Predictive Analysis of Missing Persons in Ecuador (2014-2024)

Marcos Espinoza-Mina, Alejandra Colina-Vargas*

Universidad Ecotec, Samborondón, Ecuador

Abstract

The growing issue of missing persons in Ecuador demands a data-driven approach to strengthen institutional prevention and search strategies. This study develops a predictive analysis framework using an official dataset of 68,072 missing persons records from 2014 to 2024. Applying the CRISP-DM methodology, we implemented classification and anomaly detection models to identify patterns and risk factors. To address a severe class imbalance, the SMOTE technique was applied, resulting in a robust XGBoost classifier. The model achieved an overall accuracy of 88.1% and, more importantly, a substantial improvement in the detection of minority outcomes, with recall values of 0.28 for "Missing" and 0.33 for "Deceased." Beyond predictive performance, the SHapley Additive exPlanations (SHAP) analysis conclusively identified distinct risk profiles for each outcome: short reporting delays strongly predict a "Found" outcome, while long delays increase the likelihood of a case remaining "Missing." For the "Deceased" class, older age emerged as the dominant predictor. This study establishes a fundamental quantitative baseline for the analysis of missing persons, demonstrating that machine learning can generate actionable intelligence for resource prioritization, especially when augmented with interpretability techniques and anomaly detection.

Keywords

Predictive Analytics, Missing Persons, Ecuador, Risk Factors, Model Interpretability, Class Imbalance

1. Introduction

The disappearance of people is a global problem that causes suffering and uncertainty, affecting millions of people worldwide, including children [1]. This phenomenon, often linked to human trafficking, violates human rights and has serious consequences for health [2]. Historically, forced disappearances have been used for intimidation, and the IACHR (Inter-American Commission on Human Rights) has expressed concern about their persistence in Latin America and the prevailing impunity [3]. The inability to know the whereabouts of a loved one causes distress to family and friends [4].

In Ecuador, the increase in missing persons over the last five years has revealed deficiencies in security, justice, and human rights [5, 6]. Adolescents, young people, and children are the most vulnerable [7]. Although women account for 56.7% of the cases, men constitute 80% of the unresolved or fatal ones [8]. Pichincha and Guayas concentrate 46% of the cases [4]. Despite a resolution rate of 94.0% of cases being found, 3.1% ending in deceased persons, and 2.8% remaining unresolved, the persistence of these last categories represents the central challenge [7]. The average distance between the place of disappearance and discovery is 50.69 km, with provincial disparities in resolution rates, such as Pichincha (91.02%) versus Esmeraldas (61.28%) [6].

Families of missing persons suffer a devastating psychological and economic impact, with 70.6% unable to resume their routines. There is low social awareness (47.1%) in Ecuador about this issue [5]. Public institutions, such as the Prosecutor's Office, the Ministry of Government, and the National Directorate of Crimes Against Life, Violent Deaths, Disappearances, Extortion, and Kidnapping (DINASED), are perceived as having low performance. The 24-hour delay in reporting is criticized by 64.7% of respondents, who highlight the importance of the first few hours. An uneven state response has been observed, with greater speed in high-profile media cases [3]. The main causes are family and social

ICAIW 2025: Workshops at the 8th International Conference on Applied Informatics 2025, October 8–11, 2025, Ben Guerir, Morocco

*Corresponding author.

mespinoza@ecotec.edu.ec (M. Espinoza-Mina); acolina@ecotec.edu.ec (A. Colina-Vargas)
0000-0003-1530-7243 (M. Espinoza-Mina); 0000-0003-1514-8852 (A. Colina-Vargas)

problems [4]. This underscores the need for robust information systems, standardized search protocols, and early warnings [6].

Artificial Intelligence (AI) and Machine Learning (ML) are promising solutions for public safety management [9]. AI, through predictive analysis, can improve the efficiency and decision-making in the public sector [10]. In the context of disappearances, AI and ML are advancing in the study and prediction of social problems [11]. Deep learning techniques can offer more accurate predictions [12], and models like Random Forest have improved the recall rate in location prediction [2]. ML can also optimize humanitarian aid [13]. However, the implementation of AI faces challenges such as algorithmic bias and data privacy, requiring solid ethical frameworks and a human-centered design [14, 15].

This study addresses the challenge of missing persons in Ecuador using a machine learning framework on historical data from 2014 to 2024. Critically, we tackle the inherent class imbalance of disappearing data using the SMOTE technique to improve the prediction of minority outcomes. Furthermore, we employ SHAP (SHapley Additive exPlanations) interpretability analysis to identify the key risk factors driving each prediction ('Found', 'Missing', or 'Deceased'). The objective is to generate data-driven insights to help authorities optimize search protocols and prioritize high-risk cases [9, 16].

2. State of the art

2.1. Definition and types of disappearances

Disappearances in Ecuador are a serious social problem that requires effective solutions, with a notable increase in cases in recent years [5, 17]. Ecuadorian legislation, through the Organic Law on Action in Cases of Disappeared and Missing Persons (2020), classifies disappearances as voluntary (by personal decision) and involuntary (caused by third parties and subject to police investigation) [5, 4]. Since June 2020, involuntary disappearance has been an autonomous crime in the Comprehensive Organic Penal Code (COIP) [18].

On the other hand, forced disappearance is defined in the COIP (article 84) as an act under state command, executed by agents or armed groups [5]. The Inter-American Convention on Forced Disappearance of Persons (1994) characterizes it as the deprivation of liberty by state agents or their collaborators, followed by a lack of information on the person's whereabouts [18, 3]. This act is a crime against humanity and violates fundamental rights [3].

2.2. Historical context and prevalence

Forced disappearance originated in the Second World War to intimidate and conceal the fate of detainees, arriving in Latin America in the 1960s [3]. In Mexico, the phenomenon grew with the "War on Drugs" (2006-2012), generating a concept of "disappeared person" that exceeds the traditional definition of forced disappearance [19]. In Ecuador, cases of forced, voluntary, and involuntary disappearances have been documented [5]. The Truth Commission of Ecuador registered 17 forced disappearances between 1984 and 2008 [18, 3]. Between 1970 and 2017, the country recorded 42,484 missing persons, with an average of 500 reports per month. Quito was the most affected city (34% of cases), and between 2014 and 2017, 59% of the victims were women [17].

2.3. The role of the State and legal framework

The Ecuadorian State must protect the rights of victims and their families [1, 20], establishing a normative framework with international declarations and conventions, in addition to national laws such as the Constitution and the COIP [4, 7]. State responsibility for damages caused by its agents has evolved, recognizing the obligation to repair for actions or omissions [3]. Ecuador's 2008 Constitution guarantees victims of crimes special protection and comprehensive reparation, including the right to truth, restitution, compensation, rehabilitation, and non-repetition [7, 3]. The right to truth implies the

State's obligation to investigate, judge, and sanction those responsible, ensuring access to information [3].

The Attorney General's Office and the National Police, through the DINASED (created in 2013), are the main entities in charge of locating missing persons and recovering remains [4, 3]. Between 2014 and June 2021, DINASED located 37,258 people. The Council of the Judiciary has implemented a Protocol for Action in the Search, Investigation, and Location of Missing Persons, establishing procedures for the National Police [4, 3].

2.4. Imprescriptibility and statute of limitations for crimes

Article 80 of the Ecuadorian Constitution declares that legal actions and penalties for crimes such as genocide, crimes against humanity, war crimes, forced disappearance, and aggression are imprescriptible, prohibiting their amnesty. It establishes criminal responsibility for both perpetrators and those who ordered the crime [18]. Forced disappearance is imprescriptible because it is a grave violation of human rights and a crime against humanity [21].

In contrast, the COIP (Art. 417, num. 3, literal d) indicates that the statute of limitations for the common crime of disappearance begins when the person appears or when there are elements to charge the crime. This creates a potential constitutional conflict if it is considered imprescriptible without being included in the Art. 80. The statute of limitations for criminal offenses seeks to guarantee legal certainty, limit criminal action, and prevent indefinite prosecution, contributing to legal stability [18].

2.5. Impact on families and society

The disappearance of a person causes immense suffering to their families, who face uncertainty, constant grief, and economic and health problems. Organizations like ASFADEC (Association of Families of Disappeared Persons in Ecuador), founded in 2012, provide support [5]. In Mexico, the concept of "disappeared alive", a term mainly used in the Mexican context, refers to individuals without an official record or civil existence, who are vulnerable to dangers like human trafficking or slavery. In this context, the search is vital to grant them "civil existence" and make visible that "there are lives that matter" [19].

2.6. Challenges in investigation and search

Families of missing persons often perceive deficiencies in public institutions (the Prosecutor's Office, Ministry of Government, DINASED), pointing to a lack of support, knowledge, and prejudice. The 24-hour waiting period to file a report is seen as a critical obstacle, as the first few hours are vital. Frequent changes of prosecutors and investigators, together with disrespectful treatment, re-victimize the complainants and violate their rights [5]. Traditional media often do not prioritize missing persons cases, which forces families to seek alternative channels [5]. While media exposure can help gather information, it can also negatively influence judicial decisions and the search for the truth [22].

2.7. Tools and strategies for search and visibility

The dissemination of information about missing persons through social media is fast and effective, especially for minors. Ecuador has implemented programs like the "Alerta Emilia," supported by the International Centre for Missing & Exploited Children (ICMEC), highlighting social collaboration in the recovery of children [4].

Complementing this dissemination, the analysis of large volumes of data and predictive models are crucial [1, 23]. Neutrosophic multicriteria analysis is used to handle data uncertainty [6], and platforms like datosabiertos.gob.ec are relevant sources [8]. Additionally, Deep Learning models, such as TextRNN, predict the locations of missing persons, improving accuracy with the inclusion of oral information [2]. To expand sources and improve data quality, databases with economic indicators, social problems, and humanitarian organizations are used [1, 20, 24].

A practical example is the QSSC (Quito Smart Safe City) prototype, a distributed mobile system with IoT, Crowdsensing, and cloud computing for alerting and gathering evidence. Simulations in Quito showed an average resolution time of 34.2 minutes, underscoring the importance of citizen collaboration [17].

2.8. International cooperation and human rights

The Inter-American Commission on Human Rights (IACHR) has expressed concern about the persistence of forced disappearances in Latin America, highlighting the lack of investigation and impunity. The Inter-American Court of Human Rights (IACtHR) emphasizes the state's obligation to investigate human rights violations, as negligence violates the American Convention on Human Rights (ACHR) [3].

Cases such as Barrios Altos and La Cantuta in Peru illustrate systematic extrajudicial executions and forced disappearances. The IACtHR's rulings have been key to invalidating amnesty laws and revoking pardons, reiterating the state's obligation to investigate, prosecute, punish, and provide reparations. The control of conventionality, implemented by the IACtHR and national judges, ensures the conformity of domestic laws with the ACHR and international law [21].

This overview highlights the complexity of disappearances [7, 8] and state deficiencies, characterized by a lack of support, knowledge, and prejudice. Inefficiency is evident in the absence of protocols, frequent staff turnover in prosecutorial and investigative bodies, and the disrespectful treatment that re-victimizes complainants. Despite this, the resilience of families is notable, with organizations like ASFADEC providing support [5]. Effective cooperation and the use of technology are crucial [6, 17]. Prototypes like QSSC demonstrate that citizen collaboration can significantly reduce case resolution time [17].

3. Methodology

This study was developed following the CRISP-DM (Cross-Industry Standard Process for Data Mining) model, a systematic and rigorous approach. All phases, from business understanding to implementation, were executed using Python to ensure traceability, reproducibility, and efficiency in data processing, analysis, and modeling [25, 26, 27].

3.1. Business understanding

The disappearance of people in Ecuador is a complex problem with a high social, institutional, and humanitarian impact [7, 8, 6, 17]. Its visibility has increased due to the rise in cases and pressure from organizations, media, and families [5, 18]. This study aims to generate data-driven intelligence to support decision-making in the prevention, search, and resolution of cases [1, 28, 8].

The problem is broken down into several key sub-problems: Lack of predictive capability to anticipate the evolution of a case [1, 23]. Limited understanding of the factors that influence the duration and outcome of disappearances [1, 29, 2].The need to identify patterns to distinguish between routine disappearances and atypical cases [1, 2, 12, 23]. The absence of analytical tools to optimize resources in investigation and search efforts [11, 25, 24, 23].

The research questions guiding the study are: Is it possible to accurately predict the resolution status (FOUND, MISSING, or DECEASED) of a case based on the initial report? What demographic (age, gender, ethnicity, nationality), geographic (area, province, canton), and contextual (date, motive, police sub-circuit) factors are most relevant to the duration and resolution of the case? How can atypical or anomalous cases requiring specialized attention be identified?

The impact of this study will benefit multiple institutional stakeholders [9, 30, 23]. For law enforcement, predictive models will offer early warnings, resource prioritization, and risk profiles [1, 2, 30]. For public policy makers, the findings can inform decisions on prevention, budget allocation, and protocol design [28, 2, 9, 8, 6, 27]. For civil society and families, data-driven models can improve transparency, efficiency, and institutional trust, contributing to a more timely and effective response [10, 5, 28, 9, 16, 8, 6, 17], to this urgent social phenomenon [8, 17].

3.2. Data understanding

The dataset was obtained from the Open Data portal of Ecuador, specifically from the Ministry of the Interior [7, 22, 8, 6]. It includes historical records of missing persons in Ecuador between 2014 and 2024, which guarantees its institutional validity. The main database contains variables such as: zona, provincia, cantón, distrito, circuito, subcircuito, sexo, nacionalidad, edad_aproximada, rango_edad, etnia, fecha_localizacion, motivo_desaparicion, motivo_desaparicion_observada, situacion_actual, dias_solucion, latitud_desaparicion, longitud_desaparicion, fecha_denuncia y fecha_desaparicion. These variables cover geographic, sociodemographic, and temporal dimensions, facilitating a multifactorial analysis and serving as a robust foundation for descriptive, inferential, and predictive studies.

3.3. Data preparation

Data preparation is crucial in data science, as it transforms raw data to maximize the performance of machine learning algorithms [31, 32, 33, 23]. In this study, the data underwent a series of systematic transformations (cleaning, structuring, and enrichment) using Python scripts to ensure the replicability and traceability of the workflow [25, 34, 35, 32, 33, 27].

3.3.1. Initial cleaning and field name translation

A Python script was developed to standardize the dataset, systematically translating column names from Spanish to English based on the official data dictionary. The headers of the input file (1.csv), which had an irregular structure with column names in the second row, were reconfigured, and irrelevant rows and columns were removed. Then, a mapping dictionary was applied to translate key fields such as zona, sexo, edad, motivo_desaparicion, and estado_desaparecido to zone, gender, age, disappearance_motive, and current_status, respectively. The resulting file, 1_translated.csv, has standardized field names for compatibility with international analysis and modeling tools while preserving data integrity.

3.3.2. Mapping and standardization of categorical values

After standardizing column names to English, the values of the categorical variables were homogenized. Minor inconsistencies were identified in the original Spanish values (e.g., "ENCONTRADO" and "ENCONTRADA"). To ensure consistency and prepare the data for international analysis and modeling, a comprehensive mapping was implemented.

First, an inspection script (03-mapeo-cat.py) extracted and counted the unique values of categorical columns such as gender, ethnicity, nationality, current_status, age_range, and the disappearance motives (disappearance_motive, observed_motive). This analysis allowed for the creation of a "translation dictionary" (a Python dictionary named translation_maps).

Subsequently, this dictionary was applied to translate and standardize categories; for example, 'HOMBRE' and 'MUJER' were mapped to 'Male' and 'Female', and 'MESTIZO/A' and 'INDIGENA' were mapped to 'Mestizo' and 'Indigenous'. This process also handled null values, assigning them 'N/A' (Not Applicable/Not Available) for consistent handling.

3.3.3. Integrated pipeline for exploratory processing and analysis

An automated Python pipeline (full_pipeline_with_export.py) was developed to integrate cleaning, exploratory analysis, feature engineering, and final preparation for modeling. The process began with cleaning and transforming data types, ensuring the dataset's structural integrity by converting geographical coordinates to floats and standardizing missing data to np.nan. Date columns were converted to datetime objects for temporal analysis.

Next, an Exploratory Data Analysis (EDA) was performed to validate the data quality and quantitatively understand the phenomenon. Descriptive analysis of numerical variables showed an average age of disappearance of 22.9 years. The high completeness of the dataset, with only 2.8% of missing data in key columns like location_date and days_to_resolution, justified not applying complex imputation. Frequency distributions of the categorical variables revealed important demographic findings: The most affected age group is Adolescents (34,495 cases; 50.7%), followed by Adults (26,710; 39.2%), highlighting the vulnerability of the young population (Figure 1). A notable gender disparity was found, with 43,090 reported cases of women compared to 24,982 of men. Regarding case resolution, 94.0% were "Found," 3.1% "Deceased," and 2.8% remain "Missing."

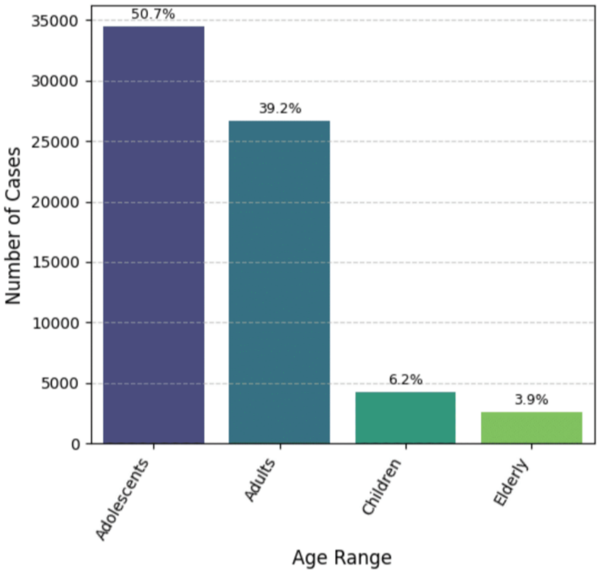


Figure 1: Distribution of age range.

The spatial and temporal dimensions were investigated through visualizations. The analysis of the annual time series of reports (Figure 2) shows a growing trend of cases over the decade, with a peak in 2023. For the spatial analysis, an interactive map (Figure 3) identified clusters of high incidence in the country's densest urban areas.

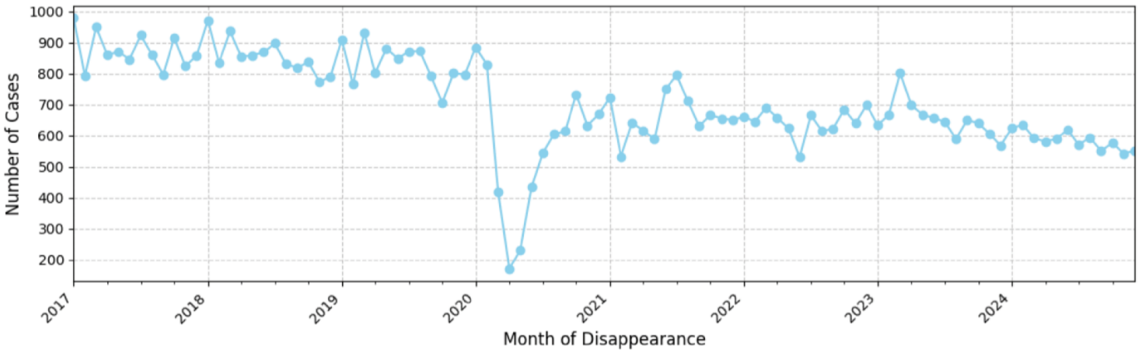


Figure 2: Disappearances over time (Monthly).

The feature engineering phase enriched the dataset with predictive variables. The disappearance_duration_days (time elapsed until location) and report_delay_days (delay in reporting the case) were calculated. Temporal features such as day_of_week and disappearance_quarter were extracted to capture cyclical patterns.

Finally, the dataset was prepared for modeling by transforming all features into a numerical format. One-Hot Encoding was applied to nominal categorical variables (gender, ethnicity) and ordinal

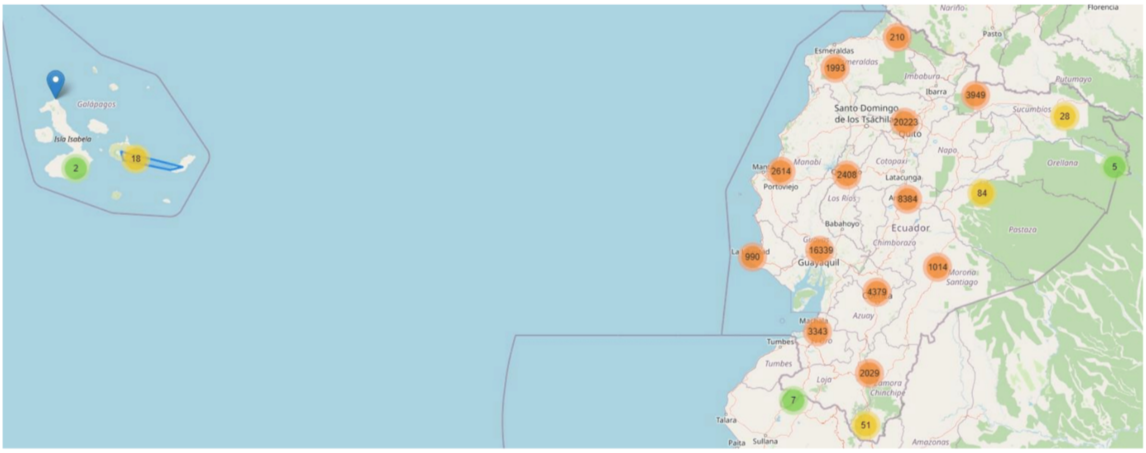


Figure 3: Geographic distribution of missing person cases in Ecuador.

encoding to age_range. All numerical features underwent standard scaling (StandardScaler) to normalize their distributions, thus optimizing the performance of the machine learning algorithms. The processed DataFrame was exported to 2_data_processed_for_modeling.csv, which served as the direct and standardized input for the Modeling phase.

3.3.4. Handling Class Imbalance with SMOTE

The exploratory data analysis revealed a severe class imbalance, with over 94% of cases belonging to the "Found" category. To address this issue, which hinders the model's ability to learn from minority classes ('Missing' and 'Deceased'), the SMOTE (Synthetic Minority Over-sampling Technique) was applied. This technique was implemented only on the training dataset to prevent data leakage, generating synthetic samples for the minority classes to create a balanced class distribution. This step is crucial for training a robust classifier capable of identifying high-risk cases.

3.4. Modeling

The modeling phase focused on two computational tasks: classification and anomaly detection [14, 25, 2, 32, 27]. Multiple algorithms were evaluated for each task. The processed dataset was split into a 70% training set and a 30% test set [1, 11, 36, 2, 37, 12, 33], using stratification on the target variable for classification to ensure a proportional representation of all classes [25, 2, 32].

3.4.1. Classification models: Predicting the final status

The main objective was to predict the final status of a case: "Found," "Missing," or "Deceased." After applying SMOTE to the training data, Random Forest, XGBoost, and a Multilayer Perceptron (MLP) were evaluated. The XGBoost model was selected as it offered the best balance between performance and interpretability, achieving a macro-averaged recall of 0.51. This indicates a much more balanced and practically useful model, correctly identifying 28% of "Missing" cases and 33% of "Deceased" cases in the test set.

While the overall accuracy is slightly lower than in the unbalanced model, the macro average recall (which measures the average recall across all classes) improved dramatically, from 0.36 to 0.51 for the XGBoost model. This indicates a much more balanced and practically useful model. The XGBoost model correctly identified 28% of "Missing" cases and 33% of "Deceased" cases in the test set, a substantial improvement over the pre-correction performance.

3.4.2. Anomaly detection

Anomaly detection was performed using an Isolation Forest model to identify atypical cases. A descriptive analysis of these anomalies revealed that they are predominantly driven by family-related motives (representing 55% of the anomalies), frequently involve individuals with disabilities or illnesses, and exhibit highly variable and prolonged resolution times, with a maximum recorded delay of 1812 days.

3.4.3. Model selection and storage

The XGBoost Classifier was selected as the best-performing model due to its superior balance in recall across all classes after the application of SMOTE to address the severe class imbalance. This improved model was saved as best_classification_model.joblib for the subsequent evaluation phase.

3.5. Evaluation

The evaluation phase focused on measuring the performance of the selected XGBoost model and interpreting the factors driving its predictions. The performance of the model was evaluated on the test set, yielding an overall accuracy of 88.1%. While this metric is slightly lower than that of the unbalanced model, a more granular analysis demonstrates the model's superior practical utility. The Confusion Matrix and detailed Classification Report confirmed the model's enhanced capability to detect critical minority classes, achieving a recall of 0.28 for "Missing" and 0.33 for "Deceased". This represents a substantial improvement in identifying high-risk cases, directly addressing the primary limitation of the initial models.

Additionally, the model's distinction capability was evaluated using Receiver Operating Characteristic (ROC) Curves and the calculation of the Area Under the Curve (AUC), using a one-vs-rest approach. This technique provides a performance measure that is insensitive to class imbalance, offering a more reliable view of the classifier's discriminative effectiveness.

For model interpretability, the SHAP technique was employed. This advanced method quantified the contribution of each feature to the model's predictions. The analysis was expanded to all three outcome classes ('Found', 'Missing', and 'Deceased') to identify the distinct factors influencing each status. This comprehensive approach allowed for a deeper understanding of how variables like reporting delay and age impact the likelihood of each specific outcome, providing valuable insights for operational decision-making. The detailed results of these evaluations (including the Confusion Matrix, ROC Curves, and SHAP plots) will be presented in Section 4

These findings have direct operational implications. The identified risk factors, particularly reporting delays and age, can be embedded into institutional early-warning dashboards, enabling law enforcement to prioritize high-risk cases in real time. By integrating the predictive model into decision-support systems, agencies such as DINASED could generate alerts that guide resource allocation, accelerate search protocols, and ultimately improve the effectiveness and timeliness of responses to disappearances.

4. Results

This section presents the empirical findings of the study, derived from the exploratory analysis and the execution of predictive models on the dataset of missing persons in Ecuador between 2014 and 2024. The results are presented through a descriptive analysis, the evaluation of model performance, and an interpretation of the most influential factors.

4.1. Findings from Exploratory Data Analysis (EDA)

The initial analysis of the 68,072 records revealed significant demographic, temporal, and geographic patterns.

4.1.1. Demographic profile

The gender analysis shows a marked preponderance of cases involving women, who account for 63.3% of the total (43,090 cases), compared to 36.7% for men (24,982 cases). Regarding the age group, adolescents (12-17 years old) are the most vulnerable, accumulating 50.7% of the reports (34,495 cases), followed by adults (26,710 cases, 39.2%). The average age of a missing person is approximately 23 years. The predominant ethnicity is Mestizo, representing 86.7% (59,028 cases) of the records with ethnic information. Regarding nationality, most cases correspond to Ecuadorian citizens (96.6%), although there is a significant presence of Venezuelans (1,156 cases) and Colombians (640 cases), reflecting migratory dynamics.

4.1.2. Motives and resolution status

The most common disappearance motive is "Family Reasons," accounting for 69.6% of cases. It is crucial to note that the vast majority of cases (94.1%) are resolved with the person being "Found." However, 2.8% of cases (1,907 records) remained in a "Missing" status at the time of data extraction, representing the core of the persistent disappearances problem.

4.1.3. Temporal and geospatial patterns

The temporal analysis (Figure 4) shows a fluctuating dynamic in the number of annual reports, with the highest levels between 2017 and 2019, exceeding 10,000 reports in that period. Since 2020, there has been a significant decrease, followed by a partial recovery, without reaching the initial peaks. The monthly analysis did not reveal a clear seasonal pattern.

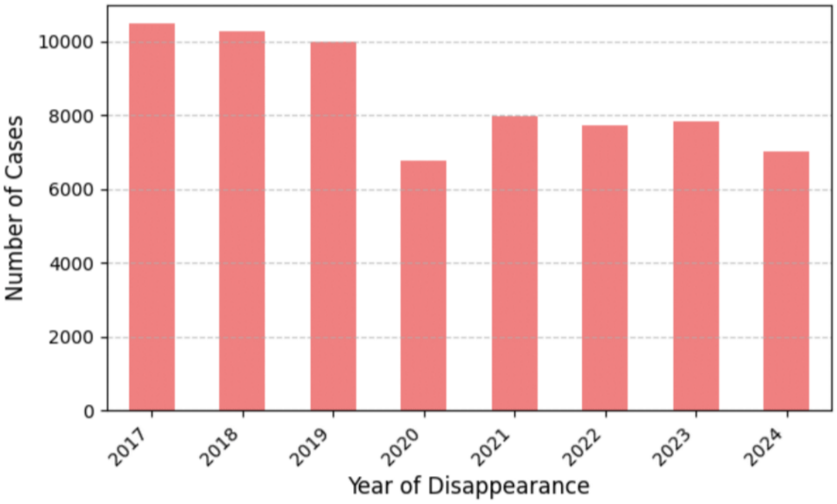


Figure 4: Disappearances over time (Yearly).

The geospatial analysis indicates a non-uniform distribution. The provinces of Pichincha and Guayas are the main "hot spots," with the highest density of incidents, which is consistent with their high population density.

4.2. Predictive model performance after addressing class imbalance

To overcome the severe class imbalance identified in the EDA, the SMOTE technique was applied to the training data. The re-evaluated XGBoost classifier was selected as the best-performing model, achieving an overall accuracy of 88.14% on the test set. The macro-averaged metrics were a precision of 0.44, a recall of 0.51, and an F1-score of 0.46. While the overall accuracy is slightly lower than that of the unbalanced model, the significant increase in macro-averaged recall demonstrates a superior ability to identify the critical minority classes. The confusion matrix (Figure 5) visually confirms this enhancement.

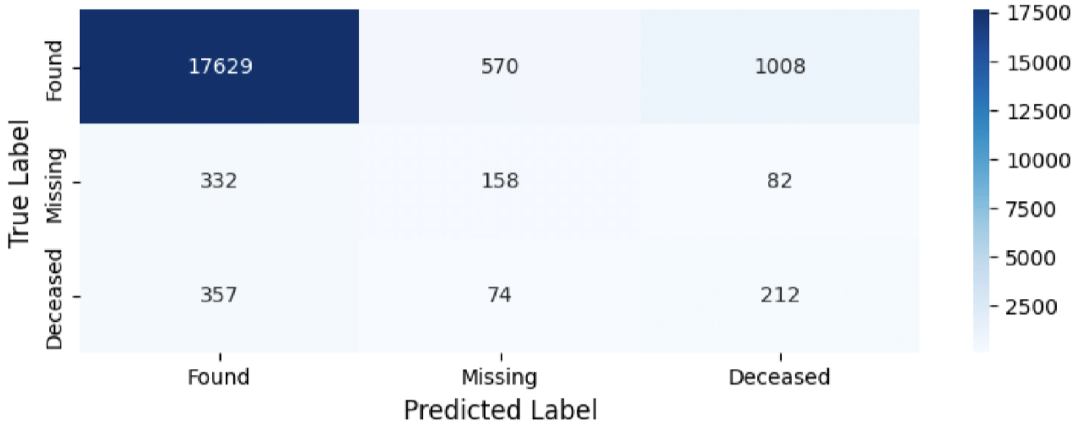


Figure 5: Confusion matrix for XGBoost classifier.

The model now correctly identifies 28% of "Missing" cases (158 instances) and 33% of "Deceased" cases (212 instances), a substantial improvement that makes the model practically useful for risk assessment. Furthermore, the model's discriminative ability is confirmed by the high Area Under the Curve (AUC) scores for all classes, as shown in the ROC curves (Figure 6).

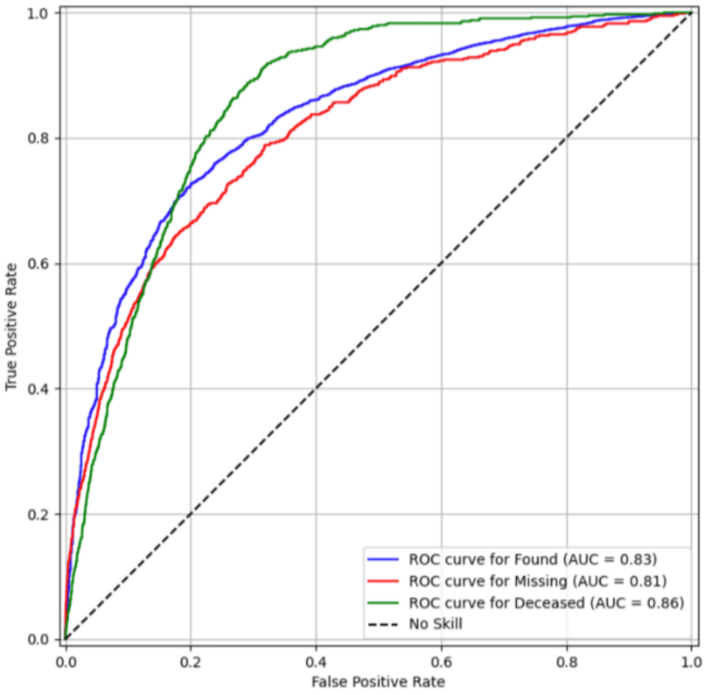
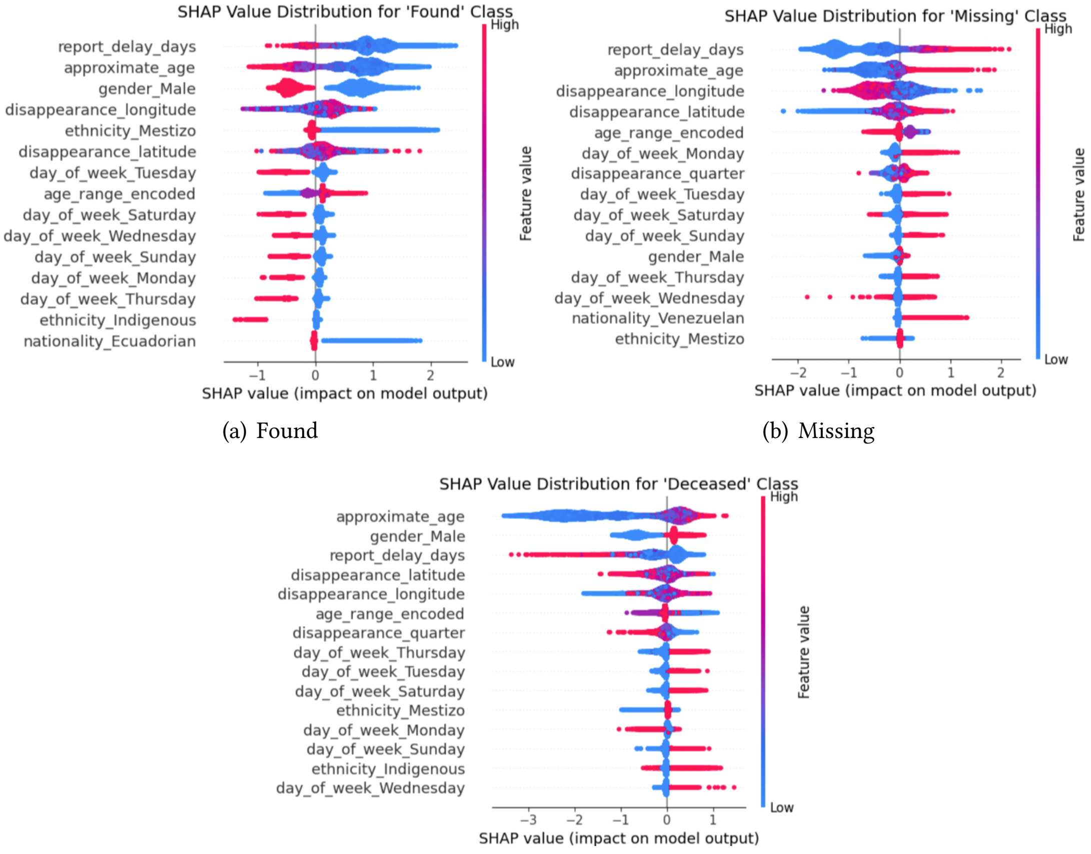


Figure 6: ROC Curves for each class of the corrected XGBoost model.

4.3. Model interpretability and risk factor analysis (SHAP)

To understand the key factors driving the model's predictions, a SHAP analysis was conducted for all three outcome classes. This approach reveals distinct risk profiles for each status by quantifying the impact of each feature on the final prediction (Figure 7).

The SHAP analysis reveals distinct risk profiles for each outcome. The variable report_delay_days



(a) Found

(b) Missing

(c) Deceased

Figure 7: SHAP value distribution showing the impact of the top features for the classes.

emerges as a critical predictor with a dual role: low values strongly predict a 'Found' outcome, indicating that cases reported quickly are more likely to be resolved successfully. Conversely, high values for the same feature are the strongest predictor that a case will remain 'Missing'. For the 'Deceased' class, however, the dominant factor shifts to approximate_age, where older individuals who go missing have a markedly higher probability of a fatal outcome. These findings provide actionable intelligence, underscoring the critical importance of immediate reporting and suggesting that age should be a primary factor in assessing the risk level of a missing person's case.

4.4. Analysis of atypical cases

The Isolation Forest algorithm was applied to the dataset, identifying 671 records (0.99%) as atypical. A descriptive analysis of these cases revealed distinct characteristics compared to the general population. The most frequent motive among anomalies was "Family Reasons," and these cases exhibited highly variable and prolonged resolution times, with a maximum recorded delay of 1,812 days. This suggests that while the primary model captures general trends, these anomalous cases represent unique dynamics or potential data entry issues that warrant separate, targeted investigation.

5. Conclusions

This study addressed the complex phenomenon of missing persons in Ecuador (2014–2024), developing a predictive framework to support institutional decision-making. The main contribution lies in the creation and evaluation of a comprehensive analytical pipeline, from data cleaning and enrichment to the development of classification and anomaly detection models. It generated the first predictive models for this context, quantifying underlying patterns and identifying both the potential and the limitations of using administrative data to model a complex social problem.

The findings are multifaceted. On one hand, the exploratory analysis of 68,072 records confirmed key demographic and temporal patterns: adolescents (50.7%) and women (63.3%) are the most vulnerable groups, with most cases concentrated in Pichincha and Guayas. Although 94.0% of cases are resolved with the person found, a persistent 2.8% remain unresolved, representing the central challenge. On the other hand, the results of predictive modeling reveal progress and challenges. After addressing the severe class imbalance with the SMOTE technique, the XGBoost classifier achieved an accuracy of 88.1% and, more importantly, a substantial improvement in the detection of minority outcomes, with recall values of 0.28 for "Missing" and 0.33 for "Deceased." These results demonstrate that predictive analytics can support risk assessment and provide actionable intelligence, even when accuracy alone may appear limited.

Beyond predictive performance, the study highlights valuable practical implications. The SHAP interpretability analysis revealed distinct risk profiles: short reporting delays strongly predict a "Found" outcome, long delays increase the likelihood of remaining "Missing," and older age is the dominant predictor of a "Deceased" outcome. These findings can be operationalized into early-warning dashboards that prioritize high-risk cases in real time, guiding resource allocation, accelerating search protocols, and improving the timeliness of institutional responses. Likewise, anomaly detection using Isolation Forest identified a small but significant set of atypical cases that differ from the general population, often involving extreme ages, uncommon motives, or unusually prolonged resolution times, flagging records that merit specialized investigation. For public policy makers, these results underscore the importance of awareness campaigns on the immediate reporting of disappearances, especially involving minors, and the need to strengthen institutional protocols for exceptional cases.

Nevertheless, this study also faced limitations. Although class imbalance was mitigated with SMOTE, recall levels for the most critical outcomes remain modest, indicating the need for further improvements through ensemble methods or cost-sensitive learning. In addition, regression models predicting resolution time proved unfeasible, reflecting the insufficiency of the current feature space. The dataset, while official and authoritative, lacks contextual variables such as investigative effort, socioeconomic conditions, media coverage, or a history of violence, factors that likely exert a decisive influence on the duration and outcome of cases.

In conclusion, this study establishes a fundamental quantitative baseline for the analysis of missing persons in Ecuador. It demonstrates that while directly predicting the most tragic outcomes remains a formidable challenge, machine learning approaches, especially those incorporating interpretability and anomaly detection, can extract valuable and actionable insights even from imperfect data. The potential of these tools to transform institutional responses is undeniable, but their full realization will depend on a sustained commitment to improving data collection, integrating richer contextual variables, and deploying predictive models into operational systems. This work is not an endpoint but a call to action to strengthen the analytical capabilities of the State and provide faster, more effective responses to one of the country's most urgent social problems.

6. Future work

Based on the findings and limitations of this study, future research will focus on three strategic axes. First, although the implementation of SMOTE significantly improved the detection of minority outcomes ("Missing" and "Deceased"), recall levels remain modest. Future work should therefore explore complementary strategies such as cost-sensitive learning, ensemble methods, or temporal validation schemes to further strengthen model robustness. Second, to overcome the infeasibility of predicting resolution time, it will be essential to enrich the current dataset with external contextual variables, such as indicators of investigative effort, media coverage, and socioeconomic factors, which could explain a greater proportion of the variance. Third, the analysis of existing data can be deepened by adopting a mixed-methods approach, combining quantitative modeling with qualitative evaluation of anomalous cases, and by applying Natural Language Processing (NLP) techniques to extract latent features from text fields. These combined efforts aim to develop a more robust predictive framework with greater operational utility for institutions responsible for the search and prevention of disappearances in Ecuador.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] D. Backer, T. Billing, Forecasting the prevalence of child acute malnutrition using environmental and conflict conditions as leading indicators, World Development 176 (2024) 106484.

[2] A. Dong, Y. Zhang, Z. Guo, P. Luo, Y. Yao, J. He, Q. Zhu, Y. Jiang, K. Xiong, Q. Guan, Predicting the locations of missing persons in china by using ngo data and deep learning techniques, International Journal of Digital Earth 17 (2024) 2304076.

[3] G. R. T. Paredes, W. E. R. Barreto, El rol del estado frente a las desapariciones forzadas de personas en ecuador, Revista Metropolitana de Ciencias Aplicadas 5 (2022) 229–238.

[4] M. A. P. Baldeón, Proceso de investigación de personas desaparecidas en ecuador, INNOVACIÓN & SABER 2 (2021) 68–75.

[5] H. R. A. Ortiz, R. A. P. Llerena, J. E. V. López, M. A. C. Pulloquinga, Las desapariciones en ecuador más allá de estadísticas, familias resilientes, inacción estatal, Revista Científica y Arbitrada de Ciencias Sociales y Trabajo Social: Tejedora. ISSN: 2697-3626 7 (2024) 166–182.

[6] J. Vera Macias, F. E. Alvarado Veliz, M. Arguello Arguello, L. J. Cevallos Torres, Neutrosophic multicriteria analysis of missing people in ecuador: An open data approach for the period 2024., Neutrosophic Computing & Machine Learning 37 (2025).

[7] M. J. Castillo Román, A. C. Espinosa Torres, J. M. González Arias, R. C. Giler Zambrano, W. L. Gallegos Pilligua, G. F. Robayo Cabrera, M. A. Quiroz Martinez, Examination of disappeared individuals in ecuador using orange data mining., Neutrosophic Computing & Machine Learning 37 (2025).

[8] E. X. Quito Recalde, J. M. Pando Farez, E. A. Riofrio Núñez, P. D. Ramírez Carrión, J. L. Pauta Riera, D. Rumbaut Rangel, Patterns and factors determinants in cases of missing persons in ecuador., Neutrosophic Computing & Machine Learning 37 (2025).

[9] G. Li, J. Wang, X. Wang, Construction and path of urban public safety governance and crisis management optimization model integrating artificial intelligence technology, Sustainability 15 (2023) 7487.

[10] C. Aldemir, T. Uçma Uysal, Artificial intelligence for financial accountability and governance in the public sector: Strategic opportunities and challenges, Administrative Sciences 15 (2025) 58.

[11] G. M. Campedelli, M. Bartulovic, K. M. Carley, Learning future terrorist targets through temporal meta-graphs, Scientific reports 11 (2021) 8533.

[12] Y. Qu, D. Tran, W. Ma, Deep learning approach to biogeographical ancestry inference, Procedia Computer Science 159 (2019) 552–561.

[13] E. Aiken, S. Bellue, D. Karlan, C. Udry, J. E. Blumenstock, Machine learning and phone data can improve targeting of humanitarian aid, Nature 603 (2022) 864–870.

[14] I. D. Apostolopoulos, P. P. Groumpos, Fuzzy cognitive maps: Their role in explainable artificial intelligence, Applied Sciences 13 (2023) 3412.

[15] A. Karatzogianni, Research design for an integrated artificial intelligence ethical framework, 2021.

[16] C. F. Luo, R. Bhambhoria, S. Dahan, X. Zhu, Prototype-based interpretability for legal citation prediction, arXiv preprint arXiv:2305.16490 (2023).

[17] A. Zambrano, E. Ortiz, X. Calderón, M. Zambrano, L. Urquiza-Aguiar, Sistema de localización de

personas desaparecidas basado en iot y cloud computing (caso de estudio: Quito), Revista Ibérica De Sistemas e Tecnologias De Informação (2019) 82–94.

[18] L. L. E. Berru, G. Y. V. Loaiza, Prescripción de acción en el delito de desaparición de personas en el código orgánico integral penal, Revista Científica Cultura, Comunicación y Desarrollo 8 (2023) 250–260.

[19] G. Gatti, I. Irazuzta, Diario de la desaparición mexicana. entre el precedente y el exceso, Disparidades. Revista de Antropología 74 (2019) e019–e019.

[20] A. Beduschi, Harnessing the potential of artificial intelligence for humanitarian action: Opportunities and risks, International Review of the Red Cross 104 (2022) 1149–1169.

[21] A. Serranò, El derecho de acceso a la justicia de las víctimas de violaciones de derechos humanos en perú: la corte interamericana de derechos humanos ante el indulto de alberto fujimori, Anuario iberoamericano de justicia constitucional (2021) 485–511.

[22] M. P. D. Masache, La vulneración del derecho a la verdad por exposición mediática en la investigación de casos de personas naturales reportadas como desaparecidas, Debate Jurídico Ecuador 4 (2021) 110–121.

[23] T. Zhang, M. Lauras, G. Zacharewicz, S. Rabah, F. Benaben, Coupling simulation and machine learning for predictive analytics in supply chain management, International Journal of Production Research 62 (2024) 8397–8414.

[24] A. Kjærum, B. S. Madsen, Pushing the boundaries of anticipatory action using machine learning, Data & Policy 7 (2025) e8.

[25] C. Christensen, T. Wagner, B. Langhals, Year-independent prediction of food insecurity using classical and neural network machine learning methods, Ai 2 (2021) 244–260.

[26] A. Fischmann, L. Z. Xu, Emotion regulation and cheap talk as signaling strategies: Evidence from crowdfunding for ukraine, Telematics and Informatics 87 (2024) 102086.

[27] H. Wilde, L. L. Chen, A. Nguyen, Z. Kimpel, J. Sidgwick, A. De Unanue, D. Veronese, B. Mateen, R. Ghani, S. Vollmer, A recommendation and risk classification system for connecting rough sleepers to essential outreach services, Data & Policy 3 (2021) e2.

[28] C. B. Casagran, G. Stavropoulos, Developing ai predictive migration tools to enhance humanitarian support: The case of eumigratool, Data & policy 6 (2024) e64.

[29] A. Berman, K. de Fine Licht, V. Carlsson, Trustworthy ai in the public sector: An empirical analysis of a swedish labor market decision-support system, Technology in Society 76 (2024) 102471.

[30] W. Zhai, A multi-level analytic framework for disaster situational awareness using twitter data, Computational Urban Science 2 (2022) 23.

[31] S. Lee, B. G. Kim, Attribute of big data analytics quality affecting business performance, Journal of Social Computing 4 (2023) 357–381.

[32] V. Shevchenko, A. Lukashevich, D. Taniushkina, A. Bulkin, R. Grinis, K. Kovalev, V. Narozhnaia, N. Sotiriadi, A. Krenke, Y. Maximov, Climate change impact on agricultural land suitability: An interpretable machine learning-based eurasia case study, IEEE Access 12 (2024) 15748–15763.

[33] L. K. Topham, W. Khan, D. Al-Jumeily, A. Waraich, A. J. Hussain, Gait identification using limb joint movement and deep machine learning, IEEE access 10 (2022) 100113–100127.

[34] S. Majumdar, A. Awasthi, From vulnerability to resilience: securing public safety gps and location services with smart radio, blockchain, and ai-driven adaptability, Electronics 14 (2025) 1207.

[35] S. H. Mohammed, A. Al-Jumaily, M. S. J. Singh, V. P. G. Jiménez, A. S. Jaber, Y. S. Hussein, M. M. A. K. Al-Najjar, D. Al-Jumeily, A review on the evaluation of feature selection using machine learning for cyber-attack detection in smart grid, Ieee Access 12 (2024) 44023–44042.

[36] F. Checchi, S. Frison, A. Warsame, K. T. Abebe, J. Achen, E. A. Ategbo, M. A. Ayoya, I. Kassim, B. Ndiaye, M. Nyawo, Can we predict the burden of acute malnutrition in crisis-affected countries? findings from somalia and south sudan, BMC nutrition 8 (2022) 92.

[37] M. Kooshki Forooshani, M. Van Den Homberg, K. Kalimeri, A. Kaltenbrunner, Y. Mejova, L. Milano, P. Ndirangu, D. Paolotti, A. Teklesadik, M. L. Turner, Towards a global impact-based forecasting model for tropical cyclones, Natural Hazards and Earth System Sciences 24 (2024) 309–329.