Eduardo Henrique Marques Ferreira

Avaliação de Técnicas de Aprendizado de Máquina para Detecção de Eventos em Sistemas Elétricos de Grande Porte a Partir de Dados de Medição Fasorial Sincronizada

Brasil

Fevereiro de 2024

Eduardo Henrique Marques Ferreira

Avaliação de Técnicas de Aprendizado de Máquina para Detecção de Eventos em Sistemas Elétricos de Grande Porte a Partir de Dados de Medição Fasorial Sincronizada

Monografia

Universidade Federal de Itajubá – UNIFEI
Programa de Pós-Graduação em Ciência e Tecnologia Computação

Orientador: Carlos Henrique Valério de Moraes
Coorientador: Alexandre Rasi Aoki

Brasil
Fevereiro de 2024

Este trabalho é dedicado à minha família, especialmente à minha esposa Kênia Aparecida Ferreira e minha filha Clarisse Marques Ferreira, fontes inesgotáveis de força e perseverança em minha jornada. Um agradecimento especial à minha mãe, Nilva Garcia Marques, e ao meu irmão, José da Silva Ferreira Junior, cujo apoio incondicional e palavras sábias foram faróis de orientação e conforto nos momentos mais desafiadores de estudo e exaustão. Estendo essa homenagem aos estimados professores do programa, cujo conhecimento e dedicação não apenas iluminaram, mas também enriqueceram meu caminho acadêmico. Este trabalho é o resultado de um amor e sacrifício coletivos, aos quais expresso minha mais profunda e sincera gratidão.

Agradecimentos

Agradeço primeiramente a Deus, fonte de toda sabedoria e entendimento, que me sustentou com o dom do conhecimento e a constante inspiração no aprendizado. Em momentos de desafio e dúvida, foi a fé que me guiou e permitiu que eu continuasse a trilhar este caminho com perseverança e esperança.

Um sincero e especial agradecimento ao Prof. Carlos Henrique Valério de Moraes, meu orientador, por sua inestimável orientação, paciência e conhecimento. Sua experiência e dedicação foram fundamentais para a realização deste trabalho. Da mesma forma, sou imensamente grato ao Prof. Alexandre Rasi Aoki, coorientador, cuja expertise e conselhos práticos enriqueceram significativamente minha jornada acadêmica.

Estendo meus agradecimentos a todo o corpo docente do Programa de Pós-graduação em Ciência e Tecnologia da Computação. Cada professor contribuiu para a minha formação e me proporcionou ferramentas essenciais para o desenvolvimento profissional e pessoal.

“Em algum lugar,
algo incrível está esperando para ser conhecido.
(Carl Sagan)

Resumo

A crescente necessidade de garantir a eficiência e a confiabilidade do Sistema Integrado Nacional (SIN), que representa a espinha dorsal da distribuição de energia do país é o que conduz este trabalho. Diante dos desafios impostos pela vastidão territorial e a complexidade da rede elétrica brasileira, a utilização de Unidade de Medição Fasorial (PMUs) emerge como uma solução promissora para monitorar a rede em tempo real. No entanto a eficácia deste monitoramento está ligada à capacidade de detectar e responder a anomalias de forma rápida e precisa, minimizando os riscos de falhas e interrupções no fornecimento. O estudo aborda o desafio de gerenciamento e avaliação da rede elétrica interligada do país onde a informação precisa é pertinente para ações preventivas e corretivas em um sistema de distribuição de extensão continental. O cerne desta pesquisa reside na exploração inovadora de técnicas avançadas de compressão de dados combinadas com algoritmos de aprendizado de máquina não supervisionado, com o intuito de otimizar a interpretação e análise dos grande volumes de dados gerados pelas PMUs. Essa abordagem aponta uma melhoria significativa na qualidade e na precisão das informações extraídas e oferece uma solução escalável para o desafio de processar e analisar os dados de um sistema de distribuição de grande escala. São avaliados a eficácia através da detecção de eventos significativos e anomalias na rede em pontos geograficamente distantes da origem do evento. Os resultados deste estudo validam a eficácia dos algoritmos propostos, evidenciando sua relevância prática e seu impacto substancial na melhoria dos padrões de qualidade e confiabilidade no fornecimento de energia elétrica. Demonstram a viabilidade de implementação dessas técnicas em cenários reais, evidenciando o potencial de transformação na prevenção de falhas e na gestão de eventos críticos, contribuindo para uma rede de distribuição de energia mais estável, eficaz e segura.

Palavras-chave: PMU. SIN. Aprendizado de máquina não supervisionado.

Abstract

The growing need to ensure the efficiency and reliability of the National Interconnected System (SIN), which represents the backbone of the country's energy distribution, is what drives this work. Faced with the challenges posed by the territorial vastness and complexity of the Brazilian electrical grid, the use of Phasor Measurement Units (PMUs) emerges as a promising solution for real-time network monitoring. However, the effectiveness of this monitoring is linked to the ability to detect and respond to anomalies quickly and accurately, minimizing the risks of failures and supply interruptions. The study addresses the challenge of managing and evaluating the country's interconnected electrical grid, where accurate information is pertinent for preventive and corrective actions in a continental-scale distribution system. The core of this research lies in the innovative exploration of advanced data compression techniques combined with unsupervised machine learning algorithms, aiming to optimize the interpretation and analysis of the large volumes of data generated by the PMUs. This approach points to a significant improvement in the quality and precision of the information extracted and offers a scalable solution to the challenge of processing and analyzing the data from a large-scale distribution system. The effectiveness is assessed through the detection of significant events and anomalies in the network at points geographically distant from the event's origin. The results of this study validate the efficacy of the proposed algorithms, highlighting their practical relevance and substantial impact on improving the quality and reliability standards in the electricity supply. They demonstrate the feasibility of implementing these techniques in real scenarios, showcasing the potential for transformation in failure prevention and critical event management, contributing to a more stable, effective, and secure energy distribution network.

Keywords: PMU. SIN. Unsupervised Machine Learning.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ACP	Análise de componentes principais
ADTK	Anomaly Detection Toolkit
ANEEL	Agência Nacional de Energia Elétrica
AP	Precisão média
AUC	Área sob a curva
CPS	Completeness Score
CSV	Comma-Separated Values
DL	Deep Learning
FMS	Fowlkes-Mallows Score
FN	Falso negativo
FP	Falso positivo
GPS	Global Positioning System
HCA	Análise de cluster hierárquica
HGS	Homogeneity Score
KB	Kilobytes
LLE	Locally Linear Embedding
LOF	LocalOutilierFactor
MAE	Erro médio absoluto
MAPE	Mean Absolute Percentage Error

MIS	Mutual Information Score
ML	Machine Learning
MSE	Erro quadrático médio
ONS	Operador Nacional do Sistema Elétrico
PDC	Concentrador de dados fasoriais
PMU	Phasor Measurement Unit
RMSE	Root Mean Squared Error
ROC	Característica de oepração do receptor
ROCOF	Taxa de variação da frequência
SCADA	Supervisory Control and Data Acquistion
SIN	Sistema Interligado Nacional
TNR	Taxa de verdadeiros negativos
TPR	Taxa de verdadeiros positivos
VN	Verdadeiro Negativo
VP	Verdadeiro positivo
WAMS	Sistemas de medição de ampla área

Sumário

1 Introdução

Desde o final do século XIX, o setor elétrico brasileiro desempenha um papel fundamental no progresso e na evolução do país. Seu impacto pode ser observado em diferentes esferas da sociedade, impulsionando o crescimento da indústria, do comércio e do bem-estar dos lares, ao suprir as demandas por melhorias e qualidade de vida. A indústria experimentou um notável crescimento, saindo de uma fase inicial tímida, com o foco na fabricação de açúcar e mineração, para se tornar uma indústria de transformação (GOMES et al., 2002) de bens de consumo. Os domicílios, por sua vez, passaram a desfrutar de maior conforto e comodidade.

O setor elétrico brasileiro revela-se crucial no desenvolvimento da sociedade como um todo, assumindo o papel de protagonista que impulsiona o progresso. Para tanto, utiliza recursos tecnológicos avançados, como equipamentos, projetos e designs. Essa característica resulta em um setor diversificado, com diferentes formas de geração e transmissão de energia, o que impõe desafios concretos a serem superados.

Um dos desafios enfrentados é a necessidade de ferramentas mais eficientes e robustas para garantir alta qualidade na análise e monitoramento em tempo real das condições de operação do sistema elétrico. Com a grande expansão da carga e evolução dos sistemas de potência, é imprescindível que os reguladores tenham acesso a dados precisos.

O extenso sistema de produção e transmissão de energia elétrica no Brasil é notavelmente caracterizado por sua abordagem hidro-termo-eólica, onde as usinas hidrelétricas predominam e uma variedade de proprietários desempenha um papel ativo. Essa complexa estrutura é interconecta por meio do Sistema Interligado Nacional (SIN), que abrange quatro subsistemas: Sul, Sudeste/Centro-Oeste, Nordeste e a maior parte da região Norte.

A intrincada rede de transmissão estabelece uma ligação crucial entre diversas fontes de geração de energia, assegurando o fornecimento necessário para atender as demandas dos consumidores. A figura 1 exemplifica claramente essa

interconexão, ilustrando a relação entre as fontes geradoras e os pontos de consumo em todo o país. Isso não somente permite a obtenção de ganhos sinérgicos, mas também aproveita a diversidade dos regimes hidrológicos das diferentes bacias (ONS, 2023c).

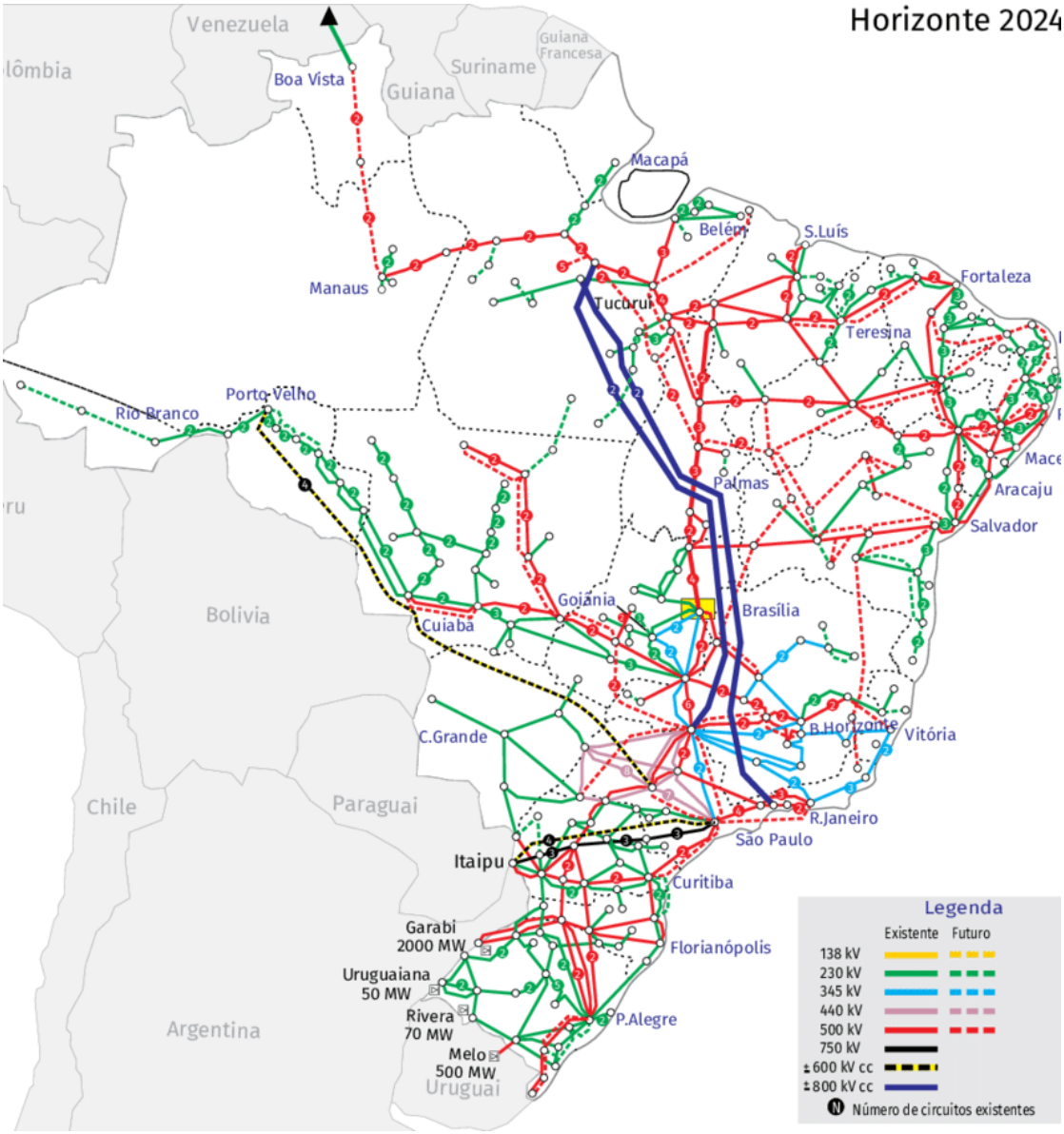

Figura 1 – Mapa do Sistema Interligado Nacional (SIN) (ONS, 2023c)

O Operador Nacional do Sistema Elétrico (ONS) elaborou critérios conhe-

cidos como Procedimentos de Rede (ONS, 2023d), homologados pela Agência Nacional de Energia Elétrica (ANEEL). Esses critérios definem parâmetros para a operação e monitoramento em tempo real, utilizando uma extensão do sistema SCADA (*Supervisory Control and Data Acquisition*). Esse sistema fornece valores eficazes das medidas de tensão, fluxo de potência ativa e reativa nas linhas de transmissão, bem como os valores de injeção de potência ativa e reativa nas barras do sistema monitorado (RIBEIRO, 2022).

O sistema SCADA utiliza uma taxa de atualização de medidas a cada cinco minutos, podendo esta ser configurada mediante a necessidade. No entanto, o desenvolvimento de novas tecnologias, como o GPS (*Global Positioning System*) colabora com o surgimento de novos equipamentos para medição fasorial sincronizado. O sistema SCADA permite medir instantaneamente o módulo e o ângulo dos fasores de tensão e corrente nas três fases da rede elétrica, bem como variações da frequência (PHADKE, 2002).

Com o advento das medições fasoriais, tornou-se possível considerar o controle baseado no valor medido de quantidades remotas, em vez de apenas em sinais locais. De acordo com Phadke e Bi (2018), essas medições permitem encontrar a diferença entre os estados do sistema que estamos realmente controlando e o estado de um modelo. Assim, o controle preditivo com retorno fasorial pode ser usado para resolver o problema de controle ótimo não linear.

Com a análise eficiente das medições fasoriais adquiridas pela *Phasor Measurement Unit* (PMU) utilizando algoritmos adequados de Inteligência Artificial, o sistema pode detectar anomalias como interrupções na linha, geração e falhas, de acordo com Amutha et al. (2021). Essa previsão tornará o sistema mais seguro e robusto ao aplicar medidas corretivas e planejar ações a serem tomadas em caso de falhas.

1.1 Motivação

Para este estudo, é utilizado dados provenientes de Unidades de Medição Fasorial (PMUs) cujo seu funcionamento é detalhado na figura 2. As PMUs fornecem medições a cada ciclo, resultando em 60 leituras por segundo. Isso equivale a um

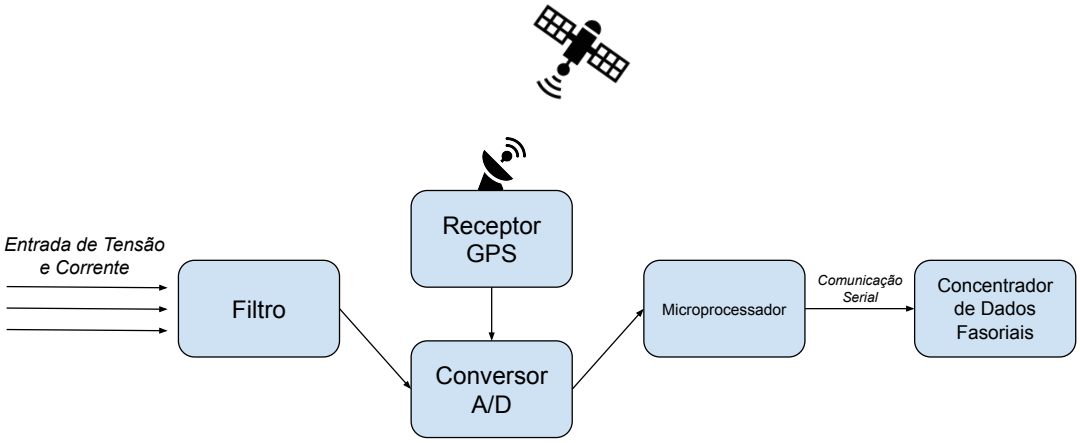

Figura 2 – Funcionamento de uma PMU
fonte: Adaptado de (PHADKE, 1993)

total de 216.000 leituras por hora, 5.184.000 por dia e 36.288.000 por semana. Cada leitura pode conter até 30 parâmetros de medição. Considerando os ciclos coletados em um único dia, temos acima de 5.000.000 de registros a serem analisados e utilizados no treinamento do sistema. Essa quantidade massiva de dados permite que o sistema identifique e aponte possíveis falhas tanto dentro quanto fora do território da concessionária.

A principal motivação deste trabalho é utilizar algoritmos de detecção de anomalias para analisar os dados coletados pelas PMUs e reforçar a importância da estabilidade e segurança do sistema elétrico. Com o aumento da demanda por energia e a integração de fontes renováveis de energia, é fundamental garantir que o sistema elétrico funcione de maneira eficiente e confiável.

Os dados coletados pelas PMUs fornecem informações valiosas sobre o estado do sistema elétrico em tempo real. No entanto, a quantidade de dados gerados pode ser esmagadora e difícil de ser analisada. Nesse cenário, a utilização de algoritmos de detecção de anomalias pode ajudar a identificar padrões incomuns nos dados e alertar os operadores do sistema sobre possíveis falhas, prevenindo ou acionando medidas de contingência nesses casos.

A exploração de diferentes algoritmos de detecção de anomalias é um campo

de pesquisa a ser explorado para apresentar resultados relevantes para a tomada de decisões. Esta análise apresenta resultados da avaliação da eficácia na detecção de anomalias pela análise dos dados, fornecendo ferramentas mais eficazes para o monitoramento do sistema elétrico.

1.1.1 Hipótese

Diante do exposto, a proposta deste trabalho, visa responder a seguinte questão: "É possível detectar anomalias dentro e fora do sistema elétrico da concessionária de energia com PMUs via algoritmos de Inteligência Artificial?"

Para responder esta questão, este trabalho pretende analisar os algoritmos de Inteligência Artificial com foco em detecção de anomalias, avaliando-os para identificar qual apresenta melhor análise e melhor assertividade na indicação de possíveis problemas.

1.2 Objetivos

O objetivo deste trabalho é realizar a detecção de eventos pelas medições das PMUs utilizando algoritmos de Inteligência Artificial no contexto de Detecção de Anomalias. Essa abordagem promissora pode solucionar o desafio proposto, à medida que a infraestrutura do sistema elétrico se torna mais complexa e interconectada. A identificação rápida e precisa de eventos anormais é essencial para garantir a estabilidade e confiabilidade do sistema como um todo.

A pesquisa foi dividida em objetivos específicos abaixo:

- Realizar a análise dos dados e compreender a melhor abordagem de leitura e estudo desses dados.
- Construir modelos de aprendizado de máquina com capacidade de identificar anomalias nos dados coletados pelas PMUs utilizando algoritmos de aprendizados de máquina não supervisionados.
- Realizar análise qualitativa dos diversos modelos aplicados, com o intuito de identificar a melhor abordagem para a detecção de eventos anormais.

Com esses objetivos específicos são essenciais para orientar o desenvolvimento da pesquisa e permitir uma análise completa dos dados e dos algoritmos aplicados. Ao atingir esses objetivos, este trabalho pode contribuir para a melhoria da estabilidade e confiabilidade do sistema elétrico, proporcionado detecção e respostas eficientes a eventos.

1.3 Organização do Documento

Neste capítulo, foram apresentadas as motivações que levaram ao desenvolvimento deste trabalho. Os próximos capítulos estão organizados da seguinte forma:

- **Capítulo 2: Referencial Teórico.** Apresenta o embasamento teórico sobre o assunto, incluindo as características intrínsecas à detecção de falhas em uma infraestrutura de rede local de grande porte.
- **Capítulo 3: Desenvolvimento.** Apresenta um referencial teórico sobre as abordagens de aprendizado de máquina supervisionado e não supervisionado, além de uma revisão bibliográfica dos trabalhos mais relevantes relacionados à detecção de falhas ou anomalias em redes de computadores.
- **Capítulo 4: Experimentos e Discussões.** Apresenta a definição dos experimentos realizados e os resultados obtidos, juntamente com uma discussão a respeito desses resultados.
- **Capítulo 5: Conclusão e trabalhos futuros.** Apresenta uma conclusão sobre o problema abordado, buscando responder à questão principal deste trabalho, que foi exposta no Capítulo 1. Além disso, são propostas ideias para trabalhos futuros relacionados ao tema.

2 Referencial Teórico

Este capítulo abordará as principais metodologias e técnicas de relevante importância para a detecção de eventos por meio de algoritmos de inteligência artificial em PMUs, apresentando as técnicas e conceitos dos algoritmos. Em seguida serão apresentadas algumas ferramentas relacionadas ao tema.

2.1 Considerações Iniciais

O setor elétrico brasileiro desempenha um papel crucial no desenvolvimento da sociedade, assumindo o protagonismo que impulsiona o progresso. Para isso, faz uso de recursos tecnológicos avançados, tornando um setor bem diversificado, abrangendo diferentes formas de geração e transmissão de energia. A busca por soluções e avanços nesse setor é essencial para atender às crescentes demandas energéticas da sociedade, garantindo um fornecimento confiável e eficiente superando os desafios impostos.

A aplicação de recursos tecnológicos proporciona uma nova perspectiva na captação, análise e estudo de informações relevantes para as redes elétricas. Esses dados processados são de grande valia estratégica, permitindo a prevenção, manutenção e tomada de decisões eficientes, o que se resulta em economia de recursos humanos, técnicos e financeiros.

2.2 Unidade de Medição Fasorial (PMU)

Um dos recursos tecnológicos utilizados para análise e controle da rede elétrica é a PMU ou Unidade de Medição Fasorial (*Phasor Measurement Unit*). Estes é um dispositivo avançado de monitoramento usado na indústria de energia elétrica, onde registra de forma precisa e sincronizada várias grandezas elétricas analógicas, como tensão, corrente, frequência e ângulo de fase convertendo-as em valores digitais possibilitando seu armazenamento e análise com precisão (LIU; BI;

YANG, 2013).

A PMU faz parte de um arcabouço de modernização do setor elétrico, implementando tecnologias de redes inteligentes com a capacidade de monitoramento, proteção e controle de uma grande área da rede elétrica (GREER et al., 2014), alterando o conceito desta rede, onde a composição era formada apenas por componentes físicos, para um sistema *ciberfísico* (LEGER; JAMES, 2018).

As PMUs são projetos com relógios de tempo altamente precisos, com a utilização do Sistema de Posicionamento Global (GPS), o que garante que todas as medições sejam sincronizadas em um único horário de referência. Faz-se importante tamanha precisão para garantir a consistência das medições em diferentes pontos da rede elétrica, permitindo uma análise precisa e uma compreensão aprofundada do comportamento do sistema.

Com sua capacidade de medir e registrar grandezas elétricas com taxas de amostragem de trinta à sessenta observações por segundo, as PMUs fornecem uma visão em tempo real do estado do sistema de energia elétrica. Destaca-se a detecção de variações na carga e ocorrências de falhas na rede. Essas medições de alta precisão permitem uma análise detalhada das respostas de tensão, ângulo de fase e frequência do sistema, gerando conhecimento valioso para os operadores do sistema, fornecendo uma visão detalhada do comportamento do sistema elétrico no aprimoramento da tomada de decisão dos operadores do sistema.

As medidas captadas permitem uma análise mais precisa e uma detecção mais rápida de eventos anormais, como oscilações de frequência ou variações abruptas de tensão, permitindo uma resposta mais eficiente e uma melhor gestão do sistema elétrico, aumentando a confiabilidade e a segurança da rede. Instalados em diferentes locais para monitoramento, os dados coletados por PMUs são enviados para um concentrador de dados fasoriais (PDC), onde são agregados e combinados com base no instante de tempo registrado pelo GPS. Estes dados consolidados são analisados e utilizados para tomada de decisões operacionais.

2.3 Aprendizado de Máquina

A inteligência é um conceito complexo que tem sido objeto de estudo por filósofos, cientistas e psicólogos ao longo dos séculos. No dicionário Michaelis (MICHAELIS, 2023), a inteligência é definida como a faculdade de entender, pensar, raciocinar e interpretar; é ter entendimento, intelecto e percepção de algo; e a capacidade de compreender e utilizar informações para resolver problemas. Essa definição abrange uma série de processos mentais complexos, como memória, raciocínio e criatividade, que são características intrínsecas dos seres humanos.

Nesta seção, serão abordadas as definições de inteligência artificial e aprendizado de máquina, bem como os diferentes tipos de aprendizado associados a essa área. Serão detalhados os conceitos fundamentais que serão utilizados ao longo deste trabalho, fornecendo uma base sólida para a compreensão desses campos em constante evolução.

2.3.1 Inteligência Artificial

A inteligência artificial (*Artificial Intelligence*) é um campo da ciência da computação dedicado à criação de agentes inteligentes, que são sistemas capazes de aprender e agir de forma autônoma. Essa área abrange uma ampla gama de conhecimentos e técnicas com o objetivo de simular a inteligência humana, explorando habilidades como processamento de linguagem natural, representação e armazenamento de conhecimento, raciocínio automatizado e aprendizado de máquina para adaptação a novas circunstâncias e detecção de padrões. Quando um sistema é capaz de demonstrar essas habilidades, ele é considerado possuidor de uma inteligência satisfatória (RUSSELL, 2010).

O marco mais significativo na história da inteligência artificial foi o desenvolvimento do Teste de Turing. O Teste de Turing é uma avaliação que visa determinar se uma máquina pode ser considerada inteligente e foi proposto por Alan Turing em seu artigo de 1950 intitulado "*Computing Machinery and Intelligence*"(TURING, 1950). Neste artigo Turing discute a possibilidade máquinas exibirem comportamento inteligente e propõe um critério prático de avaliação. O Teste desafia uma máquina a realizar uma conversa de forma convincente, onde

um observador humano não possa distinguir se as respostas são fornecidas por uma máquina ou por um ser humano, impactando profundamente o campo da inteligência artificial e influenciando pesquisas e o desenvolvimento na área até os dias atuais.

Em seu livro, Russell (2010) traça uma linha do tempo que ilustra a evolução da inteligência artificial desde os anos 1950 até os dias atuais. Essa linha do tempo revela os marcos cruciais e a expansão da aplicabilidade dessa ciência em diversas áreas de pesquisas. Destaca-se avanços significativos em campos como visão computacional, processamento de linguagem natural e robótica. Atualmente, a inteligência artificial é caracterizada por uma ampla gama de técnicas e ferramentas que são impulsionadas pela evolução da capacidade computacional pela disponibilidade grandes conjuntos de dados.

2.3.2 Aprendizado de Máquina

O aprendizado de máquina (*Machine Learning*) é uma vertente da inteligência artificial que possui uma ampla gama de aplicações. Consiste na programação de computadores para aprender a partir de dados, permitindo que melhorem seu desempenho ao longo do tempo (GÉRON, 2022d).

Existem dois tipos principais de aprendizado de máquina: o aprendizado supervisionado e o aprendizado não supervisionado. No aprendizado supervisionado, o sistema é treinado com um conjunto de dados que contém exemplos de entrada e saída, aprendendo a mapear as entradas para as saídas correspondentes. No aprendizado não supervisionado, o sistema é treinado apenas com exemplos de entradas, buscando identificar padrões nos dados sem informações prévias sobre a saída.

O aprendizado de máquina possui diversas aplicações, como reconhecimento de padrões, classificação, regressão e agrupamento. Por meio desta técnica é possível identificar padrões em imagens, texto ou áudio, classificando os dados em categorias, prevendo valores ou agrupando dados em grupos semelhantes.

Essa área de pesquisa está em constante evolução e desempenha um papel fundamental no avanço da tecnologia. A utilização dessas técnicas e ferramentas tem

o potencial de ter um impacto positivo na vida e nas atividades humanas. No entanto, é essencial ressaltar a importância de utilizar o aprendizado de máquina de maneira responsável e ética, considerando questões como privacidade, imparcialidade e transparência dos algoritmos.

2.3.2.1 Aprendizado supervisionado

O aprendizado supervisionado é um cenário em que um agente é treinado com a assistência de um professor ou supervisor. O professor fornece ao agente uma medida precisa do seu erro, comparando-o diretamente com os valores de saída desejados. Isso é geralmente obtido por meio de um conjunto de treinamento, que consiste em pares de entrada e saída esperada. Com base nesses exemplos, o agente ajusta seus parâmetros para reduzir a magnitude de uma função de perda global, buscando minimizar a diferença entre os valores previstos e os valores esperados. O objetivo final é treinar o agente para que ele possa generalizar seu aprendizado e lidar com amostras não vistas anteriormente (BONACCORSO, 2017).

De acordo com Géron (2022d), no aprendizado supervisionado, é fornecido ao algoritmo um conjunto de treinamento com soluções desejadas, também chamadas de rótulos ou *labels*. A classificação é uma tarefa típica do aprendizado supervisionado, e um exemplo citado é o filtro de *spam* presente em softwares de gerenciamento de e-mails. Esse *software* é treinado com muitos exemplos de e-mails juntamente com suas classes, e aprende a classificar novos e-mails, conforme figura 3.

Durante o processo de treinamento, ocorrem ajustes iterativos nos parâmetros do modelo com base no retorno fornecido pelo professor (BONACCORSO, 2017). Conforme o agente passa por várias iterações, sua precisão geral tende a aumentar, e a diferença entre as previsões e os valores esperados tende a diminuir. É importante ressaltar que o modelo também deve ser capaz de generalizar seu aprendizado transcendendo aos dados de treinamento. Dessa forma, o aprendizado supervisionado direciona aos modelos a capacidade de aprender a partir de exemplos rotulados, para que possam realizar previsões ou tomar decisões com base em novos dados.

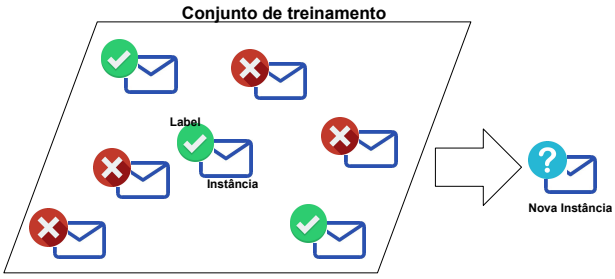

Figura 3 – Conjunto de treinamento rotulado para aprendizagem supervisionada fonte: Adaptado de (GÉRON, 2022a)

Bonaccorso (2017) apresenta as aplicações comuns de aprendizado supervisionado:

- Análise preditiva baseada em regressão ou classificação categórica;
- Detecção de *spam*;
- Detecção de padrões;
- Processamento de linguagem natural;
- Classificação de imagem;
- Processamento automático de sequências (música ou fala).

2.3.2.2 Aprendizado não supervisionado

O aprendizado não supervisionado é uma abordagem em que não há presença de um professor ou supervisor para fornecer as medidas precisas de erro. Nesse contexto, o algoritmo é responsável por encontrar padrões e estruturas nos dados

sem ter informações prévias sobre as saídas desejadas. O algoritmo explora a similaridade ou dissimilaridade entre os exemplos de entrada para agrupá-los ou identificar relações entre eles.

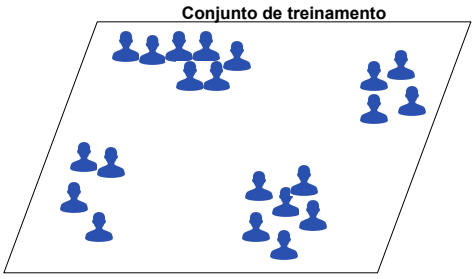

Figura 4 – Conjunto de treinamento sem rótulo para aprendizagem não supervisionada
fonte: Adaptado de (GÉRON, 2022b)

No aprendizado não supervisionado, não há rótulos ou categorias pré-definidas para guiar o processo de treinamento. O algoritmo busca identificar padrões ocultos, *clusters* ou associações nos dados, explorando a distribuição dos exemplos de entrada. É possível agrupar os dados semelhantes em *clusters*, identificar anomalias ou encontrar estruturas complexas nos dados que não seriam facilmente percebidas por um observador humano.

Géron (2022d) faz uma comparação sobre dados de visitantes em um *blog*, onde o sistema tem que classificar em grupos onde possuem características semelhantes. A figura 4 representa o conjunto de dados a serem agrupados e a figura 5 é o resultado desta clusterização aplicada. No cenário proposto a clusterização pode ajudar a direcionar as postagens do *blog*.

Existem diversos algoritmos de aprendizado não supervisionado que podem

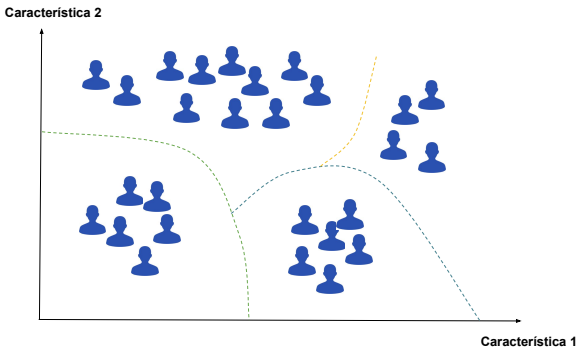

Figura 5 – Conjunto agrupado
fonte: Adaptado de (GÉRON, 2022c)

ser aplicados para a identificação de *clusters* ou anomalias. A biblioteca Scikit-learn é uma renomada ferramenta da linguagem de programação Python, amplamente empregada em tarefas de aprendizado de máquina, incluindo algoritmos de classificação. Esta biblioteca oferece um conjunto abrangente de recursos para avaliação e análise (PEDREGOSA et al., 2011). Géron (2022d) destaca alguns desses algoritmos empregados neste estudo:

- **Clusterização**:
  - **KMeans (Clusterização K-média):** algoritmo de agrupamento popular. É avaliado por meio de métricas como inércia, coeficiente de silhueta e pontuação de completude;
  - **DBSCAN:** algoritmo de agrupamento baseado em densidade. É avaliado com base em métricas como coeficiente de silhueta e índice de Davies-Bouldin;
  - **Análise de *cluster* hierárquica (HCA)**.
- **Detecção de anomalias e de novidades**:

  - **One-class SVM:** é uma implementação do algoritmo de Suporte á Maquina de Vetores utilizado para detecção de valores discrepantes. É avaliada com base na precisão da detecção de *outliers* (número de *outliers* classificados corretamente) e na precisão da detecção de novos dados;
  - **IsolationForest:** conjunto baseado em árvore de detecção de anomalias. Pode ser avaliada com base nas métricas de precisão média, curva de característica de operação do receptor (ROC) e área sob a curva (AUC);
  - **EllipticEnvelope:** algoritmo de detecção de *outlier* baseado no método de envelope elíptico. É avaliado usando precisão e revocação em tarefas de detecção de *outliers*;
  - **LocalOutlierFactor:** comumente conhecido como LOF, é uma abordagem de detecção de anomalia baseada em densidade. É avaliada por meio de medidas como precisão em k, curva *receiver operationg characteristic* (ROC) e precisão média (AP);
  - **SGDOneClassSVM:** é uma variante de **OneClassSVM** implementado, usando uma descida gradiente estocástico, onde se avalia usando a precisão da detecção de *outliers* e a precisão da detecção de novos dados.

- **Visualização e redução de dimensionalidade**:
  - **Análise de Componentes Principais (ACP)**;
  - **Nytroem:** é um método de aproximação do *kernel*. Pode ser avaliado por meio de métricas como erro médio absoluto (MAE) e o erro quadrático médio (MSE);
  - **LLE (Método de redução de dimensionalidade não linear [*Locally Linear Embedding*])**;
  - **t-SNE (Método de incorporação estocástica de vizinhos distribuídos [*Distributed Stochastic Neighbor Embedding*]**.

- **Aprendizado de regras por associação**:

– **KNeighborsClassifier:** algoritmo de classificação simples, mas poderoso. Pode ser avaliado usando exatidão, precisão, *recall* e *F1-score* (PEDREGOSA et al., 2011).

2.3.2.3 Métricas de Avaliação

Os algoritmos de classificação são amplamente utilizados no aprendizado de máquina para categorizar ou atribuir rótulos aos dados estudados. O desempenho de um algoritmo de classificação pode ser avaliado usando várias métricas de avaliação, como acurácia, precisão, *recall* e *F1-score*. Tais métricas geralmente são visualizadas por meio de uma matriz de confusão, representada na Figura 6, que fornece uma visão abrangente dos resultados do algoritmo (SOKOLOVA; LAPALME, 2009) para uma avaliação mais adequada das métricas neste estudo.

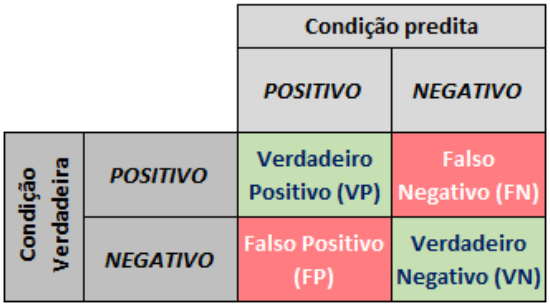

Figura 6 – Exemplo de Matriz de Confusão
fonte: Do autor (2023)

A acurácia é uma métrica empregada na avaliação de modelos de classificação. Ela é definida como a proporção de previsões corretas em relação ao total de previsões geradas por um determinado modelo. Em resumo, ela representa a taxa de acertos do modelo, fornecendo uma medida para avaliação da eficácia de suas previsões em vários contextos de classificação (JAMES et al., 2013).

O *recall* é uma métrica é aplicada em problemas de classificação, pois reflete a capacidade de um modelo em identificar corretamente exemplos que pertencem à classe positiva, considerando todos os exemplos que realmente são positivos (PROVOST; FAWCETT, 2013). O *recall* informa qual a fração dos casos verdadeiramente positivos foi corretamente detectada pelo modelo. É uma métrica

relevante em cenários onde a identificação precisa dos positivos é crítica, como em aplicações médicas para detecção de doenças.

A precisão é uma métrica que avalia a habilidade de um modelo em identificar corretamente os exemplos positivos, considerando todos os exemplos que o modelo classificou como positivos (HASTIE et al., 2009). Fornece uma medida da fração de positivos encontrados pelo modelo que são realmente positivos, buscando evitar falsos positivos, garantindo que as previsões positivas feitas pelo modelo sejam de alta confiança.

O *F1-score* é uma métrica que harmoniza a precisão e o *recall* em um único valor, atribuindo importância a ambos. Sua fórmula de cálculo envolve a média harmônica da precisão e do *recall* (HARRISON, 2019). *F1-score* pode ser utilizada quando se deseja encontrar um equilíbrio entre esses dois aspectos, sobretudo em cenários com classes desbalanceadas. Ela proporciona uma avaliação abrangente do desempenho do modelo, levando em consideração tanto a capacidade de identificar corretamente exemplos positivos quanto a capacidade de evitar falsos positivos.

A matriz de confusão nos permite analisar o número de previsões de uma condição Verdadeiro Positivo (VP), Falso Positivo (FP), Falso Negativo (FN) e Verdadeiro Negativo (VN) encontradas pelo algoritmo. Com base nas quantidades encontradas, pode-se calcular várias métricas. A precisão mede a correção geral das previsões do algoritmo. A precisão quantifica a proporção das instâncias positivas previstas corretamente de todas as instâncias positivas previstas. Com a taxa de Verdadeiro Positivo é determinado a proporção de instâncias positivas corretas gerando a média harmônica de precisão e *recall*, resultando numa medida equilibrada de ambas as métricas.

Métricas de regressão são medidas estatísticas que são usadas para avaliar o desempenho de modelos de regressão para algoritmos de previsão. Elas ajudam a determinar se o modelo de regressão está ajustado aos dados e se este modelo está fazendo previsões precisas (MONTGOMERY; PECK; VINING, 2021).

As métricas de regressão são aplicadas principalmente para lidar com problemas onde os valores de saída são contínuos, como previsão de preços, temperaturas, pontuações ou qualquer valor numérico. Abaixo foi relacionado algumas métricas

mais comuns:

- ***Mean Absolute Error (MAE)*:** Calcula a média das diferenças absolutas entre as previsões do modelo e os valores reais. Esta métrica mede o tamanho médio dos erros;
- ***Mean Squared Error (MSE)*:** Calcula a média das diferenças ao quadrado entre as previsões do modelo e os valores reais. A métrica penaliza erros maiores mais fortemente do que a MAE;
- ***Root Mean Squared Error (RMSE)*:** É a raiz quadrada do MSE. A métrica fornece uma medida da dispersão dos error em uma escala semelhante à da variável de saída;
- ***R-squared (R^2) Score*:** Coeficiente de determinação onde quantifica a proporção da variabilidade total dos valores de saída que é explicada pelo modelo. Um R^2 próximo de 1 indica que o modelo está ajustado;
- ***Mean Absolute Percentage Error (MAPE)*:** Mede a média das porcentagens das diferenças absolutas entre as previsões do modelo e os valores reais em relação aos valores reais. Fornece uma medida relativa do erro;
- ***Max Error*:** Calcula o erro máximo absoluto entre as previsões do modelo e os valores reais. Identifica o pior caso de erro.

As métricas de clusterização desempenham um papel fundamental na análise quantitativa da qualidade dos agrupamentos obtidos por algoritmos de clusterização. Elas oferecem uma maneira de avaliar tanto a eficácia do agrupamento quanto a coesão dos grupos formados. Essas métricas são valiosas ferramentas para medir a validade e a consistência dos resultados de clusterização. Quando se trabalha com algoritmos de clusterização, geralmente não sabe-se a priori a verdadeira estrutura de agrupamento dos dados, porém, especialmente em problemas de aprendizado não supervisionado, tem-se um conjunto de dados com rótulos verdadeiros disponíveis para fins de avaliação.

Duas métricas são amplamente utilizadas para avaliar *clusters* quando os rótulos verdadeiros estão disponíveis que são *adjusted rand score* e o *completeness score*. O ***adjusted rand score*** é uma métrica introduzida por Hubert e Arabie (1985) que mede a concordância entre os rótulos verdadeiros e os rótulos estimados pela solução de clusterização. Este algoritmos fornece um valor entre -1 e 1, onde 1 indica uma correspondência perfeita entre os agrupamentos e -1 indica uma discordância completa.

Voltando o olhar para o ***completeness score***, esta é uma métrica que mede a porcentagem de objetos de um mesmo *cluster* que estão corretamente rotulados. Esta métrica avalia o quão bem todos os itens pertencentes à mesma classe são atribuídos ao mesmo *cluster*, medindo a completude dos *clusters* gerados. Este valor pode variar de 0 a 1, onde 1 indica que todos os itens de uma classe estão no mesmo *cluster* (ROSENBERG; HIRSCHBERG, 2007).

A necessidade de agrupamentos de dados sem a dependência de um conjunto de dados tutor é recorrente, especialmente em cenários complexos, como o Sistema Elétrico Nacional Interligado, caracterizado por sua abrangência continental. A inviabilidade da obtenção de dados de tutor fora dos limites geográficos da concessionária de energia elétrica é um desafio adicional. Este contexto nos leva a explorar três métricas amplamente empregadas para a avaliação de agrupamentos em cenários sem rótulos verdadeiros que são *calinski harabasz score*, *davies bouldin score* e *silhouette score*.

O ***calinski harabasz score***, também conhecido como critério de variação fora e dentro dos *clusters*, é uma métrica que busca maximizar a relação entre a dispersão entre *cluster* e a dispersão dentro do *cluster*. Quanto maior o valor dessa métrica, melhor a separação entre os *clusters* (CALIŃSKI; HARABASZ, 1974).

Por sua vez o ***davies bouldin score*** é uma medida que avalia a média das razões de similaridade entre cada *cluster* e seu *cluster* mais próximo. Valores menores indicam *clusters* mais compactos e bem separados (DAVIES; BOULDIN, 1979).

Por fim o ***silhouette score*** mede o quão semelhantes os objetos em um *cluster* são entre si em comparação com outros *clusters* próximos. Valores próximos

y_true	#	#	#	#	#	#	#	#	#	#	1	#	1	#	1	#	1	#	
y_pred	0	0	0	1	0	1	0	0	0	1	0	0	1	0	1	1	0	0	
acurácia	0	0	0	0	0	0	0	0	0	0	-1	0	1	0	1	0	-1	0	

Figura 7 – Métrica de Classificação Modificada para Anomalias

de 1 indicam *clusters* bem definidos, enquanto valores próximos de -1 sugerem que os objetos podem ter sido atribuídos ao *cluster* errado (ROUSSEEUW, 1987).

A Figura 7 apresenta uma classificação modificada para anomalias, baseada nas previsões y_{pred} e nos valores verdadeiros y_{true}. Os valores em y_{pred} refletem as previsões do modelo, com 0 indicando a ausência de anomalias e 1 indicando sua presença. Os valores em y_{true} representam os dados verdadeiros, sendo 1 onde existem anomalias documentadas e '#' em outros casos. A linha final, intitulada 'acurácia', identifica onde o modelo acertou (1) e onde cometeu erros (-1) na detecção de anomalias.

2.4 Trabalhos Relacionados

Os sistemas elétricos são caracterizados pela sua complexidade e pela necessidade de uma operação confiável e eficiente. A detecção e classificação de eventos em sistemas elétricos são vitais para a realização de operações e distribuições seguras. A utilização de técnicas de aprendizado de máquina para análise de dados gerados pelas PMUs, têm se mostrado promissoras para detecção rápida e precisa de eventos elétricos. Nesta seção serão levantado alguns trabalhos relacionados à este assunto.

No trabalho de Cui et al. (2019), é introduzida uma inovadora abordagem de detecção de eventos. Nessa perspectiva, um algoritmo de compressão de dados desempenha um papel fundamental ao reduzir significativamente o *streaming* de dados, contribuindo para a resolução de problemas de otimização. Devido à notável quantidade de dados gerados pelas PMUs, a tarefa computacional associada é considerável, tornando ainda mais essencial a adoção de uma estratégia de

otimização eficaz.

Em outro artigo os autores Aalam e Shubhanga (2023) abordam a importância do monitoramento de sistemas de energia por meio de Sistemas de Medição de Ampla Área (WAMS). Devido à natureza geograficamente diversificada das PMUs e à alta taxa de amostragem usada para sincrofasor, os dados da PMU revelam o estado de um sistema de energia de forma mais precisa do que os métodos convencionais baseados em SCADA. O artigo ressalta que a detecção automatizada de eventos é necessária devido à impraticabilidade da abordagem manual. São mencionados os desafios da grande quantidade de dados gerados pelas PMUs e a importância da análise eficiente para que os operadores do sistema compreendam a condição atual do sistema. O artigo também discute a detecção de eventos em sistemas de energia modernos, a classificação de eventos e a importância da análise de eventos como parte do processo mais amplo. É mencionada a disponibilidade de algoritmos de detecção de eventos na literatura, classificados como baseados em treinamento e não baseados em treinamento. Além disso, são apresentados métodos de detecção de eventos baseados em aprendizado de máquina (ML) ou *deep learning* (DL) para detecção e categorização em tempo real.

Com a utilização de micro redes e energias renováveis, os autores Thomas, Koshy e R. (2020) afirmam em seu artigo que o cenário citado aumenta a complexidade do sistema e a detecção, estimativa e análise de eventos no sistema elétrico. A utilização das PMUs simplificou o monitoramento e controle dinâmico do sistema de energia, destacando o uso de técnicas de transformação *wavelet* para processar os dados coletados e identificar eventos transitórios, como desligamentos de geradores e injeções de potência reativa. Também menciona a necessidade de monitorar continuamente mudanças na frequência e tensão para detectar eventos. São discutidas técnicas de extração de características com base em análise *wavelet* e classificação não linear usando algoritmos de aprendizado de máquina.

No trabalho de Aligholian et al. (2020), os autores abordam a importância das unidades de medição de fasoriais em nível de distribuição, conhecidas como microPMUs, que fornecem medidas de tensão e corrente com alta resolução e precisão, melhorando a visibilidade na rede de distribuição. Esses dados são aplicados em diversas áreas, como identificação de topologia e fase, modelagem de carga,

estimativa de estado, monitoramento de ativos e cibersegurança do sistema de distribuição. O foco é investigar eventos nos sistemas de distribuição, abordando métodos estatísticos e de aprendizado de máquina para detecção de eventos. O texto introduz métodos baseados em modelos de Rede Adversárias Genereativas (GAN) para a detecção de eventos, destacando a inovação de abordagens não supervisionadas que não exigem conhecimento especializado e podem identificar marcadores individuais e em grupo nos dados das microPMUs. Resultados de avaliação com dados reais de microPMUs demonstram que os métodos propostos superam um método estatístico tradicional, especialmente em eventos de pequenas variações.

Com a crescente quantidade de dados coletados por PMUs em unidades elétricas nos EUA, o artigo de Hai et al. (2021) destaca o desafio de analisar rapidamente esses dados históricos para detecção de eventos relevantes. A detecção de eventos é essencial e muitas vezes tratada como uma tarefa de aprendizado não supervisionado, mas os dados das PMUs frequentemente desafiam essa abordagem. Para contornar a necessidade de rotulação manual intensiva, os métodos semi-supervisionados em conjunto com aprendizado ativo são usados, mas podem ser inviáveis para grandes volumes de dados. O estudo propõe o uso de aprendizado por transferência para reduzir a necessidade de rótulos, demonstrando que essa abordagem supera algoritmos de aprendizado de máquina do estado da arte em detecção de eventos com um pequeno número de instâncias rotuladas representativas, mesmo em situações que exigem esforço intenso de rotulação.

2.5 Considerações finais

Neste capítulo explorou-se alguns conceitos relacionados ao uso de algoritmos de Inteligência Artificial na detecção de eventos em sistemas elétricos. O setor elétrico desempenha um papel vital no desenvolvimento da sociedade, e a aplicação de recursos tecnológicos avançados, como as Unidades de Medição Fasorial (PMUs) oferece novas perspectivas na captura e análise de informações relevantes para a gestão das redes elétricas.

As PMUs, como dispositivos avançados de monitoramento, têm a capacidade

de capturar informações elétricas com alta precisão e sincronização, fornecendo uma visão em tempo real do sistema elétrico. Isso permite uma detecção rápida de eventos anormais, como variações na carga e falhas na rede, contribuindo para a melhoria da confiabilidade e segurança do sistema elétrico.

O Aprendizado de Máquina, por sua vez, é uma ferramenta poderosa que possibilita o desenvolvimento de algoritmos capazes de identificar padrões e realizar previsões a partir de dados. Neste trabalho, foram exploradas as vertentes do Aprendizado supervisionado e não supervisionado, destacando suas aplicações e análise de sistemas elétricos. O Aprendizado supervisionado é essencial para a classificação de eventos, como falhas e variações, enquanto o Aprendizado não supervisionado é útil para a detecção de anomalias e agrupamentos.

3 Desenvolvimento

Este capítulo irá tratar do desenvolvimento da proposta feita para este trabalho, analisando as técnicas de detecção de anomalias envolvendo o problema do Sistema Interligado Nacional (SIN) em relação à rede da concessionária do estado do Paraná. Será detalhado a composição da base de dados e como esta será utilizada para o estudo deste trabalho.

3.1 Considerações Iniciais

O Sistema Interligado Nacional (SIN) representa uma intrincada rede de sistemas interconectados de geração e transmissão de energia elétrica no Brasil, assegurando o fornecimento necessário para atender às demandas dos consumidores em extensões continentais. Diante desse cenário de produção em abordagem hidro-termo-eólica, bem como outros modelos de geração, a distribuição conecta quatro subsistemas (Sul, Sudeste/Centro-Oeste, Nordeste e a maior parte da região Norte) (ONS, 2023d), delineando um campo de pesquisa de grande relevância para o escopo deste trabalho.

Com a característica continental e interconectada do sistema SIN é remetido à hipótese levantada neste trabalho, onde a possibilidade de detecção de anomalias dentro e fora do sistema elétrico da concessionária por meio do uso de PMUs combinadas com algoritmos de Inteligência Artificial.

No contexto desse desafio, é considerado a ocorrência de um evento significativo em um subsistema distante do sistema elétrico da concessionária. A capacidade das PMUs de capturar e transmitir dados em tempo real proporciona uma oportunidade única para avaliar como esses eventos impactam a rede elétrica em diferentes locais. O desenvolvimento de algoritmos de Inteligência Artificial capazes de analisar esses dados, fornecem *insights* valiosos para a detecção precoce de anomalias, abrindo a oportunidade de tomada de decisões mais eficazes, contribuindo para a estabilidade e confiabilidade do sistema elétrico.

3.2 Base de Dados de Estudo

Os arquivos foram disponibilizados em duas pastas, separadas por data de eventos estudados ONS (2023a) e ONS (2023b), no formato CSV (*Comma-Separated Values*) para o armazenamento dos dados contendo informações representadas em formato tabular. O formato CSV facilita a manipulação e análise dos dados, favorecendo a interoperabilidade com distintas ferramentas de análise de dados.

A base de dados se constitui de diversos arquivos, cada qual abarcando um intervalo temporal específico. Cada arquivo condensa um conjunto multifacetado de informações acerca de grandezas elétricas, incluindo magnitude e ângulo de fase, com coleta realizada em diversos barramentos da rede elétrica. Além disso, são registrados dados relativos à frequência da rede, variação da frequência em taxa (ROCOF), estado das medições e detalhes temporais (hora, minuto, segundo e mili segundos).

Os parâmetros de magnitude e ângulo de fase figuram como elementos fundamentais na caracterização do comportamento elétrico do sistema. A magnitude alude à intensidade de uma grandeza elétrica, enquanto o ângulo de fase estabelece sua posição relativa em relação a uma referência. Ambos são cruciais para a avaliação da qualidade e estabilidade da rede, possibilitando a identificação de oscilações, flutuações e anomalias.

Entre as características dessa base de dados, sobressai sua natureza sincronizada e a notável resolução temporal. Essa característica confere-lhe a capacidade de registrar com detalhes cada evento ou mudança que ocorre no sistema elétrico, permitindo uma análise minuciosa em tempo real dos dados. Além disso, sua geolocalização proporciona a oportunidade de investigação de eventos específicos determinados locais da rede elétrica.

A exploração dessa base de dados assume um papel fundamental neste estudo, uma vez que por meio dela é possível identificar eventos de relevância, discernir padrões comportamentais e detectar potenciais anomalias. A análise cuidadosa dos dados desempenha um papel crucial na garantia da confiabilidade do sistema elétrico, contribuindo para a melhoria da qualidade do serviço prestado e a prevenção de incidentes que poderiam resultar em interrupções no fornecimento de

energia elétrica. Essa análise é essencial para a eficácia e a segurança da operação do sistema elétrico, garantindo seu funcionamento contínuo e a satisfação dos usuários.

A tabela 1 tem como propósito exemplificar os campos contidos no arquivo denominado C37118-5193-GER-3, que corresponde a uma parcela específica da base de dados analisada neste estudo. Este arquivo em particular, abarca registro referentes ao dia 08 de abril de 2021, data em que foi documentado um evento de desligamento na rede do Sistema Interligado Nacional (SIN) (ONS, 2023a). O arquivo fornece informações essenciais para esta pesquisa, uma vez que está relacionado a uma PMU acoplada a um gerador. Este arquivo é composto por um conjunto de 32 colunas, onde seus atributos serão detalhados na tabela supracitada.

Conforme previamente mencionado, os arquivos desta base de dados compartilham campos fixos, consistentes em dados temporais, tais como Data, Hora, Estado, Frequência e Taxa de Variação da Frequência (ROCOF). Estas informações temporalmente ancoradas servem como base para a contextualização das medições efetuadas.

Adicionalmente aos atributos mencionados, o arquivo de exemplo incorpora campos variáveis que encapsulam as medições correspondentes à Magnitude e Ângulo de cada fase. Embora a natureza desses campos possa variar entre diferentes arquivos, desempenham um papel essencial nesta pesquisa, contribuindo para uma análise detalhada e abrangente dos eventos, padrões e identificação de anomalias.

3.3 Proposta

Ao analisar os dados dos arquivos, percebe-se o imenso volume de informações a serem processadas, demandando vastos recursos computacionais e de armazenamento. Este cenário ressalta a importância de condensar esses dados, uma abordagem já explorada em diversas pesquisas. Em cada leitura por ciclo, soma-se um total de 60 leituras por segundo, o que se traduz em 216.000 leituras por hora, 5.184.000 leituras diárias e 36.288.000 leituras semanais. Considerando uma média de 30 parâmetros por leitura, resultando em aproximadamente 155 milhões de registros coletados em um único dia.

Dadas essas estatísticas, os arquivos gerados pelas PMUs podem facilmente ultrapassar os 3 Gigabytes de tamanho. Arquivos desse tamanho são notoriamente demorados para serem lidos, o que compromete a análise em tempo real, tornando-a lenta e propensa a erros. Por isso, a compressão de dados surge como uma ferramenta valiosa, capaz de aprimorar significativamente a leitura e processamento dos dados coletados.

Ao empregar a função ***to_parquet*** da biblioteca Pandas (PANDAS, 2023), observou-se uma melhoria expressiva na eficiência da leitura e análise dos dados. O ***Parquet***, como discutido no artigo de Buroni et al. (2018), é um formato de armazenamento colunar. Isso significa que os dados são armazenados por coluna e não por linha, otimizando consideravelmente operações analíticas.

Ao empregar a biblioteca ***Parquet***, é direcionado à técnica de *downsampling*, ou condensação de dados. Esta técnica tem se mostrado uma excelente ferramenta para a redução da dimensionalidade de conjuntos de dados vastos. Essa é uma importante abordagem em cenários que exigem uma administração de dados ágil e refinada. Para analisar os dados oriundos das PMUs, empregou-se esta técnica aliada ao método ***to_parquet*** da biblioteca **Pandas**, focando especificamente nos dias e horários indicados pelo Operador Nacional do Sistema, que sinalizou eventos de importância crítica para o Sistema Interligado Nacional (ONS, 2023a). Liu et al. (2022) enfatiza a pertinência deste método, ilustrando sua aplicabilidade em contextos intrincados, como o das turbinas a gás.

Os dados condensados corroboram o objetivo deste estudo, sugerindo que PMUs, mesmo quando situadas em regiões geograficamente distantes dos eventos em análise, são capazes de identificar alterações na rede. Isso habilita a implementação de medidas preventivas, minimizando o risco de incidentes em cascata na rede e contribuindo para a manutenção e preservação dos equipamentos de geração e distribuição de energia.

O pré-processamento de dados das PMUs é uma etapa crítica que garante a qualidade dos dados que serão posteriormente utilizados para monitoramento e análise. A literatura destaca uma variedade de técnicas empregadas nesta etapa, como o artigo de Vanfretti, Bengtsson e Gjerde (2015) que fornece *insights* signi-

ficativos sobre as abordagens metodológicas. Inicialmente, um procedimento de catalogação dos arquivos disponíveis é realizado e após é aplicado uma função para listar e quantificar todos os arquivos disponíveis no diretório que serão processados. Os arquivos que serão utilizados, são do formato ***Parquet*** como supracitado nesta seção. A adoção deste formato é essencial devido à sua eficiência em desempenho e compressão, oferecendo grandes vantagens de processamento e armazenamento.

É implementado também uma função para extrair e filtrar dados dos arquivos, com base em um intervalo de tempo pré-definido que corresponde aos eventos estudados. Vanfretti, Bengtsson e Gjerde (2015) enfatiza que "os procedimentos envolvidos no pré-processamento têm obviamente um papel importante", indicando que as etapas de filtragem e extração são fundamentais para assegurar que os dados utilizados reflitam precisamente as condições do sistema durante os eventos. Zhou et al. (2008) também enfatiza a importância de aplicar um método de pré-processamento para reduzir a escala do estudo.

Nesta função de extração e filtragem é também implementado manipulações para se organizar os dados de forma mais eficiente, incluindo agrupamento de colunas de várias tabelas em uma única tabela, garantindo que todas as informações relevantes sejam consolidados em um formato unificado e acessível. Este passo é importante para facilitar a análise subsequente e para manter a integridade dos dados através de um esquema organizado. A eliminação de colunas nulas é outra operação realizada para a limpeza do conjunto de dados irrelevantes ou redundantes para se manter a qualidade dos dados.

Ainda na função de extração e filtragem é adicionado uma coluna como um marcador binário indicando se a linha de dados está dentro do intervalo de tempo dos eventos estudados. Este passo é vital para a classificação subsequente dos dados em normais ou relacionados aos eventos.

A utilização de algoritmos de aprendizado de máquina para identificação de padrões *outliers* nos dados, tornou-se uma prática comum, e ferramentas como o *scikit-learn* (SCIKIT-LEARN, 2023) oferecem uma variedade de modelos para análise de dados. Modelos como *OneClassSVM, EllipticEnvelope, IsolationForest* e *LocalOutlierFactor* são frequentemente empregados para identificar pontos que não

se conformam ao padrão esperados dos dados.

Cada modelo possui seus próprios parâmetros que tornam adequado para diferentes tipos de buscas nos dados e cenários de anomalias. Após a aplicação desses modelos, é fundamental avaliar seu desempenho utilizando métricas robustas. A acurácia balanceada, por exemplo, é especialmente útil em situações onde as classes estão desequilibradas, enquanto o *F1-score* fornece uma medida de precisão e revocação. O *Fowlkes-Mallows Score, Mutual Information Score, Completeness Score* e *Homogeneity Score* oferecem perspectivas adicionais sobre a qualidade dos agrupamento realizados pelos modelos em relação às anomalias identificadas.

Avançando além do *Scikit-Learn*, a biblioteca ***PyOD*** (PYOD, 2023) é outro recurso estudado para detecção de anomalias. Esta biblioteca especializada, estende o repertório de modelos disponíveis e fornecem mecanismos adicionais para lidar com a detecção de *outliers*. Ao aplicar esses modelos avançados, o código pode calcular métricas similares para cada evento identificado, proporcionando uma compreensão abrangente da eficácia do modelo.

Complementando o arcabouço de bibliotecas empregadas, o uso da ***Anomaly Detection Toolkit*** **(ADTK)** (ANALYTICS, 2023) traz uma abordagem diferenciada com o modelo *SeasonalAD*, que é particularmente eficaz para dados com padrões sazonais. Este modelo pode capturar anomalias que são dependentes do tempo que podem não ser detectadas por métodos que não consideram a temporalidade dos dados. A visualização das anomalias detectadas é de grande valia para esta análise, permitindo discernir rapidamente padrões atípicos e a identificação de condições sob as quais as anomalias ocorrem. A integração dessas bibliotecas resultam um fluxo bem elaborado e diverso para detecção de anomalias, o qual pode ser aplicado a uma vasta gama de domínios e tipos de dados. O código foi projetado não apenas para identificar anomalias, mas também para avaliar e interpretar o significado dessas anomalias dentro do contexto específico de cada evento.

Dentro do escopo deste estudo, a figura 8 oferece uma representação gráfica do algoritmo empregado, elucidando as operações específicas realizadas na construção da base de dados utilizada para análises subsequentes. A utilização do formato de arquivo *Parquet* se revelou de extrema importância, uma vantagem já destacada

neste capítulo, devido à sua capacidade de comprimir eficientemente os dados e acelerar significativamente a leitura e o processamento dos mesmos.

Prosseguindo para além da etapa de processamento inicial, o estudo avança para o processamento analítico, detalhando e comparando os resultados gerados pelo algoritmo dedicado à detecção de anomalias. Este processo está sendo representado na figura 9, que descreve os passos e as análises específicas aplicadas aos eventos investigados. Para a implementação dos algoritmos, optou-se pela linguagem de programação Python, reconhecida por sua ampla aplicabilidade e eficácia em computação científica (OLIPHANT, 2007). Os códigos desenvolvidos foram disponibilizados de maneira aberta e acessível através de um repositório público na plataforma **GitHub** (GITHUB, 2024), conforme citado na referência de Ferreira (2023).

Em casos em que os dados de tutor não estão disponíveis, é possível utilizar uma acurácia modificada para avaliar a solução de clusterização. Essa acurácia é calculada como a porcentagem de objetos que estão corretamente alocados em um *cluster*, considerando os *clusters* estimados pela solução de clusterização.

A acurácia modificada é uma métrica menos precisa do que as métricas sem tutor tradicionais, pois não considera a compactação dos *clusters* e a separação entre eles. No entanto, ela é uma métrica útil em casos em que não há dados de tutor disponíveis.

3.4 Métricas

A avaliação dos resultados dos detectores de anomalia desempenha um papel fundamental na análise de sistemas e processos, permitindo a identificação de comportamentos incomuns ou potencialmente prejudiciais. Para uma avaliação adequada da eficácia desses detectores, diversas métricas são comumente empregadas, entre as quais se destacam a Acurácia Balanceada (*Balanced Accuracy*), o *F1-Score*, o *Fowlkes-Mallows Score* (FMS), o *Mutual Information Score* (MIS), o *Completeness Score* (CPS) e o *Homogeneity Score* (HGS).

A ***Acurácia Balanceada***, como apontada por Lavin e Ahmad (2015)

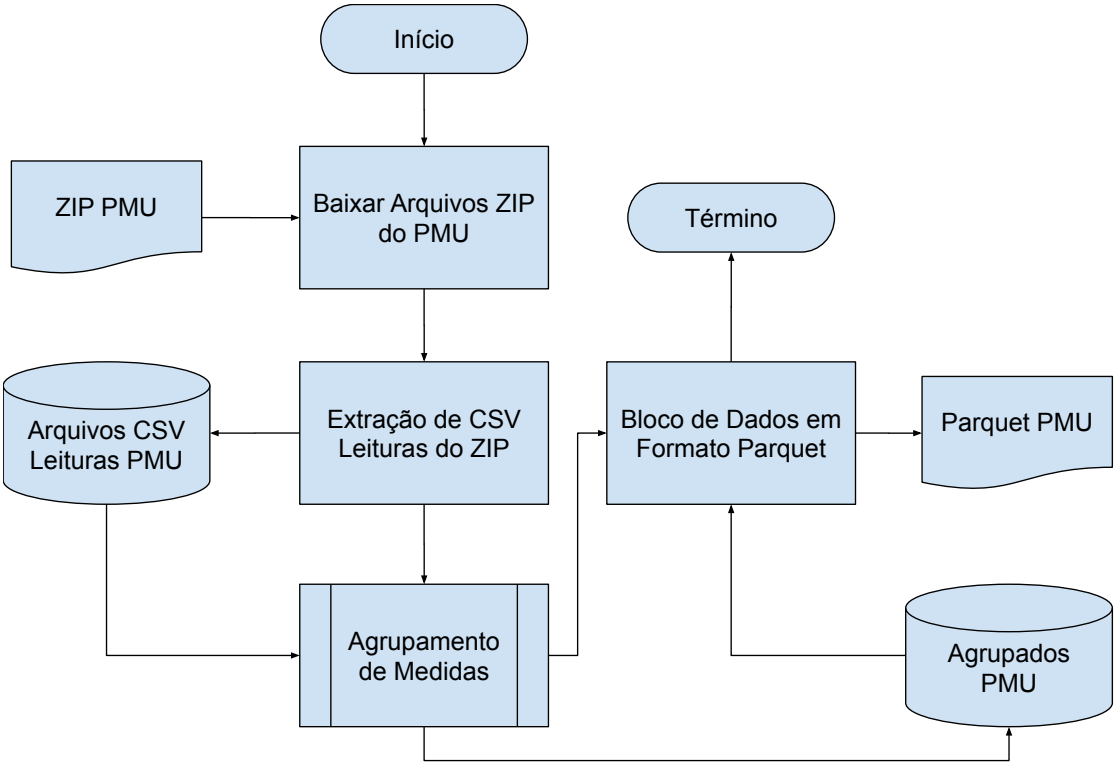

Figura 8 – Criação dos Arquivos Parquet
fonte: Do autor (2023)

representada pela equação 3.1, ilustra uma métrica crucial que avalia a habilidade do detector de anomalia em discriminar entre instâncias normais e anômalas. Essa métrica leva em consideração tanto a taxa de verdadeiros positivos (TPR) quanto a taxa de verdadeiros negativos (TNR), sendo especialmente relevante quando as classes estão desequilibradas. Um alto valor de Acurácia Balanceada indica um desempenho consistente do detector na detecção de anomalias.

$$acuraciaBalanceada = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{3.1}$$

O ***F1-Score*** é uma métrica que combina precisão e *recall* (STOCCO; TONELLA, 2020), ilustrada pela equação 3.2. A precisão mede a proporção de exemplos classificados como positivos que são realmente positivos $Precisao = \frac{TP}{TP+FP}$, enquanto o *recall* mede a proporção de exemplos verdadeiramente positivos que foram

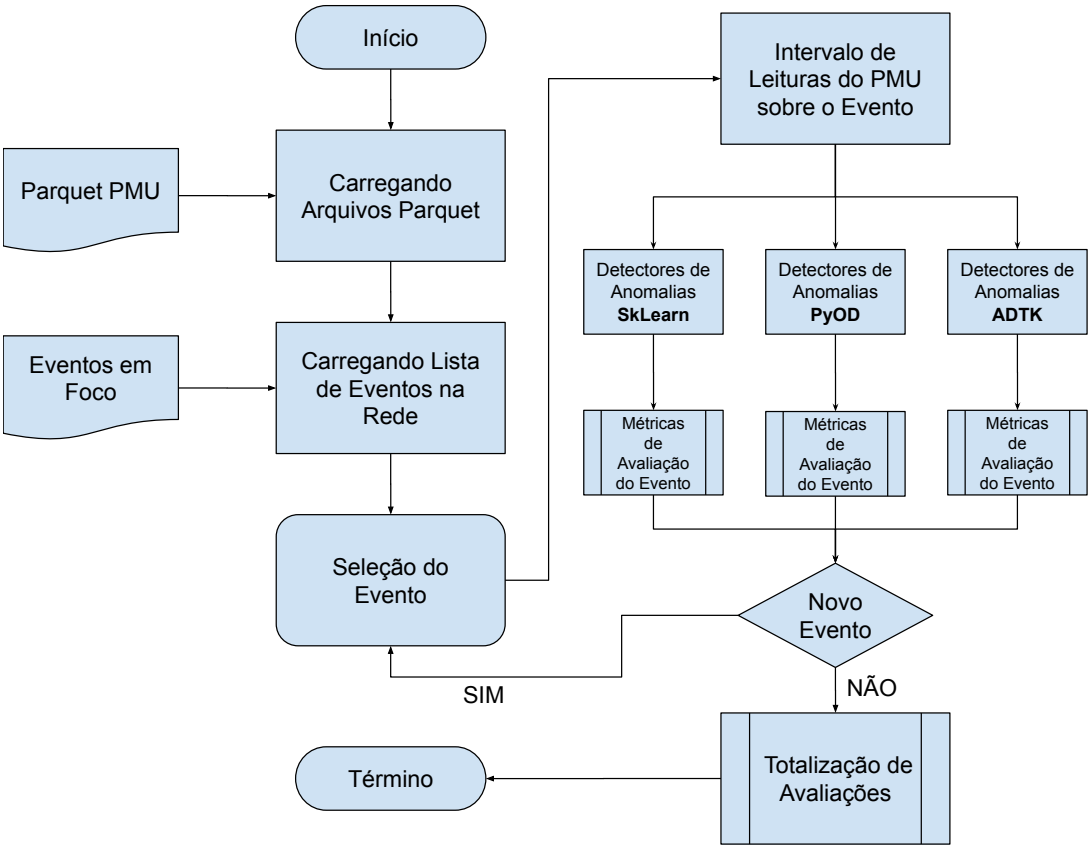

Figura 9 – Avaliação dos detectores de anomalia para os eventos nas medidas dos PMUs
fonte: Do autor (2023)

corretamente identificados $Recall = \frac{TP}{TP+FN}$. O *F1-Score* é particularmente útil quando os resultados falsos positivos e falsos negativos são desiguais, buscando um equilíbrio entre precisão e *recall*.

$$f1score = 2.\frac{Precisao.Recall}{Precisao + Recall} \tag{3.2}$$

O ***Fowlkes-Mallows Score*** (FMS), também abordado por Stocco e Tonella (2020), é utilizado para avaliar a similaridade entre os grupos de instâncias verdadeiras e as previstas. A equação 3.3 representa esta métrica, fornecendo uma medida que quantifica quão bem os grupos formados pelos resultados do detector de anomalias se aproximam dos grupos verdadeiros, indicando uma boa similaridade

quando próximo de 1. Além disso, o ***Mutual Information Score*** (MIS) quantifica a quantidade de informação compartilhada entre os rótulos verdadeiros e os rótulos previstos pelo detector, refletindo o grau de dependência entre esses rótulos.

$$FMS = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}} \tag{3.3}$$

A ***Completeness Score*** (CPS) e a ***Homogeneity Score*** (HGS), também abordadas por Stocco e Tonella (2020), permitem avaliar a extensão em que todos os exemplos verdadeiramente positivos foram recuperados pelo detector. O CPS verifica se o detector identificou todas as anomalias presentes nos dados, apontando para uma completude mais alta. Já o HGS, avalia se os grupos formados pelo detector contêm apenas pontos de dados que pertencem a uma única classe, garantindo a capacidade do detector de agrupar as anomalias de forma distintiva em relação aos exemplos normais. Essas métricas, em conjunto, fornecem uma visão abrangente do desempenho dos detectores de anomalia em diversas situações de avaliação.

$$cps = 1 - \frac{H\left(T|K\right)}{H(T)} \tag{3.4}$$

$$hgs = 1 - \frac{H\left(K|T\right)}{H(K)} \tag{3.5}$$

A equação 3.4 define o ***Completeness Score***, que mede a capacidade do modelo de agrupar todos os membros de uma classe verdadeira em um único *cluster*, através da proporção de $1 - \frac{H(T|K)}{H(T)}$, onde $H\left(T|K\right)$ representa a entropia condicional das classes verdadeiras dado os *clusters* e $H(T)$ a entropia das classes verdadeiras. Por outro lado a equação 3.5 apresenta o ***Homogeneity Score***, que avalia se cada *cluster* contém apenas membros de uma classe, calculando por $1 - \frac{H(K|T)}{H(K)}$, com $H\left(K|T\right)$ sendo a entropia condicional dos *clusters* dado as classes e $H(K)$ a entropia dos *clusters*. Ambas as métricas refletem diferentes aspectos da qualidade dos *clusters*, destacando a eficiência do detector em manter a homogeneidade e completude ao mesmo tempo.

3.5 Considerações Finais

Conforme apresentado neste capítulo é proposto uma metodologia rigorosa e aprofundada para a análise de dados de energia elétrica, empregando uma combinação poderosa de ferramentas, técnicas de análise de dados e detecção de anomalias. A imensa quantidade de dados coletados por PMUs exige a implementação de estratégias eficientes para sua gestão, análise e interpretação. Através da compressão de dados, da implementação de vários modelos de aprendizado de máquina, e do uso eficaz de métricas de avaliação relevantes, este estudo demonstrou a viabilidade e a eficácia desta abordagem para identificar eventos de importância crítica e detectar anomalias dentro de um vasto conjunto de dados.

A utilização de métricas como ***Acurácia Balanceada***, ***F1-Score***, ***Fowlkes-Mallows Score***, ***Mutual Information Score***, ***Completeness Score*** e ***Homogeneity Score*** demonstraram ser essenciais na avaliação da eficácia dos modelos de detecção de anomalia utilizados. Essas métricas em combinação forneceram uma perspectiva abrangente da performance do detector, proporcionando uma compreensão aprofundada de suas potenciais limitações e pontos fortes.

A adoção das bibliotecas ***Pandas***, ***PyOD*** e ***ADTK***, provou ser significativa no manuseio eficiente dos dados, permitindo a realização de operações complexas com eficácia e em tempo hábil. A flexibilidade dessas ferramentas fornece a oportunidade de experimentar uma variedade de abordagens analíticas e ajustar o processo à natureza específica dos dados à disposição. Como resultado, esta investigação não apenas apresentou uma estratégia viável para o estudo de grandes volumes de dados coletados pelas PMUs, mas também estabelece uma fundação sólida para futuras pesquisas nesta área.

Tabela 1 – Significado dos Campos do Arquivo C37118-5193-GER-3

Campo	**Significado**
Date	Data (Sincronizado com GPS)
Time (America/Sao_Paulo)	Hora (Hora, minuto, segundo e milissegundos, sincronizado com GPS)
Status	Estado
Frequency	Frequência
df/dt	Taxa de Variação da Frequência (ROCOF)
VA1 V-GER3 Magnitude	Magnitude da barra VA1
VA1 V-GER3 Angle	Ângulo da barra VA1
VB1 V-GER3 Magnitude	Magnitude da barra VB1
VB1 V-GER3 Angle	Ângulo da barra VB1
VC1 V-GER3 Magnitude	Magnitude da barra VC1
VC1 V-GER3 Angle	Ângulo da barra VC1
IA1 I-GER3 Magnitude	Magnitude da corrente IA1
IA1 I-GER3 Angle	Ângulo da corrente IA1
IB1 I-GER3 Magnitude	Magnitude da corrente IB1
IB1 I-GER3 Angle	Ângulo da corrente IB1
IC1 I-GER3 Magnitude	Magnitude da corrente IC1
IC1 I-GER3 Angle	Ângulo da corrente IC1
IA1 I-TRAFO-GER3 Magnitude	Magnitude da corrente IA1
IA1 I-TRAFO-GER3 Angle	Ângulo da corrente IA1
IB1 I-TRAFO-GER3 Magnitude	Magnitude da corrente IB1
IB1 I-TRAFO-GER3 Angle	Ângulo da corrente IB1
IC1 I-TRAFO-GER3 Magnitude	Magnitude da corrente IC1
IC1 I-TRAFO-GER3 Angle	Ângulo da corrente IC1
IEX-RAT-G3	Corrente de excitação do gerador
PD-G3	Dados de potência ativa do gerador
PSS-SA-G3	Dados do sistema de estabilização de potência do gerador
W-G3	Dados de velocidade angular do gerador
RPM-P-G3	Dados de RPM (rotações por minuto) do gerador (eixo principal)
RPM-S-G3	Dados de RPM (rotações por minuto) do gerador (eixo secundário)
SETPOINT-MW-G3	Valor de referência de potência ativa do gerador
VEX-RAT-G3	Tensão de excitação do gerador
G3-ONS-COMANDO	Comandos do gerador pelo ONS (Operador Nacional do Sistema Elétrico)

4 Experimentos e Discussões

Neste capítulo serão apresentados os experimentos realizados utilizando diferentes algoritmos para detecção de anomalias em grandes volumes de dados provenientes de PMUs. Para otimizar a análise, foi realizado um processo de compressão de dados, através dos conceitos de ***parquet file***, uma especificação de representação de dados colunares eficiente. Esta etapa revelou-se fundamental, reduzindo o tamanho dos arquivos e permitindo uma leitura mais rápida e eficiente dos dados. Foi explorado técnicas de aprendizado de máquina não supervisionado, tais como **OneClassSVM**, **EllipticEnvelope**, dentre outras, para identificação de anomalias. A eficácia dos métodos utilizados foi avaliada através da aplicação de diferentes métricas de desempenho.

4.1 Criação dos Arquivos Parquet

No início da análise, deparou-se com uma questão desafiadora relacionada ao processamento dos dados. Os arquivos provenientes das PMUs eram de tamanho considerável, contendo uma vasta quantidade de dados pertinentes às leituras das fases elétricas com as quais estavam associados. Conforme detalhado nos capítulos 2 e 3 deste trabalho, os conjuntos de dados alcançavam a ordem de milhões de entradas. Esse volume expressivo de dados revelou-se um obstáculo significativo para a eficiência da leitura e processamento, comprometendo a celeridade requerida pelo modelo analítico proposto. A prolixidade no manuseio dos dados apresentou-se como fator limitante para a sua aplicação em tempo real.

Visando superar esta limitação, desenvolveu-se um algoritmo para a compressão dos dados, otimizando sua utilização nas análises de anomalias. Este algoritmo, codificado na linguagem de programação **Python**, emprega a biblioteca **Pandas** (NUMFOCUS, 2023) para a leitura dos arquivos *CSV* originais, utilizando o ***constructor*** *DataFrame*, uma estrutura de dados tabular e orientada a colunas. Em seguida, aplica a função *to_parquet* para transcodificar os dados para o formato

parquet. De acordo com McKinney (2012), a biblioteca **Pandas** é equipada com estruturas e funções de alto nível, concebidas para simplificar e acelerar o trabalho com dados estruturados ou tabulares, características que são de suma importância para a presente investigação.

A função *to_parquet* é a implementação do padrão de *parquet file*, uma especificação apresentada pela ***Apache Parquet*** para disponibilizar vantagens da reapresentação de dados colunares compactada e eficiente para os projetos do ecossistema ***Hadoop***, pertencente à ***Apache Fundation*** (PARQUET, 2023). A especificação *parquet* permite a organização de estruturas de dados aninhadas e complexas em memória, tal como descrito por Melnik et al. (2010), e é projetada para suportar esquemas de compressão e codificação altamente eficazes. Essas características são fundamentais para melhorar significativamente o desempenho da aplicação que necessita de manipulação eficiente de grandes volumes de dados.

Com a integração da biblioteca em questão, o algoritmo de compressão assume um papel fundamental no contexto da análise de anomalias delineada por este estudo. Durante a implementação, configura-se uma rotina de leitura para os arquivos originais, os quais se encontram comprimidos em formato *ZIP*. Esta iteração promove a extração do conteúdo do arquivo *ZIP*, resultando em um arquivo acessível no formato *CSV* pronto para ser processado. Posteriormente, procede-se à abertura e leitura do arquivo *CSV* por meio da função *read_csv* pertencente à biblioteca **Pandas**. A utilização do parâmetro *chunksize* possibilita uma leitura em modalidade de *streaming* de dados, otimizando o procedimento para alcançar um desempenho de leitura superior e mais eficiente.

Durante a etapa de aquisição de dados, é estabelecido uma estrutura de dados colunar para catalogar os registros que serão empregados na análise de anomalias. Essa estrutura é composta por informações referentes aos arquivos individuais das PMUs em análise, assim como dados pertinentes a ângulo e frequência. A partir dessas informações, calculam-se valores mínimos, máximos e a variação dessas métricas, os quais são meticulosamente integrados ao ***constructor*** *DataFrame*.

Ilustrando a eficácia do processo de compressão e otimização de dados adotado, observou-se uma diminuição notável no volume do arquivo de dados.

Originalmente, o arquivo no formato *CSV* possuía 1.024.009 KB e, após a compressão e tratamento dos dados, foi gerado um arquivo no formato *parquet* com apenas 101 KB. Essa redução não só facilita uma leitura mais ágil dos dados como também propicia uma análise mais eficiente, demonstrando a relevância prática da metodologia implementada em grandes conjuntos de dados.

A implementação do algoritmo de compressão de dados, detalhada nesta seção, marca um avanço significativo na otimização da análise de anomalias proposta nesta pesquisa. A conversão dos arquivos no formato *CSV*, derivados das PMUs, para o compacto arquivo no formato *parquet* não apenas evidencia uma redução substancial no tamanho dos dados, mas também reflete em melhorias notáveis na velocidade de leitura e eficiência de processamento. Através desta metodologia aplicada, os dados se tornam mais acessíveis para análise preditivas e tomadas de decisões estratégicas em tempo real.

4.2 Experimentos

Com os dados preparados e armazenados em um formato otimizado para acesso, esta pesquisa avança para a fase experimental. A utilização de métodos de aprendizado de máquina não supervisionados emerge como o próximo curso de ação, aplicando técnicas avançadas aos dados acumulados para extrair padrões. Esta abordagem permite a identificação de padrões e anomalias dentro do conjunto de dados, sem a necessidade de intervenção ou direcionamento manual, o que é fundamental para a análise objetiva.

Para a execução dos códigos desenvolvidos em linguagem de programação Python e a implementação dos experimentos delineados nesta pesquisa, optou-se pela utilização do **Google Colaboratory** (GOOGLE, 2023), um ambiente de computação em nuvem provido pela empresa Google. Esta plataforma, acessível na modalidade gratuita, foi selecionada por oferecer um ambiente de desenvolvimento integrado e hospedado que facilita a execução de *scripts* e a utilização das ferramentas analíticas necessárias.

4.2.1 Detectores de Anomalias com Scikit-Learn

Após a transformação dos dados para o formato *parquet*, procedeu-se com a fase experimental, empregando algoritmos de aprendizado de máquina não supervisionados, utilizando a biblioteca ***Scikit-Learn*** (SCIKIT-LEARN, 2023). Neste contexto, empregou-se uma variável designada como *contaminantes* para quantificar a proporção estimada de *outliers*, o que é crucial para estabelecer o limiar nos escores das amostras dentro do conjunto de dados. Tal parâmetro é instrumental para calibrar os algoritmos de detecção e garantir a precisão dos resultados obtidos na identificação de desvios de anomalias.

Na etapa subsequente, o algoritmo configura uma estrutura de dados do tipo *array* listada como ***detectores***, a qual é composta por instâncias de classes de modelos analíticos escolhidos para a operação de identificação de anomalias. Essa seleção abrange classes distintas, cada uma destacando-se por suas capacidades específicas de discernimento de padrões anômalos nos dados. Entre essas classes destaca-se:

- ***OneClasseSVM*:** Aplica uma técnica de máquina de vetores focando na detecção de novidades (SCIKIT-LEARN, 2024d);
- ***EllipticEnvelope*:** Identifica *outliers* sob a premissa de uma distribuição normal dos dados (SCIKIT-LEARN, 2024a);
- Combinação entre ***Nystroem*** e ***SGDOneClassSVM*:** Uma *pipeline* combinando ***Nystroem*** para a aproximação de função *kernel* e ***SGDOneClassSVM*** para a detecção de *outliers* (SCIKIT-LEARN, 2024e);
- ***IsolationForest*:** Utiliza uma abordagem baseada em árvores para o isolamento de anomalias (SCIKIT-LEARN, 2024b);
- ***LocalOutlierFactor*:** Tem por base a densidade local dos dados par discernir valores anômalos (SCIKIT-LEARN, 2024c).

Uma lista dos eventos alvo da pesquisa foi organizada e acondicionada em uma estrutura de dados do tipo *array*, a qual é iterativamente utilizada para identificar e coletar os dados pertinentes à análise. Durante essas iterações, um *DataFrame*

é estabelecido para acondicionar as variáveis explicativas e a variável de resposta. Dentre os dados extraídos, a taxa de variação da frequência é particularmente capturada e reservada para avaliações futuras.

Cada um dos detectores, munido de um conjunto específico de dados, emite previsões acerca da presença ou ausência de anomalias. Essas previsões são tratadas e transcodificadas para um formato binário, no qual os *outliers* são assinalados pelo indicativo numeral 1 e os dados considerados normais recebem a marcação 0. Este processo gera uma fase subsequente de análise.

Prosseguindo para as análises do processamento efetuado pelos algoritmos de detecção de anomalias, conduzindo uma análise comparativa entre os algoritmos. A metodologia adotada conta com a definição de uma fração de contaminação estimada como parâmetro das métricas, configurada para refletir a proporção de *outliers* esperados no conjunto de dados. Para esta análise utilizou-se de vários algoritmos, cada qual com suas especificidades e parâmetros ajustados.

Os experimentos foram conduzidos sobre uma lista de eventos, onde, para cada um, os dados foram carregados e uma janela temporal foi estabelecida para a análise. O conjunto de dados foi separado em características e variáveis-alvo, que representam os rótulos verdadeiros para o período do evento. Os detectores foram então aplicados para prever a natureza dos dados, distinguindo entre *Normais* e *Anômalos*. Para os modelos de classificação, como ***RandomForestClassifier*** e ***GradientBoostingClassifier***, que requerem dados de treinamentos com etiquetas, foi adotado o procedimento de ajuste e predição convencional do algoritmo.

As previsões geradas por cada detector foram submetidas a uma conversão para um formato binário, facilitando a avaliação subsequente. Utilizando uma série de algoritmos de métricas de desempenho, incluindo a acurácia balanceada e a pontuação ***F1***, as predições foram minuciosamente avaliadas. O objeto era quantificar a eficácia dos detectores em identificar corretamente os *outliers* dentro do período que está sendo comparado.

Os resultados da análise de desempenho forma registrados, culminando na compilação de um *DataFrame* consolidado que reflete as métricas de avaliação para cada algoritmo. Esta abordagem permitiu uma análise comparativa dos diferentes

métodos de detecção de anomalias, contribuindo para a compreensão dos algoritmos de aprendizado de máquinas não supervisionados em contextos dinâmicos e vários parâmetros a serem considerados.

A tabela 2 consolida os desempenhos de vários métodos de detecção de anomalias aplicados ao conjunto de dados processados pela biblioteca ***Scikit-Learn***. Cada linha corresponde a um evento específico e sua análise subsequente por meio de diferentes detectores de anomalias. As métricas variam significativa entre os métodos e eventos, servido de indicador para determinar a adequação de um método ao contexto apresentado nesta pesquisa. As colunas são descritas de forma detalhada a seguir:

- **Evento:** Data e hora do evento analisado;
- **Detector:** Nome do método de detecção utilizado;
- **Acurácia balanceada:** Métrica de avaliação que considera o desempenho do modelo em todas as classes;
- **F1:** A pontuação **F1** é a média harmônica da precisão e da sensibilidade. É uma medida que avalia o modelo em termos de precisão e *recall*;
- **Precisão:** Indica a proporção e eventos identificados corretamente como anomalias (verdadeiros positivos) em relação ao total de eventos identificados como anomalias (verdadeiros positivos + falsos positivos);
- **Recall (sensibilidade):** Mede a proporção de eventos reais de anomalias que foram corretamente identificados. Esta métrica garante que a maioria das verdadeiras anomalias sejam detectadas;
- ***Fowlkes-Mallows Score (FMS)*:** Métrica da similaridade entre os grupos verdadeiros e os previstos, com base na precisão e sensibilidade;
- ***Adjuted Mutual Info (MIS)*:** Métrica de ajuste que compara a similaridade entre dois agrupamentos, ajustada pela chance;
- ***Completeness Score (CPS)*:** Mede se todos os pontos de dados que são membros de uma classe verdadeira são elementos do mesmo grupo previsto;

- ***Homogeneity Score (HGS)*:** Mede se cada grupo previsto contém apenas membros de uma única classe verdadeira;
- **Tempo de treino:** Tempo em segundos utilizado para o treinamento;

No estudo realizado por Hannon et al. (2021) sobre a detecção e classificação de anomalias em dados de PMUs em tempo real, a importância de um *framework* interpretável e eficiente é enfatizada, visando auxiliar os operadores do sistema elétrico em suas decisões críticas. Este estudo é fundamental para a metodologia de medição adotada nesta pesquisa, onde é focado particularmente nos valores das métricas de ***F1 score***, **Precisão** e ***Recall***. Embora Hannon et al. (2021) não mencione explicitamente essas três métricas, elas estão intrinsecamente alinhadas com os objetivos do estudo, que são detalhados a seguir:

- **Interpretabilidade e Informação útil para operadores:** Esta dimensão garante uma compreensão clara e mensurável do desempenho do modelo, especialmente em termos de acurácia na detecção de eventos reais e na eficácia em minimizar falsos positivos. Esta característica é essencial para que os operadores possam confiar e agir com base nas informações fornecidas pelo sistema.
- **Balanceamento entre detecção de anomalias e falsos positivos:** No contexto específico do Sistema Interligado Nacional, o não reconhecimento de uma anomalia real (falsos negativos) e a indicação errônea de uma anomalia (falsos positivos) podem ter implicações sérias. A utilização das métricas de precisão e *recall* é crucial para uma avaliação balanceada desses dois tipos de erros. O *F1 score*, ao combinar harmoniosamente a precisão e o *recall*, oferece uma medida compreensiva que equilibra estes aspectos críticos.
- **Eficiência computacional:** É imperativo que o método de detecção de anomalias seja não somente preciso, mas também eficiente e capaz de proporcionar respostas rápidas. A agilidade no processamento e na resposta é fundamental para a tomada de decisão em tempo real.

- **Generalização e transferibilidade:** O *framework* proposto por Hannon et al. (2021) almeja ser versátil e aplicável a uma ampla gama de sistemas de infraestrutura. A adoção das métricas mencionadas facilita uma avaliação consistente e comparável, possibilitando sua aplicação e validação em diferentes sistemas e contextos operacionais.

Na interpretação dos resultados obtidos através das métricas, e com base no *framework* delineado por Hannon et al. (2021), daremos ênfase particular aos indicadores de **Acurácia Balanceada**, **F1-Score** e **Precisão** para determinar o detector de anomalias mais eficiente dentre as bibliotecas analisadas. Estas métricas foram escolhidas por sua capacidade de fornecer uma avaliação abrangente e equilibrada do desempenho dos detectores, considerando tanto a precisão na identificação de anomalias reais quanto a minimização de falsos positivos. A combinação destes indicadores será fundamental para identificar qual biblioteca oferece a melhor ferramenta para detecção de anomalias, alinhando-se assim com os objetivos e critérios de eficácia estabelecidos neste trabalho.

Tabela 2 – Métricas de anomalias apuradas pela biblioteca Scikit-Learn.

Evento	Detector	Acur. Balanceada	F1	Prec.	Rec.	FMS	MIS	CPS	HGS	Tempo Treino (S)
2021-04-08 18:34:00	EllipticEnvelope	65,05	40	50	33,33	90,4	10,37	16,82	12,61	2,7038
2021-04-08 18:34:00	EllipticEnvelope	65,05	40	50	33,33	90,4	10,37	16,82	12,61	1,786
2021-04-08 18:34:00	EllipticEnvelope	65,05	40	50	33,33	90,4	10,37	16,82	12,61	1,6245
2021-04-08 18:34:00	OneClassSVM	50	16,22	8,82	100	91,34	0	100	0	0,0675
2021-04-08 18:34:00	OneClassSVM	50	16,22	8,82	100	91,34	0	100	0	0,0022
2021-04-08 18:34:00	OneClassSVM	46,77	0	0	0	85,01	-2,48	2,51	1,88	0,0028
2021-04-08 18:34:00	Pipeline	50	0	0	0	91,34	0	100	0	0,0087
2021-04-08 18:34:00	Pipeline	50	0	0	0	91,34	0	100	0	0,0185

Continuação da tabela 2

Evento	Detector	Acur. Balanceada	F1	Prec.	Rec.	FMS	MIS	CPS	HGS	Tempo Treino (S)
2021-04-08 18:34:00	Pipeline	50	0	0	0	91,34	0	100	0	0,0141
2021-04-08 18:34:00	Pipeline	50	0	0	0	91,34	0	100	0	0,0253
2021-04-08 18:34:00	IsolationForest	76,88	44,44	33,33	66,67	83,48	11,5			
2021-04-08 18:34:00	IsolationForest	93,55	60	42,86	100	86,28	36,3	31,04	52,89	0,3043
2021-04-08 18:34:00	IsolationForest	76,88	44,44	33,33	66,67	83,48	11,5	12,71	19,84	0,0627
2021-04-08 18:34:00	IsolationForest	75,27	40	28,57	66,67	80,53	8,2	9,73	16,58	0,0647
2021-04-08 18:34:00	LocalOutlierFactor	48,39	0	0	0	88,11	-3,16	2,08	0,92	0,0068
2021-04-08 18:34:00	LocalOutlierFactor	48,39	0	0	0	88,11	-3,16	2,08	0,92	0,0071
2021-04-08 18:34:00	LocalOutlierFactor	50	0	0	0	91,34	0	100	0	0,0316

Continuação da tabela 2

Evento	Detector	Acur. Balanceada	F1	Prec.	Rec.	FMS	MIS	CPS	HGS	Tempo Treino (S)
2021-04-08 18:34:00	LocalOutlierFactor	63,44	33,33	33,33	33,33	87,18	3,62	8,1	8,1	0,0322
2021-04-08 18:34:00	LocalOutlierFactor	50	0	0	0	91,34	0	100	0	0,0304
2021-04-08 18:34:00	LocalOutlierFactor	50	0	0	0	91,34	0	100	0	0,0287
2021-04-08 18:34:00	DBSCAN	50	16,22	8,82	100	91,34	0	100	0	0,0062
2021-04-08 18:34:00	KMeans	50	0	0	0	91,34	0	100	0	0,1539
2021-04-08 18:34:00	AgglomerativeClustering	50	0	0	0	91,34	0	100	0	
2021-05-28 11:06:00	EllipticEnvelope	50	0	0	0	91,34	0	100	0	0,9541
2021-05-28 11:06:00	EllipticEnvelope	50	0	0	0	91,34	0	100	0	0,8763
2021-05-28 11:06:00	EllipticEnvelope	50	0	0	0	91,34	0	100	0	1,0454

Continuação da tabela 2

Evento	Detector	Acur. Balanceada	F1	Prec.	Rec.	FMS	MIS	CPS	HGS	Tempo Treino (S)
2021-05-28 11:06:00	OneClassSVM	50	16,22	8,82	100	91,34	0	100	0	0,0521
2021-05-28 11:06:00	OneClassSVM	50	16,22	8,82	100	91,34	0	100	0	0,0029
2021-05-28 11:06:00	OneClassSVM	40,32	0	0	0	74,15	0,21	3,87	6,04	0,0027
2021-05-28 11:06:00	Pipeline	50	0	0	0	91,34	0	100	0	0,0055
2021-05-28 11:06:00	Pipeline	50	0	0	0	91,34	0	100	0	0,0146
2021-05-28 11:06:00	Pipeline	50	0	0	0	91,34	0	100	0	0,0057
2021-05-28 11:06:00	Pipeline	50	0	0	0	91,34	0	100	0	0,0089
2021-05-28 11:06:00	IsolationForest	66,67	50	100	33,33	93,74	32,6	57,67	25,64	0,5235
2021-05-28 11:06:00	IsolationForest	66,67	50	100	33,33	93,74	32,6	57,67	25,64	0,4449

Continuação da tabela 2

Evento	Detector	Acur. Balanceada	F1	Prec.	Rec.	FMS	MIS	CPS	HGS	Tempo Treino (S)
2021-05-28 11:06:00	IsolationForest	66,67	50	100	33,33	93,74	32,6	57,67	25,64	0,0888
2021-05-28 11:06:00	IsolationForest	66,67	50	100	33,33	93,74	32,6	57,67	25,64	0,0893
2021-05-28 11:06:00	LocalOutlierFactor	50	0	0	0	91,34	0	100	0	0,0473
2021-05-28 11:06:00	LocalOutlierFactor	50	0	0	0	91,34	0	100	0	0,0102
2021-05-28 11:06:00	LocalOutlierFactor	50	0	0	0	91,34	0	100	0	0,0224
2021-05-28 11:06:00	LocalOutlierFactor	50	0	0	0	91,34	0	100	0	0,0257
2021-05-28 11:06:00	LocalOutlierFactor	50	0	0	0	91,34	0	100	0	0,0244
2021-05-28 11:06:00	LocalOutlierFactor	50	0	0	0	91,34	0	100	0	0,0264
2021-05-28 11:06:00	DBSCAN	50	16,22	8,82	100	91,34	0	100	0	0,0896

Continuação da tabela 2

Evento	Detector	Acur. Balanceada	F1	Prec.	Rec.	FMS	MIS	CPS	HGS	Tempo Treino (S)
2021-05-28 11:06:00	KMeans	50	0	0	0	91,34	0	100	0	0,187
2021-05-28 11:06:00	AgglomerativeClustering	50	0	0	0	91,34	0	100	0	0,0264
2021-05-28 11:26:00	EllipticEnvelope	81,72	66,67	66,67	66,67	93,16	34,61	37,64	37,64	0,8663
2021-05-28 11:26:00	EllipticEnvelope	81,72	66,67	66,67	66,67	93,16	34,61	37,64	37,64	0,6962
2021-05-28 11:26:00	EllipticEnvelope	81,72	66,67	66,67	66,67	93,16	34,61	37,64	37,64	1,1607
2021-05-28 11:26:00	OneClassSVM	50	16,22	8,82	100	91,34	0	100	0	0,0425
2021-05-28 11:26:00	OneClassSVM	50	16,22	8,82	100	91,34	0	100	0	0,0024
2021-05-28 11:26:00	OneClassSVM	43,55	0	0	0	79,24	-1,2	3,2	3,89	0,003
2021-05-28 11:26:00	Pipeline	50	0	0	0	91,34	0	100	0	0,0049

Continuação da tabela 2

Evento	Detector	Acur. Balanceada	F1	Prec.	Rec.	FMS	MIS	CPS	HGS	Tempo Treino (S)
2021-05-28 11:26:00	Pipeline	50	0	0	0	91,34	0	100	0	0,0132
2021-05-28 11:26:00	Pipeline	50	0	0	0	91,34	0	100	0	0,0055
2021-05-28 11:26:00	Pipeline	50	0	0	0	91,34	0	100	0	0,009
2021-05-28 11:26:00	IsolationForest	96,77	75	60	100	92,92	53,58	47,77	66,84	0,4001
2021-05-28 11:26:00	IsolationForest	96,77	75	60	100	92,92	53,58	47,77	66,84	0,2972
2021-05-28 11:26:00	IsolationForest	96,77	75	60	100	92,92	53,58	47,77	66,84	0,0588
2021-05-28 11:26:00	IsolationForest	96,77	75	60	100	92,92	53,58	47,77	66,84	0,0587
2021-05-28 11:26:00	LocalOutlierFactor	65,05	40	50	33,33	90,4	10,37	16,82	12,61	0,005
2021-05-28 11:26:00	LocalOutlierFactor	50	0	0	0	91,34	0	100	0	0,0061

Continuação da tabela 2

Evento	Detector	Acur. Balanceada	F1	Prec.	Rec.	FMS	MIS	CPS	HGS	Tempo Treino (S)
2021-05-28 11:26:00	LocalOutlierFactor	83,33	80	100	66,67	96,62	62,49	74,9	56,14	0,0176
2021-05-28 11:26:00	LocalOutlierFactor	65,05	40	50	33,33	90,4	10,37	16,82	12,61	0,0195
2021-05-28 11:26:00	LocalOutlierFactor	50	0	0	0	91,34	0	100	0	0,0166
2021-05-28 11:26:00	LocalOutlierFactor	50	0	0	0	91,34	0	100	0	0,0169
2021-05-28 11:26:00	DBSCAN	50	16,22	8,82	100	91,34	0	100	0	0,0051
2021-05-28 11:26:00	KMeans	50	0	0	0	91,34	0	100	0	0,0841
2021-05-28 11:26:00	AgglomerativeClustering	50	0	0	0	91,34	0	100	0	

De acordo com as métricas apresentadas, remete que o detector ***IsolationForest*** desempenha uma melhor identificação de anomalias em vários eventos. Como exemplo, é explorado os dados do **Evento** nos intervalos de **"2021-04-08 18:34:00"** e **"2021-05-28 11:26:00"**, onde o algoritmo alcançou altas pontuações nas métricas de *F1 score*, Precisão e *Recall*, o que sugere a eficácia da detecção de anomalias.

4.2.2 Detectores de Anomalias com PyOD

PyOD é uma biblioteca de detecção de anomalias em Python, de grande relevância por sua capacidade de lidar com dados multivariados de forma escalável e eficiente. Zhao, Nasrullah e Li (2019) afirmam que esta ferramenta é uma solução versátil que engloba uma ampla variedade de técnicas de detecção de anomalias, desde métodos clássicos até abordagens inovadoras baseadas em redes neurais. Esta biblioteca emprega práticas de desenvolvimento como teste unitários e integração contínua, garantindo confiabilidade e qualidade.

Como próximo passo na experimentação da análise nesta pesquisa, a utilização da biblioteca PyOD (PYOD, 2023) inicia sua análise configurando uma variável de *contaminantes* definindo limites que indicam a proporção esperada de anomalias nos dados analisados. Uma lista de detectores é inicializada, cada um configurado com este limite de contaminação. Seguindo a execução o algoritmo cria uma iteração sob os eventos. Para cada evento, há uma carga dos dados relacionados, abordando uma janela de tempo na qual a análise é focada.

Os dados de entrada e os rótulos verdadeiros são extraídos para avaliação. Apenas os rótulos dentro da janela de tempo são considerados, indicando um interesse nos resultados para a detecção das anomalias. Dados adicionais denominados ***rocof***, são coletados para cada evento avaliado.

Para cada detectores de anomalias é conduzido em uma iteração onde este é treinado e as previsões são efetuadas mediante os dados de entrada. As previsão são então transformadas onde os valores são armazenados nos numerais 1 para anômalos e 0 para normais. O algoritmo ainda procede o cálculo de várias métricas de desempenho já utilizadas na análise de dados da biblioteca ***Sciki-Learn***. Essa

abordagem multidimensional para a avaliação de desempenho, permite uma análise compreensiva sobre como cada modelo faz a detecção de anomalias.

Os resultados são então agregados em um dicionário e anexados a uma lista de métricas, que mais tarde são convertidos em uma estrutura de dados *DataFrame* para impressão e análise dos resultados abordados na tabela 3. A tabela apresenta os resultados de uma série de experimento de detecção de anomalias utilizando diferentes algoritmos. Cada linha da tabela corresponde a um evento individual, onde diversos detectores, como ***ECOD***, ***ABOD***, ***CBLOFA***, entre outros, foram aplicados para identificar desvios padrão. As métricas supracitadas na seção 4.2.1 deste capítulo, fornecem uma visão compreensiva sobre a qualidade da detecção de valores anômalos relevantes à esta pesquisa bem como os campos desta tabela, já detalhados na biblioteca ***Scikit-Learn***.

A análise da biblioteca **PyOD** sugere que os algoritmos testados tiveram um desempenho abaixo do esperado no que se refere à detecção de anomalias. Notoriamente os resultados em todas as métricas aplicadas a cada detector os valores permaneceram em 0, apontando que o algoritmo não encontrou nenhum similaridade nas previsões de anomalias.

Tabela 3 – Métricas de anomalias apuradas pela biblioteca PyOD.

Evento	Detector	Acur. Balanceada	F1	Prec.	Rec.	FMS	MIS	CPS	HGS	Tempo Treino (S)
2021-04-08 18:34:00	ECOD	50	0	0	0	91,34	0	100	0	1,7219
2021-04-08 18:34:00	ABOD	50	0	0	0	91,34	0	100	0	5,7078
2021-04-08 18:34:00	CBLOF	50	0	0	0	91,34	0	100	0	4,7623
2021-04-08 18:34:00	COF	50	0	0	0	91,34	0	100	0	0,1851
2021-04-08 18:34:00	COPOD	50	0	0	0	91,34	0	100	0	0,0253
2021-04-08 18:34:00	KDE	50	0	0	0	91,34	0	100	0	0,0487
2021-04-08 18:34:00	HBOS	50	0	0	0	91,34	0	100	0	4,6137
2021-04-08 18:34:00	LOF	50	0	0	0	91,34	0	100	0	0,0581

Continuação da tabela 3

Evento	Detector	Acur. Balanceada	F1	Prec.	Rec.	FMS	MIS	CPS	HGS	Tempo Treino (S)
2021-04-08 18:34:00	OCSVM	50	0	0	0	91,34	0	100	0	0,0482
2021-04-08 18:34:00	PCA	50	0	0	0	91,34	0	100	0	0,1165
2021-04-08 18:34:00	SOD	50	0	0	0	91,34	0	100	0	8,6569
2021-05-28 11:06:00	ABOD	50	0	0	0	91,34	0	100	0	0,0523
2021-05-28 11:06:00	CBLOF	50	0	0	0	91,34	0	100	0	0,2038
2021-05-28 11:06:00	COF	50	0	0	0	91,34	0	100	0	0,1214
2021-05-28 11:06:00	KDE	50	0	0	0	91,34	0	100	0	0,0184
2021-05-28 11:06:00	HBOS	50	0	0	0	91,34	0	100	0	0,0556
2021-05-28 11:06:00	LOF	50	0	0	0	91,34	0	100	0	0,007

Continuação da tabela 3

Evento	Detector	Acur. Balanceada	F1	Prec.	Rec.	FMS	MIS	CPS	HGS	Tempo Treino (S)
2021-05-28 11:06:00	OCSVM	50	0	0	0	91,34	0	100	0	0,0252
2021-05-28 11:06:00	PCA	50	0	0	0	91,34	0	100	0	0,0078
2021-05-28 11:06:00	SOD	50	0	0	0	91,34	0	100	0	0,1446

4.2.3 Detectores de Anomalias com ADTK

A utilização da biblioteca ***Anomaly Detection Toolkit (ADTK)*** é aplicado métodos estatísticos e baseados em aprendizado de máquina para identificar padrões atípicos em dados temporais. Inicialmente o nível de contaminação é configurado para indicar a proporção esperada de pontos anômalos nos dados. Vários detectores são instanciados com parâmetros específicos, incluindo as funções ***SeasonalAD*** para identificar desvios de padrões sazonais, ***QuantileAD*** para detectar valores extremos usando quantis, e ***OutlierDetector*** que emprega o ***Isolation Forest***.

Adicionalmente o algoritmo faz uso da função ***Generalized ESD Test for Anomalies***, um teste estatístico robusto para *outliers* externos, e do ***MinClusterDetector***,que busca pequenos aglomerados de anomalias. Para uma abordagem baseada em redução de dimensionalidade, o ***PcaAD*** é incluído, o qual utiliza a análise de componentes principais para encontrar discrepâncias nos dados.

Apontando para cada evento, uma validação da série temporal é conduzida, garantindo que os dados estejam em um formato adequado para análise. Após a validação, cada detector é aplicado sequencialmente. Os resultados são coletados individualmente, permitindo uma comparação direta entre os métodos. Com o intuito de facilitar a interpretação dos resultados, onde os pontos anômalos são destacados e os componentes principais das séries temporais são delineados para análise comparativa.

Ao realizar a análise estatística, é convertido os resultados das detecções e os valores reais para formatos apropriados, aplicando métricas como acurácia balanceada, pontuação F1, precisão, *recall*, marcação de tempo de treinamento, dentre outras. Os resultados foram armazenados em um ***DataFrame*** da biblioteca **Pandas**, proporcionando uma visão detalhada do desempenho de cada detector listados na tabela 4.

A figura 10 exibe uma série temporal aplicado à função de ***OutlierDetector*** gerando o gráfico sobre os dados de ***rocof*** de mudança de frequência numa rede elétrica, medidos a variação ao longo do tempo. O gráfico mostra um comportamento bastante estável e periódico da variável medida até um ponto de anomalia detectada,

sombreada em vermelho com o comportamento usual da série temporal estudada. A biblioteca **ADTK** foi capaz de localizar *outliers* nos eventos já catalogados apontados pelo ONS (ONS, 2023a) nos dias 08/04/2021 às 18:34 e 28/05/2021 às 11h26, demonstrando sua eficiência na detecção de anomalias nos dados estudados. Um aspecto particularmente intrigante foi a detecção de possíveis anomalias fora do período previamente catalogado, indicando variações significativas que podem ser atribuídas a oscilações na rede em outros pontos do Sistema Interligado Nacional conforme apontado na figura 11. Essas observações apontam para a possibilidade de que tais anomalias possam originar-se de diferentes localidades ou até mesmo da própria localidade da concessionária de energia elétrica.

Tabela 4 – Métricas de anomalias apuradas pela biblioteca ADTK.

Evento	Detector	Acur. Balanceada	F1	Prec.	Rec.	FMS	MIS	CPS	HGS	Tempo Treino (S)
2021-04-08 18:34:00	OutlierDetector	93,55	60	42,86	100	86,28	36,3	31,04	52,89	0,3154
2021-04-08 18:34:00	MinClusterDetector	93,55	60	42,86	100	86,28	36,3	31,04	52,89	0,378
2021-04-08 18:34:00	PcaAD	50	0	0	0	91,34	0	100	0	0,2303
2021-05-28 11:06:00	OutlierDetector	66,67	50	100	33,33	93,74	32,6	57,67	25,64	0,3784
2021-05-28 11:06:00	MinClusterDetector	66,67	50	100	33,33	93,74	32,6	57,67	25,64	0,5885
2021-05-28 11:06:00	PcaAD	50	0	0	0	91,34	0	100	0	0,1026
2021-05-28 11:26:00	OutlierDetector	96,77	75	60	100	92,92	53,58	47,77	66,84	0,4195
2021-05-28 11:26:00	MinClusterDetector	96,77	75	60	100	92,92	53,58	47,77	66,84	0,6751

Continuação da tabela 4

Evento	Detector	Acur. Balanceada	F1	Prec.	Rec.	FMS	MIS	CPS	HGS	Tempo Treino (S)
2021-05-28 11:26:00	PcaAD	81,72	66,67	66,67	66,67	93,16	34,61	37,64	37,64	0,2176

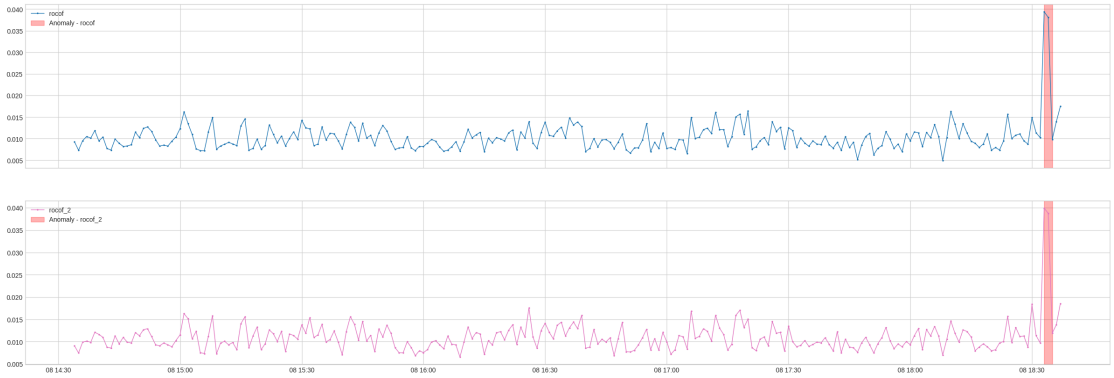

Figura 10 – Gráfico utilizando a função OutlierDetector no dia 08/04/2021 às 18h34
fonte: Do autor (2023)

4.3 Discussões

A escolha pela compressão dos arquivos extraídos das PMUs de análise para o formato ***parquet*** emergiu como uma resposta estratégica às exigências de processamento e manipulação de grandes volumes de dados. Dada a magnitude dos dados, que alcançam a ordem de milhões, a transição para um formato mais eficiente em termos de armazenamento e gerenciamento tornou-se crucial para melhorar a velocidade de processamento e leitura, otimizando assim a implementação de

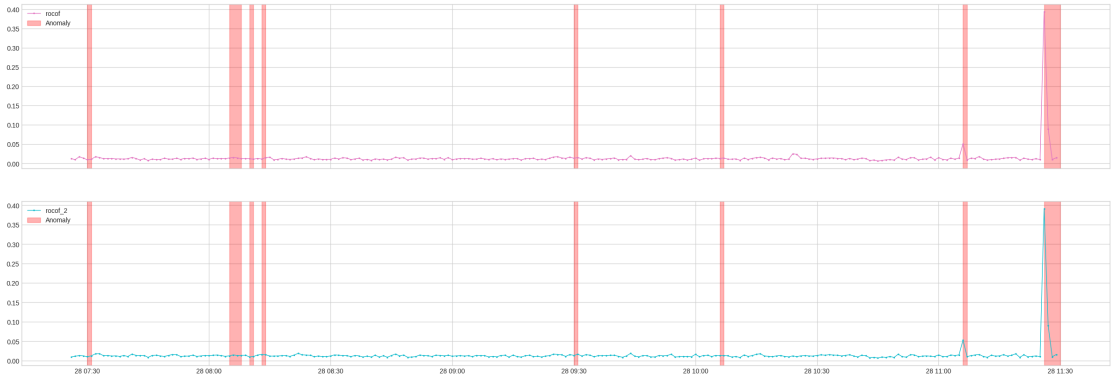

Figura 11 – Gráfico utilizando a função OutlierDetector no dia 08/05/2021 às 11h26
fonte: Do autor (2023)

modelos analíticos complexos.

A eficácia desta abordagem é claramente ilustrada na notável redução do tamanho dos arquivos. Antes da compressão, um arquivo típico em formato CSV poderia ocupar 1.024.009 KB, enquanto que, após a compressão para o formato *parquet*, o mesmo arquivo foi reduzido para meros 101 KB. Esta compressão notável não apenas economiza espaço de armazenamento, mas também facilita uma leitura mais rápida e eficiente dos dados, essencial para a análise.

Essa melhoria na gestão dos dados marcou um avanço substancial no processo de análise de informações, onde a rapidez de leitura e processamento tornou-se um fator chave. Tal aprimoramento foi particularmente relevante na detecção de anomalias, pois permitiu aos algoritmos utilizados operar com maior eficiência, realçando a importância da compactação e otimização de dados em estudos de grande escala.

A implementação dos algoritmos das bibliotecas **Scikit-Learn**, **PyOD** e **ADTK** resultou em diversas análises relevantes, contribuindo significativa para a questão central desta pequisa que trata da viabilidade da aplicação de algoritmos de detecção de anomalias em modelos não supervisionados com dados oriundos de PMUs para a identificação de problemas relevantes e tomada de decisão.

Com a utilização da biblioteca **Scikit-Learn**, foi observado uma alta precisão na identificação dos eventos previamente definidos no escopo desta pesquisa. Vários detectores formam eficientes não apenas na identificação, mas também na representação visual destes eventos. Além disso, foram identificadas outras anomalias que podem estar relacionadas a eventos no Sistema Interligado Nacional ou mesmo na área de cobertura da concessionária de energia elétrica.

A análise dos resultados alcançados através da implementação das métricas disponíveis na biblioteca **Scikit-Learn** e seguindo o *framework* proposto por Hannon et al. (2021), observa-se a efetividade dos algoritmos ***IsolationForest*** ***EllipticEnvelope***. Especificamente, nos eventos ocorridos em 08/04/2021 às 18h34 em 28/05/2021 às 11h26, respectivamente, estes algoritmos demonstraram capacidade de prever anomalias fora do domínio geográfico da concessionária de energia elétrico. Este resultado sugere que os algoritmos é capaz corroborar à

hipótese levantada neste trabalho.

Por outro lado, a aplicação da biblioteca **PyOD** mostrou-se menos eficiente na detecção de anomalias nos dados das PMUs. Os resultados indicaram *outliers* distintos dos definidos como referência, e os detectores não foram tão precisos na identificação de eventos relevantes, em comparação com os resultados obtidos pela biblioteca **Scikit-Learn**. Constatou-se que esta biblioteca não se mostrou adequada para a identificação de eventos em redes elétricas.

A análise dos dados com a biblioteca **ADTK** revelou uma consistência mais apurada com os resultados já obtidos pela biblioteca **Sciki-Learn**. Os detectores não só corroboraram os *outliers* identificados nos eventos marcados como tutores, mas também os visualizaram graficamente, proporcionando uma análise temporal clara. Esta biblioteca identificou com precisão os eventos marcados como anomalias e também revelou outras discrepâncias potencialmente significativas dentro do SIN.

As análises realizadas indicam a preferência pelas bibliotecas **Scikit-Learn** e **ADTK**, que demonstram consistência na identificação de anomalias e também revelaram *outiliers* relevantes fora do escopo previamente estabelecido. Estas bibliotecas possuem algoritmos capazes de detectar desvios no padrão de normalidade, mesmo em locais distantes ou fora da área de cobertura direta da concessionária. Esta capacidade é de extrema importância para a análise e o monitoramento eficaz de sistemas de energia elétrica abrindo caminho para futuras análises e aprimoramentos dos algoritmos, destacando a eficácia destas ferramentas na análise e na identificação de anomalias em redes de energia elétrica.

A Tabela 5 sintetiza os resultados da análise, destacando que os detectores ***IsolationForest***, ***EllipticEnvelope***, e ***OutlierDetector*** demonstraram um desempenho notável e consistente na identificação de anomalias. Estes detectores se destacaram na classificação, alinhando-se eficientemente com o *framework* proposto por Hannon et al. (2021) sendo observado as colunas de Acurácia balanceada, F1-Score e Precisão. Este alinhamento é de grande importância, pois evidencia que esses detectores não apenas identificam anomalias de forma precisa, mas também o fazem de uma maneira que corroboram com os princípios estabelecidos no *framework*. Tal consistência nos resultados reafirma a relevância desses algoritmos específicos

para a detecção de anomalias em sistemas de energia elétrica, enfatizando a sua utilidade prática e confiabilidade na análise de dados de PMU.

Tabela 5 – Métricas com dados consistentes de anomalias apontados pelas bibliotecas Scikit-Learn e ADTK.

Bibl.	Evento	Detector	Acur. Balanceada	F1	Prec.	Rec.	FMS	MIS	CPS	HGS	Tempo Treino (s)
Scikit-Learn	2021-04-08 18:34:00	IsolationForest	93,55	60	42,86	100	86,28	36,3	31,04	52,89	0,3043
Scikit-Learn	2021-05-28 11:26:00	EllipticEnvelope	81,72	66,67	66,67	66,67	93,16	34,61	37,64	37,64	0,8663
Scikit-Learn	2021-05-28 11:26:00	EllipticEnvelope	81,72	66,67	66,67	66,67	93,16	34,61	37,64	37,64	0,6962
Scikit-Learn	2021-05-28 11:26:00	EllipticEnvelope	81,72	66,67	66,67	66,67	93,16	34,61	37,64	37,64	1,1607
ADTK	2021-04-08 18:34:00	OutlierDetector	93,55	60	42,86	100	86,28	36,3	31,04	52,89	0,3154
ADTK	2021-05-28 11:26:00	OutlierDetector	96,77	75	60	100	92,92	53,58	47,77	66,84	0,4195

5 Conclusão

O Sistema Interligado Nacional (SIN) constitui um marco tecnológico notável na infraestrutura de distribuição de energia elétrica do Brasil. Sua extensão continental implica uma série de desafios significativos, especialmente em termos de gerenciamento e avaliação eficaz da rede. Dentro desse contexto, um evento em uma localidade remota do sistema elétrico pode repercutir em outras partes desta intricada rede. Assim, a demanda por informações precisas e oportunas para prevenção e ação em casos de irregularidades se torna um aspecto crucial, como evidenciado neste estudo.

As PMUs, estrategicamente instalados ao longo do SIN, são essenciais para a coleta e análise de dados elétricos. Este trabalho destacou como a compressão de dados para o formato *parquet* facilitou uma análise mais ágil e eficaz, permitindo a rápida detecção e resposta a anomalias. A redução drástica no tamanho dos arquivos não apenas otimizou o armazenamento, mas também acelerou o processo de leitura dos dados, um avanço notável para a prática de monitoramento em tempo real.

A aplicação de algoritmos avançados de aprendizado de máquina, com ênfase nas bibliotecas ***Scikit-Learn*** e ***ADTK***, provou ser particularmente eficaz. Os resultados obtidos ilustram a precisão e a eficiência desses algoritmos na identificação de eventos pré-definidos e na detecção de novas anomalias, que podem indicar problemas significativos dentro da rede ou em sua periferia. A capacidade desses algoritmos de operar com alta eficiência sublinha a sua relevância para futuras implementações e aprimoramentos nos sistemas de detecção e análise. A análise detalhada da performance de diferentes algoritmos, conforme documentado na tabela 5, confirmam a eficácia destes algoritmos na identificação de anomalias, correlacionando com eventos reais documentados no SIN.

Os detectores ***IsolationForest*** e ***EllipticEnvelope*** emergiram como ferramentas robustas para a classificação e análise de eventos, alinhando-se eficientemente com o *framework* proposto por Hannon et al. (2021). A consistência dos resultados,

abrangendo acurácia balanceada, F1-Score e precisão, não só confirma a capacidade destes algoritmos de detectar anomalias de forma precisa, mas também sua adaptação ao complexo ambiente do SIN.

Este estudo, ao confrontar a hipótese central de que é possível identificar eventos significativos inclusive em locais geograficamente distantes, validou essa suposição com sucesso através de uma análise meticulosa. Os algoritmos demonstraram uma habilidade notável para correlacionar suas detecções com eventos reais, um aspecto que reforça a viabilidade de sua aplicação para a monitoração constante e abrangente do SIN.

Em suma, a integração da compressão de dados e o uso estratégico de algoritmos avançados de detecção de anomalias representam uma abordagem promissora e eficaz para o aprimoramento das operações do SIN. As ferramentas analisadas neste estudo não apenas melhoram as capacidades preventivas e corretivas da rede, mas também estabelecem um padrão para a gestão de dados e análise em larga escala, garantindo a confiabilidade e a segurança do fornecimento de energia elétrica aos consumidores brasileiros.

Portanto, a aplicação de técnicas de aprendizado de máquina, juntamente com a compressão de dados, se configura como uma estratégia extremamente relevante e eficaz, capaz de aperfeiçoar significativamente as ações preventivas e corretivas dentro do SIN, consolidando a qualidade e a confiabilidade do serviço fornecido aos usuários finais. A continuidade dessa pesquisa é essencial para explorar ainda mais as capacidades dessas tecnologias, visando a otimização contínua das práticas de monitoramento e análise de anomalias na rede elétrica nacional.

5.1 Trabalhos Futuros

Como direção para trabalhos futuros, propõe-se expandir as análises incluindo dados de PMUs de outras concessionárias de energia elétrica. Isso permitiria uma comparação mais abrangente dos resultados e testaria a confiabilidade dos algoritmos em uma variedade maior de cenários. Além disso, seria interessante explorar técnicas de aprendizado de máquina supervisionado para prever eventos futuros com base nos padrões identificados. Uma abordagem de *ensemble learning*

como explorado por Sagi e Rokach (2018), que combina os resultados de múltiplos modelos, poderia ser investigada para melhorar a precisão da detecção de anomalias.

A busca por novas técnicas de estudo também se fazem pertinentes com a investigação do uso do ***AnoGAN*** (*Anomaly Generation*) como técnica avançada para detecção de anomalias em dados de PMUs. O ***AnoGAN***, que se baseia em Redes Geradoras Adversariais (GANs), oferece uma abordagem promissora para identificação de padrões anormais sem a necessidade de extensos conjuntos de dados rotulados (SCHLEGL et al., 2017). Esta técnica poderia ser particularmente útil para analisar os dados em tempo real, permitindo não só a detecção de anomalias de forma mais eficiente e precisa, mas também a identificação de padrões sutis que possam preceder eventos críticos. Devido ao alto custo computacional, não foi possível a utilização desta técnica no contexto deste trabalho.

Implementar o ***AnoGAN*** no contexto da análise de dados de PMUs de diferentes concessionárias permitiria uma detecção de anomalias mais robusta e adaptativa, superando alguns dos desafios impostos por métodos tradicionais. Além disso, a integração do ***AnoGAN*** em uma ferramenta de monitoramento e alerta precoce, poderia significativamente elevar a capacidade do Operador Nacional do Sistema de prever e mitigar potenciais falhas na rede elétrica de forma proativa. A exploração dessa técnica de aprendizado profundo não supervisionado acrescentaria novo ferramental para o aprimoramento da segurança e eficiência do SIN, marcando um avanço significativo na gestão e operação de redes elétricas inteligentes.

Referências

AALAM, M. K.; SHUBHANGA, K. Power system event detection and localization—a new approach. *Electric Power Systems Research*, v. 223, p. 109553, 2023. ISSN 0378-7796. Disponível em: <https://www.sciencedirect.com/science/article/pii/S037877962300442X>. Citado na página 33.

ALIGHOLIAN, A. et al. Event detection in micro-pmu data: A generative adversarial network scoring method. In: *2020 IEEE Power Energy Society General Meeting (PESGM)*. [S.l.: s.n.], 2020. p. 1–5. Citado na página 33.

AMUTHA, A. L. et al. Anomaly detection in multivariate streaming pmu data using density estimation technique in wide area monitoring system. *Expert Systems with Applications*, v. 175, p. 114865, 2021. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417421003067>. Citado na página 15.

ANALYTICS, I. A. *Anomaly Detection Toolkit (ADTK) - Docs*. [S.l.], 2023. Disponível em: <https://adtk.readthedocs.io/en/stable/>. Acesso em: 21 set. 2023. Citado na página 41.

BONACCORSO, G. *Machine learning algorithms*. [S.l.]: Packt Publishing Ltd, 2017. Citado 2 vezes nas páginas 23 e 24.

BURONI, G. et al. On-board-unit data: A big data platform for scalable storage and processing. In: *2018 4th International Conference on Cloud Computing Technologies and Applications (Cloudtech)*. [S.l.: s.n.], 2018. p. 1–5. Citado na página 39.

CALIŃSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, Taylor & Francis, v. 3, n. 1, p. 1–27, 1974. Citado na página 31.

CUI, M. et al. A novel event detection method using pmu data with high precision. *IEEE Transactions on Power Systems*, v. 34, n. 1, p. 454–466, Jan 2019. ISSN 1558-0679. Citado na página 32.

DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1, n. 2, p. 224–227, 1979. Citado na página 31.

FERREIRA, E. H. M. *OutlierDetection_PMU_UNIFEI2023*. [S.l.]: GitHub, 2023. <https://github.com/eduardohen1/OutlierDetection_PMU_UNIFEI2023>. Citado na página 42.

GÉRON, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. [S.l.]: "O'Reilly Media, Inc.", 2022. 8 p. Citado 2 vezes nas páginas 7 e 24.

GÉRON, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. [S.l.]: "O'Reilly Media, Inc.", 2022. 10 p. Citado 2 vezes nas páginas 7 e 25.

GÉRON, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. [S.l.]: "O'Reilly Media, Inc.", 2022. 11 p. Citado 2 vezes nas páginas 7 e 26.

GÉRON, A. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. [S.l.]: "O'Reilly Media, Inc.", 2022. Citado 4 vezes nas páginas 22, 23, 25 e 26.

GITHUB. *Github*. [S.l.], 2024. Disponível em: <https://github.com/>. Acesso em: 02 mar. 2024. Citado na página 42.

GOMES, A. C. S. et al. O setor elétrico. Dba, 2002. Citado na página 13.

GOOGLE. *Conheça o Colab*. [S.l.], 2023. Disponível em: <https://colab.research.google.com/>. Acesso em: 21 set. 2023. Citado na página 50.

GREER, C. et al. Nist framework and roadmap for smart grid interoperability standards, release 3.0. Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD, 2014-10-01 00:10:00 2014. Disponível em: <https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=916755>. Citado na página 20.

HAI, A. A. et al. Transfer learning for event detection from pmu measurements with scarce labels. *IEEE Access*, v. 9, p. 127420–127432, 2021. Citado na página 34.

HANNON, C. et al. Real-time anomaly detection and classification in streaming pmu data. In: *2021 IEEE Madrid PowerTech*. [S.l.: s.n.], 2021. p. 1–6. Citado 5 vezes nas páginas 54, 55, 74, 75 e 78.

HARRISON, M. *Machine learning pocket reference: working with structured data in python*. [S.l.]: O'Reilly Media, 2019. Citado na página 29.

HASTIE, T. et al. *The elements of statistical learning: data mining, inference, and prediction*. [S.l.]: Springer, 2009. v. 2. Citado na página 29.

HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of classification*, Springer, v. 2, p. 193–218, 1985. Citado na página 31.

JAMES, G. et al. *An introduction to statistical learning*. [S.l.]: Springer, 2013. v. 112. Citado na página 28.

LAVIN, A.; AHMAD, S. Evaluating real-time anomaly detection algorithms – the numenta anomaly benchmark. In: *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*. [S.l.: s.n.], 2015. p. 38–44. Citado na página 42.

LEGER, A. S.; JAMES, J. Cyber-physical systems approach for wide area control applications. In: *2018 IEEE Texas Power and Energy Conference (TPEC)*. [S.l.: s.n.], 2018. p. 1–6. Citado na página 20.

LIU, D. et al. Highly imbalanced fault diagnosis of gas turbines via clustering-based downsampling and deep siamese self-attention network. *Advanced Engineering Informatics*, v. 54, p. 101725, 2022. ISSN 1474-0346. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1474034622001835>. Citado na página 39.

LIU, H.; BI, T.; YANG, Q. The evaluation of phasor measurement units and their dynamic behavior analysis. *IEEE Transactions on Instrumentation and Measurement*, v. 62, n. 6, p. 1479–1485, 2013. Citado na página 20.

MCKINNEY, W. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. [S.l.]: "O'Reilly Media, Inc.", 2012. 4–5 p. Citado na página 49.

MELNIK, S. et al. Dremel: Interactive analysis of web-scale datasets. In: *Proc. of the 36th Int'l Conf on Very Large Data Bases*. [s.n.], 2010. p. 330–339. Disponível em: <http://www.vldb2010.org/accept.htm>. Citado na página 49.

MICHAELIS, D. B. da L. P. [S.l.], 2023. Disponível em: <https://michaelis.uol.com.br/moderno-portugues/busca/portugues-brasileiro/intelig%C3%AAncia/>. Acesso em: 16 jul. 2023. Citado na página 21.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. *Introduction to linear regression analysis*. [S.l.]: John Wiley & Sons, 2021. Citado na página 29.

NUMFOCUS, I. *Pandas Documentation*. [S.l.], 2023. Disponível em: <https://pandas.pydata.org/docs/index.html>. Acesso em: 21 set. 2023. Citado na página 48.

OLIPHANT, T. E. Python for scientific computing. *Computing in science & engineering*, IEEE, v. 9, n. 3, p. 10–20, 2007. Citado na página 42.

ONS. *Boletim de Interrupção de Suprimento de Energia no Sistema Interligado Nacional - N° do BISE: ONS 019/2021*. [S.l.], 2023. Disponível em: <https://www.ons.org.br/AcervoDigitalDocumentosEPublicacoes/BISE%20ONS%20019-21%20080421-18h34%20-%20Ocorr%C3%AAncia%20no%20Amap%C3%A1.pdf>. Acesso em: 21 set. 2023. Citado 4 vezes nas páginas 37, 38, 39 e 70.

ONS. *Boletim de Interrupção de Suprimento de Energia no Sistema Interligado Nacional - N° do BISE: ONS 031/2021*. [S.l.], 2023. Disponível em: <https://www.ons.org.br/AcervoDigitalDocumentosEPublicacoes/BISE%20ONS%20031-21%20280521-11h26%20Ocorr%C3%AAncia%20no%20SIN.pdf>. Acesso em: 21 set. 2023. Citado na página 37.

ONS. *O que é o SIN*. [S.l.], 2023. Disponível em: <https://www.ons.org.br/paginas/sobre-o-sin/o-que-e-o-sin>. Acesso em: 27 ago. 2023. Citado 2 vezes nas páginas 7 e 14.

ONS. *Procedimentos de Rede*. [S.l.], 2023. Disponível em: <https://www.ons.org.br/paginas/sobre-o-ons/procedimentos-de-rede/vigentes>. Acesso em: 10 jun. 2023. Citado 2 vezes nas páginas 15 e 36.

PANDAS. *Documentação Pandas - DataFrame.to_parquet*. [S.l.], 2023. Disponível em: <https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.to_parquet.html>. Acesso em: 21 set. 2023. Citado na página 39.

PARQUET, A. *Apache Parquet Documentation*. [S.l.], 2023. Disponível em: <https://parquet.apache.org/docs/overview/motivation/>. Acesso em: 21 set. 2023. Citado na página 49.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, JMLR. org, v. 12, p. 2825–2830, 2011. Citado 2 vezes nas páginas 26 e 28.

PHADKE, A. Synchronized phasor measurements in power systems. *IEEE Computer Applications in Power*, v. 6, n. 2, p. 10–15, 1993. Citado 2 vezes nas páginas 7 e 16.

PHADKE, A. Synchronized phasor measurements-a historical overview. In: *IEEE/PES Transmission and Distribution Conference and Exhibition*. [S.l.: s.n.], 2002. v. 1, p. 476–479 vol.1. Citado na página 15.

PHADKE, A.; BI, T. Phasor measurement units, wams, and their applications in protection and control of power systems. *J. Mod. Power Syst. Clean Energy*, v. 6, p. 619–629, 2018. Citado na página 15.

PROVOST, F.; FAWCETT, T. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. [S.l.]: "O'Reilly Media, Inc.", 2013. Citado na página 28.

PYOD. *Documentação PyOD*. [S.l.], 2023. Disponível em: <https://pyod.readthedocs.io/en/latest/>. Acesso em: 21 set. 2023. Citado 2 vezes nas páginas 41 e 64.

RIBEIRO, R. de A. *Propriedades estatísticas dos sinais de PMU - Phrasor Measurement Unit*. Dissertação (Dissertação de Mestrado) — Universidade Federal de Itajubá, 2022. Citado na página 15.

ROSENBERG, A.; HIRSCHBERG, J. V-measure: A conditional entropy-based external cluster evaluation measure. In: *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*. [S.l.: s.n.], 2007. p. 410–420. Citado na página 31.

ROUSSEEUW, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, p. 53–65, 1987. ISSN 0377-0427. Disponível em: <https://www.sciencedirect.com/science/article/pii/0377042787901257>. Citado na página 32.

RUSSELL, S. J. *Artificial intelligence a modern approach*. [S.l.]: Pearson Education, Inc., 2010. Citado 2 vezes nas páginas 21 e 22.

SAGI, O.; ROKACH, L. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Wiley Online Library, v. 8, n. 4, p. e1249, 2018. Citado na página 80.

SCHLEGL, T. et al. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: NIETHAMMER, M. et al. (Ed.). *Information Processing in Medical Imaging*. Cham: Springer International Publishing, 2017. p. 146–157. ISBN 978-3-319-59050-9. Citado na página 80.

SCIKIT-LEARN. *Documentação Scikit-Learn*. [S.l.], 2023. Disponível em: <https://scikit-learn.org/0.21/documentation.html>. Acesso em: 21 set. 2023. Citado 2 vezes nas páginas 40 e 51.

SCIKIT-LEARN. *Documentação Scikit-Learn - EllipticEnvelope*. [S.l.], 2024. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.covariance.EllipticEnvelope.html>. Acesso em: 10 mar. 2024. Citado na página 51.

SCIKIT-LEARN. *Documentação Scikit-Learn - IsolationForest*. [S.l.], 2024. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.IsolationForest.html>. Acesso em: 10 mar. 2024. Citado na página 51.

SCIKIT-LEARN. *Documentação Scikit-Learn - LocalOutLierFactor*. [S.l.], 2024. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html>. Acesso em: 10 mar. 2024. Citado na página 51.

SCIKIT-LEARN. *Documentação Scikit-Learn - OneClassSVM*. [S.l.], 2024. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html>. Acesso em: 10 mar. 2024. Citado na página 51.

SCIKIT-LEARN. *Documentação Scikit-Learn - SGDOneClassSVM*. [S.l.], 2024. Disponível em: <https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDOneClassSVM.html>. Acesso em: 10 mar. 2024. Citado na página 51.

SOKOLOVA, M.; LAPALME, G. A systematic analysis of performance measures for classification tasks. *Information Processing Management*, v. 45, n. 4, p. 427–437, 2009. ISSN 0306-4573. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0306457309000259>. Citado na página 28.

STOCCO, A.; TONELLA, P. Towards anomaly detectors that learn continuously. In: *2020 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*. [S.l.: s.n.], 2020. p. 201–208. Citado 3 vezes nas páginas 43, 44 e 45.

THOMAS, A.; KOSHY, S.; R., S. Machine learning based detection and classification of power system events. In: *2020 International Conference on Power, Instrumentation, Control and Computing (PICC)*. [S.l.: s.n.], 2020. p. 1–6. Citado na página 33.

TURING, A. M. Mind. *Mind*, v. 59, n. 236, p. 433–460, 1950. Citado na página 21.

VANFRETTI, L.; BENGTSSON, S.; GJERDE, J. O. Preprocessing synchronized phasor measurement data for spectral analysis of electromechanical oscillations in the nordic grid. *International Transactions on Electrical Energy Systems*, Wiley Online Library, v. 25, n. 2, p. 348–358, 2015. Citado 2 vezes nas páginas 39 e 40.

ZHAO, Y.; NASRULLAH, Z.; LI, Z. Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, v. 20, n. 96, p. 1–7, 2019. Disponível em: <http://jmlr.org/papers/v20/19-011.html>. Citado na página 64.

ZHOU, M. et al. A preprocessing method for effective pmu placement studies. In: IEEE. *2008 Third International Conference on Electric Utility Deregulation and Restructuring and Power Technologies*. [S.l.], 2008. p. 2862–2867. Citado na página 40.