outliers*;
5. Normalização dos dados;
6. Análise das variáveis;
7. Otimização dos hiperparâmetros;
8. Aplicação dos algoritmos;
9. Avaliação dos desempenhos dos algoritmos.

Para a primeira atividade de pesquisa foi preciso entender como os softwares de fluxo de potência organizam os dados, para extrair em um formato de tabela ".xlsx" de simples compreensão e aplicação dos passos seguintes. Na sequência, os dados operacionais da rede, como indicadores e equipamentos, foram coletados e cruzados com os dados obtidos anteriormente, para consegui-los organizar na mesma base, facilitando a aplicação dentro do ambiente dos algoritmos.

Após a montagem da base, todas as variáveis foram identificadas e nomeadas, de modo a facilitar o processo de análise de preditores. Com isso foi possível realizar a limpeza de dados que inicialmente possuíssem algum ruído ou dados faltantes, efetuando a exclusão daqueles em que não era possível uma

correção, para não impactar a modelagem. As variáveis do tipo categórica foram transformadas no tipo *dummy*, pois na regressão dos algoritmos as do tipo categóricas não conseguiriam ser processadas e agregadas ao modelo.

Uma nova análise da tabela foi realizada buscando identificar dados discrepantes e que chamassem a atenção em uma análise manual. Além disso, em uma nova avaliação gráfica foi possível verificar a distribuição dos dados e excluir os *outliers*.

A etapa seguinte constitui-se em criar novas variáveis levando em consideração dados existentes para testa-los no modelo. Com os dados pré-processados e sem *outliers*, foi possível normalizar e padronizar individualmente cada coluna de variáveis para realizar a seleção.

Ao término da montagem da base de dados, a aplicação de todos os dados ao algoritmo tomaria muito tempo de processamento, modelagem e escolha dos hiperparâmetros. Então as variáveis passaram por uma avaliação preliminar com três tipos de correlação, sendo elas a de Pearson, Spearman e Kendall. O objetivo da correlação foi identificar o nível de ligação entre os dados, assim como com a demanda dos estudos, para dar suporte na escolha das variáveis. Devido à grande quantidade de preditores disponíveis, optou-se por utilizar mais uma métrica de desempenho, a *F-measure*, para ter certeza da escolha tomada, ordenando de forma decrescente de prioridade com a pontuação obtida individualmente.

Com a definição dos preditores, um conjunto de valores dos hiperparâmetros foram carregados nos algoritmos pelo pacote *Grid Search* da biblioteca *Scikit-learn*¸ sendo esse pacote responsável por testar exaustivamente nos modelos todas as variações dos hiperparâmetros com a finalidade de encontrar a melhor combinação para obter o menor erro médio absoluto.

Os melhores hiperparâmetros foram selecionados e aplicados aos algoritmos para treinamento. Dando seguimento, o resultado de cada um deles foi avaliado e comparado por meio do erro médio absoluto, erro quadrático médio e tempo de treinamento, tanto para a base de treino quanto para a base de testes. Uma última comparação entre os algoritmos foi realizada com o objetivo de avaliar a assertividade dos modelos em geral e abertas por classe de consumidores. Por fim, o melhor modelo foi selecionado para a construção do protótipo, com base nas métricas e análises realizadas nos subcapítulos a seguir.

4.2 Análise das Variáveis

A planilha de informações com os estudos realizados por especialistas em fluxo de potência tinha originalmente 8.007 linhas e 6 colunas, com dados do período entre janeiro de 2014 e setembro de 2021. Outra base com os dados de indicadores da rede foi extraída e adicionada à primeira base cruzando os dados por alimentadores e subestações, tornando a nova planilha em 8.007 linhas e 46 colunas, totalizando 368.322 células de dados. As variáveis preditoras seguem listadas abaixo:

- %REG_DEC: Valor de DEC realizado pelo conjunto sobre a sua respectiva meta regulatória.
- %REG_FEC: Valor de FEC realizado pelo conjunto sobre a sua respectiva meta regulatória.
- %UTILIZAÇÃO: Valor de corrente máxima lida pelo alimentador no último ano sobre sua capacidade máxima.
- ALC: Representa o comprimento da média tensão dos alimentadores dividido pelo número de clientes.
- ALIMENTADOR_AJUST: Nome do alimentador ajustado na categoria de *dummy* para implementar no algoritmo.
- ANICI: Representa o número médio ponderado de clientes interrompidos por todos os dispositivos de proteção do alimentador.
- ANICI_CHAVE: Representa o número médio ponderado de clientes interrompidos por chaves no alimentador.
- ANICI_FUSÍVEL: Representa o número médio ponderado de clientes interrompidos por fusíveis no alimentador.
- ATIVIDADE_AJUST: Atividade exercida pelo empreendimento ajustada na categoria de *dummy* para implementar no algoritmo.
- CAIDI: Duração média do tempo de recomposição do alimentador em minutos.
- CAPACIDADE: Corrente máxima suportada pelo alimentador.
- CARREGAMENTO: Valor de corrente máxima lida pelo alimentador.
- CHAVES: Quantidade de chaves no alimentador.
- CI: Clientes médios interrompidos por alimentador.

- CI_CHAVES: Clientes médios interrompidos por chaves em um alimentador.
- CI_FUSÍVEIL: Clientes médios interrompidos por fusíveis em um alimentador.
- CLIENTES_ALIMENTADOR: Quantidade total de clientes em um alimentador.
- CLIENTES_CONJUNTO: Quantidade total de clientes em um conjunto.
- DELTA: Diferença entre ICC_PONTO e ICC_SED.
- DEMANDA_PREV (kW): Demanda do empreendimento solicitada pelo cliente para estudo de fluxo de carga.
- DIST_SED: Distância em km da subestação para o ponto de conexão solicitado.
- FALTAS: Quantidade de faltas no alimentador.
- FALTA_FUSÍVEL: Quantidade de faltas ocasionadas por fusíveis do alimentador.
- FALTA_VEGETAÇÃO: Quantidade de faltas ocasionadas por vegetação do alimentador.
- FRIC: Número de interrupções do alimentador sobre a quantidade total de clientes do alimentador.
- FRIC_VEGETAÇÃO: Número de interrupções ocasionadas por vegetação do alimentador sobre a quantidade total de clientes do alimentador.
- FUSÍVEIS: Quantidade de fusíveis no alimentador.
- ICC/km: Divisão da variável “DELTA” pela “DIST_SED”.
- ICC_PONTO: Curto-circuito trifásico simétrico no ponto selecionado para conexão.
- ICC_SED: Curto-circuito trifásico simétrico na subestação que irá atender o empreendimento.
- KM_TOTAL: km total da rede de média tensão do alimentador associado.
- KV_MADRUGADA: Tensão média no alimentador no período da madrugada.
- KV_MANHÃ: Tensão média no alimentador no período da manhã.

- KV_NOITE: Tensão média no alimentador no período da noite.
- KV_TARDE: Tensão média no alimentador no período da tarde.
- OBRA_AJUST: Resultado do estudo para a demanda selecionada, indicando se há ou não necessidade de obra. Variável ajustada para o tipo *dummy* visando a implementação no algoritmo.
- PASSO_TLC: Quantidade de equipamentos telecomandados distribuídos no alimentador sobre a quantidade de clientes do alimentador.
- RELIGADORES: Quantidade de equipamentos telecomandados no alimentador.
- SED_AJUST: Nome da subestação que atenderá a cliente ajustada para o tipo *dummy* visando a implementação no algoritmo.
- SAIDI_ALIMENTADOR: Duração média da interrupção para cada cliente a nível de alimentador em minutos.
- SAIDI_CONJUNTO: Duração média da interrupção para cada cliente a nível de conjunto em minutos.
- SAIFI_ALIMENTADOR: Número médio de interrupções sofridas pelos clientes do alimentador.
- SAIFI_CHAVE: Número médio de interrupções sofridas pelos clientes do alimentador por chaves.
- SAIFI_FUSÍVEL: Número médio de interrupções sofridas pelos clientes do alimentador por fusíveis.
- SAIFI_CONJUNTO: Número médio de interrupções sofridas pelos clientes do conjunto.
- TIPO_AJUST: Tipo de classe de consumidor (Residencial, Comercial, Industrial e Outros) ajustado para o tipo *dummy* visando a implementação no algoritmo.

4.3 Limpeza e Transformação de Dados

A primeira limpeza de dados foi realizada de forma visual buscando possíveis incoerências, como valores textuais em entradas que deveriam ser numéricas, valores negativos que não poderiam ser fisicamente possíveis dependendo da entrada e

dados em branco ou zerados em preditores textuais. Os dados com esses problemas foram removidos para não distorcer ou enviesar o modelo, caso fosse efetuada uma substituição dos valores com base em alguma estimativa ou outra premissa.

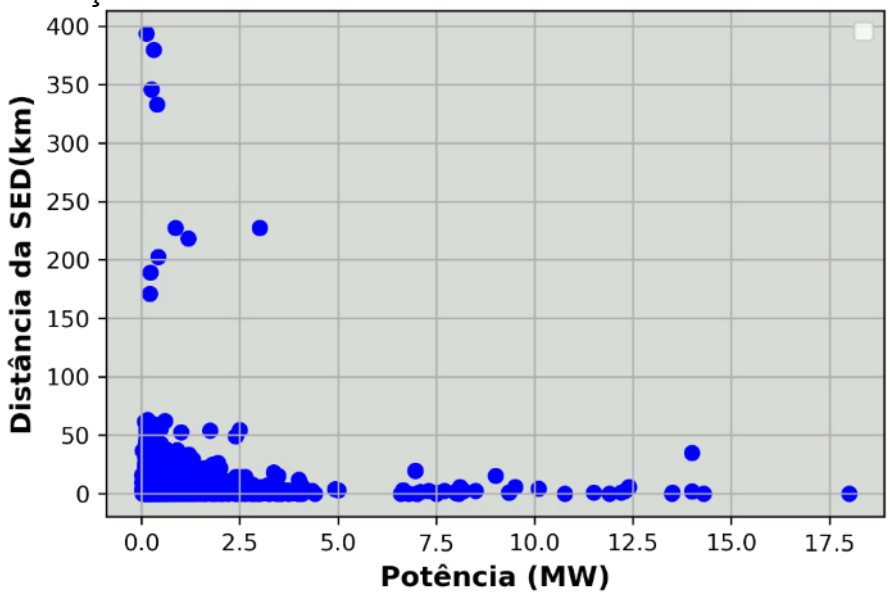
As entradas referentes a alimentador, atividade do empreendimento, obra e tipo de classe do empreendimento receberam um tratamento com a biblioteca *pandas* para transformar seus dados categóricos em equivalentes *dummy* para aplicação no algoritmo, dado que para realizar a regressão nos escolhidos todas os preditores devem ser do tipo numérico.

Apesar da grande quantidade de dados de entrada, novos previsores calculados a partir dos existentes podem resultar em um ganho no algoritmo. Dito isso, foram adicionados os previsores delta de curto-circuito trifásico simétrico e curto-circuito trifásico simétrico por quilômetro para avaliar se a variação e degradação de outros parâmetros nos alimentadores influenciam na modelagem proposta.

Para identificar dados discrepantes e que acabaram passando despercebido em tratamentos anteriores, foi construída a Figura 13 com auxílio das bibliotecas *sklearn* e *pandas*. Nela podemos observar que os dados estão mais concentrados em regiões próximas à subestação e com potência até 5MW, mas também existem potências pequenas longe da subestação e o contrário, em que potências grandes estão próximas a subestação. Nesses casos não foram considerados *outliers* pela concepção de instalações das subestações estarem sempre próximas a grandes concentrações de cargas. Além do mais, o Ceará tem a característica de possuir alimentadores muito extensos para atender pequenas cargas. Logo, nessa análise nenhumas das cargas foram consideradas como *outliers*.

No sentido de finalização da montagem e tratamento da base de dados a normalização foi realizada em todos os preditores, com exceção na demanda que é a variável de saída, para que as faixas de amplitude de cada elemento não induzam esse peso aplicado ao algoritmo, tornando-o assim tendencioso. Desta forma, todos os preditores foram alterados para variar entre 0 e 1, evitando impactos na performance e avaliação.

Figura 13 – Detecção de outliers de potência por distância da subestação.



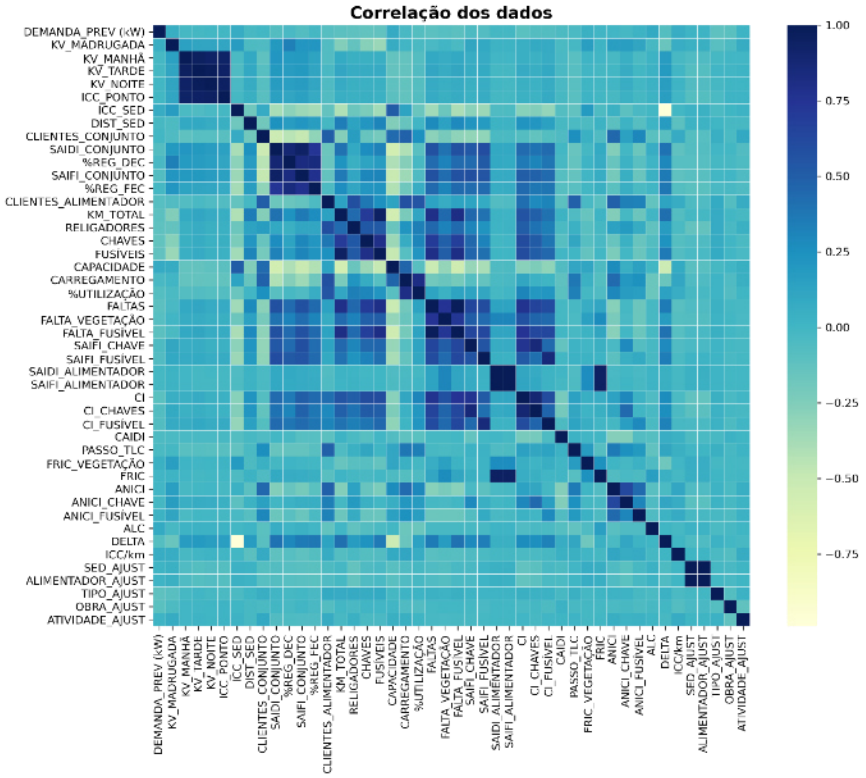
Fonte: elaborado pelo autor.

4.4 Seleção das Variáveis Preditoras

Uma análise gráfica é efetuada nessa etapa para uma primeira avaliação do comportamento e correlação das variáveis, antes da escolha das melhores que irão ser aplicadas aos algoritmos. A seleção dos dados preditores será dividida em três fases, sendo duas de correção e uma de desempenho para regressão.

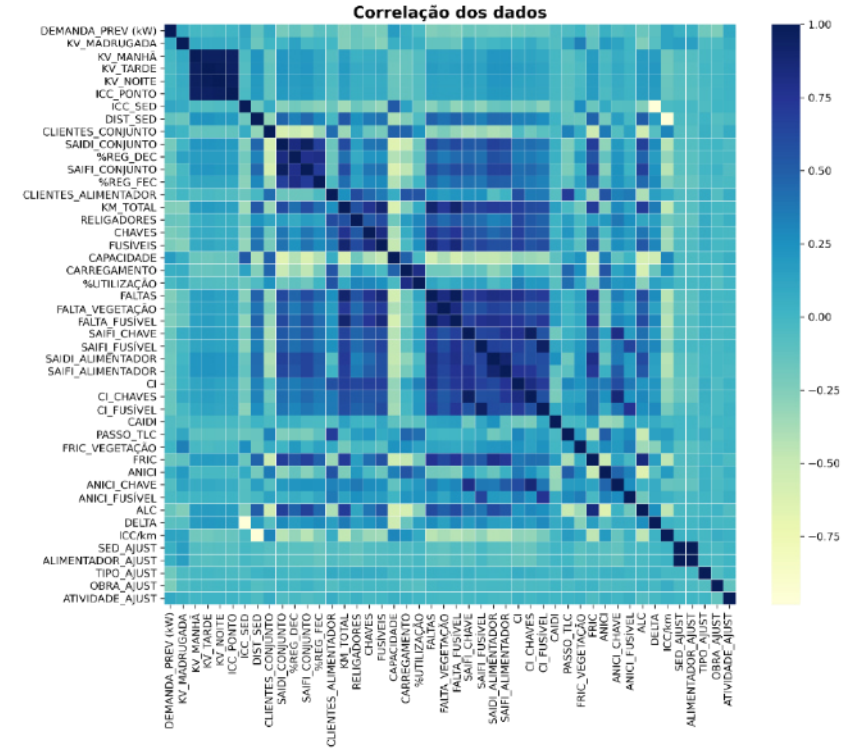
Inicialmente a correlação interna mostrará se há correlação entre todas as variáveis, utilizando um mapa de calor dos métodos de Pearson, Spearman e Kendall. Nesse mapa de calor cada linha é testada a correlação com cada coluna, valores positivos indicam que são diretamente proporcionais e valores negativos que são inversamente proporcionais, no qual o nível de correlação é mais forte quando está próximo do módulo de 1. Na Figura 14 foi utilizado o método de correlação de Pearson, na Figura 15 o método de Spearman e na Figura 16 o método de Kendall.

Figura 14 – Correlação entre as variáveis utilizando o método de Pearson.



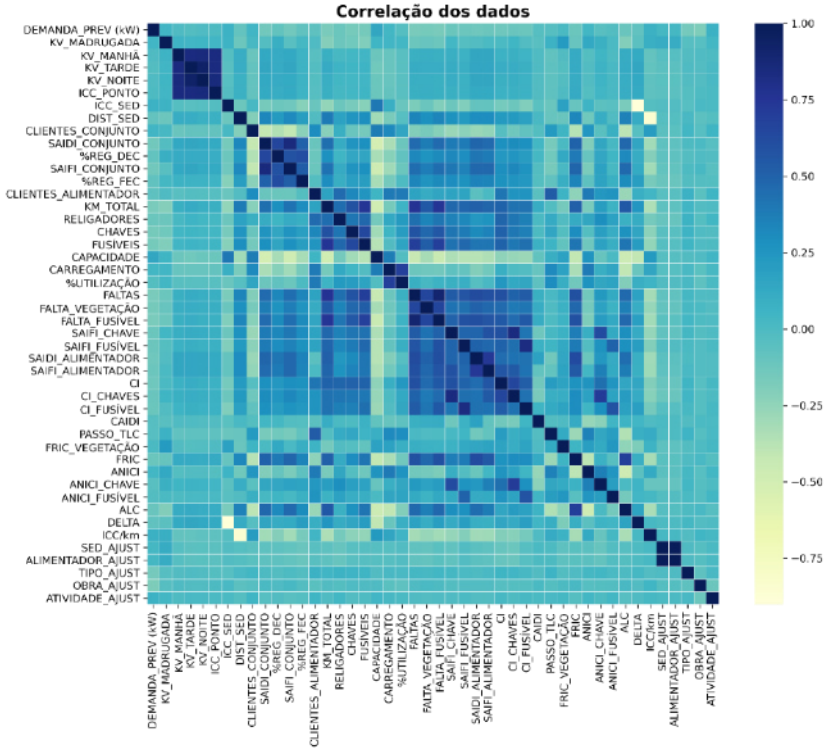
Fonte: elaborado pelo autor.

Figura 15 – Correlação entre as variáveis utilizando o método de Spearman.



Fonte: elaborado pelo autor.

Figura 16 – Correlação entre as variáveis utilizando o método de Kendall.

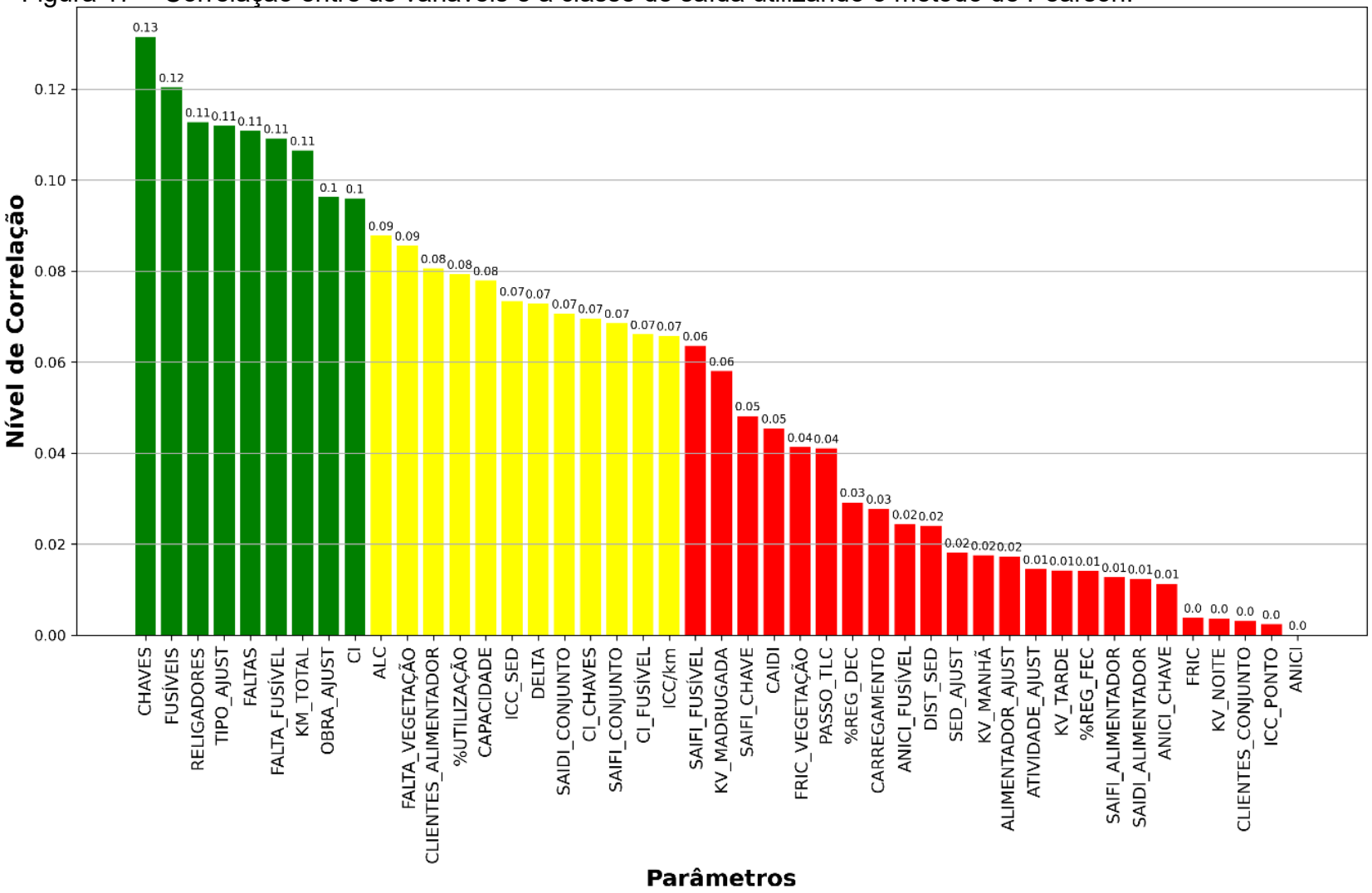


Fonte: elaborado pelo autor.

Ao comparar as Figura 14, Figura 15 e Figura 16, partindo do método com menores correlações para o mais forte, verifica-se que o método de Kendall visualmente apresenta cores tendendo ao verde e azul claro, indicando uma menor correlação entre as variáveis. O método de Pearson apresentou mais cores próximas ao azul escuro em algumas regiões, indicando forte correlação entre as variáveis em análise. Por fim o método de Spearman, entre todos os métodos, apresentou mais variáveis com essa grande correlação, com regiões escuras em destaque.

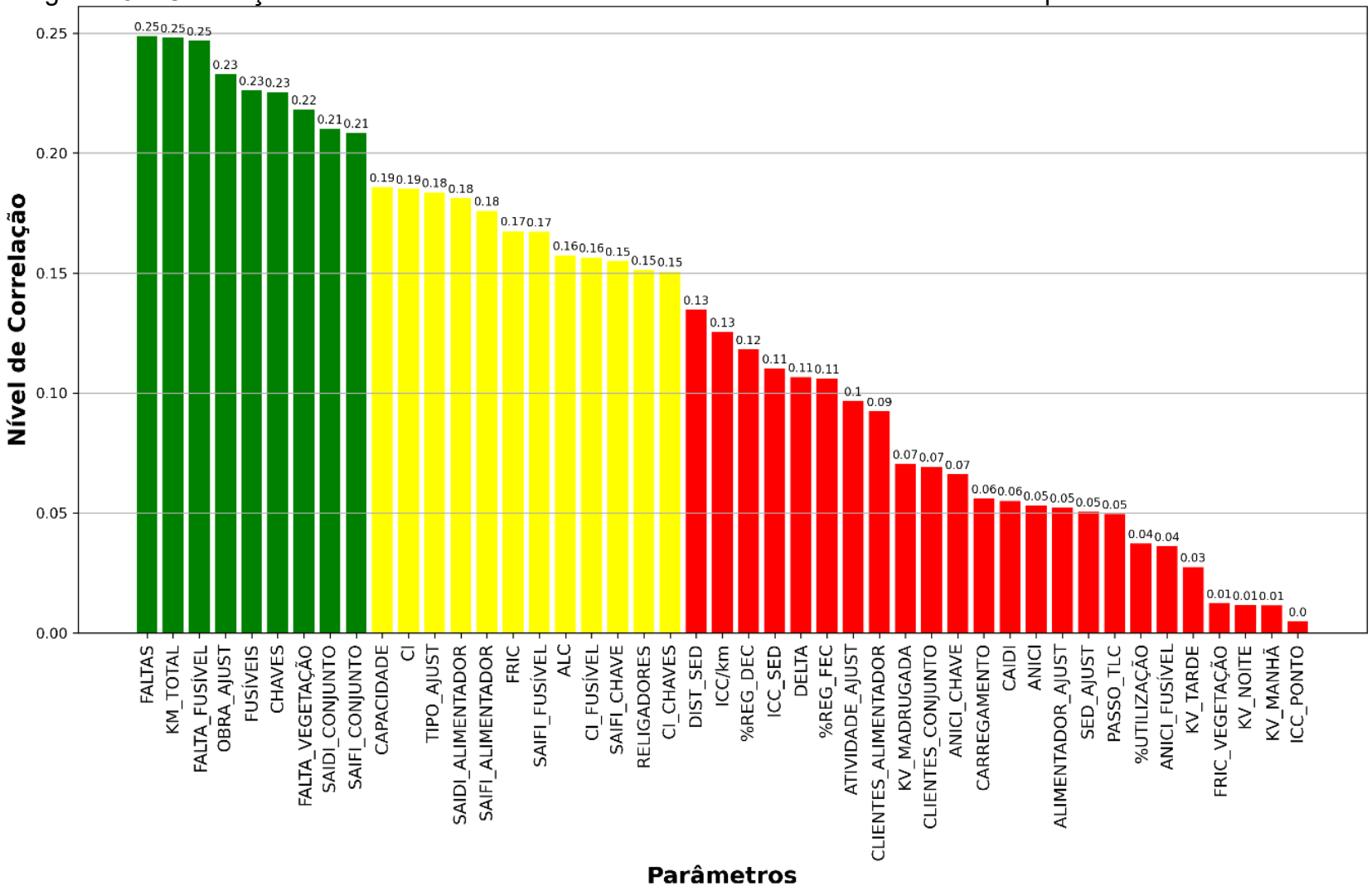
Partindo para a avaliação numérica entre as variáveis e a classe de saída (potência), os métodos de Pearson, Spearman e Kendall foram aplicados para a construção de um gráfico de barras com seus valores em módulo visando facilitar a visualização e entendimento dessa segunda análise. Nesse segundo estágio de análise serão selecionadas as colunas com maior correlação nos três métodos aplicados e em caso de dados que apresentem correlação entre si, serão avaliados a sua possível remoção já que poderiam indicar o mesmo preditor, enviesando o modelo.

Figura 17 – Correlação entre as variáveis e a classe de saída utilizando o método de Pearson.



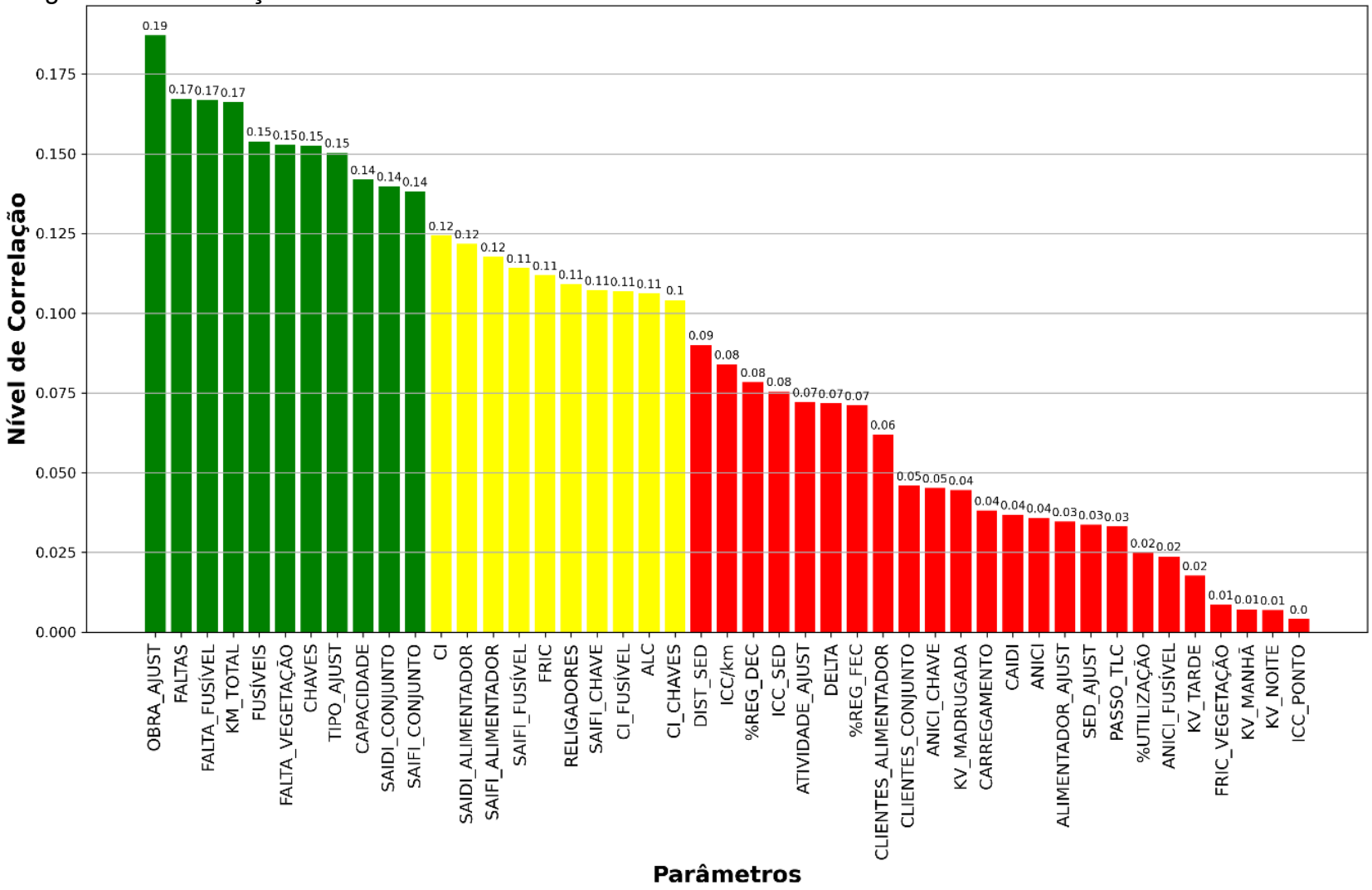
Fonte: elaborado pelo autor.

Figura 18 – Correlação entre as variáveis e a classe de saída utilizando o método de Spearman.



Fonte: elaborado pelo autor.

Figura 19 – Correlação entre as variáveis e a classe de saída utilizando o método de Kendall.



Fonte: elaborado pelo autor.

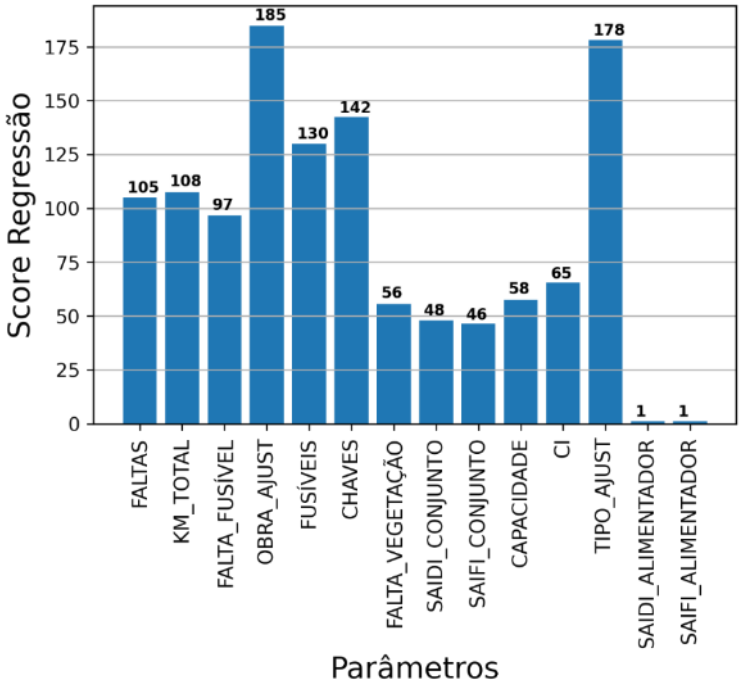
A Figura 17 apresenta o resultado da correlação entre as variáveis com a saída no método de Pearson, a Figura 18 para o método de Spearman e a Figura 19 para o método de Kendall. Em cada figura são mostrados os preditores ordenados em ordem decrescente de correlação com a potência e separados em grupos, em que verde representa o melhor grupo, amarelo os valores intermediários e o vermelho para os dados restantes.

Analisando os gráficos, a partir do coeficiente de correlação foi realizado um primeiro filtro de preditores, removendo as entradas com os piores desempenhos, permanecendo com 30% dos dados para realizar testes posteriores para maior eficiência e menor tempo de processamento durante o treinamento.

A terceira etapa de análise consiste em avaliar o impacto individual de cada variável de entrada na saída desejada com a métrica de *F-measure*. As variáveis escolhidas foram FALTAS, KM_TOTAL, FALTA_FUSÍVEL, OBRA_AJUST, FUSÍVEIS, CHAVES, FALTA_VEGETAÇÃO, SAIDI_CONJUNTO, SAIFI_CONJUNTO, CAPACIDADE, CI, TIPO_AJUST, SAIDI_ALIMENTADOR e SAIFI_ALIMENTADOR. A Figura 20 apresenta a pontuação na regressão de cada preditor, onde quanto maior o valor obtido, a probabilidade de impactar os algoritmos positivamente é alta. Apesar disso, todas as variáveis preditoras serão testadas individualmente em cada modelo, ainda com risco de ser removida em caso de piora da assertividade com sua presença ou adicionada para aumentar a precisão do modelo.

Para a fase inicial de modelagem foram escolhidos os parâmetros com maior rendimento de métricas anteriores, permanecendo com 15% dos previsores, resultado do resultado da Figura 20 e listados a seguir: FALTAS, KM_TOTAL, FALTA_FUSÍVEL, OBRA_AJUST, FUSÍVEIS, CHAVES e TIPO_AJUST.

Figura 20 – *F-measure* para avaliar as melhores preditoras.



Fonte: elaborado pelo autor.

4.5 Otimização dos Hiperparâmetros do Modelo

Após a análise das variáveis de entrada e seleção das variáveis preditoras, o próximo passo é a aplicação nos métodos escolhidos para prever a variável de saída, mas para cada algoritmo ajustes podem ser realizados em seus hiperparâmetros visando melhorar o desempenho.

A amplitude de cada hiperparâmetro em seu respectivo modelo, somado às diversas combinações possíveis, tornam essa tarefa manual praticamente impossível de ser efetuada. Por isso para busca da melhor combinação foi utilizada a ferramenta *Grid Search* da biblioteca *Scikit-learn*, em que ela permite colocar todos os hiperparâmetros e implementa todas as combinações possíveis nos testes dos algoritmos, no qual ao final escolherá o teste em que apresentar o menor erro durante a simulação.

As Tabela 4, Tabela 5, Tabela 6 e Tabela 7 mostram os quatro modelos RF, SVM, RNA e XGBoost com seus hiperparâmetros otimizados, selecionados com base nas diversas iterações da ferramenta *Grid Search* nos cenários em que foram impostos.

Tabela 4 – Valores dos hiperparâmetros para o projeto de *Random Forest*.

Parâmetro	Condição
Tipo de Treinamento	Supervisionado
Tipo de Problema	Regressão
Entrada	FALTAS, KM_TOTAL, FALTA_FUSÍVEL, OBRA_AJUST, FUSÍVEIS, CHAVES e TIPO_AJUST.
Saída	POTÊNCIA
Número de árvores	50
Quantidade de nós	100

Fonte: elaborado pelo autor.

Tabela 5 – Valores dos hiperparâmetros para o projeto de SVM.

Parâmetro	Condição
Tipo de Treinamento	Supervisionado
Tipo de Problema	Regressão
Entrada	FALTAS, KM_TOTAL, FALTA_FUSÍVEL, OBRA_AJUST, FUSÍVEIS, CHAVES e TIPO_AJUST.
Saída	POTÊNCIA
Kernel	rbf
C	9
Epsilon	0.3
Degree	1

Fonte: elaborado pelo autor.

Outros hiperparâmetros dos modelos podem ser modificados, mas à medida que colocamos mais entradas na ferramenta *Grid Search* o tempo de processamento aumenta exponencialmente para encontrar os melhores, pelo fato de testar todas as combinações possíveis. Então o foco está em alterar os principais hiperparâmetros que afetam diretamente o desempenho e não aqueles que fazem ajustes finos.

Tabela 6 – Valores dos hiperparâmetros para o projeto de RNA

Parâmetro	Condição
Tipo de Treinamento	Supervisionado
Tipo de Problema	Regressão
Entrada	FALTAS, KM_TOTAL, FALTA_FUSÍVEL, OBRA_AJUST, FUSÍVEIS, CHAVES e TIPO_AJUST.
Saída	POTÊNCIA
Ativação	relu
Tolerância	0.0001
Camadas Ocultas	50,50,50
Número Máximo de Interações	5000
Alfa	0.001
Otimização	lbfgs

Fonte: elaborado pelo autor.

Tabela 7 – Valores dos hiperparâmetros para o projeto de XGBoost.

Parâmetro	Condição
Tipo de Treinamento	Supervisionado
Tipo de Problema	Regressão
Entrada	FALTAS, KM_TOTAL, FALTA_FUSÍVEL, OBRA_AJUST, FUSÍVEIS, CHAVES e TIPO_AJUST.
Saída	POTÊNCIA
Booster	gbtree
Métrica	mae
Taxa de aprendizado	0.01
Número de árvores	350
Quantidade de nós	20
Objetivo	reg:squarederror
gamma	10

Fonte: elaborado pelo autor.

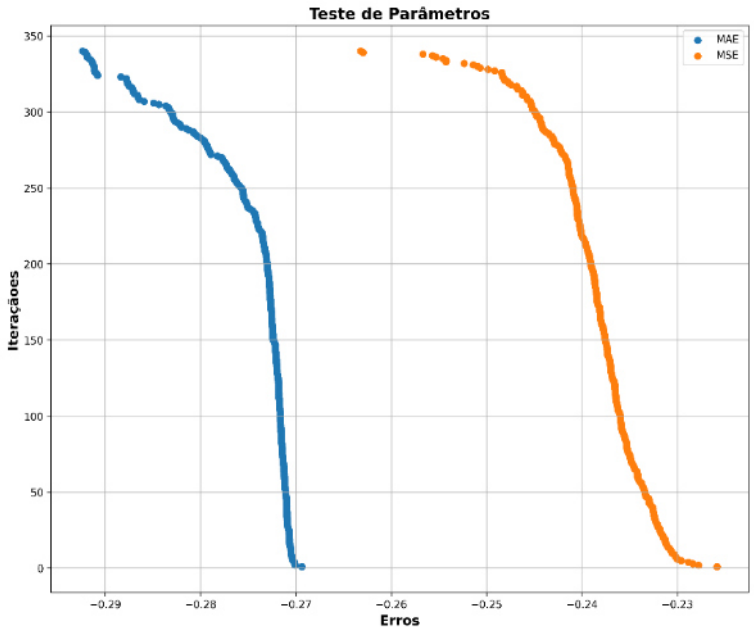
5 RESULTADOS E DISCUSSÕES

Este capítulo apresenta os resultados obtidos nos algoritmos a partir da metodologia do capítulo anterior. A princípio aplicam-se os preditores com os hiperparâmetros otimizados nos modelos e serão avaliadas métricas para seleção de um modelo que ao final do capítulo se tornará um protótipo.

5.1 Aplicação dos Algoritmos de Aprendizagem

Para avaliar o desempenho dos modelos de aprendizagem e prever a potência em todos os pontos da rede elétrica, foram utilizados cinco métricas. Inicialmente foram selecionados 80% dos dados para treino, acompanhando a evolução das suas iterações, medindo pelo erro médio absoluto e o erro médio quadrático.

Figura 21 – Nível de iteração por erros absolutos e quadráticos do RF.

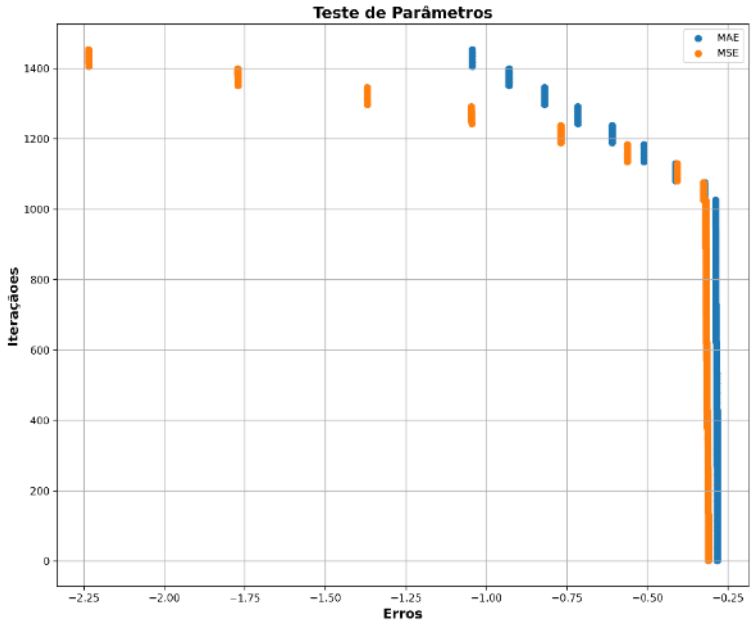


Fonte: elaborado pelo autor.

Com o auxílio da ferramenta *Grid Search* foi possível monitorar os erros durante o treinamento dos algoritmos aplicando os hiperparâmetros encontrados anteriormente. As Figura 21 a Figura 24 mostram o comportamento de cada iteração

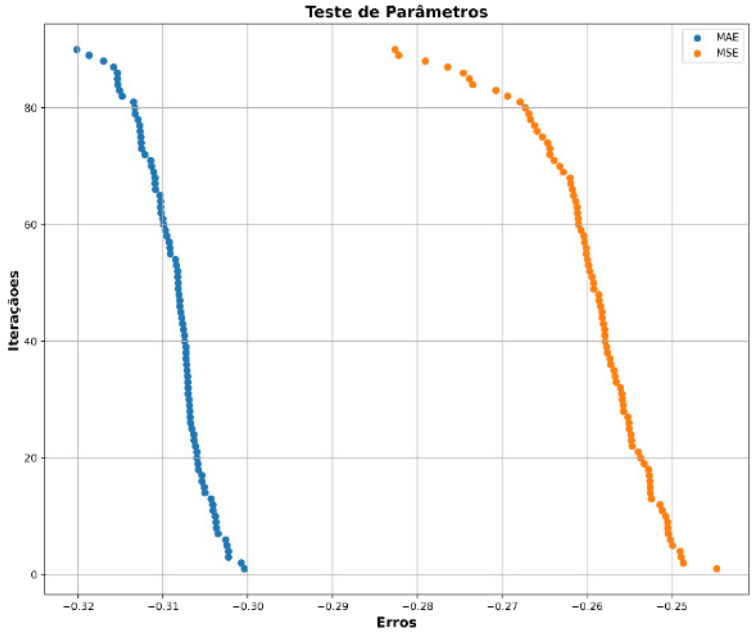
dos erros médio absoluto e erros quadráticos médios para os algoritmos de RF, SVM, RNA e XGBoost, respectivamente.

Figura 22 – Nível de iteração por erros absolutos e quadráticos do SVM.



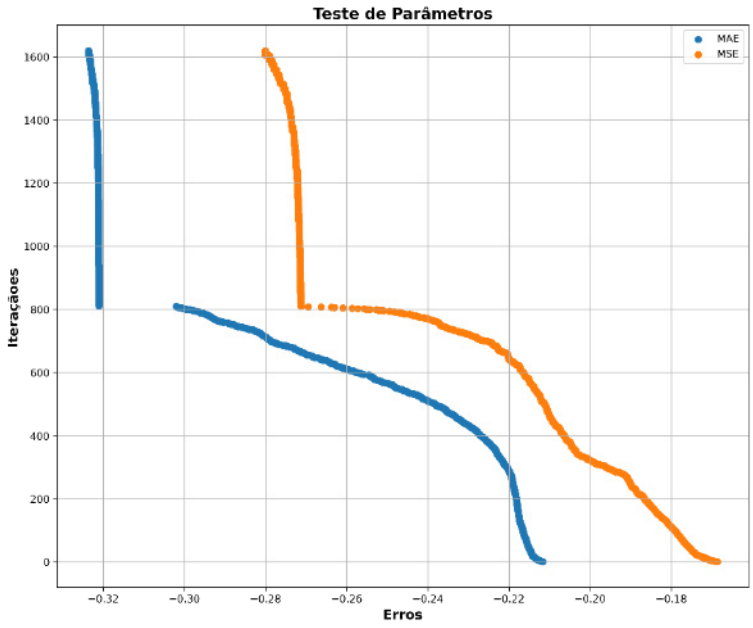
Fonte: elaborado pelo autor.

Figura 23 – Nível de iteração por erros absolutos e quadráticos do RNA.



Fonte: elaborado pelo autor.

Figura 24 – Nível de iteração por erros absolutos e quadráticos do XGBoost.



Fonte: elaborado pelo autor.

Na Tabela 8 é possível visualizar os erros dos algoritmos na base de teste que correspondem ao erro médio absoluto e ao erro quadrático médio. Verifica-se que o erro médio absoluto está distribuído em uma faixa entre 0,145 MW do algoritmo XGBoost e 0,286 MW do algoritmo RNA.

Tabela 8 – Comparação das métricas dos algoritmos.

Algoritmo	MAE	MSE
RF	0,170	0,094
SVM	0,235	0,158
RNA	0,286	0,173
XGBoost	0,145	0,072

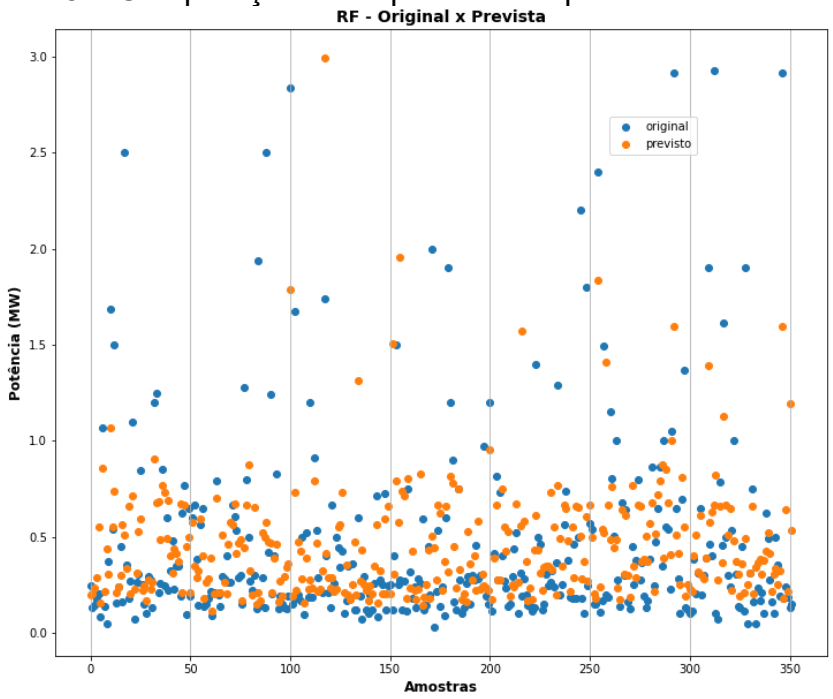
Fonte: elaborado pelo autor.

Os resultados mostram que o algoritmo XGBoost apresentou os menores erros e que suas previsões indicam 0,145 MW ou 145 kW para mais ou para menos dentro do esperado. Em segundo lugar ficou o algoritmo de RF, na sequência o SVM e por último o RNA.

Com o objetivo de exemplificar o comportamento da aprendizagem de todos os algoritmos, realizou-se uma plotagem para os quatro métodos comparando o valor previsto com o valor real da base de testes.

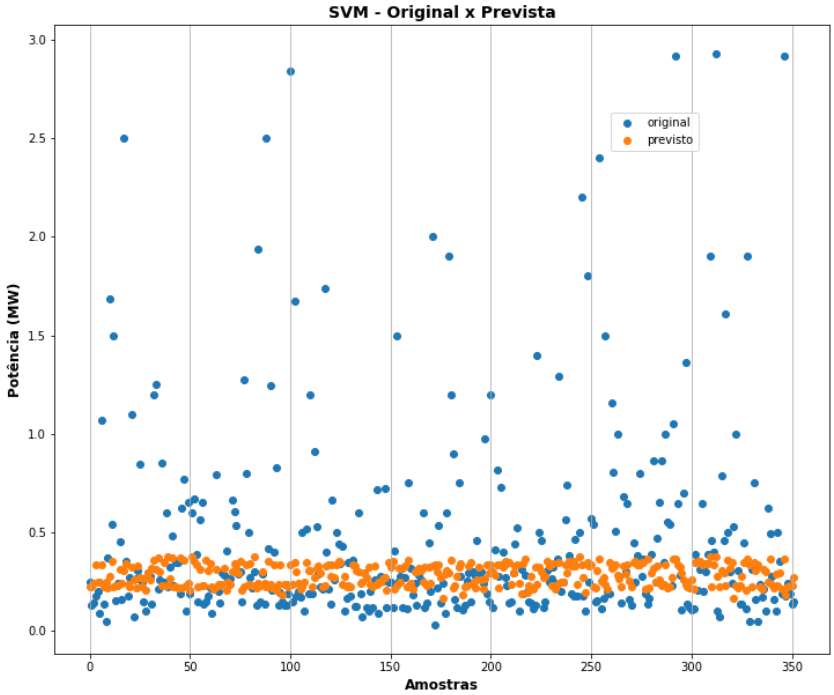
As Figura 25 a Figura 28 exibem o resultado dos algoritmos com a base de testes de estudos reais realizado por especialista da distribuidora. Essas figuras servem como métrica de avaliação de precisão e exatidão dos algoritmos.

Figura 25 – Comparação entre potência de previsão e real do RF.



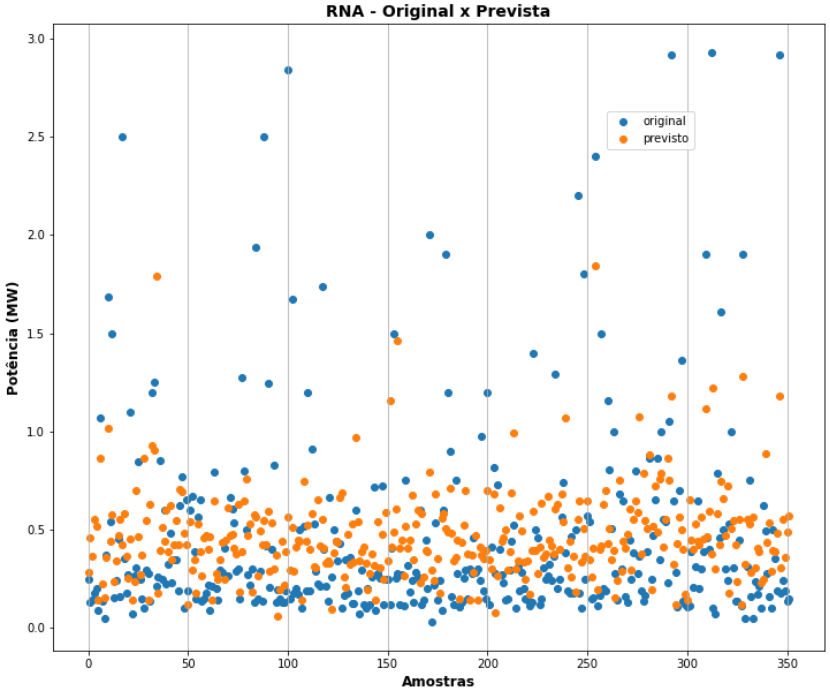
Fonte: elaborado pelo autor.

Figura 26 – Comparação entre potência de previsão e real do SVM.



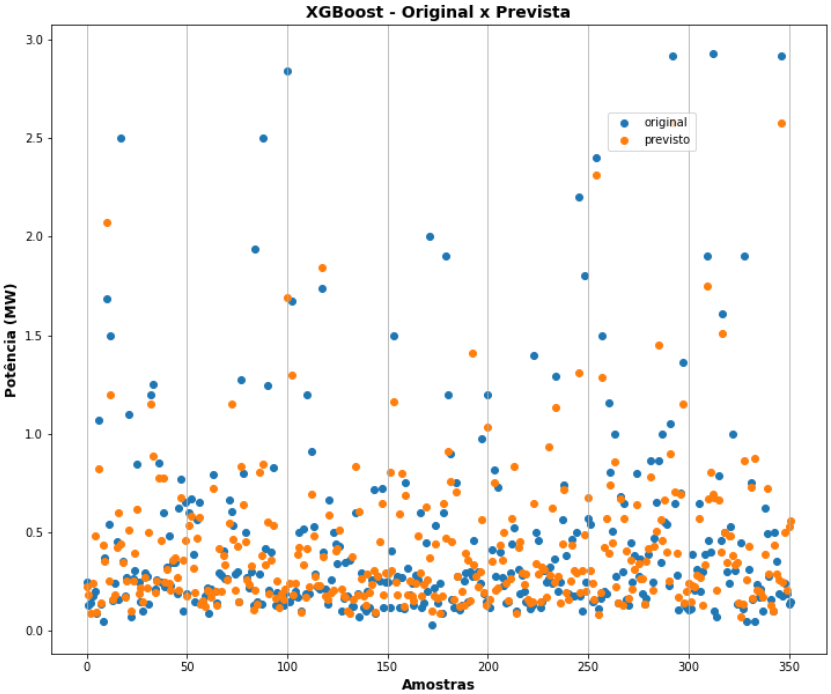
Fonte: elaborado pelo autor.

Figura 27 – Comparação entre potência de previsão e real do RNA.



Fonte: elaborado pelo autor.

Figura 28 – Comparação entre potência de previsão e real do XGBoost.



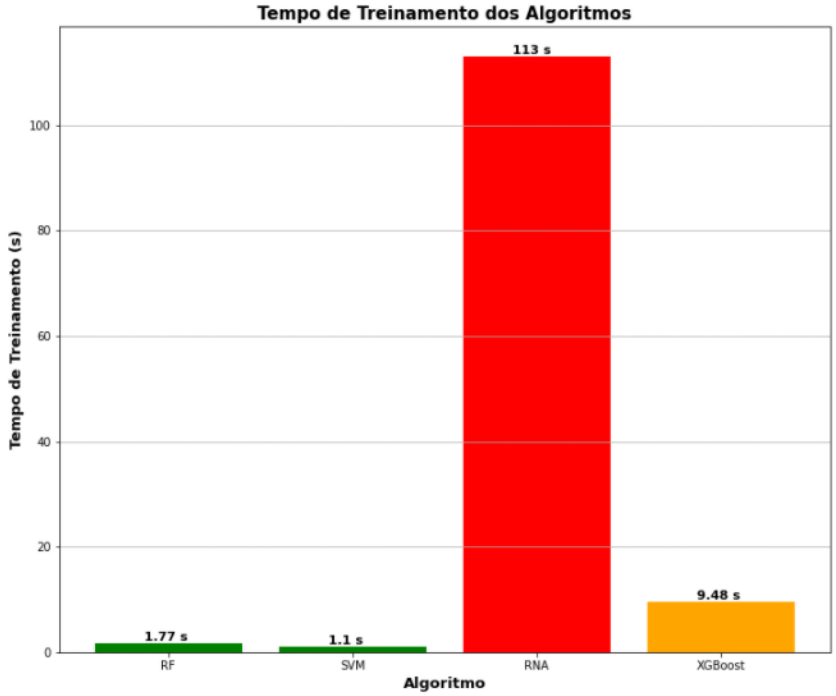
Fonte: elaborado pelo autor.

A análise gráfica confirma que a distribuição do XGBoost mais se aproximou dos valores originais, mas ainda assim existem duas regiões distintas, em

cargas até 1MW a precisão é excelente e superior a esse valor o algoritmo tende a subestimar os valores. Essa característica não é um problema dado o fato que o cliente conseguiria a conexão em indicações de uma potência menor para um ponto que suporta mais, e o contrário não é válido, uma vez que a necessidade do cliente seria maior do que a capacidade disponível da rede.

Na sequência de precisão pelo erro médio absoluto aparece o RF que está muito próximo ao XGBoost também graficamente e com características semelhantes. O próximo seria o SVM, mas graficamente vemos que realizou previsões concentradas aonde estavam a maioria das solicitações por isso apresentou um erro médio absoluto menor que o RNA, que por sua vez apresenta uma distribuição melhor para buscar as previsões, logo a escolha entre os algoritmos depende da aplicação.

Figura 29 – Tempo de treinamento dos algoritmos.



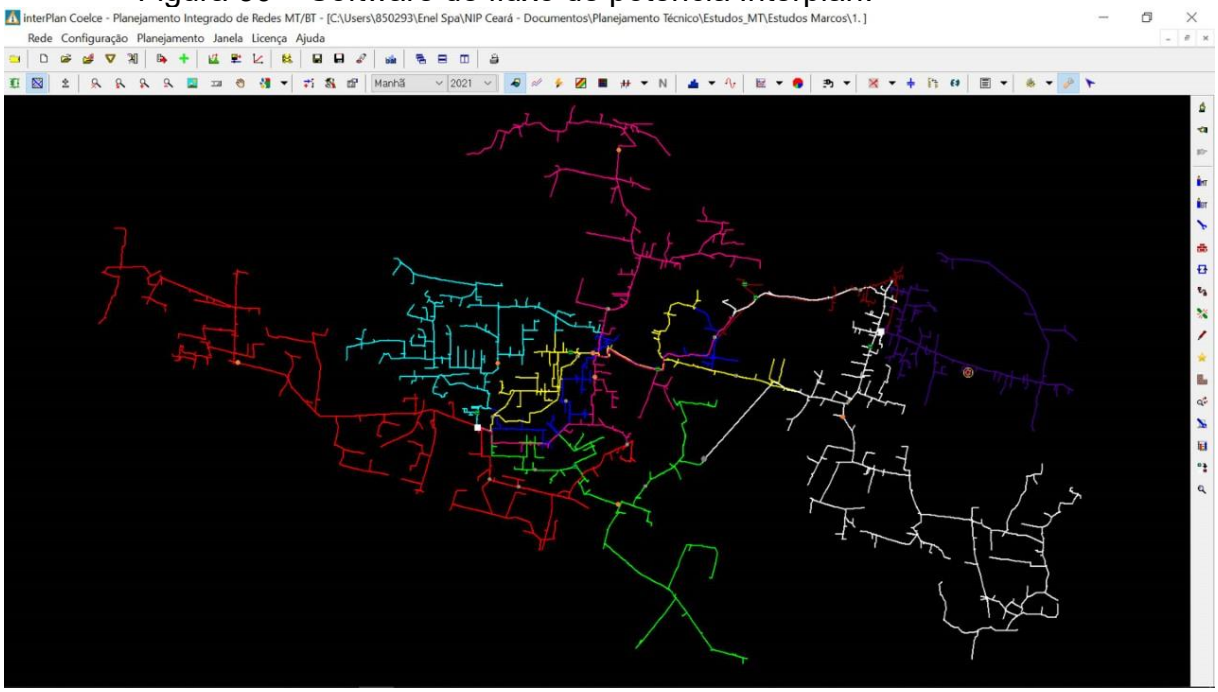
Fonte: elaborado pelo autor.

A última métrica é o tempo de treinamento onde é avaliada a velocidade para prever os valores e se essa métrica pode afetar de alguma maneira na atualização dos dados. Como observamos na Figura 29 até o pior algoritmo que é o RNA com 113 s não representa um tempo considerável para impactar na atualização do modelo, mas em comparação com o SVM sendo o melhor com 1,1 s significa um aumento de 1027%, expressando uma baixa eficiência.

A partir de todas as métricas utilizadas comprova-se que o XGBoost tem uma pequena vantagem frente ao RF em relação a previsões, no entanto, o RF tem mais velocidade para modelar. Os algoritmos SVM e RNA foram descartados para a próxima etapa de validação pelo comportamento e valores de previsão estarem mais distantes do que os outros dois concorrentes.

Para a etapa de validação foram selecionadas 80 amostras com potência e pontos aleatórios da rede em um programa de python, limitando a escolha em 20 de cada uma das quatro classes de consumo (residencial, comercial, industrial e outros). Essa nova base montada foi implementada nos algoritmos de RF e XGBoost para realizar previsões e comparar com os valores reais obtidos por um estudo de fluxo de potência.

Figura 30 – Software de fluxo de potência Interplan.



Fonte: elaborado pelo autor.

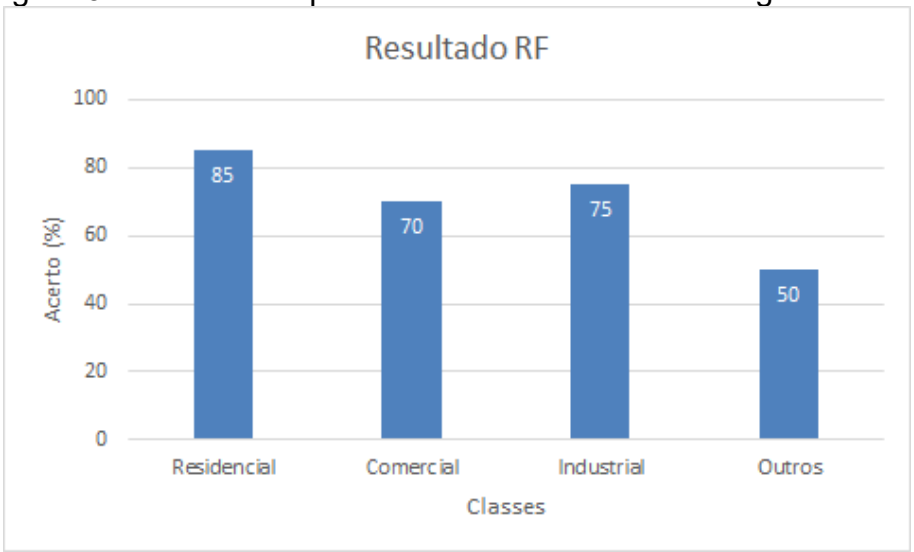
As análises dos resultados em todos os pontos foram realizadas em um programa especializado em fluxo de potência, chamado de Interplan. O Interplan, na Figura 30, é um programa que simula a rede de distribuição de média e baixa tensão utilizado pela Enel Distribuição Ceará para realizar estudos de curto-circuito, manobras, edição da rede, alterações topológicas, confiabilidade e análise de obras de atendimento em específico ou de expansão. Com o programa é possível avaliar o impacto em indicadores de qualidade, continuidade, tensões, capacidade dos cabos

e equipamentos com a inserção de cargas, gerações ou remanejamento de outros alimentadores.

No programa foram conectados à rede as cargas em todos os pontos escolhidos com curvas de cargas da sua respectiva classe escolhida previamente. Suas cargas começaram com 50 kW e aumentados gradativamente avaliando nível de tensão e corrente em todos os cabos do alimentador e nos equipamentos, nível de perdas e flexibilidade operacional em caso de uma emergência surgindo a necessidade de manobrar o alimentador. O valor de potência foi aumentado até que um dos fatores não fosse atendido, inviabilizando deste modo a conexão, onde esse valor seria considerado a potência máxima em que o sistema consegue suprir sem a necessidade de intervenções.

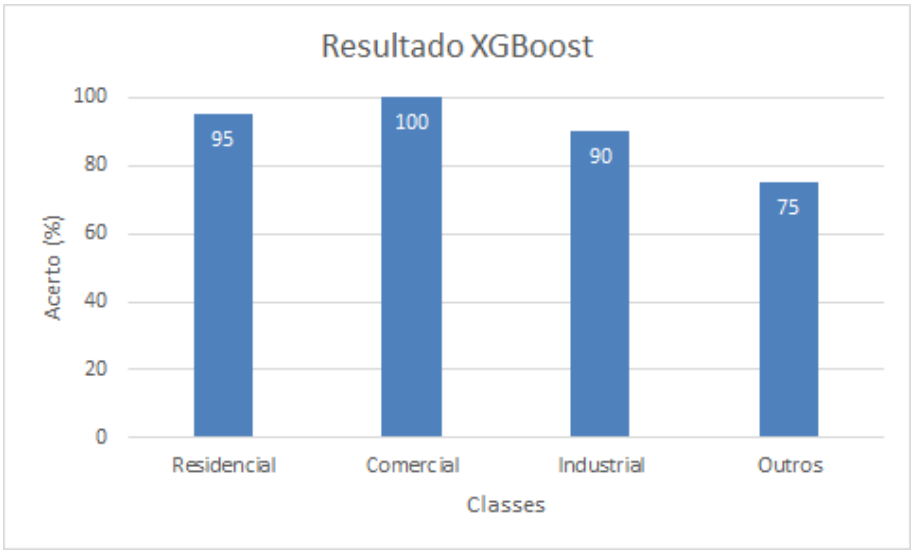
Os valores obtidos pelo estudo foram comparados com as previsões, no qual seriam considerados acertos nos casos em que a previsão for menor ou igual ao resultado da avaliação. A Figura 31 apresenta o percentual de acerto por classe do algoritmo RF e a Figura 32 do XGBoost.

Figura 31 – Resultado por classe do treinamento do algoritmo RF.



Fonte: elaborado pelo autor.

Figura 32 – Resultado por classe do treinamento do algoritmo XGBoost.



Fonte: elaborado pelo autor.

Como podemos observar, o algoritmo XGBoost apresentou resultados melhores do que o RF, com 90% de precisão total contra 70%, respectivamente. Comparando classe a classe, o XGBoost desempenhou com um percentual superior em todas as situações e o menor acerto foi de 75% na classe outros. Uma hipótese para explicar esse baixo rendimento é dada pela característica da rede de onde normalmente se conecta, pois os principais clientes dessa classe são rurais. O problema das regiões rurais é que a densidade de cliente é muito baixa, logo estão no final de alimentadores e a rede é mais velha e frágil. Mais especificamente, nesses trechos o condutor é do tipo monofilar com retorno por terra (MRT) que tem baixa capacidade de conexão de cargas, então qualquer valor previsto pelo modelo deverá ser superior ao esperado. Por outro lado, alimentadores e subestações mais robustas, que atendem às outras classes apresentaram um percentual de acerto maior do que 90%.

O resultado de desempenho dos algoritmos avaliando todas as métricas aponta que a modelagem do XGBoost é a ideal para utilização na finalidade escolhida, com os dados que estão dispostos. Deste modo, essa técnica foi escolhida para realizar o protótipo de mapear toda a rede do Ceará com suas previsões.

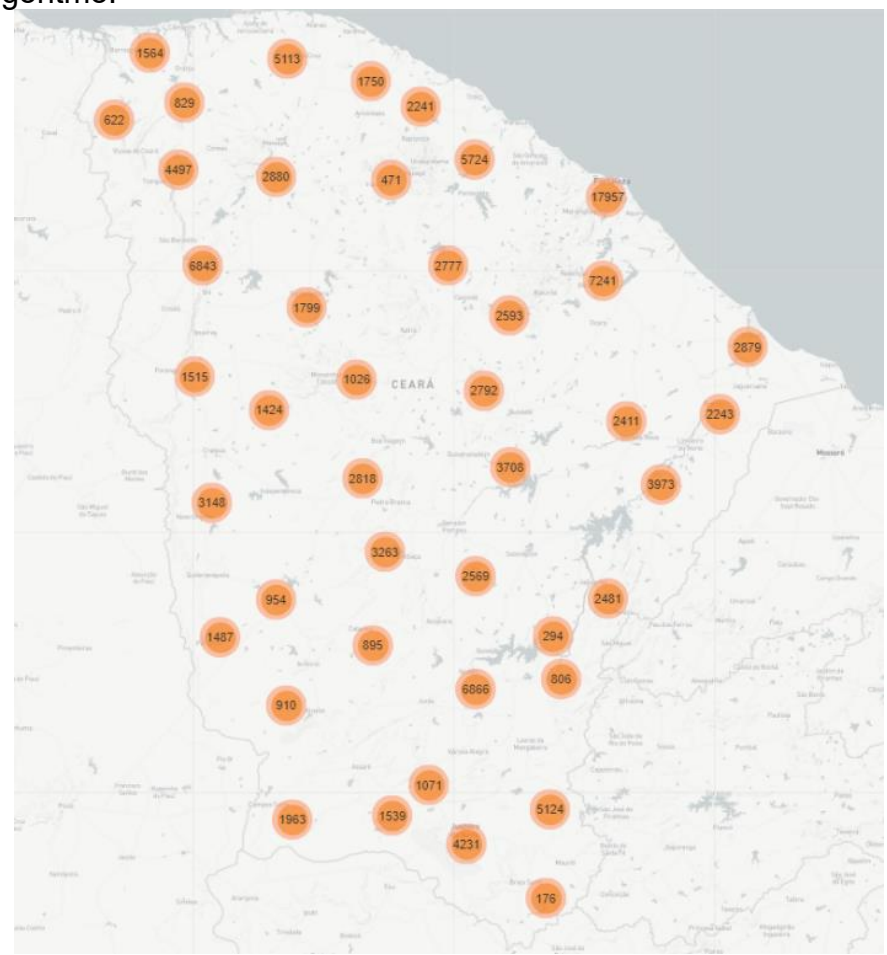
5.2 Protótipo

O protótipo consiste em realizar um mapa interativo do estado do Ceará com todos os pontos da rede elétrica de média tensão com sua respectiva previsão de potência. Seu objetivo é fornecer uma ferramenta preliminar confiável para tomada de decisões dos clientes e avaliações por parte da distribuidora.

O mapa foi desenvolvido em python e *HyperText Markup Language* (html) com o algoritmo XGBoost realizando a previsão de 127.467 pontos de transformação da rede elétrica de média tensão do Ceará, onde o usuário pode navegar pelo arquivo e visualizar a nível de rua a potência suportada pela rede sem a necessidade de obras.

Foram construídas duas versões pensando no detalhamento que a distribuidora pode oferecer como um teste inicial. A Figura 33 mostra a visão geral das duas versões, com o mapa do Ceará ao fundo.

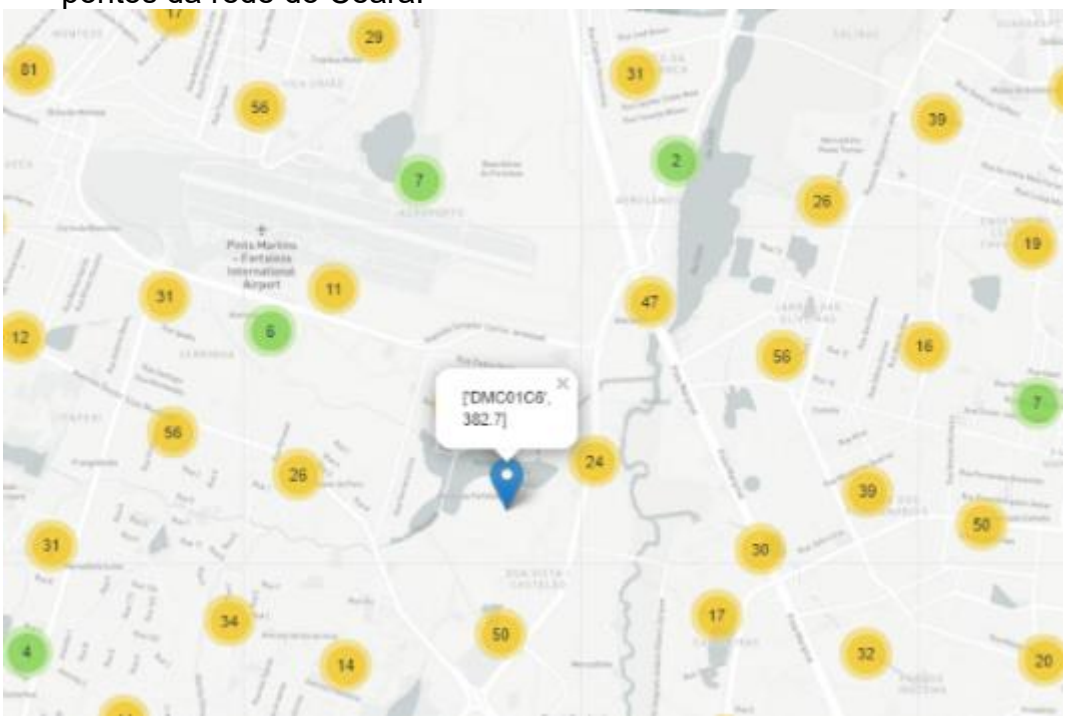
Figura 33 – Visão geral do mapa do Ceará com previsões do algoritmo.



Fonte: elaborado pelo autor.

A primeira versão na Figura 34, foi pensado como uma visão da distribuidora para identificar todos os pontos da rede, sendo eles marcadores azuis que ao pressionar é aberta uma caixa com o alimentador e qual a potência suportada.

Figura 34 – Visão da distribuidora com as previsões de todos os pontos da rede do Ceará.



Fonte: elaborado pelo autor.

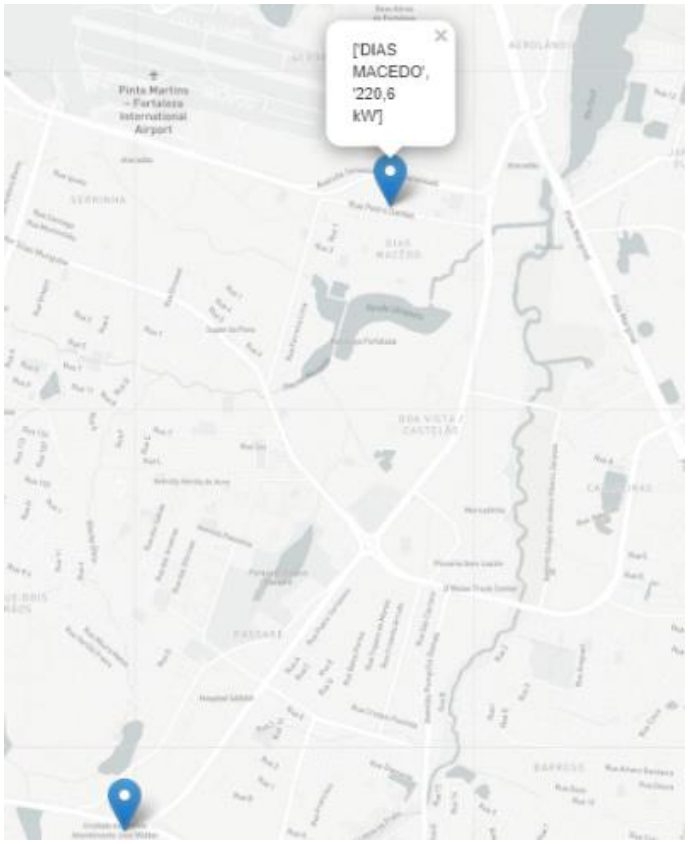
A segunda versão na Figura 35, é uma versão resumida, em que o ícone azul representa a subestação e ao pressionar o valor informado é a menor potência dentre todos os pontos da subestação, ou seja, é informado o pior caso e se o valor escolhido for menor do que o indicado, seu empreendimento pode ser atendido em qualquer lugar que a subestação alimente.

Para ambas as imagens os dados quando estão em regiões próximas são aglomerados em um único ponto com a quantidade total de barras da rede elétrica, que ao aproximar, são exibidas com indicadores azuis, conforme mencionados anteriormente.

Um ponto de atenção deve ser dado às atualizações, pois os dados representam a configuração atual da rede elétrica e quaisquer obras que sejam realizadas mudam completamente a previsão do local. Por isso para aplicação do

mapa tanto para distribuidora como para o acessante, o protótipo não deve ser tratado como uma verdade absoluta e sim como uma ferramenta de consulta.

Figura 35 – Visão do cliente com as previsões de todos as subestações do Ceará.



Fonte: elaborado pelo autor.

6 CONCLUSÃO

Este trabalho abordou métodos de aprendizado de máquina para mapear toda a rede elétrica do estado do Ceará com predições de potência para conexão de clientes sem a necessidade de intervenções por parte da distribuidora. O estudo foi norteado pelo problema da complexidade e velocidade das análises, e buscou-se avaliar o desempenho dos algoritmos para encontrar respostas precisas.

Os objetivos foram gerados como etapas para direcionar a pesquisa. A fundamentação teórica da pesquisa foi construída como a base científica para suportar a metodologia e os resultados. A metodologia foi desenvolvida a partir de procedimentos que atendesse aos modelos de análise preditiva previamente selecionados.

Os resultados foram apresentados como fruto de uma concepção intelectual, divido em etapas, buscando comprovar a efetividade da modelagem, em que o melhor resultado de predição de potência conseguiu estimar acertadamente em 90% dos casos.

Comparando os métodos de RF, SVM, RNA e XGBoost, aplicados no trabalho, verificou-se que o algoritmo XGBoost apresentou o melhor resultado com 145 kW de erro médio absoluto e na validação em uma das classes acertou 100%, mas com média geral de 90%. O RF chegou em 70% de acerto geral para o erro absoluto médio de 170 kW. Os outros dois algoritmos de RNA e SVM não foram testados em um caso real pois os seus erros de 286 kW e 235 kW, respectivamente, representam um valor significativo. Como destacado em todas as métricas o XGBoost foi o melhor algoritmo, sendo escolhido e posteriormente implementado no protótipo.

O mapa com as previsões mais precisas pode gerar impactos nos usuários que manuseiam a ferramenta. Pelo lado da distribuidora, esta tem a oportunidade de conhecer o sistema elétrico visualmente, entender as debilidades, propor ações de melhoria para reforçar o sistema e de forma proativa direcionar o crescimento da rede. Pelo lado do acessante, é possível ter uma ferramenta capaz de informar um dado que indique uma possível viabilidade técnica e financeira, de forma que a conexão também seja um critério para escolha do local e potência para o seu empreendimento.

Atualmente o AVT é enviado dentro de um prazo considerável e havendo a necessidade, o processo será repetido até obter êxito em localizar um ponto da rede que suporte o empreendimento. Essa demora ocasiona um desgaste entre cliente e

distribuidora e com a utilização da ferramenta, o usuário evitaria esse trâmite, ficando mais satisfeito pela velocidade em que a informação é gerada, refletindo na melhora da imagem da empresa. Além disso, a escolha de um local pelo consumidor sem solicitar vários orçamentos prévios reduz a carga de trabalho dos colaboradores, que podem focar em outras atividades que agreguem valor à distribuidora.

Algumas limitações foram detectadas principalmente para a atualização da base, porque a rede é dinâmica, repleta de obras e manobras que alteram os dados de entrada e assim o algoritmo mudaria sua previsão. Entretanto para a finalidade do mapa, uma atualização mensal ou trimestral atenderia a demanda de ambas as partes. Outra limitação é a capacidade computacional para processamento da grande quantidade de dados e criação do modelo de predição, sendo necessário algumas horas para gerar o resultado e havendo a necessidade de atualização o ciclo precisa ser repetido.

A partir das limitações apresentadas, indica-se como sugestão de trabalho futuro um estudo para avaliar outras possíveis variáveis preditoras que entreguem um resultado mais preciso e eficiente. Sugere-se também a aplicação de métodos de aprendizagem para prever ao mesmo tempo cargas e gerações, a partir das melhores variáveis preditoras disponíveis. Além disso, uma última sugestão seria uma avaliação de viabilidade dos mesmos preditores para aplicação em outras distribuidoras, levantando os dados e realizando um mapa de previsão para toda a rede de média tensão do Brasil.

REFERÊNCIAS

AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA. **Resolução Normativa 956**, de 7 de dezembro de 2021. Disponível em: <https://www2.aneel.gov.br/cedoc/ren2021956.html>. Acesso em: 15 jan. 2022.

AGÊNCIA NACIONAL DE ENERGIA ELÉTRICA. **Resolução Normativa 1.000**, de 7 de dezembro de 2021. Disponível em: <https://www.in.gov.br/en/web/dou/-/resolucao-normativa-aneel-n-1.000-de-7-de-dezembro-de-2021-368359651>. Acesso em: 15 jan. 2022.

ASADZADEH, S. M.; AZADEH, A.; ZIAEIFAR, A. **A neuro-fuzzy-regression algorithm for improved prediction of manufacturing lead time with machine breakdowns.** Concurrent Engineering, v. 19, n. 4, p. 269-281, 2011.

BREIMAN, Leo. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001.

BOTCHKAREV, Alexei. A NEW TYPOLOGY DESIGN OF PERFORMANCE METRICS TO MEASURE ERRORS IN MACHINE LEARNING REGRESSION ALGORITHMS. **Interdisciplinary Journal of Information, Knowledge & Management**, v. 14, 2019.

BOTCHKAREV, Alexei. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. **arXiv preprint arXiv:1809.03006**, 2018.

CEMIG. **Mapa de Disponibilidade de Minigeração**. https://www.geo.cemig.com.br/mca/Home/IndexData?tipoAcesso=1. Acesso em: 12 dez. 2021.

CHAI, T., & DRAXLER, R. R. (2014). **Root mean square error (RMSE) or mean absolute error (MAE)?** – Arguments against avoiding RMSE in the literature. Geoscientific Model Development, 7(3), 1247-1250.

CHEN, T.; GUESTRIN, C. **XGBoost: A scalable tree boosting system**, Association for Computing Machinery, pp. 785-794. 2016.

CHOK, N. S. **Pearson's Versus Spearman's and Kendall's Correlation Coefficients for Continuous Data**. Master's Thesis, University of Pittsburgh. 2010.

DHALIWAL, Sukhpreet Singh; NAHID, Abdullah-Al; ABBAS, Robert. **Effective intrusion detection system using XGBoost**. Information, v. 9. 2018.

EIA. **International Energy Outlook 2021.** Disponível em: https://www.eia.gov/outlooks/ieo/consumption/sub-topic-01.php. Acesso em: 19 dez. 2021.

ELERING. **E-gridmap.** Disponível em: https://vla.elering.ee/?lang=en. Acesso em: 12 dez. 2021.

EPE e MME. **Anuário Estatístico de Energia Elétrica 2021**. Disponível em: https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-160/topico-168/Anuário_2021.pdf. Acesso em: 27 dez. 2021.

FIEC. **Rotas Estratégicas setoriais 2025.** https://arquivos.sfiec.org.br/nucleoeconomia/files/files/rotas_estrategicas/EconomiadomarRota.pdf. Acesso em: 27 dez. 2021

FRIEDMAN, J. **Stochastic gradient boosting**. Computational Statistics & Data Analysis, 2002.

GÉRON, Aurélien. **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems**. O'Reilly Media, 2019.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

HAYKIN, Simon; NETWORK, N. A comprehensive foundation. **Neural networks**, v. 2, n. 2004, p. 41, 2004.

HASTIE; TREVOR; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. 2017.

KUHN, M.; JOHNSON, K. **Applied predictive modeling**. New York: Springer, 2013.

LEÃO, Ruth Pastôra Saraiva. **GTD – Geração, Transmissão e Distribuição de Energia Elétrica.** Apostila. 2018.

LIM, TS., LOH, WY. & SHIH, YS. A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. **Machine Learning 40**, 203–228 (2000).

NILSSON, N. J. **Introduction to Machine Learning: An Early Draft of a Proposed Textbook**. 1998.

ONS, EPE e CCEE. **Previsões de carga para o Planejamento Anual da Operação Energética 2022-2026.** Disponível em: https://www.epe.gov.br/sites-pt/publicacoes-dados-abertos/publicacoes/PublicacoesArquivos/publicacao-305/topico-603/Boletim%20Técnico%20Previsões%20de%20carga%20para%20o%20PLAN%202022-2026_Final.pdf. Acesso em: 19 dez. 2021

PEDREGOSA, F. Scikit-learn: Machine Learning in Python, **Journal of machine learning research**, v. 12, n. Oct, p. 2825-2830, 2011.

PRINCIPE, J. C.; EULIANO, N. R.; LEFEBVRE, W. C. **Neural and adaptive systems: fundamentals through simulations**. New York: Wiley, 2000.

RASCHKA, S.; MIRJALILI, V. **Python machine learning**. Birmingham-UK: Packt Publishing Ltd, 2017.

RAVADANEGH, S. N. e ROSHANAGH, R. G. **On optimal multistage electric power distribution networks expansion planning.** Int. J. Electr. Power Energy Syst., vol. 54, pp. 487–497, jan. 2014.

SILVEIRA, Gabriel Eugênio de Aguiar. **Aprendizado de máquina aplicado à predição de potência de geração distribuída na rede de distribuição de média tensão.** 2021. Trabalho de Conclusão de Curso. Universidade Federal do Ceará.

SHI, H. **Best-first Decision Tree Learning**. Thesis, Master of Science. The University of Waikato, Hamilton, New Zealand. 2007.

SVENSÉN, M.; BISHOP, C.M. **Pattern Recognition and Machine Learning**; Springer: Cham, Switzerland, 2007.

SVETNIK, Vladimir et al. Random Forest: a classification and regression tool for compound classification and QSAR modeling. **Journal of chemical information and computer sciences**, v. 43, n. 6, p. 1947-1958, 2003.

UNIVERSIDADE FEDERAL DO CEARÁ. Biblioteca Universitária. **Guia de normalização de trabalhos acadêmicos da Universidade Federal do Ceará**. Fortaleza: Biblioteca Universitária, 2013. Disponível em: https://biblioteca.ufc.br/wp-content/uploads/2019/10/guia-de-citacao-06.10.2019.pdf. Acesso em: 9 jun. 2021.

VAN RUIJVEN, Bas J.; DE CIAN, Enrica; SUE WING, Ian. **Amplification of future energy demand growth due to climate change.** Disponível em: https://doi.org/10.1038/s41467-019-10399-3. Acesso em: 19 dez. 2021

WESTERN POWER DISTRIBUTION. **Network Capacity Map**. https://www.westernpower.co.uk/our-network/network-capacity-map-application. Acesso em: 12 dez. 2021.

ZAKI, M. J.; MEIRA JR, W.; MEIRA, W. **Data mining and analysis**: fundamental concepts and algorithms. Cambridge University Press, 2014.