UNIVERSIDADE FEDERAL DO PARÁ
NÚCLEO DE DESENVOLVIMENTO AMAZÔNICO EM ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

ANDERSON FRANCISCO DE SOUSA ALMEIDA

IMPLEMENTAÇÃO DE MODELOS COMPUTACIONAIS NA PREDIÇÃO TEMPORAL E ESPAÇO-TEMPORAL DE PARÂMETROS DA QUALIDADE DE ÁGUA

Tucuruí-PA
Dezembro de 2021

UNIVERSIDADE FEDERAL DO PARÁ
NÚCLEO DE DESENVOLVIMENTO AMAZÔNICO EM ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO APLICADA

ANDERSON FRANCISCO DE SOUSA ALMEIDA

IMPLEMENTAÇÃO DE MODELOS COMPUTACIONAIS NA PREDIÇÃO TEMPORAL E ESPAÇO-TEMPORAL DE PARÂMETROS DA QUALIDADE DE ÁGUA

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada do Núcleo de Desenvolvimento Amazônico em Engenharia, da Universidade Federal do Pará, como requisito para a obtenção do título de Mestre em Computação Aplicada.

Orientador: Prof. Dr. Marcos Amaris
Coorientador: Prof. Dr. Bruno Merlin

Tucuruí-PA
Dezembro de 2021

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

A447i Almeida, Anderson.
Implementação de modelos computacionais na predição temporal e espaço-temporal de parâmetros da qualidade de água / Anderson Almeida. — 2021.
63 f. : il. color.

Orientador(a): Prof. Dr. Marcos Gonzalez
Coorientador(a): Prof. Dr. Bruno Merlin
Dissertação (Mestrado) - Universidade Federal do Pará, Núcleo de Desenvolvimento Amazônico em Engenharia, Programa de Pós-Graduação em Computação Aplicada, Tucuruí, 2021.

1. Qualidade da água. 2. Aprendizado de máquina. 3. Regressão. 4. Redes Neurais Artificiais. I. Título.

CDD 006.3

ANDERSON FRANCISCO DE SOUSA ALMEIDA

IMPLEMENTAÇÃO DE MODELOS COMPUTACIONAIS NA PREDIÇÃO TEMPORAL E ESPAÇO-TEMPORAL DE PARÂMETROS DA QUALIDADE DE ÁGUA

Dissertação apresentada ao Programa de Pós-Graduação em Computação Aplicada do Núcleo de Desenvolvimento Amazônico em Engenharia, da Universidade Federal do Pará, como requisito para a obtenção do título de Mestre em Computação Aplicada.

Data da Defesa: 14 de dezembro de 2021.
Conceito: Aprovado

Banca Examinadora

Prof. Dr. Marcos Tulio Amaris Gonzalez
PPCA/NDAE/UFPA
Orientador

Prof. Dr. Bruno Merlin
PPCA/NDAE/UFPA
Coorientador

Prof. Dr. Adam Dreyton Ferreira dos Santos
PPCA/UFPA
Membro interno

Prof. Dr. Pedro Moreira de Sousa Júnior
UFRA
Membro externo

Quando tens verdadeira disposição para vencer, atinges altos pontos, superas a ti mesmo, realizas grandes feitos e imprimes a tua marca onde atuas.

Quem se dispõe a vencer atrai a vitória.

Lourival Lopes

AGRADECIMENTOS

À Deus, pelo fornecimento de fé e força, que resultou na superação das dificuldades ocorridas durante a trajetória acadêmica.

À minha mãe, Antônia de Fátima, por sempre acreditar que a educação é o melhor caminho para o sucesso, por isso, és a maior incentivadora na minha vida.

À minha esposa, Jamilla Almeida, pelo apoio e compreensão nos momentos que mais necessitei de concentração.

Ao meu orientador, Marcos Amariz, pela confiança, por compartilhar os seus conhecimentos comigo e a sua dedicação nas orientações durante esta trajetória acadêmica.

Ao coorientador, Bruno Merlin, pelos ensinamentos durante as orientações.

Aos docentes do Programa de Pós-Graduação de Computação Aplicada pelo compartilhamento dos seus conhecimentos nas aulas.

Ao Diretor do Campus Capanema da Universidade Federal Rural da Amazônia, Ebson Cândido, e ao colega de profissão, Saulo Silva, e aos demais colegas da instituição pelo apoio e compreensão durante esta trajetória acadêmica.

Ao Prof. Pedro Moreira da Universidade Federal Rural da Amazônia - Campus Capanema, pelo incentivo desde o processo seletivo do mestrado.

À amiga de turma, Patrícia Diniz, pela parceria nas trocas de ideias e soluções durante as atividades acadêmicas.

RESUMO

A qualidade da água está diretamente relacionada com o seu nível de poluição causada pelas ações antrópicas e industriais, destacando-se como consequência a redução da disponibilidade de uma água de qualidade. Por isso, são realizados os monitoramentos limnológicos dos parâmetros básicos da qualidade da água como forma de obtenção de dados que norteiam as tomadas de decisão dos órgãos gestores de recursos hídricos. Neste contexto, o presente estudo tem a implementação de algoritmos de aprendizado de máquina para predizer de modo temporal e espaço-temporal os dados dos parâmetros da qualidade da água. As técnicas de aprendizado de máquina usadas foram regressão linear, random forest, redes neurais MLP e LSTM. Foram usados dois pontos de coleta de uma Unidade de Gerenciamento de Recursos Hidricos em São Paulo, Brasil. Os modelos são avaliados através das métricas MAPE (Erro percentual médio absoluto) e RMSE (Erro de raiz quadrada média). Portanto, na predição temporal a técnica LSTM apresentou o melhor desempenho em relação as demais técnicas, pois tem o menor resultado de MAPE médio, com 4.66%. Enquanto que o MLP apresentou o melhor desempenho em relação aos dados utilizados, pois tem o menor resultado de RMSE médio, com 2.47. Porém, na predição espaço-temporal, o MLP têm os melhores desempenho tanto em relação as demais técnicas quanto aos dados utilizados, pois têm os menores resultados médios de MAPE e RMSE, respectivamente, 5.94% e 1.34. Desse modo, estes desempenhos das redes neurais podem ser justificados pela não linearidade dos dados dos parâmetros. Além disso, os resultados dos experimentos visam contribuir com o processo de monitoramento da qualidade da água e auxiliar o planejamento da gestão hídrica de modo que atenda as legislações vigentes e possibilite a indicação de políticas públicas através de modelos de aprendizado de máquina na predição dos parâmetros de qualidade de água.

Palavras–chave: Qualidade da água; Aprendizado de máquina; Regressão; Redes Neurais Artificiais.

ABSTRACT

The quality of water is directly related to its level of pollution caused by anthropic and industrial actions, with a consequent reduction in the availability of quality water. Therefore, limnological monitoring of the basic parameters of water quality is carried out, as a way of obtaining data that guide the decision-making of water resources management bodies. In this context, the present study has the implementation of machine learning algorithms to predict temporal and spatiotemporal water quality parameter data. The ML techniques used were linear regression, random forest, MLP and LSTM neural networks. Two collection points from a Water Resources Management Unit in São Paulo, Brazil were used. Models are evaluated using MAPE (mean absolute percentage error) and RMSE (root mean squared erro) metrics. Therefore, in temporal prediction, the LSTM technique presented the best performance in relation to the other techniques, as it has the lowest average MAPE result, with 4.66%. While MLP presented the best performance in relation to the data used, as it has the lowest average RMSE result, with 2.47. However, in spatiotemporal prediction, MLP has the best performance both in relation to the other techniques and the data used, as it has the lowest average results of MAPE and RMSE, respectively, 5.94% and 1.34. Thus, these performances of neural networks can be justified by the non-linearity of the parameter data. Other than that, the results of the experiments aim to contribute to the water quality monitoring process and assist in the planning of water management, so that it meets current legislation and enables the indication of public policies, through machine learning models in prediction of water quality parameters.

Keywords: Water quality, Machine learning, Regression, Artificial Neural Networks

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ANA	Agência Nacional das Águas e Saneamento Básico
CETESB	Companhia Ambiental do Estado de São Paulo
CONAMA	Conselho Nacional do Meio Ambiente
COSANPA	Companhia de Saneamento do Pará
CRH	Conselho Estadual de Recursos Hídricos
CSELM	*Composite Semi-supervised Extreme Learning Machine*
DBO	Demanda Bioquímica de Oxigênio
DHP	*Direct Hashing and Pruning*
DI	*Discharge*
EEMD-ELM	*Ensemble Empirical Mode Decomposition - Extreme Learning Machine*
EMAX	*Maximun Error*
EMD	*Empirical Mode Decomposition*
FIQ	*Faixa interquartil*
GSP	*Generalized Sequential Pattern*
IA	Inteligência Artificial
KDD	*Knowledge Discovery in Databases*
LAI	*Lei de Acesso à Informação*
LSSVM	*Least Square Support Vector Machine*
LSTM	*Long-Short Term Memory*
M5T	*M5 Model Tree*
MAE	*Mean Absolute Error*
MAPE	*Mean Absolute Percentage Error*
MARS	*Multivariate Adaptive Regression Splines*
MLP	*Perceptron Multilayer*
NSC	*Nash Sutcliffe efficiency Coefficient*

OD	Oxigênio Dissolvido
PH	Potencial Hidrogeniônico
PIB	Produto Interno Bruto
PLS-ELM	*Partial Least Square - Extreme Learning Machine*
PNQA	Programa Nacional de Avaliação de Qualidade das Águas
RF	*Random Forest*
RL	*Regressão Linear*
RMSE	*Root Mean Square Error*
SC	*Specific Conductance*
SD	*Sensor Depth*
SEMA	Secretaria Estadual de Meio Ambiente
SEMAS	Secretaria de Meio Ambiente e Sustentabilidade
SESPA	Secretaria Estadual de Saúde do Pará
ST	Series Temporais
UFRA	Universidade Federal Rural da Amazônia
UGRHI	Unidade de Gerenciamento de Recursos Hídricos
WD-LSSVM	Wavelet Denoising - Least Square Support Vector Machine

SUMÁRIO

1 INTRODUÇÃO

A oferta de água é determinada pela dinâmica hídrica e socioeconômica das bacias dos rios de água doce, além das suas condições de qualidade. E o conhecimento dessa oferta é produzido pelo monitoramento limnológico que coleta periodicamente dados dos parâmetros básicos da qualidade da água em rios e lagos, permitindo o planejamento e a gestão dos recursos hídricos (MAROTTA; SANTOS; ENRICH-PRAST, 2008).

Desse modo, esses dados históricos dos parâmetros da qualidade da água são transformados em série temporal ordenada e sequenciada com objetivo de fazer uma análise preditiva para sintetizar e caracterizar o comportamento da série e finalmente aplicar um modelo adequado a este comportamento, encontrando possíveis valores futuros Parmezan (2014). Ainda Box et al. (2015) cita que esse tipo de análise fornece informações qualificadas para o planejamento econômico, dos negócios, da produção, do controle e otimização de processos.

Diate disso, a tecnologia da informação contribui para a qualidade da tomada de decisão usando análise preditiva sobre uma série temporal e área da computação, que trata amplamente de análises preditivas é o aprendizado de máquina, que é uma sub-área da inteligência artificial aplicada para adquirir conhecimento de forma automática (MONARD; BARANAUSKAS, 2003). No caso das series temporais dos parâmetros da qualidade da água, o algoritmo de aprendizado de máquina otimiza a análise dos resultados dos monitoramentos.

Com o propósito de auxiliar no planejamento e na gestão dos recursos hídricos quanto à classificação dos corpos hídricos de acordo com a resolução Nº 357/2005 - CONAMA e Lei Nº 9433/1997, além de possibilitar a indicação de políticas públicas, este estudo se propõe a desenvolver modelos usando técnicas de aprendizado de máquina para predizer os valores dos parâmetros da qualidade da água e comparar os seus desempenhos das técnicas nas predições. Para isso, neste estudo serão utilizados os resultados dos monitoramentos nas Unidades de Gerenciamento de Recursos Hidrícos (UGRHI) do Estado de São Paulo, sendo disponibilizados pela Companhia Ambiental do Estado de São Paulo (CETESB).

1.1 Contexto do problema de pesquisa

A contextualização do problema de pesquisa está relacionada quanto à importância, à disponibilidade e ao consumo da água. Assim, a água é um importante recurso natural para todos os seres vivos e está presente nas atividades econômicas industriais e agrícolas, no abastecimento urbano, etc. Por isso, a sua disponibilidade requer atenção, pois segundo Narciso e Jaskiu (2019) apenas 0,5% da água do planeta está disponível para o consumo e no Brasil é previsto o aumento do consumo da água em 24% até 2030 (ANA, 2019b).

Dessa forma, o desenvolvimento econômico e a urbanização do país contribuem diretamente para o crescimento do consumo de água, destacando a irrigação, abastecimento urbano e a

indústria como atividades que mais consomem água no Brasil, conforme a Figura 1.

Figura 1 – Média anual de 2018 da água consumida no Brasil.

Fonte: Adaptado de ANA (2019a).

Para preservar esse recurso natural e atender a demanda prevista é necessário realizar monitoramentos dos corpos hídricos e para tal ação há mais de 4 mil pontos de monitoramento distribuídos pelo país, em parceria com órgãos públicos estaduais (ANA, 2019a), entre os quais 471 pontos pertencem à CETESB (SEMA, 2018). Diante disso, neste estudo são utilizados os dados brutos dos parâmetros da qualidade da água das UGRHI. Desse modo, surge o problema de pesquisa que é a falta de otimização analítica dos dados dos parâmetros da qualidade da água, que solucionada pode contribuir com o planejamento, gestão e classificação dos recursos hídricos de acordo com a resolução Nº 357/2005 - CONAMA e atender um dos objetivos da Lei Nº 9433/1997, a qual visa assegurar a disponibilidade de uma água de qualidade para a atual e futuras gerações.

1.2 Justificativa

Segundo Braga, Porto e Tucci (2015), o planejamento e a gestão dos recursos hídricos dependem de informações confiáveis para garantir qualidade na tomada das decisões e consequentemente preservar a sustentabilidade do ecossistema. Ressalta-se que a falta de informação produz um custo maior do que adquirir informação por meios tecnológicos.

Braga, Porto e Tucci (2015), citam

> A informação sobre a qualidade da água é necessária para que se conheça a situação dos corpos hídricos com relação aos impactos antrópicos na bacia hidrográfica e é essencial para que se planeje sua ocupação e seja exercido o necessário controle dos impactos.

A SEMA (2018) afirma que o volume de dados coletado por rede de monitoramento contribui na condução de projetos relacionados ao saneamento no Estado de São Paulo, ao atendimento das legislações vigentes como a Resolução CONAMA 357/2005 e ao objetivo do Programa Nacional de Avaliação de Qualidade das Águas (PNQA), que é ampliar o conhecimento sobre a qualidade das águas superficiais no Brasil,

A qualidade do recurso hídrico é um fator importante na avaliação do desenvolvimento sustentável de uma região (CORDOBA; MARTINEZ; FERRER, 2010); por isso, está relacionada com as ações das atividades econômicas, como a agropecuária e a industrial e suas consequências na saúde pública.

Diante disso, na agropecuária, apesar da sua importância econômica para o Brasil, Telles e Domingues (2015) afirmam que "diferentes impactos ambientais são associados à utilização da água em sistemas de produção agrícola e pecuário", entre eles a deterioração da qualidade dos recursos hídricos, devido à aplicação de agrotóxicos nas culturas e escoamento da água da chuva em áreas de pastagens. Na indústria, Silva e Kulay (2015) afirmam que o setor é responsável por 22% de uso no mundo e o principal lançador de efluentes nos recursos hídricos, diminuindo a qualidade e oferta de água. Na saúde pública, Branco et al. (2015) afirma que no Brasil há um índice de 2,3% de óbitos relacionados a doenças de veiculação hídrica.

Assim, o produto da pesquisa visa contribuir com o processo de monitoramento da qualidade das águas, avaliando as suas tendências e diagnósticos, tornando-se um subsídio para o planejamento da gestão hídrica em manter o equilíbrio entre o desenvolvimento econômico, demográfico e a disponibilidade hídrica para diversos tipos de uso, beneficiando áreas como a saúde pública, pois com o aumento da qualidade da água há redução de ocorrências de doenças veiculadas a esta.

Portanto, considerando a importância de uma informação futura sobre os parâmetros da qualidade da água para toda a sociedade e ecossistemas que utilizam a água para várias finalidades, a pesquisa pretende responder a seguinte questão: quais algoritmos de aprendizado de máquina são adequados para predizer os parâmetros da qualidade da água nas UGRHI?

1.3 Objetivos de pesquisa

1.3.1 Objetivo geral

Analisar a aplicabilidade do aprendizado de máquina para a predição temporal e espaço-temporal de parâmetros de qualidade da água.

1.3.2 Objetivos específicos

Assim, especificamente, a pesquisa pretende:

- Obter bancos de dados, de livre acesso, afim de extrair a série histórica de dados de parâmetros de qualidade da água.
- Utilizar métodos estatísticos para a análise dos dados coletados e para a padronização e normalização do conjunto amostral.
- Analisar o desempenho dos algoritmos utilizados na predição temporal e espaço-temporal dos parâmetros da qualidade da água.

1.4 Organização do Trabalho

O restante do trabalho está divido assim:

- Capitulo 2 - Fundamentação Teórica: Neste capitulo são apresentados os conceitos de alguns parâmetros da qualidade da água, das series temporais, do processo de mineração de dados, aprendizado de máquina, métricas de desempenho e trabalhos relacionados.
- Capitulo 3 – Metodologia: Neste capitulo são apresentados os procedimentos metodológicos da coleta de dados, análise exploratória dos dados, pré-processamento dos dados e configuração dos experimentos de predição.
- Capitulo 4 – Resultados: Neste capitulo são apresentados os resultados da análise exploratória dos dados e dos experimentos de predição.
- Capitulo 5 – Conclusão: São enunciadas as considerações finais do trabalho, as limitações da pesquisa, sugestões de trabalhos futuros e as produções cientificas.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão apresentados os conceitos básicos utilizados nesta dissertação referentes aos parâmetros da qualidade da água, series temporais, mineração de dados, aprendizado de máquina e seus algoritmos para predição. Por fim, são descritos os trabalhos relacionados com o tema da pesquisa.

2.1 Parâmetros da Qualidade da Água

A qualidade da água é formada pelas características de natureza física, química e biológica representadas por diferentes parâmetros que indicam a sua qualidade, de acordo com o valor padrão estabelecido em lei para cada finalidade de uso (DERISIO, 2016). Ademais, os corpos de água podem ser potáveis, cujo padrão de potabilidade para o consumo humano é regulamentado pela portaria Nº 2.914 do Ministério da Saúde de 2011, salinas ou doces superficiais regulamentada pela resolução nº 357/2005 e suas alterações com a publicação das resoluções nº 410/2009 e 430/2011 do CONAMA que trata sobre a classificação destes tipos de corpos de água em função do uso. As águas subterrâneas são regulamentadas pela resolução nº 396/2008 do CONAMA que dispõe sobre a classificação deste tipo de corpo de água em função do padrão de qualidade.

Diante disso, a coleta dos dados sobre os parâmetros da qualidade da água torna-se de suma importância para as ações de planejamento e gestão dos recursos hídricos, pois mantém atualizada a situação das condições dos corpos de água, além de subsidiar, a descoberta de conhecimento futuro sobre os parâmetros da qualidade das águas superficiais através da análise dos dados coletados durante uma regularidade temporal (GASTALDINI et al., 2001). A seguir, serão descritos os parâmetros utilizados no presente estudo.

2.1.1 Oxigênio Dissolvido (OD)

O Oxigênio Dissolvido é vital para o metabolismo dos seres que vivem no corpo d'água (SOLANKI; AGRAWAL; KHARE, 2015). Portanto, a concentração de OD é o mais importante parâmetro de qualidade da água e adotado como principal indicador de poluição no rio (MOHAN; KUMAR, 2016), tornando-se indispensável para o equilíbrio da vida aquática. A Tabela 1 apresenta os valores de oxigênio dissolvido e a sua classificação quanto as condições da qualidade das águas doces segundo a resolução Nº357/2005 - CONAMA.

Tabela 1 – Classificação das condições da qualidade das águas doces de acordo com os valores de Oxigênio Dissolvido.

Classes	Valor de OD (mg/L)
Classe 1	> 6
Classe 2	>5 e < 6
Classe 3	>4 e < 5
Classe 4	>2 e < 4

2.1.2 Potencial Hidrogeniônico (pH)

É um dos parâmetros mais utilizados na análise da qualidade da água, determinando se a água é ácida, cujos valores influenciam diretamente na fisiologia de várias espécies. O valor de pH varia na faixa entre 0 a 14 e dependendo do valor é classificada em ácida, neutra e alcalina (ANA, 2019), conforme a Tabela 2. A variação do pH pode ser influenciada por efluentes adicionados por indústrias ou atividades humanas instaladas ao redor do corpo d'água (SOLANKI; AGRAWAL; KHARE, 2015).

Tabela 2 – Classificação do pH.

Classificação do pH	Valor de pH
Ácida	< 7
Neutra	$= 7$
Alcalina	> 7

2.1.3 Turbidez

A turbidez é uma característica física que resulta na quantidade de luz refletida mediante a presença de material em suspensão nos corpos d'água (FRAVET; CRUZ, 2007). Esses materiais provocam a dispersão e absorção da luz, deixando a condição física da água turva (ANA, 2019). Portanto, o nível de turbidez de um corpo d'água é determinante para a sua condição e produtividade. A Figura 2 apresenta diferentes amostras de turbidez da água.

Figura 2 – Exemplo de amostras de turbidez da água.

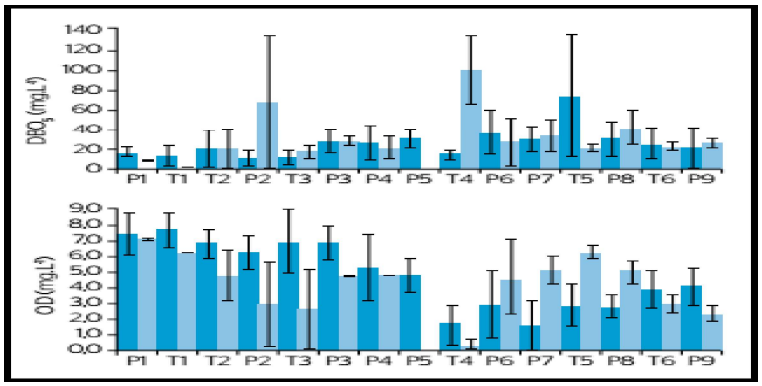
Fonte: Adaptado de Water (2019).

2.1.4 Demanda Bioquímica de Oxigênio (DBO)

O DBO é um parâmetro responsável pela identificação da quantidade de matéria orgânica biodegradável no corpo d'água (GASTALDINI; TEIXEIRA, 2001), portanto o aumento do nível de DBO indica a insuficiência de oxigênio e, consequentemente, a extinção da vida aquática,

assim como quando o nível de DBO está baixo indica aumento do nível de oxigênio. A Figura 3 apresenta a relação inversa entre os níveis de DBO e OD.

Figura 3 – Relação do DBO com OD no corpo de água.



Fonte: Adaptado de Menezes et al. (2016).

2.1.5 Fósforo

O Fósforo está presente no processo de produção da eutrofização, indicando se o recurso hídrico está poluído por fosfatos (ANA, 2019). A sua produção é por fontes naturais como rochas, partículas presentes na atmosfera ou matéria orgânica em decomposição, também por fontes artificiais como esgoto doméstico ou industrial e partículas de origem industrial na atmosfera. Os valores do parâmetro fósforo combinado com outros parâmetros são importantes para determinar a qualidade da água quanto à presença de nutrientes. A Figura 4 apresenta um recurso hídrico em processo de eutrofização.

Figura 4 – Processo de eutrofização com fósforo.

Fonte: Adaptado de Souza (2019).

2.1.6 Temperatura

É um parâmetro que influencia diretamente as reações físico-químicas dos recursos hídricos e o metabolismo dos organismos aquáticos. Além disso, a variação da temperatura ocorre devido aos fatores climáticos e a profundidade do corpo d'água (SIMONETTI, 2018).

2.1.7 Sólido

É um parâmetro que com elevados valores impactam a disponibilidade de alimentos para os organismos aquáticos, além de proporcionar a decomposição anaeróbica nos corpos d'água (SIMONETTI, 2018).

2.1.8 Coliformes fecais

É um parâmetro que indica a existência de micro-organismos patogênicos que causam doenças de veiculação hídrica, tais como a febre tifóide e cólera (ANA, 2019).

2.2 Séries Temporais

As Séries Temporais (ST) são caracterizadas como um conjunto de observações coletadas de forma ordenada e sequencial ao longo do tempo (CHATFIELD, 2013). A representação matemática de uma serie temporal é um vetor Z, de ordem $nx1$, onde n é o número de observações. Também pode ser representada por dois vetores, um vetor (Z_n) armazena as observações e o outro vetor (T_n) armazena o tempo em que a observações foram coletadas (OLIVEIRA et al., 2014), conforme a Equação 2.1.

$$Z = \begin{bmatrix} Z_1 & Z_2 & Z_3, \cdots, & Z_n \\ T_1 & T_2 & T_3, \cdots, & T_n \end{bmatrix} \tag{2.1}$$

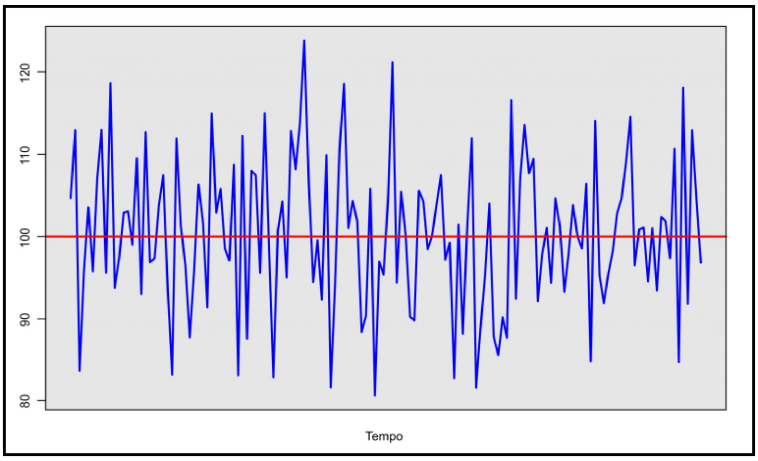
Segundo Morettin e Toloi (2018), a análise de uma série temporal tem os objetivos: investigar o seu mecanismo gerador, prever valores futuros, descrever o comportamento da série e descobrir períodos relevantes nos dados. Alguns exemplos de séries temporais são: os valores diários de poluição em uma cidade, índices diários da Bolsa de Valores de São Paulo, acidentes ocorridos nas rodovias da cidade durante um mês e outros (DAMETTO, 2018).

2.2.1 Características de uma série temporal

- **Contínua ou discreta:** Nas ST contínuas as observações dos dados são realizadas de forma contínuas sobre um intervalo de tempo específico, enquanto que na ST discreta as observações são convertidas em intervalos fixos de tempo (BROCKWELL; DAVIS, 2002).

- **Estacionaria:** Quando a ST se desenvolve ao redor de uma média constante, representando estabilidade nos dados da ST (MORETTIN; TOLOI, 2006). Na Figura 5, a linha vermelha representa a estacionariedade na ST.

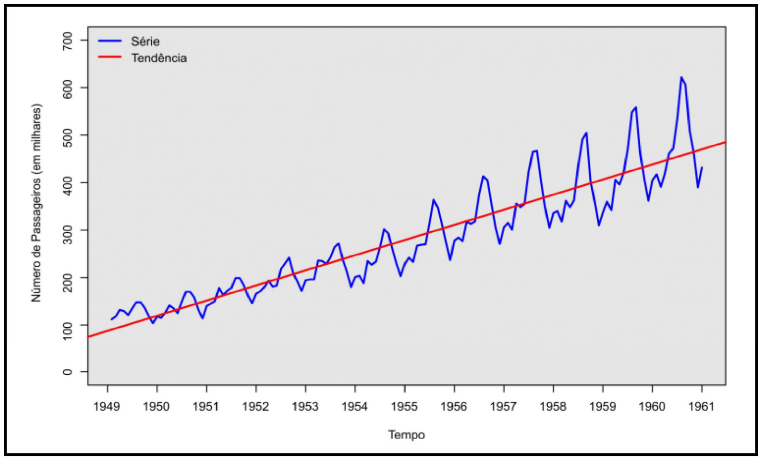
Figura 5 – Exemplo de serie temporal não estacionária.



Fonte: Adaptado de Oliveira (2019).

- **Tendência:** É representada por um movimento regular, lento e extenso ao longo da ST, podendo ser tanto crescente quanto decrescente (EHLERS, 2012). São exemplos de tendência crescente os fenômenos de desenvolvimento demográfico e hábitos de consumo, enquanto que as taxas de mortalidade e desemprego são exemplos de tendência decrescente (PARMEZAN, 2014). A Figura 6 apresenta graficamente um exemplo de ST com tendência crescente.

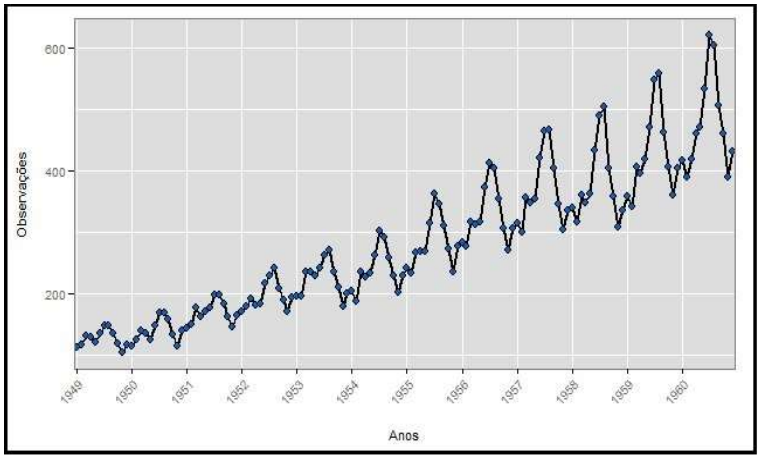
Figura 6 – Exemplo de serie temporal com tendência.



Fonte: Adaptado de Oliveira (2019).

- **Sazonalidade:** Repetição de um comportamento em diferentes períodos de tempo na ST. Como exemplo de sazonalidade é o aumento das vendas de ar condicionado no verão e agasalhos no inverno (PARMEZAN, 2014). A Figura 7 apresenta graficamente um exemplo de ST com sazonalidade.

Figura 7 – Exemplo de serie temporal com sazonalidade.



Fonte: Adaptado de Action (2019).

2.3 Processo de Descoberta de Conhecimento

Silva, Peres e Boscarioli (2017), citam

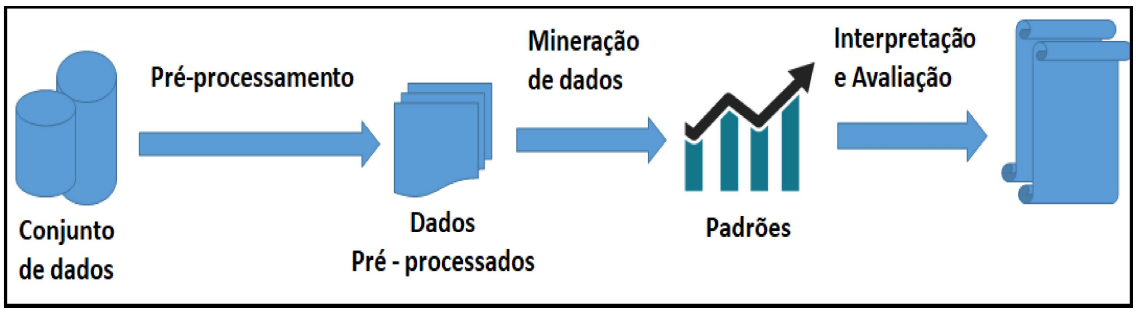
> A descoberta de conhecimento em bases de dados tem como objetivo encontrar padrões intrísecos aos dados nela contidos, apresentando-os de forma a facilitar sua assimilação como conhecimento [...] tal descoberta está associada a um processo analítico, sistemático e, até onde possível, automatizado [...] denomina-se processo de KDD.

Segundo Fayyad, Piatetsky-Shapiro e Smyth (1996), o processo de Descoberta de Conhecimento em Banco de Dados - KDD possui "[...] várias etapas, não trivial, interativo, iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis [...]"e define essas etapas em: Seleção, Pré-processamento, Transformação, Mineração de Dados e Interpretação/Validação. Outros autores como Goldschmidt e Passos (2005), defendem a simplificação do processo de KDD nas seguintes etapas: Pré-processamento, Mineração de Dados e Pós-processamento. A Figura 8 apresenta as etapas do processo KDD defendidas por Goldschmidt e Passos (2005).

2.3.1 Pré-processamento de Dados

Segundo Goldschmidt e Passos (2005), a etapa de pré-processamento de dados tem o objetivo de preparar os dados para a etapa de mineração de dados, executando as seguintes

Figura 8 – Etapas do processo KDD.



Fonte: *Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).*

procedimentos: seleção de dados, limpeza de dados, codificação dos dados e enriquecimento dos dados. Dessa forma, busca-se melhorar a qualidade dos dados que serão processados e consequentemente obter um resultado de qualidade (CORTES; PORCARO; LIFSCHITZ, 2002). A seguir, são descritos os procedimentos do pré-processamento de dados:

- **Seleção de dados:** Tem objetivo de identificar as informações que serão utilizadas durante o processo de KDD (GOLDSCHMIDT; PASSOS, 2005), utilizando técnicas de estatística descritiva como aplicação de medidas de tendência central (média, mediana e moda), dispersão (amplitude, variância, desvio-padrão e coeficiente de dispersão) e análise correlação, com visualização gráfica dos resultados (SILVA; PERES; BOSCARIOLI, 2017).
- **Limpeza de dados:** Tem objetivo de tratar os dados selecionados que possuam valores ausentes, errôneos (fora de padrão) ou inconsistentes, preservando a completude, integridade e veracidade dos dados (GOLDSCHMIDT; PASSOS, 2005), conforme a Tabela 3.

Tabela 3 – Métodos aplicáveis na limpeza de dados.

Métodos para dados ausentes	**Descrição**
Remoção do registro	Remove o registro em que ocorre a ausência de valor.
Preenchimento manual	Abordagem trabalhosa, de difícil aplicabilidade em grandes bancos de dados.
Preenchimento automático	Com um valor constante ou valor médio ou aplicação de análise preditiva considerando os valores presentes do atributo.
Métodos para dados fora de padrão	**Descrição**
Correção manual	Aplicada quando há poucos registros.
Correção automática	Utiliza técnicas de agrupamento para identificar valores fora do padrão e a regressão para encontrar o melhor ajuste entre as variáveis.

Fonte: Adaptado de Silva, Peres e Boscarioli (2017).

- **Transformação:** Tem objetivo de transformar os dados para um formato adequado para a etapa de mineração de dados (GOLDSCHMIDT; PASSOS, 2005). Entre as técnicas

existentes para a normalização de dados, há o *Min-Max*, representado pela Equação 3.3, que organiza os valores do conjunto de dados em uma mesma escala, que retornando resultados positivos o intervalo de valores é entre 0 e 1 ou quando são negativos o intervalo de valores é entre -1 e 1 (SILVA; PERES; BOSCARIOLI, 2017).

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{2.2}$$

onde, x é a variável a normalizar e x' é o resultado da normalização *min-max*.

2.3.2 Mineração de Dados

Segundo Cortes, Porcaro e Lifschitz (2002), a etapa de mineração de dados corresponde na identificação dos objetivos (tarefas) da mineração e da técnica mais adequada a ser aplicada na extração de padrões a partir dos dados.

De acordo com Goldschmidt e Passos (2005), as principais tarefas da mineração de dados são:

- **Associação:** Tem o objetivo de encontrar padrões associados ou correlacionados entre os atributos de um conjunto de dados (SILVA; PERES; BOSCARIOLI, 2017). Os algoritmos Apriori, GSP e DHP são exemplos de técnicas que implementam a associação (GOLDSCHMIDT; PASSOS, 2005).

- **Classificação:** Mapeia um conjunto de dados em um conjunto de rótulos categóricos (classe) predefinidos e quando aplicado em novos registros possam prever a sua classe. Redes neurais, algoritmos genéticos e lógica indutiva são exemplos de técnicas que implementam a classificação (GOLDSCHMIDT; PASSOS, 2005).

- **Agrupamento:** Segmenta um conjunto de dados em um número de subgrupos homogêneos, de acordo com as suas similaridades (CORTES; PORCARO; LIFSCHITZ, 2002). O K-Means e K-Modes, entre outros, são exemplos de técnicas que implementam o agrupamento (GOLDSCHMIDT; PASSOS, 2005).

- **Regressão:** Mapeia um conjunto de dados x com objetivo de prever uma variável y do tipo numérico (GOLDSCHMIDT; PASSOS, 2005), conforme a equação 2.3, que pode ser linear ou não linear. A equação da reta exemplifica a regressão linear, enquanto que a equação exponencial exemplifica a regressão não linear. Ademais, a regressão é caracterizada como simples, quando há apenas um atributo para estimar y, ou multivariada, quando há mais de um atributo para estimar y. A regressão é aplicada para resolver problemas de indicadores econômicos, séries temporais e etc (SILVA; PERES; BOSCARIOLI, 2017), utilizando

técnicas como Regressão linear, Redes neurais, Máquina de suporte vetorial e árvore de decisão (GOLDSCHMIDT; PASSOS, 2005).

$$y = f(x) \tag{2.3}$$

2.3.3 Interpretação e Avaliação

Segundo Cortes, Porcaro e Lifschitz (2002), a avaliação tem o objetivo de identificar quais padrões podem ser utilizados baseados na expressividade estatística e a interpretação tem o objetivo de tornar os padrões identificados como uma ferramenta de apoio ao processo de tomada de decisão (MACEDO; MATOS, 2010).

Segundo Campos (2008), a avaliação é realizada baseada nas métricas de erro que servem para quantificar o nível de eficiência dos modelos utilizados na predição de séries temporais. A seguir, as métricas mais utilizadas nos trabalhos relacionados:

- **Erro Absoluto Médio** (*Mean Absolute Error*)

$$MAE = \frac{1}{N}\sum_{k=1}^{N}|y - y'| \tag{2.4}$$

- **Erro Percentual Absoluto Médio** (*Mean Absolute Percentage Error*)

$$MAPE = \frac{1}{N}\sum_{k=1}^{N}\left|\frac{y - y'}{y}x100\right| \tag{2.5}$$

- **Raiz do Erro Quadrático Médio** (*Root Mean Square Error*)

$$RMSE = \frac{\sqrt{\sum_{k=1}^{N}(y - y')^2}}{N} \tag{2.6}$$

onde:

- y: representa a série temporal.
- y': representa a predição da série temporal.

- **Coeficiente de determinação** (R^2)

$$R^2 = \frac{y'X'Xy}{y'X'Xy + e'e} \tag{2.7}$$

onde:

- y: representa o vetor de parâmetros. estimados.
- X': representa a matriz de delineamento.
- e': representa o vetor de erros associado.

As métricas 2.4 e 2.5 servem para avaliar o desempenho entre os modelos de predição, a métrica 2.6 serve para avaliar o desempenho do modelo em relação aos dados utilizados (CAMPOS, 2008) e a métrica 2.7 avalia a qualidade de ajuste do modelo (CRAMER, 1987), na qual retorna valores entre 0 e 1, logo, quanto mais próximo de 1, mais ajustado é o modelo (PALA, 2019).

2.4 Aprendizado de Máquina

Segundo Russel e Norvig (2013), a Inteligência Artificial é um estudo "de agentes inteligentes capazes de perceber seu meio ambiente e de realizar ações com a expectativa de selecionar uma ação que maximize o seu desempenho".

Haykin (2007), cita

> [...] O objetivo da Inteligência Artificial (IA) é o desenvolvimento de paradigmas ou algoritmos que requeiram máquinas para realizar tarefas cognitivas, para os quais os humanos são atualmente melhores ... e deve ser capaz de fazer três coisas: (1) armazenar conhecimento, (2) aplicar o conhecimento armazenado para resolver problemas e (3) adquirir novo conhecimento através da experiência [...]

Portanto, o objetivo de armazenar conhecimento (HAYKIN, 2007) refere-se ao aprendizado de máquina que é uma subárea da inteligência artificial que utiliza técnicas computacionais para adquirir conhecimento de forma automática (MONARD; BARANAUSKAS, 2003), baseado em um conjunto de dados.

Russel e Norvig (2013) definem as formas de treinamento de aprendizado de máquina, que são:

- **Aprendizagem não supervisionada:** O agente aprende padrões sem a observação de exemplos, utilizada na tarefa de agrupamento (CAMILO; SILVA, 2009).
- **Aprendizagem por reforço:** O agente aprende a partir de uma série de reforços representados por recompensas ou punições.
- **Aprendizagem semi-supervisionada:** O agente aprende padrões com pouca observação de exemplos de um grande conjunto de dados.

- **Aprendizagem supervisionada:** Nesta aprendizagem há uma interferência indutiva, em que há um "professor"para direcionar cada saída correta relacionada a cada observação do subconjunto de treinamento (SILVA; SPATTI; FLAUZINO, 2010), ajustada pelos algoritmos que executam as tarefas de classificação ou regressão (CAMILO; SILVA, 2009).

2.4.1 Técnicas de Regressão para Séries Temporais

Nesta seção serão apresentadas algumas técnicas que implementam a tarefa de regressão na mineração de dados, como a regressão linear, random forest e redes neurais artificiais, especificamente as redes *Long-Short Term Memory* (LSTM) e *Perceptron Multilayer* (MLP).

2.4.1.1 Regressão Linear

Para Alpaydin (2010), quando a saída de um problema é um número, então temos uma regressão linear. Para exemplificar a regressão linear, vamos supor que pretendemos prever a média mensal de PH da seguinte forma: onde as entradas são os atributos do pH (desvio padrão, máximo e mínimo) e a saída é a média mensal do pH, conforme a Tabela 4. Se temos um conjunto de dados com os atributos de pH, podemos aplicar o aprendizado de máquina para treinar e prever a média mensal do pH.

Tabela 4 – Atributos de pH por ano de monitoramento.

Ano de monitoramento	Média mensal	Desvio padrão	Máximo	Mínimo
2007	6,51	0,18	6,9	6,36
2008	6,66	0,27	7,09	6,37
2009	6,42	0,18	6,62	6,26

Fonte: Adaptado de Vasconcelos e Souza (2011).

Então, x representa o conjunto dos atributos (entrada) e y a média mensal (saída). Assim, a Figura 9 representa graficamente a função

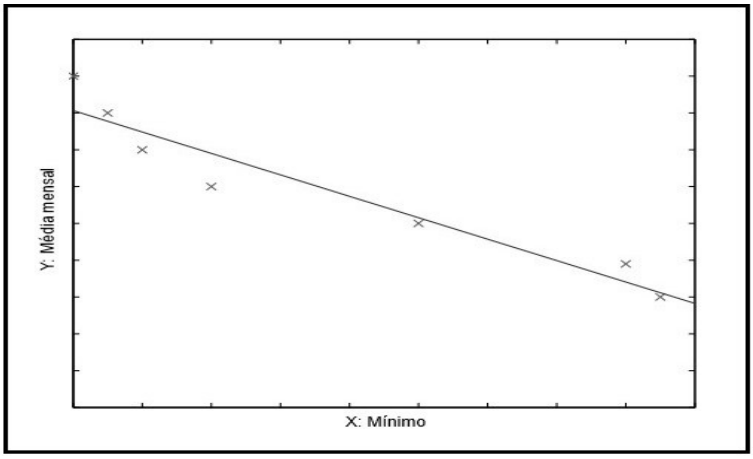
$$y = w_0 + w_1 * x \tag{2.8}$$

Onde,

- x representa a entrada (atributos);
- y representa a saída (número predito);
- w_0 e w_1 representa os pesos;

Russel e Norvig (2013) dizem que o valor de y é alterado pela mudança dos pesos relativos em cada interação. Assim, os pesos são ajustados de forma que ocorra uma proximidade

Figura 9 – Representação da função da regressão linear.



Fonte: Adaptado de Alpaydin (2010).

entre o valor real e o valor predito, conforme representado na Figura 9, onde a reta indica a saída da regressão linear e os pontos indicam os dados reais. A distância entre a reta e os pontos são os erros do peso ou perda de cada predição e para reduzir esse erro é aplicada a descida do gradiente que é uma função parametrizada por uma taxa de aprendizado utilizando os dados de treinamento de forma dinâmica até alcançar o erro mínimo possível (RUSSEL; NORVIG, 2013) produzindo como resultado um peso atualizado. A função da descida gradiente é:

$$w_i = w_i - taxa * erro * x_i \tag{2.9}$$

onde,

- w_i é o peso;
- $taxa$ é a taxa de aprendizado;
- $erro$ é a diferença entre o valor real e o valor predito;
- x_i é o valor da entrada;

Assim como em outros modelos, o ajuste do erro de predição é um dos critérios para a eficiência da regressão linear. O mesmo se aplica para a fase de validação (GUJARATI, 2003).

2.4.1.2 Random Forest

O Random Forest é um algoritmo de aprendizado de máquina aplicado nas tarefas de classificação, regressão e aprendizado não supervisionado, apresentando resultados satisfatórios (CHEN et al., 2020). Consiste no treinamento de várias árvores de decisão utilizando observações aleatórias geradas pela técnica bootstrap a partir da observação original, tendo em vista que na tarefa de regressão são assumidos valores numéricos e considera-se a média dos resultados de

cada árvore para melhorar a predição e o controle do sobreajuste (BREIMAN, 2001; ZHOU et al., 2016; OLIVEIRA et al., 2012; SCIKIT-LEARN, 2020). Segundo Roy e Larocque (2012), o random forest explora duas camadas de aleatoriedade, etapas I e II, aumentando a sua velocidade e precisão. Ainda define o funcionamento básico do algoritmo conforme a Figura 10.

Figura 10 – Etapas do processamento do random forest para regressão. I) seleciona as amostras dos dados originais através da técnica bootstrap; II) seleciona os melhores preditores de cada amostra da etapa I. Com isso, a árvore de regressão é aumentada; e III) são preditos novos dados através do cálculo médio das árvores.

Fonte: Adaptado de Roy e Larocque (2012).

2.4.1.3 Redes Neurais Artificiais

Haykin (2007) define que

> [...] uma rede neural é uma máquina que é projetada para modelar a maneira como o cérebro realiza uma tarefa particular ou função de interesse; a rede é normalmente implementada utilizando-se componentes eletrônicos ou é simulada por programação em um computador digital [...]

Diante disso, dois aspectos são semelhantes ao cérebro: conhecimento adquirido por processo de aprendizagem, que é o ajuste ordenado dos pesos sinápticos e a conexão entre neurônios para armazenar o conhecimento. A generalização é o principal benefício da rede neural, pois produz saídas adequadas para diferentes entradas usadas durante o treinamento da rede, por isso podendo ser aplicada a um grande conjunto de dados (HAYKIN, 2007). Além disso, as redes neurais resolvem problemas não-lineares (SCHULTE, 2019).

A rede neural artificial é estruturada com sinais de entrada (x) que são as variáveis da aplicação, os pesos sinápticos (w) que ajustam cada sinal de entrada, o combinador linear ($\sum$) soma os valores dos sinais de entrada com os pesos sinápticos produzindo como resultado o potencial de ativação (u); a função de ativação (g), podendo ser as equações 2.10 a 2.13, que definem o intervalo de valores na saída do neurônio e a saída (y) apresenta o valor final da rede neural (SILVA; SPATTI; FLAUZINO, 2010). A Figura 11 apresenta a estrutura da rede neural artificial.

- **Logística:** Retorna valores reais entre 0 e 1.

$$g(u) = \frac{1}{1+\theta^{-\beta u}} \tag{2.10}$$

- **Tangente Hiperbólica:** Retorna valores reais entre -1 e 1.

$$g(u) = \frac{1-\theta^{-\beta}}{1+\theta^{-\beta u}} \tag{2.11}$$

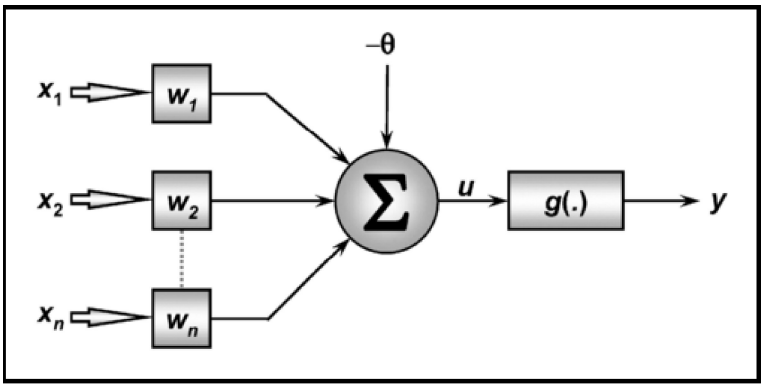
- **Gaussiana:** Retorna resultados iguais em relação ao valores do potencial de ativação equidistante .

$$g(u) = e^{-\frac{(u-c)^2}{2\theta}} \tag{2.12}$$

- **Linear:** Retorna valores iguais ao potencial de ativação.

$$g(u) = u \tag{2.13}$$

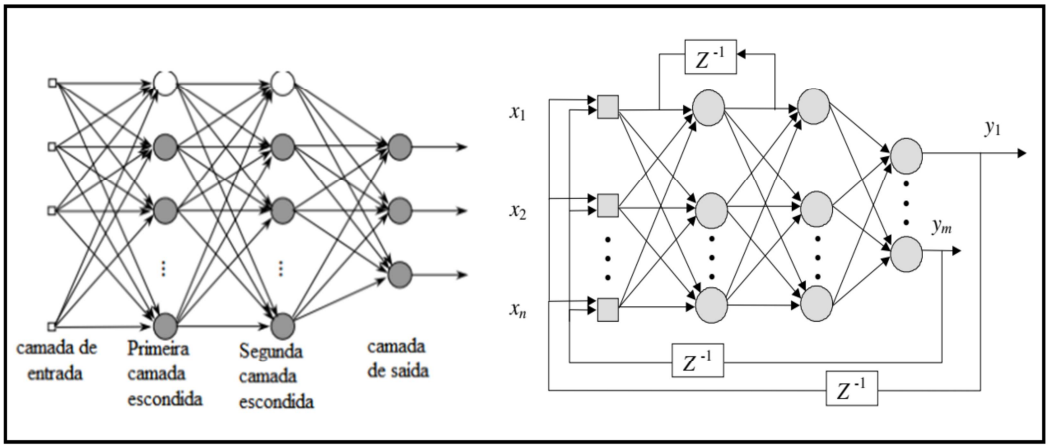
Figura 11 – Modelo de um neurônio.



Fonte: Adaptado de Silva, Spatti e Flauzino (2010).

Quanto à arquitetura da rede neural podemos citar a *Feed-Forward* e a *Feed-Backward*. A *Feed-Forward* também é chamada de redes alimentadas adiante, na qual a saída de uma camada é utilizada como entrada na próxima camada e assim por diante (HAYKIN, 2007). A sua composição é formada pela camada de entrada que recebe as variáveis do conjunto de dados por uma ou mais camadas escondidas, em que a quantidade de neurônios nessa camada depende do problema que será resolvido e do conjunto de dados, e a camada de saída, que retorna o resultado da rede. Por isso, esse tipo de arquitetura pode ser aplicado em problemas relacionados à aproximação de funções, classificação de padrões, identificação de sistemas, otimização e outros (SILVA; SPATTI; FLAUZINO, 2010). Já na *Feed-Backward*, também conhecida como rede recorrente ou acíclica, as saídas da rede realimentam os sinais de entrada de outros neurônios, sendo aplicada em problemas como previsão de séries temporais, otimização e identificação de sistemas, entre outros (SILVA; SPATTI; FLAUZINO, 2010). A Figura 12, representa as arquiteturas *Feed-Forward* e *Feed-Backward*.

Figura 12 – À Esq. Arquitetura *feedforward* múltipla camada. À Dir. Arquitetura recorrente ou realimentada.



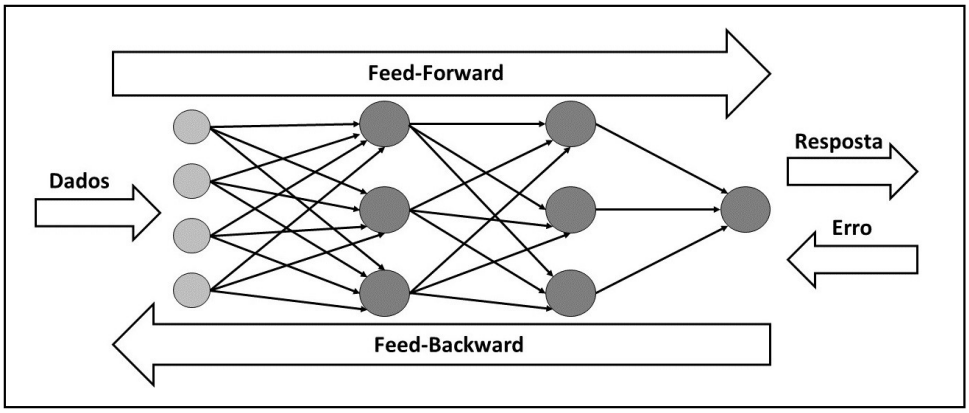
Fonte: Fleck et al. (2016) e Benini (2008).

2.4.1.3.1 Rede Neural MLP

As redes neurais MLP possuem a arquitetura *Feed-Forward* com treinamento supervisionado *BackPropagation* do erro, uma generalização do algoritmo mínimo quadrado médio para ajustar os pesos nos neurônios nas camadas escondidas (RÊGO et al., 2013) e possui duas fases: a propagação adiante (*forward*) e a propagação reversa (*backward*), conforme a Figura 13.

Na fase propagação adiante (*forward*) os sinais de entrada são propagados pelas camadas até a camada de saída, sem alteração dos pesos sinápticos e realizado o cálculo do erro entre as respostas desejadas e as produzidas pela rede (SILVA; SPATTI; FLAUZINO, 2010). Na propagação reversa (*backward*), o erro calculado na primeira fase, *forward*, é utilizado como parâmetro de ajuste dos pesos sinápticos em cada iteração da rede neural (SILVA; SPATTI; FLAUZINO, 2010).

Figura 13 – Fases do processo de treinamento Backpropagation.

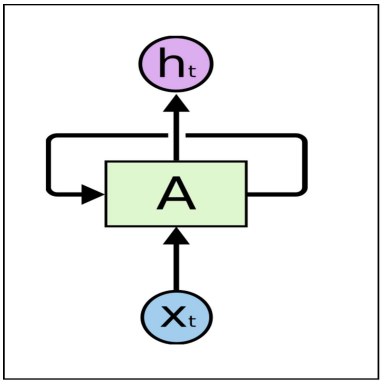


Fonte: Adaptado de Rêgo et al. (2013).

2.4.1.3.2 Rede Neural LSTM

As redes neurais LSTM possuem a arquitetura do tipo recorrente (*Feed-Backward*) com o objetivo de reparar a falta de memória de longo prazo da rede neural, retendo as principais informações dos sinais de entrada (BANDARA; BERGMEIR; SMYL, 2020) e a persistência dessas informações são realizadas com laços de repetição (OLAH, 2015). A Figura 14 exemplifica a persistência das informações na rede neural LSTM, onde X_t é um valor qualquer recebido pelo neurônio A que enviará a saída para o próximo neurônio e para si mesmo, através do laço de repetição (realimentação) com tempo de parada definido pelo número de neurônios configurados na rede.

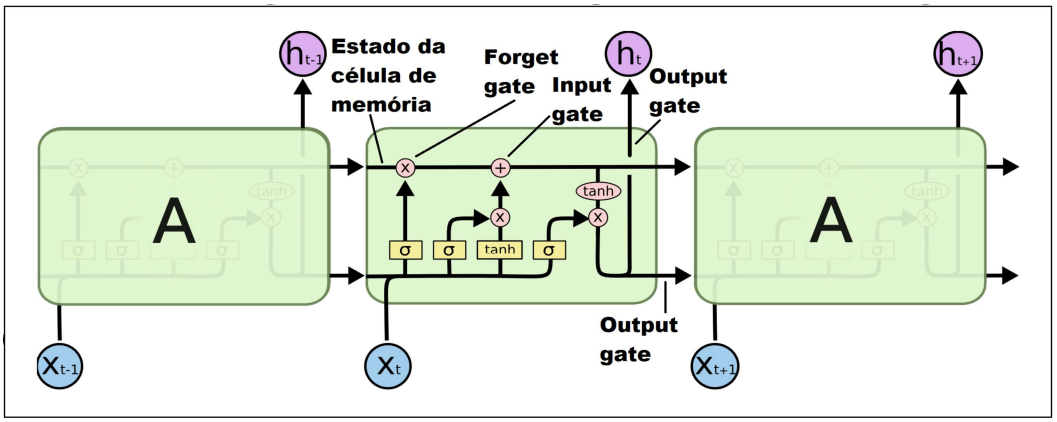
Figura 14 – Exemplo da persistência das informações na rede neural LSTM.



Fonte: Adaptado de Olah (2015).

Segundo Olah (2015), a estrutura da rede neural LSTM é composta por cadeia de células de memória (neurônios) e comportas (*Forget*, *Input* e *Output*). A cadeia das células de memórias indica o estado da célula e as comportas manipulam as memórias. A Figura 15 apresenta a estrutura da rede neural LSTM, onde *Forget gate* apaga a informação, consequentemente liberando a memória, a *Input gate* adiciona a informação e a *Output gate* lê as informações da memória.

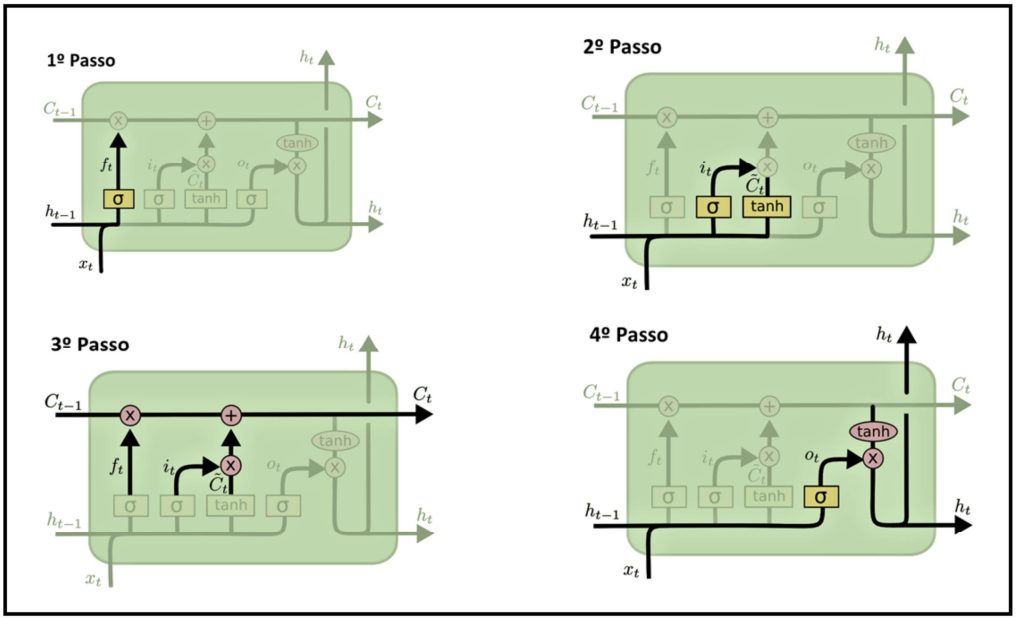
Figura 15 – Estrutura da rede neural LSTM.



Fonte: Adaptado de Olah (2015).

A rede neural LSTM funciona com 4 passos, os quais se iniciam com as entradas $h_t - 1$ e X_t, respectivamente, a saída do neurônio anterior e o valor atual, que serão repassados para a função sigmoide que selecionará as informações importantes e retorna valores entre 0 e 1. Se os valores forem iguais a 0, a informação é removida da memória. No 2º Passo são verificadas quais informações serão alteradas, criando um vetor com a função tangente hiperbólica, retornando novos valores entre -1 e 1 e adicionando-os no lugar do que foi removido no 1º passo; no 3º passo, o estado da célula de memória é atualizado executando os passos anteriores; e no 4º passo, retorna como saída as informações necessárias, aplicando a função sigmoide que coloca os valores entre -1 e 1 (OLAH, 2015). A Figura 16, apresenta os passos descritos.

Figura 16 – Passos do funcionamento da rede neural LSTM.



Fonte: Adaptado de Olah (2015).

2.5 Trabalhos Relacionados

Nesta seção serão descritos os trabalhos científicos relacionados com o tema da pesquisa, por isso durante o processo de seleção nos repositórios prevaleceram aqueles que experimentaram a predição de algum parâmetro da qualidade usando o aprendizado de máquina.

Desse modo, a Tabela 5 demonstra que os trabalhos científicos durante os experimentos fazem uso de diferentes técnicas de aprendizado de máquina como as redes neurais (ID: 2, 19 e 31), Splines Regressão Adaptativa Multivariada (ID: 20), Árvores de Decisão (ID: 20), Máquina de Vetor de Suporte (ID: 20) e Softplus Extreme (ID: 33), sendo que algumas estão associadas a métodos de pré-processamento de dados (ID: 21), visando aumentar a qualidade dos dados de entrada e contribuir para a melhora dos desempenhos das predições dos parâmetros da qualidade da água nos diferentes corpos hídricos.

Além do uso de diferentes técnicas nestes experimentos de predição, ainda, nota-se que em todos os trabalhos ocorre a predição temporal do parâmetro oxigênio dissolvido. Dessa forma, é ressaltada a importância deste parâmetro para o planejamento e gestão da qualidade dos recursos hídricos. Ademais, os dados usados nos trabalhos relacionados são de recursos hídricos localizados no continente Asiático, destacando-se os países como a China e Índia e na América do Norte, especificamente, nos Estados Unidos.

Portanto, diante do exposto, há oportunidades de lacunas a serem exploradas nesse ramo de pesquisa relacionada com a predição de parâmetros da qualidade usando aprendizado de máquina. A primeira delas é referente aos dados oriundos dos monitoramentos dos parâmetros da qualidade da água dos recursos hídricos brasileiros, visto que o Brasil é um país reconhecido pela sua abundância hídrica, principalmente na região amazônica. Com isso, torna-se importante encontrar dados que servem para os experimentos de predição dos parâmetros da qualidade da água usando aprendizado máquina. Outra lacuna é a importância da predição de outros parâmetros da qualidade da água, não somente o oxigênio dissolvido como visto na Tabela 5, porque a associação dos parâmetros define o índice da qualidade da água, contribuindo com o processo de planejamento da gestão hídrica. Além disso, outras técnicas mais simples como a regressão linear e random forest ou mais complexas como as redes neurais MLP e LSTM podem ser usadas como alternativas para a predição temporal. Essa comumente presente nos trabalhos da Tabela 5 e a predição espaço-temporal como uma opção de predição a ser experimentada. Por fim, através da comparação dos desempenhos entre as técnicas será conhecida a técnica que melhor funciona para cada parâmetro quando utiliza-se os dados dos recursos hídricos do território brasileiro.

Tabela 5 – Resumo dos trabalhos relacionados.

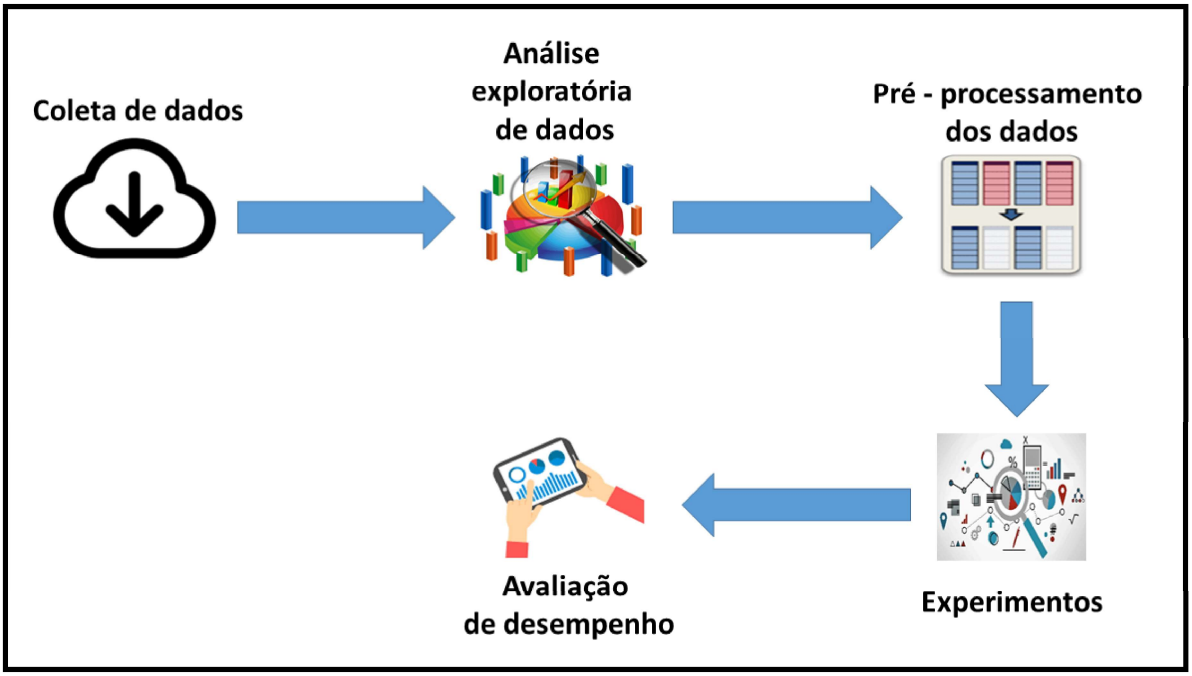
Autor	Objetivo	Técnica / Metodologia	Resultados
Sarkar e Pandey (2015).	Simular os valores do OD no Rio Yamuna (Índia).	Rede neural MLP com algoritmo Feed Forward / Utilizou 3 redes neurais com 14, 10 e 5 variáveis de entrada, ajuste de pesos usando a regra delta generalizada e 5000 épocas de iteração. Os dados com frequência mensal sobre a vazão, tempo de percurso, temperatura, pH, condutividade elétrica, DBO e DO de 3 estações diferentes. As métricas RMSE, coeficiente de correlação (R) e R^2 avaliam os modelos.	Melhor desempenho com 10 variáveis na entrada na rede neural.
Alizadeh e Kavianpour (2015).	Prever os valores diários da salinidade, temperatura e OD, além do OD por hora na baia de Hilo Bay, Oceano Pacífico.	Rede neural Feed-Backward com algoritmo de otimização Levenberg-Marquardt e Wavelets. / Os dados sobre salinidade, turbidez, OD, temperatura e a clorofila têm 650 amostras diárias e 540 amostras horárias. No experimento dados foram divididos em treinamento (70%), validação (15%) e conjuntos de testes (15%). As métricas coeficiente de correlação (R) e RMSE avaliam os modelos.	O Wavelets melhorou o desempenho das redes neurais nas predições e, sozinho, obteve o melhor desempenho na predição horária do OD.
Heddam (2016)	Modelar e prever os valores de OD no rio Klamath, EUA.	Rede neural MLP Feed Forward e rede neural de função de base radial / Os dados sobre OD, ph, temperatura SC e SD foram divididos em conjuntos de treinamento(60%), validação(20%) e teste(20%) durante a modelagem, sendo que para predizer OD, considerou-se apenas a série temporal do OD. As métricas coeficiente de correlação, RMSE e MAE avaliam os modelos.	O desempenho é melhor quando o modelo é configurado com 72h de janela deslizante.
Heddam e Kisi (2018)	Prever os valores diários de OD de três estações do Serviço Geológico dos Estados Unidos.	Splines de Regressão Adaptativa Multivariada (MARS), M5 modelo de árvore (M5T) e Máquina de vetor de suporte de mínimos quadrados (LSSVM) / Os dados sobre temperatura, ph, DI, SC e DO foram divididos em treinamento, validação e teste. As variáveis de entrada nos modelos foram combinadas com o uso do coeficiente de correlação. As métricas coeficiente de correlação (R), RMSE e MAE avaliam os modelos.	O modelo LSSVM apresenta o melhor entre todos quando usa os parâmetros TE e pH como entrada no modelo.
Huan, Cao e Qin (2018)	Prever OD no lago para aquicultura na cidade de Changzhou, China.	Máquina de vetor de suporte de mínimos quadrados (LSSVM) com Decomposição em modo empírico (EMD). / Para melhorar o desempenho do modelo LSSVM são usados o EMD, que decompõe a série temporal OD, além da estrutura de evidências bayesiana que obtêm os melhores valores dos parâmetros. As métricas RMSE, MAPE, MAE e EMAX, avaliam o modelo.	O EMD-LSSVM tem melhor desempenho de predição do que os modelos WD-LSSVM, EEMD-ELM e o LSSVM padrão.
Shi et al. (2019)	Prever a mudança de OD numa aquicultura na cidade de Wuxi, China.	Softplus Extreme baseado em Clustering (CSELM) / Os dados sobre OD, pH e temperatura, umidade, pressão atmosférica, dióxido de carbono, entre outros, possui 4464 amostras, das quais 3780 são usadas no treinamento e 684 para teste. Na fase de pré-processamento, o k-medoids agrupa as amostras por similaridade de tempo, em seguida, CSELM é treinado e testado. As métricas de RMSE, MAPE, MAE, NSC e Time, avaliam o modelo.	O CSELM tem melhor desempenho quando comparado aos modelos existentes PLS-ELM e ELM.

Fonte: Elaborada pelo autor.

3 METODOLOGIA DA PESQUISA

Neste capítulo é apresentada a metodologia utilizada na pesquisa através da descrição dos procedimentos para a coleta de dados, análise exploratória dos dados, pré-processamento dos dados, configuração dos experimentos e avaliação do desempenho dos modelos experimentados, conforme a Figura 17.

Figura 17 – Fluxo da metodologia aplicada na pesquisa.



Fonte: Elaborada pelo autor.

3.1 Coleta de dados

O procedimento metodológico da pesquisa é iniciado com a busca dos dados históricos sobre os parâmetros da qualidade da água de bacias hidrográficas localizadas no Brasil que possuam dados numéricos. Para isso, procurou-se mapear e estabelecer contato com os órgãos públicos que atuam na gestão de recursos hídricos, na gestão ambiental, no tratamento da água para abastecimento público, na vigilância à saúde ou em projeto de pesquisa relacionado à qualidade da água, tanto no governo federal quanto em governos estaduais. Na aquisição dos dados, em algumas instituições, houve contato presencial com o responsável pelo departamento, em outras, através da Lei Nº 12.527 - Lei de Acesso à Informação – LAI, que regula o acesso às informações de interesse público ou por meio de ferramenta online disponibilizada pela instituição.

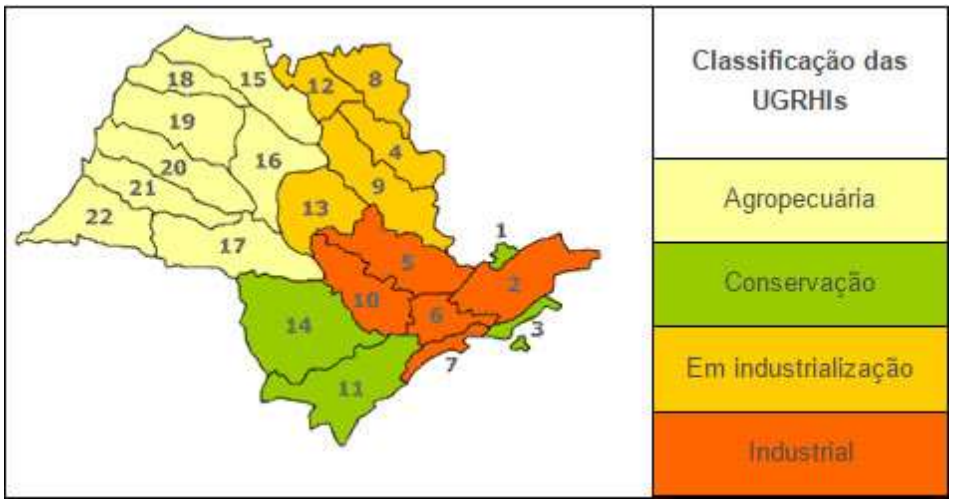
Desse modo, os conjuntos de dados recebidos possuem dados brutos e a escolha da fonte de dados para a predição seguiu os critérios relacionados com a proposta de uma pesquisa

com abordagem quantitativa. Dessa forma, o conjunto de dados deve possuir valores numéricos dos parâmetros da qualidade da água e a data da realização das observações, pois viabilizam a transformação do conjunto de dados em serie temporal e a quantidade de observações deve ser suficiente para os experimentos de predição dos parâmetros da qualidade da água usando aprendizado de máquina.

Diante disso, optou-se por utilizar os dados dos parâmetros da qualidade da água das bacias hidrográficas do Estado de São Paulo, cujas coletas de dados são realizadas pela CETESB a qual também disponibiliza a série histórica de parâmetros como coliformes, dbo, fósforo, nitrogênio, oxigênio dissolvido, sólido total, temperatura, turbidez e pH totalizando 20.890 amostras com periodicidade de monitoramento mensal entre janeiro de 1978 a dezembro de 2019, especificamente das UGRHI 02, 05, 06, 07 e 10, classificadas como industrializadas (CETESB, 2019), conforme a Figura 18.

Essas UGRHI possuem um forte potencial econômico da região com Produto Interno Bruto (PIB) estadual estimados em mais de R$ 1 trilhão. Além disso, a maior parte dos 43 milhões de habitantes do Estado de São Paulo estão nas regiões das UGRHI 02,03,05,06, 07 (CRH, 2017), portanto essas características demonstram que a qualidade da água pode ser impactada pelas atividades socioeconômicas nesta região. Esse conjunto de dados foi extraído do Sistema de Informações InfoÁGUAS[1].

Figura 18 – Unidades de Gerenciamento de Recursos Hídricos (UGRHI) no Estado de São Paulo.



Fonte: CETESB (2019).

Por fim, a análise exploratória dos dados é fundamental para o entendimento do comportamento das series temporais de cada UGRHI, de modo que auxilie na seleção da UGRHI e na definição dos seus pontos de coleta que servem para a tarefa de regressão. Além disso, no pré-processamento ocorrem os seguintes procedimentos: a transformação da frequência das series temporais para mensal, a limpeza dos dados, o tratamento de valores ausentes, a transformação logarítmica e a normalização dos valores dos parâmetros. Ademais, tem a fase dos

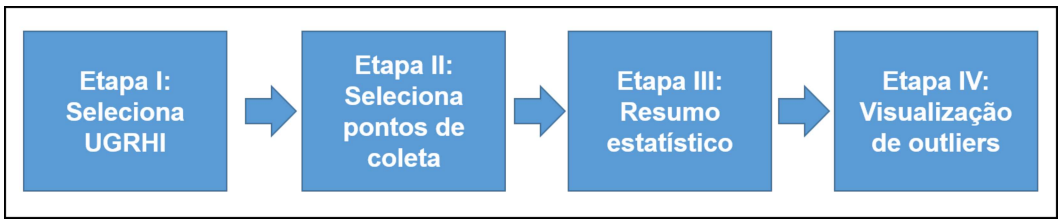
[1] CETESB (2019). Infoáguas .https://sistemainfoaguas.cetesb.sp.gov.br/. [Online; De-zembro 7 de 2019]

experimentos de predição temporal e espaço-temporal dos parâmetros da qualidade da água. Por fim, os resultados médios de MAPE e RMSE das predições são comparados com as médias de MAPE e RMSE dos baselines e das técnicas de aprendizado de máquina a partir das séries temporais dos parâmetros.

3.2 Análise Exploratória dos Dados

Para tanto, o fluxo da análise exploratória dos dados é apresentada na Figura 19. Assim, na Etapa I é selecionada a UGRHI que possui a maior quantidade de amostras para cada parâmetro permitindo eliminar parâmetros que têm o quantitativo de amostras insuficientes para o experimento de predição usando aprendizado de máquina, além de determinar o parâmetro que possui o maior quantitativo de amostras, pois este será utilizado na Etapa II. Com essas definições na Etapa I, dois pontos de coleta com maior quantidade de amostras e menor quantidade de valores ausentes são destacados na Etapa II, pois valores ausentes podem interferir na qualidade dos resultados (SILVA; PERES; BOSCARIOLI, 2017). Diante disso, na Etapa III, através da análise do resumo estatístico, que expõe as medidas de tendência central como a média e mediana, as medidas de dispersão como o desvio padrão, valor mínimo e máximo, são analisadas as distribuições dos valores dos parâmetros de cada ponto de coleta selecionado na Etapa II. Por fim, na Etapa IV, verifica-se a presença de *outlier* através da representação gráfica do diagrama de caixa ou boxplot.

Figura 19 – Fluxo da análise exploratória dos dados.



Fonte: Elaborada pelo autor.

3.3 Pré-processamento dos dados

Nesta fase, os resultados da análise dos dados serão utilizados para definir os procedimentos aplicáveis na organização dos dados, que são os seguintes: transformação da frequência das séries temporais para mensal, pois os dados dos parâmetros, ainda que monitorados mensalmente apresentam desorganização na frequência temporal; a limpeza dos dados, através da retirada dos *outliers*, sendo localizados pelo cálculo dos limites, superior e inferior utilizando a regra 1.5xFIQ, representada pelas as Equações 3.1 e 3.2 (ACADEMY, 2019); tratamento dos valores ausentes aplicando o método da interpolação linear para substituir esses valores por novos criados a partir dos dados existentes no conjunto de dados (WANDERLEY; AMORIM; CARVALHO,

2014); transformação logaritmica dos valores dos parâmetros que apresentam alta variância e a normalização dos valores dos parâmetros utilizando a Equação 3.3, implementada pela função *MinMaxScaler* do pacote *sklearn.preprocessing* da linguagem Python para organizar os dados dentro de uma escala padrão.

$$limiteSuperior = terceiroQuartil + (1.5 * FIQ) \tag{3.1}$$

$$limiteInferior = primeiroQuartil - (1.5 * FIQ) \tag{3.2}$$

Onde a faixa interquartil (FIQ) é calculada pela subtração do terceiroQuartil - primeiro-Quartil.

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{3.3}$$

onde x é a variável a normalizar e x' é o resultado da normalização *min-max*.

3.4 Configuração dos Experimentos de Predição

Nos experimentos se propõe a executar as predições temporal e espaço-temporal dos parâmetros da qualidade da água de forma individualizada. Desse modo, na predição temporal são utilizados os dados de uma série temporal para predizer o valor futuro de acordo com os valores temporalmente anteriores (FERNANDES, 1995), enquanto que na predição espaço-temporal são usadas duas séries temporais distintas para predizer o valor futuro, conforme a Figura 20. Para tanto, são utilizados algoritmos de aprendizado de máquina como a regressão linear, random forest e as redes neurais MLP e LSTM para predizer os valores dos parâmetros da qualidade da água.

Diante disso, o algoritmo MLP tem reconhecida contribuição na tarefa de classificação de imagens na pesquisa médica (LORENCIN et al., 2020), assim como na tarefa de regressão para predizer o impacto de doenças altamente infecciosa (CAR et al., 2020). Por isso, é relevante experimentar a regressão com o MLP utilizando dados dos parâmetros da qualidade da água.

Os algoritmos foram implementados com a linguagem de programação Python, através da IDE Jupyter Notebook (Versão 6.1.4) presente na Interface Gráfica do Usuário (GUI) Anaconda Navigator. O Python foi escolhido para a implementação dos algoritmos devido a sua popularidade na área de ciência de dados e a diversidade de bibliotecas para a computação científica e aprendizado de máquina (RASCHKA; MIRJALILI, 2017).

Basicamente, os modelos de regressão linear são configurados com o parâmetro *fit_intercept* = *True*, de modo que automaticamente sejam calculados os valores de interceptação que resultam

no menor erro de predição (SCIKIT-LEARN, 2020). Nos modelos de random forest são configurados com os parâmetros *n_estimators* = 100, que definem o número de árvores no modelo, *criterion* = *squared_error* para avaliar a qualidade das ramificações das árvores e *bootstrap* = *True*, que faz uma amostragem aleatória para substituir as amostras de entrada no treinamento (SCIKIT-LEARN, 2020).

Os modelos das redes neurais têm 100 épocas de iteração, monitoramento da métrica de erro (*EarlyStopping*) limitado a 5 épocas sem melhora de desempenho e função de ativação *Relu* nas camadas facilitando o treinamento (NAIR; HINTON, 2010). Também há 50 neurônios nas camadas de entrada e intermediária conforme demonstrado por Almeida et al. (2020), de que esta quantidade de neurônios produz os melhores desempenhos na predição temporal dos parâmetros e 1 neurônio na saída. A camada intermediária no LSTM é implementada com *Dropout* de 0.2 para evitar o ajuste excessivo (*overfitting*) dos dados durante o treinamento, compilado com a métrica MAE e otimizador *adam* com taxa de 0.001 para ajustar os pesos de acordo com MAE (OTT; SARTORI, 2018). Os dados são divididos em 70% para treinar e 30% para testar.

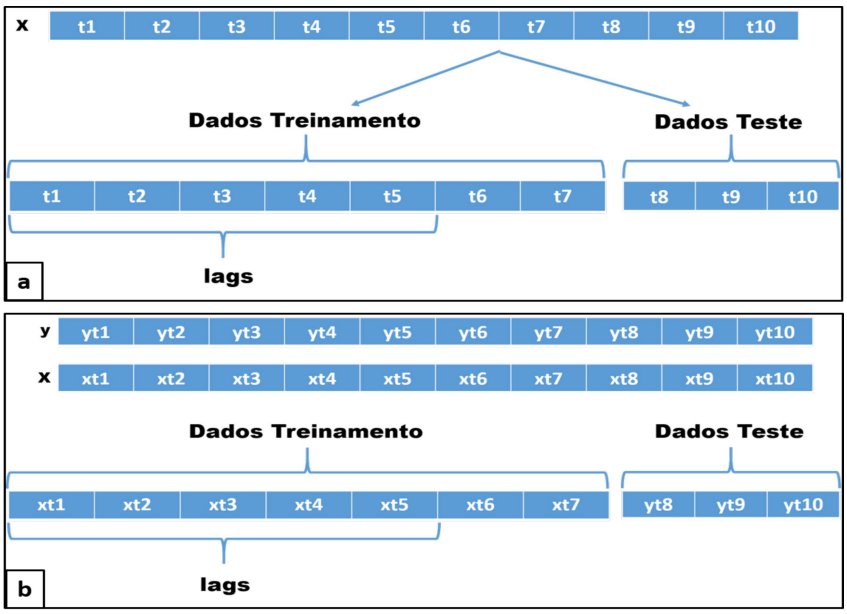
Ademais, os modelos têm um tempo (lags) definidos para predizer o valor do parâmetro e determinado pela quantidade de observações imediatamente anteriores ao valor predito (SHIRAISHI; MARUJO, 2020), transformando a série histórica em problemas de aprendizado supervisionado para a tarefa de regressão (MATSUMOTO et al., 2019), conforme a Figura 20. Desse modo, são atribuídas de 1 a 10 observações como lags e nas redes neurais, que são técnicas não determinísticas, é obtida a média do resultado de 5 simulações para cada lag.

Além disso, nas predições dos experimentos são utilizados os dados de testes, devido eles serem desconhecidos do modelo treinado, como Silva, Peres e Boscarioli (2017) citam

> Independentemente da medida de avaliação a ser usada para atestar a qualidade de um modelo, não é adequado avaliá-lo por seu desempenho em relação aos exemplares apresentados no processo de treinamento (indução). É sempre necessário saber como o modelo se comporta quando aplicado a exemplares que ainda não conhece...O motivo para essa ressalva é que modelos preditivos, a depender de como são gerados, podem levar à manifestação de um fenômeno bastante conhecido, o sobreajuste (do inglês *overfitting*).

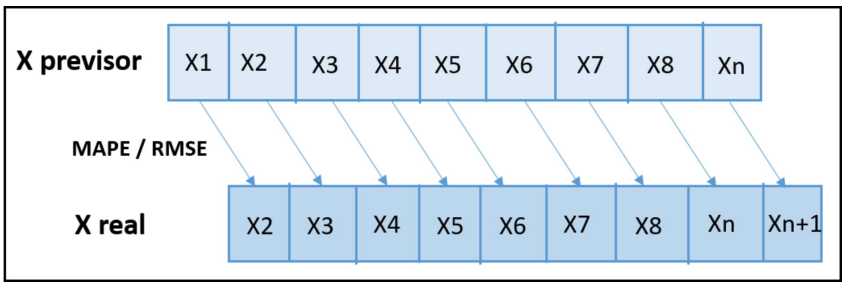
Por fim, os resultados médios de MAPE e RMSE das predições são comparados com as médias de MAPE e RMSE do baseline, conforme exemplo da Figura 21, e das técnicas a partir das séries temporais dos parâmetros.

Figura 20 – Esquema de predição. a) Predição temporal, onde x é uma série temporal que será dividida em treinamento (70%) e teste (30%), com os lags sendo os preditores de *t6* no treinamento. b) Predição espaço-temporal, onde x e y são séries temporais e 70% da série x são usados no treinamento com os lags sendo os preditores de *y* e 30% da série y são usados no teste.



Fonte: Elaborada pelo autor.

Figura 21 – O valor anterior X1 do X previsor é comparado ao valor futuro X2 do X real.



Fonte: Elaborada pelo autor.

4 RESULTADOS

Neste capítulo serão apresentados os resultados da análise exploratória dos dados e a avaliação de desempenho dos modelos regressão linear, random forest, MLP e LSTM na predição dos parâmetros da qualidade da água.

4.1 Análise exploratória dos dados

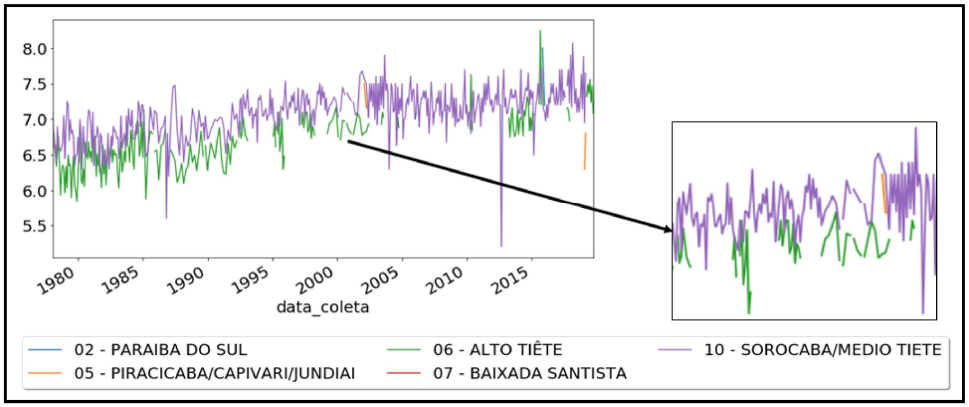
A Tabela 6 demonstra que os dados da UGRHI 06 são mais adequados para os experimentos, pois têm maior quantidade de amostras. Além disso, o parâmetro Nitrogênio possui quantidade de dados insuficiente para os experimentos de predição, por isso é descartado nesta pesquisa. Contrariamente, os dados do parâmetro pH tem a maior quantidade de amostras do que os outros parâmetros, por isso é apresentada nos resultados da seleção dos pontos de coleta, que serão utilizados nos experimentos. Diante disso, a Figura 22, através do gráfico de linhas de comportamento do valor do parâmetro pH entre os anos de 1978 a 2019, demonstra que apesar de possuir maior quantidade de amostras, a série temporal da UGRHI 06 possui falhas na amostragem dos dados ao longo do tempo de coleta, como nos anos de 1985, 1995, 2000 e, maior ainda, entre 2005 a 2015.

Tabela 6 – Quantitativo de amostras de cada UGRHI.

Parâmetros	UGRHI 02	UGRHI 05	UGRHI 06	UGRHI 07	UGRHI 10	TOTAL
Coliformes	41	127	**1663**	30	1221	3082
DBO	111	310	**2248**	72	1426	4167
Fósforo	112	213	**2240**	72	1413	4063
Nitrogênio	5	2	**52**	0	52	111
OD	112	311	**2036**	73	1421	3953
pH	**112**	**306**	**2257**	**72**	**1431**	**4188**
Sólido	112	243	**2241**	71	1435	4102
Temperatura	112	309	**2254**	73	1432	4180
Turbidez	112	309	**2218**	73	1432	4144

Fonte: Elaborada pelo autor.

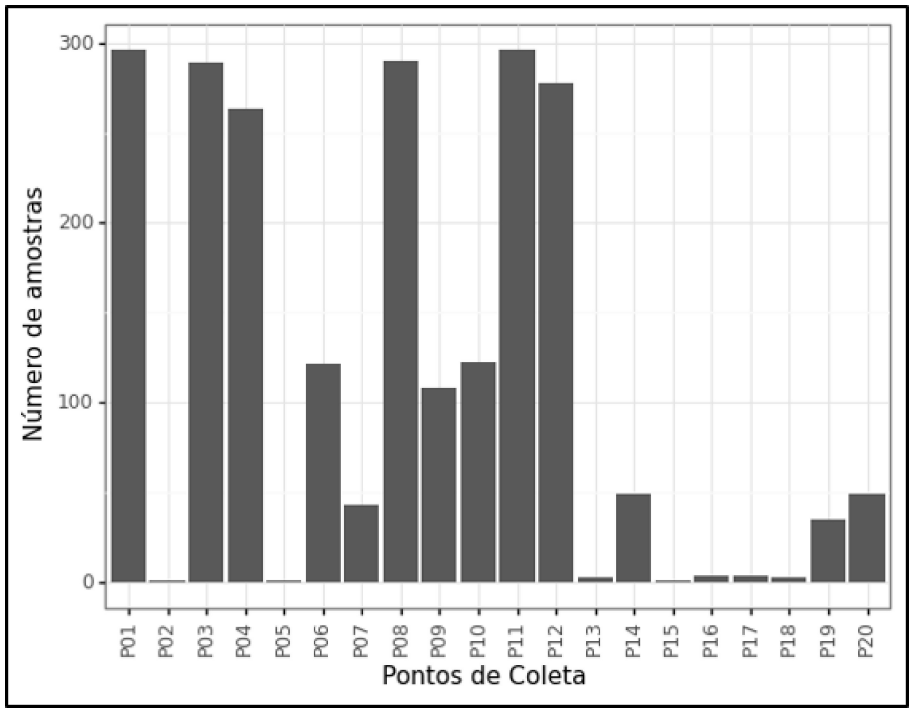
Figura 22 – Comportamento das séries temporais com média mensal do valor de pH de cada UGRHI.



Fonte: Elaborada pelo autor.

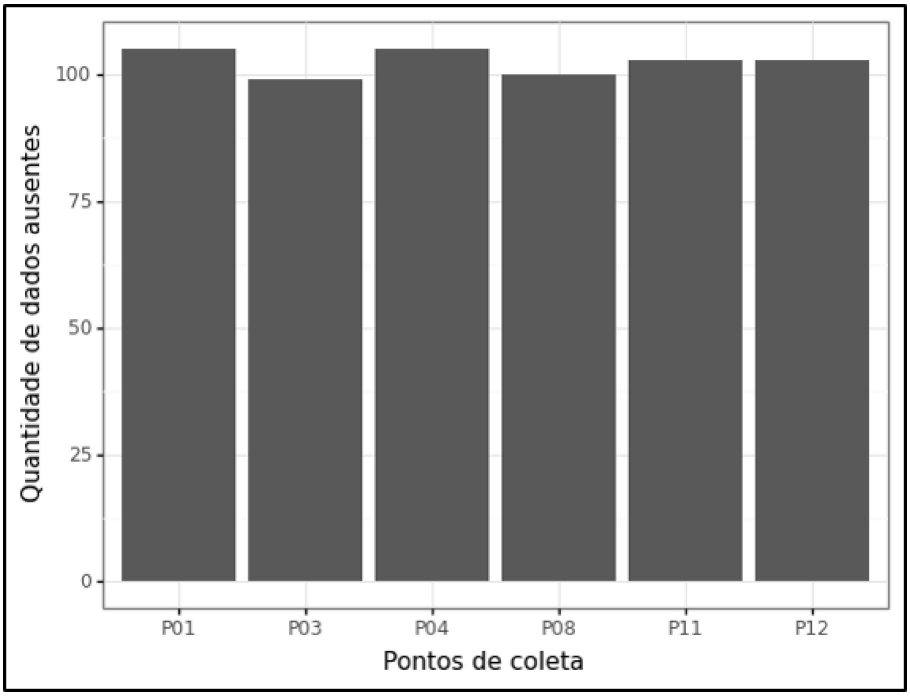
A Figura 23 através do gráfico de colunas de quantidade de amostras por pontos de coleta da UGRHI 06, demonstra que há 6 pontos de coleta com mais de 200 amostras enquanto que 3 pontos têm entre 100 e 200 amostras e os demais pontos possuem menos de 100 amostras. Portanto, os dados dos pontos de coleta P01, P03, P04, P08, P11 e P12 são candidatos para utilização nos experimentos. Assim, a Figura 24, através do gráfico de colunas de quantidade de dados ausentes por pontos de coleta, indica que os pontos P03 e P08 são aptos aos experimentos, pois apresentam os menores quantitativos de valores ausentes dentre os 6 pontos selecionados anteriormente com maiores quantidades de amostras na UGRHI 06. Por isso, nos experimentos são utilizados os dados dos monitoramentos dos parâmetros pH, coliformes, Demanda Bioquímica de oxigênio (DBO), fósforo, Oxigênio Dissolvido (OD), sólido, temperatura e turbidez dos pontos de coleta, P03 e P08, com coordenadas geográficas (latitude: 233354, longitude: 460057 altitude: 750 m) e (latitude: 233255, longitude: 460809 altitude: 735 m), respectivamente. Finalmente, a Figura 25, através do gráfico de linhas de comportamento dos valores de pH dos pontos P03 e P08 durante o tempo de coleta, demonstra que as séries temporais desses pontos têm interseções dos valores de pH nos anos de 1985, 1995, entre os anos 2000 a 2005, em 2010 e entre 2015 a 2020.

Figura 23 – Número de amostras por ponto de coleta.



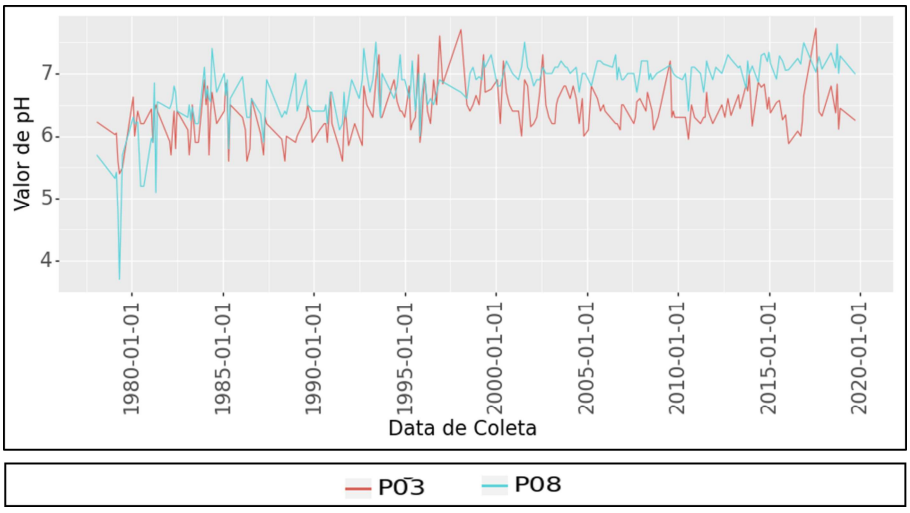
Fonte: Elaborada pelo autor.

Figura 24 – Quantidade de dados ausentes por ponto de coleta.



Fonte: Elaborada pelo autor.

Figura 25 – Séries temporais dos pontos de coleta selecionados.



Fonte: Elaborada pelo autor.

Tabela 7 – Resumo estatístico dos dados do ponto de coleta P03.

Parâmetros	Amostras	Média	STD	Min	50%	Máx
Coliformes	171.00	447.05	1685.88	0.00	36.00	17000.00
DBO	206.00	3.11	2.11	0.10	3.00	19.00
Fósforo	208.00	0.06	0.06	0.00	0.05	0.49
OD	208.00	4.62	1.92	0.00	4.87	10.20
pH	208.00	6.38	0.39	5.40	6.38	7.72
Sólido	208.00	69.08	55.82	8.00	56.00	684.00
Temperatura	208.00	20.91	2.37	15.00	21.00	29.00
Turbidez	206.00	8.56	10.45	0.00	5.04	70.00

Fonte: Elaborada pelo autor.

Tabela 8 – Resumo estatístico dos dados do ponto de coleta P08.

Parâmetros	Amostras	Média	STD	Min	50%	Máx
Coliformes	170.00	1053989.23	3483274.16	1.00	270000.00	32033833.33
DBO	206.00	25.17	20.57	2.00	18.00	93.00
Fósforo	208.00	0.86	0.97	0.04	0.52	8.70
OD	197.00	0.74	0.88	0.00	0.40	4.50
pH	207.00	6.77	0.50	3.70	6.90	7.50
Sólido	208.00	350.07	160.57	100.00	325.50	1490.00
Temperatura	207.00	21.39	2.68	14.00	21.90	27.00
Turbidez	204.00	39.49	38.74	0.37	29.00	270.00

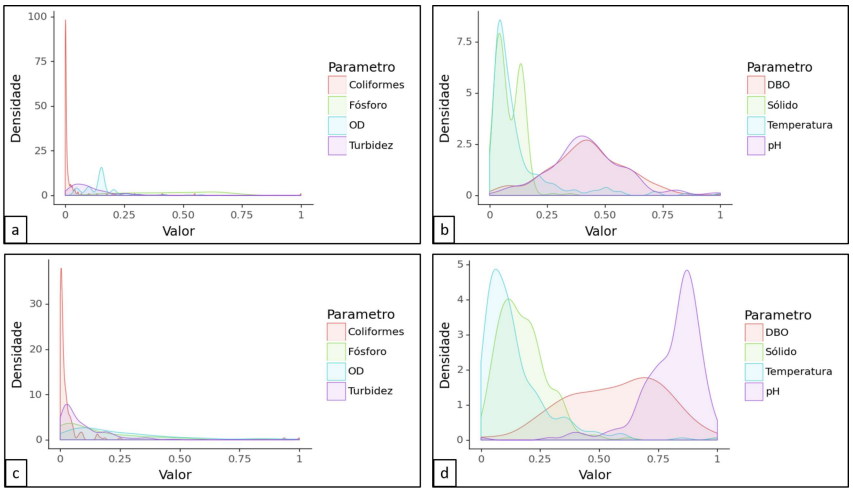
Fonte: Elaborada pelo autor.

Segundo Silva, Peres e Boscarioli (2017), as medidas de tendência central como a média e a mediana permitem identificar o comportamento dos valores no conjunto de dados. Ademais, a comparação entre essas medidas revelam informações sobre a distribuição da frequência dos valores, classificando em simétrica e assimétrica.

Assim, nas Tabelas 7 e 8, os valores indicam que há uma distribuição assimétrica dos valores dos parâmetros nos pontos de coleta, exceto os valores do parâmetro pH que é simétrico, pois a média e a mediana possuem os mesmos valores no ponto de coleta P03. Além disso, o desvio padrão indica a consistência de um fenômeno (SILVA; PERES; BOSCARIOLI, 2017), então o baixo desvio padrão, em relação à média, indica uniformidade entre os valores dos parâmetros pH, DBO, sólido e temperatura, em ambos pontos de coleta. Dessa forma, com os valores dos parâmetros normalizados, as Figuras 26b e 26d, através do gráfico de densidade dos valores de DBO, sólido, temperatura e pH, demonstram concentração e interseção entre os valores dos parâmetros no ponto P03, aproximadamente, com os valores DBO e pH entre 0.25 a 0.60 e os valores de sólido e temperatura entre 0 a 0.20. Enquanto que no ponto P08, a concentração dos valores de DBO e pH não estão interseccionados, aproximadamente, entre 0.25 a 0.75 e 0.75 a 1, respectivamente. Porém, nos parâmetros sólido e temperatura há interseção dos valores concentrados, entre 0 a 0.40 e 0 a 0.20, respectivamente. Todavia, os valores de coliformes, fósforo, OD e turbidez estão mais distribuídos entre 0.25 a 1 e mais concentrados

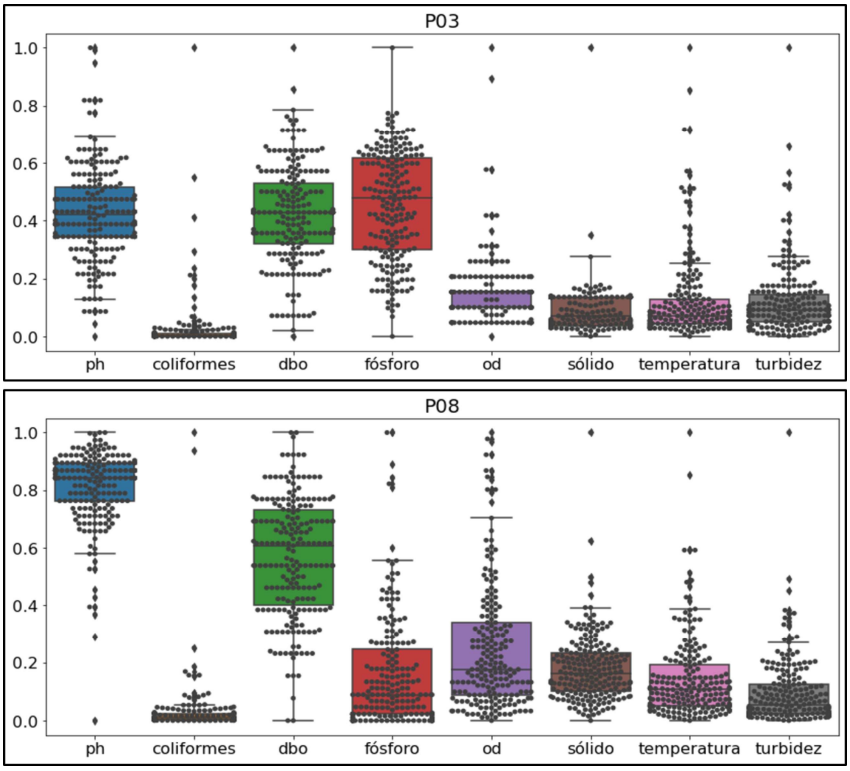
com alguma interseção entre 0 a 0.25, em ambos os pontos de coleta, conforme as Figuras 26a e 26c, através do gráfico de densidade dos valores desses parâmetros. Por fim, na Figura 27, comparando a variação entre os valores dos parâmetros, através do diagrama de caixa, temos que, em ambos pontos de coleta, há ocorrência de *outliers*, exceto nos valores do fósforo no ponto P03 e DBO no ponto P08.

Figura 26 – Em (a) e (c), respectivamente, P03 e P08 estão mais distribuídos. Enquanto, em (b) e (d), respectivamente, P03 e P08 estão mais uniformes.



Fonte: Elaborada pelo autor.

Figura 27 – Outliers nos valores dos parâmetros dos pontos de coleta selecionados.



Fonte: Elaborada pelo autor.

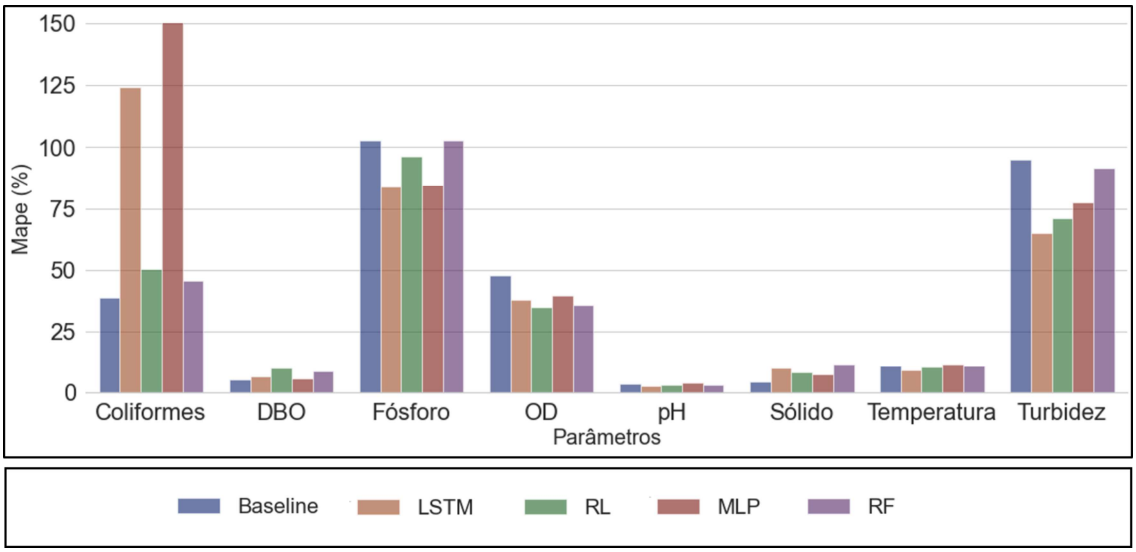
4.2 Experimentos

Nesta seção são avaliados os resultados das predições temporal e espaço-temporal individuais dos parâmetros da qualidade da água, usando as técnicas de aprendizado de máquina Regressão Linear (RL), Random Forest (RF), MLP e LSTM. Para tanto, os resultados são avaliados pelas métricas de erro MAPE e RMSE.

4.2.1 Predição Temporal

Então, na predição temporal com os dados do ponto de coleta P03, a Figura 28, através do gráfico de coluna com MAPE médio por parâmetros, demonstra que dos 8 parâmetros preditos, apenas os parâmetros DBO, sólido total, temperatura da água e pH têm as melhores predições com MAPE médio $<= 15\%$ em todas as técnicas. Diferentemente disso, o MAPE médio dos parâmetros coliformes, fósforo, OD e turbidez estão acima de 25%, devido a predominância da variância nos valores destes parâmetros conforme a Figura 26a.

Figura 28 – Desempenho da predição temporal dos parâmetros.



Fonte: Elaborada pelo autor.

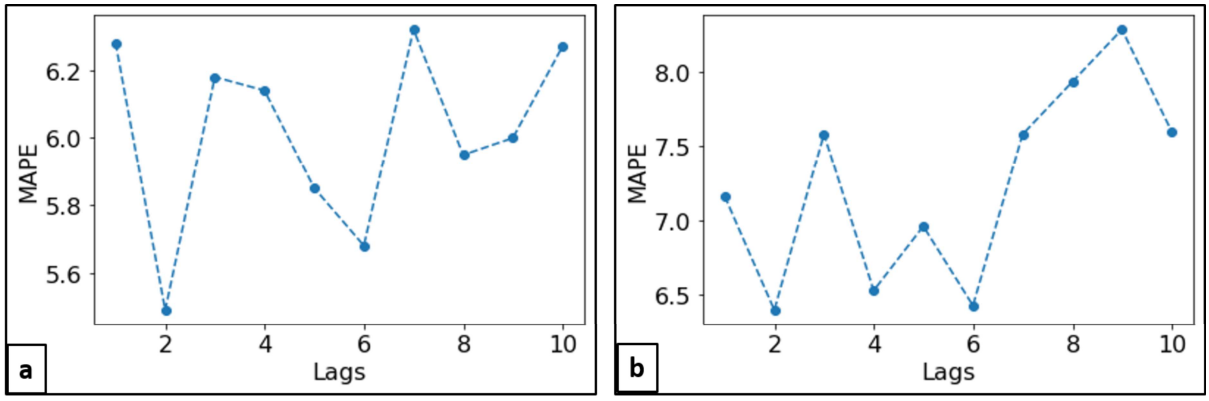
Diante disso, os melhores desempenhos de predição temporal ocorrem quando os modelos são configurados com 2 lags, pois têm o menor MAPE médio, conforme a Figura 29a e 29b, através do gráfico de linha do comportamento do MAPE médio por quantidade de lags nos modelos.

Na Figura 30, através do gráfico de coluna com MAPE médio por parâmetros, demonstra-se que os resultados da predição temporal dos parâmetros DBO, sólido, temperatura e pH, quando os modelos são configurados com 2 lag, a RL têm maior MAPE médio na predição do DBO com 11.08% e o RF tem maior MAPE médio na predição do sólido, temperatura e pH, respectivamente, 9.52%, 10.28% e 3.05%. Porém, as técnicas LSTM e MLP têm os melhores desempenhos de MAPE médio por parâmetro. Desse modo, o MLP tem o menor MAPE médio

na predição do sólido com 4.98% e o LSTM têm os menores MAPE médio nas predições dos parâmetros DBO, temperatura e pH, respectivamente, 2.33%, 8.08% e 2.89%. De outro modo, a Figura 31, através do gráfico de linha de comportamento dos valores dos parâmetros ao longo do tempo, em meses, demonstra esses desempenhos das técnicas por parâmetros, comparando as series temporais dos valores observados com os valores previstos dos parâmetros por cada técnica.

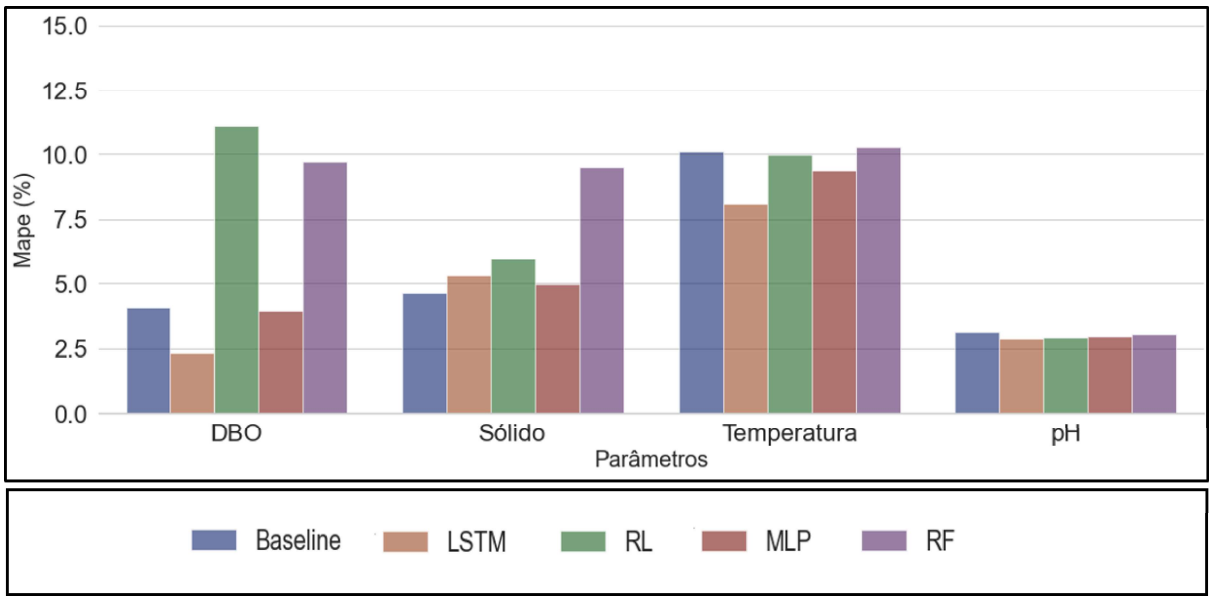
No desempenho por técnica, as técnicas regressão linear e random forest têm MAPE médio maior do que o baseline, por isso essas técnicas têm baixos desempenhos de predição. Todavia, o LSTM apresenta o melhor desempenho em relação as demais técnicas, pois tem o menor MAPE médio enquanto que o MLP apresenta o melhor desempenho em relação aos dados utilizados por ter o menor RMSE médio, conforme a Tabela 9.

Figura 29 – Desempenho de lag na predição temporal em a) Baseline e b) com as técnicas.



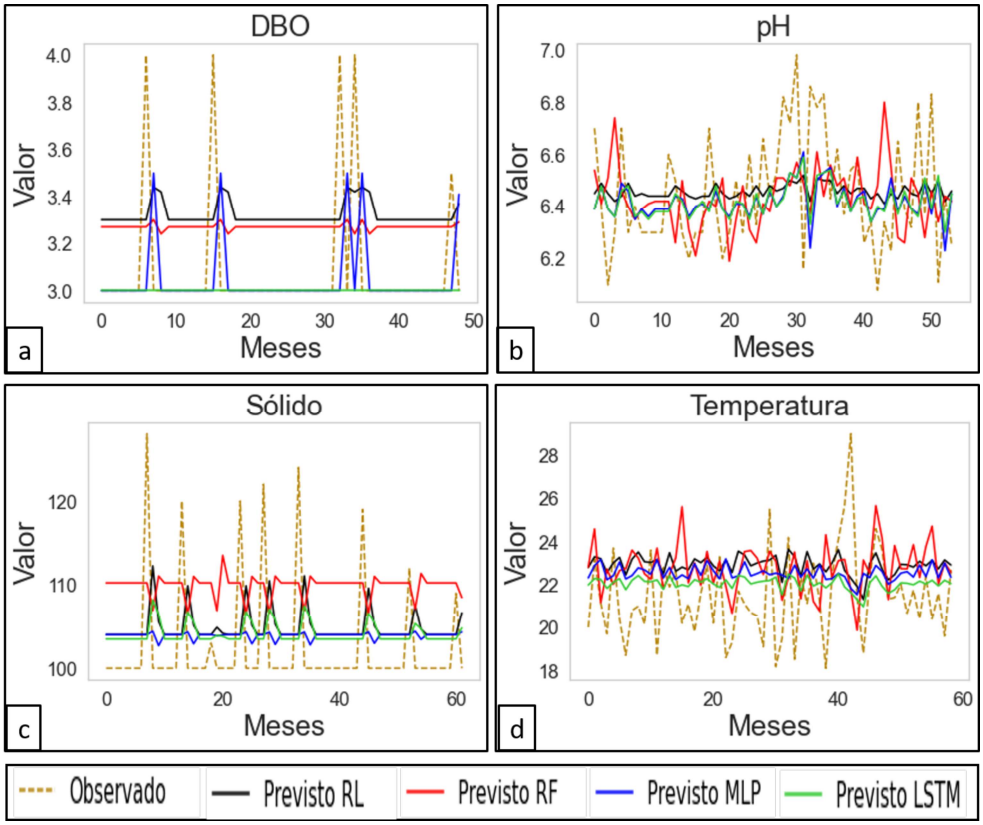
Fonte: Elaborada pelo autor.

Figura 30 – Desempenho das técnicas por parâmetros quando lag = 2.



Fonte: Elaborada pelo autor.

Figura 31 – Série temporal das predições temporal dos parâmetros com os dados de teste.



Fonte: Elaborada pelo autor.

4.2.2 Predição Espaço-Temporal

Na predição espaço-temporal com os dados dos pontos de coleta P03 e P08, a Figura 32, através do gráfico de coluna com MAPE médio por parâmetros, demonstra que dos 8 parâmetros preditos, somente temperatura e pH têm MAPE médio $<= 15\%$ em todas as técnicas. Sendo que, a Figura 33, através do gráfico de linhas de comportamento do MAPE médio por quantidade de lags do parâmetro temperatura, demonstra que todas as técnicas têm o desempenho de MAPE médio $<= 15\%$ somente a partir do 3º lag. Diferentemente disso, o MAPE médio dos parâmetros OD e turbidez estão acima de 100%, fósforo e DBO estão entre 50% e 100% e do sólido entre 16% e 50%. Quanto ao parâmetro coliformes, apresentou resultados indefinidos também chamados de infinitos no Python, devido à variância dos seus valores no ponto P08, de modo que, em comparação ao ponto P03, é ainda mais acentuada.

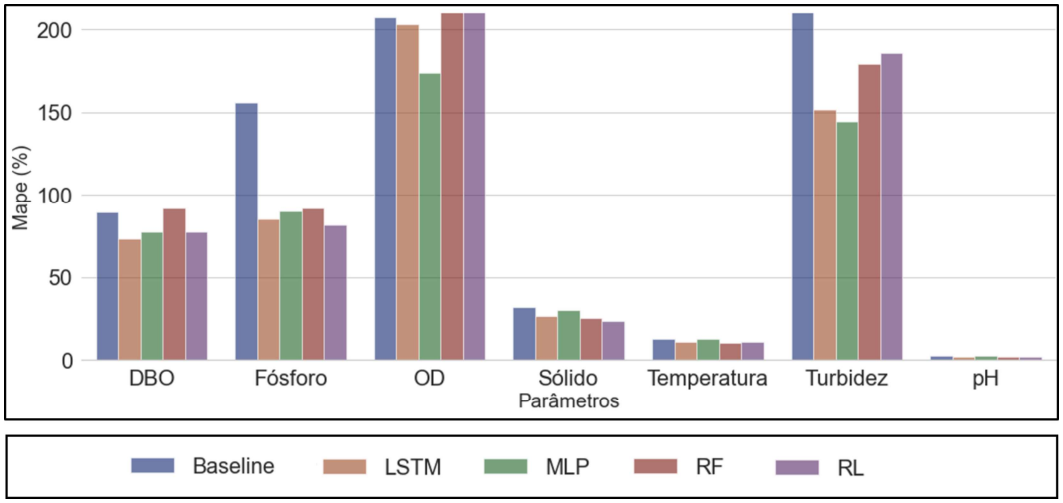
Diante disso, com as técnicas o menor MAPE médio é com 10 lags e com Baseline o menor MAPE médio é com 5 lags, conforme a Figura 34a e 34b, através do gráfico de linha do comportamento do MAPE médio por quantidade de lags nos modelos. Na Figura 35, através do gráfico de coluna com MAPE médio por parâmetros, demonstra-se que os resultados da predição espaço-temporal dos parâmetros temperatura e pH, quando os modelos são configurados com 10 lag, a RL têm os maiores MAPE médio na predição destes parâmetros com 10.67% e 2.07%, respectivamente. Porém, o MLP têm os melhores desempenhos de MAPE médio por

parâmetro, de modo que, os menores MAPE médio de temperatura e pH são 9.71% e 1.90%, respectivamente. De outro modo, a Figura 36, através do gráfico de linha do comportamento dos valores dos parâmetros ao longo do tempo, em meses, demonstra esses desempenhos das técnicas por parâmetros, comparando as séries temporais dos valores observados com os valores previstos dos parâmetros por cada técnica.

No desempenho por técnica, as técnicas de aprendizado de máquina têm MAPE médio menor do que o baseline, por isso esses desempenhos de predição são satisfatórios. Todavia, o MLP tem os melhores desempenho tanto em relação às demais técnicas quanto aos dados utilizados, pois têm os menores resultados médios de MAPE e RMSE, conforme a Tabela 9.

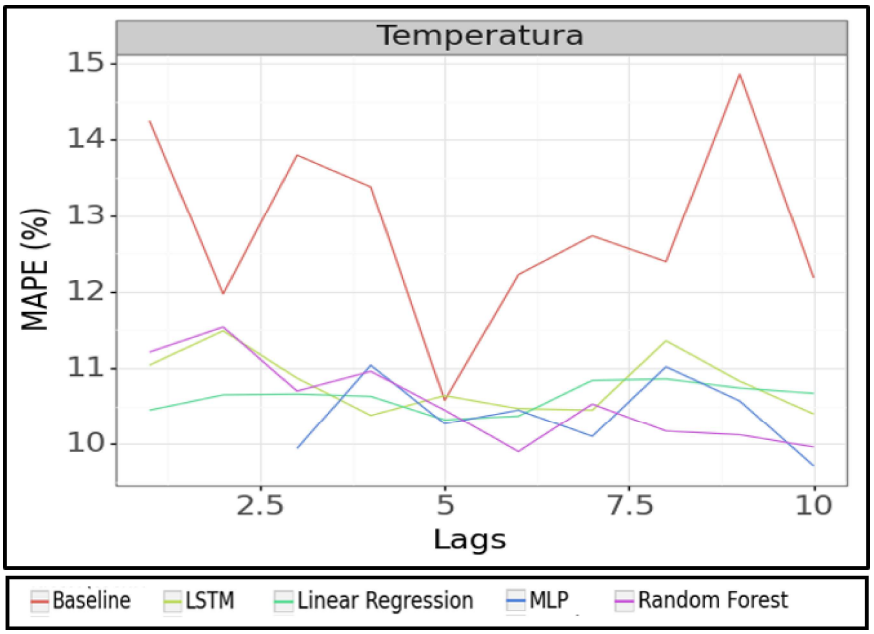
Por fim, nos resultados dos desempenhos observamos que as técnicas de redes neurais obtiveram os melhores desempenhos, pois são técnicas aptas a resolverem problemas não-lineares (SCHULTE, 2019). Ademais, apesar da literatura citar que a rede neural LSTM é adequada na predição de series temporais, porém notamos que a rede neural MLP apresenta os melhores desempenhos na predição temporal quando visualizamos os resultados da métrica de avaliação RMSE e na predição espaço-temporal tanto nas métricas MAPE e RMSE. Isto pode ser justificado por Chniti, Bakir e Zaher (2017), que demonstra nos resultados da sua pesquisa que os modelos LSTM obtêm os melhores desempenhos quando são multivariados na entrada, diferentemente das configurações dos experimentos nesta pesquisa que utilizaram apenas os valores de um parâmetro por vez na predição.

Figura 32 – Resultados da predição espaço-temporal dos parâmetros.



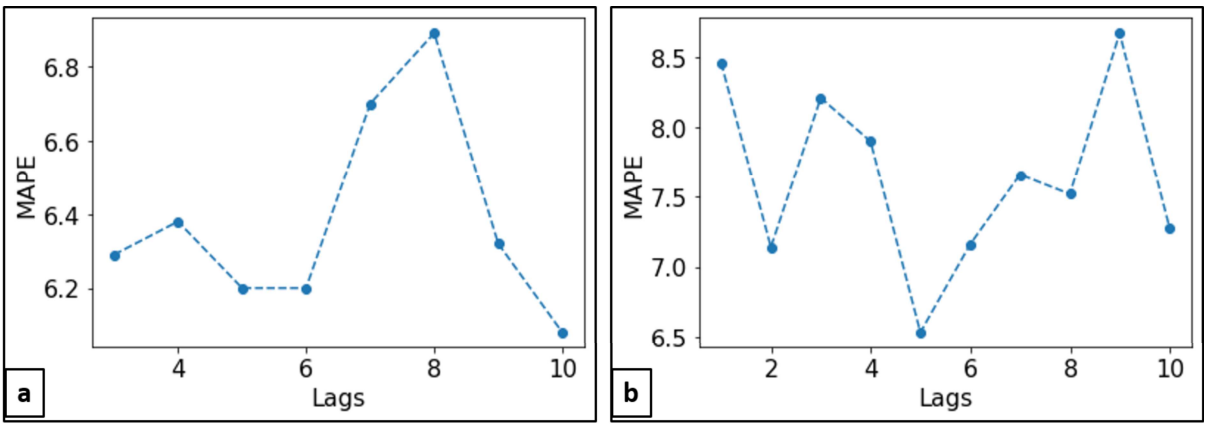
Fonte: Elaborada pelo autor.

Figura 33 – Predição espaço-temporal do parâmetro temperatura.



Fonte: Elaborada pelo autor.

Figura 34 – Desempenho de lag na predição espaço-temporal com a) as técnicas e b) Baseline.



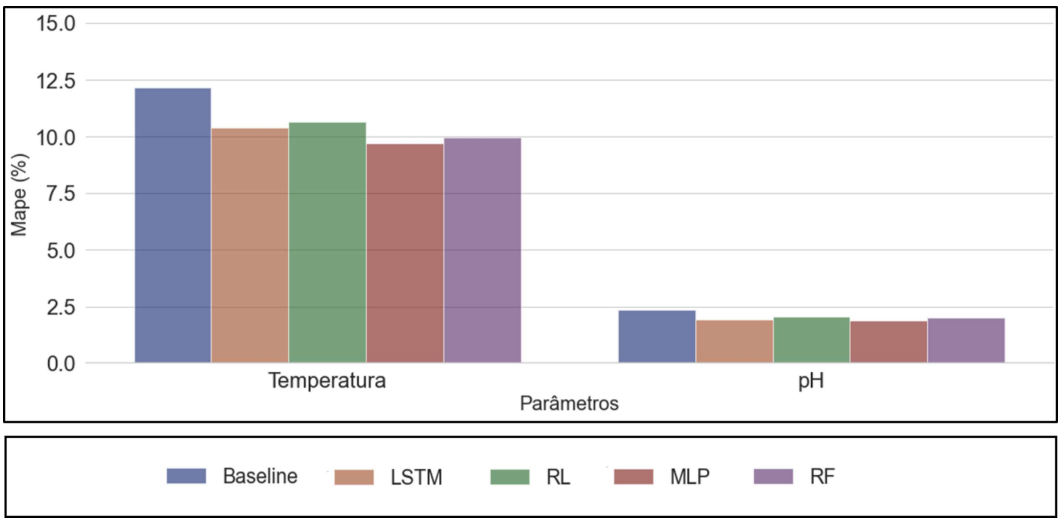
Fonte: Elaborada pelo autor.

Tabela 9 – Desempenho MAPE médio por técnica.

	Temporal		Espaço-temporal	
Técnicas	**MAPE(%)**	**RMSE**	**MAPE(%)**	**RMSE**
Baseline	5.49	3.48	6.53	1.62
LSTM	**4.66**	2.49	6.42	1.50
MLP	5.32	**2.47**	**5.94**	**1.34**
Random Forest	8.14	3.32	6.48	1.50
Regressão Linear	7.48	2.71	6.34	1.47

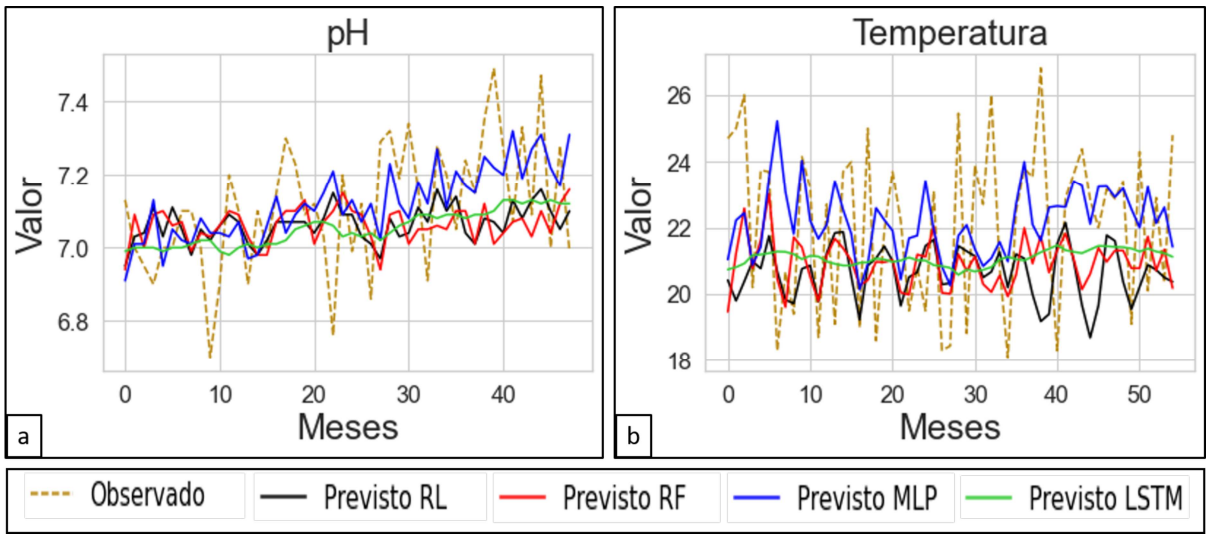
Fonte: Elaborada pelo autor.

Figura 35 – Desempenho das técnicas por parâmetro quando lag $= 10$.



Fonte: Elaborada pelo autor.

Figura 36 – Série temporal das predições espaço-temporal dos parâmetros com os dados de teste.



Fonte: Elaborada pelo autor.

5 CONCLUSÕES

Neste estudo, apesar das buscas por banco de dados com registros dos monitoramentos dos parâmetros da qualidade das águas superficiais localizadas no norte do Brasil, utilizamos os dados disponibilizados pela Companhia Ambiental do Estado de São Paulo (CETESB), devido à falta de padronização e a descontinuidade dos monitoramentos dos recursos hídricos na região norte.

Diante disso, através da análise exploratória dos dados, concluímos que os dados dos pontos de coleta P03 e P08 da UGRHI 06 - ALTO TIÊTE atendem os critérios estabelecidos na metodologia desta pesquisa, possibilitando a análise do desempenho dos algoritmos de aprendizado de máquina nos experimentos de predição temporal e espaço-temporal dos parâmetros coliformes, demanda bioquímica de oxigênio (DBO), fósforo, oxigênio dissolvido, potencial hidrogeniônico (PH), sólido, temperatura e turbidez .

Dessa forma, esses dados são caracterizados como assimétricos na distribuição dos valores, exceto do parâmetro pH. Apresenta, ainda, a ocorrência de *outliers*, exceto nos valores dos parâmetros fósforo no ponto P03 e DBO no ponto P08. Além disso, os valores do parâmetro coliformes está mais distribuído e esta variância nos dados pode ter produzido baixa acurácia nas predições temporal e espaço-temporal. Entretanto, os parâmetros pH, DBO, sólido e temperatura são uniformes nos seus valores, logo, servem para as predições.

Então, na predição temporal, com os dados do ponto de coleta P03, dos 8 parâmetros preditos, apenas os parâmetros DBO, Sólido Total, Temperatura da água e pH têm MAPE médio $<= 15\%$, devido a baixa variância nos seus valores, contrariamente, do que ocorre com os demais parâmetros. Além disso, os melhores desempenhos para este tipo de predição ocorrem quando os modelos são configurados com 2 lags, logo as técnicas LSTM e MLP têm os melhores resultados de MAPE médio nas predições destes parâmetros. No desempenho por técnica, o LSTM apresenta o melhor desempenho em relação as demais técnicas, pois tem o menor resultado de MAPE médio com 4.66%. Enquanto que o MLP apresenta o melhor desempenho em relação aos dados utilizados, pois tem o menor resultado de RMSE médio, com 2.47.

Porém, na predição espaço-temporal, com os dados dos pontos de coleta P03 e P08, dos 8 parâmetros preditos, somente temperatura e pH têm MAPE médio $<= 15\%$, sendo que a predição da temperatura apresenta esta métrica somente a partir do 3º lag. Entretanto, o parâmetro coliformes não apresentou resultado para este tipo de predição, devido à acentuada variância no ponto de coleta P08. Por conseguinte, os melhores desempenhos para este tipo de predição ocorrem quando os modelos são configurados com 10 lags, logo a técnica MLP têm os melhores desempenhos tanto em relação as demais técnicas quanto aos dados utilizados, pois possui os menores resultados médios de MAPE e RMSE, respectivamente 5.94% e 1.34.

Esses resultados demonstram que as técnicas de redes neurais obtiveram os melhores

desempenhos, pois são técnicas aptas a resolverem problemas não-lineares. Ademais, o LSTM obtêm os melhores desempenhos quando são multivariados na entrada, diferentemente das configurações dos experimentos nesta pesquisa que utilizaram apenas os valores de um parâmetro por vez na predição.

Portanto, os resultados deste estudo contribuem com o processo de monitoramento da qualidade da água e com a viabilização de implementação de um sistema de apoio à decisão para auxiliar na tomada de decisão dos gestores de recursos hídricos, de modo que a previsibilidade dos valores dos parâmetros da qualidade da água possibilite a elaboração de políticas públicas com objetivo de atendimento as legislações vigentes e, consequentemente, a sustentabilidade do recurso hídrico e o aumento da disponibilidade de água para o consumo das atuais e futuras gerações.

5.1 Limitações da Pesquisa

Apesar dos resultados, a pesquisa possui limitações quanto à quantidade e à qualidade dos dados disponíveis para este tipo de experimento.

5.2 Trabalhos Futuros

- Coleta de dados dos parâmetros da qualidade da água de forma ordenada e sequencial, de modo que novos estudos abordem a predição espacial dos parâmetros da qualidade da água e o aprofundamento do estudo da predição temporal e espaço-temporal com os parâmetros coliformes e fósforo.
- Utilização de outros algoritmos de aprendizado de máquina para predizer os parâmetros da qualidade da água.
- Implementação de um aplicativo para apoio a decisão dos gestores de recursos hídricos.

5.3 Produções

A seguir são listadas as produções, publicadas e submissões aceitas como consequência do desenvolvimento deste estudo.

5.3.1 Publicados

(ALMEIDA et al., 2020). Anderson Almeida, Marcos Amaris, Allan Veras e Bruno Merlin. Modelagem e Predição Temporal de Parâmetros de Qualidade de Água Usando Redes Neurais Profundas. *11º WCAMA – Workshop de Computação Aplicada à Gestão do Meio*

Ambiente e Recursos Naturais, evento satélite do XL Congresso da Sociedade Brasileira de Computação, Novembro 2020. 10 páginas.

(ALMEIDA et al., 2021). Anderson Almeida, Marcos Amaris, Allan Veras, Bruno Merlin e Adam Santos. Predição Temporal de Parâmetros de Qualidade de Água Usando Redes Neurais Profundas. *Brazilian Journal of Development*, Novembro 2021. 14 páginas.

5.3.2 Submetido e Aceito

(ALMEIDA; AMARIS; MERLIN, 2021). Anderson Almeida, Marcos Amaris e Bruno Merlin. Predição temporal e espaço-temporal dos parâmetros da qualidade da água. *Em: Memórias do 1º ERAMIA - NORTE2 – Escola Regional de Aprendizado de Máquina e Inteligência Artificial Norte 2, evento promovido pela Sociedade Brasileira de Computação*, Novembro 2021. 4 páginas.

REFERÊNCIAS

ACADEMY, K. **Como identificar outliers usando a regra 1,5xFIQ**. 2019. Disponível em: <https://pt.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/identifying-outliers-iqr-rule>.

ACTION, P. **Sazonalidade**. 2019. Disponível em: <http://www.portalaction.com.br/series-temporais/23-sazonalidade>.

ALIZADEH, M. J.; KAVIANPOUR, M. R. Development of wavelet-ann models to predict water quality parameters in hilo bay, pacific ocean. **Marine pollution bulletin**, Elsevier, v. 98, n. 1-2, p. 171–178, 2015.

ALMEIDA, A.; AMARIS, M.; MERLIN, B. Predição temporal e espaço-temporal de parâmetros de qualidade de água. In: SBC. **Em: Memórias do 1º ERAMIA - NORTE2 – Escola Regional de Aprendizado de Máquina e Inteligência Artificial Norte 2, evento promovido pela Sociedade Brasileira de Computação**. [S.l.], 2021.

ALMEIDA, A. et al. Modelagem e predição temporal de parâmetros de qualidade de água usando redes neurais profundas. In: SBC. **Anais do XI Workshop de Computação Aplicada à Gestão do Meio Ambiente e Recursos Naturais**. [S.l.], 2020. p. 121–130.

ALMEIDA, A. et al. Predição temporal de parâmetros de qualidade de água usando redes neurais profundas. In: BRAZILIAN JOURNAL OF DEVELOPMENT. **Em: Memórias do Brazilian Journal of Development**. [S.l.], 2021.

ALPAYDIN, E. **Introduction to Machine Learning. [Sl]**. [S.l.]: The MIT Press, 2010.

ANA. **Conjuntura dos recursos hídricos no Brasil 2019: informe anual**. [S.l.], 2019.

ANA. **Manual de usos consultivos da água no Brasil**. [S.l.], 2019.

ANA. **Monitoramento da qualidade da água em rios e reservatórios**. 2019. Disponível em: <http://capacitacao.ana.gov.br/conhecerh/bitstream/ana/76/6/Unidade_3.pdf>.

BANDARA, K.; BERGMEIR, C.; SMYL, S. Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach. **Expert Systems with Applications**, Elsevier, v. 140, p. 112896, 2020.

BENINI, F. A. V. Rede neural recorrente com perturbação simultânea aplicada no problema do caixeiro viajante. **São Carlos: Universidade de São Paulo**, 2008.

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

BRAGA, B.; PORTO, M.; TUCCI, C. E. Monitoramento de quantidade e qualidade das águas. **Águas doces no Brasil: Capital ecológico, uso e conservação**, p. 127 – 142, 2015.

BRANCO, S. M. et al. Água e saúde humana. **Águas doces no Brasil: Capital ecológico, uso e conservação**, p. 319 – 358, 2015.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, n. 1, p. 5–32, 2001.

BROCKWELL, P. J.; DAVIS, R. A. **Introduction to time series and forecasting**. [S.l.]: springer, 2002.

CAMILO, C. O.; SILVA, J. C. d. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFC)**, p. 1–29, 2009.

CAMPOS, R. J. Previsão de séries temporais com aplicações a séries de consumo de energia elétrica. Universidade Federal de Minas Gerais, 2008.

CAR, Z. et al. Modeling the spread of covid-19 infection using a multilayer perceptron. **Computational and mathematical methods in medicine**, Hindawi, v. 2020, 2020.

CETESB. **Infoáguas**. 2019. <https://sistemainfoaguas.cetesb.sp.gov.br/>. [Online; Dezembro 7 de 2019].

CHATFIELD, C. **The Analysis of Time Series: An Introduction**. [S.l.]: Chapman and Hall/CRC, 2013.

CHEN, K. et al. Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. **Water research**, Elsevier, v. 171, p. 115454, 2020.

CHNITI, G.; BAKIR, H.; ZAHER, H. E-commerce time series forecasting using lstm neural network and support vector regression. In: **Proceedings of the International Conference on Big Data and Internet of Thing**. [S.l.: s.n.], 2017. p. 80–84.

CORDOBA, E. B.; MARTINEZ, A. C.; FERRER, E. V. Water quality indicators: Comparison of a probabilistic index and a general quality index. the case of the confederación hidrográfica del júcar (spain). **Ecological Indicators**, Elsevier, v. 10, n. 5, p. 1049–1054, 2010.

CORTES, S. d. C.; PORCARO, R. M.; LIFSCHITZ, S. Mineração de dados: Funcionalidades, técnicas e abordagens. **PUC RIO Inf MCC**, 2002.

CRAMER, J. S. Mean and variance of r2 in small and moderate samples. **Journal of econometrics**, Elsevier, v. 35, n. 2-3, p. 253–266, 1987.

CRH. **Plano Estadual de Recursos Hídricos: PERH 2016-2019**. São Paulo:CRH, 2017.

DAMETTO, R. C. **Estudo da aplicação de redes neurais artificiais para predição de séries temporais financeiras**. Tese (Dissertação de mestrado) — UNIVERSIDADE ESTADUAL PAULISTA "JÚLIO DE MESQUITA FILHO", BAURU, 2018.

DERISIO, J. C. **Introdução ao controle de poluição ambiental**. [S.l.]: Oficina de Textos, 2016.

EHLERS, R. Análise de séries temporais, 2003. **Departamento de Estatística, Universidade Federal do Paraná**, 2012.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery: An overview. In: **Advances in Knowledge Discovery and Data Mining**. England: AAAI Press/The MIT Press, 1996. p. p.1–34.

FERNANDES, L. G. L. Utilização de redes neurais na análise e previsão de séries temporais. 1995.

FLECK, L. et al. Redes neurais artificiais: Princípios básicos. **Revista Eletrônica Científica Inovação e Tecnologia**, v. 1, n. 13, p. 47–57, 2016.

FRAVET, A. M. M. F. de; CRUZ, R. L. Qualidade da água utilizada para irrigação de hortaliças na região de botucatu-sp. **Irriga**, p. 144–155, 2007.

GASTALDINI, M. et al. Conceitos para a avaliação da qualidade da água. **PAIVA, JBD; PAIVA, EMCD Hidrologia aplicada à gestão de pequenas bacias hidrográficas. Porto Alegre: ABRH**, p. 428–51, 2001.

GASTALDINI, M.; TEIXEIRA, E. Avaliação da qualidade da água. **PAIVA, JBD; PAIVA, EMCD Hidrologia aplicada à gestão de pequenas bacias hidrográficas. Porto Alegre: ABRH**, p. 453–90, 2001.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining: um guia Prático**. [S.l.]: Elsevier, 2005.

GUJARATI, D. N. **Basic Econometrics**. fourth. [S.l.]: McGraw-Hill, 2003.

HAYKIN, S. **Redes neurais: princípios e prática**. [S.l.]: Bookman Editora, 2007.

HEDDAM, S. Simultaneous modelling and forecasting of hourly dissolved oxygen concentration (do) using radial basis function neural network (rbfnn) based approach: a case study from the klamath river, oregon, usa. **Modeling Earth Systems and Environment**, Springer, v. 2, n. 3, p. 135, 2016.

HEDDAM, S.; KISI, O. Modelling daily dissolved oxygen concentration using least square support vector machine, multivariate adaptive regression splines and m5 model tree. **Journal of Hydrology**, Elsevier, v. 559, p. 499–509, 2018.

HUAN, J.; CAO, W.; QIN, Y. Prediction of dissolved oxygen in aquaculture based on eemd and lssvm optimized by the bayesian evidence framework. **Computers and electronics in agriculture**, Elsevier, v. 150, p. 257–265, 2018.

LORENCIN, I. et al. Using multi-layer perceptron with laplacian edge detector for bladder cancer diagnosis. **Artificial Intelligence in Medicine**, Elsevier, v. 102, p. 101746, 2020.

MACEDO, D. C.; MATOS, S. N. Extração de conhecimento através da mineração de dados. **Revista de Engenharia e Tecnologia**, v. 2, n. 2, p. Páginas–22, 2010.

MAROTTA, H.; SANTOS, R. O. d.; ENRICH-PRAST, A. Monitoramento limnológico: um instrumento para a conservação dos recursos hídricos no planejamento e na gestão urbano-ambientais. **Ambiente & sociedade**, SciELO Brasil, v. 11, n. 1, p. 67–79, 2008.

MATSUMOTO, D. K. F. et al. Estudo em séries temporais financeiras utilizando redes neurais recorrentes. Universidade Federal de Alagoas, 2019.

MENEZES, J. P. C. et al. Relação entre padrões de uso e ocupação do solo e qualidade da água em uma bacia hidrográfica urbana. **Engenharia Sanitária e Ambiental**, SciELO Brasil, v. 21, p. 519–534, 2016.

MOHAN, S.; KUMAR, K. P. Waste load allocation using machine scheduling: model application. **Environmental Processes**, Springer, v. 3, n. 1, p. 139–151, 2016.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. **Sistemas inteligentes-Fundamentos e aplicações**, v. 1, n. 1, p. 1, 2003.

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

MORETTIN, P. A.; TOLOI, C. M. **Análise de séries temporais: modelos lineares univariados**. [S.l.]: Editora Blucher, 2018.

NAIR, V.; HINTON, G. E. Rectified linear units improve restricted boltzmann machines. In: **Icml**. [S.l.: s.n.], 2010.

NARCISO, G. A. A.; JASKIU, I. F. Compostos orgânicos e a contaminação da água: Descarte inadequado do óleo de cozinha usado. **Enaproc**, v. 1, n. 1, 2019.

OLAH, C. Understanding lstm networks, 2015. **URL http://colah. github. io/posts/2015-08-Understanding-LSTMs**, 2015.

OLIVEIRA, B. **Características das séries temporais**. 2019. Disponível em: <https://www.abgconsultoria.com.br/blog/caracteristicas-das-series-temporais/>.

OLIVEIRA, S. et al. Modeling spatial patterns of fire occurrence in mediterranean europe using multiple regression and random forest. **Forest Ecology and Management**, Elsevier, v. 275, p. 117–129, 2012.

OLIVEIRA, T. P. et al. Predição de tráfego, usando redes neurais artificiais, para gerenciamento adaptativo de largura de banda em roteadores. Universidade Federal de Uberlândia, 2014.

OTT, E. F.; SARTORI, A. Predição de peso na criação de frango de corte utilizando redes neurais artificiais. 2018.

PALA, L. O. d. O. Revisitando a estimação de coeficiente de determinação. Universidade Federal de Alfenas, 2019.

PARMEZAN, A. R. S. **Predição de séries temporais por similaridade**. Tese (Doutorado) — Universidade de São Paulo, 2014.

RASCHKA, S.; MIRJALILI, V. Python machine learning: Machine learning and deep learning with python. **Scikit-Learn, and TensorFlow. Second edition ed**, 2017.

RÊGO, A. d. S. et al. Predição de parâmetros de qualidade do biodiesel utilizando redes neurais artificiais. Universidade Federal do Maranhão, 2013.

ROY, M.-H.; LAROCQUE, D. Robustness of random forests for regression. **Journal of Nonparametric Statistics**, Taylor & Francis, v. 24, n. 4, p. 993–1006, 2012.

RUSSEL, S.; NORVIG, P. Inteligência artificial. 3a edição. **Editora Campus**, 2013.

SARKAR, A.; PANDEY, P. River water quality modelling using artificial neural network technique. **Aquatic Procedia**, Elsevier, v. 4, p. 1070–1077, 2015.

SCHULTE, L. G. Suporte à decisão em pastagens: Análise espaço-temporal e aprendizado de máquina para predição da disponibilidade de forragem no contexto de smart farming. Universidade Federal do Pampa, 2019.

SCIKIT-LEARN. 2020. Disponível em: <https://scikit-learn.org/>.

SEMA. **CETESB: 50 anos de História e estórias**. São Paulo:SEMA, 2018.

SHI, P. et al. Prediction of dissolved oxygen content in aquaculture using clustering-based softplus extreme learning machine. **Computers and Electronics in Agriculture**, Elsevier, v. 157, p. 329–338, 2019.

SHIRAISHI, D.; MARUJO, E. Predição de sepse em unidade de terapia intensiva: uma abordagem de aprendizado de máquina. In: SBC. **Anais do XX Simpósio Brasileiro de Computação Aplicada à Saúde**. [S.l.], 2020. p. 416–421.

SILVA, G. A.; KULAY, L. A. água na indústria. **Águas doces no Brasil: Capital ecológico, uso e conservação**, p. 319 – 358, 2015.

SILVA, I. N. D.; SPATTI, D. H.; FLAUZINO, R. A. **Redes NEurais Artificiais para Engenharia e Ciências Aplicadas: Curso Prático**. [S.l.]: Artliber, 2010.

SILVA, L. A. da; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados: com aplicações em R**. [S.l.]: Elsevier Brasil, 2017.

SIMONETTI, V. C. Correlação espacial e sazonal de parâmetros indicadores de qualidade da água da bacia hidrográfica do alto sorocaba associadas ao uso do solo. **Revista Brasileira de Ciência do Solo**, Universidade Estadual Paulista (UNESP), 2018.

SOLANKI, A.; AGRAWAL, H.; KHARE, K. Predictive analysis of water quality parameters using deep learning. **International Journal of Computer Applications**, Foundation of Computer Science, v. 125, n. 9, p. 0975–8887, 2015.

SOUZA, J. S. de. **Eutrofização**. 2019. Disponível em: <https://www.infoescola.com/ecologia/eutrofizacao/>.

TELLES, D. D.; DOMINGUES, A. F. água na agricultura e pecuária. **Águas doces no Brasil: Capital ecológico, uso e conservação**, p. 319 – 358, 2015.

VASCONCELOS, V. d. M. M.; SOUZA, C. F. Caracterização dos parâmetros de qualidade da água do manancial utinga, belém, pa, brasil/characterization of water quality parameters of the reservoir utinga, belém, pa, brazil. **Revista Ambiente & Água**, Instituto de Pesquisas Ambientais em Bacias Hidrográficas, v. 6, n. 2, p. 305, 2011.

WANDERLEY, H. S.; AMORIM, R. F. C. d.; CARVALHO, F. O. d. Interpolação espacial de dados médios mensais pluviométricos com redes neurais artificiais. **Revista Brasileira de Meteorologia**, SciELO Brasil, v. 29, p. 389–396, 2014.

WATER, D. **Parâmetro físico de qualidade: turbidez da água**. 2019. Disponível em: <https://www.digitalwater.com.br/parametro-fisico-de-qualidade-turbidez-da-agua/>.

ZHOU, X. et al. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. **The Crop Journal**, Elsevier, v. 4, n. 3, p. 212–219, 2016.