

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DO RIO GRANDE
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL

USO DE SÉRIES TEMPORAIS APLICADAS A ANÁLISE DE DADOS FÍSICO-
QUÍMICOS COMO INSTRUMENTO DE APOIO A GESTÃO DA EFICIÊNCIA
DA CLARIFICAÇÃO DA ÁGUA EM UMA ETA

por

Fabiano Nobre Mendes

Dissertação para
obtenção do Título de
Mestre em Modelagem
Computacional

Rio Grande, setembro de 2023.

Fabiano Nobre Mendes

“UTILIZANDO SÉRIES TEMPORAIS APLICADAS A ANÁLISE DE DADOS FÍSICO-QUÍMICOS COMO INSTRUMENTO DE APOIO A GESTÃO DA EFICIÊNCIA DA CLARIFICAÇÃO DA ÁGUA EM UMA ESTAÇÃO DE TRATAMENTO”

Dissertação apresentada ao Programa de Pós-Graduação em Modelagem Computacional da Universidade Federal do Rio Grande - FURG, como requisito parcial para obtenção do Grau de Mestre. Área de concentração: Modelagem Computacional.

Aprovada em:

BANCA EXAMINADORA

Documento assinado digitalmente
 **DIANA FRANCISCA ADAMATTI**
Data: 25/09/2023 17:13:01-0300
Verifique em <https://validar.itl.gov.br>

Profa. Dra. Diana Francisca Adamatti (Orientadora – FURG)

Documento assinado digitalmente
 **ANDRÉ ANDRADE LONGARAY**
Data: 28/09/2023 07:55:48-0300
Verifique em <https://validar.itl.gov.br>

Prof. Dr. André Andrade Longaray (FURG)



Prof. Dr. Edécio Fernando Iepsen (FATEC SENAC Pelotas)

Rio Grande - RS
2023

Ficha Catalográfica

M538u Mendes, Fabiano Nobre.

Uso de séries temporais aplicadas a análise de dados físico-químicos como instrumento de apoio a gestão da eficiência da clarificação da água em uma ETA / Fabiano Nobre Mendes. – 2023. 52 f.

Dissertação (mestrado) – Universidade Federal do Rio Grande – FURG, Programa de Pós-Graduação em Modelagem Computacional, Rio Grande/RS, 2023.

Orientadora: Dra. Diana Francisca Adamatti.

1. Tratamento de Água 2. *Machine Learning* 3. Turbidez
4. Séries Temporais I. Adamatti, Diana Francisca II. Título.

CDU 628.16

Catálogo na Fonte: Bibliotecário José Paulo dos Santos CRB 10/2344

ATA DE DEFESA DE DISSERTAÇÃO**ATA Nº 03/2023**

Ao vigésimo sexto dia do mês de setembro de dois mil e vinte e três, por meio de videoconferência, realizou-se a 170ª sessão de defesa de dissertação do Programa de Pós-Graduação em Modelagem Computacional. O mestrando **Fabiano Nobre Mendes** apresentou e defendeu a dissertação **“UTILIZANDO SÉRIES TEMPORAIS APLICADAS A ANÁLISE DE DADOS FÍSICO-QUÍMICOS COMO INSTRUMENTO DE APOIO A GESTÃO DA EFICIÊNCIA DA CLARIFICAÇÃO DA ÁGUA EM UMA ESTAÇÃO DE TRATAMENTO”** como requisito parcial para a obtenção do grau de Mestre em Modelagem Computacional. A sessão foi aberta às treze horas e trinta minutos pela Profa. Dra. Diana Francisca Adamatti, orientadora da dissertação e presidente da Comissão de Avaliação. A referida Comissão esteve integrada, além da presidente e orientadora, pelo Prof. Dr. André Andrade Longaray da FURG e pelo Prof. Dr. Edécio Fernando Iepsen da FATEC SENAC Pelotas. Após cada integrante da banca ter arguido o mestrando, e este ter tido a oportunidade de responder a cada um, os examinadores atribuíram os seguintes conceitos ao trabalho:

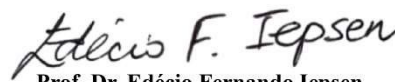
Prof. Dr. André Andrade Longaray: conceito A

Prof. Dr. Edécio Fernando Iepsen: conceito A

Tendo sido o candidato APROVADO, com conceito A, neste requisito parcial para a obtenção do grau de Mestre em Modelagem Computacional.

Documento assinado digitalmente
 **ANDRÉ ANDRADE LONGARAY**
Data: 28/09/2023 07:55:48-0300
Verifique em <https://validar.it.gov.br>

Prof. Dr. André Andrade Longaray
Membro Interno


Prof. Dr. Edécio Fernando Iepsen
Membro Externo

Documento assinado digitalmente
 **DIANA FRANCISCA ADAMATTI**
Data: 26/09/2023 15:38:19-0300
Verifique em <https://validar.it.gov.br>

Profa. Dra. Diana Francisca Adamatti
Orientadora

AGRADECIMENTOS

Obter o título de Mestre em Modelagem computacional pela Furg é um sonho realizado! Não foi uma jornada fácil, abdicar muitas vezes de estar com a família e amigos, mas o que pra muitos é apenas mais uma etapa de estudo, para mim tinha outro significado, saber que tive condições e capacidade de realizar este sonho.

Agradeço a minha orientadora Diana por acreditar na proposta e na minha capacidade em realizá-la; tu és uma pessoa especial e carregas o verdadeiro significado da palavra professor.

Agradeço as instituições que me proporcionaram o meu aprendizado, o Senac e a Furg, especialmente ao PPGMC, juntos eles mudam vidas.

Agradeço ao SANEP pela parceria, pois sem ela este trabalho não poderia ter sido realizado.

Agradeço a minha esposa Karen por todo o suporte, carinho e ajuda, sem o teu amor não conseguiria atingir os meus objetivos, aos meus pais e irmã, obrigado pelos ensinamentos da vida, vocês são exemplos de luta, alegria e amor, e aos meus filhos, Miguel e Filipe, meu singelo “Amo vocês ao infinito e além”.

Tudo posso naquele que me fortalece. (Filipenses 4:13).

RESUMO

A água é um recurso natural e essencial para a vida, devendo ser preservada. Sua potabilidade também é vital para a sobrevivência do ser humano. Neste contexto, o parâmetro turbidez é fator preponderante, pois sua presença pode “proteger” microrganismos nocivos durante a etapa de desinfecção. Considerando o extenso volume de dados coletados pelo Serviço Autônomo de Saneamento de Pelotas na Estação de Tratamento de Água Santa Bárbara ao longo dos anos, propõe-se desenvolver uma metodologia para dar suporte a tomada de decisão por parte dos gestores, através do tratamento dos dados utilizando métodos estatísticos, aprendizado de máquina e séries temporais (SARIMA), utilizando o ambiente Colab. Os estudos demonstraram que a turbidez é um parâmetro essencial no monitoramento da qualidade da água e que pode ser inferida antes da etapa de filtração a partir dos dados coletados ao longo do tempo. Além disso, utilizar dados da própria estação de tratamento oferece a vantagem de levar em conta na análise suas peculiaridades.

Palavras-chaves: Tratamento de Água, *Machine Learning*, Turbidez, Séries Temporais.

ABSTRACT

Water is a natural and essential resource for life and must be preserved. Its potability is also vital for human survival. In this context, the turbidity parameter is a preponderant factor, as its presence can “protect” harmful microorganisms during the disinfection process. Considering the extensive volume of data collected by Serviço Autônomo de Saneamento de Pelotas at the Santa Bárbara Water Treatment Station over the years, we propose to develop a methodology to support decision-making by managers, through data processing using statistical methods, machine learning and time series. Studies have shown that turbidity is an essential parameter in monitoring water quality and that it can be inferred before the filtration stage from the data collected throughout the time. In addition, using data from the treatment plant itself offers the advantage of taking into account its peculiarities in the analysis.

Keywords: Water Treatment, *Machine Learning*, Turbidity, Time Series.

ÍNDICE

1. Introdução	9
1.1 Objetivos	10
1.2 Estrutura do Texto	10
2. Embasamento Teórico	12
2.1 Tratamento convencional de água para abastecimento público	12
2.1.1 Turbidez	13
2.1.2 Cor Aparente	14
2.2 Aprendizado de máquina	14
2.3 Séries Temporais	16
2.4 Análise de componentes principais	16
2.5 Ferramentas utilizadas	17
2.5.1 Python	17
2.5.2 Google Colab	17
2.5.3 SARIMA	17
2.5.4 Função autoarima()	18
2.5.5 Análise Estacionária	18
2.5.6 Desvio Padrão	18
2.5.7 Erro Padrão	19
2.5.8 Autocorrelação e Autocorrelação Parcial (ACF e PACF)	19
2.6 Trabalhos Relacionados	19
3. Estudo de Caso	21
3.1 Metodologia	21
3.2 Etapas do Desenvolvimento	21
4. Resultados e Discussões	30
5. Conclusões	34
Referências	36
Apêndice A	38

LISTA DE FIGURAS

Figura 1. Fluxograma genérico do tratamento em ciclo completo (SANEP, 2022)	12
Figura 2. Etapas da metodologia proposta	21
Figura 3. Exemplo de planilha de dados do Sanep	22
Figura 4. Quantidades de dados faltantes por variáveis	23
Figura 5. Amostra da planilha de dados já com os dados reorganizados	23
Figura 6. <i>Heatmap</i> com todas as variáveis disponíveis na base de dados	24
Figura 7. <i>Heatmap</i> com as variáveis que apresentaram coeficientes de correlação de Pearson maiores do que 0,60.	25
Figura 8. Autovetores, autovalores e variância das 5 variáveis mais fortemente relacionadas conforme o PCA	26
Figura 9. Gráfico de % de variância explicada em função do número de variáveis	26
Figura 10. Gráfico de dispersão entre Turbidez e Cor Aparente	27
Figura 11. Número indicado de clusters conforme o método Elbow do K-means	28
Figura 12. Centróides em valores reais	28
Figura 13. Gráfico mostrando o valor escalar entre turbidez e cor aparente	29
Figura 14. Gráfico mostrando a sazonalidade dos dados ao longo dos anos	30
Figura 15. Autocorrelação entre valores vizinhos dentro do intervalo de confiança	31
Figura 16. Autocorrelação Parcial entre valores vizinhos dentro do intervalo de confiança	31
Figura 17. Valores previstos pelo modelo para os próximos 12 meses	32
Figura 18. Série temporal de turbidez para os próximos 12 meses	33

LISTA DE ABREVIATURAS

ETA	Estação de Tratamento de água
SANEP	Serviço Autônomo de Saneamento de Pelotas
PCA	<i>Principal Component Analysis</i>
ACF	Autocorrelação
PACF	Autocorrelação Parcial

LISTA DE SÍMBOLOS

Símbolos universais matemáticos

$\leq$ menor ou igual a

1. Introdução

A produção de água potável é essencial à sobrevivência de uma sociedade (WHO, 2017). Uma das principais etapas no tratamento convencional de águas superficiais visando sua potabilização é a clarificação. A clarificação consiste em uma sequência de operações cujo objetivo é a remoção de impurezas suspensas ou dissolvidas na água, tais como partículas de silte e argila, micro-organismos, plâncton, etc (Santos, 2011).

Para avaliar a eficiência da clarificação, um dos principais parâmetros considerados é a análise de turbidez, especialmente após a operação de filtração. A turbidez é um parâmetro que pode influenciar diretamente o consumidor pois, dependendo do seu teor, pode ser percebida visualmente, causando repulsa. Embora quando a turbidez é mais elevada as chances da presença de microrganismos sejam maiores, uma água com baixo teor de turbidez também pode conter contaminação, não somente microbiológica, oferecendo riscos talvez até maiores à saúde do consumidor (Libânio, 2010; Santos, 2011).

Diversos dados relativos às diferentes etapas do tratamento da água são gerados ao longo dos anos. Geralmente, sua aplicação limita-se a uma orientação momentânea de como proceder na execução do tratamento e ao preenchimento de relatórios previstos em legislação, normalmente contendo somente estatística descritiva. Nem sempre há oportunidade de desenvolver uma análise mais abrangente e metodológica dos dados para explorar seu potencial, tanto no reconhecimento quanto na modelagem de possíveis padrões de comportamento. Porém, utilizar o próprio histórico de dados de uma Estação de Tratamento de Água (ETA) para entender como suas particularidades afetam seu desempenho, pode ser uma ferramenta de gestão muito útil na determinação de suas potencialidades e vulnerabilidades.

Atualmente, com o desenvolvimento dos recursos computacionais, é possível aplicar métodos estatísticos de análise em grandes volumes de dados para diversos fins, tais como: extrair informações úteis ao entendimento de um processo ou fenômeno para subsidiar ajustes ou correções no mesmo, estimativa de resultados, determinação de padrões e/ou tendências, etc.

Nesse sentido, o aprendizado de máquina, em inglês *Machine Learning* (Tan,

2019), vem suprir uma lacuna importante na aplicação da estatística e da computação no processo de tratamento de água, pois, como visa aprender a partir dos dados, ao utilizar o histórico de dados da própria ETA, poderá indicar padrões e tendências peculiares da mesma. Essa informação extraída poderá auxiliar as decisões de gestão no momento do tratamento, nos próximos dias, meses e anos. Dentro deste cenário, tem-se a seguinte questão de pesquisa: Utilizar dados físico-químicos captados diretamente no tratamento da água ao longo dos anos na ETA, torna possível entender as causas da turbidez na qualidade da água, assim como apoiar a tomada de decisão por parte dos gestores?

1.1 Objetivos

O objetivo geral deste trabalho é desenvolver um modelo capaz de validar o parâmetro físico-químico turbidez como principal indicador de qualidade da água tratada na ETA, bem como, a partir dele, criar alternativas que auxiliem a tomada de decisão dos gestores.

Para obtenção do objetivo principal, tem-se como objetivos específicos:

- Coletar dados secundários disponibilizados pela autarquia SANEP/Pelotas e dispô-los em uma base matricial de análise;
- Avaliar e validar os dados através de métodos estatísticos e técnicas de *Machine Learning*, de forma a verificar o uso ou não da turbidez como parâmetro físico-químico preponderante para o tratamento na ETA Santa Bárbara;
- Utilizar séries temporais para prever valores da turbidez antes dos filtros, indicando se há e qual é a tendência de variação (aumento ou diminuição), de forma a dar o apoio a tomada de decisão dos gestores e técnicos.

1.2 Estrutura do Texto

O trabalho foi estruturado em 5 capítulos. No capítulo 2 encontra-se o embasamento teórico, contemplando, dentre outros, o tratamento convencional realizado na ETA Santa Bárbara e os parâmetros físico-químicos mais importantes para o tratamento da água. Ainda no capítulo 2 é possível perceber a importância do aprendizado de máquina através de seus algoritmos e métodos, assim como as

ferramentas utilizadas no desenvolvimento da metodologia e os trabalhos relacionados dando suporte ao que está sendo estudado.

No capítulo 3 temos a metodologia que foi aplicada, englobando todas suas etapas de desenvolvimento, bem como os resultados obtidos, atestados através de imagens e valores dos dados.

No capítulo 4 temos a visualização e interpretação dos resultados, através de valores e imagens das tendências, padrões e *outliers*, dando o suporte necessário à análise dos gestores, permitindo que os mesmos reflitam sobre os resultados, oferecendo embasamento teórico, prático e estatístico para a tomada de decisão.

No capítulo 5, Resultados e Discussões, é possível entender, discutir e analisar os dados gerados, sendo possível compreender a utilização do modelo SARIMA, assim como analisar os resultados dos testes aplicados na verificação de existência ou não da estacionariedade da série temporal. Ainda neste capítulo são apresentados os resultados dos testes realizados e a utilização do *autoarima()*, bem como a descrição da série temporal, mostrando os dados ao longo do tempo, nas etapas de treinamento, teste e predição, além da tendência para os próximos 12 meses. Uma avaliação técnica também é discutida a partir de um formulário preenchido pelos técnicos do SANEP, permitindo entender sua visão profissional sobre o impacto dos dados e gráficos gerados.

Finalmente, no capítulo 6, temos as conclusões, onde o texto retoma o objetivo geral idealizado para o projeto, ressaltando a utilização das técnicas, dos dados gerados e da visualização dos gráficos, de forma que possa ser discutido entre os administradores da ETA. Também é possível vislumbrar os trabalhos futuros, de forma que possam ampliar e melhorar o modelo proposto.

2. Embasamento Teórico

2.1 Tratamento convencional de água para abastecimento público

A ETA Santa Bárbara foi inaugurada em outubro de 1968, com o objetivo de reforçar o abastecimento de água da cidade de Pelotas/RS, aproveitando o barramento do Arroio Santa Bárbara. Composta por duas etapas independentes, a primeira concluída em 1968 e a segunda em 1985, sua capacidade total de produção é de 40 milhões de litros de água potável por dia (SANEP, 2022). Diversos parâmetros (físico-químicos, microbiológicos e hidrobiológicos) são analisados e registrados antes, durante e depois do tratamento.

A ETA Santa Bárbara utiliza o tratamento em ciclo completo, composto por: coagulação/floculação, decantação, filtração, correção de pH, fluoretação e desinfecção (Figura 1).



Figura 1. Fluxograma genérico do tratamento em ciclo completo (SANEP, 2022).

Cada Etapa (ETA independente) possui um sistema único para coagulação, seguido de três floco-decantadores operando em paralelo e cinco filtros que captam a saída dos três floco-decantadores juntos, também operando em paralelo.

No tratamento em ciclo completo a água bruta em geral é coagulada com sais de alumínio ou ferro no mecanismo de varredura, ocorrendo assim precipitação do metal coagulado retendo as impurezas acumuladas durante o processo. Isto irá ocorrer de forma hidráulica ou mecanizada na unidade de mistura rápida, dependendo da vazão, da variação da qualidade da água bruta e das condições disponíveis de operação e manutenção. Logo após ocorrer a coagulação da água, a mesma é submetida a uma agitação lenta até que os flocos tenham dimensão e massa suficientes para a sua

remoção por gravidade nos decantadores. Os decantadores podem ser tanto convencionais como de alta taxa; em ambos os casos, logo após a decantação, a água é submetida à filtração descendente em filtros com meios porosos granulares, optando-se sempre que possível pela filtração com taxa declinante variável.

A filtração consiste no processo cuja principal função é a remoção das partículas responsáveis pela cor e turbidez (Libânio, 2010), com a colaboração de um meio filtrante. A filtração rápida trabalha com proporções 40 vezes maiores que as da filtração lenta, abrangendo assim fenômenos físicos e químicos; já a filtração lenta envolve fenômenos físicos, químicos e biológicos.

A ETA Santa Bárbara utiliza policloreto de alumínio como coagulante, floculação hidráulica, decantadores convencionais e filtros rápidos granulares.

Após as operações de clarificação, são efetuadas a fluoretação, desinfecção e, caso necessário, correção de pH. A fluoretação das águas de abastecimento é uma exigência do Ministério da Saúde (Brasil, 2022) para assegurar o acesso ao flúor como método de prevenção de cáries dentárias às populações em vulnerabilidade social. A desinfecção, geralmente utilizando compostos clorados (cloro gás, hipoclorito de sódio, dióxido de cloro, etc), visa inativar micro- organismos potencialmente patogênicos que não foram eliminados no processo de clarificação, além de fornecer um residual do desinfetante para proteger a água de contaminação durante seu percurso para abastecimento. O pH da etapa de coagulação, que varia conforme a classe de coagulante utilizado, geralmente é ácido, mas a água tratada deve ser distribuída com pH neutro. Portanto, às vezes é necessário adicionar compostos como cal ou hidróxido de sódio para ajustar o valor do pH.

A ETA Santa Bárbara utiliza ácido fluossilícico para fluoretação, cloro gás para desinfecção e hidróxido de sódio para correção do pH.

2.1.1 Turbidez

Um dos parâmetros mais importantes de serem monitorados para acompanhar a eficiência do tratamento para potabilização da água é a turbidez, principalmente a turbidez da água filtrada. Com significado semelhante ao dos parâmetros sólidos suspensos e contagem de partículas em termos de qualidade da água, a opção por inserir a turbidez nos padrões de potabilidade nacional e internacional deve-se à

simplicidade e rapidez de sua determinação, além do menor custo global de análise quando comparado aos parâmetros citados. Fisicamente, a turbidez é uma inferência da concentração de partículas suspensas e coloidais na água (geralmente compostas por fragmentos de argila, silte, plâncton, micro-organismos, e matéria orgânica e inorgânica particulada) obtida pela passagem de um feixe de luz através da amostra e expressa em unidades de Turbidez (Libânio, 2010).

Inicialmente a turbidez era considerada um parâmetro estético, pois águas com turbidez alta causavam repulsa ao consumidor. Com o passar do tempo e a constatação de que a eficiência da desinfecção, independente do desinfetante empregado, era maior para águas com turbidez baixa, a mesma passou a ser considerada um parâmetro sanitário. A presença de partículas suspensas na massa líquida (turbidez) pode conferir aos microrganismos proteção contra os desinfetantes (efeito-escudo), além da possibilidade de adsorção de substâncias tóxicas pelas partículas transportando-as ao consumidor. Dada a importância da turbidez, a legislação mundial vem propondo valores máximos permitidos cada vez mais restritivos para fins de potabilidade (Libânio, 2010).

2.1.2 Cor Aparente

A cor aparente das águas superficiais em sua normalidade é produzida pela reflexão da luz em partículas coloidais dispersas na massa líquida, tais como compostos orgânicos e compostos de ferro e de manganês. O despejo de resíduos industriais devido à ação humana também pode conferir cor à água (Libânio, 2010).

2.2 Aprendizado de máquina

Aprendizado de máquina (*Machine Learning*) é a ciência que faz uso de uma série de procedimentos e algoritmos para identificar padrões, agrupamentos ou tendências e, então, extrair informações úteis para análise de dados, de maneira totalmente automatizada (Nelli, 2015). Em outras palavras, aprendizado de máquina engloba métodos matemáticos/estatísticos e treina algoritmos para identificar padrões.

O aprendizado de máquina pode ser dividido em supervisionado e não supervisionado. Um algoritmo de aprendizagem supervisionada poderá ser tanto de

classificação quanto de regressão. Este tipo de algoritmo irá lidar com problemas em que já existe uma classificação prévia dos dados e se deseja saber a qual categoria o dado que não está classificado irá pertencer. Em um algoritmo de aprendizagem não-supervisionado o mesmo irá aprender com os dados não-rotulados, ou seja, estes dados não tem os resultados desejados a serem estimados, então deve-se por meio de exploração dos dados verificar os possíveis relacionamentos entre eles a fim de agrupá-los e organizá-los em função de suas afinidades (Netto, 2021).

Neste trabalho são apresentados dois algoritmos utilizados durante a pesquisa: k-means e regras de associação. O algoritmo K-Means ou K-Médias, irá fornecer uma classificação de informações de acordo com os próprios dados dispostos em sua relação, estes dados classificados pelo k-means irão ser indicados e destinados a uma classe (*Cluster*).

A clusterização dos dados irá agrupar os dados de maior relevância para o problema proposto, separados-os em conjuntos, subconjuntos ou grupos de informações, os unindo de tal forma que os elementos irão compartilhar características comuns dentro dos *clusters* criados (Schmuller, 2018).

Esta classificação será apoiada em análises e comparações entre os valores numéricos dos dados, assim, o algoritmo irá fornecer uma classificação sem a necessidade de uma supervisão humana, ou seja, sem nenhuma pré-classificação. Por causa desta peculiaridade, o K-Means é visto como um algoritmo de mineração de dados não-supervisionado.

As regras de associação são uma técnica não-supervisionada que visa encontrar as associações mais relevantes entre um determinado conjunto de dados, aplicado a outro conjunto de dados, ou seja, ela irá encontrar os mais frequentes relacionamentos descrevendo suas relações e padrões entre os dados analisados (Tan, 2019). Esta técnica visa ainda encontrar os padrões mais frequentes entre os atributos de um conjunto de dados.

Toda a regra de associação deve ter medidas que sinalizem a sua valoração e a sua importância, duas destas medidas são a de suporte e confiança. O suporte irá indicar o quanto a regra de associação é importante para a base de dados. A medida de confiança irá determinar o número de ocorrências onde X e Y irão aparecer, o mesmo irá ser dividido pelo número de eventos onde X aparecerá, indicando o quanto Y será relacionado com X (Tan, 2019).

2.3 Séries Temporais

As séries temporais são conjuntos de dados coletados ao longo dos anos, estas informações são utilizadas para analisar, prever padrões e tendências. Esses conjuntos de dados são regularmente usados em várias áreas, como medicina, robótica, meteorologia e estudos comportamentais (Nielsen 2021). Ao analisar as séries temporais, é importante considerar fatores como sazonalidade, tendência e ciclos, pois eles podem influenciar os resultados. Segundo Nielsen (2021), é possível diagnosticar comportamentos passados e prever o comportamento futuro utilizando séries temporais.

Dentre os modelos estatísticos existentes, neste trabalho optou-se pelo SARIMA (*Seasonal Autoregressive Integrated Moving Average*), que é um modelo estatístico amplamente utilizado para previsão e análise de séries temporais. Este modelo combina elementos autorregressivos (AR) e médias móveis (MA), e leva em consideração a tendência e a sazonalidade dos dados, permitindo prever valores futuros com base em padrões passados, por isso ele é largamente utilizado em áreas como: economia, finanças e meteorologia, onde existem intervalos de dados regulares e específicos.

2.4 Análise de componentes principais

A Análise de Componentes Principais (*Principal Component Analysis* - PCA) é um algoritmo estatístico e está fundamentado na variância dos dados, pois ele cria uma nova representação, com um volume menor de informações, mas mantendo a variância entre os dados. Esta nova representação será descorrelacionada, mas cada representação criada irá possuir uma variância (Mingoti, 2020).

Esta técnica ainda irá agrupar os dados utilizando a sua variação, indicando que os indivíduos serão agrupados pela sua variância e suas características, possibilitando assim a redução de dimensionalidades e por consequência um número menor de variáveis a serem estudadas, mas mantendo a integridade de informações necessárias.

O PCA irá extrair $p \leq m$ de um conjunto de dados, onde m é o número de variáveis independentes e p é o número menor ou igual a quantidade de m , na qual p irá criar novas variáveis independentes, que explicarão a melhor variação na base de

dados, sem considerar a dimensão dependente.

2.5 Ferramentas utilizadas

2.5.1 Python

Python é uma linguagem de programação de alto nível, amplamente utilizada em aplicações *web*, *desktop* e *Machine Learning*. Ela tende a ter uma baixa curva de aprendizado e é gratuita para os desenvolvedores que queiram utilizá-la. Ao longo dos anos a linguagem *Python* tem adquirido uma grande aceitação junto à comunidade de cientistas de dados e desenvolvedores, encontrando-se em primeiro lugar como linguagem mais utilizada no mundo (<https://pypl.github.io/PYPL.html>).

2.5.2 Google Colab

O Google Colab é uma ferramenta que permite a utilização de código fonte com texto e imagens, gerando um resultado conhecido como *notebook*, dentro de um ambiente que poderá ser compartilhado com outros desenvolvedores, podendo receber contribuições e atualizações (<https://colab.research.google.com/>).

A linguagem *Python* é a mais utilizada nesta ferramenta, mas é possível trabalhar com outras linguagens como: R, Julia, *Swift*, etc.

2.5.3 SARIMA

SARIMA (*Seasonal Autorregressive Integrated Moving Average*), é uma extensão do ARIMA. Segundo Nielsen (2021), ele suporta dados de séries temporais univariados com um componente sazonal, sendo esta técnica usada para prever dados de séries temporais ao longo do tempo. O SARIMA modela as relações entre valores passados e presentes identificando padrões nos dados, ele usa uma combinação de modelos de auto-regressão (AR) e média móvel (MA), bem como diferenciação, para capturar os componentes sazonais dos dados. Ao configurarmos um modelo SARIMA, será necessário selecionar hiperparâmetros, tendência (p), ordem de diferença (d) e

ordem de média móvel (q). Os elementos sazonais incluem a ordem de auto-regressão sazonal (P), ordem de diferença sazonal (D) e ordem de média móvel sazonal (Q).

2.5.4 Função autoarima()

A função `autoarima()` é uma ferramenta para análise de séries temporais. Através dela, é possível automatizar a seleção dos melhores parâmetros para a criação do modelo. A função utiliza algoritmos de busca para encontrar a combinação de parâmetros que melhor se ajusta aos dados, levando em consideração critérios como o AIC (*Akaike Information Criterion*) ou o BIC (*Bayesian Information Criterion*). Além disso, a `autoarima()` também permite a inclusão de sazonalidade nos modelos, o que a torna importante quando se trabalha com dados sazonais (Nielsen, 2021).

2.5.5 Análise Estacionária

A análise estacionária tem como objetivo verificar se uma série temporal é estacionária ou não em sua média, analisando se sua variância e autocorrelação não estão mudando ao longo do tempo. Isso indica que os valores da série temporal não estão aumentando ou diminuindo ao longo dos anos, e não há correlação entre as informações observadas em diferentes pontos no tempo. Para verificar se uma série temporal é estacionária ou não utilizamos o teste de Augmented Dickey- Fuller (ADF). Segundo Barros *et al.* (2017), o teste ADF compara a estatística de teste da série temporal com uma distribuição crítica de valores. Se o teste for menor que o valor crítico e o *p-value* for menor que 0,05 então a hipótese nula é rejeitada e conclui-se que a série temporal é estacionária. Se o teste for maior que o valor crítico, então a hipótese nula não é rejeitada e conclui-se que a série temporal não é estacionária.

2.5.6 Desvio Padrão

O desvio padrão é uma medida de dispersão de dados. Segundo Azevedo (2016), ele mede a distância média entre os dados e a média. Quanto maior o desvio padrão, mais dispersos estão os dados. Quanto menor o desvio padrão, mais próximos estão os dados. O desvio padrão é uma medida útil para entender a distribuição de

dados. Ele pode ser usado para comparar diferentes conjuntos de dados e para identificar *outliers* (valores discrepantes).

2.5.7 Erro Padrão

O erro padrão é uma medida da incerteza de uma estimativa, ele é a medida de uma variação amostral em relação a uma média populacional dos dados. Sendo assim, este teste ajuda a verificar a confiabilidade amostral calculada, tendo seu resultado na mesma unidade de medida do valor amostral (Paes, 2008).

2.5.8 Autocorrelação e Autocorrelação Parcial (ACF e PACF)

A autocorrelação (ACF) e a autocorrelação parcial (PACF) são duas ferramentas estatísticas usadas para analisar séries temporais. Elas são usadas para determinar o grau de correlação entre os valores de uma série temporal em diferentes *lags*. Segundo Alves (2020), uma correlação nos diz o quão relacionado é um valor com outro, porém, isso não necessariamente significa que um valor influenciou no outro, já a ACF mede a autocorrelação entre valores vizinhos da série temporal, e a PACF calcula a autocorrelação entre valores aleatórios distribuídos em uma série temporal.

2.6 Trabalhos Relacionados

Nesta seção são apresentados alguns trabalhos relacionados com a pesquisa proposta. O trabalho de García-Ávila et. al (2020) teve o intuito de reduzir, otimizar e melhorar o consumo de água durante o processo de filtração, analisando as variáveis que afetam a saturação do leito filtrante, utilizando os métodos indicados pelo Centro Panamericano de Engenharia Sanitária e Ciências Ambientais (CEPIS). Foram avaliados parâmetros como velocidade de filtração, qualidade de filtração inicial e turbidez comparando os valores antes e depois das intervenções e utilizando o Teste t de *Student* para verificar se as diferenças foram estatisticamente significativas. O resultado indicou melhoria da qualidade da água, bem como maior segurança no

consumo da água tratada ao se medir o desempenho dos filtros utilizados no tratamento.

Já o artigo de Meyers et. al (2016) propõem uma metodologia para realizar a previsão da turbidez, dando suporte gerencial em tempo real e fornecendo um alerta do aumento ou diminuição da turbidez num sistema de tratamento de água. A metodologia sugerida utiliza *Machine Learning* aplicada a dados anteriormente salvos para aprender com a turbidez, a partir de uma rede neural, cujos principais parâmetros manuseados são vazão e turbidez. Esta metodologia é totalmente orientada a dados, não requerendo nenhum custo hidráulico ou material específico, somente as informações armazenadas ao longo do tempo. Neste contexto, os resultados obtidos indicam que, previsões poderão ser realizadas com uma razoável precisão da turbidez em até 15 minutos antes da mesma ter um aumento.

Em Ilic et al. (2022) é aplicado o algoritmo de classificação *Naive Bayes*, para a criação de um modelo de previsão baseado em parâmetros de qualidade da água, tais como: temperatura, pH, condutividade elétrica, oxigênio dissolvido, entre outros. Amostras foram coletadas de 2013 até 2019 na província de Vojvodina (Sérvia), de pontos onde a qualidade da água era notadamente diferente, devido às condições naturais do ambiente. Estas diferenças são necessárias para o treinamento do modelo, pois assim ele irá cobrir o maior número de classes de qualidade da água. Tais amostras foram utilizadas para o treinamento do modelo de previsão, onde foi possível prever a classificação da água em 64 amostras dos 68 pontos coletados.

Esses trabalhos são exemplos tanto da importância dos estudos da qualidade da água para implementação de melhorias técnicas, quanto da possibilidade de análise através de diferentes metodologias, tais como *Machine Learning* e *Naive Bayes*, para a modelagem preditiva de resultados.

3. Estudo de Caso

3.1 Metodologia

Este trabalho propõe o desenvolvimento de um modelo que valide a turbidez como principal dado físico-químico no tratamento da água, indicando a sua importância como fator preponderante em sua qualidade, assim como auxiliar os gestores na tomada de decisão. A Figura 2 ilustra os passos que deverão ser seguidos para execução da metodologia proposta.



Figura 2. Etapas da metodologia proposta.

3.2 Etapas do Desenvolvimento

1. Definição do problema: Ao desenvolvermos um projeto de análise de dados, devemos identificar inicialmente uma problemática, averiguando a sua autenticidade, podendo confirmar ou refutar o problema proposto após o mesmo ser analisado. Segundo Netto (2021), o problema, logo após identificado, poderá ser classificado como gerencial (quando é utilizado para dar suporte a uma decisão superior) ou operacional (como identificar quais fatores podem otimizar o controle de qualidade na produção). Uma das etapas essenciais à clarificação da água para abastecimento público é a filtração, cujo parâmetro a ser acompanhado para certificação da sua eficiência é a turbidez. Sendo assim, o objetivo deste trabalho é

tentar prever os valores da turbidez antes dos filtros, a partir dos valores de turbidez nas etapas anteriores, subsidiando as correções a serem feitas no decorrer do tratamento.

2. Preparação dos dados: Após a definição do problema, deve-se coletar e preparar os dados para posterior análise, tendo como objetivo garantir a precisão e consistência dos mesmos, atestando, assim, uma exploração válida e realista dos valores coletados. Segundo Netto (2021), a preparação dos dados tem as seguintes etapas:

1. Extração e carga dos dados.
2. Exploração e visualização dos dados.
3. Transformação dos dados.

Neste projeto, inicialmente foram obtidos dados secundários coletados ao longo do tratamento, no período de 2016 a 2021. Os mesmos tiveram a sua carga realizada na ferramenta *on-line* e gratuita *Google Colab*, que permite criar, executar códigos e visualizar gráficos e *dashboards* (Netto, 2021).

Em seguida os dados foram explorados e visualizados em busca de padrões e relações entre as informações. Um exemplo de planilha encontra-se na Figura 3.

ID	Ano	Mês	Alumínio Resolvel (mg Al/L)	Cloro (mg Cl/L)	Cor Aparente (ppb)	DO ₂ (mg O ₂ /L)	Ferro Total (mg Fe/L)	Fósforo Total (mg PO ₄ /L)	Manganês (mg Mn/L)	Nitrato (mg NO ₃ /L)	Níquel (mg Ni/L)	Nitrogênio Amônio (mg NH ₄ /L)	Oxigênio Dissolvido (mg O ₂ /L)	pH	Sólidos Dissolvidos Totais a 180°C	Sulfato Total (mg SO ₄ /L)	Sulfeto (mg H ₂ S/L)	Temperatura ar (°C)	Temperatura água (°C)	Turbidez (uT)	Conteúdo Bacteriano Total	Índice Coliforme a Totale100 ml	Índice Escherichia coli/100 ml
1	2016	1																					
2	2016	2	0.04	12.00	200.00	0.00	1.30	0.47	0.24	0.04	0.00	0.07	6.00	7.15	109.80	3.00	0.89	32.00	27.00	19.20	80.00	23000.00	45.00
3	2016	3	0.00	9.73	419.00	2.70	3.94	0.47	0.90	0.10	0.00	0.50	8.00	7.17	183.20	2.53	0.00	23.00	21.00	39.70	120.00	23000.00	330.00
4	2016	4	0.03	9.03	187.00	2.30	3.68	0.39	0.58	0.53	0.01	0.05		6.78	126.44	2.25	0.85	22.00	22.00	29.80	50.00	23000.00	170.00
5	2016	5	0.03	8.09	2.18	0.33	0.24	0.45	0.74	0.35	0.01	1.26	7.90	7.22	92.20	9.42	0.85	15.00	18.00	30.20	50.00	23000.00	170.00
6	2016	6	0.01	7.90	134.00	1.00	2.91	0.33	0.83	0.31	0.00	1.14	7.90	6.90	99.60	12.30	0.89	13.00	13.00	26.70	40.00	7900.00	93.00
7	2016	7	0.00	8.51	100.00	2.30	1.35	0.39	0.56	0.12	0.01	0.78	8.90	6.82	109.40	9.06	0.43	14.00	13.00	21.00	25.00	23000.00	170.00
8	2016	8		11.34		0.81		0.42	0.55	0.14	0.03	1.33	8.40	6.99		8.05	0.89	16.00	13.00	21.50	50.00	4900.00	45.00

Figura 3. Exemplo de planilha de dados do Sanep.

Ainda na etapa da preparação dos dados, foi possível visualizar ausência de dados para alguns atributos, ou seja, campos que não foram preenchidos ao longo da jornada de trabalho. Os dados ausentes correspondem a 12% do total de dados. Os mesmos foram quantificados através de codificação e visualizados (Figura 4).

↳ Temperatura água (°C)	14
Temperatura ar (°C)	14
Sólidos Dissolvidos Totais a 180°C	14
Ferro Total (mg Fe/L)	11
Alumínio Residual (mg Al/L)	9
Cor Aparente (uH)	9
Fósforo Total (mg P04/L)	8
Índice Escherichia coli/100mL	7
DBO5 (mg O2/L)	7
Manganês (mg Mn/L)	7
Índice Coliformes Totais/100mL	6
Conteúdo Bacteriano Total	6
Nitrato (mg NO3/L)	6
Turbidez (uT)	6
Oxigênio Dissolvido (mg O2/L)	6
Sulfeto (mg H2S/L)	5
Nitrogênio Amoniacal (mg NH3/L)	5
Sulfato Total (mg SO4/L)	5
pH	5
Nitrito (mg NO2/L)	5
Cloretos (mg Cl/ L)	5
Ano	0
Mês	0
ID	0
dtype: int64	

Figura 4. Quantidades de dados faltantes por variáveis.

Para suprir esta lacuna de informações foi utilizada a mediana, que segundo Vieira (2019) oferece como vantagem fornecer um valor típico do conjunto de dados, pois não é afetada por valores extremos.

ID	Ano	Mês	Alumínio Residual (mg Al/L)	Cloretos (mg Cl/ L)	Cor Aparente (uH)	DBO5 (mg O2/L)	Fósforo Total (mg P04/L)	Ferro Total (mg Fe/L)	Manganês (mg Mn/L)	pH	Sólidos Dissolvidos Totais a 180°C	Sulfato Total (mg SO4/L)	Sulfeto (mg H2S/L)	Temperatura ar (°C)	Temperatura água (°C)	Turbidez (uT)	Conteúdo Bacteriano Total	Índice Coliformes Totais/100mL	Índice Escherichia coli/100mL	
0	1	2016	1	0.04	10.04	219.77	5.50	1.40	0.39	0.50	7.16	155.20	0.00	0.00	21.0	21.0	30.20	30.0	23000.0	45.0
1	2	2016	2	0.04	12.00	204.00	0.00	1.30	0.47	0.24	7.10	150.00	0.00	0.00	22.0	21.0	19.20	30.0	20000.0	40.0
2	3	2016	3	0.00	9.70	410.00	2.70	3.54	0.47	0.30	7.17	163.20	2.20	0.00	23.0	21.0	38.70	120.0	23000.0	100.0
3	4	2016	4	0.01	9.00	187.00	2.00	0.60	0.39	0.50	6.78	130.40	2.20	0.00	22.0	22.0	24.00	30.0	15000.0	110.0
4	5	2016	5	0.03	0.00	2.10	0.20	0.24	0.40	0.74	7.22	92.20	0.40	0.00	18.0	19.0	30.20	30.0	23000.0	110.0
5	6	2016	6	0.04	0.00	219.77	5.50	1.40	0.37	0.40	7.11	122.00	0.00	0.00	22.0	20.0	17.02	30.0	20000.0	40.0
6	7	2016	7	0.04	9.30	216.00	0.00	0.64	0.43	0.30	7.21	150.00	10.10	0.00	21.0	19.0	18.00	120.0	20000.0	100.0
7	8	2017	1	0.12	12.00	950.00	1.10	0.00	0.20	0.10	6.80	161.20	0.00	6.41	21.0	21.0	30.20	180.0	23000.0	0.0
8	9	2017	3	0.01	10.00	77.00	0.10	0.34	0.24	0.13	7.17	90.00	0.27	1.30	24.0	20.0	4.00	40.0	3000.0	0.0
9	10	2017	5	0.36	14.00	410.00	0.10	0.40	0.34	0.30	7.34	161.20	7.40	44.0	24.0	21.0	6.40	80.0	30000.0	0.0

Figura 5. Amostra da planilha de dados já com os dados reorganizados.

Após o preenchimento dos dados faltantes com a mediana, foi possível observar as correlações entre as variáveis, utilizando um *heatmap*, ou seja, um mapa de calor que irá associar cores à intensidade das correlações. Para verificar a correlação entre as variáveis, foi utilizado o coeficiente de *Pearson*, que mede o grau de correlação utilizando o cálculo de direção positiva ou negativa entre -1 e 1 (Vieira, 2019). A Figura 6 mostra o *heatmap* obtido.

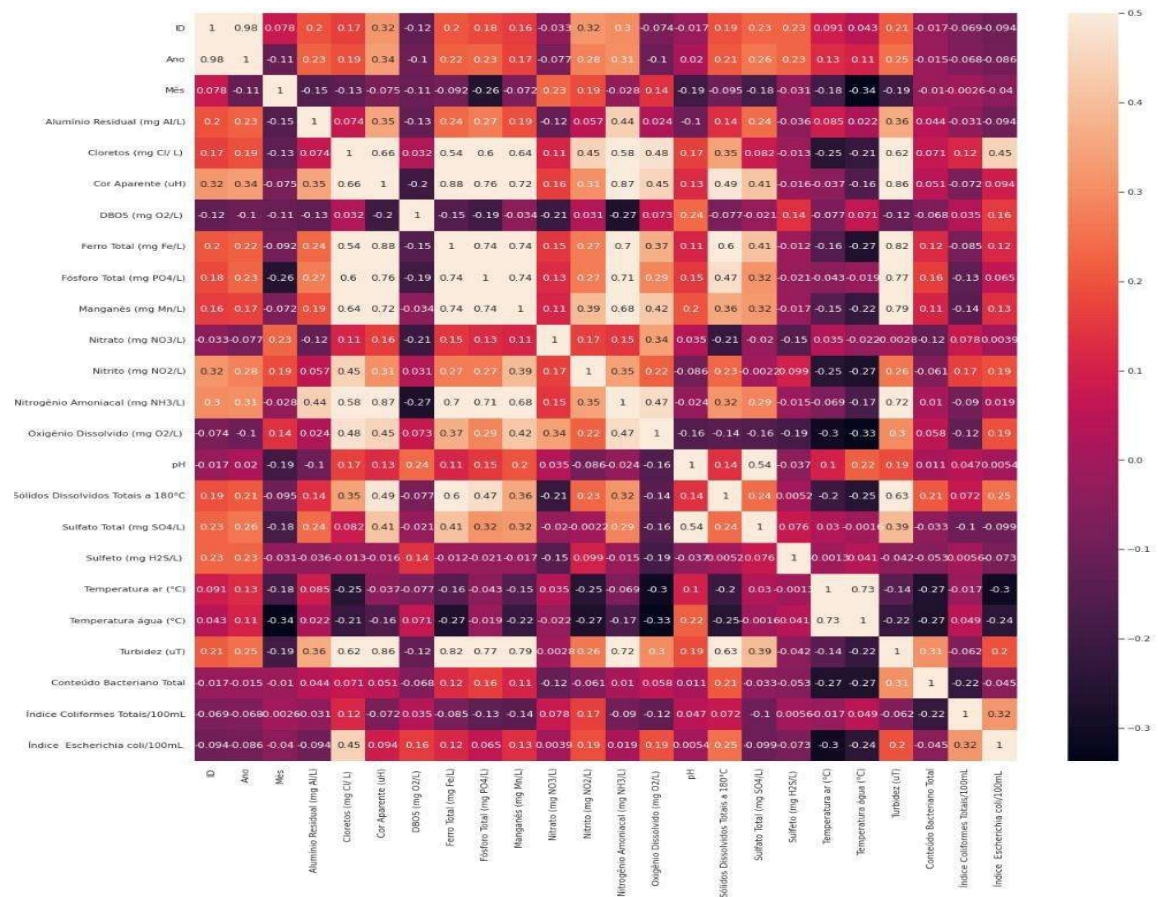


Figura 6. *Heatmap* com todas as variáveis disponíveis na base de dados.

Devido ao grande número de variáveis presentes na base de dados, inicialmente foi feita uma seleção manual com as variáveis mais fortemente correlacionadas ($r > 0,60$), construindo-se, a seguir, um novo *heatmap* com um esse número de variáveis (Figura 7). Desta forma, gera-se uma base de dados mais equilibrada para posterior redução de dimensionalidades através de um algoritmo específico.

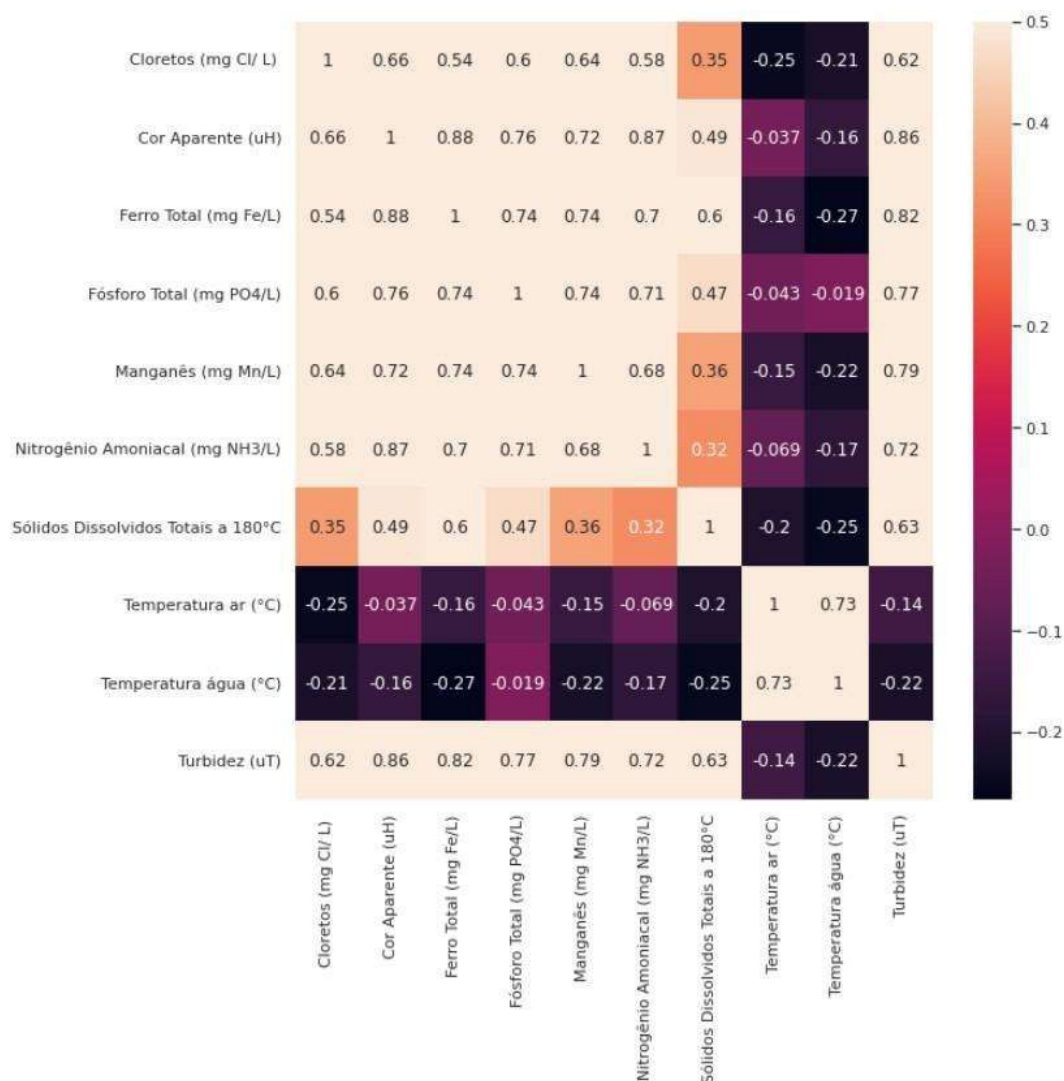


Figura 7. *Heatmap* com as variáveis que apresentaram coeficientes de correlação de Pearson maiores do que 0,60.

A fim de evitar uma quantidade excessiva de variáveis altamente correlacionadas, que possam se tornar redundantes ou que não apresentem informações úteis ao problema proposto, causando assim um *overfitting*, foi utilizada a técnica de PCA para a análise e redução dos parâmetros. Esta técnica é não paramétrica e baseia-se na variância dos dados, ou seja, ela cria uma representação dos dados, com um número de dimensionalidade menor, mas possuindo ainda uma variância entre os dados.

Inicialmente foi pensado para o modelo uma redução de variáveis para um total de cinco, porém, após a implementação do PCA, foi possível perceber que com apenas duas variáveis seria possível descrever e explicar quase 99% da variância da base de dados: a variável turbidez explica 96% e a Cor Aparente explica 0,03 % (Figura 8).

```

Variância explicada:
[8.25091860e+04 2.79693369e+03 2.21836561e+02 3.26878489e+01
 1.58038102e+01]
Variância explicada (ratio):
[0.96 0.03 0. 0. 0. ]

```

Figura 8. Variância das 5 variáveis mais fortemente relacionadas conforme o PCA.

Através do gráfico da Figura 9 é possível visualizar que com apenas duas variáveis, Turbidez e Cor aparente (representadas no eixo x), podemos descrever quase 99% das informações do banco de dados, representadas no eixo y como % de variância explicada.

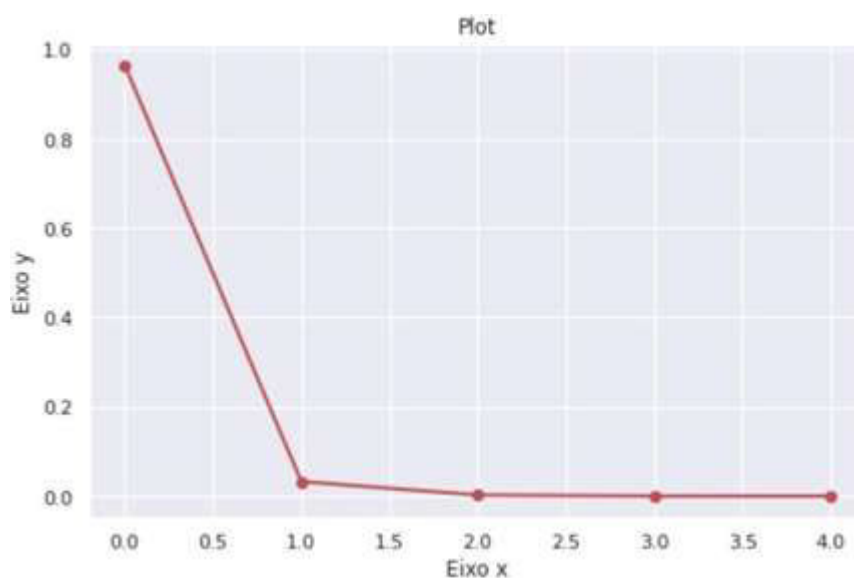


Figura 9. Gráfico de % de variância explicada em função do número de variáveis.

Ainda, para salientar a importância e a ligação entre as duas dimensões, é possível ver, observando a Figura 10, a qual representa a Turbidez no eixo x e a Cor Aparente no eixo y, que quanto maior a Turbidez, maior é a Cor Aparente, e quanto menor a Turbidez, menor a Cor aparente.

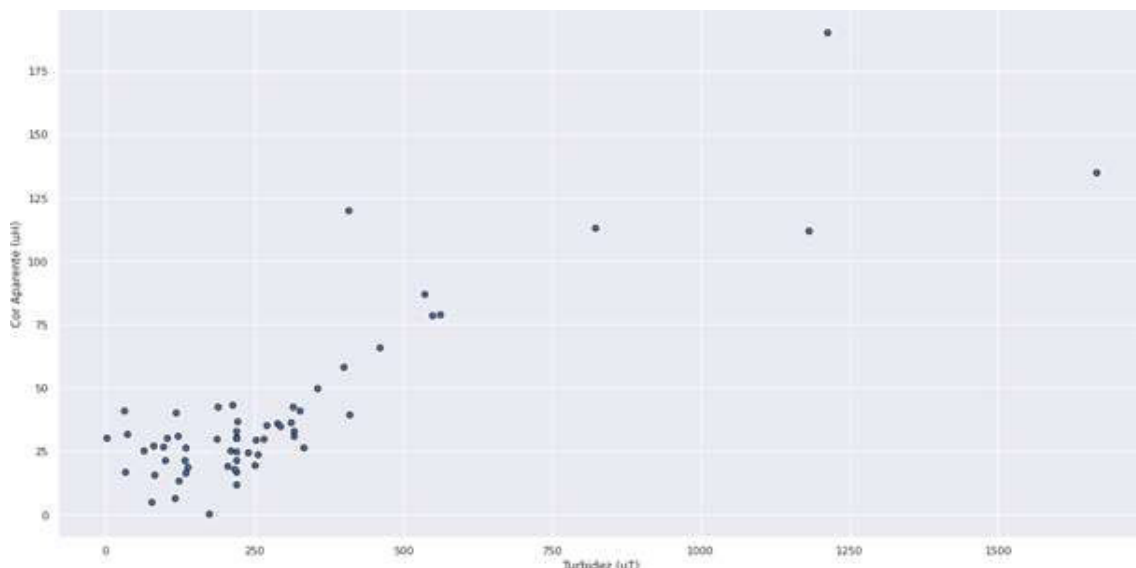


Figura 10. Gráfico de dispersão entre Turbidez e Cor Aparente.

3. Escolha do algoritmo: Após a etapa de processamento, temos uma das fases mais importantes na implementação de um modelo: a escolha do algoritmo. Esta escolha está diretamente ligada ao objetivo proposto, bem como ao sucesso do projeto, e pode ser chamada de modelagem preditiva (Netto, 2021).

Para o modelo proposto foi pensado no algoritmo não supervisionado K-means, cujo objetivo é descobrir objetos semelhantes através da clusterização, ou seja, através dos agrupamentos dos dados (Schmuller, 2019).

Inicialmente foi aplicada à base de dados (que agora consiste somente nas variáveis Turbidez e Cor Aparente, conforme elencado pelo PCA) a função *StandarScaler* da biblioteca *Sklearn* para calcular a média e o desvio padrão dos dados de treinamento.

Para verificarmos a quantidade necessária de Clusters em nosso projeto, foi utilizado o método *Elbow* do K-means, que consiste em somar a distância quadrada de cada ponto ao centro atribuído (Schmuller, 2019), cuja plotagem permitiu interpretar que o número ideal de clusters seria três (Figura 11).

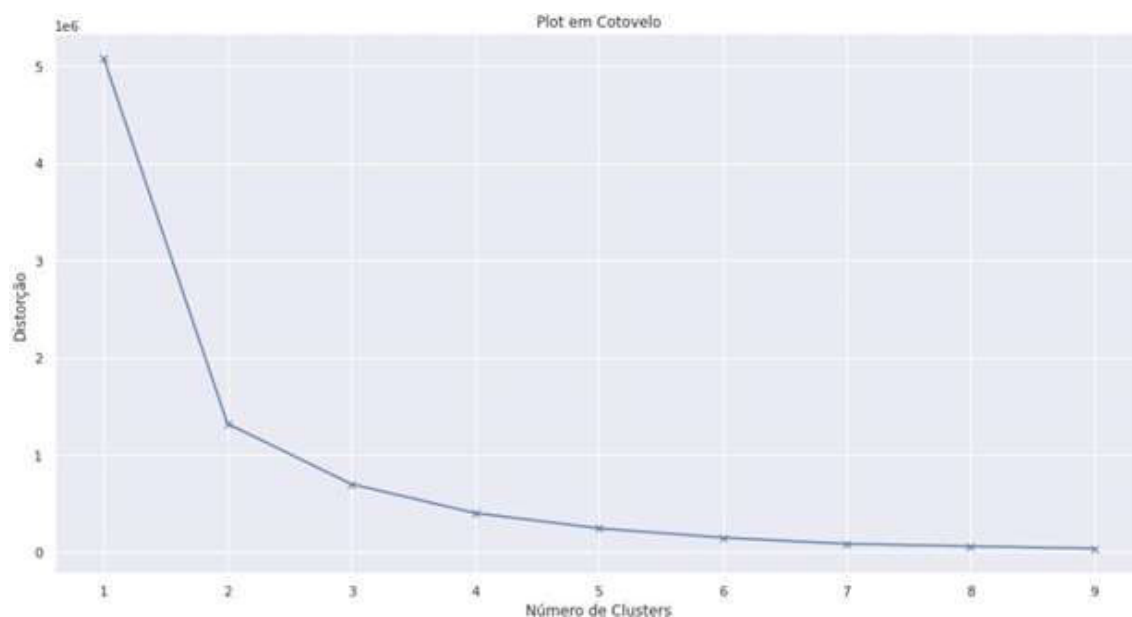


Figura 11. Número indicado de clusters conforme o método Elbow do K-means.

Considerando três clusters como quantidade ideal foi possível realizar o treinamento e criar um centróide para cada cluster. Os mesmos atuam como protótipos, onde pontos aleatórios no *dataset* são escolhidos para comporem as coordenadas dos centróides iniciais. Então, ao plotar estes dados, é possível visualizar e interpretar como os dados estão distribuídos e relacionados.

4. Visualização e interpretação: A visualização e interpretação dos dados estão intimamente ligadas, pois é neste momento que podemos analisar a exibição gráfica dos dados e identificar as tendências, padrões e *outliers*. Expor informações na sua forma bruta iria provavelmente gerar interpretações incorretas e dificultar a tomada de decisão por parte dos gestores. No entanto, a visualização dos dados irá dar uma capacitação para melhor absorver as informações por parte dos gestores, assim como conseguir transmitir ideias, conceitos e tendências complexas.

Em nosso modelo foi possível visualizar e interpretar duas informações: a primeira é que, ao trazer os dados escolhidos como centróides para valores reais é possível perceber que, quando os valores de turbidez estiverem em média 86, 145 e 27, os valores de cor aparente estarão 533, 1.352 e 194 (Figura 12).

```
array([[ 533.57142857,   86.01428571],
       [1352.86666667,  145.66666667],
       [ 194.83735849,   27.47830189]])
```

Figura 12. Centróides em valores reais.

Em nossa segunda interpretação após a inclusão dos dados e centróides é possível ter a percepção que existe um aumento escalar da relação entre a Cor Aparente e a Turbidez, ou seja, quanto maior o valor de Turbidez maior será o valor de Cor Aparente e vice-versa (Figura 13).



Figura 13. Gráfico mostrando o valor escalar entre turbidez e cor aparente.

5. Apoio à tomada de decisão: Nesta última fase, o modelo visa dar suporte a tomada de decisão dos gestores, mostrando seu funcionamento, bem como inserindo os usuários no contexto do que se quer gerar, etapa fundamental para o sucesso do trabalho. Ao acessar o *Google Colab* o usuário encontrará o modelo funcional, no qual será possível visualizar gráficos, tanto da disposição dos dados, quanto dos resultados gerados, além da relação entre os parâmetros Turbidez e Cor aparente. Com esse modelo será possível realizar a predição da turbidez, informação importante para que o gestor possa organizar a estratégia de tratamento e antecipar, quando necessário, a compra de produtos físico-químicos para o tratamento da água na ETA.

4. Resultados e Discussões

Por sua relevância (revelada pelo PCA), a Turbidez foi a variável selecionada para fazer parte da série temporal.

Após a disposição dos dados de turbidez ao longo dos anos, foram analisadas a tendência, o resíduo e a sazonalidade. Por apresentar sazonalidade optou-se pela utilização do modelo SARIMA para elaboração da série temporal (Figura 14).

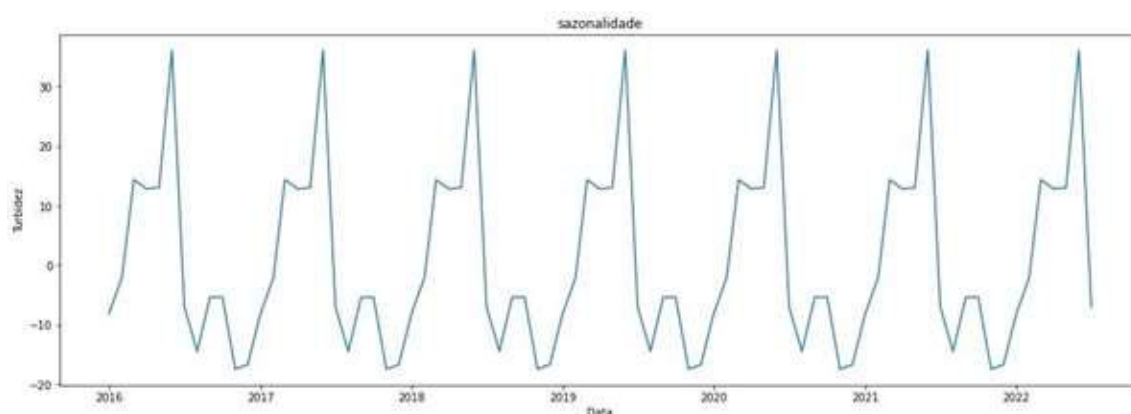


Figura 14. Gráfico mostrando a sazonalidade dos dados ao longo dos anos.

Testes como o de análise estacionária, desvio padrão, erro padrão, ACF e PACF, forneceram uma base ainda mais consolidada para a utilização do modelo proposto. A análise estacionária indicou um valor estatístico igual a -3,443, menor que o valor crítico de 5% que é igual a -2,889, indicando que a série é estacionária. Os testes ACF e PACF mostraram que não existe uma autocorrelação entre os valores da série, nem mesmo uma autocorrelação parcial, pois a maioria dos dados encontra-se dentro do intervalo de confiança, como é mostrado nas Figuras 15 e 16.

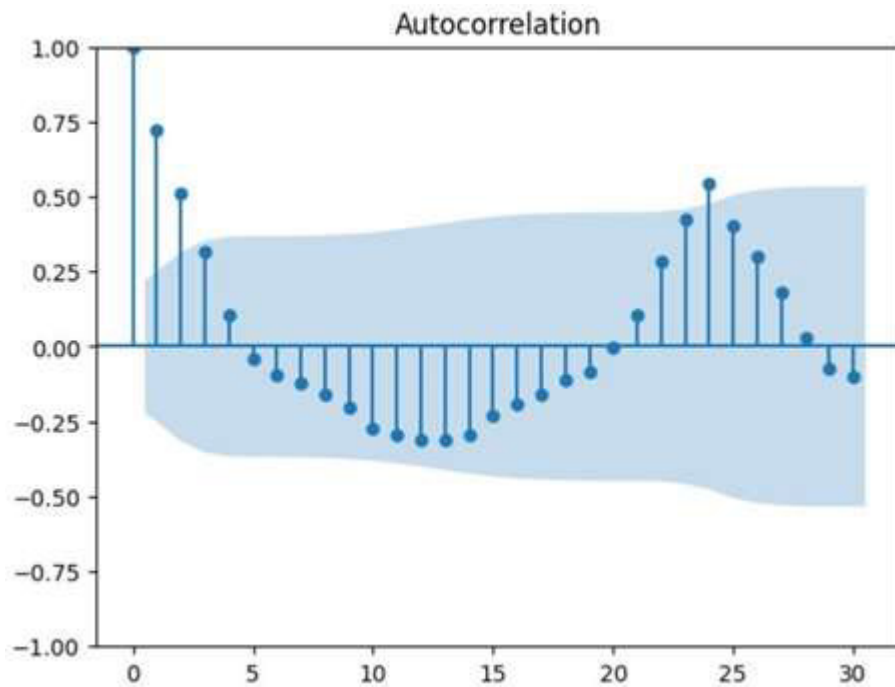


Figura 15. Autocorrelação entre valores vizinhos dentro do intervalo de confiança.

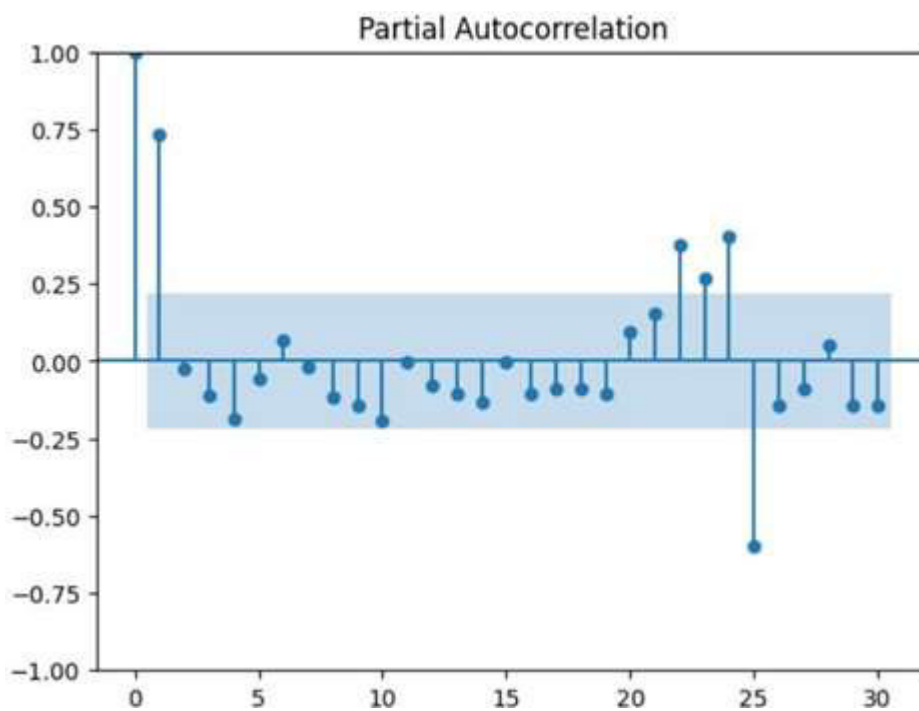


Figura 16. Autocorrelação Parcial entre valores vizinhos dentro do intervalo de confiança.

Como sugere Nielsen (2021), identificar os parâmetros de um modelo SARIMA pode ser uma tarefa complicada, por isso, o mesmo sugere a utilização da função `autoarima()` do pacote *forecast*. Sua utilização neste estudo forneceu os melhores parâmetros para o modelo proposto.

Com os parâmetros indicados pelo `autoarima()`, o modelo foi treinado, sendo

possível visualizar e analisar os valores da turbidez ao longo dos próximos 12 meses. Com uma dispersão dos dados no valor de 14.67 NTU e um taxa de erro padrão 4.42, esses valores poderão servir como base tanto para avaliação do modelo quanto controle de turbidez ao longo do ano (Figura 17).

2022-08-01	57.555598
2022-09-01	43.443275
2022-10-01	40.758482
2022-11-01	29.754973
2022-12-01	26.523560
2023-01-01	30.786198
2023-02-01	14.145444
2023-03-01	13.889863
2023-04-01	13.502688
2023-05-01	12.698688
2023-06-01	12.061118
2023-07-01	11.319523

Figura 17. Valores previstos pelo modelo para os próximos 12 meses.

Ao observarmos a Figura 18, é possível ver os dados dispostos ao longo dos anos, classificados de forma a visualizar as informações da série temporal, sendo possível diagnosticar sua tendência ao longo do tempo. Em azul temos os dados que foram utilizados para treinamento, inseridos no algoritmo para a criação do modelo. Geralmente essas informações costumam representar 70% da totalidade dos dados. Em vermelho, temos os dados de teste, que também foram inseridos no algoritmo para simular previsões reais que o modelo realizará. Os dados de teste costumam representar 30% da totalidade dos dados. É possível, ainda, verificar que há uma similaridade entre as linhas de teste e treinamento, permitindo assim que a sua performance real seja diagnosticada.

Ainda na Figura 18, em verde, temos o resultado da predição, indicando uma tendência de queda dos valores de turbidez, depois de uma alta no ano de 2022. Ter posse desse tipo de informação possibilitará uma melhor organização por parte dos gestores, que poderão analisar e tomar a melhor decisão, visando a qualidade da água e do tratamento.

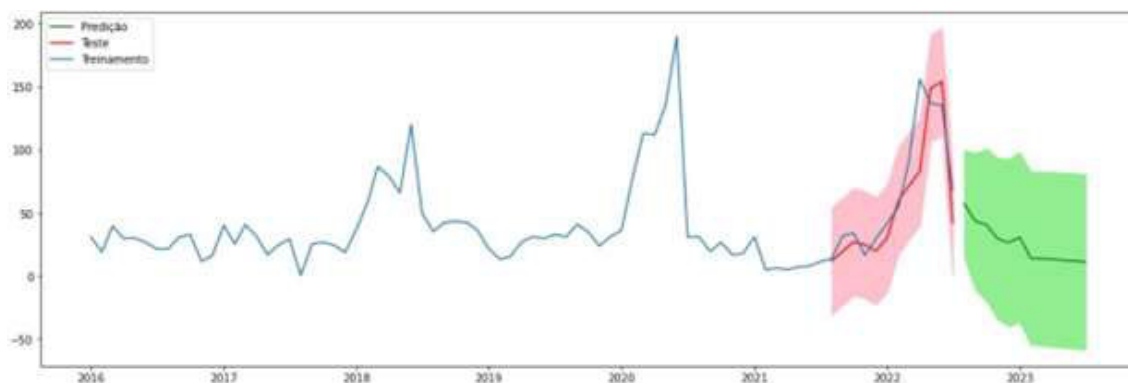


Figura 18. Série temporal de turbidez para os próximos 12 meses.

Com todas as informações, dados gerados e gráficos disponibilizados, foi possível criar um formulário que foi respondido por 3 (três) profissionais da área técnica, identificadas com a instituição e com mestrado em Engenharia Química, mestrado em Recursos Hídricos e mestrado em Ciências Biológicas, respectivamente. As respostas das três profissionais, que possuem expertise em tratamento de água com mais de 6 anos de experiência trabalhando em ETAs, se encontram no Apêndice A.

Foi possível perceber um entendimento dos dados por parte das profissionais, que na primeira pergunta mostraram entendimento sobre a relação entre a turbidez e a cor aparente, atestando assim a importância destes dados físico-químicos para o projeto. Na segunda questão duas das técnicas notaram em dados reais (centróides) o comportamento dos parâmetros e como eles se comportam dentro do corpo hídrico, porém, a terceira técnica não soube responder.

Nas questões 3 e 4 é mostrado o resultado das previsões para 12 meses tanto graficamente quanto em valores reais, onde na questão número três, duas das profissionais entenderam a importância de se ter valores previstos para turbidez ao longo do ano e o efeito que isso poderá causar na tomada de decisão por parte dos gestores, e a terceira indicou que os valores previstos estão abaixo dos geralmente obtidos na ETA. Ao observar as respostas da questão número 4, foi possível perceber que as profissionais analisaram aspectos diferentes do gráfico, a primeira analisou os picos de turbidez indicando a precipitação como fator principal do aumento da turbidez, a segunda atentou-se também para os picos de turbidez ao longo dos anos e como isso influencia na tomada de decisão dos gestores, já a terceira percebeu a tendência da previsão gerada para os próximos 12 meses, indicando que isto poderá auxiliar também na tomada de decisão dos administradores da Barragem Santa Bárbara.

5. Conclusões

O objetivo principal deste trabalho foi construir um modelo robusto e confiável capaz de validar a turbidez como um indicador preponderante da qualidade da água tratada na ETA. Desta forma, fornecer dados e gráficos para a criação de futuras alternativas que possam auxiliar os gestores na administração e gestão da água.

Inicialmente para a definição do problema se fez necessário identificar e entender quais os parâmetros físico-químicos seriam os mais importantes para o trabalho. Portanto, foi necessário realizar a coleta dos dados secundários e pré-processá-los, resultando na melhora da qualidade dos dados e na indicação dos dados mais relevantes para o modelo. Nesta etapa o destaque foi para o PCA (*Principal Component Analysis*) que indicou a Turbidez como parâmetro que mais explica a variabilidade dos dados (96%), mostrando a Turbidez como dado físico-químico fundamental a ser analisado no *dataset*.

Para que os gestores pudessem ter um real entendimento do impacto da Turbidez na qualidade da água, foi possível gerar dados, gráficos e previsões. Através da função *StandarScaler* e do método *Elbow*, foi possível gerar os dados, mostrando a proporcionalidade entre turbidez e cor aparente, tanto em valores reais quanto graficamente; estas informações nos mostram que o aumento da turbidez na água irá implicar também no aumento de sua Cor Aparente.

Outro fator importante criado e idealizado para o trabalho foi a utilização das séries temporais no modelo proposto. A utilização da função *autoarima()* juntamente com o testes realizados, mostrou valores preditos para 12 meses. Aprender com os dados passados e realizar previsões poderá ser de grande utilidade para os administradores da ETA, pois saber o valor aproximado da turbidez para os próximos meses poderá auxiliar na previsão de compra de produtos químicos para o tratamento da água.

Com o trabalho realizado até o momento, pode-se concluir que a utilização de um modelo que possa extrair informações e criar previsionamento no tratamento da água, com base no histórico de dados e da turbidez ao longo dos anos, poderá ser extremamente benéfico para a qualidade da água.

Trabalhar com dados com tanta variância e influência climática tem as suas limitações, mas não impossível de ser implementado, possibilitando serem extraídos

valores e informações relevantes para beneficiar a qualidade de um item vital para a população, a água.

Como forma de melhorar o modelo proposto, ainda poderá ser utilizado o algoritmo XGBOOST, pois tal algoritmo possui maior flexibilidade, utilizando um grande número de hiperparâmetros passíveis de aperfeiçoamento, permitindo assim se ajustar ao cenário e ao problema proposto com maior facilidade.

A criação de uma interface, onde o usuário possa interagir, realizando *upload* de planilhas com os dados físico-químicos, assim como visualizar em tempo real ou até mesmo criar simulações dos dados, poderá melhorar a interação do usuário com o modelo, o deixando ainda mais usual. Outra possibilidade seria solicitar o acesso ao ambiente de trabalho da ETA e validar *in loco* o modelo proposto, pois é crucial para evitar surpresas indesejadas ou problemas inesperados após a implementação do modelo proposto. Esta validação garantiria que todas as variáveis e complexidades do ambiente real seriam consideradas no processo, resultando em uma solução ainda mais robusta e confiável.

Referências

- ALVES, Gisely, **Séries Temporais: Parte 2**, 2020 Disponível em: <https://medium.com/@gisely.alves/s%C3%A9ries-temporais-parte-2-7162d24c5429>. Acesso 10 de maio. 2023
- AZEVEDO, Paulo Roberto Medeiros, **Introdução à estatística**, 3ª ed. Natal: 2016
- BARROS, Anna et al. **Análise de Séries Temporais em R: Curso Introductório**, 1ª ed. São Paulo: 2017.
- BRASIL. **Portaria GM/MS nº 888 de 04 de maio de 2021**, 2021. Disponível em: <https://www.in.gov.br/en/web/dou/-/portaria-gm/ms-n-888-de-4-de-maio-de-2021-318461562>. Acesso 13 de nov. 2022.
- GARCÍA-ÁVILA, Fernando et al. Optimization of water use in a rapid filtration system: a case study. **Water-Energy Nexus**, v. 3, p. 1-10, 2020.
- ILIĆ M, SRDJEVIĆ Z, SRDJEVIĆ B. Water quality prediction based on Naïve Bayes algorithm. **Water Sci Technol**, 2022 Feb;85(4):1027-1039.
- LIBÂNIO, Marcelo, **Fundamentos de qualidade e tratamento de água**, 3ª ed. Campinas 2010.
- MEYERS, Gregory et al. Short-term forecasting of turbidity in a UK water distribution system. **Procedia Engineering**, v. 154, p. 1140-1147, 2016.
- MINGOTI, Sueli Aparecida, **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**, 4ª ed. Belo Horizonte, 2020.
- NETTO, A. **Python para Data Science e Machine Learning**, 1ª ed. Rio de Janeiro: 2021. NIELSEN, A. **Análise Práticas de Séries Temporais**, 1ª ed. Rio de Janeiro: 2021.
- PAES, Ângela Tavares, **Por Dentro da Estatística**, 2008 Disponível em: <https://www2.ufjf.br/ppgsaude/files/2018/11/Desvio-Padrao-erro-e-erro-Padrao-qual-a-diferenca.pdf>. Acesso 19 de maio.2023.
- SANTOS, Gabriel Rosa. **Estudo de Clarificação de Água de Abastecimento Público e Otimização da Estação de Tratamento de Água. Dissertação (Mestrado em Tecnologia dos Processos Químicos e Bioquímicos)** – Escola de Química, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2011.
- SCHMULLER, Joseph, **Projetos em R para leigos**, 1ª ed. Rio de Janeiro 2018. TAN, Pang-Ning, **Introduction to data mining**, 2ª

ed. New York, 2019.

VIEIRA, Sonia, **Fundamentos de estatística**, 6^a ed. São Paulo 2019.

WHO, World Health Organization 2017. **Guidelines for drinking-water quality: fourth edition**, 2017

APÊNDICE A - FORMULÁRIO DE VALIDAÇÃO

05/09/2023, 11:21

Análise de dados sobre a clarificação da água

Análise de dados sobre a clarificação da água

Este formulário apresenta alguns gráficos sobre a clarificação da água e como sua análise pode ser realizada. Essa pesquisa faz parte da dissertação de mestrado em Modelagem Computacional da Universidade Federal do Rio Grande.

Quaisquer dúvidas ou questionamentos, entre em contato com Fabiano Nobre Mendes (fabianonobre@gmail.com).

E-mail *

karperes@gmail.com

Formação acadêmica ou técnica: *

Engenheira Química

Você trabalha com gestão e tratamento de água? *

☒ Sim

☐ Não

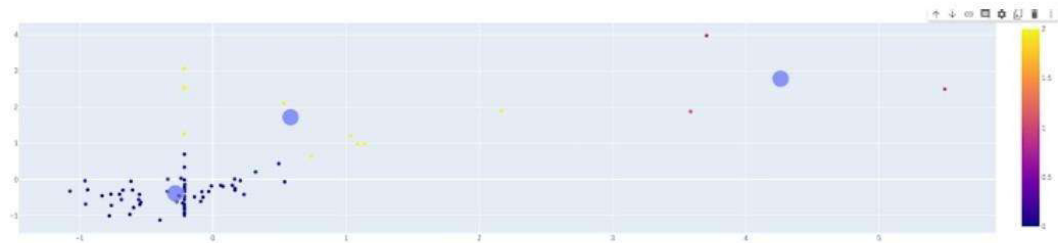
Se você respondeu sim, há quanto tempo você trabalha na área?

Trabalhei 7 anos como técnica e 3 como engenheira química no SANEP, agora leciono na área.

05/09/2023, 11:21

Análise de dados sobre a clarificação da água

1 – Sendo o eixo x turbidez e o y Cor aparente, o que podemos concluir visualizando o gráfico abaixo? *



Que turbidez e cor aparente são diretamente proporcionais.

2 – Sendo a primeira coluna a média da cor aparente da água e a segunda coluna a média da turbidez, onde os três pontos abaixo representam os círculos azuis maiores na figura da questão 1, que podemos concluir analisando estes dados? *

```
array([[ 533.57142857,   86.01428571],
       [1352.86666667,  145.66666667],
       [ 194.83735849,   27.47830189]])
```

Que quando a cor aparente for 533,57, a turbidez será 86,01 e assim, sucessivamente.

05/09/2023, 11:21

Análise de dados sobre a clarificação da água

3 – Utilizando uma técnica de series temporais (SARIMA), foram preditos os valores para os últimos 12 meses. Podemos concluir que são valores compatíveis com a realidade da turbidez da água na barragem Santa Barbara? *

Caso os valores sejam congruentes, como estes poderiam ajudar os gestores da ETA?

```

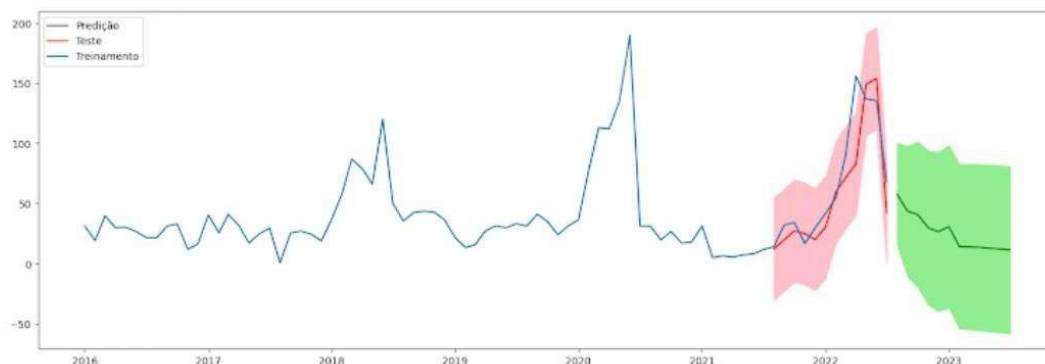
2022-08-01    57.555598
2022-09-01    43.443275
2022-10-01    40.758482
2022-11-01    29.754973
2022-12-01    26.523560
2023-01-01    30.786198
2023-02-01    14.145444
2023-03-01    13.889863
2023-04-01    13.502688
2023-05-01    12.698688
2023-06-01    12.061118
2023-07-01    11.319523
Freq: MS, Name: predicted_mean, dtype: float64

```

São valores possíveis de ocorrer. Essa predição pode auxiliar no planejamento do tratamento, como na aquisição de insumos, por exemplo.

4 – Ao visualizar o gráfico com os valores para o período de 2016 a 2022, com a predição de 12 meses em verde, pode-se concluir qual a tendência da turbidez? E esta seria normal para a época do ano? *

Esta informação seria útil para os administradores da ETA?



Sim, é possível perceber uma tendência. Isso poderia auxiliar no planejamento do tratamento, como citado na questão anterior.

05/09/2023, 11:21

Análise de dados sobre a clarificação da água

Você autoriza que seu nome e suas respostas a esse formulário sejam divulgados como apêndice da dissertação correspondente? *

- ☒ Sim, autorizo.
- ☐ Não, não autorizo.

Este conteúdo não foi criado nem aprovado pelo Google.

Google Formulários

05/09/2023, 11:24

Análise de dados sobre a clarificação da água

Análise de dados sobre a clarificação da água

Este formulário apresenta alguns gráficos sobre a clarificação da água e como sua análise pode ser realizada. Essa pesquisa faz parte da dissertação de mestrado em Modelagem Computacional da Universidade Federal do Rio Grande.

Quaisquer dúvidas ou questionamentos, entre em contato com Fabiano Nobre Mendes (fabianonobre@gmail.com).

E-mail *

dmartinsampaio@gmail.com

Formação acadêmica ou técnica: *

Mestra em Recursos Hídricos

Você trabalha com gestão e tratamento de água? *

☒ Sim

☐ Não

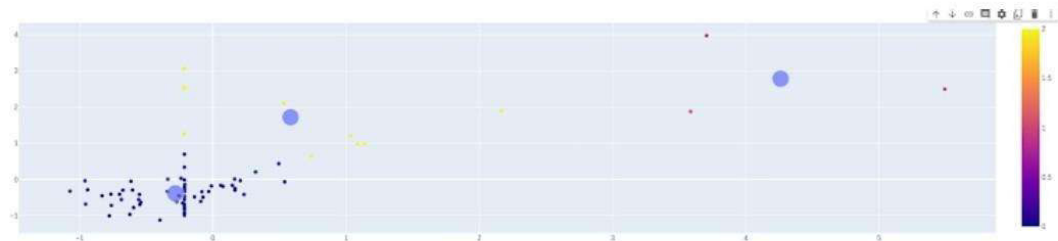
Se você respondeu sim, há quanto tempo você trabalha na área?

6 anos

05/09/2023, 11:24

Análise de dados sobre a clarificação da água

1 – Sendo o eixo x turbidez e o y Cor aparente, o que podemos concluir visualizando o gráfico abaixo? *



Podemos concluir que há uma forte relação entre cor aparente e turbidez.

2 – Sendo a primeira coluna a média da cor aparente da água e a segunda coluna a média da turbidez, onde os três pontos abaixo representam os círculos azuis maiores na figura da questão 1, que podemos concluir analisando estes dados? *

```
array([[ 533.57142857,   86.01428571],
       [1352.86666667,  145.66666667],
       [ 194.83735849,   27.47830189]])
```

Que eles confirmam o visualizado no gráfico anterior, já que percebe-se uma menor cor quando é encontrada uma menor turbidez, e uma maior cor com uma maior turbidez. Essas médias nos dão uma ideia do comportamento de ambos parâmetros para o corpo hídrico em questão.

05/09/2023, 11:24

Análise de dados sobre a clarificação da água

3 – Utilizando uma técnica de series temporais (SARIMA), foram preditos os valores para os últimos 12 meses. Podemos concluir que são valores compatíveis com a realidade da turbidez da água na barragem Santa Barbara? *

Caso os valores sejam congruentes, como estes poderiam ajudar os gestores da ETA?

```

2022-08-01    57.555598
2022-09-01    43.443275
2022-10-01    40.758482
2022-11-01    29.754973
2022-12-01    26.523560
2023-01-01    30.786198
2023-02-01    14.145444
2023-03-01    13.889863
2023-04-01    13.502688
2023-05-01    12.698688
2023-06-01    12.061118
2023-07-01    11.319523
Freq: MS, Name: predicted_mean, dtype: float64

```

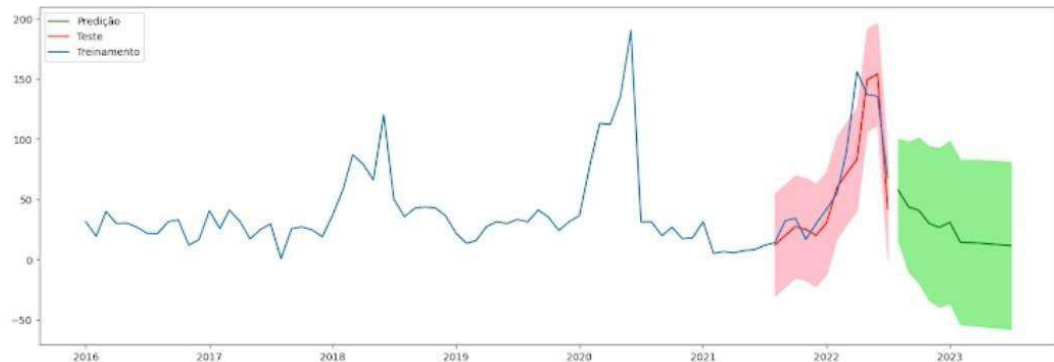
São realmente compatíveis. Essas previsões podem beneficiar os gestores da ETA Santa Barbara ao permitir o planejamento eficaz de operações durante o ano, melhoria do tratamento de água e redução de custos. Isso ajuda no dia-a-dia da operação do tratamento enquanto também viabiliza a avaliação de estratégias, resultando em decisões mais fundamentadas e na manutenção da qualidade da água.

05/09/2023, 11:24

Análise de dados sobre a clarificação da água

4 – Ao visualizar o gráfico com os valores para o período de 2016 a 2022, com a predição de 12 meses em verde, pode-se concluir qual a tendência da turbidez? E esta seria normal para a época do ano? *

Esta informação seria útil para os administradores da ETA?



Percebe-se picos de turbidez mais ou menos na metade do ano, um período chuvoso na nossa região, já que essa chuva acaba causando erosão do solo e ao carreamento de partículas para a água. É uma informação útil para os administradores e gestores da ETA, principalmente por fundamentar várias tomadas de decisão no tratamento de água.

Você autoriza que seu nome e suas respostas a esse formulário sejam divulgados como apêndice da dissertação correspondente? *

☒ Sim, autorizo.

☐ Não, não autorizo.

Este conteúdo não foi criado nem aprovado pelo Google.

Google Formulários

05/09/2023, 11:24

Análise de dados sobre a clarificação da água

Análise de dados sobre a clarificação da água

Este formulário apresenta alguns gráficos sobre a clarificação da água e como sua análise pode ser realizada. Essa pesquisa faz parte da dissertação de mestrado em Modelagem Computacional da Universidade Federal do Rio Grande.

Quaisquer dúvidas ou questionamentos, entre em contato com Fabiano Nobre Mendes (fabianonobre@gmail.com).

E-mail *

levn791ster@gmail.com

Formação acadêmica ou técnica: *

Ciências Biológicas

Você trabalha com gestão e tratamento de água? *

☒ Sim

☐ Não

Se você respondeu sim, há quanto tempo você trabalha na área?

13 anos

05/09/2023, 11:24

Análise de dados sobre a clarificação da água

1 – Sendo o eixo x turbidez e o y Cor aparente, o que podemos concluir visualizando o gráfico abaixo? *



As variáveis são proporcionais.

2 – Sendo a primeira coluna a média da cor aparente da água e a segunda coluna a média da turbidez, onde os três pontos abaixo representam os círculos azuis maiores na figura da questão 1, que podemos concluir analisando estes dados? *

```
array([[ 533.57142857,   86.01428571],
       [1352.86666667,  145.66666667],
       [ 194.83735849,   27.47830189]])
```

Não sei

05/09/2023, 11:24

Análise de dados sobre a clarificação da água

3 – Utilizando uma técnica de series temporais (SARIMA), foram preditos os valores para os últimos 12 meses. Podemos concluir que são valores compatíveis com a realidade da turbidez da água na barragem Santa Barbara? *

Caso os valores sejam congruentes, como estes poderiam ajudar os gestores da ETA?

```

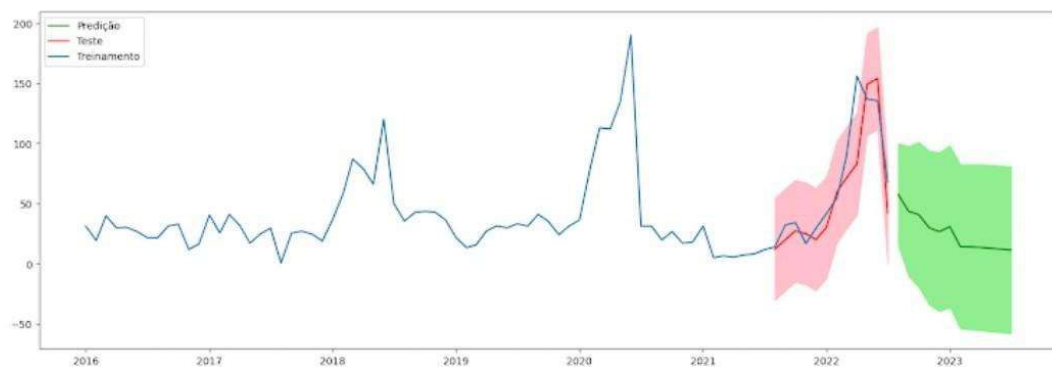
2022-08-01    57.555598
2022-09-01    43.443275
2022-10-01    40.758482
2022-11-01    29.754973
2022-12-01    26.523560
2023-01-01    30.786198
2023-02-01    14.145444
2023-03-01    13.889863
2023-04-01    13.502688
2023-05-01    12.698688
2023-06-01    12.061118
2023-07-01    11.319523
Freq: MS, Name: predicted_mean, dtype: float64

```

Os valores de turbidez estão baixos.

4 – Ao visualizar o gráfico com os valores para o período de 2016 a 2022, com a predição de 12 meses em verde, pode-se concluir qual a tendência da turbidez? E esta seria normal para a época do ano? *

Esta informação seria útil para os administradores da ETA?



Picos de turbidez no meio do ano, pode ser por causa do período de maior precipitação.

05/09/2023, 11:24

Análise de dados sobre a clarificação da água

Você autoriza que seu nome e suas respostas a esse formulário sejam divulgados como apêndice da dissertação correspondente? *

- ☒ Sim, autorizo.
- ☐ Não, não autorizo.

Este conteúdo não foi criado nem aprovado pelo Google.

Google Formulários