INSTITUTO LATINO-AMERICANO DE INFRAESTRUTUA E TERRITÓRIO (ILATIT)

ENGENHARIA QUÍMICA

APRENDIZADO DE MÁQUINA PARA A SIMULAÇÃO DE OXIGÊNIO DISSOLVIDO DA ÁGUA SUPERFICIAL DA REPRESA BILLINGS

HAIDEN ARLEY ARDILA JOVÉL

Foz do Iguaçu - PR
2024

HAIDEN ARLEY ARDILA JOVÉL

APRENDIZADO DE MÁQUINA PARA A SIMULAÇÃO DE OXIGÊNIO DISSOLVIDO DA ÁGUA SUPERFICIAL DA REPRESA BILLINGS

Trabalho de Conclusão de Curso apresentado ao Instituto Latino-Americano de Tecnologia, Infraestrutura, e Território da Universidade Federal da Integração Latino-Americana, como requisito parcial à obtenção do título de Bacharel em Engenharia Química.

Orientadora: Prof. Dra Marlei Roling Scariot
Coorientador: Luís Antonio Lourenço

Foz do Iguaçu - PR
2024

HAIDEN ARLEY ARDILA JOVÉL

APRENDIZADO DE MÁQUINA PARA A SIMULAÇÃO DE OXIGÊNIO DISSOLVIDO DA ÁGUA SUPERFICIAL DA REPRESA BILLINGS

Trabalho de Conclusão de Curso apresentado ao Instituto Latino-Americano de Tecnologia, Infraestrutura, e Território da Universidade Federal da Integração Latino-Americana, como requisito parcial à obtenção do título de Bacharel em Engenharia Química.

BANCA EXAMINADORA

_____

Orientadora: Prof. Dra. Marlei Roling Scariot
UNILA

_____

Coorientador: Dr. Luís Antonio Lourenço
UFSC

_____

Prof. Dra. Leila Roling Scariot da Silva
IF GOIANO

Foz do Iguaçu, 19 de abril de 2024

Dedico este trabalho a minha família e amigos, os quais foram pilares fundamentais neste processo.

AGRADECIMENTOS

Primeiramente, agradeço a Deus por me permitir cumprir minha meta, de me tornar Engenheiro Químico, foram inumeráveis os desafios que abrangi longe da minha família, num país desconhecido, com uma cultura totalmente nova, uma linguagem desconhecida, e um milhão de experiências, emoções, sensações sobre tudo conhecimento aprendido de pessoas maravilhosas.

A minha mãe, Claudia Jovél, a mulher que mais amo nesse mundo, pessoa pela qual tive a motivação de chegar até o final, quem me brindou seu apoio incondicional, e quem sempre soube me encaminhar para ser a pessoa de bem quem eu sou.

A meu pai Miguel Ardila, meus irmãos Edward Jovél e Gina Jovél, meus avós Freddy Jovél e Maria Elba Escobar, meu tio Danny Jovél, minha cunhada Julieth López, minha mãe putativa Matty Bonilla, e ao meu pai de coração Efraín Caballero, conselheiros que encontrei, os quais sempre me apoiaram na distância.

A minha orientadora, a professora Marlei Roling Scariot, quem forjou em mim todas as ferramentas para ser um bom profissional, me inspirou em escrever este trabalho, me apoiou inúmeras vezes e sempre aportou amplamente na minha formação pessoal e sobre tudo profissional.

Ao Professor Luís Antonio Lourenço, quem me inspirou na docência e me direcionou da melhor forma ao longo da minha trajetória na graduação me transmitindo da melhor forma possível os conhecimentos que me permitiram consolidar este trabalho de conclusão de curso.

À professora Aline Toci e Carlos Bauer, que me permitiram me adentrar e iniciar no maravilhoso mundo da docência e a iniciação científica.

À família Fretes Romero e à família Ruíz Díaz, em especial a minhas mães de coração Antolina Romero e Zoraida Barboza, quem sempre me acolheram como um filho durante todos estes anos e me fizeram sentir como um membro mais família.

Ao Fernando Peralta, quem me acompanhou neste processo durante todos estes anos, esteve sempre comigo nos momentos fáceis e difíceis, e quem me ensinou a amar incondicionalmente.

À Alba Teresa Aguilera por todos os conselhos e me brindar sua amizade sincera. À Gissella Aguilera por sempre me apoiar na distância e em momentos que precisei. À Noelia Sandra por sempre estar presente quando precisei de conselhos de vida e xicaras

de realidade, à Maria Elena Zabala amiga de sempre, à Bertha Montenegro por todos os momentos felizes e nossas aventuras no Subway, à Diana Britez por estar sempre para mim.

Ao “Team Roberto Carlos”, Edgar e Fiorella, amigos que sempre levarei no meu coração.

Às melhores amigas de faculdade I Lien e Kelly Borne, quem sempre lembrarei com carinho, admiração e respeito.

À Ghessyca Bonfim, Marcia Pereira, Leticia Lisik, Karla Andrade e Lourdes Ayala por ser dessas amigas que sei que sempre poderei contar, independente da situação e/ou momento.

Ao Carlos Flecha, meu psicólogo não certificado, ao Carmelo García e Matias Fretes, por ser amigos muito especiais, companheiros de tertúlias e tererê.

À Francieli Mrozcka, Karen Luciano, Jessica Machado e Adriana Pires, amigas maravilhosas, as quais cada dia me ensinam mais sobre o “mundo real” e me ajudam a crescer como pessoa tanto profissionalmente quanto pessoalmente.

Finalmente, à Camila Muller Minato, minha primeira líder, quem nunca irei esquecer todos os aprendizados nos meus primeiros passos como profissional, pessoa quem sempre lembrarei com todo meu carinho e respeito.

“Encuentra lo que amas y deja que te mate”
Charles Bukowski

RESUMO

Na atualidade, são muitas as tecnologias implementadas com o objetivo de mitigar os impactos ambientais que ocorrem no mundo, em vista disso, diversas áreas do conhecimento têm agido de forma interdisciplinar com a finalidade de aplicar técnicas e/ou ferramentas que permitam melhorar os processos tradicionais tornando-os mais eficientes. A ciência de dados é uma disciplina com um campo abrangente de atuação nas Engenharias permitindo a otimização de processos rotineiros de forma mais simples utilizando modelos matemáticos e um amplo conjunto de dados. Órgãos públicos como a SABESP realizam monitoramento rotineiro dos parâmetros de qualidade da água da represa Billings disponibilizando um relatório com os resultados no decorrer de um ano, além disso a NASA realiza monitoramento por satélite de diversos parâmetros atmosféricos. O presente trabalho teve como objetivo a utilização de modelos de aprendizado de máquina para simular a quantidade de oxigênio dissolvido (OD) no braço Taquacetuba (BL-105) da represa Billings, os resultados apontam que devido à natureza não linear das variáveis que compõem o sistema natural, o modelo floresta aleatória (*Random Forest*) consegue ajustes e predições melhores do que o modelo de regressão linear. Foi possível a obtenção de predições coerentes com a diminuição significativa das variáveis da SABESP para os dois modelos. O modelo floresta aleatória conseguiu diferençar a disponibilidade de OD no meio hidrográfico levando em consideração as especificidades sazonais além dos períodos diurnos e noturnos ao longo de dois anos. O ano 2017 apresentou valores mais precisos do que o ano 2019, quando comparado com os dados reais. Por fim, a pesquisa sugere a possibilidade de tornar o processo de monitoramento da qualidade da água mais econômico, devido à capacidade do modelo de aprendizado de máquina em prever resultados satisfatórios com uma quantidade reduzida de parâmetros de monitoramento do complexo hidrográfico Billings.

Palavras-chave: modelos de predição; oxigênio dissolvido; monitoramento ambiental; qualidade da água; aprendizado de máquina.

RESUMEN

En la actualidad, son numerosas las tecnologías implementadas con el objetivo de mitigar los impactos ambientales que ocurren en el mundo; por tanto, diversas áreas del conocimiento han actuado de forma interdisciplinaria con la finalidad de aplicar técnicas y/o herramientas que permitan mejorar los procesos tradicionales haciéndolos más eficientes. La ciencia de datos es una disciplina con un campo amplio de acción en las ingenierías, permitiendo la optimización de procesos rutinarios de manera más sencilla utilizando modelos matemáticos y un amplio conjunto de datos. Organismos públicos como la SABESP realizan monitoreo rutinario de los parámetros de calidad del agua de la represa Billings, ofreciendo un informe con los resultados a lo largo de un año; además, la NASA lleva a cabo un monitoreo por satélite de diversos parámetros atmosféricos. El presente trabajo tuvo como objetivo la utilización de modelos de aprendizaje de máquina para simular la cantidad de oxígeno disuelto (OD) en el brazo Taquacetuba (BL-105) de la represa Billings; los resultados señalan que debido a la naturaleza no lineal de las variables que componen el sistema natural, el modelo de bosque aleatorio (*Random Florest*) logra ajustes y predicciones mejores que el modelo de regresión lineal. Se pudo obtener predicciones coherentes con la disminución significativa de las variables de SABESP para los dos modelos. El modelo de bosque aleatorio logró diferenciar la disponibilidad de OD en el medio hidrográfico teniendo en cuenta las especificidades estacionales además de los periodos diurnos y nocturnos a lo largo de dos años. El año 2017 presentó valores más precisos que el año 2019 al compararlos con los datos reales. Por último, la investigación sugiere la posibilidad de hacer el proceso de monitoreo de la calidad del agua más económico debido a la capacidad del modelo de aprendizaje automático para prever resultados satisfactorios con una cantidad reducida de parámetros de monitoreo del complejo hidrográfico Billings.

Palabras clave: modelos de predicción, oxígeno disuelto, monitoreo ambiental, calidad del agua, aprendizaje automático.

ABSTRACT

Currently, there are many technologies implemented with the aim of mitigating environmental impacts occurring worldwide; as a result, various areas of knowledge have been acting in an interdisciplinary manner to apply techniques and/or tools that improve traditional processes, making them more efficient. Data science is a discipline with a broad field of action in Engineering, allowing for the optimization of routine processes in a simpler way using mathematical models and a wide range of data. Public organizations such as SABESP carry out routine monitoring of water quality parameters in the Billings reservoir, providing a report with results over the course of a year; additionally, NASA conducts satellite monitoring of various atmospheric parameters. This study aimed to use machine learning models to simulate the amount of dissolved oxygen (DO) in the Taquacetuba arm (BL-105) of the Billings reservoir; the results indicate that due to the nonlinear nature of the variables that compose the natural system, the Random Forest model achieves better adjustments and predictions than the linear regression model. Coherent predictions were obtained with a significant decrease in SABESP variables for both models. The random forest model was able to differentiate the availability of DO in the hydrographic environment, taking into account seasonal specificities as well as daytime and nighttime periods over two years. The year 2017 showed more accurate values than 2019 when compared to real data. Finally, the research suggests the possibility of making the water quality monitoring process more economical, due to the machine learning model's ability to predict satisfactory results with a reduced number of monitoring parameters for the Billings hydrographic complex.

Key words: prediction model, dissolved oxygen, environmental monitoring, water quality, machine learning.

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

RMSP	Região Metropolitana de São Paulo
OD	Oxigênio Dissolvido
ORP	Potencial de Oxidorredução
N_2	Nitrogênio
CH_4	Metano
NTU	Unidade Nefelométrica de Turbidez
pH	Potencial de Hidrogênio
ppm	Partes por milhão
CO_2	Dióxido de Carbono
$H_2CO_3^-$	Ácido Carbônico
O_2	Oxigênio
SiO_2	Dióxido de Silício
MAE	Mean Absolute Erro
MSE	Mean Square Erro
RMSE	Root Mean Square Erro
R^2	Coeficiente de Determinação
W	Watts
m	Metro
L	Litro
s	Segundo
mg	Miligrama
V	Volts
RF	Random Forest

LISTA DE SIGLAS

NASA	National Aeronautics and Space Administration
SABESP	Companhia de Saneamento Básico do Estado de São Paulo
OECD	Organization for Economic Cooperation and Development
CETESB	Companhia Ambiental do Estado de São Paulo
OMS	Organização Mundial da Saúde
PMA	Programa Metropolitano de Água

LISTA DE FIGURAS

SUMÁRIO

1. INTRODUÇÃO

O estudo dos recursos naturais, suas propriedades e características tem sido um destaque na atualidade para diversos campos do conhecimento que pretendem desenvolver tecnologia com a finalidade de garantir e preservar o meio ambiente. A qualidade da água é um assunto de ampla e fundamental importância devido a sua versatilidade, sua indispensabilidade e sua utilidade. Nesse sentido, tendo em conta o panorama atual, o qual está caracterizado pela escassez deste recurso valioso, tem-se tornado indispensável realizar estudos que permitam estudar mais fundo seus parâmetros e a relação que estes têm com a saúde humana, animal e sobre tudo seu impacto no meio ambiente de forma generalizada.

A represa Billings, localizada na Região Metropolitana de São Paulo (RMSP) a uma altitude de 746,5 m e cobrindo uma área de 127mi^2, foi construída com o intuito de gerar energia hidroelétrica a partir das bacias hidrográficas da região de Guarapiranga e Serra do Mar. Entretanto, embora seja considerada o maior reservatório de água da RMSP, na atualidade se encontra preenchida com águas altamente poluídas que são recebidas dos esgotos das cidades vizinhas (ALVIM et al., 2022).

A problemática ambiental que apresenta a Represa Billings é grave, devido ao alto grau de eutrofização na água, ocasionada pelo lançamento de matéria orgânica por parte da população localizada ao redor das bacias hídricas que preenchem a represa (CARDOSO-SILVA et al., 2014). Em consequência disso, no decorrer dos anos tem sido percebida uma degradação gradativa na qualidade da água de forma generalizada, a qual é prejudicial para a saúde humana, animal e diretamente para o meio ambiente.

A proliferação de algas cianofíceas representa uma preocupação significativa para a saúde pública devido à sua natureza de difícil remoção. Além disso, os corpos d'água que contêm essas algas demandam tratamentos dispendiosos e frequentemente inviáveis. Por esse motivo, as condições de qualidade da água proveniente da represa Billings têm gerado preocupações de cunho socioeconômico. Dificuldades relacionadas ao tratamento dessas algas tornam complexa a tarefa de mitigar os danos tanto à saúde quanto ao meio ambiente quando essas águas são utilizadas para abastecimento público (FERRARA, 2018).

Devido aos altos índices de contaminação atuais nos recursos hídricos e mananciais, tornou-se necessário o desenvolvimento e implementação de tecnologias que

permitam o monitoramento dos sistemas hídricos (HANISCH; SOUZA FREIRE-NORDI, 2015). Nesse contexto, o emprego de modelos de simulação tem desempenhado um papel de extrema relevância na análise de parâmetros tanto qualitativos quanto quantitativos da qualidade da água. Isto, devido ao fato que permite a previsão tendencial das características e padrões que os ecossistemas aquáticos poderão apresentar em períodos específicos.

A modelagem e consequente simulação de modelos matemáticos que descrevem sistemas naturais são uma ferramenta indispensável na contemporaneidade, pois tratam-se de disciplinas científicas transversais que permitem uma maior compreensão da dinâmica dos elementos dos sistemas naturais (atmosférico, aquático, terrestre, etc.) e suas interações com os diversos parâmetros aos quais se encontram expostos (ALDANA et al., 2017). Porém, o fato mais interessante e importante do estudo e aplicação destas disciplinas está relacionado à possibilidade de prever cenários futuros e obter dados que podem ser analisados de diversas formas, permitindo caracterizar um determinado ecossistema, e em base a suas condições e/ou previsão, propor métodos alternativos para evitar sua degradação ou garantir sua preservação, mantendo seus padrões e visando a melhora da sua qualidade como um todo.

Nos anos recentes, no âmbito contemporâneo e em consonância com o que foi previamente exposto, a ciência de dados, também referida como mineração de dados, tem manifestado um rápido destaque em diversos domínios do conhecimento. Isso engloba a Engenharia e suas variadas disciplinas e subáreas, incluindo a Engenharia Química e a Engenharia Ecológica (LOURENÇO, 2021). A amplificação e aplicação da ciência de dados nesse campo emergiram como resposta ao imperativo e à necessidade de instaurar uma estratégia para otimização de processos por meio da utilização de um repositório de dados, o qual, em conjunção com modelos matemáticos e algoritmos, capacita a projeção de eventos por meio de aprendizado de máquina.

A generalização entre as relações de entrada (fatores) e as varáveis de saída (respostas) permite uma maior versatilidade devido a sua natureza de inferência indutiva, o qual é muito importante devido à ampla abrangência de sistemas que podem ser avaliados, assim como à possibilidade de determinar a relação intrínseca entre parâmetros físicos e químicos a partir de um determinado banco de dados, o que pode ter um resultado muito interessante no estudo de sistemas naturais, como é o caso do complexo Billings.

O *software* “Orange Data Mining” permite realizar todos os passos de mineração de dados desde a visualização e aplicação de métodos de aprendizado de máquina, tornando-o ideal para realizar a simulação do complexo Billings, tendo em conta que algumas instituições como a SABESP e a NASA, fornecem um grande volume de dados dos parâmetros de qualidade da água para alimentar e/ou treinar o algoritmo, possibilitando a simulação de variáveis da qualidade da água da represa.

Nesse sentido, o presente estudo pretende utilizar os métodos de aprendizado de máquina utilizando o *software* Orange Data Mining. O modelo será treinado com bancos de dados fornecidos por distintas entidades/instituições que realizam monitoramento das condições da água da represa Billings e da região de forma geral.

O intuito da aplicação é a criação de um modelo de inteligência artificial que consiga prever o comportamento dinâmico do oxigênio dissolvido da água superficial do complexo Billings.

Nessa predição de variáveis, é viável analisar as inter-relações existentes entre diferentes parâmetros (pH, temperatura, radiação solar, turbidez, condutividade, velocidade do vento, potencial de oxirredução radiação fotossintética, precipitação, etc.) e a variável alvo (oxigênio dissolvido), levando em consideração as condições da represa, e as dinâmicas temporais/sazonais correspondentes a esses parâmetros.

Isso abre oportunidades para explorar a interdisciplinaridade entre a engenharia ecológica e o conhecimento da engenharia química, em conjunto com a aplicação de técnicas de mineração de dados. Sendo o principal propósito aperfeiçoar a capacidade de predizer o comportamento da qualidade da água na represa Billings, por meio da simulação de variáveis cruciais, além de tornar possível antecipar cenários futuros. Isso tem como meta a conservação e o aprimoramento da qualidade dos recursos naturais e do ambiente como um todo.

2 OBJETIVOS

2.1. OBJETIVO GERAL

Utilizar a técnica de aprendizado de máquina para predizer o comportamento dinâmico da variável oxigênio dissolvido da água superficial da represa Billings.

2.2. OBJETIVOS ESPECÍFICOS

- Levantar o banco de dados fornecido pela SABESP, do complexo Billings, e adaptá-lo ao *software* “Orange Data Mining”;
- Levantar o banco de dados fornecido pela NASA, do complexo Billings, e adaptá-lo ao *software* “Orange Data Mining”;
- Escolher e treinar o(s) modelo(s) de aprendizado de máquina a partir dos bancos de dados do complexo Billings no *software* “Orange Data Mining”;
- Analisar as predições geradas pelo *software* “Orange Data Mining” em relação aos parâmetros carregados no modelo;
- Identificar as relações que existem entre os parâmetros no complexo Billings a partir das predições geradas no *software* “Orange Data Mining”;
- Descobrir a dinâmica do modelo de aprendizado de máquina em relação aos parâmetros do banco de dados do complexo Billings;
- Caracterizar os parâmetros gerados pelo modelo de aprendizado de máquina a partir do banco de dados do complexo Billings para explicar o comportamento ou sua tendência em função do tempo.

3 JUSTIFICATIVA

Dentre as problemáticas ambientais atuais da humanidade, a poluição das águas é uma das mais graves e prejudiciais a longo prazo. Apesar do desenvolvimento tecnológico e das técnicas inovadoras que visam mitigar os danos causados pela intervenção humana, existem muitos desafios a serem enfrentados para garantir o abastecimento de água com qualidade e a manutenção de ecossistemas equilibrados. Por esse motivo, a água, ao ser um recurso hídrico indispensável e limitado para a biosfera, torna-se um objeto de estudo de extrema importância, inclusive desde o ponto de vista da engenharia.

A represa Billings é muito conhecida e importante na Região Metropolitana de São Paulo (RMSP). Apesar de não ter sido construída com a finalidade de abastecimento, ela vem atendendo à crescente demanda de água ocasionada pelo crescimento populacional e industrial, principalmente em épocas secas ou caracterizadas por altas temperaturas (CARDOSO-SILVA et al., 2014). Analogamente, a reversão das águas do rio Pinheiros e o lançamento de efluentes industriais e domésticos foi mais frequente nas bacias que compõem este complexo, o que levou a constatar um alto índice de poluição do sistema aquático e, portanto, uma perda generalizada da qualidade da água da represa, que também era utilizada para a geração de energia (HERREIRA; PIZELLA, 2020).

A mineração de dados nos últimos anos tem sido bastante empregada em diversas disciplinas, e em especial na engenharia ecológica e suas áreas transversais (DA COSTA FILHO et al., 2019). Isto é devido a que a ciência de dados pode ser aplicada de diversas formas, como é o caso da simulação, onde a finalidade é reproduzir o comportamento futuro das variáveis. Devido à necessidade de monitoramento constante dos parâmetros de qualidade de fontes hidrográficas, tem-se tornando comum a criação de banco de dados, os quais resultam em ferramentas de alto interesse no ramo em questão, devido ao fato, que a partir deles é possível fazer uma análise dos possíveis panoramas que podem ser enfrentados no decorrer do tempo, quando aplicado um modelo de aprendizado de máquina.

Desde o ponto de vista da simulação de sistemas naturais, a possibilidade de reprodução de variáveis da qualidade da água a partir de modelos de aprendizado de máquina torna-se uma ferramenta muito eficaz, pois, a partir das inferências geradas pelo algoritmo é possível avaliar o comportamento dinâmico de variáveis importantes, interpretar gráficos e estabelecer relações entre parâmetros de forma simples e com dinâmicas caracterizadas por uma alta precisão em relação ao sistema natural estudado em questão.

Nesse sentido, a aplicação de modelos de aprendizado de máquina no complexo Billings torna-se uma estratégia interdisciplinar que envolve várias áreas do conhecimento que podem ser aplicadas como objeto de estudo na Engenharia Química, e que permite ter uma visão mais estratégica e estruturada da aplicação da modelagem e simulação em sistemas naturais, e o papel do engenheiro na análise e otimização de processos que podem precisar ou não ajustes e/ou ou intervenções para sua melhora de forma geral.

Além disso, aproveitando a essência e as necessidades inerentes ao paradigma de aprendizado de máquina, instituições como a SABESP e a NASA disponibilizam conjuntos de dados reais que espelham as condições ambientais em que o sistema opera. Esse acesso facilita a obtenção de uma base de dados abrangente, o que, por sua vez, capacita o modelo de treinamento a desenvolver resultados preditivos de forma mais refinada.

O fato mais interessante relaciona-se com a possibilidade de avaliar panoramas complexos, que é muito importante, pois ao possuir um alto grau de poluição, a conscientização ambiental torna-se urgente e a intervenção de órgãos públicos ainda mais. Dessa forma, é possível coletar informações relacionadas à deterioração progressiva do ecossistema aquático do complexo Billings e consequentemente do meio ambiente da região que tem causado impactos ambientais na região, gerando prejuízos sanitários, sociais, econômicos, ambientais, etc.

O objeto de estudo da Represa Billings será o braço Taquacetuba, do qual se tem uma ampla variedade de informações fornecidas pela Companhia Ambiental do Estado de São Paulo (CETESB). Este braço da Represa Billings é caracterizado por uma dinâmica diferenciada em relação ao corpo central, pois, possui uma comunidade fitoplanctônica gerada por diversos fatores meteorológicos e ambientais. Nesse sentido, as condições hidrodinâmicas são bastante diferentes com o corpo central, inclusive em aspectos físico-químicos, já que os dados apontam a que as densidades máximas registradas neste braço são consideravelmente menores do que as encontradas no corpo central (CARDOSO-SILVA et al., 2014).

Além disso, a seleção dessa seção da Represa Billings está vinculada a elementos que instigam cenários mais intrincados, dado o seu entorno próximo com a Represa Guarapiranga, a qualidade das águas envolvida e o seu significativo volume, que alcança a marca de 67 milhões de metros cúbicos, com uma descarga regularizada de 1,7

m^3/s (SABESP, 2023). Finalmente, é imprescindível destacar que a captação de água nesse braço da Represa Billings, encontra-se bastante afastado do corpo inicial e, consequentemente, as condições da água serão de maior qualidade devido à ausência de substancias indesejáveis de grande quantidade de lodo no sedimento. Outros aspectos relacionados à topografia e geologia das margens foram considerados na escolha do local para a instalação do sistema de captação e recalque.

Ademais, a partir do modelo treinado será possível contribuir às áreas da pesquisa em relação à análise dos comportamentos e tendências dos parâmetros da represa em função de tempo, identificando e analisando a relação que pode existir ou não entre os mesmos e como estes podem afetar os sistemas aquáticos, e dessa forma estabelecer relações entre os parâmetros do sistema e sua variação no tempo.

A análise de sistemas envolvendo várias reações bioquímicas assim como suas taxas, envolvem conhecimentos interdisciplinares, dentre eles alguns conhecimentos clássicos da área de Engenharia Química, como na área de engenharia bioquímica (Cinética de Michelis-Mentem), os quais são necessários para compreender os processos de produção e respiração de algas.

Por fim, aplicam-se conhecimentos relacionados à transferência de massa (absorção) para compreender os processos de transferência do oxigênio dissolvido entre a coluna d'água e o ar atmosférico como uma função da temperatura, nas áreas transversais como é a área da química, estudam-se assuntos relacionados a reações químicas, reações químicas em meios aquosos, assim como os processos de nitrificação e desnitrificação. Assim, são vários os exemplos de processos clássicos da Engenharia Química, que estarão presentes nas análises dos resultados deste estudo.

4 REFERENCIAL TEÓRICO

4.1 MEIO AMBIENTE

Em geral, o termo “meio ambiente” é associado à descrição de todo o que seja “natural”, ou resumidamente a todos os componentes vivos e abióticos que rodeiam a um organismo, ou um grupo de organismos (ZAVALA GUILLEN, 2018). O meio natural, está compreendido por todos os componentes físicos tais como o ar, temperatura, relevo, solo e corpos d’água, assim como também todos os componentes vivos como plantas, animais e microrganismos. Analogamente, também existe o “meio ambiente construído” o qual está constituído por todos os elementos e processos realizados pelo homem. Entretanto, embora seja utilizado para fazer referência a qualquer uma das duas terminologias, é mais comum se referir ao meio natural estritamente.

Os elementos que compõem o meio ambiente não atuam de forma isolada, sendo caracterizados por agir como partes de um sistema de processos que são vinculados entre si (SOLÓRZANO CHAMORRO; VERA BASURTO; BUÑAY CANTOS, 2022). Assim, o meio ambiente opera através de um complexo dinâmico de comunidades vegetais, animais e microrganismos que atuam de forma funcional e se complementam entre si.

Os assuntos relacionados com meio ambiente são estudados pela ecologia, um ramo da biologia especializada nos seres vivos e sua relação com o meio (GÓMEZ PUERTO, 2021). Os expertos e cientistas da área destacam a florestação e a preservação de corpos hídricos como uma questão fundamental, isto é, devido ao fato de que as arvores e a água cumprem funções vitais para um amplo número de fauna existente e também para os seres humanos, pois, tratam-se dos principais “recursos naturais” fundamentais para a preservação da flora e a fauna de forma generalizada.

Em síntese, qualquer organismo obtém do meio ambiente o sustento necessário para garantir sua supervivência, não apenas por uma questão alimentaria, se não também como fonte de energia, ar e inclusive refúgio (SCARDUA, 2021). Para os seres humanos, o meio ambiente trata-se do “fator chave para a vida”, no qual deve-se trabalhar para manter seu equilíbrio e assegurar uma qualidade de vida igual ou melhor com a que se tem atualmente.

Segundo o Banco Mundial, a correta administração dos recursos naturais e o meio ambiente permite a consolidação de bases fortes de um crescimento sustentável e inclusivo (MUNDIAL, 2020), contribuindo drasticamente na redução de muitos problemas

sociais e ambientais. Dessa forma, o meio ambiente, cria uma responsabilidade social que relaciona outros conceitos tais como sustentabilidade, ecossistemas, economia circular, entre outros, os quais caracterizam as condições de um meio ambiente de qualidade.

Entre os principais componentes do meio ambiente, encontra-se a água, pois, ao ser um líquido vital para a sobrevivência dos seres vivos, é indispensável para manter o equilíbrio do meio ambiente. Assim também, o ar é fundamental na identificação do grau de poluição atmosférica que pode influenciar na qualidade de oxigênio que é participe na respiração, e a temperatura, parâmetro pelo qual é possível monitorar o frio e o calor, que podem em muitas ocasiões afetar organismos que sobrevivem apenas em faixas de temperatura muito específicas (SOLÓRZANO CHAMORRO; VERA BASURTO; BUÑAY CANTOS, 2022).

Os desastres e o meio ambiente encontram-se diretamente relacionados, uma vez que existem os “desastres de origem natural”, os quais são eventos extremos que ocorrem de forma natural dentro de um ecossistema (PNUD, 2018). Estes eventos naturais são a consequência de uma mudança nas condições de um ecossistema podendo gerar fenômenos como derretimento da neve, desborde de rios, inundações, desertificação, etc. Embora, alguns desses desastres aconteçam de forma natural, a degradação do meio ambiente gerada pela atividade humana tem promovido um incremento na frequência e intensidade dos desastres de origem natural.

No decorrer da história tem-se evidenciado que os ecossistemas são muito resilientes. Sendo que muitos deles hão contribuído notavelmente no desenvolvimento e sustentabilidade do homem no percorrer do tempo. Entretanto, a industrialização, o crescimento populacional e a gestão não sustentável dos recursos naturais acarretaram numa deterioração generalizada do meio ambiente e consequentemente problemáticas ambientais muito graves e de amplo interesse e conscientização social.

Atualmente nas sociedades, a relação com o meio ambiente está diretamente relacionada, pois, independentemente de qual seja o proposito (econômico, social, político, etc.) ocorrerá um impacto direto no meio ambiente, gerando perturbações indesejáveis e danos ambientais (SCARDUA, 2021). Embora existam diferentes políticas públicas em prol do meio ambiente e com a finalidade de evitar as problemáticas de poluição e contaminação, não tem sido suficiente na restauração e recuperação dos ecossistemas acarretando numa problemática social que precisa ser solucionada antes que os danos sejam irreparáveis.

A interrogativa relacionada ao meio ambiente e sua compatibilidade com o crescimento econômico possui uma variedade de respostas atualmente, e são muitas as opiniões respeito ao assunto. Os ecologistas radicais ressaltam que o crescimento econômico levará à destruição de nosso habitat (meio ambiente) e os sépticos fundamentam que o crescimento econômico é a solução mais viável para mitigar os problemas ambientais (SOLÓRZANO CHAMORRO; VERA BASURTO; BUÑAY CANTOS, 2022).

4.2 RECURSOS NATURAIS

Os recursos naturais são definidos como os materiais e a energia contida na natureza e que são indispensáveis e uteis para os humanos (PEREIRA; FARIA; DO NASCIMENTO LIMA, 2021). Em termos mais técnicos, os recursos naturais são também distinguidos como os bens e serviços que proporciona a natureza sem intervenção direta do homem. Assim, os recursos naturais incluem os produtos animais, vegetais e climáticos (ar, temperatura, vento, água, etc.) os quais são gerados pela natureza e surgem independentemente da existência ou não do homem.

Estes recursos são colocados à disposição do homem, o qual tem como finalidade emprega-los de diversas formas dependendo da necessidade que tenha (SÁNCHEZ et al., 2019). Nesse sentido, os recursos naturais são utilizados e transformados pelo homem para satisfazer suas necessidades e seu aproveitamento ocorre por meios e tecnologias específicas.

Embora os recursos naturais sejam indispensáveis para o desenvolvimento tecnológico e para o funcionamento das sociedades, estes são limitados e sem um planejamento adequado e uma organização personalizada, alguns destes podem desaparecer no percorrer do tempo (BODIN; CRONA; ERNSTSON, 2017). Os recursos naturais podem ser classificados em várias categorias, dependendo de suas características, a classificação mais comum é por sua fonte de origem, por seu estado de desenvolvimento e por sua capacidade de renovação.

A classificação por fonte de origem divide os recursos naturais em orgânicos e inorgânicos. Aqueles que provém da matéria orgânica tais como plantas, animais e seus produtos, derivados da decomposição da matéria (carvão) são considerados como bióticos (orgânicos), já os produtos abióticos (inorgânicos) não derivam da matéria orgânica (solo, água, ar, vento, etc.) (SÁNCHEZ et al., 2019).

Outra forma de classificação está relacionada ao estado de desenvolvimento, ou seja, se eles estarão ou não disponíveis no futuro. Assim, dividem-se em “potenciais”, os quais são caracterizados por se encontrar em abundância numa determinada região e não estar sendo explorados atualmente (seja pelo pouco interesse ou pela dificuldade de extração). Os “atuais” que estão disponíveis numa zona que está sendo explorada e que existem informações relacionadas a sua quantidade, qualidade e disponibilidade no futuro e as “reservas” que se refere ao recurso natural cuja exploração se deixa para o futuro (SÁNCHEZ et al., 2019).

Finalmente, outra classificação muito conhecida está relacionada à capacidade de renovação do recurso natural, como seu nome diz, esta depende da disponibilidade do recurso no decorrer do tempo até seu esgotamento definitivo. Este tipo de recursos naturais possui duas subdivisões, podendo ser “renováveis” onde sua taxa de renovação é relativamente superior a sua taxa de uso e consequentemente enquanto se consume o recurso, é possível renova-lo para evitar sua desaparição no transcurso do tempo, e os “não renováveis”, os quais sua taxa de extração ou de consumo é maior que sua renovação, motivo pelo qual esgotam-se no decorrer do tempo (SÁNCHEZ et al., 2019).

O estudo e análise dos recursos naturais é de ampla importância devido ao fato que sua disponibilidade garante uma melhoria na qualidade de vida das sociedades, e seu uso racional é fundamental para sua preservação (PEREIRA; FARIA; DO NASCIMENTO LIMA, 2021). Nessa ordem de ideias, no momento que o homem utiliza os recursos naturais não obtém apenas benefícios pessoais, se não também comunitários que favorecem intrinsicamente o desenvolvimento tecnológico, social e o estilo de vida em geral.

Dessa forma, em prol da preservação e manutenção dos recursos naturais, a sustentabilidade é uma ação fundamental que permite o desenvolvimento ambiental, econômico e social tendo em conta que os recursos naturais são limitados e que eles devem estar disponíveis no futuro para as novas gerações e para as novas necessidades da sociedade (ORELLANA SALAS; LALVAY PORTILLA, 2018). Assim, ao proporcionar uma relação mais amena e eco amigável com a natureza, é possível criar um ambiente harmônico entre o meio ambiente (natureza) e as sociedades.

A América Latina possui uma ampla variedade de recursos naturais e um patrimônio natural muito diferenciado em comparação a outras regiões do mundo. Ao longo da história, tem enfrentado vários desafios importantes no que diz respeito à gestão e governança dos recursos naturais, onde a finalidade geral é promover um desenvolvimento

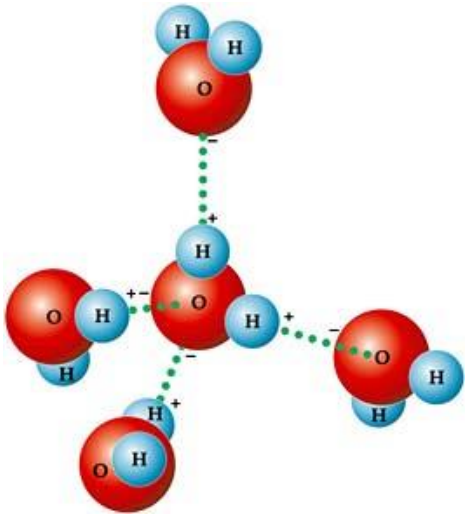
inclusivo e sustentável de suas sociedades num contexto econômico, social e ambiental (GRANT, 2001).

O território latino-americano possui um terço das reservas de água doce de todo mundo e uma quinta parte dos bosques naturais e aproximadamente 12% de seus solos são cultiváveis e favorecidos por sua dinâmica climática (SÁNCHEZ et al., 2019). Na américa latina, os recursos naturais, incluídos o setor agroindustrial, além da mineração, hidrocarbonetos e outros, representam mais de 70% das exportações totais, situando-se na última década entre os maiores distribuidores do mundo de recursos naturais de alta qualidade (MARACLE, 2020).

4.3 PARÂMETROS E QUALIDADE DA ÁGUA

A água em termos gerais, é uma molécula polar cuja conformação é disposta por dois átomos de hidrogênio e um de oxigênio (Figura 1), a suas ligações formam uma rede de ligações de hidrogênio que lhe tornam uma molécula bastante instável, cuja característica é sua capacidade de suportar altas temperaturas até atingir seu ponto de ebulição e fusão (SHRIVER, 1998).

Figura 1. Estrutura molecular da água



Fonte: SHRIVER, 2008.

Este recurso hídrico pode ser encontrado no estado sólido (gelo), gasoso (vapor) e líquido (água), sendo que cada uma de suas fases possui propriedades físicas e químicas muito diferenciadas e muito importantes na sobrevivência dos ecossistemas. As

características da água podem ser químicas, físicas ou biológicas e segundo sua composição pode ter diferentes classificações (doce, salgada, mole e dura as mais comuns) (PEREZ, 2015).

As características e versatilidade da água permitem que seja possível seu emprego em uma ampla variedade de áreas e ramos de pesquisa, isto devido às suas propriedades físico-químicas que lhe tornam uma substância "simples e versátil". Nos últimos séculos, a água tem-se tornado padrão na comparação com outras substâncias líquidas, e tem auxiliado na resolução de problemas rotineiros em base a suas características físico químicas, isto é devido a que sua densidade é de aproximadamente 997 $kg/m^3 \approx 1$ g/cm^3, além de outras propriedades de interesse, como é o caso da tensão superficial, que permite sua comparação com outros fluidos de forma mais genérica e mais simples, sendo a principal causa, que a água possui uma tensão muito alta (ACS, 2017).

A fundamentalidade deste recurso hídrico permite que seja objeto de alto interesse não apenas ambiental, mas também em outras áreas, tais como a social, econômica, energias renováveis, etc. Assim, a necessidade de que a mesma esteja nos parâmetros adequados respeito a sua qualidade, está intrinsicamente relacionada ao desenvolvimento social e tecnológico das regiões, estando inclusive relacionada com a redução de pobreza e às desigualdades (SERRAGLIO et al., 2021). Ao realizar um correto gerenciamento e disposição dos recursos hídricos, é possível realizar uma avaliação mais rigorosa do impacto do meio ambiente, pois, qualquer mudança nos padrões dos parâmetros será notável, e consequentemente será possível procurar soluções para combater o problema.

O emprego constante da água combinado com o desenvolvimento tecnológico de forma acelerada a nível mundial tem evidenciado uma influência notável nos padrões dos parâmetros da água e consequentemente na sua qualidade (causa da maior poluição). Assim, as legislações vigentes tem sido adaptadas com o intuito de suprir essas mudanças e tentar incitar um tratamento mais rigoroso da água, tendo em conta padrões que são benéficos para consumo humano e animal e portanto podem classificar o recurso, como de "qualidade" (SERRAGLIO et al., 2021).

Nesse sentido, a qualidade da água pode ser estimada tendo em conta um conjunto de parâmetros estabelecidos, legislados e adequados previamente (SÁNCHEZ MARTÍNEZ, 2017). A característica mais imprescindível que deve ter este líquido está relacionada a sua potabilidade, isto se deve ao fato de que ela possa garantir que os seres que a consumam estejam em condições de segurança e bem-estar. A qualidade da água representa

um dos papeis mais importantes na vida humana e nas características do ecossistema, e a partir de seus parâmetros físico-químicos e biológicos, é possível inferir sobre a qualidade ambiental de uma bacia hidrográfica e, portanto, de uma região em específico.

Além disso, as características de cor, turbidez e condutividade são utilizadas como parâmetros para determinar a qualidade da água (SÁNCHEZ MARTÍNEZ, 2017). A avaliação destes parâmetros é de ampla importância, pois, permite avaliar o comportamento dinâmico de parâmetros de qualidade da água importantes além de articular propostas para sua adequação. O monitoramento dos recursos hídricos cujo intuito seja avaliar a qualidade da água no Brasil são legislados por agências reguladoras que trabalham a níveis federais, estaduais, intermunicipais e em muitos casos municipais (SERRAGLIO et al., 2021). Nesse sentido, dependendo da região, os parâmetros tendem a ser diferentes, mesmo que suas faixas não apresentem variações tão abruptas.

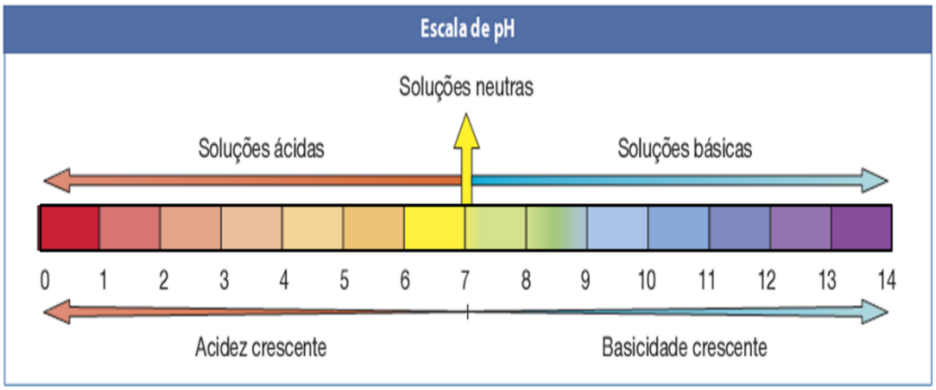
Os critérios de caracterização das águas no Brasil são estabelecidos de acordo com a classificação do curso hídrico. Dessa forma, tem-se em conta parâmetros numéricos com faixas estabelecidas e termos qualitativos que descrevem o índice de qualidade da água e sua preservação (PASSOS; DE FREITAS MUNIZ; FILHO, 2018). Em geral, estes parâmetros possuem uma ampla influência de padrões internacionais, porém, muitos são adaptados devido ao fato que existe uma diferenciação entre as águas nacionais e os corpos d'água internacionais.

A seguir estão apresentados alguns principais parâmetros de qualidade da água que serão utilizados ao longo deste trabalho.

4.3.1 pH

O pH (potencial de hidrogênio) é uma medida que indica as propriedades relacionadas à acidez ou alcalinidade da água (ACS, 2017b). Assim, em termos genéricos se define como a concentração de íons hidrogênio na água. Este tipo de medida possui uma escala logarítmica (Figura 2) com valores do 0 até o 14, o que diz que, conforme avança uma unidade na escala, há uma diminuição dez vezes maior na concentração de íons hidrogênio. Nesse sentido, com uma diminuição do pH, a água torna-se mais ácida, e analogamente, com o aumento do pH a água fica mais básica (VINET; ZHEDANOV, 2011).

Figura 2. Escala de pH



Fonte: "Blog de Físico-Química do André: Escala de pH", 2012

A escala de pH foi desenvolvida por Danes Sorensen no ano 1909. A partir dessa escala, é possível determinar a acidez ou alcalinidade de uma substância. Os valores estão fixados desde 0 até 14, onde o pH igual a 7 é neutro, abaixo de 7 é acima de 7 é alcalino ou básico. No ano de 2002, a IUPAC recomendou utilizar as convenções atuais da medida do pH baseadas na convenção Bates-Guggenheim (KOTZ; TREICHEL; WEAVER, 2003).

A importância do pH envolve análises em muitas áreas do conhecimento, e encontra-se presente na cotidianidade forma geral. Por exemplo, muitas reações químicas do corpo humano e animal acontecem numa faixa estreita de pH (metabolismo celular), e são necessárias para garantir sua sobrevivência (ZITA, 2021). Entretanto, assim como é benéfico, pode ser prejudicial quando não estiver ajustado para os organismos, como é o caso por exemplo dos peixes, que quando se encontram num meio com pH muito ácido ou básico, começam a ter danos físicos nas brânquias, esqueletos e nadadeiras (ACS, 2017b).

Quando as mudanças no pH não forem ajustadas, podem alterar a concentração de outras substâncias na água, modificando o nível de toxicidade. Este é um caso muito comum que é legislado no lançamento de efluentes que contém mercúrio. Isto se deve ao fato que se o pH do mercúrio for diminuído, este pode-se tornar solúvel na água e consequentemente é muito tóxico para a saúde pública e para o ecossistema (ROCHA; PEREIRA; DE PADUA, 1985).

Na área da bioquímica, a medição de pH é imprescindível, pois, muitas enzimas e proteínas possuem uma faixa específica para ter um alto índice de atividade e gerar um produto melhor. Assim também, os cultivos celulares devem ter um controle rigoroso da faixa pH, devido a que algumas células apenas sobrevivem nela, caso contrário morrem (ZITA, 2021). Na indústria alimentícia, o pH também é amplamente controlado para

prevenir o crescimento de organismos patógenos, os exemplos mais comuns são na indústria de lácteos e bebidas fermentadas.

Já na agricultura e na análise de solos, o pH determina que tipo de plantas sobreviveria em determinada região, além de tornar possível o monitoramento do crescimento de plantas, pois dependendo desta propriedade, é possível inferir sobre sua velocidade de crescimento além da conveniência de semear no lugar em questão (ZITA, 2021). Assim como ocorre com o solo, o pH tem uma alta influencia na avaliação da qualidade da água, pois esta característica permite determinar a solubilidade do liquido assim como a biodisponibilidade de substâncias químicas como nutrientes (fósforo, nitrogênio e carbono) e metais pesados (chumbo, cobre, cadmio, etc.) (CARBOTECNIA, 2010).

Nesse sentido, o pH da água está regido pelo equilíbrio do meio e sua concentração de compostos dissolvidos. Assim, em águas naturais este equilíbrio é caracterizado por um pH de 7 e deve ser avaliada a presença de dióxido de carbono (CO_2), ácido carbônico ($H_2CO_3^-$), além de outros componentes naturais, como os ácidos húmicos e os fúlvicos os quais são o resultado da degradação da matéria orgânica (SCARIOT, 2008). O pH também está relacionado às condições ambientais e climáticas das fontes hídricas, assim, dependendo do dia e a hora, este terá uma variação significativa, a qual é justificada por conta dos processos bioquímicos que ocorrem em diferentes lapsos de tempo.

A condição de acidez ou alcalinidade depende também da intervenção de outros fenômenos, como é o caso da radiação solar no processo de fotossíntese, dessa forma, como o dióxido de carbono reage com as moléculas de água produzindo o íon hidrônio, o pH se torna então mais ácido no meio (SCARIOT, 2008).

4.3.2 Radiação Solar

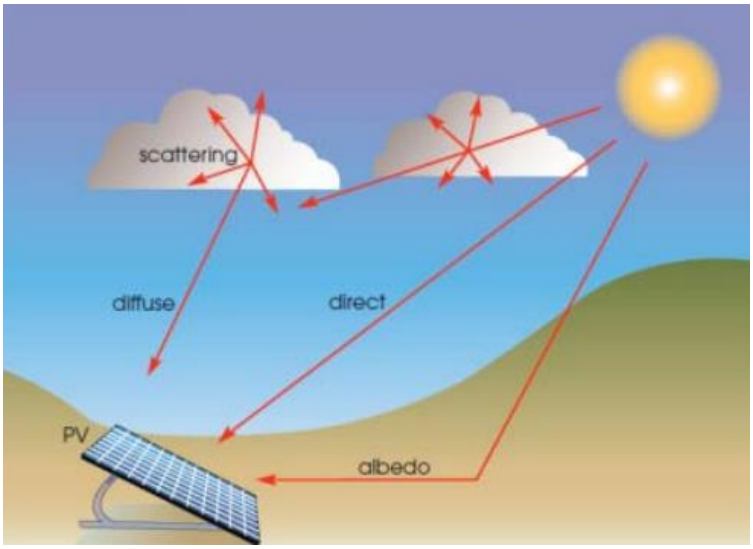
A radiação solar é a energia emitida pelo sol, a qual é propagada em todas as direções através do espaço mediante ondas eletromagnéticas, e sua geração está relacionada às reações de hidrogênio no núcleo do sol por fusão nuclear, sendo emitida pela superfície solar. Esta energia é a principal responsável pela dinâmica dos processos atmosféricos e o clima (RADIACI, 2012).

O sol é conhecido como a principal fonte de energia para todos os processos no sistema terra. Mais do 99% da energia que recebe a atmosfera, o ar e a água

provêm do sol, estes componentes do sistema terra absorvem energia solar a qual é irradiada na forma de calor. A energia recebida pela terra num ano atinge até $5{,}46x10^{24}$ watts (W) (VALDÉS; RIVEROS; ARACINBIA, 2012). A densidade de potência do sol (potência por unidade de área) acima da atmosfera terrestre é denominada como constante solar e é equivalente a 1,366 Wm^{-2}, entretanto, este valor é reduzido em aproximadamente 30% quando ingressa na atmosfera, dando uma insolação na superfície da terra de aproximadamente 1000 Wm^{-2}, sob nível do mar num dia claro (VALDÉS; RIVEROS; ARACINBIA, 2012).

A radiação solar que chega à superfície da terra possui três componentes (Figura 3). O primeiro é a "radiação direta" a qual procede em linha reta desde o sol, a segunda, a "radiação difusa" a qual provém de todas as direções, excetuando o sol e a terceira, o "albedo" que faz referência à radiação refletida pela superfície terrestre. A somatória dos 3 componentes constitui a radiação global como um todo.

Figura 3. Componentes da radiação global e sua incidência num painel fotovoltaico



Fonte: (LYNN, 2010).

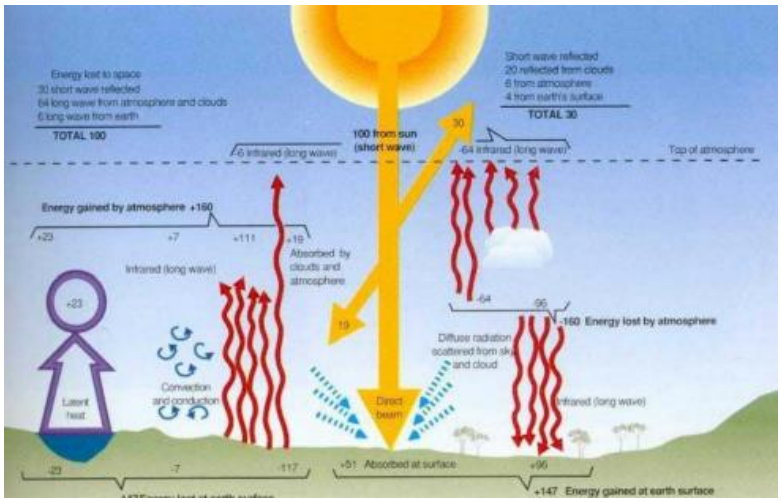
As informações relacionadas a radiação solar são de amplo interesse e possuem uma grande variedade de aplicações nas áreas de engenharia, arquitetura, agricultura, saúde, meteorologia, etc. Assim, na atualidade este tipo de energia é muito empregada como fonte alternativa de produção de energia na geração de eletricidade, em projeto de sistemas de aquecimento de água, projeção de edifícios e infraestrutura, monitoramento de crescimento de plantas, desidratação de alimentos, implicações na saúde,

análise de evaporação e irrigação, qualidade do ar, etc. (MAGARREIRO; FREITAS; BRITO, 2016).

O sol, em termos gerais, emite energia em forma de radiação de "onda curta". Após passar pela atmosfera, ocorre o processo de enfraquecimento por difusão, reflexão das nuvens e absorção pelas moléculas de gases (ozônio e vapor de água) e também pelas partículas em suspensão, assim, a radiação solar alcança a superfície terrestre oceânica e continental que reflete ou absorve este tipo de energia (RADIACI, 2012). A quantidade de radiação absorvida pela superfície é devolvida em direção ao espaço exterior na forma de radiação "longa" a qual é responsável direta pela transmissão de calor na atmosfera.

A atmosfera é majoritariamente transparente à radiação solar entrante, assim, no topo da atmosfera chega um 100% da radiação solar e aproximadamente 26% é dispersado pela atmosfera como radiação difusa até a superfície (Figura 4). Dessa forma, aproximadamente 51% da radiação atinge a superfície terrestre, sendo 19% absorvida pelas nuvens e gases atmosféricos (INZUNZA, 2019). Embora alguns componentes do sistema terra façam uma absorção notável da radiação solar, aproximadamente 30% da mesma é perdida no espaço, a qual é dispersa nas nuvens e dispersa no solo, o qual infere que a radiação solar que chega na atmosfera pode ser dispersada, refletida ou absorvida por seus componentes dependendo do comprimento de onda da energia transmitida e o tamanho e natureza da substância que modifica a radiação.

Figura 4. Distribuição da radiação no sistema terra

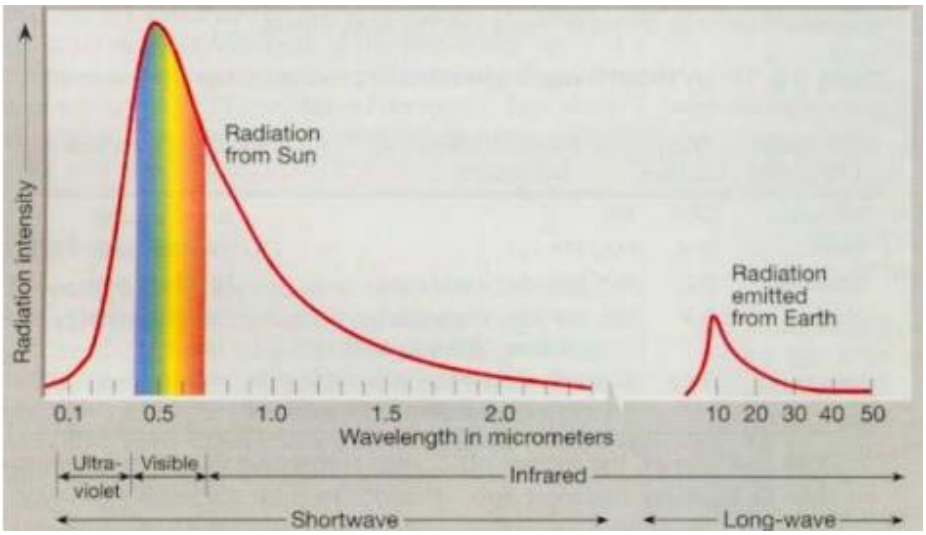


Fonte: (INZUNZA, 2019).

Como o planeta terra possui uma temperatura muito inferior que do sol, a radiação terrestre é emitida em comprimentos de onda muito mais longos que a radiação

solar de “onda curta” (LEAL IGA; LEAL IGA; SOL, 2015). Nessa ordem de ideias, a radiação terrestre emite comprimento de ondas numa faixa de 1 e 30µm, no range infravermelho do espectro, com um máximo de 10µm, por esse motivo, é conhecida também como radiação de “onda longa” ou radiação infravermelha (Figura 5).

Figura 5. Esquema da quantidade da radiação solar e terrestre



Fonte: (INZUNZA, 2019).

Embora o desenvolvimento tecnológico tenha avançado de forma acelerada atualmente, as medições de radiação solar em áreas específicas nem sempre estão disponíveis devido ao alto custo e aos requerimentos tecnológicos (JIMENEZ; WILL; RODRÍGUEZ, 2017). Assim também, a falta de base de dados climática é uma problemática comum no mundo, além das falhas e a incerteza na coleta de dados das estações meteorológicas, pois ao estar localizadas em locais remotos, seu acesso, reparação e manutenção muitas vezes torna-se difícil. Na literatura, é comum encontrar uma ampla gama de dados com os quais é possível estimar a radiação solar, entretanto, os modelos empíricos possuem limitações, pois, muitas vezes a modelagem carece de dados indispensáveis para montar modelos precisos de inteligência artificial.

Nesse sentido, a avaliação de modelos que incluam este parâmetro a partir de modelos de aprendizado de máquina torna-se de amplo interesse, uma vez que existe um amplo banco de dados disponibilizado por instituições confiáveis como é a *National Aeronautics and Space Administration* (NASA), e que a partir desses dados, há a possibilidade de reproduzir comportamentos em sistemas naturais, além de analisar

tendências nos sistemas que podem levar a inferir sobre a relação entre a radiação solar e outros parâmetros, e como estes podem afetar o ecossistema no decorrer do tempo.

4.3.3 Precipitação

A precipitação é a fase do ciclo hidrológico que consiste na queda da água desde a atmosfera até a superfície terrestre (SILVA et al., 2020). Este fenômeno é produzido como consequência da condensação, ou seja, da acumulação de vapor de água na atmosfera que propicia a formação das nuvens. Quando as nuvens acumulam um alto teor de vapor d'água, o peso das gotas faz que a água desça até a superfície. Este fenômeno também é conhecido como precipitação atmosférica ou precipitação pluvial.

Dependendo da temperatura em que esteja a atmosfera, a água que cai durante a fase de precipitação pode chegar à superfície em estado líquido ou sólido. As precipitações líquidas são denominadas "chuvas" de forma geral, e estas variam dependendo do tamanho das gotas e da intensidade na qual caem, assim como também do tempo de duração das mesmas (LOZANO-RIVAS, 2018).

As precipitações também podem ser apresentadas no formato de granizo, que são bolas de água congeladas que podem ter diferentes tamanhos e texturas. Este fenômeno ocorre quando a temperatura encontra-se muito próxima aos zero graus (0°C), provocando uma mudança de fase de líquido a sólido (processo de solidificação) (SILVA et al., 2018).

A precipitação é objeto de estudo da meteorologia por ser um fator importante na observação do clima e dos fenômenos atmosféricos. Sua medição permite estabelecer períodos de duração das chuvas e registrar a quantidade de água que chega à superfície (LOZANO-RIVAS, 2018). A unidade utilizada para medir a precipitação é o milímetro (mm) o qual é equivalente a um litro por metro quadrado de superfície (L/m^2). Estes valores permitem expressar o volume das precipitações que tenham caído sobre uma determinada área durante um período.

A atmosfera tem uma relação intrínseca com o fenômeno de precipitação, pois a partir de sua altitude e estação do ano é possível inferir sobre as propriedades pluviais de uma determinada região (CORREA; GALVANI, 2017). Sendo assim, a atmosfera tem características de reservatório, pois pode distribuir vapor de água e consequentemente influenciar as propriedades de precipitação pluvial em termos gerais.

A atmosfera está constituída por ar seco, vapor d'água e partículas em suspensão, e a formação de precipitações está relacionada à ascensão de massas de ar úmido, sendo que este fenômeno produz um resfriamento adiabático que promove que o vapor atinja seu ponto de saturação e que o ar seja expandido nas zonas de menor pressão. Dessa forma, favorecido por sua condensação e supondo condições favoráveis (com a existência de núcleos higroscópicos), o vapor de água sofrerá condensação e gerará gotas pequenas em torno dos núcleos higroscópicos (COSTA; BECKER; BRITO, 2013).

Devido a isso, as gotas precisam ter um peso maior do que a resistência do ar, pois caso contrário, permanecerão mantidas em suspensão nos formatos de nuvens e nevoeiros. Assim, quando atinjam um tamanho que ultrapasse à resistência do ar, ocorrerá um deslocamento em direção ao solo. Os crescimentos de gotas mais comuns destacados são os mecanismos de coalescência e de difusão de vapor (TUCCI; BERTONI, 1993)

As precipitações podem ser classificadas dependendo das condições nas quais é produzido o movimento vertical do ar (ascensão). Estas condições são caracterizadas tendo em conta fatores como convecção térmica, relevo e ação frontal de massas de ar. Sendo então classificadas principalmente como "precipitações convectivas", "precipitações orográficas" e "precipitações ciclônicas" (SILVA et al., 2018). No Quadro 1, encontra-se ilustrado as principais classificações das precipitações e suas características mais destacadas.

Quadro 1. Classificação e características das precipitações.

N°	Classificação	Característica(s) mais destacada(s)
1.	**Precipitação Convectiva**	Quando o aquecimento da superfície terrestre é desigual provoca o aparecimento de camadas de ar com densidades diferentes, gerando consequentemente uma estratificação térmica da atmosfera "equilíbrio instável". Estas são de grande intensidade e curta duração.
2.	**Precipitação Orográfica**	São o resultado da ascensão mecânica de correntes de ar úmido horizontais sob barreiras montais, tais como rochas e montanhas, assim, quando os ventos úmidos são resfriados de forma adiabática haverá geração de nuvens e consequente ocorrência de chuvas.
3.	**Precipitação Ciclônica**	Ocorrem ao longo da superfície de descontinuidade que separa duas massas de temperatura e umidade diferentes.

Fonte: Adaptado de (PEREZ, 2015).

A represa Billings possuí as características para classificar a sua precipitação do tipo “orográfica”, ao se encontrar numa região altamente montanhosa, a possibilidade de ocorrer correntes de ar úmido ao longo do ano é alta.

A dinâmica hidrológica de uma bacia hidrográfica é determinada por fatores naturais climatológicos (precipitação, evaporação, umidade do ar, vento, etc.) assim como suas características físicas. Dessa forma, a precipitação destaca um papel importante nas características das regiões e seus eventuais fenômenos, logo, a partir de seu comportamento é possível realizar análises sobre os sistemas estudados e torna-se interessante relaciona-la com outros parâmetros envolvidos no sistema estudado.

4.3.4 Velocidade do Vento

O vento desempenha um papel muito importante no estudo da dinâmica dos sistemas, isto devido a que dependendo de sua velocidade, o volume de ar que desloca pode ser maior ou menor e consequentemente, pode ocorrer uma variação na emissão de corpos atmosféricos que existirão nos sistemas (VENEGAS; MAZZEO, 2012).

A direção do vento mede a componente horizontal da velocidade do mesmo, e esta informação é muito útil na meteorologia, seja para fins informativos, quanto para gerar um prognostico mais preciso e adequado. Além disso, o vento é caracterizado por duas grandezas principalmente: a velocidade e a direção, sendo definidas em síntese, como a componente horizontal do deslocamento do ar num ponto e um instante determinado (IDEAM, 2018). A medição deste parâmetro é realizada a partir de um anemômetro, equipamento calibrado para obter a velocidade do vento numa unidade de medida em específico (que em geral está no SI). A velocidade média do vento varia entre 3 e 7 m/s, dependendo das situações meteorológicas (IDEAM, 2018).

A velocidade média do vento é menor durante o período noturno, onde é caracterizada por pouca variação, devido a que uma das principais causas do aumento da velocidade do vento é a saída do sol, onde suas mudanças mais destacáveis são dadas entre as 12 e 16 horas solares. A partir das informações obtidas tanto do vento, como de sua velocidade, é possível estabelecer os “perfis do vento” (ERASO CHECA; ESCOBAR ROSERO, 2018). Climas locais do vento são influenciados pela orografia e rugosidade do território, sendo assim, em geral, a velocidade do vento aumenta conforme aumenta a altura sobre o solo.

A caracterização do vento não é constante, e, portanto, o parâmetro está relacionado ao “vento instantâneo”, “vento médio” e rachas. Diante isso, na meteorologia, o vento instantâneo é medido por um período de 3 minutos, enquanto o vento médio é calculado em intervalos de 10 em 10 minutos (MARTINS DOS SANTOS; DA SILVA OLIVEIRA, 2021). O fenômeno estrutural (como as edificações, rochas e montanhas) também afetam à passagem do vento, o que leva a uma perda na sua velocidade, motivo pelo qual, em regiões onde há muita construção a tendência é a ter uma velocidade do vento menor em comparação a zonas menos construídas.

Nesse sentido, os “obstáculos” impedem o movimento uniforme do vento e consequentemente afetam a distribuição do mesmo e sua velocidade. Então, o vento ao ter um comportamento bastante modificado por causa das perturbações, poderá ter influência no sistema, uma vez que será responsável pelo lançamento de partículas em regiões especificas.

4.3.5 Temperatura

A temperatura é uma variável de estado da matéria que relaciona-se com a energia cinética média de suas partículas (BELENDEZ, 2017). Este parâmetro aporta uma ampla variedade de informações sobre as propriedades físicas dos corpos e está intrinsicamente relacionada a outras variáveis no sistema. Em termos mais técnicos, a temperatura é definida como uma magnitude que é relacionada com a velocidade média do movimento molecular da matéria. Logo, quanto maior seja o movimento das mesmas, maior será sua energia cinética, já que estas moléculas podem rotacionar e se deslocar com maior rapidez, consequentemente, a temperatura será maior (COLUNGA; OLGUIN GRANADOS; VARELA TOVAR, 2020).

A temperatura é uma magnitude que mede o nível térmico ou o calor que um determinado corpo possui. Na natureza, os corpos encontram-se num estado de agregação (sólido, liquido ou gás), em todos esses estados, as moléculas encontram-se em continuo movimento.

O estudo físico da temperatura relaciona-se com o “equilíbrio térmico”, isto devido ao fato que um sistema mecânico possui várias configurações que dependem da distribuição da energia de seus subsistemas. Dessa forma, dentre suas configurações, há uma na qual todos os subsistemas encontram-se em equilíbrio térmico, e que pode ser estudada

e/ou avaliada a partir das técnicas da mecânica estatística (SERWAY, 2018).

Existem várias escalas para expressar a temperatura, assim, as mais convencionais são o Celsius, a qual define a temperatura do ponto de gelo como 0°C e a temperatura do ponto de vaporização como 100°C. A escala Fahrenheit, a qual estabelece 32°F para a temperatura do gelo e 212°F ao ponto de ebulição de água, e por último a escala Kelvin, que fixa o zero absoluto (-273,15°C) (SERWAY, 2018).

O estudo da temperatura é fundamental em sistemas aquáticos, pois este parâmetro é função de muitos processos físicos, químicos e bioquímicos do sistema. Assim, conforme aumenta a temperatura, haverá alteração em muitas das características padrão do sistema hídrico tais como: viscosidade, tensão superficial, compressibilidade, calor especifico, constante de ionização, calor latente e calor de vaporização. Nesses casos, aumentos na temperatura geram como consequência a diminuição significativa dessas propriedades, e analogamente, outras propriedades como condutividade térmica, pressão de vapor e solubilidade (sólidos) aumentam (ROCHA; PEREIRA; DE PADUA, 1985).

Os parâmetros relacionados à temperatura, são determinantes e fundamentais no direcionamento da maioria dos processos físicos e bioquímicos dos sistemas aquáticos. Assim, quando um sistema é exposto a altas temperaturas, as taxas de reações químicas também aumentam de forma drástica. Além disso, outros aspectos também são afetados diretamente, como é a diminuição da solubilidade gases tais como oxigênio (O_2), dióxido de carbono (CO_2) e nitrogênio (N_2) (SCARIOT, 2008).

Dessa forma, a temperatura possui um efeito direto na cinética das reações químicas e, portanto, altera os sistemas aquáticos nas suas estruturas proteicas, enzimáticas e biológicas em geral. Alterações mínimas (> 5,0°) podem acarretar já na liberação de efeitos tóxicos de certas substâncias presentes na água, que podem reduzir o tempo de sobrevivência dos peixes. (ROCHA; PEREIRA; DE PADUA, 1985). Analogamente, o frio também pode causar a morte de várias espécies devido ao retardamento nos seus movimentos devido à queda brusca na temperatura.

Na análise hidrográfica haverá variações de temperatura conforme exista influencia na latitude, altitude e nas condições climáticas e ambientais. Porém, estas variações ocorrem de forma gradual e não uniforme, assim, o incremento da temperatura dos corpos hidrográficos gera o aumento da atividade biológica dos organismos vivos presentes na água. Nesse sentido, a biodiversidade que se encontra presente na água (tais como plantas aquáticas, animais, insetos, etc.) possuem um limite de tolerância à temperatura (SCARIOT,

2008). Embora exista esse limite, em muitos casos, quando é recorrente a alteração da temperatura, os organismos optam pela migração do ecossistema, acarretando em outra série de consequências ambientais.

A temperatura também influi nas propriedades químicas da água e consequentemente na sua qualidade. Um exemplo disso é que com o aumento da temperatura, ocorrerá uma alteração do sistema ocasionada por o aumento de reações químicas, que favorecem os processos de evaporação e volatilização das substâncias presentes na água, e que em muitos casos, podem ser substâncias prejudiciais ou indesejáveis, podendo também afetar a densidade da água, tornando-a mais densa devido às precipitações que são promovidas pelo aumento da temperatura (CARDOSO-SILVA et al., 2014).

Outros aspectos são afetados conforme a temperatura é aumentada, o mais comum está relacionado à taxa metabólica dos organismos, onde em muitos casos promove o crescimento acelerado de algas indesejáveis, que podem alterar a qualidade da água e tornam-se uma problemática ambiental de forma geral (GIORDANO, 2004).

4.3.6 Turbidez

A turbidez é um termo utilizado para descrever a transparência da água. Na análise de recursos hídricos, este parâmetro diz sobre o nível de lodo que possui a água, embora, seja um parâmetro relacionado à clareza da água, sua determinação está relacionada com as propriedades óticas da água, avaliando sua dispersão e absorção da luz na mesma (GONZÁLEZ TORO, 2014). O aumento da turbidez é destacado a partir do fato que muda a direção da luz, assim, se a turbidez for baixa, a luz não terá mudanças na direção, entretanto, se esta não for uniforme no passo da luz, infere-se que se encontra bastante turva. A regra geral diz que quanto menor for a clareza da água, maior será sua turbidez, porém, em muitos casos não é aplicada esta generalidade, pois existem águas transparentes, cuja turbidez é alta (GOYENOLA, 2007).

Este fenômeno ocorre de forma natural e se produz na maioria de corpos d'água tais como rios, lagoas, oceanos, etc. A fonte mais destacada de turbidez nas águas abertas é o fitoplâncton, espécie muito comum em fontes hidrográficas que podem estar expostas a partículas que contém argilas e lodos procedentes da erosão, sedimentos de leito em suspensão e resíduos orgânicos de forma geral (ROCHA; PEREIRA; DE PADUA, 1985).

Esta turbidez também pode ser ocasionada pela fauna aquática, como é o caso dos peixes, quando se alimentam e geram sedimentos que tornam a água mais turva.

A característica de turbidez também pode ter origem antropogênica, como é o caso do lançamento de efluentes com alto conteúdo de matéria orgânica, as descargas das águas residuais, os sistemas de pesca, etc. Isto, além de gerar uma problemática ambiental relacionada à poluição de recursos hídricos, também torna a água mais turva, e consequentemente requer tratamento para adequar esta característica.

Embora a turbidez seja uma característica que pode ser promovida pelo homem, a mesma pode ser ocasionada por causas naturais como tormentas, chuvas torrenciais e inundações, as quais criam correntes de águas rápidas que podem transportar um maior número de partículas e sedimentos de maior tamanho. Logo, haverá um maior arrastre de areia, lodo, argila e partículas orgânicas na superfície da água e a turbidez será maior (GONZÁLEZ TORO, 2014).

O estudo da turbidez nos diferentes corpos hidrográficos é de amplo interesse, pois, seu aumento pode deteriorar o entorno natural e alterar o crescimento de flora e fauna. Esta consequência ocorre devido a que quando a penetração de luz é bloqueada através da água, uma alta concentração de partículas do material podem modificar a penetração afetando os organismos de forma geral. Analogamente, se a penetração da luz for reduzida de forma significativa, é possível que se reduza a fotossíntese e a sua vez pode ser produzido um descenso na liberação de oxigênio na água durante o dia (METRÓLOGOS METAS & ASOCIADOS, 2010). A redução da penetração da luz também pode ter um impacto sensorial já que impede que alguns organismos possam visualizar seus alimentos, suas presas, seus predadores, etc. Isto é produzido se o aumento de turbidez é gerado de forma natural ou em eventos naturais, também se destaca que muitas espécies de flora e fauna estão habituadas à variabilidade natural da turbidez e podem sobreviver sem problemas com menos luz solar durante períodos específicos (dependendo da espécie).

As partículas em suspensão também podem absorver calor da luz solar, tornando as águas turvas, muito mais quentes, e consequentemente gerando uma redução na concentração de oxigênio na água (o oxigênio se dissolve mais simplesmente na água fria). Além disso, alguns organismos serão afetados por essa mudança de temperatura no meio no qual habitam e terão dificuldade para sobreviver no mesmo (GOYENOLA, 2007).

No tratamento de água, o impacto da turbidez considera-se meramente estético, entretanto, sua eliminação é indispensável para desinfetar efetivamente a água e

poder torna-la potável. As partículas em suspensão também promovem a adesão de metais pesados e uma variedade de compostos tóxicos que são prejudiciais para a saúde (GONZÁLEZ TORO, 2014). Os valores de turbidez em geral, são expressos em Unidades Nefelométricas de Turbidez (NTU) ou em mg/L de SiO_2 (miligramas por litro de sílica). No Quadro 2 ilustra-se a faixa de turbidez recomendada na avaliação da qualidade da água.

Quadro 2. Limites de turbidez recomendados na água

Uso da Água	**NTU**
Água potável	< 0,5 a 5,0
Água subterrânea	< 1,0
Piscicultura	10 a 40

Fonte: adaptado de GONZÁLEZ, 2014**.**

A represa Billings, atualmente, é caracterizada por um alto nível de turbidez, ocasionada pelo lançamento de efluentes tanto domésticos como industriais na região, nesse sentido, a problemática ambiental é muito destacada, pois, esta turbidez impede o efeito fotossintético das plantas no sistema aquático, além de tornar a água mais quente em termos gerais. (KOHATSU et al., 2018). Embora, a turbidez seja uma problemática bastante atual, é importante destacar que a RMSP, é caracterizada por um amplo período de chuvas nas estações invernais e outonais, portanto, a turbidez também é promovida por causas naturais.

4.3.7 Oxigênio Dissolvido (OD)

O oxigênio dissolvido (OD) é um parâmetro muito importante na análise de recursos hídricos, e define-se como a quantidade de oxigênio gasoso que se encontra dissolvido na água (GUERRERO, 2015). O oxigênio livre é uma característica fundamental para a vida dos peixes, plantas, algas e outros organismos aquáticos, portanto, é um parâmetro indicador da capacidade de um sistema hidrográfico para manter a vida aquática. A concentração deste elemento é resultado do oxigênio que entra no sistema e aquele que é consumido pelos organismos vivos. A entrada de oxigênio pode ser provocada por muitas fontes, entretanto, a principal fonte provém do oxigênio absorvido da atmosfera.

Em geral, o oxigênio é dissolvido com facilidade até atingir o ponto de saturação na água. Uma vez ocorre sua dissolução, este passa pelo processo de difusão de

forma lenta e sua distribuição vai depender do movimento da água (SIERRA, 2013). Este é considerado um processo natural e continuo, logo, haverá uma troca de oxigênio entre a água e o ar. A direção e a velocidade dependem da facilidade ou dificuldade de contato entre os dois componentes. Uma água turbulenta (torrente de montanha ou lagoa) estará caracterizada por uma maior absorção já que a superfície da água estará exposta ao ar, analogamente, as águas que se encontrem estagnadas irão reter e absorver menor quantidade de oxigênio.

As plantas possuem um papel bastante destacado na distribuição de oxigênio, e influenciam diretamente na quantidade do mesmo. Embora, seja uma problemática ambiental o controle da concentração de oxigênio dissolvido, sua insuficiência pode causar morte de alguns organismos, redução no crescimento de algumas plantas, problemas relacionados a larvas e mudanças que apresenta as espécies em diversas massas de água (GUERRERO, 2015). Além disso, a ausência de OD promove a migração generalizada de muitas espécies, o qual afeta drasticamente o ecossistema em questão. O oxigênio dissolvido (OD) é caracterizado em termos de concentração (mg/L) ou como a quantidade de oxigênio que pode ter a água a uma determinada temperatura.

Em geral este parâmetro é expresso nas unidades de mg/L (miligramas por litro) ou mais frequentemente (ppm) e avalia-se sua porcentagem de saturação, que é um termo utilizado na determinação da quantidade de oxigênio que é possível dissolver num volume específico a uma determinada temperatura (GARCÍA; MIRANDA, 2018), O oxigênio dissolvido é função de outros parâmetros na água, sendo a temperatura, o OD das fontes (entrada), OD das saídas, etc. A principal consideração realizada é o efeito da temperatura no meio aquático, pois, o aumento da temperatura diminui a quantidade de oxigênio dissolvido na água, podendo provocar situações indesejáveis como é o caso de que as águas recebam descargas de dejetos orgânicos.

A análise do OD é extremamente importante nos corpos hidrográficos, pois, é considerado um parâmetro na avaliação da qualidade da água. Ainda, o OD é indispensável para garantir o correto funcionamento das condições aeróbicas em aguas superficiais, e é considerado um dos principais indicadores primários na avaliação da sustentabilidade das águas superficiais para dar suporte à vida aquática (SCARIOT, 2008).

A salinidade, também é uma característica que afeta diretamente (em menor proporção quando comparada com a temperatura) na capacidade da água dissolver oxigênio. Em termos gerais, conforme aumenta a salinidade diminui a solubilidade de O_2 na água, logo, a quantidade de minerais e sais causantes de poluição podem alterar o teor de

OD na água, nesse sentido, é possível explicar porque nas aguas salgadas (mares e oceanos), os valores de OD são menores em relação às águas doces (FIORUCCI; FILHO, 2005) mesmo estando nas mesmas condições de pressão e temperatura.

Além dos fatores mencionados anteriormente, a disponibilidade e concentração de OD nos sistemas aquáticos também é afetada por outros fatores bioquímicos e climáticos. As principais fontes de oxigênio para a água são a atmosfera e a fotossíntese, entretanto, algumas perdas podem ser o resultado da decomposição de matéria orgânica (gerada pelo lançamento de efluentes), respiração de organismos aquáticos, nitrificação e oxidação química abiótica de substâncias como íons metálicos (FIORUCCI; FILHO, 2005).

Existem situações nas quais os sistemas hidrográficos estão expostos a situações ambientais severas, e consequentemente as implicações afetam significativamente o sistema. Este é o caso de quando a matéria orgânica é lançada nos corpos d'água, originários de esgotos domésticos, efluentes industrias ricos em matéria orgânica, etc. O aumento da matéria orgânica ocasiona uma diminuição na taxa de respiração de uma ampla variedade de microrganismos, além de gerar altas quantidades de dióxido de carbono (CO_2) e metano (CH_4), componentes químicos que são prejudiciais no sistema hidrográfico e que diminuem a quantidade de OD no meio, afetando de forma drástica tanto a flora como a fauna em questão (ANDRÉ LIMA DA COSTA et al., 2022).

4.3.8 Eutrofização

Na atualidade, a eutrofização é destacada como um dos maiores problemas de qualidade da água e de maior importância, uma vez que afeta tanto o sistema ambiental como a saúde. Este fenômeno é provocado pelo excesso de nutrientes na água, principalmente nitrogênio e fosforo, o qual é lançado majoritariamente como resíduo doméstico ou industrial (GÓMEZ, 2012).

O crescimento populacional de forma acelerada e o abrangente crescimento industrial, tem sido os principais responsáveis pela poluição dos sistemas aquáticos. As principais atividades que promovem a eutrofização dos recursos hídricos são a agricultura (quando se empregam fertilizantes nitrogenados que se filtram na terra até atingir águas subterrâneas), no setor pecuário (lançamento de dejetos animais ricos em nutrientes), resíduos urbanos e domésticos, contaminação atmosférica e a atividade florestal. (GÓMEZ, 2012). O excesso de nutrientes promove que as plantas e os organismos cresçam

de forma acelerada, e ao longo de seu ciclo de vida, consumem uma grande quantidade de oxigênio dissolvido, aportando um alto teor de matéria orgânica no sistema aquático em questão.

O nitrogênio e o fosforo encontram-se presentes na natureza em diversos formatos incluindo nos corpos aquáticos, embora cumpram uma função muito importante nos ecossistemas, quando são lançados em altas concentrações nas águas superficiais, provocam o fenômeno da eutrofização e consequentemente, ocorre uma variação nas propriedades relacionadas ao sabor, odor, turbidez, coloração da água e o mais importante, a redução do oxigênio dissolvido, provocando crescimento excessivo de algas, mortandade de peixes e tornando o meio aquático poluído de forma geral (BARRETO et al., 2013).

Nesse sentido, a eutrofização afeta a qualidade das águas, já que ao aumentar o número de plantas e organismos, a tendência é que o oxigênio se esgote, o qual afeta diretamente à fauna aquática e também a suas características físico-químicas, pois, gera odores indesejáveis que acarretam em perdas tanto econômicas, quanto problemas de saúde e sanitários às pessoas da região (SERRAGLIO et al., 2021). Com o desenvolvimento tecnológico atual, são várias as técnicas empregadas para a determinação da eutrofização nos corpos d'água, e é uma característica amplamente monitorada para evitar problemáticas no sistema aquático.

Monitorar a qualidade da água disponível é fundamental na gestão de recursos hídricos. No Brasil, estão implementados vários índices e indicadores ambientais para avaliação da mesma (BARRETO et al., 2013). Estes indicadores estão baseados nas propriedades físico-químicas e biológicas da água, assim, são estabelecidos os parâmetros relacionados aos níveis de trofia, os quais tem em conta a concentração de fosforo total, à clorofila e ao disco de Secchi, permitindo então, classificar as águas dependendo de suas características tróficas.

Uma das preocupações mais relevantes, enquanto aos problemas de salubridade gerados pela eutrofização, está relacionada ao crescimento de algas cianobactérias, as quais são potencialmente tóxicas e sua remoção e/ou tratamento é de alto custo. Este tipo de plantas, além de afetar a qualidade da água, é um problema para a saúde das pessoas e/ou animais que consumem este recurso (REZENDE et al., 2013).

A maioria de problemáticas de eutrofização encontra-se relacionada à alta concentração de fosforo e nitrogênio nos corpos hídricos, entretanto, não apenas por uma questão de produção de algas e ausência de oxigênio dissolvido. O alto teor de fosforo e

nitrogênio levam a acelerar o processo de biodegradação de produtos petroquímicos, hidrocarbonetos aromáticos e pesticidas, pois, o estado trófico da água acelera o aumento da biomassa bacteriana, e consequentemente, acarretará um aumento na diversidade de substratos orgânicos, os quais as bactérias têm facilidade de metabolizar.

Além de seu impacto ambiental, a eutrofização acarreta um aumento generalizado nos custos de tratamento da água, cuja finalidade é o abastecimento público, uma vez que são requeridos uma ampla quantidade de coagulantes e alcalinizantes para realizar um ajuste de pH de coagulação. Além disso, são adicionados bastante reagentes de alto custo para evitar a flotação, remover os flocos, etc. (BARRETO et al., 2013). Diante isso, a eficiência na desinfecção é muito cara e possui uma baixa eficiência devido à possibilidade de crescimento de bactérias no sistema de distribuição.

4.3.9 Potencial de Oxirredução (ORP) e Condutividade Elétrica

O potencial de oxidação e redução ou também denominado potencial de oxirredução (ORP) é um parâmetro muito utilizado na área de tratamento de águas devido a sua capacidade de deduzir e/ou caracterizar a fonte hidrográfica em termos de qualidade e potabilidade (VALLS, 2019). A partir das características relacionadas à oxirredução no recurso hídrico, é possível deduzir os níveis de desinfecção do mesmo.

De forma generalizada, é possível definir a condutividade elétrica como o fluxo de eletricidade promovido pelo transporte de elétrons podendo ser de dois tipos, condutores metálicos e/ou eletrônicos ou condutor iônicos (ANÓNIMO, 2014). Para realizar as análises de fonte hidrográficas, o estudo é enfatizado em condutores iônicos. A condutividade elétrica de uma solução é definida como uma propriedade de transmissão de corrente elétrica e tensão a partir do número, carga e movimentação de íons, além da viscosidade do meio, assim, em sistemas hídricos, a condutividade elétrica tem um comportamento definido em função da temperatura e a viscosidade do meio, onde, entre maior seja o aumento da temperatura, maior será o transporte iônico e consequentemente maior será a condutividade elétrica (BOYD, 2017).

Analogamente, o potencial de oxirredução mede a energia química proveniente das reações de oxirredução mediante um elétrodo, tornando-a em energia elétrica (ROCHA et al., 2023). Em geral, este tipo de energia elétrica é expresso na unidade de milivolts (mV), e dependendo de sua grandeza, é possível caracterizá-la.

O potencial de oxidação e redução (ORP) assim como na química, é

interpretado a partir de seus sinais, assim, quando o mesmo for positivo deduz-se que ocorreu oxidação (cátions) e se for negativo infere-se que ocorreu uma redução (ânions). Dessa forma, o princípio geral de medição de ORP nas fontes hídricas está baseado na avaliação dos sinais gerados por este parâmetro, onde a partir das cargas presente na água é possível deduzir a "saúde ambiental" de dito corpo hidrográfico (JARDIM, 2014).

A interpretação do ORP nas fontes hidrográficas está intrinsicamente relacionada a seu meio, caso este for um meio "redutor", o ORP será negativo, analogamente, se o meio for majoritariamente "oxidante", o ORP será positivo. No tratamento de águas, um residual desinfetante tipo oxidante na água sempre irá manter valores de ORP positivos. Assim, quanto maior for a quantidade de desinfetante oxidante no meio redox, maior será o valor captado no monitoramento do mesmo (VALLS, 2019).

Devido aos constantes desafios que abrangem os sistemas de tratamento de efluentes e em especial de águas, o parâmetro relacionado ao potencial de oxirredução tem sido bastante estudado, pois, existe uma correlação direta entre este e a atividade bacteriana (LUKASSEN et al., 2022). Um valor redox ao redor de 0 ou inclusive negativo caracteriza um excesso de matéria orgânica na fonte hidrográfica, portanto, a poluição será muito maior.

Semelhante ao ORP, a qualidade da água constantemente é avaliada em função da condutividade elétrica, na maioria dos casos em base a seu sistema redutor. A "linha divisória" ou faixa determinante está fixada nos 75 mV, assim, um sistema será oxidante quando seu potencial redox atinja ou supere 75 mV, caso contrário, o mesmo será redutor (JARDIM, 2014). A classificação a partir desta faixa determinante estabelece que águas cujo ORP seja $\leq$ 75,0 mV, a água é predominantemente redutora, e consequentemente encontra-se caracterizada por microrganismos e matéria orgânica. Analogamente, se for $\geq$ 75,0 mV, será predominantemente oxidante, portanto, a quantidade de matéria orgânica e microrganismos será menor.

No ano 1971 a Organização Mundial da Saúde (OMS) (ORGANIZACIÓN MUNDIAL DE LA SALUD, 2015), adaptou um valor padrão de ORP para águas potáveis. Assim, águas com um ORP de 650,0 mV são consideradas potáveis devido a seu alto grau desinfeção. Neste mesmo ano, pesquisadores chegaram na conclusão que existe uma relação exponencial entre a velocidade de inativação de vírus em águas e o ORP, faixas iguais ou superior a 650,0 mV provocam quase instantaneamente a desativação de vírus inclusive se estiverem em altas concentrações (VALLS, 2019).

As condições atuais do complexo Billings nos últimos anos têm demostrado que os parâmetros de qualidade da água relacionados ao ORP e a condutividade elétrica caracterizam a fonte hídrica como “altamente poluída”, valores de monitoramento do ORP próximos a 0 e condutividade elétrica baixa são sinais de que se tem perdido gradativamente a qualidade da água e consequentemente sua potabilidade.

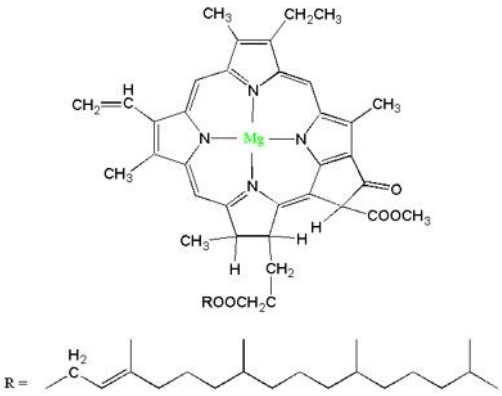
Por este motivo, torna-se de amplo interesse avaliar modelos ambientais do braço Taquacetuba da Represa Billings a partir da simulação destes parâmetros de monitoramento e avaliação da qualidade da água.

4.3.10 Clorofila

A clorofila é um pigmento de coloração verde capaz de absorver luz solar ou artificial a qual se encontra em todos os organismos que fabricam seu próprio alimento a partir do processo de fotossíntese, os quais são mais comumente denominados como organismos “autótrofos” (ROBERTIS; HIB, 2017). A clorofila formada em folhas, talos e outras partes das plantas é encontrada em estruturas denominadas “cloroplastos”, e é neste local onde ocorre o processo de fotossíntese.

A pesar que comumente associa-se a clorofila apenas para se referir à pigmentação nas plantas, ela encontra-se presente também em algas e bactérias como é o caso das cianobactérias (CARDOSO-SILVA et al., 2014). Este conjunto de células orgânicas dos cloroplastos são característicos das células vegetais, e seus pigmentos vegetais promovem as reações fotoquímicas necessárias para a fotossíntese. Na sua estrutura molecular, a clorofila possui um anel de porfirina que contém magnésio (Figura 6), cuja principal função é absorver a luz. Devido a sua estrutura molecular, a clorofila através do processo de fotossíntese permite a conversão de energia inorgânica (dióxido de carbono e água) em energia orgânica (hidratos de carbono), sendo o principal receptor de energia luminosa neste processo.

Figura 6. Estrutura molecular da clorofila



Fonte: (ROBERTIS; HIB, 2017).

A clorofila é classificada em dois principais grupos, a clorofila a e a clorofila b. A clorofila a é o pigmento principal responsável pela captura da luz solar e pela sua transformação em energia química nas células vegetais, já a clorofila b, trata-se de uma pigmentação adicional, cujo objetivo é complementar a ação da clorofila a, ampliando a faixa de luz absorvida (abrangendo um espectro de regiões de luz maior) (ROBERTIS; HIB, 2017). A clorofila a é considerada como diretamente proporcional à concentração de biomassa das algas (SCARIOT, 2008).

Nos corpos d'água, bacias hidrográficas, e em geral todos os sistemas pluviais, a quantidade de clorofila é um indicador de ampla importância em relação a sua qualidade (COND, 2004). Em geral, quando uma fonte hidrográfica apresenta altos níveis de clorofila infere-se que trata-se de uma água com índices notáveis de eutrofização, e portanto haverá um desvio do regime natural do ecossistema, estando caracterizado pelo alto crescimento de algas promovido pelo acúmulo excessivo de nutrientes como é o caso do fosforo e o nitrogênio na água (PERDOMO; BARBAZÁN, 2007).

O monitoramento de clorofila é bastante habitual e rotineiro nos sistemas de tratamento de águas, embora a clorofila não seja sempre distinguida como um parâmetro crítico na avaliação do tratamento das águas, trata-se de um indicador importante na determinação da potabilidade da água (GIORDANO, 2004). Assim, águas com níveis de clorofila muito altos, tendem a estar caracterizadas por baixa potabilidade e consequentemente torna inviável seu consumo. Além disso, ao ser uma característica intrinsicamente relacionada à eutrofização, deduz-se que águas com níveis altos de clorofila estão consequentemente eutrofizadas, tornando seu tratamento muito mais caro e menos eficiente (devido à aparição de algas cianofíceas).

A represa Billings, no decorrer dos anos tem apresentado uma perda generalizada na qualidade de sua água, os monitoramentos realizados pela CETESB demostram que os índices de clorofila são muito elevados, o que infere consequentemente de seu grau de eutrofização. O lançamento de matéria orgânica pela população e pelas industrias das cidades vizinhas tem acarretado no incremento gradativo deste parâmetro.

As informações atuais da CETESB/SABESP estabelecem que conforme o percorrer dos anos, tem-se tornado mais complicado e desafiador o tratamento das águas do complexo Billings, chegando atingir inclusive ao braço Taquacetuba, cuja característica principal era possuir parâmetros de qualidade da água maiores quando comparado com outros braços (CETESB - SP, 2020).

A seguir estão apresentas as principais características da represa Billings além de sua localização geográfica.

4.4 REPRESA BILLINGS

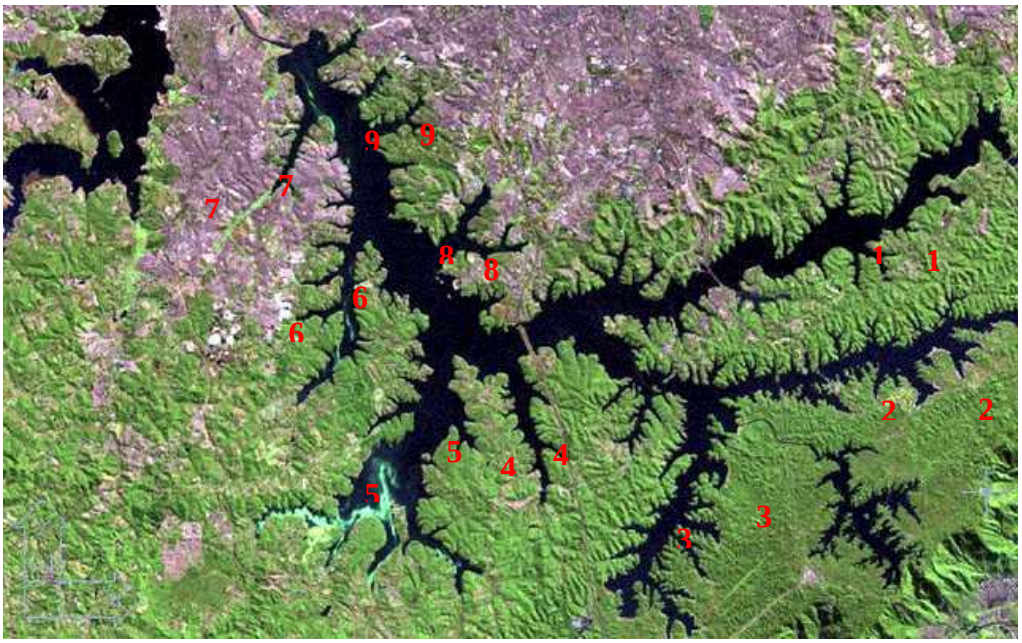
A Represa Billings é o maior reservatório de água da RMSP, entretanto não pode ter seus recursos hídricos plenamente aproveitados, por conta do alto grau de poluição e degradação da qualidade de suas águas e sedimentos (ALVIM et al., 2022) contaminados ao longo de décadas, principalmente, pela reversão do Rio Tietê para o canal do Rio Pinheiros em direção ao reservatório Billings. Situa-se na porção sul da RMSP em território dos municípios de Diadema, São Bernardo do Campo, São Paulo, Santo André, Ribeirão Pires e Rio Grande da Serra, fazendo limite, a oeste, com a Bacia Hidrográfica da Represa do Guarapiranga e, ao sul, com a Serra do Mar.

Devido a seu formato peculiar, a Represa Billings está subdividida em braços ou sub-bacias, que são as seguintes: Braço do Rio Grande (1), separado do corpo central pela barragem da Rodovia Anchieta; Braço do Rio Pequeno (2); Braço do Rio Capivari (3); Braço do Rio Pedra Branca (4); Braço do Taquacetuba (5); Braço do Bororé (6); Braço do Cocaia (7); Braço do Alvarenga (8) e Braço do Grota Funda (9) (Figura 7).

A população em torno da Represa Billings está mal distribuída sendo que a maior densidade demográfica ocorre no Braço do Cocaia, com 114 hab/ha, seguido por Grota Funda, Alvarenga e Corpo Central. Já as menores, dão-se nos braços Capivari com 0,3 hab/ha e Rio Pequeno com 0,5 hab/ha. E apesar das densidades habitacionais serem distintas, todos possuem taxas de urbanização elevadas, variando de 92 a 99%, com a população

concentrada nas aras urbanas.

Figura 7. Localização geral da área de estudo – Represa Billings.



Fonte: (MEHTA; KANANI; LANDE, 2019).

Atualmente, a água da represa continua degradada (DE LIMA, 2023), porém a crescente ocupação irregular e desordenada ao seu redor é outro fator que ameaça a qualidade de suas águas. Aproximadamente 700 mil pessoas vivem no entorno do reservatório Billings, gerando problemas ambientais devido ao lançamento de esgotos domésticos e ao desmatamento das áreas verdes remanescentes.

As florações de algas cianofíceas têm sido cada vez mais frequentes em reservatórios brasileiros, como é o caso da represa Billings. Essa espécie foi classificada como tropical ou subtropical, mas hoje é também encontrada em regiões temperadas, dada a sua capacidade de adaptação e à sua competitividade. Essas florações consistem, de fato, grande problema para a saúde pública pois esta alga produz toxinas diversas que atuam no fígado, rins e no sistema neuromuscular, além do fato que é extremamente caro o tratamento das águas eutrofizadas destinadas ao abastecimento público (PADIAL, 2009).

Com o objetivo de regularizar o abastecimento de água da RMSP, a Companhia de Saneamento Básico do Estado de São Paulo (SABESP), iniciou-se em 1996 o Programa Metropolitano de Água (PMA), que contemplava uma série de obras com o objetivo de atingir a meta estabelecida, garantindo o suprimento de água na região até o ano 2000 (KUSSAMA et al., 1997).

Uma das obras prioritárias da SABESP consistiu na implantação do Sistema Produtor Taquacetuba-Guarapiranga que, visava à transposição das águas do Braço Taquacetuba da Represa Billings para a Represa Guarapiranga para incrementar o suprimento de água à população. Segundo o Plano Integrado de Aproveitamento e Controle dos Recursos Hídricos das Bacias Alto Tietê, Piracicaba e Baixada Santista (HIDROPLAN, 1995), o aporte de água do reservatório Billings pode variar de 1,7 a 4,5 m^3/s para o abastecimento público.

O braço Taquacetuba é uma região de 8.102 hectares, possui aproximadamente 5 km de extensão e abriga uma população total de 32.278 habitantes (26.247 na área urbana). Localiza-se nos municípios de São Bernardo do Campo e São Paulo, com presença da Área de Proteção Ambiental Municipal Capivari/Monos e terras indígenas ao sul. Segundo os critérios do IBGE, inexistem favelas nessa região e parte das águas deste braço é bombeada para o Reservatório do Guarapiranga como suplemento de vazão para abastecimento público.

4.5. SIMULAÇÃO

A simulação é um tema de destaque na atualidade que se encontra no processo de disseminação ao redor do mundo todo. O termo simulação possui vários significados dependendo do contexto, entretanto, em termos gerais, define-se simulação como a imitação de um processo, assim como o processo de desenho de um modelo que forma parte de um sistema real ou imaginário (DIECKMANN, 2020). O objetivo da simulação é reproduzir as capacidades e o comportamento de um determinado sistema, sem a necessidade de reproduzi-lo ou experimentar com ele em escala real, seja por uma questão de custo, riscos ou por outros fatores limitantes.

As aplicações mais comuns das simulações são dadas nos campos de projeto para visualizar um sistema inexistente, encontrar soluções para um problema em específico, ou mesmo avaliar o comportamento de materiais, quantificar a produção de um produto, na análise de sistemas complexos, etc. Assim, a simulação busca delimitar capacidades e avaliar comportamentos de um sistema em questão, seja por mudanças na estrutura do sistema ou modificação de uma de suas partes (GONZÁLEZ et al., 2011).

Além disso, a simulação também é útil para avaliar modelos complexos que possam chegar a ter possíveis incidências de fenômenos externos. No treinamento das

simulações, recriam-se possíveis cenários que podem ser enfrentados em determinadas situações. O processo de simulação é desenvolvido a partir de uma série de passos que estruturam um modelo e seu funcionamento num sistema, assim também validam, operam e analisam os resultados do mesmo a partir de situações supostas (COUTINHO; FIGUEIREDO, 2020).

Na simulação é muito importante definir o modelo conceitual, ou seja, explicitar os algoritmos, teorias e limites que podem descrever o sistema, definindo também as entradas de informação requeridas e as saídas (*input* e *output*). Logo, reúne-se a informação que será utilizada na parametrização dos dados de entrada e na avaliação do rendimento da simulação até finalmente chegar na construção de um modelo de *software,* cujo objetivo é a execução de tarefas de simulação (GONZÁLEZ et al., 2011).

São aplicados vários critérios de verificação, validação e acreditação do modelo para comprovar seu funcionamento correto, considerando os dados de entrada inseridos, assim como os parâmetros estabelecidos. Dependendo da situação, nos desenhos experimentais, são tidas em conta limitações que partem das diferentes faces do processo, garantindo dessa forma mais confiabilidade na simulação e nos dados de saída (resultados obtidos da simulação) (COUTINHO; FIGUEIREDO, 2020).

Na simulação, os estados são grupos de variáveis que descrevem o sistema num intervalo de tempo, ou seja, é o conjunto de características temporais que tornam reconhecível uma parte do ciclo do sistema. A importância dos componentes da simulação é definida por sua prioridade, a forma com a qual o tempo é representado na simulação e a relevância nos critérios utilizados na configuração do estado (COUTINHO; FIGUEIREDO, 2020).

A aleatoriedade é um conceito muito ligado à simulação, isto devido a que a maioria de cenários reais não estão subscritos à exatidão. Existem faixas de erro e possibilidades mínimas de mudanças súbitas na simulação. Por causa disso, muitos modelos requerem valores aleatórios para introduzir a variabilidade ao sistema (IRRGANG; SAYNISCH-WAGNER; THOMAS, 2020).

Na engenharia química, a modelagem e simulação são termos indispensáveis na compreensão do comportamento das partes de um sistema e do sistema como um todo (FRANCO, 2021). A necessidade da aplicação destas ferramentas, encontra-se relacionada à possibilidade de estudar sistemas sem a necessidade de realizar testes experimentais, permitindo avaliar modelos e/ou comportamentos de sistemas naturais que

podem ser prejudiciais numa determinada faixa de tempo, propondo assim, a possibilidade de aplicação de técnicas para a otimização de processos, tendo em conta um conjunto de dados de entrada já estabelecido.

4.5.1 Simulação de Sistemas Naturais

A simulação de sistemas naturais é uma disciplina transversal que permite obter um nível de compreensão dos sistemas naturais, seus comportamentos, fenômenos e tendências utilizando um enfoque interdisciplinar. Assim como qualquer outro tipo de simulação, na simulação de sistemas naturais, empregam-se técnicas matemáticas formais, conceituais e computacionais para o estudo de sistemas naturais, nas áreas de simulação de sistemas climáticos, modelagem e simulação de sistemas de cultivos, simulação de crescimento de cultivos, dinâmica populacional, etc.

A simulação de sistemas naturais permite prever cenários a partir de técnicas matemáticas e numéricas de uma ampla gama de possíveis fenômenos de sistemas naturais, destacando um papel muito importante na engenharia ecológica e áreas afins, pois, permite avaliar panoramas de forma geral e relacionar parâmetros tendo em conta o modelo planteado.

Entender o funcionamento das bacias hidrográficas torna-se de amplo interesse devido à necessidade de prevalecer a água como recurso hídrico fundamental para à sobrevivência. Assim, a simulação de sistemas naturais permite realizar um estudo mais aprofundado sobre os balanços hídricos e os processos que controlam o movimento da água, assim como os impactos das mudanças do uso do solo sob a quantidade e qualidade da água (VIANA et al., 2018).

A modelagem e simulação são ferramentas importantes na gestão e à tomada de decisões sobre o uso da água. Os modelos são capazes de representar os processos físicos de um sistema e dessa forma trazer um conjunto de informações que permitem representar os processos físicos do mesmo, assim como gerar informações normalmente não disponíveis e/ou não previveis. Uma vez modelado o sistema e posteriormente simulado, é possível a obtenção de panoramas, cenários, modelos, tendências e informações importantes de forma geral, permitindo dessa forma avaliar e/ou estabelecer soluções a partir das observações e inclusive validar o modelo ou em seu defeito calibrá-lo de tal forma que permita sua implementação em outros tipos de ecossistemas (ALDANA et al., 2017).

4.6 APRENDIZADO DE MÁQUINA

O aprendizado de máquina ou também conhecido como *machine learning*, é uma disciplina da área da computação, a qual se encontra como um subconjunto trabalhado na inteligência artificial (IA), e que serve para criar sistemas "que podem aprender por si mesmos" (F. FRANCO; J. RAMOS, 2019).

Assim, a denominação de aprendizado é dada devido à capacidade do sistema para identificar uma ampla serie de padrões complexos, os quais são determinados por uma grande quantidade de parâmetros, ou seja, a máquina só consegue aprender utilizando algoritmos específicos para tal atividade, os quais podem ser modificados conforme muda a entrada de dados, e que pode, nesse sentido, predizer cenários futuros ou tomar ações de forma automatizada segundo condições previamente estabelecidas (REDACCIÓN APD, 2019).

Os algoritmos utilizados no aprendizado de máquina, realizam boa parte das ações por sua conta além de terem a capacidade de obter seus próprios cálculos segundo os dados que são recopilados no sistema, e quanto maior for a quantidade e qualidade dos dados, mais precisas e melhores serão as ações resultantes (F. FRANCO; J. RAMOS,2019).

No aprendizado de máquina, também há uma influência muito significativa da aplicação de fórmulas heurísticas, as quais podem-se tornar ferramentas na elaboração de um sistema de inferências. O sistema de aprendizado de máquina precisa de um amplo volume de dados de relevância para poder subministrar respostas realmente válidas (REDACCIÓN APD, 2019). Os sistemas de aprendizado de máquina, permitem a predição do comportamento dinâmico de parâmetros importantes ou iniciar operações que são a solução para uma tarefa em específico. A partir de um amplo volume de dados também é possível deduzir e generalizar um comportamento e baseado nele, realizar predições para casos totalmente novos. Atualmente, existem três tipos de aprendizado de máquina mais comuns (Quadro 3).

Quadro 3. Tipos de aprendizado de máquina mais comuns

Tipo de aprendizado de máquina	Breve descrição
Aprendizagem supervisionado	Conhece-se a informação do treinamento, proporcionando uma quantidade de dados e definindo-os com etiquetas. Assim, ao ser adicionados novos dados, poderão ser introduzidos sem necessidade das mesmas.
Aprendizagem não supervisionado	São treinados em dados não rotulados, sua finalidade é a compreensão e abstração de padrões de informação de maneira direta, permitindo a verificação de dados novos e estabelecendo conexões entre a entrada desconhecida e as saídas padrão.
Aprendizagem por reforço	Os sistemas aprendem a partir da experiência. É uma técnica baseada no teste e o erro, e no uso de funções que otimizam o comportamento do sistema.

Fonte: adaptado de (REDACCIÓN APD, 2019)

Na aplicação da técnica de aprendizado de máquina, são empregados diversos modelos que possuem diferentes aplicações dependendo da avaliação e/ou sistema analisado, entre esses modelos encontramos o modelo de regressão linear e o modelo floresta aleatória (*Random Forest*). O modelo de regressão linear está caracterizado principalmente na identificação de variáveis que se encontram correlacionadas num determinado conjunto de dados, ou seja, de forma sintética, este modelo permite estabelecer a relação existente entre a variável dependente (*target*) e as variáveis independentes (*features*) procurando a reta que mais se aproxima das observações e com menor número de erros (MAULUD, 2020).

Por outro lado, o modelo floresta aleatória, trata-se de um algoritmo de aprendizagem de máquina cuja funcionalidade é obtida a partir da combinação do resultado de múltiplas árvores de decisão que tem como objetivo atingir um único resultado (LIU, 2012). Este modelo tem-se tornado uma inovação em diversos campos do conhecimento devido a sua capacidade de adaptação tanto em problemas de classificação quanto de regressão.

As informações relacionadas aos parâmetros de qualidade da água são fundamentais para a implementação de modelos de aprendizado de máquina e no Brasil, devem ser divulgadas por parte dos órgãos que são encarregados de seu processamento, e são de livre acesso ao público. Assim, são vários os bancos de dados que existem (e estão

disponíveis) de muitas represas, reservatórios e corpos hídricos, permitindo então, a aplicação de modelos de aprendizado de máquina. Isto devido às facilidades que acarreta sua implementação, pois, ajuda a processar, extrair e resumir informações uteis, além da vantagem de prever o comportamento dinâmico de parâmetros importantes e o e panorama dos recursos hidrográficos a partir das informações fornecidas.

5 METODOLOGIA

Para a execução do plano de trabalho, foram necessários *softwares* apropriados para a estruturação e respectiva organização das tabelas (banco de dados), além de um visualizador de dados e/ou *software* destinado à aplicação de aprendizado de máquina. Os principais *softwares* utilizados foram o Orange Data Mining (cuja natureza é gratuita, *software* livre) e o *software* Excel ® (pacote office).

Para a organização e estruturação das planilhas foi utilizado o *software* Excel ®, e em geral todo o pacote office. Todas as atividades relacionadas à organização, estruturação, tratamento de dados, conversão de dados, disposição de variáveis e plotagem de gráficos foi realizada neste *software*. O principal motivo desta escolha está relacionado à praticidade que possui o Excel ® no seu manuseio e aplicação, permitindo realizar todas as atividades e aplicar todas suas ferramentas matemáticas e visuais de forma intuitiva e simples.

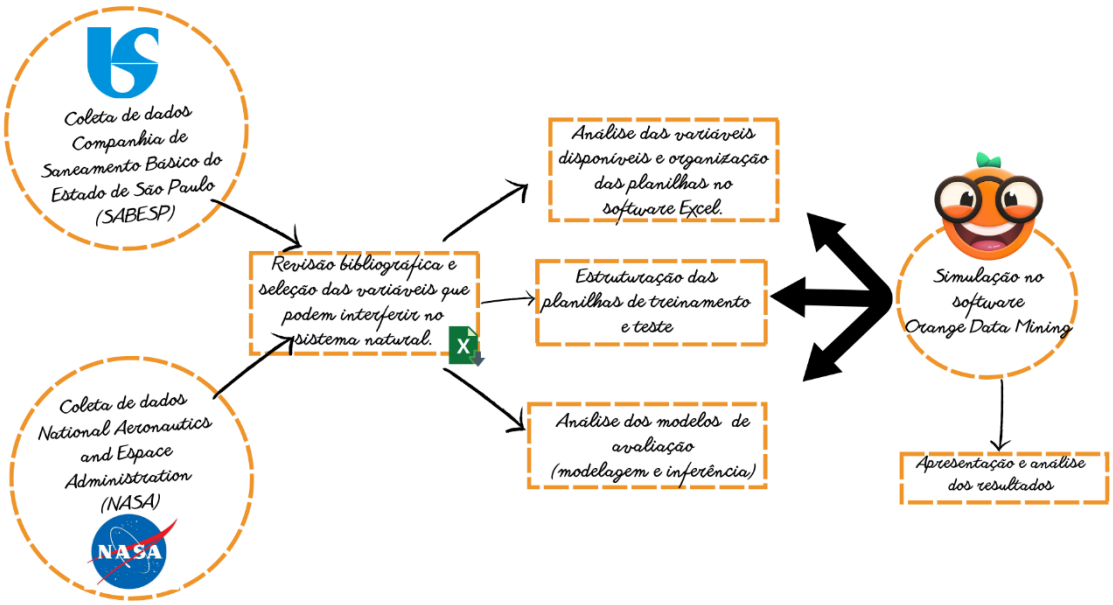
Na realização da simulação e respectiva aplicação do aprendizado de máquina, utilizou-se o *software* Orange Data Mining, o qual possui uma ampla quantidade de ferramentas e modelos de aprendizado de máquina que são compatíveis com o comportamento dinâmico dos sistemas naturais. Entre as principais características que tornaram o Orange um *software* ótimo para aplicação de aprendizado de máquina encontra-se a praticidade do mesmo, simplicidade no *layout* e nas configurações, além da disponibilidade de ferramentas de análise de dados, a qual permitiu a visualização e simulação de um amplo conjunto de dados de forma simples e compatível com o *software* de organização e estruturação do banco de dados (Excel ®).

A utilização desses dois *softwares* permitiu realizar análise do comportamento e tendências sazonais de sistemas naturais, tendo em conta as especificidades do sistema estudado em questão quando comparado com os dados reais.

5.1. FLUXOGRAMA DE TRABALHO

Para a execução do plano de trabalho, foi proposto o fluxograma de trabalho disponível na Figura 8. De forma sintética, o trabalho passou pelas etapas de coleta, organização e estruturação dos dados, seguido da análise de dados (pré-processamento) e finalmente foi submetido à fase de modelagem e inferência, momento no qual são lançados os resultados no *software* Orange Data Mining.

Figura 8. Fluxograma de Trabalho



Fonte: o Autor, 2024.

A primeira etapa deste trabalho consistiu no estudo e compreensão da relação sistêmica entre as variáveis envolvidas nos processos que determinam a qualidade da água de um sistema lótico.

Para realizar essa avaliação, foram pesquisadas diversas fontes de revistas científicas e artigos (AHMED et al., 2019; ANDRÉ LIMA DA COSTA et al., 2022; CARDOSO-SILVA et al., 2014; CORREIA et al., 2008; GARCÍA; MIRANDA, 2018; KOHATSU et al., 2018; KORANGA et al., 2022; PADIAL; POMPÊO; MOSCHINI-CARLOS, 2009; SENDACZ et al., 2005; SERRAGLIO et al., 2021; SHIN et al., 2020; SIERRA, 2013; VIEIRA, 2019), além de teses e dissertações (BARRETO et al., 2013; FERREIRA; CUNHA-SANTINO; BIANCHINI JÚNIOR, 2015; SCARIOT, 2008) que trazem informações sobre a relação que tem alguns parâmetros dos complexos naturais na qualidade da água e como eles podem interferir diretamente na disponibilidade do oxigênio dissolvido (OD), parâmetro fundamental para a avaliação dos níveis de eutrofização e poluição da água.

Posteriormente, foi realizada a coleta de dados nos sites oficiais da SABESP e a NASA, e consequentemente, a estruturação e organização do banco de dados

do sistema natural como um todo. Seguidamente, foi realizada a etapa de pré-processamento dos dados, a qual consistiu na análise do banco de dados antes de realizar o processo de visualização e simulação (aplicação de aprendizado de máquina) no *software* Orange Data Mining.

Uma vez realizada essa atividade, escolheram-se os modelos de aprendizado de máquina mais adequados para realizar a avaliação do sistema natural (modelagem e inferência), e finalmente, realizou-se a simulação no *software* Orange Data Mining para avaliação dos resultados. Os processos estão detalhados nos itens a seguir:

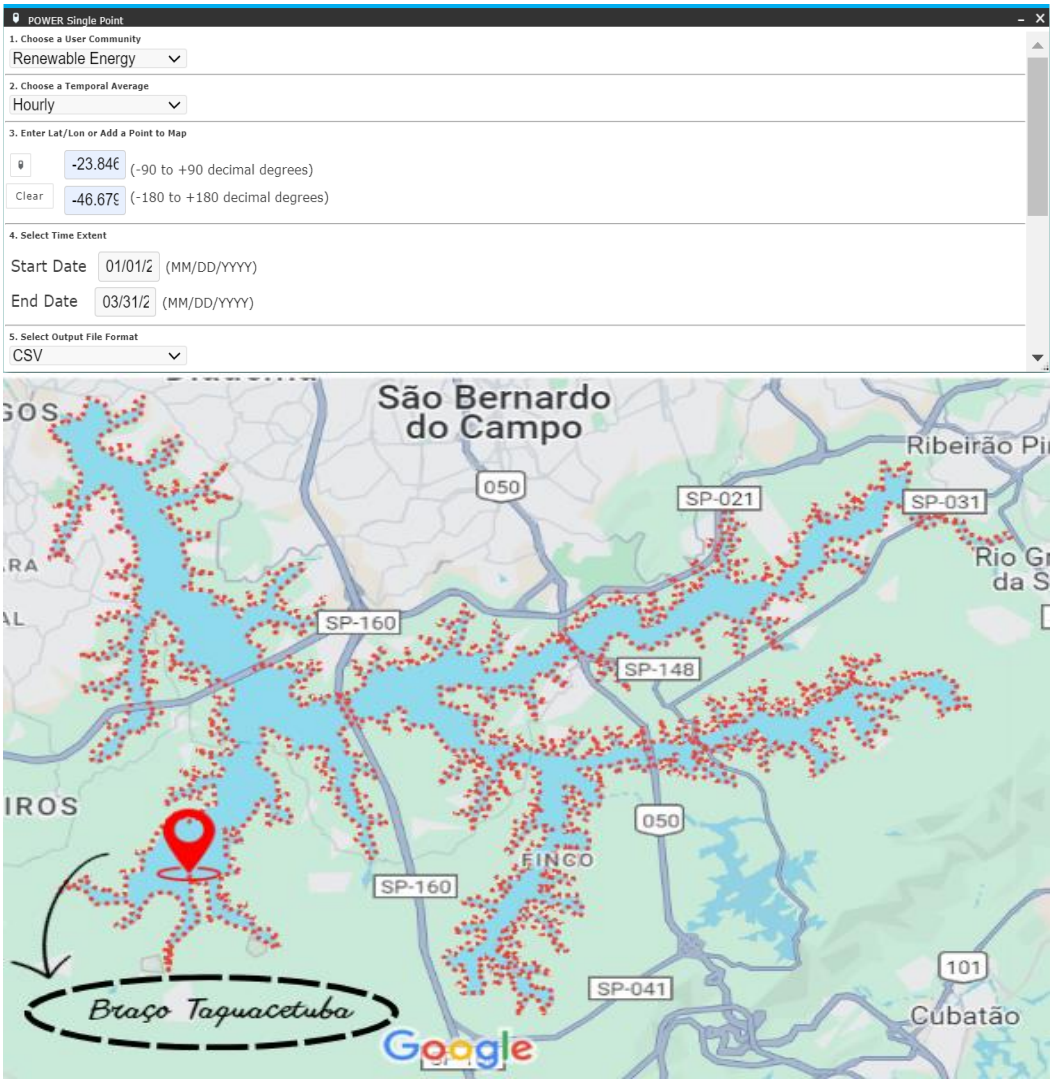
5.2. COLETA DE DADOS

Dados experimentais da qualidade da água do reservatório Billings são divulgados pela SABESP anualmente no seu site oficial (SABESP, 2024). Devido à grande importância do Braço Taquacetuba para o abastecimento e fornecimento de água potável, além da disponibilidade de várias sondas de monitoramento remoto na represa Billings, escolheu-se o Braço Taquaceutaba – "BL 105" como objeto de avaliação. Outro motivo da seleção está relacionado a sua localização geográfica, que ao estar afastada das regiões mais poluídas da represa (CARDOSO-SILVA et al., 2014), permite uma avaliação mais adequada da disponibilidade de OD no sistema hídrico.

Após a revisão bibliográfica relacionada aos parâmetros que afetam a qualidade da água e a disponibilidade de OD nos sistemas hídricos, foram selecionadas as variáveis: condutividade elétrica, oxigênio dissolvido, pH, temperatura e turbidez. Essas variáveis são monitoradas em tempo real todos os dias do ano, a cada hora. A SABESP disponibiliza os dados de monitoramento anuais no seu site apenas no formato de PDF, e devido a essa característica, foi necessária a utilização de um conversor de PDF para Excel ® (formato .csv) para a manipulação de dados.

Ademais, dados ambientais, obtidos a partir de monitoramento por satélite, são divulgados pela NASA, foram selecionados alguns parâmetros de interesse nos sistemas naturais e que são monitorados pela NASA, estes foram: radiação solar, precipitação, velocidade do vento e pressão superficial. Para a obtenção das informações referentes a estes parâmetros, foi necessário primeiramente selecionar a área geográfica de monitoramento no seu site oficial (NASA, 2024). Na 9 ilustram-se as coordenadas e seleção geográfica selecionada para a geração de dados no site da NASA.

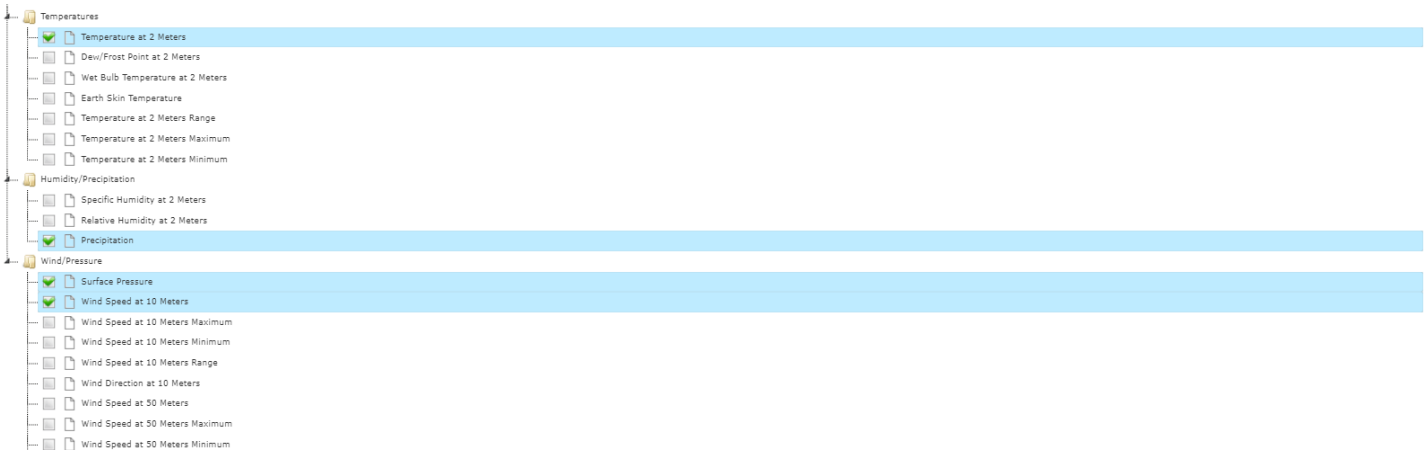
Figura 9. Coordenadas e seleção geográfica – Dados Satélite da NASA



Fonte: NASA; MAPS, 2024

O site da NASA permite a geração do banco de dados de monitoramento por satélite a partir de pontos específicos. No caso da área de interesse, foram selecionadas as coordenadas -23.846, -46.679, região que abrange o braço Taquacetuba da represa Billings. No mesmo navegador, é possível configurar a medida temporal desejada (monitoramento diário, por horas, mensal, etc.), devido à necessidade de juntar os pares ordenados (equivalência de dados SABESP vs NASA), foi selecionada a opção “Hourly”, a qual gerou resultados de hora em hora, além disso, o site também permite a seleção de datas específicas para a geração de dados, no caso, foi gerado o relatório correspondente a um ano inteiro. Os parâmetros selecionados para a geração do relatório anual encontram-se disponíveis na Figura 10.

Figura 10. Parâmetros selecionados para a geração do Banco de Dados Anual – NASA



Fonte: NASA, 2024

O relatório (banco de dados) da NASA permitiu a exportação do relatório no formato .csv no Excel ®, tornando mais simples o tratamento de dados. Uma vez recopilado o conjunto de dados da SABESP e a NASA, os mesmos foram transferidos a uma única planilha (Figura 11) com a finalidade de gerar o banco de dados final. Após a obtenção do banco de dados, os mesmos foram padronizados (de hora em hora) em função do tempo e salvos em formato adequado (.csv) para serem importados no Orange Data Mining. Nesse ponto é importante destacar, que devido às especificidades do *software* Orange Data Mining, a padronização das colunas referentes a tempo foi de indispensável importância, uma vez que o *software* precisa ter intervalos de tempos iguais (mesma quantidade de dados em tempo real) para realizar as predições de forma correta.

Figura 11. Banco de dados SABESP vs NASA

Sistema Remoto de Monitoramento da Qualidade da Água - Represa Billings							National Aeronautics and Space Administration (NASA)						
Captação Taquacetuba - UMR BL 105 (Sonda Superfície)													
Companhia de Saneamento Básico do Estado de São Paulo (SABESP)													
Ano 20XX							Ano 2018 - Região Metropolitana de São Paulo						
Data	Hora	Condutividade (µS/cm)	OD (mg/L)	pH	Temperatura (°C)	Turbidez (NTU)	Data	Hora	All Sky Surface PAR Total (W/m^2)	Temperatura a 2 m (°C)	Precipitação (mm/hora)	Velocidade do vento (m/s)	Pressão superficial (kPa)

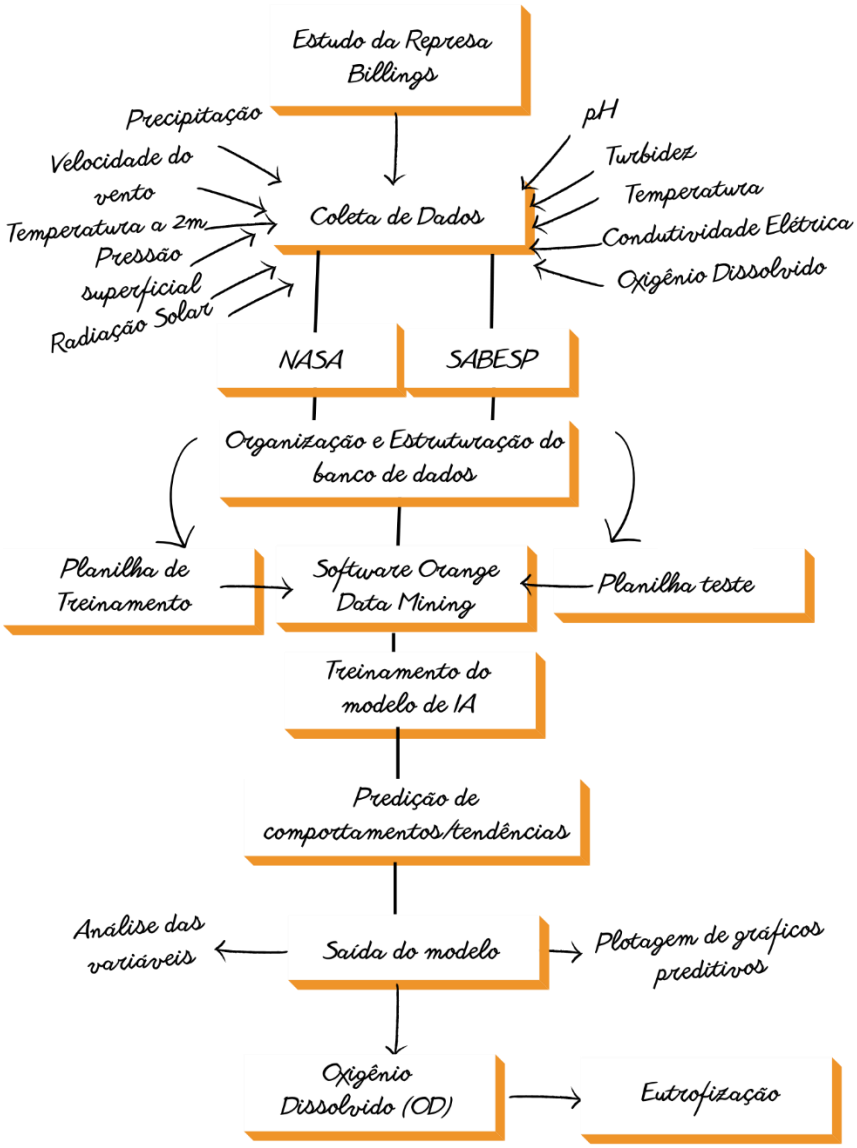
Fonte: o Autor, 2024.

Apesar da SABESP monitorar a qualidade da água no transcurso de um ano, muitas vezes o sensor apresentou falhas ou era submetido a manutenções, motivo pelo qual os dados não se encontravam disponíveis em sua totalidade, já o satélite da NASA foi capaz de obter a totalidade das informações ao longo do tempo, por hora.

Com a finalidade de realizar a análise mais atual das condições e níveis de eutrofização do braço Taquacetuba da represa Billings, selecionou-se o ano 2018 para compor a planilha de treinamento, uma vez que o mesmo possui o monitoramento mais completo nos últimos 5 anos, já as planilhas de teste foram compostas pelo ano anterior (2017) e o ano subsequente (2019).

Devido às particularidades de interpretação do Orange Data Mining, os dados foram convertidos a numéricos e temporários. Na Figura 12, ilustra-se o fluxograma metodológico após a coleta de dados.

Figura 12. Fluxograma de Simulação Orange Data Mining



Fonte: o Autor, 2024.

5.3. AVALIAÇÃO DO BANCO DE DADOS

Um dos "passos" mais críticos para a aplicação de aprendizado de máquina está intrinsecamente relacionado à avaliação dos bancos de dados. Isto é, determinar quais serão as análises estatísticas/lógicas e/ou experimentais que poderão comprovar que o modelo em questão (conjunto de dados) poderá gerar variáveis de saída confiáveis e precisas.

Com o intuito de aproveitar ao máximo as ferramentas disponibilizadas no *software* Orange Data Mining, utilizou-se o coeficiente de correlação de Pearson para avaliar o banco de dados em questão. O coeficiente de correlação de Pearson é uma ferramenta estatística que mede a relação que existe entre duas variáveis continuas, a faixa de correlação encontra-se desde o -1 até 1, onde 0 indica que não há associação entre as variáveis, um valor maior do que zero indica que enquanto aumenta o valor de uma variável, a outra também aumenta e analogamente, um valor menor do que 0 sugere que a associação é negativa, ou seja, à medida que aumenta o valor de uma variável a outra diminui (DISPERSI; PUNTOS, 2015a).

Na área da ciência de dados, o coeficiente de correlação de Pearson é muito utilizado devido ao fato que permite mensurar a relação linear entre duas variáveis, além da sua capacidade de interpretação intuitiva (valores de -1 até 1). No caso da sua aplicação na avaliação dos dados de sistemas naturais, a sua importância radica na capacidade de demonstrar a relação que pode existir entre as variáveis de entrada e de saída , além da sua versatilidade, uma vez que o mesmo é "indiferente" às escalas de medidas utilizadas nas variáveis, ou seja, as diferenças na escala (unidades de medida) não afetam a medida de correlação (HERNÁNDEZ et al., 2018). No quadro 4 ilustram-se as avaliações do coeficiente de correlação de Pearson.

Quadro 4. Avaliação do teste de correlação de Pearson

Faixa de Correlação de Pearson	**Avaliação**
0,0 – 0,2	Correlação fraca ou nula
0,2 – 0,4	Correlação fraca
0,4 – 0,6	Correlação moderada
0,6 – 0,8	Correlação forte
0,8 – 1,0	Correlação muito forte

Fonte: Adaptado de (DISPERSI; PUNTOS, 2015b)

Nesse sentido, e em base às características especificas e intuitivas desse coeficiente, foi realizada a aplicação do mesmo na avaliação dos resultados simulados no *software* Orange Data Mining.

No *software* Orange Data Mining, utilizou-se o *widget* "Correlations" (Figura 13) com a finalidade de visualizar o coeficiente de correlação de Pearson de todas as variáveis avaliadas no sistema natural e a variável *target* (objetivo) – oxigênio dissolvido.

Figura 13. *Widget* de Correlações no *software* Orange Data Mining

Fonte: o Autor, 2024.

O *widget* permite calcular os coeficientes de correlação de Pearson ou Spearman para todos os pares ordenados num conjunto de dados. A principal aplicação do mesmo teve como objetivo conhecer o comportamento (coeficiente de correlação) dos parâmetros avaliados na simulação com a variável *target*.

Assim, variáveis cujo valor fosse negativo eram submetidas ao "*skip*", isto com o intuito de diminuir a quantidade de parâmetros da SABESP requeridos para obter uma resposta satisfatória na simulação, tendo como argumento principal o comportamento linear de cada um dos parâmetros em relação à variável objetivo (OD), portanto, variáveis cujo valor fosse positivo, eram submetidas a uma avaliação de "tentativa e erro", metodologia na qual as variáveis eram submetidas ao "*skip*" conforme seu fator de correlação de Pearson diminuir.

A decisão de avaliar o coeficiente de correlação de Pearson foi apenas como um argumento de exclusão de variáveis na busca de estabelecer a menor quantidade de parâmetros da SABESP necessários para simular o modelo, além de conhecer de melhor forma o perfil do banco de dados avaliado em questão.

Devido à natureza não linear da dinâmica dos sistemas naturais, estando caracterizado pelo comportamento senoidal entre o dia e a noite e cossenoidal ao longo do ano na formação das estações do ano (SCARIOT, 2008), o coeficiente de correlação de Pearson não é adequado na avaliação do desempenho da simulação e/ou no modelo de aprendizado de máquina em sistemas não lineares (DUFERA; LIU; XU, 2023).

5.4. BANCO DE DADOS PLANILHA DE TREINAMENTO

Um dos principais desafios que abrange a criação de banco de dados para a aplicação de aprendizado de máquina está relacionado à recopilação, unificação e preparação de dados (estruturação), o principal motivo disso está relacionado a que embora existam muitos dados disponíveis, nem todos são críticos ou permitem avaliar com precisão o sistema analisado em questão (ALVES, 2022). Além disso, muitas vezes são coletados dados que apresentam um desvio abrupto no sistema e que não condizem com a realidade.

Para realizar as aplicações de aprendizado de máquina, de forma mais precisa e concisa, requer-se um amplo conjunto de dados constantes e que não possuam muitos ruídos (erros). Essa característica tornou o complexo Billings ideal para realizar a aplicação das técnicas de aprendizado de máquina, uma vez que as sondas monitoram em tempo real os parâmetros e condições que apresenta a represa Billings e em consequência seus braços ao longo do ano.

Para a estruturação do banco de dados da planilha de treinamento, foi utilizado o *software* Excel ®. O banco de dados utilizado na planilha de treinamento foi do ano 2018 devido a estar caracterizado pela maior quantidade de dados no percorrer do ano. A planilha de treinamento está composta por 8760 colunas e 12 filas de dados contendo as informações tanto da SABESP quanto da NASA. Todas as informações foram compostas por dados capturados em tempo real a cada hora durante os 365 dias do ano do braço Taquacetuba da Represa Billings. No quadro 5 ilustram-se os dados disponibilizados no banco de dados.

Quadro 5. Dados - Planilha de Treinamento

Dados SABESP	Unidade	Dados NASA	Unidade
Condutividade	µS/cm	Radiação fotossintética	(W/m²)
Oxigênio Dissolvido	mg/L	Temperatura a 2 m	°C
pH	-	Precipitação	mm/h
Temperatura	°C	Velocidade do Vento	m/s
Turbidez	NTU	Pressão Superficial	kPa

Fonte: o Autor, 2024.

A escolha dos dados (parâmetros) disponibilizados na planilha de treinamento foi realizada a partir de uma avaliação prévia de referências bibliográficas confiáveis que destacam a relação entre as variáveis em questão e a disponibilidade de oxigênio dissolvido no meio hídrico.

A importância de ter um banco de dados anual esteve intrinsicamente relacionada a um dos objetivos, o qual era a verificação de que o algoritmo pode interpretar ou reconhecer as mudanças sazonais, assim como as mudanças do comportamento do oxigênio dissolvido durante o dia e a noite. Apesar, que o ano 2018 possuir a maior quantidade de dados do ano encontrou-se uma faixa na qual a sonda apresentou falhas, esse período foi do dia 06/03/2018 até o dia 12/04/2018. Devido a essa "perda de dados", foram inseridos, nesse trecho faltante, dados do ano 2022, com a finalidade de diminuir a quantidade de "ruídos" que poderia acontecer com a falta dessas informações, além de melhorar o sistema de treinamento no Orange Data Mining.

Na planilha de treinamento realizaram-se as respectivas classificações relacionadas à "natureza" do dado. No *software* Excel ® foram caracterizadas as informações e respectivamente classificadas conforme apresentado no Quadro 6.

Quadro 6. Classificação da "natureza" das variáveis em Excel ®

Variável	Classificação no Excel ®
Data	Data
Hora	Data
Condutividade	Inteiro
OD	Real
pH	Real
Temperatura	Real
Turbidez	Real

Radiação Fotossintética	Real
Precipitação	Real
Velocidade do Vento	Real
Pressão Superficial	Real

Fonte: o Autor, 2024.

A importância de realizar essa classificação no *software* Excel ® está relacionada à interpretação que o Orange tem que dar a cada variável. Apesar, que no *software* Orange Data Mining é possível realizar a "conversão de domínio", ou seja, alterar a interpretação das células, a realização desse passo prévio permite evitar erros ou *bugs* quando a simulação é realizada.

Em vista de que o modelo de simulação é do tipo "caixa preta", cuja principal característica é a descrição das relações funcionais entre as variáveis de entrada e saída, formatou-se a planilha de treinamento (Figura 14), de forma semelhante ao que foi a planilha teste. Assim, ao ter as mesmas variáveis disponíveis de entrada e saída tanto na planilha de treinamento quanto na planilha teste (diferente ano), foi possível avaliar a menor quantidade de parâmetros que permitem estimar oxigênio dissolvido de forma precisa e com comportamento semelhante às condições reais, tanto em termos diurnos e noturnos quanto em termos sazonais.

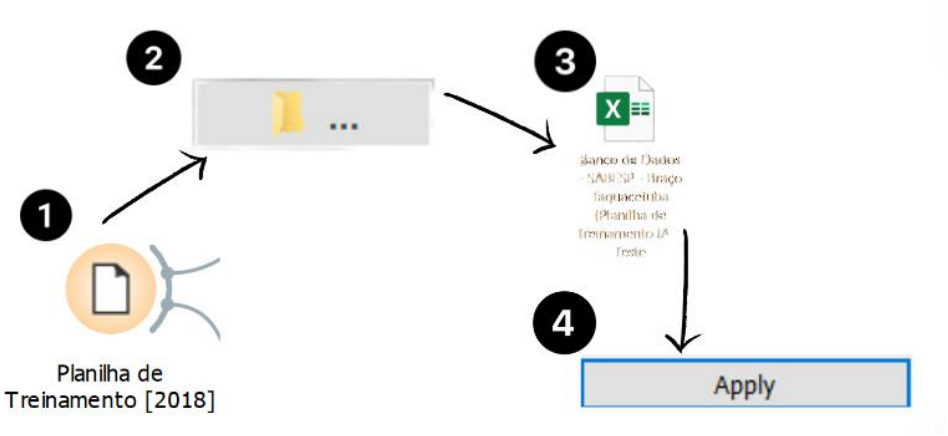
Figura 14. Planilha de Treinamento (Ano 2018)

Data	Hora	Condutividade (µS/cm)	OD (mg/L)	pH	Temperatura (°C)	Turbidez (NTU)	All Sky Surface PAR Total (W/m²)	Temperatura a 2 m (°C)	Precipitação (mm/hora)	Velocidade do vento (m/s)	Pressão superficial (kPa)
01/01/2018	0:00:00	179,00	7,22	8,10	24,60	20,50	0,00	21,74	0,04	2,11	95,30
01/01/2018	1:00:00	179,00	7,17	8,10	24,50	21,00	0,00	21,58	0,04	2,52	95,23
01/01/2018	2:00:00	179,00	6,97	8,00	24,50	21,00	0,00	21,40	0,04	2,69	95,17
01/01/2018	3:00:00	179,00	7,22	8,10	24,50	21,80	0,00	21,25	0,04	2,70	95,15
01/01/2018	4:00:00	178,00	7,33	8,10	24,50	25,20	0,00	21,09	0,04	2,62	95,15
01/01/2018	5:00:00	178,00	7,23	8,10	24,50	24,00	4,95	21,11	0,04	2,65	95,19
01/01/2018	6:00:00	179,00	7,18	8,00	24,50	23,80	53,10	22,26	0,06	3,52	95,23
01/01/2018	7:00:00	179,00	7,32	8,10	24,50	24,40	103,80	23,71	0,14	4,15	95,26
01/01/2018	8:00:00	178,00	7,66	8,20	24,60	24,20	166,88	25,59	0,24	3,91	95,29
01/01/2018	9:00:00	178,00	7,90	8,30	24,80	21,50	221,15	27,15	0,01	3,25	sy61
01/01/2018	10:00:00	178,00	7,88	8,30	24,80	23,40	312,52	28,35	0,01	2,72	95,28
01/01/2018	11:00:00	179,00	7,93	8,40	24,90	22,10	274,23	29,05	0,01	2,50	95,26
01/01/2018	12:00:00	179,00	7,66	8,30	24,90	21,20	344,67	29,30	0,01	2,45	95,21
01/01/2018	13:00:00	180,00	7,46	8,20	25,00	17,70	365,80	29,47	0,01	2,20	95,14
01/01/2018	14:00:00	180,00	7,25	8,10	25,00	17,40	323,85	29,45	0,01	1,89	95,07
01/01/2018	15:00:00	180,00	7,70	8,30	25,10	19,10	275,77	29,29	0,01	1,63	95,03
01/01/2018	16:00:00	180,00	7,92	8,40	25,10	20,50	190,85	28,67	0,01	1,41	95,02
01/01/2018	17:00:00	179,00	8,17	8,50	25,10	22,00	93,67	27,59	0,01	1,50	95,04
01/01/2018	18:00:00	179,00	8,23	8,50	25,00	21,70	15,23	26,05	0,00	2,16	95,07
01/01/2018	19:00:00	178,00	9,48	8,90	25,50	30,60	0,00	24,61	0,00	2,60	95,10
01/01/2018	20:00:00	177,00	9,55	9,00	25,70	27,30	0,00	24,12	0,00	2,53	95,13
01/01/2018	21:00:00	177,00	9,30	8,90	25,60	24,10	0,00	23,85	0,00	2,34	95,15
01/01/2018	22:00:00	177,00	8,43	8,50	25,30	23,90	0,00	23,32	0,01	2,23	95,13
01/01/2018	23:00:00	178,00	8,09	8,40	25,20	23,60	0,00	22,96	0,05	1,91	95,06
02/01/2018	0:00:00	178,00	8,11	8,50	25,10	21,90	0,00	22,90	0,55	1,77	94,99

Fonte: o Autor, 2024.

Após a estruturação do banco de dados da planilha de treinamento, procedeu-se à importação da mesma no *software* Orange Data Mining com a finalidade de realizar as configurações preliminares. Para realizar a importação da planilha de treinamento foi utilizado o *widget* "File" (Figura 15). Após a inserção do *widget* (1), realizou-se o carregamento ou *upload* (2) do arquivo .csv (3) referente à planilha de treinamento, finalmente, foi finalizado o processo de "carregamento" utilizando a função "Apply" (4).

Figura 15. Processo de importação da planilha de treinamento



Fonte: o Autor, 2024.

Uma vez realizado este processo, o *software* Orange Data Mining disponibilizou as informações contidas no arquivo .csv (Figura 16). Esse *widget* classifica as variáveis como: "categorical" quando se refere a uma categoria quaisquer (em geral de classificação), "numerical" quando a variável é interpretada como um número, "text" quando a mesma é apenas interpretada como texto, e "datetime" quando a célula traz informações relacionadas a tempo de forma geral.

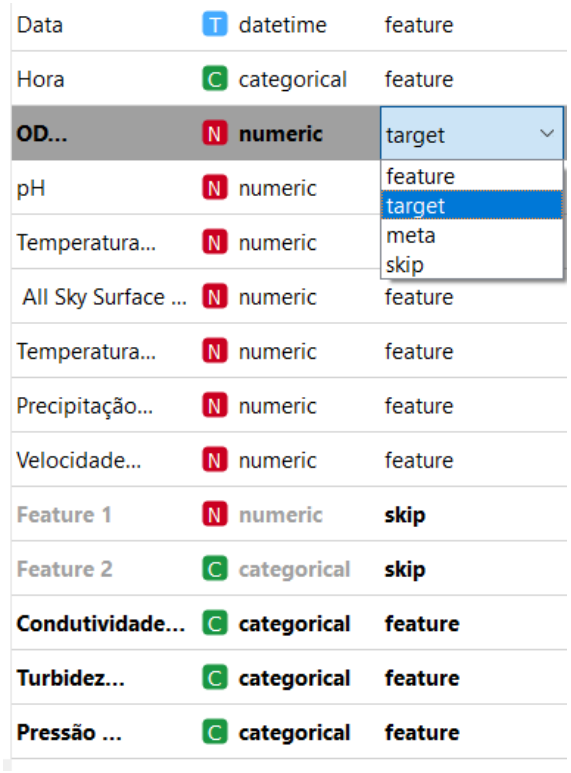
Figura 16. *Layout* do *widget* "File" no carregamento da planilha de treinamento

	Name	Type
1	Data	T datetime
2	Hora	C categorical
3	OD...	N numeric

Fonte: o Autor, 2024.

Além disso, o *widget* classifica as variáveis como “Feature” equivalente à variável preditora, “Target” referente à identificação da variável alvo, “Meta” como variáveis de cadeia e “Skip” caso a informação carregada não seja de interesse ou precise ser “ignorada” na simulação. No caso da configuração da planilha de treinamento, foi realizada a configuração conforme apresentado na Figura 17.

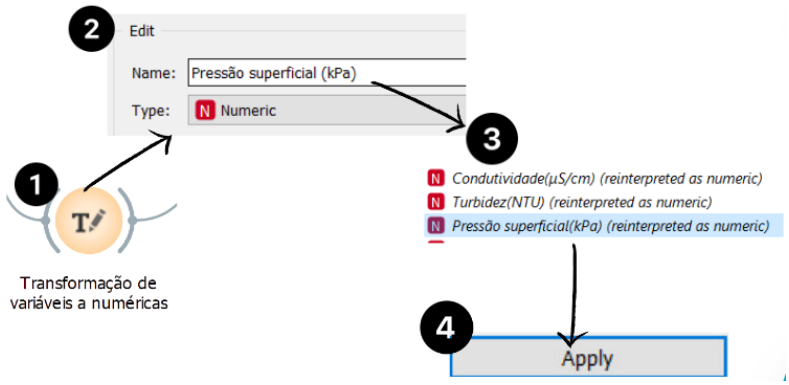
Figura 17. Configuração *widget* “File” – Planilha de Treinamento



Fonte: o Autor, 2024.

As variáveis do tipo tempo foram configuradas como “datetime” e todos os outros parâmetros numéricos foram classificados como “numeric”, variáveis ou células que o Orange Data Mining interpreta como “vazio” foram configuradas como “skip” com o intuito de ignorá-las na simulação. Para ajustar a interpretação do *software* das variáveis condutividade, turbidez e pressão superficial como numéricas, foi necessária a utilização do *widget* “Edit domain” (Figura 18).

Figura 18. Edição de domínio de variáveis da planilha de treinamento



Fonte: o Autor, 2024.

Para utilizar o *widget* "Edit domain" foi simplesmente ligado ao "File" que continha a planilha de treinamento e em consequência o banco de dados de treinamento (1), seguidamente, foi clicado acima do parâmetro que possuía o problema de interpretação (2) e na sequência foi alterado o domínio para "Numeric", após isso, o parâmetro mostra que ele foi reinterpretado como número (3) e após realizar esse processo para todos os parâmetros com a mesma problemática foi executada a função "Apply" (4) para salvar as mudanças no arquivo. O *widget* de "Data Table" foi acoplado ao "File" com a finalidade de visualizar os dados contidos no arquivo .csv importado no Orange.

Em vista que segundo a ciência de dados, quanto maior seja o volume de dados, melhor será a precisão na simulação (DA COSTA FILHO et al., 2019), na configuração do *widget* "File" escolheram-se todas as variáveis como "recursos" (excetuando o oxigênio dissolvido, a qual era o target), com o intuito de "treinar" com maior qualidade o algoritmo, permitindo estabelecer e/ou interpretar todas as possíveis relações que existem entre as variáveis num sistema hidrográfico e seu impacto com a variável objetivo.

5.5. BANCO DE DADOS DE TESTE

O processo de montagem das planilhas de teste foi semelhante à planilha de treinamento. Entretanto, as informações contidas nelas, não foram correspondentes ao ano da planilha de treinamento (ano 2018). As planilhas de teste foram conformadas pelo ano 2017 (ano anterior) e o ano 2019 (ano subsequente). O banco de dados referente a esses dois anos, apesar de possuírem uma quantidade de dados menor do que o ano 2018, estava

caracterizado pela disponibilidade de um amplo volume de dados de vários parâmetros, entre eles o OD, o qual permitiu, uma comparação abrangente do comportamento dinâmico do sistema real e o simulado.

A planilha de teste do ano 2017 (Figura 19) foi conformada com as informações dos parâmetros: Condutividade, pH, ORP, temperatura, turbidez, clorofila, radiação fotossintética, precipitação, velocidade do vento e pressão superficial. Essa planilha possuía um total de 7783 linhas e 13 colunas. Numa análise generalizada, o ano 2017 esteve caracterizado por possuir bastante falhas na sonda, impedindo ter a coleta de dados de vários meses no transcurso do ano, sobretudo dos parâmetros de pH e turbidez.

Porém, mesmo com a ausência desses dados, encontraram-se disponíveis informações que contemplam as quatro (4) estações do ano, permitindo então realizar a comparação entre os dados reais e os dados simulados.

Figura 19. Planilha de teste (Ano 2017)

Data	Hora	Condutividade (µS/cm)	pH	ORP (mV)	Temperatura (°C)	Turbidez (NTU)	Clorofila (µg/L)	All Sky Surface PAR Total (W/m²)	Temperatura a 2 m (°C)	Precipitação (mm/hora)	Velocidade do vento (m/s)	Pressão superficial (kPa)
1/1/2017	0:00:00	187,0	0,0	0,0	27,4	0,0	4,6	0,0	23,2	0,0	2,6	95,2
1/1/2017	1:00:00	188,0	0,0	0,0	27,3	0,0	6,4	0,0	22,9	0,0	2,7	95,2
1/1/2017	2:00:00	186,0	0,0	0,0	27,3	0,0	7,0	0,0	22,5	0,0	2,7	95,4
1/1/2017	3:00:00	187,0	0,0	0,0	27,3	0,0	6,9	0,0	22,3	0,0	2,6	95,4
1/1/2017	4:00:00	185,0	0,0	0,0	27,4	0,0	5,6	0,0	22,1	0,0	2,6	95,3
1/1/2017	5:00:00	187,0	0,0	0,0	27,3	0,0	6,6	4,2	22,1	0,0	2,5	95,3
1/1/2017	6:00:00	188,0	0,0	0,0	27,3	0,0	6,5	57,5	23,0	0,0	2,7	95,4
1/1/2017	7:00:00	188,0	0,0	0,0	27,2	0,0	6,1	148,1	24,0	0,0	3,0	95,5
1/1/2017	8:00:00	187,0	0,0	0,0	27,2	0,0	6,1	255,8	26,0	0,1	3,3	95,5
1/1/2017	9:00:00	187,0	0,0	0,0	27,3	0,0	7,1	358,5	28,2	0,0	3,8	95,6
1/1/2017	10:00:00	187,0	0,0	0,0	27,4	0,0	7,3	432,7	29,7	0,0	4,1	95,6
1/1/2017	11:00:00	188,0	0,0	0,0	27,6	0,0	6,4	478,1	30,6	0,0	4,2	95,6
1/1/2017	12:00:00	189,0	0,0	0,0	27,8	0,0	7,0	469,0	31,1	0,0	4,2	95,6
1/1/2017	13:00:00	189,0	0,0	0,0	27,8	0,0	5,7	388,1	31,3	0,0	4,2	95,5
1/1/2017	14:00:00	189,0	0,0	0,0	27,8	0,0	5,9	341,4	31,0	0,0	4,1	95,4
1/1/2017	15:00:00	189,0	0,0	0,0	27,7	0,0	6,9	244,5	30,4	0,0	3,9	95,4
1/1/2017	16:00:00	189,0	0,0	0,0	27,7	0,0	6,0	124,6	29,4	0,1	3,2	95,3
1/1/2017	17:00:00	189,0	0,0	0,0	27,6	0,0	5,8	64,5	28,0	0,1	2,2	95,3
1/1/2017	18:00:00	188,0	0,0	0,0	27,4	0,0	6,1	5,2	26,3	0,2	1,6	95,3
1/1/2017	20:00:00	191,0	0,0	0,0	27,8	0,0	6,0	0,0	25,1	0,4	0,9	95,4
1/1/2017	21:00:00	191,0	0,0	0,0	27,8	0,0	6,4	0,0	24,6	0,5	0,1	95,4
1/1/2017	22:00:00	189,0	0,0	0,0	27,8	0,0	5,8	0,0	24,2	0,6	1,4	95,5
1/1/2017	23:00:00	188,0	0,0	0,0	27,7	0,0	6,3	0,0	23,7	0,5	2,5	95,5

Fonte: o Autor, 2024.

Analogamente à planilha de teste do ano 2017, a planilha de teste do ano 2019 (Figura 20), foi conformada pelos parâmetros de monitoramento da água: Condutividade, pH, temperatura, turbidez, radiação fotossintética, precipitação, velocidade do vento e pressão superficial. Esse banco de dados possuía 5634 linhas e 11 colunas. Uma das características desse banco de dados foi a ausência do mês de agosto e algumas semanas do mês de julho, porém, semelhante à planilha de teste do ano 2017, ela abrange as quatro (4) estações do ano.

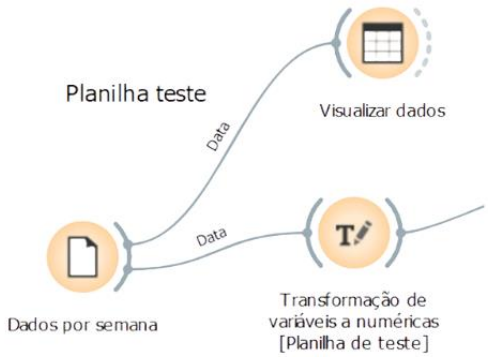
Figura 20. Planilha de teste (Ano 2019)

Data	Hora	Condutividade (µS/cm)	pH	Temp. (°C)	Turbidez (NTU)	All Sky Surface PAR Total (W/m²)	Temperatura a 2 m (°C)	Precipitação (mm/hora)	Velocidade do vento (m/s)	Pressão superficial (kPa)
1/1/19	0:00:00	190,00	9,80	27,60	8,20	0,00	22,43	0,01	3,28	95,63
1/1/19	1:00:00	189,70	9,70	27,50	8,40	0,00	22,19	0,01	3,00	95,61
1/1/19	2:00:00	189,20	9,60	27,10	7,90	0,00	21,98	0,00	2,76	95,55
1/1/19	3:00:00	189,80	9,60	26,70	7,60	0,00	21,70	0,00	2,48	95,47
1/1/19	4:00:00	189,10	9,70	27,10	8,10	0,00	21,51	0,00	2,37	95,41
1/1/19	5:00:00	188,80	9,70	27,10	7,80	0,00	21,33	0,00	2,26	95,41
1/1/19	6:00:00	189,10	9,70	27,00	8,00	0,00	21,23	0,00	2,17	95,47
1/1/19	7:00:00	189,00	9,70	26,80	8,10	0,00	21,15	0,00	2,11	95,55
1/1/19	8:00:00	189,40	9,70	26,70	9,10	5,73	21,34	0,00	1,99	95,63
1/1/19	9:00:00	189,20	9,80	27,00	11,10	66,90	22,90	0,00	2,39	95,70
1/1/19	10:00:00	189,20	9,80	27,20	8,60	155,80	24,51	0,00	2,18	95,74
1/1/19	11:00:00	188,70	9,70	27,00	10,30	250,02	26,62	0,00	1,67	95,74
1/1/19	12:00:00	189,20	9,70	26,80	9,10	289,23	28,58	0,01	1,03	95,73
1/1/19	13:00:00	190,80	9,90	28,10	10,60	342,40	30,04	0,01	0,74	95,70
1/1/19	14:00:00	192,80	9,90	28,80	9,30	367,38	30,90	0,03	1,41	95,64
1/1/19	15:00:00	192,20	9,90	28,80	9,60	338,27	31,17	0,11	2,55	95,58
1/1/19	16:00:00	192,30	9,90	28,80	9,60	232,88	30,60	0,26	4,23	95,54
1/1/19	17:00:00	192,20	9,80	28,70	9,40	191,38	29,12	0,34	5,22	95,52
1/1/19	18:00:00	191,50	9,80	28,70	9,30	94,42	27,59	0,37	5,34	95,53
1/1/19	19:00:00	191,20	9,80	28,60	8,30	66,40	26,20	0,46	5,01	95,55
1/1/19	20:00:00	190,90	9,70	28,00	9,60	36,45	24,97	0,47	4,39	95,60
1/1/19	21:00:00	191,00	9,80	27,90	8,00	11,38	23,78	0,32	3,49	95,67
1/1/19	22:00:00	190,80	9,70	27,50	8,00	0,00	23,03	0,24	2,61	95,74
1/1/19	23:00:00	190,90	9,80	27,30	8,10	0,00	22,83	0,17	2,27	95,79
2/1/19	0:00:00	190,60	9,80	27,10	7,80	0,00	22,77	0,12	2,15	95,85

Fonte: o Autor, 2024.

Durante o processo de preparação das planilhas teste, foram excluídas as informações correspondentes à variável objetivo (target), estando então sem informações referentes ao oxigênio dissolvido. Para a montagem da planilha de teste no *software* Orange Data Mining, de forma similar à planilha de treinamento. Na Figura 21 encontra-se ilustrado os *widgets* que conformaram o processo de visualização e simulação de dados do Orange. Primeiramente, foi utilizado o *widget* "File", o qual foi conectado aos *widgets* "Data Table" e "Edit Domain". O *widget* "Data Table" permite visualizar os dados contidos no arquivo .csv no próprio *software* Orange, já o *widget* "Edit Domain", conforme apresentado anteriormente permite a transformação e/ou reinterpretação das variáveis "não lidas corretamente" pelo Orange, para um domínio de interesse.

Figura 21. *Widgets* utilizados nas planilhas de treinamento



Fonte: o Autor, 2024.

As configurações do *widget* "File" encontram-se disponibilizadas na

Figura 22. Assim como no caso da planilha de treinamento, a planilha teste também foi acoplada ao *widget* "Correlations" com a finalidade de avaliar o coeficiente de correlação de Pearson. Similar à planilha de treinamento, o *software* Orange teve erros ao interpretar a natureza das variáveis (numéricas, temporárias ou categóricas), entretanto, foi ajustado com o *widget* "Edit Domain".

Figura 22. Configuração widget "File" - Planilha teste

	Name	Type	Role
1	Hora	T datetime	feature
2	Condutividade...	N numeric	**skip**
3	pH	N numeric	feature
4	Temperatura...	N numeric	**skip**
5	All Sky Surface ...	N numeric	feature
6	Temperatura...	N numeric	feature
7	Precipitação...	N numeric	feature
8	Velocidade...	N numeric	feature
9	Pressão ...	N numeric	feature
10	Feature 1	N numeric	**skip**
11	Feature 2	N numeric	**skip**
12	Feature 3	C categorical	**skip**
13	**Data**	C **categorical**	**feature**
14	ORP...	C categorical	**skip**
15	Turbidez...	C categorical	**skip**
16	Clorofila...	C categorical	**skip**

Fonte: o Autor, 2024.

Tendo em vista que um dos intuitos na análise do comportamento dinâmico de oxigênio dissolvido na represa Billings está relacionado à melhoria continua e barateamento do processo de monitoramento da qualidade da água, a planilha de teste passou por um processo de avaliação após inferência, ou seja, conforme eram avaliados os coeficientes de correlação de Pearson e o comportamento dinâmico do sistema (comparação da dinâmica real vs simulada), algumas variáveis da SABESP eram configuradas para ser ignoradas (skip). Variáveis cujo coeficiente de correlação de Pearson for alto, eram mais susceptíveis (em termos de argumentos) a estabelecer uma predição mais adequada do que

as variáveis que apresentaram comportamento não linear.

O objetivo principal de aplicar o “skip” nas variáveis da SABESP foi compreender quais eram as variáveis (lineares) mais críticas que permitiam a avaliação do sistema natural, ou seja, possibilitou estabelecer a simulação com a menor quantidade de variáveis da SABESP, contudo que permitisse simular adequadamente o comportamento do sistema lótico.

A razão pela qual foram excluídos apenas os dados da SABESP nas simulações é justificada pelo fato de que os dados da SABESP ao serem obtidos a partir de sondas, as quais representam altos custos de aquisição e de operação para diversas empresas de saneamento brasileiras, enquanto que os dados da NASA são dados obtidos por imagens de satélite e disponibilizados gratuitamente. Portanto, um dos objetivos está em diminuir o custo operacional dos sistemas de monitoramento existentes, além de permitir que outros sistemas hídricos, que não possuem sondas de monitoramento, tenham a possibilidade de obter informações de oxigênio dissolvido a partir de dados de monitoramento por satélite.

As configurações das duas planilhas de teste (2017 e 2019) foram realizadas da mesma forma, caracterizadas pelas mesmas configurações e mesmos processos de importação de arquivo, assim como a mesma quantidade de *widgets* acoplados na simulação. As únicas especificidades tiveram relação à quantidade de parâmetros disponíveis nos dois anos.

5.6. MODELO DE APRENDIZADO DE MÁQUINA

A próxima etapa tratou-se da escolha do modelo ideal de aprendizado de máquina. Para isso, foram empregados os próprios modelos disponíveis no *software* Orange Data Mining.

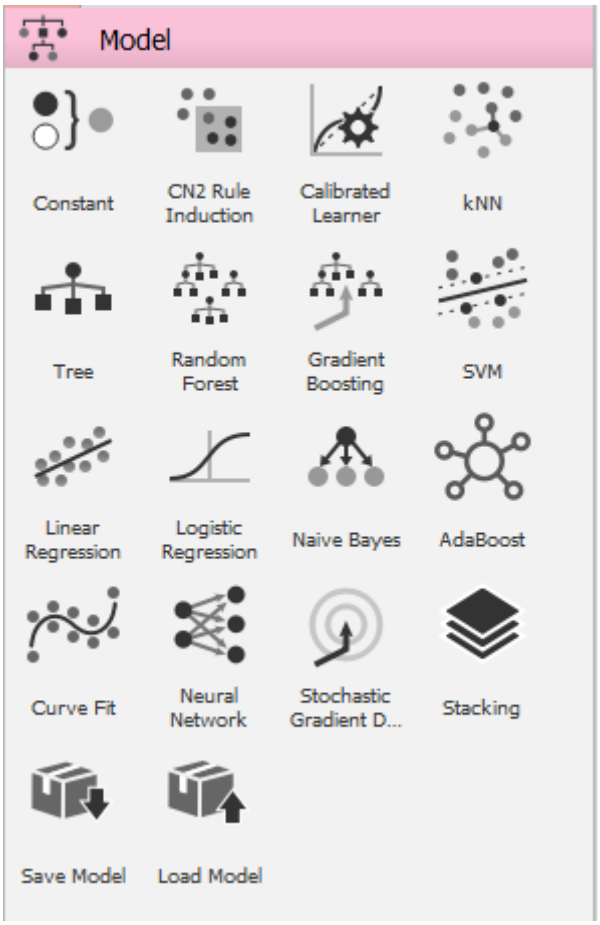
Modelos de caixa preta são comuns de ser utilizados na avaliação de sistemas caracterizados por um alto nível de complexidade, é dizer, aqueles que possuem uma ampla variedade de variáveis de entrada e cujas variáveis de saída podem estar relacionadas entre si (ALVES; ANDRADE, 2022). Em termos mais específicos, um modelo de “caixa preta” na área de aplicação de aprendizado de máquina refere-se a um modelo que é caracterizado por gerar resultados sem uma “explicação” intuitiva de como foram obtidos os resultados.

Em geral, os modelos de caixa preta são processados por complexas

demonstrações e deduções matemáticas que são empregadas na inferência de um algoritmo "adequado" para o sistema estudado. No caso da simulação de sistemas naturais, o modelo de caixa preta permite abranger as possibilidades existentes e/ou relações entre os distintos parâmetros que compõem um sistema hídrico e no caso da represa Billings, deduzir, como eles podem permitir realizar simulações da disponibilidade e/ou quantidade de oxigênio dissolvido no sistema lótico.

Para a aplicação da técnica de aprendizado de máquina e consequentemente simulação do sistema hídrico da represa Billings, foi utilizado o *software* Orange Data Mining. O objetivo principal na simulação é avaliar a capacidade de diferentes modelos de aprendizado de máquina disponíveis no *software* na simulação de variáveis, nesse caso, a variável *target*, ou seja, o OD. Na Figura 23 ilustram-se os modelos de aprendizado de máquina disponíveis no *software* Orange Data Mining.

Figura 23. Modelos de aprendizado de máquina no Orange Data Mining



Fonte: o Autor, 2024.

Apesar de existir uma ampla variedade de modelos disponíveis, foram escolhidos apenas dois modelos para realizar a avaliação, sendo o modelo de “Linear Regression” ou regressão linear e o modelo de “Random Forest” ou floresta aleatória.

O motivo da seleção desses dois modelos para a execução da simulação esteve relacionado a sua facilidade de implementação (ampla variedade de bibliotecas no Orange) , interpretabilidade (simples) e treinamento rápido (tempos de treinamento mais curtos, além de estudos prévios de aprendizado de máquina que destacam o modelo floresta aleatória como um dos mais adequados para a simulação de sistemas hidrográficos (KORANGA et al., 2022).

Na navegação do sistema de modelos (*models*) no *software* Orange, é possível visualizar a definição básica de todos os modelos disponíveis. No Quadro 7, ilustra-se a descrição de cada um dos modelos de aprendizado de máquina empregados na simulação do sistema dinâmico do braço Taquacetuba da represa Billings.

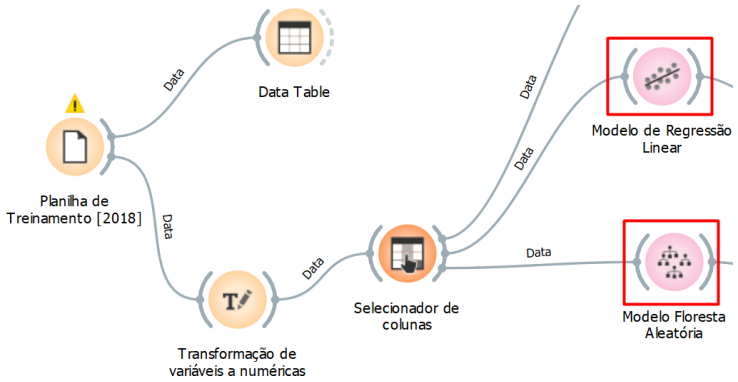
Quadro 7. Descrição dos modelos de aprendizado de máquina utilizados na simulação

Ícone	Modelo	Descrição
	Regressão Linear	O *widget* de regressão linear construí uma predição que simula uma função linear a partir dos dados de entrada.
	Floresta Aleatória	O *widget* do modelo floresta aleatória desenvolve amostras a partir dos dados de treinamento, extraindo um subconjunto arbitrário de atribuições que permitem a escolha do “melhor” atributo para a divisão.

Fonte: o Autor, 2024.

Os *widgets* referentes aos modelos no *software* Orange, foram acoplados ou “conectados” ao banco de dados de treinamento (Figura 24), o qual continha todas as informações do ano 2018. O *widget* “select columns” permitiu a visualização dos dados de processamento após a conversão do domínio de alguns parâmetros (reinterpretação dos dados para numéricos).

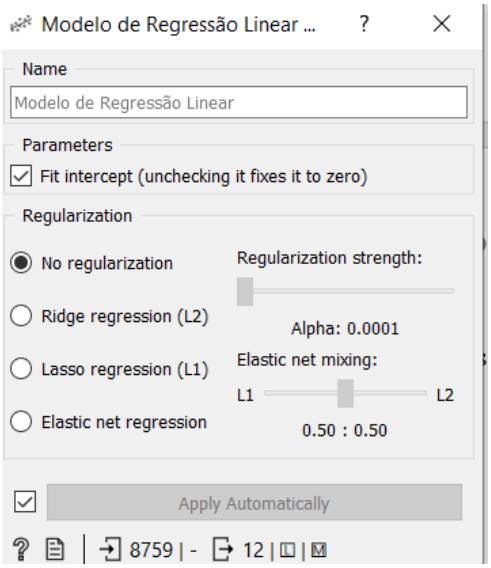
Figura 24. Modelos utilizados na simulação do sistema lótico



Fonte: o Autor, 2024.

Na sequência, foram realizadas as configurações dos modelos tendo em conta às especificidades dos mesmos e do sistema como um todo. Para o caso do modelo de regressão linear (Figura 25), as configurações foram "padrão", ou seja, pré-estabelecidas pelo modelo em si. A razão pela qual essas configurações se adequavam no modelo empregado, estava relacionada ao fato de que é feita apenas uma avaliação do comportamento linear entre as variáveis, sendo então, desnecessário a regularização da reta ou aplicar métodos que promovam essa normalização. A única configuração de interesse foi o ajuste da interceptação da reta.

Figura 25. Configurações do modelo de regressão linear

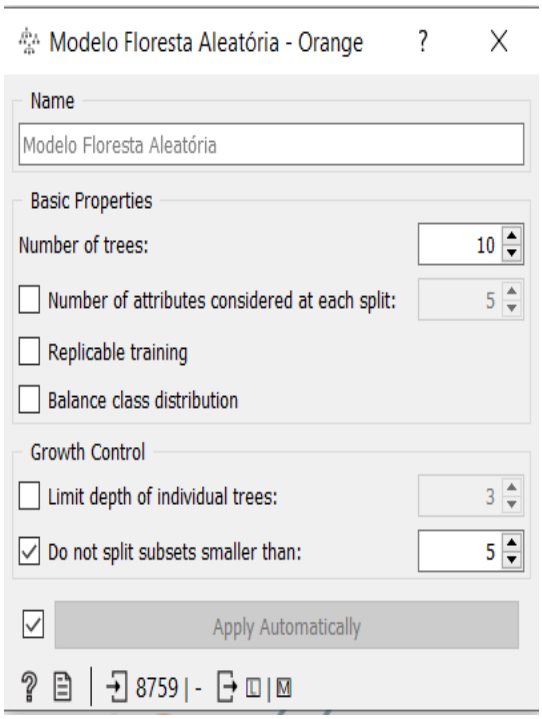


Fonte: o Autor, 2024.

A complexidade dos sistemas naturais pode alterar o desempenho na

simulação dependendo do modelo utilizado (KORANGA et al., 2022). Assim, modelos lineares podem "não ter um ajuste adequado" quando são avaliadas as relações que podem existir entre os parâmetros do sistema hidrográfico. Já para o caso da configuração do *widget* do modelo de floresta aleatória (Figura 26), foi necessário realizar ajustes relacionados à quantidade de árvores a simular, os atributos de cada divisão (estando limitado a no máximo 5) e finalmente estabelecer a quantidade de subconjuntos que o modelo iria processar (sendo no caso menor do que 5).

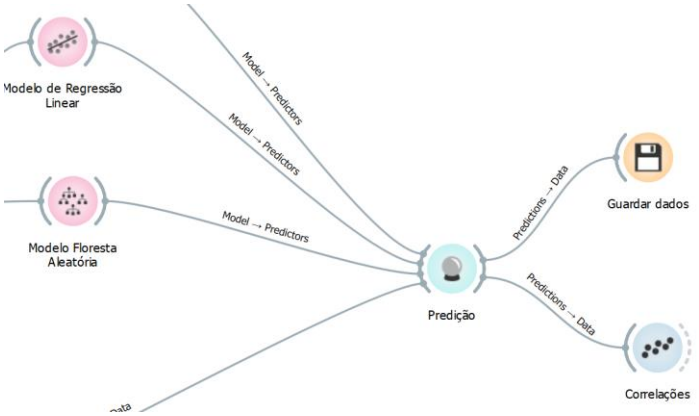
Figura 26. Configurações do modelo floresta aleatória



Fonte: o Autor, 2024.

Finalmente, os *widgets* referentes ao modelo após sua configuração foram conectados ao *widget* de "predição – *prediction"* (Figura 27) com a finalidade de emprega-lo para simular o comportamento do OD a partir do banco de dados da planilha teste e dos modelos propostos para realizar a simulação.

Figura 27. Layout dos modelos após "conexão" com o *widget* de predição

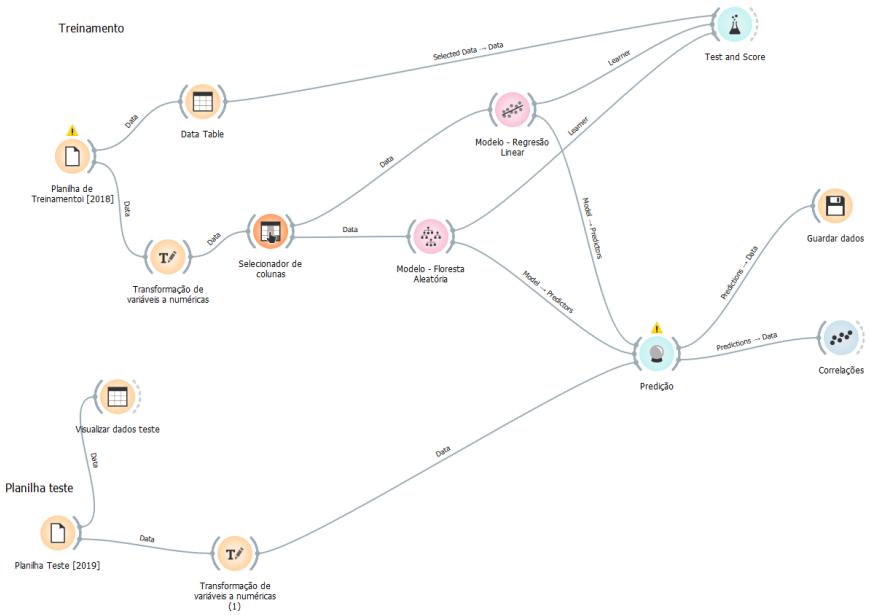


Fonte: o Autor, 2024.

5.7. MONTAGEM DO SISTEMA DE SIMULAÇÃO NO ORANGE DATA MINING

De forma geral, o modelo de aprendizado de máquina (Figura 28), foi realizado em base de três (3) fatores ou estruturas principalmente. O primeiro ramo foi conformado pela planilha de treinamento cuja função é treinar o algoritmo com todas as informações e/ou variáveis que podem interferir ao longo do ano na disponibilidade de oxigênio dissolvido. No caso, foi selecionado o ano 2018.

Figura 28. *Layout* do sistema de aprendizado de máquina implementado no Orange



Fonte: o Autor, 2024.

A planilha de treinamento foi composta pelo *widget* “File” cuja função relacionou-se à importação do arquivo .csv que continha o banco de dados de 2018 do Braço Taquacetuba (BL-105), ao mesmo tempo, com a finalidade de realizar uma “reinterpretação” do domínio ou natureza das variáveis, foi utilizado o *widget* “Edit Domain”, cuja função é a alteração do domínio das variáveis, a sua vez, utilizou-se o *widget* “Data Table” com a finalidade de visualizar o conjunto de dados após a troca de domínio. Na sequência foram utilizados os *widgets* dos modelos de aprendizado de máquina aplicados e finalmente utilizou-se o *widget* de “Simulation” ou simulação, o qual lançou os resultados de OD após a simulação (Figura 29).

Figura 29. Dados simulados no Orange Data Mining

	Modelo de Regressão Linear	Modelo Floresta Aleatória	Hora	pH	Surface PAR Total	Temperatura a 2 m (°C)	Precipitação (mm/hora)	Velocidade do vento (m/s)	Pressão superficia (kPa)
3070	7.92	8.99	18:00:00	8.08	0.00	17.34	0.02	1.99	95.60
3071	7.68	8.63	19:00:00.029000	8.00	0.00	17.09	0.02	2.04	95.65
3072	7.70	8.63	19:59:59.971000	8.02	0.00	16.83	0.01	2.20	95.67
3073	7.57	8.63	21:00:00	7.99	0.00	16.48	0.01	2.44	95.66
3074	7.44	8.64	22:00:00.029000	7.96	0.00	16.15	0.00	2.66	95.64
3075	7.27	8.51	22:59:59.971000	7.91	0.00	15.85	0.00	2.72	95.60
3076	7.25	8.51	00:00:00	7.91	0.00	15.49	0.00	2.53	95.55
3077	7.34	8.63	01:00:00.029000	7.96	0.00	15.06	0.01	2.34	95.51
3078	7.20	8.51	01:59:59.971000	7.93	0.00	14.59	0.01	2.49	95.51
3079	7.02	8.52	03:00:00	7.89	2.50	14.16	0.01	2.85	95.55
3080	7.03	8.47	04:00:00.029000	7.87	58.80	16.34	0.02	3.57	95.61
3081	6.99	8.28	04:59:59.971000	7.82	115.48	18.50	0.00	3.69	95.63
3082	7.06	8.28	06:00:00	7.82	211.98	21.01	0.00	3.35	95.62
3083	6.96	8.28	07:00:00.029000	7.79	304.50	22.83	0.00	3.44	95.64
3084	6.86	7.83	07:59:59.971000	7.74	335.92	23.68	0.00	3.20	95.69
3085	7.13	8.30	09:00:00	7.81	342.83	24.32	0.00	2.73	95.68
3086	7.48	8.50	10:00:00.029000	7.89	313.12	24.71	0.00	2.22	95.63
3087	8.02	8.67	10:59:59.971000	8.03	253.73	24.80	0.00	1.86	95.57
3088	8.62	9.15	12:00:00	8.19	171.12	24.55	0.00	1.59	95.49
3089	8.75	9.16	13:00:00.029000	8.19	74.73	23.53	0.00	1.32	95.46
3090	8.58	9.01	13:59:59.971000	8.13	4.92	21.90	0.00	1.18	95.46
3091	8.68	9.14	15:00:00	8.20	0.00	21.08	0.00	1.32	95.49
3092	8.47	8.99	16:00:00.029000	8.14	0.00	20.50	0.00	1.28	95.53
3093	8.33	8.99	16:59:59.971000	8.11	0.00	19.97	0.00	1.28	95.58
3094	9.43	9.19	18:00:00	8.56	0.00	19.30	0.00	1.52	95.65
3095	9.16	8.83	19:00:00.029000	8.49	0.00	18.61	0.00	1.78	95.73
3096	8.86	8.61	19:59:59.971000	8.40	0.00	18.21	0.00	2.05	95.80
3097	8.65	9.05	21:00:00	8.34	0.00	17.81	0.00	2.19	95.83
3098	8.42	9.05	22:00:00.029000	8.27	0.00	17.53	0.01	2.35	95.85
3099	8.27	9.14	22:59:59.971000	8.22	0.00	17.42	0.01	2.50	95.85
3100	8.24	9.14	00:00:00	8.21	0.00	17.42	0.03	2.49	95.84
3101	8.05	8.99	01:00:00.029000	8.14	0.00	17.38	0.05	2.49	95.83
3102	7.73	8.63	01:59:59.971000	8.03	0.00	17.26	0.07	2.59	95.83
3103	7.74	8.64							

Fonte: o Autor, 2024.

Já no caso da planilha de teste (ramo 2), a principal função foi a verificação da precisão dos modelos ao avaliar a disponibilidade de OD no sistema lótico. Assim, ela foi caracterizada por possuir dados de outros anos (ano anterior e subsequente) com a finalidade de realizar uma comparação do comportamento do sistema ao avaliar os dados reais quanto os dados simulados.

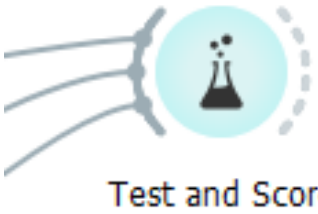
Similar à planilha de treinamento, foram utilizados os *widgets* File para

importação de arquivos, “Data Table” para a visualização de dados, “Edit Domain” para a reinterpretação das variáveis e finalmente “Simulation” para realizar a simulação. Alguns dos resultados gerados pelo Orange, foram utilizados para avaliar a precisão na resposta do modelo de aprendizado de máquina, para isso, foi utilizada a ferramenta analítica de erro relativo percentual, cujo principal objetivo é avaliar os erros existentes entre os valores reais e experimentais num determinado conjunto de dados.

Após a simulação (ramo 3), utilizaram-se os *widgets* de “Correlations” para avaliar a correlações entre as variáveis a partir do coeficiente de correlação de Pearson e o *widget* “Save Data” cuja função é a exportação dos dados simulados no formato .csv para simplificar a gestão dos dados obtidos.

Finalmente, utilizou-se o *widget* “Test and Score” (Figura 30). De forma sintética, essa ferramenta permite realizar testes estatísticos de algoritmos de aprendizagem de dados. Nele se encontram disponíveis várias estruturas de amostragem e análise de resultados (medidas de desempenho).

Figura 30. Widget "Teste and Score" - Orange Data Mining



Fonte: o Autor, 2024.

Além disso, o *widget* permite a realização de análises estatísticas amplamente utilizadas na ciência de dados, estas são: MSE, RMSE, MAE e R^2. O *software* Orange Data Mining define na sua biblioteca de ferramentas a utilização de cada um desses parâmetros assim como sua relação na avaliação do desempenho do modelo. No Quadro 8 ilustram-se as definições de cada uma das variáveis estatísticas utilizadas na avaliação do desempenho do modelo ao utilizar o *widget* Teste and Score.

Quadro 8. Definição dos parâmetros estatísticos do widget “Teste and Score”

Nomenclatura	Definição	Descrição
MSE (*Mean Square Erro)*	Erro Quadrático Médio	Permite mensurar a média dos quadrados dos erros ou desvios (diferença entre estimado e esperado)
RMSE (*Root Mean Square Erro*)	Raiz do Erro Quadrático Médio	Trata-se da raiz quadrada da média aritmética dos quadrados dos erros de um conjunto de números (uma medida imperfeita do ajuste dos dados esperados)
MAE (*Mean Absolute Erro*)	Erro Absoluto Médio	É utilizado para medir a proximidade das predições dos resultados finais quando comparado com os dados simulados.
R^2 (*Coefficient of Determination*)	Coeficiente de Determinação	É interpretado como a proporção da variância na variável dependente, a qual é explicada a partir da variável independente.

Fonte: o Autor, 2024.

O *software* Orange, possui uma biblioteca virtual a qual contém todas as formulas matemáticas e modelos estatísticos requeridos para obter esses dados. Dessa forma, foi possível a obtenção da avaliação desses resultados utilizando simplesmente o *widget* em questão.

Devido a que a análise de sistemas naturais possuem uma dinâmica complexa caracterizada por um comportamento não linear, as análises estatísticas no *widget* não foram alteradas, mantendo as configurações “padrão” do *software* na avaliação do desempenho do sistema a partir de modelos estatísticos.

O erro médio absoluto ou *mean absolute erro* é amplamente utilizado na ciência de dados devido às informações que traz na avaliação do desempenho dos modelos de regressão (FARIAS; CARVALHO, 2023), a principal característica que torna versátil esta variável é a capacidade de estabelecer a diferença absoluta entre os valores previstos pelo modelo e aqueles que foram simulados. Sendo assim, uma das avaliações do desempenho do modelo teve em consideração o MAE, cujo príncipio de avaliação foi relacionado ao seu

valor, no qual, quanto menor for, melhor será o desempenho do modelo.

A diferença do MAE, O MSE (*Mean Square Erro*) permite avaliar de uma forma mais rigrosa e/ou notável os erros e/ou ruidos mais destacados do sistema (WANG; SONG; ZHOU, 2015), assim, valores abruptos de MSE podem destacar erros na simulação do sistema e de fato no modelo de aprendizado de máquina que precisam ser avaliados antes da interpretação dos resultados finais.

Analogamente, o RMSE (*Root Mean Square Erro*) é conhecido na ciência de dados por ser um, dos principais indicadores de rendimento nos modelos de regressão, permitindo estabelecer estatisticamente a diferença entre a média dos valores simulados por um modelo e os valores reais (CAMILOTTI et al., 2023), assim, semelhante à avaliação do MAE, quanto menor o valor numérico referente ao RMSE, melhor será o modelo, tendo como base modelos perfeitos iguais a zero.

Finalmente, o coeficiente de determinação (R^2), é amplamente utilizado na ciência de dados e consequentemente nas diversas aplicações de aprendizado de máquina devido a sua capacidade de mensurar a qualidade de predição de um modelo estatístico (CHICCO; WARRENS; JURMAN, 2021). A faixa de valores do coeficiente de determinação abrange o conjunto de números de 0 até 1, e de forma geral, sua avaliação é feita em base ao valor numérico, onde quanto maior o valor R^2, melhor será o modelo ao realizar as predições no sistema avaliado, outra forma de avaliar o resultado relacionado ao R^2 é como a proporção das variâncias que são comaprtilhas entre as variáveis independentes e dependentes (CHICCO; WARRENS; JURMAN, 2021).

Assim, os parâmetros destacados na Tabela 4 foram utilizados na avaliação do desempenho dos modelos tanto do ano 2017 quanto do ano 2019 levando em consideração as especificidades de cada ano além dos modelos de aprendizagem de máquina avaliados.

6 RESULTADOS E DISCUSSÕES

6.1. AVALIAÇÃO E COMPARAÇÃO DOS MODELOS DE APRENDIZADO DE MÁQUINA

Primeiramente, para realizar uma verificação mais abrangente e adequada dos resultados obtidos, foi necessário avaliar o desempenho dos dois modelos de aprendizado de máquina implementados na simulação no *software* Orange Data Mining. Estes foram o modelo de regressão linear e o modelo floresta aleatória.

Para realizar essa análise, foram tidos em conta dois principais fatos: o primeiro, relacionado à natureza do sistema e o segundo, tendo em base uma análise estatística comparativa dos sistemas (análise estatístico pelo *widget* Test and Score disponível no Orange).

A dinâmica dos sistemas naturais é caracterizada pela sua complexidade no que diz respeito à relação que existe entre cada um dos parâmetros do meio, assim, modelos hidrográficos são caracterizados por possuírem um comportamento não linear (KORANGA et al., 2022), sendo assim, estimou-se a partir da revisão bibliográfica, que o modelo de regressão linear poderia ter dificuldades quando comparado com o modelo floresta aleatória ao simular a qualidade da água superficial do braço Taquacetuba da Represa Billings.

Contudo, a revisão bibliográfica não foi um argumento eliminatório do modelo de regressão linear, pois, a pesar que os sistemas hídricos apresentam comportamento não linear, é possível que a relação entre algumas variáveis fosse linear. Assim, no processo de verificação do desempenho dos modelos, foi utilizado o *widget* Teste and Score, o qual teve como principal objetivo realizar análises estatísticas do desempenho do modelo em função do banco de dados.

As variáveis estatísticas utilizadas para a avaliação do desempenho e consequentemente verificação do modelo de aprendizado de máquina mais adequado para o braço Taquacetuba da represa Billings foram: MSE (*Mean Square Erro*), RMSE (*Root Mean Square Erro*), MAE (*Mean Absolute Erro*) e o coeficiente de determinação R^2.

Após acoplar o *widget* Teste and Score no Data Table para o cálculo dos valores estatísticos (banco de dados de 2018), foram gerados os valores tanto para o modelo de regressão linear quanto para o modelo floresta aleatória. Na Tabela 1 ilustram-se os

valores obtidos para os modelos.

Tabela 1. Resultados estatísticos na avaliação dos modelos para o banco de dados (2018)

Modelo	MSE	RMSE	MAE	R^2
Regressão Linear	> 1,0	> 1,0	> 1,0	< 0,0
Floresta Aleatória	0,102	0,319	0,210	0,981

Fonte: o Autor, 2024.

De forma geral, essas variáveis estatísticas são amplamente utilizadas na ciência de dados devido à simplicidade que apresentam na avaliação do desempenho dos modelos de aprendizado de máquina (CHICCO; WARRENS; JURMAN, 2021).

O MAE foi o primeiro valor de avaliação de desempenho do modelo, uma vez que quanto menor for, melhor será o desempenho do modelo (AHMED et al., 2019), nesse sentido, quanto mais próximo estiver o MAE de zero, mais adequado o modelo para a avaliação do sistema, nesse caso, o modelo floresta aleatória apresentou um melhor desempenho no modelo de aprendizado de máquina quando comparado com o modelo de regressão linear, o qual teve valores maiores a 1,0.

O erro absoluto médio (MAE) permite estabelecer relação entre o valor real e o valor simulado, e em base à média do erro absoluto é possível inferir a margem de ruídos e erros para cada modelo (CHICCO; WARRENS; JURMAN, 2021), analogamente, a avaliação do MAE deve ser realizada em paralelo à avaliação do RMSE.

O RMSE permite determinar a proximidade do valor previsto e o valor real no modelo, assim, se o RMSE foi muito superior que o MAE, é provável que o modelo apresente predições erradas ou caracterizadas por ruídos (CAMILOTTI et al., 2023). Para o caso do modelo de regressão linear, o RMSE foi muito maior ao MAE, portanto, o modelo será caracterizado pela presença de predições erradas (valores abruptos), já no caso do modelo floresta aleatório o valor de RMSE é levemente maior ao MAE, o que pode caracterizar a possível presença de fenômenos de *overfitting* e *underfitting*, porém, torna-lo mais adequado do que o modelo de regressão linear na simulação do braço Taquacetuba da represa Billings.

O *overfitting* refere-se a um fenômeno que caracteriza uma “boa predição” para os dados de treino, dificultando a generalização em sistemas novos, ou seja, que não

fazem parte dos dados de treinamento, por outro lado, o *underfitting* consiste na incapacidade que possuí o modelo na aprendizagem de dados, gerando um erro elevado tanto nos dados de treino quanto nos dados de teste (JABBAR H, 2015).

O MSE permitiu visualizar que não há a presença de erros abruptos no modelo floresta aleatória. Por outro lado, o modelo de regressão linear apresentou valores muito altos, portanto, não apresentou um bom desempenho na simulação do sistema hidrográfico. No caso do modelo floresta aleatória, o resultado indica que quanto maior for o RMSE do MAE, o tamanho da amostra teste aumenta (SAGAN et al., 2020), ou seja, o modelo possui amostra de teste de tamanhos diferentes, fato que é frequente na modelagem e simulação de casos reais.

Finalmente, o coeficiente de determinação (R^2) permitiu obter informações relacionadas à relação que existe entre as variáveis, de forma geral, esse parâmetro é avaliado a partir de sua proximidade ao 0 ou a o 1, assim, variáveis próximas a 0 indicam baixa correlação, já variáveis próximas a 1 indicam uma correlação mais forte (THIEMANN; SCHEIBLER; WIEGAND, 2000).

Nos resultados obtidos no *widget* Teste and Score do *software* Orange, foi identificado que para o modelo de regressão linear, o coeficiente de determinação foi menor do que 0 (negativo), segundo a biblioteca de instruções do *widget* de pontuações do próprio Orange, o coeficiente de determinação pode ser negativo como consequência de respostas de saída ruins ou como resultado de predições que não conseguem ter em conta as características de entrada do modelo, esse fenômeno também é conhecido como “predição imperfeita”.

Por outro lado, o modelo floresta aleatória obteve resultados próximos a 1, o que indicou que às variáveis possuem uma tendência de crescer ao mesmo tempo, fato que indica um ajuste mais adequado quando comparado com o modelo de regressão linear na avaliação do desempenho de modelos.

O comportamento do coeficiente de determinação (R^2) é independente da linearidade do modelo de regressão, assim, esse parâmetro poderia ser próximo a 0 mesmo se for caracterizado como um modelo linear e próximos a 1 quando o modelo for caracterizado por sua não linearidade (CHICCO; WARRENS; JURMAN, 2021).

Diante isso, foi possível escolher o modelo floresta aleatória como o modelo de aprendizado de máquina mais adequado na simulação de OD da água superficial do braço Taquacetuba da represa Billings, já para o caso do modelo de regressão linear, foi

possível inferir que possui tendência a gerar predições ruins, caracterizadas por muitos ruídos e/ou erros que podem levar a deduções erradas no caso que forem avaliadas.

A pesar do modelo de regressão linear não ter uma boa adequação na avaliação do sistema estudado, sua simulação no *software* Orange Data Mining permitiu estabelecer a menor quantidade de parâmetros necessários para simular corretamente a qualidade da água superficial do braço Taquacetuba da represa Billings, para isso, foi empregado o *widget* Correlations, o qual estimou os valores de coeficiente de correlação de Pearson para os dois sistemas de regressão (linear e floresta aleatória).

Em base ao valor numérico do coeficiente de correlação de Pearson, foi possível determinar quais parâmetros de qualidade da água do sistema Billings possuem comportamento linear em função dos dois modelos, assim, aplicou-se a metodologia de tentativa e erro com a finalidade de escolher a menor quantidade de parâmetros possíveis que permitem simular o complexo hidrográfico de forma satisfatória, ou de fato, contendo a menor quantidade de ruídos.

Além de permitir estabelecer relações mais abrangentes entre os parâmetros que comporiam o modelo de aprendizado de máquina do sistema Billings, o intuito de reduzir a quantidade de parâmetros requeridos para simular a disponibilidade de oxigênio dissolvido, foi baratear o processo de monitoramento da qualidade da água.

Na Tabela 2, ilustram-se os valores de coeficiente de correlação de Pearson obtidos após utilização do *widget* "Correlations" no *software* Orange Data Mining para o modelo regressão linear do ano 2017, já na Tabela 3, encontram-se estabelecidos os valores relacionados ao modelo floresta aleatória para o mesmo ano.

Tabela 2. Coeficientes de correlação de Pearson – Modelo Regressão Linear (2017)

Parâmetro	**Coeficiente de Correlação de Pearson**	**Classificação**
Clorofila	+0,435	Correlação fraca
Condutividade	+0,056	Correlação nula
ORP	-0,204	Correlação nula
pH	+0,972	Correlação muito forte
Precipitação	-0,165	Correlação nula
Pressão Superficial	+0,073	Correlação nula
Radiação Fotossintética	+0,183	Correlação nula
Temperatura	+0,369	Correlação fraca
Temperatura a 2m	+0,377	Correlação fraca

Turbidez	+0,460	Correlação moderada
Velocidade do Vento	-0,303	Correlação nula

Fonte: o Autor, 2024.

Tabela 3. Coeficientes de correlação de Pearson – Modelo Floresta Aleatória (2017)

Parâmetro	**Coeficiente de Correlação de Pearson**	**Classificação**
Clorofila	+0,445	Correlação fraca
Condutividade	+0,078	Correlação nula
ORP	-0,282	Correlação nula
pH	+0,964	Correlação muito forte
Precipitação	-0,143	Correlação nula
Pressão Superficial	+0,123	Correlação nula
Radiação Fotossintética	+0,184	Correlação nula
Temperatura	+0,358	Correlação fraca
Temperatura a 2m	+0,307	Correlação fraca
Turbidez	+0,525	Correlação moderada
Velocidade do Vento	-0,262	Correlação nula

Fonte: o Autor, 2024.

Na análise dos resultados relacionados ao coeficiente de correlação de Pearson para o ano 2017, constatou-se semelhanças nos resultados estimados para os dois modelos, assim, parâmetros cuja tendência em relação à variável *target* estivesse classificada como “nula” e “fraca” foram submetidas ao “skip” na simulação, seguidamente variáveis cuja classificação fosse “moderada”, “forte” ou “muito forte” foram submetidas ao teste de “tentativa e erro”, no qual era realizada a simulação e na sequência comparados os dados reais com os simulados após a diminuição “skip” de uma das variáveis.

6.2. RESULTADO DOS DADOS DE TREINAMENTO E SIMULAÇÃO DO OD PARA O ANO DE 2017

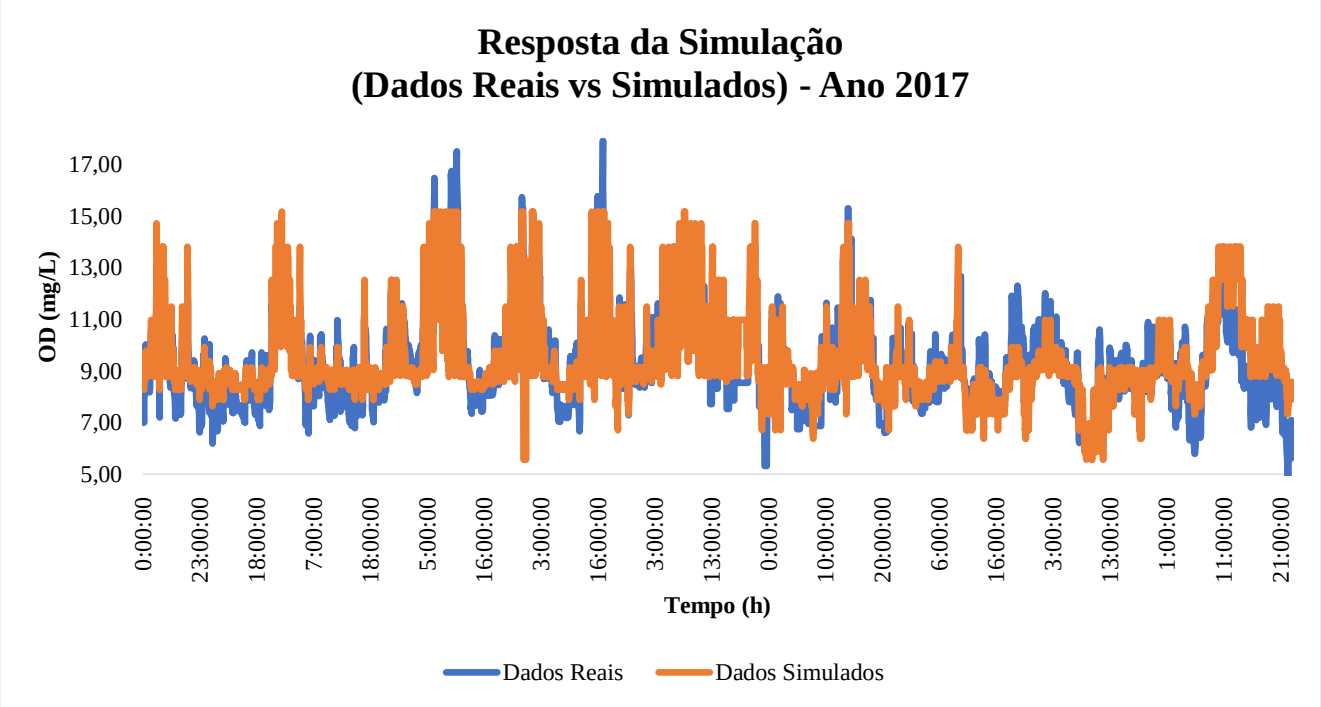
Para a simulação dos resultados do ano de 2017, partiu-se do fato de que a variável mais crítica fornecida pela SABESP na previsão de oxigênio dissolvido é o pH, a qual combinada com os parâmetros de monitoramento por satélite da NASA, permitiu a obtenção de parâmetros de saída muito precisos quando comparados com os dados reais, caracterizados pela ausência de ruídos abruptos e pela distinção do comportamento

(disponibilidade de OD) diário (dia e a noite) quanto nas estações do ano.

O pH ao ser uma propriedade intensiva determinante na determinação da qualidade da água de sistemas hidrográficos (OTHMAN et al., 2020), tornou-se a variável mais importante ao avaliar o modelo de aprendizado de máquina para o ano 2017, permitindo obter resultados altamente precisos como resposta de saída (simulação de OD).

Na Figura 31 encontra-se o gráfico com as respostas de saída após a simulação, nele, encontram-se disponíveis dos valores simulados e reais em função do tempo (horas) no decorrer do ano 2017. Visualmente, foi possível observar que os dados simulados (azul) sobrepõem os dados reais (laranja), o que permite uma visão mais abrangente dos resultados do modelo e da sua dinâmica anual, além de se verificar que o modelo de aprendizado de máquina conseguiu simular respostas dentro da faixa esperada, semelhante aos dados reais (em azul).

Figura 31. Resposta da Simulação Anual (Dados Reais vs simulados) - Ano 2017



Fonte: o Autor, 2024.

Embora as respostas de saída tenham demonstrado que o modelo consegue simular de forma satisfatória as faixas de disponibilidade de oxigênio dissolvido no decorrer de um ano tendo em contas as especificidades das estações, foi possível observar que há a

presença de alguns pontos (no inverno) nos quais ocorre o fenômeno de *underfitting*, indicando que o modelo não conseguiu reduzir o erro de treinamento (OTHMAN et al., 2020).

Já no mês de dezembro (primavera) foi detectado numa pequena faixa (18 até 21 de dezembro) a geração de *overfitting*, o qual indicou que, a pesar de o modelo estatístico ter se ajustado muito bem aos dados de treinamento, ele teve dificuldades na generalização de resultados nesse período, entretanto, conseguiu se ajustar adequadamente nos outros dias do mesmo mês, caracterizando essa faixa como um "ruído" na simulação.

6.2.1 Avaliação da dinâmica diária dos resultados da simulação de 2017, para cada estação do ano.

Na sequência, foi realizado o fracionamento dos períodos sazonais para avaliar de forma mais "minuciosa" a disponibilidade de OD (real vs simulada) nas diferentes estações, além de analisar a capacidade do modelo na criação de respostas precisas ao diferenciar a quantidade no dia e na noite.

Ao avaliar o comportamento diário e sazonal, foram utilizadas informações referentes à classificação das estações do ano a partir da sua duração mensal. No Quadro 2, encontram-se disponíveis as faixas estacionais e sua respectiva duração no Brasil.

Quadro 9. Estações no Brasil - Classificação por períodos

Estação	Período
Outono	21 de março – 21 de junho
Inverno	21 de junho – 23 de setembro
Primavera	23 de setembro – 21 de dezembro
Verão	21 de dezembro – 21 de março

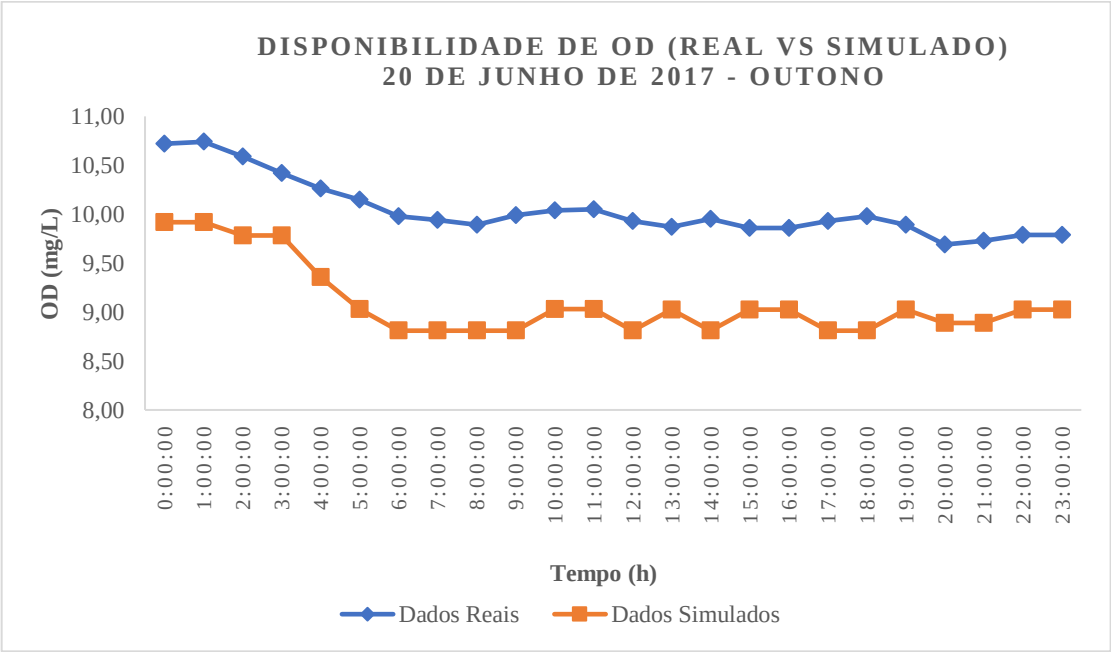
Adaptado de: (RIBEIRO FORTES; ARAÚJO DINIZ, 2023)

Para plotar os gráficos, foi utilizado o padrão com *layout* de linhas de dispersão, no eixo y foram plotados os valores relacionados ao oxigênio dissolvido (mg/L) e no eixo x a variação dele ao longo do tempo, assim, foram feitas duas classificações principalmente, a primeira em função da disponibilidade de oxigênio dissolvido diária (dia e noite) nas quatro (4) estações do ano e a segunda, em base a uma análise da disponibilidade

de oxigênio dissolvido ao longo de 15 dias nas quatro estações do ano.

Na Figura 32 ilustra-se o comportamento de oxigênio dissolvido (real e simulado) em função do tempo para a estação de outono. Neste caso, foi selecionada a data de 20 de junho de 2017, onde foi possível constatar que os dados simulados apresentaram uma precisão e tendência muito semelhante aos dados reais, estando caracterizados pelo reconhecimento nas mudanças de OD no transcurso do dia e da noite no outono.

Figura 32. Disponibilidade de OD (real vs simulado) – 20 de junho de 2017



Fonte: o Autor, 2024.

No outono, a disponibilidade de oxigênio dissolvido é caracterizada por começar a apresentar valores altos que quando comparada com estações mais quentes (primavera e verão). O modelo conseguiu interpretar as “oscilações” que ocorrem entre o dia e a noite tendo em conta as especificidades da estação. Na Tabela 4, apresenta-se uma comparação entre os valores reais e os valores simulados para a estação de outono de 2017 além do erro relativo percentual.

Tabela 4. Comparação dados reais vs simulados (Ano 2017) – Outono

Data	Hora	OD (Real)	OD (Simulado)	Erro (%)
20/06/2017	0:00:00	10,72	9,918295	9,92
20/06/2017	1:00:00	10,74	9,918295	9,92
20/06/2017	2:00:00	10,59	9,780412	9,78
20/06/2017	3:00:00	10,42	9,780412	9,78
20/06/2017	4:00:00	10,26	9,357817	9,36
20/06/2017	5:00:00	10,15	9,0334	9,03
20/06/2017	6:00:00	9,98	8,810107	8,81
20/06/2017	7:00:00	9,94	8,810107	8,81
20/06/2017	8:00:00	9,89	8,810107	8,81
20/06/2017	9:00:00	9,99	8,810107	8,81
20/06/2017	10:00:00	10,04	9,0334	9,03
20/06/2017	11:00:00	10,05	9,0334	9,03
20/06/2017	12:00:00	9,93	8,810107	8,81
20/06/2017	13:00:00	9,87	9,025129	9,03
20/06/2017	14:00:00	9,95	8,810107	8,81
20/06/2017	15:00:00	9,86	9,025129	9,03
20/06/2017	16:00:00	9,86	9,025129	9,03
20/06/2017	17:00:00	9,93	8,810107	8,81
20/06/2017	18:00:00	9,98	8,810107	8,81
20/06/2017	19:00:00	9,89	9,025129	9,03
20/06/2017	20:00:00	9,69	8,887152	8,89
20/06/2017	21:00:00	9,73	8,887152	8,89
20/06/2017	22:00:00	9,79	9,025129	9,03
20/06/2017	23:00:00	9,79	9,025129	9,03

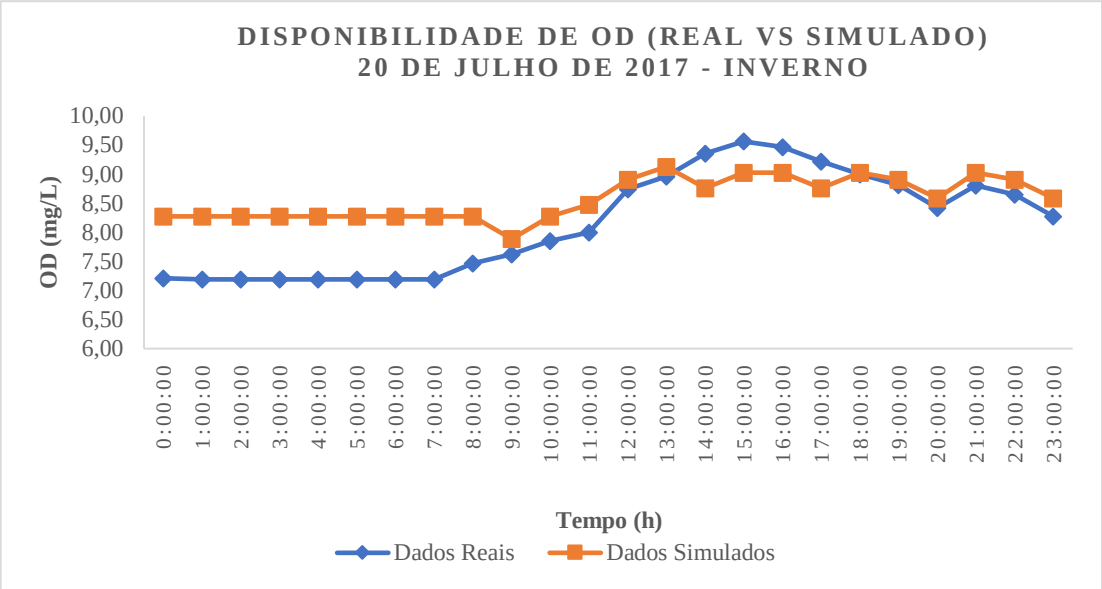
Fonte: o Autor, 2024.

Observou-se que o erro relativo percentual foi mantido numa faixa inferior ao 15%, portanto, a resposta é caracterizada por uma boa precisão (JONES et al., 2014; S.; A., 2021). As casas decimais dos valores simulados foram mantidas em 6 devido a que o *software* Orange gera até 15 decimais, e achou-se pertinente destacar a diferença (mesmo sendo apenas casas decimais) dos valores reais e simulados. Nos sistemas naturais, no outono a capacidade de retenção de oxigênio dissolvido na água pode ser afetada pela queda de matéria orgânica características da época (folhas, poeira, árvores, etc.), podendo ter oscilações na sua quantidade no começo da estação.

No inverno, a disponibilidade de oxigênio dissolvido é maior devido a fatores relacionados à temperatura do meio, a qual, ao ser tão baixa, promove uma menor atividade biológica e consequentemente uma maior quantidade de OD no lote hidrográfico.

Na Figura 33, é possível observar o gráfico comparativo (real vs simulado) da simulação do dia 20 de julho de 2017 (inverno).

Figura 33. Disponibilidade de OD (real vs simulado) – 20 de julho de 2017



Fonte: o Autor, 2024.

O algoritmo conseguiu predizer de forma satisfatória e precisa as tendências e variações no transcurso de um dia na disponibilidade de oxigênio dissolvido do dia em questão, apresentando assim, resultados dentro da faixa esperada. Foi detectado um pequeno fenômeno de *overfitting* das 14:00 hs até às 17:00 hs, onde o modelo obteve respostas com a mesma tendência dos dados reais, entretanto, com valores ligeiramente menores. Na Tabela 5, ilustram-se os dados comparativos assim como o erro percentual relativo para o presente caso.

Tabela 5. Comparação dados reais vs simulados (Ano 2017) – Inverno

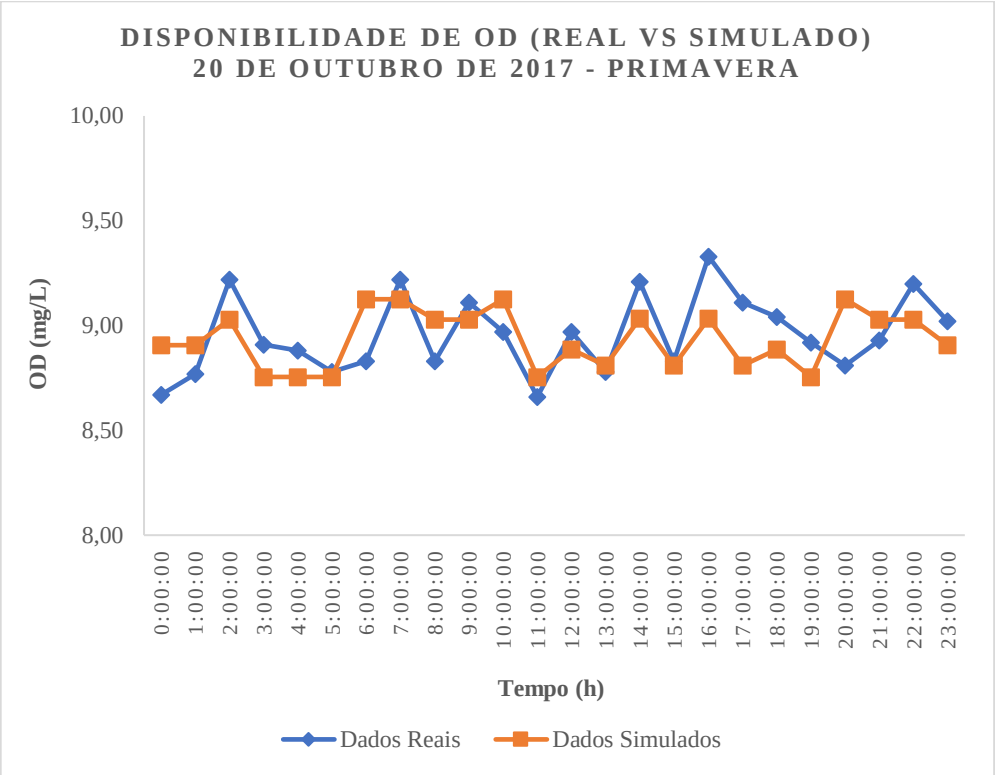
Data	Hora	OD (Real)	OD (Simulado)	Erro (%)
20/07/2017	0:00:00	7,21	8,272414	15,21
20/07/2017	1:00:00	7,19	8,272414	14,74
20/07/2017	2:00:00	7,19	8,272414	15,05
20/07/2017	3:00:00	7,19	8,272414	15,05
20/07/2017	4:00:00	7,19	8,272414	15,05
20/07/2017	5:00:00	7,19	8,272414	15,05
20/07/2017	6:00:00	7,19	8,272414	15,05
20/07/2017	7:00:00	7,19	8,272414	15,05
20/07/2017	8:00:00	7,47	8,272414	15,05
20/07/2017	9:00:00	7,62	7,8837	5,54
20/07/2017	10:00:00	7,85	8,272414	8,56

20/07/2017	11:00:00	8,00	8,476229	7,98
20/07/2017	12:00:00	8,74	8,906964	11,34
20/07/2017	13:00:00	8,96	9,125643	4,41
20/07/2017	14:00:00	9,35	8,754783	2,29
20/07/2017	15:00:00	9,56	9,025129	3,47
20/07/2017	16:00:00	9,46	9,025129	5,59
20/07/2017	17:00:00	9,22	8,754783	7,45
20/07/2017	18:00:00	9,00	9,027617	2,09
20/07/2017	19:00:00	8,81	8,906964	1,03
20/07/2017	20:00:00	8,42	8,585386	2,55
20/07/2017	21:00:00	8,80	9,027617	7,22
20/07/2017	22:00:00	8,65	8,906964	1,22
20/07/2017	23:00:00	8,27	8,585386	0,75

Fonte: o Autor, 2024.

Observou-se que o erro relativo lançou resultados ligeiramente maiores a 15%, portanto há características de possíveis erros de precisão na resposta de saída, entretanto, as características dos mesmos coincidem com a faixa de OD disponível no meio (dado real). Na sequência, foi realizada a avaliação da primavera no mesmo ano, tendo em conta que primavera assim como o verão, são caracterizadas pela diminuição gradativa de oxigênio dissolvido por fatores relacionados à temperatura (aumento da temperatura superficial da água), devido à adsorção do oxigênio do ar atmosférico ser inversamente dependente da temperatura, ou seja, quanto maior a temperatura da água menor a adsorção do oxigênio dissolvido na coluna d'água (SCARIOT, 2008). Na Figura 34 encontra-se disponível o gráfico comparativo (real vs simulado) da simulação executada para o dia 20 de outubro de 2017 (primavera).

Figura 34. Disponibilidade de OD (real vs simulado) – 20 de outubro de 2017



Fonte: o Autor, 2024.

Na avaliação desta estação, foi possível constar que há uma diminuição significativa na disponibilidade de oxigênio dissolvido no transcurso do dia quando comparada com outras estações. O modelo conseguiu predizer de forma precisa e satisfatória a pequena diferença que existe entre o dia e a noite de OD nessa estação, apenas um lapso de duas horas o modelo reconheceu valores ligeiramente menores à faixa real.

A primavera é reconhecida por ser uma das estações nas quais os processos fotossintéticos ocorrem majoritariamente (maior intensidade da radiação fotossintética), e consequentemente a atividade microbiológica, além do crescimento de plantas (algas) também (ALVES, 2010), sendo assim, um dos principais argumentos para a distinção da "variação" de OD neste braço da represa Billings.

Semelhante às outras estações, foi estimado o erro relativo percentual com a finalidade de avaliar a precisão na resposta de saída da simulação. Na Tabela 6, é possível visualizar os dados comparativos entre os dados reais e simulados, além do erro percentual relativo para a estação em questão.

Tabela 6. Comparação dados reais vs simulados (Ano 2017) – Primavera

Data	Hora	OD (Real)	OD (Simulado)	Erro (%)
20/10/2017	0:00:00	8,67	8,906964	2,73
20/10/2017	1:00:00	8,77	8,906964	1,56
20/10/2017	2:00:00	9,22	9,027617	2,09
20/10/2017	3:00:00	8,91	8,754783	1,74
20/10/2017	4:00:00	8,88	8,754783	1,41
20/10/2017	5:00:00	8,78	8,754783	0,29
20/10/2017	6:00:00	8,83	9,125643	3,35
20/10/2017	7:00:00	9,22	9,125643	1,02
20/10/2017	8:00:00	8,83	9,027617	2,24
20/10/2017	9:00:00	9,11	9,027617	0,90
20/10/2017	10:00:00	8,97	9,125643	1,74
20/10/2017	11:00:00	8,66	8,754783	1,09
20/10/2017	12:00:00	8,97	8,887152	0,92
20/10/2017	13:00:00	8,78	8,810107	0,34
20/10/2017	14:00:00	9,21	9,0334	1,92
20/10/2017	15:00:00	8,83	8,810107	0,23
20/10/2017	16:00:00	9,33	9,0334	3,18
20/10/2017	17:00:00	9,11	8,810107	3,29
20/10/2017	18:00:00	9,04	8,887152	1,69
20/10/2017	19:00:00	8,92	8,754783	1,85
20/10/2017	20:00:00	8,81	9,125643	3,58
20/10/2017	21:00:00	8,93	9,027617	1,09
20/10/2017	22:00:00	9,20	9,027617	1,87
20/10/2017	23:00:00	9,02	8,906964	1,25

Fonte: o Autor, 2024.

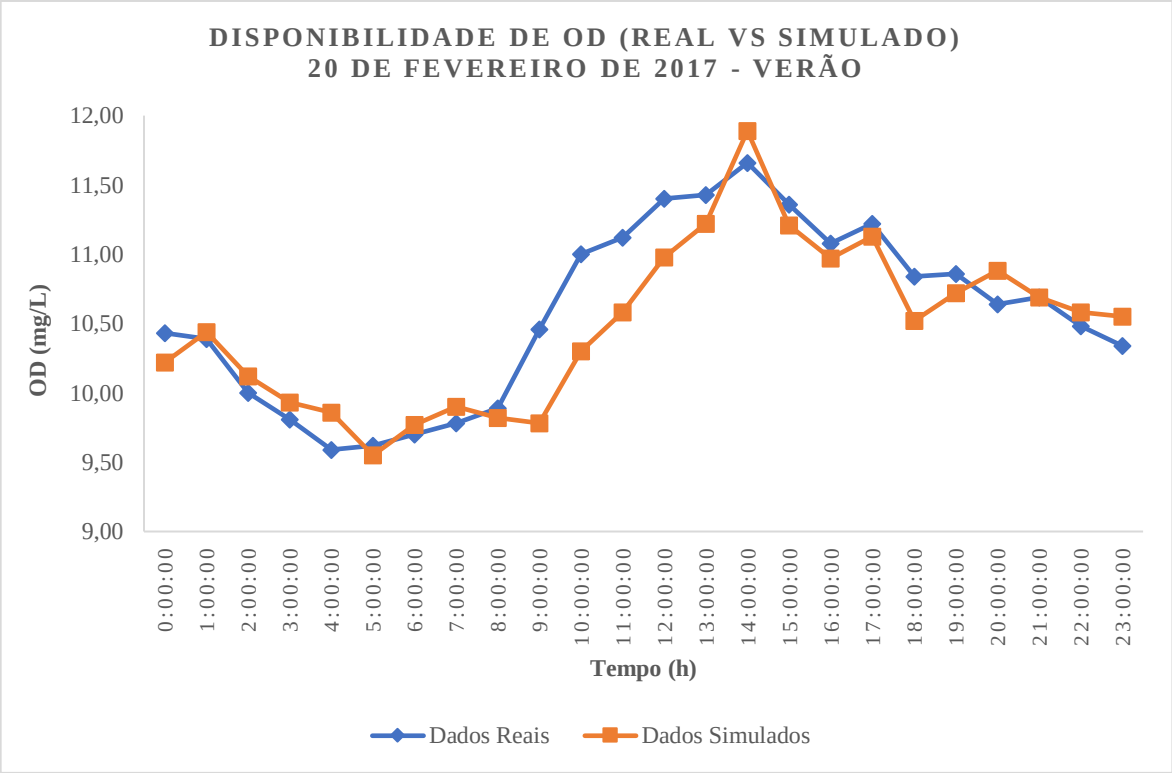
Verificou-se que o modelo apresentou alta precisão, pois os erros percentuais foram menores a 15%, além disso, a faixa de estimação de OD foi coerente aos dados reais, atingindo valores muito próximos além das mesmas tendências específicas da primavera. Finalmente, na avaliação dos resultados obtidos para a estação de verão do ano 2017 (20 de fevereiro de 2017), observaram-se mudanças significativas na disponibilidade de oxigênio dissolvido, as quais não são características de dita estação.

Durante o verão, as temperaturas da água geralmente aumentam devido ao clima mais quente, em consequência, a capacidade que possuem as bacias hidrográficas para ter oxigênio também diminui, acarretando então na diminuição notável de oxigênio dissolvido na coluna d'água (SCARIOT, 2008), por outro lado, a atividade biológica é destacada amplamente no ciclo de oxigênio dissolvido durante o verão, isto é devido a que durante o dia, a fotossíntese das plantas aquáticas pode aumentar a produção de oxigênio na água, porém, nos períodos noturnos a respiração dos microrganismos aquáticos pode

consumir oxigênio, reduzindo dessa forma os níveis disponíveis (ALVES, 2010), além do modelo ter conseguido predizer esta dinâmica de OD durante o verão, conforme observa-se na Figura 35.

O principal indício dessa mudança não característica da disponibilidade de OD no verão, poderia estar relacionada ao incremento dos níveis de eutrofização devido à agricultura intensiva característica da época na região de São Paulo e redondezas da RMSP (ALVAREZ et al., 2005). Na Figura 35, ilustra-se o comportamento (real vs simulado) da disponibilidade de oxigênio dissolvido em função do tempo num período de 24 horas no dia 20 de fevereiro de 2017 (verão).

Figura 35. Disponibilidade de OD (real vs simulado) – 20 de fevereiro de 2017



Fonte: o Autor, 2024.

A resposta do modelo de regressão gerou resultados precisos quando comparados aos reais, foi evidenciado uma pequena divergência entre os valores reais e os preditos no período da manhã, entretanto, muito próximo ao valor real. De forma geral, o algoritmo conseguiu predizer satisfatoriamente a tendência nos valores de OD no transcurso de um dia na estação, tendo em conta sua natureza "não comum", pois o período não é caracterizado pela alta disponibilidade de OD devido aos fatores de temperatura

característicos da estação. Na Tabela 7, é possível visualizar as estimativas de erro relativo percentual neste caso.

Tabela 7. Comparação dados reais vs simulados (Ano 2017) – Verão

Data	Hora	OD (Real)	OD (Simulado)	Erro (%)
20/2/2017	0:00:00	10,43	10,22	2,01
20/2/2017	1:00:00	10,39	10,44	0,48
20/2/2017	2:00:00	10,00	10,12	1,20
20/2/2017	3:00:00	9,81	9,93	1,22
20/2/2017	4:00:00	9,59	9,86	2,82
20/2/2017	5:00:00	9,62	9,55	0,73
20/2/2017	6:00:00	9,70	9,77	0,72
20/2/2017	7:00:00	9,78	9,9	1,23
20/2/2017	8:00:00	9,89	9,82	0,71
20/2/2017	9:00:00	10,46	9,78	6,50
20/2/2017	10:00:00	11,00	10,3	6,36
20/2/2017	11:00:00	11,12	10,58	4,86
20/2/2017	12:00:00	11,40	10,98	3,68
20/2/2017	13:00:00	11,43	11,22	1,84
20/2/2017	14:00:00	11,66	11,89	1,97
20/2/2017	15:00:00	11,36	11,21	1,32
20/2/2017	16:00:00	11,08	10,97	0,99
20/2/2017	17:00:00	11,22	11,13	0,80
20/2/2017	18:00:00	10,84	10,52	2,95
20/2/2017	19:00:00	10,86	10,72	1,29
20/2/2017	20:00:00	10,64	10,88	2,26
20/2/2017	21:00:00	10,69	10,69	0,00
20/2/2017	22:00:00	10,48	10,58	0,95
20/2/2017	23:00:00	10,34	10,55	2,03

Fonte: o Autor, 2024.

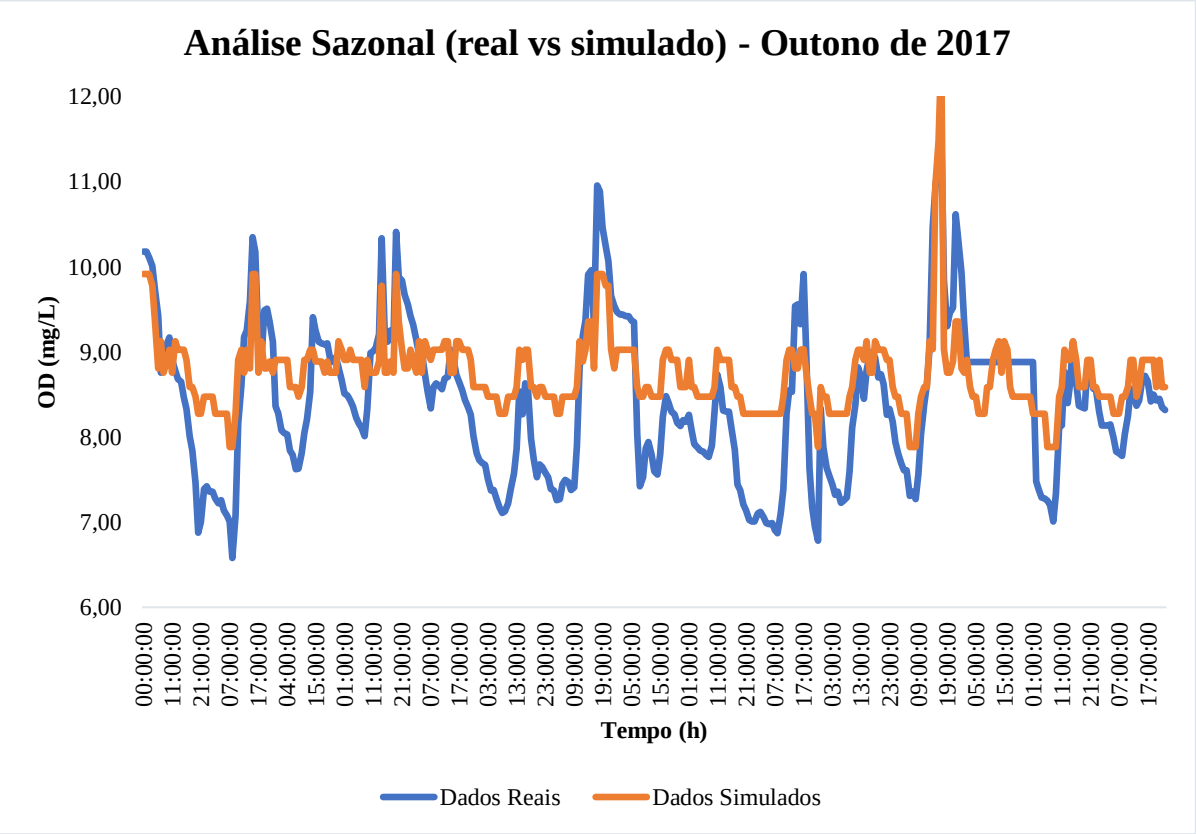
A tabela 7 mostra que o erro percentual foi sempre inferior a 15%, demonstrando a alta precisão que teve a simulação. Durante os picos de calor 09:00 às 15:00 hs, a tendência na resposta do modelo é gerar resultados ligeiramente menores do que os dados reais, porém, os mesmos encontraram-se dentro da faixa esperada de oxigênio dissolvido, tanto quando comparados com os dados reais, quanto com às especificidades do sistema.

6.2.2 Avaliação da dinâmica semanal dos resultados da simulação de 2017 para cada estação do ano

Com a finalidade de generalizar de forma mais abrangente as respostas

obtidas no modelo de aprendizado de máquina para o ano 2017, utilizando apenas os parâmetros de monitoramento por satélite da NASA e os dados de pH da SABESP, foram gerados gráficos representativos contendo 15 dias consecutivos para cada estação. Nas Figuras 36, 37. 38 e 39 ilustram-se os gráficos contendo os comportamentos sazonais (real vs simulado) num período de 15 dias (dia e noite) no ano 2017.

Figura 36. Análise Sazonal (real vs simulado) - Outono de 2017



Fonte: o Autor, 2024.

A análise para a estação de outono foi realizada nos primeiros 15 dias do mês de junho. Percebeu-se que para a estação de outono, cuja característica padrão e a retenção de oxigênio dissolvido no meio devido às quedas de temperatura e da atividade biológica, o modelo lançou resultados que se encontram dentro da faixa real, entretanto, reproduziu valores menores do que os reais, sobretudo os picos inferiores, em que os valores simulados aparentam estar normalizados, não conseguindo reproduzir a dinâmica real dos picos inferiores.

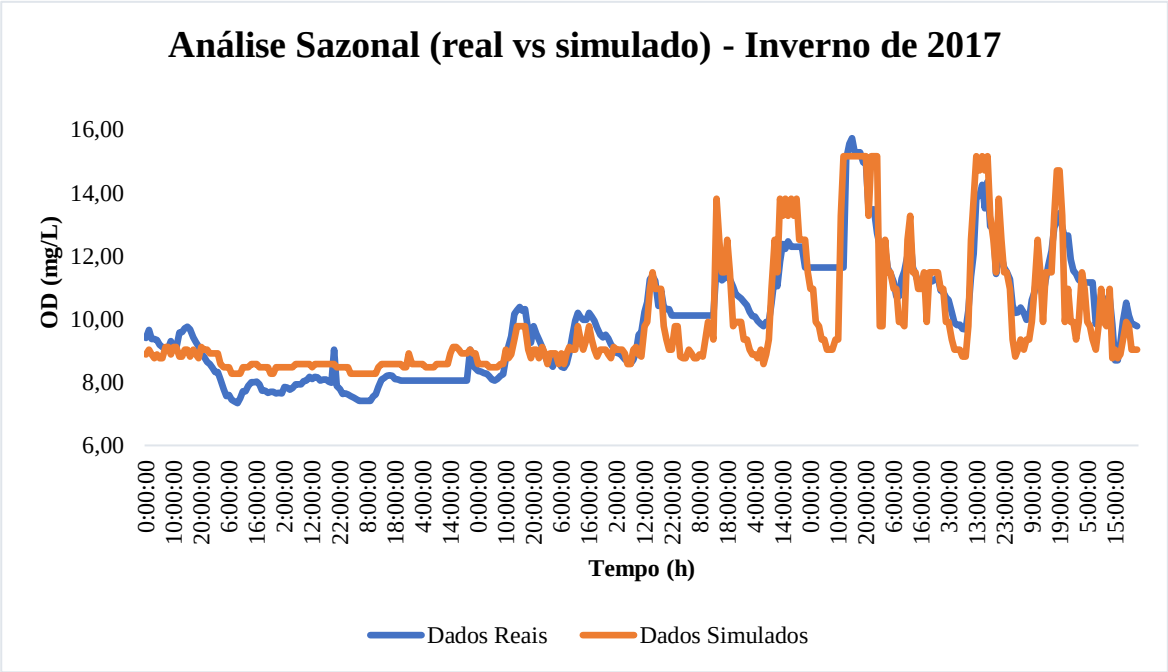
Embora, os resultados da simulação sejam maiores do que os valores reais, o modelo ainda conseguiu reproduzir o comportamento tendencial ao longo dos quinze

dias tendo em conta os “picos” de aumento e em menor escalas os picos de diminuição de OD no meio, até atingir valores médios próximos (precisos) aos valores reais.

No que diz respeito aos níveis de eutrofização da água do braço Taquacetuba no período do outono, é possível destacar que houve uma queda não característica de OD que poderia ser explicada pela queda de materiais orgânicos como folhas, árvores, madeira dos arredores da represa que levaram à diminuição na disponibilidade de OD, no presente caso, ao ter vários picos nos quais houve valores próximos a 5,0 mg/L (valor mínimo estabelecido pela resolução CONAMA 357/05) destacou-se uma tendência à eutrofização na maioria do tempo até sua estabilização (HERREIRA; PIZELLA, 2020).

O fenômeno não característico de diminuição na quantidade de OD no braço Taquacetuba nesse período só conseguiu ser estabilizado após a segunda semana do mês de julho (Figura 37), período em que ocorreu a transição do outono para o inverno. O modelo de regressão conseguiu reproduzir respostas de saída muito mais precisas nesse período, as quais se encontraram dentro da faixa referente aos dados reais.

Figura 37. Análise Sazonal (real vs simulado) – Inverno de 2017



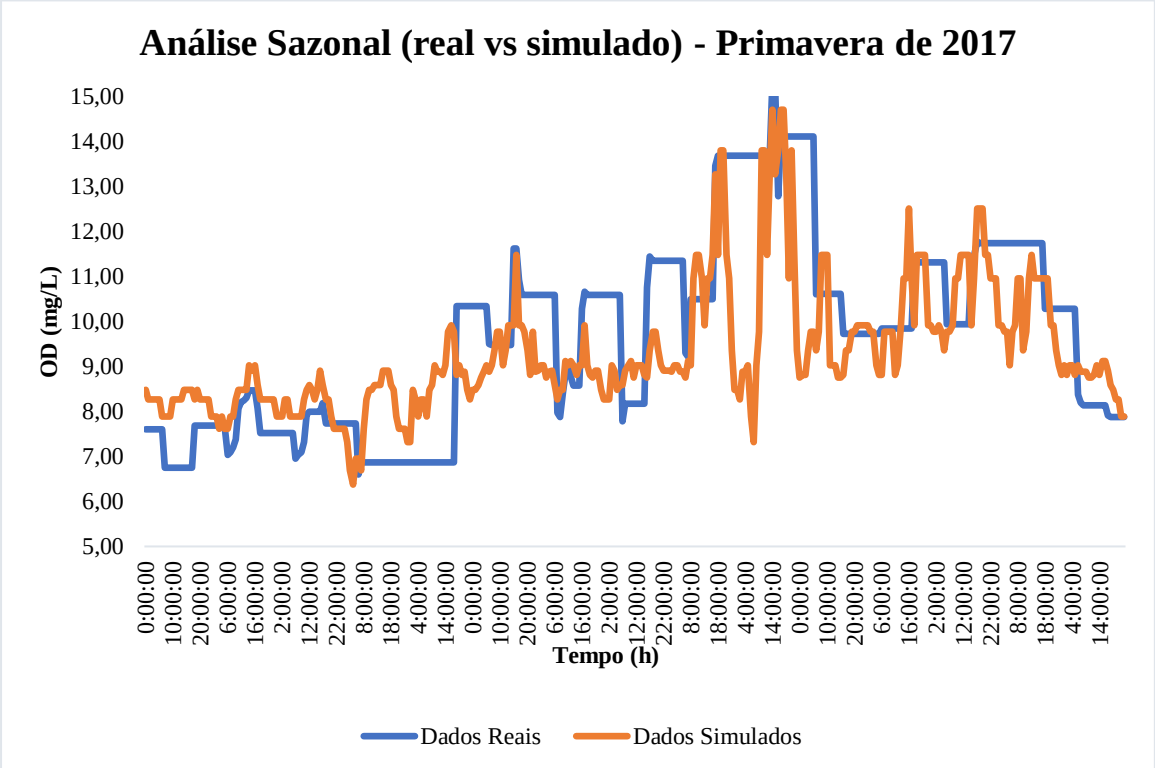
Fonte: o Autor, 2024.

Além disso, o comportamento tendencial correspondente à quantidade de OD no meio foi semelhante aos dados reais, conseguindo reconhecer os períodos nos quais

houve mudanças notáveis no mesmo, por outro lado, o modelo continuou gerando valores ligeiramente menores do que os dados reais. A pesar que desde o outono, na análise de sistemas naturais, é comum ter um aumento gradativo na disponibilidade de OD, neste ano, só no inverno foi possível perceber este aumento. Devido às quedas de temperatura características desta estação, a adsorção de OD é maior.

O modelo, de forma geral, conseguiu predizer as relações entre os parâmetros que levaram à queda de OD nesse braço da represa Billings, permitindo caracterizar os dados do inverno como irregulares, cuja tendência foi se ajustar em níveis abaixo dos valores esperados (tendência a atingir a eutrofização). Na primavera (Figura 38), foi constatado que o modelo gerou resultados "padrões" do que seria o comportamento natural na primavera, nessa análise sazonal, foi escolhido o mês de outubro.

Figura 38. Análise Sazonal (real vs simulado) – Primavera de 2017



Fonte: o Autor, 2024.

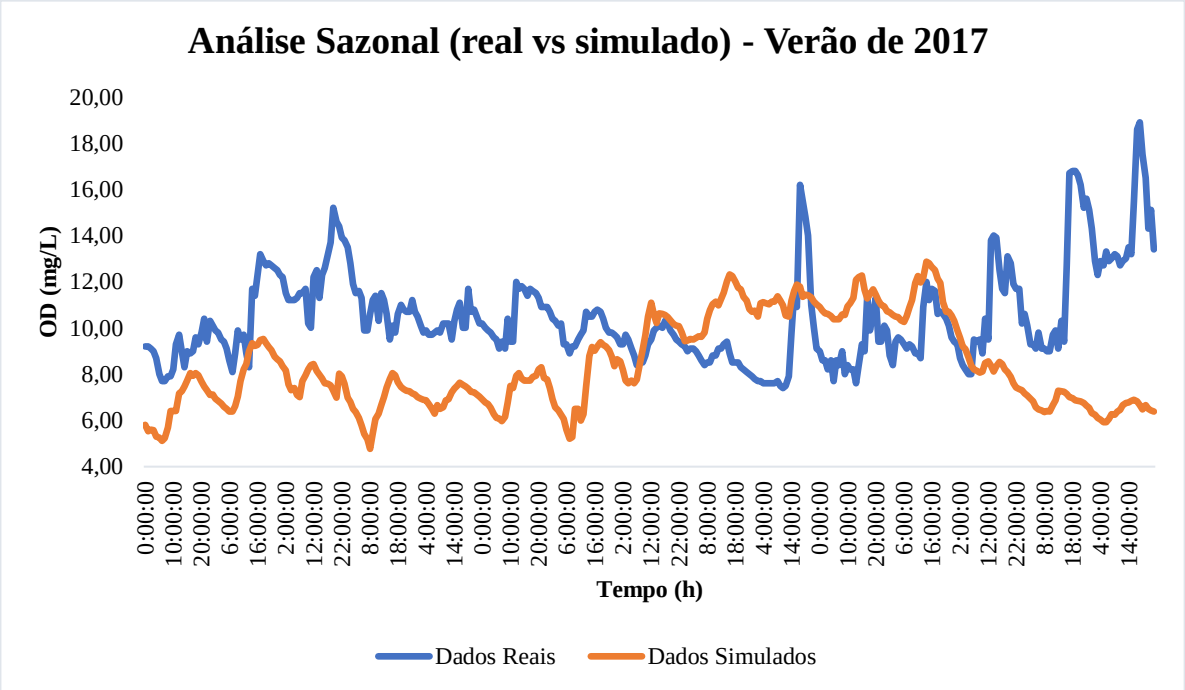
A "divergência" entre os dados reais e os dados simulados pode estar relacionada a algum comportamento atípico nas saídas e uma possível disfunção na sonda de monitoramento da represa, pois, foram detectados vários períodos nos quais a variável de OD não teve mudanças, criando consequentemente, momentos nos quais os dados reais

representaram o comportamento de uma linha reta.

Entretanto, embora o comportamento tendencial não tenha caracterizado alguns picos no período avaliado, a precisão entre os valores reais e simulados foi bem próxima, de tal forma que o modelo conseguiu gerar valores de saída de OD característicos da estação de forma satisfatória, pudendo destacar os períodos nos quais há uma presença maior de OD, revelando uma utilidade potencial do modelo de aprendizado de máquina para preencher lacunas de dados reais, em que as sondas apresentam problemas.

Além disso, o modelo conseguiu replicar a transição de uma estação à outra no que diz respeito à disponibilidade de OD, com o aumento da temperatura e ao ser caracterizada como a estação onde mais ocorrem processos fotossintéticos, a tendência é haver menos disponibilidade de OD, e de fato, foi possível detectar esse fenômeno. Finalmente, na avaliação sazonal do verão, foi escolhido o período correspondente aos 15 primeiros dias do mês de janeiro do ano 2017 (Figura 39).

Figura 39. Análise Sazonal (real vs simulado) – Verão de 2017



Fonte: o Autor, 2024.

Nesse período, destacou-se que o modelo de regressão conseguiu predizer respostas de saída menores aos dados reais, tendo características de fenômeno de *underfitting*. A pesar do modelo de aprendizado de máquina conseguir ajustes semelhantes aos dados reais, e em muitos períodos as faixas reais e simuladas “coincidem”, a tendência

generalizada foi a reprodução de valores menores aos dados reais.

Porém, a natureza sazonal do período avaliado não é comum, pois, devido às altas temperaturas do mês, a faixa de disponibilidade de OD não deveria ultrapassar os 12,0 mg/L (SENDACZ et al., 2005). Nesse sentido, o fato da divergência no comportamento tendencial da quantidade de OD no braço Taquacetuba para essa faixa poderia estar relacionado a dois principais motivos.

O primeiro, em função de diversas falhas que apresentou a sonda de monitoramento na coleta de dados de pH, o qual poderia ter levado à presença de pequenas divergências no modelo, ou à possibilidade de o modelo ter conseguido replicar dados "padrão" do comportamento normal no verão a partir da planilha de treinamento.

Em relação à avaliação dos níveis de eutrofização do meio, foi possível constatar que a resposta da simulação gerou valores próximos a 5,0 mg/L, o que não é o ideal, pois erroneamente levaria a inferir que houve uma queda drástica na disponibilidade de OD e consequentemente uma maior probabilidade de caracterização de água com altos teores de nitrogênio e fosforo, o que não é confirmado pela realidade.

De forma geral, o modelo apresentou resultados muito satisfatórios ao compará-los com os dados reais, permitindo abrir a possibilidade de monitoramento por aprendizado de máquina, tornando mais barato o processo e diminuindo custos operacionais.

6.3 RESULTADO DOS DADOS DE TREINAMENTO E SIMULAÇÃO DO OD PARA O ANO DE 2019

Analogamente ao processo de avaliação de modelos e de comportamento sazonal no ano 2017, foi realizada a verificação dos resultados para o ano 2019 (ano subsequente à planilha de treinamento). Em base ao fato de que a análise estatística lançou resultados que inferem que o modelo de regressão linear não é adequado para avaliar a quantidade de oxigênio dissolvido na água superficial da represa Billings (devido à natureza não linear do sistema natural), foram avaliados apenas os parâmetros referentes à correlação de Pearson com a finalidade de verificar quais variáveis teriam maior chance de possuir uma relação intrínseca com a variável *target*.

Os dados disponibilizados pela SABESP do ano 2019 possuem muitas falhas nas sondas de monitoramento, e em muitos casos os períodos e inclusive meses não houve captura de dados, motivo pelo qual o banco de dados da planilha de treinamento é

menor quando comparada com a do ano 2017, no entanto, o conjunto de dados consegue abranger as quatro estações do ano.

Na Tabela 8, encontram-se ilustrados os coeficientes de correlação de Pearson para cada uma das variáveis que conformaram a planilha de treinamento e de teste em relação à variável *target* (OD).

Tabela 8. Coeficientes de correlação de Pearson – Modelo Regressão Linear (2019)

Parâmetro	Coeficiente de Correlação de Pearson	Classificação
Condutividade	+0,577	Correlação moderada
pH	+0,016	Correlação nula
Precipitação	+0,006	Correlação nula
Pressão Superficial	-0,257	Correlação nula
Radiação Fotossintética	+0,020	Correlação nula
Temperatura	+0,751	Correlação forte
Temperatura a 2m	+0,650	Correlação forte
Turbidez	-0,372	Correlação nula
Velocidade do Vento	-0,342	Correlação nula

Fonte: o Autor, 2024.

O modelo de regressão linear destacou apenas como correlações moderadas e fortes os parâmetros: condutividade, temperatura e temperatura a 2m, as demais variáveis foram classificadas como variáveis cuja correlação é nula (não crescem em função do OD). Na Tabela 9 ilustram-se os coeficientes de correlação de Pearson obtidos no modelo floresta aleatória.

Tabela 9. Coeficientes de correlação de Pearson – Modelo Floresta Aleatória (2019)

Parâmetro	Coeficiente de Correlação de Pearson	Classificação
Condutividade	+0,300	Correlação fraca
pH	+0,413	Correlação fraca
Precipitação	+0,117	Correlação nula
Pressão Superficial	-0,312	Correlação nula
Radiação Fotossintética	+0,408	Correlação moderada
Temperatura	+0,513	Correlação moderada
Temperatura a 2m	+0,694	Correlação forte
Turbidez	-0,157	Correlação nula
Velocidade do Vento	-0,005	Correlação nula

Fonte: o Autor, 2024.

No caso dos coeficientes de correlação de Pearson obtidos no modelo

floresta aleatória, foram classificados como valores de correlação moderada e forte as variáveis: pH, radiação fotossintética, temperatura e temperatura a 2m. Assim, comparando os resultados obtidos para ambos os casos, foram submetidas ao teste de “tentativa e erro” as variáveis da SABESP de pH e temperatura.

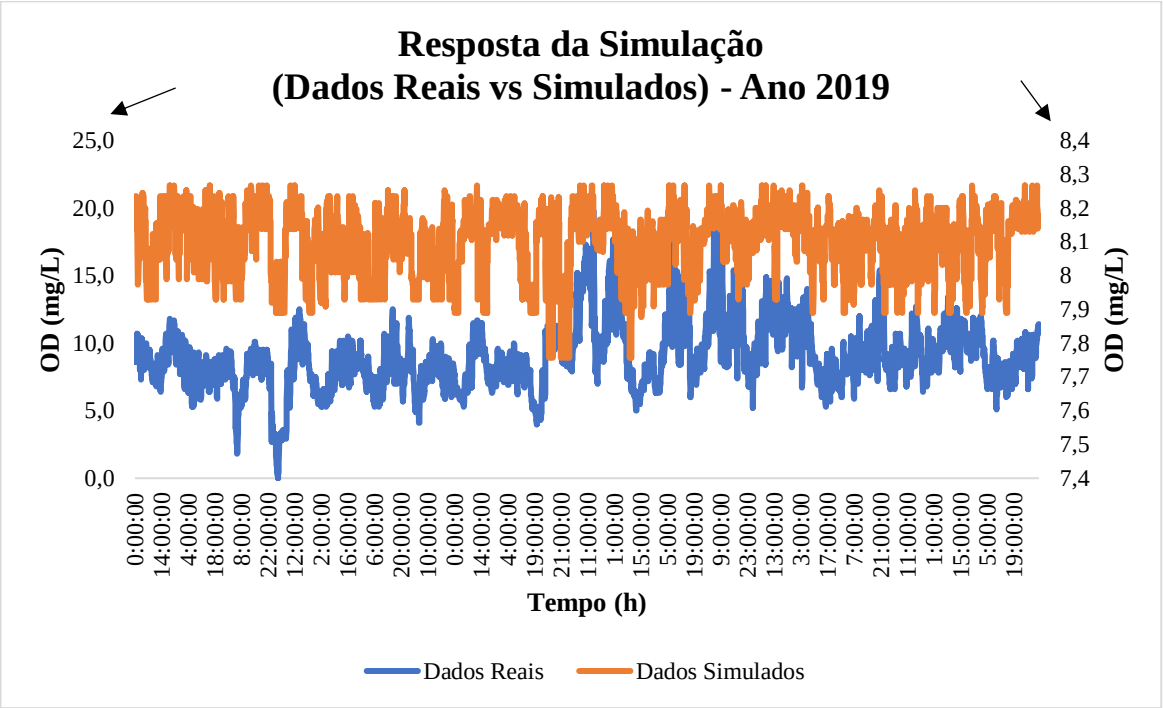
Após a simulação, obteve-se um resultado ligeiramente diferente do obtido para o ano de 2017, foi detectado que apenas o valor de pH e os dados de monitoramento por satélite da NASA não permitiram ter uma boa resposta de saída, no entanto, ao combinar a variável pH com temperatura (monitoramento da SABESP) foi possível ter um modelo de regressão adequado quando comparado com os dados reais.

Destaca-se que o banco de dados de 2019 não apresenta a variável clorofila a qual apresentou uma correlação fraca no banco de dados de treinamento de 2017. Devido à ausência da variável clorofila no banco de dados correspondente ao ano 2019, provavelmente o modelo destacou as variáveis de temperatura e pH como as mais críticas em relação à variável *target*.

Neste ano, tendo como base aos dados de monitoramento de qualidade da água da SABESP, foi possível constatar uma perda generalizada na qualidade da água do braço Taquacetuba, onde, em alguns períodos os valores de oxigênio dissolvido disponível na água quase foram nulos (≈ 0,0 mg/L).

A Figura 40, ilustra a resposta de saída global ao longo do ano de 2019, após executar a simulação com as variáveis: pH, temperatura, além das variáveis de monitoramento por satélite da NASA. Os dados reais devem ser lidos na ordenada à esquerda do gráfico, enquanto, que os dados simulados devem ser lidos na ordenada à direita.

Figura 40. Resposta da Simulação (Dados Reais vs simulados) - Ano 2019



Fonte: o Autor, 2024.

Detectou-se a presença do fenômeno de *overfitting* na resposta de saída do modelo, onde percebeu-se que o modelo de regressão produz resultados precisos em base aos dados de treinamento, porém não para os dados novos.

Uma das possíveis explicações para ter ocorrido o fenômeno de *overfitting* está relacionada à ausência da variável clorofila no banco de dados de treinamento do ano de 2019, pois a clorofila é uma variável que está relacionada de forma sistêmica com as demais variáveis que apresentaram alguma correlação, principalmente ao pH e, também indiretamente ao OD, pois a clorofila reflete a atividade fotossintética do corpo hídrico, responsável pela produção de oxigênio pelas algas, ou seja, a quantidade de clorofila no meio permite estabelecer relações diretas com a quantidade de OD, além de outras variáveis como é o caso da radiação fotossintética (parâmetro de monitoramento por satélite da NASA).

Outro fato que teria levado ao fenômeno de *overfitting* poderia estar relacionado às falhas intermitentes na sonda de monitoramento da SABESP também podem ter contribuído para o fenômeno.

O oxigênio dissolvido possui uma relação intrínseca com a clorofila, já que sua disponibilidade e/ou quantidade no meio depende do processo de fotossíntese relacionado por algumas algas, plantas e bactérias, dessa forma, quanto maior for a quantidade de clorofila no meio, maior será a disponibilidade de OD no mesmo (BARRETO

et al., 2013).

A pesar desta limitação, é possível afirmar que o modelo conseguiu reproduzir o comportamento tendencial, ao longo do ano, ou seja, ele previu as oscilações da disponibilidade de oxigênio dissolvido na coluna d'água de forma satisfatória, permitindo destacar a diferenciação sazonal e reconhecendo as mudanças nos períodos diurnos e noturnos.

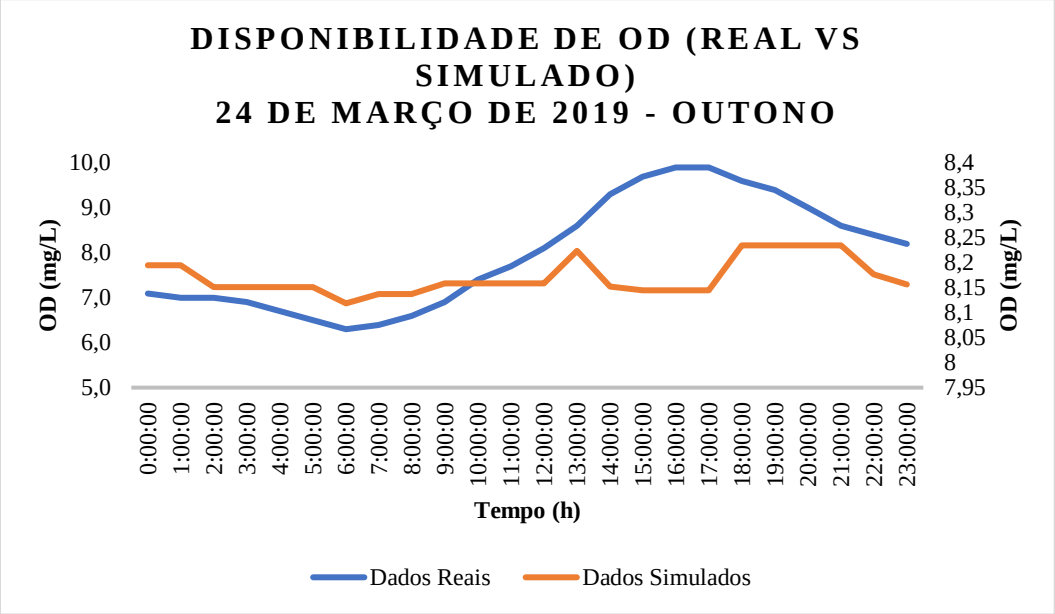
No caso do modelo de aprendizado de máquina aplicado para o ano 2019, foi possível destacar que foram reconhecidas como variáveis críticas o pH e a temperatura, o que condiz com a literatura, uma vez que os processos bioquímicos nos meios aquáticos são determinados por uma faixa determinada de pH e no caso da temperatura, dependendo da sazonalidade, haverá uma maior ou menor quantidade de OD (SCARIOT, 2008).

6.3.1 Avaliação da dinâmica diária dos resultados da simulação de 2019 para cada estação do ano

De forma semelhante à análise realizada para o ano 2017, foi realizada a avaliação do modelo de 2019 em questões de sazonalidade (reconhecimento das mudanças sazonais ademais da quantidade de OD do dia e a noite), além da sua precisão na predição de valores de OD quando comparado com os dados reais.

Diante isso, foram plotados os gráficos de forma amostral de um dia para cada uma das estações do ano, na Figura 41 ilustra-se o gráfico amostral para avaliação da resposta do modelo de regressão no outono do ano 2019 (24 de março de 2019).

Figura 41. Disponibilidade de OD (Real vs *si*mulado) – 24 de março de 2019



Fonte: o Autor, 2024.

O modelo conseguiu predizer de forma satisfatória o comportamento tendencial da disponibilidade de OD no braço Taquacetuba da represa Billings, entretanto, com "picos" caracterizados pela predição de valores menores quando comparados com os dados reais, neste caso, a resposta do modelo de regressão conseguiu se ajustar em alguns períodos à faixa de dados reais, apresentando apenas faixas específicas de "ruídos" nos resultados de saída.

Outro fato importante a destacar é que o modelo conseguiu reconhecer as diferenças relacionadas à disponibilidade de oxigênio dissolvido na coluna d'água no transcurso do dia, destacando acréscimos desta propriedade no período noturno. Uma das maiores dificuldades constatadas no modelo é a capacidade de identificar picos de acréscimo de OD.

Além disso, foi possível detectar a partir dos dados reais uma perda na qualidade da água no braço Taquacetuba, quando comprado com anos anteriores, pois, a disponibilidade de OD no meio oscila entre 7,0 mg/L até 8,0 mg/L, característica "irregular" para o padrão da estação devido à queda na temperatura do ambiente e consequentemente no meio aquático. Para avaliar se o modelo conseguiu reproduzir esta perda de qualidade e a precisão dos resultados do modelo, foi utilizado o erro percentual. Na Tabela 10 ilustra-se a comparação dos resultados para essa estação e seu respectivo erro percentual.

Tabela 10. Comparação dados reais vs simulados (Ano 2019) – Outono

Data	Hora	OD (Real)	OD (Simulado)	Erro (%)
24/3/19	0:00:00	7,1	8,195521	15,43
24/3/19	1:00:00	7,0	8,195521	17,08
24/3/19	2:00:00	7,0	8,151321	16,45
24/3/19	3:00:00	6,9	8,151321	18,14
24/3/19	4:00:00	6,7	8,151321	21,66
24/3/19	5:00:00	6,5	8,151321	25,40
24/3/19	6:00:00	6,3	8,11875	28,87
24/3/19	7:00:00	6,4	8,13735	27,15
24/3/19	8:00:00	6,6	8,13735	23,29
24/3/19	9:00:00	6,9	8,158779	18,24
24/3/19	10:00:00	7,4	8,158779	10,25
24/3/19	11:00:00	7,7	8,158779	5,96
24/3/19	12:00:00	8,1	8,158779	0,73
24/3/19	13:00:00	8,6	8,223779	4,37
24/3/19	14:00:00	9,3	8,151929	12,34
24/3/19	15:00:00	9,7	8,1445	16,04
24/3/19	16:00:00	9,9	8,1445	17,73
24/3/19	17:00:00	9,9	8,1445	17,73
24/3/19	18:00:00	9,6	8,234921	14,22
24/3/19	19:00:00	9,4	8,234921	12,39
24/3/19	20:00:00	9,0	8,234921	8,50
24/3/19	21:00:00	8,6	8,234921	4,25
24/3/19	22:00:00	8,4	8,177071	2,65
24/3/19	23:00:00	8,2	8,156271	0,53

Fonte: o Autor, 2024.

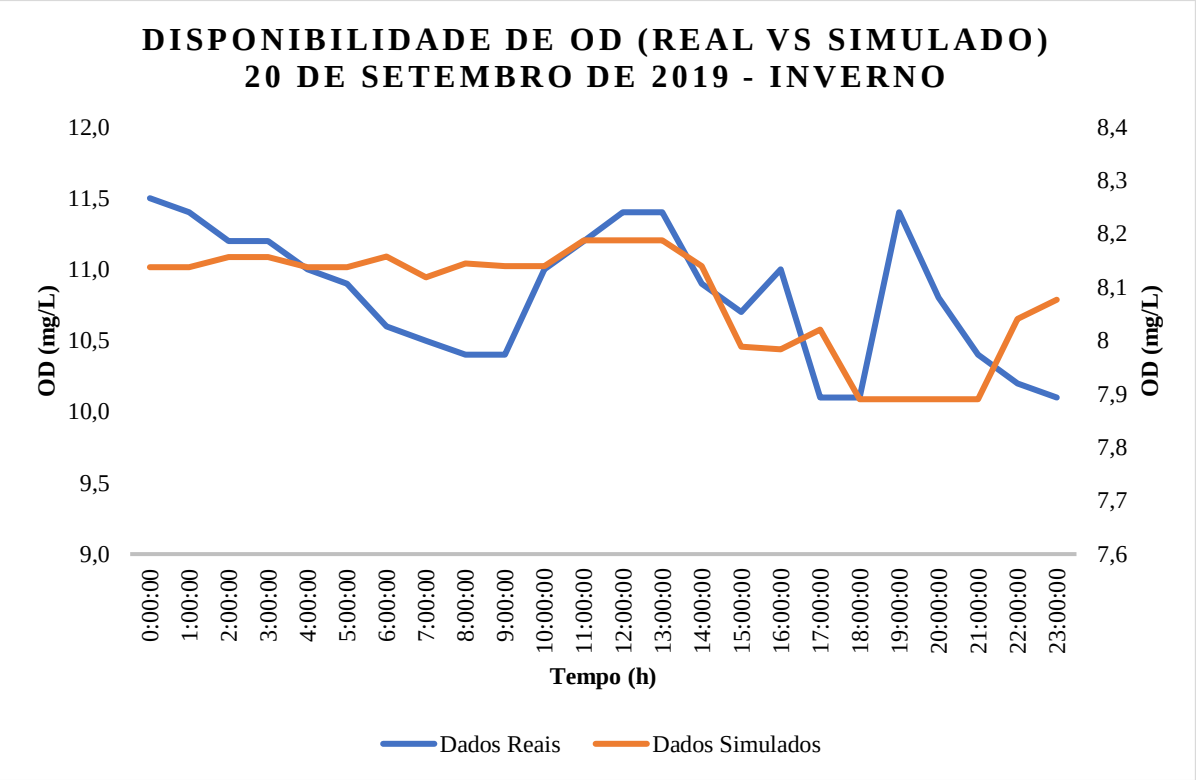
Em base aos resultados, foi possível determinar que devido a que os erros percentuais são maiores que 15%, houve uma diminuição generalizada na precisão da resposta de saída do OD disponível na coluna d'água quando comparada com o ano 2017. Porém, embora a resposta do modelo de regressão careça de propriedades de precisão, o modelo foi capaz de distinguir as mudanças de OD ocorridas durante o dia de forma satisfatória.

A pesar que os erros relativos percentuais de forma geral refletem uma diminuição na precisão dos valores, é importante destacar que o mesmo atingiu em vários momentos faixas menores ou próximas a 15%, o qual infere que a divergência dos valores na resposta de saída aconteceu em momentos ou períodos específicos e não na sua totalidade.

Na sequência, realizou-se a análise do comportamento tendencial para o inverno do ano 2019, foi selecionado de forma amostral o dia 20 de setembro de 2019,

representando de forma sintética, os últimos dias do inverno ao estar numa data próxima para a transição do inverno para a primavera. Na Figura 42 ilustra-se o gráfico que representa a disponibilidade de oxigênio dissolvido em função do tempo na data escolhida.

Figura 42. Disponibilidade de OD (Real vs simulado) – 20 de setembro de 2019



Fonte: o Autor, 2024.

Constatou-se que o modelo gerou respostas de saída com resultados menores quando comparadas aos dados reais, oscilando nas faixas de 8,0 mg/L até 9,0 mg/L, entretanto, tendo dificuldades na predição de valores maiores a essa faixa. No inverno, espera-se um aumento gradativo na quantidade de OD disponível (conforme os dados reais), entretanto o modelo apresentou dificuldades em ultrapassar a faixa dos 9,0 mg/L.

Contudo, o modelo de regressão caracterizou-se pela reprodução do comportamento tendencial de forma semelhante aos dados reais, reconhecendo as mudanças e características no transcurso do dia, tendo em conta as especificidades da estação.

Para avaliar a precisão das respostas de saída do modelo de regressão, assim como nos outros casos, foi utilizado o erro percentual, ferramenta estatística utilizada com a finalidade de estabelecer a porcentagem de erro existente ao comparar os dados reais e os dados experimentais. Na Tabela 11 encontram-se ilustrados os resultados referentes a

erro percentual obtidos na avaliação do modelo no inverno de 2019.

Tabela 11. Comparação dados reais vs simulados (Ano 2019) – Inverno

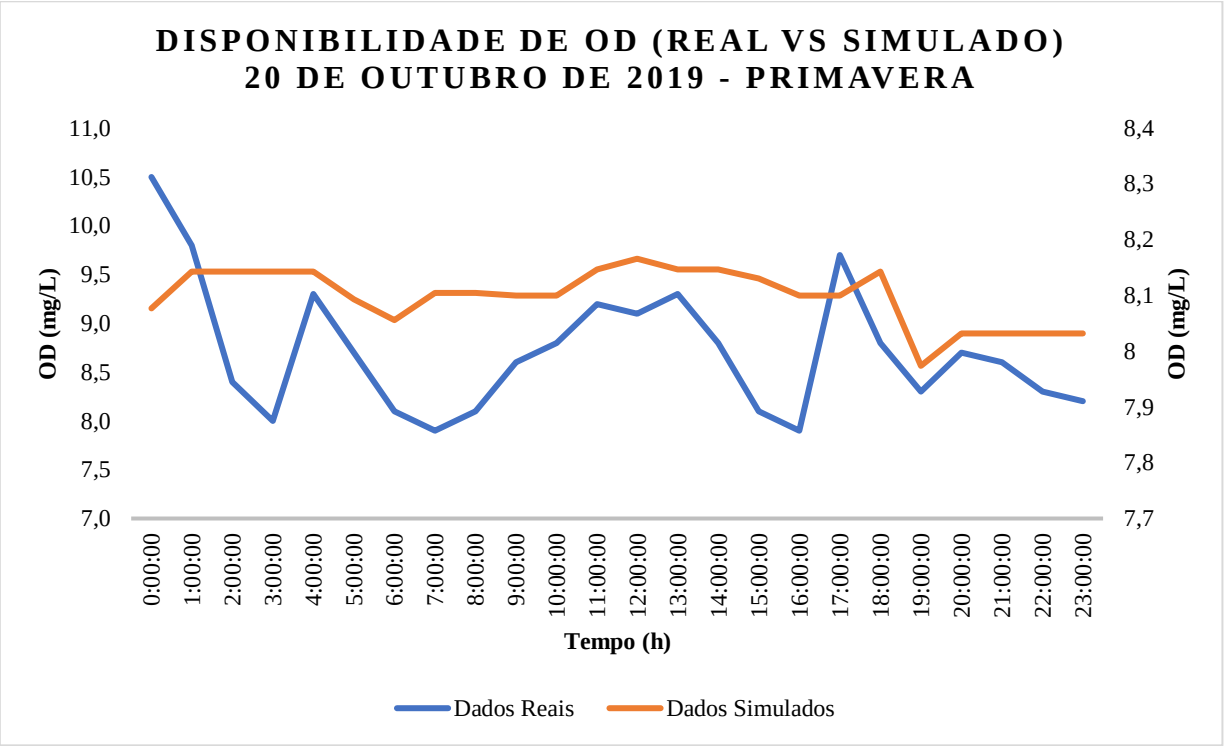
Data	Hora	OD (Real)	OD (Simulado)	Erro (%)
20/9/19	0:00:00	11,5	8,137671	29,24
20/9/19	1:00:00	11,4	8,137671	28,62
20/9/19	2:00:00	11,2	8,156271	27,18
20/9/19	3:00:00	11,2	8,156271	27,18
20/9/19	4:00:00	11,0	8,137671	26,02
20/9/19	5:00:00	10,9	8,137671	25,34
20/9/19	6:00:00	10,6	8,15745	23,04
20/9/19	7:00:00	10,5	8,11875	22,68
20/9/19	8:00:00	10,4	8,1445	21,69
20/9/19	9:00:00	10,4	8,1395	21,74
20/9/19	10:00:00	11,0	8,1395	26,00
20/9/19	11:00:00	11,2	8,18775	26,90
20/9/19	12:00:00	11,4	8,18775	28,18
20/9/19	13:00:00	11,4	8,18775	28,18
20/9/19	14:00:00	10,9	8,139	25,33
20/9/19	15:00:00	10,7	7,988267	25,34
20/9/19	16:00:00	11,0	7,983267	27,42
20/9/19	17:00:00	10,1	8,020838	20,59
20/9/19	18:00:00	10,1	7,890105	21,88
20/9/19	19:00:00	11,4	7,890105	30,79
20/9/19	20:00:00	10,8	7,890105	26,94
20/9/19	21:00:00	10,4	7,890105	24,13
20/9/19	22:00:00	10,2	8,040838	21,17
20/9/19	23:00:00	10,1	8,076838	20,03

Fonte: o Autor, 2024.

Neste caso, foi detectada uma menor precisão dos dados simulados quando comparados com os dados reais, e inclusive houve um aumento gradativo no erro relativo percentual chegando a atingir valores superiores a 30%, o que caracteriza a distinção de "ruídos" na resposta do modelo de regressão. Esse evento demonstra que o modelo possui dificuldades na reprodução de faixas de OD que atingem valores maiores a 9,0 mg/L ou de fato a picos de acréscimo dessa propriedade.

Seguidamente, foi realizada a plotagem dos dados da amostragem da primavera do ano 2019. Foi selecionado de forma amostral o dia 20 do mês de outubro para realizar essa análise, na Figura 43 encontra-se disponível o gráfico da quantidade de OD em função do tempo (real vs simulado) com os dados gerados após a simulação.

Figura 43. Disponibilidade de OD (Real vs simulado) – 20 de outubro de 2019



Fonte: o Autor, 2024.

Após ter realizado a simulação e plotado os gráficos, foi possível visualizar que as respostas de saída do modelo de regressão foram bastante semelhantes aos dados reais, conseguindo predizer a tendência do comportamento de OD em função do tempo de forma satisfatória com a presença de *overfitting* na maioria do tempo.

Em comparação com as outras respostas de aprendizado de máquina previamente avaliadas, constatou-se um ajuste adequado em relação aos dados reais, conseguindo a reprodução de valores próximos aos reais com a distinção da propriedade característica da estação, que é o começo da diminuição de OD devido ao aumento das temperaturas, além da atividade fotossintética e biológica no meio.

A tendência do comportamento relacionado à quantidade de oxigênio dissolvido simulado é observada a partir da obtenção de valores levemente maiores do que os dados reais, caracterizado pela criação de “picos” em alguns períodos do dia. A faixa de valores obtidos na simulação enquadrou-se nos valores reais, prevendo valores de OD coerentes à estação (característico da transição do inverno para a primavera).

A priori, nesse período foi possível constatar que os níveis de OD tiveram um comportamento padrão em comparação às outras estações do mesmo ano. Para realizar

uma verificação da precisão nas respostas do modelo de aprendizado de máquina, foi estimado o erro relativo percentual. Na Tabela 12 ilustram-se os valores obtidos após ter realizado a estimativa do erro percentual para a primavera de 2019 quando comparado com os dados reais (monitoramento da SABESP).

Tabela 12. Comparação dados reais vs simulados (Ano 2019) – Primavera

Data	Hora	OD (Real)	OD (Simulado)	Erro (%)
20/10/19	0:00:00	10,5	8,076838	23,08
20/10/19	1:00:00	9,8	8,142421	16,91
20/10/19	2:00:00	8,4	8,142421	3,07
20/10/19	3:00:00	8,0	8,142421	1,78
20/10/19	4:00:00	9,3	8,142421	12,45
20/10/19	5:00:00	8,7	8,093033	6,98
20/10/19	6:00:00	8,1	8,055461	0,55
20/10/19	7:00:00	7,9	8,10485	2,59
20/10/19	8:00:00	8,1	8,10485	0,06
20/10/19	9:00:00	8,6	8,1001	5,81
20/10/19	10:00:00	8,8	8,1001	7,95
20/10/19	11:00:00	9,2	8,146929	11,45
20/10/19	12:00:00	9,1	8,165929	10,26
20/10/19	13:00:00	9,3	8,146929	12,40
20/10/19	14:00:00	8,8	8,146929	7,42
20/10/19	15:00:00	8,1	8,13015	0,37
20/10/19	16:00:00	7,9	8,1001	2,53
20/10/19	17:00:00	9,7	8,1001	16,49
20/10/19	18:00:00	8,8	8,142421	7,47
20/10/19	19:00:00	8,3	7,973794	3,93
20/10/19	20:00:00	8,7	8,031644	7,68
20/10/19	21:00:00	8,6	8,031644	6,61
20/10/19	22:00:00	8,3	8,031644	3,23
20/10/19	23:00:00	8,2	8,031644	2,05

Fonte: o Autor, 2024.

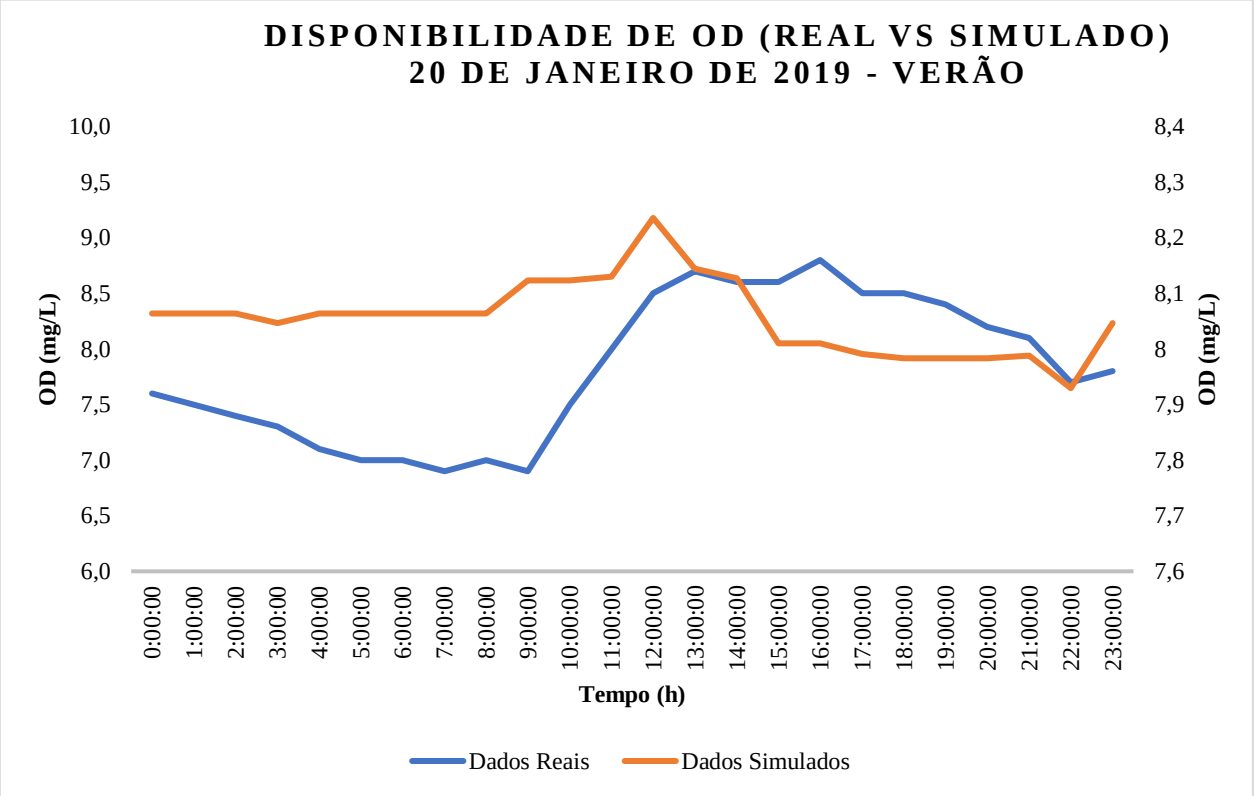
Observou-se que salvo casos específicos. os erros relativos percentuais do dia amostrado não apresentaram valores maiores a 15% o que infere sobre a precisão na resposta do modelo, além disso, constatou-se que os valores de OD adequam-se coerentemente na faixa característica da primavera conseguindo diferençar satisfatoriamente as mudanças de OD no meio no dia e na noite.

Conforme esperado nos sistemas naturais, e sobretudo na avaliação dos sistemas hidrográficos, na primavera ocorre uma queda gradativa na quantidade de OD disponível na coluna d'água devido a vários fatores bioquímicos, climáticos e biológicos de

forma generalizada.

Para finalizar, foi realizada a análise amostral do verão de 2019 para verificar a eficácia no reconhecimento das variações sazonais características desta estação, diante isso, foi selecionado o dia 20 de dezembro de 2019 como representante amostral deste período. Na Figura 44 ilustra-se o gráfico dos valores de OD (real vs simulado) em função do tempo no verão do ano 2019.

Figura 44. Disponibilidade de OD (Real vs simulado) – 20 de janeiro de 2019



Fonte: o Autor, 2024.

Destacou-se que a resposta do modelo de regressão obteve comportamento semelhante quando comparado com os dados reais, entretanto, sua tendência foi a reprodução de valores ligeiramente maiores aos mesmos, assim como nos outros casos, a resposta do modelo de regressão conseguiu reconhecer a mudança na disponibilidade de OD num período de 24 horas.

Embora tenham sido destacados resultados positivos nas respostas de saída ao reconhecer as diferenças na quantidade de OD em diversos horários do dia, foram percebidos alguns picos característicos de “ruídos” na resposta da simulação, oscilando entre valores maiores e menores no algoritmo ao comparar os dados simulados com os dados reais.

Em seguida, foram estimados os erros relativos percentuais com a finalidade de verificar a precisão das respostas de saída do modelo. Na Tabela 13, encontram-se disponíveis os valores referentes ao erro relativo percentual estimado para o dia 20 de janeiro de 2019 (verão) ao comparar os dados reais e os dados simulados.

Tabela 13. Comparação dados reais vs simulados (Ano 2019) – Verão

Data	Hora	OD (Real)	OD (Simulado)	Erro (%)
20/1/19	0:00:00	7,6	8,064255	6,11
20/1/19	1:00:00	7,5	8,064255	7,52
20/1/19	2:00:00	7,4	8,064255	8,98
20/1/19	3:00:00	7,3	8,046755	10,23
20/1/19	4:00:00	7,1	8,064255	13,58
20/1/19	5:00:00	7,0	8,064255	15,20
20/1/19	6:00:00	7,0	8,064255	15,20
20/1/19	7:00:00	6,9	8,064255	16,87
20/1/19	8:00:00	7,0	8,064255	15,20
20/1/19	9:00:00	6,9	8,123017	17,72
20/1/19	10:00:00	7,5	8,123017	8,31
20/1/19	11:00:00	8,0	8,130445	1,63
20/1/19	12:00:00	8,5	8,235779	3,11
20/1/19	13:00:00	8,7	8,144445	6,39
20/1/19	14:00:00	8,6	8,126945	5,50
20/1/19	15:00:00	8,6	8,009695	6,86
20/1/19	16:00:00	8,8	8,009695	8,98
20/1/19	17:00:00	8,5	7,990695	5,99
20/1/19	18:00:00	8,5	7,983267	6,08
20/1/19	19:00:00	8,4	7,983267	4,96
20/1/19	20:00:00	8,2	7,983267	2,64
20/1/19	21:00:00	8,1	7,988267	1,38
20/1/19	22:00:00	7,7	7,929505	2,98
20/1/19	23:00:00	7,8	8,046755	3,16

Fonte: o Autor, 2024.

Observou-se que as respostas do modelo de regressão lançaram valores cujo erro relativo percentual estimado na maioria dos casos foi menor que 15%, o que infere que as respostas de saída foram bastantes precisas em relação aos dados reais, ademais, o modelo conseguiu prever as principais diferenças diurnas e noturnas (além das sazonais) na quantidade de OD, tendo em conta as especificidades que caracterizam o verão (tendência a possuir uma menor quantidade de OD devido ao incremento da temperatura do ambiente).

O modelo floresta aleatória teve uma diminuição na precisão dos resultados para o ano 2019, sobretudo nas estações onde a disponibilidade de OD é maior

(outono e inverno), ademais, em comparação às respostas do ano 2017, o modelo possui dificuldades na reprodução de valores não característicos da estação, os quais podem ser originários de fontes pontuais de poluição não previsíveis e muitas vezes ilegais. Porém, em todos os casos o modelo conseguiu reconhecer as diferenças sazonais na quantidade de OD além da distinção desse parâmetro nos períodos diurnos e noturnos, gerando em alguns casos valores bastante precisos quando comparado aos dados reais.

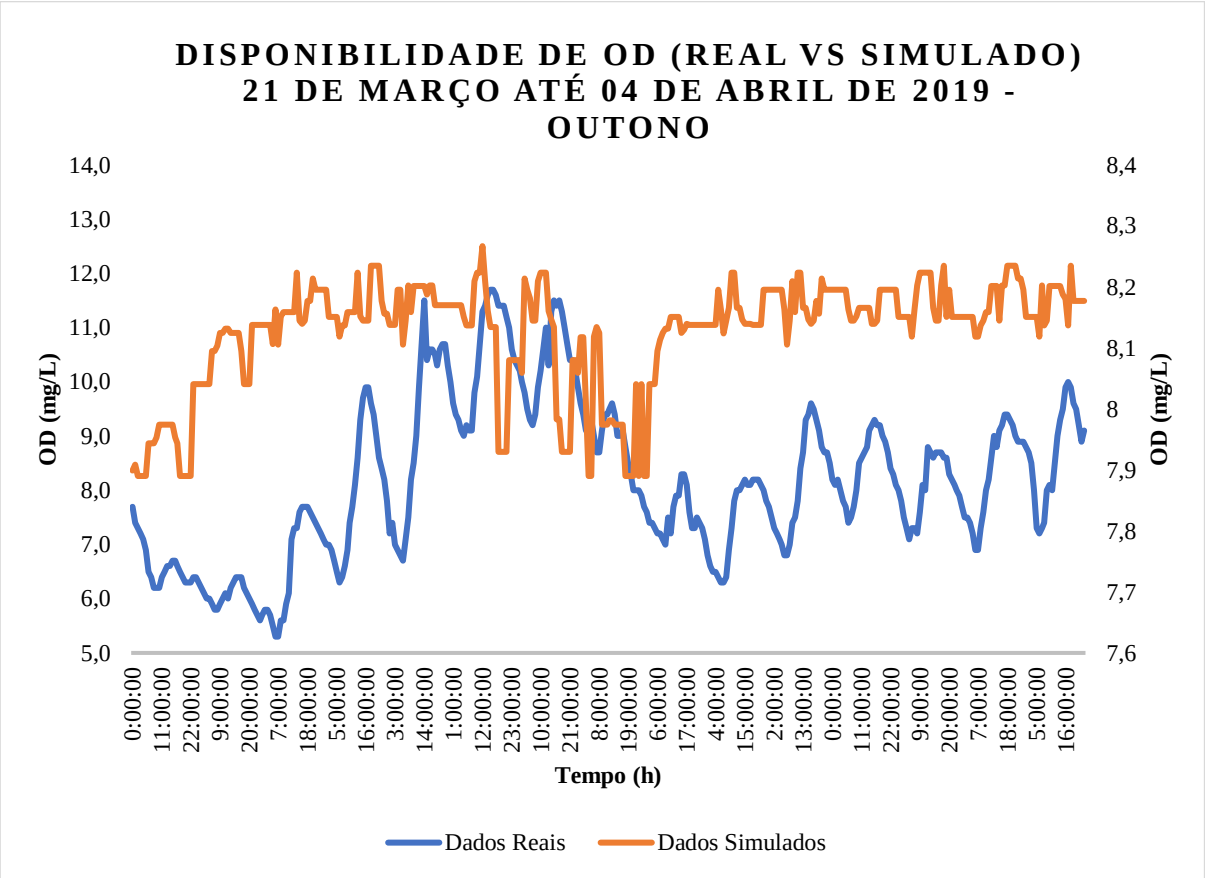
Em termos gerais, pode-se afirmar que o modelo conseguiu se ajustar melhor sempre que os valores reais eram mais baixos, demonstrando que a limitação do modelo se deu em termos de faixas limitadas de resultados, contudo ele conseguiu mostrar um bom desempenho em relação à reprodução das dinâmicas diárias e sazonais, sugerindo a necessidade de ampliação do banco de dados de treinamento a fim de buscar maior variabilidade de situações reais às quais os sistemas lóticos estão submetidos.

6.3.2. Avaliação da dinâmica semanal dos resultados da simulação de 2019 para cada estação do ano

Semelhante à análise sazonal realizada para o ano 2017, no ano 2019 foram plotados gráficos de quinze dias consecutivos (duas semanas) em cada uma das estações para realizar uma avaliação mais abrangente das respostas de saída do modelo de regressão, e dessa forma avaliar possíveis mudanças no que respeita à quantidade de OD e consequentemente à qualidade da água do braço Taquacetuba do complexo Billings.

Para avaliar o comportamento sazonal, plotaram-se os dados referentes ao período de 21/03/2019 até 04/04/2019 para totalizar 15 dias representativos do outono do ano 2019. Na Figura 45 encontra-se disponível o gráfico comparativo (real vs simulado) do período selecionado para a avaliação da análise sazonal característica do outono.

Figura 45. Análise Sazonal (real vs simulado) – Outono de 2019



Fonte: o Autor, 2024.

De forma generalizada, constatou-se a existência do fenômeno de *overfitting* ou sobreposição dos valores nas respostas de saída do modelo de regressão ao avaliar o outono de 2019, somente após uma semana o modelo conseguiu se ajustar na faixa de OD real, oscilando entre valores ligeiramente menores e maiores até atingir a metade da segunda semana da quinzena avaliada.

Nessa avaliação, constatou-se um fenômeno atípico não característico da estação de outono, o qual estava relacionado à diminuição abrupta de OD no meio, chegando a atingir valores próximos a 5,0 mg/L, diante isso, foi possível destacar que o modelo apresentou um atraso para mostrar uma queda nos valores e não conseguiu predizer os valores mais baixos de OD, os quais estavam disponíveis nos dados reais, o que precisa ser melhor compreendido.

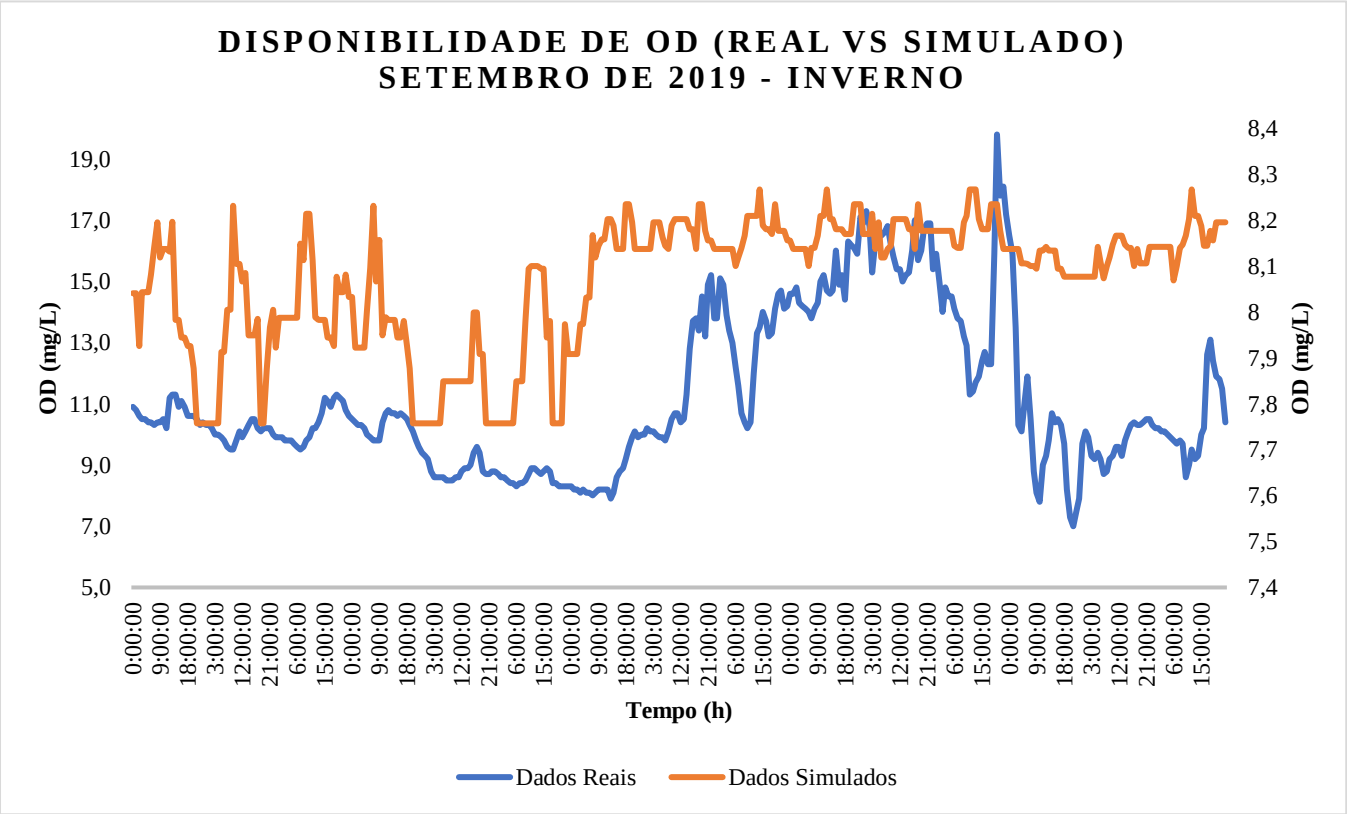
Por outro lado, nos resultados do modelo de regressão, as respostas de saída indicam uma permanência dentro da faixa padrão do inverno, sendo capaz de reconhecer a tendência e comportamento relacionado à disponibilidade de oxigênio dissolvido disponível na coluna d'água, diferençando as oscilações que poderiam ocorrer

diariamente, além disso, devido à presença do fenômeno de *overfitting*, foram detectados vários picos que a pesar de serem semelhantes aos dados reais, ocorrem de forma mais frequente, caracterizando "ruídos" em alguns períodos em específico.

A queda nos níveis de OD no outono não é característica da época, e poderiam indicar altos níveis de poluição na água, resultando em uma deterioração da qualidade da água, além de apresentar um alerta no que respeita à flora e fauna no braço Taquacetuba, com tendência à eutrofização.

Após a avaliação sazonal (comportamento e tendência) do outono do ano 2019, procedeu-se a realizar a mesma verificação para o inverno do mesmo ano. A tendência no inverno pressupõe um aumento na disponibilidade de OD em função da queda da temperatura, nesta verificação, foi utilizado como conjunto de amostragem os primeiros 15 dias do mês de setembro do ano 2019. Na Figura 46 ilustra-se o comportamento (simulado vs real) na análise sazonal do inverno do ano 2019.

Figura 46. Análise Sazonal (real vs simulado) – Inverno de 2019



Fonte: o Autor, 2024.

Foi constatado que a resposta do modelo de regressão apresentou a

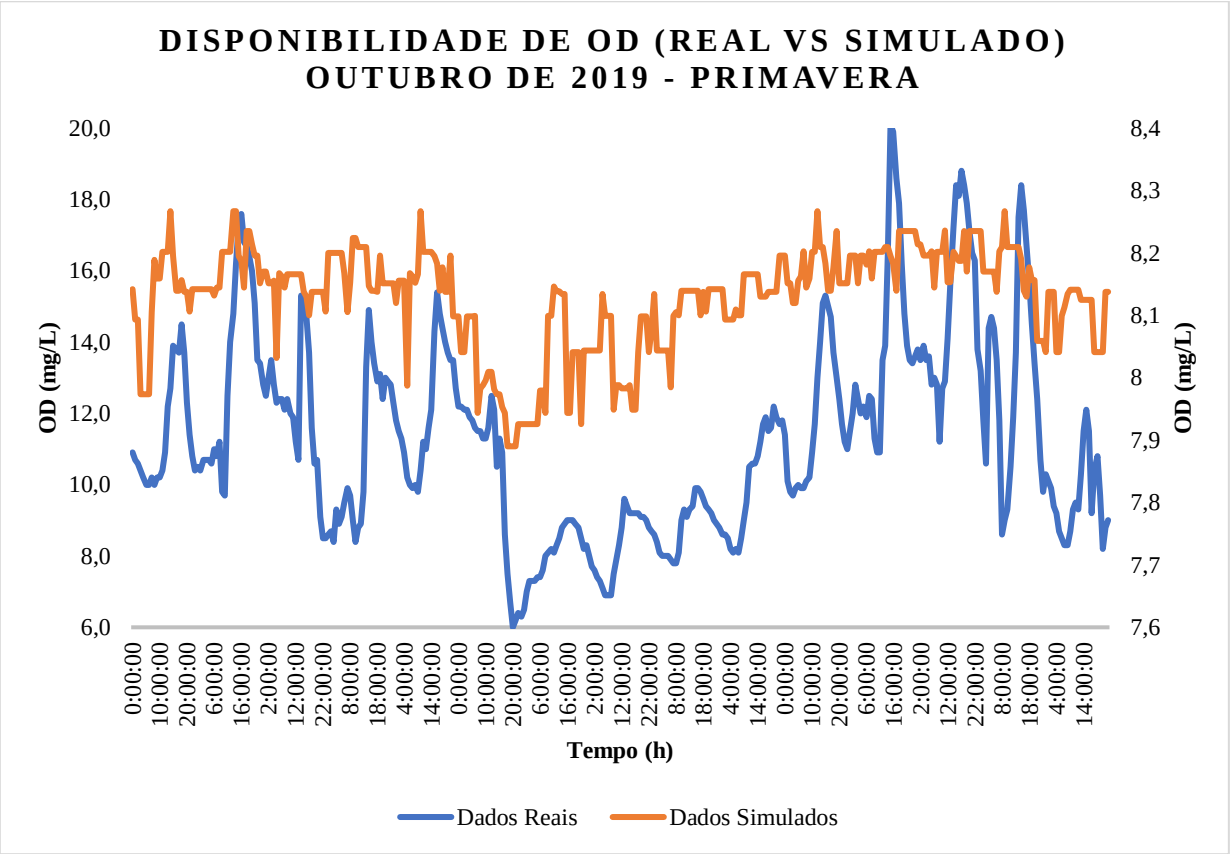
reprodução de valores menores numa escala inferior aos dados reais, tendo como característica principal a dificuldade em simular picos de aumento ou diminuição de OD, a pesar do modelo conseguir “entender” relativamente as diferenças diárias que existem num período de 15 dias relacionadas à disponibilidade de OD, a faixa de valores na qual o modelo se ajusta é sempre menor que a faixa de valores reais.

Os dados reais refletem uma melhora destacada na qualidade da água do braço Taquacetuba da represa Billings, uma vez que os valores de OD aumentaram quando comparado com a estação de outono. A pesar da falta de precisão notável das respostas do modelo de regressão, ele consegue perceber as diferenças desde a perspectiva sazonal que existem ao longo dos 15 dias, pois apresentou um aumento relativo dos valores de OD em relação à estação anterior, conseguindo reproduzir um comportamento muito semelhante aos dados reais, entretanto, com valores menores.

Embora, este vicio não seja desejado, e de fato seja inadequado devido à falta de precisão, tanto os dados reais (sonda de monitoramento da SABESP), quanto os dados simulados conseguem mostrar uma informação valiosa sobre a qualidade da água da represa Billings, uma vez que ambos os resultados sugerem que a qualidade da água do complexo hidrográfico encontra-se dentro dos padrões estabelecidos na legislação.

Seguidamente, realizou-se a avaliação sazonal (tendências e comportamento) para a primavera do ano 2019, a qual está caracterizada por um aumento generalizado na atividade fotossintética, aumento na atividade biológica, além do aumento gradativo de temperatura. Para essa avaliação escolheu-se a faixa representativa pertencente aos primeiros 15 dias do mês de outubro. Na Figura 47 ilustra-se o gráfico contendo o comportamento (real vs simulado) para a análise sazonal da primavera do ano 2019.

Figura 47. Análise Sazonal (real vs simulado) – Primavera de 2019



Fonte: o Autor, 2024.

De forma geral, o modelo conseguiu reproduzir valores com uma maior precisão em vários períodos da primavera, conseguindo se ajustar em vários momentos às faixas de OD reais. Contudo, semelhante às outras estações, a reprodução de valores não conseguiu reproduzir a alta significativa nos valores de OD e nem distinguir as quedas abruptas na quantidade de OD em períodos específicos.

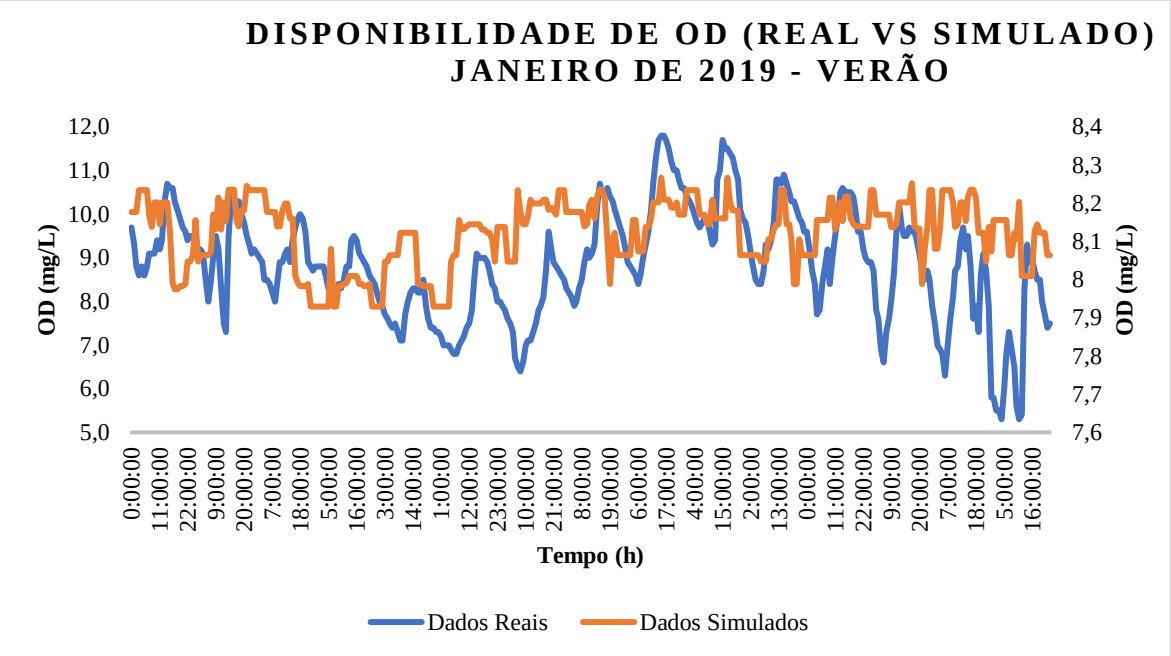
O modelo conseguiu predizer o comportamento tendencial "padrão" da primavera ao longo do tempo, sendo capaz de diferençar levemente às diferenças relacionadas à quantidade de OD tendo em conta às especificidades sazonais. A pesar das dificuldades do modelo em reconhecer as quedas ou aumentos abruptos (picos), as conclusões relacionadas à qualidade de água tanto para os dados reais, quanto para os dados simulados é a mesma, conseguindo se encaixar de forma satisfatória na faixa de valores mínimos estabelecidos pela legislação.

Os picos negativos na quantidade de OD na primavera está intrinsicamente relacionada à dinâmica dos sistemas naturais, uma vez que a primavera possui um comportamento oscilante, onde ocorrem ao mesmo tempo vários processos fotossintéticos,

reinicia-se a atividade biológica, além do fato deste período, se encontrar numa faixa de transição sazonal, com oscilações mais significativas nas temperaturas diárias.

Finalmente, realizou-se análise sazonal (comportamento e tendências reais vs simulados) para o verão do ano 2019, para esta verificação, utilizaram-se de forma amostral os primeiros 15 dias do mês de janeiro. Na Figura 48 encontra-se disponível o gráfico representando o comportamento (real vs simulado) na análise sazonal realizada para o verão do ano 2019.

Figura 48. Análise Sazonal (real vs simulado) – Verão de 2019



Fonte: o Autor, 2024.

No caso do verão de 2019, a diferença das outras estações, o modelo conseguiu reproduzir de forma muito precisa às quantidades de OD disponíveis na coluna d'água, permitindo se ajustar dentro das faixas reais de forma satisfatória tendo em conta às especificidades diurnas e noturnas. No entanto, de forma semelhante aos outros casos de análise sazonal do ano 2019, ele possui dificuldade em distinguir quedas abruptas, picos ou comportamentos não característicos da estação, reproduzindo valores apenas que se encontram dentro da faixa padrão, ou seja, nos valores que iriam acontecer normalmente seguindo a planilha de treinamento.

Além disso, foi possível perceber a capacidade do modelo em replicar valores de OD capazes de diferençar a quantidade do mesmo no dia e na noite. A pesar da

presença de picos abruptos caracterizando uma queda generalizada de OD em diversos pontos do gráfico, a resposta do modelo não promove uma má interpretação nos resultados ou na faixa permitida na legislação, pois tanto os valores reais, quanto os valores simulados são superiores a 5,0 mg/L.

O monitoramento por sonda da SABESP do ano 2019 foi caracterizado por possuir muitas falhas no decorrer do ano, apresentando muitos meses incompletos e inclusive a ausência generalizada de um amplo conjunto de dados no dia a dia, o que possivelmente veio a dificultar o ajuste do modelo de regressão e consequentemente promovendo o *overfitting* devido à limitação dos dados e o *loop* em alguns períodos do ano.

No entanto, salvo o caso do inverno do ano 2019, o modelo conseguiu predizer de forma satisfatória a tendência e o comportamento "padrão" que teria a qualidade da água a partir dos parâmetros de pH, temperatura, além das informações do monitoramento por satélite da NASA.

Sugere-se que maiores estudos sejam realizados coma finalidade de resolução dos problemas de *overfitting* de modo à oferecer uma alternativa segura para baratear o processo de monitoramento da qualidade da água da represa Billings, que poderia utilizar uma menor quantidade de ferramentas e instrumentais industriais, conseguindo então, tornar mais eficiente o processo de acompanhamento e monitoramento da qualidade da água.

Recomenda-se em trabalhos futuros a utilização de um banco de dados na planilha de treinamento com maiores informações, de tal forma que os modelos de aprendizado de máquina possam distinguir de forma mais adequadas as quedas ou aumentos abruptos ou não característicos de alguns períodos anuais ou sazonais.

7 CONSIDERAÇÕES FINAIS

A ciência de dados tem permitido o desenvolvimento abrangente de diversos campos do conhecimento, entre eles, a Engenharia Química, área de atuação que envolve um amplo conjunto de ramos que permitem a aplicação desta disciplina e de aprendizado de máquina, permitindo a otimização de processos e a implementação de estratégias que visam a melhorar as condições ou eficiência dos métodos tradicionais.

No presente trabalho aplicaram-se modelos de aprendizado de máquina com o objetivo de simular a quantidade de oxigênio dissolvido (OD) utilizando a menor quantidade possível dados experimentais das sondas da SABESP, num período de 2 anos.

A ampla disponibilidade de dados disponibilizadas em sites governamentais de acesso público, permitiram a criação e estruturação de banco de dados que permitem sua aplicação utilizando modelos de aprendizado de máquina, cabendo a apertura de amplas possibilidades de aplicação em diversas áreas do conhecimento, como é o caso do monitoramento da qualidade da água (controle de processo).

Constatou-se que devido à natureza não linear da maioria das variáveis que compõem os sistemas naturais além de análise estatística, o modelo mais adequado no ajuste e nas predições foi o modelo Floresta Aleatória.

Os resultados de oxigênio dissolvido obtidos na verificação de 2017, utilizando apenas a variável de pH e os dados de monitoramento por satélite da NASA, inferem uma alta precisão ao compará-los com os dados reais.

Para a simulação do ano 2019 foram utilizados os parâmetros de pH, temperatura e os dados de monitoramento por satélite da NASA, obtendo para duas estações valores mais precisos, já para as outras duas, achou-se o fenômeno de *overfitting* e a presença de *loops* que podem indicar dificuldade no modelo para interpretar picos ou comportamentos atípicos, não característicos da estação

O modelo de aprendizado de máquina foi capaz de detectar as diferenças sazonais existentes na quantidade de OD para os dois anos avaliados, além de destacar a dinâmica entre o dia e na noite de forma satisfatória.Na maioria dos casos estudados, os resultados do modelo estão próximos da faixa real e conseguiram reproduzir as dinâmicas sazonais e diárias, o que abre caminho para pesquisas visando à redução dos custos associados ao monitoramento da qualidade da água na represa Billings, por meio da diminuição do número de sondas requeridas para tal procedimento. Embora maiores estudos

sejam necessários, o estudo realizado fornece subsídios para estudos futuros, é um indicativo significativo dessa possibilidade além de antecipar avanços futuros nessa área.

REFERÊNCIAS BIBLIOGRÁFICAS

ACS. **¿Qué es el estigma? ¿Qué es el estigma?iAgua Respuestas**, 2017a. Disponível em: <https://unlp.edu.ar/frontend/media/98/27598/3f23fc987dbbeda82587753c9796000a.pdf>

AHMED, U. et al. Efficient water quality prediction using supervised machine learning. **Water (Switzerland)**, v. 11, n. 11, 2019.

ALDANA, C. R. B. et al. **Modelado y simulación de sistemas naturales**. [s.l.] Universidad Jorge Tadeo Lozano, 2019.

ALVAREZ, R. D. C. F. et al. Análise de crescimento de duas cultivares de amendoim (Arachis hypogaea L.). **Acta Scientiarum. Agronomy**, v. 27, n. 4, 2005.

ALVES, M. A. S.; ANDRADE, O. M. DE. Da "caixa-preta" à "caixa de vidro": o uso da explainable artificial intelligence (XAI) para reduzir a opacidade e enfrentar o enviesamento em modelos algorítmicos. **Direito Público**, v. 18, n. 100, 2022.

ALVIM, A. T. B. et al. **Meio ambiente, urbanização e assentamentos precários: desafios para os projetos urbanos contemporâneos no Brasil**. 2022

ANDRÉ LIMA DA COSTA et al. Dinâmica Da Qualidade Da Água Superficial No Campus Da Ufpa Em Belém-Pa. **Studies in Environmental and Animal Sciences**, v. 2, n. 3, p. 81–119, 2022.

ANÓNIMO. CONDUCTIVIDAD ELÉCTRICA Aspectos teóricos. **Quimica**, 2014.

ANTONIO LOURENÇO, L. CIÊNCIA DE DADOS NA ENGENHARIA QUÍMICA: Um estudo bibliométrico. [s.d.].

ATKINS, P. W. **Química Inorgânica**. [s.l: s.n.].

BARRETO, L. V. et al. Eutrofização Em Rios Brasileiros. **ENCICLOPÉDIA BIOSFERA, Centro Científico Conhecer**, v. 9, n. 16, p. 2165–2179, 2013.

BELENDEZ, A. Calor y temperatura. **Departamento de Física, Ingeniería de Sistemas y Teoría de la Señal**, p. 1–21, 2017.

Blog de Físico-Química do André: Escala de pH. Disponível em: <http://andre-godinho-cfq-8a.blogspot.com/2012/12/escala-de-ph.html>. Acesso em: 11 jul. 2022.

BODIN, Ö.; CRONA, B.; ERNSTSON, H. Las redes sociales en la gestión de los recursos naturales : ¿ Qué hay que aprender de una perspectiva estructural ? **Redes**, v. 28, p. 1–8, 2017.

BOYD, C. Conductividad eléctrica del agua, parte 1. **Global Aquaculture Advocate**, 2017.

CAMILOTTI, E. et al. Redes neurais artificiais para o gerenciamento da indústria avícola: uma simulação baseada na cadeia de produção de frangos de corte. **Ciência Animal Brasileira**, v. 24, 2023.

CARBOTECNIA. **¿Qué es el pH del agua?** Disponível em: <https://www.carbotecnia.info/encyclopedia/que-es-el-ph-del-agua/#disqus_thread>.

CARDOSO-SILVA, S. et al. Compartimentalização e qualidade da água: o caso da Represa Billings Compartmentalization and water quality: Billings reservoir case Viviane Moschini-Carlos 3. **Bioikos**, v. 28, n. 1, p. 31–43, 2014.

CETESB - SP. **CETESB**.

CHICCO, D.; WARRENS, M. J.; JURMAN, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. **PeerJ Computer Science**, v. 7, p. 1–24, 2021.

COLUNGA, A.; OLGUIN GRANADOS, V. D.; VARELA TOVAR, O. Mecanismos de transferencia de calor. **TEPEXI Boletín Científico de la Escuela Superior Tepeji del Río**, v. 7, n. 14, p. 58–61, 2020.

COND, M. Graus de trofia em corpos d ' água do E s tado de S ão P aulo : **Instituto de Biociências**, v. Tese de Do, 2004.

CORREA, M. G. G.; GALVANI, E. Variabilidade Espacial e Sazonal da Precipitação Pluviométrica na Bacia Hidrográfica no Rio Piquiri-PR. **Geography Department University of Sao Paulo**, v. 34, p. 21, 2017.

CORREIA, A. et al. Análise da Turbidez da Água em Diferentes Estados de Tratamento 5 . Determinação da Turbidez em. **Encontro Regional de Matemática Aplicada e Computacional - ERMAC**, v. 1, p. 2–6, 2008.

COSTA, M. N. DE M.; BECKER, C. T.; BRITO, J. I. B. DE. Análise Das Séries Temporais De Precipitação Do Semiárido Paraibano Em Um Período De 100 Anos - 1911 A 2010 (Analysis Of The Time Series Of Precipitation Of The Paraiba Semi Arid In A Period Of 100 Years - 1911 To 2010). **Revista Brasileira de Geografia Física**, v. 6, n. 4, p. 680–696, 2013.

COUTINHO, C. DE Q. E S.; FIGUEIREDO, A. DE C. Simulação computacional. **Zetetike**, v. 28, p. e020017, 2020.

DE LIMA[1], Thainara Munhoz Alexandre; CARVALHO[1], Rômulo Marques; DE ALMEIDA[1], Cláudia Maria. **IMPACT OF THE EXPANSION OF URBAN IMPERVIOUS SURFACES ON CYANOBACTERIAL BLOOM IN THE BILLINGS RESERVOIR**, SP-BRAZIL. 2023.

DIECKMANN, P. La simulación es más que Tecnología : el ambiente de la simulación. **Universidad de Copenhague**, 2011.

DISPERSI, D. D. E.; PUNTOS, D. E. L. O. S. Coeficiente de correlación de pearson. **Glosarios@Servidor-Alicante.Com**, p. 1–6, 2015a.

DISPERSI, D. D. E.; PUNTOS, D. E. L. O. S. **Coeficiente de correlación de pearson**. Disponível em: <https://glosarios.servidor-alicante.com/terminos-estadistica/coeficiente-de-correlacion-de-pearson>.

DOMÍNGUEZ, R. et al. **Recursos naturales, medio ambiente y sostenibilidad.** [s.l: s.n.].

DUFERA, A. G.; LIU, T.; XU, J. Regression models of Pearson correlation coefficient. **Statistical Theory and Related Fields**, v. 7, n. 2, p. 97–106, 2023.

ERASO CHECA, F.; ESCOBAR ROSERO, E. Metodología para la determinación de características del viento y evaluación del potencial de energía eólica en Túquerres-Nariño. **Revista científica**, v. 1, n. 31, p. 19–31, 2018.

F. FRANCO, E.; J. RAMOS, R. Aprendizaje de máquina y aprendizaje profundo en biotecnología: aplicaciones, impactos y desafíos. **Ciencia, Ambiente y Clima**, v. 2, n. 2, p. 7–26, 14 dez. 2019.

FARIAS, A. H. T. DE; CARVALHO, J. B. DE. **Machine learning e análise estatística de formas: uma aplicação no espaço tangente**. 2023

FERREIRA, C. DOS S.; CUNHA-SANTINO, M. B. DA; BIANCHINI JÚNIOR, I. Eutrofização: aspectos conceituais, usos da água e diretrizes para a gestão ambiental. **Revista Ibero-Americana de Ciências Ambientais**, v. 6, n. 1, p. 65–77, 2015.

FIORUCCI, A. R.; FILHO, E. B. Oxigênio dissolvido: propriedades e solubilidade. **Química Nova na Escola**, v. 22, n. 4, p. 10–16, 2005.

FRANCO, I. C. Modelagem e Simulação: Processos Químicos. **The Journal of Engineering and Exact Sciences**, v. 7, n. 4, p. francoic2021- 01, 2021.

GARCÍA, F.; MIRANDA, V. Eutrofización, una amenaza para el recurso hídrico. **Volumen II de la Colección: Agenda pública para el desarrollo regional, la metropolización y la sostenibilidad**, p. 35–367, 2018.

GIORDANO, G. Tratamento e controle de efluentes industriais. **Depto de Engenharia Sanitária e do Meio Ambiente, …**, 2004.

GÓMEZ, J. **Feminicidios: causas, consecuencias y soluciones**. Disponível em: <http://hoy.com.do/feminicidios-causas-consecuencias-y-soluciones/>.

GÓMEZ PUERTO, Á. B. Estado social y medio ambiente. **Revista de Fomento Social**, p. 409–429, 2021.

GONZÁLEZ TORO, C. La turbidez. **Servicio de Extension Agricola**, p. 11, 2014.

GOYENOLA, G. Transparencia, color y turbidez. **Guía para la utilización de las Valijas Viajeras - Transparencia, etc**, p. 2, 2007.

GRANT, ET AL. **Ecología Y Manejo De Recursos Naturales; Análisis De Sistemas De SimulaciónEd. IICA**, 2001.

GUERRERO, W. M. Folleto Informativo Oxígeno Disuelto (OD). **Folleto Informativo**, 2015.

HANISCH, W.; FREIRE-NORDI, C. S. Monitoramento remoto em tempo real de

mananciais visando às florações de cianobactérias. **Ecologia de reservatórios e interfaces**, p. 190–211, 2015.

HERNÁNDEZ, J. et al. Sobre el uso adecuado del coeficiente de correlación de Pearson. **Archivos Venezolanos de Farmacología y Terapéutica**, v. 37, n. 5, p. 586–601, 2018.

HERREIRA, B. N.; PIZELLA, D. G. Gestão Da Qualidade Hídrica Na Bacia Hidrográfica Do Rio Tietê (Sp): Dificuldades Para O Enquadramento Das Águas Doces Superficiais. **Revista Gestão & Sustentabilidade Ambiental**, v. 9, n. 2, p. 332, 2020.

IDEAM. Atlas Velocidad Del Viento. **Renovables**, 2018.

INZUNZA, J. Radiación solar y terrestre. **Ciencias Integradas**, 2019.

IRRGANG, C.; SAYNISCH-WAGNER, J.; THOMAS, M. Machine Learning-Based Prediction of Spatiotemporal Uncertainties in Global Wind Velocity Reanalyses. **Journal of Advances in Modeling Earth Systems**, v. 12, n. 5, 2020.

JABBAR, H.; KHAN, Rafiqul Zaman. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). **Computer Science, Communication and Instrumentation Devices**, v. 70, n. 10.3850, p. 978-981, 2015.

JARDIM, W. F. Medição e interpretação de valores do potencial redox (EH) em matrizes ambientais. **Quimica Nova**, v. 37, n. 7, p. 1233–1235, 2014.

JIMENEZ, V. A.; WILL, A.; RODRÍGUEZ, S. Estimación De Radiación Solar Horaria Utilizando Modelos Empíricos Y Redes Neuronales Artificiales. **Ciencia y Tecnología**, v. 1, n. 17, p. 29, 2017.

JONES, D. G. et al. Monitoring of near surface gas seepage from a shallow injection experiment at the CO2 Field Lab, Norway. **International Journal of Greenhouse Gas Control**, v. 28, p. 300–317, 2014.

KOHATSU, M. Y. et al. Fitotoxicidade de água superficial da Região Metropolitana de São Paulo utilizando bioensaio com Sinapis alba. **Acta Brasiliensis**, v. 2, n. 2, p. 58, 2018.

KORANGA, M. et al. Efficient water quality prediction models based on machine learning algorithms for Nainital Lake, Uttarakhand. **Materials Today: Proceedings**, v. 57, p. 1706–1712, 1 jan. 2022.

KOTZ, J. C.; TREICHEL, P. M.; WEAVER, G. C. Química y Reactividad Química. In: **Química y Reactividad Química**. [s.l: s.n.]. p. 1292.

LEAL IGA, C.; LEAL IGA, J.; SOL, E. Radiacion Solar. **Universidad Autonoma de Nuevo Leon**, v. 76, p. 7, 2015.

LIU, Yanli; WANG, Yourong; ZHANG, Jian. New machine learning algorithm: Random forest. In: **Information Computing and Applications: Third International Conference, ICICA 2012**, Chengde, China, September 14-16, 2012. Proceedings 3. Springer Berlin Heidelberg, 2012. p. 246-252.

LOZANO-RIVAS, W. A. Precipitación. In: **Clima, hidrología y meteorología**. [s.l: s.n.]. p. 251–298.

LUKASSEN, M. B. et al. Impact of water quality parameters on geosmin levels and geosmin producers in European recirculating aquaculture systems. **Journal of Applied Microbiology**, v. 132, n. 3, p. 2475–2487, 2022.

LYNN, P. A. **Electricity from Sunlight: An Introduction to Photovoltaics**. [s.l: s.n.].

MAGARREIRO, C.; FREITAS, S.; BRITO, M. C. Radiação e energia solar. **Gazeta de Física**, v. 39, n. 1/2, p. 57–59, 2016.

MARACLE, L. Agua. **Revista Colombiana de Antropologia**, v. 56, n. 2, p. 229–234, 2020.

MARTINS DOS SANTOS, D.; DA SILVA OLIVEIRA, E. **Investigação de redes neurais arficiais para predição da velocidade do vento**. 2021

MAULUD, Dastan; ABDULAZEEZ, Adnan M. A review on linear regression comprehensive in machine learning. **Journal of Applied Science and Technology Trends,** v. 1, n. 2, p. 140-147, 2020.

MEHTA, H.; KANANI, P.; LANDE, P. Google Maps. **International Journal of Computer Applications**, v. 178, n. 8, p. 41–46, 2019.

METRÓLOGOS METAS & ASOCIADOS. Medición de Turbidez en la Calidad del Agua. **La Guía Metas**, v. 10, n. 01, p. 1–6, 2010.

MUNDIAL, B. Banco Mundial, informe anual 2020. **Banco Mundial, informe anual 2020**, 2020.

NEWMAN, B. **ÁguaDancing Times**, out. 2010.

ORELLANA SALAS, J. A.; LALVAY PORTILLA, T. D. C. Uso e importancia de los recursos naturales y su incidencia en el desarrollo turístico. Caso Cantón Chilla, El Oro, Ecuador. **Revista interamericana de ambiente y turismo**, v. 14, n. 1, p. 65–79, 2018.

ORGANIZACIÓN MUNDIAL DE LA SALUD. **OMS | Agua**.

OTHMAN, F. et al. Efficient river water quality index prediction considering minimal number of inputs variables. **Engineering Applications of Computational Fluid Mechanics**, v. 14, n. 1, p. 751–763, 2020.

PADIAL, P.; POMPÊO, M.; MOSCHINI-CARLOS, V. Heterogeneidade espacial e temporal da qualidade da água no reservatório Rio das Pedras (Complexo Billings, São Paulo). **Ambiente e Agua - An Interdisciplinary Journal of Applied Science**, v. 4, n. 3, p. 35–53, 2009.

PASSOS, A. L. L.; DE FREITAS MUNIZ, D. H.; FILHO, E. C. O. Criteria for assessing water quality in Brazil: A Questioning about the parameters used. **Fronteiras**, v. 7, n. 2, p. 290–303, 2018.

PERDOMO, C.; BARBAZÁN, M. Nitrógeno. **Área De Suelos Y Aguas**, v. 1, 2007.

PEREIRA, É. DOS S.; FARIA, R. R.; DO NASCIMENTO LIMA, T. Recursos Naturais. **Revista Acta Ambiental Catarinense**, v. 18, n. 1, p. 239–252, 2021.

PEREZ, G. Manual de Hidrologia Aplicada. **UFOP - Universidade de Ouro Preto**, p. Cap.2, 17, 2015.

PNUD. Documento de apoyo: Medio Ambiente. **International Strategy for Disaster Reduction**, v. 1, n. 1, p. 38, 2018.

RADIACI, S. D. E. L. A. Características de la radiación solar. p. 1–21, 2012.

REDACCIÓN APD. ¿Qué es Machine Learning y cómo funciona'. **¿Qué es Machine Learning y cómo funciona?**, 2019.

REZENDE, K. F. O. et al. Histopatologia das brânquias de Tilápia do nilo *Oreochromis niloticus*, provenientes da represa Billings, área de proteção ambiental Bororé- Colônia. **Atas de Saúde ambiental**, v. 1, n. 1, p. 57–68, 2013.

RIBEIRO FORTES, M.; ARAÚJO DINIZ, A. C. Reflexões Sobre As Estações Do Ano: **Boletim Paulista de Geografia**, v. 109, n. 1, p. 40–64, 2023.

ROBERTIS, E. M.; HIB, J. Biologia celular e molecular. **Rev. Elet. Cient.**, v. 3, n. 4, p. 694–703, 2017.

ROCHA, A. A.; PEREIRA, D. N.; DE PADUA, H. B. Fishing yield and chemical contamination of the water of the Billings Reservoir, S. Paulo (Brazil). **Revista de Saude Publica**, v. 19, n. 5, p. 401–410, 1985.

ROCHA, E. S. et al. Relação entre padrões de uso e ocupação do solo e qualidade da água no rio Itanhém, entre Medeiros Neto e Teixeira de Freitas, Bahia. **Revista Brasileira de Geografia Física**, v. 16, n. 2, p. 688–702, 2023.

RUIZ ESPARZA GONZÁLEZ, H. et al. Capítulo 11-Ingeniería y gestión de sistemas SIMULACIÓN: CONCEPTOS Y EVOLUCIÓN. p. 12, 2010.

S., Y.; A., M. Efficient Water Quality Prediction for Indian Rivers Using Machine Learning. **Asian Journal of Applied Science and Technology**, v. 05, n. 01, p. 100–109, 2021.

SAGAN, V. et al. **Monitoring inland water quality using remote sensing: potential and limitations of spectral indices, bio-optical simulations, machine learning, and cloud computingEarth-Science Reviews**, 2020.

SÁNCHEZ, J. **Recursos naturales , medio ambiente y sostenibilidad : 70 años de pensamiento de la CEPAL**. [s.l: s.n.]. v. 4

SÁNCHEZ MARTÍNEZ, D. V. Calidad del agua. **Boletín Científico de la Escuela Superior Atotonilco de Tula**, v. 4, n. 7, 2017.

SCARDUA, F. P. Meio Ambiente. **Brasil em Números**, v. 29, p. 449–470, 2021.

SCARIOT, M. R. Modelagem e Simulação Sistêmica de Rios : Avaliação dos Impactos Ambientais no Rio Mogi-Guaçu/SP. **Universidade Estadual de Campinas**, p. 200, 2008.

SENDACZ, S. et al. Cargas de nutrientes (nitrogênio e fósforo) na bacia do Alto Tietê (Cabeceiras e Guarapiranga). **Hydrological Dynamics**, 2005.

SERRAGLIO, R. D. et al. Water Quality Parameters in Brazil. **Revista Brasileira de Geografia Fisica**, v. 14, n. 7, p. 3814–3830, 31 dez. 2021.

SERWAY, R. A. Y C. V. **Fundamentos de física**. [s.l: s.n.].

SHIN, Y. et al. Prediction of chlorophyll-a concentrations in the Nakdong river using machine learning methods. **Water (Switzerland)**, v. 12, n. 6, 2020.

SIERRA, J. Oxígeno disuelto. **Ciencias Con Lo Mejor De Vernier**, 2013.

SILVA, E. C. DA et al. **Precipitação Pluviométrica**. [s.l: s.n.].

SOLÓRZANO CHAMORRO, J. J.; VERA BASURTO, J. S.; BUÑAY CANTOS, J. P. Crecimiento económico y medio ambiente. **Reciamuc**, v. 6, n. 1, p. 203–212, 23 jan. 2022.

THIEMANN, M.; SCHEIBLER, E.; WIEGAND, K. W. Nitric Acid, Nitrous Acid, and Nitrogen Oxides. In: **Ullmann's Encyclopedia of Industrial Chemistry**. [s.l: s.n.].

TUCCI; BERTONI. Precipitação. **Universidade Federal da Bahia - Departamento de Hidraulica e Saneamento**, 1993.

VALDÉS, M.; RIVEROS, D.; ARACINBIA, R. Radiación Solar. **La Radiaciòn Solar**, 2012.

VALLS, J. L. **El potencial Redox (ORP)aviNews**, 2019.

VENEGAS, L.; MAZZEO, N. La velocidad del viento y la dispersión de contaminantes en la atmósfera. **Consejo Nacional de Investigaciones Científicas y Técnicas**, p. 1–11, 2012.

VIANA, J. F. DE S. et al. Modelagem hidrológica da Bacia Hidrográfica do Rio Pirapama-PE utilizando o modelo SWAT. **Journal of Environmental Analysis and Progress**, v. 3, n. 1, p. 155–172, 31 jan. 2018.

VIEIRA, R. Os principais parâmetros monitorados pelas sondas multiparâmetros são: pH, condutividade, temperatura, turbidez, clorofila ou cianobactérias e oxigênio dissolvido. **Agencia Nacional das Aguas–ANA-2015**, 2019.

VINET, L.; ZHEDANOV, A. A "missing" family of classical orthogonal polynomials. **Journal of Physics A: Mathematical and Theoretical**, v. 44, n. 8, p. 374, 2011.

WANG, J.; SONG, H.; ZHOU, X. A collaborative filtering recommendation algorithm based on biclustering. **2015 IEEE International Conference on Cyber Technology in Automation, Control and Intelligent Systems, IEEE-CYBER 2015**, v. 22, n. 4, p. 803–807, 2015.

ZAVALA GUILLEN, A. K. Documento de apoyo: Medio Ambiente. **International Strategy for Disaster Reduction**, v. 1, n. 1, p. 38, 2018.

ZITA, A. **¿Qué es el pH? - Toda Materia**.