UNIVERSIDADE
FEDERAL RURAL
DE PERNAMBUCO

Fábio Alves de Freitas

Development of Machine Learning Models for the Prediction of Dissolved Oxygen in Aquaculture 4.0

Recife

2021

Fábio Alves de Freitas

Development of Machine Learning Models for the Prediction of Dissolved Oxygen in Aquaculture 4.0

Monografia apresentada ao Curso de Bacharelado em Ciências da Computação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Ciências da Computação.

Universidade Federal Rural de Pernambuco – UFRPE

Departamento de Computação

Curso de Bacharelado em Ciências da Computação

Orientador: Prof. Dr. Obionor de Oliveira Nóbrega

Coorientador: Prof. Dr. Fernando Antonio Aires Lins

Recife

2021

Dados Internacionais de Catalogação na Publicação
Universidade Federal Rural de Pernambuco
Sistema Integrado de Bibliotecas
Gerada automaticamente, mediante os dados fornecidos pelo(a) autor(a)

F866d Freitas, Fábio Alves de
Development of Machine Learning Models for the Prediction of Dissolved Oxygen in Aquaculture 4.0 / Fábio Alves de Freitas. - 2021.
25 f. : il.

Orientador: Obionor de Oliveira Nobrega.
Coorientador: Fernando Antonio Aires Lins.
Inclui referências.

Trabalho de Conclusão de Curso (Graduação) - Universidade Federal Rural de Pernambuco, Bacharelado em Ciência da Computação, Recife, 2021.

1. IoT. 2. Aquacultura 4.0. 3. Dissolved Oxygen. 4. Machine Learning. 5. Prediction. I. Nobrega, Obionor de Oliveira, orient. II. Lins, Fernando Antonio Aires, coorient. III. Título

CDD 004

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO (UFRPE)
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

http://www.bcc.ufrpe.br

Ficha de Aprovação do Trabalho de Conclusão de Curso

Trabalho defendido por Fábio Alves de Freitas às 17 horas do dia 24 de fevereiro de 2021, no link http://meet.google.com/fah-pkyy-bob, como requisito para conclusão do curso de Bacharelado em Ciência da Computação da Universidade Federal Rural de Pernambuco, intitulado *Development of Machine Learning Models for the Prediction of Dissolved Oxygen in Aquaculture 4.0*, orientado por Obionor de Oliveira Nóbrega e aprovado pela seguinte banca examinadora:

Obionor de Oliveira Nóbrega
DC/UFRPE

Jeísa Pereira de Oliveira Domingues
DC/UFRPE

Agradecimentos

A Deus Pai, todo poderoso, pelo dom da vida, pela capacidade de planejar, sonhar e concretizar mais um sonho em minha caminhada.

Aos meus pais, Joana Dar'c e Joselito Alves, minha irmã, Eduarda Elvira, e toda a minha família, pelos ensinamentos, apoio e por acreditarem na minha capacidade de superação e realização de meus sonhos e objetivos.

Ao professor Dr. Obionor de Oliveira Nóbrega, por ter me convidado a este projeto na metade do curso, onde arcávamos com todos os custos do nosso bolso e hoje estamos colhendo os frutos do nosso trabalho.

Ao professor Dr. Fernando Antonio Aires Lins, pelos ensinamentos em diversas disciplinas e pela coorientação deste trabalho de conclusão de curso.

Ao professor Dr. Rafael Dueire Lins, pela recomendação a vaga do meu primeiro emprego, de onde obtive diversos aprendizados, que foram aplicados neste trabalho.

Ao coordenador e a secretaria do curso de bacharelado em ciência da computação, o professor Dr. Ruan Vasconcelos Bezerra Carvalho e Sandra Xavier respectivamente, por todo o apoio na resolução em tempo recorde das burocracias da universidade.

Aos meus amigos pela amizade, confiança e pela parceria em incontáveis projetos dentro e fora do curso.

Resumo

O mundo enfrenta o problema de alimentar uma população crescente, que chegará a mais de 9 bilhões de pessoas até 2050. Desta forma, existe a necessidade do desenvolvimento de atividades que promovam a produção de alimentos, nas dimensões da sustentabilidade (social, técnico-econômica, e ambiental). Neste contexto destacam-se os sistemas de IoT voltados à aquicultura 4.0, que possibilitam o cultivo de altas produções por unidade de volume, com baixo impacto ambiental. Entretanto, esses sistemas precisam ser extremamente controlados, necessitando de sensores para realização de leituras em tempo real das métricas da água, com destaque para o sensor de oxigênio dissolvido (OD), que desempenha um papel essencial na determinação da qualidade e quantidade de “habitat” disponível para os organismos presentes no sistema. Mesmo com essa importância, esse sensor é muitas vezes não utilizado, devido a seu alto custo associado. Como solução alternativa para este problema, foram propostos modelos de aprendizagem de máquina para a predição do OD, e que utilizam as leituras da temperatura e do pH como entradas. Foram realizados experimentos comparando diferentes técnicas de escalonamento de dados e o desempenho da predição em diferentes estações do ano e foram utilizadas métricas de regressão para avaliação dos modelos implementados. Os resultados mostraram que o modelo LSTM proposto pode realizar predições OD e ser aplicado em sistemas de IoT e aquicultura 4.0.

Palavras-chave: IoT, Aquacultura 4.0, Oxigênio Dissolvido, Aprendizado de Máquina, Predição.

Abstract

The world faces the problem of feeding a growing population, which will reach more than 9 billion people by 2050. Thus, there is a need to develop activities that promote food production, within the dimensions of sustainability (social, technical-economic, and environmental). In this context, IoT systems focused on aquaculture 4.0 stand out, which allows the cultivation of high productions per unit of volume, with low environmental impact. However, these systems need to be extremely controlled, requiring sensors to perform real-time readings of water metrics, with emphasis on the dissolved oxygen (DO) sensor, which plays an essential role in determining the quality and quantity of available habitat for the organisms present in the system. Even with this importance, this sensor is often not used, due to its high associated cost. As an alternative solution to this problem, machine learning models have been proposed to predict DO, using temperature and pH readings as inputs. Experiments were carried out comparing different data scaling techniques and the prediction performance in different seasons of the year and regression metrics were used to evaluate the implemented models. The results showed that the proposed LSTM model is capable of making OD predictions and being applied in IoT and aquaculture 4.0 systems.

Keywords: IoT, Aquacultura 4.0, Dissolved Oxygen, Machine Learning, Prediction.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Atualmente o mundo enfrenta o desafio de alimentar uma população que cresce exponencialmente. De acordo com a *Food and Agriculture Organization of United Nations* (FAO) (FOOD; NATIONS, 2018), será necessário alimentar mais de 9 bilhões de pessoas até 2050, em um cenário global onde ocorre a diminuição dos campos agrícolas (na Ásia e África, por exemplo) e uma diminuição (40% menor) impactante na pesca marinha nas regiões tropicais. Agregado a estes fatos, acrescenta-se a necessidade do desenvolvimento de atividades que promovam a produção de alimentos, dentro das dimensões da sustentabilidade (social, técnico-econômica, e ambiental).

Neste contexto pode-se destacar a aquicultura na prática da produção de alimentos marinhos. A aquicultura é uma sub-área do agronegócio e a ciência que estuda e desenvolve técnicas de cultivo e reprodução de organismos aquáticos. É uma atividade exercida já há bastante tempo, e que tem suas técnicas mais tradicionais caracterizadas por serem manuais e fazerem o uso direto de recursos hídricos, o que pode gerar impactos negativos para a qualidade ambiental (FONSECA et al., 2021).

Neste cenário surge a aquicultura 4.0, ou aquicultura inteligente (DUPONT; COUSIN; DUPONT, 2018), caracterizada pela utilização de tecnologia de ponta nas atividades da área, possibilitando: a aplicação de técnicas mais eficientes para os cultivos hiperintensivos, com altas produções por unidade de volume; menor utilização de recursos hídricos, reduzindo o impacto ambiente, como nos sistemas de ”biofloco”, que fazem uso de tanques; e tornando a tecnologia computacional uma importante ferramenta no auxílio da produção, por meio da otimização de processos, redução de custos e aumento da produtividade (CLERCQ; VATS; BIEL, 2018).

A Internet das coisas (*Internet of Things* - IoT) é um dos conceitos mais utilizados para o desenvolvimento de sistemas voltados para aquicultura inteligente, possuindo uma ampla aplicabilidade, por meio do controle e monitoramento remoto (YAURI; RIOS; LEZAMA, 2017) e tomadas de decisão automáticas e inteligentes (DZULQORNAIN; RASYID; SUKARIDHOTO, 2018). Esses sistemas precisam ser extremamente controlados, necessitando de leituras em tempo real das métricas do ambiente, obtidas a partir da utilização de sensores, dispositivos eletrônicos responsáveis pela aquisição de dados, onde tem-se a temperatura, o potencial hidrogeniônico (pH) e o oxigênio dissolvido (OD) como as métricas mais utilizadas. Essas leituras auxiliam nos demais processos do sistema, como a tomada de decisão e exibição de relatórios em tempo real do estado do sistema.

Por outro lado, um dos desafios para o desenvolvimento destes tipos de sis-

temas são os custos associados a aquisição, instalação e manutenção dos sensores (DABROWSKI; RAHMAN; GEORGE, 2018), com destaque para o sensor de OD, que desempenha o papel de essencial de determinar a qualidade e quantidade de habitat disponível para os organismos (ZHANG et al., 2019), e mesmo com curtos intervalos de tempo com os níveis de OD fora dos intervalo ideal pode ser fatal aos animais presentes no sistema. Mesmo com todo esta importância, o sensor de OD é muita vezes omitido desses sistemas, por ter um custo associado mais elevado que os demais sensores, reduzindo assim o potencial dos sistemas.

A partir disto, torna-se necessário a busca de métodos alternativos para a obtenção do OD, a fim de mitigar a falta do sensor de OD. Em trabalhos como (AHMED, 2017) e (EMAMGHOLIZADEH et al., 2014) são propostos modelos de aprendizagem de máquina, para prever os valores do OD de rios e estimar a qualidade da água. Estes modelos utilizam leituras de métricas da água como: temperatura, ph, nitrito, nitrato, condutividade, turbidez, etc. O conjunto de dados utilizados neles foram registrados mensalmente, o que torna inviável a utilização dos modelos propostos para a predição do OD na aquicultura, que necessita de múltiplas leituras diárias do ambiente.

Pouco se investigou acerca da predição de OD voltada para a aquicultura. Trabalhos como (DZULQORNAIN; RASYID; SUKARIDHOTO, 2018) e (ZHANG et al., 2019) propuseram variações do modelos de predição de qualidade da água, a fim de aplicá-los à aquicultura, utilizando bases de dados mais adequadas a este contexto, com múltiplas leituras diárias das métricas da água. Entretanto, o processo de treinamento desses modelos utilizam as leituras do OD como entrada, justificando a utilização dos mesmos como prova de conceito, porém tornando os modelos dependentes de leituras do OD, que consequentemente gera a necessidade do sensor de OD. Desta forma, o objetivo deste trabalho é o de desenvolver uma solução de predição de OD utilizando algoritmos de aprendizagem de máquina baseados apenas no pH e temperatura para sistemas de IoT para aquicultura 4.0.

Este trabalho está organizado da seguinte forma. Na Seção II são discutidos os trabalhos relacionados e o estado da arte da área; na Seção III é explicada a metodologia utilizada; na Seção IV são discutidos os resultados obtidos; e por fim, na Seção V são apresentadas as conclusões da pesquisa.

2 Trabalhos Relacionados

Em (ROUNDS, 2002) é proposta a predição do OD para avaliar a qualidade da água rio Tualatin (Oregon - EUA). Para isto foi construído um modelo de rede neural artificial (*Artificial Neural Network - ANN*), com mecanismo de *feedfoward*. Foram utilizadas a temperatura do ar, a radiação solar, a frequência de chuvas e o fluxo de corrente da água com entradas do modelo. Na aquicultura, as métricas mais comuns para os processos de análise de dados são oriundas da água, o que torna o esse modelo não adequado a esse contexto.

Outros trabalhos relacionados à predição de OD, como (EMAMGHOLIZADEH et al., 2014), (AHMED, 2017), (SCHTZ et al., 2015) e (WANG et al., 2017), foram observados as seguintes características em comum: todos utilizam modelos de ANN de predição de OD voldados a avaliação da qualidade da água, o que sugere que esses modelos tenham se tornado um padrão na área; a realização da avaliação de desempenho por meio das métricas de regressão *mean absolut error* (MAE), *root-mean-square error* (RMSE) e *r-squared* (R^2); e a utilização de bases de dados contendo leituras de métricas da água, que são coletadas mensalmente. Devido a frequência das coleta de dados, as bases de dados utilizadas não são adequadas para o treinamento de modelos voltados à aquicultura, que necessitam de múltiplas leituras diárias, para uma maior precisão dos processos de análises de dados e desempenho dos modelos.

Em (ANTANASIJEVIĆ et al., 2014) ocorreu um aprofundamento dos requisitos técnicos da predição de OD voltada para a avaliação da qualidade da água. Nele foi proposto avaliação dos diferentes técnicas de escalonamento de dados e a técnicas seleção de entradas para os modelos de ANN implementados. Para o escalonamento, foram comparadas as técnicas: normalização (min-max), mediana, z-score, sigmoide e tangente hiperbólica. Para a seleção das entradas, foram comparadas as técnicas: variação da inflação, correlação de pearson e algoritmo genético. Os melhores resultados foram obtidos a partir da utilização da normalização, como técnica de escalonamento ,e a correlação de pearson, para seleção das entradas. Entretanto, a correlação de pearson, mesmo indicando bons resultados, é uma técnica direcionada a avaliação do relacionamento linear entre variáveis, que não é o caso das métricas de qualidade da água, que não apresentam esse comportamento. A fim de identificar a correlação entra variáveis não lineares, (KRASKOV; STÖGBAUER; GRASSBERGER, 2004) propôs a técnica da informação mútua (*MI - mutual information*) para a quantificação da correlação entre variáveis, identificando suas relações lineares e não lineares, e (ROSS, 2014) aprimorou essa técnica, permitindo que o MI seja calculado tanto para variáveis discretas e contínuas.

Em (DABROWSKI; RAHMAN; GEORGE, 2018) é proposta a predição do OD voltada para a aquicultura, onde é utilizada uma base de dados com múltiplas leituras diárias das métricas pH, temperatura e OD. Foi implementado o algoritmo *linear dynamical system* (LDS), que utiliza estas três métricas durante seu treinamento, para identificar as dependências entre as variáveis, e em seguida é capaz de estimar uma variável faltante a partir das demais presentes. Também foram implementados modelos baseados em *long-short term memory* (LSTM) e ANN, e utilizam o pH, a temperatura e a própria predição do LDS como entrada. Os melhores resultados foram obtidos com o LSTM. Entretanto o trabalho discute brevemente esses resultados e utiliza apenas o *normalized root-mean-square error* (NRMSE) como métrica de avaliação dos algoritmos, tornando difícil a comparação de seus modelos propostos com os de outros trabalhos. Adicionalmente, além dos modelos dependerem das leituras do OD, devido ao processo de treinamento do LDS, a base de dados utilizada possui registros faltantes, o que pode ter gerado resultados equivocados.

Em (ZHANG et al., 2019) também foi proposta a predição do OD voltada a aquicultura. A base de dados utilizada possui leituras da temperatura, pH, turbidez, condutividade, clorofila e OD. As leituras ocorreram a cada meia hora, num período de um ano, resultando em 17520 registros salvos. A proposta trazida pelos autores é a de utilizar modelos multicamadas do ANN, implementando desta forma o *multi-layer artificial neural network* MANN. É realizado o comparativo do MANN com o MANN-dropout, MANN-pearson e ANN e a partir de modelos não baseados em redes neurais, como o *support vector regression* (SVR) e o *linear regression model* (LRM). Foi utilizado o conceito de mutual information (KRASKOV; STÖGBAUER; GRASSBERGER, 2004; ROSS, 2014) para a determinação das métricas mais apropriadas para servirem de entrada aos modelos, sendo selecionadas a temperatura, o pH, a clorofila e o OD. Devido a utilização do OD como entrada, os modelos são dependentes das leituras do sensores de OD, não sendo apropriados para a substituição do mesmo.

A análise destes trabalhos mostrou que existem estudos envolvendo a predição de OD para avaliação da qualidade da água de rios, e que não podem ser diretamente aplicados à aquicultura, devido a limitações como a frequência de leitura de dados das bases de dados e as metodologias empregadas para a seleção das entradas. A predição de OD diretamente ligada a aquicultura foi investigada nos trabalhos (ZHANG et al., 2019) e (DABROWSKI; RAHMAN; GEORGE, 2018), entretanto os valores dessa métrica são utilizados como entrada dos modelos, tornando-os dependentes do sensor de OD e não adequados para substituí-lo.

3 Materiais e Métodos

As implementações do presente trabalho podem ser encontradas através do link presente na referência (IMPLEMENTAçõES..., 2021).

3.1 Base de Dados

Foi utilizada a base de dados disponibilizada por (ZHANG, 2019) (ZHANG et al., 2019). Este conjunto de dados trata-se de uma série temporal de leituras das métricas temperatura, pH, condutividade, turbidez, clorofila e oxigênio dissolvido, coletadas a cada meia hora por um período de um ano (1/11/2013 à 31/10/2014), resultando num total de 17520 registros de cada métrica. As coletas de dados ocorreram no sistema estuário de *Baffle Creek* (Queensland - Austrália), utilizando o módulo sensor YSI Model 6600, posicionado a 20cm de profundidade da água.

De acordo com (ZHANG et al., 2019), o clima nesta região é caracterizado por uma estação seca, de maio a outubro, e uma chuvosa, novembro a abril. A partir desse fato, a base de dados foi subdividida em seis diferentes formatos (Tabela 1), variando as duas estações do ano e a janela de tempo entre os registros para 30, 90 e 120 minutos. Cada formato foi um cenário avaliado pelo treinamento e testes dos algoritmos, com o intuito de determinar a estação e intervalo de tempo que trouxesse maior ganho à predição.

Tabela 1 – Base de dados modificadas utilizadas nos experimentos.

Nome da base de dados	Estação do ano	Intervalo de tempo	Número de registros
dry_30	Seca	30 minutos	8842
dry_90	Seca	90 minutos	2948
dry_120	Seca	120 minutos	2212
rain_30	Chuvosa	30 minutos	8680
rain_90	Chuvosa	90 minutos	2896
rain_120	Chuvosa	120 minutos	2172

3.2 Seleção das métricas de entrada

De acordo com (ROUNDS, 2002) (ROUNDS, 2002), modelos baseados em ANN são particularmente bons em identificar padrões em conjuntos não lineares de dados. Esta funcionalidade é aprimorada caso o conjunto de dados de entrada possua um maior grau de correlação entre si, o que aumenta o grau de generalização dos mo-

delos (VERGARA; ESTÉVEZ, 2014), ou seja, a capacidade dos mesmos de preverem com maior precisão dados que não foram utilizados no processo de treinamento.

Como mencionado no trabalhos relacionados, (KRASKOV; STÖGBAUER; GRASSBERGER, 2004) propõe a técnica de MI para a avaliar o grau de correlação entre variáveis e (ROSS, 2014) aprimora o MI para que seja utilizado em variáveis discretas e contínuas. Visto que os dados da base dados utilizada são contínuos e não lineares, a aplicação dessa técnica é de grande valia. De acordo com estes trabalho, quanto maior a pontuação do MI, maior a correlação entre as variáveis. Este critério foi utilizado para determinar as métricas a serem selecionadas como entrada dos modelos. Foram comparando os registros de temperatura, pH, turbidez, condutividade e clorofila com os registros do oxigênio dissolvido da base de dados em seu formato original. As pontuações do MI obtidas podem ser observadas na Fig 1, e foram calculadas utilizando a linguagem *Python* (ROSSUM; DRAKE, 2009), por meio da biblioteca *scikit-learn* (PEDREGOSA et al., 2011), que possui a implementação proposta por (ROSS, 2014). A partir de seus resultados, foi decidido utilizar a temperatura e pH como métricas de entradas, por retornarem as maiores pontuações do MI.

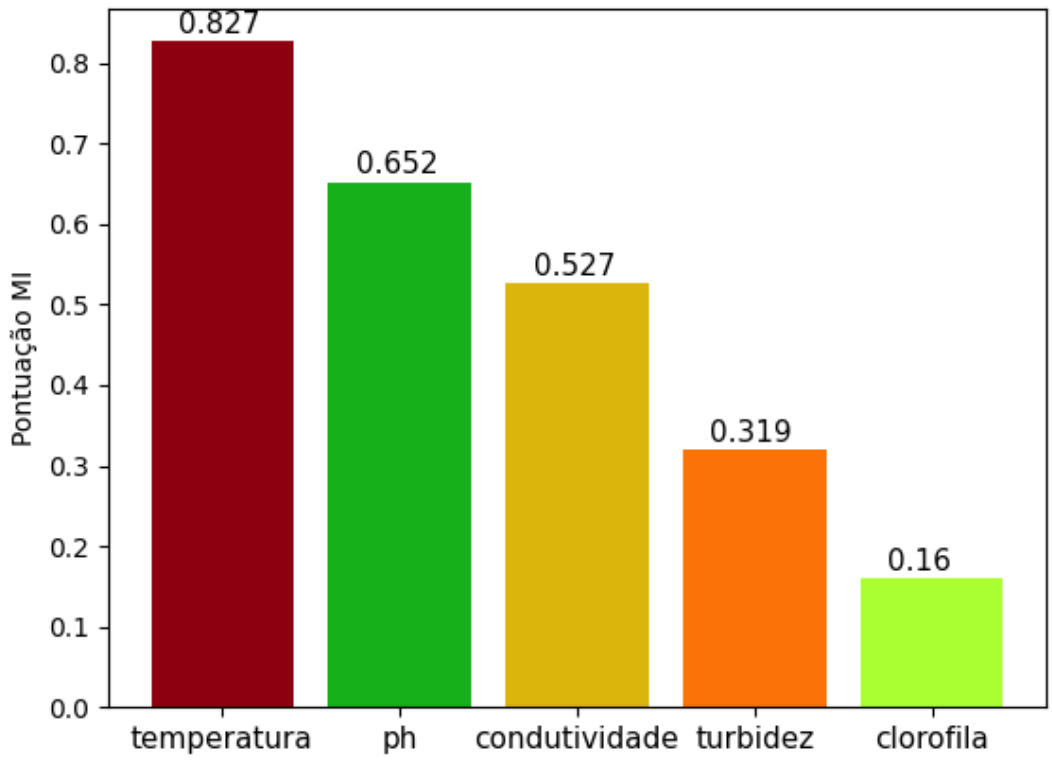

Figura 1 – Pontuações do MI entre o OD e as demais variáveis da base de dados.

3.3 Modelos implementados

Foram implementados modelos dos algoritmos LSTM, MANN, ANN e SVR, inspirados nos modelos propostos por (DABROWSKI; RAHMAN; GEORGE, 2018; ZHANG et al., 2019). Foram comparados os resultados desses trabalho com as versões dos modelos do presente trabalho. Todos os modelos utilizam apenas a temperatura e pH como entrada e como saída geram as predições do OD.

Todos os modelos foram implementados na linguagem *Python* (ROSSUM; DRAKE, 2009). Para o LSTM, MANN e ANN foi utilizada a biblioteca *keras* (CHOLLET et al., 2015), que disponibiliza a implementação de diversos algoritmos baseados em redes neurais. O LSTM foi escolhido para confirmar os bons resultados obtidos por (DABROWSKI; RAHMAN; GEORGE, 2018). O MANN e ANN foram escolhidos por serem os modelos mais mais bem sucedidos em (ZHANG et al., 2019) e pela semelhança a diversos outros modelos baseados em ANN, que são amplamente utilizados na área da predição de OD. Os parâmetros e arquitetura para modelos baseados em ANN podem ser observados na Tabela 2. Para o SVR foi utilizada a biblioteca *scikit-learn* (PEDREGOSA et al., 2011), com os parâmetros no formato padrão definidos pela biblioteca. Além disso, este modelo foi selecionado para observar o comportamento das predições com um modelo não baseado em redes neurais.

Tabela 2 – Parâmetros utilizados nos modelos LSTM, MANN e ANN.

Parâmetros	MANN	ANN	LSTM
N.º de neurônios na camada de entrada	3	9	128
N.º de camadas escondidas	2	0	0
N.º de neurônios nas camadas escondidas	3	0	0
N.º de neurônios na camada de saída	1	1	1
Algoritmo de otimização	SGD	Adam	Adam
Função de ativação	Relu	Tanh	Relu

3.4 Escalonamento dos dados

O escalonamento dos dados diz respeito a técnicas que buscam colocar os dados numa mesma escala numérica. É uma técnica de pré-processamento amplamente utilizada para a aprendizagem de máquina, e auxilia o treinamento dos modelos impedindo que os modelos sejam influenciados por disparidades nas escalas dos dados de entrada. Na base de dados utilizada, a temperatura possui um mínimo de 15,73ºC e um máximo de 32,73ºC, e o pH possui um mínimo de 6,59 e um máximo de 8,47. A fim de evitar a perda de desempenho dos modelos, foram utilizadas as técnicas de padro-

nização e normalização, para criar um outro critério de comparação entre os modelos e determinar a técnica mais adequada para o cenário estudado no presente trabalho. A padronização (equação 3.1) utiliza a média (μ) e o desvio padrão (σ) do conjuntos de dados em seu cálculo. Já a normalização (equação 3.2), conhecida como a técnica min-max em (ANTANASIJEVIĆ et al., 2014), faz um mapeamento para o intervalo real entre 0 e 1, por meio das operações envolvendo os valores máximo e mínimo dos conjunto de dados.

$$z = \frac{x - \mu(x)}{\sigma(x)} \tag{3.1}$$

$$z = \frac{x - min(x)}{max(x) - min(x)} \tag{3.2}$$

3.5 Treinamento e Métricas de Avaliação

Para o treinamento de modelos de aprendizagem de máquina pode-se utilizar as técnicas de *holdout* e a validação cruzada (*cross validation*). O *holdout* consiste na separação de uma porcentagem da base de dados para treinamento e outra para teste e tem a vantagem de ser mais rápida. Entretanto, caso os dados utilizados nos testes sejam muito semelhantes aos utilizados nos treinos, o que indicará uma boa generalização, e os dados utilizados em produção serem muito distintos dos que já foram vistos pelo modelo, então a análise de desempenho modelo pode estar exibindo resultados equivocados, visto que o processo de treinamento não teve abrangência suficiente.

Já na validação cruzada, a base de dados é divida num número *x* de partições (*folds*). Em seguida são realizadas *x* rodadas de treinamento e teste, de forma que cada partição seja utilizada tanto para treino quanto para teste. Ao final de cada rodada são calculadas as métricas de avaliação de desempenho do modelo e o seu desempenho final é avaliado a partir da média das métricas obtidas em cada rodada. Desta forma, essa técnica é capaz de trazer uma maior abrangência ao modelo. Sua desvantagem está no custo computacional, visto que o número de execuções do treinamento é o mesmo do número de partições, enquanto que o *holdout* realiza apenas um treinamento.

A fim de aumentar a abrangência dos modelos implementados foi utilizada a técnica de validação cruzada no treinamento. Baseado em (DABROWSKI; RAHMAN; GEORGE, 2018), que também utilizou esta técnica, foram utilizadas 20 partições para o treinamento dos modelos.

Com relação a avaliação de desempenho dos modelos, foram utilizadas as métricas de regressão *root-mean-square error* (RMSE), *mean absolut error* (MAE) e *r-squared* (R^2), que de acordo com os trabalhos relacionados, são as métricas mais comuns para este propósito na área da predição de OD. O RMSE e MAE tem a função de indicar o erro médio dos modelos, ou seja, quanto mais próximos de zero, mais preciso o modelo é. O intervalo destas métricas varia de 0 a e$+\infty$. Já o R^2 indica a capacidade dos modelos de seguirem a tendência dos dados preditos, ou seja, determina o quão próxima é a curva associada as predições do modelo da curva real dos dados de teste. O intervalo de valores desta métrica é compreendido entre 1 e $-\infty$, onde 1 indica uma maior capacidade do modelo de seguir a tendência dos dados.

4 Resultados e Discussão

Ao total foram realizados 48 experimentos, combinando as formatações das bases de dados (Tabela 1), os quatro modelos de aprendizagem de máquina e as duas técnicas de escalonamento de dados. Para cada experimento foram aplicadas as metodologias de treinamento e avaliação de desempenho descritas na seção 3.5. Os resultados foram separados em seis figuras, a fim de simplificar sua análise, com as Figuras 2, 3 e 4 apresentando os resultados relacionados a normalização e com as Figuras 5, 6 e 7 apresentando os resultados relacionados a padronização. No canto esquerdo destas figuras pode ser observada a respectiva métrica de avaliação de desempenho utilizada. Além disso, em cada uma delas há seis grupos de gráficos em barras, cada um representando uma das bases de dados, onde em cada grupo foram utilizadas quatro cores para representar os quatro algoritmos implementados.

Analisando as técnicas de escalonamento no MANN, foi observado que a padronização é a técnica mais adequada para este modelo. As pontuações do RMSE e MAE desse modelo utilizando a padronização (Figuras 5 e 6), chegam aos seus maiores valores com a base de dados *rain_120*, comportamento observado também nos demais modelos. Entretanto, as seu RMSE e MAE com a normalização têm seus maiores valores nas bases de dados *dry_30* e *dry_120*, diferente do que ocorre nos demais modelos, que apresentam RMSE e MAE minimizados nessas bases de dados. Outro fator que evidencia isto são as pontuações do R^2, que chegam a -12,67 com a normalização (Figura 4) e a -4,17 a padronização (Figura 7), indicando que o MANN segue melhor a tendência dos dados por meio da padronização.

Comparando as técnicas de escalonamento para os modelos LSTM, ANN e SVR, tem-se que: o SVR não apresentou diferenças entre as pontuações de nenhuma métrica; o LSTM apresentou uma média de 0,01 de variação no R^2 e 0,1 de variação no RMSE e MAE nas bases de dados *dry_90* e *rain_30*; e o ANN apresentou uma média de 0,13 de variação no R^2 e variação 0,1 no RMSE e MAE nas bases de dados *dry_90*, *rain_30* e *rain_120*. Desta forma, como não foram observadas diferenças significativas entre os escalonamentos para estes modelos, tanto a normalização quando padronização podem ser utilizadas por eles sem perda de desempenho.

Realizando um comparativo das bases de dados, de acordo com a estação do ano, tem-se que os experimentos entre as três bases de dados de cada estação obtiveram resultados relativamente próximos entre si para os modelos LSTM, ANN e SVR. Para esses modelos na estação seca as diferenças entre os valores mínimo e máximo do RMSE e MAE são de 0,02 e do R^2 diferença de 0,3, enquanto que na

estação chuvosa as diferenças entre os valores mínimo e máximo do RMSE e MAE são de 0,06 e do R^2 diferença de 0,42. Desta forma, o intervalo de leitura indicado para sistemas de IoT e aquicultura 4.0 depende apenas da frequências de leitura requisitas para a temperatura e o pH, uma vez que não há grande diferença no desempenho dos modelos a partir dos intervalos de tempo analisados (30, 90 e 120). Esta afirmação não se aplica ao MANN, que não apresenta uma padronização de resultados por estação.

Dado que as pontuações do RMSE e MAE do LSTM, ANN e SVR são menores na estação seca, poderia-se afirmar que estes três modelos possuem um melhor desempenho nessa estação, devido ao erro médio inferior. Entretanto, a pontuação do R^2 desses modelos varia entre -1,2 e -1,52 na estação seca, enquanto que na estação chuvosa o R^2 varia entre -0.33 e 0.09, indicando uma maior capacidade de seguir a tendência dos dados reais na estação chuvosa. Desta forma, mesmo com erro médio menor na estação seca, é na estação chuvosa que esses modelos apresentam um desempenho superior.

Comparando os quatro modelos implementados entre si, tem-se que o que mais adequado à predição do OD para aquicultura 4.0 foi o LSTM, com as pontuações do MAE e RMSE inferiores aos dos demais modelos para a estação chuvosa, indicando um erro médio menor, e empatando com o ANN e SVR na estação seca. Esta afirmação é reforçada comparando o LSTM, que obteve RMSE médio de 0,14 e R^2 médio de -0,6, com o MANN proposto por (ZHANG et al., 2019), que realizou o treinamento de seu modelo com a mesma base de dados utilizada no presente trabalho, e obteve RMSE médio de 0,12 e R^2 médio de 0,89. Fica claro que ambos os modelos que obtiveram resultados relativamente próximos, indicando que o LSTM também é capaz de realizar predições do OD. Os modelos ANN e SVR empataram, com variação máxima de apenas 0,2 para o RMSE e MAE e 0,01 para o R^2. Já o MANN proposto no presente trabalho apresentou um desempenho inferior aos demais modelos implementados.

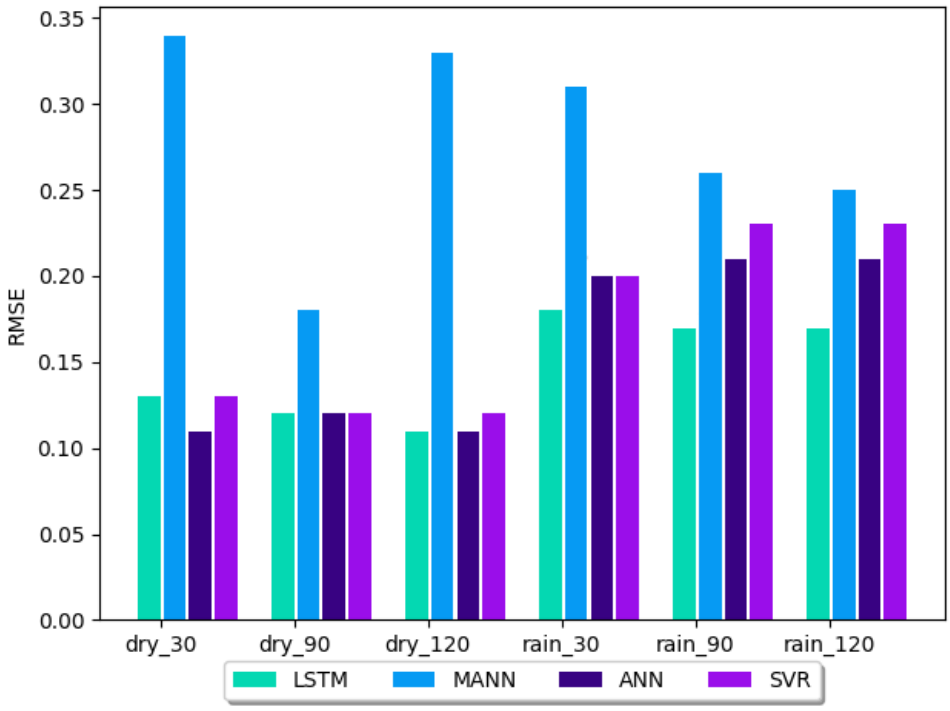

Figura 2 – Pontuação do RMSE dos treinamentos utilizando normalização.

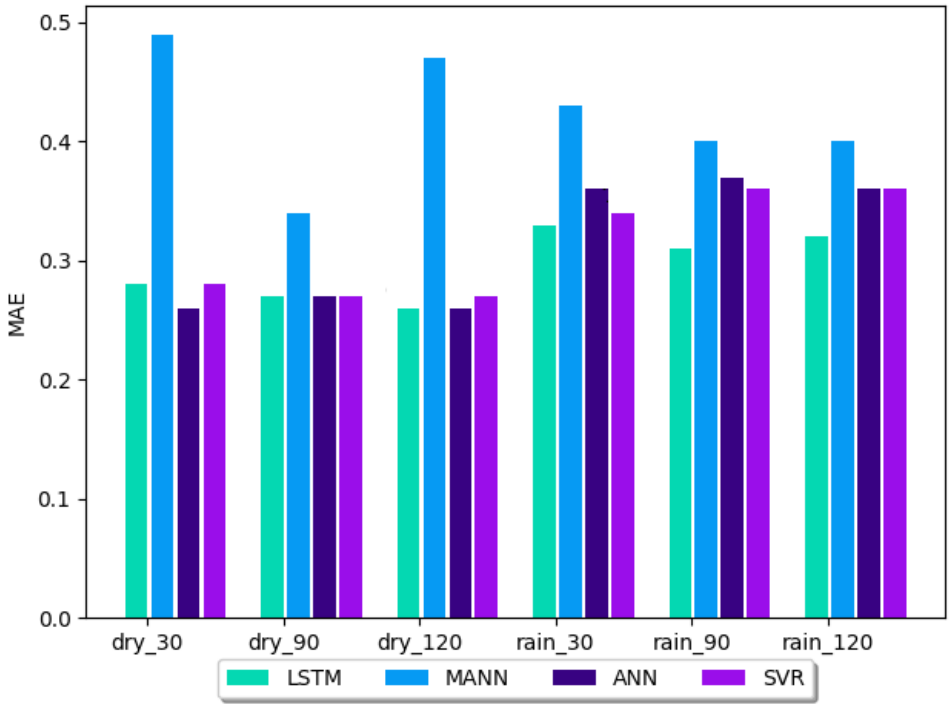

Figura 3 – Pontuação do MAE dos treinamentos utilizando normalização.

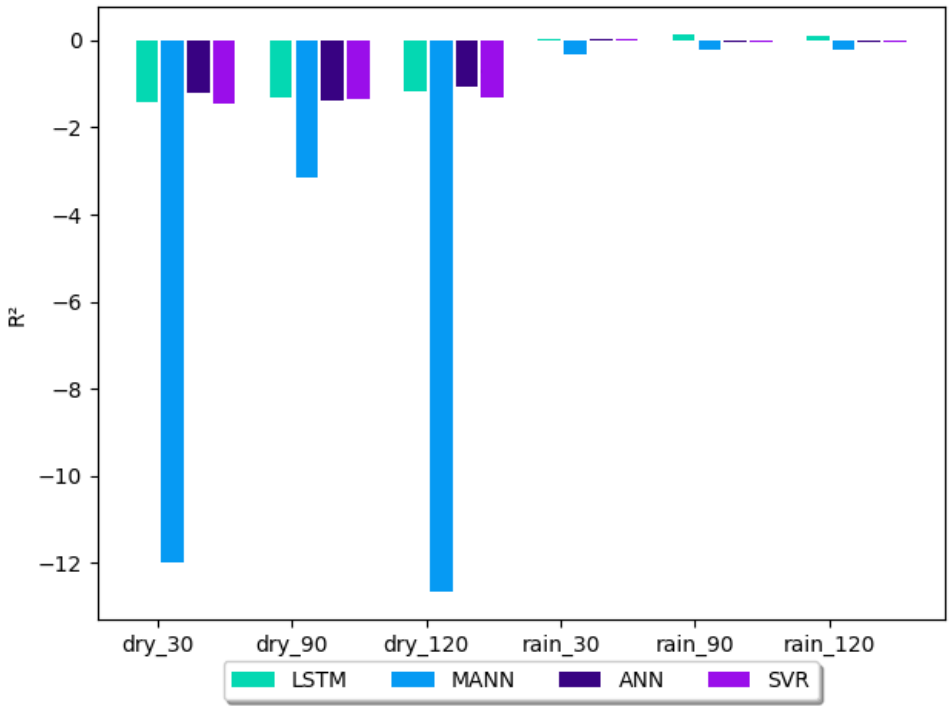

Figura 4 – Pontuação do R² dos treinamentos utilizando normalização.

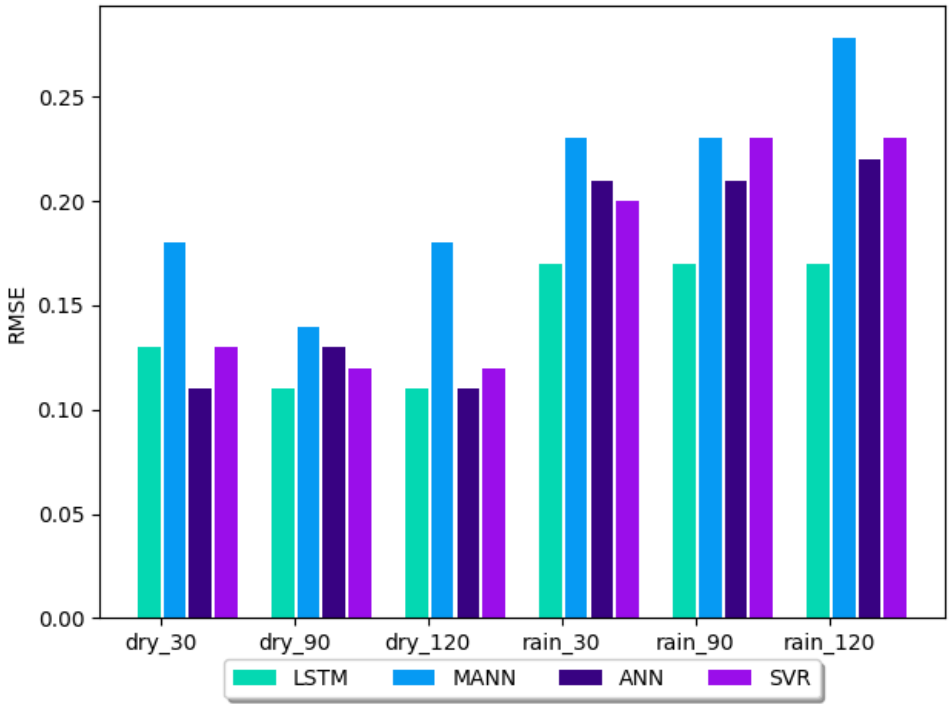

Figura 5 – Pontuação do RMSE dos treinamentos utilizando padronização.

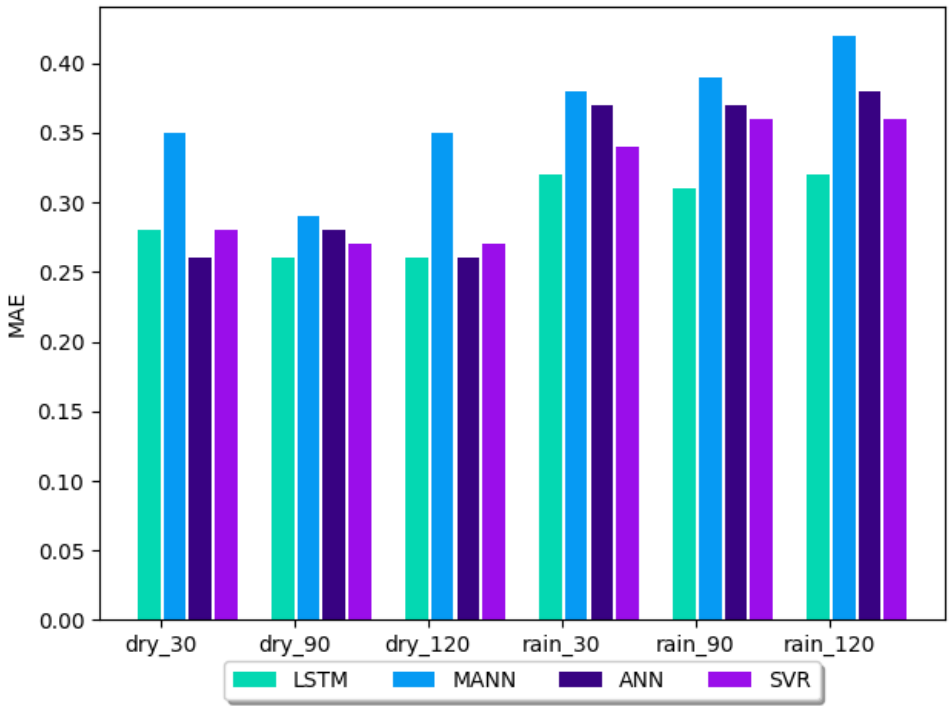

Figura 6 – Pontuação do MAE dos treinamentos utilizando padronização.

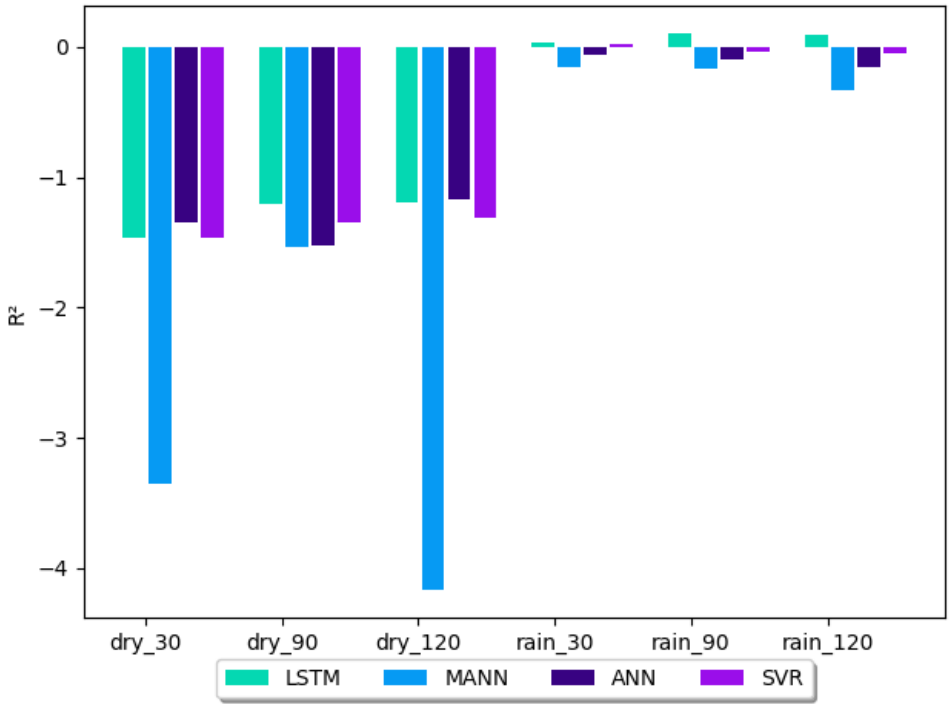

Figura 7 – Pontuação do R² dos treinamentos utilizando padronização.

5 Conclusões e Trabalhos Futuros

Neste trabalho fora desenvolvidos modelos de aprendizagem de máquina voltados a predição de oxigênio dissolvido para sistemas de aquicultura 4.0 baseados em IoT. Foram analisados trabalhos relacionados ao tema, a fim de compreender o estado da arte área, e aprimorar modelos que já apresentassem bons resultados. Nesses trabalhos, foi observado que os modelos de mais sucesso na predição do OD são baseados em redes neurais artificiais (ANN), porém não podem ser aplicados a sistemas de aquicultura 4.0, por utilizarem leituras passadas do OD como entrada, tornando-os dependentes do sensor de OD.

Desta forma, os modelos propostos foram baseados nos algoritmos LSTM, ANN, MANN e SVR, e foram projetados com a capacidade de serem independentes do sensor de OD, utilizando outras métricas da água como entrada. Foi utilizada uma base de dados contendo leituras da temperatura, pH, condutividade, turbidez, clorofila e OD para o processo de treinamento. A coleta do conjunto de dados ocorreu num estuário da região de Baffle Creek (Queensland - Austrália), em intervalos de 30 minutos por um período de um ano. A seleção das métricas de entrada por meio da técnica de informação mútua (Mutual Information - MI), onde foram calculadas as pontuações de correlação entre o OD e as demais métricas da água presentes nessa base. A métricas selecionadas como entrada dos modelos foram a temperatura e o pH, por obterem as maiores pontuações de MI.

Foram realizados 48 experimentos, combinando: 6 diferentes formatações da base de dados, envolvendo as estações seca e chuvosa e diferentes intervalo de tempo entre os registros; os quatro modelos de aprendizagem de máquina; e duas técnicas de escalonamento de dados, a normalização e padronização. A avaliação de desempenho dos modelos foi realizada por meio das métricas de regressão RMSE, MAE e R^2. As análises dos resultados evidenciaram que os modelos LSTM, ANN e SVR são capazes de realizar as predições de OD em sistemas de aquicultura 4.0, utilizando apenas a temperatura e pH como entradas, sendo capazes de substituir o sensor de OD. Nesse cenário destaca-se o LSTM, que apresentou o melhor desempenho dentre os quatro modelos avaliados.

Como trabalhos futuros pode-se apontar: o aprimoramento dos modelos propostos com relação a métrica R^2, que obteve valores negativos em alguns experimentos; e desenvolver um sistema IoT voltado a aquicultura 4.0 integrado aos modelos propostos.

Referências

AHMED, A. M. Prediction of dissolved oxygen in surma river by biochemical oxygen demand and chemical oxygen demand using the artificial neural networks (anns). *Journal of King Saud University-Engineering Sciences*, Elsevier, v. 29, n. 2, p. 151–158, 2017. Citado 2 vezes nas páginas 9 e 10.

ANTANASIJEVIĆ, D. et al. Modelling of dissolved oxygen in the danube river using artificial neural networks and monte carlo simulation uncertainty analysis. *Journal of Hydrology*, Elsevier, v. 519, p. 1895–1907, 2014. Citado 2 vezes nas páginas 10 e 15.

CHOLLET, F. et al. *Keras*. 2015. <https://keras.io>. Citado na página 14.

CLERCQ, M. D.; VATS, A.; BIEL, A. Agriculture 4.0: The future of farming technology. *Proceedings of the World Government Summit, Dubai, UAE*, p. 11–13, 2018. Citado na página 8.

DABROWSKI, J. J.; RAHMAN, A.; GEORGE, A. Prediction of dissolved oxygen from ph and water temperature in aquaculture prawn ponds. In: *Proceedings of the Australasian joint conference on artificial intelligence-workshops*. [S.l.: s.n.], 2018. p. 2–6. Citado 4 vezes nas páginas 9, 11, 14 e 15.

DUPONT, C.; COUSIN, P.; DUPONT, S. Iot for aquaculture 4.0 smart and easy-to-deploy real-time water monitoring with iot. In: IEEE. *2018 Global Internet of Things Summit (GIoTS)*. [S.l.], 2018. p. 1–5. Citado na página 8.

DZULQORNAIN, M. I.; RASYID, M. U. H. A.; SUKARIDHOTO, S. Design and development of smart aquaculture system based on ifttt model and cloud integration. In: EDP SCIENCES. *MATEC Web of Conferences*. [S.l.], 2018. v. 164, p. 01030. Citado 2 vezes nas páginas 8 e 9.

EMAMGHOLIZADEH, S. et al. Prediction of water quality parameters of karoon river (iran) by artificial intelligence-based models. *International Journal of Environmental Science and Technology*, Springer, v. 11, n. 3, p. 645–656, 2014. Citado 2 vezes nas páginas 9 e 10.

FONSECA, R. A. et al. Aquicultura: Impactos ambientais negativos e a mitigação com práticas agroecológicas. *in: Tópicos em recuperação de áreas degradadas [recurso eletrônico …*, 2021. Citado na página 8.

FOOD; NATIONS, A. O. of the U. The state of world fisheries and aquaculture 2018–meeting the sustainable development goals. *FAO*, 2018. Citado na página 8.

IMPLEMENTAçõES do presente trabalho. 2021. <https://github.com/fabioafreitas/DissolvedOxygenPrediction>. Citado na página 12.

KRASKOV, A.; STÖGBAUER, H.; GRASSBERGER, P. Estimating mutual information. *Physical review E*, APS, v. 69, n. 6, p. 066138, 2004. Citado 3 vezes nas páginas 10, 11 e 13.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado 2 vezes nas páginas 13 e 14.

ROSS, B. C. Mutual information between discrete and continuous data sets. *PloS one*, Public Library of Science, v. 9, n. 2, p. e87357, 2014. Citado 3 vezes nas páginas 10, 11 e 13.

ROSSUM, G. V.; DRAKE, F. L. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN 1441412697. Citado 2 vezes nas páginas 13 e 14.

ROUNDS, S. A. Development of a neural network model for dissolved oxygen in the tualatin river, oregon. In: *Proceeding of the second Federal Interagency Hydrologic Modeling Conference, Las Vegas, Nevada*. [S.l.: s.n.], 2002. Citado 2 vezes nas páginas 10 e 12.

SCHTZ, F. C. de A. et al. Simulation of the concentration of dissolved oxygen in river waters using artificial neural networks. In: IEEE. *2015 11th International Conference on Natural Computation (ICNC)*. [S.l.], 2015. p. 1252–1257. Citado na página 10.

VERGARA, J. R.; ESTÉVEZ, P. A. A review of feature selection methods based on mutual information. *Neural computing and applications*, Springer, v. 24, n. 1, p. 175–186, 2014. Citado na página 13.

WANG, Y. et al. Water quality prediction method based on lstm neural network. In: IEEE. *2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*. [S.l.], 2017. p. 1–5. Citado na página 10.

YAURI, R.; RIOS, M.; LEZAMA, J. Water quality monitoring of peruvian amazon based in the internet of things. In: IEEE. *2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON)*. [S.l.], 2017. p. 1–4. Citado na página 8.

ZHANG, Y. et al. Applying multi-layer artificial neural network and mutual information to the prediction of trends in dissolved oxygen. *Frontiers in Environmental Science*, Frontiers, v. 7, p. 46, 2019. Citado 5 vezes nas páginas 9, 11, 12, 14 e 18.