# Introducing a Data-Driven Solution for Predicting Causes of Death Among Missing Migrants with Real-World Impact

Md Siam
*Institute of Information Technology*
*University of Dhaka*
Dhaka, Bangladesh
bsse1104@iit.du.ac.bd

Md Arif Hasan
*Institute of Information Technology*
*University of Dhaka*
Dhaka, Bangladesh
bsse1112@iit.du.ac.bd

Shartaz Sajid Nahid
*Institute of Information Technology*
*University of Dhaka*
Dhaka, Bangladesh
bsse1123@iit.du.ac.bd

BM Mainul Hossain
*Institute of Information Technology*
*University of Dhaka*
Dhaka, Bangladesh
mainul@iit.du.ac.bd

***Abstract*—This paper addresses the challenge of improving predictive precision in data analysis, specifically in predicting the cause of death among missing migrants using the "Global Missing Migrants dataset". The study employs a meticulous methodology encompassing data preprocessing, refined feature selection, and thorough model optimization, with a primary focus on decision tree algorithms. The final decision tree model, after extensive hyperparameter tuning, attains an impressive accuracy of approximately 88%, with high precision and recall rates. By enhancing predictive precision through machine learning techniques, the paper provides actionable insights for policymakers and humanitarian organizations to improve safety measures and reduce migrant fatalities. This work underscores the significance of comprehensive data analysis and offers a practical framework for addressing the complex challenges faced by missing migrants. The outcomes of this paper not only contribute to the field of data analysis but also have real-world implications for safeguarding the lives of vulnerable migrant populations and informing future policies and interventions.**

***Index Terms*—Missing Migrants, Cause of Death, Predictive Modeling, Data Analysis, Humanitarian Issues**

## I. Introduction

The growing number of missing migrants is a serious problem with far-reaching effects. People who travel at great risk, frequently to escape extreme situations, encounter numerous risks while migrating. Unfortunately, a number of these people have unexpected and traumatic deaths, which emphasizes how important it is to identify the causes of these deaths.

The obvious lack of study dedicated to the field of data analysis and predictive modeling concerning the cause of death among missing migrants adds to the intensity of the situation. While previous research has focused on wider migration trends [1]–[3], and the obligations of state and government towards the missing migrants from a legal, policy, and psychological perspective [1], specific drivers of death in this vulnerable demographic have remained largely unknown. This huge gap in existing information highlights the need for a concerted effort to develop a model that might enhance our understanding of the events that led to these unfortunate results.

This research aims to bridge that information gap by presenting a thorough technique for greatly enhancing the accuracy of forecasting the causes of death among missing migrants. This methodology consists of some major components like data preprocessing, feature selection, model selection, and model optimization. We fine-tuned a decision tree model by expertly handling class imbalances, deleting redundant data, and picking predictive characteristics with care. This model is notable for its simplicity and precision, and it represents a substantial leap in the field.

The significance of our research extends beyond academic circles, as it offers invaluable insights for policymakers and humanitarian organizations. Our solution enhances our capacity to identify the causes of death among missing migrants, providing a data-driven foundation for the development of targeted interventions and safety measures. The combination of rigorous data processing, astute feature curation, effective encoding strategies, and the prudent selection and optimization of our model fills a significant research gap. This, in turn, provides a solid foundation for future studies and the development of policies aimed at addressing the complex challenges faced by missing migrants.

## II. Literature Review

The phenomenon of missing migrants is a pressing issue with far-reaching social, political, and humanitarian implications. Over the years, researchers and scholars have explored various aspects of this complex problem, ranging from the

classification of migration to the challenges faced by families of the missing migrants. While significant progress has been made in understanding this issue, there is a need for enhanced accuracy in forecasting the causes of death among missing migrants, particularly in cases where migrants go missing due to a wide range of circumstances. In this literature review, we delve into key studies that shed light on the phenomenon of missing migrants and identify the existing gaps in knowledge, which our paper aims to address.

The research of Sinha et al. [4] contributes to a foundational understanding of migration. It offers a concise yet informative definition of migration and provides a classification of migration based on various factors, such as political boundaries, length of time, distance, number of migrants, and decision-making approaches. This classification aids in categorizing and understanding the diverse forms of migration, laying the groundwork for more in-depth analyses.

Several studies have been conducted that focuses on the issue of missing migrants in detail. For example, Nyberg et al. [5] acknowledges that while missing migrants due to armed enforcement and related global laws are recognized, there is a significant gap in addressing situations where people go missing for other reasons, particularly those subject to enforced disappearances. This study emphasizes the need for a comprehensive understanding of the diverse causes of migrant disappearances, highlighting the complexities of the problem. Furthermore, Citroni et al. [6] explores the mechanics of migration in the context of Central America and Mexico. It examines the impacts and obstacles faced by families of missing migrants who are seeking information about the whereabouts of their loved ones and striving for justice and redress. This research offers insights into the challenges faced by these families and provides a regional perspective on the issue.

In the European context, the publication found at Eda et al. [1] scrutinizes state responses and duties concerning missing migrants and their surviving families. It adopts a multidisciplinary approach, considering legal, policy, and psychological aspects. This research underscores the importance of comprehensive state responses and the need to address the issue from various angles.

Moreover, Klochok et al. [3] addresses the impact of missing migrants on their families. This study delves into the profound emotional, psychological, and social effects that the disappearance of a loved one can have on families and highlights the importance of supporting these affected families.

These seminal works in the field of missing migrants have significantly advanced our understanding of the complexities and implications of this humanitarian issue. However, there exists a substantial research gap in the realm of forecasting the causes of death among missing migrants, particularly in cases where the reasons for disappearance are diverse and not adequately recognized. Our paper aims to bridge this information gap by presenting a thorough technique for greatly enhancing the accuracy of such forecasts. By doing so, we contribute to a deeper understanding of the missing migrant phenomenon and facilitate more effective responses to this pressing global challenge.

## III. Methodology

The methodology employed in this research was a systematic approach to improving the accuracy of a predictive model. The dataset primarily consisted of categorical variables, creating challenges related to data preprocessing, class imbalance, and feature selection. The steps undertaken are as follows:

### *A. Data Collection*

Our dataset comes from the "Missing Migrants Project" run by the International Organization for Migration (IOM) since 2014. This project aims to document the tragic deaths and disappearances of people during their journey to international destinations. However, collecting this data is quite challenging, leading to some undercounting. The dataset includes records of deaths at borders, transportation accidents, shipwrecks, violent attacks, and medical problems during journeys. It also accounts for cases where migrant bodies are found. The data is gathered from various sources, such as official records, media reports, NGOs, and surveys, offering a comprehensive view of the challenges faced by missing migrants.

### *B. Data Encoding and Initial Model Training*

Following data collection, the data underwent the process of encoding because of the presence of categorical variables. The initial data encoding involved using a label encoder, a common technique to transform categorical variables into numerical labels, making them suitable for machine learning algorithms. Recognizing the limitations of label encoding, particularly for categorical variables with numerous categories, a transition to dummy encoding was considered. Dummy encoding was applied to provide a more comprehensive representation of categorical data. [7] However, this transition did positively impact model performance but the impact was very little.

### *C. Data Balancing Strategies*

Class imbalance within the dataset posed a significant challenge. Data skewness and class imbalance were addressed to enhance model accuracy. Duplicate rows within the dataset were identified and removed, leading to data improvement. To mitigate the risk of biased model predictions, a combination of undersampling and oversampling techniques was tried and applied. [8]

### *D. Feature Selection and Refinement*

Feature selection and refinement were pivotal for improving model performance. An initial assessment of feature importance was conducted to identify the variables that had the most influence on the target variable. [9] Subsequently, a heat map was constructed to visualize feature correlations and guide the selection of relevant features. The refined feature set was employed to rerun the predictive model to evaluate the impact of feature selection on performance.

### E. Model Selection and Optimization

In the course of this research, the process of selecting and optimizing the machine learning model proved to be a critical factor in achieving the research goals. After careful consideration, the decision tree algorithm emerged as the optimal choice, primarily due to its compatibility with handling categorical data and its capacity for providing interpretability.

*1) Decision Tree Model:* The decision tree algorithm is defined as:

$$y = f(x, \theta)$$

Where: - $y$: The predicted value or class label - Cause of Death. - $x$: The input features - Migration route, Region of Origin, Region of Incident, Location of death, Information Source. - $\theta$: Model parameters and splitting criteria.

The model underwent an extensive optimization process, involving the fine-tuning of hyper-parameters essential for its performance. [10] Hyper-parameter tuning involves the adjustment of specific settings that govern how the machine learning algorithm functions, enhancing its ability to make accurate predictions. This optimization included the adjustment of parameters like tree depth (max-depth), minimum samples required for node splitting (min-samples-split), and the splitting criterion (e.g., Gini impurity or entropy).

*2) Hyperparameter Tuning:* The hyperparameter tuning process is mathematically represented as:

$$OptimizedModel = f(Hyperparameters, Data)$$

Where: - $OptimizedModel$: The final decision tree model with tuned hyperparameters. - $Hyperparameters$: Parameters like max-depth, min-samples-split, etc. - $Data$: The training data used to train the model.

In tandem with hyperparameter tuning, cross-validation techniques were utilized to ensure a

*3) Cost-Complexity Pruning:* Cost-Complexity Pruning, often referred to simply as Pruning, is a fundamental technique in machine learning, particularly for decision tree algorithms. [11] Its primary purpose is to prevent overfitting by simplifying a decision tree, thus improving its generalization capabilities. This process entails systematically removing branches from the tree to reduce its complexity and enhance its predictive performance.

In the context of a decision tree, Cost-Complexity Pruning can be mathematically represented as follows:

$$R_\alpha(T) = R(T) + \alpha|T|$$

Where: - $R_\alpha(T)$ represents the cost-complexity measure of tree $T$ after pruning. - $R(T)$ denotes the misclassification error or impurity measure of the original tree. - $\alpha$ is a hyper-parameter that regulates the trade-off between tree simplicity and accuracy. - $|T|$ signifies the number of terminal nodes (leaves) in the tree $T$.

By varying the hyper-parameter $\alpha$, the trade-off between simplicity and accuracy can be controlled, enabling the creation of pruned decision trees that better align with the data and avoid overfitting.

*4) Gini Impurity and Entropy:* Gini Impurity and Entropy are two essential metrics employed in decision tree algorithms to assess the impurity or disorder of a dataset. These metrics serve as criteria for selecting the best attribute to split the data at each node in a decision tree. They are instrumental in making decisions that result in well-structured and informative trees.

The Gini Impurity ($Gini(T)$) is mathematically represented as:

$$Gini(T) = 1 - \sum_{i=1}^{c} (p_i)^2$$

Where: - $c$ is the number of classes or categories. - $p_i$ represents the proportion of instances belonging to class $i$ in the data-set.

A lower Gini Impurity score indicates a purer data-set with a lower level of disorder, making it an attractive metric for classification tasks.

Entropy ($Entropy(T)$) is another metric used in decision trees and is calculated as:

$$Entropy(T) = -\sum_{i=1}^{c} p_i \cdot \log_2(p_i)$$

Where: - $c$ is the number of classes or categories. - $p_i$ is the proportion of instances belonging to class $i$.

Entropy measures the level of disorder or impurity in a data-set. Lower entropy values indicate a more organized and homogeneous data-set.

Both Gini Impurity and Entropy are pivotal in the decision-making process of decision trees. By assessing these metrics at each node, the algorithm can determine the attribute that results in the purest child nodes, leading to effective data partitioning for accurate classification.

### F. Performance Metrics

To identify the hyper-parameter combinations that maximize model accuracy while mitigating over-fitting, the research leveraged a set of essential performance metrics, including:

- Accuracy: $\frac{\text{True Positives+True Negatives}}{\text{Total}}$
- Precision: $\frac{\text{True Positives}}{\text{True Positives+False Positives}}$
- Recall: $\frac{\text{True Positives}}{\text{True Positives+False Negatives}}$
- F1 Score: $\frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{\text{Precision+Recall}}$

### G. Model Comparison and Selection

It is noteworthy that the research did consider the random forest algorithm; however, it was observed that this alternative did not yield a substantial improvement over the decision tree model. Consequently, the decision tree model was deemed the most suitable choice, primarily because of its demonstrated ability to achieve the desired predictive accuracy. [12]

By considering these essential formulas and mathematical concepts, the research was able to systematically select and optimize the decision tree model for its specific research objectives. This methodological approach ensured that the decision tree model was fine-tuned to its best possible performance and validated through rigorous evaluation processes.

TABLE I: Global Missing Migrants Dataset Column Details with Acronyms

| Type | Variable Name | Acronym | Description |
|---|---|---|---|
| Predictor | Incident Type | ti | Type of the incident |
| | Incident Year | iy | The month in which the incident occurred. |
| | Reported Month | rm | The year in which the incident occurred. |
| | Region of Origin | ro | Region of origin of the decedent(s). In some incidents, region of origin may be marked as "Presumed" or "(P)" if migrants traveling through that location are known to hail from a certain region. |
| | Region of Incident | ri | The region in which an incident took place. |
| | Country of Origin | co | Country of birth of the decedent. |
| | Number of Dead | ndead | The total number of people confirmed dead in one incident, i.e. the number of bodies recovered. |
| | Minimum Estimated Number of Dead and Missing | mendm | The total number of those who are missing and are thus assumed to be dead. This variable is generally recorded in incidents involving shipwrecks. |
| | Total Number of Dead and Missing | tndm | The sum of the 'number dead' and 'number missing' variables. |
| | Number of Survivors | ns | The number of migrants that survived the incident, if known. |
| | Number of Females | nf | Indicates the number of females found dead or missing. |
| | Number of Males | nm | Indicates the number of males found dead or missing. |
| | Number of Children | nc | Indicates the number of individuals under the age of 18 found dead or missing. |
| | Migration Route | mr | Name of the migrant route on which incident occurred, if known. |
| | Location of Death | ld | Place where the death(s) occurred or where the body or bodies were found. Nearby towns or cities or borders are included where possible. |
| | Coordinates | coord | Place where the death(s) occurred or where the body or bodies were found. |
| | Information Source | is | Name of source of information for each incident. Multiple sources may be listed. |
| | UNSD Geographical Grouping | ggroup | Geographical region in which the incident took place, as designated by the United Nations Statistics Division (UNSD) geoscheme. |
| Response | Cause of Death | cd | The determination of conditions resulting in the migrant's death i.e. the circumstances of the event that produced the fatal injury. |

## IV. Data

The name and description of the predictors and the response are given in Table I. The data-set utilized in this study originates from the Missing Migrants Project, an initiative led by the International Organization for Migration (IOM) since 2014. This comprehensive data-set meticulously documents the tragic experiences of migrants as they embark on perilous journeys towards international destinations. It is important to note that the data-set encompasses those migrants who have met with unfortunate fates at the external borders of states or during their migration processes to foreign lands, irrespective of their legal status. Specifically, the data includes records of migrants who have lost their lives due to transportation accidents, shipwrecks, violent attacks, or medical complications encountered en route. Additionally, the data-set accounts for unidentified individuals found in areas associated with migration routes, allowing for the inclusion of such cases. Conversely, the data-set excludes deaths occurring within immigration detention facilities, post-deportation in the migrant's homeland, or those loosely related to irregular migrant status, such as fatalities resulting from labor exploitation. Furthermore, migrants who have established new residences and subsequently meet unfortunate ends are not part of this data-set.

Within our data-set, encompassing 19 distinct columns, we aimed to predict the cause of death. Our initial step involved examining the correlation heat-map of these features, as illustrated in Figure 2. This analysis shed light on the relationships between these features and the cause of death. In this correlation assessment, we considered values on a scale from -1, indicating a high negative correlation, to 1, signifying a high positive correlation, with 0 representing no correlation. The correlation analysis revealed the following associations with the cause of death: "it" (-0.01), "iy" (-0.088), "rm" (0.022), "ro" (-0.22), "ri" (0.31), "ro" (-0.061), "ndead" (-0.057), "mendm" (-0.15), "tndm" (-0.15), "ns" (-0.075), "nf" (-0.08), "nm" (-0.061), "nc" (-0.04), "mr" (0.11), and "ld" (0.0066). In light of these correlations, we opted to focus on the two most prominent features, namely "ri" (region of incident) and "mr" (migration route), as these displayed the strongest associations with the cause of death.

In our feature selection process, we extended our analysis beyond the initial heatmap exploration. We considered all the features at our disposal and employed a decision tree algorithm to determine the most significant predictors within the feature set. The decision tree analysis highlighted several important features, which included "Incident type," "Region of origin," "Number of females," "Number of children," and "Information

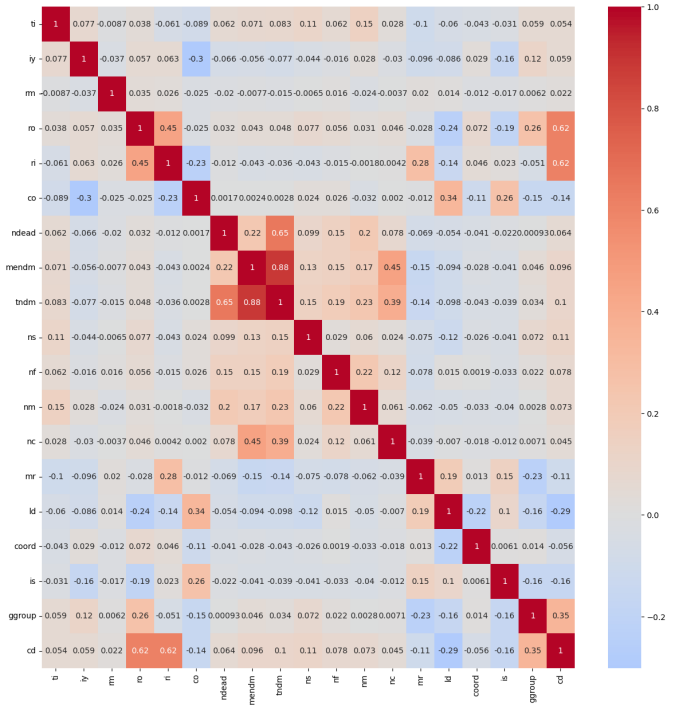



Fig. 1: Correlation heat-map of the data-set for all the features

source," in that order.

However, upon closer examination, it became evident that the distribution of the "Incident type" feature was highly skewed [Fig 3]. Among the 13,020 recorded occurrences of "Incident type," an overwhelming 12,670 belonged to a single class, while the remaining three classes collectively represented a mere 350 occurrences. This skewed distribution raised concerns regarding potential bias and led to the decision to exclude "Incident type" as an important feature in our predictive model.

A similar pattern was observed with the "Number of females" and "Number of children" features. Consequently, these features were also deemed unsuitable as significant predictors. After careful consideration, we elected to retain "Region of origin," "Region of incident," "Migration route," and "Information source" as the final set of predictor features for our cause of death prediction model.

In the course of our analysis, it was evident that the features we intended to use for prediction, namely "Region of origin," "Region of incident," "Migration route," and "Information source," were categorical in nature. Before employing these features in our predictive model, it was imperative to convert them into a format suitable for analysis. To achieve this, we opted for label encoding, a technique that transformed the categorical values into corresponding integers.

Initially, when we employed a decision tree based on the label-encoded data, the model's performance yielded unsatisfactory results, with notably low accuracy. A more in-depth analysis of the data-set revealed skewed class distributions within the target feature. To address this imbalance, we decided to focus on the top five classes of the target feature and retrained the decision tree. Despite this effort, the model's performance remained sub-optimal. As a solution, we resorted to oversampling the minority classes, effectively re-balancing the class distribution. In our pursuit of a solution, we recognized the presence of 621 duplicate entries within the data-set. To rectify this issue, we systematically removed these duplicate records and subsequently reran the algorithm. This cleansing process notably enhanced the accuracy of our model, resulting in a significant improvement and achieving an acceptable level of performance.

However, the interpret-ability of the decision tree generated from this over-sampled data remained challenging. Consequently, we revisited the encoding technique and transitioned to one-hot encoding. This change not only led to a more interpret-able model but also improved the model's accuracy, providing more reliable results. Moreover, optimizations such as hyper-parameter tuning and cost complexity pruning helped us improve the accuracy of the model.

## V. Result Analysis and Discussion

### A. *Impact of Data Prepossessing*

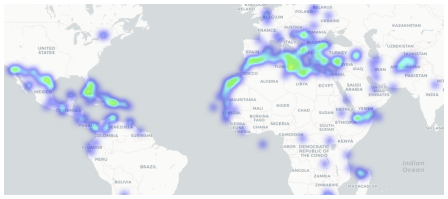
In our quest to classify the Missing Migrants data-set with a focus on predicting the "Cause of Death," we embarked on a comprehensive analysis, adopting various techniques and strategies to enhance our classification model. Our journey began with addressing class imbalance through the use of oversampling and under-sampling alongside a Decision Tree classifier.

However, the analysis revealed an unexpected outcome, where a feature importance approach with a focus on the top 5 features did not result in performance improvements; instead, it led to a minor dip in overall accuracy.

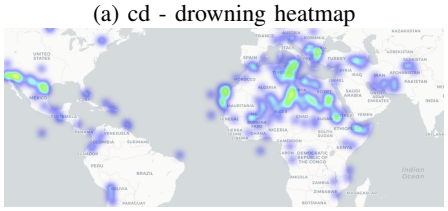
To refine our model, we transitioned to heat-map-based feature selection, culminating in the selection of the most pertinent features, including 'Migration route,' 'Region of Origin,' 'Information Source,' 'Region of Incident,' and 'Location of death.' We notably excluded 'Incident Type' due to skewness concerns that could introduce bias. Additionally, we delved into encoding techniques, observed issues with one-hot encoding, and addressed class imbalance by exploring oversampling methods. The identification and removal of duplicate columns further enhanced our prepossessing steps. This analysis collectively informs our understanding of how these measures impact predictive accuracy and offers valuable insights for similar classification tasks in the future.

### B. *Feature Importance Analysis*

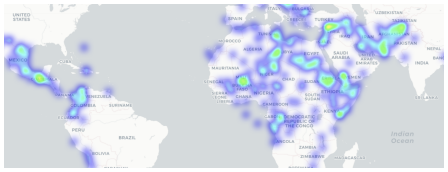
The feature importance analysis reveals the significance of various factors in predicting the cause of death among missing migrants. Incident Type emerges as the most influential predictor (Importance: 0.476), followed by Migration Route (Importance: 0.190), Region of Origin (Importance: 0.133), Number of Survivors (Importance: 0.056), and Total Number of Dead and Missing (Importance: 0.038514).



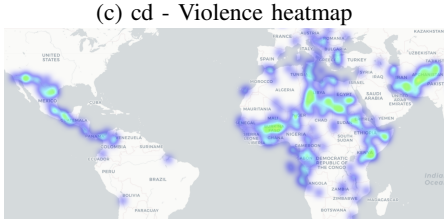
(a) cd - drowning heatmap



(b) cd - Harsh environmental conditions heatmap



(c) cd - Violence heatmap



(d) cd - Sickness/lack of access healthcare heatmap

Fig. 2: Geographical Heat-map of Missing Migrants Grouped by Cause of Death

Lower-importance features encompass Total Number of Dead and Missing, Number of Males, Number of Females, Number of Children, Number of Dead, and Minimum Estimated Number of Missing, each contributing with lesser impact. This analysis provides valuable insights into the determinants of cause of death among missing migrants, underscoring the pivotal roles of Incident Type and Migration Route within this humanitarian context.

### *C. . Refining Feature Selection Based on Heat-map Analysis*

In our quest to refine feature selection for predicting 'Cause of Death' among missing migrants, initial feature importance analysis didn't improve accuracy. So, we turned to heat-map analysis, identifying 'Migration route,' 'Region of Origin,' 'Information Source,' 'Region of Incident,' and 'Location of death' as the key features. We excluded 'Incident Type' due to its skewed behavior and potential bias concerns, aiming for an equitable representation of features. This strategic selection enhances accuracy and reliability in predicting the cause of death for this humanitarian issue, emphasizing the critical role of specific features in this context.

### *D. Impact of Encoding Techniques*

Our research delved into the impact of encoding techniques on model performance, specifically exploring label encoding and one-hot encoding. The outcome was unexpected, as one-hot encoding resulted in a significant accuracy drop, with the model achieving only 46.12% accuracy. This drop can be attributed to the complex nature of the data and the intricacies of the classification problem. One-hot encoding, which creates binary columns for each category within a feature, substantially increased the data set dimensional, challenging the model's ability to discern meaningful patterns within the data. Moreover, the resulting sparse matrix posed challenges to the model's generalization capabilities. This encoding dilemma is further evident in the classification report, where various classes exhibited low precision, recall, and F1-scores, emphasizing the model's struggle in making accurate predictions. In conclusion, our study underscores the vital role of selecting an appropriate encoding method, tailored to the data and the specific problem context. The unexpected accuracy drop with one-hot encoding highlights the need for careful consideration of encoding techniques during data prepossessing to optimize model performance effectively.

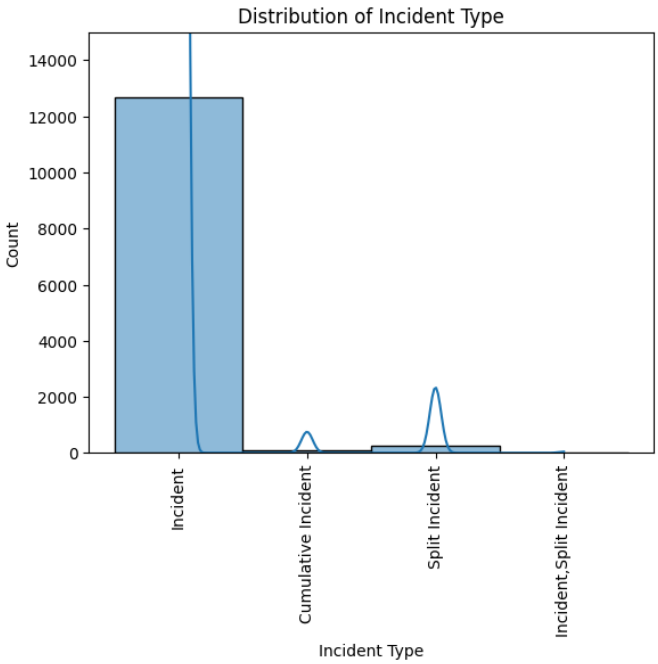



Fig. 3: Distribution of Incident Type

### *E. Handling Class Imbalance and Sample Size Variations*

Our research undertook a multifaceted approach to handle the issues of class imbalance and sample size variations within the dataset. This process involved two primary steps:

1. Identification of Duplicate Records: During data preprocessing, we identified and removed 641 duplicate records present in the dataset. The existence of duplicate records can introduce bias and adversely impact the model's performance.

2. Balancing Sample Sizes: To mitigate the challenges associated with class imbalance and varying sample sizes, we implemented oversampling as a key strategy. This technique involves generating additional instances of the minority class to align its size more closely with that of the majority class. By doing so, we aimed to enhance the model's capacity to make precise predictions, particularly for the minority class.

Our focus on eliminating duplicate records and incorporating oversampling aimed to create a more balanced class distribution. This, in turn, sought to enhance the model's overall performance and effectively address the inherent issues of class imbalance and sample size variations. The analysis of how these measures impacted the model's accuracy and predictive capabilities is crucial to assessing their effectiveness in this context.

### *F. Effect of Duplicate Column Removal*

641 duplicate columns were identified in the data-set. These duplicates were systematically removed to enhance model performance. This step contributed to improved accuracy, with the Decision Tree classifier achieving approximately 88% accuracy. The removal of duplicate columns streamlined the data-set and demonstrated the significance of data prepossessing in enhancing model accuracy.

### *G. Hyper parameter Tuning and Model Performance*

The final Decision Tree model underwent a meticulous tuning process, where we optimized key hyper parameters, including 'max-depth: None,' 'max-features: sq-rt,' 'min-samples-leaf: 1,' and 'min-samples-split: 2.' Additionally, we identified the best cost-complexity pruning alpha (ccp-alpha) as 0.0, a critical aspect of model refinement.

Notably, the model exhibited a remarkable leap in performance, culminating in an impressive accuracy of approximately 88%. The accompanying classification report for diverse classes indicates robust precision, recall, and F1-score values. Table II shows the performance metrics obtained by our model. These findings collectively affirm the model's competence in effectively classifying the causes of death among missing migrants, emphasizing its practical utility.

The obtained high accuracy and resilient performance underscore the pivotal role of meticulous data prepossessing, the judicious selection of features, and the fine-tuning of model parameters. This collaborative effort results in a Decision Tree classifier with substantial predictive power, offering an invaluable instrument for comprehending and addressing the multifaceted challenges confronting missing migrants.

This outcome stands as a testament to the successful synthesis of data prepossessing, feature engineering, and model optimization, further advancing our capacity to engage with and mitigate the humanitarian issues associated with missing migrants.

### *H. Interpreting Model Performance in the Context of Missing Migrants*

The enhanced model, utilizing a refined data set and advanced encoding techniques, underscores its increased relevance and practical utility in the context of missing migrants and the Missing Migrants Project.

| | **precision** | **recall** | **f1-score** | **support** |
|---|---|---|---|---|
| **accuracy** | | | 0.88 | 9879 |
| **macro avg** | 0.89 | 0.88 | 0.88 | 9879 |
| **weighted avg** | 0.88 | 0.88 | 0.88 | 9879 |

TABLE II: Performance Metrics

Data-Driven Insights: Utilizing a data-driven approach, our refined model, which centers its analysis on critical attributes such as 'Migration route,' 'Region of Origin,' 'Cause of Death,' 'Region of Incident,' 'Location of death,' and 'Information Source,' emerges as a valuable source of insights into the causes of death among missing migrants. By focusing on these essential variables, the model offers highly relevant information for comprehending the challenges and perils encountered by migrants.

Tragically, 2797 migrants lost their lives due to drowning, the geograohical heatmap can be found in Fig.2(a). This perilous fate was most pronounced along the Libya to Tunisia to Greece route, claiming the lives of 1335 individuals. Additionally, the Mexico to the USA route saw 877 fatalities, while the Morocco to Spain route recorded 585 deaths. Harsh environmental condition resulted in 815 fatalities, with the majority occurring on the Mexico to USA route (643), and the remaining cases transpiring on the Chad to Libya route (172), which is depicted in Fig.2(b). Fig.2(c) represents the fatalities stemming from sickness numbered 487, with 232 tragic deaths on the Sudan to Eritrea route, 142 on the Afghanistan to Iran route, and 113 on the Afghanistan to Tajikistan route. The loss of 324 lives was attributed to violence, with the most significant counts observed on the Syria to Turkey route (182). Furthermore, the Mexico to Guatemala route experienced 128 fatalities, while the Myanmar to Bangladesh route reported a smaller count of 14, according to the repressentation of Fig.2(d).

These counts underscore the profound challenges and risks faced by migrants on their arduous journeys. They emphasize the vital role of humanitarian efforts in improving the safety and well-being of these individuals, further underscoring the urgency of addressing the complex issues surrounding migration.

Significant Accuracy Gain: With the model optimization efforts, we have realized a substantial boost in accuracy, reaching approximately 88.9%. This elevated level of accuracy equips the model with exceptional performance in classifying diverse causes of death, accompanied by high precision and recall rates across multiple classes. This improved accuracy is invaluable for the Missing Migrants Project, as it empowers the project to more effectively identify and respond to the causes of death.

Practical Applications: The model's utility extends to a broad spectrum of stakeholders, including migrant welfare organizations, policymakers, and researchers. By effectively identifying causes of death, it serves as a foundational tool in

shaping policies and interventions that enhance migrant safety and diminish mortality rates.

Enhanced Understanding: Beyond accurate classification, the model delves deep into the intricate web of issues surrounding missing migrants. It unravels the influence of various factors such as migration routes and regions of origin, thus enabling stakeholders to make well-informed decisions and strategies.

Improved Humanitarian Efforts: At its core, this research strengthens humanitarian endeavors, focusing on preserving the lives and welfare of migrants. The insights gained are instrumental in preventing future migrant fatalities and bolstering support systems for those traversing perilous journeys.

The refined data-set and encoding techniques have amplified the model's relevance and potential impact within the Missing Migrants Project, equipping it to deliver actionable insights for the reduction of missing migrants and the advancement of safety and well-being for individuals embarking on precarious odysseys.

### *I. Limitations and Future Directions*

In the course of our analysis, several critical considerations emerged. Firstly, it's imperative to recognize potential data quality concerns, including inaccuracies and underreporting, which may impact result interpretation. Second, addressing possible biases stemming from uneven representation of incidents or regions is essential for a more robust analysis. Moreover, future research endeavors could expand the scope of features for a more comprehensive investigation. Continuous efforts to refine data collection methods are crucial to ensure the accuracy of insights. Exploring advanced modeling techniques, such as Random Forest and deep learning, may offer opportunities to enhance predictive accuracy. Interdisciplinary collaboration with experts from various fields can enrich our understanding of the subject matter. Additionally, incorporating spatial and temporal analysis can uncover valuable patterns. Ethical considerations should underpin our research to ensure the dignity of individuals is respected. Lastly, forging partnerships with migrant welfare organizations can facilitate the translation of research findings into practical policy initiatives.

## VI. Conclusion

The research objective was to enhance our capacity to identify the causes of death among missing migrants, leveraging data from the Missing Migrants Project. This involved a meticulous methodology spanning data preprocessing, feature selection, and model optimization. Key features influencing predictive accuracy were systematically identified, while measures to ensure data integrity included addressing class imbalance and removing duplicate records. Following rigorous tuning, the decision tree model delivered substantial accuracy improvements.

In summary, the study underscores the critical role of rigorous data analysis, selective feature choice and encoding techniques. It provides actionable insights for policymakers and humanitarian organizations striving to address the multifaceted challenges faced by missing migrants, ultimately contributing to their improved safety and well-being. This work also sets the stage for future research, promoting advancements in modeling techniques with the goal of further reducing migrant fatalities.

## References


[1] L. N. Eda, "Missing migrants: legal obligations and psychosocial implications for families." Ph.D. dissertation, Bournemouth University, 2021.

[2] J. Sarkin, "Respecting and protecting the lives of migrants and refugees: the need for a human rights approach to save lives and find missing persons," *The International Journal of Human Rights*, vol. 22, no. 2, pp. 207–236, 2018.

[3] G. Klochok and C. Herrera-Espiñeira, "The grief of relatives of missing migrants and supportive interventions: a narrative review," *Clinical Nursing Research*, vol. 30, no. 7, pp. 1023–1029, 2021.

[4] B. Sinha, "Human migration: concepts and approaches," *Foldrajzi Ertesito*, vol. 3, no. 4, pp. 403–414, 2005.

[5] N. Nyberg Sørensen and L. Huttunen, "Missing migrants and the politics of disappearance in armed conflicts and migratory contexts," *Ethnos*, vol. 87, no. 2, pp. 321–337, 2022.

[6] G. Citroni, "The first attempts in mexico and central america to address the phenomenon of missing and disappeared migrants," *International Review of the Red Cross*, vol. 99, no. 905, pp. 735–757, 2017.

[7] C. Seger, "An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing," 2018.

[8] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, p. 20–29, jun 2004.

[9] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *2014 Science and Information Conference*, 2014, pp. 372–378.

[10] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020.

[11] A. L. Prodromidis and S. J. Stolfo, "Cost complexity-based pruning of ensemble classifiers," *Knowledge and Information Systems*, vol. 3, no. 4, pp. 449–469, Nov 2001.

[12] K. Kim and J. sik Hong, "A hybrid decision tree algorithm for mixed numeric and categorical data in regression analysis," *Pattern Recognition Letters*, vol. 98, pp. 39–45, 2017.