UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE

CENTRO DE TECNOLOGIA

CURSO DE ENGENHARIA DE COMPUTAÇÃO

Ítalo Oliveira Fernandes

Análise de dados para geração de indicadores de um sistema de produção de água purificada na indústria farmacêutica

Natal – RN

Fevereiro de 2022

Ítalo Oliveira Fernandes

Análise de dados para geração de indicadores de um sistema de produção de água purificada na indústria farmacêutica

Trabalho de Conclusão de Curso de Engenharia de Computação da Universidade Federal do Rio Grande do Norte, apresentado como requisito parcial para a obtenção do grau de Bacharel em Engenharia de Computação

Orientador: Dr. Heitor Medeiros Florêncio

Natal – RN

Fevereiro de 2022

Universidade Federal do Rio Grande do Norte - UFRN
Sistema de Bibliotecas - SISBI
Catalogação de Publicação na Fonte. UFRN - Biblioteca Central Zila Mamede

Fernandes, Italo Oliveira.
Análise de dados para geração de indicadores de um sistema de produção de água purificada na indústria farmacêutica / Italo Oliveira Fernandes. - 2022.
67 f.: il.

Monografia (graduação) - Universidade Federal do Rio Grande do Norte, Centro de Tecnologia, Curso de Engenharia de Computação, Natal, RN, 2022.
Orientador: Prof. Dr. Heitor Medeiros Florêncio.

1. Água purificada - Monografia. 2. Análise de dados - Monografia. 3. Indústria farmacêutica - Monografia. I. Florêncio, Heitor Medeiros. II. Título.

RN/UF/BCZM CDU 681.3

Elaborado por Ana Cristina Cavalcanti Tinôco - CRB-15/262

Ítalo Oliveira Fernandes

Análise de dados para geração de indicadores de um sistema de produção de água purificada na indústria farmacêutica

Trabalho de Conclusão de Curso de Engenharia de Computação da Universidade Federal do Rio Grande do Norte, apresentado como requisito parcial para a obtenção do grau de Bacharel em Engenharia de Computação

Orientador: Dr. Heitor Medeiros Florêncio

Trabalho aprovado. Natal – RN, 11 de Fevereiro de 2022:

Prof. Dr. Heitor Medeiros Florêncio - Orientador
UFRN

Prof. Dr. Jefferson Doolan Fernandes
IFRN

Ma. Gisliany Lillian Alves de Oliveira
ALRN

Natal – RN
Fevereiro de 2022

Dedico esse trabalho primeiramente à minha mãe Regilânia, que sempre lutou enquanto professora para conseguir me criar em meio às dificuldades e para que eu tivesse acesso a uma educação de qualidade. Dedico também a todos os educadores, familiares e amigos que me apoiaram durante toda a minha trajetória escolar e acadêmica.

AGRADECIMENTOS

Agradeço a minha mãe, que além de ter me dado todas as condições para estudar, me deu apoio e suporte quando eu quis mudar de estado a fim de realizar a graduação que eu desejava. Agradeço a ela também por todos os ensinamentos de vida.

Agradeço a Deus que me deu forças em momentos de fraquezas e me ajudou a seguir em frente.

Agradeço a meu professor orientador Heitor Medeiros Florêncio, que além de ser um excelente profissional e líder foi também um grande orientador e amigo durante a realização desse trabalho, sempre disposto a ajudar de qualquer forma e a qualquer momento.

Agradeço a Ítalo Moisés, que foi quem me deu o maior suporte para que eu pudesse focar na preparação desse trabalho na reta final e que me incentivou a continuar todas as vezes que parecia muito difícil.

Agradeço a minha família e a todos os amigos que tornaram a minha jornada acadêmica mais agradável e significativa, em especial a Ana Rute, Dorgival, Gabriel, Inaldo, Jadson, Lara e Lucas Lima, que compartilharam das minhas conquistas e desafios mais significativos.

Agradeço a Include Engenharia, a RN Júnior e a todos os amigos e aprendizados que conquistei no Movimento Empresa Júnior.

Agradeço aos amigos e colegas que conheci no NUPLAM, profissionais incríveis que me ensinaram muitas coisas para a realização desse trabalho e que me deram apoio contínuo desde que entrei na equipe. Agradeço especialmente a Talison, que foi um líder, um tutor, um amigo e uma grande de inspiração profissional.

RESUMO

Este trabalho tem como objetivo gerar indicadores para monitorar os processos da estação de tratamento de água de uma indústria farmacêutica a partir da criação de um modelo de análise de dados, que inclui desde a coleta até a visualização de dados. A água purificada é uma matéria-prima central na indústria farmacêutica e sua produção precisa seguir normas de qualidade estabelecidas por órgãos sanitários reguladores. Com isso, é essencial monitorar essa produção e extrair informações através de ferramentas inteligentes 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 que possam ser considerados de qualidade ao estudo e então explorar e visualizar informações extraídas desses dados. O resultado esperado é a melhoria no acompanhamento da unidade de tratamento de água através da visualização de dados e criação de indicadores.

Palavras-chaves: Água Purificada; Análise de dados; Indústria farmacêutica.

ABSTRACT

This work aims to generate indicators to monitor a pharmaceutical industry water treatment plant processes from the creation of a data analysis model, which includes stages from data collection to data visualization. Purified water is a central raw material in the pharmaceutical industry and its production has to follow quality standards established by regulatory health authorities. Thus, it is essential to monitor this production and extract information through intelligent data tools. 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 an improvement in the monitoring of the water treatment unit through the visualization of data and the creation of indicators.

Keywords: Purified Water; Data Analytics; Pharmaceutical Industry.

LISTA DE ILUSTRAÇÕES

Figura 1 – Visualização de dados com as bibliotecas Matplotlib e Seaborn. 20
Figura 2 – Exemplo didático para entendimento do *Boxplot* - notas finais de uma turma universitária. 25
Figura 3 – Tabela dos tipos de água para uso farmacêutico e parâmetros de qualidade. 28
Figura 4 – Dashboard da ETA no SCANUPLAM. 29
Figura 5 – Arquitetura de automação da ETA. 30
Figura 6 – Pipeline para modelagem de análise de dados da ETA. 33
Figura 7 – Dado armazenado na base de dados do SCANUPLAM. 38
Figura 8 – Informações do *DataFrame* preparado. 39
Figura 9 – Operações na coluna *createdAt*. 40
Figura 10 – Conversão de tipo na *feature pH*. 40
Figura 11 – Resultado das conversões de tipos. 41
Figura 12 – Análise de completude de dados. 41
Figura 13 – Correlação entre as variáveis quantitativas. 43
Figura 14 – Script para remoção de features não selecionadas. 44
Figura 15 – Visualização do *DataFrame* após o pré-processamento. 44
Figura 16 – Descrição estatística das variáveis contínuas. 45
Figura 17 – Histograma e gráfico de violino do nível no tanque PW. 45
Figura 18 – Histograma e gráfico de violino da condutividade no tanque PW 46
Figura 19 – Histograma e gráfico de violino da temperatura no tanque PW. 46
Figura 20 – Histograma e gráfico de violino do TOC no tanque PW. 47
Figura 21 – Histograma e gráfico de violino da condutividade na osmose reversa. . 47
Figura 22 – Tabela de frequências do estado da válvula de realimentação no tanque PW. 48
Figura 23 – Gráfico de pizza do estado da válvula de realimentação no tanque PW. 48
Figura 24 – Visão do *DataFrame* após a adição das novas *features*. 49
Figura 25 – Valores limites para identificar potenciais outliers. 50
Figura 26 – Gráfico da condutividade no tanque PW com *outliers*. 50
Figura 27 – Gráfico da temperatura no tanque PW com *outliers*. 51
Figura 28 – Gráfico da condutividade na osmose reversa com *outliers*. 51
Figura 29 – Gráfico do TOC no tanque PW com *outliers*. 51
Figura 30 – Decomposição dos componentes da série temporal da temperatura no tanque PW. 52

Figura 31 – Decomposição dos componentes da série temporal da condutividade no tanque PW. 53
Figura 32 – Decomposição dos componentes da série temporal da condutividade na osmose reversa. 54
Figura 33 – Decomposição dos componentes da série temporal do TOC no tanque PW. 54
Figura 34 – Decomposição dos componentes da série temporal do nível no tanque PW. 55
Figura 35 – Matriz de gráfico de pontos entre as variáveis quantitativas. 56
Figura 36 – Indicadores de histórico semanal dos estados das máquinas. 58
Figura 37 – Indicador do histórico semanal de produção de água no SCANUPLAM. 59

LISTA DE TABELAS

Tabela 1 – Exemplos de funções importantes para manipulação de *DataFrames* do *pandas*. 19
Tabela 2 – Exemplo de média móvel com número fixo de amostras. 22
Tabela 3 – Exemplo de média móvel com número dinâmico de amostras. 22
Tabela 4 – Resumo sobre algumas variáveis da ETA. 36
Tabela 5 – Mapeamento das variáveis *diaNome* e *diaNumero*. 49
Tabela 6 – Variáveis da ETA obtidas do SCANUPLAM e suas descrições. 66

LISTA DE ABREVIATURAS E SIGLAS

Anvisa	*Agência Nacional de Vigilânica Sanitária*
BPF	*Boas Práticas de Fabricação*
CLP	*Controlador Lógico Programável*
ETA	*Estação de Tratamento de Água*
HVAC	*Heating, Ventilating and Air Conditioning*
IoT	*Internet of Things*
JSON	*JavaScript Object Notation*
MQTT	*Message Queuing Telemetry Transport*
MWNN	*Morlet wavelet Neural Network*
NUPLAM	*Núcleo de Pesquisa em Alimentos e Medicamentos*
SOM	*Self-Organizing Map*
SVM	*Support Vector Machine*
PW	*Purified Water*
REST	*Representational State Transfer*
TOC	*Total Organic Carbon*

SUMÁRIO

1 INTRODUÇÃO 13
1.1 Trabalhos Relacionados 14
1.2 Objetivos 16
1.3 Estrutura do Trabalho 17

2 REFERENCIAL TEÓRICO 18
2.1 *Análise de dados* 18
2.1.1 Python para análise de dados 18
2.1.1.1 Biblioteca NumPy 19
2.1.1.2 Biblioteca pandas 19
2.1.1.3 Bibliotecas matplotlib e seaborn 20
2.1.2 Séries temporais 20
2.1.3 Pré-processamento de dados 23
2.1.3.1 Completude de dados 23
2.1.3.2 Tratamento de *Outliers* 24

3 ANÁLISE DE DADOS: CASE DA PRODUÇÃO DE ÁGUA PURIFICADA NO NUPLAM 27
3.1 Industria farmacêutica: produção de água para uso farmacêutico . . 27
3.1.1 SCANUPLAM 28
3.1.2 Processos da ETA 30
3.2 Pipeline - Modelo de análise de dados da ETA 32
3.2.1 Coleta de dados 32
3.2.2 Pré-processamento 33
3.2.3 Análise exploratória e visualização dos dados 34

4 COLETA DE DADOS E PRÉ-PROCESSAMENTO 36
4.1 Informações do processo 36
4.2 Armazenamento e coleta de dados 37
4.2.1 Obtenção de dados do SCANUPLAM 38
4.3 Pré-processamento 38
4.3.1 Preparação do *DataFrame* 38
4.3.2 Conversão de tipos 39
4.3.3 Completude dos dados 41
4.3.4 Limpeza de dados 42

4.3.5 Seleção de *features* 42

5 ANÁLISE EXPLORATÓRIA E VISUALIZAÇÃO DE DADOS 45
5.1 Interpretações estatísticas 45
5.2 Criação de novas features 48
5.3 Exploração e visualização de dados 49

6 RESULTADOS 58

7 CONCLUSÃO 60

REFERÊNCIAS 62

APÊNDICES 64

APÊNDICE A – VARIÁVEIS DO SCANUPLAM 65

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, de acordo com Brynjolfsson e McAfee (2014), essa grande quantidade de dados tem a capacidade de impactar organizações, sistemas de produção e a sociedade. Nesse contexto, novas tecnologias e conceitos que lidam com esses dados vem ganhando força, como a Internet das Coisas (*Internet of Things* - IoT), Computação em Nuvem, Big Data e Análise de Dados.

Um dos setores que pode evoluir a partir do uso inteligente dos dados é a industria farmacêutica. Dentro do contexto dos laboratórios farmacêuticos, os sistemas de controle das utilidades, como a produção de água para uso farmacêutico, são atividades essenciais para a fabricação dos medicamentos e precisam de acompanhamento contínuo a fim de garantir os padrões de qualidade estabelecidos por órgãos regulamentadores.

Um dos principais insumos na indústria farmacêutica é a água purificada, que é utilizada, por exemplo, na manufatura de medicamentos, limpeza de equipamentos e laboratórios e como solvente em estudos. Dada sua importância, os sistemas de produção e tratamento de água para uso farmacêutico, considerados utilidades nessa indústria, necessitam de cuidados para evitar a contaminação por microrganismos, uma vez que a água exige padrões de qualidade rigorosos para ser considerada adequada para uso e que ela pode agregar compostos e sofrer recontaminação facilmente, conforme afirma a Anvisa (2013).

No Brasil, esses padrões são definidos pela Agência Nacional de Vigilância Sanitária (Anvisa), que é o órgão público responsável por regulamentar a produção de água purificada na indústria farmacêutica através das regras de Boas Práticas de Fabricação (BPF) e por fiscalizar o cumprimento das exigências a partir de processos de validação, nos quais são avaliadas todas as etapas de processos produtivos.

Para atender a essas regulamentações, é importante que essas indústrias utilizem métodos e ferramentas que as permitam analisar e interpretar dados e indicadores de todo o processo de tratamento de água, incluindo coleta, tratamento, armazenamento e distribuição. Dessa forma é possível identificar erros e anomalias rapidamente, prevenir mal funcionamentos e melhorar o embasamento para as tomadas de decisões, de maneira a permitir melhorias na garantia de qualidade na produção.

Nas indústrias, os sistemas de supervisão e aquisição de dados (*Supervisory Control And Data Acquisition* - SCADA) são comumente utilizados para realizar atividades de operação de sistemas, coleta e armazenamento de dados e visualização de dados históricos,

de acordo com Lamb (2015). Contudo, Lee, Kao e Yang (2014) afirmam que por falta de ferramentas inteligentes para análise de dados, muitos sistemas de manufatura ainda não estão preparados para lidar com grandes volume de dados produzidos. Dessa forma, existem organizações que produzem e armazenam grandes volumes de dados, mas não conseguem extrair informações úteis ao processo produtivo. Para lidar com essa questão, vem ganhando força o conceito da Indústria 4.0, que, segundo Frank, Dalenogare e Ayala (2019), tem como elemento central a manufatura inteligente e utiliza tecnologias digitais para coletar e processar dados heterogêneos de diversos sensores e dispositivos IoT em tempo real, criando informações úteis ao processo industrial.

Nesse contexto, é possível aproveitar essas tecnologias como um método para auxiliar as indústrias do ramo farmacêutico a melhorar seus processos e garantir a qualidade de sua produção diante dos padrões estabelecidos por órgãos regulamentadores. Em vista dessa oportunidade, esse trabalho propõe a geração de indicadores de monitoramento da produção de água purificada para uso em laboratórios e indústrias farmacêuticas a partir da criação de um modelo de análise de dados. Para isso, o trabalho foi aplicado a partir de dados obtidos da Estação de Tratamento de Água (ETA) do laboratório farmacêutico NUPLAM (Núcleo de Pesquisa em Alimentos e Medicamentos) da Universidade Federal do Rio Grande do Norte (UFRN).

1.1 Trabalhos Relacionados

Alguns trabalhos se propõem a criar modelos analíticos nas indústrias das áreas farmacêutica e geral, principalmente para a predição de erros. Além disso, também existem estudos relacionados ao tratamento de água em outros contextos, mas ainda tendo a análise de dados como um elemento essencial.

Em Garmaroodi et al. (2021), os autores desenvolvem dois modelos de detecção de anomalias baseados em mineração de dados e aprendizado de máquina aplicados a uma estação de tratamento de água *CHRIST Osmotron*. Com a criação desses modelos, eles esperam que a detecção de anomalias seja mais precisa e possua um tempo de resposta menor, permitindo a prevenção de novos erros e ações rápidas para evitar danos na estação ou até mesmo o desligamento dela. Foram realizadas as etapas de coleta de dados de sensores IoT, pré-processamento dos dados, treinamento do modelo e visualização de dados.

Durante a etapa de coleta de dados foram utilizados seis sensores IoT ligados a diferentes etapas do processo formando um conjunto de amostras normais e com erros. Já no pré-processamento, foram aplicadas as técnicas de limpeza de dados, normalização, remoção de ruídos e então os dados passaram para etapa de treinamento, na qual foram utilizadas as abordagens a seguir. A primeira é baseada em dados e utiliza técnicas de

classificação supervisionada de aprendizado de máquina como máquinas de vetores de suporte (*Support Vector Machine* - SVM) e árvores de decisão para identificar possíveis erros e é recomendada em casos nos quais o comportamento do sistema é desconhecido, muito complexo ou que o modelo está incompleto ou inacessível. Já a segunda é baseada no modelo do sistema e utiliza apenas o conjunto de dados normais, aplicando uma rede neural artificial que identifica o comportamento normal do sistema e define limites adaptativos para os quais, além deles, os dados são considerados como anomalias. Ambos os resultados foram positivos, porém o segundo apresentou a vantagem de conseguir identificar problemas sem a necessidade de conhecer os dados com erro previamente. Por fim, são apresentados gráficos com resultados de cada modelo proposto.

Nesse trabalho, o contexto principal é do tratamento de água para uso farmacêutico e o foco é no desenvolvimento das abordagens propostas para a predição de dados e apesar de haver uma seção de pré-tratamento, ela não é muito detalhada. Além disso, não há nenhuma etapa voltada à parte de análise exploratória dos dados com o objetivo de se aprofundar na descrição analítica das amostras utilizadas e avaliar o conjunto de amostras antes da aplicação no modelo.

Já em Zhang et al. (2017), é proposto o uso de um algoritmo utilizando janelas móveis duplas para detecção de anomalias em um sistema de tratamento de água de um rio com transmissão de dados em tempo real para um *data center*. Esse ambiente apresenta desafios diferentes em relação ao ambiente farmacêutico, com grandes chances de problemas como falta de bateria dos equipamentos, comprometimento do hardware dos sensores e influência do meio ambiente nas medições.

A pesquisa de Zhang et al. (2017) coletou dados do pH da água durante um período de 3 meses para propor um algoritmo de detecção de anomalias é baseado em um modelo de combinação linear autorregressiva. Esse modelo utiliza um intervalo preditivo com janelas móveis duplas e uma estratégia de verificação de retrocesso. Para verificar a eficiência da proposta o algoritmo é comparado a outros dois que também são utilizados em processos de detecção de anomalias, o AD e o ADAM. Os resultados apresentaram que a técnica desenvolvida conseguiu diminuir o número de alarmes falso positivos e apresentou uma taxa de detecção de anomalias melhor que os outros algoritmos usados na comparação.

De forma semelhante ao anterior, esse trabalho não se aprofunda em descrever as etapas do projeto até a parte de análise dos dados, tendo como foco o desenvolvimento da estratégia para detecção de anomalias. Além disso, nele só é monitorada uma variável, por isso não existe um estudo de possíveis relações entre ela e outros fatores que afetam a qualidade da água.

Além desses, outras áreas também desenvolvem trabalhos com inteligência de dados

para tornar suas indústrias mais eficiente. Wen e Xie (2021) realizam um estudo para avaliar a performance da geração de energia em duas turbinas de vento chinesas a partir dos dados de um sistema SCADA. Os autores combinam técnicas de modelos preditivos e de análise do desempenho das turbinas.

A coleta dos dados nesse caso foi feita a partir do sistema supervisório que monitora mais de 50 parâmetros para cada turbina e foram utilizados dados de aproximadamente 3 meses. Após a coleta, os dados passaram pelo processo de pré-processamento, no qual as variáveis que seriam utilizadas no trabalho foram selecionadas a partir de análise do coeficiente de correlação de Pearson. Em seguida, foi realizada uma limpeza nos dados para remover registros fora da faixa de operação dos sensores. Os autores utilizaram uma rede neural *Morlet wavelet* (*Morlet wavelet Neural Network* - MWNN) para realizar a geração do modelo preditivo de geração de energia nas turbinas. Por fim, a partir do modelo preditivo, os modelos *self-organizing map* (SOM) e de Markov foram utilizados para realizar uma análise de dados de performance. Como resultado, foi proposta uma nova variável chamada de índice de anormalidade, que pode ser utilizada para metrificar o índice de anormalidade em turbinas de vento.

Diferente dos outros, o trabalho de Wen e Xie (2021) se aprofunda em todas as etapas do processo analítico e não apenas na construção de um modelo preditivo, abordando de foram clara os métodos utilizados para tratar e explorar os dados. Além disso, o principal resultado obtido foi a criação de um novo indicador de performance, ao invés da avaliação de uma técnica de predição de anomalias. Entretanto, o seu escopo se distancia do monitoramento de unidades de produção água para uso farmacêutico.

A partir dessas referências, é possível sugerir que há pesquisas na área de análise de dados para o tratamento de água, porém elas variam entre diferentes campos de aplicação, o que também muda as exigências dos estudos. Outro ponto importante é que essas pesquisas aparentam estar focadas principalmente em desenvolver técnicas para detectar anomalias nos sistemas de tratamento. Também existem outros estudos que vão além da modelagem preditiva para detecção de anomalias e permitem descrever, analisar e melhorar a performance de sistemas industriais, entretanto ainda há poucos deles na meio industrial farmacêutico, que precisa seguir um alto padrão de confiabilidade nos processos.

1.2 Objetivos

O objetivo geral deste trabalho é gerar indicadores para monitoramento contínuo de uma ETA a partir de um modelo de análise de dados que utiliza as principais variáveis do processo de produção de água purificada e os conhecimentos obtidos com os especialistas da ETA.

Ademais, os objetivos específicos do trabalho são:

- Obter informações sobre o processo de tratamento de água a partir de entrevistas com especialistas;
- Discutir e aplicar técnicas de pré-processamento e análise exploratória de dados a partir da base de dados do sistema de supervisão central do NUPLAM;
- Propor indicadores que possam fornecer *insights* para melhorar os processos de controle de qualidade da água do NUPLAM.

1.3 Estrutura do Trabalho

Este capítulo introduz o trabalho desenvolvido com o contexto de aplicação, trabalhos relacionados e a definição dos objetivos de pesquisa. Em seguida, o Capítulo 2 descreve os conceitos teóricos ligados a projetos de análise de dados, bem como as principais técnicas e ferramentas utilizadas para esse trabalho. O Capítulo 3 discute a importância da água na indústria farmacêutica e descreve o funcionamento do processo realizado por um sistema de tratamento de água purificada e então apresenta o *pipeline* utilizado para a modelagem analítica dos dados. Na sequência, o Capítulo 4 aborda a execução e resultados das etapas de coleta e pré-processamento de dados, enquanto o Capítulo 5 faz o mesmo para as etapas de análise exploratória e visualização de dados. O Capítulo 6, então, apresenta os principais resultados obtidos na modelagem de análise de dados. Por fim, o Capítulo 7 traz as considerações finais, com as principais conclusões, as contribuições do trabalho e os possíveis trabalhos futuros.

2 REFERENCIAL TEÓRICO

Neste capítulo serão apresentados os conceitos teóricos sobre análise de dados necessários para o desenvolvimento desse trabalho. Para isso, é importante discutir o ciclo de vida padrão de um projeto de análise de dados, algumas técnicas de análise e visualização de dados relevantes para esse estudo e o uso da linguagem de programação Python e suas bibliotecas como ferramentas de apoio para a aplicação dessas técnicas.

2.1 Análise de dados

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. De acordo com Ahmed et al. (2017), exitem três tipos de análises de dados: descritivas, preditivas e prescritivas.

O modelo de análise de dados descritiva tem como objetivo a compreender o passado e o presente, descrevendo o que já aconteceu e o que está acontecendo. Já a análise preditiva tem a finalidade de prever o futuro, tentando descobrir o que irá acontecer e o porquê. Por fim, a análise prescritiva tem como intenção sugerir aos tomadores de decisão o que deve ser feito e o porquê.

2.1.1 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). Além disso, para McKinney (2019), ela se tornou uma das principais linguagens para ciência de dados e aprendizado de máquina frente a outros concorrentes devido ao suporte melhorado dela para bibliotecas, o que em conjunto com sua robustez permite a criação de aplicações poderosas para dados.

O Python possui diversas bibliotecas implementadas para lidar com tratamento, processamento e visualização de dados de forma simples e ao mesmo tempo robusta. Dessa forma, seu uso se torna atrativo dadas às facilidades e ao poder que fornece para cientistas de dados.

A seguir, serão apresentadas algumas das principais bibliotecas Python utilizadas para análise de dados.

2.1.1.1 Biblioteca NumPy

O pacote *NumPy* é uma das ferramentas centrais para a computação numérica científica com Python (HARRIS et al., 2020). Sua principal função é permitir criar de estruturas de dados em formatos de *arrays* multidimensionais chamados de *ndarrays*. Essas estruturas se diferenciam dos vetores *built-in* do Python por serem mais rápidas e eficientes e por fornecer diversas operações como transformação de dados, álgebra linear e manipulação de todos os elementos de uma única vez.

Dada sua importância na construção de estruturas de dados eficientes, é comum que outras bibliotecas de ciência de dados utilizem os elementos do tipo *ndarray* como base para suas funcionalidades.

2.1.1.2 Biblioteca pandas

O pandas é uma poderosa ferramenta para manipulação e análise de dados (REBACK et al., 2020; MCKINNEY, 2010). De acordo com McKinney (2019), o *pandas* oferece estruturas de dados de alto nível e funções projetadas para tornar o trabalho com dados estruturados ou tabulares rápido, fácil e expressivo.

Essa biblioteca possui duas estruturas de dados principais: *Series* e *DataFrame*. As *Series* funcionam como *arrays* unidimensionais rotulados, diferindo dos *ndarrays* do *NumPy* por aqueles possuírem índices explícitos e que podem ser de qualquer tipo, enquanto esses possuem índices implícitos do tipo inteiro. Já os *DataFrames* são utilizados como estrutura de dados tabular com o uso de linhas e colunas indexadas. A Tabela 1 apresenta algumas instruções para manipular *DataFrames*.

Função	Descrição
pd.DataFrame	Cria um objeto *DataFrame*.
pd.DataFrame.head	Retorna os n primeiros elementos do *DataFrame*.
pd.DataFrame.info	Exibe informações do *DataFrame* como colunas e valores não nulos.
pd.DataFrame.describe	Exibe as principais medidas de estatística descritiva para as colunas do *DataFrame*.
pd.DataFrame.merge	Mescla dois *DataFrames* em um.
pd.DataFrame.dropna	Remove linhas ou colunas com valores nulos.
pd.DataFrame.fillna	Preenche valores nulos.
pd.DataFrame.rolling	Fornece uma janela móvel para percorrer as colunas do *DataFrame*.

Tabela 1 – Exemplos de funções importantes para manipulação de *DataFrames* do *pandas*.

A biblioteca fornece funções de alto nível para trabalhar com dados de forma semelhante às tabelas de planilhas ou bancos de dados relacionais, dentre elas estão atividades essenciais como ler, filtrar, manipular, limpar e agregar dados. Além disso, também é fornecido suporte para o trabalho com séries temporais, que são os principais tipos de dados quando se trabalha com internet da coisas. Com isso, ela é de grande auxílio para tratar e preparar as bases de dados para a etapa de análise exploratória.

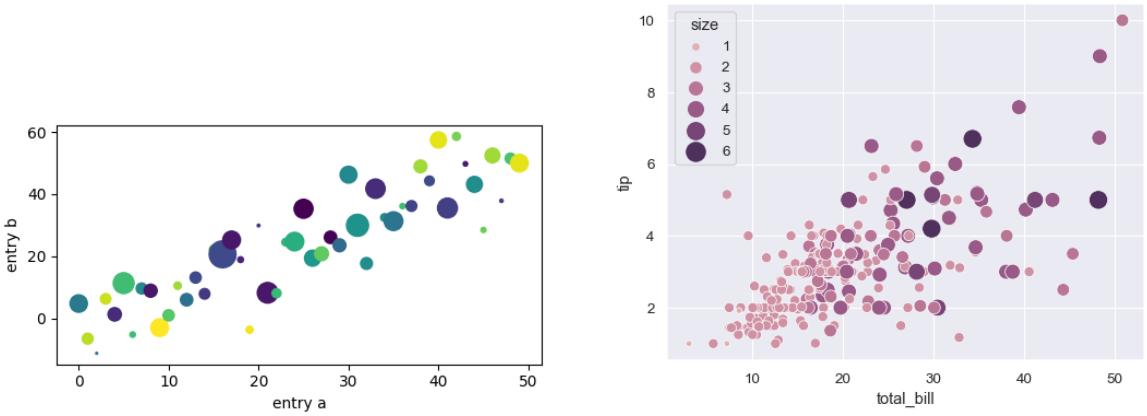
Ao trabalhar com o *pandas* para ciência de dados, alguns termos são comumente utilizados. As variáveis são chamadas de colunas ou *features* e cada elemento da amostra é uma linha ou registro. Para cada registro, existe um rótulo chamado de *index*. Ao longo desse trabalho essas nomenclaturas serão amplamente utilizadas.

2.1.1.3 Bibliotecas matplotlib e seaborn

Tanto a *matplotlib* quanto a *seaborn* são bibliotecas Python utilizadas para visualização de dados (HUNTER, 2007; WASKOM, 2021). Elas permitem plotar gráficos essenciais como os gráficos de linha, gráfico de barras, gráficos de pontos, dentre outros.

A primeira é a biblioteca de visualização mais popular da linguagem Python, o que segundo McKinney (2019) faz com que ela tenha uma boa integração com o ecossistema de análise de dados do Python. Além disso, ela também usa comandos similares ao da linguagem matlab, auxiliando assim usuários que já conhecem essa ferramenta.

Seaborn é uma ferramenta construída em cima da *matplotlib* e a complementa na exibição estatística de dados, além de fornecer mais temas visuais e permitir um embelezamento dos gráficos. A Figura 1 apresenta um exemplo comparativo da plotagem de gráficos de pontos nas duas bibliotecas.



(a) Gráfico de pontos Matplotlib (b) Gráfico de pontos Seaborn

Figura 1 – Visualização de dados com as bibliotecas Matplotlib e Seaborn.

Fonte: (HUNTER; DALE, 2007; WAKSOM, 2017).

2.1.2 Séries temporais

Nessa seção, serão abordados os conceitos fundamentais para analisar e descrever séries temporais com base em dados históricos.

Conforme Box et al. (2015), 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.

Essas especificidades permitem que as séries temporais sejam utilizadas para armazenar dados históricos em diversos contextos, por exemplo no campo da saúde, economia e engenharia. 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** - representa o comportamento de crescimento ou decrescimento dos dados ao longo do tempo, não necessariamente de forma linear. Quando existe uma visualização clara desse comportamento ele é chamado de padrão de tendência, mas além dele a componente de tendência também pode incorporar padrões cíclicos, que são representados por momentos de ascensão e queda nos valores que não necessariamente possuem frequências fixas.
- **Sazonalidade** - representa o comportamento de sazonalidade de como uma série é afetada por fatores sazonais, isto é, eventos que acontecem constantemente e em frequências iguais e conhecidas.
- **Resíduo** - representa os elementos restantes após extrair as componentes de tendência e sazonalidade da série original.

A equação 2.1 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.

$$y_t = T_t + S_t + R_t \tag{2.1}$$

Para realizar a extração da componente de tendências uma estratégia utilizada é suavização das curvas através de uma função de médias móveis, como aborda Hyndman e Athanasopoulos (2021). A equação de médias móveis simples é descrita na equação 2.2, 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.

$$M_t = \frac{1}{k} \sum_{j=-k+1}^{0} X_{t+j} \tag{2.2}$$

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. A tabela 2 mostra um exemplo de aplicação dessa função com $k = 3$.

Data	Série original	Médias móveis
01/01/2021	1	NaN
02/01/2021	2	NaN
03/01/2021	3	2
04/01/2021	4	3
05/01/2021	5	4

Tabela 2 – Exemplo de média móvel com número fixo de amostras.

Quanto maior o número de valores utilizados para calcular a média, 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. Isso significa que se por um lado é possível usar um valor alto para k para isolar ainda mais a tendência, por outro isso pode acarretar na perda de muitas informações referentes às amostras iniciais eliminadas.

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.

Para exemplificar, a tabela 3 mostra o cálculo de medias móveis com uma janela de três dias. Nela, a primeira média utiliza apenas o valor do primeiro dia, na segunda, as duas primeiras medições são utilizadas e a partir do terceiro dia as médias são calculadas para o período de três dias.

Data	Série original	Médias móveis
01/01/2021	1	1
02/01/2021	2	1.5
03/01/2021	3	2
04/01/2021	4	3
05/01/2021	5	4

Tabela 3 – Exemplo de média móvel com número dinâmico de amostras.

Dessa forma há uma suavização progressiva e uma perda menor dos dados iniciais, apesar de eles utilizarem menos amostras para o cálculo da média.

Para a extração da componente de sazonalidade, é preciso remover a tendência da série e, conforme Morettin e Toloi (2006), 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.

A equação 2.3 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.

$$D_t(t) = X_t(t) - X_t(t-1) \tag{2.3}$$

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 2.1.

A partir da biblioteca *pandas* é possível acessar diversas funções de manipulação de tempo que auxiliam na análise de *DataFrames* de séries temporais. Por exemplo, existem as funções *pd.DataFrame.rolling*, que cria uma janela móvel para percorrer séries, auxiliando no cálculo das médias móveis e *pd.DataFrame.diff*, que calcula a diferença entre os dados consecutivos da mesma coluna do *DataFrame*.

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 nos próximos capítulos desse trabalho será realizada utilizando apenas os conceitos descritos nesse capítulo.

2.1.3 Pré-processamento de dados

Por vezes, ao coletar os dados salvos em uma base de dados, eles podem não estar adequados para o uso em um trabalho de análise de dados. Alguns dos problemas que podem ser encontrados são: dados faltantes, variáveis salvas com tipos inadequados ou registros duplicados. Com isso, antes de realizar análises exploratórias ou criar modelos de dados é preciso limpar o conjunto de dados para evitar que possíveis inconsistências possam interferir nos resultados do trabalho, tornando-o menos preciso e confiável.

Nos itens a seguir serão apresentadas algumas das principais técnicas utilizadas na etapa de pré-processamento de dados.

2.1.3.1 Completude de dados

Um dos problemas possíveis para conjuntos de dados não tratados é que podem existir registros incompletos com valores salvos para algumas variáveis e para outras não. Para identificar essa situação é possível ver o número de registros vazios para cada coluna de um *DataFrame df* utilizando a função *df.isnull().count()*.

Depois de identificados, os dados nulos precisam ser analisados e tratados. Algumas estratégias para lidar com esses dados faltantes são remover os registros incompletos,

remover o atributo que possui dados faltantes da análise ou imputar os valores ausentes por um valor fixo, como o valor médio da variável ou o último valor não nulo registrado.

Na primeira e segunda opção, um problema da estratégia é que são perdidas informações de registros que foram medidos, enquanto na terceira opção um ponto negativo é que os valores inseridos não são reais, o que pode adicionar incerteza na análise. A escolha de qual método utilizar depende de cada caso e dos objetivos do analista de dados.

Em Python, é possível remover linhas inteiras com valores nulos em um *DataFrame* através do comando *dropna()* ou colunas adicionando o parâmetro *"axis=1"* para a função. De forma similar, existe o comando *fillna()* que mapeia valores nulos e os substitui por um novo.

2.1.3.2 Tratamento de *Outliers*

De acordo com Moreira, Carvalho e Horváth (2019), *outliers* são valores ou objetos anômalos em um conjunto de dados. Uma outra forma de entendê-los é como medições de um atributo que diferem muito de outras medições do mesmo atributo.

Apesar disso, não é correto assumir de imediato que um *outlier* represente um erro. Conforme Nesa, Ghosh e Banerjee (2018), 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 apresentaram 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.

Segundo Bruce, Bruce e Gedeck (2020), uma forma simples de identificar *outliers* é o uso de gráficos *boxplots*, que permitem visualizar onde a maioria dos dados de uma amostra estão localizados, os seus quartis e os valores extremos. A Figura 2 mostra um exemplo de *boxplot* das notas do exame final de uma turma de alunos universitários, considerando que o professor estabelece as notas no intervalo de 0 a 10.

Nesse exemplo, é possível identificar dois valores anômalos, que se encontram além dos limites externos da caixa. Para o dado mais a esquerda, de valor zero, é provável que represente um evento real relacionado a um aluno que apresentou baixo desempenho no exame e deve ser levado em consideração. Entretanto, o ponto extremo a direita provavelmente é um erro, já que o valor 50 foge do intervalo estabelecido pelo professor inicialmente. Nesse caso, o erro pode ter ocorrido por um erro de digitação.

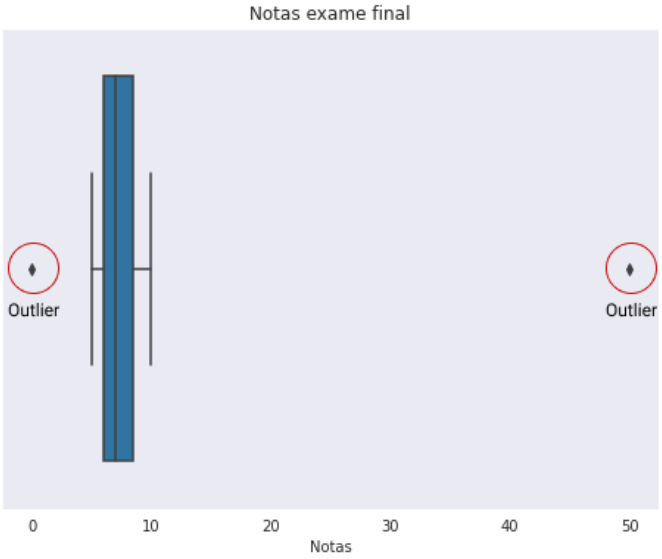


Figura 2 – Exemplo didático para entendimento do *Boxplot* - notas finais de uma turma universitária.

Fonte: Próprio autor

Outro método simples de identificar esses valores e que não precisa de visualização gráfica é utilizando a distância interquartil, que é o mesmo método que os *boxplots* utilizam para identificá-los, como explica Bruce, Bruce e Gedeck (2020). A distância interquartil (IQ) é a diferença entre o terceiro quartil (Q3), valor para o qual 75% das amostras estão abaixo dele e o primeiro quartil (Q1), valor para o qual 25% das amostras estão abaixo dele, de acordo com a equação 2.4.

$$IQ = Q3 - Q1 \tag{2.4}$$

Conhecendo o valor IQ, considerasse que os valores que estejam n vezes IQ abaixo do primeiro quartil e n vezes IQ acima do terceiro quartil são *outliers*, sendo n um número real comumente utilizado como 1,5. As equações 2.5 e 2.6 descrevem a fórmula pra cálculo dos limites superiores e inferiores para definição de *outliers*.

$$Limite\ superior = Q3 + 1,5 * IQ \tag{2.5}$$

$$Limite\ inferior = Q1 - 1,5 * IQ \tag{2.6}$$

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.

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 similares às realizadas no tratamento de dados faltantes, removendo todos os registros que possuam uma medida anômala ou substituindo 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 as técnicas de limpeza descritas nesta seção e a seleção de *features*, que será descrita em seções posteriores.

3 ANÁLISE DE DADOS: CASE DA PRODUÇÃO DE ÁGUA PURIFICADA NO NUPLAM

Este capítulo busca discutir a importância da água purificada para a indústria farmacêutica e descrever o funcionamento dos processos industriais na estação de tratamento de água purificada do NUPLAM, incluindo o funcionamento do sistema de supervisão da ETA. Em seguida, será proposto um pipeline para a definição de um modelo de análise de dados descritivo para a ETA que servirá como base para a realização deste trabalho.

3.1 Industria farmacêutica: produção de água para uso farmacêutico

Um conceito importante dentro da indústria é o de utilidades, que representam sistemas responsáveis pela produção de insumos essenciais no desenvolvimento de outras atividades produtivas. Na indústria farmacêutica, algumas utilidades são: os sistemas de ar comprimido, sistemas de HVAC (*Heating, Ventilating and Air Conditioning*) e os sistemas de tratamento de água para uso farmacêutico, como a água purificada.

Enquanto insumo farmacêutico, a água purificada é utilizada tanto para o processo de fabricação de medicamentos não parentais, ou seja, que não são injetáveis, quanto na limpeza de equipamentos e laboratórios, solução de reagentes e outras aplicações. Conforme a Anvisa (2013), o controle da contaminação da água para uso farmacêutico é fundamental, uma vez que a água tem grande susceptibilidade para agregar compostos diversos e para sofrer recontaminação, mesmo após a etapa de purificação. Com isso, processos inadequados podem resultar em produção de água com qualidade pobre, capaz de colocar em risco a eficácia e a segurança dos fármacos e as indústrias farmacêuticas precisam garantir um processo de tratamento de água eficiente para alcançar níveis de qualidade aceitáveis.

No Brasil, o órgão responsável por definir e fiscalizar os padrões de qualidade da água para uso farmacêutico é a Anvisa, através das normas técnicas de Boas Práticas de Fabricação (BPF) e da 5° Farmacopéia Brasileira, com base em recomendações internacionais. Para garantir que essas normas sejam cumpridas, a agência exige que os sistemas de purificação de água para uso farmacêutico passem por um processo de validação para garantir o controle e qualidade em todo processo produtivo, que inclui as etapas de produção,

armazenamento e distribuição.

Existe mais de um tipo de água para uso farmacêutico: água purificada, água ultrapurificada e água para injetáveis. A água purificada possui três parâmetros críticos: a condutividade, o TOC (*Total Organic Carbon*) e a contagem de microbiológicos. A Figura 3 apresenta um comparativo entre os diferentes tipos de águas utilizados no meio farmacêutico e as recomendações da Anvisa para os seus parâmetros de qualidade, bem como exemplos de aplicação para cada tipo.

Tipo de Água	*Características*	*Parâmetros críticos sugeridos*	*Exemplos de Aplicação*
Água Potável	Obtida de mananciais ou da rede de distribuição pública.	Possui legislação específica.	Limpeza em geral e fonte de alimentação de sistemas de tratamento.
Água Reagente	Água potável tratada por deionização ou outro processo. Possui baixa exigência de pureza.	Condutividade de 1 a 5,0 µS/cm a 25,0 °C ± 0,5 °C (resistividade > 0,2 MΩ-cm) COT < 0,20 mg/L	Lavagem de material, abastecimento de equipamentos, autoclaves, banho-maria, histologia, usos diversos.
Água purificada	Níveis variáveis de contaminação orgânica e bacteriana. Exige cuidados de forma a evitar a contaminação química e microbiológica. Pode ser obtida por osmose reversa ou por uma combinação de técnicas de purificação a partir da água potável ou da reagente.	Condutividade de 0,1 a 1,3 µS/cm a 25,0 °C ± 0,5°C (resistividade > 1,0 MΩ-cm); COT < 0,50 mg/L; Contagem total de bactérias < 100 UFC/mL Ausência de Pseudomonas e outros patogênicos.	Produção de medicamentos e cosméticos em geral, farmácias, lavagem de material, preparo de soluções reagentes, meios de cultura, tampões, diluições, microbiologia em geral, análises clínicas, técnicas por Elisa, radioimunoensaio, aplicações diversas na maioria dos laboratórios, principalmente em análises qualitativas ou quantitativas menos exigentes (em %). Em CLAE (em %).
Água para injetáveis	Água purificada tratada por destilação ou processo similar.	Atende aos requisitos químicos da água purificada e exige controle de endotoxina, partículas e esterilidade. Contagem microbiológica < 10UFC/100 mL. Endotoxinas < 0,25 UI de endotoxina/mL; COT < 0,50 mg/L.	Como veículo ou solvente de injetáveis, fabricação de princípios ativos de uso parenteral, lavagem final de equipamentos, tubulação e recipientes usados em preparações parenterais. Usada como diluente de preparações parenterais.
Água ultrapurificada	Para análises que exigem mínima interferência e máxima precisão e exatidão. Baixa concentração iônica, baixa carga microbiana e baixo nível de carbono orgânico total. Água purificada tratada por processo complementar.	Condutividade de 0,055 a 0,1 µS/cm a 25,0 °C ± 0,5 °C (resistividade > 18,0 MΩ-cm) COT < 0,05 mg/L (alguns casos < 0,003 mg/L) Contagem total de mesófilos < 1 UFC/100mL(se utilizada para fins farmacêuticos).	Dosagem de resíduos minerais ou orgânicos, endotoxinas, preparações de calibradores, controles, SQR, espectrometria de absorção atômica, ICP/IOS, ICP/MS, espectrometria de massa, procedimentos enzimáticos, cromatografia a gás, CLAE (ppm ou ppb), biologia molecular e cultivo celular etc.Eventualmente em preparações farmacêuticas que requeiram água de alta pureza

COT = Carbono orgânico total;
UFC/100 mL = Unidades formadoras de colônias; população microbiológica viável

Figura 3 – Tabela dos tipos de água para uso farmacêutico e parâmetros de qualidade.

Fonte: (ANVISA, 2010)

Em vista do rigor estabelecido por essas normas, é importante que as indústrias farmacêuticas sejam capazes de monitorar constantemente o processo produtivo da água purificada. Dessa maneira, é plausível assumir que o uso de ferramentas voltadas para a coleta e monitoramento contínuo de dados da produção é uma estratégia importante para essas organizações, bem como a aplicação de técnicas de análise de dados para transformá-los em informações que permitam avaliar o desempenho produtivo, interpretar os eventos, identificar falhas e anomalias rapidamente, e tomar melhores decisões de gerenciamento.

A seguir serão apresentados o SCANUPLAM, o sistema de supervisão dos processos industriais do NUPLAM, e as etapas de fabricação de água purificada na ETA NUPLAM.

3.1.1 SCANUPLAM

O SCANUPLAM é um sistema de supervisão que monitora todos os processos industriais nos diversos setores e áreas produtivas do NUPLAM. Esse software foi cri-

ado como parte de um projeto de digitalização do laboratório farmacêutico. A Figura 4 mostra a interface do SCANUPLAM com o *dashboard* da ETA, apresentando os dados monitorados em tempo real.

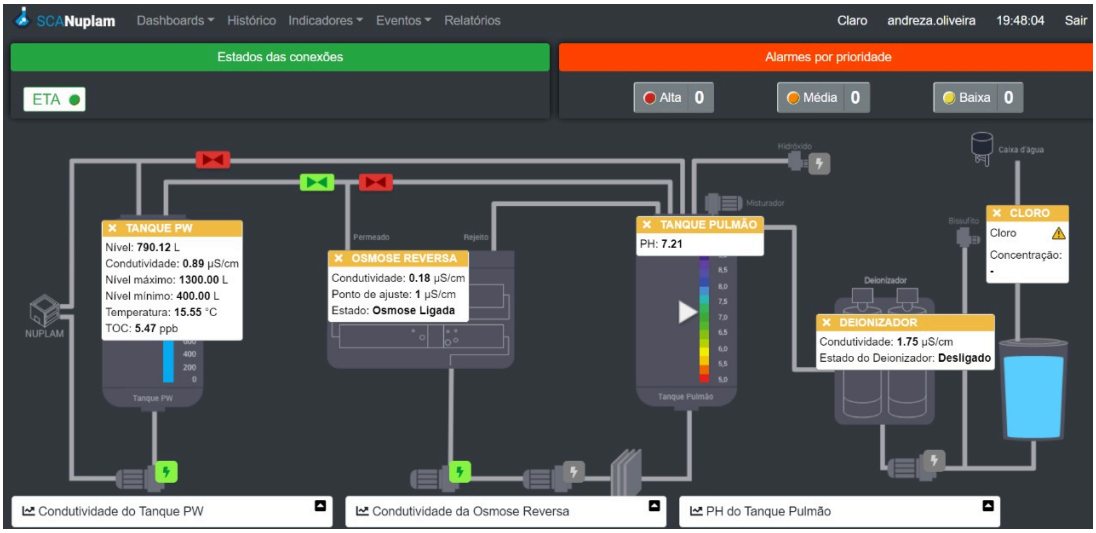


Figura 4 – Dashboard da ETA no SCANUPLAM.

Fonte: Captura de tela do software SCANUPLAM.

Na Figura 5 é possível compreender a arquitetura de automação da qual o software faz parte. Os processos físicos do chão de fábrica, onde ficam os sensores, podem ser conectados a dispositivos de controle e coleta de dados industriais, como controladores lógicos programáveis (CLP) e analisadores, ou a dispositivos IoT, denominados pela equipe do NUPLAM de *Nodes IoT*. Os equipamentos são conectados então a um *Node IoT* central através de protocolos industriais cabeados, enquanto os outros se conectam com o mesmo dispositivo central através da rede utilizando o protocolo MQTT (Message Queuing Telemetry Transport).

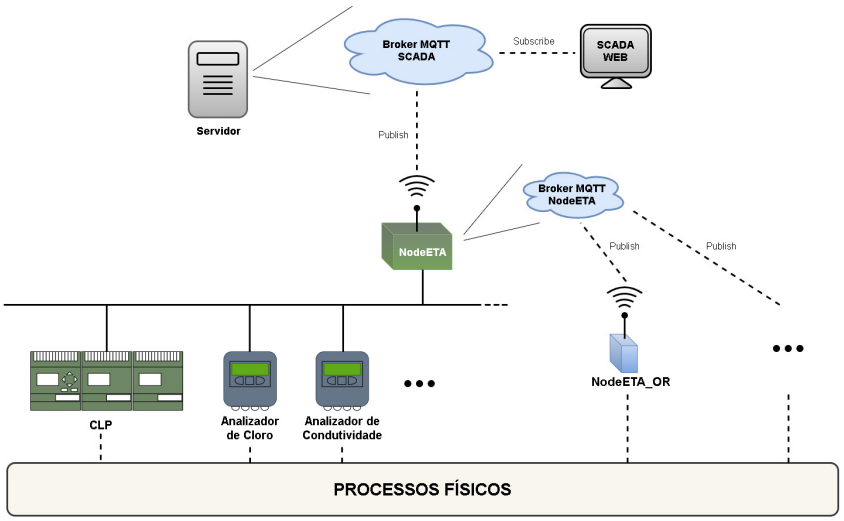


Figura 5 – Arquitetura de automação da ETA.

Fonte: (CORTEZ, 2021)

O NodeIoT central, chamado de NodeETA, recebe os dados de todos os dispositivos e então os trata e envia para o servidor que hospeda o sistema SCADA: o SCANUPLAM. O sistema então armazena os dados coletados em um banco de dados e exibe-os em interfaces intuitivas.

Em geral, as principais funcionalidades do sistema são: coletar dados em tempo real dos controladores, armazenar os dados, exibir *dashboard* dos processos em tempo real, exibir gráficos históricos de variáveis, gerar alarmes e notificar os usuários, e exibir indicadores operacionais.

3.1.2 Processos da ETA

Não existe uma única forma de produção de água purificada. Alguns sistemas empregados para essa atividade geralmente são os deionizadores, osmose-reversa, ultrafiltração ou eletro-deionização, sendo mais comum a união dessas tecnologias, de forma que as configurações do sistema evitem a proliferação microbiológica, conforme explica a Anvisa (2013). Para este trabalho, o foco será na ETA do NUPLAM e em sua configuração.

A produção de água purificada na ETA envolve quatro etapas principais:

- Coleta e pré-tratamento
- Tratamento
- Armazenamento
- Distribuição

Na etapa de coleta e pré-tratamento, a estação recebe a água impura de um poço de água externo ao laboratório e o primeiro método de pré-tratamento é a cloração da água com o objetivo de eliminar e impedir a multiplicação de matérias orgânicas. Ao final da cloração, a água é armazenada temporariamente em um tanque de entrada.

A partir desse tanque de entrada, a água é direcionada para o deionizador e no caminho uma bomba de bissulfito de sódio é utilizada para o remoção de oxigênio, a fim de controlar o processo corrosivo na ETA. Em seguida, a água entra no sistema de deionização tem como objetivo remover os sais inorgânicos dissolvidos na água e faz isso a partir da liberação de resinas catiônicas e aniônicas, que capturam os íons existentes no líquido e produzem íons de H^+ e OH^- formadores de moléculas de água.

Após o processo de deionização, a água é armazenada no tanque pulmão, que é um elemento intermediário entre o pré-tratamento e o tratamento da água. O tanque pulmão é responsável por armazenar a água deionizada além de parte da água rejeitada no processo de tratamento (osmose reversa) para manter a vazão de entrada necessária para o processo de tratamento: sistema de osmose reversa. O único controle realizado no tanque pulmão é o controle do pH da água pré-tratada.

Após a água sair do tanque pulmão, é iniciado o processo de tratamento, que ocorre no sistema osmose reversa de duplo passo. Esse sistema contém uma bomba responsável por criar pressão para empurrar a água de um meio mais concentrado para um menos concentrado através de membranas semipermeáveis que retém as impurezas da água, eliminando componentes iônicos, compostos orgânicos e micro-organismos como vírus e bactérias. Esse processo utiliza duas membranas para que a água de saída da primeira membrana seja utilizada como entrada da segunda e assim o nível de eficiência da filtragem seja ainda maior.

Essas membranas utilizadas na osmose reversa possuem poros muito pequenos, por isso as etapas anteriores ao tratamento são essenciais para o funcionamento do equipamento, evitando a obstrução desses poros. No pré-tratamento a maioria das impurezas de maior dimensão são eliminadas e o controle de pH no tanque pulmão evita a precipitação dos sais inorgânicos rejeitados no tratamento.

Ao final do tratamento na osmose reversa, existem duas válvulas de controle que determinam o destino no líquido e que são controladas de acordo o valor de condutividade da água de saída. Se esse valor for menor que $1\mu S/cm$ o produto de saída é chamado de água permeada e a válvula de permeado se abre, enquanto a válvula de rejeito fica fechada, para que ele passe à etapa de armazenamento no tanque PW (purified water). Caso contrário, a água é considerada imprópria para o armazenamento e é chamada de água de rejeito, então a válvula de rejeito é aberta e a de permeado fechada, fazendo com que a água volte ao processo de pré-tratamento no tanque pulmão.

A água armazenada no tanque PW é o produto final da ETA e deve estar pronta para ser distribuída para uso farmacêutico. Nessa etapa é imprescindível monitorar os níveis de carga microbiológica, condutividade e teor de carbono orgânico total (TOC), de forma a prevenir que a água sofra recontaminação. É importante também que o tanque mantenha sempre um volume mínimo para garantir as necessidades de consumo de água a curto prazo mesmo durante períodos de manutenção ou de sanitização.

Para evitar que esse produto seja recontaminado e que os equipamentos de tratamento sejam danificados ou fiquem suscetíveis ao acúmulo de materiais orgânicos, é obrigatório que a produção de água se mantenha constante por longos períodos de tempos. A partir dessa necessidade, é utilizado um sistema de distribuição formado por um anel de circulação ligado a duas válvulas, uma voltada para a distribuição na indústria e outra que permite a recirculação para o tanque pulmão. Esse anel de circulação percorre todos os setores da fábrica que precisam de água purificada.

Esse processo de recirculação deve acontecer constantemente, para que os equipamentos continuem a operar tanto em períodos de baixa disponibilidade de fornecimento de água quanto em períodos de pouca ou nenhuma demanda de água purificada. No segundo caso, a válvula de recirculação é aberta sempre que o nível de água no tanque atinge um valor limite máximo de armazenamento e fecha quando o nível atinge o limite mínimo.

O sistema de supervisão coleta dados a partir da etapa de pré-tratamento até os dados da água purificada no tanque PW. Todas as variáveis essenciais para monitoramento e operação da sistema são armazenadas no banco do sistema supervisório.

Com base nesses dados e no entendimento do sistema, foi desenvolvido um pipeline de análise de dados para avaliar o desempenho da produção de água na ETA.

3.2 Pipeline - Modelo de análise de dados da ETA

A partir do conhecimento teórico discutido, é proposto um pipeline com a definição de etapas para criação de um modelo de análise de dados descritiva para a ETA. O diagrama da Figura 6 representa esse modelo e, em seguida, cada passo será descrito. No capítulo seguinte, serão apresentados os resultados para cada etapa da aplicação desse pipeline.

3.2.1 Coleta de dados

Entender a fundo o contexto que será analisado é essencial para o analista de dados realizar seu trabalho de forma mais assertiva. Por isso, na primeira etapa o objetivo é compreender como ocorre o tratamento de água no NUPLAM, quais as principais particu-

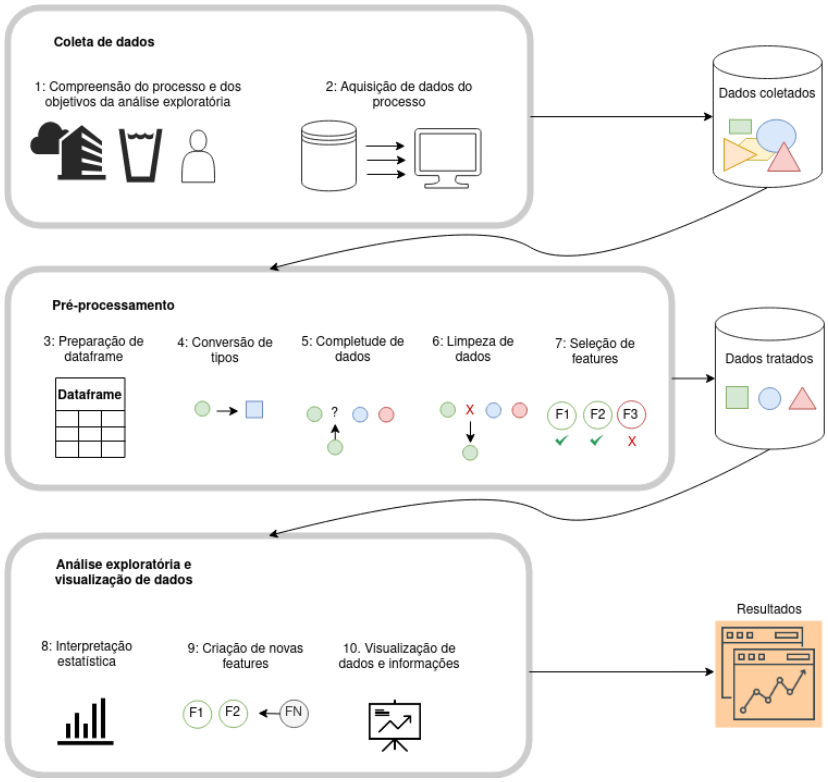


Figura 6 – Pipeline para modelagem de análise de dados da ETA.

Fonte: Próprio autor

laridades do processo e dos equipamentos utilizados pelo setor e definir quais os objetivos da análise de dados.

O modelo proposto inicia com a compreensão do processo através da revisão da literatura sobre os processos de tratamento de água purificada utilizados na ETA do NUPLAM e a execução de uma etapa essencial: a realização de entrevistas com os especialistas da área para obter detalhes dos processos e, consequentemente, dos dados a serem coletados.

Em seguida, o segundo passo consiste em selecionar um conjunto de dados do processo da ETA que será utilizado para o validação do modelo. Nessa fase, a extração dos dados se dá a partir da base armazenada pelo SCANUPLAM. Além disso, é preciso conhecer e compreender o significado de cada variável salva na base de dados a partir das informações dos especialistas.

3.2.2 Pré-processamento

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

A fase de Pré-processamento inicia com a preparação do *DataFrame* a partir dos dados coletados do banco. Esses dados são carregados em objetos *DataFrames*, de modo que seja possível utilizar as funções da biblioteca pandas para manipulá-los com mais facilidade nas etapas seguintes.

O quarto passo executa as conversões dos tipos de algumas variáveis, que podem estar inconsistentes ou não serem significativas para os valores representados. O objetivo dessa operação é garantir que seja possível manipular os dados de forma adequada nas etapas seguintes do pré-processamento e da análise exploratória.

Em seguida, o *DataFrame* será analisado para verificar a completude dos dados, ou seja, verificar se existem colunas com registros ausentes. A parir dessa verificação, as colunas que possuírem mais de 20% de dados faltantes serão consideradas impróprias para utilização e excluídas do *DataFrame*. E quanto aos outros casos, será analisado se os dados faltantes serão preenchidos com o valor médio da *feature* ou com o último registro não nulo.

O sexto passo consiste na limpeza dos dados, onde será feita a remoção de possíveis dados duplicados, que por algum erro podem possuir o mesmo registro de tempo, e o tratamento de *outliers* que estejam fora da faixa de operação dos dispositivos, de acordo com as informações dos especialistas. Por exemplo, não é possível receber dados de condutividade negativos.

Por fim, a sétima etapa consistirá em selecionar as *features* que permanecerão no *DataFrame*, de acordo com sua relevância para o processo. Essa escolha é feita a partir das informações coletadas com os especialistas da ETA e do estudo das variáveis de processo.

3.2.3 Análise exploratória e visualização dos 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.

No passo oito, as variáveis serão divididas de acordo com sua classificação em quantitativas e qualitativas. Para as variáveis 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 variáveis.

Com a interpretação estatística dos dados e conhecendo o processo de tratamento de água, a etapa nove consiste em utilizar essas informações como base para criar novas *features* a partir dos dados existentes e adicioná-las ao *DataFrame*. Com essas novas variáveis, espera-se extrair informações que estavam implícitas nos dados existentes e

utilizá-las de forma explícita na análise exploratória.

Os indicadores gerados ao final deste trabalho são baseados nessa etapa de criação de novas *features* e aprimoram a representação do desempenho de produção de água purificada no NUPLAM.

No entanto, o resultado final do trabalho só é factível e acessível aos usuários da ETA se forem utilizadas as diferentes técnicas de visualização de dados presentes nas bibliotecas. O último passo do modelo é a geração de gráficos e representações visuais dos dados gerados como resultados.

4 COLETA DE DADOS E PRÉ-PROCESSAMENTO

4.1 Informações do processo

A primeira etapa do modelo de análise de dados é compreender os detalhes do processo realizado na ETA do NUPLAM e quais as principais informações que precisam ser extraídas. Para isso, foram realizadas reuniões com os especialistas responsáveis por acompanhar a produção de água purificada a fim de compreender as regras de negócio e definir os objetivos da análise exploratória.

Para compreender alguns detalhes técnicos do processo, a Tabela 4 resume as informações referentes aos parâmetros críticos de algumas variáveis e limites de operação dos sensores utilizados na ETA, que representam respectivamente, as faixas de valores aceitáveis para garantia da qualidade do processo e as faixas aceitáveis para considerar que um sensor pode ter realizado uma medição correta. A lista completa de variáveis disponíveis será apresentada nas próximas seções.

Variável	**Parâmetros críticos**	**Limites de operação do sensor**
TOC (ppb)	$0 < \text{TOC} < 500$	$\text{TOC} > 0$
Condutividade no tanque PW ($\mu S/cm$)	$0,1 < \text{condutividade} < 1,3$	$0 < \text{condutividade} < 2$
Temperatura no tanque PW ($°C$)	temperatura <= 25	–
Condutividade na osmose reversa ($\mu S/cm$)	$\text{condutividade} < 1$	$0 < \text{condutividade} < 2$
PH no tanque pulmão	–	$1 < \text{pH} < 14$

Tabela 4 – Resumo sobre algumas variáveis da ETA.

Fonte: Autoria própria.

Ainda sobre a Tabela 4 é importante notar que na Figura 3 o TOC é medido em mg/L e no NUPLAM essa medida é feita em partes por bilhão (ppb), porém $0,5mg/L$ são equivalentes a $500ppb$, então as regras também são equivalentes.

Quanto à eficiência do processo, existem dois fatores que se destacam para avaliar a produção de água: o volume produzido e o atendimento às exigências de qualidade da Anvisa, que determinam que a água pura é definida através de seus valores de condutividade a uma dada temperatura, TOC e teor microbiológico.

Outro ponto importante é que a ETA no NUPLAM passa por um regime sazonal semanal, no qual o processo de sanitização do deionizador é realizado em todas as segundas-feiras e o de sanitização da osmose-reversa é feito em todas as terças-feiras. A mudança no estado dessas máquinas afeta o estado da estação como um todo e por isso é possível que exista influência desses processos nos valores coletados ao longo da semana.

Quanto aos dados disponíveis, existem problemas conhecidos nas leituras dos sensores de pH e cloro, portanto é preciso avaliar com cuidado a qualidade dos dados para essas duas variáveis. Além disso, não há informações disponíveis sobre o estado de operação do deionizador, o estado de operação da osmose reversa, o teor microbiológico e sobre a vazão de entrada e saída dos tanques.

Considerando essas informações dos especialistas, foi definido que as principais variáveis de estudo são a condutividade, o TOC, a temperatura e o nivel no tanque PW e que o objetivo da etapa de análise exploratória e visualização de dados é identificar possíveis relações entre as variáveis e como a sazonalidade nos processos da ETA interferem nas medições ao longo das semanas. A partir o cumprimento desses objetivos, espera-se que seja possível atingir também os objetivos gerais e específicos desse trabalho.

4.2 Armazenamento e coleta de dados

O SCANUPLAM é responsável pela comunicação com os controladores e, consequentemente, sensores utilizados na estação de tratamento e também pelo armazenamento de dados em tempo real em um banco de dados MongoDB, com periodicidade de 10 segundos entre cada coleta. Com isso, os dados necessários para esse trabalho foram coletados a partir de buscas na base de dados do sistema.

Deve-se notar, por motivos de transparência, que no período de realização da pesquisa o SCANUPLAM ainda estava em fase de desenvolvimento, o que significa que ainda não estava em uma versão de produção completamente estável e testada. No entanto, as funcionalidades de coleta e armazenamento já estavam desenvolvidas e em funcionamento. Dado esse cenário, considerou-se que os dados são válidos para a pesquisa, mas devem ser tratados com atenção na etapa de pré-processamento para mitigar possíveis erros na análise.

Com isso, o intervalo de tempo escolhido para obtenção de dados abrange o período de 10 semanas entre os dias 28/06/2021, uma segunda-feira, e 05/09/2021, um domingo. Um dos motivos dessa escolha foi o fato de que entre os dias 28/06/2021 e 25/07/2021 a ETA passou por uma vistoria, portanto não foi realizada nenhuma manutenção nesse intervalo. Com isso, é possível que a amostra consiga descrever bem a produção de água purificada, em diferentes condições, por um período significativo de tempo.

4.2.1 Obtenção de dados do SCANUPLAM

Na obtenção dos dados, foi utilizada a ferramenta *MongoDB Compass* para acessar a base de dados de desenvolvimento do SCANUPLAM para filtrar e exportar os dados em um arquivo de formato JSON.

Figura 7 – Dado armazenado na base de dados do SCANUPLAM.

Esse arquivo é formado por um vetor de objetos que possuem todas as informações de variáveis em um determinado tempo divididas a partir da máquina ou processo a que pertencem na ETA. O formato dos objetos e das variáveis podem ser vistos na Figura 7, enquanto mais informações e detalhes sobre cada variável são encontradas na tabela do apêndice A.

4.3 Pré-processamento

A etapa de pré-processamento tem o objetivo de preparar os dados para garantir que eles estejam tratados e adequados para as etapas seguintes.

4.3.1 Preparação do *DataFrame*

Para lidar com o tratamento e manipulação de dados, foi utilizado o pacote *pandas* do Python, que possui objetos do tipo *DataFrame* com diversas funcionalidades desenvolvidas para lidar com essas atividades.

O primeiro passo é converter os dados que se encontram originalmente em formato JSON para o formato de *DataFrame*. Entretanto, o Pandas não consegue lidar diretamente com objetos aninhados, que é o caso do modelo de dados utilizado, com *features* separadas

de acordo com a respectiva máquina a qual estão associadas. A solução encontrada foi criar quatro *DataFrames* para os diferentes processos da ETA (tanque_pw, osmose_reversa, deionizador e tanque_pulmao). A coluna *createdAt* foi replicada para cada DataFrame e as colunas com mesmo nome entre as estruturas foram renomeadas param serem únicas.

```
Int64Index: 753542 entries, 0 to 753541
Data columns (total 16 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   nivel                   753542 non-null  float64
 1   condutividade_pw        753542 non-null  float64
 2   estado_valvula          753542 non-null  int64
 3   estado_bomba            753542 non-null  int64
 4   nivel_min               753542 non-null  int64
 5   nivel_max               753542 non-null  int64
 6   temperatura             753542 non-null  float64
 7   toc                     753542 non-null  float64
 8   createdAt               753542 non-null  int64
 9   pH                      655387 non-null  object
 10  condutividade_or        753542 non-null  float64
 11  estado_valvulaSP1_TC404 753542 non-null  float64
 12  estado_valvulaSP2_TC404 753542 non-null  float64
 13  condutividade_deio      753542 non-null  float64
 14  estado_deionizador      753542 non-null  int64
 15  cloro                   753542 non-null  float64
dtypes: float64(9), int64(6), object(1)
memory usage: 97.7+ MB
```

Figura 8 – Informações do *DataFrame* preparado.

Por fim, foi realizada uma operação de *merge* a partir da coluna *createdAt* para unir os objetos em uma única estrutura. A Figura 8 mostra as informações do *DataFrame* resultante dessa operação, que será utilizado nas próximas etapas. No total, ele possui 753.542 registros.

4.3.2 Conversão de tipos

Para garantir que todos os dados sejam consistentes, é importante que esses dados sejam representados por variáveis com tipos significativos. Com isso, foram verificados os tipos de cada variável, como representado na Figura 8.

Dentre as *features* do conjunto de dados, a coluna *createdAt* é formada por números inteiros que representam os milissegundos passados entre 1 de Janeiro de 1970 00:00:00 UTC e o horário do registro. Porém, como as séries analisadas são temporais, é mais fácil trabalhar com o tipo *datetime64* do *NumPy*, que possui funcionalidades para manipulação de tempo e um formato mais próximo do que é utilizado para representar datas no cotidiano.

```
createdAt = pd.to_datetime(dados.createdAt, unit='ms', utc=True)
createdAt = createdAt.dt.tz_convert('America/Recife')
dados["createdAt"] = createdAt

dados.set_index("createdAt", inplace=True)

dados = dados.sort_index()
```

Figura 9 – Operações na coluna *createdAt*.

Cada linha do *DataFrame* deve ser representada por um tempo único e ser ordenada de forma crescente de acordo com esse parâmetro. Por esse motivo, optou-se por transformar a *feature createdAt* no *index* do *DataFrame* e ordenar esse *index* no sentido crescente. A Figura 9 mostra os passos utilizados para essas operações.

Outra variável que também apresenta uma tipagem inconsistente é a do pH, que deveria ser numérica mas é apresentada com o tipo *object*. Uma investigação nos dados revelou que na coluna do pH existiam tanto valores numéricos, como esperado, quanto alguns objetos no formato de dicionário do Python que não possuíam informações sobre o pH. Para resolver o problema, os registros do tipo dicionário foram convertidos em valores nulos, como mostra a Figura 10, já que não representavam nenhuma informação útil . Em seguida, toda a coluna foi convertida para o tipo *float64* do *NumPy*.

```
def selectOnlyNumbers(x):
  if isinstance(x, dict):
    x = None
  else:
    x= x

  return x

dados['pH']
dados['pH'] = dados['pH'].transform(selectOnlyNumbers)
dados['pH'] = dados['pH'].astype("float64")
```

Figura 10 – Conversão de tipo na *feature pH*.

A Figura 11 apesenta o estado do *DataFrame* após essas mudanças. É possível perceber que a coluna *createdAt* não é mais exibida, pois ela passou a ser a coluna *index*, que mudou do tipo *Int64Index* para *DatetimeIndex*, e a *feature pH* se tornou do tipo *float64*.

```
DatetimeIndex: 753542 entries, 2021-06-28 00:00:01.093000-03:00 to 2021-09-05 23:59:54.795000-03:00
Data columns (total 15 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   nivel                   753542 non-null  float64
 1   condutividade_pw        753542 non-null  float64
 2   estado_valvula          753542 non-null  int64
 3   estado_bomba            753542 non-null  int64
 4   nivel_min               753542 non-null  int64
 5   nivel_max               753542 non-null  int64
 6   temperatura             753542 non-null  float64
 7   toc                     753542 non-null  float64
 8   pH                      613414 non-null  float64
 9   condutividade_or        753542 non-null  float64
 10  estado_valvulaSP1_TC404  753542 non-null  float64
 11  estado_valvulaSP2_TC404  753542 non-null  float64
 12  condutividade_deio      753542 non-null  float64
 13  estado_deionizador      753542 non-null  int64
 14  cloro                   753542 non-null  float64
dtypes: float64(10), int64(5)
memory usage: 92.0 MB
```

Figura 11 – Resultado das conversões de tipos.

4.3.3 Completude dos dados

A etapa de completude dos dados iniciou com uma observação da porcentagem de registros faltantes em cada *feature* para avaliar a completude dos dados e então decidir como lidar com essa possíveis colunas incompletas. Na Figura 12 essa operação é apresentada no espaço de fundo cinza e logo abaixo estão os resultados.

```
totalDadosNulos = dados.isnull().sum()
numeroLinhas = dados.shape[0]

totalDadosNulos/numeroLinhas * 100
```

```
nivel                     0.000000
condutividade_pw          0.000000
estado_valvula            0.000000
estado_bomba              0.000000
nivel_min                 0.000000
nivel_max                 0.000000
temperatura               0.000000
toc                       0.000000
pH                       18.595911
condutividade_or          0.000000
estado_valvulaSP1_TC404   0.000000
estado_valvulaSP2_TC404   0.000000
condutividade_deio        0.000000
estado_deionizador        0.000000
cloro                     0.000000
dtype: float64
```

Figura 12 – Análise de completude de dados.

É possível observar que a *feature pH* apresentou aproximadamente 18,60% de dados faltantes. No entanto, a porcentagem 18% não ultrapassa o limiar de 20%, estabelecido neste trabalho como limiar para remoção da *feature*. Mesmo considerando a informação do especialista sobre falhas na leitura do valor de pH, foi escolhido manter a *feature* pH para a etapa de análise de outliers.

4.3.4 Limpeza de dados

A primeira técnica utilizada no processo de limpeza dos dados foi a remoção de registros duplicados. Para isso, foi utilizada a função *DataFrame.drop_duplicates* e o resultado foi a eliminação de 157.635 registros, o que fez com que o novo total de dados fosse igual a 595.907. Além disso, a porcentagem de dados ausentes na coluna pH subiu para 21,78%.

Em seguida, foram feitas análises de *outliers* de erro a partir das informações obtidas com os especialistas da ETA para as variáveis cujos sensores apresentavam limites de operação bem estabelecidos, conforme a Tabela 4.

Dentre as variáveis em que foram observados valores fora da faixa limite, a primeira *feature* que apresentou problemas foi o *pH*, com 64,74% dos dados fora dos limites aceitáveis. Com isso, somado aos outros problemas já identificados, ficou entendido que essa variável não apresentava qualidade suficiente para ser utilizada nesse trabalho e por isso deveria ser removida.

Já a feature *condutividade_pw* apresentou apenas 27 registros de erros, o que representa aproximadamente 0,005% da amostra. Além disso, foram identificados 7.249 erros na coluna *condutividade_deio*, o que equivale a cerca de 1,22% do total de linhas do *DataFrame*. Para as duas últimas variáveis, a escolha de tratamento foi a remoção das linhas com registros de erros, dado que a quantidade de problemas identificados foi pequena quando comparada ao tamanho total da amostra.

Como resultado dos procedimentos executados nessa seção, a quantidade de dados da tabela diminuiu para 588.277 amostras.

4.3.5 Seleção de *features*

A seleção de *features* levou em consideração as informações obtidas com os especialistas da ETA para entender quais variáveis 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 variáveis quantitativas com valores normalizados entre 0 e 1 foi utilizado para avaliar os dados com base no critério 2. A Figura 13 apresenta um gráfico de correlação das seis variáveis com base no coeficiente de Pearson.

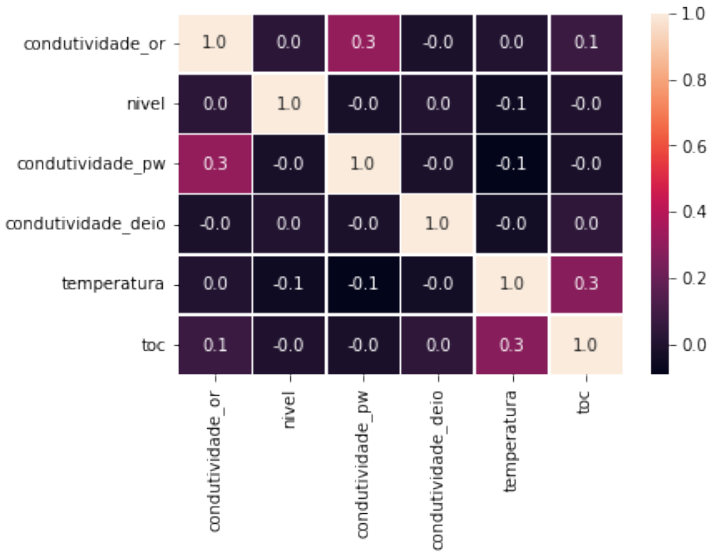


Figura 13 – Correlação entre as variáveis quantitativas.

Fonte: Próprio autor

Observa-se que a variável *condutividade_deio* (condutividade da água após o processo de deionização) é a única variável que apresenta baixo grau de correlação com todas as outras variáveis quantitativas. Com isso, essa variável será removida na seleção de *features* seguindo o critério 2.

A partir da lista de critérios e do resultado da análise de correlação, a lista a seguir apresenta as colunas que não foram selecionados para continuar no *DataFrame* após o pré-processamento.

- **condutividade_deionizador**: remoção pelo critério 2;
- **estado_bomba**: remoção pelo critério 3;
- **nivel_min**: remoção pelo critério 3;
- **nivel_max**: remoção pelo critério 3;
- **estado_valvulaSP1_TC404**: remoção pelo critério 3;
- **estado_valvulaSP2_TC404**: remoção pelo critério 3;
- **pH**: remoção pelo critério 4;
- **cloro**: remoção pelo critério 4;

- **estado_deionizador**: remoção pelo critério 4.

Com base nessas decisões a Figura 14 mostra o script que foi utilizado para remover as colunas não selecionadas.

```
# Seleção de features
dados = dados.drop(columns=['estado_bomba', 'nivel_min', 'nivel_max', 'pH',
                            'cloro', 'estado_valvulaSP1_TC404', 'estado_valvulaSP2_TC404',
                            'estado_deionizador', 'condutividade_deio'])
```

Figura 14 – Script para remoção de *features* não selecionadas.

Após a execução desses passos, finaliza a etapa de pré-tratamento dos dados do modelo e as primeiras linhas do *DataFrame* resultante podem ser visualizadas na Figura 15, gerada a partir do comando *DataFrame.head()*.

createdAt	nivel	condutividade_pw	estado_valvula	temperatura	toc	condutividade_or
2021-06-28 00:00:01.093000-03:00	872.236572	0.902416	0	17.444300	6.323409	0.456795
2021-06-28 00:00:11.092000-03:00	875.810364	0.906937	0	17.404514	6.211285	0.453157
2021-06-28 00:00:21.095000-03:00	878.684875	0.909650	0	17.440683	6.175116	0.459602
2021-06-28 00:00:31.105000-03:00	880.627197	0.896991	0	17.408131	6.811690	0.454609
2021-06-28 00:00:41.101000-03:00	884.900146	0.919596	0	17.426214	6.334259	0.454609

Figura 15 – Visualização do *DataFrame* após o pré-processamento.

Os dados estão prontos para serem utilizados na etapas seguintes de análise exploratória, geração dos indicadores e visualização dos dados.

5 ANÁLISE EXPLORATÓRIA E VISUALIZAÇÃO DE DADOS

5.1 Interpretações estatísticas

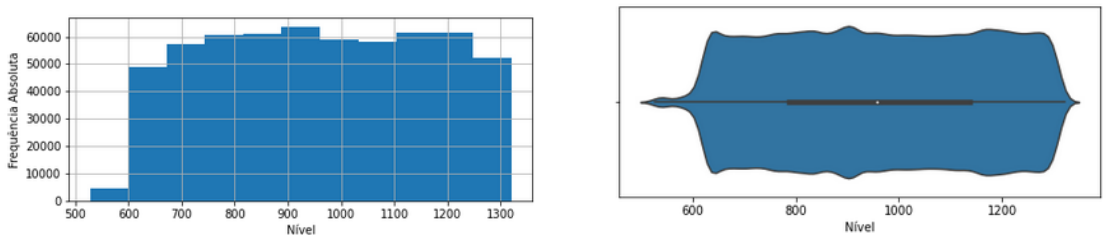
Na primeira etapa da análise exploratória foram utilizadas técnicas para análise e interpretação dos dados através de estatística descritiva. Para isso, as variáveis foram divididas em dois grupos de acordo com a classificação delas em quantitativa contínuas ou qualitativas nominais.

	nivel	condutividade_pw	temperatura	toc	condutividade_or
count	588631.000000	588631.000000	588631.000000	588631.000000	588631.000000
mean	959.982256	0.863865	18.761691	10.428016	0.393022
std	203.078674	0.094748	1.100999	9.075006	0.147446
min	527.754761	0.000904	17.277922	1.603356	0.103627
25%	787.475708	0.804760	18.102575	4.225607	0.293270
50%	957.618958	0.868960	18.402777	6.855093	0.399662
75%	1138.172729	0.925022	19.024885	13.553588	0.480470
max	1320.513428	1.250543	25.654659	112.794044	1.999991

Figura 16 – Descrição estatística das variáveis contínuas.

Fonte: Próprio autor

Para as variáveis quantitativas, foram obtidas as principais medidas de tendência central e dispersão para variáveis quantitativas contínuas através da função DataFrame.describe(). A Figura 16 mostra os resultados dessa operação.

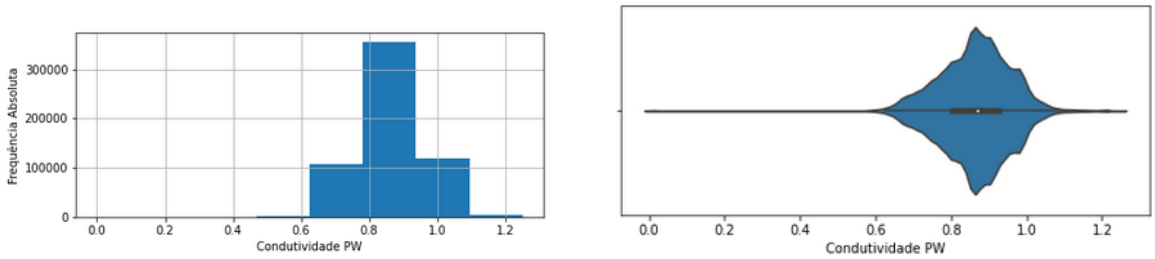


(a) Histograma (b) Gráfico violino

Figura 17 – Histograma e gráfico de violino do nível no tanque PW.

Fonte: Próprio autor

A partir desses resultados, é possível notar que a variável *nivel* possui uma média próxima do valor da mediana, representado pelo percentil 50%, o que indica uma distribuição de probabilidade simétrica. Entretanto, o valor do desvio padrão indica um alto nível de dispersão dos dados. O histograma e gráfico de violino na Figura 17 mostram que essa variável possui uma distribuição aproximadamente uniforme. É provável que esse comportamento ocorra pelo fato de a industria não estar consumindo a água produzida, então ela permanece recirculando continuamente pela ETA.

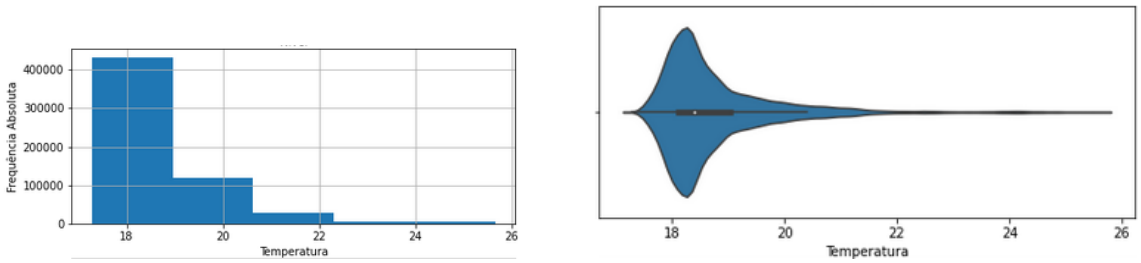


(a) Histograma (b) Gráfico violino

Figura 18 – Histograma e gráfico de violino da variável da condutividade no tanque PW.

Fonte: Próprio autor

Além disso, a feature *condutividade_pw* possui média e medianas próximas e um baixo desvio padrão e o valor máximo atingido na amostra foi de $1.25 \mu S/cm$, que está dentro das conformidades exigidas pela Anvisa. Nos gráfico da Figura 18 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 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 variável é baixo.



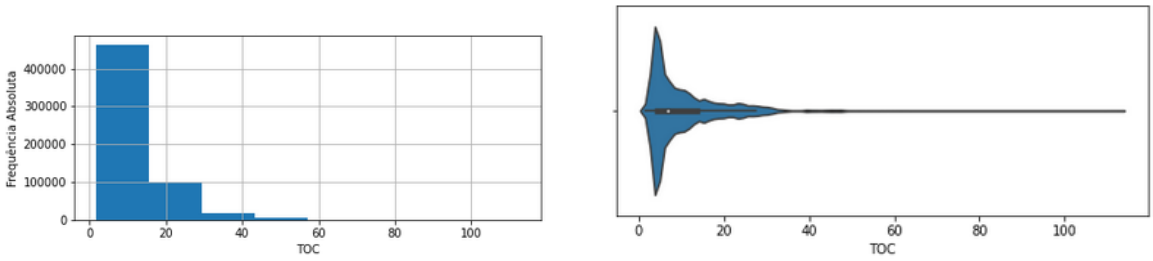
(a) Histograma (b) Gráfico violino

Figura 19 – Histograma e gráfico de violino da temperatura no tanque PW.

Fonte: Próprio autor

Para as variáveis de TOC e temperatura, as Figuras 19 e 20 mostram que ambas seguem uma distribuição assimétrica a direita. Isso significa que os dados estão mais concentrados abaixo da média, porém existe uma alta dispersão de dados acima das

medianas, principalmente para o TOC, e existem algumas medições com valores altos que deslocam os valores das médias para a direita. Novamente, é preciso entender o significado desses outliers para entender se eles se referem a falhas ou anomalias.

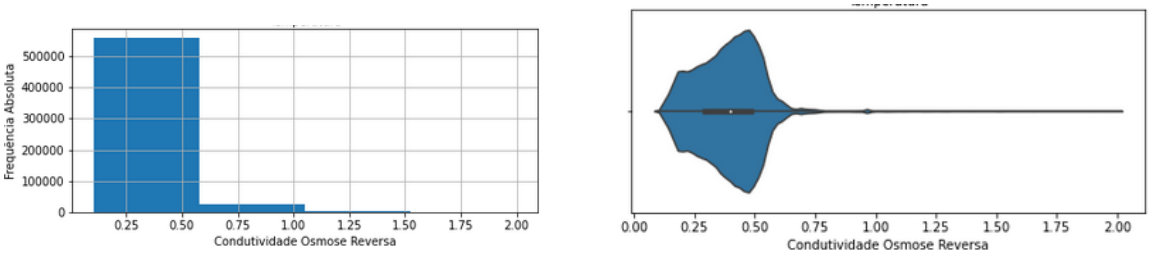


(a) Histograma (b) Gráfico violino

Figura 20 – Histograma e gráfico de violino do TOC no tanque PW.

Fonte: Próprio autor

Quanto à coluna *condutividade_osmose_reversa*, na Figura 21, ela possui um valor médio próximo à mediana, com uma distribuição assimétrica a esquerda e um baixo desvio padrão, o que se significa que existem mais valores acima da média, mas a maioria dos registros se encontra ao redor dela. Assim como nos casos anteriores, existem valores extremos a serem analisados.



(a) Histograma (b) Gráfico violino

Figura 21 – Histograma e gráfico de violino da condutividade na osmose reversa.

Fonte: Próprio autor

Em seguida, para o segundo grupo, que possui apenas a feature *estado_valvula*, foi criada a tabela de frequências da Figura 22 e o gráfico de pizza da Figura 23 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 22 – Tabela de frequências do estado da válvula de realimentação no tanque PW.

Fonte: Próprio autor

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.

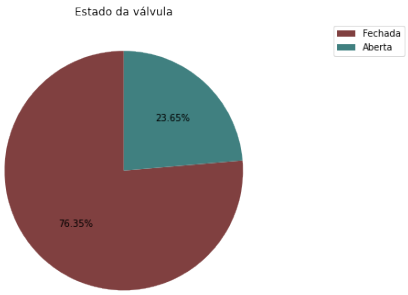


Figura 23 – Gráfico de pizza do estado da válvula de realimentação no tanque PW.

Fonte: Próprio autor

5.2 Criação de novas features

A partir dos conhecimentos sobre as variáveis 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 variáveis surgiu a partir das informações obtidas na seção 4.1 de que os equipamentos da ETA passam por procedimentos de sanitização que acontecem semanalmente em dias fixos. Portanto, essas duas variáveis podem ajudar na visualização de dados temporais agrupados por dias da semana.

Como o processo de sanitização da osmose reversa acontece nas terças-feiras, o dia escolhido para representar o valor 0 foi o que vem após dele, a quarta-feira. Dessa forma,

é possível visualizar os gráficos começando no primeiro dia após as sanitizações e como o sistema se comporta durante a semana. A Tabela 5 apresenta um mapeamento entre o dia da semana e o valor das variáveis *diaNome* e *diaNumero*.

	Quarta-feira	**Quinta-feira**	**Sexta-feira**	**Sábado**	**Domingo**	**Segunda-feira**	**Terça-feira**
diaNome	wednesday	thursday	friday	saturday	sunday	monday	tuesday
diaNumero	0	1	2	3	4	5	6

Tabela 5 – Mapeamento das variáveis *diaNome* e *diaNumero*.

Fonte: Próprio autor

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.

Também foi adicionada ao *DataFrame* a variável *tempo_valvula_aberta*, que conta a quantidade de tempo em segundos que a válvula de recirculação do tanque PW fica aberta ou fechada após mudar o estado. Valores positivos representam o tempo em que a válvula ficou aberta e valores negativos representam a quantidade de tempo que a válvula permaneceu fechada.

Com isso, ao fim dessas operações o *DataFrame* teve 3 colunas adicionadas e suas características gerais podem ser identificadas na Figura 24.

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 588631 entries, 2021-06-28 00:00:01.093000-03:00 to 2021-09-05 23:59:54.795000-03:00
Data columns (total 10 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   nivel                588631 non-null  float64
 1   condutividade_pw     588631 non-null  float64
 2   estado_valvula       588631 non-null  int64
 3   temperatura          588631 non-null  float64
 4   toc                  588631 non-null  float64
 5   condutividade_or     588631 non-null  float64
 6   tempo_valvula_aberta 588631 non-null  float64
 7   diaNome              588631 non-null  object
 8   diaNumero            588631 non-null  int64
 9   semana               588631 non-null  int64
dtypes: float64(6), int64(3), object(1)
memory usage: 49.4+ MB
```

Figura 24 – Visão do *DataFrame* após a adição das novas *features*.

5.3 Exploração e visualização de dados

O primeiro passo executado nessa etapa foi investigar quais dos outliers identificados na seção 5.1 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 espo-

rádicos, o que pode indicar que são medições falhas. A Figura 25 apresenta os limite de valores superiores e inferiores das variáveis analisadas para outliers.

condutividade_pw	0.624367
temperatura	16.719110
toc	-9.766363
condutividade_or	0.012471

(a) Limites inferiores

condutividade_pw	1.105414
temperatura	20.408350
toc	27.545559
condutividade_or	0.761269

(b) Limites superiores

Figura 25 – Valores limites para identificar potenciais outliers.

Dentre essas variáveis, apenas a condutividade no tanque PW apresentou registros de valores abaixo do limite inferior, além de apresentar *outliers* acima do limite superior. Para tentar identificar a natureza dessas medições, o gráfico da Figura 26 foi plotado para identificar o comportamento dos pontos abaixo e acima dos limites delimitados. Nele, 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.

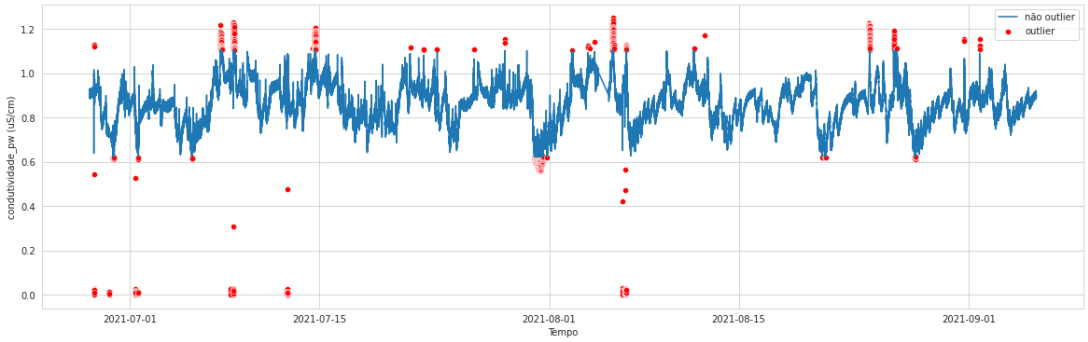


Figura 26 – Gráfico da condutividade no tanque PW com *outliers*.

Fonte: Próprio autor

Para as variáveis *temperatura* e *condutividade_or* os *outliers* se mostraram ajustados às curvas ao longo do tempo, como extensões do comportamento esperado das variáveis, o que é visível nas Figuras 27 e 28, portanto foi considerado que não havia necessidade de tratá-los.

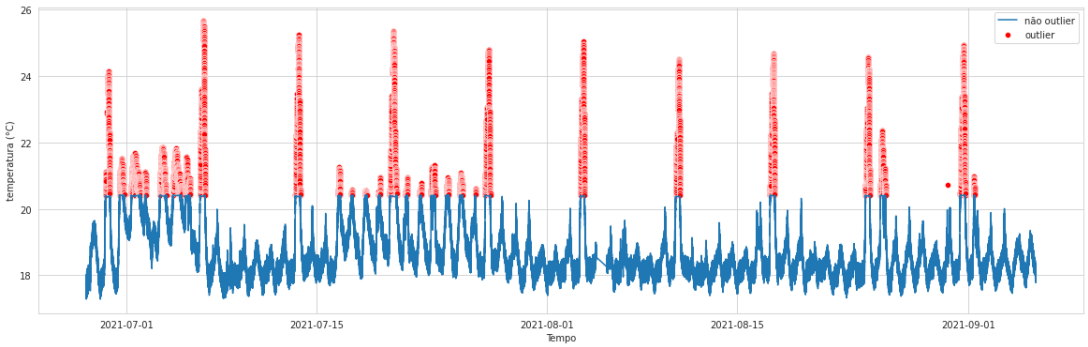


Figura 27 – Gráfico da temperatura no tanque PW com *outliers*.

Fonte: Próprio autor

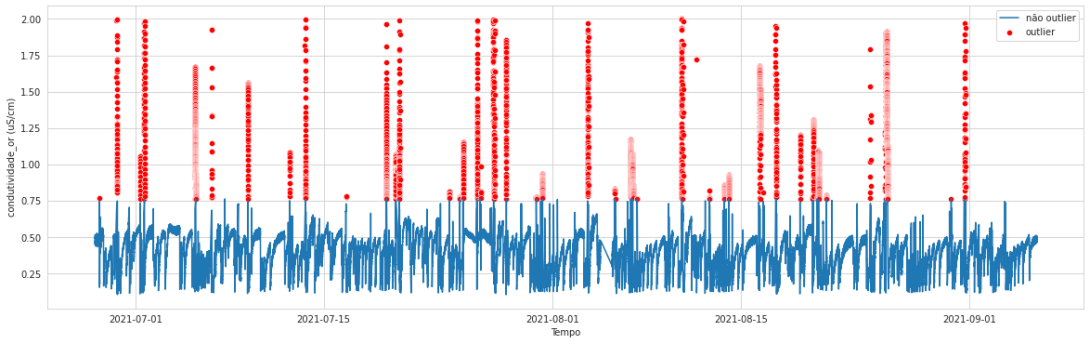


Figura 28 – Gráfico da condutividade na osmose reversa com *outliers*.

Fonte: Próprio autor

Já para o TOC, o gráfico na Figura 29 mostra que os outliers acima de 80*ppb* também se apresentaram como poucos e espalhados, além de não serem contínuos em relação à curva da variável. A partir disso, esse valor foi utilizado como limite para que os valores acima dele fossem excluídos.

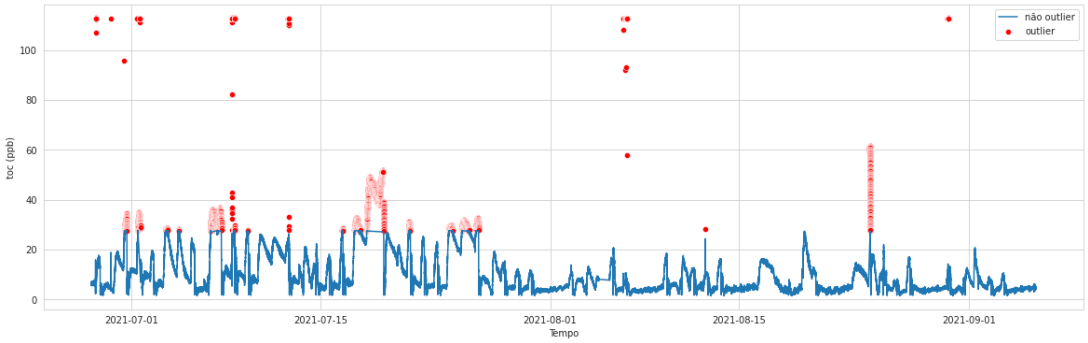


Figura 29 – Gráfico do TOC no tanque PW com *outliers*.

Fonte: Próprio autor

Com essas operações, foram removidos 450 dados, o que representava aproximadamente 0% da amostra, que passou a ter 588.181 registros no total.

Em seguida, foi realizada uma análise das séries temporais das variáveis 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 são coletados a cada 10 segundos, no entanto essa frequência era muito 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 30 está o gráfico da temperatura no tanque PW com amostragem de 1 dia e ao seu lado está a componente de tendência dessa variável, extraída com a janela de 7 dias de amostras. Abaixo, o gráfico de linhas do lado esquerdo representa a sazonalidade da série após removida a tendência da função original e ao lado esquerdo a sazonalidade é exibida novamente, mas com valores agrupados através da média por dia da semana. Os outros gráficos de decomposição das séries temporais seguem a mesma estrutura.

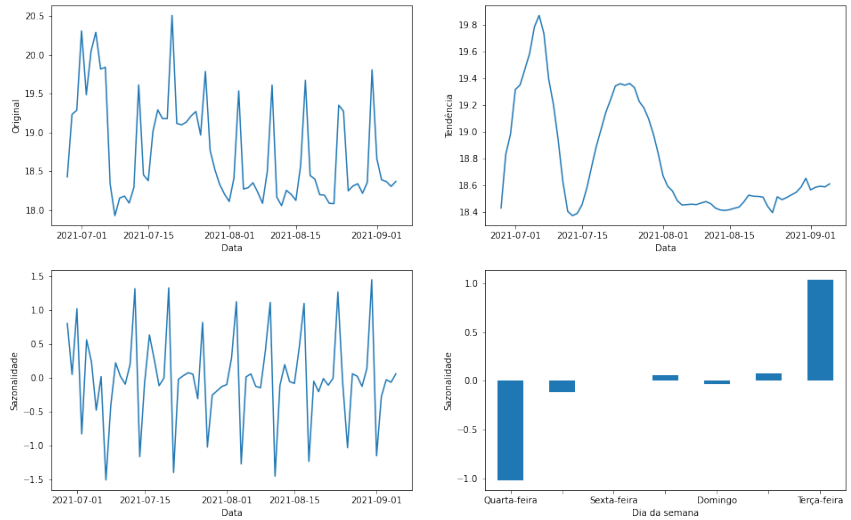


Figura 30 – Decomposição dos componentes da série temporal da temperatura no tanque PW.

Fonte: Próprio autor

A partir da análise da tendência é possível perceber que no mês de julho a temperatura passou por dois ciclos de aproximadamente 15 dias nos quais a média subiu alguns graus e depois caiu e que nos períodos seguintes ela tendeu a permanecer mais estável. Essa diferença entre os meses pode indicar que em algum dos dois meses houve alguma

anomalia no processo da ETA. Quanto à sazonalidade, a temperatura costuma aumentar consideravelmente nas terças-feiras, então na quarta-feira ela diminui e se mantém estável durante a semana. Esse comportamento pode estar relacionado ao fato de que nas terças-feiras é realizado o processo de sanitização da osmose reversa e por isso a recirculação de água para e a água fica em repouso por algum tempo.

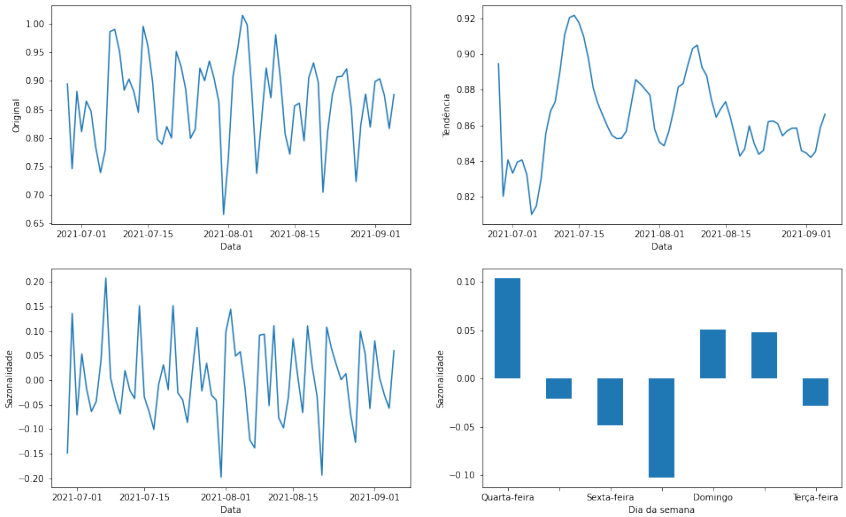


Figura 31 – Decomposição dos componentes da série temporal da condutividade no tanque PW.

Fonte: Próprio autor

Para a condutividades da água no tanque PW, nota-se a partir da Figura 31 que ela 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.

Já para a condutividade da água na saída da osmose reversa, na Figura 32, o comportamento da componente de tendência é similar ao da variável anterior, aparentando não seguir uma padrão. Além disso, pela análise da sazonalidade é possível sugerir que a influência dessa componente é mais forte nas terças-feiras e o comportamento durante a semana também segue um padrão similar ao comentado para a condutividade no tanque PW.

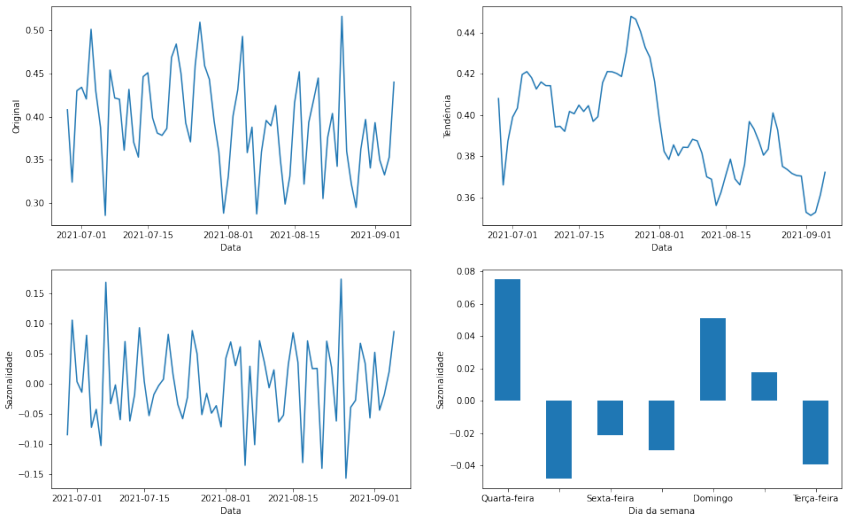


Figura 32 – Decomposição dos componentes da série temporal da condutividade na osmose reversa.

Fonte: Próprio autor

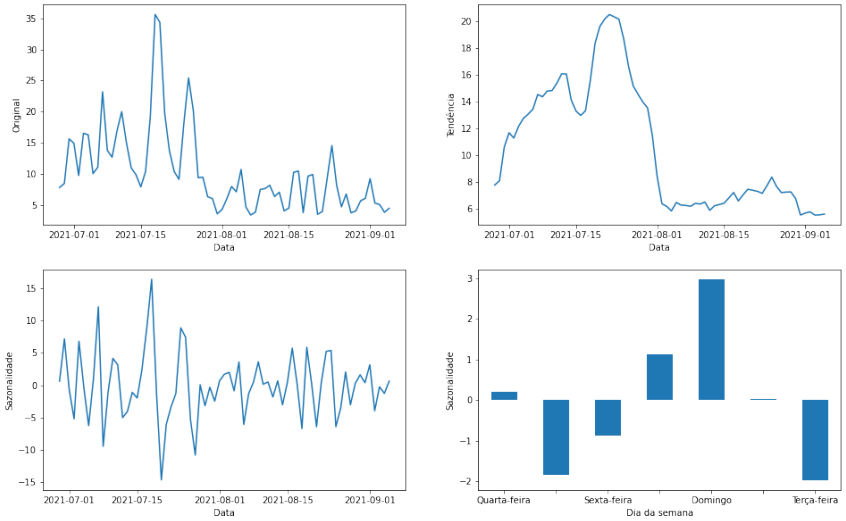


Figura 33 – Decomposição dos componentes da série temporal do TOC no tanque PW.

Fonte: Próprio autor

Quanto ao TOC, na Figura 33, a componente de tendência seguiu um comportamento parecido com o da temperatura, com dois ciclos de aproximadamente 15 dias no qual os valores subiram e desceram significativamente em julho e uma tendência mais constante a partir de agosto. Sobre a sazonalidade, os dias que aparentam ter maior influência na variação do TOC são as terças-feiras, quintas-feiras e domingos. A partir da terça-feira, as variações mais significativas são negativas e aumentam para o lado positivo gradativamente durante o decorrer da semana.

Em relação ao nível no tanque PW, na Figura 34, a análise mostra que não aparenta haver uma tendência no valor do nível do tanque, o que era esperado, dado que dentro do

período a água estava recirculando a todo momento. Se houvesse demanda de consumo de água na indústria esses valores de nível iriam variar de forma diferente e é plausível supor que nesse caso seria mais adequada uma análise de tendência. Para a componente de sazonalidade, os dias que sofrem maior influência na variação são as terças-feiras e quarta-feiras, pois na terça-feira é realizada uma pausa na produção de água que altera o nível médio do tanque e na quarta-feira a variação retorna ao normal. É importante notar que apesar de o índice de sazonalidade na terça-feira ter se mostrado negativo, esse fator pode depender do estado do nível do tanque quando a válvula de recirculação é fechada para a sanitização da osmose reversa, pois não há garantia de que ele estará sempre a baixo da média. Com isso, é possível que para amostras com um período de tempo maior, considerando o mesmo estado de recirculação contínua da água, o efeito da sazonalidade esteja menos presente.

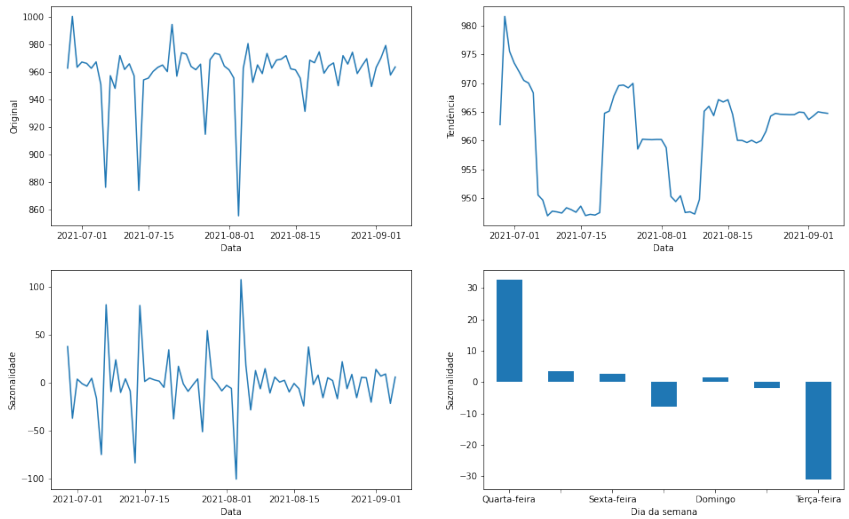


Figura 34 – Decomposição dos componentes da série temporal do nível no tanque PW.

Fonte: Próprio autor

Em seguida, foi criada uma matriz de gráficos de pontos para analisar visualmente a correlação entre as variáveis quantitativas, entretanto, essa visualização ficou prejudicada e incompreensível devido ao grande número de amostras. Como solução para esse problema, a matriz foi refeita a partir dos dados reamostrados, com periodicidade de 6 horas e valores agregados pela média. Dessa forma foi possível visualizar de forma mais clara os pontos, como exibido na Figura 35.

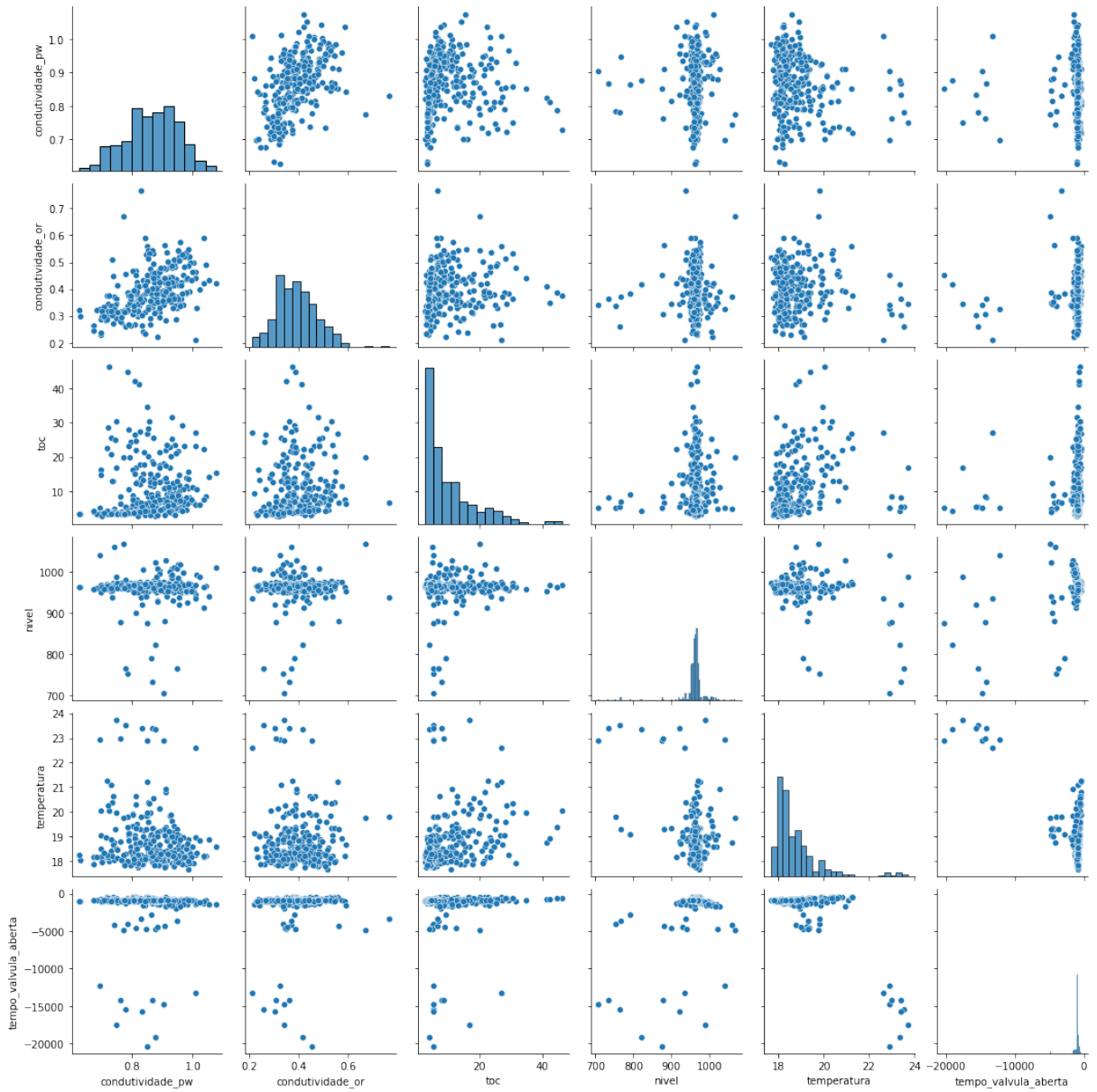


Figura 35 – Matriz de gráfico de pontos entre as variáveis quantitativas.

Fonte: Próprio autor

A partir da matriz da figura anterior, são identificados 3 pares de pontos que aparentam ter algum grau de correlação entre sim. O primeiro é o par *condutividade_pw condutividade_or*, que aparentam ter uma correlação positiva moderada, pois as duas parecem aumentar juntas em um comportamento linear. O segundo par é da *temperatura* e do *toc*, que parecem ter uma correlação positiva, porém fraca. Por fim, é possível perceber uma correlação negativa forte entre as variáveis *temperatura* e *tempo_valvula_aberta*, pois a medida que a o valor da variável *tempo_valvula_aberta* diminui, ou seja, a válvula passa mais tempo fechada, a temperatura tende a aumentar.

No geral, não é possível afirmar com certeza que existe uma relação de causa entre os pares correlacionados. Entretanto, para o terceiro par, é plausível supor que nos

períodos em que a válvula de recirculação do tanque PW passa muito tempo fechada a temperatura no tanque aumenta por não haver água recirculando.

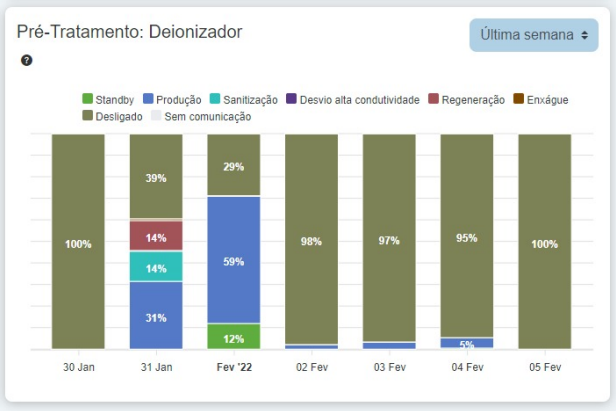
Com o trabalho apresentado nesse capítulo, compreendeu-se mais a fundo a essência e o comportamento dos dados coletados pelo SCANUPLAM através de análises estatísticas, criação de novas features e realização de análise exploratória e visualização de dados, utilizando técnicas de diferenciação entre outliers de erros ou de eventos anômalos, de decomposição e análise de séries temporais e de observações gráficas para identificar correlações entre as variáveis quantitativas. Os resultados obtidos com esse estudo serão apresentados no próximo capítulo.

6 RESULTADOS

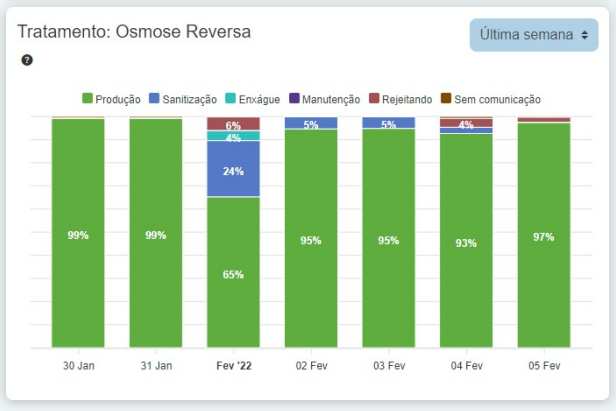
O principal resultado esperado deste trabalho é a criação de novos indicadores para permitir o monitoramento do desempenho da produção de água na ETA do NUPLAM.

Como resultado das análises, é possível afirmar que o principal fator responsável por definir o comportamento das variáveis 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 esses processos e sempre que o valor das variáveis começa a subir ao longo da semana, eles voltam a se estabilizar após as sanitizações.

Foram sugeridos dois novos indicadores: o histórico semanal do estado da osmose reversa e o histórico semanal do estado do deionizador, permitindo identificar a porcentagem de tempo em cada processo ao longo dos dias e como a duração deles afeta o sistema. Com isso, esses indicadores foram integrados ao sistema de supervisão SCANUPLAM na página de indicadores, conforme mostra a Figura 36.



(a) Deionizador



(b) Osmose reversa

Figura 36 – Indicadores de histórico semanal dos estados das máquinas.

Fonte: Captura de tela do software SCANUPLAM.

Outro resultado alcançado, apesar de não estar ligado à criação de indicadores, foi que o sistema do SCANUPLAM possuía uma funcionalidade de alarmes que apenas identificava e notificava quando uma variável ultrapassava um valor limite. Após os resultados da pesquisa, a equipe de desenvolvimento do software percebeu que alguns parâmetros iriam passar do limite durante a sanitização e isso não representava um erro, por exemplo a condutividade na osmose reversa. Com isso, a equipe alterou essa funcionalidade para evitar alarmes falso positivos durantes períodos que as máquinas não estavam em seu estado normal de produção.

Também foi sugerida a criação de um indicador de histórico semanal da produção de água purificada. Além de indicar quanto a ETA está produzindo por dia, esse indicador pode ajudar na identificação de anomalias, pois mesmo que o sistema não seja capaz de identificar isso de forma automática, operadores ou especialistas podem ver uma alteração no comportamento esperado e agirem rapidamente para identificar o motivo dessa anomalia e como tratá-la.

A partir disso, foi implementada uma função no sistema que captura diariamente a produção de água do dia anterior e o indicador foi adicionado ao SCANUPLAM, conforme sugerido. A Figura 37 mostra a tela do indicador no software.

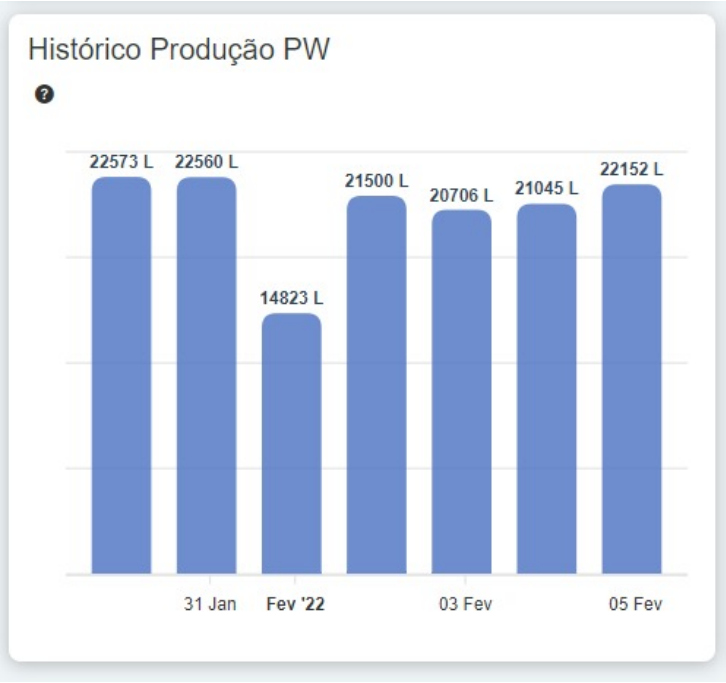


Figura 37 – Indicador do histórico semanal de produção de água no SCANUPLAM.

Fonte: Captura de tela do software SCANUPLAM.

A partir desses resultados obtidos foi possível contribuir para o desenvolvimento de ferramentas para análise de dados dentro do NUPLAM e atingir os objetivos estabelecidos para esse trabalho. Na seção seguinte, serão apresentadas as conclusões finais sobre esses resultados e sobre o trabalho de forma geral.

7 CONCLUSÃO

Neste trabalho foi proposto e aplicado um modelo de análise de dados descritivo a partir dos dados coletados de uma estação de tratamento de água purificada de um laboratório farmacêutico. Dada a importância de garantir os padrões de qualidade da água purificada, o uso de ferramentas de análise de dados é uma boa estratégia para garantir um monitoramento constante e eficaz da produção desse insumo.

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 da fábrica (SCANUPLAM) 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, extraindo informações das séries temporais e identificando padrões nas variáveis.

O trabalho atingiu os objetivos específicos estabelecidos: discutir e aplicar técnicas de um projeto de análise de dados na base de dados existente no supervisório do NUPLAM de forma a contribuir para a análise de dados históricos e obter informações do desempenho da unidade de produção de água purificada.

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. Com isso, foi possível atingir o objetivo geral e o último dos objetivos específicos.

Além de contribuir para a cultura do uso de dados para tomada de decisões dentro do NUPLAM, este trabalho acrescenta à literatura um modelo de análise de dados descritivo na área de tratamento de água, visto que outros trabalhos com aplicações em contexto semelhante aparentam focar no uso de técnicas de aprendizagem de máquina para identificação de falhas.

Enquanto impedimentos para a realização do trabalho, os principais identificados foram a falta de alguns parâmetros importantes para a avaliação de desempenho da ETA, como o nível microbiológico da água purificada e a vazão nos tanques, e a possível existência de vieses inconscientes durante a análise dos dados. Apesar disso, ainda foi possível realizar um trabalho promissor com base na base de dados disponível e evitar vieses ao consultar os especialistas para ter um aprofundamento no entendimento do processo.

Dada a amplitude de possibilidades na área da análise de dados, é possível utilizar este estudo como base para trabalhos futuros. Como sugestão, esses novos trabalhos podem utilizar novas variáveis de processo que passaram a ser coletadas após a execução desse projeto ou que ainda virão a ser, de modo a realizar uma análise ainda mais completa. Além disso, é possível aplicar novas técnicas de análise de dados, incluindo técnicas preditivas e prescritivas que auxiliem a melhorar o uso inteligente de dados para monitoramento do processo de purificação de água.

REFERÊNCIAS

AHMED, E. et al. The role of big data analytics in internet of things. *Computer Networks*, Elsevier, v. 129, p. 459–471, 2017.

ANVISA. *Farmacopeia Brasileira*. fifth. [S.l.]: Agência Nacional de Vigilância Sanitária, 2010. v. 1.

ANVISA. Guia de qualidade para sistemas de purificação de Água para uso farmacêutico. *Agência Nacional de Vigilância Sanitária*, 2013.

BOX, G. E. et al. *Time series analysis: forecasting and control*. [S.l.]: John Wiley & Sons, 2015.

BRUCE, P.; BRUCE, A.; GEDECK, P. *Practical statistics for data scientists: 50+ essential concepts using R and Python*. [S.l.]: O'Reilly Media, 2020.

BRYNJOLFSSON, E.; MCAFEE, A. *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. [S.l.]: WW Norton & Company, 2014.

CORTEZ, H. B. d. C. *Projeto de um dispositivo IoT para monitoramento do estado de produção de água para fins farmacêuticos*. 51 p. Monografia (Trabalho de Conclusão de Curso (Graduação em Engenharia Mecatrônica)) — Centro de Tecnologia, Universidade Federal do Rio Grande do Norte, Centro de Tecnologia, Universidade Federal do Rio Grande do Norte, Natal, 2021.

FRANK, A. G.; DALENOGARE, L. S.; AYALA, N. F. Industry 4.0 technologies: Implementation patterns in manufacturing companies. *International Journal of Production Economics*, Elsevier, v. 210, p. 15–26, 2019.

GARMAROODI, M. S. S. et al. Detection of anomalies in industrial iot systems by data mining: Study of christ osmotron water purification system. *IEEE Internet of Things Journal*, v. 8, n. 13, p. 10280–10287, 2021. Cited By :2. Disponível em: <www.scopus.com>.

HARRIS, C. R. et al. Array programming with NumPy. *Nature*, Springer Science and Business Media LLC, v. 585, n. 7825, p. 357–362, set. 2020. Disponível em: <https://doi.org/10.1038/s41586-020-2649-2>.

HUNTER, J.; DALE, D. The matplotlib users guide. *Matplotlib 0.90. 0 users guide*, 2007.

HUNTER, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007.

HYNDMAN, R.; ATHANASOPOULOS, G. *Forecasting: Principles and Practice*. 3rd. ed. Melburn, Australia: OTexts, 2021. Disponível em: <https://OTexts.com/fpp3>. Acesso em: 08/02/2021.

LAMB, F. *Automação industrial na prática-série Tekne*. [S.l.]: AMGH Editora, 2015. 267 p.

LEE, J.; KAO, H.-A.; YANG, S. Service innovation and smart analytics for industry 4.0 and big data environment. *Procedia Cirp*, Elsevier, v. 16, p. 3–8, 2014.

MCKINNEY, W. *Python para análise de dados: Tratamento de dados com Pandas, NumPy e IPython*. Novatec Editora, 2019. ISBN 9788575226476. Disponível em: <https://books.google.com.br/books?id=Oj5FDwAAQBAJ>. Acesso em: 03/02/2021.

MCKINNEY Wes. Data Structures for Statistical Computing in Python. In: WALT Stéfan van der; MILLMAN Jarrod (Ed.). *Proceedings of the 9th Python in Science Conference*. [S.l.: s.n.], 2010. p. 56 – 61.

MOREIRA, J. M.; CARVALHO, A. C. P. d. L. F. d.; HORVáTH, T. *A general introduction to data analytics*. [S.l.]: Wiley, 2019.

MORETTIN, P. A.; TOLOI, C. Análise de séries temporais. In: *Análise de séries temporais*. [S.l.: s.n.], 2006. p. 538–538.

NESA, N.; GHOSH, T.; BANERJEE, I. Outlier detection in sensed data using statistical learning models for iot. In: *2018 IEEE Wireless Communications and Networking Conference (WCNC)*. [S.l.: s.n.], 2018. p. 1–6.

REBACK, J. et al. *pandas-dev/pandas: Pandas*. Zenodo, 2020. Disponível em: <https://doi.org/10.5281/zenodo.3509134>.

VANDERPLAS, J. *Python data science handbook: Essential tools for working with data*. [S.l.]: "O'Reilly Media, Inc.", 2016. xi - xiii p.

WAKSOM, M. An introduction to seaborn. v. 26, 2017. Disponível em: <https://seaborn.pydata.org/introduction.html>. Acesso em: 06/02/2021.

WASKOM, M. L. seaborn: statistical data visualization. *Journal of Open Source Software*, The Open Journal, v. 6, n. 60, p. 3021, 2021. Disponível em: <https://doi.org/10.21105/joss.03021>.

WEN, X.; XIE, M. Performance evaluation of wind turbines based on scada data. *Wind Engineering*, v. 45, n. 5, p. 1243–1255, 2021. Disponível em: <www.scopus.com>.

ZHANG, J. et al. A real-time anomaly detection algorithm/or water quality data using dual time-moving windows. In: *7th International Conference on Innovative Computing Technology, INTECH 2017*. [s.n.], 2017. p. 36–41. Cited By :7. Disponível em: <www.scopus.com>.

Apêndices

APÊNDICE A – VARIÁVEIS DO SCANUPLAM

Variável	Unidade	Tipo	Descrição
setor	–	string	Setor monitorado
periodicidade	–	string	Período de tempo entre as amostras
createdAt	ms	int	Valor numérico correspondente ao horário do registro de acordo com o horário universal.
tanque_pw.nivel	L	float	Volume de água no tanque PW.
tanque_pw.condutividade	$\mu S/cm$	float	Condutividade da água no tanque PW.
tanque_pw.estado_valvula	–	int	Estado da válvula de realimentação da ETA. 0 - Válvula fechada; 1 - Válvula aberta.
tanque_pw.estado_bomba	–	int	Estado da bomba de alimentação do tanque PW. 0 - Bomba desligada; 1 - Bomba ligada.
tanque_pw.nivel_min	L	float	Valor de limite mínimo para o volume no tanque PW.
tanque_pw.nivel_max	L	float	Valor de limite máximo para o volume no tanque PW.
tanque_pw.temperatura	$°C$	float	Temperatura da água no tanque PW.
tanque_pw.toc	ppb	float	Carbono orgânico total na água do tanque PW.
deionizador.condutividade	$\mu S/cm$	float	Condutividade da água no deionizador.
deionizador. estado_deionizador	–	int	Valor que representa o estado de operação no deionizador.
deionizador.cloro	ppm	float	Contração de cloro na água de entrada da ETA.
osmose_reversa. condutividade	$\mu S/cm$	float	Condutividade da água na saída da osmose reversa.
osmose_reversa. estado_valvulaSP1_TC404	–	float	Limite para o qual a válvula de permeado fecha quando a condutividade na osmose reversa fica acima dele.
osmose_reversa. estado_valvulaSP2_TC404	–	float	Limite para o qual a válvula de rejeito abre quando a condutividade na osmose reversa fica acima dele.
tanque_pulmao.pH	–	float	PH da água no tanque pulmão.

Tabela 6 – Variáveis da ETA obtidas do SCANUPLAM e suas descrições.

Fonte: Próprio autor