Aplicação de técnicas de mineração em dados sintéticos para manutenção preditiva: um estudo de caso

Rafael Schena[1], **João Cesar Netto**[1], **Karin Becker**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS

{rafael.schena, netto, karin.becker}@inf.ufrgs.br

Abstract. *This paper presents a case study of application data mining techniques in synthetic data for predictive maintenance of a naval propulsion system, with the objective of analyzing its applicability and suitability in the construction of predictive models for maintenance. In the first stage, we applied data mining techniques to the original dataset, and raised hypotheses about the results obtained with synthetic data. In the second and third stages, respectively, we tested the hypotheses raised in the initial stage by inserting class imbalance and measurement uncertainties. This way, we made the synthetic data more useful for building failure predicting models for real industrial scenarios.*

Resumo. *Este artigo apresenta um caso de aplicação de técnicas de mineração de dados em dados sintéticos para manutenção preditiva de um sistema de propulsão naval, com o objetivo de analisar sua aplicabilidade e adequação na construção de modelos preditivos para manutenção. Na primeira etapa, aplicamos técnicas de mineração de dados no* dataset *original, e levantamos hipóteses sobre os resultados obtidos com dados sintéticos. Na segunda e terceira etapas, respectivamente, testamos as hipóteses levantadas na etapa inicial através da inserção de desbalanceamento de classes e incertezas de medição. Desta forma, tornamos os dados sintéticos mais úteis à construção de modelos preditivos de falhas em cenários industriais reais.*

1. Introdução

A manutenção preditiva é uma estratégia de manutenção amplamente utilizada na indústria que busca o instante ótimo de execução das intervenções nos equipamentos, de modo a diminuir o número de intervenções, reduzir custos e aumentar a confiabilidade dos ativos. De acordo com a definição em [Mobley 2002], é um programa de manutenção preventiva com base na condição dos equipamentos, eficiência do sistema ou outros indicadores que possam retratar o nível de desgaste da máquina.

Os avanços nas últimas décadas em tecnologias de sensores possibilitaram o desenvolvimento das técnicas de *Prognostics and Health Monitoring* (PHM), um tipo de manutenção preditiva executada em tempo real através do monitoramento de dados de sensores instalados em equipamentos industriais. Com a integração de tecnologias como IIoT (*Industrial Internet of Things*), o gerenciamento de ativos de produção instalados em diferentes localidades pode ser realizado em tempo real, viabilizando inclusive o processamento remoto dos dados e intercâmbio de informações entre equipamentos similares.

Para a construção de modelos de manutenção preditiva PHM, existem três abordagens possíveis: (a) um modelo físico-matemático do equipamento em análise para comparação entre os dados reais coletados e os teóricos; (b) um modelo estatístico com base nos dados de uma população de equipamentos similares e aplicação destas estatísticas para prever a vida de um equipamento em particular; e, (c) abordagens orientadas a dados (*data-driven*), que fazem uso de uma massa de dados de falhas de equipamentos para treino de modelos de aprendizado de máquina para predição da falha.

A construção de modelos preditivos PHM tem sido intensamente estudada do ponto de vista da construção dos modelos usando técnicas de aprendizagem de máquina [Zhang et al. 2019, Susto et al. 2015, Mathew et al. 2017, Carchiolo et al. 2019, Mahmoodzadeh et al. 2020]. Contudo, para que retratem fidedignamente o processo de degradação das máquinas, dependem de dados reais de operação e falhas destes ativos, o que representa atualmente um grande desafio na área [Mauthe et al. 2021]. Tal dificuldade é devida principalmente a questões estratégicas das empresas e fabricantes de equipamentos, que não desejam que tais dados se tornem públicos. Além disso, as próprias indústrias têm dificuldade em coletar dados de falhas reais que formem um histórico consistente de indicadores de integridade do equipamento, tanto pelos altos custos e riscos gerados por falhas de determinadas máquinas, como também pela raridade estatística de determinados tipos de eventos catastróficos. Portanto, muitas vezes não há histórico de dados suficiente para se construir um modelo confiável no tocante às falhas que se deseja evitar.

Neste cenário, a geração de dados sintéticos se apresenta como uma solução técnica e economicamente viável para construção de modelos preditivos para PHM. Porém, para construção de um modelo de PHM eficaz para os cenários industriais relevantes deve-se considerar uma série de requisitos do conjunto de dados, entre eles a cobertura do espaço de estados possível das variáveis [Mauthe et al. 2021]. Logo, a simples geração de dados por simuladores pode não ser suficientemente verossímil para a construção de modelos preditivos úteis em cenários práticos da indústria. Neste contexto existe uma lacuna de trabalhos que explorem as condições aceitáveis e os limites de aplicação de dados sintéticos na construção de modelos PHM para manutenção preditiva.

O presente trabalho desenvolve um estudo de caso de manutenção preditiva orientada a dados onde aplicam-se técnicas de mineração a dados sintéticos, propondo melhorias ao conjunto de dados para torná-lo mais útil à construção de modelos preditivos de falhas em cenários reais. O objetivo inicial foi entender quão bem um *dataset* sintético representa a realidade do problema de predição do estado de saúde dos equipamentos, e, posteriormente propor alternativas para torná-lo mais fiel à realidade de operação dos equipamentos através da inserção de incertezas nos dados e desbalanceamento de classes. O estudo explorou um *dataset* público [Coraddu et al. 2016] de uma turbina de propulsão naval e foi desenvolvido em três etapas. Na primeira, criou-se um *baseline* envolvendo a modelagem preditiva do estado de manutenção do sistema utilizando a metodologia CRISP-DM com os dados originais divididos em três classes estatisticamente equilibradas (normal, alerta e emergência). Com base nos resultados da primeira etapa, testamos hipóteses de adequação do conjunto original de dados a situações típicas de cenários reais de operação, a saber, situações de desbalanceamento do *dataset* e perturbações causadas pela presença de ruídos e incertezas de medição nos dados. Esta abordagem contribui com *insights* para a geração de dados sintéticos para modelos de manutenção preditiva

mais condizentes com a realidade da operação da indústria.

Os resultados mostram que modelos computacionais geradores de dados sintéticos, apesar de capazes de descrever fisicamente as relações existentes entre as variáveis de um sistema complexo como um sistema de propulsão naval, podem deixar de retratar nuances típicas de um cenário industrial prático (ruído nos dados, incertezas de medição, falhas nos instrumentos). Os experimentos também apontam que é possível tratar tais limitações em um ciclo que permite ajustes tanto nos modelos preditivos quanto no conjunto de dados. Todos os códigos, dados utilizados, resultados intermediários e demais desenvolvimentos estão disponíveis em um repositório público.

O restante deste manuscrito se organiza da seguinte forma: a Seção 2 apresenta os trabalhos relacionados abordando os desafios da manutenção preditiva orientada a dados; a Seção 3 apresenta a visão geral do estudo de caso; a Seção 4 apresenta os experimentos iniciais seguindo as etapas da metodologia CRISP-DM; as seções 5 e 6 apresentam os testes relativos à inserção de desbalanceamento e de perturbações na distribuição dos dados e; por fim, a Seção 7 apresenta as conclusões do trabalho.

2. Trabalhos relacionados

A evolução da IIoT, com tecnologias de sensores industriais viabiliza a geração em tempo real de uma enorme massa de dados relativos à saúde de equipamentos, motivou muitos trabalhos que exploram técnicas de aprendizado de máquina para manutenção preditiva. Em [Zhang et al. 2019], é feita uma comparação de diversas técnicas e métodos supervisionados de classificação para manutenção preditiva. Uma nova abordagem orientada a dados para manutenção baseada na condição é apresentada em [Susto et al. 2015], em que são utilizados multi-classificadores para possibilitar a implantação de regras de decisão dinâmicas para gestão de manutenção com base em indicadores de saúde dos equipamentos, com dados censurados e com alta dimensionalidade. Em [Mathew et al. 2017], são aplicados e comparados algoritmos de regressão para previsão do tempo restante de vida útil até a falha do equipamento. A utilização de técnicas de processamento de linguagem natural para minerar padrões em relatórios de manutenção é explorada em [Carchiolo et al. 2019] para previsão de falhas em plantas de geração de energia. Em [Mahmoodzadeh et al. 2020] é utilizado o aprendizado por reforço para otimização da estratégia de prevenção de corrosão em dutos de gás, visando a otimização do sistema para minimizar falhas e otimizar o custo do ativo ao longo da sua vida útil. Nenhum destes trabalhos, contudo, aborda as verificações necessárias aos dados para certificação de que as situações reais que se deseja prever são bem caracterizadas nos dados de entrada dos modelos.

De acordo com [Mauthe et al. 2021], a baixa disponibilidade de dados reais de falhas é um dos grandes desafios de pesquisa na área de PHM. Os autores analisam *datasets* públicos de manutenção preditiva, concluindo que a vasta maioria advêm de simulações e ensaios de laboratório e alertam que a maioria não cobre de forma significativa todo o espaço de estados relevante para a previsão de falhas. No entanto, não é feita uma análise estatística exploratória detalhada explicitando as deficiências apontadas.

Atacando o problema da falta de dados reais de falha para construção de modelos preditivos, [Rao 2020] propõe a utilização de dados sintéticos gerados por um gêmeo digital, onde os dados são gerados a partir de um modelo físico em um simulador com

defeitos modelados, para que fossem simuladas situações de desgaste extremo do equipamento não contempladas pelo histórico de operação. O trabalho compara as curvas de dados simulados e dados reais medidos em campo, mas não explora a cobertura dos dados sintéticos para os estados relevantes para a indústria.

Dados sintéticos são gerados em [Coraddu et al. 2016] para treino de modelos de predição do nível de desgaste dos equipamentos de propulsão naval. Apesar do processo de geração dos dados ser descrito, não é apresentada uma descrição estatística dos dados mostrando compatibilidade com cenários de dados reais encontrados na indústria.

Neste trabalho desenvolvemos um estudo de caso de manutenção preditiva orientada a dados em que são utilizadas técnicas de mineração de dados para extração de padrões em um conjunto de dados sintéticos para manutenção preditiva e inserção de perturbações nos dados para torná-los mais compatíveis a medições observadas em uma instalação real.

3. Visão Geral

Este trabalho adota o dataset sintético criado em [Coraddu et al. 2016] para predição de falhas em sistema de propulsão naval acionado por turbina a gás, tanto em termos de índice de degradação da compressão como de turbina. Para o desenvolvimento do estudo de caso, usamos a metodologia CRISP-DM (*CRoss-Industry Standard Process for Data Mining*) [Wirth and Hipp 2000], que se tornou um padrão independente da indústria para desenvolvimento de projetos de mineração de dados, inclusive nas áreas de engenharia e manufatura [Schröer et al. 2021]. As fases da metodologia (mostradas na figura 1) foram executadas de tal forma que as propriedades estatísticas dos dados pudessem ser evidenciadas, e que hipóteses sobre a influência dos dados nos modelos preditivos pudessem ser formuladas, contribuindo assim para um melhor entendimento de propriedades a serem consideradas na geração de dados sintéticos para manutenção preditiva.

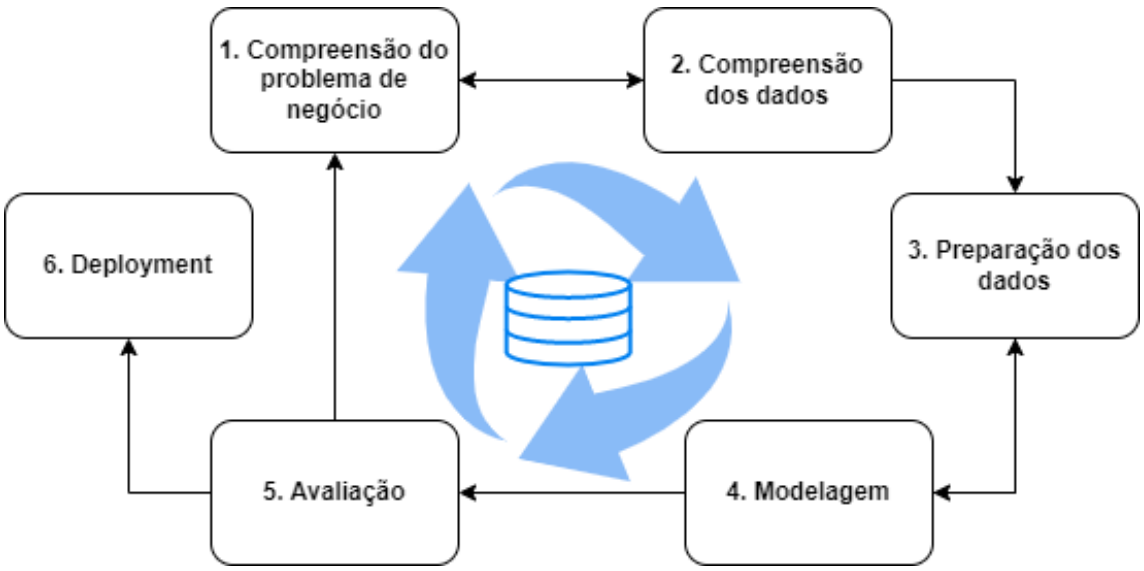

Figura 1. Fases da metodologia CRISP-DM [Wirth and Hipp 2000].

Em uma primeira etapa, analisamos os dados e desenvolvemos um *baseline* para os modelos preditivos PHM testando diferentes algoritmos de aprendizado supervisio-

nado, e analisando seu desempenho. Face aos resultados, em seguida testamos duas hipóteses de alteração nos dados, e seu consequente efeito nos modelos preditivos. As hipóteses testadas orientaram o dataset a cenários mais realistas da indústria, a saber, casos de falha mais esparsos e perturbações nos dados coletados.

Cada etapa é detalhada nas seções que se seguem. Para os experimentos, foram utilizadas a linguagem Python, juntamente com os pacotes pandas, numpy, matplotlib, seaborn e scikit-learn. O código está disponível em um repositório público[1].

4. Etapa 1 - Aplicação de técnicas de mineração de dados aos dados sintéticos originais

Geramos e avaliamos modelos preditivos treinados sobre o dataset sintético original seguindo a metodologia CRISP-DM, de acordo com cada uma das fases do modelo de referência, como descrito no restante desta seção.

4.1. Compreensão do problema de negócio

A primeira etapa do modelo de referência CRISP-DM visa estabelecer os objetivos do negócio e de mineração correspondentes. Com base nos dados adotados, temos como objetivo criar modelos preditivos que classificam os níveis de degradação do compressor e da turbina de um sistema de propulsão naval. Os dados estão disponíveis publicamente no repositório da UCI[2], onde também pode ser encontrado o dicionário de dados com a descrição de todas variáveis.

É importante destacar que o trabalho original associado a esses dados [Coraddu et al. 2016] visa a construção de um modelo de regressão que quantifica as variáveis que definem níveis de degradação dos estágios de compressão (kMc) e da turbina (kMt), os quais caracterizam a perda de eficiência dos equipamentos ao longo do uso devido ao desgaste, erosão e acúmulo de sujeira nos componentes. Portanto foi necessário definir níveis arbitrários de degradação com base na distribuição das variáveis alvo kMc e kMt. Como primeira abordagem, foram definidos os seguintes níveis arbitrários de degradação:

- Emergência: faixa de 33% piores valores possíveis para o intervalo observado;
- Alerta: faixa entre 33% e 66% piores valores possíveis para o intervalo observado;
- Normal: faixa de 33% melhores valores possíveis para o intervalo observado.

Desta forma, definimos o problema de mineração como a classificação dos estados de operação do compressor e da turbina de um motor a reação de um sistema de propulsão naval em três possíveis categorias: *Emergência*, *Alerta* ou *Normal*.

4.2. Compreensão dos Dados

O objetivo desta fase do CRISP-DM é verificar a compatibilidade do *dataset* com os objetivos de negócio/mineração definidos, e avaliar as propriedades e sanidade dos dados. O *dataset* é composto por 11.934 observações e 18 variáveis. Não há valores faltantes, e todas as variáveis são contínuas (float64). Em se tratando de grandezas físicas que são

[1] `https://github.com/rafaelschena/naval_predictive_maintenance`

[2] `https://archive.ics.uci.edu/ml/datasets/condition+based+maintenance+of+naval+propulsion+plants`

medidas por meio de sensores em diferentes pontos do sistema (velocidade, pressão, temperatura, rotação), não foram verificadas inconsistências com a representação dos dados. Os histogramas com a distribuição das variáveis são mostrados na Figura 2.

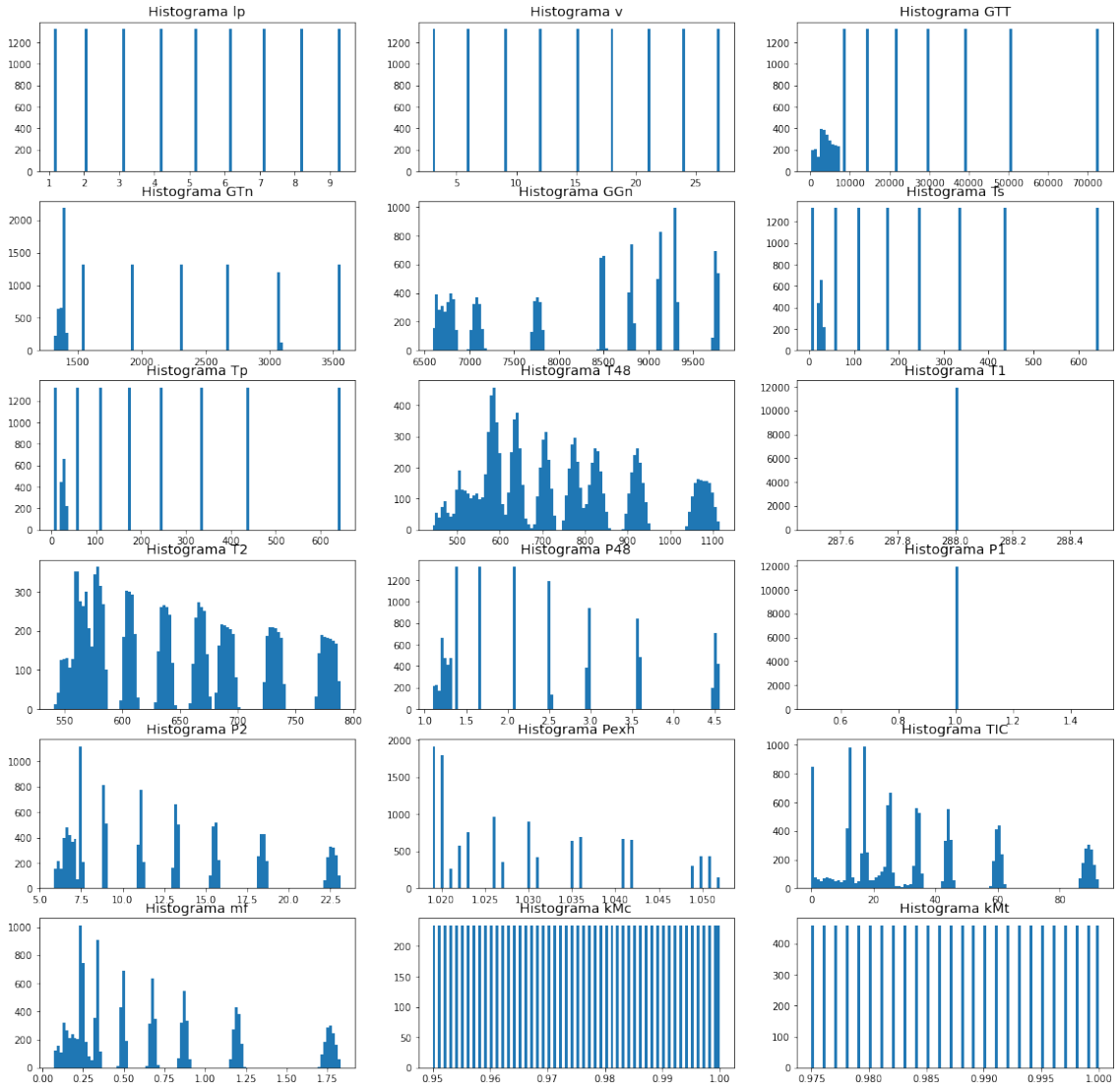

Figura 2. Histogramas das variáveis do *dataset*

Dos gráficos da Figura 2 e das observações iniciais do *dataset*, pôde-se observar que:

- Os dados não se apresentam na forma de séries temporais, uma vez que não têm associado um timestamp e nenhuma indicação de que estão em uma sequência no tempo;
- Na distribuição estatística dos dados são identificadas várias lacunas, indicando que existem faixas de valores que não constam no conjunto de dados, apesar de se tratar de um *dataset* composto somente por variáveis contínuas. Tal característica da distribuição dos dados permite que se efetue uma discretização destas grandezas com uma menor perda de informação (erro de quantização);
- Há faixas de variação muito diferentes entre as variáveis;

- Existem variáveis que não apresentam variabilidade e podem ser eliminadas do *dataset*, porque não contribuem para a previsibilidade das variáveis alvo.

Fizemos ainda uma análise de correlações entre as variáveis do conjunto de dados, utilizando o método de Pearson. Encontramos correlações baixas entre as variáveis preditoras e as variáveis alvo, como mostrado na figura 3, o que indica que ou não há relação entre elas ou a relação existente não é linear. Tal característica do *dataset* corrobora, em um primeiro momento, a utilização de modelos de classificação para previsão por níveis de degradação, uma vez que a tendência para este tipo de relação é que os erros de previsão sejam altos e a utilidade da previsão gerada seja baixa.

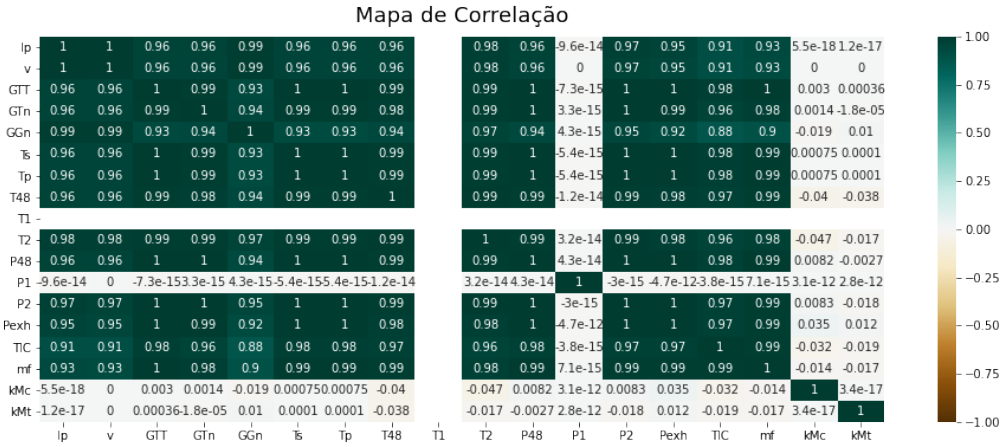

Figura 3. Mapa de correlação entre as variáveis do dataset.

4.3. Preparação dos Dados

Esta fase do CRISP-DM envolve o pré-processamento necessário às atividades de mineração de dados. Executamos as seguintes tarefas:

- Exclusão das variáveis T1 e P1, que não apresentam variabilidade e não contribuem para o processo;
- Criação de rótulos com status '*Normal*', '*Alerta*' e '*Emergência*' para as variáveis alvo, conforme indicado na Seção 4.1;
- Exclusão das variáveis kMc e kMt que foram utilizadas para criação dos rótulos discretos;
- Divisão dos dados em conjunto de treino (70% das observações) e conjunto de teste (30% das observações).
- Criação de versões alternativas do *dataset* com dados normalizados para investigar o impacto sobre a correlação entre as variáveis preditoras e as variáveis alvo. Posteriormente, verificou-se que mesmo utilizando dados padronizados e normalizados, o problema das baixas correlações persistiu, e os novos *datasets* gerados acabaram não sendo utilizados para treinamento dos modelos preditivos.

4.4. Modelagem

Esta fase do CRISP-DM envolve a mineração de dados propriamente dita. Testamos seis algoritmos de classificação baseados em estratégias diferentes para abordar o problema

proposto de detecção do nível de degradação do compressor e da turbina do sistema de propulsão naval. Os algoritmos selecionados foram: Decision Tree Classifier, Random Forest Classifier, K-Nearest-Neighbors, Naive-Bayes, Support Vector Machine Classifier e AdaBoost Classifier[3]. A Figura 4 mostra os resultados obtidos após o treinamento dos modelos para cada uma das variáveis alvo, e apuração das métricas utilizando os dados de teste.

Foram computadas para avaliação as métricas de acurácia geral do modelo e a revocação do rótulo *Emergência*, que indica o percentual dos casos realmente classificados como emergência os modelos foram capazes de prever. Como o caso da manutenção de emergência é sempre mais crítico, este indicador é o mais importante a ser observado, pois uma situação de emergência não detectada pelo modelo pode causar um prejuízo maior à operação da embarcação. No entanto, a revocação não pode ser avaliada isoladamente, pois um modelo enviesado que atribuísse estado de emergência a todos os casos que avaliasse teria uma revocação de 100%, porém erraria todas as demais previsões, de modo que o ideal é realizar uma avaliação conjunta com a acurácia geral do modelo.

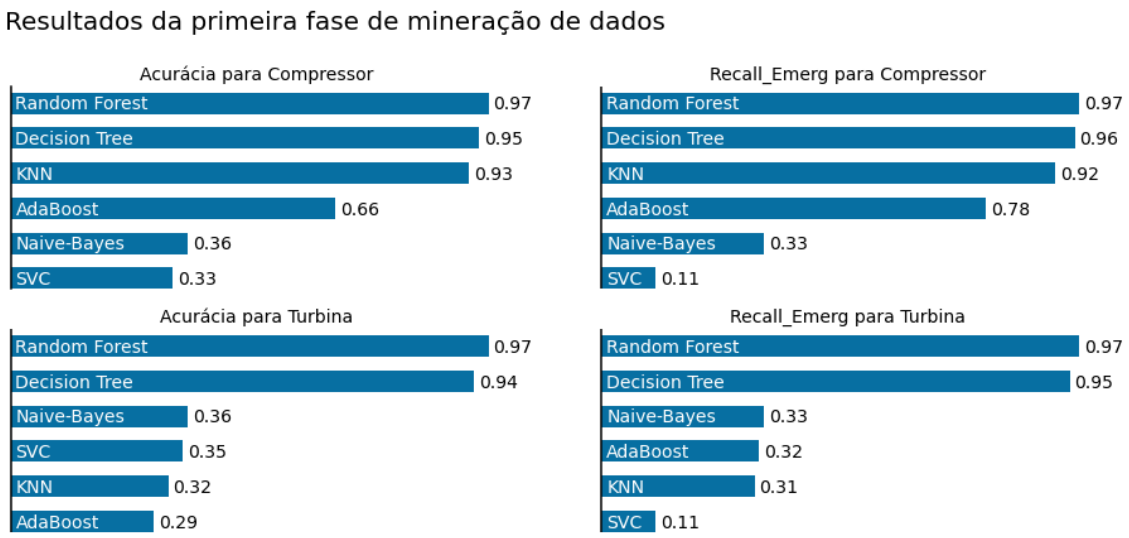

Figura 4. Métricas dos modelos de classificação para a Etapa 1.

4.5. Avaliação

Nesta etapa do CRISP-DM os resultados da modelagem são avaliados em relação aos objetivos. Considerando ambas as variáveis alvo, os modelos treinados com o algoritmo Random Forest apresentaram os melhores resultados, com revocação e acurácia de 97% tanto para o compressor como para turbina, como pode ser observado na Figura 4. Com resultados um pouco abaixo destes, vêm os modelos baseados no algoritmo Decision Tree. Todos os demais modelos tiveram desempenho inferior, com alguns ficando próximos do *baseline* da aleatoriedade.

Algumas hipóteses foram levantadas durante a apresentação dos resultados parciais do projeto a respeito dos primeiros resultados. São elas:

[3]Documentação com detalhes de implementação e utilização disponíveis em: `https://scikit-learn.org/stable/`

- **Hipótese 1**: faixas arbitrárias definidas para níveis de normalidade, alerta e emergência podem ser muito amplos e facilitar o trabalho de classificação, o que justifica os resultados excelente obtidos. Contudo, em um cenário normal de operação, os equipamentos atuam em condições normais a maior parte do tempo, sugerindo que é necessário que os dados sintéticos representem o desbalanceamento natural dos dados. O teste desta hipótese é detalhado na Seção 5;
- **Hipótese 2**: alto desempenho dos primeiros modelos, juntamente com a distribuição dos dados mostrada nos histogramas da Figura 3, podem ser um indício que a separação das variáveis contínuas em faixas bem delimitadas do espectro facilitou a tarefa de classificação por algoritmos do tipo árvore de decisão, explicando o bom desempenho observados nos modelos de Decision Tree e Random Forest. Tal hipótese pode ser corroborada pelo fato de o *dataset* ser simulado e, por definição, ser "livre" de erros de medição e ruídos típicos de uma instalação real. O teste desta hipótese é descrito na Seção 6.

As hipóteses levantadas foram testadas redefinindo-se os níveis arbitrários de classificação e inserindo erros aleatórios de medição nas variáveis preditoras que podem ser passíveis de erros, em níveis compatíveis com os sensores aplicáveis a cada medida. Tais soluções dispensam o uso do simulador reparametrizado para geração de novos dados.

5. Etapa 2: Resposta dos modelos preditivos ao desbalanceamento do dataset

Foram testados dois cenários alternativos para faixas alvo. Os cenários são os seguintes:

Cenário 1:

- Emergência: faixa de 10% piores valores possíveis para o intervalo observado;
- Alerta: faixa entre 10% e 20% piores valores possíveis para o intervalo observado;
- Normal: faixa de 80% melhores valores possíveis para o intervalo observado.

Cenário 2:

- Emergência: faixa de 5% piores valores possíveis para o intervalo observado;
- Alerta: faixa entre 5% e 10% piores valores possíveis para o intervalo observado;
- Normal: faixa de 90% melhores valores possíveis para o intervalo observado.

A Figura 5 mostra a evolução do indicador composto Acurácia x Revocação_Emerg para cada um dos algoritmos elencados na Seção 4.4 testados para previsão dos estágios do compressor e da turbina.

Observa-se nesses resultados uma queda na performance dos algoritmos de classificação quando as faixas de valores para classificação como níveis de '*Alerta*' e '*Emergência*' são estreitadas. Porém, há que se observar que os algoritmos RandomForest, DecisionTree e KNN se mostraram bastante robustos a este estreitamento e consequente desbalanceamento do *dataset*, com acurácias gerais acima de 90% e revocação da classe *Emergência* acima de 70%. Por outro lado, os algoritmos AdaBoost, SVC e Naive-Bayes tiveram a revocação da classe *Emergência* tendendo a zero no pior cenário.

6. Etapa 3 – Inserção de incertezas no *dataset*

Para testar a hipótese de falta de erros de medição em um *dataset* simulado foram inseridos erros gaussianos para cada uma das variáveis oriundas de medidas físicas (dado

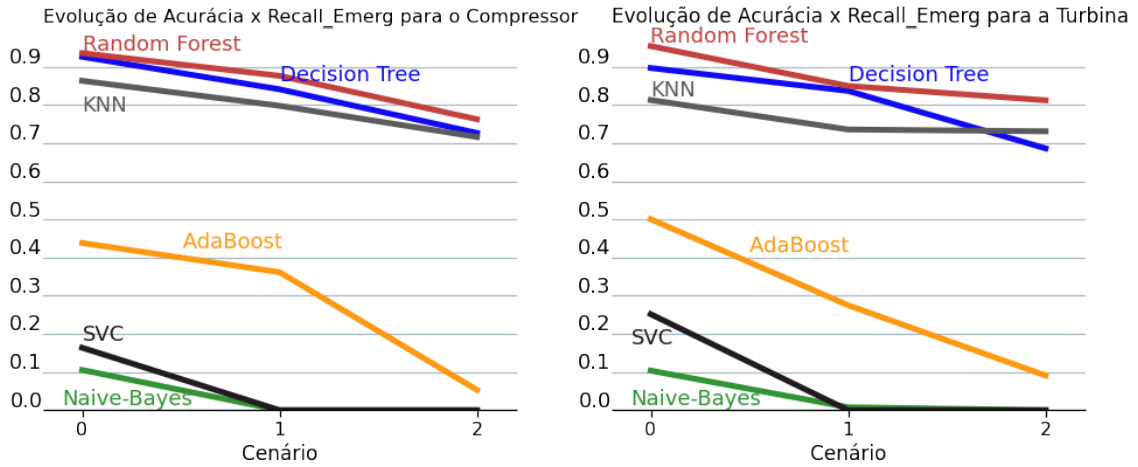

Figura 5. Resultados do indicador composto Acurácia x Recall_Emerg para os modelos testados nos três cenários de balanceamento de classes para o compressor e turbina.

que os erros são desvios para mais ou para menos em torno de uma média, assumida como o valor correto). Tentou-se emular um cenário de pior caso para cada tipo de erro, cuidando para não se extrapolar a ordem de grandeza da incerteza referente a cada tipo de medida. O trabalho de [Dalheim and Steen 2021] foi utilizado como referência para incerteza nas medidas de velocidade (3,6%), torque (1%) e rotação (1%). Já as incertezas das medidas de temperatura (0,01%), pressão (1%) e fluxo de massa (2%) foram baseadas em [Brun and Kurz 1998].

Aplicados ao *dataset*, os erros descritos acima provocaram alterações na distribuição dos dados, como observado na Figura 6, onde a título de exemplo as distribuições dos dados de duas variáveis são comparadas antes e depois da inserção das incertezas. Nos histogramas com incertezas inseridas pode ser visto um maior espalhamento dos valores devido ao ruído inserido, muitas vezes cobrindo lacunas de valores que não foram cobertos no *dataset* original.

O mesmo procedimento para treinamento e teste dos seis modelos de classificação foi aplicado aos novos dados com erros incertezas inseridas. Dos resultados mostrados na Figura 7 pode-se constatar que, como esperado, o acréscimo de incertezas aos dados piora o desempenho dos modelos. Quando estas incertezas são combinadas com uma faixa mais estreita de valores de referência para os níveis de *Alerta* e *Emergência*, o desempenho dos modelos de classificação diminui ainda mais. No entanto, dois modelos se sobressaíram em relação aos demais em quaisquer cenários. São eles os modelos baseados nos algoritmos RandomForest e DecisionTree. Tomando como base o cenário 1, com classes desbalanceadas (10% Emergência e 10% Alerta), para o estágio compressor os modelos apresentaram acurácia superior a 89% e revocação da classe de emergência superior a 74% (Acurácia x Revocação_Emerg 66%). Já para o estágio da turbina, o desempenho foi pior, com acurácia acima de 78% e revocação da classe *Emergência* acima de 47% (Acurácia x Revocação_Emerg 37%).

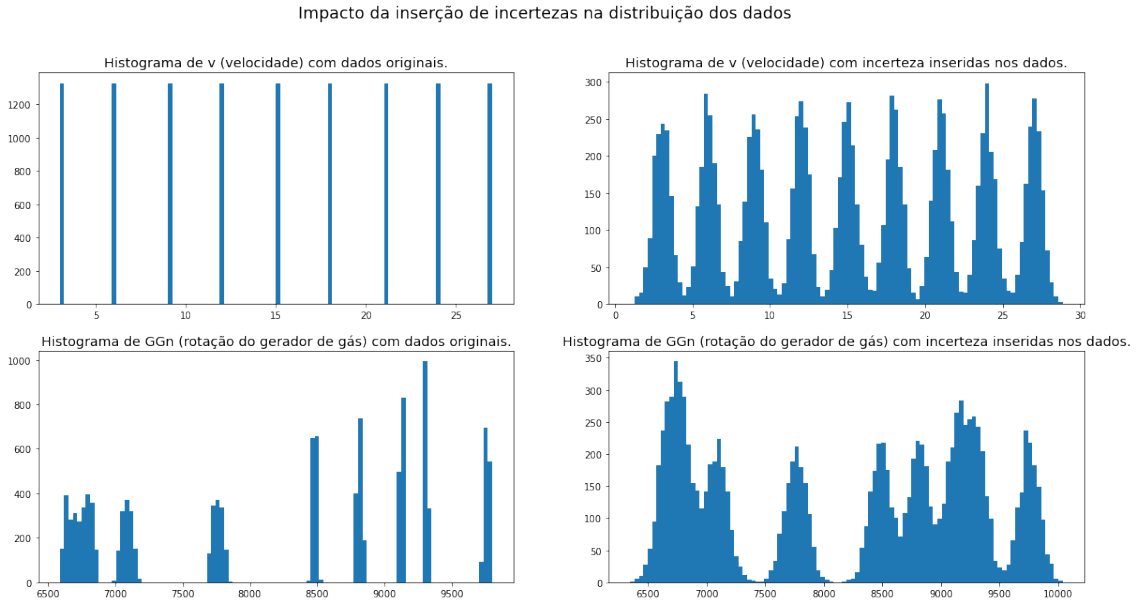

Figura 6. Comparativo dos histogramas das variáveis v (velocidade) e GGn (rotação do gerador de gás) com dados originais e com incertezas inseridas.

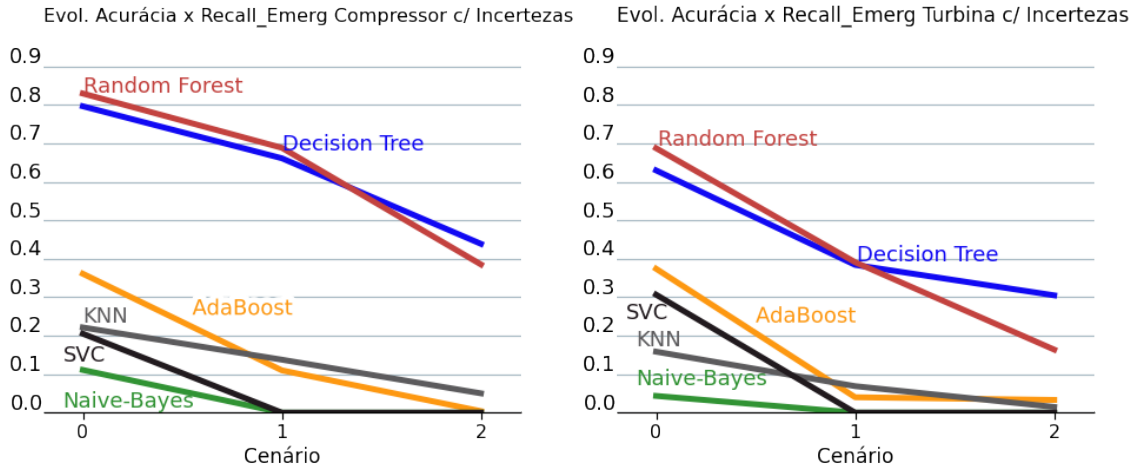

Figura 7. Resultados do indicador composto Acurácia x Recall_Emerg para os modelos testados nos três cenários de balanceamento de classes e com incertezas inseridas para o compressor e turbina.

7. Conclusões

Neste trabalho desenvolvemos um estudo de caso para investigação sobre a aplicação de dados sintéticos para manutenção preditiva, quais são suas limitações e condições necessárias para se simular cenários verossímeis e sobre como tornar os dados sintéticos mais semelhantes a dados industriais reais. Vimos que o problema de manutenção preditiva em tempo real (PHM) possui um cenário de dados característico, onde os dados de falha são escassos por natureza, os dados de grandezas medidas estão sujeitos à presença de incertezas de medição, ruídos e falhas de sensores. Para suprir esta demanda, a geração de dados sintéticos se mostra como uma alternativa técnica e economicamente viável para

construção de modelos preditivos de manutenção. No entanto, ainda existe um *gap* de pesquisa quanto à caracterização dos dados sintéticos para manutenção preditiva.

Neste sentido, o presente estudo contribui com uma abordagem para avaliação da qualidade e e geração de *insights* para melhorar o processo de criação de dados sintéticos de manutenção preditiva. Atingimos tal resultado utilizando técnicas de mineração de dados para análise estatística e extração de padrões em um conjunto de dados simulados, com a finalidade de avaliar sua aplicabilidade para a construção de modelos preditivos de manutenção.

O trabalho foi desenvolvido em três etapas. Na primeira etapa do estudo utilizamos a metodologia CRISP-DM, com a utilização de algoritmos de classificação para previsão do estado de manutenção do equipamento em questão em relação aos seus dois estágios: compressor e turbina. Para definição das variáveis alvo na etapa 1 do trabalho foram definidos limites arbitrários dividindo o *dataset* em 3 classes de iguais tamanhos em relação ao índice de degradação de cada estágio calculado pelo simulador: estado normal, alerta e emergência. Ao final da primeira rodada de modelagem, foram obtidos bons resultados para previsão das classes dos equipamentos.

Foram levantadas hipóteses que pudessem explicar o cenário encontrado na etapa 1. A primeira delas com respeito às faixas de valores atribuídas arbitrariamente como níveis de alerta e emergência. Na etapa 2, foram redistribuídos os dados de acordo com outros limites arbitrários em dois cenários alternativos que geraram classes desbalanceadas para previsão. Os mesmos modelos de classificação foram testados com os dados dos novos cenários e as métricas foram apuradas. Três algoritmos se mostraram relativamente mais robustos ao desbalanceamento das classes: DecisionTree, RandomForest e KNN.

A outra hipótese levantada para explicar os resultados do cenário base foi a de que os dados, por serem sintéticos, não continham ruídos e incertezas relativas ao processo de medição por sensores. Para testar esta hipótese, na etapa 3, foram pesquisados valores compatíveis para incertezas de medição para cada uma das grandezas físicas do *dataset*, e um erro compatível foi inserido nos dados. Neste caso, observou-se uma significativa redução no desempenho dos modelos de classificação, especialmente em relação à métrica de revocação da classe emergência, mais acentuada quanto mais desbalanceados eram os dados, de acordo com os mesmos cenários propostos na hipótese anterior. Ainda assim, foram observados resultados relativamente robustos para os modelos de RandomForest e DecisionTree, com acurácia geral em torno de 80% e revocação da classe de emergência em torno de 50%. Em um cenário prático, tal resultado já seria um indicativo forte de que tal equipamento deveria ir para manutenção imediata, sendo capaz de detectar metade das falhas graves iminentes.

Pôde-se concluir que dados sintéticos podem ser aplicados na construção de modelos preditivos de manutenção. Contudo, há que se ter os devidos cuidados quanto à cobertura do espaço de estados possíveis e quanto à presença de distorções inerentes ao cenário prático na indústria, como a presença de ruídos e erros de medição.

Um possível experimento complementar seria a aplicação das mesmas técnicas sobre um conjunto de dados reais amostrados em ensaios de laboratório em condições controladas e avaliados comparativamente para condições de campo ou de longa duração. Um dos objetivos futuros deste projeto de pesquisa é a integração entre dados de sensores

em tempo real, simuladores e modelos de aprendizado de máquina em um gêmeo digital, possibilitando complementar os dados reais com dados sintéticos para aprimorar os modelos preditivos de falhas.

Agradecimentos Pesquisa parcialmente apoiada pela CAPES e Projeto PeTwin (financiamento FINEP e Libra Consortium).

Referências

Brun, K. and Kurz, R. (1998). Measurement uncertainties encountered during gas turbine driven compressor field testing. In *Turbo Expo: Power for Land, Sea, and Air*, volume 78644, page V003T07A001. American Society of Mechanical Engineers.

Carchiolo, V., Longheu, A., Di Martino, V., and Consoli, N. (2019). Power plants failure reports analysis for predictive maintenance. In *WEBIST*, pages 404–410.

Coraddu, A., Oneto, L., Ghio, A., Savio, S., Anguita, D., and Figari, M. (2016). Machine learning approaches for improving condition-based maintenance of naval propulsion plants. *Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment*, 230(1):136–153.

Dalheim, Ø. Ø. and Steen, S. (2021). Uncertainty in the real-time estimation of ship speed through water. *Ocean Engineering*, 235:109423.

Mahmoodzadeh, Z., Wu, K.-Y., Lopez Droguett, E., and Mosleh, A. (2020). Condition-based maintenance with reinforcement learning for dry gas pipeline subject to internal corrosion. *Sensors*, 20(19):5708.

Mathew, V., Toby, T., Singh, V., Rao, B. M., and Kumar, M. G. (2017). Prediction of remaining useful lifetime (rul) of turbofan engine using machine learning. In *2017 IEEE International Conference on Circuits and Systems (ICCS)*, pages 306–311.

Mauthe, F., Hagmeyer, S., and Zeiler, P. (2021). Creation of publicly available data sets for prognostics and diagnostics addressing data scenarios relevant to industrial applications. *International Journal of Prognostics and Health Management*, 12(2).

Mobley, R. K. (2002). *An introduction to predictive maintenance*. Elsevier.

Rao, S. V. (2020). Using a digital twin in predictive maintenance. *Journal of Petroleum Technology*, 72(08):42–44.

Schröer, C., Kruse, F., and Gómez, J. M. (2021). A systematic literature review on applying crisp-dm process model. *Procedia Computer Science*, 181:526–534.

Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., and Beghi, A. (2015). Machine learning for predictive maintenance: A multiple classifier approach. *IEEE Transactions on Industrial Informatics*, 11(3):812–820.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–40. Manchester.

Zhang, W., Yang, D., and Wang, H. (2019). Data-driven methods for predictive maintenance of industrial equipment: A survey. *IEEE Systems Journal*, 13(3):2213–2227.