Criação de uma rede neural para classificação de amostras d'água conforme normativa CONAMA 357/05 a partir de um *dataset* da SEMASA

Eduardo Caldeira Vicente
Universidade do Vale do Itajaí - UNIVALI
eduardo.c@edu.univali.br

Anita Maria da Rocha
Universidade do Vale do Itajaí - UNIVALI
anita.fernandes@univali.br

ABSTRACT

This work presents the creation of a neural network trained using a dataset from SEMASA (Serviço Municipal de Água, Saneamento Básico e Infraestrutura) from Itajaí, containing pH, turbidity, and apparent color data of water bodies from various neighborhoods in the municipality, capable of classifying water samples according to the water types defined by the CONAMA (Conselho Nacional do Meio Ambiente) 357/05 standard for embedding in an IoT device. The dataset is composed of 8,928 records from 2018 to 2023, covering neighborhoods São Roque I and II, Arapongas, and Limoeiro. The selected variables are critical for water quality assessment, where pH indicates the acidity or alkalinity of the water, turbidity measures the presence of suspended particles, and apparent color is a visual indicator of dissolved and suspended materials.

KEYWORDS

Artificial Intelligence, Neural Network, Water Monitoring, IoT.

1 Introdução

A crescente preocupação com a qualidade dos recursos hídricos tem impulsionado o desenvolvimento de métodos avançados para monitoramento e classificação da água. Nesse contexto, as Redes Neurais Artificiais (RNAs) emergem como ferramentas promissoras, capazes de lidar com a complexidade e variabilidade dos parâmetros de qualidade da água [1]. Além disso, técnicas de aprendizado de máquina têm sido aplicadas para prever parâmetros de qualidade da água, fornecendo subsídios para o planejamento e gestão hídrica [2].

A Resolução CONAMA nº 357/2005 estabelece diretrizes para a classificação dos corpos d'água e padrões de lançamento de efluentes, servindo como referência normativa para esses estudos [3].

A aplicação de técnicas de aprendizado de máquina no monitoramento e análise da qualidade da água tem se mostrado promissora, permitindo a predição de parâmetros essenciais e a detecção de contaminantes de forma eficiente. Como por apresentado por Silva, 2022 ([4]), que utilizou redes neurais artificiais para analisar a dinâmica do uso e cobertura da terra em bacias hidrográficas, contribuindo para a compreensão dos impactos ambientais na qualidade da água.

Além disso, estudos como o de Pacheco e Pereira, 2018 ([5]), exploraram o uso de *deep learning* em diversas áreas do conhecimento, incluindo a modelagem de recursos hídricos, onde os autores destacam a utilização de algoritmos de aprendizado profundo para a previsão de parâmetros críticos, como vazão de rios, qualidade da água e detecção de poluentes. Eles apontam que as redes neurais profundas têm capacidade de modelar relações complexas entre variáveis climáticas, físicas e químicas, que afetam diretamente os recursos hídricos.

Esses trabalhos demonstram a eficácia de algoritmos de aprendizado de máquina na predição e classificação de parâmetros da qualidade da água, oferecendo alternativas eficientes aos métodos tradicionais de monitoramento. Portanto, a integração de RNAs no monitoramento da qualidade da água, alinhada às normativas vigentes, a qual é o principal objetivo do presente trabalho, representa um avanço significativo na preservação dos recursos hídricos e na promoção da saúde pública.

2 Solução Proposta

Este trabalho se caracteriza como pesquisa aplicada [6], focando na criação de uma rede neural para classificar amostras de água conforme a normativa CONAMA 357/05 a partir de uma base de dados disponibilizada pela SEMASA de Itajaí. A abordagem qualitativa [7] visa avaliar a acurácia do modelo gerado para uma posterior implantação em um microcontrolador para a criação de um dispositivo AIoT.

2.1 Base de dados

A base de dados disponibilizada pela SEMASA do município de Itajaí é composta por 8.928 registros dos anos de 2018 a 2023, segregados de hora em hora, conforme demonstrado na Figura 1. Dentre as variáveis presentes estão pH, cor aparente e turbidez de corpos d'água dos bairros São Roque I e II, Arapongas e Limoeiro. As variáveis foram escolhidas por sua relevância na avaliação da qualidade da água: o pH reflete o equilíbrio ácido-base, a turbidez

indica a presença de partículas em suspensão, e a cor aparente está associada à presença de compostos dissolvidos e materiais orgânicos.

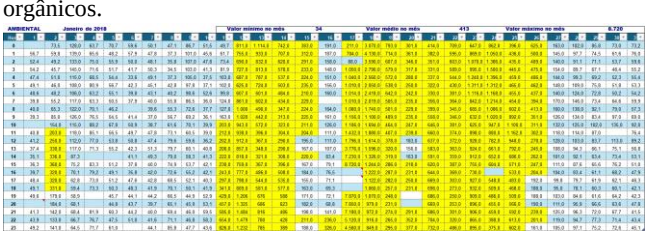

Figura 1: Base de dados da SEMASA

O pré-processamento da base de dados foi realizado na linguagem R, em que foram analisadas a relação de dados faltantes, a simulação de preenchimento de dados utilizando algoritmos de *predictive mean matching*, a matriz de correlação, a detecção de outliers por meio da técnica z-score, além das estatísticas de média, moda e desvio padrão de cada variável envolvida. A técnica z-score consiste em padronizar os dados de acordo com sua média e desvio padrão, atribuindo um valor que representa a distância de cada observação em relação à média, em termos de desvios padrão. Observações que apresentam valores absolutos de z-score superiores a um determinado limiar, comumente 3, são identificadas como potenciais outliers.

Para o treinamento da rede neural os dados precisam estar organizados em uma estrutura simples e de fácil processamento, desta forma foi necessário ajustar a planilha em uma tabela única com as colunas de: ano, mes, dia, hora, ph, cor_aparente, turbidez e classe_corpo_agua. Junto a isto foi necessário realizar o efetivo preenchimento dos dados faltantes, ao qual foi feito a partir da média de cada uma das variáveis.

2.1 Rede Neural Desenvolvida

A criação da rede neural foi realizada na linguagem Python utilizando como base a biblioteca *tensorflow*. A base foi separada em 70% para treinamento e 30% para validação. A construção da rede foi definida como modelo de predição e elaborada com três camadas, sendo a primeira com 16 neurônios e função de ativação ReLu, a segunda com 8 neurônios utilizando também função de ativação ReLu e por último uma camada de saída com função de ativação *softmax* para classificação.

A função de ativação ReLu foi utilizada nas duas primeiras camadas por sua eficácia em mitigar o problema do desaparecimento do gradiente e por acelerar a convergência durante o treinamento. O número de neurônios foi definido empiricamente após testes exploratórios para otimizar a acurácia sem superdimensionar o modelo.

A compilação do modelo foi realizada utilizando o algoritmo de otimização “adam” devido à sua capacidade de adaptação dinâmica da taxa de aprendizado, otimizando a convergência para mínimos locais. A função de perda “*categorical_crossentropy*” foi escolhida por ser adequada para problemas de classificação multiclasse, como o abordado neste trabalho. Na sequência o treinamento da base foi realizado com 50 épocas em lotes de 16 registros cada.

Por fim foi realizada a avaliação do modelo, o que inclui a verificação da acurácia geral da rede, a geração da matriz de confusão relacionado a classe prevista e a classe real e a exibição de gráficos de acurácia e perda conforme a evolução das épocas de treinamento.

3 Resultados e Discussões

Nesta seção são descritos os resultados obtidos a partir da aplicação das técnicas citadas no tópico 2.

3.1 Pré-Processamento

Dentre as análises relacionadas ao pré-processamento da base é importante evidenciar os dados faltantes, e matriz de correlação das variáveis.

A base apresenta pouca ausência de dados quando verificamos as variáveis de turbidez e pH, sendo ambas por volta dos 4%, já ao verificar a falta de dados relacionado a cor aparente, podemos observar uma porcentagem de 50% de dados faltantes conforme visto na Figura 2.

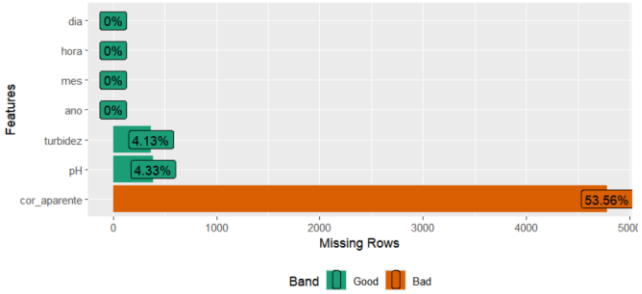

Figura 2: Dados faltantes

Por fim ao verificar a matriz de correlação das variáveis, podemos notar que de forma geral além da relação entre as mesmas variáveis (exemplo: pH com pH) nota-se uma maior relação entre cor aparente e turbidez da água como demonstrado na Figura 3. Além desses, observa-se uma pequena correlação entre pH, cor aparente e turbidez com o mês do ano, acredita-se que esta relação se deve principalmente à meses mais chuvosos e meses de pouca chuva.

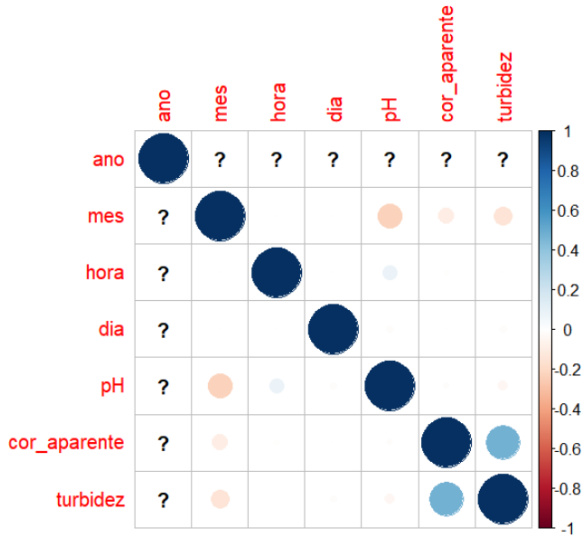

Figura 3: Matriz de correlação de variáveis

3.2 Modelo de rede neural

A rede apresentou uma acurácia de 98% na classificação dos registros disponíveis utilizados para a validação (30% da base). Para verificar a divergência na classificação, foi utilizada a matriz de confusão relacionando a classe prevista pela rede e a classe real do registro, de acordo com a Figura 4, a qual indica que os erros de classificação ocorrem, em sua maioria, entre classes vizinhas, refletindo uma transição gradual entre as classes de qualidade da água, como esperado na natureza dos dados ambientais.

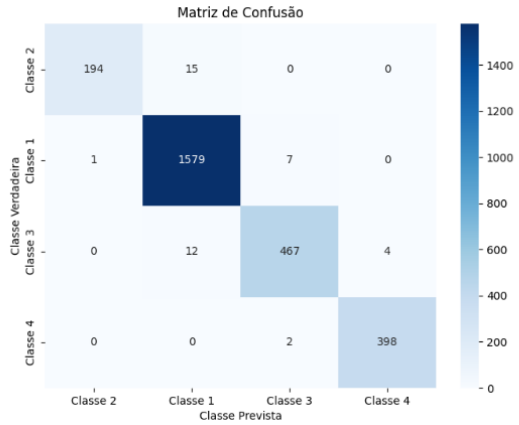

Figura 4: Matriz de confusão classe prevista x real

Então foi verificada a evolução da acurácia e da perda da rede conforme a evolução do treinamento, a partir disso, pode-se determinar que mais épocas de treinamento não ajudariam a aumentar a acurácia do modelo, uma vez que a progressão nas épocas após a época 30 é muito baixa como observa-se na Figura 5.

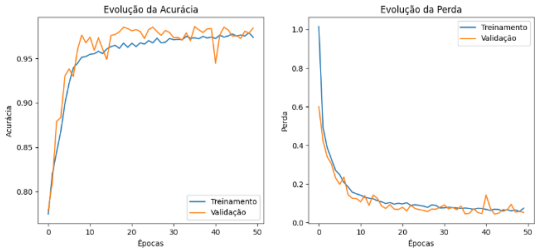

Figura 5: Gráficos de acurácia e perda

4 Considerações Finais

Considera-se que a rede neural desenvolvida teve êxito na classificação de amostras a partir da base disponibilizada segundo as classes descritas na normativa CONAMA 357/05. Para aumentar a acurácia do modelo um aumento de épocas ou a alteração de outros parâmetros da rede não resultou em uma melhora significativa. Desta forma para obter um modelo com uma maior acurácia, é necessária uma base com mais dados para aumentar os dados de treinamento para a rede neural.

REFERÊNCIAS

[1] Almeida, Anderson Francisco de Sousa; Veras, Adonney Allan de Oliveira; Merlin, Bruno; Santos, Adam; Amaris, Marcos. Predição temporal de parâmetros da qualidade da água usando redes neurais profundas. Brazilian Journal of Development, Curitiba, v. 7, n. 11, p. 107662-107678, nov. 2021. DOI: 10.34117/bjdv7n11-410. Disponível em: https://ojs.brazilianjournals.com.br/ojs/index.php/BRJD/article/view/40051. Acesso em: 20 nov. 2024.

[2] Rocha, Maria de Jesus Delmiro; SOUZA FILHO, Francisco de Assis. Aplicação de redes neurais para a classificação e avaliação do grau de degradação da qualidade da água de reservatórios rurais no semiárido brasileiro. In: XXIV Simpósio Brasileiro de Recursos Hdricos, 2021, Brasil. Anais [...]. ISSN 2318-0358. Disponível em: https://anais.abrhidro.org.br/job.php?Job=12851. Acesso em: 25 nov. 2024.

[3] Brasil. Resolução CONAMA nº 357, de 17 de março de 2005. Disponível em: https://www.icmbio.gov.br/cepsul/images/stories/legislacao/Resolucao/2005/res_conama_357_2005_classificacao_corpos_agua_rtfcda_altrd_res_393_2007_397_2008_410_2009_430_2011.pdf. Acesso em: 01 dez. 2024.

[4] SILVA, C. V. *Uso de Redes Neurais Artificiais para Análise Multitemporal da Dinâmica do Uso e Cobertura da Terra em Bacias Hidrográficas*. 2022.

[5] PACHECO, C.; PEREIRA, N. *Deep Learning: Conceitos e Utilização nas Diversas Áreas do Conhecimento*. Revista Ada Lovelace, v. 2, p. 34-49, 2018.

[6] Prodanov, Cleber Cristiano; De Freitas, Ernani Cesar. Metodologia do trabalho científico: métodos e técnicas da pesquisa e do trabalho acadêmico-2ª Edição. Editora Feevale, 2013.

[7] Neves, José Luis. Pesquisa qualitativa: características, usos e possibilidades. Caderno de pesquisas em administração, São Paulo, v. 1, n. 3, p. 1-5, 1996.