CHAPTER I
INTRODUCTION

1.1 Background

Water is an indispensable resource that supports human survival by fulfilling domestic consumption needs, maintaining sanitation, and enabling various economic activities [1]. However, the global availability of clean water remains a serious challenge. According to the World Health Organization (WHO), more than two billion individuals live in areas where freshwater resources are under significant pressure, and this challenge is expected to intensify as climate change continues and the global population grows [2]. In addition, around 1.7 billion people continue to rely on drinking water supplies affected by fecal contamination, thereby increasing the likelihood of waterborne health problems on a global scale [3]. These findings demonstrate that equitable provision of potable water continues to be difficult to achieve worldwide, particularly across developing countries where water treatment infrastructure is still inadequate [4]. Inadequate water quality contributes to the transmission of multiple infectious diseases, including cholera, diarrhea, dysentery, fever, polio, and typhoid. The WHO estimates that microbiologically unsafe drinking water is responsible for approximately 505,000 diarrhea-related deaths annually worldwide [5]. Beyond its adverse health consequences, unsafe drinking water also imposes substantial social and economic burdens, with the greatest impact experienced by populations that are more vulnerable, particularly children and economically disadvantaged communities. Consequently, systematic monitoring and periodic evaluation of water quality constitute key measures for supporting advance the implementation of Sustainable Development Goal (SDG) 6, whose objective is to achieve universal access to safe water through its sustainable availability and management [2].

According to the WHO, drinking water is considered safe and acceptable if it falls within reasonable quality parameter ranges. These include a pH of approximately 6.5-8.5, total dissolved solids (TDS/Solids) generally less than 600 mg/L (becoming unpalatable above 1000 mg/L), moderate and non-excessive hardness (approximately below 200 mg/L to prevent excessive scaling), an ideal sulfate level not exceeding 500 mg/L, chloramine levels used as a residual disinfectant ranging from 0.5-2 mg/L and not exceeding the health guideline value of 3 mg/L, total trihalomethanes (THMs) in treated water generally below 100 μg/L, and low turbidity with a value of approximately 0.5 NTU prior to disinfection and an average of 0.2 NTU or less, ensuring it appears visually clear and acceptable to consumers [6]. Along with technological advancements, conventional water quality testing methods, which are manual and require laboratory processes, are considered slow, expensive, and unable to provide real-time information [7]. Consequently, advanced technologies, particularly machine learning techniques, are increasingly employed to assess and predict water quality more rapidly and efficiently, thereby supporting timely interventions to protect public health and ecosystems [8]. Intelligent systems based on artificial intelligence employing a machine learning approach are capable of autonomously learning patterns from water quality data to provide more accurate, reliable, and efficient identification of drinking water quality [9].

A study conducted by Özsezer and Mermer reported that the Extreme Gradient Boosting (XGBoost) algorithm outperformed several competing methods in predicting drinking water quality. Their model was developed using the "Water Quality & Potability" dataset, which consists of nine physical and chemical water quality attributes, including pH, Hardness, Solids, and Sulfate. The comparative evaluation involved six alternative classification algorithms, with XGBoost obtaining the best predictive performance by reaching an accuracy of 0.79 and an AUC value of 1.00. These results demonstrate its outstanding ability to distinguish potable water samples from non-potable ones [10]. These findings suggest that XGBoost is capable of modeling complex relationships within heterogeneous and nonlinear water quality data, thereby producing reliable classification results. From a methodological perspective, XGBoost belongs to the family of gradient boosting ensemble algorithms, in which multiple decision trees are combined sequentially to construct a highly predictive model. It has gained widespread recognition because it provides both computational efficiency and strong predictive capability across a broad range of classification and regression tasks [11]. Owing to these characteristics, XGBoost has become a widely adopted approach in predictive modeling, including water quality assessment and classification, where accurate predictions, model stability, and computational efficiency are essential requirements.

Although XGBoost has consistently demonstrated strong predictive performance in water quality studies, its effectiveness is strongly influenced by the configuration of its hyperparameters. In particular, settings including learning rate, max_depth, and n_estimators play a crucial role in determining the model's ability to generalize to unseen data [12]. To address this issue, several automated hyperparameter optimization techniques have been introduced. One widely adopted approach is Optuna, whose optimization strategy relies on the Tree-structured Parzen Estimator (TPE) to adaptively search for promising hyperparameter combinations within the defined search space. [13]. By utilizing information obtained from previous optimization trials, Optuna adaptively identifies promising hyperparameter combinations, enabling a faster search process while simultaneously improving computational efficiency [12]. Recent investigations have provided substantial evidence that integrating Optuna into the optimization process can significantly improve the predictive performance of XGBoost models. For example, Rai and Kumar found that applying Optuna for hyperparameter tuning enabled XGBoost to outperform optimization approaches based on Genetic Algorithm (GA) and Particle Swarm Optimization (PSO). Their results showed superior predictive accuracy together with reduced training time, which was primarily attributed to Optuna's adaptive sampling and pruning mechanisms [14]. Consistent evidence was presented by Liu et al., who reported that the combination of Optuna and XGBoost generated highly accurate and stable predictions for estimating heavy metal concentrations in soil, including datasets characterized by limited sample sizes and nonlinear relationships [15]. Overall, incorporating Optuna into the XGBoost modeling framework enhances predictive accuracy while improving computational efficiency. This integration also increases the robustness and reliability of machine learning-based classification systems designed for environmental applications, particularly drinking water quality assessment.

Based on the phenomena described above, this study aims to optimize the XGBoost algorithm using Optuna for drinking water quality classification. It is anticipated that this research will contribute to the advancement of scientific knowledge, particularly in demonstrating the accuracy improvement of the Optuna-optimized XGBoost algorithm in classifying drinking water quality.

1.2 Problem Formulation

Based on the background described above, The research problems addressed in this study are formulated into the following research questions:

1. How does the XGBoost algorithm perform in classifying drinking water quality before hyperparameter optimization?
2. How can the XGBoost algorithm be optimized using Optuna for drinking water quality classification?
3. How do the evaluation results of the XGBoost algorithm compare before and after optimization using Optuna for drinking water quality classification?
4. How can the performance comparison of the XGBoost algorithm, before and after optimization using Optuna, be visualized in an interactive web dashboard based on accuracy, precision, recall, and F1-score evaluation metrics?

1.3 Research Objectives

In alignment with the research questions outlined above, the specific objectives to be achieved in this study are:

1. To apply the XGBoost algorithm for drinking water quality classification.
2. To apply Optuna hyperparameter optimization to the XGBoost algorithm for drinking water quality classification.
3. To compare the evaluation results of the XGBoost algorithm optimized with Optuna for drinking water quality classification.
4. To build an interactive web dashboard capable of displaying the performance comparison results of the XGBoost algorithm before and after hyperparameter optimization using Optuna, based on accuracy, precision, recall, and F1-score.

1.4 Research Benefits

The expected benefits of this research are as follows:

1. To provide an understanding of the application of the XGBoost algorithm in the drinking water quality classification process and its performance prior to hyperparameter optimization.
2. To provide a reference regarding the application of the Optuna hyperparameter optimization method to improve the performance of the XGBoost algorithm in classifying drinking water quality.
3. To provide information regarding the comparison of the XGBoost algorithm's evaluation results before and after optimization using Optuna.
4. To produce an interactive web dashboard that assists in visualizing the performance comparison results of the XGBoost algorithm before and after optimization, making the information easier to understand and analyze.

1.5 Research Limitations

To ensure that the discussion remains focused and the research outcomes are interpreted within the appropriate context, the author defines the scope and limitations based on the methods and environments utilized in the experimental setup:

1. The dataset used in this study is "Water Quality & Potability," obtained from Kaggle, comprising 3,276 data points, and is freely accessible via the link https://www.kaggle.com/datasets/tusharpaul2001/crop-water-management-and-quality.
2. The dataset consists of 9 input features (pH, Hardness, Solids, Chloramines, Sulfate, Conductivity, Organic Carbon, Trihalomethanes, and Turbidity) and 1 target output (potability).
3. This research is limited to data classification regarding drinking water suitability, specifically its potential to be drinkable (potable) or undrinkable (not potable).
4. The approach utilized in this research is the XGBoost algorithm with hyperparameter optimization using Optuna.
5. The algorithm is implemented in Python utilizing relevant libraries.

6. Model performance evaluation is conducted using a Confusion Matrix with accuracy, precision, recall, and F1-Score metrics to measure the classification performance of each tested model.
7. The web dashboard visualization is solely focused on the drinking water quality classification results and model evaluation, using a Streamlit-based application deployed on the Streamlit Community Cloud platform.