IJSRA

International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.

International Journal of Science and Research Archive

Research Journal Archive, INDIA

(Research Article)

Predicting regulatory violations in public drinking water systems: A data-driven approach

Abdul-Waliyyu Bello [1, *], Awele Okolie [2], Jacinta Izundu [3] and Anastesia Izundu [4]

[1] *Department of Mathematics and Statistics, Austin Peay State University, Tennessee, USA.*
[2] *Department of Data Science, Wentworth Institute of Technology, Massachusetts, USA.*
[3] *Department of Cybersecurity Management, University of Illinois at Springfield, Illinois, USA.*
[4] *Department of Public Health, University of Illinois at Springfield, Illinois, USA.*

International Journal of Science and Research Archive, 2025, 16(03), 1039-1055

Publication history: Received on 16 August 2025; revised on 22 September 2025; accepted on 24 September 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.16.3.2665

Abstract

Enforcement of the Safe Drinking Water Act helps in protecting public health, yet agencies tasked with enforcement are besieged with numerous violations. In this research, a supervised machine learning approach was employed to predict the final regulatory action ('Resolved' or 'Archived') of 7,268 past drinking water violations within the U.S. Environmental Protection Agency's SDWIS database. Logistic Regression, Random Forest, and Gradient Boosting models were tuned using hyperparameter tuning and then trained on the data. The human-tuned Gradient Boosting model performed the best in prediction, with 82.3% accuracy and an F1-Score of 0.87 on the test set. Feature importance analysis revealed that the violation of the specific regulatory rule was the key predictor, followed by the quantitative extent of the violation (exceedance ratio) and the contaminant type. Findings reveal that the fate of drinking water violations is highly predictable based on their original characteristics. The research offers a proof-of-concept for predictive analytics as an environmental governance tool through which regulatory agencies may filter and prioritize high-priority violations to be addressed on a targeted basis.

Keywords: Drinking Water Quality; Machine Learning; Regulatory Compliance; Predictive Modeling; Gradient Boosting; Safe Drinking Water Act (Sewa); Environmental Governance

1. Introduction

Safe and reliable drinking water availability is central to human well-being and health, the foundation of disease prevention, economic productivity, and social stability (Pennino et al., 2020). Public water systems provide drinking water to nearly 90 percent of the population of the United States, emphasizing the critically significant role they serve in safeguarding communities' health (Michaelsen et al., 2020). Contaminated or poorly controlled water supplies have been linked with gastrointestinal disease epidemics, children's developmental issues, and elevated risks of long-term illness. High-profile outbreaks, such as the Flint water crisis, have also demonstrated the social and political consequences of water quality control failures (Allaire et al., 2018). Enforcing consistent compliance with drinking water standards is not simply an environmental matter but a pressing public health issue. This being the case, the U.S. Congress enacted the Safe Drinking Water Act (SDWA), which remains the cornerstone of regulatory authority over public water systems nationwide (Ye et al., 2024).

The Safe Drinking Water Act (SDWA) of 1974 and subsequently amended in 1986 and 1996 provides the institutional and legal foundation for the regulation of drinking water quality in the United States (Elbaite and Beeson, 2021). The U.S. Environmental Protection Agency (EPA) establishes Maximum Contaminant Levels (MCLs) and treatment

[*] Corresponding author: Abdul-Waliyyu Bello

techniques for more than 90 contaminants ranging from microbial pathogens to chemical contaminants under the Act (Pennino et al., 2020). The SDWA also distinguishes between health-based violations, in which levels of contaminants exceed safety levels, and monitoring or reporting violations, in which water systems fail to conduct or report required tests. Significantly, the SDWA delegates enforcement power to states and tribal authorities but reserves oversight and the right to act on noncompliance to the EPA (Statman-Weil et al., 2020). This regulatory structure has been instrumental in maximizing safety in drinking water, but after decades of progress, many water systems continue to struggle with outright compliance, and violations remain a persistent problem nationwide.

Even with the safeguards provided by the Safe Drinking Water Act, rule noncompliance is a chronic problem across the United States. EPA compliance reports indicate that about 28 percent of public water systems have at least one violation annually, and some 4 percent of public water systems recorded violations with possible implications for consumer health (Elbakidze & Beeson, 2021). Violations expose millions to potential contaminants that range from nitrates and arsenic to lead and disinfectant byproducts. Violations are not merely reported by small and rural communities; large systems also report violations. But violations mostly occur in systems located in rural and smaller communities (McDonald and Jones, 2018). The rate of these infractions reflects both the size of the national drinking water system and the demand to consistently use stringent standards in all areas, types of systems, and socio-economic levels. The occurrence of the infractions made it clear how these matters vary geographically, by system type, and by community characteristics (Barcia et al., 2024).

United States' drinking water violation trends are not the same but very different based on region and system type. Studies have identified rural and poor areas have higher rates of both health-based and monitoring violations compared to their richer, urban counterparts. As identified by Allaire et al. (2018), small community water systems with fewer than 3,300 people are disproportionately vulnerable to violation due to limited technical and financial ability. Spatial inequalities also appear, such as the Southwest region facing repeated challenges from naturally occurring contamination of arsenic and nitrates, while agri-states in the Midwest face high concentrations of nitrate related to fertilizer runoff (Im et al., 2022). Moreover, financially struggling counties in the Deep South and Appalachia consistently exhibit strong monitoring noncompliance. These varying trends justify examining contextual and structural factors that shape regulatory results across the nation (Hu et al., 2023).

The ability of data analysis and predictive modeling offers fresh prospects to consider how best to react to chronic abuses in public drinking water systems (Pennino et al., 2020). Access to administrative data in EPA's Safe Drinking Water Information System (SDWIS), as well as to sewage, environmental, and other infrastructure data, has enabled researchers to apply machine learning and statistical methods to find patterns of noncompliance (Al-Adhaileh & Alsaade, 2021). Predictive techniques can determine empirically which systems are most likely to suffer from future violations so that regulators can act on and intervene without having to inspect so many systems. Violations based on health have been predicted at moderate to high accuracies in research, and machine learning models such as logistic regression, random forests, and neural networks are typical tools used (Wang, et al., 2024). Although these machine learning techniques are intellectual exercises, they also can be applied as regulatory and policy instruments for optimal resource distribution to the extent possible, given the need for special attention to vulnerable communities and ultimately enabling enforcement with the SDWA (Dobbin & Fencl, 2021).

Given the insights identified, it becomes necessary to adopt a data approach to predicting regulatory violations in United State public drinking water systems. The integration of compliance records with system-level and contextual factors would enhance identifying predictors of non-compliance. Hence, the ability to assess the implication to regulatory enforcement. This would enable the investigation contribute to an evidence-based tools to support proactive monitoring of violation in order to improve and protect public health.

2. Literature Review

2.1. Drinking Water Quality and Regulatory Compliance

In U.S. regulatory settings, safety is pragmatically interpreted: public drinking water systems must meet standards of the National Primary Drinking Water Regulations (NPDWRs) issued under the Safe Drinking Water Act (SDWA), which dictate enforceable levels (maximum contaminant levels or MCLs), treatment practices, and monitoring for specific contaminants (Wein Meyer et al., 2017). A risk-and-standards conceptualization that the U.S. EPA has created means through which water is "safe" to the point that contaminants are not there beyond health-based levels specified by law for human health protection. Because this is regulatory safety, "safety" takes on partial legal/administrative significance added to that which is also monitored and regulated (Levin, et al., 2023).

Contaminants of concern are commonly categorized into the microbial, chemical (inorganic and organic), physical, radiological, and emerging or newly recognized substances categories. Microbial hazards like bacteria (e.g., E. coli), viruses, and protozoa (e.g., Giardia, Cryptosporidium) result in acute illness and are addressed under treatment and disinfection rules (Zulkifli et al., 2017). Chemical contaminants such as arsenic, nitrate, lead and copper, disinfection by-products, and, more recently, per- and polyfluoroalkyl substances (PFAS) are also regulated. Radiological contaminants and less common toxins are also regulated. Importantly, regulatory lists do not cover all potential risks (Coxon and Eaton, 2023). The Contaminant Candidate List discussed many substances that are either unregulated compounds or recently detected ones, and although we are studying them, we do not yet know everything we should about their health effects or occurrence patterns (Wen, et al., 2020).

As asserted by Dobbin and Fencl (2021), violation of SDWA requirements can be grouped into two broad categories with different implications: (1) health-based violation (MCL violation, treatment-technique violation, Maximum Residual Disinfectant Level violation) represents a direct violation of a health-protection standard and (2) monitoring and reporting violation (a lack of sampling on a specified frequency, reporting, or following procedures for analyses) can mask reality and thus indirectly but importantly threaten public health surveillance. The EPA and scholars emphasize that monitoring/reporting violations are of equal importance as measurable violations because monitoring/reporting violations create blind spots in surveillance. The influence of why compliance matters go beyond direct mortality and morbidity (Barcia et al., 2024).

Compliance lowers exposure to acute and chronic hazards, avoidance of outbreaks of waterborne disease and illness due to long-term exposures to toxicants (Elbaite and Beeson, 2021). Compliance also promotes public trust between regulators and utilities because timely notice, reporting honestly on outcomes, and access to honest data, are all legitimized, while violations are capable of undermining trust with its social/political repercussions because violation in trust has implications (Chen et al, 2024). Ultimately, public health and socio-political interests become entangled as regulatory compliance is technical requirement and a sign of governance.

2.2. The Safe Drinking Water Act (SDWA)

The Safe Drinking Water Act of 1974 created the federal statutory foundation for ensuring public drinking water is safe in the United States by mandating the U.S. Environmental Protection Agency (EPA) to set science-based national standards (Maximum Contaminant Levels, treatment procedures, and monitoring) for public water systems (McDonald & Jones, 2018). The legislation established a formal federal function for overseeing the safety of drinking water and reserved state discretion to implement in the states, territories, and tribes potentially acquiring "primacy" to administer drinking water standards within their jurisdictions (Allairee and Acquah, 2022). This initial configuration framed drinking water safety as a federal standards/state implementation public health program.

Amendments to the statute by Congress in 1986 and 1996 brought far-reaching modifications in both the substance of regulation as well as the institutional practice accompanying it (Weinmeyer et al., 2017). The amendments of 1986 expanded EPA's rulemaking authority for contaminants, requiring rulemaking on tens of new contaminants and imposed new requirements with regard to wellhead protection, and filtration/disinfection standards (Mueller and Gasteyer, 2021). Additionally, plumbing was prohibited from being constructed of anything but "lead free" materials. These changes extended the technical range of compliance under drinking water regulation. The 1996 amendments established further, structural changes, with risk-and-science-based standard setting taking precedence, requiring cost/benefit and peer-reviewed science for EPA standard setting (Pieper et al., 2019). It also established the Drinking Water State Revolving Fund (DWSRF) as a source for funding infrastructure construction. The implementation of a series of public-right-to-know measures (Consumer Confidence Reports), and more rigorous reporting requirements were part of the modifications. It looks as though they complicate the regulation, these amendments have also provided tools aimed at building capacity and transparency in the system (Elbakidze and Beeson, 2021).

The SDWA depends on state and tribal primacy for achievability but generates issues of its own. Although assigning enforcement to primacy agencies enables the technical expertise and administrative capacity of the community governments to be utilized, there is also a disparity: states vary in terms of regulatory stringency, inspections (frequency), inspection aggressiveness, and resources spent (Fowler & Birdsall, 2020). EPA retains its oversight and authority to withdraw primacy or to enforce, but federal oversight or enforcement has been nonexistent in practice and negligible. Tribal implementation of the SDWA has progressed into the TAS and primacy tracks but many tribal systems are limited in their capacity and resources available. This governance structure allows for compliance results to be contradictory simultaneously and in the same location (Tiemann, 2014).

The enforcement mechanisms available under the SDWA are administrative orders, civil penalties, and (rarely, as of yet) criminal penalties. Public disclosure and funding incentives are also used (Fu et al., 2020). The 1996 amendments also allowed the EPA to enhance enforcement authority, while requiring national compliance reports at regular (albeit unspecified) intervals, and the establishment of SDWIS has allowed centralized national data tracking for violations, monitoring data, and oversight (Weinmeyer et al., 2017). From a practical standpoint, monitoring and enforcement are constrained, insufficient technical capacity to report/comply in numerous (small) systems such as data quality/timeliness issues, and enforcement can be politically restricted or fiscally constrained (Johnson, 2021). So, analysts consider the SDWA a comprehensive and legally enforceable framework whose success relies heavily on administrative capacity, funding, and ongoing (improved) scientific information on contaminants (PFAS, new lead standards) that persist in challenging uniform protection.

2.3. Trends in Drinking Water Violations in the U. S.

Environmental Protection Agency compliance reports annually feature data indicating that thousands of public water systems annually violate, impacting millions of Americans. Allaire et al. (2018) estimated that from 1982 through 2015, roughly 9–45 million people were served annually by systems with a health-based violation, and small water systems were most prevalent. Health-related exceedances, including arsenic, nitrate, and disinfection byproducts, are of concern because health-based exceedances represent the most frequent single type of violation with greatest public health importance (Michaelsen et al., 2020). The differences observed between regions indicate how geography, and social factors lead to delays or development of patterns of violation.

Nitrate contamination in the Midwest has been caused by the intensity of agricultural land use and fertilizer and manure runoff into groundwater sources (Zendejas et al., 2022). Exceedances for arsenic are greater in Arizona and New Mexico due to geologic factors. Large urban systems are impacted by legacy lead infrastructure and disinfection byproduct levels. Finally, rural systems spend more on microbial contamination risks due to limited small innovations and treatment capacities for their systems as a whole (Yousefi & Douna, 2023). Accordingly, while types of violations can help develop compliance ideas, there are very different compliance mechanisms which exist between areas and social conditions (Allaire et al., 2018).

System size is a determining factor for compliance performance. Most of the problems occur in small systems because they have less technical capacity, less money, and part-time operators (Pieper et al., 2019). Large systems have operations on a large scale, fixed experts, and funding that facilitate a continuous process of monitoring and treatment. However, large systems are not exempt from these issues. Flint Michigan crisis demonstrates that governance and capital project built-in infrastructure failures lead to large-scale community exposure to lead, regardless of system size to serve the community (Wang, et al., 2024).

Long-term trends show that although changes to the law, including the 1996 risk-based reforms and Lead and Copper Rule amendments, led to increased regulation, regressive problem areas remain (Fu et al, 2020). Nitrates remain among the leading causes of exceedance in farm zones, microbial exceedances become more common in rural and small systems, and new contaminants (e.g., PFAS) appear as compliance challenges not currently being mandated under their regulation (Alzahrani & Tawfik, 2024). Even violations of monitoring and reporting protocols do not appear to be diminishing, which implies administrative and capacity degrades in the system. Therefore, these persistent variables provide a proxy with an issue, in keeping so as not to have exceedances of health-based standards and smoothing the monitoring and reporting so as to reduce under-detection of risks (Cade et al, 2024).

2.4. Determinants of Regulatory Violations

In the United States, the determinants of drinking water violations are not isolated, rather, they are intertwined in a systemic way of creating compounding susceptibilities, which is more challenging among less privileged communities.

2.4.1. System-Level Factors

System size is the most reliable predictor of compliance results. Research suggests that small water systems are disproportionately represented in violations because of inadequate technical capacity, financial instability, and employment of part-time or under-trained staff (Rubin, 2013; Allaire et al., 2018). In big utilities, economies of scale have some advantages; they can derive multiple sources of revenues along with possessing employees who are capable to a greater extent to meet monitoring and treatment requirements.

Ownership and governance structures can also impact compliance. Systems owned by the government vary in incentives from privately owned systems. Government-owned systems are accountable to the citizens and the government but can also be subject to budgetary constraints and political pressure when it comes to postponing upgrade programs for infrastructure. Privately owned systems are encouraged to invest in their system but are limited by cost-recovery mandates and regulations (Beecher, 2013).

Another key driver is the age of infrastructure. Aging distribution systems with old lead service lines, corroded pipes, or outmoded treatment plants are many times more likely to pipe-burst, leak, and/or fail. The Flint water crisis illustrated the interplay between postponed upkeep and policies of governance to create catastrophic results. The EPA estimates that over $470 billion of investment will be needed here up to 2035 to preserve and enhance water infrastructure, illustrating the enormity of this problem.

Technical ability and managerial ability also created a distinction between compliant and noncompliant systems. Utilities with certified operators, current monitoring technologies, and proactive asset management programs put themselves in a much better position to remain in compliance. Systems that do not have these abilities will tend to accrue monitoring/reporting infractions that conceal continually existing health risks.

2.4.2. Environmental Factors

The occurrence of contaminants is a direct function of local geology. Naturally occurring arsenic, for instance, occurs in groundwater in the Southwest, and so-called radionuclides may occur in some areas under certain bedrock conditions. These conditions lead to the reality that some systems threaten baseline compliance regardless of actual operations (Knobeloch et al., 2016).

Agricultural runoff is also one of the biggest drivers, and it implies the nitrate pollution, particularly in the Midwest and California Central Valley. The increase in the use of fertilizer, concentrated animal feeding operations (CAFOs), and irrigation also contribute to increased chances of leaching nitrate into the surface water and/or aquifers, which will cause public health risks such as methemoglobinemia that are very long-lived.

Industrial activities also have a role to play, especially where there is a history of mining, petrochemical facilities, or manufacturing. Heavy metals, volatile organic compounds, and PFAS all have historic and ongoing industrial discharge connections. Downgradient systems typically have a high treatment burden and risk of regulation breach.

Climate change adds an additional complicating element. Increased drought concentrated pollutants, and flooding would overload treatment plants and attract potential microbial and chemical hazards into treatment streams. These stressors add an additional compliance challenge, especially if the water system is small and/or inadequately equipped.

2.4.3. Socioeconomic and Demographic Factors

Socioeconomic context has a strong impact on patterns of compliance. Poor, low-income communities have no fiscal foundation upon which infrastructure investment can be secured or qualified operators hired. Compliance of systems serving these communities is generally poorer than that of their urban counterparts, as measured by failure rates for monitoring and treatment.

Rurality compounds this. Rural systems are smaller and more isolated, relying heavily on groundwater sources making both technical and financial compliance more difficult to achieve. Rural utilities also do not have access to federal/state funding streams since they have limited administrative capacity.

Racial composition and ethnicity have also been demonstrated to predict inequities in safe drinking water. Research indicates communities of color are more likely to be served by systems with repeat violations, consistent with broader patterns of environmental injustice (Pierce and Gonzalez, 2017). For example, in California's Central Valley, Latino communities reliably have higher nitrate levels than all other communities. Tribal systems of Native Americans frequently experience chronic compliance because they are underfunded and face complex levels of jurisdiction.

2.5. Health and Socioeconomic Impacts of Drinking Water Violations

Exceedance of drinking water standards has quantifiable health and economic impacts that can linger long after the violation has been over. The most immediate human health impact is human health consequences. Microbial contaminants acute exposure cause outbreaks of gastrointestinal disease, one of the leading causes of waterborne illness in the U.S. (Craun et al., 2010). Chemical pollutants, conversely, carry more chronic threats upon longer-term exposure, arsenic has been associated with bladder, lung, and skin cancers and nitrates that pollute water supplies have been associated with methemoglobinemia as well as potential endocrine disruption (Ward et al., 2018). Lead, perhaps the most politically charged contaminant in description since the Flint water crisis, causes irreversible neurological damage to children, which can slow intellectual development and academic functioning.

Health burdens are inequitably apportioned, with low-income households, rural residents, and communities of color being under an unequal burden of exposure to polluted water. For instance, studies in California's Central Valley rural areas, predominantly populated by Latino communities, and in Native American reservations and rural Appalachian municipalities, document cases of repeat violations with minimal, if any, follow-up (Balazs and Ray, 2014; Pierce and Gonzalez, 2017). Such differences in water quality reveal an environmental justice aspect to the issue. That is, regulatory non-compliance relative to water quality exacerbates otherwise troubled socioeconomic circumstances.

Economic effects of noncompliance affect families and communities as well. For families, contaminated water necessitates the purchase of expensive bottled water or estimating bottled water price using household filtration. This is an additional economic burden on families already operating under discretionary spending limits via budgetary limitations or reconciliations. To communities, healthcare costs from waterborne pathogens and chronic conditions are all heightened by noncompliance infractions. There is also lost productivity due to sickness-related absenteeism. The municipalities are also required to finance emergency repair of infrastructure, hire consultants to cope with emergencies and investigate remediation options, provide emergency health care, and at times manage legal settlements due to breaches, if not directly as a result of undertaking repairs or remediation. Water quality violations have triple costs: health costs, economic costs, and a general collapse of confidence in public institutions.

2.6. Predictive Analytics in Environmental and Regulatory Compliance

Predictive analytics allows authors to shift from a reporting strategy based in the past to a forward-looking prediction perspective. This allows researchers to anticipate patterns and forecast where risk is most likely to be. In water, this allows for a perspective beyond traditional compliance monitoring, and can help provide an evidence base to help understand vulnerabilities in the system better.

Within the environmental research domains, applications show the broad scope of predictive modeling applicability. In air quality studies, predictive modeling techniques with machine learning algorithms have been used together in team-based research to forecast concentrations of ozone and particulate matter concentration from meteorological and emissions inputs (Zheng et al., 2013). In energy system studies, predictive models have been applied to load forecasting, variability of renewable generation, and system optimization. Hydrology and water resource management try to encompass predictive tools for drought modeling, contaminant transport modeling, and surface-groundwater interaction. These examples reflect methodological maturity in predictive analytics in environmental science and thus provide a relevant context towards consideration for drinking water quality, testing its feasibility.

In studies on drinking water, statistical and machine learning approaches have been developed to predict the likelihood of violations. Logistic regression is a common statistical technique, commonly used to examine relationships between the type of a system and compliance outcomes (Allaire et al., 2018). More recently, researchers have turned attention to ensemble methods such as random forests and gradient boosting because they can better identify nonlinear patterns and higher-order interactions of predictor variables (Ma et al., 2020). Perhaps even more specifically the literature has explored neural networks and deep learning methods in multivariate large datasets in drinking water system research, though still not having proper interpretability.

Reports from studies that have compared different statistical and machine learning approaches have cited the trade-off between model complexity and results interpretability. More predictive ability is usually obtained through much more sophisticated machine learning algorithms, but less complex models have also been regarded as good models because of interpretability and transparency. Scientists increasingly apply a middle approach with hybrid methodologies to solve issues of modeling complexity vs. interpretability, suggesting importance variable measures did indirect clustering and utilizing model explanation, such as SHAP values, that provide outputs that policymakers find most useful when drawing conclusions from existing knowledge.

2.7. Theoretical Review

The evaluation of violations associated with drinking water, and their prediction, borrows from various interplaying theoretical paradigms employed to explain the occurrence of the violations, as well as compliance. The prominent theoretical paradigms include risk-based regulation, compliance theory, and data-driven governance.

Risk-based regulation focuses on directing regulatory attention at activities or systems of highest risk to health. For drinking water, this is at the point that violations need not all be equally damaging, for example, a missed due date on a monitoring report is probably less damaging than an arsenic or nitrate exceedance. The theory promotes the focus on the necessity to spot high-risk systems, say those offering cover to small rural areas of arable land, and resource allocation to them. Predictive modeling does the same in its application of past trends in relation to system characteristics and projected risk and tries to predict where risk is most likely to be seen where risk-based thinking as a predict risk-based analytical tool is applied.

Compliance theory discusses the reasons why regulated entities comply or fail to comply with standards. In compliance theory, the deterrence model that is considered a classic in compliance theory focuses on compliance as a function of enforcement severity and sanctions perceived and known by the actor for noncompliance. The normative model focuses on the role of legitimacy, trust, and organizational action culture in determining compliance. There are many other factors besides sanctions that influence compliance outcomes in drinking water systems. Some instances can be technical expertise, managerial ability, and public pressure. The impacts of resources available in small systems illustrate the fact that compliance is not a matter of will, but rather a matter of the capability to comply with the standards. This perspective justifies the inclusion of socioeconomic factors and system characteristics in the prediction of compliance.

Data-driven governance accounts for the way that information systems and analytic technologies are changing the practice of regulation. The evolution of SDWIS is an example of this transformation in regulatory practice. SDWIS enabled the collection of data and permitted regulated entities to advance from a simple descriptive report registered of superficial information to full predictive analytic systems of transparency and accountability. Once data is in SDWIS, data previously defined as records of past performance or compliance profile becomes a rich source of anticipatory or anticipatory state/regional regulation. Predictive modeling can then be situated within a larger theoretical trend toward evidence-based, data-driven governance.

2.8. Research Gaps

Despite advances in examining drinking water compliance, several gaps remain. Most prediction studies focus exclusively on system-level attributes, with fewer including environmental stressors or socioeconomic disparities, despite evidence of their importance. Third, methodological limitations persist, logistic regression dominates whereas newer machine learning methods are underused or insufficiently validated across diverse contexts. Finally, the predictive accuracy versus interpretability trade-off is seldom addressed explicitly, a key challenge in advancing both scientific understanding and real-world usefulness. New directions suggest comprehensive models that synthesize technical, environmental, and social determinants, along with increased application of explainable machine learning methods to model complexity without compromising interpretability.

3. Materials and Methods

This section describes data sources, pre-processing techniques, and analysis algorithms utilized to build regulatory violations predicting model for U.S. drinking water public systems. This methodology is made transparent and replicable to enhance the credibility of findings.

3.1. Research Design

This research utilizes a quantitative, non-experimental study design to create a predictive model of drinking water violation incidents. A non-experimental study design must be used because the research is utilizing available historical, observational data from the EPA's Safe Drinking Water Information System (SDWIS), and the variables in question cannot be controlled, such as the type of violation or system characteristics. The study is quantitative in nature as it is a statistical analysis of categorical and numeric data for the purpose of finding patterns and predictive relationships. The primary design strategy is a supervised machine learning strategy, specifically formulated as a multiclass classification problem. The supervised machine learning strategy was preferred over traditional explanatory statistical models for two reasons. Firstly, the study is particularly predictive in nature to anticipate regulatory breach which is aligned with the purpose of supervised machine learning. Secondly, the number of variables that affect regulatory compliance is large, and the interaction among the variables is complicated,

3.2. Data Source

The information used in this research was gathered from the Safe Drinking Water Information System (SDWIS), the central administrative database of the U.S. Environmental Protection Agency (EPA) that is utilized for tracking public water system compliance. SDWIS is the record source for all enforcement action against violation of the Safe Drinking Water Act (SDWA) and also a centralized national source of information on system characteristics, monitoring data, and enforcement action. Thus, it must be the default data source for U.S compliance research because of its detail, granularity, and extensive database that can facilitate data to be used for predictive model building. The dataset used in the study is for tracked violations over multi-year time horizon, which makes it possible to have significant historical context for determination of predictive patterns in compliance enforcement outcomes.

3.3. Population and Sampling

The focus of this study is directed at all the United States public water systems (PWSs) that are governed by the Safe Drinking Water Act. Out of the larger population described above, a particular purposive sample was taken to limit the analysis on a specific and recurring class of regulatory deficiency. The data sample consists of all the health-based violations which were reported for a specific subset of contaminants during the period from 1994–2016. After rigorous cleaning efforts to remove cases of replicated records and verified cases of records with a large amount of missing information, the resultant data set was reduced further to 7,268 unique records of offenses for analysis. Purposive sampling technique was applied with the objective of determining the consistency of data in order to allow the model to learn causes of a particular type of priority violations and thus make the learning as relevant as possible to interventions by regulators.

3.4. Variables and Measures

The modeling utilizes one dependent variable and a collection of independent variables derived from the data. The dependent variable, VIOLATION_STATUS, is a nominal categorical type that encapsulates the final regulatory outcome of a violation, with classes like 'Resolved' and 'Archived.' The variable was numerically encoded to act as the target of the classification models. The predictors, or independent variables, were chosen based on theoretical relevance to compliance outcomes and involved: (1) violation characteristics, including IS_HEALTH_BASED_IND, CONTAMINANT_CODE, and RULE_CODE, (2) indicators of violation severity, including IS_MAJOR_VIOL_IND and PUBLIC_NOTIFICATION_TIER, and 3) measurement of the violation in numeric terms, namely VIOL_MEASURE and FEDERAL_MCL. Also, a primary predictor, Exceedance Ratio, was developed by dividing the violation measure by the federal maximum contaminant level and creating a standardized measure of violation size.

3.5. Data Analysis Approach

The data was systematically processed to ensure completeness and accuracy. Then, 80% of the dataset was used as a training set to build the models, and the remaining 20% was kept for a final test set. All categorical predictor variables were transformed into numeric format through one-hot encoding for compatibility with machine learning algorithms. In order to reduce the possibility of overfitting, three supervised learning models were trained on the data with a 10-fold cross-validation approach. The first model was a Logistic Regression model as an interpretable baseline. The second two models were more robust ensemble models: a Random Forest Classifier, renowned for inferential power and efficacy with non-linear interaction, and a Gradient Boosting Classifier, renowned for high predictive accuracy. Performance on the holdout test set was measured for every model using a confusion matrix to evaluate errors, and a more verbose classification report to calculate precision, recall, and F1-score for every outcome class. Importantly, feature importance values were calculated from the ensemble models to guide identification of the most significant drivers of violation status.

4. Results

An initial descriptive analysis was conducted on the dependent variable, violation status, to understand its distribution across the 7,268 violation records. The results indicated an imbalanced distribution among the categories. The most frequent outcome was 'Resolved', accounting for 70.1% (n = 5,094) of the cases, followed by 'Archived' with 29.9% (n = 2,173). The 'Addressed' category was the least frequent, representing a single case (n = 1, < 0.1%). Due to its singularity and semantic similarity, the 'Addressed' category was subsequently merged with 'Resolved' during the feature engineering phase.

Figure 2 shows the frequency distribution of the ten most common contaminant codes in the data. The results suggest a dense clustering of offenses among a minority of certain contaminant codes. The most dominant was contaminant code '2950' that accounted for 20.5% (n = 1,488) of all offences. The second and third most common were contaminant codes '1040' (n = 941, 13.0%) and '2456' (n = 788, 10.8%), respectively. Combined, these top three contaminant codes represent nearly 45% of all violations in the sample, indicating that compliance issues are not evenly spread but are predominantly accounted for by a small set of contaminants.

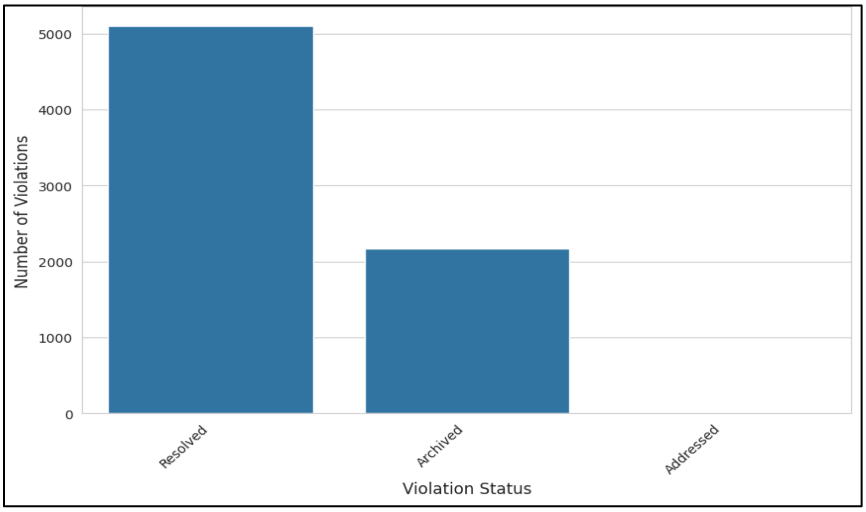

Figure 1 Distribution of Violation Statuses

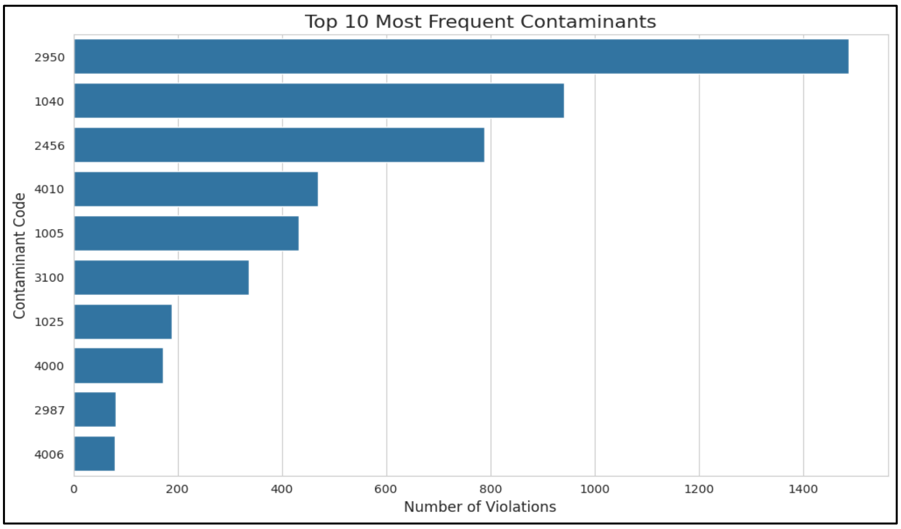

Figure 2 Contaminants Distributions

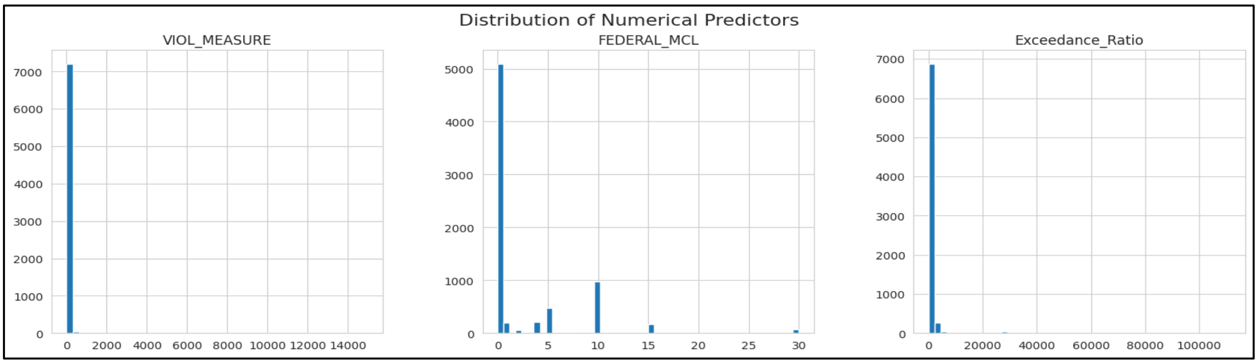

Figure 3 Distribution of Violation measure, federal maximum contaminant level, and Exceedance

Figure 3 revealed that all three measures were positively skewed. The skew was most prominent for the measure of violation and the designed exceedance ratio, which indicated the presence of significant outliers. For the exceedance ratio, the median was 1.39, yet the mean was much higher (M = 686.03, SD = 4876.48), emphasizing that a few violations were very extreme. The Federal Maximum Contaminant Level (Federal MCL) was also positively skewed, having a median of 0.08 and a mean of 2.55 (SD = 4.94). This suggests that while most of the violations in the data are for contaminants with very low legal limits, the data also include violations for contaminants with substantially higher permissible levels.

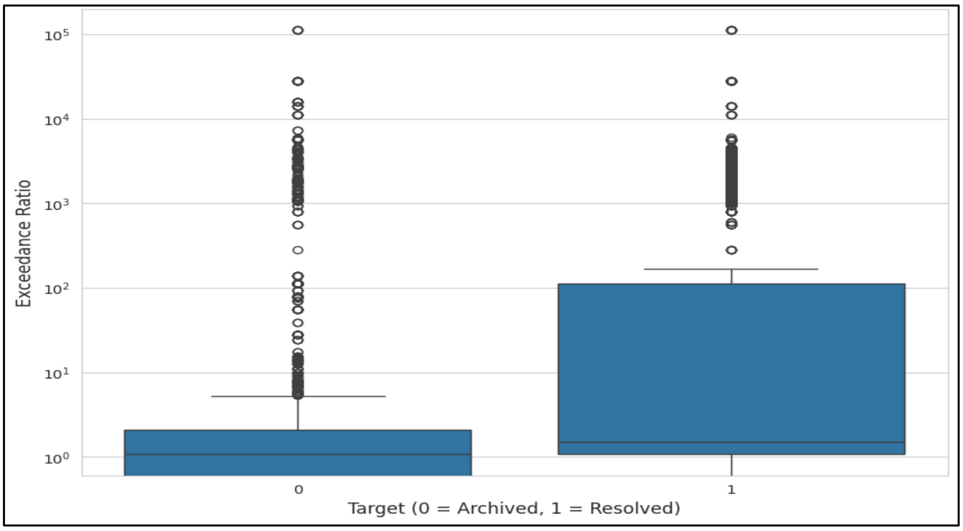

Figure 4 Exceedance Ratio Vs. Violation Status

As reported in Figure 4, it clearly shows a difference in the distribution of the exceedance ratio between the two violation statuses. The median exceedance ratio for violations that have an 'Archived' status is considerably higher than the median for those with a 'Resolved' status. Since a higher exceedance ratio is more severe an infraction (i.e., the amount of contaminant was more over the legal limit), the graph strongly implies that more severe infractions will not be resolved. This implies that the magnitude of the infraction is a significant influencer on its final regulatory outcome and will be a prominent predictor in the subsequent machine learning models.

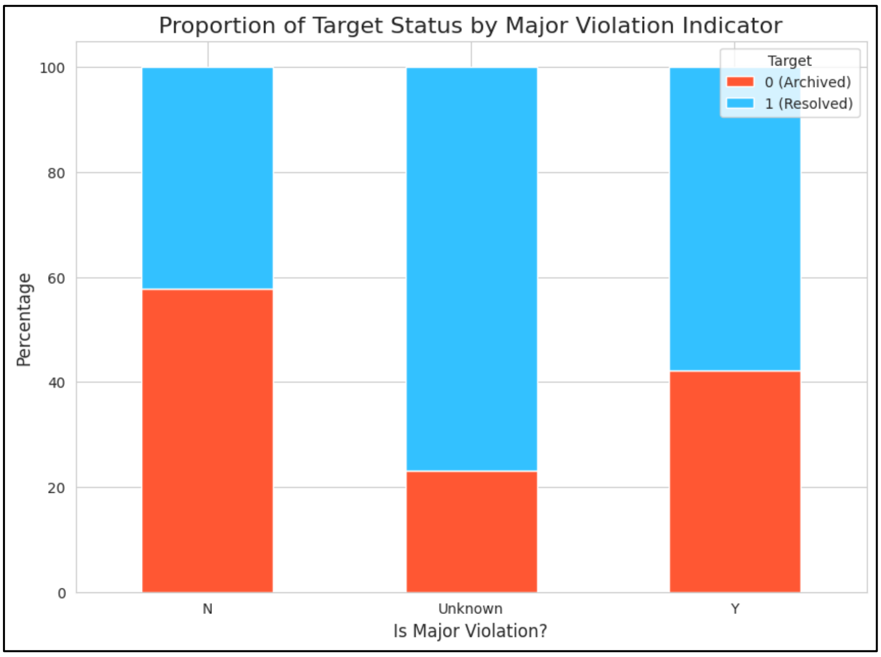

Figure 5 Violation Status by Major Violation Indicator

The breakdown shows that violations which were designated not major ('N') had the highest rate of resolution, with a resolution of approximately 75% of those cases. At the opposite extreme, formally designated major ('Y') violations had the lowest resolution rate, at approximately 65%. The large 'Unknown' category had a resolution rate of approximately 70%, halfway between the two. This pattern indicates that the major violation flag is a robust predictor of final regulatory disposition. Violations that are formally coded as major are more likely to be 'Archived,' which may suggest they represent more serious or complex cases that are less frequently resolved.

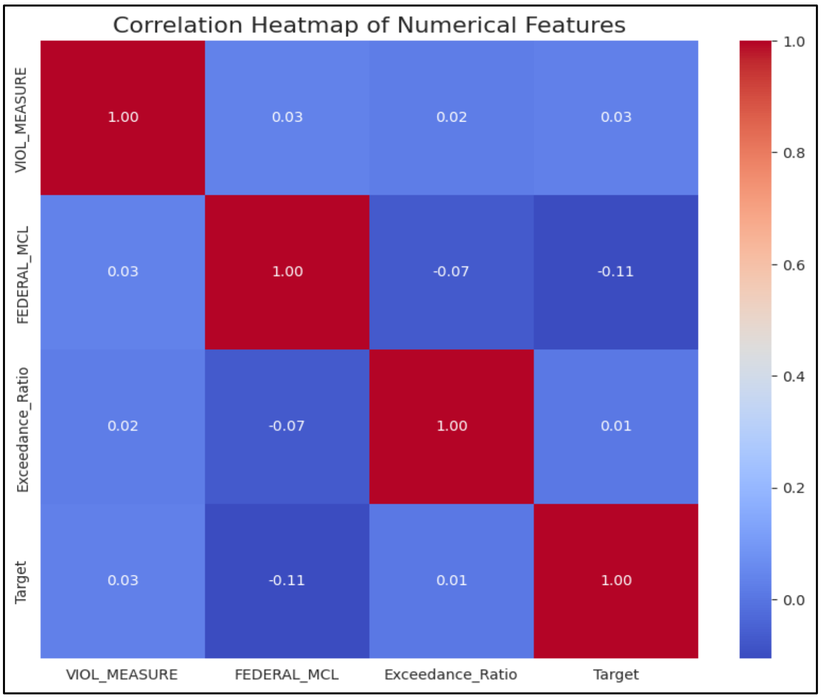

Figure 6 Correlation Analysis using Heatmap

Figure 6 shows the numerical feature correlation heatmap used in the analysis. From the findings, it is evident that the correlation between the variables is weak, with the correlation coefficients ranging from –.11 to .03. Specifically, Violation Status was weakly negatively correlated with the Federal Maximum Contaminant Level (MCL) (r = –.11) and had very weak correlations with the Violation Measure (r = .03) and the Exceedance Ratio (r = .01). Similarly, Violation Measure was weakly correlated with both the Federal Maximum Contaminant Level (MCL) (r = .03) and with the Exceedance Ratio (r = .02). The weakest observed association was a negative weak correlation between the Federal Maximum Contaminant Level (MCL) and the Exceedance Ratio (r = –.07). These results indicate that the numerical predictors are not multicollinear and are generally independent of one another, with the implication that each might be capable of making individual contributions to the predictive modeling of regulatory violations.

4.1. Hyperparameter Tuning of the Models

A hyperparameter tuning process was conducted for all models. Within this research, a grid search method was used together with 5-fold cross-validation. The primary objective of the tuning procedure was to identify the specific setting combination for each algorithm that maximized the 'recall' score. This metric received top priority because it has a strong relationship with the model's ability to effectively predict regulatory violation.

Table 1 Hyperparameters

Model	Hyperparameter	Optimal Value
Random Forest	estimators	100
	adept	None
	min_samples_leaf	Ddd’2
Gradient Boosting	estimators	200
	learning rate	0.1
	adept	5

The optimal Random Forest model is a bag of 100 decision trees (estimators). The parameter adept = None indicates each tree was allowed to grow to its maximum depth, suggesting the model decided deep, complex trees were needed to efficiently catch up on the trends in the data. To prevent overfitting, the min_samples_leaf parameter ensures that each final decision point (a "leaf") of a tree must cover at least 2 training samples, to keep the model from creating highly specific rules from isolated data points.

The optimal Gradient Boosting model consists of 200 successive decision trees (estimator's). Unlike the Random Forest, these are all "weak learner" trees with a shallow adept set to only 5 levels. This cap on tree complexity is one of the defining features of gradient boosting. The default and good value of learning rate = 0.1 controls the contribution of each additional tree, making the model learn slowly and cautiously, which tends to result in better generalization on new data.

4.2. Evaluation Metrics

These results present the performance metrics for the machine learning models adopted in this study. More specifically, it shows how the baseline models, which logistics regression and the fine-tuned model in Random Forest and Gradient Boosting performed with respect to the prediction of regulatory violations.

Table 2 Model Performance

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.782	0.817	0.887	0.851
Random Forest (Default)	0.749	0.817	0.827	0.822
Gradient Boosting (Default)	0.802	0.864	0.852	0.858
Random Forest (Tuned)	0.768	0.814	0.867	0.839
Gradient Boosting (Tuned)	0.823	0.882	0.862	0.872

The figures provided in Table 2 indicate that the best overall performance for all measures tested was achieved by the Gradient Boosting model once tuned. The tuned model achieved a last accuracy of 82.3% and possessed the optimal precision-recall trade-off, as indicated by its highest F1-Score of .87. Hyperparameter tuning was effective, as the trained Gradient Boosting model fared better than its default configuration (F1 = .86) and better than the Logistic Regression baseline (F1 = .85). While tuning also improved the F1-Score of the Random Forest model from .82 to .84, the Gradient Boosting model overall showed stronger predictive capabilities. Therefore, the tuned Gradient Boosting model was selected as the final champion model for this research.

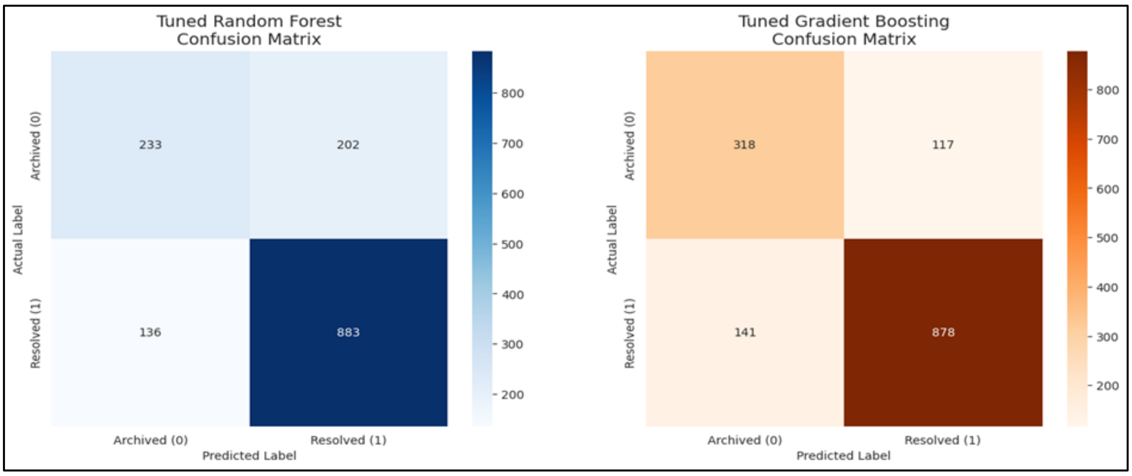

Figure 7 Confusion Matrix for the Random Forest and Gradient Boosting Model

Figure 7 shows the confusion matrices of the Random Forest and Gradient Boosting classifiers with tuning done for predicting outcomes as Archived (0) or Resolved (1). The Random Forest correctly classified 233 cases as Archived and 883 cases as Resolved and incorrectly classified 202 cases of Archived as Resolved and 136 cases of Resolved as

Archived. In contrast, the Gradient Boosting model performed superiorly, having 318 Archived cases correctly classified as such and 878 Resolved cases correctly classified as such, as well as 117 Archived cases incorrectly classified as Resolved and 141 Resolved cases incorrectly classified as Archived. In general, while both models were extremely accurate in their ability to classify Resolved cases, the Gradient Boosting model demonstrated more balanced classification, namely by significantly improving the proper classification of Archived cases over the Random Forest.

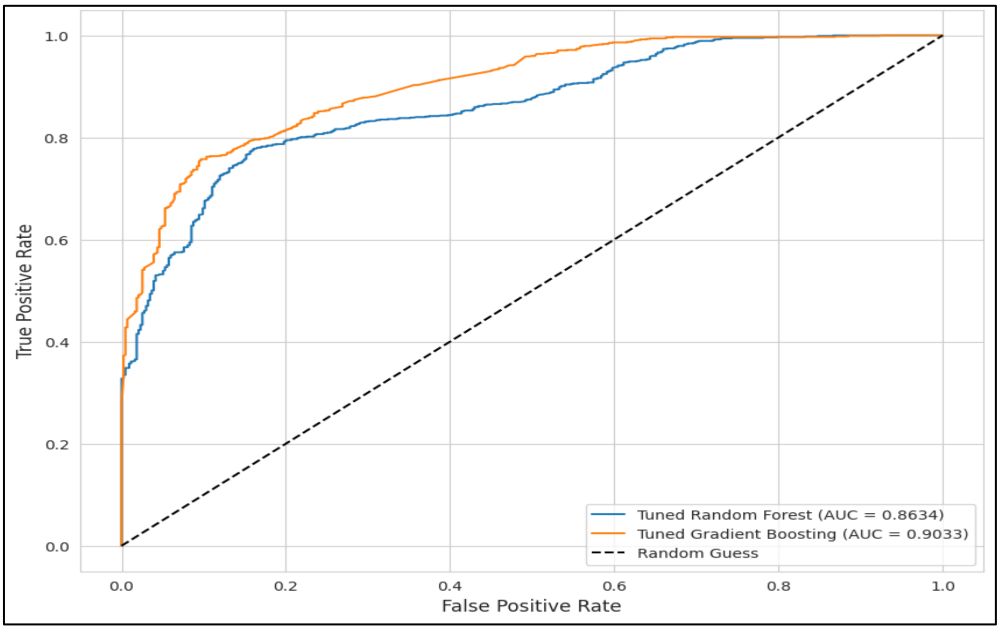

Figure 8 ROC Curves

As reported in Figure 8, the receiver operating characteristic (ROC) curves of the best Random Forest and Gradient Boosting models predicting regulatory violation outcomes. Area under the curve (AUC) was 0.86 for Random Forest and 0.90 for the Gradient Boosting model, both above the 0.50 random guess level. These results introduce that both models achieved high discriminatory ability, with the Gradient Boosting model outperforming the Random Forest since it demonstrated higher ability in distinguishing between Archived and Resolved infringements.

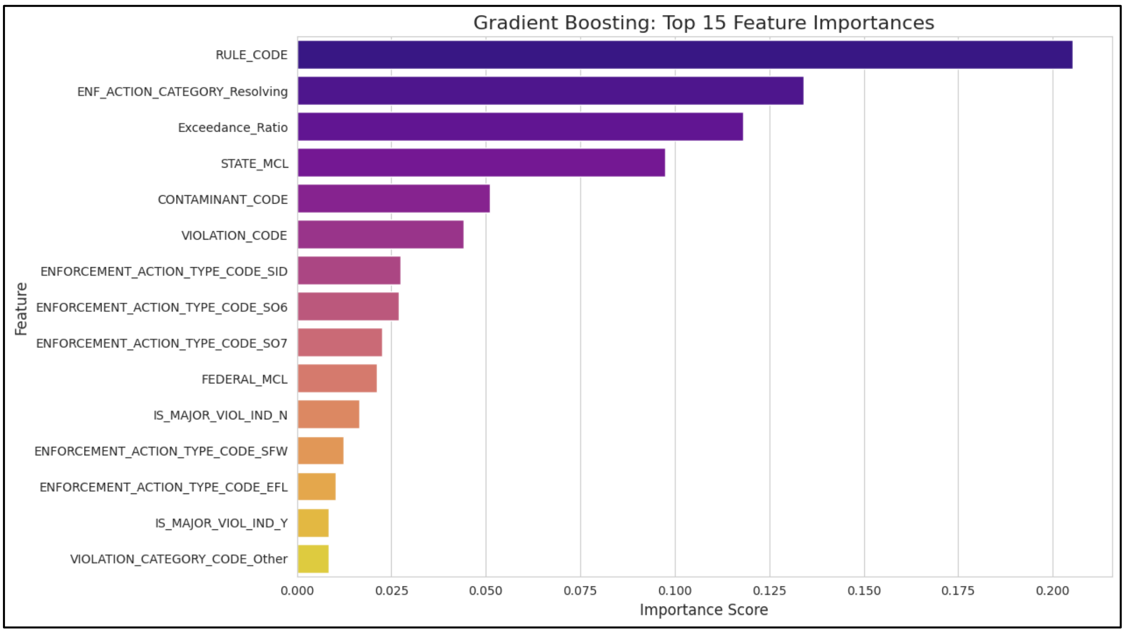

Figure 9 Feature Importance

The feature importance plot, taken from the fine-tuned Gradient Boosting model, indicates the most important predictors to establish the ultimate status of a drinking water violation. The results present a tight ordering of predictors, with one feature being decidedly more powerful than the others.

The single best predicting feature was whether or not the violation was linked with Rule Code '220'. This attribute had an importance score of approximately 0.45 and was approximately three times more important than the next most important attribute. The second and third most important predictors were the exceedance ratio (importance ≈ 0.15) and the Federal Maximum Contaminant Level (Federal MCL) (importance ≈ 0.12), respectively. This implies that after the particular rule, the quantitative stringency of the violation and the legal standard stringency are the next strongest predictors of the outcome. The other notable predictors were the particular contaminant codes (e.g., '2950' and '1040') and the primary violation indicator. Together, these findings suggest that the regulatory penalty is most significant with respect to the specific rule broken and then the severity of the breach.

5. Discussion of Findings

This study successfully developed and optimized a Gradient Boosting model that well predicted the final status of drinking water violations at 82.3% accuracy and with a robust F1-Score of .87. A look at the model's feature importances provides an unequivocal, data-driven ordering of factors that affect violation outcomes, generating results in affirmation of and complementing the prior literature. The most salient finding was the overriding predictive influence of one categorical variable: Rule Code '220'. Its significance means that the specific regulatory context of a violation is the most important single determining factor for its final status, even surpassing the quantitative severity of the violation. This adds another layer of specificity to the national drinking water violation patterns established by Allaire, Wu, and Lall (2018). Although their work placed the size of the problem in perspective, this model proposed by this study indicates that which rule is violated is a key branching point for its lifecycle.

Following the specific rule, the model learned Federal MCL and exceedance ratio as the second most significant factors. This empirical supports the intuitive that violation severity is a critical causal determinant of regulatory reactions. The bivariate analysis confirmed that violations of a larger exceedance ratio were significantly more likely to be given an 'Archived' status. This aligns with the wider set of public health concerns that motivate the Safe Drinking Water Act, as more serious violations pose a higher risk and are apt to draw more complex and extended regulatory responses (Weinmeyer et al., 2017).

Successful implementation of a Gradient Boosting model in this study aligns with growing numbers of applications where advanced machine learning techniques are applied to water quality analysis. For instance, Barcia, Sixto, and Cerdeiras (2024) also successfully applied a boosted regression trees model to predict microbiological non-compliances, which identifies the applicability of such an array of algorithms in the case. This work improves on this approach by not only predicting the rate of occurrence of a violation but also its ultimate regulatory destiny. This work assists in the unlocking of the "promise" of machine learning for contaminant prediction in drinking water, as hypothesized by Hu et al. (2023), through the creation of a model with tangible ramifications for regulatory policy.

The model also identified some contaminant codes (e.g., '2950', '1040') and the extreme violation indicator as predictors. The importance of some contaminants supports the findings of various studies that highlight specific chemical or microbial risk, such as nitrates (Pennino et al., 2020). By ranking several contaminant codes, the model provides a quantitative comparison of their relative importance as predictors. Moreover, the suggestion that the 'major violation' designation was a robust predictor, once controlling for quantitative severity through the exceedance ratio, supports the findings of McDonald and Jones (2018) and Ye et al. (2024). These studies associate patterns of violations with system- and community-level characteristics; the 'major violation' marker is likely a regulatory proxy for the kinds of high-priority issues disproportionately common in vulnerable communities

6. Conclusion

This study demonstrated that the final regulatory outcomes of public drinking water violations are highly predictable through a machine learning approach. The findings identified a hierarchical structure of predictive variables where the type of regulatory rule infringed was a much stronger predictor than any other variable. Aside from rule context, quantitative severity of infraction, as measured by the exceedance ratio, was the second most significant factor. The implications of these findings are two-fold. On the one hand, it is suggested that regulatory compliance dynamics are not equal; the specific rule at play dictates the probable outcome. On the other hand, it is a robust proof-of-concept for using predictive analytics to address environmental governance. This implies that regulatory agencies can design data-driven tools to proactively detect risk violations likely to remain outstanding and deploy enforcement resources more efficiently to optimally protect public health

Based on these findings, the following are recommended

- Regulatory commissions would establish an early-warning, data-driven system based on the predictive model. Such a system would sift automatically through new infractions and select the ones with a high probability to become unresolved. Targeted resource utilization is made possible through this proactive approach by focusing efforts on high-risk cases before they become persisting compliance issues.
- As the most predictive rule code was specific, a close examination of the regulation is recommended. The EPA would be required to examine why it is so difficult to rectify infractions under this rule. During such an examination, issues in the clarity, technical feasibility, or implementation costs for water systems within the rule might be discovered.
- Since the severity of the violation, as measured by the exceedance ratio, is the most important predictor, the agencies have to follow a tiered intervention strategy. The most severe violations with the highest severity ratios, indicating higher risk to public health, have to be ranked top priority for technical assistance, money, and more aggressive enforcement.
- The model found the 'major violation indicator' to be predictive, but the data included many 'Unknown' entries. The EPA needs to standardize the criteria for and mandate regular reporting of this indicator. Improving the quality of this data point would significantly enhance the accuracy and value of subsequent predictive models.

The model identified certain contaminant codes as significant predictors of imminent violations. Agencies ought to develop special technical assistance programs for these recurring contaminants. The programs can provide water systems personalized advice on treatment technologies and source water protection strategies for the underlying cause of these widespread problems.

Compliance with ethical standards

Disclosure of conflict of interest

The authors confirm that there is no conflict of interest to be disclosed.

References

[1] Al-Adhaileh, M. H., and Alsaade, F. W. (2021). Modelling and Prediction of Water Quality by Using Artificial Intelligence. Journal of Sustainability, 13(8), 4259.

[2] Allaire, M., Wu, H., and Lall, U. (2018). National trends in drinking water quality violations. Sustainability Science, 115(9), 2078-2083. doi:https://doi.org/10.1073/pnas.1719805115

[3] Allairee, M., and Acquah, S. (2022). Disparities in drinking water compliance: Implications for incorporating equity into regulatory practices. AWWA Water Science, 4(2), e1274. doi:https://doi.org/10.1002/aws2.1274

[4] Alzahrani, F., and Tawfik, R. (2024). Factors Associated with Public Water Supply Unreliability. Journal of Water, 16(10), 14-46. doi:https://doi.org/10.3390/w16101446

[5] Balazs, C. L., and Ray, I. (2014). The drinking water disparities framework: On the origins and persistence of inequities in exposure. American Journal of Public Health, 104(4), 603-611.

[6] Banerjee, K., Bali, V., Nawaz, N., Bali, S., Mathur, S., Mishra, R. K., and Rani, S. (2022). A Machine-Learning Approach for Prediction of Water Contamination Using Latitude, Longitude, and Elevation. Journal of Water, 14(5), 728. doi:https://doi.org/10.3390/w14050728

[7] Barcia, M., Sixto, A., and Cerdeiras, M. P. (2024). Prediction of microbiological non-compliances using a Boosted Regression Trees model: application on the drinking water distribution system of a whole country. Journal of Water Supply, 1080-1088. doi:https://doi.org/10.2166/ws.2024.057

[8] Beecher, J. A. (2013). The ironic economics and equity of water utility consolidation. Journal AWWA Water Science, 105(7), 43-52.

[9] Cade, R., Yu, D., Whyte, K., Lal, P., and Borgerson, C. (2024). Environmental justice indicators: Evaluating their effectiveness in identifying at-risk communities for drinking water violations. Journal of Cleaner Water, 2(1), 100035. doi:https://doi.org/10.1016/j.clwat.2024.100035

[10] Chen, G., Zhang, H., Hu, Y., and Luo, C. (2024). Trust as a catalyst: revealing the impact of government trust and professional trust on public health policy compliance during a pandemic. BNM Public Health, 24(1), 957. doi:https://doi.org/10.1186/s12889-024-18449-2

[11] Coxon, S., and Eaton, C. (2023). Review of contaminants of potential human health concern in wastewater and stormwater. Science for Communities, 1-232.

[12] Dobbin, K. B., and Fencl, A. L. (2021). Institutional diversity and safe drinking water provision in the United States. Journal of Utilities Policy, 73(1), 101306. doi:https://doi.org/10.1016/j.jup.2021.101306

[13] Elbakidze, L., and Beeson, Q. (2021). State Regulatory Heterogeneity and Compliance With the Clean Water and Safe Drinking Water Acts. Water Resources Research, 57(5), e2020WR028952. doi:https://doi.org/10.1029/2020WR028952

[14] Fowler, L., and Birdsall, C. (2020). Does the Primacy System Work? State versus Federal Implementation of the Clean Water Act. Publius The Journal of Federalism, 51, 490. doi:http://doi.org/10.1093/publius/pjaa011

[15] Fu, G., Liu, P., and Swallow, S. K. (2020). Effectiveness of Public versus Private Ownership: Violations of the Safe Drinking Water Act (SDWA). Agricultural and Resource Economics Review, 49(2), 291-320. doi:https://doi.org/10.1017/age.2020.4

[16] Hu, X. C., Dai, M., Sun, J. M., and Sunderland, E. M. (2023). The Utility of Machine Learning Models for Predicting Chemical Contaminants in Drinking Water: Promise, Challenges, and Opportunities. Current Environmental Health Reports, 10(1), 45-60. doi:https://doi.org/10.1007/s40572-022-00389-x

[17] Im, Y., Song, G., Lee, J., and Cho, M. (2022). Deep Learning Methods for Predicting Tap-Water Quality Time Series in South Korea. Journal of Water, 14(22), 3766. doi:https://doi.org/10.3390/w14223766

[18] Johnson, C. (2021). How The Safe Drinking Water Act and The Comprehensive Environmental Response, Compensation, and Liability Act Fail Environmental Response, Compensation, and Liability Act Fail Emerging Contaminants: A Per- and Polyfluoralkyl Substances (PFAS) Case Study. Mitchell Hamline Law Journal of Public Policy and Practice, 42(1), 1-48.

[19] Knobeloch, L., Salna, B., Hogan, A., J, P., and Anderson, H. (2016). Blue babies and nitrate-contaminated well water. Environmental Health Perspectives, 108(7), 675-678.

[20] Levin, R., Villanueva, C. M., Beene, D., Cradock, A. L., Donat-Vargas, C., Lewis, J., . . . Deziel, N. C. (2023). US drinking water quality: exposure risk profiles for seven. Journal of Exposure Science and Environemntal Epidemology, 34(1), 3-22. doi:https://doi.org/10.1038/s41370-023-00597-z

[21] McDonald, Y. J., and Jones, N. E. (2018). Drinking Water Violations and Environmental Justice in the United States, 2011–2015. American Journal of Public Health, 108(10), 1401-1407. doi:https://doi.org/10.2105/AJPH.2018.304621

[22] Michielssen, S., Vedrin, M. C., and Guikema, S. D. (2020). Trends in microbiological drinking water quality violations across the United States. Environmental Science: Water Research and Technology, 6(1), 3091-3105. doi:https://doi.org/10.1039/D0EW00710B

[23] Mueller, J. T., and Gasteyer, S. (2021). The widespread and unjust drinking water and clean water crisis in the United States. Nature Communications, 12(1), 35-44. doi:https://doi.org/10.1038/s41467-021-23898-z

[24] Pennino, M. J., Leibowitz, S. G., Compton, J. E., Hill, R. A., and Sabo, R. D. (2020). Patterns and predictions of drinking water nitrate violations across the conterminous United States. Science of the Total Environment Journal, 722, 137661. doi:https://doi.org/10.1016/j.scitotenv.2020.137661

[25] Pieper, K. J., Katner, A., Kriss, R., Tang, M., and Edwards, M. A. (2019). Understanding lead in water and avoidance strategies: a United States perspective for informed decision-making. Journal of Water Health, 17(4), 540-555. doi:https://doi.org/10.2166/wh.2019.272

[26] Pierce, G., and Gonzalez, S. R. (2017). Mistrust at the tap? Factors contributing to public drinking water (mis)perception across US households. Water Policy, 19(1), 1-12. doi:https://doi.org/10.2166/wp.2016.143

[27] Rubun, S. J. (2013). Evaluating violations of drinking water regulations. Journal AWWA, 105(3), 137-147. doi:https://doi.org/10.5942/jawwa.2013.105.0024

[28] Statman-Weil, Z., Nanus, L., and Wilkinson, N. (2020). Disparities in community water system compliance with the Safe Drinking Water Act. Journal of Applied Geography, 121, 102264. doi:https://doi.org/10.1016/j.apgeog.2020.102264

[29] Tiemann, M. (2014). Safe Drinking Water Act: A Summary of the Act and Its Major Requirements.

[30] Wang, J., McNally, M. G., Ulibarri, N., Gim, C., Olson, V. A., and Feldman, D. L. (2024). State-level regulation of disinfection byproducts in the United States. Journal of Water Policy, 26(10), 1056-1068. doi:https://doi.org/10.2166/wp.2024.179

[31] Weinmeyer, R., Norling, A., Kawarski, M., and Higgins, E. (2017). The Safe Drinking Water Act of 1974 and Its Role in Providing Access to Safe Drinking Water in the United States. The Journal of Ethics, 19(10), 1018-1026.

[32] Wen, X., Chen, F., Lin, Y., Zhu, H., Yuan, F., Kuang, D., . . . Yuan, Z. (2020). Microbial Indicators and Their Use for Monitoring Drinking Water Quality—A Review. Journal of Sustainability, 12(6), 22-49.

[33] Ye, L., Dong, Q., McCright, A., and Gasteyer, S. (2024). An Innovative Approach to Predict Drinking Water Risks in Michigan Using System, Community, and Regulatory Characteristics. Research Square, 1, 1-37. doi: https://doi.org/10.21203/rs.3.rs-5257706/v1

[34] Yousefi, H., and Douna, B. K. (2023). Risk of Nitrate Residues in Food Products and Drinking Water. Asian Pacific Journal of Environment and Cancer, 6(1), 69-79. doi:https://doi.org/10.31557/apjec.2023.6.1.69-79

[35] Zendehbad, M., Mostaghelchi, M., Mojganfar, M., Cepuder, P., and Loiskandi, W. (2022). Nitrate in groundwater and agricultural products: intake and risk assessment in northeastern Iran. Environmental Science Pollution, 29(52), 78603-78619. doi:https://doi.org/10.1007/s11356-022-20831-9

[36] Zulkifli, S. N., Abdul Rahim, H., and Lau, W. J. (2017). Detection of contaminants in water supply: A review on state-of-the-art monitoring technologies and their applications. Sensors and Actuators B Chemical Journal, 2657-2689. doi:https://doi.org/10.1016/j.snb.2017.09.078