APLICAÇÃO DE TÉCNICAS DE APRENDIZADO DE MÁQUINA PARA IDENTIFICAÇÃO DO POTENCIAL HIDROGENIÔNICO (pH) EM AMBIENTE MONITORADO

Resumo:. *O monitoramento da qualidade da água é crucial para a proteção dos recursos hídricos em larga escala e tem grande importância para a manutenção do meio ambiente. Este trabalho aplica Aprendizado de Máquina para melhorar o monitoramento da qualidade da água com base em Multi Layer Perceptron Regression, Support Vector Regression e Random Forest para a definição do Potencial Hidrogeniônico (pH) em ambiente monitorado para a produção de peixes. São utilizados dados de monitoramento de pH fornecidos pela Empresa Brasileira de Pesquisa Agropecuária (Embrapa) de pesquisas realizadas no ano de 2016. Os resultados obtidos mostram que Random Forest obteve melhor desempenho com valores de MAE, MAPE e RMSE iguais a 0,182, 2,4% e 0,058, respectivamente.*

Palavras-chave: *Qualidade da Água. Aprendizado de Máquina. Potencial Hidrogeniônico.*

1 INTRODUÇÃO

Os sistemas de abastecimento de água no mundo são uma das infraestruturas mais importantes do cotidiano das pessoas em geral. De acordo com o relatório da Organização Mundial da Saúde (OMS), ainda há mais de 681 milhões de pessoas na Terra lutando para receber água potável suficiente. Nos sistemas de distribuição e fornecimento existentes, a qualidade da água é um fator chave em todo o processo, desde a fonte de água, tratamento e tubulações distribuídas. A qualidade da água predominante é controlada usando uma série de parâmetros físicos e biológicos. Eles são diferentes de países ou regiões com base nas condições geográficas e de desenvolvimento da região (WU; WANG; SEIDU, 2019).

Em muitas regiões no Brasil e no Mundo a qualidade da água é monitorada por diversas razões. A hipótese é que os diversos indicadores utilizados na avaliação da qualidade da água estão correlacionados. Dentre os principais indicadores utilizados estão diretamente ligados ao potencial hidrogeniônico (pH), temperatura, salinidade, oxigênio, íons, quantidade de sais minerais entre outros (DANIELS, 2021).

A previsão de parâmetros importantes na qualidade da água é baseada nos dados históricos de monitoramento da seção de controle de uma determinada região, usando tecnologia moderna para estimar e especular a tendência de mudança futura da concentração do índice de qualidade da água. Prever a concentração dos principais poluentes na água refletindo o grau de poluição é de grande importância para a manutenção qualidade da água em adequação ao ambiente hídrico, especialmente quando as mudanças na qualidade da água são relacionadas a fatores multivariados (ZHANG; JIN, 2020).

A poluição hídrica é um fator de grande importância em geral, pois os impactos da poluição da água podem acarretar a danos em muitos setores, no entanto, podem ser tratados de forma produtiva se as informações forem examinadas e a qualidade da água for antecipada. Logo, é extremamente essencial conceber novos métodos e técnicas para a deterioração da qualidade da água e determinar a tendência futura da qualidade da água. Por isso, a previsão da qualidade da água é importante, e um modelo de previsão apropriado tem valor prático positivo e influência social (RAGI et al. 2019).

Na literatura científica, diversos autores discutem meios de se identificar parâmetros e índices relevantes para especificar os níveis ideais de qualidade da água, dado que esta é uma realidade que vem se deteriorando a um ritmo alarmante como resultado direto da rápida industrialização e urbanização. As soluções existentes, partem desde alternativas automatizadas em centros de controle de distribuição hídrica, a aplicação de modelos inteligentes e de baixo custo operacional.

Em Li (2022, p. 127659), por exemplo, os autores propõem um modelo de previsão de parâmetros de qualidade da água combinando ao Discrete Hidden Markov Model (DHMM) e métodos de agrupamento K-means. O modelo foi usado para prever a saturação de oxigênio dissolvido e parâmetros para identificar graus de qualidade de turvo, opaco ou espesso com matéria em suspensão presentes nas fontes de água de seis fazendas no interior de Bohai Rim na China.

O processo de predição da qualidade da água é extremamente importante para diversos ecossistemas. Diante disso, os autores de (ISLAM; IRSHAD, 2022) aplicam um modelos de Deep Learning Enabled Water Quality Prediction and Classification (AEODL-WQPC) para prever o índice de qualidade da água em um conjunto de dados de qualidade de água de referência obtido do repositório Kaggle. O modelo prevê diferentes níveis de qualidade da água, elegendo ao final os melhores parâmetros para se especificar a qualidade da água.

Dentre os parâmetros avaliados para se definir o potencial de qualidade da água, o potencial hidrogeniônico, ou pH é um dos parâmetros mais importantes para caracterização da qualidade da água , tanto para o fornecimento de água potável, quanto para utilização na agricultura em geral, podendo influenciar a microbiologia do solo e no processo de troca catiônica entre solo e planta (DE JESUS GUIMARÃES et al. 2021). Nesse contexto, os modelos de previsão da qualidade da água, tornaram-se importantes meios de monitoramento da segurança dos recursos hídricos e garantia essencial para o desenvolvimento sustentável no mundo atual.

Nos últimos anos, com o aumento de dados, o aprendizado de máquina tornou-se um dos algoritmos mais indicado para previsão de medidas, e é amplamente utilizado em ambientes aquáticos. Este trabalho foca-se em prever o pH a partir de outros parâmetros importantes de qualidade da água coletados de um projeto da Empresa Brasileira de Pesquisa Agropecuária (Embrapa) voltado para a piscicultura. No experimento, três modelos de regressão são implementados, dentre eles Multilayer Perceptron Regression (MLP Regression), Support Vector Regression (SVR) e Random Forest Regression (RF Regression).

2 MÉTODO DE AVALIAÇÃO

2.1 Conjunto de Dados

O conjunto de dados utilizado foi extraído de um banco de dados que contém informações sobre a Avaliação da Qualidade da Água na Piscicultura (BioAqua). O sistema apresenta informações coletadas em campo no ano de 2016 em quatro propriedades do interior paulista com cultivo de tilápia, pelo projeto da Embrapa intitulado "Uso de bioindicadores para avaliação da qualidade da água no cultivo da tilápia". Não existe banco de dados similar, no contexto da tilapicultura. É a primeira vez que são reunidos dados dos bioindicadores que colonizam os viveiros escavados com utilização de coletores com substrato artificial, além de dados organizados para a parasitofauna de peixes no monitoramento ambiental.

O banco compreende três matrizes de dados: variáveis físicas e químicas de qualidade de água, ectoparasitos de peixes e organismos bentônicos. A última apresenta uma coleção de fotos e desenhos dos organismos, facilitando a identificação do animal encontrado. Os dados gerados podem ser exportados para planilhas Excel com possibilidade de produzir tabelas e gráficos. Neste trabalho foram usadas apenas as variáveis físicas e químicas de qualidade da água.

2.2 Aprendizado de Máquina

Modelos baseados em M, Support Vector Regression e Random Forest foram aplicados para prever o pH da água, dados as características ambientais e hídricas do ambiente de teste. Para implementação dos modelos, os scripts usaram Python.7.6 e a biblioteca scikit-learn.

2.3 Multilayer Perceptron Regression

Deep Learning é um ramo do aprendizado de máquina que tenta modelar abstrações de dados de alto nível usando várias camadas de neurônios que consistem em estruturas complexas ou transformações não lineares. Com o aumento da quantidade de dados e o poder da computação, redes neurais com estruturas mais complexas têm despertado ampla atenção e sido aplicadas em diversos campos.

O aprendizado não depende de processamento prévio de dados e extrair recursos automaticamente. Para usar um exemplo simples, uma Deep Learning encarregada de interpretar formas aprenderia a reconhecer arestas simples na primeira camada e, em seguida, adicionaria o reconhecimento das formas mais complexas compostas por essas arestas.

Neste trabalho foi utilizada uma arquitetura de Deep Learning chamada MLP (Multi Layer Perceptron) (BISHOP et al. 1995). O modelo de MLP possui uma ou mais camadas ocultas entre suas camadas de entrada e saída, os neurônios são organizados por camadas, as conexões são sempre direcionadas das camadas inferiores para as camadas superiores e os neurônios de uma mesma camada não são interconectados. Para aplicar o MLP em um problema de regressão, basta alterar a função de ativação dos neurônios, formando assim um modelo chamado de MLP Regression.

2.4 Support Vector Regression

Support Vector Regression é uma técnica analítica para investigar a relação entre uma ou mais variáveis preditoras e uma variável dependente de valor real (contínua). Ao contrário dos

métodos tradicionais de regressão que dependem de suposições do modelo que podem não ser precisas (por exemplo, distribuição linear de dados), o SVR é uma técnica de aprendizado de máquina na qual um modelo aprende a importância de uma variável para caracterizar a relação entre entrada e saída.

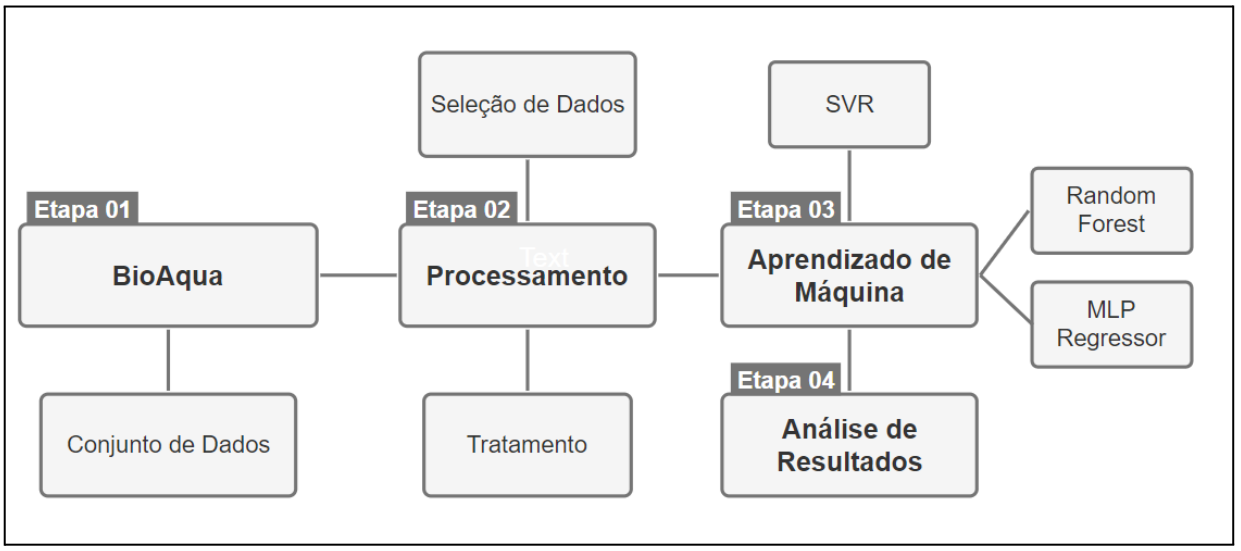
2.5 Random Forest

Trata-se de um termo geral que abrange classificadores baseados em comitês de árvores de decisão. O método básico segue os passos do algoritmo bagging para construir os classificadores base, isto é, as árvores de decisão. Além da amostragem *bootstrap*, na qual amostras aleatórias do conjunto de treinamento são geradas com reposição a partir do conjunto original, e da votação por maioria, utilizadas no bagging, o Random Forest aplica ainda métodos de subespaços aleatórios (MIENYE; SUN; WANG, 2020) na construção do conjunto de treinamento, o que promove diversidade nos classificadores base, podendo melhorar o desempenho de generalização do comitê.

2.6 Modelo de Avaliação

O modelo de avaliação desenvolvido para este estudo consiste em uma sequência de etapas que abrange desde a coleta de informações no banco de dados BioAqua, processamento de informações, aplicação dos modelos de aprendizado de máquina até a etapa final de análise dos resultados obtidos.

Figura 1 – Modelo de Avaliação.



Conforme visto na Figura 1, o modelo de avaliação contém 4 etapas de desenvolvimento. Na Etapa 01, é feita a coleta de informações e alimentação do conjunto de dados no BioAqua. Na Etapa 2, os dados obtidos são tratados através de técnicas de *Data Science* (tratamento de features, limpeza, seleção e identificação de Insights), identificando as variáveis de maior sensibilidade para o experimento, conforme visto na Tabela 01.

Com os dados processados, o código[1] dos modelos de aprendizado de máquina (SVR, MLP e Random Forest) são aplicados para previsão do pH que é um dos indicadores para a qualidade ideal da água. Na etapa final (Etapa 04), é realizada a análise dos resultados e o melhor modelo é especificado, com base nas métricas de avaliação de performance.

[1] https://github.com/albert-santos/water-ph-prediction

O conjunto de variáveis, expresso pela Tabela 1, apresenta o nome e o tipo das variáveis, ou features, presentes no *dataset*. Além disso, não há valores nulos no conjunto de dados utilizados.

Tabela 1 – Variáveis do conjunto de dados

Nome	Descrição	Tipo
T	Temperatura em graus celsius (ºC)	Contínua
ORP	Potencial de oxidação/redução em miliVolt (mv)	Contínua
Condutividade	Condutividade em miliSiemens/cm (mS/cm)	Contínua
Turbidez	Turbidez em Unidade de Turbidez Nefelométrica (ntu)	Contínua
OD	Oxigênio dissolvido (mg/L)	Contínua
TDS	Total de sólidos dissolvidos (ppm)	Contínua
Salinidade	Salinidade da água (%)	Contínua
pH	pH da água	Contínua

A normalização de uma amostra x do conjunto de dados é calculada de acordo com a Equação (1), sendo u a média das amostras do conjunto de treino e s o desvio padrão. A centralização e o dimensionamento acontecem independentemente em cada *feature*.

$$z = \frac{(x-u)}{s} \quad (1)$$

Foram utilizadas três métricas de avaliação para examinar o desempenho dos modelos aplicados neste trabalho: MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error) e RMSE (Root Mean Square Error). Estas métricas são definidas da seguinte forma:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y'_i - y_i| \quad (2)$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y'_i - y_i}{y_i}\right| \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y'_i - y_i)^2} \quad (4)$$

A definição dos parâmetros de cada modelo foi realizada com base em um *grid search* utilizando o valor de MAPE para classificação do melhor resultado. Para o melhor MLP *regression* selecionado a quantidade de neurônios na camada escondida foi de 10 utilizando o otimizador LBFGS. Para o SVR, C (parâmetro de regularização) foi igual a 3 com o kernel RBF. Por fim, para o RF *regression* a função selecionada para medir a qualidade de divisão foi a *poisson* com o número de árvores igual a 10. A Tabela 2 apresenta o intervalo e os parâmetros selecionados pelo *grid search.*

Tabela 2 – Parâmetros utilizados em cada modelo

Modelo	Intervalo dos parâmetros	Parâmetros selecionados
MLP Regression	hidden_layer_size=[8-40], optimizer=[‘lbfgs’, ‘sgd’, ‘adam’]	hidden_layer_size = 10, optimizer= ‘lbfgs’
SVR	C=[1.0-40.0], kernel=[‘linear’,’poly’, ’rbf’, ‘sigmoid’]	C=3.0, kernel=’rbf’
RF Regression	criterion=[‘squared_error’,‘absolute_error’,‘ poisson’], n_estimators=[1-100]	criterion=’poisson’, n_estimator=10

Além disso, o *dataset* foi dividido antes da normalização na proporção de 75% para treinamento e 25% para teste e o processo de validação cruzada *k-fold* foi realizado durante o *grid search* com k=3.

3 RESULTADOS

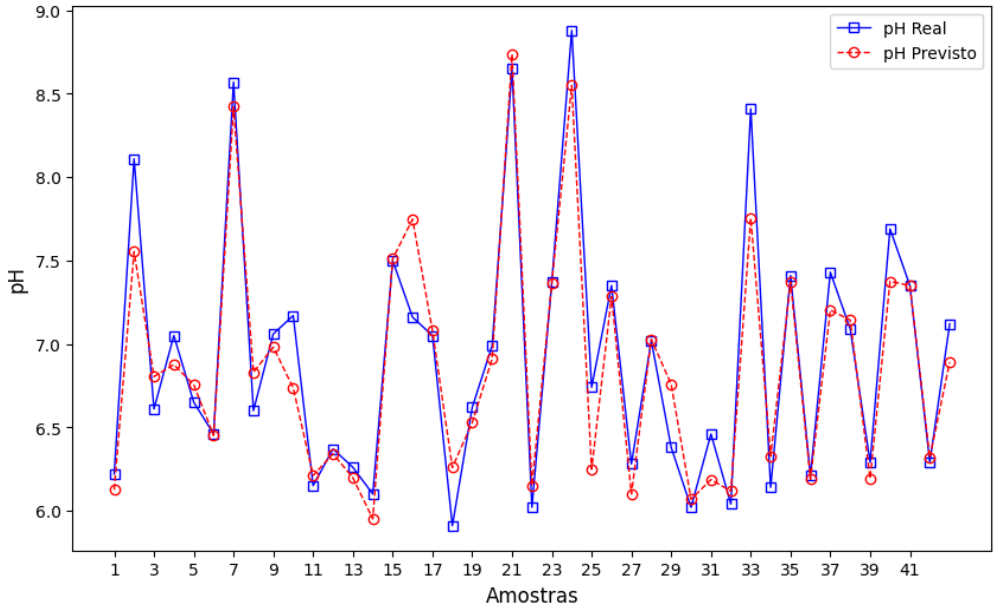
Observa-se que os modelos implementados obtiveram bom desempenho para prever o pH da água, conforme visto na Tabela 03. No entanto, o RF Regression apresentou um desempenho ligeiramente melhor nas três métricas de avaliação utilizadas. A melhora em relação ao MLP regressor, segundo melhor modelo, foi de 9,4%, 0,4% e 24,6% em relação aos valores de MAE, MAPE e RMSE, respectivamente.

Tabela 3 – Desempenho dos modelos

Modelo	MAE	MAPE	RMSE
MLP Regression	0.201	2.8 %	0.077
SVR	0.206	2.9 %	0.079
RF Regression	**0.182**	**2.4 %**	**0.058**

O MLP regressor, expresso na Tabela 3, utiliza apenas uma camada intermediária e a quantidade de amostras é relativamente pequena, o que faz com que esse modelo não seja capaz de superar o RF regression.

Figura 1 – pH previsto pelo RF regression no conjunto de tes



Fonte: Imagem produzida pelos autores

A Figura 1 apresenta os valores observados de pH no conjunto de teste e as previsões feitas pelo modelo RF Regression. Além disso, demonstra a eficiência desse modelo em prever o

parâmetro de pH da água. Nota-se que os valores previstos e observados se sobrepõem muito bem, o que assegura a acurácia do modelo. No geral, o modelo fornece uma boa previsão para o parâmetro de qualidade da água considerado.

4 CONSIDERAÇÕES FINAIS

O presente trabalho fornece uma aplicação de modelos MPL regression, SVR e RF regression para a previsão do pH, um importante parâmetro de qualidade da água. A previsão foi baseada em um conjunto de dados reais coletados pela Embrapa. Uma correlação muito boa foi observada entre os valores medidos e previstos quando os modelos foram aplicados. Entre os três modelos aplicados, o RF regressor obteve o melhor desempenho. Os resultados revelam que é totalmente viável prever o pH da água a partir de outros parâmetros importantes como temperatura, OD e salinidade.

Agradecimentos

Os autores gostariam de agradecer à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e a Universidade Federal do Pará pelo apoio financeiro e suporte técnico que ajudaram a melhorar significativamente a qualidade deste trabalho.

REFERÊNCIAS

BISHOP, Christopher M. et al. **Neural networks for pattern recognition**. Oxford university press, 1995.

DANIELS, Alexis. **Predicting Water Quality Parameters in Lake Pontchartrain Using Machine Learning**. 2021. Tese de Doutorado. Southeastern Louisiana University.

DE JESUS GUIMARÃES, João et al. **Effect of irrigation water pH on the agronomic development of hops in protected cultivation**. Agricultural Water Management, v. 253, p. 106924, 2021.

ISLAM, Nazrul; IRSHAD, Kashif. **Artificial ecosystem optimization with Deep Learning Enabled Water Quality Prediction and Classification model**. Chemosphere, p. 136615, 2022.

RAGI, Nikhil M.; HOLLA, Ravishankar; MANJU, G. **Predicting water quality parameters using machine learning**. In: 2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT). IEEE, 2019. p. 1109-1112.

LI, Dashe et al. **An advanced approach for the precise prediction of water quality using a discrete hidden markov model**. Journal of Hydrology, v. 609, p. 127659, 2022.

MIENYE, Ibomoiye Domor; SUN, Yanxia; WANG, Zenghui. **An improved ensemble learning approach for the prediction of heart disease risk**. Informatics in Medicine Unlocked, v. 20, p. 100402, 2020.

WU, Di; WANG, Hao; SEIDU, Razak. **A Tensor Model for Quality Analysis in Industrial Drinking Water Supply System**. In: 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech). IEEE, 2019. p. 1090-1092.

ZHANG, Huiqing; JIN, Kemei. **Research on water quality prediction method based on AE-LSTM**. In: 2020 5th International Conference on Automation, Control and Robotics Engineering (CACRE). IEEE, 2020. p. 602-606.