Análise de dados para geração de indicadores de uma planta de tratamento de água

Ítalo O. Fernandes * **Heitor M. Florencio** **

* *Programa de Pós-Graduação em Tecnologia da Informação, Instituto Metrópole Digital, Universidade Federal do Rio Grande do Norte, RN, (e-mail: italo.oliveira.029@ufrn.edu.br).*
** *Instituto Metrópole Digital, Universidade Federal do Rio Grande do Norte, RN, (e-mail: heitorm@imd.ufrn.br)*

Abstract: This work aims to generate indicators to monitor a pharmaceutical laboratory water treatment plant processes from the creation of a data analysis model, which includes stages from data collection to data visualization. The proposed methodology includes understanding the process and defining analysis objectives, collecting data from controllers, preparing them in the pre-processing stage to make them suitable for the study and then exploring and visualizing information extracted from these data. The expected result is to generate monitoring indicators that provide information to improve water treatment unit monitoring.

Resumo: Este trabalho tem como objetivo gerar indicadores para monitorar os processos da estação de tratamento de água em um laboratório farmacêutico a partir da criação de um modelo de análise de dados, que inclui desde a coleta até a visualização de dados. A metodologia utilizada inclui compreender o processo e definir objetivos de análise, coletar dados dos controladores dos processos, prepará-los no pré-processamento para melhorar a confiabilidade dos dados para o estudo e então explorar e visualizar informações extraídas desses dados. O resultado esperado é a geração de indicadores de monitoramento que forneçam informações para melhorar o acompanhamento da unidade de tratamento.

Keywords: Indicators; Data analysis; Monitoring; Industry; Pharmaceutical.
Palavras-chaves: Indicadores; Análise de dados; Monitoramento; Indústria; Farmacêutica.

1. INTRODUÇÃO

O avanço da digitalização da informação faz com que o número de dados produzidos pelas mais diversas fontes cresçam a todo momento e essa grande quantidade de dados tem a capacidade de impactar organizações, sistemas de produção e a sociedade (Brynjolfsson e McAfee, 2014). No contexto da indústria, são utilizadas tecnologias digitais para coletar e processar dados heterogêneos de diversos sensores e dispositivos IoT em tempo real, criando informações úteis ao processo industrial (Frank et al., 2019).

Contudo, muitos sistemas de manufatura ainda não estão preparados para lidar com grandes volume de dados produzidos por falta de ferramentas inteligentes para análise de dados (Lee et al., 2014). Dessa forma, existem organizações que produzem e armazenam grandes volumes de dados, mas não conseguem extrair informações úteis ao processo produtivo.

Esse trabalho propõe a geração de indicadores de desempenho de uma planta de tratamento de água a partir criação de um modelo de análise de dados. O modelo de análise dados, ou *pipeline*, contempla a coleta dos dados, as técnicas de processamento dos dados, a análise descritiva e estatística dos dados até a geração do indicadores a partir de análise de sazonalidade dos dados.

2. REFERENCIAL TEÓRICO

A análise de dados se baseia em extrair informações significativas a partir de um conjunto de dados e então utilizá-las como base para melhorar processos de tomada de decisão em um determinado contexto, tornando-os mais eficientes e precisos.

2.1 Python para análise de dados

A linguagem de programação *Python* é uma ferramenta de código aberto popularmente utilizada para trabalhos na área de análise de dados, trazendo vantagens como gratuidade, simplicidade e uma grande comunidade ativa de desenvolvedores (VanderPlas, 2016). Devido ao seu suporte melhorado para bibliotecas e sua robustez, ela se tornou uma das principais linguagens para ciência de dados e aprendizado de máquina frente a outros concorrentes e permite a criação de aplicações poderosas para dados (McKinney, 2019).

Dentre as principais bibliotecas do ambiente *Python* utilizadas para análise de dados e presentes nesse trabalho estão: o *NumPy*, utilizado para computação científica, permite criar estruturas de dados em formatos de *arrays* multidimensionais chamados de *ndarrays*; o *Pandas*, utilizado para trabalhar com dados tabulares; a *Matplotlib* e a

Seaborn, utilizadas para visualização de dados através da geração de gráficos.

2.2 Análise de séries temporais

As séries temporais são sequências de registros organizados de forma sequencial no tempo. Essa característica de sequencialidade faz com que esse tipo de dado tenha características particulares, como a exigência de indexação única no tempo, o fato de que sempre são crescentes e a de que, normalmente, observações adjacentes apresentam uma relação de dependência.

Na indústria, é comum que esse tipo de séries de dados seja utilizado para armazenar os dados históricos coletados a partir de seus processos produtivos. Além disso, alguns contextos de aplicação importantes são a análise desses dados históricos, a previsão de valores futuros a partir de dados antigos e atuais, a detecção de anomalias e a criação de modelos para descrever sistemas.

Segundo Hyndman e Athanasopoulos (2021), as séries temporais podem ser decompostas em três componentes: tendência, sazonalidade e resíduo. O primeiro representa o comportamento de crescimento ou decrescimento dos dados ao longo do tempo, não necessariamente de forma linear. O segundo representa como uma série é afetada por fatores sazonais, isto é, eventos que acontecem constantemente e em frequências iguais e conhecidas. Por fim, o resíduo representa os elementos restantes após extrair as componentes de tendência e sazonalidade da série original.

A equação 1 representa um modelo clássico para decomposição das componentes de séries temporais, chamado de modelo aditivo. Nele, a série y_t é formada pela adição da componentes de tendência T_t, de sazonalidade S_t e de ruído R_t.

$$y_t = T_t + S_t + R_t \tag{1}$$

Para realizar a extração da componente de tendências uma estratégia utilizada é suavização das curvas através de uma função de médias móveis (Hyndman e Athanasopoulos, 2021). A equação de médias móveis simples é descrita na equação 2, na qual M_t representa função de médias móveis, X_t é a função original e k é o número de amostras utilizadas para calcular as médias. O funcionamento dessa técnica é baseado em percorrer a série temporal e calcular as médias dos k últimos valores para cada ponto e com isso as componentes de sazonalidade e de resíduo são eliminados, restando apenas a tendência.

$$M_t = \frac{1}{k} \sum_{j=-k+1}^{0} X_{t+j} \tag{2}$$

Quanto maior o valor de k utilizado, mais suave será a função resultante. Entretanto uma desvantagem desse método é que ao determinar um número fixo para k, as $k-1$ primeiras amostras serão sempre perdidas, já que não possuem valores anteriores suficientes para serem calculadas. Uma alternativa para minimizar esse problema é utilizar uma quantidade variável de amostras para o cálculo das médias, baseada em um período de tempo ou em um determinado número de amostras. Dessa forma, os dados usarão apenas os valores anteriores que estiverem disponíveis, há uma suavização progressiva e uma perda menor dos dados iniciais.

Para a extração da componente de sazonalidade, é preciso remover a tendência da série e uma técnica simples mas eficaz para realizar essa remoção é a das diferenças sucessivas, que percorre a série temporal calculando a diferença entre os pontos adjacentes (Morettin e Toloi, 2006). A equação 3 descreve matematicamente essa função, em que D_t representa a função de diferenças divididas, que pode ser considerada como a componente de sazonalidade e X_t é a série original.

$$D_t(t) = X_t(t) - X_t(t-1) \tag{3}$$

Quanto a componente de resíduo, ela é obtida através da eliminação dos outros dois componentes da série original, sendo necessário apenas fazer uma subtração a partir da equação 1.

Existem diversas outras técnicas para análise de séries temporais, como o uso de modelos de aprendizado de máquina para previsão de valores futuros ou a decomposição de componentes utilizando um modelo multiplicativo, ao invés do modelo aditivo apresentado. Entretanto, a análise de séries temporais ao longo desse trabalho será realizada utilizando apenas os conceitos apresentados nesse capítulo.

2.3 Pré-processamento de dados

Ao coletar os dados salvos em uma base de dados, eles podem estar inadequados para o uso em trabalhos de análise de dados. Alguns dos problemas que podem ser encontrados são: dados faltantes, *features* salvas com tipos inadequados, registros duplicados e presença de *outliers*. Antes de realizar análises exploratórias ou criar modelos de dados é preciso tratar o conjunto de dados para evitar que possíveis inconsistências interfiram nos resultados do trabalho, tornando-o menos preciso e confiável.

Um possível problema em base de dados é a existência de dados faltantes ou registros incompletos, com valores salvos para algumas *features* e para outras não. Para lidar com essa situação, primeiro é necessário investigar o motivo da falta de dados e com base nisso decidir a melhor estratégia, que pode incluir remover uma linha ou coluna inteira com um dado faltante ou realizar uma imputação de dados através de técnicas como uso do valor médio, uso de valores aleatórios, interpolação, uso do valor mais frequente, dentre outras várias possibilidades.

Outro fator que pode também pode afetar os dados negativamente é a presença de *outliers*, que podem ser entendidos como valores ou objetos anômalos em um conjunto de dados ou como medições de um atributo que diferem muito de outras medições do mesmo atributo. Contudo, não é correto assumir de imediato que um *outlier* represente um erro. Esse tipo de medição pode surgir devido a um erro ou um evento. No primeiro caso, as anomalias pode acontecer a partir de um problema de medição ou por falhas humanas. Em uma aplicação IoT, por exemplo, um sensor pode falhar e gerar valores ruidosos. Já os *outliers* de eventos representam valores que realmente aconteceram, mas que por algum fenômeno do mundo real apresenta-

ram medições que se diferenciam do seu estado padrão. Normalmente, as medições extremas causadas por falhas costumam apresentar valores que se diferenciam consideravelmente das outras amostras, enquanto os valores advindos de anomalias não costumam apresentar mudanças abruptas em relação ao restante dos dados e geralmente possuem uma duração maior (Nesa et al., 2018).

De acordo com Bruce et al. (2020), uma forma de identificar *outliers* é através de gráficos *boxplots*, que permitem visualizar onde a maioria dos dados de uma amostra estão localizados, os seus quartis e os valores extremos. Outro método, que não precisa de visualização gráfica, funciona utilizando a distância interquartil, que é o mesmo método utilizado pelos *boxplots* para identificar os limites para definir os *outliers*, como explica Bruce et al. (2020).

Para realizar o tratamento de *outliers* é importante entender primeiro o contexto de aplicação dos dados analisados e o significado dos dados avaliados. Dessa forma, é possível compreender o que causou a anomalia e se ela representa um erro ou um evento fora da normalidade e, a partir dessa avaliação, o analista pode escolher manter os *outliers*, o que geralmente acontece quando eles representam eventos que descrevem a realidade e precisam ser estudados, ou pode realizar operações de remoção ou imputação para substituir os valores extremos.

As técnicas de verificação de completude dos dados e tratamento de *outliers* fazem parte do conjunto de técnicas de limpeza dos dados na fase de pré-processamento. Além das técnicas de limpeza, existem técnicas de transformação dos dados, como normalização, seleção de *features* e discretização, ou até mesmo uso de técnicas de redução dos dados. No entanto, neste trabalho serão utilizadas apenas os artifícios descritos nesta seção e a seleção de *features*, que será descrita em seções posteriores.

3. METODOLOGIA

Este trabalho apresenta um *pipeline* para criação de um modelo de análise de dados para geração de indicadores desempenho de uma estação de tratamento de água. O diagrama da Figura 1 representa esse modelo e, em seguida, cada passo será descrito.

Para realizar um trabalho de análise de dados é preciso definir objetivos claros e entender o contexto que será analisado. A primeira etapa foi compreender como ocorre o processo industrial a ser analisado, quais as principais particularidades dele, dos equipamentos utilizados e as principais saídas. Logo após, foram realizadas a coleta de dados e as entrevistas com os especialistas da área.

Em seguida, foi selecionado um conjunto de dados do processo para ser utilizado na validação do modelo. Nessa fase, a extração dos dados foi feita a partir de uma base de dados do processo com valores históricos já armazenados. A partir desses dados, as *features* de processo são identificadas e estudadas a partir das informações dos especialistas.

O uso adequado de técnicas de pré-processamento aumenta a confiabilidade dos dados analisados e, consequentemente, dos resultados da modelagem. Portanto, os dados coletados na fase anterior foram tratados para que os resultados do modelo proposto neste trabalho sejam efetivos e confiáveis.

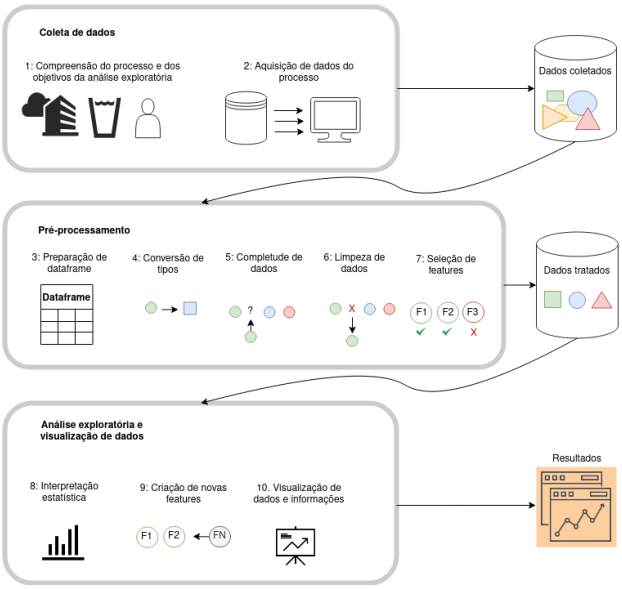


Figura 1. *Pipeline* para o modelo de análise de dados.

Após os dados serem tratados, eles podem ser explorados com mais segurança de que os resultados obtidos serão confiáveis. Com isso, inicia-se a etapa de exploração dos dados.

Nessa etapa, as *features* serão divididas de acordo com sua classificação em quantitativas e qualitativas. Para as *features* quantitativas será realizada uma análise estatística descritiva do conjunto de dados a fim de interpretar as principais medidas de tendência central e de dispersão, e para as qualitativas a interpretação se dará a partir da análise das frequências das *features*.

Por fim, o último passo deste trabalho foi a geração de novas *features* a partir da análise exploratória dos dados e criação dos indicadores do processo no sistema supervisório da planta.

Neste trabalho, os dados analisados foi do processo de produção de água purificada em uma Estação de Tratamento de Água (ETA) em um laboratório farmacêutico.

3.1 Coleta de dados

A primeira etapa do modelo de análise de dados foi realizar entrevistas com os especialistas do laboratório farmacêutico responsáveis pela produção de água para compreender os detalhes do processo realizado na ETA e definir os objetivos da análise exploratória. Como resultado, além de entender o processo, as principais informações obtidas foram as variáveis de saída mais importantes, as máquinas e equipamentos utilizados, os limites operacionais de alguns sensores e a definição de questões de análise.

O processo de produção de água é composta de 4 equipamentos principais: um deionizador, responsável pelo pré-tratamento; um tanque pulmão, que armazena a água entre a etapa de pré-tratamento e a de tratamento; uma osmose reversa, responsável pelo tratamento da água; e o tanque PW (água purificada, do inglês *purified water*), que armazena a água produzida e a disponibiliza para distribuição. As saídas mais importantes do processo são a condutividade, o TOC e o teor microbiológico da água

no tanque PW, dado que essas são as características utilizadas na regulamentação da produção de água para uso farmacêutico.

As principais questões levantadas foram:

- Quais dados influenciam o processo de produção de água purificada?
- O comportamento das variáveis de saída sofre influência de fatores sazonais?

A partir dessas informações, foi feita então a coleta de dados diretamente através da base de dados do laboratório farmacêutico do período de 10 semanas. Os dados foram obtidos em formato *json* e cada registro possuía um registro de tempo e *features* de processo divididas em quatro grupos de acordo com os equipamentos da ETA a qual se referiam.

3.2 Pré-processamento

O primeiro passo do pré-processamento foi ler os dados coletados em formato *json* e transformá-los para o formato de *DataFrame* para trabalhar com manipulação e processamento de dados.

Em seguida, foi feita uma análise dos tipos de dados das colunas do *DataFrame* para identificar necessidades ou oportunidades de realizar conversões de tipo. As mudanças consistiram em corrigir tipos que foram identificados errados na conversão de *json* para *DataFrame*, como valores numéricos lidos como *strings* e em transformar os registros de tempo que estavam em formato numérico, representados em milissegundos, para um formato *datetime64* do pandas, que auxilia na manipulação de dados temporais. Além disso, o tempo deixou de ser uma coluna comum da tabela e foi transformada em índice, para que os dados possam ser representados por ele.

A etapa de completude dos dados iniciou com uma observação da porcentagem de registros faltantes em cada *feature* para avaliar como lidar com possíveis colunas incompletas. Dentre as *features* avaliadas, uma delas apresentou cerca de 18% de dados faltantes, que foi a *feature pH*, que indica o valor do pH da água do tanque pulmão.

Logo após, o processo de limpeza de dados inciou. A primeira ação realizada foi a de remover registros duplicados, ou seja, que foram registrados no mesmo instante de tempo, o que pode ter ocorrido por algum erro no processo de coleta de dados por parte do laboratório.

Também foi feita uma análise de *outliers* de erros através das informações obtidas com os especialistas sobre os limites de operação. Nessa análise, dentre as *features* identificadas, um destaque é que a *feature pH* possuía aproximadamente 65% dos registros fora dos seus limites de operação. Somando isso aos valores faltantes identificados anteriormente, foi entendido que essa *feature* apresentava problemas de coleta no banco e não era ideal para nosso modelo, por isso optou-se por removê-la do estudo. Outras *features* também apresentaram *outliers* fora da faixa de operação, porém em quantidades pequenas, por isso apenas os registros com problemas foram removidos.

A seleção de *features* levou em consideração as informações obtidas com os especialistas da ETA para entender quais delas são essenciais ao processo. Com isso, os critérios a seguir foram definidos:

(1) *Features* consideradas importantes pelos especialistas da ETA devem ser mantidas;
(2) *Features* não indicadas como importantes pelos especialistas e que apresentarem baixa correlação com *features* de maior importância devem ser removidas;
(3) *Features* com valores constantes devem ser removidas;
(4) *Features* com falhas de medição conhecidas devem ser removidas.

Um *DataFrame* auxiliar contendo as seis *features* quantitativas com valores normalizados entre 0 e 1 foi utilizado para avaliar os dados com base no critério 2. A Figura 2 apresenta um gráfico de correlação das seis *features* com base no coeficiente de Pearson.

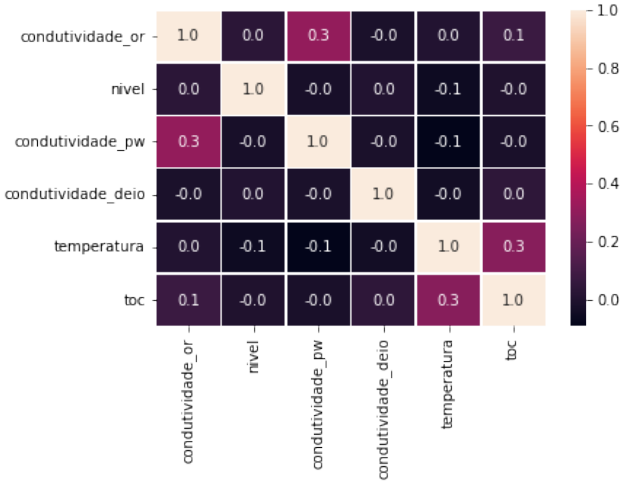


Figura 2. Correlação entre as *features* quantitativas.

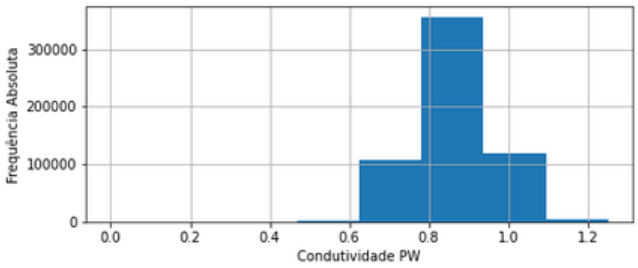
Após avaliar os dados de acordo com os critérios estabelecidos, foram eliminadas 9 *features* dos dados iniciais. Com isso, o *DataFrame* resultante se manteve com 7 colunas.

3.3 Análise exploratória e visualização dos dados

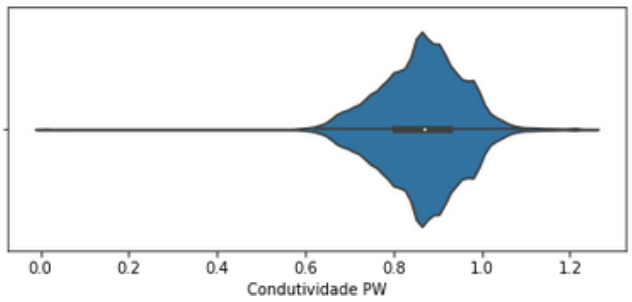
Na primeira etapa da análise exploratória foram utilizadas técnicas para análise e interpretação dos dados através de estatística descritiva. Para isso, as *features* foram divididas em dois grupos de acordo com a classificação delas em quantitativa contínuas ou qualitativas nominais.

Para as *features* quantitativas, foram obtidas as principais medidas de tendência central e dispersão para valores quantitativos contínuos através da função *DataFrame.describe()* e também foram plotados histogramas e gráficos violinos para conseguir visualizar a distribuição dos valores das *features*. A Figura 3 mostra o resultados da geração desses gráficos para a *feature condutividade_pw*, que descreve a condutividade da água no tanque PW.

A *feature condutividade_pw* possui média e mediana próximas e um baixo desvio padrão e o valor máximo atingido na amostra foi de $1.25 \mu S/cm$. É possível observar que a condutividade possui uma distribuição aproximadamente normal, com a concentração de valores sendo um pouco maior acima da média. Outro fator importante é a presença de alguns valores extremos identificados próximos de



(a) Histograma



(b) Gráfico violino

Figura 3. Histograma e gráfico de violino do nível no tanque PW.

zero, o que é muito abaixo da média e deve ser analisado para identificar se eles acontecem por erro de medição ou por eventos anômalos, considerando que o desvio padrão para essa *feature* é baixo. Os mesmos procedimentos e análises foram realizados para as outras *features* contínuas.

Para o grupo das *features* discretas, que possui apenas a *feature estado_valvula*, foi criada a tabela de frequências da Figura 4 e o gráfico de pizza da Figura 5 para observar a incidência com que cada valor ocorre.

estado_valvula	f	fr	fr(%)
0	449419	0.763499	76.349869
1	139212	0.236501	23.650131

Figura 4. Tabela de frequências do estado da válvula de realimentação no tanque PW.

A válvula do tanque PW permanece a maior parte do tempo fechada, estando aberta em aproximadamente 24% dos registros. É provável que isso aconteça por a vazão de saída do tanque ser maior que a de entrada e, consequentemente, ele seca mais rápido do que enche.

A partir dos conhecimentos sobre as *features* utilizadas e dos objetivos de análise do projeto, a segunda etapa da análise exploratória consiste em criar novas *features* que possam auxiliar na análise de processo de produção de água purificada.

As duas primeiras *features* criada foram as *diaNome* e *diaNumero*, que representam, respectivamente, uma *string* com o nome do dia da semana em que um registro foi coletado, em inglês, e um valor numérico inteiro para representar cada dia da semana. A proposta dessas *features* surgiu a partir das informações obtidas pelos especialistas do processo de que os equipamentos passam por determinados procedimentos de limpeza em dias fixos da semana que se repetem semanalmente. Portanto, elas podem ajudar na visualização de dados temporais agrupados por dias da semana.

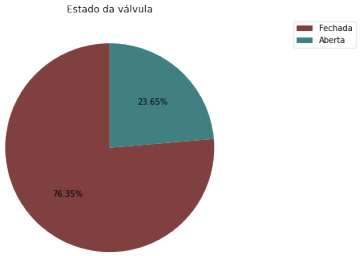


Figura 5. Gráfico de pizza do estado da válvula de realimentação no tanque PW.

Além disso, outra nova *feature* é a *semana*, uma coluna do tipo inteiro que indica em qual das 10 semanas os dados foram coletados, em uma faixa de 1 a 10. O motivo da criação dela é diferenciar e observar o comportamento da produção de água purificada ao longo de cada semana.

Em seguida, foi realizada uma investigação mais aprofundada de quais dos *outliers* identificados através da análise de distribuição são erros ou representam eventos anômalos. Para isso foram realizados dois passos: identificar os valores extremos utilizando a distância interquartil e observar graficamente os pontos para identificar se eles representam um evento que se demora por algum tempo, podendo representar um evento de anomalia nos valores, ou se são esporádicos, o que pode indicar que são medições falhas.

A figura 6 apresenta o resultado desses procedimentos aplicados para a *feature* de condutividade da água no tanque PW. Nela, os pontos em vermelho representam os valores extremos, ou seja, que estão além dos intervalos interquartis. É possível perceber que os *outliers* da parte superior apresentam, em sua maioria, uma continuidade no tempo, por isso foram considerados como eventos anômalos e que devem ser considerados no trabalho. Entretanto, parte dos os valores extremos da parte inferior são exibidos de forma descontínua no tempo, principalmente abaixo de um valor próximo a $0,55 \mu S/cm$ e, portanto, esse valor ficou estabelecido como o limite inferior dos dados e os registros abaixo dele foram removidos.

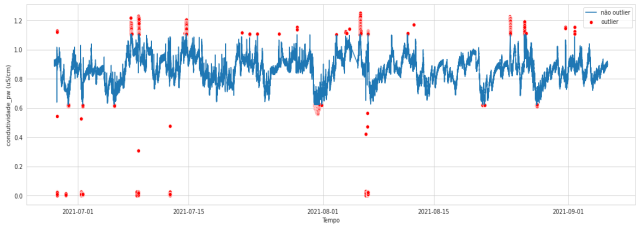


Figura 6. Gráfico da condutividade no tanque PW com *outliers*.

O passo seguinte foi a realização de uma análise das séries temporais das *features* contínuas a partir da extração das componentes de tendência e sazonalidade. Um dos objetivos dessa análise era identificar o comportamento do

sistema a cada semana já que em todas as semanas são realizados processos de sanitização das máquinas.

Originalmente, os dados possuem uma amostragem de 10 segundos, no entanto isso gera uma frequência alta para a extração das componentes das séries por semana, principalmente para a sazonalidade. Para resolver esse problema as séries foram reamostradas para uma periodicidade de 1 dia e os valores foram agregados pela média.

Na extração das componentes de tendência, foi utilizado o comando *DataFrame.rolling().mean()* para calcular as médias móveis das series reamostradas utilizando uma janela com amostras de 7 dias. Já na extração da componente de sazonalidade, foi utilizada a função *DataFrame.diff* para calcular as diferenças sucessivas entre as amostras.

No canto superior da Figura 7 está o gráfico da condutividade no tanque PW com amostragem de 1 dia e ao seu lado está a componente de tendência dessa *feature*, extraída com a janela de 7 dias de amostras. Em baixo, o gráfico de linhas do lado esquerdo representa a sazonalidade da série após removida a tendência da função original e ao lado direito a sazonalidade é exibida novamente, mas com valores agrupados através da média por dia da semana.

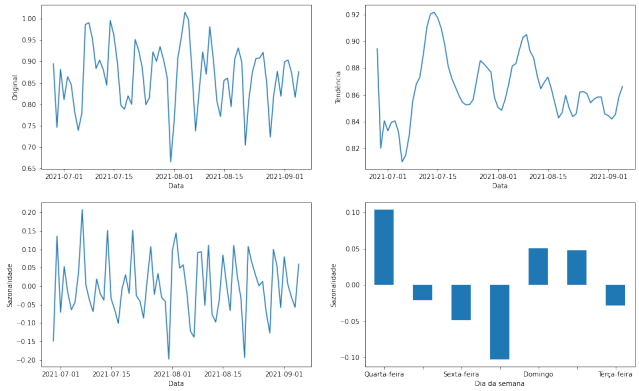


Figura 7. Decomposição dos componentes da série temporal da condutividade no tanque PW.

É possível notar que a *feature* não demonstra seguir uma tendência bem definida, tendo subidas e descidas constantes e irregulares nos valores. Quanto à sazonalidade, ela aparenta ser mais significativa nas quarta-feiras e sábados, com variações crescentes na primeira metade da semana, entre domingo e quarta-feira, com exceção das terças feiras e uma variação decrescente na segunda metade, entre quinta-feira e sábado.

No geral, a avaliação de sazonalidade revelou que existe sim uma relação aparente entre os dias das semana e o comportamento dos dados da produção de água. Essa relação é visível ao perceber que determinados padrões se repetem semanalmente.

4. RESULTADOS

O principal resultado esperado deste trabalho é a criação de novos indicadores através do modelo de análise de dados proposto.

Como resultado das análises, é possível afirmar que o principal fator responsável por definir o comportamento das *features* do processo de produção de água é a realização dos processos de sanitização do deionizador e da osmose reversa semanalmente. As maiores variações estão nos dias que acontecem mudanças de estado nessas duas máquinas.

Com base nisso, foram sugeridos dois novos indicadores: o histórico semanal da porcentagem de tempo dos estados por dia da semana para a osmose reversa e para o deionizador, permitindo identificar a porcentagem de tempo em cada processo ao longo dos dias e como a duração deles afeta o sistema.

Na Tabela 1, é possível visualizar o resultado do indicador de histórico de estados da osmose reversa coletado durante uma semana, que indica a porcentagem de tempo que a máquina passou em cada estado durante cada dia da semana.

Tabela 1. Histórico de estados da osmose reversa durante uma semana

	Seg	Ter	Qua	Qui	Sex	Sáb	Dom
E1(%)	95.0	65.8	90.3	89.7	92.6	51.2	0.0
E2(%)	4.9	27.7	5.7	7.4	4.6	0.0	0.0
E3(%)	0.0	5.6	0.0	0.0	0.0	0.0	0.0
E4(%)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
E5(%)	0.1	0.9	2.8	2.7	2.6	48.8	99.9
E6(%)	0.0	0.0	1.3	0.0	0.1	0.0	0.1
E7(%)	0.0	-0.01	0.0	-0.01	0.1	0.0	-0.01

Além disso, a Tabela 2 apresenta a média e o desvio padrão das porcentagens de tempo que a máquina passou em cada estado, por dia da semana, durante um período de 2 meses de dados coletados.

Tabela 2. Média e o desvio padrão do histórico de estados da osmose reversa durante 2 meses

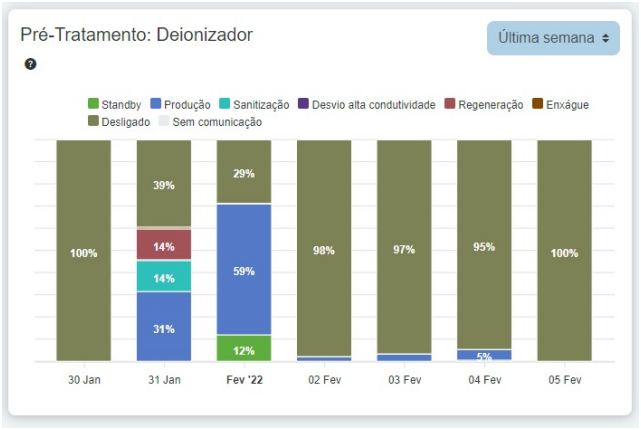
		Seg	Ter	Qua	Qui	Sex	Sáb	Dom
E1(%)	$\bar{X}$	88.2	63.8	84.1	83.8	82.4	82.9	87.1
	σ_x	14.8	18.6	34.1	34.0	36.5	40.6	35.2
E2(%)	$\bar{X}$	1.8	13.4	1.8	1.9	2.5	0.0	0.0
	σ_x	2.2	11.6	1.9	2.1	2.5	0.0	0.0
E3(%)	$\bar{X}$	0.0	2.5	0.0	0.0	0.0	0.0	0.0
	σ_x	0.0	2.4	0.0	0.0	0.0	0.0	0.0
E4(%)	$\bar{X}$	0.0	0.0	0.6	0.0	0.0	0.0	0.0
	σ_x	0.0	0.0	1.6	0.0	0.0	0.0	0.0
E5(%)	$\bar{X}$	1.3	5.1	0.6	1.6	0.6	0.3	0.3
	σ_x	2.0	4.1	1.4	1.1	0.8	0.7	0.9
E6(%)	$\bar{X}$	8.6	15.1	13.0	12.7	14.5	16.8	12.6
	σ_x	14.9	27.6	35.2	35.3	37.7	40.8	35.3
E7(%)	$\bar{X}$	0.1	0.1	0.0	0.0	0.0	0.0	0.0
	σ_x	0.2	0.2	0.0	0.0	0.0	0.0	0.0

Ao apresentar os resultados dos indicadores aos supervisores do processo, os indicadores também foram implementados de forma gráfica no supervisório da planta. A Figura 8 apresenta essa implementação mostrando o histórico de estados em uma semana.

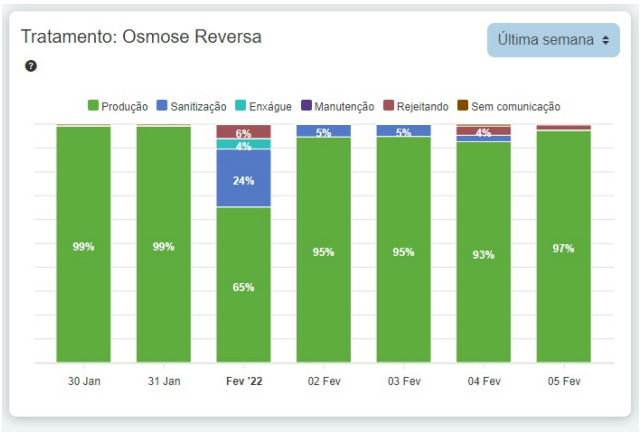
A partir desses resultados obtidos foi possível contribuir para o desenvolvimento de ferramentas para análise de dados para geração de indicadores de um sistema de supervisão da estação de tratamento de água.

5. CONCLUSÃO

Neste trabalho foi proposto e aplicado um modelo de análise de dados descritivo para geração de indicadores



(a) Deionizador



(b) Osmose reversa

Figura 8. Indicadores de histórico semanal dos estados das máquinas.

de monitoramento de processos industriais. Para isso, foi realizado um estudo baseado no case de uma estação de tratamento de água purificada de um laboratório farmacêutico.

Inicialmente, houve um estudo sobre os dados e os processos de produção de água purificada a partir de reuniões com especialistas da área. Com isso, os dados foram coletados a partir da base de dados do sistema de supervisão do laboratório e então passaram por uma etapa de pré-processamento na qual foram utilizadas técnicas para adaptar a base de dados como: preparação do *DataFrame*, conversão de tipos, avaliação de completude dos dados, remoção de dados duplicados e *outliers*, e seleção de *features*. Em seguida, na etapa de análise exploratória e visualização foi feita uma análise estatística dos dados, seguida da criação de novas *features* para análise. Então, foram utilizadas técnicas para explorar os dados, identificando possíveis falhas nas medições e extraindo informações das séries temporais.

Os resultados apresentaram a geração de novos indicadores de desempenho da unidade de produção de água purificada, que fornecem *insights* aos supervisores e operadores do setor.

Os indicadores foram implementados dentro no sistema supervisório para monitoramento contínuo do processo.

AGRADECIMENTOS

Ao Núcleo de Pesquisa em Alimentos e Medicamentos (NUPLAM) da UFRN pela contribuição na pesquisa.

REFERÊNCIAS

Bruce, P., Bruce, A., e Gedeck, P. (2020). *Practical statistics for data scientists: 50+ essential concepts using R and Python*. O'Reilly Media.

Brynjolfsson, E. e McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.

Frank, A.G., Dalenogare, L.S., e Ayala, N.F. (2019). Industry 4.0 technologies: Implementation patterns in manufacturing companies. *International Journal of Production Economics*, 210, 15–26.

Hyndman, R. e Athanasopoulos, G. (2021). *Forecasting: Principles and Practice*. OTexts, Melburn, Australia, 3rd edition. URL `https://OTexts.com/fpp3`.

Lee, J., Kao, H.A., e Yang, S. (2014). Service innovation and smart analytics for industry 4.0 and big data environment. *Procedia Cirp*, 16, 3–8.

McKinney, W. (2019). *Python para análise de dados: Tratamento de dados com Pandas, NumPy e IPython*. Novatec Editora. URL `https://books.google.com.br/books?id=Oj5FDwAAQBAJ`.

Morettin, P.A. e Toloi, C. (2006). Análise de séries temporais. In *Análise de séries temporais*, 538–538.

Nesa, N., Ghosh, T., e Banerjee, I. (2018). Outlier detection in sensed data using statistical learning models for iot. In *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, 1–6. doi:10.1109/WCNC.2018.8376988.

VanderPlas, J. (2016). *Python data science handbook: Essential tools for working with data*. "O'Reilly Media, Inc.".