Previsão de Quantidade de Clientes e Potência Instalada de Geração Distribuída em Pernambuco

Iago Gade Gusmão Carrazzoni
Escola Politécnica de Pernambuco
Universidade de Pernambuco
Recife, Brasil
iggc@poli.br

João Fausto Lorenzato de Oliveira
Dr. em Ciências da Computação
Prof. Titular na Universidade de Pernambuco
Recife, Brasil
fausto.lorenzato@upe.br

***Resumo*—As fontes de energia renováveis vêm crescendo muito nos últimos anos, por reduzir os impactos ambientais e por haver incentivos do governo para sua participação na produção de energia elétrica, como o sistema de compensação de energia para geração de energia renovável que existe no Brasil. Mas a compensação acaba injetando muita energia na rede, acarretando em problemas na qualidade do fornecimento de energia elétrica. Este projeto tem como objetivo conduzir um estudo sobre a previsibilidade da quantidade de clientes de geração distribuída (GD) e a sua potência instalada (em kW) em Pernambuco, com o auxílio do aprendizado de máquina, para que seja possível auxiliar no processo de tomada de decisão sobre a geração distribuída no estado. A partir da base de "Relação de empreendimentos de Geração Distribuída", disponibilizada no site da Agência Nacional de Energia Elétrica (ANEEL), foi utilizado o modelo *Support Vector Regression* (SVR) para fazer a previsão. Os resultados mostram que a quantidade de clientes e a potência instalada das GDs em Pernambuco tem a tendência de continuar crescendo. Portanto, com as previsões feitas nesse trabalho, espera-se alertar os responsáveis pelo planejamento da rede de distribuição de eletricidade de Pernambuco sobre o crescimento da geração distribuída no estado, para assim evitar os problemas que a grande quantidade de injeção de energia excedente da GD na rede possa causar no futuro.**

Palavras-Chave*—Previsão, Clientes, Potência Instalada, Geração Distribuída

I. Introdução

As fontes de energia renováveis têm sido o foco de muitas nações para sua participação na produção de energia elétrica, com os governos alterando legislações para facilitar a sua integração no mercado (Doyle *et al.* [1]). Entre elas, a energia solar apresenta um crescimento exponencial, por reduzir os impactos ambientais e por haver incentivos do governo, como o sistema de compensação de energia para geração e a isenção de impostos para fontes renováveis (Doyle *et al.* [1] e Rigo *et al.* [2]). No Brasil, a geração solar fotovoltaica possui a maior taxa de crescimento entre as fontes de energia renováveis (Rigo *et al.* [2]), já que pode ser instalada em residências.

Com as fontes de energia renováveis, uma nova abordagem de produção de eletricidade surgiu, a geração distribuída (GD), que são sistemas de produção de energia em pequena escala conectados às redes de distribuição, baseados principalmente em energias renováveis. Uma unidade consumidora, por um sistema de GD, pode produzir eletricidade para autoconsumo e distribuir a energia excedente para a rede local após o consumo próprio (Freitas [3]). Com isso, a GD traz benefícios comerciais (cliente economiza na conta de energia) e ambientais (redução do impacto no ambiente).

Porém, apesar dos benefícios que a GD trouxe, alguns desafios surgiram com o crescimento acelerado no Brasil, o qual mostrou um crescimento substancial de 2013 a 2020 (Rigo *et al.* [2] e Carstens *et al.* [5]), que não foi estagnado mesmo após a pandemia do COVID-19 (Costa *et al.* [4]). Carstens *et al.* [5] destacam alguns: a falta de desenvolvimento de novas tecnologias, a escassez de profissionais qualificados e o pequeno mercado interno no Brasil. Já Souza *et al.* [9] mostram que, por conta da crescente integração do sistema de geração distribuída, o controle de tensão das redes de distribuição se tornou um grande desafio enfrentado pelos engenheiros das distribuidoras, porque quando uma unidade GD é conectada a uma rede de distribuição, a energia excedente da GD está sendo injetada na rede. Isso pode causar alterações nos níveis de tensão do sistema e provocar problemas na qualidade do fornecimento de energia elétrica, como, por exemplo, danificar os equipamentos que fazem o controle da regulação de tensão do sistema.

Com isso, surge o questionamento: qual será o cenário futuro da GD em Pernambuco? Ou seja, quantos consumidores adotarão o sistema de GD e quanta energia será injetada na rede nos próximos anos no estado de Pernambuco? Isso mostra que a previsão de quantidade de clientes e de energia é essencial no planejamento, gerenciamento e conservação de energia e, segundo Albuquerque *et al.* [10], as abordagens de inteligência artificial são os métodos mais avançados para fazer essas previsões.

Este projeto tem como objetivo conduzir um estudo sobre a previsibilidade da quantidade de clientes de GD e a potência instalada (kW) por eles em Pernambuco para os próximos anos, com o auxílio do aprendizado de máquina, para que seja possível ter uma noção de como será o cenário futuro da geração distribuída, com o intuito de auxiliar no planejamento da rede elétrica e evitar os problemas causados pelo rápido aumento da GD em Pernambuco.

O restante do artigo está dividido entre Revisão Teórica na seção 2, Metodologia na seção 3, Resultados na seção 4 e Conclusão na seção 5.

II. Revisão Teórica

A literatura sobre a geração distribuída, especialmente sobre impactos regulatórios e econômicos da sua adoção no Brasil, tem trazido conhecimento importante para o processo de tomada de decisão de corporações e governos. Então, para entender a situação atual da GD no país, é necessário realizar um estudo sobre trabalhos relacionados ao assunto.

Doyle *et al.* [1] mostram que, no Brasil, até o ano 2000, a energia solar era apenas usada em pequenas instalações rurais, não conectadas a rede, então não era economicamente viável fazer sua expansão. Mas isso mudou com as Resoluções Normativas da Agência Nacional de Energia Elétrica (ANEEL): Resolução Normativa n° 482/2012 e Resolução Normativa n° 687/2015. A 482 estabeleceu as condições para compensação do excedente de produção de eletricidade, no qual se a energia gerada pela unidade consumidora for maior que a consumida, essa energia é cedida (injetada na rede), por meio de empréstimo gratuito à distribuidora local e posteriormente compensada no consumo de energia elétrica do consumidor. Já a 687 estabeleceu que a energia gerada por consumidores de geração distribuída pode ser compartilhada entre várias unidades consumidoras, desde que dentro da mesma área de concessão. Com isso, houve um crescimento na nova abordagem de produção de eletricidade, a geração distribuída.

Costa *et al.* [4] mostra um estudo semelhante a [1], mas analisando também como o cenário das energias renováveis foi alterado depois da pandemia do COVID-19 e destacando os problemas de difusão de energia solar e eólica no Brasil, tais como a falta de infraestrutura da rede elétrica e o alto custo de importação de materiais, por conta da baixa produção de equipamentos nacionais. É mostrado que mesmo com a pandemia do COVID-19, a GD continuou crescendo.

Para uma unidade consumidora produzir eletricidade para autoconsumo e distribuir a energia excedente para a rede local após o consumo próprio foi um grande motivo para o crescimento na adoção do sistema de GD, pois ocorre uma considerável redução na conta de energia elétrica da unidade residencial no qual a GD está instalada e até mesmo de outras unidades residenciais incluídas para receber os créditos de energia, como mostra Freitas [3].

Um estudo semelhante ao escopo deste trabalho é mostrado em Carstens *et al.* [5], que dispõe estimativas da ANEEL para o mercado de GD, no qual o número de sistemas solares fotovoltaicos residenciais e comerciais instalados no Brasil até o ano de 2024 será de aproximadamente 886 mil unidades. Apesar das estimativas da ANEEL serem semelhantes ao objetivo desse artigo, não é especificada a quantidade para Pernambuco, assim como não é mostrada a energia injetada na rede pelas unidades fotovoltaicas.

Tudo isso mostra a importância de haver estudos sobre previsão do cenário de GD no Brasil, para que seja possível estimar o planejamento de investimentos em geração e distribuição de eletricidade. Com isso, foi revisada a literatura relacionada a previsões de consumo de energia, para checar se houve trabalhos semelhantes sobre o assunto e para auxiliar na avaliação de qual modelo de aprendizado de máquina é mais adequado e eficiente para trabalhos com previsão de energia.

Li *et al.* [6] usam o método de *Support Vector Machine* (SVM), para fazer previsão de consumo de energia. Esse método é baseado em regressão estatística e é uma ferramenta importante no campo de reconhecimento de padrões e predição de regressão para resolver problemas de alta dimensão, não lineares e pequenas amostras.

Para melhorar a eficiência das previsões de consumo de energia, o modelo de *Support Vector Regression* (SVR) é proposto por Zhong *et al.* [7]. O SVR é o SVM direcionado para o método de regressão. Essa proposta é semelhante a [6], pois ambos utilizam o mesmo modelo de aprendizado supervisionado para fazer a previsão do consumo de energia. O estudo mostra que o modelo com SVR atingiu uma alta precisão.

Assim como Li *et al.* [6] e Zhong *et al.* [7], Xu *et al.* [8] utiliza o SVM, mas em um modelo híbrido com algoritmo genético. É destacado que o SVM é utilizado por conseguir realizar uma previsão eficaz mesmo com uma base de dados insuficiente, também pode ser aplicado para prever amostras não lineares e é capaz de evitar *overfitting* (o modelo tem um desempenho excelente no teste, mas apresenta um resultado ruim na previsão de novos valores). O algoritmo genético é utilizado para encontrar encontrar soluções eficientes com requisitos computacionais relativamente modestos, sendo usado principalmente para determinar de forma adaptativa dois parâmetros do SVM, que influenciam no desempenho da previsão. No trabalho, eles fazem uma previsão de médio prazo do consumo de energia elétrica para oleodutos.

Albuquerque *et al.* [10] e Sujan *et al.* [11] usam diversos modelos estatísticos, incluindo modelos de machine learning, para fazer a previsão precisa do consumo de eletricidade. Entre eles: *Autoregressive integrated moving average* (ARIMA); *Random walk*; *Least absolute shrinkage and selection operator* (Lasso); *Least angle regression* (Lars); *Lasso Lars*; *Ridge Regression*; *Elastic Net*; *Random Forest*; *Long Short Term Memory*, Árvore de Decisão (ou árvore de classificação e regressão). É destacado em [10] que os algoritmos que obtiveram as melhores previsões foram os modelos de aprendizado de máquina, especialmente o *Random Forest* e o *Lasso Lars*.

A previsão de eletricidade demandada por hora de dois edifícios educacionais utilizando Random Forest é feito por Wang *et al.* [12]. Além disso, as previsões também são feitas com Árvore de Decisão (CART) e SVR. Analisando os resultados, os três modelos foram eficientes nas previsões, mas o Random Forest mostrou uma performance superior do que o CART e o SVR.

O estudo de Kim *et al.* [13] compara a previsão de consumo de energia elétrica em um prédio utilizando dois métodos diferentes: regressão linear (RL), um método estatístico de *machine learning* mais tradicional; e algoritmo de rede neural artificial (ANN). Apesar do ANN ter obtido uma precisão maior nas previsões, ambos os métodos foram eficientes e capazes de atender os requisitos da previsão temporal a longo prazo e em tempo real com base nas taxas de ocupação e

condições ambientais locais do prédio.

Logo, este trabalho se difere dos outros por realizar um estudo de previsão de geração distribuída especificamente para Pernambuco. Vale ressaltar que, diferente deste trabalho, a maioria dos estudos de previsão de energia se referem ao consumo de energia elétrica e não a potência instalada de GD. A potência instalada é medida em kW e se refere a soma das potências nominais da unidade geradora, ou seja, representa a capacidade caso todos os equipamentos estejam ligados e operando ao mesmo tempo.

III. Metodologia

A metodologia se divide em: Base de dados utilizada; Algoritmos de aprendizado de máquina utilizados; Treinamento do modelo; Teste e validação do modelo.

A. Base de dados

A base de dados utilizada foi a "Relação de empreendimentos de Geração Distribuída", disponibilizada pela ANEEL em [14]. Essa base possui os dados referentes às gerações distribuídas, abrangidos pela Resolução Normativa 482/2012. A relação dos empreendimentos é classificada pelas variáveis que compõem sua identificação, incluindo: distribuidora conectada, código do empreendimento, nome do titular, classe de produção, quantidade de unidades consumidoras que recebem os créditos, data da conexão, tipo de unidade produtora, potência instalada (kW), município e unidade de federação onde está localizada. Cada linha representa uma unidade geradora, ou seja, um cliente de geração distribuída.

O site da ANEEL no qual a base se encontra (Relação de empreendimentos de Geração Distribuída [14]) indica que essa base é atualizada diariamente, devido a alta demanda por essas informações. Para esse projeto, a base foi extraída em 14/09/2022, sendo 20/08/2013 a data mais antiga dos dados e 31/07/2022 a data mais recente.

A Tabela 1 exemplifica a extração de uma parte da tabela, com 4 clientes e algumas colunas necessárias para este estudo: "NomAgente", a qual caracteriza o nome da distribuidora; "DthAtualizaCadastralEmpreend", que possui a data de cadastro da GD; "MdaPotenciaInstaladaKW", que possui a potência instalada em kW; e "SigUF", o qual representa a unidade de federação da GD cadastrada. Logo, a Tabela 1 mostra 4 clientes, sendo um de Tocantis (SigUF = TO), dois de Pernambuco (SigUF = PE) e um de Minas Gerais (SigUF = MG), com suas respectivas datas de conexão, potência instalada e nome da distribuidora responsável pelo cliente.

Então, para esse projeto, inicialmente a coluna "SigUF", com a unidade de federação da GD cadastrada, foi filtrada pela UF igual a PE, para pegar os clientes de geração distribuída de Pernambuco. Após isso, foram coletadas as colunas com as datas de cadastro e potência instalada das GDs, a partir das colunas "DthAtualizaCadastralEmpreend" e "MdaPotenciaInstaladaKW", mostradas na Tabela 1. Após as manipulações nos dados, eles foram agrupados em quantidade de clientes e potência instalada (em kW) por mês, ficando em um formato "mês/ano"como mostrado nas Tabela 2 e 3.

Sendo "data_conexao"a data que os clientes conectaram à geração distribuída, a Tabela 2 mostra a série temporal da quantidade de clientes conectados por mês e a Tabela 3 mostra a série temporal da soma da potência instalada, em kW, desses clientes conectados por mês. Por conta da extensão das tabelas, foram dispostos apenas os dados do mês de dezembro de cada ano mais a primeira data (08/2013) e o última data (07/2022).

Tabela 1
PARTE DA TABELA "RELAÇÃO DE EMPREENDIMENTOS DE GERAÇÃO DISTRIBUÍDA.

NomAgente	DthAtualiza CadastralEmpreend	MdaPotencia InstaladaKW	SigUF
ENERGISA TOCANTINS DISTRIBUIDORA DE ENERGIA S.A.	2022-08-26	8,00	TO
COMPANHIA ENERGÉTICA DE PERNAMBUCO	2022-04-26	8,50	PE
COMPANHIA ENERGÉTICA DE PERNAMBUCO	2016-09-01	2,50	PE
CEMIG DISTRIBUIÇÃO S.A	2022-01-23	13,00	MG

Tabela 2
PARTE DA QUANTIDADE DE CLIENTES DE GD CONECTADOS POR MÊS.

data_conexao	quantidade_clientes
2013-08-01	1
2013-12-01	1
2014-12-01	2
2015-12-01	6
2016-12-01	7
2017-12-01	22
2018-12-01	51
2019-12-01	247
2020-12-01	499
2021-12-01	1869
2022-07-01	2309

Tabela 3
PARTE DA POTÊNCIA INSTALADA (KW) DE CLIENTES DE GD CONECTADOS POR MÊS.

data_conexao	potencia_instalada_kw
2013-08-01	3.00
2013-12-01	967.00
2014-12-01	11.51
2015-12-01	61.68
2016-12-01	42.26
2017-12-01	302.90
2018-12-01	803.10
2019-12-01	5170.72
2020-12-01	6533.30
2021-12-01	17344.26
2022-07-01	21892.55

B. Algoritmos de aprendizado de máquina utilizados

O fator principal na escolha do algoritmo foi que o mesmo tivesse uma eficiente precisão da previsão, ou seja, um erro pequeno. Então, para decidir qual algoritmo usar no estudo, foram escolhidos dois modelos, no qual a previsão foi feita com o modelo que apresentou menor erro na fase de teste.

A partir da revisão teórica, foi possível observar que o *Support Vector Regression* (SVR) é muito utilizado nas pesquisas e se mostrou eficiente nas previsões de consumo de energia em Li *et al.* [6], Zhong *et al.* [7], Xu *et al.* [8] e Wang *et al.* [12], logo foi o primeiro modelo selecionado. O segundo selecionado foi o modelo de regressão linear (RL), por ser um modelo estatístico tradicional, simples de ser utilizado e com uma boa precisão de séries temporais (Kim *et al.* [13]).

Um terceiro modelo foi cogitado, o *Random Forest* (RF), por também possuir uma alta precisão nas previsões e também ser muito utilizado nas pesquisas, como mostrado em Albuquerque *et al.* [10], Sujan *et al.* [11] e Wang *et al.* [12]. Mas, por conta do tempo de execução dele, como comentado por Wang *et al.* [12], ele não foi utilizado na comparação. O RF ficou separado para ser utilizado apenas caso o SVR e a regressão linear não fossem validados após o treinamento.

1) Support Vector Regression: O *Support Vector Regression* (SVR) é uma técnica eficaz para aplicações de regressão. O SVR tem origem das *Support Vector Machine* (SVM), que são métodos de aprendizado supervisionado os quais analisam os dados e reconhecem padrões, usado para classificação e análise de regressão (Li *et al.* [6]). Ele é baseado em estatística e combinado com a teoria de minimização de risco estrutural, que engloba o máximo de valores do *dataset* utilizando vetores auxiliares, buscando reduzir o erro. O SVR possui boa eficácia em previsões, como mostrado por Li *et al.* [6], Zhong *et al.* [7], Xu *et al.* [8] e Wang *et al.* [12].

2) Regressão linear: A regressão linear (RL) é uma equação que descreve a relação estatística entre uma ou mais variáveis independentes ou preditoras (entrada ou X) e a variável resposta ou dependente (saída ou Y). O Y é a variável dependente pois varia de acordo com o valor de X. No geral, a RL encontra a linha que melhor representa as variáveis de entrada com a variável de saída (Kim *et al.* [13]).

3) Random Forest: O *Random Forest* (RF) é um modelo de predição do tipo *ensemble learning*, um tipo de modelo mais robusto e complexo de aprendizado de máquina, o qual combina o resultado de múltiplos modelos para produzir um melhor modelo preditivo. O RF consiste em uma coletânea de diferentes Árvores de Decisão (CART). O CART usa regras de decisão simples, comparando os valores do nó raiz (atributos dos dados) e seguindo pelos nós folhas (valor ou classe) correspondentes a partir do resultado da comparação. O RF utiliza um conjunto de árvores em vez de uma única árvore, reduzindo o problema de instabilidade do CART com a combinação da previsão de várias árvores diferentes, como explicado por Wang *et al.* [12]. Por conta de sua precisão, é bastante utilizado na maioria dos casos, mesmo possuindo um tempo de execução maior.

C. Treinamento dos modelos

Para a previsão, foi utilizado o scikit-learn, uma biblioteca de aprendizado de máquina de código aberto para o Python. Para o SVR o ”sklearn.svm.SVR”, para a regressão linear o “sklearn.linear_model.LinearRegression”. Apenas para conhecimento, a regressão com RF utiliza o “sklearn.ensemble.RandomForestRegressor”.

Inicialmente, é necessário normalizar os dados, ou seja, deixar em uma escala de valores entre 0 e 1, pois dessa forma o treinamento do modelo se torna mais eficaz. As figuras 1 e 2 mostram os gráficos das séries normalizadas, sendo a Figura 1 para quantidade de clientes e a Figura 2 para potência instalada:

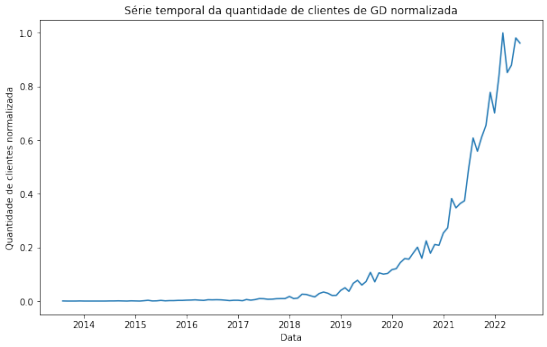

Figura 1. Série temporal da quantidade de clientes de GD normalizada.

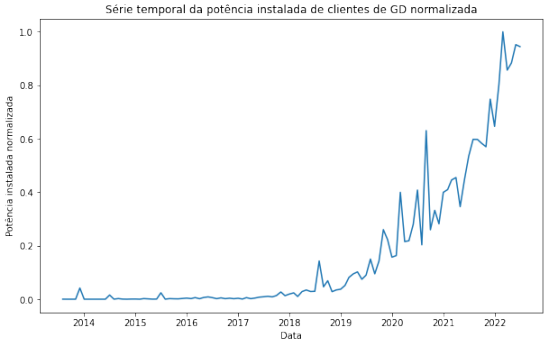

Figura 2. Série temporal da potência instalada de clientes de GD normalizada.

Com a Figura 1, é possível observar o crescimento na quantidade de clientes conectados por mês que a geração distribuída teve nos últimos anos, com a curva possuindo uma natureza menos descontínuo em comparação com a Figura 2.

Analisando a Figura 2, nota-se a natureza mais dispersa dos dados de potência instalada. Isso acontece pois a potência instalada varia muito com os clientes de geração distribuída, podendo o cliente ser de Minigeração ou Microgeração. Na Microgeração, o sistema fotovoltaico tem uma potência de até 75kW e na Minigeração o sistema possui uma potência entre 76kW e 5MW. Com uma análise na base ”Relação de empreendimentos de Geração Distribuída”, é possível filtrar se o cliente é de Micro ou Minigeração pela coluna ”DscPorte”, como mostrado na Tabela 4.

Tabela 4
QUANTIDADE DE MINIGERAÇÃO E MICROGERAÇÃO DE PERNAMBUCO DE 20/08/2013 ATÉ 31/07/2022.

DscPorte	Quantidade	Porcentagem
Microgeração	36324	99,1294%
Minigeração	319	00,8706%

Então, para ser possível aplicar os modelos de aprendizado de máquina em problemas de previsão de séries temporais, a série foi transformada em uma matriz na qual cada valor está relacionado à janelas de tempo (*lags*) que o precede. Ou seja, em um contexto de série temporal, os *lags* em relação a um tempo t são definidos como os valores da série em etapas de tempo anteriores. Por exemplo, o *lag* 1 é o valor em $t-1$.

Para aplicar o *lag*, primeiro é necessário conhecer a autocorrelação dos dados. Em um modelo para prever uma série temporal, seja considerando métodos estatísticos ou através de redes neurais artificiais, é necessário conhecer a relação entre as observações atuais e as anteriores. Com isso, aplicando a função *plot_acf* aos dados, foram gerados os dois gráficos a das figuras 3 e 4, um para a autocorrelação da quantidade de clientes e o outro para a autocorrelação da potência instalada. A análise do gráfico funciona da seguinte forma: a quantidade de *lags* será o valor do eixo x no qual as barras correspondentes ficarem mais próximas ao lado de fora curva. Exemplificando, através do gráfico da Figura 3, é possível notar que sua dimensionalidade de *lags* é aproximadamente 6, enquanto pelo gráfico da Figura 4, a quantidade de é *lags* aproximadamente 7.

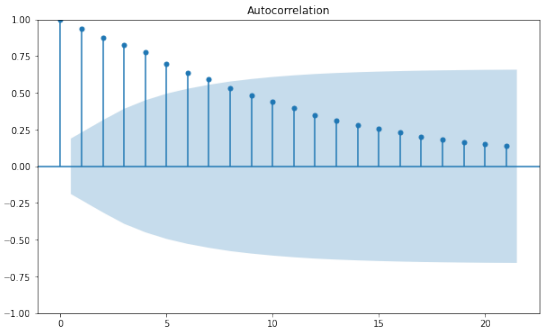

Figura 3. Autocorrelação da quantidade de clientes de GD.

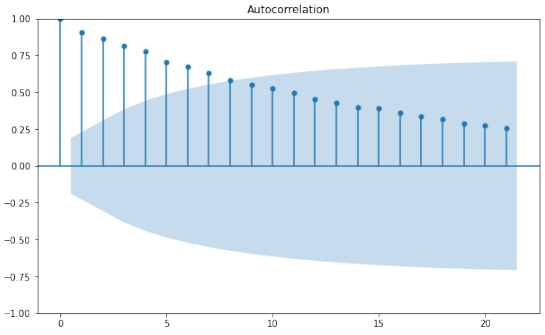

Figura 4. Autocorrelação da potência instalada de clientes GD.

Logo, foram aplicados 6 *lags* nos dados da quantidade de clientes e 7 *lags* nos dados da potência instalada. Apenas para exemplificar, a Tabela 5 se refere a quantidade de clientes com a aplicação dos *lags*, no qual a variável de saída Y é a primeira coluna e as variáveis de entrada X são as colunas restantes com os *lags*. A Tabela 6 é semelhante a tabela 5, mas se refere a potência instalada (kW). Ambas as tabelas mostram apenas uma parte das tabelas completas utilizadas na previsão, com 3 *lags* para os últimos 4 meses dos dados.

Tabela 5
PARTE DA TABELA PARA PREVISÃO DE QUANTIDADE DE CLIENTES GD COM OS LAGS.

quantidade de clientes (t)	lag1(t-1)	lag2(t-2)	lag3(t-3)
0.852145	1.000000	0.838817	0.701374
0.879633	0.852145	1.000000	0.838817
0.981258	0.879633	0.852145	1.000000
0.961683	0.981258	0.879633	0.852145

Tabela 6
PARTE DA TABELA PARA PREVISÃO DE POTÊNCIA INSTALADA DE CLIENTES GD COM OS LAGS.

potência instalada (t)	lag1(t-1)	lag2(t-2)	lag3(t-3)
0.856772	1.000000	0.803225	0.646259
0.883526	0.856772	1.000000	0.803225
0.951892	0.883526	0.856772	1.000000
0.944416	0.951892	0.883526	0.856772

Após finalizar toda a preparação da base de dados, para quantidade e potência instalada de clientes GD agrupados por data de conexão, a base foi separada em dados de treino e dados de teste, para então iniciar o treinamento dos modelos de regressão linear e SVR separadamente.

Para iniciar o treino da regressão linear, não são necessários parâmetros específicos, enquanto o SVR necessita do preenchimento de alguns parâmetros. Os parâmetros do SVR utilizados foram:

- C ou Parâmetro de penalidade do termo de erro, define a penalidade para uma classificação incorreta. Um valor grande de C resulta em uma penalidade maior e um valor menor de C resulta em uma penalidade menor. Logo, não pode ser um valor muito alto para não ocorrer *overfitting* e nem um valor muito baixo para não ocorrer *underfitting* (o modelo não consegue aprender as relações durante o treinamento, resultando em um mau desempenho);
- ϵ especifica o tubo ϵ dentro do qual nenhuma penalidade é associada na função de perda do treinamento, ou seja, os pontos que caem dentro deste tubo são considerados como previsões corretas e não são penalizados pelo algoritmo;
- γ ou Coeficiente de kernel, define até onde a influência de um único exemplo de treinamento chega, sendo os valores

intermediários de γ que dão um modelo com bons limites de decisão.

Após fazer o treinamento com alguns valores diferentes para cada parâmetro, os parâmetros do SVR que tiveram o menor erro foram C = 1000, ϵ = 0.001 e γ = 0.001.

D. Teste e validação do modelo

Foi utilizado o método iterativo de previsão, no qual cada nova previsão é baseada na anterior, ou seja, a primeira previsão fora dos dados conhecidos é utilizada como $t_n - 1$ no próximo ciclo de previsão.

Após a execução das previsões para os dados de teste, para definir a precisão e selecionar o modelo para prever ciclos futuros, inicialmente foram calculados o erro médio absoluto (MAE) e erro médio quadrático (MSE), tanto para a quantidade de clientes quanto para a potência instalada. O MAE é definido pela diferença entre os valores da previsão e os valores reais e o MSE indica a média de diferença quadrática entre a predição do modelo e o valor de destino. Então, o resultado ficou como disposto nas tabelas 7 e 8:

Tabela 7
COMPARAÇÃO ENTRE MAE E MSE DA QUANTIDADE DE CLIENTES DOS TESTES DOS TRÊS MODELOS.

Modelo	MAE	MSE
Support Vector Regression	0.0583	0.0056
Regressão linear	0.0657	0.0072

Tabela 8
COMPARAÇÃO ENTRE MAE E MSE DA POTÊNCIA INSTALADA DOS TESTES DOS TRÊS MODELOS.

Modelo	MAE	MSE
Support Vector Regression	0.1133	0.0206
Regressão linear	0.2415	0.0698

Logo, o modelo escolhido para as previsões dos ciclos futuros foi o SVR, por apresentar os menores MAE e MSE, tanto para a quantidade de clientes quanto para a potência instalada, como mostrado nas tabelas 7 e 8. Além disso, como ambos o SVR e a regressão linear tiveram erros baixos, os modelos tiveram resultados satisfatórios e não foi necessário utilizar o *Random Forest*.

Percebe-se que o erros foram maiores para a previsão da potência instalada. O motivo é a natureza mais aleatória dos dados, como mostrado na Figura 2. Isso se deve ao fato da potência instalada variar muito de cliente para cliente de geração distribuída, pois o cliente pode ser de Minigeração ou de Microgeração, como explicado anteriormente. Ademais, mesmo o cliente sendo de Mini ou Microgeração, a variação da necessidade de energia varia muito. Isso é possível de visualisar através da variância (quanto maior o valor da variância, mais longe os dados estão do valor médio) e desvio padrão (quanto mais próximo de zero, mais homogêneo são os dados) da base, dispostos na tabela 9:

Tabela 9
VARIÂNCIA E DESVIO PADRÃO DOS DADOS

	Variância	Desvio padrão
Quantidade de clientes	1.214991e+06	1102.266308
Potência instalada	5.922662e+07	7695.883313

A seguir, é possível ver as imagens dos testes, no qual a Figura 5 representa o teste para quantidade de clientes de GD e a Figura 6 o teste para a potência instalada de GD. A linha azul representa os dados reais do teste, a linha amarela a previsão para a regressão linear e a linha vermelha a previsão para o SVR.

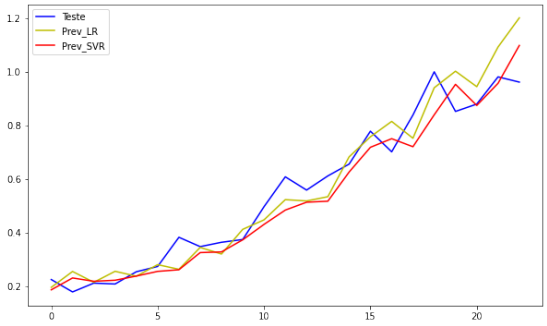

Figura 5. Teste da quantidade de clientes de GD.

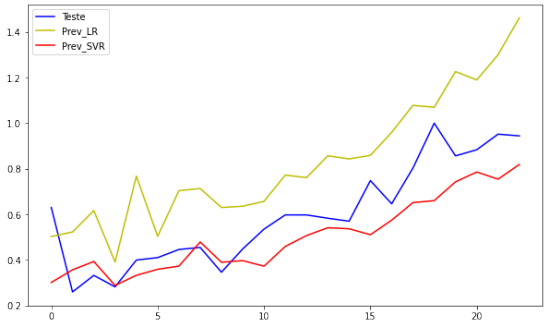

Figura 6. Teste da potência instalada de clientes de GD.

IV. RESULTADOS

Utilizando o *Support Vector Regression*, foram previstos 17 ciclos após o final dos dados conhecidos. Como o último dado conhecido tinha data de dezembro de 2022, para 5 ciclos tem-se a previsão para dezembro de 2022 e para 17 ciclos a previsão para dezembro de 2023, sendo do dado inicial (dezembro de 2013) até o final (dezembro de 2023) um total de 125 ciclos.

Abaixo estão os gráficos ilustrando os resultados, sendo a Figura 7 os valores para quantidade de clientes de GD e a Figura 8 os valores para a potência instalada em kW de GD, com a linha azul indicando os dados conhecidos e a linha vermelha indicando os dados previstos além dos dados conhecidos.

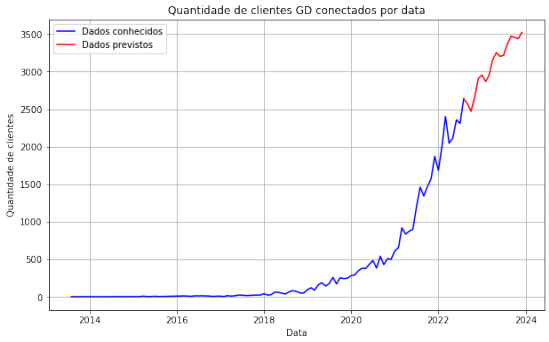

Figura 7. Previsão da quantidade de clientes de GD em Pernambuco até dezembro de 2023.

Figura 8. Previsão da potência instalada (kW) de clientes de GD em Pernambuco até dezembro de 2023.

Selecionando os valores finais para os 5 ciclos (dezembro de 2022) e os 17 ciclos (dezembro de 2023), é possível observar os valores previstos para o final de 2022 e final de 2023 nas tabelas 10 e 11. A Tabela 10 possui os valores de quantidade de clientes de GD conectados por data mais uma coluna adicional, com os valores acumulados. A Tabela 11 é semelhante a Tabela 10, mas com os dados de potência instalada em kW de GD.

Tabela 10
VALORES DE QUANTIDADE DE CLIENTES DE GD EM PERNAMBUCO.

Data	Quantidade de clientes	Quantidade de clientes acumulada
dez/2013	1	2
dez/2014	2	9
dez/2015	6	51
dez/2016	7	158
dez/2017	22	344
dez/2018	51	956
dez/2019	247	3090
dez/2020	499	8029
dez/2021	1869	21721
dez/2022	2909	49891
dez/2023	3522	88738

Tabela 11
VALORES DE POTÊNCIA INSTALADA (KW) DE CLIENTES DE GD EM PERNAMBUCO.

Data	Potência instalada (kW)	Potência instalada (kW) acumulada
dez/2013	967.0	970.0
dez/2014	11.51	1408.8
dez/2015	61.68	2253.04
dez/2016	42.26	3395.29
dez/2017	302.9	5847.26
dez/2018	803.1	17315.04
dez/2019	5170.72	49806.62
dez/2020	6533.3	132085.13
dez/2021	17344.26	274331.81
dez/2022	19409.28	511649.22
dez/2023	18886.50	742635.84

Analisando o gráfico disposto na figura 7, assim como observando os valores da tabela 10, é possível observar que a quantidade de clientes de GD em Pernambuco tende a crescer com o passar do tempo. Isso se torna um ponto de atenção para os responsáveis pelo planejamento da rede de distribuição de energia de Pernambuco, pois com o aumento da quantidade de clientes, vai haver um aumento da energia injetada na rede, ocasionando em um aumento nas alterações nos níveis de tensão do sistema, que, segundo Souza *et al.* [9] provocam problemas na qualidade do fornecimento de energia elétrica.

Sobre a previsão da potência instalada, é possível notar, no gráfico disposto na figura 8 e pelos valores da tabela 11, que a curva não permaneceu crescendo e sim houve uma pequena queda, mesmo que o esperado fosse um crescimento. Essa queda ocorreu por conta do método iterativo de previsão, o qual possui uma característica de acúmulo de erros. Isso é justificado devido a potência instalada ter um erro maior em comparação com a quantidade de clientes.

Esses resultados são importantes pois mostram a importância de haver estudos e trabalhos relacionados a previsão de demanda de energia do cenário de geração distribuída não só em Pernambuco, mas em todo o Brasil, para que seja possível estimar o planejamento de investimentos nas redes de geração e distribuição de energia elétrica.

V. CONCLUSÕES

Portanto, observando as previsões feitas a partir dos dados fornecidos pela ANEEL sobre a quantidade de clientes e a potência instalada (em kW) por clientes de geração distribuída em Pernambuco, percebe-se a tendência de crescimento da GD no estado com o passar do tempo.

Tudo isso mostra a importância de haver estudos sobre previsão de demanda de energia do cenário de geração distribuída em todo o Brasil, para que seja possível estimar o planejamento de investimentos nas redes de geração e distribuição de eletricidade para evitar os problemas causados pelo aumento acelerado da GD, diminuindo assim os seus impactos negativos, anteriormente citados.

Com as previsões feitas nesse trabalho, espera-se alertar os responsáveis pelo planejamento da rede de distribuição de eletricidade de Pernambuco sobre o crescimento da geração

distribuída no estado, para assim evitar os problemas que a grande quantidade de injeção de energia excedente da GD na rede possa causar no futuro.

Trabalhos futuros

Para trabalhos futuros, pode-se aprofundar esse estudo de previsão das gerações distribuídas especificando os seus tipos, como previsão para energia solar e energia eólica. Outra opção é analisar as variáveis para alguma região mais específica, como algum município. Existem muitas possibilidades, já que a base "Relação de empreendimentos de Geração Distribuída" possui muitos campos interessantes que podem ser estudados.

Em relação à previsão, podem ser feitos estudos utilizando outros métodos de *machine learning*, utilizando até o próprio *Random Forest* comentado anteriormente, para comparar as métricas de MSE e MAE e analisar a possibilidade de haver um algoritmo com uma precisão melhor para prever as variáveis utilizadas neste estudo.

Agradecimentos

Agradeço primeiramente ao professor João Fausto Lorenzato de Oliveira, por ter sido meu orientador e ter desempenhado a função com competência e dedicação; e agradeço aos amigos e familiares, em especial a Douglas Azevedo Pereira Dantas, por todo o apoio, que contribuiu para a realização deste trabalho.

Referências

[1] DOYLE, Gabriel Nasser Doile de; ROTELLA JUNIOR, Paulo; ROCHA, Luiz Célio Souza; CARNEIRO, Priscila França Gonzaga; PERUCHI, Rogério Santana; JANDA, Karel; AQUILA, Giancarlo, "Impact of regulatory changes on economic feasibility of distributed generation solar units in Brazil". Sustainable Energy Technologies and Assessments 48, 2021.

[2] RIGO, Paula D.; SILUK, Julio Cezar M.; LACERDA, Daniel P.; SPELLMEIER Júlia P., "Competitive business model of photovoltaic solar energy installers in Brazil". Renewable Energy 181, 2022.

[3] FREITAS, Bruno Moreno Rodrigo de., "What's driving solar energy adoption in Brazil? Exploring settlement patterns of place and space". Energy Research & Social Science 89, 2022.

[4] COSTA, Evaldo; TEIXEIRA, Ana Carolina Rodrigues; COSTA, Suellen Caroline Silva; CONSONI, Flavia L. "Influence of public policies on the diffusion of wind and solar PV sources in Brazil and the possible effects of COVID-19". Renewable and Sustainable Energy Reviews 162, 2022.

[5] CARSTENS, Danielle Denes dos Santos; CUNHA, Sieglinde Kindl da. "Challenges and opportunities for the growth of solar photovoltaic energy in Brazil". Energy Policy 125, 2019.

[6] LI, Ling-Ling; WEN, Shi-Yu; TSENG, Ming-Lang; WANG, Cheng-Shan. "Renewable energy prediction: A novel short-term prediction model of photovoltaic output power". Journal of Cleaner Production 228, 2019.

[7] ZHONG, Hai; WANG, Jiajun; JIA, Hongjie; MU, Yunfei; LV, Shilei. "Vector field-based support vector regression for building energy consumption prediction". Applied Energy 242, 2019.

[8] XU, Lei; HOU, Lei; ZHU, Zhenyu; LI, Yu; LIU, Jiaquan; LEI, Ting; WU, Xingguang. "Mid-term prediction of electrical energy consumption for crude oil pipelines using a hybrid algorithm of support vector machine and genetic algorithm". Energy 222, 2021.

[9] SOUZA, Valéria Monteiro de; VIEIRA, João Paulo Abreu; BRITO, Hugo Rodrigues de. "Mitigating the impact of high-capacity dispatchable distributed generation on reconfigurable distribution networks with step voltage regulators: A real case study on voltage issues". Electric Power Systems Research 207, 2022.

[10] ALBUQUERQUE, Pedro C.; CAJUEIRO, Daniel O.; ROSSI, Marina D.C. "Machine learning models for forecasting power electricity consumption using a high dimensional dataset". Expert Systems With Applications 187, 2022.

[11] Sujan Reddy A.; Akashdeep S.; Harshvardhan R.; Sowmya Kamath S. "Stacking Deep learning and Machine learning models for short-term energy consumption forecasting". Advanced Engineering Informatics 52, 2022.

[12] WANG, Zeyu; WANG, Yueren; ZENG, Ruochen, SRINIVASAN, Ravi S.; AHRENTZEN, Sherry. "Random forest based hourly building energy prediction". Energy Build 171, 2018.

[13] KIM, Moon Keun; KIM, Yang-Seon; SREBRIC, Jelena. "Predictions of electricity consumption in a campus building using occupant rates and weather elements with sensitivity analysis: Artificial neural network vs. linear regression". Sustainable Cities and Society 62, 2020.

[14] Relação de empreendimentos de Geração Distribuída. Agência Nacional de Energia Elétrica (ANEEL).