INSTITUTO FEDERAL
DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
Minas Gerais
Campus Bambuí

Lucas dos Anjos de Castro

APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS PARA GERAÇÃO DE MODELOS PREDITIVOS DE GERAÇÃO E CONSUMO DE ENERGIA ELÉTRICA: UM ESTUDO DE CASO NO IFMG – *CAMPUS* BAMBUÍ

Bambuí – MG
2021

LUCAS DOS ANJOS DE CASTRO

APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS PARA GERAÇÃO DE MODELOS PREDITIVOS DE GERAÇÃO E CONSUMO DE ENERGIA ELÉTRICA: UM ESTUDO DE CASO NO IFMG – *CAMPUS* BAMBUÍ

Trabalho de conclusão de curso apresentado ao Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais – *Campus* Bambuí como requisito parcial para obtenção do grau de Bacharel em Engenharia de Computação.

Orientador: Prof. Me. Felipe Lopes de Melo Faria
Coorientador: Prof. Me. Calebe Giaculi Júnior

Bambuí – MG
2021

C355a 2021

Castro, Lucas dos Anjos de.
Aplicação de técnicas de mineração de dados para geração de modelos preditivos de geração e consumo de energia elétrica: um estudo de caso no IFMG – *Campus* Bambuí / Lucas dos Anjos de Castro. - Bambuí, 2021.
54 f. : il. (algumas color.).

Orientador: Felipe Lopes de Melo Faria.
Trabalho de Conclusão de Curso (Graduação em Engenharia da Computação) - Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais - *Campus* Bambuí.

1. Mineração de dados - Computação. I. Faria, Felipe Lopes de Melo (orientador). II. Instituto Federal de Educação Ciência e Tecnologia de Minas Gerais – *Campus* Bambuí. III. Título.

CDD: 005.3

Lucas dos Anjos de Castro

APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS PARA GERAÇÃO DE MODELOS PREDITIVOS DE GERAÇÃO E CONSUMO DE ENERGIA ELÉTRICA: UM ESTUDO DE CASO NO IFMG – *CAMPUS* BAMBUÍ

Trabalho de conclusão de curso apresentado ao Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais – *Campus* Bambuí como requisito parcial para obtenção do grau de Bacharel em Engenharia de Computação.

Aprovada em 30 de setembro de 2021, pela banca examinadora:

Prof. Me. Felipe Lopes de Melo Faria - IFMG - *Campus* Bambuí (Orientador)

Prof. Me. Calebe Giaculi Junior - IFMG - *Campus* Bambuí (Coorientador)

Prof. Dr. Marcos Roberto Ribeiro - IFMG - *Campus* Bambuí

Prof. Me. Francisco Heider Willy dos Santos - IFMG - *Campus* Bambuí

Documento assinado eletronicamente por **Felipe Lopes de Melo Faria**, **Professor**, em 30/09/2021, às 11:02, conforme art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **Calebe Giaculi Junior**, **Professor**, em 30/09/2021, às 11:03, conforme art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **Marcos Roberto Ribeiro**, **Professor**, em 30/09/2021, às 11:06, conforme art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **Francisco Heider Willy dos Santos**, **Professor**, em 30/09/2021, às 11:06, conforme art. 1º, III, "b", da Lei 11.419/2006.

A autenticidade do documento pode ser conferida no site https://sei.ifmg.edu.br/consultadocs informando o código verificador **0967768** e o código CRC **63BEE789**.

BAMBUÍ-MG
2021

Dedico este trabalho aos meus pais, Magnólia e André, que sempre me incentivaram nos estudos e tornaram possível esta jornada.

AGRADECIMENTOS

Aos meus pais, meus tios Lindoval e Roberto, agradeço por sempre acreditarem em mim, incentivarem e ajudarem em tudo o que foi possível.

Aos meus orientadores, agradeço não somente a dedicação ao trabalho, com conselhos e ideias, mas como exemplos que levarei para a vida.

“E então, uma quinta-feira, quase dois mil anos depois que um homem foi pregado num pedaço de madeira por ter dito que seria ótimo se as pessoas fossem legais umas com as outras para variar...”

(Douglas Adams)

RESUMO

O planejamento de gastos com energia elétrica vem se mostrando cada vez mais uma tarefa de difícil realização. Já que o ambiente em que grandes instituições operam está constantemente sujeito a mudanças inesperadas, é comum que busquem entender seu consumo de recursos para encaixá-lo em suas receitas. O presente trabalho teve como objetivo a aplicação de técnicas de Mineração de Dados, como KNN, RNA, Regressões e outras, a fim de gerar modelos preditivos que possam contribuir na análise de dados sobre a geração da usina fotovoltaica e consumo de energia elétrica do IFMG – *Campus* Bambuí. Para a análise da predição, foram utilizadas métricas de avaliação dos modelos gerados, tais como MAPE, MAE e RMSE. O trabalho apresenta pesquisas correlatas, a fim de se observar métricas, técnicas e parâmetros empregados no estado da arte. Emprega-se a ferramenta para Mineração de Dados WEKA, explorando seus recursos para edição e visualização de dados, além de sua biblioteca de algoritmos. Ressalta-se destaque positivo para o algoritmo de RNA e Regressão Linear, e negativo para o KNN e Random Forest. Além disso, os dados de consumo de energia elétrica mostraram-se mais difíceis de serem preditos, e a adição de atributos mostrou-se ora contribuir para melhorar, ora contribuir para piorar os resultados. Por fim, revela-se uma associação entre a presença de precipitação e a precisão dos modelos gerados.

Palavras-chave: Mineração de Dados. Inteligência Artificial. Energia Elétrica. Energia Fotovoltaica.

ABSTRACT

Planning for electrical energy expenses is increasingly proving to be a difficult task. Since the environment in which major institutions operate is constantly subject to unexpected changes, it is common for them seeking to understand their resources consumption in order to fit it into their revenues. The present work aims to apply data mining techniques such as KNN, ANN, Regressions and others, in order to generate predictive models that may contribute to the analysis of data on the production of photovoltaic power plants and the electrical energy consumption at IFMG – Campus Bambuí. For the prediction analysis, metrics will be used to evaluate the generated models, such as MAPE, MAE and RMSE. The paper presents related research in order to observe metrics, techniques, and parameters employed in the state of the art. The data mining tool WEKA is used, exploring its features for data editing and visualization as well as its library of algorithms. The highlights stand as positive for the ANN and Linear Regression algorithms and negative for KNN and Random Forest. In addition, the electricity consumption data have proven more difficult to predict and the addition of attributes proved to either improve or worsen the results. Finally, an association is revealed between the presence of precipitation and the accuracy of the generated models.

Keywords: Data Mining. Artificial Intelligence. Electric Power. Photovoltaic Energy.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ºC - Graus Celsius

DDR - *Double Data Rate*

GeForce - *General Electronic Facsimile Optimized for Repair and Ceaseless Exploration*

GTX - *Giga Texel Shader eXtreme*

HFP - Horário Fora de Ponta

HP - Horário de Ponta

ID - *Identity*

IFMG - Instituto Federal de Educação, Ciência e Tecnologia de Minas Gerais

KNN - k *Nearest Neighbor*

KWh - Kilowatt-hora

MAE - *Mean Absolute Error*

MAPE - *Mean Absolute Percentage Error*

mB - Milibar

mm - Milímetros

MSE - *Mean Squared Error*

m/s - Metros por segundo

PIB - Produto Interno Bruto

RAM - *Random Access Memory*

RF - *Random Forests*

RLM - Regressão Linear Múltipla

RLS - Regressão Linear Simples

RMSE - *Root Mean Squared Error*

RNA - Redes Neurais Artificiais

R^2 - Coeficiente de Determinação

SVM - *Support Vector Machine*

WEKA - *Waikato Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO

O planejamento de gastos vem se mostrando cada vez mais ser uma tarefa de difícil realização. Já que o ambiente em que grandes instituições operam está constantemente sujeito a mudanças inesperadas, é comum que busquem entender seu consumo de recursos para encaixá-los em suas receitas. Em organizações de ensino, ocorre um paralelo entre constantes demandas por vagas e resultados com a oferta de novos cursos e recursos atrelados a suas fontes de receita (QUEIROZ; QUEIROZ; HÉKIS, 2011).

Métricas para planejamento em si apresentam todo um campo de estudo robusto e grande gama de problemáticas. Ressaltam-se, na medida que avança a ciência, preocupações com o meio ambiente através da utilização de fontes de energias limpas. Outro ponto importante é que, na medida em que as civilizações crescem e prosperam, acompanha-se uma crescente demanda por energia elétrica. Constata-se que, em instituições com um relevante tempo de existência, ocorrem volumes consideráveis de informações armazenadas em consequência de documentação; porém, muitas vezes, esses dados não são aproveitados após seu uso burocrático. Com os avanços tecnológicos atuais, cada vez mais, as corporações buscam armazenar de forma digital o maior número possível de informações, encontrando, assim, um problema recorrente: grandes bases de dados de difícil entendimento. Mesmo com a atual grande capacidade de processamento de informação, é trabalhoso atrelar sentido ao que é analisado, de maneira a tornar o processamento útil. Um processo de Mineração de Dados apresenta resultados que técnicas estatísticas habituais não são capazes de identificar (SCHUCH *et al*., 2010).

O presente trabalho propõe a aplicação de técnicas de Mineração de Dados à base existente no Instituto Federal de Minas Gerais (IFMG) - *Campus* Bambuí, referente à geração de energia da microusina fotovoltaica e a dados a serem coletados sobre o consumo energético do *campus*. Busca-se predizer o consumo e a geração de energia do *campus*, podendo servir de subsídio para auxiliar no planejamento dos gastos de consumo energético e também em uma possível elaboração de um plano de autossuficiência. Em suma, objetiva-se a criação de modelos preditivos eficientes que possam vir a auxiliar o IFMG.

1.1 Objetivo geral

O objetivo geral do presente trabalho é aplicar técnicas de Mineração de Dados para gerar modelos preditivos que possam auxiliar na economia de energia e na projeção da produção de energia da usina fotovoltaica do IFMG - *Campus* Bambuí.

1.2 Objetivos específicos

Os objetivos específicos do presente estudo são:

- Caracterizar os dados de consumo e geração de energia em bases de dados;
- Gerar modelos de regressão para predição de dados de consumo e produção de energia elétrica a partir de dados históricos, utilizando-se de técnicas de Mineração de Dados, e medir sua acurácia.

1.3 Justificativa

Como explora Becker (2012), após a Revolução Industrial, a atividade humana passa a modificar o meio ambiente de maneira semelhante a fenômenos naturais, sempre consumindo de maneira crescente, aliada ao crescimento demográfico e à urbanização. Em escala global, destaca-se o crescimento da necessidade e da escassez de recursos para a geração de energia elétrica de maneira tão intensa que, em alguns casos, resulta em guerras. Diante de tamanhas adversidades na obtenção de energia elétrica, destaca-se a importância do seu uso de maneira consciente e também preocupações em economizá-la.

Outro fator a ser considerado é o financeiro. Frente a um cenário de redução de receita, quaisquer possibilidades de economia tornam-se interessantes. Olhando por outra perspectiva, com um entendimento melhor sobre possibilidades de economia, é possível direcionar recursos para outras áreas de relevância em uma instituição.

Dentro do escopo da geração de energia limpa, enquadra-se a redução de emissão de gases poluentes como um dos pontos da redução de agressão ao meio ambiente, proporcionada pela energia fotovoltaica. Este meio de produção de energia elétrica tem ganhado muita força nos últimos anos ao mostrar eficiência e custos atrativos de implantação. Compreende-se a Mineração de Dados como uma área muito rica e útil na resolução de uma gama enorme de problemáticas. O presente trabalho contribuirá com a aplicação de

algoritmos do segmento a um âmbito de relevante interesse - o consumo de energia elétrica. Futuramente, outras pesquisas poderão usar como base este trabalho para estudos parecidos aplicados a outros *campi* do IFMG, assim como para análises do próprio *Campus* Bambuí em possíveis investimentos na expansão do sistema fotovoltaico.

2 REFERENCIAL TEÓRICO

Como abordado por Schuch *et al*. (2010), tempos atrás, fatores limitadores do armazenamento digital de dados eram o processamento e a capacidade física. Com relevantes avanços tecnológicos, a dificuldade agora encontra-se em extrair conhecimento de um amontoado de informações que, em um primeiro momento, não mostra utilidade alguma. Neste âmbito, surgem os métodos de Mineração de Dados, mostrando-se muito úteis em auxiliar pesquisas a evidenciar informações proveitosas dentro de um conjunto de dados.

Mineração de Dados trabalha diretamente ligada a outras áreas do conhecimento, como estatística e computação, não tendo em vista a substituição destas, mas sim apoderando-se de suas métricas para trabalhar. É interessante pontuar que a Mineração mostra-se um campo que evolui ao passo das necessidades dos demais, tais como agricultura, comércio, física, entre outros. Ressalta-se que este campo abrange a área de predição, fornecendo capacidades de, baseando-se em dados analisados, prever comportamentos através de um modelo (TAN; STEINBACH; KUMAR, 2009).

Tan, Steinbach e Kumar (2009) verificam a existência de duas principais categorias dentro da Mineração de Dados: tarefas de predição e tarefas descritivas. Também abordam exemplos de aplicações, como o típico caso de prever o tipo de uma flor baseando-se em suas características físicas, predição em oportunidades de vendas de produtos analisando itens relacionados, determinação do assunto de um artigo levando em conta as ocorrências de determinadas palavras dentro do texto, ou até mesmo explorar registros de transações bancárias para detectar fraudes em cartões de crédito. Dentro da predição, existem dois tipos de tarefas: classificação e regressão. Na classificação, são usados os dados das variáveis disponíveis para se determinar a classe de uma variável-alvo, por exemplo, basear-se em padrões de comportamentos de clientes online para prever se determinado cliente fará uma compra em uma loja ou não. Já na regressão, o objetivo é basear-se nos valores das variáveis disponíveis, valendo-se de seus dados históricos para prever o valor futuro de uma variável, por exemplo, basear-se no histórico das ações de uma empresa para determinar o seu valor futuro.

Um dos algoritmos mais utilizados dentro da mineração é o de Redes Neurais Artificiais (RNA), tendo sua origem na tentativa de criar um algoritmo que simulasse o funcionamento de sistemas neurais biológicos. Como explorado por Tan, Steinbach e Kumar (2009), “análoga à estrutura do cérebro humano, uma RNA é composta de um conjunto

interconectado de nodos e ligações direcionadas". Estas estruturas vão desde modelos mais simples de RNA, como o *perceptron*, até redes mais complexas, como um RNA Multicamadas.

Verifica-se, também, como exploram Han, Pei e Kamber (2011), o algoritmo k - *Nearest Neighbors* (KNN), que utiliza medidas de proximidade (distância Euclidiana ou de Manhattan, por exemplo) para determinar os k elementos mais próximos do valor que se quer classificar ou prever e o associa aos valores mais frequentes dentro da vizinhança.

Outra técnica interessante é a Árvore de Decisão, que consiste em uma estrutura baseada em uma série de questionamentos cuidadosamente organizados sobre as características do elemento de teste, "cada vez que recebemos uma resposta, uma questão seguinte é feita até que cheguemos a uma conclusão sobre o rótulo da classe do registro" (TAN; STEINBACH; KUMAR, 2009, p. 177). Da Árvore de Decisão, surgem as Florestas Aleatórias, ou *Random Forests* (RF), um classificador que se baseia na combinação dos resultados de diversas árvores de decisão, com a geração de cada árvore embasada nos valores independentes de vetores aleatórios.

Ainda assim, existem os modelos de Regressão Linear Simples (RLS) e Múltipla (RLM), que consistem em representações por meio de modelos matemáticos para previsões que determinam relações entre uma variável-alvo (que se deseja prever) e variáveis independentes. No algoritmo de Regressão Linear, define-se o modelo baseado na relação com uma variável independente, diferentemente do algoritmo de Regressão Linear Múltipla, em que se determina o modelo com base na relação com diversas variáveis independentes (RODRIGUES, 2012).

Um aspecto de relevância quando são empregadas técnicas de previsões é a medida de acurácia. De Myttenaere *et al*. (2016) apontam que modelos preditivos são, muitas vezes, escolhidos baseando-se em medidores de acurácia, como é o caso do *Mean Absolute Percentage Error* (MAPE), que é amplamente utilizado em predições por sua praticidade no entendimento da porcentagem de erro.

Os dados utilizados no trabalho dividem-se em duas classes: consumo e geração de energia elétrica. Consumo consiste no histórico de demanda de energia elétrica presente em contas de energia, documentadas pelo *campus* e disponíveis para acesso público pelo Portal da Transparência. Por questões de praticidade, os dados foram obtidos por concessão do Setor de Planejamento do *campus*. Já a geração faz jus à energia produzida dentro do *campus* pela microusina fotovoltaica. Os dados são de acesso público, através do site portal IFMG -

Campus Bambuí[1].

Sistemas de energia solar fotovoltaica convertem diretamente a radiação solar em energia elétrica (GASPARIN; KRENZINGER, 2017), destacando-se tanto nacional quanto internacionalmente. Com o passar dos anos, essa tecnologia vem ganhando eficiência e tornando-se economicamente viável, com quedas cada vez maiores nos preços de instalação (MACHADO; MIRANDA, 2015). Um sistema fotovoltaico é uma ferramenta composta de diversos elementos visando à conversão da energia solar em eletricidade. Seu principal elemento é a célula fotovoltaica, ou módulo fotovoltaico, que pode ser agrupada e disposta de inúmeras maneiras, conforme a necessidade e finalidade do sistema (SILVA; COSTA, 2018).

2.1 Estado da arte

O uso eficiente de técnicas de Mineração de Dados oferece benefícios às mais diversas instituições. Aponta-se que, das dificuldades de gestão competente, nascem as necessidades de se entender as demandas presentes no funcionamento das organizações. Enfatizam-se os casos de instituições públicas, em que os recursos são cada vez mais escassos, e conseguir economias em um setor pode possibilitar acréscimos em outros, enriquecendo seus serviços ao público-alvo. Constata-se a relevância de estudos neste âmbito em períodos de crises e necessidades, como o vivido atualmente devido à pandemia do novo Coronavírus. A situação é de profunda crise socioeconômica, contração drástica das atividades produtivas, além de previsões de decréscimo no Produto Interno Bruto (PIB) a ponto de se tornar a pior recessão da história brasileira, sendo, inclusive, comparada a um cenário de guerra (SARAIVA; OLIVEIRA; MOREJON, 2020).

Outra vertente expressiva é a preocupação com recursos para a geração de energia. Visto que os recursos naturais, como carvão e petróleo, são finitos, é importante buscar por fontes renováveis e limpas, destacando-se sistemas fotovoltaicos por sua eficiência e praticidade. Brito e Silva (2006) evidenciam que a energia fotovoltaica apresenta alta viabilidade econômica, pois, operando em condições favoráveis, módulos fotovoltaicos levam apenas de 2 a 3 anos para apresentar retorno energético e, ao longo de sua vida, produzem cerca de dez vezes mais energia do que foi gasto para sua fabricação.

Na sequência, serão abordados temas relacionados à energia elétrica, como encontrados em Terra (2003); Silva (2012); Alves, Lotufo e Lopes (2013); Abreu *et al.*

[1] https://www.bambui.ifmg.edu.br/sunnyportal.html

(2017); Kopiler *et al.* (2019) e Jawad *et al*. (2020).

Em Kopiler *et al.* (2019), abordam-se características que evidenciam o porquê da utilização de técnicas de Aprendizado de Máquina para gestão de sistemas elétricos; dentre elas, destaca-se a área de predição. O trabalho fomenta que, à medida que os sistemas elétricos evoluem, são necessárias técnicas mais aprimoradas para se obter "funcionamento seguro e confiável" (KOPILER *et al*., 2019, p. 27). Como resultado, é necessário lidar com um volume cada vez maior de informações, inviabilizando, assim, técnicas simples de análise, tornando atrativos os recursos da Mineração de Dados.

Um estudo sobre predições envolvendo energia elétrica, promovido por Terra (2003), constata a dificuldade de se trabalhar na produção de modelos quando se consideram dados históricos de consumo de energia. Atribui-se o impasse a incertezas decorrentes de medições incorretas, problemas de fornecimento de energia elétrica ou, ainda, padrões inconsistentes de consumo. Outro ponto abordado no trabalho é o enriquecimento da base de dados através da adição de dados meteorológicos, como os atributos de temperaturas máxima e mínima. Além disso, o autor mostra o comportamento e as particularidades da aplicação de algoritmos de Redes Neurais Artificiais e também de Modelos Lineares empregando as métricas de validação *Mean Squared Error*[2] (MSE), *Root Mean Squared Error*[3] (RMSE), *Mean Absolute Error*[4] (MAE), MAPE e Coeficiente de Determinação (R^2).

Silva (2012) aborda o desenvolvimento de um modelo de predição híbrido de Regressão com Redes Neurais Artificiais, utilizando treinamento através do algoritmo de Levenberg-Marquardt. Tal dissertação visa melhorar os resultados de previsões de cargas elétricas obtidos com técnicas de predição mais tradicionais dentro da Mineração de Dados por meio da aplicação do modelo híbrido. A autora aponta que diversas literaturas abordam modelos de predição híbridos e destacam ótimos resultados quando comparados a modelos únicos. Unindo as especialidades de cada modelo, foi possível aliar as características desejadas para um problema oportuno. Ao examinar os resultados das predições, Silva (2012) aponta, por meio das medidas de acurácia MAPE e Erro Máximo, que, apesar de bons resultados para todas as técnicas, em todas as aplicações abordadas, o modelo híbrido apresenta o melhor resultado. Por fim, visando aperfeiçoar os resultados obtidos, a autora sugere para futuros trabalhos abordar diferentes técnicas de validação, aplicar o modelo apresentado em previsões de maior extensão, utilização de uma rede neural de outra família,

[2] Erro Médio Quadrático
[3] Raiz do Erro Médio Quadrático
[4] Erro Absoluto Médio

entre outros.

No artigo de Alves, Lotufo e Lopes (2013), é explorado um método chamado de *stepwise*, onde se estudam quais os atributos de maior impacto sobre os resultados da predição, como técnica de filtragem. Neste trabalho, são aplicadas Regressão Linear Múltipla e Redes Neurais Artificiais para a predição de demanda de cargas elétricas, com o diferencial do uso do método *stepwise* visando a melhores resultados. Tentando prever a demanda de carga sobre intervalos de 24 e 48 horas, os autores utilizaram as medidas de acurácia MAPE e Erro Máximo, relatando, assim, ótimas previsões em relação aos valores reais. Por fim, Alves, Lotufo e Lopes (2013) concluem que o método *stepwise* agiliza a seleção de atributos, apontam resultados satisfatórios nas predições e sugerem para futuros trabalhos o uso do algoritmo de Levenberg-Marquardt.

Em Abreu *et al.* (2017), estuda-se a predição de cargas elétricas através de uma Rede Neural Artificial de tipo *Fuzzy*, com o diferencial da utilização de treinamento continuado. O trabalho tem como foco a comparação dos resultados de predição entre um modelo com treinamento convencional e outro com treinamento continuado. Mais uma vez, utilizando-se das medidas de acurácia MAPE e Erro Máximo, os autores apontam resultados satisfatórios para ambos os modelos preditivos; porém, destacando melhores resultados com a utilização do treinamento continuado. Por fim, Abreu *et al.* (2017) comparam os aspectos tanto da utilização de treinamento tradicional quanto continuado, mostrando como cada técnica se encaixa em distintas situações.

Jawad *et al*. (2020) propõem um modelo otimizado para previsão de demanda elétrica de menor custo fazendo uma correlação entre cargas elétricas e atributos meteorológicos. Os autores exploram a importância da relação entre energia elétrica, com suas formas de produção e fatores impactantes, além de uma análise das implicações do quanto os modelos gerados podem ajudar em planejamento, organização e benefícios financeiros. São aplicados os algoritmos Regressão Linear Múltipla, *Support Vector Machine*[5] (SVM), KNN, *Random Forests* e *AdaBoost*, e a validação dos modelos é feita através das métricas MAE, RMSE e MAPE. Os atributos meteorológicos analisados foram a Temperatura, em Graus Celsius (ºC); o Ponto de Orvalho, também em Graus Celsius; a Porcentagem de Umidade do Ar; Precipitação, em milímetros (mm); Velocidade do Vento, em metros por segundo (m/s); e a Presença de Nuvens. Por fim, os autores evidenciam que os atributos que melhor se relacionaram com a demanda elétrica foram o ponto de orvalho e a temperatura.

[5] Máquinas de Vetores de Suporte

Ao se estudar a aplicação destas técnicas, entende-se o poder de seus resultados, podendo-se observar oportunidades de empregá-las para ajudar no planejamento energético do IFMG. De maneira semelhante, este trabalho objetiva aplicar métodos voltados à predição, e assim indicar valores próximos dos futuros para auxiliar o *campus* a entender o seu consumo e produção de energia elétrica. Destaca-se a possibilidade de observar onde os modelos aplicados por autores anteriores falharam na busca de aprimoramentos e avanços, como a geração de novos modelos e a utilização de técnicas mais atuais, buscando-se obter melhores resultados.

O presente trabalho busca alinhar as técnicas de predições mais adequadas à temática para predição de demanda e geração de energia elétrica. Como abordaram os diversos autores apresentados previamente, a predição de demanda é amplamente explorada, traz resultados relevantes nos modelos propostos, além de possibilitar amplo entendimento das demandas do ambiente estudado. Este trabalho estende a análise da demanda energética aos dados existentes sobre a geração advinda da microusina fotovoltaica do IFMG - *Campus* Bambuí, incitando, também, futuras análises nos mais diversos *campi* que dispõem de usinas fotovoltaicas. Compreende-se que, como resultados, serão possíveis análises sobre os dados e modelos abordados que possam vir a beneficiar o *campus*, mais uma vez atentando para a necessidade de estudos promovendo possibilidades de economia em períodos difíceis como o vivido atualmente, com a pandemia do Coronavírus.

3 METODOLOGIA

O presente capítulo apresenta as classificações do trabalho, assim como um detalhamento dos processos metodológicos empregados para a conquista dos objetivos propostos. A Seção 3.1 apresenta as classificações da pesquisa, seguida pela Seção 3.2, que explora uma sequência de passos para a solução proposta. Já a Seção 3.3 ilustra as ferramentas utilizadas, e o capítulo se encerra com a Seção 3.4, indicando a metodologia de desenvolvimento empregada.

3.1 Classificação do trabalho

O presente trabalho caracteriza-se, quanto à natureza, como uma pesquisa aplicada, visando ao emprego prático e imediato do conhecimento (MORESI, 2003). Os objetivos se enquadram, na classificação de Gil (2002), como pesquisa descritiva, uma vez que se visa descrever e entender as relações da usina fotovoltaica do IFMG - *Campus* Bambuí com o seu gasto de energia, de maneira a estabelecer relações entre as variáveis que compõem o sistema, além da utilização da coleta dos dados de geração e consumo de energia elétrica.

Quanto aos procedimentos, primeiramente, classifica-se a pesquisa como experimental, por trabalhar com variáveis que influenciam os objetos de estudo de maneira a observar os efeitos decorridos (GIL, 2002). Posteriormente, os procedimentos classificam-se também como um estudo de caso, por terem como objeto de estudo a usina fotovoltaica e o consumo de energia elétrica exclusivamente do IFMG - *Campus* Bambuí (GIL, 2002).

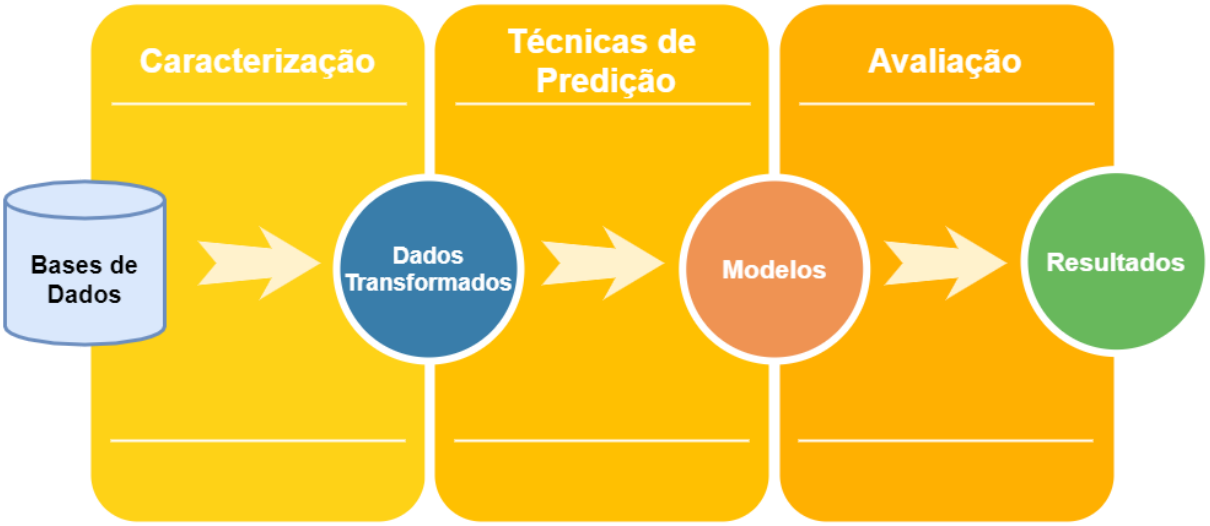
A problemática do trabalho se categoriza, conforme explorado por Moresi (2003), como quantitativa, por lidar com dados numéricos, conter a aplicação de algoritmos, avaliar a precisão por meio de métodos estatísticos, como as medidas de acurácia, e promover a análise desses números para se obter conclusões. Por fim, conforme explicado por Wazlawick (2017), o trabalho se adequa à apresentação de algo presumivelmente melhor por comparar a aplicação de diferentes algoritmos para predição.

3.2 Solução

Para a realização do trabalho, definiram-se três etapas (Figura 1). Na etapa “Caracterização”, foram realizados a escolha, o estudo e a caracterização das bases de dados,

com uma parte contendo os dados de geração de energia elétrica, e outra, os dados de consumo de energia elétrica do *campus*. Na etapa "Técnicas de Predição", foram aplicadas técnicas computacionais para predição, utilizando os algoritmos k *Nearest Neighbor*s (KNN), *Random Forests*, Redes Neurais Artificiais e Regressão Linear Simples e Múltipla. Na etapa "Avaliação", efetuou-se a análise dos resultados, utilizando-se medidas de acurácia dos modelos de predição gerados pelos algoritmos.

Figura 1 - Etapas para realização do trabalho



Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth, 1996.

3.2.1 Caracterização das bases de dados

Os dados utilizados dividem-se em duas categorias: informações de geração de energia elétrica e informações de consumo de energia elétrica, ambas em Kilowatt-hora (KWh). Os dados de geração de energia encontram-se no portal do *campus*, situando-se entre dezembro de 2018 até os dias atuais. Eles possibilitam a análise de várias granularidades diferentes de tempo, contendo informações de consultas diárias, mensais e anuais. Já os dados de consumo de energia dispõem-se do ano de 2013 a 2020, retirados de contas de energia elétrica, de acordo com sua disponibilidade, estando disponíveis apenas como valores mensais. Ainda sobre os dados de consumo, observa-se a distinção de dois perfis - o consumo dentro do horário de ponta (HP) e fora do horário de ponta (HFP), que trazem diferenças quanto ao preço da tarifa e comportamento de consumo de energia elétrica. A partir da análise destas duas dimensões do consumo, identificam-se mais oportunidades de análises sobre os dados preditos.

Dados meteorológicos foram introduzidos no estudo como uma tentativa de enriquecer a análise proposta, em sintonia com o estado da arte. As informações são dadas em milímetros, graus celsius, milibar (mB), porcentagem e metros por segundo, advindas de uma estação automática instalada no *campus* do IFMG de Bambuí, disponibilizadas através do Banco de Dados Meteorológicos[6] do Instituto Nacional de Meteorologia[7]. A estação denominada A565 disponibiliza informações como o número de dias com precipitação, precipitação total, umidade do ar, pressão atmosférica, ponto de orvalho, temperatura média, mínima e máxima e velocidade do vento.

Para determinação de quais atributos utilizar, foram consultados os estudos de Terra (2003) e Jawad *et al*. (2020), que evidenciam a eficácia do uso das seguintes informações: número de dias com precipitação, precipitação total, ponto de orvalho, umidade relativa do ar, temperatura máxima, média e mínima e velocidade média do vento. Distingue-se que o número de dias com precipitação é disponibilizado apenas para granularidade de tempo superior a mensal, assim como umidade do ar e ponto de orvalho são viabilizados apenas nas bases de granularidade diária.

Visando a uma boa organização, as bases de dados utilizadas no trabalho são listadas a seguir:

A: Consumo de energia em HFP com medições mês a mês do período de 2019 a 2020;
B: Base A com adição dos atributos meteorológicos;
C: Consumo de energia em HP com medições mês a mês do período de 2019 a 2020;
D: Base C com adição dos atributos meteorológicos;
E: Geração de energia com medições mês a mês do período de 2019 a 2020;
F: Base E com adição dos atributos meteorológicos;
G: Geração de energia com medições dia a dia do mês de fevereiro de 2020;
H: Base G com adição dos atributos meteorológicos;
I: Geração de energia com medições dia a dia do mês de julho de 2020;
J: Base I com adição dos atributos meteorológicos.

Os experimentos realizados buscaram a comparação dos comportamentos dos algoritmos sobre os dados de geração e consumo de energia elétrica. A escolha se deu pela disponibilidade, de forma a encontrar dificuldades para obtenção de um período de tempo significativo com dados ininterruptos.

[6] https://bdmep.inmet.gov.br
[7] https://portal.inmet.gov.br

A seguir, a Tabela 1 retrata como estão organizados os dados, representando a base de consumo de energia elétrica mês a mês. Cada instância das bases representa a medição de um mês e tem os valores de atributo “Consumo”, para as bases de consumo de energia, e “Geração”, para as bases de geração de energia, ambos em KWh. O mesmo formato é estendido para bases semelhantes ao longo de todo o trabalho, variando apenas a granularidade entre mensal e diária.

Tabela 1 - Representação dos dados

Data	**Consumo (KWh)**
01-2019	70080
02-2019	58320
03-2019	68400
...	...
...	...
07-2020	42720
08-2020	43920
09-2020	44880

Fonte: Elaborado pelo autor, 2021.

De maneira análoga, a Tabela 2 exibe a organização dos dados, ilustrando uma base de consumo de energia elétrica com a adição dos atributos meteorológicos. A mesma configuração se apresenta para bases semelhantes ao longo de todo o trabalho, variando para bases de Geração de energia e também para a disponibilidade dos atributos: dias com precipitação e Temperatura Média (ºC), para a granularidade mensal, e Ponto de Orvalho (ºC), Temperaturas Mínimas, Máximas (ºC) e a Umidade Relativa do Ar (%), para a granularidade diária.

Tabela 2 - Representação dos dados complementada de atributos meteorológicos

Data	Consumo (KWh)	Dias com precipitação	Precipitação total (mm)	Temperatura média (ºC)	Velocidade máxima do vento (m/s)
01-2019	70080	13	115	24,42	7,1
02-2019	58329	16	304,6	23,56	7,3
03-2019	68400	17	155,4	23,02	10,1
...	...	...	...	...	...
...	...	...	...	...	...
07-2020	42720	0	0	17,8	6,1
08-2020	43920	1	8,8	17,65	5,4
09-2020	44880	2	13	21,48	5,2

Fonte: Elaborada pelo autor, 2021.

3.2.2 Técnicas de predição

Como evidenciado previamente no Capítulo 2, os algoritmos que se mostram promissores dentro do estado da arte e que foram empregados no trabalho são o KNN, Redes Neurais Artificiais do tipo *Perceptron* de Multicamadas, Regressão Linear Simples e Múltipla e *Random Forests*. Para o algoritmo KNN, conforme explorado por D. W. (1992), categoriza-se que os valores mais adequados para o parâmetro k são: 1, 3, 5, 7 e 10. Já na aplicação do *Random Forests*, mostra-se necessária a regulagem de dois parâmetros: *ntree*, definindo o número de árvores da floresta, e o *mtry*, que regula a quantidade de atributos avaliados para construção das árvores. No trabalho de Breiman (2001), sugerem-se os valores para o *ntree* de 500 e 1000; e, para o *mtry*, recomendam-se as seguintes configurações: 1/6 da quantidade de atributos preditores, 2/3 da quantidade de atributos preditores e 1/3 da quantidade de atributos preditores.

3.2.3 Avaliação

A avaliação da precisão preditiva dos modelos de regressão foi aferida pela técnica de validação por divisão dos dados em conjuntos de treinos e testes. Como abordado por Santos

et al. (2019), este tipo de validação expõe se o modelo avaliado obtém boa performance tanto nos treinamentos quanto em testes, visando aprimorar sua capacidade na predição de valores em situações mais generalizadas. Os autores apontam que, dentre as proporções mais utilizadas para a divisão dos dados, estão 60:40, 70:30 e 80:20, de maneira que o primeiro valor representa a porcentagem de divisão dos dados para a criação do conjunto de treino, e o segundo, para o conjunto de testes. A medição da acurácia preditiva é necessária para se verificar quanto os modelos se ajustam aos dados. No presente estudo, a proporção de 70:30 foi empregada por ser a indicada pela ferramenta utilizada como divisão padrão, por meio da opção *Evaluate on held out training*, conforme a documentação (PENTAHO, 2014).

Sendo evidenciadas pelo seu amplo uso na literatura, as medidas de acurácia MAPE (DE MYTTENAERE *et al*., 2016), MAE (TAREEN *et al*., 2019) e RMSE (RAFIQUE *et al*., 2020) foram utilizadas. De Myttenaere *et al.* (2016) explicam que o MAPE é amplamente empregado por apresentar uma interpretação muito intuitiva do erro, possuindo ampla relevância nos campos de finanças e na definição de preços para produtos, sendo que, quanto menor o valor indicado, menor o erro apresentado pelo modelo. Os autores ainda fomentam o seu uso como medida de validação para predição do consumo de energia elétrica - ponto reforçado ao se analisar a quantidade de trabalhos científicos apresentados no estado da arte que empregam a métrica. O MAE, como explorado por Tareen *et al*. (2019), avalia a extensão média dos erros sobre as predições, ignorando sua direção e atribuindo peso igual para as diferenças entre os valores reais e preditos, sendo expresso na grandeza do valor predito. Os autores indicam que, quanto menor o valor obtido de MAE, mais preciso é o modelo. Já em Rafique *et al*. (2020, p. 4, tradução nossa), é evidenciado que "O RMSE mede a discrepância entre os valores de concentração de radon medidos e preditos. Valores menores de RMSE indicam incongruências menores"[8]. Além disso, de maneira semelhante ao MAE, o RMSE é medido na grandeza da variável predita, ou seja, em um estudo de demanda de energia elétrica expressa em KWh. Cada medida tem suas particularidades e se destaca em um aspecto diferente, sendo difícil eleger se uma pode ser superior à outra. Em um trabalho comparativo entre medidas, Chai e Draxler (2014) destacam que o ideal é realizar uma combinação entre métricas diferentes. As medidas empregadas são definidas como:

[8] No original: The RMSE measures the discrepancy between measured and predicted values of radon concentration. Smaller RMSE values indicate lower incongruities.

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^{N} \frac{|\bar{y}i - yi|}{yi} \quad (1)$$

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\bar{y}i - yi| \quad (2)$$

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\bar{y}i - yi)^2} \quad (3)$$

Sendo que N representa a quantidade de registros de teste, e $\bar{y}i$ e yi são, respectivamente, os valores preditos e reais.

3.3 Ferramentas

Nesta seção, são definidas as ferramentas, de maneira detalhada, a fim de permitir a sua reprodutibilidade por outros pesquisadores/autores.

3.3.1 Weka

A ferramenta *Waikato Environment for Knowledge Analysis* (WEKA), versão 3.8.5 (WEKA, 2016), para Mineração de Dados foi empregada neste trabalho. A escolha da aplicação se deu por sua utilização em disciplinas da graduação e trabalhos científicos, além de ser um *software* livre. Outros fatores são a simplicidade de se trabalhar com a ferramenta, a sua grande gama de algoritmos suportados e também por se tratar de uma aplicação robusta para análise de informações, contendo diversas ferramentas para filtragem, edição e visualização dos dados. Além do mais, o WEKA possui fácil integração com as atuais e mais populares ferramentas de ciência dos dados, como as linguagens R e Python (WEKA, 2016).

Para a análise de séries temporais, objeto de estudo principal deste trabalho, utilizou-se a extensão disponibilizada pela Pentaho (2014), denominada *Forecast*, contando com um extenso catálogo de algoritmos, ferramentas de validação e exploração dos dados e atributos.

3.3.2 Configuração da máquina

A máquina utilizada durante o trabalho é de posse do discente, possuindo as especificações: processador Intel Core I5 - 9400f 2.90 *Gigahertz*, 16 *Gigabytes* DDR4 de *Random Access Memory* (RAM), com *clock* de 2666 *Megahertz*, placa de vídeo NVIDIA General Electronic Facsimile Optimized for Repair and Ceaseless Exploration (GeForce) Giga Texel Shader eXtreme (GTX) 1060 de 3 *Gigabytes* e sistema operacional de 64 *bits* Windows 10.

3.4 Metodologia de desenvolvimento do trabalho

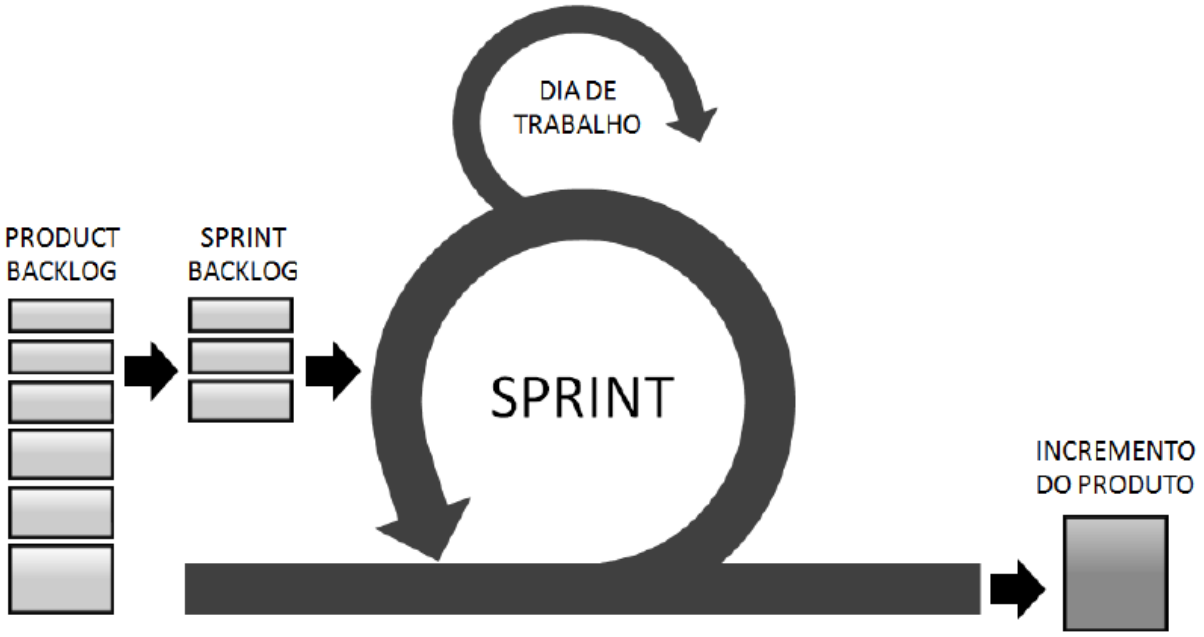
Para o desenvolvimento do trabalho, foi empregada a metodologia Scrum, de desenvolvimento ágil. Sabbagh (2014) aponta que, embora a metodologia escolhida tenha sido originalmente designada para o desenvolvimento de *software,* seu uso encontra poucos limites; desde projetos multimilionários até projetos mais simples explanaram a eficácia de suas aplicações. Além disso, o autor salienta que o Scrum reduz os riscos de insucesso, possibilita entregas de maneira rápida - como qualquer metodologia ágil - adapta-se bem a mudanças que podem surgir no decorrer do trabalho, além de aumentar a qualidade do produto final e impulsionar a produtividade. Em suma, o Scrum se adequa com inúmeros pontos positivos para o desenvolvimento de qualquer pesquisa. Para melhor visualização das etapas metodológicas, a Figura 2 apresenta o ciclo resumido do Scrum.

Alguns elementos principais do método utilizados foram os *Sprints* e o *Product Backlog*. *Sprints* se resumem a jornadas para a realização de trabalho e entrega de tarefas, e o *Product Backlog* pode ser entendido como os itens e tarefas a serem efetuados em determinados estágios do trabalho. O *Product Backlog* deste trabalho é apresentado na Tabela 3, a seguir, contendo as colunas - *Identity* (ID): identificando cada atividade do trabalho; Nome: denominando cada tarefa; Importância: atribuindo valores de 0 a 10 para a importância da atividade; Estimativa: apresentando a quantidade de horas estimadas para o desempenho das atividades; Como demonstrar: exibindo ações a serem executadas para realizar a tarefa.

A metodologia se deu pela definição e execução de atividades semanais dispostas em *sprints,* sempre com uma reunião entre os orientadores e orientando no início da semana para a discussão das entregas do *sprint* anterior e definição das etapas presentes no próximo *sprint*.

Os itens que compõem o *Product Backlog* são as etapas do trabalho, como a apresentação da proposta, revisão bibliográfica, estudo das técnicas de mineração, entre outras. Exemplificando, apresenta-se o primeiro *Sprint* do trabalho na Tabela 4, que contemplou a definição do tema da pesquisa na apresentação da proposta e o levantamento bibliográfico. As demais *Sprints* são apresentadas no APÊNDICE A.

Figura 2 - O ciclo do Scrum



Fonte: Adaptado de Sabbagh (2014).

Tabela 3 - *Product Backlog*

ID	Nome	Importância (0 - 10)	Estimativa (horas)	Como demonstrar
1	Apresentação da Proposta	10	32	Escrever e apresentar a proposta inicial do trabalho.
2	Revisão Bibliográfica	10	40	Levantar bibliografias com o tema alinhado ao trabalho.
3	Estudo de técnicas de Mineração	10	24	Definir técnicas a serem utilizadas.
4	Estudo da Ferramenta WEKA	8	22	Definir métricas, parâmetros e ferramentas da aplicação a serem utilizadas.
5	Coleta dos dados	10	12	Construir as bases de dados para análise.
6	Análise e Preparação dos dados	10	15	Caracterizar a base de dados.
7	Aplicação dos Algoritmos	9	20	Aplicar os algoritmos para gerar os modelos preditivos.
8	Análise dos Resultados	10	30	Aferir a acurácia preditiva dos modelos gerados.
9	Escrita da Monografia	10	40	Desenvolver a monografia a ser entregue ao fim do trabalho.
10	Defesa	10	34	Apresentar à banca de avaliadores todo o trabalho monográfico desenvolvido.

Fonte: Elaborada pelo autor, 2020.

Tabela 4 - *Sprint* 1

ID	Tarefas	Horas
1	Apresentação da Proposta	32
2	Revisão Bibliográfica	20

Fonte: Elaborada pelo autor, 2020.

4 RESULTADOS E DISCUSSÕES

O presente capítulo apresenta os experimentos relacionados ao estudo dos modelos preditivos gerados utilizando técnicas de regressão. Este foi desenvolvido sobre as bases de dados contendo dois escopos: dados de consumo e dados de geração de energia elétrica. Na Seção 4.1, é evidenciada uma análise sobre intervalos semelhantes de tempo, a fim de comparar os resultados para os dados de consumo, geração de energia e informações meteorológicas. Posteriormente, na Seção 4.2, efetua-se um experimento comparativo sobre os dados de geração de energia e meteorológicos, analisando o comportamento dos algoritmos em diferentes situações. Já nas seções posteriores, 4.3 e 4.4, os resultados obtidos são comparados e discutidos

4.1 Análise mensal

O experimento realizado sobre as bases A, B, C, D, E e F utilizou informações contidas no intervalo de 2019 a 2020 e foi escolhido levando-se em consideração que os dados apresentam-se de maneira consistente, sem ruídos ou faltando informações. Tais características inerentes a esta base também estarão presentes nos dados das demais seções.

As informações foram dispostas com os meses como instâncias, e a medida, em KWh, como atributo para formar as séries iniciais. Foram empregadas as medidas de consumo e geração de energia, tanto como atributo preditor quanto como atributo-classe para as bases de dados sem informações meteorológicas. Já para a análise posterior, contendo atributos meteorológicos, foram empregados o número de dias com precipitação, a precipitação total (mm), a temperatura média (ºC) e a velocidade do vento (m/s) como atributos preditores, e o consumo e geração de energia, como atributos-classe.

As tabelas a seguir dispõem as métricas de avaliação (em linhas) para cada técnica utilizada (em colunas). Em todas as tabelas desta e das próximas seções, os melhores resultados foram destacados em negrito, e, nas tabelas posteriores, sobre resultados, seguiu-se a mesma configuração. O conjunto das Tabelas 5 a 8 refere-se aos resultados das bases de dados A, B, C e D, respectivamente, tratando-se de informações sobre consumo de energia elétrica.

Tabela 5 - Resultados Base de Dados A

Métricas	**Regressão Linear**	**RNA**	**KNN**		***Random Forest***	
			k = 1	k = 10	*ntree* = 500	*ntree* = 1000
MAE (KWh)	5725,5	**5009,26**	13880,00	20168,00	13517,82	13490,47
MAPE	12,08%	**9,0%**	31,60%	45,30%	30,20%	30,12%
RMSE (KWh)	**6135,24**	7234,81	15961,65	21478,93	14190,32	14139,91

Fonte: Elaborada pelo autor, 2020.

Ao serem analisados os resultados para os dados de horário fora de ponta (Tabela 5), evidenciou-se que o modelo que melhor se ajustou aos dados foi o obtido pela técnica RNA, apresentando MAE = 5009,26 KWh, MAPE = 9,50% e RMSE = 7234,81 KWh, seguida pela técnica Regressão Linear, com MAE = 5725,5 KWh, MAPE = 12,08% e RMSE = 6135,24 KWh. Ressalta-se a distinção destes resultados comparados aos demais modelos, que obtiveram erros percentuais três ou quatro vezes maiores. Na distinção dos resultados em algoritmos com variados parâmetros, no KNN, o melhor se mostrou para o parâmetro k igual a 1, com amplas diferenças quando comparado ao modelo gerado por k igual a 10. Já na análise do *Random Forest,* o melhor resultado foi evidenciado para o parâmetro *ntree* igual a 1000, apresentando baixa distinção para o mesmo parâmetro ajustado a 500, embora ambos não representem um bom desempenho.

Tabela 6 - Resultados Base de Dados B

Métricas	**Regressão Linear**	**RNA**	**KNN**		***Random Forest***	
			k = 1	k = 10	*ntree* = 500	*ntree* = 1000
MAE (KWh)	20828,06	**12000,7**	13720	20528	17206,99	17254,52
MAPE	46,84%	**24,85%**	31,17%	45,9%	38,89%	39%
RMSE (KWh)	23033,44	**13892,65**	16776,84	21606,79	18663,33	18719,36

Fonte: Elaborada pelo autor, 2021.

Com os resultados para os dados de horário fora de ponta com atributos meteorológicos (Tabela 6), evidencia-se que o modelo que melhor se ajustou aos dados foi o obtido pela técnica RNA, apresentando MAE = 12000,7 KWh, MAPE = 24,85% e RMSE =

13892,65 KWh, seguida pela técnica KNN, com k ajustado a 1, obtendo MAE = 13720 KWh, MAPE = 31,17% e RMSE = 16776,84 KWh. Distinguindo-se os resultados em algoritmos com variados parâmetros, o KNN, com o parâmetro k = 10, evidenciou consideráveis diferenças, diferentemente da análise do *Random Forest,* em que o melhor resultado se mostrou para o parâmetro *ntree* igual a 500, mas apresentou baixa distinção para o mesmo parâmetro ajustado a 1000. Ressalta-se que todos os modelos gerados apontaram valores de erros elevados.

Tabela 7 - Resultados Base de Dados C

Métricas	**Regressão Linear**	**RNA**	**KNN**		***Random Forest***	
			k = 1	k = 10	*ntree* = 500	*ntree* = 1000
MAE (KWh)	1303,52	**1047,71**	1280,00	2556,00	1738,55	1752,32
MAPE	29,68%	**21,05%**	28,53%	57,18%	39,34%	39,67%
RMSE (KWh)	1483,21	**1163,59**	1446,65	2669,20	1885,51	1902,56

Fonte: Elaborado pelo autor, 2020.

Para a análise com os dados de consumo em horário de ponta (Tabela 7), destaca-se que os modelos que melhor se ajustaram aos dados foram os obtidos pelo RNA com MAE = 1047,71 KWh, MAPE = 21,05% e RMSE = 1163,59 KWh e KNN com k ajustado a 1, obtendo MAE = 1280 KWh, MAPE = 28,53% e RMSE = 1446,65 KWh. Explorando o KNN com k ajustado a 10, obteve-se um erro percentual superior a duas vezes o apresentado pelo modelo com k igual a 1, de maneira a não evidenciar um bom desempenho. Experimentando a variação do atributo *ntree* para o *Random Forest*, quando ajustado a 500, obtiveram-se resultados ligeiramente melhores e com baixa distinção para os alcançados para *ntree* igual a 1000.

Tabela 8 - Resultados Base de Dados D

Métricas	**Regressão Linear**	**RNA**	**KNN**		***Random Forest***	
			k = 1	k = 10	*ntree* = 500	*ntree* = 1000
MAE (KWh)	**837,21**	2131,85	1880	2568	1895,69	1908,18

MAPE	**16,46%**	48,58%	42,80%	57,54%	42,91%	43,20%
RMSE (KWh)	**1091,90**	2475,75	2162,22	2687,71	2043,48	2059,74

Fonte: Elaborada pelo autor, 2021.

Analisando os dados de consumo em horário de ponta com atributos meteorológicos (Tabela 8), destaca-se que o modelo que melhor se ajustou aos dados foi o obtido pela Regressão Linear com MAE = 837,21 KWh, MAPE = 16,46% e RMSE = 1091,90 KWh, sendo o único a apresentar baixos erros, pois todos os outros exibiram erros percentuais duas ou até três vezes superiores. No KNN, o melhor modelo foi encontrado com k ajustado a 1, ficando um pouco distante para a configuração com k igual a 10. Experimentando-se a variação do atributo *ntree* para o *Random Forest*, quando ajustado a 500, obtiveram-se resultados ligeiramente melhores e com baixa distinção para os alcançados para *ntree* igual a 1000.

Na sequência, o conjunto das Tabelas 9 e 10 refere-se aos resultados das bases de dados E e F, respectivamente, tratando-se de informações sobre geração de energia elétrica.

Tabela 9 - Resultados Base de Dados E

Métricas	**Regressão Linear**	**RNA**	**KNN**		***Random Forest***	
			k = 1	k = 10	*ntree* = 500	*ntree* = 1000
MAE (KWh)	485,89	347,62	308,04	322,71	250,34	**246,37**
MAPE	20,5%	14,8%	13,07%	13,63%	10,34%	**10,26%**
RMSE (KWh)	535,99	525,85	362,05	372,68	314,46	**304,67**

Fonte: Elaborada pelo autor, 2020.

Para os dados de geração de energia elétrica (Tabela 9), observou-se que os modelos obtidos por ambas as configurações do algoritmo *Random Forest*, com MAE = 246,37 KWh, MAPE = 10,26% e RMSE = 304,67 KWh, para *ntree* = 1000 e MAE = 250,34 KWh, MAPE = 10,34% e RMSE = 314,46 KWh, para ntree = 500, evidenciaram os melhores resultados. Ainda para a técnica do *Random Forest,* observa-se baixa distinção entre os resultados diante da variação do parâmetro *ntree*. Analisando-se os comportamentos dos modelos gerados pelo algoritmo KNN, com k = 1, obtiveram-se resultados melhores e muito próximos dos de k = 10. Para esta base de dados, ressalta-se que todos os modelos exibiram bons resultados, com o

maior valor de MAPE obtido pelo algoritmo Regressão Linear e igual a 20,5%.

Tabela 10 - Resultados Base de Dados F

Métricas	**Regressão Linear**	**RNA**	**KNN**		***Random Forest***	
			k = 1	k = 10	*ntree* = 500	*ntree* = 1000
MAE (KWh)	260,43	580,21	259,75	289,56	213,45	**212,10**
MAPE	11,25%	23,97%	11,34%	12,32%	9,48%	**9,43%**
RMSE (KWh)	392,29	812,79	300,57	335,10	258,65	**257,52**

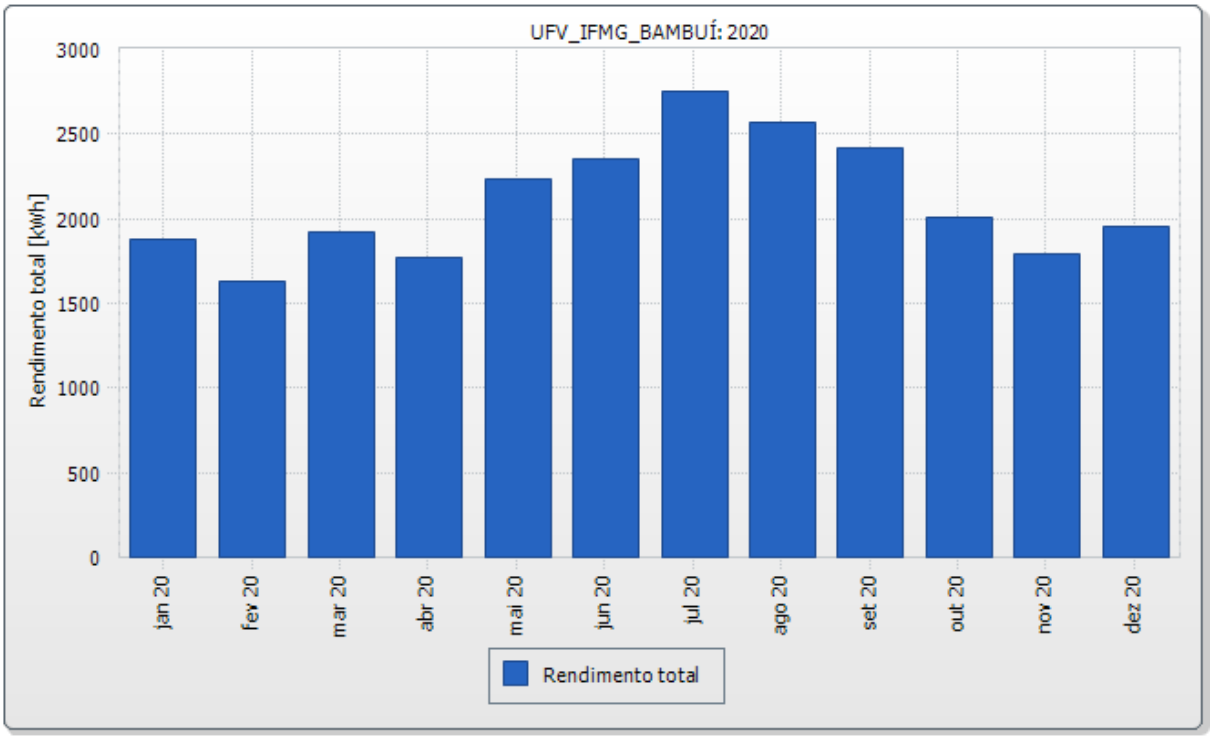
Fonte: Elaborada pelo autor, 2021.

Observa-se, pela Tabela 10, que os melhores modelos para a base de geração de energia com adição dos atributos meteorológicos foram alcançados pelas duas configurações do algoritmo de *Random Forest,* com o *ntree* igual a 1000, obtendo MAE = 212,10 KWh, MAPE = 9,43% e RMSE = 257,52 KWh, e com *ntree* igual a 500, com MAE = 213,45 KWh, MAPE = 9,48% e RMSE = 258,65 KWh, embora ambos os resultados tenham sido muito próximos. Avaliando-se os comportamentos dos modelos gerados pelo algoritmo KNN, com k = 1, obtiveram-se resultados melhores, porém próximos dos de k = 10. Para esta base de dados, destaca-se que todos os modelos apresentaram bons resultados, com o maior valor de MAPE obtido pelo algoritmo RNA e igual a 23,97%.

4.2 Análise diária

Nesta seção, o ponto principal foi buscar uma disposição diferente dos dados, variando a granularidade analisada. A fonte das informações de consumo possibilita sua visualização apenas mês a mês, diferentemente das bases de geração, que fornecem informações diárias, mensais e anuais. Portanto, os dados abordados foram os de geração de energia, e a análise se deu comparando leituras diárias do mês de maior e de menor geração em 2020. No Gráfico 1, pode-se visualizar a representação de toda a geração de energia elétrica mensalmente disposta em 2020, mostrando fevereiro como o mês de menor geração, e julho, como o de maior.

Gráfico 1 - Geração de energia mensal do ano de 2020



Fonte: SMA Solar Technology AG (2021).

Os experimentos foram realizados sobre as bases G, H, I e J, sendo G e H referentes a fevereiro, e I e J, a julho de 2020. Dispuseram-se as informações com os dias como instâncias, e a medida, em KWh, tanto como atributo preditor quanto como classe, para as bases de dados sem informações meteorológicas (Bases G e I). Já para a análise posterior, contendo atributos meteorológicos (Bases H e J), foram empregados a precipitação total (mm), o ponto de orvalho (ºC), as temperaturas máxima e mínima (ºC), a umidade relativa do ar (em porcentagem) e a velocidade do vento (m/s) como atributos preditores, e a geração de energia como atributo-classe. As tabelas a seguir dispõem as métricas de avaliação (em linhas) para cada técnica utilizada (em colunas).

Tabela 11 - Resultados Base de Dados G

Métricas	**Regressão Linear**	**RNA**	**KNN**		***Random Forest***	
			k = 1	k = 10	*ntree* = 500	*ntree* = 1000
MAE (KWh)	14,8	37,42	31,31	**11,57**	36,07	36,18
MAPE	31,76%	77,84%	60,47%	**24,95%**	76,02%	76,20%
RMSE (KWh)	17,67	40,66	44,64	**12,45**	36,71	36,77

Fonte: Elaborada pelo autor, 2021.

Analisando-se os dados durante o mês de fevereiro de 2020, que apresentou menor geração de energia elétrica (Tabela 11), destaca-se que o modelo gerado que melhor se ajustou a eles foi obtido através da técnica KNN, com o parâmetro k ajustado a 10, obtendo MAE = 11,57 KWh, MAPE = 24,95% e RMSE = 12,45 KWh. Explorando a técnica do KNN com o parâmetro k ajustado a 1, obteve-se um erro percentual superior a duas vezes o exibido pelo modelo com k igual a 10, de maneira a não apresentar um bom desempenho. Experimentando a variação do atributo *ntree* para o *Random Forest*, ambos os valores do atributo (500 e 1000) evidenciaram modelos de baixa confiabilidade, obtendo-se erros percentuais superiores a 75%.

Tabela 12 - Resultados Base de Dados H

Métricas	**Regressão Linear**	**RNA**	**KNN**		***Random Forest***	
			k = 1	k = 10	*ntree* = 500	*ntree* = 1000
MAE (KWh)	**10,9**	15,56	11,37	11,29	15,06	15,23
MAPE	23,64%	31%	**23,5%**	24,44%	32,36%	32,80%
RMSE (KWh)	**11,95**	18,80	12,54	12,30	16,03	16,26

Fonte: Elaborado pelo autor, 2021.

Avaliando-se os dados para o mês de fevereiro de 2020, que retratou menor geração de energia elétrica e passou pela adição dos atributos meteorológicos (Tabela 12), destaca-se que os modelos de predição que melhor se ajustaram aos dados foram os obtidos por meio da técnica de Regressão Linear, apresentando MAE = 10,9 KWh, MAPE = 23,64% e RMSE = 11,95 KWh e através do algoritmo KNN com o parâmetro k ajustado a 1, obtendo MAE = 11,37 KWh, MAPE = 23,5% e RMSE = 12,54 KWh. Explorando o algoritmo do KNN com o parâmetro k ajustado a 10, alcançaram-se resultados muito próximos ao do modelo com o parâmetro k igual a 1. Experimentando-se a variação do parâmetro *ntree* para o algoritmo do *Random Forest*, para os dois valores do parâmetro (500 e 1000), os modelos apresentaram resultados próximos e mostraram baixa confiabilidade, obtendo erros percentuais superiores a 30%.

Tabela 13 - Resultados Base de Dados I

Métricas	**Regressão Linear**	**RNA**	**KNN**		***Random Forest***	
			k = 1	k = 10	*ntree* = 500	*ntree* = 1000
MAE (KWh)	**9,56**	18,84	11,54	11,84	12,87	12,94
MAPE	**10,68%**	21,84%	12,43%	12,39%	13,30%	13,38%
RMSE (KWh)	**11,09**	22,44	13,77	13,14	14,92	14,98

Fonte: Elaborada pelo autor, 2021.

Para os resultados sobre o mês de julho de 2020, de maior geração de energia elétrica (Tabela 13), evidenciou-se que o modelo que melhor se ajustou aos dados foi o obtido pela técnica da Regressão Linear, apresentando MAE = 9,56 KWh, MAPE = 10,68% e RMSE = 11,09 KWh, seguida pelo algoritmo KNN com o parâmetro k ajustado a 10, evidenciando MAE = 11,84 KWh, MAPE = 12,39% e RMSE = 13,14 KWh. Para os resultados em algoritmos com variados parâmetros, no KNN, o modelo gerado por k igual a 1 obteve diferenças mínimas quando comparado ao modelo gerado por k igual a 10. De maneira semelhante, na análise do *Random Forest*, o melhor resultado se mostrou para o parâmetro *ntree* = 500, apresentando baixa distinção para o mesmo parâmetro ajustado a 1000, sendo que ambos representam bom desempenho.

Tabela 14 - Resultados Base de Dados J

Métricas	**Regressão Linear**	**RNA**	**KNN**			***Random Forest***	
			k = 1	k = 5	k = 10	*ntree* = 500	*ntree* = 1000
MAE (KWh)	9,87	10,86	12,02	**9,03**	10,89	11,84	12,16
MAPE	11,14%	12,02%	13,74%	**9,74%**	11,51%	12,23%	12,56%
RMSE (KWh)	14,1	15,32	15,77	**10,21**	12,09	13,48	13,79

Fonte: Elaborada pelo autor, 2021.

No que diz respeito aos resultados dos dados do mês de julho de 2020, de maior geração de energia elétrica com atributos meteorológicos (Tabela 14), evidenciou-se que não foi suficiente a apresentação de apenas duas configurações para o atributo k no algoritmo KNN (1 e 10 apresentados anteriormente), já que o modelo que melhor se ajustou aos dados

foi o obtido por esta técnica com o k ajustado a 5, apresentando MAE = 9,03 KWh, MAPE = 9,74% e RMSE = 10,21 KWh. Na distinção dos resultados em algoritmos com variados parâmetros, para o KNN, os modelos gerados por k igual a 1 e 10 atingiram valores maiores, porém com diferenças pequenas em comparação ao modelo gerado por k igual a 5. De maneira semelhante, na análise do *Random Forest*, o melhor resultado se mostrou para o parâmetro *ntree* igual a 500, apresentando baixa distinção para o mesmo parâmetro ajustado a 1000. Destaca-se que todos os modelos gerados para esta base de dados obtiveram bons resultados, com o maior erro igual a 13,74%.

4.3 Comparação dos resultados

Para a comparação dos resultados obtidos, a análise se dividiu em 3 segmentos. O primeiro busca um paralelo entre os comportamentos das técnicas empregadas nas duas naturezas de dados utilizados - geração e consumo de energia elétrica (Seção 4.3.1). Já o segundo segmento enfatiza a comparação entre os resultados antes e depois da adição dos dados meteorológicos aos experimentos (Seção 4.3.2). Na sequência, objetiva-se analisar o comportamento dos algoritmos entre o mês de menor e de maior geração de energia elétrica (Seção 4.3.3). Por fim, explora-se qual algoritmo obteve a melhor performance (Seção 4.3.4).

4.3.1 Consumo e geração de energia elétrica

O primeiro ponto analisado foi entre o perfil do consumo de energia elétrica para as informações dentro e fora do horário de ponta. Comparando-se as Tabelas 5 e 7, destaca-se que o melhor resultado foi obtido com o RNA sobre os dados de HFP com MAE = 5009,26 KWh, MAPE = 9,5% e RMSE = 7234,81 KWh. Já o pior foi alcançado pelo KNN com k ajustado a 10 sobre os dados de HP, apresentando MAE = 2556 KWh, MAPE = 57,18% e RMSE = 2669,2 KWh. As tabelas evidenciam resultados ligeiramente melhores para os modelos gerados a partir das informações em HFP.

Adicionando-se a Tabela 9 (geração de energia) à comparação, observa-se que, apesar do melhor resultado manter-se no RNA sobre os dados de HFP, diferentemente do conjunto dos modelos gerados a partir de informações de consumo de energia elétrica, todos os modelos apresentaram baixos erros, e o destaque vai para a técnica *Random Forest*, por apresentar o modelo com maior precisão.

4.3.2 Presença de dados meteorológicos

Analisando os primeiros experimentos, as Tabelas 5 e 6 evidenciam que, para os dados de consumo de energia elétrica em HFP, a adição dos atributos meteorológicos, majoritariamente, piorou o desempenho dos algoritmos. O modelo que menos errou foi o obtido antes da adição de atributos (Tabela 5) pelo RNA, obtendo MAE = 5009,26 KWh, MAPE = 9,5% e RMSE = 7234,81 KWh, e o que mais errou, depois (Tabela 6), obtendo MAE = 20528 KWh, MAPE = 45,9% e RMSE = 21606,79 KWh. Para ambas as tabelas, o melhor resultado foi o alcançado pela técnica RNA, e o pior, com o KNN com k igual a 10.

Para as informações sobre HP, as Tabelas 7 e 8 apontam que a base com adição dos dados meteorológicos atingiu o melhor resultado, e o modelo foi obtido com a Regressão Linear, apresentando MAE = 837,21 KWh, MAPE = 16,46% e RMSE = 1091,9 KWh. Os piores resultados foram alcançados em ambas as bases pelo KNN com k igual a 10, expressando valores bem próximos, com MAE = 2556 KWh, MAPE = 57,18% e RMSE = 2669,2 KWh, para a base original, e MAE = 2568 KWh, MAPE = 57,54% e RMSE = 2687,71 KWh com a adição dos dados meteorológicos. Excluindo-se os resultados de destaque, observa-se que, em ambas bases, os modelos obtidos evidenciam baixa confiabilidade em virtude de porcentagens de erro acima de 28%.

Avaliando-se as Tabelas 9 e 10, destaca-se, em ambas, o desempenho do algoritmo *Random Forest* com *ntree* ajustado a 1000, atingindo o melhor resultado para a base com adição de informações meteorológicas e MAE = 246,37 KWh, MAPE = 10,26% e RMSE = 304,67 KWh. O pior resultado também é apresentado para esta base de dados, alcançada por meio do RNA obtendo MAE = 580,21 KWh, MAPE = 23,97% e RMSE = 812,79 KWh. Ressalta-se que, nas duas bases de dados, todos os resultados foram mais baixos se comparados aos dos outros experimentos.

4.3.3 Geração: análise diária

Para o último estudo, as Tabelas 11 e 13 deixam claro que o mês de fevereiro de 2020, de menor geração de energia elétrica, obteve, majoritariamente, mais incertezas nos modelos gerados. Em contraste, o mês de julho de 2020, de maior geração de energia elétrica, evidenciou, em todos os modelos gerados, baixos valores para a porcentagem de erro. O

modelo de maior erro foi o alcançado pela técnica RNA sobre as informações do mês de fevereiro, com MAE = 37,42 KWh, MAPE = 77,84% e RMSE = 40,66 KWh, e o de menor erro, o obtido pela técnica Regressão Linear sobre os dados do mês de julho, com MAE = 9,56 KWh, MAPE = 10,68% e RMSE = 11,09 KWh. Nota-se que, em ambas as tabelas, os piores resultados individuais foram conquistados pelo algoritmo RNA.

4.3.4 Performance dos algoritmos

Para eleger os algoritmos que obtiveram os melhores desempenhos neste trabalho, a Tabela 15, a seguir, contabiliza, dentre as técnicas empregadas, a quantidade de vezes em que cada uma apresentou o modelo melhor pontuado em ao menos duas métricas de avaliação dentre as aplicações nas 10 bases de dados (bases A, B, C, D, E, F, G, H, I e J). Observou-se que as melhores performances foram, da melhor para a pior, obtidas na sequência: RNA, Regressão Linear, KNN e Random Forest.

Tabela 15 - Performance dos Algoritmos

Algoritmos	Quantidade de Destaques
RNA	3
Regressão Linear	3
KNN	2
Random Forest	2

Fonte: Elaborada pelo autor, 2021.

4.4 Discussão dos resultados

Analisando os comportamentos dos algoritmos para os perfis de consumo de energia elétrica (HFP e HP), aponta-se destaque para os resultados do Horário Fora de Ponta. Este pode ser melhor compreendido quando se explora a natureza de seus perfis. Como indicado por Masseroni e Oliveira (2012), o Horário de Ponta é o momento em que ocorre o pico de consumo de energia elétrica com a maioria de seus usuários ativos, gerando, assim, uma situação de medições atípicas de comportamento caótico, dificultando a construção dos modelos preditivos. Assim, corroboram-se os melhores resultados obtidos sob as informações

em Horário Fora de Ponta, apresentando uma média de porcentagem de erro (MAPE médio) igual a 26,47%, enquanto, para Horário de Ponta, a média foi de 35,91%.

Comparando-se os dados de geração e consumo de energia elétrica, constatou-se que todos os algoritmos aplicados sobre a base de geração de energia apresentaram baixos erros e resultados muito próximos. A média dos valores de erro percentual (MAPE médio) para a Tabela 9 é de 13,77% - que é quase metade da média obtida dentro do melhor resultado para o consumo de energia (26,47%). Terra (2003) apresenta, em sua análise, fontes para determinar erros e inconsistências nos dados de consumo de energia, sendo estas a ocorrência de registros incorretos, falhas no fornecimento e eventos esporádicos. Dados os resultados da Tabela 9 e o tamanho reduzido da instalação da usina fotovoltaica do IFMG - *Campus* Bambuí, sugere-se que os dados de geração estão menos suscetíveis às fontes de erro apontadas por Terra (2003), apresentando perfis de medições mais consistentes e lineares.

Partindo para a experimentação dos dados meteorológicos, comparando os resultados para o consumo de energia elétrica em HFP, a partir das Tabelas 5 e 6, evidenciou-se que a adição dos atributos meteorológicos, majoritariamente, piorou os resultados dos modelos. A Tabela 5 apresenta média de porcentagem de erro (média de MAPE) igual a 26,47% enquanto, após a adição dos novos atributos, na Tabela 6, obteve-se média de 37,77%. Para as bases de dados sobre consumo de energia elétrica em HP, os resultados são semelhantes. A partir das Tabelas 7 e 8, identifica-se uma piora de desempenho após a adição dos novos atributos. Na Tabela 7, destaca-se uma média de porcentagem de erro (média de MAPE) igual a 35,91%, e, após a adição dos novos atributos, na Tabela 8, a média sobe para 41,91%.

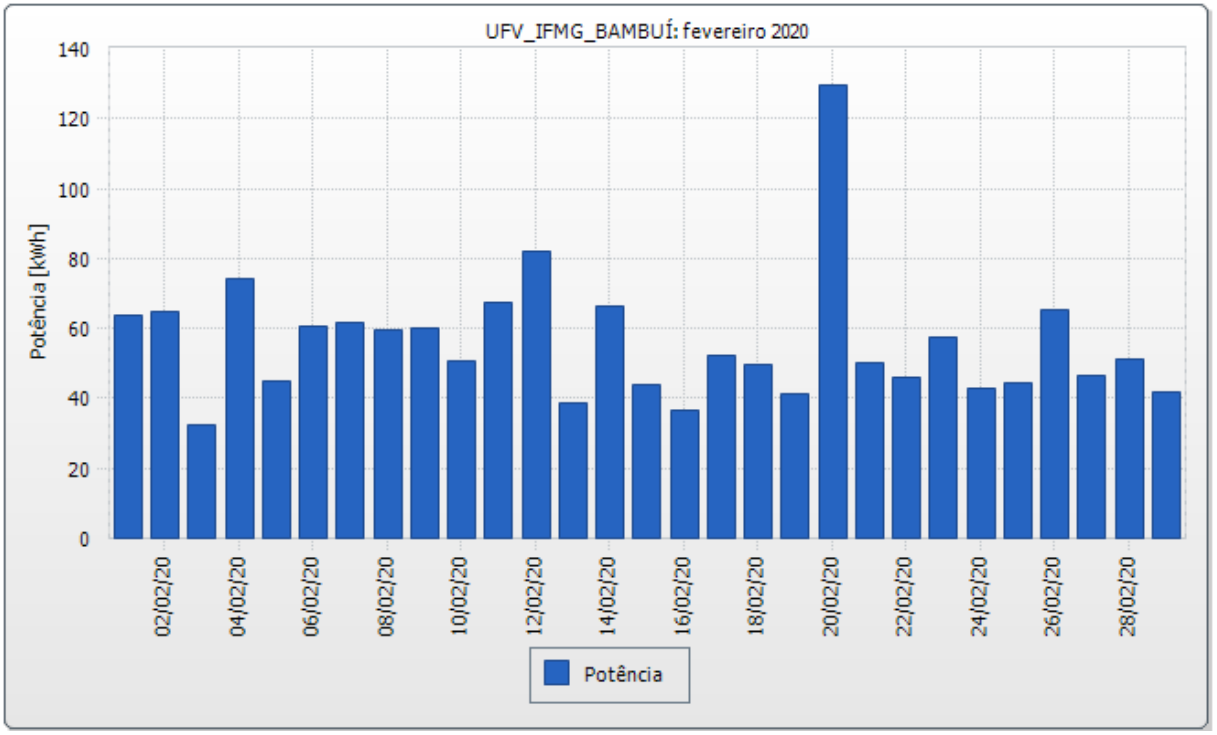
Estendendo-se a comparação para as bases de geração de energia elétrica mensal, a partir das Tabelas 9 e 10, percebe-se um comportamento contrário aos dados de consumo. Na Tabela 9, alcançou-se uma média de erros percentuais (média de MAPE) igual a 13,77%, e, após a adição dos novos atributos, na Tabela 10, a média foi levemente reduzida para 12,96%. Por fim, analisando-se as bases de dados de geração de energia elétrica diária, para o mês de fevereiro de 2020 (Tabela 11), obteve-se uma média de erros percentuais igual a 57,87%, e, após a adição dos novos atributos (Tabela 12), a média caiu para menos da metade, ficando igual a 27,96%. Já para o mês de julho de 2020, as Tabelas 13 e 14 mostram uma leve redução da média de erros percentuais após a adição dos atributos meteorológicos, obtendo média de 14% para a Tabela 13, e de 11,85%, para a Tabela 14.

Uma possível explicação para esta relação entre os atributos é a correlação, explorada por Jawad *et al*. (2020), em que a correlação entre um atributo e outro é mensurada através

dos métodos de Spearman e de Person. Dessa maneira, é avaliado se a presença de determinado atributo terá impacto positivo ou negativo na construção dos modelos preditivos. Para a comparação em questão, observou-se que a presença dos atributos meteorológicos piorou o desempenho dos modelos gerados sobre os dados de consumo de energia elétrica, podendo indicar uma baixa correlação entre estes. Já para os modelos obtidos sobre dados de geração de energia elétrica, notou-se um incremento do desempenho, podendo indicar uma boa correlação entre os atributos.

Para a análise diária sobre dados de geração de energia elétrica, aponta-se que os perfis de medições do sistema fotovoltaico refletem-se nos modelos de predição construídos. O Gráfico 2, a seguir, apresenta a geração de energia elétrica durante o mês de fevereiro de 2020, sendo nítida a visualização de um volume menor de geração de energia, além de um perfil caótico.

Gráfico 2 - Geração de energia Fevereiro de 2020

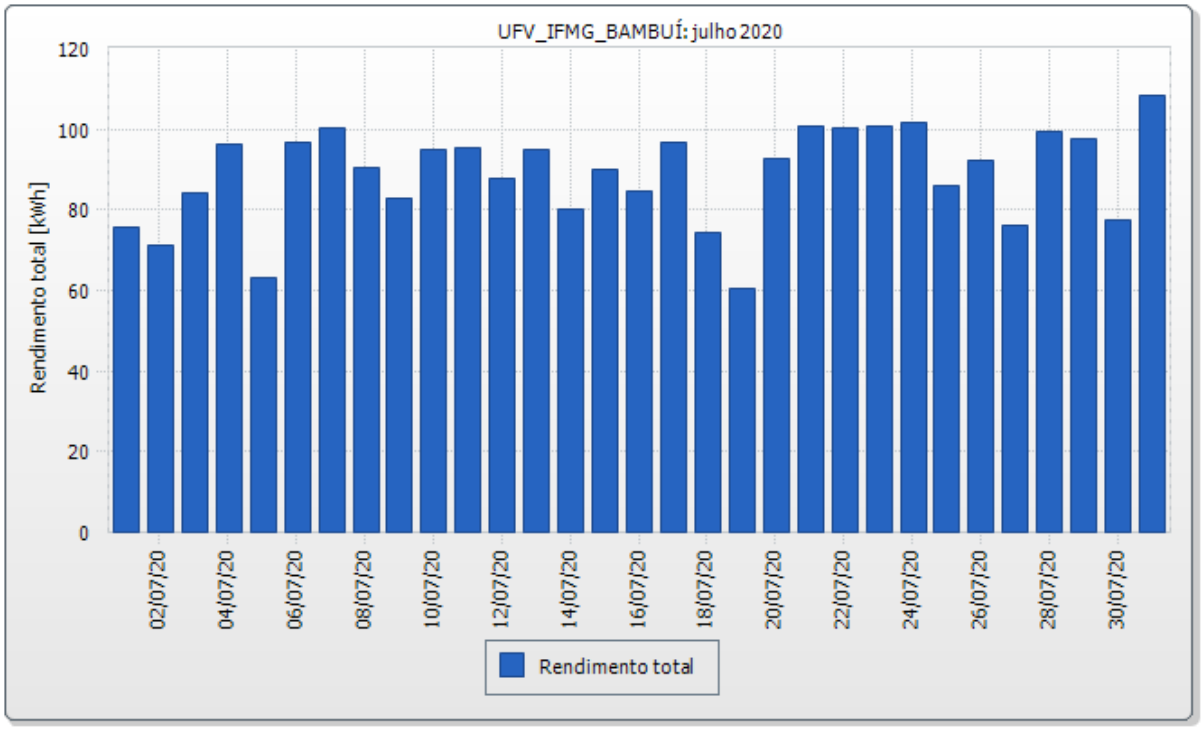


Fonte: SMA Solar Technology AG (2021).

Os modelos dispostos na Tabela 11 apresentaram uma média de erro percentual (MAPE) igual a 57,87%, obtendo-se apenas um modelo com esta métrica inferior a 30% (KNN com k ajustado a 10). Demonstram-se, assim, resultados insatisfatórios das técnicas empregadas sobre os dados do mês de menor geração de energia elétrica. O Gráfico 3 exibe a geração de energia elétrica durante o mês de julho de 2020, a partir da qual se constatam

maiores níveis de geração de energia elétrica e também um perfil mais organizado. Os modelos dispostos na Tabela 13 evidenciam uma média de erro percentual (MAPE) igual a 14%, valor cerca de quatro vezes inferior ao do mês de fevereiro.

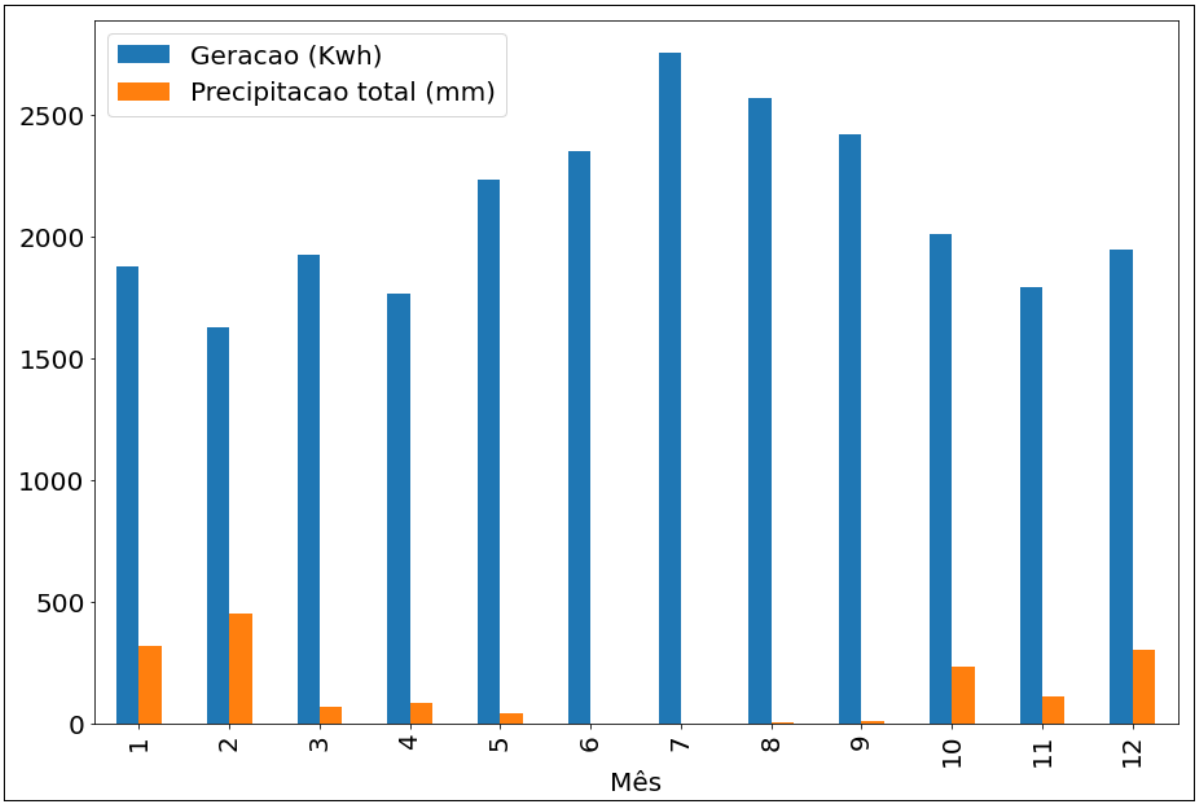
Gráfico 3 - Geração de energia Julho de 2020



Fonte: SMA Solar Technology AG (2021).

Outra maneira de se entender os resultados obtidos é por meio da análise do volume de precipitação associado à geração de energia elétrica em cada mês. A seguir, o Gráfico 4 dispõe estas medidas, evidenciando os efeitos da presença ou não de precipitação sobre as predições realizadas nos meses escolhidos. A partir do Gráfico 4, identifica-se que o mês de fevereiro, que obteve os modelos com maiores erros percentuais, foi também o de maior precipitação de todo o ano de 2020. De maneira semelhante, o mês de julho, que atingiu os modelos com menores erros percentuais, apresentou precipitação muito baixa ou inexistente. Dessa maneira, compreende-se que a presença da chuva pode ser um indicativo que influencie nos desempenhos dos modelos preditivos.

Gráfico 4 - Geração de energia e Precipitação Mensal em 2020



Fonte: Adaptado de SMA Solar Technology AG (2021).

5 CONCLUSÃO

Planejar o gasto de recursos é indispensável em qualquer escopo, desde uma grande organização, composta por diversos colaboradores, até um trabalhador que gerencia o seu salário com as contas do mês. Por outro lado, planejar-se pode ser uma tarefa não tão simples, ressaltando o caso de instituições públicas de ensino onde os recursos são cada vez mais limitados. Este tipo de organização armazena volumes consideráveis de informações, em detrimento de processos burocráticos, mas nem sempre atribui outros usos aos dados armazenados. Este trabalho aplicou técnicas de Mineração de Dados sobre informações de consumo e geração de energia elétrica, buscando avaliar o poder preditivo dos modelos e a possibilidade de auxiliar o *campus* a entender o perfil destes recursos.

Na busca de enriquecer a análise desta pesquisa, os dados avaliados foram caracterizados em 10 bases diferentes, variando a granularidade de tempo, quantidades de instâncias e atributos preditores. Para a realização do estudo, foram empregados os algoritmos KNN, RNA, Regressões e *Random Forests*, e os modelos obtidos foram avaliados através das técnicas MAE, MAPE e RMSE. Dentre os resultados alcançados, os algoritmos que apresentaram maior destaque foram o RNA e a Regressão Linear, com 3 modelos de melhor precisão cada dentre as 10 bases de dados avaliadas, e os de menor destaque, o KNN e o *Random Forest*, apresentando apenas 2 destaques cada (Tabela 15). Outro fator ressaltado foi o desempenho geral dos algoritmos em diferentes naturezas dos dados. De acordo com as métricas empregadas, as informações de consumo de energia elétrica mostraram ser mais difíceis de serem preditas que as de geração de energia elétrica. Em sequência, a adição dos atributos meteorológicos à análise proposta revelou que, para informações de consumo de energia elétrica, o desempenho dos algoritmos piorou; enquanto, para as informações de geração, o desempenho melhorou, como apontado na Seção 4.3.2. Por fim, a análise diária revelou uma relação entre a presença de chuva e o desempenho obtido pelos algoritmos, destacando que, em um mês sem precipitação, os modelos são mais precisos quando comparados ao mês de maior precipitação para o ano analisado.

Para trabalhos futuros, sugere-se a utilização de outras ferramentas na construção dos modelos preditivos, tais como as linguagens de programação Python e R, visando a uma maior escalabilidade dos dados. Além disso, seria interessante expandir o escopo do estudo para outros consumidores de energia com usinas fotovoltaicas instaladas possuindo dados disponíveis para análise, explorando as informações de gasto e consumo em ambientes com

características similares ou não. Outro ponto interessante seria a realização de um estudo de correlação entre os atributos-classe e os meteorológicos, de maneira a determinar até onde seria viável a adição de tais informações. Mais uma possibilidade seria a determinação do valor de uma futura cobrança a partir de modelos preditivos gerados com dados de consumo.

REFERÊNCIAS BIBLIOGRÁFICAS

ABREU, Thays; JUNIOR, Carlos R. Santos; LOPES, Mara L. M.; LOTUDO, Anna Diva. Previsão de Cargas Elétricas utilizando uma Rede Neural ARTMAP Fuzzy com Treinamento Continuado. In: **Anais XIII Brazilian Congress on Computational Intelligence**. 2017. Disponível em: http://cbic2017.org/papers/cbic-paper-88.pdf. Acesso em: 3 maio 2020.

ALVES, Marleide F.; LOTUFO, Anna Diva P.; LOPES, Mara Lúcia M. Seleção de variáveis stepwise aplicadas em redes neurais artificiais para previsão de demanda de cargas elétricas. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, v. 1, n. 1, 2013. Disponível em: https://proceedings.sbmac.org.br/sbmac/article/view/144. Acesso em: 26 abr. 2020.

BAMBUÍ. INSTITUTO FEDERAL DE EDUCAÇÃO CIÊNCIA E TECNOLOGIA DE MINAS GERAIS. Portal. Disponível em: https://www.bambui.ifmg.edu.br/portal/index.php/. Acesso em: 10 maio 2020.

BECKER, Bertha Koiffmann. Reflexões sobre hidrelétricas na Amazônia: água, energia e desenvolvimento. **Boletim do Museu Paraense Emílio Goeldi. Ciências Humanas**, Belém, v. 7, n. 3, p. 783-790, dez. 2012 . Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1981-81222012000300011&lng=pt&nrm=iso. Acesso em 28 fev. 2020.

BREIMAN, Leo. Random forests. **Machine learning**, v. 45, n. 1, p. 5-32, 2001. Disponível em: https://link.springer.com/article/10.1023/A:1010933404324. Acesso em: 16 set. 2020.

BRITO, Miguel C.; SILVA, José A. Energia fotovoltaica: conversão de energia solar em eletricidade. **Lisboa, Faculdade de Ciências da Universidade de Lisboa**, 2006. Disponível em: http://solar.fc.ul.pt/i1.pdf. Acesso em: 27 abr. 2020.

CHAI, Tianfeng; DRAXLER, Roland R. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. **Geoscientific model development**, v. 7, n. 3, p. 1247-1250, 2014. Disponível em: https://gmd.copernicus.org/articles/7/1247/2014/. Acesso em 15 mar. 2021.

D. W. Aha. Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. **International Journal of Man-Machine Studies**, 36(2):267-287, 1992. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/002073739290018G. Acesso em: 17 set. 2020.

DE MYTTENAERE, Arnaud De; GOLDEN, Boris; LE GRAND, Bénédicte; ROSSI, Fabrice. Mean absolute percentage error for regression models. **Neurocomputing**, v. 192, p. 38-48, 2016. Disponível em: https://arxiv.org/abs/1605.02541. Acesso em 15 ago. 2020.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37-37, 1996. Disponível em: https://www.aaai.org/ojs/index.php/aimagazine/article/view/1230. Acesso em 14 de set. 2020.

GASPARIN, Fabiano Perin; KRENZINGER, Arno. Desempenho de um sistema fotovoltaico em dez cidades brasileiras com diferentes orientações do painel. **Revista Brasileira de Energia Solar** , v. 8, n. 1, p. 10-17, 2017. Disponível em: https://rbens.emnuvens.com.br/rbens/article/view/169/160. Acesso em: 12 ago. 2020.

GIL, Antonio Carlos. **Como elaborar projetos de pesquisa**. São Paulo: Atlas, 2002.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. **Data mining: concepts and techniques**. Elsevier, 2011.

JAWAD, Muhammad; NADEEM, Malik Sajjad Ahmed; SHIM, Seong-O; KHAN, Ishtiaq Rasool; SHAHEEN, Aliya; HABIB, Nazneen; HUSSAIN, Lal; AZIZ, Wajid. Machine Learning Based Cost Effective Electricity Load Forecasting Model Using Correlated Meteorological Parameters. **IEEE Access**, [S.L.], v. 8, p. 146847-146864, 20 ago. 2020. Institute of Electrical and Electronics Engineers (IEEE). http://dx.doi.org/10.1109/access.2020.3014086. Acesso em: 10 mar. 2021.

KOPILER, Alberto A; SILVA, Victor Navarro A. L. da; OLIVEIRA, Luiz Antônio A. de; LINDEN, Ricardo; SILVA, Luis Renato A. de A.; FONSECA, Bruno L. da C. Redes Neurais Artificiais e suas aplicações no setor elétrico. **Revista de Engenharias da Faculdade Salesiana**, n. 9, p. 27-33, 2019. Disponível em: http://www.fsma.edu.br/RESA/Edicao9/FSMA_RESA_2019_1_04.pdf. Acesso em: 14 abr. 2020.

MACHADO, Carolina T.; MIRANDA, Fabio S. Energia Solar Fotovoltaica: uma breve revisão. **Revista virtual de química**, v. 7, n. 1, p. 126-143, 2015. Disponível em: http://rvq-sub.sbq.org.br/index.php/rvq/article/view/664/508. Acesso em: 12 ago. 2020.

MASSERONI, James; OLIVEIRA, Cristina Maria de. Utilização de grupos geradores diesel em horário de ponta. **Revista Modelos–FACOS/CNEC**, v. 2, n. 2, p. 52-56, 2012. Disponível em: http://facos.edu.br/publicacoes/revistas/modelos/agosto_2012/pdf/utilizacao_de_grupos_geradores_diesel_em_horario_de_ponta.pdf. Acesso em: 12 jun. 2021.

MORESI, Eduardo. Metodologia da pesquisa. Brasília: **Universidade Católica de Brasília**, v. 108, p. 24, 2003. Disponível em: http://www.inf.ufes.br/~pdcosta/ensino/2010-2-metodologia-de-pesquisa/MetodologiaPesquisa-Moresi2003.pdf. Acesso em: 9 set. 2020.

PENTAHO. **Time Series Analysis and Forecasting with Weka**. Software. Versão 1.0.27. [S.L.], 24 mar. 2014. Disponível em: https://wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka. Acesso em: 17 set. 2020.

PHAM-GIA, T.; HUNG, T. L. The mean and median absolute deviations. **Mathematical and Computer Modelling**, v. 34, n. 7-8, p. 921-936, 2001. Disponível em: https://www.sciencedirect.com/science/article/pii/S0895717701001091. Acesso em: 15 set. 2020.

QUEIROZ, Jamerson Viegas; QUEIROZ, Fernanda Cristina Barbosa Pereira; HÉKIS, Hélio Roberto. Gestão estratégica e financeira das Instituições de ensino superior: um estudo de caso. **Iberoamerican Journal Of Industrial Engineering** , Florianópolis, v. 3, n. 1, p.98-117, jul. 2011. Semestral. Disponível em: http://incubadora.periodicos.ufsc.br/index.php/IJIE/article/view/504/pdf. Acesso em: 26 fev. 2020.

RAFIQUE, Muhammad; TAREEN, Aleem Dad Khan; MIR, Adil Aslim; NADEEM, Malik Sajjad Ahmed; ASIM, Khawaja M.; KEARFOTT, Kimberlee Jane. Delegated Regressor, A Robust Approach for Automated Anomaly Detection in the Soil Radon Time Series Data. **Scientific Reports**, [S.L.], v. 10, n. 1, 20 fev. 2020. Springer Science and Business Media LLC. http://dx.doi.org/10.1038/s41598-020-59881-9. Acesso em: 16 mar. 2021.

RODRIGUES, Sandra Cristina Antunes. **Modelo de regressão linear e suas aplicações**. 2012. Tese de Doutorado. Universidade da Beira Interior. Disponível em: https://ubibliorum.ubi.pt/handle/10400.6/1869. Acesso em: 13 ago. 2020.

SABBAGH, Rafael. **Scrum: Gestão ágil para projetos de sucesso**. Editora Casa do Código, 2014.

SANTOS, Hellen Geremias dos; NASCIMENTO, Carla Ferreira do; IZBICKI, Rafael; DUARTE, Yeda Aparecida de Oliveira; CHIAVEGATTO FILHO, Alexandre Dias Porto. Machine learning para análises preditivas em saúde: exemplo de aplicação para predizer óbito em idosos de São Paulo, Brasil. **Cadernos de Saúde Pública**, [S.L.], v. 35, n. 7, 2019. UNIFESP (SciELO). http://dx.doi.org/10.1590/0102-311x00050818. Acesso em: 27 mar. 2021.

SARAIVA, Illyushin Zaak; OLIVEIRA, Nadja Simone Menezes Nery; MOREJON, Camilo Freddy Mendoza. Impactos das Políticas de Quarentena da Pandemia Covid-19, Sars-Cov-2, sobre a CT&I Brasileira: prospectando cenários pós-crise epidêmica. **Cadernos de Prospecção**, v. 13, n. 2 COVID-19, p. 378, 2020. Disponível em: https://cienciasmedicasbiologicas.ufba.br/index.php/nit/article/view/36066. Acesso em:12 abr. 2020.

SCHUCH, Regis; DILL, Sérgio Luis; SUASEN, Paulo Sérgio; PADOIN, Edson Luis; CAMPOS, Mauricio de. Mineração de dados em uma subestação de energia elétrica. In: **Proceedings of the 9th Brazilian Conference on Dynamics, Control and Their Applications–dincon** . 2010. p. 804. Disponível em: http://sbmac.locaweb.com.br/dincon/trabalhos/PDF/energy/68015.pdf. Acesso em: 27 fev. 2020.

SILVA, Gabriel Allan de Araujo; COSTA, Lucas Gabriel Mindêlo da. **Viabilidade econômica de sistemas fotovoltaicos integrados a residências unifamiliares em João Pessoa**. 2018. 24 f. TCC (Graduação) - Curso de Engenharia Civil, Uninassau, João Pessoa, 2018. Cap. 3. Disponível em: https://bityli.com/VdVDm. Acesso em: 13 ago. 2020.

SILVA, Thays Aparecida de Abreu. Previsão de cargas elétricas através de um modelo híbrido de regressão com redes neurais. 2012. Disponível em:

https://repositorio.unesp.br/handle/11449/87107. Acesso em: 20 abr. 2020.

SMA SOLAR TECHNOLOGY AG. **Sunny Portal**. 2021. Disponível em: https://www.bambui.ifmg.edu.br/sunnyportal.html. Acesso em: 21 jan. 2021.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao datamining: mineração de dados**. Ciência Moderna, p. 3, 2009.

TAREEN, Aleem Dad Khan; ASIM, Khawaja M.; KEARFOTT, Kimberlee Jane; RAFIQUE, Muhammad; NADEEM, Malik Sajjad Ahmed; IQBAL, Talat; RAHMAN, Saeed Ur. Automated anomalous behaviour detection in soil radon gas prior to earthquakes using computational intelligence techniques. **Journal of Environmental Radioactivity**, [S.L.], v. 203, p. 48-54, jul. 2019. Elsevier BV. http://dx.doi.org/10.1016/j.jenvrad.2019.03.003. Acesso em: 15 mar. 2021.

TERRA, Guilherme Saad. **Uma Metodologia de Mineração de Dados para previsão de Cargas**. Rio de Janeiro, 2003. Disponível em: http://www.coc.ufrj.br/pt/teses-de-doutorado/147-2003/943-guilherme-saad-terra. Acesso em: 28 fev. 2020.

WAZLAWICK, Raul. **Metodologia de pesquisa para ciência da computação**. Elsevier Brasil, 2017.

WEKA. **The University of Waikato**. Software. Versão 3.8.5. [S.L.], 2016. Disponível em: http://www.cs.waikato.ac.nz/ml/weka/. Acesso em: 19 abr. 2020.

APÊNDICE A - Disposição dos *Sprints* do trabalho

Tabela 16 - *Sprint* 2

ID	Tarefas	Horas
2	Revisão Bibliográfica	7
3	Estudo de técnicas de Mineração	24
4	Estudo da Ferramenta WEKA	22
5	Coleta dos dados	12

Fonte: Elaborada pelo autor, 2020.

Tabela 17 - *Sprint* 3

ID	Tarefas	Horas
2	Revisão Bibliográfica	7
6	Análise e Preparação dos dados	15
7	Aplicação dos Algoritmos	20
8	Análise dos Resultados	30

Fonte: Elaborada pelo autor, 2020.

Tabela 18 - *Sprint* 4

ID	Tarefas	Horas
2	Revisão Bibliográfica	6
9	Escrita da Monografia	40
10	Defesa	34

Fonte: Elaborada pelo autor, 2020.