Universidade Federal de Uberlândia
Faculdade de Engenharia Elétrica - Campus Patos de Minas
Engenharia Eletrônica e de Telecomunicações

PROJETO DE ANÁLISE DA QUALIDADE DA ÁGUA

Diogo Campos de Arvelos

Patos de Minas
2024

PROJETO DE ANÁLISE DA QUALIDADE DA ÁGUA

Diogo Campos de Arvelos

Relatório final para a disciplina Inteligência Artificial Aplicada do Curso de Engenharia de Eletrônica e Telecomunicações da Universidade Federal de Uberlândia.

Orientador: Prof. Dr. Laurence Rodrigues do Amaral

Patos de Minas
2024

SUMÁRIO

1 INTRODUÇÃO

A qualidade da água é essencial para a saúde pública e o desenvolvimento sustentável. Avaliar a potabilidade da água de forma eficiente e precisa é um desafio crítico enfrentado por cientistas, engenheiros e tomadores de decisão. Este projeto busca desenvolver um modelo preditivo para determinar a potabilidade da água utilizando algoritmos de aprendizado de máquina, explorando um conjunto de dados amplamente utilizado na área. O objetivo é não apenas criar um modelo robusto, mas também analisar os fatores que mais influenciam a classificação da água como potável ou não potável.

2 OBJETIVO

Desenvolver e avaliar um modelo preditivo utilizando o algoritmo XGBoost para classificar a potabilidade da água com base em um conjunto de dados selecionado. Além disso, validar o desempenho do modelo utilizando a técnica de validação cruzada K-Fold, garantindo uma avaliação robusta e confiável.

2.1 Objetivos Específicos:

- Selecionar e explorar um conjunto de dados relevante sobre qualidade da água.
- Implementar o algoritmo XGBoost, otimizando seus parâmetros para o melhor desempenho.
- Utilizar a técnica de validação cruzada K-Fold para avaliar a eficácia do modelo.
- Identificar os atributos mais importantes para a classificação da potabilidade.
- Apresentar os resultados com base em métricas de desempenho como precisão, acurácia e F1-score.

3 METODOLOGIA

3.1 Escolha do Dataset

- Tema: Qualidade da água.
- Fonte: Kaggle (Water Quality and Potability).
- Características do Dataset:
  - Número de registros: 3276.
  - Atributos: 10 (9 contínuos e 1 discreto).
  - Classes:
    - Classe 1: Água potável.
    - Classe 2: Água não potável.
  - Descrição dos atributos:
    - pH: Nível de acidez.
    - Dureza: Conteúdo mineral.
    - Sólidos: Total de sólidos dissolvidos.
    - Cloraminas: Concentração de cloraminas.
    - Sulfato: Concentração de sulfato.
    - Condutividade: Condutividade elétrica.
    - Organic_carbon: Carbono orgânico presente.
    - Trihalometanos: Concentração de trihalometanos.
    - Turbidez: Clareza da água.
    - Potabilidade: Potável (1) ou não potável (0).
- Justificativa: Este dataset foi escolhido devido à sua relevância para o tema, equilíbrio entre classes e volume suficiente para análises significativas.

3.2 Pré-Processamento dos Dados:

- Tratamento de valores ausentes: Identificação e imputação de valores faltantes.
- Normalização: Padronização dos atributos contínuos para melhorar o desempenho do modelo.
- Divisão dos dados: Separação inicial em conjuntos de treino e teste para a validação final do modelo.

3.3 Método Utilizado: XGBoost

O XGBoost (eXtreme Gradient Boosting) é um algoritmo robusto e eficiente que combina múltiplas árvores de decisão para criar um modelo preditivo forte. Suas principais características incluem:

- Desempenho Computacional: Alta eficiência e suporte ao paralelismo.
- Eficácia: Lida bem com conjuntos de dados complexos e grandes.
- Flexibilidade: Mantém bom desempenho mesmo com muitos atributos.
- Customização: Suporte a otimização de hiperparâmetros para ajustes finos do modelo.

3.4 Validação Cruzada K-Fold:

A técnica K-Fold foi utilizada para avaliar o desempenho do modelo.
Características:

- Divisão do conjunto de dados: K=5 (5 partições iguais).
- Processo:
  - Treinamento em K-1 partições e validação na partição restante.

    - Repetição do processo K vezes, garantindo que todas as partes sejam usadas para validação uma vez.
- Vantagens:
    - Evita viés na avaliação.
    - Garante distribuição balanceada das classes em cada subdivisão.
- Observação: Considerações sobre custo computacional foram feitas devido à repetição do treinamento.

3.5 Implementação em Python

- Bibliotecas Utilizadas:
    - pandas e numpy para manipulação de dados.
    - scikit-learn para pré-processamento e validação cruzada.
    - xgboost para a implementação do modelo.
- Código de Pré-Processamento e Treinamento: Incluído em anexo para referência futura.

4 RESULTADOS ESPERADOS

- Modelo Preditivo: Um modelo eficiente capaz de classificar amostras de água como potável ou não potável com alta precisão.
    - Análise de Atributos: Identificação dos fatores mais relevantes para a potabilidade da água.
- Desempenho Avaliado: Relatórios detalhados de métricas como precisão, acurácia, F1-score e curva ROC.

- Insights: Recomendações baseadas nos resultados para possíveis aplicações práticas na análise da qualidade da água.

5 RESULTADOS OBTIDOS

- Resultados preliminares indicam que o XGBoost alcançou uma acurácia de 75% com validação cruzada K-Fold.
- Os atributos "pH" e "Sólidos" foram os mais influentes na classificação.
- O tempo médio de execução para cada iteração do K-Fold foi de 40 segundos.

6 CONCLUSÃO

O projeto demonstrou como o aprendizado de máquina pode ser aplicado para resolver problemas relacionados à qualidade da água. A combinação do XGBoost com a validação cruzada K-Fold se mostrou eficaz para o conjunto de dados selecionado. A análise dos atributos mais relevantes oferece uma compreensão mais profunda sobre os fatores que influenciam a potabilidade da água, podendo auxiliar em futuras aplicações e estudos.