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

THIAGO AZEVEDO CAMPOS COSTA

SISTEMA DE REGRAS PARA ACOMPANHAMENTO DE PERFORMANCE EM USINAS FOTOVOLTAICAS EMPREGANDO TÉCNICAS DE APRENDIZAGEM DE MÁQUINA

FORTALEZA
2021

THIAGO AZEVEDO CAMPOS COSTA

SISTEMA DE REGRAS PARA ACOMPANHAMENTO DE PERFORMANCE EM USINAS FOTOVOLTAICAS EMPREGANDO TÉCNICAS DE APRENDIZAGEM DE MÁQUINA

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia Elétrica do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia Elétrica.

Orientador: Prof. Dr. Arthur Plínio de Souza Braga

FORTALEZA

2021

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

C876s Costa, Thiago Azevedo Campos.
Sistema de regras para acompanhamento de performance em usinas fotovoltaicas empregando técnicas de aprendizagem de máquina / Thiago Azevedo Campos Costa. – 2021.
81 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Tecnologia, Curso de Engenharia Elétrica, Fortaleza, 2021.
Orientação: Prof. Dr. Arthur Plínio de Souza Braga.

1. Plantas fotovoltaicas. 2. Aprendizagem de máquina. 3. Monitoramento de performance. I. Título.
CDD 621.3

THIAGO AZEVEDO CAMPOS COSTA

SISTEMA DE REGRAS PARA ACOMPANHAMENTO DE PERFORMANCE EM USINAS FOTOVOLTAICAS EMPREGANDO TÉCNICAS DE APRENDIZAGEM DE MÁQUINA

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia Elétrica do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia Elétrica.

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. Arthur Plínio de Souza Braga (Orientador)
Universidade Federal do Ceará (UFC)

Prof. Dr. Paulo Cesar Marques de Carvalho
Universidade Federal do Ceará (UFC)

Profa. Ma. Tatiane Carolyne Carneiro
Universidade Federal do Maranhão (UFMA)

À minha mãe, Ivonni Azevedo.

Ao meu pai, Juarez Campos.

À minha namorada, Bárbara Thais.

A todos os amigos que fizeram parte dessa trajetória.

AGRADECIMENTOS

Aos meus pais, que estiveram juntos de mim e me apoiaram incondicionalmente em todos os momentos mais importantes da minha vida pessoal e profissional.

À Bárbara Thais Araújo, minha namorada, companheira e amiga, que por muitos anos tem sido meu apoio nos meus principais desafios, e há de continuar sendo por muitos anos a vir.

Ao Prof. Dr. Arthur Plínio de Souza Braga, por me orientar em meu trabalho de conclusão de curso e em outros projetos ao longo da minha graduação.

À Delfos Intelligent Maintenance e a todos meus amigos de trabalho, por viabilizarem esse projeto, pela ótima convivência diária e por me impulsionarem a melhorar cada vez mais enquanto profissional.

A todos os meus amigos de longa data que sempre me apoiaram e estiveram juntos nessa caminhada.

Aos meus amigos da faculdade por todo o apoio, amizade e por todos os momentos vividos no "quadrado".

Ao Movimento Empresa Júnior e à Tecsys Jr., pelas amizades, momentos e aprendizagens únicas, permitindo minhas primeiras experiências profissionais em Engenharia Elétrica e de trabalho em empresa, me agregando experiência e amadurecimento nos âmbitos pessoal e profissional.

Ao Grupo de Automação, Controle e Robótica (GPAR), onde aprendi bastante com meus companheiros e tive meus primeiros contatos com temas que hoje me são bastante pertinentes em minha vida profissional.

Ao Programa Brafitec, por ter me proporcionado fazer novas amizades, conhecer novas culturas e viver uma das maiores experiências da minha graduação e da minha vida.

Aos professores convidados à banca avaliadora, Prof. Dr. Paulo Cesar Marques de Carvalho e Profa. Ma. Tatiane Carolyne Carneiro, pela avaliação e sugestões ao trabalho.

"Se enxerguei mais longe, é porque me apoiei em ombros de gigantes."

(Isaac Newton)

RESUMO

A geração fotovoltaica continua crescendo sua participação nas matrizes energéticas do Brasil e do mundo, justificando o desenvolvimento de tecnologias que permitem uma maior eficiência e melhor monitoramento de performance. Nesse contexto, o presente trabalho objetiva desenvolver um sistema de regras baseado na modelagem de energia esperada para auxiliar no acompanhamento de performance em usinas fotovoltaicas. Técnicas de Aprendizagem de Máquina também são utilizadas para aproximar a geração esperada dos sistemas fotovoltaicos com base em dados de séries temporais históricas de potência elétrica e variáveis ambientais (ex.: irradiância, temperatura ambiente, velocidade do vento e umidade do ar). Com base em dados de uma usina fotovoltaica real localizada no nordeste brasileiro, filtros e técnicas de pré-processamento são aplicados para amplificar o desempenho dos modelos de Aprendizagem de Máquina que foram comparados: Multilayer Perceptron, k-Nearest Neighbors e Random Forest. O desempenho dos modelos de geração obtidos a partir das técnicas de aprendizagem de máquina são comparados a um modelo físico disponível na PVlib, uma biblioteca em Python para simulação de sistemas fotovoltaicos. Os resultados mostram uma acurácia maior dos modelos obtidos a partir de técnicas de aprendizagem, em que o Random Forest apresenta uma Raiz do Erro Médio Quadrático de 57,42 no conjunto de teste. O sistema proposto para identificar ocorrências de performance abaixo do esperado, é descrito em seus detalhes de implementação e apresentado na forma de uma interface gráfica. Um estudo de caso é conduzido para validar a aplicação do sistema de regras, totalizando 667 ocorrências de baixa performance identificadas para 32 inversores em um período de 31 dias.

Palavras-chave: Plantas Fotovoltaicas. Aprendizagem de Máquina. Monitoramento de Performance.

ABSTRACT

Photovoltaic power generation continues growing its share on energy matrices in Brazil and worldwide, justifying the development of new technologies that allow better efficiency and performance monitoring. Given that, this work proposes a performance assessment rule system for photovoltaic plants, based on expected power estimation. Machine Learning Techniques were also applied aiming to approximate a photovoltaic system's expected energy, based on historical time-series of electric power and environmental variables (e.g.: irradiance, ambient temperature, wind speed and air humidity). Using real data collected from a photovoltaic plant located in Brazil's Northeast region, filters and preprocessing techniques are applied to improve the Machine Learning models compared on the task of regression: Multilayer Perceptron, k-Nearest Neighbors and Random Forest. The performance of the energy models obtained by the Machine Learning techniques are compared to a physical model available in PVlib, a Python library for photovoltaic systems simulation. The results show a greater accuracy for the models obtained with the Machine Learning, with Random Forest presenting 57.42 as Root Mean Square Error on the test data. The proposed system for identifying low performance occurrences is described in its implementation details and presented by a graphical user interface. A case study is conducted to validate the proposed system, obtaining 667 low performance occurrences detected for 32 inverters in a 31 days period.

Keywords: Photovoltaic plants. Machine Learning. Performance Monitoring.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

FV	Fotovoltaica
IQR	*Interquartile Range*
kNN	*k-Nearest Neighbors*
MAE	*Mean Absolute Error*
MLP	*Multilayer Perceptron*
MMQ	Método dos Mínimos Quadrados
PR	*Performance Ratio*
R2	Coeficiente de Determinação
RF	*Random Forest*
RMSE	*Root-mean-square Error*
RNA	Redes Neurais Artificiais
SCADA	Supervisory Control and Data Acquisition
SIN	Sistema Integrado Nacional

LISTA DE SÍMBOLOS

E	Energia
ϵ	Erro
P	Potência
$\hat{P}$	Potência Esperada
$\bar{P}$	Potência Média

SUMÁRIO

1 INTRODUÇÃO

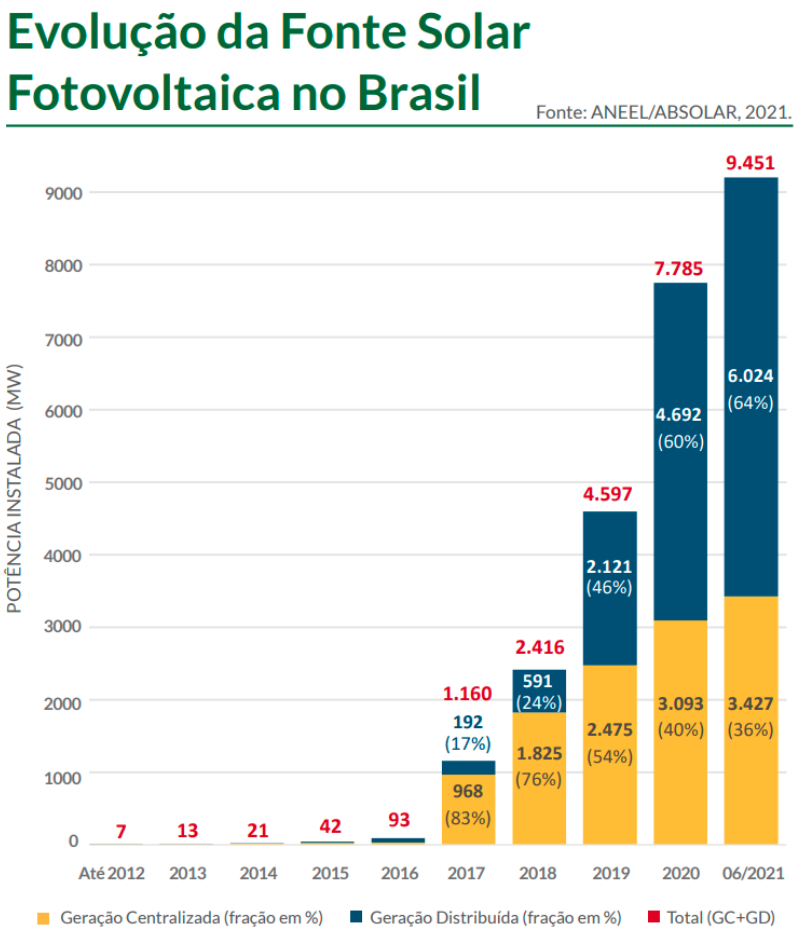
A produção de energia elétrica sempre foi um dos pilares da sociedade moderna, em um cenário de crescentes demandas energéticas para atender às mais diversas necessidades: domiciliares, hospitalares, industriais, entre outras. Os avanços tecnológicos sempre buscaram a otimização desse processo de geração, diminuindo custos, perdas e impactos ambientais, ao ritmo em que a capacidade produtiva aumenta. No contexto contemporâneo, existe uma ascensão de matrizes energéticas renováveis e limpas, visando a substituição de métodos que, outrora, figuraram como as principais fontes energéticas mundiais, tais como a utilização de combustíveis fósseis (IRENA, 2019).

Dentre estas novas tecnologias emergentes, encontra-se a geração de energia elétrica por utilização de painéis capazes de transformar a energia contida na irradiação solar em energia elétrica, a Geração FV. Segundo dados atualizados em 1 de agosto de 2021 pela Associação Brasileira de Energia Solar Fotovoltaica (ABSOLAR, 2021), exibidos pela Figura 1, a porção FV da matriz energética brasileira tem crescido constantemente ao longo dos anos, apresentando um potencial bastante promissor. Este fenômeno deve-se, principalmente, pela fonte renovável e limpa que possibilita o processo, além da praticidade de instalação e adequação ao SIN, abrindo portas para que mesmo o consumidor domiciliar possa possuir sua geração própria.

Dessa forma, grandes usinas de geração FV estão surgindo em território nacional, aumentando, consequentemente, a demanda por um acompanhamento minucioso da performance destas usinas, de forma a estimar um comportamento esperado e compará-lo a observações durante a operação real. Existem normas internacionais (IEC 61724-1, 2017) que definem diversas metodologias para acompanhamento de performance, envolvendo comparações entre valores medidos e esperados idealmente. Este processo visa a diminuição de perdas energéticas, maximizando a energia disponível para o SIN e o retorno financeiro para as empresas responsáveis.

Neste contexto, o capítulo de Introdução busca prover o cenário atual em que o problema proposto se insere, explicitar a motivação que embasa o tema do presente trabalho, com contextualização de operações reais executadas em usinas FV. Desta feita, a motivação resulta no planejamento do trabalho, contendo os objetivos traçados, a metodologia que será aplicada para a solução do problema proposto e a estrutura do trabalho.

Figura 1 – Evolução da geração FV na matriz energética brasileira



Fonte: ABSOLAR (2021)

1.1 Motivação

O acompanhamento de uma usina de geração FV é uma etapa fundamental para o sucesso da operação como um todo. Monitorar de forma objetiva e clara o cotidiano de uma usina é imprescindível para bons gestores, do ponto de vista técnico e financeiro. Desta feita, o uso de ferramentas estatísticas e análise de dados está cada vez mais em evidência como aliados poderosos, em que o estudo dos dados históricos pode embasar o comportamento esperado por uma planta FV em condições adequadas.

Nesta perspectiva, de um ponto de vista gerencial, uma das estimativas mais importantes é a de potência esperada pelo sistema dadas as condições específicas nas quais este opera, e, consequentemente, a definição de um parâmetro que mensure a performance do sistema, em comparação com suas operações históricas. Para atingir este objetivo, gestores podem utilizar modelos suficientemente fiéis à realidade para determinar se seus

sistemas operam de forma satisfatória (IEC 61724-1, 2017).

No entanto, imprecisões nesses modelos adotados podem causar falsas conclusões sobre performance. Quanto maior a precisão um modelo possuir em determinar o comportamento esperado, maior será a confiança na informação recebida por um gestor e maior será o grau de detecção de anomalias na operação, falhas diversas que podem passar despercebidas ao monitoramento mas que afetam diretamente o desempenho da geração. Uma vez que uma performance baixa é precisamente identificada, mesmo sem falhas a uma primeira vista, uma investigação minuciosa pode desencadear-se em torno de resolver o problema e retomar a operação normal o quanto antes, aumentando sua eficiência.

Portanto, o monitoramento de performance é uma das maiores demandas de gestores de usinas, impulsionando novos produtos e novas empresas dedicadas a esta finalidade. Assim, um sistema capaz de realizar esse acompanhamento e apontar sugestões operacionais para contornar situações de baixa performance, da forma mais precisa possível, é um produto de grande valor para estes gestores.

1.2 Objetivos

O objetivo geral do presente trabalho é desenvolver um sistema capaz de realizar o monitoramento de performance e indicar possíveis causas para episódios de desempenho de geração abaixo do esperado, utilizando-se de ferramentas de modelagem, tais como modelos de Aprendizagem de Máquina e outras técnicas estatísticas.

Objetiva-se, portanto, os seguintes resultados:

- Modelos capazes de executar suficientemente bem (de acordo com métricas de validação) a predição de potência elétrica esperada para plantas de geração FV;
- Comparações entre diferentes modelos, de forma a destacar as melhores precisões a partir de métricas pré-definidas;
- Desenvolvimento de um sistema de regras para classificação de ocorrências de baixa performance;
- Análise da viabilidade de uso prático e direto do sistema em uma operação real de geração FV.

1.3 Metodologia

A metodologia aplicada de forma a alcançar os objetivos traçados seguirá de forma a:

- Estudar as características de dados reais de geração FV;
- Aplicar filtros para remover amostras com baixa significância estatística;
- Aplicar tratamentos estatísticos ao conjunto de dados;
- Utilizar modelos físicos e técnicas de Aprendizagem de Máquina para desenvolver diferentes modelos;
- Definir métricas de validação e aplicá-las de forma analítica aos modelos;
- Propor um sistema de monitoramento baseado nas modelagens do sistema.

1.4 Estrutura do trabalho

Portanto, a estrutura do trabalho será da forma:

- Capítulo 1: visão geral do contexto e do problema a ser analisado, objetivos traçados, metodologia proposta e estrutura de organização do trabalho;
- Capítulo 2: revisão bibliográfica de trabalhos relacionados e fundamentação teórica sobre os temas que circundam Sistemas Fotovoltaicos, Modelagem Física e Modelagem por Aprendizagem de Máquina;
- Capítulo 3: detalhamento da metodologia proposta para os objetivos traçados, tais como a coleta, exploração, filtragem e processamento dos dados disponíveis, detalhamento dos modelos expostos e comparação utilizando métricas de validação;
- Capítulo 4: detalhamento do sistema de regras proposto para o monitoramento de performance, incluindo as regras utilizadas e a proposta de interface gráfica para auxiliar sua utilização;
- Capítulo 5: exposição dos resultados obtidos pela comparação entre os modelos analisados e dos resultados oriundos da aplicação do sistema de regras, exibindo um estudo de caso;
- Capítulo 6: conclusões obtidas pelo estudo e análise do cumprimento dos objetivos traçados e sugestões para trabalhos futuros que possam dar continuidade à linha de pesquisa explorada no estudo.

2 FUNDAMENTAÇÃO E DEFINIÇÕES

O capítulo de Fundamentação e Definições busca detalhar todo o embasamento necessário para a compreensão dos capítulos de Metodologia e Resultados. Dessa forma, será realizada uma Revisão Bibliográfica, contextualizando o presente trabalho dentre outros estudos similares, destacando convergências e inspirações entre eles. Além disso, será explicada toda a base teórica para tratar do tema proposto, desde todas as características de um sistema de geração FV e modelagem de sistemas, com seus detalhes, embasamento e funcionamento.

2.1 Revisão Bibliográfica

A aplicação de Aprendizagem de Máquina e técnicas estatísticas em problemas relacionados a geração FV é um campo cada vez mais em evidência, dado o avanço das tecnologias envolvendo a operação.

Neste contexto, Kaaya and Ascencio-Vásquez (2021) fazem uma comparação ampla entre diferentes métodos para modelagem de sistemas FV, entre eles modelos físicos, modelos heurísticos e modelos estatísticos. Dentre as conclusões obtidas, a comparação mostrou que os modelos físicos possuem o maior grau de incerteza ao prever a geração do sistema. Embora o foco da referência seja a predição futura de geração (o *Forecasting*), a análise comparativa entre diferentes tipos de modelos também foi objeto de estudo para o trabalho desenvolvido, ressaltando a importância de modelos precisos para suprir uma demanda cada vez maior por modelos de performance confiáveis para geração FV.

Ademais, Al-Dahidi et al. (2019) fazem um estudo específico para aplicação de uma Rede Neural Artificial para a modelagem de um sistema FV real. No estudo, são comparados diferentes algoritmos de aprendizagem aplicados à rede neural, além de diferentes combinações de variáveis ambientais para determinar quais se adequam melhor à modelagem. Desse modo, o uso de Redes Neurais Artificiais (RNA) para este fim é cada vez mais validado, abrindo portas, também, para outros modelos de Aprendizagem de Máquina.

Seguindo na mesma temática, Simal Pérez and Batlles (2021) procuram estimar as perdas por sujidade (*Soiling Losses*) presentes em uma planta FV, esta que é uma das causas mais estudadas para perda de performance em plantas reais. O estudo é voltado

especificamente para a sujidade, mas assemelha-se aos demais por também propor uma modelagem utilizando RNA e variáveis de ambiente. No entanto, a modelagem foca em determinar as perdas advindas desse problema bastante pertinente dentre as causas mais comuns de perda de performance em usinas FV.

Rodrigues et al. (2018) propuseram um sistema para análise de performance que procura generalizar para diversas localizações e questões sazonais em que a planta FV esteja inserida. Desse modo, 5 técnicas diferentes de Aprendizagem de Máquina foram comparadas em diferentes plantas, que estão localizadas em diferentes localizações e diferentes condições ambientais. O foco do estudo é na generalização dos modelos, comparando as diferentes técnicas quanto à capacidade de se adaptar a diferentes sistemas e conseguir realizar a predição de performance proposta.

Por fim, Al-Dahidi et al. (2020) buscam experimentar o uso de RNA para modelar a potência de sistemas FV, mas treinadas para horários específicos do dia. Dessa forma, cada modelo será responsável por atuar em cada uma das 24 horas do dia. O estudo utiliza uma planta real para aquisição dos dados e busca mostrar que a estratégia de treino local pode ser mais eficiente com relação a tempo e recursos computacionais, quando comparados a uma estratégia composta por um modelo único.

Portanto, esses trabalhos ajudaram a embasar a literatura sobre aplicação de Aprendizagem de Máquina em problemas relacionados a plantas FV. Por mais que alguns trabalhos foquem não somente na modelagem, mas sim na predição de valores para tempos futuros, eles puderam contribuir com abordagens válidas para compreender sistemas FV e suas características.

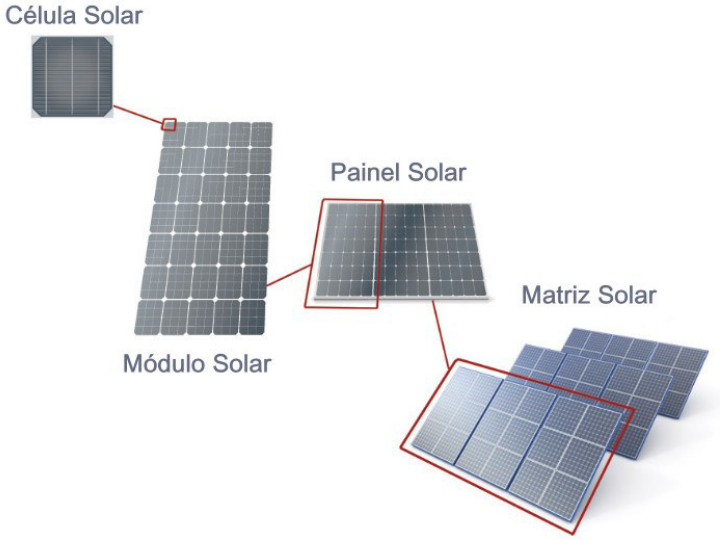
2.2 Sistemas Fotovoltaicos

Usinas de geração FV são sistemas subdivididos em diversos componentes essenciais para a operação, que possui como produto a produção de energia elétrica a partir da irradiação solar.

O principal dispositivo responsável pela conversão da energia irradiada pelo Sol em energia elétrica é o painel FV, coração da operação, que é composto por módulos solares. Estes são compostos por partições menores, as chamadas células solares, que possuem propriedades físicas propícias à ocorrência do Efeito Fotovoltaico (CRESESB, 2014), que é a criação de tensão elétrica a partir da luz. A Figura 2 demonstra essa

subdivisão fundamental desde um conjunto de painéis FV, até a menor unidade, a Célula Solar.

Figura 2 – Configuração de Células Solares



Fonte: Reis (2015)

O Efeito Fotovoltaico é o fator chave para a geração de energia elétrica a partir da luz emitida pelo Sol, mas existem muitas outras etapas necessárias para um bom funcionamento da operação. Nesta seção, serão apresentados os principais elementos que compõem o sistema completo, tal como grandezas físicas associadas e conceitos intrínsecos à operação. Como principal referência para a seção, foi utilizada uma norma internacional (IEC 61724-1, 2017) que trata especificamente do monitoramento de performance para Sistemas FV, elencando os principais dispositivos e grandezas presentes na operação.

2.2.1 Grandezas Ambientais

Juntamente com as grandezas elétricas, existem também as grandezas ambientais, determinantes no resultado final da produção FV.

2.2.1.1 Irradiância Solar

Irradiância Solar (IEC 61724-1, 2017) representa o fluxo radiante instantâneo recebido por unidade de área, medida na unidade W/m^2. Dependendo da forma como é medida, esta grandeza pode apresentar variações. As mais comuns são:

- Irradiância em Plano Inclinado: medida com um sensor em paralelo ao plano dos painéis FV. Em caso do uso de Rastreadores, o sensor deve manter-se sempre alinhado ao plano;
- Irradiância Horizontal Global: medida com um sensor orientado horizontalmente.

Como a medição de irradiância em plano inclinado é, essencialmente, uma medida mais precisa da luz que efetivamente chega às células solares, esta é principal medida de irradiância utilizada durante o estudo.

2.2.1.2 Temperatura Ambiente

Temperatura ambiente (IEC 61724-1, 2017) medida no local da operação, em $^{\circ}C$. Essa medição é realizada por sensores que buscam representar, da melhor forma possível, as condições as quais os módulos FV estão estão expostos.

2.2.1.3 Velocidade do Vento

Velocidade do Vento (IEC 61724-1, 2017), medida em m/s, observada no ambiente da usina FV. Esta velocidade do vento relaciona-se, geralmente, à temperatura dos módulos FV, estando relacionada, portanto, à eficiência destes. Sua medição deve ser realizada em uma localização e altura que represente as condições as quais os módulos estão expostos.

2.2.1.4 Umidade do Ar

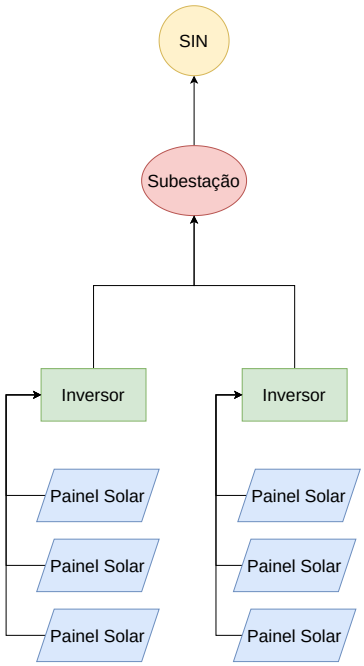
Umidade do Ar (IEC 61724-1, 2017) descreve a quantidade de água que está presente por volume de ar, em forma de vapor, medida em porcentagem. Esta grandeza é mais uma medição comum a ser realizada em plantas FV, contribuindo, também, com a compreensão das características do sistema.

2.2.2 Topologia e Dispositivos

Existem diversas topologias possíveis para a construção de uma usina FV, permitindo adaptações para as necessidades específicas de cada projeto. No entanto, existem dispositivos presentes na maioria (ou totalidade) das usinas. A Figura 3 demonstra

uma estrutura simplificada que poderia ser aplicada em um sistema real.

Figura 3 – Exemplo de topologia simplificada para um sistema FV, desde os painéis fotovoltaicos até a entrega ao SIN



Fonte: Autor

Como é possível observar, existe uma cadeia de dispositivos envolvidos no processo de geração de energia elétrica, a partir dos painéis solares até a ponta final, que seria representada pela Subestação do sistema.

Ademais, detalha-se cada um dos dispositivos presentes no sistema e as suas funções individuais.

2.2.2.1 Painel Solar

Como exemplificado pela Figura 2, o Painel Solar (CRESESB, 2014) é o dispositivo responsável por realizar a conversão entre energia da radiação solar incidente e energia elétrica, por meio do Efeito Fotovoltaico.

Esse fenômeno ocorre na menor unidade presente no painel, que é a célula FV, muitas vezes composta pelo silício, material semicondutor. As células são mecanicamente agrupadas, formando os módulos, que compõem o painel FV.

A potência gerada nesse processo possui um caráter de Corrente Contínua, similar ao de baterias. Essa característica pode ser útil para determinadas aplicações,

como por exemplo para uso em carros elétricos ou para sistemas abastecidos por baterias. No entanto, para produção de energia destinada a grandes sistemas de potência (como o caso do SIN), é necessário que haja uma conversão para Corrente Alternada, de forma a compatibilizar essa potência com o resto do sistema.

2.2.2.2 Rastreador

Em muitas plantas FV, existem os rastreadores (CRESESB, 2014), estruturas de sustentação para conjuntos de painéis solares que ajustam sua angulação no intuito de maximizar a incidência solar no plano dos painéis ao longo do dia. Esses dispositivos geralmente atuam de forma automática, possuindo um controlador próprio para ajuste de sua angulação.

2.2.2.3 Inversor

O inversor (CRESESB, 2014) é um dispositivo eletrônico responsável por realizar a conversão de uma potência de entrada em Corrente Contínua para uma potência de saída em Corrente Alternada. Existem no mercado diversos fabricantes produzindo o dispositivo, em diferentes níveis de potência e tensão suportados, geralmente possuindo interfaces para configurações.

Além da aplicação em sistemas FV, existem outras aplicações para inversores, como, por exemplo, no acionamento e controle de motores elétricos industriais.

2.2.2.4 Piranômetro

Responsável por medir a irradiância solar, os piranômetros (CRESESB, 2014) são dispositivos essenciais para o monitoramento de sistemas FV. Eles possuem a capacidade de mensurar a irradiância global horizontal de um ambiente ou mesmo a irradiância incidente em um superfície plana, como a superfície de um painel solar.

2.2.3 Operação

Além da estrutura, dispositivos e grandezas físicas associadas ao sistema FV como um todo, existem também fatores relacionados à operação cotidiana da usina, aspectos que influenciam diretamente na performance geral.

2.2.3.1 Sistema SCADA

O Sistema SCADA (do inglês, *Supervisory Control and Data Acquisition*) é um sistema centralizado para acompanhamento de diversas informações e dados em tempo real sobre uma planta específica (Khatri, 2018).

Pelo sistema, é possível acompanhar a produção em tempo real, dados históricos, alarmes e todas as informações disponíveis sobre uma usina. Desta forma, são necessários sensores de medição para alimentar o sistema, realizando todas as medições necessárias. Esses dados geralmente são armazenados em estruturas centralizadoras de fácil acesso, como por exemplo bancos de dados hospedados em servidores.

Ademais, o sistema pode ser configurado para emitir alarmes sempre que medições específicas apresentarem valores incoerentes ou indicarem falhas em dispositivos da operação, podendo levar a interrupção do funcionamento de parte ou da totalidade de uma usina.

Desta feita, um acompanhamento operacional eficiente sempre começa pelo estabelecimento de um sistema SCADA, presente em diversas aplicações industriais, e, sobretudo, usinas de produção energética das mais variadas.

2.2.3.2 Indisponibilidades Técnicas

Como demonstrado na seção anterior, o próprio sistema SCADA emite alertas sobre mau funcionamento de dispositivos da operação, configurando indisponibilidades técnicas, ou interrupções momentâneas de parte ou toda capacidade produtiva.

Estas indisponibilidades, na realidade de um sistema FV, podem ser causadas por diversos fatores, entre eles:

- Falha no funcionamento interno de inversores;
- Sobretensões em pontos do circuito interno da usina;
- Sobretemperatura/Sobrecarga de dispositivos;
- Acionamento de dispositivos de proteção.

Estes eventos são recorrentes em operações complexas, sendo bastante comum que ocorram diversas vezes por dia. Um dos grandes desafios da gestão de usinas FV é reduzir esse número de ocorrências ao mínimo possível, o que significaria um maior tempo de operação em condições normais, resultando em uma maior eficiência e, assim, maiores

níveis produtivos.

2.2.3.3 Perda de performance

Apesar do monitoramento exercido pelo sistema SCADA, existem problemas típicos de uma operação FV que possuem uma maior dificuldade de monitoramento via sensores e alarmes. Alguns desses cenários problemáticos podem afetar diretamente a performance da planta sem que haja uma pronta identificação do problema. A citar alguns problemas mais comuns:

- Sombreamento momentâneo dos painéis FV, impedindo que a irradiação solar atinja as células solares;
- Sujeira, chuva, neve e outras intempéries sob os painéis, reduzindo a área de exposição à irradiação solar;
- Falhas em componentes de circuitos anteriores ao inversor, por exemplo:
  - Falhas quanto ao funcionamento dos painéis FV;
  - Falhas quanto ao funcionamento de combinadores de painéis;
  - Falhas quanto ao funcionamento dos rastreadores.

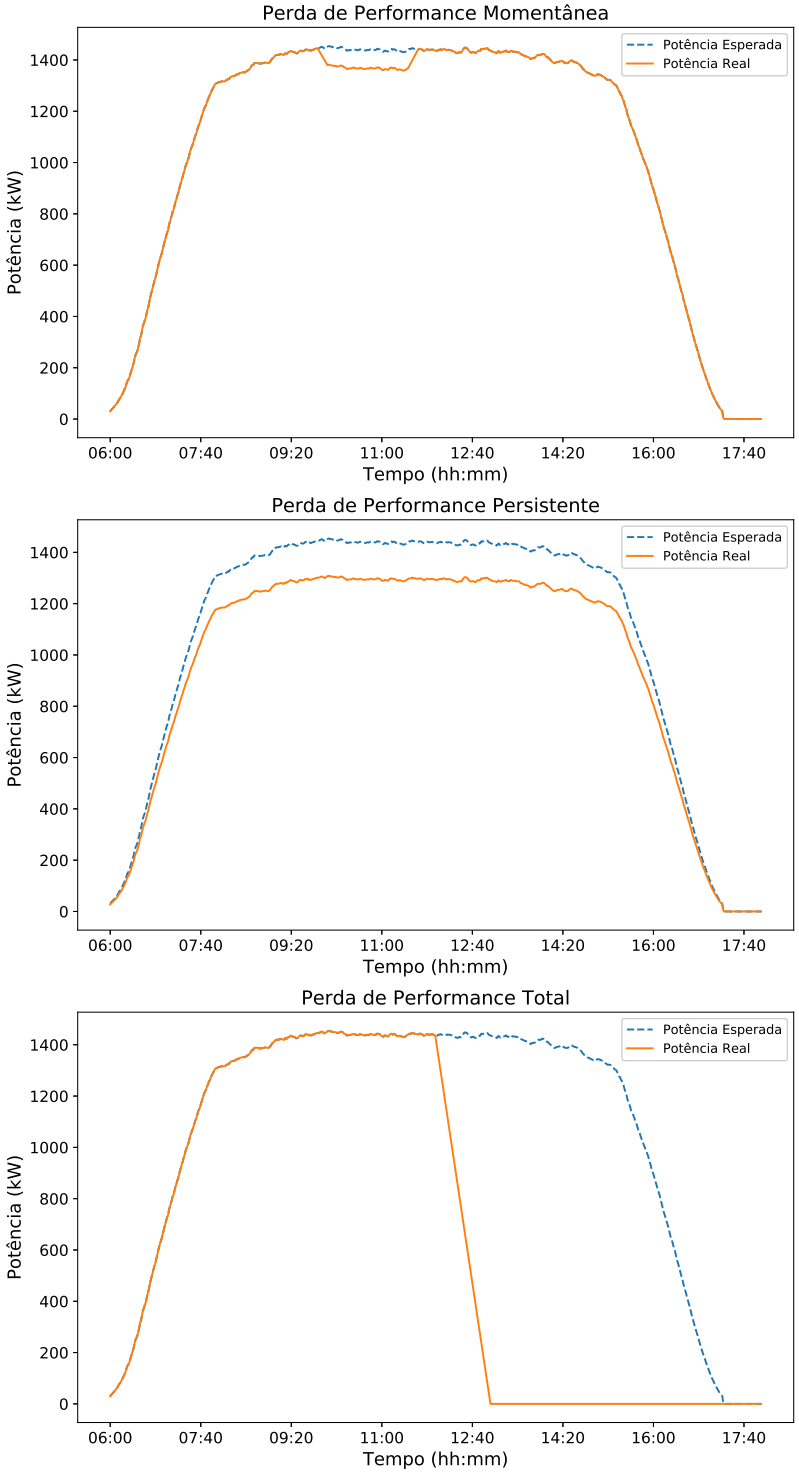
Desse modo, o acompanhamento da operação precisa ir além de monitorar apenas alarmes do SCADA, é preciso identificar e compreender situações mais complexas que, mesmo de difícil detecção, afetam a operação.

Essas ocorrências podem ser dividas em três tipos diferentes, baseado nas suas intensidades e durações:

- Perda de performance momentânea: um episódio de desvio leve ou moderado de performance que logo retoma os patamares normais da operação;
- Perda de performance dissolvida: um episódio de desvio leve ou moderado de performance que persiste mesmo após longo período de tempo;
- Perda de performance total: um episódio de desvio pesado de performance, representando uma falha relevante no sistema, muitas vezes gerando alarmes no SCADA.

Essa classificação estipulada pode auxiliar na identificação da ocorrência responsável por afetar o desempenho da planta e na tomada de medidas competentes para sanar o problema. A Figura 4 exemplifica os tipos característicos de perda de performance.

Figura 4 – Exemplos meramente ilustrativos para diferentes tipos de perdas de performance



Fonte: Autor

2.2.3.4 Parâmetros de performance

Existem parâmetros usualmente utilizados para mensurar, de forma objetiva e direta, a performance de plantas FV. Essencialmente, a performance está associada a capacidade produtiva observada em comparação a capacidade esperada.

Desse modo, é possível definir uma métrica (bastante utilizada no cotidiano operacional) chamada *Performance Ratio* (PR), que, de forma simples, compara a energia real produzida E_{real} e a energia teórica esperada $E_{esperada}$ para um mesmo período (IEC 61724-1, 2017).

$$PR = \frac{E_{real}}{E_{esperada}} \tag{2.1}$$

Através dessa métrica, gestores traçam metas para considerar que o sistema está operando da maneira esperada e de forma satisfatória. No entanto, a estimação de energia esperada não é tão simples e direta, podendo ser alcançada por diferentes métodos e com diferentes precisões. Quanto maior a precisão da estimação com relação ao sistema real, maior será a precisão do indicador e melhor será o monitoramento da performance por parte dos gestores.

2.3 Fundamentos de Aprendizagem Estatística

Esta seção será endereçada a desenvolver conceitos relacionados à Aprendizagem Estatística, bases fundamentais para a compreensão do trabalho proposto. Serão cobertos desde conceitos fundamentais de Estatística a conceitos de Aprendizagem de Máquina e técnicas mais apuradas.

Como principal referência no assunto, foi utilizado o livro *The Elements of Statistical Learning* (Hastie et al., 2001), uma das referências mais consolidadas na literatura de Aprendizagem de Máquina.

2.3.1 Correlação Linear

A Correlação entre diferentes conjuntos de dados é uma medida extremamente relevante para a exploração de dados em geral. Sejam dois conjuntos independentes de dados $X \in \mathbb{R}^n$ e $Y \in \mathbb{R}^n$, em que x_i e y_i sejam a i-ésima amostra de X e Y, respectivamente,

é possível observar medidas de correlação entre os conjuntos. Em outras palavras, a correlação observa o quanto esses conjuntos possuem características em comum e, portanto, variam de forma similar.

Uma das maneiras de determinar essa correlação, é calculando os Coeficientes de Pearson (Shalev-Shwartz and Ben-David, 2014), que são indicadores de correlação linear calculados pela covariância entre os conjuntos dividida pelo produto de seus desvios-padrão.

$$c_{Pearson}(X,Y) = \frac{cov(X,Y)}{\sigma_X \cdot \sigma_Y} \tag{2.2}$$

Em que σ_X e σ_Y são os desvios-padrão de X e Y, respectivamente. Em que a covariância entre dois conjuntos de dados $cov(X,Y)$ é calculada da forma:

$$cov(X,Y) = \frac{\sum_{i=1}^{N}(X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{N-1} \tag{2.3}$$

Em que $\bar{X}$ e $\bar{Y}$ são as médias de X e Y, respectivamente. Desta feita, as correlações calculadas são normalizadas entre -1 e 1, em que -1 indica uma máxima correlação inversa e 1 indica uma máxima correlação direta.

2.3.2 Amplitude Interquartil (IQR)

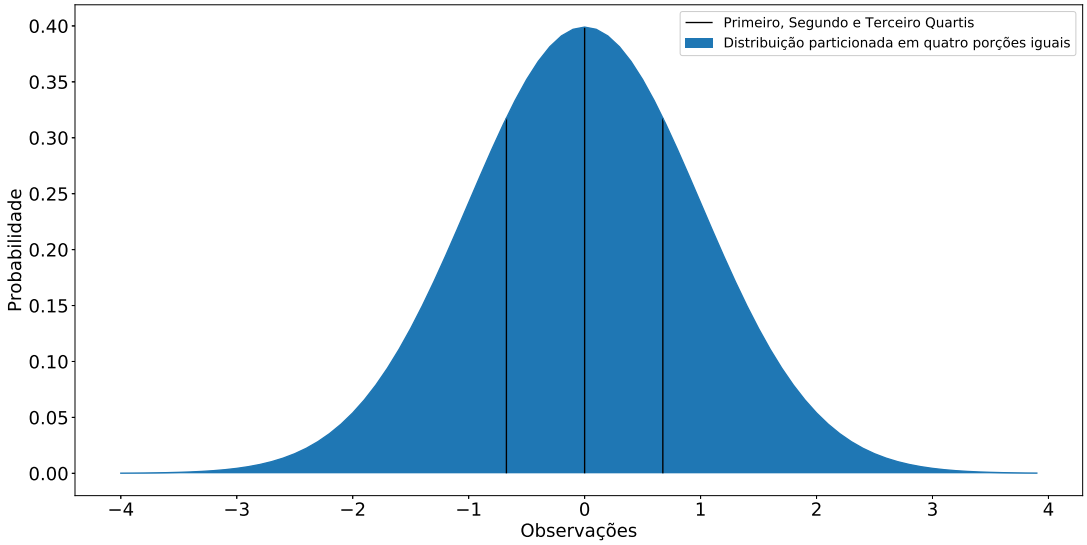
Os quartis são medidas separatrizes que buscam dividir os dados em quatro parcelas contendo quantidades iguais de amostras.

Dessa forma, o primeiro quartil Q_1 divide os dados em uma proporção (a esquerda e a direita, respectivamente) de 25% e 75%, o segundo quartil Q_2 divide os dados em uma proporção de 50% e 50% (coincidindo com a mediana do conjunto) e o terceiro quartil Q_3 divide os dados em uma proporção de 75% e 25%. Ao fim, existem quatro porções iguais, como representado pela Figura 5.

A Amplitude Interquartil (*Interquartile Range* (IQR)) é uma medida de dispersão que utiliza os quartis. A amplitude é calculada como a diferença entre o Terceiro Quartil e o Primeiro Quartil, sendo, portanto, o comprimento da metade mais central dos dados (Han et al., 2011).

É uma medida bastante utilizada para determinar *Outliers* (pontos discrepantes com relação aos demais), definindo uma relação entre os quartis e a amplitude interquartil

Figura 5 – Quartis separando uma Distribuição Normal em quatro porções



Fonte: Autor

para evidenciar esses pontos destoantes. Em geral, admite-se que, dado um conjunto $X \in \mathbb{R}^n$:

$$Outliers = \{x \in X : x < Q_1 - 1.5 \cdot IQR \text{ } ou \text{ } x > Q_3 + 1.5 \cdot IQR\} \tag{2.4}$$

Em que Q_1 é o primeiro quartil de X, Q_3 é o terceiro quartil de X e IQR é sua Amplitude Interquartil.

2.3.3 Regressão

Dentro do domínio da Aprendizagem de Máquina, uma das principais tarefas desempenhadas é a da Regressão, que objetiva a modelagem, por meio de um modelo estatístico e observações históricas, de uma função desconhecida.

Dessa forma, é preciso aproximar uma função f da forma:

$$f : \mathbb{R}^n \to \mathbb{R} \tag{2.5}$$

Sem explicitamente defini-la, uma vez que ela é desconhecida e, muitas vezes, impossível de ser determinada analiticamente.

Serão tratados de diversos modelos de regressão na Seção 2.4.2, modelos estes que serão utilizados para o cerne do trabalho proposto, que é aproximar a produção esperada de inversores em usinas fotovoltaicas.

2.3.3.1 Regressão Linear

O modelo de Regressão Linear é o exemplo mais consolidado e simples para regressão. Trata-se de estimar uma função f como sendo linear, com parâmetros definidos pelo conjunto de dados à disposição. Assim, assume-se que esse relacionamento entre a variável dependente e a variável independente (a ser estimada) possui um caráter linear, com a adição de um erro ϵ intrínseco.

Então, sejam os conjuntos de dados $X \in \mathbb{R}^n$ e $Y \in \mathbb{R}^n$, em que $x_i \in \mathbb{R}$ e $y_i \in \mathbb{R}$ são, respectivamente, o i-ésimo elemento de X e Y, o modelo de Regressão Linear assume que:

$$y_i = \theta_0 + \theta_1 \cdot x_i + \epsilon_i \tag{2.6}$$

Em que y_i é o valor que precisa ser estimado pela regressão, θ_0 e θ_1 são parâmetros a ser calculados e ϵ_i é um ruído, adicionando incerteza ao modelo linear com relação às observações que estão sendo modeladas.

Um exemplo de Regressão Linear, para o caso específico unidimensional, é representado pela Figura 6

De forma generalizada, considerando agora os conjuntos de dados $X \in \mathbb{R}^{n \times p}$ e $Y \in \mathbb{R}^n$, em que $x_i \in \mathbb{R}^p$ e $y_i \in \mathbb{R}$, com $i \in \{1, ..., n\}$. O modelo de Regressão Linear, assume a forma generalizada:

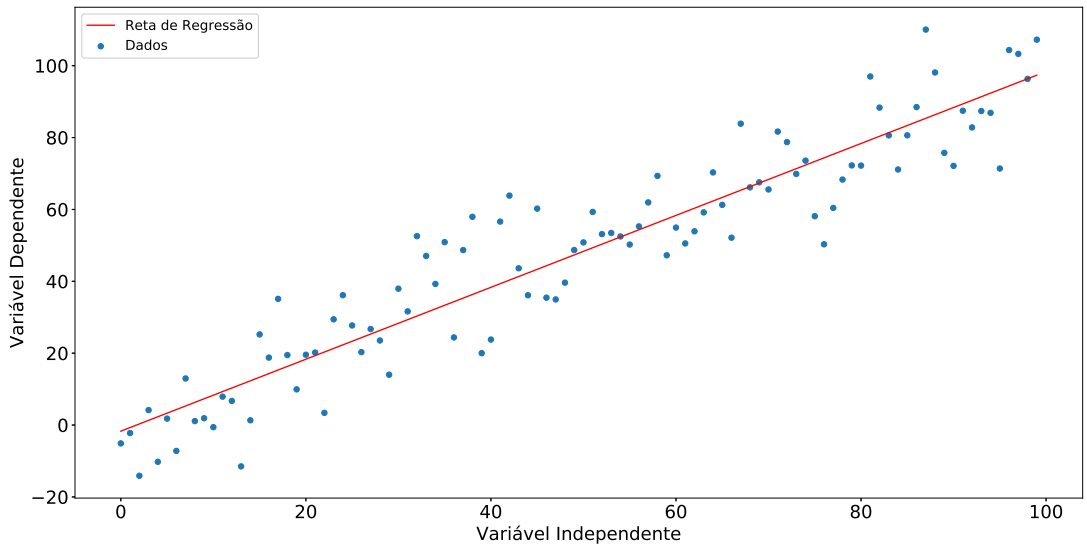
$$y_i = \theta_0 + \theta \cdot x_i + \epsilon_i \tag{2.7}$$

Em que $\theta_0 \in \mathbb{R}$, $\theta \in \mathbb{R}^p$ e $\epsilon_i \in \mathbb{R}$.

Uma das formas possíveis de definir os parâmetros da Regressão Linear, é utilizando o Método dos Mínimos Quadrados (MMQ), que consiste em minimizar o erro ϵ associado ao problema (Hastie et al., 2001). Esse erro, pode ser explicitado da forma:

$$\epsilon_i = y_i - (\theta_0 + \theta \cdot x_i) \tag{2.8}$$

Figura 6 – Exemplo de Regressão Linear



Fonte: Autor

Então, o problema de otimização, em sua forma matricial, é definido pela Equação 2.9.

$$\min_{\theta,\theta_0} ||Y - \theta_0 - \theta \cdot X||_2^2 \tag{2.9}$$

Em que a Norma Euclidiana de um $x \in \mathbb{R}^n$ é definida como:

$$||x||_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2} \tag{2.10}$$

Como X é uma matriz no $\mathbb{R}^{n \times p}$ podemos observá-la, também, como da forma:

$$X = [x_1, x_2, ..., x_p] \tag{2.11}$$

Em que cada $x_i \in \mathbb{R}^n$.

Considerando, agora, um $\hat{X}$ como sendo a matriz X concatenada por uma coluna unitária, ou seja:

$$\hat{X} = [1, x_1, x_2, ..., x_p] \tag{2.12}$$

Em que 1 representa o vetor unitário em $\mathbb{R}^n$. Considera-se, também, um $\hat{\theta} \in \mathbb{R}^{p+1}$ como sendo o θ concatenado com θ_0, da forma:

$$\hat{\theta} = [\theta_0, \theta] \tag{2.13}$$

O problema de otimização pode ser rearranjado da forma representada pela Equação 2.14.

$$\min_{\hat{\theta}} ||Y - \hat{X} \cdot \hat{\theta}||_2^2 \tag{2.14}$$

A solução para este problema de otimização resulta nos coeficientes θ e θ_0 adequados para o conjunto de dados observado. Os coeficientes de regresão, segundo o Método dos Mínimos Quadrados, são obtidos pela Equação 2.15, representando sua forma matricial.

$$\hat{\theta} = (\hat{X}^T \hat{X})^{-1} \hat{X}^T Y \tag{2.15}$$

Existem ainda outros métodos de obtenção dos coeficientes para o problema de Regressão Linear. A citar, outros métodos bastante vistos na literatura são os métodos *Ridge* e *Lasso*.

2.4 Modelagem

Para o desenvolvimento do sistema de acompanhamento de performance proposto, a modelagem da produção esperada de inversores em usinas FV se faz necessária. Nesta seção, serão cobertos os princípios básicos de funcionamento para as diferentes técnicas de modelagem implementadas e comparadas. Estas técnicas foram separadas em duas categorias: Modelagem Física e Aprendizagem de Máquina. A primeira busca representar um sistema real utilizando-se de equações físicas pré-estabelecidas para aproximar as características do objeto de estudo, enquanto a segunda baseia-se em utilizar o comportamento histórico do sistema para obter uma função de regressão adequada, utilizando diferentes técnicas e algoritmos do domínio conhecido como Aprendizagem de Máquina.

2.4.1 Modelagem Física

Quando trata-se de modelagem física para sistemas FV, existem *softwares* bastante consolidados na indústria, sendo utilizados cada vez mais para modelar os sistemas físicos e realizar projeções de performance para o futuro.

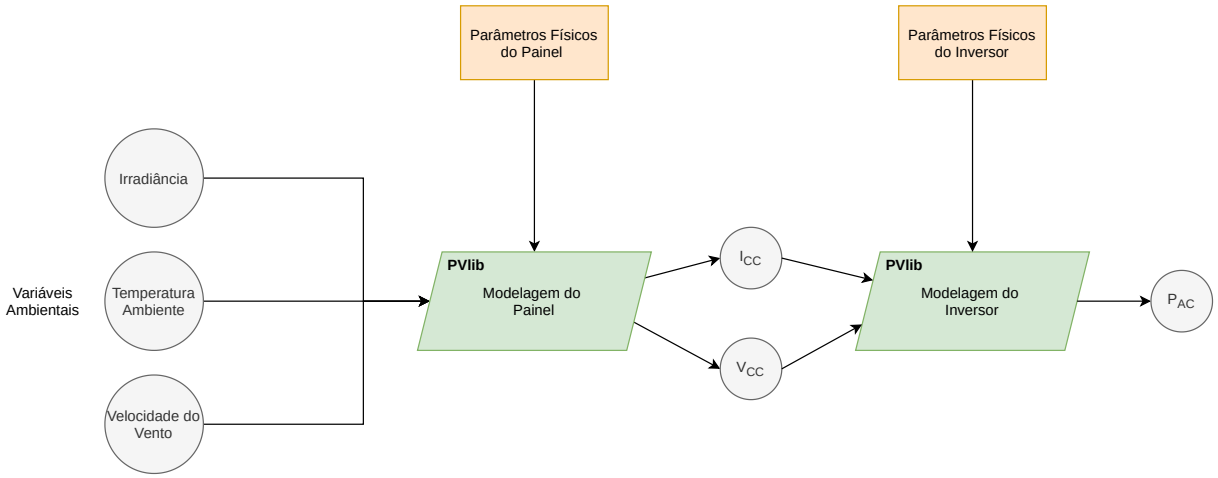
Entre os *softwares* mais utilizados para a função, encontra-se o PVsyst, uma das principais referências quando o tema é modelagem de sistemas FV, amplamente utilizado na indústria. No entanto, existem outras alternativas que apresentam um desempenho bastante satisfatório, como é estudado por Gurupira and Rix (2017), que comparam três dos principais programas de simulação e modelagem disponíveis para a indústria FV.

Dentre estas opções principais, optou-se pela utilização da PVlib (Holmgren et al., 2018), tendo vista que é um pacote disponível em Python e MATLAB de fácil utilização e possui um desempenho bom quando comparado às outras grandes opções da indústria.

O funcionamento do PVlib baseia-se na modelagem dos painéis utilizando equacionamento físico e na utilização de parâmetros para os dispositivos envolvidos (tais como rastreadores, painéis e inversores), como resistências internas e valores de referência. Esses parâmetros podem ser encontrados para diversos modelos e fabricantes diferentes, e são disponibilizados por um banco de dados mantido pela *Sandia National Laboratories*, responsável pelo desenvolvimento da ferramenta. Após a adequação da planta no *software* de acordo com as condições reais, é possível ter acesso a gráficos e valores de performance esperados para o sistema real. A Figura 7 representa uma utilização simplificada do *software*, utilizando as medições de variáveis de ambiente, os parâmetros dos dispositivos especificados pelo fabricante e a modelagem, que pode ser encontrada com mais detalhes nas referências da biblioteca.

Existe uma extensa documentação disponibilizada para auxiliar quanto ao uso das funções, classes e estruturas pré-prontas disponibilizadas, de forma a tornar mais fácil a experiência do usuário e suprir a falta de uma interface gráfica, exigindo que o utilizador possua alguma experiência com programação.

Figura 7 – Exemplo simplificado de utilização do *software* PVlib



Fonte: Autor

2.4.2 Aprendizagem de Máquina

Nesta seção, serão descritos em detalhes os algoritmos que serão objetos de comparação, e pertencem ao campo de estudo da Aprendizagem de Máquina, este que possui suas fundações no campo do Aprendizagem Estatística, e que nos últimos anos tem recebido uma grande relevância, visto que o alto poder computacional de computadores modernos permite diversas aplicações nunca antes vistas, que estão cada vez mais presentes no cotidiano comum.

Os modelos em questão possuem suas próprias particularidades (que serão detalhadas a seguir), mas em geral possuem um ajuste de parâmetros de acordo com dados históricos, que são expostos ao modelo de maneira sistemática durante uma fase chamada de treino, para que este adapte-se da melhor maneira ao sistema a ser modelado.

Além dos parâmetros de modelo (por vezes chamados de pesos) que são definidos durante a fase de treino e estão diretamente relacionados ao resultado final, existem também os chamados hiperparâmetros, que são definidos de antemão ao treino e configuram características pontuais intrínsecas a cada modelo. Para cada um dos modelos comparados, serão explicitados os principais hiperparâmetros, juntamente com uma visão geral do seu funcionamento e treino.

Foram escolhidos três algoritmos de Aprendizagem de Máquina para o desenvolvimento do estudo. O *Multilayer Perceptron* (MLP) é um algoritmo bastante consolidado, sendo utilizado na solução de diversos problemas envolvendo inteligência

artificial. Já o *k-Nearest Neighbors* (kNN), também é um algoritmo bastante consolidado na literatura, mas possui uma estrutura mais simples e de fácil interpretação, sendo também bastante utilizado. Por fim, o *Random Forest* (RF) é um algoritmo pertencente a uma categoria de Comitê de Máquinas (*Ensemble Learning*), em que são utilizados diversos modelos para determinar a resposta final. Todos estes algoritmos são, portanto, bons candidatos para atingir o objetivo de regressão.

2.4.2.1 Multilayer Perceptron Network (MLP)

Redes Neurais Artificiais (Hastie et al., 2001) são modelos que tornaram-se bastante enfatizados pelos resultados práticos que apresentam, principalmente no subdomínio da Aprendizagem Profunda (do inglês, *Deep Learning*), onde é empregada para o uso em reconhecimento de imagens, sons e entre outras representações de conhecimento até então inéditas para máquinas.

Desse modo, a popularização das RNA popularizou, também, diversas novas aplicações focadas na tarefa de classificação, muito comum em Aprendizagem de Máquina. No entanto, esta também pode ser utilizada com o intuito de realizar uma regressão, esta que será estudada em mais detalhes a seguir.

Dentre as várias topologias de RNA existentes, uma das mais comuns é a MLP, que como o nome sugere, é uma rede de várias camadas de Perceptrons, estes que são as menores unidades do modelo.

Como exemplifica a Figura 8, um Perceptron é uma combinação linear entre parâmetros (chamados de pesos) e a aplicação de uma Função de Ativação $f_{ativação}$.

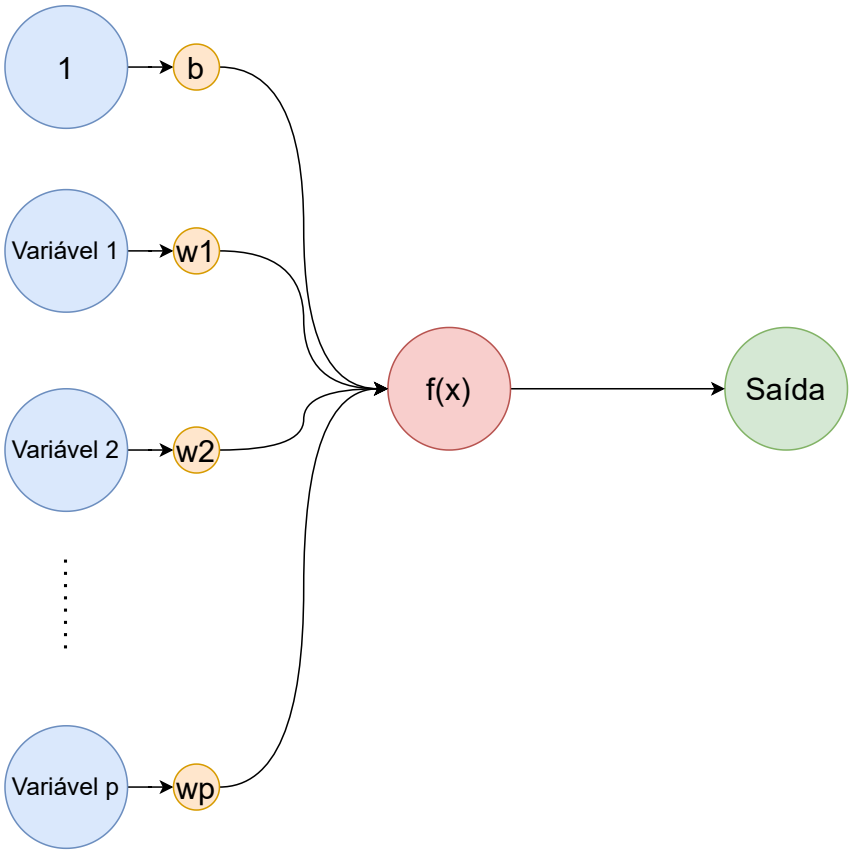
$$\hat{y}_{perceptron} = f_{ativação}(\sum_{i=1}^{n} x_i \cdot w_i + b) \tag{2.16}$$

O valor b, por vezes chamado de *bias*, pode ser considerado como um w_0, um peso independente à entrada do neurônio para contribuir com a combinação linear que será parâmetro para a função de ativação.

A função de ativação, por sua vez, pode ser arbitrariamente escolhida para aplicações específicas. Uma das funções mais utilizadas é a função Sigmoide.

$$Sig(x) = \frac{1}{1+e^{-x}} \tag{2.17}$$

Figura 8 – Exemplo de Perceptron, explicitando as entradas, pesos, bias, função de ativação e saída



Fonte: Autor

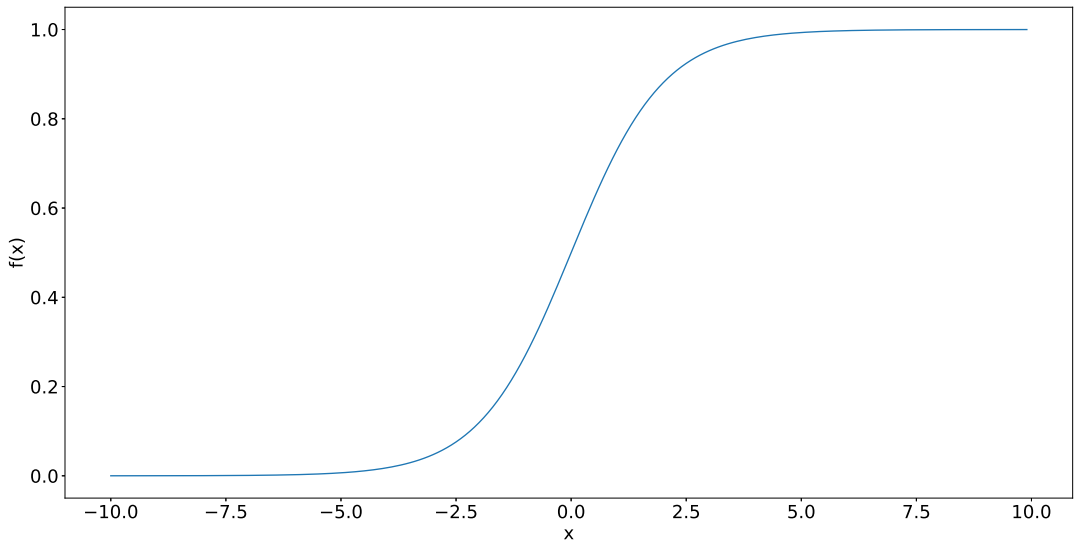
O comportamento da Sigmoide é representado pela Figura 9.

Como é possível observar, a função possui limites entre 0 e 1, saturando nestes valores a depender da sua entrada e determinando, portanto, se estes neurônios serão ativados ou não, transmitindo informação para o resto da rede ou não. Dada a natureza das combinações lineares e das funções de ativação utilizadas, faz-se necessária uma normalização das variáveis de entrada, para que estas estejam sempre em níveis de grandeza similares. Caso contrário, é fácil observar que, durante a operação linear entre pesos e variáveis, aquelas de maior magnitude tenderam a uma maior importância no resultado final.

No entanto, tratando-se de uma única combinação linear, o Perceptron possui apenas a capacidade de resolver problemas lineares. Para sanar esta limitação, utiliza-se uma rede de vários Perceptrons interligados, como na topologia representada pela Figura 10.

Com essa agregação, estamos mapeando os elementos de uma camada para

Figura 9 – Função Sigmoide



Fonte: Autor

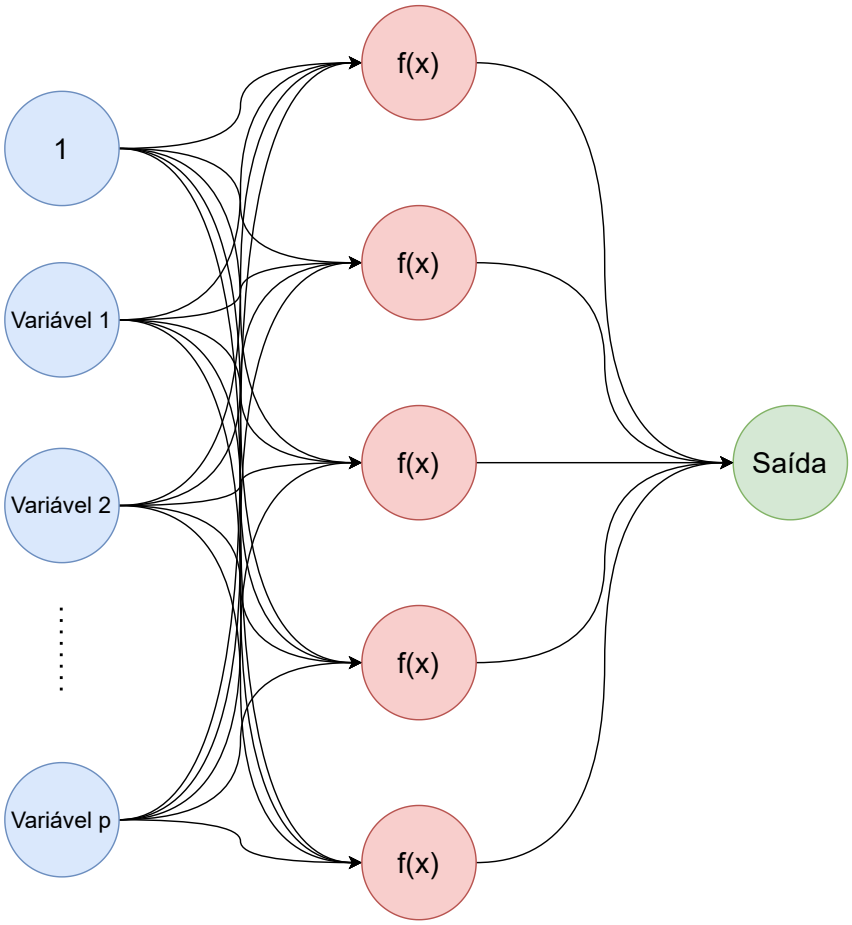
outra dimensão na camada seguinte, podendo agora resolver problemas não-lineares e aproximar funções complexas para regressão.

Existem, portanto, as camadas de entrada e de saída, que se ocupam, respectivamente, das variáveis que vão ser introduzidas ao modelo de regressão e da saída final do modelo, que para o caso da regressão possuirá apenas um neurônio com função de ativação Identidade, ou unitária, que equivale à simplesmente uma combinação linear.

As demais camadas são chamadas de camadas ocultas, estas que adicionam a complexidade necessária para que o modelo possa aproximar funções mais complexas. Não existe regra para a escolha do número de camadas ocultas ou para o número de neurônios presentes em cada uma dessas camadas, devendo ser escolhidos empiricamente. No entanto, existem referências (Heaton, 2008) que apontam que não há necessidade para mais que duas camadas ocultas para a maioria das aplicações, e que, apesar de não ser um método exato, existem algumas regras comuns para determinar o número de neurônios em cada cada oculta.

Este é um ponto importante para o desenho da topologia, uma vez que ao adicionar mais camadas e neurônios (e portanto complexidade ao modelo), o modelo pode ser induzido a um Sobreajuste (fenômeno observado em modelos estatísticos que ficam especializados em um conjunto específico de dados, sem generalizar), armazenando informações em excesso sobre o conjunto de treino, evitando que o modelo generalize bem.

Figura 10 – Estrutura de uma rede Multilayer Perceptron (os pesos e bias foram ocultados por simplicidade)



Fonte: Autor

Ao mesmo passo, a remoção de complexidade do modelo pode causar um efeito contrário, o Subajuste (fenômeno observado em modelos estatísticos que não possuem acurácia suficiente para executar determinada tarefa), em que o modelo não possui a capacidade mínima necessária para executar seu objetivo, seja ele uma classificação ou regressão. Dessa forma, esses parâmetros precisam ser escolhidos de forma bastante criteriosa.

A fase de treino para o modelo é responsável por determinar os pesos de cada ligação entre neurônios. Para isso, é preciso definir uma função de custo, utilizada para mensurar a performance da regressão com relação às observações reais. Em geral, para problemas de regressão com RNA, uma escolha comum para função de custo é o Erro Médio Quadrático (MSE).

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2.18}$$

Esse treino é realizado utilizando a técnica chamada *Backpropagation*.

Para encontrar os pesos ideais, configura-se um problema de otimização. É preciso encontrar os valores de pesos que minimizam a função de custo. Existem diversas estratégias para resolução desse problema de otimização, sendo o Gradiente Descendente Estocástico uma das mais utilizadas, que consiste em atualizar os pesos iterativamente da forma:

$$w^{t+1} = w^t - \eta \nabla_w L_{MSE}(w) \tag{2.19}$$

Em que ∇_w é o gradiente em função dos pesos w e η é chamada de Taxa de Aprendizagem, controla o quão rápido a direção do gradiente será seguida.

Em uma rede com camadas ocultas, este gradiente será composto por uma regra da cadeia (Hastie et al., 2001).

O treino, então, roda iterativamente por um número pré-definido de épocas ou até que uma condição de parada seja atingida. A condição de parada mais comum é uma comparação entre os valores de perda:

$$|L(w^t) - L(w^{t-1})| < \delta_{min} \tag{2.20}$$

Em que δ_{min} é o menor valor de variação permitido para a variação de custo. Abaixo disso, é considerado que o algoritmo convergiu.

2.4.2.2 k-Nearest Neighbors (kNN)

A k-Nearest Neighbors (kNN), ou k Vizinhos Mais Próximos, é uma das técnicas mais elementares e intuitivas para o objetivo de Classificação ou Regressão no domínio da Aprendizagem de Máquina (Hastie et al., 2001). Apesar de sua aparente simplicidade, é constantemente utilizada em problemas reais e apresenta resultados bastante sólidos.

A ideia central do modelo baseia-se na premissa de que, dado um conjunto de dados em um espaço arbitrário $\mathbb{R}^p$ e dada uma métrica de distância definida, os pontos mais próximos (ou de menor distância entre si) possuem característica semelhantes. Dessa forma, para um novo ponto fora do conjunto de dados a ser classificado, é possível observar os k vizinhos mais próximos a esse ponto e, assim, determinar a classificação ou regressão desse novo ponto, em que k é um valor arbitrária entre 0 e o número total de amostras.

Desta feita, o valor de k é um hiperparâmetro para o modelo, este deve ser definido previamente ao treinamento do modelo. Outra definição importante é a métrica de distância a ser adotada. Em geral, existem várias métricas de distância para possível utilização, sendo a mais utilizada destas a Distância Euclidiana. Sejam $x \in \mathbb{R}^p$ e $y \in \mathbb{R}^p$, sendo x_i e y_i os i-ésimo termos de x e y, respectivamente, a Distância Euclidiana entre x e y é definida como:

$$d_{euclidiana}(x,y) = ||x-y||_2 = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2} \tag{2.21}$$

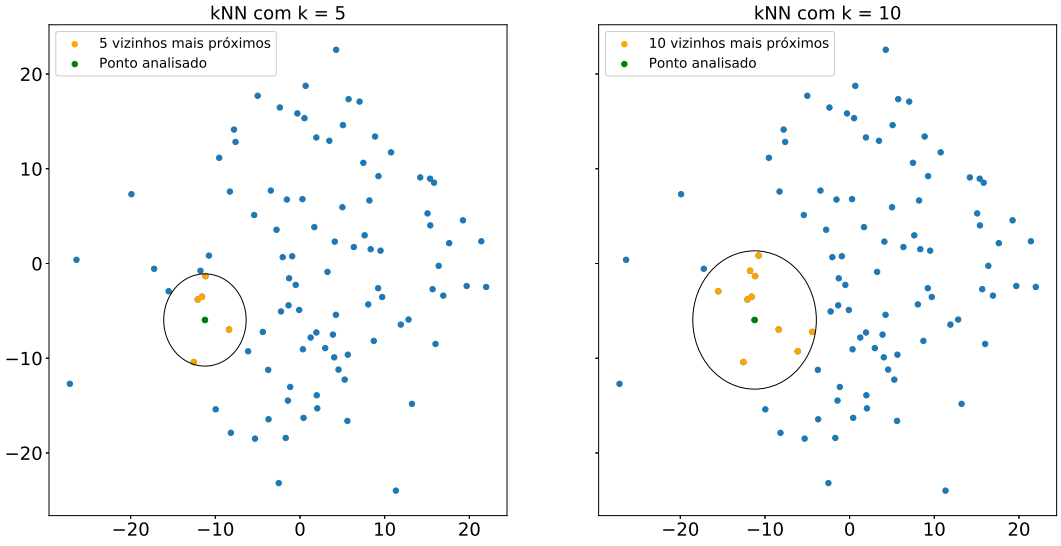
De forma mais generalizada, existe a Distância Minkowski, definida da forma:

$$d_{Minkowski}(x,y) = (\sum_{i=1}^{p}|x_i - y_i|^M)^{\frac{1}{M}} \tag{2.22}$$

Em que o parâmetro M (geralmente representado por p, o que não foi feito para evitar confusão com o número de variáveis) controla a potência a elevar a diferença entre os pontos. Observe que para $M = 2$, a Distância Euclidiana é obtida.

A Figura 11 mostra uma visualização do efeito do parâmetro k, que controla o número de vizinhos mais próximos.

Figura 11 – Comparação entre a escolha de 5 e 10 vizinhos para o kNN



Fonte: Autor

Definidos a distância e o valor de k a serem adotados, o modelo de regressão pode ser definido. Dado o conjunto de treino contendo n amostras no total, cada amostra $x_i \in \mathbb{R}^p$, da forma:

$$X_{treino} \in \mathbb{R}^{n \times p} \tag{2.23}$$

Cada amostra associada a uma variável dependente $y_i \in \mathbb{R}$, representando o valor a ser aproximado pelo modelo de regressão.

$$Y_{treino} \in \mathbb{R}^{n} \tag{2.24}$$

Seja novo ponto $x_{novo} \in \mathbb{R}^p$ fora do conjunto de treino, as distâncias $d(x_{novo}, x_i)$ para todos $x_i \in X_{treino}$ são calculadas. O conjunto dos k vizinhos mais próximos X^{kNN} será composto pelos k pontos que possuem menor distância com relação ao novo ponto.

Assim, imaginando esse conjunto $X^{kNN} \in \mathbb{R}^{k \times p}$ dos pontos mais próximos, um vetor $d \in \mathbb{R}^k$ contendo as distâncias correspondentes e $Y^{kNN} \in \mathbb{R}^k$ os valores correspondentes da variável a ser aproximada, o valor de regressão $\hat{y}_{novo}$ é calculado da forma:

$$\hat{y}_{novo} = \frac{1}{k} \sum_{i=1}^{k} \frac{y_i}{d(x_{novo}, x_i)} \tag{2.25}$$

Em outras palavras, será uma média ponderada pelo inverso da distância entre os pontos. Desse modo, os pontos com maior proximidade possuirão, também, maior poder em determinar o valor final da regressão.

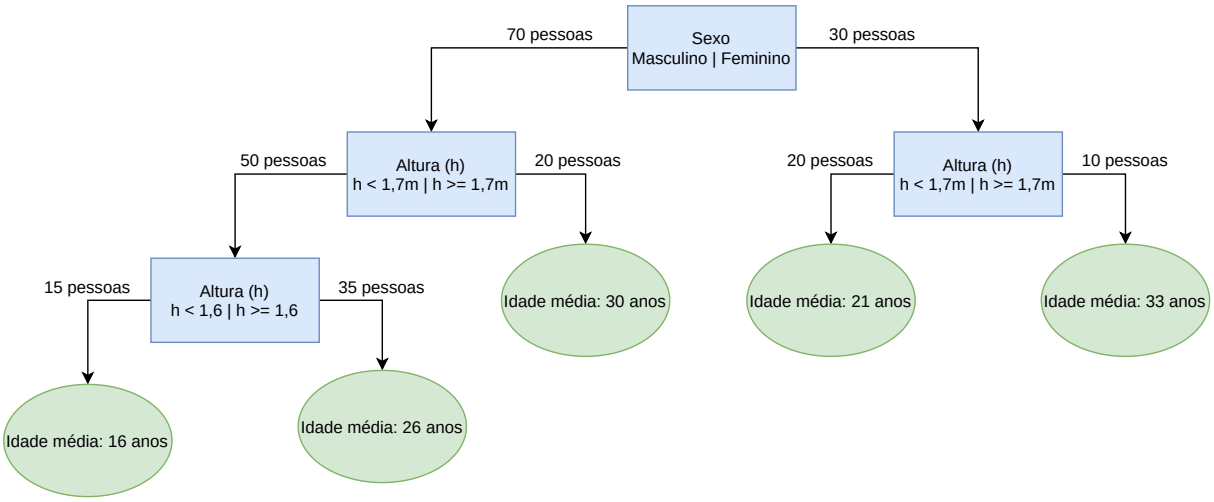
2.4.2.3 Random Forest (RF)

O Random Forest (RF) é um modelo pertencente à categoria de *Ensemble Methods* (Métodos de Comitê), que consiste em criar vários modelos simples e adotar um valor final como uma colaboração entre os vários modelos (Hastie et al., 2001). Essa estrutura tem se mostrado como sendo de mais fácil treino e ajuste, ao passo em que busca minimizar a incerteza da predição realizada, em que vários estimadores não correlacionados são levados em consideração ao mesmo tempo, buscando minimizar o grau de incerteza

do geral. Nesse contexto, o Random Forest é composto por um conjunto de Árvores de Decisão ou Regressão, um modelo simples que pode funcionar tanto para Classificação quanto para Regressão, respectivamente.

A ideia central para uma Árvore de Decisões (ou Árvore de Regressões), é criar regras aplicadas às variáveis de forma que seja possível criar subdivisões ou caminhos (os ramos da árvore) que ajudem a diferenciar os dados em cada um dos ramos. A Figura 12 mostra um exemplo simples de Árvore de Regressão, em que deseja-se estimar a idade de uma pessoa com base em características como Sexo e Altura, hipoteticamente utilizando 100 pessoas como base para construção da árvore.

Figura 12 – Exemplo simples de Árvore de Regressão para estimar a idade de uma pessoa



Fonte: Autor

Os nós são pontos da árvore que possuem subdivisões, enquanto os pontos finais (que não possuem subdivisões) são chamados de folhas.

Cada divisão objetiva criar dois subgrupos de dados, definam-se S_1 e S_2, de forma a minimizar uma métrica de custo pré-definida. Para a tarefa de regressão, usualmente se utiliza a Soma dos Erros Quadráticos. Dessa forma, sejam $\bar{y}_1$ e $\bar{y}_2$ os valores médios de S_1 e S_2, respectivamente, a Soma dos Erros Quadráticos pode ser definida como:

$$SSE = \sum_{i \in S_1} (y_i - \bar{y}_1)^2 + \sum_{i \in S_2} (y_i - \bar{y}_2)^2 \tag{2.26}$$

E a divisão que minimiza essa expressão é dita como uma divisão ótima. Assim, observa-se que as divisões buscam manter agrupados os termos com similaridades.

As folhas são os agrupamentos finais apresentadas pelo modelo. Ao fim da árvore, após todos os nós, existem as folhas, destino final dos dados. Para a regressão de um novo valor x_{novo}, procura-se qual é a folha em que esse valor se encaixa e seu valor predito será a média dos valores observados pela folha. Digamos que, seguindo a árvore, o valor x_{novo} estaria alocado em uma folha F_k com m observações, o valor de sua regressão é calculado como:

$$\hat{y} = \frac{1}{m} \sum_{i \in F_k} y_i \tag{2.27}$$

Ao decorrer das divisões realizadas, a árvore cresce verticalmente seguindo algumas premissas:

- Existe uma profundidade máxima para árvore, medida pela quantidade de nós criados verticalmente;
- Existe um número mínimo de amostras para que haja uma divisão;
- Existe um número mínimo de amostras para que um ponto possa ser considerado uma folha.

Para a Random Forest, a criação de cada árvore é precedida por um procedimento chamado *Bootstrap*, que é uma amostragem aleatória de apenas uma parte do conjunto de dados, para que as árvores tenham contato com conjuntos de dados diferentes e isso possa diminuir a correlação entre elas.

Dado um conjunto de dados com p variáveis, em cada etapa de construção de uma árvore, a melhor divisão só pode levar em consideração um número m com $m < p$ de variáveis escolhidas aleatoriamente. Ou seja, há também uma amostragem aleatória das variáveis que serão consideradas em cada divisão, aumentando ainda mais o fator estocástico do processo, este que se repetirá até que uma das condições de parada sejam atingidas.

Ao fim, haverá um conjunto de Árvores de Regressão que irão compor o resultado final, da forma:

$$\hat{y}_{RF}(x_{novo}) = \frac{1}{A} \sum_{i=1}^{A} T_i(x_{novo}) \tag{2.28}$$

Em que A é o número de árvores construídas e $T_i(x_{novo})$ é a previsão da i-ésima árvore com relação ao valor x_{novo}.

Dessa forma, os valores que controlam as condições de parada, o número m de variáveis escolhidas para cada divisão e a quantidade de árvores presentes na floresta devem ser definidos anteriormente ao desenvolvimento do modelo, sendo estes considerados hiperparâmetros que devem ser escolhidos de forma empírica para melhor adaptar-se ao conjunto de dados.

2.5 Comentários Parciais

Neste capítulo, foram cobertos todos os conceitos fundamentais que embasam o presente trabalho desenvolvido. Para as seções posteriores, destaca-se a importância das características operacionais de usinas FV, de modo a fornecer um maior contexto e interpretabilidade do procedimento, a importância dos pontos relacionados a filtragem e exploração de dados (como a Amplitude Interquartil) que serão utilizados como parte do procedimento, além dos conceitos envolvendo a modelagem em si, como a ferramenta PVlib e os algoritmos de Aprendizagem de Máquina.

3 METODOLOGIA

Neste capítulo, será explicitado todo o procedimento adotado visando os objetivos, especificando detalhes sobre os dados utilizados, exploração, filtragem, pré-processamento e, finalmente, o treinamento de modelos por Aprendizagem de Máquina, capazes de calcular a potência esperada de um inversor, dadas as medições meteorológicas.

Por motivos de simplicidade, os exemplos a aplicações do capítulo de Metodologia utilizarão apenas um inversor. Todo o procedimento será replicado para os demais inversores e suas conclusões expostas no capítulo de Resultados.

3.1 Coleta de dados

Os dados utilizados para o presente trabalho foram coletados de uma usina FV real localizada no Brasil, cedidos pela Delfos Intelligent Maintenance (www.delfosim.com), empresa especializada em monitoramento de performance e predição de falhas para Parques Eólicos, Usinas FV e outras fontes geradores de energia elétrica. Mais detalhes sobre a instalação real, as características dos dados e as variáveis utilizadas serão tratados nas subseções a seguir.

3.1.1 Planta

Para o estudo, foram utilizadas séries temporais de uma planta FV real, no período entre Setembro de 2020 e Junho de 2021, amostradas em intervalos de 10 minutos. O parque fica localizado na região Nordeste do Brasil, possui, no total, 32 inversores, somando uma potência instalada de 63,26 MWp.

Os dados são disponibilizados via Sistema SCADA e armazenados em um banco de dados historiador. A nível de inversores, existe informação sobre valores de correntes elétricas (para o lado de corrente contínua e de corrente alternada), valores de tensão elétrica e potência de saída instantânea.

Ademais, o parque conta com piranômetros e estações de meteorologia, capazes de medir as condições ambientais, tais como a irradiância global horizontal, irradiância no plano dos painéis, velocidade do vento, temperatura ambiente e umidade local, em W/m^2, W/m^2, m/s, $^\circ C$ e porcentagem, respectivamente.

3.1.2 Variáveis Utilizadas

Desta forma, tendo em vista as variáveis ambientais disponíveis para o estudo e, também, metodologias adotadas em trabalhos similares, as variáveis meteorológicas mais relevantes serão utilizadas como variáveis de entrada para o modelo, enquanto a potência ativa do inversor será a variável de saída. Da forma:

$$X = [I_{POA}, v_{vento}, T_{amb}, H_{amb}] \tag{3.1}$$

$$Y = P_{act} \tag{3.2}$$

Em que I_{POA} é a Irradiância Solar no plano inclinado, v_{vento} é a velocidade do vento, T_{amb} é a temperatura ambiente, H_{amb} é a umidade e P_{act} é a potência ativa do inversor. Ademais, $X \in \mathbb{R}^{n \times 4}$, n é o número de amostras.

Para validação da metodologia, entende-se que é preciso treinar modelos específicos para cada inversor, visto que estes podem originar de fabricantes diferentes e possuir especificações próprias. Desta forma, será treinado um modelo para cada cada inversor e um compilado dos resultados será realizado posteriormente para efeitos de validação.

3.2 Exploração dos dados

De antemão, é preciso realizar uma exploração prévia dos dados, no intuito de melhor compreender as características intrínsecas do conjunto de dados a ser trabalhado e, portanto, as características do sistema a ser modelado. De forma a melhor compreender essas características, foram observadas as correlações lineares entre diferentes variáveis e a distribuição estatística de cada uma delas, objetivando fundamentação de informações importantes para o decorrer do trabalho.

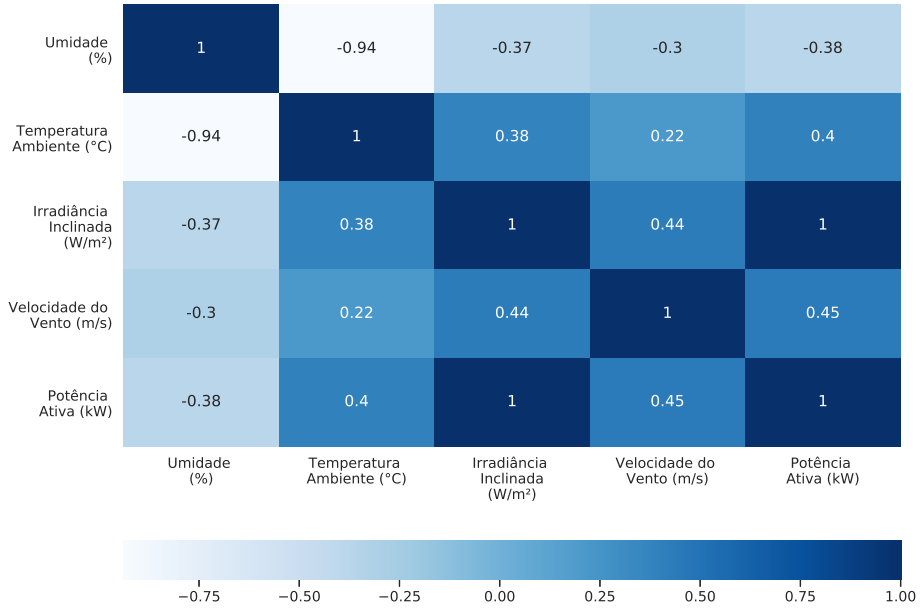
3.2.1 Correlação Linear

No intuito de melhor compreender o comportamento dos dados à disposição, uma abordagem bastante comum é o cálculo de correlação entre as variáveis. Isto é,

entender a influência das variáveis entre si, e, dessa forma, proporcionar conhecimento prévio que pode auxiliar no processo.

Dessa forma, foram calculados os Coeficientes de Correlação de Pearson, segundo relatado pela Seção 2.3.1. Os coeficientes foram, portanto, calculados para cada par de variáveis disponíveis no conjunto de dados e o resultado pode ser observado visualmente por um gráfico da calor, como representado pela Figura 13.

Figura 13 – Mapa de Calor com as correlações de Pearson



Fonte: Autor

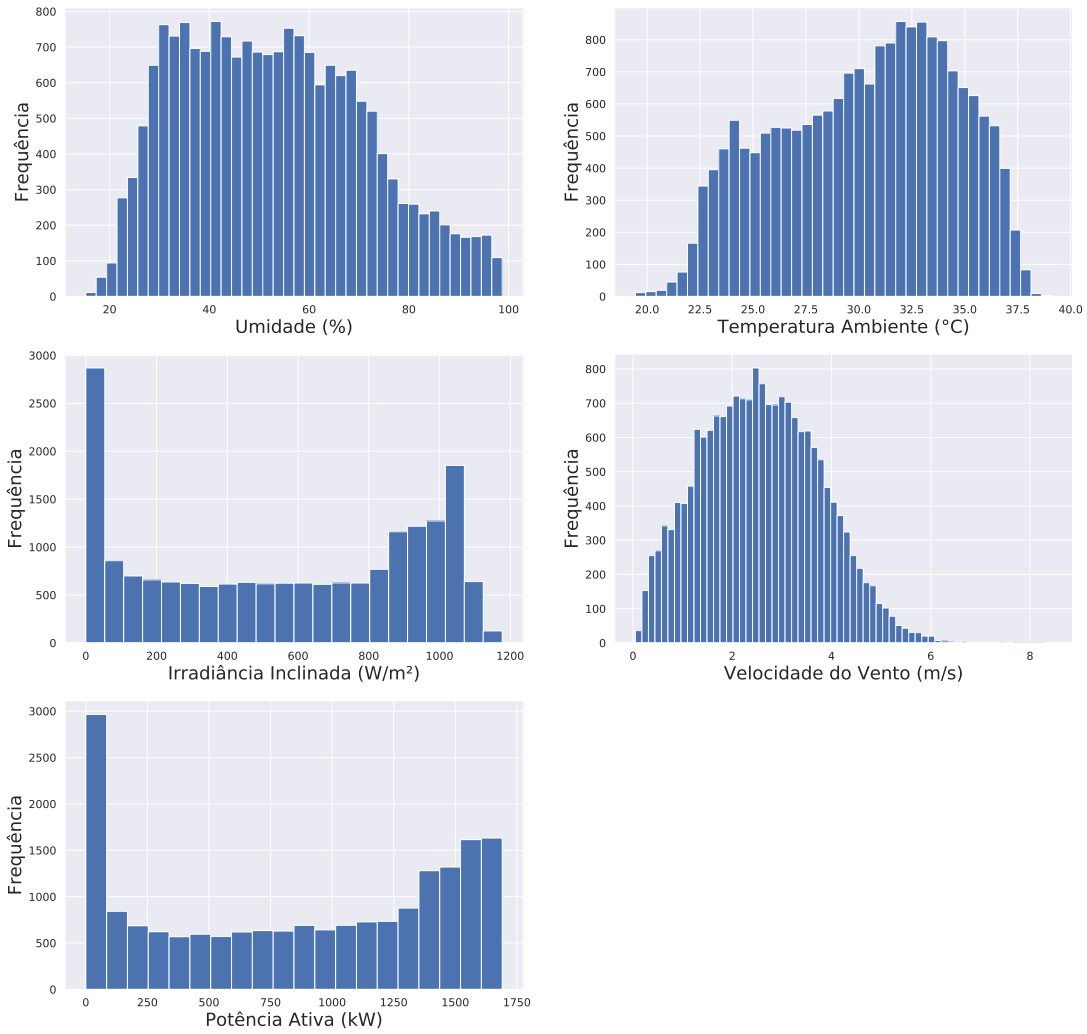
Com base na figura anterior, é possível observar traços importantes sobre os dados. A maior correlação, como é esperado, encontra-se entre a Potência Ativa e a Irradiância Inclinada, uma vez que a incidência solar influencia diretamente na produção dos painéis FV, o que eleva a correlação entre as grandezas a um nível quase unitário. No entanto, observa-se, também, que as demais variáveis também possuem correlação considerável com a variável de saída, mostrando, dessa forma, que estas guardam informação sobre a variável dependente, podendo, portanto, serem utilizadas pelos algoritmos.

3.2.2 Distribuição Estatística

Outro artifício que pode ser utilizado para investigar o caráter os dados, é a observação dos histogramas de cada uma das variáveis. Isso permite que a distribuição dos

dados possa ser visualizada de maneira direta. Os histogramas para os dados disponíveis (no período entre Setembro de 2020 e Junho de 2021) estão representados pela Figura 14.

Figura 14 – Histogramas das Distribuições das variáveis



Fonte: Autor

De posse destas distribuições, observa-se que as distribuições de Umidade, Temperatura Ambiente e Velocidade do Vento tendem a ser mais centradas, enquanto as distribuições de Irradiância Inclinada e Potência Ativa tendem a possuir dois centros polarizados, um para valores pequenos e outro para valores mais altos, o que é uma característica típica de uma planta de geração FV, que tende a operar, na maior parte do tempo, com valores de alta irradiância e, portanto, alta potência ativa.

Ademais, estas observações ressaltam, mais uma vez, uma alta correlação entre

Irradiância Inclinada e Potência Ativa, como apontada pelos histogramas.

3.3 Filtragem de dados

A filtragem de dados é uma etapa primordial para o treinamento de um algoritmo de Aprendizagem de Máquina. Consiste em eliminar amostras consideradas improváveis (e, portanto, menos previsíveis) do conjunto de dados, para que o modelo concentre-se em observações que melhor descrevam a real distribuição estatística responsável pela amostragem dos dados.

Dessa forma, filtros de dados podem ajudar na eficiência (mantendo apenas os pontos de maior interesse) e eficácia do treino, uma vez que os pontos considerados *outliers*, ou pontos fora da curva, podem, até mesmo, induzir um sobreajuste no modelo.

3.3.1 Filtros sugeridos pela IEC TS 61724-3

Foram aplicados filtros sugeridos pela norma internacional IEC TS 61724-3 (2016), em sua Tabela 3, que diz respeito a critérios de filtragem para qualidade de dados em uma planta FV.

Tabela 1 – Exemplos de critérios para filtragem de dados, a ser ajustado de acordo com condições locais

Critérios sugeridos (dados amostrados em 15 minutos)					
Filtro	Descrição	Irradiância (W/m^2)	Temperatura ($°C$)	Velocidade do Vento (m/s)	Potência CA
Intervalo	Valores fora de limites razoáveis	<-6 ou >1500	>50 ou <-30	>32 ou <0	>102% da Nominal CA ou <-1% da Nominal CA
Faltantes	Valores faltantes ou duplicados	N/A	N/A	N/A	N/A
Mortos	Valores presos em um mesmo valor ao longo do tempo (usa derivada)	<0,0001 com valor >5	<0,0001	?	?
Variações abruptas	Variações irreais entre pontos consecutivos (usa derivada)	>800	>4	>10	>80% da Nominal CA

Fonte: Adaptado da Tabela 3 da norma IEC TS 61724-3

Dessa forma, para os dados de Irradiância Inclinada, Temperatura Ambiente, Velocidade do Vento e Potência Ativa, foram aplicados os seguintes filtros:

- Intervalos: medições devem estar entre limites admissíveis (sugeridos na Tabela 3 da norma);
- Dados faltantes: deletar medições duplicadas ou com qualquer uma das variáveis faltantes para um instante de tempo;
- Dados mortos: deletar valores constantes ao longo do tempo utilizando cálculo de derivada;
- Mudanças abruptas: deletar valores que contenham mudanças abruptas entre instantes de tempo consecutivos, utilizando cálculo de derivada;
- Inversor indisponível: deletar valores para instantes de tempo em que existem alarmes de indisponibilidade do inversor ativos, adquiridos via SCADA.

Contudo, o filtro para dados mortos não foi aplicado em função da baixa precisão de casas decimais dos dados disponíveis, sendo aplicados, apenas, os filtros restantes sugeridos.

3.3.2 Filtro de Intervalo Interquartil (IQR)

Além destes, foi aplicado um filtro estatístico direcionado aos *outliers* da relação entre Potência Ativa e Irradiância. De acordo com observações de correlação obtidas na Seção 3.2.1, potência e irradiância possuem uma alta correlação linear durante uma faixa de operação normal. Desta feita, é esperado que pontos suficientemente distantes desse comportamento durante uma operação normal sejam considerados anormalidades e, portanto, sua exclusão incrementa a qualidade do conjunto de dados.

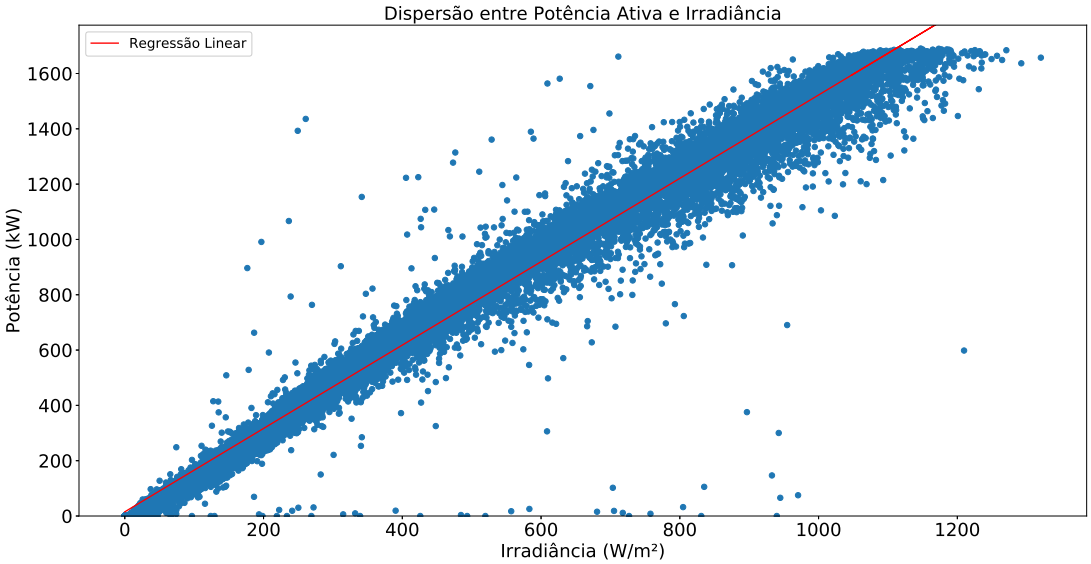
Baseando-se nessa premissa, é possível calcular uma Regressão Linear relacionando a irradiância solar inclinada $I_{POA}^{(t)}$ e a potência esperada $\hat{P}^{(t)}$ para um instante de tempo t, da forma:

$$\hat{P}^{(t)}(I_{POA}^{(t)}) = \theta_1 \cdot I_{POA}^{(t)} + \theta_0 \tag{3.3}$$

Em que $\hat{P}^{(t)}(I_{POA}^{(t)})$ é o valor esperado da potência para um instante de tempo t em função da Irradiância no Plano Inclinado I_{POA}, considerando a regressão linear. A

reta de regressão calculada, bem como os pontos disponíveis, estão exibidos na Figura 15.

Figura 15 – Dispersão entre os dados e a regressão linear antes do filtro



Fonte: Autor

Desta feita, podemos definir um erro ϵ da forma:

$$\epsilon^{(t)} = P^{(t)} - \hat{P}^{(t)} \tag{3.4}$$

Em que $P^{(t)}$ é a potência medida para o instante de tempo t e $\hat{P}^{(t)}$ é o resultado da Regressão Linear para o mesmo instante de tempo t. Este erro explicita, ponto a ponto, a diferença entre o valor real de potência ativa e o valor esperado pela premissa de linearidade.

Com estas definições, torna-se preciso definir o limiar que determina um *outlier* por este critério. Para este objetivo, será observada a distribuição da variável de erro ϵ, utilizando-se o método de Intervalo Interquartil (relatado na Seção 2.3.2). Os chamados quartis são um caso especial de quantil, caracterizando três valores que dividem o conjunto de dados em quatro partes iguais. O segundo quartil divide o conjunto de dados pela metade (coincidindo, portanto, com a mediana), enquanto os outros dois dividem em uma proporção de 25% e 75%.

Segundo este método, dada a distribuição do erro calculado, uma amostra x pode ser considerada fora do comportamento padrão caso obedeça a uma das condições:

- $x < Q_1 - 1.5 \cdot IQR$

- $x > Q_3 + 1.5 \cdot IQR$

Em que:

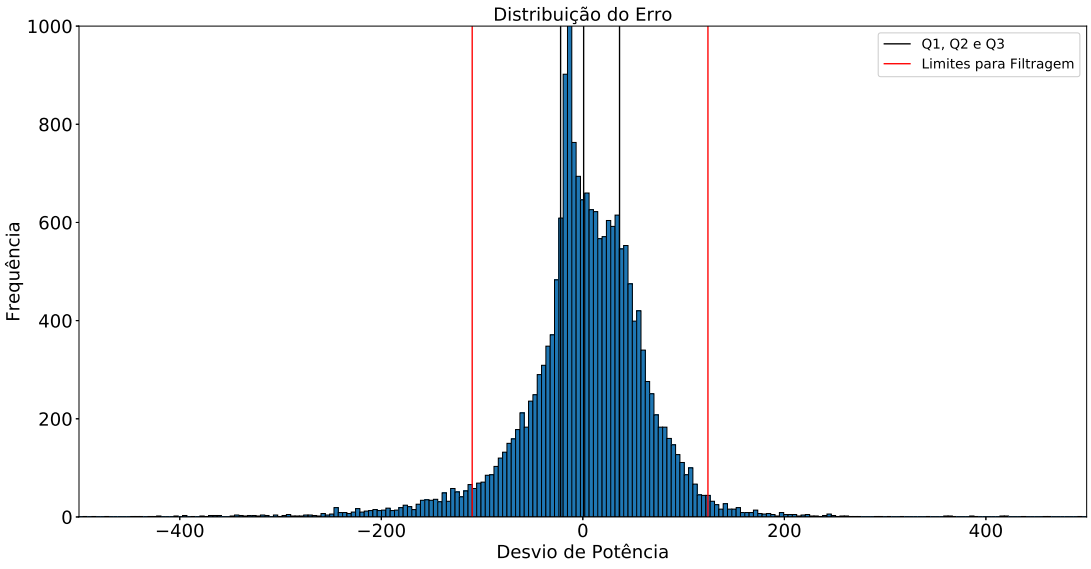
- Q_1: primeiro quartil;
- Q_2: segundo quartil (mediana);
- Q_3: terceiro quartil.

Ademais:

$$IQR = Q_3 - Q_1 \tag{3.5}$$

O cálculo dos quartis e do intervalo que define *outliers* foi calculado para a variável de erro ϵ e explicitada por um histograma, como na Figura 16, em que as linhas verticais pretas indicam, da esquerda para a direita, o primeiro, o segundo e o terceiro quartil, e as linhas verticais vermelhas indicam os intervalos inferior e superior para o filtro.

Figura 16 – Histograma da distribuição do erro ϵ e os pontos de filtragem de acordo com o método de Intervalo Interquartil

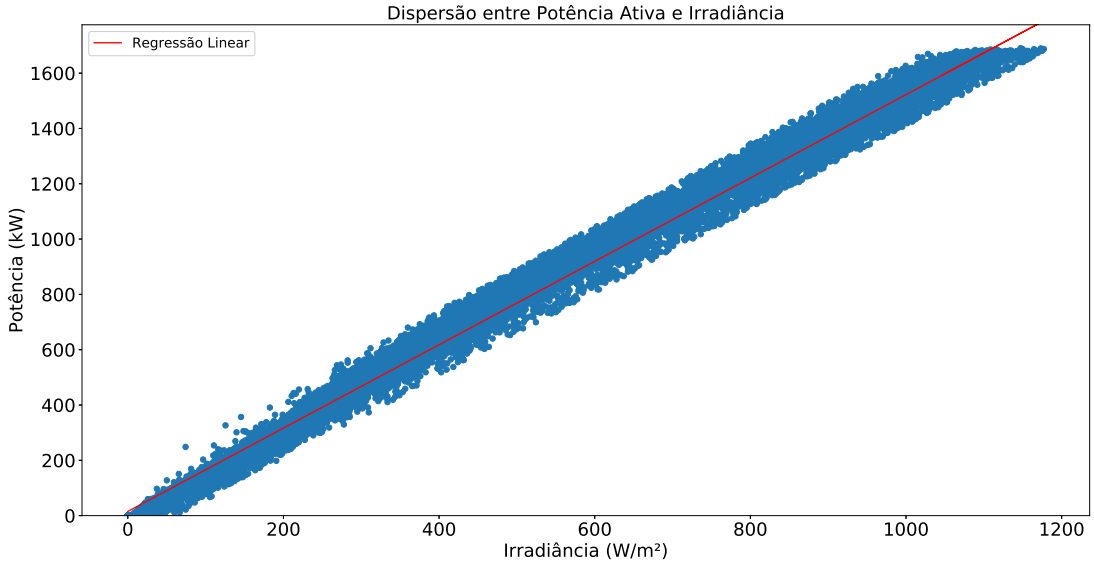


Fonte: Autor

Retirando os pontos sugeridos pelo método, a dispersão dos pontos diminui com relação à regressão linear calculada à priori, como mostra a Figura 17.

Dessa forma, garantimos que possuímos à disposição apenas os pontos que obedecem a premissa da alta correlação linear entre estas grandezas, retirando os pontos

Figura 17 – Dispersão entre os dados e a regressão linear depois do filtro



Fonte: Autor

que mais destoavam e, portanto, são menos previsíveis para um algoritmo.

Ao fim de todos os métodos de filtragem (incluindo os métodos sugeridos pela norma), resta a quantidade de dados representada pela Tabela 2, para o inversor em questão:

Tabela 2 – Resumo dos filtros

Etapa	**Quantidade de Pontos**	**Porcentagem (%)**
Início	52445	100
Disponibilidade e Dados Faltantes	17737	33,82
Intervalo de valores aceitáveis	17737	33,82
Variações Abruptas	17732	33,81
Filtro IQR	16712	31,87

Fonte: Autor

3.4 Processamento

Após a etapa de filtragem dos dados, há uma etapa também bastante importante para o resultado final: o processamento (ou pré-processamento) dos dados.

São cálculos ou adequações realizadas para que os dados estejam em uma representação ótima do ponto de vista do modelo que será treinado.

3.4.1 Normalização

A normalização (Shalev-Shwartz and Ben-David, 2014) é um dos tratamentos de dados mais importantes do ponto de vista de algoritmos de Aprendizagem de Máquina.

Existem modelos de aprendizagem que são sensíveis à grandeza dos dados, de modo que dados que apresentam maior ordem de grandeza possuem uma maior relevância para o processo de predição ou regressão. Desse modo, é preciso que haja um escalonamento das variáveis, de modo igualar a relevância de todas, independendo da ordem de grandeza inicial.

Uma das técnicas de escalonamento mais utilizadas na literatura é a Normalização Mínimo-Máximo, que consiste em remapear um conjunto de dados a um intervalo normalizado entre 0 e 1. Desta forma, seja $x \in \mathbb{R}^n$, onde n é o número de amostras da variável:

$$x_{normalizado} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{3.6}$$

Em que x_{min} é o valor mínimo e x_{max} é o valor máximo assumidos x dentre as n amostras.

Este método, no entanto, é altamente sensível à presença de *outliers*, uma vez que lida-se com valores de mínimo e máximo de uma amostra de pontos. Desta feita, ressalta-se a importância de um processo de filtragem antecedendo ao escalonamento.

3.4.2 Conjuntos de Treino e Teste

Uma prática, também, bastante comum tratando-se de Aprendizagem de Máquina é a separação de um conjunto de dados para treino do modelo e outro para teste (ou validação), para que o modelo possa ter um treino não enviesado.

Essa separação garante que os dados utilizados para aplicar métricas de validação nunca antes foram expostos ao modelo durante treinamento, de forma que a ocorrência de um sobreajuste do modelo ficará mais evidente. A proporção entre a quantidade de amostras para treino é arbitrária, geralmente em torno de 70%, sujeito à escolha, e a amostragem desses conjuntos geralmente é de forma aleatória, de modo que o risco de viés em um dos conjuntos de dados seja minimizado.

3.5 Modelos Comparados

Após todo o tratamento prévio dos dados, diferentes modelos serão comparados no objetivo predizer a potência ativa do inversor com melhor precisão. Entre estes modelos, estão modelos físicos e modelos de Aprendizagem de Máquina, que possuem suas divergências essenciais e abordagens diferentes.

Desse modo, os modelos analisados serão:

- Modelo físico do PVlib;
- k-Nearest Neighbors (kNN);
- Multilayer Perceptron Network (MLP);
- Random Forest (RF).

Ressalta-se que a motivação para escolha dos modelos analisados e suas características com mais detalhes estão relatadas na Seção 2.4.2.

A implementação do modelo do MLP foi realizada por meio da biblioteca Keras (Chollet et al., 2015), e os demais modelos de aprendizagem foram implementados com a biblioteca Scikit-learn (Pedregosa et al., 2011).

O modelo físico foi projetado pela Delfos Intelligent Maintenance, utilizando a versão estável 0.8.1 do software PVlib (Will Holmgren et al., 2021) e as características dos painéis FV providas pelo fabricante. Portanto, o modelo físico em estudo é propriedade da Delfos e é utilizado para análises dessa planta FV.

Aprendizagem de Máquina, em geral, utiliza-se de dados históricos para tentar modelar comportamentos. Desse modo, é preciso expor o modelo ao conjunto de pontos (em uma fase de treinamento) para que este possa se adaptar às características do dado.

3.5.1 Validação Cruzada

Validação Cruzada (Hastie et al., 2001) é um método utilizado para testar a generalização de um modelo, treinando e testando com diferentes subconjuntos de dados. Esta é uma alternativa à criação de um terceiro conjunto de dados (o conjunto de validação), utilizando os dados à disposição de maneira mais eficiente.

Alguns modelos possuem parâmetros variáveis e que devem ser especificados antes do início do treino, são os chamados hiperparâmetros, como explicitados na Seção 2.4.2. Como na maioria das vezes esses hiperparâmetros não possuem uma regra clara para

escolha, eles usualmente são determinados empiricamente por tentativa e erro. Ou seja, alguns valores possíveis são determinados e testados, aqueles que apresentarem melhores resultados serão os hiperparâmetros adotados.

Dessa forma, a Validação Cruzada é um método bastante utilizado para determinação de hiperparâmetros, avaliando a forma como um modelo generaliza bem utilizando diferentes hiperparâmetros.

O procedimento de validação cruzada se dá com o processo:

1. O conjunto de dados é dividido aleatoriamente em K partições com quantidade de dados aproximadamente iguais;
2. Para cada partição de dados k_i em K:
   a) A partição k_i é definida como conjunto de validação;
   b) O modelo é treinado utilizando as $K-1$ partições restantes;
   c) O valor de custo é calculado para o modelo utilizando o conjunto de validação.
3. O custo médio apresentado pelas K partições é calculado.

Assim, esse procedimento será repetido para cada uma das configurações possíveis de hiperparâmetros, de modo a determinar a configuração com menor valor de custo e, portanto, a que melhor se adequa à aplicação.

Os hiperparâmetros candidatos para cada modelo são exibidos na Tabela 3.

Tabela 3 – Valores disponíveis para os hiperparâmetros

	Hiperparâmetro	Valores candidatos
Multilayer Perceptron Network	Nº neurônios na primeira camada oculta	{1; 2; 3; 4}
	Nº neurônios na segunda camada oculta	{0; 1; 2; 3; 4}
k-Nearest Neighbors	Número de vizinhos (k)	[1, 100]
	Parâmetro (p) para a distância Minkowski	{1; 1,5; 2; 2,5; 3}
Random Forest	Máxima profundidade	{80; 90; 100; 110}
	Máximo número de variáveis por divisão	{2; 3}
	Mínimo número de amostras por folha	{3; 4; 5}
	Mínimo número de amostras por divisão	{8; 10; 12}
	Número de estimadores	{100; 200; 300; 1000}

Fonte: Autor

Para a escolha do número de camadas ocultas e número de neurônios para cada

camada da MLP, procurou-se manter o número de neurônios por camada menor que o número de variáveis de entrada (para controlar a complexidade da rede) e o número de camadas ocultas em no máximo 2 (Heaton, 2008).

Já para os algoritmos de kNN e RF, procurou-se definir a gama dos valores dos hiperparâmetros de forma empírica, observando a influência do comportamento dos modelos pela alteração desses parâmetros dos algoritmos.

3.6 Métricas de Validação

Para validar o desempenho dos modelos comparados, métricas de validação precisam ser definidas. Tratando-se de modelos de regressão, existem métricas bastante consolidadas na literatura, que possibilitam uma melhor observação do resultado final.

Para as métricas avaliadas a seguir, admitem-se $\hat{y}$ como o vetor de predições para y e $\bar{y}$ como a média dos valores de y.

A seguir, as métricas de validação adotadas no presente trabalho.

3.6.1 Coeficiente de Determinação (R2)

O Coeficiente de Determinação (R2) é uma medida de comparação entre a incerteza da saída de um modelo de regressão e a incerteza dos dados reais (Dodge, 2008). Dessa forma, é uma métrica que varia entre 0 e 1, e, quanto mais próxima de 1, melhor é o resultado do modelo e mais correlacionado o resultado está do real.

Dessa forma, a soma quadrática dos resíduos é definida como:

$$SS_r = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.7}$$

Ademais, a soma quadrática total:

$$SS_t = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{3.8}$$

E, logo, o Coeficiente de Determinação pode ser definido como uma relação entre eles, da forma:

$$R^2 = 1 - \frac{SS_r}{SS_t} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{3.9}$$

3.6.2 Raiz do Erro Quadrático Médio (RMSE)

A Raiz do Erro Quadrático Médio (RMSE) é uma métrica não-negativa da diferença entre os valores preditos pelo modelo e os dados reais (Hyndman and Koehler, 2006).

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{1}{n}(y_i - \hat{y}_i)^2} \tag{3.10}$$

Observa-se que, devido ao comportamento quadrático, grandes erros possuem contribuições maiores, desproporcionais ao efeito de erros menores. Desta forma, é uma métrica bastante sensível a *outliers*.

3.6.3 Erro Absoluto Médio (MAE)

O Erro Absoluto Médio, ou *Mean Absolute Error* (MAE), é similar à Raiz do Erro Quadrático Médio, com exceção do comportamento quadrático observado na métrica anterior (Hyndman and Koehler, 2006).

$$MAE = \sum_{i=1}^{n} \frac{1}{n}|y_i - \hat{y}_i| \tag{3.11}$$

No caso do MAE, cada erro contribui de forma proporcionalmente direta para o erro geral, contornando o problema de maior significância para *outliers*, além de ser uma métrica mais direta.

3.7 Comentários Parciais

Neste capítulo, detalhou-se todo o procedimento proposto para obtenção dos resultados almejados, desde a filtragem e tratamento dos dados até as métricas de validação que serão utilizadas para comparar os diferentes resultados obtidos. Desta feita, a sequência do trabalho buscará aplicar toda a metodologia no conjunto de dados analisado, visando os objetivos previamente traçados para o estudo.

4 SISTEMA DE MONITORAMENTO DE PERFORMANCE

Após as comparações entre os diferentes métodos de modelagem e os resultados obtidos, este capítulo será dedicado a propor um sistema de regras que utiliza os valores de energia esperada providos pela modelagem para identificar casos de operação com baixa performance, divididos em três tipos de ocorrência. Este sistema objetiva prover uma detecção direta de possíveis pontos de atenção no acompanhamento de performance de uma usina real, disponibilizando, também, análises visuais para facilitar a gestão de desempenho de usinas.

Além disso, será proposta, junto ao sistema de regras, uma interface visual para exemplificar um uso prático da ferramenta. Nesta interface visual, será possível analisar os dados dos inversores individualmente, observando os possíveis pontos de atenção e facilitando o processo para que a tomada de decisão sobre atitudes corretivas seja realizada de maneira otimizada.

4.1 Regras do Sistema

Baseando-se nos perfis comuns de perda de performance apontados na Seção 2.2.3.3 e representados pela Figura 4, foram desenvolvidas regras de comparação entre variáveis de operação, no intuito de categorizar possíveis ocorrências de baixa performance em um dos perfis listados anteriormente.

As variáveis utilizadas para as definições das regras são:

- Razão de Performance (PR);
- Potência Esperada ($P_{esperada}$);
- Tempo de duração (Δt).

Em que a Razão de Performance é como especificada na Seção 2.2.3.4, mas calculada em termos de potência, analogamente à energia.

Os dados da série temporal serão analisados ao longo do tempo. Os pontos que obedecerem às características impostas por uma das regras, será rotulada como pertencente à esta regra e gerará uma ocorrência. Uma ocorrência é definida por um instante de tempo de início t_{inicio} e um instante de tempo final t_{fim}, indicando o instante de tempo em que o inversor começou a apresentar a perda de performance e o instante de tempo em que este voltou a operar com normalidade, respectivamente.

4.1.1 Regra 1: Perdas momentâneas

Para a determinação de episódios de perdas momentâneas, foi adotada a seguinte regra:

- $0,1 < PR < 0,95$
- $\Delta t > 1h$

Em outras palavras, estarão sendo considerados pontos que consecutivamente apresentarem valores de PR dentro da faixa especificada, com uma duração mínima de 1 hora, adicionando uma tolerância temporal. O limite inferior foi estabelecido de forma e não considerar episódios com característica de Perda Total de operação e o limite superior adiciona uma tolerância de 5% para admitir as perdas de performance.

4.1.2 Regra 2: Perdas totais

Para especificar episódios de perda total, a seguinte regra foi adotada:

- $PR \leq 0,1$
- $P_{esperada} > 100$
- $\Delta t > 10min$

Desse modo, será considerado como uma ocorrência um conjunto de pontos consecutivos que apresentar $PR \leq 0,1$ por pelo menos 10 minutos. Adicionalmente, considera-se a condição de $P_{esperada} > 100$, no intuito de evitar pontos noturnos, que, naturalmente, apresentarão um $PR = 0$.

4.1.3 Regra 3: Perdas dissolvidas

Já a terceira regra, utilizará critérios mais específicos. Como mostra a Figura 4, idealmente um caso de perdas dissolvidas ao longo do dia seria um episódio contínuo de baixa performance. No entanto, utilizando dados reais, essa caracterização pode não ser tão constante, apresentando diversos pequenos episódios de baixa performance durante o dia.

Desse modo, é preciso considerar uma métrica diária para identificar esse tipo de perda. Para isso, o critério principal para a terceira regra será a quantidade de pontos que possuem $PR < 1$ durante um dia inteiro.

Logo, foi considerada a regra como:

- Se mais da metade das amostras de um dia apresentarem $PR < 1$:
  - $PR < 1$
  - $P_{esperada} > 100$
  - $\Delta t > 10min$

Em outras palavras, dentro de um dia em que mais da metade dos instantes de tempo observados apresentam $PR < 1$, as ocorrências serão formadas por estes pontos de perda de performance consecutivos que apresentem $P_{esperada} > 100$ e duração maior que 10 minutos. Isso poderá originar várias pequenas ocorrências ao longo do dia, que podem sinalizar alguma problemática de perda de performance estacionária, tal como sujeira nos painéis, exigindo, portanto, ações mitigadoras.

4.2 Interface Visual

A interface visual proposta para o sistema de acompanhamento utilizará o Streamlit (Sehmi et al.), uma ferramenta que permite a criação de um *Dashboard* (quadro interativo para visualizações de dados) em Python. Esta ferramente é extensamente utilizada por empresas que trabalham com análise de dados e é compatível com várias bibliotecas de visualização de dados.

Dentro da plataforma, será possível observar diversas análises e informações sobre os inversores de uma usina FV, todas estas informações baseadas em cálculos de performance. As análises são subdivididas em quatro seções principais, dentro da plataforma.

As análises presentes serão:

- Tabela de Ocorrências: página exibindo todas as ocorrências categorizadas pelo conjunto de regras do sistema;
- Séries Temporais: página exibindo as séries temporais de Potência Ativa e Potência Esperada para inversores e faixas de tempo específicas, destacando as ocorrências visualmente;
- PR Diário: página exibindo o cálculo de PR consolidado diariamente, representado por um gráfico de barras;
- Dispersão entre Potência Real e Potência Esperada: página exibindo um gráfico de dispersão entre as duas grandezas principais, destacando o horário específico para cada ponto.

Para cada uma das análises, filtros poderão ser especificados, entre eles: tempo de início da análise, tempo final da análise e inversor analisado.

4.3 Comentários Parciais

O sistema de regras proposto foi detalhado e será aplicado em prática na seção de Resultados, utilizando os modelos estudados nos capítulos anteriores e os aplicando em dados inéditos, de forma a simular uma aplicação de monitoramento em uma usina real.

5 RESULTADOS

Neste capítulo, serão tratados os resultados obtidos segundo a Metodologia apontada e o sistema de regras proposto. Estes resultados buscam validar os objetivos previamente traçados para o trabalho, visando, portanto, mensurar de forma objetiva o produto do estudo.

Para cada inversor na planta FV em estudo, foram treinados modelos preditivos para serem comparados. Antes do treinamento, os filtros foram aplicados para cada um dos inversores.

A topologia e os hiperparâmetros dos modelos foram definidos a partir do método de Validação Cruzada, explicado na Seção 3.5.1, utilizando $K = 4$ subconjuntos. Esse método foi aplicado apenas para um inversor e seu resultado (os hiperparâmetros selecionados) replicado para os demais, devido ao alto esforço computacional necessário e partindo da premissa de que o comportamento dos dados e, portanto, os hiperparâmetros dos modelos serão similares entre si. Ressalta-se, no entanto, que os modelos foram treinados separadamente para cada inversor. Portanto, cada inversor possui seus próprios modelos treinados especificamente, e, dessa forma, a premissa adotada foi aplicada apenas para topologia e determinação de hiperparâmetros.

Por fim, foi realizada a comparação dos modelos em estudo utilizando as métricas propostas, em conjunto com interpretações para essas métricas e os resultados observados.

5.1 Filtragem total

Para cada um dos inversores no estudo, foram aplicados os filtros sugeridos pelo capítulo de Metodologia. A Tabela 4 mostra o percentual final de dados para cada inversor, além da quantidade de dados para cada uma das etapas do processo de filtragem.

Como é possível observar, em geral o filtro responsável pela retenção da maior quantidade de dados é o que diz respeito a dados faltantes, que ocorrem devido a falhas no sistema de medição local do parque. No entanto, é possível observar, também, que o filtro IQR possui sua parcela de contribuição, retirando *outliers* do conjunto de dados, o que melhora a qualidade do treino, como já explicado.

Tabela 4 – Resultados dos filtros aplicados a todos os inversores

Inversor	Início	Dados faltantes	Disponi-bilidade	Intervalos aceitáveis	Mudanças abruptas	IQR	Final (%)
1	33179	20378	20378	20356	20341	18464	55,65%
2	33179	18708	18703	18661	18652	17820	53,71%
3	33179	18753	18753	18751	18745	17525	52,82%
4	33179	20666	20630	20630	20626	18720	56,42%
5	33179	20367	20356	19834	19817	19391	58,44%
6	33179	20372	20367	20367	20360	19540	58,89%
7	33179	20315	20314	20314	20310	18284	55,11%
8	33179	20370	20354	20354	20346	19535	58,88%
9	33179	20309	20303	19746	19720	19242	57,99%
10	33179	20473	20473	20473	20469	19647	59,22%
11	33179	20478	20472	20472	20466	19739	59,49%
12	33179	18762	18734	18732	18725	17600	53,05%
13	33179	18681	18673	18631	18622	17515	52,79%
14	33179	20654	20639	20639	20634	19675	59,30%
15	33179	20310	20304	19747	19721	18671	56,27%
16	33179	20365	20349	19827	19810	19312	58,21%
17	33179	20477	20469	20469	20463	19733	59,47%
18	33179	19970	19863	19841	19823	18138	54,67%
19	33179	20385	20375	20366	20359	18501	55,76%
20	33179	20443	20435	20435	20427	19617	59,12%
21	33179	20666	20660	20660	20656	19081	57,51%
22	33179	18678	18666	18624	18615	17393	52,42%
23	33179	18708	18699	18657	18648	17871	53,86%
24	33179	20259	20226	20226	20221	18938	57,08%
25	33179	20422	20413	20413	20407	19539	58,89%
26	33179	20667	20657	20657	20652	19506	58,79%
27	33179	20315	20302	20302	20298	18317	55,21%
28	33179	20473	20466	20466	20462	19546	58,91%
29	33179	18752	18752	18750	18744	17511	52,78%
30	33179	18762	18756	18754	18748	17377	52,37%
31	33179	20386	20385	20376	20370	18515	55,80%
32	33179	20317	20297	20297	20292	19163	57,76%

Fonte: Autor

5.2 Validação Cruzada

A validação cruzada foi realizada de forma a determinar os melhores hiperparâmetros para os algoritmos de aprendizagem. A função de custo utilizada no processo foi o RMSE, explicitado na seção 3.6.2.

Desta forma, os parâmetros escolhidos estão representados pela Tabela 5.

Apresentando, durante o processo de validação cruzada, os valores de custo

Tabela 5 – Hiperparâmetros selecionados por Validação Cruzada

	Hiperparâmetro	Valor
Multilayer Perceptron Network	Nº neurônios na primeira camada oculta	4
	Nº neurônios na segunda camada oculta	1
k-Nearest Neighbors	Número de vizinhos (k)	22
	Parâmetro (p) para a distância Minkowski	1,5
Random Forest	Máxima profundidade	80
	Máximo número de variáveis por divisão	3
	Mínimo número de amostras por folha	5
	Mínimo número de amostras por divisão	8
	Número de estimadores	1000

Fonte: Autor

ilustrados na Figura 18.

Como para a validação cruzada do algoritmo de Random Forest foram avaliados 5 hiperparâmetros diferentes, a representação visual desses resultados precisou ser simplificada, representando a Máxima Profundidade (max_depth) por cores, o Máximo Número de variáveis por divisão ($max\_features$) por símbolos e os demais hiperparâmetros implícitos ao decorrer do eixo das abscissas.

5.3 Métricas de Validação

Após a determinação dos melhores hiperparâmetros pelo método de Validação Cruzada, os modelos de aprendizagem foram criados e treinados separadamente para cada um dos inversores, de modo que para cada inversor, há um modelo de Multilayer Perceptron, k-Nearest Neighbors, Random Forest e o modelo físico utilizando o PVlib.

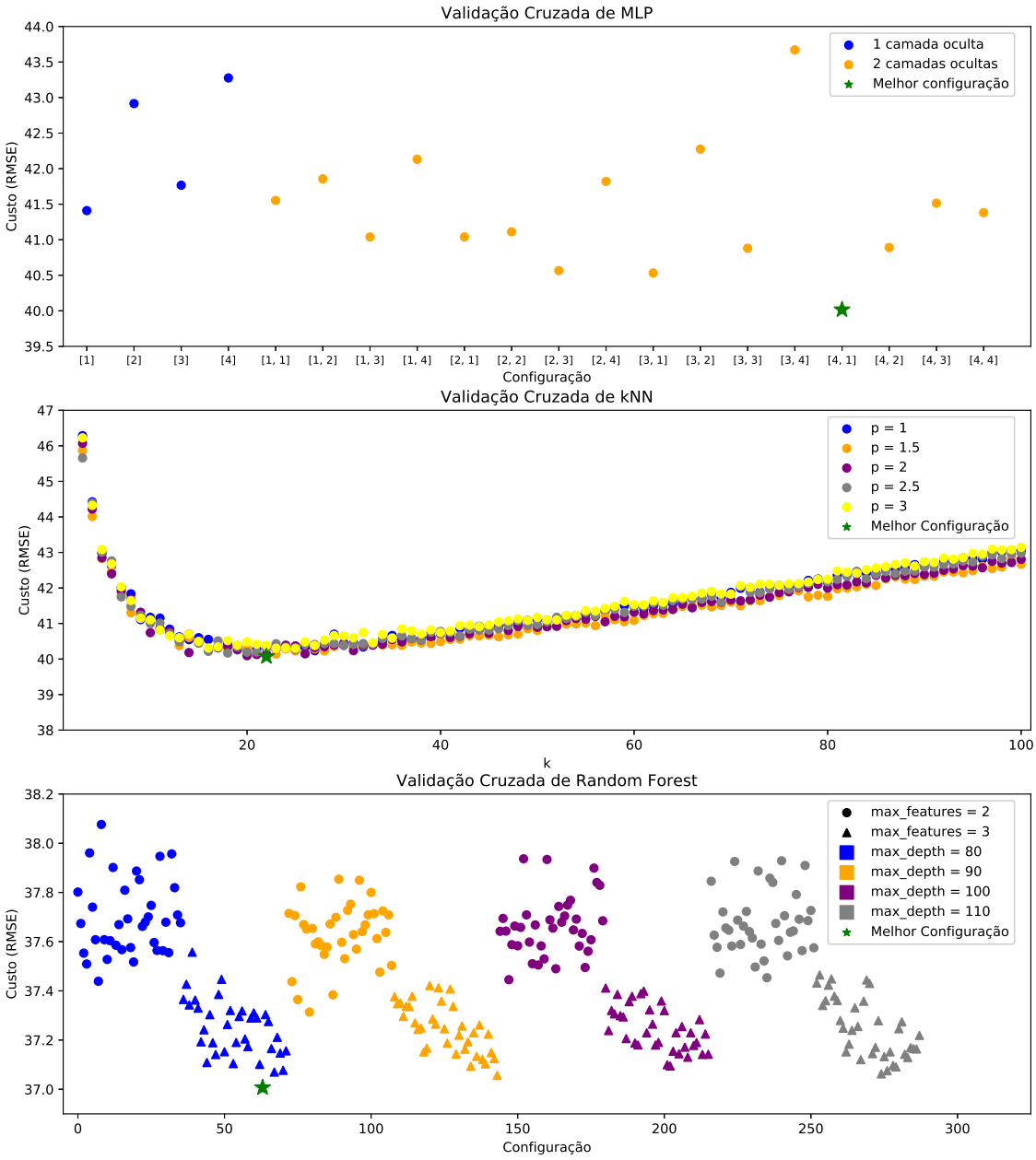
Para este treino definitivo, foi utilizada apenas a divisão simples entre conjunto de treino e conjunto de teste, na proporção de 70% de dados de treino, que serão utilizados durante a fase de treino de cada modelo, enquanto o conjunto de teste será utilizado na obtenção das métricas de validação.

Desse modo, a Tabela 6 explicita os resultados gerais para cada modelo utilizando cada métrica, unificando todos os inversores do estudo.

De posse dos resultados de validação, é possível observar que os modelos de Aprendizagem de Máquina tendem a apresentar uma performance superior quando comparados ao modelo físico. Em especial, o modelo de Random Forest apresentou as melhores performances para as três métricas avaliadas.

Para visualizar o resultado de forma mais gráfica e prática, pela Figura 19

Figura 18 – Valores de custo para procedimento de Validação Cruzada



Fonte: Autor

é possível analisar os resultados em um dia de céu claro para um inversor específico, mostrando um cenário ideal de produção dos inversores. Para cada modelo, também são expostas as métricas relativas ao dia específico.

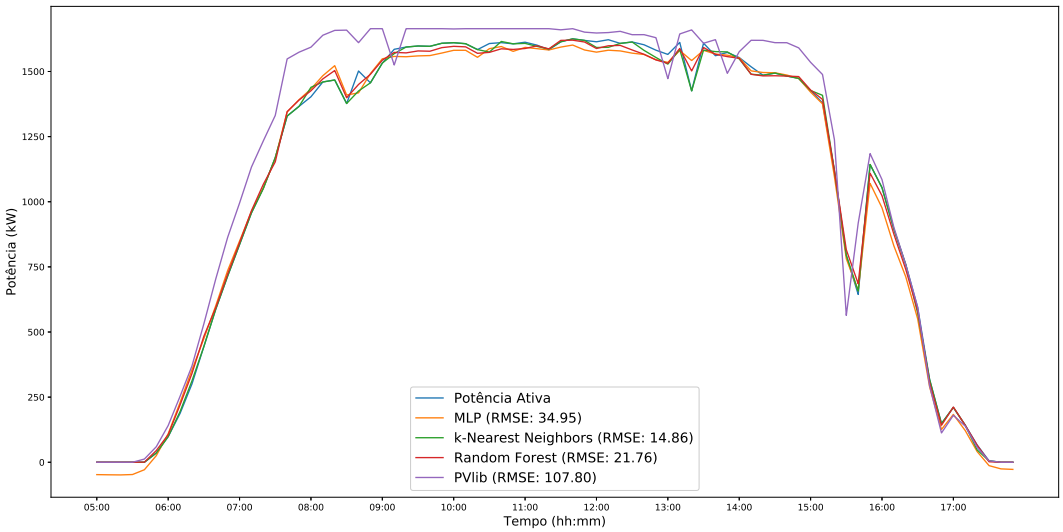
Para este dia relativamente sem interrupções de irradiação solar, é possível perceber, visualmente, que nas porções de operação com altos valores de irradiância, o modelo físico apresenta um erro maior, apresentando valores maiores do que o da operação real. Para este dia específico, o modelo de kNN apresentou o valor de 14,86 para a métrica

Tabela 6 – Métricas de validação modelos comparados

	RMSE	R2	MAE
Multilayer Perceptron Network	61,05	0,9890	42,17
k-Nearest Neighbors	58,80	0,9898	38,85
Random Forest	57,42	0,9903	36,26
Modelo Físico (PVlib)	172,83	0,9104	108,30

Fonte: Autor

Figura 19 – Visualização das predições e seus respectivos RMSE em comparação com a potência real medida



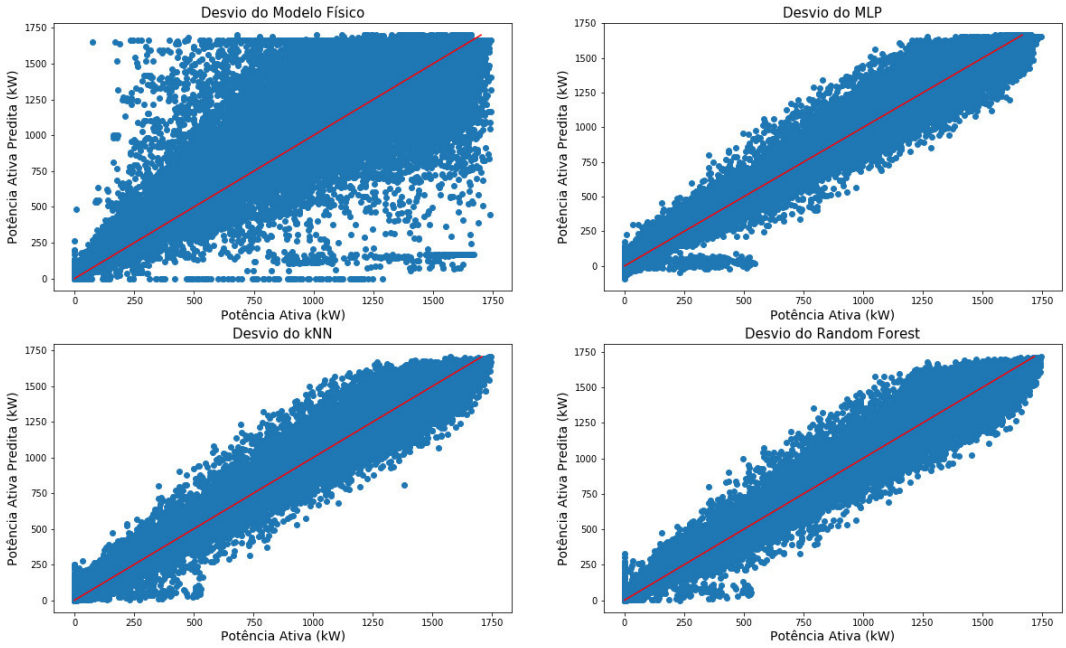
Fonte: Autor

de RMSE (sendo esta a melhor métrica entre os modelos para o dia específico), mas o modelo de Random Forest continuou apresentando bons resultados (com um valor de 21,76 para o RMSE), enquanto o modelo físico demonstra a menor performance na métrica adotada (107,80 de RMSE).

Outra forma de avaliar graficamente a performance dos modelos é visualizando a sua dispersão com relação a variável dependente, a potência ativa. Esta que é uma maneira de observar visualmente o efeito da métrica de Erro Absoluto Médio, e está exibida pela Figura 20.

Validando, mais uma vez, o resultado das métricas de validação, em que os modelos de Aprendizagem de Máquina possuem um MAE relativamente parecido, enquanto o modelo físico apresenta uma dispersão maior.

Figura 20 – Dispersão visual para todos os modelos



Fonte: Autor

5.4 Estudo de Caso

Após a validação dos modelos, foi separado um novo conjunto de dados independente ao treino e validação para a mesma usina FV. O intuito é simular um caso real de utilização do sistema proposto para o acompanhamento de performance.

O novo conjunto de dados comporta medições realizadas no mês de Julho de 2021 (um período de 31 dias) para os mesmos 32 inversores analisados. Serão aplicadas as regras propostas na Seção 4.1 para a identificação de ocorrências de baixa performance. Ademais, a plataforma visual será implementada e exibida em mais detalhes, mostrando suas funcionalidades e seus potenciais de auxílio em operações reais.

Como o algoritmo de RF foi o que apresentou as melhores métricas de validação, esse modelo foi utilizado para calcular a energia esperada para esse novo conjunto de dados, para que, com base nessas regressões, sejam aplicadas as regras do sistema.

5.4.1 Aplicação das Regras do Sistema

As regras de classificação para ocorrências de baixa performance foram aplicadas ao novo conjunto de dados, resultando em ocorrências que serão exemplificadas a seguir e poderão ser exploradas na plataforma proposta.

No total, foram registradas 667 ocorrências para os 32 inversores no período de 31 dias, sendo destas 171 pela Regra 1, 10 pela Regra 2 e 486 pela Regra 3. A grande quantidade de ocorrências de Regra 3 é justificada pela própria natureza do teste, que tende a detectar várias pequenas ocorrências espalhadas durante um dia. A Tabela 7 mostra em detalhes a quantidade de ocorrências detectada por cada regra para cada um dos inversores.

Tabela 7 – Ocorrências registrada para cada regra e cada inversor durante os 30 dias

Inversor	**Regra 1**	**Regra 2**	**Regra 3**
1	14	1	119
2	17	1	109
3	8	0	38
4	6	1	48
5	9	0	23
6	3	0	8
7	2	1	0
8	2	0	0
9	2	1	8
10	0	2	6
11	0	1	41
12	0	0	7
13	0	0	0
14	1	0	0
15	2	0	0
16	0	0	0
17	3	0	28
18	1	2	0
19	13	0	29
20	7	0	12
21	7	0	0
22	10	0	0
23	1	0	0
24	1	0	0
25	11	0	0
26	10	0	0
27	14	0	0
28	11	0	0
29	7	0	0
30	4	0	5
31	3	0	0
32	2	0	5

Fonte: Autor

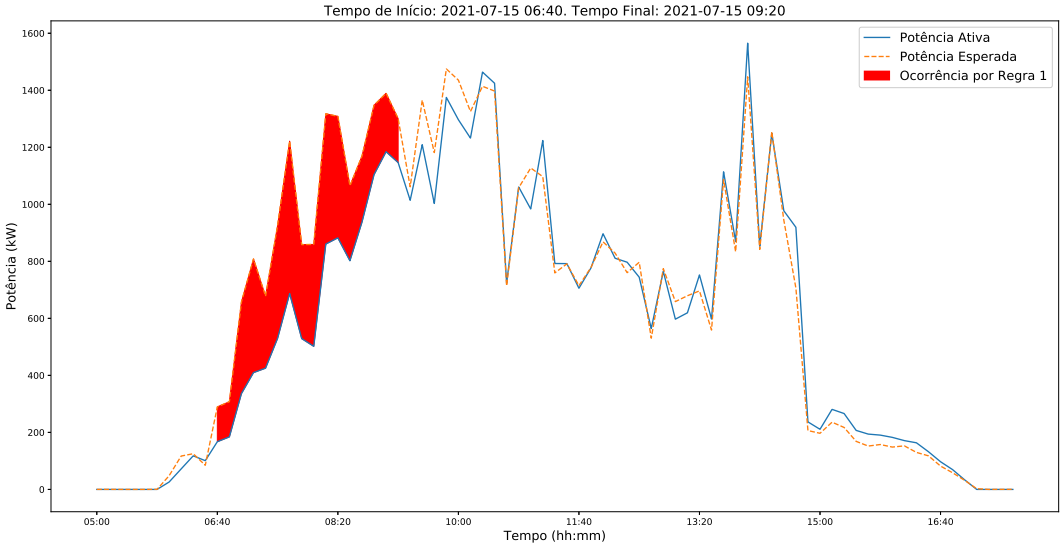
Ademais, para cada uma das regras, exemplificam-se a seguir algumas

ocorrências detectadas.

5.4.1.1 Regra 1

Para a regra de perda momentânea, destaca-se um exemplar pela Figura 21.

Figura 21 – Exemplo de ocorrência por Regra 1



Fonte: Autor

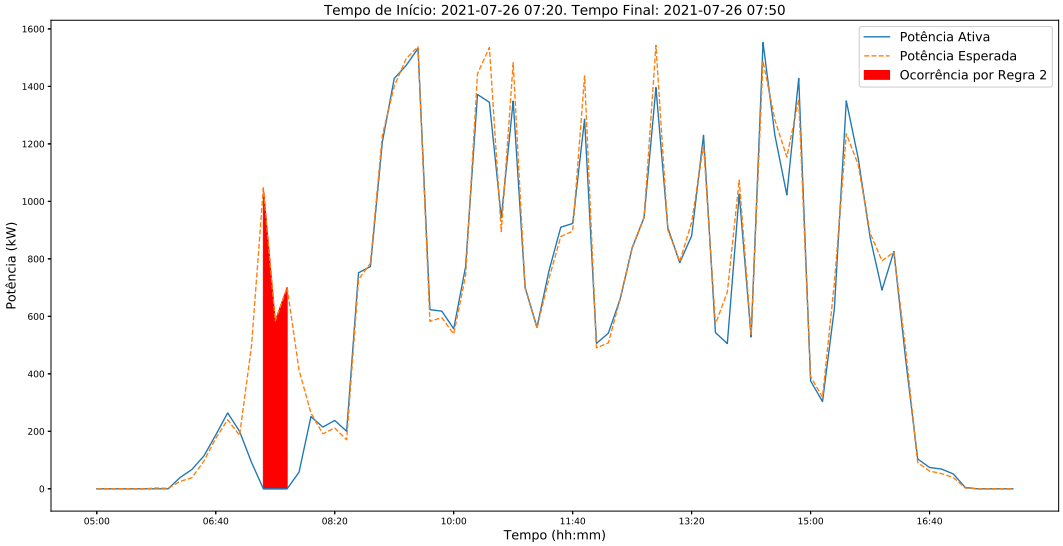
Na figura, é possível observar que, de fato, houve uma perda de performance que perdurou mas logo retomou patamares normais. Na maioria dos casos, sombreamentos momentâneos podem ser a causa principal para este tipo de ocorrência. No entanto, caso episódios com este se tornem mais recorrentes, outras fontes de falhas devem ser investigadas, como por exemplo falhas em rastreadores.

5.4.1.2 Regra 2

Para a regra de perda total, a Figura 22 exemplifica.

Este tipo de falha é mais facilmente detectada pelo sistema SCADA. Assim, o sistema proposto atuará como uma redundância, para caso o sistema principal falhe ao detectar o ocorrido. Esse tipo de comportamento geralmente acontece por falhas gerais nos circuitos internos ou paradas para manutenção de equipamentos da operação.

Figura 22 – Exemplo de ocorrência por Regra 2

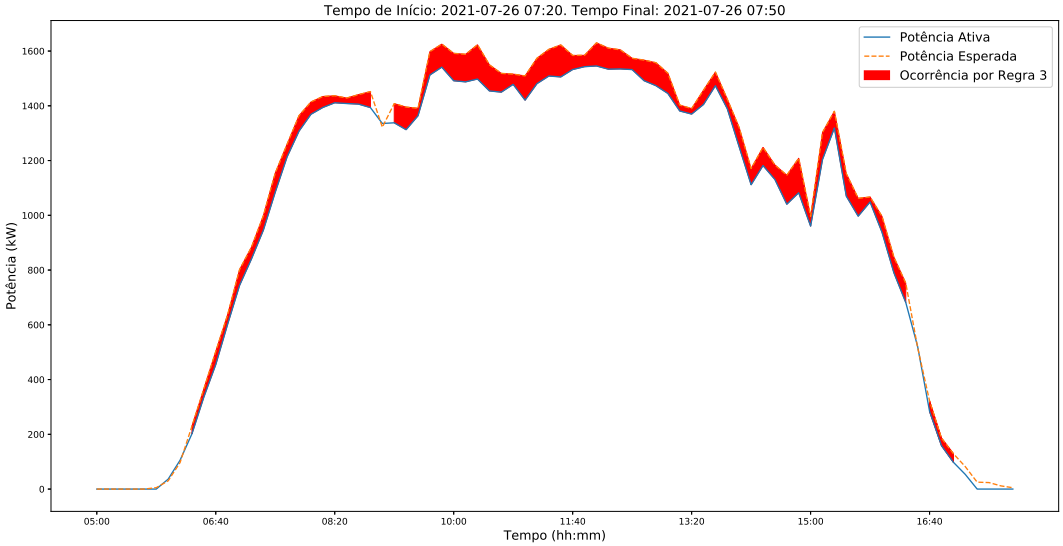


Fonte: Autor

5.4.1.3 Regra 3

Por último, a Figura 23 demonstra um comportamento detectado pela terceira regra.

Figura 23 – Exemplo de ocorrência por Regra 3



Fonte: Autor

Como é possível observar, o comportamento de baixa performance perdura, dissolvendo-se por grande parte do dia. Esse episódio requer uma atenção especial, podendo indicar fenômenos de natureza constante como sujeira em painéis ou falhas em componentes

da operação, tais como painéis ou mesmo na caixa de junção.

5.4.2 Interface Gráfica

Para exibir todos os resultados do Estudo de Caso e exemplificar um uso prático simulando uma operação real, um exemplo de interface visual básica foi desenvolvido. Na página inicial, estão destacadas as opções de páginas disponíveis.

Cada página será exemplificada a seguir.

5.4.2.1 Tabela de Ocorrências

Página que permite a observação de uma tabela contendo as ocorrências registradas para todos os inversores dentro de um intervalo de tempo. Esta tabela poderá ser filtrada para inversores específicos, para uma regra específica ou intervalo de tempo a ser considerado.

Assim, é possível ter uma observação direta de quais e quantas ocorrências foram registradas na operação, exemplificada pela Figura 24.

Figura 24 – Página com a tabela de ocorrências e filtros para exibição

	Inversor	Início	Fim	Duração	Regra
110	9	2021/07/01 10:40:00	2021/07/01 11:50:00	0 days 01:10:00	1
91	7	2021/07/02 06:20:00	2021/07/02 06:50:00	0 days 00:30:00	2
566	3	2021/07/02 09:20:00	2021/07/02 10:30:00	0 days 01:10:00	1
303	1	2021/07/02 09:20:00	2021/07/02 10:40:00	0 days 01:20:00	1
101	30	2021/07/03 05:50:00	2021/07/03 07:00:00	0 days 01:10:00	1
455	2	2021/07/03 06:10:00	2021/07/03 07:00:00	0 days 00:50:00	3
456	2	2021/07/03 08:00:00	2021/07/03 09:50:00	0 days 01:50:00	3
457	2	2021/07/03 10:20:00	2021/07/03 10:40:00	0 days 00:20:00	3
458	2	2021/07/03 10:50:00	2021/07/03 11:30:00	0 days 00:40:00	3
459	2	2021/07/03 11:40:00	2021/07/03 12:00:00	0 days 00:20:00	3
460	2	2021/07/03 14:20:00	2021/07/03 16:40:00	0 days 02:20:00	3

Fonte: Autor

Como especificado anteriormente, cada ocorrência será caracterizada por um

tempo de início, tempo de fim e a regra que a gerou.

5.4.2.2 Séries Temporais

Página que exibe séries temporais de Potência Ativa e Potência Esperada por inversor para um intervalo de tempo específico, além de destacar visualmente as ocorrências presentes nesse período.

A Figura 25 demonstra um exemplo de série temporal para a página.

Figura 25 – Página com a exibição de séries temporais e filtros

Fonte: Autor

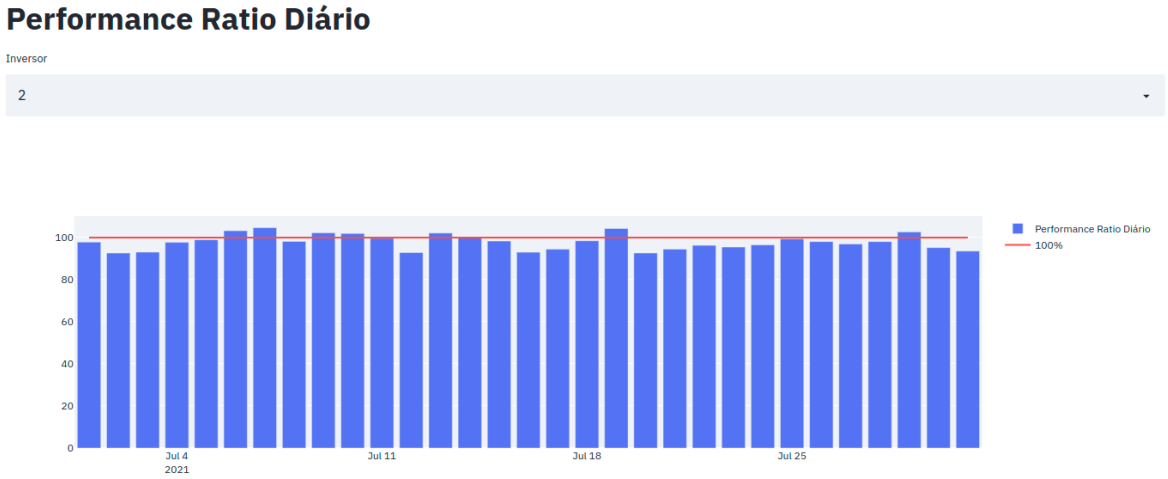
Esta análise permite que o usuário tenha uma melhor compreensão do comportamento de um inversor específico ao longo de um dia.

5.4.2.3 PR Diário

Esta análise busca consolidar o cálculo de PR diariamente e fazer uma comparação diária, observando se este se aproxima do 100% para todos os dias.

A análise representada pela Figura 26 dá uma visão mais geral da operação de um inversor específico, espaçado durante o mês completo, facilitando a observação de

Figura 26 – Página com a exibição dos cálculos de PR consolidados diariamente, para um inversor específico



Fonte: Autor

falhas mais duradouras.

5.4.2.4 Dispersão entre Potência Real e Potência Esperada

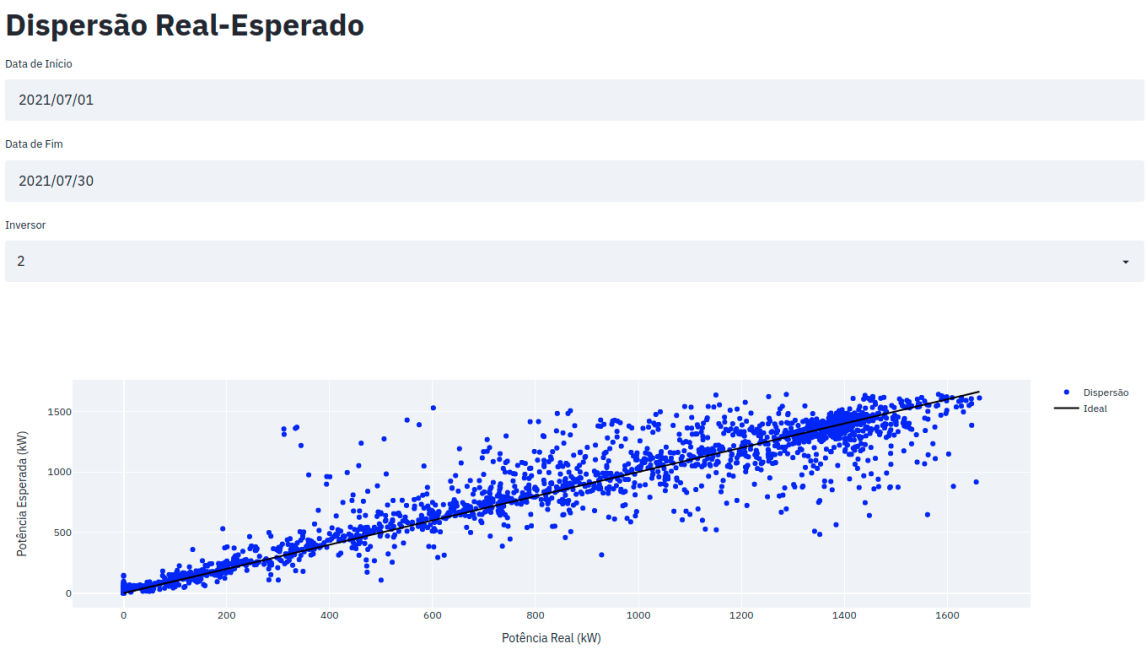
Por fim, esta página mostra o grau de dispersão entre o real e o esperado para cada instante de tempo analisado, como exibido na Figura 27.

Esta configura-se, portanto, como mais uma ferramenta para analisar dispersões pontuais no tempo para um inversor específico.

5.5 Comentários Parciais

Neste capítulo foram demonstrados e debatidos todos os resultados obtidos quanto à comparação entre modelos para a estimação da energia esperada. Além disso, uma aplicação prática do sistema de acompanhamento proposto demonstraram resultados satisfatórios no objetivo de utilizar regras simples para a identificação de episódios de baixa performance em plantas FV.

Figura 27 – Página com a exibição de dispersão e filtros



Fonte: Autor

6 CONCLUSÕES

De posse dos resultados obtidos e analisados, tornou-se possível a obtenção de diversos modelos capazes de aproximar a energia esperada de um sistema FV, com base em seu comportamento histórico. Assim, o trabalho apresenta uma maneira viável de avaliar a performance de um sistema real, comparando seu desempenho com uma representação histórica.

No mesmo contexto, a metodologia proposta permitiu uma comparação entre os modelos estudados, de forma a determinar qual deles melhor adequa-se às características dos dados, observando métricas objetivas e apresentando conclusões baseadas na interpretação destas. Com destaque, o algoritmo de RF apresentou as melhores métricas de validação, apresentando-se como a técnica mais adequada para a estimação de energia esperada, dentre os métodos comparados.

Ademais, o Sistema de Regras para monitoramento de performance mostrou-se capaz de identificar ocorrências de forma automática utilizando suas regras baseadas em potência esperada, classificando estas ineficiências por meio das três regras em um Estudo de Caso, que mostrou exemplos práticos de aplicação do sistema proposto.

Finalmente, a interface gráfica auxilia a interpretabilidade do sistema, exemplificando uma aplicação visual para acompanhamento constante de performance, fortalecendo a viabilidade da Metodologia proposta para aplicações em usinas reais, auxiliando o acompanhamento operacional dessas plantas fotovoltaicas.

6.1 Trabalhos Futuros

Como sugestões para trabalhos futuros, é possível destacar os seguintes direcionamentos que objetivam um amadurecimento do tema trabalhado:

- Como sugere a Figura 14, pode ser relevante explorar técnicas de *oversampling*, visando a criação virtual de mais dados para utilização durante treinamento do modelo, além de buscar uma distribuição estatística mais equilibrada para as variáveis utilizadas;
- Considerar os efeitos de sazonalidades dentro do conjunto de dados, efeitos estes que foram desconsiderados durante o presente trabalho;
- Comparação de uma gama maior de modelos físicos, de forma a enriquecer

ainda mais a discussão entre modelos físicos e técnicas de Aprendizagem de Máquina;

- Evolução do projeto para utilização de técnicas de *forecasting*, passando a utilizar técnicas capazes de realizar predição para tempos futuros, com base em comportamentos históricos;
- Gerar novas regras de avaliação da performance da geração utilizando técnicas de Aprendizagem de Máquina (ex.: Árvores de Decisão).

REFERÊNCIAS

ABSOLAR: *Infográfico ABSOLAR, Atualizado em 01/08/2021.* 2021. – URL <https://www.absolar.org.br/mercado/infografico/>

Al-Dahidi, Sameer ; Ayadi, Osama ; Adeeb, Jehad ; Louzazni, Mohamed: Assessment of Artificial Neural Networks Learning Algorithms and Training Datasets for Solar Photovoltaic Power Production Prediction. In: *Frontiers in Energy Research* 7 (2019), p. 130

Al-Dahidi, Sameer ; Louzazni, Mohamed ; Omran, Nahed: A Local Training Strategy-Based Artificial Neural Network for Predicting the Power Production of Solar Photovoltaic Systems. In: *IEEE Access* 8 (2020), p. 150262–150281

Chollet, Francois et al.: *Keras*. 2015. – URL <https://github.com/fchollet/keras>

CRESESB. Centro de Referência para as Energias Solar e Eólica Sérgio de S. Brito. Manual de Engenharia para Sistemas Fotovoltaico. Rio de Janeiro, Brasil, mar. 2014

Dodge, Y.: *The Concise Encyclopedia of Statistics*. Springer New York, 2008 (The Concise Encyclopedia of Statistics)

Gurupira, Tafadzwa ; Rix, Arnold: PV Simulation Software Comparisons: PVSYST, NREL SAM and PVlib, 2017

Han, Jiawei ; Kamber, Micheline ; Pei, Jian: *Data Mining: Concepts and Techniques*. 3rd. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 2011

Hastie, Trevor ; Tibshirani, Robert ; Friedman, Jerome: *The Elements of Statistical Learning*. New York, NY, USA : Springer New York Inc., 2001 (Springer Series in Statistics)

Heaton, Jeff: *Introduction to Neural Networks for Java, 2nd Edition*. Heaton Research, Inc., 2008

Holmgren, William F. ; Hansen, Clifford W. ; Mikofski, Mark A.: pvlib python: a python package for modeling solar energy systems. In: *Journal of Open Source Software* 3 (2018), n. 29, p. 884

Hyndman, Rob J. ; Koehler, Anne B.: Another look at measures of forecast accuracy. In: *International Journal of Forecasting* 22 (2006), n. 4, p. 679–688

IEC 61724-1. International Electrotechnical Commission. Photovoltaic system performance - Part 1: Monitoring. Geneva, Switzerland, mar. 2017

IEC TS 61724-3. International Electrotechnical Commission. Photovoltaic system performance - Part 3: Energy evaluation method. Geneva, Switzerland, jul. 2016

IRENA. International Renewable Energy Agency. Global energy transformation: A roadmap to 2050 (2019 edition). Abu Dhabi, 2019

Kaaya, Ismail ; Ascencio-Vásquez, Julián: Solar Radiation - Measurements, Modeling and Forecasting for Photovoltaic Solar Energy Applications. In: *Photovoltaic Power Forecasting Methods* (2021)

Khatri, Pooja: Review of SCADA system for photovoltaic power plant. (2018)

Pedregosa et al.: Scikit-learn: Machine Learning in Python. In: *Journal of Machine Learning Research* 12 (2011), p. 2825–2830

PVsyst: *PVsyst - Photovoltaic Software*. – URL <https://www.pvsyst.com/>

Reis, Pedro: *Como funcionam as células solares fotovoltaicas*. 2015. – URL <https://www.portal-energia.com/como-funcionam-celulas-solares-componentes-operacoes/>

Rodrigues, Sandy ; Ramos, Helena ; Morgado-Dias, F.: Machine learning PV system performance analyser. In: *Progress in Photovoltaics: Research and Applications* 26 (2018), 08, p. 675–687

Sehmi, Arvindra ; Shukla, Ashish et al.: *Streamlit*. – URL <https://docs.streamlit.io/>

Shalev-Shwartz, Shai ; Ben-David, Shai: *Understanding Machine Learning: From Theory to Algorithms*. USA : Cambridge University Press, 2014

Simal Pérez, Joaquín Alonso-Montesinos ; Batlles, Francisco J.: Estimation of Soiling Losses from an Experimental Photovoltaic Plant Using Artificial Intelligence Techniques. In: *Applied Sciences* 11 (2021), n. 4, p. 1516

Will Holmgren, Calama-Consulting et al.: *pvlib/pvlib-python: v0.8.1*. 2021. – URL <https://doi.org/10.5281/zenodo.4417742>