Predição de Oxigênio Dissolvido em amostras de água

Daniel Ambrosim Falqueto

FGV - Fundação Getúlio Vargas, junho de 2023

1 Introdução

Encontrar maneiras boas e baratas de medir o oxigênio dissolvido na água pode ser um desafio devido à necessidade de equipamentos especializados e à complexidade da análise. Métodos precisos, como sondas de oxigênio dissolvido, requerem equipamentos mais avançados e podem ser mais caros. Além disso, a calibração e manutenção desses dispositivos também podem ser exigentes. Existem alternativas mais econômicas, mas as medições podem não ser tão precisas. É importante avaliar cuidadosamente as opções disponíveis e considerar o equilíbrio entre custo, precisão e praticidade ao medir o oxigênio dissolvido na água para que a quantidade seja necessária numa determinada situação seja mantida.

A medição do oxigênio na água, sem a necessidade de métodos complicados, desempenha um papel crucial na criação e transporte de peixes. O oxigênio dissolvido na água é essencial para a respiração dos peixes, garantindo o bom funcionamento de seus sistemas respiratórios. Manter níveis adequados de oxigênio é vital para a sobrevivência e o bem-estar dos peixes em ambientes aquáticos. A medição precisa e contínua do oxigênio permite aos aquicultores e profissionais de transporte de peixes monitorar a qualidade da água em tempo real, identificar possíveis problemas de oxigenação e tomar medidas corretivas imediatas, como aeração adequada ou ajuste dos níveis de oxigênio. Isso ajuda a prevenir doenças, estresse e mortalidade dos peixes, promovendo uma produção saudável e sustentável, além de garantir a segurança durante o transporte dos animais. Simplificar os métodos de obtenção da quantidade de oxigênio dissolvido na água facilita o monitoramento regular e acessível, contribuindo para o sucesso da criação e transporte de peixes de forma eficiente e responsável.

Diferentes espécies de peixes têm necessidades diferentes de oxigênio dissolvido na água. Alguns peixes requerem níveis mais altos de oxigênio para sobreviver, enquanto outros são mais tolerantes a baixos níveis de oxigênio. Em geral, peixes de água fria, como trutas e salmões, tendem a exigir mais oxigênio dissolvido na água. Essas espécies estão adaptadas a habitats com água rica em oxigênio, como rios de montanha e lagos profundos.

A quantidade específica de oxigênio dissolvido necessária, pode variar de espécie para espécie e também depende das condições ambientais. No entanto, em águas bem oxigenadas, a maioria dos peixes prospera em níveis de oxigênio dissolvido de pelo menos 5 a 6 mg/L (miligramas por litro). Alguns peixes mais exigentes podem precisar de níveis acima de 8 mg/L.

É importante observar que, em condições de baixo oxigênio dissolvido na água, os peixes podem enfrentar estresse, doenças e até morte. Portanto, é fundamental monitorar e manter níveis adequados de oxigênio dissolvido em aquários, lagoas ou outros habitats aquáticos onde os peixes são criados ou mantidos.

Por isso, neste presente artigo, vamos tentar encontrar o melhor modelo para a predição da quantidade de oxigênio dissolvido na água com base nos parâmetros que podem ser medidos com facilidade, pH, temperatura e turbidez. Além disso, trazemos outro modelo para casos onde seja possível obter a condutividade da água.

Os parâmetros da água obtidos no dataset são explicados e Apresentação dos principais procedimentos de medição são descritos a seguir:

Condutividade (Conductivity): A condutividade da água é uma medida da capacidade da água de transmitir corrente elétrica devido à presença de íons dissolvidos. As maneiras de medição envolvem o uso de sondas de condutividade, que requerem equipamentos especializados, como condutivímetros, eletrodos de alta precisão e calibração frequente. Além disso, é necessário um ambiente controlado e técnicas de amostragem cuidadosas para evitar interferências e obter resultados precisos. Essas abordagens são mais comumente aplicadas em laboratórios de pesquisa e indústrias que exigem medições altamente precisas da condutividade da água.

Oxigênio Dissolvido (Dissolved Oxygen): Esta variável mede a quantidade de oxigênio dissolvido na água, que é essencial para a vida aquática. Altos níveis de oxigênio dissolvido são essenciais para a sobrevivência de peixes e outras formas de vida marinha. As maneiras complexas e caras de medir o oxigênio dissolvido na água envolvem o uso de sondas de oxigênio dissolvido, também conhecidas como medidores de OD.

Esses equipamentos possuem eletrodos sensíveis que detectam e quantificam o oxigênio dissolvido na água. As sondas de OD requerem calibração frequente e manutenção adequada para garantir resultados precisos. Além disso, são necessários dispositivos de leitura e registradores de dados para obter leituras contínuas e precisas. Esses métodos são mais comumente utilizados em laboratórios de pesquisa, indústrias e estações de monitoramento ambiental, onde medições precisas e em tempo real do oxigênio dissolvido são essenciais.

pH: O pH mede a acidez ou alcalinidade de um líquido em uma escala de 0 a 14. Um valor inferior a 7 indica condições ácidas, um valor superior a 7 indica uma necessidade primária e um valor de 7 indica um estado neutro. Existem maneiras simples e baratas de medir o pH da água. Uma opção comum é o uso de tiras de teste de pH, que são pequenas tiras de papel impregnadas com indicadores químicos sensíveis ao pH. Basta mergulhar a tira na água e comparar a cor resultante com uma escala de cores fornecida nas embalagens. Outra opção é o uso de kits de teste de pH líquido, que contêm uma solução indicadora e um frasco de teste. Basta adicionar algumas gotas da solução à água e observar a mudança de cor. Esses métodos são econômicos, simples de usar e amplamente disponíveis em lojas de produtos químicos, laboratórios ou lojas especializadas em jardinagem.

Temperatura (Temperature): A temperatura afeta muitos processos físicos, químicos e biológicos em corpos d'água. É um fator importante que afeta a taxa de muitos processos aquáticos. Existem opções simples e acessíveis para medir a temperatura da água, como termômetros de mercúrio, termômetros digitais e termômetros de infravermelho. Esses dispositivos fornecem leituras rápidas e precisas, sendo ideais para uso em poços de criação de peixes, aquários e meios de transporte de peixes. Termômetros de uso único também são uma alternativa econômica e conveniente. É importante seguir as instruções e garantir a calibração correta para resultados confiáveis.

Turbidez (Turbitidy): A turbidez da água é uma medida da quantidade de partículas suspensas, como sedimentos, argila, matéria orgânica ou microorganismos, que afetam a transparência visual da água. As maneiras de medição incluem o uso de turbidímetros portáteis, que são dispositivos compactos e acessíveis que emitem luz e medem a quantidade de luz dispersa pelas partículas na água. Além disso, kits de teste de turbidez, que usam produtos químicos e um comparador visual para determinar a turbidez, também são uma opção econômica. Ambos os métodos fornecem resultados relativos de turbidez e são amplamente utilizados para monitorar a qualidade da água em diversas aplicações.

2 Métodos

No presente estudo, foi adotado um modelo linear para descrever a relação entre as variáveis investigadas. Inicialmente, foi observado um padrão visual nos dados que sugeria uma tendência linear. Para encontrar o modelo que melhor se ajusta e explica os dados foi utilizado o Modelo Generalizado Linear e a família gaussiana (ou normal) na linguagem R.

Para escolher o modelo vamos obervar pricipalmente as 3 seguintes métricas: MAE, AIC e Erro Residual, nessa ordem de significância.

Para obter erro médio absoluto, MAE (Mean Absolute Error), utilizamos Leave-One-Out Cross Validation (LOOCV) que é um método de amostragem usado para avaliar o desempenho de um modelo preditivo. Este método funciona da seguinte forma:

São feitas a mesma quantidade de iterações do que de amostras no conjunto de dados. Em cada iteração, um ponto dos dados é excluído do conjunto de dados, o modelo é treinado usando os pontos de dados restantes, excluindo o ponto deixado de fora. O modelo é então usado para prever a variável de resposta para o ponto de dados deixado de fora no passo inicial e esse valor é somado. Esses processos são repetidos para cada ponto do conjunto de dados, deixando de fora um ponto de dados de cada vez. Com isso, após a última iteração, o erro de previsão geral é calculado tirando a média dos erros de teste obtidos.

O MAE é calculado a partir da média dos erros absolutos, ou seja, utilizamos o módulo de cada erro. O erro, pode ser interpretado como a diferença em valor absoluto entre Y_i e sua média e assim, temos:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \bar{Y}|$$

A métrica MAE é sensível a outlier, ou seja valores muito distantes da média, porém como nos dados todos os valores variam pouco não teremos problemas com isso. Outra razão para utilizar MAE ao invés de

RSME (Root Mean Squared Error), além de os dados serem bem comportados é o fato da baixa variação dos dados, que em grande parte dos valores a difererença em relação a média é menor que 1, ou seja elevando esse erro ao quadrado o tornaríamos menor do que realmente é, e os erros maiores do que 1 ficariam maiores, tornando a análise mais imprecisa.

Neste estudo não vamos utilizar a métrica R-squared (R^2) porque essa pode não ser uma métrica ideal quando os dados têm pouca variação, pois o R^2 mede a proporção da variabilidade total explicada pelo modelo. Se a variabilidade total for pequena, é possível que o R^2 também seja baixo, mesmo que o modelo esteja ajustado adequadamente aos dados.

A métrica AIC (Akaike's Information Criterion), é um critério estatístico utilizado para comparar e selecionar modelos estatísticos. Ele foi desenvolvido por Hirotugu Akaike e fornece uma medida relativa de qualidade do ajuste de um modelo, levando em consideração a complexidade do modelo e o número de parâmetros estimados.

O AIC é calculado a partir da função de verossimilhança do modelo e do número de parâmetros estimados. Quanto menor o valor do AIC, melhor é o ajuste do modelo aos dados. O critério penaliza modelos mais complexos, evitando o sobreajuste, pois leva em conta não apenas a qualidade do ajuste, mas também a quantidade de informação que o modelo utiliza.

Ao comparar diferentes modelos usando o AIC, o modelo com o menor valor de AIC é considerado o melhor ajuste entre as opções consideradas. Além disso, a base lógica do AIC se encaixa no princípio da Navalha de Occam. Segundo este princípio, dadas duas hipóteses (modelos estatísticos) de mesmo poder explicativo para determinado fenômeno, a hipótese mais simples têm maior chance de estar correta. O AIC leva em conta e penaliza a complexidade dos modelos e tende a favorecer a escolha de modelos mais simples.

O Erro Residual, também conhecido como resíduo, refere-se à diferença entre o valor observado e o valor previsto por um modelo estatístico. Em outras palavras, é a discrepância entre os dados reais e as estimativas fornecidas pelo modelo. O erro residual é a medida da variabilidade não explicada pelo modelo.

Para cada ponto no conjunto de dados, o erro residual é calculado subtraindo-se o valor previsto pelo modelo do valor observado. Um erro residual positivo indica que o valor observado é maior do que o valor previsto, enquanto um erro residual negativo indica que o valor observado é menor do que o valor previsto. Novamente, quanto mais próximo de zero o Erro Residual, melhor ajustado aos dados estará o modelo.

Além disso vamos olhar também para as previsões do intercepto em cada modelo. O intercepto representa o valor previsto da variável dependente quando todas as variáveis independentes são iguais a zero. Ele captura o efeito médio não explicado pelas variáveis independentes no modelo. Como nos nossos dados a variável dependente (Oxigênio dissolvido) varia pouco e é pequena, não queremos um modelo cujo intercepto seja muito grande nem um modelo cujo intercepto seja negativo.

Os coeficientes das variáveis independentes fornecem informações sobre o efeito individual dessas variáveis na variável dependente, controlando os outros efeitos no modelo. Os coeficientes indicam o tamanho da mudança esperada na variável dependente quando a variável independente correspondente aumenta em uma unidade, mantendo as demais variáveis constantes. Ao analisar esses coeficientes podemos identificar quais variáveis têm um impacto significativo na previsão e compreender como as variáveis independentes influenciam a variável dependente e como elas se relacionam entre si.

3 Resultados

Os dados observados são de 500 amostras diferentes de água contendo a medição de 5 parâmetros, dentre eles Oxigênio dissolvido, que queremos prever com base nos outros. Abaixo temos o histograma da quantidade de Oxigênio dissolvido medido em miligramas por litro (mg/L). Observe que os dados variam pouco, apenas de 6 a 10.

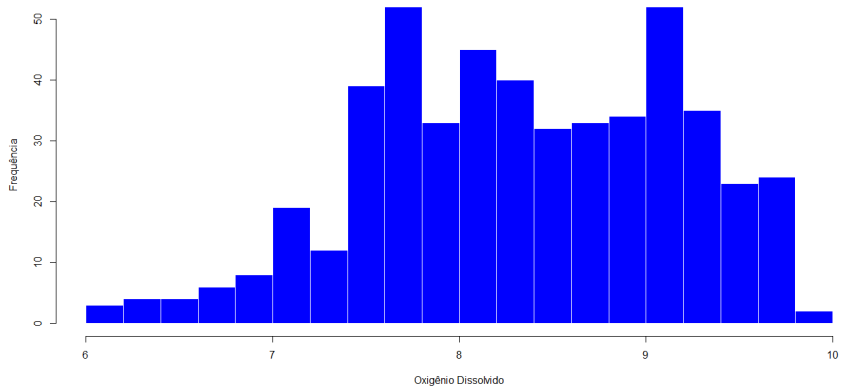

Figura 1: Histogrâma dos dados de Oxigênio Dissolvido

Dentre os parâmetros observados nos dados o que melhor aparenta, à primeira vista, ter correlação com a quantidade de Oxigênio dissolvido é o pH.

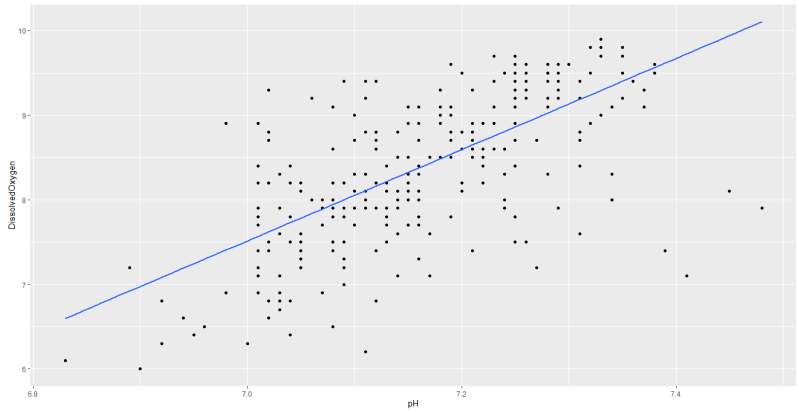

Figura 2: Plot pH e Oxigênio dissolvido

Para simplificar a escrita dos modelos vamos simplificar o nome das variáveis:
Dissolved Oxygen = DO; pH = pH; Turbidity = Tb; Temperature = Tp; Condutividade = Co.

Sabendo que estamos lidando com um modelo linear, como explicado anteriormente, vamos comparar algumas métricas para qualificar qual é o melhor modelo, como AIC, Erro Residual e MAE. Para obter o modelo que melhor se ajustasse e prevesse os dados foram testados grande maioria dos possíveis modelos, que continham combinações dos parâmetros, parâmetros ao quadrado, combinações lineares dos parâmetros dois a dois e até mesmo a combinação linear dos três parâmetros.

O modelo que obteve menor AIC e Erro Residual foi a combinação de todos os parâmetros citados anteriormente com excessão dos parâmetros ao quadrado, ou seja:

$$DO = pH + Tb + Tp + pH \cdot Tb + pH \cdot Tp + Tp \cdot Tb + pH \cdot Tb \cdot Tp.$$

O modelo obteve AIC = 798,18 e Erro Residual = 139,36. Entretanto, o indicador MAE de 0,4102 foi maior em comparação com vários outros modelos. Além disso, o intercepto obtido foi de 1387 algo que está muito distante dos dados reais, que variam de 6 a 10. Ademais, a estimação dos coeficientes e de seus desvios padrões eram muito altos. Por isso esse modelo foi rejeitado.

Outro modelo que se ajustou bem aos dados foi:

$$DO = pH + (Tb)^2 + pH \cdot Tb + pH \cdot Tp + Tp \cdot Tb.$$

O modelo obteve AIC = 812,96 e Erro Residual = 144,70. O indicador MAE foi de 0,4087, um dos menores até então e Intercepto igual a 3,39. Porém o grau de significância de $(Tb)^2$ era muito baixo e apresentava pouca diferença nos dados pois sua estimativa era muito próxima de zero.

Por isso o modelo final escolhido foi:

$$DO = pH + pH \cdot Tb + pH \cdot Tp + Tp \cdot Tb.$$

Este modelo é similar ao modelo anterior, com a diferenção que o modelo escolhido não tem o termo $(Tb)^2$. O modelo obteve AIC = 811,83, pouco menor que o anterior, Erro Residual = 144,90, ligeiramente maior que o anterior e MAE = 0,4065, menor que o anterior. Para curiosidade do leitor, o RSME também foi um dos melhores entre os modelos estudados, sedo de 0,546. As estimativas para os coeficientes também são similares sendo 1,19 para o intercepto.

Segue abaixo as estimativas para os valores dos coeficientes, juntamente com seus erros padrões:

Tabela 1: Coeficientes do modelo escolhido

Coeficientes	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	1.19750	6.42413	0.186	0.8522
pH	-3.37584	1.79130	-1.885	0.0601
Turbidity:Temperature	-0.33137	0.06945	-4.771	2.42e-06
pH:Turbidity	0.97452	0.21580	4.516	7.89e-06
pH:Temperature	0.20709	0.04087	5.068	5.70e-07

Para melhor visualização, abaixo temos a predição dos coeficientes com 95% de confiança.

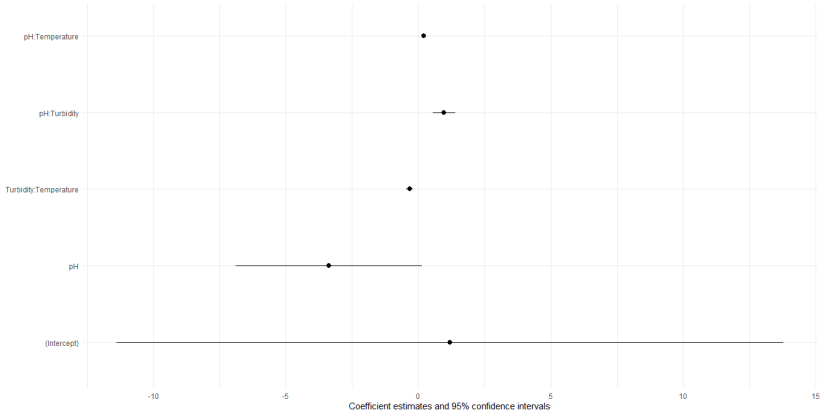

Figura 3: Predição dos coeficientes com 95% de confiança

A Condutividade da água, também tem boa correlação com a quantidade de Oxigênio Dissolvido como é possível ver no plot a seguir:

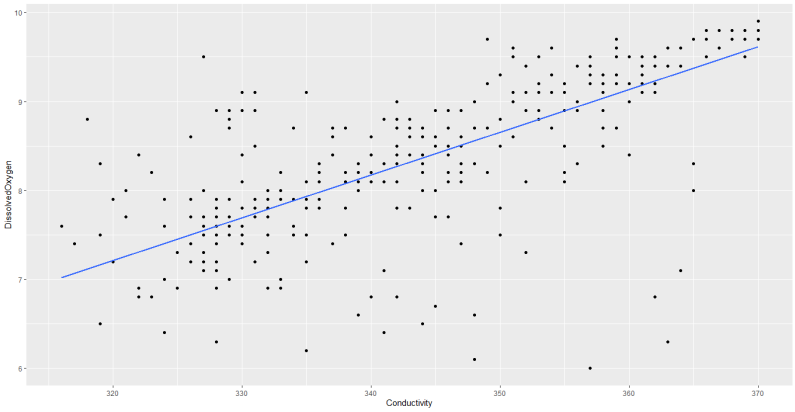

Figura 4: Plot Condutividade e Oxigênio dissolvido

Portanto, foi desenvolvido um modelo que incorporasse os dados relacionados à condutividade da água, com o intuito de abordar essa variável como parte do processo de análise. O modelo que continha condutividade e obteve as melhores métricas foi:

$$Do = Tb + Co + pH \cdot Tb$$

Para escolha do modelo foram observadas as mesmas métricas do modelo sem condutividade. O modelo obteve AIC = 642,81, Erro Residual = 103,79, MAE = 0,3206. Além disso os coeficientes foram:

Tabela 2: Coeficientes do modelo com condutividade

Coeficientes	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.883166	0.766673	-1.152	0.25
Turbidity	-5.051644	0.437932	-11.535	$< 2 \times 10^{-16}$
Conductivity	0.031871	0.002114	15.077	$< 2 \times 10^{-16}$
Turbidity:pH	0.648240	0.061066	10.615	$< 2 \times 10^{-16}$

Para melhor visualização, abaixo temos a predição dos coeficientes com 95% de confiança.

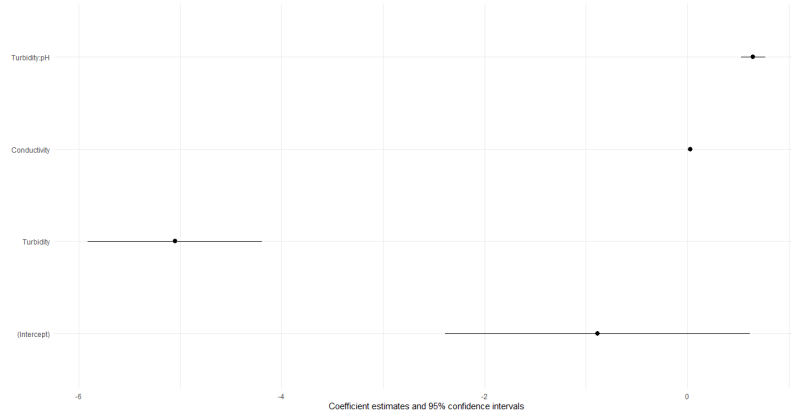

Figura 5: Predição dos coeficientes com condutividade com 95% de confiança

4 Conclusão

A partir dos modelos podemos observar que os parâmetros combinados podem ser melhor ajustados aos em comparação aos parâmetros isolados em si. No modelo escolhido sem condutividade, apenas o parâmetro pH ficou isolado e o modelo que obteve menor AIC e Erro residual foi o que dispunha de todas as possíves combinações dos parâmetros, por mais que esse não tenha sido o modelo escolhido.

Em média o modelo prevê com erro aproximado de 0,4. Supondo que esse modelo fosse utilizado para estudo e predição de dissolução de oxigênio em resenvatórios de água, poços ou laboratórios de experiência, seria aconselhado corrigir os níveis de oxigênio sempre que a quantidade prevista esteja próxima da quantidade de risco para o experimento.

Ao comparar-mos o modelo que contém Condutividade ao modelo sem Condutividade e avaliar suas métricas de desempenho, observamos que o modelo que contém Condutividade apresenta resultados melhores, indicando que ele é mais eficaz e possui mais informações relevantes para o problema em questão.

Isso acontece pelo fato que o modelo com Condutividade tem mais informação, e a Condutividade tem forte correlação com a quantidade de Oxigênio dissolvido. Então como podemos reparar, todas as métricas utilizadas para combarar a bondade do ajuste estão melhores no modelo com condutividade. O AIC foi 20,8% menor, o Erro Residual foi 28,4% menor e MAE 21,1% menor, tornando o modelo significativamente mais acertivo para previsão e ajuste do conjunto de dados.

A partir disso podemos concluir que quanto mais variáveis num conjunto de dados maiores são as chances do modelo estar melhor ajustado. Adicionar mais variáveis pode permitir que o modelo capture melhor a complexidade e a diversidade dos dados. Ao incluir mais variáveis relevantes, o modelo tem a capacidade de capturar mais nuances e relações entre as variáveis dependentes e independentes, levando a uma melhor representação dos padrões presentes nos dados e assim à melhores predições.

Além disso, o aumento do número de variáveis pode fornecer mais informações ao modelo, permitindo uma melhor estimativa dos parâmetros. Com mais dados disponíveis, o modelo tem uma base mais sólida para calcular os coeficientes e ajustar-se adequadamente às observações. Isso resulta em estimativas mais precisas e confiáveis dos parâmetros do modelo.

Os principais impecílios para a modelagem foram a pouca variação dos dados, o baixo nível de conhecimento prático em modelagem, similaridade nos resultados obtidos nos modelos e a falta de tempo.

Dentre os dados apresentados, para modelagem, os parâmetros de Temperatura, Turbidez e pH variavam pouco. Dados com pouca variação podem resultar em estimativas imprecisas dos parâmetros do modelo. Com menos variação, pode haver menos informações disponíveis para estimar os efeitos das variáveis independentes sobre a variável dependente. Isso pode levar a estimativas enviesadas ou pouco confiáveis dos coeficientes do modelo, tornando as previsões e inferências menos precisas.

Baixos níveis de conhecimento sobre práticas de modelagem podem ser prejudiciais de várias maneiras. Primeiro, a falta de familiaridade com técnicas e métodos de modelagem pode levar a uma seleção inadequada de modelos estatísticos ou algoritmos de aprendizado de máquina. Isso pode levar a modelos mal especificados, ineficientes ou com baixo desempenho. Além disso, a falta de familiaridade com a linguagem pode limitar a compreensão das características e recursos disponíveis para modelagem. Isso pode levar a uma subutilização dos recursos da linguagem que dificulta a exploração de técnicas mais avançadas ou a implementação de modelos mais complexos.

Foi observado também grande proximidade nos resultados obtidos nos diferentes modelos. Em um grande número de modelos o AIC, Erro Residual e MAE obtidos eram muito próximos, tornando difícil a escolha do modelo que melhor se adequava aos dados.

Para continuação e aprofundamento do estudo buscando um modelo que melhor se ajuste e explique os dados, podemos exploração de abordagens mais eficientes, métodos de otimização, técnicas de aprendizado de máquina avançadas e algoritmos mais sofisticados para melhorar a precisão e a eficácia dos modelos.

Nesse contexto, é possível aprimorar modelos existentes através da incorporação de novas variáveis, melhoria da seleção de variáveis, incorporação de estruturas mais complexas ou melhorias nos algoritmos de ajuste do modelo. Essas melhorias podem levar a modelos mais precisos e robustos, capazes de lidar com uma variedade maior de situações e gerar resultados mais confiáveis. Podemos também buscar novos pacotes que para auxiliar a modelagem. A evolução contínua das linguagens de programação e dos pacotes de modelagem oferece oportunidades para aprimorar a eficiência e a funcionalidade dos modelos. Isso pode envolver a adição de recursos avançados, otimizações de desempenho, suporte a novos tipos de dados e integração com

outras ferramentas e bibliotecas relevantes.

Além disso, podemos incluir dados adicionais para testagem de predição e diferença de modelos. Isso pode envolver a exploração de dados não estruturados. As informações de redes sociais e outras fontes de dados estão em constante expansão.

Podemos ainda aprimorar a capacidade de interpretação e visualização dos resultados do modelo. Isso pode ser alcançado através do desenvolvimento de técnicas mais avançadas de visualização, métricas de avaliação mais robustas e abordagens inovadoras para comunicar os resultados de forma clara e compreensível para diferentes públicos.

Referências

[de Jesus 2020] Jesus, Gutelvam R. de: Métricas para avaliação de Modelos de Regressão. https://blog.brq.com/metricas-para-avaliacao-de-modelos-de-regressao-variaveis-continuas-numericas. 2020

[Pregibon] Pregibon, Hastie .: Fitting Generalized Linear Models. https://stat.ethz.ch/R-manual/ R-devel/library/stats/html/glm.html

[ShreyanshVerma27 2023] ShreyanshVerma27: Water Quality Testing | Kaggle. https://www.kaggle.com/datasets/shreyanshverma27/water-quality-testing. 2023

[STHDA 2018] STHDA: Cross-Validation Essentials in R. http://www.sthda.com/english/articles/ 38-regression-model-validation/157-cross-validation-essentials-in-r. 2018