UNIVERSIDADE ABERTA

Modelação estatística aplicada à avaliação da condição de turbinas eólicas com base em dados SCADA.

José Jonas Nhantingo

Mestrado em Estatística, Matemática e Computação, Especialização em Estatística Computacional

2026

UNIVERSIDADE ABERTA

Modelação estatística aplicada à avaliação da condição de turbinas eólicas com base em dados SCADA.

José Jonas Nhantingo

Mestrado em Estatística, Matemática e Computação, Especialização em Estatística Computacional

Dissertação orientada pela Professora Doutora Alda Cristina Jesus Valentim Nunes de Carvalho, Unviversidade Aberta e Co-orientada por Professor Doutor Tiago Alexandre Narciso da Silva, Universidade Nova de Lisboa

2026

Dedicatória

Dedico este trabalho em memória ao meus Jonas Jossefa Nhantingo e Maria Mateus Mahosse, pelos ensinamentos que mesmo com deseparecimento físico deles continuam me guiar.

Dedico à minha esposa Elsa Mariza Covela, aos meus Filhos Keiser Dúmissan José Nhantingo, Cailana José Nhantingo e Eunísia José Nhantingo por me apoiarem e acreditar em mim. Ao meu primo Amândio Simão Mahone, pelo apaio prestado em todo meu percurso estudantil.

Agradecimentos

A Deus pelo dom da vida.

À professora Doutora Alda Carvalho e ao professor Doutor Tiago Silva, pela simplicidade, paciência, compreensão e competência demonstrada na orientação desde a fase da escolha do tema até a efectivação desta dissertação.

Agradeço ao meu primo Rodriques Abel Nhatingo pelo apoio moral e financeiro prestado durante a minha formação.

A todos os docentes que me acompanharam durante a parte curricular do Mestrado em Estatística, Matemática e Computação por transmitirem suas experiências agradáveis no ramo científico.

Aos meus colegas de Turma de Mestrado especialmente Armindo Saia, Wilson Tiago, Fernando Majante e Paulo Mussa pelo apoio prestado durante o curso. De forma particular agradeço ao colega Paulo Mussa pelo suporte informático.

Agradeço o meu colega de trabalho Manuel Chichonque pelo apoio linguístico.

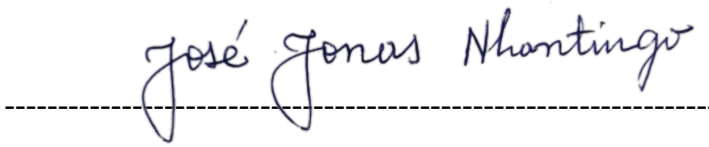
Os meus agradecimentos estendem-se para todos que directa ou indirectamente contribuíram para o sucesso deste trabalho.

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração da presente dissertação. Confirmo que em todo trabalho conducente à sua elaboração não recorri à prática de plágio ou qualquer outra forma de falsificação de resultados.

Mais declaro que tomei conhecimento integral do Regulamento Disciplinar da Universidade Aberta, publicado no Diario da República, 2.ª série, n.º 215, de 6 de novembro de 2013.

Universidade Aberta, 11 de fevereiro de 2026



José Jonas Nhantingo

Resumo

A monitorização do estado de turbinas eólicas a partir de dados SCADA, constitui um desafio significativo devido à elevada variabilidade operacional, à natureza não estacionária dos regimes do vento e à ausência de registos históricos de falhas devidamente rotulados. Neste contexto, este trabalho propõe uma abordagem probabilística integrada para a modelação do estado de condição de turbinas eólicas sob incerteza, combinando regressão quantílica e modelos bayesianos orientados à detecção de novidade. O termo "deteção de novidade" foi adoptado na formulação de Bishop (1994), uma vez que o objetivo central não é era identificação de falhas previamente conhecidas, mas sim a identificação de padrões de comportamento que se desviam do regime normal aprendido, podendo ou não estar associados a condições de degradação ou falha. Esta formulação permite uma abordagem mais robusta e flexível, especialmente adequada a ambientes industriais onde a definição explícita de classes de falha é limitada ou inexistente.

Neste trabalho assume-se que não há informação prévia fiável sobre a frequência relativa de falhas, devido à ausência de registos históricos completos e de rotulagem supervisionada nos dados SCADA. Numa primeira etapa, a regressão quantílica é aplicada à curva de potência com o objectivo de caracterizar o comportamento operacional esperado da turbina em diferentes regimes de vento, permitindo a definição de bandas estatísticas de funcionamento normal por meio de quantis inferior e superior. Esta etapa possibilita uma pré-classificação robusta de dados SCADA, mitigando os efeitos de heterocedasticidade, valores extremos e elevada variabilidade operacional.

Com base neste enquadramento, o problema de controlo de condição é formulado como um problema de classificação binária, no qual o estado de saúde da turbina é modelado como uma variável latente. São então aplicados e comparados três modelos estatísticos para estimar a verosimilhança dos dados sob hipótese de funcionamento normal: (i) um método dos *bins* baseado em intervalos de velocidade do vento e limites quantílicos, (ii) modelos baseados na distribuição normal multivariada e (iii) modelos de cópulas, incluindo as cópulas Gaussianas e *t-Student*, considerando versões estáticas e com actualização temporal.

Assumindo probabilidades *a priori* uniformes, a decisão de classificação baseia-se exclusivamente na improbabilidade conjunta das observações face ao regime operacional saudável, seguindo o princípio de detecção de novidade. O desempenho dos métodos é avaliado por meio de métricas estatísticas clássicas de classificação, permitindo uma comparação sistemática da sua capacidade de identificar estados normais e anómalos.

Os resultados mostraram que métodos simples baseados em limiares por *bins* são insuficientes para captar a complexidade das dependências multivariadas dos dados SCADA. Em contraste, os modelos probabilísticos multivariados apresentam melhorias significativas em métricas como acurácia, especificidade, F1- score e coeficiente de correlação de Matthews, evidenciando maior robustez em ambientes ruidosos e não estacionários. A incorporação da dependência temporal (*one-step-ahead*), revelou-se crucial para a detecção precoce de trajectórias anómalas e padrões evolutivos de degradação, destacando os modelos baseados na distribuição normal multivariada *one-step-ahead* como a abordagem mais eficaz para o controlo de condição e suporte à manutenção preditiva de turbinas eólicas.

Palavras-chave: Turbinas eólicas; Regressão quantílica; Controlo de condição; Dados SCADA; Inferência bayesiana; Detecção de novidade.

Abstract

Monitoring the condition state of wind turbines using SCADA data poses a significant challenge due to high operational variability, the non-stationary nature of regimes, and the frequent lack of properly labeled historical failure records. In this context, this work proposes an integrated probabilistic approach for modeling wind turbine condition states under uncertainty, combining quantile regression with Bayesian models oriented toward novelty detection.

In a first stage, quantile regression is applied to the power curve to characterize the expected operational behavior of the turbine under different wind regimes, enabling the definition of statistical bands of normal operation through conditional quantiles. This step allows a robust pre-classification of SCADA data, mitigating the effects of heteroscedasticity, extreme values, and high operational variability.

Based on this framework, the condition monitoring problem is formulated as binary probabilistic task, in which the turbine health state is modeled as a latent variable. Three statistical models are then applied and compared to estimate the data likelihood under the normal operating hypothesis: (i) a bin-based method relying on wind speed intervals and quantile limits, (ii) models based on the multivariate normal distribution, and (iii) copula-based models, including Gaussian and t-Student copulas, considering both static and one-step-ahead versions.

Assuming uniform prior probabilities, the classification decision relies exclusively on the joint improbability of observations with respect to the healthy operational regime, following the novelty detection principle. The performance of the methods is assessed using classical statistical classification metrics, enabling a systematic comparison of their ability to identify normal and abnormal states.

The results show that simple bin-based threshold methods are insufficient to capture the complexity of multivariate dependencies present in SCADA data. In contrast, multivariate probabilistic models achieve significant improvements in metrics such as accuracy, specificity, F1-score, and the Matthews correlation coefficient, demonstrating greater robustness in noisy and non-stationary environments. The incorporation of temporal dependence proves to be crucial for the early detection of anomalous trajectories and evolving degradation patterns, with temporally updated multivariate normal models emerging as the most effective approach for wind turbine condition monitoring and predictive maintenance support.

Keywords: Wind turbines; Quantile regression; Condition monitoring; SCADA data; Bayesian inference; Novelty detection

Índice

Índice de tabelas

Índice de figuras

Lista de abreviaturas

IEA- International Energy Agency

GWEC- Global Wind Energy Council

CDF-Função distribuição acumulada

CMS-Condition monitoring system

RQ- Regressão quantílica

VN- Verdadeiros negativos

VP- Verdadeiros positivos

FP- Falsos positivos

FN- Falsos negativos

SCADA- Supervisory Control And Data Acquisition

MCC- Matthews Correlation coeficiente

MW- Megawatt

kW- quilowatt

CAPÍTULO I: APRESENTAÇÃO DO TEMA E CONTEXTUALIZAÇÃO

1. Introdução

1.1. **Enquadramento e Contextualização do estudo**

A crescente preocupação com a sustentabilidade energética, a segurança do abastecimento energético e a mitigação das alterações climáticas tem impulsionado, nas últimas décadas, a adopção em larga escala de fontes de energia renovável. Entre estas, a energia eólica assume um papel de destaque, sendo actualmente uma das tecnologias mais maduras, competitivas e amplamente integradas nos sistemas eléctricos modernos, tanto em aplicações *onshore* como *offshore* (*International Energy Agency* [IEA], 2023).

Como resultado desse crescimento, a capacidade global instalada de energia eólica ultrapassou 1.136 GW em 2024, impulsionada por uma adição anual recorde de 117 GW, consolidando a energia eólica como um dos pilares de transição energética global (*Global Wind Energy Council* [GWEC], 2025). No entanto, a rápida expansão observada desde o início dos anos 2000 resultou num parque eólico global heterogéneo, composto por turbinas de diferentes gerações tecnológicas, idades e níveis de desempenho.

Ainda o mesmo relatório, refere que o crescimento de 2024 foi de cerca de 109 GW de novas instalações de energia eólica *onshore* e de 8 GW de energia eólica *offshore*, elevando a capacidade acumulada global de energia eólica para 1.136 GW, distribuída por todos continentes, com 55 países instalando energia eólica.

Em termos de capacidade instalada, a China destacou-se em 2024 como o país com maior potência eólica acumulada, seguida pelos Estados Unidos, Alemanha, Índia e Brasil. A Tabela 1 apresenta os cinco países com maior capacidade instalada de energia eólica, considerando apenas este indicador absoluto. No entanto, esta abordagem deve ser interpretada com cautela, uma vez que não considera as diferenças significativas de dimensão territorial, populacional e estrutura dos sistemas elétricos entre países. Assim, uma avaliação mais equilibrada do desenvolvimento da energia eólica requer a utilização de indicadores relativos, como a capacidade instalada per capita ou a percentagem de penetração na matriz elétrica, que permitem uma comparação mais representativa do grau de integração desta tecnologia em cada contexto nacional.

Tabela 1: País com maior potência instalada

Nome do país	Potência instalada em 2024 (MW)	Total (MW)
China	79824	520600
EUA	4058	154258
Alemanha	4022	72760
Índia	3420	48156
Brasil	3278	33727

Fonte: GWEC,2025. Adaptado

A Europa é actualmente a região com mais participação da energia eólica em sua matriz eléctrica, com a Dinamarca liderando o caminho com cerca de 50% da sua electricidade vindo da energia eólica em 2020 (Song et al., 2018).

De acordo com IEA (2024), a energia eólica é uma das tecnologias de energia renovável mais baratas e competitivas das actualmente disponíveis, com custos de produção e manutenção em queda devido ao aumento da eficiência dos equipamentos, ao aprimoramento das tecnologias eólicas. A previsão é que continuará a crescer globalmente nos próximos anos, com instalação de novos parques *onshore* e *offshore* em diversos países.

Neste contexto, uma parcela significativa das turbinas actualmente em operação aproxima-se ou já ultrapassou a sua vida útil de projecto, estimada entre 20 e 25 anos (TWI, 2023). A medida que esses equipamentos envelhecem, verifica-se um aumento progressivo das taxas de falha, da frequência de intervenções de manutenção e dos custos associados à operação, afectando directamente a disponibilidade e a rentabilidade dos parques eólicos (Carroll et al., 2016).

As turbinas eólicas são sistemas electromecânicos complexos, constituídos por múltiplos subsistemas críticos, como o sistema de transmissão, o gerador, os rolamentos, as pás e os sistemas de controlo, que operam sob condições ambientais severas e regimes de carga altamente variáveis. A exposição a ventos turbulentos, ciclos de carga mecânica, variações térmicas e ambientais agressivas acelera os mecanismos de degradação e fadiga estrutural, tornando inevitável a ocorrência de falhas ao longo do seu ciclo de vida (Tavner et al.,2007).

Face a esse cenário, estratégias como a extensão de vida útil e *repowering* têm sido consideradas como alternativas para prolongar a exploração dos activos existentes, melhorar o desempenho energético e

reduzir o custo de energia. No entanto, a viabilidade técnica e económica dessas estratégias depende fortemente da capacidade de avaliar, de forma contínua e confiável, o estado de condição real das turbinas eólicas (*Renewable Energy World*, 2023).

Nesse sentido, o controlo de condição (*condition monitoring*) e a detecção precoce de falhas assumem um papel central na operação moderna de parques eólicos. A identificação antecipada de comportamentos anómalos permite a transição de estratégias de manutenção correctiva para abordagens preventivas, reduzindo paragens não planeadas, custos de reparação e riscos de falhas catastróficas (Hameed et al., 2009).

Segundo Tautz-Weinert e Watson (2017), os sistemas *Supervisory Control And Data Acquisition* (SCADA) constituem uma das principais fontes de informação para esse fim, ao disponibilizarem grandes volumes de dados operacionais recolhidos continuamente durante a operação das turbinas. Diversos estudos têm demonstrado o potencial de análise de dados SCADA para a monitorização do estado condição de turbinas eólicas, permitindo a detecção de desvios de comportamento associados a falhas incipientes e à degradação progressiva dos componentes (Schlechitingen & Santos, 2011; Tautz-Weinert & Watson, 2017).

Dentro desse contexto, as abordagens bayesianas têm-se destacado como soluções particularmente promissoras para o controlo de condição baseado em dados SCADA, uma vez que permitem a modelagem explícita de incertezas inerentes aos dados operacionais, bem como a incorporação de conhecimento prévio e experiências anteriores no processo de inferência (Gelman et al., 2014).

Neste trabalho, são adoptadas três abordagens distintas para a detecção de anomalias nos dados SCADA: (i) método dos *bins*, (ii) método baseado em distribuição normal multivariada e (iii) método de cópula. A escolha desses três métodos insere-se na busca por técnicas, genéricas e robustas, de avaliação ou detecção de anomalias para aplicação no controlo de condição, uma vez que cada abordagem apresenta características próprias, vantagens específicas e limitações inerentes. Ao utilizar esses métodos de forma complementar, torna-se possível obter uma avaliação mais consistente do estado de saúde das turbinas eólicas, contribuindo para o desenvolvimento de estratégias eficazes de controlo e apoio à decisão na operação e manutenção de parques eólicos.

1.2. **Motivação**

Tradicionalmente, o controlo de condição de turbinas eólicas tem recorrido a abordagens baseadas em limites fixos de alarme, inspecções periódicas e sistemas dedicados de monitorização de vibrações, como o *Condition Monitoring System* (CMS). Embora essas técnicas sejam eficazes na detecção de falhas já desenvolvidas, apresentam limitações significativas quando aplicadas à identificação de falhas incipientes ou de comportamentos anómalos subtis, os quais são frequentemente ocultados pela elevada variabilidade operacional e ambiental, em particular, pelas condições de vento e pelos regimes operacionais não estacionários das turbinas eólicas (Harmeed et al., 2009).

Além disso, sistemas baseados exclusivamente em vibração implicam custos adicionais de instalação e manutenção, tornando-se menos atractivos para parques eólicos envelhecidos ou activos fora de garantia. Nesse contexto, cresce o interesse por abordagens alternativas que explorem dados já disponíveis nos parques eólicos, sem necessidade de instrumentação adicional (Tautz-Weinert & Watson, 2017).

Neste cenário, segundo Schlechtingen e Santos (2011), métodos baseados na análise dados SCADA, aliados a modelos estatísticos e probabilísticos, surgem como alternativas particularmente promissoras para o controlo de condição de turbinas eólicas. Os dados SCADA fornecem medições contínuas de múltiplas variáveis operacionais, como velocidade de vento, potência gerada, rotações, medidas de vibração, temperaturas e pressões, permitindo caracterizar o comportamento global da turbina ao longo do tempo e identificar desvios em relação ao seu funcionamento normal, expectável.

Em particular, os métodos bayesianos oferecem vantagens significativas às abordagens determinísticas tradicionais, uma vez que permitem: (i) a modelação explícita da incerteza inerente às variáveis nos dados SCADA; (ii) a representação das dependências estatísticas em múltiplas variáveis operacionais; e (iii) a actualização contínua do conhecimento do sistema à medida que novos dados se tornam disponíveis (Gelman et al., 2014). Estas características tornam os métodos bayesianos especialmente adequados para ambientes complexos e ruidosos, como os sistemas eólicos.

A aplicação de técnicas estatísticas avançadas, tais como a regressão quantílica, distribuições normais multivariadas e cópulas, possibilitam uma análise mais abrangente do estado do funcionamento das turbinas. Em particular, as cópulas permitem modelar de forma flexível a dependência entre as variáveis, independentemente das distribuições marginais, o que se revela especialmente vantajoso na

análise de dados SCADA, frequentemente não gaussianos e assimétricos (Nelsen, 2006; Joe, 2014). Estas abordagens contribuem para uma detecção de anomalias mais robusta, reduzindo falsos alarmes e aumentado a sensibilidade à degradação progressiva dos componentes.

Assim, este estudo justifica-se pela necessidade crescente de desenvolver metodologias mais eficazes, robustas e economicamente viáveis para o controlo de condição de turbinas eólicas. Ao explorar o potencial dos métodos probabilísticos aplicados à análise de dados SCADA, o presente trabalho procura contribuir para o avanço do estado de arte em monitorização de condição, oferecendo uma abordagem comparativa e integrada que apoie a tomada de decisão na operação e manutenção de parques eólicos.

1.3. **Formulação do problema**

Apesar da ampla disponibilidade de dados SCADA, a identificação precisa e confiável do estado de funcionamento das turbinas eólicas, continua a ser um desafio. A elevada variabilidade operacional, combinada com ruído de medição ou variabilidade aleatória, dados incompletos, dificulta a distinção entre comportamento normal e anómalo (Kusiak & Li, 2011).

Além disso, as variáveis operacionais apresentam frequentemente correlações complexas, lineares e não lineares, e distribuições marginais não gaussianas, o que limita a eficácia de métodos tradicionais baseados em análises univariadas.

Deste modo, o problema central desta investigação consiste em determinar como métodos probabilísticos podem ser aplicados de forma eficaz para a modelação do comportamento das turbinas eólicas e, na identificação dos estados anormais e antecipar falhas, utilizando exclusivamente dados SCADA.

Com a resolução deste problema pretende-se contribuir para metodologias de controlo de condição mais confiáveis, adaptáveis e economicamente viáveis, apoiando a tomada de decisão em parques eólicos.

1.4. **Objectivos**

Objectivo geral

- Desenvolver um modelo robusto para avaliação e controlo de condição de turbinas eólicas, capaz de identificar desvio ao comportamento normal da curva de potência.

Objectivos Específicos

- Identificar o estado normal/anormal do funcionamento das turbinas eólicas por meio da Regressão quantílica aplicada à curva de potência;
- Aplicar as três abordagens seleccionadas (método bin, método baseado em distribuição normal multivariada e método de cópula) para modelar a dependência entre as variáveis e controlar a condição das turbinas;
- Comparar os diferentes métodos em termos da sua capacidade de detectar anomalias e prever falhas em turbinas eólicas.

1.5. **Metodologia geral do estudo**

A metodologia adotada neste trabalho baseia-se na análise quantitativa de dados históricos provenientes de sistemas SCADA. Uma vez que os dados disponibilizados não apresentam rotulagem prévia quanto ao estado de funcionamento, tornou-se necessário estabelecer um critério inicial que permita distinguir entre comportamento normal e anómalo da turbina eólica. Neste contexto, recorreu-se à regressão quantílica aplicada à curva de potência, por se tratar de uma abordagem capaz de concretizar o comportamento esperado da turbina ao longo de diferentes regimes operacionais, considerando a variabilidade natural dos dados SCADA. Ao estimar quantis condicionais da potência em função da velocidade do vento, a regressão quantílica possibilita a definição de limites estatísticos robustos, menos sensíveis a valores discrepantes e à heterocedasticidade, que servem de referência para a identificação preliminar dos estados normal e anómalo.

A partir desta etapa de pré-classificação, são então implementados e avaliados três métodos probabilísticos para o controlo de condição de turbinas eólicas, visando a modelação de incerteza e a detecção precoce de falhas incipientes.

O critério de classificação adaptado neste trabalho encontra suporte directo na abordagem proposta por Song et al. (2018), que enquadram a monitorização do estado de saúde de turbinas eólicas como um problema de inferência probabilística baseado em dados SCADA. Tal como no presente estudo, os autores consideram uma variável latente representando o estado de condição da turbina e recorrem ao

teorema de Bayes para estimar probabilidades *a posteriori* associadas cada estado de funcionamento, a partir da verosimilhança dos dados observados.

Ainda os mesmos autores, enfatizam a vantagem dos métodos bayesianos na quantificação explícita da incerteza e na transição de classificação binária rígida para uma avaliação probabilística contínua do estado de saúde. Embora, neste trabalho, a decisão final seja operacionalizada através de limiares probabilísticos ou quantílicos, a regra de decisão permanece matematicamente equivalente à maximização da probabilidade *a posteriori* sob hipótese a priori uniformes ou priori não informativa.

Deste modo, a metodologia proposta pode ser interpretada como um caso particular de um classificador bayesiano orientado à detecção de novidade, estendido à comparação sistemática de diferentes modelos probabilísticos para estimação de verosimilhança. Esta integração reforça a coerência teoria do enquadramento metodológico e evidencia a relevância do paradigma bayesiano como base comum para o controlo de condição turbinas eólicas a partir de dados SCADA.

O desempenho dos métodos é avaliado por meio de métricas estatísticas e de classificação, que foram calculadas a partir da matriz de confusão, permitindo uma comparação sistemática das abordagens consideradas. A análise resultados visa identificar as vantagens e limitações de cada método, bem como a sua aplicabilidade prática em sistemas reais de monitorização.

1.6. **Estrutura da dissertação.**

A presente dissertação encontra-se organizada em seis capítulos. No capítulo I apresenta-se introdução, contextualizando o tema, definindo o problema da pesquisa e os objectivos do estudo. O capítulo II aborda a fundamentação teórica e revisão da literatura relevante. O capítulo III, descreve detalhadamente a metodologia adoptada. No capítulo IV apresentam-se os resultados obtidos, enquanto o capítulo V discute e analisa esses resultados de forma crítica. E no capítulo VI são sintetizadas as principais conclusões do estudo e perspectivas para trabalhos futuros.

CAPÍTULO II: REVISÃO DA LITERATURA.

2. Revisão dos estudos e teorias relacionados ao tema

2.1. **Energia eólica e funcionamento das Turbinas eólicas**

As turbinas eólicas são dispositivos electromecânicos projectados para converter a energia cinética do vento em energia mecânica rotacional, e posteriormente em energia eléctrica por meio de um gerador acoplado (Manwell et al., 2010). Em termos estruturais, uma turbina é composta por pás, responsáveis por captação de vento, o rotor que transmite o movimento, a gôndola (*nacelle*) que abriga os sistemas de conversão e controlo, e a torre de sustentação. As turbinas podem ser classificadas de acordo com o eixo de rotação que pode ser horizontal ou vertical, a potência instalada (pequena, média ou grande escala) e aplicação (*onshore* ou *offshore*), sendo as turbinas de eixo horizontal as mais utilizadas pela elevada eficiência e escalabilidade. Na Figura 2.1, é possível visualizar o sistema de uma turbina eólica e seus principais componentes.

Figura 2.1: Esquema de uma turbina eólica.

Fonte: https://www.mundodaeletrica.com.br/energia-eolica-como-funciona/

De acordo com Letcher (2023), o desenvolvimento de energia eólica em escala significativa teve suas origens nos Estados Unidos de América na década de 1970, como resposta à dependência dos combustíveis fósseis e para diversificar a matriz energética. Entre as diferentes fontes renováveis, as energias solares e eólicas se destacaram pelo crescimento expressivo, impulsionado tanto por factores conjunturais, como a crise de petróleo e a sustentabilidade, quanto por avanços tecnológicos que permitem a ampliação da sua competitividade.

Esse movimento histórico encontra continuidade no presente, em que a ênfase deixou de ser apenas viabilidade de geração em larga escala e passou a incluir a eficiência operacional e confiabilidade das turbinas modernas. Actualmente, o uso de sistemas de monitoramento baseados em dados SCADA é fundamental, pois permite acompanhar variáveis em tempo real, como velocidade de vento, velocidade de rotor, potência gerada, temperatura de componentes, entre outras variáveis, possibilitando a detecção precoce de falhas e anomalias.

O SCADA correlaciona múltiplos conjuntos de variáveis, para treinar modelos de estados operacionais normais e utilizar esses modelos para detectar comportamento anómalo e *outliers* (Rezamand et al., 2020 citado por Pinna, 2024, p.43). Portanto, técnicas de análise estatística aplicadas a esses dados possibilitam identificar padrões de falhas e implementar estratégias de manutenção preditiva, consolidando a energia eólica não apenas como alternativa sustentável, mas como um sector altamente tecnológico e orientado por dados (Tuatz-Weinert & Watson, 2017).

Assim, o que começou na década 1970 como uma resposta à crise energética, evoluiu para um campo que combina a produção limpa de energia e análise avançada de dados, reforçando a energia eólica como elemento central na transição energética global.

Curva de potência duma turbina eólica

O comportamento operacional de uma turbina eólica é tradicionalmente descrito pela sua curva de potência, que estabelece a relação entre a velocidade de vento e potência eléctrica gerada. Esta curva constitui uma ferramenta fundamental tanto para avaliação de desempenho como para controlo de condição e detecção de falhas (IEC, 2017).

Segundo Manwell et al. (2010) a curva de potência típica apresenta quatro regimes principais de operação:

(i) Região abaixo da velocidade de arranque (*cut-in*), na qual a energia do vento é insuficiente para iniciar a rotação do rotor e, consequentemente, não há produção de energia eléctrica;

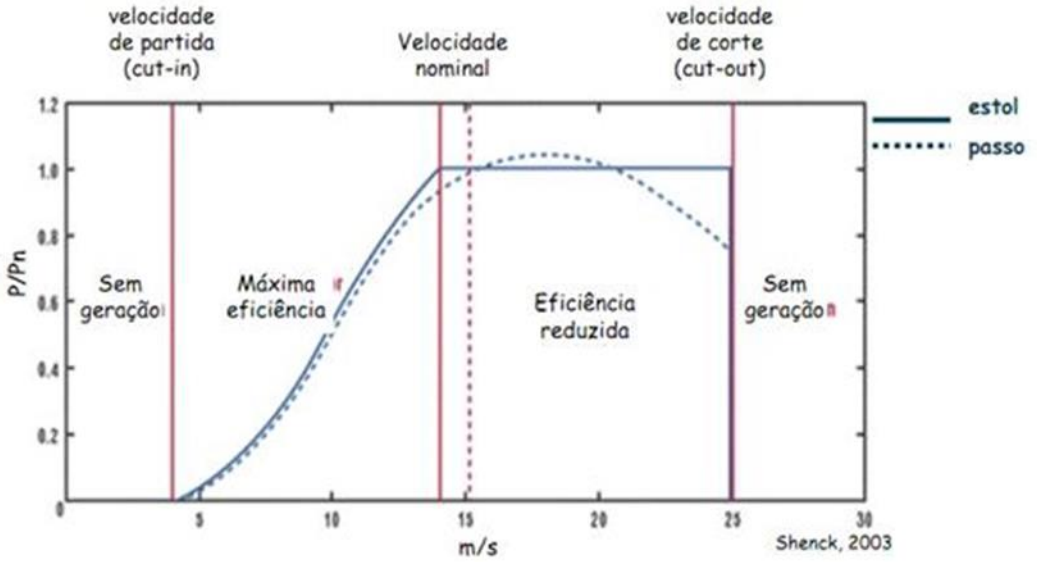
(ii) Região de carga parcial, onde a potência gerada aumenta aproximadamente de forma cúbica com a velocidade do vento, sendo fortemente influenciada pela eficiência aerodinâmica e pelo controlo de ângulo das pás;

(iii) Região de potência nominal, na qual a turbina opera à potência máxima especificada, sendo o excesso de energia regulado pelos sistemas de controlo de *Pitch* e/ou de *stall* das pás e;

(iv) Região de corte por segurança (*cut-out*), em que a turbina é desligada para evitar danos estruturais em condições de vento extremo.

A Figura 2.2, mostra curva de potência com os 4 (quatro) regimes de operação.

Figura 2.2: Curva de potência característica com 4 regimes.



Fonte: https://energiaeolicaufabc.blogspot.com/2011/12/

Desvios persistentes ou alterações significativas na forma de curva de potência podem indicar anomalias operacionais, perda de eficiência aerodinâmica, desgaste mecânico ou falhas incipientes em componentes críticos, como pás, rolamentos, caixa multiplicadora ou gerador (Tautz-Weinert & Watson, 2017). Por essa razão, a análise da curva de potência, especialmente baseada em dados

SCADA, tornou-se uma das abordagens mais utilizadas para o controlo de estado de saúde de turbinas eólicas.

Além disso, devido à elevada variabilidade natural do vento, à turbulência atmosférica e às incertezas inerentes às variáveis SCADA, a curva de potência observada apresenta dispersão significativa em torno da curva teórica ou de referência (Schlechtingen & Santos, 2011).

Segundo Bandi e Apt (2016), os valores instantâneos de potência e velocidade do vento exibem uma dispersão em torno da curva média de potência, reflectindo a variabilidade natural de vento e limitações dos modelos teóricos. Este facto, reforça a necessidade de uso de métodos estatísticos e baseados em dados reais, como análise de dados SCADA e modelos probabilísticos, para uma caracterização mais realista do desempenho das turbinas e para aplicações como controlo de condição e detecção de anomalias, fornecendo a base teórica para as abordagens desenvolvidas nos capítulos seguintes.

2.2. Sistemas SCADA aplicados a turbinas eólicas

Os sistemas SCADA são utilizados na operação moderna de parques eólicos, desempenhando um papel central na monitorização, controlo e supervisão do desempenho das turbinas eólicas. Estes sistemas permitem a aquisição contínua de dados operacionais e ambientais, bem como a execução de comandos de controlo, assegurando a operação segura e eficiente dos aerogeradores.

Tavner (2012) fundamenta o papel dos sistemas SCADA no contexto da fiabilidade, disponibilidade e manutenção de turbinas eólicas, destacando a sua importância na operação segura e na gestão do desempenho ao longo do ciclo de vida das turbinas. O autor enquadra os sistemas SCADA como uma ferramenta essencial de suporte à operação e manutenção, especialmente em parques de grande escala e em ambientes *offshore*, onde intervenções físicas são dispendiosas e logisticamente complexas.

Por sua vez, Tautz-Weinert e Watson (2017) complementam esta perspectiva ao analisar criticamente o potencial e limitações dos dados SCADA para o controlo de condição. Os autores demonstram que, embora os sistemas SCADA não tenham sido originalmente concebidos para detecção de falhas, a aquisição contínua de dados operacionais e ambientais torna-os uma fonte valiosa para a análise do comportamento das turbinas, desde que sejam aplicadas técnicas estatísticas adequadas.

Os sistemas SCADA registam, em intervalos regulares, tipicamente 10 minutos, um conjunto alargado de variáveis, incluindo velocidade e direcção do vento, potência activa e reactiva, velocidade do rotor, ângulo de *pitch*, posição de *yaw*, temperaturas de componentes críticos (gerador, rolamentos, óleo de caixa multiplicadora), pressões, bem como estados operacionais e alarmes. Para Schlechtingen e Santos (2011), este vasto conjunto de medições reflecte o comportamento dinâmico da turbina ao longo do tempo e sob diferentes condições ambientais.

Ainda na mesma abordagem estes autores afirmam que, apesar da elevada disponibilidade e do grande volume de dados, os sistemas SCADA apresentam limitações importantes. Entre estas destacam-se a presença de ruído nas medições, dados ausentes, valores atípicos, erros de calibração e agregação temporal das variáveis, que pode mascarar fenómenos de curta duração ou comportamentos transitórios relevantes. Além disso, a elevada variabilidade natural das condições de vento introduz dispersão significativa nos dados, dificultando a distinção entre os comportamentos normais e anormais.

Estas limitações tornam inadequadas abordagens baseadas em análises univariadas, reforçando a necessidade de métodos estatísticos e probabilísticos robustos capazes de lidar com incertezas, dependência entre as variáveis e distribuições não gaussianas. Neste contexto, abordagens multivariadas e bayesianas, têm demonstrado elevado potencial na exploração eficaz dos dados SCADA para fins de controlo de condição e detecção de falhas.

Assim, embora os sistemas SCADA não tenham sido originalmente concebidos como ferramentas dedicadas ao controlo de condição, sua ampla cobertura operacional, aliada a técnicas estatísticas avançadas, torna-os uma fonte valiosa e economicamente viável para a implementação de estratégias de manutenção preditiva e gestão de saúde de turbinas eólicas, especialmente em parques eólicos envelhecidos ou fora de garantia.

2.3. Controlo de condição e manutenção preditiva

O controlo de condição *(condition monitoring*) constitui um elemento central nas estratégias modernas de operação e manutenção de turbinas eólicas, uma vez que permite a monitorização contínua do estado de funcionamento dos principais subsistemas do aerogerador. Conforme salientado por Hameed et al. (2009), a detecção precoce de falhas por meio de técnicas de controlo de condição

contribui de forma significativa para a redução dos custos de operação e manutenção (O&M), bem como para o aumento da disponibilidade dos equipamentos. Esta abordagem está alinhada com a definição clássica de controlo de condição, que visa identificar anomalias e processos de degradação progressiva antes que estes evoluam para falhas críticas, evitando paragens não planeadas e danos severos aos componentes. No contexto específico das turbinas eólicas, o papel estratégico do controlo de condição é reforçado pela elevada complexidade dos aerogeradores, pela variabilidade das condições de operação impostas pelo vento e pelos elevados custos associados às intervenções de manutenção, especialmente em parques eólicos de grande escala ou *offshore*, conforme referenciado por Tavner (2012).

Esta abordagem está directamente associada à manutenção preditiva, cujo princípio fundamental consiste em planear intervenções de manutenção com base no estado real dos equipamentos, em vez de calendários fixos ou estratégias puramente correctivas. De acordo com Mobley (2002), a manutenção preditiva permite a redução significativa dos custos operacionais, a minimização do tempo de indisponibilidade e o prolongamento da vida útil dos componentes. No contexto da energia eólica, estes benefícios traduzem-se num aumento da fiabilidade das turbinas e numa melhoria da competitividade global da produção de energia, assegurar uma gestão mais eficiente dos recursos de operação e manutenção.

As técnicas de controlo de condição aplicadas a turbinas eólicas podem ser genericamente classificadas em três grandes categorias:

(i) Abordagens baseadas em modelos físicos, que utilizam descrições matemáticas detalhadas do comportamento dinâmico e estrutural dos componentes da turbina. Embora apresentem elevada precisão teórica, estas abordagens exigem conhecimento aprofundado do sistema, de valores de parâmetros difíceis de obter e elevado esforço computacional, o que limita a sua aplicação em ambientes reais complexos (García Márquez et al., 2012).

(ii) Abordagens baseadas em sinais, frequentemente associadas à análise de vibração, acústica ou correntes eléctricas, que permitem detecção de falhas em componentes específicos, como rolamentos e engrenagens. Apesar da elevada sensibilidade, estes métodos requerem sensores delicados, custos adicionais de instalação e manutenção, e nem sempre são economicamente viáveis para todos os parques eólicos (Hameed et al., 2009).

(iii) Abordagens baseadas em dados, exploram dados, exploram o histórico de dados operacionais para identificar padrões normais e inferir sobre o comportamento anómalo. Entre estas, as abordagens baseadas em dados SCADA têm ganho destaque, devido à sua ampla disponibilidade, cobertura de todo o sistema e ausência de necessidade de instrumentação adicional (Tautz-Weinert &Watson, 2017). A utilização de dados SCADA permite analisar simultaneamente múltiplas variáveis operacionais e ambientais, reflectindo o comportamento global da turbina ao longo do tempo. No entanto, a elevada variabilidade operacional, o ruído e complexa dependência entre variáveis tornam necessária a utilização de métodos estatísticos, capazes de distinguir entre variações normais de operação e comportamentos anómalos associados a falhas incipientes (García Márquez et al., 2012).

Neste contexto, métodos estatísticos probabilísticos, apresentam-se como ferramentas adequadas para o controlo de condição baseado em dados SCADA, fornecendo indicadores quantitativos do estado de funcionamento das turbinas eólicas e apoiando estratégias de manutenção preditiva. Estas abordagens constituem a base teórica para os métodos desenvolvidos e avaliados neste trabalho.

2.4. Abordagens estatísticas e probabilísticas relacionadas ao tema.

A utilização de dados SCADA para o controlo de condição e monitorização de estado de saúde de turbinas eólicas tem sido amplamente investigada, recorrendo a abordagens estatísticas, probabilísticas e de aprendizagem automatizadas (Tautz-Weinert & Watson, 2017). A elevada disponibilidade de dados operacionais possibilita o desenvolvimento de modelos orientados capazes de identificar padrões normais de funcionamento e detectar desvios associados a degradação ou falhas incipientes.

Abordagens iniciais basearam-se predominantemente na análise da curva de potência, explorando desvios estatísticos entre a potência observada e a potência esperada para um determinado regime de vento. Métodos baseados em *binning* de velocidade do vento e estatísticas condicionais demonstraram ser eficazes na identificação de perdas do desempenho, mantendo elevada interpretação física (Pandit & Infield, 2018). No entanto, estas abordagens apresentam limitações na representação de dependências multivariadas e na integração de incerteza de forma explícita.

Com o avanço da modelação estatística, modelos probabilísticos multivariados, como a distribuição normal multivariada, passaram a ser utilizados para descrever o comportamento conjunto de múltiplas variáveis SCADA sob condições normais de operação. Contudo, segundo Song et al. (2018), a suposição da normalidade conjunta pode ser restrita quando os dados apresentam assimetrias, não linearidades ou dependências de cuada.

Como alternativa à hipótese da normalidade multivariada, métodos baseados em cópulas têm sido introduzidos para capturar estruturas de dependências mais complexas entre variáveis SCADA. Fundamentadas no teorema de Sklar, as cópulas permitem modelar separadamente as distribuições marginais e a dependência, sendo particularmente adequadas para sistemas onde falhas afectam simultaneamente múltiplas variáveis (Nelsen, 2006). Estudos recentes indicam que cópulas Gaussianas e *t-Student* melhoram a sensibilidade na detecção de anomalias, especialmente em cenários com dependência de cauda (Embrechts et al., 2022).

Neste âmbito, Song et al. (2018) propõem uma abordagem explicitamente bayesiana e orientada a dados para a monitorização de estado de saúde de turbinas eólicas, utilizando dados SCADA. Os autores formulam o problema como uma inferência probabilística do estado de saúde, integrando conhecimento prévio e informação observacional através do teorema de Bayes. Em vez de uma classificação rígida, o método estima probabilidades a posteriori associadas a diferentes estados de condição, permitindo uma avaliação gradual da degradação e uma quantificação explícita da incerteza. Os resultados demonstram que a abordagem bayesiana apresenta maior robustez face a ruído e variações operacionais, além de oferecer melhor interpretação para apoiar à tomada de decisão em manutenção preditiva.

Paralelamente, técnicas de aprendizagem automática e *deep learning* têm sido aplicadas ao controlo de condição de turbinas eólicas (Verma et al., 2022). Apesar do elevado desempenho preditivo, estas abordagens apresentam limitações relacionadas com a necessidade de grandes conjuntos de dados rotulados e com a reduzida interpretabilidade dos modelos, aspectos frequentemente apontados como desvantagens face a métodos probabilísticos bayesianos (Aggarwal, 2013).

De forma geral, a literatura evidencia que, embora existam múltiplas abordagens eficazes para detecção de anomalias em turbinas eólicas, a comparação sistemática de métodos probabilísticos sob enquadramento bayesiano comum permanece limitada. Assim, o presente trabalho diferencia-se ao analisar, de forma comparativa, métodos baseados em *binning*, distribuição normal multivariada e

cópulas, permitindo uma avaliação consistente do seu desempenho no controlo de condição de turbinas eólicas a partir de dados SCADA.

CAPÍTULO III: METODOLOGIA: DESCRIÇÃO DA METODOLOGIA UTILIZADA

3. Metodologia

Este capítulo descreve a metodologia adoptada para o desenvolvimento, implementação e avaliação dos métodos de controlo de condição aplicados a turbinas eólicas com base em dados SCADA. Apresenta-se neste capítulo, tipo e abordagem de pesquisa, área do estudo e objecto de pesquisa, dados utilizados e procedimentos metodológicos.

3.1. Tipo e abordagem da pesquisa

A presente pesquisa caracteriza-se como quantitativa e descritiva, uma vez que se baseia na análise estatística de dados numéricos provenientes da operação real de turbinas eólicas, recolhidos por sistemas SCADA. De acordo com Gil (2008), a pesquisa quantitativa é apropriada quando o objectivo central consiste na mensuração, análise e interpretação de fenómenos observáveis, permitindo identificar padrões, relações e comportamentos a partir de dados empíricos. Neste contexto, a utilização de métodos estatísticos e probabilísticos possibilita uma avaliação objectiva do estado de condição das turbinas eólicas.

Sob ponto de vista dos objectivos, o estudo assume uma natureza descritiva, pois procura caracterizar o comportamento operacional das turbinas eólicas em diferentes regimes de funcionamento, bem como descrever a relação entre variáveis ambientais e operacionais, tais como velocidade do vento, potência gerada, velocidade do rotor e torque. Segundo Lakatos e Marconi (2017), a pesquisa descritiva tem como finalidade principal a descrição sistemática das características de determinada população ou fenómeno, sem interferência directa do investigador sobre o objecto do estudo, o que se adequa ao presente trabalho, baseado em dados históricos operacionais.

Adicionalmente, a investigação adopta uma abordagem comparativa, na medida em que avalia e compara o desempenho de diferentes métodos estatísticos e probabilísticos aplicados ao controlo de condição e à detecção de falhas em turbinas eólicas. Entre os métodos analisados incluem-se abordagens clássicas, como o método bin, técnicas mais avançadas, como modelos baseados na distribuição normal multivariada e em cópulas. Esta comparação permite identificar vantagens, limitações e níveis de robustez de cada método face às características dos dados SCADA, conforme recomendado em estudos metodológicos aplicados à engenharia (Gil, 2008).

Quanto à sua finalidade, a pesquisa enquadra-se como aplicada, uma vez que visa o desenvolvimento, a adaptação e avaliação de métodos como potencial de aplicação prática em sistemas reais de monitorização e manutenção preditiva de aerogeradores. Conforme salientado por Lakatos e Marconi (2017), a pesquisa aplicada distingue-se por procurar soluções para problemas concretos, contribuindo directamente para melhoria de processos de tomada de decisão em contextos industriais. Neste caso, os resultados obtidos podem apoiar estratégias de manutenção baseadas na condição, promovendo a redução de custos operacionais e aumento da disponibilidade das turbinas eólicas.

Por fim, destaca-se que a abordagem metodológica adoptada estabelece uma articulação entre conceitos de estatística clássica e inferência bayesiana, proporcionando um enquadramento consistente para análise de incertezas e para a tomada de decisão probabilística no âmbito de controlo de condição de turbinas eólicas.

3.2. Área de estudo e objecto de pesquisa

A área de estudo desta pesquisa insere-se no domínio da engenharia eólica, com enfoque específico no controlo de condição e detecção de falhas em turbinas eólicas a partir da análise de dados operacionais (SCADA). O estudo centra-se em turbinas eólicas de eixo horizontal, tecnologia amplamente utilizada em parques eólicos comerciais, cuja operação é fortemente influenciada por condições ambientais variáveis, particular a velocidade e direcção do vento.

O objecto da pesquisa consiste na identificação do estado de condição em turbinas eólicas controladas por sistemas SCADA, os quais permitem a recolha contínua e sistemática de um grande conjunto de variáveis operacionais e ambientais. Estes sistemas registam medições em intervalos regulares de tempo de 10 minutos, possibilitando a construção de bases de dados históricos extensas, fundamentais para estudos de controlo da condição e manutenção preditiva.

Neste trabalho, foram analisadas duas turbinas eólicas, designadas WTG01 e WTG02, seleccionadas por apresentarem um histórico operacional suficientemente longo e consistente, bem como variabilidade representativa das condições de vento e regimes de funcionamento. A escolha destas turbinas permitiu assegurar a aplicação e comparação dos diferentes métodos estatísticos considerados, garantindo relevância prática e validade empírica aos resultados obtidos.

A análise das turbinas WTG01 e WTG02 abrange diferentes estados operacionais, incluindo regimes de operação normal e períodos associados a comportamento anómalo, o que possibilita avaliar a capacidade dos métodos propostos na identificação de desvios de desempenho e potenciais falhas.

3.3. Dados utilizados

Os dados utilizados neste estudo consistem em dados reais de operação de turbinas eólicas, provenientes de sistemas SCADA, recolhidos durante o ano 2012, ao longo de um período contínuo de funcionamento das turbinas eólicas analisadas. Estes dados, são de natureza quantitativa e representam medições agregadas em intervalos temporais regulares, descrevendo o comportamento operacional das turbinas sob diferentes condições ambientais e regimes de funcionamento.

O conjunto de dados inclui variáveis ambientais e operacionais relevantes para o controlo de condição, tais como a velocidade do vento, a potência activa, a velocidade do rotor e o torque. As observações encontram-se organizadas sob a forma de uma matriz $T \times p$, que cada linha corresponde a um instante temporal e cada coluna representa uma variável controlada.

3.3.1. Pré-processamento dos dados

O pré-processamento dos dados constitui uma etapa fundamental para garantir a qualidade da análise estatística. As principais etapas seguidas neste trabalho incluem a remoção de observações inválidas ou incompletas e filtragem de regimes operacionais não relevantes (paragens e regimes de arranque).

3.4. Regressão quantílica como método de rotulagem dos dados

Uma característica relevante dos dados SCADA analisados, é a ausência de registos de falhas e rotulagem supervisionada que identifique directamente os estados de funcionamento normal e anómalo das turbinas eólicas. Esta limitação impossibilita a aplicação directa de métodos de classificação supervisionada convencionais e exige a adopção de estratégias alternativas para a definição do estado de condição. Neste contexto recorreu-se à regressão quantílica como método estatístico para caracterização

do comportamento operacional esperado da turbina eólica e para a rotulagem indirecta dos dados SCADA, servindo de base à posterior aplicação dos métodos apresentados nas secções seguintes.

Neste trabalho, foram estimados quantis inferiores e superiores (0.1 e 0.9), os quais definem uma banda de funcionamento considerada normal da turbina eólica para cada regime do vento. A escolha destes quantis inferior e superior, fundamenta-se na recomendação de Koenker & Hallock (2001), que salientam que quantis extremos podem conduzir a estimativas instáveis devido à escassez de observações nas caudas da distribuição. Assim, os quantis de 10% e 90% permitem capturar a variabilidade natural do processo com maior robustez estatística, ao mesmo tempo que preservam sensibilidade à detecção de desvios relevantes. Observações cujas variáveis operacionais, nomeadamente a potência, se situam de forma persistente fora desta banda são interpretadas como potenciais estados anormais ou indícios de degradação do desempenho da turbina.

Com base neste critério, as observações SCADA foram classificados em dois estados de condição:

i) **Estado norma**l **($Y = 1$):** observações que se encontram dentro da banda definida pelos quantis;
ii) **Estado anómalo ($Y = 0$):** observações que se encontram fora da banda definida pelos quantis.

Esta rotulagem indirecta permitiu transformar o conjunto de dados originais num conjunto de dados pseudo-supervisionado, possibilitando a aplicação dos métodos de classificação descritos nas secções seguintes.

A regressão quantílica fornece uma aproximação empírica da região de funcionamento normal, enquanto os métodos probabilísticos subsequentes modelam explicitamente as densidades condicionais $p(x|Y = 1)$ e $p(x|Y = 0)$, bem como as probabilidades a priori $p(Y = 1)$ e $p(Y = 0)$. A combinação destas quantidades permite inferir a probabilidade a posteriori do estado de condição da turbina, de acordo com o teorema de Bayes.

Esta estratégia metodológica apresenta várias vantagens: é robusta à presença de valores discrepantes, não requer pressupostos de normalidade, adapta-se à heterocedasticidade típica dos dados SCADA e reflecte de forma realista a ausência de informação de falhas em ambientes industriais. O conjunto de

dados assim construídos constitui, portanto, a base para a avaliação comparativa do desempenho dos métodos probabilísticos considerados neste estudo.

3.5. Métodos aplicados no controlo de condição

No contexto do controlo de condição de turbinas eólicas, a detecção de estados anómalos pode ser formulada como um problema de classificação probabilística binária. Define-se a variável latente do estado de condição:

$$y \in \{0,1\}, \qquad (3.1)$$

em que $Y = 1$ representa o estado normal (saudável) da turbina e $Y = 0$, corresponde a um estado anómalo ou potencialmente associado a falha.

Dada uma observação multivariada $x \in \mathbb{R}^d$, composta por d variáveis operacionais medidas pelo sistema SCADA.

De acordo com teorema de Bayes, a probabilidade a prosteriori do estado normal condicionada à observação x é dada por:

$$p(Y = 1|x) = \frac{p(x|Y=1)p(Y=1)}{p(x)}, \qquad (3.2)$$

Em que $p(x|Y = 1)$ representa a verosimilhança da observação sob hipótese de funcionamento narmal da turbina eolica, $p(Y = 1)$ corresponde a probabilidade a priori associada a esse estado, e $p(x)$ correspende ao termo de normalização, definido como probabilidade marginal da observação, dada por

$$p(x) = p(x|Y = 1)p(Y = 1) + p(x|Y = 0)p(Y = 0). \qquad (3.3)$$

Uma vez que o termo $p(x)$ é comum a todas as hipóteses e não depende do estado Y, a inferência pode ser formulada, a menos de uma constante de proporcionalidade, como:

$$p(Y = 1|x) \propto p(x|Y = 1)p(Y = 1). \qquad (3.4)$$

Neste trabalho assume-se que não há informação prévia fiável sobre a frequência relativa de falhas, devido à ausência de registos históricos completos e de rotulagem supervisionada nos dados SCADA. Assim, são consideradas probabilidades a priori constantes para os estados, isto é:

$$p(Y = 1) = p(Y = 0), \qquad (3.5)$$

o que conduz a uma regra de decisão baseada exclusivamente na verosimilhança condicional:

$$p(Y = 1|x) \propto p(x|Y = 1) \qquad (3.6)$$

Neste caso, o critério de classificação reduz-se à avaliação da plausibilidade da observação x face ao comportamento normal previamente, aprendido a partir dos dados. Observações com elevada verosimilhança são interpretadas como compatíveis com regime normal de operação, enquanto observações com baixa verosimilhança indicam desvios estatisticamente improváveis, sendo associadas a estados anormais.

Esta formulação é baseada em princípio de detecção de novidade ***(****novelty detection*) proposta por Bishop (1994), no qual apenas a classe normal é explicitamente modelada. Neste caso, a função de verosimilhança $p(x|Y = 1)$ é estimada utilizando exclusivamente dados considerados normais, e a classificação resulta da comparação dessa verosimilhança com um limiar probabilístico definido a partir da distribuição empírica da classe normal.

No âmbito desta pesquisa, a verosimilhança condicional $p(x|Y = 1)$, associada ao funcionamento normal da turbina eólica é aproximada por três modelos estatísticos distintos, cada um deles explorando diferentes níveis de complexidade na representação do comportamento operacional a partir de dados SCADA. Especificamente, são considerados: (i) o método dos *bins*, baseado na discretização da velocidade do vento e na definição de regiões empíricas de elevada densidade por quantis; (ii) o modelo baseado na distribuição normal multivariada, que assume uma estrutura gaussiana para dependência conjunta entre as variáveis operacionais; e (iii) os modelos de cópulas, nomeadamente a cópula Gaussiana e a cópula *t-Student*, que permitem uma modelação mais flexível da dependência, incluindo efeitos não lineares e de cauda.

Apesar das diferenças na modelação estatística e nos pressupostos subjacentes a cada abordagem, os três métodos partilham um princípio comum de decisão: a classificação do estado de condição da turbina baseia-se na improbabilidade das observações face ao regime operacional normal previamente

aprendido. Adicionalmente, para os métodos de distribuição normal multivariado e de cópulas é adoptada uma estratégia de variação temporal (*onde-step-ahead*), na qual cada observação é avaliada sequencialmente utilizando apenas a informação disponível até ao instante anterior, de modo a simular um cenário realista de monitorização em tempo real. Nas secções seguintes, cada método é descrito em detalhe, bem como os respectivos critérios de classificação.

3.5.1. Método dos *Bins*

No âmbito deste trabalho, o método bin foi utilizado como abordagem probabilística de referência para a identificação do estado de condição da turbina eólica a partir de dados SCADA, sendo integrado num enquadramento bayesiano de classificação. A metodologia baseia-se na hipótese de que, para cada faixa de velocidade do vento, existe um comportamento operacional esperado das restantes variáveis monitoradas, o qual pode ser caracterizado estatisticamente a partir de dados históricos.

Inicialmente, as observações de velocidade do vento são discretizadas em intervalos predefinidos (*bins*), definidos por limites consecutivos (v_k, v_{k+1}) . Cada observação i é atribuída ao respectivo *bin*, de acordo com regra:

$$bins(i) = k \quad se \;\; v_k \leq WindSpeed_i < v_{k+1} \qquad (3.7)$$

Este agrupamento permite que todas as análises subsequentes sejam realizadas sob condições aerodinâmicas aproximadamente homogéneas, reduzindo a influência directa da variabilidade do vento sobre as restantes variáveis operacionais.

i. **Cálculo dos quantis por intervalo**

Para cada *bin* k, são estimados limites estatísticos para as outras variáveis monitoradas: potência (P_i), Velocidade do rotor (R_i) ou torque (T_i). Particularmente, são utilizados os quantis 0.1 e 0.9, que definem uma banda de operação considerada normal para cada variável sob aquela condição de velocidade do vento. Estes limites são interpretados como aproximação empírica das distribuições condicionais das variáveis dado estado normal da turbina, isto é, $p(x|Y = 1)$.

Neste contexto, a classificação do estado de condição da turbina baseia-se na estimação das probabilidades a posteriori associadas aos estados normal ($Y = 1$) e anormal ($Y = 0$), dadas novas

observações SCADA. A regra de decisão é fundamentada no teorema de Bayes, conforme a equação em (3.4).

A densidade condicional associada ao estado normal é aproximada de forma não paramétrica através do critério por quantis. Assim para uma observação x_i pertencente ao bin k, assume-se que:

$$p(x_i|Y=1) \text{ é } elevada\ se \begin{cases} Q_{0.1}^{P}(k) \leq P_i \leq Q_{0.9}^{P}(k) \\ Q_{0.1}^{R}(k) \leq R_i \leq Q_{0.9}^{R}(k) \end{cases}, \qquad (3.8)$$

E significativamente reduzida fora dos limites. Consequentemente, a regra de decisão pode ser expressa como:

ii. **Regra de decisão (classificação)**

$$\hat{y}_i = \begin{cases} 1, & se\ p(Y=1|x_i) > p(Y=0|x_i) \\ 0, & caso\ contrário \end{cases} \qquad (3.9)$$

O que na prática, corresponde à verificação do comprimento simultâneo dos limites por quantis definido para o *bin* correspondente.

Desta forma, uma observação é classificada como normal quando apresenta elevada verosimilhança sob o modelo empírico do estado saudável, enquanto observações fora da região definidas pelos quantis são associadas a baixa probabilidade *a posteriori* de funcionamento normal, sendo classificados como anormais. Este procedimento permite interpretar o método bin como um caso particular de classificação bayesiana, no qual as distribuições condicionais são aproximadas por região empírica de elevada densidade.

3.5.2. Modelação pela distribuição normal multivariada

Nesta etapa, foi aplicada a distribuição normal multivariada para modelar a dependência conjunta entre as variáveis operacionais. A partir da média vectorial e da matriz de co-variância estimadas, foi possível calcular a densidade da probabilidade associada a cada observação. Valores com baixa probabilidade foram interpretados como potenciais indicadores de comportamento anómalo.

A caracterização probabilística de regimes operacionais saudáveis em sistemas físicos por meio modelos gaussianos multivariados constitui um fundamento estatístico clássico e detecção de novidade, pois permite modelar simultaneamente a dispersão individual das variáveis e sua dependência estrutural (Mardia et al., 1979). Em aplicações a turbinas eólicas, esta formulação revela-se particularmente adequada, dado o comportamento físico aproximadamente linear e correlacionado entre variáveis como potência, torque e velocidade do rotor quando condicionadas pela velocidade do vento em regimes de operação estáveis (Manwell et al., 2010).

Nesta abordagem, a distribuição normal multivariada é utilizada para modelar a dependência conjunta entre as variáveis operacionais controladas pelo sistema SCADA, permitindo calcular a densidade de probabilidade associada a cada observação. Valores associados a baixas probabilidades são interpretados como potenciais indicadores de comportamento anormal ou degradação do sistema.

Hipótese

Assume-se que os dados correspondentes ao estado normal da turbina seguem uma distribuição normal multivariada. Define-se o vector de variáveis observadas pelo sistema SCADA da turbina como:

$$X = \begin{bmatrix} Potência \\ Velocidade\ do\ vento \\ Velocidade\ do\ rotor \end{bmatrix}, \qquad X \sim N_3(\mu, \sum)$$

onde $\mu \in \mathbb{R}^d$ é vetor de médias das variáveis sob condição normal (sem falha) e $\sum \in \mathbb{R}^{3\times3}$ é matriz de co-variância que representa a variância e correlação das variáveis sob condição normal.

Ajuste do modelo

Seja n_1 o número de observações consideradas normais ($Y = 1$). O modelo é ajustado exclusivamente com esta classe, seguindo o princípio da detecção de novidade, no qual apenas o comportamento saudável do sistema é explicitamente modelado. Os parâmetros da distribuição são estimados por meio dos estimadores amostrais clássicos:

Vector de médias:

$$\hat{\mu} = \frac{1}{n_1} \sum_{i:y_i=1} X_i \qquad (3.10)$$

Matriz de co-variância:

$$\hat{\Sigma} = \frac{1}{n_1 - 1} \sum_{i:y_i=1} (X_i - \hat{\mu})(X_i - \hat{\mu})^T \qquad (3.11)$$

Esta abordagem assume que a distribuição estimada representa adequadamente a verosimilhança condicional dos dados, dado o estado saudável da turbina.

Cálculo da densidade de probabilidade conjunta

Com os parâmetros ajustados, cada observação x é avaliada pela função densidade da distribuição normal multivariada, dada por:

$$f(x) = \frac{1}{(2\pi)^{\frac{k}{2}} |\sum|^{\frac{1}{2}}} exp\left[-\frac{1}{2}(x-\mu)^T \sum^{-1} (x-\mu)\right] \qquad (3.12)$$

em que *k* corresponde ao número de variáveis consideradas. Em implementações práticas, este cálculo é realizado por rotinas computacionais eficientes de densidade multivariada, como as disponibilizadas na biblioteca "*mvtorm*" em R ou outras equivalentes (Genz & Bretz, 2009).

Segundo Aggarwal (2013), observações com baixa densidade multivariada não são necessariamente valores extremos em variáveis individuais, mas sim combinações improváveis sob a estrutura de dependência estimada para o estado normal.

Definição do limiar probabilístico

O critério de decisão baseia-se na definição um limiar L, obtido a partir do quantil inferior das densidades associadas às observações normais:

$$L = Q_{0.1}[f(x)|\hat{y} = 1] \qquad (3.13)$$

Este limiar corresponde ao 10.º percentil da densidade da classe normal, admitindo que uma fracção reduzida de observações normais pode apresentar baixa verosimilhança devido à variabilidade natural

do sistema ou a ruído de operação (Hodge & Austin, 2004). A escolha do quantil inferior representa um equilíbrio entre rejeitar variações normais muito improváveis e evitar classificar ruído operacional saudável como falha, conforme discutido em (Goldstein & Uchida, 2016).

Regra de classificação

A classificação baseia-se na probabilidade a posteriori do estado normal $Y = 1$, proporcional à verosimilhança $p(x|Y = 1)$ e à probabilidade a priori $\mathrm{p}(Y = 1)$. Admitindo priors constantes, a regra de decisão reduz-se à comparação direta da densidade com limiar definido:

Cada novo vector observado x é classificado segundo a regra:

$$\hat{y} = \begin{cases} 0, se\ f(x) < L\ (\text{anómalo/provável falha}) \\ 1, se\ f(x) > L\ (\text{normal/sem falha}) \end{cases} \quad (3.14)$$

Este tipo de decisão é matematicamente análogo a um detector de novidade baseado na improbabilidade conjunta sob a hipótese gaussiana (Bishop, 1994).

Interpretação física

Durante a operação normal, as variáveis mantêm coerência aerodinâmica e electromecânica, potência ou torque crescem e rotor acelera de forma correlacionada conforme o vento aumenta (Manwell et al., 2010). Falhas mecânicas ou eléctricas quebram essa coerência, deslocando o vector para regiões de baixa probabilidade conjunta, mesmo quando nenhuma variável isolada é extrema (Hodge & Austin, 2004).

3.5.3. Modelação por cópulas

Como alternativa mais flexível, foram utilizados modelos de cópulas, em particular a cópula Gaussiana e a cópula *t-Student*, permitindo modelar a dependência entre as variáveis independentemente das suas distribuições marginais. As marginais foram estimadas separadamente, e os parâmetros das cópulas foram ajustados a partir dos dados observados.

A modelagem de dependência em dados SCADA por cópulas é um pilar metodológico na análise multivariada moderna, pois permite capturar relações não-lineares e dependência de cauda sem exigir que a distribuição conjunta pertença a uma família paramétrica (Nelsen, 2006). Em particular, o teorema de Sklar garante a factorização de qualquer distribuição conjunta contínua nas suas marginais e numa cópula que codifica exclusivamente a dependência entre as variáveis. Esta propriedade torna as cópulas especialmente adequadas para problemas de detecção de novidade e identificação de anomalias em sistemas industriais complexos, incluindo turbinas eólicas controladas por dados SCADA.

Fundamentação

Considere o vetor das observações contínuas da turbina eólica na *i*-ésima amostra:

$$x_i = (p_i, v_i, r_i) \qquad (3.15)$$

em que p_i, v_i, r_i representam, respectivamente, potência ativa, velocidade do vento e velocidade do rotor. Pelo Teorema de Sklar, a função de distribuição conjunta pode ser expressa como:

$$F(p, v, r) = C(F_p(p), F_v(v), F_r(r), \theta) \qquad (3.16)$$

onde $F_p(.)$, $F_v(.)$ e $F_r(.)$ denotam as funções de distribuição acumulada (CDF) marginal das variáveis potência, velocidade do vento e velocidade do rotor, respectivamente, $C(\cdot)$ é a **função cópula**, que descreve a dependência entre as variáveis e θ representa o conjunto de parâmetros da cópula.

Ajustamento das distribuições marginais

O ajustamento das distribuições das marginais é realizado exclusivamente com observações associadas ao estado normal do funcionamento da turbina ($Y = 1$), seguindo o princípio de detecção de novidade. Considere-se o conjunto de observações normais $\{x_i\}_{i=1}^{n}$, com $x_i = (p_i, v_i, r_i)$.

Segundo Song et al. (2018), para cada variável ajusta-se uma função de distribuição marginal, assumindo uma distribuição normal, dada por:

$p \sim N(\mu_1, \sigma_1^2)$ (Potência activa)

$v \sim N(\mu_2, \sigma_2^2)$ (velocidade do vento)

$r \sim N(\mu_3, \sigma_3^2)$ (velocidade do rotor)

Cada observação $xi = (pi, vi, ri)$ é então transformada em pseudo-observações uniformes no intervalo $(0,1)$ por meio das respectivas funções de distribuição acumulada:

$$u_i = \begin{bmatrix} F_P(p_i) \\ F_V(v_i) \\ F_R(r_i) \end{bmatrix} = \begin{bmatrix} \Phi\left(\frac{p_i - \hat{\mu}_p}{\hat{\sigma}_p}\right) \\ \Phi\left(\frac{v_i - \hat{\mu}_v}{\hat{\sigma}_v}\right) \\ \Phi\left(\frac{r_i - \hat{\mu}_r}{\hat{\sigma}_r}\right) \end{bmatrix} \in (0,1)^3 \qquad (3.17)$$

Este mapeamento projecta os dados para o espaço uniforme, no qual a dependência entre as variáveis passa a ser descrita exclusivamente pela cópula.

Ajustamento das cópulas

A densidade conjunta da observação original $xi = (pi, vi, ri)$ sob o modelo de cópulas é dada por:

$$F(x_i) = C(u_i, \theta).f_P(p_i).f_V(v_i).f_R(r_i) \qquad (3.18)$$

em que $c(.)$ representa a densidade da cópula, obtida pela derivada mista de terceira ordem:

$$c(u_P, u_V, u_R, \theta) = \frac{\partial^3}{\partial u_P \partial u_V \partial u_R} C(u_P, u_V, u_R, \theta) \qquad (3.19)$$

Esta densidade codifica exclusivamente a dependência estatística entre as variáveis.

Famílias de cópulas consideradas

- **Cópula gaussiana,** adequada para modelar dependência linear simétrica:

$$c_G(u, \sum) = \frac{1}{|\sum|^{\frac{1}{2}}} exp\left[-\frac{1}{2} z^T (\sum^{-1} - I) z\right] \qquad (3.20)$$

onde:

$\sum$ é a matriz de correlação do regime saudável e;

$z_j = \Phi^{-1}(u_j)$ (inverso da CDF normal univariada), formando $z = (z_P, z_V, z_\sum)^T$

$$z^T(\sum^{-1} - I)z = \left(\frac{\rho_{12}^2(x_1^2+x_2^2)+\rho_{13}^2(x_1^2+x_3^2)+\rho_{23}^2(x_2^2+x_3^2)-2\rho_{12}x_1x_2-2\rho_{13}x_1x_3-2\rho_{23}x_2x_3)}{(|\Sigma|)}\right) \qquad \textbf{(3.21)}$$

- **Cópula *t-Student,*** capaz capturar dependência com cauda pesada**:**

$$c_t(u, \sum, df) = \frac{t_{R,df}(z)}{\prod_{j=1}^{3} t_{df}(z_j)} |\sum|^{-\frac{1}{2}} \qquad (3.19)$$

$$C_t(u_1.u_2.u_3: \Sigma, df) = |\sum|^{-\frac{1}{2}} \frac{\Gamma(\frac{df+3}{2})\left[\Gamma(\frac{df}{2})\right]^2 (1+\frac{1}{df} z^T \Sigma^{-1} z)^{-\frac{df+3}{2}}}{\left[\Gamma(\frac{df+1}{2})\right]^3 \prod_{i=1}^{3} (1+\frac{z_i^2}{df})^{-\frac{df+1}{2}}} \qquad (3.22)$$

Onde:

$t_{\sum,v(.)}$ é a densidade t multivariada;

$t_{df}(.)$ é a densidade t univariada;

df é o número de graus de liberdade (controla espessura da cauda);

$z_j = {t_{df}}^{-1}(u_j)$ (inverso da CDF t univariada).

$$z = (z_P, z_V, z_R)^T$$

A cópula t é vantajosa por capturar dependência de cauda (*tail dependence*), útil quando falhas que induzem desvios extremos e coerentes nas variáveis (Embrecht et al., 2001).

Cálculo do score

Para cada observação x_i, o interesse reside em avaliar a sua probabilidade sob o modelo normal. O score é definido baseado na log-verosimilhança da densidade da cópula:

$$\ell_i = log\, p\left(x_i | Y = 1, \hat{\theta}\right) = \log(c\left(u_{Pi}, u_{Vi}, u_{Pi}, \hat{\theta}\right)) \qquad (3.23)$$

onde $\hat{\theta}$ representa a estimativa *a posteriori* dos parâmetros. Este score é proporcional ao log da verosimilhança marginal da observação em regime normal, sendo utilizado como estatística de decisão.

Valores elevados de ℓ_i indicam observações compatíveis com estrutura de dependência do regime normal, enquanto valores reduzidos sugerem combinações improváveis, portanto, potenciais valores anormais.

Definição do limiar probabilístico

A detecção de anomalia é realizada por detecção de novidade. Define-se um limite probabilístico (10.º percentil) dos scores ℓ_i associados à classe normal.

Para todas as observações i:

$$score_i = \ell_i = log(c(u_{1i}, u_{2i}, u_{3i}, \theta)) \qquad (3.24)$$

O limiar de classificação é o percentil inferior (10%) dos scores do estado Normal:

$$L = \mathrm{Q}_{0.1}(\ell i \mid \text{Normal}) \qquad (3.25)$$

Critério de classificação final

$$\hat{y}_i = \begin{cases} 1, se\ \ell i \geq L \text{ (Normal)} \\ 0, se \ell i < L \text{ (Anormal)} \end{cases} \qquad (3.26)$$

Interpretação física

As cópulas modelam explicitamente a distribuição geradora dos dados sob funcionamento normal, permitindo quantificar a incerteza associada à dependência entre variáveis SCADA. Falhas operacionais introduzem padrões de dependência incompatíveis como o modelo aprendido, resultando numa redução significativa da probabilidade *a posteriori* das observações, permitindo representar relações lineares (cópula gaussina), dependências de cauda pesada (cópula-t) e padrões multivariados complexos. Falhas mecânicas ou eléctricas tendem a quebrar a coerência física entre potência, vento velocidade do rotor ou torque, produzindo combinações estatisticamente improváveis sob dependência normal. Este desacoplamento torna os modelos de cópulas particularmente eficazes na detecção desvios operacionais, características de falhas probabilísticas em sistemas mecânicos acoplados (Joe, 2014).

3.5.4. Estratégia de classificação em actualização temporal "*one-step-ahead*"

De modo a reproduzir um cenário realista de monitorização em tempo real, foi adaptada uma estratégia de classificação *one-step-ahead*, sugerida por Song et al. (2018). Nesta abordagem, a decisão relativa ao estado de condição da turbina no instante i é tomada utilizando exclusivamente a informação disponível até ao instante $i - 1$, evitando qualquer utilização de dados futuros.

Considera-se uma sequência temporal de observações SCADA $\{x_1, x_2, \dots, x_n\}$. O classificador estatístico é constituído com base no conjunto histórico $D_{i-1} = \{x_1, x_2, \dots, x_{i-1}\}$, sendo a previsão do estado no instante i dada por:

$$\hat{y}_i = f(x_i | D_{i-1}) \qquad (3.27)$$

No método Normal Multivariado, esta estratégia traduz-se na reestimação adaptativa dos parâmetros do modelo sempre que existe um número suficiente de observações normais, permitindo acompanhar as variações lentas no regime operacional. No método baseado em cópulas, a observação corrente é avaliada com base na estrutura de dependência aprendida a partir do comportamento normal histórico, assegurando que a classificação reflecte apenas informação passada.

A adopção da estratégia *one-step-ahead* garante uma avaliação realista do desempenho dos métodos propostos, elimina o risco de fuga de informação temporal e assegura a aplicabilidade prática dos modelos em sistemas de controlo de condição baseados em dados SCADA, onde as decisões devem ser tomadas sequencialmente à medida que novas observações se tornam disponíveis.

3.6. Avaliação de desempenho

A avaliação do desempenho dos métodos de detecção de estados de condição foi realizada com base em métricas clássicas utlizadas em problemas de classificação binária, especialmente em contexto de detecção de falhas ou anomalias em sistemas industriais. Estas métricas permitem quantificar, de forma objectiva, capacidade de cada abordagem em identificar correctamente estados normais e anormais da turbina eólica.

No presente trabalho foram consideradas 5 métricas: especificidade, erro e acurácia, definidas de acordo com Song et al. (2018) e F1-score e MCC sugeridas por Pinna (2024). A especificida de mede a capacidade de identificar correctamente os pontos anormais, sendo considerada a métrica mais crítica para detecção de falhas (Kusiak & Verma, 2011); o erro representa a taxa global de classificações incorrectas; a acurácia reflecte o grau geral de acerto do classificador, F1-score mede o equilíbrio entre a sensibilidade e a precisão, e o MCC que indica a concordância entre os rótulos reais e previstos. Todas as métricas foram derivadas da matriz de confusão, ferramenta reconhecida por sintetizar acertos e erros, e permitir comparações sistemáticas entre diferentes abordagens.

3.6.1. Matriz de confusão

"Um método bastante utilizado para analisar resultados produzidos pelos classificadores é a matriz de confusão e as medidas de desempenho que dela resultam" (Pinna, 2024, p.38). Matriz de confusão é uma tabela que organiza as previsões de modelo em relação aos valores reais, distribuindo-as em quatro categorias principais: verdadeiros positivos (VP), verdadeiros negativos (VN), falsos positivos (FP) e falsos negativos (FN). Essa estrutura permite avaliar de forma clara o número de acertos e de erros para cada classe, servindo como base para o cálculo das métricas. A tabela 2, ilustra o esquema da matriz de confusão.

Tabela 2: Esquema de matriz de confusão.

		Classe verdadeira (referência)	
		0	1
Classe prevista	0	VN	FN
	1	FP	VP

Fonte: Autoria própria

onde:

Verdadeiros positivos (VP): número de dados normais classificados correctamente como normais

Verdadeiros negativos (VN): número de dados anormais classificados correctamente como anormais

Falsos positivos (FP): número de dados anormais classificados incorrectamente como normais

Falsos negativos (FN): número de dados normais classificados incorrectamente como anormais, correspondente. Algumas medidas de desempenho ou métricas baseadas na matriz de confusão são descritas a seguir:

Acurácia é uma métrica global que indica a proporção de classificações corretas em relação ao total de casos avaliados. Sua importância reside no facto de fornecer uma visão geral do desempenho do modelo, embora possa ser enganosa quando há desbalanceamento entre as classes.

$$acuracy = \frac{VN+VP}{VN+VP+FP+FN} \tag{3.28}$$

Especificidade: mede a capacidade do classificador identificar correctamente os casos negativos ou simplesmente estados anómalos. É fundamental quando o custo de não detectar o estado anormal é elevado, pois assegura que o sistema seja eficaz na detecção de falhas ou condições adversas.

$$especificity = \frac{VN}{VN+FP} \tag{3.29}$$

F1-score: é média harmónica entre a precisão e sensibilidade (revocação). É útil quando se deseja um equilíbrio entre precisão e sensibilidade.

$$F1 - score = 2 \times \frac{precisão \times sensibilidade}{precisão + sensibilidade} \tag{3.30}$$

Coeficiente de correlação de Matthews

O *Matthews Correlation coeficiente* (MCC) assume valores entre -1 e 1. Uma pontuação de 1 indica concordância perfeita entre os valores previstos e reais. O MCC leva em conta todos os quatro valores da matriz de confusão: VN, FP, VP e FN. Isso proporciona uma medida mais equilibrada do desempenho do modelo, especialmente útil em contextos de classes desbalanceadas.

$$MCC = \frac{VN \times VP - FP \times FN}{\sqrt{(VP+FP)(VP+FN)(VN+FP)(VN+FN)}} \tag{3.31}$$

Erro: indica a percentagem total de classificações incorrectas, abrangendo tanto os falsos positivos quanto falsos negativos. É relevante porque oferece uma medida das limitações do modelo, permitindo identificar a necessidade de ajustes ou optimizações no processo de classificação.

$$erro = \frac{FP+FN}{FP+FN+VN+VP} \tag{3.32}$$

Em conjunto, a interpretação destas métricas fornece uma avaliação equilibrada, combinando tanto a detecção de estados anormais quanto a minimização dos falsos alarmes, aspectos críticos para a operação e eficiente das turbinas eólicas.

3.6.2. Cenários

Três cenários são implementados nomeadamente: cenário 1, considera três variáveis (potência activa, velocidade do vento e velocidade do rotor), o cenário 2, tem como variáveis, potência activa e

velocidade do vento e o cenário 3 é referente às variáveis velocidade do vento e torque, onde esta última resulta de quociente entre a potência activa e velocidade do rotor. Estes cenários são explicados com detalhe na Secção 4.2.2.

Em conjunto, estas métricas e os cenários propostos, permitem uma avaliação abrangente do desempenho os métodos estudados, evidenciando não apenas a taxa global de acerto, mas principalmente a capacidade de discriminação entre os estados anormais e normais, aspecto crítico em aplicações de controlo de condição e manutenção preditiva de turbinas eólicas.

3.7. Ferramentas computacionais

As análises desenvolvidas neste estudo foram realizadas utilizando o *software* estatístico R, reconhecido pela sua robustez e flexibilidade na análise de dados, modelação estatística e implementação de métodos probabilísticos. A escolha desta plataforma deve-se, em particular, à sua ampla disponibilidade de pacotes especializados para análise de dados SCADA, detecção de anomalias e avaliação de classificadores.

Para a modelação probabilística multivariada foram utilizados pacotes dedicados à estimação de distribuições e ao cálculo eficiente de densidades conjuntas, nomeadamente para distribuição normal multivariada e modelos baseados detecção de novidade. A implementação de modelos de cópulas, incluindo a cópula gaussiana e cópula *t-Student*, recorreu a bibliotecas específicas que permitem o ajustamento de distribuições marginais, a estimação dos parâmetros de dependência e o cálculo de log-verosimilhança associada a cada observação.

A avaliação do desempenho dos métodos foi conduzida com recurso a pacotes destinados à análise de classificadores, possibilitando a construção de matrizes de confusão e cálculo de métricas estatísticas referenciadas na Secção 3.6.1. Estes recursos garantiram uma comparação consistente e reprodutível entre as diferentes abordagens metodológicas consideradas.

A utilização integrada destas ferramentas computacionais permitiu assegurar rigor estatístico, transparência metodológica e reprodutibilidade das análises realizadas, aspectos essenciais em estudos baseados em dados SCADA e em aplicações de controlo de condição de turbinas eólicas.

CAPÍTULO IV: ESTUDO DO CASO

4. Apresentação dos resultados

4.1. Descrição dos dados

O controlo da condição de turbinas eólicas utilizando dados SCADA é uma abordagem consolidada para garantir eficiência operacional e redução de custos de manutenção em parques eólicos (Maldonado-Correa et al., 2020). Sistemas SCADA registam, em intervalos regulares, uma ampla gama de variáveis, como velocidade de vento, potência de saída, velocidade de rotor, temperatura de componentes, ângulo das pás. Esses dados fornecem suporte para análises estatísticas avançadas, modelagem preditiva e estratégias de manutenção baseada em condição (Wang et al., 2014).

Neste estudo, o objectivo é avaliar a eficácia de uma abordagem baseada em dados SCADA para inferir o estado de condição de duas turbinas eólicas, designadas como WTG01 e WTG02, operando em um parque eólico situado na região centro de Portugal. Os dados foram colectados ao longo de todo ano de 2012, com uma periodicidade de 10 minutos.

Embora o sistema SCADA registe dezenas de variáveis, para este trabalho foram seleccionadas três variáveis principais: velocidade de vento, potência de saída e velocidade de rotor. Essa selecção baseia-se no significado físico e na relevância dessas variáveis para o desempenho operacional das turbinas eólicas, bem como recomendações da literatura para reduzir a dimensionalidade dos dados e, evitar redundâncias e melhorando a performance dos modelos estatísticos (Maldonado-Correa et al., 2020, Wang et al., 2014).

A velocidade do vento é a variável externa mais relevante, determinando a energia cinética disponível para conversão eléctrica. A potência de saída indica a eficiência do sistema na conversão dessa energia, enquanto a velocidade do rotor fornece informações sobre a resposta mecânica da turbina.

Modelos baseados em redes bayesianas têm sido amplamente aplicados em diagnóstico de falhas de turbinas eólicas, pois possibilitam representar de forma probabilística as relações entre variáveis de operação e estados de condição (Wang et al., 2014). Essa abordagem permite a actualização dinâmica das probabilidades condicionais à medida que novos dados são colectados, oferecendo uma estimativa confiável do estado de condição dos equipamentos.

4.1.1. Variável derivada: Torque

Com base nas variáveis seleccionadas, foi criada uma variável derivada denominada Torque, definida como o quociente entre a potência de saída (P) e velocidade de rotor (ω):

$$Torque = \frac{P}{\omega}$$

O torque representa a força rotacional exercida no eixo da turbina, reflectindo a interacção entre a produção eléctrica e a resposta mecânica do equipamento (Verma et al., 2022). A inclusão do torque no modelo, busca aumentar a sensibilidade para detectar anomalias e antecipar falhas mecânicas e criando um novo cenário na modelação estatística proposta, de modo a permitir a comparação dos resultados.

A utilização do torque apresenta três vantagens principais. Primeiramente, ele é altamente sensível a anomalias mecânicas, como desgaste em engrenagens ou rolamentos, que podem não ser detectadas apenas pela potência ou rotação do rotor (Verma et al., 2022). Em segundo lugar, fornece informação integrada entre aspectos eléctricos e mecânicos, melhorando a interpretação do comportamento dinâmico da turbina eólica (Maldonado-Correa et al., 2020). Por fim, a sua utilização pode melhorar a acurácia dos modelos preditivos, permitindo identificar padrões sutis de degradação antes que resultem em falhas significativas (Wang et al, 2014; Wei et al., 2022).

Além disso, as variáveis derivadas, como o torque, são eficazes na identificação de condições operacionais anormais e aumentam a confiabilidade de estratégias de manutenção baseada em condição (Maldonado-Correa et al., 2020). Essa abordagem transforma dados brutos em informações estratégicas para operação e manutenção, permitindo acções preventivas, aumentando a disponibilidade das turbinas e garantindo a continuidade na geração de energia.

4.1.2. Estatísticas descritivas dos dados brutos, Turbina WTG01

As tabelas 3 apresentam as estatísticas descritivas básicas (média, mediana, primeiro e terceiro quartis, mínimo e máximo) dos três principais parâmetros SCADA analisados antes do pré-processamento: velocidade do rotor, velocidade de vento e potência de saída. Esses indicadores fornecem uma visão geral do comportamento operacional da turbina eólica WTG01 ao longo do período de colecta de informação.

Velocidade do rotor

A velocidade do rotor variou de 0 rotações por minuto (rpm) a 14.95 rotações por minuto (rpm). O valor mínimo igual a zero revela que, em determinados momentos, a turbina esteve totalmente parada, fato comum em períodos de calmaria e manutenção (Wang et al., 2014). O primeiro quartil ($Q_1 = 10.08\ \text{rpm}$) e terceiro quartil ($Q_3 = 14.19\ \text{rpm}$) indicam que em pelo menos 50% do tempo a turbina operou dentro da faixa considerada normal de rotação. A média é menor que a mediana, que por sua vez é menor que Q_3 evidencia assimetria à esquerda, ou seja, a presença de valores baixos puxam a média para valores reduzidos. O máximo de 14.95 rpm mostra que a turbina atinge rotações próximas da condição nominal compreendida entre 9.6 rpm a 14.9 rpm, demonstrando capacidade de operar em regime pleno quando as condições de vento são favoráveis.

Velocidade do vento

A velocidade do vento apresentou um mínimo de 0 m/s, caracterizando períodos de completa calma. O primeiro quartil ($Q_1 = 3.54\ \text{m/s}$) mostra que, em 25 % do tempo, a velocidade do vento esteve abaixo de 4 m/s que corresponde à velocidade de corte (*curt-in*), situação em que a turbina não gera energia. A mediana de 5.96 m/s indica vento moderado, enquanto a média é superior à mediana, sugerindo uma distribuição assimétrica à direita, com tendência a episódios de vento mais forte. O máximo de 24.760 m/s aproxima-se da velocidade de corte superior, estimada em 25 m/s, ponto em que a turbina poderia entrar em modo de protecção para evitar sobrecarga, indicando que a máquina esteve exposta a rajadas significativas.

Potência de saída

A potência de saída variou de $-9.97\ \text{kW}$ a $2415.79\ \text{kW}$. O valor mínimo negativo provavelmente decorre da paragem da turbina, mas com consumo energético nas máquinas ou de períodos em que a turbina esteve em modo de espera (*standby*). O primeiro quartil ($Q_1 = 7.56\ \text{kW}$) mostra que, em 25 % do tempo, a turbina praticamente não produziu energia, consistente com os registos de ventos fracos. A mediana de 265.33 kW muito inferior à média de 657 kW, revela uma distribuição fortemente

assimétrica à direita, indicando que grande parte do tempo a produção foi baixa, com picos de geração em momentos específicos. O terceiro quartil ($Q_3 = 1103.52$ kW) demonstra que 25% de tempo a turbina gerou mais de 1 MW, um desempenho positivo. O pico máximo próximo de 2.3 MW, valor que coincide com a potência nominal indicada por fabricante, confirma que a turbina atinge plena capacidade em determinadas condições.

Do modo geral, os dados demonstram que a produção de energia da WTG01 é bastante variável. Embora haja períodos em que a turbina opere com elevada eficiência e atinja a potência nominal, ela passa grande parte do tempo em regime de geração reduzida, resultado directo da variabilidade natural do vento. Essa análise estatística reforça a importância de estratégias de operação e manutenção que considerem tanto os momentos de alta produção quanto os de períodos de baixa geração, garantindo a confiabilidade e o aproveitamento máximo dos recursos eólicos disponíveis.

Tabela 3: Estatísticas descritivas de dados brutos da turbina 1

Variáveis	WTG01_RotorSpeed	WTG01_WindSpeed	WTG01_ActivePower
Mínimos	Min. : 0.00	Min. : 0.000	Min. : -9.973
1º quartil	1st Qu.:10.08	1st Qu.: 3.535	1st Qu.: 7.559
Mediana	Median :10.96	Median : 5.967	Median : 265.331
Média	Mean : 9.82	Mean : 6.700	Mean : 657.244
3º quartil	3rd Qu.:14.19	3rd Qu.: 9.211	3rd Qu.:1103.516
Máximos	Max. :14.95	Max. :24.760	Max. :2415.786
Número de observações		52704	

Fonte: Software R. Adaptado

4.1.3. Estatísticas descritivas de dados brutos da turbina WTG02

As estatísticas descritivas da turbina WTG02 (tabela 4), obtidas a partir dos dados SCADA brutos, incluem também as mesmas variáveis refenciadas na Secção 4.1.2. Análise destes indicadores permite caracterizar a variabilidade operacional da turbina e identificar diferenças subtis face ao comportamento observado na WTG01.

Velocidade do rotor

A velocidade do rotor da turbina 2 variou de 0 rpm a 14.98 rpm. A ocorrência de valores nulos indica períodos em que a turbina eólica esteve inactiva, os quais podem estar associados tanto a condições de vento insuficiente como paragens operacionais programadas. A distribuição apresenta uma assimetria negativa, evidenciada pela relação média <mediana <terceiro quartil sugerindo que episódios de baixa rotação ocorrem com frequência suficiente para influenciar a média global.

Os valores do primeiro quartil ($\mathrm{Q}_1 = 9.98\,\mathrm{rpm}$) e o terceiro quartil ($Q_3 = 14.23\,\mathrm{rpm}$), indicam que metade das observações se concentram numa faixa de rotação compatível com o funcionamento normal da turbina. O valor máximo próximo de 15 rpm confirma que a turbina 2 atinge regimes de rotação muito próxima do nominal, que segundo as especificações do frabricante é de 16.9 rpm, reflectindo um comportamento estável quando sujeita a condições eólicas adequadas (Kusiak et al., 2009).

Velocidade do vento

A velocidade do vento registada na turbina 2 apresenta um mínimo de 0 m/s, caracterizando situações de completa ausência de vento. O primeiro quartil ($Q_1 = 3.83\,\mathrm{m/s}$) revela que, em 25% do tempo a turbina operou sob velocidades inferiores à velocidade de arranque, limitando a geração de energia. A mediana de 6.37 m/s aponta para um regime do vento é predominantemente moderado, enquanto a média superior à mediana sugere a presença de eventos esporádicos de vento mais intenso. Destaca-se o valor máximo de 23.96 m/s, significativamente elevado, o que evidencia a exposição da turbina a rajadas fortes e potenciais condições de operação próximos do limiar de protecção, com impacto relevante sobre a variabilidade dos restantes parâmetros operacionais.

Potência de saída

A potência activa da turbina 2 variou entre -9.28 kW e 2419.40 kW. À semelhança do comportamento observado na primeira turbina, os valores mínimos negativos registados são atribuídos ao consumo dos equipamentos auxiliares, durante períodos em que a turbina se encontra fora de operação. O primeiro quartil ($Q_1 = 5.08$ kw) mostra que, em aproximadamente um quarto de período analisado, a produção de energia foi praticamente nula, em concordância com a frequência de regimes de vento fraco identificados anteriormente.

A diferença acentuada da mediana de (274.20 kW) e a média (667.45 kW) evidencia uma distribuição assimétrica à direita, indicando que a maior parte do tempo a turbina opera em regimes de baixa a média produção, sendo os valores elevados concentrados em intervalos relativamente curtos. O terceiro quartil ($Q_3 = 1133.99$ kW), revela que, em 25 % do tempo, a potência excedeu 1 MW, enquanto o valor próximo de 2.3 MW confirma a capacidade da turbina em alcançar a potência nominal sob condições favoráveis.

De forma agregada, os resultados evidenciam que a turbina WTG02 apresenta um comportamento operacional marcado por elevada dispersão e sensibilidade às condições de vento. Embora a turbina seja capaz de operar em plena capacidade, a predominância de períodos de produção reduzida reforça o impacto da variabilidade eólica sobre o desempenho energético global. Esta caracterização estatística constitui um elemento fundamental para a definição de estratégias de operação, manutenção e monitorização de condição adaptadas à dinâmica real de funcionamento da turbina eólica.

Tabela 4: Estatísticas descritivas de dados brutos da turbina 2

Variáveis	WTG02_RotorSpeed	WTG02_WindSpeed	WTG02_ActivePower
Mínimos	Min. : 0.000	Min. : 0.000	Min. : -9.277
1^{0}s quartis	1st Qu.: 9.976	1st Qu.: 3.831	1st Qu.: 5.081
Mediana	Median :11.004	Median : 6.368	Median : 274.195
Média	Mean : 9.689	Mean : 7.084	Mean : 667.453
3^{o}s quartis	3rd Qu.:14.233	3rd Qu.: 9.629	3rd Qu.:1133.985
Máximos	Max. :14.979	Max. :23.959	Max. :2419.401
Número de observações		52704	

Fonte: Software R. Adaptado

Comparação entre as turbinas WTG01 e WTG02

A análise comparativa das estatísticas descritivas das duas turbinas revela comportamentos operacionais globalmente semelhantes, mas com diferenças relevantes em termos de variabilidade e exposição a regimes extremos de vento. Ambas as turbinas apresentam períodos recorrentes de paragem, evidenciados por valores mínimos nulos de velocidade do rotor e de vento, bem como distribuições assimétricas da potência activa, marcadas por longos intervalos de baixa produção e picos próximos da

potência nominal, um padrão típico de sistemas eólicos operando sob regimes de vento não estacionários (Manwell et al., 2010).

A análise de coeficiente de variação, indicador que permite comparar a dispersão relativa entre séries com diferentes níveis médios, revela elevados níveis de variabilidade em ambas as turbinas, com valores de 45.15% para a turbina 1 e 43.25% para a turbina 2. Observa-se que, em termos relativos, a turbina 1 apresenta uma variabilidade ligeiramente superior. Contudo, a reduzida diferença entre os valores sugere que ambas as turbinas operam sob condições de variabilidade globalmente equivalentes, reflectindo a forte influência da natureza intermitente e não estacionária do recurso eólico.

Por sua vez, o desvio padrão da velocidade do vento foi de 3.56 para a turbina 1 e de 3.63 para a turbina 2, indicando níveis de dispersão absoluta também semelhantes. Ainda assim, o valor ligeiramente superior observado na turbina 2 sugere maior variabilidade das condições de vento em torno da média, indincando exposição marginalmente mais elevada a flutuações e rajadas de vento. A leitura conjunta destes dois indicadores permite concluir que, embora a turbina 1 apresente maior variabilidade relativa, a turbina 2 está sujeita a uma dispersão absoluta ligeiramente superior.

Nesse contexto, a turbina 2 distingue-se por apresentar maior dispersão nos dados de velocidade do vento, incluindo valores máximos de rotação, significativamente mais elevados, o que sugere maior exposição a rajadas intensas e potenciais regimes de operação próximos dos limiares de protecção. Essa maior variabilidade reflecte-se também na potência activa, indicando um comportamento operacional mais heterogéneo quando comparado com a turbina 1, fenómeno também reportado em estudos baseados em dados SCADA de múltiplas turbinas (Pandit & Infield, 2018).

Apesar destas diferenças, ambas as turbinas demonstram capacidade de atingir regimes de operação plena, reforçando que a variabilidade eólica constitui o principal factor condicionante do desempenho energético e justificando a necessidade de abordagens probabilísticas e multivariadas para a monitorização de estado de condição.

4.1.4. Estatísticas descritivas de dados pré-processados das duas turbinas

Foi aplicado um filtro à velocidade do rotor, considerando apenas valores compreendidos entre 9.6 rpm e 16.9 rpm, com o objetivo de restringir a análise ao regime de operação nominal da turbina. Este

intervalo foi definido de forma a isolar a região em que a dinâmica do sistema é predominantemente estável e caracterizada por relações físicas consistentes entre a velocidade do vento, a velocidade do rotor e a potência gerada. Valores inferiores a este limiar estão tipicamente associados a regimes transitórios, como fases de arranque ou operação abaixo da velocidade de cut-in, indicando elevada variabilidade e reduzida representatividade estatística. Por sua vez, valores superiores refletem frequentemente a atuação dos sistemas de controlo ativo, nomeadamente o controlo de passo (*pitch control*), que introduz não linearidades e limita a resposta aerodinâmica natural da turbina. Assim, esta opção metodológica permite excluir períodos de funcionamento instável, assegurando maior coerência física entre as variáveis analisadas e mitigando a influência de ruído e de valores atípicos nos dados SCADA, em linha com as recomendações da literatura para pré-processamento de dados em turbinas eólicas (Song et al., 201).

No seguimento desta etapa de pré-processamento, as tabelas 5 e 6 apresentam as estatísticas descritivas dos dados das turbinas WTG01 e WTG02, proporcionando uma caracterização quantitativa das principais variáveis operacionais e ambientais. Estas estatísticas, que incluem medidas de tendência central e de dispersão, refletem um conjunto de dados previamente depurado, resultante da remoção de valores espúrios e da exclusão de períodos de inatividade.

Desta forma, a análise das referidas tabelas permite avaliar não apenas a consistência interna dos dados, mas também a variabilidade residual após filtragem, constituindo um passo fundamental para garantir a fiabilidade das etapas subsequentes de modelação estatística, monitorização de condição e deteção de comportamentos anómalos.

Tabela 5: Estatísticas descritivas de dados pré-processados da turbina 1

Variáveis	WTG01_RotorSpeed	WTG01_WindSpeed	WTG01_ActivePower
Mínimos	Min. : 9.602	Min. : 1.740	Min. : -9.973
1^{o}s quartis	1st Qu.:10.530	1st Qu.: 5.058	1st Qu.: 149.687
Mediana	Median :12.133	Median : 7.322	Median : 524.424
Média	Mean :12.440	Mean : 7.903	Mean : 850.669
3^{o}s quartis	3rd Qu.:14.421	3rd Qu.:10.155	3rd Qu.:1478.421
Máximos	Max. :14.952	Max. :24.760	Max. :2415.786
Número de observações		40644	

Fonte: Software R. Adaptado

Tabela 6: Estatísticas descritivas de dados pré-processados da turbina 2

Variáveis	WTG02_RotorSpeed	WTG02_WindSpeed	WTG02_ActivePower
Mínimos	Min. : 9.605	Min. : 1.805	Min. : -9.277
1°s quartis	1st Qu.:10.600	1st Qu.: 5.540	1st Qu.: 164.585
Mediana	Median :12.383	Median : 7.791	Median : 568.150
Média	Mean :12.530	Mean : 8.404	Mean : 880.884
3°s quartis	3rd Qu.:14.478	3rd Qu.:10.647	3rd Qu.:1538.897
Máximos	Max. :14.979	Max. :23.884	Max. :2419.401
Número de observações		40644	

Fonte: Software R. Adaptado

As estatísticas descritivas dos dados pré-processados das turbinas WTG01 e WTG02 permitem comparar de forma integrada o comportamento operacional das máquinas com base em informações limpas e consistentes, um passo essencial em análises de manutenção preditiva (Lorenzo Gigoni et al., 2019). Inicialmente, o conjunto bruto de dados da WTG01 apresentava 52704 observações, das quais 40644 permaneceram após o pré-processamento, enquanto a WTG02 possuía o mesmo volume bruto (52704) e 39836 registos válidos depois da limpeza.

O pré-processamento também afectou directamente os valores mínimos de velocidade de rotor e de vento, que foram ajustados para 9.6 rpm e 1.74 m/s na turbina WTG01, e para 9.6 rpm e 1.805 m/s na WTG02. Mesmo com esse refinamento, as velocidades mínimas de vento continuam abaixo da velocidade nominal de operação, reflectindo a realidade de que ambas as turbinas enfrentam momentos de vento insuficiente para a geração ideal (Pandit & Infield, 2018).

Quando se cruzam essas informações com as estatísticas dos dados brutos, nota-se que a WTG01 apresentou velocidade de rotor variando de 0 a 14.95 rpm, com distribuição assimétrica à esquerda (média inferior à mediana), sugerindo períodos em que a turbina esteve parada ou operou a baixas rotações. A velocidade do vento teve a mediana de 5.97 m/s e picos de 24.76 m/s um regime de ventos moderados com rajadas ocasionais fortes. A potência de saída foi altamente variável, de -9.97 kW a 2415 kW, com muitos momentos de geração reduzida e picos próximos da capacidade nominal de 2,5 MW.

A WTG02 apresentou padrão semelhante: velocidade de rotor entre 0 e 14.98 rpm, também com assimetria à esquerda, indicando que a média foi puxada para baixo por valores de baixa rotação. A

velocidade de vento atingiu 23.96 m/s, nível capaz de accionar o modo de protecção, embora a mediana de 6.37 m/s revele condições geralmente moderadas. Sua potência variou de -9.28 kW a 2419 kW novamente evidenciando longos períodos de baixa geração alternados com momentos de produção plena.

Comparando os dados brutos e pré-processados, fica claro que o tratamento removeu registos irrelevantes ou fora do escopo de operação normal, concentrando a análise nas condições efectivas geração. A elevação dos valores mínimos da velocidade rotor para 9.6 rpm, por exemplo, demonstra que apenas períodos de rotação efectiva foram considerados ou mantidos. Essa filtragem é essencial para identificar padrões de anormalidade e para a construção de modelos de monitoramento de diagnósticos que reflictam a performance real de cada turbina eólica (Song et al., 2018).

Assim, a combinação dos resultados brutos e pré-processados indica que ambas as turbinas operam de forma intermitente, com boa capacidade de atingir o máximo nominal, porém, sujeitas a variações de vento que impactam significativamente a geração de energia. A análise comparativa entre as duas máquinas reforça a necessidade de visualizações adicionais e de métricas estatísticas para detectar desvios, anomalias e oportunidades de optimização na operação e manutenção.

Para além da apresentação das estatísticas descritivas e do cálculo das métricas (apresentadas mais adiante), a etapa que se segue envolve a visualização dos dados SCADA, um processo fundamental para compreender e identificar padrões de anormalidade no funcionamento das turbinas WTG01 e WTG02 (Song et al., 2018; Maldonado-Correa et al., 2020). A construção de curvas de potência e histogramas possibilita avaliar, de forma intuitiva, as variações e distribuições dos três principais parâmetros monitorados: velocidade do rotor, velocidade do vento e potência de saída.

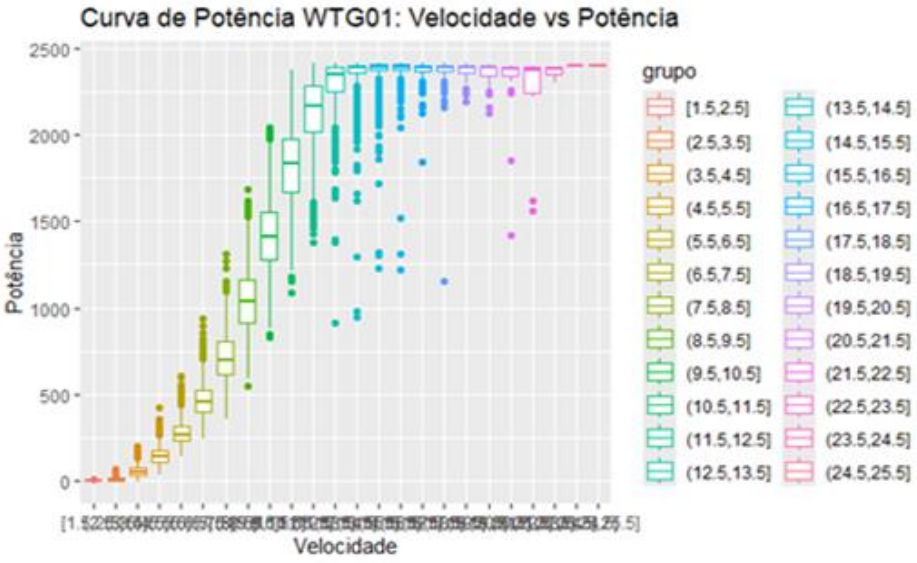
4.1.5. Curvas de Potência

A curva de potência expressa a relação entre a velocidade do vento e energia eléctrica gerada, sendo um parâmetro-chave para avaliação de desempenho (Pandit &Infield, 2018). A seguir apresentam-se as análises das turbinas WTG01 e WTG02.

4.1.5.1. Curva de potência da turbina 1

A análise das curvas de potências permite compreender a relação entre a velocidade de vento e a energia eléctrica efectivamente gerada, sendo um indicador fundamental de desempenho. A curva de potência da turbina WTG01 (figura 4.1) exibe o comportamento típico esperado para aerogeradores de 2.5 MW. Na região inicial, entre aproximadamente 1.5 e 3.5 m/s, a potência permanece nula, reflectindo a ausência de geração significativa. A faixa de 3.5 a 12.5 m/s caracteriza a zona de operação eficiente, onde a potência aumenta de forma quase proporcional à velocidade do vento, ainda que com alguma dispersão vertical nos *boxplots*, possivelmente associada a ajustes operacionais ou turbulência local. Entre 12.5 e 21 m/s a potência estabiliza em torno de 2300 a 2500 kw, representando o regime nominal da turbina eólica. Acima de 22.5 m/s observa-se queda gradual da potência, indicando a aproximação da velocidade de corte superior (*cut-out*) e a actuação dos mecanismos de protecção.

Figura 4.1: Curva de potência da turbina 1.



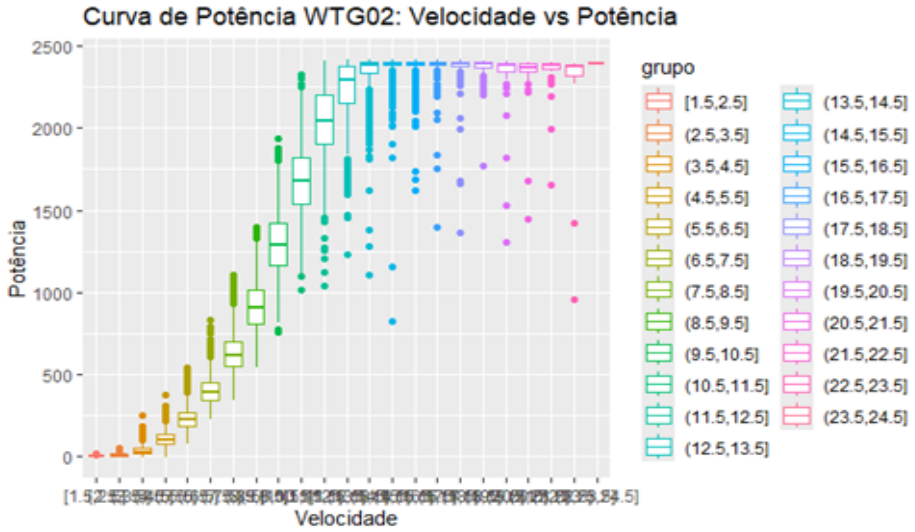
Fonte: Software R.

4.1.5.2. Curva de potência da turbina 2

A turbina WTG02 apresenta padrão geral semelhante à turbina WTG01, mas com algumas particularidades. A região de velocidade insuficiente para geração situa-se em torno de 1 a 3 m/s. Entre 3 e 11 m/s a potência cresce com a velocidade, exibindo dispersões que podem reflectir turbulência ou

actuação activa das pás para o controlo de carga. O regime nominal estende-se de 11 a 17 m/s, alcançando valores próximos de 2500 kW, porém com *boxplots* de maior amplitude, indicando uma instabilidade operacional. Na faixa de 17 a 23 m/s a potência tende à estabilidade, seguida de leve redução e presença de valores discrepantes, sinalizando o início do controlo de segurança. A partir de 23 m/s ocorre a velocidade de corte, quando a turbina reduz ou interrompe a geração para se proteger.

Figura 4.2: Curva de potência da turbina 2.



Fonte: Software R.

Comparação entre as duas turbinas

A comparação directa das duas curvas revela que ambas crescem até cerca de 14 m/s, atingem potência nominal próxima de 2500 kw e apresentam queda ou maior dispersão acima de 22 m/s. Entretanto, a turbina WTG02 evidencia dispersão mais acentuada na faixa de 12 a 22 m/s, com diversos pontos abaixo da potência esperada, enquanto a WTG01 mantém-se mais estável e próxima do limite nominal. Essa diferença sugere melhor desempenho e menores perdas operacionais da WTG01. Além disso, a WTG02 exibe maior número de valores discrepantes em velocidades intermediarias (10 a 17 m/s), possivelmente associados a falhas intermitentes ou controlo menos eficiente, ao posso que a WTG01 apresenta distribuição mais compacta e regular. Esses resultados reforçam a importância de pré-

processamento rigoroso e monitoramento contínuo para detectar desvios e optimizar a operação (Maldonado-Correa et al., 2020).

4.2. Apresentação das métricas estatísticas: estudo computacional.

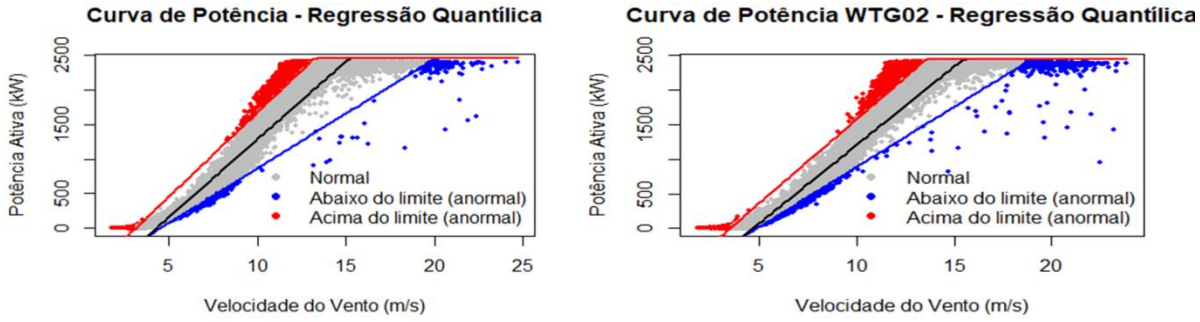
O objectivo central desta análise é identificar de forma robusta os estados normal e anormal das turbinas eólicas WTG01 e WTG02, avaliando e comparando a capacidade de diferentes conjuntos de variáveis em apoiar essa tarefa crítica para manutenção preditiva. Os rótulos, anómalo e normal no conjunto de dados SCADA, foram atribuídos através da regressão quantílica e curva de potência, conforme descrito na secção seguinte.

4.2.1. Aplicação da regressão quantílica na identificação dos estados normais e anormais das turbinas

A regressão quantílica (RQ) é uma técnica estatística que permite modelar diferentes quantis da distribuição condicional de uma variável dependente em função de variáveis independentes. Ao contrário da regressão linear tradicional, centrada na média condicional, a RQ captura toda a distribuição condicional dos dados, sendo particularmente adequada para monitorização de turbinas eólicas, devido à variabilidade operacional e aos extremos de potência observados (Bessa et al., 2012; Schlechtingen & Santos, 2011).

Sendo que os dados SCADA fornecidos não possuem rotulagem supervisionada, a regressão quantílica foi inicialmente aplicada para identificar o estado **normal** e **anormal** do funcionamento das turbinas a partir da curva de potência. A técnica estima limites inferiores e superiores da potência condicionada à velocidade de vento, utilizando os quantis 0.1 e 0.9. Esses limites, permitiram classificar os estados como normal (1), quando a potência observada está dentro dos limites; ou anormal (0) quando a potência excede os limites, indicando possíveis desvios do comportamento normal ou condições atípicas. A escolha destes quantis equilibra a detecção precoce de anomalias com a robustez contra ruído de medição, evitando falsos positivos ou negativos (Bessa et al., 2012).

Figura 4.3: Curvas de potência (RQ), WTG01 e WTG02



Fonte: Software R.

Na figura acima, apresentam-se as curvas de potência das turbinas WTG01 e WTG02, obtidas por meio de regressão quantílica, nas quais observa-se a relação entre velocidade de vento e potência activa. As curvas correspondentes aos quantis inferior ($\tau = 0.10$) e superior ($\tau = 0.90$), delimitam a região de operação considerada normal (faixa cinza). Essa abordagem permite caracterizar o comportamento esperado da turbina sob diferentes condições de vento, sem pressupor a normalidade dos resíduos, o que confere maior robustez à análise (Koenker & Basset, 1978).

Pontos localizados dentro do intervalo entre as curvas quantílicas ($\tau = 0.10$ e $\tau = 0.90$) representam o estado normal de operação, indicando que a turbina converte a energia cinética do vento em potência de forma eficiente e estável. Por outro lado, os pontos abaixo do limite inferior ($\tau = 0.10$), destacados a azul, indicam condições anormais associados a subdesempenho, podendo estar relacionadas a falhas mecânicas, desalinhamento da turbina, sujeita nas pás ou degradação de componentes eléctricos. Já os pontos acima do limite superior ($\tau = 0.90$), a vermelho, também representam anomalias, frequentemente associadas a leituras incorrectas de sensores, ruído de medição ou condições atmosféricas atípicas (Pandit & Infield, 2018).

Ainda os mesmos autores, afirmam que, a RQ permite detectar e diagnosticar desvios significativos da potência observada em relação aos limites estatísticos, distinguindo falhas emergentes de flutuações normais de operação, contribuindo para manutenção preditiva e fiabilidade operacional.

Após a identificação dos estados, foi criada uma coluna “estado”, com os rótulos 1 para operação **normal** e 0 para **anormal**, e posteriormente, adicionada aos dados SCADA de cada turbina como

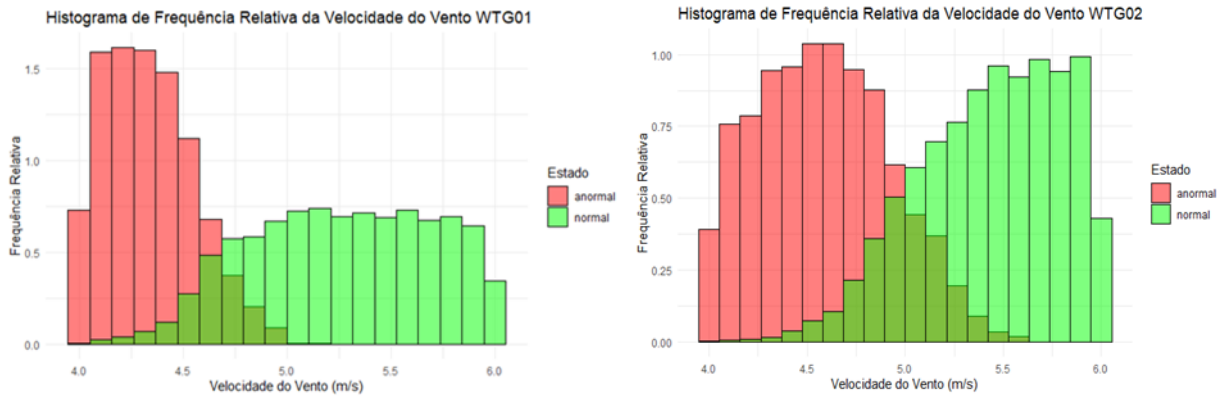
referência para treino dos modelos. Três cenários experimentais foram definidos, aplicando três métodos complementares considerados neste trabalho.

4.2.1.1 Histogramas

Os histogramas da figura 4.4 completam a avaliação das estatísticas descritivas, ao mostrar a distribuição dos três principais parâmetros SCADA classificados como normais e anormais. Na WTG01, a maioria dos estados anormais concentram-se abaixo de 4.5 m/s, como transição clara para o estado normal acima dessa faixa, padrão esperado em turbinas que podem apresentar instabilidades momentâneas em baixas rotações, conforme descrito por Kusiak et al. (2009). Já na WTG02, os estados anormais estendem-se até aproximadamente 5 m/s, com sobreposição considerável entre condições normais e anormais. De acordo com Tautz-Weinert e Watson (2017), essa sobreposição pode reflectir degradação de componentes, como pás ou sistema de yaw, ou problemas mecânicos intermitentes que comprometem o desempenho mesmo em velocidades de vento moderadas.

Em suma, as análises das curvas de potência e dos histogramas indicam que a turbina WTG01 apresenta melhor estabilidade e eficiência, sobretudo na região de potência nominal, enquanto a turbina WTG02 demostra maior variabilidade e presença de anomalias, sugerindo necessidade de inspecções preventivas e estratégias de manutenção mais rigorosas.

Figura 4.4: Histogramas das turbinas 1 e 2.



Fonte: Software R.

4.2.2. Descrição dos cenários experimentais

Após a identificação dos estados normal e anómalo por meio de regressão quantílica aplicada à curva de potência, são definidos três cenários experimentais baseados em diferentes combinações de variáveis operacionais e ambientais provenientes dos dados SCADA. Estes cenários têm como objectivo analisar o impacto de selecção de variáveis e do nível de complexidade do modelo na capacidade de detecção de anomalias e de representação do estado operacional da turbina. A utilização de configurações bivariadas e multivariadas permite uma avaliação comparativa consistente dos métodos propostos, fornecendo um enquadramento estruturado para a análise dos resultados apresentados nas secções subsequentes.

No **cenário 1**, foram utlizadas as variáveis: potência activa, velocidade do vento e velocidade do rotor, de modo a capturar a relação directa entre a energia efectivamente gerada, o movimento mecânico e o recurso eólico disponível. Essa abordagem tem sido destacada como fundamental para compreender a conversão de energia e detectar desvios operacionais, pois anomalias podem surgir quando a potência produzida não corresponde ao regime esperado do vento e rotação (Soe & Htet, 2024). Estudos apontam que a correlação entre potência e velocidade de vento é um dos indicadores mais eficazes para a identificação precoce de falhas aerodinâmicas e eléctricas (Carrol et al., 2016).

No **cenário 2** são consideradas apenas as variáveis: potência activa e velocidade, mantendo o foco na relação fundamental entre o recurso eólico disponível e a energia efectivamente convertida pela turbina. Esta combinação, embora mais simples do que o cenário 1, continua a ser reconhecida por alguns investigadores, como uma das bases sólidas para modelação e detecção de anomalias, uma vez que a potência gerada é fortemente condicionada pelo regime de vento incidente (Burton et al., 2021). Assim, os desvios significativos entre a velocidade de vento e potência correspondente podem sinalizar perdas de eficiência aerodinâmica, problemas no controlo de passo das pás (*pitch*), limitações no sistema de conversão ou até estados iniciais eléctrica (Carrol et al., 2016).

A utilização de um conjunto reduzido de variáveis também permite evidenciar de forma mais clara a estrutura estatística subjacente ao comportamento normal da turbina, dado que a relação potência- vento tende a apresentar padrões bem definidos em condições operacionais saudáveis. Por esse motivo, vários autores defendem que, os modelos bivariados podem ser suficientes para capturar desvios operacionais relevantes, principalmente em cenários onde o fluxo de dados é limitado ou onde se pretende reduzir a complexidade computacional sem comprometer a capacidade detecção (Burton et al., 2021; Astolfi et

al., 2020). Neste contexto, o cenário 2 assume particular importância, pois avalia se uma selecção mínima de variáveis é capaz de representar adequadamente o estado operacional da turbina e identificar anomalias com robustez estatística semelhante aos cenários mais completos.

Por sua vez, **o cenário 3** concentrou-se nas variáveis: torque e velocidade de vento, com foco na análise da resposta mecânica da turbina frente às variações do vento. O torque está directamente associado à integridade estrutural do conjunto de gerador, sendo capaz de revelar esforços excessivos e desgaste de componentes antes que estes se reflictam na potência de saída (Lind et al., 2014). Essa perspectiva tem ganhado destaque, pois combina indicadores de carga mecânica com variáveis ambientais para detectar condições anormais quando o desempenho energético aparente é satisfatório.

A análise comparativa dos três cenários é fundamental para verificar qual conjuntos de variáveis se mostra mais eficiente na identificação de estados normal e anormal. Enquanto o cenário 1 enfatiza o balanço energético e é sensível a desvios de desempenho aerodinâmico e eléctrico, o cenário 2 explora a relação entre a potência gerada e o potencial eólico disponível e o cenário 3 aprofunda-se na integridade mecânica, detectando anomalias de torque mesmo em condições de vento são aparentemente estáveis. Resultados em estudos similares indicam que combinar indicadores de carga mecânica com variáveis ambientais tende a aumentar a capacidade de detecção precoce de falhas estruturais (Remigius & Natarajan, 2021). Dessa forma, a integração de ambos os cenários fornecem uma visão abrangente do comportamento das turbinas, contribuindo para estratégias de operação mais seguras e eficientes.

Em ambos os cenários, aplicou-se a previsão de um passo à frente (*one-step-ahead*) da condição da turbina, baseado no método de actualização temporal. Essa abordagem é amplamente utilizada na análise de séries temporais em dados SCADA, uma vez que assume que o estado futuro imediato da turbina será semelhante ao estado actual. Essa suposição funciona como linha de base simples, porém robusta, especialmente em sistemas com forte correlação temporal, como é o caso das turbinas eólicas. Além disso, a persistência permite identificar de forma clara desvios abruptos entre o valor observado e o valor previsto, os quais podem sinalizar o início de condições anormais ou degradação operacional (Song et al., 2018). Assim, a técnica não apenas facilita a interpretação dos resultados, mas também fornece um ponto de referência sólido para avaliar o desempenho de métodos mais avançados, como método normal multivariado e método de cópulas.

Na perspectiva bayesiana, a previsão *one-step-ahead* é entendida como uma estimativa *a priori*, baseada no estado actual da turbina. A observação subsequente do valor real atua como evidência, permitindo actualizar de forma interativa a crença sobre o estudo futuro da turbina por meio do teorema de Bayes.

Essa abordagem bayesiana permite não apenas calcular a probabilidade do estado futuro ser normal ou anormal, mas também quantificar a incerteza à previsão, algo que os métodos determinísticos simples não fornecem directamente. Em outras palavras, qualquer desvio significativo entre o previsto e observado pode ser interpretado em termos da sua probabilidade, ajustando a confiança na ocorrência de uma condição normal. Além disso, a formulação facilita a integração de informações adicionais (leitura de torque ou histórico de falhas) como evidencia *a priori*, melhorando a detecção de anomalias e fornecendo uma base estatística rigorosa para decisões operacionais.

4.3 Apresentação dos resultados do método dos *bins*

Nas tabelas 7 e 8 apresentam-se os resultados de classificação do método dos *bins* com base nos conjuntos de dados das turbinas WTG01 e WTG02, considerando três cenários distintos destacados na Secção 4.2.2 e métricas descritas na Secção4.2.3.1. Observa-se que o desempenho do método variou entre os cenários, tanto para a turbina 1 quanto para a turbina 2.

Tabela 2: Métricas da turbina 1, Método dos *bins*

Cenário	Erro	Especificidade	Acurácia	F1-scorre	MCC
1	31,5%	35.2%	68.5%	79.6%	0.1107
2	28.5%	28.8%	71.5%	82.2%	0.1104
3	21.2%	11%	78.8%	87.9%	0.1183

Fonte: Autoria própria

Tabela 3: Métricas da turbina 2, Método dos *bins*

Cenário	Erro	Especificidade	Acurácia	F1-scorre	MCC
1	23.9%	15.4%	76.1%	85.9%	0.892
2	29.3%	26.8%	70.7%	81.7%	0.849
3	28.4%	29%	71.6%	82.2%	0.887

Fonte: Autoria própria

A especificidade, que reflecte a capacidade de o sistema detectar correctamente os estados anormais da turbina, é uma métrica de particular interesse no contexto da manutenção preditiva de turbinas eólicas. Conforme argumentam Schlechitingen & Santos (2011) e Aggarwal (2013), a detecção precoce de padrões anómalos em dados operacionais permite antecipar falhas incipientes, reduzir o risco de ocorrências críticas e optimizar os custos associados à operação e manutenção dos sistemas eólicos.

Em relação aos resultados da **turbina** 1, o cenário 3 apresentou o menor erro (21.2%) e maior acurácia (78.8%), embora sua especificidade tenha sido a menor (11%). Já o cenário 2 reduziu o erro de 31.5% no cenário 1 para 28.5% e aumentou a acurácia de 68.5% para 71.5%, com especificidade de 28.8%, oferecendo um desempenho mais equilibrado, o *F1-score* (82.2% no cenário 2) indica que o modelo manteve boa precisão e sensibilidade na classificação de ambos os estados, enquanto o MCC (0.1104) confirma que o equilíbrio entre verdadeiros positivos e negativos se manteve estável, mesmo com a alteração das variáveis.

Na turbina WTG02, observa-se comportamento diferente. O cenário 3 atingiu a maior especificidade (29%) com acurácia de 71.6% e o erro de 28.4%. O cenário 2, embora apresentando o erro mais elevado (29.3%) e acurácia menor de (70.7), apresentou o aumento de especificidade de 15.4% (no cenário 1) para 26.8%, melhorando significativamente a detecção dos estados anormais. O F1-*score* (81.7%) e o MCC (0.849) indicam que o modelo mantém um desempenho robusto na classificação, equilibrando precisão e sensibilidade, apesar do aumento do erro global.

De um modo geral, os resultados indicam que o cenário 2 proporcionou o desempenho mais equilibrado do método dos *bins*, com aumento de especificidade, manutenção de boas métricas de F1 e MCC e redução do erro em algumas situações. A combinação das variáveis relevantes (potência activa e velocidade do vento) torna o modelo mais sensível à detecção de estados anormais, enquanto a turbina WTG01 manteve melhor desempenho global na detecção de anomalias devido à maior especificidade em todos os cenários. Esses resultados reforçam a importância da escolha adequada das variáveis para optimizar o desempenho do método dos *bins* na classificação de condições anormais e normais das turbinas.

4.4. Resultados do método baseado em distribuição normal multivariada

4.4.1. Resultado de modelo estático

Os resultados obtidos (tabelas 9 e 10) com os modelos baseados em distribuição normal multivariada tradicional (estacionários) evidenciam uma melhoria substancial na detecção de anomalias quando comparado ao método dos *bins*.

Tabela 4: Métricas da turbina 1, Método de distribuição normal multivariado

Cenário	Erro	Especificidade	Acurácia (*acuracy*)	*F1-scorre*	MCC
1	19.7%	50.9%	80.3%	88%	0.074
2	18.2%	54.9%	81.8%	88.8%	0.127
3	18.4%	54.5%	81,6%	88.7%	0.126

Fonte: Autoria própria

Tabela 5: Métricas da turbina 2, Método de distribuição normal multivariado

Cenário	Erro	Especificidade	Acurácia (*acuracy*)	*F1-scorre*	MCC
1	18.9%	53.1%	81.3%	88.4%	0.53
2	17.8%	56%	82.1%	89%	0.57
3	17.6%	56.6%	82.3%	89.1%	0.58

Fonte: autoria própria

De modo geral, as métricas revelam que a consideração conjunta das variáveis operacionais contribui para representação mais fiel da dinâmica da turbina, reflectindo-se num desempenho mais robusto em todas as métricas avaliadas.

A especificidade foi a métrica que mais se beneficiou da abordagem multivariada. Para a turbina WTG01, os valores situaram-se entre 50.9% e 54.9%, com destaque para o cenário 2, que apresentou o melhor desempenho entre os cenários analisados. Para a turbina WTG02, a especificidade variou entre 53.1% e 56.6% sendo o cenário 3 aquele que proporcionou os maiores ganhos. Esses resultados demonstram que diferentes combinações de variáveis fornecem contribuições diferenciadas para a

discriminação entre estados normais e anormais, reforçando a importância de selecção adequada de indicadores de condição.

No que diz respeito ao erro, os valores observados permanecem entre 17.6% e 19.7%, representado uma redução significativa em relação ao método dos *bins*. Na turbina WTG01, o melhor desempenho foi obtido no cenário 2 (18.2%), seguido muito perto pelo cenário 3 (18.4%), enquanto o cenário 1, embora consistente, apresentou o maior valor entre os três (19.7%). Para a turbina WTG02, a tendência de melhoria foi ainda mais pronunciada, com o menor erro registado no cenário 3 (17.6%), seguido pelo cenário 2 (17.8%) e pelo cenário 1 (18.9%). Redução do erro nestes modelos indica que a modelação multivariada reduz substancialmente a probabilidade de classificações incorrectas, em particular para estados normais em regimes operacionais variáveis.

O F1-score, reforça esta conclusão, uma vez que apresentou valores elevados em todos os cenários, variando entre 88% e 89.1 %. Na turbina WTG01, o cenário 2 (88.8%) e cenário 3 (88.7%) apresentaram desempenhos similares e superiores ao cenário 1 (88%). Na WTG02, verificou-se tendência semelhante, com cenario3 (89.1%) e cenário 2 (89%), apresentando os melhores resultados. O F1-score elevado indica a capacidade do modelo em manter equilíbrio entre a detecção correta das anomalias e controlo dos falsos positivos, algo fundamental em sistemas de manutenção preditiva.

Por fim, o MCC, métrica que integra todos componentes da matriz de confusão, demonstrou diferenças relevantes entre as turbinas. Na turbina WTG01, os valores permaneceram relativamente baixos (0.074 a 0.127), ainda que o cenário 2 e cenário 3 tenham sugerido maior coerência entre as previsões e estados reais, comparativamente com o cenário 1. Já na WTG02, o MCC apresentou valores significativamente superiores (0.53 a 0.58), com destaque para o cenário 3, que alcançou o melhor desempenho (0.58). Essa diferença pode ser atribuída a menor variabilidade operacional da turbina WTG02 ou padrões mais bem definido entre classes.

De modo geral, os modelos de distribuição normal multivariados estacionários demonstram ganhos expressivos e consistentes em todas as métricas, destacando que a incorporação conjunta das variáveis aumenta a capacidade discriminação entre estados normais e anormais.

4.4.2. Método de distribuição normal multivariado *one-step-ahead*.

Nas tabelas 11 e 12, são apresentados os resultados da variante com actualização temporal.

Tabela 6: Métricas de turbina 1, Método de distribuição normal multivariado *one-step-ahead*.

Cenário	Erro	Especificidade	Acurácia (*acuracy*)	F1-scorre	MCC
1	35.4%	70.9%	64.6%	74%	0.086
2	27.5%	74.1%	72.5%	80.7%	0.121
3	29.4%	70.1	70.%	79.4%	0.779

Fonte: Autoria própria

Tabela 7: Métricas da turbina 2, Método de distribuição normal multivariado *one-step-ahead*.

Cenário	Erro	Especificidade	Acurácia (*acuracy*)	F1-scorre	MCC
1	25.1%	57.3%	74.9%	83.5%	0.49
2	25%	58.1%	74.9%	83.5%	0.50
3	25.7%	60.2%	74.0%	82.9%	0.48

Fonte: Autoria própria

Ao incorporar a componente temporal, os modelos *one-step-ahead* apresentam melhorias ainda mais expressivos na detecção de anomalias, sobretudo no que diz respeito à especificidade. Segundo a visão de Bessa et al. (2012), a inclusão de informação histórica permite capturar padrões transientes que antecedem o processo de falha, contribuindo para uma resposta mais sensível às alterações operacionais.

Na turbina WTG01, especificidade aumentou significativamente, situando-se entre 70.1% e 74.1%. O cenário 2 apresentou o melhor desempenho com 74.1%, seguido pelo cenário 1 com (70.9%) e pelo cenário 3 (70.1%). Na turbina WTG02, os valores variaram entre 57.3% e 60.2%, com cenário 3 novamente apresentando o melhor resultado. Essa melhoria demonstra os modelos temporais são particularmente eficazes em identificar estados anormais, mesmo em condições de maior variabilidade operacional.

O aumento da sensibilidade do estado anómalo, no entanto, resultou num acréscimo do erro total, que oscilou entre 25% e 35%. Na WTG01, o menor erro foi observado no cenário 2 (27.5%), enquanto o cenario1 apresentou o maior valor (35.4%). Na WTG02, os erros apresentaram-se mais equilibrados

entre os cenários, variando apenas entre 25% e 25.7%. Segundo Hodge & Austin (2004), este efeito pode ser interpretado como um *trade off* típico de detecção de falhas: ao priorizar a captura de anomalias, os modelos tendem a aumentar o número de falsos positivos, elevando ligeiramente o erro global.

Quanto à acurácia, mesmo com a maior taxa de erro, o desempenho dos modelos temporais manteve-se satisfatório. Na turbina WTG01, o cenário 2 apresentou novamente o melhor resultado (72.5%), seguido pelo cenário 3 (70%) e pelo cenário 1 (64.6%). Na turbina WTG02, a acurácia manteve-se estável e elevada nos três cenários, com valores em torno de 74% a 74.9%. Estes resultados demonstram que, mesmo privilegiando a detecção de falhas, os modelos temporais continuam a classificar correctamente a maioria das observações.

O F1-score manteve-se elevado, variando entre 74% e 83.5%. Na WTG01, o melhor desempenho foi verificado no cenário 2 (80.7%) seguido pelo cenário 3 (79.4%) e pelo cenário 1 (74%). Para a WTG02, os valores foram ainda mais homogéneos, com os cenários 1 e 2 apresentando o melhor desempenho (83.5%). O F1-score elevado confirma que os modelos temporais conseguem manter o equilíbrio entre a precisão e sensibilidade, mesmo com maior sensibilidade ao estado anormal.

A métrica MCC, evidenciou os impactos mais marcantes da incorporação de informação temporal. Na WTG01, o cenário 3 apresentou um MCC de 0.779, o valor mais elevado entre todos os modelos e cenários analisados, demonstrando forte correlação entre as previsões e os estados reais. Os restantes cenários apresentaram valores mais modestos (0.086 e 0.1210), ainda que superiores que aos modelos dos *bins.* A WTG02, o MCC manteve-se relativamente estável (0.48 a 0.50), revelando que a actualização temporal melhora a coerência das previsões, mas de forma menos acentuada que na WTG01.

Em suma, os resultados indicam que os modelos de actualização temporal apresentam a melhor capacidade de detecção de falhas, ainda que com custos moderados em termos do erro global. A análise comparativa entre os cenários demonstra que a combinação de potência e velocidade de vento é particularmente eficaz para a WTG01, enquanto a utilização de torque em conjunto com velocidade de vento contribui de forma relevante para a WTG02. Estes resultados são coerentes com estudo recentes, que enfatizam o papel de métodos probabilísticos multivariados com componentes temporais como uma abordagem metodologicamente mais sólida e estatisticamente consistente para o controlo de condição e

detecção de anomalias em turbinas eólicas (Kusiak & Li, 2011). Tais métodos, ao incorporarem perspectivas bayesianas, permitem uma representação mais realista de incerteza e das interdependências dos dados, promovendo diagnósticos mais sensíveis e confiáveis em contextos industriais complexos.

A abordagem da modelagem bayesiana torna-se ainda mais clara ao considerar a natureza de incerteza e estocástica do comportamento aerodinâmico e electrodinâmico de turbinas eólicas. Abordagem que incorpora estrutura probabilística explicita e mecanismos revisão contínua de crenças permitem não apenas melhorar o desempenho classificador, mas também quantificar incertezas associadas às decisões, o que é particularmente importante para a manutenção preditiva. Tal perspectiva é coerente com o enquadramento teórico discutido por Paulino et al. (2018), que destacam a importância de modelos hierárquico e mecanismos de actualização sequencial para capturar dependências multivariadas e dinâmicas temporais em sistemas complexos.

Em consonância com este enquadramento teórico, os resultados obtidos reforçam a vantagem de modelos que integram dependências multivariadas com mecanismos de actualização temporal, uma vez que estes conseguem capturar melhor a distribuição condicional dos estados operacionais, corrigindo previsões à medida novas evidências são observadas.

4.5. Resultados do método baseado em cópulas

4.5.1. Modelos de cópulas estáticos

A análise de modelos de cópulas estáticos (tabelas 13 e 14) comprova que o desempenho depende fortemente da combinação de variáveis escolhidas e do tipo de cópula adoptada.

Tabela 8: Métricas da turbina 1, Método de cópulas

Cenário	Tipo de cópula	Erro	Especificidade	Acurácia (*acuracy*)	*F1-scorre*	MCC
1	gaussiana	24.65%	16.72%	75.35%	85.38%	0.0848
1	*t-Student*	24.7%	16.61%	75.3%	85.4%	0.0839
2	gaussiana	14.28%	68.43%	85.72%	91%	0.672
2	*t-Student*	14.13%	69.17%	85.87%	91.05%	0.679
3	gaussiana	16.61%	56.88%	83.39%	89.67%	0.606
3	*t-Student*	16.55%	57.15%	83.45%	89.69%	0.607

Fonte: Autoria própria

Tabela 9: Métricas da turbina 2, Método de cópulas

Cenário	Tipo de cópula	Erro	Especificidade	Acurácia (*acuracy*)	*F1-scorre*	MCC
1	gaussiana	25.2%	14.1%	74.8%	85.1%	0.004
1	*t-Student*	25.3%	13.6%	74.4%	85.1%	0.0038
2	gaussiana	15.4%	62.9%	84.6%	90.3%	0.13
2	*t-Student*	15.3%	63.3%	87.7%	91.5%	0.13
3	gaussiana	16.9%	55.4%	83.1%	89.5%	0.081
3	*t-Student*	16.8%	55.6%	83.2%	89.5%	0.08

Fonte: Autoria própria

Para a turbina WTG01, observa-se que o cenário 1 apresenta desempenho limitado: a especificidade situa-se apenas entre 16.72 % (gaussiana) e 16.61 *% (t-Student*), com erro de 24.65% a 24.7%, acurácia de 75.3% a 75.35%, F1-score próximos de 85% e MCC inferior a 0.09 (0.0848 para cópula gaussiana e 0.0839 para *a t-Student*). A diferença entre as duas cópulas é mínima, incluindo na especificidade, onde ambas permanecem praticamente iguais. Isso mostra que, neste cenário, a escolha da família de cópula não melhora a capacidade de distinguir estados anormais. Esses resultados indicam que, quando múltiplas variáveis fortemente correlacionados são incluídas simultaneamente, o modelo não consegue captar adequadamente a estrutura de dependência necessária para distinguir estados anormais.

O desempenho melhora substancialmente no cenário 2. A especificidade aumenta de forma expressiva para 68.43% a 69.17%, acompanhada de erro reduzido (entre 14.28% e 14.13), acurácia entre 85.72% e 85.87%, F1-scores de 91% a 91.05% e MCC entre 0.672 e 0.692. Aqui torna-se evidente a vantagem da cópula *t-Student*, cuja especificidade de 69.17% supera a da cópula gaussiana (68.48%), indicando melhor capacidade de identificação de anormais pela via de dependências de cauda pesadas e pequenos desvios operacionais. Este cenário revela que a dependência bivariada entre potência e velocidade do vento é particularmente informativa para detectar desvios do comportamento normal, reflectindo uma estrutura de cópula mais estável e representativa da componente física da turbina.

No cenário 3, observa-se desempenho intermediário: a especificidade situa-se entre 56.88%% e 57.15%%, com erro entre 16.61% e 16.55%, acurácia de 83.39% a 83.45%, F1-score próximo de 89.67% a 89.69%, e MCC em torno de 0.606 a 0.607. Embora sejam muito próximas, a cópula t apresenta especificidade ligeiramente maior (57.15%) comparativamente à gaussiana, mostrando

pequena, mas consistente vantagem na identificação de estados anormais. Assim, embora menos eficaz que o cenário 2, esse cenário ainda captura dependências relevantes entre torque aerodinâmico e regime de vento.

Para a turbina WTG02, verifica-se um padrão semelhante. No cenário 1, a especificidade é baixa 13.6% a 14.1%, com erro entre 25.2% e 25.35, acurácia entre 74.4% e 74.8%, F1-score de 85.1%, e MCC muito reduzido (entre 0.004 e 0.0038). Aqui as diferenças entre as cópulas também são mínimas também em termos de especificidade, indicando que nenhum dos modelos está capturando adequadamente a estrutura de dependência multivariada, reforçando que a inclusão de muitas variáveis não necessariamente amplia a capacidade discriminativa do modelo estático e que a limitação deriva do cenário e não do tipo de cópula.

O cenário 2 novamente apresenta melhor performance, alcançando especificidade de 62.9% a 63.3%, erro entre 15.4% e 15.3%, acurácia entre 84 e 87.7%, F1-scores entre 90.3% e 91.5%, e MCC de 0.13 em ambos os tipos de cópula. A cópula *t-Student* apresenta a maior especificidade (63.3%), ligeiramente superior à gaussiana (62.9%), demonstrando vantagem na capacidade de detecção de anomalias, embora a seja menor do que na turbina WTG01.

O cenário 3 apresenta desempenho intermediário, com especificidade em torno de 55.4% (gaussiana) e 55.6% *(t-Student*), o erro entre 16.9% 16.8%, acurácia de 83.1% a 83.2%, F1-score por volta de 89.5%, e MCC entre 0.08 e 0.081. Mais uma vez, a cópula t obtém ligeira vantagem na especificidade (55.6%) quando comparada à gaussiana (55.4%), sugerindo que, mesmo em cenários com dependências menos intensas, a *t-Student* mantém desempenho marginalmente superior na detecção de estados anormais.

Em síntese, os resultados dos modelos estáticos demonstram que eficácia depende fortemente da escolha de variáveis. Cenários com estruturas bivariadas bem definidas, em particular potência e velocidade de vento forneceram maior estabilidade e melhor capacidade de detecção de falhas. Entre as famílias de cópulas, a *t-Student* apresentou as combinações entre baixo erro, melhores valores de especificidade, acurácia, F1-score e MCC, indicando melhor capacidade de identificação de anormais nos cenários analisados.

4.5.2. Modelos de cópulas com actualização temporal (*one-step-ahead*)

A incorporação da dimensão temporal nos modelos de cópula *one-step-ahead* (tabelas 15 e 16) melhora substancialmente o desempenho, mostrando que a dinâmica entre observações sucessivas contribui de forma crítica para a detecção de anomalias.

Tabela 10: Métricas da turbina 1, Método de cópulas *one-step-ahead*

Cenário	Tipo de cópula	Erro	Especificidade	Acurácia (acuracy)	F1-scorre	MCC
1	gaussiana	42%	52.8%	58%	71.8%	0.201
1	*t-studant*	28.6%	21.1%	71.4%	83%	-0.018
2	gaussiana	11.3%	67.1%	88.7%	93.6%	0.462
2	*t-student*	8.8%	46.1%	91.2%	95.3%	0.425
3	gaussiana	11.7%	50.4%	88.3%	93.5%	0.368
3	*t-student*	8.8%	37.2%	91.3%	95.3%	0.382

Fonte: autoria própria

Tabela 11: Métricas da turbina 2, Método de cópulas one-step-ahead

Cenário	Tipo de cópula	Erro	Especificidade	Acurácia (*acuracy*)	*F1-score*	MCC
1	gaussiana	41.8%	42.3%	58.2%	91.2%	0.012
1	*t-studant*	26.9%	15.0%	73.1%	84.3%	-0.043
2	gaussiana	10.2%	49.4%	89.8%	94.4%	0.131
2	*t-student*	7.9%	37.5%	92.1%	95.7%	0.136
3	gaussiana	10%	44.0%	90.0%	94.5%	0.119
3	*t-student*	7.5%	35.3%	92.5%	96.0%	0.684

Fonte: autoria própria

Para a WTG01, no cenário 1, cópula gaussiana apresenta especificidade de 52.8% (erro de 42%, acurácia 58%, F1-score, MCC 0.201), um salto expressivo em comparação com o modelo estático, enquanto a cópula *t-Student* exibe desempenho reduzido, com especificidade de 21.1%, mas cometeu menos erro (28.6%), acurácia de 71.4%, F1-*score* (83%) e MCC (-0.018) muito reduzido. Esse contraste reforça que a escolha da família copular torna-se ainda determinante quando se incorpora a evolução temporal, influenciando capacidade de capturar a transição entre estados. Conforme destacado por

Nelsen (2006), diferentes famílias de cópulas apresentam capacidades distintas de modelar estruturas de dependência, o que influencia directamente o desempenho na identificação de transições entre estados. Assim, a superioridade da cópula gaussiana em termos de especificidade (52.8%) torna-se consistente com a sua maior estabilidade sob actualizações temporais.

Nos cenários 2 e 3, a inclusão da informação temporal produz ganhos ainda mais evidentes. Ambos os tipos de cópulas apresentam F1-scores superiores a 93%, especificidade variando entre 37.2% e 67.1%, MCC na 0.368 a 0.462. Estes resultados demonstram que a actualização temporal confere maior estabilidade à estrutura de dependência, permitindo ao modelo ajustar-se à variabilidade condicional das séries SCADA.

Para a WTG02, observa-se um comportamento similar. No cenário 1, a especificidade varia entre 15% e 42.3%, dependendo do tipo de cópula. Já nos cenários 2 e 3, obtém-se especificidade entre 35.3% e 49.4%, F1-scores acima de 94%, e MCC entre 0.119 e 0.684, reforçando que a modelagem temporal amplia significativamente a capacidade do modelo em detectar anomalias operacionais em variáveis-chave.

Esses resultados mostram que a inclusão de actualização temporal permite capturar a evolução da dependência entre potência, vento, torque e rotor, aspecto já destacado em estudos como Song et al. (2018), que demonstram que a estrutura temporal é essencial para separar mudanças abrutas associadas a falhas de variações normais. De forma comparativa, observa-se que o impacto da actualização temporal não é uniforme entre os cenários, tipos de cópulas e turbinas. Para a WTG01, os ganhos mais expressivos ocorrem nos cenários 2 e 3, onde a cópula gaussiana atinge especificidade entre 50.4% e 67.1%, superando consistentemente a *t-Student*, o que indica maior capacidade da gaussiana em modelar dependências locais e identificar anomalias em condições operacionais mais complexas. Já na turbina 2, o padrão repete-se, embora com amplitudes menores: a gaussiana mantém melhor especificidade em todos os cenários, especialmente no cenário 2 (49.4%), enquanto a cópula-t exibe valores inferiores, entre 15% e 37.5%. A comparação entre as turbinas ainda mostra que a WTG01 responde de maneira mais sensível à actualização temporal, exibindo maiores variações entre os cenários e famílias de cópulas, enquanto a WTG02 apresenta comportamento mais estável, mas ainda beneficiando-se claramente da componente temporal. Assim, verifica-se que a escolha da cópula e o contexto operacional (cenário) influência de forma decisiva no desempenho do modelo, reforçando que abordagens dinâmicas são fundamentais para detecção robusta de anomalias em séries SCADA.

A comparação entre os modelos de cópulas estáticos e os modelos com actualização temporal (*one-step-ahead*), mostram diferenças marcantes no desempenho, indicando que a introdução da dimensão temporal transforma profundamente a capacidade de detecção de anomalias.

As diferenças entre as cópulas e turbinas tornam-se evidentes quando se compara o desempenho entre os modelos estáticos e modelos com actualização temporal. A cópula gaussiana mostra-se mais eficaz nos modelos temporais, apresentando consistentemente maior especificidade, especialmente nos cenários 1 e 2, onde a incorporação da dinâmica entre observações sucessivas favorece estruturas de dependência mais suaves e localmente ajustadas (Tautz-Weinert & Watson, 2017). Por outro lado, a cópula *t-Student* mantém uma vantagem leve nos modelos estáticos, sobretudo por capturar dependências de caudas pesadas e flutuações operacionais, mas perde desempenho de forma sistemática quando a dinâmica temporal é incorporada, sugerindo que sua flexibilidade em caudas não traduz em melhor adaptação a mudanças sucessivas no tempo (Soe & Htet, 2024).

Do ponto de vista das turbinas, a WTG01 demonstra maior sensibilidade à actualização temporal, exibindo saltos expressivos entre cenários e famílias de cópulas. Isso indica que a estrutura de dependência desta turbina responde fortemente à inclusão da componente temporal, reorganizando as relações entre potência, vento, torque e rotor de forma mais pronunciada. Já a WTG02 apresenta comportamento mais estável, mas mesmo assim beneficia-se claramente da actualização temporal, com melhorias sistemáticas em todas as métricas, mostrando que a evolução temporal reforça o desempenho mesmo em máquinas com dinâmicas menos voláteis.

De forma integrada, observa-se que os modelos estáticos dependem fortemente da escolha das variáveis e são limitadas pela ausência de dependência temporal, revelando dificuldades para identificar mudanças sutis no regime operacional. A actualização temporal melhora quase todas as métricas, permitindo capturar transições entre estados e pequenas alterações que permanecem invisíveis em modelos puramente estáticos (Maldonado-Correa et al., 2020). Nesse contexto, a cópula gaussiana destaca-se como a mais eficaz nos modelos temporais, enquanto a *t-Student* se mostra ligeiramente superior entre os modelos estáticos. Além disso, confirma-se que a WTG01 reage mais fortemente à introdução temporal, enquanto a WTG02 apresenta ganhos mais moderados, mas ainda significativos. Assim, conclui-se que os modelos dinâmicos, especialmente baseados na cópula gaussiana, apresentam superioridade clara e consistente na detecção de anomalias em séries temporais SCADA, reforçando a importância da dependência temporal como componente central em sistemas de monitoramento.

CAPÍTULO V: DISCUSSÃO DOS RESULTADOS

5. Discussão e Análise dos Resultados

A análise dos resultados obtidos na aplicação dos diferentes métodos avaliados, revela de forma clara que o desempenho dos algoritmos de detecção de estados de condição das turbinas em dados SCADA depende fortemente da estrutura probabilística utilizada, da escolha das variáveis e da capacidade de cada abordagem, relações multivariadas e dinâmicas ao longo do tempo.

Essas conclusões estão alinhadas às análises contemporâneas sobre a manutenção preditiva em turbinas eólicas, que destacam a importância de técnicas multivariadas e sensíveis à dinâmica temporal para antecipar estados anormais (Maldonado-Correa et al.,2020; Kusiak & Li.,2011; Tautz-Weinert & Watson., 2017).

5.1. Comparação geral entre os métodos

5.1.1. Método dos *bins*

O método dos *bins*, apesar de simples, apresentou desempenho modesto e alta sensibilidade ao conjunto de variáveis utilizadas. Conforme referem Bessa et al. (2012), métodos determinísticos são frequentemente insuficientes para capturar padrões subtis em sistemas dinâmicos, o que se reflectiu nas especificidades reduzida, particularmente na turbina WTG01. Embora cenários com menor dimensionalidade (por exemplo, potência e velocidade do vento) tenham produzido resultados mais equilibrados, o método mostrou limitações claras em distinguir estados coerentes sob variabilidade operacional.

O desempenho relativamente pior no cenário 1, tanto em WTG01 quanto em WTG02, reforça que a adição indiscriminada de variáveis não melhora necessariamente a capacidade discriminativa, questão amplamente discutida em modelação de sistemas multivariadas (Embrechts et al., 2002).

5.1.2. Modelos baseados em distribuição normal multivariada (estáticos)

A aplicação de modelos baseados na distribuição normal multivariada resultou numa melhoria substancial da capacidade de detecção de anomalias em ambas as turbinas analisadas. Em particular, observou-se um aumento consistente da especificidade, com valores superiores a 50% em todos os cenários avaliados indicando maior capacidade do modelo em reconhecer correctamente os estados anormais de operação. Conforme defendido por Manwell et al. (2010) e Astolfi et al. (2021), comportamento aerodinâmica e electromecânica das turbinas eólicas é intrinsecamente determinado por interdependências entre as múltiplas variáveis operacionais. Consequentemente, modelos estatísticos capazes de representar explicitamente essa estrutura de dependência conjunta revelam-se mais eficazes na identificação de desvios face ao regime normal de funcionamento.

A forte redução do erro e aumento do F1-score (entre 88-89%) demonstram que a modelação multivariada é mais apropriada para cenários de operação real, enquanto MCC mais elevado na turbina WTG02, sugere uma estrutura comportamental mais estável e menos ruidosa nesta turbina, como também observado em análise de campo (Carrol et al., 2016).

5.1.3. Modelos multivariados em actualização temporal (*one-step-ahead*)

A incorporação de dependências temporais produziu o avanço mais expressivo entre os métodos. Como destacado por Mckinnon et al. (2020), a inclusão de informação histórica captura variações transitórias que precedem falhas, permitindo identificar padrões anómalos antes de se consolidarem.

Este efeito é particularmente evidente pelo aumento substancial de especificidade:

WTG01: até 74.1%

WTG02: até 60.2%

Todos os cenários analisados apresentam incremento significativo da capacidade de detectar estados anormais. Embora o erro global tenha aumento, comportamento típico quando se favorece a detecção precoce (Fawcett, 2006). O equilíbrio geral do modelo permaneceu elevado, sustentado pelos altos valores de *F1-score* (74-83%) e MCC (particularmente elevado na WTG01, atingindo 0.779).

A incorporação explícita de dinâmica temporal é coerente com a perspectiva bayesiana defendida por de Finetti (1974), Muteira (1995) e Paulino et al. (2018), segundo os quais a actualização contínua de crenças gera representações mais realistas de incerteza e melhora a acurácia em sistemas estocásticos complexos.

5.1.4. Modelos de cópulas (estáticos)

Os resultados com cópulas estáticas revelam que o desempenho depende intensamente do tipo de variáveis empregues e da família copular. Em cenários com três variáveis (cenário 1), o desempenho foi consistentemente fraco, com especificidade abaixo de 17% nas duas turbinas, mostrando que a dependência multivariada não foi bem captada.

Esse resultado confirma problemas de sobreparametrização (*overparameterization*) e dificuldades na estimação de dependências multivariada discutida por Nelsen (2006).

Por outro lado, os cenários bivariados (cenário 2, especialmente potência e velocidade do vento), mostraram desempenho substancialmente melhor, com especificidade acima de 60% em ambas as turbinas.

A cópula *t-Student* obteve desempenho substancialmente superior à gaussiana em termos de especificidade, erro e MCC, o que se alinha à literatura sobre dependência de cauda e detecção de desvios operacionais extremos (Demarta & McNeil, 2005).

5.1.5. Modelos de cópulas com actualização temporal (*one-ste-ahead*)

A inclusão da componente temporal nos modelos de cópulas mostrou melhorias significativas, sobretudo na turbina WTG01. No cenário 1, a cópula gaussiana alcançou especificidade de 52.8%, um salto expressivo em relação ao modelo estático.

Contudo, um comportamento inesperado foi observado: a cópula *t-Student* apresentou especificidade menor (21.1%), mas erro significativamente reduzido (28.6%) e acurácia elevada (71.4%). Esse contraste mostra que a capacidade de modelar transições temporais não depende de força de cauda, mas

da estabilidade da estrutura de dependência ao longo do tempo, questão ressaltada por Patton (2012) e Nelsen (2006).

Os cenários 2 e 3 apresentam melhorias ainda mais expressivas, com:

F1-score acima de 93%

Ganhos notáveis de especificidade.

MCC superior ao encontrado nos modelos estáticos.

Isso demonstra que a inclusão da dinâmica temporal melhora a coerência estrutural do modelo independentemente da família copular, reforçando conclusões recentes de Soe & Htet (2024) e Verma et al. (2022) sobre a criticidade da informação temporal para prognóstico de falhas.

5.2. Discussão integrada dos resultados

A análise integrada dos métodos revela um padrão consistente: abordagens simples baseadas em método dos *bins* mostram-se insuficientes para representar a complexidade inerente aos dados SCADA, que envolvem relações não lineares e forte variabilidade operacional. Em contraste, os modelos multivariados oferecem um avanço substancial ao incorporar estruturas de correlação entre as variáveis-chave, reforçando princípios físicos amplamente discutidos por Burton et al. (2021) e Manwell et al. (2010), segundo os quais o comportamento aerodinâmico e electromecânico das turbinas depende de interacções simultâneas entre potência, velocidade do vento, temperatura e carga.

Os resultados também demonstram que a inclusão de dependência temporal é o factor que mais contribui para a melhoria do desempenho dos modelos, revelando que estados anormais emergem de forma abrupta. Em geral, tais eventos são precedidos por trajectórias operacionais degradadas, um fenómeno amplamente reconhecido em alguns estudos de prognóstico de falhas. Nesse contexto, enquanto as cópulas estáticas enfrentam limitações associadas à sobreparametrizacão, sobretudo quando aplicadas a conjuntos multivariados extensos. Elas ainda apresentam bom desempenho quando se trabalha com pares de variáveis claramente definidos, preservando a estrutura de dependência de forma robusta.

Por outro lado, as cópulas *one-step-ahead*, se destacam por capturar de forma mais eficaz as transições entre estados normais e anormais, ainda que seu desempenho dependa tanto da família copular escolhida quanto da estabilidade temporal de dependência entre variáveis. Os resultados mostram que, nos cenários estáticos bivariados, a cópula *t-Student* tende a ser superior, especialmente devido à sua capacidade de modelar caudas pesadas. Entretanto, ao incorporar actualizações temporais, a cópula gaussiana apresenta maior estabilidade e desempenho mais consistente, possivelmente pela suavização natural que seu núcleo impõe à estrutura de dependência.

Em síntese, esta análise converge com conclusões de trabalhos anteriores (Song et al., 2018), reforçando que métodos probabilísticos multivariados enriquecidos com componente temporal constituem a abordagem mais eficaz para a detecção precoce de falhas em turbinas eólicas, especialmente em sistemas operacionais complexos como aqueles monitorados via SCADA.

CAPÍTULO VI: CONCLUSÃO: SÍNTESE DOS RESULTADOS

6.Conclusões e considerações finais.

6. 1. Conclusões

O presente trabalho surge da necessidade de avaliar o desempenho de diferentes abordagens estatísticas: método dos *bins*, modelos baseados em distribuição normal multivariados e modelos baseados em cópulas, os dois últimos, em versões estáticos e com actualização temporal (*one-step-ahead*), aplicados à detecção precoce de anomalias em turbinas eólicas a partir de dados SCADA. A análise comparativa dos resultados obtidos permitiu estabelecer um conjunto de conclusões que contribuem para o avanço do controlo de condição baseado em modelação probabilística em sistemas eólicos.

Os resultados demonstram, em primeiro lugar, que o método dos bins apresenta limitações significativas na caracterização do comportamento operacional das turbinas eólicas. Embora simples e computacionalmente eficiente, este método revelou fraca capacidade de discriminação de estados anormais, reflectida em valores de especificidade sistematicamente baixos, frequentemente inferiores a 36 % em ambas a s turbinas analisadas. Mesmo nos cenários mais favoráveis, como o cenário 2 da turbina 1, a especificidade não ultrapassou 28.8 %, enquanto o MCC permaneceu próximo de zero (0.1104), indicando a fraca correlação entre as previsões e estados reais. Estes resultados confirmam que abordagens baseadas exclusivamente em limiares unvariados são insuficientes para captar dependências estruturais e variabilidade operacional complexa.

A adopção de modelos probabilísticos multivariados baseados em distribuição normal representou um avanço substancial face ao método dos *bins*. Os modelos de distribuição normal multivariados estacionários mostraram melhorias consitentes em todas as métricas avaliadas, com reduções significativas do erro global (valores entre 17.6 % e 19.7 %) e aumentos expressivos da especificidade, que passou para intervalos entre 50.9 % e 56.6 %. Paralelamente, os valores elevados de F1-score (entre 88 % e 89 %) e o aumento do MCC, particularmente na turbina 2 (até 0.58), confirmam uma melhor capacidade de equilíbrio entre a detecção de anomalias e o controlo de falsos alarmes. Estes resultados compravam que a modelação conjunta das variáveis operacionais permite representar de forma mais fiel as relações físicas subjacentes ao funcionamento das turbinas, conforme fundamentos apresentados por Manwel et al. (2010) e Burton et al. (2021).

Um dos contributos centrais desta investigação reside na demonstração do impacto decisivo da dependência temporal. Os modelos da distribuição normal multivariados *one-step-ahead* apresentam ganhos particularmente expressivos da detecção de estados anómalos, aumentos substanciais da especificidade, que atingiu valores entre 70.1 % e 74.1 % na WTG01 e até 60.2 % na WTG02. Estes ganhos foram acompanhados por valores de F1-score (até 83.5%) e por um aumento significativo do MCC, destacando-se o valor de 0.779 observado no cenário 3 da turbina 1. Embora a incorporação da componente temporal tenha conduzido a um aumento do erro global (até cerca de 35 %), este comportamento reflecte um trade-off típico de sistemas de detecção de falhas, nos quais a priorização da sensibilidade ao estado anómalo implica um maior número de falsos positivos. Ainda assim, os resultados confirmam que os modelos temporais são mais eficazes na identificação de trajectórias de degradação progressivas, alinhando-se com evidências por Marvuglia e Messineo (2012) e Song et al. 2018).

No que se refere aos modelos baseados em cópulas, os resultados indicam que o seu desempenho depende fortemente da escolha das variáveis e da família copular adoptada. Nos modelos estáticos, os melhores resultados foram obtidos em cenários bivariados bem definidos, em particular com combinação potência- velocidade do vento. Nestes casos, a cópula t-Student destacou-se ligeiramente, alcançando especificidades até 69.17 % na WTG01 e 63.3 % na WTG02, com F1-scores superiores a 91 % e MCC próximos de 0.69. Em contrapartida, cenários com maior dimensionalidade revelam sinais de sobreparametrização e perda de capacidade discriminativa. Nos modelos de cópulas com actualização temporal observou-se uma melhoria global do desempenho, especialmente para cópula gaussiana, que apresentou maior estabilidade e especificidade em cenários dinâmicos, atingindo valores superiores a 60 % em alguns casos (concretamente 67.1 %). Estes resultados confirmam que a eficácia das cópulas está intimamente ligada a natureza de dependência temporal e à estrutura de correlação entre as variáveis, conforme discutido por Nelsen (2006).

De forma global, os resultados convergem para uma conclusão abrangente: modelos probabilísticos multivariados que incorporam dependência temporal, em particular os baseados em distribuição normal multivariada one-step-ahead, constituem as estratégias mais eficazes para detecção precoce de anomalias em turbinas eólicas. Estes modelos combinam elevada capacidade discriminativa (especificidade superior a 70 %), bom equilíbrio entre precisão e sensibilidade (F1-score acima de 80 %) e maior coerência global de previsões (MCC elevado), oferecendo bases sólidas para sistemas de

manutenção preditiva mais sensíveis, interpretáveis e alinhados com a realidade operacional dos parques eólicos modernos.

6.2. Considerações finais e trabalhos futuros

Apesar dos resultados promissores obtidos, este estudo apresenta algumas limitações que abrem espaço para investigações futuras. Em primeiro lugar, os modelos avaliados baseiam-se predominanantemente em prossupostos de normalidade e estrutura de dependência relativamente simples, o que pode limitar a sua capacidade de capturar comportamentos altamente não lineares ou regimes operacionais extremos. Neste contexto, a extensão para cópulas dinâmicas multivariadas de maior dimensão surge como uma linha de investigação natural, permititndo representar dependências complexas e não estacionárias de forma mais flexíveis.

Adicionalmente, a incorporação explícita de modelos bayesianos hierárquicos e mecanismos de actualização sequencial poderá reforçar a capacidade de quantificação da incerteza associada às decisões de diagnóstico, aspecto particularmente relevante em aplicações industriais críticas. A integração de modelos híbridos, combinando abordagens estatísticas baseadas em dados SCADA com modelos físicos simplificados das turbinas, constitui igualmente uma direcção promissora, com potencial para melhorar a interpretabilidade e a robustez dos sistemas de detecção.

Por fim trabalhos futuros poderão explorar a validação destes métodos em conjuntos de dados mais extensos, envolvendo diferentes tecnologias de turbinas, condições ambientais distintas e eventos reais de falhas documentados. Tais extensões permitirão avaliar a generalização dos modelos propostos e contribuir para o desenvolvimento de estratégias de operação e manutenção mais eficientes, sustentáveis e orientadas para o aumento da disponibilidade dos sistemas eólicos.

7. REFERÊNCIAS BIBLIOGRÁFICAS

Aggarwal, C. C. (2013). *Outlier Analysis* (2nd ed.). Springer.

Al-Naser, W., Khan, M. R., & Pecht, M. (2017). *Use of SCADA data for wind turbine condition monitoring: A review and empirical analysis*. Renewable and Sustainable Energy Reviews, 76, 165-175.

Astolfi, D., Castellani, F., Lombardi, A., & Terzi, L. (2021). *Multivariate SCADA Data Analysis Methods for Real-World Wind Turbine Power Curve Monitoring*. Energies, 14(4), 1105.

Bandi, M. M., & Apt, J. (2016). *Variability of the wind turbine power curve*. Applied Sciences, 6(9), 262. https://doi.org/10.3390/app6090262.

Barbetta, P. A. (2018). *Estudo de fatores associados através de regressão quantílica hierárquica: evidências do desempenho escolar no Brasil*. Estudos em Avaliação Educacional, 29(67), 50–68. https://publicacoes.fcc.org.br/eae/article/view/4973.

Bessa, R. J., Miranda, V., Botterud, A., Zhou, Z., & Wang, J. (2012). *Time-adaptive quantile-copula for wind power probabilistic forecasting*. Renewable Energy, 40(1), 29–39. https://doi.org/10.1016/j.renene.2011.08.019.

Brechmann, E. C., & Schepsmeier, U. (2013). *Modeling dependence with C- and D-vine copulas: The R package VineCopula*. Journal of Statistical Software, 52(3), 1–27.https://doi.org/10.18637/jss.v052.i03.

Bishop, C. M. (1994). *Novelty Detection and Neural Network Validation*. IEE Proceedings - Vision, Image and Signal Processing, 141(4), 217–222.

Boyle, G. (2004). *Renewable Energy: Power for a Sustainable Future*. Oxford University. Press.

Burton, T., Sharpe, D., Jenkins, N., & Bossanyi, E. (2021). *Wind energy handbook (3rd ed.)*. Wiley.

Cade, B. S., & Noon, B. R. (2003). *A gentle introduction to quantile regression for ecologists*. Frontiers in Ecology and the Environment, 1(8), 412–420. https://doi.org/10.1890/1540-9295(2003)001[0412:AGITQR]2.0.CO;2.

Carroll, J., McDonald, A., & McMillan, D. (2016). *Failure rate, repair time and unscheduled O&M cost analysis of offshore wind turbines.* Wind Energy, 19(6), 1107–1119. https://doi.org/10.1002/we.1887.

De Finetti, B. (1974). *Theory of probability (Vol. 1)*. New York: Wiley.

Demarta, S., & McNeil, A. J. (2005). *The t copula and related copulas*. International Statistical Review, 73(1), 111–129.

Embrechts, P., McNeil, A., & Straumann, D. (2002). *Correlation and dependence in risk management: Properties and pitfalls.* In Risk Management: Value at Risk and Beyond (pp. 176–223). Cambridge University Press.

Fawcett, T. (2006). *An introduction to ROC analysis*. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

Fowlie, A. (2020). *Objective Bayesian approach to the Jeffreys-Lindley paradox*. Link: https://arxiv.org/abs/2012.04879.

García Márquez, F. P., Tobias, A. M., Pinar Pérez, J. M., & Papaelias, M. (2012). *Condition monitoring of wind turbines: Techniques and methods.* Renewable Energy, 46, 169–178. https://doi.org/10.1016/j.renene.2012.03.003.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis (3rd ed.)*. Chapman and Hall/CRC.

Genz, A., & Bretz, F. (2009). *Computation of Multivariate Normal and t Probabilities*, Springer.

Genz, A., Bretz, F., & Hothorn, T. (2008). mvtnorm: Multivariate normal and t distributions (R package version). R Foundation for Statistical Computing.https://cran.r-project.org/package=mvtnorm.

Geraci, M., & Bottai, M. (2007). *Quantile regression for longitudinal data using the asymmetric Laplace distribution*. Biostatistics, 8(1), 140–154. https://doi.org/10.1093/biostatistics/kxj039.

Lorenzo Gigoni, et al. (2019). *Wind turbine main bearing fault detection: A scada-based framework for condition monitoring*. Applied Sciences, 9(15), 3038. https://doi.org/10.3390/app9153038

Gil, A. C. (2008). *Métodos e técnicas de pesquisa social* (6ª ed.). Atlas.

Global Wind Energy Council. (2025). Global wind report 2025. GWEC.https://www.gwec.net.

Goldstein, M., & Uchida, S. (2016). "*A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data"*, PLOS ONE, 11(4).

Hameed, Z., Hong, Y. S., Cho, Y. M., Ahn, S. H., & Song, C. K. (2009). *Condition monitoring and fault detection of wind turbines and related algorithms: A review*. Renewable and Sustainable Energy Reviews, 13(1), 1–39. https://doi.org/10.1016/j.rser.2007.05.008.

Hofert, M., Kojadinovic, I., Mächler, M., & Yan, J. (2020). *copula: Multivariate dependence with copulas*. R package.https://cran.r-project.org/package=cópula.

Hodge, V., & Austin, J. (2004). *"A Survey of Outlier Detection Methodologies"*. AI Review, 22, 85–126.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: principles and practice* (3rd ed.). OTexts. Disponível em: https://otexts.com/fpp3/.

International Energy Agency (2023). Renewables 2023. IEA. https://www.iea.org.

Joe, H. (1997). *Multivariate models and dependence concepts*. Chapman & Hall. https://doi.org/10.1201/9780367803896.

Joe, H. (2014). *Dependence modeling with copulas*. CRC Press.

Kass, R. E., & Wasserman, L. (1996). *Noninformative Bayesian Priors*. Journal of the American Statistical Association, 91(435), 1343-1370. Link: https://www.jstor.org/stable/2291711.

Koenker, R. (2005). *Quantile regression*. Cambridge University Press.

Koenker, R., & Bassett, G. (1978). *Regression quantiles. Econometrica*, 46(1), 33–50. https://doi.org/10.2307/1913643.

Koenker, R., & Hallock, K. F. (2001). Quantile regression. Journal of Economic Perspectives, 15(4), 143–156. https://doi.org/10.1257/jep.15.4.143.

Kusiak, A., & Li, W. (2011). *The prediction and diagnosis of wind turbine faults.* Renewable Energy, 36(1), 16–23. https://doi.org/10.1016/j.renene.2010.05.014.

Kusiak, A., Zhang, Z., & Verma, A. (2009). *Prediction, operations, and condition monitoring in wind energy*. Energy,34(12),1835-1845. https://doi.org/10.1016/J.RENENE.2008.10.022.

Lakatos, E. M., & Marconi, M. A. (2017). *Fundamentos de metodologia científica* (8ª ed.). Atlas.

Letcher, T. M. (2023). *Future energy*: *Improved, sustainable and clean options for our planet (4th ed.)*. Elsevier.

Lind, P. G., Wächter, M., & Peinke, J. (2014). Fatigue loads estimation through a simple stochastic model. arXiv. https://arxiv.org/abs/1410.8005.

Lindley, D. V. (1972). *Bayesian statistics*: *A review*. Philadelphia: Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611970654.

Machado, J. A. F., & Santos Silva, J. M. C. (2005). *Quantiles for counts. Journal of the American Statistical Association*, 100(472), 1226–1237. https://doi.org/10.1198/016214505000000330.

Maldonado-Correa, J., Martín-Martínez, S., Artigao, E., & Gómez-Lázaro, E. (2020). *Using SCADA data for wind turbine condition monitoring: A systematic literature review*. Energies, 13(12), 3132. https://doi.org/10.3390/en13123132.

Manwell, J. F., McGowan, J. G., & Rogers, A. L. (2010). *Wind energy explained: Theory, design and application (2nd ed.)*. Wiley.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press.

McKinnon, C., Turnbull, A., Koukoura, S., Carroll, J., & McDonald, A. (2020). *Effect of time history on normal behaviour modelling using SCADA data to predict wind turbine failures*. Energies, 13(18), 4745. https://doi.org/10.3390/en13184745.

Mobley, R. K. (2002). *An Introduction to Predictive Maintenance*. Elsevier.

Marvuglia, A., & Messineo, A. (2012). *Monitoring of wind farms' power curves using machine learning techniques.* Applied Energy, 98, 574–583. https://doi.org/10.1016/j.apenergy.2012.04.037.

Murteira, B. J. (1995). *Introdução à Inferência bayesiana*. Working paper n° 21. Universidade Tecnica de Lisboa.

Nelsen, R. B. (2006). *An introdution to copulas*. Springer. Second edition. Portland. USA.

Pandit, R., & Infield, D. (2018). *Gaussian process operational curves for wind turbine condition monitoring*. Energies, 11(7), 1631. https://doi.org/10.3390/en11071631.

Patton, A. J. (2012). *A review of copula models for economic time series*. Journal of Multivariate Analysis, 110, 4–18.

Paulino, C. D., Amaral Turkman, M. A., Murteira, B., & Silva, G. L. (2018). *Estatística Bayesiana: 2ª edição revista e ampliada*. Fundação Calouste Gulbenkian.

Paulino, C. D., Turkman, M. A. A. (2015). *Estatística bayesiana computacional- uma introdução*. Sociedade Portuguesa de Estatística. Lisboa.

Pinna, D. R. (2024). *Identificação de falhas em turbinas eólicas: uma abordagem de aprendizado de máquina centrado em dados*. Dissertação de Mestrado.CEFET/RJ. Rio de Janeiro.

Remigius, W. D., & Natarajan, A. (2021). *Identification of wind turbine main-shaft torsional loads from high-frequency SCADA measurements using an inverse-problem approach*. Wind Energy Science, 6, 1401-1412. https://doi.org/10.5194/wes-6-1401-2021.

Renewable Energy World. (2023). *The repowering mission: Breathing new life into our aging wind turbine fleet*. https://www.renewableenergyworld.com.

Resende, D. V. (2000). *Inferência bayesiana e simulação estocástica (amostragem GIBBS) na estimação de componentes de variância e valores genéticos em plantas perenes*. Embrapa. Colombo.

Rodriguez, C. L. B. (2005). *Inferência bayesiana no modelo normal assimétrico*. Dissertação de Mestrado. Universidade de São Paulo.

Schepsmeier, U., Stöber, J., & Brechmann, E. C. (2017). *CDVine: Statistical inference of C- and D-vine copulas.* R package documentation. https://cran.r-project.org/package=CDVine.

Schlechtingen, M., & Santos, I. F. (2011). *Comparative analysis of neural network and regression based condition monitoring approaches for wind turbine fault detection*. Mechanical Systems and Signal Processing, 25(5), 1849–1875. https://doi.org/10.1016/j.ymssp.2010.12.007.

Sheng, S. (2015). *Prognostics and health management of turbines*. Current status and Future opportunities. National Renewable Energy Laboratory golden. NREL/PR-5000-65605.

Soe, H., M., & Htet, A. (2024). *A comprehensive review of SCADA-based wind turbine performance and reliability modeling with machine learning approaches*. Renewable and Sustainable Energy Reviews, 158, 112077. https://www.jescae.com/index.php/jtie/article/view/1028/325.

Song, Z., Zhang, Z., Jiang, Y., & Zhu, J. (2018). *Wind turbine health state monitoring based on a Bayesian data-driven approach.* Renewable Energy, 125, 172-181.

Tautz-Weinert, J., & Watson, S. J. (2017). *Using SCADA data for wind turbine condition monitoring – A review*. IET Renewable Power Generation, 11(4), 382–394. https://doi.org/10.1049/iet-rpg.2016.0248.

Tavner, P. J. (2012). *Offshore wind turbines: Reliability, availability and maintenance*. Institution of Engineering and Technology.

Tavner, P. J., Xiang, J., & Spinato, F. (2007). *Reliability analysis for wind turbines*. Wind Energy, 10(1), 1–18. https://doi.org/10.1002/we.204.

Tax, D. M., & Duin, R. P. W. (2004). *Support Vector Data Description, Machine Learning*. 54, 45–66.

TWI. (2023). *How long do wind turbines last?* TWI – The Welding Institute. https://www.twi-global.com/technical-knowledge/faqs/how-long-do-wind-turbines-last.aspx.

Verma, A., Zappalá, D., Sheng, S., & Watson, S. J. (2022). *Wind turbine gearbox fault prognosis using high-frequency SCADA data. Journal of Physics: Conference Series*, 2265(3), 032067. https://doi.org/10.1088/1742-6596/2265/3/032067.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S (4th ed.).* Springer. https://www.springer.com/gp/book/9780387954578.

Wang, K., Sharma, V. S., & Zhang, Z. Y. (2014). *SCADA data based condition monitoring of wind turbines. Advances in Manufacturing*, 2(1), 61–69. https://doi.org/10.1007/s40436-014-0067-0.

Wei, L., Qian, Z., Pei, Y., & Wang, J. (2022). *Wind turbine fault diagnosis by the approach of SCADA alarms analysis*. Applied Sciences, 12(1), 69. https://doi.org/10.3390/app12010069.

Yu, K., & Moyeed, R. A. (2001). *Bayesian quantile regression. Statistics and Probability Letters*, 54(4), 437–447. https://doi.org/10.1016/S0167-7152(01)00124-9.

Yu, K., Lu, Z., & Stander, J. (2003). *Quantile regression: Applications and current research areas*. The Statistician, 52(3), 331–350. https://doi.org/10.1111/1467-9884.00363.

ANEXOS

Nestes anexos descrevem-se os procedimentos adotados para a leitura, selecção das variáveis e análise estatística descritiva dos dados SCADA das turbinas eólicas escolhidas, bem como os *scripts* de classificação utilizados na avaliação do estado de condicao. A análise foi realizada no *software* R, iniciando com a importação dos dados brutos a partir de um ficheiro em formato *csv* e selecção das principais variáveis operacionais.

ANEXO I

LEITURRA DOS DADOS E ESTATÍSTICAS DESCRITIVAS DOS ANTES DE FILTRAGEM E VISUALIZAÇÃO DOS DADOS

ANEXO 1.1: IMPORTAÇÃO DOS DADOS

```
# LEITURA DOS DADOS

dados <- read.csv("_frt_12.csv", sep=";")
```

ANEXO 1.2: CÁLCULO DAS ESTATÍSTIVCAS DESCRITIVAS

1. **Estatísticas descritivas de dados Dados brutos**

 # Carregamento de pacote para manipulação dos dados

```
library(dplyr)
```

Seleccao das variáveis

```
dados_selecionados <- dados %>% select(tstamp,WTG02_RotorSpeed, WTG02_WindSpeed, WTG02_ActivePower)
```

Cálculo das estatísticas descritivas

```
summary(dados_selecionados)
```

Tabela de estatístcas descritivas de dados brutos

```
tstamp          WTG02_RotorSpeed WTG02_WindSpeed  WTG02_ActivePower
 Length:52704       Min.   : 0.000   Min.   : 0.000   Min.   :  -9.277
 Class :character   1st Qu.: 9.976   1st Qu.: 3.831   1st Qu.:   5.081
 Mode  :character   Median :11.004   Median : 6.368   Median : 274.195
                    Mean   : 9.689   Mean   : 7.084   Mean   : 667.453
                    3rd Qu.:14.233   3rd Qu.: 9.629   3rd Qu.:1133.985
                    Max.   :14.979   Max.   :23.959   Max.   :2419.401
```

Fonte : software R

2. **Estatísticas descritivas de dados pré-processados**

 # Filtro dos dados usando velocidade do rotor

```
WTG02 <- dados[9.6<=dados$WTG02_RotorSpeed & dados$WTG02_RotorSpeed<=16.9,1:15]
```

 # Selecção das variáveis

dados_selec <- WTG02 %>% select (tstamp,WTG02_RotorSpeed,WTG02_WindSpeed, WTG02_ActivePower)

 # Cálculo das estatísticas descritivas

summary(dados_selec)

Tabela de estatístcas descritivas de dados brutos

```
  tstamp             WTG02_RotorSpee   WTG02_WindSpeed   WTG02_ActivePower

   Length:52704       Min.   : 0.000   Min.   : 0.000    Min.   :  -9.277
   Class :character   1st Qu.: 9.976   1st Qu.: 3.831    1st Qu.:   5.081
   Mode  :character   Median :11.004   Median : 6.368    Median : 274.195
                      Mean   : 9.689   Mean   : 7.084    Mean   : 667.453
                      3rd Qu.:14.233   3rd Qu.: 9.629    3rd Qu.:1133.985
                      Max.   :14.979   Max.   :23.959    Max.   :2419.401
```

Fonte : software R

ANEXO 1.3: Curva de potência e hitograma

 # Curva de potência por intervalo de velocidade do vento

Carregando os pacotes

install.packages("ggplot2")

library(ggplot2)

Criando um novo data frame para plotar a curva

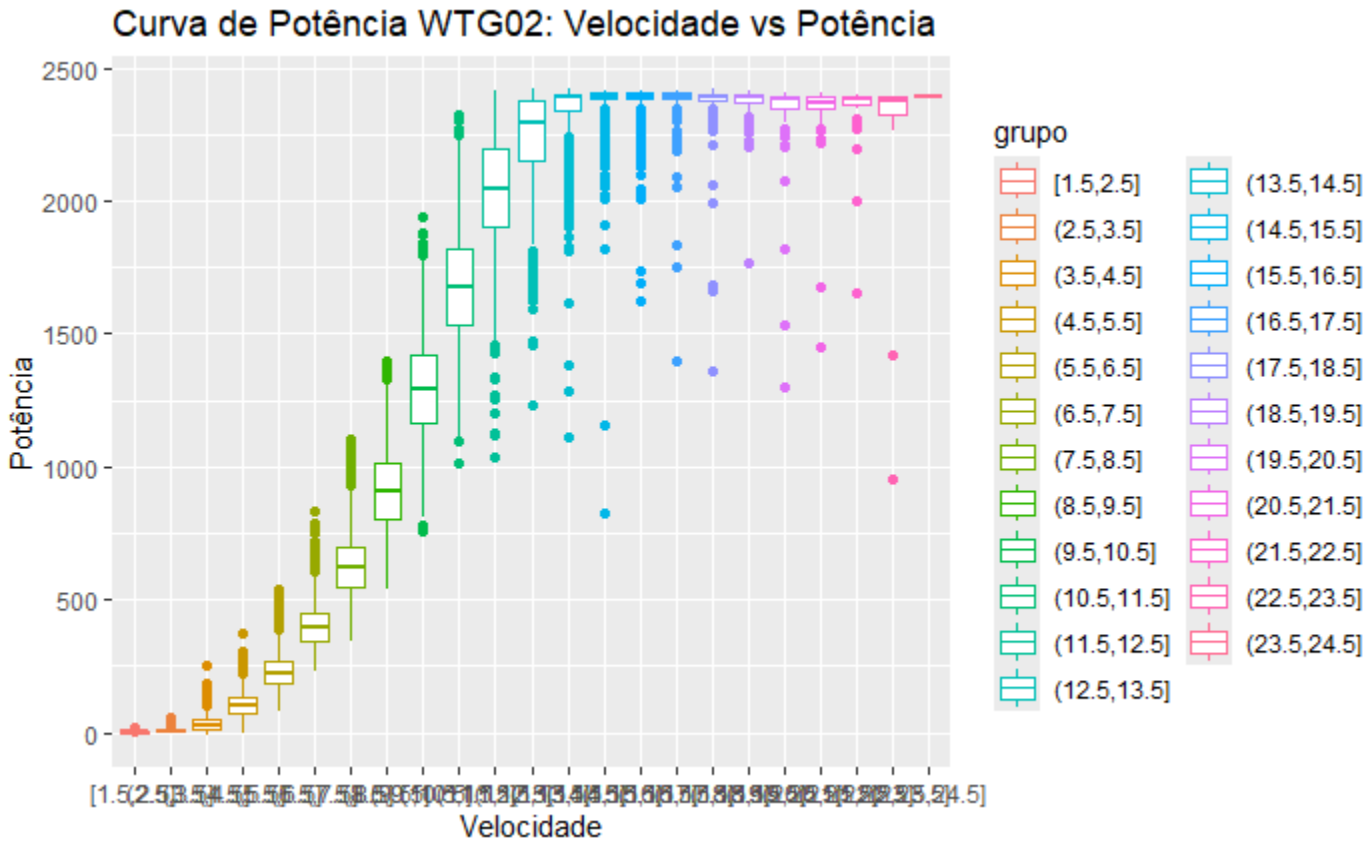
dados_filtrados <- WTG02%>%

+ filter(WTG02$WTG02_WindSpeed > 1 & WTG02$WTG02_WindSpeed < 25) %>%

+ mutate(grupo = cut_width(WTG02_WindSpeed, width = 1))

ggplot(dados_filtrados, aes(x = grupo, y = WTG02_ActivePower,color=grupo)) +

+ geom_boxplot() + labs(title = "Curva de Potência: Velocidade vs Potência", x = "Velocidade", y = "Potência")

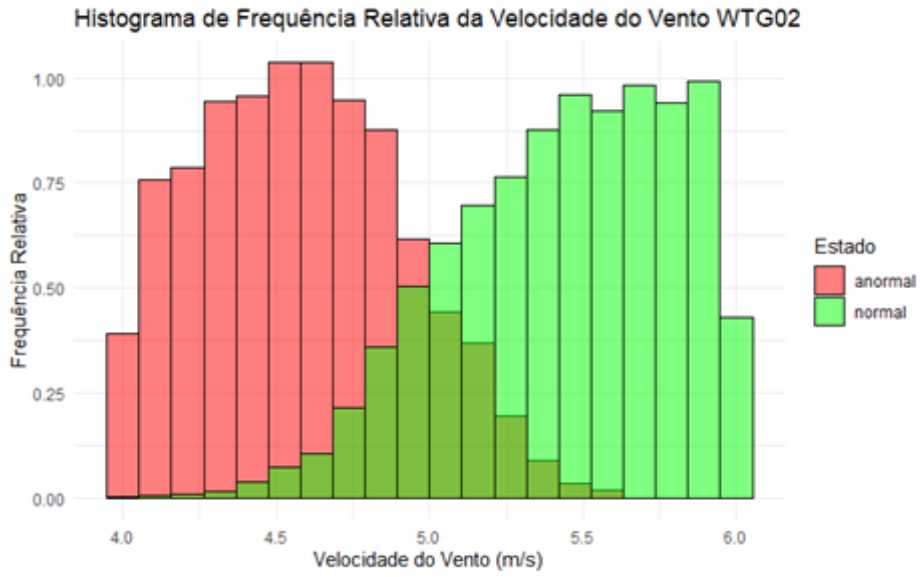


Fonte : Software R

 # Histogramas

```
dados_filtr <- subset(WTG02, WTG02_WindSpeed >= 4 & WTG02_WindSpeed  <= 6)

estado <- numeric(nrow(dados_filtr))

dados_filtr$estado <- ifelse(dados_filtr$WTG02_ActivePower >= 100, "normal", "anormal")

ggplot(dados_filtr, aes(x = WTG02_WindSpeed, fill = estado)) +

  geom_histogram(aes(y = ..density..),

                 position = "identity",  bins = 20, alpha = 0.5,

                 color = "black") + scale_fill_manual(values = c("normal" = "green", "anormal" = "red")) +
```

```
labs( title = "Histograma de Frequência Relativa da Velocidade do Vento (4-6 m/s)",
  x = "Velocidade do Vento (m/s)",
  y = "Frequência Relativa",
  fill = "Estado" ) + theme_minimal()
```



Fonte: Software R

ANEXO II

ROTULAGEM DOS DADOS POR MEIO DE REGRESSÃO QUANTÍLICA E CURVA DE POTÊNCIA

ANEXOS 2.1: Rotulagem de estado de condicão com Regressão Quantílica

 # carregamento de pocote necessário

 # Library(quantreg)

1. Ajuste de modelos

```
# escolher quantis para construir o "intervalo aceitável"

tau_infer <- 0.1

tau_super <- 0.90

# 1) Ajustar os modelos de regressão quantílica (ActivePower ~ WindSpeed)

modelo_infer <- rq(WTG01_ActivePower~ WTG01_WindSpeed, tau = tau_infer, data = WTG01)

modelo_super <- rq(WTG01_ActivePower~ WTG01_WindSpeed, tau = tau_super, data = WTG01)

# 2) Prever os quantis para todas as linhas (vetorizado, mais rápido)

pred_infer <- predict(modelo_infer, newdata = WTG01)

pred_super <- predict(modelo_super, newdata = WTG01)

# 3) Iniciando vetor de estados com NA (para distinguir ausentes)

estado<- rep(NA_real_, nrow(WTG01))

# 4) Loop linha-a-linha que decide o estado com base no intervalo dos quantis

for (i in seq_len(nrow(WTG01))) {  obs_power <- WTG01$WTG01_ActivePower[i]  # potência observada na linha i

low_q  <- pred_infer [i]            # quantil inferior previsto na linha i

 high_q  <- pred_super [i]             # quantil superior previsto na linha i

# se qualquer valor for NA, manter NA no estado (ou poderia optar por 0)

if (is.na(obs_power) || is.na(low_q) || is.na(high_q)) {  estado[i] <- NA_real_  next  }

# classificar: normal (1) se potência estiver dentro do intervalo [low_q, high_q]
```

if (obs_power >= low_q && obs_power <= high_q) { estado[i] <- 1 # normal } else { estado[i] <- 0 # anormal

 # Resultado final

Código:

WTG01$estado <- estado

Explicação: Adiciona o vetor de estados ao dataframe WTG01.

ANEXO 2.2: **Extrato das primeiras 10 observações**

Tabela de Extrato das primeiras 10 observações

	tstamp	WTG01_ActivePower	WTG01_WindSpeed	WTG01_RotorSpeed	estado_qr
1	01/01/2012 00:00	38.32384000	3.783989	10.143170	1
2	01/01/2012 00:10	14.97024000	3.522001	10.072130	0
3	01/01/2012 00:20	19.65428000	3.888776	10.083350	1
4	01/01/2012 00:30	6.20262100	3.552188	10.042860	0
5	01/01/2012 00:40	-0.08620552	3.084853	10.203940	1
6	01/01/2012 03:00	2.36983300	2.490167	10.337400	1
7	01/01/2012 04:00	22.80718000	3.398706	10.239220	1
8	01/01/2012 04:10	16.81236000	3.486639	10.145380	1
9	01/01/2012 04:20	106.84500000	4.666214	10.365440	1
10	01/01/2012 04:30	108.75880000	4.604634	10.362300	1

Fonte: software R. Adaptado

ANEXO 2.3: Curva de potência com limites quantílicos acopulados.

Truncando a curva via potência nominal

P_nom <- 2500

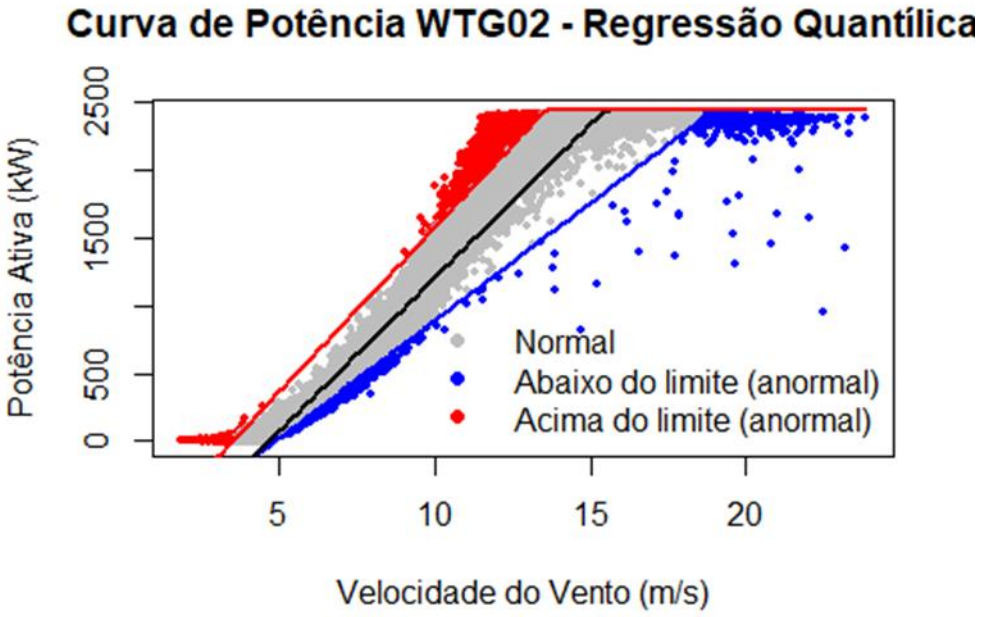
 # --- 3. Ajustar modelos quantílicos ---

modelo_lower <- rq(WTG02_ActivePower ~ WTG02_WindSpeed, tau = 0.1, data = WTG02)

```
modelo_med   <- rq(WTG02_ActivePower ~ WTG02_WindSpeed, tau = 0.50, data = WTG02)

modelo_upper <- rq(WTG02_ActivePower ~ WTG02_WindSpeed, tau = 0.90, data = WTG02)

# --- 4. Grid para curvas suaves ---

grid_wind <- data.frame(

  WTG02_WindSpeed = seq(

    min(WTG02$WTG02_WindSpeed, na.rm = TRUE),

    max(WTG02$WTG02_WindSpeed, na.rm = TRUE),

    length.out = 100))

# --- 5. Previsões no grid ---

grid_wind$pred_lower <- predict(modelo_lower, newdata = grid_wind)

grid_wind$pred_med   <- predict(modelo_med,   newdata = grid_wind)

grid_wind$pred_upper <- predict(modelo_upper, newdata = grid_wind)


# --- 6. Limitar previsões à potência nominal ---

grid_wind <- grid_wind %>%  mutate( pred_lower = pmin(pred_lower, P_nom),

    pred_med   = pmin(pred_med,   P_nom),

    pred_upper = pmin(pred_upper, P_nom) )

# --- 7. Previsões ponto a ponto (para classificação) ---

WTG02$pred_lower <- predict(modelo_lower, newdata = WTG02)

WTG02$pred_upper <- predict(modelo_upper, newdata = WTG02)

# --- 8. Classificar observações ---
```

```
WTG02 <- WTG02 %>%
  mutate( estado = case_when(WTG02_ActivePower < pred_lower ~ "Abaixo do limite (anormal)",
      WTG02_ActivePower > pred_upper ~ "Acima do limite (anormal)",
      TRUE ~ "Normal" ),cor = case_when(
      estado == "Normal" ~ "gray",
      estado == "Abaixo do limite (anormal)" ~ "blue",
      estado== "Acima do limite (anormal)" ~ "red"  ))
# --- 9. Gráfico da curva de potência ---
plot(WTG02$WTG02_WindSpeed, WTG02$WTG02_ActivePower,
  pch = 16, cex = 0.6, col = WTG02$cor,
  xlab = "Velocidade do Vento (m/s)",
  ylab = "Potência Ativa (kW)",
  main = "Curva de Potência WTG02 - Regressão Quantílica")
# --- 10. Linhas dos quantis (limitadas pela potência nominal) ---
lines(grid_wind$WTG02_WindSpeed, grid_wind$pred_lower, col = "blue",  lwd = 2)
lines(grid_wind$WTG02_WindSpeed, grid_wind$pred_med,   col = "black", lwd = 2)
lines(grid_wind$WTG02_WindSpeed, grid_wind$pred_upper, col = "red",   lwd = 2)
# --- 12. Legenda ---
legend(  "bottomright",  legend = c(   "Normal",  "Abaixo do limite (anormal)", "Acima do limite (anormal)"),col = c("gray", "blue", "red", "blue", "black", "red"),
  pch = c(16, 16, 16, NA, NA, NA, NA),
```

```
lwd = c(NA, NA, NA, 2, 2, 2, 2),

lty = c(NA, NA, NA, 1, 1, 1, 2),

bty = "n")
```



Fonte: Software R

ANEXO III

DETECÇÃO DE ESTADO DE CONDICAO E MATRIZ DE CONFUSÃO USANDO OS TRÊS MODELOS

ANEXO 3.1: CLASSIFIÇÃO USANDO MÉTODO BIN TURBINA WTG01

CENÁRIO 1

```
# 1. Definir limites dos bins (ajuste conforme seu dataset)
# ----------------------------------------------------------
bin_limits <- seq(0, 25, by = 1)  # Bins de 1 m/s de velocidade do vento
num_bins <- length(bin_limits) - 1
# ----------------------------------------------------------
# 2. Criar bins e calcular estatísticas por bin (quantis)
# ----------------------------------------------------------
WTG01_bins <- WTG01 %>%
  mutate(bin = cut(WTG01_WindSpeed, breaks = bin_limits, include.lowest = TRUE)) %>%
  group_by(bin) %>%summarise(vento_medio = mean(WTG01_WindSpeed, na.rm = TRUE),
    # Potência
    q_pot_low  = quantile(WTG01_ActivePower, 0.1, na.rm = TRUE),
    q_pot_high = quantile(WTG01_ActivePower, 0.9, na.rm = TRUE),
    # Velocidade do rotor
    q_rot_low  = quantile(WTG01_RotorSpeed, 0.1, na.rm = TRUE),
    q_rot_high = quantile(WTG01_RotorSpeed, 0.9, na.rm = TRUE),
      .groups = "drop")
# ----------------------------------------------------------
# 3. Predição baseada nos intervalos inter-quantis
# ----------------------------------------------------------
WTG01_pred <- WTG01 %>%
```

```
  mutate(bin = cut(WTG01_WindSpeed, breaks = bin_limits, include.lowest = TRUE)) %>%
  left_join(WTG01_bins, by = "bin") %>%  mutate( pred_classe = ifelse(
     WTG01_ActivePower >= q_pot_low & WTG01_ActivePower <= q_pot_high &
     WTG01_RotorSpeed  >= q_rot_low & WTG01_RotorSpeed  <= q_rot_high,  1,   # normal
     0    # anormal   ) )
# ------------------------------------------------------------
# 4. Matriz de Confusão
# ------------------------------------------------------------
conf_mat <- confusionMatrix( factor(WTG01_pred$pred_classe, levels = c(0, 1)),factor(WTG01_pred$estado,      levels = c(0, 1)), positive = "1")
cat("\n=== MATRIZ DE CONFUSÃO
===\n")
print(conf_mat)
                Referência
Predicto       0          1
              0  2864  7532
              1  5263 24985
```

CENÁRIO 2

```
# ------------------------------------------------------------
# 1. Definir limites dos bins (ajuste conforme seu dataset)
```

```
# ----------------------------------------------------------
bin_limits <- seq(0, 25, by = 1)  # Bins de 1 m/s de velocidade do vento
num_bins <- length(bin_limits) - 1
# ----------------------------------------------------------
# 2. Criar bins e calcular estatísticas por bin (quantis)
# ----------------------------------------------------------
WTG01_bins <- WTG01 %>%
  mutate(bin = cut(WTG01_WindSpeed, breaks = bin_limits, include.lowest = TRUE)) %>%
  group_by(bin) %>%summarise(vento_medio = mean(WTG01_WindSpeed, na.rm = TRUE),
    q_low  = quantile(WTG01_ActivePower, 0.1, na.rm = TRUE),   # Quantil 10
    q_high = quantile(WTG01_ActivePower, 0.9, na.rm = TRUE),   # Quantil 90
    .groups = "drop"  )
# ----------------------------------------------------------
# 3. Predição baseada nos intervalos inter-quantis
# ----------------------------------------------------------
WTG01_pred <- WTG01 %>%
  mutate(bin = cut(WTG01_WindSpeed, breaks = bin_limits, include.lowest = TRUE)) %>%
  left_join(WTG01_bins, by = "bin") %>%
  mutate(pred_classe = ifelse( WTG01_ActivePower >= q_low & WTG01_ActivePower <= q_high,
      1,   # normal
      0    # anormal))
```

```
# ----------------------------------------------------------
# 4. Matriz de Confusão
# ----------------------------------------------------------
conf_mat <- confusionMatrix(factor(WTG01_pred$pred_classe, levels = c(0, 1)),
  factor(WTG01_pred$estado,      levels = c(0, 1)),
  positive = "1")
cat("\n=== MATRIZ DE CONFUSÃO ===\n")
print(conf_mat)
```

```
          Referência
Prevista     0        1
         0  2347  5799
         1  5780 26718
```

CENÁRIO 3

```
# ----------------------------------------------------------
# 1. Definir limites dos bins (ajuste conforme seu dataset)
# ----------------------------------------------------------
bin_limits <- seq(0, 25, by = 1)  # Bins de 1 m/s de velocidade do vento
num_bins <- length(bin_limits) - 1
# ----------------------------------------------------------
# 2. Criar bins e calcular estatísticas por bin (quantis)
```

```
# ----------------------------------------------------------

WTG01_bins <- WTG01 %>%

  mutate(bin = cut(WTG01_WindSpeed, breaks = bin_limits, include.lowest = TRUE)) %>%

  group_by(bin) %>% summarise( vento_medio = mean(WTG01_WindSpeed, na.rm = TRUE),

    # Torque

    q_torque_low  = quantile(torque, 0.10, na.rm = TRUE),  # Quantil 10%

    q_torque_high = quantile(torque, 0.90, na.rm = TRUE),  # Quantil 90%  .groups = "drop")

# ----------------------------------------------------------

# 3. Predição baseada nos intervalos inter-quantis

# ----------------------------------------------------------

WTG01_pred <- WTG01 %>%

  mutate(bin = cut(WTG01_WindSpeed, breaks = bin_limits, include.lowest = TRUE)) %>%

  left_join(WTG01_bins, by = "bin") %>%

  mutate( pred_classe = ifelse(torque >= q_torque_low & torque <= q_torque_high,

      1,   # normal

      0    # anormal ))

# ----------------------------------------------------------

# 4. Matriz de Confusão

# ----------------------------------------------------------

conf_mat <- confusionMatrix(factor(WTG01_pred$pred_classe, levels = c(0, 1)),

  factor(WTG01_pred$estado,      levels = c(0, 1)), positive = "1")
```

```
cat("\n=== MATRIZ DE CONFUSÃO - MÉTODO DOS BINS (Torque & WindSpeed) ===\n")
print(conf_mat)
          Referência
Previsto     0       1
         0   894  1370
         1    7233 31144
```

CLASSIFIÇÃO USANDO MÉTODO BIN TURBINA WTG02

CENÁRIO 1

```
# ------------------------------------------------------------
# 1. Definir limites dos bins (ajuste conforme seu dataset)
# ------------------------------------------------------------
bin_limits <- seq(0, 25, by = 1)  # Bins de 1 m/s de velocidade do vento
num_bins <- length(bin_limits) - 1
# ------------------------------------------------------------
# 2. Criar bins e calcular estatísticas por bin (quantis)
# ------------------------------------------------------------
WTG02_bins <- WTG02 %>%
  mutate(bin = cut(WTG02_WindSpeed, breaks = bin_limits, include.lowest = TRUE)) %>%
  group_by(bin) %>% summarise(vento_medio = mean(WTG02_WindSpeed, na.rm = TRUE),
    # Potência
```

```
    q_pot_low  = quantile(WTG02_ActivePower, 0.1, na.rm = TRUE),
    q_pot_high = quantile(WTG02_ActivePower, 0.9, na.rm = TRUE),
    # Velocidade do rotor
    q_rot_low  = quantile(WTG02_RotorSpeed, 0.1, na.rm = TRUE),
    q_rot_high = quantile(WTG02_RotorSpeed, 0.9, na.rm = TRUE), .groups = "drop")
# ----------------------------------------------------------
# 3. Predição baseada nos intervalos inter-quantis
# ----------------------------------------------------------
WTG02_pred <- WTG02 %>%
  mutate(bin = cut(WTG02_WindSpeed, breaks = bin_limits, include.lowest = TRUE)) %>%
  left_join(WTG02_bins, by = "bin") %>%
  mutate( pred_classe = ifelse(WTG02_ActivePower >= q_pot_low & WTG02_ActivePower <=
q_pot_high &WTG02_RotorSpeed  >= q_rot_low & WTG02_RotorSpeed  <= q_rot_high,
      1,   # normal
      0    # anormal))
# ----------------------------------------------------------
# 4. Matriz de Confusão
# ----------------------------------------------------------
conf_mat <- confusionMatrix(
  factor(WTG02_pred$pred_classe, levels = c(0, 1)),
  factor(WTG02_pred$estado,      levels = c(0, 1)),
```

```
  positive = "1")

cat("\n=== MATRIZ DE CONFUSÃO - ===\n")

print(conf_mat)

               Referência

Previsto     0       1

         0  1227  2775

         1  6741 29093
```

CENÁRIO 2

```
# ------------------------------------------------------------

# 1. Definir limites dos bins (ajuste conforme seu dataset)

# ------------------------------------------------------------

bin_limits <- seq(0, 25, by = 1)  # Bins de 1 m/s de velocidade do vento

num_bins <- length(bin_limits) - 1



# ------------------------------------------------------------

# 2. Criar bins e calcular estatísticas por bin (quantis)

# ------------------------------------------------------------

WTG02_bins <- WTG02 %>%

  mutate(bin = cut(WTG02_WindSpeed, breaks = bin_limits, include.lowest = TRUE)) %>%

  group_by(bin) %>%summarise(vento_medio = mean(WTG02_WindSpeed, na.rm = TRUE),
```

```
    q_low  = quantile(WTG02_ActivePower, 0.1, na.rm = TRUE),
    q_high = quantile(WTG02_ActivePower, 0.9, na.rm = TRUE),
    .groups = "drop" )
# ---------------------------------------------------------
# 3. Predição baseada nos intervalos inter-quantis
# ---------------------------------------------------------
WTG02_pred <- WTG02 %>%
  mutate(bin = cut(WTG02_WindSpeed, breaks = bin_limits, include.lowest = TRUE)) %>%
  left_join(WTG02_bins, by = "bin") %>%
  mutate(
    pred_classe = ifelse(
      WTG02_ActivePower >= q_low & WTG02_ActivePower <= q_high,
      1,   # normal
      0    # anormal ) )
# ---------------------------------------------------------
# 4. Matriz de Confusão
# ---------------------------------------------------------
conf_mat <- confusionMatrix(factor(WTG02_pred$pred_classe, levels = c(0, 1)),
  factor(WTG02_pred$estado,      levels = c(0, 1)),
  positive = "1")
cat("\n=== MATRIZ DE CONFUSÃO ===\n")
```

```
print(conf_mat)

          Referência

Previsto     0         1

           0  2137  5841

           1  5831 26027
```

CENÁRIO 3

```
# ----------------------------------------------------------

# 1. Definir limites dos bins (ajuste conforme seu dataset)

# ----------------------------------------------------------

bin_limits <- seq(0, 25, by = 1)  # Bins de 1 m/s de velocidade do vento

num_bins <- length(bin_limits) - 1

# ----------------------------------------------------------

# 2. Criar bins e calcular estatísticas por bin (quantis)

# ----------------------------------------------------------

WTG02_bins <- WTG02 %>%

  mutate(bin = cut(WTG02_WindSpeed, breaks = bin_limits, include.lowest = TRUE)) %>%

  group_by(bin) %>% summarise(vento_medio = mean(WTG02_WindSpeed, na.rm = TRUE),

    # Torque

    q_torque_low  = quantile(torque, 0.10, na.rm = TRUE),  # Quantil 10%

    q_torque_high = quantile(torque, 0.90, na.rm = TRUE),  # Quantil 90% .groups = "drop")
```

```
# ----------------------------------------------------------

# 3. Predição baseada nos intervalos inter-quantis

# ----------------------------------------------------------

WTG02_pred <- WTG02 %>%mutate(bin = cut(WTG02_WindSpeed, breaks = bin_limits, include.lowest = TRUE)) %>%

  left_join(WTG02_bins, by = "bin") %>% mutate(   pred_classe = ifelse( torque >= q_torque_low & torque <= q_torque_high,

      1,   # normal

      0    # anormal ))

# ----------------------------------------------------------

# 4. Matriz de Confusão

# ----------------------------------------------------------

conf_mat <- confusionMatrix(factor(WTG02_pred$pred_classe, levels = c(0, 1)),

  factor(WTG02_pred$estado,      levels = c(0, 1)),  positive = "1")

cat("\n=== MATRIZ DE CONFUSÃO ===\n")

print(conf_mat)

              Referência

Previsto     0        1

          0  2309  5669

          1    5661 26197
```

ANEXO 3.2: CLASSIFIÇÃO USANDO MÉTODO DISTR. NORMAL MULTIVARIADA TURBINA WTG01

CENÁRIO 1

```
# ----------------------------
# Script: Detecção de Falhas WTG01
# Método: Distribuição Normal Multivariada
# ----------------------------
# 1. Pacotes utilizados
# Funções estatísticas
library(MASS)     # mvrnorm, dmvnorm
# Matriz de confusão e métricas
library(caret)    # confusionMatrix
# Manipulação de dataframes
library(dplyr)
# Cálculo de densidade multivariada
library(mvtnorm)
# ----------------------------
# 1. Selecção das variáveis numéricas e rótulo
# ----------------------------
```

```
# Seleciona variáveis numéricas relevantes

variaveis <- WTG01[, c("WTG01_ActivePower", "WTG01_WindSpeed","WTG01_RotorSpeed")]

# Define a coluna de classe (0=anormal, 1=normal)

classe <- factor(WTG01$estado, levels = c(0, 1))

# Cria o dataframe final para modelagem

dados <- data.frame(variaveis, classe)

# ----------------------------

# 2. Modelo Normal Multivariado

# ----------------------------

# Treino apenas com dados normais

# Calcula média e covariância das variáveis normais

mu <- colMeans(dados[dados$classe == 1, 1:3], na.rm = TRUE)

Sigma <- cov(dados[dados$classe == 1, 1:3], use = "complete.obs")

# Cálculo  densidade multivariada para cada observação

dens <- dmvnorm(dados[,1:3], mean = mu, sigma = Sigma)

# ----------------------------

# 3. Definir limiar para classificar falhas

# ----------------------------

# Threshold = percentil da densidade dos dados normais (10%)

limiar <- quantile(dens[dados$classe == 1], probs = 0.1, na.rm = TRUE)

#Classifica observações como falha (0) ou normal (1)
```

```
dados$previsto <- ifelse(dens < limiar, 0, 1)

dados$previsto <- factor(dados$previsto, levels = c(0, 1))

# ----------------------------

# 4. Matriz de confusão

# ----------------------------

# Compara valores reais x previstos

matriz <- confusionMatrix( dados$classe,dados$previsto, positive = "1")

# Mostra VN, FP, FN, VP

print(matriz)

Confusion Matrix and Statistics

                Referência

Previsto          0              1

             0  (VN)3373  (FN)4754

             1    (FP)3252  (VP)29265
```

CENÁRIO 2

```
# ----------------------------

# Script: Detecção de Falhas WTG01

# Método: Distribuição Normal Multivariada

# ----------------------------

# Pacotes
```

```
library(MASS)     # mvrnorm, dmvnorm

library(caret)    # confusionMatrix

library(dplyr)

# install.packages("mvtnorm")

library(mvtnorm)

# 1. Seleção das variáveis numéricas e rótulo

# ----------------------------

# Supondo que já exista a coluna 'estado' com classes "Normal" / "Anormal"

variaveis <- WTG01[, c("WTG01_ActivePower", "WTG01_WindSpeed")]

classe <- factor(WTG01$estado, levels = c(0, 1))

dados <- data.frame(variaveis, classe)

# ----------------------------

# 2. Modelo Normal Multivariado

# ----------------------------

# Treino apenas com dados normais

mu <- colMeans(dados[dados$classe == 1, 1:2], na.rm = TRUE)

Sigma <- cov(dados[dados$classe == 1, 1:2], use = "complete.obs")

# Cálculo da densidade para todos os pontos

dens <- dmvnorm(dados[,1:2], mean = mu, sigma = Sigma)

# ----------------------------

# 3. Definir limiar para classificar falhas
```

```
# ----------------------------

# Threshold = percentil da densidade dos dados normais (10%)

limiar <- quantile(dens[dados$classe == 1], probs = 0.1, na.rm = TRUE)

dados$previsto <- ifelse(dens < limiar, 0, 1)

dados$previsto <- factor(dados$previsto, levels = c(0, 1))

# ----------------------------

# 4. Matriz de confusão

# ----------------------------

matriz <- confusionMatrix( dados$classe,dados$previsto, positive = "1")

print(matriz)
```

```
          Referência

Previsto     0      1

       0  3966  4161

       1  3252 29265
```

CENÁRIO 3

```
# ----------------------------

# Script: Detecção de Falhas WTG01

# Método: Distribuição Normal Multivariada

# ----------------------------
```

```
# Pacotes
library(MASS)     # mvrnorm, dmvnorm
library(caret)    # confusionMatrix
library(dplyr)
# install.packages("mvtnorm")
library(mvtnorm)
# ----------------------------
# Script: Detecção de Falhas WTG01
# Método: Distribuição Normal Multivariada
# ----------------------------
# Pacotes
library(MASS)     # mvrnorm, dmvnorm
library(caret)    # confusionMatrix
library(dplyr)
# install.packages("mvtnorm")
library(mvtnorm)
# ----------------------------
# 1. Seleção das variáveis numéricas e rótulo
# ----------------------------
# Supondo que já exista a coluna 'estado' com classes "Normal" / "anormal"
variaveis <- WTG01[, c("torque", "WTG01_WindSpeed")]
```

```
classe <- factor(WTG01$estado, levels = c(0, 1))

dados <- data.frame(variaveis, classe)

# ----------------------------

# 2. Modelo Normal Multivariado

# ----------------------------

# Treino apenas com dados normais

mu <- colMeans(dados[dados$classe == 1, 1:2], na.rm = TRUE)

Sigma <- cov(dados[dados$classe == 1, 1:2], use = "complete.obs")

# Cálculo da densidade para todos os pontos

dens <- dmvnorm(dados[,1:2], mean = mu, sigma = Sigma)

# ----------------------------

# 3. Definir limiar para classificar falhas

# ----------------------------

# Threshold = percentil da densidade dos dados normais (10%)

limiar <- quantile(dens[dados$classe == 1], probs = 0.1, na.rm = TRUE)

dados$previsto <- ifelse(dens < limiar, 0, 1)

dados$previsto <- factor(dados$previsto, levels = c(0, 1))

# ----------------------------

# 4. Matriz de confusão

# ----------------------------

matriz <- confusionMatrix( dados$classe,dados$previsto, positive = "0")
```

```
print(matriz)

Confusion Matrix and Statistics


          Referência
Prevista     0       1
        0  3895  4235
        1  3252 29262
```

ANEXO 3.3: CLASSIFIÇÃO USANDO MÉTODO DISTR. NORMAL MULTIVARIADA, TURBINA WTG01:

VARIANTE "*ONE-STEP-AHEAD*"

CENÁRIO 1

```
#------------------------------------------------------------

# MÉTODO NORMAL MULTIVARIADO – ONE-STEP-AHEAD

# Modelagem de ActivePower, WindSpeed e RotorSpeed

#------------------------------------------------------------

# Pacotes necessários

library(MASS)    # Para mvrnorm e funções MVN

library(caret)   # Para matriz de confusão

library(dplyr)
```

```
library(mvtnorm)

#-------------------------------------------------------------

# 1. Preparar os dados

#-------------------------------------------------------------

# Seleciona apenas colunas relevantes

dados <- WTG01 %>% select(WTG01_ActivePower, WTG01_WindSpeed, WTG01_RotorSpeed, estado) %>%  na.omit()

Renomeia colunas para facilitar a manipulação

colnames(dados) <- c("ActivePower", "WindSpeed", "RotorSpeed", "estado")

#-------------------------------------------------------------

# 2. Estimar parâmetros iniciais

#-------------------------------------------------------------

# Seleciona apenas dados normais

dados_normais <- dados %>% filter(estado == 1)

# Calcula média e covariância das três variáveis

mu <- colMeans(dados_normais[, 1:3])

sigma <- cov(dados_normais[, 1:3])

#-------------------------------------------------------------

# 3. Modelagem em actualização temporal

#-------------------------------------------------------------

n <- nrow(dados)
```

```
predito <- numeric(n)

for (i in 2:n) {

  # Histórico até a observação anterior (para simular tempo real)

  hist <- dados[1:(i - 1), ]

  hist_normais <- hist %>% filter(estado == 1)

  # Reestimar parâmetros (atualização adaptativa)

  if (nrow(hist_normais) > 10) {

    mu <- colMeans(hist_normais[, 1:3])

    sigma <- cov(hist_normais[, 1:3])  }

  # Cálculo da densidade MVN para o ponto atual

  x_i <- as.numeric(dados[i, 1:3])

  densidade <- dmvnorm(x_i, mean = mu, sigma = sigma)

  # Threshold (pode ser ajustado)

  # Usando o 10º percentil das densidades normais observadas como limite

  if (i == 2) {limite <- quantile(dmvnorm(as.matrix(hist_normais[, 1:3]), mean = mu, sigma = sigma), 0.1) }

  predito[i] <- ifelse(densidade < limite, 0, 1)}

#-------------------------------------------------------------

# 4. Matriz de confusão

#-------------------------------------------------------------

verdadeiro <- dados$estado[2:n]
```

```
previsto <- predito[2:n]

matriz_confusao <- confusionMatrix( factor(previsto, levels = c(0, 1)),factor(verdadeiro, levels = c(0, 1)), positive = "1")

print(matriz_confusao)

                Referência

Prevista       0        1

         0    5763    12025

         1    2364     20491
```

CENÁRIO 2

```
#------------------------------------------------------------

# MÉTODO NORMAL MULTIVARIADO – one-step-ahead

# Modelagem de ActivePower, WindSpeed e RotorSpeed

#------------------------------------------------------------

# Pacotes necessários

library(MASS)    # Para mvrnorm e funções MVN

library(caret)   # Para matriz de confusão

library(dplyr)

library(mvtnorm)

#------------------------------------------------------------

# 1. Preparar os dados
```

```
#-----------------------------------------------------------

# O dataframe deve conter: ActivePower, WindSpeed,  estado


dados <- WTG01 %>%

  select(WTG01_ActivePower, WTG01_WindSpeed, estado) %>%

  na.omit()

colnames(dados) <- c("ActivePower", "WindSpeed","estado")

#-----------------------------------------------------------

# 2. Separar dados normais para estimar parâmetros

#-----------------------------------------------------------

dados_normais <- dados %>% filter(estado == 1)

mu <- colMeans(dados_normais[, 1:2])

sigma <- cov(dados_normais[, 1:2])

#-----------------------------------------------------------

# 3. Modelagem um passo à frente

#-----------------------------------------------------------

n <- nrow(dados)

predito <- numeric(n)

for (i in 2:n) { # Histórico até a observação anterior (para simular tempo real)

  hist <- dados[1:(i - 1), ]

  hist_normais <- hist %>% filter(estado == 1)
```

```
  # Reestimar parâmetros (atualização adaptativa)
  if (nrow(hist_normais) > 10) {
    mu <- colMeans(hist_normais[, 1:2])
    sigma <- cov(hist_normais[, 1:2])}
  # Cálculo da densidade MVN para o ponto atual
  x_i <- as.numeric(dados[i, 1:2])
  densidade <- dmvnorm(x_i, mean = mu, sigma = sigma)
  # Threshold (pode ser ajustado)
  # Usando o 10º percentil das densidades normais observadas como limite
  if (i == 2) { limite <- quantile(dmvnorm(as.matrix(hist_normais[, 1:2]), mean = mu, sigma = sigma), 0.1)}
  predito[i] <- ifelse(densidade < limite, 0, 1)}
#------------------------------------------------------------
# 4. Matriz de confusão
#------------------------------------------------------------
verdadeiro <- dados$estado[2:n]
previsto <- predito[2:n]
matriz_confusao <- confusionMatrix( factor(previsto, levels = c(0, 1)),factor(verdadeiro, levels = c(0, 1)), positive = "1")
print(matriz_confusao)
                Referência
  Prevista     0         1
```

```
  0  6020   9079
  1  2107 23437
```

CENÁRIO 3

```
# Variáveis: Torque e WindSpeed
#----------------------------------------------------------
# Pacotes necessários
library(MASS)      # mvrnorm e funções estatísticas
library(mvtnorm)   # dmvnorm() para densidade normal multivariada
library(caret)     # matriz de confusão
library(dplyr)     # manipulação de dados
#----------------------------------------------------------
# 1. Preparação dos dados
#----------------------------------------------------------
# Ele deve conter colunas: Torque, WindSpeed e estado (0 ou 1)
dados <- WTG01 %>%
  select(torque, WTG01_WindSpeed, estado) %>%
  rename(Torque = torque, WindSpeed = WTG01_WindSpeed) %>%
  na.omit()
#----------------------------------------------------------
# 2. Inicialização com dados normais (classe = 1)
```

```
#-----------------------------------------------------------
dados_normais <- dados %>% filter(estado == 1)
# Média e covariância iniciais das variáveis sob regime normal
mu <- colMeans(dados_normais[, c("Torque", "WindSpeed")])
sigma <- cov(dados_normais[, c("Torque", "WindSpeed")])
#-----------------------------------------------------------
# 3. Loop “one-step-ahead”
#-----------------------------------------------------------
n <- nrow(dados)
predito <- numeric(n)

for (i in 2:n) {
  # Histórico até o ponto anterior
  hist <- dados[1:(i - 1), ]
  hist_normais <- hist %>% filter(estado == 1)
  # Atualiza parâmetros se houver dados suficientes
  if (nrow(hist_normais) > 20) {
    mu <- colMeans(hist_normais[, c("Torque", "WindSpeed")])
    sigma <- cov(hist_normais[, c("Torque", "WindSpeed")])}
  # Ponto atual
  x_i <- as.numeric(dados[i, c("Torque", "WindSpeed")])
```

```
  # Densidade de probabilidade segundo o modelo normal multivariado
  densidade <- dmvnorm(x_i, mean = mu, sigma = sigma)
  # Define o limiar como o 10º percentil das densidades normais históricas
  if (i == 2) {
    ref_dens <- dmvnorm(as.matrix(hist_normais[, c("Torque", "WindSpeed")]),mean = mu, sigma = sigma)
    limite <- quantile(ref_dens, 0.1) }
  # Classificação: 1 = normal, 0 = anômalo
  predito[i] <- ifelse(densidade < limite, 0, 1)}
#------------------------------------------------------------
# 4. Avaliação - Matriz de confusão
#------------------------------------------------------------
verdadeiro <- dados$estado[2:n]
previsto <- predito[2:n]
matriz_confusao <- confusionMatrix(
  factor(previsto, levels = c(0, 1)),
  factor(verdadeiro, levels = c(0, 1)),
  positive = "1")
#------------------------------------------------------------
# 5. Resultados
#------------------------------------------------------------
```

```
cat("\n===== MÉTODO NORMAL MULTIVARIADO ====\n")

print(matriz_confusao)

                Referência

Prevista     0         1

          0  5699  9501

          1 2431 23012
```

ANEXO 3.4: CLASSIFIÇÃO USANDO MÉTODO DISTR. NORMAL MULTIVARIADA TURBINA WTG02

Como o procedimento é mesmo da turbina 1, para a turbina 2 são apresentadas as matrizes de confusão correspondentes a cada cenário.

Tabela de matrizes de confusão, método de disribuicao normal multivariado turbina 2

Cenário	Modelo estático	Modelo one-step-ahead
1	Referência Prevista 0 1 0 3616 4352 1 3187 28681	Referência Prevista 0 1 0 4566 6590 1 3402 25277
2	Referência Prevista 0 1 0 4068 3900 1 3187 28681	Referência Prevista 0 1 0 4630 6643 1 3338 25224
3	Referência Prevista 0 1 0 4158 3812 1 3187 28679	Referência Prevista 0 1 0 4797 7054 1 3173 24811

Fonte: autoria própria

ANEXO 3. 5: CLASSIFIÇÃO USANDO MÉTODO DE CÓPULA, TURBINA WTG02

CENÁRIO 1

```
# Instalar pacotes se necessário

# install.packages(c("copula", "fitdistrplus", "caret"))

library(copula)

library(fitdistrplus)

library(caret)

set.seed(123)  # para reprodutibilidade

# Selecionar variáveis relevantes

# -------------------

vars <- c("WTG02_ActivePower", "WTG02_WindSpeed", "WTG02_RotorSpeed")

if (!all(vars %in% names(WTG02))) {

  stop("O dataframe WTG01 precisa ter as colunas: ", paste(vars, collapse = ", "), " e a coluna 'estado'")}

X <- WTG02[, vars]

X_norm <- X[WTG02$estado == 1, ]

# -------------------

# Ajuste das marginais

# -------------------

marginais <- list()

u_data <- matrix(NA, nrow = nrow(X_norm), ncol = length(vars))
```

 # 1. Potência - Normal

```
fit_pot <- fitdist(X_norm$WTG02_ActivePower, "norm")
marginais[["potencia"]] <- fit_pot
u_data[, 1] <- pnorm(X_norm$WTG02_ActivePower,
                     mean = fit_pot$estimate["mean"],
                     sd   = fit_pot$estimate["sd"])
# 2. Velocidade do vento -normal
fit_wind <- fitdist(X_norm$WTG02_WindSpeed, "norm")
marginais[["wind"]] <- fit_wind
u_data[, 2] <- pnorm(X_norm$WTG02_WindSpeed,
                     mean = fit_wind$estimate["mean"],
                     sd   = fit_wind$estimate["sd"])
# 3. Velocidade do rotor - Normal
fit_rotor <- fitdist(X_norm$WTG02_RotorSpeed, "norm")
marginais[["rotor"]] <- fit_rotor
u_data[, 3] <- pnorm(X_norm$WTG02_RotorSpeed,
                     mean = fit_rotor$estimate["mean"],
                     sd   = fit_rotor$estimate["sd"])
# Truncar valores extremos (0 ou 1)
u_data <- pmin(pmax(u_data, 1e-6), 1 - 1e-6)
# ------------------
# Função segura para ajustar cópula
```

```
# -------------------

safe_fit_copula <- function(cop, data) {

  fit <- try(fitCopula(cop, data = data, method = "itau"), silent = TRUE)

  if (inherits(fit, "try-error")) {

    message("Método itau falhou, tentando ML...")

    fit <- fitCopula(cop, data = data, method = "ml")

  } else { message("Ajuste bem sucedido com método itau.")  } return(fit)}
```

 # Ajustar cópulas

```
cop_gauss <- normalCopula(dim = 3, dispstr = "un")

fit_gauss <- safe_fit_copula(cop_gauss, u_data)

cop_t <- tCopula(dim = 3, dispstr = "un")

fit_t <- safe_fit_copula(cop_t, u_data)

# -------------------
```

 # Função para calcular log-verossimilhança

```
# -------------------

loglik_obs <- function(obs, fit_copula, marginais) {

  u <- c( pnorm(obs[1], mean = marginais$potencia$estimate["mean"],

                sd   = marginais$potencia$estimate["sd"]),

    pnorm(obs[2], rate = marginais$vento$estimate["rate"]),

    pnorm(obs[3], mean = marginais$rotor$estimate["mean"],

                sd   = marginais$rotor$estimate["sd"])  )
```

```
  u <- pmin(pmax(u, 1e-6), 1 - 1e-6)  # truncar extremos
  dCopula(u, fit_copula@copula, log = TRUE}
# -------------------
# Classificação
# -------------------
scores_gauss <- apply(X, 1, loglik_obs, fit_copula = fit_gauss, marginais = marginais)
scores_t     <- apply(X, 1, loglik_obs, fit_copula = fit_t,     marginais = marginais)
# Limiar: 10% piores casos normais
limiar_gauss <- quantile(scores_gauss[WTG02$estado == 1], 0.1, na.rm = TRUE)
limiar_t     <- quantile(scores_t[WTG02$estado== 1], 0.1, na.rm = TRUE)
pred_gauss <- ifelse(scores_gauss >= limiar_gauss, 1, 0)
pred_t     <- ifelse(scores_t >= limiar_t, 1, 0)
# -------------------
# Matrizes de confusão
# -------------------
WTG02$estado <- factor(WTG02$estado, levels = c(0,1))
pred_gauss <- factor(pred_gauss, levels = c(0,1))
pred_t     <- factor(pred_t, levels = c(0,1))

=== Copula Gaussiana ===
 print(confusionMatrix(pred_gauss, WTG02$estado))
```

Referência

Previsto 0 1

0 1125 3187

1 6843 28681

=== Copula t ===

print(confusionMatrix(pred_t, WTG02$estado))

Referência

Previsto 0 1

0 1083 3187

1 6885 28681

CENÁRIO 2

```
library(copula)
library(caret)
library(dplyr)
library(fitdistrplus)
set.seed(123)
```

```
vars <- c("WTG02_ActivePower", "WTG02_WindSpeed")

if (!all(vars %in% names(WTG02))) {

  stop("O dataframe WTG02 precisa ter as colunas: ", paste(vars, collapse = ", "), " e a coluna 'estado'")}

X <- WTG02[, vars]

X_norm <- X[WTG02$estado== 1, ]  # dados normais

# -------------------

# Ajuste das marginais (normais)

# -------------------

marginais <- list()

u_data <- matrix(NA, nrow = nrow(X_norm), ncol = length(vars))

# 1. Potência - Normal

fit_potencia <- fitdist(X_norm$WTG02_ActivePower, "norm")

marginais[["WTG02_ActivePower"]] <- fit_potencia

u_data[, 1] <- pnorm(X_norm$WTG02_ActivePower,

                     mean = fit_potencia$estimate["mean"],

                     sd   = fit_potencia$estimate["sd"])

# 2. WindSpeed - Normal

fit_wind <- fitdist(X_norm$WTG02_WindSpeed, "norm")

marginais[["wind"]] <- fit_wind

u_data[, 2] <- pnorm(X_norm$WTG02_WindSpeed,

                     mean = fit_wind$estimate["mean"],
```

```
                sd   = fit_wind$estimate["sd"])
```

 # Truncar valores extremos

```
u_data <- pmin(pmax(u_data, 1e-6), 1 - 1e-6)

# -------------------

# Função segura para ajustar cópula

# -------------------

safe_fit_copula <- function(cop, data) {

  fit <- try(fitCopula(cop, data = data, method = "itau"), silent = TRUE)

  if (inherits(fit, "try-error")) {

    message("Método itau falhou, tentando ML...")

    fit <- fitCopula(cop, data = data, method = "ml")

  } else {   message("Ajuste bem sucedido com método itau.") }  return(fit)}
```

 # Ajuste das cópulas

```
cop_gauss <- normalCopula(dim = 2, dispstr = "un")

fit_gauss <- safe_fit_copula(cop_gauss, u_data)

cop_t <- tCopula(dim = 2, dispstr = "un")

fit_t <- safe_fit_copula(cop_t, u_data)

# -------------------
```

 # Função log-verossimilhança

```
# -------------------

loglik_obs <- function(obs, fit_copula, marginais) {
```

```
  u <- c( pnorm(obs[1], mean = marginais$WTG02_ActivePower$estimate["mean"],
                sd   = marginais$WTG02_ActivePower$estimate["sd"]),
    pnorm(obs[2], mean = marginais$wind$estimate["mean"],
                sd   = marginais$wind$estimate["sd"])  )
  u <- pmin(pmax(u, 1e-6), 1 - 1e-6)
  dCopula(u, fit_copula@copula, log = TRUE)}
# -------------------
```

 # Classificação

```
# -------------------
scores_gauss <- apply(X, 1, loglik_obs, fit_copula = fit_gauss, marginais = marginais)
scores_t     <- apply(X, 1, loglik_obs, fit_copula = fit_t,     marginais = marginais)

# Limiar: 10% piores casos normais
limiar_gauss <- quantile(scores_gauss[WTG02$estado == 1], 0.1, na.rm = TRUE)
limiar_t     <- quantile(scores_t[WTG02$estado== 1], 0.1, na.rm = TRUE)
pred_gauss <- ifelse(scores_gauss >= limiar_gauss, 1, 0)
pred_t     <- ifelse(scores_t >= limiar_t, 1, 0)
# -------------------
# Matrizes de confusão
# -------------------
WTG02$estado <- factor(WTG02$estado, levels = c(0,1))
```

```
pred_gauss <- factor(pred_gauss, levels = c(0,1))
pred_t     <- factor(pred_t, levels = c(0,1))
cat("\n=== Copula Gaussiana ===\n")
print(confusionMatrix(pred_t, WTG02$estado))
```

```
          Referência
Previsto    0       1
        0  5013  3187
        1  2955 28681
```

=== Copula t ===

```
print(confusionMatrix(pred_t, WTG02$estado))
```

```
          Referência
Previsto     0      1
        0     5043  3187
        1    2925 28681
```

CENÁRIO 3

```
library(copula)

library(caret)

library(dplyr)

library(fitdistrplus)

set.seed(123)

vars <- c("torque", "WTG02_WindSpeed")

if (!all(vars %in% names(WTG02))) {

  stop("O dataframe WTG02 precisa ter as colunas: ", paste(vars, collapse = ", "), " e a coluna 'estado'")}

X <- WTG02[, vars]

X_norm <- X[WTG02$estado== 1, ]  # dados normais

# -------------------

# Ajuste das marginais (normais)

# -------------------

marginais <- list()

u_data <- matrix(NA, nrow = nrow(X_norm), ncol = length(vars))


# 1. Torque - Normal

fit_torque <- fitdist(X_norm$torque, "norm")

marginais[["torque"]] <- fit_torque

u_data[, 1] <- pnorm(X_norm$torque,

                     mean = fit_torque$estimate["mean"],
```

```
                sd   = fit_torque$estimate["sd"])
# 2. WindSpeed - Normal
fit_wind <- fitdist(X_norm$WTG02_WindSpeed, "norm")
marginais[["wind"]] <- fit_wind
u_data[, 2] <- pnorm(X_norm$WTG02_WindSpeed,
                mean = fit_wind$estimate["mean"],
                sd   = fit_wind$estimate["sd"])
# Truncar valores extremos
u_data <- pmin(pmax(u_data, 1e-6), 1 - 1e-6)


# -------------------
# Função segura para ajustar cópula
# -------------------
safe_fit_copula <- function(cop, data) {
  fit <- try(fitCopula(cop, data = data, method = "itau"), silent = TRUE)
  if (inherits(fit, "try-error")) {message("Método itau falhou, tentando ML...")
    fit <- fitCopula(cop, data = data, method = "ml")
  } else {message("Ajuste bem sucedido com método itau.")} return(fit)}


# Ajuste das cópulas
cop_gauss <- normalCopula(dim = 2, dispstr = "un")
```

```
fit_gauss <- safe_fit_copula(cop_gauss, u_data)

cop_t <- tCopula(dim = 2, dispstr = "un")

fit_t <- safe_fit_copula(cop_t, u_data)

# -------------------

# Função log-verossimilhança

# -------------------

loglik_obs <- function(obs, fit_copula, marginais) {

  u <- c(pnorm(obs[1], mean = marginais$torque$estimate["mean"],

               sd   = marginais$torque$estimate["sd"]),

    pnorm(obs[2], mean = marginais$wind$estimate["mean"],

               sd   = marginais$wind$estimate["sd"]) )

  u <- pmin(pmax(u, 1e-6), 1 - 1e-6)

  dCopula(u, fit_copula@copula, log = TRUE)}

# -------------------

# Classificação

# -------------------

scores_gauss <- apply(X, 1, loglik_obs, fit_copula = fit_gauss, marginais = marginais)

scores_t     <- apply(X, 1, loglik_obs, fit_copula = fit_t,     marginais = marginais)

# Limiar: 10% piores casos normais

limiar_gauss <- quantile(scores_gauss[WTG02$estado == 1], 0.1, na.rm = TRUE)

limiar_t     <- quantile(scores_t[WTG02$estado== 1], 0.1, na.rm = TRUE)
```

```
pred_gauss <- ifelse(scores_gauss >= limiar_gauss, 1, 0)

pred_t     <- ifelse(scores_t >= limiar_t, 1, 0)

# -------------------
```

 # Matrizes de confusão

```
# -------------------

WTG02$estado <- factor(WTG02$estado, levels = c(0,1))

pred_gauss <- factor(pred_gauss, levels = c(0,1))

pred_t     <- factor(pred_t, levels = c(0,1)

=== Copula Gaussiana ===

 print(confusionMatrix(pred_gauss, WTG02$estado))


          Referência

Previsto     0     1

        0  4416  3187

        1  3554 28679

 === Copula t ===

> print(confusionMatrix(pred_t, WTG02$estado))


          Referência

Previsto     0     1
```

0 4428 3187

1 3542 28679

ANEXO 3.6: CLASSIFIÇÃO USANDO MÉTODO DE CÓPULAS, *ONE-STEP-AHEAD*, TURBINA WTG02

CENÁRIO 1

```
# ============================================================

# MATRIZ DE CONFUSÃO – CÓPULAS GAUSSIANA E t

# MÉTODO: MARGINAIS NORMAIS | VARIANTE: UM PASSO À FRENTE

# ============================================================


# Pacotes necessários

install.packages(c("copula", "MASS", "fitdistrplus", "caret", "dplyr"))

library(copula)

library(MASS)

library(fitdistrplus)

library(caret)

library(dplyr)

# ------------------------------------------------------------

# 1. Seleção das variáveis e filtragem dos dados
```

```
# ----------------------------------------------------------

vars <- c("WTG02_ActivePower", "WTG02_WindSpeed","WTG02_RotorSpeed")

# Verificar se as colunas existem

if (!all(vars %in% names(WTG02))) {

  stop("O dataframe WTG02 precisa conter as colunas: ", paste(vars, collapse = ", "), " e 'estado'")}

# Remover NA

WTG02 <- na.omit(WTG02[, c(vars, "estado")])

# Separar dados normais (estado 1)

dados_normais <- WTG02 %>% filter(estado == 1)

# ----------------------------------------------------------

# 2. Ajuste das marginais normais

# ----------------------------------------------------------

# Estimar média e desvio padrão

param_marginais <- lapply(dados_normais[vars], function(x)

  list(mean = mean(x), sd = sd(x)))

# Converter dados normais para U[0,1] via CDF normal

U_norm <- as.data.frame(mapply(function(x, p)

  pnorm(x, mean = p$mean, sd = p$sd), dados_normais[vars], param_marginais))

# ----------------------------------------------------------

# 3. Ajuste das cópulas (Gaussiana e t)

# ----------------------------------------------------------
```

```
cop_gauss <- normalCopula(dim = length(vars))

cop_t     <- tCopula(dim = length(vars))

fit_gauss <- fitCopula(cop_gauss, data = as.matrix(U_norm), method = "itau")

fit_t     <- fitCopula(cop_t,     data = as.matrix(U_norm), method = "itau")

# ----------------------------------------------------------

# 4. VARIANTE “UM PASSO À FRENTE” – PREVISÃO

# ----------------------------------------------------------

n <- nrow(WTG02)

pred_gauss <- numeric(n)

pred_t     <- numeric(n)

for (i in 2:n) { if (WTG02$estado[i-1] == 1 && all(!is.na(WTG02[i-1, vars]))) {

   obs <- as.numeric(WTG01[i, vars])

   # Converter observação para U via marginais normais

   U_obs <- mapply(function(x, p) pnorm(x, mean = p$mean, sd = p$sd),

                   as.list(obs), param_marginais)


   # Densidades sob cada cópula

   dens_gauss <- dCopula(U_obs, fit_gauss@copula)

   dens_t     <- dCopula(U_obs, fit_t@copula)

   # Classificação: densidade < limiar => anômalo (0)

   limiar <- 0.1
```

```
    pred_gauss[i] <- ifelse(dens_gauss < limiar, 0, 1)
    pred_t[i]     <- ifelse(dens_t < limiar, 0, 1)
  } else {pred_gauss[i] <- NA
    pred_t[i]     <- NA}}
# ----------------------------------------------------------
# 5. MATRIZES DE CONFUSÃO
# ----------------------------------------------------------
validos_gauss <- which(!is.na(pred_gauss))
validos_t     <- which(!is.na(pred_t))

verdadeiro <- WTG02$estado

# --- Cópula Gaussiana ---
matriz_gauss <- confusionMatrix(
  factor(pred_gauss[validos_gauss], levels = c(0,1)),
  factor(verdadeiro[validos_gauss], levels = c(0,1)),
  positive = "1")
# --- Cópula t ---
matriz_t <- confusionMatrix(
  factor(pred_t[validos_t], levels = c(0,1)),
  factor(verdadeiro[validos_t], levels = c(0,1)),
```

```
  positive = "1")

# ------------------------------------------------------------
```

 # 6. RESULTADOS

```
# ------------------------------------------------------------

cat("\n==============================")

cat("\n MATRIZ DE CONFUSÃO - CÓPULA GAUSSIANA")

cat("\n==============================\n")

print(matriz_gauss$table)

          Referência

Previsto   0       1

        0  1223 11663

       1    1668 17314


 MATRIZ DE CONFUSÃO - CÓPULA t> cat("\n==============================\n")

print(matriz_t$table)

             Referência

Previsto     0     1

        0   433  6107

       1    2458 22870
```

CENÁRIO 2

```
# ============================================================
# MATRIZ DE CONFUSÃO – CÓPULAS GAUSSIANA E t
# MÉTODO: MARGINAIS NORMAIS | VARIANTE: UM PASSO À FRENTE
# ============================================================

# Pacotes necessários
install.packages(c("copula", "MASS", "fitdistrplus", "caret", "dplyr"))
library(copula)
library(MASS)
library(fitdistrplus)
library(caret)
library(dplyr)
# ------------------------------------------------------------
# 1. Seleção das variáveis e filtragem dos dados
# ------------------------------------------------------------
vars <- c("WTG02_ActivePower", "WTG02_WindSpeed")
# Verificar se as colunas existem
if (!all(vars %in% names(WTG02))) {
  stop("O dataframe WTG02 precisa conter as colunas: ", paste(vars, collapse = ", "), " e 'estado'")}
# Remover NA
WTG02 <- na.omit(WTG02[, c(vars, "estado")])
```

```
# Separar dados normais (estado 1)

dados_normais <- WTG02 %>% filter(estado == 1)

# ----------------------------------------------------------

# 2. Ajuste das marginais normais

# ----------------------------------------------------------

# Estimar média e desvio padrão

param_marginais <- lapply(dados_normais[vars], function(x)

  list(mean = mean(x), sd = sd(x)))

# Converter dados normais para U[0,1] via CDF normal

U_norm <- as.data.frame(mapply(function(x, p)

  pnorm(x, mean = p$mean, sd = p$sd), dados_normais[vars], param_marginais))

# ----------------------------------------------------------

# 3. Ajuste das cópulas (Gaussiana e t)

# ----------------------------------------------------------

cop_gauss <- normalCopula(dim = length(vars))

cop_t     <- tCopula(dim = length(vars))

fit_gauss <- fitCopula(cop_gauss, data = as.matrix(U_norm), method = "itau")

fit_t     <- fitCopula(cop_t,     data = as.matrix(U_norm), method = "itau")

# ----------------------------------------------------------

# 4. VARIANTE “ONE-STEP-AHEAD” – PREVISÃO

# ----------------------------------------------------------
```

```
n <- nrow(WTG02)

pred_gauss <- numeric(n)

pred_t     <- numeric(n)

for (i in 2:n) { if (WTG02$estado[i-1] == 1 && all(!is.na(WTG02[i-1, vars]))) {

   obs <- as.numeric(WTG02[i, vars])

   # Converter observação para U via marginais normais

   U_obs <- mapply(function(x, p) pnorm(x, mean = p$mean, sd = p$sd),

                   as.list(obs), param_marginais)

   # Densidades sob cada cópula

   dens_gauss <- dCopula(U_obs, fit_gauss@copula)

   dens_t     <- dCopula(U_obs, fit_t@copula)

   # Classificação: densidade < limiar => anômalo (0)

   limiar <- 0.1

   pred_gauss[i] <- ifelse(dens_gauss < limiar, 0, 1)

   pred_t[i]     <- ifelse(dens_t < limiar, 0, 1)

  } else { pred_gauss[i] <- NA

   pred_t[i]     <- NA }}

# ---------------------------------------------------------

# 5. MATRIZES DE CONFUSÃO

# ---------------------------------------------------------

validos_gauss <- which(!is.na(pred_gauss))
```

```
validos_t     <- which(!is.na(pred_t))

verdadeiro <- WTG02$estado

# --- Cópula Gaussiana ---

matriz_gauss <- confusionMatrix(

  factor(pred_gauss[validos_gauss], levels = c(0,1)),

  factor(verdadeiro[validos_gauss], levels = c(0,1)),

  positive = "1")

# --- Cópula t ---

matriz_t <- confusionMatrix(

  factor(pred_t[validos_t], levels = c(0,1)),

  factor(verdadeiro[validos_t], levels = c(0,1)),

  positive = "1")

# ------------------------------------------------------

# 6. RESULTADOS

# ------------------------------------------------------

cat("\n MATRIZ DE CONFUSÃO - CÓPULA GAUSSIANA")

cat("\n==============================\n")

print(matriz_gauss$table)
```

Referência

```
Previsto     0     1
         0  1428  1778
         1    1463 27199


 MATRIZ DE CONFUSÃO - CÓPULA t> cat("\n============================\n")
print(matriz_t$table)
                Referência
Previsto     0     1
         0  1083   731
         1  1808 28246
```

CENÁRIO 3

```
# ============================================================
# MATRIZ DE CONFUSÃO – CÓPULAS GAUSSIANA E t
# MÉTODO: MARGINAIS NORMAIS | VARIANTE: UM PASSO À FRENTE
# ============================================================
# Pacotes necessários
install.packages(c("copula", "MASS", "fitdistrplus", "caret", "dplyr"))
library(copula)
library(MASS)
```

```
library(fitdistrplus)

library(caret)

library(dplyr)

# ----------------------------------------------------------

# 1. Seleção das variáveis e filtragem dos dados

# ----------------------------------------------------------

vars <- c("torque", "WTG02_WindSpeed")

# Verificar se as colunas existem

if (!all(vars %in% names(WTG02))) {

  stop("O dataframe WTG02 precisa conter as colunas: ", paste(vars, collapse = ", "), " e 'estado'")}

# Remover NA

WTG02 <- na.omit(WTG02[, c(vars, "estado")])

# Separar dados normais (estado 1)

dados_normais <- WTG02 %>% filter(estado == 1)

# ----------------------------------------------------------

# 2. Ajuste das marginais normais

# ----------------------------------------------------------

# Estimar média e desvio padrão

param_marginais <- lapply(dados_normais[vars], function(x)

  list(mean = mean(x), sd = sd(x)))

# Converter dados normais para U[0,1] via CDF normal
```

```
U_norm <- as.data.frame(mapply(function(x, p)

  pnorm(x, mean = p$mean, sd = p$sd), dados_normais[vars], param_marginais))

# ----------------------------------------------------------

# 3. Ajuste das cópulas (Gaussiana e t)

# ----------------------------------------------------------

cop_gauss <- normalCopula(dim = length(vars))

cop_t     <- tCopula(dim = length(vars))


fit_gauss <- fitCopula(cop_gauss, data = as.matrix(U_norm), method = "itau")

fit_t     <- fitCopula(cop_t,     data = as.matrix(U_norm), method = "itau")

# ----------------------------------------------------------

# 4. VARIANTE "ONE-STEP-AHEAD" – PREVISÃO

# ----------------------------------------------------------

n <- nrow(WTG02)

pred_gauss <- numeric(n)

pred_t     <- numeric(n)

for (i in 2:n) { if (WTG02$estado[i-1] == 1 && all(!is.na(WTG02[i-1, vars]))) {

    obs <- as.numeric(WTG02[i, vars])

    # Converter observação para U via marginais normais

    U_obs <- mapply(function(x, p) pnorm(x, mean = p$mean, sd = p$sd),

                    as.list(obs), param_marginais)
```

```
    # Densidades sob cada cópula
    dens_gauss <- dCopula(U_obs, fit_gauss@copula)
    dens_t     <- dCopula(U_obs, fit_t@copula)
    # Classificação: densidade < limiar => anômalo (0)
    limiar <- 0.1
    pred_gauss[i] <- ifelse(dens_gauss < limiar, 0, 1)
    pred_t[i]     <- ifelse(dens_t < limiar, 0, 1)
  } else {  pred_gauss[i] <- NA
    pred_t[i]     <- NA }}
# ----------------------------------------------------------
# 5. MATRIZES DE CONFUSÃO
# ----------------------------------------------------------
validos_gauss <- which(!is.na(pred_gauss))
validos_t     <- which(!is.na(pred_t))
verdadeiro <- WTG02$estado
# --- Cópula Gaussiana ---
matriz_gauss <- confusionMatrix(
  factor(pred_gauss[validos_gauss], levels = c(0,1)),
  factor(verdadeiro[validos_gauss], levels = c(0,1)), positive = "1")
# --- Cópula t ---
matriz_t <- confusionMatrix(
```

```
  factor(pred_t[validos_t], levels = c(0,1)),
  factor(verdadeiro[validos_t], levels = c(0,1)),
  positive = "1")
# ------------------------------------------------------------
# 6. RESULTADOS
# ------------------------------------------------------------
cat("\n==============================")
cat("\n MATRIZ DE CONFUSÃO - CÓPULA GAUSSIANA")
cat("\n==============================\n")
print(matriz_gauss$table)
               Referência
Previsto     0     1
         0  1222  1635
         1  1555 27454

cat("\n MATRIZ DE CONFUSÃO - CÓPULA t")
cat("\n==============================\n")
print(matriz_t$table)

           Referência
```

Previsto 0 1

0 981 600

1 1796 28489

ANEXO 3.7: CLASSIFIÇÃO USANDO MÉTODO DE CÓPULA, TURBINA WTG01

Tabela de matrizes de confusão, método de cópula (estático) turbina 1

Cenário	Cópula gaussiana	Cópula t-Student
1	Referência Prevista 0 1 0 1359 3252 1 6768 29265	Referência Prevista 0 1 0 1350 3252 1 6777 29265
2	Referência Prevista 0 1 0 5561 3252 1 2566 29265	Referência Prevista 0 1 0 5621 3252 1 2506 29265
3	Referência Prevista 0 1 0 4625 3252 1 3505 29262	Referência Prevista 0 1 0 4646 3252 1 3484 29262

Fonte: autoria própria

ANEXO3.8: CLASSIFIÇÃO USANDO MÉTODO DE CÓPULA, "*ONE-STEP-AHEAD*" TURBINA WTG01

Tabela de matrizes de confusão, método de cópula (one-step-ahead) turbina 1

Cenário	Cópula gaussiana	Cópula t-Student
1	Referência Previsto 0 1 0 1458 12351 1 1305 17403	Referência Previsto 0 1 0 583 7125 1 2180 22629
2	Referência Previsto 0 1 0 1852 2762 1 911 26992	Referência Previsto 0 1 0 1275 1359 1 1488 28395
3	Referência Previsto 0 1 0 1410 2421 1 1387 27296	Referência Previsto 0 1 0 1040 1078 1 1757 28639

Fonte: autoria própria