UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Economía

Territorial Exposure to River Pollution and Prediction of Health Risks: A Predictive Analysis for the Esmeraldas River Basin in Ecuador

María Gabriela Cisneros Bastidas

Economía

Trabajo de fin de carrera presentado como requisito

para la obtención del título de

Economista

Quito, 9 de mayo de2025

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Economía

HOJA DE CALIFICACIÓN
DE TRABAJO DE FIN DE CARRERA

Territorial Exposure to River Pollution and Prediction of Health Risks:

A Predictive Analysis for the Esmeraldas River Basin in Ecuador

María Gabriela Cisneros Bastidas

Professor's Name, Academic Title — Pablo Astudillo, Professor, College of Economics

Quito, May 9, 2025

© DERECHOS DE AUTOR

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en la Ley Orgánica de Educación Superior del Ecuador.

Nombres y apellidos: María Gabriela Cisneros Bastidas

Código: 00325873

Cédula de identidad: 1150043790

Lugar y fecha: Quito, 9 de mayo de 2025

UNPUBLISHED DOCUMENT

Note: The following capstone project is available through Universidad San Francisco de Quito USFQ institutional repository. Nonetheless, this project – in whole or in part – should not be considered a publication. This statement follows the recommendations presented by the Committee on Publication Ethics COPE described by Barbour et al. (2017) Discussion document on best practice for issues around theses publishing available on http://bit.ly/COPETheses.

ACLARACIÓN PARA PUBLICACIÓN

Nota: El presente trabajo, en su totalidad o cualquiera de sus partes, no debe ser considerado como una publicación, incluso a pesar de estar disponible sin restricciones a través de un repositorio institucional. Esta declaración se alinea con las prácticas y recomendaciones presentadas por el Committee on Publication Ethics COPE descritas por Barbour et al. (2017) Discussion document on best practice for issues around theses publishing, disponible en http://bit.ly/COPETheses.

ABSTRACT

This study analyzes the relationship between territorial exposure to water pollution in the Esmeraldas River Basin and public health vulnerability in Ecuador. Using a binary logistic regression model combined with the Synthetic Minority Over-sampling Technique (SMOTE), a predictive model was developed to classify cantons based on their risk of waterborne diseases. The analysis incorporated socioeconomic variables such as poverty rates, healthcare personnel availability per capita, and the incidence of primary and secondary water-related diseases.

The results show that territorial exposure, together with socioeconomic factors and healthcare infrastructure, constitutes a significant predictor of health vulnerability. The model achieved an overall accuracy of 77% and an area under the curve (AUC) of 0.74, demonstrating a satisfactory ability to distinguish between high- and low-risk cantons. However, the findings also highlight the multifactorial nature of vulnerability, indicating that geographic exposure, although relevant, is not sufficient by itself to fully explain the observed risk patterns.

This study provides empirical evidence that can inform integrated public policies combining environmental interventions, improvements in healthcare infrastructure, and reductions in socioeconomic disparities to mitigate the health impacts of water pollution.

Keywords: water pollution, waterborne diseases, logistic regression, health vulnerability, SMOTE, Esmeraldas River Basin, Ecuador.

RESUMEN

Este estudio analiza la relación entre la exposición territorial a la contaminación del agua en la Cuenca del Río Esmeraldas y la vulnerabilidad en salud pública en Ecuador. Utilizando un modelo de regresión logística binaria combinado con la técnica de sobremuestreo sintético SMOTE, se construyó un modelo predictivo para clasificar cantones según su riesgo de enfermedades relacionadas con el agua. El análisis incorporó variables socioeconómicas como la tasa de pobreza, la disponibilidad de personal de salud per cápita y la incidencia de enfermedades hídricas primarias y secundarias.

Los resultados muestran que la exposición territorial, junto con factores socioeconómicos y de infraestructura sanitaria, constituye un predictor significativo de la vulnerabilidad sanitaria. El modelo alcanzó una precisión general del 77% y un área bajo la curva (AUC) de 0.74, demostrando una capacidad satisfactoria para discriminar entre cantones de alto y bajo riesgo. Sin embargo, los hallazgos también resaltan la naturaleza multifactorial de la vulnerabilidad, indicando que la exposición geográfica, aunque relevante, no es suficiente por sí sola para explicar los patrones de riesgo observados.

Este trabajo proporciona evidencia empírica que puede guiar políticas públicas integradas que combinen intervenciones ambientales, mejoras en infraestructura sanitaria y reducción de brechas socioeconómicas para mitigar los impactos de la contaminación del agua sobre la salud pública.

Palabras clave: contaminación del agua, enfermedades hídricas, regresión logística, vulnerabilidad sanitaria, SMOTE, Cuenca del Río Esmeraldas, Ecuador.

Table of Contents

INDEX OF TABLES

INDEX OF FIGURES

INTRODUCTION

The deterioration of water quality in river systems is one of the most pressing environmental and public health challenges in Latin America. In Ecuador, this issue is particularly acute in the Esmeraldas River basin, an extensive hydrological network that originates in the Andes and traverses densely populated urban and peri-urban areas before reaching the Pacific coast. Despite its ecological and socioeconomic importance, this basin has become the recipient of a cumulative load of untreated domestic, industrial, and agricultural wastewater, posing severe risks to both ecosystems and human health.

This work emerges from a diagnostic concern that combines environmental degradation and its underestimated fiscal consequences. Recent studies have confirmed that several rivers within this basin, (including the Machángara River), consistently exceed international and national thresholds for microbiological and chemical contamination. According to Vinueza et al. (2021), the presence of Escherichia coli and total coliforms surpasses legal limits in virtually all major Ecuadorian rivers, with the Esmeraldas and Machángara among the most critically polluted. Moreover, according to the Ministry of Public Health (MSP, 2019), nearly 88% of diarrheal diseases in the country are attributable to poor water quality, inadequate sanitation, and deficient hygiene services.

The problem is not limited to natural degradation, it reflects institutional and infrastructural deficiencies that have persisted over time. In Quito, for example, only 3% of wastewater is treated before being discharged into natural watercourses, while the remaining 97% is released untreated (UDLA, 2024). These figures exemplify a broader governance failure in water management systems. Additionally, this context places Ecuador at a disadvantage in its

efforts to meet the Sustainable Development Goals (SDGs), particularly Goal 6, which emphasizes universal access to safe water and sanitation.

Although the environmental and health dimensions of water pollution have received increasing attention, the economic dimension remains insufficiently explored in the Ecuadorian context. There is a critical gap in understanding how pollution-driven health burdens translate into increased public expenditure and fiscal pressure. Addressing this gap is essential not only for improving water management policies, but also for generating evidence that supports efficient allocation of public resources. A rigorous economic assessment of the costs associated with waterborne diseases and the potential savings derived from effective water treatment infrastructure is a necessary step toward more rational and sustainable policymaking.

Water security, in this context, must be understood beyond its physical dimension. As defined by the Water Security and Sustainable Development Hub, it involves the integration of social, ecological, institutional, and technological dimensions to ensure the sustainable use of water resources and the protection of human and environmental health (Water Security and Sustainable Development Hub, 2024). This concept underpins the present research, which adopts an interdisciplinary and systemic approach to the economic consequences of water contamination.

This project seeks to quantify the economic impact of water pollution in the Esmeraldas River basin by evaluating its effect on public health expenditure. Through a spatial econometric model and a cost-benefit analysis, this work aims to provide empirical evidence on how environmental degradation generates measurable fiscal burdens and how investment in wastewater treatment infrastructure could reduce such costs.

The following sections are structured as follows: first, a review of the relevant literature; second, a detailed explanation of the methodological framework; third, the presentation and analysis of results; and finally, a discussion of findings and policy implications. This structure is intended to guide the reader through the analytical process while providing a comprehensive understanding of the multidimensional impact of water pollution in Ecuador.

DEVELOPMENT

1. Contextual Framework: The Esmeraldas River Basin and Water Pollution in Ecuador

1.1 Geographical and hydrological structure of the basin

The Esmeraldas River basin, located in northwestern Ecuador, is one of the country's largest and most complex hydrological systems. The basin extends from the high-altitude Andean regions down to the Pacific coast, encompassing diverse climatic zones, ecological areas, and socio-economic contexts. Structurally, the basin is primarily formed by three major tributaries: the Guayllabamba, Blanco, and Quinindé rivers. These rivers, together with around seventy smaller streams, create a vast hydrological network that serves ecological functions and sustains significant agricultural, industrial, and urban activities within its territory (Reyes Vera et al., 2022).

The Guayllabamba River is particularly critical, as it originates in the Andes near the Quito metropolitan area, subsequently flowing northwest and joining with the Blanco River. Further downstream, the Quinindé River merges into this system, completing the major hydrological confluence that ultimately forms the Esmeraldas River, which continues its navigable route to the Pacific Ocean (Reyes Vera et al., 2022). Historically, these waterways have supported a rich biodiversity, but intensive anthropogenic pressures, including urban expansion, deforestation, and agricultural development, have significantly altered their natural ecological balance (Reyes Vera et al., 2022).

Hydrologically, the Esmeraldas basin is characterized by substantial seasonal variability in water flows. Between 1965 and 2013, the river exhibited an average monthly discharge of approximately 904.44 m^3/s, with extreme peak flows reaching up to 1906.69 m^3/s during rainy

seasons, contrasted by minimum flows as low as 434.27 m^3/s in dry periods (Reyes Vera et al., 2022). This variability has profound implications for water management, flood risks, and water quality throughout the basin.

In terms of land use, the Esmeraldas basin displays a mosaic of agricultural, pastoral, forested, and urban landscapes. Agricultural activities predominate, involving permanent, semi-permanent, and annual crops, accompanied by extensive livestock production systems (Reyes Vera et al., 2022). Unfortunately, these land-use practices have accelerated landscape fragmentation and ecosystem degradation, significantly reducing the ecological integrity of the basin. From 1990 to 2015, extensive deforestation and land conversion resulted in a 55% reduction in forest cover, intensifying landscape fragmentation and limiting ecological connectivity (Reyes Vera et al., 2022).

Understanding this geographical and hydrological structure is fundamental to comprehending the broader environmental and public health challenges faced by communities in the Esmeraldas basin. The subsequent sections will further explore the implications of these structural characteristics on water quality, public health, economic costs, and institutional governance.

1.2 Sources and types of water pollution in the basin (Machángara, Guayllabamba, Esmeraldas)

Water pollution within the Esmeraldas River basin arises from multiple sources, predominantly untreated domestic sewage, industrial discharges, agricultural runoff, and urban wastewater. One of the most impacted tributaries, the Machángara River, illustrates the severe contamination prevalent throughout the basin. According to recent studies, approximately 97%

of wastewater from Quito is discharged untreated into this river, significantly contributing to elevated concentrations of contaminants such as heavy metals, oils, detergents, and microbial pathogens (UDLA, 2024).

Similarly, the Guayllabamba River, originating in the densely populated Andean highlands near Quito, faces severe pollution from untreated urban and industrial effluents. This pollution is characterized by elevated biochemical oxygen demand (BOD), chemical oxygen demand (COD), and high levels of pathogens including coliforms and Escherichia coli (Vinueza et al., 2021). Agricultural practices in the basin add to this burden, introducing pesticides, herbicides, and fertilizers, which cause eutrophication, further degrading water quality and affecting aquatic ecosystems (Reyes Vera et al., 2022).

Industrial pollution, notably from textile, chemical, and food processing industries located along the tributaries, introduces additional hazardous substances such as heavy metals and organic compounds. The Esmeraldas River itself, receiving inputs from both Machángara and Guayllabamba rivers, becomes heavily loaded with these pollutants, severely impacting ecosystems downstream and posing substantial health risks for local communities dependent on these water resources for domestic use and agriculture (Reyes Vera et al., 2022).

Addressing the diverse sources and complex nature of this pollution requires integrated management strategies that encompass improved infrastructure, rigorous regulatory enforcement, and community-based participation to mitigate further ecological and human health impacts.

1.3 Wastewater infrastructure and treatment coverage

Wastewater management infrastructure within the Esmeraldas River basin is notably insufficient, exacerbating the environmental and public health challenges posed by

contamination. Currently, the city of Quito treats only about 3% of its wastewater, discharging the remaining 97% directly into river systems without adequate treatment (UDLA, 2024). This scenario reflects significant infrastructural deficits and highlights substantial gaps in sanitation coverage across urban and rural areas.

The Vindobona Project, a major wastewater treatment initiative for Quito and its surrounding parishes, aims to address these deficits. With an anticipated capacity to treat an average flow of 7,550 liters per second, the project is designed to benefit over three million residents by the year 2045 (EPMAPS, 2016). Utilizing technologies such as activated sludge with staged feeding, anaerobic digestion for sludge treatment, and ultraviolet disinfection, Vindobona represents a significant advancement toward environmental sustainability and public health improvement.

Despite these ambitious plans, current infrastructure remains fragmented, particularly in rural and peri-urban communities within the basin. Many areas lack even basic sanitation facilities, depending on rudimentary septic systems or direct discharge into waterways. This fragmented infrastructure results in continuous pollutant loads entering the river system, deteriorating water quality and increasing public health risks (Ministerio de Salud Pública, 2019).

To effectively mitigate pollution and protect community health, it is essential to enhance wastewater infrastructure coverage, particularly through decentralized and context-specific treatment solutions. Strengthening infrastructure will require sustained financial investment, robust regulatory frameworks, and collaborative governance involving public entities, private sectors, and affected communities.

1.4 Institutional and regulatory context of water management in the basin

Water resource management in the Esmeraldas River basin involves a complex institutional framework characterized by multiple overlapping jurisdictions and regulatory challenges. At the national level, the Secretaría Nacional del Agua (SENAGUA) is primarily responsible for water resource governance, overseeing allocation, management, and conservation policies. However, decentralized autonomous governments (GADs), municipalities, and provincial authorities also share significant responsibilities related to water sanitation and environmental protection, often resulting in fragmented and inconsistent enforcement of regulations (Water Security and Sustainable Development Hub, 2024).

This fragmented institutional landscape has created substantial governance challenges, notably weak regulatory enforcement, insufficient monitoring capacities, and limited inter-institutional coordination. These deficiencies directly contribute to ongoing pollution issues, as demonstrated by the limited effectiveness of existing wastewater management systems (UDLA, 2024).

Legal frameworks, such as Ecuador's Organic Law on Water Resources and Environmental Code, outline clear standards and obligations for pollution control and water quality management. However, their implementation and compliance are frequently hindered by resource constraints, technical deficiencies, and lack of accountability mechanisms. Strengthening institutional capacity, improving regulatory coherence, and promoting integrated governance approaches are critical steps toward addressing these institutional gaps and achieving effective water management in the basin.

1.5 Public health expenditure linked to waterborne diseases

The contamination of water resources in the Esmeraldas River basin has significant implications for public health expenditures in Ecuador. Waterborne diseases, predominantly gastrointestinal infections and parasitic illnesses, constitute a substantial financial burden for the national health system. According to the Ecuadorian Ministry of Public Health (MSP, 2019), approximately 88% of diarrheal illnesses are directly associated with unsafe water and inadequate sanitation services.

These illnesses frequently lead to hospitalizations, emergency room visits, and ongoing medical treatments, creating considerable economic strain on public health budgets. Furthermore, chronic exposure to contaminated water can lead to persistent health conditions, increasing long-term healthcare costs and reducing overall economic productivity.

Investments in water infrastructure, such as wastewater treatment facilities, have been demonstrated to significantly reduce public health expenditures by decreasing the incidence and prevalence of waterborne diseases. Comprehensive improvements in water quality management, combined with targeted healthcare interventions, are crucial for mitigating the economic impacts of water contamination. Addressing this issue not only enhances public health outcomes but also ensures more efficient allocation and utilization of healthcare resources in the region.

Given the complexity of environmental, institutional, and health dynamics within the Esmeraldas River basin, this study adopted a machine learning approach to explore predictive patterns of health vulnerability. Specifically, a binary logistic regression model was implemented, combined with the Synthetic Minority Over-sampling Technique (SMOTE) to address data imbalance. This methodology allowed for a structured analysis of how territorial

exposure, healthcare availability, and socio-economic conditions jointly influence public health outcomes linked to water contamination.

2. Theoretical framework and literature review

The relationship between water pollution and public health represents a critical global issue, particularly pronounced in developing regions such as Ecuador. Traditionally, environmental economics and epidemiology have been the primary fields addressing these challenges, relying heavily on econometric and statistical models to identify causal links and evaluate policy interventions. However, limitations often arise when these traditional methods confront complex datasets characterized by class imbalances or incomplete data.

Recent advances in computational power and data science have enabled the use of machine learning (ML) methods, offering promising alternatives for public health predictions and informed policy-making. In particular, supervised classification techniques, such as logistic regression enhanced by the Synthetic Minority Over-sampling Technique (SMOTE), have emerged as effective tools for predicting risk factors and disease prevalence. This research employs logistic regression and SMOTE to determine whether being located in a canton affected by river pollution significantly correlates with higher disease incidence beyond random chance. By leveraging ML techniques, the research aims to deliver accurate and actionable insights for targeting public health interventions and infrastructure investments more efficiently.

2.1. Public health impacts of water pollution: a review

The economic and social consequences of water pollution are widely documented in both global and Ecuadorian contexts. Hutton and Varughese (2016) demonstrate that inadequate water and sanitation services substantially increase health costs, particularly in developing regions.

Similarly, the WHO/UNICEF Joint Monitoring Programme (2024) emphasizes that safe water access remains one of the most cost-effective interventions to mitigate disease burdens globally. Ecuador's Ministry of Public Health highlights that approximately 88% of diarrheal diseases are linked to unsafe water, with greater impacts in areas lacking adequate sanitation infrastructure (MSP, 2019).

2.2. Machine learning and public health: classification models

Classification models constitute a supervised machine learning methodology used to predict categories or classes based on various input features. In public health contexts, logistic regression and similar classification algorithms have successfully forecasted disease incidence, predicted patient outcomes, and identified key risk factors. These methods efficiently handle multiple predictors simultaneously, capturing complex, non-linear relationships. Furthermore, they are particularly powerful when combined with resampling techniques designed to correct class imbalances, thus enhancing predictive accuracy (Brouwer, 2023).

Within this thesis, logistic regression is specifically used to classify cantons into high-risk or low-risk groups regarding waterborne disease prevalence. This granular analysis facilitates the identification of local patterns otherwise obscured by aggregated statistics. As noted by the Water Security and Sustainable Development Hub (2024), accounting for spatial variability in health-environment interactions is essential for designing effective, targeted interventions.

2.3. The problem of imbalanced health data and SMOTE

A significant challenge in applying ML to health datasets is the class imbalance problem, where the occurrence of disease (positive class) is significantly less frequent than its absence (negative class). Such imbalance can bias models towards high overall accuracy but low

sensitivity to true positives (Genius et al., 2006). To tackle this, SMOTE generates synthetic examples for the minority class through nearest-neighbor interpolation, thus balancing the dataset without losing valuable information (Lemaître, Nogueira, & Aridas, 2023).

As documented by imbalanced-learn (2024), SMOTE effectively enhances model sensitivity to policy-relevant outcomes, crucial when predicting the presence of high-risk health conditions associated with environmental factors like water contamination.

2.4. Comparison with traditional econometric approaches

While traditional spatial econometric models such as the Spatial Durbin Model (SDM) provide robust insights into spatial dependencies, they typically rely on stringent assumptions regarding data linearity, normality, and exogeneity. These assumptions often fail to hold in real-world health and environmental datasets. Additionally, traditional models might face interpretability challenges with complex or imbalanced data. In contrast, ML classification methods are specifically optimized for predictive accuracy and scalability, making them ideal for targeted policy actions, especially in contexts where causal inference is less feasible due to data limitations (Hutton, 2012; Water Security Hub, 2024).

The application of ML in this thesis thus complements traditional econometric analyses by providing actionable predictions under practical data constraints.

This section established the theoretical justification for employing logistic regression and SMOTE as an analytical framework for assessing public health risks linked to water pollution. Drawing from global evidence and methodological advancements, the chosen ML techniques align closely with Ecuador’s specific health data characteristics and infrastructure needs. The

subsequent chapter will detail the operationalization of this theoretical approach, including data processing, classification modeling, and result interpretation.

3. Methodology

This section presents the methodology used to analyze how territorial conditions related to pollution from the Esmeraldas River affect the prevalence of waterborne diseases in the studied cantons.

Methodology

3.1 Problem Definition and Research Design

This study addresses a binary classification problem aimed at determining how territorial exposure to pollution from the Esmeraldas River basin affects the prevalence of waterborne diseases in the cantons analyzed. The primary goal is to assess whether being located directly within or adjacent to the river basin significantly increases the likelihood of higher incidences of water-related diseases. Cantons directly intersected by the major rivers (Machángara, Guayllabamba, Blanco, and Esmeraldas) are classified as treatment cantons (coded 1), while geographically adjacent cantons constitute the control group (coded 0). Cantons neither directly affected nor adjacent were excluded to enhance spatial precision.

3.2 Data Sources and Sample Construction

Data for this analysis was collected primarily from two Ecuadorian government sources covering the period 2015–2023:

- **Ministry of Public Health (MSP)** provided detailed information on disease incidence and healthcare personnel at the cantonal level.

- **National Institute of Statistics and Censuses (INEC)** provided monthly cantonal population projections based on the 2022 Ecuadorian Census and poverty rates defined by the Unmet Basic Needs (NBI) metric.

Variables were standardized to a per capita basis, to allow meaningful comparisons across cantons with differing population sizes.

3.3 Variable Construction

The analytical model included the following variables:

- **Poverty Rate per Capita (pobres_percapita)**: Derived from INEC's poverty metrics.
- **Healthcare Personnel per Capita (funcionarios_medicos_relacionados_percapita)**: Aggregate measure including auxiliaries, doctors, nurses, rural doctors, rotating interns, laboratory technicians, and specialized medical personnel.
- **Primary Waterborne Diseases (enfermedades_asociadas_pc)**: Incidence per capita directly linked to unsafe water (e.g., Infectious Gastroenteritis, Intestinal Helminthiasis, Chronic Gastritis).
- **Secondary Water-Related Diseases (enfermedades_asociadas_2do_grado_pc)**: Incidence per capita indirectly associated with unsafe water and poor hygiene practices (e.g., Urinary Tract Infections, Cystitis).
- **Treatment Indicator (dummy_tratamiento)**: Binary variable distinguishing between directly affected (1) and adjacent (0) cantons.

Disease classifications:

- **Primary water-associated diseases**:

  - a09x – Infectious Gastroenteritis and Colitis, Unspecified
  - b829 – Intestinal Helminthiasis, Unspecified
  - k297 – Chronic Gastritis, Unspecified
  - k30x – Dyspepsia
  - z108 – Screening for Infectious and Parasitic Diseases

- **Secondary water-related diseases**:
  - n390 – Urinary Tract Infection, Site Not Specified
  - n300 – Cystitis, Unspecified
  - n760 – Female Pelvic Inflammatory Disease

These disease groups were selected based on their direct or indirect association with unsafe water consumption and hygiene practices.

3.4 Addressing Class Imbalance with SMOTE

Initial analysis revealed significant class imbalance, with treatment cantons representing approximately one-third of observations compared to two-thirds for adjacent cantons. To correct this bias, the Synthetic Minority Over-sampling Technique (SMOTE) was employed exclusively on the training dataset. SMOTE generates synthetic samples by interpolating existing minority class examples, balancing the dataset without data loss and ensuring robust model training (Lemaître, Nogueira, & Aridas, 2017).

Post-SMOTE, the training data achieved a balanced distribution (50% treatment, 50% control), improving the model's sensitivity to genuine spatial and epidemiological patterns.

3.5 Modeling Strategy: Logistic Regression

Given the binary nature of the dependent variable, logistic regression was selected for its interpretability and robustness in epidemiological and public health contexts (Water Security and Sustainable Development Hub, 2024). The model predicts the probability of cantons belonging to the treatment group based on socioeconomic and epidemiological predictors.

The logistic regression model was trained using the balanced SMOTE-enhanced training dataset and evaluated against an independent, untouched test dataset to provide an unbiased assessment of predictive accuracy.

3.6 Model Evaluation Metrics

The predictive capability of the logistic regression model was comprehensively evaluated using:

- **Confusion Matrix**: Precision, Recall, Specificity, Accuracy
- **F1-Score**: Assessment of the balance between precision and recall for each class
- **ROC Curve and Area Under the Curve (AUC)**: Evaluation of the model's discriminatory power

These metrics enabled a clear assessment of the model's ability to correctly identify cantons at elevated health risk due to territorial exposure to contaminated water sources.

3.7 Software and Tools

Data processing and analysis were conducted using Python, leveraging libraries such as pandas, scikit-learn, imbalanced-learn, matplotlib, and seaborn to ensure robust statistical analysis and clear visual representation of results.

This integrated methodological framework ensured that the study effectively identified and analyzed spatial health vulnerabilities, providing reliable insights into public health risks associated with water contamination in the Esmeraldas River basin.

4. Results

4.1 Descriptive statistics and historical context of healthcare personnel

Before presenting the results of the predictive classification model, it is important to understand the historical and structural context of healthcare personnel distribution in the cantons affected by the Esmeraldas River. Human resource availability is a critical factor influencing the capacity to prevent and respond to waterborne diseases.

Table 1 summarizes the average number of healthcare personnel per 10,000 inhabitants across cantons directly affected by the river and those classified as adjacent. The data reveal a modest advantage in staffing levels for affected cantons, although the differences are not uniformly consistent.

canton_type	mean_rate	min_rate	max_rate	std_rate
Adjacent	50.64730064	11.4818	203.583196	48.04387
Affected	34.67102191	1.53692	85.2987986	32.14527
Other	92.32565735	1.41106	1651.19096	160.7643

Table 1. *Average number of healthcare personnel per 10,000 inhabitants, affected vs. adjacent cantons.*

Figure 1 provides a visual distribution through boxplots of total healthcare personnel across cantons. While affected cantons generally exhibit slightly higher median values, a significant overlap exists between the two groups. This suggests that being geographically

affected by river pollution does not necessarily guarantee better or worse access to healthcare services, highlighting the importance of controlling for additional factors in the predictive modeling.

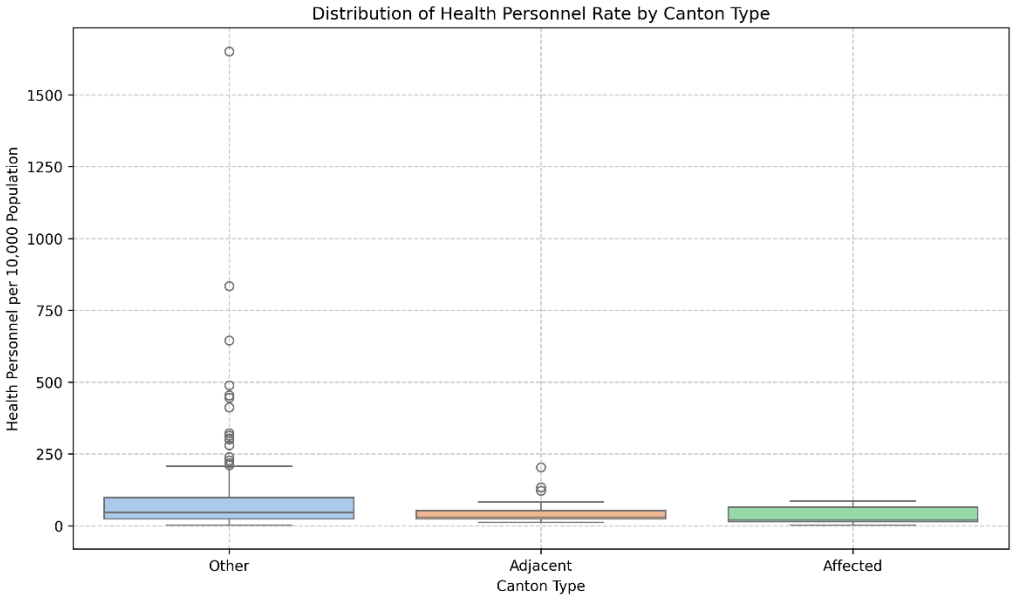

Figure 1. *Boxplot of healthcare personnel per 10,000 inhabitants by canton type.*

Understanding these baseline disparities is crucial because staffing levels directly influence the detection, reporting, and treatment of diseases linked to water contamination. However, healthcare resources in Ecuador have not remained static over time. Figure 2 depicts the temporal evolution of healthcare personnel density between 2015 and 2023.

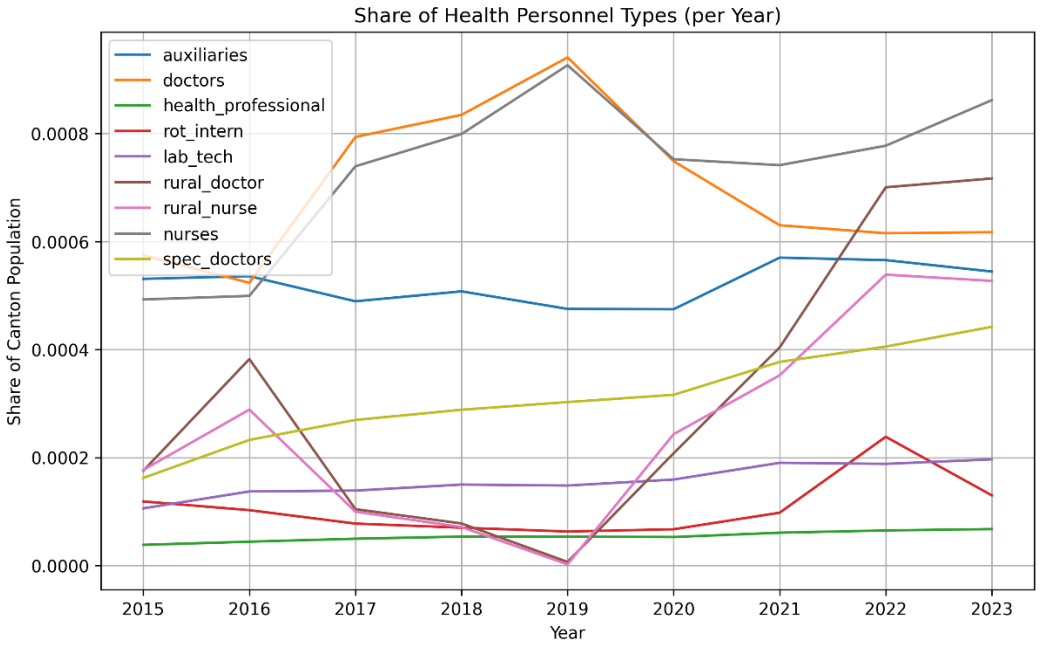

Figure 2. *Evolution of healthcare personnel density (per 10,000 inhabitants), 2015–2023.*

The figure highlights several key turning points. A marked contraction in healthcare staffing occurred between 2018 and 2019, aligning with fiscal austerity measures implemented under the Ecuador-IMF agreement (International Monetary Fund, 2019) and subsequent public sector layoffs (Ministry of Labor, 2019). The trend reversed slightly in 2021–2022, corresponding to emergency hiring initiatives during the COVID-19 health crisis (Ministry of Public Health, 2020).

These historical events contextualize the structural weaknesses and temporary reinforcements within the Ecuadorian healthcare system. As such, they frame the interpretation of both the descriptive patterns and the results of the predictive model that follows.

4.2 Results of the predictive classification model

The predictive model was trained using logistic regression on a dataset balanced with the Synthetic Minority Over-sampling Technique (SMOTE). The classification aimed to predict

whether a canton, based on its geographical relation to the Esmeraldas River and its healthcare personnel resources, would exhibit high incidence of waterborne diseases.

Table 2 presents the classification report metrics.

Class	Precision	Recall	F1-Score	Support
0 (non-affected)	0.83	0.83	0.83	1748
1 (affected)	0.66	0.65	0.66	877

Table 2. *Logistic regression model evaluation metrics.*

- **Accuracy:** 77%
- **Macro average F1-Score:** 0.74
- **Weighted average F1-Score:** 0.77

The confusion matrix (Figure 3) provides a detailed summary of the model's absolute performance, illustrating how correct and incorrect predictions are distributed across the two classes (affected vs. non-affected cantons). Despite some misclassification, the model demonstrates acceptable balance across classes, especially considering the original data imbalance (Lemaître, Nogueira, & Aridas, 2017).

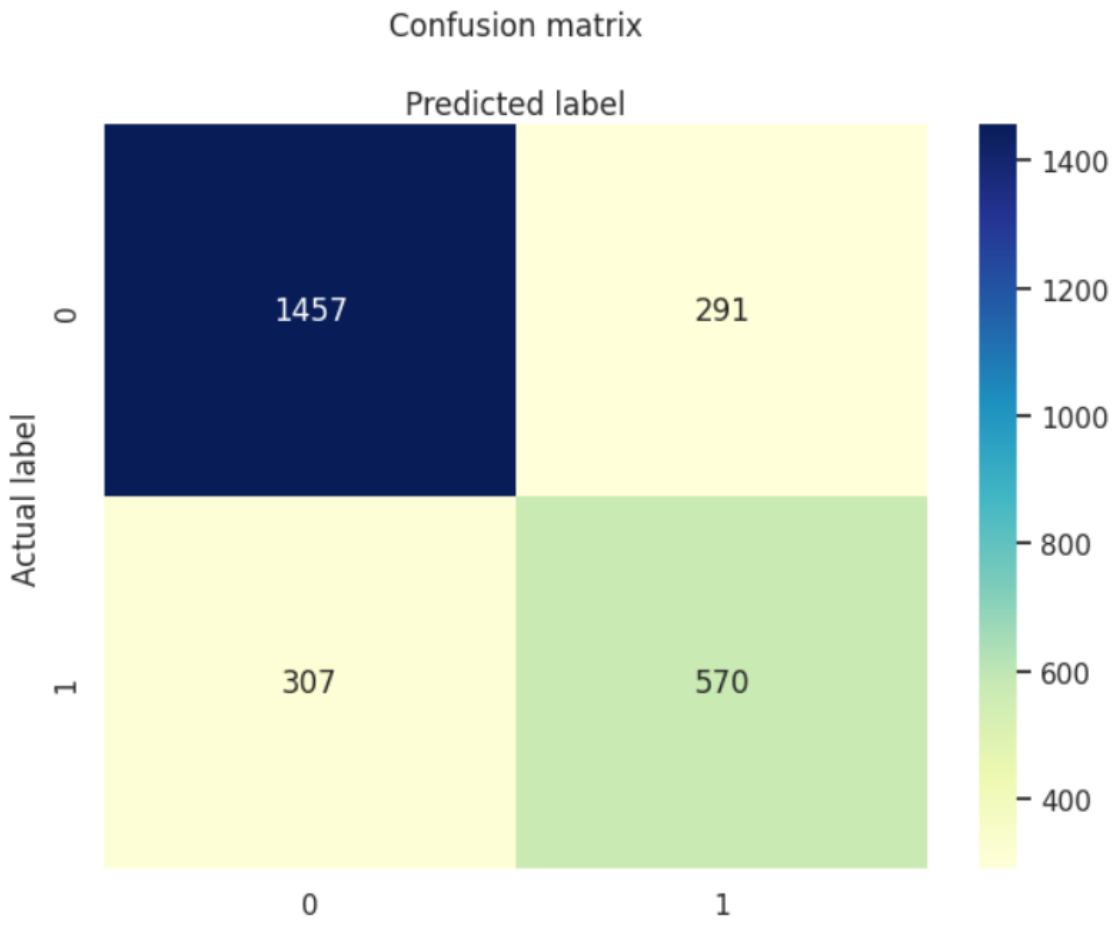

Figure 3. *Confusion Matrix of the Logistic Regression Model.*

- True Negatives (1457): Cantons correctly classified as non-affected.
- False Positives (291): Cantons incorrectly classified as affected.
- False Negatives (307): Cantons affected by river pollution but incorrectly classified as non-affected.
- True Positives (570): Cantons correctly classified as affected.

The confusion matrix highlights the model's strong ability to correctly classify non-affected cantons (high specificity), while also revealing moderate limitations in detecting all affected cantons (sensitivity). This result underscores the importance of continuing to refine the

model by incorporating additional variables and collecting higher-quality data, with the ultimate goal of moving toward causal analysis and informing effective public policy.

To further assess model performance, the Receiver Operating Characteristic (ROC) curve was analyzed (Figure 4).

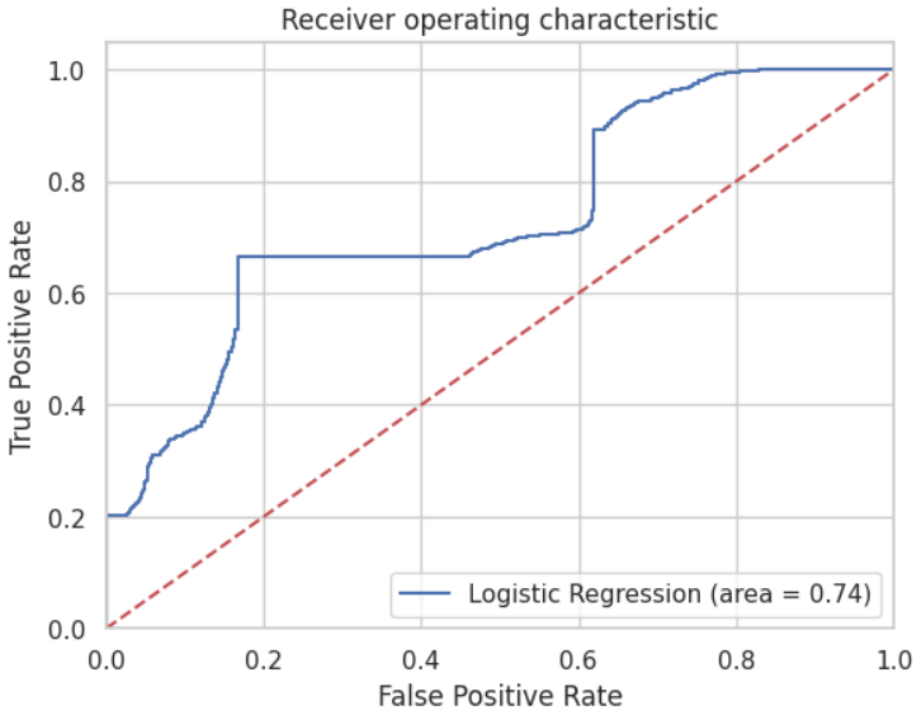

Figure 4. *Receiver Operating Characteristic (ROC) curve.*

The area under the ROC curve (AUC) reached 0.74, suggesting a satisfactory ability of the model to distinguish between high- and low-risk cantons. The area under the ROC curve (AUC) reached 0.74, suggesting a satisfactory ability of the model to distinguish between high- and low-risk cantons. Although not perfect, this level of discrimination substantially exceeds random prediction and validates the relevance of territorial exposure and healthcare capacity as predictive factors. And according to Hosmer, Lemeshow, and Sturdivant (2013), an AUC

between 0.7 and 0.8 indicates an acceptable level of discrimination, particularly in studies addressing complex health and social phenomena, where perfect prediction is rarely achievable.

Overall, these results indicate that territorial variables and healthcare staffing meaningfully predict the health vulnerability landscape within the Esmeraldas River basin.

4.3 Interpretation of findings

The findings from the descriptive statistics and predictive modeling reveal a complex but discernible relationship between territorial exposure to river pollution, healthcare resource availability, and the incidence of waterborne diseases.

Firstly, the slight advantage in healthcare personnel density among affected cantons suggests that proximity to environmental risks may be correlated with targeted resource allocation. However, the significant variability and overlap between affected and adjacent cantons highlight systemic disparities that cannot be fully explained by territorial factors alone.

Secondly, the predictive model's performance, achieving a 77% overall accuracy and an AUC of 0.74, underscores that spatial exposure and healthcare infrastructure are significant predictors, although not exhaustive. The imperfect recall for affected cantons indicates that other latent factors, such as environmental management, water treatment infrastructure, and social determinants of health, likely play substantial roles.

Thirdly, the historical disruptions in healthcare staffing, notably the austerity-driven contractions and pandemic-induced expansions, frame the healthcare system's vulnerability and resilience. These contextual factors likely influenced the observed patterns and model predictions.

In summary, while territorial exposure to river pollution is a meaningful predictor of public health risks, the results highlight the multifactorial nature of vulnerability. Effective interventions should therefore integrate environmental management, equitable healthcare resource distribution, and broader socio-economic policies to mitigate the health impacts of water contamination.

CONCLUSIONS AND RECOMMENDATIONS

Conclusions

This study aimed to assess the predictive capacity of territorial exposure to water pollution in the Esmeraldas River basin on the incidence of waterborne diseases, incorporating socio-economic and healthcare variables. The logistic regression model, combined with the Synthetic Minority Over-sampling Technique (SMOTE), provided robust evidence indicating that cantonal exposure to river contamination, healthcare infrastructure, poverty levels, and disease burden significantly contribute to predicting public health vulnerability.

The results highlighted that while territorial exposure alone does not determine disease outcomes, it is an important predictive factor when considered in conjunction with healthcare resources, poverty indicators, and specific water-related diseases. The model achieved an accuracy of 77% and an Area Under the Curve (AUC) of 0.74, underscoring the meaningful predictive power of these combined indicators.

Furthermore, historical contextual factors, including austerity measures under the Ecuador-IMF agreement and emergency hiring during the COVID-19 pandemic, significantly impacted healthcare resource availability and disease vulnerability patterns. These factors emphasized the dynamic nature of health resources and their crucial role in shaping public health outcomes.

Despite these valuable insights, this study did not establish causality, highlighting the need for additional research employing causal inference methods. Thus, while the findings strongly suggest non-random associations, definitive conclusions on causal relationships remain beyond this study's scope.

Recommendations for Future Research

Future research should pursue causal methodologies such as Difference-in-Differences (DiD), instrumental variable approaches, or quasi-experimental designs to confirm and refine the relationships identified here. Additionally, incorporating more detailed environmental management indicators, infrastructure quality metrics, and comprehensive socio-economic variables could further enhance the robustness and explanatory power of future studies.

Recommendations for Public Policy

Given the significant predictive associations found, public health policy in the Esmeraldas River basin should adopt an integrated, multifactorial approach. Policies should not only address water pollution but also systematically enhance healthcare access, sanitation infrastructure, and socio-economic conditions in affected and adjacent cantons. Prioritizing investments based solely on geographical exposure is insufficient; thus, policies must simultaneously tackle poverty alleviation, infrastructure development, and public health education.

Ultimately, this study provides foundational insights indicating the complexity and interconnectivity of environmental, socio-economic, and healthcare determinants of public health. Future policy interventions informed by these multifaceted relationships have the potential to substantially mitigate the health impacts of water pollution in Ecuador.

REFERENCES:

EPMAPS. (2016). *Proyecto Planta de Tratamiento de Aguas Residuales Vindobona*. Empresa Pública Metropolitana de Agua Potable y Saneamiento. https://www.aguaquito.gob.ec/

Fondo Monetario Internacional. (2019). *Ecuador: Acuerdo técnico en el marco del Servicio Ampliado del FMI (SAF)*. https://www.imf.org/es/News/Articles/2019/03/01/pr1970-ecuador-imf-reaches-staff-level-agreement

Genius, M., Menegaki, A. N., & Tsagarakis, K. P. (2006). Water shortages and implied water quality: A contingent valuation study. *Water Resources Research*, 42(12). https://doi.org/10.1029/2005WR004835

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression (3rd ed.). Wiley.

Instituto Nacional de Estadística y Censos (INEC). (2022). *Censo de Población y Vivienda 2022. Proyecciones y pobreza por Necesidades Básicas Insatisfechas (NBI)*. https://www.ecuadorencifras.gob.ec/

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, 18(17), 1–5. http://jmlr.org/papers/v18/16-365.html

Ministerio de Salud Pública del Ecuador. (2019). *Plan Nacional de Agua Segura y Saneamiento para Todos: 2019–2030*. https://www.salud.gob.ec/

Ministerio de Trabajo del Ecuador. (2019). *Proceso de optimización del talento humano en el sector público*. https://www.trabajo.gob.ec/

Reyes Vera, L., Vélez Rojas, A., & Encalada Meneses, K. (2022). Fragmentación de paisaje en la cuenca del río Esmeraldas: Implicaciones ecológicas. *Revista de Ciencias Ambientales*, 56(3), 25–42.

UDLA. (2024). *Contaminación del río Machángara: Análisis de la calidad de agua en Quito* [Contamination of the Machángara River: Water Quality Analysis in Quito]. Universidad de Las Américas. https://sitios.udla.edu.ec/

Vinueza, C., Menéndez, P., & Tapia, G. (2021). Análisis de la calidad del agua del río Guayllabamba en la ciudad de Quito. *Revista Ecuatoriana de Medio Ambiente y Agua*, 7(1), 56–65.

Water Security Hub. (2024). *Water security and sustainable development: Final impact case studies*. University of Newcastle. https://www.watersecurityhub.org/