Previsão de Qualidade de Água

Bruno Ferreira de Lima 112389
Eng.Computação.
Fundação Hermínio Ometto
Araras, Brasil
brunoferreira48@alunos.fho.edu.br

Marina de Souza Pina Oliveira 111838
Eng.Computação.
Fundação Hermínio Ometto.
Araras, Brasil
marina.oliveira@alunos.fho.edu.br

***Abstract*—This work presents a classification model based on Support Vector Machines (SVM) to predict surface water quality using monitoring data from the Brazilian National Water Agency (ANA). The proposed approach uses physicochemical and biological parameters that compose the Water Quality Index (WQI), such as dissolved oxygen, turbidity, total phosphorus and thermotolerant coliforms. After preprocessing and normalizing the historical data series, different SVM kernels (linear and RBF) were trained and evaluated using accuracy, precision, recall, F1-score and confusion matrices. The RBF kernel achieved better performance, especially for intermediate and low-quality classes, indicating that non-linear decision boundaries are more suitable for this problem. The final model was integrated into a graphical user interface that allows users to input new samples and obtain the predicted water quality class in real time.**

***Index Terms*—Water quality, Support Vector Machines, ANA, WQI, environmental monitoring.**

I. Introdução

A água é um recurso vital para a vida, essencial não apenas para o consumo humano, mas também para a agricultura, a indústria e a manutenção dos ecossistemas. No entanto, a crescente poluição causada por atividades antrópicas, como o descarte inadequado de resíduos e o uso excessivo de produtos químicos, tem comprometido a qualidade da água em diversas regiões. Essa situação representa um sério risco à saúde das populações e ao equilíbrio ambiental, tornando urgente o desenvolvimento de métodos eficazes para monitorar e prever a qualidade da água.

Nesse cenário, o uso de técnicas avançadas de aprendizado de máquina, como as Máquinas de Vetores de Suporte (Support Vector Machines – SVM), surge como uma solução promissora. Esse tipo de modelo é capaz de analisar grandes volumes de dados complexos e identificar padrões que ajudam a antecipar mudanças na qualidade da água, possibilitando ações preventivas mais eficientes. Este trabalho tem como objetivo explorar o potencial do SVM na previsão da qualidade da água, utilizando como variáveis de entrada parâmetros que compõem o Índice de Qualidade da Água (IQA), como pH, turbidez, oxigênio dissolvido, coliformes termotolerantes, entre outros.

II. Trabalhos Relacionados

O trabalho fundamental para este projeto aborda a aplicação de diversas técnicas de aprendizado de máquina com o objetivo de analisar o índice de qualidade da água, considerando múltiplos indicadores de qualidade. A pesquisa avalia o desempenho de cada modelo e os compara utilizando variáveis de controle, como a taxa de acerto.

O principal objetivo deste estudo consiste no desenvolvimento de um modelo preditivo com base na técnica de Support Vector Machine (SVM). A escolha deste algoritmo foi fundamentada na análise comparativa do artigo de referência, que demonstrou a sua eficácia na classificação de águas em níveis de ”regular” e ”ótima”.

As variáveis selecionadas para a análise da qualidade da água são baseadas nos parâmetros que compõem o Índice de Qualidade das Águas (IQA), conforme definido pela Agência Nacional de Águas (ANA). Os indicadores utilizados incluem oxigênio dissolvido, coliformes termotolerantes, potencial hidrogeniônico (pH), temperatura da água, nitrogênio total, fósforo total, turbidez e sólidos totais, por meio desses índices o modelo SVM será capaz de classificar corretamente a qualidade da água.

III. Metodologia

Este trabalho foi desenvolvido com o objetivo de construir um modelo preditivo capaz de avaliar a qualidade da água com base em parâmetros físico-químicos e biológicos. O processo foi dividido em quatro etapas principais: coleta e preparação dos dados, tratamento e normalização, treinamento do modelo SVM, e avaliação dos resultados.

A. Coleta e preparação dos dados

Os dados utilizados para o treinamento do modelo foram obtidos de bases públicas de monitoramento ambiental disponibilizadas pela Agência Nacional de Águas (ANA) e complementadas por conjuntos de dados abertos referentes ao Índice de Qualidade da Água (IQA). As variáveis consideradas incluem os principais parâmetros de classificação: oxigênio dissolvido, coliformes termotolerantes, pH, temperatura da água, nitrogênio total, fósforo total, turbidez e sólidos totais. Cada registro representa uma amostra de um corpo d’água analisado em determinado local e data, acompanhado do rótulo de classificação de qualidade, categorizado em níveis como ótima, boa, regular e ruim.

B. Tratamento e normalização dos dados

Antes do treinamento, os dados passaram por uma etapa de limpeza, removendo registros incompletos e valores inconsistentes. Em seguida, todas as variáveis numéricas foram

normalizadas para o intervalo [0,1], garantindo que nenhuma característica dominasse o processo de aprendizado devido à diferença de escala. Também foi realizada uma análise exploratória de dados (EDA), com o objetivo de compreender a distribuição das variáveis, identificar correlações e verificar a relevância de cada parâmetro para o modelo.

C. Treinamento do modelo SVM

O algoritmo Support Vector Machine (SVM) foi escolhido pela sua eficiência na classificação de dados não lineares e pela capacidade de generalização em conjuntos de dados complexos. Para este trabalho, utilizou-se a implementação disponível na biblioteca scikit-learn da linguagem Python, com experimentos realizados nos modos linear e RBF (Radial Basis Function) para comparar o desempenho de diferentes funções de kernel. Os dados foram divididos em conjuntos de treinamento, validação e teste, em proporções aproximadas de 70%, 15% e 15%, respectivamente, utilizando estratificação por classe. Essa divisão permitiu ajustar os hiperparâmetros e comparar diferentes configurações de kernel na base de validação, preservando um conjunto independente para avaliação final.

D. Avaliação e métricas de desempenho A etapa final consistiu na avaliação do modelo por meio de métricas amplamente utilizadas em problemas de classificação, tais como acurácia global, precisão, recall e F1-score por classe. Adicionalmente, foram geradas matrizes de confusão para os conjuntos de validação e teste, permitindo uma análise visual dos acertos e erros em cada categoria de qualidade da água.

IV. Resultados

Utilizando o conjunto de dados obtidos a partir dos indicadores de qualidade da água disponibilizados pela Agência Nacional de Águas e Saneamento Básico(ANA), foi gerado um dataset final contendo milhares de amostras rotuladas em classes de qualidade (ótima, boa, regular, ruim e péssima) e compostas pelos parâmetros físico-químicos e biologicos que são utilizados na analise de qualidade.

Após a coleta dos dados base e definição do dataset, foi realizado o tratamento e normalização dos dados, nesta etapa o algoritmo remove dados desnecessários e parametriza valores antes de serem utilizados no modelo SVM, em seguida separa o conjunto de dados em duas partes treinamento e avaliação, sendo os dados de avaliação dividos entre teste e validação.

Para a etapa de treinamento do modelo foram utilizados dois classificadores SVM, um com kernel linear e outro com kernel RBF, ambos treinados sobre o mesmo conjunto de dados. O desempenho de cada classificador foi avaliado por meio da acurácia global, das métricas por classe (precisão, recall e F1-score) e das matrizes de confusão geradas automaticamente pelo código.

A partir desta matriz figura 1, é possivel observar um bom desempenho na identificação da classe boa, onde está concentrada a maioria dos acertos. Entretanto, outras amostras das classes vizinhas, como ótima e regular, são classificadas como boa, indicando uma dificuldade em definir fronteiras de decisão entre as categorias o mesmo comportamento se repete para as classes ruim e péssima. a figura 2 exemplifica esta

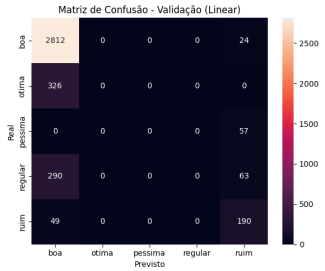

Fig. 1: Matriz de Confusão Kernel Linear

distribuição, onde a classe boa apresenta valores elevados de precisão, recall e F1-score, enquanto a classe “ruim” apresenta métricas mais modestas e as demais classes praticamente não são reconhecidas.

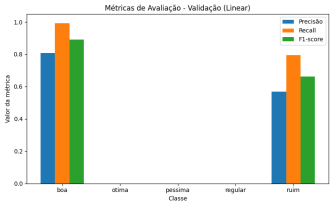

Fig. 2: Metricas Validação Kernel Linear

Quando aplicado o kernel RBF, o comportamento muda de forma significativa. como mostrado na matriz de confusão na figura 3, é possível verificar um aumento na quantidade de acertos nas outras classes de dados, especialmente péssima, regular e ruim, que passam a ser reconhecidas com maior frequência do que no modelo anterior.

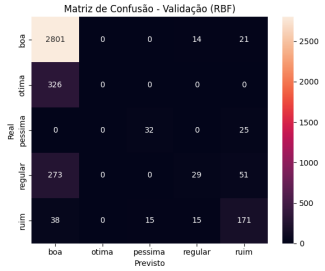

Fig. 3: Matriz de confusão kernel RBf

As métricas de precisão reforçam esse resultado. Embora a classe boa continue com valores elevados, há um ganho de precisão na definição das outras classes figura 4, que passam a apresentar F1-scores mais equilibrados em comparação ao modelo linear.

A partir desta comparação entre o modelo com kernel linear e o kernel RBF, o RBF foi definido como mais adequado para a aplicação. Sendo assim, foi submetido a avaliação no conjunto de teste, composto por amostras não utilizadas no treinamento ou validação figura 5 e figura 6.

Após avaliação pelo conjunto de teste, o modelo apresentou uma taxa de acerto coerente com a observada nas etapas de

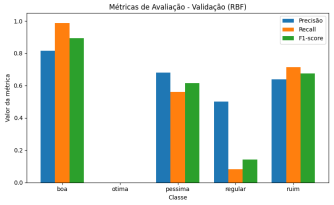

Fig. 4: Metricas de Validação RBF

treinamento e validação, com confusões concentradas principalmente entre as classes adjacentes em termos de qualidade. Este comportamento assemelha-se ao esperado em problemas reais de qualidade da água, onde a transição entre as categorias de classificação ocorre de maneira gradual e depende de pequenas variações nos parâmetros monitorados.

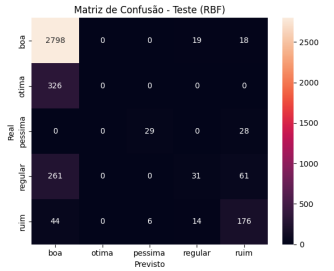

Fig. 5: Matriz de Confusão Teste RBF

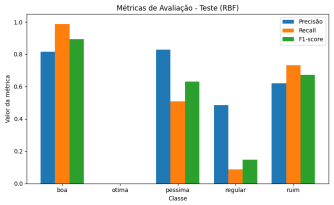

Fig. 6: Metricas de Validação Teste RBF

De maneira geral, os resultados indicam que o modelo utilizado consegue atender de maneira satisfatória a análise dos parâmetros para classificação da qualidade de água e fornecer previsões consistentes para cada categoria de qualidade, o desempenho observado sugere que a abordagem é promissora e pode ser aprimorada com o ajuste de hiperparâmetros e balanceamento de classes.

V. Conclusão

A partir do desenvolvimento deste trabalho utilizando Máquinas de Vetores de Suporte(SVM), aplicado a um conjunto de dados com milhares de amostras, contendo parâmetros físico-químicos e biológicos associados ao Índice de Qualidade de Água (IQA), foi possível treinar e avaliar modelos capazes de classificar corpos hídricos em diferentes faixas de qualidade.

Os experimentos realizados após o treinamento utilizando diferentes kernels, linear e RBF, mostraram um desempenho superior do RBF em relação ao linear, especialmente na destinção das classes nas faixas intemediárias. Ainda que persistam erros em fronteiras de decisão próximas os resultados demonstram que o modelo apresenta potencial para uso em cenários práticos.

O modelo enfrentou algumas limitações devido a qualidade dos dados disponiveis para análise, os dados utilizados são oriundos da ANA desta maneiras alguns índices estão disponiveis em quantidades diferentes, o que influencia na capacidade do modelo.

O trabalho cumpriu o objetivo de implementar um modelo SVM para classificação da qualidade da água baseado em dados reais, bem como demonstrar seu desempenho para uma aplicação prática, podendo ser expandindo para um novo trabalho onde está análise seja feita de modo autônomo utilizando sensores para coleta de dados e utilizando-os dentro do algoritmo para fazer análise constante de um determinado corpo de água.

References

[1] NASCIMENTO, Mário Elias Carvalho do. Previsão da qualidade da água utilizando técnicas de aprendizado de máquina. 2020. Trabalho de Conclusão de Curso (Graduação em Engenharia Ambiental) – Universidade Federal do Paraná, Curitiba, 2020. Disponível em: https://acervodigital.ufpr.br/xmlui/bitstream/handle/1884/89304/R%20-%20T%20-%20MARIO%20ELIAS%20CARVALHO%20DO%20NASCIMENTO.pdf?sequence=1. Acesso em: 20 set. 2025.

[2] AGÊNCIA NACIONAL DE ÁGUAS (ANA). Índice de Qualidade das Águas (IQA). Brasília, [2024?]. Disponível em: https://www.ana.gov.br/portalpnqa/indicadores-indice-aguas.aspx. Acesso em: 29 set. 2025.

[3] AGÊNCIA NACIONAL DE ÁGUAS E SANEAMENTO BÁSICO (Brasil). Indicadores de Qualidade da Água – IQA – série histórica – até 2021. Brasília, 2021. Conjunto de dados em formato aberto, com valores anuais de IQA para estações de monitoramento em corpos hídricos brasileiros. Disponível em: https://dadosabertos.ana.gov.br/maps/7a278de90bd14330ab014c9b5db350e0_17/about. Acesso em: 29 set. 2025.

[4] AGÊNCIA NACIONAL DE ÁGUAS E SANEAMENTO BÁSICO (Brasil). Indicadores de Qualidade da Água – Fósforo Total – série histórica – até 2021. Brasília, 2021. Conjunto de dados em formato digital. Disponível em: https://dadosabertos.ana.gov.br/datasets/0419dd6718cb4be0a331c7589c57ea2b_5/about. Acesso em: 29 set. 2025.

[5] AGÊNCIA NACIONAL DE ÁGUAS E SANEAMENTO BÁSICO (Brasil). Indicadores de Qualidade da Água – Turbidez – série histórica – até 2021. Brasília, 2021. Conjunto de dados em formato digital. Disponível em: https://dadosabertos.ana.gov.br/datasets/97e46167e18c4fb0bda9dd5f8ed7783b_8/explore. Acesso em: 29 set. 2025.

[6] AGÊNCIA NACIONAL DE ÁGUAS E SANEAMENTO BÁSICO (Brasil). Indicadores de Qualidade da Água – DBO – série histórica – até 2021. Brasília, 2021. Conjunto de dados em formato digital. Disponível em: https://dadosabertos.ana.gov.br/datasets/d82c795398754609b0a8b4a550ef6c57_14/about. Acesso em: 29 set. 2025.